# Controlled measurement and comparative analysis of cellular components in *E. coli* under different growth conditions

Mehmet U. Caglar, John R. Houser, Walter F. Lenoir, Drew Vander Wood, Dariya K. Sydykova, … other authors … Edward M. Marcotte, Jeffrey E. Barrick, Claus O. Wilke

Affiliations

## Abstract

In which ways Bacteria epigenetically response to several external factors is an important question that needs to be investigated. Here we present a detailed genome wide datasets of *E. coli* under several different stress conditions such as osmotic stresses as sodium and Magnesium, within a time wide range of time intervals up to 2 weeks and with four different carbon sources such as glucose, gluconate, lactate, and glycerol. Data contains multiple different unique combinations of these conditions and includes measurements in both proteomics and transcriptomics. So the data set is rich in all number of samples, variety of carbon sources, variety of size of the time span, in addition to those data includes RNA expressions and protein abundances under the same controlled conditions. This data can be used to build detailed epigenetic models related with *E. coli* that contains changes related with osmotic stresses, starvation and behavior with different carbon sources. In addition to that we characterize and discuss general trends in data and compare the effects of several variables.

## Introduction

A goal of systems biology has been to understand how phenotype originates from genotype. The phenotype of a cell is determined by complex regulation of 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 (Botstein and Risch 2003). Computational models are particularly well suited to studying this problem as they can synthesize and organize diverse and complex data in a predictive framework, but there is a need for more detailed experimental studies including more samples in this area to understand interactions between different levels of omics (Zhang, Li, and Nie 2010). For example much effort is currently being spent on understanding how to best integrate multiple cellular subsystems. For example, there are many proposed approaches to combining gene expression with protein abundances focusing on integrative, whole-cell models.

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 under many different external conditions. *E. coli* is an ideal organism to study high quality genome wide multi level epigenetic effects of external conditions, since it is well adapted to laboratory environment (Lee 1996) and was one of the first organisms whose genome is sequenced (Blattner et al. 1997) and because of the mentioned reasons itsgenome is one of the most studied genomes. In addition to that the external conditions spanned with our data set gives a detailed picture of the effects of a variety of important external parameters.

To demonstrate our data sets usefulness in the following sections report the results of our characterization. To characterize the data we firstly compare how well our data is clustered in terms of different. We also look at the number of genes that are significantly up and down regulated under different conditions and analyze clustering of up and down regulation of genes under different conditions in order to figure out similarities and differences between conditions under investigation. We finally look at the common genes between different parameters that are significantly changed with respect to changing variables of the given parameter.

## Results

### Experimental design and data collection

We grew multiple cultures of E. coli REL606, from the same stock, under a variety of different growth conditions, and we measured RNA and protein abundances under all conditions (Figure 1). Results from one of these conditions, long-term glucose starvation, have been presented previously (Houser et al. 2015). Conditions not previously described include one additional starvation experiment, using glycerol instead of glucose as carbon source, exponential and stationary phase clutures using gluconate and lactate as carbon source, respectively, and conditions including Mg+2 and Na+1 stress. More details of the data can be found in Supplementary Table 1.

RNA and protein abundances were measured as described (Houser et al. 2015). All resulting data sets were carefully checked for quality, normalized, and log-transformed. Our raw RNA-seq data before quality control consisted of 149 data sets covering read for 4279 distinct genes. After quality control, we retained 143 data sets covering 3698 distinct genes. Similarly, our raw protein data consisted of 81 data sets for 3278 distinct proteins, and we retained all 81 data sets and 3278 proteins after quality control. All raw data files are available in appropriate repositories (see Methods for details), and final processed data are available as Supporting Tables X-Y.

### Broad trends of gene expression differ between mRNA and proteins

To identify broad trends of gene expression among the different growth conditions, we performed hierarchical clustering on both mRNA and protein abundances (Figures 2 and 3). For mRNA, we found that differences in gene expression were primarily driven by the growth phase (exponential vs. stationary/late stationary). Nearly all exponential samples clustered together in one group, separate from the vast majority of stationary and late-stationary samples (Figure 2). Mg+2 levels, Na+1 levels, and carbon source had less influence on the clustering results. By contrast, we did not find a similarly strong effect for protein abundances (Figure 3). Protein abundances seemed to cluster primarily by experimental batch (see below).

To quantify the clustering patterns of mRNA and protein abundances, we defined a metric that measured how strongly clustered a given variable of the growth environment (growth phase, Mg+2 level, Na+1 level, carbon source) was relative to the random expectation of no clustering. For each variable, we calculated the mean cophenetic distance between all pairs corresponding to the same condition (e.g., for growth phase, all pairs sampled at exponential phase and all pairs sampled at stationary/late stationary phase). The cophenetic distance is defined as the height of the dendrogram (produced by the hierarchical clustering) from the two selected leafs to the point where the two branches merge. We then converted each mean cophenetic distance into a *z*-score, by resampling mean cophenetic distances from dendograms with reshuffled leaf assignments. A *z*-score below -2.0 indicates that the mRNA or protein abundances are clustered significantly by the corresponding variable.

We found that mRNA abundances were significantly clustered by growth phase, with a *z*-score of -3.34 (Table 1). Mg+2 levels displayed the next-largest *z*-scores by magnitude, of -0.79, but not significantly different from zero. The *z*-scores for all other variables did not exceed 0.20 in magnitude (Table 1). Importantly, when we calculated a *z*-score for batch number *(how exactly is that defined?)*, we found that batch effects were virtually absent in the mRNA abundances, *z* = -0.16.

For protein abundances, on the other hand, clustering was primarily driven by batch effects (Table 1). Only batch number had a significant *z*-score, of -4.62. The next largest *z*-score by magnitude was -1.55 for carbon source, followed by -1.22 for Na+1 levels. Growth phase, the dominant effect for mRNA levels, had a *z*-score of only -0.66 (Table 1).

### Identification of differentially expressed genes

*This entire section needs to be redone. You’re using a non-standard way of analyzing differentially expressed genes. This will not pass peer review. I’m leaving the text in blue for now, but it will have to be rewritten.*

The next step is, determining under which conditions RNA and protein levels are altered and in what strength these changes occur. Our large data set enables us to look for changes in genes with respect to several conditions and we analyze these changes on RNA and protein levels and look for similarities and differences between conditions.

One can look at the significance of changes in normalized RNA read counts and in normalized protein read counts, with respect to multiple conditions. So each time we either categorize data with two labels "a" data with given carbon source "b" rest of the data, or focus on varying continuous parameters such as Na, Mg+2 concentrations or growth time in hours than we look at the change in the read levels and measure the significance of change. For RNA we calculate significance of change for 4279 genes and for 3278 proteins with respect to conditions related with four variables. *P* generated from Spearman correlation between normalized reads and the investigated conditions represents significance of change for each gene or protein. This *P* represents how good each gene or protein is as a separator for given condition. The next step is to adjust these *P’* by considering multiple hypothesis testing. For each condition we adjust p-values obtained for genes and proteins by using “fdr”. Then we pick the genes or proteins that have a *P’* under 0.05 as significantly varying ones.

So as can be seen from figure 4a most significant changes happen due to growth time and Mg+2 levels follow it. With increasing time almost four times of genes are down regulated than the genes that are up regulated, and by putting into account that we are analyzing 3698 genes almost 2/3rds of the genes of the genes are significantly altered (either up regulated or down regulated) with growth time. With increasing Mg+2 concentration a significant portion of the genes are up regulated and a few of them are down regulated. For glucose there are thirty-two genes that are significantly down regulated and six genes that are up regulated and for glycerol eight genes are down regulated and sixteen of them are up regulated significantly. We do not have any significantly up regulated or down-regulated genes for increasing Na+1 concentration and for carbon sources gluconate and lactate. So there is a large variation in responses of RNA data to applied stresses. Similarly as can be seen from figure 4b most significant changes in protein levels happen due to Na+1 concentration and glucose-glycerol levels follow it. There are almost 1300 up regulated and 1000 down regulated proteins associated with increasing Na+1 concentration. There are 880 up regulated and 569 down regulated proteins associated with a carbon source switch from glucose to glycerol. There are 260 proteins that are up regulated significantly with increasing time and 267 down regulated proteins associated with increasing time. There are only 2 proteins significantly down regulated with increasing Mg+2 concentration.

The next thing we investigate is how the significantly changed genes are shared between categories for both RNA-Seq data and protein data. For defining the significantly changed genes and proteins we use the same threshold we used before (P’<0.05) and count the genes or proteins that change significantly with respect to that threshold. Figure 4c shows a Venn diagram of three categories growth phase, carbon source, and Mg+2 level representing the number of significantly changed genes with respect to categories. We do not have any significantly up regulated or down-regulated genes for increasing Na+1 concentration and for carbon sources gluconate and lactate so we do not have set for Na+1 levels in Venn diagram and the carbon source set is composed of two carbon sources glucose and glycerol. As can be seen there are fifteen genes that are significantly affected from all three categories investigated. The whole list of these genes is in table 2. 31% of all genes that are significantly altered by changing carbon source are only affected from changes in carbon source, and 32% of all genes that are significantly altered by changing Mg+2 concentration are only affected from changes in Mg+2 concentration. Unlike carbon source and Mg+2 concentration 77% of all genes that are significantly altered by changing growth phase are only affected from changes in growth phase. 62% of genes that are altered by carbon sources are also altered by growth phase, similarly 65% of genes that are altered by Mg+2 concentration are also altered by growth phase. Unlike carbon source and Mg+2 concentration 77% of all genes that are significantly altered by changing growth phase are only affected from changes in growth phase. In addition to that 90% of all genes that are significantly altered in our experiments are related with growth phase. Similarly figure 4d shows a Venn diagram of three categories growth phase, carbon source, and Na+1 level representing the number of significantly changed proteins with respect to categories. We only have two significantly down regulated and no up regulated proteins for increasing Mg+2 concentration they are not significantly affected from any other variable and so not included in figures. As can be seen there are sixty-three genes that are significantly affected from all three categories investigated. The whole list of these proteins is also in table 2. 12% of all proteins that are significantly altered by changing carbon source are only affected from changes in carbon source, and 42% of all proteins that are significantly altered by changing Na+1 concentration are only affected from changes in Na+1 concentration. 39% of proteins that are altered by growth phase are only affected from changes in growth phase. 83% of the proteins affected by changes in carbon source and 49% of proteins affected by changes in Na+1 stress are affected from changes in both carbon source and Na+1 stress.

## Discussion

The aim of the project is to investigate the epigenetic behavior of *E. coli* under different stress conditions. We subject *E. coli* to different experimental conditions and measure the RNA and protein abundances at different time points. Experimental conditions include four different carbon sources and variations of Na and Mg stress levels. As a result, we obtain a comprehensive dataset that can be used to understand the relationship between external conditions and cellular responses.

The data gives a comprehensive picture of *E. coli* both in terms of number, range and depth of different stresses. For example (Schmidt et al. 2015) only focuses on protein levels and measures (>2300) genes and deals with 22 unique conditions or (Soufi et al. 2015) again measures (>2300) proteins and deals with 10 unique conditions. Similarly, (Lewis et al. 2010) deals with 3 different carbon sources and three to six replicas of adaptive process on them. Although the (Lewis et al. 2009) study focuses on 70 unique conditions in 213 expression profiles it does not look into protein expressions.

We observe several differences between RNA and protein patterns. The differences between proteins and RNA’s are observed in multiple terms. We observe the correlation between proteins and RNA observations by using corresponding data sets and find the Pearson correlation xx and if we only focus on stationary state data we find the Pearson correlation as xx, these values are in range the similar range with previous studies. Due to (Vogel and Marcotte 2012), the differences between RNA and Protein abundances are due to translation or protein degradation other than mRNA abundances which is determined by transcription and mRNA degradation. On the other hand, contribution of experimental noise to this variation is extremely low. One other the significant observations is the difference between response patterns for RNA and proteins. We measured the clustering quality of data sets with respect to variables we investigate for both RNA and protein data. Resulted order of clustering is different for proteins and RNA. For proteins the most significant response is to change in carbon sources on the other hand for RNA the most significant response is to change in growth time. So we can conclude that the effects of time in RNA level does not transferred to protein level that strong. On the other hand effects of carbon source, which is more distributed than random when we consider RNA data, causes distinguished most significant clusters in protein data. This might indicate other sources of regulation on proteins depending on the carbon source.

Another difference between RNA and protein responses is related with the distribution of number of significantly responding genes and proteins. Number of significantly responding elements decreased from thousand five hundred to five hundreds for growth time related data. So most of the significantly varying RNA abundances do not cause significantly varying protein levels. A similar effect is observed for responses to Mg+2 changes between RNA and protein levels. Around seven hundred significantly responding RNA does only cause two significantly varying proteins. For Mg+2 responses this differences might be related with cells response time.

An inverse effect is occurred related with responses variations in Na+1 levels and carbon sources. There are just few RNA abundances significantly response to those two variables, but there are thousands of protein abundances that response to the same variables. These observations might be related with other regulation mechanisms that are involved in the control of protein abundances

When it comes to clustering of variables due to responses proteins and RNA shows similar behaviors except growth time. All carbon sources clustered together and Mg+2 and Na+1 stands different ends of carbon sources for both RNA and protein concentrations. Growth time is the most significant variable for carbon sources but it is not the most separated out variable for proteins.

There are some issues related with data we collected. One of the biggest problems is the difference between the number of data samples for different variables for both RNA and protein data. These differences might cause some biases in *z-score* and *P* calculations.

Another problem is related with the number of different values used for spanning variable intervals. For carbon sources the change is binary it is either the carbon source we are focusing or not. For Na+1 concentrations we have four different values and for Mg+2 concentrations we have 10 different values. These differences might cause some biases in *p value* calculations of spearman correlations used for measuring the significance of responses with respect to proteins or RNAs

A third problem is the batch effects; the experiments were done in a large time intervals and in different laboratories if there are some affects that are related with these uncontrolled variables it is hard to fix them since batches also correlated with experiments. The batch effect for RNA data is small (with *z-score* -0.16), but the effect is pretty strong in protein data (with *z-score* -4.62).

As a summary our study provides a large data set that investigates the epigenetic responses of *E. coli* under different external conditions, to our knowledge this is one of the biggest studies both in terms of number of different variables investigated and number of sample collected, a basic data analysis was done to compare and contrast the similarities and differences between different variables for both RNA and protein abundances.

## Materials and Methods

### Cell Growth

Growth and harvesting of *E. coli* B REL606 cell pellets for the multiomic analysis was as previously described with the following additional details. For tests of different carbon sources, the Davis Minimal (DM) medium used was supplemented with 0.5 g/L of the specified compound (glycerol, lactate, or gluconate) instead of glucose. Mg2+ concentrations were varied by changing the amount of MgSO4 added to DM media from the concentration of 0.83 mM that is normally present. For tests of different Na+ concentrations, NaCl was added to achieve the final concentration. The base recipe for DM already contains ~5 mM Na+ due to the inclusion of sodium citrate, so 95 mM NaCl was added for the 100 mM Na+ condition, for example. Exponential phase samples were taken during growth when the OD600 reached 20-60% of the maximum achieved after saturating growth. Stationary phase samples were collect 20-24 hours after the corresponding exponential sample. The exact sampling times for each condition are provided in Supplementary Table S1.

### RNA-seq

Total RNA was isolated from cell pellets using the RNAsnap method (Stead et al. 2012). After extraction, RNA was ethanol precipitated and resuspended in μl H2O. Each sample was then DNase treated and purified using the on-column method for the Zymo Clean & Concentrator-25 (Zymo Research). RNA concentrations were determined throughout the purification using a Qubit 2.0 fluorometer (Life Technologies). DNase-treated total RNA (≤5 μg) was then pro-cessed with the gram-negative bacteria RiboZero rRNA removal kit (Epicentre). After rRNA depletion, each sample was ethanol precipitated and resuspended in H2O again. A fraction of the RNA was then fragmented to ~250 bp using NEBNext Magnesium RNA Fragmentation Module (New England Biolabs). After fragmentation, RNA was ethanol precipitated, resuspended in 20 μl ultra-pure water, and phosphorylated using T4 PNK (New England Biolabs). After another ethanol precipitation cleanup step, sequencing library preparation was performed using the NEBNext Small RNA Library Pre Set and Multiplex Oligos for Illumina, Multiplex Compatible (New England Biolabs). Samples were ethanol precipitated again after library preparation and separated on a 4% agarose gel. All DNA fragments greater than 100 bp were excised from the gel and isolated using the Zymoclean Gel DNA Recovery kit (Zymo Research). Libraries were sequenced using an Illumina HiSeq 2500 at the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin to generate 2×101-base paired-end reads.

For RNA-seq analysis, we implemented a custom analysis pipeline using the REL606 *E. coli* B genome (GenBank: NC\_012967.1) as the reference sequence (Jeong et al. 2009). We updated annotations of sRNAs in this genome sequence using the Rfam 11.0 database (Burge et al. 2013). Prior to mapping, we trimmed adapter sequences from Illumina reads using Flexbar 2.31 (Dodt et al. 2012). Mapping was carried out in single-end mode using Bowtie2 2.1.0 with the –k 1 option to achieve one unique mapping location per read (Langmead and Salzberg 2012). The raw number of reads mapping to each gene were counted using HTSeq 0.6.0 (Anders, Pyl, and Huber 2015). Exact details for the full computational pipeline are available at https://github.com/wilkelab/AG3C\_starvation\_tc\_RNAseq.

### Proteomics

*E. coli* cell pellets were resuspended in 50 mM Tris-HCl pH 8.0, 10 mM DTT. 2,2,2- trifluoroethanol (Sigma) was added to 50% (v/v) final concentration and samples were incubated at 56°C for 45 min. Following incubation, iodoacetamide was added to a concentration of 25 mM and samples were incubated at room temperature in the dark for 30 min. Samples were diluted 10-fold with 2 mM CaCl2, 50 mM Tris-HCl, pH 8.0. Samples were digested with trypsin (Pierce) at 37°C for 5 h. Digestion was quenched by adding formic acid to 1% (v/v). Tryptic peptides were filtered through Amicon Ultra 30 kD spin filtration columns and bound, washed, and eluted from HyperSep C18 SpinTips (Thermo Scientific). Eluted peptides were dried by speed-vac and resuspended in Buffer C (5% acetonitrile, 0.1% formic acid) for analysis by LC-MS/MS.

For LC-MS/MS analysis, peptides were subjected to separation by C18 reverse phase chromatography on a Dionex Ultimate 3000 RSLCnano UHPLC system (Thermo Scientific). Peptides were loaded onto an Acclaim C18 PepMap RSLC column (Dionex; Thermo Scientific) and eluted using a 5-40% acetonitrile gradient over 250 min at 300 nl/min flow rate. Eluted peptides were directly injected into an Orbitrap Elite mass spectrometer (Thermo Scientific) by nano-electrospray and subject to data-dependent tandem mass spectrometry, with full precursor ion scans (MS1) collected at 60,0000 resolution. Monoisotopic precursor selection and charge-state screening were enabled, with ions of charge >+1 selected for collision-induced dissociation (CID). Up to 20 fragmentation scans (MS2) were collected per MS1. Dynamic exclusion was active with 45s exclusion for ions selected twice within a 30s window.

Spectra were searched against an *E. coli* strain REL606 protein sequence database and common contaminant proteins (MaxQuant using SEQUEST (Proteome Discoverer 1.4; Thermo Scientific). Fully-tryptic peptides were considered, with up to two missed cleavages. Tolerances of 10 ppm (MS1) and 0.5 Da (MS2), carbamidomethylation of cysteine as static modification, and oxidized methionine as dynamic modification were used. High-confidence peptide-spectral matches (PSMs) were filtered at <1% false discovery rate determined by Percolator (Proteome Discoverer 1.4; Thermo Scientific).

### Data Preparation

Our data is generated in two different sets proteins, and RNA Seq. For RNA data we do not have any NAs and the reads are exact integers on the other hand the protein data initially contains non-integer values because disturbing same reads to different associated proteins. For the pipeline protein data is rounded and converted to integer values. The unreadable protein values which initially shown by blanks in the raw data sets are converted to zeros. One more additional problem in the protein data is the in matching rows associated with different experiments done in different dates. We only use the intersected parts of the protein data rows in the analysis. Then both the RNA Seq data and protein data is treated in the same way.

The first step is to find some outliers in the data sets. The idea is to find concentration distributions for protein and RNA concentrations if all the four variables are same, i.e the only difference is batch. We investigated the conditions where we have at least 3 batches with the same condition and look at the differences between these 3. If one of these 3 batches is significantly different than the other two we assume that batch has a problem, and is not included in the further analysis. The differences are investigated by analyzing histograms of logarithms of point wise division of concentration vectors in pairs of two. So in cases where we have 3 samples, we look at the log ratios of three possible couples and generate histograms of their ratios. The expected histogram should have a significant peak around 0, and if two of the three histograms do not have this property we claim that one sample out of three is different than other two and should not be used in the further analysis.

After removing the odds, the both data sets are normalized with DeSeq algorithm, which aims to normalize the effects of search depth in different experiments. In regular DeSeq routine size factor calculation is composed of three steps; at first step algorithm calculates geometric mean of each gene count through all different data sets. In the second step it normalizes each gene in a data set with respect to *geometric mean* of that gene. In the third step it calculates *median value* of reads with respect to all genes measured and obtains a value called size factor for that data set. This method has a downside; if for a gene we have zero in any of 143 data sets then the geometric mean of that gene is zero and it is impossible to normalize the read with respect to geometric mean for that gene. Out of 3698 genes only 176 have non-zero reads for all samples. To prevent this problem we obtained size factors by adding “+1” to all measurements, which helps to use all genes even the ones that have zero value for some of the data sets. Obtained size factors are used to normalize the original (not +1 added) dataset. This procedure of adding +1 helps to generate more robust results. (Supporting information).

Finally, DeSeq normalized data sets are re-normalized with variable stabilizing transformation (*vst*) function in DeSeq2 package, that is similar to calculation of logarithm of the measured values. These normalized datasets are used in the further steps of the analysis.

### Measuring Clustering

The normalized data sets are than clustered based on Euclidian distance between them and by using the complete linkage method for *‘hclust’* function in R. This method defines the cluster distance between two clusters as the maximum distance between their individual components (Madhulatha 2012). At every stage of the clustering process, the two closest clusters are merged into a new cluster. The repetitive process continues until the whole data set is bundle up into one single cluster. After this process we ended up with a dendogram that measures closeness of different samples, which is generated without focusing on the external variables. The next step is to investigate how the external variables are clustered with respect to this dendogram. For this we calculate *cophenetic distance* for data pairs in each category in a variable and calculate the mean of them. Than compare the mean value with mean values generated with the same method for artificially mixed labels to calculate the *z-score* of the mean value.

### Quantification of Cell Response

The response of an individual RNA or protein count to change in a single variable is quantified by calculation of Spearman correlation between normalized reads and the investigated condition. To calculate the individual effects of carbon sources, data is labeled as “1” for the investigated carbon source and labeled as “0” for the rest of data. Then the *P* value of Spearman correlation of normalized counts with respect to this “1” and “0” is calculated. If the *P* value is small enough one can say the change of that specific protein or RNA concentration with respect to the carbon source is significant. A similar method is used for continuous variables of Mg+2 concentration, Na+1 concentration and growth phase as well. This time, the *P* value of Spearman correlation of normalized counts with respect to the continuous variable is calculated. All the *P* values for individual genes or proteins are adjusted to *P’* values by false discovery rate “fdr” method before further analysis.

The significance of responses to “carbon source” is calculated as finding the minimum of *P* values computed for response of individual carbon source with respect to the rest of the data. These values are used to find the number of significantly varying RNA and protein concentrations with respect to “carbon source” in the *Venn diagrams*.

To see the clustering of cell responses to different variables we generate a new variable named as *score* which is derived from *P’* values. It is calculated as

In the equation, *rho* is Spearman's rank correlation coefficient and *P’* is the adjusted *P value* for Spearman correlation. The *score* is a positive number if the change is in the positive direction and a negative number if the change is in the negative direction. The magnitude of *score* is bigger if the change is more significant and smaller if the change is less significant.

## References

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics (Oxford, England)* 31 (2): 166–69. doi:10.1093/bioinformatics/btu638.

Blattner, Frederick R., Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, et al. 1997. “The Complete Genome Sequence of Escherichia Coli K-12.” *Science* 277 (5331): 1453–62. doi:10.1126/science.277.5331.1453.

Botstein, David, and Neil Risch. 2003. “Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease.” *Nature Genetics* 33 (March): 228–37. doi:10.1038/ng1090.

Burge, Sarah W., Jennifer Daub, Ruth Eberhardt, John Tate, Lars Barquist, Eric P. Nawrocki, Sean R. Eddy, Paul P. Gardner, and Alex Bateman. 2013. “Rfam 11.0: 10 Years of RNA Families.” *Nucleic Acids Research* 41 (Database issue): D226–32. doi:10.1093/nar/gks1005.

Dodt, Matthias, Johannes T. Roehr, Rina Ahmed, and Christoph Dieterich. 2012. “FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms.” *Biology* 1 (3): 895–905. doi:10.3390/biology1030895.

Houser, John R., Craig Barnhart, Daniel R. Boutz, Sean M. Carroll, Aurko Dasgupta, Joshua K. Michener, Brittany D. Needham, et al. 2015. “Controlled Measurement and Comparative Analysis of Cellular Components in E. Coli Reveals Broad Regulatory Changes in Response to Glucose Starvation.” *PLoS Computational Biology* 11 (8): e1004400. doi:10.1371/journal.pcbi.1004400.

Jeong, Haeyoung, Valérie Barbe, Choong Hoon Lee, David Vallenet, Dong Su Yu, Sang-Haeng Choi, Arnaud Couloux, et al. 2009. “Genome Sequences of Escherichia Coli B Strains REL606 and BL21(DE3).” *Journal of Molecular Biology* 394 (4): 644–52. doi:10.1016/j.jmb.2009.09.052.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

Lee, Sang Yup. 1996. “High Cell-Density Culture of Escherichia Coli.” *Trends in Biotechnology* 14 (3): 98–105. doi:10.1016/0167-7799(96)80930-9.

Lewis, Nathan E., Byung-Kwan Cho, Eric M. Knight, and Bernhard O. Palsson. 2009. “Gene Expression Profiling and the Use of Genome-Scale In Silico Models of Escherichia Coli for Analysis: Providing Context for Content.” *Journal of Bacteriology* 191 (11): 3437–44. doi:10.1128/JB.00034-09.

Lewis, Nathan E, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, et al. 2010. “Omic Data from Evolved E. Coli Are Consistent with Computed Optimal Growth from Genome-Scale Models.” *Molecular Systems Biology* 6 (July): 390. doi:10.1038/msb.2010.47.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550. doi:10.1186/s13059-014-0550-8.

Madhulatha, T. Soni. 2012. “An Overview on Clustering Methods.” *arXiv:1205.1117 [cs]*, May. http://arxiv.org/abs/1205.1117.

Schmidt, Alexander, Karl Kochanowski, Silke Vedelaar, Erik Ahrné, Benjamin Volkmer, Luciano Callipo, Kèvin Knoops, Manuel Bauer, Ruedi Aebersold, and Matthias Heinemann. 2015. “The Quantitative and Condition-Dependent Escherichia Coli Proteome.” *Nature Biotechnology* advance online publication (December). doi:10.1038/nbt.3418.

Soufi, Boumediene, Karsten Krug, Andreas Harst, and Boris Macek. 2015. “Characterization of the E. Coli Proteome and Its Modifications during Growth and Ethanol Stress.” *Frontiers in Microbiology* 6 (February). doi:10.3389/fmicb.2015.00103.

Stead, Mark B., Ankit Agrawal, Katherine E. Bowden, Rakia Nasir, Bijoy K. Mohanty, Richard B. Meagher, and Sidney R. Kushner. 2012. “RNAsnapTM: A Rapid, Quantitative and Inexpensive, Method for Isolating Total RNA from Bacteria.” *Nucleic Acids Research*, July, gks680. doi:10.1093/nar/gks680.

Vogel, Christine, and Edward M. Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews Genetics* 13 (4): 227–32. doi:10.1038/nrg3185.

Zhang, W., F. Li, and L. Nie. 2010. “Integrating Multiple ‘Omics’ Analysis for Microbial Biology: Application and Methodologies.” *Microbiology* 156 (2): 287–301. doi:10.1099/mic.0.034793-0.