Transcriptional Analysis of the Developmental Stages of *Chlamydia trachomatis*

Hilda Delgadillo

Katrina Sherbina

Dillon Williams

BIOL 367 Biological Databases

December 13, 2013

Introduction

Chlamydia trachomatis is among the most clinically significant human pathogens. It is responsible for the leading cause of infectious blindness which is rather serious in developing countries. The particular strain that was studied within the microarray paper is Serovar L2 which causes sexually transmitted diseases. However, the DNA with which this serovar was compared to on the microarray chips were that of Serovar A/HAR-13 which causes blinding trachoma. Moreover, *Chlamydia trachomatis* has an obligate intracellular nature which makes it hard to research since it requires host cells to survive. Furthermore, there has been a rise in the number of infections in recent years, so there is an increasing need for understanding the virulence of *Chlamydia trachomatis*. Although, there is much still waiting to be uncovered about this particular pathogen we are familiar with the primary and fundamental developmental stages of this pathogen. The Elementary Body (EB) of the pathogen attaches to the cell through chemoreceptors. This EB developmental body goes through phagocytosis by the host cell. Within the phagosome there is a conversion of EB to a Reticulate Body (RB) developmental infectious form of the pathogen. There is replication of the RB by binary fission and about 40 to 48 hours later the RBs are converted to EBs. At this time, the host cell lyses where the infectious EBs are released and begin to spread its virulence. In addition, *Chlamydia trachomatis* takes the nutrients of its host, during an infection, such as nucleotides, amino acids, and lipids since it is devoid of several enzymes and some entire metabolic pathways. Hence, these mentioned nutrients are essential for *Chlamydia trachomatis* to develop.

Omsland *et al.* (2012) investigated the developmental stages and the necessary requirements needed for protein synthesis in the EBs and RBs of *Chlamydia*. Since this pathogen has an obligate intracellular nature as mentioned previously, it was necessary for the researchers

to provide a cell-free (axenic) media, so there could be a concise distinguishing between *Chlamydia trachomatis* and its host cells performance of metabolic pathways. This axenic media is known as *Chlamydia* intracellular phosphate-1 (CIP-1) which consists of nutrients including DTT which specifically reduces the EB disulfide bonds promoting an axenic metabolic activity. This same medium was used for both developmental bodies, EBs and RBs. Furthermore, the antibiotic Rifimpacin was introduced to the EBs and RBs along with the medium. Along with this antibiotic, which inhibits transcription, there was an incorporation of [^{35}S]Cys-Met which is a Sulfur isotope that can bind specifically to proteins. Additionally, since the presence of this isotope was only seen by chance it indicates that *Chlamydia trachomatis* displays de novo transcription. Therefore, we would like to analyze the genes and their significance in transcription metabolism using GenMAPP and MAPPFinder, so to do so we have created our database with GenMAPP builder along with XMLPipeDB. With the following process we are able to view that indeed Elementary Bodies are not so metabolically inert after-all just as the paper concludes due to its de novo protein synthesis while Reticulate Bodies also display performance in de novo synthesis.

Methods

Determination of the Model Organism Database for *C. trachomatis* A/HAR-13

The EnsemblBacteria database page for *C. trachomatis* A/HAR-13 (http://bacteria.ensembl.org/chlamydia_trachomatis_a_har_13/Info/Index) was determined to be the model organisms database for the species. Browsing through the chromosome maps and other files in the EnsemblBacteria database for the species of interest, it was determined that

"CTA_####" was the primary gene ID that should appear within the gene database to be created for the species.

Creation of a GenMAPP-compatible Gene Database

GenMAPP Builder 2.0b71 (available from https://sourceforge.net/projects/xmlpipedb/) and PostgreSQL for Windows (available from http://www.postgresql.org/) were downloaded in order to export a gene database for *C. trachomatis* A/HAR-13. A new database was created in PostgreSQL into which a collection of tables was imported by running the queries in the file *gmbuilder.sql* located in the GenMAPP Builder program folder. Three files were imported into GenMAPP Builder in the following order with PostgreSQL running in the background: UniProt XML file specific to *C. trachomatis* A/HAR-13, OBO-XML file, and Uniprot-GOA file specific to *C. trachomatis* A/HAR-13. The OBO-XML file was processed after its import. The UniProt XML file contains the proteomic information for the strain. The Uniprot-GOA file associates the gene products of the species with the Gene Ontology terms specified in the OBO-XML file. After successfully importing the aforementioned files, a gene database export was performed through GenMAPP Builder. This first step in this process involves exporting information in the XML files into the relational database tables that were created in PostgreSQL as previously described. The second step involves the actual creation of the GenMAPP compatible database by extracting the information in the relational database and associating each gene with gene ontology terms (LMU BioDB 2013, 2013). A more detailed protocol of the GenMAPP Builder process can be found in the Loyola Marymount University's Fall 2013 Biological Databases wiki (refer to https://xmlpipedb.cs.lmu.edu/biodb/fall2013/index.php/Running_GenMAPP_Builder).

Inspection and Validation of the Gene Database

Five methods were used to validate the newly created gene database. First, the Tally Engine in GenMAPP Builder was run to check if all of the gene products, ordered locus names and GO terms contained in the XML files were transferred successfully to the relational database in PostgreSQL. Of the five counts listed for the XML files and the database, the counts for the UniProt IDs and the Ordered Locus IDs were used in final validation of the gene database (Fig. 1). Within PostgreSQL, a query was run to count the number of gene IDs in the newly created database. These gene IDs were designated as "ordered locus names" in the table genenametype. This number of gene IDs was then verified using the XMLPipeDB Match utility (available from https://sourceforge.net/projects/xmlpipedb/) through the command line. After using the aforementioned tools, the database was opened in Access to view the number of ordered locus names listed in the OriginalRowCounts table. All of the aforementioned counts were then compared to the number of coding genes listed in the statistics provided within the EnsemblDatabase for *C. trachomatis* A/HAR-13 (refer to http://bacteria.ensembl.org/chlamydia_trachomatis_a_har_13/Info/Annotation/#assembly). A more detailed protocol for performing the counts using TallyEngine, SQL queries in PostgreSQL, XMLPipeDB Match, and Access can be found within Loyola Marymount University's Fall 2013 Biological Databases wiki (refer to https://xmlpipedb.cs.lmu.edu/biodb/fall2013/index.php/How_Do_I_Count_Thee%3F_Let_Me_Count_The_Ways).

Modification of the GenMAPP Builder 2.0b71 Code

To address any concerns that arose during the inspection of the gene database, alterations or additions to the GenMAPP Builder 2.0b71 code were made within Eclipse IDE for Java EE

Developers (available from http://www.eclipse.org/downloads/) with Subclipse (available from http://subclipse.tigris.org/) and Java SE 7u45 (available from http://www.oracle.com/technetwork/java/javase/downloads/index.html).

Acquiring, Normalizing, and Organizing the Microarray Data for GenMAPP

The microarray raw data was downloaded from ArrayExpress (accession number E-GEOD-39530). The raw data was stored on Affymetrix chips that required special software to read. The Affymetrix RML Custom Pathogenic chip 3 CDF file was needed to analyze data with dChip from NCBI GEO (platform GPL4692). The dChip software was downloaded from the dChip website (available at: https://sites.google.com/site/dchipsoft/home). The tutorial guide was then used to normalize the data. The dChip software was used to open the raw data and the raw data was then exported into an excel spreadsheet. The sdrf file (also found on ArrayExpress website) had to be used in to locate the experimental groups in the raw data. The groups to be compared were: EB in axenic media (which contained 2 replicates) to RB in axenic media (which contained 4 replicates), EB + rifampicin (which contained 3 replicates) to RB + rifampicin (which contained 3 replicates), there were also 4 replicates labeled "carry over of EB" that were not relevant to analysis being done in this experiment, so they were ignored. The following calculations were carried out: the average of each group (EB, RB, EB in presence of rifampicin, RB in presence of rifampicin) was computed; the rations of EB to RB and EB in the presence of rifampicin to RB in the presence of rifampicin were computed; the resulting ratios were Log2 transformed, the p values comparing the EB and RB groups and EB in the presence of rifampicin and RB in the presence of rifampicin groups using the TTEST function in excel were computed. A new sheet was created for GenMAPP. The tags that Affymetrix appended to each probe were separated from the actual gene ID. All gene ID's other than those pertaining to *C.*

trachomatis were deleted. GenMAPP and MAPPFinder were then ran using the gene database. A Created color set for data analysis in both the presence and absence of rifampicin was created.

GenMAPP and MAPPFinder Analysis of the Microarray Data

Two color sets were created in GenMAPP: one for the absence of rifampicin and one for the presence of rifampicin. Two criterion were constructed for each color set: increased gene expression denoted by an average log fold change that is less than 0.25 with p value less than 0.05 and decreased gene expression denoted by an average log fold change of less than -0.25 with p value less than 0.05.

Results

Quantity and Identity of the Genes in the Gene Database

Upon vetting our *Chlamydia trachomatis* database with the Tally Engine, the number of Ordered Locus matched with both the XML count and the database created for *Chlamydia trachomatis* A/HAR-13 (Fig. 1). Furthermore, we did a count through the SQL Query and it matched the 917 count of the Tally Engine. However, when inspecting the OriginalRowCounts table in Access the number of unique IDs found for OrderedLocusNames was 919 which deviated from the previous counts . Moreover, the number of unique IDs were also inspected through XMLPipeDB Match which gave us a count of 911 (Fig. 2), which brought us to inspect Access again to see why there were more reported counts in comparison to the rest of the quality assurance methods. We found gene IDs with as such, pCTA_####, which refer to genes in the plasmid, also present in addition to the grand majority of gene IDs, CTA_####. Hence, we went back to XMLPipeDB Match and inserted a "[p|]…" in our search which resulted in a count of 8 gene IDs (Fig. 3). We found the last eight IDs that would sum up to 919 which correspond to the previous count result from OriginalRowCounts table in Access. Additionally, the discrepancy in

the counts were due to the separation or lack thereof of the gene ID CTA_0406/CTA_0407/CTA_0408 which due to the joining of these names by slashes indicates that these ordered locus names actually refer to one gene, but were counted as three individual genes in XMLPipeDB Match, in the OriginalRowCountsTable in Access, and our Model Organism Database (MOD) (Table 1).

XML Path	XML Count	Database Table	Database Count
UniProt	917	UniProt	917
Ordered Locus	917	Ordered Locus	917
RefSeq	926	RefSeq	926
GeneID	926	GeneID	926
GO Terms	40071	GO Terms	40071

Fig. 1. The TallyEngine tool in GenMAPP Builder 2.0b71 was used to compare the total ID count in the XML files to that of the gene database created for *Chlamydia trachomatis* A/HAR-13. Only the UniProt ID and Ordered Locus ID counts were considered in the validation of the gene database.

```
java -jar xmlpipedb-match-1.1.1.jar "CTA_[0-9][0-9][0-9][0-9]" <
Uniprot_XML_C.trachomatis_serovar_A_KS_20131114.xml
```

Fig. 2. This is the query used in XMLPipeDB Match which presented a count of 911, so we were prompted to go back to Access to figure out where the discrepancy lies.

```
java -jar xmlpipedb-match-1.1.1.jar "[p|]CTA_[0-9][0-9][0-9][0-9]" <
Uniprot_XML_C.trachomatis_serovar_A_KS_20131114.xml
```

Fig. 3. This figure demonstrates the new XMLPipeDB Match query with the addition on the "[p|]…" which corresponds to the additional IDs found in Access. The new query came up with 8 unique IDs which once added to 911 the queries would consist of 919 unique IDs which is validated by Access.

Table 1. Five different tools were used to determine the number of genes in the relational gene database created for *Chlamydia trachomatis* A/HAR-13. The differences in the counts are a result of the separation of the gene ID CTA_0406/CTA_0407/CTA_0408 into separate IDs in some databases but not others.

Counting Method	**Count**
TallyEngine (UniProt IDs and Ordered Locus Names)	917
SQL Query	917
XMLPipeDB Match	919
OriginalRowCountsTable in Access	919
Model Organism Database (EnsemblBacteria)	919

Customization of the GenMAPP Builder 2.0b71 to Accommodate the Requirements for Exporting a Gene Database for *C. trachomatis* A/HAR-13

A custom species profile was added to the existing GenMAPP Builder 2.0b71 code in order to customize the gene database export process for *C. trachomatis*. This custom species profile documents the species name as well as a species specific database link: http://bacteria.ensembl.org/chlamydia_trachomatis_a_har_13/Gene/Summary?g=~. This URL is a general link to a gene entry in the EnsemblBacteria database for *C. trachomatis*, where the ~ stands in for an actual gene ID. With this link in the GenMAPP Builder code, it is possible to click on a gene in a pathway visualized within GenMAPP and navigate to the entry within the EnsemblBacteria database for that particular gene.

In addition to the custom species profile, the TallyEngine within GenMAPP Builder was customized for *C. trachomatis*. To perform this customization, it was necessary to determine how the genes are tagged within the UniProt XML file and where those gene tags are located within the relational database. Through searching for the sample gene tag CTA_0407, it was

found that the UniProt XML identifies all the genes with the label “ordered locus”. From previous SQL queries that were run to vet the gene database, it was determined that the location of the genes is in the table genenametype within the relational database.

After the creation of the custom species profile and the customization of the TallyEngine, the gene database export was performed again through GenMAPP Builder using the same PostgreSQL database as previously described. The testing report for this export (Fig. S1) and a schema describing the connections between all the tables in the relational gene database (Fig. S2) are included in the Supplementary Data. The testing report documents the discrepancy between the different counting methods due to the separation of the ID CTA_0406/CTA_0407/CTA_0408 in some but not all of the counting methods used to validate the database. However, no modifications were made to the GenMAPP Builder 2.0b71 code to address this discrepancy because only one of the ordered locus names in the ID, specifically CTA_0407, appeared in the microarray data. It is likely that this will remain the case for other microarray platforms for *C. trachomatis.*

Statistical Analysis of Microarray Data

Based on the results (Table 2), it would appear that rifampicin seems to have a considerable effect on gene expression. One speculation is that the larger number of genes up-regulated in the presence of rifampicin in comparison to the absence of rifampicin is a result of rifampicin's transcriptional inhibition of the differentiation from the EB stage to the RB stage. A possible explanation for the smaller number of genes down-regulated in the presence of rifampicin in comparison to the absence of rifampicin could be the transcriptional inhibition of the differentiation of RB back to EB in the lifecycle of *C. trachomatis.*

There were concerns regarding the statistical analysis of the data collected in the absence of rifampicin, as an incomplete data set was used to procure the results. The microarray data for EB in the absence of rifampicin only contained two replicates of the original four that had been performed by Olmsland *et al.* (2012). As the t-test that was run compares the mean of the EB population to the mean of the RB population, the concern was that, with only two replicates, it is difficult to make an accurate comparison of the two means. The average in this case had the potential to be thrown off by outliers, and the only way to mitigate this would be by having more replicates. As a result, the lack of a sufficient number of replicates for EB decreased the confidence in the statistical analysis.

Table 2. Four different p value cut-offs as well as two average log fold change (one for increased gene expression and one with decreased gene expression) cut-off with p values of less than 0.05 were used to determine the significant number of genes changed in the presence and absence of rifampicin. These statistics show that rifampicin has a considerable effect on gene expression.

Filter(s) Used	**No Rifampicin**	**Percent Changed***	**Rifampicin**	**Percent Changed°°**
p value < 0.05	864	93.40540541	471	50.91891892
p value < 0.01	794	85.83783784	308	33.2972973
p value< 0.001	676	73.08108108	93	10.05405405
p value < 0.0001	235	25.40540541	26	2.810810811
average log fold change > 0.25 & p value < 0.05	82	8.864864865	169	18.27027027
average log fold change < -0.25 & p value < 0.05	778	84.10810811	302	32.64864865

GenMAPP and MAPPFinder Analysis Results

Significant GO terms were determined for increased and decreased gene expression using the data collected in the presence and absence of rifampicin. In both the presence and absence of rifampicin, there were more GO terms that showed up for the decreased criterion (see Table 3,4,5, and 6).

The MAPP for cellular carbohydrate metabolic process was analyzed. This MAPP, which resulted from the analysis of the data collected in the presence of rifampicin, was chosen as a result of the concerns regarding the lack of a sufficient number of EB replicates in the data collected in the absence of rifampicin. In the MAPP for the cellular carbohydrate metabolic process (Fig. 4), the genes involved in the glycogen biosynthesis pathway are up-regulated. As glucose-6-phosphate is a key intermediate in this pathway (Ahern, 2009), the increase in gene expression supports the conclusion presented by Omsland et al. (2012) that EBs preferentially use glucose-6-phosphate.

Table 3. The top GO terms pertaining to genes up-regulated in rifampicin. The following filters were used on the original MAPPFinder results to obtain this table: Z score is greater than 2, p Value is less than 0.05, number changed is greater than or equal to 3 and less than 150, and percent changed is greater than 13.

GO Terms Significantly Increased (No Rif.)	Z Score	P Value	Number Changed	Percent Changed
pathogenesis	3.59	0.017	3	27.27273
multi-organism process	3.59	0.017	3	27.27273
macromolecule modification	2.707	0.018	5	13.88889

Table 4. The top GO terms pertaining to genes down-regulated in rifampicin. The following filters were used on the original MAPPFinder results to obtain this table: Z score is greater than 2, p Value is less than 0.05, number changed is greater than or equal to 3 and less than 150, and percent changed is greater than 50.

GO Terms Significantly Decreased (No Rif.)	Z Score	P Value	Number Changed	Percent Changed
macromolecule biosynthetic process	2.967	0.002	146	96.68874
cellular macromolecule biosynthetic process	2.944	0.002	145	96.66666
translation	2.902	0.004	88	98.8764
gene expression	2.705	0.004	135	96.42857
metal ion binding	2.76	0.013	82	98.79518
cation binding	2.435	0.021	84	97.67442
cytoplasmic part	2.004	0.034	68	97.14286
lipid metabolic process	2.027	0.035	37	100
cellular lipid metabolic process	2.027	0.035	37	100
carbohydrate derivative metabolic process	2.252	0.04	77	97.46835
organonitrogen compound metabolic process	2.098	0.043	123	95.34884
macromolecular complex	2.039	0.043	83	96.51163
organophosphate biosynthetic process	2.169	0.045	42	100

Table 5. The top GO terms pertaining to genes up-regulated in the presence of rifampicin. The following filters were used on the original MAPPFinder results to obtain this table: Z score is greater than 2, p Value is less than 0.05, number changed is greater than or equal to 4 and less than 100, and percent changed is greater than 25.

GO Terms Significantly Increased (Rif.)	**Z Score**	**P Value**	**Number Changed**	**Percent Changed**
isomerase activity	2.664	0.013	8	29.62963
cellular carbohydrate metabolic process	2.852	0.021	4	44.44444
intramolecular transferase activity	2.351	0.038	4	36.36364
pathogenesis	2.351	0.041	4	36.36364
multi-organism process	2.351	0.041	4	36.36364
energy derivation by oxidation of organic compounds	2.351	0.044	4	36.36364

Table 6. The top GO terms pertaining to genes down-regulated in the presence of rifampicin. The following filters were used on the original MAPPFinder results to obtain this table: Z score is greater than 2, p Value is less than 0.05, number changed is greater than or equal to 4 and less than 100, and percent changed is greater than 25.

GO Terms Significantly Decreased	Z Score	P Value	Number Changed	Percent Changed
translation	9.501	0	74	83.14606
ribonucleoprotein complex	9.252	0	52	96.2963
structural constituent of ribosome	9.147	0	51	96.22642
ribosome	9.147	0	51	96.22642
non-membrane-bounded organelle	9.097	0	54	93.10345
intracellular non-membrane-bounded organelle	9.097	0	54	93.10345
organelle	8.92	0	54	91.52542
intracellular organelle	8.92	0	54	91.52542
cellular protein metabolic process	8.711	0	84	73.68421
structural molecule activity	8.704	0	52	91.22807
cytoplasmic part	7.966	0	57	81.42857
rRNA binding	7.075	0	33	94.28571
ribosomal subunit	4.285	0	13	92.85714
organelle part	3.94	0	13	86.66666
intracellular organelle part	3.94	0	13	86.66666
protein folding	3.028	0.002	9	81.81818
small ribosomal subunit	2.91	0.002	7	87.5

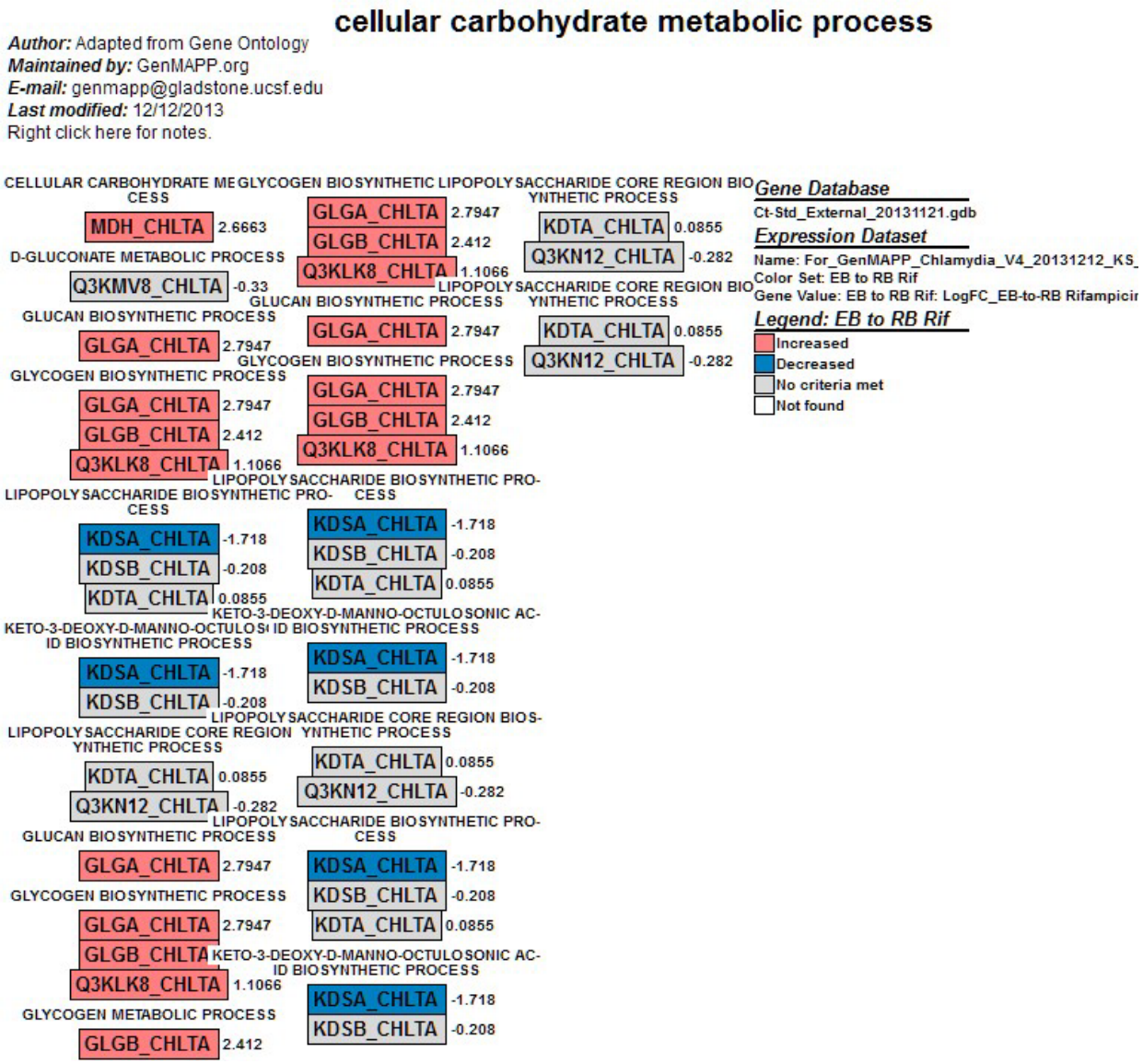

Fig. 4. MAPPP of the cellular carbohydrate metabolic process. This was one of the top ranked terms among the up-regulated genes in the presence of rifampicin.

Discussion

The GenMAPP Builder process for *C. trachomatis* A/HAR-13 was completed with few errors and modifications to the source code. The import of the UniProt and GOA association files and the export of the gene database were executed with few problems. After inspection of the gene database, few customizations were necessary to make to the GenMAPP Builder code to

improve the gene database export process for *C. trachomatis* A/HAR-13. The custom species profile that was created for *C. trachomatis* A/HAR-13 followed a standard template to which the only modifications were to the species name and a species-specific database link. In addition, the customization to the Tally Engine were not necessary because the discrepancy in the gene ID counts noticed during the validation process for the gene database was easily explained by the unique construction of one of the gene IDs.

The inspection of the gene database revealed that one gene ID, namely CTA_0406/CTA_0407/CTA_0408, is a combination of three ordered locus names. Different validation methods produced different gene counts depending on whether or not the method separated the ordered locus names in the ID. After observing that only one of those ordered locus names appeared in the microarray data provided by Omsland *et al.* (2012), it was decided that the GenMAPP Builder code did not require any modifications specific to this ID. It is likely that only one but not all of the ordered locus names in the ID appear in any microarray platform.

As stated earlier, based on the results from Table 2, it would appear that rifampicin seems to have a considerable effect on gene expression. There are two speculations as to why this is: one speculation is that the larger number of genes up-regulated in the presence of rifampicin in comparison to the absence of rifampicin is a result of rifampicin's transcriptional inhibition of the differentiation from EB to RB; the other is that the smaller number of genes down-regulated in the presence of rifampicin in comparison to the absence of rifampicin could be the transcriptional inhibition of the differentiation of RB back to EB in the lifecycle of *C. trachomatis*.

In regards to the MAPPFinder results, significant GO terms were determined for increased and decreased gene expression using the data collected in the presence and absence of rifampicin and in both the presence and absence of rifampicin, there were more GO terms that

showed up for the decreased criterion (see Table 3,4,5, and 6). The MAPP for cellular carbohydrate synthesis, which was analyzed, resulted from the analysis of the data collected in the presence of rifampicin. This MAPP was chosen as a result of the concerns regarding the lack of a sufficient number of EB replicates in the data collected in the absence of rifampicin. In the MAPP for the cellular carbohydrate metabolic process (Fig. 1), the genes involved in the glycogen biosynthesis pathway were up-regulated. As glucose-6-phosphate is a key intermediate in this pathway (Ahern, 2009), the increase in gene expression corroborates the conclusion presented by Omsland *et al*. (2012) that EBs preferentially use glucose-6-phosphate. This refutes the notion that EB's are metabolically inactive. Future improvement for this experiment may include using a P Value correction in order to mitigate the false positive rate that may be the result of the lack of replicates performed on EB in the absence of rifampicin. Future analysis for improving the results of the experiment may also include analyzing significant MAPP's for the down-regulated genes.

References

Ahern, K. (2009) Glycogen Metabolism Notes < http://oregonstate.edu/instruction/bb450/summer09/lecture/glycogennotes.html>. Accessed 12 December 2013.

LMU BioDB 2013 (2013) How Do I Count Thee? Let Me Count The Ways <https://xmlpipedb.cs.lmu.edu/biodb/fall2013/index.php/How_Do_I_Count_Thee%3F_Let_Me_Count_The_Ways#XMLPipeDB_Match>. Accessed 12 December 2013.

Omsland, A., Sager, J., Nair, V., Sturdevant, D.E., Hackstadt, T. (2012) Developmental stage-specific metabolic and transcriptional activity of Chlamydia trachomatis in an axenic medium. PNAS 109: 19781-19785. doi: 10.1073/pnas.1212831109.

Acknowledgments

The authors would like to thank Dr. Kam D. Dahlquist for her assistance and guidance in analyzing the data using statistical tests as well as GenMAPP and MAPPFinder. In addition, the authors would like to thank Dr. John David N. Dionisio for his assistance and guidance throughout the process of modifying the GenMAPP Builder code to accommodate the requirements to export a gene database for *C. trachomatis* A/HAR-13.Lastly, the authors would like to acknowledge Dr. Daniel E. Sturdevant for his communications with the authors regarding the microarray data used in the analysis.

Supplementary Data

Testing Report

Version of GenMAPP Builder: 2.0b71

Computer on which export was run: Personal computer

Postgres Database name: CT_KS_20131119_32bit_gmb2b71

UniProt XML filename: Uniprot_XML_C.trachomatis_serovar_A_KS_20131114.xml

- UniProt XML version (The version information can be found at the UniProt News Page): UniProt release 2013_11
- Original file name from UniProt site : uniprot-organism%3A315277+keyword%3A181.xml
- Time taken to import: 1.20 min

GO OBO-XML filename: Go_daily-termdb_v2_HD_20131107.obo-xml

- GO OBO-XML version (The version information can be found in the file properties after the file downloaded from the GO Download page has been unzipped): 11/06/2013
    - Original file name as listed in beta.geneontology.org: go_daily-termdb.obo-xml.gz
- Time taken to import: 13.05 min
- Time taken to process: 10.80 min

GOA filename: 22183.C_trachomatis_A_KS_20131114.goa

- GOA version (News on this page records past releases; current information can be found in the Last modified field on the FTP site): 11/12/13
    - Original file name as listed in the FTP site: 22183.C_trachomatis_A.goa
- Time taken to import: 0.04 min
- Name of .gdb file: Ct-Std_v2_KS_20131121.gdb
- Time taken to export .gdb:
    - Start Time: 10:56 AM (approximately)
    - End Time: 11:14:26 AM
- Upload your file and link to it here: Ct-Std_External_20131121.gdb

Note: There is one ID "CTA_0406/CTA_0407/CTA_0408" that is a combination of three predicted genes that are actually one gene. The Tally Engine and PostgreSQL count this as one gene resulting in a total gene count of 917. However, through Access and xmlpipedb*match*, the ID is separated into three separate genes bringing the total count to 919.

Fig. S1. Testing report for the finalized gene database for *C. trachoamtis* A/HAR-13. The report notes that the ID CTA_0406/CTA_0407/CTA_0408 in the database is counted differently depending on the validation method that is used.

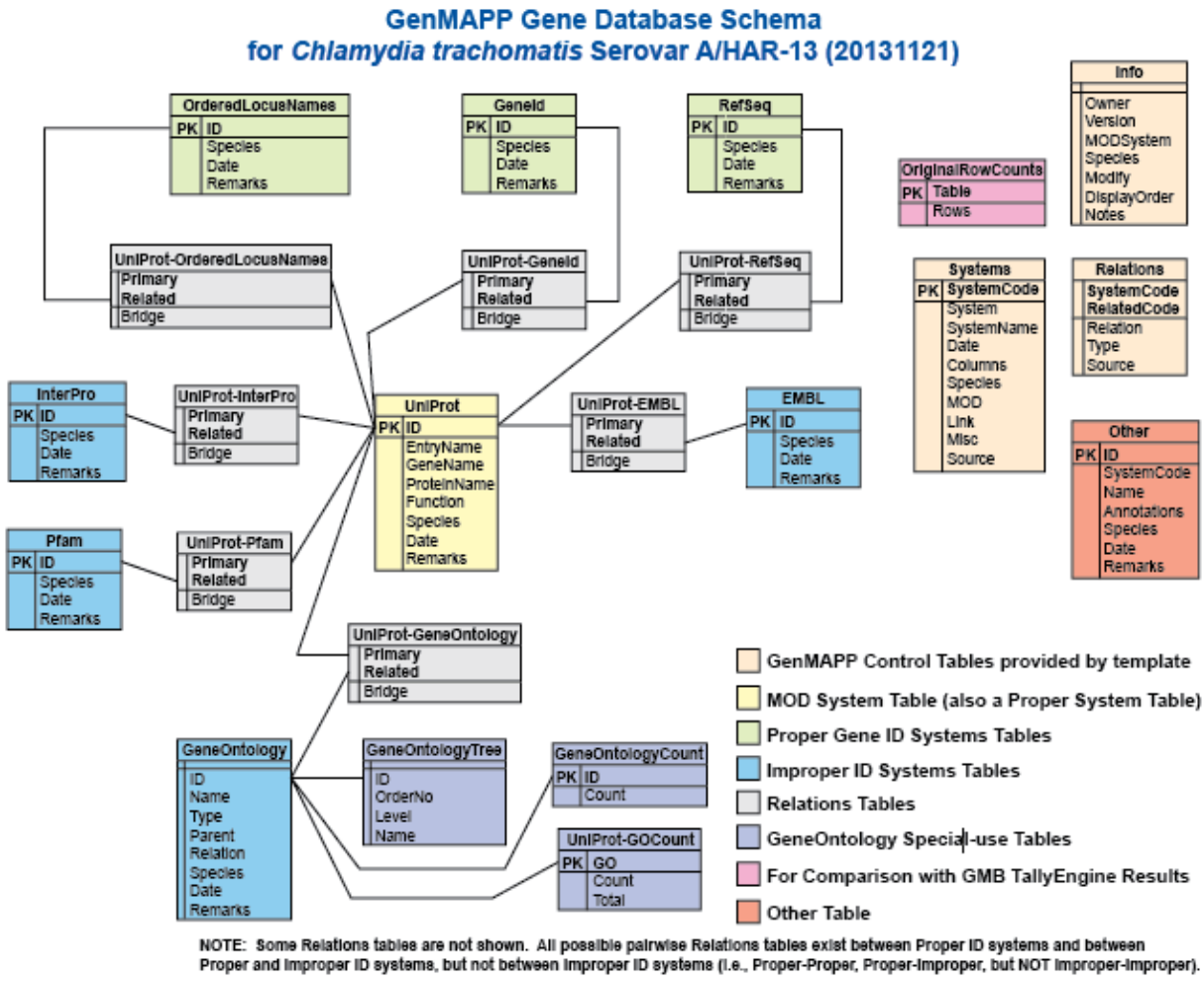

Fig. S2. The relational database schema for *C. trachomatis* A/HAR-13 depicts all of the tables within the relational database and how they are connected to each other with the exception of the tables OriginalRowCounts, Systems, Relations, Other, and Info.