Large-scale multi-omic data integration and analysis: challenges and opportunities

Biomedical Informatics and Computational Biology
Research Symposium
January 17, 2014

Tim Griffin
tgriffin@umn.edu

Outline

- Historical perspective on multi-omics: yesterday and today
- Informatic challenges in multi-omics
- A solution: The Galaxy framework
- Galaxy in use
  - Proteogenomics
  - Metaproteomics
- Concluding thoughts

Acknowledgements

University of Minnesota
Ebbing de Jong
Joel Kooren
Sri Bandhakavi

Dr. Joel Rudney

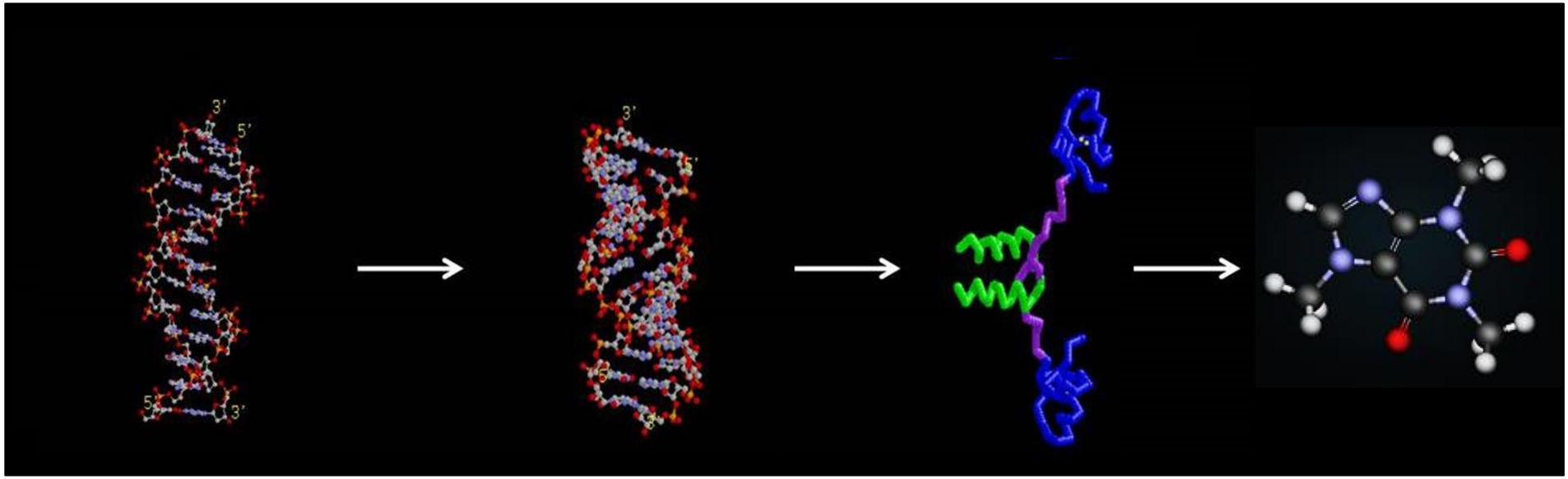
University of Minnesota Supercomputing Institute
John Chilton (Penn State)
Ben Lynch
James Johnson
Getiria Onsongo
Bart Gottschalk

University of Wisconsin-Madison
Gloria Shenykman
Dr. Lloyd Smith

International collaborators
Dr. Ira Cooke (La Trobe University)
Dr. Lennart Martens (Ghent University)
Dr. Conrad Bessant (Queen Mary University of London)

Funding
NSF, NIH

Starting point: connecting the “-omes” of biology



DNA
Genome

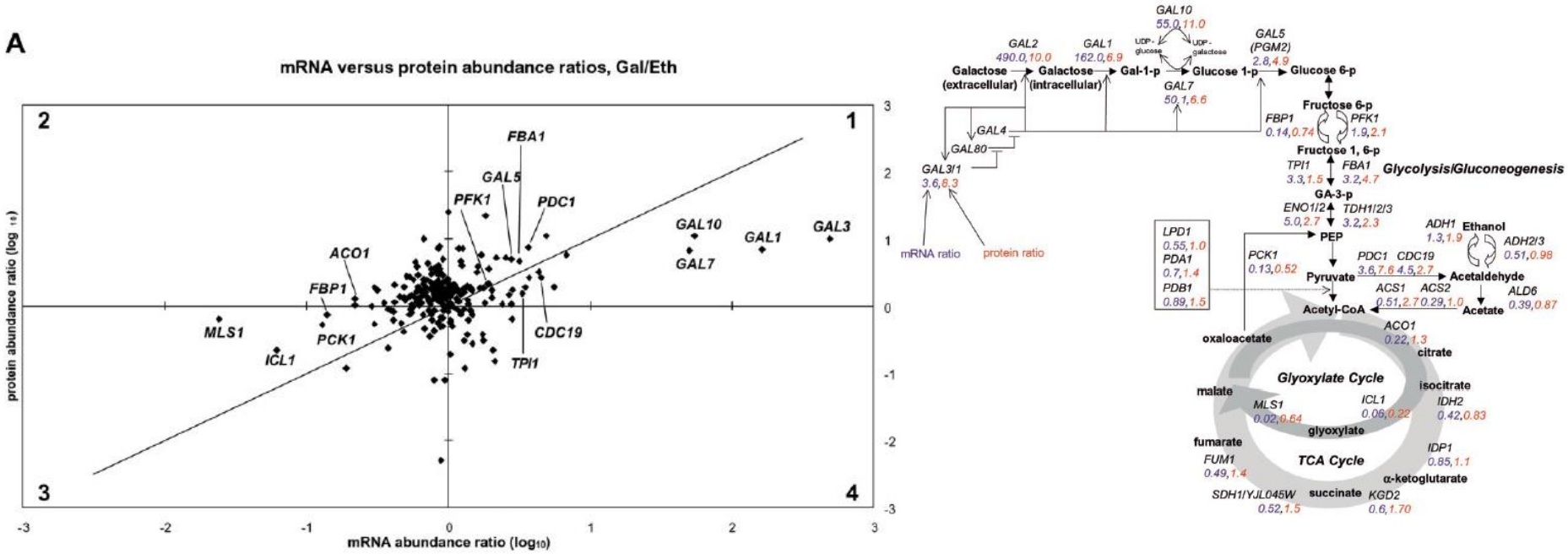
RNA
Transcriptome

Protein
Proteome

Metabolite
Metabolome

- Integrating ‘omic data (i.e. multi-omic data) reveals new molecular connections and cause/effect relationships

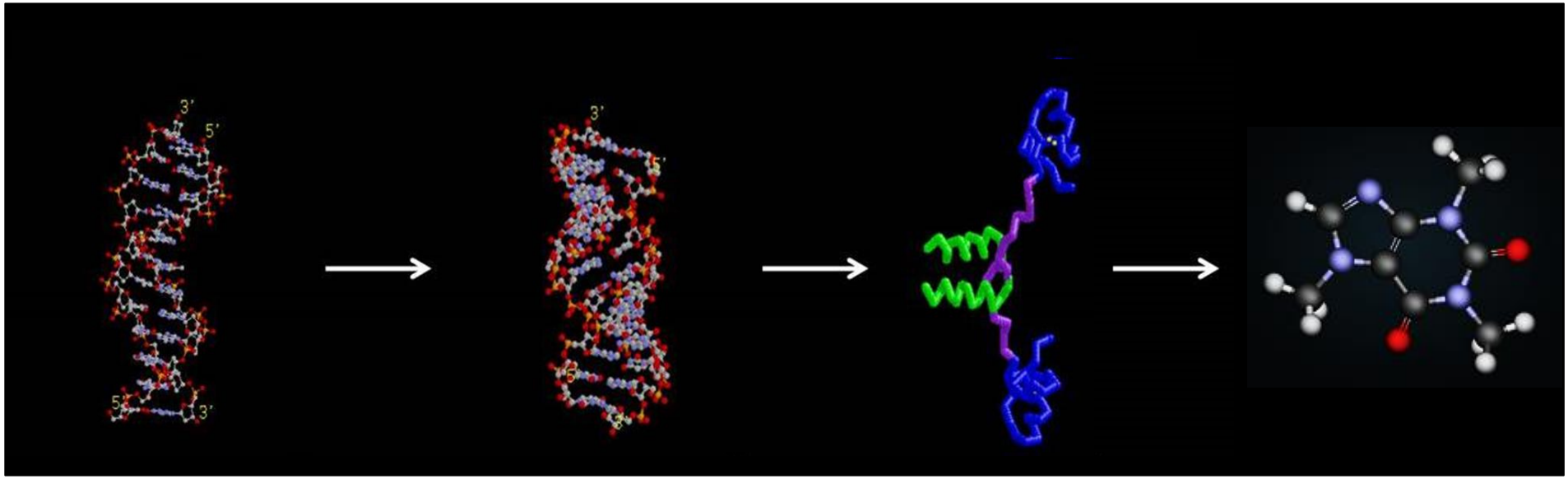
Historical perspective: multi-omics circa 2002



Molecular & Cellular Proteomics 1:323–333, 2002.

- ICAT labeling for quantitative proteomics
- LCQ mass spectrometer
- DNA microarray containing ~6200 yeast ORFs

Flash-forward: New and improved ‘omics technologies



DNA
Genome

RNA
Transcriptome

High-throughput sequencing

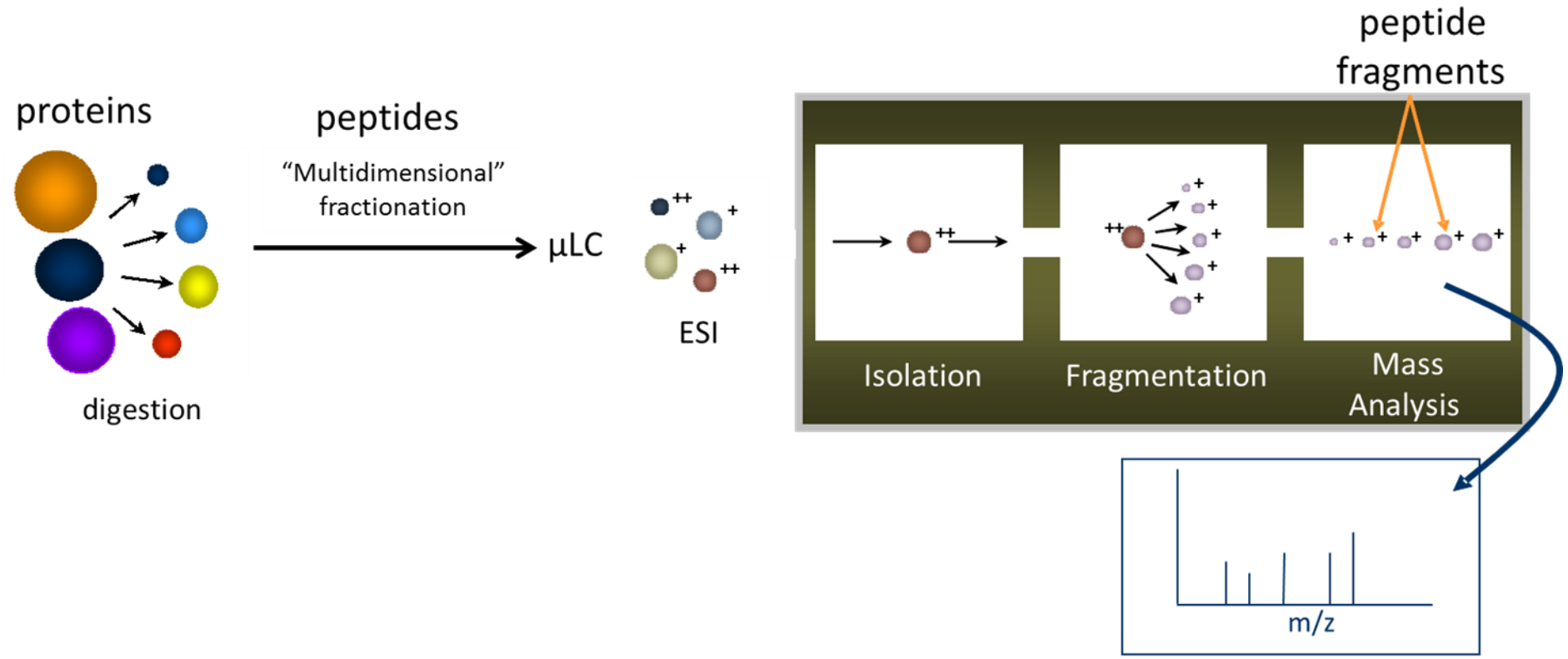
Protein
Proteome

Metabolite
Metabolome

High resolution mass spectrometry

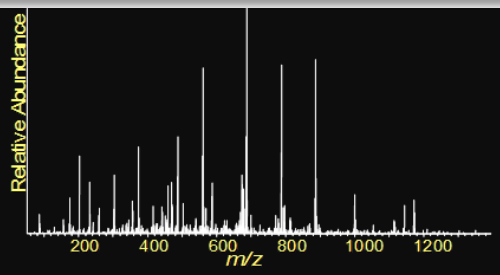
Technology example: MS-based proteomics

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)



Protein identification from MS data

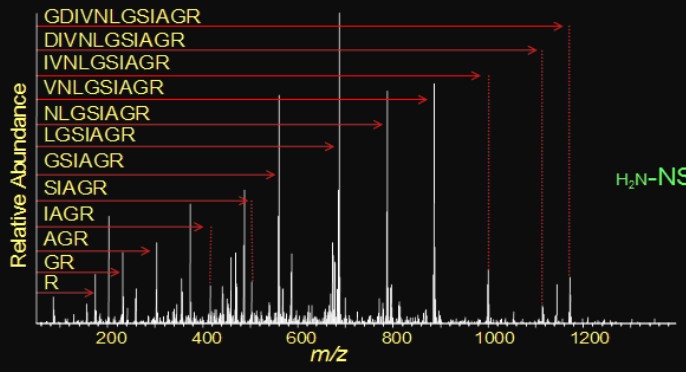
Raw MS/MS spectrum



Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

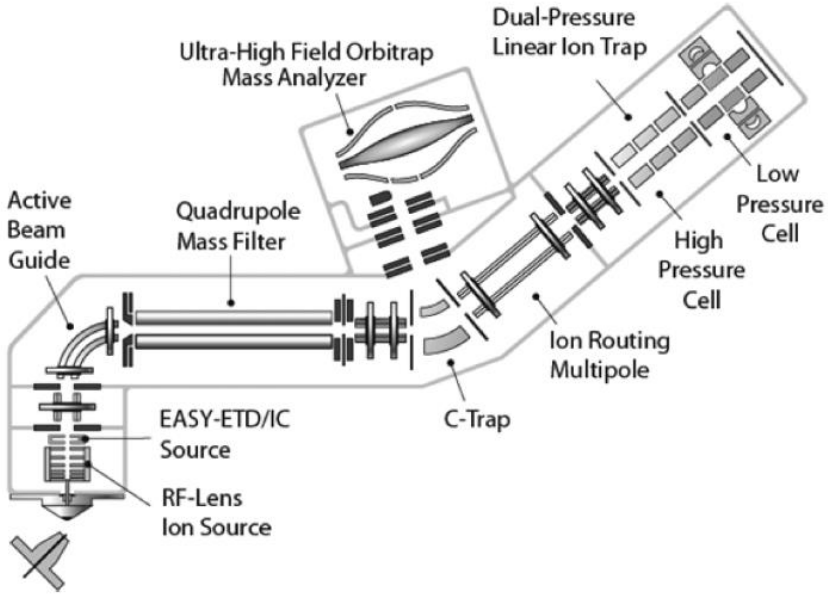
Peptide sequence match



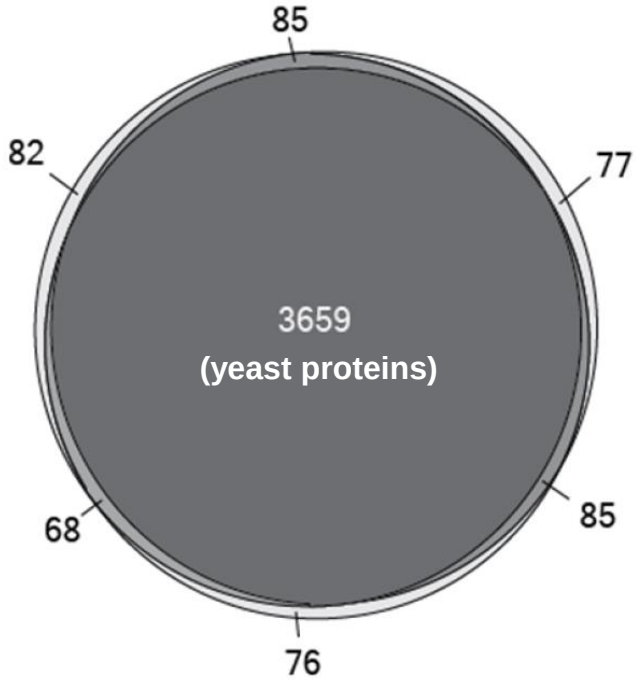
H_2N-NSGDIVNLGSIAGR-COOH

Protein identification

Realizing comprehensive/reproducible proteome?



Orbitrap Fusion mass spectrometer



Single LC-MS data acquisition in triplicate!

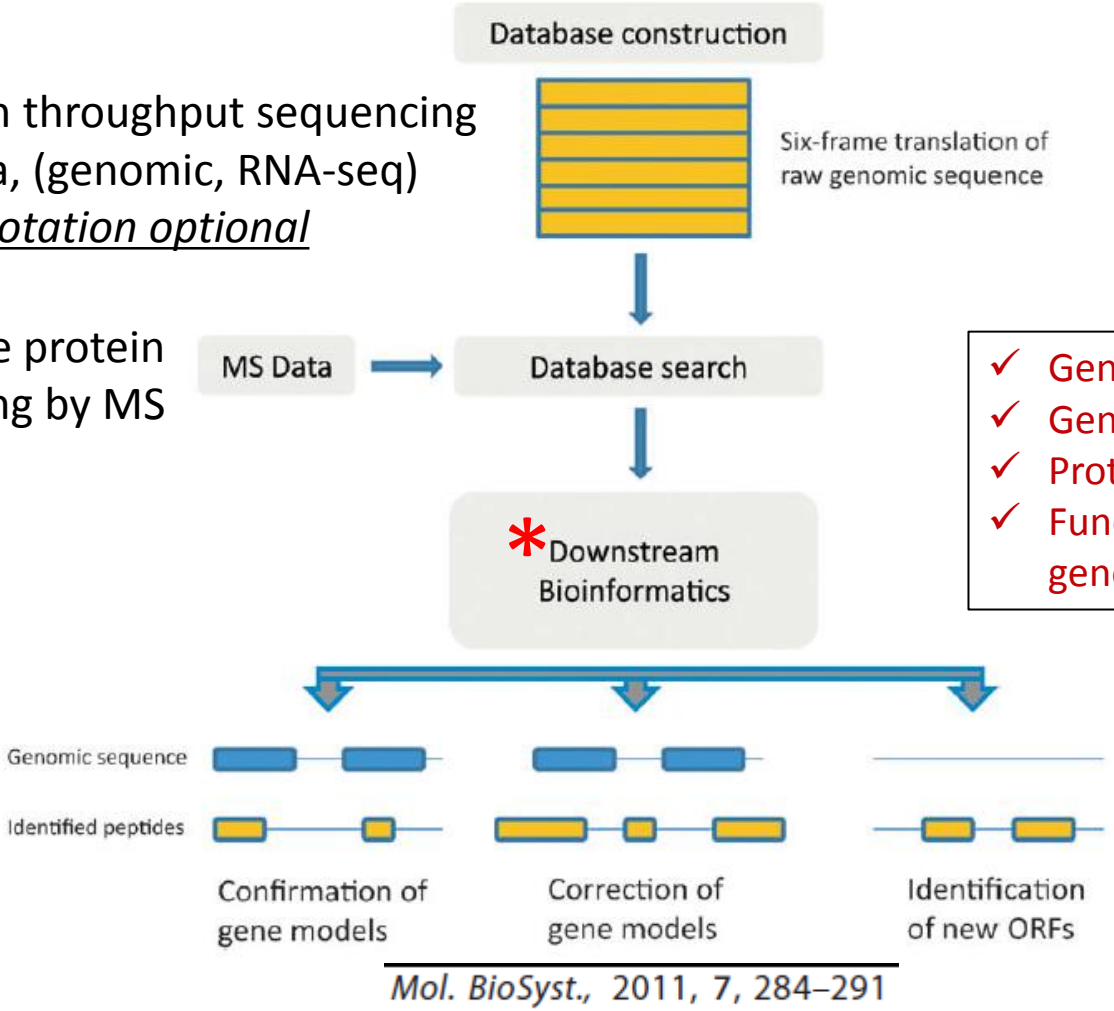
Anal. Chem. 2013, 85, 11710−11714

Converging technologies lead to new multi-omic possibilities

Proteogenomics

High throughput sequencing data, (genomic, RNA-seq) *annotation optional*

Comprehensive protein sampling by MS



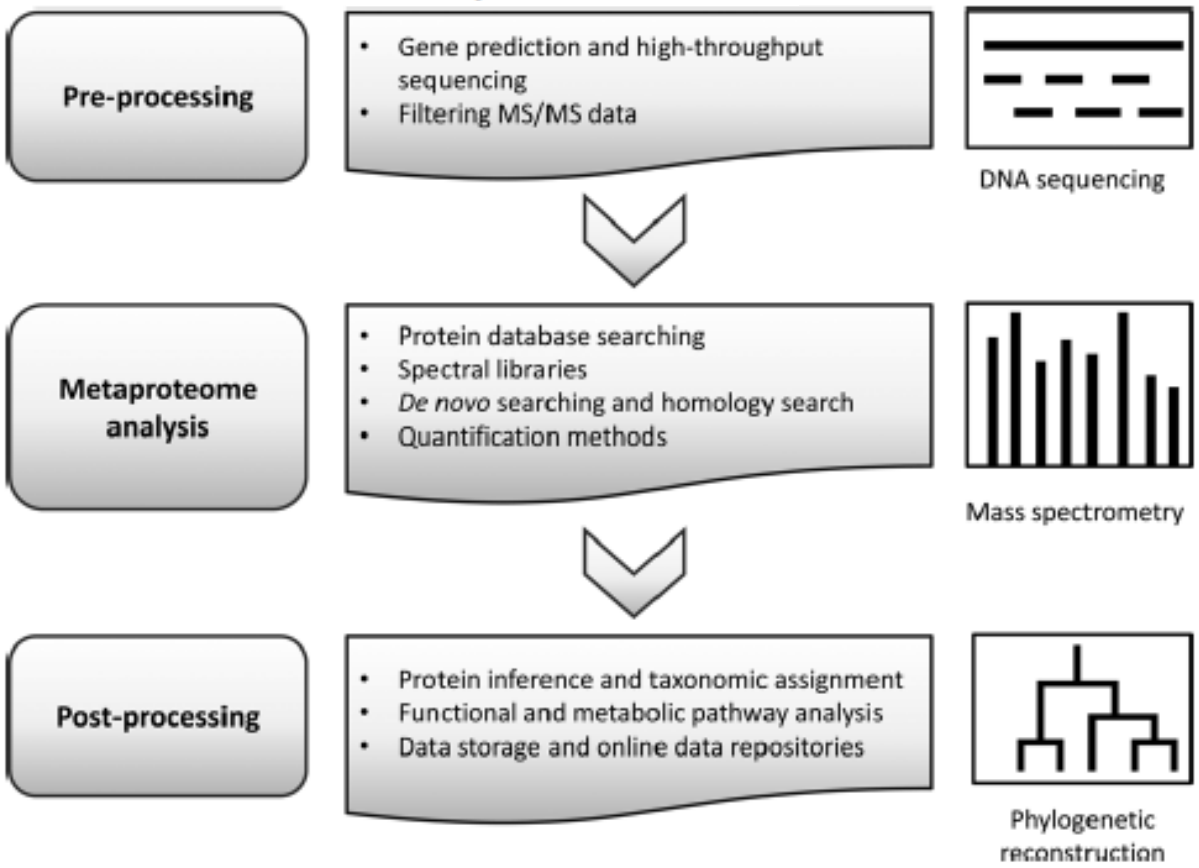
- ✓ Genome annotation
- ✓ Gene expression regulation
- ✓ Protein variants in disease
- ✓ Functional outcomes of genome mutation

Mol. BioSyst., 2011, 7, 284–291

Converging technologies lead to new multi-omic possibilities

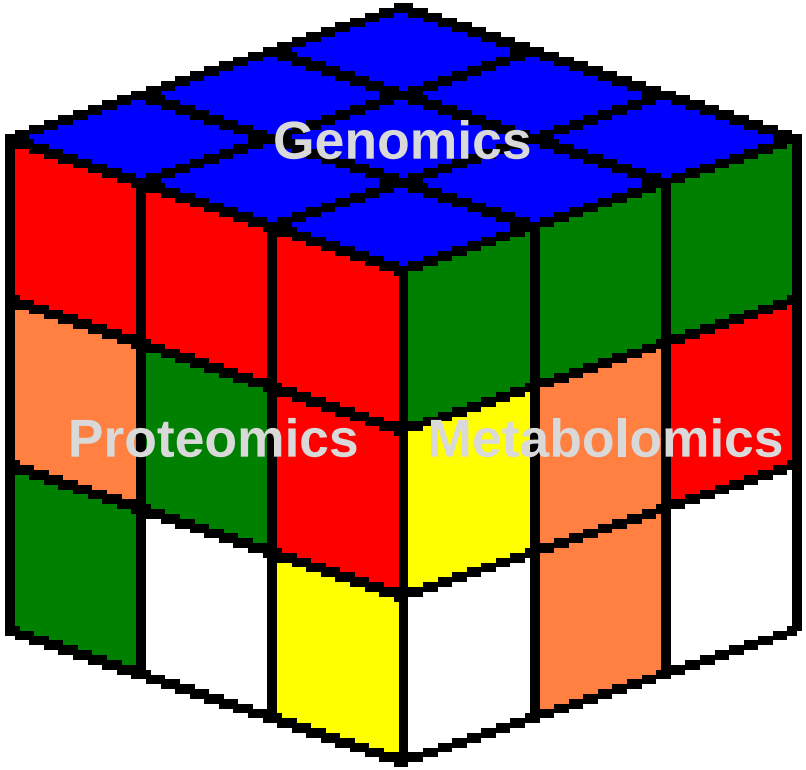
Metaproteomics (aka Community Proteomics)

Mol. BioSyst., 2013, **9**, 578–585



- Characterizes collection of proteins expressed by the community offering insight into conferred biochemical functions

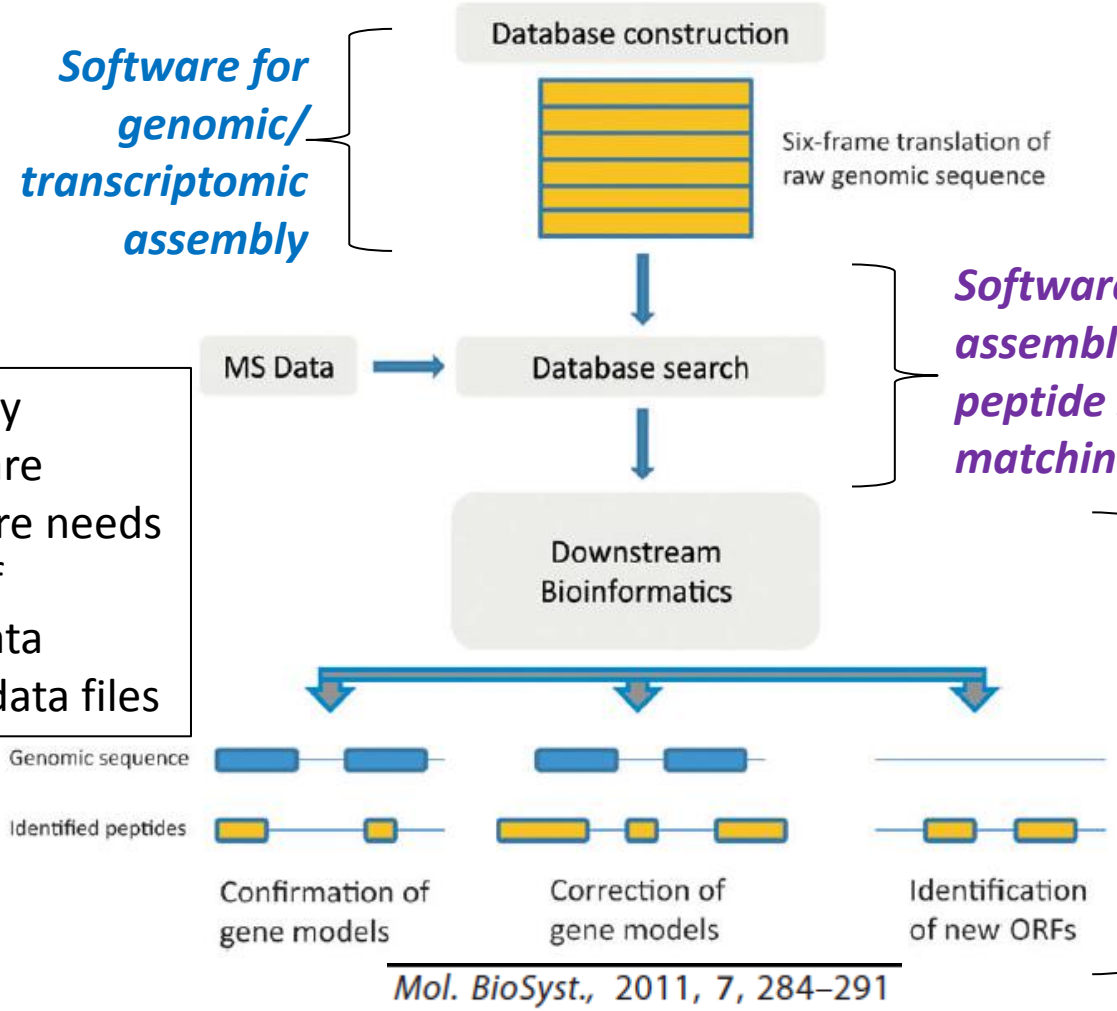
The tie that binds: informatics and computing



- Tools needed to solve a multidimensional, integrated puzzle

Challenge: use and integration of disparate software

- Mastery of many different software
- Diverse hardware needs
- Compatibility of input/output data
- Handling large data files



Software for genomic/ transcriptomic assembly

Software for dB assembly and peptide sequence matching

Software for integration of data and interpretation

A solution: The Galaxy Framework

Galaxy

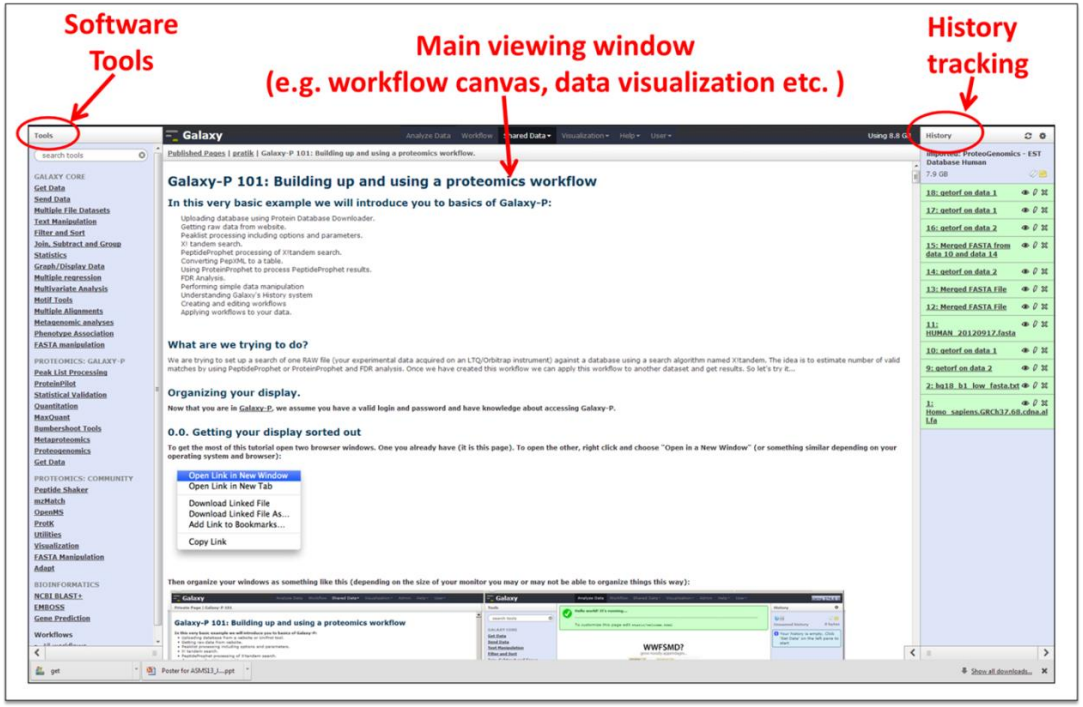
- A web-based, community developed bioinformatics framework/platform/workbench

- Originally designed to address issues in *genomic* informatics including:
  - Software accessibility and usability
  - Analytical transparency
  - Reproducibility
  - Scalability
  - Share-ability: complete sharing of even complex workflows

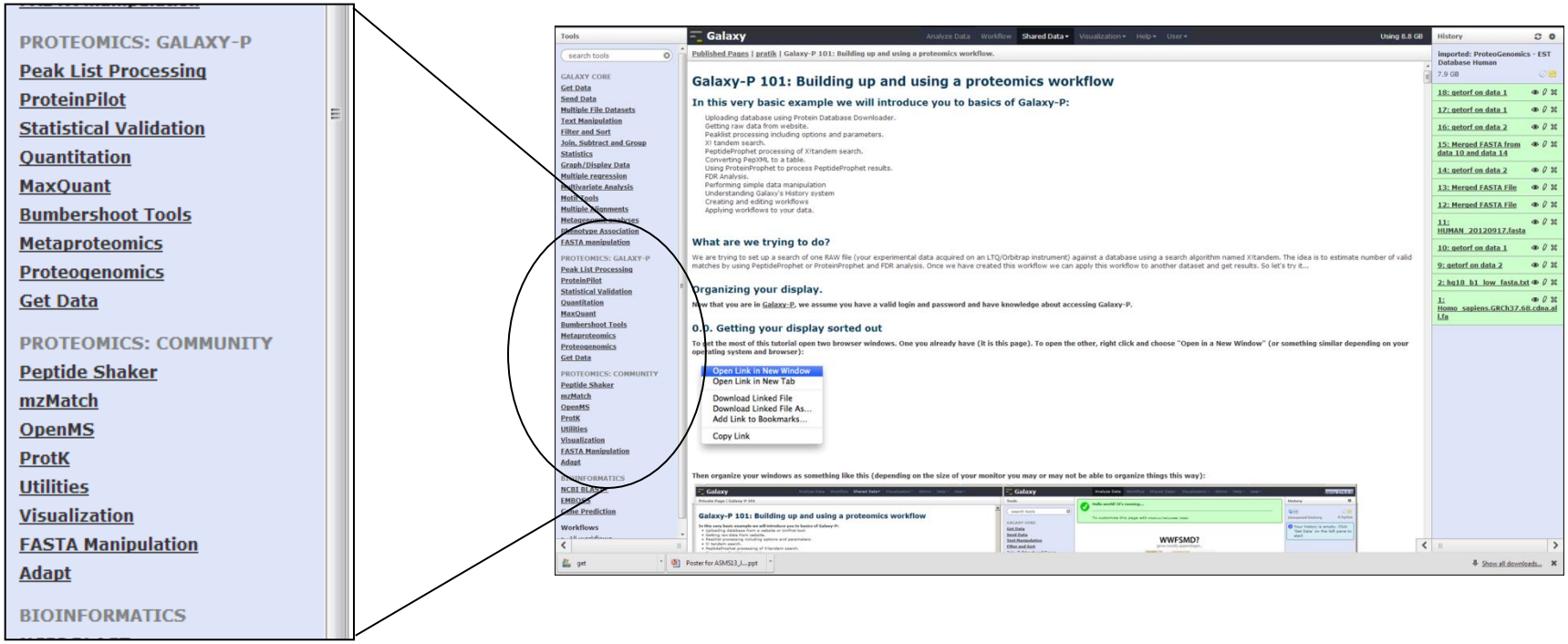
- **In a nutshell:** Galaxy provides an open framework into which disparate software programs can be deployed, integrated into customized workflows for typical to advanced applications, which can be shared in their entirety with other users

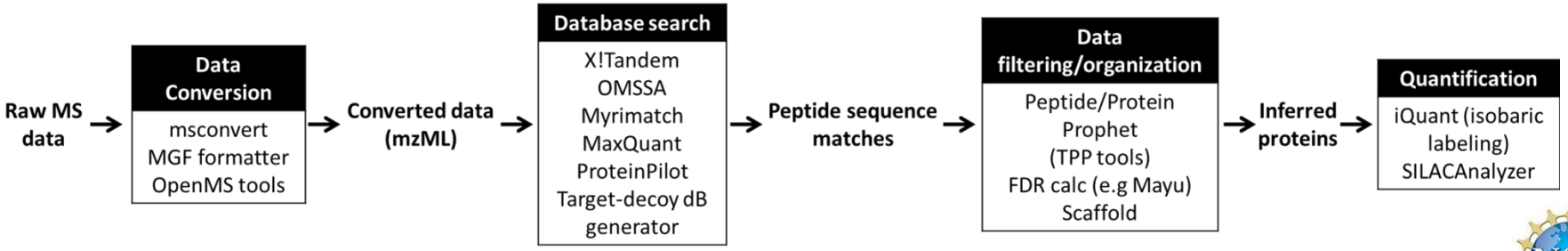
Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. ***Genome Biol.*** 2010, **11**: R86.

A (free) supermarket for ‘omics software?

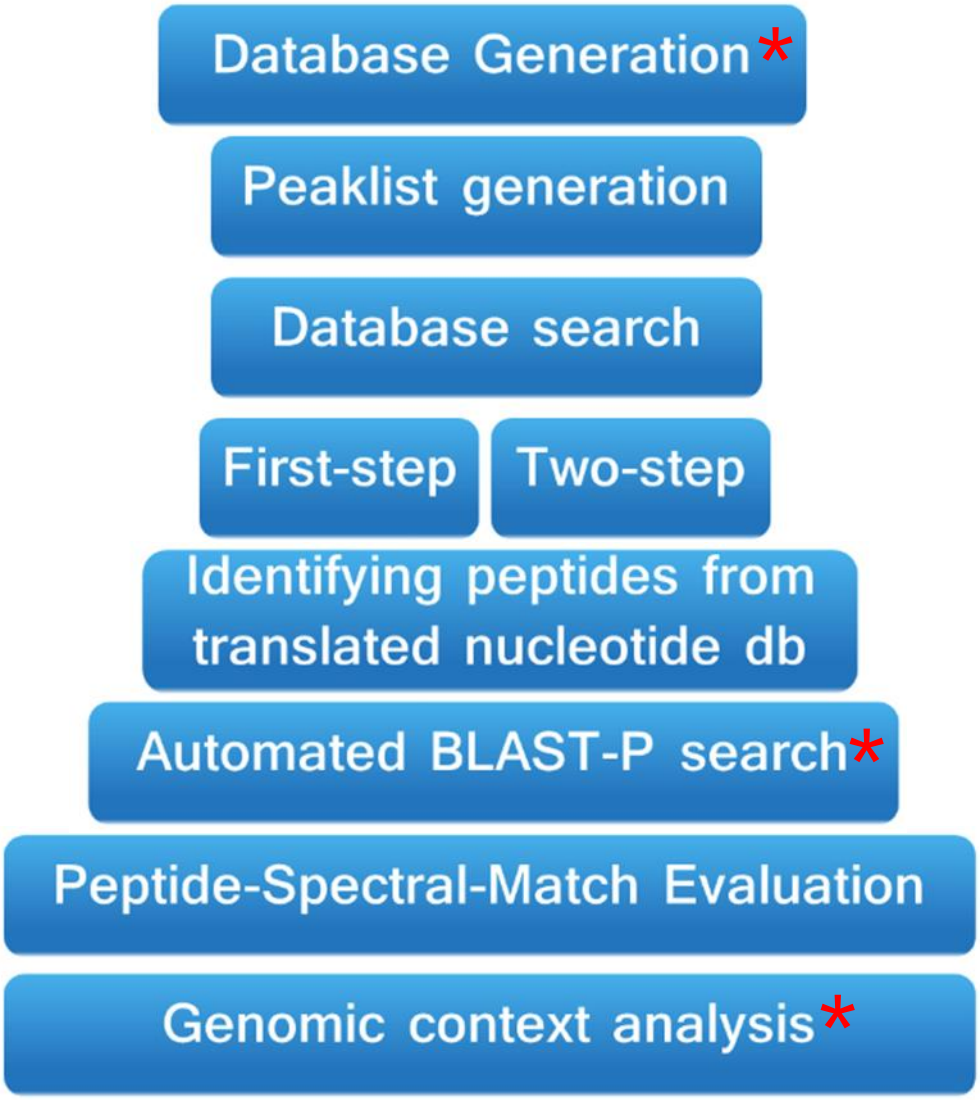


Extending Galaxy for multi-omics: GalaxyP



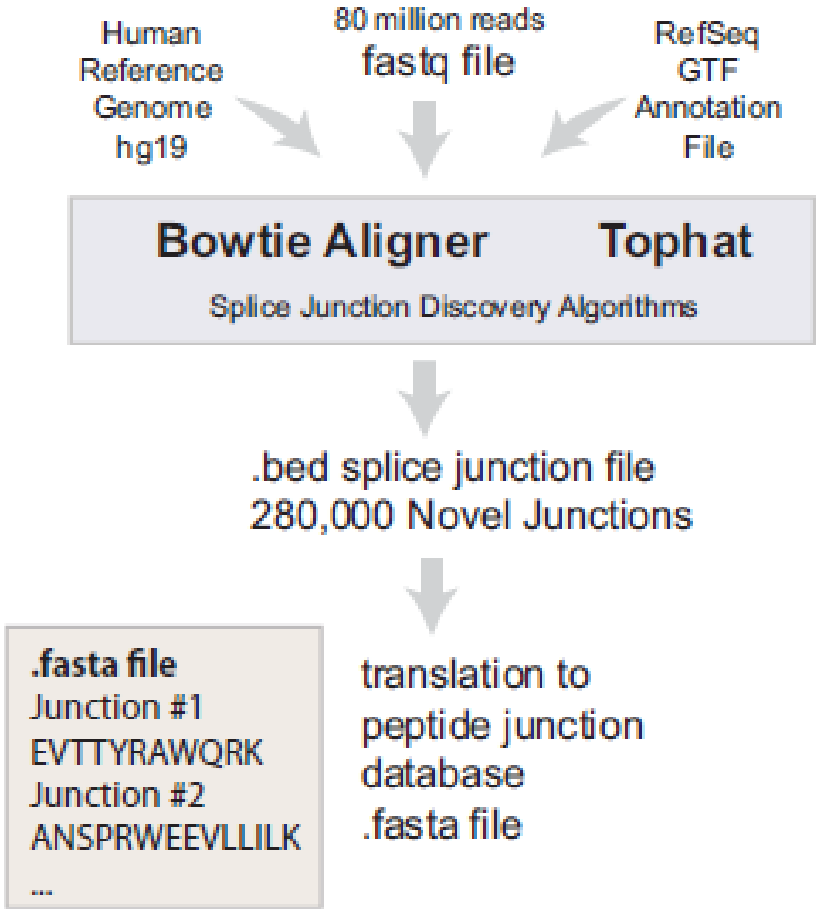


Example application: proteogenomics



Proteogenomics: protein database generation

Galaxy genomic software

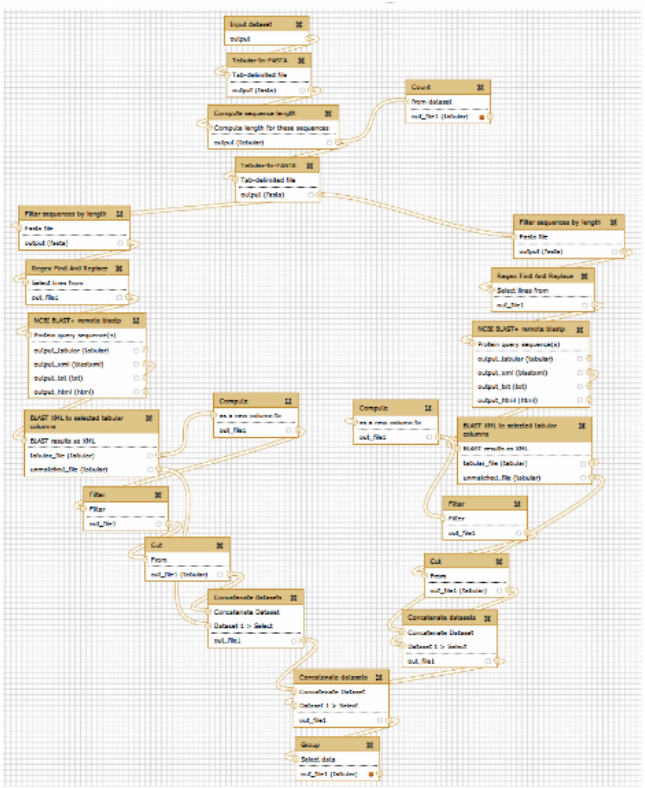


Gloria Sheynkman
UW-Madison

Mol Cell Proteomics (2013) 12, 2341-53

Assessing novelty: automated BLAST-P processing

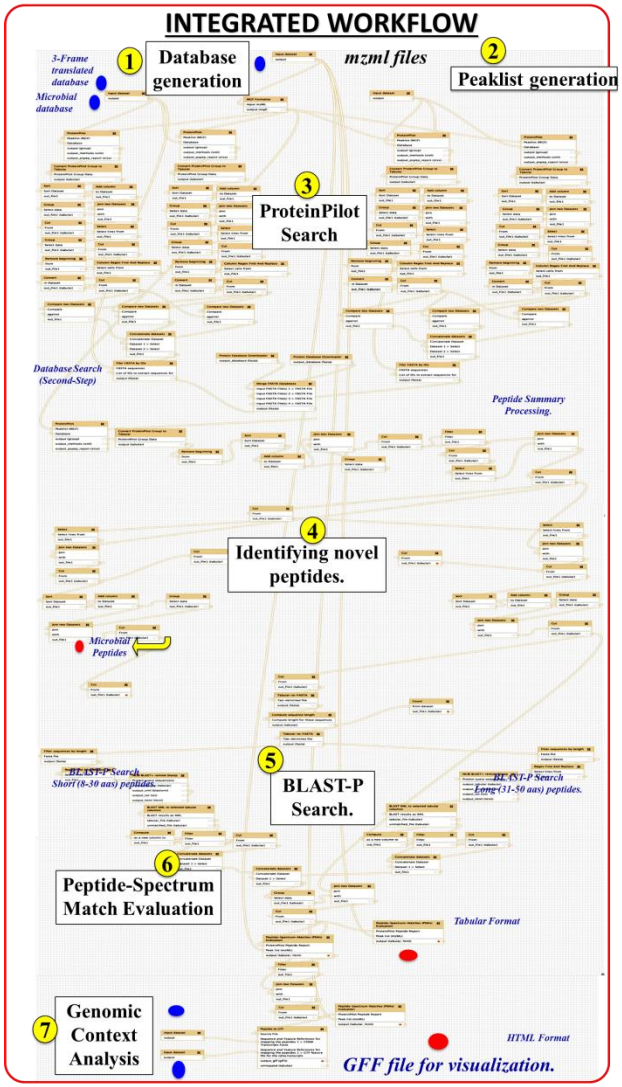
- Automatic searching of thousands of peptides against BLAST-P using criteria for small peptides (8-30 aa) and large (> 30 aa); flexible to different stringencies for “novel” sequences



Visualizing novel peptide hits

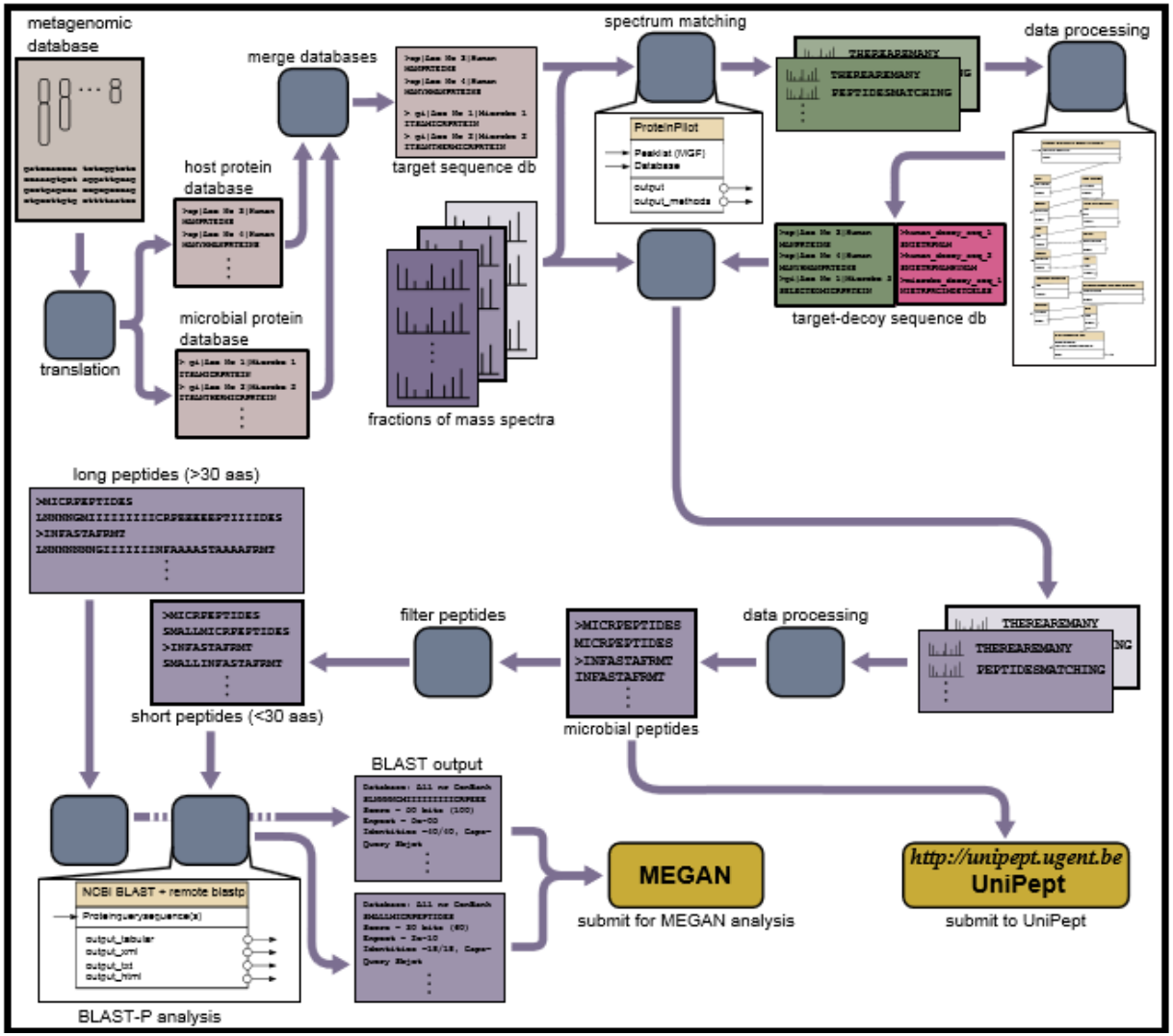
- **IGV compatible: Peptide-to-genome viewer**

Putting it all together



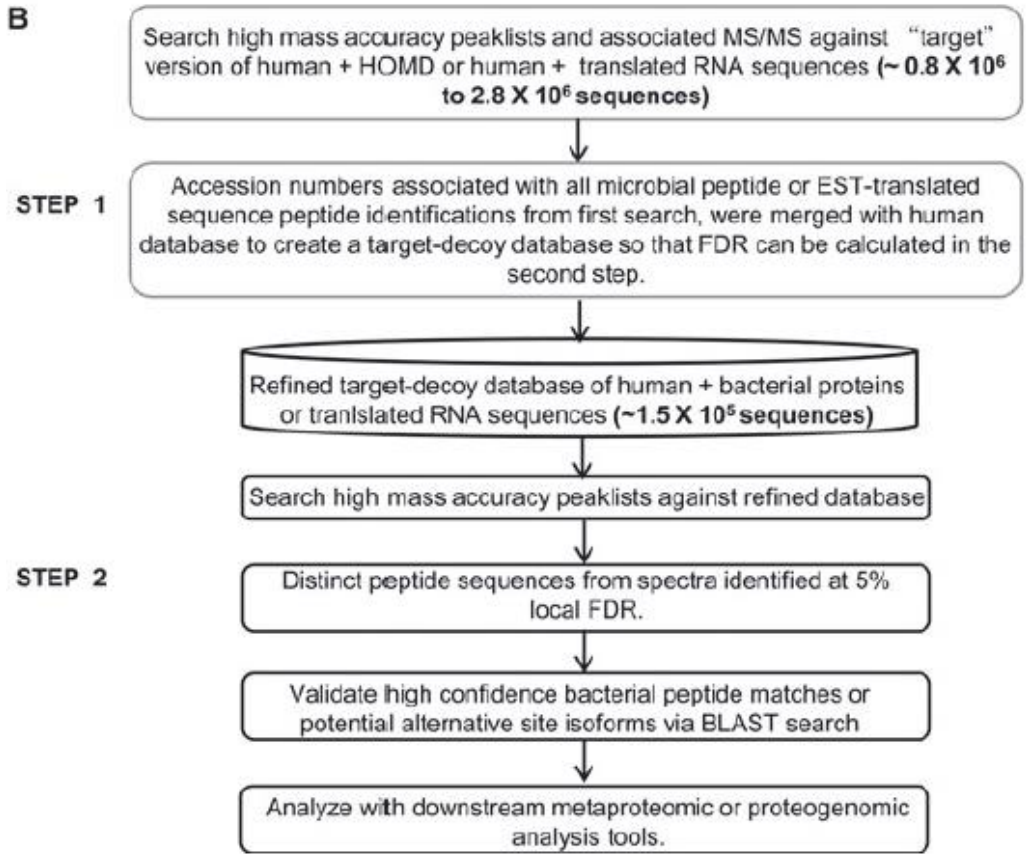
- 150 step workflow using diverse software, integrated and automated in Galaxy

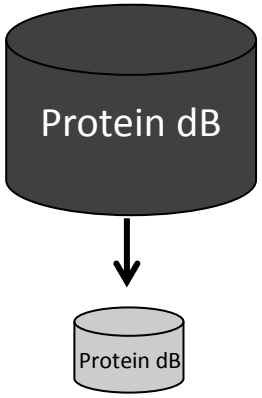
Metaproteomic workflow

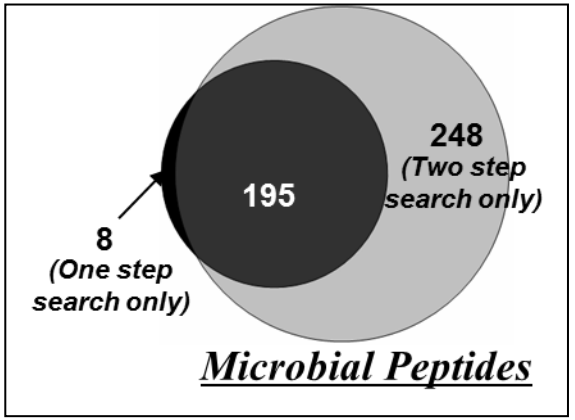


Increasing microbial peptide identifications

- **Addressing the large database challenge: 2-step database searching**



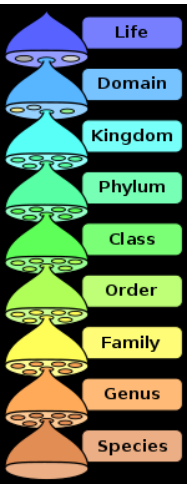




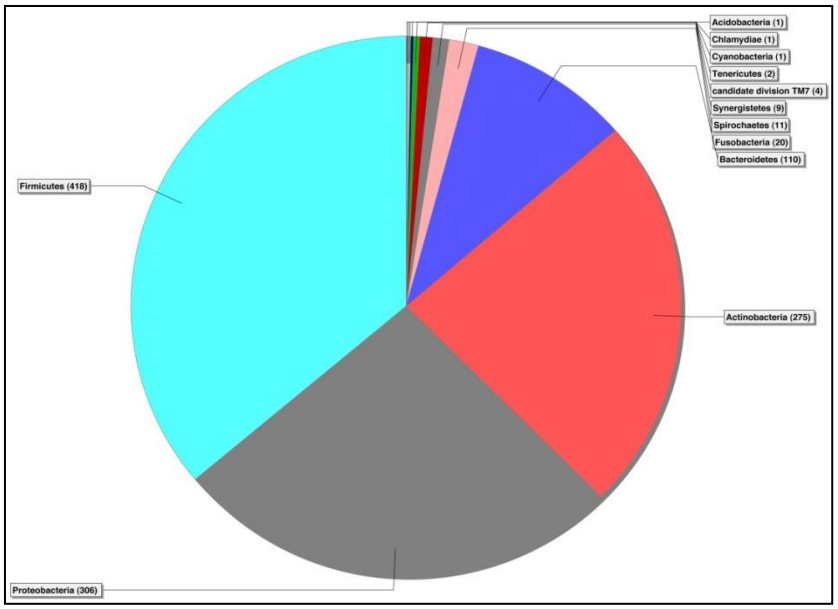
Jagtap et al *Proteomics*. 2013 (8):1352-7

Taxonomic analysis

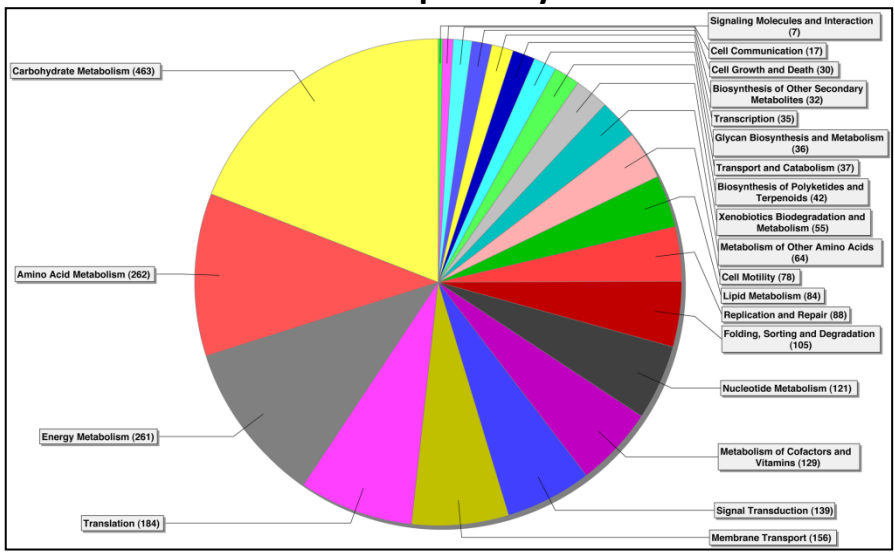
- **Output compatible with bioinformatic tools (MEGAN)**



Bacterial phyla



KEGG pathways



(Joel Rudney)

Proteomics 2012, *12*, 992–1001

Concluding thoughts: A new paradigm in publishing?

Old paradigm

2 Materials and methods

2.1 Salivary supernatant dataset

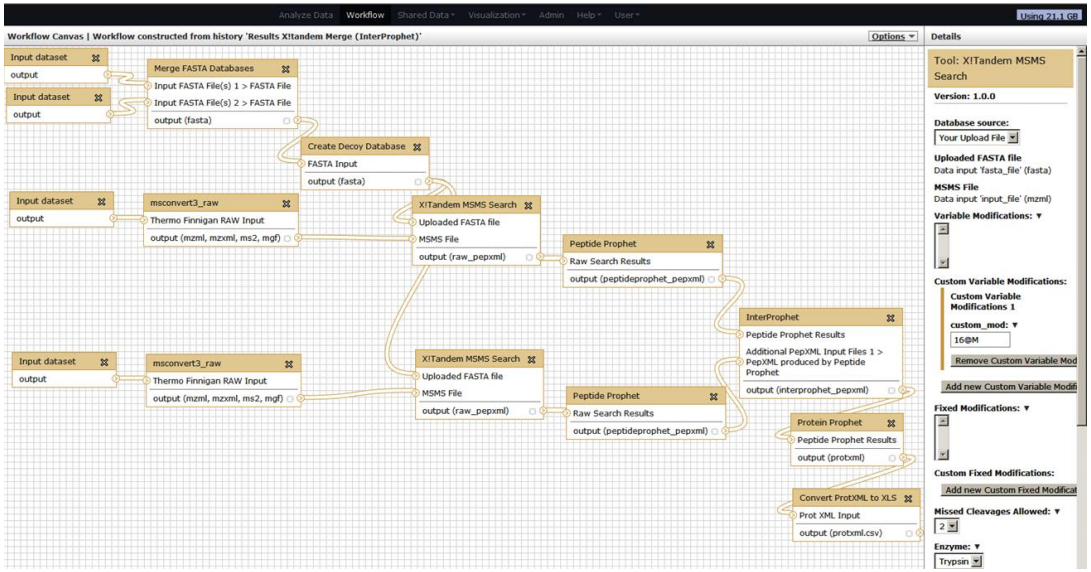
Salivary supernatant was collected and pooled from six healthy subjects who refrained from eating or drinking for 90 min. Proteins were analyzed using ProteoMiner™ (Bio-Rad Laboratories, Hercules, CA, USA) for DRC, multidimensional peptide fractionation, and an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) as described in Bandhakavi et al. 2009 [8]. Additional 45 RAW files generated from ProteoMiner ™ Library-2-treated saliva were also analyzed.

2.2 Two-step method for peptide sequence matching and protein identification

RAW files generated (200 total) from the LTQ-Orbitrap salivary supernatant dataset were processed using the MaxQuant (v1.0.13.13) "Quant" module to generate .MSM files [8, 22–24]. Individual Peak and iso MSM files corresponding to each .RAW file were converted to Mascot generic format (MGF) and searched using ProteinPilot v 4.0 (ProteinPilot Software 4.0; Revision: 148085; Paragon Algorithm: 4.0.0.0. 148083; AB SCIEX, Framingham, MA). Paragon searches [34] were conducted using LTQ-Orbitrap subppm instrument settings. Other parameters used for the search were as follows: Sample Type: Identification; Cys alkylation: None; Digestion: Trypsin; ID Focus: Biological Modifications; Search effort: Thorough.

In the first step, all 200 RAW files were searched against a database consisting of all the translated human oral microbial genomic sequences from the Human Oral Microbiome Database (HOMD) [25], along with the human IPI v3.52 database and contaminant proteins (1 687 426 total protein sequences) [26]. The ProteinPilot searches generated a .group file that was used to generate a Protein Report from the peptide sequence matches. All nonhuman protein sequences that were identified at the threshold of at least 66% Conf score (0.47 ProtScore) in the first step were merged with the Human IPI v3.52 database along with contaminant proteins to generate a "refined" FASTA database for the second step.

In the second step, all 200 RAW files were searched against a "Target Decoy" version of the FASTA database mentioned above, by appending the reversed protein sequences to the forward sequences, resulting in a database containing 152 724 total protein sequences. Parameters for ProteinPilot



New paradigm: Transparent, complete and usable by others

Concluding thoughts

- “Big Data” respositories: Workflow framework (e.g. Galaxy) offers a way to store and use analytical tools/workflows with raw multi-omic data
- Better ways needed to integrate ‘omic data repositories to realize benefits of multi-omics
- Academic-industry partnerships: a way forward in solving data analysis challenges in multi-omics and Big Data?