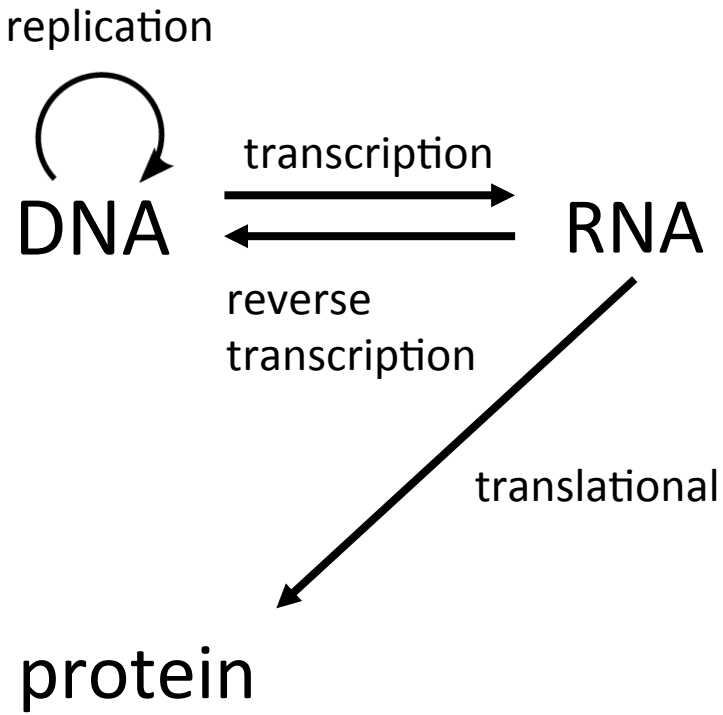
Building a low-budget public resource for large-scale proteomic analysis

Anoop Mayampurath
Computation Institute
April 17, 2014

THE UNIVERSITY OF CHICAGO

The world of proteomics



Proteins carry out **most** cellular activity, including **control** (regulation) of transcription, translation, and replication of DNA.

Alzheimer's Disease

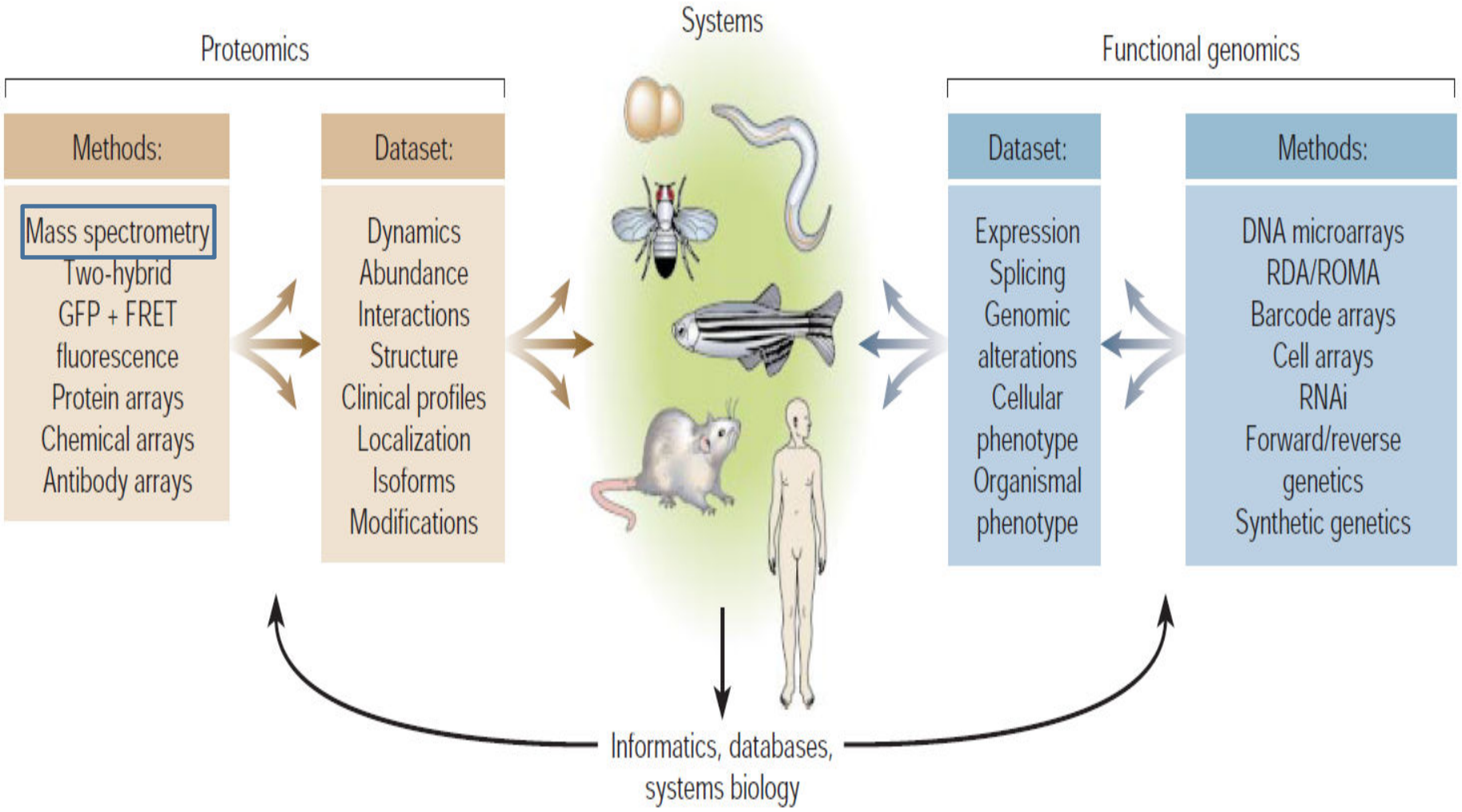
- Most common late-onset Alzheimer's gene : APOE (e2,e3, and e4)
- Plaque accumulation but do not develop Alzheimer's
- Repressor element-1 silencing transcription factor

ARTICLE

doi:10.1038/nature13163

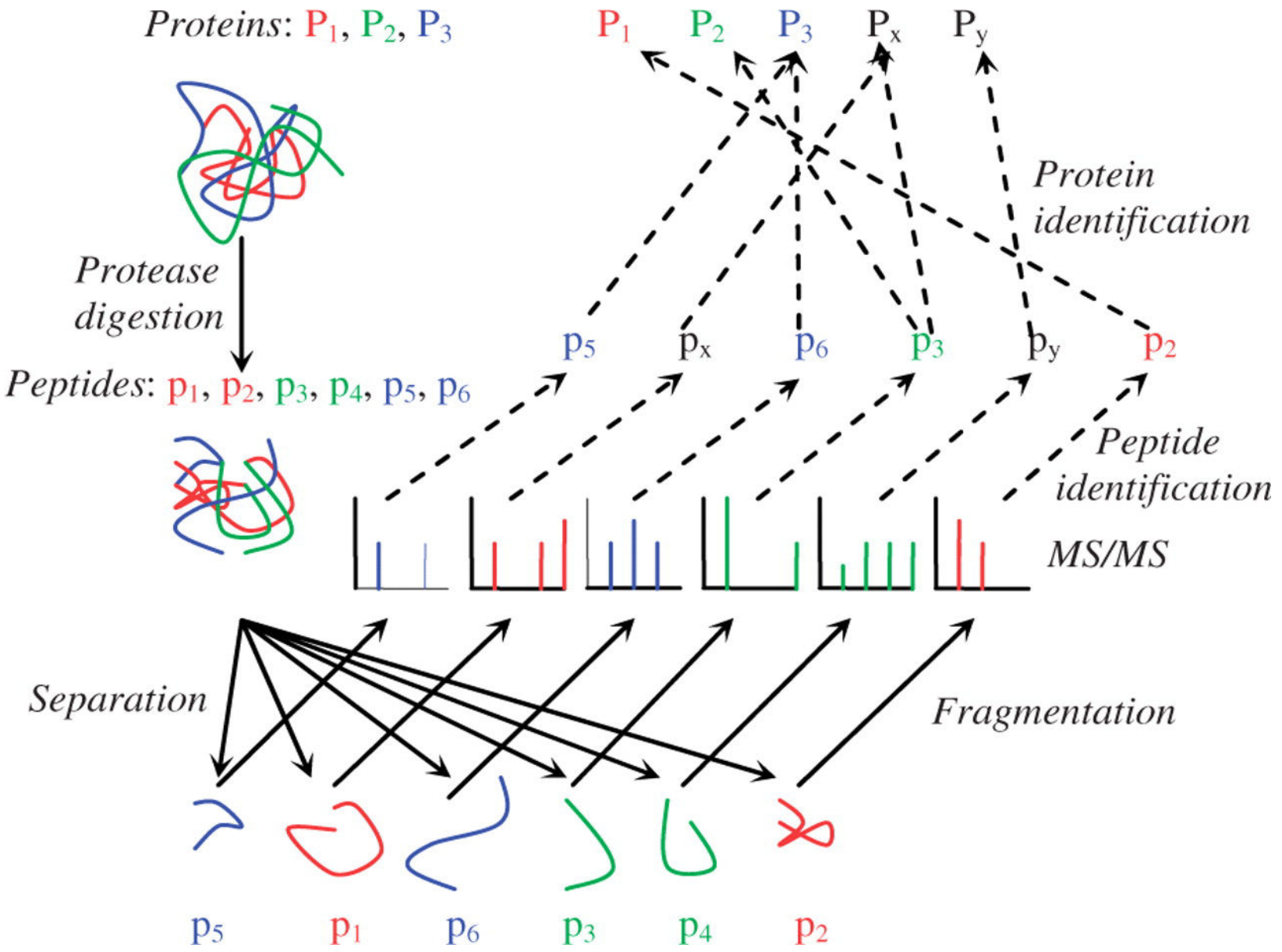
REST and stress resistance in ageing and Alzheimer's disease

Tao Lu[1], Liviu Aron[1], Joseph Zullo[1], Ying Pan[1], Haeyoung Kim[1], Yiwen Chen[2], Tun-Hsiang Yang[1], Hyun-Min Kim[1], Derek Drake[1], X. Shirley Liu[2], David A. Bennett[3], Monica P. Colaiácovo[1] & Bruce A. Yankner[1]



Tyers & Mann Nature 422, 2003

A little background on mass spectrometry-proteomics

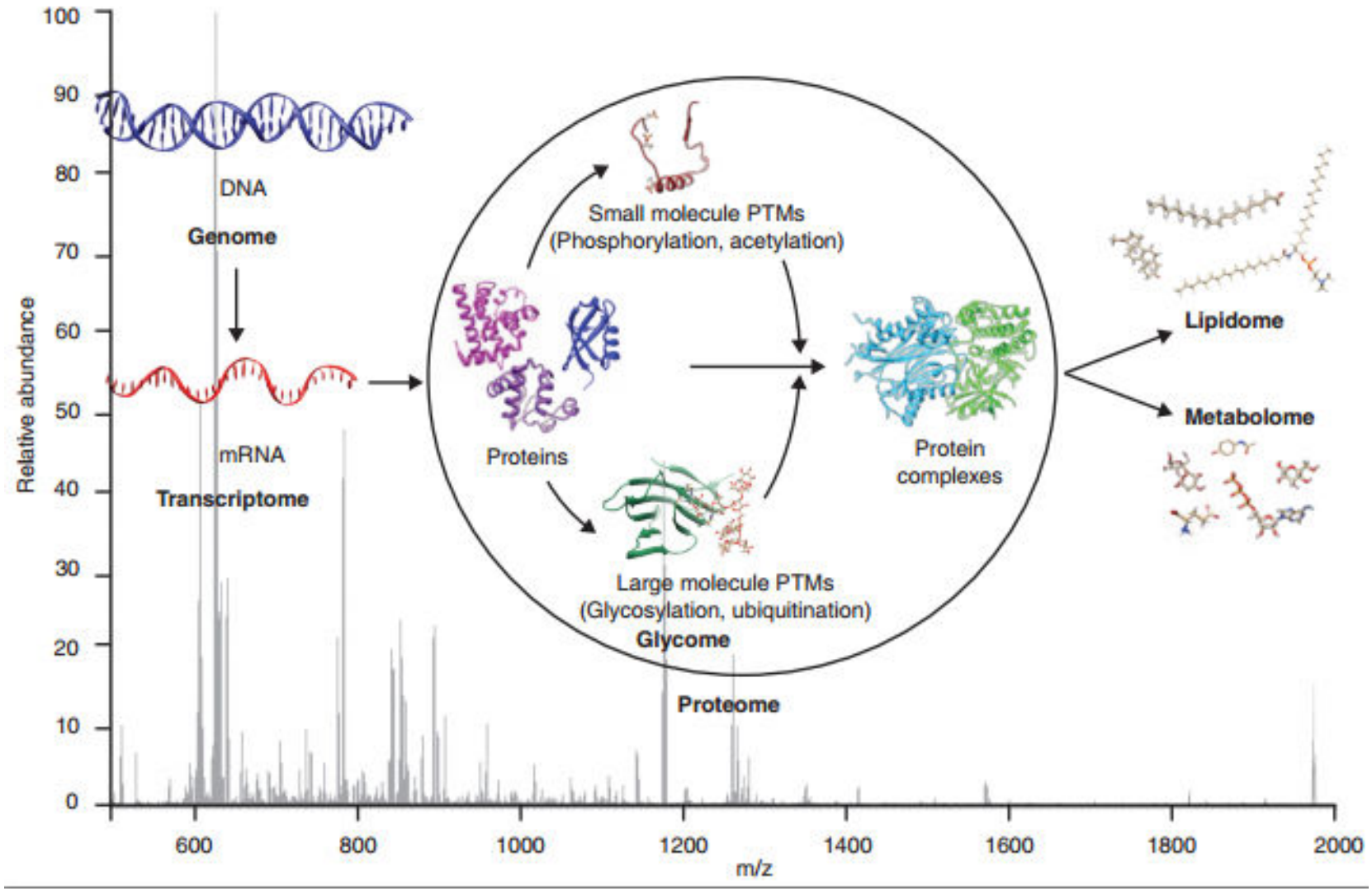


Shen C et al. Bioinformatics 2008;24:202-208

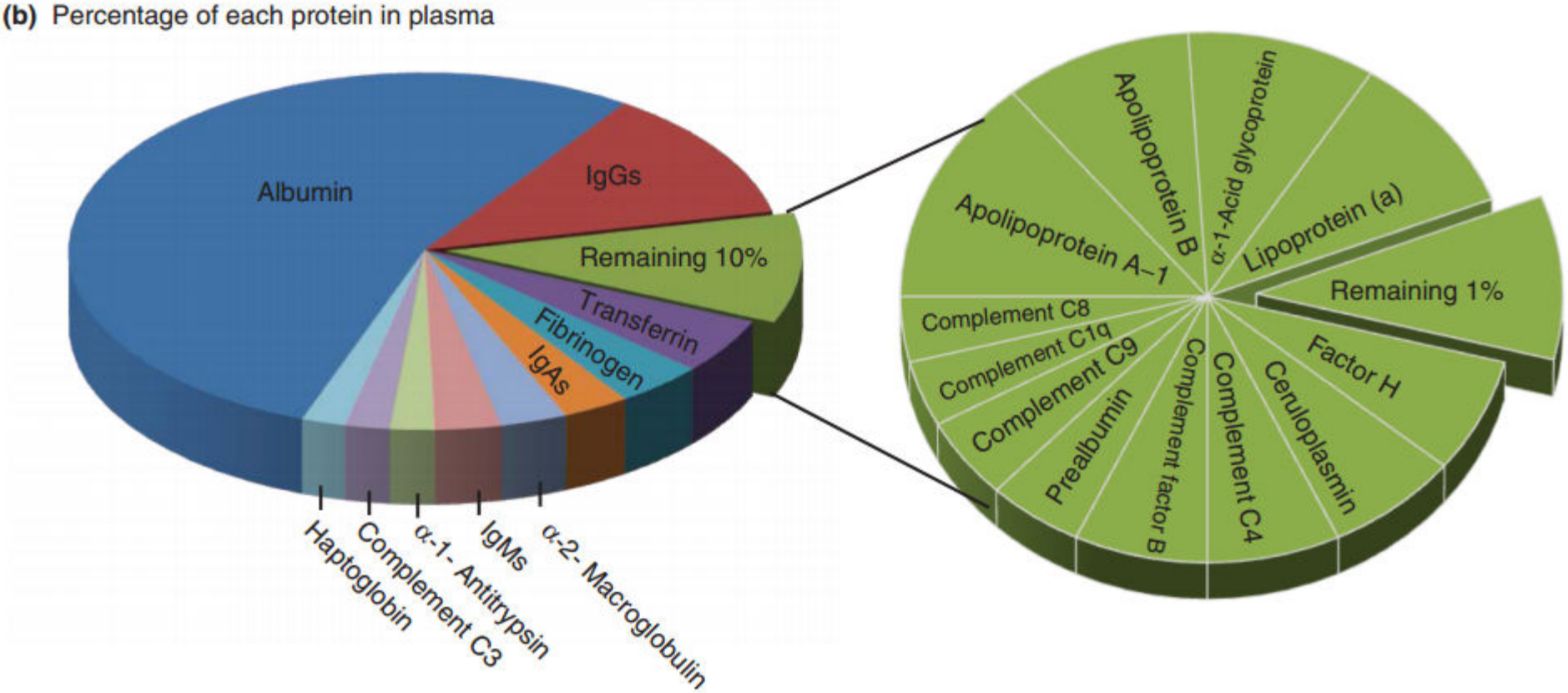
LC-MS/MS clinical proteomic challenges

Biology

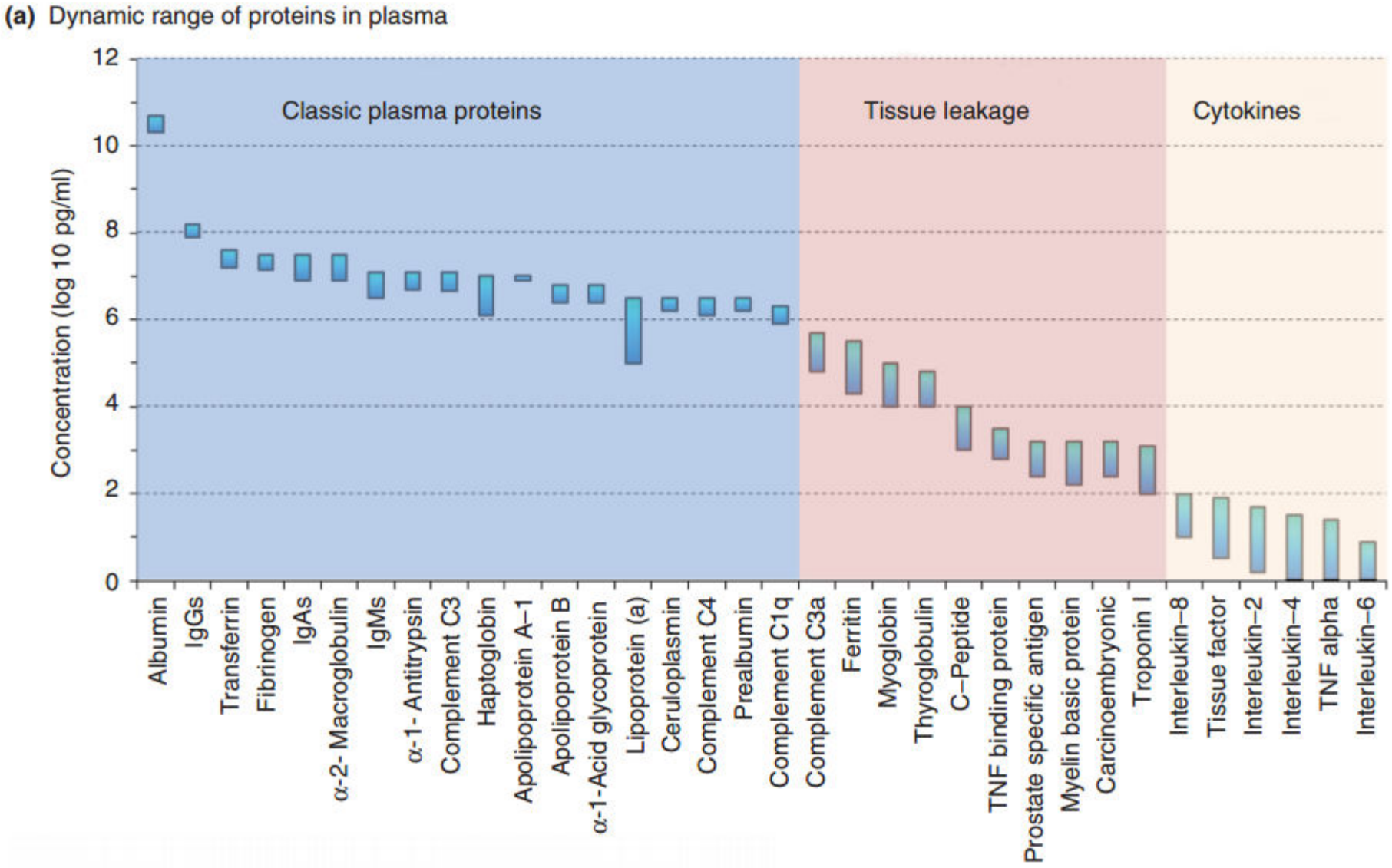
- Complexity of the proteome
- Dynamic range of plasma proteins



Baker *et al. Genome Medicine* 2012 **4**:63



Baker *et al. Genome Medicine* 2012 **4**:63



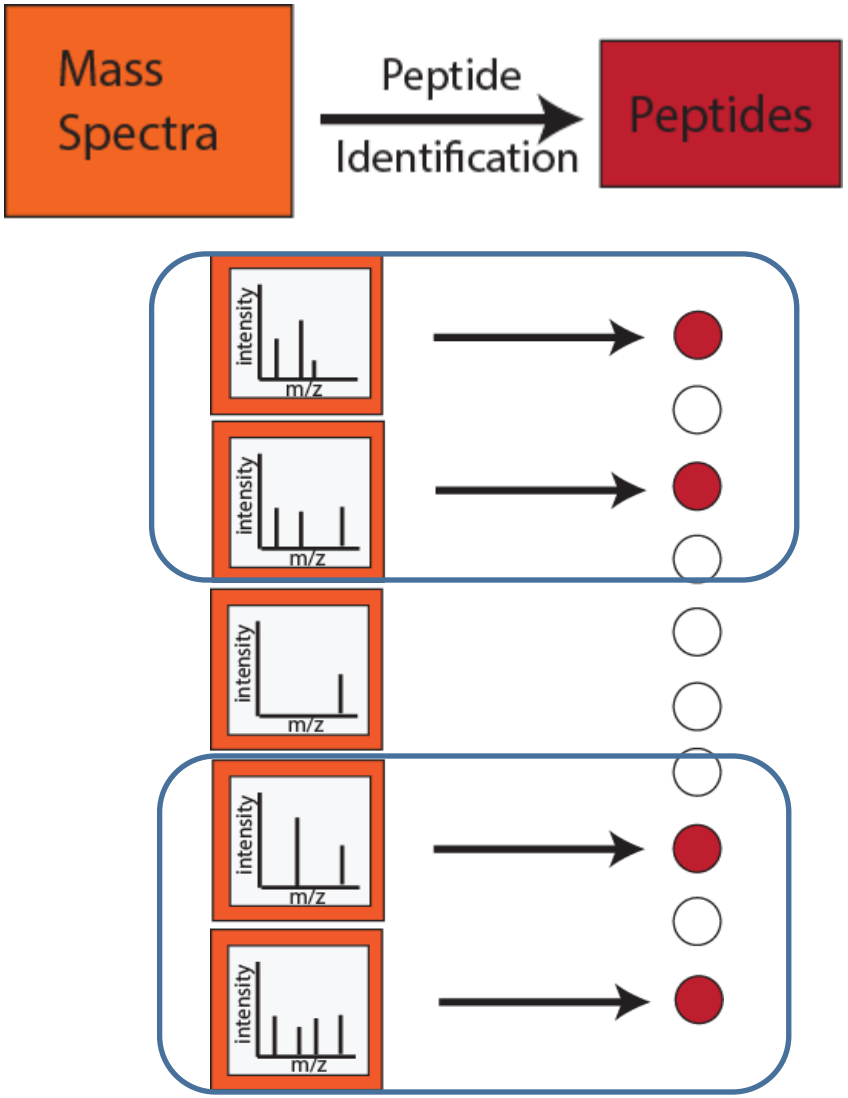
Baker *et al. Genome Medicine* 2012 **4**:63

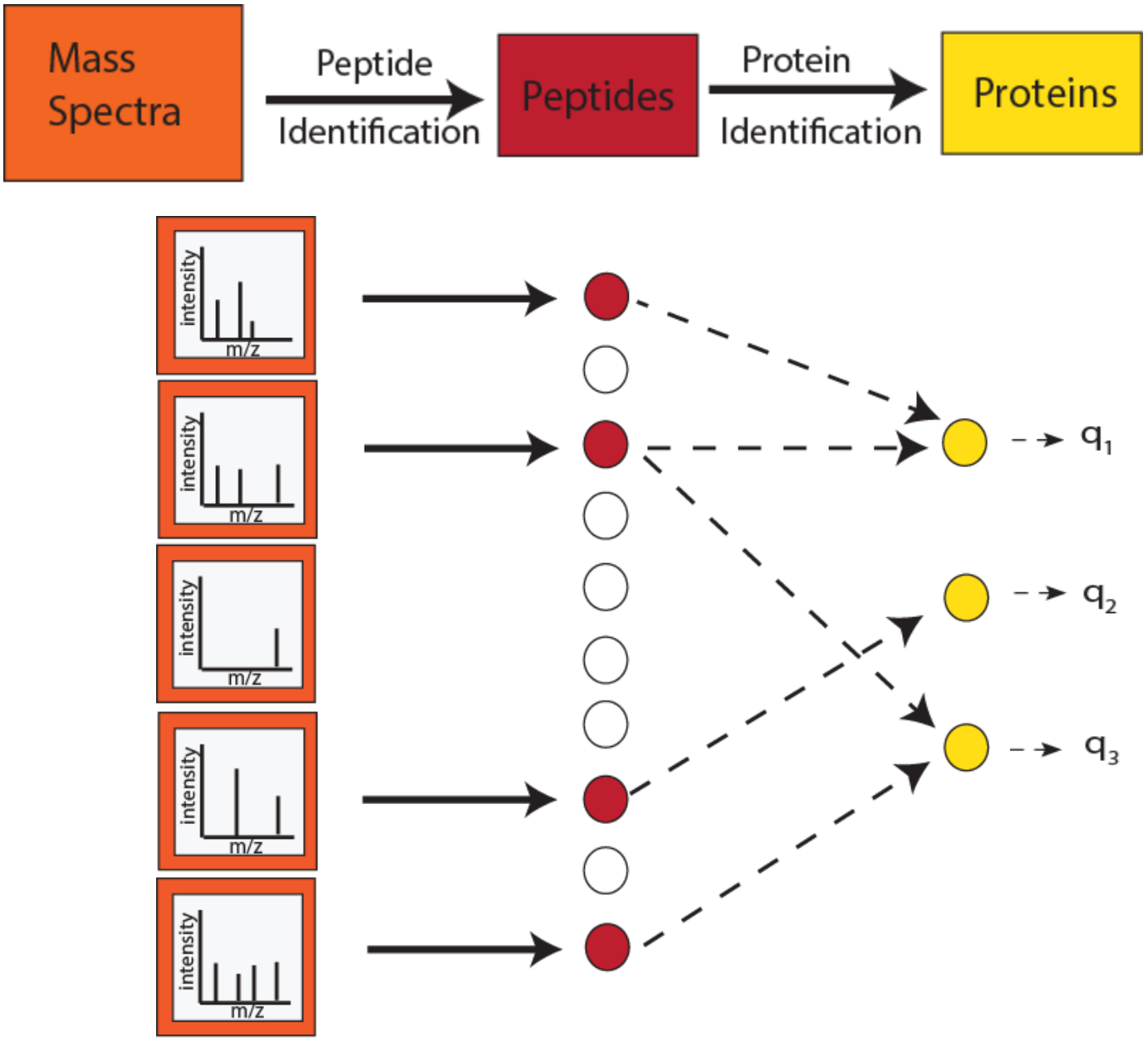
LC-MS/MS clinical proteomic challenges

Computational

- lack of a framework capable of performing large-scale proteomic analysis

The workflow




Mass Spectra
Peptide Identification
Peptides
Protein Identification
Proteins
intensity
m/z
q1
q2
q3

Software List

platforms, pipelines and libraries

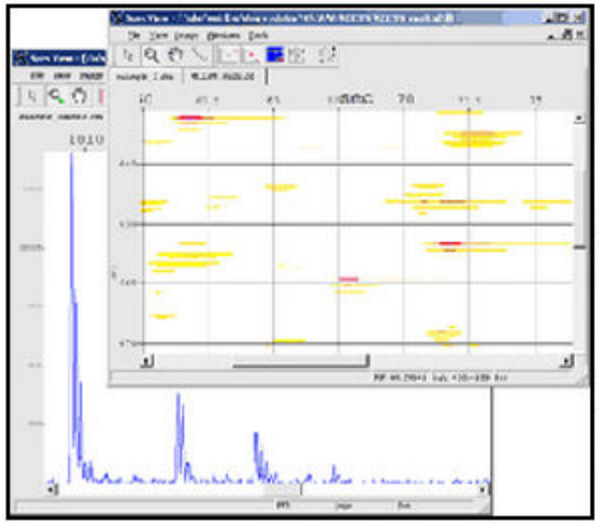
CPAS	LIMS and analysis tools for proteomics data (includes msInspect)		
CPFP	Central Proteomics Facilities Pipeline [1] (demo here)		
GenePattern	platform for integrative genomics and proteomics (includes PEPPeR [2] and other tools for proteomics)	Java	
InSilicoSpectro	open source proteomics library (of Perl functions) [3]	Perl	
libfbi	a fast implementation of box intersection for correspondence estimation in peak picking, alignment, etc.	C++	
Mass-up	utility with full GUI for proteomics data analysis, particularly MALDI-TOF	Java	
MASSyPup	a lightweight Linux live distribution prepackaged with XTandem, mMass, MZmine, PepNovo/UniNovo, PeptideShaker, mscovert, XCMS etc.		
mspire	MS data processing in Ruby, including mzML reader/writer, *in-silico* digestion, isotopic pattern calculation etc. [4]	Ruby	
OpenMS	library for the analysis, reduction and visualization of LC-MS(/MS) data	C++	
PAPPSO	Plateforme d'Analyse Protéomique de Paris Sud-Ouest	Java	
Proteios	pipeline/LIMS for proteomics experiments and analysis	Java	
Proteomatic	platform for creating MS/MS data analysis workflows using scripts [5]	C++	
ProteoWizard	open source library for proteomics tools development (supports mzML) [6]	C++	
pymzML	Python module to parse mzML data based on cElementTree [7]	Python	
Pyteomics	framework for proteomics data analysis, supporting mzML, MGF, pepXML and more [8]	Python	
QuPE	integrated environment for storage, analysis and integration of proteomics data (requires login) [9]	Java	web
Rproteomics	set of routines for analyzing proteomics data, an XML database to store the results and a user interface	R	
TOPP	the OpenMS protein identification/quantitation pipeline	C++	
TPP	Institute for Systems Biology "Trans-Proteomic Pipeline"		
XCMS	software package (in R) for metabolite profiling from LC-MS data	R	

OPENMS

About

OpenMS is an open-source software C++ library for **LC/MS data management and analyses**. It offers an infrastructure for the rapid development of mass spectrometry related software. OpenMS is free software available under the three clause BSD license and runs under Windows, MacOSX and Linux.

It comes with a vast variety of pre-built and ready-to-use tools for proteomics and metabolomics data analysis (TOPPTools) and powerful 2D and 3D visualization(TOPPView).



OpenMS offers analyses for various quantitation protocols, including **label-free quantitation, SILAC, iTRAQ, SRM, SWATH, ...**

It provides built-in algorithms for **de-novo identification and database search,** as well as adapters to other state-of-the art tools like **XTandem**, **Mascot**, **OMSSA**, etc.

It supports easy integration of OpenMS built tools into workflow engines like **Knime**, **Galaxy**, **WS-Pgrade**, and **TOPPAS** via the TOPPtools concept and a unified parameter handling (CTD).

News

- OpenMS 1.11.1 released
- 7th OpenMS User Meeting – High--performance software for high--throughput proteomics and metabolomics
- OpenMS 1.11 released
- OpenMS 1.10 released
- 6th OpenMS User Meeting – High--performance software for high--throughput proteomics and metabolomics

Get Galaxy-P

The Minnesota Supercomputing Institute provides a public Galaxy-P server capable of limited analyses, testing, and demor

For heavy use you will likely need to install your own instance of Galaxy-P or spin up a Galaxy-P cluster on the cloud.

Public Server

Advantages Immediately accessible. Easiest option for publically sharing data and pages.

Limitations Limited computational and disk resources. Potential problems associated with uploading protected or sensitive data to any public resource.

usegalaxyp.org

Install Your Own

Advantages Full control of computational resources. Easy to modify existing tools or add your own. Use our open source Galaxy-P tools targetting commercial applications that are not available on the public server. Right now these include ProteinPilot and Scaffold.

Limitations Because of its flexiblity, Galaxy can be time-consuming to install and maintain. Galaxy-P adds more tools and can be configured to utilize remote Windows resources adding additional complexity.

Install Instructions

Take to the Clo

Advantages BioCloudCentral provid interface for creating a Galaxy-P clus CloudMan a likewise easy-to-use inte the disk and computational resource P.

Limitations Currently no access to t Windows such as MaxQuant or vend

Launch Now

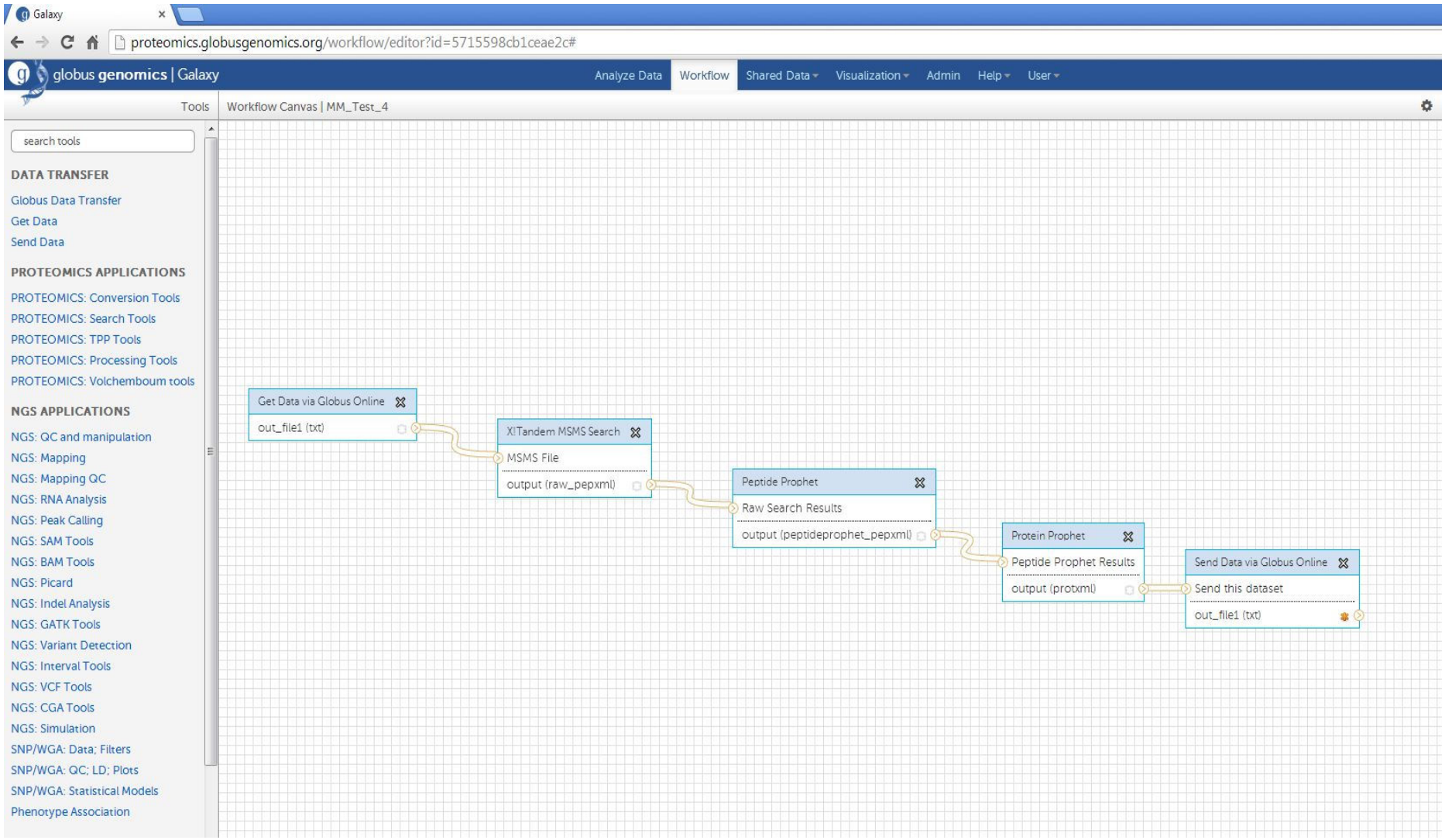
What is Galaxy-P?

Galaxy-P is a multiple 'omics' data analysis platform with particular emphasis on mass

News

For the latest Galaxy-P news, please follow us on Twitter.

proteomics.globusgenomics.org



An example : multiple myeloma

Metric	One fraction (~800 MB)	All 22 fractions (~ 16GB)
Wall Time	19 minutes	24 minutes
Total CPU Time	19 minutes	418 minutes
AWS Nodes Used (m2.4xlarge)	1	3
On-demand Cost ($1.64)	$1.64	$4.92
Spot instance Cost ($0.14 per hour)	$0.14	$0.42*

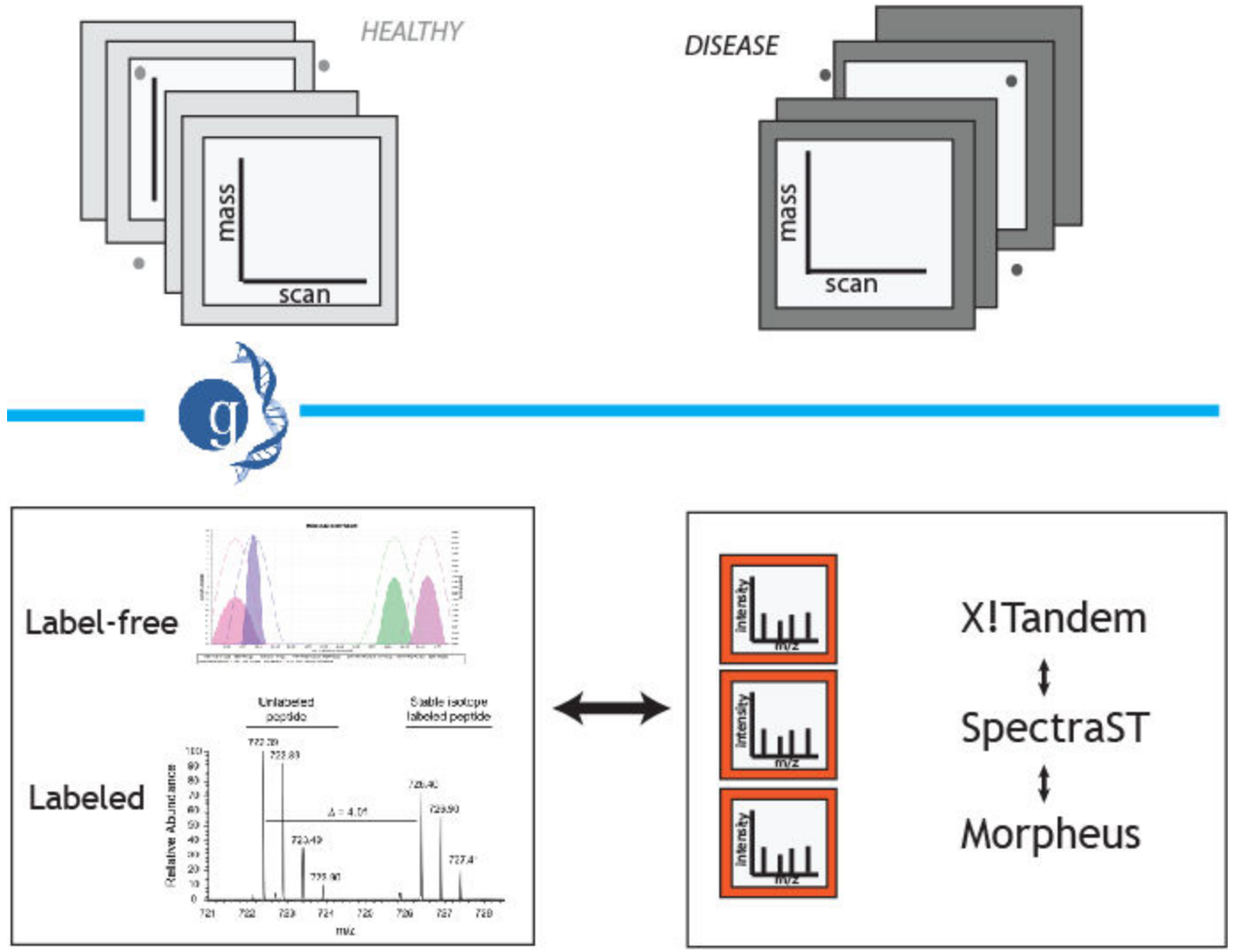
Current pricing for off-campus LC-MS/MS is as follows:

University of Illinois at Chicago (UIC): Orbitrap-Velos ($100/sample run), ($75/hourly)
Mayo Clinic Proteomics Resource Center (MPRC): Orbitrap-LTQ ($100/sample run)
Northwestern University (NU) Proteomics Core: Orbitrap-Velos ($140/sample run)
Northwestern University (NU) Proteomics Center of Excellence (PCE): Velos Orbitrap Elite with ETD, Velos-FT12T Ultra.

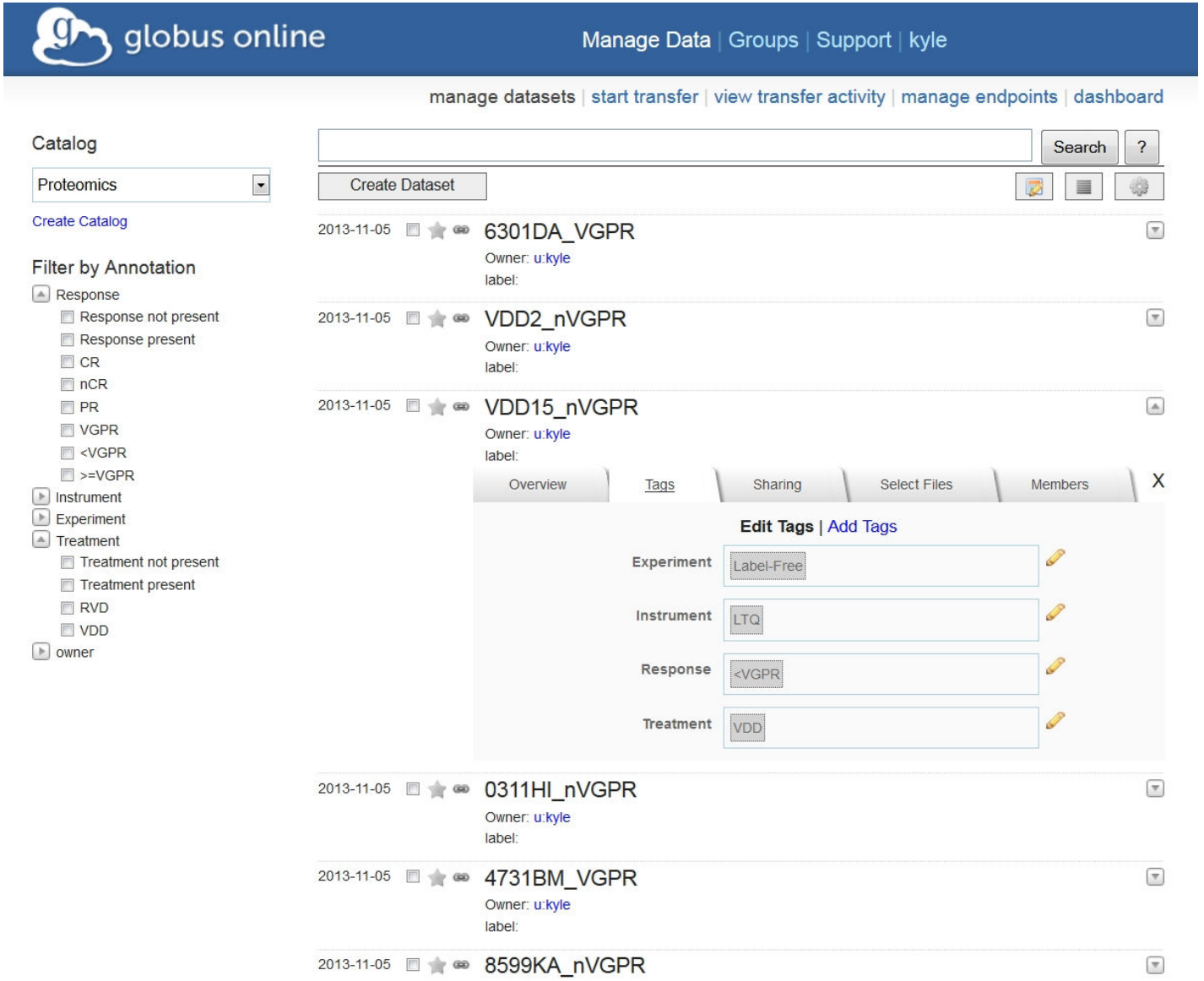
Future work

Functional category	Tools currently available	Tools to be incorporated in future
Dataset Tools	uf-mzML, , MGF-DTA file converters	msconvert Decon2LS (console version), DeconMSn, MultiAlign(command line), SuperHirn
Search Tools	X! Tandem, OMSSA, Mascot, MSGF+	SpectraST, Morpheus (command line), MyriMatch
Identification Tools	PeptideProphet, ProteinProphet , iProphet, Validator-MAX	MaxQuant, IDPicker3
Quantification Tools	Quantifier, XPRESS, Libra, ASAPRatio	MaxQuant, IDPQuantify
Other	PepXML/ProtXML to Table	Tools for: processing proteomics catalogs, combining outputs from different workflows, intelligent inclusion predictors.

Workflow integration



Globus Proteomics catalog



Acknowledgements

- Sam Volchenboum
- Kolbrun Kristjansdottir
- Don Wolfgeher
- Alex Rodriguez
- Kyle Chard
- Ravi Madduri
- Paul Dave