Staphylococcus aureus subsp. aureus MRSA252: understanding the genome through the application of a strain specific database

Andrew Hirning, Kelly Parks, and Michael Piña
Department of Biology
Loyola Marymount University

Introduction:

Staphylococcus aureus is an opportunistic bacterium and one of the major causes of community-acquired and hospital-acquired infections.[1] It produces numerous toxins and is resistant to practically all antibiotics, most notably meticillin and vancomycin.[1] MRSA, or meticillin resistant Staphylococcus aureus, has a very complex genome. It has the ability to acquire genes by lateral gene transfer from distantly related organisms, and it can adapt to environmentally selective pressures, such as antibiotics and the human immune system. These characteristics make MRSA not only diverse and robust, but they are also the driving force for its resistance abilities.[1]

The sequencing of the genome revealed important information about the cell wall biosynthesis of MRSA, and how it contributes to its antibiotic resistance. Traditional penicillin-binding proteins in the cell walls of bacteria, for the purpose of peptidoglycan (a component of the cell wall) synthesis, have a gycotransferase domain at the N-.[1] Interestingly, the sequencing showed that, in MRSA, these proteins either lacked the domain or had no glycosyltransferase activity, which is the reason for its resistance to late stage cell wall biosynthesis inhibition by penicillin derived (in our case meticillin) antibiotics.[1]

What the microarray experiment conducted by O'Neill et. al. attempts to accomplish is the foundation of a transcription signature of the inhibition of early stage CWB, instead of focusing on late stage CWB.[2] For the purpose of our project, we focused on only one condition of the inhibition of early CWB, namely the inhibition of enzyme MurA by the antibiotic fosfomycin (Figure 1). To best obtain these results, we wanted to use GenMAPP and MAPPFinder to analyze the microarray data, but were unable to because no database existed for the specific strain of *S. aureus*, namely MRSA252, in our microarray data.

XMLPipeDB and GenMAPP are utilities that allow one to take publicly available data and build relational databases from that material. These tools would allow us to take MRSA252 XML data and create a specific gene database for use in GenMAPP. By using the microarray data with our strain specific database, we hope to learn what genes had their expression significantly changed in response to the presence of fosfomycin, and what this means for the inhibition of early stage CWB.

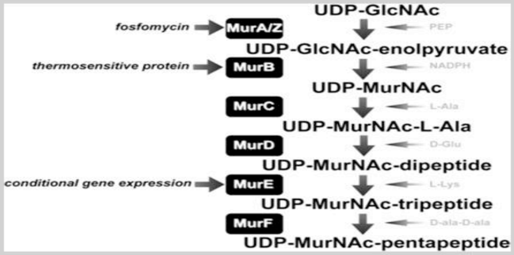


FIG. 1. The stage I cell wall biosynthesis pathway in *Staphylococcus aureus* involves the biosynthesis of UDP-MurNAc-pentapeptide from UDP-GlcNAc, mediated by the Mur enzymes. The three points at which inhibition of the pathway was achieved in this study are shown in italics. PEP, phosphoenolpyruvate.

Methods:

In order to examine the increased and decreased gene expressions seen in our microarray data we needed to create a database of *S. aureus* for use with GenMAPP. GenMAPP (Gene Map Annotator and Pathway Profiler) is a free computer application for viewing and analyzing DNA microarray and other genomic and proteomic data on biological pathways. MAPPFinder is an accessory program that works with GenMAPP and Gene Ontology to identify global biological trends in gene expression data. The GenMAPP Gene Database (file with the extension .gdb) is used to relate gene IDs on MAPPs.

To create the .gdb file, source files were acquired from online repositories. Specifically, an XML file containing all UniProt gene information for *S. aureus* MRSA 252 was acquired from Integr8, as well as a GOA file which related the annotated genes within the XML file to different ID systems.[3] Also, a OBO-XML file was obtained from the Gene Ontology (GO) Consortium which contains standardized terms for annotating genes.[4] A new database was created in PostgreSQL[5], and the database was filled with tables using a standardized SQL command distributed with GenMAPP Builder[6]. GenMAPP Builder was used to populate the tables in the PostgreSQL relational database with the data contained within the UniProt XML file, as well as the GO OBO-XML file. Once the data was imported, GenMAPP Builder processed the data for correct formatting. Once the database was filled correctly, the database could be exported into a GenMAPP compatible GDB file. This required not only the database which had been compiled from the two XML files, but also the GOA file which was acquired from Integr8. GenMAPP Builder formatted the database for use with GenMAPP.

Several techniques were used in order to ensure the validity and consistency of the various IDs across the entire export cycle of the .gdb. TallyEngine - a part of GenMAPP Builder - provided an automatic table of counts for OrderedLocusNames, RefSeq, GeneID, UniProt, and GO Terms. The microarray data analyzed referred to genes by their OrderedLocusNames IDs in the form of SAR#### or SAR####.# and thusly, OrderedLocusNames were a primary concern for checking consistency across the export process. The format for the OrderedLocusNames was originally thought to also include SAR####a and SAR####b, but it was subsequently discovered that these IDs refer to certain open reading frames and will be excluded from any future versions of the .gdb.

Another tool of the XMLPipeDB suite is the match utility. Using the command: java -jar xmlpipedb-match-1.1.1.jar "\"ordered locus\"\>SAR....(.?)(.?)\<\/name\>" < 19583.S_aureus_MRSA252_20091123.xml in a command line environment gave the number of OrderedLocusNames IDs located in the UniProt XML file. A similar query command was used within PostgreSQL as follows: select count (*) from genenametype where type = 'orderedlocus'; after the XML data had been imported in to the newly created PostgreSQL database. After the GenMAPP compatible database was exported with GenMAPP Builder, a final step included obtaining OrginalRowCounts information from the .gdb (viewed in Microsoft Access).

While GenMAPP Builder is capable of generating a gene database for almost any organism without modification, specific changes were made to GenMAPP Builder as a program so that it could make species specific changes to the gene database. Source code for GenMAPP Builder was checked out from the XMLPipeDB project page at SourceForge[7]. Coding environment was Eclipse, with Subclipse add-on for code checkout. A species profile was added to the program to facilitate the creation of species specific database tables if needed. No tables were needed, but the species profile was necessary to allow for other modifications. A change were also made to TallyEngine, GenMAPP Builder's internal counting script. This modification added the OrderedLocusNames to the IDs counted by TallyEngine. As mentioned above, this allows for the validation of the gene database, in terms of the number of different IDs that are in the database. Also, *S. aureus* MRSA 252 was included in the catalog of known species in GenMAPP Builder, to allow for accelerated development of new gene databases at the advent of new protein data.

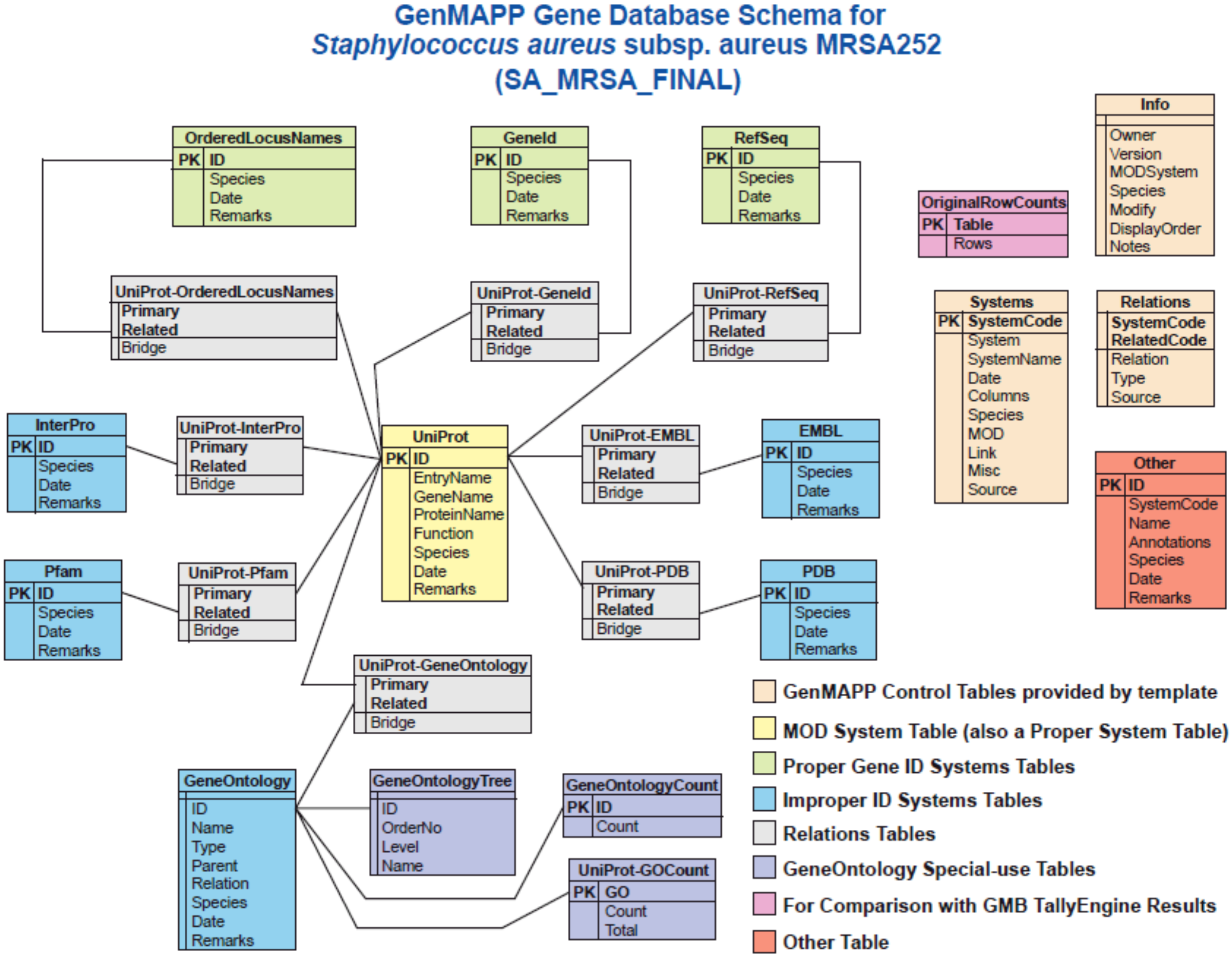
Simultaneous to the development of the strain specific gene database, the raw microarray data obtained from the experiment performed by O'Neill et. al. required processing and analysis before being into GenMAPP. As mentioned prior, the initial microarray data had three separate sets of conditions: the inhibition of MurA, MurB, and MurE enzymes involved in early stage CWB. For our project, we focused on just the set of data for the inhibition of MurA by the antibiotic fosfomycin. Our condition was cultured in triplicate and hybridized in duplicate for a total of 36 original files. However, our microarray data presented Cy3 and Cy5 data separately, so we had to create an organization for how these two sets came together for our ratios. We made the assumption that the files that ended with the same digit in the tens place for the same group of data (ex. SA14_080_Cy3_38 and SA14_080_Cy5_37) were the corresponding Cy3 and Cy5 data. 36 differences between signal and background medians (signal - background) were then processed for each Cy3 and Cy5 file. These differences were then processed using the ratio formula (experimental ratio (with fosfomycin) divided by the control ratio (w/o fosfomycin)). Next, we found the log2 values for the ratios. Then we found the average and standard deviation for each column of log2 values. Next the scaled centered values were found for each replicate. After that, the average LogFC for all the replicates (in our case all 18) were found. Lastly, we calculated TSTAT and PValue statistics based on the averages and standard deviations for all the replicates together. After all the statistics were performed, all important information was arranged in a final sheet ready to be imported into GenMAPP. This final step included adding a system code column after the Gene IDs and formatting the Gene IDs so that they matched the OrderedLocusNames in our gene database (MRSA252-#### to SAR####). The final processed data, named Final Sheet, was then imported into GenMAPP using the Expression Dataset Manager. Here, we were able to create a colorset, called s_aureus_colorset. The colorset established a set of criteria by which to color genes whose expressions were significantly changed. These criteria included: increased gene expression ([Avg_log_FC_all] > 0.25 AND [pvalue] < 0.05) to be colored red and decreased gene expression ([Avg_log_FC_all] < -0.25 AND [pvalue] < 0.05) to be colored purple.

Once the Expression Dataset (Final Sheet.gex) had been imported, the colorset had been created, and the correct .gdb had been selected (because of the timing of the project, all our microarray data and results were processed and found with the Sa-Std_20091123.gdb), we generated new results using MAPPFinder, which provided us with a list of GO terms related to genes that were significantly expressed either more (sa_20091203-Criterion0-GO) or less (sa_20091203-Criterion1-GO). Since the number of GO terms was large, we filtered them down to a workable number (between 15-20) with the filters z value > 2 and permute p <0.05 for both. Additionally, for increased criterion, GO terms were filtered with the number of genes changed for a GO term being greater than or equal to 4 and less than 100, and the percentage of genes changed for a GO term being greater than 35%. The decreased data was additionally filtered with the number of genes changed being greater than or equal to 4 and less than 100. With these filtered GO terms, a relevant biological pathway, affected at the genetic level, was chosen, and a visual representation of that pathway, or a MAPP, with the genes colored by their expression values was created using GenMAPP.

Results:

A visual schema edited with Adobe InDesign was used in order to show the relations between the various tables in the .gdb. Since the database created was UniProt-centric, a table for

UniProt is shown in the middle of the schema. It is important to note that not all relations in the database are represented here (Figure 2).



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

FIG. 2 GenMAPP Gene Database Schema for the final database.

The various counts of OrderedLocusNames reconciled across the XML, TallyEngine, and PostgreSQL database, which came to a total of 2659 IDs, but did not in the final .gdb, which came to a total of 2690. This is ok at this time because it is acceptable to have more IDs in the .gdb than in the XML data, but not vice versa. This incongruity is likely due to the fact that GenMAPP Builder is still sorting these OrderedLocusName look-a-likes into this table. In addition, OriginalRowCounts agreed with the other IDs in TallyEngine. There were 4 other IDs located throughout the XML that were not located inside ordered locus tags. Research of these IDs confirmed that they were not valid and our group decided not to include them in future versions of the .gdb. In order to achieve this, GenMAPP Builder will need to be altered to only recognize those IDs specifically inside ordered locus tags because it is still accepting not only the 4 invalid IDs, but other similar IDs that are not in fact OrderedLocusNames. All ID checks can be seen here in the respective figures (Figures 3-6).

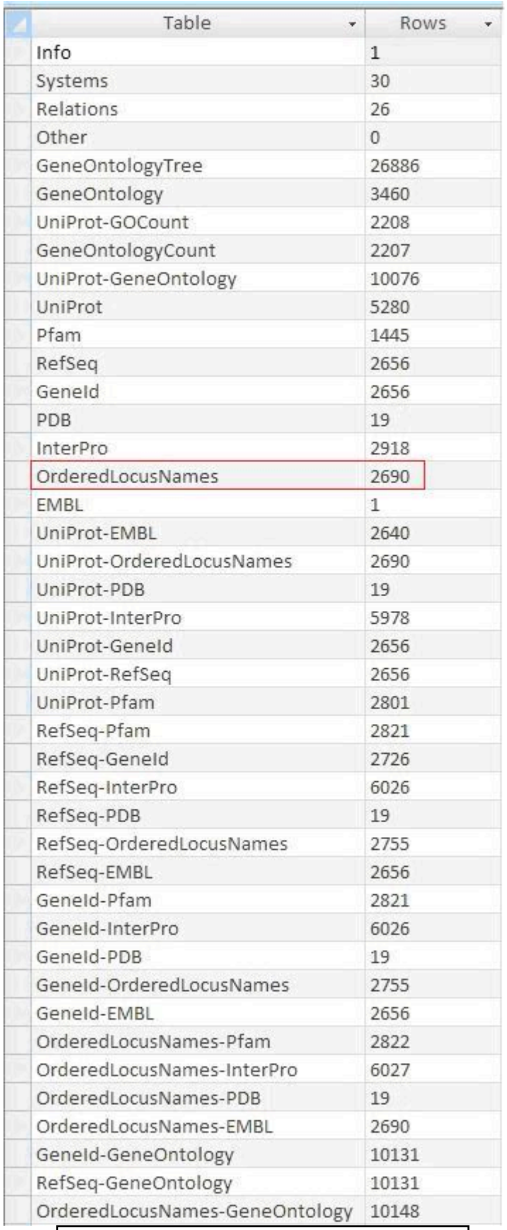
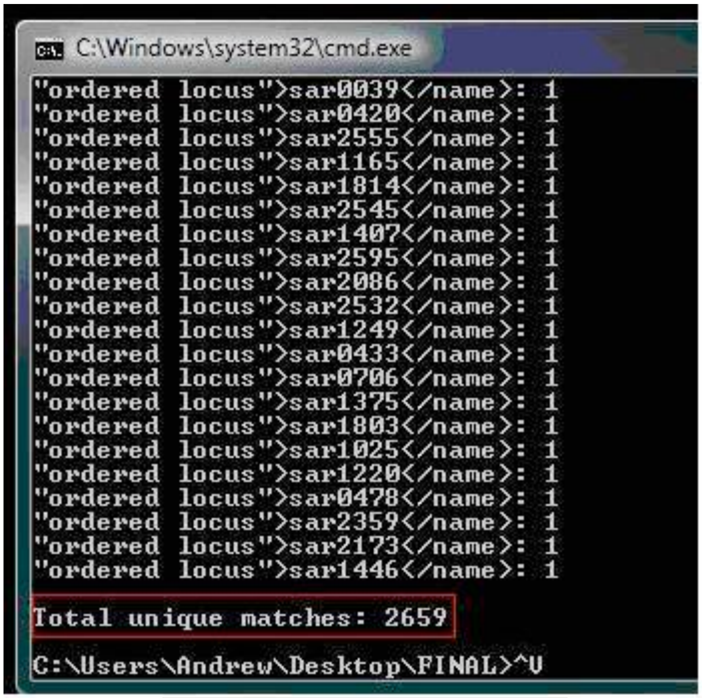


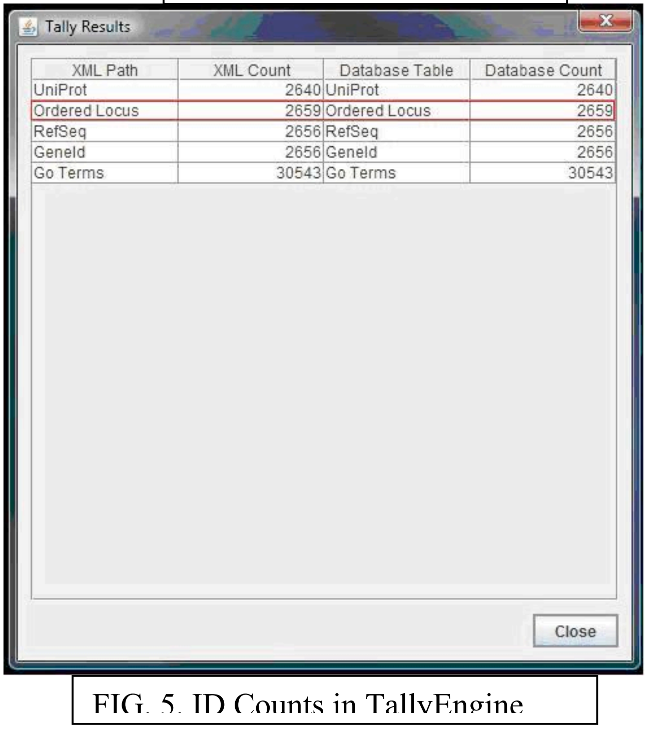
Table	Rows
Info	1
Systems	30
Relations	26
Other	0
GeneOntologyTree	26886
GeneOntology	3460
UniProt-GOCount	2208
GeneOntologyCount	2207
UniProt-GeneOntology	10076
UniProt	5280
Pfam	1445
RefSeq	2656
GeneId	2656
PDB	19
InterPro	2918
OrderedLocusNames	2690
EMBL	1
UniProt-EMBL	2640
UniProt-OrderedLocusNames	2690
UniProt-PDB	19
UniProt-InterPro	5978
UniProt-GeneId	2656
UniProt-RefSeq	2656
UniProt-Pfam	2801
RefSeq-Pfam	2821
RefSeq-GeneId	2726
RefSeq-InterPro	6026
RefSeq-PDB	19
RefSeq-OrderedLocusNames	2755
RefSeq-EMBL	2656
GeneId-Pfam	2821
GeneId-InterPro	6026
GeneId-PDB	19
GeneId-OrderedLocusNames	2755
GeneId-EMBL	2656
OrderedLocusNames-Pfam	2822
OrderedLocusNames-InterPro	6027
OrderedLocusNames-PDB	19
OrderedLocusNames-EMBL	2690
GeneId-GeneOntology	10131
RefSeq-GeneOntology	10131
OrderedLocusNames-GeneOntology	10148

FIG. 3. ID Counts in Access



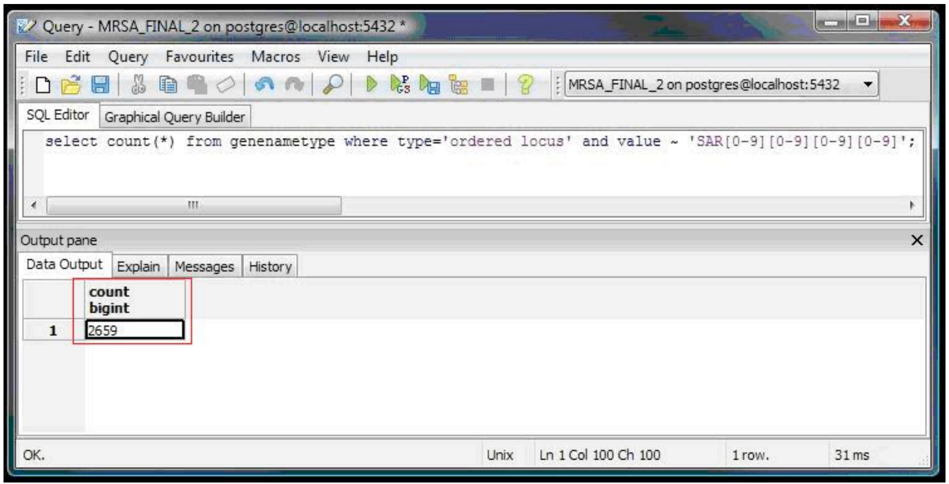
```
"ordered locus">sar0039</name>: 1
"ordered locus">sar0420</name>: 1
"ordered locus">sar2555</name>: 1
"ordered locus">sar1165</name>: 1
"ordered locus">sar1814</name>: 1
"ordered locus">sar2545</name>: 1
"ordered locus">sar1407</name>: 1
"ordered locus">sar2595</name>: 1
"ordered locus">sar2086</name>: 1
"ordered locus">sar2532</name>: 1
"ordered locus">sar1249</name>: 1
"ordered locus">sar0433</name>: 1
"ordered locus">sar0706</name>: 1
"ordered locus">sar1375</name>: 1
"ordered locus">sar1803</name>: 1
"ordered locus">sar1025</name>: 1
"ordered locus">sar1220</name>: 1
"ordered locus">sar0478</name>: 1
"ordered locus">sar2359</name>: 1
"ordered locus">sar2173</name>: 1
"ordered locus">sar1446</name>: 1

Total unique matches: 2659

C:\Users\Andrew\Desktop\FINAL>^V
```

FIG. 4. ID Counts in Match



XML Path	XML Count	Database Table	Database Count
UniProt	2640	UniProt	2640
Ordered Locus	2659	Ordered Locus	2659
RefSeq	2656	RefSeq	2656
GeneId	2656	GeneId	2656
Go Terms	30543	Go Terms	30543

FIG. 5. ID Counts in TallyEngine



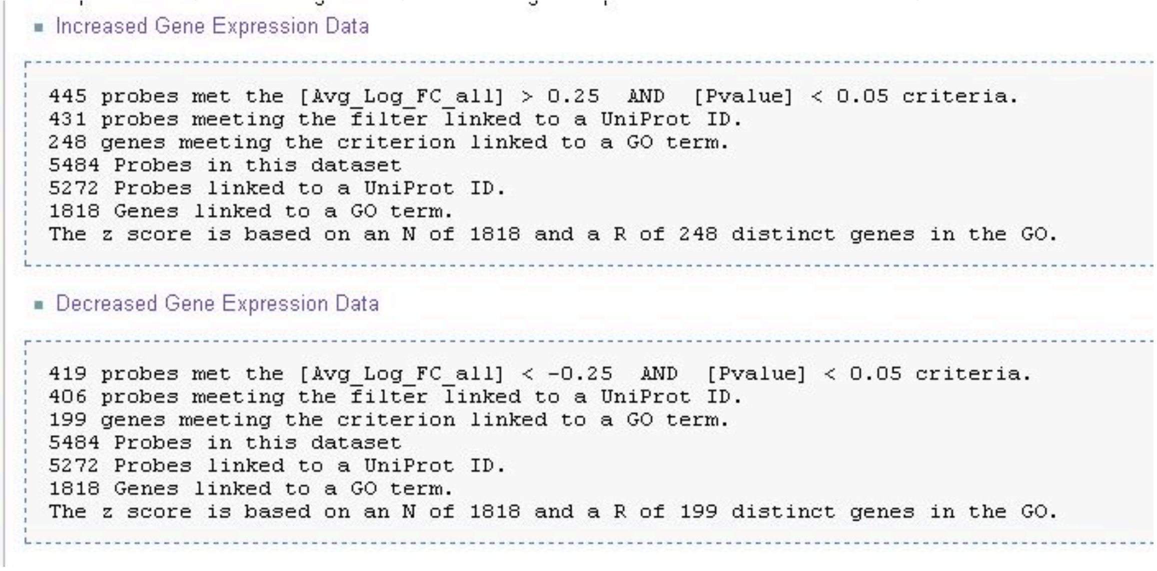
```
select count(*) from genenametype where type='ordered locus' and value ~ 'SAR[0-9][0-9][0-9][0-9]';
```

	count bigint
1	2659

FIG. 6. ID Counts in SQL

In the final GDB, there is no link to the Model Organism Database when one searches for an OrderedLocusName in GenMAPP. This may be due to a discrepancy in how the species is identified within the XML and how GenMAPP Builder determines species.

As for the results of the processed microarray data, 445 genes met the criteria [Avg_log_FC_all] > 0.25 AND [pvalue] < 0.05 for increased gene expression, and 419 genes met the criteria [Avg_log_FC_all] < -0.25 AND [pvalue] < 0.05 for decreased gene expression out of a total 5484 genes in the dataset. This means that roughly about 16% of the total genes in the dataset were significantly changed either more or less in expression (Figure 7).



▪ Increased Gene Expression Data

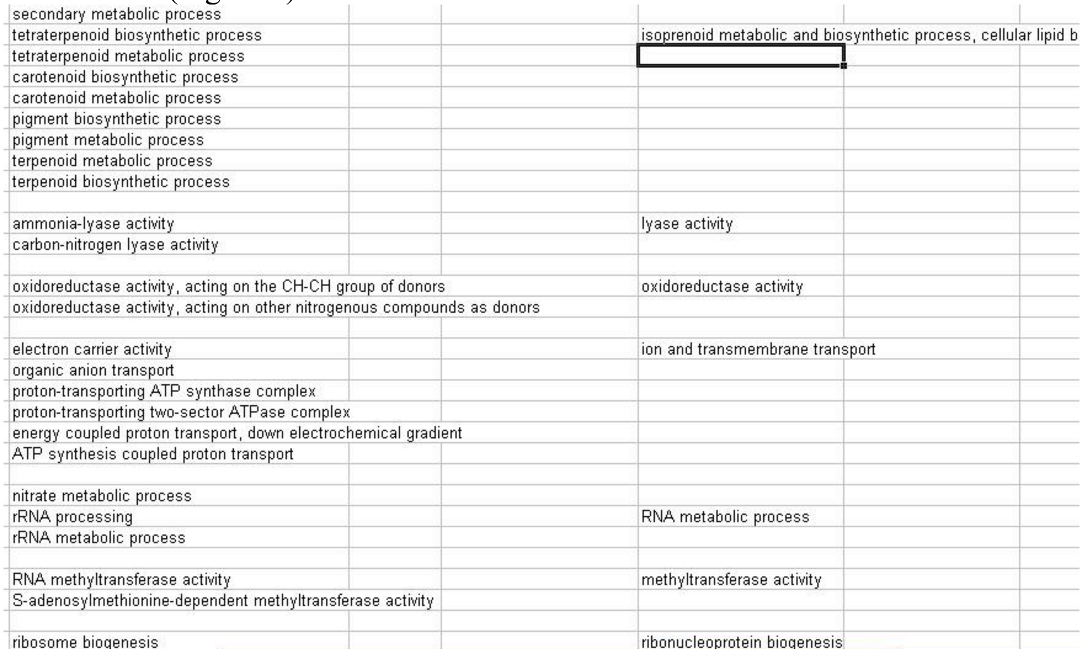
```
445 probes met the [Avg_Log_FC_all] > 0.25  AND  [Pvalue] < 0.05 criteria.
431 probes meeting the filter linked to a UniProt ID.
248 genes meeting the criterion linked to a GO term.
5484 Probes in this dataset
5272 Probes linked to a UniProt ID.
1818 Genes linked to a GO term.
The z score is based on an N of 1818 and a R of 248 distinct genes in the GO.
```

▪ Decreased Gene Expression Data

```
419 probes met the [Avg_Log_FC_all] < -0.25  AND  [Pvalue] < 0.05 criteria.
406 probes meeting the filter linked to a UniProt ID.
199 genes meeting the criterion linked to a GO term.
5484 Probes in this dataset
5272 Probes linked to a UniProt ID.
1818 Genes linked to a GO term.
The z score is based on an N of 1818 and a R of 199 distinct genes in the GO.
```

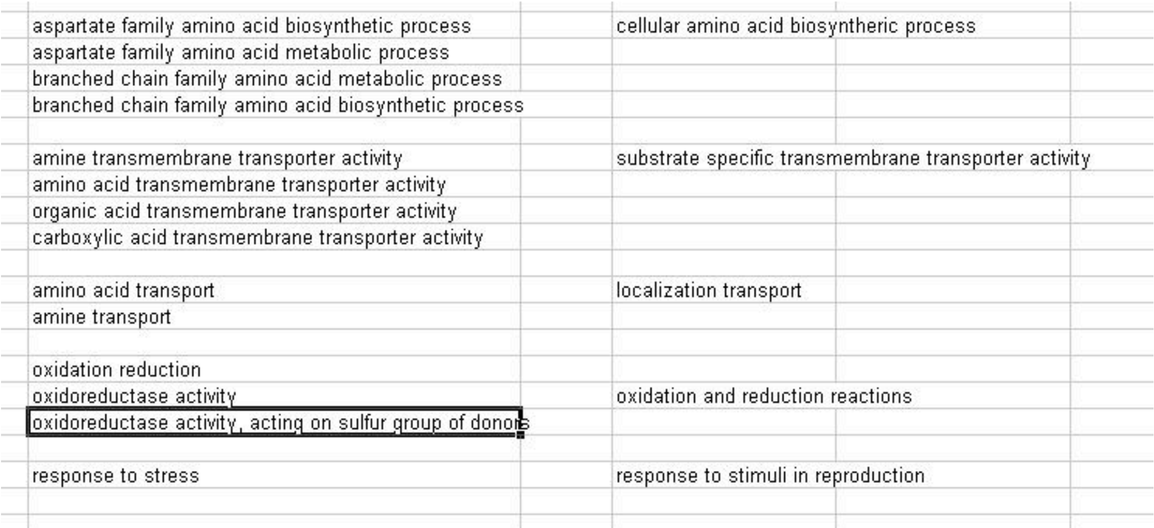
FIG. 7. Counts of genes significantly increased and decreased in the database.

Our MAPPFinder results provided us with a list of GO terms that were related to those genes that matched the criteria for both increased and decreased gene expression. By filtering the GO terms down even more to reflect only those terms that were most affected, how MRSA's early stage CWB was affected by the presences of fosfomycin could be better understood. Terms related to those genes that increased in gene expression fell within the categories of isoprenoid metabolic and biosynthetic process, lyase activity, oxidoreductase activity, ion and transmembrane transport, RNA metabolic process, methyltransferase activity, and ribonucleoprotein biogenesis (Figure 8). Terms related to those genes that decreased in gene expression fell within the categories cellular amino acid biosynthetic process, substrate specific transmembrane transporter activity, localization transport, oxidation and reduction reactions, and response to stimuli (Figure 9).



secondary metabolic process	
tetraterpenoid biosynthetic process	isoprenoid metabolic and biosynthetic process, cellular lipid b
tetraterpenoid metabolic process	
carotenoid biosynthetic process	
carotenoid metabolic process	
pigment biosynthetic process	
pigment metabolic process	
terpenoid metabolic process	
terpenoid biosynthetic process	
ammonia-lyase activity	lyase activity
carbon-nitrogen lyase activity	
oxidoreductase activity, acting on the CH-CH group of donors	oxidoreductase activity
oxidoreductase activity, acting on other nitrogenous compounds as donors	
electron carrier activity	ion and transmembrane transport
organic anion transport	
proton-transporting ATP synthase complex	
proton-transporting two-sector ATPase complex	
energy coupled proton transport, down electrochemical gradient	
ATP synthesis coupled proton transport	
nitrate metabolic process	
rRNA processing	RNA metabolic process
rRNA metabolic process	
RNA methyltransferase activity	methyltransferase activity
S-adenosylmethionine-dependent methyltransferase activity	
ribosome biogenesis	ribonucleoprotein biogenesis

FIG. 8. GO terms related to genes with increased expression.



aspartate family amino acid biosynthetic process	cellular amino acid biosyntheric process
aspartate family amino acid metabolic process	
branched chain family amino acid metabolic process	
branched chain family amino acid biosynthetic process	
amine transmembrane transporter activity	substrate specific transmembrane transporter activity
amino acid transmembrane transporter activity	
organic acid transmembrane transporter activity	
carboxylic acid transmembrane transporter activity	
amino acid transport	localization transport
amine transport	
oxidation reduction	
oxidoreductase activity	oxidation and reduction reactions
oxidoreductase activity, acting on sulfur group of donors	
response to stress	response to stimuli in reproduction

FIG. 9. GO terms related to genes with decreased expression.

GenMAPP also enable us to make a visual representation of a relevant biological pathway that was affected by a change in gene expression. For our project, a MAPP depicting branched chain family amino acid biosynthesis, which was a specific GO term dealing with the synthesis of amino acids Isoleucine, Leucine, and Valine within the category of cellular amino acid biosynthetic process, was created, highlighting those genes, specifically SAR2145, SAR0559, and SAR2144, that were lowered in gene expression due to the presence of fosfomycin (Figure 10).

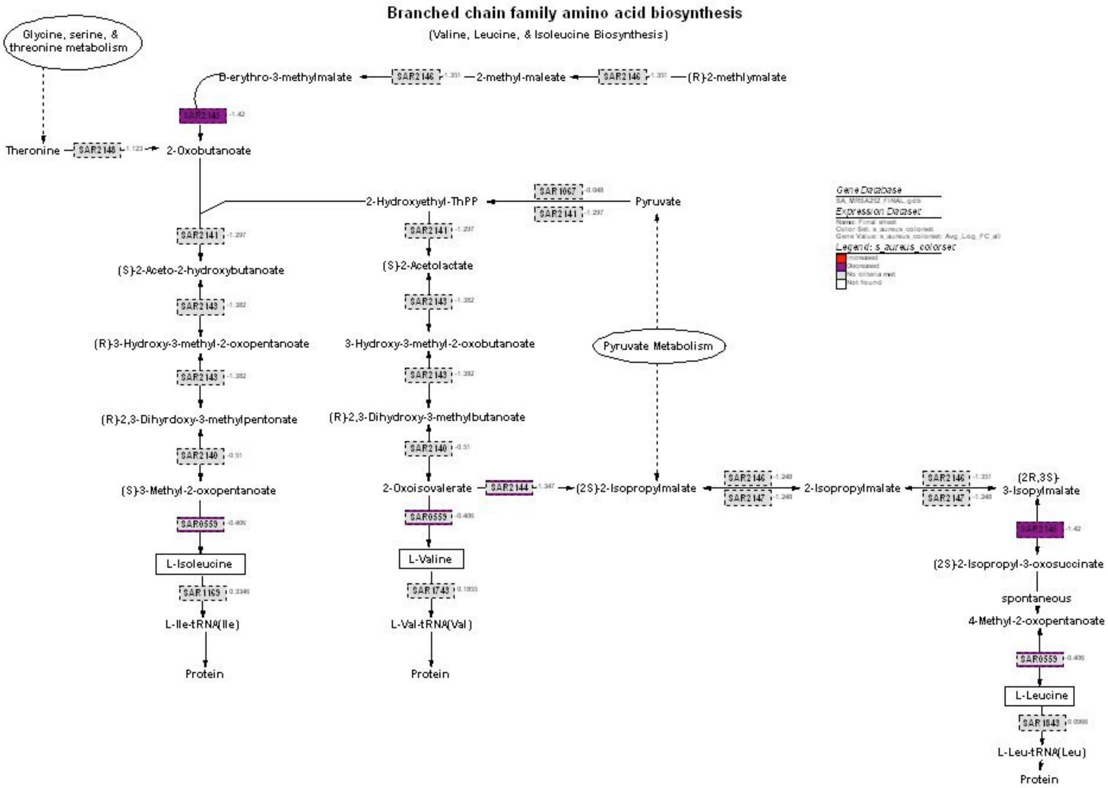


FIG. 10. GenMAPP MAPP of branched chain family amino acid biosynthesis.

Discussion:

While much time was spent configuring GenMAPP builder and preparing microarray data, the actual aim of this project was to create a gene database for *Staphylococcus aureus* MRSA 252. The creation of this database was intended to allow easier study of transcriptional and translational changes on an organism wide scale. So to conclude what we learned from this project, a comparison of our biological results will be made with the results from the associated paper for the microarray data that we used. The gene database allows for the processed microarray results to be input into MAPPFinder and the results can then be examined based on whether transcription/translation was increased or decreased. The most notable MAPPFinder results from our data included the increased expression of those genes related to GO terms ion and transmembrane transport resulting in ATP synthesis, riboneucleoprotein biogenesis, and the decreased expression of those genes related to GO terms cellular amino acid biosynthesis, substrate-specific transmembrane transporter activity, and response to stress. Analysis of these results and the definitions of the individual GO terms gives us a deeper insight into what is possibly happening to the MRSA cells being affected by fosfomycin. For instance, in the case of the data increased in gene expression, the MRSA cells may be focusing the energies on the processes of ion and transmembrane transport because of the ATP synthesis that it yields; since the cells are under attack by the antibiotic fosfomycin, they may be increasing the energy supply in order to counteract the damaging effects of the antibiotic. The increase in expression of those genes involved in the biogenesis of the ribosome and related proteins may also be for the purpose of increased production of ribosome units for the synthesis of more proteins due to those the damaged by the presence of fosfomycin, as is reflected in the decreased gene expression of amino acid biosynthesis. In terms of those genes that were downregulated related to the GO term response to stimuli, these genes may be expressed less because of the change in the state of the cell due to the disturbance to homeostasis caused by fosfomycin interaction. Because fosfomycin enters a bacteria cell through glycerophosphate transporters, decreased gene expression in those processes related to the GO term substrate-specific transmembrane transporter activity may have been affected as a result of attempting to keep fosfomycin from entering the cell. However, without further testing or experimentation, these results can only be substantiated by inference. In comparison to the results put forth by O'Neill et. al., who reported that overall little deregulation in expression of the Mur enzymes, including the MurA enzyme that we focused on, was observed, suggesting that CWB could not be inhibited in its early stage by the antibiotic fosfomycin, was supported by our results. No specific GO terms or decreased gene expression data obtained in our results from MAPPFinder pointed to the fact that there was any deregulation in enzymatic activity in the early stage of CWB. Although no new information was obtained with respect to the early stage inhibition of CWB in MRSA, we were able to accomplish the creation of a strain specific database that allowed us to look at this microarray data in a new and detailed way, revealing more that could have been studied prior without the use of GenMAPP and MAPPFinder.

References:

1. M.Kuroda, T.Ohta, I.Uchiyama, T.Baba, H.Yuzawa, I.Kobayashi, L.Cui, A.Oguchi, K.Aoki, Y.Nagai. Whole genome sequencing of meticillin-resistant Staphylococcus aureus. The Lancet, Volume 357, Issue 9264, Pages 1225-1240.

2. O'Neill, A. J., Lindsay, J. A., Gould, K., Hinds, J., Chopra, I. Transcriptional Signature following Inhibition of Early-Stage Cell Wall Biosynthesis in Staphylococcus aureus. Antimicrob. Agents Chemother. 2009: 53: 1701-1704.

3. http://www.ebi.ac.uk/integr8/

4. http://www.geneontology.org/GO.downloads.ontology.shtml

5. http://www.postgresql.org/

6. http://xmlpipedb.cs.lmu.edu/

7. http://sourceforge.net/projects/xmlpipedb/

8. http://www.geneontology.org/ontology/GO.defs

<u>Acknowledgements</u>:

As a team, we would like to thank our two professors, Dr. Dahlquist and Dr. Dionisio, for their help; without their assistance, this project would never have made it to this point. We would also like to thank the other members of the Biological Databases class, who carried out their work right along side of us, but were always open to discussion and willing to help.