



WORKFLOWS TO FOLLOW UP ON RNA-SEQ ANALYSIS

HELENE R. MCMURRAY, PH.D.

EDWARD G. MINER LIBRARY

7 NOVEMBER 2016

TOPICS TO BE COVERED TODAY

- High-throughput sequencing basics
- Pattern recognition by hierarchical clustering
- Gene Ontology mapping
- Canonical pathway mapping

NEXT-GENERATION GAP

JOHN D. MCPHERSON
NATURE METHODS 6, S2 - S5
(2009)

[http://www.urmc.rochester.edu/libraries/
miner/mdl.aspx?U=http://dx.doi.org/doi:
10.1038/nmeth.f.268](http://www.urmc.rochester.edu/libraries/miner/mdl.aspx?U=http://dx.doi.org/doi:10.1038/nmeth.f.268)

- Discusses high-throughput sequencing platforms with references to more information on each technology
 - Roche 454
 - Illumina GAllx
 - Applied Biosystems SOLID
 - Helicos HeliScope

SENSE FROM SEQUENCE READS: METHODS FOR ALIGNMENT AND ASSEMBLY

PAUL FLICEK & EWAN BIRNEY
NATURE METHODS **6**, S6 - S12 (2009)

[http://www.urmc.rochester.edu/libraries/
miner/mdl.aspx?U=http://dx.doi.org/doi:
10.1038/nmeth.1376](http://www.urmc.rochester.edu/libraries/miner/mdl.aspx?U=http://dx.doi.org/doi:10.1038/nmeth.1376)

- Discusses alignment methods
 - Hash-based methods
 - Burrows-Wheeler transform methods
- Explains assembly based on those alignments

COMPUTATION FOR CHIP-SEQ AND RNA- SEQ STUDIES

SHIRLEY PEPKE, BARBARA WOLD &
ALI MORTAZAVI
NATURE METHODS 6, S22 - S32
(2009)

[http://www.urmc.rochester.edu/libraries/
miner/mdl.aspx?U=http://dx.doi.org/doi:
10.1038/nmeth.1371](http://www.urmc.rochester.edu/libraries/miner/mdl.aspx?U=http://dx.doi.org/doi:10.1038/nmeth.1371)

- Provides overview of ChIP-seq and RNA-seq analyses
- ChIP-seq:
 - Classes of ChIP-seq signals
 - Peak-finders, regions, summits and sources
 - Publicly available ChIP-seq software
- RNA-seq:
 - Approaches to handle spliced reads
 - Quantifying gene expression
 - Publicly available RNA-seq software

VISUALIZATION OF OMICS DATA FOR SYSTEMS BIOLOGY

N. GEHLENBORG, S.I.

O'DONOGHUE, N.S. BALIGA, A.

GOESMANN, M.A. HIBBS, H.

KITANO, O. KOHLBACHER, H.

NEUWEGER, R. SCHNEIDER, D.

TENENBAUM & A. GAVIN

NATURE METHODS 7, S56 - S68
(2010)

<http://www.urmc.rochester.edu/libraries/miner/mdl.aspx?U=http://dx.doi.org/doi:10.1038/nmeth.1436>

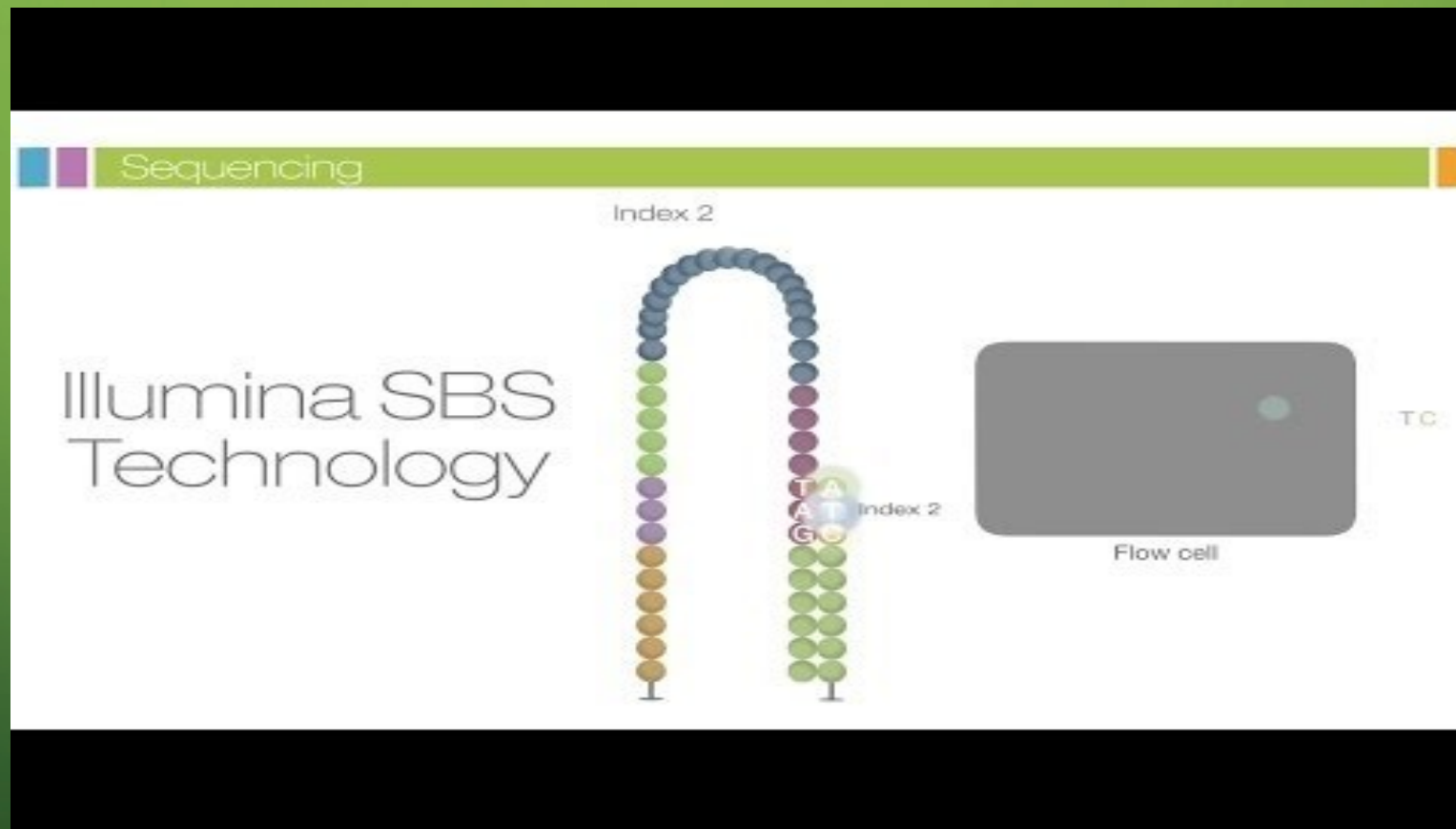
- Visualization based on networks vs. pathways

- Protein interaction networks
- Spatial information
- Expression profile data
- Multivariate -omics analysis
- Metabolic network visualization
- Pathway editing

The background is a solid green gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles connecting them.

HIGH-THROUGHPUT SEQUENCING WORKFLOW

HOW DOES ILLUMINA SEQUENCING BY SYNTHESIS WORK?



BASIC HIGH-THROUGHPUT SEQUENCING WORKFLOW: UPSTREAM

[http://en.wikipedia.org/wiki/
List_of_sequence_alignment_software](http://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Sample preparation

Genomic RNA

Library construction

Sequencing

Illumina, Roche (454), SOLiD, etc

Raw reads

FASTQ or color space FASTQ

Quality control & filtering

HTSeq etc

Alignment

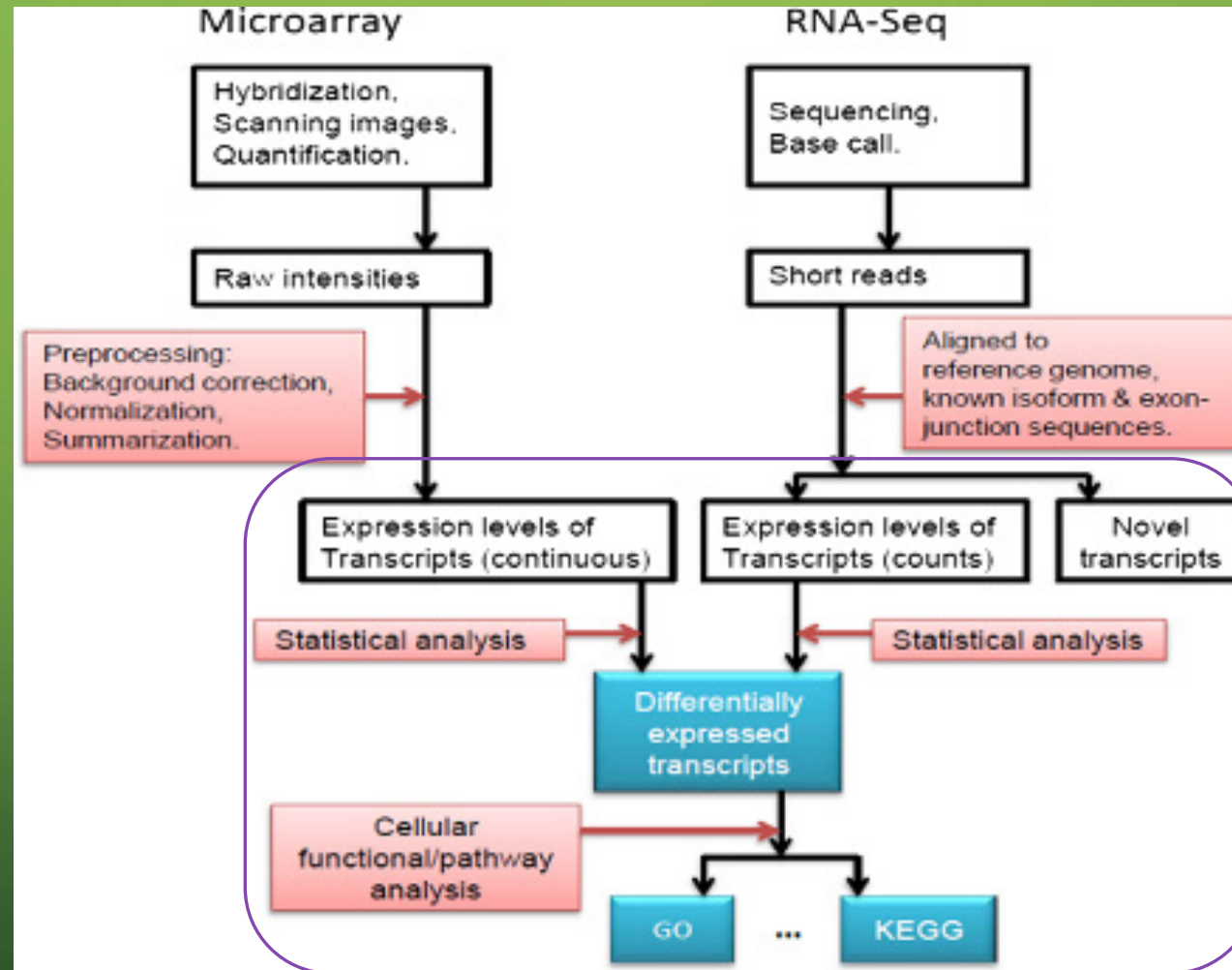
BWA, TopHat, etc

Alignment statistics & filtering

Mapped reads

SAM or BAM format

BASIC HIGH-THROUGHPUT SEQUENCING WORKFLOW: DOWNSTREAM



The screenshot displays the Galaxy web interface. At the top, a navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A status message indicates intermittent unavailability on Saturday, Sept 20th. The left sidebar, titled 'Tools', contains a search bar and a list of tool categories: Get Data, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, and Genome Diversity. The main content area features an announcement: 'Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).' Below this is a graphic stating 'Galaxy-User is now Biostars' with a globe icon and 'GALAXY EXPLAINED'. To the right of the graphic is a 'Tweets' section showing three tweets from the Galaxy Project, OpenHelix Staff, and Hans-Rudolf Hotz. The far right panel, titled 'History', shows 'Unnamed history' with 0 bytes and a message: 'This history is empty. You can [load your own data](#) or [get data from an external source](#)'.

GALAXY SUITE OF TOOLS

<http://galaxyproject.org/>

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.

Bioconductor uses the R statistical programming language, and is open source and open development. It

Install »

Get started with *Bioconductor*

- [Install *Bioconductor*](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)

BIOCONDUCTOR AND R

<https://www.bioconductor.org/>

<https://www.r-project.org/>

The background is a green gradient. In the corners, there are decorative circuit-like patterns made of thin white lines and small circles, resembling a stylized PCB or data network.

IDENTIFICATION OF PATTERNS IN THE DATA

COMMON WAYS TO MAKE SENSE OF –OMICS DATA

- Identification of patterns in the data

- Advantage: Unbiased; Disadvantage: Can be hard to interpret, easy to over-interpret
- Hierarchical clustering
- Principle Components Analysis
- Other graphical representations

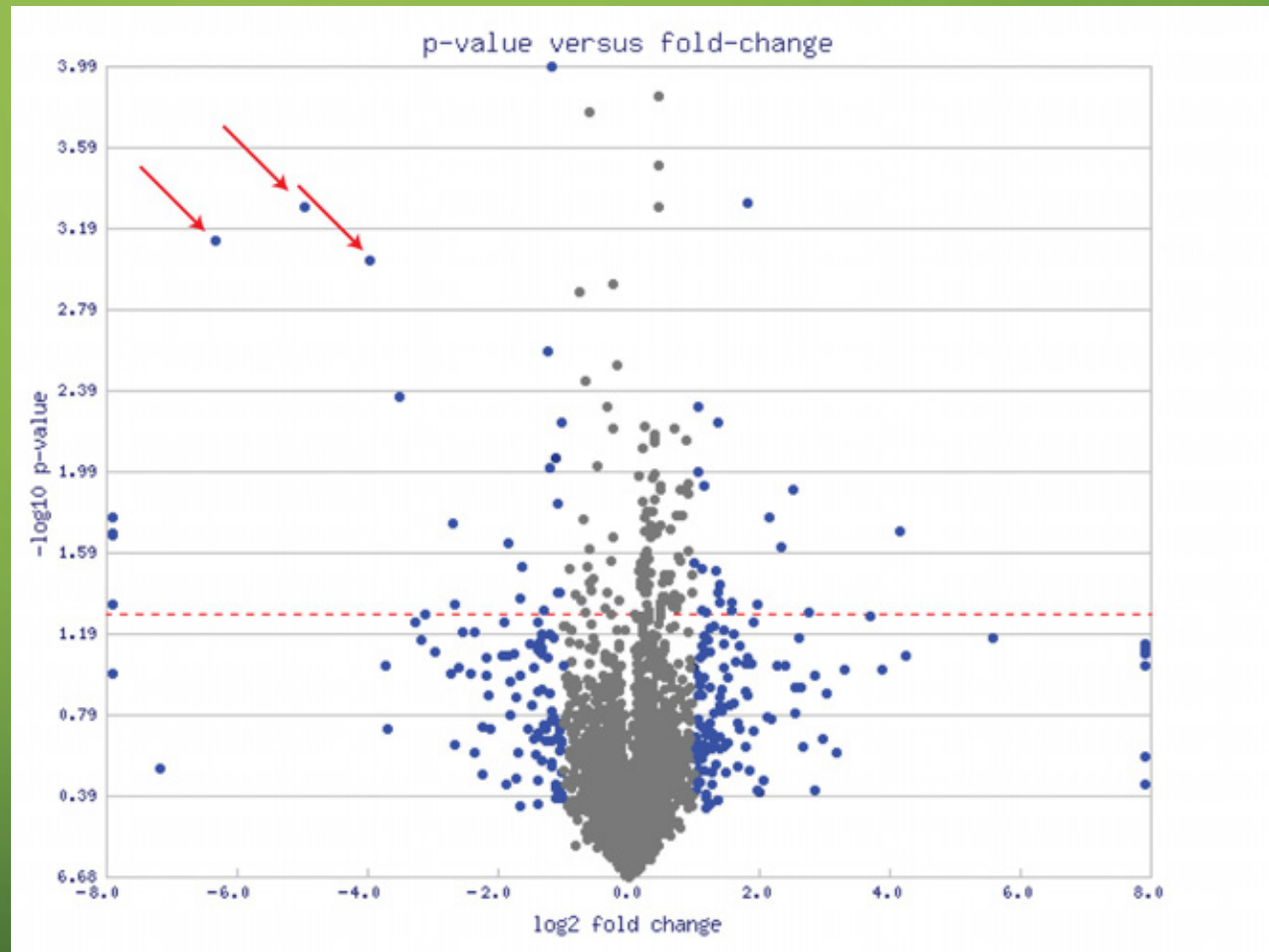
- Mapping new information on to prior knowledge

- Gene Ontology mapping
- Canonical pathway mapping
- Set-level analysis (GSEA, e.g.)
- Protein interaction analysis
- Common functions of compounds

VOLCANO PLOT

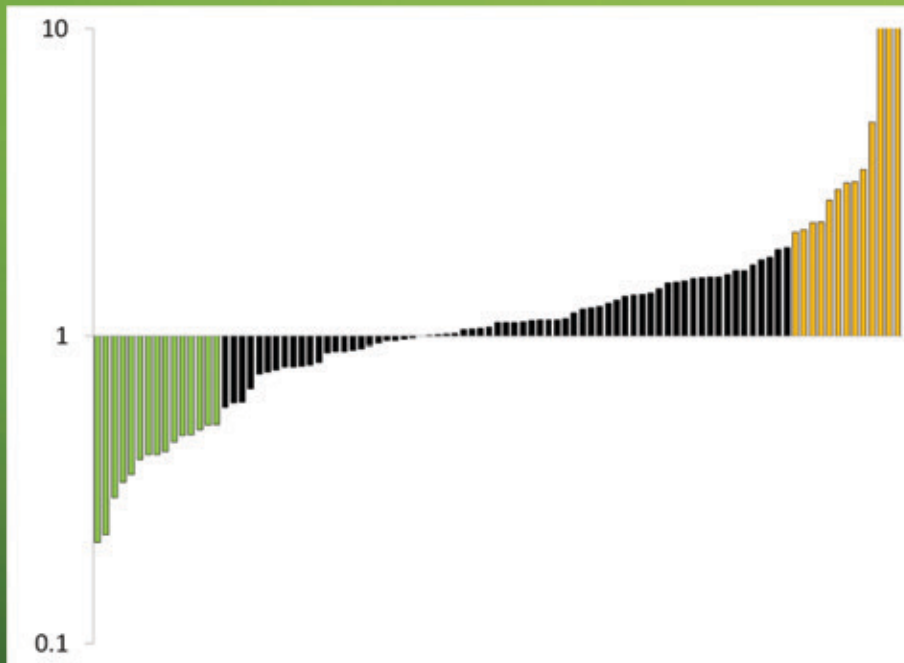
By Roadnottaken - Own work, Public Domain,

<https://commons.wikimedia.org/wiki/index.php?curid=8901192>

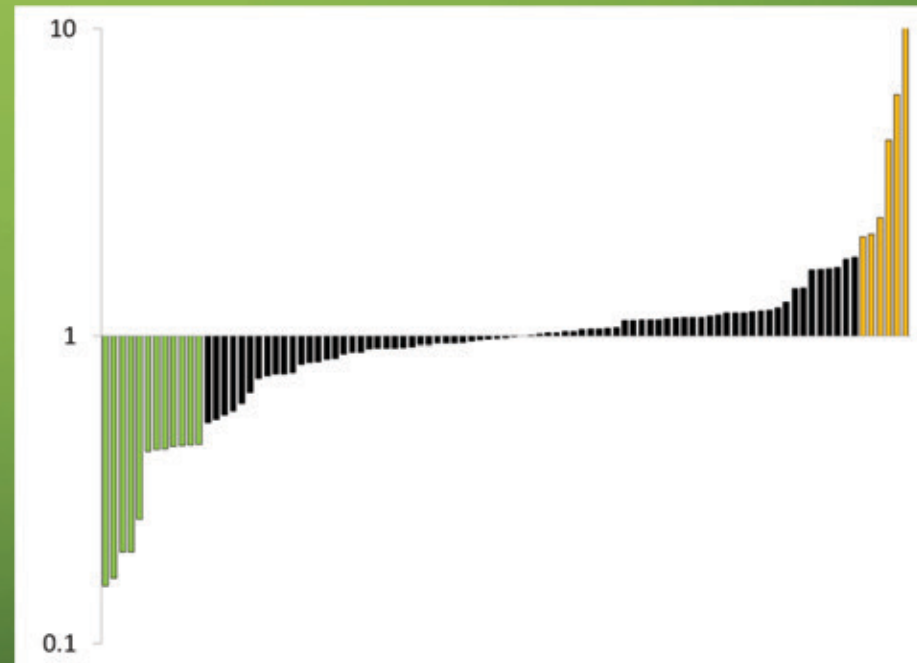


WATERFALL PLOTS

Cell Type 1



Cell Type 2



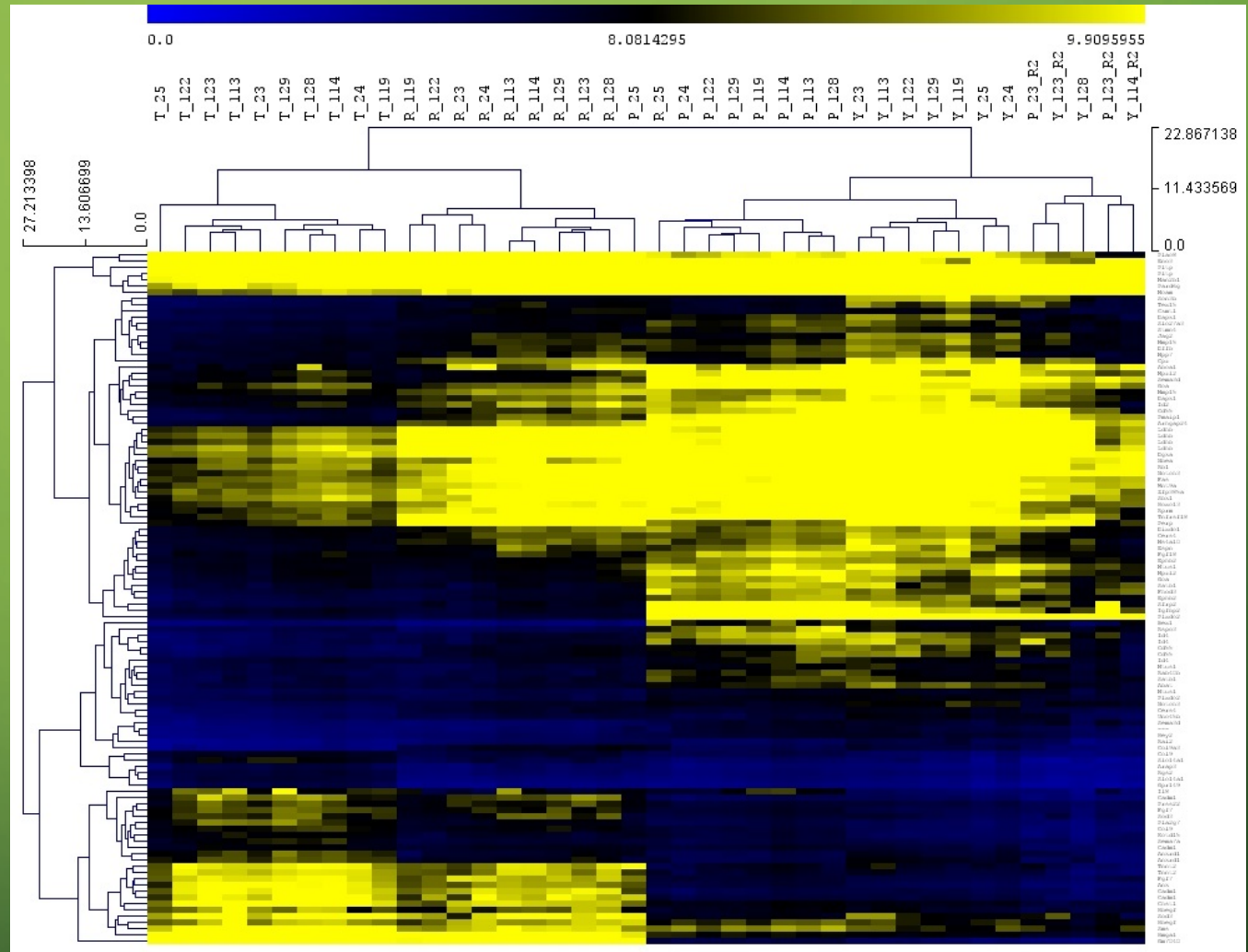
Gene order on these graphs is different!

HIERARCHICAL CLUSTERING

- Hierarchical clustering relates the samples or elements to one another in hierarchical fashion
 - Based on various metrics (correlation, ranking or distance between samples)
- There are other types of clustering, used less commonly but may be important in your project
 - k-means clustering, self-organizing maps (SOM), etc.

HIERARCHICAL CLUSTERING

Clustering and visualization via
TM4 MeV software, Version 4.9.0



New in JMP® Pro 12

As the advanced analytics version of our software, JMP Pro contains everything users know and love about JMP – and more. With the release of this latest version – JMP Pro 12 – users will enjoy new capabilities and see a leap in performance in almost all of its earlier platforms.

➤ [Learn how to get JMP Pro 12](#)



JMP: BETTER STATISTICAL ANALYSIS

http://www.jmp.com/en_us/home.html

<http://tech.rochester.edu/services/software-site-licensing/>



Open your own file

Choose a file...

GCT 1.3, GCT 1.2, MAF, GMT, a tab-delimited text file, or an Excel spreadsheet
All data is processed in the browser and never sent to any server

Contact
Linking
Tutorial
Source Code

Or select a preloaded dataset

Name	Gene Expression	Copy Number By Gene	Mutations	Gene Essentiality	
Cancer Cell Line Encyclopedia (CCLE), Project Achilles	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Open

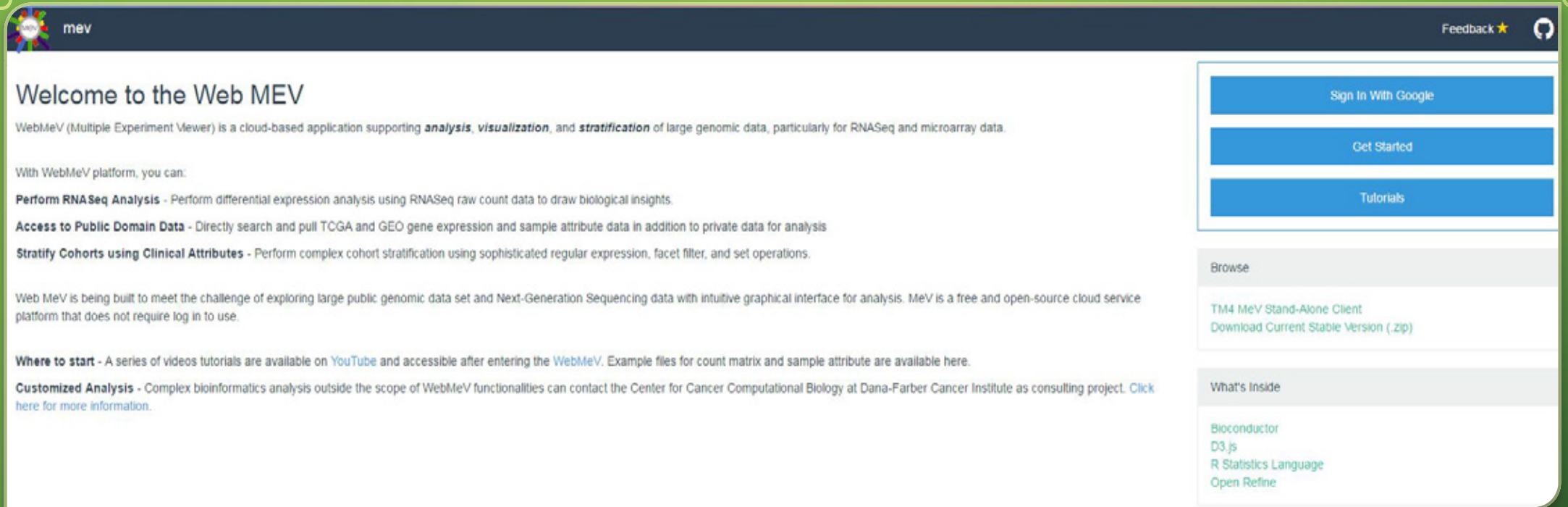
TCGA data version 1/11/2015

Please adhere to the [TCGA publication guidelines](#) when using TCGA data in your publications.

Disease	Gene Expression	GISTIC Copy Number	Copy Number By Gene	Mutations	Proteomics	Methylation
---------	-----------------	--------------------	---------------------	-----------	------------	-------------

MORPHEUS

<https://software.broadinstitute.org/morpheus/>



mev Feedback ★ ↻

Welcome to the Web MEV

WebMeV (Multiple Experiment Viewer) is a cloud-based application supporting **analysis**, **visualization**, and **stratification** of large genomic data, particularly for RNASeq and microarray data.

With WebMeV platform, you can:

- Perform RNASeq Analysis** - Perform differential expression analysis using RNASeq raw count data to draw biological insights.
- Access to Public Domain Data** - Directly search and pull TCGA and GEO gene expression and sample attribute data in addition to private data for analysis
- Stratify Cohorts using Clinical Attributes** - Perform complex cohort stratification using sophisticated regular expression, facet filter, and set operations.

Web MeV is being built to meet the challenge of exploring large public genomic data set and Next-Generation Sequencing data with intuitive graphical interface for analysis. MeV is a free and open-source cloud service platform that does not require log in to use.

Where to start - A series of videos tutorials are available on [YouTube](#) and accessible after entering the [WebMeV](#). Example files for count matrix and sample attribute are available here.

Customized Analysis - Complex bioinformatics analysis outside the scope of WebMeV functionalities can contact the Center for Cancer Computational Biology at Dana-Farber Cancer Institute as consulting project. [Click here for more information.](#)

Sign In With Google

Get Started

Tutorials

Browse

- TM4 MeV Stand-Alone Client
- Download Current Stable Version (.zip)

What's Inside

- Bioconductor
- D3.js
- R Statistics Language
- Open Refine

MULTI-EXPERIMENT VIEWER (MEV)

<http://www.tm4.org/#/welcome>

The background is a solid green gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural network connections. These elements consist of thin lines that branch out and terminate in small circles.

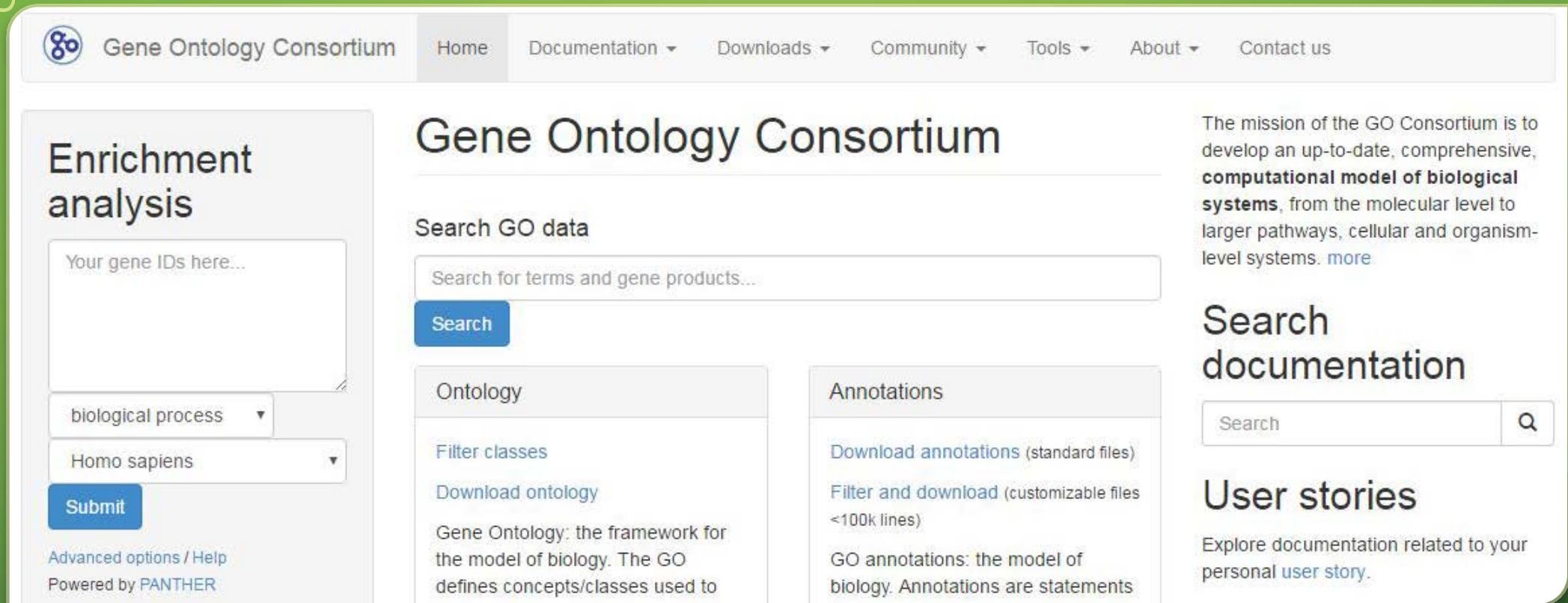
MAPPING NEW INFORMATION ONTO PRIOR KNOWLEDGE

SET-LEVEL / ENRICHMENT ANALYSIS

- Tests for over-representation of elements of a given set in a larger set
 - Gene ontology analysis
 - Hierarchical system of categorizing genes
 - Controlled vocabulary
 - Assess over- or under-represented ontology terms
 - Pathway analysis
 - Allows integration of novel experimental data with known biological function
 - Relies on accurate mapping of pathways in public domain
 - Assumes that pathways are the same in all contexts



GENE ONTOLOGY ANALYSIS



The screenshot shows the Gene Ontology Consortium website. The header includes the logo and navigation links: Home, Documentation, Downloads, Community, Tools, About, and Contact us. The main content area is divided into three columns. The left column features an 'Enrichment analysis' section with a text input for gene IDs, dropdowns for 'biological process' and 'Homo sapiens', and a 'Submit' button. The middle column has a 'Search GO data' section with a search bar and a 'Search' button, followed by an 'Ontology' section with links to 'Filter classes' and 'Download ontology', and a brief description of the GO framework. The right column contains a mission statement, a 'Search documentation' section with a search bar, and a 'User stories' section. The website is powered by PANTHER.

Gene Ontology Consortium

Home Documentation Downloads Community Tools About Contact us

Enrichment analysis

Your gene IDs here...

biological process

Homo sapiens

Submit

[Advanced options / Help](#)
Powered by [PANTHER](#)

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Search

Ontology

[Filter classes](#)

[Download ontology](#)

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to

Annotations

[Download annotations](#) (standard files)

[Filter and download](#) (customizable files <100k lines)

GO annotations: the model of biology. Annotations are statements

The mission of the GO Consortium is to develop an up-to-date, comprehensive, **computational model of biological systems**, from the molecular level to larger pathways, cellular and organism-level systems. [more](#)

Search documentation

Search

User stories

Explore documentation related to your personal [user story](#).


THE GENE ONTOLOGY CONSORTIUM

<http://geneontology.org/>

The screenshot shows the PANTHER Classification System website. At the top left is the PANTHER logo with a panther silhouette and the text "PANTHER Classification System". To the right are links for "LOGIN", "REGISTER", and "CONTACT US". Below this is a navigation bar with links: "Home", "About", "PANTHER Data", "PANTHER Tools", "Workspace", "Downloads", and "Help/Tutorial". A banner below the navigation bar says "PANTHER 9.0 released" with a link "Click to view details". On the left side, there is a "Search" section with a dropdown menu set to "All" and a "Go" button. Below the search section is a "Quick links" section with links to "Whole genome function views", "Genome statistics", "How to cite PANTHER", and "NEW! Recent publication describing PANTHER". The main content area has tabs for "Gene List Analysis", "Browse", "Sequence Search", "cSNP Scoring", and "Keyword Search". The "Gene List Analysis" tab is active, showing a message: "Please refer to [Nature Protocol](#) publication for details to how to use page." Below this message is a "Help Tips" section with "Steps:" and a list: 1. Select list and list type to analyze, 2. Select Organism, 3. Select operation. To the right of the steps is a form for "Enter IDs: Supported IDs" with a text input field and a "Choose File" button. Below the input field is the text "No file chosen". To the right of the input field is the text "separate IDs by a space or comma". Below the "Enter IDs" section is a section for "Upload IDs: File format".

PROTEIN ANALYSIS THROUGH EVOLUTIONARY RELATIONSHIPS (PANTHER)

<http://www.pantherdb.org/>




DAVID Bioinformatics Resources 6.8(Beta)
National Institute of Allergy and Infectious Diseases (NIAID), NIH


[Home](#) [Start Analysis](#) [Shortcut to DAVID Tools](#) [Technical Center](#) [Downloads & APIs](#) [Term of Service](#) [Why DAVID?](#) [About Us](#)

*** Welcome to DAVID 6.8 Beta with updated Knowledgebase ([more info](#)). ***

Shortcut to DAVID Tools

 **Functional Annotation**

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)

 **Gene Functional Classification**


Provide a rapid means to reduce large lists of genes into functionally related groups of genes to

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8 Beta

2003 - 2016

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 Beta [comprises a full Knowledgebase update to the sixth version of our](#)

 **What's Important in DAVID?**

- [Version 6.7 release note](#)
- [New requirement to cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)

DATABASE FOR ANNOTATION, VISUALIZATION AND INTEGRATED DISCOVERY (DAVID)

<https://david.ncifcrf.gov/>



WebGestalt

WEB-based GENE SeT AnaLysis Toolkit

Translating gene lists into biological insights...

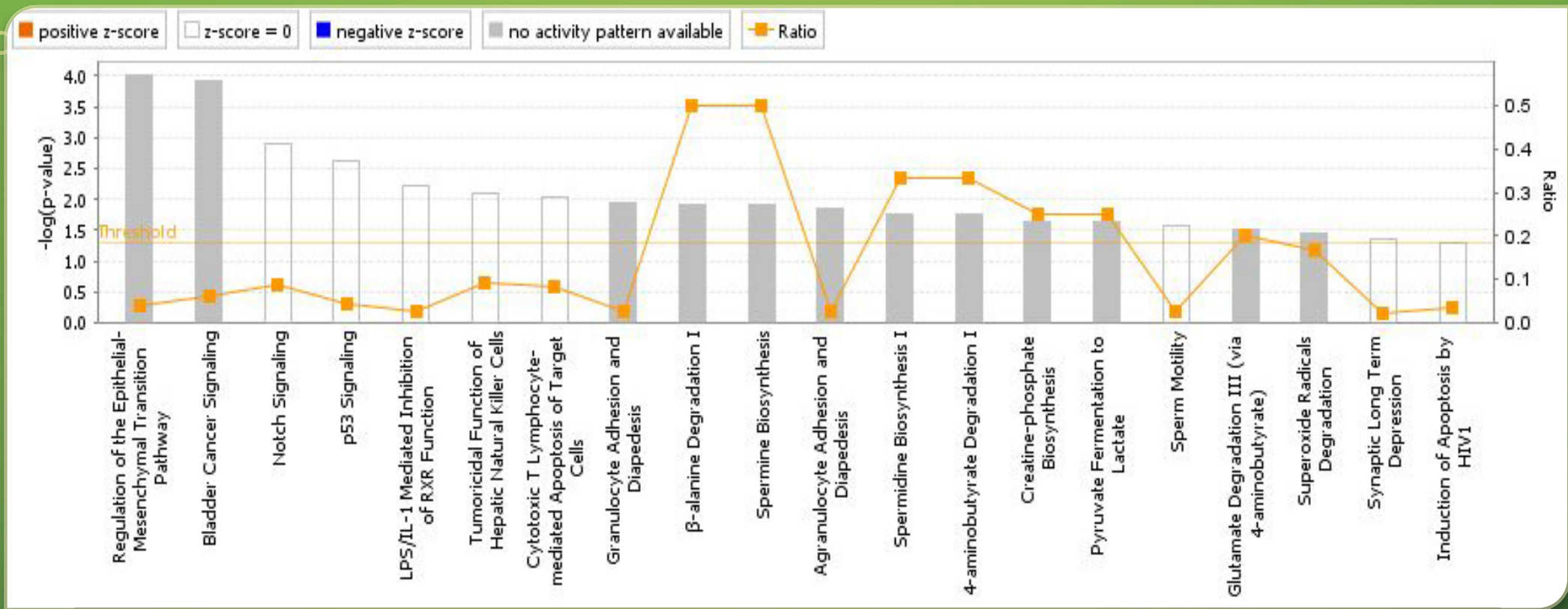
START | [Sample data](#) | [Manual](#) | [Citation](#) | [User Forum](#)

To visualize and compare multiple GO term lists, please use [GOView](#).

To discuss the use and development of WebGestalt or GOView, please join the new [User Forum](#).

WEB-BASED GENE SET ANALYSIS TOOLKIT (WEBGESTALT)

<http://www.webgestalt.org/>

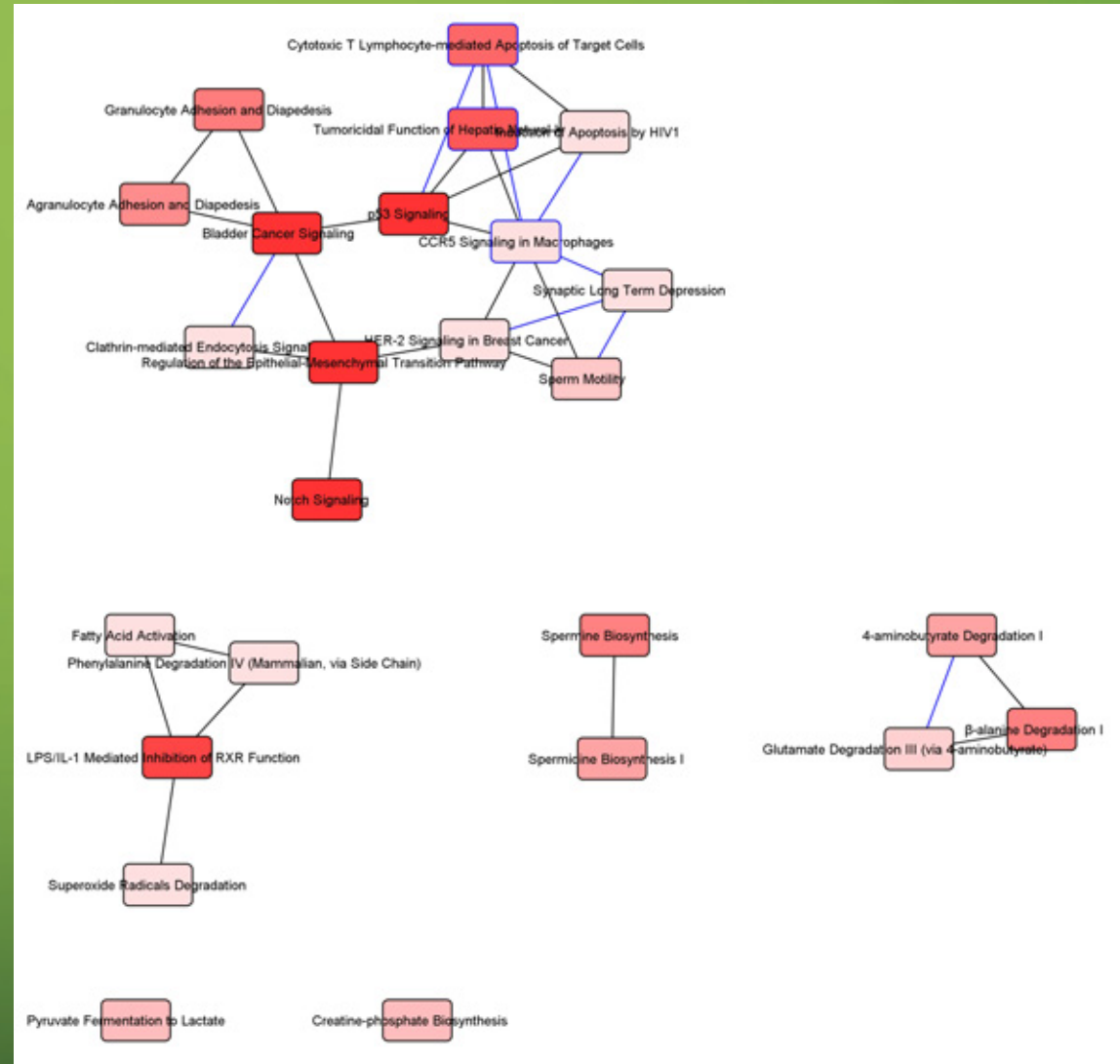


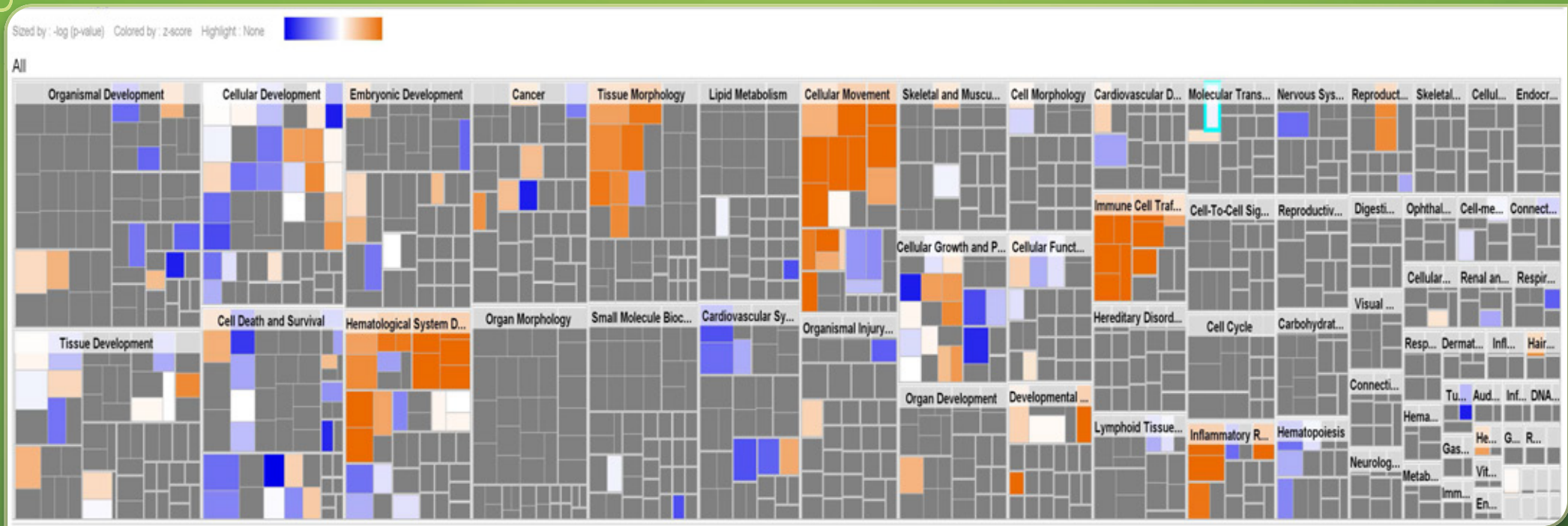
GENE ONTOLOGY ANALYSIS

Created with Ingenuity Pathway Analysis software, March 2016

GENE ONTOLOGY ANALYSIS

Created with Ingenuity Pathway Analysis software, March 2016





GENE ONTOLOGY ANALYSIS

Created with Ingenuity Pathway Analysis software, March 2016

The background is a green gradient. In the corners, there are decorative elements resembling circuit board traces or neural network connections, consisting of thin white lines and small circles.

PATHWAY ANALYSIS

The banner features the Reactome logo on the left, which includes a diagram of a metabolic pathway with white rounded rectangles and arrows. To the right of the logo is a large, detailed illustration of a biological pathway. It shows a red oval (possibly a lipid or protein) interacting with a green protein structure, with various colored dots (blue, orange, yellow) representing molecules and arrows indicating the flow of the reaction. The background is a dark blue gradient.

REACTOME
A CURATED PATHWAY DATABASE

About Content Documentation Tools Community Download Contact

e.g. O95631, NTN1, signalin Search

 Browse Pathways

 Analyze Data

 Reactome FI Network

Tweets

Current Version: Reactome hits 50

 IntAct at EBI 14 Oct

THE REACTOME PROJECT

<http://www.reactome.org/>

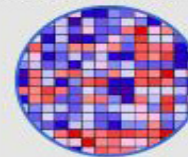
Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- ▶ **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- ▶ **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- ▶ **View documentation** describing GSEA and MSigDB.

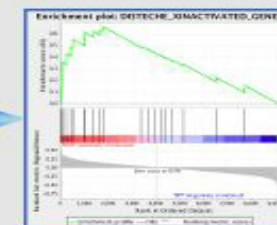
Molecular Profile Data



Gene Set Database



Enriched Sets



GENE SET ENRICHMENT ANALYSIS (GSEA) & THE MOLECULAR SIGNATURES DATABASE (MSIGDB)

<http://www.broadinstitute.org/gsea/index.jsp>



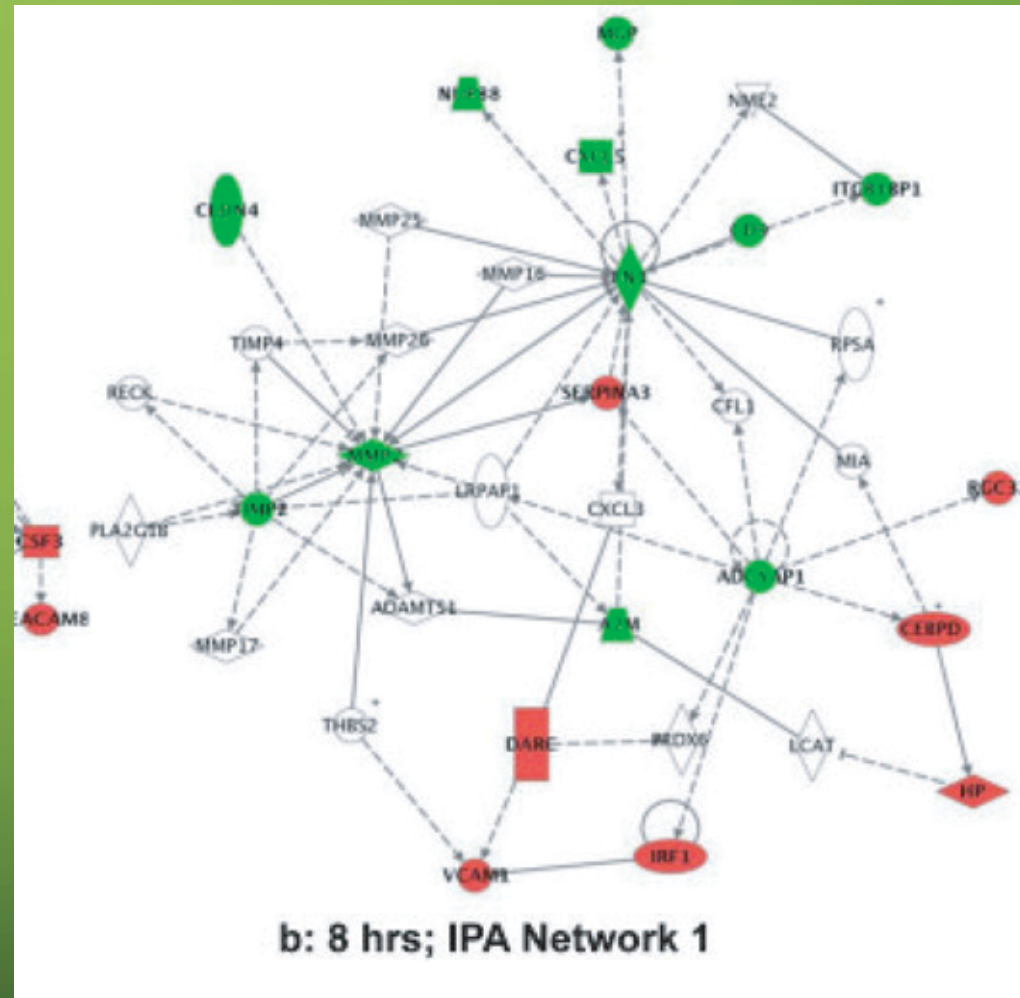
INGENUITY PATHWAY ANALYSIS

<http://www.ingenuity.com/products/ipa>

<https://www.urmc.rochester.edu/libraries/Miner/research/MolecularBiologyTools.cfm>

INGENUITY PATHWAY ANALYSIS OUTPUT

Hypothetical “network” generated
by Ingenuity Pathway Analysis



The background is a solid green gradient. In the corners, there are abstract, light green line art designs that resemble circuit boards or neural network connections. These designs consist of vertical and horizontal lines of varying lengths, some ending in small circles, creating a technical and scientific aesthetic.

PROTEIN INTERACTION ANALYSIS



Welcome to the Biological General Repository for Interaction Datasets

BioGRID is an interaction repository with data compiled through comprehensive curation efforts. Our current index is version **3.4.134** and searches **55,957** publications for **1,055,196** protein and genetic interactions, **27,501** chemical associations and **38,559** post translational modifications from major model organism species. All data are **freely** provided via our search index and available for download in standardized formats.

INTERACTION STATISTICS

LATEST DOWNLOADS

Search the BioGRID

Search by identifiers, keywords, and gene names...

All Organisms

SUBMIT GENE SEARCH Q



Advanced Search



Search Tips



Featured Datasets

By Gene

By Publication

BIOLOGICAL GENERAL REPOSITORY FOR INTERACTION DATASETS (BIOGRID)

<http://thebiogrid.org/>


IntAct Molecular Interaction Database

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. The IntAct Team also produce the [Complex Portal](#).

Search in IntAct

Enter search term(s)...

Search

 Search Tips

Examples

- Gene, Protein, RNA or Chemical name: [BRCA2](#), [Staurosporine](#)
- UniProtKB or ChEBI AC: [Q06609](#), [CHEBI:15996](#)
- UniProtKB ID: [LCK_HUMAN](#)
- RNACentral ID: [URS00004C95F4_559292](#)

INTACT MOLECULAR INTERACTION DATABASE

<http://www.ebi.ac.uk/intact/>

[Search](#)[Download](#)[Help](#)[My Data](#)

Welcome to STRING


Protein-Protein Interaction Networks

SEARCH TOOL FOR THE RETRIEVAL OF INTERACTING GENES/PROTEINS (STRING)

<http://string-db.org/>

The background is a solid green gradient. In the corners, there are abstract, light green line art designs that resemble circuit traces or molecular structures. These designs consist of thin lines connecting small circles, creating a network-like pattern. The top-left and bottom-left corners have more complex, branching structures, while the top-right and bottom-right corners have simpler, more linear traces.

METABOLOMICS DATA ANALYSIS



MetaboAnalyst 3.0
– a comprehensive tool suite for metabolomic data analysis

[Home](#)
[Overview](#)
[Data Formats](#)
[FAQs](#)
[Tutorials](#)

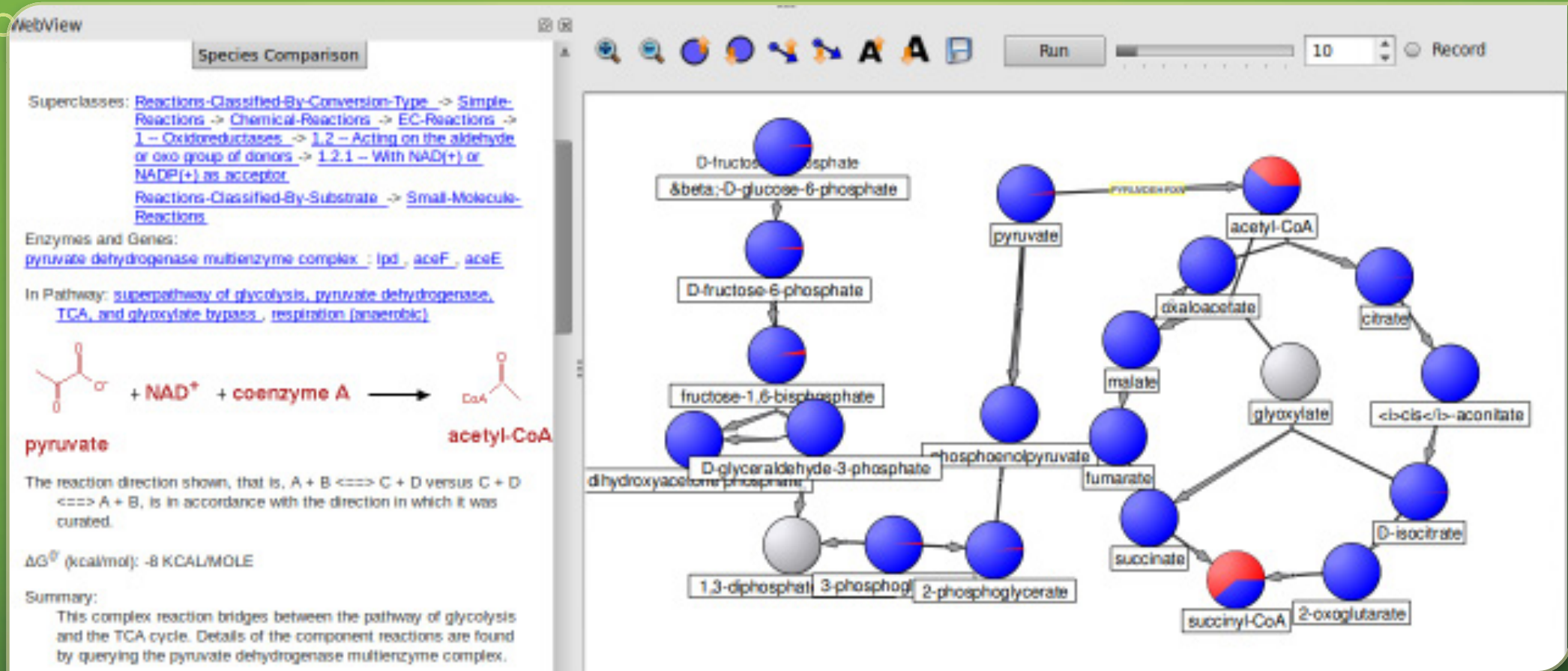
Welcome [click here to start](#)

News & Updates

- Added support for batch effect correction for multiple data sets (**Other Utilities** module) (02/22/2016); **NEW**
- Updated the web framework for better performance (02/18/2016); **NEW**
- Upgraded the Google Cloud server for improved performance (10/30/2015); **NEW**
- Added support for detailed ROC curve analysis of individual biomarkers (10/29/2015); **NEW**
- Several feature improvements and bug fixes based on user feedback (10/16/2015); **NEW**

METABOANALYST 3.0

<http://www.metaboanalyst.ca/>



MAVEN: OPEN-SOURCE METABOLOMICS DATA ANALYZER

<http://genomics-pubs.princeton.edu/mzroll/index.php?show=index>

IMPALA: Integrated Molecular Pathway Level Analysis

pathway over-representation and enrichment analysis with expression and / or metabolite data

genes/proteins

- example input for over-representation analysis
- example input for enrichment analysis

paste genes or proteins below



metabolites

- example input for over-representation analysis
- example input for enrichment analysis

paste metabolites below



INTEGRATED MOLECULAR PATHWAY LEVEL ANALYSIS (IMPALA)

<http://impala.molgen.mpg.de/>

The background is a solid green gradient. In the four corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles connecting them.

HIGH-THROUGHPUT CHEMICAL SCREENING



PUBCHEM

<https://pubchem.ncbi.nlm.nih.gov/>