Gene Database Testing Report- cw20151210

From LMU BioDB 2015

Contents

Files Asked for in the Gene Database Testing Report

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

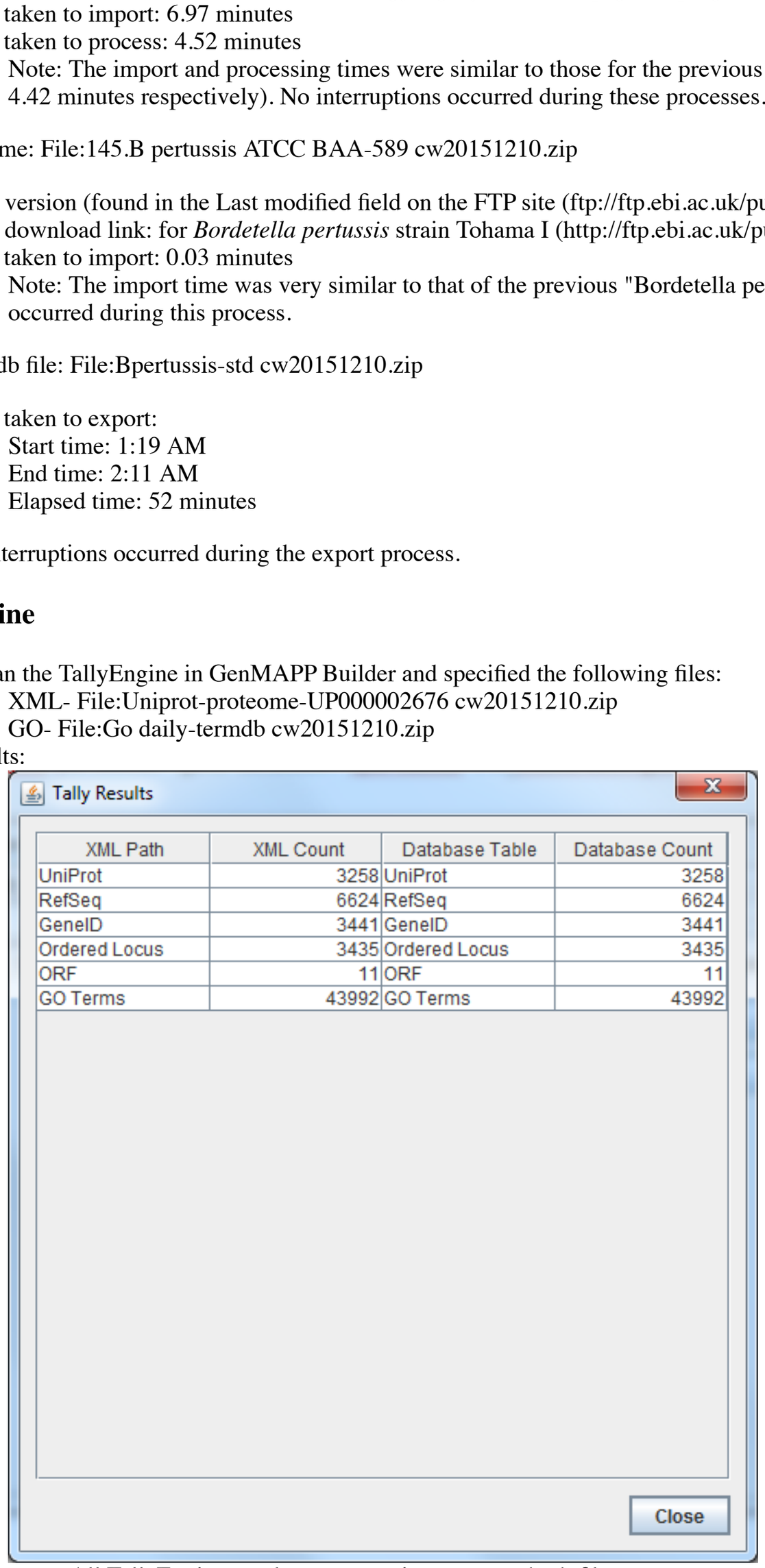
- GOA version (found in the Last modified field on the FTP site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/)): Last Modified- 08-Dec-2015 02:45
- GOA download link: for *Bordetella pertussis* strain Tohama I (http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/145.B_pertussis_ATCC_BAA-589.goa)
- 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 interruptions occurred 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).

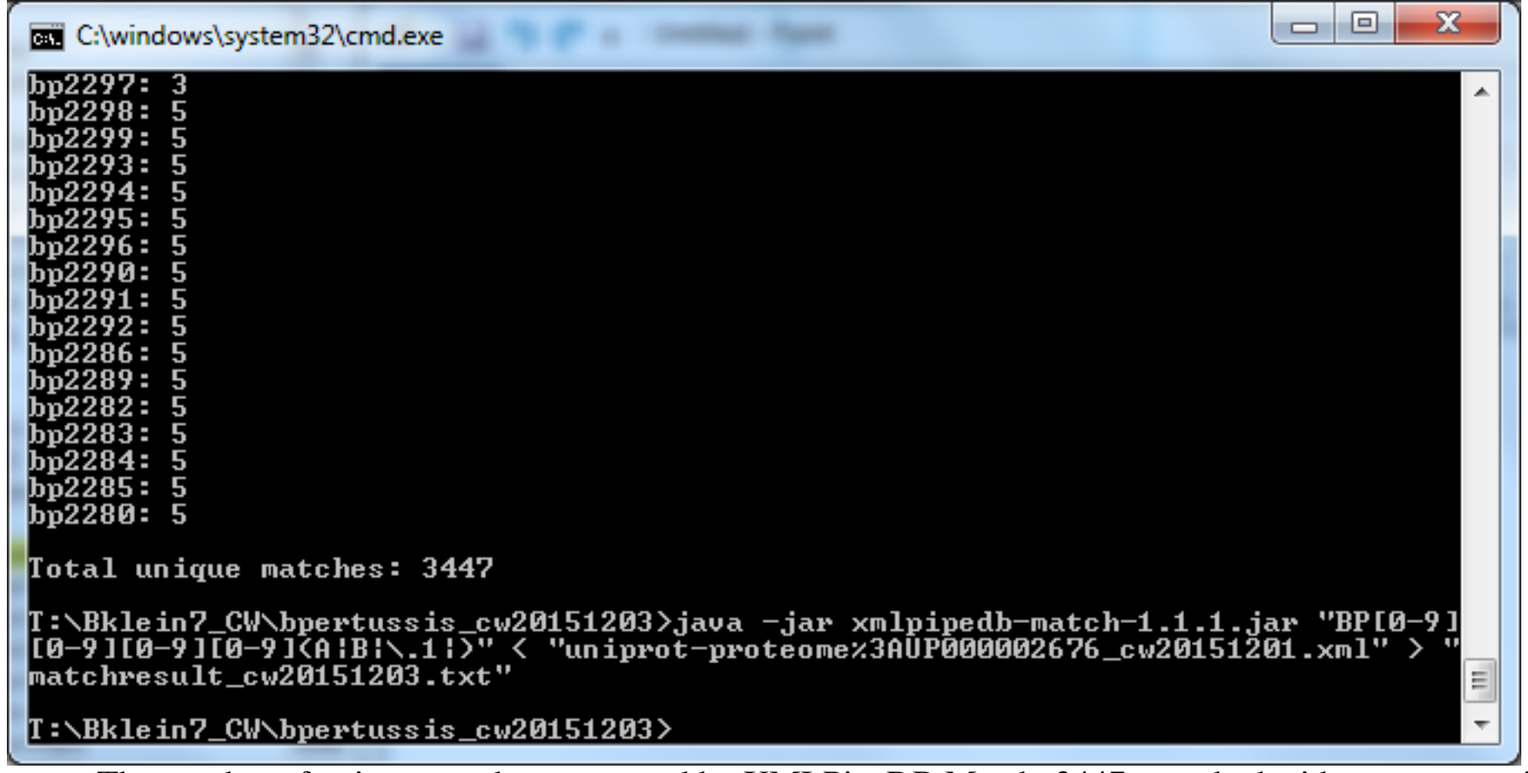
- We entered the project folder using the following command:

```
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".

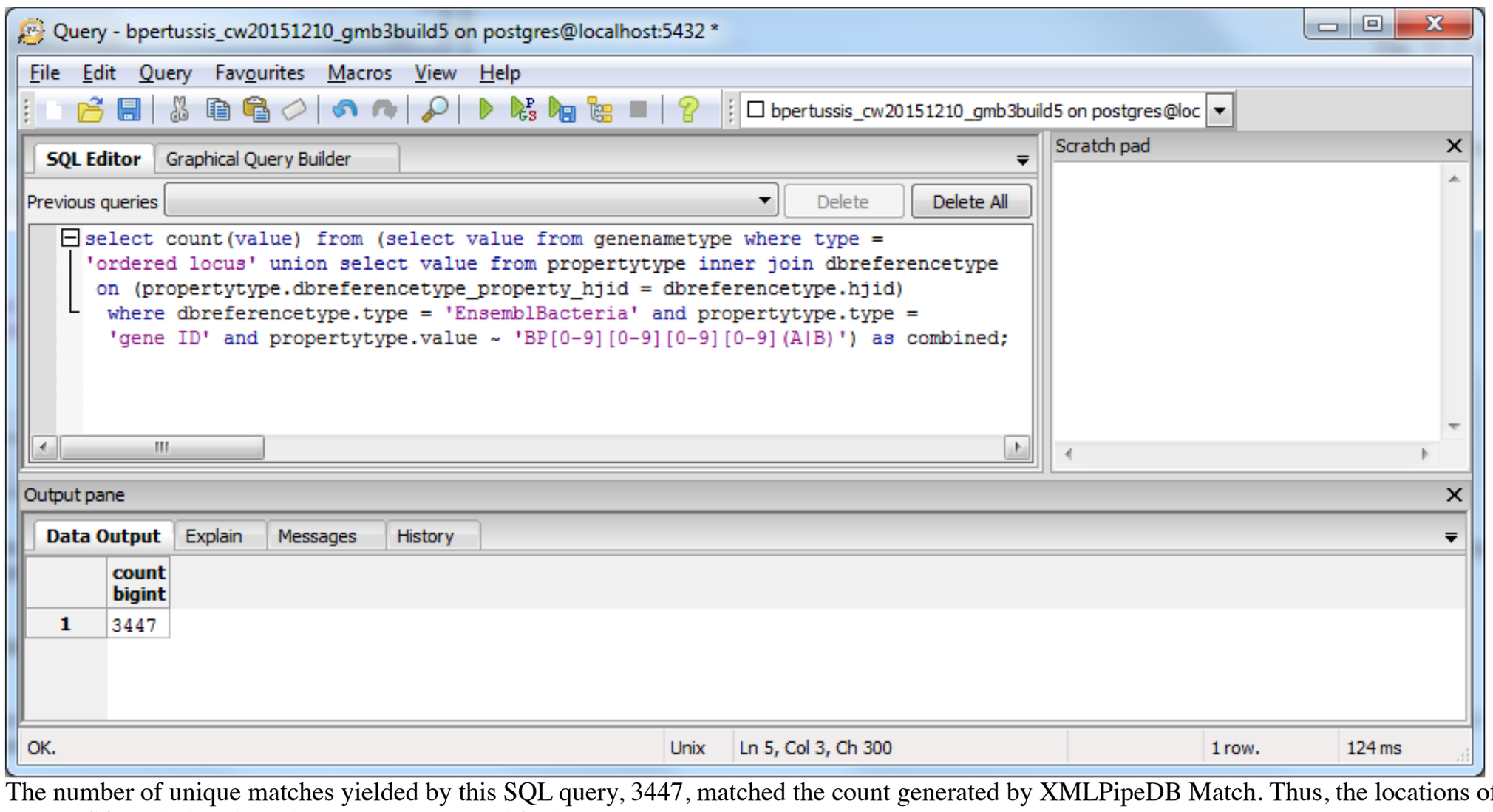
Using SQL Queries to Validate the PostgreSQL Database Results from the TallyEngine

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:

bpertussis-std_cw20151210.gdb		Vc-std_20151027_TR.gdb	
Table	Rows	Table	Rows
Info	1	Info	1
Systems	35	Systems	35
Relations	35	Relations	35
Other	0	Other	0
GeneOntologyTree	79609	GeneOntologyTree	127188
GeneOntology	4836	GeneOntology	6563
UniProt-GOCount	2756	UniProt-GOCount	3693
GeneOntologyCount	2755	GeneOntologyCount	3692
UniProt-GeneOntology	13072	UniProt-GeneOntology	22269
UniProt	3258	UniProt	3789
RefSeq	6624	RefSeq	6697
EMBL	67	EMBL	176
Pfam	1794	Pfam	2116
InterPro	3681	InterPro	4452
GeneID	3441	GeneID	3338
EnsemblBacteria	3444	EnsemblBacteria	3828
PDB	43	PDB	297
OrderedLocusNames	3447	OrderedLocusNames	7664
UniProt-PDB	85	UniProt-PDB	414
UniProt-OrderedLocusNames	3447	UniProt-OrderedLocusNames	7664
UniProt-EnsemblBacteria	3493	UniProt-EnsemblBacteria	4069
UniProt-GeneID	3491	UniProt-GeneID	3576
UniProt-InterPro	9536	UniProt-InterPro	10688
UniProt-Pfam	4130	UniProt-Pfam	4596
UniProt-EMBL	3589	UniProt-EMBL	5461
UniProt-RefSeq	6728	UniProt-RefSeq	7159
RefSeq-EMBL	8532	RefSeq-EMBL	7216
RefSeq-Pfam	8423	RefSeq-Pfam	8145
RefSeq-InterPro	19639	RefSeq-InterPro	19009
RefSeq-GeneID	14207	RefSeq-GeneID	6814
RefSeq-EnsemblBacteria	14010	RefSeq-EnsemblBacteria	6848
RefSeq-OrderedLocusNames	14010	RefSeq-OrderedLocusNames	13700
RefSeq-PDB	125	RefSeq-PDB	627
GeneID-EMBL	5196	GeneID-EMBL	3607
GeneID-Pfam	4437	GeneID-Pfam	4071
GeneID-InterPro	10462	GeneID-InterPro	9474
GeneID-EnsemblBacteria	10774	GeneID-EnsemblBacteria	3452
GeneID-OrderedLocusNames	10774	GeneID-OrderedLocusNames	6904
GeneID-PDB	55	GeneID-PDB	314
EnsemblBacteria-EMBL	5164	EnsemblBacteria-EMBL	4103
EnsemblBacteria-Pfam	4432	EnsemblBacteria-Pfam	4291
EnsemblBacteria-InterPro	10447	EnsemblBacteria-InterPro	9966
EnsemblBacteria-OrderedLocusNames	10643	EnsemblBacteria-OrderedLocusNames	7944
EnsemblBacteria-PDB	55	EnsemblBacteria-PDB	317
OrderedLocusNames-EMBL	5168	OrderedLocusNames-EMBL	8224
OrderedLocusNames-Pfam	4436	OrderedLocusNames-Pfam	8594
OrderedLocusNames-InterPro	10454	OrderedLocusNames-InterPro	19962
OrderedLocusNames-PDB	55	OrderedLocusNames-PDB	634
GeneID-GeneOntology	13652	GeneID-GeneOntology	20412
RefSeq-GeneOntology	26423	RefSeq-GeneOntology	41064
OrderedLocusNames-GeneOntology	13645	OrderedLocusNames-GeneOntology	44858
EnsemblBacteria-GeneOntology	13626	EnsemblBacteria-GeneOntology	22412

- 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).**

OrderedLocusNames

ID	Species	Date
BP1285	Bordetella pe	12/10/2015
BP0213	Bordetella pe	12/10/2015
BP3410	Bordetella pe	12/10/2015
BP0063A	Bordetella pe	12/10/2015
BP0101A	Bordetella pe	12/10/2015
BP0101B	Bordetella pe	12/10/2015
BP0684A	Bordetella pe	12/10/2015
BP0970A	Bordetella pe	12/10/2015
BP1165A	Bordetella pe	12/10/2015
BP1188A	Bordetella pe	12/10/2015
BP1545A	Bordetella pe	12/10/2015
BP1757A	Bordetella pe	12/10/2015
BP2125A	Bordetella pe	12/10/2015
BP3167A	Bordetella pe	12/10/2015
BP3239A	Bordetella pe	12/10/2015

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.

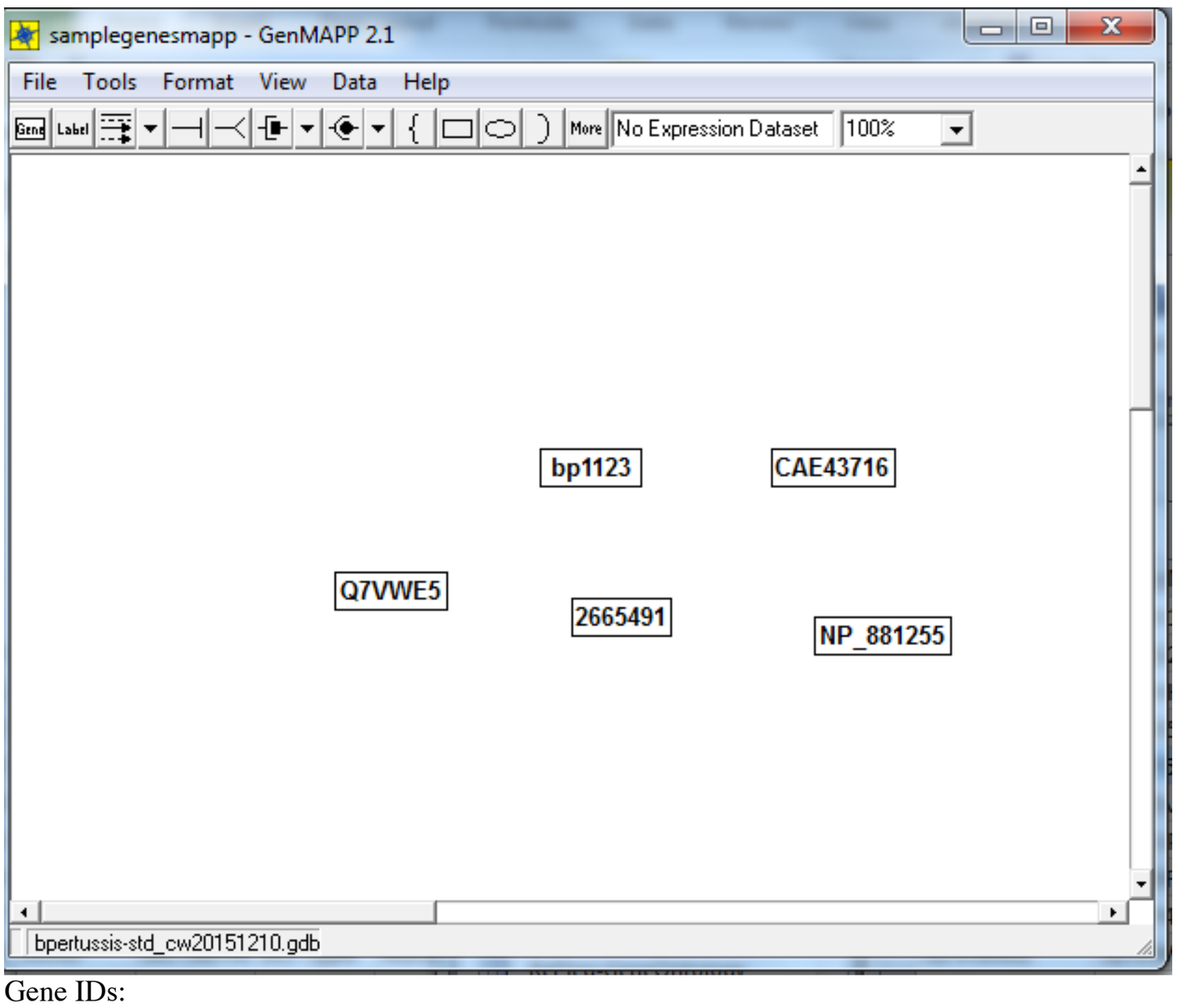
- Systems Table
  - 35 gene ID systems were listed, 11 of which were used in the creation of this .gdb file and listed the appropriate import date (12/10/2015).
    - All gene ID systems relevant to *B. pertussis* were listed. This includes: EMBL, EnsemblBacteria, GeneID, GeneOntology, InterPro, OrderedLocusNames, 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:
1	2	3	4	5	6
[O,P,Q]	[0-9]	[A-Z,0-9]	[A-Z,0-9]	[A-Z,0-9]	[0-9]
- 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:



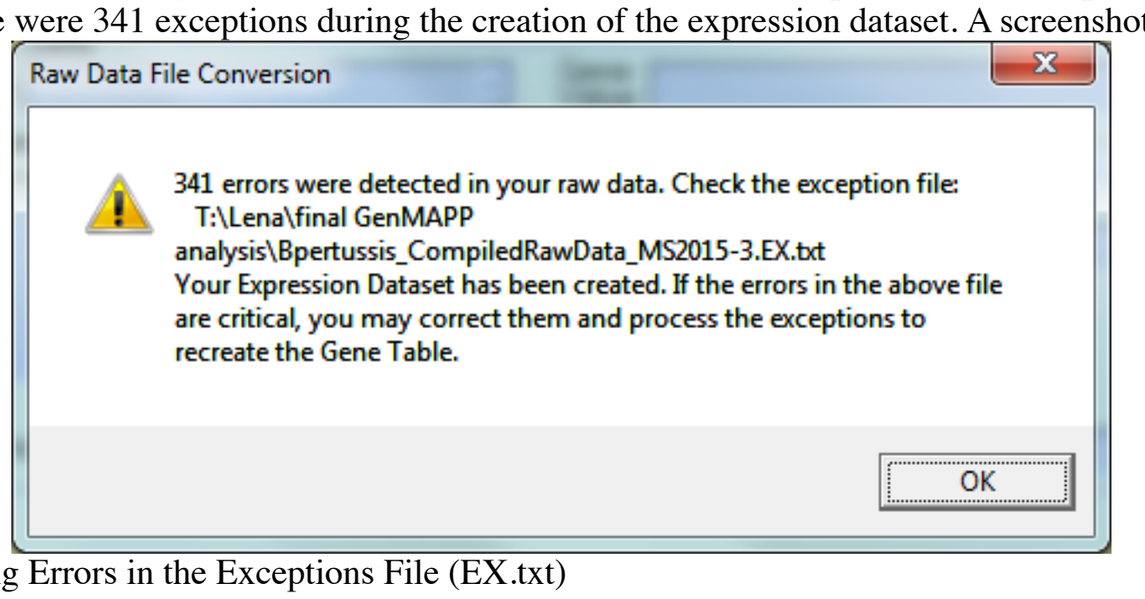
- Gene IDs:
  - **bp1123** refers to the OrderedLocusNames gene ID system.
  - **CAE43716** refers to the EmsemblBacteria gene ID system.
  - **Q7VWE**5 refers to the UniProt gene ID system.

- **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.**

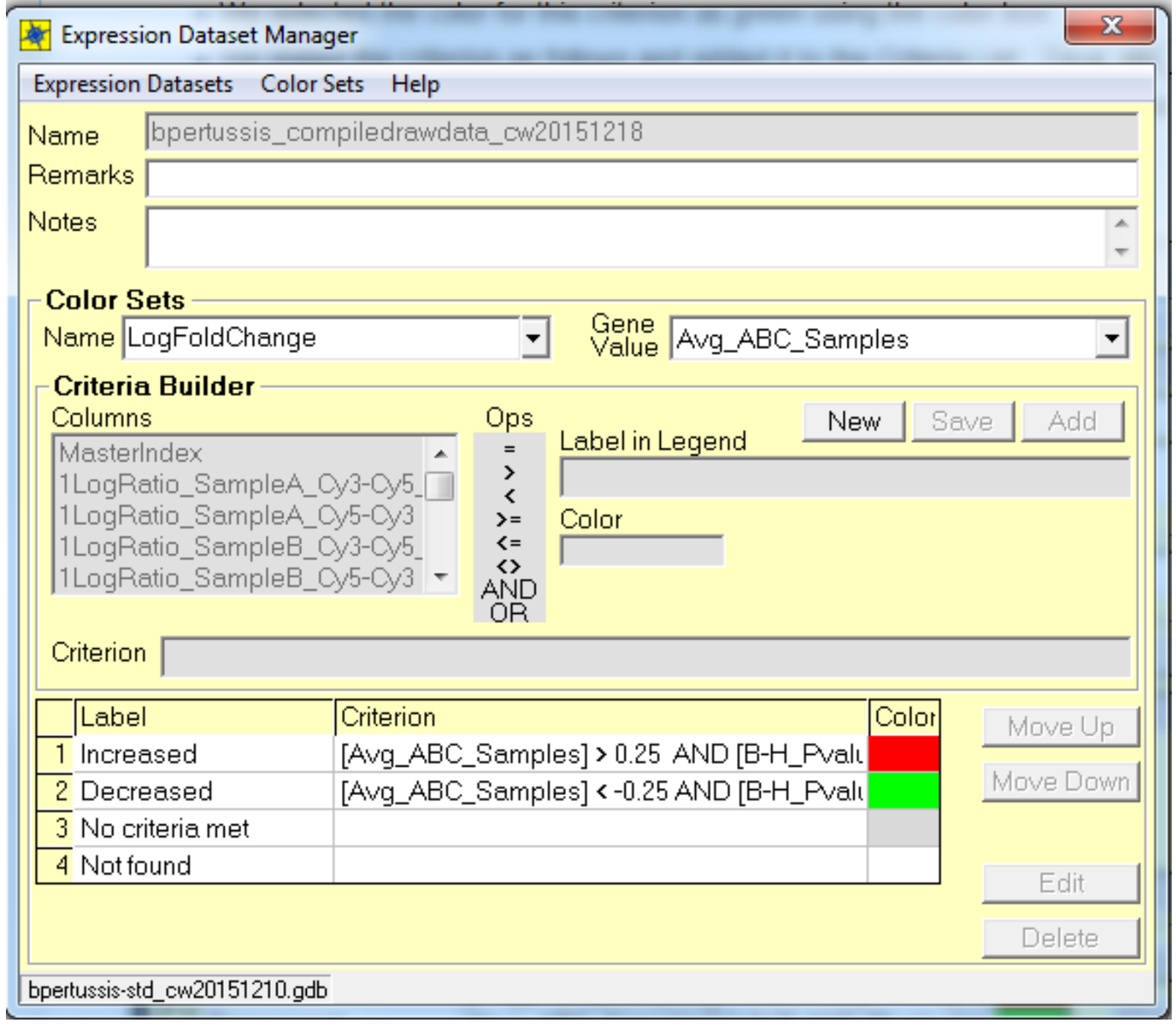
Coloring a MAPP with Expression Data

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:

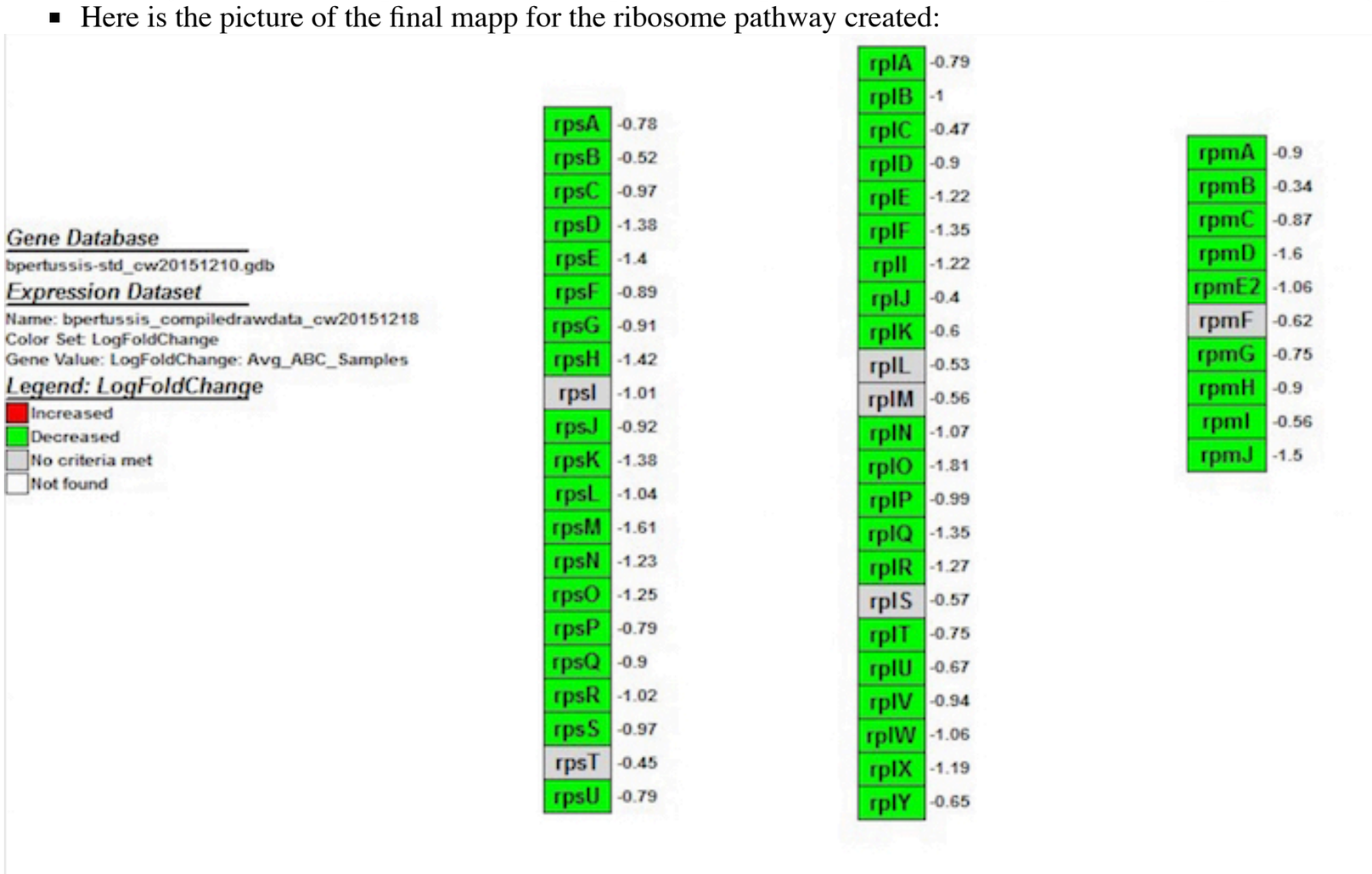


Note: No errors were encountered in the creation of the Color Set.

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:



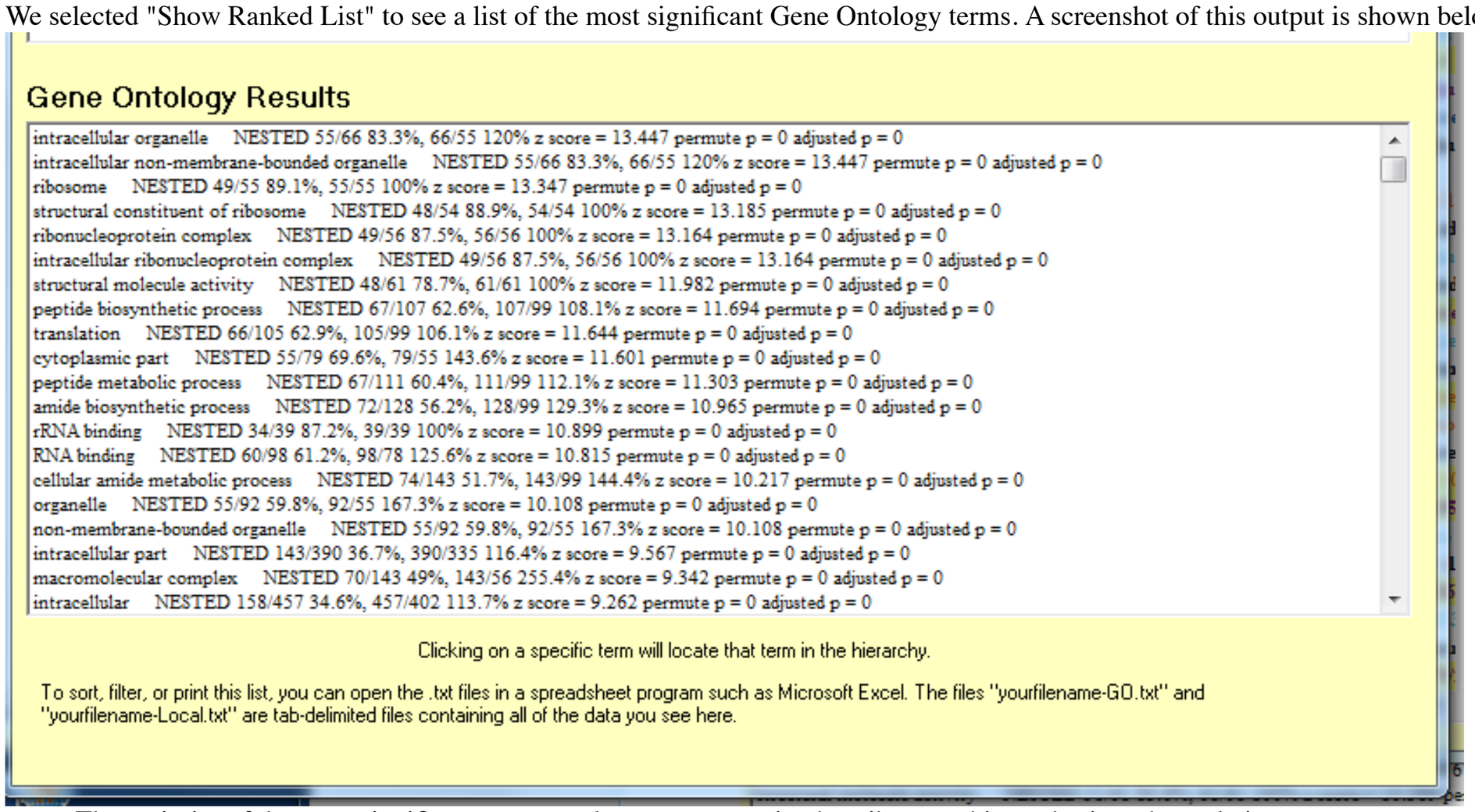
  - Most of the ribosome genes that were generated on this mapp appeared to be the color green, symbolizing a decrease, except for the grey colored genes that were not significantly changed in this experiment. Since the genes mapped for the ribosome pathway all appeared to be green, this means that the 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:
  - 

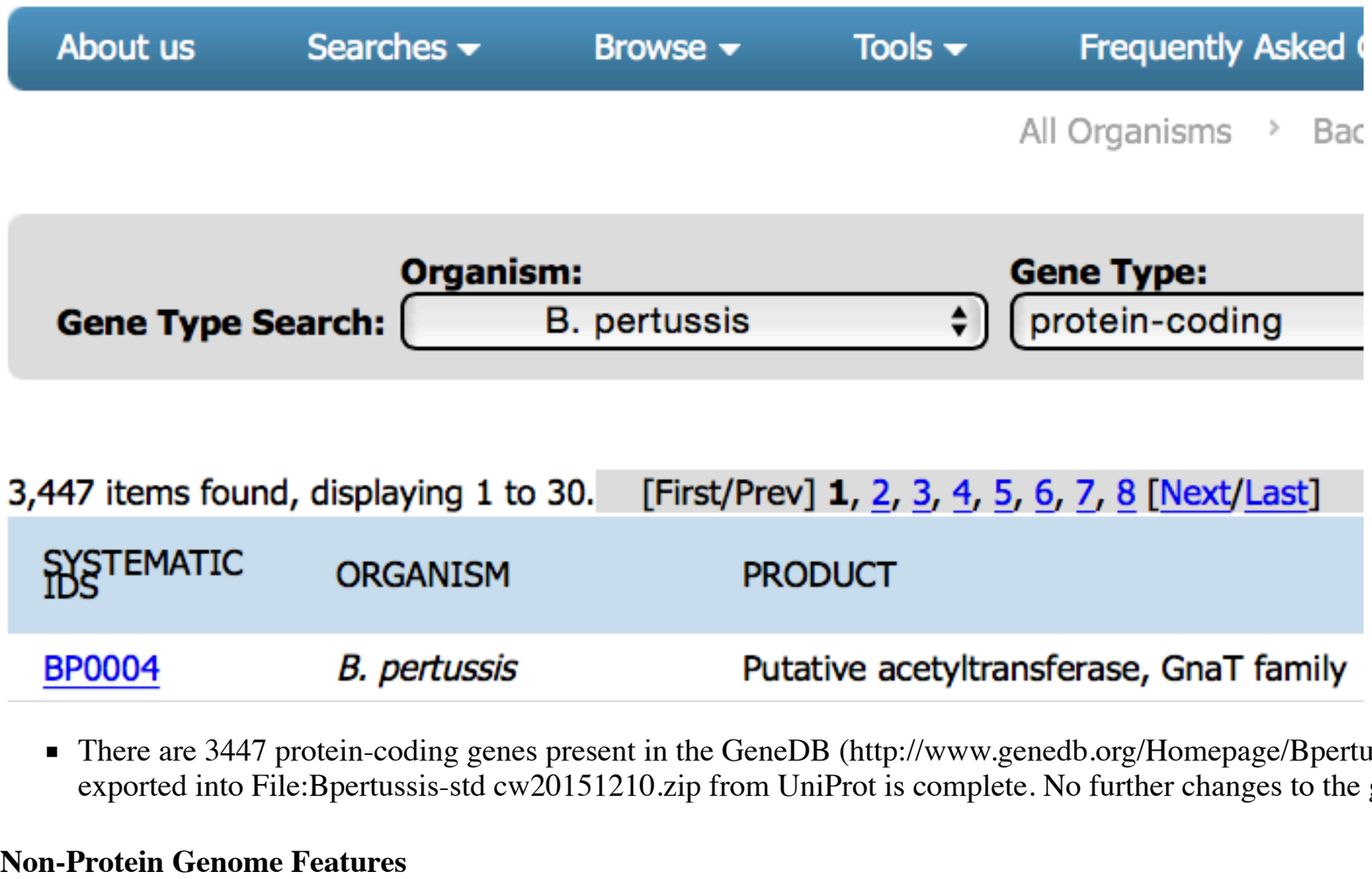
    - The majority of the most significant gene ontology terms pertained to ribosome biosynthesis and translation.

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.

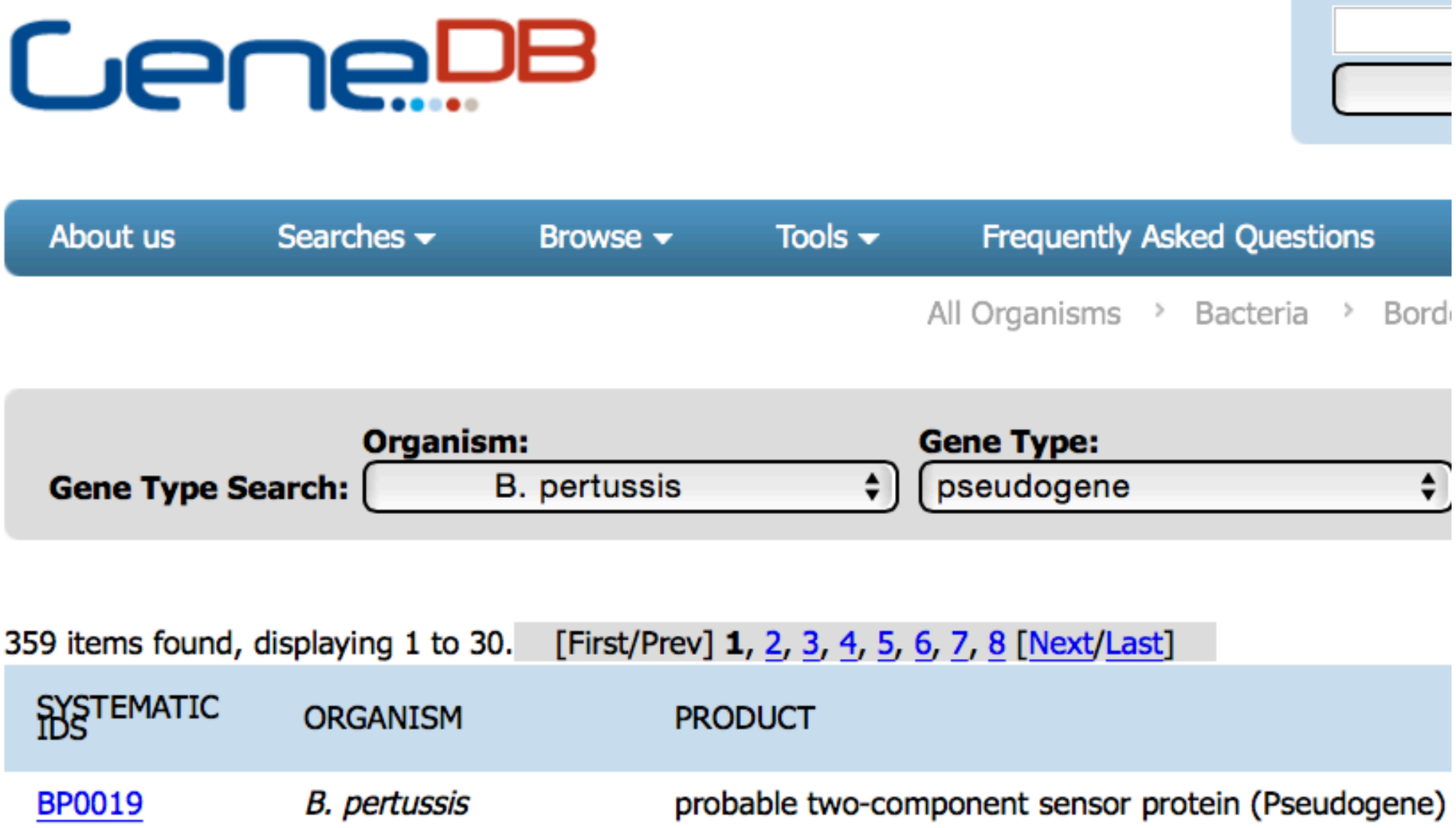
Protein-Coding Genes



- 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.

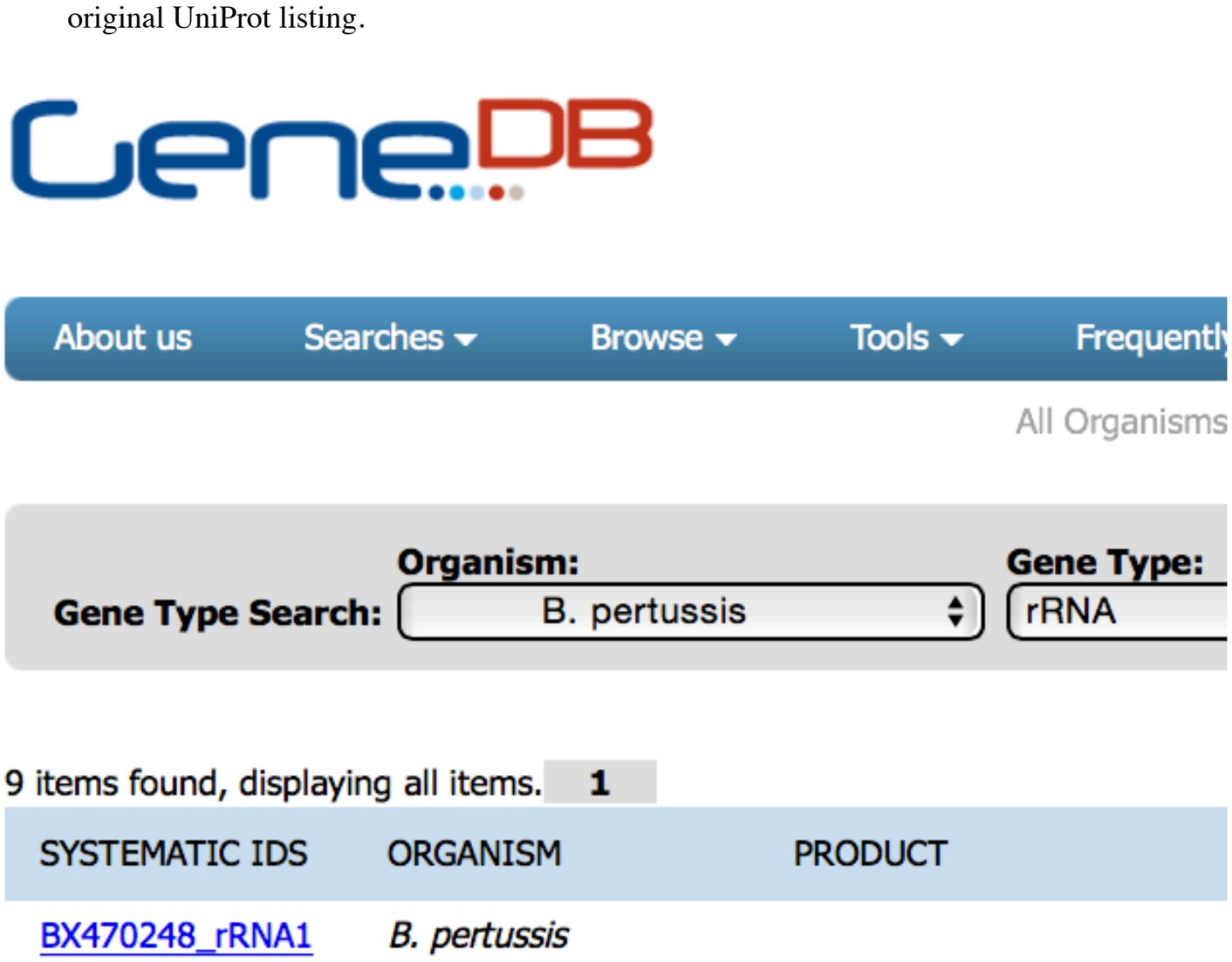
Non-Protein Genome Features

1. Pseudogenes



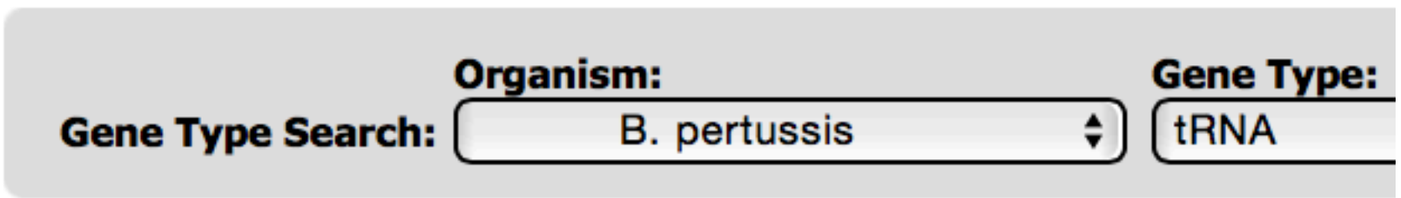
- GeneDB indicated that 359 pseudogenes are present in the *B. pertussis* genome. Pseudogenes do not code for proteins and were therefore not included in the original UniProt listing.

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



51 items found, displaying 1 to 30. [First/Prev] **1**, 2 [Next/Last]

SYSTEMATIC IDS	ORGANISM	PRODUCT
BX470248_tRNA1	*B. pertussis*	

- 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. snRNA
    - GeneDB retrieved 0 genes that encode for snRNA.
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.

Team Information & Links

GenMAPP Analysis of *Bordetella pertussis* Microarray Data

Gene Database Project Links							
Overview	**Deliverables**	Reference Format (https://peerj.com/about/author-instructions/#reference-format)	**Guilds**	Project Manager	GenMAPP User	Quality Assurance	Coder
			Teams	Heavy Metal HaterZ	The Class Whoopers	GÉNialOMICS	Oregon Trail Survivors

Journal Entries

Class Whoopers Individual Journal Entries				
Brandon Klein	**Week 11**	**Week 12**	**Week 14**	**Week 15**
Lena Olufson	**Week 11**	**Week 12**	**Week 14**	**Week 15**
Mahrad Saeedi	**Week 11**	**Week 12**	**Week 14**	**Week 15**
Team Entries	**Week 11**	**Week 12**	**Week 14**	**Week 15**

Group Members

- Project Manager & Coder: Brandon Klein
- Quality Assurance: Mahrad Saeedi
- GenMAPP User: Lena Olufson
- Team Page

Team Weekly Assignments

- Week 10 Creation of page and combined annotated bibliography (midnight 11/10)
- Week 11 (midnight 11/17)
- Week 12 (midnight 11/24)
- Week 14 (midnight 12/8)
- Week 15 (midnight 12/15)

Categories: Group Projects | Class Whoopers