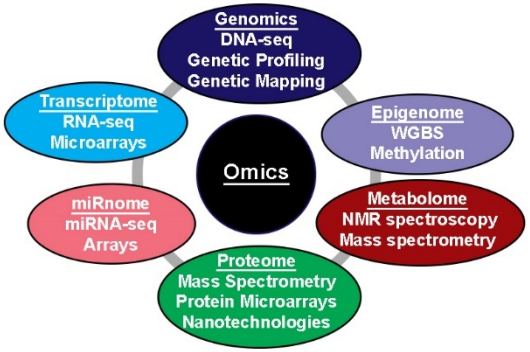
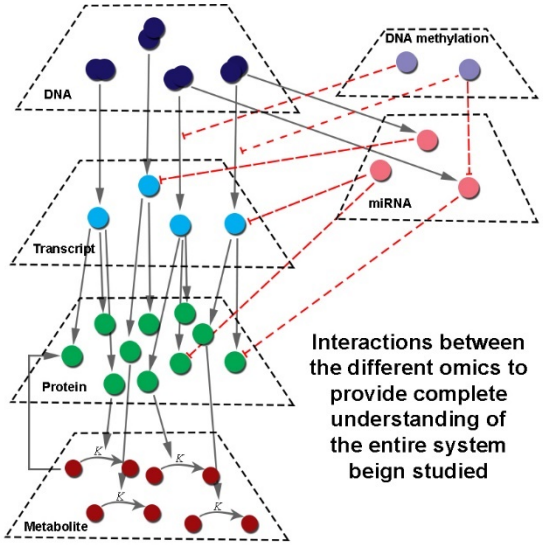
Transcriptomics and Proteomics Discussion

Afshin Beheshti, PhD
Wyle Labs, Space Biosciences Division
NASA Ames Research Center, Moffett Field, CA
afshin.beheshti@nasa.gov

GeneLab
Open Science for Exploration

Omics Available on GeneLab and Challenges to Overcome





Interactions between the different omics to provide complete understanding of the entire system beign studied

Public Databases
GeneLab
TCGA
etc

Use Public databases with mulitple omics datasets to determine overall behavior of the system

Challenges to multi-omics datasets

- **Different Omics Platforms**
  Integration of Data
- **Computing Power**
  Lack of Computer Clusters
  Time to Preprocess Data
- **Datasets**
  Poor Controls
  Bad biological replicates
  Bad experimental setup
- **Analysis Methods**
  Normalization
  Biased analysis
  No uniform data processing

Possible Solutions

- Develop Software Tools to integrate different platforms
- Supercomputer clusters made available on the cloud to process data
- From analysis done on public data demonstrate how to setup proper future experiments
- Create universal pipeline for unbiased analysis of all omics data

GeneLab
Open Science for Exploration

Present Pipelines and Discussion

- Goal of this session:
  - Provide a consensus on the pipelines for the different omics to implement on GeneLab
  - Discuss issues that might occur with the consensus pipelines
  - Other options to the pipelines presented

GeneLab
Open Science for Exploration

Transcriptomic Pipelines: Microarray

Pre-processing

Background Subtraction and Normalization

Limma
Limma-voom
RMA (for Affymetrix)
GCRMA (for Affymetrix)
Oligo (for newer Exon Arrays)
PIMENTo (for Illumina)
GenePattern

Higher-Order Analysis

Differential Expression Analysis

Limma
Limma-voom
MeV

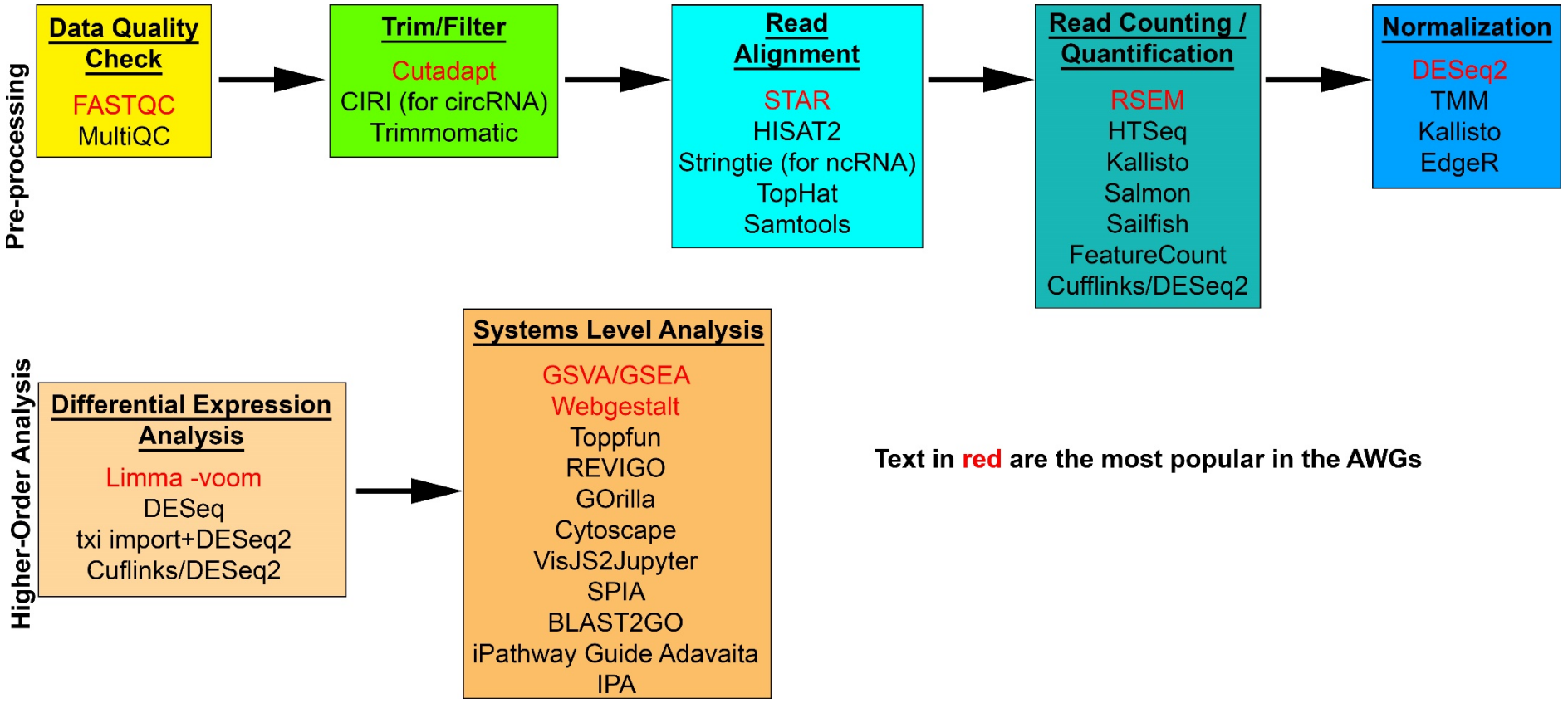
Systems Level Analysis

GSVA/GSEA
Webgestalt
Toppfun
REVIGO
GOrilla
Cytoscape
VisJS2Jupyter
SPIA
BLAST2GO
iPathway Guide Adavaita
IPA

Text in red are the most popular in the AWGs

GeneLab
Open Science for Exploration

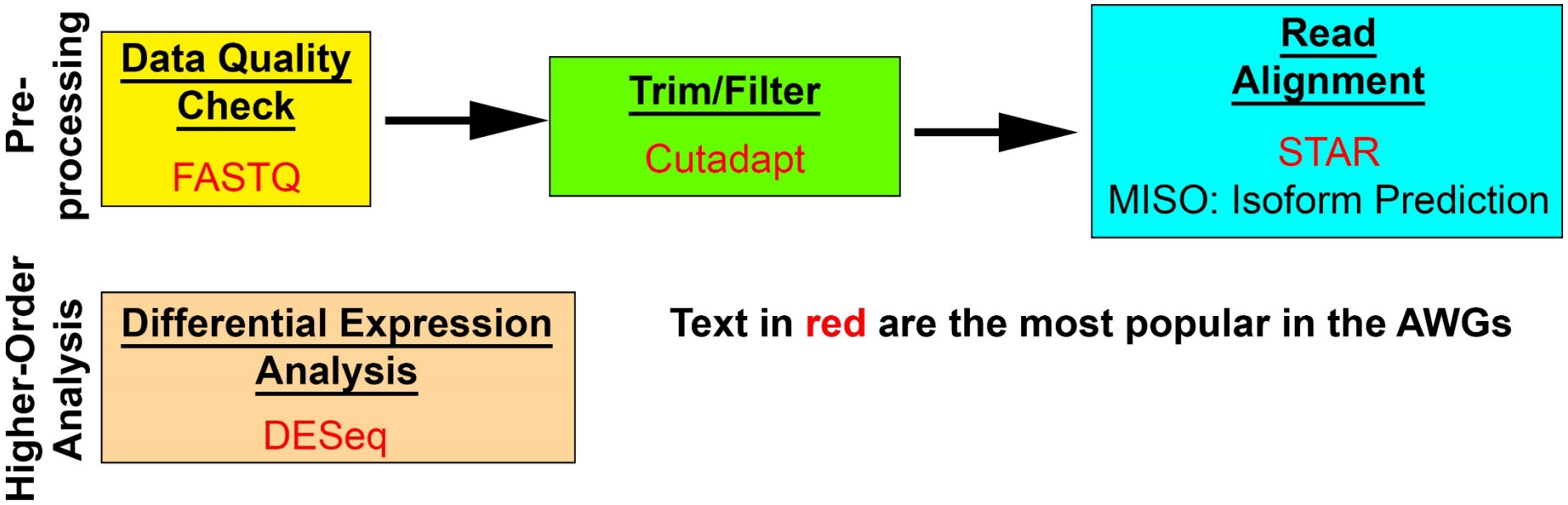
Transcriptomic Pipelines: RNA-Seq



Text in red are the most popular in the AWGs

GeneLab
Open Science for Exploration

Transcriptomic Pipelines: RNA-seq differential splicing



Text in red are the most popular in the AWGs

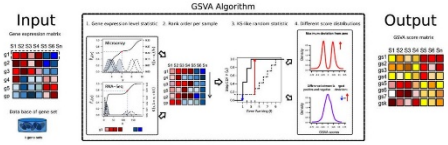
GeneLab
Open Science for Exploration

Transcriptomic Pipelines: Systems Level Analysis

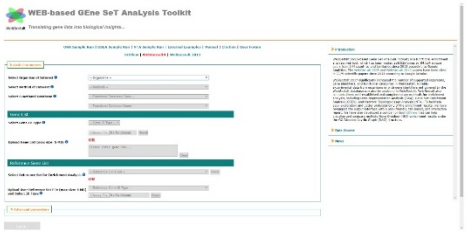
Systems Level Analysis and Visualization Tools

Top Picks from AWG

GSVA/GSEA



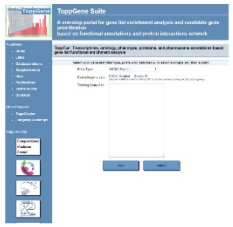
Webgestalt



Other Open Source Tools

REVIGO

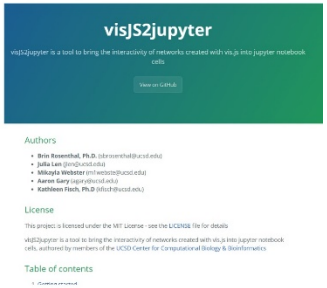
Toppfun



GOrilla

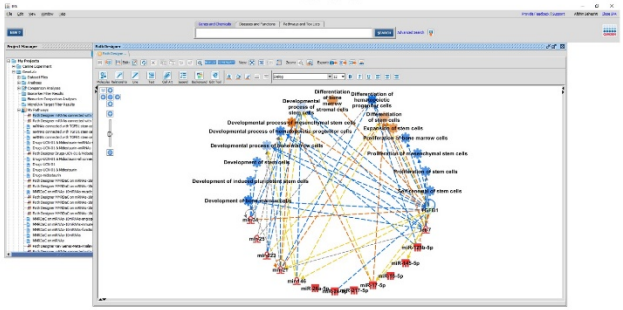
Cytoscape

VisJS2Jupyter

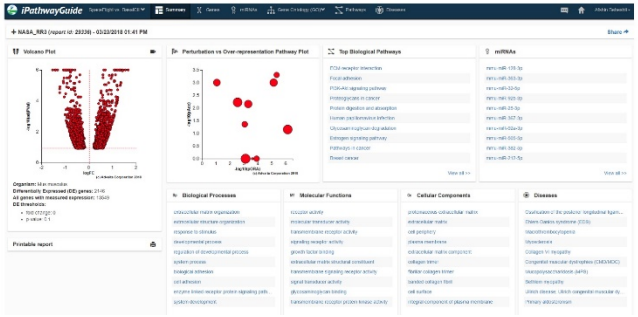


Proprietary Software

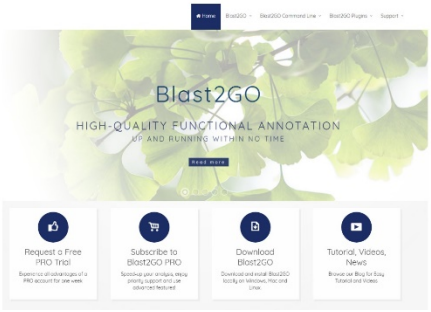
IPA



iPathwayGuide

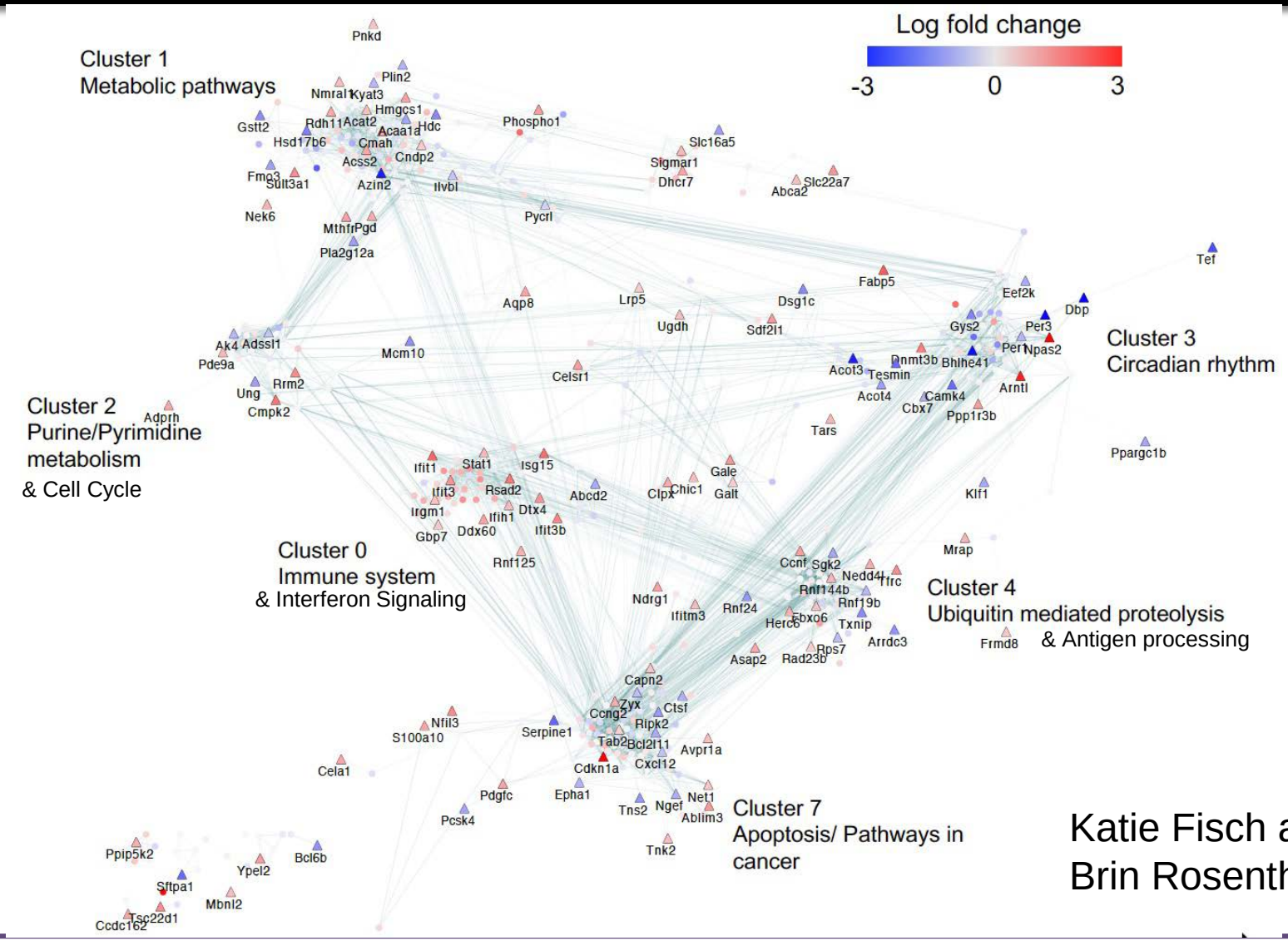


Blast2Go



GeneLab
Open Science for Exploration

Examples of Transcriptomic Analysis

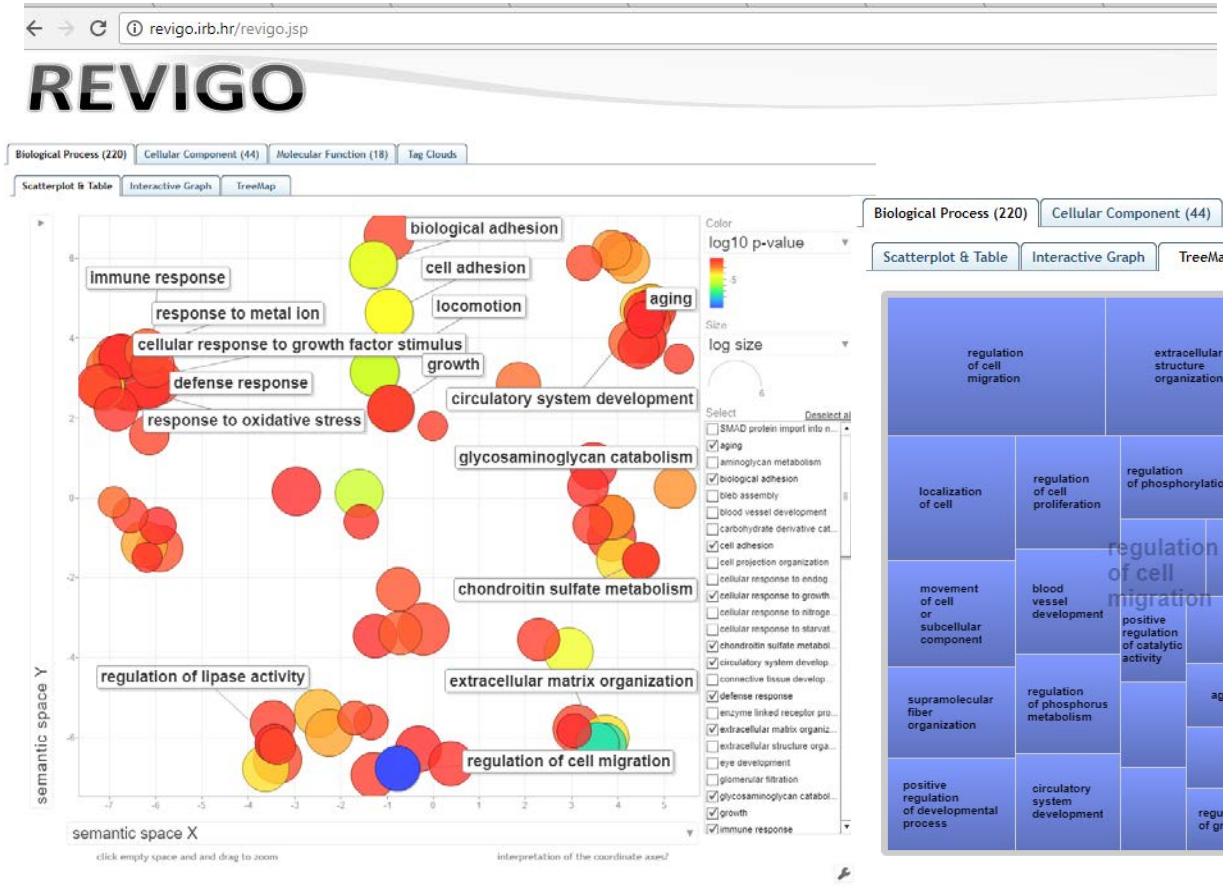


Katie Fisch and
Brin Rosenthal

GeneLab
Open Science for Exploration

Examples of Transcriptomic Analysis

RR3 – Over-representation analysis - 549 unique genes (FLT vs BC)

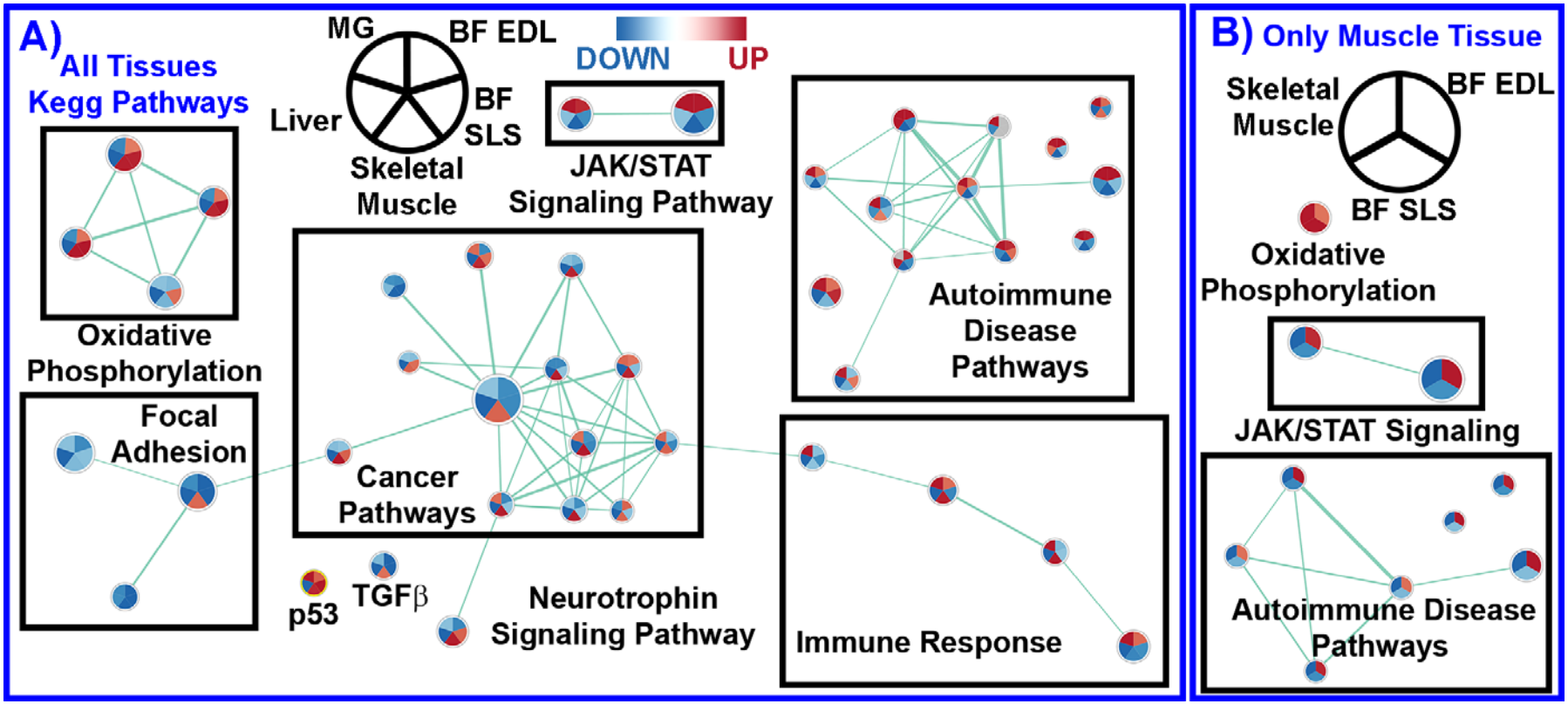


Willian Abraham da Silveira

GeneLab
Open Science for Exploration

Examples of Transcriptomic Analysis

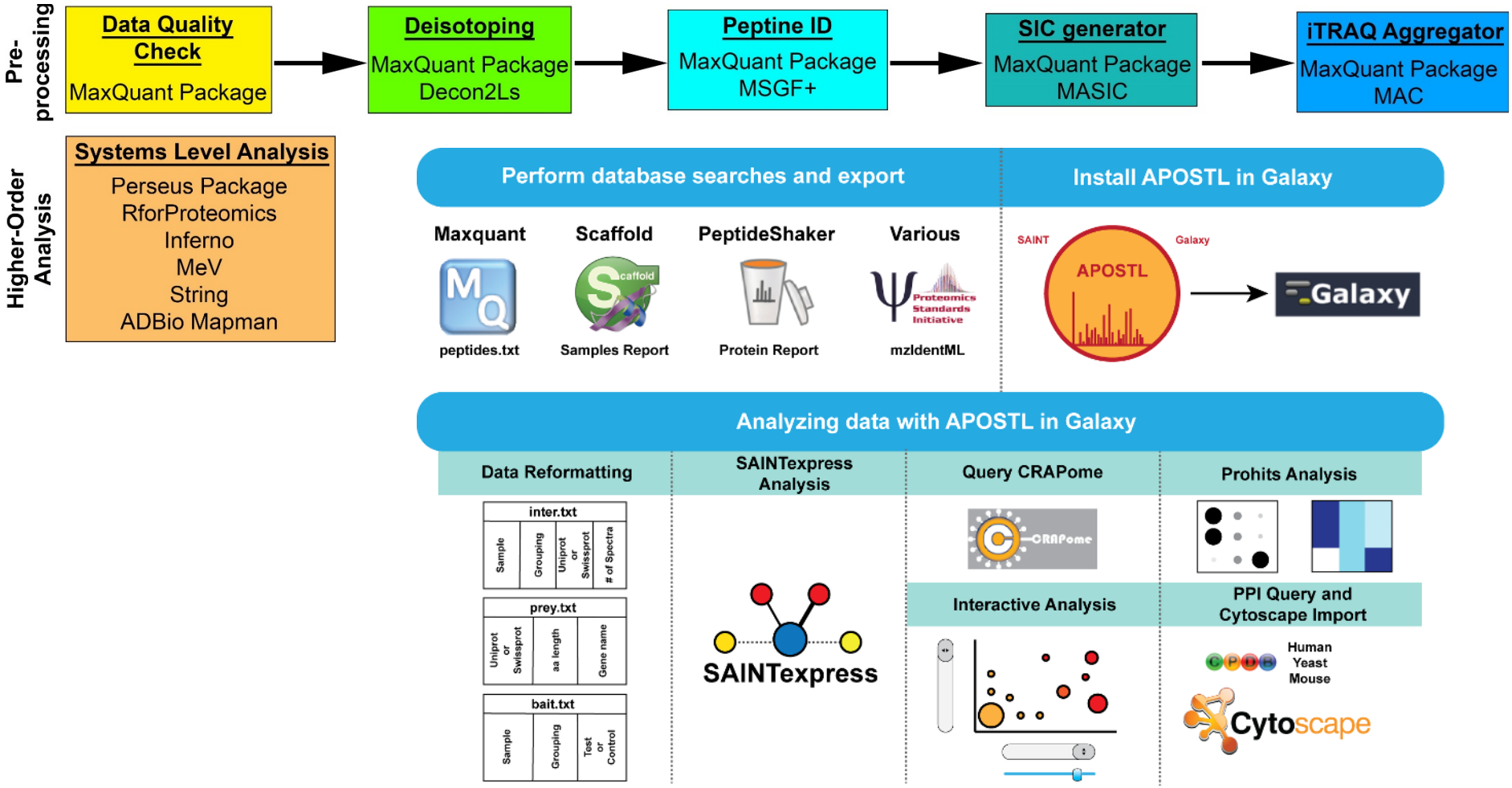
GSEA analysis in a network representation with a Cytoscape plugin



Afshin Beheshti

GeneLab
Open Science for Exploration

Proteomic Pipelines



https://github.com/bornea/APOSTL/blob/master/Readme.md

Proteomic Pipelines

Byonic		Database search program released in 2011 by Protein Metrics Inc. with original developments at PARC[1] that searches MS/MS data from all types of instruments and internally employs the program Combyne,[2] which combines peptide identifications to produce protein scores and identification probabilities.	server
Greylag	open source	Database search program developed at the Stowers Institute for Medical Research designed to perform large searches on computational clusters having hundreds of nodes.	server
InsPecT		A MS-alignment search engine available at the Center for Computational Mass Spectrometry at the University of California, San Diego[3]	download
Mascot	proprietary	Performs mass spectrometry data analysis through a statistical evaluation of matches between observed and projected peptide fragments.[4]	
MassMatrix		MassMatrix is a database search software package for tandem mass spectrometric data. It uses a mass accuracy sensitive probabilistic scoring model to rank peptide and protein matches.	download
MassWiz	open source	Search algorithm developed at Institute of Genomics and Integrative Biology available as a windows commandline tool.	download, server
MyriMatch	open source	Database search program developed at the Vanderbilt University Medical Center designed to run in a single-computer environment or across an entire cluster of processing nodes.[5]	download
OMSSA	open source	database search program developed at the National Center for Biotechnology Information.[6][7]	
PEAKS DB	proprietary	Database search engine, run in parallel with de novo sequencing to automatically validate search results, allowing for a higher number of found sequences for a given false discovery rate. In addition to providing an independent database search, results can be incorporated as part of the software's multi-engine (Sequest, Mascot, X!Tandem, OMSSA, PEAKS DB) consensus reporting tool, inChorus.[8] The tool also provides a list of sequences identified exclusively by de novo sequencing.	
Phenyx		Developed by Geneva Bioinformatics (GeneBio) in collaboration with the Swiss Institute of Bioinformatics (SIB). Phenyx incorporates OLAV, a family of statistical scoring models, to generate and optimize scoring schemes that can be tailored for all kinds of instruments, instrumental set-ups and general sample treatments.[9]	
ProteinPilot Software		Uses Paragon database search algorithm that combines the generation of short sequence tags ('taglets') for computation of sequence temperature values and estimates of feature probabilities to enable the peptide identification considering hundreds of modifications, non-tryptic cleavages and amino acid substitutions. Uses the Pro Group Algorithm for protein inference analysis to report the minimal set of proteins justified based on the peptide evidence. Supports quantification for label-based workflows (iTRAQ reagents, mTRAQ reagents and SILAC labeling). A translation layer translates user interface controls in the language of the proteomics experimental scientist to underlying complex informatics parameters.[10] pate/content/6/9/1638.short	download
Protein Prospector	open source	Protein Prospector is a package of about twenty proteomic analysis tools developed at the University of California San Francisco. The tandem mass spectrometry searching software is Batch-Tag / Batch-Tag Web, with the results processed and displayed using Search Compare. It uses scoring systems tailored to instrument and fragmentation mode to optimize analysis of different types of fragmentation data.	server, source
RAId		Developed at the National Center for Biotechnology Information, Robust Accurate Identification (RAId)[11] is a suite of proteomics tools for analyzing tandem mass spectrometry data with accurate statistics.[12]	
SEQUEST	proprietary	Identifies collections of tandem mass spectra to peptide sequences that have been generated from databases of protein sequences.[13]	
SIMS		SIMS (Sequential Interval Motif Search) is a software tool design to perform unrestrictive PTM search over tandem mass spectra; users do not have to characterize the potential PTMs. Instead, users only need to specify the range of modification mass for each individual amino acid.	server
SimTandem	freeware	A database search engine for identification of peptide sequences from LC/MS/MS data; the engine can be used as an external tool in OpenMS/TOPP.[14]	download
SQID	open source	SeQuence IDentification (SQID) is an intensity-incorporated protein identification algorithm for tandem mass spectrometry.	download
X!Tandem	open source	Matches tandem mass spectra with peptide sequences.	server
pFind		pFind Studio is a computational solution for mass spectrometry-based proteomics, it germinated in 2002 in Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.	download

Proteomic Pipelines

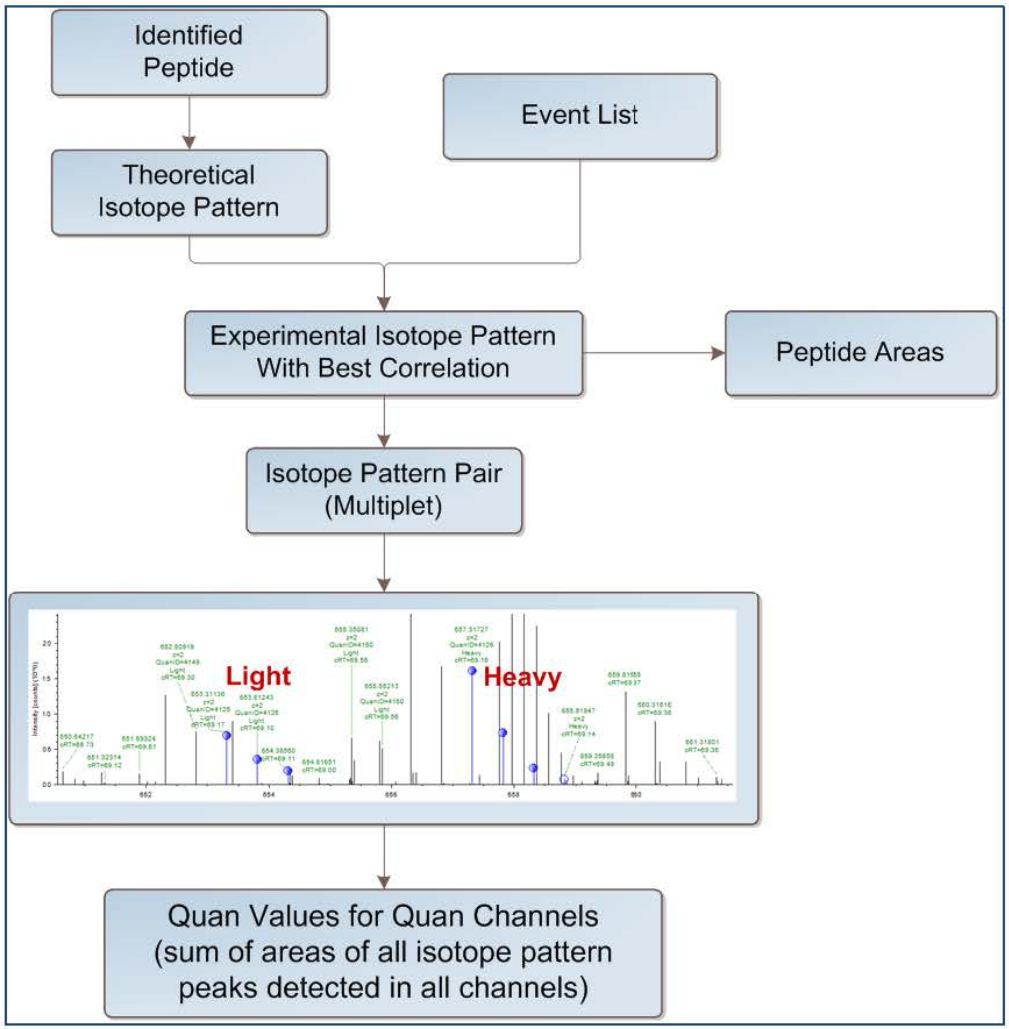
Software solutions that Sucharita Dutta Utilizes: **Proteome Discoverer and Pinnacle** Proteome Discoverer--Identify

- **Workflow-based system for proteomics "deep sequencing"**
- **Key Features**
  - Workflow editor enables highly flexible searches
  - Comprehensive set of tools for data mining of CID, HCD and ETD Spectra
  - Easily searches complex experiments produced by LTQ Velos Orbitrap systems (e.g. data dependent decision tree, HCD-triggered ETD for glycan analysis, etc.)
  - SEQUEST and Mascot
  - Automated calculation of false discovery rate
  - Labeled quantification (SILAC, TMT)
  - Node SDK (in development) enables 3rd parties develop nodes

GeneLab
Open Science for Exploration

Proteomic Pipelines

Workflow for performing data analysis on Genelab datasets



Questions and Discussion about GeneLab and AWG

Questions and Discussion???