

**Generating a New Gene
Database for *Leishmania*
major using GenMAPP and
XML pipedb**

Gabriel Leis, Viktoria Kuehn, Lena
Hunt, Kevin McGee

Outline

- Intro to Leishmania Major
- Methods
 - Database Preparation
 - Preparation of Microarray data
 - GenMAPP and MAPPFinder
- Results and Analysis
 - Database testing and analysis
 - MAPPFinder and Microarray analysis

Leishmania major needs a gene database

- Tropical and subtropical protozoan causing leishmaniases
 - 2 million infections annually
- Microarray data from Ivans *et al.*
- Analysis of Biological pathways
 - GenMAPP (Gene Map Annotator and Pathway Profile)
 - MAPPFinder

→ Create Gene Database with GenMAPP Builder



Gene expression comparison between *L. major* and *L. infantum* in developmental life stages

- Multispecies DNA oligonucleotide microarray
- Compared promastigotes (sandfly) and amastigotes (mouse macrophage)
- Alexa 647/Alexa 555 signal intensities (amastigote/promastigote) to signal mean intensities
- 91-93% of genes no change between stages

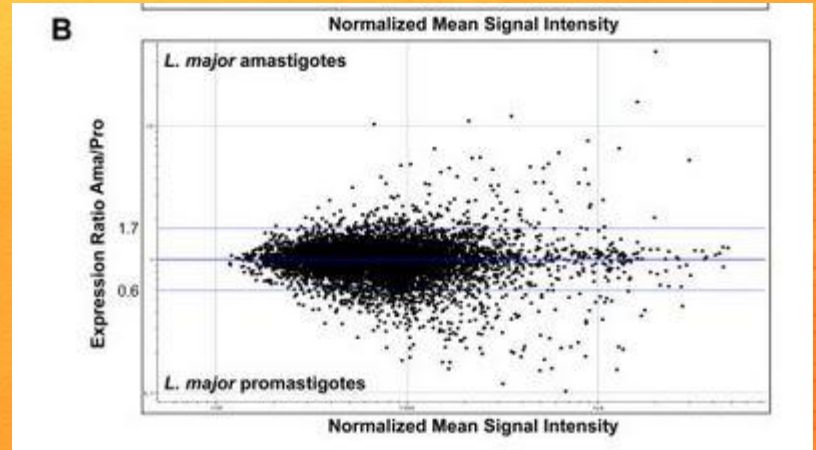


Figure 1: Whole-genome expression profiling of amastigote vs. promastigote signal intensities

- Differences in expression in RNA level
- Main variations: metabolism, cellular organization and transport

Datasources and Generating a GeneDB

- Uniprot XML from Uniprot's complete proteomes page
 - GOA from Uniprot Gene Ontology Associations downloads page
 - GO OBO-XML from Gene Ontology downloads page
-
- New Database created in Postgres SQL
 - Linked **GenMAPP Builder** to new database
 - Imported Data files, and exported a new Gene Database





Preparing Microarray Data

- Downloaded and organized data: Raw Data and SDRF
- Matched name to chip and labeled with species name
- Only kept name and expression ratio
- Accounted for swapped dyes
- Centered and scaled, found average log fold change
- Statistical analysis: P-value and T-test
- Formatted to be compatible with GenMAPP (.txt file)

Running GenMAPP using the Gene Database

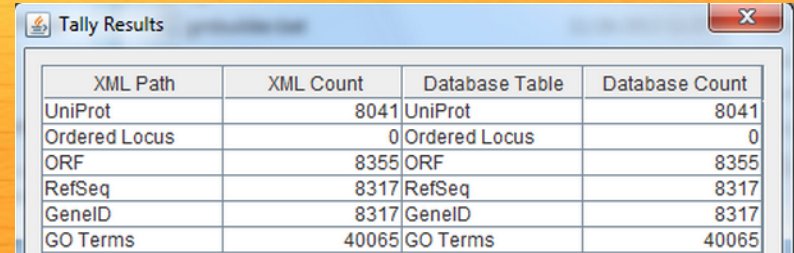
Import of microarray data yielded many errors

- Several changes were made to database to accomodate
- Many IDs simply missing in XML
- Ran MAPPFinder from GenMAPP homepage
 - Ran for 1 ½ hours
 - Had 1,820 errors
- Placed Genes under the go term “aromatic compound catabolic process” and compared them on a MAPP

Database Testing Report

General ID format: LmjF##[._]####

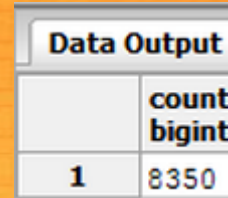
- 8355 ORFs in TallyEngine
- 8353 ORFs in XMLPipeDB Match
- 8350 ORFs in Postgres Queries
 - + 5 stragglers
- 16662 Original Row Counts
- Visual inspection
 - Date available for ½ of genes



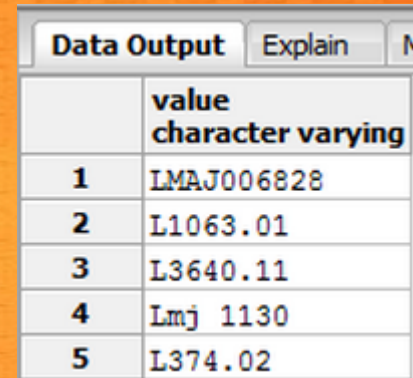
XML Path	XML Count	Database Table	Database Count
UniProt	8041	UniProt	8041
Ordered Locus	0	Ordered Locus	0
ORF	8355	ORF	8355
RefSeq	8317	RefSeq	8317
GeneID	8317	GeneID	8317
GO Terms	40065	GO Terms	40065

Total unique matches: 8353

C:\Users\keckuser\Downloads>



Data Output	
	count bigint
1	8350



Data Output		Explain	M
	value character varying		
1	LMAJ006828		
2	L1063.01		
3	L3640.11		
4	Lmj 1130		
5	L374.02		

Report on quantity and identity of gene IDs that did not make it into the database

- 1753 gene IDs were not present in the XML
- IDs not present in XML follow the form LmjF01.[0160-1983]
- Except IDs ending in zero are found in XML
- All IDs successfully uploaded into Postgres
- Certain IDs not exported into GenMAPP follow the form LmjF01.
[01][0-9][0-9]0
 - i. Database export did not properly assimilate IDs to the form LMJF_##_##### or LMJF.##_#####

Report on what changes need to be made to the GenMAPP Builder code in order to accommodate the second and third type of missing gene IDs

Initial change to Postgres was grabbing IDs from ORF
Code currently needs expansion to accommodate form:

LmjF01.[01][0-9][0-9]0

As well as a few outliers:

- LMAJ006828
- L1063.01
- L3640.11
- Lmj 11430
- L374.02

DNA microarray analysis results and statistical analysis

- Sanity check
- More than 960 results out of 19,201 t-tests performed had p-value of less than 0.05
- 2861 increased relative to control
- 2431 decreased relative to control

Average log fold change of > 0.25 and $p < 0.05$: 2861
Average log fold change of < -0.25 and $p < 0.05$: 2431

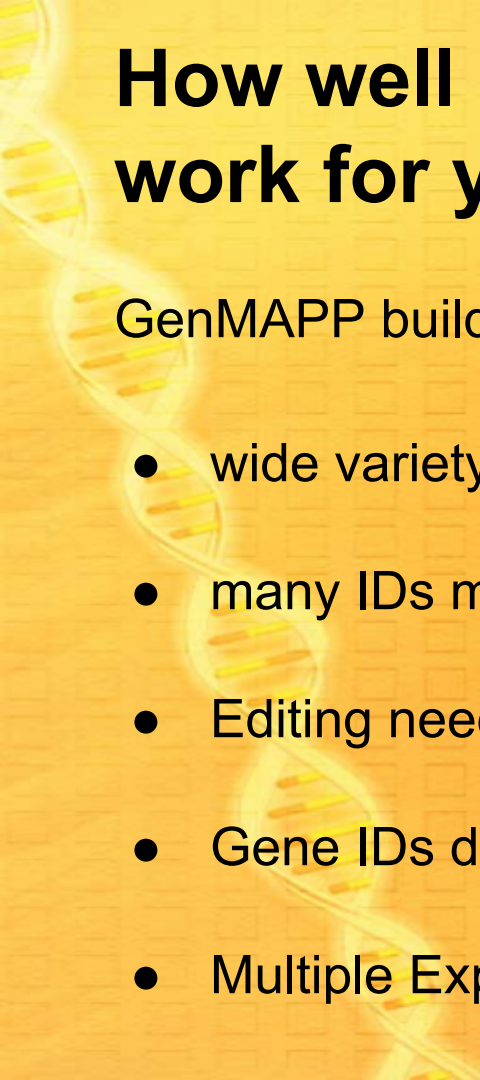
P-value less than 0.05: 5303

P-value less than 0.01: 2130

P-value less than 0.001: 317

P-value less than 0.0001: 0

- Genes for analysis in GenMAPP
- Increase in expression colored blue
- Decrease in gene expression colored purple



How well did the GenMAPP Builder process work for your species?

GenMAPP builder process was difficult:

- wide variety of ID formats
- many IDs missing in the XML
- Editing needed for Tally Engine
- Gene IDs difficult to place in XML
- Multiple Exports necessary to produce final database

MappFinder results

Top GO Terms

- 6 Terms were found to be significantly increased
 - Two groups were in the same family(membrane family).
- 8 groups were found to be significantly decreased.
 - All of these had to do with catabolic processes

Gene Ontology Results

catalytic activity	NESTED	46/187	24.6%	187/2113	8.8%	z score = 3.153	permute p = 0.002	adjusted p = 0.416
endonuclease activity	NESTED	3/3	100%	3/38	7.9%	z score = 3.636	permute p = 0.003	adjusted p = 0.293
DNA catabolic process	NESTED	3/3	100%	3/9	33.3%	z score = 3.636	permute p = 0.003	adjusted p = 0.293
aromatic compound catabolic process	NESTED	7/12	58.3%	12/144	8.3%	z score = 3.599	permute p = 0.003	adjusted p = 0.31
cellular nitrogen compound catabolic process	NESTED	7/12	58.3%	12/144	8.3%	z score = 3.599	permute p = 0.003	adjusted p = 0.31
nucleobase-containing compound catabolic process	NESTED	7/12	58.3%	12/142	8.5%	z score = 3.599	permute p = 0.003	adjusted p = 0.31
organic cyclic compound catabolic process	NESTED	7/13	53.8%	13/145	9%	z score = 3.328	permute p = 0.003	adjusted p = 0.377
heterocycle catabolic process	NESTED	7/13	53.8%	13/145	9%	z score = 3.328	permute p = 0.003	adjusted p = 0.377
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	NESTED	3/3	100%	3/32	9.4%	z score = 3.636	permute p = 0.003	adjusted p = 0.293
macromolecular complex	NESTED	1/42	2.4%	42/483	8.7%	z score = -2.88	permute p = 0.008	adjusted p = 0.914
cellular_component	NESTED	16/132	12.1%	132/1340	9.9%	z score = -2.439	permute p = 0.009	adjusted p = 0.949
cellular macromolecule catabolic process	NESTED	4/6	66.7%	6/71	8.5%	z score = 3.051	permute p = 0.011	adjusted p = 0.512
macromolecule catabolic process	NESTED	4/6	66.7%	6/81	7.4%	z score = 3.051	permute p = 0.011	adjusted p = 0.512
non-membrane-bounded organelle	NESTED	1/33	3%	33/295	11.2%	z score = -2.414	permute p = 0.011	adjusted p = 0.952
intracellular non-membrane-bounded organelle	NESTED	1/33	3%	33/295	11.2%	z score = -2.414	permute p = 0.011	adjusted p = 0.952
hydrolase activity	NESTED	20/68	29.4%	68/863	7.9%	z score = 2.564	permute p = 0.012	adjusted p = 0.944
nuclease activity	NESTED	3/4	75%	4/68	5.9%	z score = 2.914	permute p = 0.015	adjusted p = 0.91
intramolecular oxidoreductase activity	NESTED	3/4	75%	4/17	23.5%	z score = 2.914	permute p = 0.019	adjusted p = 0.91
dodecenoyl-CoA delta-isomerase activity	NESTED	2/2	100%	2/2	100%	z score = 2.964	permute p = 0.024	adjusted p = 0.896
intramolecular oxidoreductase activity, transposing C=C bonds	NESTED	2/2	100%	2/6	33.3%	z score = 2.964	permute p = 0.024	adjusted p = 0.896

Clicking on a specific term will locate that term in the hierarchy.

Paper found variations in genes involving metabolism, cellular organization and biogenesis, and transport

- 25% of differentially expressed genes b/w life stages involved in metabolism (both species)
- Promastigotes upregulated genes involving carbohydrate and glucose metabolism
- Studies found that differentiating parasite shifts main energy source:
 - Promastigote= glucose
 - Amastigote= fatty acids and amino acids

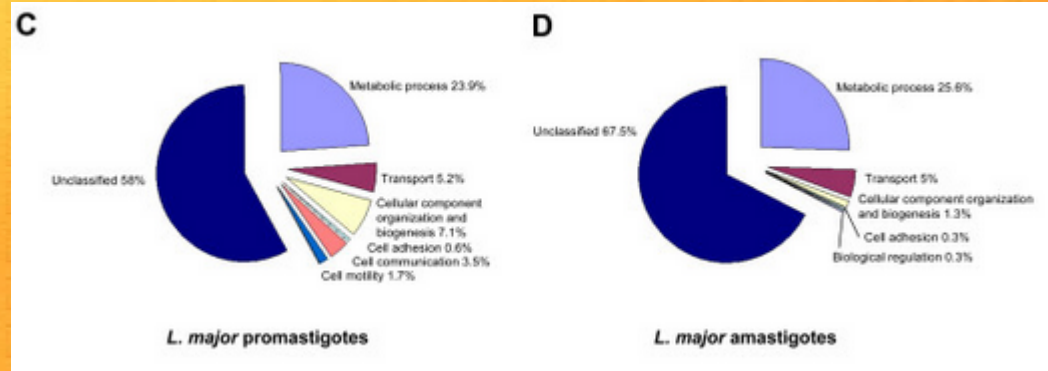
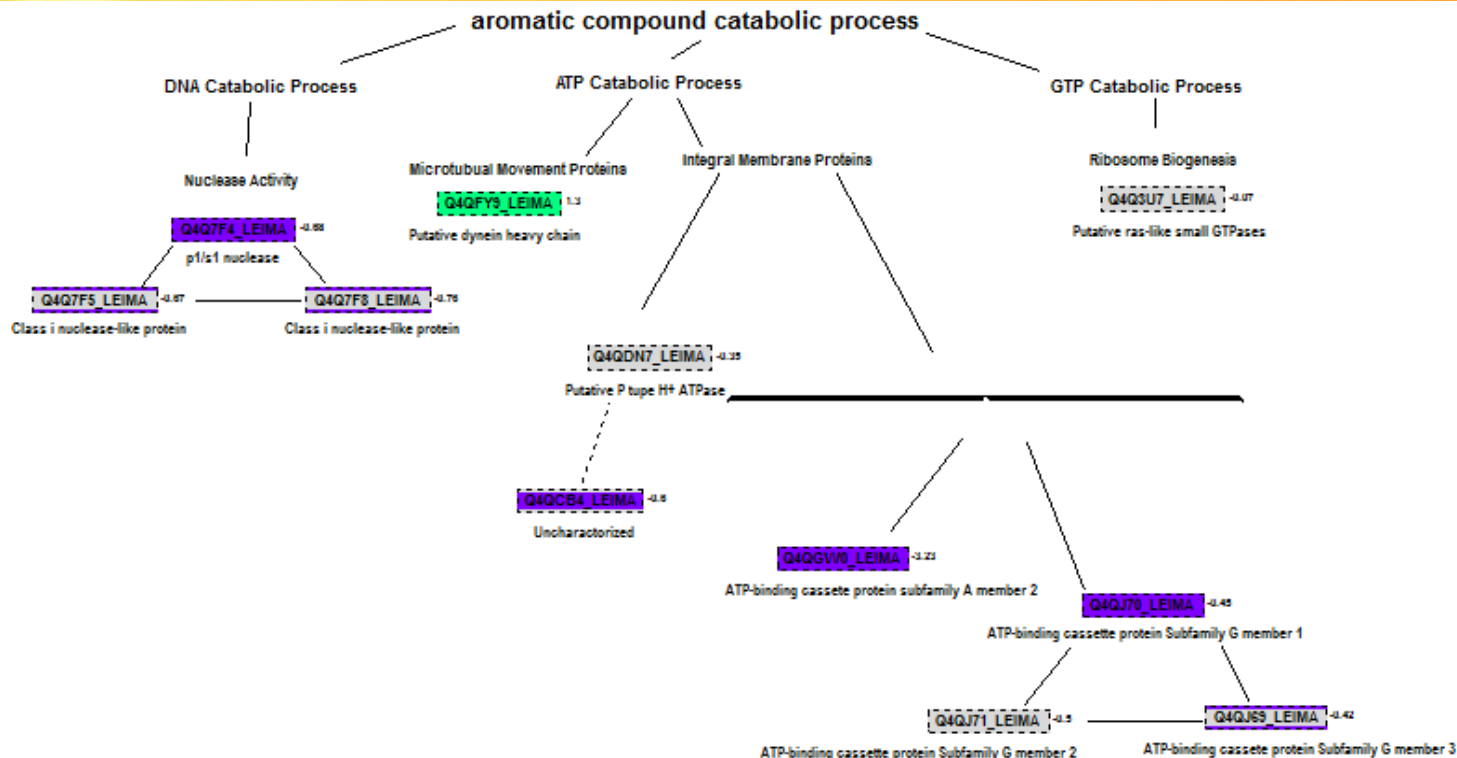


Figure 2: Distribution of *L. major* differentially expressed genes according to Gene Ontology function categories

- Differences in gene expression: cellular organization, biogenesis and cell motility between life stages
 - Motile flagellated promastigotes
 - Amastigotes increased expression in lysosomal proteins

GenMAPP map



Gene Database

Leshmans/CDB 28112013 Lema_Caba.gdb

Expression Dataset

Name: Leshmans/ComplidStatAnalysis_MJH filtered

Color Set: Amasigote vs. PRoma

Gene Value: Amasigote vs. PRoma: Avg. Log₂ AL

Legend: Amasigote vs. PRoma

- Increased
- Decreased
- No criteria met
- Not found

References

Rochette, A., Raymond F., Ubeda JM., Smith M., Messier N., Boisvert S., Rigault P., Corbeil J., Ouellette M., Papadopoulou B. Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species. *BMC Genomics*, (2008 May 29). **9**:255.



Acknowledgments

Dr. Dionisio

Dr. Dahlquist