# 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)

## Abstract

Predicting bacterial phenotype from external perturbations is a problem investigated many times; on the other hand, the question of predicting external conditions by using phenotype is not investigated much. Here we use an *E.coli* transcriptomic and proteomic dataset to predict the growth conditions. Our findings indicate that we can make far from random predictions related with carbon source, salt levels and growth phase of the bacteria from both transcriptome and proteome data. We can also predict multiple conditions for a sample up to 75% accuracy. There is a small synergistic effect related to combining transcriptome and proteome data. It is also possible to make predictions for continuous parameters like the time that the sample was taken and salt concentrations. The analysis indicates two clear outcomes; exponential data have a higher predictive value and combining mRNA and protein data results in a small but consistent increase in prediction power independent of the machine learning model used.

## Introduction

Figure out the changes in the organisms, while responding to external conditions is a big challenge. This problem has two distinct directions the first and very well investigated question asks whether we can predict the changes in the organism with respect to external conditions. The second and not that much-investigated condition asks the opposite question; whether the growth conditions can be predicted by using information from the cell content.

Another component of the challenge is to build the link between external conditions and cell content by using multiple layers of genomic data. Adding more layers supposed to give better outcomes, but there are multiple challenges based on biases of distinct methods used to collect information1 from different -omics layers and biases related with batches which are inevitable 2,3 if one runs a study based on a big dataset that is generated in multiple labs within a long time span.

There are some studies that focus on predicting external conditions by using the cells internal variables4,5, but the main focus of those studies is to generate a theoretical framework for the problem. Here we use a large dataset6 composed of transcriptomic and proteomic data to predict the growth conditions. Our unique dataset that includes paired mRNA and protein measurements of *E.coli* under multiple different growth conditions used to train machine learning algorithms with the aim of predicting the growth conditions of the bacteria. This large paired database also allows us to compare the predictability of external conditions by using transcriptomic and proteomic datasets either individually or together. By using the framework, we also asked which changes in the environment have the biggest impact on the organism under investigation in transcriptomics and proteomics levels. We observe by combining transcriptomics and proteomics datasets we obtain more predictability associated with external conditions. In addition to those we also investigate the parameters that increase and decrease the predictability of those conditions and figure out bacterial phase has a huge effect on predictability.

## Results

### Data structure and pipeline design

We used a previously published 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 asked here to what extent these parameters could be predicted through machine learning from the mRNA abundances, the protein abundances, or both.

We employed four different machine learning models, three based on SVMs with different kernels (radial kernel, sigmoidal kernel, and linear kernel) and the fourth using random forests. We used *C-Classification* for training classification model and *eps-regression* for training regression models which we use for predicting growth rates. We adjusted weights of samples in a way that each class ends up with equal weight in order to prevent the prediction bias in favor of more populated classes.

We also generate a tuning loop to fine tune the parameters that the machine learning algorithms use. We optimize the parameter "cost" for three models; SVM with linear, radial, and sigmoidal kernel. We optimize the parameter "gamma" for SVM with radial and sigmoidal kernels. For random forest algorithm, we optimize three parameters; "mtry", "ntrees", and "nodesize". We use the multi-conditional f1 score10, in order to weight all conditions equally and assign same importance to false positives and false negatives.

Before SVM we apply DeSeq211 for size factor normalization, fSVA12 to normalize batch effects and PCA13 to obtain the principal components of the data. Despite the fact that our dataset has strong batch effects 6, we believe we get rid of the batch effects as much as possible before training our data with the help of the fSVA algorithm. We write our tuning algorithm that enables us to divide the dataset into subsets semi-randomly in a way that the ratios of samples tried to be constant as much as possible between training tuning and test sets. We also calculate the conditional class weights for each training data in tuning process [figure 2].

So overall our pipeline is designed for tuning four different models, SVM with the linear kernel, SVM with the radial kernel, SVM with the sigmoidal kernel, and random forest; independent of weight and batch effects and by using principal components in order to prevent overfitting and assigning same importance to false positives, false negatives through all individual conditions.

### 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 considered growth condition as described by 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, 16 distinct combinations of these four variables were present in the data (Figure 1).

We trained the models to predict the entire four-dimensional condition vector at once for a given sample, and we used the multi-conditional *F*1 score10 to assess 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. We first 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 fitted to mRNA data, and the random forest model outcompeted the other models (though by a smaller margin) when fitted to protein data (Table 1). The best parameter sets after tuning for mRNA and protein samples are xx, yy as cost and xx, yy as gamma, for radial and sigmoid kernels respectively [Supplementary figure 1-2].

We next compared the *F*1 scores we obtained when the tuned models were used to make predictions on the test datasets. For predictions based on mRNA abundances, we found that model performance on test data was general quite good, with *F*1 scores centered around 0.7 (Figure 3). The *F*1 scores were virtually identical for the three SVM based models, independent of the chosen kernel, and they were somewhat lower for the random forest model. For predictions based on protein abundances, model performance on test data was slightly worse than for the mRNA case, but still very good. The SVM with a radial kernel performed the best and the random forest model performed the worst.

In addition to assessing overall prediction performance with *F*1 scores, we also recorded the percentage of times specific growth conditions were accurately predicted or mispredicted in a confusion matrix (Figure 4). The large numbers/dark colorings along the diagonal highlight that predictions were generally accurate. For example, for predictions made from mRNA abundances, stationary-phase samples are frequently mispredicted as having glucose as carbon source, base Na+ concentration, and either base or low Mg2+ concentration. These mispredictions are driven in part 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).

### Prediction accuracy improves when predictions are made jointly from mRNA and protein abundances

We next asked whether predictions could be improved by considering mRNA and protein abundances at the same time. To address this question, we focused on the subset of 102 samples for which both mRNA and protein abundances were available, and run our analysis pipeline three times, once for mRNA abundances only, once for protein abundances only, and once for the combined dataset containing both mRNA and protein abundances. First, we found that for all four machine-learning algorithms, protein abundances alone yielded significantly better predictions than mRNA abundances alone (Figure 5 and Table 2). Second, predictions made with the combined dataset were consistently better than predictions made using either mRNA or protein abundances alone, irrespective of machine-learning algorithm used (Figure 5 and Table 2).

When considering the confusion matrices for the three scenarios, we found that many of the mispredictions arising from mRNA abundances were not that common when using protein abundances and vice versa (Supplementary Figures 4 and 5). For example, when using mRNA abundances, many conditions were mispredicted as exponential phase, glycerol, base Mg2+, base Na+, or as stationary phase, glucose, base Mg2+, high Na+, but these particular mispredictions were rare or absent when using protein abundances. By contrast, when using protein abundances, a few conditions were mispredicted as stationary phase, glycerol, base Mg2+, base Na+, and these mispredictions were virtually absent when using mRNA abundances. For predictions made from the combined dataset, mispredictions unique to either mRNA or protein abundances were generally suppressed, and only those mispredictions that arose for both mRNA and protein abundances alone arose also in the combined dataset (Supplementary Figure 6).

### Prediction accuracy differs among conditions and data source

We also analyzed how prediction accuracy differed among conditions. First, considering only samples that had either exponential or stationary phase samples for both protein and mRNA data, 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 or protein abundances) used (Figure 6). This observation implies that *E. coli* gene expression levels are less affected by growth conditions when the bacteria are in stationary phase than when they are in exponential phase.

We also found that when making predictions based on mRNA abundances, growth phase was most accurately predicted and carbon source least accurately, with Mg2+ and Na+ concentration falling between these two extremes. By contrast, when making predictions based on protein abundances, carbon source was most accurately predicted and Mg2+ concentration least accurately, with Na+ concentration and growth phase falling between these two extremes (Figure 7 and Supplementary Figure 7). These results are consistent with the clustering observed among individual conditions in a previous analysis of the same dataset6. Finally, for the combined data set, we found that for carbon source and Mg2+ concentration, where protein abundances yielded better predictions than mRNA abundances, prediction accuracy generally fell between the prediction accuracies of the individual datasets (Supplementary Figure 8), whereas for Na+ concentration and growth phase, where mRNA abundances yielded better predictions than protein abundances, prediction accuracy was generally as good as or better than between the prediction accuracies of the individual datasets (Supplementary Figure 8).

## Discussion

We try to find methods to classify growth conditions of bacteria by analyzing its mRNA and protein concentrations, this is the inverse problem of finding concentrations of specific mRNAs and proteins under given growth conditions. Overall results indicate that we can predict the growth conditions up to xx percent of the time when we try multi-category classifications. We used the dataset6, which includes samples that have paired mRNA and protein reads for our study.

We can predict the external conditions that the bacteria growth even after batch effect correction, and even just by using 152 samples for mRNA data and 105 samples for protein data and for 16 distinct conditions. This indicates *E.coli* strongly responses to external conditions in a measurable and consistent way.

Results suggest that protein concentrations include more information than mRNA concentrations and combined data includes more information than proteins after the data is filtered in order to keep only paired mRNA and protein samples. If we equate the number of samples, protein data includes more information than mRNA data; but mRNA data still have some information about external conditions that protein data do not have so that the combined dataset can make stronger predictions than both from mRNA and protein data alone. This difference indicates the post-translational regulation is controlled by external conditions up to a degree. Increase in predictive power on internal cellular state after combining multiple omics datasets have been theoretically hypothesized 14 and experimentally observed before1,15,16. Post-transcriptional regulation that triggered by an external stimulus that aims to arrange cells response time to external stimuli might be a result of post-translational phosphorylation17, or might be associated with stress coping mechanisms18.

The prediction power of the tests decrease from exponential to stationary phase, which indicates the concentration of mRNA and proteins become indistinguishable from one external condition to another, this behavior is independent of external conditions or the machine learning algorithms used. One reason for this behavior might be associated with endogenous metabolis; 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 provided carbon source, suppress the differences between the cells in different external carbon source environments19,20. Another reason for this behavior might be related to strong coupling between gene expression noise and growth rate; multiple studies conclude higher gene expression noise is associated with in lower growth rate which might be a survival strategy in harsh enviroments21. Negative correlation between population average gene expression and noise was shown for *E. coli* and *Saccharomyces cerevisiae*22,23. Positive correlation between growth rate and population average gene expression was also shown for micro-organisms 24–26. The direct consequence of those two is the negative correlation between growth and noise.

Study of the relation between the predictability of specific external conditions with mRNA and protein abundances generate two statistically significant results. First; growth phase can be reliably predicted from mRNA concentrations, stronger than any other condition that we investigate, on the other hand, the predictability of the growth phase from protein concentrations are limited. The reason for this might be related with different life-cycles of mRNAs and proteins27. Because of low degradation rates of proteins, a large portion of stationary phase protein composition of microorganisms was transcribed in exponential phase. This increases the similarity between exponential and stationary phase protein compositions which might decrease the predictability of growth phase from protein data. On the contrariety, high mRNA degradation rates limit the presence of exponential phase mRNAs in stationary phase28. This generates more distinct exponential and stationary phase mRNA compositions, which might increase the predictability of growth phase from mRNA data. Second; carbon sources can be reliably predicted from protein concentrations, stronger than any other condition that we investigate, on the other hand, the predictability of carbon sources from mRNA concentrations are limited. Carbon assimilation is heavily regulated with post-translational regulation29–31, and this might be the reason of the greater predictive power of the E. coli proteome compared to *E. coli* transcriptome. [figure 7 - Supplementary figure 8]. Similar predictability of Mg+2 and Na+1 from mRNA and protein datasets indicate small to none post-translational regulation related with environmental salt content. Those results are in agreement with the previous investigation of the same dataset6. In that study, we looked for cluster purity with the help of cophenetic distance of the dendrogram generated for mRNA and protein datasets 32,33. Comparison of those results with the predictability of the external conditions indicates, more the sample clusters are isolated more predictable they are. If we focus on paired data and look for the prediction performance for intersected mRNA, intersected protein and intersected mRNA & protein data we see that most of the time combined data performance is in between mRNA performance and protein performance and closed to the higher one [Supplementary figure 8].

The biggest limitation of the work is the sample size, although the sample size is comparable or larger compared to similar multi-conditional transcriptomic and/or proteomic studies 34–37, the comparison between all data and intersection data analyses [figure 4a vs supplementary figure 4 and figure 4b vs. supplementary figure 5] indicates the prediction power decreases as the training set gets smaller for both mRNA and protein data. This trend indicates we are not in stationary predictability regime in terms of the number of samples.

The second problem seems to be associated with sample number bias38–40, although we made corrections with weight factors41,42 and, we chose the multi-conditional macro F1 score43 for cost function in the optimization process. it seems, there is still a correlation between sample size and prediction performance; i.e. the predictability of individual conditions increases with the number of training samples (supplementary figure 3).

## Materials and Methods

### 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 DeSeq211 package and apply variance stabilizing transformation44 (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 Analysis21 (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 analysis13 (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 xx 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 xx times with independent divisions between training&tune, and test data sets.

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. 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 package8 for implementing support vector machines with linear, radial, and sigmoid kernels that based on the libSVM45 library and the randomForest package9 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 xx\*4=xx 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 F1 score 10,43,46 that normalized F1 scores over individual conditions; i.e. each condition has equal weight instead of each sample. Out of two distinct macro F1 score calculation methods in the literature

1. ; where 43
2. 10

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

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

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

11. 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).

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

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

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

15. 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).

16. 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).

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

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

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

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

21. 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

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

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

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

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

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

27. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. in *BioNumbers—the database of key numbers in molecular and cell biology* **38,** (2010).

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

29. 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).

30. 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).

31. 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).

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

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

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

35. 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).

36. 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).

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

38. 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).

39. Chawla, N. V. 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

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

41. 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).

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

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

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

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

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.