THE GALAXY FRAMEWORK AS A UNIFYING BIOINFORMATICS SOLUTION FOR MULTI-OMIC DATA ANALYSIS

PRATIK JAGTAP
University of Minnesota
Twin-cities, MN

MULTI-OMICS

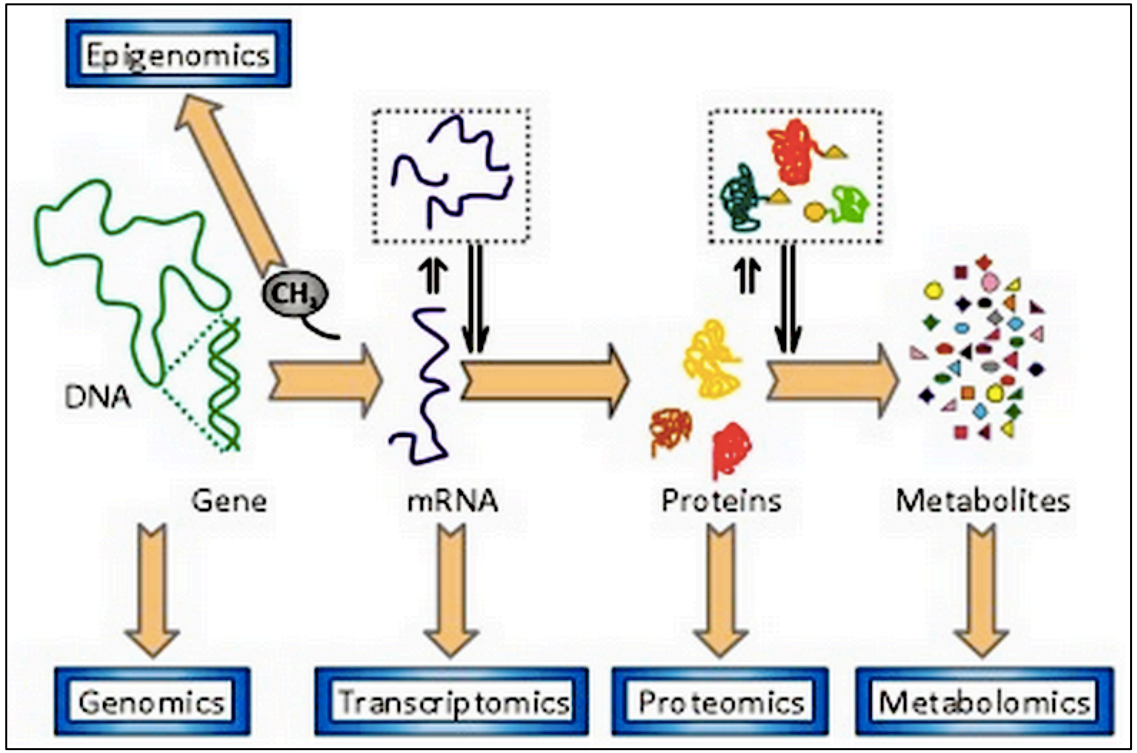


Image Source: Goodacre, J. Exp. Bot 2005.

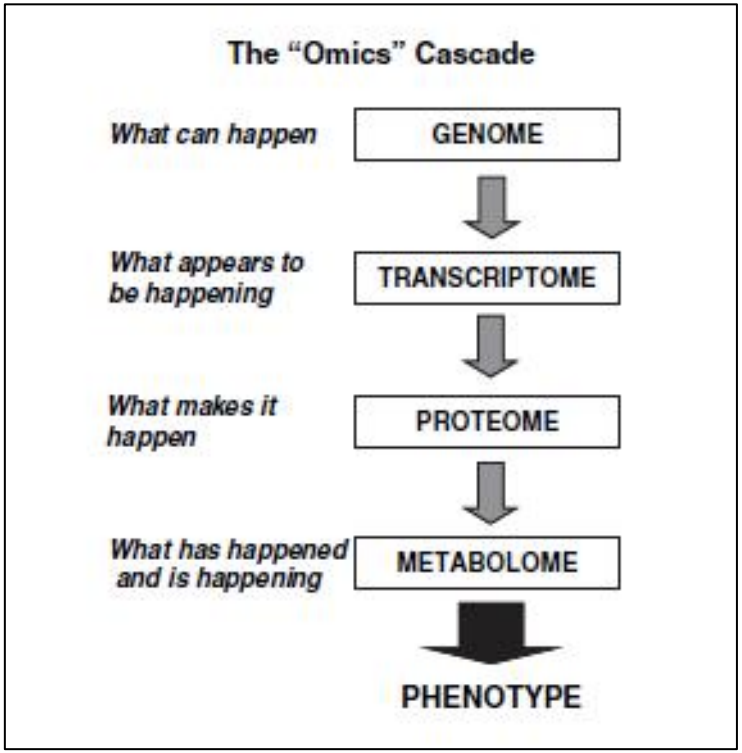
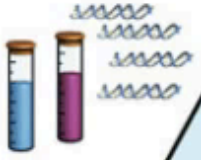


Image Source: http://fluorous.com/images/omics.JPG

GENOME



Sample library preparation

Next gen sequencing
Illumina, Complete Genomics, PacBio, etc.

Raw data
size ~TB, e.g., .fastq files

Quality control
FastQC, trimming, Picard

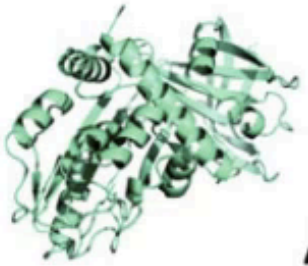
Mapping of reads
ELAND II, Bowtie, SOAP, Novoalign MAQ, BWA, GATK

SNV identification and effect
GATK, HugeSeq, SAMTools, SIFT, Polyphen, Annovar

SVs identified
(Paired-end, read depth, split-read algorithms)
BreakSeq, DELLY, Pindel

PROTEINS

Quantitative Mass Spectrometry Sample Workflow



Quantitation sample preparation
SILAC, iTRAQ, TMT, label-free

Liquid chromatography - mass spectrometry profiling (LC-MS/MS)

Convert to .mzML for open source
MSConvert, TPP

MS/MS spectra identification
X!Tandem, SEQUEST, OMMSA, Mascot

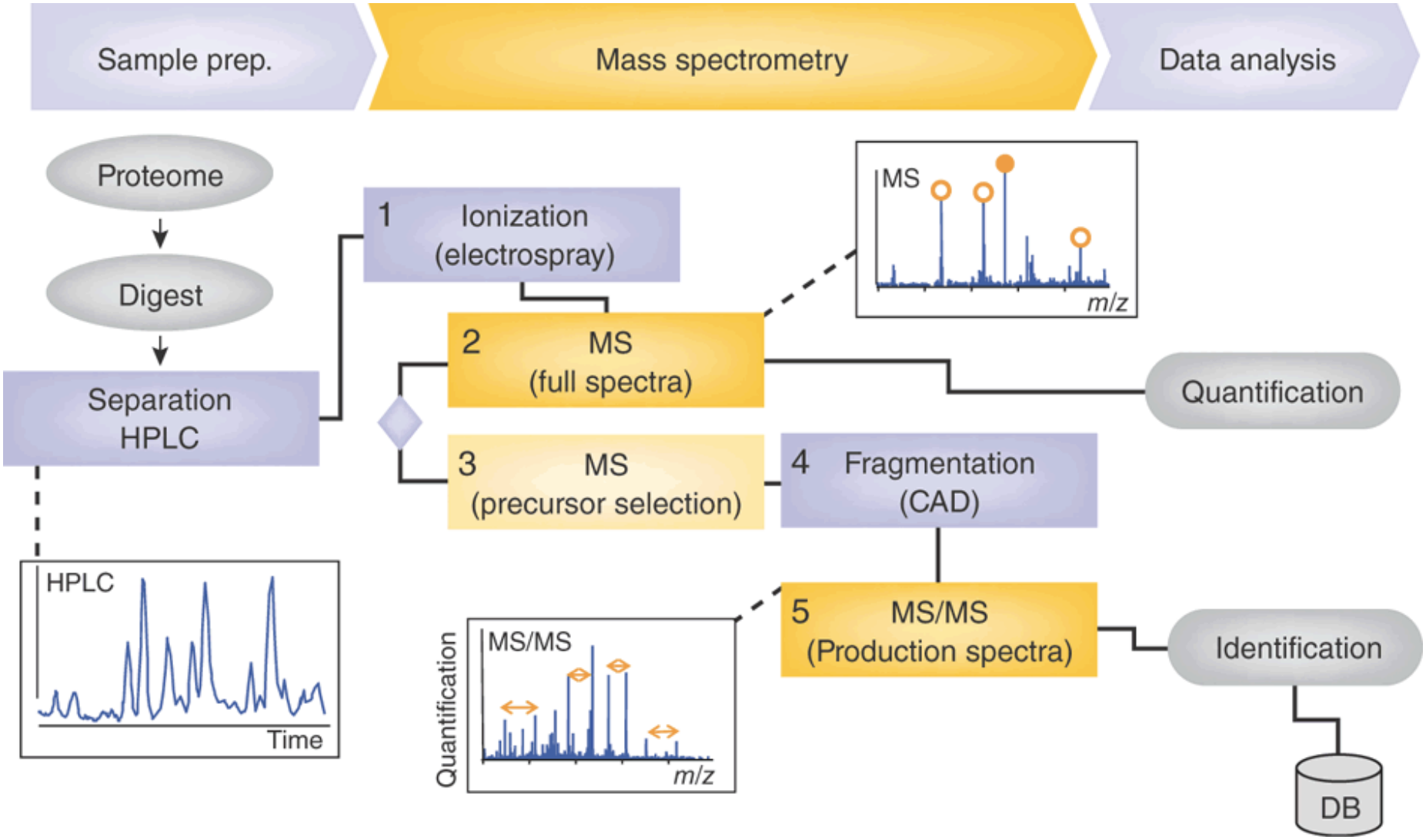
Quality control, false discovery rate estimations

Normalization
Ratios ($\mu = 1$); distinct runs

Data annotation/ consolidation
Uniprot/NCBI annotation

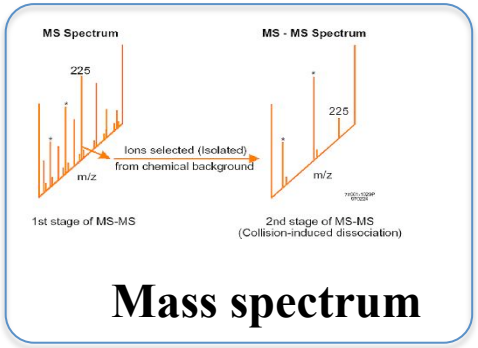
goo.gl/iamZth

PROTEOMICS WORKFLOW

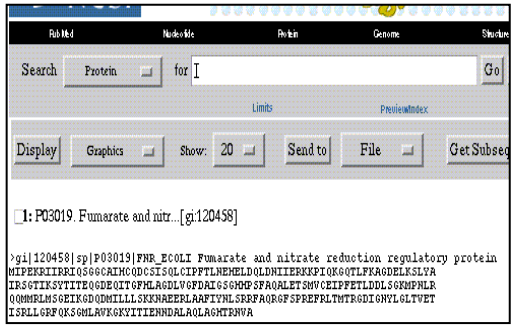


Bruno Domon & Ruedi Aebersold
Nature Biotechnology 28, 710–721 (2010)

DEFINING PROTEOGENOMICS & METAPROTEOMICS : LOOKING WITHIN AND WITHOUT

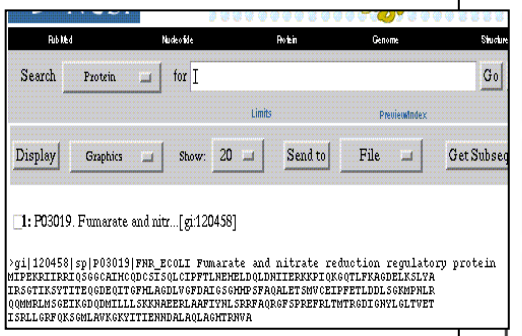


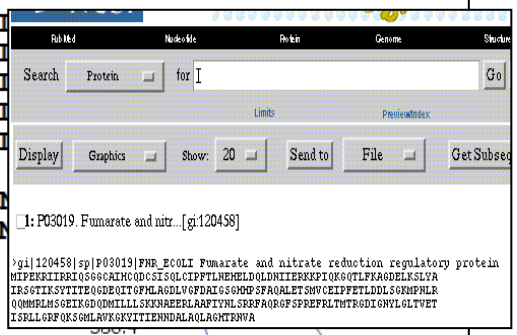
Mass spectrum



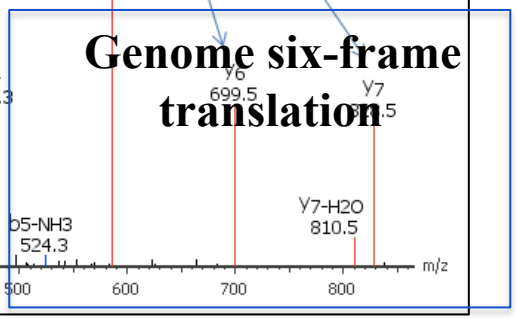
Reference Protein Database from genomic annotation

Peptide Spectral Match





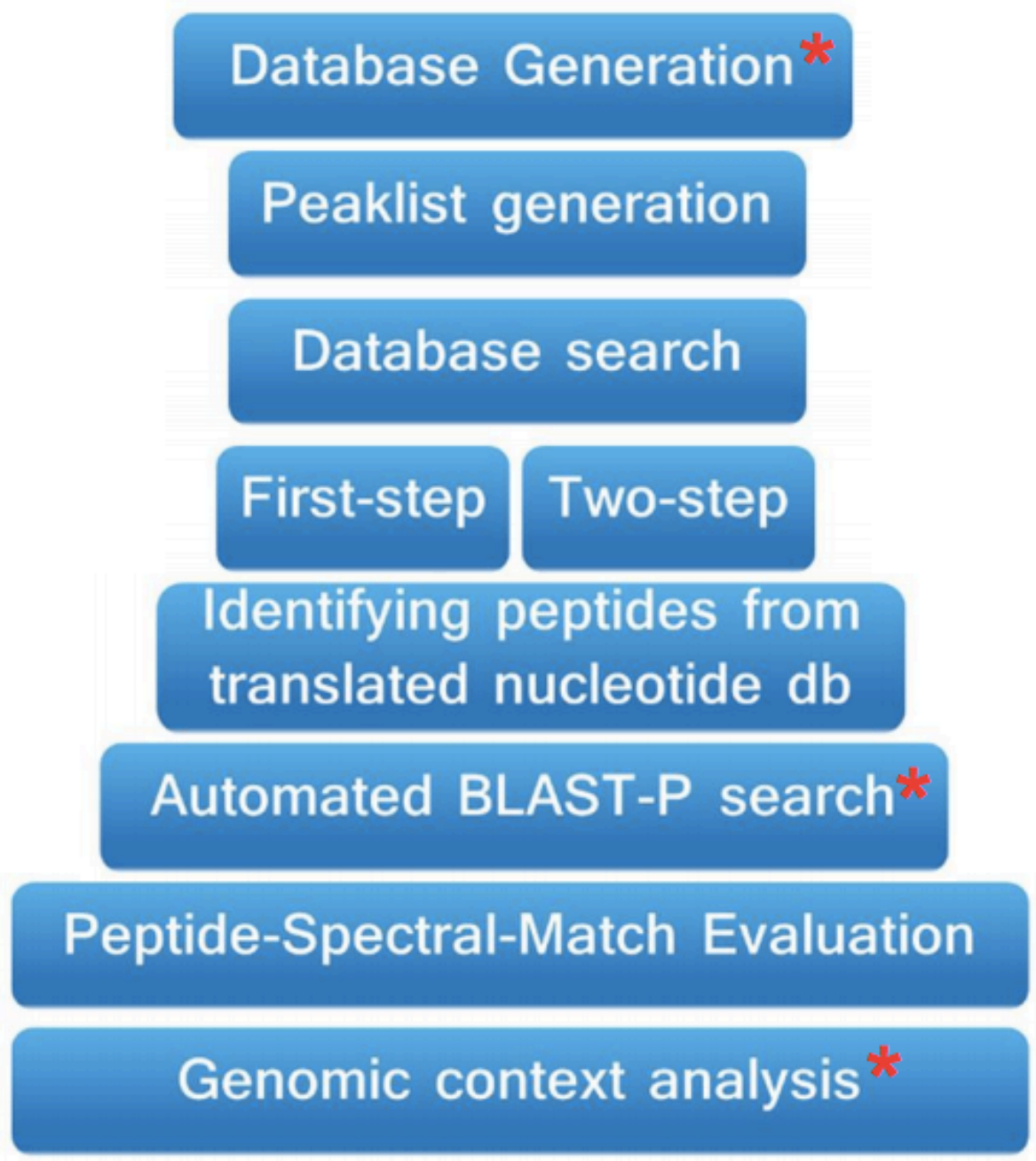
RNASeq data



cDNA three-frame translation

Metagenomic sequences

DEFINING PROTEOGENOMICS: STEPS INVOLVED



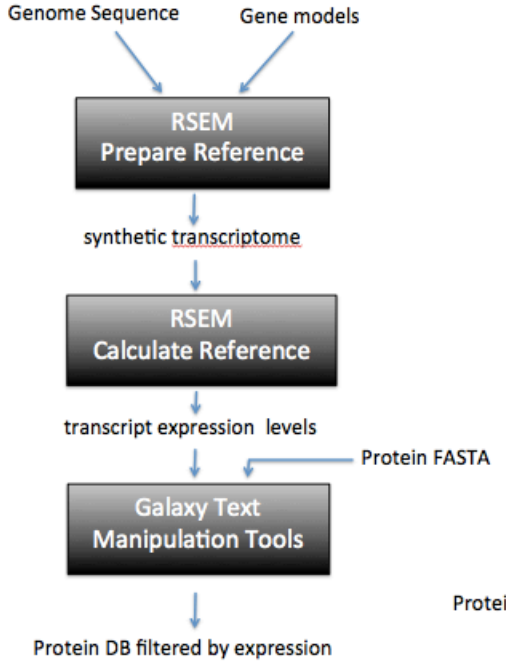
RNASeq DERIVED PROTEOMIC DATABASES


Gloria Sheynkman


James Johnson

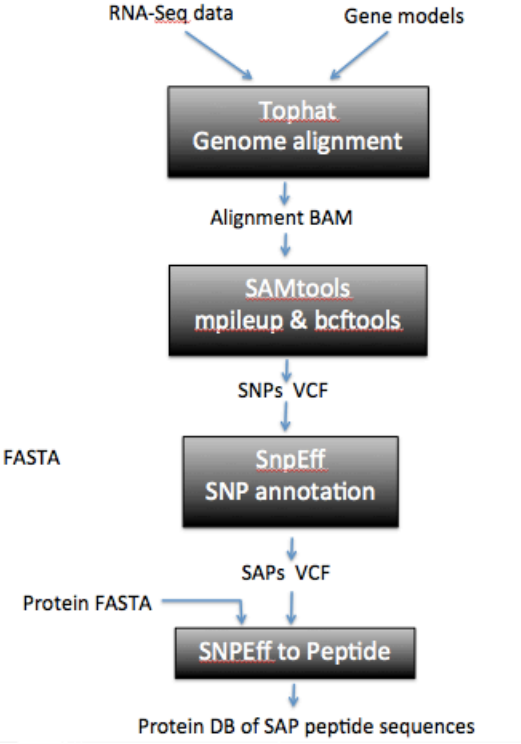
Reduced Database

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



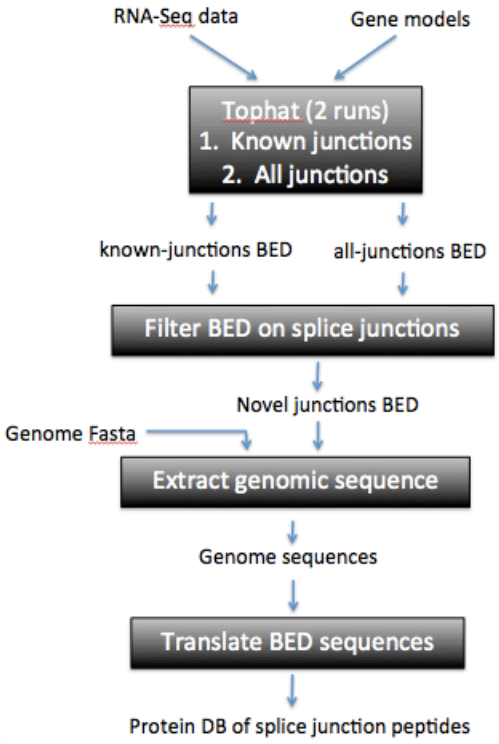
SAP Database

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.

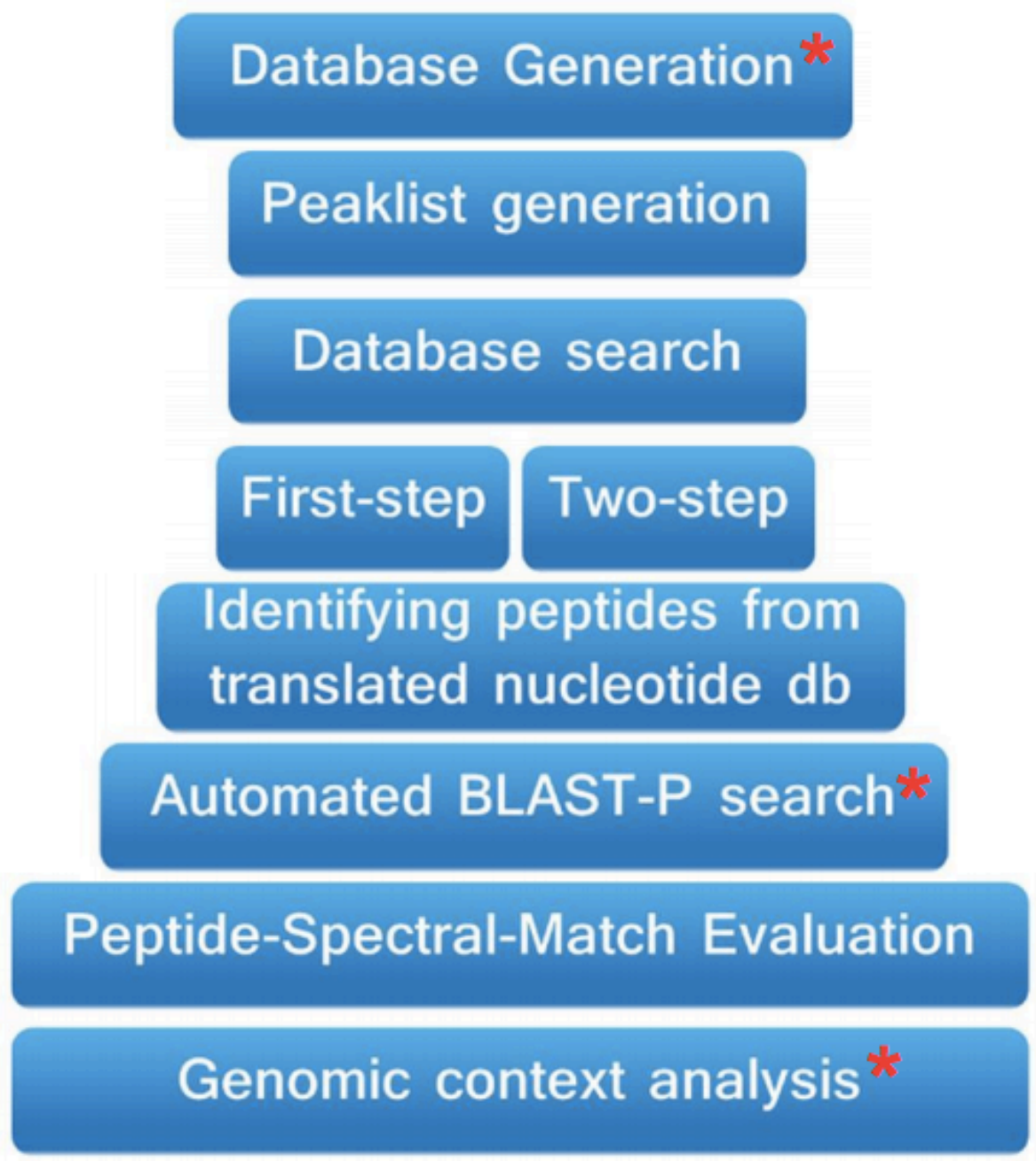


Splice Database

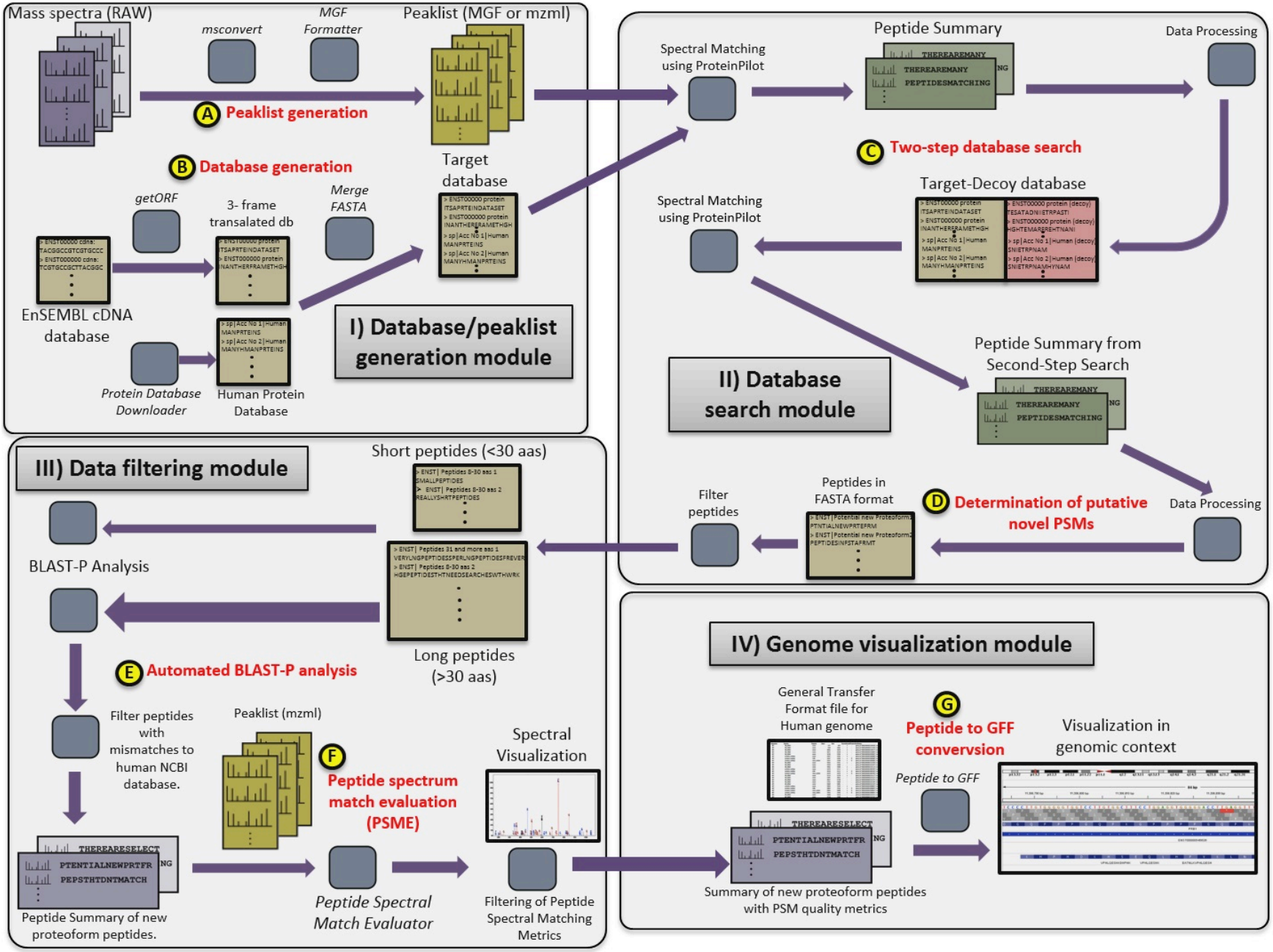
Tophat alignments are used to find evidence of novel splice variant transcripts. The novel splice junctions are translated into a protein database.



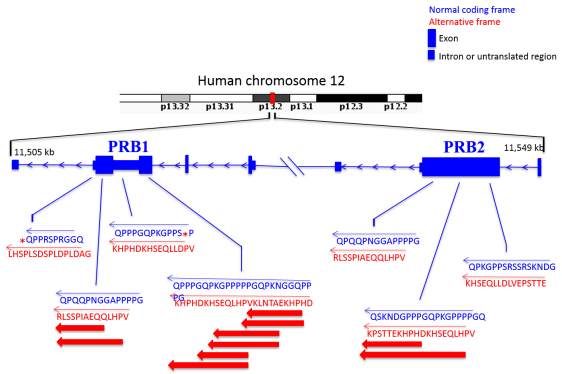
DEFINING PROTEOGENOMICS: STEPS INVOLVED



PROTEOGENOMICS WORKFLOW

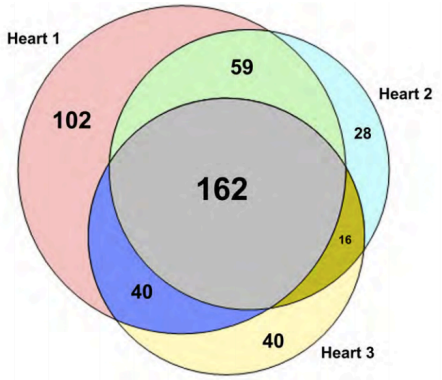


PROTEOGENOMICS : BIOLOGICAL INSIGHTS



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.

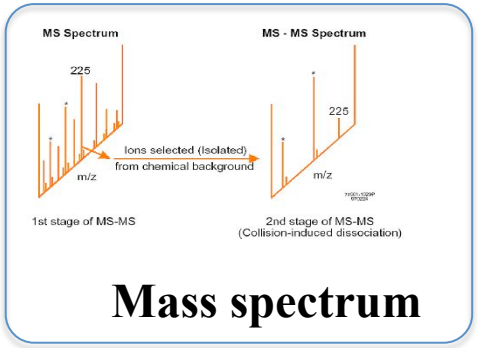


NON-MODEL ORGANISM PROTEOGENOMICS

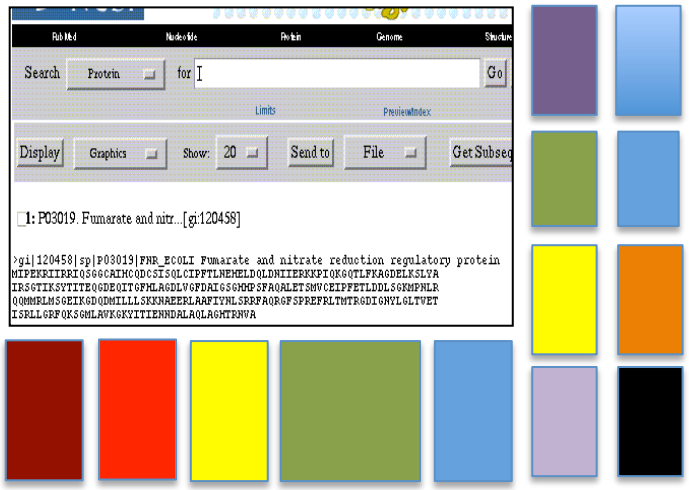
Katie Vermillion

- 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.

DEFINING METAPROTEOMICS



Mass spectrum



Metagenomic sequences

Database Generation*

Peaklist generation

Database search

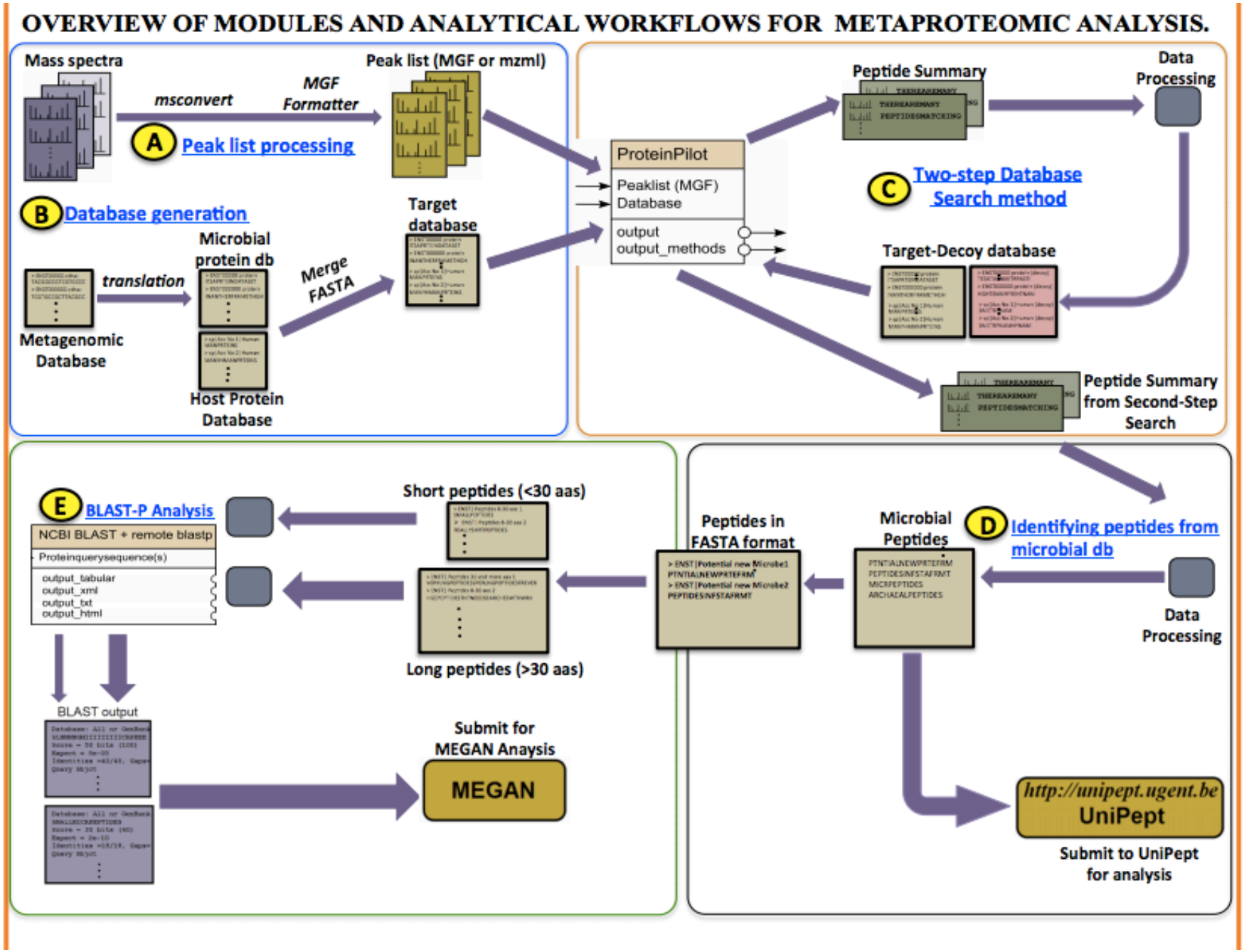
First-step | Two-step

Identifying peptides from translated nucleotide db

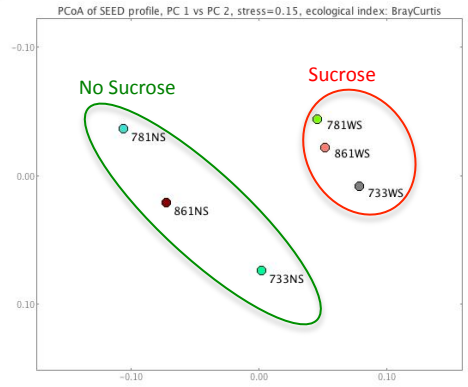
microbial

Automated BLAST-P search*

DEFINING METAPROTEOMICS: STEPS INVOLVED



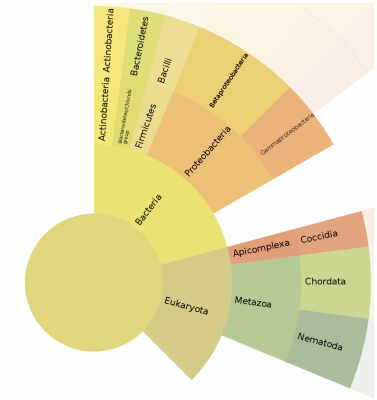
METAPROTEOMICS : BIOLOGICAL INSIGHTS



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.**
- **Six replicates currently being analyzed.**

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.**

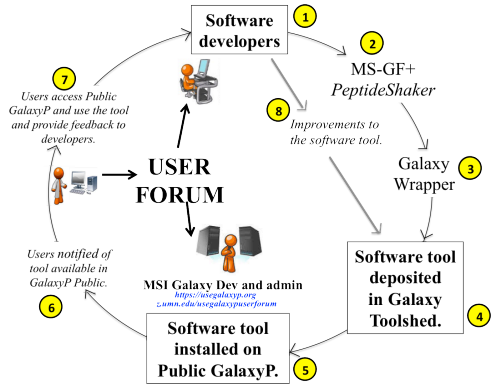
Brian Sandri

GALAXYP : ONGOING PROJECTS

	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.

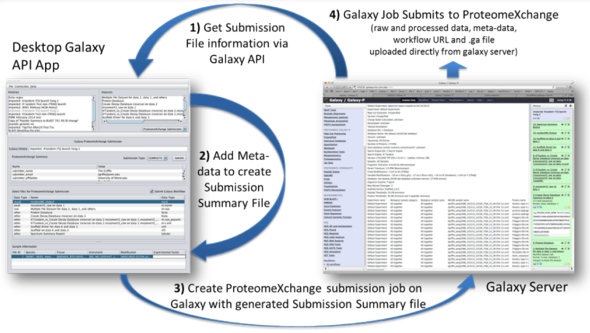
REPERTOIRE OF WORKFLOWS

- **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.**



COMMUNITY BASED SOFTWARE DEVELOPMENT

- **Community-based software development model should prove effective for future implementation, testing and continued improvement of command-line driven software tools.**
- **We plan to offer the many functionalities of MS-GF+ and PeptideShaker in Galaxy, along with opportunities for integration with other software tools via use of workflows.**



From Talk by Tim Griffin @ ASMS 2014.
'Public sharing of complex MS-based qualitative and quantitative proteomic data analysis workflows: adding value to big data repositories.'

SUBMITTING DATASETS TO DATA REPOSITORIES

- **Modification of ProteomeExchange to communicate with Galaxy API.**
- **Deployment of existing tools in Galaxy for ProteomeExchange submission (e.g. PeptideShaker tools).**
- **Automated data retrieval – re-analysis and mining of public data for new discoveries.**

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Biochemistry, Molecular Biology & Biophysics

Tim Griffin

Department of Medicine

Brian Sandri
Kevin Viken
Maneesh Bhargava
Chris Wendt

Department of Biology (Duluth)

Matt Andrews
Katie Vermillion

School of Dentistry

Joel Rudney

Center for Mass Spectrometry and Proteomics

LeeAnn Higgins
Ebbing de Jong
Todd Markowski

UNIVERSITY OF MINNESOTA
SUPERCOMPUTING INSTITUTE

James Johnson
Bart Gottschalk
Getiria Onsongo
Trevor Wennblom
Anne Lamblin
Ben Lynch

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Gloria Sheynkman
Lloyd Smith
Michael Shortreed

AB SCIEX

Sean Seymour

BIO-RAD

Sricharan Bandhakavi

COMMUNITY BASED SOFTWARE DEVELOPMENT

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

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

Ira Cooke
La Trobe University, Melbourne , Australia

Bjoern Gruening
University of Freiburg, Freiburg, Germany

Sangtae Kim
Pacific Northwest National Laboratory, Richland, WA

John Chilton
Galaxy Team
Penn State University

Funding
NSF, NIH

NSF