Module 3, Lesson 17
Introduction to Pathway Analysis

Alexandra Emmons, PhD

December 1, 2022

Bioinformatics Training and Education Program

NIH NATIONAL CANCER INSTITUTE
Center for Cancer Research

Module 2 Survey

- Let’s take a second to take a brief Webex poll
- This will help us improve the class and clear up any misunderstandings by the final lesson.

Module 3 Overview

Lesson 1: Introduction to Pathway Analysis

Lesson 2: Functional gene enrichment with DAVID

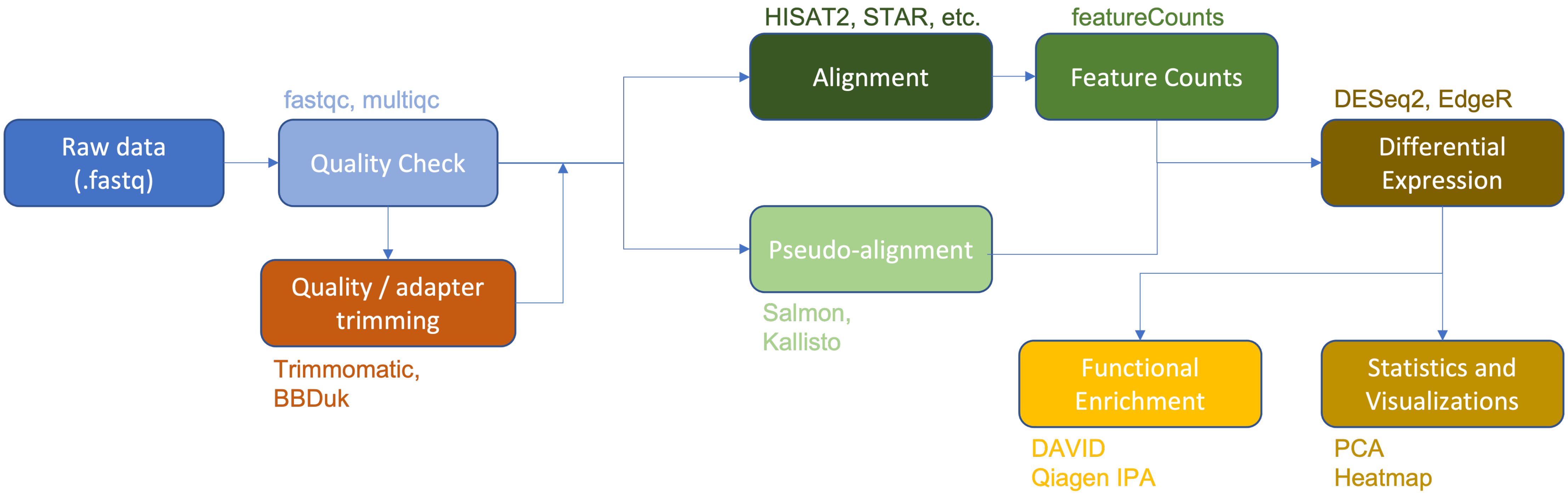
Lesson 3: Qiagen IPA

Lesson 4: Course Wrap-Up

Lesson 17 objectives

- Determine potential next steps following differential expression analysis.
- Tour geneontology.org and understand the three main ontologies.
- Learn about different methods and tools related to functional enrichment and pathway analysis.
- Get familiar with databases commonly used by popular functional enrichment tools.

Overview



HISAT2, STAR, etc.

featureCounts

fastqc, multiqc

DESeq2, EdgeR

Raw data (.fastq)

Quality Check

Alignment

Feature Counts

Differential Expression

Pseudo-alignment

Quality / adapter trimming

Salmon, Kallisto

Trimmomatic, BBDuk

Functional Enrichment

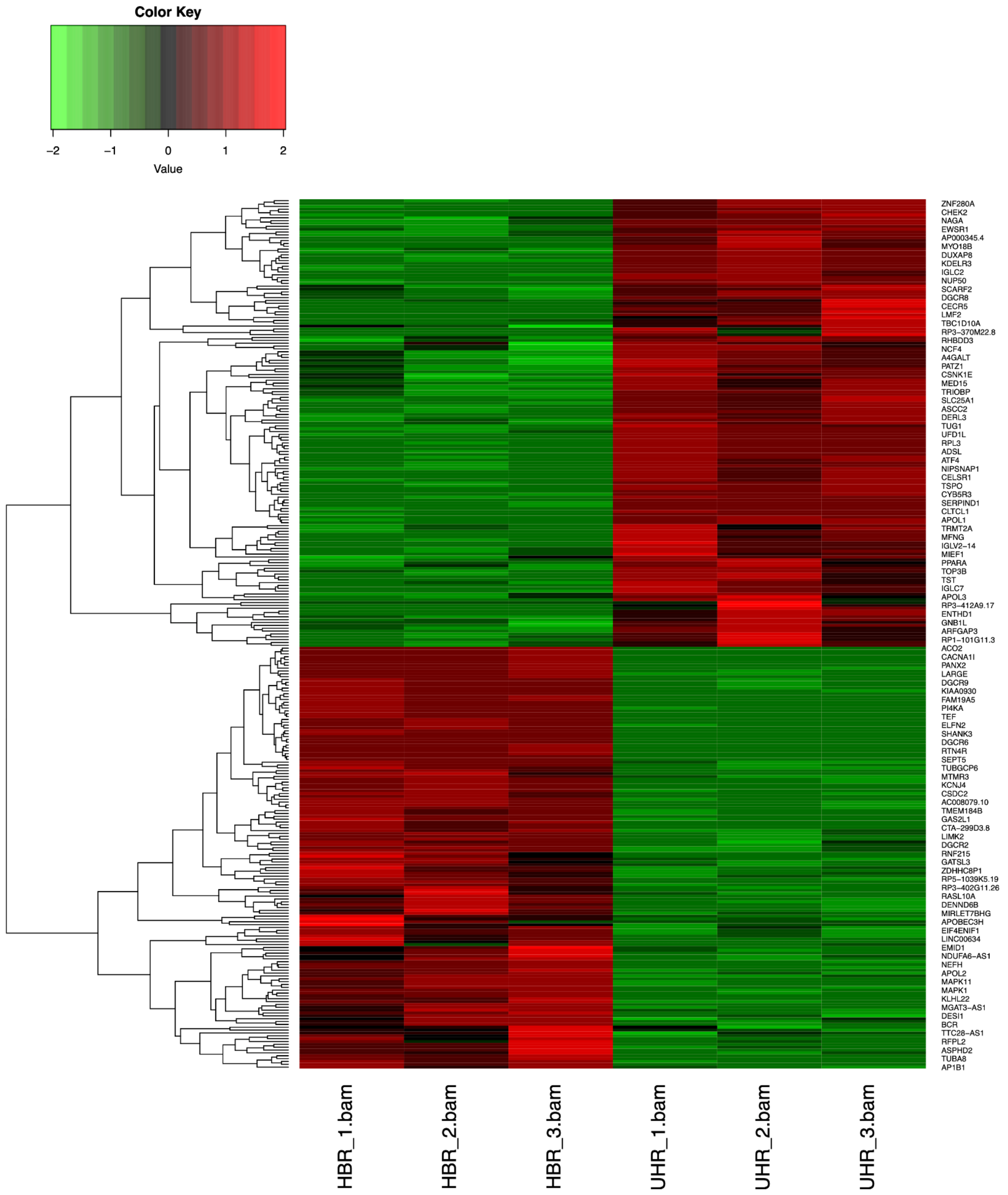
Statistics and Visualizations

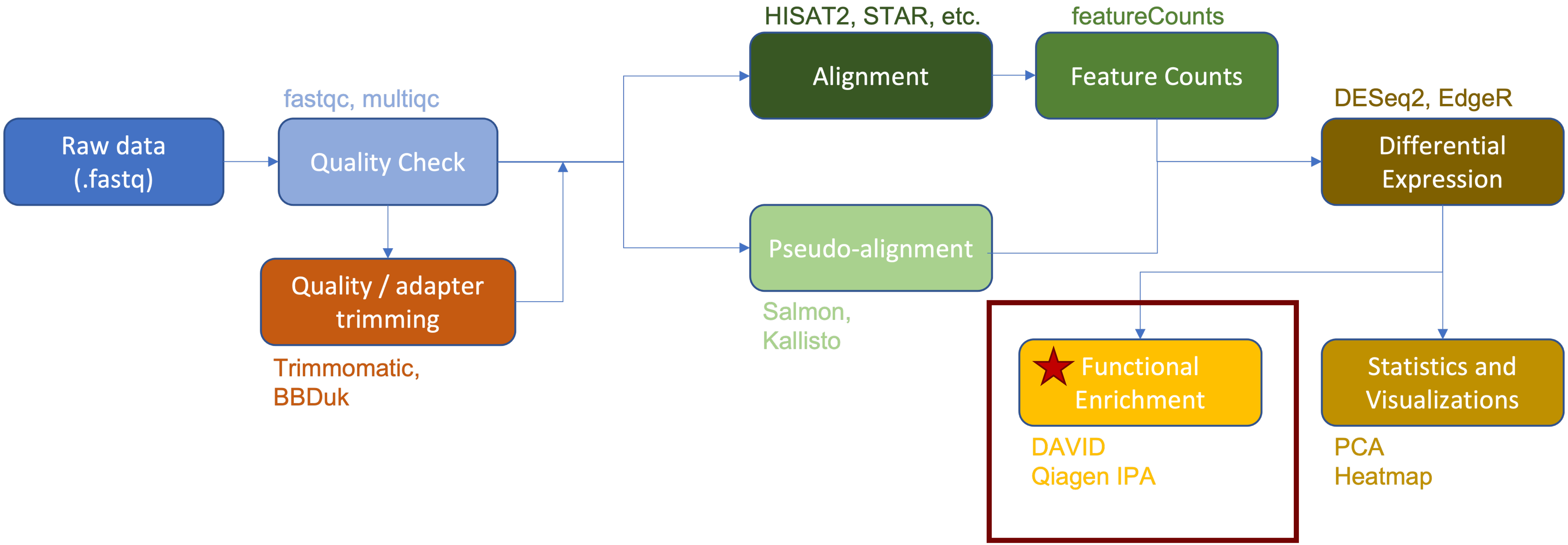
DAVID
Qiagen IPA

PCA
Heatmap

SO MANY GENES….

We generated a heatmap, so what’s next?




Raw data (.fastq)
fastqc, multiqc
Quality Check
Quality / adapter trimming
Trimmomatic, BBDuk
HISAT2, STAR, etc.
Alignment
Pseudo-alignment
Salmon, Kallisto
featureCounts
Feature Counts
DESeq2, EdgeR
Differential Expression
★ Functional Enrichment
DAVID
Qiagen IPA
Statistics and Visualizations
PCA
Heatmap

Why gene set / pathway analysis?

1

Increase the statistical power in our analysis

2

Ease interpretation

3

Predict new roles for genes

4

Better integrate data from different methods

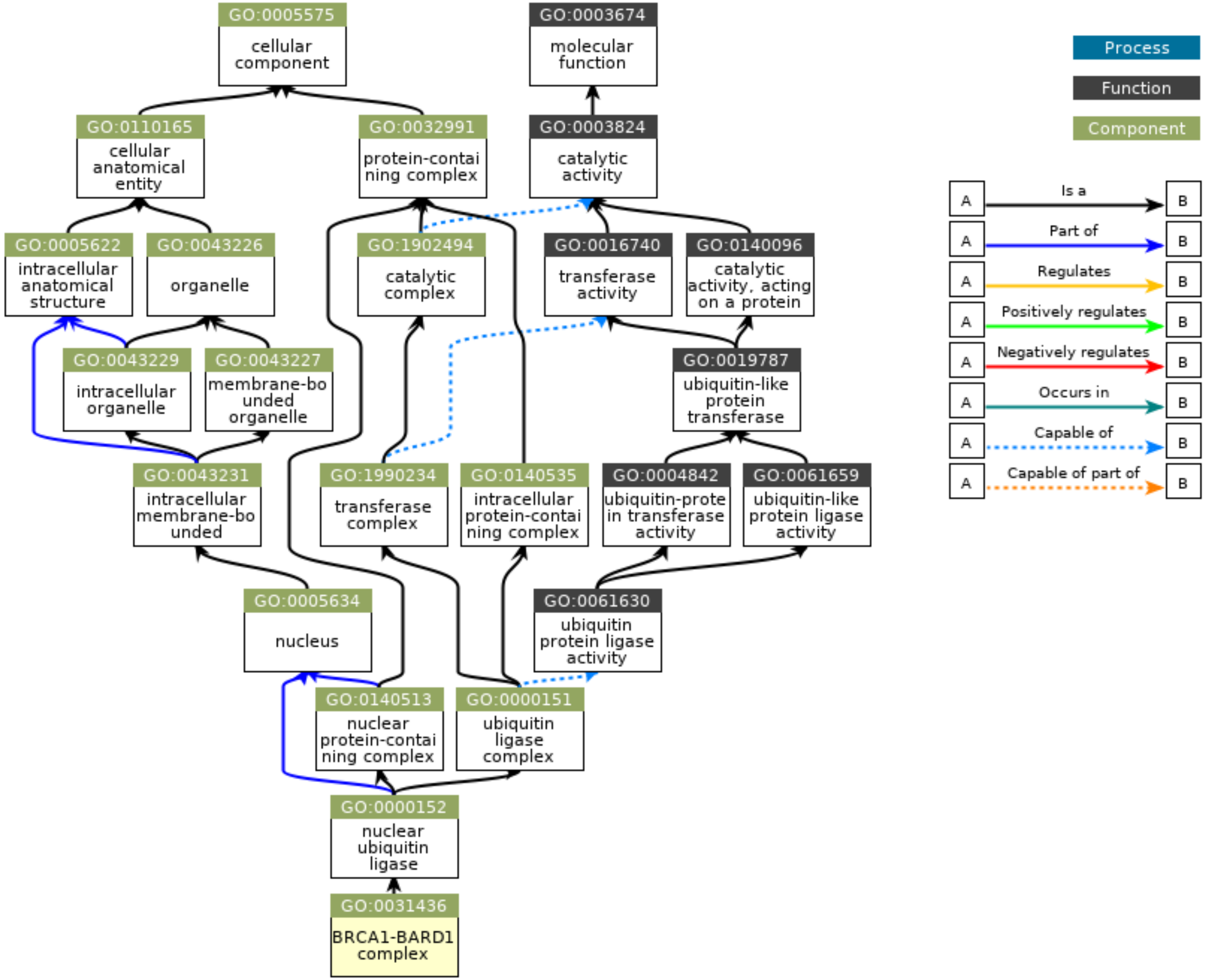
What is gene ontology?

- The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms. --- https://www.ebi.ac.uk/ols/ontologies/go
  - Controlled vocabulary
  - Maintained by the Gene Ontology Consortium
    - Updated regularly
- Two parts:
  - the ontology (the GO terms and their hierarchical relationship)
  - the annotations (the annotated genes linked to various GO terms)

What is gene ontology?

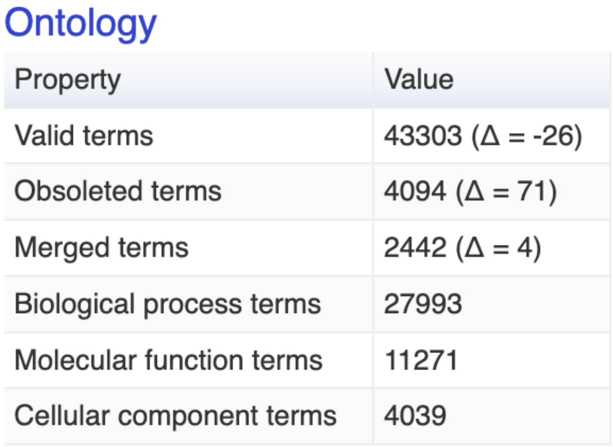
GO integrates information about gene product function in the context of three domains:

- Molecular function (MF) - "the molecular activities of individual gene products”
- Cellular component (CC) - "where the gene products are active"
- Biological process (BP) - "the pathways and larger processes to which that gene product’s activity contributes”



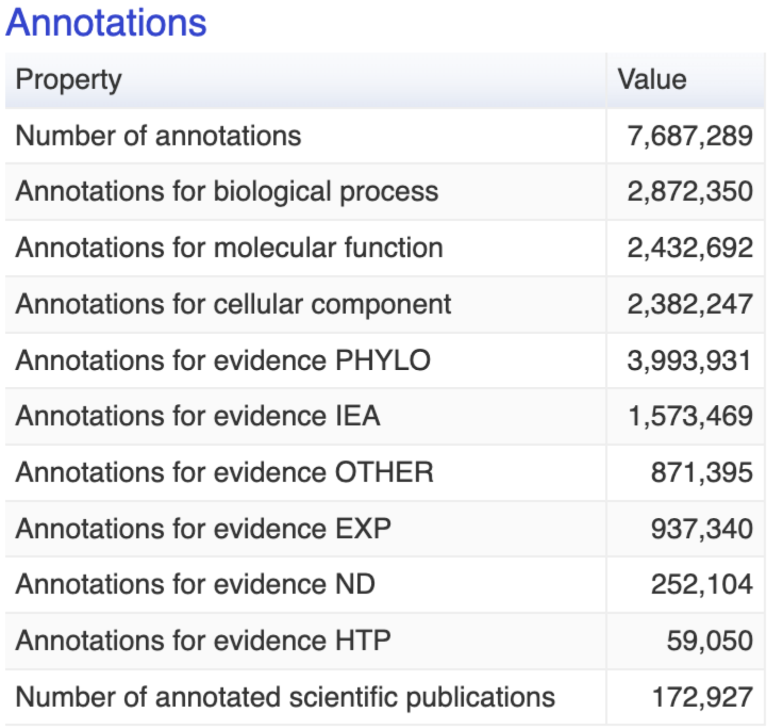
QuickGO - https://www.ebi.ac.uk/QuickGO

GO Statistics



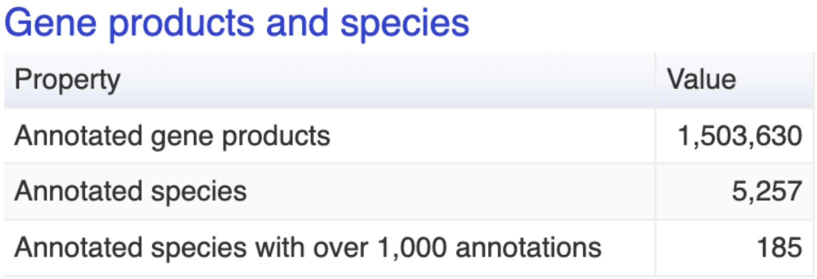
Ontology

Property	Value
Valid terms	43303 (Δ = -26)
Obsoleted terms	4094 (Δ = 71)
Merged terms	2442 (Δ = 4)
Biological process terms	27993
Molecular function terms	11271
Cellular component terms	4039



Annotations

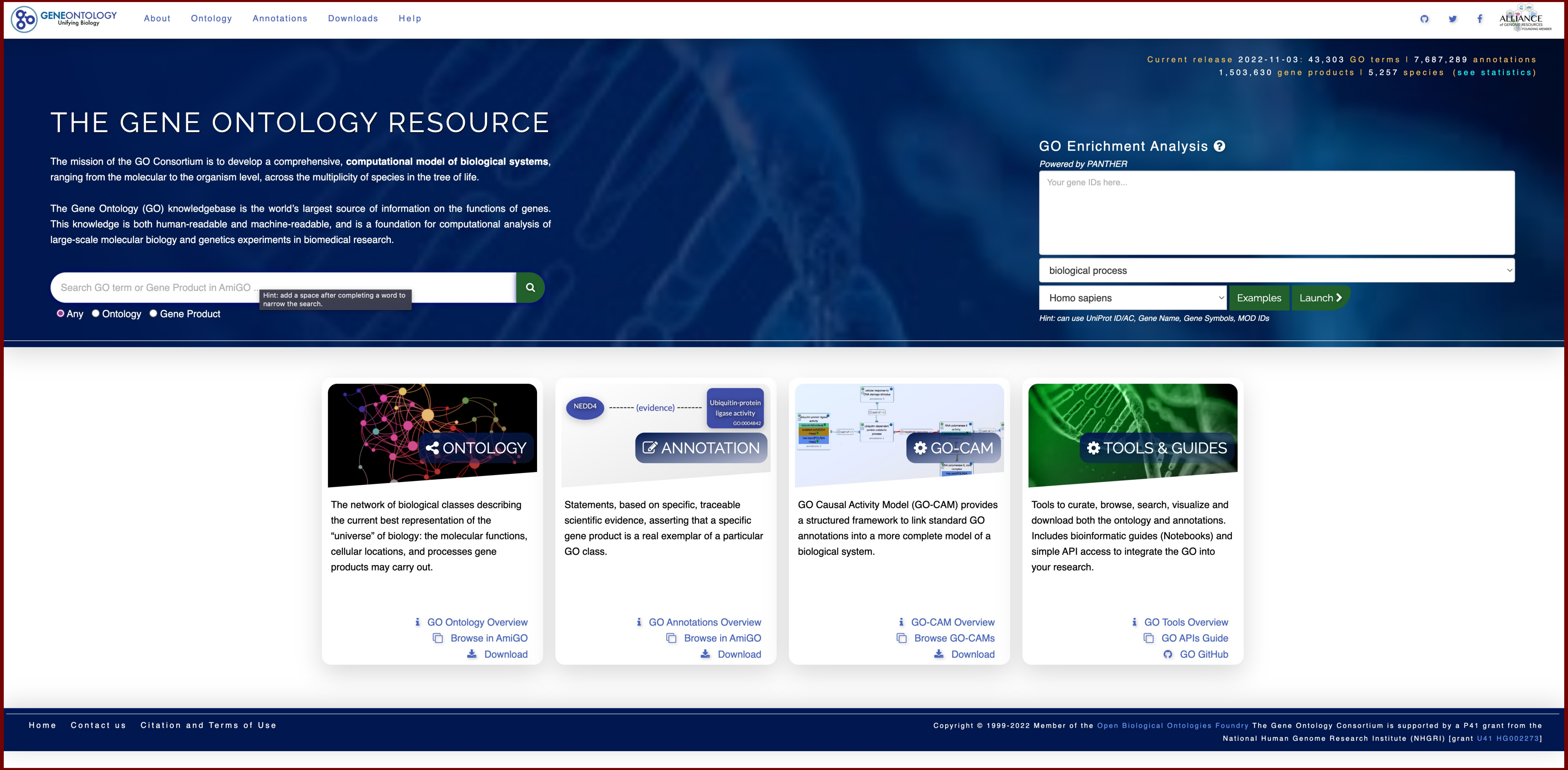
Property	Value
Number of annotations	7,687,289
Annotations for biological process	2,872,350
Annotations for molecular function	2,432,692
Annotations for cellular component	2,382,247
Annotations for evidence PHYLO	3,993,931
Annotations for evidence IEA	1,573,469
Annotations for evidence OTHER	871,395
Annotations for evidence EXP	937,340
Annotations for evidence ND	252,104
Annotations for evidence HTP	59,050
Number of annotated scientific publications	172,927



Gene products and species

Property	Value
Annotated gene products	1,503,630
Annotated species	5,257
Annotated species with over 1,000 annotations	185

Check out these tips for working with GO terms!



Current release 2022-11-03: 43,303 GO terms | 7,687,289 annotations
1,503,630 gene products | 5,257 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...

Hint: add a space after completing a word to narrow the search.

● Any ○ Ontology ○ Gene Product

GO Enrichment Analysis

Powered by PANTHER

Your gene IDs here...

biological process

Homo sapiens

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs

The network of biological classes describing the current best representation of the "universe" of biology: the molecular functions, cellular locations, and processes gene products may carry out.

Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular GO class.

GO Causal Activity Model (GO-CAM) provides a structured framework to link standard GO annotations into a more complete model of a biological system.

Tools to curate, browse, search, visualize and download both the ontology and annotations. Includes bioinformatic guides (Notebooks) and simple API access to integrate the GO into your research.

Copyright © 1999-2022 Member of the Open Biological Ontologies Foundry The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant U41 HG002273]

Other databases

- Kyoto Encyclopedia of Genes and Genomes (KEGG)
  - Curated database
  - biological pathways
  - Molecular interaction networks
  - Very nice pathway maps
  - Restricted licenses
- Pathway Commons
  - a meta-database of pathways from other pathway databases
- PANTHER
  - Database of signaling pathways
- WikiPathways
  - community driven meta-database of pathways

Other databases

- NDEx
  - an open-source framework where scientists and organizations can store, share, manipulate, and publish biological network knowledge
- HumanCyc
  - an encyclopedic reference on human metabolic pathways, the human genome, and human metabolites.

Check out Pathguide to get an idea of how many databases are available.

3 general approaches to pathway analysis:

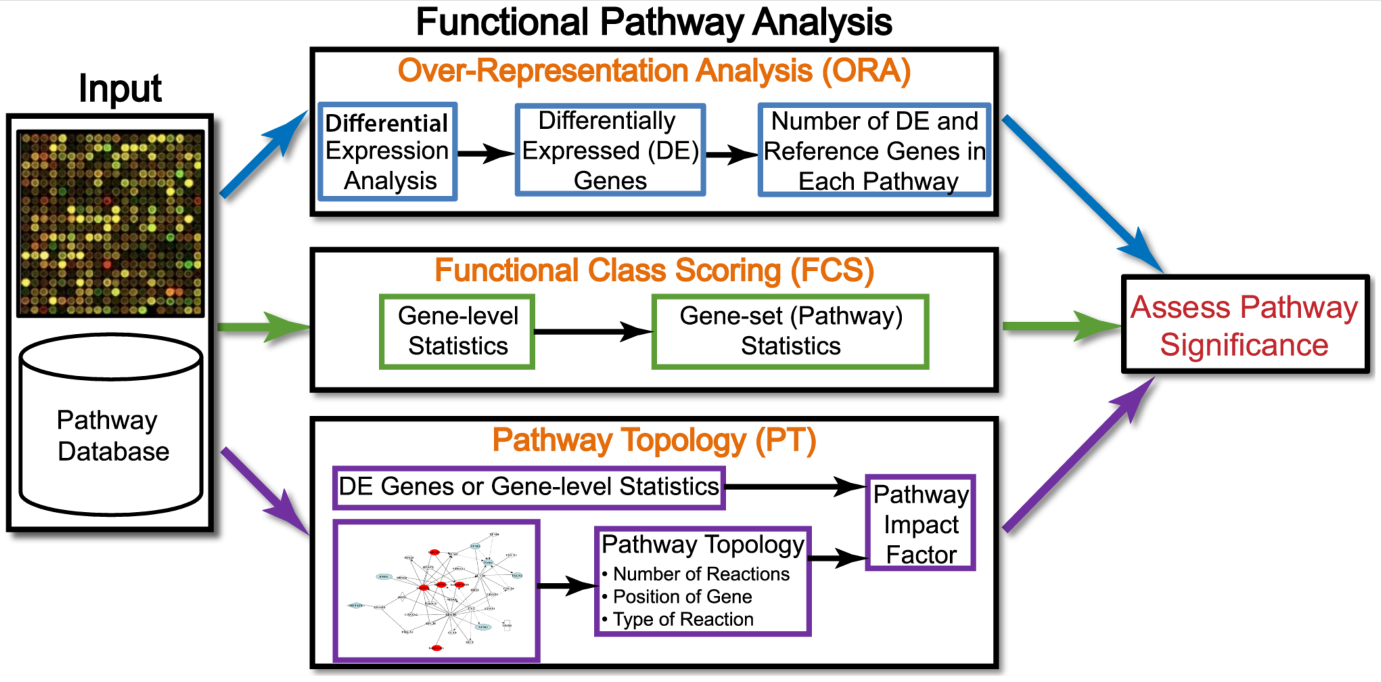

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

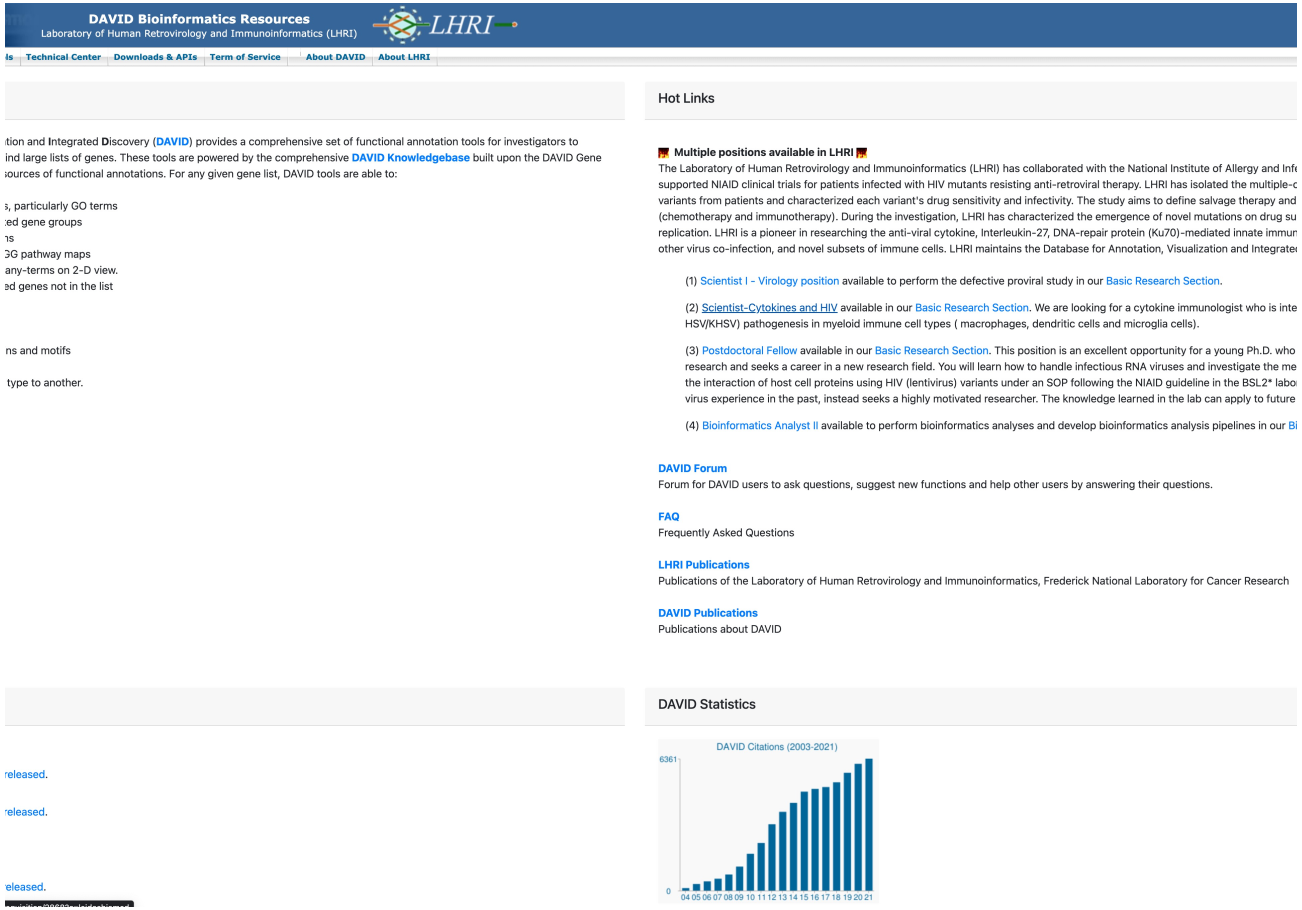
Image from Khatri et al. 2012

- Over-representation analysis
- Functional class scoring
- Pathway topology

Over-representation Analysis

- Statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression --- (Khatri et al. 2012)
- Strategy
  - Provide an input list of gene IDs (uses a threshold)
  - Input genes for each pathway are counted
  - The same counting step is applied to a background set of genes
  - Pathways are tested for over or under representation using tests based on hypergeometric, chi-square, or binomial distribution

DAVID



DAVID Bioinformatics Resources
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)
LHRI
ls
Technical Center
Downloads & APIs
Term of Service
About DAVID
About LHRI

ıtion and **I**ntegrated **D**iscovery (DAVID) provides a comprehensive set of functional annotation tools for investigators to
ind large lists of genes. These tools are powered by the comprehensive **DAVID Knowledgebase** built upon the DAVID Gene
sources of functional annotations. For any given gene list, DAVID tools are able to:

ṡ, particularly GO terms
ːed gene groups
ıs
ƷG pathway maps
any-terms on 2-D view.
ed genes not in the list

ns and motifs

type to another.

released.

released.

eleased.

Hot Links

Multiple positions available in LHRI

The Laboratory of Human Retrovirology and Immunoinformatics (LHRI) has collaborated with the National Institute of Allergy and Infe
supported NIAID clinical trials for patients infected with HIV mutants resisting anti-retroviral therapy. LHRI has isolated the multiple-c
variants from patients and characterized each variant's drug sensitivity and infectivity. The study aims to define salvage therapy and
(chemotherapy and immunotherapy). During the investigation, LHRI has characterized the emergence of novel mutations on drug su
replication. LHRI is a pioneer in researching the anti-viral cytokine, Interleukin-27, DNA-repair protein (Ku70)-mediated innate immun
other virus co-infection, and novel subsets of immune cells. LHRI maintains the Database for Annotation, Visualization and Integrate

(1) Scientist I - Virology position available to perform the defective proviral study in our Basic Research Section.

(2) Scientist-Cytokines and HIV available in our Basic Research Section. We are looking for a cytokine immunologist who is inte
HSV/KHSV) pathogenesis in myeloid immune cell types (macrophages, dendritic cells and microglia cells).

(3) Postdoctoral Fellow available in our Basic Research Section. This position is an excellent opportunity for a young Ph.D. who
research and seeks a career in a new research field. You will learn how to handle infectious RNA viruses and investigate the me
the interaction of host cell proteins using HIV (lentivirus) variants under an SOP following the NIAID guideline in the BSL2* labo
virus experience in the past, instead seeks a highly motivated researcher. The knowledge learned in the lab can apply to future

(4) Bioinformatics Analyst II available to perform bioinformatics analyses and develop bioinformatics analysis pipelines in our Bi

DAVID Forum
Forum for DAVID users to ask questions, suggest new functions and help other users by answering their questions.

FAQ
Frequently Asked Questions

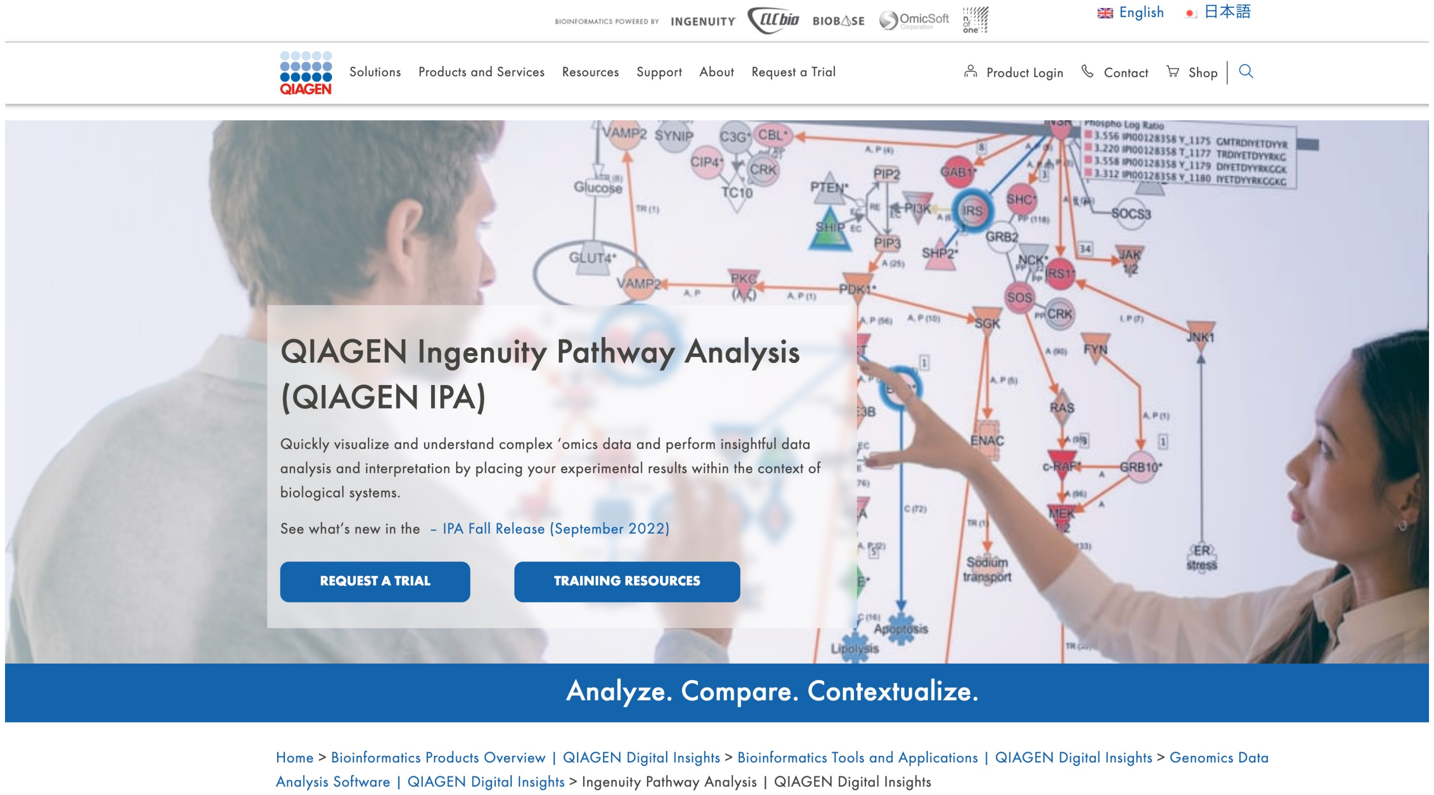
LHRI Publications
Publications of the Laboratory of Human Retrovirology and Immunoinformatics, Frederick National Laboratory for Cancer Research

DAVID Publications
Publications about DAVID

DAVID Statistics

- **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery
- Very popular
- Easy to use and produces a lot of output
- Uses a variety of databases (NCBI, Uniprot, Ensembl, Gene Ontology, KEGG, Reactome, etc.).

QIAGEN IPA



QIAGEN Ingenuity Pathway Analysis (QIAGEN IPA)

Quickly visualize and understand complex ’omics data and perform insightful data analysis and interpretation by placing your experimental results within the context of biological systems.

See what’s new in the – IPA Fall Release (September 2022)

Analyze. Compare. Contextualize.

Home > Bioinformatics Products Overview | QIAGEN Digital Insights > Bioinformatics Tools and Applications | QIAGEN Digital Insights > Genomics Data Analysis Software | QIAGEN Digital Insights > Ingenuity Pathway Analysis | QIAGEN Digital Insights

Functional Class Scoring

- “The hypothesis of functional class scoring (FCS) is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes (i.e., pathways) can also have significant effects.” --- Khatri et al. 2012
- Strategy:
    - Compute a gene level statistic (differential expression)
    - Create a pathway level statistic by aggregating gene level stats
    - Determine statistical significance from pathway stat
        - Competitive vs self-contained methods

Functional Class Scoring: GSEA

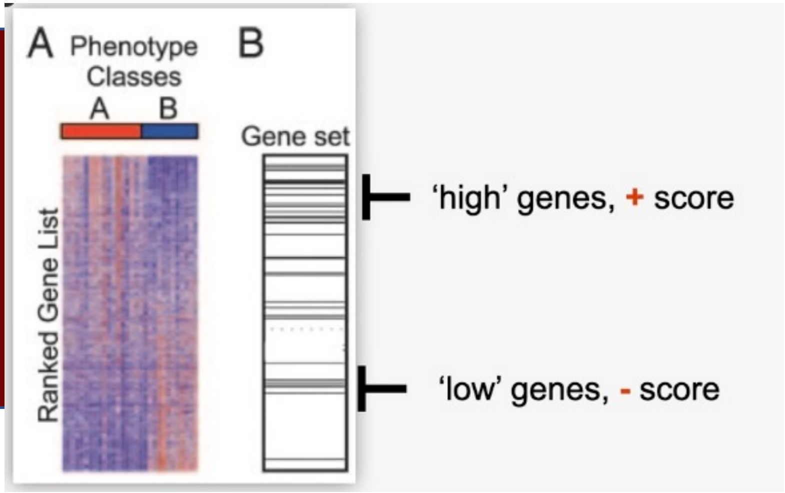
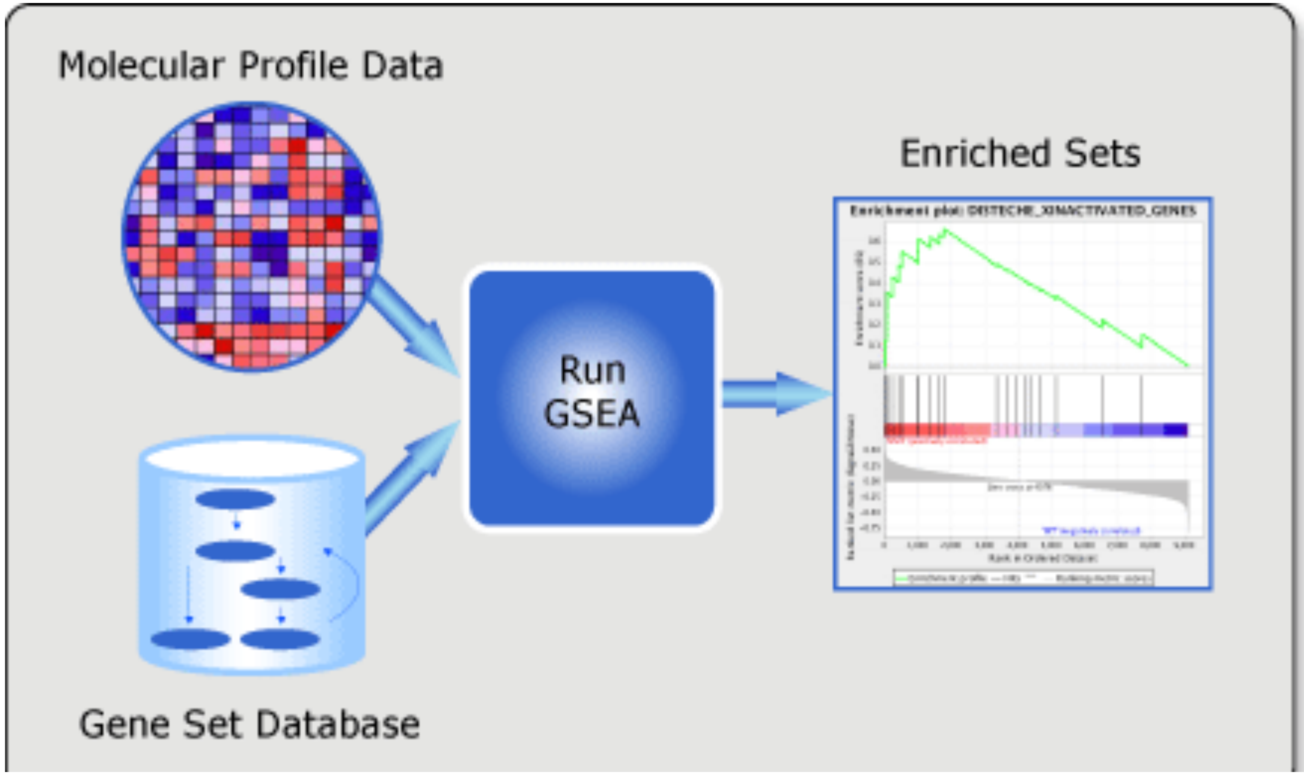


Image from https://diytranscriptomics.com/project/lecture-10



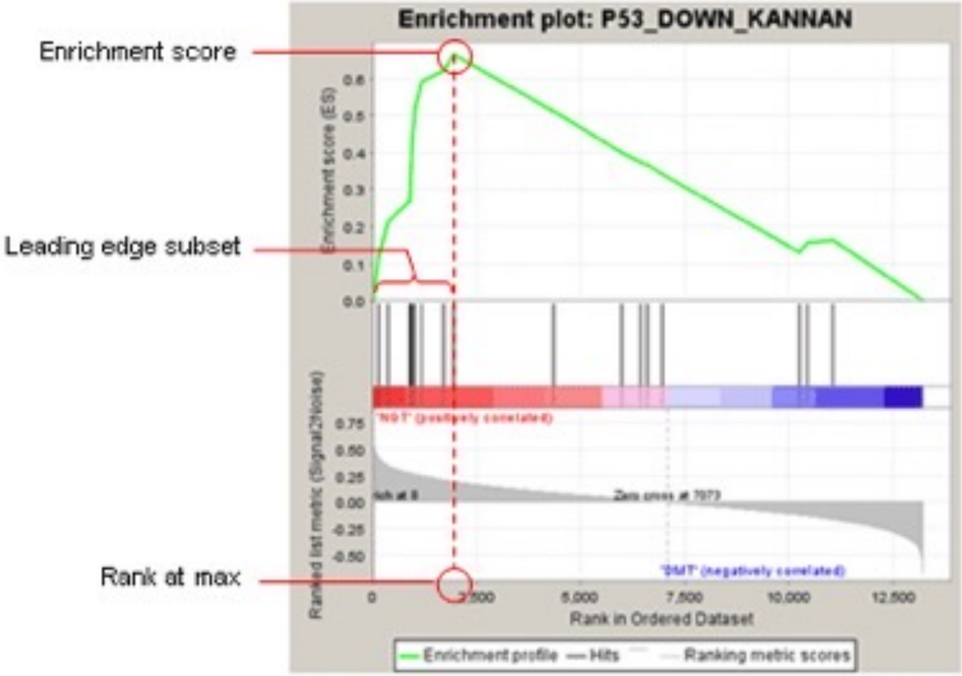
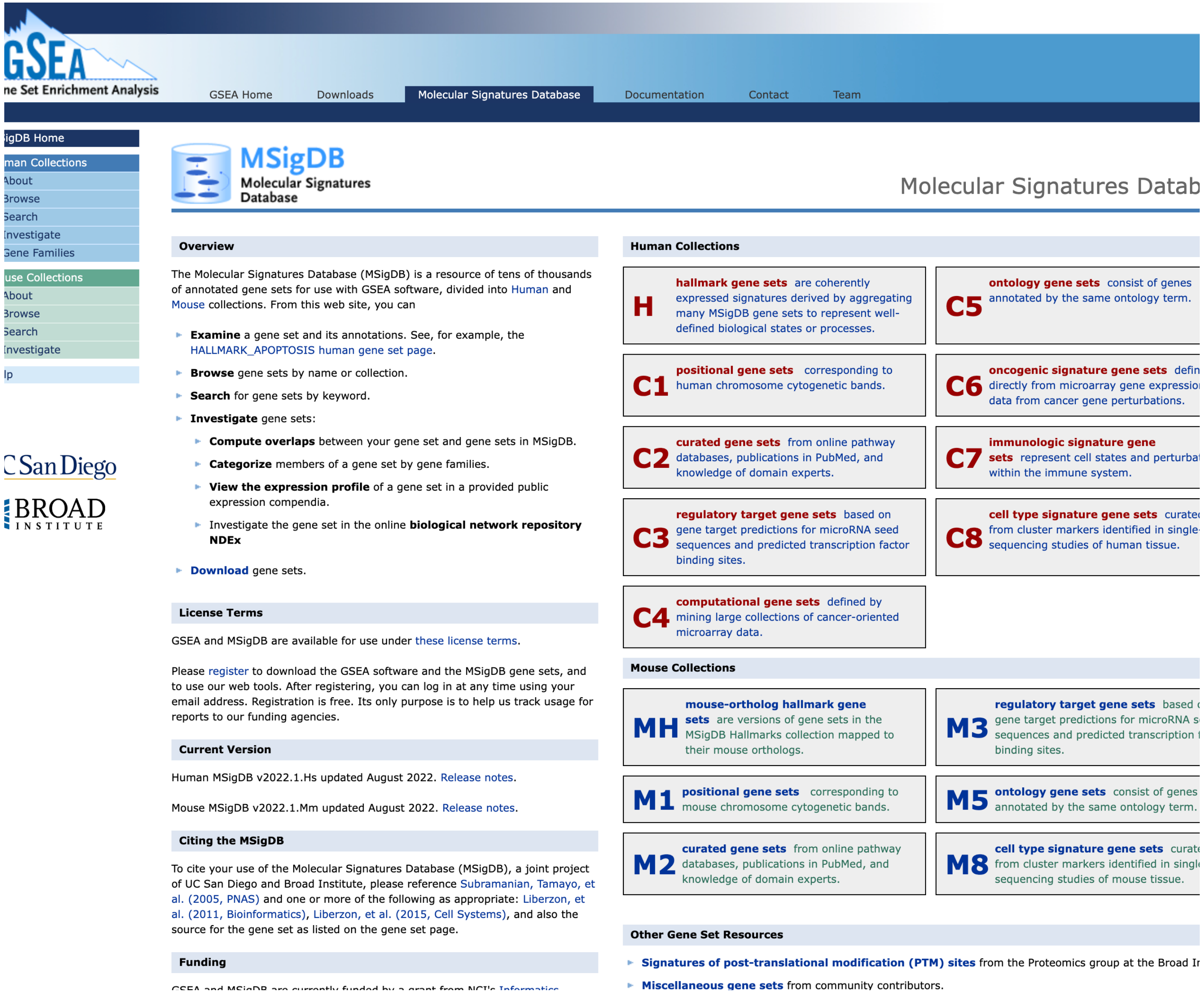


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

MSigDB

- Curated signatures (gene sets) database by UC San Diego and the Broad Institute
- 33196 gene sets
- Themed mouse and human collections



Molecular Signatures Datab

Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software and the MSigDB gene sets, and to use our web tools. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

Human MSigDB v2022.1.Hs updated August 2022. Release notes.

Mouse MSigDB v2022.1.Mm updated August 2022. Release notes.

Citing the MSigDB

To cite your use of the Molecular Signatures Database (MSigDB), a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et al. (2005, PNAS) and one or more of the following as appropriate: Liberzon, et al. (2011, Bioinformatics), Liberzon, et al. (2015, Cell Systems), and also the source for the gene set as listed on the gene set page.

Funding

GSEA and MSigDB are currently funded by a grant from NCI's Informatics

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defin directly from microarray gene expressio data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbat within the immune system.

C8 **cell type signature gene sets** curatec from cluster markers identified in single- sequencing studies of human tissue.

Mouse Collections

MH **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

M1 **positional gene sets** corresponding to mouse chromosome cytogenetic bands.

M2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

M3 **regulatory target gene sets** based c gene target predictions for microRNA s sequences and predicted transcription f binding sites.

M5 **ontology gene sets** consist of genes annotated by the same ontology term.

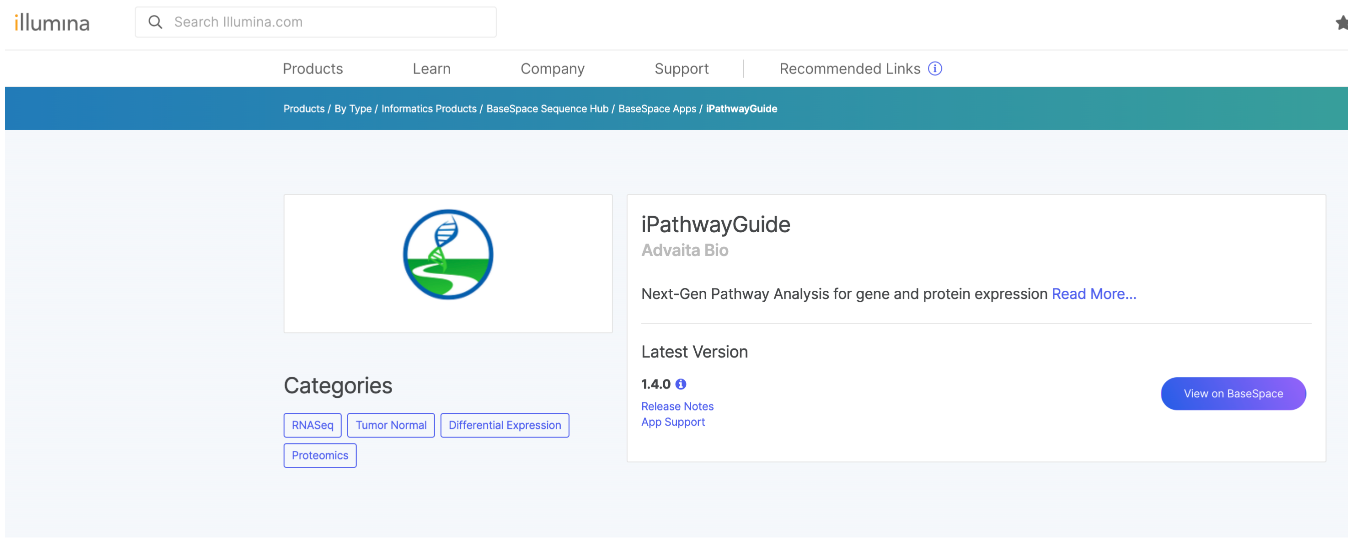
M8 **cell type signature gene sets** curate from cluster markers identified in singl sequencing studies of mouse tissue.

Other Gene Set Resources

- **Signatures of post-translational modification (PTM) sites** from the Proteomics group at the Broad Ir
- **Miscellaneous gene sets** from community contributors.

Pathway Topology

- PT methods generally use network based modeling
- They consider the information ignored in ORA and FCS methods: gene product interactions, positions of genes, and types of genes
- Examples iPathwayGuide, Pathway-Express, SPIA, NetGSA



iPathwayGuide

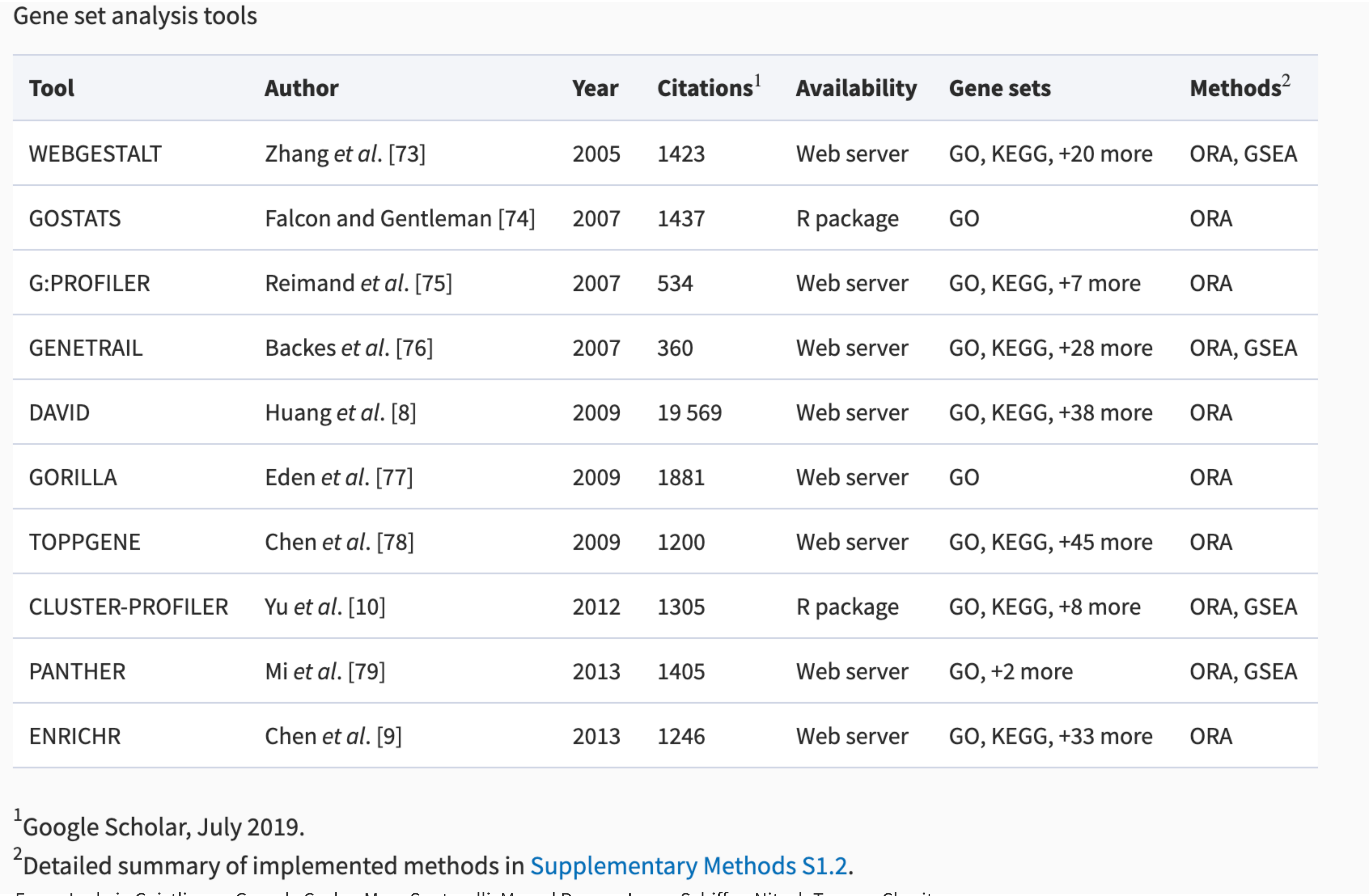
Advaita Bio

Next-Gen Pathway Analysis for gene and protein expression Read More...

Latest Version

1.4.0

Some possible tools



Gene set analysis tools

Tool	Author	Year	Citations[1]	Availability	Gene sets	Methods[2]
WEBGESTALT	Zhang *et al.* [73]	2005	1423	Web server	GO, KEGG, +20 more	ORA, GSEA
GOSTATS	Falcon and Gentleman [74]	2007	1437	R package	GO	ORA
G:PROFILER	Reimand *et al.* [75]	2007	534	Web server	GO, KEGG, +7 more	ORA
GENETRAIL	Backes *et al.* [76]	2007	360	Web server	GO, KEGG, +28 more	ORA, GSEA
DAVID	Huang *et al.* [8]	2009	19 569	Web server	GO, KEGG, +38 more	ORA
GORILLA	Eden *et al.* [77]	2009	1881	Web server	GO	ORA
TOPPGENE	Chen *et al.* [78]	2009	1200	Web server	GO, KEGG, +45 more	ORA
CLUSTER-PROFILER	Yu *et al.* [10]	2012	1305	R package	GO, KEGG, +8 more	ORA, GSEA
PANTHER	Mi *et al.* [79]	2013	1405	Web server	GO, +2 more	ORA, GSEA
ENRICHR	Chen *et al.* [9]	2013	1246	Web server	GO, KEGG, +33 more	ORA

[1] Google Scholar, July 2019.

[2] Detailed summary of implemented methods in Supplementary Methods S1.2.

From: Ludwig Geistlinger, Gergely Csaba, Mara Santarelli, Marcel Ramos, Lucas Schiffer, Nitesh Turaga, Charity Law, Sean Davis, Vincent Carey, Martin Morgan, Ralf Zimmer, Levi Waldron, Toward a gold standard for benchmarking gene set enrichment analysis, *Briefings in Bioinformatics*, Volume 22, Issue 1, January 2021, Pages 545–556, https://doi.org/10.1093/bib/bbz158

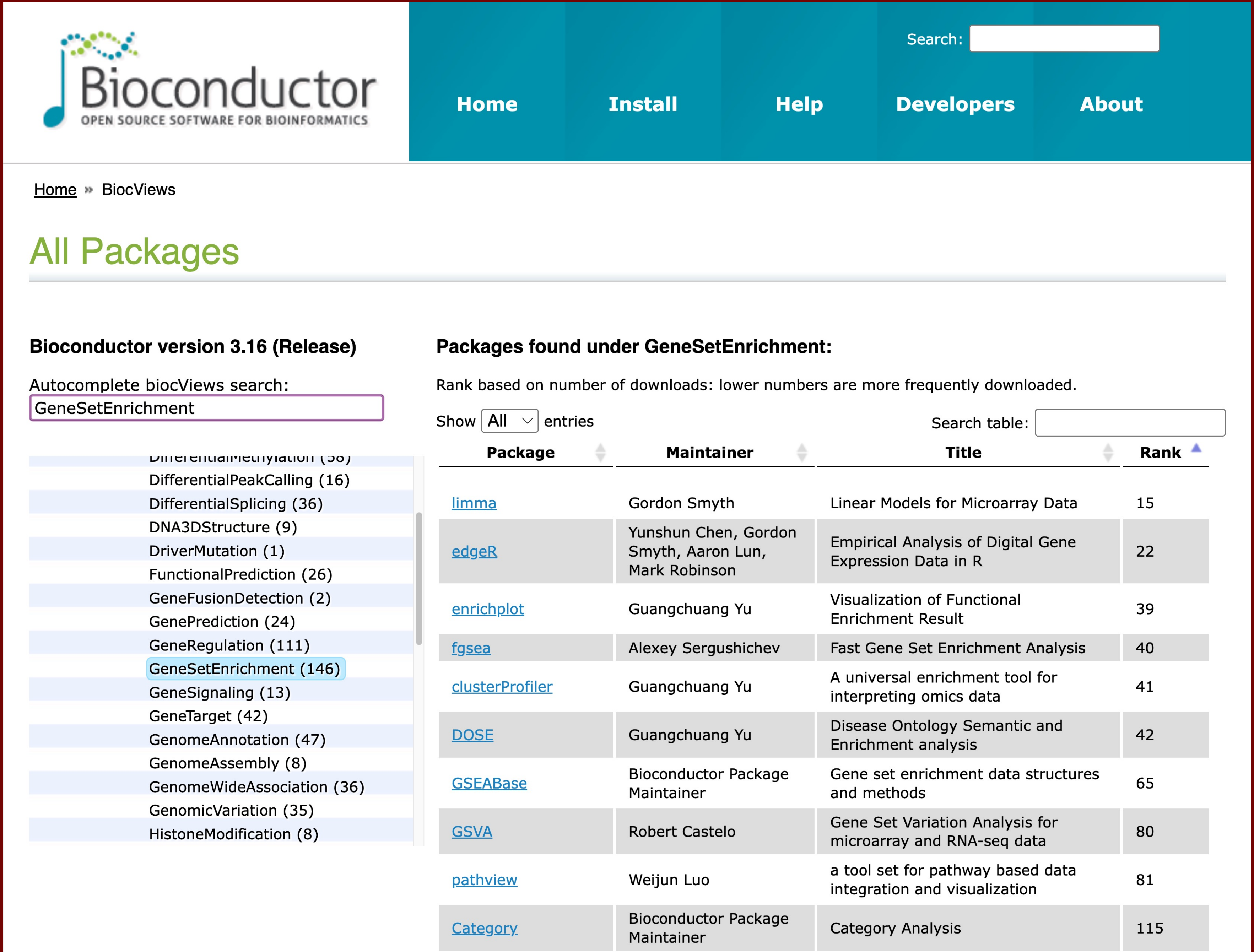
A note about tools and databases

- Not all databases survive
  - Check to see when information was last updated
- Tools also frequently fall to the wayside
  - Check to see if the tool is maintained
  - If the tool is not readily updated, it’s likely it is using outdated versions of databases.

Importance of gene IDs

- Different tools use different gene ID annotations
- May need to convert between annotations using available programs
- Annotations can be genome build specific
- See linked tutorial in the course docs
- Some annotation programs/databases:
  - g:convert
  - BioMart
  - AnnotationHub

R packages



Home » BiocViews

All Packages

Bioconductor version 3.16 (Release)

Autocomplete biocViews search: GeneSetEnrichment

DifferentialMethylation (58)
DifferentialPeakCalling (16)
DifferentialSplicing (36)
DNA3DStructure (9)
DriverMutation (1)
FunctionalPrediction (26)
GeneFusionDetection (2)
GenePrediction (24)
GeneRegulation (111)
GeneSetEnrichment (146)
GeneSignaling (13)
GeneTarget (42)
GenomeAnnotation (47)
GenomeAssembly (8)
GenomeWideAssociation (36)
GenomicVariation (35)
HistoneModification (8)

Packages found under GeneSetEnrichment:

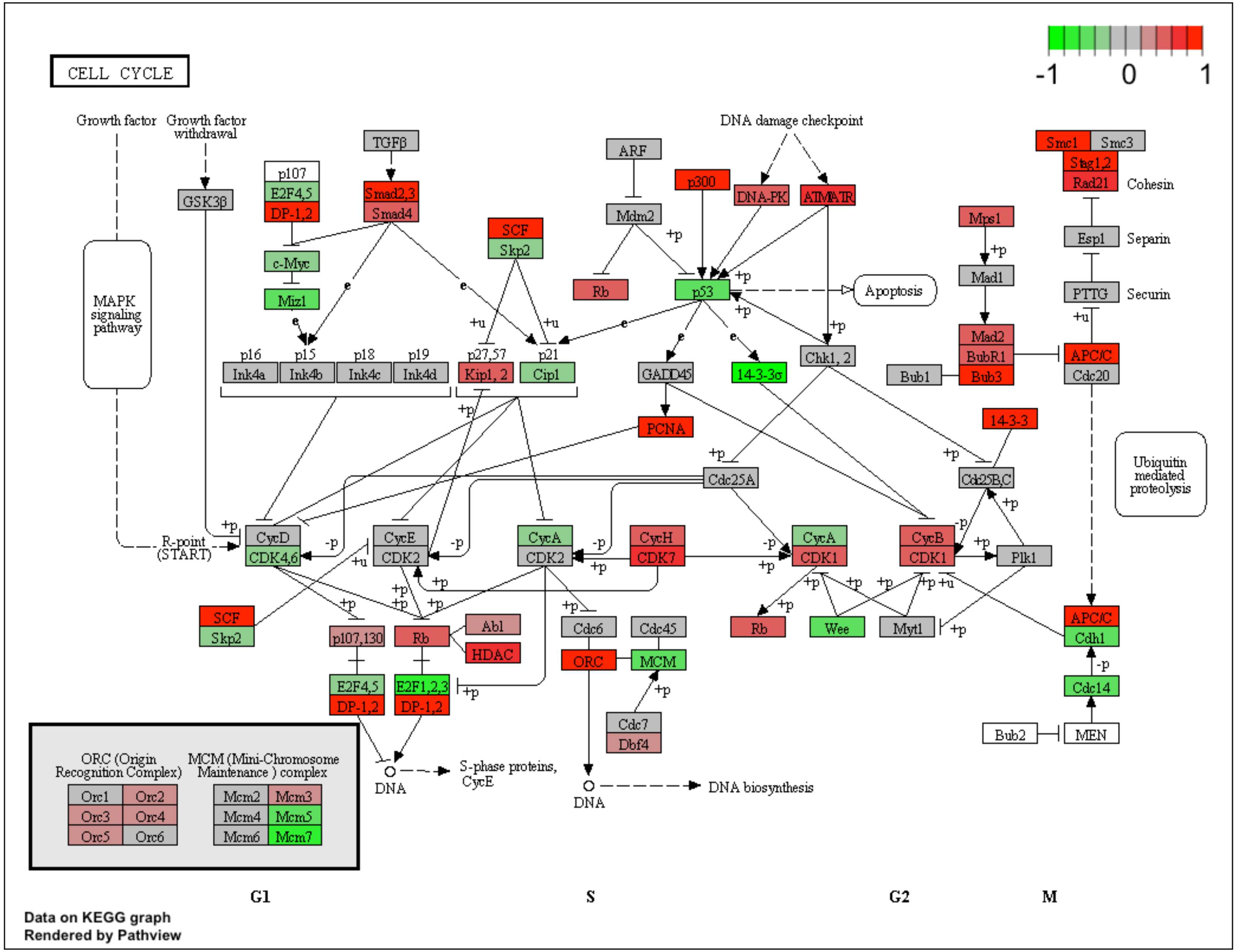
Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries

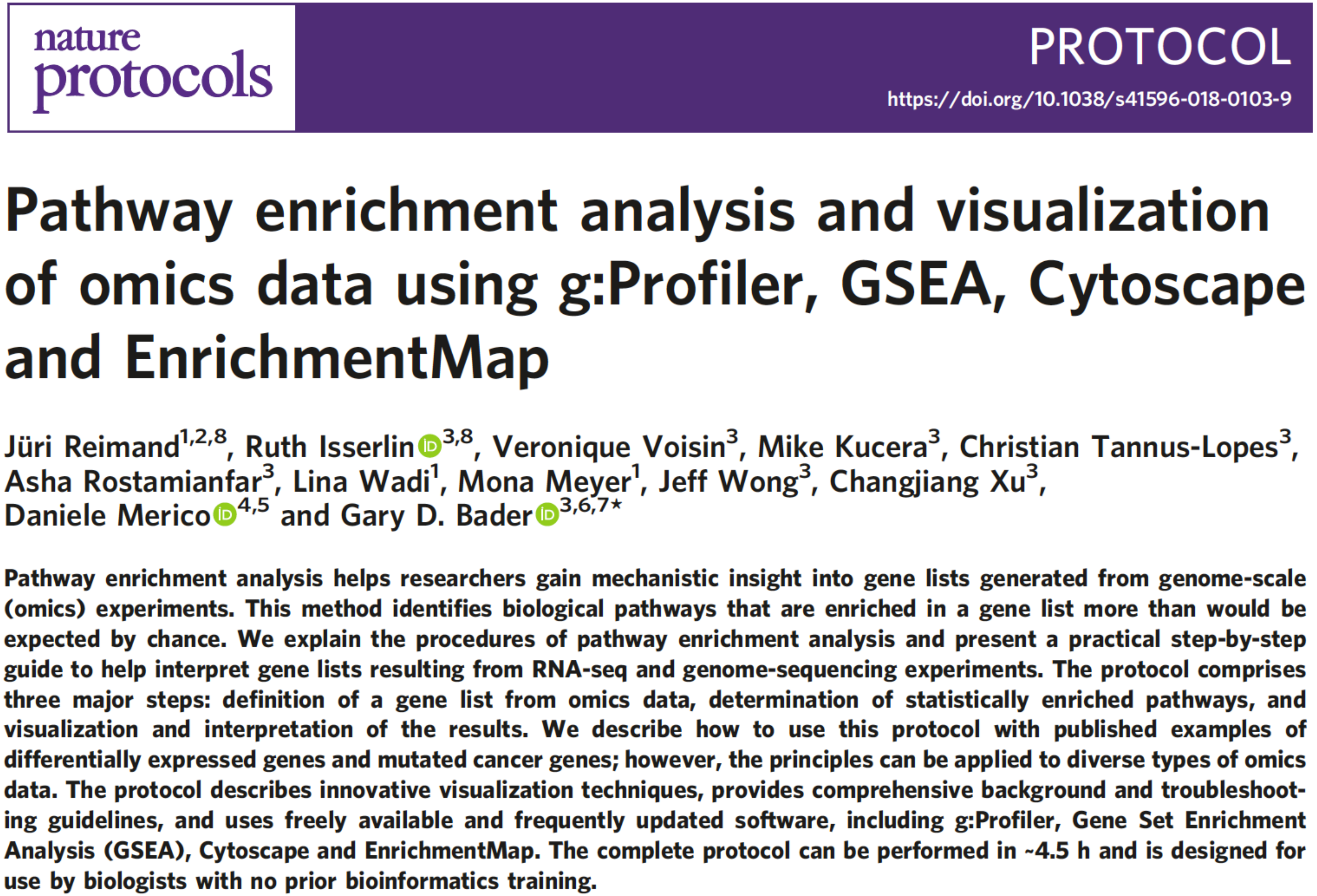
Search table:

Package	Maintainer	Title	Rank
limma	Gordon Smyth	Linear Models for Microarray Data	15
edgeR	Yunshun Chen, Gordon Smyth, Aaron Lun, Mark Robinson	Empirical Analysis of Digital Gene Expression Data in R	22
enrichplot	Guangchuang Yu	Visualization of Functional Enrichment Result	39
fgsea	Alexey Sergushichev	Fast Gene Set Enrichment Analysis	40
clusterProfiler	Guangchuang Yu	A universal enrichment tool for interpreting omics data	41
DOSE	Guangchuang Yu	Disease Ontology Semantic and Enrichment analysis	42
GSEABase	Bioconductor Package Maintainer	Gene set enrichment data structures and methods	65
GSVA	Robert Castelo	Gene Set Variation Analysis for microarray and RNA-seq data	80
pathview	Weijun Luo	a tool set for pathway based data integration and visualization	81
Category	Bioconductor Package Maintainer	Category Analysis	115

Pathview


CELL CYCLE
-1
0
1
Growth factor
Growth factor withdrawal
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
GSK3β
c-Myc
Miz1
SCF
Skp2
ARF
Mdm2
Rb
p300
DNA damage checkpoint
DNA-PK
ATM/ATR
p53
Apoptosis
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Mps1
Mad1
Esp1
Separin
PTTG
Securin
MAPK signaling pathway
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
GADD45
14-3-3σ
Chk1, 2
Mad2
BubR1
Bub3
Bub1
APC/C
Cdc20
PCNA
14-3-3
Cdc25A
Cdc25B,C
Ubiquitin mediated proteolysis
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
Cdc6
Cdc45
ORC
MCM
Rb
Wee
Myt1
APC/C
Cdh1
Cdc14
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc7
Dbf4
Bub2
MEN
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
G1
S
G2
M
Data on KEGG graph
Rendered by Pathview

Nature Protocols 14, 482–517 (2019)



nature protocols

PROTOCOL

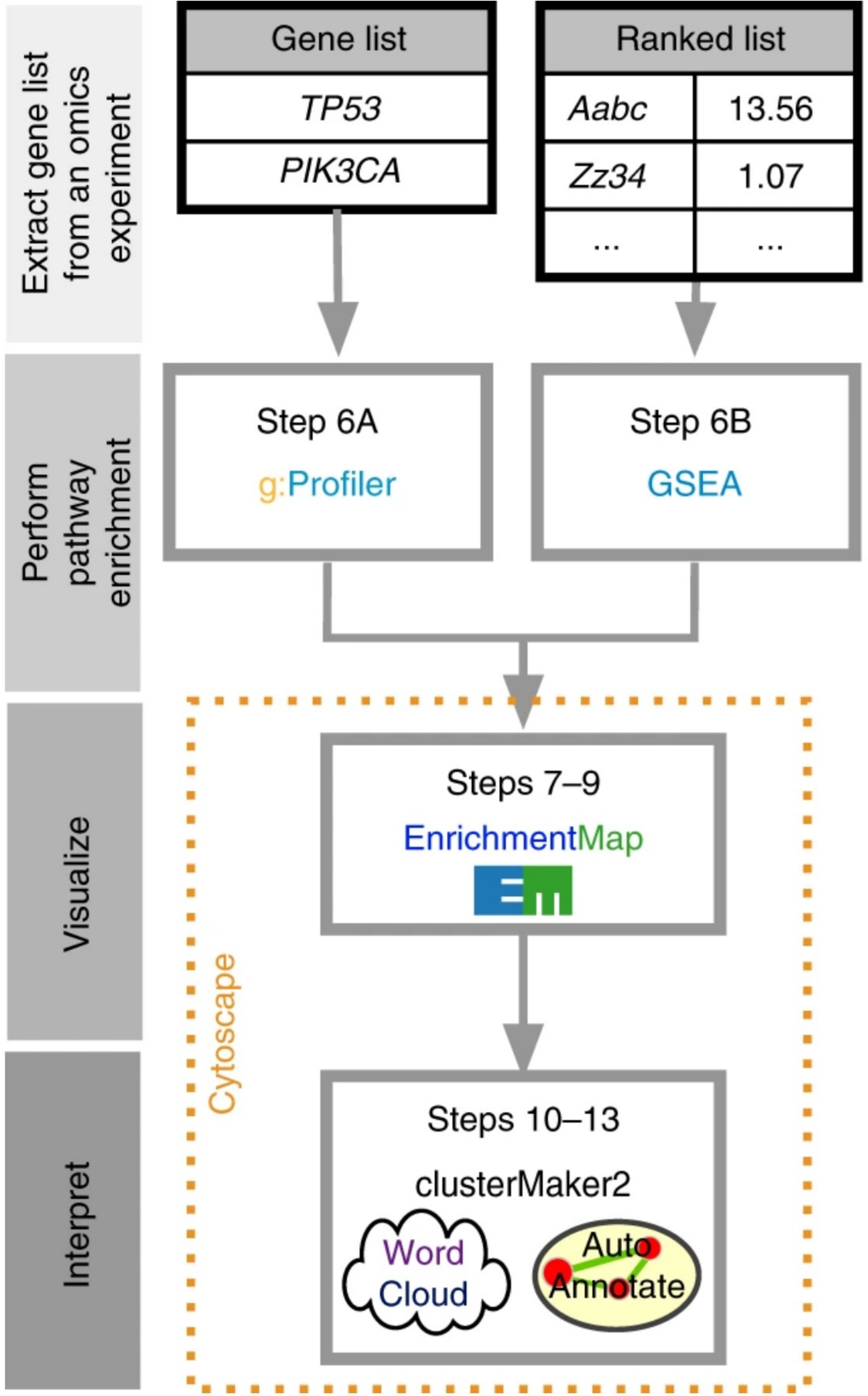
https://doi.org/10.1038/s41596-018-0103-9

Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

Jüri Reimand[1,2,8], Ruth Isserlin[3,8], Veronique Voisin[3], Mike Kucera[3], Christian Tannus-Lopes[3], Asha Rostamianfar[3], Lina Wadi[1], Mona Meyer[1], Jeff Wong[3], Changjiang Xu[3], Daniele Merico[4,5] and Gary D. Bader[3,6,7*]

Pathway enrichment analysis helps researchers gain mechanistic insight into gene lists generated from genome-scale (omics) experiments. This method identifies biological pathways that are enriched in a gene list more than would be expected by chance. We explain the procedures of pathway enrichment analysis and present a practical step-by-step guide to help interpret gene lists resulting from RNA-seq and genome-sequencing experiments. The protocol comprises three major steps: definition of a gene list from omics data, determination of statistically enriched pathways, and visualization and interpretation of the results. We describe how to use this protocol with published examples of differentially expressed genes and mutated cancer genes; however, the principles can be applied to diverse types of omics data. The protocol describes innovative visualization techniques, provides comprehensive background and troubleshooting guidelines, and uses freely available and frequently updated software, including g:Profiler, Gene Set Enrichment Analysis (GSEA), Cytoscape and EnrichmentMap. The complete protocol can be performed in ~4.5 h and is designed for use by biologists with no prior bioinformatics training.

Nature Protocols 14, 482–517 (2019)



Extract gene list from an omics experiment

Gene list
TP53
PIK3CA

Ranked list	
Aabc	13.56
Zz34	1.07
...	...

Perform pathway enrichment

Step 6A
g:Profiler

Step 6B
GSEA

Visualize

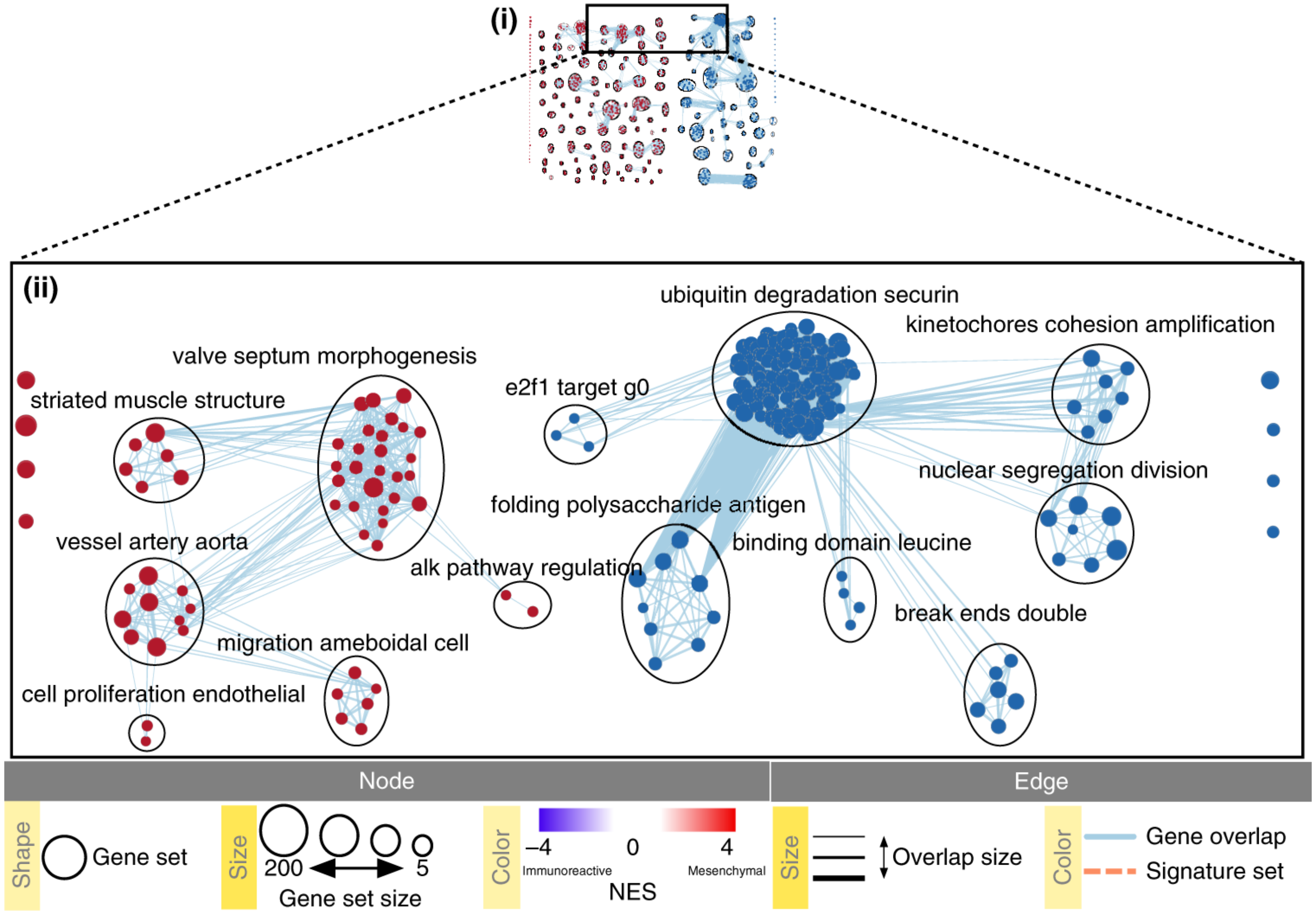
Steps 7–9
EnrichmentMap

Interpret

Steps 10–13
clusterMaker2

Cytoscape

Nature Protocols 14, 482–517 (2019)


(i)
(ii)
ubiquitin degradation securin
kinetochores cohesion amplification
valve septum morphogenesis
e2f1 target g0
striated muscle structure
nuclear segregation division
folding polysaccharide antigen
vessel artery aorta
binding domain leucine
alk pathway regulation
break ends double
migration ameboidal cell
cell proliferation endothelial
Node
Edge
Shape
Gene set
Size
200
5
Gene set size
Color
–4
0
4
Immunoreactive
Mesenchymal
NES
Size
Overlap size
Color
Gene overlap
Signature set