# Controlled Measurements of Multiple Cellular Components in *E. coli* as a Resource for Integrative Computational Modeling of Cellular Subsystems

John R. Houser1,2, Craig Barnhart1,2, Daniel R. Boutz1,2, Sean M. Carroll3, Joshua K. Michener3, Brittany D. Needham1,2, Ophelia Papoulas1,2, Viswanadham Sridhara4, Dariya K. Sydykova1, Christopher J. Marx3,5, M. Stephen Trent1,2,6, Jeffrey E. Barrick1,2,4,6#, Edward M. Marcotte1,2,4,6#, Claus O. Wilke1,2,4,7#

1Institue of Cellular and Molecular Biology, The University of Texas at Austin  
2Center for Systems and Synthetic Biology, The University of Texas at Austin  
3Department of Organismic and Evolutionary Biology, Harvard University  
4Center for Computational Biology and Bioinformatics, The University of Texas at Austin  
5Department of Biological Sciences, University of Idaho  
6Department of Molecular Biosciences, The University of Texas at Austin  
7Department of Integrative Biology, The University of Texas at Austin

#Address correspondence to: Jeffrey E. Barrick, [jbarrick@cm.utexas.edu](mailto:jbarrick@cm.utexas.edu); Edward M. Marcotte, [marcotte@icmb.utexas.edu](mailto:marcotte@icmb.utexas.edu); Claus O. Wilke, [wilke@austin.utexas.edu](mailto:wilke@austin.utexas.edu)

## Abstract

Genome wide models are a useful tool in synthetic engineering of biology. Most genome-wide models however focus on individual subsystems, or networks, of the cell, such as in flux balance models of metabolism. Recent effort has been focused on how to best integrate models of separate subsystems (e.g. transcription, metabolism, etc.). Thus, there is a need for consistent and comparable measurements of these different cellular components. Though many large data sets exist for specific types of cellular components (e.g. RNA expression, protein abundances, and metabolites) there is a lack of data that span multiple different sub-systems under the same controlled conditions. Here we present genome-wide measurements on RNA and proteins as well as lipids and metabolic fluxes from E. coli grown under identically controlled conditions. These data serve as a resource for building integrative models of cellular processes. Specifically, we present data from a long-term glucose starvation time course. These data comprise the first completed results from a series of planned experiments considering different environmental conditions. Furthermore we characterize and discuss the general trends seen in this time course comparing and contrasting our data sets on different sub-processes.

# Introduction

A goal of systems biology has been to understand how phenotype originates from geneotype. The phenotype of a cell is determined by complex regulation of cell signaling, gene regulation, metabolism, and lipid biochemistry. Understanding the connection between phenotype and genotype is crucial to understanding disease and for synthetic engineering of biology. Computational models are particularly well suited to studying this problem as they can synthesize and organize diverse and complex data in a predictive framework. Though computational models of individual component subsystems, such as flux models of metabolism1–3, have enjoyed a long history of success, they are still limited in their application. Much effort is currently being spent on understanding how to best integrate multiple subsystems. For example, there are many proposed approaches to combining gene expression with metabolic flux networks4–12 with other studies focusing on integrative, whole-cell, models13,14. Given the growing interest in integrative modeling approaches there is a pressing need for high quality genome-wide data that is comparable across cellular subsystems.

Here we present a long-term glucose starvation experiment with data on genome-wide RNA and protein levels as well as metabolic flux and lipid modification under identical, controlled, experimental conditions. The resultant data set serves as a rich resource for computational models that span and integrate cellular sub-systems and for looking up responses to specific genes and/or molecules across cellular subsystems.

To demonstrate our data sets usefulness, in the following sections we report the results of our in-depth characterization. To characterize the data we compared and contrasted differences between mRNA and protein expression profiles, differences between expression profiles within an operon, differences between metabolic flux and gene expression, and differences between lipid modification and gene expression.

# Results

## Controlled measurements of multiple cellular components

In this section we outline our approach to measure multiple cellular components under identically controlled experimental conditions. We grew multiple cultures, from the same stock, under identical growth conditions of long-term glucose starvation, in the same media, and the cultures were distributed to the different labs to measure RNA, protein, lipids, and metabolic flux ratios (Figure 1A). Freezer stocks of the *E. coli REL606* strain where revived for 24 hrs, diluted and preconditioned for another 24 hrs, and diluted again at the start of the experimental time course (Figure 1B). In a pilot experiment a growth curve was taken to determine informative time points for analysis (Figure 1C). Time points spanning 3 hrs-2 weeks where collected and equitably distributed to measure RNA via RNAseq, proteins via LC/MS, Lipids via Maldi-TOF MS and ESI MS, and Metabolic Fluxes via U-13C labeled glucose and GC-MS (Figure 1D). Below we characterize our data outlining major trends and comparing and contrasting the response of different cellular subsystems.

## General trends of RNA and Protein time course

Gene expression is often differentially regulated in order to cope with a changing environment. Due to translational and post-translational regulation we expected differences in the response of proteins and their transcripts in response to glucose starvation. To visualize these differences we compared and contrasted the general trends in the response of RNA and proteins by way of K-means clustering. To simplify analysis we focus on only those RNAs and proteins that are changing significantly (as measured by false discovery rate and fold change cutoff, respectively) throughout the time course, yielding ~1900 significantly changing transcripts/proteins. In *K*-means clustering it is up to the user to pick the number of clusters that is appropriate such that the profiles are well separated into groups with unique and distinct behaviors. We varied the number of clusters for both RNA and protein profiles to find a satisfactory grouping of the data, and we found the best clustering performance to be around 15 clusters for the RNA profiles and 25 clusters for the protein profiles. Thus, the RNAs respond in a much more uniform manner than the proteins. This is illustrated by the heatmap of the cluster centroids of RNA and protein (Figure 2A and B, respectively). A vast majority of the RNAs are down regulated while the protein response is much less uniform, likely due to differences in post-translational modification, such as degradation.

It is important to note that the above clustering of the RNA and protein are independent and thus we cannot directly compare individual clusters between Figures 2A and B. The next section addresses the correlation between absolute and relative changes in abundance of individual proteins and their transcripts.

## Correlations between individual mRNAs and proteins

It is well known that absolute levels of proteins do not correlate strongly with their corresponding transcripts (Vogel and Marcotte, 2012). In addition to a weak absolute correlation between mRNA and protein at a given time point we also expected, due to potentially long protein degradation rates in stationary phase, a weak relative correlation within a single time course between a protein and its transcript. To put proteins and RNA within a given sample on comparable scales the protein counts were normalized using the apex method (Lu et al., 2007) for absolute quantification and the mRNA was normalized to the length of each transcript. Both protein and mRNA levels were averaged across all three biological replicates. Additionally, all proteins and mRNAs were scaled by the average of all proteins and mRNA. To relate the relative levels of a protein and its transcript we had to account for the underlying dynamics of the time courses by considering two limiting cases. At one extreme we assumed each protein has a degradation rate slower than the time scale of the experiment. At the other extreme we assumed each protein is degraded on a time scale that is fast compared to the time scale of the experiment. In the first limiting case proteins integrate their transcript levels over time. In the second limiting case (relative) protein levels track with their (relative) transcript level. If a protein is not integrally or proportionally related to its transcript it could be due to an intermediate protein degradation rate, saturation effects, or some other post-translational regulation or nonlinearity. The strongest absolute correlation, across the time course, between mRNA and protein was at three hours (Figure 2F, Spearman ρ=0.59, *P*=1e-224). This number is in-line with measurements of correlations between RNA and protein for other organisms (Abreu et al., 2009; Gygi et al., 1999; Vogel and Marcotte, 2012; Vogel et al., 2010; Washburn et al., 2003). Plotted in Figure 2C and D are histograms of the Spearman correlation coefficients (ρ) calculated for protein vs. the integral of its transcript and protein vs. its transcript, respectively. Approximately 15% of the proteins correlated highly (ρ>0.75) with the integral of their transcripts whereas approximately 20% correlated highly with proportional levels of RNA. There is no overlap between the two sets as can be seen by the strong anti-correlation in the 2D histogram in Figure 2E of protein vs the integral and proportional levels of mRNA.

## Correlation of expression between genes in the same operon

Genes within an operon are co-transcribed as a single RNA and thus are likely to be under the same transcriptional control. Differences in translational efficiency between genes, localization, feedback, or other post-translational modifications often lead to differences in protein expression of genes in the same operon (Lim et al., 2011; Mattheakis and Nomura, 1988; Wek et al., 1987; Yamada and Saier Jr, 1988). It has been shown that there is a correlation between the translational efficiency of a gene and its genetic distance from the end of the operon (Lim et al., 2011). We expected to see very high correlation between RNAs within an operon as they are under the same transcriptional control; however, we expect there to be less correlation between proteins within an operon as proteins within an operon are not guaranteed to be under the same translational/post-translational regulation.

As a measure of correlation of gene expression within an operon we took the average of the pairwise Pearson correlation coefficient for all possible pairs of transcripts and proteins within an operon. Approximately eighty percent of transcripts had a mean pairwise correlation coefficient greater than 0.8 within an operon (Figure 3A). On the other hand less than fourteen percent of proteins had a mean pairwise correlation coefficient greater than 0.8 within an operon (Figure 3B). Genes closer together within an operon are more likely to have correlated protein profiles (see Figure 3C) evidence that distance between genes is a strong indicator of translational regulation.

We also show a few examples of highly correlated transcript and protein proteins for individual operons (Figure 3 E and C/D respectively). Even though the proteins in these plots have, on average, a high pairwise correlation there are still some differences in the expression profiles, unlike the transcripts that have nearly identical expression profiles for a given operon.

## RNA and Protein time course classification

Typical analysis of RNA expression data often involves performing a hierarchical clustering of profiles followed by a term enrichment of subsets of genes found in the emerging patterns. In this approach the patterning that comes from hierarchical clustering can be arbitrary, dependent on what level of the hierarchy one chooses to focus on. Thus, we sought to sort the time courses into general behaviors in an unbiased manner. To accomplish this we fit each individual mRNA and protein to a piecewise continuous curve (Figure 4A). This curve is defined by 4 free time parameters and 3 free amplitude parameters. To fit the curve we used a population based differential evolution (DE) algorithm. Furthermore the fitness function used in minimization is scaled to the experimental error (see methods). Thus, our algorithm provides confidence intervals for our fit based upon the variability in biological replicates. To demonstrate the effectiveness of our fitting strategy we randomly selected five mRNA profiles and their respective fits (Figure 2B-F). Green circles show the average of three biological replicates with their standard deviations (green bars) and the blue line and bar show the average and standard deviation of the population of fits respectively. Both the data and fit have been normalized to the average of the time course.

As can be seen in Figure 4B-E, for mRNAs we generally get good agreement between the data and model. Thus, the fits give us reasonable estimates of the distribution of time scales involved in the response. Figure 4F shows the distribution of *t*1, the time to first inflection. Most of the mRNAs respond between 3-8 hrs, during the transition between exponential to stationary growth. To better understand the regulation of cellular processes (and mRNAs) in our dataset, we sorted the mRNA profiles, based upon the estimated parameters of our fit, into five general categories: up regulated, down regulated, temporarily up regulated, temporarily down regulated, or unknown. The confidence intervals for our fits allowed sorting individual mRNAs into these five categories with high confidence. The mRNAs in the categories “down regulated” and “up regulated” showed significant enrichment for GO terms. The average of the mRNAs in each of these terms can be seen in Figure 5A, B. Terms enriched in the set of down-regulated transcripts are involved in translation, carboxylic acid biosynthetic process, and nitrogen compound biosynthetic process. These processes are, likely, down regulated for energy conservation purposes in the face of limiting resources. Terms enriched in the set of up-regulated transcripts are involved in carbohydrate catabolic process and alditol metabolic process. These two terms are involved with the breaking down and processing of carbohydrates.

To characterize the protein response we followed the same general strategy of fitting, classification, and GO enrichment as we did for the RNA profiles. The distribution of the time to first inflection for the proteins is a little broader than the mRNAs. However, they are still mostly in the range between 3-8 hrs with very few proteins that have not responded by the time the cells enter stationary phase. There are many proteins that are turned on for the duration of the time course, compared to the mRNAs where very few remain on for the entire duration of the experiment. Figure 5D shows the average of the proteins in a given GO term that are enriched in the set of proteins that are being up regulated. As in the case of down-regulated RNAs these proteins are down regulated, likely, to conserve energy and include proteins involved in translation, and locomotion. Up-regulated proteins are, like the up-regulated transcripts, involved in carbohydrate catabolism but also include terms involved in stress response and metabolism of glycerol. The average of proteins in GO terms being down regulated have a much wider distribution of decay times compared to the RNAs being down regulated likely due to differing protein degradation rates (and/or thermodynamic stability) (Figure 5C).

As a complementary approach we also averaged all proteins in a given Kegg pathway regardless of their behavior. Many pathways showed little to no differential regulation, on average, in their protein levels. Pathways that did change cohesively are plotted in Figure 5E and F depending on if they were down or up regulated, respectively. As in the previous term enrichment analysis, we see motility down-regulated as well as other energy consuming processes involved in metabolism and biosynthesis. Interestingly biosynthesis of siderophores was up-regulated, likely to do increased demands for iron.

## Flux ratio analysis.

We used flux ratio analysis to measure the relative metabolic fluxes passing through different branches of the central metabolism. To measure flux we relied on the amino acid labeling pattern. As there is little ab-initio protein synthesis after the cells stop growing (after ~8hrs) the computed flux ratio after eight hours represent a cumulative flux ratio. Thus after this time we expect that most of the flux ratios to remain relatively constant and this is certainly the case. The only significant changes in the flux ratio that occur during the time course are in P5P from G3P and S7P and P5P from G6P lower branch (Figure 6G). P5P from the G6P lower branch decreases from 1 to 0.5 into the second week. P5P from G6P and S7P is one minus the flux from P5P from the G6P lower branch. All other measured flux ratios showed little change through the course of the experiment (Figure 6A-I).<representative more of low protein turnover than anything else?> We also compared the flux ratios with the corresponding ratio of enzymes responsible for catalyzing those reactions both at the level of RNA and protein. Generally speaking, the protein ratios changed little, consistent with the low change in flux ratio <Note: I’m not sure if this is true, Viswanadan will have to speak to this>. The flux ratio that did change (P5P from G6P lower branch), however, showed a strong correlation with its corresponding protein ratio (ρ=0.85, *P*=…) and a weak (but significant?) correlation between its corresponding mRNA ratio (ρ=0.5, *P*=…).

## Lipid profiles

Using negative ion MALDI-TOF and ESI mass spectrometry we analyzed Lipid A and phospholipids, respectively. Around 8hr the ms peaks associated with an addition of the C16 chain to lipid A and the presence of a ~702.5 m/z corresponding to cyclopropanation of one unsaturated double bond within a PE molecule containing acyl chains increased. Both modifications to lipid A and phospholipid continued to increase up to the 2 week period with the 702.5 m/z peak associated with phospholipid modification dominating, by 2 weeks, over the basal major peak at ~688.5 m/z. The enzymes relevant to the above lipidA and phospholipid modifications are pagP and cycloproponated fatty acid synthase (cfa), respectively. PagP transcript levels do not change significantly during the time course, neither did PhoP levels that drive transcription of PagP. Neither PagP or PhoP was observed at the protein level. CFA synthase protein levels increase between 3-6 hrs before decreasing again and CFA synthase RNA levels increase again around 1 week (Figure 7B).

# Discussion

Our data set is a rich resource for comparing and contrasting the response of multiple cellular subsystems. It is well known that mRNA and protein levels do not strongly correlate in an absolute or relative sense. This is a fact clearly demonstrated by our data. Thus it is crucial to have data on both RNA and protein . This gives us information on both transcriptional regulation, as well as post-transcriptional/translational regulation.

We also saw a late modification of Lipids and phospholipids. These modifications are associated with acid resistance in starved cells. Additionally the modification doesn’t really follow CFA synthatase levels this potentially highlights the fact that activation doesn’t necessarily follow abundance measurements.

# Methods

## Growth procedure

*E. coli REL606* was inoculated from freezer stock in 50mL DM500 and incubated at 37C overnight. 500 uL of the overnight culture was diluted in 50mL of DM500 at 37C and grown for 24hrs. On the day of the experiment, 500UL of the 24hr culture was added to 10 flasks containing 50mL DM500 each, grown at 37C. At each time point 1ml was removed from the flasks, washed with 0.7% NACL, spun down, the supernatant was removed, and the remaining cell pellet was flash frozen using liquid nitrogen and stored at –80C.

## RNAseq

Total RNA was isolated from cell pellets using RNAsnap method (Stead et al., 2012). After extraction, RNA was ethanol precipitated and resuspended in 100ul ultra pure water. The RNA was then cleaned and DNase treated using Zymo Clean and Concentrator 25 (Zymo Research). RNA was then eluted in 50ul ultra pure water. Total RNA concentration was determined using Qubit 2.0 Fluorometer (Life Technologies). Ribosomal RNA was depleted using Gram-negative bacteria RiboZero rRNA removal kit (Epicentre). After depletion the RNA was ethanol precipitated and resuspended in 20ul ultra pure water. RNA concentration was determined using Qubit 2.0 Fluorometer (Life Technologies). RNA was then fragmented to approximately 250 bp using NEBNext Magnesium RNA Fragmentation Module (New England Biolabs). After fragmentation the RNA was ethanol precipitated and resuspended in 20ul ultra pure water. After fragmentation, the RNA was Kinase treated using T4 PNK (New England Biolabs). The RNA was ethanol precipitated and resuspended in 20ul ultra pure water. Library preparation was done using NEBNext Small RNA Library Prep Set for Illumina, Multiplex Compatible (New England Biolabs). Each sample was multiplexed using NEBNext Multiplex Oligos for Illumina (New England Biolabs). After library preparation samples were ethanol precipitated and then run on a 4% agarose gel. All DNA greater than 100 bp was excised and cleaned using Zymoclean Gel DNA Recovery kit (Zymo Research). Final library concentration was determined using Qubit 2.0 Fluorometer (Life Technologies). Libraries were sequenced using an Illumina HiSeq 2500 at the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin.

For RNA-seq analysis, we used our own custom analysis pipeline using the REL606 Escherichia coli B strain genome as the reference sequence (Jeong et al., 2009). Prior to mapping, all of the single-end reads have had the adapters trimmed using FLEXBAR 2.31 (Dodt et al., 2012). Mapping was carried out using bowtie2 2.1.0 with the –k 1 option to achieve one unique mapping location per read (Langmead and Salzberg, 2012). Raw number of reads mapping per gene were counted using htseq-count 0.6.0 (Anders et al., 2014). The exact pipeline is available at <https://github.com/clauswilke/Ecoli_RNAseq>.

## Proteomics

Frozen pellets where resuspended in 300 uL of buffer (50mM Tris-HCL pH 8.1,100mM KCL, 5mM MgCl2). 50UL of sample was removed for preparation,treated with 50uL trifluoroethanol (TFE), and placed on ice for 15min. DTT was then added to a final concentration of 5mM and incubated at 55C for 45 min. Samples where alkylated by addition of Iodoacetamide (IAM) to a final concentration of 15mM. Trypsin digest was performed by addition of 800uL of 50mM Tris pH 8.0, 2mM CaCl2 followed by 2ug of trypsin. Digestion took place at 37C for 4-5 hrs and was stopped by 10uL of formic acid (1% vol./vol). Samples where then Ultrafiltrated to remove insoluble and undigested material using Amicon Ultra MWCO 10kD spin-caps and finally concentrated and purified by C18 filtration.

Liquid chormotography and mass spectrometry (LC/MS) was carried out on a LTQ-Orbitrap (Thermo Fisher) as has been described previously <ref>. <particular settings need to be filled in>

We searched the raw MS-spectra using a Sequest search against a database for *E. Coli* REL606 that contained reverse, decoy, sequences to help control for false positives.

## Flux analysis

Flux ratios were obtained using the methods of (Zamboni et al., 2009). Cell pellets were resuspended in 200 mL of 6 N HCl, hydrolyzed at 105°C overnight, and dried at 95°C for up to 24 hours. To the hydrolyzed cell material we added 40 mL of dimethylformamide (DMF) and gently mixed until a “light straw” color was obtained. The DMF resuspension was transferred to a GC-MS vial with plastic insert and 40 mL of N-tert-butyldimethylsilyl-N-methyltrifluoroacetamide with 1% tert- butyldimethyl-chlorosilane; vials were capped and baked at 85°C for 2 hours, and samples were analyzed within 2 days of derivitization.

Analysis of derivitized samples was performed on a Shimadzu QP2010 Plus GC-MS (Columbia, MD) with autosampler. The GC-MS protocol included: 1 mL of sample injected with 1:10 split mode at 230°C; an oven gradient of 160°C for 1 min, ramp to 310°C at 20°C/min, and hold at 310°C for 0.5 min; and flow rate was 1 mL/min in helium. A total of five runs were performed for each sample: a blank injection of DMF to waste, a blank injection of DMF to the column, and three technical replicates of each vial.

Details of flux inference?

## Lipid analysis

Lipid A and phospholipids were isolated from bacterial pellets containing 3-9 x 109 cells. Pellets were resuspended in 5ml 1:2:08 CHCl3: MeOH: H2O for 20 minutes and spun at 10,000 xg for 10 minutes. Pellets containing lipid A were further purified by the Bligh/Dyer method as previously described (Hankins et al., 2013). Phospholipids in the supernatant were further purified by extractions as previously described (Giles et al., 2011). Mass analysis of purified lipid A fractions was performed using a MALDI-TOF/TOF (ABI 4700 Proteomics Analyzer) mass spectrometer in the negative ion linear mode as previously described (Hankins et al., 2013). Phospholipid analysis was performed by liquid chromatography/ESI-mass spectrometry at Duke University as previously described (ref?).

## **Expression profile data analysis**

This section outlines our process for analyzing raw counts from the proteomics and RNAseq experiments. Initially, proteins with low counts (<10) over the entire duration of the time course where filtered out. Each time point was then normalized to the read depth (e.g. the sum of all counts for that particular time point). Only proteins with a fold change of >=1.5 where considered for further analysis. Protein profiles where then normalized to the maximum value for a given protein time course. To estimate the absolute protein abundance we made use of the APEX normalization method described in detail elsewhere<vogel 2008>.

Raw RNA counts where normalized using the Deseq method <Anders 2010>. RNAs with a p-value less than 0.05, at some point during the time course, where considered to have changed significantly and kept for further analysis. To compare RNA abundances within a single time point, raw RNA counts where normalized to their read depth (sum of all counts for that time point). Finally, normalized RNA and protein profiles, both relative and absolute, where averaged across all three biological replicates.

Clustering of protein profiles was performed using the python library scipy<Jones E. 2001> k-means clustering algorithm with the number of protein clusters set to 25 and RNA clusters to 15. To compare relative protein profiles with the integral of their, relative, transcript levels we integrated each of the transcript profiles, from the initial time to each additional time point, using the trapezoidal method implemented by the python numpy library<van der Walt 2011>.

We used a piecewise continuous curve to fit both RNA and protein profiles. This curve was defined by seven free parameters, 4 free time parameters, and 3 free amplitude parameters. To fit the profiles we used a custom implementation of a differential evolution (DE) algorithm <Price 2005>. Briefly, the DE algorithm generates an ensemble of random parameter guesses within a predefined range, at each iteration vectors of individual parameter sets (sometimes called agents) are mixed together at a predefined crossover rate,only those crossover events that yield a smaller error (defined by a predefined cost function) are kept, and the process is iterated until a convergence criteria is met. In our fits we used an ensemble of 15 agents with a crossover frequency of 0.75 and an mixing strength of 0.6. The crossover frequency determines the probability that an agent will be changed at any given iteration and the mixing strength determines how large a change an agent undergoes if it is chosen to be altered. The crossover frequency and mixing strength where picked based upon an empirical study of the dependence of convergence efficiency on these parameters<Pederson 2010> for some standard optimization problems. The cost function is given by

where di(tj), σi(tj), and si(tj) are the average of all experimental repeats of protein (or mRNA) i at time tj, the standard deviation of the experiments of the protein (or mRNA) i at time tj, and the average of the ensemble simulations i at time tj respectively. Scaling by the standard deviation places a relatively lower weight on data points with relatively larger errors for a given protein or mRNA.

Some of the profiles may be slightly over fit by our curve (e.g. profiles that are upregulated or down regulated once during the time course without further modulation of expression). Thus care needs to be exercised in the interpretation of some of the parameters. However, we found t1 to reliably give the time to first inflection, that the sum of t2-t4 was a decent proxy to how long it took a RNA/protein to reach a steady state after entering stationary phase, and that we could reliable sort the behavior into four categories based upon the amplitude parameters. The four categories we used where that of up-regulated, down-regulated, temporarily up-regulated or temporarily down regulated. The sorting into categories was aided by our estimate of the distribution of parameters that allow for a good fit within in the population of fits. A fit was considered good if it was on average (across the time course) one standard deviation, or less, away from the experimental average.

We used the DAVID database(david.abcc.ncifcrf.gov)<Haung 2009 A&B> to perform Gene Ontology term enrichment on each subset of sorted genes: up-regulated, down-regulated, temporarily up-regulated or temporarily down regulated . Specifically we made use of DAVID's API, instead of the web interface, to generate the GO-enrichment through a python script. GO terms where clustered based upon genes in a given term to reduce redundancy in the returned results.

As a complementary approach we also enriched, for Kegg pathway terms, the entire set of significantly changing proteins (without presorting) using, again, the DAVID database API. The proteins within each returned Kegg pathway where then averaged to see if there was any consistent response across the entire pathway. Those Kegg terms that gave in-cohesive responses across proteins in that pathway returned a relatively flat average and where filtered out.

All of the scripts used to perform the above analysis can be downloaded at github.com/marcotte/AG3C <I should merge Daryia's code with mine>.

# References

Abreu, R. de S., Penalva, L.O., Marcotte, E.M., Vogel, C., 2009. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 5, 1512–1526. doi:10.1039/B908315D

Åkesson, M., Förster, J., Nielsen, J., 2004. Integration of gene expression data into genome-scale metabolic models. Metab. Eng. 6, 285–293. doi:10.1016/j.ymben.2003.12.002

Anders, S., Pyl, P.T., Huber, W., 2014. HTSeq - A Python framework to work with high-throughput sequencing data (No. 002824).

Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., Tagkopoulos, I., 2014. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli. Mol. Syst. Biol. 10, n/a–n/a. doi:10.15252/msb.20145108

Colijn, C., Brandes, A., Zucker, J., Lun, D.S., Weiner, B., Farhat, M.R., Cheng, T.-Y., Moody, D.B., Murray, M., Galagan, J.E., 2009. Interpreting Expression Data with Metabolic Flux Models: Predicting Mycobacterium tuberculosis Mycolic Acid Production. PLoS Comput Biol 5, e1000489. doi:10.1371/journal.pcbi.1000489

Collins, S.B., Reznik, E., Segrè, D., 2012. Temporal Expression-based Analysis of Metabolism. PLoS Comput Biol 8, e1002781. doi:10.1371/journal.pcbi.1002781

Dodt, M., Roehr, J.T., Ahmed, R., Dieterich, C., 2012. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology 1, 895–905. doi:10.3390/biology1030895

Duarte, N.C., Herrgård, M.J., Palsson, B.Ø., 2004. Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. Genome Res. 14, 1298–1309. doi:10.1101/gr.2250904

Fang, X., Wallqvist, A., Reifman, J., 2012. Modeling Phenotypic Metabolic Adaptations of Mycobacterium tuberculosis H37Rv under Hypoxia. PLoS Comput Biol 8, e1002688. doi:10.1371/journal.pcbi.1002688

Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.Ø., 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3, n/a–n/a. doi:10.1038/msb4100155

Feist, A.M., Palsson, B.Ø., 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat. Biotechnol. 26, 659–667. doi:10.1038/nbt1401

Giles, D.K., Hankins, J.V., Guan, Z., Trent, M.S., 2011. Remodelling of the Vibrio cholerae membrane by incorporation of exogenous fatty acids from host and aquatic environments. Mol. Microbiol. 79, 716–728. doi:10.1111/j.1365-2958.2010.07476.x

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between Protein and mRNA Abundance in Yeast. Mol. Cell. Biol. 19, 1720–1730.

Hankins, J.V., Madsen, J.A., Needham, B.D., Brodbelt, J.S., Trent, M.S., 2013. The Outer Membrane of Gram-Negative Bacteria: Lipid A Isolation and Characterization, in: Delcour, A.H. (Ed.), Bacterial Cell Surfaces, Methods in Molecular Biology. Humana Press, pp. 239–258.

Jeong, H., Barbe, V., Lee, C.H., Vallenet, D., Yu, D.S., Choi, S.-H., Couloux, A., Lee, S.-W., Yoon, S.H., Cattolico, L., Hur, C.-G., Park, H.-S., Ségurens, B., Kim, S.C., Oh, T.K., Lenski, R.E., Studier, F.W., Daegelen, P., Kim, J.F., 2009. Genome Sequences of Escherichia coli B strains REL606 and BL21(DE3). J. Mol. Biol. 394, 644–652. doi:10.1016/j.jmb.2009.09.052

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival Jr., B., Assad-Garcia, N., Glass, J.I., Covert, M.W., 2012. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell 150, 389–401. doi:10.1016/j.cell.2012.05.044

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. doi:10.1038/nmeth.1923

Lee, D., Smallbone, K., Dunn, W.B., Murabito, E., Winder, C.L., Kell, D.B., Mendes, P., Swainston, N., 2012. Improving metabolic flux predictions using absolute gene expression data. BMC Syst. Biol. 6, 73. doi:10.1186/1752-0509-6-73

Lim, H.N., Lee, Y., Hussein, R., 2011. Fundamental relationship between operon organization and gene expression. Proc. Natl. Acad. Sci. 108, 10626–10631. doi:10.1073/pnas.1105692108

Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E.M., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25, 117–124. doi:10.1038/nbt1270

Mahadevan, R., Edwards, J.S., Doyle III, F.J., 2002. Dynamic Flux Balance Analysis of Diauxic Growth in Escherichia coli. Biophys. J. 83, 1331–1340. doi:10.1016/S0006-3495(02)73903-9

Mattheakis, L.C., Nomura, M., 1988. Feedback regulation of the spc operon in Escherichia coli: translational coupling and mRNA processing. J. Bacteriol. 170, 4484–4492.

O’Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., Palsson, B.Ø., 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. 9, n/a–n/a. doi:10.1038/msb.2013.52

Stead, M.B., Agrawal, A., Bowden, K.E., Nasir, R., Mohanty, B.K., Meagher, R.B., Kushner, S.R., 2012. RNAsnapTM: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. Nucleic Acids Res. 40, e156. doi:10.1093/nar/gks680

Van Berlo, R.J.P., de Ridder, D., Daran, J.-M., Daran-Lapujade, P.A., Teusink, B., Reinders, M.J.T., 2011. Predicting Metabolic Fluxes Using Gene Expression Differences As Constraints. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 206–216. doi:10.1109/TCBB.2009.55

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., Penalva, L.O., 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. 6, n/a–n/a. doi:10.1038/msb.2010.59

Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227–232. doi:10.1038/nrg3185

Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., Yates, J.R., 2003. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. 100, 3107–3112. doi:10.1073/pnas.0634629100

Wek, R.C., Sameshima, J.H., Hatfield, G.W., 1987. Rho-dependent transcriptional polarity in the ilvGMEDA operon of wild-type Escherichia coli K12. J. Biol. Chem. 262, 15256–15261.

Yamada, M., Saier Jr, M.H., 1988. Positive and negative regulators for glucitol (gut) operon expression in Escherichia coli. J. Mol. Biol. 203, 569–583. doi:10.1016/0022-2836(88)90193-3

Zamboni, N., Fendt, S.-M., Rühl, M., Sauer, U., 2009. 13C-based metabolic flux analysis. Nat. Protoc. 4, 878–892. doi:10.1038/nprot.2009.58

# Figure Captions

**Figure 1.** **Overview of experimental work flow.**

A) Measurements of RNA, protein, lipids, and metabolic flux are taken under uniform growth and environmental conditions. B) Long-term glucose starvation experiment. E. coli REL606 strain was taken from freezer stock and revived in 10ml DM500 media with minimal glucose for 24 hrs (day -2). The next day this culture was diluted 1:10 in 500mL DM500 and grown for 24 hrs(day-1). On the day of the experiment (day 0) the previously conditioned culture was diluted again into several individual cultures. C) The OD @ 600 was measured to asses growth and optimal collection of time points. Nine time points where selected for this experiment spanning 3hr-2wks. D) For each sample 1ml was removed from the culture for each experiment to be done, spun down, flash frozen, and distributed to the experimental labs in our consortium to measure RNA via RNA seq, protein via LC/MS, lipids via Maldi-TOF MS and ESI MS, and metabolic flux via GC-MS.

**Figure 2. *K*-means clustering of RNA and protein profiles in long-term glucose starvation**

RNA and Protein profiles, normalized to each molecules maximum value, where clustered by geometric distance using K-means clustering with 15 and 25 clusters respectively. The cluster centroids where then plotted as heat maps with darker blue representing higher RNA, or protein, levels. A) RNA levels are largely shutoff upon glucose starvation with some RNA being temporarily up regulated during the transition between exponential and stationary growth. B) Protein levels show a much wider range of behaviors with some being up or down regulated for the duration of the experiment as well as for a short period of time during the transition from exponential to stationary growth. C) Histogram of the correlation coefficients between individual protein levels and the time integral of RNA. This is the limit of slow protein degradation where protein levels are proportional to the cumulative sum of their respective transcripts. D) Histogram of the correlation coefficient between relative protein levels and their corresponding (relative) transcripts. This is a measure of how proportional proteins are to their respective RNAs. E) 2-D histogram of the correlation coefficient of protein vs. RNA (on the x-axis) and the protein vs. the time integral of RNA (y-axis). Darker colors indicate more genes in that given bin. There is a strong anti-correlation between the two measures of dynamic correlation indicating that these two quantities are largely mutually exclusive. F) The correlation between all RNA and protein levels for a single time point is weak ( Spearman correlation coefficient of 0.59).

**Figure 3. Intra operon correlation.**

(A) and (B) histograms of the median pairwise correlation coefficient between all possible pairs of RNA and Protein profiles, respectively, within an operon. Proteins that have a smaller inter-gene distance are more likely to have correlated profiles (C). 2D Histogram of the pairwise correlation between proteins in the same operon, y-axis, and the inter-gene distance between the protein coding regions, x-axis. (D) and (E) Examples of proteins in the same operon that are highly correlated. (F) and (G) Examples of RNAs in the same operon that are highly correlated.

**Figure 4. Analysis of RNA and protein profiles for long-term glucose starvation**

We group RNA, and protein, time courses based on general qualitative behaviors. In response to glucose starvation RNA can be shut off, turned on, pulsed up, and pulsed down. A) To sort the profiles, a piecewise continuous curve is fit to our data. The parameter, t0, is the time at which we start to collect data at 3hrs into growth. The curve was fit using a differential evolution fitting algorithm that is gradient free and population based, allowing for a range of possible parameter sets that can explain our data given the experimental error. B-F) Four random examples of measured RNA time courses averaged across 3 biological replicates (green circles) with their standard deviations (green bars) along with the corresponding fits (blue). The blue bars represent the standard deviation of the range of fits that agree with our data. Both experimental time courses and fits are normalized by the average of the time course. G) Most of the RNA’s begin to change between 6-8 hrs, when the cells begin to enter stationary phase. This is demonstrated by the histogram of t1, the time to the first inflection point.

**Figure 5. Term enrichment and general behaviors.**

Fitting the mRNA and protein profiles allowed us to estimate the underlying dynamics and differential regulation of each gene sorting them into high confidence bins describing their behavior. Genes where put into bins based upon if they where up regulated, down regulated, temporarily up regulated, or temporarily down regulated. The mRNA, or proteins, in each bin where then tested for enrichment of GO terms. A) and B) shows the average of the mRNAs in a given enriched term for mRNAs that are down and up regulated respectively. C) and D) shows the average of the proteins in a given enriched term for proteins that are down and up regulated respectively. There was no significant enrichment for either proteins or mRNAs that where temporarily up or down regulated. As a complementary approach we took the average of all proteins in a given Kegg pathway annotation. E) and F) show the average of the proteins in the Kegg pathway that are changing significantly. All the other terms where flat.

**Figure 6. Flux ratio profiles in long-term glucose growth.**

Flux ratios where computed using fiat flux software from GC-MS derived 13C constraints. Only one independent flux ratio showed a significant change over the course of the experiment. Flux ratios for (A)SER from GLY, (B) OYR from MAL upper branch, (C) PEP through TK upper branch, (D) PEP through PPP upper branch, E) PEP from OAA, (F) OAA from PEP, (G) P5P from G6P lower branch, (H) E4P through TK, and (I) GLY through serine. Of the flux ratios that changed significantly (P5P from G6P lower branch) they showed a large correlation (~0.85) with the corresponding ratio of proteins that catalyze those reactions while the corresponding mRNA ratio shows a much lower correlation (~0.5) with the flux ratios.

**Figure 7. Lipid A and phospholipid profiles in long term glucose growth.**

Lipid A and phospholipids were extracted from all samples for analysis by negative ion MALDI-TOF and ESI-MS mass spectrometry, respectively, and the 6 hour and 2 week representative samples are shown in this figure. A) Activation of the acyl-transferase PagP, adds a C16 chain to lipid A on the 2-position primary acyl chain, resulting in a m/z of ~2035. B) Modification of phospholipids by cyclopropanation of one unsaturated double bond is catalyzed by CFA synthase. Transcripts of CFA synthase increase at late times (green) consistent with modification of PE where as the level of CFA protein stays relatively flat at late time points. C) As represented here by the 6 hr sample, lipid A from all samples collected between 3 and 48 hrs contained one major peak at ~1797 m/z corresponding to wild type, hexa-acylated lipid A. As illustrated on the right by the 2 week sample, the 1 and 2 week (168 and 336 hr) time points show the addition of the C16 chain to lipid A. D) Phosphatidylethanolamine (PE) is shown here as a representation of similar results obtained in phosphatidylglycerol (PG). The phospholipid profiles of the samples remained relatively consistent with wild type *E. coli* phospholipid profiles until hour 8, when a gradual increase in a peak ~702.5 m/z began and became the predominant species by 2 weeks. This mass corresponds to the cyclopropanation of one unsaturated double bond within a PE molecule containing acyl chains totaling 33 carbons distributed between the two acyl chains.

**Figure 1- supplemental figure 1** **Correlations of raw Protein and mRNA counts between different biological repeats.**

Scatter plot between biological repeats 1 and 2 (A,C) and 1 and 3 (B,D) along with their associated Spearman correlation coefficients.

**Figure 1-supplemental figure 2. Vin diagram of the protein IDs across all three biological replicates.** The overlap between all three sets is 2825 proteins identified.

**Figure 1-supplemental figure 3. RNA fractions**

For each time point the fraction of total RNA (excluding rRNA) that is either non-coding (yellow), tRNA (Red), or mRNA (blue).

**Figure 2-supplemental figure 1. Correlations between the absolute abundance between Protein and mRNA for all time points.** Both the RNA and protein levels are scaled to their respective averages across all RNAs or proteins for each time point and then log transformed. All P values are <1e-46.

**Figure 4-supplemental figure 1. Distributions of timescales found by fitting the piecewise continuous curve, described in the main text, to the mRNA and protein profiles.**

(A-C) mRNA distributions of t1, time to first inflection, the time between the first inflection and the time it takes to stop to changing, and the total time it takes for a given profile to stop changing.

# Supplemental

**Consistency of biological repeats**

To demonstrate the repeatability of our proteomic and RNAseq pipelines we investigated the correlation between individual biological repeats. To this end, we calculated the spearman correlation coefficient, and plotted the scatter, between the first and second and the first and third biological repeats for both protein and RNA raw counts (Figure 1-supplemental figure 1). Biological repeats show ~0.93± 0.02 <P value is so low as to be 0> correlation in the raw counts for both protein and RNA. Both RNAseq and mass spectorometery where comparably repeatable. Additionally the proteins ID by our mass spec pipeline in a given biological repeat are highly consistent. Just over 2800 proteins are identified across all three biological replicates, with ~3200 protein IDs for any given individual experiment (Figure 1-supplemental figure 2). This is an improvement over the previous record of ~2600 e. coli protein IDs in a mass spec experiment <Krug 2012>.

**RNA composition**

The overall pool of mRNA decreases over the time course as demands for new RNA and protein is lessened in stationary phase. To visualize this we plot the mRNA, tRNA, and non-coding (NC) RNA, as a fraction of the total RNA (excluding ribosomal RNA) for a given time point(Figure 1-supplemental figure 3). Ribosomal RNA is excluded from our analysis as it is depleted in our RNA prep. In early time points the RNA pool is dominated by mRNA. For late time points tRNAs dominate the total RNA pool.

**Absolute Correlations**

Absolute correlation between proteins and their corresponding transcripts are relatively strong for time points <=8hrs with a correlation coefficient of ~0.59. After 8hrs, when cells have entered stationary phase, the correlation is much weaker. At 48 hrs the correlation is the weakest (0.29). In E. coli mRNAs typically have relatively fast degradation rates with ½ lives around 5-10min <ref> while proteins half-lives are typically much longer (>30min). During entry into stationary phase mRNA expression is differentially regulated before the 10 hr mark, for the most part, protein levels however take a while to catch up, due to there slower degradation rate. We see this effect in the fact that correlation increases as time goes on after the initial decrease at 24-48 hrs ending up with a correlation of 0.39.

**Distributions of the mRNA and protein profile timescales**

Here we further investigate the range of responses to long-term glucose starvation seen in our RNA and protein expression experiments. To do this we plot histograms of the time scale parameters we found by fitting the piecewise continuous curve to our data(Figure 4-supplemental figure 1). The most informative time scales are t1, the time to first inflection, and t2+t3+t4 the time it takes for the profile to stop changing. The majority of proteins and their transcripts begin changing before the 10 hr mark (or just after the cells enter stationary phase). Once the profiles begin to change it takes >10 hrs before it stops changing again. However, the apparent long time scale of proteins and transcripts changing could be due to the low time resolution of our experiment after the cells have entered stationary phase.