Automated integration of Mass-Spectrometry based Proteomics evidence for improvement of gene annotations

Ritesh Krishna
Institute of Integrative Biology
University of Liverpool

July 12, 2011
British Society for Proteome Research Meeting 2011, Cambridge, UK

Objective

- Use of mass-spectrometry data to confirm that predicted gene models are translated into proteins
- Use the dataset to examine whether there are supporting evidences for alternative gene predictions at particular loci
- To check whether the same data can be used to identify novel genes

EuPathDB
Version 2.8
22 Nov 10
Eukaryotic Pathogen Database Resources

A EuPathDB Project

Gene ID: cgd7_230 Gene Text Search: synth*

About EuPathDB | Help | Contact Us | Login | Register

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community

Data Summary

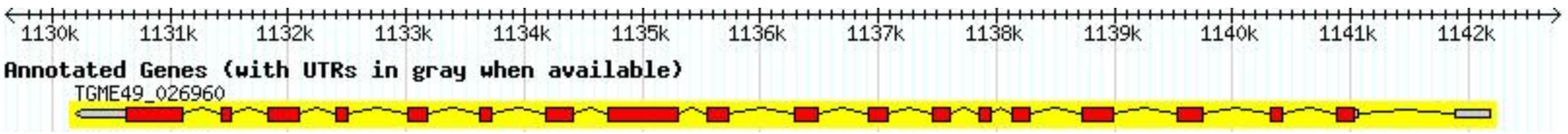
News

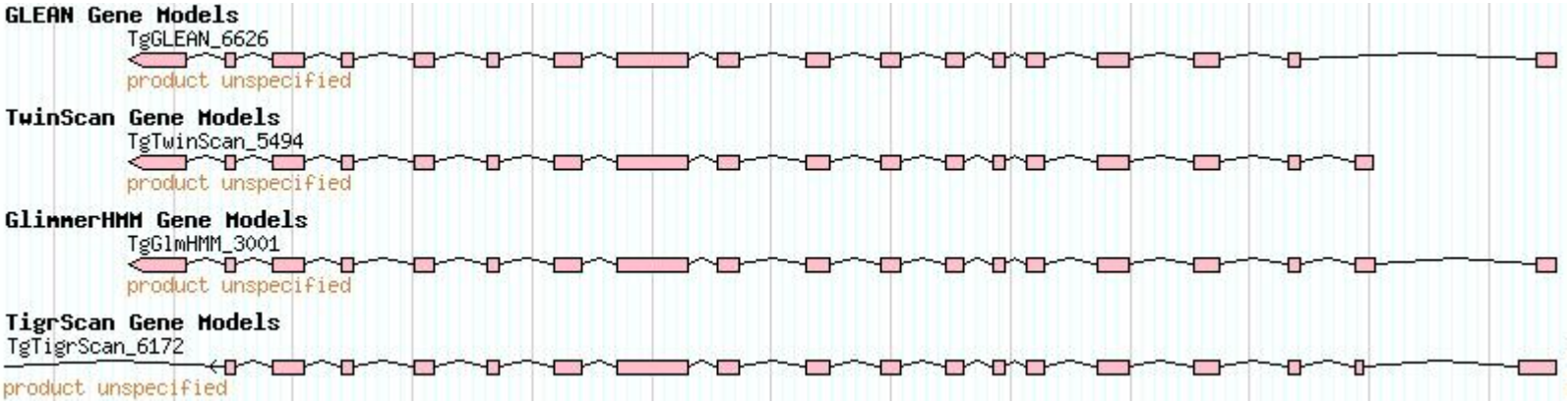
- **18 November 2010** AmoebaDB 1.3 Released
- **18 November 2010** PlasmoDB 7.1 Released
- **18 November 2010** TriTrypDB 2.5 Released

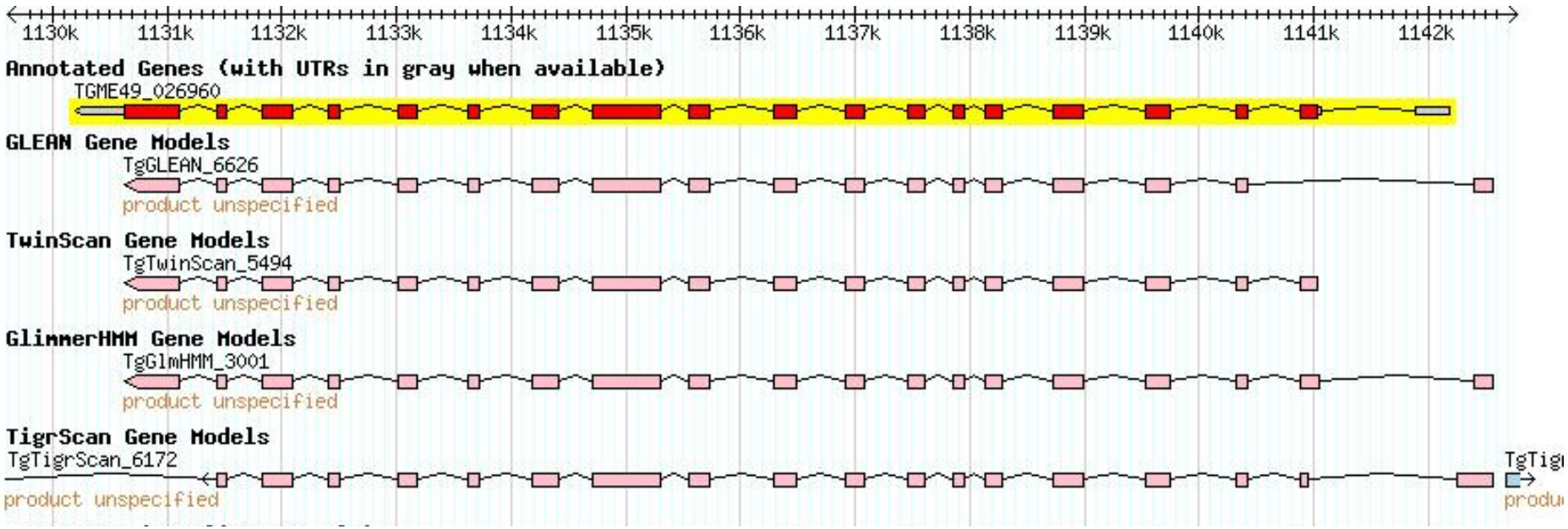
We are pleased to announce our 2011 **EuPathDB Workshop**, June 12-15, 2011 in Athens, GA, USA. For more information and to apply click here. **Application deadline is February 1, 2011**.

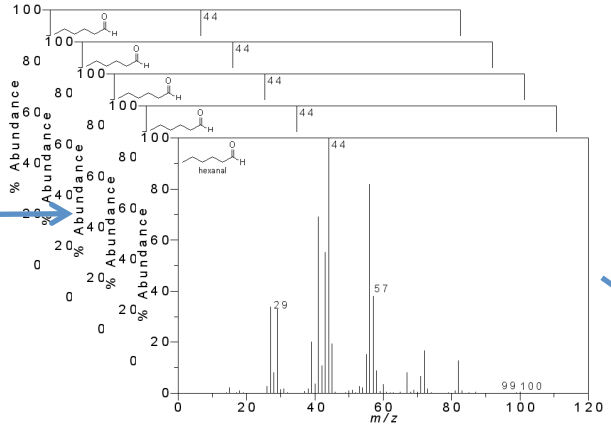
EuPathDB Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Diseases is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens (*Cryptosporidium*, *Encephalitozoon*, *Entamoeba*, *Enterocytozoon*, *Giardia*, *Leishmania*, *Neospora*, *Plasmodium*, *Toxoplasma*, *Trichomonas* and *Trypanosoma*).

AmoebaDB CryptoDB GiardiaDB MicrosporidiaDB PlasmoDB ToxoDB TrichDB TriTrypDB






1130k
1131k
1132k
1133k
1134k
1135k
1136k
1137k
1138k
1139k
1140k
1141k
1142k
Annotated Genes (with UTRs in gray when available)
TGME49_026960
GLEAN Gene Models
TgGLEAN_6626
product unspecified
TwinScan Gene Models
TgTwinScan_5494
product unspecified
GlimmerHMM Gene Models
TgGlmHMM_3001
product unspecified
TigrScan Gene Models
TgTigrScan_6172
product unspecified



Search Engines

Mascot

MATRIX SCIENCE

Ommsa

OMSSA

X!Tandem

x!

...etc.

Gene Models

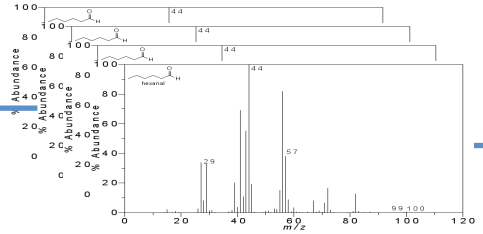
swissprot SIB EBI

EuPathDB
Eukaryotic Pathogen Database Resources
http://EuPathDB.org

... Any protein sequence database of your choice

Protein Sequence Database

Peptide-Spectrum matches and Protein Identification



OMSSA

X!

.csv

Protein Sequence Database

.xml

mzIdentML Omssa

mzIdentML X!Tandem

False Discovery Rate control and consensus between different search engines

Reranked Peptide-Spectrum Matches and adjusted FDR Scores

Modes of operation

Single input file

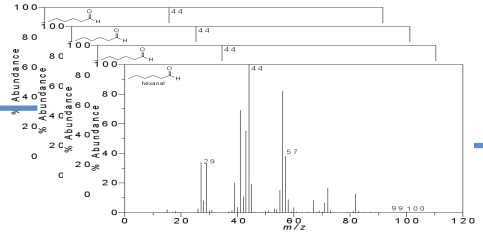
Whole directory of input files

Single machine

A cluster of machines

FASTA

GFF



OMSSA

X!

.csv

Protein Sequence Database

.xml

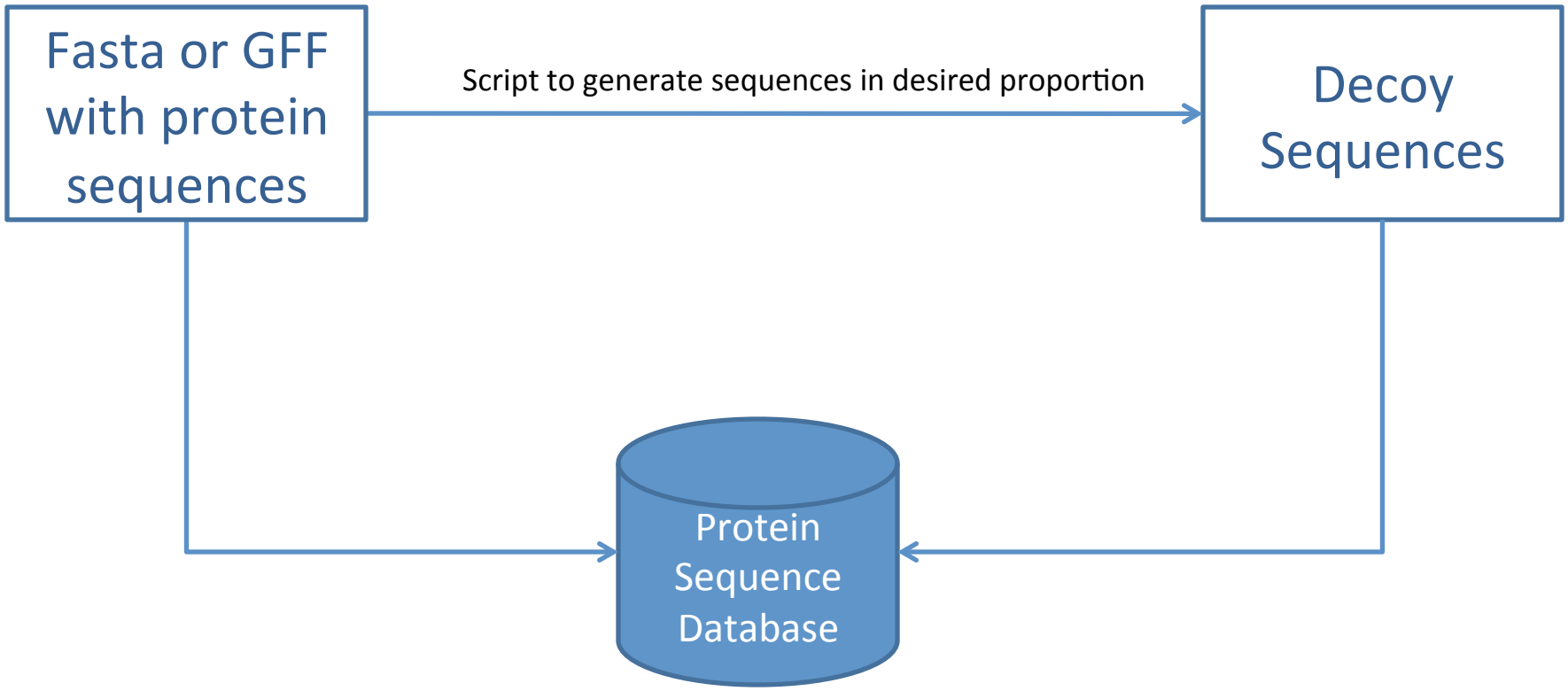
mzIdentML Omssa

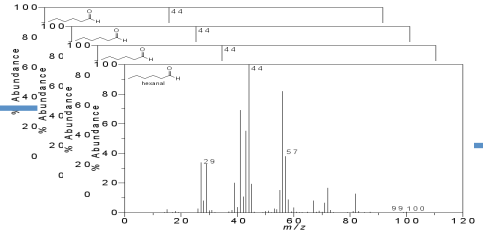
mzIdentML X!Tandem

False Discovery Rate control and consensus between different search engines

Reranked Peptide-Spectrum Matches and adjusted FDR Scores

Protein Sequence Database





OMSSA

X!

.csv

Protein Sequence Database

.xml

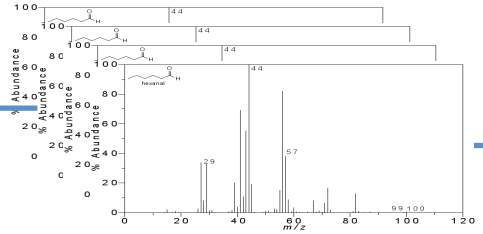
mzIdentML Omssa

mzIdentML X!Tandem

False Discovery Rate control and consensus between different search engines

Reranked Peptide-Spectrum Matches and adjusted FDR Scores

Controlling the search engines –
- Specification of digestive enzyme used
- Specification of parent and fragment tolerances
- Specification of maximum missed cleavage allowed
- Specification of Post translational modifications



OMSSA

X!

.csv

Protein Sequence Database

.xml

mzIdentML Omssa

mzIdentML X!Tandem

False Discovery Rate control and consensus between different search engines

Reranked Peptide-Spectrum Matches and adjusted FDR Scores

-- Increase confidence by creating a consensus between Omssa and X!Tandem results

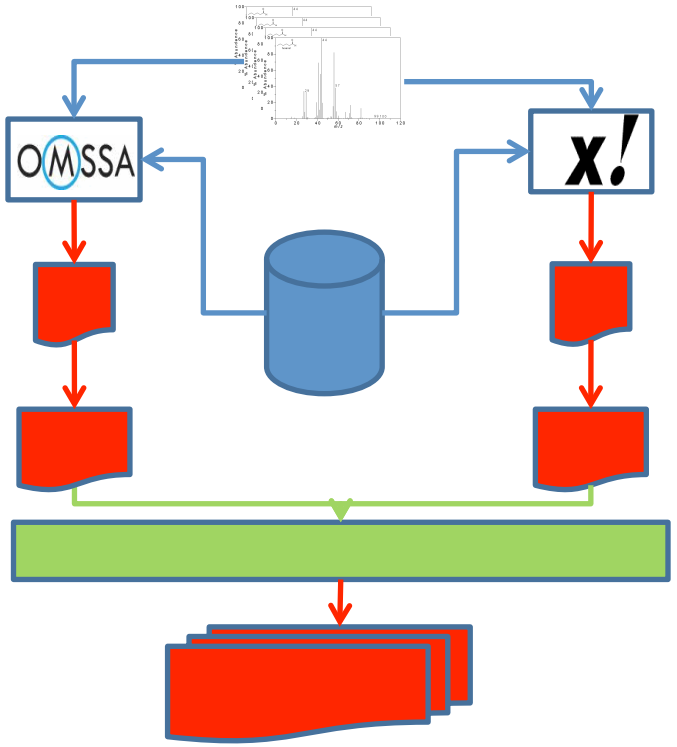
-- Use of Decoy database to compute False Discovery Rate

-- Details of the algorithm available at the following publications

Jones, A. R., Siepen, J. A., Hubbard, S. J. and Paton, N. W. (2009), Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *PROTEOMICS*, 9: 1220–1229.

Wedge,D. C., Krishna, R., Blackhurst, P., Siepen, J.A., Jones, A.R., Hubbard, S.J. (2011), **FDRAnalysis: A Tool for the Integrated Analysis of Tandem Mass Spectrometry Identification Results from Multiple Search Engines.** *Journal of Proteome Research, 10* (4), 2088-2094

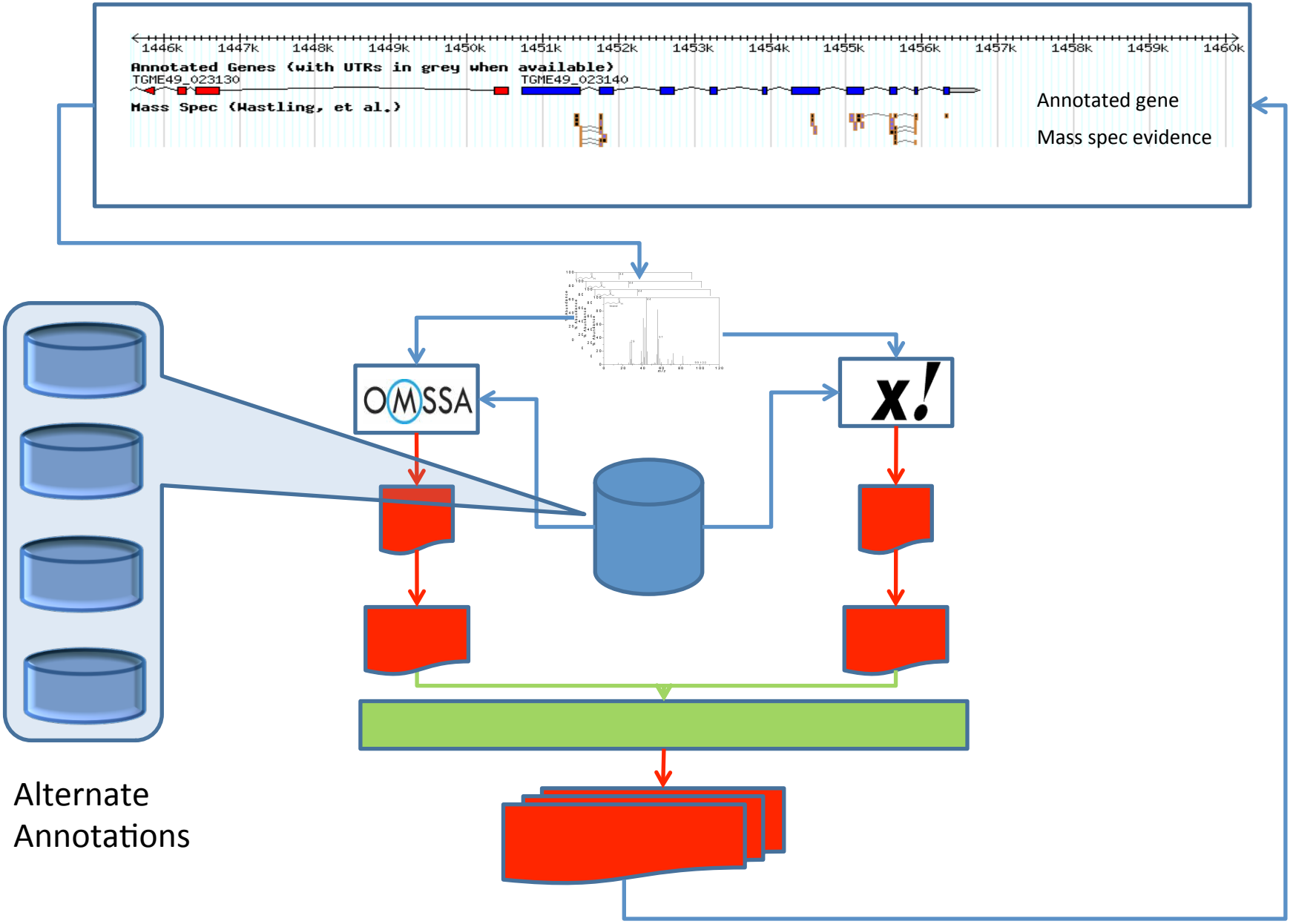
FASTA → Text Files

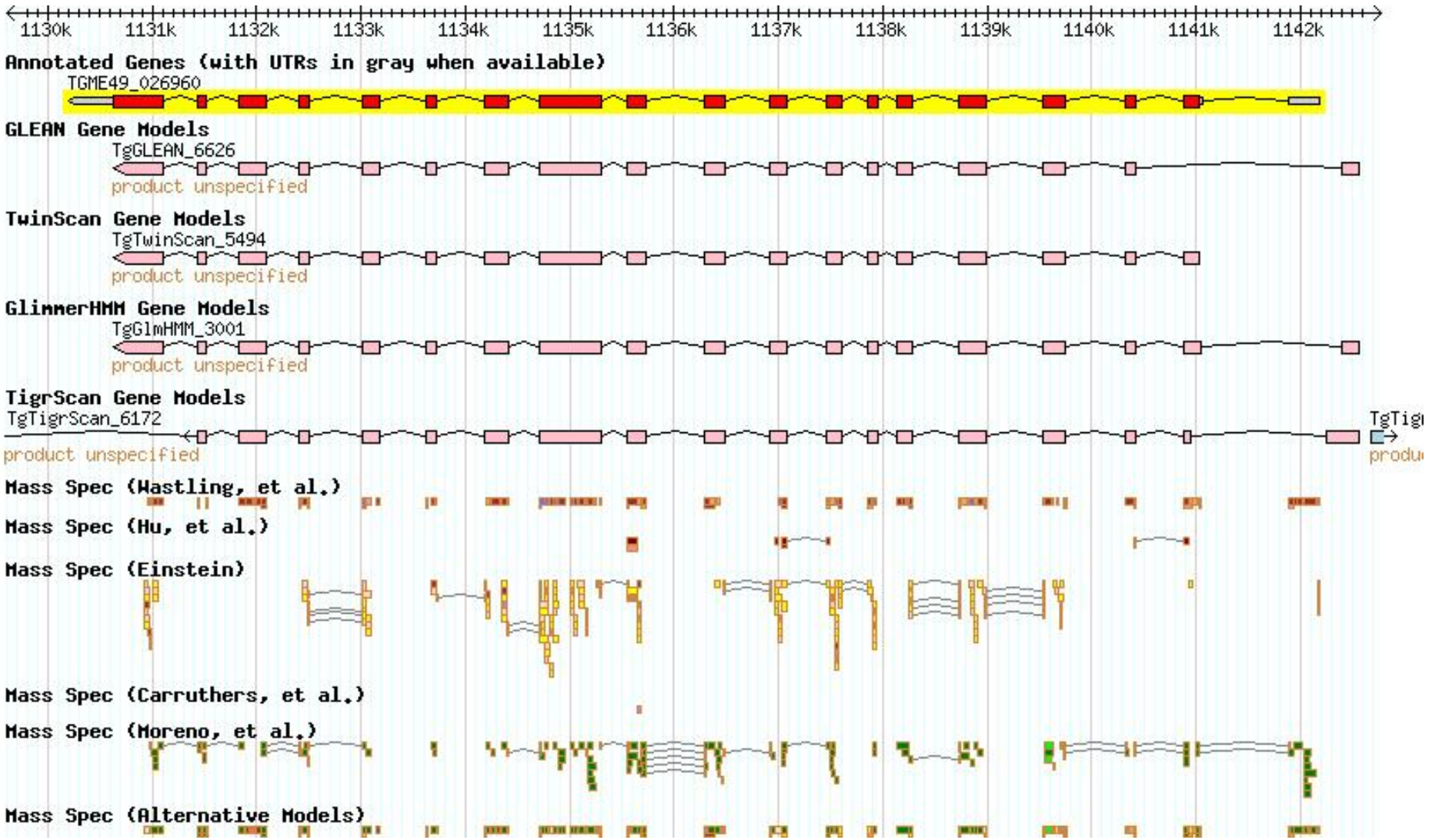


GFF → GFF

Output for each input peak file
Omssa output in native CSV format
Omssa output converted to mzIdentML format
X!Tandem output in native XML format
X!Tandem output converted to mzIdentML format
Summary file reporting the consensus between Omssa and X! Tandem results – Re-ranked peptide-spectrum matches and adjusted FDR scores in a tab delimited text file

Output for complete dataset
Each input peak file has its own output directory with above listed files
A text file summarizing all the identified proteins, respective peptides and relevant statistics
A GFF3 file if operating in the GFF mode


1446k 1447k 1448k 1449k 1450k 1451k 1452k 1453k 1454k 1455k 1456k 1457k 1458k 1459k 1460k
Annotated Genes (with UTRs in grey when available)
TGME49_023130
TGME49_023140
Mass Spec (Wastling, et al.)
Annotated gene
Mass spec evidence
OMSSA
X!
Alternate
Annotations


1130k
1131k
1132k
1133k
1134k
1135k
1136k
1137k
1138k
1139k
1140k
1141k
1142k
Annotated Genes (with UTRs in gray when available)
TGME49_026960
GLEAN Gene Models
TgGLEAN_6626
product unspecified
TwinScan Gene Models
TgTwinScan_5494
product unspecified
GlimmerHMM Gene Models
TgGlmHMM_3001
product unspecified
TigrScan Gene Models
TgTigrScan_6172
product unspecified
Mass Spec (Wastling, et al.)
Mass Spec (Hu, et al.)
Mass Spec (Einstein)
Mass Spec (Carruthers, et al.)
Mass Spec (Moreno, et al.)
Mass Spec (Alternative Models)

Current status

Test organisms

Developers resources

Java

perl

Google code

Acknowledgements

UNIVERSITY OF LIVERPOOL

Dr. Andy Jones

Prof. Jonathan Wastling

Dr. Dong Xia

Prof. David Roos

Dr. Brian Brunk

Dr. Omar Harb

Contact e-mail – Ritesh.Krishna@liverpool.ac.uk