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

The pathogenic character of *Bordetella pertussis*, the causative agent of whooping cough, is controlled by the BvgA/S two-component regulatory system. A recent microarray experiment unexpectedly implicated the polysaccharide (PS) capsule synthesized by *B. pertussis* with the proper activation of the virulent Bvg+ phase. In this study, we used the programs GenMAPP and MAPPFinder to characterize global trends in expression data captured in this microarray experiment. Our analysis differed from those of the original study, revealing a lower amount of significant expression changes and more detailed pathway-level responses. In particular, we found consistent down-regulation of nearly all genes involved in the *B. pertussis* ribosome biogenesis pathway and transcriptional changes associated with a distinct metabolic pathway shift.

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.

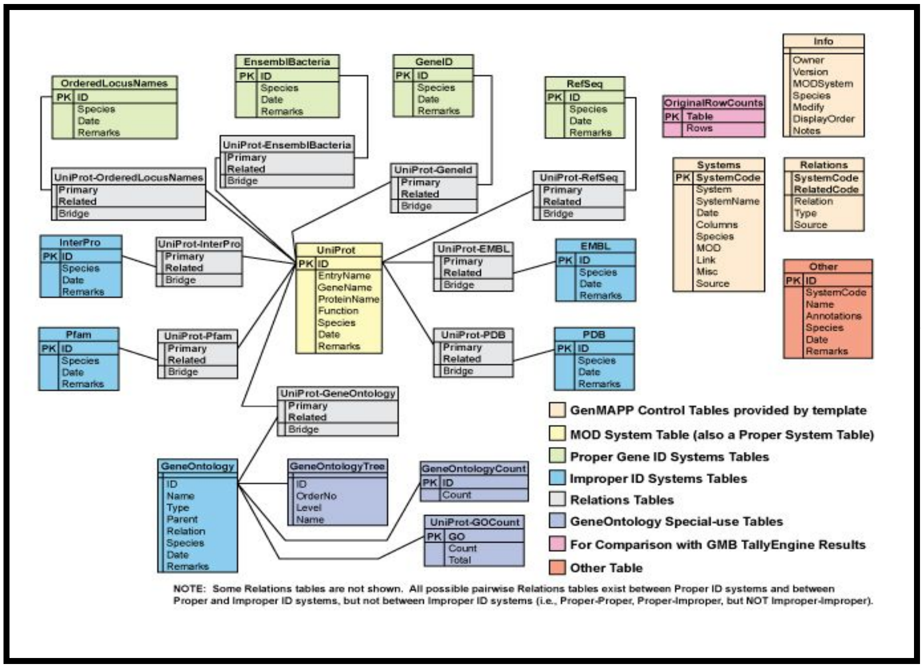
Two pathway-based MAPPs were created, one representing the Ribosome Kegg pathway and one for the Nitrogen Cycle Kegg pathway.

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 (Figure 1). 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.



NOTE: Some Relations tables are not shown. All possible pairwise Relations tables exist between Proper ID systems and between Proper and Improper ID systems, but not between Improper ID systems (i.e., Proper-Proper, Proper-Improper, but NOT Improper-Improper).

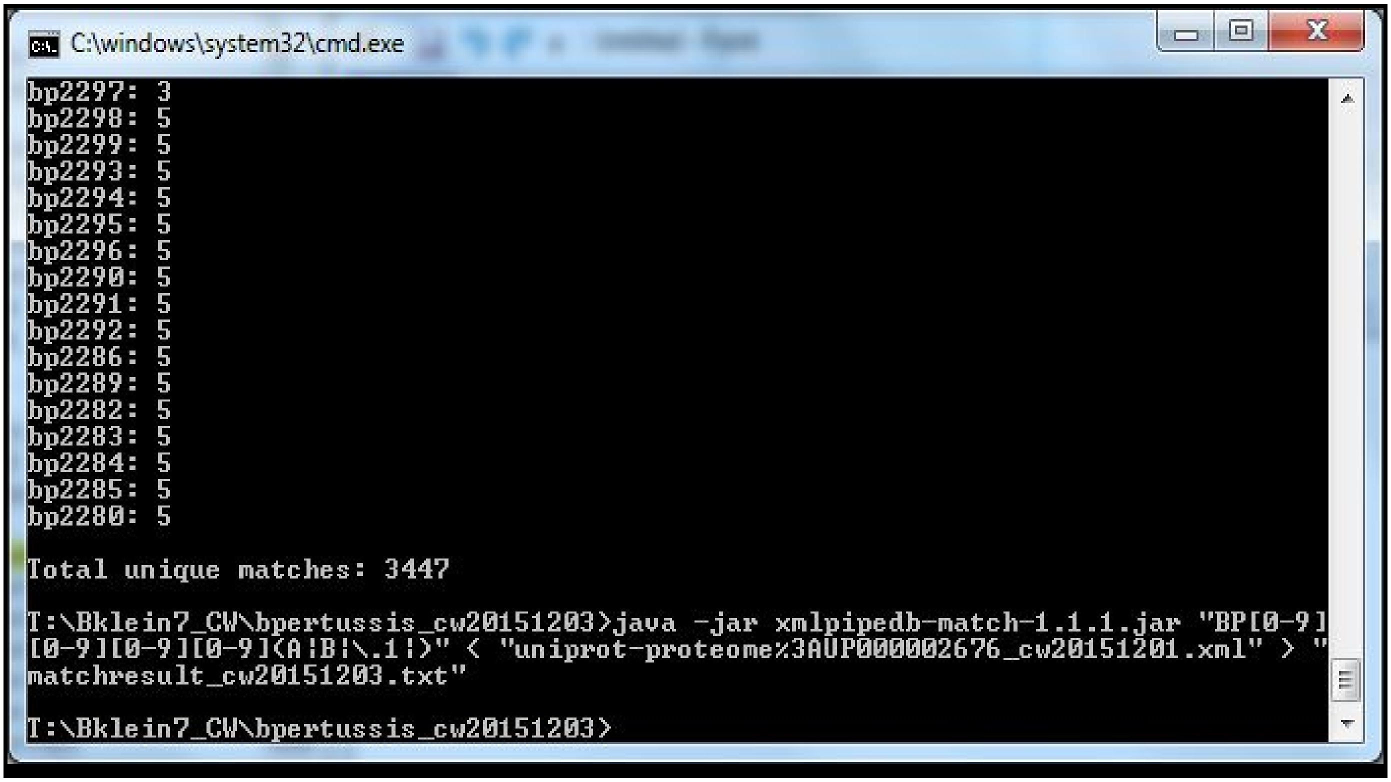
Figure 1 depicts the GenMAPP Gene Database Schema for *Bordetella pertussis* Tohama I.

The GeneDB Model Organism Database (MOD) for *Bordetella pertussis* demonstrated that this organism's genome contains a total of 3447 protein-coding genes. Upon exporting the final version of our *B. pertussis* gene database, we confirmed the presence of all the gene IDs using a series of counting systems. 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 (Table 1). 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]**	**Ensembl Bacteria 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.

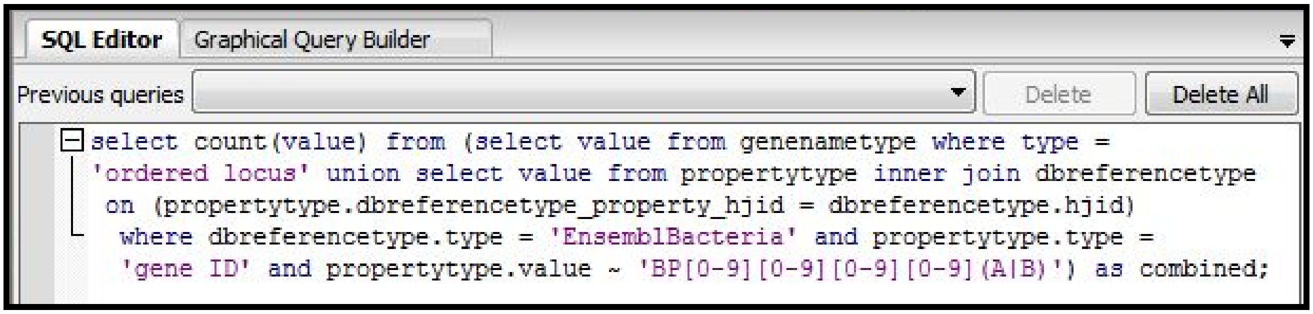
A command was created and inputted into XMLPipeDB Match in order to count the number of gene IDs in the original XML file. Through the process of developing the command, a regex was crafted to capture all identified gene ID patterns: BP#### | BP####.1 | BP####A | BP####B. Putting all of the pieces together resulted in the final command used (Figure 1).



```
bp2297: 3
bp2298: 5
bp2299: 5
bp2293: 5
bp2294: 5
bp2295: 5
bp2296: 5
bp2290: 5
bp2291: 5
bp2292: 5
bp2286: 5
bp2289: 5
bp2282: 5
bp2283: 5
bp2284: 5
bp2285: 5
bp2280: 5

Total unique matches: 3447

T:\Bklein7_CW\bpertussis_cw20151203>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" > "
matchresult_cw20151203.txt"

T:\Bklein7_CW\bpertussis_cw20151203>
```

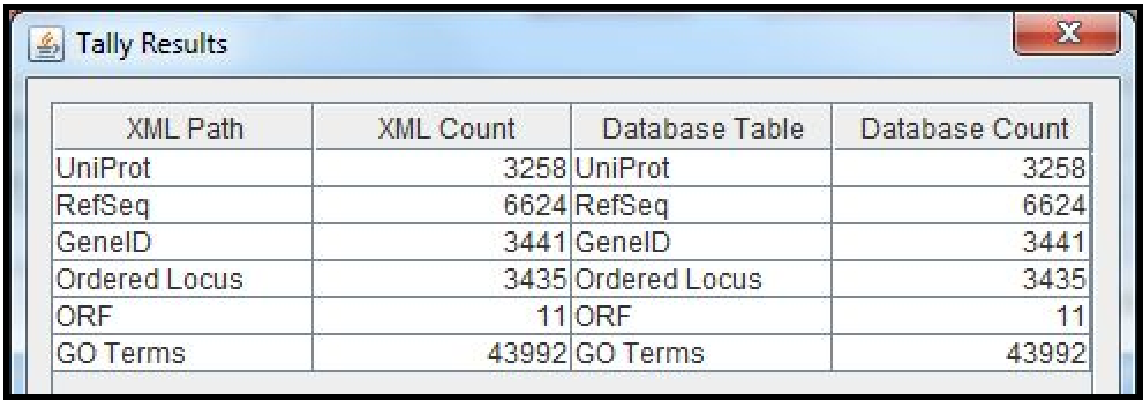
Figure 1 shows the exact code inputted into XMLPipeDB Match, followed by the resulting count.

The query used in PGAdmin III was coded to include all of the gene IDs falling under these three categories: *OrderedLocusNames, ORF, and EnsemblBacteria.* (Figure 2). Furthermore, the count from this query was re-confirmed using TallyEngine. (Figure 3).



```sql
select count(value) from (select value from genenametype where type =
'ordered locus' union select value from propertytype inner join dbreferencetype
 on (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)') as combined;
```

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

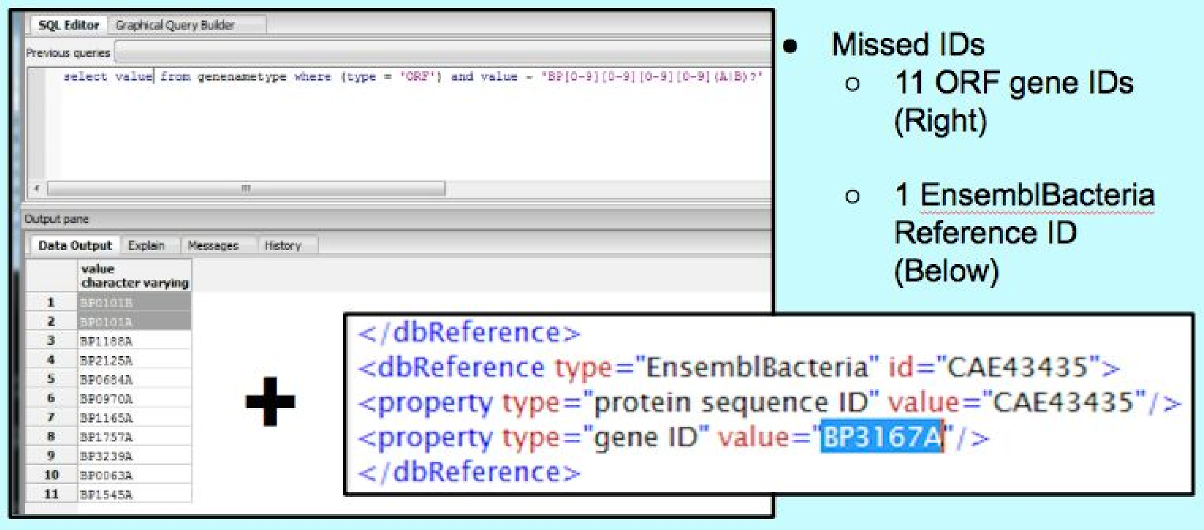


XML Path	XML Count	Database Table	Database Count
UniProt	3258	UniProt	3258
RefSeq	6624	RefSeq	6624
GeneID	3441	GeneID	3441
Ordered Locus	3435	Ordered Locus	3435
ORF	11	ORF	11
GO Terms	43992	GO Terms	43992

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). Necessary changes were made to the initial GenMAPP Builder code in order to successfully import the missing gene IDs. Specifically, new method blocks were added to the code to import *ORF* listings (figure 5). Using TallyEngine, PostgreSQL, and by inspecting the *OriginalRowCounts* Table in the database, we confirmed a total of 3447 gene IDs. This change worked to incorporate the *ORF* gene IDs, that had previously been missing.



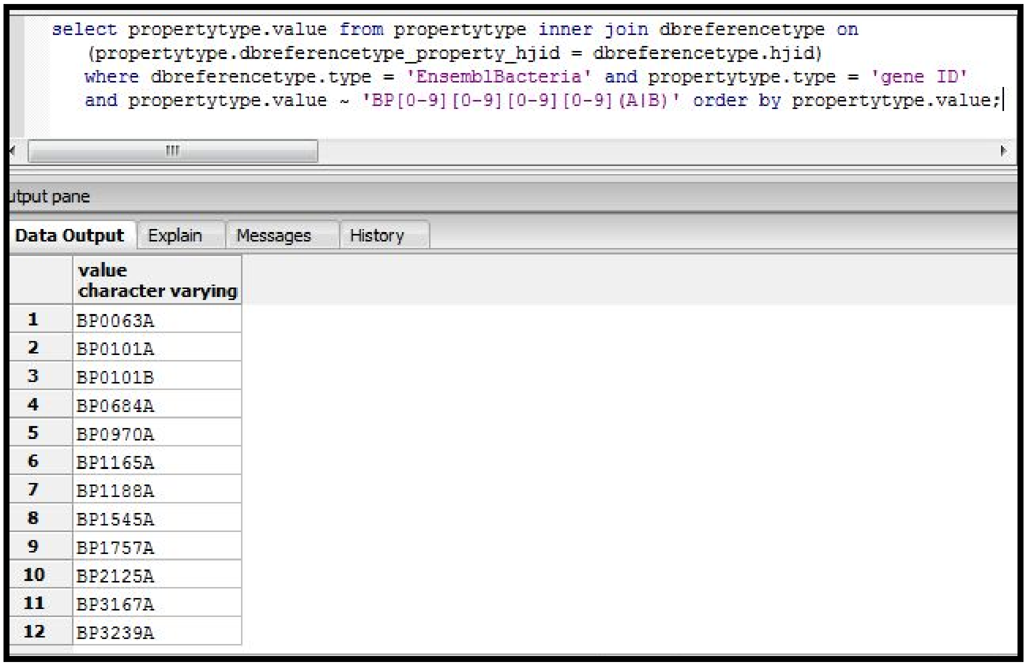
```
select value from genenametype where (type = 'ORF') and value ~ 'BP[0-9][0-9][0-9][0-9](A|B)?'
```

	value character varying
1	BP0101B
2	BP0101A
3	BP1188A
4	BP2125A
5	BP0684A
6	BP0970A
7	BP1165A
8	BP1757A
9	BP3239A
10	BP0063A
11	BP1545A

- Missed IDs
  - 11 ORF gene IDs (Right)
  - 1 EnsemblBacteria Reference ID (Below)

```
</dbReference>
<dbReference type="EnsemblBacteria" id="CAE43435">
<property type="protein sequence ID" value="CAE43435"/>
<property type="gene ID" value="BP3167A"/>
</dbReference>
```

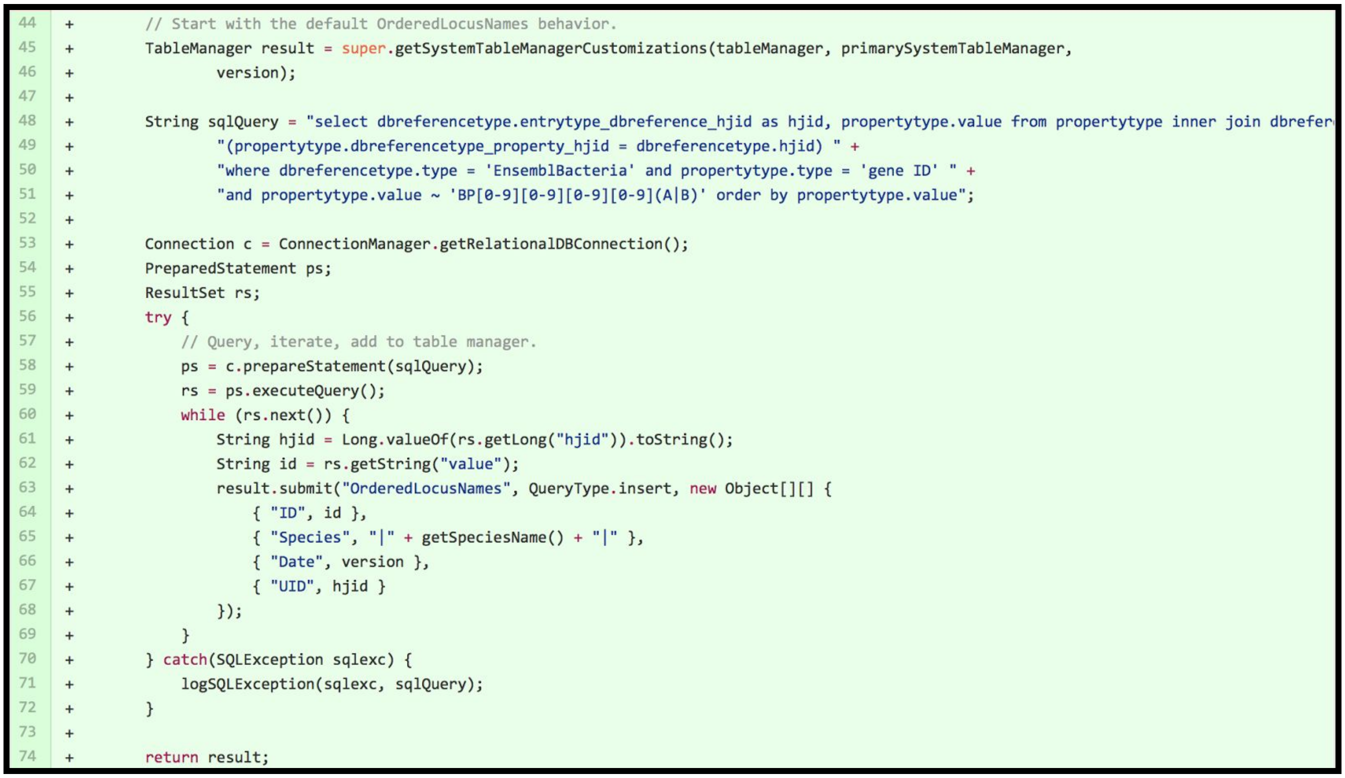
Figure 4 presents the unique patterns amongst missing IDs that allowed for their selective retrieval using an SQL Query.



```
select propertytype.value from propertytype inner join dbreferencetype on
   (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;
```

	value character varying
1	BP0063A
2	BP0101A
3	BP0101B
4	BP0684A
5	BP0970A
6	BP1165A
7	BP1188A
8	BP1545A
9	BP1757A
10	BP2125A
11	BP3167A
12	BP3239A

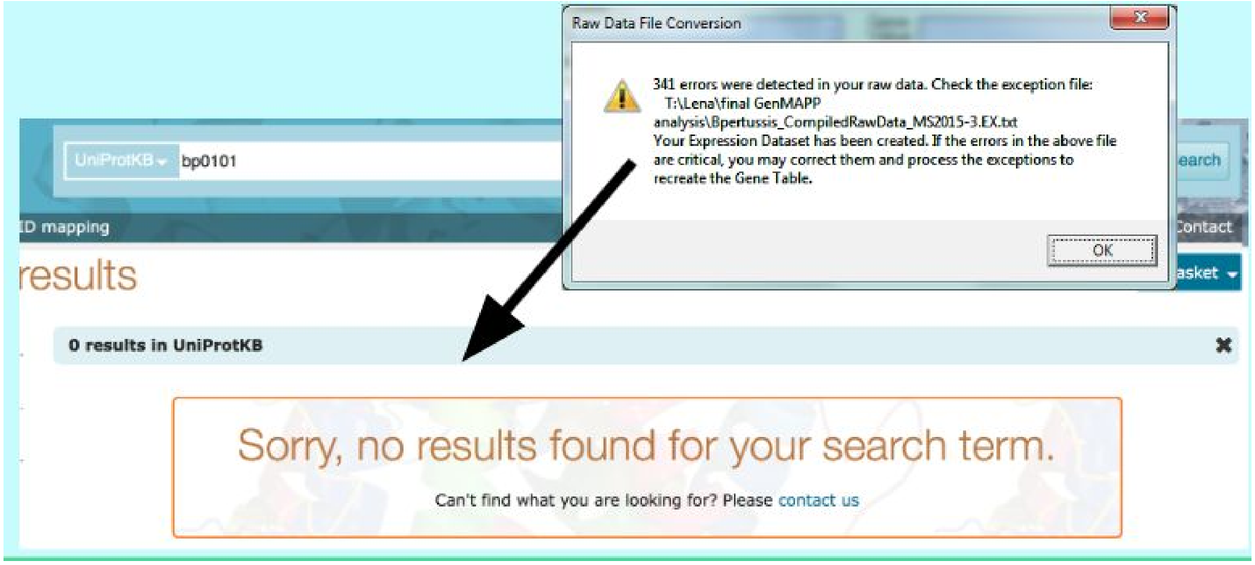
Figure 5 depicts the new method block to the *B. pertussis* custom profile code in order to successfully import the missing IDs into PostgreSQL.



```
+       // 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.

Upon importing the microarray data into the new gene database, 341 errors were detected in the data. Through further investigation, and in examining the Uniprot XML file, it was made clear that these 341 errors referred to gene IDs that were in the microarray data, but not in the XML file (Figure 7). Because none of these IDs were present in the XML file, this demonstrates that the 341 errors arise from out of date or non-protein coding genes in the microarray data.



341 errors were detected in your raw data. Check the exception file:
T:\Lena\final GenMAPP analysis\Bpertussis_CompiledRawData_MS2015-3.EX.txt
Your Expression Dataset has been created. If the errors in the above file are critical, you may correct them and process the exceptions to recreate the Gene Table.

0 results in UniProtKB

Sorry, no results found for your search term.

Can't find what you are looking for? Please contact us

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

After the statistical tests and GenMAPP preparation actions were performed on the microarray data, a sanity check was run before moving onto the GenMAPP/MAPPFinder analysis. A sanity check was done in order to make sure that the data analysis was performed correctly (Table 2). By applying filters to the data, the number of genes that were significantly changed at the various p value cut-offs were calculated. The Bonferonni P value was the most stringent restriction placed on the genes and thus yielded nine specific genes that were deemed as most significant.

	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%

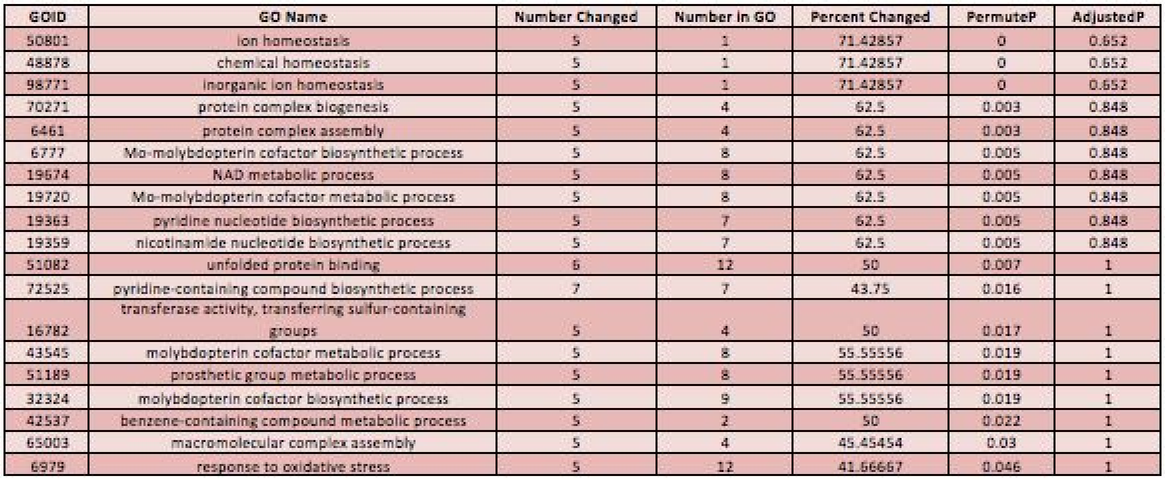
Table 2 lists the sanity check results performed on the data analysis for specific P-value filters.

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

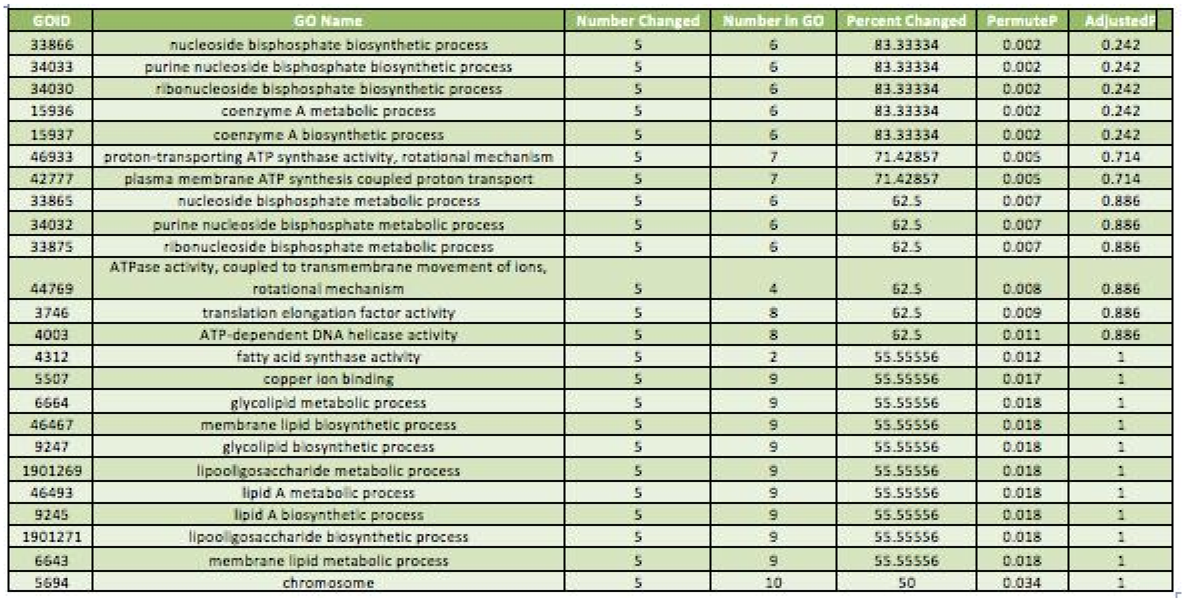
Decrease: [Avg_ABC_Samples] < -0.25 AND [Pvalue] < 0.05 GREEN

After applying the above criteria, MAPPFinder was run in order to generate the significant results according to the specific restrictions set for the values. After these results were produced, both of the criterion GO files were filtered to display the top 20 gene ontology terms. The top increased GO terms were placed in a table (Table 3), and the top decreased GO terms were placed in a table (Table 4) that included the respective P values and percent changed for each GO term. An analysis of a few of the key GO terms for the increased and decreased criteria is presented below.



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.

The ribosome pathway generated in GenMAPP (Figure 8) depicts the down-regulation of the selected genes from the ribosome Kegg pathway. The MAPP displays the small (rps) and large (rpl) ribosomal subunit genes in their respective columns organized alphabetically A-Z. The down-regulation of the genes is represented by the color green, thus displaying the specific genes that were decreased in accordance with the criteria set in the expression dataset.

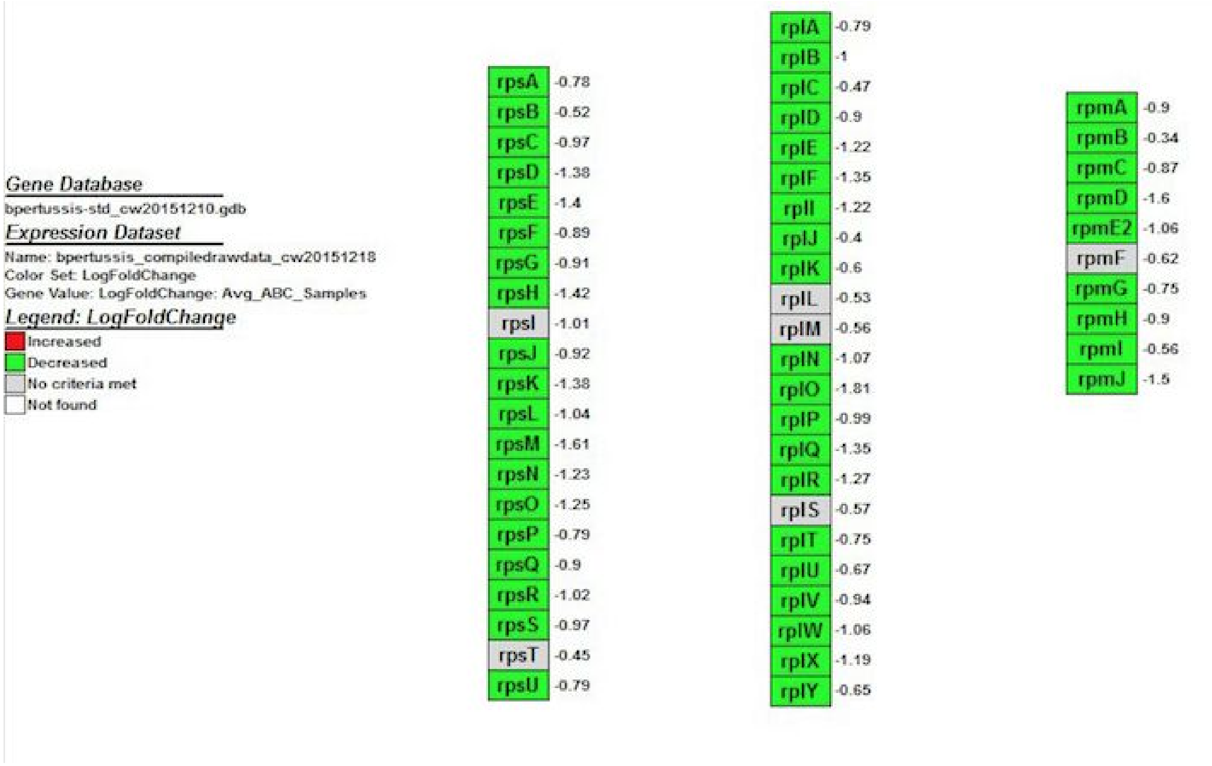


Figure 8 presents the MAPP created using the selected ribosome pathway genes applied to the expression dataset in GenMAPP.

Discussion

The process of building a *Bordetella pertussis* gene database for GenMAPP analysis revealed the efficacy of programs offered by the XMLPipeDB project. Although the program GenMAPP Builder successfully constructed a relational gene database for *B. pertussis* during our vanilla export, missing gene IDs prevented this file from being useful in assessing microarray data. Modifications to GenMAPP Builder were necessary to import both open reading frame (ORF) genes and one specific ID, BP3167A, from the *DBReference* table. The method block used to import ORF genes for the *B. pertussis* custom class was in fact adapted from code that existed in several other existing classes, suggesting that this is a common override to the default export processes. Integrating the option to import ORF genes into this default or as a possible option thus appears warranted. Further, the need to import values from the *DBReference* table presents a more deeply rooted coding issue than was explored in our study. In our case, the unique regex of the EnsemblBacteria reference ID needed from this table allowed us to retrieve it by adding an override that used a pattern-based PostgreSQL query. However, future IDs warranting inclusion from the *DBReference* table may not be isolatable in this manner. The GenMAPP Builder programming could therefore be expanded to create a new table that isolates listings in a new *Reference* table that are distinct from values present in the *OrderedLocusNames* table. Such expansions would facilitate the application of GenMAPP to visualize microarray data on species that do not presently have gene databases.

The statistical analysis we conducted on the existing microarray data yielded results that differed from those produced in the original study. Microarray data downloaded from the original study did not include average log fold changes for the six data replicates. However, calculation of average log fold changes using their reported fold changes differed from those calculated in our study. For the genes *rplX*, *rplQ*, and *rplR* involved in the ribosome biogenesis pathway, our average log fold changes followed by theirs were -1.1886 vs. 0.627, -1.353 vs. 0.735, and -1.022 vs. 0.651 respectively. Because the methods used to calculate log fold changes in the original study were not reported, it is difficult to explain why this discrepancy exists. Further, the original study did not quantify how many genes they observed to have significant expression changes without using relative terms. They did claim to use the Benjamini-Hochberg p-value correction as well, but this cannot be verified.

The final GenMAPP analysis of existing microarray data on *Bordetella pertussis* capsule deletion elucidated pathway-level changes that were not previously characterized. Most notably, a mapp was created demonstrating the nearly global down-regulation of the ribosome biogenesis pathway in Δ*kpst* deletion mutants (Figure 8). Down-regulation of ribosome biosynthesis in an organism is associated with the conservation of energy by reducing protein synthesis in response to environmental changes. Thus, *Bordetella pertussis* isolates without capsules exhibited decreased protein synthesis upon colonizing hosts. The original study reported findings on a separate experiment in which Δ*kpst* mutants that were locked in the Bvg+ phase did not experience significant transcriptional changes when compared to wild type (WT) *Bordetella pertussis* after both were intranasally infected in mice. This suggests that capsule deletion decreased virulence by preventing activation of the BvgA/S regulatory system. It is possible that capsule deletion prevents proper function of proteins in the perisplamic space such as BvgA, preventing the Bvg+ response to host colonization. Regardless, demonstrated inactivation of the BvgA/S complex by capsule deletion provides an explanation for observed down-regulation of ribosome biogenesis. Entering the Bvg+ virulent stage upon host colonization is associated with the up-regulation of various vags that encode *B. pertussis* virulence determinants such as the pertussis toxin and adhesins. Therefore, by effectively locking Δ*kpst* mutants in either the Bvg-intermediate or Bvg- phase, down-regulation of the experimental group's ribosome biogenesis pathway can instead be interpreted as up-regulation of ribosome production in the WT upon entering the Bvg+ phase. This demonstrates that the virulence response in *B. pertussis* requires significant up-regulation of ribosome production and protein synthesis when compared to its seemingly lower energy avirulent phases.

Although we did not entirely mapp metabolic pathways, significant down-regulation of processes associated with pyruvate dehydrogenase was reported based on our filtered MAPPFinder results (Table 4). Notably, down-regulation of coenzyme A biosynthetic and metabolic processes indicates a shift from the TCA cycle in Δ*kpst* mutants. The end result of this metabolic shift is yet to be fully characterized, but preliminary results suggest a possible shift to the nitrogen cycle. This finding links the high-energy yielding process of aerobic respiration to activation of the Bvg+ phase. This response may be dependent on establishment of the presence of a human host and generation of higher yields of ATP to fuel increased protein synthesis as was previously concluded.

Conclusions

Our GenMAPP analysis of microarray data produced from a *B. pertussis* capsule deletion experiment suggested improvements to the XMLPipeDB Project in addition to generating new findings of biological interest. Although GenMAPP is a powerful tool for visualizing global expression changes from microarray array data, it is limited by the number of species-specific gene databases that are currently available. Coding changes that internalized integration of ORF gene listings into gene database exports and allowed for easier retrieval of reference IDs would make this tool more accessible to users with limited access or coding experience. Through circumventing the above issues with coding changes to a custom profile, a comprehensive gene database for *Bordetella pertussis* was crafted. This database enabled a new GenMAPP analysis of existing microarray findings regarding deletion of the *kpst* gene involved with capsule development in *B. pertussis*. This analysis revealed nearly global down-regulation of the ribosome biogenesis pathway and metabolic shift away from aerobic pathways as part of the transcriptional response to capsule deletion. In context, these findings provide new information regarding activation of the Bvg+ virulence phase in *B. pertussis* and the mechanism of this bacterium's pathogenesis.

Moving forward, we would like to further investigate the transcriptional response of *B. pertussis* to BVGA/S activation and its involvement in enabling pathogenesis. Such studies would be designed to approach pathogenic behavior from a systems biology and genomics perspective. Further, we would like to expand the XMLPipeDB project to facilitate the creation of new gene databases. This would increase the accessibility of new users to the tools provided by GenMAPP, allowing its proliferation. Ideally, such modifications to programs associated with GenMAPP would enable future studies similar to the one we conducted to occur more frequently and for new bacteria of interest.

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

Appendix

Gene Database Testing Report- cw20151210

From LMU BioDB 2015

Contents

For convenience, all of the files explicitly asked for in the sections below were compressed together in this file: File:Testingreport cw20151210.zip

Pre-requisites

The following set of software was used in the creation and testing of the Bordetella pertussis gene database:

1. 7-zip (http://www.7-zip.org/)tool that for unpacking .gz and .zip files
2. PostgreSQL (http://www.postgresql.org) on Windows (version 9.4.x)
3. GenMAPP Builder (https://sourceforge.net/projects/xmlpipedb/files/)
4. Java JDK 1.8 64-bit
5. GenMAPP 2 (https://github.com/GenMAPPCS/genmapp)
6. XMLPipeDB match utility (https://sourceforge.net/projects/xmlpipedb/files/) for counting IDs in XML files
7. Microsoft Access for reading .mdb files

Gene Database Creation

Downloading Data Source Files and GenMAPP Builder

We download the UniProt XML, GOA, and GO OBO-XML files for Bordetella Pertussis along with the GenMAPP Builder program.

All files were saved to the folder Bklein7_CW\bpertussis_cw20151210 on our computer's ThawSpace.

Files that required extraction were unzipped using 7-zip (http://www.7-zip.org/).

Data files that remained in a folder after unzipping were removed from their folders to facilitate organization and command line processing.

UniProt XML

We went to the UniProt Complete Proteomes

(http://www.uniprot.org/taxonomy/complete-proteomes) page.

From there, we navigated to the complete proteome download page for Bordetella pertussis

(strain Tohama I / ATCC BAA-589 / NCTC 13251)

(http://www.uniprot.org/proteomes/UP000002676).

We clicked on the "Download" button at the top of the page above and selected the following options:

"Download all"

"XML" from the "Format" drop-down menu

"Compressed" format

We extracted the file using 7-zip (http://www.7-zip.org/).

GOA

UniProt-GOA files can be downloaded from the UniProt-GOA ftp site

(http://ftp.ebi.ac.uk/pub/databases/GO/goa/).

Within the above site, we navigated to the for Bordetella pertussis strain Tohama I

(http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/145.B_pertussis_ATCC_BAA-589.goa).

This text file was automatically opened by the browser. Therefore, we had to manually download the file.

GO OBO-XML

We downloaded the GO OBO-XML formatted file from the Gene Ontology legacy download page (http://geneontology.org/page/download-ontology#Legacy_Downloads).

We extracted the file using 7-zip (http://www.7-zip.org/).

Downloaded GenMAPP Builder

1. 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 - cw20151210): File:Dist cw20151210.zip.

2. We extracted the GenMAPP Builder folder using 7-zip (http://www.7-zip.org/).

Creating the New Database in PostgreSQL

We launched pgAdmin III 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.

We clicked on the Open File icon and selected the file gmbuilder.sql. This imported a series of SQL commands into the editor tab.

We clicked on the Execute Query icon to run this command.

In viewing the schema for this database, we confirmed that there were 167 tables after running the above command.

Configuring GenMAPP Builder to Connect to the PostgreSQL Database

To begin, we launched gmbuilder.bat. We selected the "Configure Database" option and entered the following information into the fields below:

Host or address: localhost

Port number: 5432

Database name: bpertussis_cw20151210_gmb3build5

Username: postgres

Password: Welcome1

Importing Data into the PostgreSQL Database

The downloaded data files for Bordetella pertussis were specified and imported into the database by clicking on the following buttons:

Selected File > Import UniProt XML...

Selected File > Import GO OBO-XML...

Clicked OK to the message asking to process the GO data.

Selected File > Import GOA...

Exporting a GenMAPP Gene Database (.gdb)

We selected File > Export to GenMAPP Gene Database... to begin the export process.

We typed in our coder's name in the owner field (Brandon Klein).

We selected the custom profile "Bordetella pertussis, Taxon ID 257313" as the gene database species and then clicked Next.

The database was saved as bpertussis-std_cw20151210.

We checked the boxes for exporting all Molecular Function, Cellular Component, and Biological Process Gene Ontology Terms.

Finally, we clicked the "Next" button to begin the export process.

Gene Database Testing Report

Export Information

Version of GenMAPP Builder: Version 3.0.0 Build 5 - cw20151210

Computer on which export was run: Seaver 120- Last computer on the right in the row farthest from the front of the room

Postgres Database name: bpertussis_cw20151210_gmb3build5

UniProt XML filename: File:Uniprot-proteome-UP000002676 cw20151210.zip

UniProt XML version (The version information was found at the UniProt News Page (http://uniprot.org/news)): 2015_12

UniProt XML download link: Bordetella pertussis (strain Tohama I / ATCC BAA-589 / NCTC 13251) (http://www.uniprot.org/proteomes/UP000002676)

Time taken to import: 2.88 minutes

Note: The import time was similar to that when creating the previous "Bordetella pertussis" gene database: bpertussis-std_cw20151203.gdb (2.59 minute). No interruptions occurred during this process.

GO OBO-XML filename: File:Go daily-termdb cw20151210.zip

GO OBO-XML version (The version information was found in the file properties): Last Modified- December 10, 2015 (TIME?)

GO OBO-XML download link: Gene Ontology legacy download page (http://geneontology.org/page/download-ontology#Legacy_Downloads)

Time taken to import: 6.97 minutes

Time taken to process: 4.52 minutes

Note: The import and processing times were similar to those for the previous "Bordetella pertussis" gene database: bpertussis-std_cw20151203.gdb (7.08 minutes and 4.42 minutes respectively). No interruptions occurred during these processes.

GOA filename: File:145.B pertussis ATCC BAA-589 cw20151210.zip

GOA version (found in the Last modified field on the FTP site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/)): Last Modified- 08-Dec-2015 02:45

GOA download link: for Bordetella pertussis strain Tohama I (http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/145.B_pertussis_ATCC_BAA-589.goa)

Time taken to import: 0.03 minutes

Note: The import time was very similar to that of the previous "Bordetella pertussis" gene database: bpertussis-std_cw20151203.gdb (0.04 minutes). No interruptionsoccurred during this process.

Name of .gdb file: File:Bpertussis-std cw20151210.zip

Time taken to export:

Start time: 1:19 AM

End time: 2:11 AM

Elapsed time: 52 minutes

Note: No interruptions occurred during the export process.

TallyEngine

We ran the TallyEngine in GenMAPP Builder and specified the following files:

XML- File:Uniprot-proteome-UP000002676 cw20151210.zip

GO- File:Go daily-termdb cw20151210.zip

Results:

All TallyEngine results were consistent across both files.

The TallyEngine was not customized to reflect the coding changes made to GenMAPP Builder Version 3.0.0 Build 5 - cw20151210.

Therefore, the total count for "Ordered Locus Names" and "ORF" gene IDs remained 3446. The extra ID that was imported in this build, "BP3167A", was not listed in either of these categories. Further TallyEngine customization is necessary to raise the count to 3447 gene IDs.

Using XMLPipeDB Match to Validate the XML Results from the TallyEngine

The following functions were performed using the Windows command line (cmd).

We entered the project folder using the following command:

```
cd /d T:\Bklein7_CW\bpertussis_cw20151210
```

We used XMLPipeDB match to identify matches of gene IDs in the UniProt XML file that conformed to the following the patterns: "BP####", "BP####.1", "BP####A", and "BP####B". The command used was as follows: 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"

Match Results:

The number of unique matches generated by XMLPipeDB Match, 3447, matched with our expectation. The count includes the total number of ordered locus (3435)
and ORF (11) gene IDs along with the unique EnsemblBacteria reference ID "BP3167A".

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

We used the SQL "union" operation to count the number of "ordered locus" gene IDs, which conform to the pattern "BP####", in addition to all gene IDs that matched the patterns "BP####A" & "BP####B" (including 11 "ORF" gene IDs and 1 EnsemblBacteria reference ID):

```
select count(value) from (select value from genenametype where type =
```

'ordered locus' union select value from propertytype inner join dbreferencetype

on (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)') as combined;

Note: This query was crafted by Dr. Dionisio.

Results:

The number of unique matches yielded by this SQL query, 3447, matched the count generated by XMLPipeDB Match. Thus, the locations of all 3447 gene IDs in the PostgreSQL relational database were accounted for here.

OriginalRowCounts Comparison

We opened the gene database file File:Bpertussis-std cw20151210.zip in Microsoft Access and assessed the "OriginalRowCounts" table to see if the expected tables were listed with the expected number of records. The contents of this table were compared to the OriginalRowCounts table of an existing .gdb file created during Week 9. Benchmark .gdb file: File:Vc-Std 20151027 TR.gdb

"OriginalRowCounts" table from the benchmark and new gdb:

All 52 tables present in the 2015 Vibrio cholerae database were also present in the B. pertussis gene database, bpertussis-std_cw20151210. This confirmed that all expected tables were successfully created.

The "OrderedLocusNames" table count is listed as 3447. This count demonstrates that the missing ID, "BP3167A", was successfully added to the export (confirmed below).

Note: The "OriginalRowCounts" tables were too large to screenshot. To circumvent this problem and facilitate the comparison, I copied the "OriginalRowCounts" tables from both gene databases into an Excel file and zoomed out. The above screenshot was taken from this Excel file. The "OrderedLocusNames" row count for bpertussis-std_cw20151210 is highlighted in yellow.

Visual Inspection

We visually inspected individual tables within File:Bpertussis-std cw20151210.zip using Microsoft Access.

Systems Table

35 gene ID systems were listed, 11 of which were used in the creation of this .gdb file and listed the appropriate import date (12/10/2015). All gene ID systems relevant to B. pertussis were listed. This includes: EMBL, EnsemblBacteria, GeneID, GeneOntology, InterPro, OrderedLocusNames, Pfam, RefSeq, and UniProt. This result corresponded with that of the benchmark .gdb file listed in the "OriginalRowCounts Comparison" section. The "OrderedLocusNames" listing properly displayed customizations to the Bordetella pertussis species profile.

In this row, the species was listed correctly as "Bordetella pertussis".

In this row, the link corresponded to the Bordetella pertussis database at GeneDB. The link was as follows:

http://www.genedb.org/gene/~;jsessionid=A06A0EFE93C64E476380393D4CBEFA69?actionName=%2FQuery%2FquickSearch&resultsSize=1&taxonNodeName=Bpertussis.

UniProt Table

This table contained 3258 entries with 6 character IDs.

All ID's in the UniProt table conform to the following pattern:

RefSeq Table

This table contained 6627 entries. All IDs began with one of three prefixes: "NP_", "YP_", or "WP_". The meanings of these prefixes can be found in the RefSeq documentation found here (http://www.ncbi.nlm.nih.gov/books/NBK50679/).

"NP_" and "YP_" Prefixes

Refer to proteins. There are 3410 "NP_" IDs and 7 "YP_" IDs.

"WP_" Prefixes

Refer to "autonomous non-redundant proteins that are not yet directly annotated on a genome".

There were 3210 IDs with the "WP_" prefixes.

Overall, every entry in the ID column was an expected value.

OrderedLocusNames Table

This table contained 3447 entries (consistent with the XMLPipeDB Match result).

The IDs were copied into an Excel document for analysis:

3434 IDs conformed to the pattern "BP####".

11 IDs conformed to the pattern "BP####A".

This included 10 ORF gene IDs & "BP3167A" (reference to an EnsemblBacteria ID).

1 ID exhibited the pattern "BP####B".

This corresponded to an ORF gene ID.

1 ID exhibited the pattern "BP####.1".

This ID was the manner in which UniProt classified "BP3167A".

bpertussis-std_cw20151210.gdb Use in GenMAPP

The following analysis was conducted in GenMAPP Version 2.1. Within GenMAPP, the Bordetella pertussis gene database was loaded by selecting Data > Choose Gene Database and then selecting the file bpertussis-std_cw20151210.gdb.

Putting a Gene on the MAPP Using the GeneFinder Window

We made a sample MAPP in which gene IDs conforming to the naming conventions of the 5 major gene databases containing Bordetella pertussis genome data were added. A screenshot of the resulting MAPP is provided below:

Gene IDs:

bp1123 refers to the OrderedLocusNames gene ID system.

CAE43716 refers to the EmsemblBacteria gene ID system.

Q7VWE5 refers to the UniProt gene ID system.

2665491 refers to the GeneID system.

NP_881255 refers to the RefSeq gene ID system.

Note: Gene IDs tested from the above gene ID systems all had complete Backpages and were successfully placed on the MAPP.

Creating an Expression Dataset in the Expression Dataset Manager

The file File:Bpertussis compiledrawdata cw20151208.txt was used to create an expression dataset in GenMAPP.

Total Number of Gene IDs Imported

3211 of the 3552 gene IDs from the microarray dataset were imported into the expression dataset. There were 341 exceptions during the creation of the expression dataset. A screenshot of the error message is shown here:

Investigating Errors in the Exceptions File (EX.txt)

All 341 exceptions triggered the following error message: "Gene not found in OrderedLocusNames or any related system."

Gene IDs that triggered this error message conformed to the patterns "BP####" and "BP####A", indicating that no unique gene ID patterns were the cause of these errors.

Example gene IDs that triggered this error are the following: BP0101, BP1677, BP0910A, and BP2029A. Searching for any of these gene IDs in UniProt returns the message "Sorry, no results found for your search term.":

The 341 gene IDs were copied into a new Excel file and compared to the gene IDs present in the file File:Bpertussis-std cw20151210.zip (adapted from the "OrderedLocusNames" table in Microsoft Access).

None of the 341 gene IDs were present in the .gdb file.

The 341 gene IDs were each individually searched for in UniProt.

None of the 341 gene IDs retrieved results in UniProt.

Conclusion: All gene IDs that triggered errors were not present in the original UniProt XML file.

Coloring a MAPP with Expression Data

Creating a New Color Set

We customized the new Expression Dataset by creating a new color set entitled "LogFoldChange".

1. We created a criterion for this color set to label genes that demonstrated a significant increase in their expression. We specified the gene value as "Avg_ABC_Samples" for the Bordetella pertussis microarray dataset. We activated the Criteria Builder by clicking the New button and named the criterion "Increased". We selected the color for this criterion as red using the color box. We stated the criterion as follows and added it to the Criteria List: [Avg_ABC_Samples] > 0.25 AND [B-H_Pvalue] < 0.05.

2. Second, we created a criterion for this color set to label genes that demonstrated a significant decrease in their expression. We specified the gene value as "Avg_ABC_Samples" for the Bordetella pertussis microarray dataset. We activated the Criteria Builder by clicking the New button and named the criterion "Decreased". We selected the color for this criterion as green using the color box. We stated the criterion as follows and added it to the Criteria List: [Avg_ABC_Samples] < -0.25 AND [B-H_Pvalue] < 0.05

3. Upon entering these color sets, we saved the entire Expression Dataset by selecting Save from the Expression Dataset menu. This effectively updated our .gex file with the new Color Set.

Screenshot of Color Set criteria:

Note: No errors were encountered in the creation of the Color Set.

Creating a Pathway-Based MAPP Using Colored Genes

Ribosome Kegg Pathway

We were able to create a mapp of the ribosome pathway by using the genes provided from the http://www.genome.jp/kegg/ website.

Once accessing the website, we selected KEGG PATHWAY from the main page.

Next, we scrolled down to "Ribosome" that was under section 2.2 Translation and selected it.

Then, we searched our organism in the drop down menu at the top of the page, and we selected the Bordetella pertussis Tomaha I organism, and clicked "Go".

This lead us to a page of the ribosome pathway with the gene IDs that pertained to our specific organism. We were then able to create a mapp using these genes in

GenMAPP. Each of the green highlighted genes on the ribosome pathway were entered into the GenMAPP mapp by entering each gene ID and the name given from the Kegg pathway, and then the expression dataset "bpertussis_expressiondataset_cw20151218" was applied to the genes to color code them.

Here is the picture of the final mapp for the ribosome pathway created:

Most of the ribosome genes that were generated on this mapp appeared to be the color green, symbolizing a decrease, except for the grey colored genes that were not significantly changed in this experiment. Since the genes mapped for the ribosome pathway all appeared to be green, this means that the expression levels of the genes pertaining to the ribosome category all decreased during the microarray experiment. Ribosomes play a key role in the translation process in cells and without them genes are often repressed and unable to perform their proper functions as they are unable to complete the replication processes. The microarray experiment analysis revealed that the absence of a membrane-associated protein named KpsT in B. pertussis, resulted in global down-regulation of gene expression including key virulence genes. The ribosome pathway depicted genes that were decreasing in gene expression, thus linking the translation process to the down-regulated key genes from the experiment because since these genes were lacking a necessary protein to help them perform the proper replication processes, translation did not occur in these genes and thus the ribosomes were not involved, ultimately leading to the decrease in expression of the genes mapped in the ribosome pathway.

Nitrogen Cycle Kegg Pathway

We were also able to create another mapp using the nitrogen cycle pathway genes provided from the http://www.genome.jp/kegg/ website.

Once accessing the website, we selected KEGG PATHWAY from the main page.

Next, we scrolled down to "Nitrogen Metabolism" that was under section 1.2 Energy Metabolism and selected it. Then, we searched our organism in the drop down menu at the top of the page, and we selected the Bordetella pertussis Tomaha I organism, and clicked "Go".

This lead us to a page of the nitrogen metabolism pathway with the gene IDs that pertained to our specific organism. We were then able to create a mapp using these genes in GenMAPP.

Each of the green highlighted genes on the nitrogen metabolism pathway were entered into the GenMAPP mapp by entering each gene ID and the name given from the Kegg pathway, and then the expression dataset "bpertussis_expressiondataset_cw20151218" was applied to the genes to color code them.

Here is the picture of the final mapp for the nitrogen cycle pathway created:

This mapp displayed both red and green colored genes; the green highlighted genes symbolizing a decrease and the red highlighted genes symbolizing an increase, as well a couple of gray genes that were not significant to the criterion. This nitrogen cycle mapp was created due to the important metabolic processes that occur in order to keep cells alive and reproducing, and specifically the nitrogen metabolism cycle. The genes that displayed red in this mapp had increased expression during the microarray experiment, and from the kegg pathway given for nitrogen metabolism, these genes can be seen to specifically aid in the metabolism of glutamate. Glutamate is important to cells as it plays a role in providing energy to allow the cells to operate correctly, and since the glutamate-related genes that we mapped were increased, it can be determined that glutamate plays a role in supplying the underlying energy to allow for the Bordetella pertussis strains to produce the polysaccharide capsule transport proteins, as studied in the microarray experiment.

Running MAPPFinder

MAPPFinder Procedure

We launched the MAPPFinder program from within GenMAPP and ensured that the bpertussis-std_cw20151210.gdb gene database was still loaded into GenMAPP.

We clicked on the button "Calculate New Results" followed by "Find File", at which point I specified the .gex file updated during the creation of the "LogFoldChange" color set.

We chose to apply both the "Increased" and "Decreased" criteria present within the LogFoldChange color set to the data.

We checked the boxes next to "Gene Ontology" and "p value", specified the results file, and then clicked "Run MAPPFinder".

This analysis took several minutes to complete.

MAPPFinder Analysis Results

We selected "Show Ranked List" to see a list of the most significant Gene Ontology terms. A screenshot of this output is shown below:

The majority of the most significant gene ontology terms pertained to ribosome biosynthesis and translation.

Note: The MAPPFinder analysis took approximately 8 minutes to complete. No errors were encountered in the process. MAPPFinder thus was confirmed to work with the Bordetella pertussis gene database.

Compare Gene Database to Outside Resource

To assess the completeness of this version of the Bordetella pertussis gene database, we explored the original genome sequencing data from Parkhill et al. (2003) that was deposited at the GeneDB Model Organism Database (MOD) (http://www.genedb.org/Homepage/Bpertussis). From the GeneDB Home Page, we accessed a Gene Type search function that was used to quantify the number of gene listings present under each provided gene category. The results of this investigation are presented below.

Protein-Coding Genes

There are 3447 protein-coding genes present in the GeneDB (http://www.genedb.org/Homepage/Bpertussis) database. This result verified that the set of protein-coding genes exported into File:Bpertussis-std cw20151210.zip from UniProt is complete. No further changes to the gene database export procedures are necessary at this time.

Non-Protein Genome Features

1. Pseudogenes

GeneDB indicated that 359 pseudogenes are present in the B. pertussis genome. Pseudogenes do not code for proteins and were therefore not included in the original UniProt listing.

2. rRNA

GeneDB indicated that 9 genes that encode for rRNA are present in the B. pertussis genome. These genes do not code for proteins and were therefore not included in the original UniProt listing.

3. tRNA

GeneDB indicated that 51 genes that encode for tRNA are present in the B. pertussis genome. These genes do not code for proteins and were therefore not included in the original UniProt listing.

4. snoRNA

GeneDB retrieved 0 genes that encode for snoRNA.

5. snRNA

GeneDB retrieved 0 genes that encode for snRNA.

6. "miscRNA"

GeneDB retrieved 0 genes that encode for "miscRNA".

A total of 419 non-protein coding genes were identified in the Bordetella pertussis genome in addition to the 3447 protein-coding genes captured in our gene database.