# THE GALAXY FRAMEWORK AS A BIOINFORMATICS SOLUTION FOR PROTEOMICS AND MULTI-OMICS STUDIES.

**Pratik Jagtap**
*Managing Director,*
*Center for Mass Spectrometry and Proteomics*
**UNIVERSITY OF MINNESOTA**

# ‘OMICS’ RESEARCH

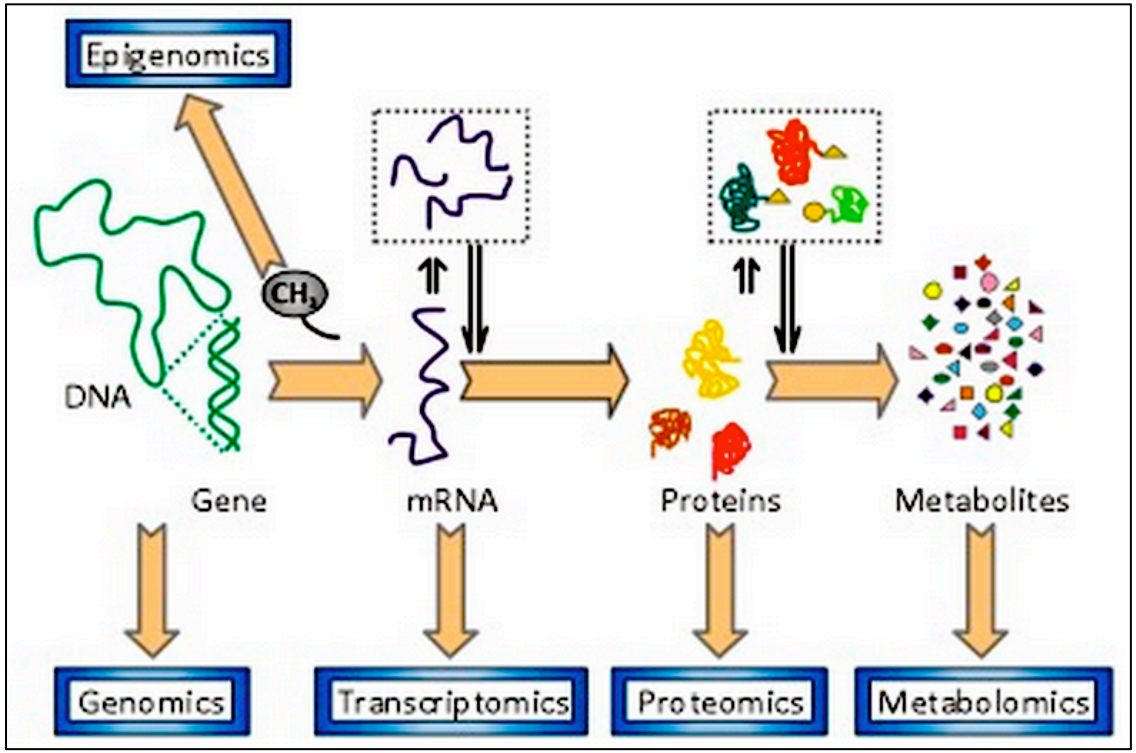



*Image Source: Goodacre, J. Exp. Bot 2005.*

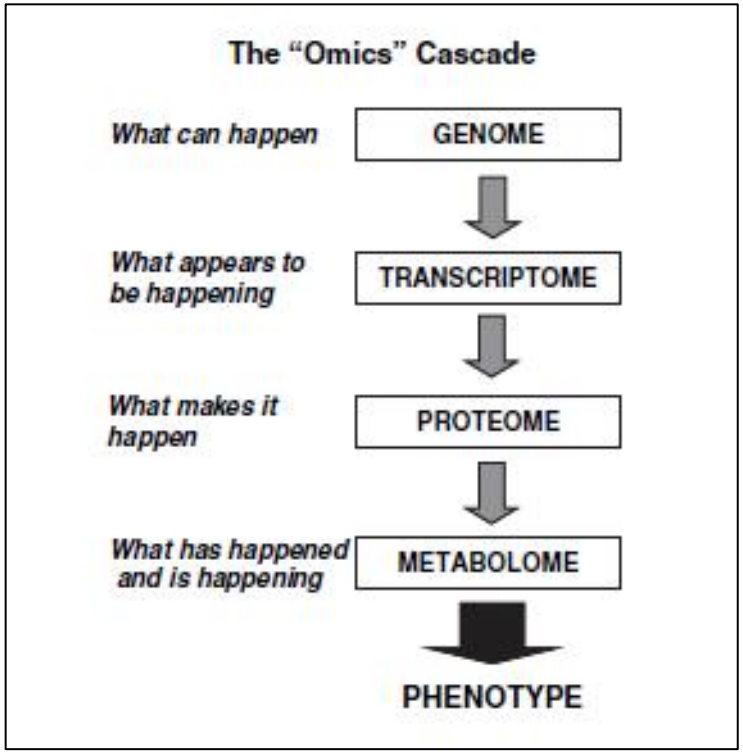
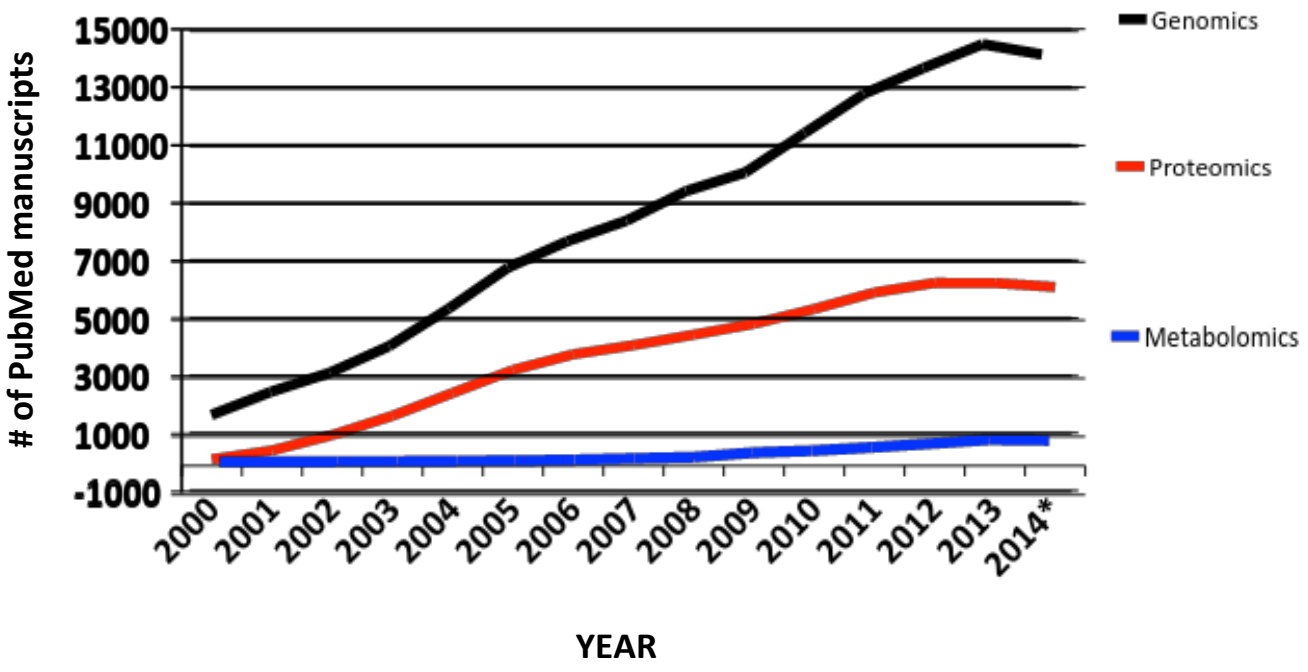


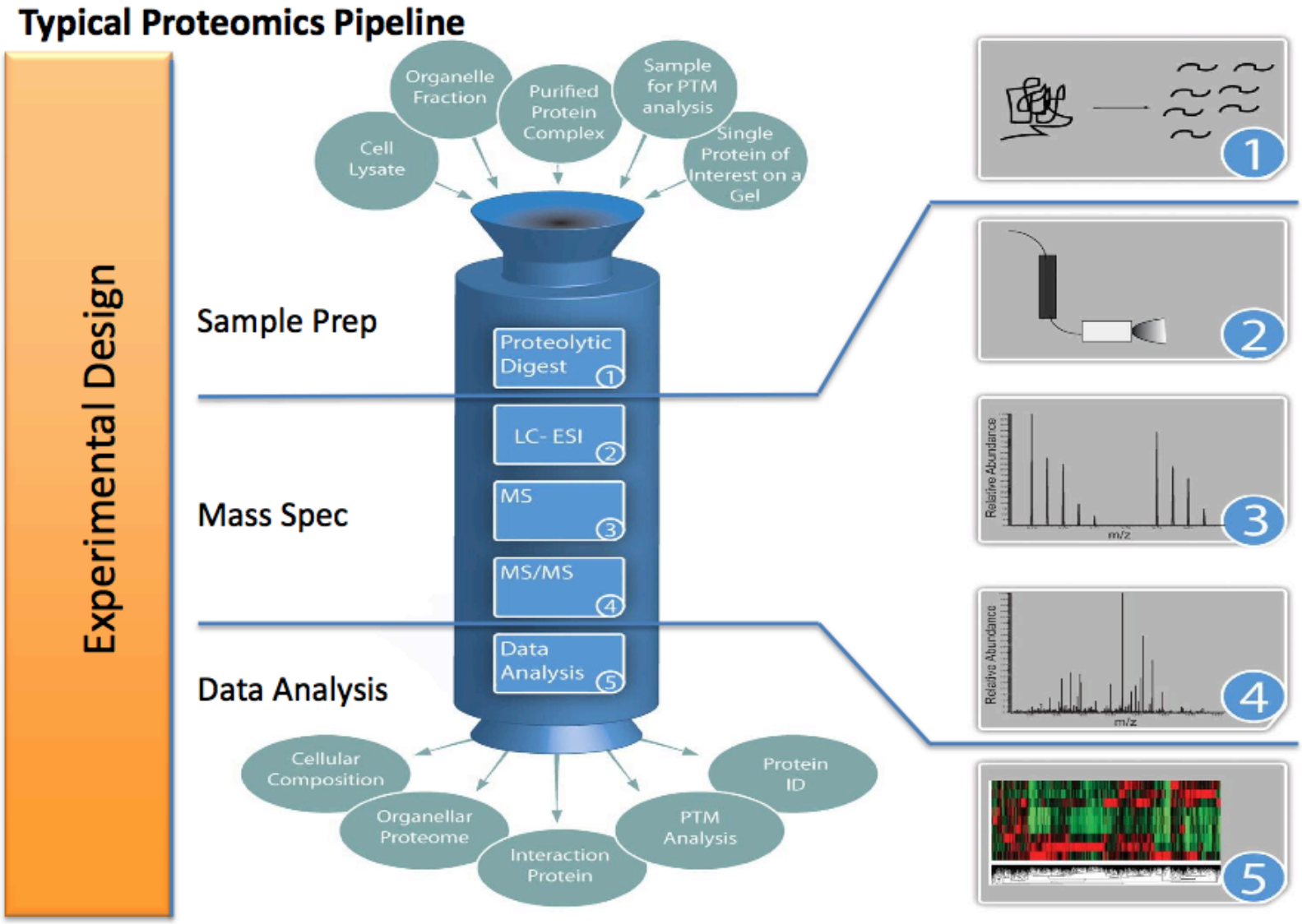

*Image Source: http://fluorous.com/images/omics.JPG*

# TRENDS IN OMICS RESEARCH



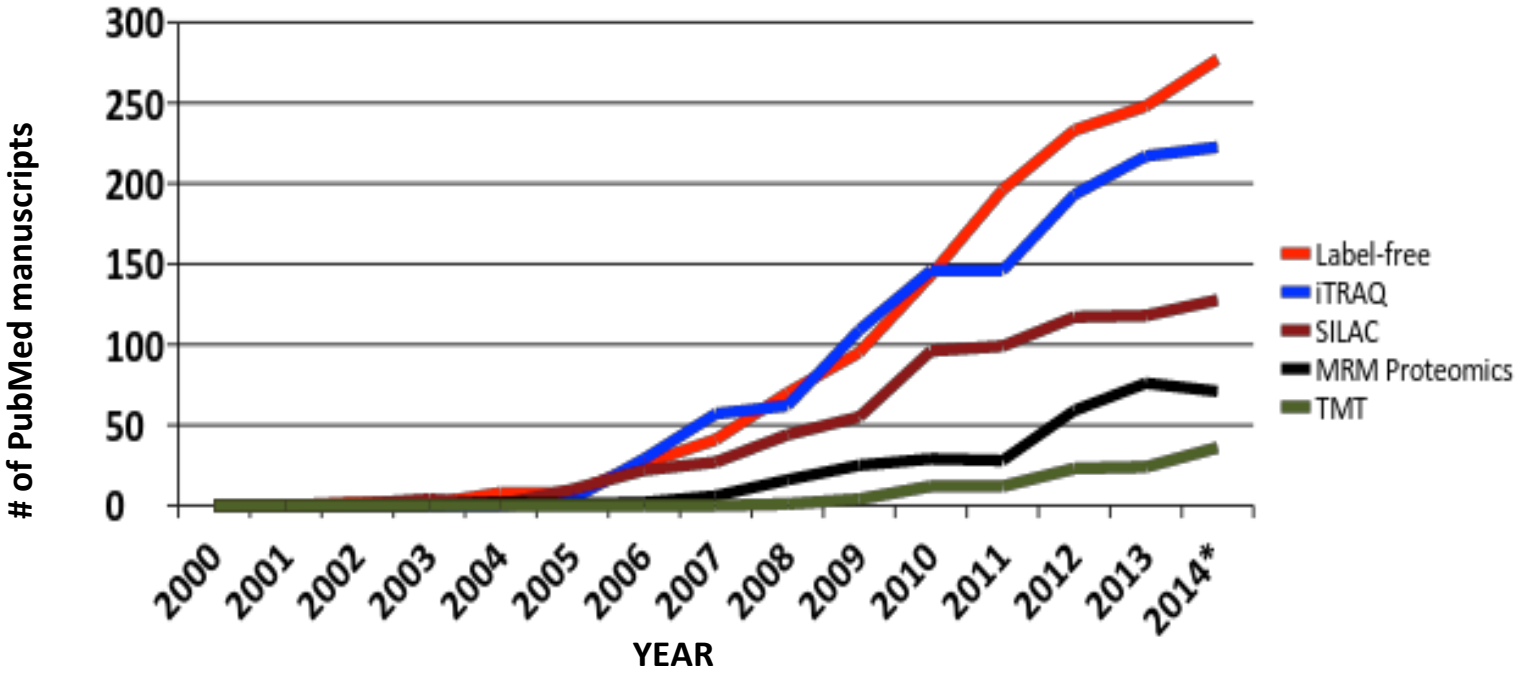

- **Genomics: Established Technology.**
- **Proteomics: Standard Technology.**
- **Metabolomics: Emerging Technology.**

# PROTEOMICS WORKFLOW



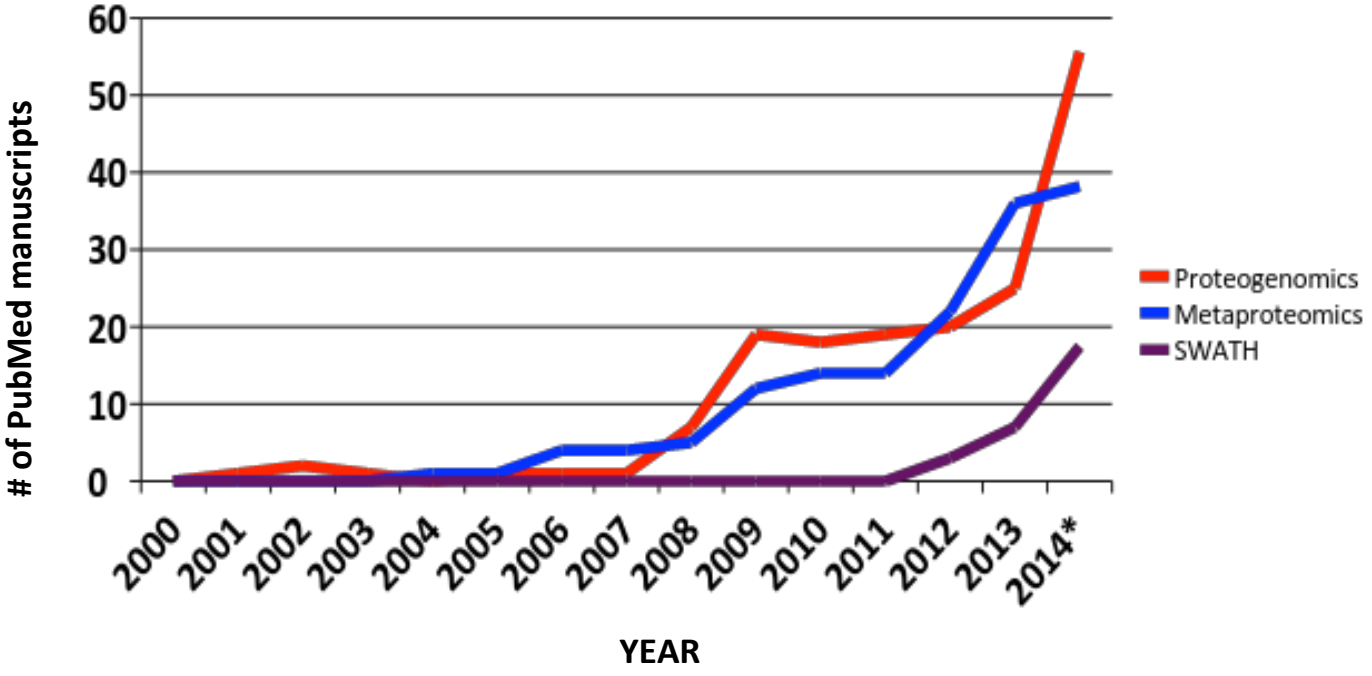

Adapted from Walther T, Mann M. JCB 2010;190:491-500

# QUANTITATIVE PROTEOMICS



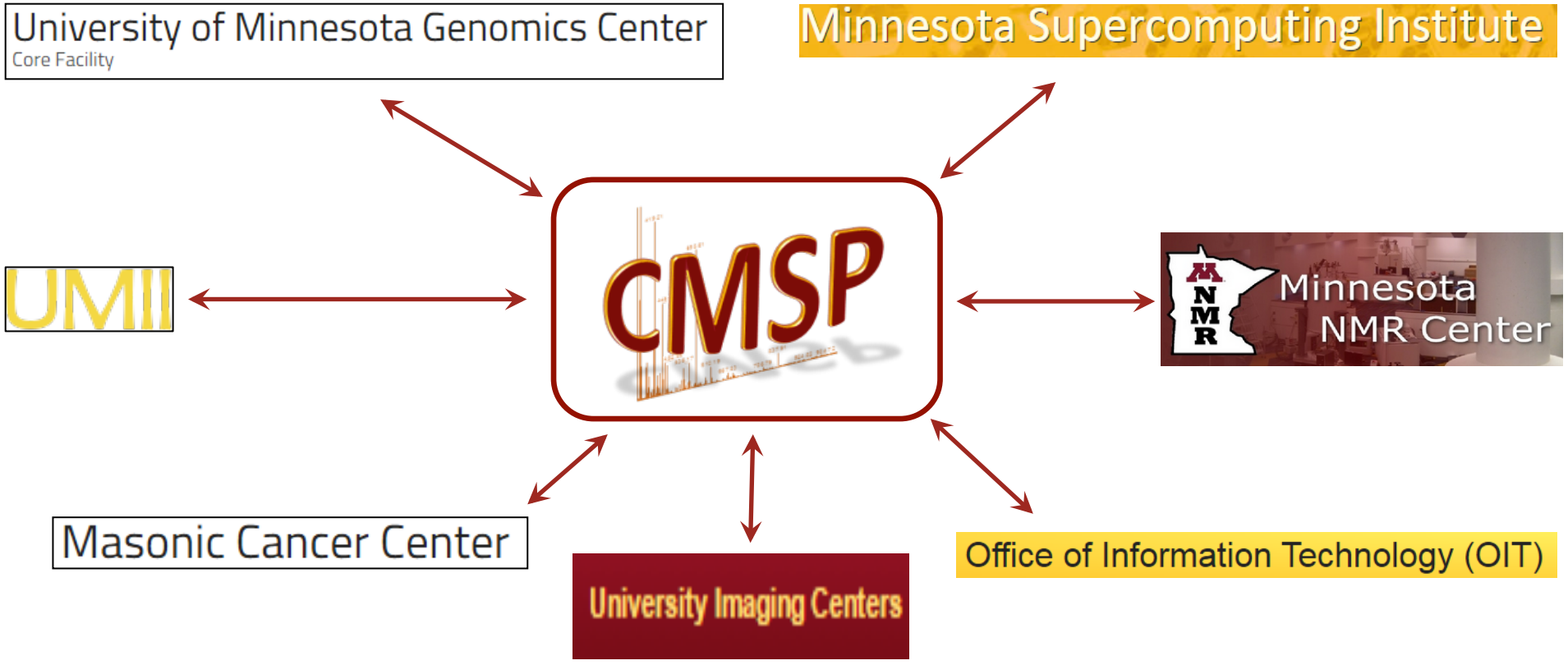

- Label-free quantitation. (AUC for MS1)
- iTRAQ (MS/MS reporter ion)
- SILAC (precursor MS ion)
- TMT (MS/MS reporter ion)
- MRM (Targeted Proteomics)

# EMERGING FIELDS IN PROTEOMICS RESEARCH




- **Next Generation Proteomics.**
  - **Proteogenomics (Uses data from RNASeq data)**
  - **Metaproteomics (uses metagenomics data)**
  - **Data-independent acquisition (For example SWATH)**

# PARTNERSHIP WITH UNIVERSITY OF MINNESOTA RESEARCH UNITS






Center for Mass Spectrometry and Proteomics | Phone | (612)625-2280 | (612)625-2279

# GALAXY-P : IMPLEMENTATION OF PROTEOMICS TOOLS WITHIN GALAXY ENVIRONMENT.


## Galaxy-P: A new community-based informatics paradigm for MS-based proteomics

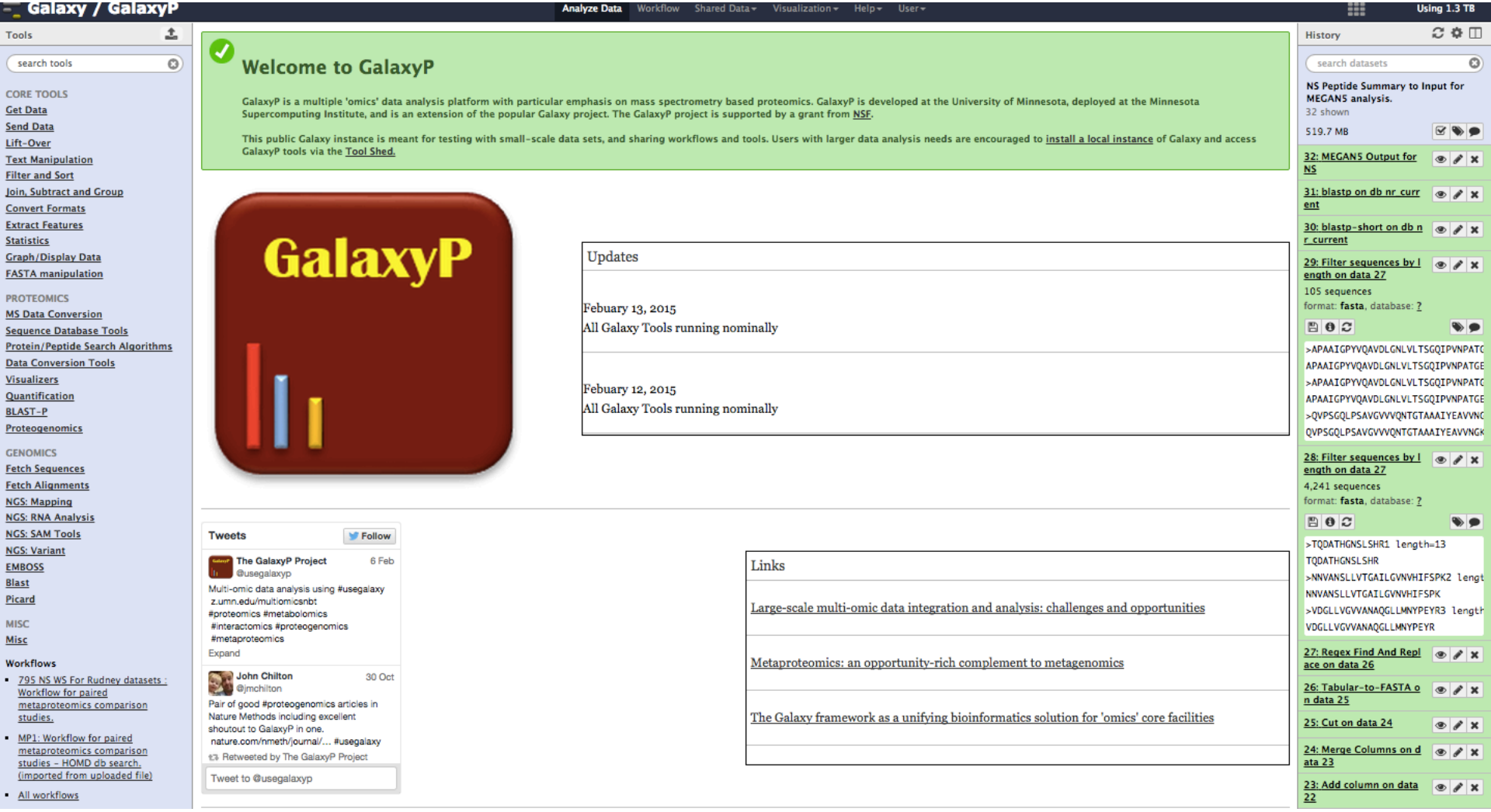
- Funded via the NSF Advances in Biological Informatics program
- 3 years of funding; effective July 15, 2012-June 30, 2015

**Grant objective in a nutshell:**

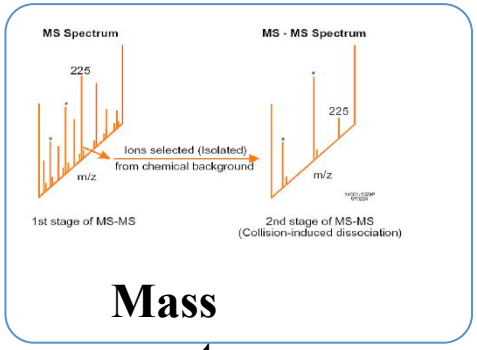
We propose to extend the Galaxy framework for genomics by deploying and integrating a series of key software programs for MS-based proteomics data analysis, thus creating Galaxy Tool Modules for Proteomics which we refer to as Galaxy-P

Project-based strategy for Galaxy-P development:
Collaborate with biological researchers with “real” projects to guide developments.

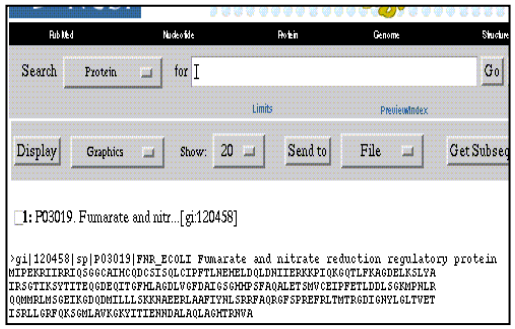
# GALAXY-P : IMPLEMENTATION OF PROTEOMICS TOOLS WITHIN GALAXY ENVIRONMENT.



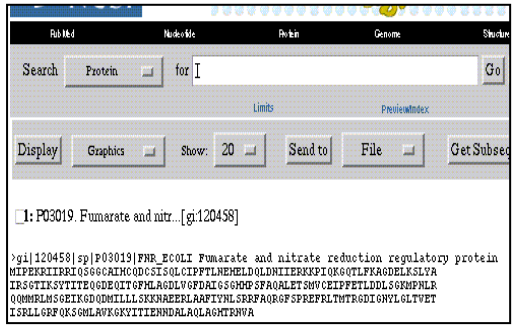
# PROTEOGENOMICS AND METAPROTEOMICS




**Mass spectrum**



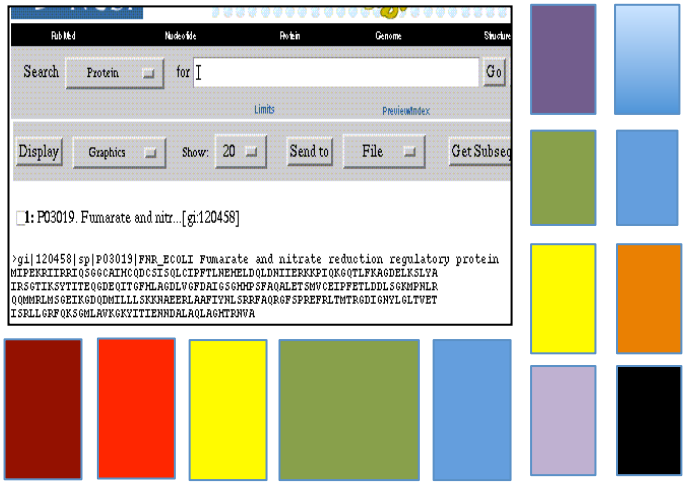

**Reference Protein Database from genomic annotation**




**RNASeq data**

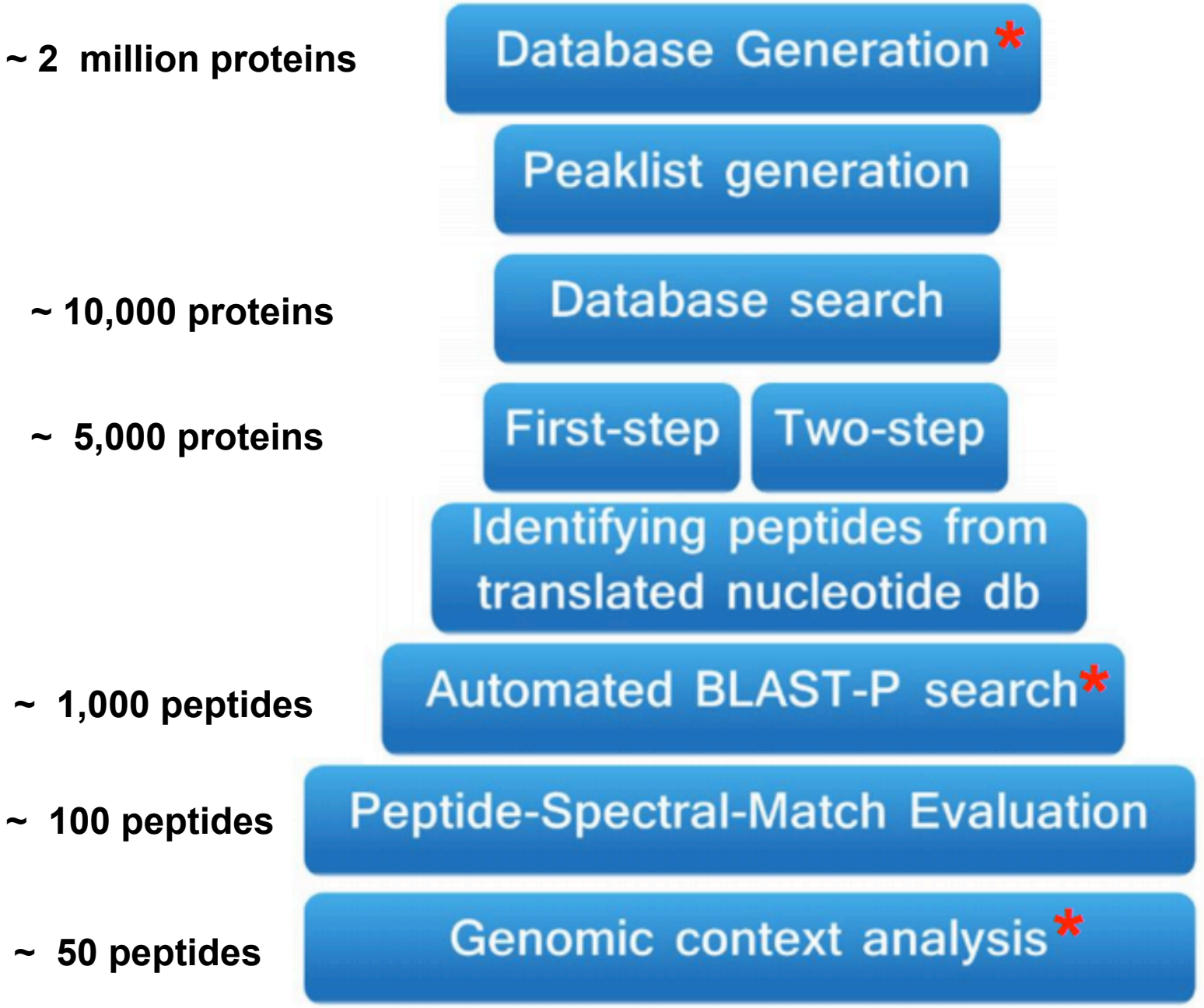
**Genome six-frame translation**

**cDNA three-frame translation**

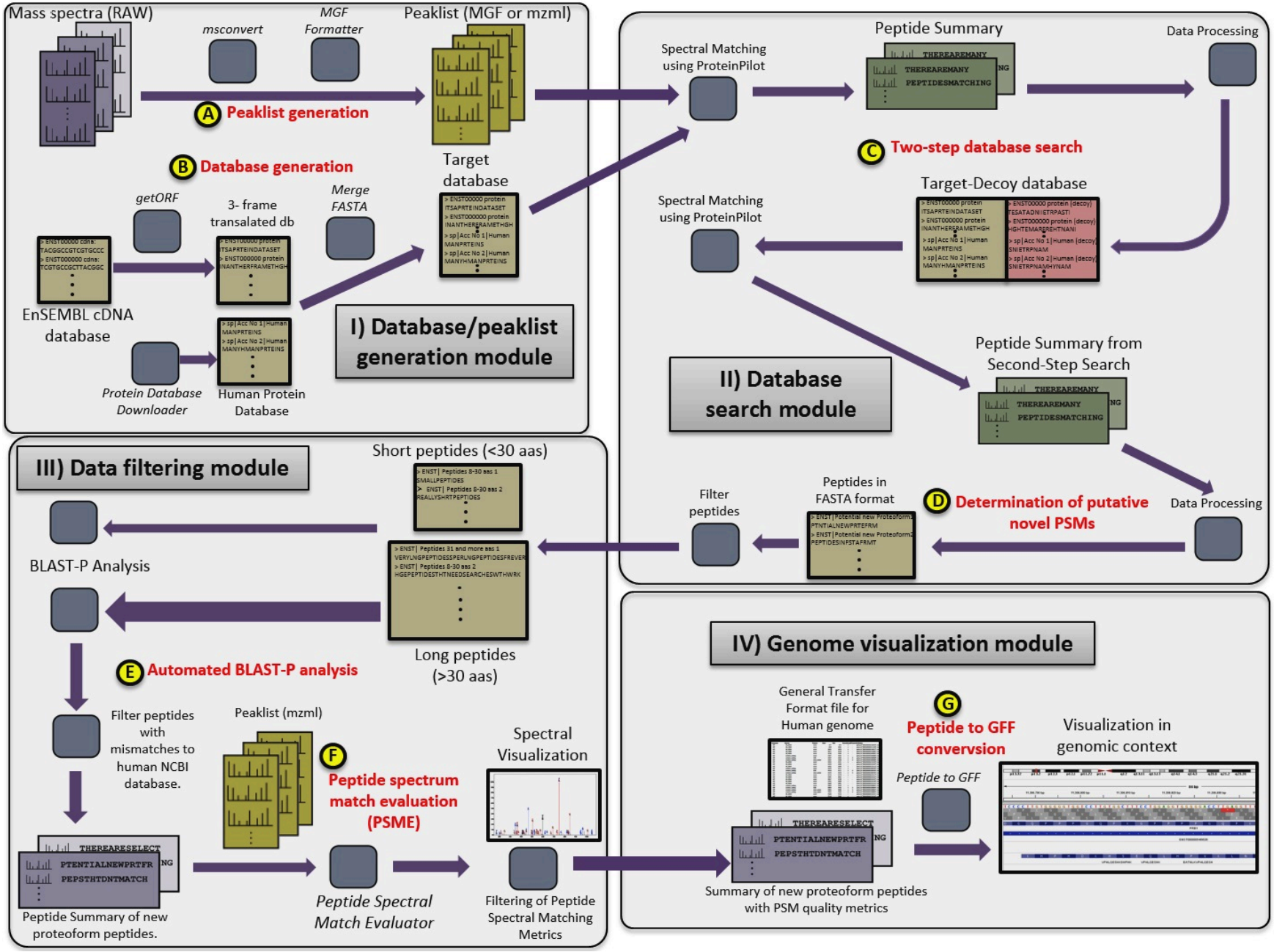



**Metagenomic sequences**

# PROTEOGENOMICS: STEPS INVOLVED



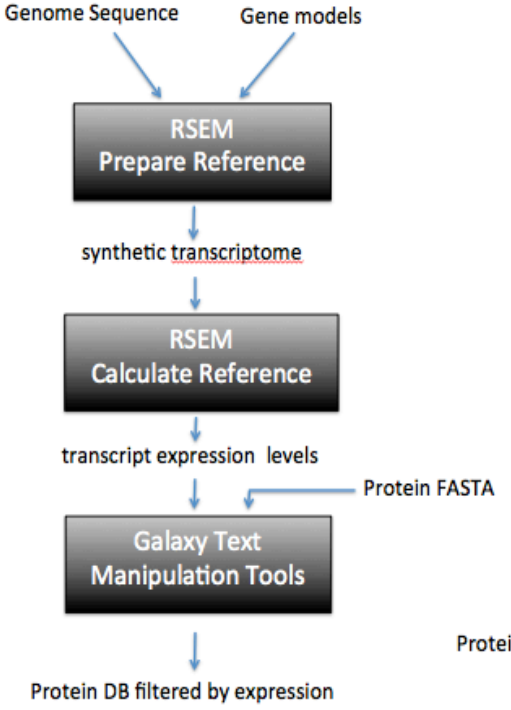
# PROTEOGENOMICS WORKFLOW



# RNASeq DERIVED PROTEOMIC DATABASES

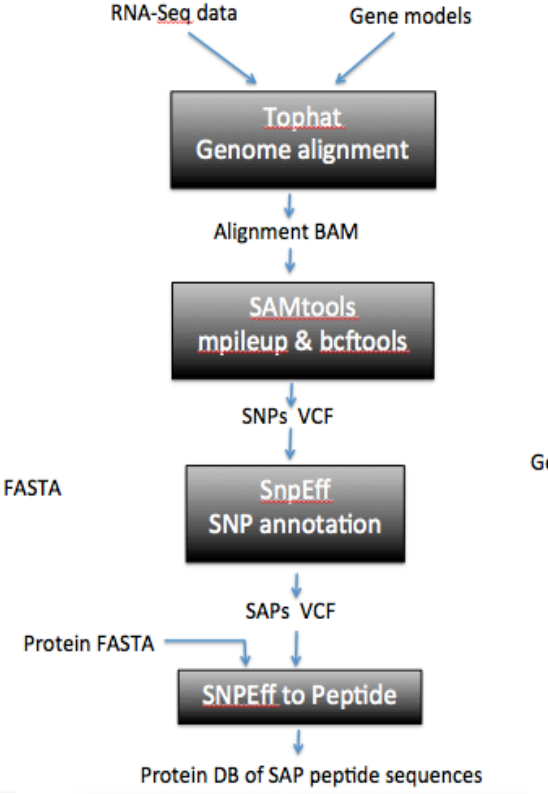
## Reduced Database

RSEM determines the RNA-Seq transcripts expressed at detectable levels. Proteins from transcripts that are not expressed are filtered out.



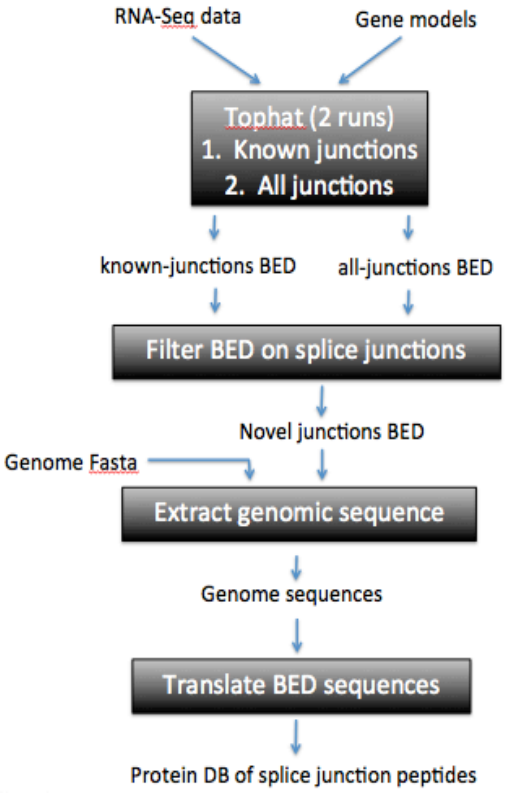

## SAP Database

RNA-Seq reads are aligned to the reference genome with tophat. SAMtools identifies variant DNA bases. SNPEff annotates the variants with effects to genes and proteins.




## Splice Database

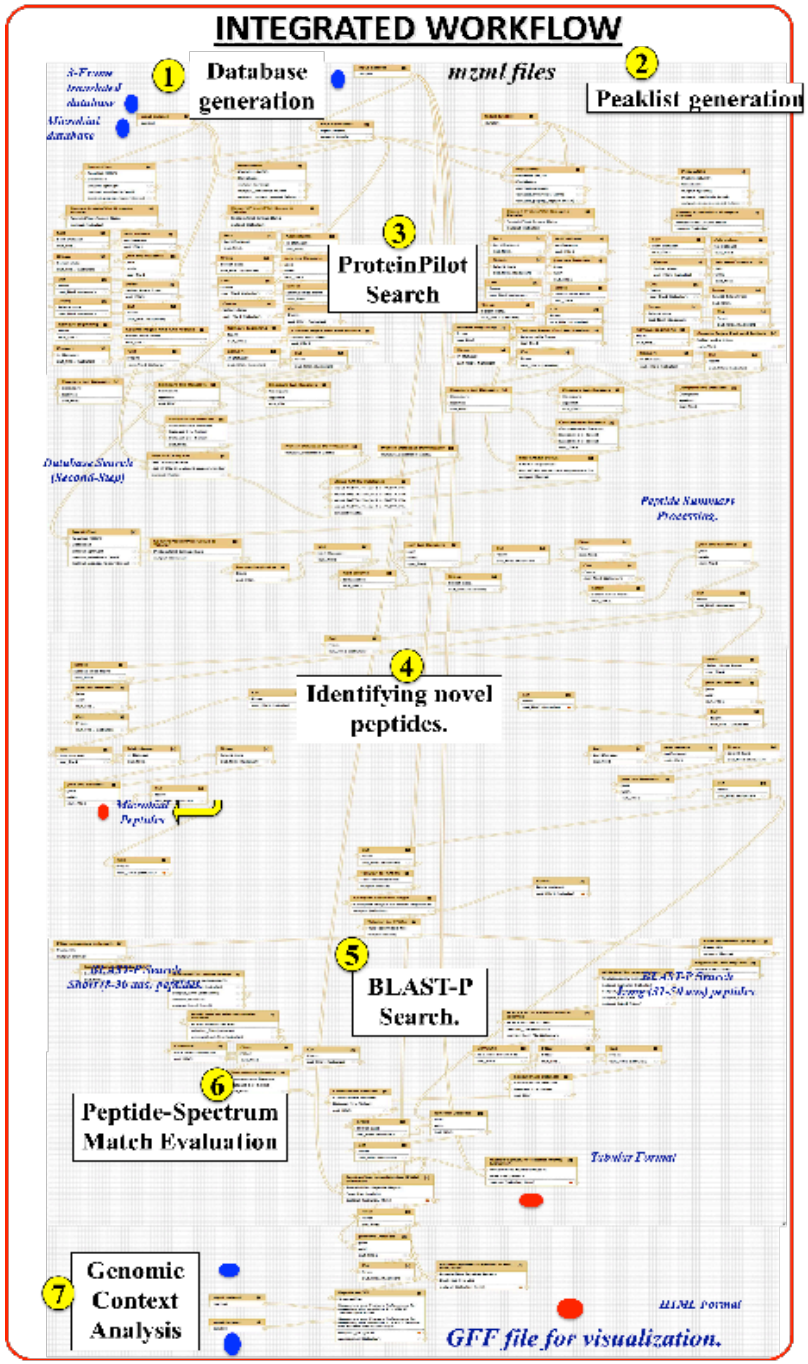
Tophat alignments are used to find evidence of novel splice variant transcripts. The novel splice junctions are translated into a protein database.




**"Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations."**
***Sheynkman G et al BMC Genomics. doi: 10.1186/1471-2164-15-703.***

*Gloria Sheynkman*

*James Johnson*



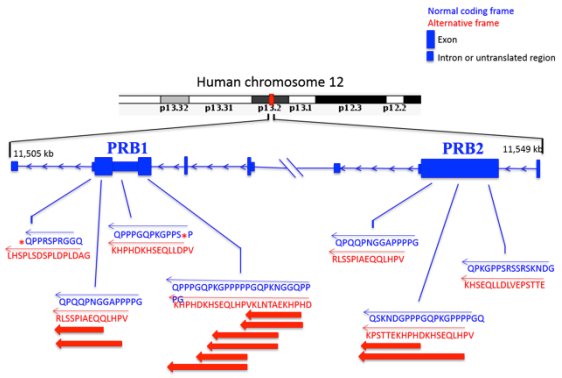

# PROTEOGENOMICS WORKFLOW

**Galaxy-P provides an integrated platform for every step of proteogenomic analysis.**

- **Build target database – download and translate EST databases or perform gene prediction with Augustus.**
- **Numerous tools for identification and text manipulation.**
- **Workflow utilizing BLAST to identify novel peptides.**
- **Tool to assess peptide-spectrum matches and visualize spectra.**
- **Visualize identified peptides on the genome.**
- **140 steps: Seamless, integrated proteogenomic workflow.**

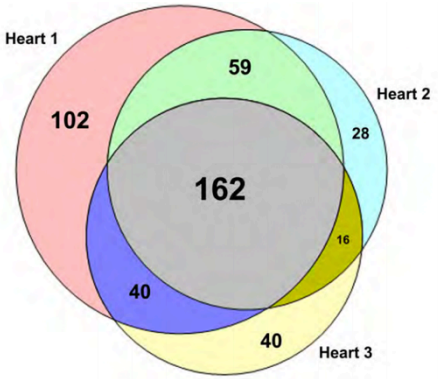
**Flexible and accessible workflows for improved proteogenomic analysis using Galaxy framework.**
**J. Proteome Res. (2014) DOI: 10.1021/pr500812t**
**Link: z.umn.edu/pgfirstlook**

# PROTEOGENOMIC INSIGHTS USING GALAXYP




## SALIVARY PROTEOGENOMICS

- 52 novel proteoforms were identified in a 3D-fractionated salivary dataset.
- Alternate frame translation was identified in PRB1 and PRB2 (12p13) region of human genome.
- PRB proteins are cleaved and secrete peptides and are known to have implications in synovial sarcoma and gastric acid secretion.

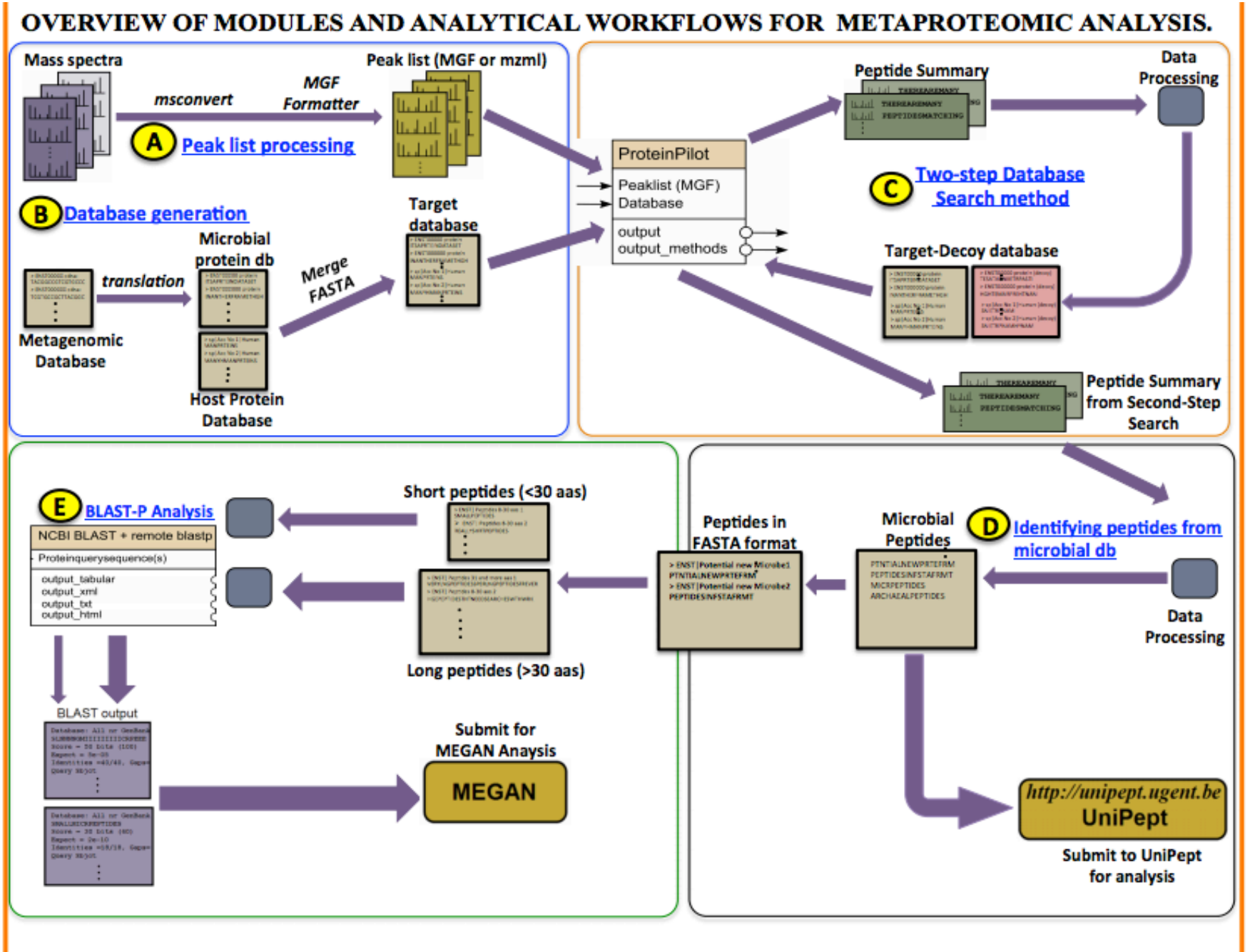



## NON-MODEL ORGANISM PROTEOGENOMICS

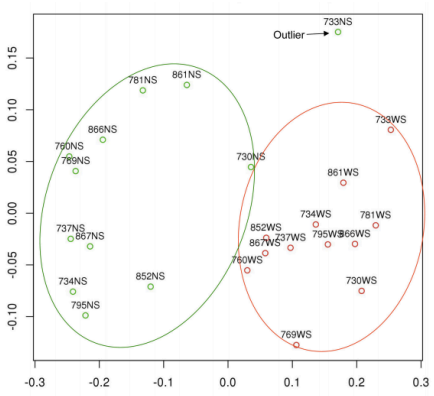

*Katie Vermillion*

- Hibernation proteogenomics in 13-lined ground squirrel.
- Identified multiple novel proteoforms across three replicates.
- Plans for improving on genome annotation; correlation of RNASeq quantitative data with proteomic quantitative data and identification of the role of both known and novel proteoforms in hibernation.

# METAPROTEOMICS: STEPS INVOLVED



# METAPROTEOMICS : BIOLOGICAL INSIGHTS

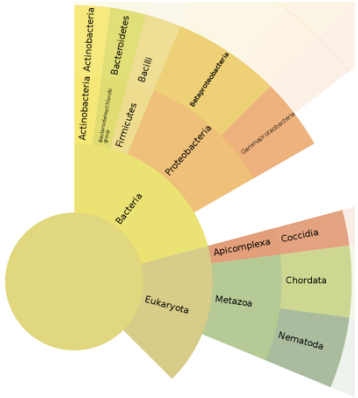



## METAPROTEOMICS OF CHILDHOOD CARIES


*Prof. Joel Rudney*

- ***In vitro*** **investigation of sucrose-induced changes in the metaproteomes of children with caries.**
- **Major shifts in taxonomy and function in paired microcosm oral biofilms grown without and with sucrose respectively.**

**. Twelve replicates currently being analyzed. Targeted proteomics on certain candidates**

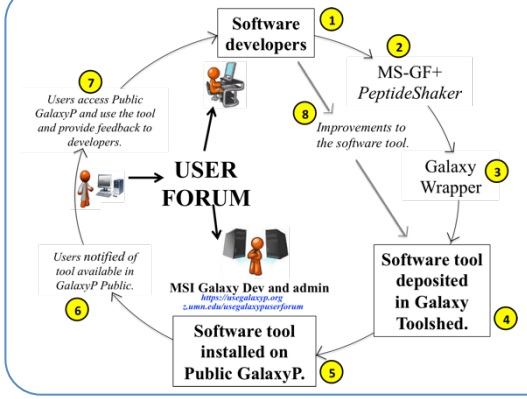



## LUNG CANCER METAPROTEOMICS


*Brian Sandri*

- **Human lung cancer associated dataset subjected to proteogenomic & metaproteomic analysis.**
- **Lung-infection causing species from *Achromobacter*, *Actinomyces*, *Stenotrophomonas* and *Streptococcus* genera were identified.**
- **Data from 16s rRNA will be used to generate databases for further analysis.**

# GALAXYP : ONGOING PROJECTS




## COMMUNITY BASED SOFTWARE DEVELOPMENT

- Community-based software development model is proving effective for implementation, testing and continued improvement of command-line driven software tools.
- We have added tools such as SearchGUI and PeptideShaker in Galaxy, along with opportunities for integration with other software tools such as OpenSWATH via use of workflows.

| | WORKFLOW | INPUT | TOOLS | OUTPUT |
|---|---|---|---|---|
| 1 | Peaklist Generation | RAW File. | msconvert, MGF Formatter | mzml and MGF files |
| 2 | Database Generation | cDNA database, Protein FASTA files. | getORF, get data, merge FASTA | Merged Protein FASTA file |
| 3 | Database Search by Two-Step Method | MGF Files, Search database. | ProteinPilot, Text processing tools | .group file, peptide summary and PSPEP FDR report. |
| 4 | Identifying peptides from translated nucleotide database. | Peptide Summary. | Text processing tools | Peptide List with accession numbers within cDNA database. |
| 5 | BLAST-P Analysis | Peptide List with accession numbers within cDNA database. | BLAST-P and short BLAST-P; Text processing tools | List of peptides that do not match with current human proteome. |
| 6 | Peptide Spectral Match Evaluation | Peptide Summary, mzml files. | PSM Evaluator, Text processing tools | PSM Evaluation metric and HTML Links. |
| 7 | Peptide to GTF conversion | Peptide Summary, cDNA database, GTF file. | Peptides to GTF | GTF file. |

## REPERTOIRE OF WORKFLOWS

- Sharing of analytical workflows that can be reused, shared and creatively modified for multiple studies.
- Multiple workflows for metaproteomics, quantitative proteomics, proteogenomics, RNASeq workflows, are being developed, shared and used.

# CONCLUDING REMARKS

- **Galaxy offers an excellent resource for reproducible workflows that can be shared with users.**

- **We have developed workflows for proteogenomics and metaproteomics analysis that can be creatively modified by users for their projects.**

- **We are also working on improving on our published blueprint workflows for proteogenomics and metaproteomics workflows by adding visualization capabilities, etc.**

- **We are working on adding new tools and workflows for emerging fields in proteomics (such as data independent acquisition / SWATH analysis).**

*Biochemistry, Molecular Biology & Biophysics*

**Tim Griffin**

Candace Guerrero

Kevin Murray

UNIVERSITY OF MINNESOTA
SUPERCOMPUTING INSTITUTE

James Johnson
Tom McGowan
Trevor Wennblom
Getiria Onsongo
Bill Gallip
Ben Lynch

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Gloria Sheynkman
Lloyd Smith
Michael Shortreed

UNIVERSITY OF MINNESOTA
Driven to Discover℠

**Department of Medicine**

Brian Sandri
Kevin Viken
Maneesh Bhargava
Chris Wendt

**Department of Biology (Duluth)**

Matt Andrews
Katie Vermillion
Kyle Anderson

**School of Dentistry**

Joel Rudney

**Laboratory Medicine and Pathology**

Somi Afiuni
Amy Skubitz

**Biochemistry, Molecular Biology and Biophysics**

Laurie Parker
Tzu-Yi Yang

Sarah Parker
Sean Seymour
Sricharan Bandhakavi

## COMMUNITY BASED SOFTWARE DEVELOPMENT

**Ira Cooke**
*La Trobe University, Melbourne , Australia*

**Bjoern Gruening**
*University of Freiburg, Freiburg, Germany*

**Lennart Martens**
*VIB Department of Medical Protein Research, Ugent, Belgium*

**Harald Barsnes and Marc Vaudel**
*University of Bergen, Bergen, Norway*

**John Chilton**
*Galaxy Team*
*Penn State University*

**Center for Mass Spectrometry and Proteomics**

Ebbing de Jong
LeeAnn Higgins
Todd Markowski

**Funding**

**NSF, NIH**