# The *E. coli* molecular phenotype 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

Text in blue still needs to be copy-edited.

## Abstract

Modern systems biology requires extensive, carefully curated measurement of cellular components in response to different environmental conditions. While modern high-throughput methods have made transcriptomics and proteomics data sets widely accessible and relatively cheap to generate, systematic measurements of both mRNA and protein abundances under a wide range of different conditions are still relatively rare. Here we present a detailed, genome wide transcriptomics and proteomics dataset of *E. coli* grown under 34 different conditions. We manipulate concentrations of sodium and magnesium in the growth media, and we consider four the different carbon sources glucose, gluconate, lactate, and glycerol. Moreover, samples are taken both in exponential and stationary phase, and we include two extensive time-courses, with multiple samples taken between 3h and 2 weeks. We find that exponential-phase samples systematically differ from stationary-phase samples, in particular at the level of mRNA. Regulatory responses to different carbon sources or salt stresses are more moderate, but we find numerous differentially-expressed genes for growth on gluconate and under salt and magnesium stress. Our data set provides a rich resource for future computational modeling of *E. coli* gene regulation, transcription, and translation.

## Introduction

A goal of systems biology has been to understand how phenotype originates from genotype. The phenotype of a cell is determined by complex regulation of 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). 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 [refs?].

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 regulatory effects of external conditions, since it is well adapted to laboratory environment (Lee 1996) and was one of the first organisms whose genome was sequenced (Blattner et al. 1997). *Put here 2-3 sentences on what has and what has not been done with E. coli in terms of systematic large-scale measurements, with references.*

Here, we provide a systematic analysis of *E. coli* gene expression under a wide variety of different conditions. We measure both mRNA and protein abundances, at exponential and stationary phases, for growth conditions including different carbon sources and different salt stresses. We find that mRNAs and proteins display divergent responses to the different growth conditions. Further, growth phase yields more systematic differences in gene expression than either carbon source or salt stress, though this effect is more pronounced in mRNAs than in proteins. We expect that our data set will provide a rich resource for future modeling work.

## Results

### Experimental design and data collection

We grew multiple cultures of *E. coli* REL606, from the same stock, under a variety of different growth conditions. We measured RNA abundances under all conditions and matching protein abundances approximately 2/3 of the 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 cultures using either gluconate or lactate as carbon source, and conditions including Mg+2 and Na+1 stress.

Measurements of RNA and protein abundances were carried out as previously described (Houser et al. 2015). All resulting data sets were carefully checked for quality, normalized, and log-transformed. We discarded 6 RNA samples from further analysis due to inconsistencies in biological replicates, so that our final data set consisted of 143 RNA samples and 101 protein samples. (Lists of all RNA and protein samples are available as Supplementary Table 1.)

Our raw RNA-seq data covers 4279 distinct mRNAs and our protein data covers 4201 distinct proteins. All raw data files are available in appropriate repositories (see Methods for details), and final processed data are available as Supporting Tables 3 and 4.

**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. We also found a similar result for protein abundances (Figure 3). The exponential-phase samples grouped together, separated from stationary and late stationary samples. Similarly, Na+1 levels and carbon sources also seemed to be grouped together upon clustering.

To quantify the clustering patterns of mRNA and protein abundances, we defined a metric that measured how strongly clustered a given variable of the growth environment (growth phase, 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 −1.96 indicates that the mRNA or protein abundances are clustered significantly by the corresponding variable.

### We found that mRNA abundances were significantly clustered by growth phase, with a *z*-score of −23.99 (Table 1). Na+1 and Mg+2 levels displayed the next-largest *z*-scores by magnitude, of −1.54 and −1.46, but these were not significantly different from zero. The *z*-score for carbon source was 1.16, which implies that there is no significant clustering by carbon source in the mRNA data. Importantly, when we calculated a *z*-score for batch number. All bacterial samples with the same batch number were started growing at the same time. (Ones with different batch numbers were grown on different days.) So, if there was an effect of incubator temperature slightly changing, media formulation, whatever other day-to-day variation, it would occur between batches but not within them. We found that batch effects also significantly influenced mRNA abundances, with *z* = −2.82.

### For protein abundances, the variables Na+1 level, growth phase, and carbon source were all significantly clustered, with *z*-scores of −4.78, −4.21, and −3.15, respectively (Table 1). Batch number had a *z*-score of −23.29, which implies that there were strong batch effects present in the protein data.

### In summary, the largest effect in mRNA abundances, growth phase, was similarly present in proteins. However, protein abundances clustered also by Na+1 and carbon source, effects that weren’t present in the mRNA data. Finally, both mRNA and protein data were influenced by batch effects, and the effect was much more pronounced for proteins than for mRNA (Table 1).

### Identification of differentially expressed genes

We next asked under which conditions and to what extent RNA and protein expression were altered. To identify differentially expressed mRNAs and proteins, we used DeSeq2 (Love, Huber, and Anders 2014). Since our data clustered significantly by growth phase, we analyzed RNA and protein expression separately for exponential and stationary phase. For each growth phase, we defined a reference condition of glucose as carbon source, base level, 5mM Na+1, and base level, 0.8mM Mg+2. We then compared RNA and protein abundances between this reference condition and the alternative conditions (different carbon sources, elevated Na+1, and elevated or reduced Mg+2). Note that a detailed comparison of reference exponential phase vs. reference stationary phase has already been published (Houser et al. 2015).

We defined significantly differentially expressed genes as those whose abundance had at least a two-fold change (log2 fold change > 1) between the reference condition and a chosen experimental condition, at an FDR-corrected *P* value < 0.05. We found that the number of significantly differentially expressed mRNAs and proteins varied substantially between exponential and stationary phase and between mRNAs and proteins (Figure 4). In general, there were fewer differentially expressed genes in stationary phase than in exponential phase. Further, protein abundances showed the most differential regulation for high Na+1 and for the carbon source glycerol, whereas mRNA showed the most differential regulation for altered Mg+2 levels and again for the carbon source glycerol.

For mRNA concentrations all investigated salt stresses mostly causes down regulation compared to the base levels with an exception of stationary phase high Mg concentrations. The effect of changing carbon sources causes mostly down regulation in mRNA levels with an exception of stationary phase lactate experiments. If we focus on protein concentrations low or high Mg concentrations either cause no or few significant change. On the other hand high Na causes lots of proteins to change their concentrations significantly both in exponential and stationary phases. Glycerol causes mostly down regulation of proteins, gluconate causes no change in protein concentrations and lactate causes up regulation of some proteins.

Next, we asked how much overlap there was among differentially expressed genes between the various growth conditions. To simplify this analysis, we did not distinguish between up- or down-regulated genes, and we combined low and high Mg+2 into one group “Mg stress” and glycerol, lactate, and gluconate into one group “carbon source”. (Note that differentially expressed genes were still identified for individual conditions, as described above, and were combined into “Mg stress” and “carbon source” only for the final comparison.)

Most of significant differentially expressed mRNAs are related with Mg stress (73% of all mRNAs) in exponential phase and 75% of all data in stationary phase) and for exponential phase a big portion of them are also differentially expressed with changing carbon sources (30% of all mRNAs). Ratio of significantly expressed mRNAs associated with changing carbon sources are decreased in stationary phase compared to exponential phase (50% of all data in exponential phase and 27% of all data in stationary phase). Ratio of significantly expressed mRNAs associated with changing Na increased in stationary phase compared to exponential phase (12% of all data in exponential phase and 30% of all data in stationary phase). The dominant factor that affects the protein concentrations is Na stress in both exponential and stationary phase (85% of all data in exponential phase and 89% of all data in stationary phase), it is not only the biggest one but also shares its proteins with both carbon source Mg stress.

When we look at significantly altered KEGG pathways & GO annotations by using DAVID (Huang, Sherman, and Lempicki 2008) we find several significantly up and down regulated pathways & annotations under different conditions. Supplementary tables 6 and 7 shows top 5 significantly altered KEGG pathways and molecular function related GO annotations under different conditions. We also individually display the changed pathways/annotations and associated, significantly altered mRNAs and proteins under supplementary figures. The regulation of KEGG pathways associated with mRNA data ordered with respect to their significance under high Mg and high Na stress in exponential phase is shown in figures 6 and 7. As one can see, almost all of the most altered pathways are significantly down regulated in both of the situations, contrast to this in stationary phase high Mg causes up regulation of most altered pathways figure 8.

We also found that a lot of metabolism related pathways are down regulated with salt stresses in both exponential and stationary phase with an exception of stationary phase high Mg mRNAs. In stationary phase with high Mg levels there is clear trend of up regulation of multiple mRNAs associated with metabolism (Figure 6), which seems like an exception when considering other up and down regulation trends. (Figure 9)

## Discussion

We have studied the regulatory response of *E. coli* under a wide variety of different growth conditions. The experimental conditions we have considered include four different carbon sources, variations of Na+1 and Mg+2 stress levels, and growth into deep stationary phase, up to two weeks post inoculation. We have found that gene regulation changes the most with respect to growth phase; in general, the exponential phase under one condition is more similar to the exponential phase under another condition than to the stationary phase under the same condition. Further, we have found that the smallest number of differentially expressed genes for different carbon sources, and larger numbers for different salt stresses. Under salt stress, most genes tend to be down-regulated, with the exception of elevated Mg+2 in stationary phase, where we have found over twice as many up-regulated mRNAs as down-regulated mRNAs. Finally, our measurements provide a comprehensive dataset that will be useful for future modeling of the *E. coli* metabolism.

Our data provides a comprehensive picture of *E. coli* 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. That researchs focus os more on distribution of protein mass within clusters of orthologous groups under different stress conditions and they try to find how cells are allocated under different conditions in our study instead of COG allocations we more focused on KEGG pathways. (Soufi et al. 2015) again measures (>2300) proteins and deals with 10 unique conditions. They more focused on up and down regulated proteins under different ethanol stresses. They applied different stresses than our study they found some similar results like down regulation of genes associated with arginine biosynthesis and metallic ion binding. 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 and the focus of the study is the change in protein concentrations in evolved strains. In comparison, we considered 34 unique conditions, measured 143 mRNAs and 101 proteins.

We observe several differences between RNA and protein patterns. 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 mRNA and proteins. We measured the clustering quality of data sets with respect to variables we investigate for both mRNA and protein data. Resulted order of clustering is different for proteins and mRNA. For mRNAs the most significant response is to change in growth time (*z* = −23.99) on the other hand for protein the significantly responding variables from most significant to least significant are Na levels (*z* = −4.78), growth phase (*z* = −4.21) and carbon source (*z* = −3.25). So we can conclude that the effect of time in mRNA level does projected to protein level in a statistically significant way. On the other hand effects of Na levels and carbon source, which is more distributed than random when we consider mRNA data, seems to amplify in protein data. This might indicate other sources of regulation on proteins depending on the Na levels and 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 related with Mg+2 stresses decrease from mRNA to proteins on the other hand significantly responding elements related with Na+1 stresses increases from mRNA to proteins. When it comes to carbon sources the patterns look more similar between mRNAs and proteins.

So most of the significantly varying RNA abundances associated with salt stresses do not cause significantly varying protein levels. For Mg+2 and Na responses this differences might be related with cells response time.

When we discuss the individual effects of stresses we can observe for most of the stresses most of the genes are down regulated and some are up regulated. Some of the up regulated genes are consisted with the previous literature for example (Weber, Kögl, and Jung 2006) found up regulation of tktB and talA under NaCl stress which is consistent with our results. Changes in salt concentrations in this study also caused many changes in ABC transporter expression. In particular, low Mg2+ concentrations cause a decrease in the fec operon, involved in iron transport. This may be because, at low magnesium concentrations, magnesium transporters are permeable to ferrous iron (Hantke 1997) and a downregulation of iron transporters is needed to compensate. The fact that many ABC transporters are affected by salt concentrations may be indicative of other transporters that are leaky to off-target ions which have not yet been characterized.

*There needs to be a paragraph relating this paper to the previous Houser paper. Maybe about here.*

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*-scoreand *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 mRNA and proteins are both significant with z scores of −2.82 and −23.39 respectively. There are plenty of methods pre or post computational discussed associated with reducing the batch effects. One of the recent studies discuss multiple methods related with removing batch effects (Kim, Zorraquino, and Tagkopoulos 2015), where they generate an iterative machine learning algorithm to cope with batch effects.

In summary, our study provides a large and comprehensive data set investigating the gene-regulatory response of *E. coli* under different growth conditions, both at the mRNA and the protein level. We have found systematic differences in gene-expression response between exponential and stationary phase, and between mRNAs and proteins. We believe that our data set provides a rich resource for future modeling of the *E. coli* metabolism.

## 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 . 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. Then both the RNA Seq data and protein data is treated in the same way.

**Sample filtering explain**

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 determination of differentially expressed mRNA and proteins contains a process that can be determined in multiple steps.

Firstly we filter out the samples that biological replicates are not poses similar distributions with in each other and sum all the technical replicates to generate one sample, which is the suggested way to handle technical replicates in DeSeq2 [ref]. The next step is to filter out the unnecessary conditions from the data file. For example if we are trying to figure out the affects of high Mg+2 on *E.Coli* in exponential phase we first pick the samples that only have glucose as carbon source, Na levels are equal to 0.8mM (base value), and pick the base and high Mg samples in exponential phase. We do not apply any filtering on specific mRNAs or proteins based on raw counts.

Then we apply DeSeq2 to calculate the size factors for samples in order to get rid of the affects of search depth inequalities. While calculating size factors we add "+1" to all samples to get non-zero results for the mRNAs and proteins that have at least one zero read.

After finishing the DeSeq2 procedure we have log 2 fold change and adjusted p values for individual mRNAs and proteins. Picking up the samples based on *P'<0.05* and *log2 fold change >2* we end up with lists of mRNAs and proteins which responds to external change significantly. We use these lists and send them to the database for annotation, visualization and integrated discovery tool (DAVID) (Huang, Sherman, and Lempicki 2008) and look for results related with KEGG pathways and molecular function related GO annotations [ref].

## 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.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2008. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols* 4 (1): 44–57. doi:10.1038/nprot.2008.211.

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.

Kim, Minseung, Violeta Zorraquino, and Ilias Tagkopoulos. 2015. “Microbial Forensics: Predicting Phenotypic Characteristics and Environmental Conditions from Large-Scale Gene Expression Profiles.” *PLOS Comput Biol* 11 (3): e1004127. doi:10.1371/journal.pcbi.1004127.

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.

Weber, Arnim, Stephanie A. Kögl, and Kirsten Jung. 2006. “Time-Dependent Proteome Alterations under Osmotic Stress during Aerobic and Anaerobic Growth in Escherichia Coli.” *Journal of Bacteriology* 188 (20): 7165–75. doi:10.1128/JB.00508-06.

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.