# New GenMAPP Analysis of *Bordetella pertussis* Microarray Data Reveals Pathway Level Responses to Polysaccharide Capsule Deletion

Brandon J. Klein, Elena E. Olufson, & Mahrad R. Saeedi

Loyola Marymount University

BIOL/CMSI 367: Biological Databases

December 18, 2015

**Abstract**

Provide an abstract of no more than 500 words.

## Introduction

The gram negative bacterium *Bordetella pertussis* is a strictly human pathogen and the causative agent of whooping cough or pertussis, an acute respiratory disease that is particularly prominent in children. World Health Organization statistics from 2014 report approximately 140,000 cases of pertussis worldwide that caused nearly 90,000 deaths. Expansion of global vaccination programs over the past 60 years have reduced the incidence and mortality rates of whooping cough. However, an increasing number of adult pertussis infections has been recently reported, suggesting that new pertussis vaccination candidates and strategies may be necessary.

Comparative genome sequencing has demonstrated that *B. pertussis* acquired its host specificity and unique virulence factors, such as the pertussis toxin, through co-evolution with the expansion of human populations. Expression of *B. pertussis* virulence factors is regulated by the BvGA/S two-component system in response to environmental stimuli. BvGA/S activation is characterized by autophosphorylation of the inner membrane sensor histidine kinase BvgS, which triggers a complex His-Asp-His-Asp phosphorylation cascade that relays this signal to BvGS, a cytoplasmic transcriptional activator. The gene regulatory response controlled by BvGA/S activation results in up-regulation of virulence-activated genes (vags) and down-regulation of virulence-repressed genes (vrgs). The resulting transcriptional profile is referred to as the virulent or Bvg+ phase, which is distinct from the avirulent Bvg- phase. *In vitro*, this system is activated at 37℃ and silenced below 27℃ or by adding sulfate ions or nicotinic acid to the growth medium regardless of temperature. A third distinct phase, referred to

as Bvg-intermediate (Bvgi), exists in which vrgs are not expressed an only some vags are expressed, but the mechanisms controlling its expression are poorly understood.

Recent reports have identified that *B. pertussis* produces a polysaccharide (PS) microcapsule that does not function as a classical virulence factor. The expression of the PS capsule is virulence-repressed, occurring at its highest rate during the avirulent Bvg- phase. However, a recent microarray experiment concluded that deletion of a the transport protein *kpst* involved in export of PS polymers involved in capsule development inhibited the virulence of *B. pertussis*. This unexpected finding suggests that the transcriptional response to the BvGA/S regulatory-system is not fully understood and warrants further investigation.

We wanted to further elucidate the pathway-level transcriptional response to *kpst* deletion using the program GenMAPP (Gene Map Annotator and Pathway Profiler). GenMAPP is a free application for visualizing and analyzing microarray data, which can be used with the accessory program MAPPFinder to identify global trends in gene expression data. However, GenMAPP analyses depend on the existence of a relational gene database connecting known gene IDs and Gene Ontology (GO) terms for the study organism. Prior to our study, no gene database had been created for *Bordetella pertussis*.

The XMLPipeDB project provides a framework through which information from UniProt and the Gene Ontology Consortium can be integrated into species-specific relational databases for use in GenMAPP. This project centers on development of the program GenMAPP Builder, which largely automates constructs of the gene databases based on information from the imported files. By editing GenMAPP Builder code, custom species profiles can be created to reliably export complete gene databases for use in GenMAPP analysis.

In this study, we conducted a GenMAPP analysis of microarray data involving the deletion of *kpst* and prevention of *B. pertussis* capsule development to characterize pathway-level transcriptional changes that occurred. This analysis was enabled by the customization of GenMAPP Builder code to accommodate a new *B. pertussis* custom class and the subsequent creation of a new GenMAPP gene database for *B. pertussis.* The results of our analysis yielded new information regarding transcriptional profiles controlled by the BvgA/S regulatory system and its mechanism.

## Materials & Methods

### Downloading Files

The UniPRot XML proteome set and GOA (GO association) files for the bacterial strain *B. pertussis* Tohama I were downloaded and compressed on December 10, 2015. Then on the same day, we proceeded in obtaining the GO OBO-XML format and downloading the list of GO terms from this file. Finally, we downloaded the custom version of GenMAPP Builder including the most recent version of the *Bordetella pertussis* custom class (Version 3.0.0 Build 5). All files, including the GenMAPP builder folder, were extracted using 7-zip.

### Creating the GenMAPP Builder Tables in PostgreSQL

Once the necessary files had been downloaded and extracted, *pgAdmin III* was launched and connected to the PostgreSQL 9.4 server (localhost:5432). On this server, we created a new database: *bpertussis_cw20151210_gmb3build5*. We opened the SQL Editor tab to use an XMLPipeDB query to create the tables in the database. In the SQL Editor, we clicked on the

Open File icon and selected the file *gmbuilder.sql*. The query was executed in order to import a series of SQL commands into the editor tab.

**Loading Files into PostgreSQL Database Using GenMAPP Builder**

We launched gmbuilder.bat. and selected the "Configure Database" option to create and customize a new database called bpertussis_cw20151210_gmb3build

**Importing Data into the PostgreSQL Database**

The following downloaded data files: UniProt XML, GO OBO-XML, and the GOA for *Bordetella pertussis,* were specified and imported into the database individually.

**Exporting into a GenMAPP Gene Database**

The imported data in PostegreSQL was then exported to the GenMAPP Gene Database. The custom profile "Bordetella pertussis, Taxon ID 257313" was chosen as the gene database species. Once the database had been saved as *bpertussis-std_cw20151210*, we were able to export it to include all Molecular Function, Cellular Component, and Biological Process Gene Ontology Terms.

**Inspecting the Gene Database**

Once the database had been created and configured, several tests were run in order to confirm that all 3447 gene encoding IDs had been properly exported. First, TallyEngine was used to compare the number of Gene IDs in the XML file to the original amount in the database. To validate these results, the same was done using XMLPipeDB Match. Then, several SQL queries were run to, again, ensure that all gene IDs in the PostgreSQL relational database were accounted for. Finally, the database was opened using Microsoft Access for a visual inspection to see if the expected tables were listed with the expected number of records.

**Preparing Microarray Data for GenMAPP Analysis**

The microarray paper to be used was selected and the six individual data samples were downloaded and compiled into one excel spreadsheet. Statistical analysis of the data was performed in Excel. The gene expression data was then formatted and prepared for import into GenMAPP.

**Running GenMAPP Using the Gene Database**

Within GenMAPP, the *Bordetella pertussis* gene database was loaded. A sample MAPP was made, in which gene IDs conforming to the naming conventions of the 5 major gene databases containing the *Bordetella pertussis* genome data were added.

**Importing the Microarray Data into GenMAPP**

An Expression Dataset was created and all gene IDs from the microarray data were imported into it. Errors in the expression file were investigated and reported before continuing in creating a MAPP. The Expression Data was customized using specific color sets to label genes and visually differentiate amongst them, whether an individual gene's expression had increased or decreased. Two pathway-based MAPPs were created, one representing the Ribosome Kegg pathway and one for the Nitrogen Cycle Kegg pathway.

**Running the MAPPFinder Analysis**

The MAPPFinder program was launched from within GenMAPP and we ensured that the *bpertussis-std_cw20151210.gdb* gene database was still loaded into GenMAPP. We applied the specified criteria to represent "Increased" or "Decreased" and matched each set of data with a specific color set. MAPPFinder was then run and the significant Gene Ontology terms were analyzed.

**Results**

The gene database schema which was created, depicted the different aspects involved with the research and demonstrated the cohesion of these various parts. Customizations were made specifically for the OrderedLocusNames and Systems tables present in the schema in order to ensure retrieval of a the full list of gene IDs for *Bordetella Pertussis* from the Uniprot XML file.

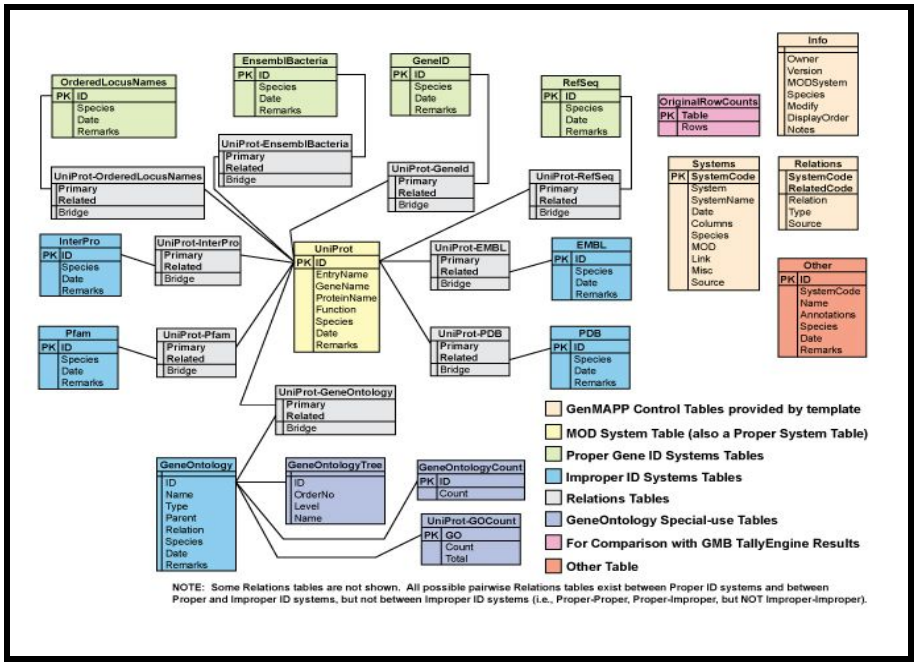



Figure 1 depicts the GenMAPP Gene Database Schema for *Bordetella pertussis* Tohama I.

Our gene model database demonstrated that there was a total of 3447 gene IDs. Once our complete set of protein coding genes were quantified, we confirmed the presence of all the gene IDs using the respective counting systems available. Through this method, along with visual inspection, it was determined that there existed 3 specific categories of gene IDs, with independant identification patterns for each category of genes. This provided confirmation that 3447 gene IDs were accounted for as depicted in the following table.

| | OrderedLocus Names [BP####\|.1] | Open Reading Frame [BP####A\|B] | EnsemblBacteria Reference ID [BP3167A] | Totals |
|---|---|---|---|---|
| **XMLPipeDB Match** | 3435 | 11 | 1 | **3447** |
| **TallyEngine-XML** | 3435 | 11 | 0 | **3446** |
| **TallyEngine-PostgreSQL** | 3435 | 11 | 0 | **3446** |
| **OriginalRowCounts [.gdb]** | 3435 | 11 | 1 | **3447** |
| **GeneDB MOD** | - | - | - | **3447** |

Table 1 lists specific gene ID counts obtained for each ID pattern (ordered locus, open reading frame, EnsemblBacteria) from independent systems and databases.

The command that was used in XMLpipeDB Match to generate results was:

```
COMMAND java -jar xmlpipedb-match-1.1.1.jar "BP[0-9][0-9][0-9][0-9](A|B|\.1|)" <"uniprot-proteome%3AUP000002676_cw20151201.xml"
OUTPUT
Total unique matches: 3447
```

The query used in PGAdmin III to generate results came out to be:

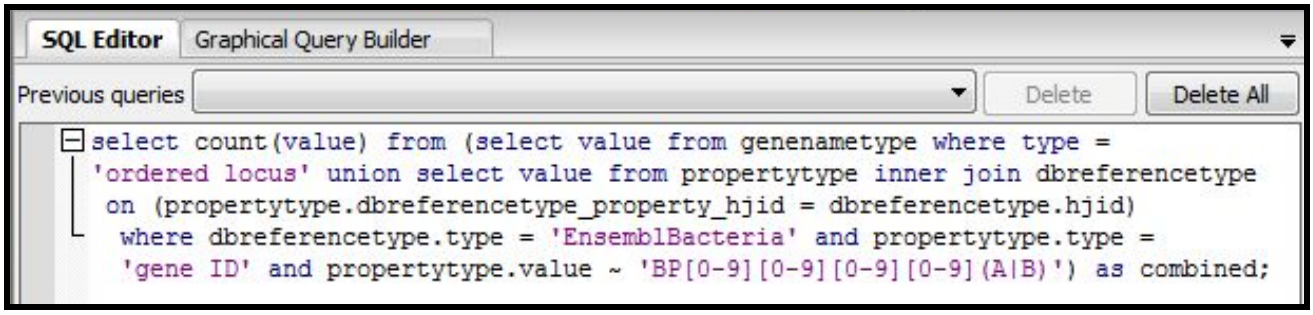
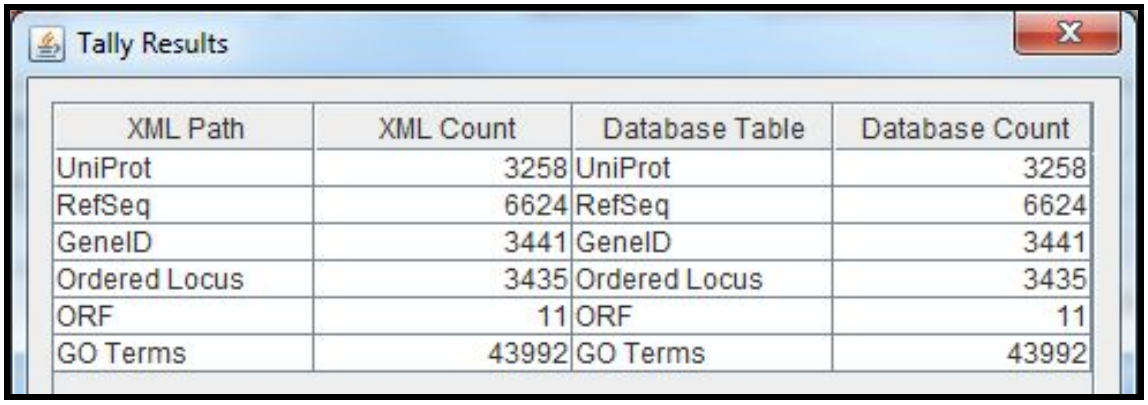



Figure 2 shows how the PGAdmin III was used to count the number of gene IDs in the relational database (20151210).




| XML Path | XML Count | Database Table | Database Count |
|---|---|---|---|
| UniProt | 3258 | UniProt | 3258 |
| RefSeq | 6624 | RefSeq | 6624 |
| GeneID | 3441 | GeneID | 3441 |
| Ordered Locus | 3435 | Ordered Locus | 3435 |
| ORF | 11 | ORF | 11 |
| GO Terms | 43992 | GO Terms | 43992 |

Figure 3 shows the results for the GenMAPP Builder TallyEngine as it counted gene IDs in the XML and PostgreSQL Database.

Gene Database

Initially, in the first version of the *B. pertussis* Gene Database, there were 12

OrderedLocusNames IDs that were present in the original XML file, but were missing when

imported into PostgresSQL.

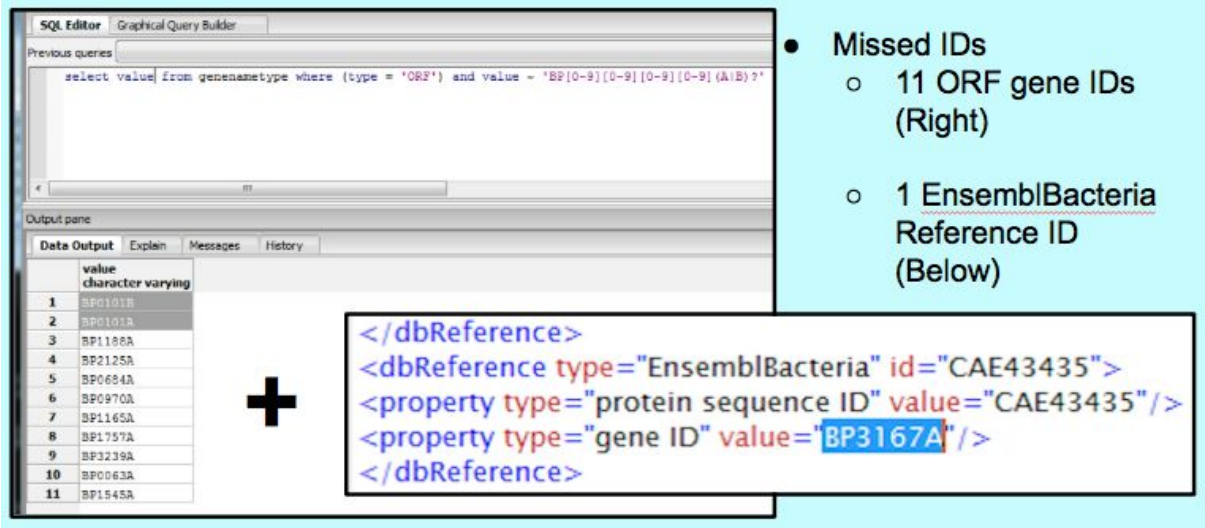



Figure 4 presents the unique patterns amongst missing IDs that allowed for their selective retrieval using an SQL Query.

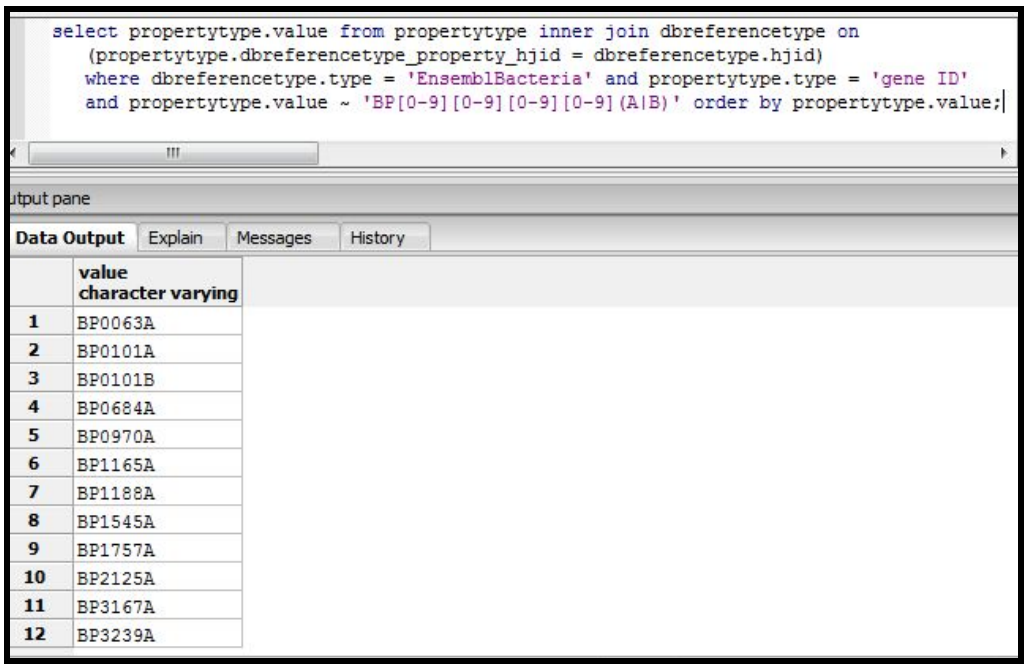



Figure 5 depicts the new method block to the *B. pertussis* custom profile code in order to successfully import the missing IDs into PostgreSQL.

```
+        // Start with the default OrderedLocusNames behavior.
+        TableManager result = super.getSystemTableManagerCustomizations(tableManager, primarySystemTableManager,
+                version);
+
+        String sqlQuery = "select dbreferencetype.entrytype_dbreference_hjid as hjid, propertytype.value from propertytype inner join dbrefer
+                "(propertytype.dbreferencetype_property_hjid = dbreferencetype.hjid) " +
+                "where dbreferencetype.type = 'EnsemblBacteria' and propertytype.type = 'gene ID' " +
+                "and propertytype.value ~ 'BP[0-9][0-9][0-9][0-9](A|B)' order by propertytype.value";
+
+        Connection c = ConnectionManager.getRelationalDBConnection();
+        PreparedStatement ps;
+        ResultSet rs;
+        try {
+            // Query, iterate, add to table manager.
+            ps = c.prepareStatement(sqlQuery);
+            rs = ps.executeQuery();
+            while (rs.next()) {
+                String hjid = Long.valueOf(rs.getLong("hjid")).toString();
+                String id = rs.getString("value");
+                result.submit("OrderedLocusNames", QueryType.insert, new Object[][] {
+                    { "ID", id },
+                    { "Species", "|" + getSpeciesName() + "|" },
+                    { "Date", version },
+                    { "UID", hjid }
+                });
+            }
+        } catch(SQLException sqlexc) {
+            logSQLException(sqlexc, sqlQuery);
+        }
+
+        return result;
```

Figure 6 displays the new method block that was added to the *B. pertussis* custom profile code to import the missing IDs.

In

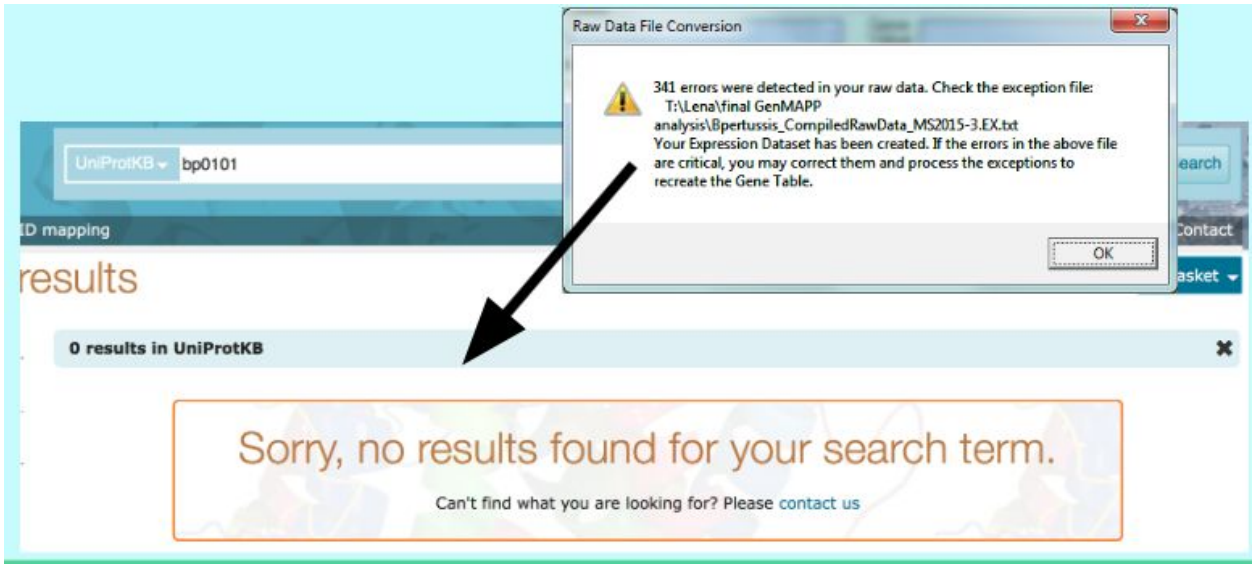



Figure 7 depicts the lack of presence of the 341 ID errors in the UniProt XML file.

* Report on what changes were made to the GenMAPP Builder code in order to to accommodate the second and third type of missing gene IDs and the result of those changes

| | P < 0.05 | P < 0.01 | P < 0.001 | P < 0.0001 | Bonferroni [P < 0.05] | Benjamini & Hochberg [P < 0.05] |
|---|---|---|---|---|---|---|
| **Number of Genes** | 1923/3552 | 1028/3552 | 242/3552 | 40/3552 | 9/3552 | 1365/3552 |
| **Percent of Total Genes** | 54% | 29% | 7% | 1% | 0.20% | 38% |

Table 2 lists the sanity check results performed on the data analysis for specific P-value filters.

The criteria used to demonstrate a significant increase or decrease in expression for the GenMAPP Expression Dataset was as follows:
Increase: [Avg_ABC_Samples] > 0.25 AND [Pvalue] < 0.05 RED
Decrease: [Avg_ABC_Samples] < -0.25 AND [Pvalue] < 0.05 GREEN

| GOID | GO Name | Number Changed | Number in GO | Percent Changed | PermuteP | AdjustedP |
|---|---|---|---|---|---|---|
| 50801 | ion homeostasis | 5 | 1 | 71.42857 | 0 | 0.652 |
| 48878 | chemical homeostasis | 5 | 1 | 71.42857 | 0 | 0.652 |
| 98771 | inorganic ion homeostasis | 5 | 1 | 71.42857 | 0 | 0.652 |
| 70271 | protein complex biogenesis | 5 | 4 | 62.5 | 0.003 | 0.848 |
| 6461 | protein complex assembly | 5 | 4 | 62.5 | 0.003 | 0.848 |
| 6777 | Mo-molybdopterin cofactor biosynthetic process | 5 | 8 | 62.5 | 0.005 | 0.848 |
| 19674 | NAD metabolic process | 5 | 8 | 62.5 | 0.005 | 0.848 |
| 19720 | Mo-molybdopterin cofactor metabolic process | 5 | 8 | 62.5 | 0.005 | 0.848 |
| 19363 | pyridine nucleotide biosynthetic process | 5 | 7 | 62.5 | 0.005 | 0.848 |
| 19359 | nicotinamide nucleotide biosynthetic process | 5 | 7 | 62.5 | 0.005 | 0.848 |
| 51082 | unfolded protein binding | 6 | 12 | 50 | 0.007 | 1 |
| 72525 | pyridine-containing compound biosynthetic process | 7 | 7 | 43.75 | 0.016 | 1 |
| 16782 | transferase activity, transferring sulfur-containing groups | 5 | 4 | 50 | 0.017 | 1 |
| 43545 | molybdopterin cofactor metabolic process | 5 | 8 | 55.55556 | 0.019 | 1 |
| 51189 | prosthetic group metabolic process | 5 | 8 | 55.55556 | 0.019 | 1 |
| 32324 | molybdopterin cofactor biosynthetic process | 5 | 9 | 55.55556 | 0.019 | 1 |
| 42537 | benzene-containing compound metabolic process | 5 | 2 | 50 | 0.022 | 1 |
| 65003 | macromolecular complex assembly | 5 | 4 | 45.45454 | 0.03 | 1 |
| 6979 | response to oxidative stress | 5 | 12 | 41.66667 | 0.046 | 1 |

Table 3 lists the top 24 GO terms generated from the “Increased” criteria applied to the data.

| GOID | GO Name | Number Changed | Number in GO | Percent Changed | PermuteP | AdjustedP |
|---|---|---|---|---|---|---|
| 33866 | nucleoside bisphosphate biosynthetic process | 5 | 6 | 83.33334 | 0.002 | 0.242 |
| 34033 | purine nucleoside bisphosphate biosynthetic process | 5 | 6 | 83.33334 | 0.002 | 0.242 |
| 34030 | ribonucleoside bisphosphate biosynthetic process | 5 | 6 | 83.33334 | 0.002 | 0.242 |
| 15936 | coenzyme A metabolic process | 5 | 6 | 83.33334 | 0.002 | 0.242 |
| 15937 | coenzyme A biosynthetic process | 5 | 6 | 83.33334 | 0.002 | 0.242 |
| 46933 | proton-transporting ATP synthase activity, rotational mechanism | 5 | 7 | 71.42857 | 0.005 | 0.714 |
| 42777 | plasma membrane ATP synthesis coupled proton transport | 5 | 7 | 71.42857 | 0.005 | 0.714 |
| 33865 | nucleoside bisphosphate metabolic process | 5 | 6 | 62.5 | 0.007 | 0.886 |
| 34032 | purine nucleoside bisphosphate metabolic process | 5 | 6 | 62.5 | 0.007 | 0.886 |
| 33875 | ribonucleoside bisphosphate metabolic process | 5 | 6 | 62.5 | 0.007 | 0.886 |
| 44769 | ATPase activity, coupled to transmembrane movement of ions, rotational mechanism | 5 | 4 | 62.5 | 0.008 | 0.886 |
| 3746 | translation elongation factor activity | 5 | 8 | 62.5 | 0.009 | 0.886 |
| 4003 | ATP-dependent DNA helicase activity | 5 | 8 | 62.5 | 0.011 | 0.886 |
| 4312 | fatty acid synthase activity | 5 | 2 | 55.55556 | 0.012 | 1 |
| 5507 | copper ion binding | 5 | 9 | 55.55556 | 0.017 | 1 |
| 6664 | glycolipid metabolic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 46467 | membrane lipid biosynthetic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 9247 | glycolipid biosynthetic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 1901269 | lipooligosaccharide metabolic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 46493 | lipid A metabolic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 9245 | lipid A biosynthetic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 1901271 | lipooligosaccharide biosynthetic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 6643 | membrane lipid metabolic process | 5 | 9 | 55.55556 | 0.018 | 1 |
| 5694 | chromosome | 5 | 10 | 50 | 0.034 | 1 |

Table 4 lists the top 19 GO terms generated from the "Decreased" criteria applied to the data.

Increased GO results: Many of the increased GO terms that were produced by GenMAPP can be seen to relate to the processes and construction of proteins in the cell. A couple of up-regulated GO terms that can be categorically related to this concept include: protein complex biogenesis, protein complex assembly, and unfolded protein binding. The protein complex biogenesis term includes the cellular process that results in the biosynthesis of constituent macromolecules, assembly, and arrangement of constituent parts of a protein complex; this can thus be tied into the GO term protein complex assembly as the biogenesis is the creation of the protein complex. Since the construction of proteins in the cell is being up-regulated, the cell is demanding the production and use of proteins at a rate higher than normal. Relating this concept to the microarray experiment on *B. pertussis*, the up-regulation of the protein assembly can be attributed to the cell trying to develop transport membrane proteins needed for the polysaccharide microcapsule made by the cell. In the experiment, the polysaccharide microcapsule is deleted, and so the up-regulation of the production of proteins would be expected as the cell is trying to compensate for its initial loss of proteins.

Decreased GO results: The decreased GO terms that were generated by GenMAPP included many variations of biosynthetic processes that occur in cells in order for the energy to be generated to support the correct functioning of the cell. A few of these down-regulated key GO terms that can be categorically related to one another include: coenzyme A biosynthetic process, coenzyme A metabolic process, and nucleoside biphosphate biosynthetic process. All of these terms relate to the processes that occur inside of a cell to act as the cells' main source of energy-generation. Specifically, the coenzyme A terms are important as coenzyme A acts as the major driving force for the cell's citric acid cycle, once the coenzyme A is transformed into acetyl-CoA to be used directly in the cycle. The citric acid cycle is the cell's main process by which it obtains it energy in ATP form, and thus when the processes that are the components of the citric acid cycle are decreased, this means that the cell is using a different method of

generating its energy. Since the citric acid cycle is not supplying the cell with its energy as it usually does, it can be said that oxidative phosphorylation has been significantly decreased and the cell is using another method or process to produce its energy.

Creating a GenMAPP MAPP of the ribosome biogenesis pathway depicted the apparent down regulation of these genes.

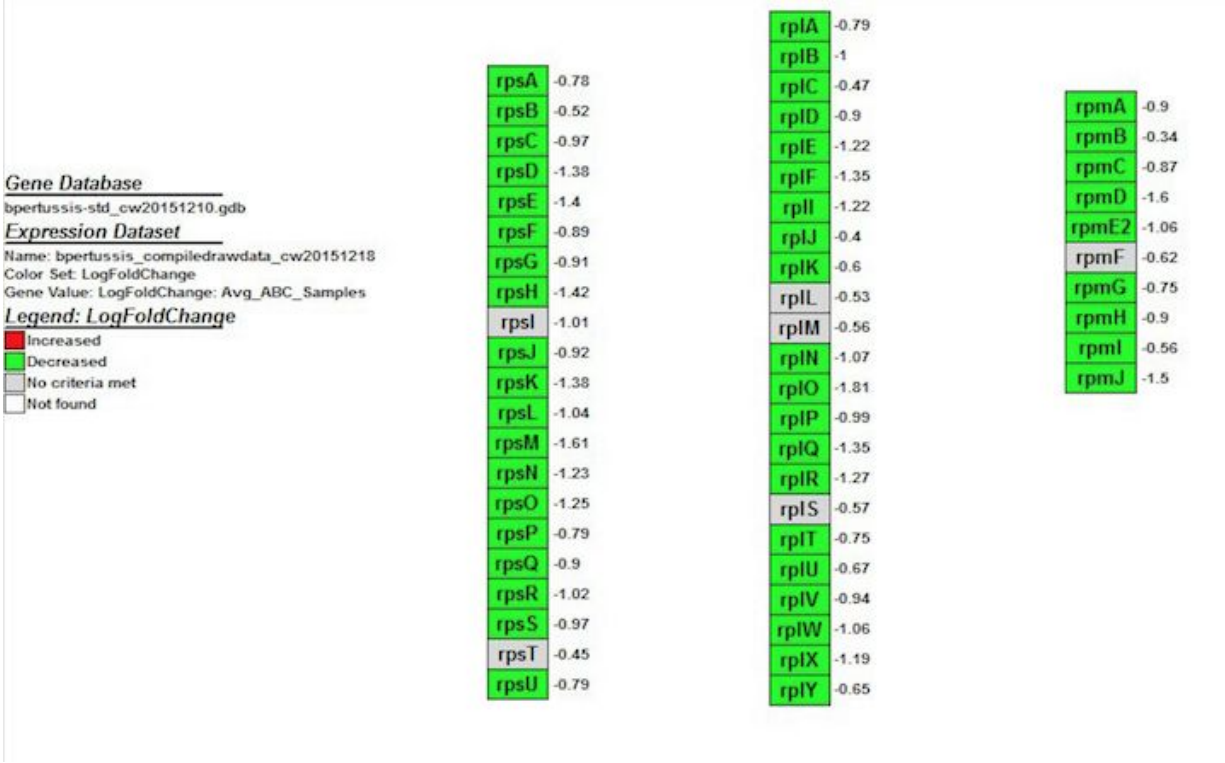



Figure 8 presents the MAPP created using the selected ribosome pathway genes applied to the expression dataset in GenMAPP.

* Gene Database Testing Report on final version of Gene Database (can be put at the end of the report as an Appendix)

**=== Discussion ===**

* How well did the GenMAPP Builder process work for your species (just comment on the technical aspects here, you will discuss the teamwork/process aspects in your individual assessment).

* Discuss the statistical analysis and MAPPFinder results for your microarray dataset. Compare it to what was reported in the original paper from which you got the microarray data.

** In particular, compare directly the log fold change value of a couple of key genes mentioned in the paper with what you found for those genes.

** Compare the criteria the journal article used for a significant expression change to the criteria that you used. How many genes met the criterion for the article vs. how many met the criterion for your analysis.

**=== Conclusions ===**

Write a concluding paragraph that summarizes the overall project and your findings.

* How closely do your findings correspond to the original study?

* Are there significant differences?

* Did you discover anything new?

* What future directions would you take if you were to continue this project?

**Acknowledgments**

We would like to acknowledge two specific people who greatly assisted us in the development and performance of our project, both of which are professors at Loyola Marymount University. These instructors are Dr. Kam D. Dahlquist from the Biology Department and Dr. John David N. Dionisio from the Computer Science Department. We would like to show our gratitude to them for sharing their pearls of wisdom with us during the course of this project and assisting us with any obstacles we faced. Their comments and feedback greatly improved the operation of our project and we thank them immensely.

## References

Hoo, R., Lam, J.H., Huot, L., Pant, A., Li, R., Hot, D., & Alonso, S. (2014). Evidence for a Role of the Polysaccharide Capsule Transport Proteins in Pertussis Pathogenesis. PLoS ONE, 9(12):e115243. doi: 10.1371/journal.pone.0115243

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., et al. (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nature genetics, 35(1), 32-40. doi:10.1038/ng1227