Extending XML PipeDB to Create a Gene Database for the Analysis of *Mycobacterium tuberculosis*

Kevin Paiz-Ramirez, Reid Oldenburg, Cydnee Charles, and Jaime Bittner

Dahlquist Labs

Introduction

Mycobacterium tuberculosis is a pathogenic bacterium that infects primarily the mammalian respiratory system (Cole et al. 1998), causing tuberculosis. MT is classified as an acid-fast gram-positive bacillus-shaped bacterium due to the absence of an outer cell membrane (Camus et al 2002). Upon inoculation into the lungs, MT bacilli are phagocytosed by alveolar macrophages, in which they reside as intracellular parasites (Cole et al. 1998). MT persists in the intracellular compartment of macrophages by disrupting phagolyosomal fusion. MT avoids digestion in phagolysomes due to the waxy, hydrophobic properties of their cell wall, which necessitates a high number genes involved in fatty acid metabolism (Murray et al. 2005). MT can exist in a host's lungs for many years without becoming "active," or virulent, and it is estimated that about one third of the world population is infected (Wooldridge 2009). The worldwide distribution of MT is irregular, with a much higher caseload in the undeveloped world (Gagneux 2009). It is estimated that individuals with an active, not latent, TB infection can infect about 15 other individuals per year (World Health Organization 2009).

Tuberculosis commonly develops in people with compromised immune function, and with the emergence of the AIDS epidemic, the number of tuberculosis cases has resurged (WHO 2009). Patients develop active TB infections when endosome-bound MT begins to divide and consume alveolar macrophages. If the infection persists, the host generates a potent inflammatory response resulting in the formation of tissue granulation and necrosis at the site of active MT. Progressive tissue granulation can be identified upon histological examination, and gross pathologic examination reveals a caseous, cheesy and diffuse, necrosis (Martinko . 2005). If untreated, an active TB infection is lethal in 50% of patients (Goa et al. 2005), as infection can progress to further complications such as lobar pneumonia (Reddy et al. 2004).

Given the high mortality and infection rate of TB, a search for an efficacious treatment is still continuing and in 1993, the WHO declared TB a global health emergency. An MT vaccination provides partial protection against TB in children, but no effective vaccine exists for adults. TB-infected individuals are treated with long-term antibiotics (WHO 2009), but the emergences of multi-drug resistant TB strains are making effective treatment more difficult (Marinko, 2005).

Gene profiling of MT would likely yield better modeling and potential targeted therapies for active TB infections. By understanding the gene expression of MT at various stages of infection, tailored treatments could be engineered that are specific to particular strains or stages of infection. Bacterial activity is a mosaic of different genetic components, and

an understanding of each component can be done with microarray analyses. Microarrays provide a full description of individual gene activity, and can be done relatively quickly and at a manageable cost. Also, an understanding of strain to strain variations can lead to the identification of vaccine antigens. Therefore, a comprehensive, standardized and thorough approach to MT microarray analysis would improve the progress of MT research by making data accessible and understandable to all interested parties

The largest issue with understanding *M. tuberculosis* centers upon the limited resources available to analyze the data compiled from other researchers. The Gene Map Annotator and Pathway Profiler, or GenMAPP is a free, open source bioinformatics software tool designed to visualize and analyze genomic data in the context of pathways connecting gene-level datasets to biological processes and diseases (Dahlquist et al. 2002). GenMAPP would be a valuable tool to analyze raw data from *Mycobacterium tuberculosis* however; currently this is not yet possible because there is no gene database for MT. The solution would be to create a database for *Mycobacterium tuberculosis* using open source tool chains for building relational databases from available XML sources. XMLPipeDB is an open source suite of Java-based tools for automatically building relational databases from an XML schema (Loyola Marymount University. 2007 http://xmlpipedb.cs.lmu.edu) This program coupled with GenMAPP builder, would asses XML proteome set and GOA (GO association) files from integr8 and UniProt would offer the possibility of creating a gene database specifically for *Mycobacterium tuberculosis*. Having a unique database for MT, would allow for an easier navigation of known genes, and would provide for a powerful platform to diagram microarray data.

Dr. Qain Goa and colleagues explored the gene expression diversity among *Mycobacterium tuberculosis* clinical isolates in their microarray research by surveying 10 clinical isolates and 2 laboratory strains of tuberculosis. They measured gene expression under well-controlled *in vitro* conditions with RNA extracted under the exponential growth phase. The data was submitted to the Stanford Microarray Database where of the 3778 unique sequences repressed on the array, 3595 were retained after filtering. (Gao et al. 2005) The results concluded that variability in gene expression likely had an effect on the pathogenicity and the identification of candidate genes for drug targets and diagnostic assays between different strains of *M. tuberculosis*. This presented genetic variability as essential for bacterial survival and should be considered before proceeding with drug development. (Gao et al. 2005) By assessing the raw data from the Stanford microarray database the purpose of our investigation was to discover new information about the microarray data using GenMAPP by focusing on the MT laboratory strain H37Rv and the clinical isolate strain G, which was most significantly expressed.

Methods

Selection of proteome set and GO association files for *Mycobacterium tuberculosis*

In the acquisition of the proteome set and GOA files for MT, we were sure to record the version types and updates. Uniprot XML proteome set and GO association (GOA) files were downloaded from Integr8 (UniProt 15.10**)** for *Mycobacterium tuberculosis* (strain

H37Rv / ATCC 25618) Tax ID: 83332. UniProt XML file was downloaded from Integr8: 30.M_tuberculosis_ATCC_25618, UniProt 15.10 (November 4, 2009). GOA filename: 30.M_tuberculosis_ATCC_25618.goa, UniProt GOA Proteome Sets version 76 (October 8, 2009). The links to download the files can be found at the bottom of the strain description with the link: Downloads.

Acquisition of GO terms from the Gene Ontology

GO terms were downloaded from the site: http://www.geneontology.org/ GO.downloads.ontology.shtml on November 3, 2009 14:00 PST in the OBO-XML format, which comes in this form: go_daily-termdb.obo-xml.gz. The file was unzipped using 7zip which yielded: go_daily-termdb.obo-xml. This file provides terms, definitions and ontology structure to the PostgreSQL tools.

Formation of GenMAPP Builder tables in PostgreSQL

In order to form a GenMAPP Builder table set PgAdmin III was installed, for PostgresSQL its capabilities as administration and management tools software from http://www.pgadmin.org. Additionally pgAdmin v1.10.0 was installed (June 6,2009). In pgAdmin III, a new database was created with the name: TB_5RPO, owner: postgres, tablespace: pg_default, and character type: English, United States. The text from gmbuilder.sql was placed into TB_5RPO and then added the info by running (green arrow). This enabled the database and added all of the necessary tables.

Export of M. tuberculosis data into GenMAPP Gene Database

GenMAPP builder version: gmbuilder-20b37 (November 2, 2009) was installed and launched from http://sourceforge.net/projects/xmlpipedb/files. GenMAPP builder was configured with the TB_5RPO: database, postgres: username and password customized. For UniProt XML, 30.M_tuberculosis_ATCC_25618 was imported, and it took approximately 7 minutes. For import of GO XML, go_daily-termdb.obo-xml was imported, and it took approximately 9 minutes. Processing of the raw gene ontology data was queued up by GenMAPP builder and took approximately 23 minutes. Exporting the GenMAPP file used the file 30.M_tuberculosis_ATCC_25618.goa, and took approximately 160 minutes.

For code editing and commitment to add and make changes to the species profile for *Mycobacterium tuberculosis* changes were made to the java code by opening the code using eclipse (eclipse-jee-galileo-SR1-win32), http://www.eclipse.org. Some of the changes made and committed by Reid Oldenburg are:

```
# Mycobacterium tuberculosis
mycobacteriumtuberculosis_level_amount=2

mycobacteriumtuberculosis_element_level0=uniprot/entry/gene/name&type&o
rdered locus
mycobacteriumtuberculosis_element_level1=uniprot/entry/gene/name&type&O
RF
```

```
mycobacteriumtuberculosis_query_level0=select count(*) from
genenametype where type = 'ordered locus';
mycobacteriumtuberculosis_query_level1=select count(*) from
genenametype where type = 'ORF';

mycobacteriumtuberculosis_table_name_level0=Ordered Locus
mycobacteriumtuberculosis_table_name_level1=ORF
```

This is not a full listing of all of the new changes made and committed to the GenMAPP Builder through OpenSource work.

Inspection and validation of the Gene Database integrity

In order to determine the completion of the gene database file, or gdb, a TallyEngine was run to determine if GenMAPP builder picked up the correct records. A Fully-Monty TallyEngine table is shown below (Figure 2b). The TallyEngine results showed the presence of many different gene types, i.e. orderedlocusnames and Open Reading frames, ORFs.

This was checked to see if the appropriate IDs had been placed under the correct tables, with the expected number of records. This was done by opening the gdb (Rpo5Mt-Std_20091202.gdb) with Microsoft Office Access 2007. Each table was identified, and its interior searched for the appropriate gene IDs. Most importantly, UniProt, RefSeq, GeneId, and OrderedLocusNames tables were visually inspected to ensure that correct IDs were associated with the correct ID types for *Mycobacterium tuberculosis* strain H37Rv gene IDs, and this procedure was followed after the export of a new gdb file (Testing Report).

Preparation of Microarray Data

Preparing the microarray data involved reviewing the raw data from the Stanford microarray database. The raw data contained 48 chips, which were not in a particular order. This involved checking against the Stanford Microarray Database list of strains to separate the chips into 4 replicates of the 12 strains. From these chips, we assessed the ID column as well as the $\log_2$[R/G Normalized Ratio (Mean)]. This presented the raw data in excel with 5665 columns of data. Once these columns were composed into a main workbook additional data such as “PRIMERS”, “INTRONS,” “Unknowns” and “Empty” were deleted. Once the final deletions were completed there were 4750 columns of data. The Cy3 was the same for each strain as we were measuring mRNA levels of H37Rv versus the genome and since each gene is only represented once the log fold changes would all have one in the denominator. Inserting two rows in between the top row of headers and the first data row for “Average as well as StdDev” followed this. This compiled the averages as well as the standard deviations, which were followed by calculating the average log fold changes for each one of the twelve strains.

This was followed by a sanity check to make sure the data was analyzed correctly by determining the number of genes that were significantly changed at p value cut offs of <0.05. This provided the filtered data for the strain with the most significant change. The results from the TTEST were as follows, with strain G showing the most expression.

T test A	57 of 4750
T test c	888 of 4750
T test D	1438 of 4750
T test E	1193 of 4750
T test F	1134 of 4750
T test G	1272 of 4750
T test H	863 of 4750
T test I	1146 of 4750
T test J	1223 of 4750
T test K	1254 of 4750

Figure 1: T-test results of the 11 strains.

Running GenMAPP with the Gene Database

The next step was to open up and run GenMAPP using the gene database. Once opened, the database that was created by the coder was loaded into GenMAPP. Then a new Expression Dataset Manager was created with the manipulated data done in the excel sheet. This dataset was titled "Tuberculosis" and the Gene value was set to "Avg_logFC_G" because strain G was the most changed when run through the TTEST. The increased criterion was given the color red and in the criteria box it read, "[Avg_logFC_G]>0.25 AND [TTEST G]<0.05." The decreased criterion was given the color blue and in the criteria box it read, "[Avg_logFC_G]<-0.25 AND [TTEST G]<0.05." Once the criteria were set, we saved this expression dataset and exited the Expression Dataset Manager to go back to GenMAPP. This gave us a .gex file, which was saved to the desktop. We then ran MAPPfinder with our database that was uploaded to GenMAPP, and clicked on "calculate new results." We chose both "increased" and "decreased" criteria in the box on the right and then checked the boxes for "Gene Ontology" and "P value."

Microarray data (import using Expression Dataset Manager)

The Mapp was created and the next step was to "show ranked list" in order to get the top 10 GO terms. This gives more of an insight into what the genes are doing in the cell. The top 10 GO terms for the criterion0 (Increased gene expression) were macromolecular complex, fatty acid metabolic process, coA carboxylase activity, fatty acid biosynthetic process, cytoplasmic part, hydrogen ion transmembrane transporter activity, generation of precursor metabolites and energy, ligase activity forming carbon-carbon bonds, lipid metabolic process, and protein complex. The top 10 GO terms for the criterion1 were Transposition, DNA recombination, transposition, DNA-mediated, transposase activity, DNA binding, DNA metabolic process, cellular_component, intracellular part, cytoplasm, and glycerol-3-phosphate metabolic process.

MAPPFinder analysis

An analysis of the genes can be performed from these two lists of top 10 GO terms. The GO terms relate to which genes are being expressed more significantly, either increased or decreased gene expression. Criterion0 refers to those genes that are being increased and based on the GO terms, it seems that there is a lot of metabolic activity and perhaps the bacterial cells are eating a lot and making a lot of energy. Criterion1 refers to the genes that are being decreased in expression. They have to do with more DNA formation and this would mean they are not doing as much growth numbers wise. The bacteria may be doing a lot of metabolic processing but not as much DNA replication, i.e. the cells are not dividing.

Gene Placement on MAPP and Pathway Illustrations

Through these two collections of the GO terms, we could decipher which pathways were relatively significant in the bacterial cell and then draw a pathway for our specific strain. The fatty acid metabolic pathway is increased in the Mycobacterium tuberculosis and this is the pathway that we chose to draw a map of. First, the Kegg pathway database (http://www.genome.jp/kegg/pathway.html) was opened and the fatty acid metabolism was chosen. Here you can find the pathway we used: http://www.genome.jp/kegg/pathway/map/map00071.html. Next we chose to highlight the genes in the generic pathway for the specific strain H37Rv. Once GenMAPP was started, our database was opened up as well as our expression dataset "tuberculosis2." Next, a gene was placed on the main window of genMAPP, using the "gene" button and by clicking the first gene that was highlighted in green in the keg website, an ID was obtained. The gene in GenMAPP opens up another window when double clicked, and into this we put the ID, which gave back the gene name. Now the gene on the GenMAPP window had a name and when we applied the expression dataset, it would become a color that correlated to whether the gene had increased (red), decreased (blue), or no criteria was found (gray). Using lines and arrows in the GenMAPP window, the map for the fatty acid metabolic pathway was drawn to completion. The actual pathway had a number of genes repeating so we noted this and simply drew the areas where new genes were being introduced.

Results

Gene Database Schema

GenMAPP builder created the gene database file (Rpo5Mt-Std 20091202.gdb) for *Mycobacterium tuberculosis* strain H37Rv with the customized species profile. Figure 2a shows the Schema Diagram that illustrates the format to which the ID types are configured. The Schema Diagram had previously been created and supplied by the Dahlquist and Dionisio group, and was adapted to MT after cross-reference with the gdb file.

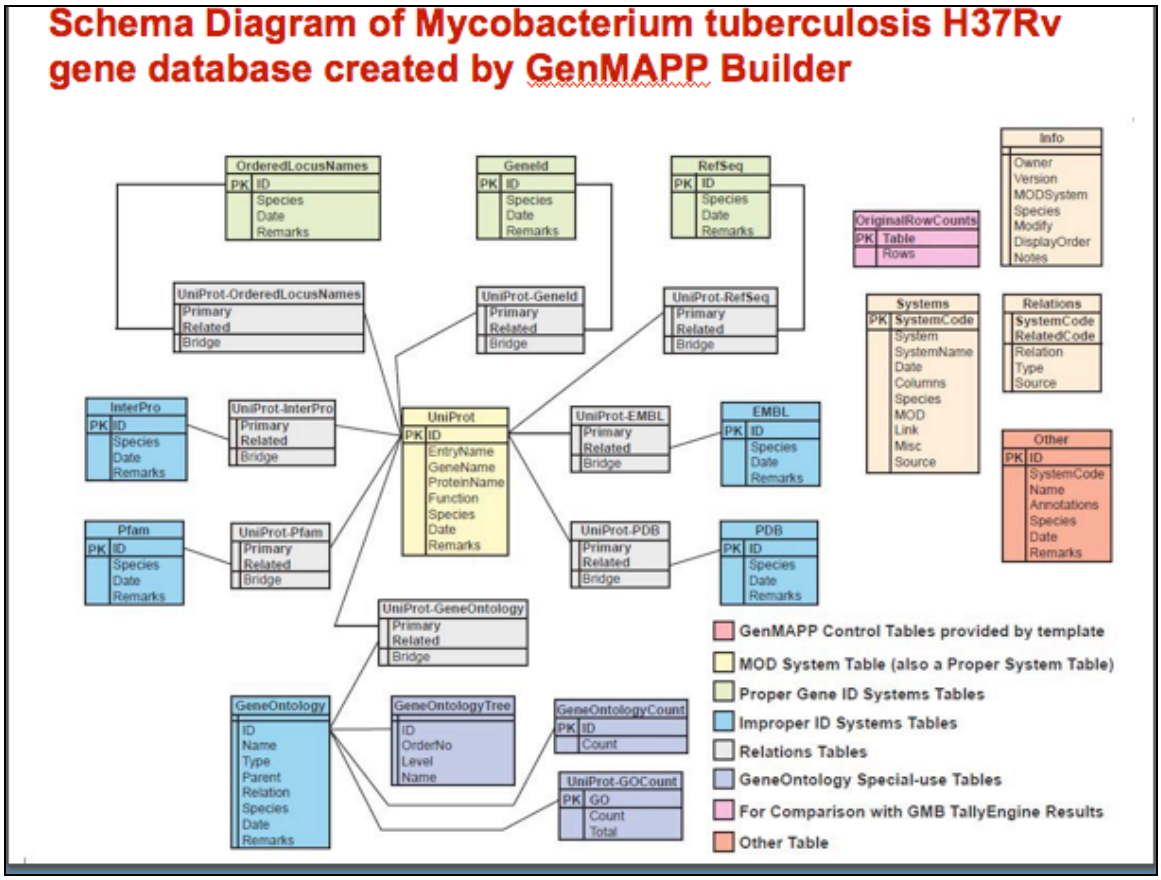


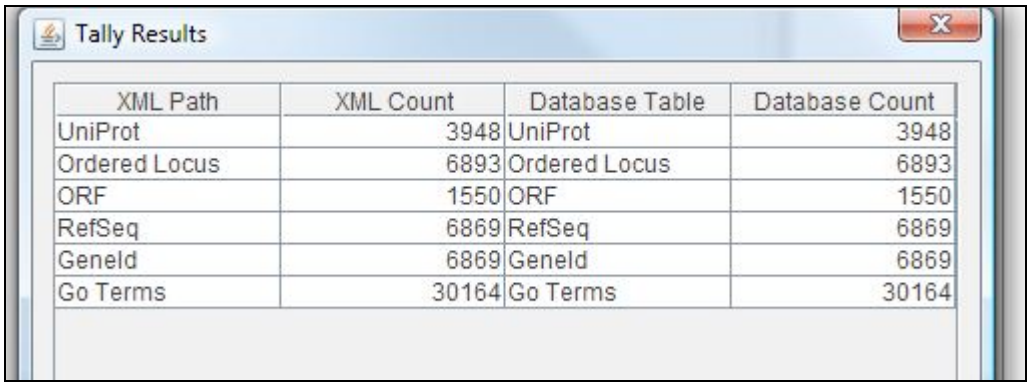
Figure 2a: Schema Diagram of *Mycobacterium tuberculosis* strain H37Rv Gene Database.

Final version of Gene Database Testing Report

The import and export of the gene IDs for Mycobacterium tuberculosis required the editing of the java code in order to create a version of GenMAPP Builder with an individualized species profile. Changes were made by the team's coder and ID minder, Reid Oldenburg and Cydnee Charles, respectively; with the guidance of Drs. Dionisio and Dahlquist. File download and use is covered in the methods section.

To check the appropriate uptake of the IDs by GenMAPP Builder, a TallyEngine was run after an export of the gene database (figure 2b), which is shown below. All of the fields match, which demonstrates the capability of the MT customized version of GenMAPP Builder to identify and include gene IDs.

An OrginalRowCounts comparison was done within the gdb file to see if the database maintained the correct tables and records; a description of the programs used can be found in the methods section. All of correct tables were maintained in the gdb file.



XML Path	XML Count	Database Table	Database Count
UniProt	3948	UniProt	3948
Ordered Locus	6893	Ordered Locus	6893
ORF	1550	ORF	1550
RefSeq	6869	RefSeq	6869
GeneId	6869	GeneId	6869
Go Terms	30164	Go Terms	30164

Figure 3b: TallyEngine results for Rpo5Mt-Std_20091202.gdb.

To check the use of the .gdb in GenMAPP we created an expression dataset with our microarray data. The creation of an Expression Dataset was relatively easy to complete. The first run through there were a few mistypes in the expression dataset manager, but once fixed it ran through MAPPfinder smoothly. This means that the .gdb works with GenMAPP/MAPPFinder and can produce results for different strains of *Mycobacterium tuberculosis*.

In order to make sure that it is possible to put a gene on the MAPP using the GeneFinder window, the first thing to do was open GenMAPP and place a gene box on the main window by using the “gene” button. Once double clicked, a backpage opens up and the gene ID is entered in as well as the option “OrderedLocusName.” This opens up a window with all of the cross-referenced ID’s for that gene. This was performed with our database and the cross-referenced ID’s were there. This process was also used for drawing the MAPP of fatty-acid metabolic pathway.

When we loaded the .txt file into GenMAPP we got 84 exceptions out of 4750 IDs total, meaning that 4666 IDs were imported. The error message, for the ID’s that were not imported, in the EX.txt file reads "Gene not found in OrderedLocusNames or any related system." This is a lower number of exceptions, meaning that our database worked to create a MAPP with our expression dataset.
The MAPP turned out well and had a complete tree with the GO terms for each gene involved in the different pathways. It was easy to find the top 10 GO terms for the Increased criteria and the Decreased criteria. There were nested pathways and each GO term was colored depending on how it fell under the criteria created in the Expression Dataset Manager. The Increased criteria were given the color red and Decreased criteria were given the color blue.

When clicking a Gene Ontology term from MAPPFinder, the MAPP file opens up. This could be seen in our MAPP. The map file is colored with the increased, decreased, or no criteria met colors. Each gene is either red (Increased) blue (decreased) or gray (no criteria met). This helps to explain the expression of each gene that is related to a specific go term.

Database gene ID identification and Query

Figure 2c describes the IDs found with the various queries ran on the finished database (gdb) file. IDs were searched in the XML, SQL, TallyEngine and gdb to keep track of the IDs that need to be accounted for between programs and the database itself.

Query type	Match (XML file)	SQL	TallyEngine	gdb
	MT: 2910	MT: 2905	Ordered Locus: 6893	
	MTCI: 41	MTCI: 41	ORF: 1550	
	MTCY: 1205	MTCY: 1203		
	MTV: 227	MTV: 226		
	Rv: 4066	Rv: 4062		
	Outlier IDs: 7			
Total	**8456**	**8437**	**8443**	**8450**

Figure 2c: Match table with query results for different gene IDs and file types. IDs were searched in the XML file using java.jar queries. SQL searches were done in PGAdmin III resulting in smaller ID field. TallyEngine was done with GenMAPP Builder, with a total comprised of ordered locus and ORF genes (customized). The gdb total is also shown.

ID sub-type	**Query**
Rv	**Rv[0-9]{4}[ABc]?(\.[0-9])?**
MTCY	**MTCY[0-9A-Z]+\.[0-9][0-9]c?**
MT	**MT[0-9]{4}(\.[0-9][0-9]?)?**
MTV	**MTV[0-9][0-9][0-9]\.[0-9][0-9][0-9]?c?**
MTCI	**MTCI[0-9][0-9]?[0-9]?[AB]?\.[0-9][0-9]c?**
Outlier IDs	**Picked out individually**
Sample query for the Rv	**select count(*) from genenametype where(type='ordered locus' or type = 'ORF') and value ~ 'Rv[0-9]{4}[ABc]?(\.[0-9]?)?';**

Figure 2d: Queries. A diagram of each of the used queries is listed above for each of the ID subtypes.

MCB1222.32c	MTC22G8.22	MTY16F9.01	MYV014.28	u0002e	u0002kc	*55c

Figure 2e: The Outlier IDs. IDs that did not adhere to particular subgroups (ie. Rv) are listed here, they did not get pulled in by the Match query (terminal).

Gene IDs in the XML

As expected, every ID in the gdb is located in the XML, since GenMAPP Builder takes IDs out of the XML, and the number of IDs identified in the XML is the largest. However, there were some gene IDs that were located in the XML, but were not being picked up by our java.jar queries done in terminal for the match. Java.jar was unable to match these IDs (figure 2f), because they did not fit into the query appropriately. Therefore a more comprehensive query needs to be generated to include the IDs listed in the figure 2f.

Rv	MTCY	MT
Rv2307.2	MTCY338.11Bc	MT3573.15
Rv2307.4		
Rv2306.2		
Rv3770.2		
Rv3274.2		
Rv3224.3		

Figure 2f: The IDs not identified by the java.jar query.

Gene IDs in the XML Source not Found in the Postgres

In order to truly identify the IDs that are *missing* from the Postgres, the queries need to be double-checked with Microsoft Excel LOOKUP and MATCHUP functions to generate an exhaustive list of the IDs missing. The SQL queries were different in execution from those done with java.jar. In other words, java.jar commands in terminal did not lead to the replicated results; the queries are listed above (Figure 2d). SQL queries should therefore be customized to create an authentic representation of the full set of gene IDs in Postgres. This can be done retroactively or proactively in relation to the Microsoft Excel LOOKUP and MATCHUP functions. The new ID counts can then be compared to the GenMAPP Gene Database IDs.

Gene IDs in the Postgres but not in the GenMAPP Gene Database (gdb)

There are currently 8450 IDs in the gdb, after subtraction of 55c from the figure, which is greater than the number of IDs in the SQL query. However, identification of the gene IDs in postgres requires more computer lab time and resources. Upon trouble-shooting, the queries will lead to an efficient means to compare all of the values from each of the different gene ID quantities. A list of all of the IDs can be accumulated with Microsoft Access, and another Microsoft Excel LOOKUP and MATCHUP round would yield comprehensive results. An exhaustive list can then be made to work on identifying the causation between ID quantity discrepancies.

Future GenMAPP Builder Coding

GenMAPP Builder is an elegant program, capable of placing vast amounts of information into a usable system for microarray analysis. The resilience of the program has led to a high yield of correctly processed gene IDs. The species profile for *Mycobacterium tuberculosis* has led to more customized results, as seen in figure 2a, b, c and d. After a full, exhaustive list is diagrammed for the IDs missing between group fields, it will be possible to fine tune the java programming in GenMAPP Builder. The first update to the GenMAPP Builder would address the misappropriation of gene ID 55c, which exists in the XML file for MT as <fullName>Rv3346/**55c** fusion protein</fullName>. By finding the search statements in the java programming (with the appropriate tags), it will be possible to customize a search function describing the Rv3346 as an exceptional ID that has a slash in its name. This customization must be made particularly for Rv3346/**55c,** since XML files have a great number of slashes. An edit of this nature to the java would not be difficult after finding the appropriate search function listing in java.

An identification of the other missing IDs—after appropriate ID minding—would require knowledge of the tagging used for each exceptional ID in the XML. That is, finding the whereabouts of each missing ID from the gdb could be done with a simple search through the XML text. After finding each ID, search functions—as previously described in this section—can be added to the GenMAPP Builder java. Before code commitment, however, it is essential to run an import/export cycle with the edited version, so as not to disrupt the opensource code for other coders.

DNA Microarray analysis

The results from GenMAPP were then filtered according to Criterion 00, the increased genes and Criterion 01, the decreased. This was compiled into an expression dataset and set to specific criteria. The increased values had a P value < 0.002 and a Z score of less >2. This yielded a result of 18 significantly increased genes with functions ranging on fatty acid metabolic process, fatty acid biosynthetic process, and translation. Likewise the decreased criteria was set at a P value of < 0.05, and a Z score of >2. This also yielded a result of 18 significantly decreased genes, yet these functions were on DNA binding, DNA metabolic process, DNA integration and other binding processes.

18	GOID	GO Name	GO Type	Number	Number	Number	Percent	Percent	Number	Number	Number	Percent	Percent	Z Score
19	32991	macromolecular complex	C	0	0	0	0	0	31	121	123	25.61983	98.37399	5.184
20	6631	fatty acid metabolic process	P	5	14	14	35.71429	100	12	30	30	40	100	5.051
22	6633	fatty acid biosynthetic process	P	7	14	14	50	100	7	14	14	50	100	4.631
23	8610	lipid biosynthetic process	P	12	29	29	41.37931	100	18	62	62	29.03226	100	4.53
24	44444	cytoplasmic part	C	0	0	0	0	0	22	83	85	26.50602	97.64706	4.521
25	15078	hydrogen ion transmembrane transporter activity	F	3	5	5	60	100	7	15	15	46.66667	100	4.384
26	6091	generation of precursor metabolites and energy	P	0	0	0	0	0	20	75	75	26.66667	100	4.335
28	6629	lipid metabolic process	P	5	32	32	15.625	100	25	104	104	24.03846	100	4.266
29	43234	protein complex	C	0	1	1	0	100	16	61	61	26.22951	100	3.789
30	44424	intracellular part	C	0	0	0	0	0	60	360	363	16.66667	99.17355	3.595
31	6412	translation	P	21	95	97	22.10526	97.93814	21	95	97	22.10526	97.93814	3.459
32	44267	cellular protein metabolic process	P	0	2	2	0	100	31	164	166	18.90244	98.79518	3.264
33	5198	structural molecule activity	F	1	1	1	100	100	16	57	59	28.07018	96.61017	4.106
34	4312	fatty-acid synthase activity	F	2	2	2	100	100	5	10	10	50	100	3.911
35	3735	structural constituent of ribosome	F	15	55	57	27.27273	96.49123	15	55	57	27.27273	96.49123	3.842
36	6066	alcohol metabolic process	P	0	2	2	0	100	15	58	59	25.86207	98.30508	3.602
37	5840	ribosome	C	15	58	60	25.86207	96.66666	15	58	60	25.86207	96.66666	3.602
38	5737	cytoplasm	C	34	261	262	13.02682	99.61832	54	334	337	16.16767	99.10979	3.131

Figure 4: List of genes with increased expression.

GOID	GO Name	GO Type	Number	Number	Number	Percent	Percent	Number	Number	Number	Percent	Percent	Z Score	Permute
32196	transposition	P	2	3	3	66.66666	100	10	21	21	47.61905	100	5.248	0
6310	DNA recombination	P	8	29	29	27.58621	100	16	47	47	34.04255	100	4.936	0
6313	transposition, DNA-mediated	P	9	20	20	45	100	9	20	20	45	100	4.751	0
4803	transposase activity	F	8	18	18	44.44444	100	8	18	18	44.44444	100	4.431	0
3677	DNA binding	F	55	300	306	18.33333	98.03922	55	305	311	18.03279	98.07074	3.899	0
6259	DNA metabolic process	P	2	7	7	28.57143	100	25	127	128	19.68504	99.21875	3.021	0.001
3676	nucleic acid binding	F	10	61	62	16.39344	98.3871	64	416	422	15.38461	98.5782	2.811	0.008
15074	DNA integration	P	5	13	14	38.46154	92.85714	5	13	14	38.46154	92.85714	3.081	0.015
6807	nitrogen compound metabolic process	P	1	16	16	6.25	100	97	709	716	13.68124	99.02235	2.271	0.016
43283	biopolymer metabolic process	P	0	0	0	0	0	87	613	619	14.1925	99.03069	2.516	0.017
3700	transcription factor activity	F	23	131	132	17.55725	99.24242	23	131	132	17.55725	99.24242	2.284	0.018
43170	macromolecule metabolic process	P	0	0	0	0	0	90	641	647	14.04056	99.07264	2.452	0.021
9069	serine family amino acid metabolic process	P	0	1	2	0	50	5	17	18	29.41176	94.44444	2.347	0.027
43176	amine binding	F	0	0	0	0	0	7	27	27	25.92593	100	2.391	0.032
16597	amino acid binding	F	3	12	12	25	100	7	27	27	25.92593	100	2.391	0.032
16903	oxidoreductase activity, acting on the aldehyde	F	0	0	0	0	0	6	24	24	25	100	2.11	0.037
6793	phosphorus metabolic process	P	0	0	0	0	0	13	66	66	19.69697	100	2.154	0.048
6796	phosphate metabolic process	P	0	1	1	0	100	13	66	66	19.69697	100	2.154	0.048

Figure 5: List of genes with decreased expression (downregulated).

The MAPP of the fatty acid metabolic pathway was drawn in GenMAPP. The final result is seen in figure 5. The pathway starts with Hexadecanoate (fatty acid) and the first gene to be expressed is fadD15. The fatty acid becomes Hexa-decanoyl-CoA and the gene fadE1 is expressed. Genes echA1 and fadB are expressed and when gene ltpl is expressed, CoA is a byproduct. All of these genes are expressed five more times until fadA2 becomes expressed as well. In the last leg of the pathway, fadE7 is also expressed. The end of this pathway produces Acetyl-CoA, which goes into the Citrate Cycle and alanine and aspartate metabolism. Other parts of this pathway are the long-chain fatty acid using gene mbtM to become long-chain-acyl-[acyl-carrier protein]. Alkane is using genes alkB, adhE1, and Rv0147 to end up with a Hydroxy fatty-acid. Overall in the fatty acid metabolic pathway, genes echA1 and Rv0147 are decreased and gene fadA2 is increased. The other genes used did not meet criteria.

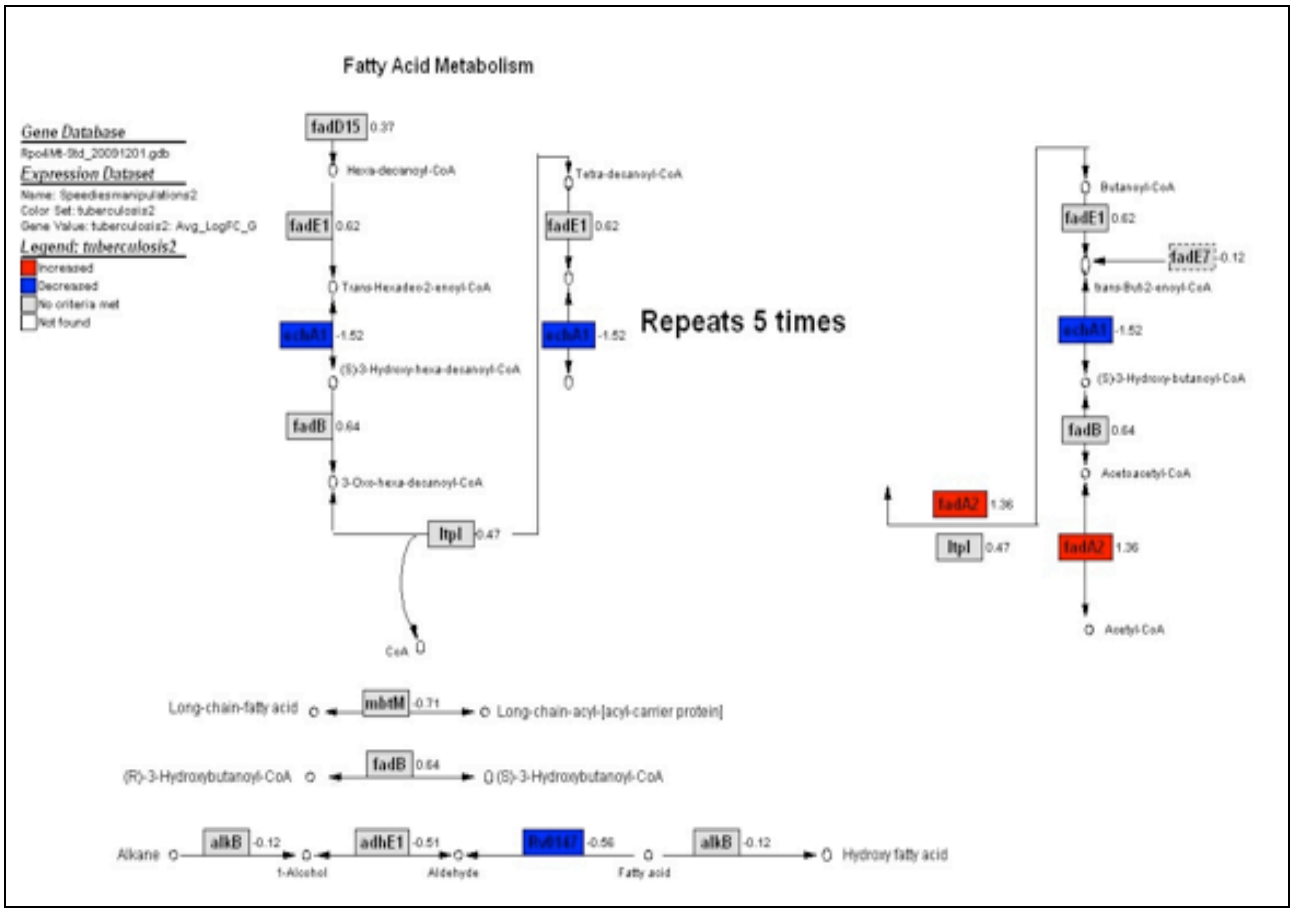


Figure 6: Finalized .MAPP of Fatty Acid Metabolism for *Mycobacterium tuberculosis* strain H37Rv.

Discussion

The results from MAPPFinder showed significant change in gene expression in the pathogenic strain G compared to the laboratory strain H37Rv. Through MAPPfinder we analyzed the dataset that was retrieved from the Stanford Microarray paper (http://smd.stanford.edu/cgi-bin/exptsets/viewExptSets.pl?exptset_no=2885&del=no). The laboratory strain H37Rv was chosen by the research team because it is the most stable strain and there was less possibility of the genome sequence changing. As a result was the best strain to compare pathogenic strains. Based on the TTEST results, Strain G had the most change in gene expression compared to H37Rv.

Based on the results from MAPPfinder strain G did show an increase in metabolic activity and a decrease in DNA synthesis and replication processes. As the genome paper (Cole et. al 1998) discussed, the cell wall of Mycobacterium tuberculosis is the main defense system against antibiotics because of its hydrophobic state, and so this makes sense that our data shows the bacterial cell is increasing the expression of genes that aid in this process. Cole's studies also describe how important the degradation of the host-cell membrane is to the Mycobacterium tuberculosis in order to build up its own cell wall. It can break down lipids and use enzymes in the host-cell to create Acetyl-CoA,

which then gets made into energy for the bacterial cell to use during the increase of production of its cell wall (Ryan 2004). The fact that the top 10 GO terms, from our MAPPfinder results, for an increase in gene expression were having to do with lipid metabolic processes as well as one for hydrogen ion transmembrane transporter activity, presents Strain G has activity in the process of building it's cell wall. This is in comparison to the gene expression of the lab strain H37Rv, so we can conclude that Strain G is more pathogenic because it spends more time building up its cell wall, which makes it even more resistant to antibiotics.

Another gene that is being increased in expression is for generation of precursor metabolites and energy. This could mean that the bacterial cell is using genes that are leading the development of energy and metabolites, fuel for the cell. This means that, compared to the H37Rv strain, this strain is using more energy and fuel and so is focusing more on the build up of its cell wall. Strain G is a more virulent strain than H37Rv and in order to combat it, the genes in cell wall production—those found in the increase criteria—should be targeted. Perhaps there is a way to decrease the expression of these genes and so antibiotics can break down the cell walls and kill the infectious bacterial cell.

References

Camus, JC, Pryor MJ, Medigue Cole. "Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv" in Microbiology vol 148, 2867-73

Cole, S.t, et al. "Deciphering the Biology of *Mycobacterium tuberculosis* from the complete gemone sequence." 1998 Nature vol 393, 537-544

Murray PR, Rosentha KS, Faller P, Medical Microbiology Academic Press 2005

Wooldridge, K (editor) Bacterial Secreted Proteins: Secretory Mechanism and Role in Pathogenesis. Academic Press 2009

Gagneux, S (editor) Strain Variation and Evolution. Mycobacterium: Genomics and Molecular Biology. Caister Academic Press 2009

Martinko, Madigan M. (editor) Brock Biology of Microorganisms (11th Ed) Prentice Hall 2005.

Reddy JR, Kwang J Lechtenberg, KF Khan. "An Immunochromatographic serological assay for the diagnosis of *Mycobacterium tuberculosis*" In Comparative Immunological Infectious Disease Vol. 25 72-27

Gao, Qian, et al. "Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates" 2005 in Microbiology vol 151, 5-14

Dionisio JDN, Dahlquist KD. “Improving the computer science in bioinformatics through open source pedagogy.” *ACM SIGCSE Bulletin* 40(2):115-119, June 2008. http://portal.acm.org/citation.cfm?id=1383602.138”3648

Open Bioinformatics Foundation (2006). http://www.open-bio.org.

XMLPipeDB home page. http://xmlpipedb.cs.lmu.edu.

XMLPipeDB SourceForge project site. http://sourceforge.net/projects/xmlpipedb