Determining Changes in S*treptococcus pneumoniae* Carbohydrate Transporter Expression During Biofilm Formation

Vilgalys, T., Meilak, K., & Vreeland, A.

Biology Department, Loyola Marymount University

BIOL/CMSI 367

13 December 2013

Introduction:

*Streptococcus pneumonia*e, a gram positive bacteria, is a human pathogen involved in many pneumococcal infections. This species is most commonly known for being the major cause of pneumonia, but it is also involved in other invasive diseases such as meningitis, otitis media, conjunctivitis, and sepsis. It is one of the top ten causes of death in the United States, primarily affecting the elderly, children, and those with a compromised immune system. *S. pneumoniae* has the relatively unique ability to switch between pathogenic and avirulent phases. The bacteria can reside in the nasopharynx region without causing any disease but becomes pathogenic in some individuals. Because of this, *S. pneumoniae* is a model organism for many immunity studies. The importance of understanding *S. pneumoniae* is amplified by its tendency to become resistant to penicillin and other antibiotics. Studying this species is important not only because of the high death rate caused by it, but also because of the high rate of drug resistance. In hospital patients, one third of *S. pneumoniae* strains were resistant to at least one common antibiotic and many had developed multidrug resistance (Hoskins et al. 2001). Because of this, *S. pneumoniae* research has focused on virulence, drug resistance, and identify possible drug targets.

The *Streptococcus pneumoniae* str. TIGR4 is a highly invasive and virulent strain used in mouse studies (Aaberge et al. 1995). The strain genome is 2,160,837 base pairs and comprises 2,236 genes, values which closely resemble other *S. pneumoniae* strains such as R6 (Hoskins et al. 2001, Tettelin et al. 2001). The TIGR4 genome contains several genetic motifs which may underlie different life stages, multiple and varied cell-surface receptors to receive information from the environment, and 25 virulence genes that may only be active during specific stages of bacterial infection (Tettelin et al. 2001).

One recent focus of study has been different life stages in *Streptococcus pneumoniae.* Recently, Sanchez et al. (2011) examined the role of biofilms in the pathogenicity of the *S.*

pneumoniae and found that only planktonic infect and kill laboratory mice. Alternately, biofilm and biofilm-derived samples inhabit the lungs, but do not spread into the bloodstream or infect their hosts. Sanchez et al. also examined gene expression, comparing planktonic bacteria to biofilm samples. Biofilm production was induced in three strains of *S. pneumoniae* (TIGR4, G54, and R6) and cells were extracted at 4, 12, 24, and 48 hours to measure gene expression. Using their gene expression data, Sanchez et al. confirmed biofilm cells showed decreased expression for virulence related genes including pneumolysin and choline binding protein.

We are interested in what other gene expression changes may accompany biofilm production. However, initial attempts were limited due to the lack of a gene database for *Streptococcus pneumonaie* which would allow us to examine gene expression throughout the genome. We used XMLPipeDB and GenMAPP Builder to create a database for *Streptococcus pneumoniae* TIGR4 and then used GenMAPP and MAPPfinder to compare gene expression in biofilm and planktonic cells. Using this method, we hope to discover how *S. pneumoniae* gene expression differs between life stages.

Methods:

Creating the *Streptococcus pneumoniae* str. TIGR4 database:

Data Sources:

We downloaded the most recent version of GenMAPP builder (gmbuilder 2.0b73) from <http://sourceforge.net/projects/xmlpipedb/files/?source=navbar>. This version of GenMAPP builder includes customized code in the *Streptococcus pneumonia* TIGR4 UniProt Species Profile database to adjust Gene IDs from SP_#### (as found in the UniProt XML file) to SP#### (the format used in the microarray data) and link to

<http://www.streppneumoniae.com/gene_detail_output.asp?id=2741&name=~>, the ensemble bacteria page for each gene (the code for which are available in Appendix 1).

We also downloaded the UniProt complete proteome set for *Streptococcus pneumoniae* serotype 4 (strain ATCC BAA-334 / TIGR4) from <http://www.uniprot.org/uniprot/?query=organism%3a170187+keyword%3a181&format=*>. The downloaded XML file last updated as part of UniProt Release 2013_11 on November 13, 2013. The downloaded file name was "uniprot-organism%3A170187+keyword%3A181.xml" and it was renamed "20131118_UniProtXML_tATK_TIGR4_TPV.xml".

The gene ontology associations were downloaded from the GOA project (<http://www.ebi.ac.uk/GOA/>) as a tab-delimited text file. The *S. pneumoniae* str. TIGR4 GOA file was downloaded from <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>, was last updated 11/12/2013 at 2:49:00 PM, and had an original file name of "57.S_pneumoniae_TIGR4.goa". It was renamed "20131118_GOA_tATK_TIGR4_TPV.goa".

The gene ontology (GO) terms were found in an OBO-XML file downloaded from http://beta.geneontology.org/page/download-ontology on November 20, 2013. The file was initial titled "go_daily-termdb.obo-xml" and was renamed "20131120_OBOXML_tATK_TPV.obo-xml".

Running GenMAPP Builder:

In PostgreSQL, we created a new database titled tATK_TIGR4_2013NOV25. GenMAPP builder tables were imported into this database by running the SQL command lines contained in gmbuilder.sql. Then we opened gmbuilder-32bit.bat, connected to the database tATK_TIGR4_2013NOV25, and imported the UniProt XML, OBO-XML, and GOA files listed earlier. The database was exported to "Streptococcus_pneumoniae_TIGR4_20131125.gdb".

Validating the Database:

We validated the gene database by using several methods to count the number of Ordered Locus Gene IDs in the UniProt XML file and exported database. TallyEngine (part of gmbuilder) was used to count the number of genes in the UniProt XML and database tables. Then, using the command prompt, the program XMLpipeDB match (downloadable from: <http://sourceforge.net/projects/xmlpipedb/>) was used to query the UniProt XML for gene IDs with the name SP_[0-9][0-9][0-9][0-9]. Next, the database tables were queried using SQL and the command: *select count(*) from genenametype where type = 'ordered locus' and value ~ 'SP_[0-9][0-9][0-9][0-9]*. Original row counts for the Ordered Locus Names in the database were then calculated in Microsoft Access and compared to the other values. The systems table, ordered locus names table, UniProt table, and reference sequence table were examined by visual inspection to verify the Gene ID system.

Examining Gene Expression

Preparing Microarray Data:

The raw microarray data was downloaded from Array Express and there were a total of 18 raw data files, one for each technical replicate. Data taken from *S. pneumoniae* 4, 12, 24, and 48 hours of biofilm formation. All raw data, including gene IDs and strain information, was compiled into one Excel spreadsheet. The data was log transformed on a second worksheet, and then scaled and centered on a third. A T-statistic and P-value were calculated for each biological replicate. Any function at any step that resulted in an error, typically a divide by zero error, was replaced with a single space character to facilitate import into GenMAPP. When copying data into a new

spreadsheet, only values were pasted rather than functions. We counted the number of genes that were significant ($p<0.05$) for each time point.

Running GenMAPP and MAPPfinder:

The compiled raw data, along with the results of the statistical tests, was imported into GenMAPP using the Expression Dataset Manager and produced 333 exceptions. XMLPipeDB Match program was performed on the UniProt XML file to export gene IDs as a text (.txt) file. This file was listed next to the exceptions file in an Excel spreadsheet and find and replace was used to replace underscore with a 'no space' in order to convert ID form from SP_#### to SP####. The origin of these errors was determined using the Excel match command.

In GenMAPP, color sets were made for the 12 hour time point which became the main focus for the remainder of the project. There were two requirements in the color code: 1) a $p<0.25$ and 2) a fold change either greater than 0.25 or less than -0.25. The 12 hour time point had the following color code: green signified a significant decrease in gene expression at that time point, red signified a significant increase in gene expression, grey signified no significant difference in gene expression, and white signified that the gene was not found in the UniProt database. A MAPPFinder analysis was run, showing the gene expression at that time, and a filtered list, ordered by p-value and z-score, was generated. The gene map of the sugar transmembrane transporter activity pathway was generated from the significant results and the functions of individual genes were looked up in UniProt.

Results:

Gene Database:

In TallyEngine, the UniProt XML file and gene database were shown to have identical counts for all values and an OrderedLocus value of 2126 (Fig. 1). Using XMLpipeDB, the XML

file was found to contain 2126 unique matches of the form SP_#### (Fig. 2), validating the TallyEngine result. The PostgreSQL query found 2126 unique matches (Fig. 3), verifying all ordered locus values from the XML file were also in the database. In Microsoft Access, the database was found to have 4252 Ordered Locus Pairs because each gene was represented two times, as SP#### and SP_#### (Table 1). Therefore, there were 2126 unique ordered locus pairs. Between the different tables, all values for the ordered locus pairs were matched (Table 2).

By visual inspection of the systems table, many different species were seen in the data set including, most importantly, *Streptococcus pneumonaie* and its accompanying information. For *S. pneumoniae* str. TIGR4, all IDs took the form of SP_#### or SP#### as a result of the changes to GenMAPP builder. These IDs also appeared in the RefSeq table, as expected. The Gene Database Schema was almost identical to that of *Vibrio cholera* (Fig. 4). Full gene database testing results can be found in Appendix 2.

Changes to GenMAPP Builder were successful in reducing the number of errors to 333 gene identifications that were not found in the database. These genes had the same SP_#### format as ones that appeared in the database. Using the Excel match command, all results were #N/A. This means that the exceptions were not present in the UniProt XML file. In earlier trials, there were no matches, but these were corrected for by incorporating an underscore after the SP (part of the GenMAPP Builder modification, Appendix 1).

Microarray Results:

Out of the 4689 genes imported into GenMAPP, 1218 were significant at the 4 hour time point, 1902 at the 12 hour time point, 918 at the 24 hour time point, and 129 at the 48 hour time point (Table 3). The 12 hour time point was selected for further analysis because it had the

highest percentage of significantly changed genes (Table 3). After 12 hours, 215 genes were significantly increased (avgLogFC>0.25) and 1,671 genes had decreased expression (avgLogFC<-0.25). This means that 16 genes had significant p-values but small changes in the avgLogFC value.

Based on filtered MAPPfinder results, fourteen pathways mostly devoted to carbohydrate transport demonstrated significant changes in gene expression based on comprehensive p-value and z-score (Table 4). Out of these pathways, we focused on changes in gene expression within the sugar transmembrane transporter pathway for the remainder of the project (Fig. 5).

Out of 31 genes in the sugar transmembrane transporter pathway (Fig. 5), six genes demonstrated significantly decreased activity and nine genes that demonstrated significantly increased activity at the 12 hour time point. Additionally, nine other genes showed no significant change in expression levels and seven genes were not found in the UniProt database. There were no consistent relationships between gene function and whether it was increased or decreased. Genes showing increased and decreased expression were often both involved in transmembrane transport and sugar phosphorylation.

Discussion:

In this project, GenMAPP builder successfully produced a working gene database. We went through four import and export cycles due to some ambiguity in the gene systems and needs to keep consistent files. In particular, GenMAPP builder required modification to match the Ordered Locus Names between UniProt which used a SP_#### notation and the microarray information which used a SP#### notation, but this was easily accomplished by minor modification to the *Vibrio cholera* code. In general, there could have been less import-export

cycles if we had collected all information from the online profiles and microarray data before proceeding with import-export cycles, but that would have increased the total length of the project.

In our statistical analysis, we found that 1902 genes or 41 percent of genes were significantly increased or decreased at the 12 hour time point, much more than the five percent that would be predicted by chance. Additionally, approximately 25 percent of genes differed at the 4 and 24 hour time points, also more than would be predicted by chance. At 48 hours, less than three percent of genes showed any significance, leading to the conclusion that any variation was likely the result of random effects. In Sanchez et al. (2001), they found that 6.2 percent of the 1674 TIGR4 genes tested showed significant changes. The discrepancy in results could be because of the amount of genes used; we had a larger amount of genes because more gene information was digitally available and because of genes replicated within the dataset. Additionally, Sanchez et al. used a more stringent p-value of 0.01 when determining whether gene expression was significantly changed which can account for the decrease in percent changed. However, in both cases, the amount of changes was more than would be predicted by chance.

The gene ontology terms we found to be significantly changed at 12 hours were mostly gene pathways related to sugar and carbohydrate gene transport (Table 4). This added to the results of Sanchez et al. (2011) which emphasized changes in virulence genes although they did mention some changes in energy production and transport. This finding shows that, in addition to undergoing virulence and metabolic changes during biofilm production, *S. pneumoniae* also change patterns of carbohydrate transport within the cell. Viewing the genes within one of these pathways, we observed no clear patterns in how individual genes changed and there were near

equal numbers of genes involved in carbohydrate transformation that were increased, decreased, and not changed (Fig. 5). The lack of clarity may be due to the limited information available for each gene. More detailed information about gene function may help determine trends in gene expression that are not available when only examining the general type of gene activity. Unfortunately, this type of information is not available for most of the genes.

Although our analysis of differences was limited to the 12 hour samples, we observed significant changes in carbon transport genes, particularly genes involved with sugar phosphorylation. These results help give a more complete view of how *S. pneumoniae* gene expression changes during biofilm formation and, although more opportunities to examine differences are available, helps us understand how virulent and avirulent life stages of *S. pneumoniae* function.

References:

Aaberge, I. S., Eng, J., Lermark, G., & Løvik, M. (1995). Virulence of *Streptococcus pneumoniae* in mice: a standardized method for preparation and frozen storage of the experimental bacterial inoculum. *Microbial pathogenesis*, *18*(2), 141-152.

Hoskins, J., Alborn, W.E. Jr., Arnold, J., Blaszczak, L.C., Burgett, S., Bradley, S., DeHoff, S.T., Estrem, L.F., Fu, D.J., Fuller, W., Geringer, C., Gilmour, R., Glass, J.S., Khoja1, H., Kraft, A.R., Lagace, R.E., LeBlanc, D.J., Lee, L.N., Lefkowitz, E.J., Lu, J., Matsushima, P., McAhren, S.M., McHenney, M., McLeaster, K., Mundy, C.W., Nicas, T.I., Norris, F.H., O'Gara1, M.J., Peery, R.B., Robertson, G.T., Rockey, P., Sun, P.M., Winkler, M.E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C.A., Baltz, R.H., Jaskunas, S.R., Rosteck Jr., P.R., Skatrud, P.L., and Glass, J.I. (2001) Genome of the bacterium Streptococcus pneumoniae strain R6. Journal of Bacteriology 183(19): 5709-5717. doi: 10.1128/JB.183.19.5709-5717.2001

Sanchez, C.J., Kumar, N., Lizcano, A., Shivshankar, P., Dunning Hotopp, J.C., Jorgensen, J.H., Tettelin, H., and Orihuela, C.J. (2011) Streptococcus pneumoniae in Biofilms Are Unable to Cause Invasive Disease Due to Altered Virulence Determinant Production. PLoS ONE 6(12): e28738. doi:10.1371/journal.pone.0028738

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson, W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C., Dougherty, B.A., Morrison, D.A., Hollingshead, S.K., Fraser, C.M. (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, *293*(5529), 498-506.

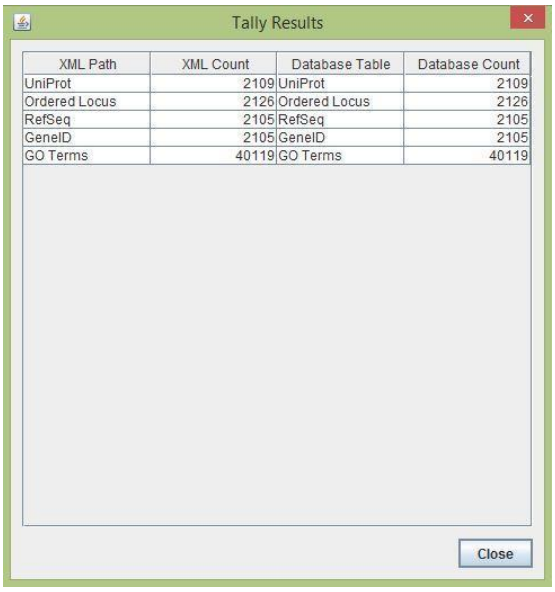

Figure 1: TallyEngine results showing the UniProt XML file and the files uploaded into GenMAPP Builder each had 2126 unique Ordered Locus Pairs.

Figure 2: An XMLpipeDB query of the UniProt XML file found 2126 gene IDs that matched the format SP_####. This number was consistent with the results of TallyEngine in GenMAPP Builder.

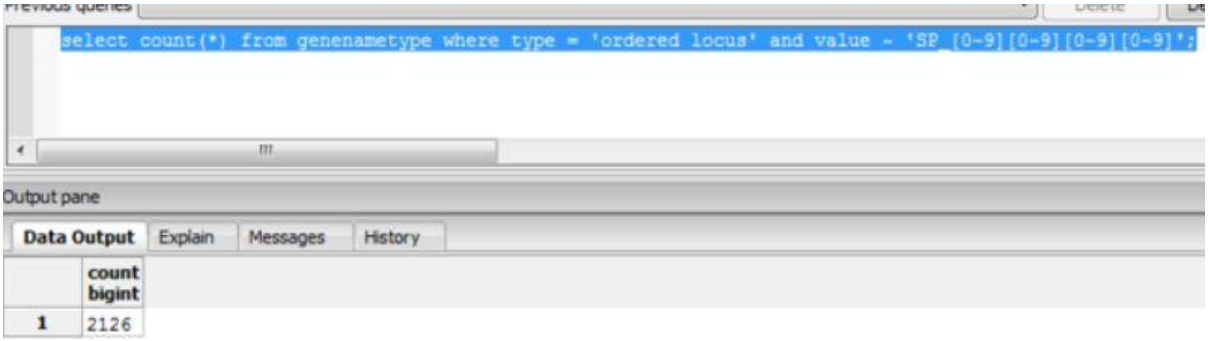

Figure 3: A PostgreSQL query for ordered locus values of the form "SP_####" found 2126 gene IDs, the same number as was found in TallyEngine and XMLpipeDB.

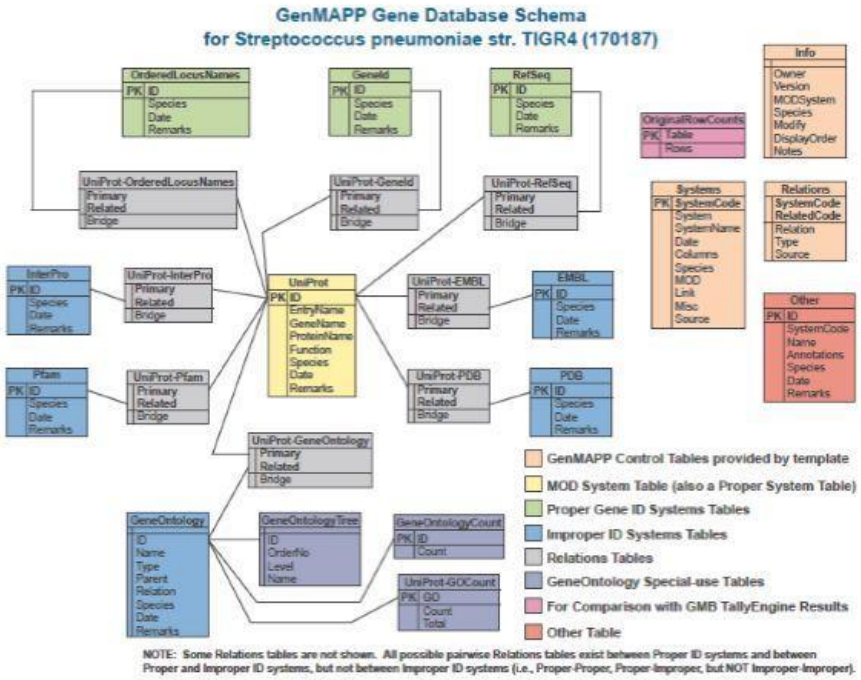

Figure 4: The gene database schema for *Streptococcus pneumoniae* str. TIGR4 demonstrating the relationships and references between tables in the gene database for *S. pneumoniae*. All relationships were as expected and there were no differences from the *Vibrio cholera* database schema.

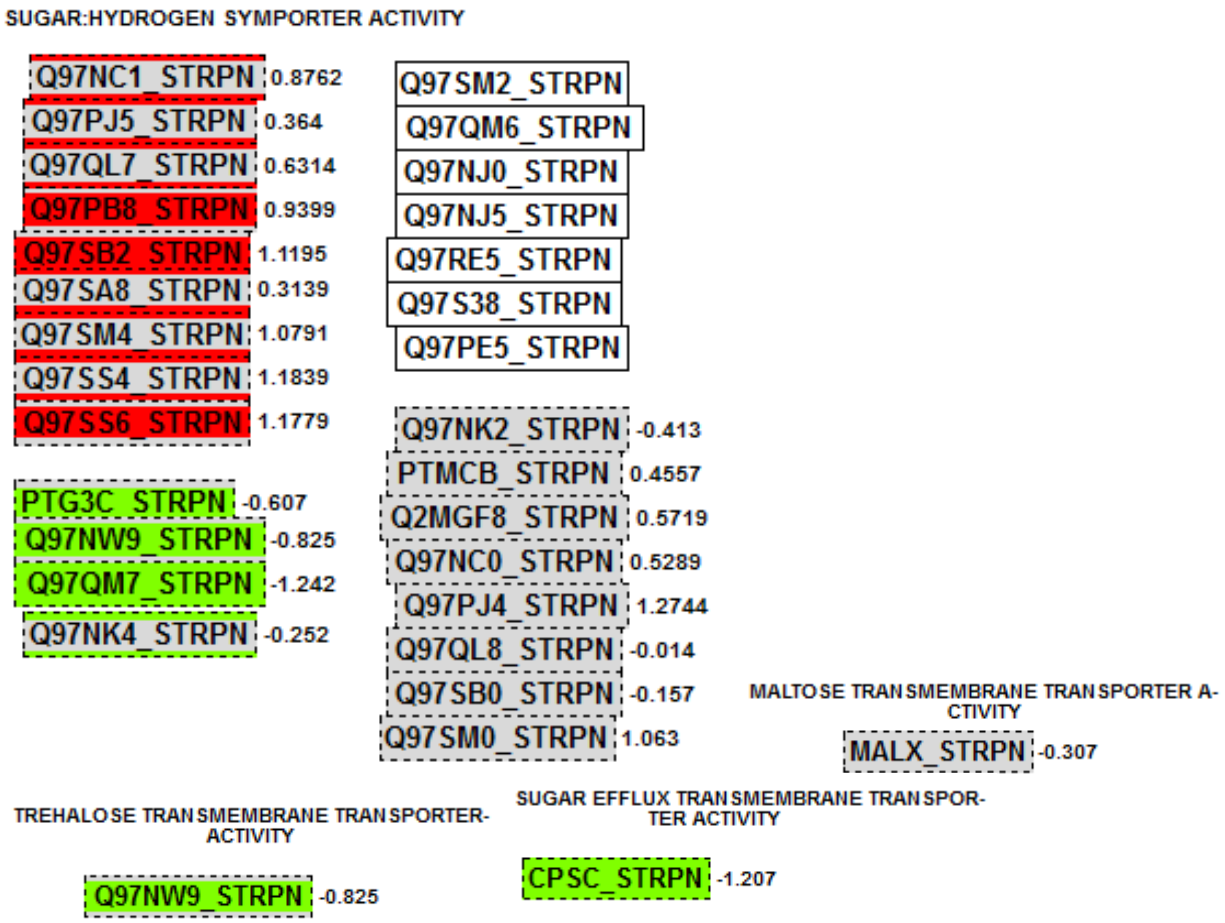

Figure 5: The reorganized MAPP of the sugar transmembrane transporter activity pathway. In this gene map, red represents a significant increase in gene expression and green represents a significant decrease. Significance was determined by a p-value of less than 0.05 and a average log fold change whose absolute value was greater than 0.25. Partially colored genes met the p-value criteria, but not the average log fold change criteria. Gray shading meant that no criteria were met and white meant the gene was not found in the UniProt database.

Table 1: Ordered row count results from the gene database as determined by Microsoft Access. The number of Ordered Locus Names was double that of other counting tallies because it contained both SP#### and SP_#### versions of each gene ID.

ID System	**ID Count**
EMBL	201
GeneOntology	3648
InterPro	2641
OrderedLocusNames	4252*
PDB	225
Pfam	1277
RefSeq	2105
UniProt	2109

*There are 2126 unique genes/proteins in the current version of the Gene Database; the 4252 count represents the total number of IDs due to duplicate IDs of the form SP#### and SP_####.

Table 2: Total Ordered Locus gened IDs counted by each method. All methods had an agreed value of 2126 except for the Microsoft Access which gave a value of 4252 because it contained both SP_#### and SP#### gene IDs.

Source	Ordered Locus Totals	Gene ID Format
Tally Engine	2126	SP_####
XMLPipeDB Match	2126	SP_####
Postgres SQL	2126	SP_####
Access (gdb)	4252	SP_####, SP####

Table 3: The number of genes showing significantly altered expression at each time point along with the percent of all genes changed.

Time	Significant P-values	Percent of all genes
4hr	1218	26%
12hr	1902	41%
24hr	918	20%
48hr	129	2.8%

Table 4: The ranked filtered list of most significantly different gene pathways generated by MAPPFinder. P-values were considered significant if they were less than or equal to 0.05. Most of the pathways relate to carbohydrate modification and transport.

Gene_Ontology_Result	**p-value**	**z-score**
phosphoenolpyruvate-dependent sugar phosphotransferase system	0.004	6.665
transporter activity	0.004	6.335
protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	0.004	6.213
carbohydrate transport	0.004	6.153
carbohydrate transporter activity	0.033	5.744
carbohydrate transmembrane transporter activity	0.033	5.744

transport	0.033	5.674
establishment of localization	0.033	5.674
localization	0.033	5.674
cation:sugar symporter activity	0.033	5.631
sugar:hydrogen symporter activity	0.033	5.631
solute:hydrogen symporter activity	0.033	5.631
sugar transmembrane transporter activity	0.039	5.447
solute:cation symporter activity	0.039	5.273

Appendix 1:

```
@Override
        public TableManager getSystemsTableManagerCustomizations(TableManager tableManager,
DatabaseProfile dbProfile) {
                super.getSystemsTableManagerCustomizations(tableManager, dbProfile);
                tableManager.submit("Systems", QueryType.update, new String[][] {
                { "SystemCode", "N" },
                { "Species", "|" + getSpeciesName() + "|" }
                });

                tableManager.submit("Systems", QueryType.update, new String[][] {
                { "SystemCode", "N" },
                { "Link",
"http://bacteria.ensembl.org/streptococcus_pneumoniae_tigr4/Gene/Summary?g=~" }
                });

                return tableManager;
        }

          /**
    * @see
edu.lmu.xmlpipedb.gmbuilder.databasetoolkit.profiles.UniProtSpeciesProfile#getSystemTableManagerC
ustomizations(edu.lmu.xmlpipedb.gmbuilder.databasetoolkit.tables.TableManager,
    *       edu.lmu.xmlpipedb.gmbuilder.databasetoolkit.tables.TableManager,
    *       java.util.Date)
    */
   @Override
   public TableManager getSystemTableManagerCustomizations(TableManager tableManager,
TableManager primarySystemTableManager, Date version) throws SQLException,
InvalidParameterException {
        // Start with the default OrderedLocusNames behavior.
        TableManager result = super.getSystemTableManagerCustomizations(tableManager,
primarySystemTableManager, version);

        // We want to grab all of the legal OrderedLocusNames Ids and
        // remove the '_', adding them to the OrderedLocusNames table
        final String vcID = "SP_*";
        String sqlQuery = "select d.entrytype_gene_hjid as hjid, c.value " + "from genenametype c inner join
genetype d " + "on (c.genetype_name_hjid = d.hjid) " + "where (c.value similar to ?)" + "and type <>
'ordered locus names' " + "group by d.entrytype_gene_hjid, c.value";

        String dateToday = GenMAPPBuilderUtilities.getSystemsDateString(version);
        Connection c = ConnectionManager.getRelationalDBConnection();
        PreparedStatement ps;
        ResultSet rs;
```

```
        try {
            // Query, iterate, add to table manager.
            ps = c.prepareStatement(sqlQuery);
            ps.setString(1, vcID);
            rs = ps.executeQuery();
            while (rs.next()) {
            String hjid = Long.valueOf(rs.getLong("hjid")).toString();

              // We want to remove the '_' here
            String id = rs.getString("value");

            String[] substrings = id.split("/");
            String new_id = null;
            for (int i = 0; i < substrings.length; i++) {

                    new_id = substrings[i].replace("_", "");

                    _Log.debug("Remove '_' from " + id + " to create: " + new_id + " for surrogate " + hjid);
                    result.submit("OrderedLocusNames", QueryType.insert, new String[][] { { "ID", new_id
}, { "Species", "|" + getSpeciesName() + "|" }, { "\"Date\"", dateToday }, { "UID", hjid } });
            }
            }
        } catch(SQLException sqlexc) {
            logSQLException(sqlexc, sqlQuery);
        }

        return result;
    }

    /**
     * Helper method for logging an SQL exception.
     */
    private void logSQLException(SQLException sqlexc, String sqlQuery) {
        _Log.error("Exception trying to execute query: " + sqlQuery);
        while (sqlexc != null) {
            _Log.error("Error code: [" + sqlexc.getErrorCode() + "]");
            _Log.error("Error message: [" + sqlexc.getMessage() + "]");
            _Log.error("Error SQL State: [" + sqlexc.getSQLState() + "]");
            sqlexc = sqlexc.getNextException();
        }
    }

    private static final Log _Log =
LogFactory.getLog(StreptococcusPneumoniaeTIGR4UniProtSpeciesProfile.class);
}
```

Appendix 2: Complete version of the testing report is available online

Testing Report

Contents

[hide]

[edit]Export Information

Version of GenMAPP Builder: 2.0b73

Database called: tATK_TIGR4_2013NOV25

Computer on which export was run: Tauras' Personal Computer

Postgres Database name: tATK_TIGR4_2013NOV25

UniProt XML filename: 20131118_UniProtXML_tATK_TIGR4_TPV.xml

- UniProt XML version: UniProt Release 2013_11; 2013Nov13
- Time taken to import: 3.15min

GO OBO-XML filename: 20131120_OBOXML_tATK_TPV.obo-xml

- GO OBO-XML version: 2013Nov20
- Time taken to import: 10.59min
- Time taken to process: 9.25min

GOA filename: 20131118_GOA_tATK_TIGR4_TPV.goa

- GOA version: 2013Nov12 14:49
- Time taken to import: 0.03min

Name of .gdb file: Streptococcus_pneumoniae_TIGR4_20131125.gdb

- Time taken to export .gdb: less than 1 hour
  - Started at 22:53
  - Finished by 23:50
- Upload your file and link to it here. Streptococcus_pneumoniae_TIGR4_20131125.gdb

TallyEngine

- Tally Engine run on Tauras' personal computer.
- **Final Results:**
  - Ordered Locus XML Count: 2126
- Ordered Locus Database Count: 2126

Using XMLPipeDB match to Validate the XML Results from the TallyEngine

- XMLPipeDB match program was downloaded from Sourceforge [[1]]
- Moved xmlmatch jar file to Downloads folder on personal computer
- Ran cmd program on personal computer
- Ran query: *cd Downloads file*
- Searched for pattern: SP_[0-9][0-9][0-9][0-9]
- Total unique matches found: 2126
- This total matched results found in Tally Engine count

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

- Ran pgAdmin III through personal computer to run SQL query
- Command used:
  - *select count(*) from genenametype where type = 'ordered locus' and value ~ 'SP_[0-9][0-9][0-9][0-9]';*
- Unique matches found: 2126

- These results matched those of Tally Engine and XMLPipeDB match, confirming values

Follow the instructions on this page to query the PostgreSQL Database.

OriginalRowCounts Comparison

- Original Row Counts for the gdb file contained had a UniProt Ordered Locus count of 4252
- This was a result of the databases including Gene IDs with and and without underscore

Visual Inspection

Systems Table

- There are numerous missing dates for the gene ID systems.

OrderedLocusNames Table

- ID's take the forms SP_#### and SP####

UniProt Table

- ID's are all in expected form SP_####

RefSeq Table

- IDs in form NP_######
- This is expected form for RefSeq, refers to protein accession number.

.gdb Use in GenMAPP

Note:

Putting a gene on the MAPP using the GeneFinder window

- Try a sample ID from each of the gene ID systems. Open the Backpage and see if all of the cross-referenced IDs that are supposed to be there are there.

Note:

- no criteria met: Q97RY3 (Q97RY3_STRPN) matched with uniprot page
- not found: Q97NJ5 (Q97NJ5_STRPN) matched with uniprot page
- decreased: Q97SJ6 (CPSC_STRPN) matched with uniprot page
- increased: Q97SB2 (Q97SB2_STRPN) matched with uniprot page

Creating an Expression Dataset in the Expression Dataset Manager

- How many of the IDs were imported out of the total IDs in the microarray dataset? How many exceptions were there? Look in the EX.txt file and look at the error codes for the records that were not imported into the Expression Dataset. Do these represent IDs that were present in the UniProt XML, but were somehow not imported? or were they not present in the UniProt XML?

Note: 4689 out of 5022 IDs were imported. There were 333 exceptions, all due to not being present in UniProt.

Coloring a MAPP with expression data

Note: increased was colored red, decreased was colored green, no criteria met was colored grey, and not found was colored white

Running MAPPFinder

Note: MAPPFinder worked successfully for all of the data used in this project.

Compare Gene Database to Outside Resource

- Could not find downloadable gene ID list on Ensembl, which was used as MOD for the TIGR4 strain.
- A 'Coding Gene Count' was given, a value of 2125
- This total is one less than our expected value of 2126
- Inability to find ID list kept us from being able to identify reasons for differences between the values.

Gene counts

Coding genes:	2,125
Short Non coding genes:	58
Gene transcripts:	2,183