

# Creation of a GenMAPP Gene Database for *Bordetella pertussis* Tohama I

Brandon Klein  
Mahrad Saeedi  
Elena Olufson

BIOL/CMSI 367: Biological Databases  
Loyola Marymount University  
December 15, 2015

# Overview

- *Bordetella pertussis* is a well-documented respiratory pathogen and is commonly known as whooping cough
- Creating a new gene database for *B. pertussis* and testing for accuracy and completeness
- Discussing and analyzing the microarray experimental procedure
- Analyzing the results of the DNA microarray experiment

# Overview

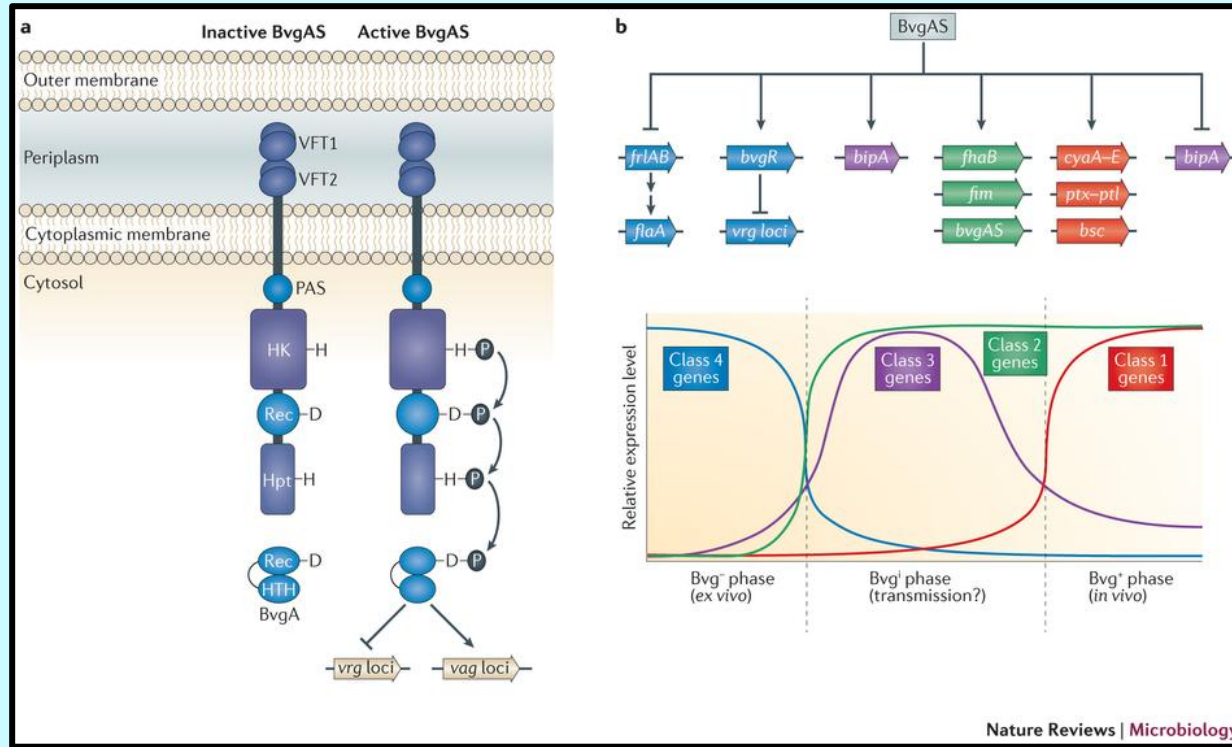
- *Bordetella pertussis* is a well-documented respiratory pathogen and is commonly known as whooping cough
- Creating a new gene database for *B. pertussis* and testing for accuracy and completeness
- Discussing and analyzing the microarray experimental procedure
- Analyzing the results of the DNA microarray experiment

# ***Bordetella pertussis* is the Causative Agent of the Whooping Cough Respiratory Infection**

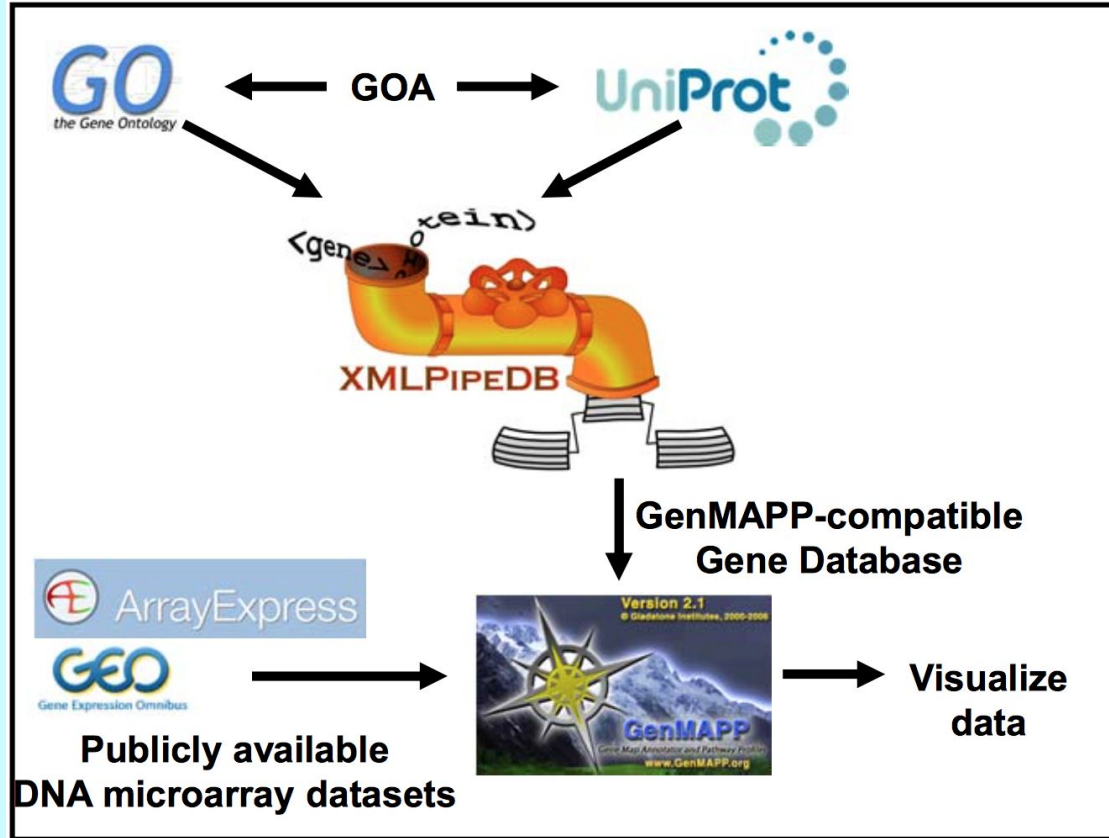
- Whooping Cough (pertussis)
  - Top 10 infectious disease
  - Leading cause of vaccine-preventable deaths
  - 2015 Statistics
    - ~16 million cases
    - 195,000 deaths in children



# The Pathogenesis of *Bordetella pertussis* is Controlled by the BvgAS Regulatory System



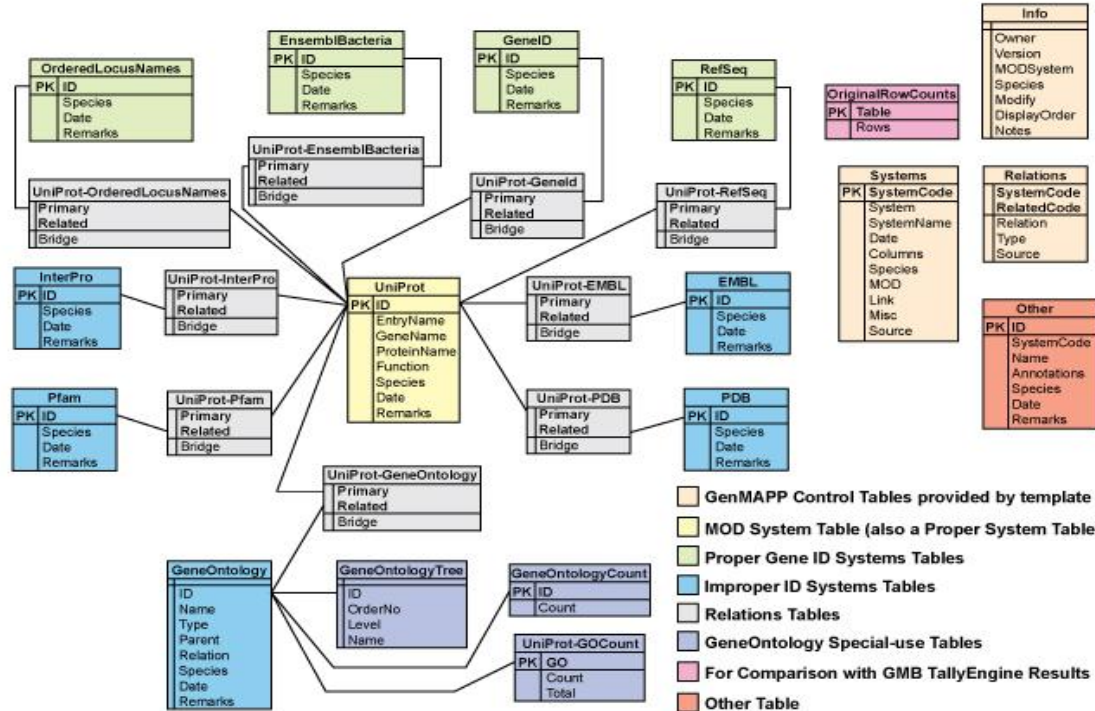
# A GenMAPP Gene Database for *Bordetella pertussis* was Created Using XMLPipeDB Tools



# Overview

- *Bordetella pertussis* is a well-documented respiratory pathogen and is commonly known as whooping cough
- Creating a new gene database for *B. pertussis* and testing for accuracy and completeness
- Discussing and analyzing the microarray experimental procedure
- Analyzing the results of the DNA microarray experiment

# GenMAPP Gene Database Schema for *Bordetella pertussis* Tohama I



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).

*bpertussis-std\_cw20151210.gdb*

- Created with the program GenMAPP Builder
  - XMLPipeDB Project
- B. pertussis* customizations:
  - Systems
  - OrderedLocusNames



# The *B. pertussis* Gene Database Contains a Complete Set of 3,447 Protein-Coding Genes

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

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

# XMLPipeDB Match Was Used to Count the Number of Gene IDs in the Original XML File

- A regex was crafted to capture all identified gene ID patterns:
  - BP##### BP#####.1 BP#####A BP#####B

## COMMAND

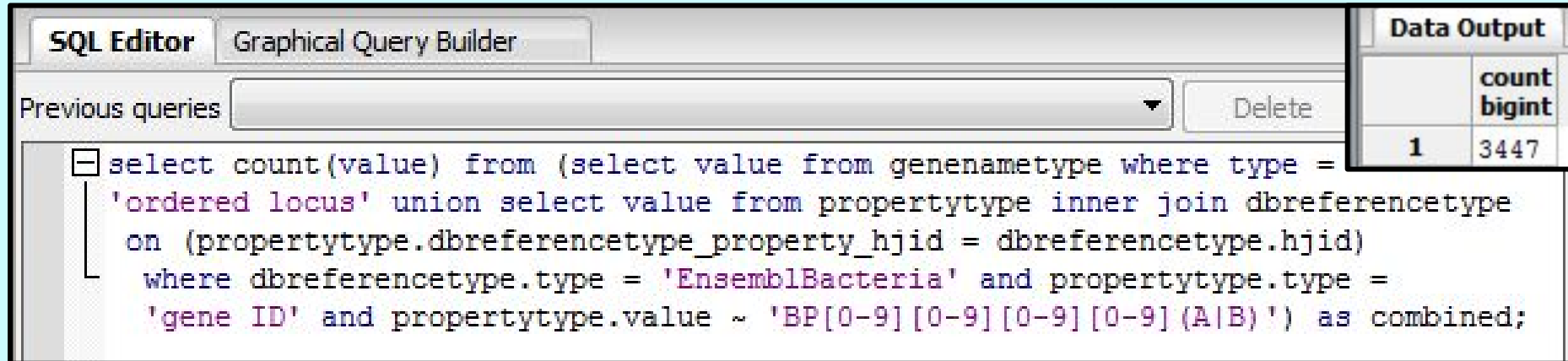
```
java -jar xmlpipedb-match-1.1.1.jar "BP[0-9][0-9][0-9][0-9](A|B|\.1|)"  
<"uniprot-proteome%3AUP000002676_cw20151201.xml"
```

## OUTPUT

Total unique matches: 3447

# PGAdmin III Was Used to Count the Number of Gene IDs in the Relational Database (20151210)

- PostgreSQL Database: *bpertussis\_cw20151210\_gmb3build5*
- SQL Query: the “union” operation was used to count . . .
  - “ordered locus” gene IDs (3435)
  - EnsemblBacteria reference IDs with the pattern “BP####(A|B)” (12)



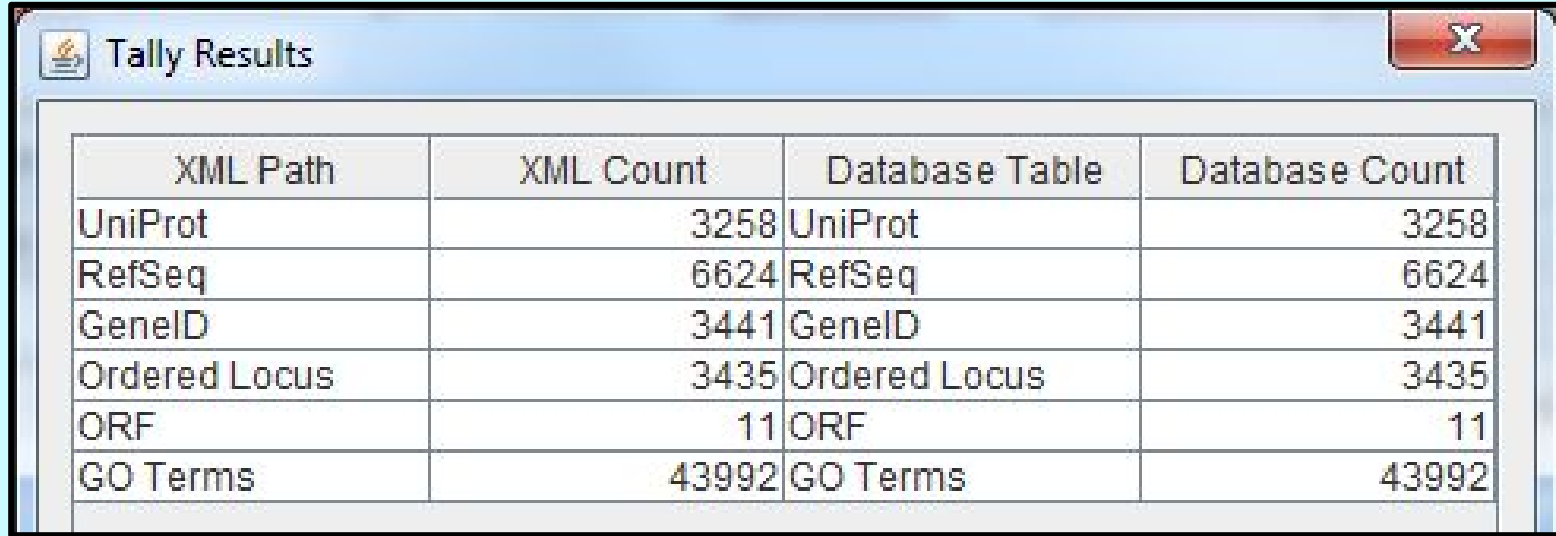
The screenshot shows the PGAdmin III interface. The top bar has two tabs: "SQL Editor" (selected) and "Graphical Query Builder". Below the tabs is a "Previous queries" dropdown menu and a "Delete" button. The main area contains a SQL query:

```
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;
```

On the right side, a "Data Output" window is open, displaying the results of the query:

	count bigint
1	3447

# The GenMAPP Builder TallyEngine Counted Gene IDs in the XML and PostgreSQL Database



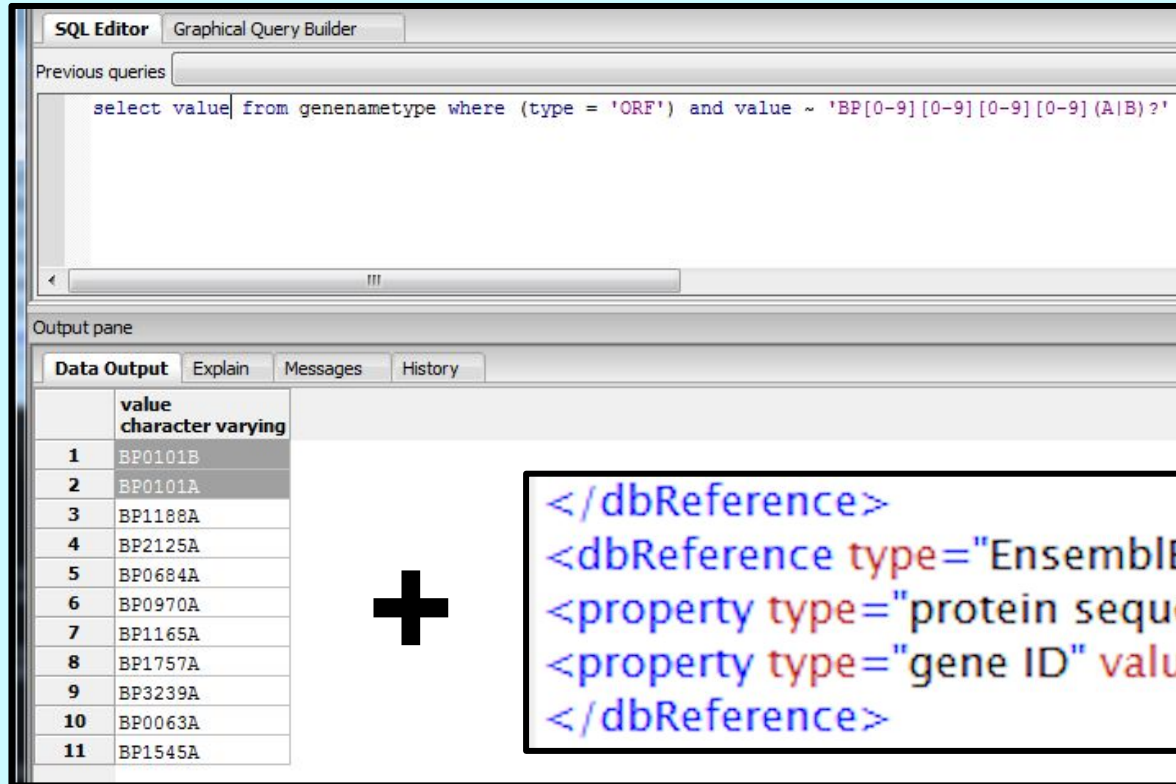
The screenshot shows a window titled 'Tally Results' with a table comparing XML and PostgreSQL data. The table has four columns: XML Path, XML Count, Database Table, and Database Count. The data is as follows:

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

Generated using GenMAPP Builder version 3.0.0 Build 5 - cw20151210

\*Customized to count ORF Values

# Initial Versions of the *B. pertussis* Gene Database Were Missing 12 OrderedLocusNames IDs



SQL Editor Graphical Query Builder

Previous queries

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

Output pane

Data Output Explain Messages History

	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

+

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

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

# Unique Patterns Amongst Missing IDs 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;
```

Output pane

Data Output

Explain

Messages

History

	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

Output Includes:

- All 11 ORF gene IDs
- 1 Reference ID (BP3167A)
- No extraneous gene IDs

# A New Method Block Was Added to the *B. pertussis* Custom Profile Code to Import Missing IDs

```
44 + // Start with the default OrderedLocusNames behavior.
45 + TableManager result = super.getSystemTableManagerCustomizations(tableManager, primarySystemTableManager,
46 + version);
47 +
48 + String sqlQuery = "select dbreferencetype.entrytype_dbreference_hjid as hjid, propertytype.value from propertytype inner join dbrefer
49 + "(propertytype.dbreferencetype_property_hjid = dbreferencetype.hjid) " +
50 + "where dbreferencetype.type = 'EnsemblBacteria' and propertytype.type = 'gene ID' " +
51 + "and propertytype.value ~ 'BP[0-9][0-9][0-9][0-9](A|B)' order by propertytype.value";
52 +
53 + Connection c = ConnectionManager.getRelationalDBConnection();
54 + PreparedStatement ps;
55 + ResultSet rs;
56 + try {
57 + // Query, iterate, add to table manager.
58 + ps = c.prepareStatement(sqlQuery);
59 + rs = ps.executeQuery();
60 + while (rs.next()) {
61 + String hjid = Long.valueOf(rs.getLong("hjid")).toString();
62 + String id = rs.getString("value");
63 + result.submit("OrderedLocusNames", QueryType.insert, new Object[][] {
64 + { "ID", id },
65 + { "Species", "|" + getSpeciesName() + "|" },
66 + { "Date", version },
67 + { "UID", hjid }
68 + });
69 + }
70 + } catch (SQLException sqlexc) {
71 + logSQLException(sqlexc, sqlQuery);
72 + }
73 +
74 + return result;
```

SQL  
Query

# All 341 IDs that Triggered Errors During Analysis Were Not Present in the UniProt XML File

UniProtKB bp0101

ID mapping

results

0 results in UniProtKB

Raw Data File Conversion

341 errors were detected in your raw data. Check the exception file:  
T:\Lena\final GenMAPP  
analysis\Bpertussis\_CompiledRawData\_MS2015-3.EX.bt  
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.

OK

Sorry, no results found for your search term.

Can't find what you are looking for? Please [contact us](#)



# Overview

- *Bordetella pertussis* is a well-documented respiratory pathogen and is commonly known as whooping cough
- Creating a new gene database for *B. pertussis* and testing for accuracy and completeness
- Discussing and analyzing the microarray experimental procedure
- Analyzing the results of the DNA microarray experiment

# Polysaccharide capsules in *B. pertussis* to determine virulence

- *B. pertussis* produces an intact polysaccharide (PS) microcapsule
- Testing  $\Delta$ KpsT mutant against the wild-type
- Experiment determining the impact of PS capsules on the virulence of *B. pertussis*

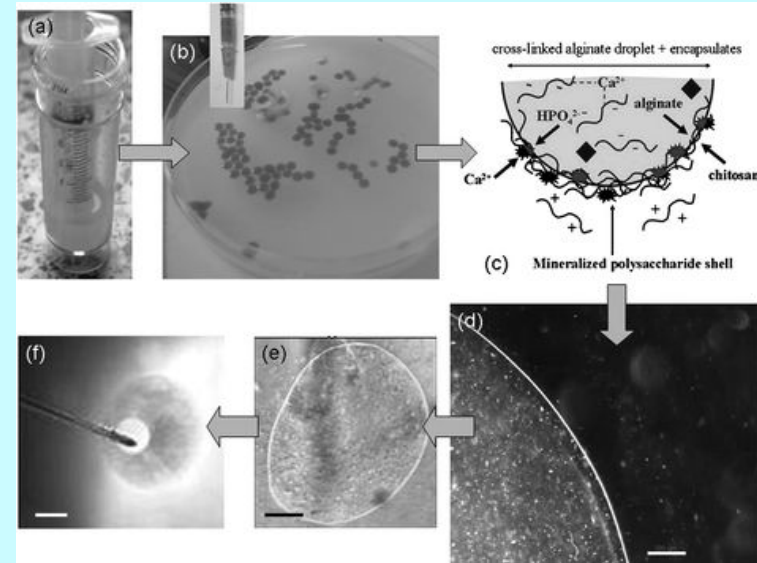
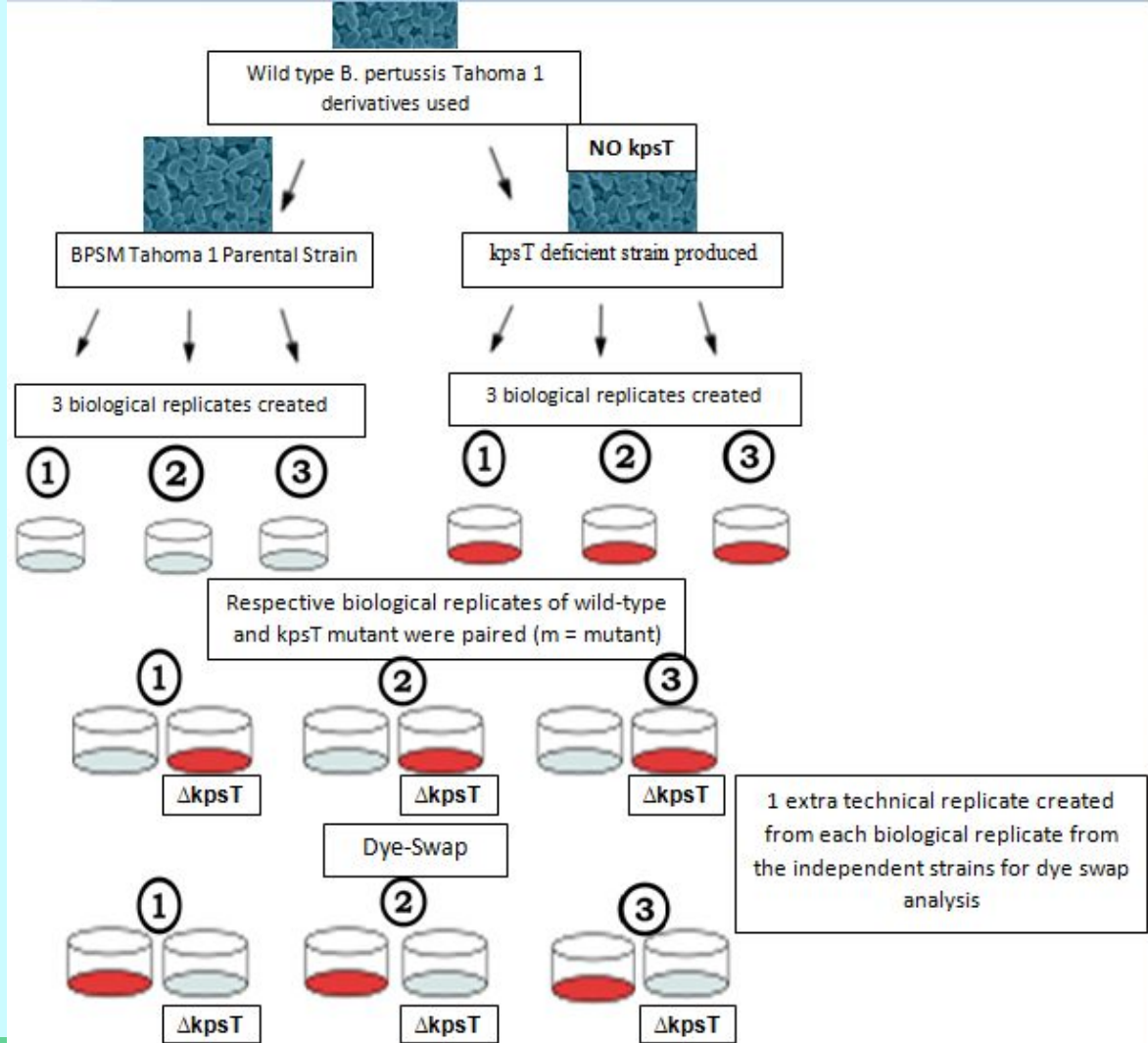


Figure: Depiction of polysaccharide microcapsule using various visual instruments.

# Process of Replicate Production



# DNA Microarray Analysis Sanity Check Reveals the Significantly Changed Genes

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

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

# Overview

- *Bordetella pertussis* is a well-documented respiratory pathogen and is commonly known as whooping cough
- Creating a new gene database for *B. pertussis* and testing for accuracy and completeness
- Discussing and analyzing the microarray experimental procedure
- Analyzing the results of the DNA microarray experiment

# Filtering MAPPFinder Results Indicated Significantly Up-Regulated Genes

GOID	GO Name	#_Changed_Local	#_GO_Local	%_Changed_Local	#_Changed	#_GO	%_Changed	Pvalue	AdjustedPvalue
16226	iron-sulfur cluster assembly	5	6	83.33334	6	7	85.71429	0.001	0.186
31163	metallo-sulfur cluster assembly	0	0	0	6	7	85.71429	0.001	0.186
70887	cellular response to chemical stimulus	0	0	0	6	4	75	0.001	0.594
71941	nitrogen cycle metabolic process	0	0	0	5	1	83.33334	0.002	0.536
19363	pyridine nucleotide biosynthetic process	4	6	66.66666	6	7	75	0.003	0.594
19674	NAD metabolic process	1	1	100	6	8	75	0.003	0.594
19359	nicotinamide nucleotide biosynthetic process	0	0	0	6	7	75	0.003	0.594
9435	NAD biosynthetic process	5	7	71.42857	5	7	71.42857	0.011	0.973
16782	transferase activity, transferring sulfur-containing groups	0	0	0	6	4	60	0.014	0.992
50801	ion homeostasis	0	0	0	5	1	71.42857	0.017	0.973
98771	inorganic ion homeostasis	0	0	0	5	1	71.42857	0.017	0.973
48878	chemical homeostasis	0	0	0	5	1	71.42857	0.017	0.973
6461	protein complex assembly	1	1	100	5	4	62.5	0.026	0.997
70271	protein complex biogenesis	0	0	0	5	4	62.5	0.026	0.997
19720	Mo-molybdopterin cofactor metabolic process	0	1	0	5	8	62.5	0.026	0.997
6777	Mo-molybdopterin cofactor biosynthetic process	5	8	62.5	5	8	62.5	0.026	0.997
72329	monocarboxylic acid catabolic process	0	0	0	6	1	54.54546	0.031	1
6979	response to oxidative stress	4	8	50	6	12	50	0.046	1
51189	prosthetic group metabolic process	0	0	0	5	8	55.55556	0.049	1
32324	molybdopterin cofactor biosynthetic process	1	2	50	5	9	55.55556	0.049	1
43545	molybdopterin cofactor metabolic process	0	0	0	5	8	55.55556	0.049	1

Table lists the top 21 GO terms generated from the “Increased” criteria applied to the data.

# Filtering MAPPFinder Results Indicated Significantly Down-Regulated Genes

GOID	GO Name	#_Changed_Local	#_GO_Local	%_Changed_Local	#_Changed	#_GO	%_Changed	Pvalue	AdjustedPvalue
15937	coenzyme A biosynthetic process	4	4	100	6	6	100	0	0.035
15936	coenzyme A metabolic process	0	0	0	6	6	100	0	0.035
33866	nucleoside bisphosphate biosynthetic process	0	0	0	6	6	100	0	0.035
34030	ribonucleoside bisphosphate biosynthetic process	0	0	0	6	6	100	0	0.035
34033	purine nucleoside bisphosphate biosynthetic process	0	0	0	6	6	100	0	0.035
45261	proton-transporting ATP synthase complex, catalytic core F(1)	5	5	100	5	5	100	0.001	0.161
33178	proton-transporting two-sector ATPase complex, catalytic domain	2	2	100	5	5	100	0.001	0.161
33865	nucleoside bisphosphate metabolic process	0	0	0	6	6	75	0.005	0.714
34032	purine nucleoside bisphosphate metabolic process	0	0	0	6	6	75	0.005	0.714
33875	ribonucleoside bisphosphate metabolic process	0	0	0	6	6	75	0.005	0.714
4312	fatty acid synthase activity	0	0	0	6	2	66.66666	0.009	0.972
46493	lipid A metabolic process	0	0	0	6	9	66.66666	0.016	0.972
6643	membrane lipid metabolic process	0	0	0	6	9	66.66666	0.016	0.972
6664	glycolipid metabolic process	0	0	0	6	9	66.66666	0.016	0.972
46467	membrane lipid biosynthetic process	0	0	0	6	9	66.66666	0.016	0.972
2E+06	lipooligosaccharide metabolic process	0	0	0	6	9	66.66666	0.016	0.972
2E+06	lipooligosaccharide biosynthetic process	0	0	0	6	9	66.66666	0.016	0.972
9245	lipid A biosynthetic process	6	9	66.66666	6	9	66.66666	0.016	0.972
9247	glycolipid biosynthetic process	0	0	0	6	9	66.66666	0.016	0.972
6119	oxidative phosphorylation	1	1	100	5	7	71.42857	0.02	0.992
4003	ATP-dependent DNA helicase activity	5	8	62.5	5	8	62.5	0.022	1
5694	chromosome	2	6	33.33333	6	10	60	0.024	0.996
6261	DNA-dependent DNA replication	1	4	25	6	8	54.54546	0.04	1
3746	translation elongation factor activity	5	8	62.5	5	8	62.5	0.044	1
5507	copper ion binding	5	9	55.55556	5	9	55.55556	0.045	1

Table lists the top 25 GO terms generated from the “Decreased” criteria applied to the data.

# Genes Involved in the Ribosome Biosynthesis Pathway Were Consistently Down-Regulated

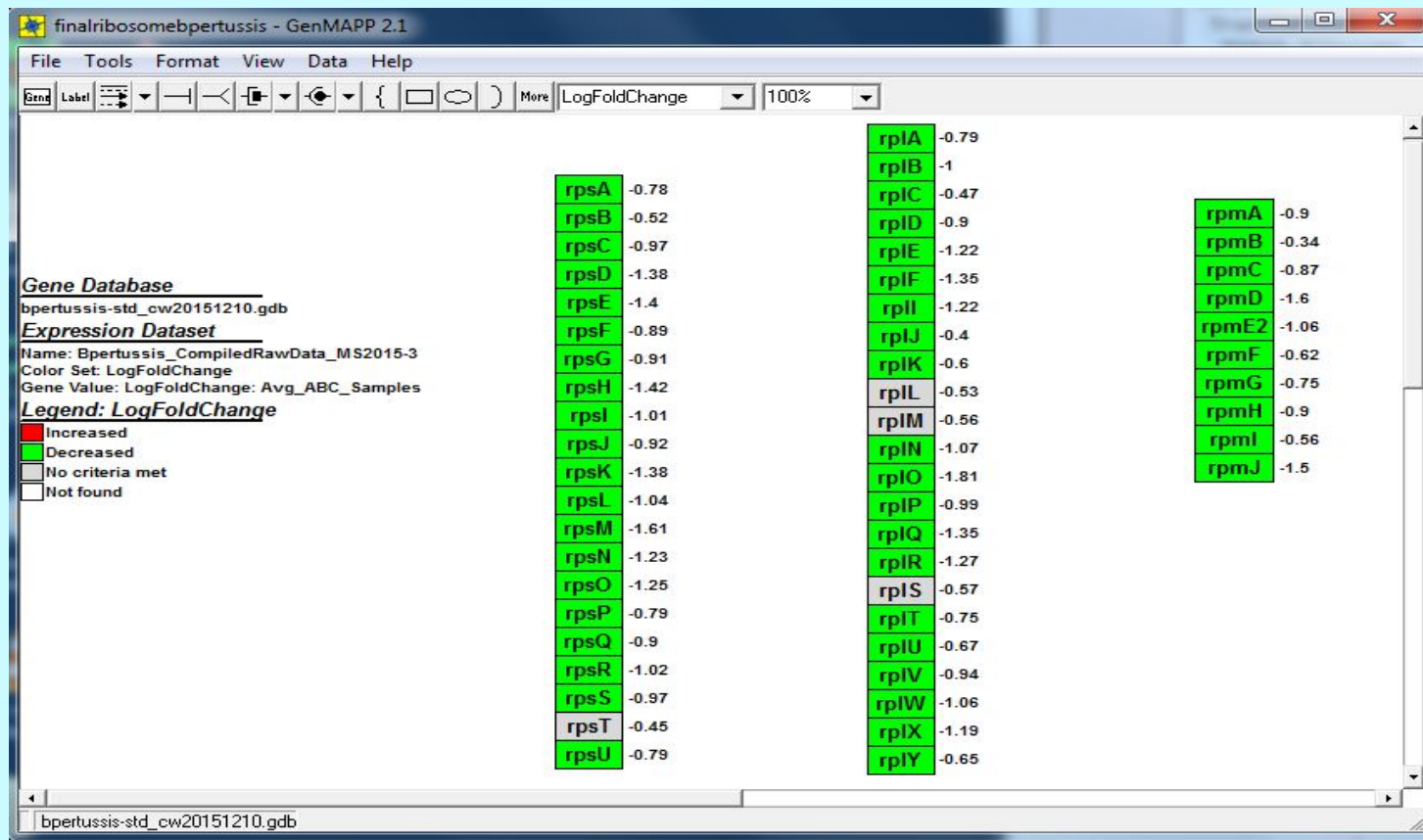


Figure presents the MAPP created using the selected ribosome pathway genes applied to the expression dataset in GenMAPP.



# Acknowledgments

- Dr. Dionisio
- Dr. Dahlquist
- Seaver College of Science & Engineering
- Biological Databases Students
  - Thank you for listening!



# 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

**Questions?**

