

Analysis of Microarray Data Reveals Biofilm-Grown *B. Cenocepacia* str. J2315 Responds to Tobramycin Treatment by repressing the electron transport chain and protein synthesis - but not the tricarboxylic acid cycle - while inducing assembly of iron-sulfur clusters.

Anindita Varshneya, Kevin W. Wyllie, Brandon O. Litvak, Veronica A. Pacheco

BIOL/CMSI 367 Biological Databases

Loyola Marymount University, Los Angeles

December 18, 2015

Abstract

B. cenocepacia is a highly infectious pathogen among cystic fibrosis (CF) patients. Because this pathogen is difficult to treat due to high levels of antibiotic resistance, *B. cenocepacia* infection is associated with significant decline in lung function and increased levels of mortality among affected CF patients. *B. cenocepacia* belongs to the *B. cepacia* complex (BCC), a group of gram-negative bacteria with high antibacterial resistance properties. Biofilm formation is a major characteristic among highly resistant infections such as those associated with BCC species due to its impermeability to antibiotics. In *B. cenocepacia*, a small population of “persister” cells are found within the biofilm, where these cells neither grow nor die in the presence of antibacterial agents. To investigate the underlying mechanisms of antibiotic tolerance in persister cells in *B. cenocepacia* str. J2315, Van Acker et al (2013) conducted a transcriptomic analysis comparing biofilm persister cells treated with tobramycin versus untreated biofilm cells. As a result of their statistical and biological analysis, they reported downregulation of the tricarboxylic acid (TCA) cycle and the electron transport chain, with upregulation of the glyoxylate shunt.

In hopes of confirming these results, pathway analysis was conducted using GenMAPP and MAPPFinder after creating a comprehensive GenMAPP-compatible Gene Database specifically for *B. cenocepacia* str. J2315 using GenMAPP Builder, a subproject of XMLPipeDB. Our analysis corresponded to Van Acker et al (2013) to the extent of recognizing the downregulation of the electron transport chain. Beyond this, several notable differences were found primarily based on the differences in stringency placed on statistical analysis. Because our analysis was based on more stringent criteria for significance, the data was analyzed more

critically, leading us to new biological conclusions. Our analysis suggested the upregulation of iron-sulfur cluster synthesis and downregulation of protein synthesis, which were not recorded by Van Acker et al (2013). Recognition of these processes logically parallel tobramycin's proposed antimicrobial mechanism which would be compensated by the biological conclusions we determined. These discoveries stand as the first step into the investigation of the antibiotic-tolerance of *B. cenocepacia* persister cells, which may lead to new approaches toward the treatment of BCC infections among cystic fibrosis patients.

Introduction

B. cenocepacia is an opportunistic pathogen which causes lung infections in cystic fibrosis (CF) patients; infection by *B. cenocepacia* is extremely difficult to treat due to a high level of antibiotic resistance, and thus, infection is tied to a significant decline in the functioning of the lung and is associated with increased levels of mortality. *B. cenocepacia* is a clinically relevant part of the *B. cepacia* complex (BCC), which is a group of gram-negative bacteria that typically reside in water or soil with high antibacterial resistance properties.^[4] *B. cenocepacia* strain J2315, is a member of the ET12 lineage subgroup IIIA, a recently emerged (1990s) epidemic lineage of *B. cenocepacia* that is extremely transmissible, particularly among CF patients.^[4] J2315, specifically, is an isolate derived from a CF patient and it exhibits strong levels of antibiotic resistance.^[4] Unlike those associated with the other subgroups, strains associated with subgroup IIIA are rarely encountered in a natural environment, suggesting that these strains have strongly adapted to a host-associated pathogen lifestyle.^[4] There also exist many virulence markers that are encountered more frequently with IIIA strains than with other subgroups.^[4] The

ET12 isolates, additionally, are known to have a cable pilus which permits binding to molecules within the host environment, such as mucins, which are abundant in the lung.^[4] In short, J2315 represents a unique and extremely significant pathogen of CF patients as it possesses properties that allow it to thrive within the lung environment and significantly resist treatment.

Biofilm formation is a widely-recognized, major factor in many recalcitrant infections, such as those of the BCC species.^[1] This is in part due to the insufficient level of antibiotic diffusion through the established biofilm.^[1] However, another more nuanced phenomenon accounts for the tenacity of BCC species. In *B. cenocepacia*, phenotypic variants called “persister” cells, have found in biofilms.^[7] These persister cells, existing as a relatively small proportion of a population, enter a state of metabolic dormancy, which allows them neither to grow nor die during antibiotic treatment.^[7] Because of their inability to grow in the presence of antibiotics, these persister cells should be seen as antibiotic tolerant, rather than resistant.^[5] To investigate the underlying mechanisms of antibiotic tolerance in persister cells in *B. cenocepacia* strain J2315, Van Acker et al (2013) conducted a transcriptomic analysis comparing biofilm persister cells treated with tobramycin versus untreated biofilm cells.^[7] They reported, most notably, downregulation of the tricarboxylic acid (TCA) cycle and the electron transport chain, with upregulation of the glyoxylate shunt.^[7] These expression changes were interpreted as compensatory measures against tobramycin’s proposed antimicrobial mechanism, which hyper-activates the electron transport chain, ultimately resulting in damage to DNA, lipids and proteins due to heightened levels of reactive oxygen species (ROS).^[7]

The previously mentioned study provided numerous data points available for statistical analysis. In order to determine if any other discoveries could be made regarding *B. cenocepacia*

str. J2315's response to tobramycin treatment, this lab elected to run this data through GenMAPP and MAPPFinder after appropriate processing of the data. GenMAPP is a computer application "designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes"^[6]. It incorporates MAPPFinder, another application that performs an analysis of gene expression using the data provided by pathway MAPPs and Gene Ontology Terms.^[6] In order to use these programs, however, a comprehensive database incorporating information from UniProt and Gene Ontology must exist especially for *B. cenocepacia* str. J2315. Due to a lack of access to a database containing this information, other programs such as GenMAPP Builder, a subproject of XMLPipeDB Project, were utilized to create such a database.^[3] XMLPipeDB is a multifaceted project that was originally created to produce gene databases incorporating biological information from multiple sources. GenMAPP Builder was created to address this issue and produce GenMAPP-compatible gene databases including data from the UniProt and Gene Ontology databases.^[3] Using these programs and a new database specifically for *B. cenocepacia* str. J2315, statistical and biological analysis of the data produced by Van Acker et al, 2013 may reveal new information about the cellular process modified within str. J2315 in response to antibacterial products such as tobramycin.

Materials and Methods

Construction of GenMAPP Builder and a GenMAPP-Compatible Gene Database

GenMAPP Builder is able to export a UniProt-centric Gene Database using the data from the UniProt complete proteome set for a species, Gene Ontology data, and Gene Ontology gene associations (Figure 2). The complete UniProt proteome set for *Burkholderia cenocepacia* str.

J2315, from UniProt release 2015_11, was downloaded from UniProtKB in XML format. The Gene Ontology association file (GOA) for J2315 was downloaded from the proteomes directory of the EMBL-EBI UniProt-GOA FTP server; this retrieved file was last updated on November 10, 2015. Gene Ontology terms were downloaded in OBO-XML format from the Gene Ontology Consortium website; the retrieved OBO-XML file was last updated on November 19, 2015. PostgreSQL Version 9.4.5, GenMAPP 2, GenMAPP Builder 3.0.0-build-5, and XMLPipeDB Match 1.1.1 were also downloaded and utilized in this project. GenMAPP 2 and GenMAPP Builder 3.0.0-build-5 were retrieved from the XMLPipeDB releases page on GitHub. XMLPipeDB Match 1.1.1 was retrieved from the XMLPipeDB releases page on SourceForge. PostgreSQL Version 9.4.5 is available through the EnterpriseDB website. All retrieved files were downloaded on November 19, 2015. The gene ID patterns for *B. cenocepacia* str. J2315 were determined by referencing the Burkholderia Genome Database^[9] and UniProtKB. The gene ID patterns for J2315 were found to be BCAL#####, BCAM#####, BCAS#####, and pBCA####; IDs occasionally included a letter, A-Z, at the end (Figure 1).

Several exports of a *B. cenocepacia* str. J2315 GenMAPP Gene Database were conducted. A PostgreSQL database was created for the export and was populated with GenMAPP Builder tables using pgAdmin III and the PSQL query stored in gmbuilder.sql, which is included with GenMAPP Builder. Using GenMAPP Builder, a connection to the PostgreSQL database was established and the UniProt XML, GOA, and OBO-XML data were all imported into the PostgreSQL database. A GenMAPP Gene Database export was initialized after the file import was complete with the purpose of exporting the PostgreSQL database data to a GenMAPP Gene Database (Figure 2). For the initial export, the only addition made to

GenMAPP Builder was a customized species profile that only contained code that connected each gene ID to its corresponding page on the model organism database, Burkholderia.com.^[9]

The exported GenMAPP Gene Database was then inspected for quality. The validity of each exported Gene Database was determined using the TallyEngine utility of GenMAPP builder, XMLPipeDB Match, SQL queries, a visual inspection of the individual tables of the Gene Database, and a comparison of the Gene Database to external resources. TallyEngine was run for the purpose of finding the XML and PostgreSQL database counts for the UniProt and GO data; TallyEngine allows a verification of the data transfer from the UniProt XML to the PostgreSQL database. XMLPipeDB Match is an executable Java program, invoked through the command line, that is able to scan an XML file and provide counts of relevant data; in this project, the program was utilized in order to establish a count of the gene IDs that are present within the UniProt proteome set XML (as a note, Match queries take the form of `java -jar (location-of-jar) "(pattern)" < (name of XML file)`). SQL queries, in pgAdmin III, were executed for the purpose of counting the number of gene IDs that were imported from the XML to the *genenametype* table of the export's corresponding PostgreSQL database. A visual inspection of the exported GenMAPP Gene Database tables, using Microsoft Access, was done to find any issues with the tables. Given that the OrderedLocusNames IDs (gene IDs) of the exported database were sourced from the UniProt XML, the total IDs covered by each export was compared to the information present in UniProtKB and the Burkholderia Genome Database.

The initial export, completed with no additional modifications to the *B. cenocepacia* species profile in GenMAPP Builder (Figure 3), resulted in a successful transfer of the data from

the UniProt XML to the PostgreSQL database, however, TallyEngine indicated that there existed a count of 337 OrderedLocusNames IDs within the UniProt XML and within the Postgres database; SQL queries utilizing the PostgreSQL database for this export confirmed what was found in TallyEngine. A visual inspection of the exported database showed that all necessary tables were present and it confirmed the findings of TallyEngine and PostgreSQL by indicating the presence of 337 rows (or individual IDs) within the OrderedLocusNames table; the OriginalRowCounts table of the exported database was referenced to confirm these counts and the number of rows within each Gene Database table. XMLPipeDB Match was then utilized with a regex that was designed to catch all of the possible gene ID patterns ("p?BCA[LMS]?[0-9][0-9][0-9][Aa]?[0-9]?[A-Z,a-z]?"). XMLPipeDB Match resulted in a count of 7126 matches which significantly differs from the number of IDs that were previously found with other resources. Burkholderia Genome Database, and UniProt, were then referenced and it was found that 6994 UniProt entries and 7114 CDSs are associated with J2315; these external resource counts are much closer to what was found with XMLPipeDB Match than what was found with TallyEngine, PostgreSQL, and the OriginalRowCounts table of the exported Gene Database. SQL queries were also utilized to observe the captured OrderedLocusNames IDs; the data represented by the *genenametype* table of the PostgreSQL database, where the type is "ordered locus", was selected with a query and it was found that the 337 IDs were in the form of BceJ2315_#####. This ID pattern deviated from the patterns observed in Burkholderia Genome Database and in the majority of UniProtKB entries. At this point, it was suspected that the utilized version of GenMAPP Builder was not capturing the majority of the gene ID data stored within the UniProt proteome set for J2315.

A second build of GenMAPP Builder was created which incorporated a customized species profile for *B. cenocepacia* str. J2315. This build did not involve any significant changes to the actual functioning of GenMAPP Builder/TallyEngine and it led to gene ID counts that were identical to what was found with the initial export Gene Database. TallyEngine, PostgreSQL, and the OriginalRowCounts table of the second build Gene Database all led to an indication that 337 OrderedLocusNames IDs were present. Given these results with the second export, the UniProt proteome set XML was opened with an XML editor in order to explore the types of data, and data entries, that resided there. Through the observation of several entries within the XML file, it was noticed that two distinct types of gene ID data were present in numerous entries: “ORF” type gene names and “ordered locus” type gene names. It was noticed that all of the gene names of the “ordered locus” type were in the form of BceJ2315_##### while all of the “ORF” type gene names were in the accustomed form that was present in Burkholderia Genome Database and in other resources (BCAL#####, BCAM#####, BCAS#####, and pBCA###, with each pattern occasionally followed by a letter, A-Z).

Checking the postgresQL database for this second export, it was noticed that the *genenametype* table actually contains “ORF” type data, in addition to “ordered locus” type data. A PSQL query was executed (utilizing a regex that included all of the known gene ID patterns) for the purpose of finding the number of ORF gene IDs present in *genenametype*; it was found that 7121 unique gene IDs were present in the postgresQL database that were in the format represented by the ORF type gene IDs. By searching through the “ordered locus” type IDs present in the XML, and by referencing UniProtKB, it was noticed that every entry with an “ordered locus” name was accompanied by an “ORF” gene name; it was also noticed that a

minority of protein entries contained associated “ordered locus” gene IDs (337 entries). Given that the “ORF” type gene IDs were present in all UniProt protein entries (and the “ordered locus” type was not), it was determined that the next developed build of GenMAPP Builder should capture the “ORF” gene name data and ignore the “ordered locus” type gene names.

A GenMAPP Gene Database export was conducted using a third build of GenMAPP Builder that incorporated modifications that allowed it to utilize the ORF type data as the OrderedLocusNames IDs that are present in the exported database. Referencing the OriginalRowCounts in the Gene Database that was exported with this build, it was found that a count of 7121 IDs were present in the OrderedLocusNames table of this Gene Database; visually observing this table via Microsoft Access it was noticed that all IDs were in the usual form that was represented by the ORF type gene names. However, it was noticed with this export that TallyEngine was not functioning properly; it appeared to exclude the counts for the “ORF” type gene IDs but it retained the counts for the “ordered locus” type genes names. Through this observation, it was noticed that there existed an error in the current version of GenMAPP Builder that prevented TallyEngine from functioning properly.

The last GenMAPP Gene Database export incorporated fixes to the TallyEngine utility that led to a TallyEngine output that properly provided the counts of the “ORF” gene IDs that are present in the UniProt XML file and the postgresSQL database. Validity testing, through an observation of the Gene Database tables, found that the data associated with this exported database corresponded with what was found with third export that utilized Build 3 of a customized GenMAPP Builder. GenMAPP Builder then underwent several final customizations to allow all data to pass through as expected. These included changes to allow the code to collect

gene IDs from the ORF tag within the XML file instead of the OrderedLocusNames tag and modify the presentation of TallyEngine to include counts from the gene IDs associated with the ORF tag. This final export resulted in a Gene Database that represented 7121 OrderedLocusNames IDs for J2315, however, this count differed slightly from what was previously found by XMLPipeDB Match (Table 1). The XMLPipeDB Match count indicates the existence of 5 discrepant ID matches; these ID matches were investigated using an application of the Excel MATCH function. The SQL query `select * from genenametype where type = 'ORF' order by value` was executed in order to export the ORF gene IDs that are present in the postgresQL database. The previously mentioned XMLPipeDB Match command was also utilized, with the addition of `> MATCHIDS_GEN_BL14_20151203` to the command, in order to export the Match results into a text file. These two sets of IDs were then imported into Excel and analyzed in order to find instances of non-matching IDs. The 5 discrepant IDs were identified as bca199f, bca5253f, bca636c, bcal0235a, and bcal0239a. By referencing the UniProt XML via an XML editor, it was found that bca199f, bca5253f, and bca636c were accidental matches that happened to correspond to the utilized regex. The last two discrepant IDs, bcal0235a and bcal0239a, were found to be database reference IDs to STRING, a protein-interactions database. After making these changes to GenMAPP Builder, a gene database was created, and pathway analysis of microarray data was ready to be done.

Statistical Analysis of Van Acker et al. DNA Microarray Data

Before compiling the data, it is imperative to understand the experimental design of Van Acker et al (2013) to organize the data efficiently. Figure 4 outlines the procedure for the microarray experiment. Under conditions of 37 degrees Celsius for a period of 24 hours,

Burkholderia cenocepacia cells were grown in 96-well plates. The supernatant was removed and treated with tobramycin. Persister cells were harvested in 37 degrees Celsius for a 24 hour period. The RNA was extracted and purified. Next, the cDNA was fluorescently tagged red and was hybridized with genomic DNA, which was fluorescently tagged green. The normalization of the data involved using two-color microarrays and a T-test was performed. The resulting data is separated into text (.txt) files. The five untreated biofilms each had their individual corresponding text file and the three tobramycin-treated biofilms also had their own respective text files of data. There was a total of eight files of raw data to compile.

The raw data that corresponds with the microarray experiment was downloaded from the Array Express^[8] web page . The data for GeneName and LogRatio is copied over to Excel and is sectioned by technical replicates, with 4 technical replicates in total. Each technical replicate section contains the biological replicates for the (untreated) biofilm and tobramycin-treated biofilm samples. The average of each biological replicate within the technical replicates were taken and put into another worksheet. The averages for the untreated biofilms and the tobramycin treated biofilms were kept separate for analytical purposes. The standard deviations for each biological replicate were also added to this worksheet. Both the averages and the standard deviations were calculated using the functions for each respective formula in Excel. These two columns of information are used to scale and center the data. This takes the first log ratio for sample Biofilm_1, subtracts the average log ratio, and divides by the standard deviation of the log ratios. This function was then pasted for the remaining data, using the drag feature.

Once the data has been scaled and centered, the average across the five untreated samples were calculated. The same was done using the three treated samples. Therefore, these two

columns named, the AVG_Biofilm_scaled_centered and AVG_Tobramycin_scaled_centered, contained the information to proceed with statistical analysis. The following column is designated for the ratio between the biofilm average and the tobramycin. The reference sample is genomic DNA which means that the ratio of the averages for the biofilm and tobramycin samples is needed to get the ratio of tobramycin to control (tobramycin over biofilm). Subtracting the biofilm average from the tobramycin average works because the numbers are in log space. Next, the P-value's for each fold change ratio was computed by performing a type-3, 2-tailed T-test through the TTest function in Excel.

This P-value is used to generate two other P-values, the Bonferroni and the Benjamini and Hochberg (BH) P-values. For the Bonferroni P-value, each previous P-value was multiplied by 7,251 (the number of genes on the sheet). BH P-values require a couple more steps. Starting from the initial P-values, the genes (rows) were ordered by ascending P-value. After, they were ranked in by adding values 1 and 2 to the adjacent column and right-clicking on the black square at the bottom-right of the highlighted region. This will automatically fill in the rank (an ascending count). Then, each unadjusted P-value was multiplied by its own rank to get the BH P-value.

With all these different calculations, there needs to be a specific format on the Excel worksheet in order for GenMAPP to read the file. A new sheet was made and it was named "forGenMAPP". The rows for this worksheet are organized as follows: gene ID's, consolidated fold change values, averages for treated and untreated samples, fold change ratios between treated and untreated samples, and corresponding P-values. These were followed by the calculated Bonferroni and BH P-values respectively. In addition, all fold change values in the

"forGenMAPP" sheet were formatted to include two decimals, while P-values were formatted to include four. The worksheet has now been finalized and is ready to be loaded into GenMAPP.

Pathway Analysis using GenMAPP and MAPPFinder

The preliminary step is to load a gene database into GenMAPP. By choosing the .gdb file, it signifies to GenMAPP that this data will be analyzed using the expression dataset. To create a new expression dataset, the forGenMAPP(.txt) file, which is made in Excel with all the compiled, normalized and statistically analyzed data, is selected in the file dialog box that appears in the Expression Datasets menu. Initially, a message will appear to notify that lines in the text files that had errors were not added into the expression dataset. Therefore, these errors are consolidated on an exception file (.EX.txt). There were 284 errors within the exception file when the *B.cenocepacia* for GenMAPP text file was run in GenMAPP. In order to customize the Expression Dataset that was just created, Color Sets are applied to communicate to GenMAPP the instructions for displaying that data via MAPPs. The Color Sets are set through given criterion that correspond to a chosen color. The color red was assigned to the "Increased" criteria and the parameters are set as $[\text{Biofilm_Tobramycin_Ratio}] > 0.25$ AND $[\text{BH_Pvalue}] < 0.05$. Thus, when there is a gene whose BH pvalue lies within these parameters, the map would color the gene with the color red. For "Decreased" criteria, the color green was designated as the marker with the parameters of $[\text{Biofilm_Tobramycin_Ratio}] < -0.25$ AND $[\text{BH_Pvalue}] < 0.05$. The criterion is used as the color set for when the MAPP is generated.

After all the necessary criterion are added to the Expression Dataset, the next step is to proceed to run MAPPFinder. Running MAPPFinder took several minutes and once the results were calculated, a Gene Ontology tree appeared. The tree contained the Gene Ontology(GO)

terms that are considered significant. Significant results are defined by a p value less than 0.05. A text file with the results is generated using the naming system XXX-CriterionX-GO where the XXX is the file name and the CriterionX is the selected criterion. Both increased and decreased criteria are important for analysis thus two separate text files were produced.

Selecting a GO term would ideally open a MAPP with the list of genes that are associated with that particular term. The genes are symbolized by a box with a gene name based on the UniProt identification system. However, MAPPFinder was not responding when the GO terms were selected thus a MAPP was not automatically generated. As a result, the MAPP (Figure 5) was created manually.

First, the GO text files are opened through Excel. Opening the files through excel enables filters to be added in order to narrow down the list of terms to about 20 terms. For both the Increased and Decreased GO files, the following filters were added to the designated columns: the Z score filtered by 'greater than 2', PermuteP filtered by 'less than 0.05, Number Changed filtered by 'greater than or equal to 4 AND less than 100 and Percent Change filtered by 'greater than or equal to 25'. Inspecting the filter lists, the pathway is selected for mapping in GenMAPP.

In Van Acker et al (2013), the main focus for downregulation was the TCA cycle and briefly mentions the downregulation of the electron transport chain. However, there were no terms in the filter GO list that indicated TCA downregulation. Oxidative phosphorylation was a GO term that was listed in the Decreased file and was selected for mapping based on its association with the electron transport chain. For a polished format of the MAPP, the KEGG web page^[10] was used as a template. KEGG contains the specific gene information for the pathway for oxidative phosphorylation in *B.cenocepacia str. J2315*. The genes in KEGG were organized by

sections of where the gene are present specifically on the pathway. This format was mirrored in GenMAPP.

Empty gene boxes were placed on the GenMAPP window according to section. The empty gene boxes were then edited to have their gene ID to confirm that the gene was found in the gene database. Once confirmed, the gene ID is substituted for the gene name from UniProt. Once all the genes in the sections have been checked and renamed with the gene name, they can be color-coded with the gene expression data from the microarray experiment.

For example, under the section NADH dehydrogenase, an empty gene box will be labeled BCAL2341. The gene ID will be checked in the gene database and there will be a message notifying that the gene was found. The ID can then be changed to the gene name NuoD. The color set can then be applied in order to see whether the gene was increased, colored red or decreased, colored green. If the gene did not meet either of these criteria, the gene box will be colored grey. To the right of the colored gene box, there is a number that indicates the BH P-value for the gene. The colored MAPP gives a useful visual of the data in which the genes are readily compared with one another in terms of upregulation or downregulation and BH P-value.

Results

According to the third export of the GenMAPP Gene Database and the build 3 of GenMAPP Builder, it was found that two major problems existed within the database. The database was collecting gene IDs from a tag that did not contain all of the genes associated with *B. cenocepacia*, and did not contain the gene IDs that follow the general pattern of genes associated with the *B. cenocepacia* genome. In order to fix this issue, it was found that gene IDs

needed to be collected from the ORF tag as opposed to the OrderedLocusNames tag, and so code would need to be added to the *B. cenocepacia* customized species profile to reflect this change. Because this gene tagging issue was previously addressed for a different species, code from the *Leishmania major* customized species profile that specifically addressed collected gene IDs from the ORF tag was added to the customized species profile of *B. cenocepacia* (Figure 6). This change, however, required additional changes to Tally Engine to ensure that all desired genes are caught and represented within the final table within the database that collected all of the gene IDs.

In order to make this fix, the Tally Engine table was modified to present all data found with the ORF tag under the “ORF” data table. This modification requires additional customization as the final database does not contain an ORF table, but instead includes a Ordered Locus table that acts as a comprehensive collection of all relevant gene IDs for any given species. Currently, TallyEngine is reporting the Ordered Locus datatable to contain 337 genes, though these 337 genes are genes identified with the OrderedLocusNames tag and are not represented in the comprehensive list of gene IDs associated with *B. cenocepacia* (Figure 7). The gene IDs of importance are represented under the “ORF” datatable, containing 7121 genes. A change will need to be made to remove the current “Ordered Locus” datatable count within the TallyEngine that is reporting gene IDs tagged with OrderedLocusNames and replace that count with the count of gene IDs tagged with ORF. Then, the “ORF” datatable and its corresponding gene count can be completely removed from within TallyEngine as it provides misleading information considering there is no ORF datatable created using our database. Because these

changes just require modifications to data labeling within TallyEngine, they in no way affect the quality of the gene database itself.

A GenMAPP Gene Database for *B. cenocepacia* str. J2315 was exported, utilizing a modified build of GenMAPP Builder, that covers 7121 valid gene IDs. The tables and data that make up this Gene Database are primarily sourced from UniProt; the exported database also reflects data that was acquired from the Gene Ontology Project (Figure 8). The TallyEngine results associated with the final export indicate the presence of 7121 “ORF” gene IDs and 337 “ordered locus” gene IDs within the UniProt XML and in the related postgresSQL database; given that these counts are consistent between the XML and database column, the final modified build of GenMAPP Builder appears to have successfully imported the data from the XML to postgresSQL (Figure 7). The final count of 7121 valid gene IDs is reflected by the TallyEngine utility and by the OriginalRowCounts table of the Gene Database, however, the XMLPipeDB Match utility resulted in a count of 7126 unique matches (Table 1). The XMLPipeDB counts were found using this command, via the windows command line, after entering the directory that housed the UniProt XML data:

```
java -jar xmlpipedb-match-1.1.1.jar  
"p?BCA[LMS]?[0-9][0-9][0-9][Aa]?[0-9]?[A-Z,a-z]?" <  
"uniprot-taxonomy%3A216591_GEN_BL12_20151119.xml"
```

The postgresSQL count of 7121 was verified through the execution of this SQL query with the postgresSQL database associated with the final export:

```
select count(*) from genenametype where type = 'ORF' and  
value ~ 'p?BCA[LMS]?[0-9][0-9][0-9][Aa]?[0-9]?[A-Z,a-z]?';
```

This XMLPipeDB Match command and PSQL query were utilized, in addition to TallyEngine and an observation of the OriginalRowCounts Gene Database table, in every Gene Database export in order to verify the counts that are present within the UniProt XML, postgresQL, and the exported Gene database.

Table 2 shows the number of genes deemed significant at P-value thresholds of varying stringency. Using a significance criteria of BH-adjusted P-value < 0.05 a total of 605 genes were significantly changed. These genes make up 8.3% of the 7251 genes in the original microarray data. However, we added another significance criteria in GenMapp: $\log(\text{fold change}) > 0.25$ or < -0.25 . By this criteria, 274 genes (3.8%) were significantly upregulated and 300 (4.1%) were significantly downregulated.

Table 3 displays some of the top gene ontology terms, generated by MAPPFinder. Amongst the upregulated (or “increased”) pathways are two which involve iron-sulfur clusters: “iron-sulfur cluster assembly” and “anion transmembrane-transporting ATPase activity”. These are likely due to tobramycin’s destructive effects on the iron-sulfur clusters embedded in the protein complexes making up the electron transport chain. “DNA integration” may have occurred as a GO term due to the harmful effects of ROS on DNA. Similarly, “outer membrane-bounded periplasmic space” may be explained by the damage of ROS on lipids, which compose the cell membranes. “Response to abiotic stimulus” is quite broad, though tobramycin can of course be considered an abiotic stimulus. After taking a closer look at this GO term on the Gene Ontology Consortium’s website^[2], we speculated that it could include some or many of the transcription factors responsible for other expression changes seen in this data. However this is, again, solely speculation. “Serine-type carboxypeptidase activity” appears to

refer to an enzyme which can hydrolyze peptide bonds (again after investigation on GOC's website), which could plausibly be advantageous to a metabolically dormant cell, as these peptide bonds store a considerable amount of chemical energy. "Methionine biosynthetic process" is more difficult to interpret. Other than being the first amino in any given protein, methionine's other unique property is its formation of disulfide bonds, which may somehow be of use to a persister cell under antibiotic stress.

The gene ontology terms returned for downregulation share more uniformity in overall function. "Intracellular non-membrane-bounded organelle," "rRNA binding," "translation" and "ribosomal subunit," all relate to protein synthesis, which fits into the context of a metabolically dormant cell, as protein synthesis is quite energetically costly. "Branched-chain amino acid biosynthetic process" may relate to the resulting lower demand for amino acids. "Oxidative phosphorylation," "quinone binding," and "oxidoreductase activity," all relate to the electron transport chain, which a persister cell would certainly benefit from downregulating in the presence of tobramycin due to its mechanism, which accelerates the oxidation of NADH.

The MAPP of genes involved in the electron transport chain, shown in Figure 5, validates the suggestion made by the gene ontology results that the electron transport chain, as a pathway, is downregulated in response to tobramycin. Genes shown in green are those that were downregulated in response to tobramycin, while those in grey were unchanged. Upregulated genes would be shown in red, though none were present.

Discussion

The final customized build of GenMAPP Builder successfully captured all of the relevant gene ID data present in the UniProt XML and had exported it, without issue, to a GenMAPP Gene Database; this final database covers 7121 OrderedLocusNames IDs that were previously identified as being tagged as type “ORF” within the UniProt XML file. As was previously discussed, the XMLPipeDB Match command resulted in an output that included 5 more IDs than were present in the final GenMAPP Gene Database (Table 1). The 5 discrepant IDs were identified and, since they did not refer to a UniProt protein entry, they were found to be irrelevant to the final Gene Database. The model organism database for *B. cenocepacia*, Burkholderia Genome Database, also reports that 7114 gene IDs are present for protein coding sequence DNA. It is believed that this difference of 7 IDs is due to the fact that Burkholderia Genome Database^[9] is manually curated, and thus, is potentially missing some gene IDs that are covered by UniProt.

Referencing the Burkholderia Genome Database^[9], it was noticed that 7384 annotated genes are present with unique gene IDs, within the genome of *B. cenocepacia* str. J2315; this indicates that 263 gene IDs are not represented by the final GenMAPP Gene Database. The missing gene IDs were investigated within the Burkholderia Genome Database and it was found that these 263 genes are either pseudogenes or functional RNA encoding genes. Given that these missing genes are not protein-encoding, they would not be represented in the UniProt XML/UniProtKB and, as a result, would not be represented in an exported GenMAPP Gene Database. The processed microarray data yielded 284 exceptions when imported into GenMAPP while utilizing the final GenMAPP Gene Database; this indicates that 284 genes, present in the microarray data, were not present in the final GenMAPP Gene Database. The exceptions were

investigated and it was found that the vast majority were genes that were not present in UniProt; these genes that lacked a UniProt entry were present in Burkholderia Genome Database as a pseudo or functional RNA encoding gene. Some of the exceptions were due to the fact that some gene IDs were slightly modified from their original form. It was concluded that none of the exceptions involved a fault with the final export GenMAPP Gene Database for *B. cenocepacia* str. J2315.

The findings associated with the last Gene Database export suggest that the final GenMAPP Gene Database for *B. cenocepacia* str. J2315 accounts for all protein-associated gene IDs covered by UniProt (Appendix); we can report with a high degree of certainty that all of the protein-associated genes covered by Van Acker et al. in their microarray paper are represented in our final GenMAPP Gene Database.

Our statistical practices, and consequently our results, differ considerably from those of Van Acker et al. First and foremost, the previous authors chose to use a rather lenient significance criteria of (unadjusted) P-value < 0.05, yielding 4318 significantly changed genes (using the log-fold-changes adjusted by our own normalization protocol). This translates to an alarming 59.6% of the total genes. Though differences in normalization protocol could account for slight differences in the amount of significantly changed genes for a given significance criteria, Van Acker et al did report similar numbers in their manuscript. Moreover, they do not state a significance threshold for the log-fold-changes. This level of leniency in significance criteria interferes with biological analysis. When the number of significantly differentially expressed genes is nearly 60%, any given gene is more likely to see an expression change than not, so it becomes difficult to come to conclusions about what pathways or processes the cell is

actually inducing or repressing. See Table 1 for a range of significance criteria and their corresponding numbers of genes with significant differential expression.

We chose to use a significance criteria of BH P-value > 0.05 , and, $|\log(\text{fold change})| > 0.25$. By this criteria, a total of 574 (8.0%) genes were significantly upregulated (274 - 3.8%) or downregulated (300 - 4.1%). With approximately 52% fewer genes being assigned a significant change in expression than in the analysis of Van Acker et al, our results, though sharing some overlap, differ as well.

Despite other experiments done by Van Acker et al, which examined tobramycin-tolerance of persister cells with deleted or inhibited enzymes of interest - and despite a sound biological explanation for the results seen therein - absent entirely from our filtered (most significant) MAPPFinder results are gene ontology terms which directly or obviously relate to either the TCA cycle or the glyoxylate shunt (Table 3), the two most emphasized pathways in the Van Acker et al manuscript. The suggestion that these pathways are not in fact differentially expressed is mirrored by log-fold-change and significance data for the individual genes (Table 4). While our analysis found one gene in the glyoxylate shunt (BCAL2122) and three in the TCA cycle (BCAM0968, BCAM0969 and BCAM0970) to be differentially expressed, this differential expression is not strikingly overrepresented within these groups of genes, in light of our data's overall significance rate of 8.0%.

Table 4 also illustrates that, though the directions of regulation (induction or repression) seen in our data are in agreement with those of the previous manuscript, there is a considerable discrepancy between the magnitudes of our log-fold-changes and those reported by Van Acker et al. However, because the previous manuscript does not include any information on how the

microarray data was processed, we can only speculate that the root of these stark discrepancies was our scaling and centering protocol.

We also saw upregulation of iron-sulfur cluster processes. This logically follows from tobramycin's proposed antimicrobial mechanism which does include destruction of structures; it's plausible that the cell would benefit from assembling more as a compensatory measure. Similarly, downregulation of protein-synthesis was seen, which fits into the context of a metabolically dormant persister cell, as translation is a highly energetically-costly process. Both of these pathways, aside from a brief mention of rRNA downregulation, are absent from the previous manuscript.

Finally, our findings did validate those of Van Acker et al in terms of electron transport chain downregulation, which act as the cell's primary counteractive measure against ROS production. Figure 5 shows a MAPP of the pertinent genes in this pathway.

Conclusion

The findings in our analysis do correspond to those of Van Acker et al, to a degree. Van Acker *et al* (2013) reported downregulation of the electron transport chain, which was confirmed by our MAPPFinder analysis as well as a MAPP of pertinent genes.

The notable differences involve the percent of genes that were found significantly changed. Our analysis shows about 8% of the total genes to be significantly changed while Van Acker *et al* (2013) reported about 60% of the total genes. The key to this difference is our choice of a more stringent criteria. As a result, when the criteria was applied to the data, the top GO terms produced highlighted pathways tailored for this criteria. For example, using the unadjusted

P-value, Van Acker *et al* (2013) focused on the downregulation of the TCA cycle and reported many significantly changed genes . However, our analysis, as illustrated by Tables 2 and 3, did not confirm the downregulation of the TCA cycle. The methodological choice of choosing a strict significance criteria enables us to analyze the data with a more critical eye, leading to new biological conclusions. Namely, our analysis suggested the upregulation of iron-sulfur cluster synthesis and downregulation of protein synthesis, which were not observed, or at least not emphasized by Van Acker *et al* (2013).

Of course, additional areas of interest within this topic may include potential differences in the cellular response to other types of antibiotics. Another important measure in the investigation of persister cells might be to examine the expression profiles of BCC species collected from actual cystic fibrosis patients (if it is possible to collect such samples), as findings from *in vitro* experiments may not always apply to the living system of interest. Further investigation into the antibiotic-tolerance of *B. cenocepacia* persister cells may empower new approaches toward the treatment of chronic infections in cystic fibrosis patients.

Acknowledgments

We'd like to acknowledge Dr. Kam D. Dahlquist and Dr. John David N. Dionisio, faculty of Loyola Marymount University, College of Science and Engineering, for their help and support throughout this project. We'd also like to acknowledge the students of Biological Databases (BIOL/CMSI 367) for their advice and contributions to the construction and analysis of this project.

Tables and Figures

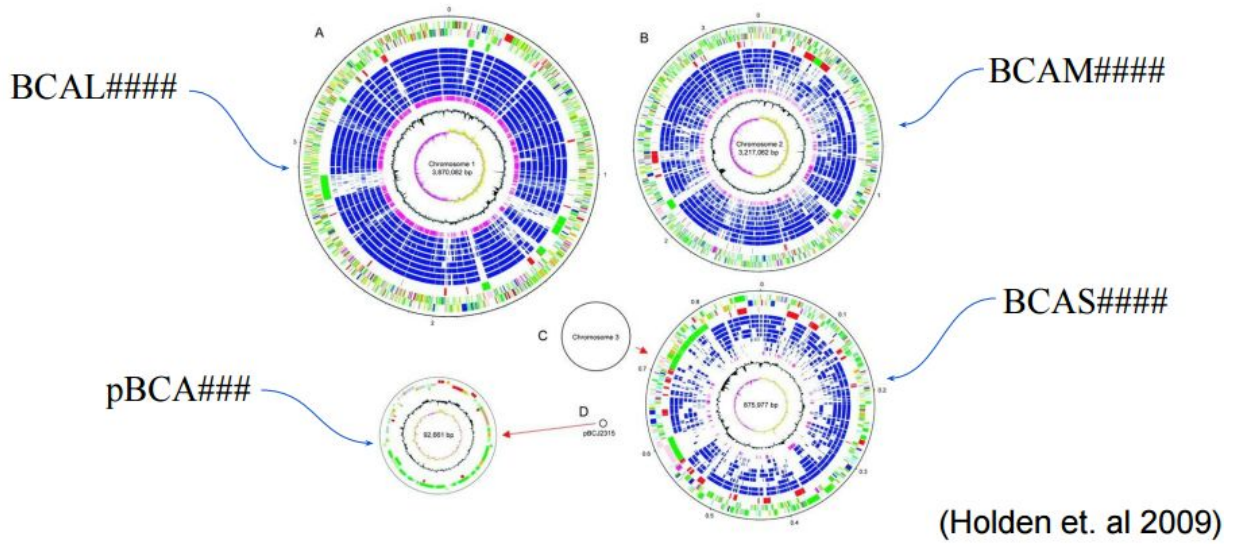


Figure 1. The three chromosomes and one plasmid that make up the J2315 genome, and the common gene ID patterns that are associated with them. This image is a modified version of that which appears as Figure 1 in Holden et al (2009).^[4]

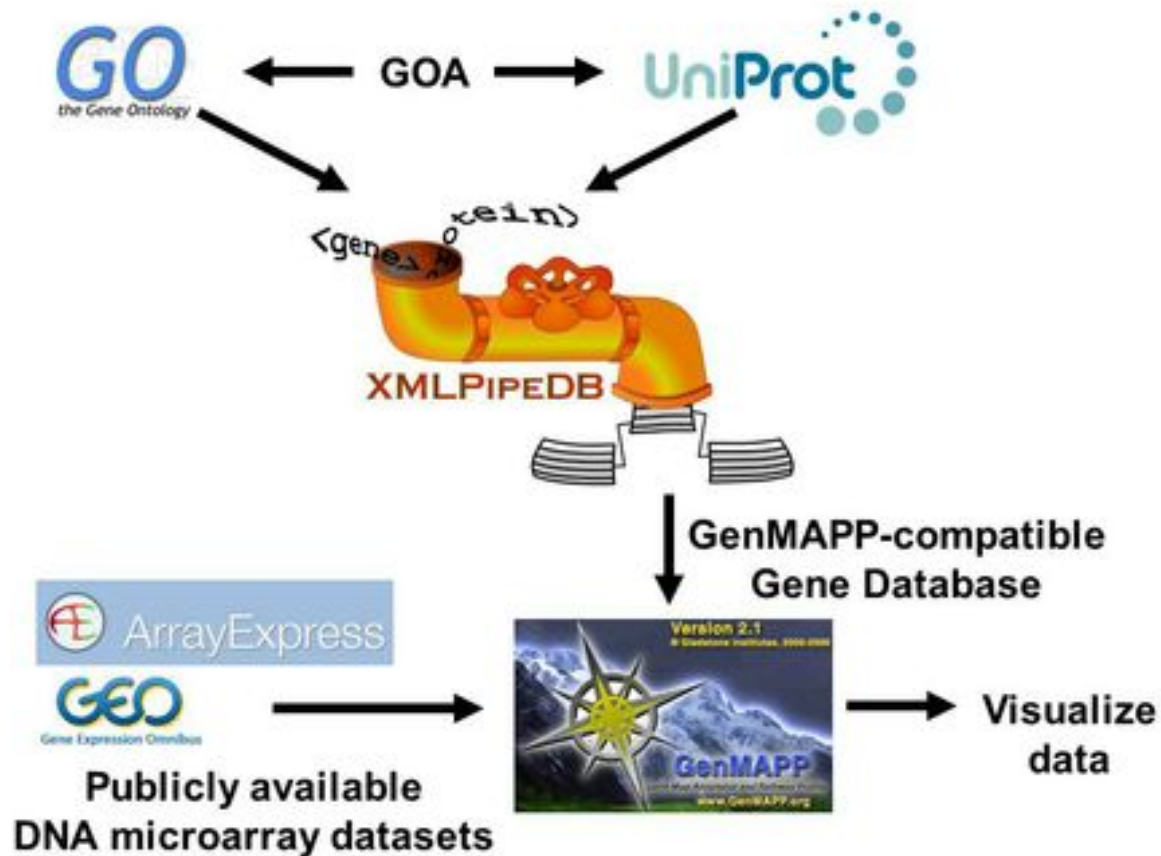


Figure 2. A summary of the entire workflow for this project. The UniProt complete proteome for *B. cenocepacia*, Gene Ontology terms, and a Gene Ontology association file were all imported into GenMAPP Builder, a program sourced from XMLPipeDB. An export is run through GenMAPP Builder that returns a GenMAPP-compatible Gene Database. After data from a publically available DNA microarray dataset have had statistical modifications and tests made, they are run through GenMAPP along with the previously created database. GenMAPP then returns colored gene data indicating an increased, decreased, or negligible change in gene expression. This graphic was created by Dr. Kam D. Dahlquist and Dr. John David N. Dionisio as a precursor to this project.

```

BurkholderiaCenocepaciaUniProtSpeciesProfile.java UniProtDatabaseProfile.java
1 package edu.lmu.xmlpipepdb.gmbuilder.databasetoolkit.profiles;
2
3 import edu.lmu.xmlpipepdb.gmbuilder.databasetoolkit.tables.TableManager;
4 import edu.lmu.xmlpipepdb.gmbuilder.databasetoolkit.tables.TableManager.QueryType;
5
6 public class BurkholderiaCenocepaciaUniProtSpeciesProfile extends UniProtSpeciesProfile {
7
8     public BurkholderiaCenocepaciaUniProtSpeciesProfile() {
9         super("Burkholderia cenocepacia",
10             216591,
11             "This profile customizes the GenMAPP Builder export for " +
12             "Burkholderia cenocepacia" +
13             " data loaded from a UniProt XML file.");
14     }
15
16     @Override
17     public TableManager getSystemsTableManagerCustomizations(TableManager tableManager, DatabaseProfile dbProfile) {
18         super.getSystemsTableManagerCustomizations(tableManager, dbProfile);
19         tableManager.submit("Systems", QueryType.update, new String[][] {
20             { "SystemCode", "N" },
21             { "Species", "|" + getSpeciesName() + "|" }
22         });
23
24         tableManager.submit("Systems", QueryType.update, new String[][] {
25             { "SystemCode", "N" },
26             { "Link", "http://www.burkholderia.com/getAnnotation.do?locusID=~" }
27         });
28
29         return tableManager;
30     }
31 }

```

Figure 3. The initial set of code added into the customized species profile within XMLPipeDB for *Burkholderia cenocepacia*. This initial version of code was used to create a dry build and determine the additional customizations that needed to be made to GenMAPP Builder to catch all of the genes and ultimately produce a comprehensive database for *B. cenocepacia*.

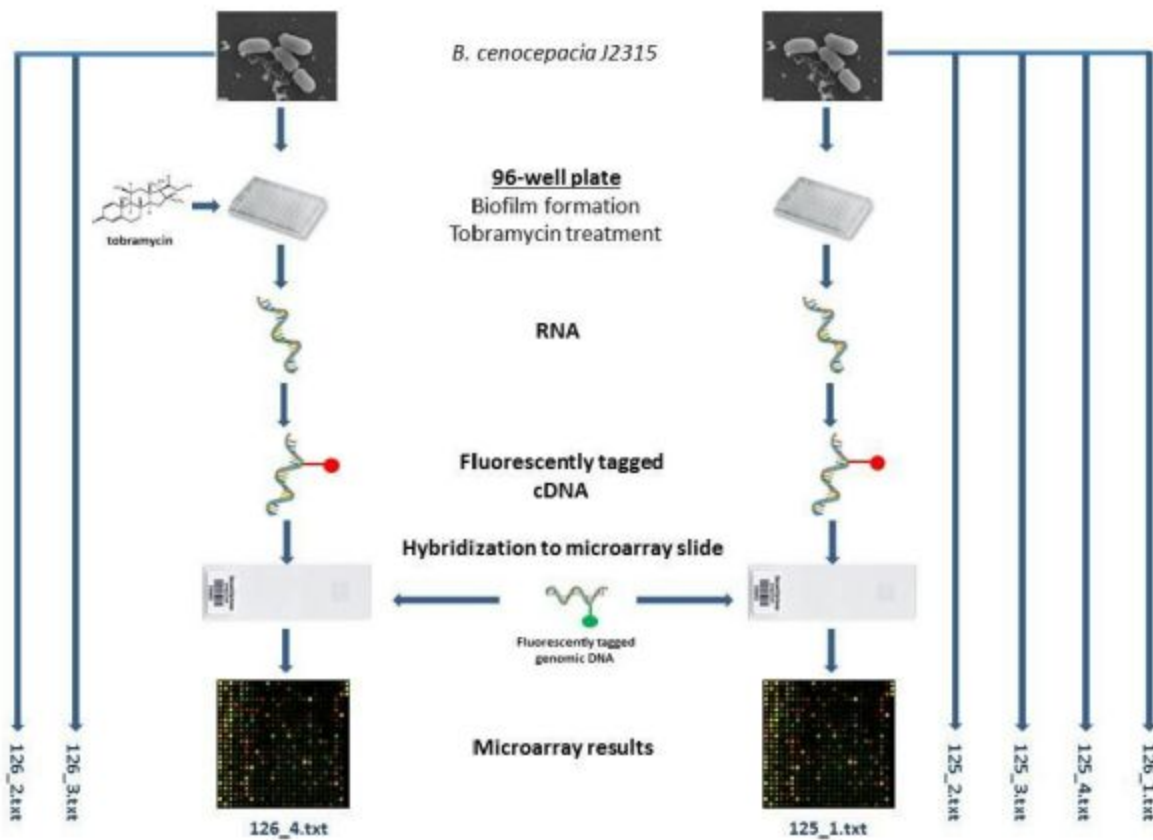


Figure 4. Van Acker *et al* performed a transcriptomic analysis on *Burkholderia cenocepacia* persister cells treated with tobramycin, versus untreated cells. A workflow of the experimental design is shown, beginning with growth of the cells and ending with the raw data files available on ArrayExpress^[8].

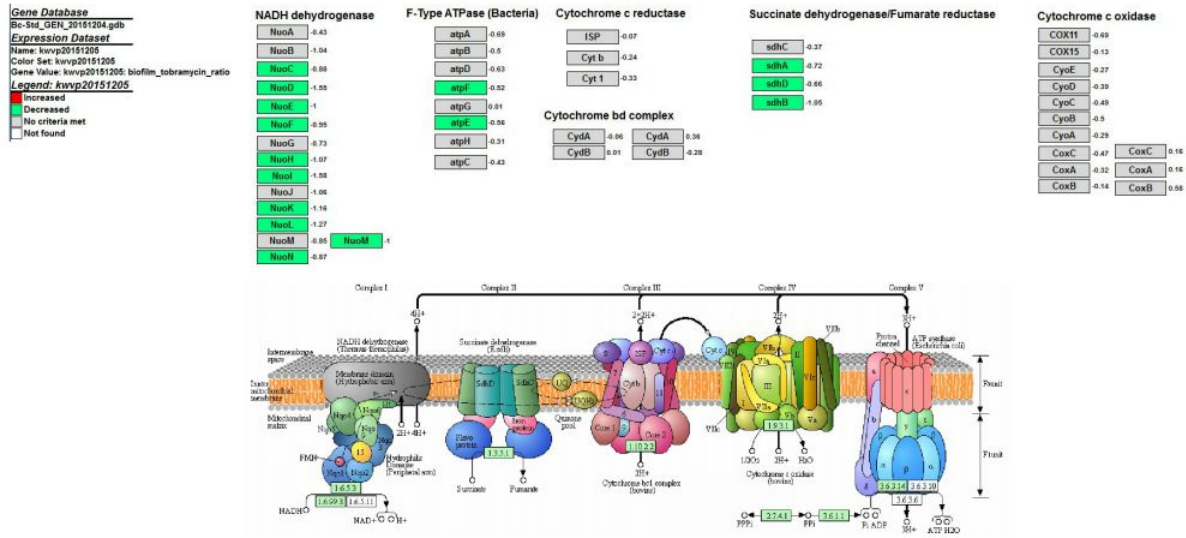


Figure 5. A MAPP of oxidative phosphorylation shows that it was significantly downregulated.

```

@Override
public TableManager getSystemTableManagerCustomizations(TableManager tableManager, TableManager primarySystemTableManager, Date version)
    throws SQLException, InvalidParameterException {
    List<String> comparisonList = new ArrayList<String>(1);
    comparisonList.add("ORF");

    return systemTableManagerCustomizationsHelper(tableManager, primarySystemTableManager, version, "OrderedLocusNames", comparisonList);
}

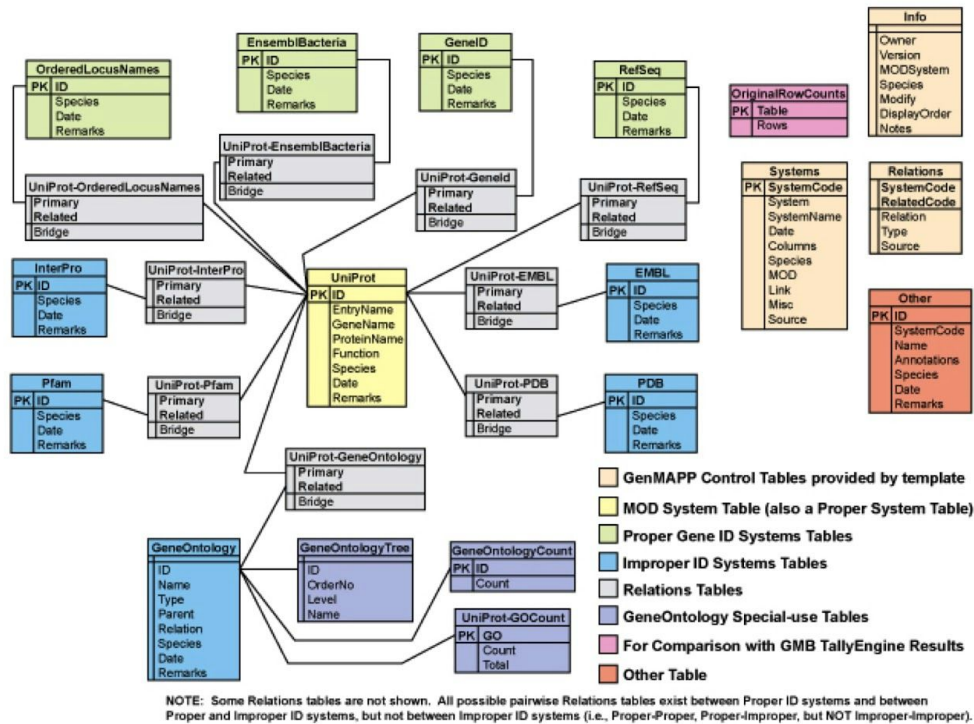
```

Figure 6. The additional code added to *B. cenocepacia*'s customized species profile. This code collects gene IDs from the ORF tag in the XML file instead of from the OrderedLocusNames tag, thereby allowing all 7121 gene IDs to be caught and represented in the final exported database.

XML Path	XML Count	Database Table	Database Count
UniProt	6994	UniProt	6994
RefSeq	5953	RefSeq	5953
GeneID	5953	GeneID	5953
Ordered Locus	337	Ordered Locus	337
ORF	7121	ORF	7121
GO Terms	43954	GO Terms	43954

Figure 7. The TallyEngine results associated with the final GenMAPP Gene Database export.

The tally results indicate the counts of UniProt and GO data that are present in the UniProt XML and the postgresQL database that was created for this export. The parity of the XML counts and database counts suggests that the data was successfully imported, from the XML, into the created postgresQL database. Note: the TallyEngine results for the final exported database indicates the number of “ordered locus” type gene IDs, in addition to “ORF” type gene IDs, in its output. However, only the “ORF” type IDs are incorporated into the final GenMAPP Gene Database.



Resource	Count of OrderedLocusNames IDs
XMLPipeDB Match, in the UniProt proteome set for J2315 (XML)	7126
TallyEngine, in the UniProt proteome set for J2315 (XML)	7121
TallyEngine, in the last created PostgreSQL database	7121
OriginalRowCounts table, in the last exported gene database for J2315	7121
Model Organism Database for <i>B. cenocepacia</i> (CDS Gene IDs)	7114

Table 1. Summary of the counts of OrderedLocusNames IDs associated with Burkholderia Genome Database, UniProt XML, and the PostgreSQL database and the OriginalRowCounts table of the final GenMAPP Gene Database. UniProt XML counts were found using TallyEngine and XMLPipeDB Match. PostgreSQL counts were found using TallyEngine and were verified using SQL queries. Regarding Burkholderia Genome Database (the model organism database for J2315), the count of coding sequence gene IDs was utilized as the count of OrderedLocusNames IDs.

Sanity Check	Number of Genes	Percentage out of 7251 genes (%)
P-value <0.05	4318	59.6
P-value <0.01	2350	32.4
P-value <0.001	1460	20.1
BH P-value <0.05	605	8.3
Bonferroni P-value < 0.05	179	2.5

Table 2. Number of genes which exhibited an expression change considered to be statistically significant by the given criteria.

	GOID	GO Name	Number Changed	Number Measured	Number in GO	Percent Changed	Percent Present	PermuteP	AdjustedP
	16226	iron-sulfur cluster assembly	7.00	8.00	8.00	87.50	100.00	0.0000	0.0090
	30288	outer membrane-bounded periplasmic space	11.00	20.00	20.00	55.00	100.00	0.0000	0.0720
	43225	anion transmembrane-transporting ATPase activity	5.00	8.00	1.00	62.50	800.00	0.0080	0.8660
Increase	15074	DNA integration	10.00	27.00	27.00	37.04	100.00	0.0200	1.0000
	9628	response to abiotic stimulus	4.00	7.00	1.00	57.14	700.00	0.0250	1.0000
	4185	serine-type carboxypeptidase activity	4.00	8.00	8.00	50.00	100.00	0.0340	1.0000
	9086	methionine biosynthetic process	4.00	9.00	9.00	44.44	100.00	0.0590	1.0000
	43232	intracellular non-membrane-bounded organelle	25.00	68.00	57.00	36.76	119.30	0.0000	0.0000
	19843	rRNA binding	18.00	39.00	39.00	46.15	100.00	0.0000	0.0000
	6412	translation	30.00	113.00	106.00	26.55	106.60	0.0000	0.0000
	44391	ribosomal subunit	9.00	14.00	7.00	64.29	200.00	0.0000	0.0000
Decrease	16655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	10.00	19.00	15.00	52.63	126.67	0.0000	0.0000
	48038	quinone binding	6.00	10.00	10.00	60.00	100.00	0.0000	0.0010
	51276	chromosome organization	7.00	26.00	14.00	26.92	185.71	0.0000	0.5080
	6119	oxidative phosphorylation	4.00	7.00	7.00	57.14	100.00	0.0010	0.0540
	9082	branched-chain amino acid biosynthetic process	4.00	16.00	15.00	25.00	106.67	0.0020	1.0000

Table 3. The table shows a list of gene ontology terms with corresponding pathways and/or functions considered by MAPPFinder to be significantly upregulated or downregulated in *Burkholderia cenocepacia* biofilm cells, as a response to tobramycin treatment.

Gene number	Annotation	Log Fold Change - GENI/DMICS	Log Fold Change - Van Acker et al. Microarray	Log Fold Change - Van Acker et al. qPCR
Glyoxylate shunt				
BCAL2122	Malate synthase	0.24	1.4	-3.3
BCAL2118	isocitrate lyase AceA	0.59	2.3	1.9
BCAM1588	isocitrate lyase	0.78	3.1	1.6
BCAL0813	RNA polymerase factor sigma 54	-0.25	-1.5	-
BCAL1945	Glyoxylate carboxylase	-0.30	-1.5	-
Tricarboic acid cycle				
BCAM0961	Aconitate hydratase	-0.09	-1.1	-
BCAM2701	Aconitate hydratase	-0.16	-1.3	-1.3
BCAM1833	Aconitate hydratase/methylisocitrate dehydratase	0.27	1.5	1.1
BCAL2735	isocitrate dehydrogenase	0.20	1.3	-2.5
BCAL2736	isocitrate dehydrogenase	0.02	1.4	1.5
BCAL1515	L-isoleucine dehydrogenase E1	-0.49	-2.0	-
BCAL1516	Dihydropyrimidine succinyl transferase	-0.31	-3.3	-
BCAL1517	Dihydropyrimidine dehydrogenase	-0.68	-2.5	-
BCAL2207	Purative dihydropyrimidine dehydrogenase	-0.20	-1.3	-
BCAL1215	Dihydropyrimidine dehydrogenase	-0.25	-1.4	-
BCAL0956	Succinyl-CoA synthetase beta chain	-0.50	-2.0	-
BCAL0957	Succinyl-CoA synthetase subunit alpha	-0.83	-3.3	-10.0
BCAM0958	Purative succinate dehydrogenase	-0.37	-1.7	-
BCAM0960	Succinate dehydrogenase flavoprotein	-0.72	-2.5	-
BCAM0970	Succinate dehydrogenase iron-sulfur subunit	-1.05	-5.0	-25.0
BCAL2908	Fumarate hydratase	-0.16	-1.3	1.9
BCAL2207	Purative fumarate dehydrogenase	0.02	1.0	-
BCAM0955	Malate dehydrogenase	-0.03	1.0	-2.0
BCAL2745	Purative citrate synthase	-0.14	-1.3	-
BCAM0964	Purative lyase	0.23	-1.4	-
BCAS0023	HcpH/HcpJ aldolase/citrate lyase family	-0.71	-2.5	-
BCAM0972	Type II citrate synthase	-1.15	-5.0	-
Oxidative phosphorylation				
BCAL2742	Cytochrome c ubiquinol oxidase subunit II	-0.49	-2.0	-
BCAL2743	Ubiquinol oxidase polypeptide I	-0.50	-2.0	-
BCAL0750	Cytochrome c oxidase polypeptide I	-0.32	-1.7	-
BCAL0752	Cytochrome c oxidase assembly protein	-0.69	-2.5	-
BCAL0753	Hypothetical protein	-0.61	-2.5	-
BCAL0754	Purative cytochrome c oxidase subunit II	-0.47	-2.0	-
NAD(P)H production				
BCAL3276	NAD-kinase	0.22	1.4	-
BCAL0672	isocitrate dehydrogenase kinase/phosphatase	0.21	1.3	-
BCAL3355	Glutamate dehydrogenase	1.01	4.2	-
BCAL3395	Malo enzyme	0.35	1.7	-
Response to oxidative stress				
BCAL1250	Purative glutathione S-transferase	0.35	1.6	-
BCAL3331	Purative glutathione S-transferase	0.86	3.4	-
BCAL0463	Purative thioredoxin	0.35	1.6	-
BCAL2013	AlpC/TSA family protein	0.47	1.9	-
BCAL2795	Glutathione peroxidase	0.32	1.6	-
BCAM2318	Purative thioredoxin oxidoreductase	-1.96	-10.0	-33.3
Fe-storage				
BCAM2827	Purative hemin ABC transporter protein	1.15	5.2	-
BCAM2830	Hemin importer ATP binding subunit	0.73	2.8	-
BCAM2224	Purative pyochelin receptor protein PprA	0.69	2.7	-
BCAL1790	Purative iron-transport protein	0.64	2.5	-
BCAL1347	Purative Fe uptake system extracellular binding protein	0.63	2.5	-
BCAM2228	Purative pyochelin synthetase PpHF	0.01	2.1	-
BCAL1789	Purative iron-transport protein	0.48	2.0	-
BCAL1371	Purative TonB-dependent siderophore receptor	0.47	2.0	-
BCAL1702	Purative ornithine biosynthesis protein	-0.56	-2.2	-

Table 4. Calculated log fold changes for the genes of interest in a study by Van Acker et al. (2013). Values shown in red were deemed significant by their respective parties.

References

1. Caraher, E., Reynolds, G., Murphy, P., McClean, S., & Callaghan, M. (2007). Comparison of antibiotic susceptibility of *Burkholderia cepacia* complex organisms when grown planktonically or as biofilm in vitro. *European journal of clinical microbiology & infectious diseases*, 26(3), 213-216.
2. Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., & Lewis, S. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2), 288-289.
3. Dionisio, J. D. N., & Dahlquist, K. D. (2008). Improving the computer science in bioinformatics through open source pedagogy. *ACM SIGCSE Bulletin*, 40(2), 115-119.
4. Holden, M. T., Seth-Smith, H. M., Crossman, L. C., Sebahia, M., Bentley, S. D., Cerdeño-Tárraga, A. M., ... & Cherevach, I. (2009). The Genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *Journal of bacteriology*, 191(1), 261-277.
5. Keren, I., Kaldalu, N., Spoering, A., Wang, Y., & Lewis, K. (2004). Persister cells and tolerance to antimicrobials. *FEMS microbiology letters*, 230(1), 13-18.
6. Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., ... & Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC bioinformatics*, 8(1), 217.

7. Van Acker, H., Sass, A., Bazzini, S., De Roy, K., Udine, C., Messiaen, T., ... & Coenye, T. (2013). Biofilm-grown *Burkholderia cepacia* complex cells survive antibiotic treatment by avoiding production of reactive oxygen species. *PLoS One*, 8(3), e58943.
8. Van Acker, Helen. "E-MEXP-3532 -Transcription Profiling by Array of Tobramycin Tolerance in *Burkholderia cenocepacia* J2315 Biofilms." *ArrayExpress*. European Molecular Biology Laboratory, 1 Feb. 2012. Web.
9. Winsor, G. L., Khaira, B., Van Rossum, T., Lo, R., Whiteside, M. D., & Brinkman, F. S. (2008). The *Burkholderia* Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics*, 24(23), 2803-2804.
10. "Oxidative Phosphorylation - *Burkholderia cenocepacia* J2315." *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Kanehisa Laboratory, 16 June 2015. Web.

Appendix

Testing Report for Final Version of GenMAPP Builder

Version of GenMAPP Builder: GenMAPP Builder Custom, Build 4

Computer on which the export was run: Home Workstation

Postgres Database name: B.cenocepacia_J2315_20151204_BUILD4_genialomics

UniProt XML filename: [uniprot-taxonomy%3A216591_GEN_BL12_20151119.xml](#)

- UniProt XML version: UniProt release 2015_11 - November 11, 2015

- UniProt XML download link: [UniProtKB link for the complete proteome of J2315](#)
- Time taken to import: 3.46 minutes
 - Note: Time taken appears to be slightly shorter than previous exports.

GO OBO-XML filename: [go_daily-termdb_GEN_BL12_20151119.obo-xml](#)

- GO OBO-XML version (derived from the date modified on the file, itself): *Date Modified: 11/19/2015 2:24 AM*
- GO OBO-XML download link: [Link from GO website](#)
- Time taken to import: 5.05 minutes
- Time taken to process: 3.75 minutes
 - Note: Time taken appears to be slightly shorter than previous exports.

GOA filename: [31277.B_cepacia_GEN_BL12_20151119.goa](#)

- GOA version: *Date Modified: 11/10/15, 1:47:00 PM* (information sourced from FTP site)
- GOA download link: [FTP site file](#)
- Time taken to import: 0.04 Minutes
 - Note: No issues were found with the import of this file.

Name of .gdb file: [Bc-Std GEN Build4 20151204.gdb](#)

- Time taken to export: 11 hours 6 minutes
 - Start time: 7:51 am
 - End time: 6:57 pm

- Note: File was exported without any major issues, however, the export appeared to take significantly longer than the previous exports. It is likely that the export took so long because the workstation had, for some period of time, entered a "sleep" mode (export was delayed, as the computer had to be taken off of "sleep").

Using TallyEngine

- PostgreSQL was initialized through pgAdmin III and the database B.cenocepacia_J2315_20151204_BUILD4_genialomics was left running
- GenMAPP builder was booted and *Run XML and Database Tallies for UniProt and GO* was selected under the *Tallies* menu item; the UniProt XML and GO files that were imported were chosen
- **Results of TallyEngine**

XML Path	XML Count	Database Table	Database Count
UniProt	6994	UniProt	6994
RefSeq	5953	RefSeq	5953
GeneID	5953	GeneID	5953
Ordered Locus	337	Ordered Locus	337
ORF	7121	ORF	7121
GO Terms	43954	GO Terms	43954

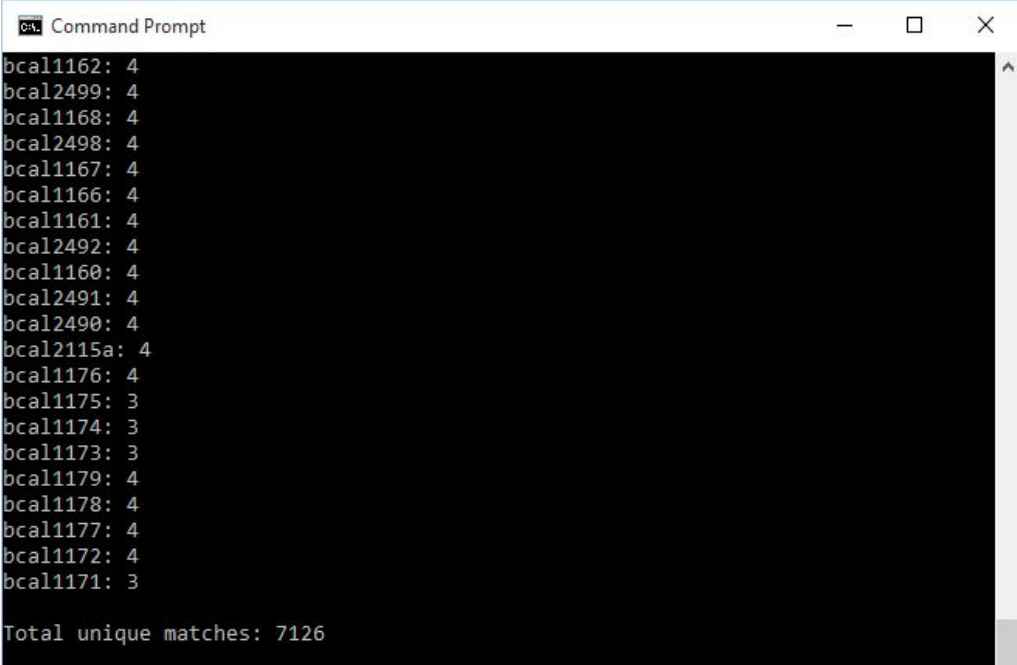
Close

- Note: These results differ significantly from what was found in previous exports. The 337 Ordered Locus gene names are now distinct from the 7121 ORF gene names (and are represented, as such, by TallyEngine). All of the counts related to external references (like UniProt) remain the same. The major and crucial change is the inclusion and representation of the ORF data.

Using XMLPipeDB match to Validate the XML Results from the TallyEngine

- The Windows command line was launched (cmd.exe)

- This set of commands was inputted into the command line in order to utilize XMLPipeDB match to verify the OrderedLocusNames count:
- `java -jar xmlpipedb-match-1.1.1.jar`
`"p?BCA[LMS]?[0-9][0-9][0-9][Aa]?[0-9]?[A-Z,a-z]?" <`
`"uniprot-taxonomy%3A216591_GEN_BL12_20151119.xml"`
 - NOTE: Prior to executing the command, the folder that held the files and `xmlpipedb-match-1.1.1.jar` was entered through the Windows command line (a set of CD commands was used in order to enter the correct directory).



```
Command Prompt
bcal1162: 4
bcal2499: 4
bcal1168: 4
bcal2498: 4
bcal1167: 4
bcal1166: 4
bcal1161: 4
bcal2492: 4
bcal1160: 4
bcal2491: 4
bcal2490: 4
bcal2115a: 4
bcal1176: 4
bcal1175: 3
bcal1174: 3
bcal1173: 3
bcal1179: 4
bcal1178: 4
bcal1177: 4
bcal1172: 4
bcal1171: 3
Total unique matches: 7126
```

- 7126 unique matches were found through XMLPipeDB match

Are your results the same as you got for the TallyEngine? Why or why not?

- These results vary slightly from what was found by TallyEngine due to the presence of 5 discrepant IDs
- The discrepant IDs, previously identified, are: bca199f, bca5253f, bca636c, bca837b, bcal0235a, and bcal0239a
- bca199f, bca5253f, bca636c, and bca837b were found to be a part of a sequence of letters and numbers under the label of "checksum"; these appeared to have been accidentally captured by the utilized Match command.
- bcal0235a and bcal0239a follow the previous identified gene name patterns, however, they both show up as database reference IDs (database reference to STRING, which is a database of known and predicted protein interactions; these data will be ignored as they do not refer to a UniProt entry.
- Excluding these 5 accidental matches, the results found using the Match utility are the same as what was found using TallyEngine

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

- *pgAdmin III* was booted and all of the necessary connections were made
- In *pgAdmin III*, the query `select count(*) from genenametype where type = 'ordered locus' and value ~ 'BceJ2315_[0-9][0-9][0-9][0-9]'`; was issued via the SQL Query menu in order to validate the TallyEngine count for "Ordered Locus" for the PSQL database.
 - 337 unique matches were found in *pgAdmin III* (postgres database results).
This lines up with what was found in TallyEngine.

- Additionally, the query `select count(*) from genenametype where type = 'ORF' and value ~ 'p?BCA[LMS]?[0-9][0-9][0-9][Aa]?[0-9]?[A-Z,a-z]?'`; was run via SQL in order to verify the ORF counts
 - 7121 counts were found which is identical to what was found through XMLPipeDB match (ignoring the discrepant IDs) and to what was reported by TallyEngine (for the ORF data).
- Are your results the same as reported by the TallyEngine? Why or why not?
 - The results are the same as what was reported by TallyEngine; this is due to the fact that the most recent build incorporated code fixes that allowed GenMAPP builder, and TallyEngine, to properly include the ORF data in their analysis/work.

OriginalRowCounts Comparison

- The newly created J2315 .gdb file was opened with a program that is able to explore a .mdb file (such as Microsoft Access); in this case, [MDB Viewer Plus](#) was utilized.
- Using the program, the OriginalRowCounts table was looked at, which contained summaries regarding each of the tables within the database (and the # of rows/entries in each of the tables)
- **OriginalRowCounts for Build 4 export of J2315**

Table	Rows
▶ Info	1
Systems	35
Relations	35
Other	0
GeneOntologyTree	87876
GeneOntology	5665
UniProt-GOCount	3279
GeneOntologyCount	3278
UniProt-GeneOntology	23396
UniProt	6994
RefSeq	5953
EMBL	8
Pfam	2496
InterPro	4894
GeneID	5953
EnsemblBacteria	7121
PDB	67
OrderedLocusNames	7121
UniProt-PDB	67
UniProt-OrderedLocusNames	7121
UniProt-EnsemblBacteria	7122
UniProt-GeneID	5954
UniProt-InterPro	17707
UniProt-Pfam	7717
UniProt-EMBL	7023
UniProt-RefSeq	5954
RefSeq-EMBL	5976
RefSeq-Pfam	6742
RefSeq-InterPro	15547
RefSeq-GeneID	5953
RefSeq-EnsemblBacteria	6069
RefSeq-OrderedLocusNames	6069
RefSeq-PDB	62
GeneID-EMBL	5976
GeneID-Pfam	6742
GeneID-InterPro	15547
GeneID-EnsemblBacteria	6069
GeneID-OrderedLocusNames	6069
GeneID-PDB	62
EnsemblBacteria-EMBL	7271
EnsemblBacteria-Pfam	7898
EnsemblBacteria-InterPro	18090
EnsemblBacteria-OrderedLocusNames	7753
EnsemblBacteria-PDB	70
OrderedLocusNames-EMBL	7271
OrderedLocusNames-Pfam	7898

- It was decided that a good reference or "benchmark" would be the database that was created using Build 3 of the customized GenMAPP builder; comparing the two

should bring into light any issues or differences that could be the result of utilizing an updated version of the modified GenMAPP builder.

- Benchmark .gdb file: [compressed Bc-Std_GEN_Build3_20151203.gdb](#)
- **OriginalRowCounts for the Build 3 export of J2315**

Table	Rows
▶ Info	1
Systems	35
Relations	35
Other	0
GeneOntologyTree	87876
GeneOntology	5665
UniProt-GOCount	3279
GeneOntologyCount	3278
UniProt-GeneOntology	23396
UniProt	6994
RefSeq	5953
EMBL	8
Pfam	2496
InterPro	4894
GeneID	5953
EnsemblBacteria	7121
PDB	67
OrderedLocusNames	7121
UniProt-PDB	67
UniProt-OrderedLocusNames	7121
UniProt-EnsemblBacteria	7122
UniProt-GeneID	5954
UniProt-InterPro	17707
UniProt-Pfam	7717
UniProt-EMBL	7023
UniProt-RefSeq	5954
RefSeq-EMBL	5976
RefSeq-Pfam	6742
RefSeq-InterPro	15547
RefSeq-GeneID	5953
RefSeq-EnsemblBacteria	6069
RefSeq-OrderedLocusNames	6069
RefSeq-PDB	62
GeneID-EMBL	5976
GeneID-Pfam	6742
GeneID-InterPro	15547
GeneID-EnsemblBacteria	6069
GeneID-OrderedLocusNames	6069
GeneID-PDB	62
EnsemblBacteria-EMBL	7271
EnsemblBacteria-Pfam	7898
EnsemblBacteria-InterPro	18090
EnsemblBacteria-OrderedLocusNames	7753
EnsemblBacteria-PDB	70
OrderedLocusNames-EMBL	7271
OrderedLocusNames-Pfam	7898

Note: It was noticed that the OriginalRowCounts table in this export is identical to the one that came from the Build 3 export. This seems to suggest that the only fundamental difference between the two builds of GenMAPP builder lies with TallyEngine (this makes sense, considering that build 4 focused upon fixing problems with TallyEngine and improper code).

Visual Inspection

Perform visual inspection of individual tables to see if there are any problems.

- Look at the Systems table. Is there a date in the Date field for all gene ID systems present in the database?
 - Yes, there are dates present for GeneOntology, InterPro, GeneID, RefSeq, UniProt, EMBL, PDB, Pfam, OrderedLocusNames, and EnsemblBacteria.
- Open the UniProt, RefSeq, and OrderedLocusNames tables. Scroll down through the table. Do all of the IDs look like they take the correct form for that type of ID?
- In the UniProt table, like before, it is apparent that only gene names of the type "ordered locus" are represented (no signs of gene names that begin with something like "BCA"). The RefSeq table appears to not have any problems. The ordered locus names table, like in Build 3, only reflects gene names in the form of `p?BCA[L,M,S]?[0-9][0-9][0-9][A,a]?[0-9]?[A-Z, a-z]`; it appears that the "ORF" data replaced the "ordered locus" gene names in this table (these IDs appear to be in the correct and common form).

Note: Visually, no changes seem apparent between the Build 3 and Build 4 export.

.gdb Use in GenMAPP[[edit](#)]

- Some of the protocol from *Part 2 of the Vibrio cholerae Microarray Data Analysis* was used as a reference for this portion of the assignment
- *Bc-Std_GEN_Build4_20151204.gdb* was placed within the Gene Databases folder of the GenMAPP directory (the folder is within the GenMAPP 2 Data folder)
- GenMAPP (Version 2.1) was launched
- The new gene database was loaded by going into *Data > Choose Gene Database*
- The tab delimited GenMAPP formatted **data** sourced from the microarray paper was loaded into GenMAPP through *Data > Expression Dataset Manager > Expression Datasets > New Dataset > GenMAPP formatted microarray data_GEN_B14_20151207.txt*

Note: There were no glaring issues with loading the files into GenMAPP (no crashes). However, this gene database led to the detection of 284 errors in the loaded raw data; this error count is identical to what was seen with the build 3 export.

Putting a gene on the MAPP using the GeneFinder window

- A test expression data-set was created in order to observe the behavior of GenMAPP with the exported database
- GeneFinder was loaded by placing a blank *Gene* element on the drafting board of GenMAPP and right-clicking it.
- The genes BCAL0001,BCAL0002, BCAM0005, and BCAS0105 were searched in the Gene ID box, with the Gene ID System set to OrderedLocusNames

- All genes were successfully found and reference pages with links successfully appeared

Note: All cross-referenced IDs were present for all of these sample Gene IDs. No crashing or issues at this step.

Creating an Expression Dataset in the Expression Dataset Manager

- The IDs in the microarray dataset were imported into GenMAPP using the new database; there existed 284 exceptions.
- The EX.txt file was opened through Excel and it was found out that the exceptions were identical to what was found with the Build 3 export.

Exceptions Analysis

- **Note:** This analysis is sourced from the [Build 3 Export](#)
- The EX.txt file was opened through Excel and it was found out that the error code for all of the exceptions was: *Gene not found in OrderedLocusNames or any related system*. The Gene IDs were sorted by error and the problematic IDs were analyzed. It was found, through the find function, that 101 of the exceptions were due to alterations in the usual formatting of the gene name (these gene names contained underscores, Js, and numbers). The rest of the exceptions, it was found (via UniProt KB searches), represented genes that are not present in the UniProt database. Several exceptions (BCAL2591, BCALr0080, BCAM0787, BCAM1951, BCASr0743a) were checked for their presence in UniProt KB or in the MOD:

- BCAL2591: No results in UniProt KB. Found in MOD; gene has no product.
- BCALr0080: No results in UniProt KB. Found in MOD; product: tRNA-Arg.
- BCAM0787: No results in UniProt KB. No results in MOD.
- BCAM1951: No results in UniProt KB. Found in MOD; gene has no product.
- BCASr0743a: No results in UniProt KB. No results in MOD.
- Note: The exceptions file contained error inducing genes that either lack a known product (protein/functional RNA), lack a MOD entry, or code for functional RNA (such as tRNA). Some gene names that contained unusual formatting (BCAL0563_J_0, and BCAL0563_J_1, for example) were found to represent genes that were covered by the MOD/UniProt (these entries were found by removing the unusual underscores/letters and searching the "fixed" gene names).
- [Excel Workbook utilized in visualizing the GenMAPP exceptions](#)

Running MAPPFinder

- Protocol sourced from [the week 8 assignment](#)
- The MAPPFinder program was launched within GenMAPP (Tools > MAPPFinder)
- "Calculate New Results" was clicked in the window that appeared by launching MAPPFinder
- For "Find File", the Expression Dataset file (with a .gex extension) was selected, and OK was clicked

- The Test criteria was selected
- The boxes corresponding to "Gene Ontology" were checked
- "Browse" button was clicked to add a name to the file that will be created
- "Run MAPPFinder" was clicked and the program was allowed to complete its analysis

Running MAPPFinder

- Protocol sourced from [the week 8 assignment](#)
- The MAPPFinder program was launched within GenMAPP (Tools > MAPPFinder)
- "Calculate New Results" was clicked in the window that appeared by launching MAPPFinder
- For "Find File", the Expression Dataset file (with a .gex extension) was selected, and OK was clicked
- The Test criteria was selected
- The boxes corresponding to "Gene Ontology" were checked
- "Browse" button was clicked to add a name to the file that will be created
- "Run MAPPFinder" was clicked and the program was allowed to complete its analysis

Note: MAPPFinder successfully loaded and provided an output with this gene database.

Compare Gene Database to Outside Resource

Outside Resource: [Burkholderia Genome DB](#), [UniProt KB](#)

- The strain page for J2315 was looked up: [1]

- 7121 OrderedLocusNames were found within the exported gdb file. 6,994 entries corresponding to J2315 proteins were found in UniProt KB, and 7114 coding sequences were found in the MOD. The count of 7121 genes that is represented by the exported database appears to make sense, in light of the data represented by the MOD and by UniProt. Since UniProt is protein-centric, the count of 6994 corresponds to only protein; it is likely that some proteins have several related gene names (which explains the reason why more gene names were found than proteins). The MOD was found, earlier, to be manually curated and it is possible that the difference between the MOD count of 7114 and the found count of 7121 is due to the MOD missing a few genes (that are present in other databases, like UniProt).
- Note: The IDs and counts covered by this export appear to be consistent with outside resources.

