Gene Database Testing Report for *Shigella flexneri* 2a str. 301

Export Information:

Version of GenMAPP Builder:

- **gmbuilder-3.0.0-build-5**

Computer on which export was run:

- **LMU Seaver 120 computer: front of the room, 3rd computer from the right**

Postgres Database name:

- **Shigella_flexneri_20151208**

UniProt XML filename:

- UniProt XML version (The version information can be found at the UniProt News Page < http://uniprot.org/news>):
  - **UniProt release 2015_11**
- UniProt XML download link:
  - <**http://www.uniprot.org/uniprot/?query=proteome:UP000001006>**
- Time taken to import:
  - **4.43 minutes**

GO OBO-XML filename:

- GO OBO-XML version (The version information can be found in the file properties after the file downloaded from the GO Download page < http://archive.geneontology.org/latest-termdb/go_daily-termdb.obo-xml.gz> has been unzipped):
  - **Version created on 11/19/2015 (at 2:24 AM)**
- GO OBO-XML download link:
  - <**http://archive.geneontology.org/latest-termdb/go_daily-termdb.obo-xml.gz>**
- Time taken to import:
  - **6.84 minutes**
- Time taken to process:
  - **5.49 minutes**

GOA filename (give filename and upload and link to compressed file):

- GOA version (News on the UniProt – GOA page < http://www.ebi.ac.uk/GOA> records past releases; current information can be found in the Last modified field on the FTP site <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>):
  - **Version released on 11/11/2015.**
- GOA download link:
  - **<http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/103.S_flexneri_301.goa>**
- Time taken to import:
  - **0.06 minutes**

Name of .gdb file (give filename and upload and link to compressed file): **Sf-Std 20151208.gdb**

- Time taken to export:
  - **2 hour, 0 minutes, 27 seconds**
- Start time:
  - **9:35:00 PM PDT**

- End time:
  - **11:35:27 PM PDT**

Using TallyEngine:

- With the necessary files import to PostgreSQL, TallyEngine was run and the following table is the result of the export:

Tally Results

XML Path	XML Count	Database Table	Database Count
UniProt	4103	UniProt	4103
RefSeq	7518	RefSeq	7518
GeneID	4406	GeneID	4406
Ordered Locus	15134	Ordered Locus	7567
Ordered Locus	0	Ordered Locus	7663
GO Terms	43954	GO Terms	43954

Using XMLPipeDB match to Validate the XML Results from the TallyEngine:

- Two separate, almost identical regex, were used in order to find more ordered locus names in the XML file from just within the <gene/> tag. The one below found 7567 IDs:

```
java -jar xmlpipedb-match-1.1.1/xmlpipedb-match-1.1.1.jar
">(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?(/|</name>)" < uniprot-
proteome%3AUP000001006.xml > shigella_flexneri_results
```

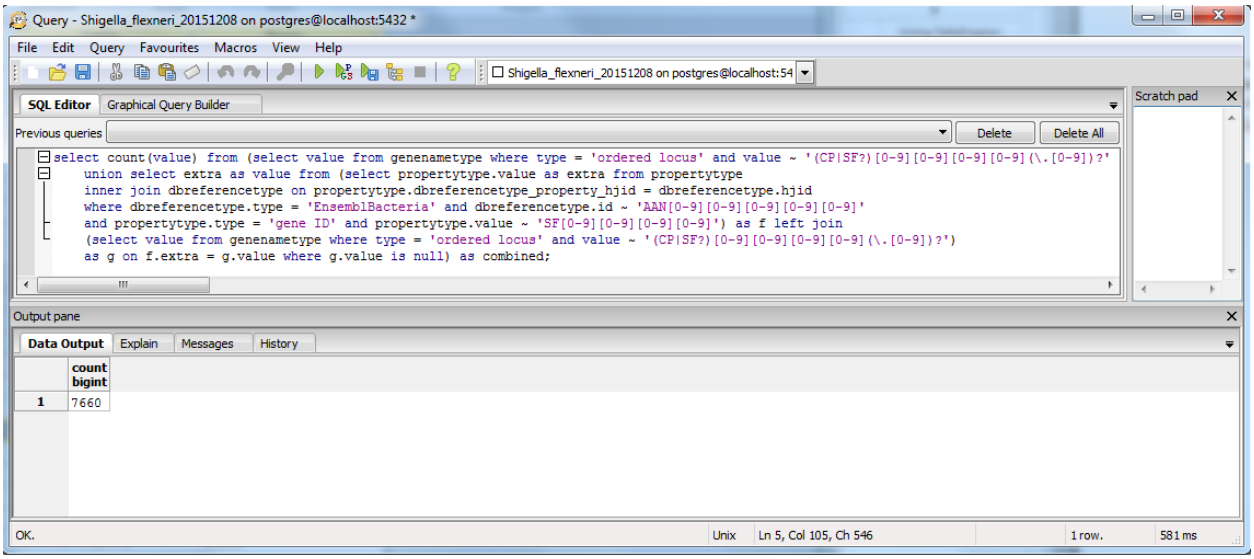
- The one below found 3 IDs:

```
java -jar xmlpipedb-match-1.1.1/xmlpipedb-match-1.1.1.jar
"/(CP|SF?)[0-9][0-9][0-9][0-9](\.[0-9])?(/|</name>)" < uniprot-
proteome%3AUP000001006.xml > shigella_flexneri_results
```

- When added together, the results become 7566 + 3 = 7569.
- Since there were ID duplicates between the <gene/> and <dbReference/> tags, there was no easy way to actually find newer IDs without miscounting. Therefore, this initial regex count was kept since it is the closest number we could get to our database export count (7569 vs. 7663).

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

- For our specific specie, a specialized PSQL query was made thanks to our advisor, Dr. John David Dionisio. Below is a screenshot of the query in action, which produced 7660 entries:



OriginalRowCounts Comparison

- In the .gdb file, which was opened in Microsoft Access, the OriginalRowCounts table was inspected in case the export did not work as we intended. With the additional ~92 IDs that was found from the <dbReference/> tag, the total count should be 7569 + 92.
- In this table, the OrderedLocusNames row was determined to contain the number of IDs that we were expecting, as seen from the table below.
- Other than the table that we were expecting to be changed, the rest of the rows seemed to be kept intact when compared with the “benchmark” build that we made previously (Build 2).

Table	Rows
Info	1
Systems	35
Relations	35
Other	0
GeneOntologyTree	109896
GeneOntology	6478
UniProt-GOCount	3674
GeneOntologyCount	3673
UniProt-GeneOntology	18524
UniProt	4103
RefSeq	7501
EMBL	121
Pfam	2357
InterPro	5124
GeneID	4389
EnsemblBacteria	9524
PDB	95
OrderedLocusNames	7661
UniProt-PDB	225
UniProt-OrderedLocusNam	7661
UniProt-EnsemblBacteria	19075
UniProt-GeneID	7931
UniProt-InterPro	21210
UniProt-Pfam	8929
UniProt-EMBL	17189
UniProt-RefSeq	14017
RefSeq-EMBL	18330
RefSeq-Pfam	9338
RefSeq-InterPro	21592
RefSeq-GeneID	15995

Visual Inspection

A visual inspection was performed on the individual tables to see if there are any problems. Primarily, the Systems table was checked for dates. There were no problems in the following tables:

- GeneOntology
- InterPro
- GeneID
- RefSeq
- UniProt
- EMBL
- PDB
- Pfam

- OrderedLocusNames
- EnsemblBacteria

Additionally, the following tables seemed to all have the correct forms of IDs:

- UniProt
- Refseq
- OrderedLocusNames

Download .gdb File

The resulting .gdb file can be downloaded in the link provided below:

- <https://xmlpipedb.cs.lmu.edu/biodb/fall2015/images/b/b8/Sf-Std_20151214.gdb>.