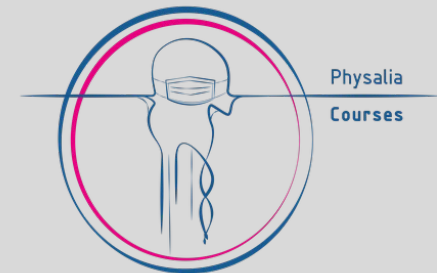


Gene ontology, gene set over-representation analyses and public databases

Peeking into functional roles of gene sets

Epigenomics Data Analysis
Jacques Serizay
Physalia 2023



Why do we need GO analyses?

After RNA-seq (or ATAC, ChIP, ...), we generally end up with a list of genes (or genomic features with the nearest gene).

→ We want to investigate the biological relevance of this set of genes.

GO over-representation analyses: what

An ontology is a **formal representation** of a **body of knowledge** **within a given domain**

GO over-representation analyses: what

An ontology term primarily consists of:

- A definition of a concept
- A representation of this concept
- A formal naming of this concept

GO over-representation analyses: what

An ontology term primarily consists of:

- A definition of a concept
- A representation of this concept
- A formal naming of this concept

=== Example term ===

:id:	GO:0000016
:name:	lactase activity
:ontology:	molecular_function
:def:	"Catalysis of the reaction: lactose + H ₂ O=D-glucose + D-galactose." [EC:3.2.1.108]
:synonym:	"lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
:synonym:	"lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
:xref:	EC:3.2.1.108
:xref:	MetaCyc:LACTASE-RXN
:xref:	Reactome:20536
:is_a:	GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds

GO over-representation analyses: what

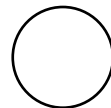
An ontology term primarily consists of:

- A definition of a concept
- A representation of this concept
- A formal naming of this concept

=== Example term ===

:id: GO:0000016
:name: lactase activity
:ontology: molecular_function
:def: "Catalysis of the reaction: lactose + H₂O=D-glucose + D-galactose." [EC:3.2.1.108]
:synonym: "lactase-phlorizin hydrolase activity"
BROAD [EC:3.2.1.108]
:synonym: "lactose galactohydrolase activity" EXACT
[EC:3.2.1.108]
:xref: EC:3.2.1.108
:xref: MetaCyc:LACTASE-RXN
:xref: Reactome:20536
:is_a: GO:0004553 ! hydrolase activity,
hydrolyzing O-glycosyl compounds

GO:0000016



GO over-representation analyses: what

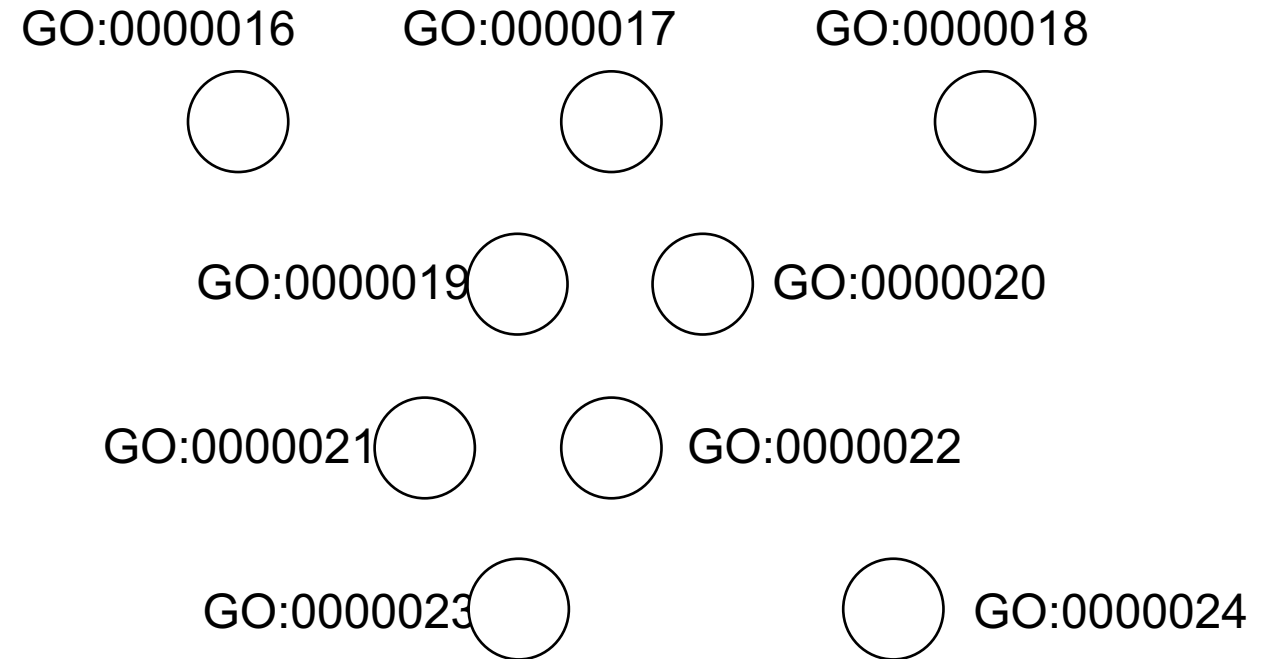
An ontology term can be further enriched with additional information:

- Elements can be annotated to individual terms



GO over-representation analyses: what

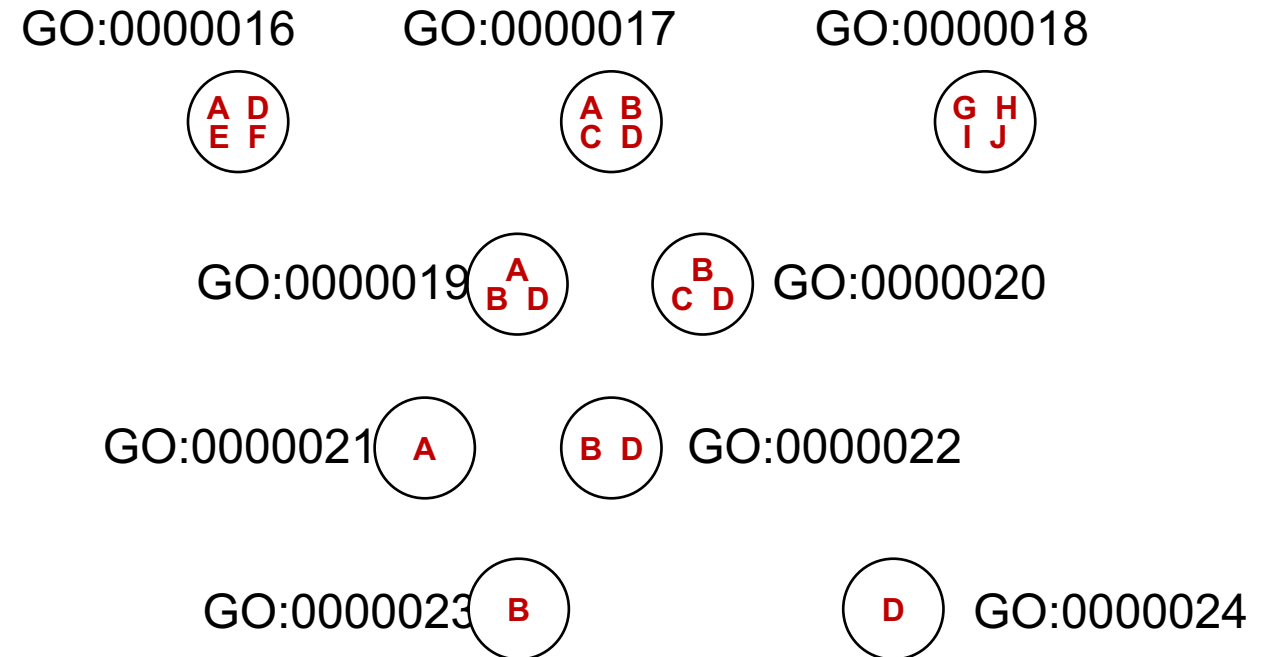
Ontology terms



GO over-representation analyses: what

Gene Ontology terms

+ GO annotations

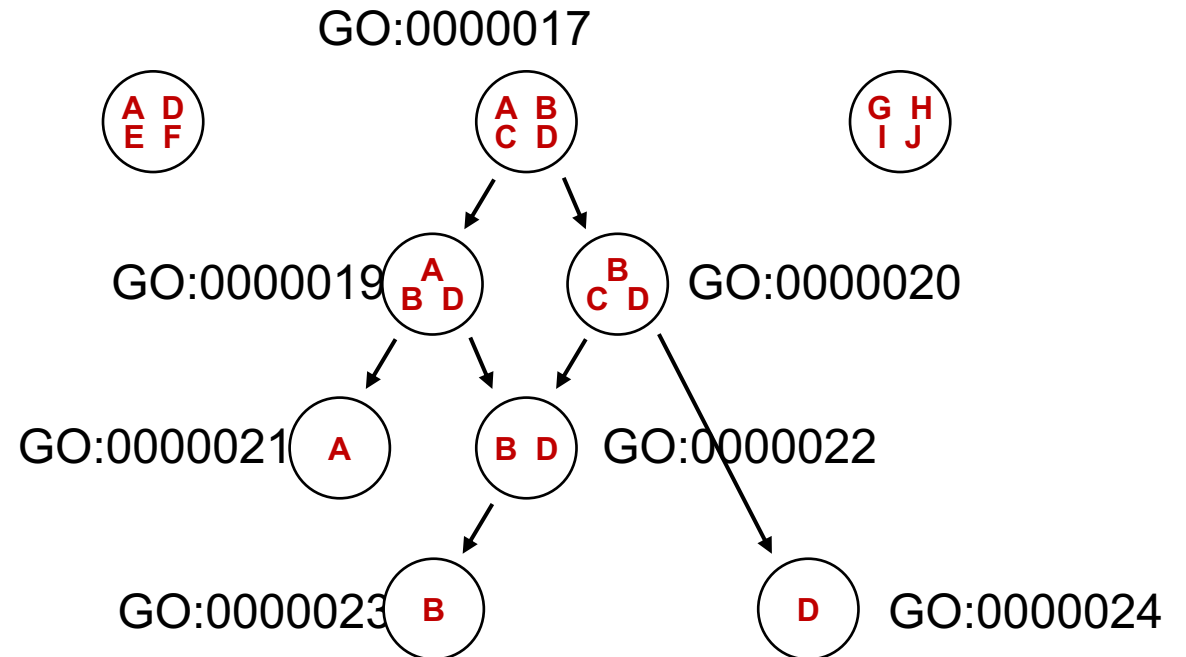


GO over-representation analyses: what

Gene Ontology terms

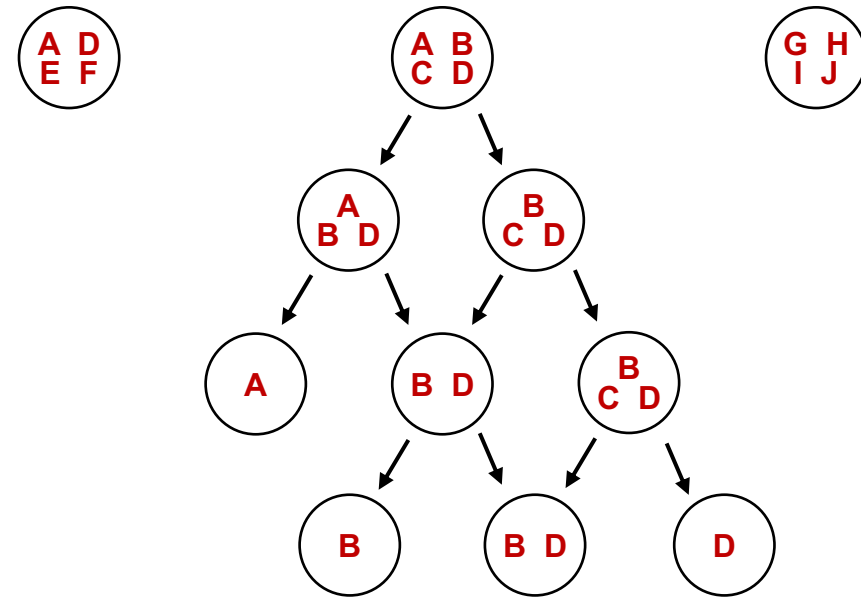
+ GO annotations

+ hierarchy



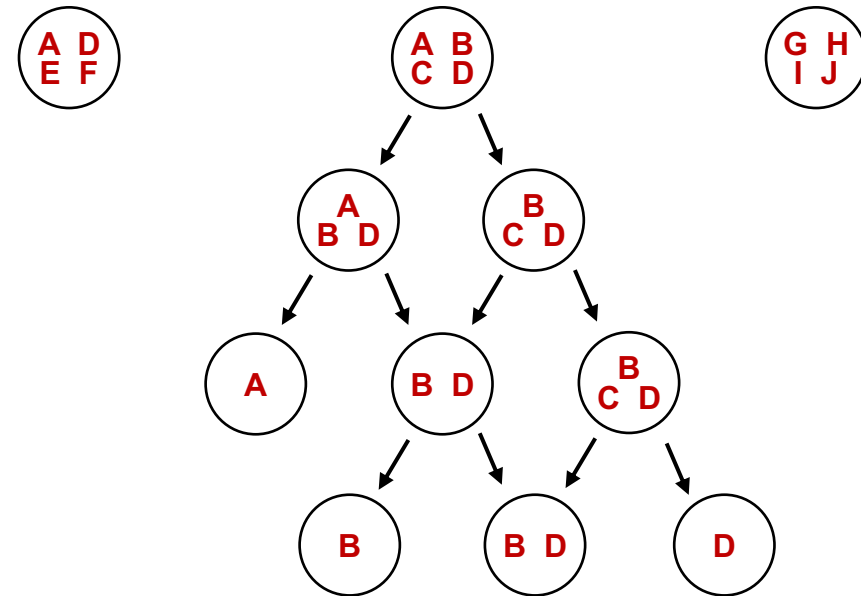
GO over-representation analyses: what

Different **Ontology terms** can contain the same sets of annotations



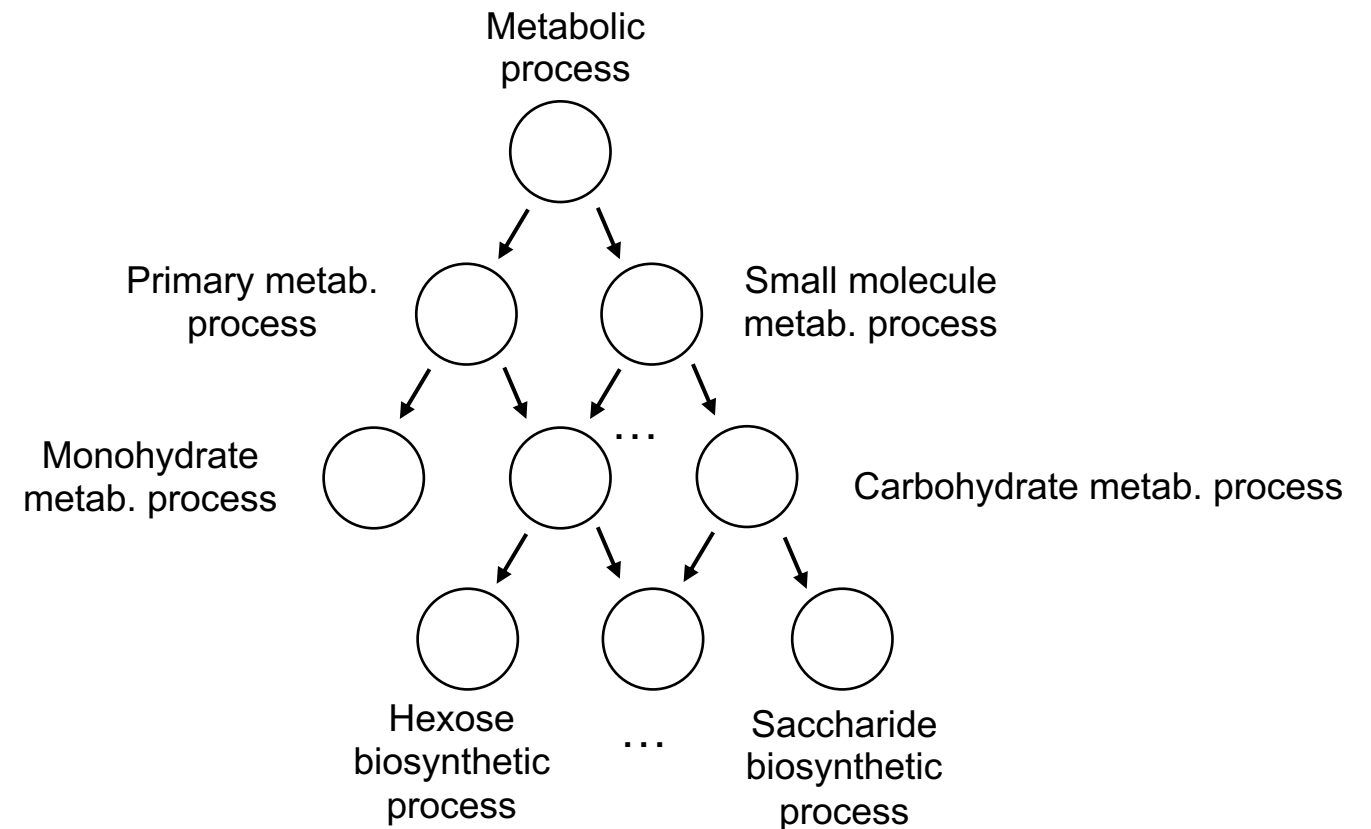
GO over-representation analyses: what

In our case, the Gene Ontology (GO) describes the current state of knowledge of the three main biological domains



GO over-representation analyses: what

In our case, the Gene Ontology (GO) describes the current state of knowledge in biology

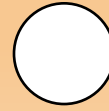


GO over-representation analyses: what

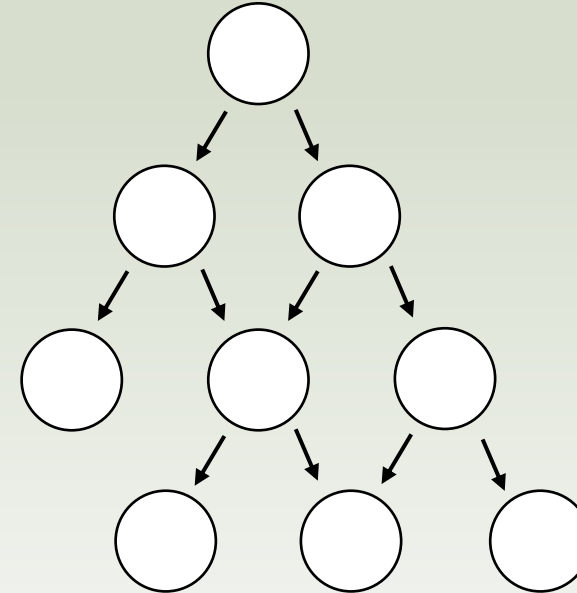
Gene Ontology (GO) is divided in three domains

- Biological Processes (BP)
- Cellular Components (CC)
- Molecular Functions (MF)

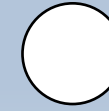
**Cellular
components**



**Biological
processes**

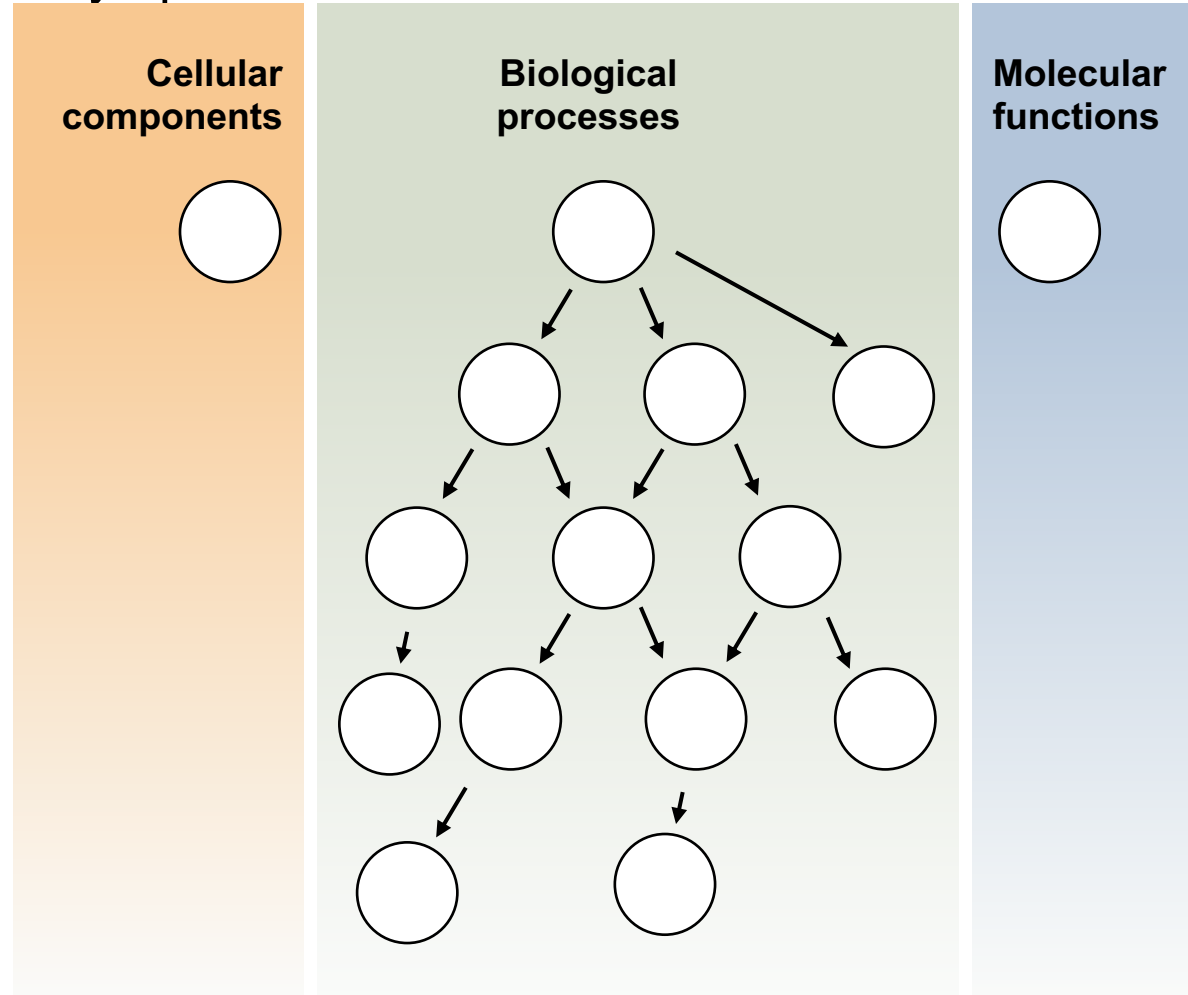


**Molecular
functions**



GO over-representation analyses: what

The Gene Ontology (GO) is a dynamic, frequently updated database

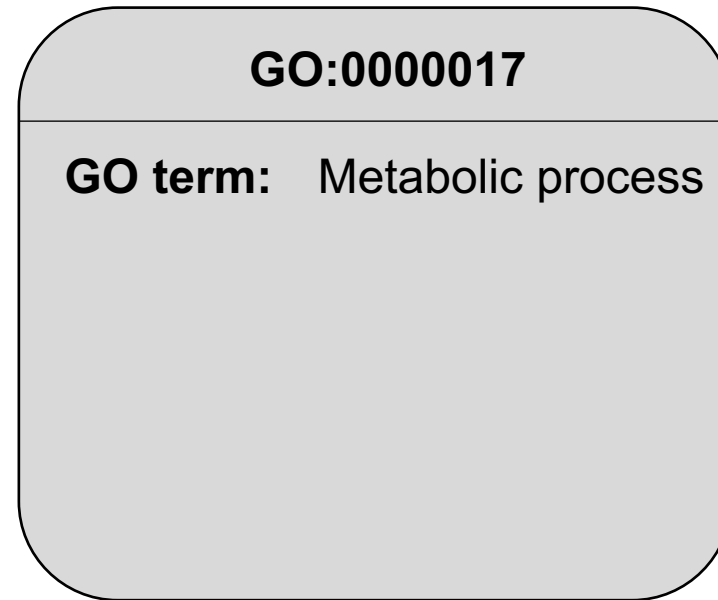
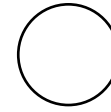


GO over-representation analyses: what

IMPORTANT:

A GO term (e.g. GO:0000017) is different from its annotations (i.e. the association of some genes to this term)

GO:0000017



GO over-representation analyses: what

IMPORTANT:

A GO term (e.g. GO:0000017) is different from its annotations (i.e. the association of some genes to this term)

GO:0000017



GO:0000017

GO term: Metabolic process

GO annotations: Gene A
Gene B
Gene C
Gene D

GO over-representation analyses: what

IMPORTANT:

GO consortium organizes GO terms and their hierarchy

External providers manage GO term annotations

GO over-representation analyses: what

IMPORTANT:

GO consortium organizes GO terms and their hierarchy

External providers manage GO term annotations

- Mouse annotations are provided by MGI (Mouse Genome Informatics)
- C. elegans annotations are provided by Wormbase
- Yeast annotations are provided by SGD (Saccharomyces Genome Database)

GO over-representation analyses: what

IMPORTANT:

GO consortium organizes GO terms and their hierarchy

External providers manage GO term annotations

- Mouse Genome Informatics (Mouse Genome Informatics)

Annotations are provided by Wormbase

Yeast annotations are provided by SGD (Saccharomyces Genome Database)

KNOW YOUR ORGANISM!

Official GO database

GO Consortium is the provider of official Gene Ontology.



About the GO

Mission Statement: 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.

The Gene Ontology resource provides a computational representation of our current scientific knowledge about the functions of genes (or, more properly, the protein and non-coding RNA molecules produced by genes) from many different organisms, from humans to bacteria. It is widely used to support scientific research, and has been cited in tens of thousands of publications.

Official GO database

GO Consortium is the provider of official Gene Ontology.

Additional refined gene ontologies exist, either from GO Consortium or from independent providers, e.g. :

- “Slim”-ed versions
- Non-model organisms

Downloading official GO database

- Versioned database
- Easy access to entire database
- OBO format

July 2, 2018 (2020-12-08) Dataset Open Access

Gene Ontology Data Archive

 Carbon, Seth;  Mungall, Chris;

Archival bundle of GO data release.

Uploaded on December 11, 2020

July 2, 2018 (2020-11-17) Dataset Open Access

Gene Ontology Data Archive

 Carbon, Seth;  Mungall, Chris;

Archival bundle of GO data release.

Uploaded on November 20, 2020

July 2, 2018 (2020-10-09) Dataset Open Access

Gene Ontology Data Archive

 Carbon, Seth;  Mungall, Chris;

Archival bundle of GO data release.

Uploaded on October 12, 2020

July 2, 2018 (2020-09-10) Dataset Open Access

Gene Ontology Data Archive

 Carbon, Seth;  Mungall, Chris;

Archival bundle of GO data release.

Uploaded on September 16, 2020

Downloading official GO database

OBO format

wget <http://purl.obolibrary.org/obo/go.obo>

```
1  > head -n 100 go.obo
2  format-version: 1.2
3  data-version: releases/2020-12-08
4  ontology: go
5
6  [Term]
7  id: GO:0000001
8  name: mitochondrion inheritance
9  namespace: biological_process
10 def: "The distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton." [GOC:mcc, PMID:10873824, PMID:11389764]
11 synonym: "mitochondrial inheritance" EXACT []
12 is_a: GO:0048308 ! organelle inheritance
13 is_a: GO:0048311 ! mitochondrion distribution
14
15 [Term]
16 id: GO:0000002
17 name: mitochondrial genome maintenance
18 namespace: biological_process
19 def: "The maintenance of the structure and integrity of the mitochondrial genome; includes replication and segregation of the mitochondrial chromosome." [GOC:ai, GOC:vw]
20 is_a: GO:0007005 ! mitochondrion organization
```

Downloading GO annotations

GO annotations:

- Also versioned
- All most recent annotations for individual species available at:

<http://current.geneontology.org/annotations/>

Downloading GO annotations

Parent

..

aspgd.gaf.gz
aspgd.gpad.gz
aspgd.gpi.gz
cgd.gaf.gz
cgd.gpad.gz
cgd.gpi.gz
dictybase.gaf.gz
dictybase.gpad.gz
dictybase.gpi.gz
ecocyc.gaf.gz
ecocyc.gpad.gz
ecocyc.gpi.gz
fb.gaf.gz
fb.gpad.gz
fb.gpi.gz
genedb_lmajor.gaf.gz
genedb_lmajor.gpad.gz
genedb_lmajor.gpi.gz
genedb_tbrucei.gaf.gz
genedb_tbrucei.gpad.gz
genedb_tbrucei.gpi.gz
goa_chicken.gaf.gz
goa_chicken.gpad.gz
goa_chicken.gpi.gz
goa_chicken_complex.gaf.gz

goa_dog_rna.gpad.gz
goa_dog_rna.gpi.gz
goa_human.gaf.gz
goa_human.gpad.gz
goa_human.gpi.gz
goa_human_complex.gaf.gz
goa_human_complex.gpad.gz
goa_human_complex.gpi.gz
goa_human_isoform.gaf.gz
goa_human_isoform.gpad.gz
goa_human_isoform.gpi.gz
goa_human_rna.gaf.gz
goa_human_rna.gpad.gz
goa_human_rna.gpi.gz
goa_pig.gaf.gz
goa_pig.gpad.gz
goa_pig.gpi.gz
goa_pig_complex.gaf.gz
goa_pig_complex.gpad.gz
goa_pig_complex.gpi.gz
goa_pig_isoform.gaf.gz
goa_pig_isoform.gpad.gz
goa_pig_isoform.gpi.gz
goa_pig_rna.gaf.gz
goa_pig_rna.gpad.gz
goa_pig_rna.gpi.gz
goa_uniprot_all.gaf.gz
goa_uniprot_all_noiea.gaf.gz
goa_uniprot_all_noiea.gpad.gz
goa_uniprot_all_noiea.gpi.gz

goa_chicken_complex.gpad.gz
goa_chicken_complex.gpi.gz
goa_chicken_isoform.gaf.gz
goa_chicken_isoform.gpad.gz
goa_chicken_isoform.gpi.gz
goa_chicken_rna.gaf.gz
goa_chicken_rna.gpad.gz
goa_chicken_rna.gpi.gz
goa_cow.gaf.gz
goa_cow.gpad.gz
goa_cow.gpi.gz
goa_cow_complex.gaf.gz
goa_cow_complex.gpad.gz
goa_cow_complex.gpi.gz
goa_cow_isoform.gaf.gz
goa_cow_isoform.gpad.gz
goa_cow_isoform.gpi.gz
goa_cow_rna.gaf.gz
goa_cow_rna.gpad.gz
goa_cow_rna.gpi.gz
goa_dog.gaf.gz
goa_dog.gpad.gz
goa_dog.gpi.gz
goa_dog_complex.gaf.gz
goa_dog_complex.gpad.gz
goa_dog_complex.gpi.gz
goa_dog_isoform.gaf.gz
goa_dog_isoform.gpad.gz
goa_dog_isoform.gpi.gz
goa_dog_rna.gaf.gz

mgf.gaf.gz
mgf.gpad.gz
mgf.gpi.gz
pombase.gaf.gz
pombase.gpad.gz
pombase.gpi.gz
pseudocap.gaf.gz
pseudocap.gpad.gz
pseudocap.gpi.gz
reactome.gaf.gz
reactome.gpad.gz
reactome.gpi.gz
rgd.gaf.gz
rgd.gpad.gz
rgd.gpi.gz
sgd.gaf.gz
sgd.gpad.gz
sgd.gpi.gz
sgn.gaf.gz
sgn.gpad.gz
sgn.gpi.gz
tair.gaf.gz
tair.gpad.gz
tair.gpi.gz
wb.gaf.gz
wb.gpad.gz
wb.gpi.gz
zfin.gaf.gz
zfin.gpad.gz
zfin.gpi.gz

Downloading GO annotations

GAF format:

It's in the name: **GO Annotation Format**

Downloading GO annotations

GAF format:

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgd-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgd_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
...
SGD S000004103 HOG1 GO:0003682 PMID:24508389 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0004707 PMID:10805732 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0005516 PMID:27421986 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:10805732 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006972 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0007231 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0016241 PMID:16874103 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0033262 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0045944 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
...
```

Downloading GO annotations

GAF format

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgd-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgd_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
!
***
SGD S000004103 HOG1 GO:0003682 PMID:24508389 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0004707 PMID:10805732 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0005516 PMID:27421986 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:10805732 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006972 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0007231 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0016241 PMID:16874103 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0033262 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0045944 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
***
```

Annotation provider

Peeking into functional roles of gene sets



Downloading GO annotations

GAF format

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgf source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgf-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgf_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
...
SGD S00000410 HOG1 GO:0003682 PMID:24508389 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0004707 PMID:10805732 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0005516 PMID:27421986 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0006468 PMID:10805732 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0006468 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0006468 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0006972 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0007231 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0016241 PMID:16874103 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0033262 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S00000410 HOG1 GO:0045944 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
...
```

DB object symbol

Peeking into functional roles of gene sets



Downloading GO annotations

GAF format

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgd-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgd_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
...
SGD S000004103 HOG1 GO:0005882 PMID:24508389 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0004707 PMID:10805732 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0005516 PMID:27421986 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:10805732 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006972 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0007231 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0016241 PMID:16874103 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0033262 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0045944 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
...
```

GO term

Peeking into functional roles of gene sets



Downloading GO annotations

GAF format

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgd-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgd_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
...
SGD S000004103 HOG1 GO:0003682 PMID:124308369 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0004707 PMID:10805732 IDA F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0005516 PMID:27421986 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:10805732 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006972 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0007231 PMID:7681220 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0016241 PMID:16874103 IMP P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0033262 PMID:23178807 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0045944 PMID:12743037 IDA P Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
...
```

Downloading GO annotations

GAF format

```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=====
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgf source association file:
!=====
!Date: 20201207
!From: Saccharomyces Genome Database (SGD)
!URL: https://www.yeastgenome.org/
!Contact Email: sgf-helpdesk@lists.stanford.edu
!Funding: NHGRI at US NIH, grant number U41-HG001315
!
!=====
!
!Header copied from paint_sgf_valid.gaf
!=====
!Created on Mon Dec 7 11:33:04 2020.
!generated-by: PANTHER
!date-generated: 2020-12-07
!PANTHER version: v.15.0.
!GO version: 2020-11-17.
!
!=====
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
...
SGD S000004103 HOG1 GO:0003682 PMID:2450836 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0004707 PMID:1080573 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0005516 PMID:2742198 IPI UniProtKB:P06787 F Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:1080573 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:1274303 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006468 PMID:2317880 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0006972 PMID:7681220 IMP Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0007231 PMID:7681220 IMP Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0016241 PMID:1687410 IMP Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0033262 PMID:2317880 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
SGD S000004103 HOG1 GO:0045944 PMID:1274303 IDA Mitogen-activated protein kinase involved in osmoregulation YLR113W|SSK3|mitogen-activated protein kinase HOG1
...
```

Type of evidence

Peeking into functional roles of gene sets



GO over-representation analyses: when

When you have a **defined** set of tens to hundreds of genes

- Genes significantly over-expressed (e.g. fold-change > 2 and adj. p-value < 0.01) in one condition vs a control
- Genes whose promoter is bound by a specific combination of transcription factors
- Thousands are probably too many genes...

GO over-representation analyses: why

- Essentially, to know whether GO terms are over-enriched in a specific list of genes
- To get an idea of the functional/structural role of your set of genes
- To bring a piece of evidence that your treatment triggers some BP/MF/CC
- To know how much of the genes involved in a specific BP/MF/CC are present in your set of interest.

GO over-representation analyses: how

Finding over-represented GO terms in a given set of genes is one of the most common tasks in genomics.

GO over-representation analyses: how

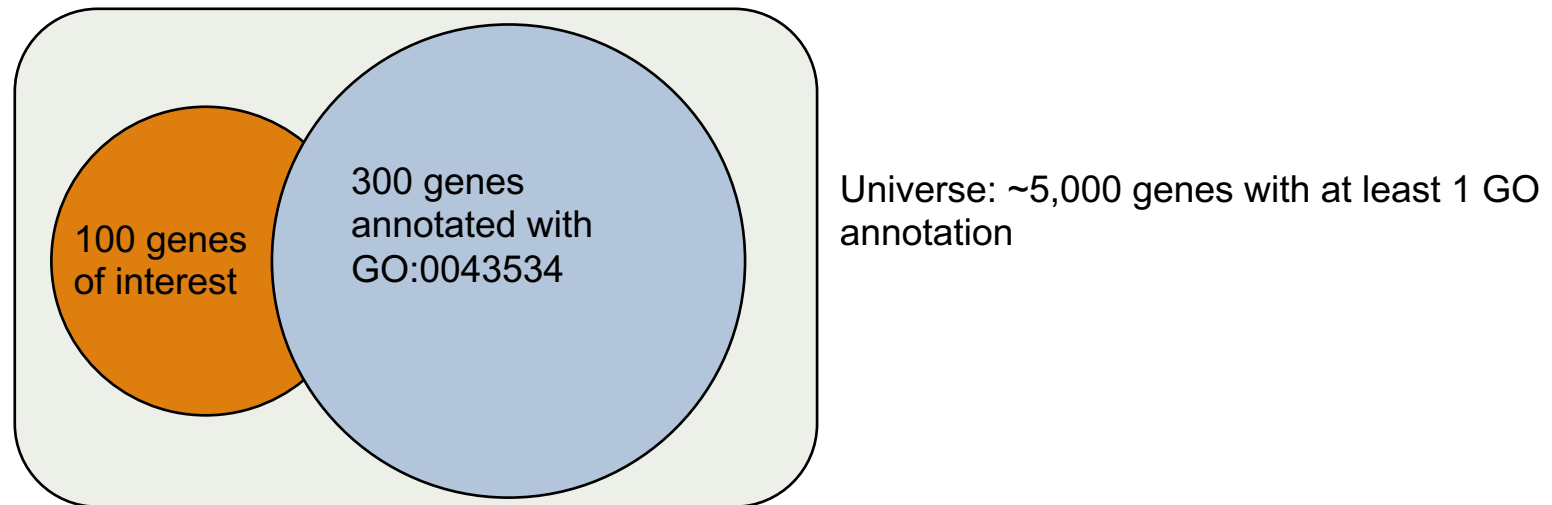
Finding over-represented GO terms in a given set of genes is one of the most common tasks in genomics.

It usually relies on a straightforward Fisher test

GO over-representation analyses: how

Finding over-represented GO terms in a given set of genes is one of the most common tasks in genomics.

It usually relies on a straightforward Fisher test



GO over-representation analyses: how

Finding over-represented GO terms in a given set of genes is one of the most common tasks in genomics.

It usually relies on a straightforward Fisher test

Think about it in terms of contingency tables

GO over-representation analyses: how

UNIVERSE = All Yeast genes annotated in the Biological Processes (5067 genes)	Genes over-expressed <i>in an assay</i> (152)	<u>Genes over-expressed <i>in an assay</i></u>	
Genes annotated <i>in GO:0006836</i> (243)	89	154	→ Sum = number of genes in GO:0006836 (243)
<u>Genes annotated <i>in GO:0006836</i></u>	63	$5067 - 89 - 154 - 63$	↓
↓ Sum = number of genes over- expressed in an assay (152)		→	Total sum = number of genes in BP (5067)

GO over-representation analyses: how

UNIVERSE = All Yeast genes annotated in the Biological Processes (5067 genes)	Genes over-expressed <i>in an assay</i> (152)	<u>Genes over-expressed <i>in an assay</i></u>
Genes annotated <i>in GO:0006836</i> (243)	89	154
<u>Genes annotated <i>in GO:0006836</i></u>	63	$5067 - 89 - 154 - 63$

➔ Now repeat that for the 44,945 GO terms in the GO database.....

GO over-representation analyses: how

UNIVERSE = All Yeast genes annotated in the Biological Processes (5067 genes)	Genes over-expressed <i>in an assay</i> (152)	Genes over-expressed <i>in an assay</i>
Genes annotated <i>in GO:0006836</i> (243)	89	154
Genes annotated <i>in GO:0006836</i>	63	$5067 - 89 - 154 - 63$

➔ AND DON'T FORGET TO CORRECT FOR MULTIPLE TESTING (because testing 44,945 times is multiple testing...)

GO over-representation analyses: how

- Fortunately, there are many tools already out there to efficiently perform these calculations
- Some web-based, some with programmatic access
- They function with a range of “autonomy”. Some need you to download the GO database, the GO annotations, or are doing all the work for you in the background

Programs to run GO over-representation analyses: gProfiler

g:Profiler

[News](#)[Archives](#)[Beta](#)[API](#)[R client](#)[FAQ](#)[Docs](#)[Contact](#)[Cite g:Profiler](#)[Services using g:P](#)[List of organisms](#)

g:Profiler has been updated with new data from Ensembl.

Show more...Close

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

Run query

random example

mixed query example

Options

Organism: ?

Homo sapiens (Human)

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼



Physalis
Courses

Peeking into functional roles of gene sets

44

Programs to run GO over-representation analyses: gProfiler

- Also available in R!
- Simple, but many optional parameters to optimize your search

```
gprofiler2::gost(  
  geneList,  
  organism = 'scerevisiae'  
)
```

Staying up-to-date...



✧ Ontologies

- [PANTHER™ GO slim](#) (version 16.0, based on GO release 2020-11-16, released 2020-12-01)
 - 3336 total terms
 - 2235 biological process terms
 - 543 cellular component terms
 - 558 molecular function terms
- PANTHER™ Protein Class (version 16.0, released 2020-12-01)
 - 210 total terms
- [Gene Ontology](#) (from GO database released 2020-10-09, DOI: 10.5281/zenodo.4081749)
 - 47308 total terms
 - 12103 molecular function terms
 - 30816 biological process terms
 - 4389 cellular component terms

Staying up-to-date...



News

Archives

Beta

API

R client

FAQ

Docs

Contact

Cite g:Profiler

Services using g:P

List of organisms

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

g:Profiler has been updated with new data from Ensembl.

The database versions in this release are:

- Ensembl 101
- Ensembl Genomes 48
- Wormbase ParaSite 14

Changes in g:GOST

There are now evidence codes available for Human Protein Atlas (HPA) results.

There is now an option to exclude evidence codes from query results. When the "No evidence codes" option is chosen, the gene-term matrix will not be shown and for large queries, the results will be somewhat faster.

Published on Tue Oct 20 2020

g:Profiler has been updated with new data from Ensembl.

The database versions in this release are:

- Ensembl 100
- Ensembl Genomes 47
- Wormbase ParaSite 14

Published on Wed Jul 22 2020

g:Profiler has been updated with new data from Ensembl.

The database versions in this release are:

- Ensembl 99
- Ensembl Genomes 46
- Wormbase ParaSite 14

There have been many minor UI improvements and small feature enhancements. One of the larger updates is the overhaul of our g:GOST multiquery interface. Comparing the enrichment of two or more gene lists should be easier and less cluttered.

There is a new API endpoint for accessing the metadata about available species. It provides the names, ID-s and available namespaces for all our organisms. The endpoint is documented on our [API page \(https://biit.cs.ut.ee/gprofiler/page/apis\)](https://biit.cs.ut.ee/gprofiler/page/apis).

It's now possible to export high-DPI (300 or 600) images of your results for use in presentations or publications.

We have updated the documentation of our [r client \(https://biit.cs.ut.ee/gprofiler_beta/page/r\)](https://biit.cs.ut.ee/gprofiler_beta/page/r) and created an instructional vignette (<https://cran.r-project.org/web/packages/gprofiler2/vignettes/gprofiler2.html>).

Published on Mon Mar 09 2020

Staying up-to-date...

GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. [ChIP-seq](#)) and by computational methods (e.g. [comparative genomics](#)). For more see our [Nature Biotech Paper](#).

News

-  Aug. 19, 2019: GREAT version 4 [adds support for human hg38 assembly and updates ontology datasets for all supported assemblies](#).
-  Sep. 8, 2018: GREAT has served over 1 million job submissions.
- Oct. 23, 2017: GREAT is moved to a VM to eliminate proxy errors.
- June 22, 2017: GREAT hardware upgrade to meet increasing submission volume.
- Nov. 16, 2015: The [GREAT user help forums](#) are frozen.
- Feb. 15, 2015: GREAT version 3 [switches to Ensembl genes, adds support for zebrafish danRer7 and mouse mm10 assemblies, and adds new ontologies](#).
- Apr. 3, 2012: GREAT version 2 [adds new annotations to human and mouse ontologies and visualization tools for data exploration](#).
- Feb. 18, 2012: The [GREAT user help forums](#) are opened.
- May 2, 2010: GREAT version 1 is launched, concurrent to [Nature Biotechnology publication](#) (reprint, Faculty of 1000 "Must Read"). [How to Cite GREAT?](#)

[More news items...](#)

Species Assembly

- ☐ Human: GRCh38 ([UCSC hg38, Dec. 2013](#))
- ☐ Human: GRCh37 ([UCSC hg19, Feb. 2009](#))
- ☐ Mouse: GRCm38 ([UCSC mm10, Dec. 2011](#))
- ☐ Mouse: NCBI build 37 ([UCSC mm9, Jul. 2007](#))

[Can I use a different species or assembly?](#)

Staying up-to-date...

The DAVID Knowledgebase (DAVID 6.8, Current version available at <https://david.ncifcrf.gov>)

Main Annotation Sources

Data Sources	Release / Download Date	DAVID Update Date
ENSEMBL	Mar 2016	May 2016
ENTREZ	May 2016	May 2016
UNIPROT	May 2016	May 2016

Secondary Sources

Data Sources	Release / Download Date	DAVID Update Date
AFFYMETRIX	Jun 2015	May 2016
AGILENT	Dec 2013	May 2016
BBID	Sep 2009	May 2016
BIOCARTA	Nov 2014	May 2016
CGAP_EST_QUARTILE	Oct 2006	May 2016
CGAP_SAGE_QUARTILE	Oct 2006	May 2016
COG_ONTOLOGY	Sep 2009	May 2016
GENE ONTOLOGY	Apr 2016	May 2016
GNF_U133A_QUARTILE	Oct 2006	May 2016
KEGG	Dec 2015	May 2016
UCSC_TFBS	Sep 2009	May 2016
UP_SEQ_FEATURE	Sep 2009	May 2016
UP_TISSUE	Sep 2009	May 2016
ZFIN_ANATOMY	Sep 2009	May 2016

Staying up-to-date...

GOrilla News

March. 8th 2013

- Added option to supply an e-mail address to which a link to the results will be sent.

December 3rd 2012

- Try our new tool [miTEA](#) for miRNA target enrichment analysis

October. 29th 2012

- Added option to supply a name for the analysis which will appear in the results page

May. 28th 2012

- A false discovery rate (FDR) column added to the results table
- Maximum input size increased to 1MB

Dec. 29th 2010

- GOrilla has been moved to a new and faster server

Aug. 30th 2010

- The GOrilla GO database is now automatically updated weekly
- The analysis results can now be exported to [REViGO](#) for further visualization

Feb. 7th 2010

- You can now run all 3 GO ontologies (Process, Function and Cellular component) in a single run.
(We thank Ben Gordon for the idea)
- GO and gene files were updated

Oct. 15th 2009

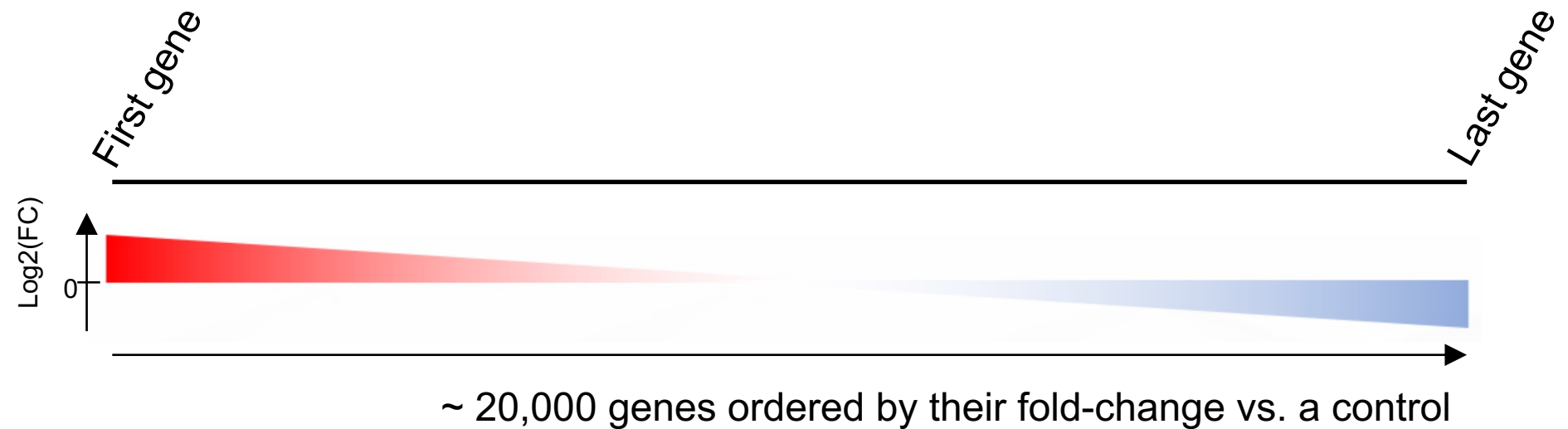
- GOrilla now supports Danio rerio (Zebrafish)
- GO and gene files were updated

What if I don't have a gene set of interest?

- Sometimes, you cannot really decide what is significant or not
- You don't like the idea of taking the top 100 genes differently expressed genes
- How to set a threshold for your genes? $FC > 2$? $FC > 5$? ???

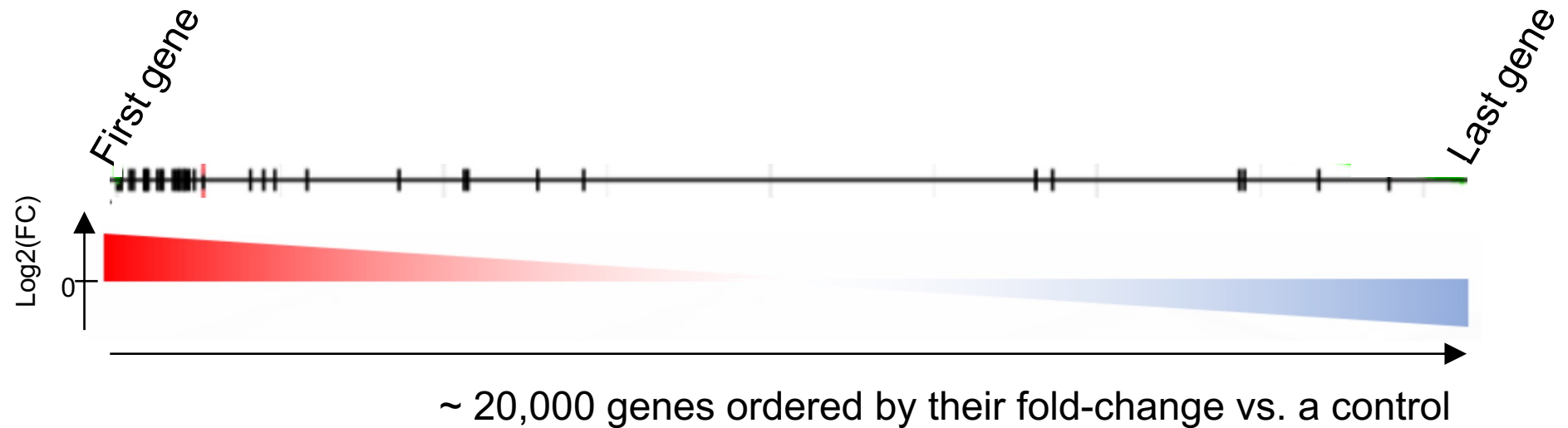
Gene Set Enrichment Analysis: cutoff-free functional analysis of ranked lists of genes

GSEA (Gene Set Enrichment Analysis) uses a ranked list of genes as input



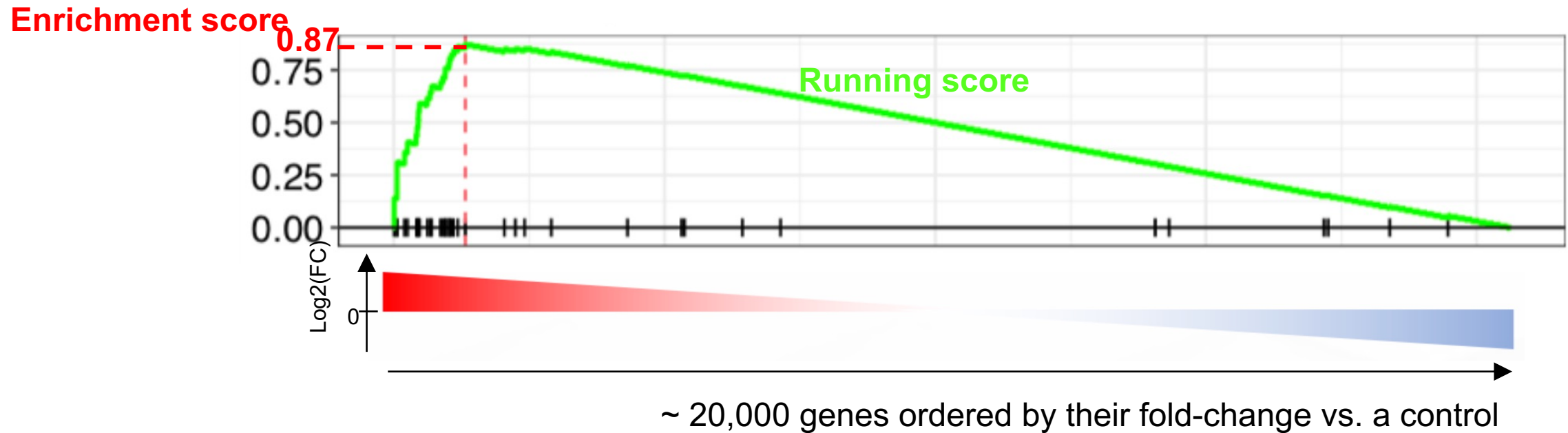
Gene Set Enrichment Analysis: cutoff-free functional analysis of ranked lists of genes

Within this list, it flags the genes belonging to a gene set (e.g. genes annotated in “centriole assembly” GO term)



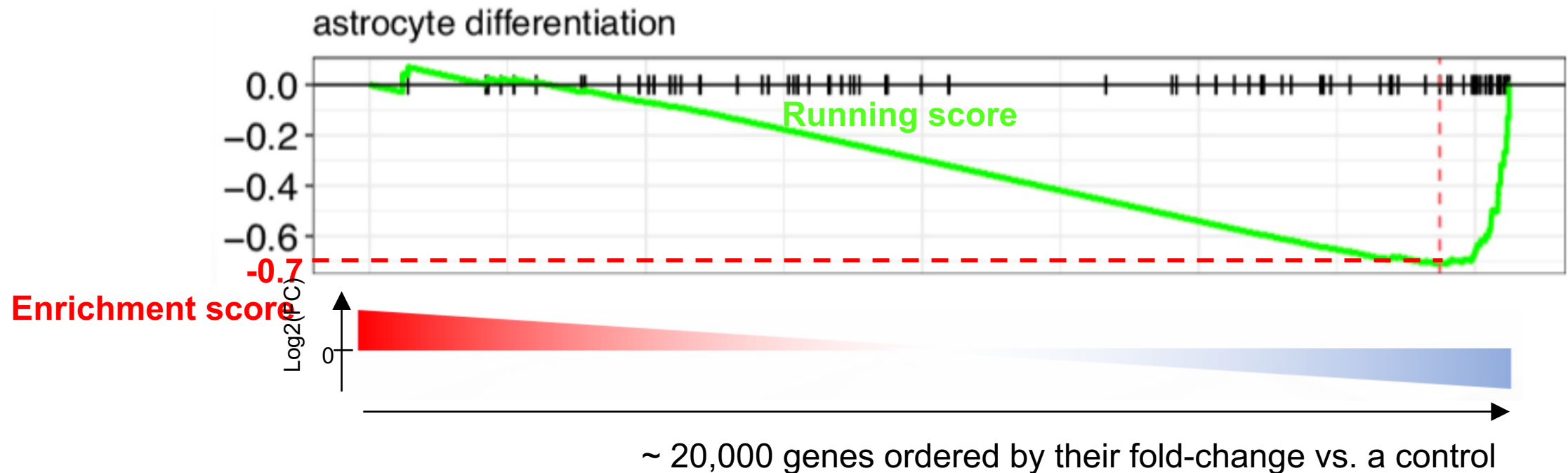
Gene Set Enrichment Analysis: cutoff-free functional analysis of ranked lists of genes

Based on the distribution of the flagged genes, it computes a “running score” and an “enrichment score”



Gene Set Enrichment Analysis: cutoff-free functional analysis of ranked lists of genes

It can also find negative enrichment scores (indicating a depletion of genes of interest in the top of a ranked list)



How to perform GSEA?

Original software: in JAVA

- I never managed to use it...

How to perform GSEA?

Original software: in JAVA

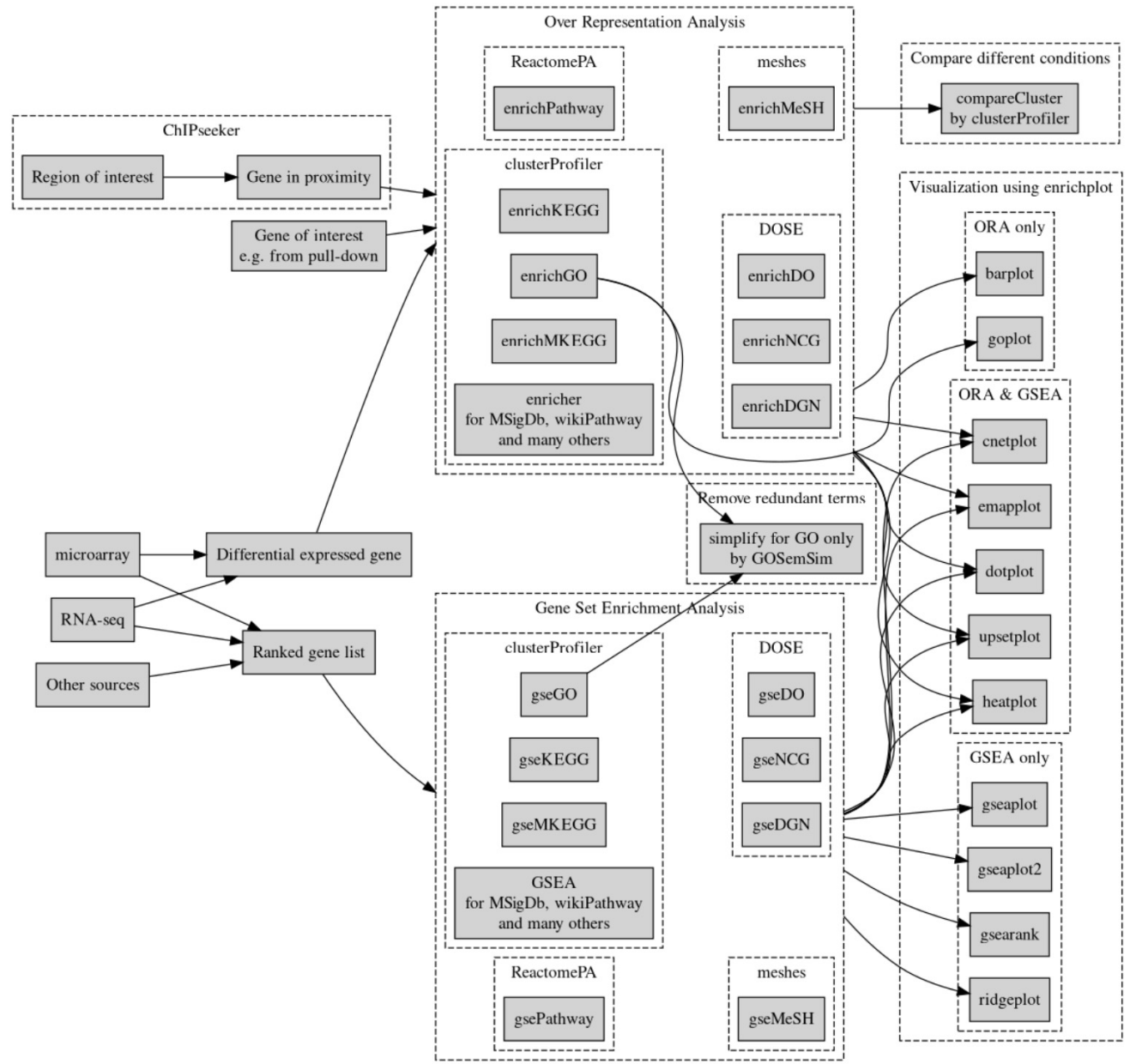
- I never managed to use it...

Since then: many programmatic implementations notably in R

- clusterProfiler is my personal favorite
- Based on fgsea, the original GSEA implementation in R
- Very complete and extensive doc
- Nice visualization outputs
- Well-integrated with GO ecosystem and other databases (disease ontology, Reactome, ...)

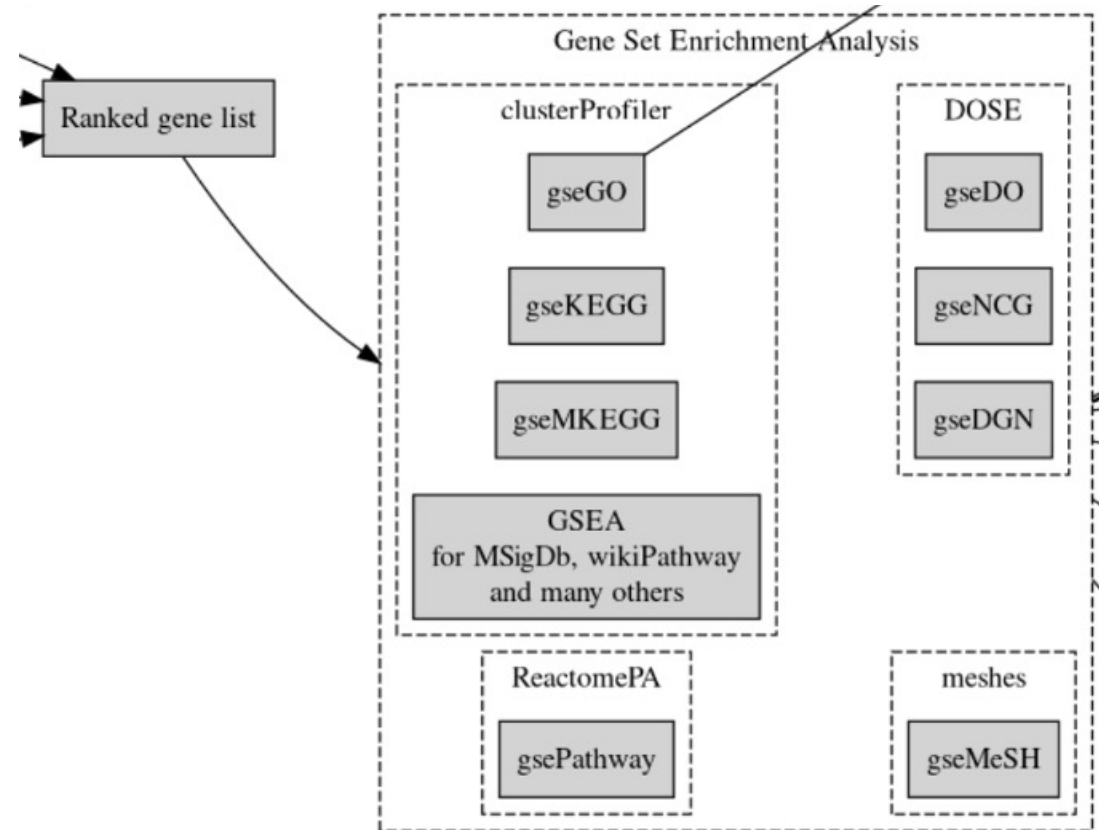
clusterProfiler

clusterProfiler is a rich set of tools to assess and visualize enrichment of a set of genes of interest compared to different databases



clusterProfiler

clusterProfiler provides multiple `gse*()` functions, based on the type of gene sets you want to use



clusterProfiler

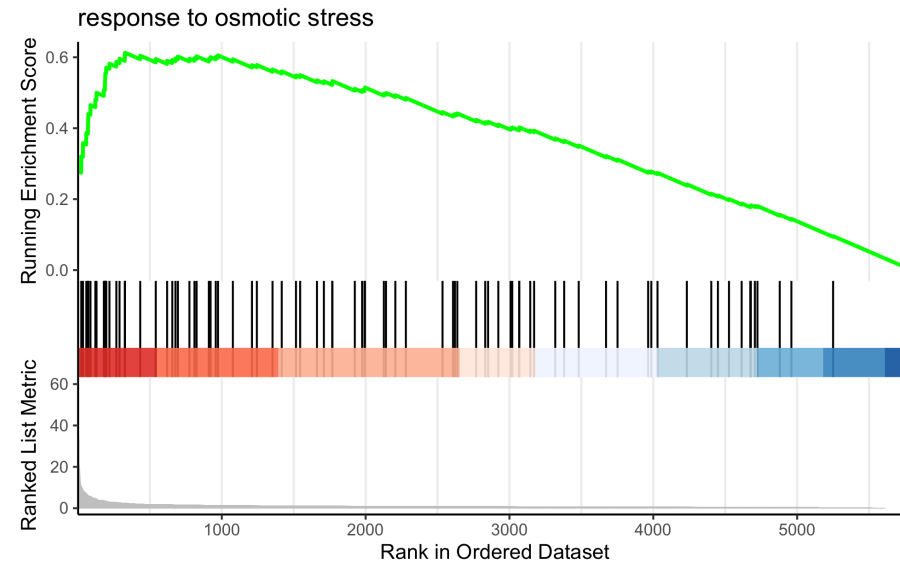
- clusterProfiler provides multiple gse*() functions, based on the type of gene sets you want to use
- gseGO() compares your ranked list of genes to the reference up-to-date GO annotations

```
clusterProfiler::gseGO(  
  rankedGeneLit,  
  keyType = "ENSEMBL",  
  OrgDb = org.Sc.sgd.db::org.Sc.sgd.db  
)
```

clusterProfiler

clusterProfiler provides plotting functions to check the results for a given GO term

```
TERM <- "response to osmotic stress"  
enrichplot::gseaplot2(  
  gsea_results,  
  title = TERM,  
  geneSetID = which(gsea_results$result$Description == TERM)  
)
```



Ten Quick Tips for Using the Gene Ontology, Blake PLoS Comp. Biol. 2013

Tip 1: Know the Source of the GO Annotations You Use

clusterProfiler: universal enrichment tool for functional and comparative study, Guangchuang Yu
(<http://yulab-smu.top/clusterProfiler-book/>)