Gene Set Analysis

GSA

What is a Gene Set?

Most common gene sets

How can we relate our results to gene sets?

Differential Expression → GSEA

What is a Gene Set?

Pathway

Wikipedia: "In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions.

Other meanings (uses of the word): e.g. gene regulatory networks etc…

A first intuitive idea of gene set, others…

- Genes involved in a pathway
- Genes corresponding to a Gene Ontology term
- Genes associated to a disease

Gene Set Analysis

Differential Expression → GSEA

What is a Gene Set?

GO

the Gene Ontology

Search

gene or protein name | go!

Downloads | Tools | Documentation | Projects | About | Contact

Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology...

Quick Links

- Tools
- AmiGO browser
- Submit GO Annotations

Search t

AmiGO is

The Gene
annotatior
also very

Cellular component

A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer). See the Documentation on the cellular component ontology for more details.

Biological process

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Further information can be found in the process ontology documentation.

Molecular function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word "activity". The documentation on the function ontology explains more about GO functions and the rules governing them.

Gene Set Analysis

library(GOstats)

GOstats

Tools for manipulating GO and microarrays.

Bioconductor version: Release (2.12)

A set of tools for interacting with GO and microarray data. A variety of basic manipulation tools for graphs, hypothesis testing and other simple calculations.

Author: R. Gentleman and S. Falcon

library(GO.db)

GO.db

A set of annotation maps describing the entire Gene Ontology

Bioconductor version: Release (2.12)

A set of annotation maps describing the entire Gene Ontology assembled using data from GO

Author: Marc Carlson

Gene Set Analysis

Differential Expression → GSEA

Most common gene sets

others

KEGG

Reactome

MSigDB

reactome.db

A set of annotation maps for reactome

Bioconductor version: Release (2.12)

A set of annotation maps for reactome assembled using data from reactome

Author: Willem Ligtenberg

GSEA
Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

- MSigDB Home
- About Collections
- Browse Gene Sets
- Search Gene Sets
- Investigate Gene Sets
- View Gene Families
- Help

MSigDB
Molecular Signatures Database

Molecular Signatures Database v3.1

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

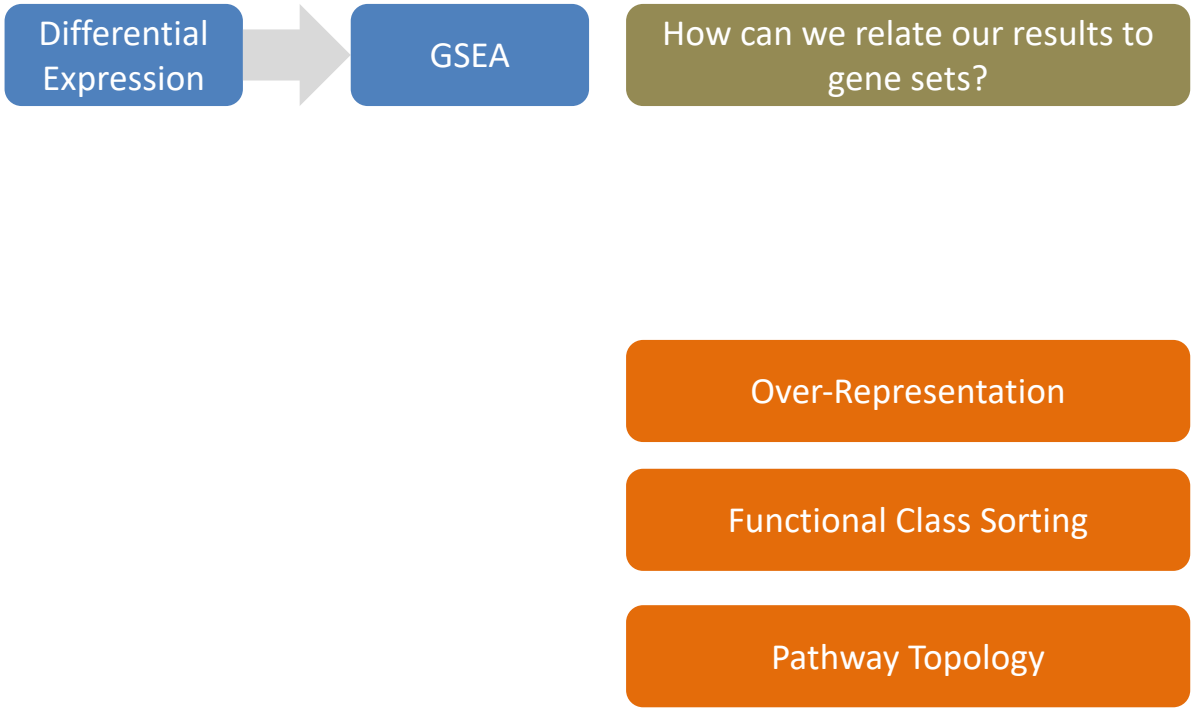
Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Collections

The MSigDB gene sets are divided into 6 major collections:

- **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.
- **c6** **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

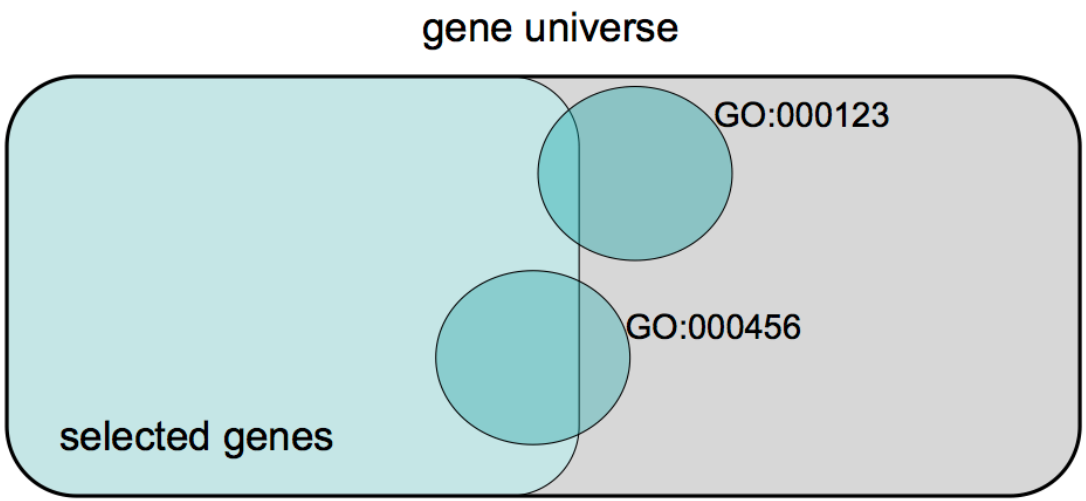
Gene Set Analysis



Over-Representation

hyperGTest

http://marray.economia.unimi.it/2007/material/day4/Lecture7.pdf



... the hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement.

hyperGTest

```
> hgDfList
[[1]]
      GOBPID       Pvalue OddsRatio   ExpCount Count Size                                                          Term
1 GO:0002566 0.0001916482 130.81988 0.02130077     2    9 somatic diversification of immune receptors via somatic mutation
2 GO:0016446 0.0001916482 130.81988 0.02130077     2    9                     somatic hypermutation of immunoglobulin genes
3 GO:0006298 0.0008972148  53.81586 0.04496829     2   19                                                   mismatch repair
4 GO:0010039 0.0009954607  50.82126 0.04733504     2   20                                                 response to iron ion

[[2]]
      GOMFID       Pvalue OddsRatio   ExpCount Count Size                   Term
1 GO:0030983 0.0005042207  74.68531 0.03395839     2   13 mismatched DNA binding
```

Gene Set Analysis

Over-Representation

Functional Class Sorting

Functional Class Sorting

geneSetTest

Similar to Gene Set Enrichment Analysis introduced by Mootha et al (2003), but the statistical tests used are different

Detect differential expression for a group of genes, even when the effects are too small or there is too little data to detect the genes individually

- alternative=="**up**" means the genes in the set tend to be up-regulated, with positive t-statistics.
- alternative=="**down**" means the genes in the set tend to be down-regulated, with negative t-statistics.
- alternative=="**either**" means the set is either up or down-regulated as a whole.
- alternative=="**mixed**" test whether the genes in the set tend to be differentially expressed, without regard for direction

ranks.only=TRUE only the ranks of the statistics are used.

- p-value is obtained from a Wilcoxon test.
- ranks.only is FALSE, then the p-value is obtained by simulation using nsim random selected sets of genes.

Gene Set Analysis

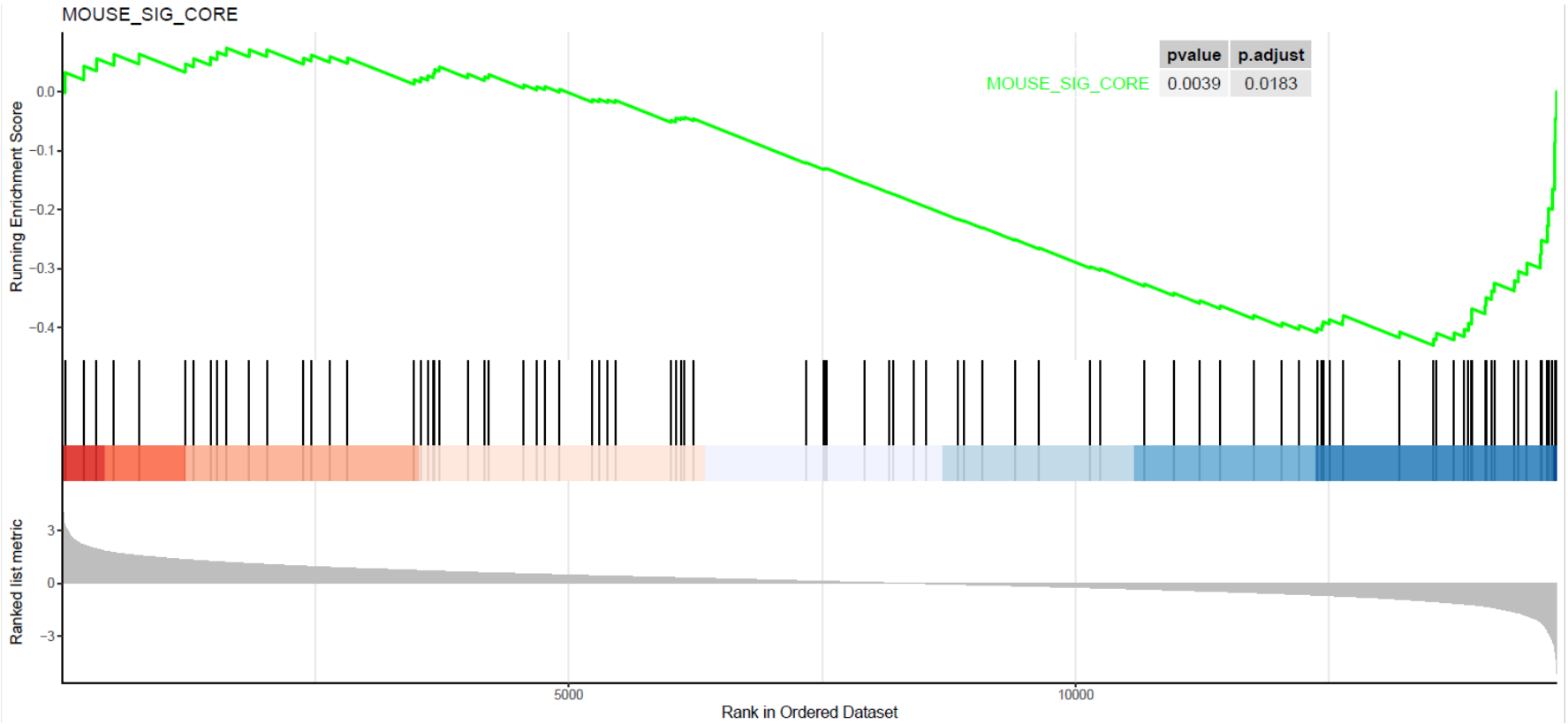


Figure generated by Miren

GSEA
Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

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.

What's New

08-Apr-2013: Version 2.0.12 of the GSEA desktop application is now available. Version 3.87 of the public web site is now available, which includes a number of bug fixes and enhancements of the Compute Overlaps tool. Please refer to the release notes for further details.

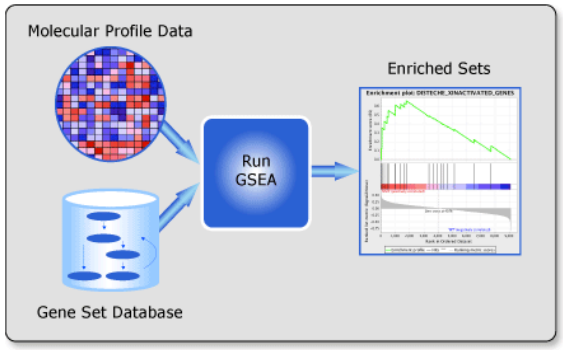
17-Jan-2013: Version 2.0.10 of the GSEA desktop application is now available. This version fixes recent FTP access issues.

15-Oct-2012: Version 3.1 of the Molecular Signatures Database (MSigDB) is now available. Highlights include:

1. more than 1,000 new gene sets curated from publications,
2. a new collection of gene sets representing oncogenic pathway activation modules,
3. two new sources of gene sets representing canonical pathways, and
4. an improved mapping to common gene identifiers for all gene sets.

See the MSigDB 3.1 Release Notes for details. A minor update of the GSEA desktop application has also been released. See the GSEA 2.0.8 Release Notes for details.

01-Oct-2012: We recently submitted a manuscript to Statistical Methods in Medical Research which provides a systematic comparison of the GSEA method with other methods employing a "simpler" t-test assessment of enrichment.



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.

Contributors

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Gene Set Analysis

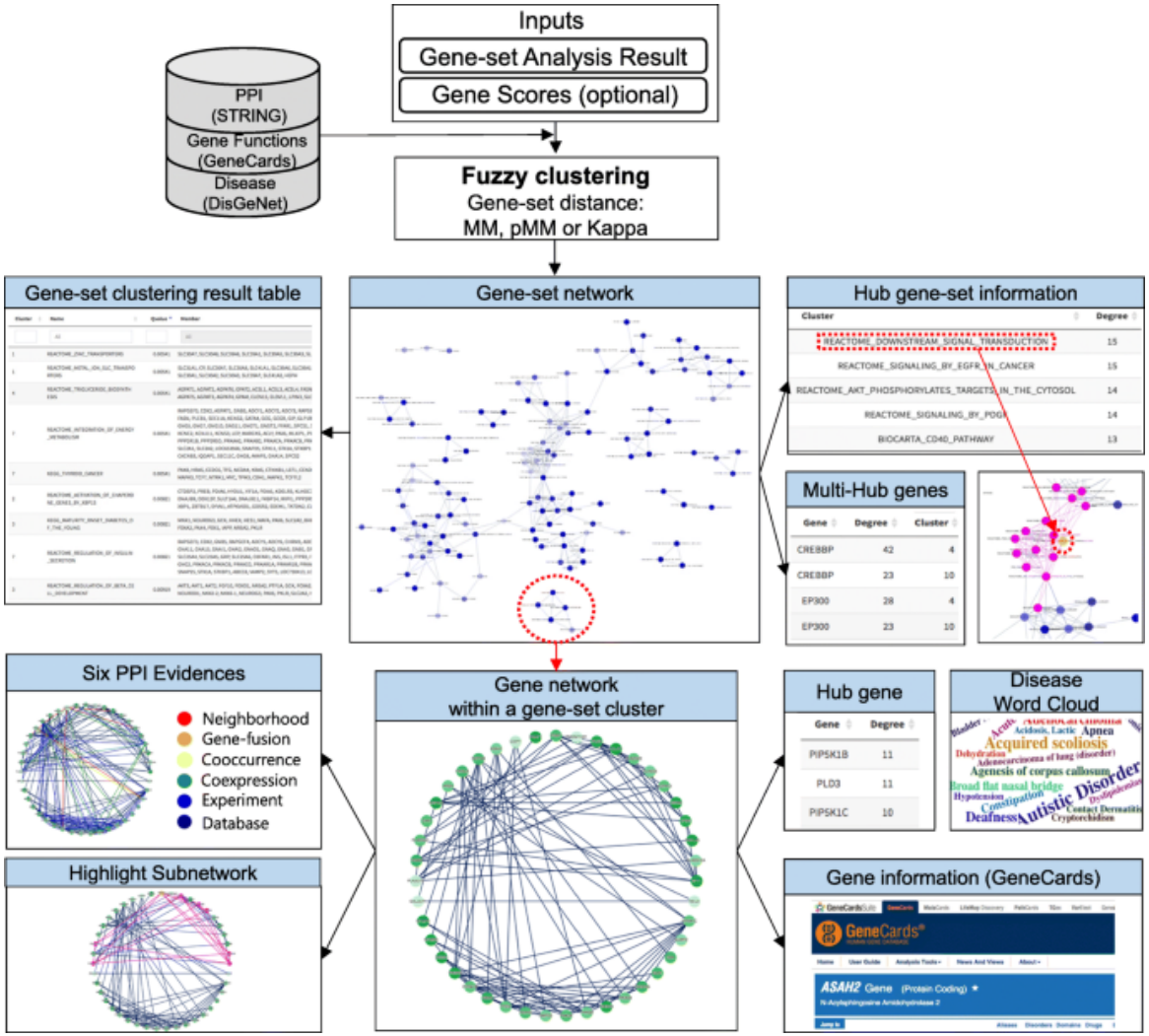
Over-Representation

Functional Class Sorting

Pathway Topology

Gene Set Analysis

”Pathway” Topology

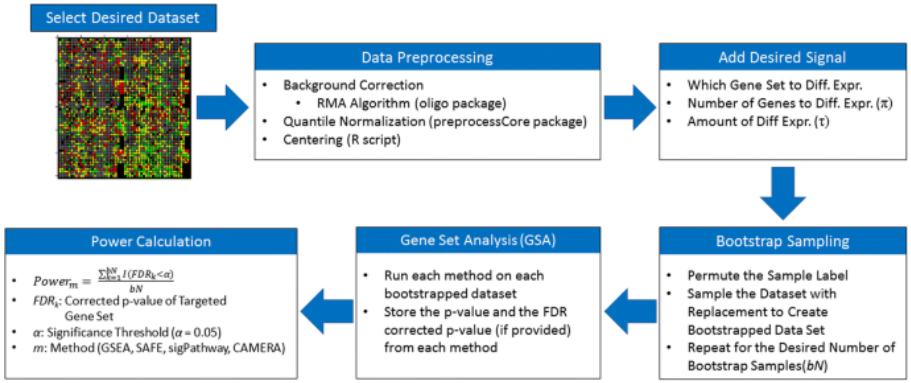


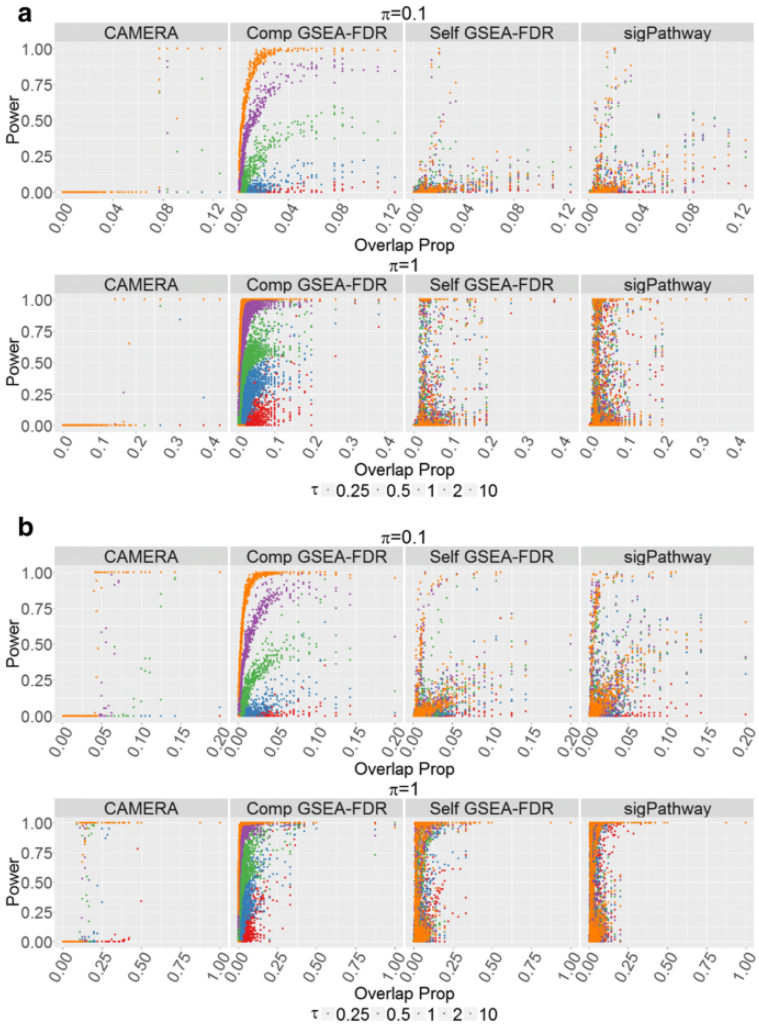
Research | Open Access | Published: 31 May 2018

Gene set analysis methods: a systematic comparison

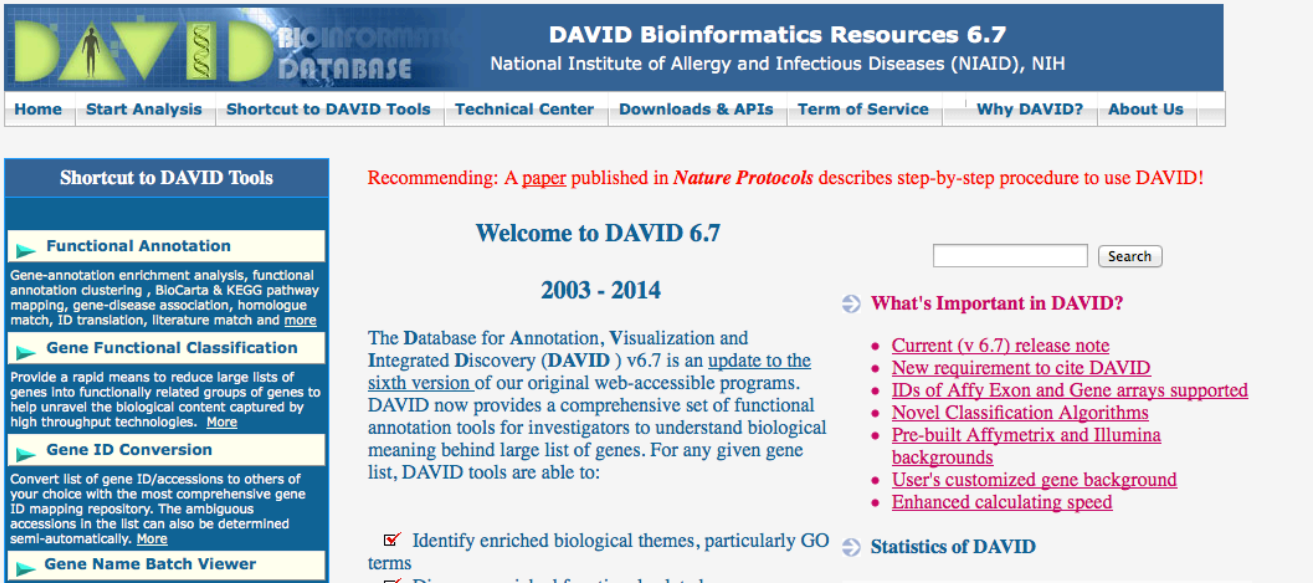
Ravi Mathur, Daniel Rotroff, Jun Ma, Ali Shojaie & Alison Motsinger-Reif

BioData Mining **11**, Article number: 8 (2018) | Cite this article





Gene Set Analysis



Multicontrast: keep contrast separated -> summarize

Software | Open Access | Published: 07 October 2020

GeneSetCluster: a tool for summarizing and integrating gene-set analysis results

Ewoud Ewing, Nuria Planell-Picola, Maja Jagodic & David Gomez-Cabrero

BMC Bioinformatics **21**, Article number: 443 (2020) | Cite this article

Distances

The pipeline then calculates the distance between gene-sets using *CombineGeneSets*. The pipeline default setting is the relative risk (RR), taken from comorbidity statistics [13], using the formula $RR_{ij} = \frac{C_{ij}/N}{(P_i P_j - C_{ij})/N} = \frac{C_{ij} N}{P_i P_j - C_{ij}}$. Where C_{ij} is the overlap between molecules of pathway 1 and pathway 2, N is the total number of genes in the experiments, Pi is the molecules of pathway 1 and Pj is the molecules of pathway 2. The other options available are the Jaccard index, which represents percentage overlap, and Cohen's Kappa, which represents the level of agreement between the gene sets. Moreover, the pipeline allows the user to supply their own distancing function if desired.