

# 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

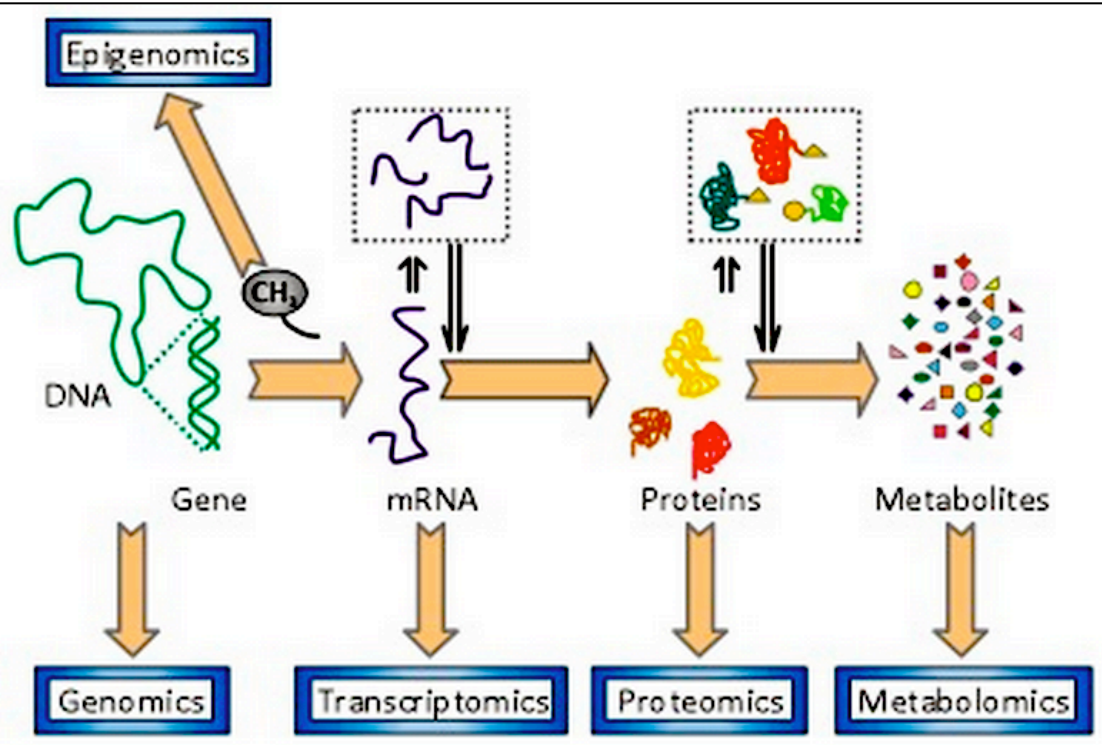


Image Source: Goodacre, J. Exp. Bot 2005.

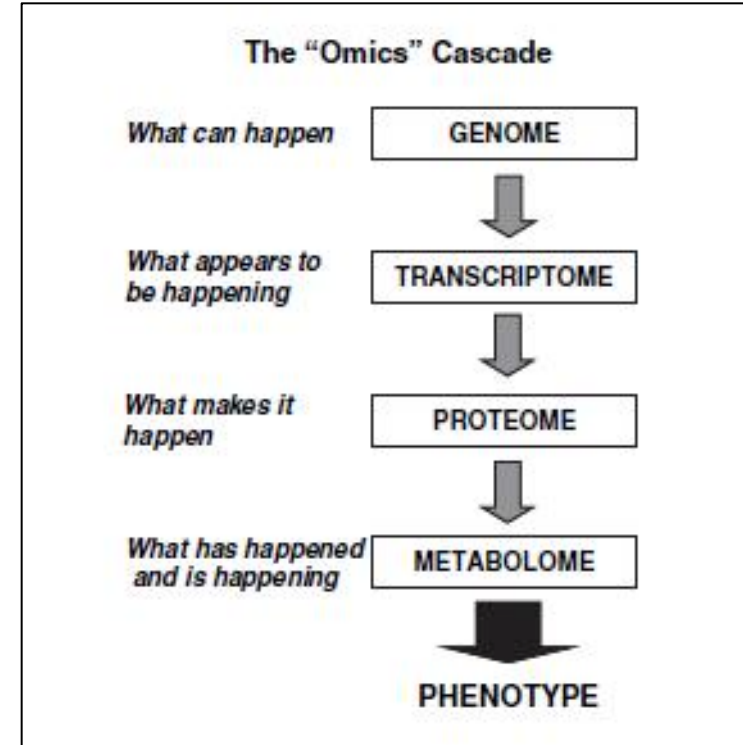
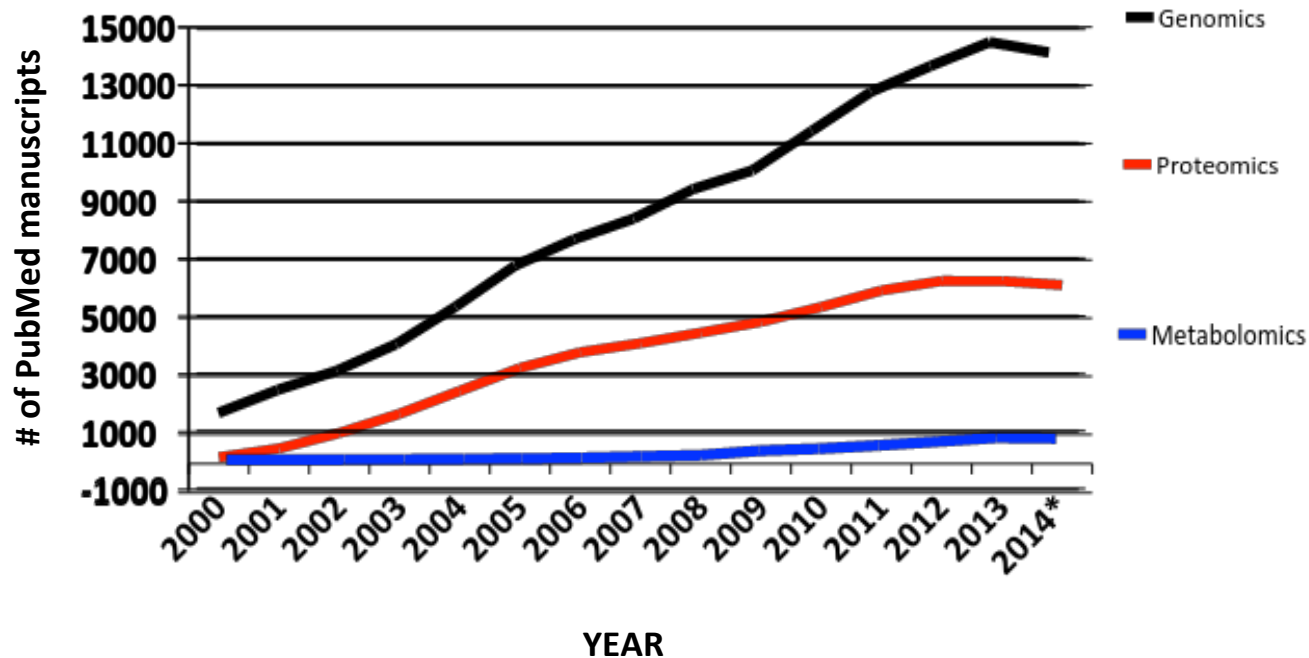


Image Source:  
<http://fluorous.com/images/omics.JPG>

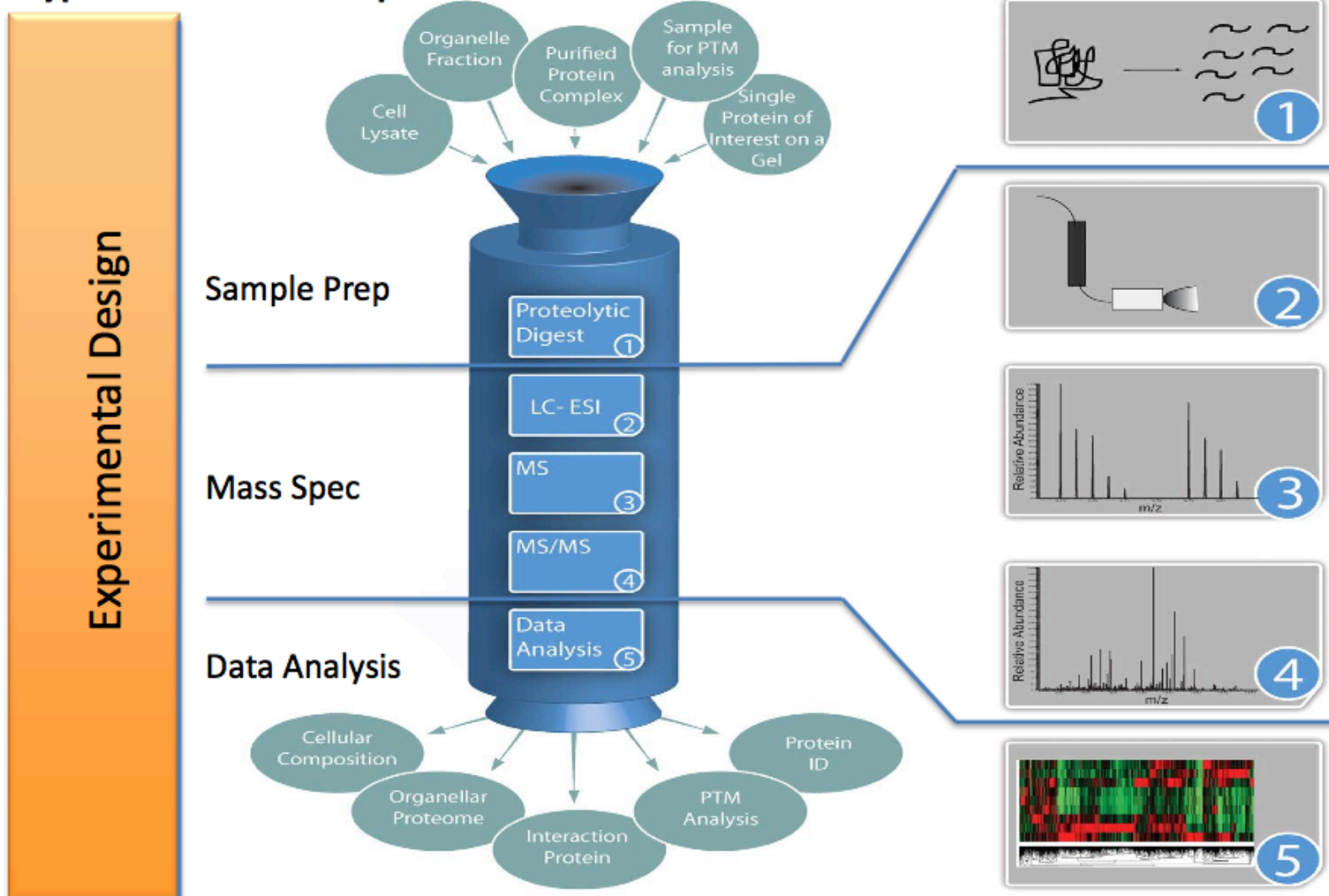
# TRENDS IN OMICS RESEARCH



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

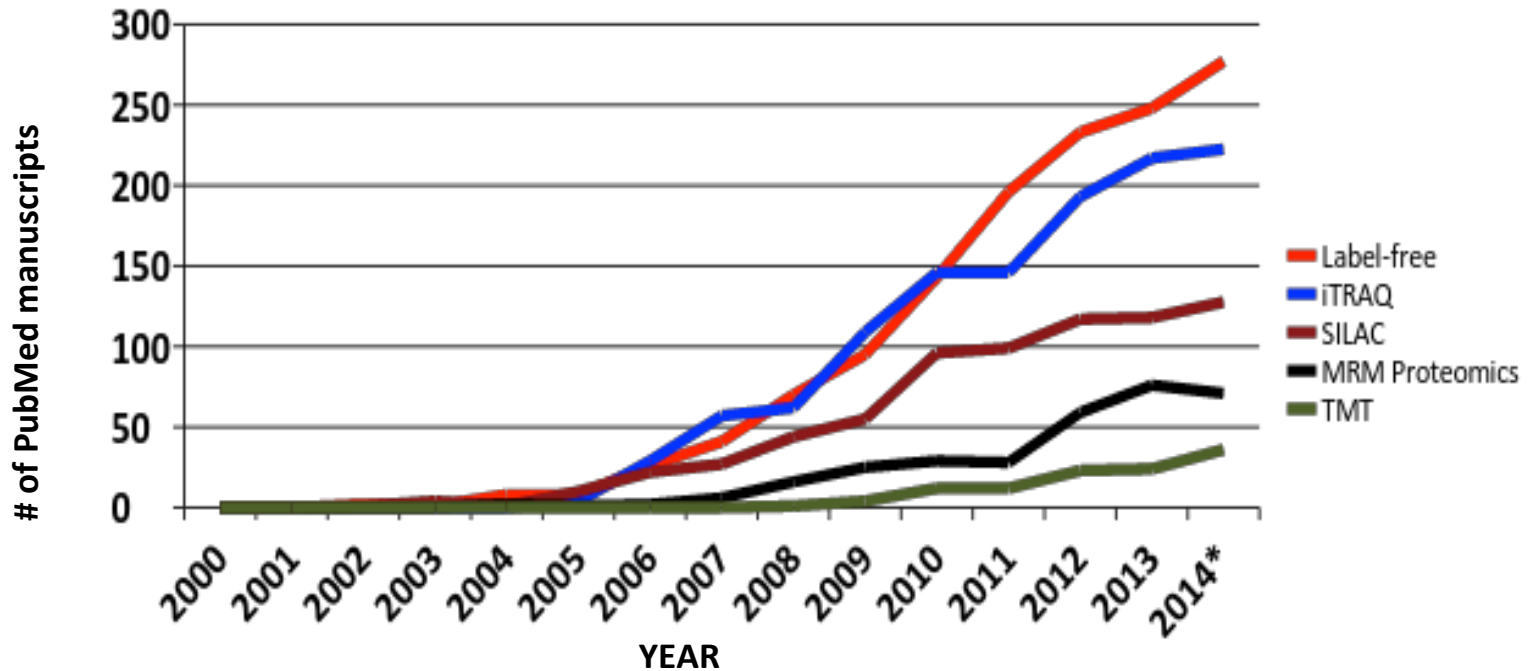
# PROTEOMICS WORKFLOW

## Typical Proteomics Pipeline



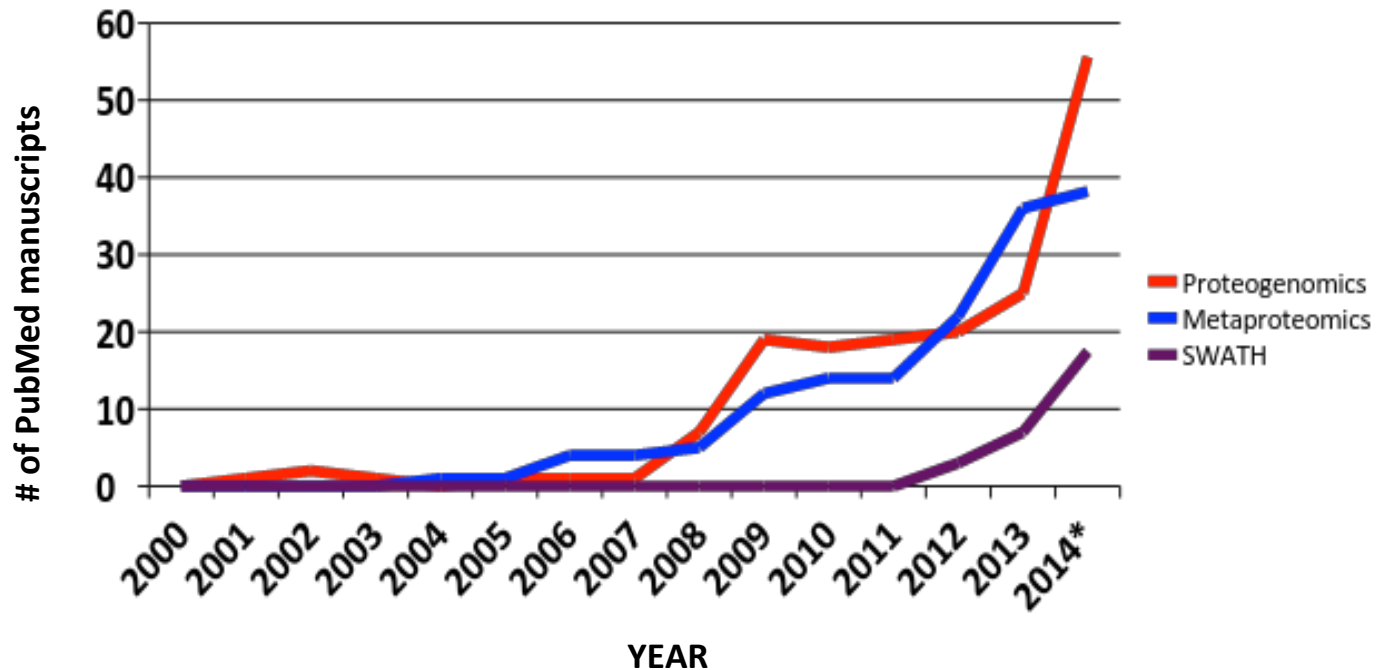
Adapted from Walther T, Mann M. JCB 2010;190:491-500

# QUANTITATIVE PROTEOMICS



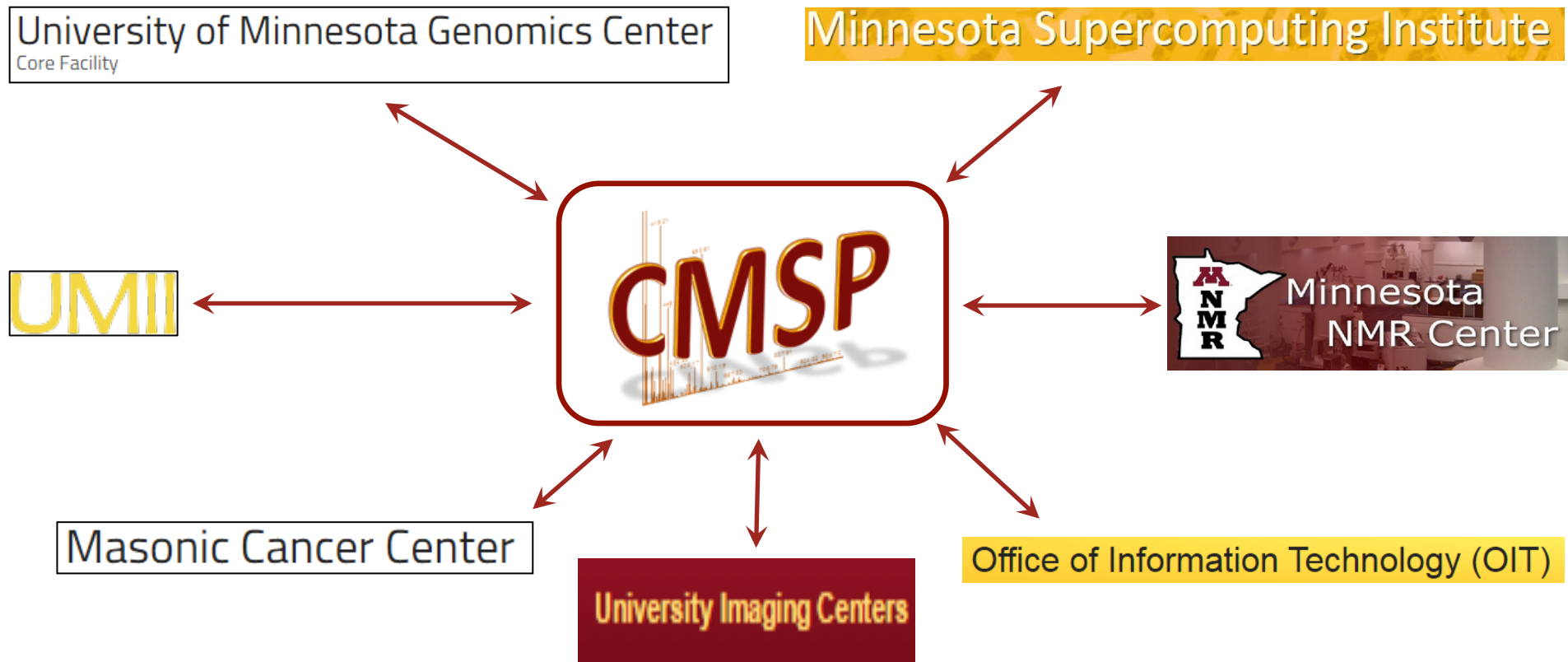
- **Label-free quantitation. (AUC for MS1)**
- **iTRAQ (MS/MS reporter ion)**
- **SILAC (precursor MS ion)**
- **TMT (MS/MS reporter ion)**
- **MRM (Targeted Proteomics)**

# 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



# **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.

**Galaxy / GalaxyP** Analyze Data Workflow Shared Data Visualization Help User Using 1.3 TB

Tools search tools

CORE TOOLS  
Get Data  
Send Data  
Lift-Over  
Text Manipulation  
Filter and Sort  
Join, Subtract and Group  
Convert Formats  
Extract Features  
Statistics  
Graph/Display Data  
FASTA manipulation

PROTEOMICS  
MS Data Conversion  
Sequence Database Tools  
Protein/Peptide Search Algorithms  
Data Conversion Tools  
Visualizers  
Quantification  
BLAST-P  
Proteogenomics

GENOMICS  
Fetch Sequences  
Fetch Alignments  
NGS: Mapping  
NGS: RNA Analysis  
NGS: SAM Tools  
NGS: Variant

EMBOSS  
Blast  
Picard

MISC  
Misc

Workflows  
795 NS WS For Rudney datasets : Workflow for paired metaproteomics comparison studies.  
MP1: Workflow for paired metaproteomics comparison studies - HOMD db search, (imported from uploaded file)  
All workflows

**Welcome to GalaxyP**

GalaxyP is a multiple 'omics' data analysis platform with particular emphasis on mass spectrometry based proteomics. GalaxyP is developed at the University of Minnesota, deployed at the Minnesota Supercomputing Institute, and is an extension of the popular Galaxy project. The GalaxyP project is supported by a grant from [NSF](#).

This public Galaxy instance is meant for testing with small-scale data sets, and sharing workflows and tools. Users with larger data analysis needs are encouraged to [install a local instance](#) of Galaxy and access GalaxyP tools via the [Tool Shed](#).

**Updates**

February 13, 2015  
All Galaxy Tools running nominally

February 12, 2015  
All Galaxy Tools running nominally

**Tweets** Follow

**The GalaxyP Project** @usegalaxy 6 Feb  
Multi-omic data analysis using #usegalaxy z.umn.edu/multiomicsnbt #proteomics #metabolomics #interactomics #proteogenomics #metaproteomics  
Expand

**John Chilton** @jmchilton 30 Oct  
Pair of good #proteogenomics articles in Nature Methods including excellent shoutout to GalaxyP in one. nature.com/nmeth/journal/... #usegalaxy Retweeted by The GalaxyP Project  
Tweet to @usegalaxy

**Links**

- [Large-scale multi-omic data integration and analysis: challenges and opportunities](#)
- [Metaproteomics: an opportunity-rich complement to metagenomics](#)
- [The Galaxy framework as a unifying bioinformatics solution for 'omics' core facilities](#)

**History** search datasets

NS Peptide Summary to Input for MEGANS analysis. 32 shown 519.7 MB

32: MEGANS Output for NS

31: blastp on db nr current

30: blastp-short on db nr current

29: Filter sequences by length on data 27  
105 sequences  
format: fasta, database: ?

28: Filter sequences by length on data 27  
4,241 sequences  
format: fasta, database: ?

27: Regex Find And Replace on data 26

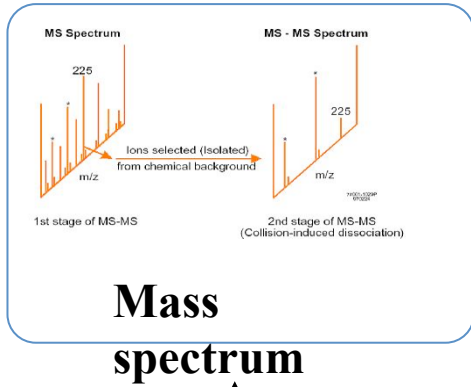
26: Tabular-to-FASTA on data 25

25: Cut on data 24

24: Merge Columns on data 23

23: Add column on data 22

# PROTEOGENOMICS AND METAPROTEOMICS



Search Protein for [ ] Go

Limits Preview/Index

Display Graphics Show: 20 Send to File Get Subseq

[1]: P03019. Fumarate and nitr... [gi:120458]

```
>gi|120458|sp|P03019|FNR_ECOLI Fumarate and nitrate reduction regulatory protein
MIPKFRIRRIQSGGCAIHCQDCISLQICPFLNEMELDQLDNIERRKPKQK@OTLFRAGDELKSLYA
IRSGTIKSYTITQGEQITQFPLAGDLVGFDAI65GHPGFPAOLETSMVCEIFPFDL65GHPFLR
QQMRLMSGEIKGDQDMILLISKNRAEERLARTVNLSEFPAQGFSPREFLTMTRGDIQWYGLTVET
ISELLGPFQKSGHLAVKGYITTIENNDALAGLGHTEHVA
```

**Reference Protein Database  
from genomic annotation**

Search Protein for [ ] Go

Limits Preview/Index

Display Graphics Show: 20 Send to File Get Subseq

[1]: P03019. Fumarate and nitr... [gi:120458]

```
>gi|120458|sp|P03019|FNR_ECOLI Fumarate and nitrate reduction regulatory protein
MIPKFRIRRIQSGGCAIHCQDCISLQICPFLNEMELDQLDNIERRKPKQK@OTLFRAGDELKSLYA
IRSGTIKSYTITQGEQITQFPLAGDLVGFDAI65GHPGFPAOLETSMVCEIFPFDL65GHPFLR
QQMRLMSGEIKGDQDMILLISKNRAEERLARTVNLSEFPAQGFSPREFLTMTRGDIQWYGLTVET
ISELLGPFQKSGHLAVKGYITTIENNDALAGLGHTEHVA
```

**Metagenomic sequences**

Search Protein for [ ] Go

Limits Preview/Index

Display Graphics Show: 20 Send to File Get Subseq

[1]: P03019. Fumarate and nitr... [gi:120458]

```
>gi|120458|sp|P03019|FNR_ECOLI Fumarate and nitrate reduction regulatory protein
MIPKFRIRRIQSGGCAIHCQDCISLQICPFLNEMELDQLDNIERRKPKQK@OTLFRAGDELKSLYA
IRSGTIKSYTITQGEQITQFPLAGDLVGFDAI65GHPGFPAOLETSMVCEIFPFDL65GHPFLR
QQMRLMSGEIKGDQDMILLISKNRAEERLARTVNLSEFPAQGFSPREFLTMTRGDIQWYGLTVET
ISELLGPFQKSGHLAVKGYITTIENNDALAGLGHTEHVA
```

**Genome six-frame  
translation**

**RNASeq  
data**

**cDNA  
three-  
frame  
translation**

# PROTEOGENOMICS: STEPS INVOLVED

~ 2 million proteins

Database Generation\*

Peaklist generation

~ 10,000 proteins

Database search

~ 5,000 proteins

First-step

Two-step

Identifying peptides from translated nucleotide db

~ 1,000 peptides

Automated BLAST-P search\*

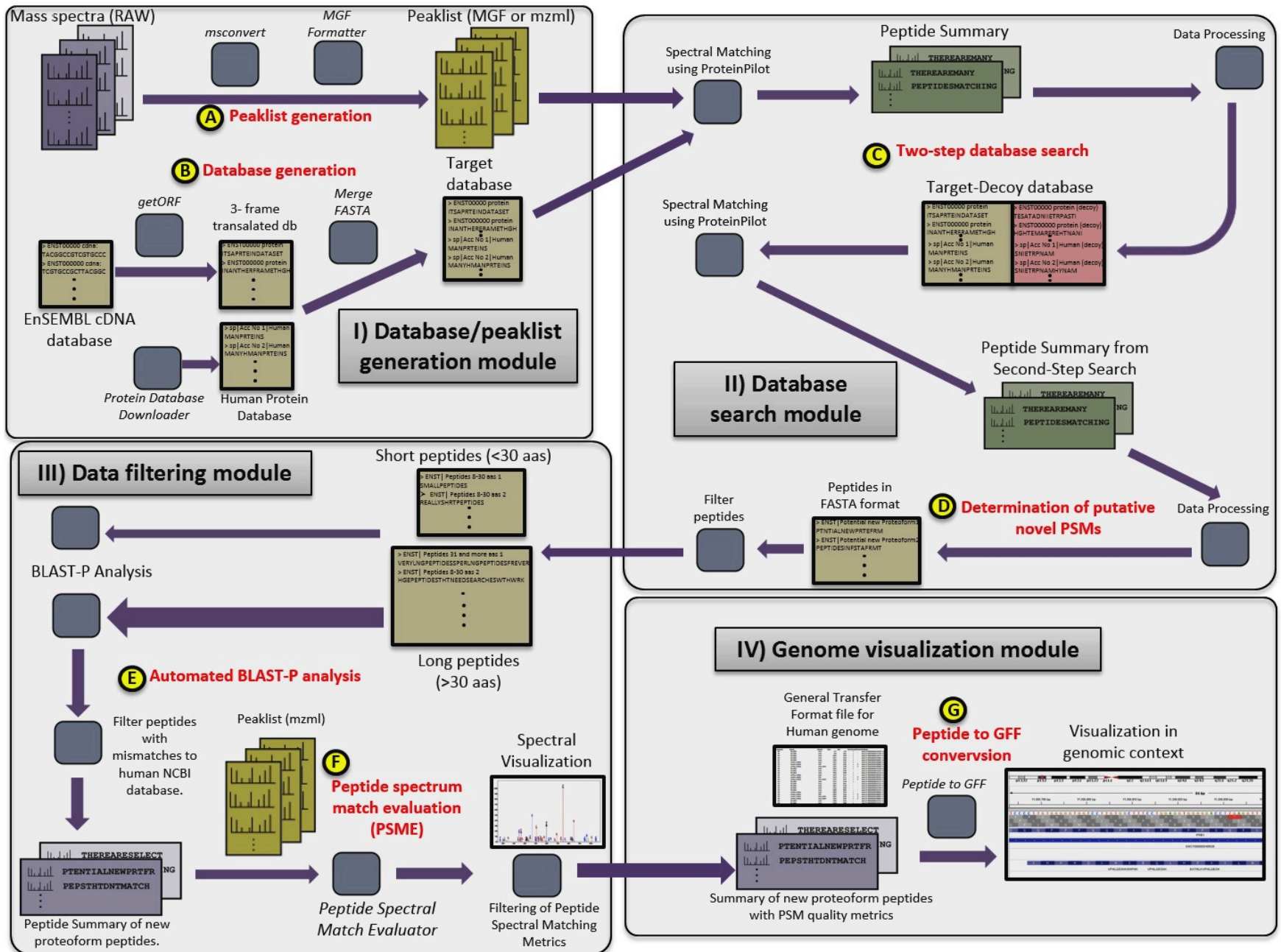
~ 100 peptides

Peptide-Spectral-Match Evaluation

~ 50 peptides

Genomic context analysis\*

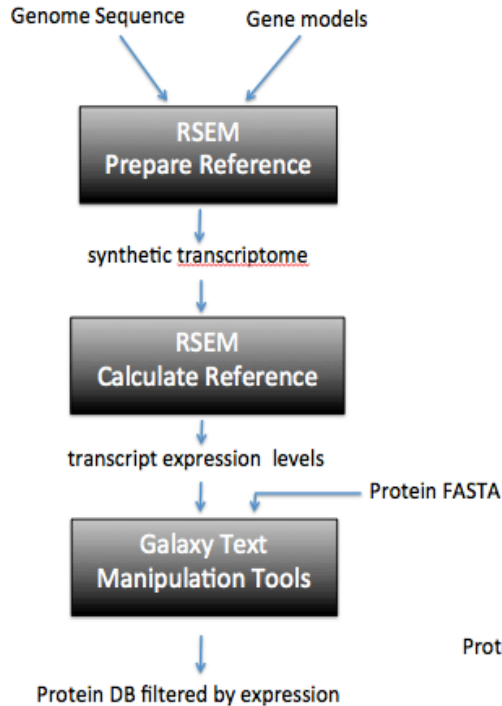
# PROTEOGENOMICS WORKFLOW



# RNASeq DERIVED PROTEOMIC DATABASES

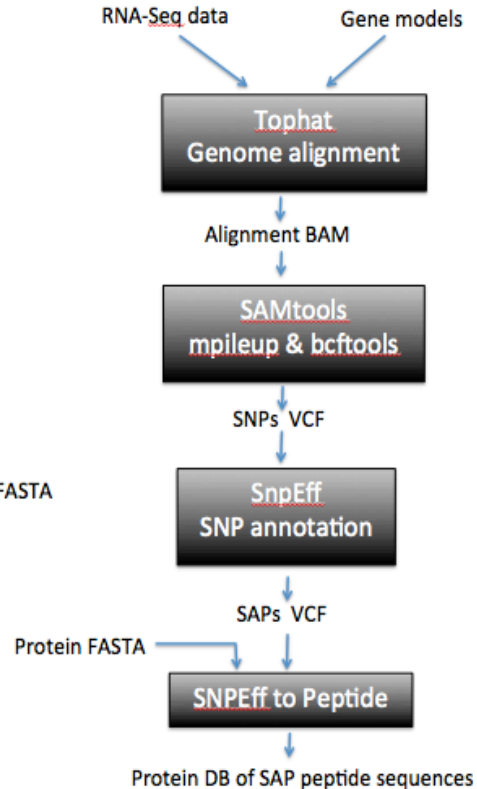
## Reduced Database

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



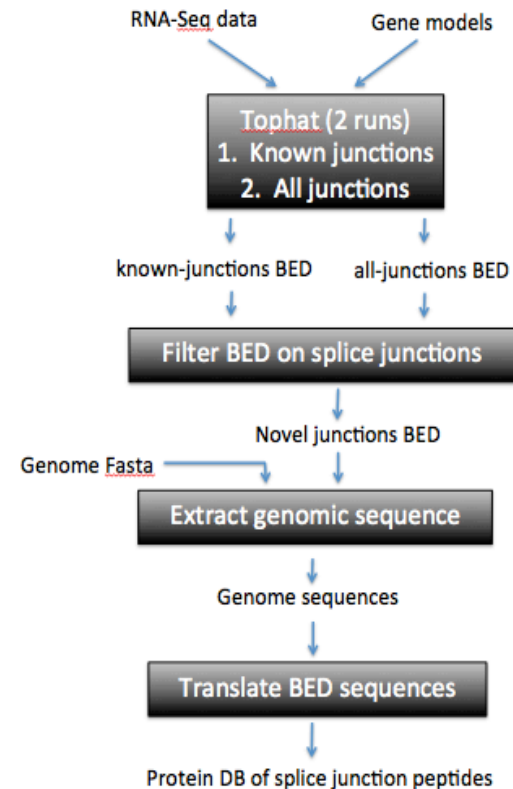
## SAP Database

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

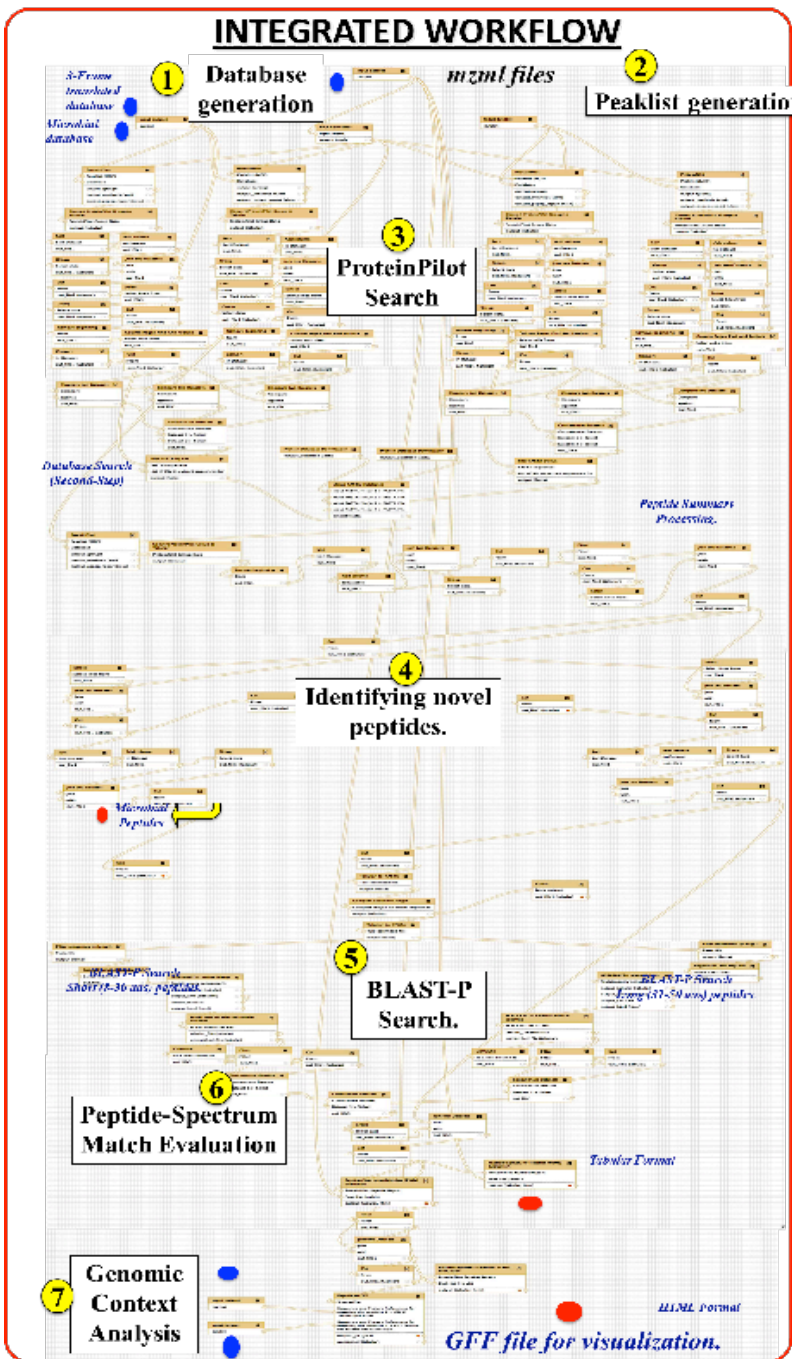


James Johnson

# PROTEOGENOMICS WORKFLOW

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

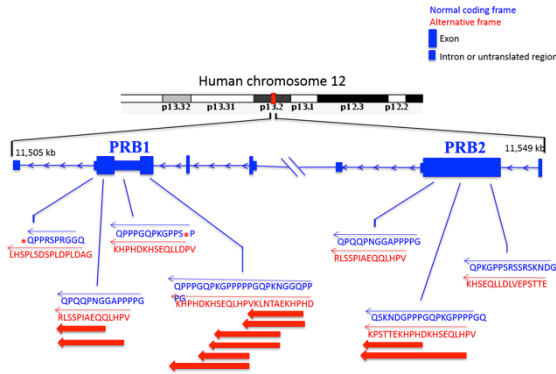
- Build target database – download and translate EST databases or perform gene prediction with Augustus.
- Numerous tools for identification and text manipulation.
- Workflow utilizing BLAST to identify novel peptides.
- Tool to assess peptide-spectrum matches and visualize spectra.
- Visualize identified 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](http://z.umn.edu/pgfirstlook)

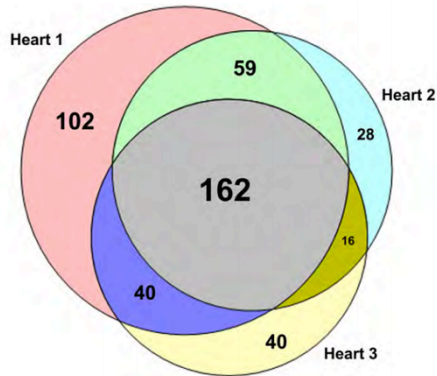
# PROTEOGENOMIC INSIGHTS USING GALAXYP

## 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



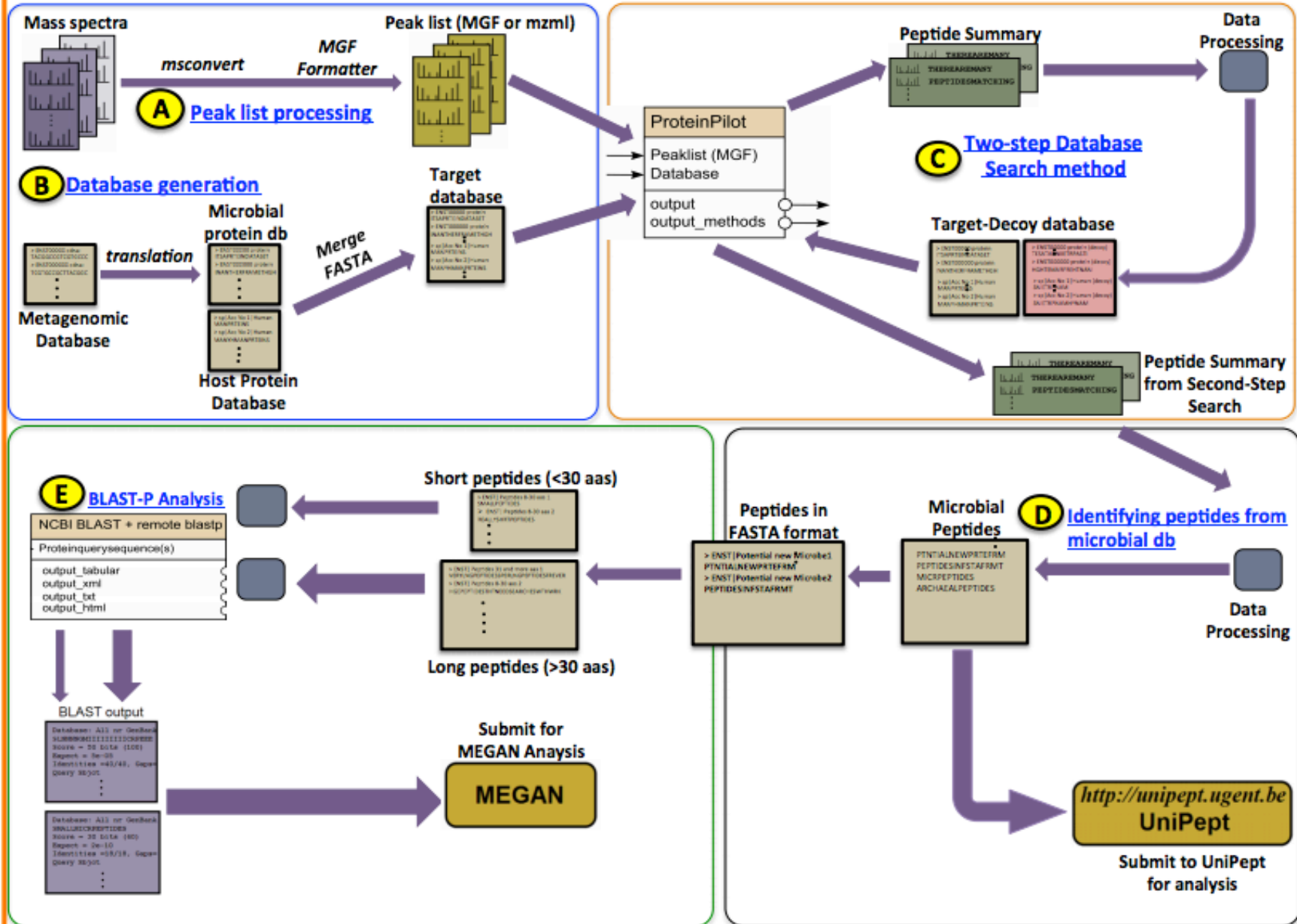
- Hibernation proteogenomics in 13-lined ground squirrel.
- Identified multiple novel proteoforms across three replicates.
- Plans for improving on genome annotation; correlation of RNASeq quantitative data with proteomic quantitative data and identification of the role of both known and novel proteoforms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



Katie Vermillion

# METAPROTEOMICS: STEPS INVOLVED

## OVERVIEW OF MODULES AND ANALYTICAL WORKFLOWS FOR METAPROTEOMIC ANALYSIS.





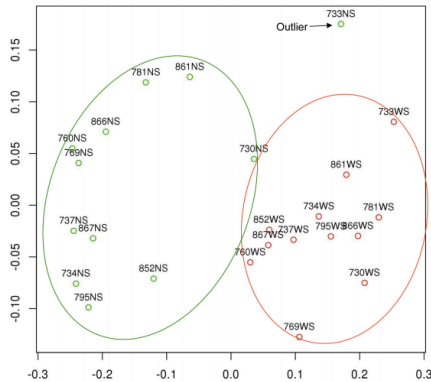
# METAPROTEOMICS : BIOLOGICAL INSIGHTS

## METAPROTEOMICS OF CHILDHOOD CARIES

- *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



Prof. Joel Rudney

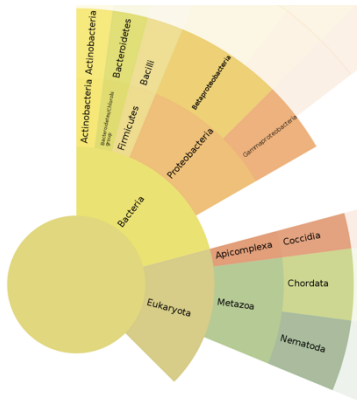


## LUNG CANCER METAPROTEOMICS

- 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.

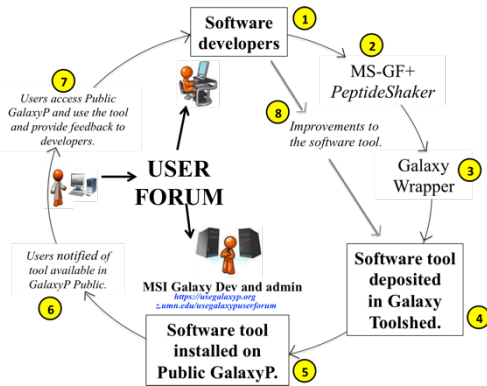


Brian Sandri



# 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.**

## REPertoire OF WORKFLOWS

WORKFLOW	INPUT	TOOLS	OUTPUT
1 <a href="#">Peaklist Generation</a>	RAW File.	msconvert, MGF Formatter	mzml and MGF files
2 <a href="#">Database Generation</a>	cDNA database, Protein FASTA files.	getORF, get data, merge FASTA	Merged Protein FASTA file
3 <a href="#">Database Search by Two-Step Method</a>	MGF Files, Search database.	ProteinPilot, Text processing tools	.group file, peptide summary and PSPEP FDR report.
4 <a href="#">Identifying peptides from translated nucleotide database.</a>	Peptide Summary.	Text processing tools	Peptide List with accession numbers within cDNA database.
5 <a href="#">BLAST-P Analysis</a>	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 <a href="#">Peptide Spectral Match Evaluation</a>	Peptide Summary, mzml files.	PSM Evaluator, Text processing tools	PSM Evaluation metric and HTML Links.
7 <a href="#">Peptide to GTF conversion</a>	Peptide Summary, cDNA database, GTF file.	Peptides to GTF	GTF file.

- **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**

