### 

### Analysis via New Gene Database Verifies Gene Expression Changes in Response to Rifamycin Drug in *Shigella flexneri*

Trixie Anne M. Roque, Jake A. Woodlee, Erich M. Yanoschik, Kristin H. Zebrowski

Biological Databases

BIOL 376-01

Loyola Marymount University

Submitted 17 December 2015

### 

### 

### 

### 

### **Abstract**

*Shigella flexneri* is the leading cause of dysentery and shigellosis in humans and has exhibited a remarkable ability to build resistance to different antibiotics. A 2012 study by Fu et al. concluded that two rifamycin antibiotics, rifaximin and rifampin, effectively cause cessation of ribosomal activity and translation but are also associated with an up-regulation in virulence genes. To verify this and further understand how *Shigella* builds resistance, a newly customized database was built with XMLPipeDB’s GenMAPP Builder for the species using genomic data from a 2002 study by Jin et al. The microarray data from the Fu et al. experiment was run through this database using GenMAPP and criteria was created to analyze the effects of the drugs at different concentrations and time points. The data obtained through MAPPFinder supported results from the Fu et al. study while also providing more insight on how rifamycin works in *Shigella*.

### **Introduction**

*Shigella flexneri* is a gram-negative facultative bacillus that is responsible for most of the shigellosis and bacillary dysentery deaths in humans worldwide. It invades the epithelial tissue of the colon and rectum, resulting in inflammation as well as intestinal symptoms ranging from short-lasting watery diarrhea to acute inflammatory bowel disease (3). There is no vaccine for *Shigella* in existence and current treatments and preventative strategies are inadequate or are being met with significant resistance from the species. Approximately 163 million cases and 1.1 million *Shigella*-related deaths are reported in developing countries each year and most of these cases are children under 5 years old, so understanding the mechanisms behind *Shigella*’s virulence properties and how its adverse effects can best be combatted are critical (4).

Pathogenic strains of *Shigella* are known to contain a virulence plasmid that contains genes necessary for invading intestinal epithelial cells and survival in host environments, i.e. surviving the proteases and acids of intestinal tract as well as restricting the immune response by T-lymphocytes. The virulence plasmid encodes a type III secretion system that allows the bacteria to enter into vacuoles in colonic epithelial cells, lyse the cell membrane, and rapidly replicate in the cytoplasm (6). In the last 50 years this plasmid has rendered previous therapy such as sulfonamides, ampicillin, and tetracycline nearly useless, leading to a surge in research to understand why resistance is building so quickly so that a proper vaccine can be created (4). One such antimicrobial agent that *Shigella* is gradually undermining is rifamycin, which has previously been an effective treatment because binds to the beta-subunit of RNA polymerase, thus inhibiting RNA synthesis. Two compounds in the rifamycin group, rifampin (RP) and rifaximin (RX), are FDA-approved antibiotics used to treat a variety of intestinal and bowel disorders. Their effects on gene expression in *Shigella* were compared to provide insight on how the bacteria circumvent treatment by Fu et al. in 2012 (2). The study found that while ribosomal and metabolism-related genes are consistently down-regulated over time with different concentrations of RF and RP, virulence genes such as those encoding flagellar assembly are commonly up-regulated.

In 2002, Jin et al. conducted a gene sequencing study on *Shigella flexneri* 2a str. 301 (Sf301) and proved that this serotypeis closely related phylogenetically to pathogenic strains of *Escherichia coli*. They discovered that both possessed the same backbone and pathogenicity island mosaic structure, although Sf301 lacks some genes present in *E. coli* and has some additional, less-understood adaptations for intracellular replication and growth (3). Using the information provided in the paper, as well as an external model organism database, we utilized XMLPipeDB’s GenMAPP Builder to construct a newly customized gene database for Sf301. Using this new database, we ran the microarray data through GenMAPP and MAPPFinder to analyze how RF and RP may be affecting gene expression levels and therefore virulence in *Shigella* in response to the antibiotic.

### **Materials & Methods**

**Downloading Necessary Files**

Files that were imported into TallyEngine were taken from the UniProt, GO Consortium page, and GOA pages. In the UniProt page, the specific taxonomy (623) and proteome IDs (UP000001006) had to be manually inputted into the search bar so that the proper strain of *Shigella flexneri* can be downloaded. All the files that were used in this project were downloaded on November 19, 2015: the UniProt XML version 2015\_11, the GO-OBO XML updated on 11/19/2015, and the GOA released on November 11, 2015. XMLPipeDB’s GenMAPP Builder was also needed to be downloaded.

**Creating the Database using PostgreSQL**

Our gene database was created using PGAdmin III. The commands used to execute the creation of the database were taken from XMLPipeDB’s GenMAPP Builder *gmbuilder.sql* file. Once the query finishes, GenMAPP Builder had to be configured to look for the PostgreSQL database that was just created.

**Exporting into a GenMAPP Gene Database**

The requirements at this stage were to ensure that all of the necessary files were imported into TallyEngine. Then, we also ensured that the name of the database matched the one created in PostgreSQL, and the file name that we would want the exported file will be called is designated. Then, we ran the export for approximately 1 to 2 hours.

**Validating the Gene Database**

Microsoft Office was primarily used to inspect the contents of the exports. Using Microsoft Access to open the .gdb file, the Systems table was observed to contain the necessary information, such as the date of the export and the URL of the external model organism database that we would link to our own gene database. Other tables that were inspected for validity were the OrderedLocusNames, UniProt, and RefSeq. The latter two were quickly skimmed for consistency, because the OrderedLocusNames table is the one that matters the most in this project.

### **Creating a Compiled Raw Data File**

The GenMAPP users accessed the microarray data from ArrayExpress and downloaded the raw files. This file was unzipped and expanded to show the treatment and control hybridized and the columns were re-labeled to indicate that they were the hybridized log ratio samples with the format LogFC [drug-concentration-time point-rep #]. For convenience, one GenMAPP user did all RP samples and the other did all RX samples. In Microsoft Excel, all ID columns were deleted except for column A, which was re-named “MasterIndex”, and rows labeled “empty” or “blank####” were deleted, leaving behind 3,926 total files. The word “error” was replaced with a space character. The samples were then combined into a master compiled raw data file.

**Normalization of Log Ratios and Statistical Analysis**

Data was averaged in Microsoft Excel and the standard deviation was calculated for all data columns indicating a different treatment in a new worksheet called “Statistics.” The average log fold change of the repetitions per treatment and time period was calculated and these columns were named Avg\_LogFC [drug-concentration-time point]. The T statistic and P value (degrees of freedom=2) were calculated that were representative of all three replicates per treatment. Bonferroni P values were calculated as well as the corrected Bonferroni P value, which replaced any number greater than 1 with the number 1 in the spreadsheet.

**Sanity Check**

Columns containing P values were separately filtered by p < 0.05, p < 0.01, and p < 0.001 and Bonferroni p value column was filtered by p < 0.05. Results for each were recorded as numbers and percentages out of 3,926 total files.

**Preparing file for GenMAPP and Running GenMAPP**

GenMAPP, a gene map annotator and pathway profiler, was used for data analysis by the GenMAPP users through the polished database created by the QA and Coder (5). A new worksheet was created for import into GenMAPP. The statistics sheet was copied and pasted into the new worksheet. All fold changes were changed to show 2 decimal places and all P value columns were edited to show 4 decimal places. A “SystemCode” column was created so that each ID had a SystemCode value of N. In order for the file to be recognized and read accurately by the program, the “0.5” headers were changed to “0pt5” and all “#DIV/0!” messages were replaced with a space character. This was saved as a “Text (Tab-delimited) (\*.txt)” format. The new database created by the QA and Coder was uploaded into GenMAPP and the compiled raw data .txt file was chosen in the Expression Dataset Manager.

**Creation of Color Sets and Running MAPPFinder**

In the Expression Dataset Manager the name of each treatment was typed in the format [drug-concentration-time period] in the “Name” section. Under “Gene Value”, the same treatment was added with the prefix Avg\_LogFC. A new color set was added with the label “Increased” and colored red and the criterion was written as “[Avg\_LogFC treatment] > 0.25 AND [PValue\_LogFC treatment] < 0.05”. A second criterion was added for that treatment with the label “Decreased” and colored green. The criterion was written as “[Avg\_LogFC treatment] < -0.25 AND [PValue\_LogFC treatment] < 0.05.” This color set was saved and the process was repeated for every drug concentration/time point combination for a total of 12 color sets.

**Analysis of MAPPFinder Results**

After each color set was saved, the resulting Criterion.GO files were opened with Excel. The criterion files for increased and decreased RX and RP at -0pt5-10 and -1-60 were filtered by the following criteria: “Number Changed” column filtered to show only between 4 and 100, “Z score” column filtered by greater than 2, “Permute P” column filtered to show less than 0.05 only, and “Adjusted P” column filtered to show less than 0.1. Out of the GO terms listed after applying each of the filters, the top 10 (or 9) were selected for these four treatments and added to a compiled file of filtered MAPPFinder results. In this compiled filtered results file all columns were deleted save for GO ID, name of GO term, number changed/present and percentage, number present/number in GO and percentage, regular p value, and adjusted p value.

**Placing Genes on MAPP and Drawing Pathway**

Receiving the analysis of MAPPFinder results with relevant GO terms regarding the induced and repressed expression of genes comparing the different drugs at 0.5xMIC at 10 minutes and 1xMIC at 60 minutes provided insight into what pathway to create on the MAPP(5). The genes associated with the pathways were found at <http://www.genome.jp/kegg/pathway.html>. The organism of interest and strain, *Shigella flexneri* 301 2a, were selected and pathways were searched regarding relevant GO terms. The maps provided information with gene labels and IDs which were used to create the MAPP. GenMAPP would search the IDs by OrderedLocusNames showing which genes were in the directory and which ones were absent. Genes were placed in categories regarding pathway and organized based on alphabetical order and function. Multiple maps were made in reference to commonly induced genes, commonly repressed genes, and the filtered criterion-GO files generated by GenMAPP.

### **Results**

**Gene Database Schema Figure for *Shigella flexneri* 2a str. 301**

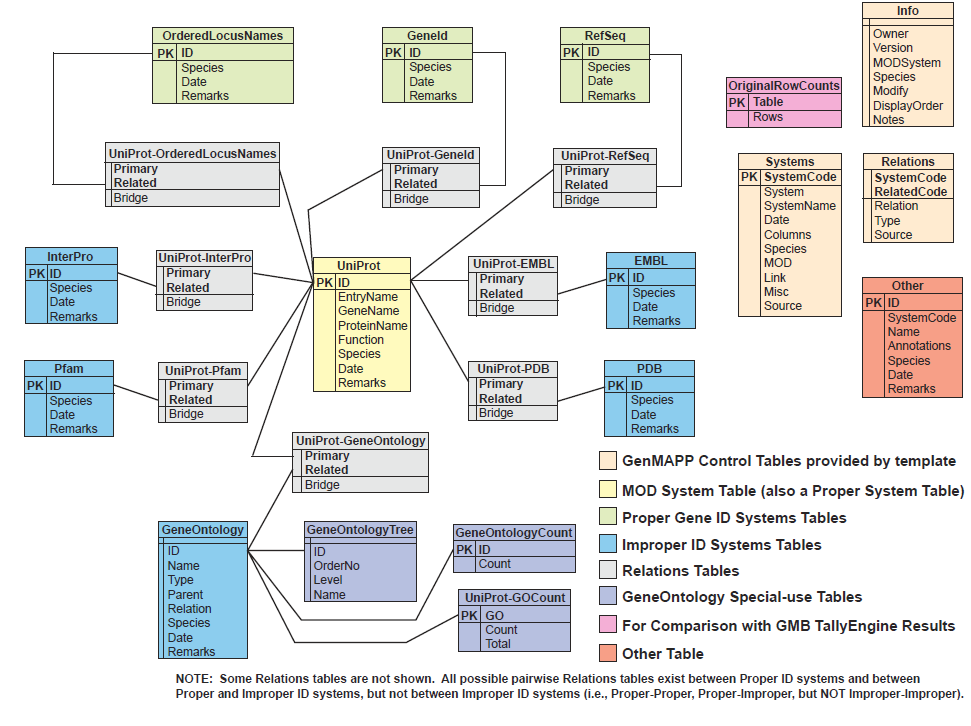
**Gene Database Testing Report**

Figure 1 – Gene database schema

See Appendix section.

**Gene Database Summary**

The ordered locus names listed below are more than the external MOD since each of the genes in the chromosome had two ordered locus names:

|  |  |
| --- | --- |
|  | # of OrderedLocusNames IDs |
| XMLPipe DB Match | 7569 |
| TallyEngine (UniProt) | 15134 = 2 \* 7567 |
| TallyEngine (PostgreSQL) | 7660 |
| OriginalRowCounts | 7661 |
| external MOD | 4446 (coding genes - chromosomes) |

**Table 1 – Summary of overall results**

The discrepancy of ~3200 IDs stems from the fact that the exported OrderedLocusNames were in fact representing the same genes. The IDs of the form SF#### can also be represented with the form S####.

**XMLPipeDB Match Command**

Two slightly different commands were used on match that was then combined to get the total number of IDs:

* java -jar xmlpipedb-match-1.1.1/xmlpipedb-match-1.1.1.jar “>(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?(/|</name>)” < uniprot-proteome%3AUP000001006.xml
* java -jar xmlpipedb-match-1.1.1/xmlpipedb-match-1.1.1.jar “/(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?(/|</name>)” < uniprot-proteome%3AUP000001006.xml

**PGAdmin III Query**

A query specific for our specie was created in order to ignore IDs that were already imported. The union of the default command for finding the ordered locus names and the command for capturing the dbReferences was used:

select count(value) from (select value from genenametype where type = 'ordered locus' and value ~ '(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?' union select extra as value from (select propertytype.value as extra from propertytype inner join dbreferencetype on propertytype.dbreferencetype\_property\_hjid = dbreferencetype.hjid where dbreferencetype.type = 'EnsemblBacteria' and [dbreferencetype.id](http://dbreferencetype.id/) ~ 'AAN[0-9][0-9][0-9][0-9][0-9]' and propertytype.type = 'gene ID' and propertytype.value ~ 'SF[0-9][0-9][0-9][0-9]') as f left join (select value from genenametype where type = 'ordered locus' and value ~ '(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?') as g on f.extra = g.value where g.value is null) as combined;

**TallyEngine Results**

The number of OrderedLocusNames IDs that were not in the XML source at all was about 322. The missing ones are of the form SF#### and S####. The number of OrderedLocusNames IDs that were not exported are those that are misnomers. A quick search using *firstObject XML Editor* revealed that some of the IDs do exist, however, they are not the ones that we needed since they did not appear in the form SF####, but as parts of some other strings (i.e., some appeared as SSF#####). From inspecting the .gdb file using Microsoft Access, the ~92 genes were indeed exported. In fact, we had exported a slightly higher number of genes than what we expected.

**Changes to the GenMAPP Builder Code**

The results above were done after we have successfully modified the GenMAPP Builder code specifically for our specie. We added a subclass of UniprotSpeciesProfile called ShigellaflexneriUniprotSpeciesProfile. This class initially had a custom constructor and overrode only one method entitled getSystemsTableManagerCustomizations. This method was customized to aid in the export of *Shigella flexneri* genes. These customizations resulted in a .gdb file that was missing 416 of the genes as seen in the Microarray raw data. It was discovered that only 92 of these 416 genes were actually in the XML file. To capture these missing 92 genes another method was created that overloaded the previous method “getSystemsTableManagerCustomizations()”. A SQL query was added that aided in the exportation of the remaining genes. The full customization, as seen below, allowed us to get all of the genes we could out of the Uniprot XML file, which in turn aided the work done in GenMAPP.



Figure 2 - Final Version of custom species profile *Shigella flexneri pt.1*

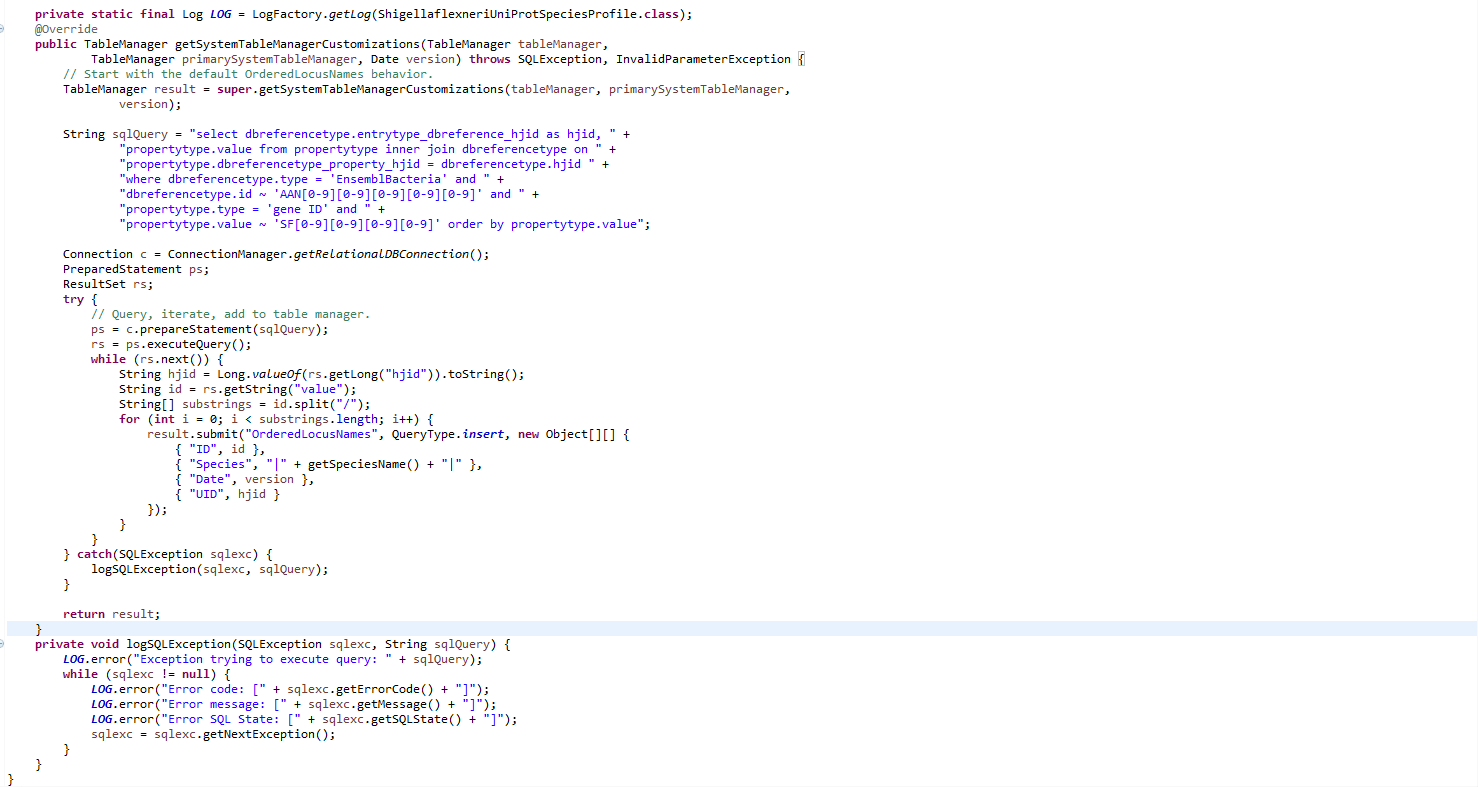
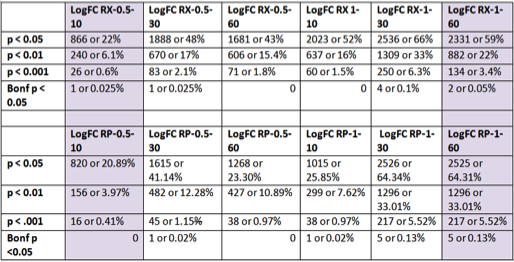


Figure 3 Final Version of custom species profile *Shigella flexneri pt.2*

**Sanity Check**



**Table 2 - Amount of genes that exhibit significance at different p-values and Bonferroni results**

Results of the sanity check indicate the amount of gene expression within the given parameters, including both up-regulated and down-regulated genes.

**Criteria for GenMAPP Expression Dataset**

Color sets were created using the criteria of log fold change greater than 0.25 for increased (red) and log fold change less than -0.25 for decreased (green). The p value cutoff for significance was set as less than 0.05 (Table #).

|  |  |  |
| --- | --- | --- |
| **Label in Legend** | **Color** | **Criteria** |
| Increased |  | Increased [Avg\_LogFC] > 0.25 and [P value] < 0.05 |
| Decreased |  | Decreased [Avg\_LogFC] < -0.25 and [P value] < 0.05 |

### **Table 3 – Criterion for Gene inhibition and induction within MAPPFinder**

**Filtered MAPPFinder Results**

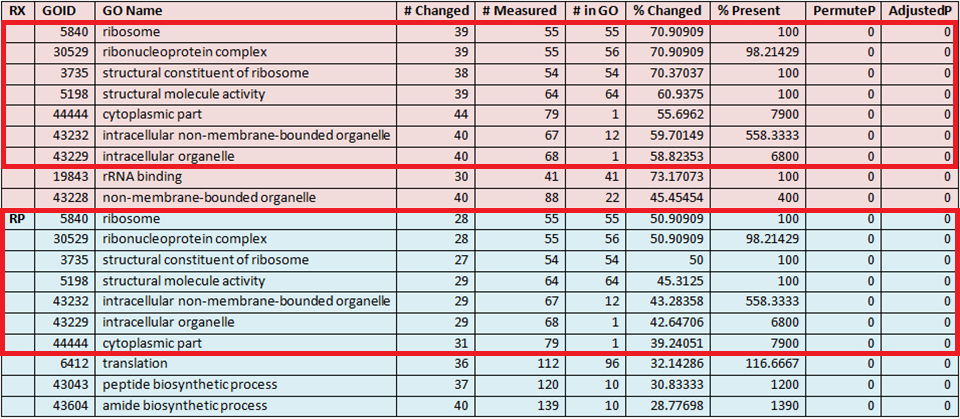


Table 4 – Down-regulation CriterionGO results from RXRP 0.5xMIC at 10 min

Results from MAPPFinder for RX and RP at 0.5 x MIC after 10 minutes indicated a drastic down-regulation of genes encoding ribosome, ribonucleoprotein complex, structural constituent of ribosome, structural molecule activity, cytoplasmic part, intracellular non-membrane-bounded organelle, and intracellular organelle. RX caused a greater percentage of genes to be significantly changed than RP but both had at least around 40% of genes encoding those processes to be changed (Table 4**)**. This is consistent with the findings of Fu et al., who found that RX has more of an effect than RP, particularly as time increases.

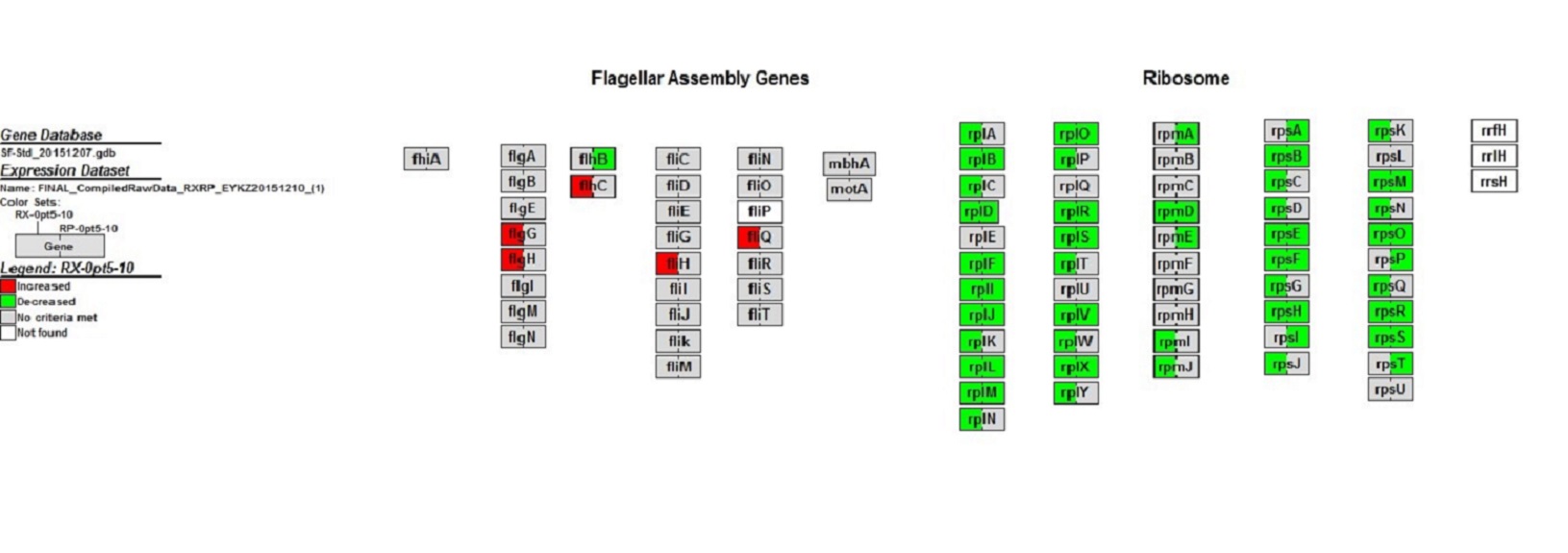


Figure 4 – MAPP of flagellar assembly and ribosomal genes RXRP 0.5xMIC at 10 min

GenMAPP was utilized to create MAPPs of pathways corresponding with gene pathways indicated relevant within CriterionGO filtered results. The above MAPP is associated with production of proteins located in pathways associated with flagellar assembly and the large and small subunit of the ribosome (Figure 4**)**. The MAPP indicates a substantial amount of genes associated with ribosomal protein production at 0.5 x MIC RP vs RX at 10 minutes are repressed.

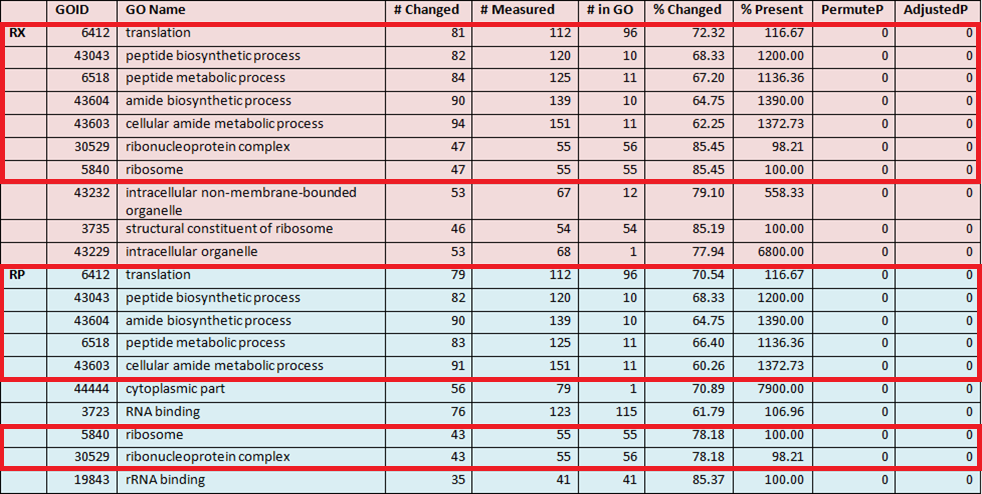
At 1 x MIC after 60 minutes both RX and RP showed significant down-regulation of translation, peptide biosynthetic process, peptide metabolic process, amide biosynthetic process, cellular amide metabolic process, ribonucleoprotein complex, and ribosome (Table 5**)**. In light of the Fu et al. paper these results are consistent because the drug acts to bind to the beta subunit of RNA polymerase, inhibiting translation of RNA into amino acids and peptide-related processes that occur in the ribosomal complex. At least 60% of genes were changed for both drugs, with RX generally having slightly more genes changed out of the total number.

Table 5 – Down-regulation CriterionGO results from RXRP 1xMIC at 60 min

### 

Table 6 – Up-regulation CriterionGO results from RXRP 1xMIC at 60 min

### At 1 x MIC after 60 minutes both RX and RP showed significant up-regulation of carbohydrate transport, cell projection, bacterial-type flagellum part, and cell projection part. This is consistent with the Fu et al. paper in that after 60 minutes of exposure to the drug there was common up-regulation seen in virulence genes from the plasmid such as those encoding flagellar production and motility. RP had a greater effect in relation to up-regulation of flagellar genes, indicating that RP increases virulence to a greater extent. This was not addressed by Fu et al.

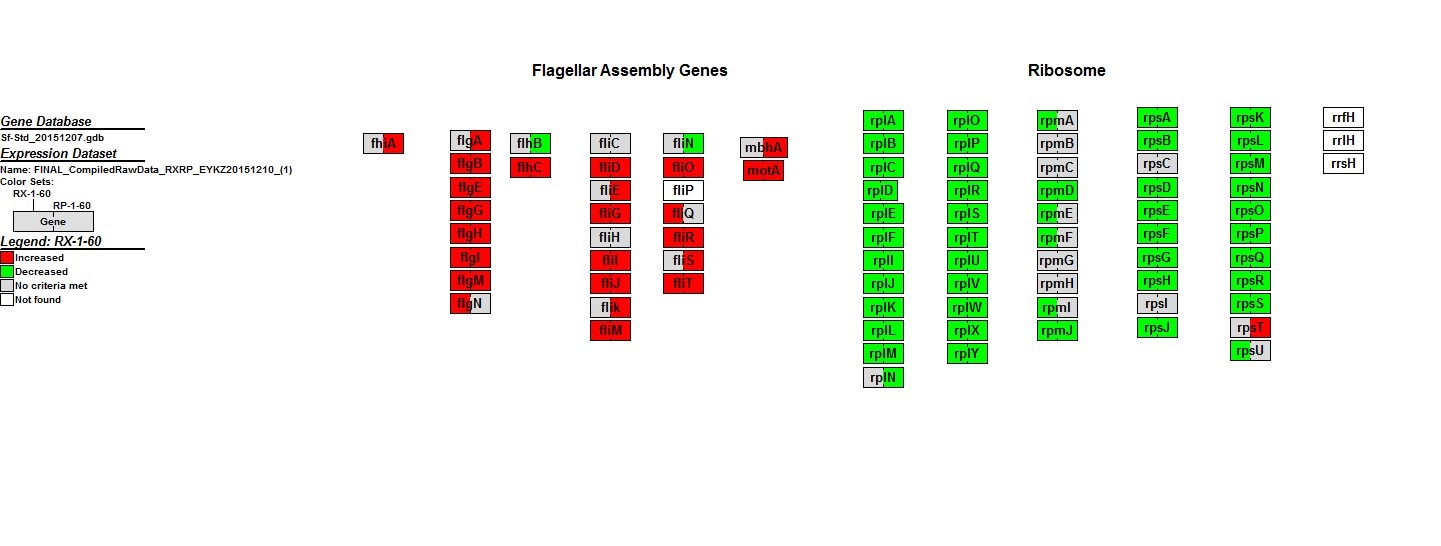


Figure 5 – MAPP of flagellar assembly and ribosomal genes RXRP 1xMIC at 60 min

The MAPP of RX vs RP 1xMIC at 60 minutes shows genes exhibit an increased expression during both treatments within the flagellar assembly pathway and decreased expression for genes located in the ribosomal protein pathway **(**Figure 5**)**. Flagellar assembly genes code for proteins associated with other pathways such as cellular processes and cell motility.

One of the features of GenMAPP is the ability to analyze and cross reference multiple data sets at a time. MAPPs were made to compare the analyzed data with commonly induced or repressed genes described by Fu et al. Both drugs at 0.5xMIC and 1xMIC were viewed at each time point to analyze and record data correlations with that of Fu et al.

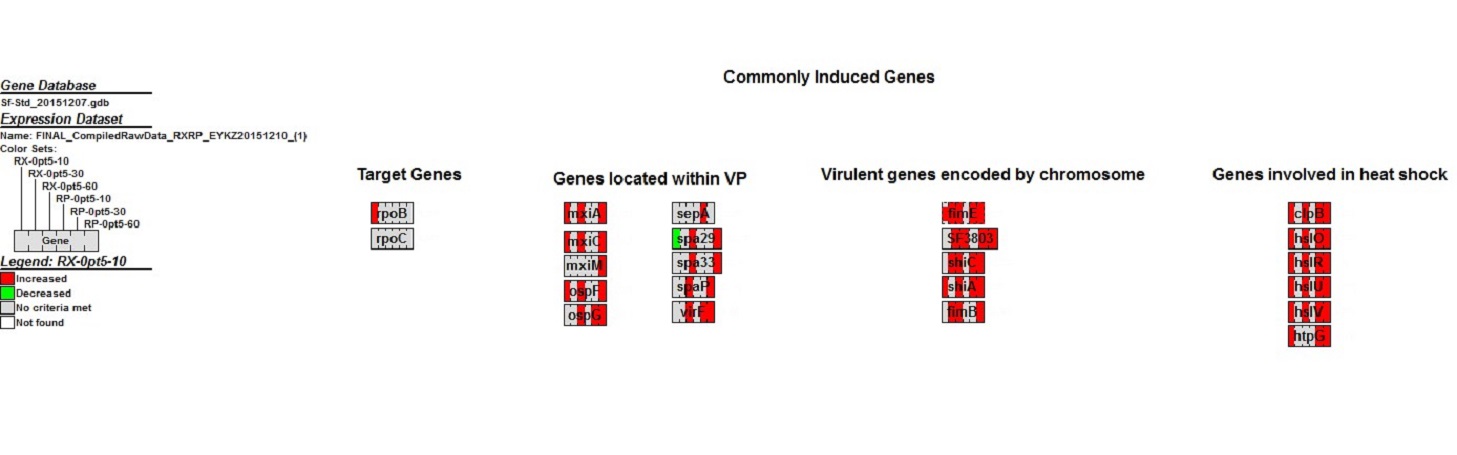


Figure 6 – MAPP of commonly induced genes RXRP 1xMIC 10 – 60 min

GenMAPP produced MAPPs analyzing pathways of commonly induced genes stated by Fu et al. The map shows a strong presence of induced genes throughout all time points of the 0.5 x MIC trials of RP and RX **(**Figure 6**)**. Fu et al. noted that some of the genes located within the VP showed time dependent rates of induction which is shown within the MAPP, *spa29* proves as an example(2).

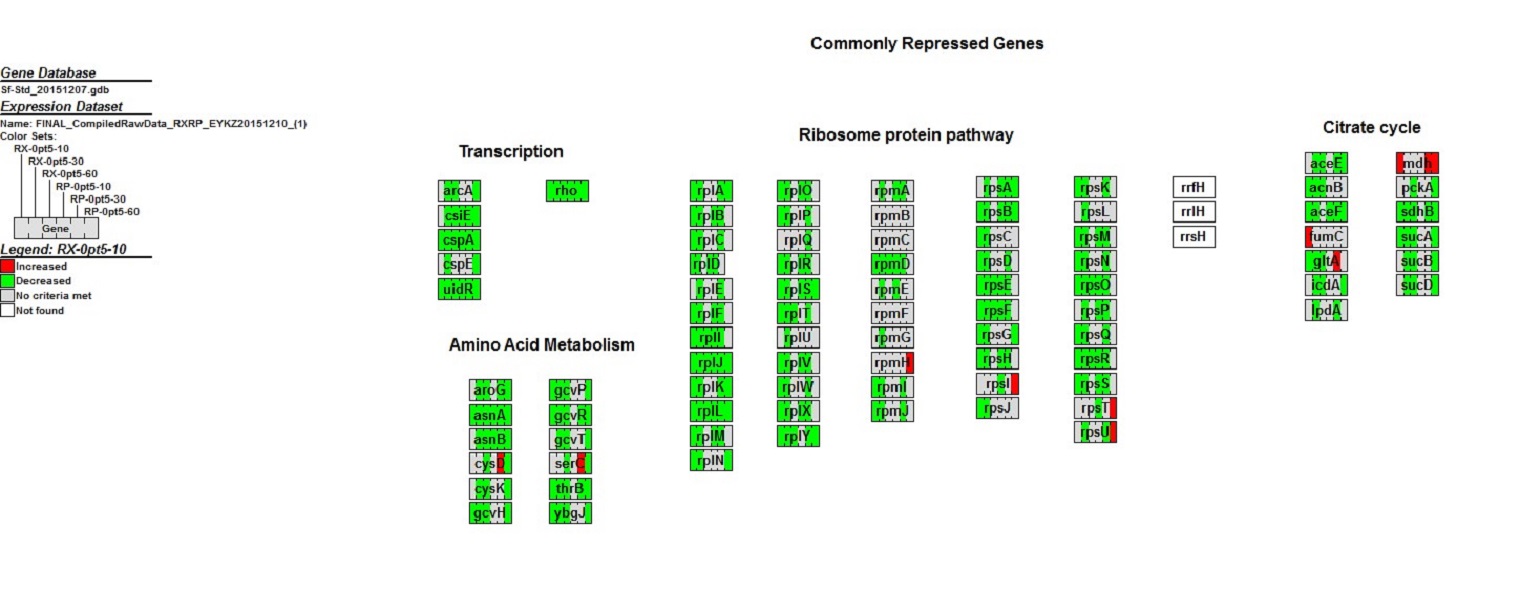


Figure 7 – MAPP of commonly repressed genes RXRP 1xMIC 10 – 60 min

MAPP analysis of pathways of commonly repressed genes stated by Fu. et al. There is a strong presence of repressed genes through all time points of the 0.5 x MIC RP and RX **(Figure 7)**. The gene *mdh* stands out for being induced within a generally repressed gene pathway, this gene codes for enzymes utilized within other pathways such as pyruvate metabolism, microbial metabolism in diverse environments, carbon metabolism, and biosynthesis of antibiotics.

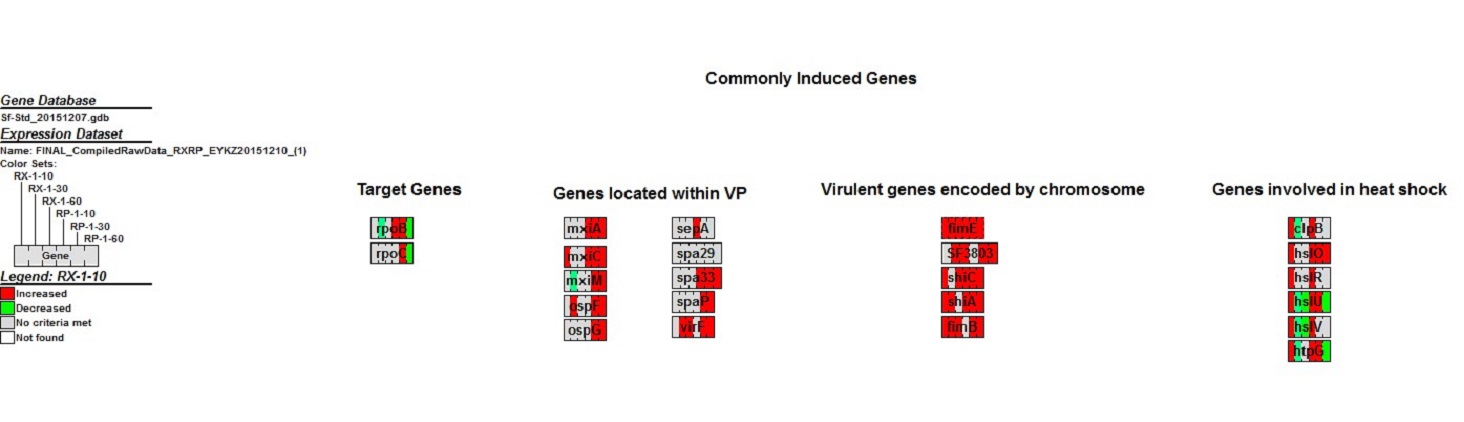


Figure 8 – MAPP of commonly induced genes RXRP 0.5xMIC 10 - 60 min

There is a strong reinforced presence of induced genes through all time points of the 1 x MIC trials of RP and RX, especially within genes that code for virulence. Genes involved in heat shock, however, indicate a time-dependent repression. Genes involved with heat shock exhibiting an initial induction followed by a time dependent repression during the 1xMIC trials include *hslU*, which codes for the protein ATP-dependent protease ATP-binding subunit; *hslV,* which codes for ATP-dependent protease peptidase subunit; and *htpG,* which codes for heat shock protein 90 (2, Figure 8).

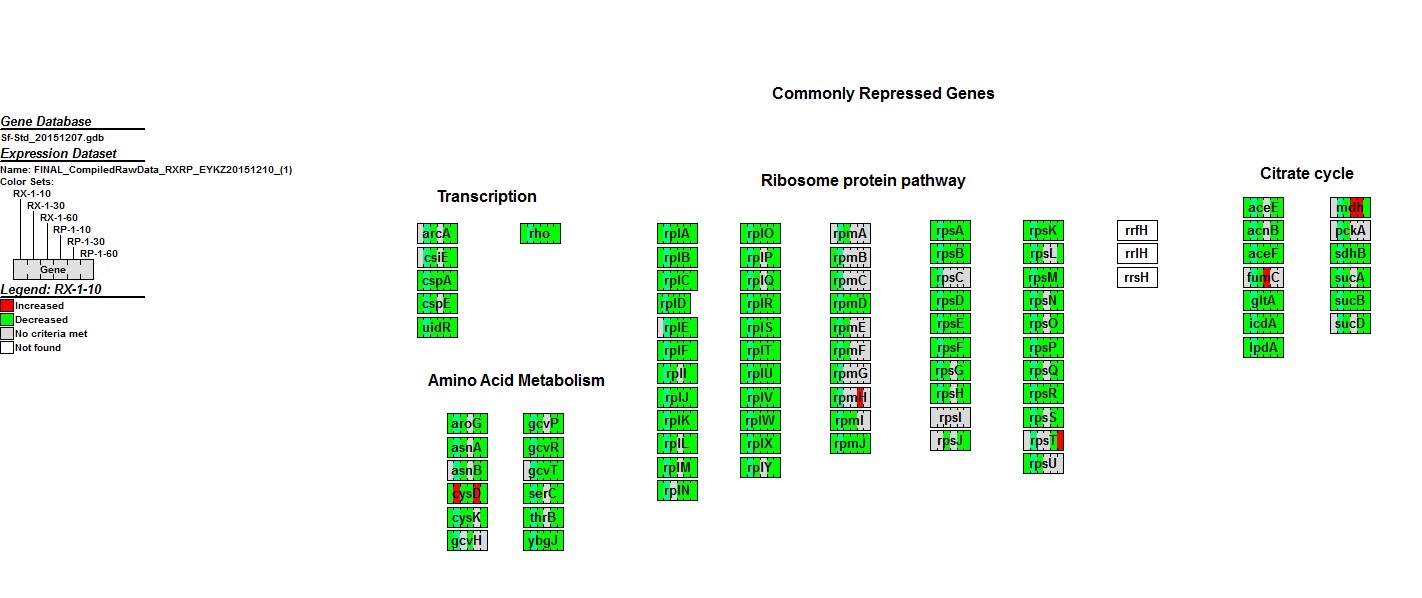


Figure 9 – MAPP of commonly repressed genes RXRP 0.5xMIC 10 – 60 min

There is a strong presence of repressed genes through all time points of the 1 x MIC trials of RP and RX, indicating that pathways regarding transcription, production of ribosomal proteins, citrate cycle, and amino acid metabolism are repressed **(**Figure 9**)**. Gene *mdh* stands out for being induced within a generally repressed gene pathway. This gene codes for enzymes utilized within other pathways such as pyruvate metabolism, microbial metabolism in diverse environments, carbon metabolism, and biosynthesis of antibiotics. *FumC* (fumarate hydratase) is an enzyme utilized within the citrate cycle and is present within the cytosol for amino acid metabolism when the cell experiences DNA impairment(7).

### **Discussion**

For the most part, the GenMAPP Builder process captured most of the valid IDs during the initial export, producing an overall count of 7567 ordered locus names from the <gene/> tag in the XML file. However, there was a bit of a setback on the testing side after some XML files were unknowingly imported twice to the initial gene database. Though the duplicated count of genes in TallyEngine did not affect the rest of the group, it still proved to be troublesome since the exports took more than twice as long as it should have been, and the cause of the extra counts was yet undetermined at the time. The UniProt and GOA files had to be re-imported and the previous build had to be re-run in order to produce the actual results.

Aside from the longer timeframe, the GenMAPP Builder export had no hiccups. It is to be expected that not every gene was found within the XML file using the initial customization, so when we figured out which genes were missing we customized GenMAPP Builder accordingly. We were missing 416 genes initially and were able to find 92 genes in the XML file that needed to be exported. Those 92 genes were included in the export by adding a single method within the source code of GenMAPP Builder. With this added method all the genes in the XML file that we could find were exported successfully. Specific changes that were made can be seen in the Results section.

One major discrepancy, however, still exists within the build. Ensembl Bacteria reports that there are only 4446 coding genes, but we have exported 7661 ordered locus names. The cause of this is that the IDs of the form SF#### represent the same genes as the ones represented by the IDs of the form S####.

When comparing values used to distinguish significant changes were compared in the average log-fold change along with T-stat values that were used to create P values. A value of less than 0.05 for statistical significance. Fu et al. used mean log ratios and within their calculations a change in gene expression was significant if the mean ratio was higher than a twofold change up or down (2). The criteria for the mean ratios is not clear, nor are their calculations identified, and only the results are shown. If a gene exhibited a twofold increase or decrease from the previous time point then it was considered significantly altered.

Our analysis considered genes significant if they showed greater than a 0.25 increase or decrease in average LogFC values and had a P value less than 0.05 (Table 3**)**. Given this criteria 18.8% of genes were significantly changed for RP at 0.5 x MIC after 10 minutes, with 22% noted as significant in the sanity check using only the P value. Additionally, 18.2% of genes were significantly changed for RX with the same treatment with the criteria of 0.25 increase or decrease in comparison to 20.8% in the sanity check. For RP at 1 x MIC after 60 minutes, 61% of genes were considered significantly changed in MAPPFinder and 64% in the sanity check, and for RX 56% were considered changed in MAPPFinder and 59% in the sanity check. Overall, between our analysis via MAPPFinder and the sanity check done in Excel, most changes in gene expression were similar, with the MAPPFinder analysis being slightly more stringent (Table 2).

In the sanity check a similar trend was seen with the percent of genes changed at the different time points. With both RX and RP the highest percent of genes changed was seen after *Shigella* was harvested at 30 minutes, regardless of the drug concentration added and even as the p value cutoff became smaller (Table 2). The general trend shows the smallest percentage of genes changed for RX and RP at all p values after 10 minutes, the largest at 30 minutes, and a middle-level percentage at 60 minutes. The exception is seen in RP-1-30 and RP-1-60, where there is almost no change or none at all, which provides a possibility for further analysis for this treatment.

The study noted genes that had altered expression levels within at least two experimental conditions for each drug. RX had a total of 535 genes altered and RP showed 367 genes altered (2). MAPPFinder results bolster claims made in Fu. et al while providing new insight into their data. A significant change of expression was identified if the mean ratio of a gene is higher than twofold up- or down-regulated within their study. The data for *rpoBC* was not logged within the Excel file named within the paper (Table S3) as a search function of the gene came up empty (2). These genes were thought to be induced 1.6 to 4.0 fold. Genes *rpoBC* and *rpoD* are the only two genes described with a fold change (2). All others are simply upregulated or downregulated.

Within the transcription pathway several genes were continually repressed. The gene *rho* was targeted because of the expectation that as there is less transcription occurring, so the free *rho* factors binding to mRNA will show a decreased gene expression throughout the trials (1). This study found that there was reduced gene expression throughout the experiment along with *csiE* coding for stationary phase inducible protein, major cold shock protein gene *cspA*, and *uidR*, that codes for a transcriptional repressor for *uid* operon(2). Amino acid metabolism shows a continual repression as well, which is expected because of the downregulation in transcriptional genes (Figures 9**,** 7). Genes associated with ribosomes, amino acids, transcription, and the citrate cycle exhibited a continual inhibition of gene expression throughout the trials **(**Figures 9**,** 7**)**.

The Criterion-GO files provided insight into the pathways with greater rates of expression and inhibition for 0.5 x MIC at 10 minutes and 1 x MIC at 60 minutes. These pathways involved flagellar assembly and motility, ribosomal activities, and metabolism. Genes that were typically induced or repressed within these pathways showed corresponding data at different time points as well. There were several genes that showed time dependency including *hslU, hslV,* and *htpG*,which displayed decreased rates of expression overtime. Heat shock genes met the criteria as a group of genes cited as commonly upregulated within the paper that showed altered levels of expression(2). These results of 1 x MIC RX vs RP show the initial upregulation with later time points gene expression significantly decreasing (Figure 8**)**. Genes encoded with virulence were typically induced, such as genes within the virulence plasmid and on the chromosome. Known genes associated with virulence located on the chromosome such as *fimE, shiC, shiA, fimB,* and *SF3803* show increasing rates of induced gene expression as stated in the paper **(**Figures3**,** 8**)**. Genes located within the flagellar assembly pathway including *flgG, flgI, flhC, fliD, fliR,* and *fliT* were found to be commonly induced in a consistent manner with the Fu et al. study (2, Figures 4, 5).

Due to the similarity in mechanisms of action for both drugs, one would expect similar reactions among pathways. There are many instances where gene expression differs among the drugs and this warrants future possible research regarding the mechanism of action and what can be hypothesized about such proteins interaction within the system. Notably, the genes encoded within the virulence plasmid showed indications of induced expression along with other virulent genes on the chromosome, which appeared to be induced at an even higher level. Along with the flagellar assembly pathway these are strong virulent factors that could possibly be harmful to a host. There are many hypothetical genes that were altered, shedding light on potential protein functions within already established and well-known pathways that be used to understand and predict mechanisms of resistance in *Shigella flexneri.*

### **Conclusion**

With respect to the gene database creation, the number of genes seemed to closely match the ones that the external gene database has if the IDs of the form S#### is not taken into account. Since the SF#### and S#### ordered locus names represent the same IDs in the UniProt XML file, which would mean that we are counting the same genes twice, if we dismiss ~3200 ordered locus names with the form S####, then we would get approximately the same number of coding genes reported on Ensembl Bacteria, i.e., 4446 coding genes. If the project were to be continued, we would have to adjust more of the existing GenMAPP Builder code to ignore the ordered locus names of the form S####. Additionally, in terms of the processes involved in the microarray data, we would like to see a dye swap with the Cy5-dCTP and Cy3-dCTP dyes. Swapping the dyes and averaging the ratios would allow for compensation in dye bias that may have been present. This study revealed that rifamycin treatments are shown to be inhibitory but at the cost of the bacteria enhancing processes related to pathogenicity. Our findings support the conclusions of the previous study by Fu et al., while providing greater insight into gene pathways affected by rifampin and rifaximin (2). Due to the parameters set within GenMAPP we were able to see the activity of genes and make comparisons not only based on the data but a visual representation of the data. The RX and RP trials showed that the steric hindrance caused by the binding of the drugs to the beta subunit of RNA polymerase has a strong effect on *Shigella flexneri* but the antibiotic also causes genes associated with motility and secretion systems to become upregulated, increasing the virulence of the bacteria.

### **Acknowledgments**

First and foremost, we would like to thank Dr. Kam Dahlquist and Dr. John Dionisio for their advice and direction on this project. We have learned a great deal over the course of this endeavor and are very grateful for their constant encouragement and collaboration. We would also like to thank Seaver College of Science and Engineering at Loyola Marymount University, specifically the Biology and Computer Science Departments, for providing us with this opportunity to contribute to the scientific community. Lastly, we would like to acknowledge our peers in the Fall 2015 Biological Databases class for their moral support.

### **References**

### 1. Alifano, P., Rivellini, F., Limauro, D., Bruni, C. B. & Carlomagno, M. S. (1991). A consensus motif common to all Rho-dependent prokaryotic transcription terminators. Cell 64, 553–563

### 

### 2. Fu H, Liu L, Zhang X, Zhu Y, Zhao L, Peng J, et al. (2012) Common Changes in Global Gene Expression Induced by RNA Polymerase Inhibitors in ''Shigella flexneri''. PLoS ONE 7(3): e33240. doi:10.1371/journal.pone.0033240

3. Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., … Yu, J. (2002). Genome sequence of ''Shigella flexneri'' 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Research*, 30(20), 4432–4441.

4. Kotloff K.L., Winickoff, J.P., Ivanoff, B., Clemens, J.D., Swerdlow, D.L., Sansonetti, P.J., Adak, G.K., Levine, M.M. (1999) Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ*,

5. Salomonis, K Hanspers, AC Zambon, K Vranizan, SC Lawlor, KD Dahlquist, SW Doniger, J Stuart, BR Conklin, and AR Pico. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics, Jun 2007; 8: 217* 7, 651–666.

6. Waligora, E.A., Fisher, C.R., Hanovice, N.J., Roudou, E.E., Wyckoff, Payne, S.M. (2014). Role of Intracellular Carbon Metabolism Pathways in *Shigella flexneri* Virulence. *Infection and Immunity* 82(7):, 2746-2755.

7. Yogev O, Singer E, Shaulian E, Goldberg M, Fox TD, Pines O. Fumarase: a mitochondrial metabolic enzyme and a cytosolic/nuclear component of the DNA damage response. PLoS Biol. 2010;8:e1000328.

Appendix

**Gene Database Testing Report for *Shigella flexneri* 2a str. 301**

**Export Information:**

Version of GenMAPP Builder:

* **gmbuilder-3.0.0-build-5**

Computer on which export was run:

* **LMU Seaver 120 computer: front of the room, 3rd computer from the right**

Postgres Database name:

* **Shigella\_flexneri\_20151208**

UniProt XML filename:

* UniProt XML version (The version information can be found at [the UniProt News Page](http://uniprot.org/news) < http://uniprot.org/news>):
  + **UniProt release 2015\_11**
* UniProt XML download link:
  + <**http://www.uniprot.org/uniprot/?query=proteome:UP000001006>**
* Time taken to import:
  + **4.43 minutes**

GO OBO-XML filename:

* GO OBO-XML version (The version information can be found in the file properties after the file downloaded from the [GO Download page](http://beta.geneontology.org/page/download-ontology) < http://archive.geneontology.org/latest-termdb/go\_daily-termdb.obo-xml.gz> has been unzipped):
  + **Version created on 11/19/2015 (at 2:24 AM)**
* GO OBO-XML download link:
  + <**http://archive.geneontology.org/latest-termdb/go\_daily-termdb.obo-xml.gz>**
* Time taken to import:
  + **6.84 minutes**
* Time taken to process:
  + **5.49 minutes**

GOA filename (give filename and upload and link to compressed file):

* GOA version (News on the UniProt – GOA page <<http://www.ebi.ac.uk/GOA>> records past releases; current information can be found in the Last modified field on the FTP site <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>):
  + **Version released on 11/11/2015.**
* GOA download link:
  + **<http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/103.S\_flexneri\_301.goa>**
* Time taken to import:
  + **0.06 minutes**
* Name of .gdb file (give filename and upload and link to compressed file): [**Sf-Std 20151208.gdb**](https://xmlpipedb.cs.lmu.edu/biodb/fall2015/images/f/fe/Sf-Std_20151208.gdb)
* Time taken to export:
  + **2 hour, 0 minutes, 27 seconds**
* Start time:
  + **9:35:00 PM PDT**
* End time:
  + **11:35:27 PM PDT**

**Using TallyEngine:**

· With the necessary files import to PostgreSQL, TallyEngine was run and the following table is the result of the export:

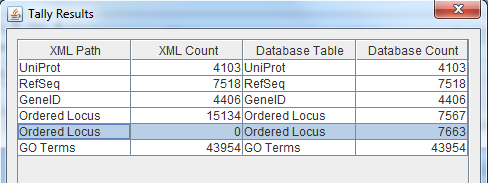


Table 7 - Final result of TallyEngine Match and PostgreSQL counts

**Using XMLPipeDB match to Validate the XML Results from the TallyEngine:**

· Two separate, almost identical regex, were used in order to find more ordered locus names in the XML file from just within the <gene/> tag. The one below found 7567 IDs:

regex1.png

· The one below found 3 IDs:

regex2.png

· When added together, the results become 7566 + 3 = 7569.

· Since there were ID duplicates between the <gene/> and <dbReference/> tags, there was no easy way to actually find newer IDs without miscounting. Therefore, this initial regex count was kept since it is the closest number we could get to our database export count (7569 vs. 7663).

**Using SQL Queries to Validate the PostgreSQL Database Results from the TallyEngine**

· For our specific specie, a specialized PSQL query was made thanks to our advisor, Dr. John David Dionisio. Below is a screenshot of the query in action, which produced 7660 entries:

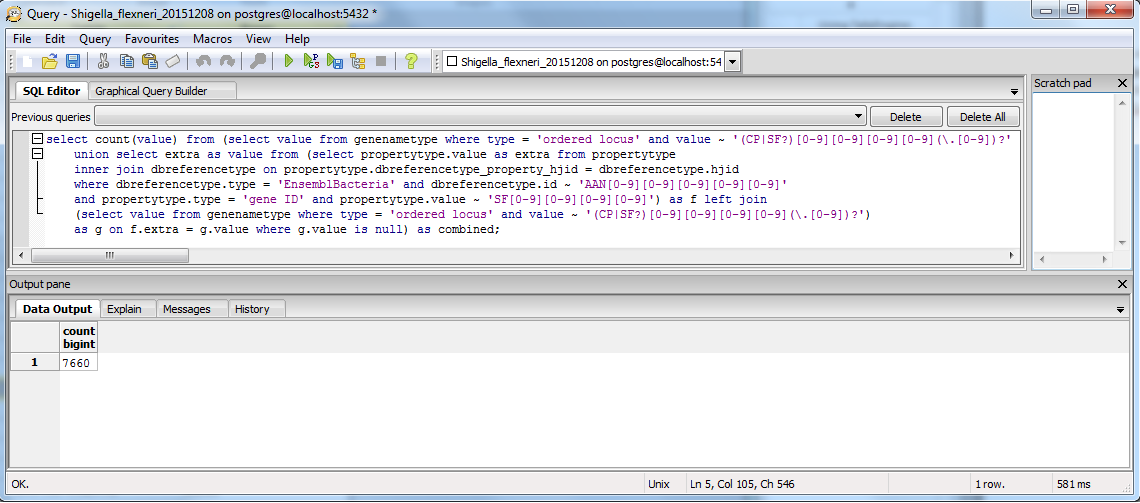


Figure 10 - PostgreSQL query specific to *Shigella flexneri*

**OriginalRowCounts Comparison**

· In the .gdb file, which was opened in Microsoft Access, the OriginalRowCounts table was inspected in case the export did not work as we intended. With the additional ~92 IDs that was found from the <dbReference/> tag, the total count should be 7569 + 92.

· In this table, the OrderedLocusNames row was determined to contain the number of IDs that we were expecting, as seen from the table below.

· Other than the table that we were expecting to be changed, the rest of the rows seemed to be kept intact when compared with the “benchmark” build that we made previously (Build 2).

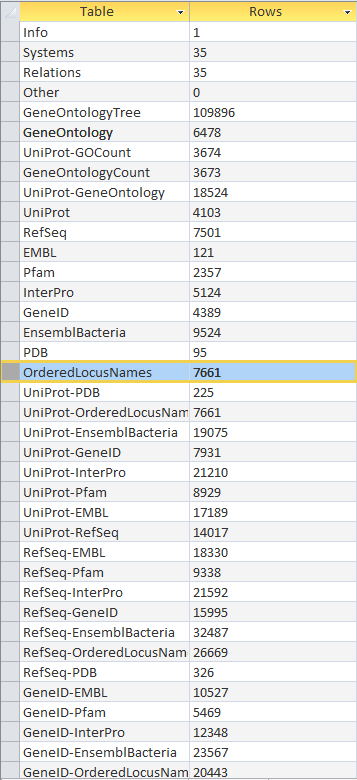


Table 8 - OriginalRowCounts table as seen in Microsoft Access

**Visual Inspection**

A visual inspection was performed on the individual tables to see if there are any problems. Primarily, the Systems table was checked for dates. There were no problems in the following tables:

· GeneOntology

· InterPro

· GeneID

· RefSeq

· UniProt

· EMBL

· PDB

· Pfam

· OrderedLocusNames

· EnsemblBacteria

Additionally, the following tables seemed to all have the correct forms of IDs:

· UniProt

· Refseq

· OrderedLocusNames

**Download .gdb File**

The resulting .gdb file can be downloaded in the link provided below:

· <https://xmlpipedb.cs.lmu.edu/biodb/fall2015/images/b/b8/Sf-Std\_20151214.gdb>.