Gene set enrichment

Jeff Leek

@jtleek

www.jtleek.com

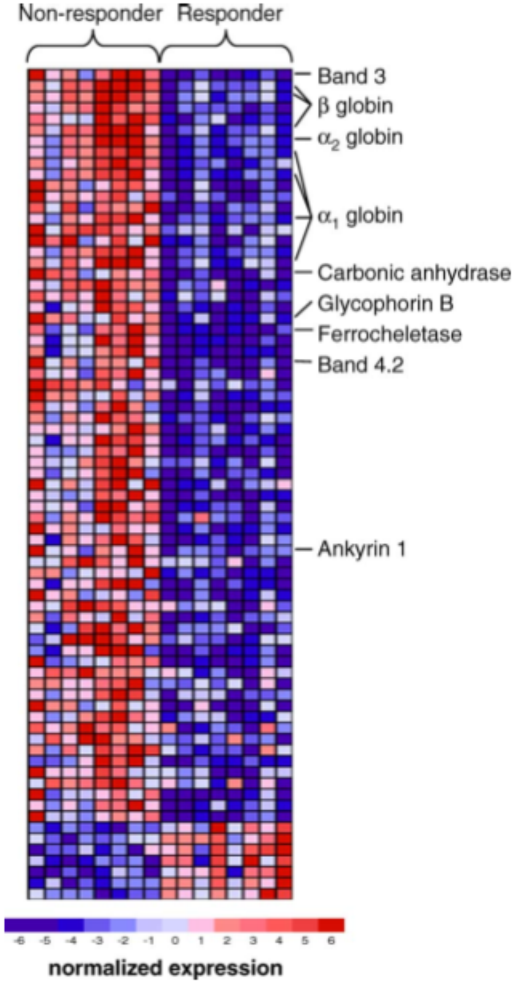
OPEN ACCESS Freely available online

PLoS MEDICINE

An Erythroid Differentiation Signature Predicts Response to Lenalidomide in Myelodysplastic Syndrome

Benjamin L. Ebert[1,2,3☯], Naomi Galili[4☯], Pablo Tamayo[1], Jocelyn Bosco[1,2], Raymond Mak[1,2], Jennifer Pretz[1,2], Shyam Tanguturi[1], Christine Ladd-Acosta[1], Richard Stone[2,3], Todd R. Golub[1,2,5,6], Azra Raza[4*]

1 Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, United States of America, **3** Brigham and Women's Hospital, Department of Medicine, Boston, Massachusetts, United States of America, **4** St. Vincent's Comprehensive Cancer Center, New York, New York, United States of America, **5** Childrens's Hospital, Boston, Massachusetts, United States of America, **6** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

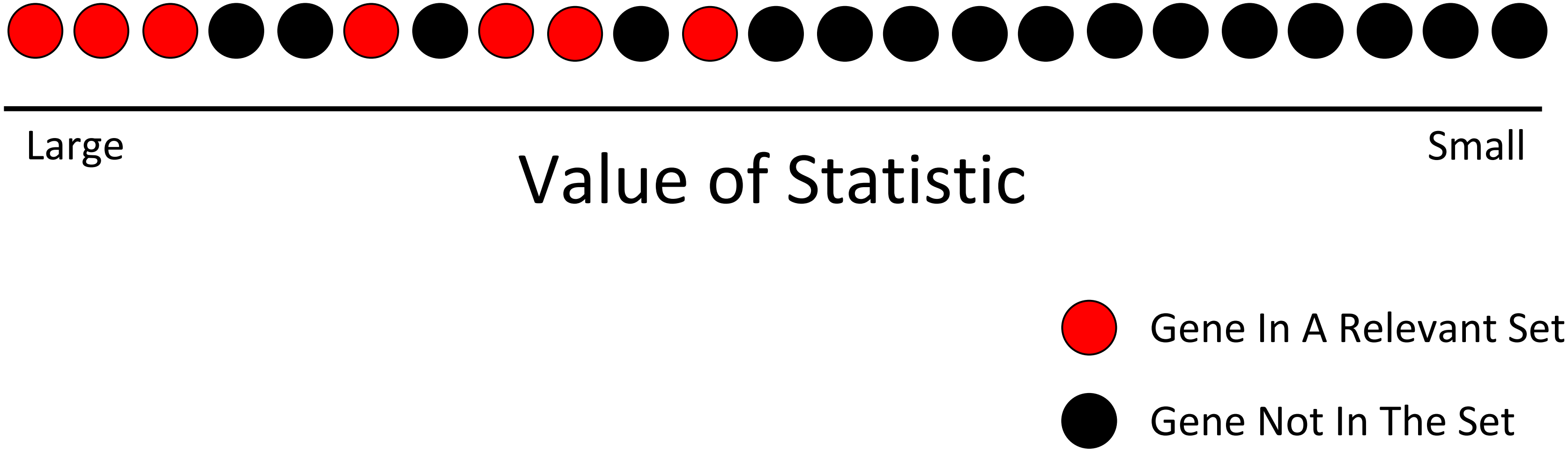


47 Genes at FDR $\leq$ 10%

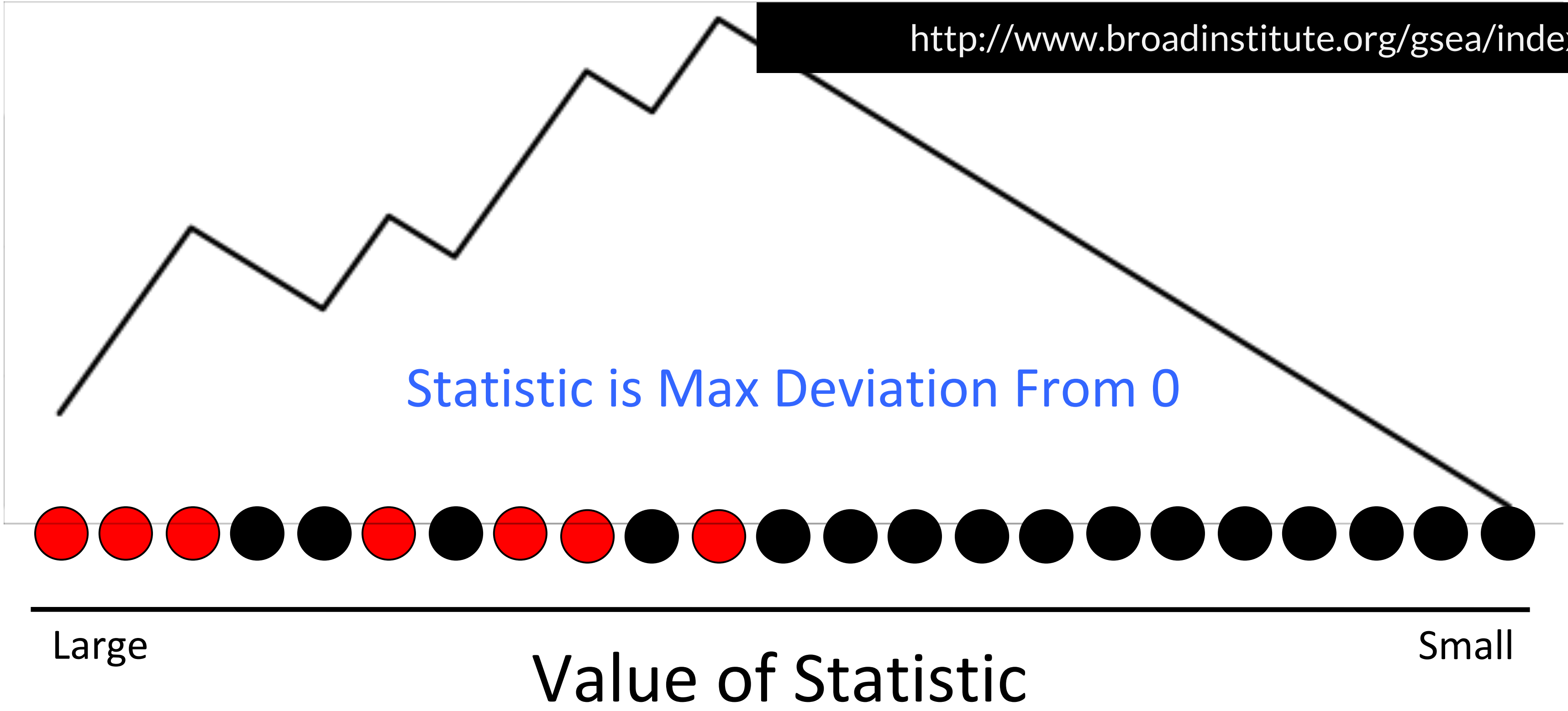
Gene set enrichment analysis

- Looks for groups of genes that share function or other characteristics

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



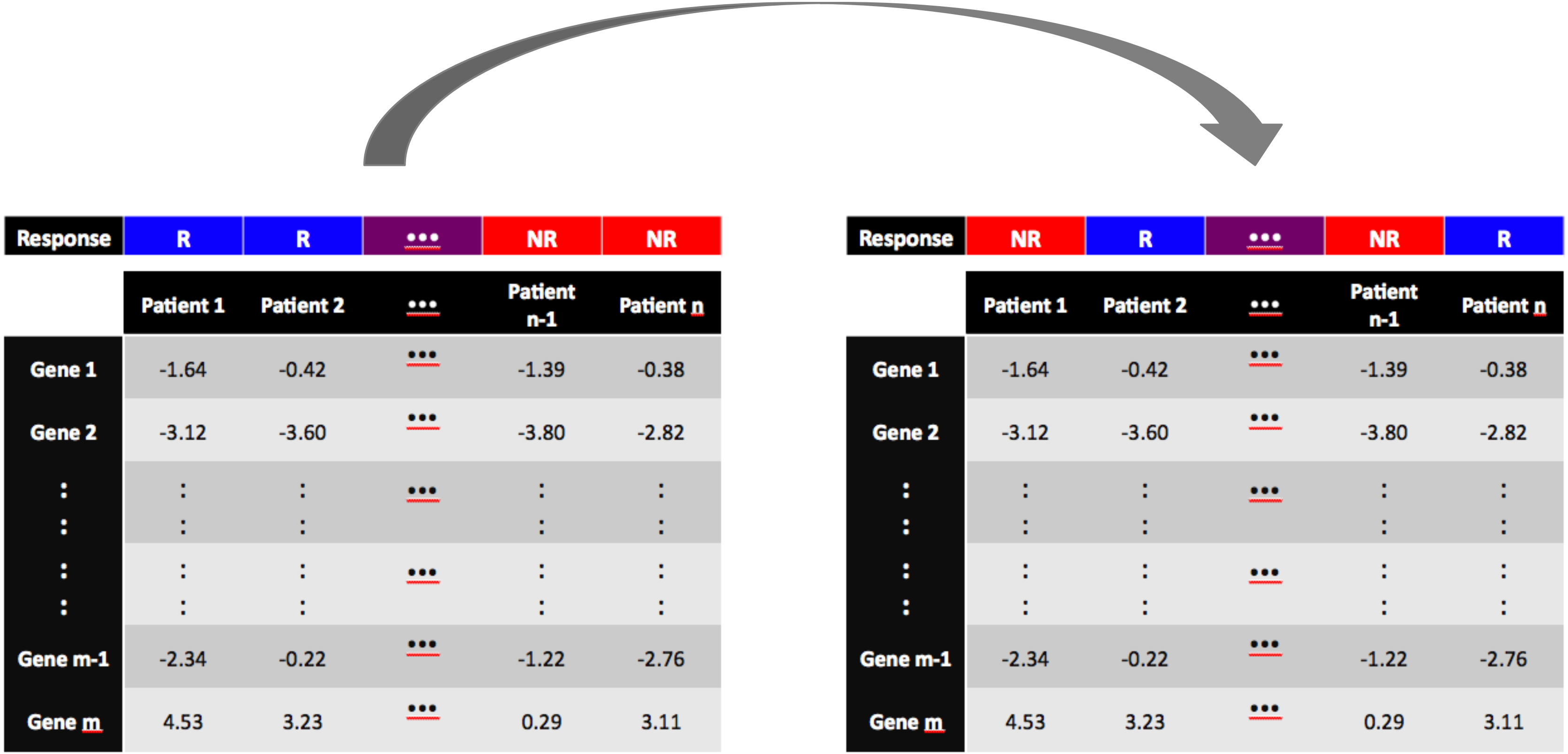
http://www.broadinstitute.org/gsea/index.jsp



🔴 Gene In A Relevant Set

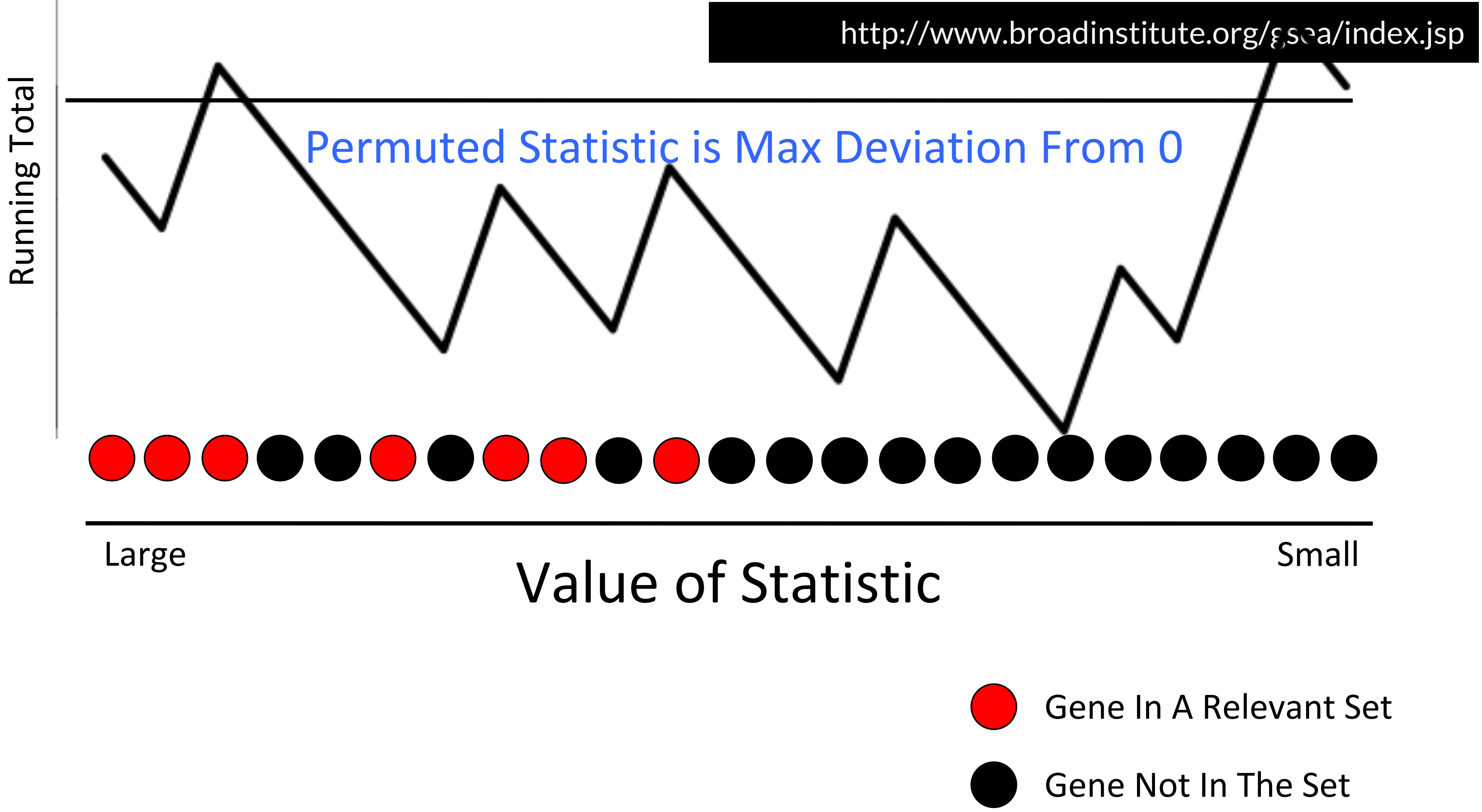
● Gene Not In The Set

Measuring significance with permutation



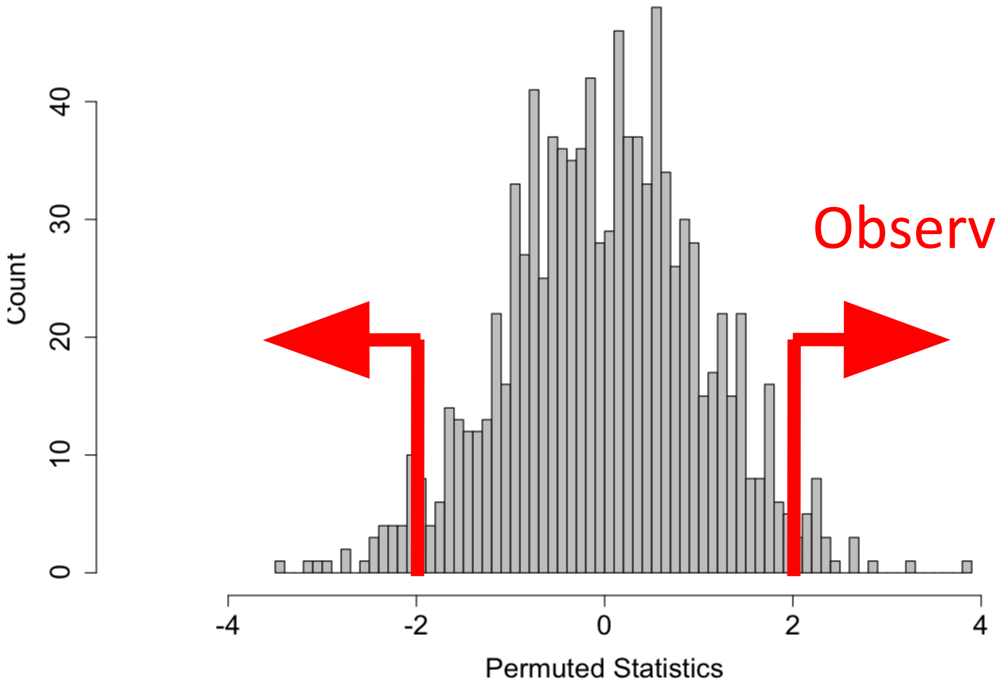
Response	R	R	•••	NR	NR
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
: :	: :	: :	•••	: :	: :
: :	: :	: :	•••	: :	: :
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11

Response	NR	R	•••	NR	R
	Patient 1	Patient 2	•••	Patient n-1	Patient n
Gene 1	-1.64	-0.42	•••	-1.39	-0.38
Gene 2	-3.12	-3.60	•••	-3.80	-2.82
: :	: :	: :	•••	: :	: :
: :	: :	: :	•••	: :	: :
Gene m-1	-2.34	-0.22	•••	-1.22	-2.76
Gene m	4.53	3.23	•••	0.29	3.11



Permuted Statistic is Max Deviation From 0

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

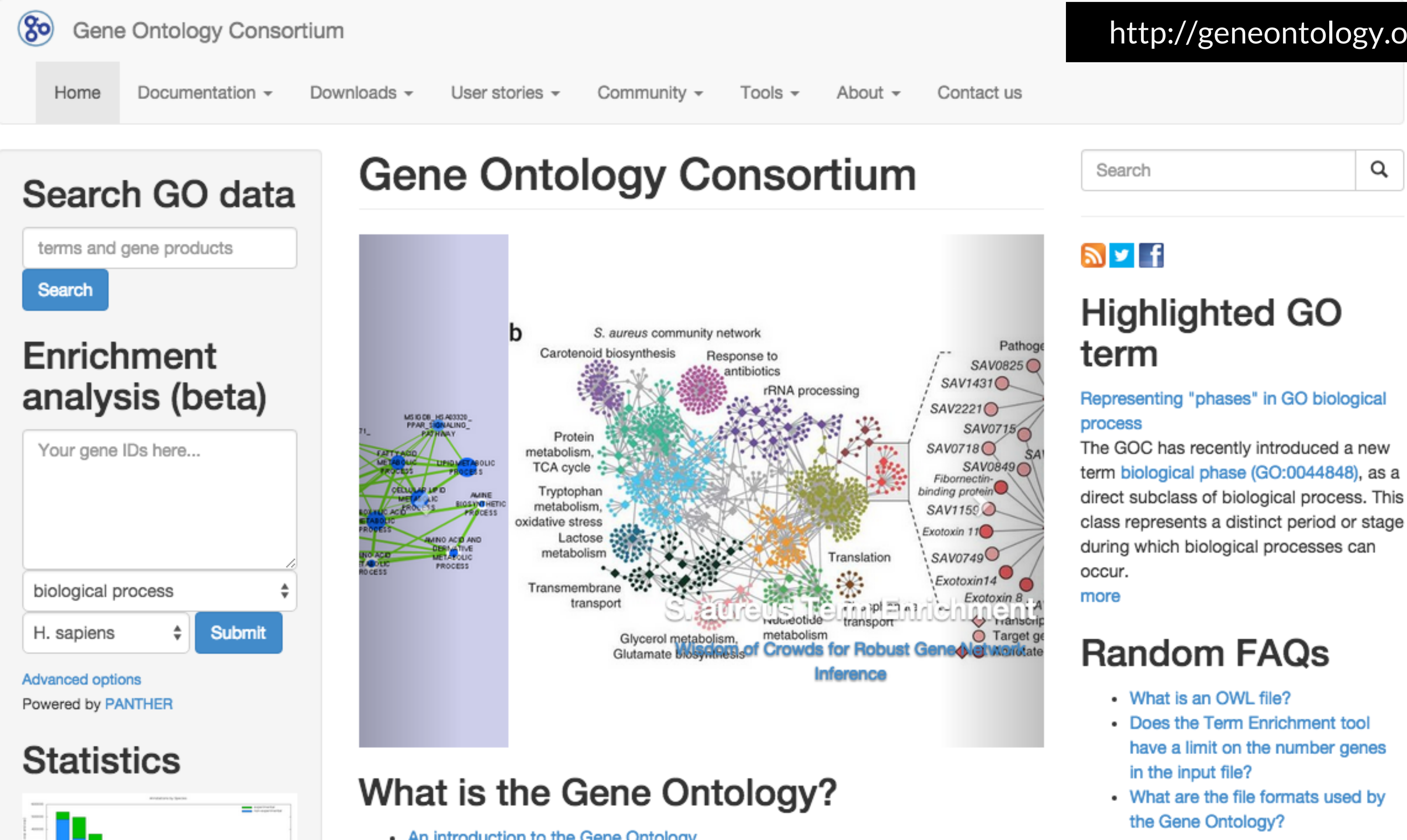


Observed Statistic = 2

$$\text{P-value} = \frac{\{\# \mid S^{perm} \mid \geq \mid S^{obs} \mid \}}{\text{\# of Permutations}}$$

What gene sets?

http://geneontology.org/



Search GO data

Enrichment analysis (beta)

Powered by PANTHER

Statistics

Gene Ontology Consortium

b
S. aureus community network
Carotenoid biosynthesis
Response to antibiotics
rRNA processing
Protein metabolism, TCA cycle
Tryptophan metabolism, oxidative stress
Lactose metabolism
Transmembrane transport
Glycerol metabolism, Glutamate biosynthesis
Nucleotide metabolism
Translation
transport
Pathoge
SAV0825
SAV1431
SAV2221
SAV0715
SAV0718
SAV0849
Fibronectin-binding protein
SAV1159
Exotoxin 11
SAV0749
Exotoxin14
Exotoxin 8
Target ge
FATTY ACID METABOLIC PROCESS
LIPID METABOLIC PROCESS
CELLULAR LIPID METABOLIC PROCESS
AMINE BIOSYNTHETIC PROCESS
AMINO ACID AND DERIVATIVE METABOLIC PROCESS
MSIGDB_HSA03320_PPAR_SIGNALING_PATHWAY

S. aureus Term Enrichment

Wisdom of Crowds for Robust Gene Network Inference

What is the Gene Ontology?

- An introduction to the Gene Ontology

Highlighted GO term

Representing "phases" in GO biological process

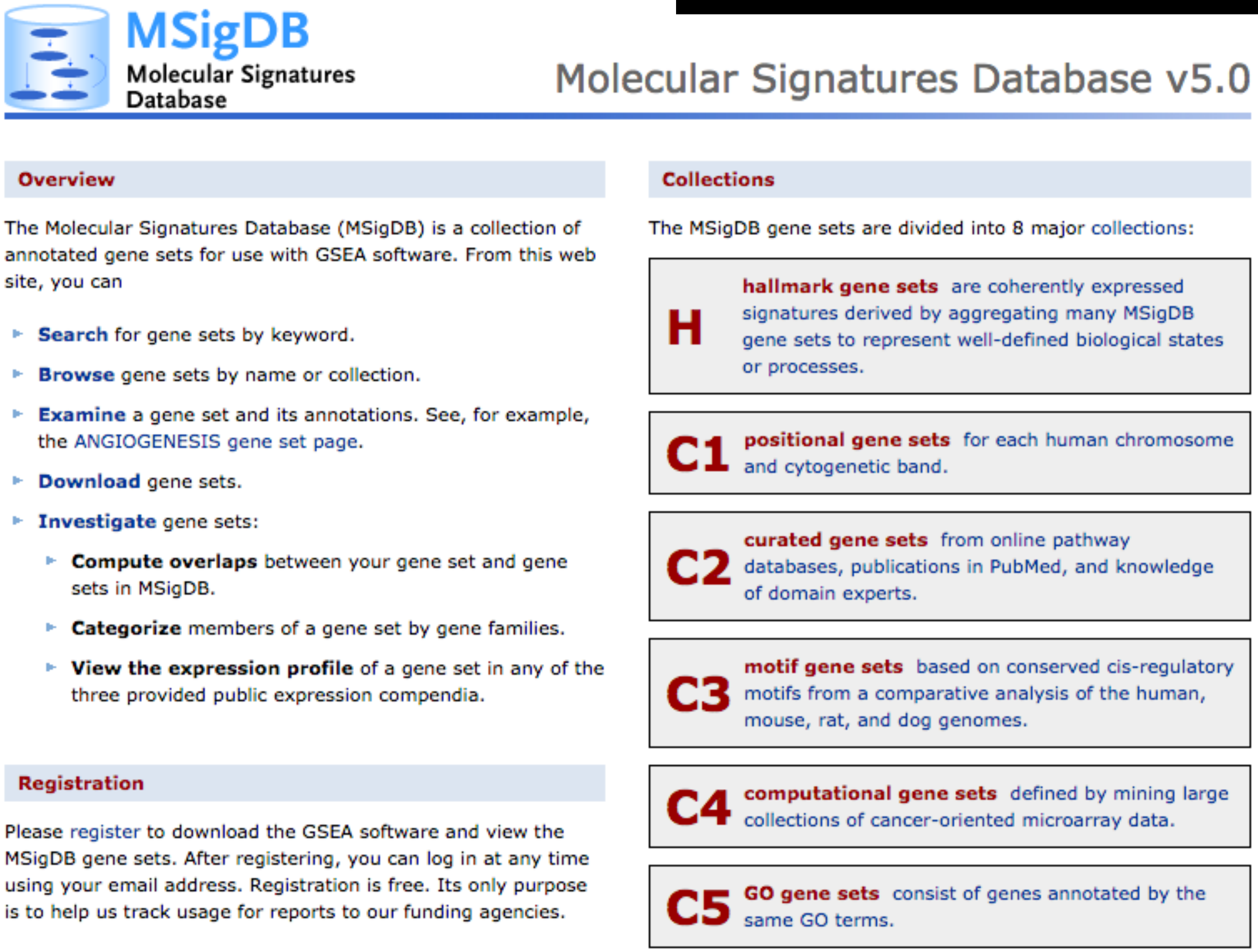
The GOC has recently introduced a new term biological phase (GO:0044848), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur.

more

Random FAQs

- What is an OWL file?
- Does the Term Enrichment tool have a limit on the number genes in the input file?
- What are the file formats used by the Gene Ontology?

http://www.broadinstitute.org/gsea/msigdb/index.jsp



MSigDB Molecular Signatures Database

Molecular Signatures Database v5.0

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **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 any of the three provided public expression compendia.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. 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.

Collections

The MSigDB gene sets are divided into 8 major 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** for each human chromosome and cytogenetic band.

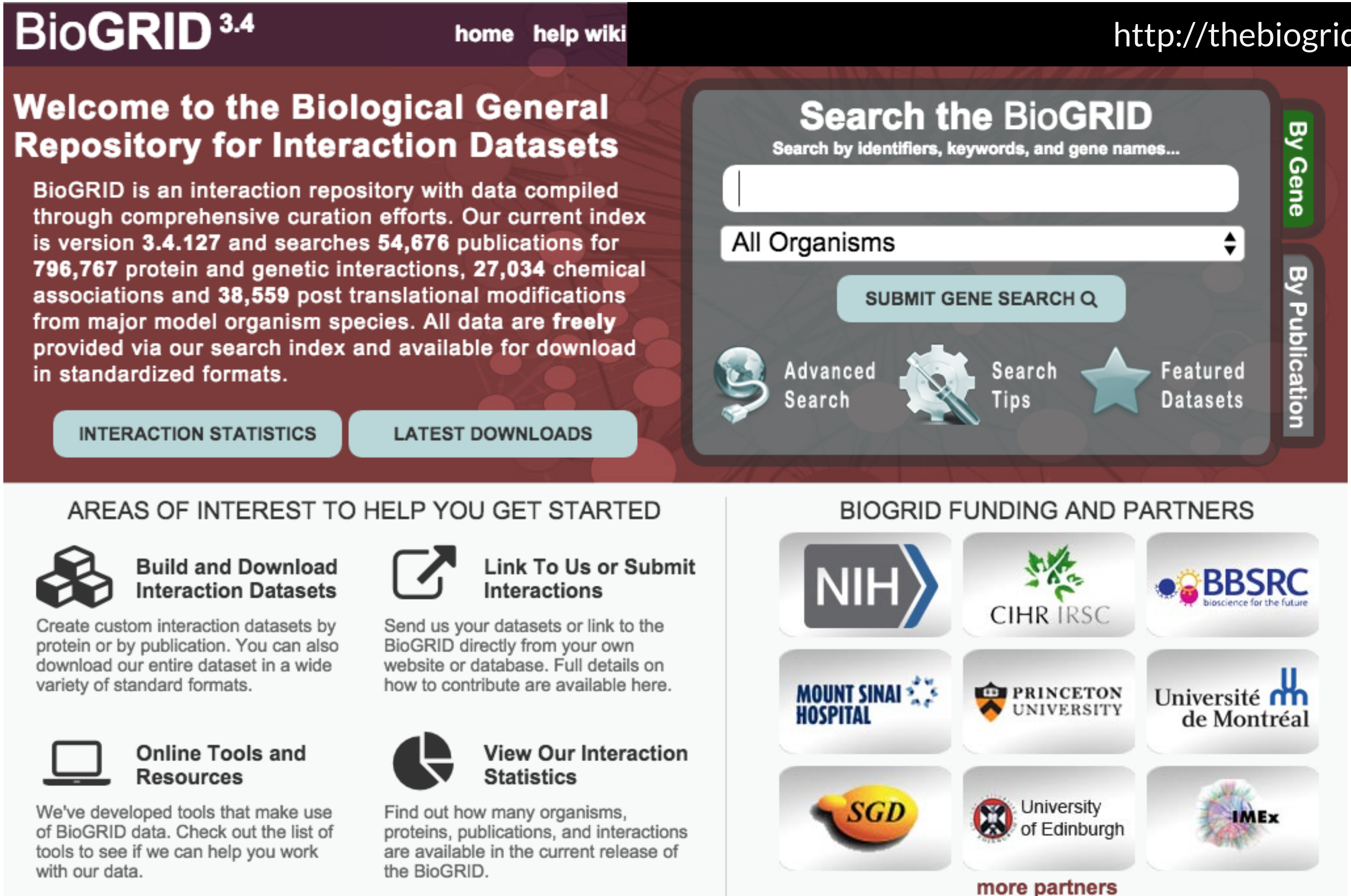
C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

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

C5 **GO gene sets** consist of genes annotated by the same GO terms.

http://thebiogrid.org/



BioGRID 3.4

home help wiki

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 **3.4.127** and searches **54,676** publications for **796,767** protein and genetic interactions, **27,034** 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

Advanced Search

Search Tips

Featured Datasets

By Gene

By Publication

AREAS OF INTEREST TO HELP YOU GET STARTED

Build and Download Interaction Datasets

Create custom interaction datasets by protein or by publication. You can also download our entire dataset in a wide variety of standard formats.

Link To Us or Submit Interactions

Send us your datasets or link to the BioGRID directly from your own website or database. Full details on how to contribute are available here.

Online Tools and Resources

We've developed tools that make use of BioGRID data. Check out the list of tools to see if we can help you work with our data.

View Our Interaction Statistics

Find out how many organisms, proteins, publications, and interactions are available in the current release of the BioGRID.

BIOGRID FUNDING AND PARTNERS

NIH

CIHR IRSC

BBSRC bioscience for the future

MOUNT SINAI HOSPITAL

PRINCETON UNIVERSITY

Université de Montréal

SGD

University of Edinburgh

IMEx

more partners

Notes and further reading

- Sometimes still very hard to interpret, especially if the categories are broad/vague
- It is easy to “tell stories” if you aren’t careful
- Incurs a second multiple testing problem
- Can be simplified
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3134237/