Streptococcus pneumoniae TIGR4 20131125 GeneTestingReport

From LMU BioDB 2013

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

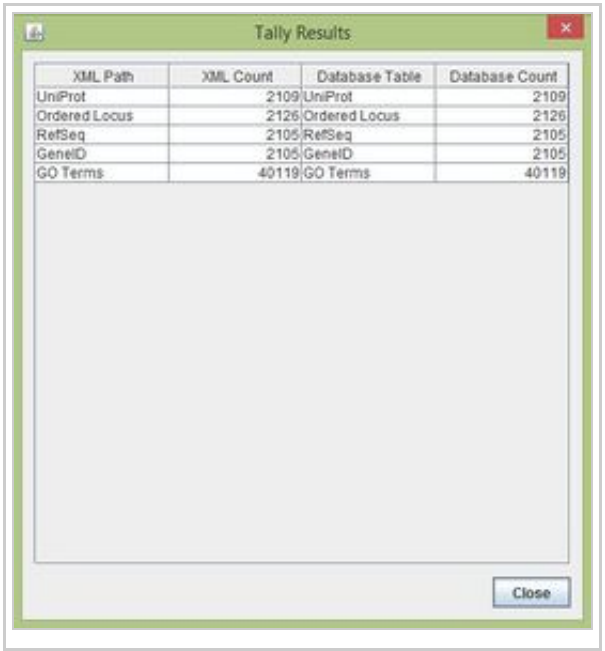
Name of .gdb file: Streptococcus_pneumoniae_TIGR4_20131125.gdb

- 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

Contents

TallyEngine

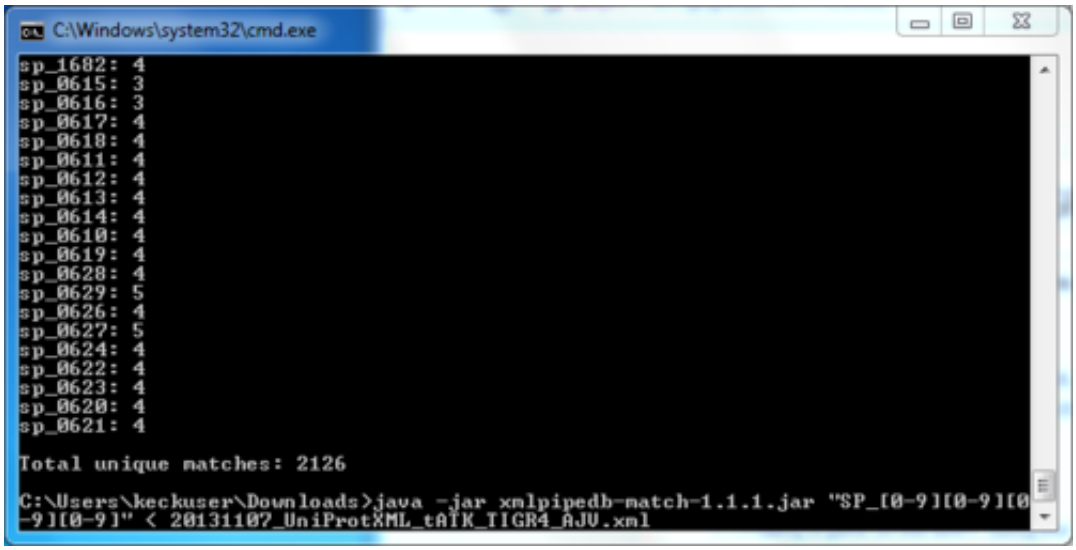
- Tally Engine run on Tauras' personal computer.
- **Final Results:**
  - Ordered Locus XML Count: 2126
  - Ordered Locus Database Count: 2126



Using XMLPipeDB match to Validate the XML Results from the TallyEngine

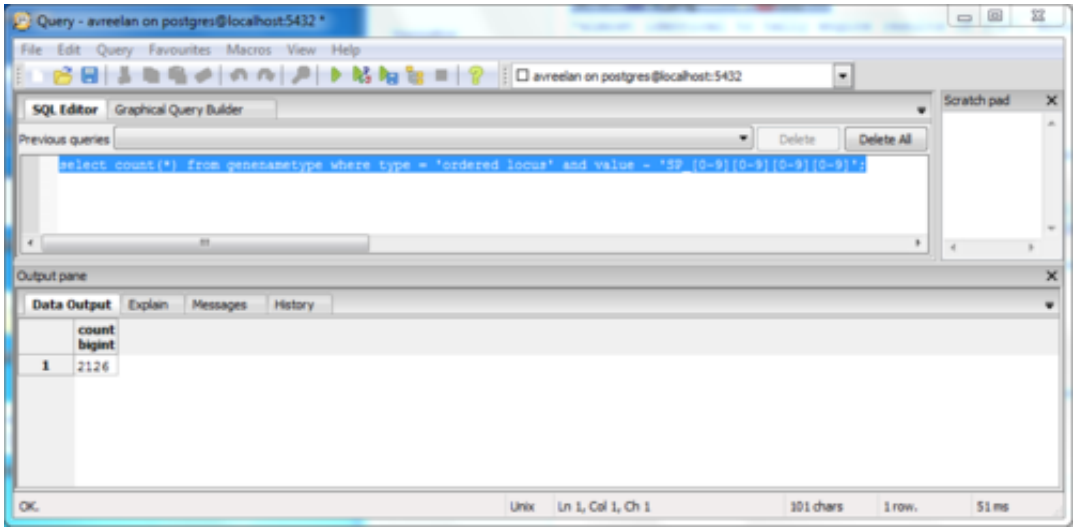
- XMLPipeDB match program was downloaded from Sourceforge [[1]]
- Moved xmlmatch jar file to Downloads folder on personal computer

- 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.

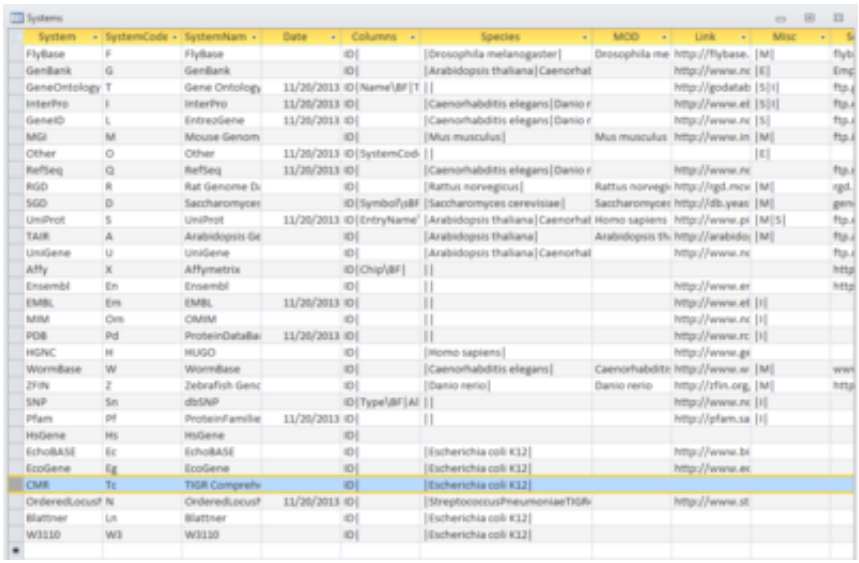
OriginalRowCounts Comparison

- 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.



System	SystemCode	SystemNam	Date	Columns	Species	MOD	Link	Misc
FlyBase	F	FlyBase		ID|	|Drosophila melanogaster|	Drosophila me	http://flybase.	|M|
GenBank	G	GenBank		ID|	|Arabidopsis thaliana|Caenorhat		http://www.nc	|E|
GeneOntology	T	Gene Ontology	11/20/2013	ID|Name|BF|T	||		http://godatab	|S|I|
InterPro	I	InterPro	11/20/2013	ID|	|Caenorhabditis elegans|Danio r		http://www.ebi	|S|I|
GeneID	L	EntrezGene	11/20/2013	ID|	|Caenorhabditis elegans|Danio r		http://www.nc	|S|
MGI	M	Mouse Genom		ID|	|Mus musculus|	Mus musculus	http://www.in	|M|
Other	O	Other	11/20/2013	ID|SystemCode	||			|E|
RefSeq	Q	RefSeq	11/20/2013	ID|	|Caenorhabditis elegans|Danio r		http://www.nc	
RGD	R	Rat Genome D		ID|	|Rattus norvegicus|	Rattus norvegi	http://rgd.mcw	|M|
SGD	D	Saccharomyces		ID|SymbolFullF	|Saccharomyces cerevisiae|	Saccharomyces	http://db.yeas	|M|
UniProt	S	UniProt	11/20/2013	ID|EntryName'	|Arabidopsis thaliana|Caenorhat	Homo sapiens	http://www.pi	|M|S|
TAIR	A	Arabidopsis Ge		ID|	|Arabidopsis thaliana|	Arabidopsis Th	http://arabido	|M|
UniGene	U	UniGene		ID|	|Arabidopsis thaliana|Caenorhat		http://www.nc	
Affy	X	Affymetrix		ID|Chip|BF|	||			
Ensembl	En	Ensembl		ID|	||		http://www.en	
EMBL	Em	EMBL	11/20/2013	ID|	||		http://www.ebi	|I|
MIM	Om	OMIM		ID|	||		http://www.nc	|I|
PDB	Pd	ProteinDataBa	11/20/2013	ID|	||		http://www.rc	|I|
HGNC	H	HUGO		ID|	|Homo sapiens|		http://www.ge	
WormBase	W	WormBase		ID|	|Caenorhabditis elegans|	Caenorhabditi	http://www.w	|M|
ZFIN	Z	Zebrafish Genc		ID|	|Danio rerio|	Danio rerio	http://zfin.org	|M|
SNP	Sn	dbSNP		ID|Type|BF|Al	||		http://www.nc	|I|
Pfam	Pf	ProteinFamilie	11/20/2013	ID|	||		http://pfam.sa	|I|
HsGene	Hs	HsGene		ID|				
EchoBASE	Ec	EchoBASE		ID|	|Escherichia coli K12|		http://www.bi	
EcoGene	Eg	EcoGene		ID|	|Escherichia coli K12|		http://www.ec	
CMR	Tc	TIGR Comprehv		ID|	|Escherichia coli K12|			
OrderedLocusN	N	OrderedLocusN	11/20/2013	ID|	|StreptococcusPneumoniaeTIGR4|		http://www.st	
Blattner	Ln	Blattner		ID|	|Escherichia coli K12|			
W3110	W3	W3110		ID|	|Escherichia coli K12|			

OrderedLocusNames Table

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

UniProt Table

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

ID	Species	Date
SP_0002	|StreptococcusPneumoniaeTIGR4|	11/25/2013
SP_0274	|StreptococcusPneumoniaeTIGR4|	11/25/2013
SP_0291	|StreptococcusPneumoniaeTIGR4|	11/25/2013
SP_0458	|StreptococcusPneumoniaeTIGR4|	11/25/2013
SP_1644	|StreptococcusPneumoniaeTIGR4|	11/25/2013

RefSeq Table

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

ID	Species	Date
NP_346215	|Streptococcus	11/25/2013
NP_345814	|Streptococcus	11/25/2013
NP_344971	|Streptococcus	11/25/2013
NP_345500	|Streptococcus	11/25/2013
NP_346430	|Streptococcus	11/25/2013
NP_345826	|Streptococcus	11/25/2013
NP_345408	|Streptococcus	11/25/2013
NP_345370	|Streptococcus	11/25/2013
NP_346204	|Streptococcus	11/25/2013
NP_345063	|Streptococcus	11/25/2013

.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

Template

Alina's User Page Kevin's User Page Tauras's User Page

Biological Databases Class Page Gene Database Project Gene Database Project Report Guidelines

Streptococcus pneumoniae

Import Export Cycle 1: tATK Export One: TIGR4 Testing Report
Import Export Cycle 2: tATK E2: TIGR4 Testing Report
Import Export Cycle 3: tATK E3: TIGR4 Testing Report
Import Export Cycle 4: tATK E4: TIGR4 Testing Report

Data Information

Project Roles: Project Manager Coder GenMAPP User Quality Assurance

- Kmeilak Week 10, Kmeilak Week 11, Kmeilak Week 12, Kmeilak Week 13
- Taur.vil Week 10, Taur.vil Week 11, Taur.vil Week 12, Taur.vil Week 13, Taur.vil Week 14, Taur.vil Week 15
- Ajvree Week 10, Ajvree Week 11, Ajvree Week 12, Ajvree Week 13, Ajvree Week 14, Ajvree Week 15

Categories: Streptococcus pneumoniae | Group Projects | Team ATK | Journal Entry

---