Gene Database Testing Report

Export Information

Version of GenMAPP Builder: gmbuilder2.0-b72

Computer on which export was run: Front row , second computer from the left

Postgres Database name:Leishmania_major_11262013

UniProt XML filename: UniprotXML Leishmania 05112013 Gabe Lena.xml

- UniProt XML version (The version information can be found at the UniProt News Page): UniProt release 2013_10 - October 16, 2013
- Time taken to import: 7.12 minutes

GO OBO-XML filename: Leishmania 05112013 Gabe Lena.obo-xml.gz

- GO OBO-XML version (The version information can be found in the file properties after the file downloaded from the GO Download page has been unzipped):
  Monday, November 04, 2013, 2:03:38 AM
- Time taken to import:6.32 minutes
- Time taken to process: 0.4 minutes

GOA filename:LeishmaniaGOA 19112013 Lena Gabe.goa

- GOA version (News on this page records past releases; current information can be found in the Last modified field on the FTP site): 14 November, 2013

12-Nov-2013 11:47 3.0M

- Time taken to import: 4.54 minutes

Name of .gdb file: Media:LeishmaniaGDB Lena Gabe 20131203.gdb

- Upload your file and link to it here.

Note:

TallyEngine

Tally Results

XML Path	XML Count	Database Table	Database Count
UniProt	8041	UniProt	8041
Ordered Locus	0	Ordered Locus	0
ORF	8355	ORF	8355
RefSeq	8317	RefSeq	8317
GeneID	8317	GeneID	8317
GO Terms	40065	GO Terms	40065

Close

Using XMLPipeDB match to Validate the XML Results from the TallyEngine

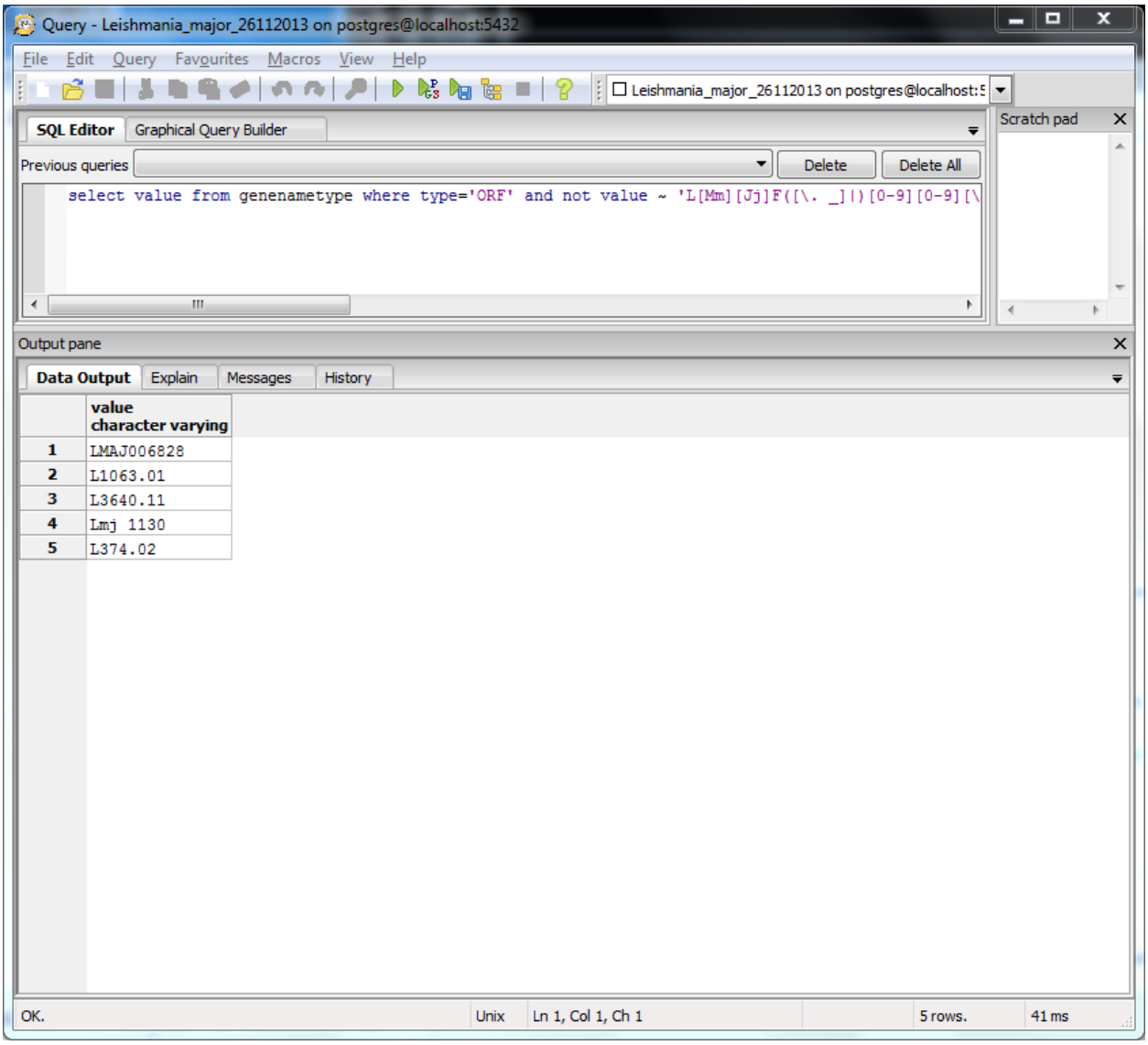
```
Command Prompt
lmjf_01_0130: 2
lmjf_26_2010: 2
lmjf_14_0310: 2
lmjf_23_1270: 2
lmjf_36_6000: 2
lmjf_29_1980: 2
lmjf_13_1250: 2
lmjf_16_1000: 2
lmjf_18_0710: 2
lmjf_32_3940: 2
lmjf_16_1005: 2
lmjf_29_0210: 2
lmjf_36_5220: 2
lmjf_15_1320: 2
lmjf_12_0820: 2
lmjf_32_3980: 2
lmjf_35_5040: 2
lmjf_23_1280: 2
lmjf_29_1970: 2
lmjf_01_0120: 2
lmjf_32_2050: 2
lmjf_14_0320: 2
lmjf_26_2020: 2
lmjf_22_0350: 2
lmjf_35_1460: 2
lmjf_19_1580: 2
lmjf_18_0720: 2
lmjf_13_1260: 2
lmjf_29_0200: 2
lmjf_32_2040: 2
lmjf_32_3970: 2
lmjf_12_0810: 2
lmjf_36_5210: 2
lmjf_15_1330: 2
lmjf_35_1470: 2
lmjf_35_5050: 2
lmjf_29_1960: 2
lmjf_01_0110: 2
lmjf_23_1290: 2
lmjf_26_2030: 2
lmjf_14_0330: 2
lmjf_33_2090: 2
lmjf_19_1590: 2
lmjf_26_0800: 2
lmjf_13_1270: 2
lmjf_18_0730: 2

Total unique matches: 8353

C:\Users\keckuser\Downloads>
```

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

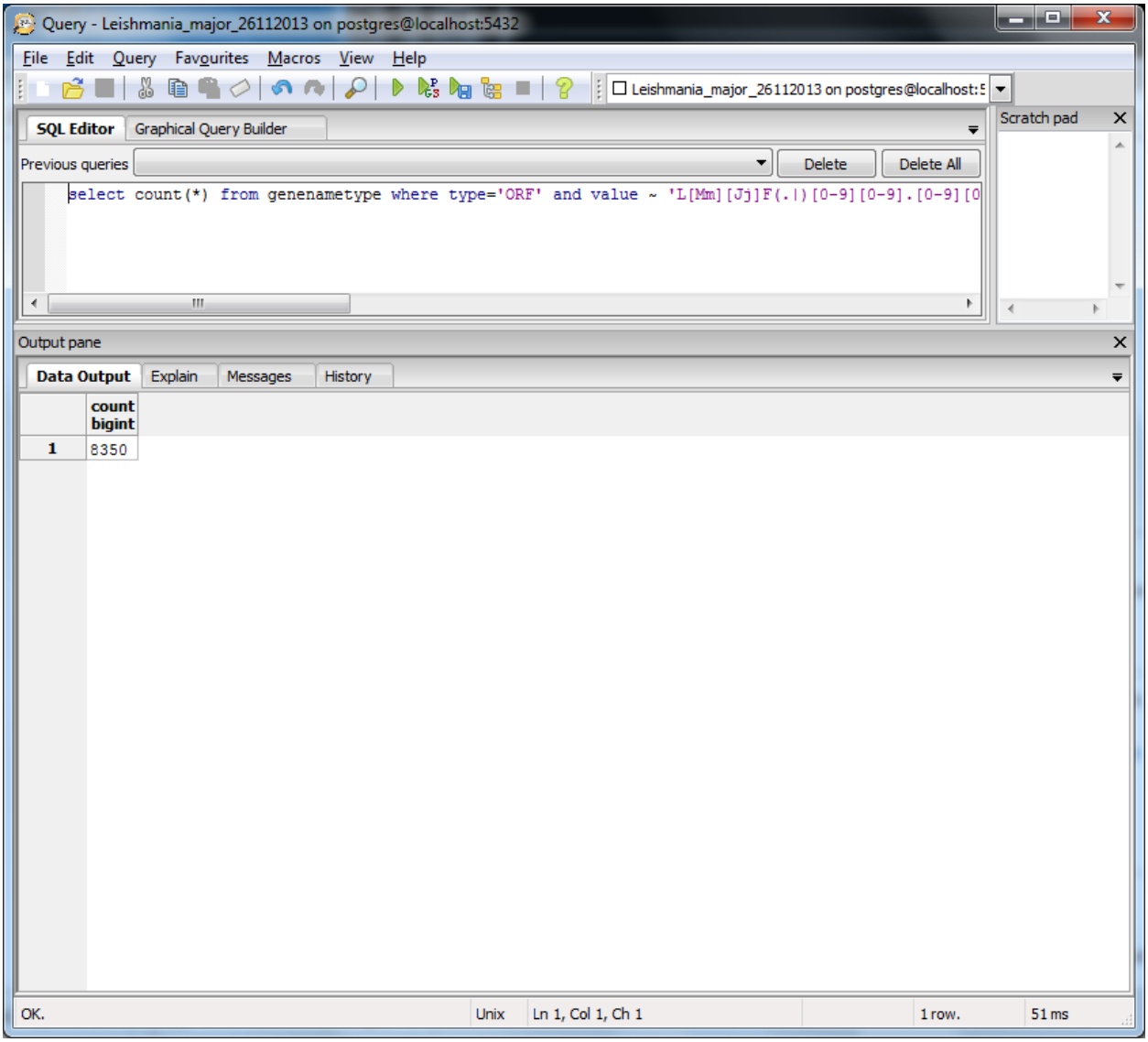
- We found 8353, we are missing 2 ORFs, but cannot reach the stragglers via coding.

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

select value from genenametype where type='ORF' and not value ~ 'L[Mm][Jj]F([\. _]|)[0-9][0-9][\. _][0-9][0-9][0-9][0-9]';



select count(*) from genenametype where type='ORF' and value ~ 'L[Mm][Jj]F(.|)[0-9][0-9].[0-9][0-9][0-9][0-9]';

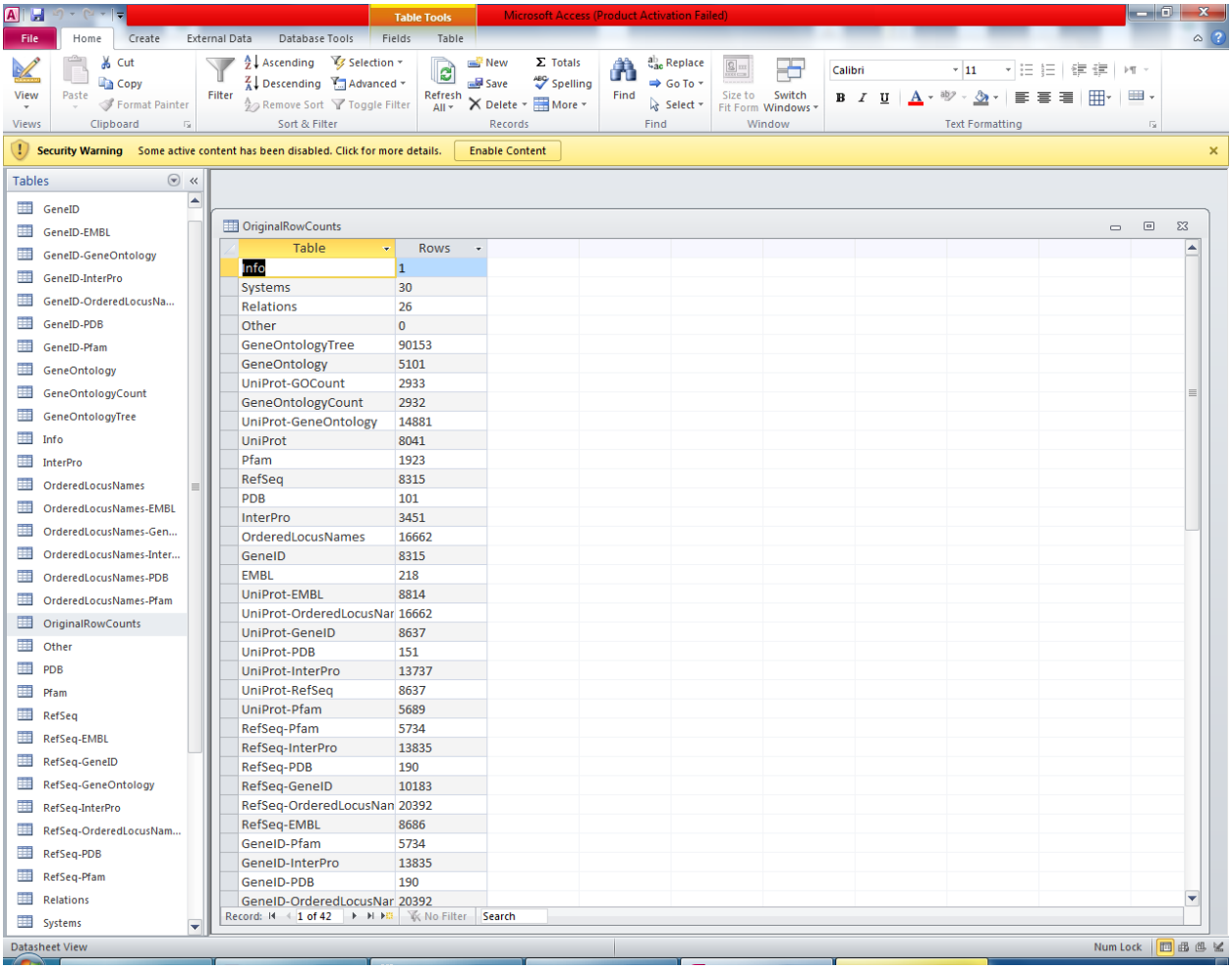


8350 that match original pattern, 5 stragglers. There is an inconsistancy in patterns where some have underscores, some have periods, some have spaces. We were unable to capture all IDs at once.

OriginalRowCounts Comparison

Within the .gdb file, look at the OriginalRowCounts table to see if the database has the expected tables with the expected number of records. Compare the tables and records with a benchmark .gdb file.

Copy the OriginalRowCounts table and paste it here:



OriginalRowCounts

Table	Rows
Info	1
Systems	30
Relations	26
Other	0
GeneOntologyTree	90153
GeneOntology	5101
UniProt-GOCount	2933
GeneOntologyCount	2932
UniProt-GeneOntology	14881
UniProt	8041
Pfam	1923
RefSeq	8315
PDB	101
InterPro	3451
OrderedLocusNames	16662
GeneID	8315
EMBL	218
UniProt-EMBL	8814
UniProt-OrderedLocusNar	16662
UniProt-GeneID	8637
UniProt-PDB	151
UniProt-InterPro	13737
UniProt-RefSeq	8637
UniProt-Pfam	5689
RefSeq-Pfam	5734
RefSeq-InterPro	13835
RefSeq-PDB	190
RefSeq-GeneID	10183
RefSeq-OrderedLocusNan	20392
RefSeq-EMBL	8686
GeneID-Pfam	5734
GeneID-InterPro	13835
GeneID-PDB	190
GeneID-OrderedLocusNar	20392

Visual Inspection

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

- Look at the Systems table. Is there a date in the Date field for all gene ID systems present in the database?
- No there are only dates for approximately 50% of the ID systems in the database
- Open the UniProt, RefSeq, and OrderedLocusNames tables. Scroll down through the table. Do all of the IDs look like they take the correct form for that type of ID?
- All IDs appear to take the correct form for that type of ID.

.gdb Use in GenMAPP

Note: Most GDB use in GenMAPP is profiled on the GenMAPP users page

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: All IDs seem to be present.

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?
- There are 1820 errors in the exceptions file. Approximately 1758 not found in XML the rest not imported into database due to species specific coding not complete

Coloring a MAPP with expression data

Note: Coloring successfully yielded graphical depiction of expression data

Running MAPPFinder

Note: See final project deliverables (.mapp)and (.xls)