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

Peeking into functional roles of gene sets

Epigenomics Data Analysis

Jacques Serizay

Physalia 2025

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 + 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

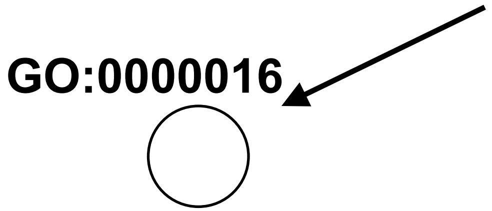
GO over-representation analyses: what

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 over-representation analyses: what

An ontology term can be further enriched with additional information:

- Elements can be annotated to individual terms

B
D

GO:0000022

GO over-representation analyses: what

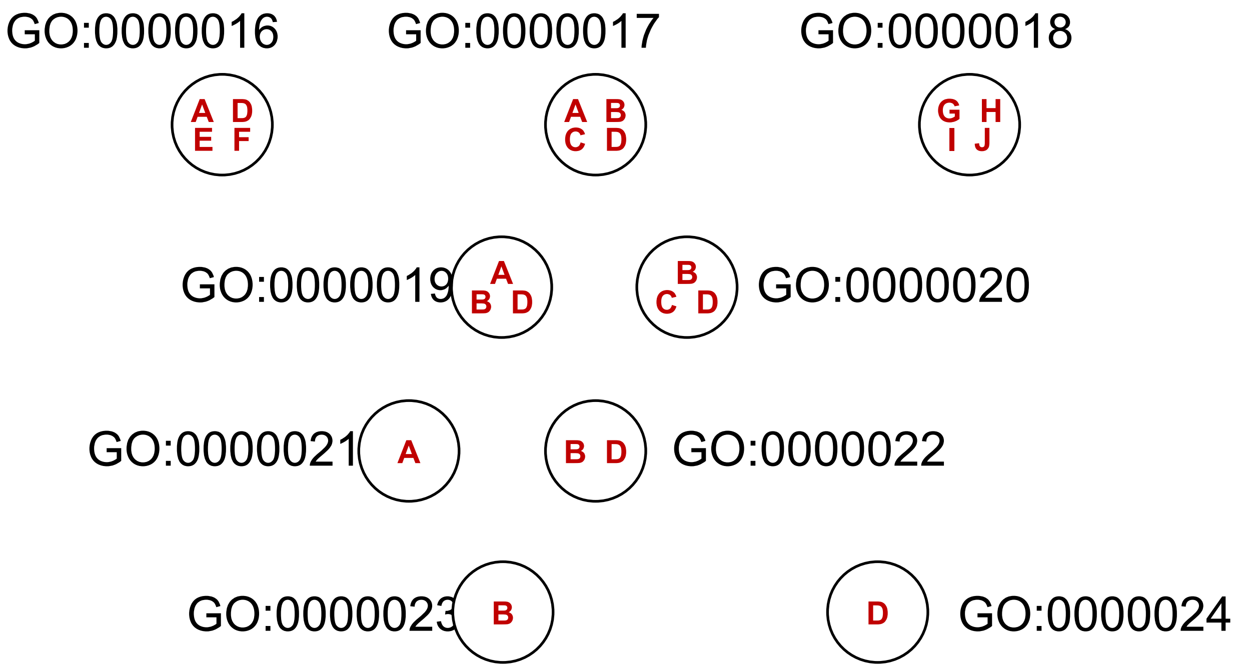
Ontology terms

GO:0000016
GO:0000017
GO:0000018
GO:0000019
GO:0000020
GO:0000021
GO:0000022
GO:0000023
GO:0000024

GO over-representation analyses: what

Gene Ontology terms

+ GO annotations

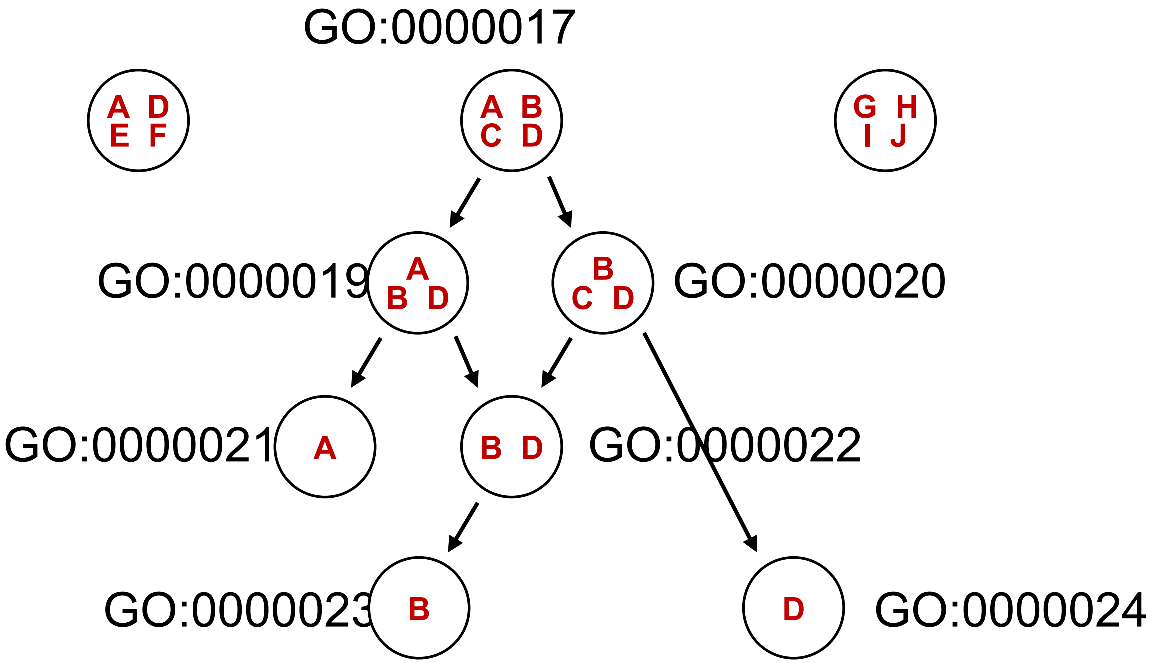


GO over-representation analyses: what

Gene Ontology terms

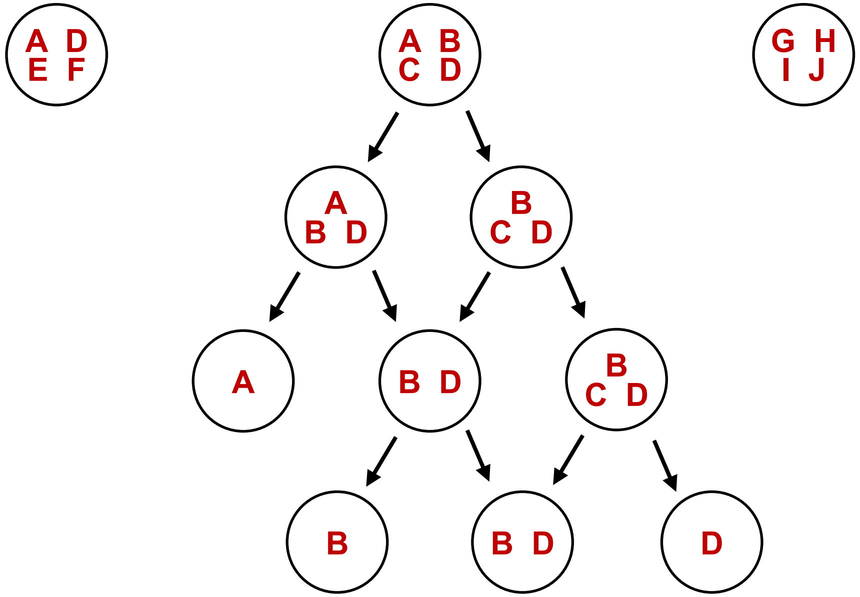
+ GO annotations

+ hierarchy



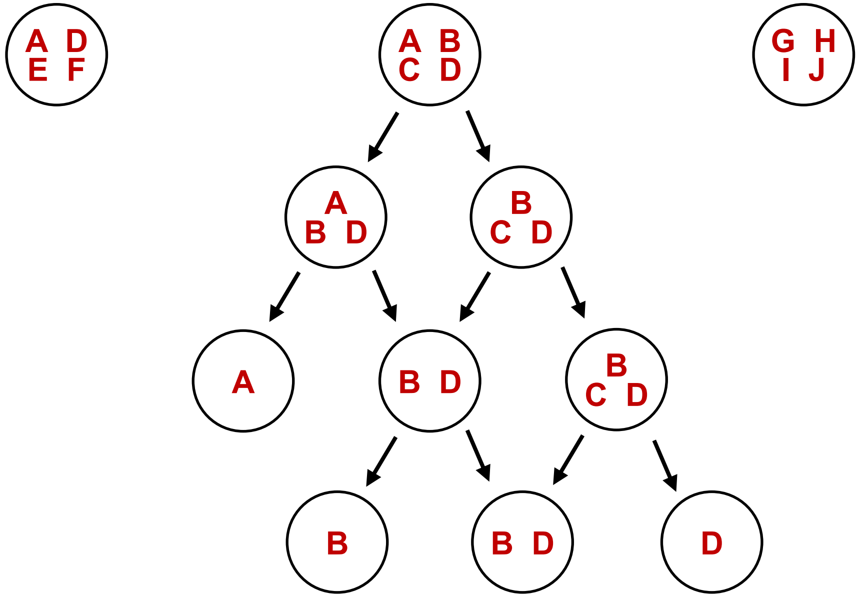
GO over-representation analyses: what

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



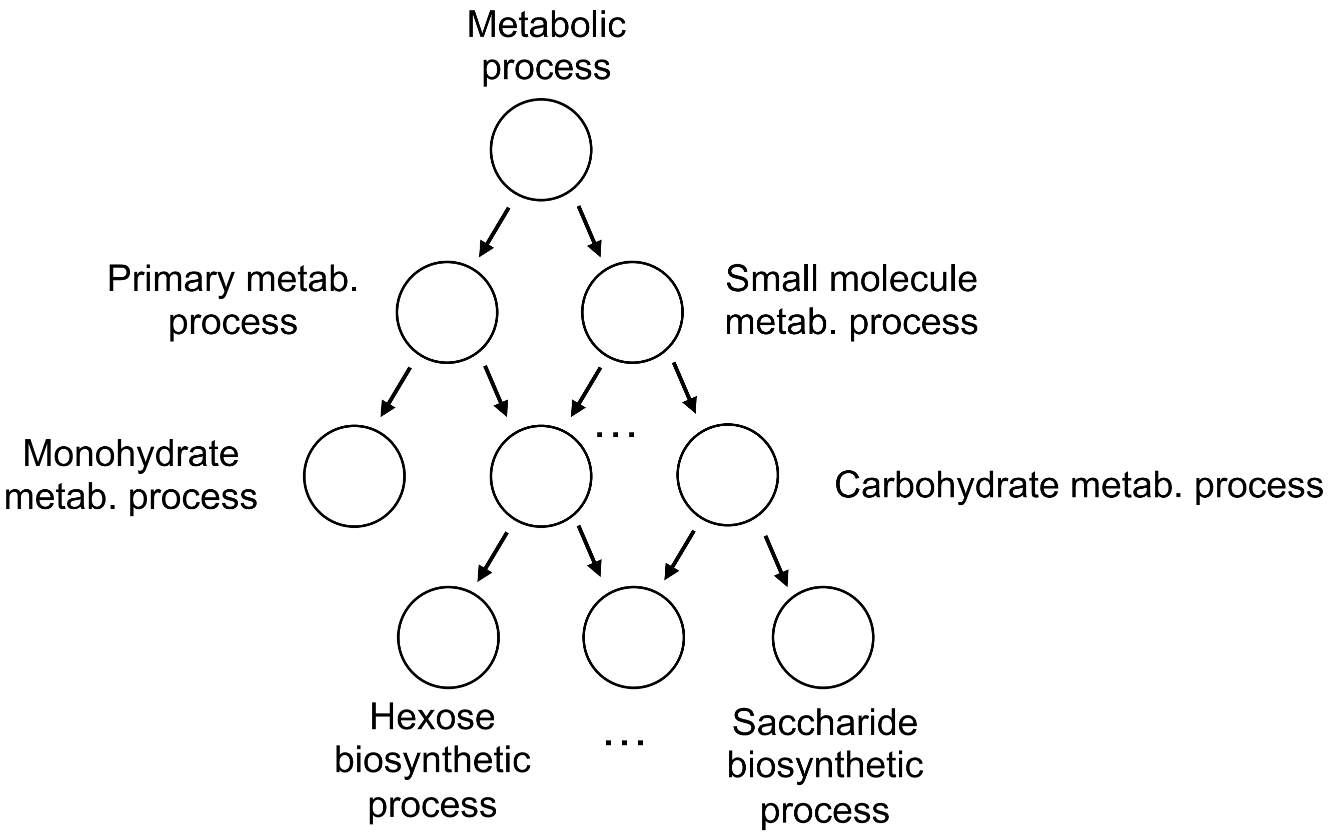
GO over-representation analyses: what

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



GO over-representation analyses: what

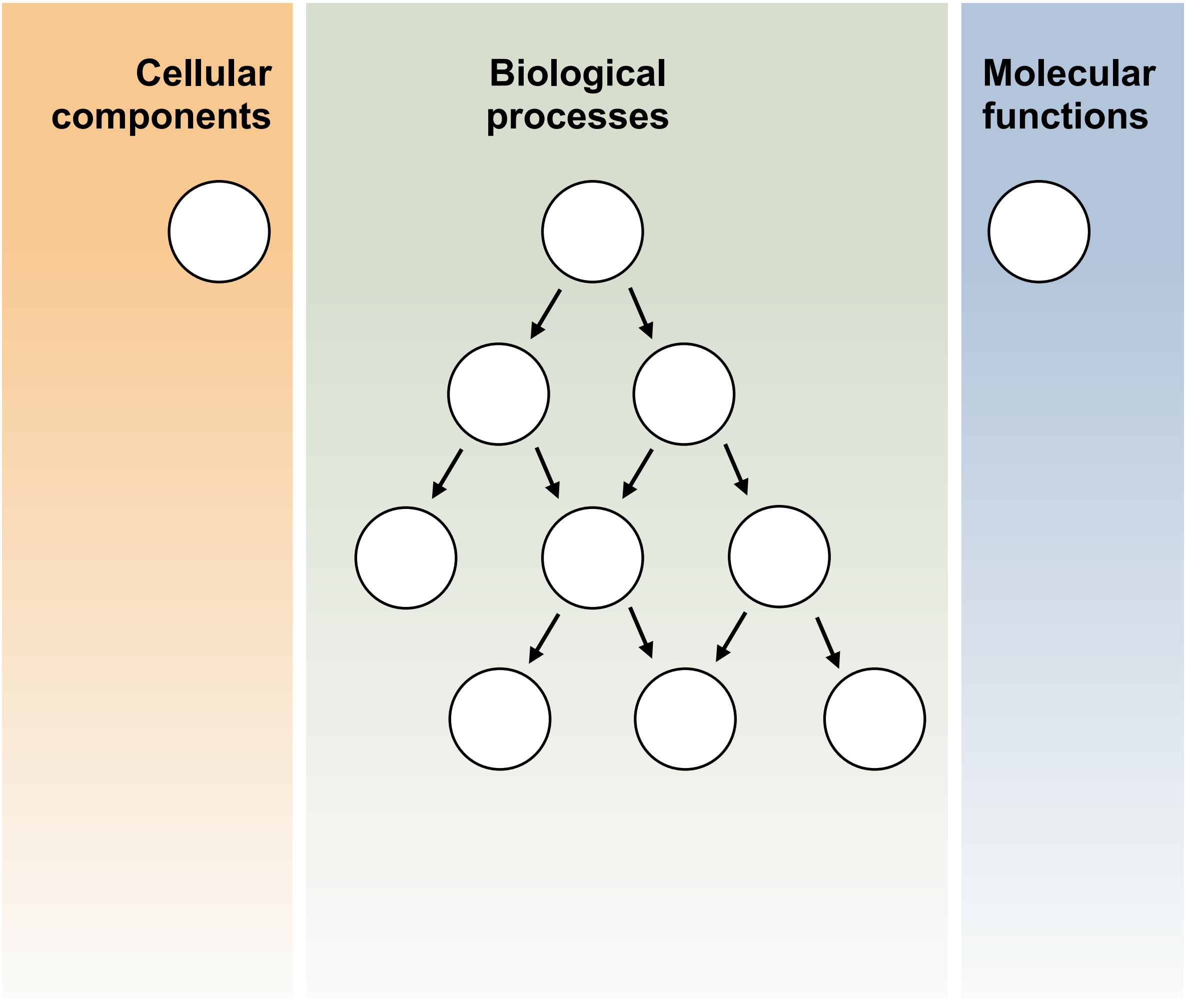
In our case, the Gene Ontology (GO) describes the **current state** of knowledge in biology


Metabolic process
Primary metab. process
Small molecule metab. process
Monohydrate metab. process
…
Carbohydrate metab. process
Hexose biosynthetic process
…
Saccharide biosynthetic process

GO over-representation analyses: what

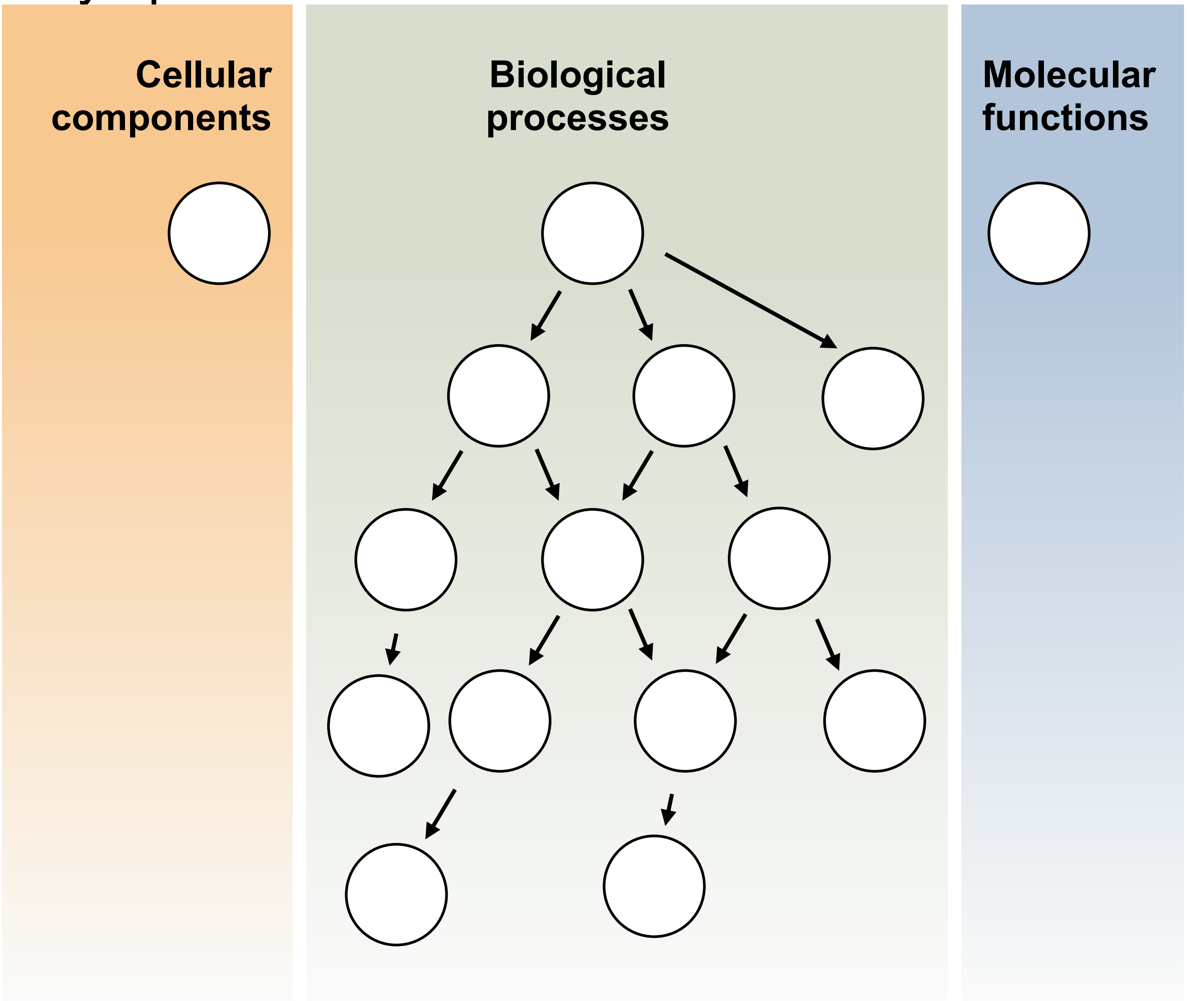
Gene Ontology (GO) is divided in three domains

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



GO over-representation analyses: what

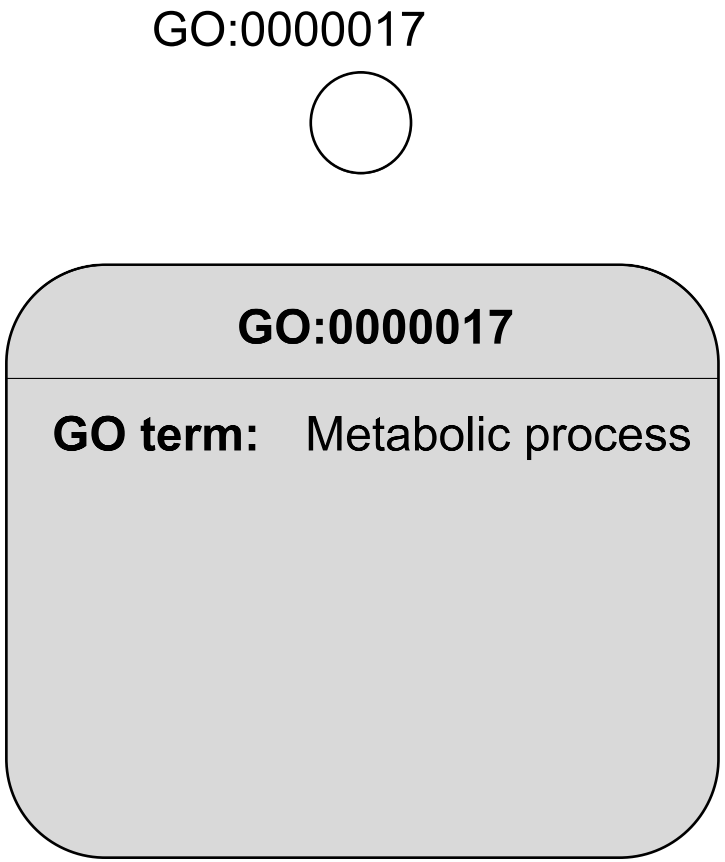
The Gene Ontology (GO) is a dynamic, frequently updated database



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:0000017

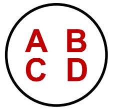
GO term: Metabolic process

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:

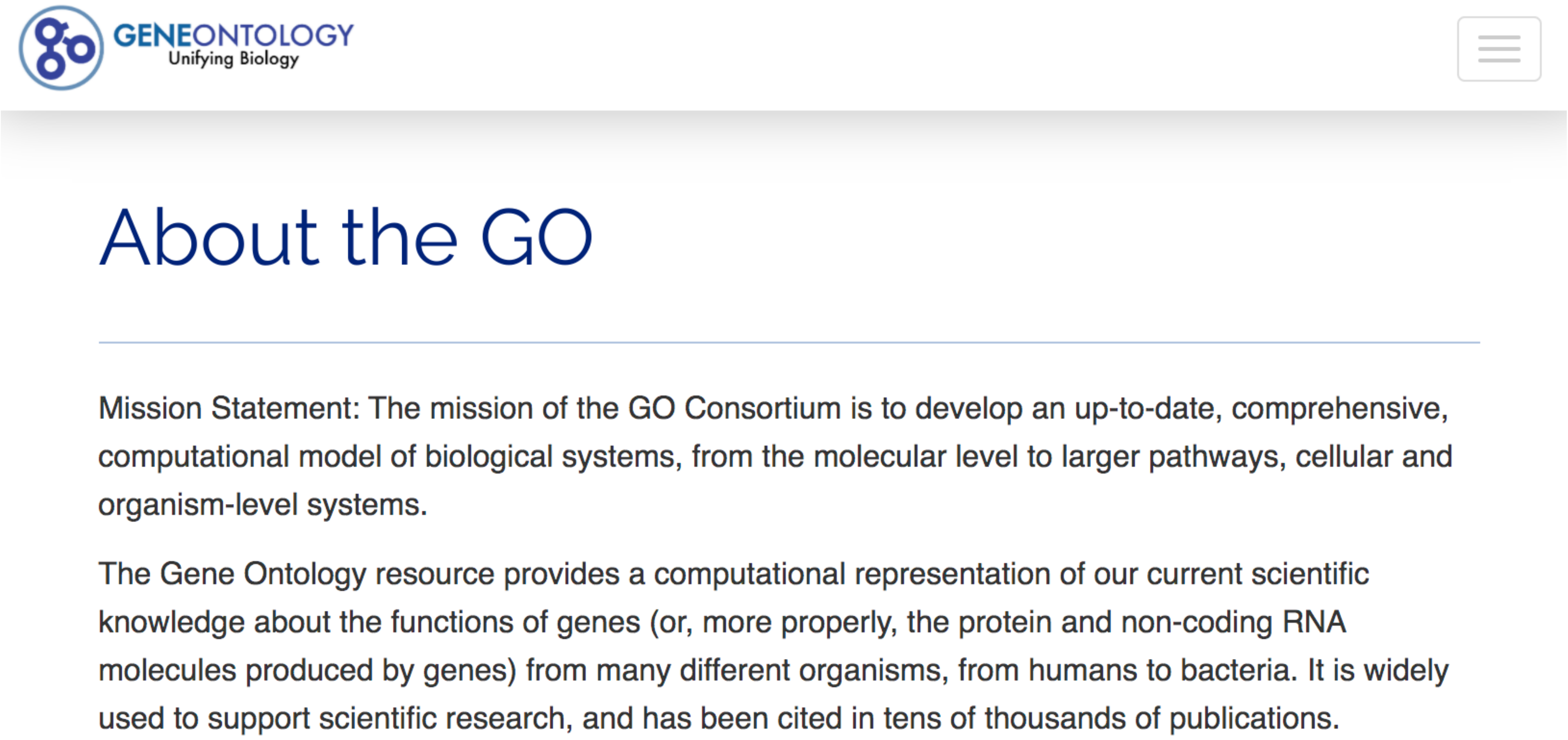
GO consortium organizes GO terms and their hierarchy

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)

Official GO database

GO Consortium is the provider of official Gene Ontology.



GENEONTOLOGY
Unifying Biology

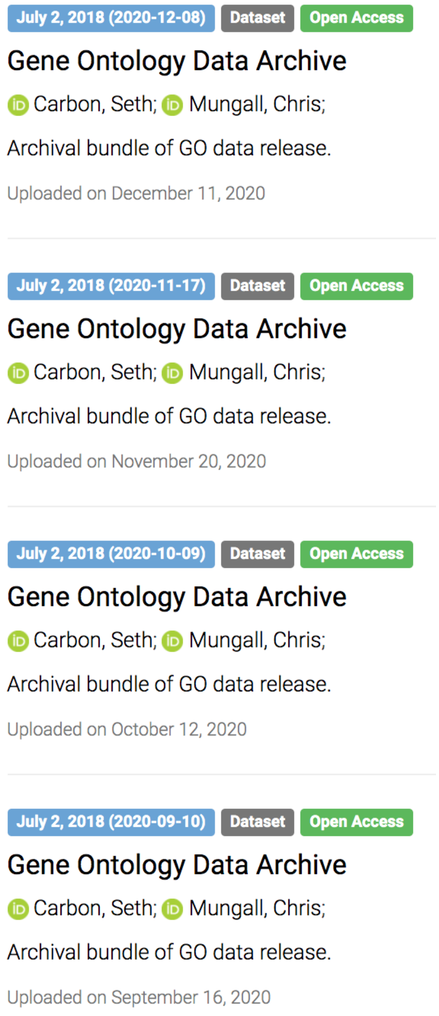
About the GO

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.

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;

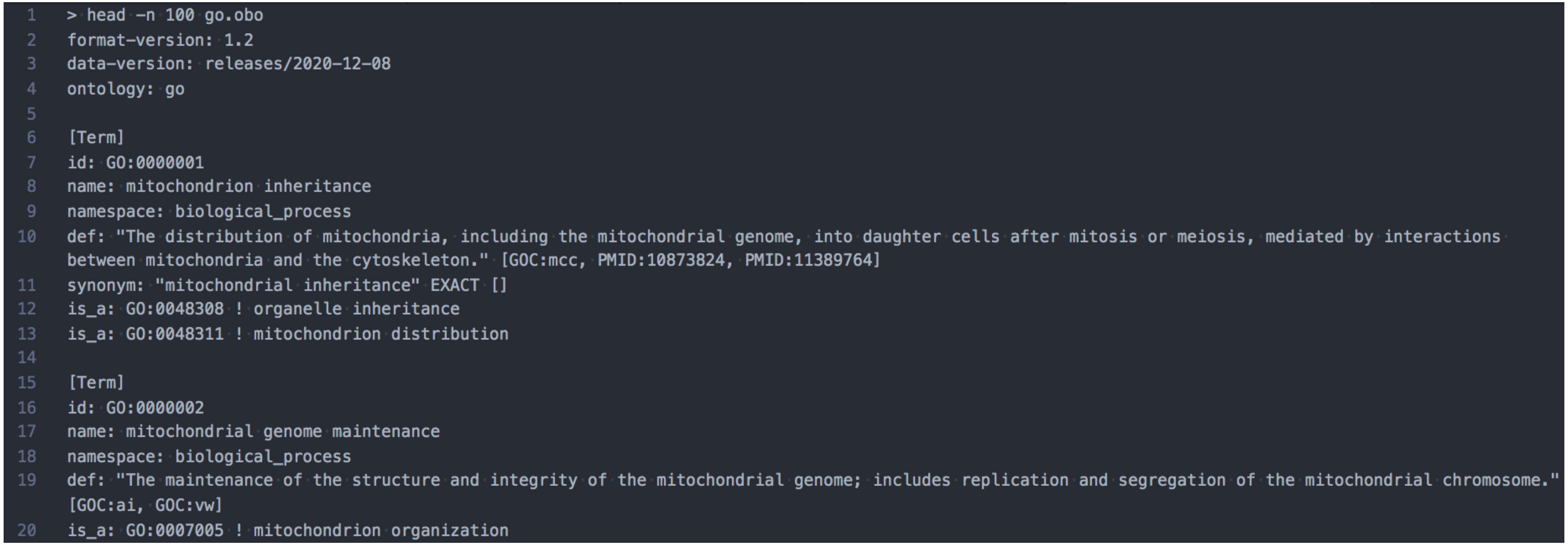
Archival bundle of GO data release.

Uploaded on September 16, 2020

Downloading official GO database

OBO format

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



```
> head -n 100 go.obo
format-version: 1.2
data-version: releases/2020-12-08
ontology: go

[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
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]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution

[Term]
id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial genome; includes replication and segregation of the mitochondrial chromosome." [GOC:ai, GOC:vw]
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
mgi.gaf.gz
mgi.gpad.gz
mgi.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
```

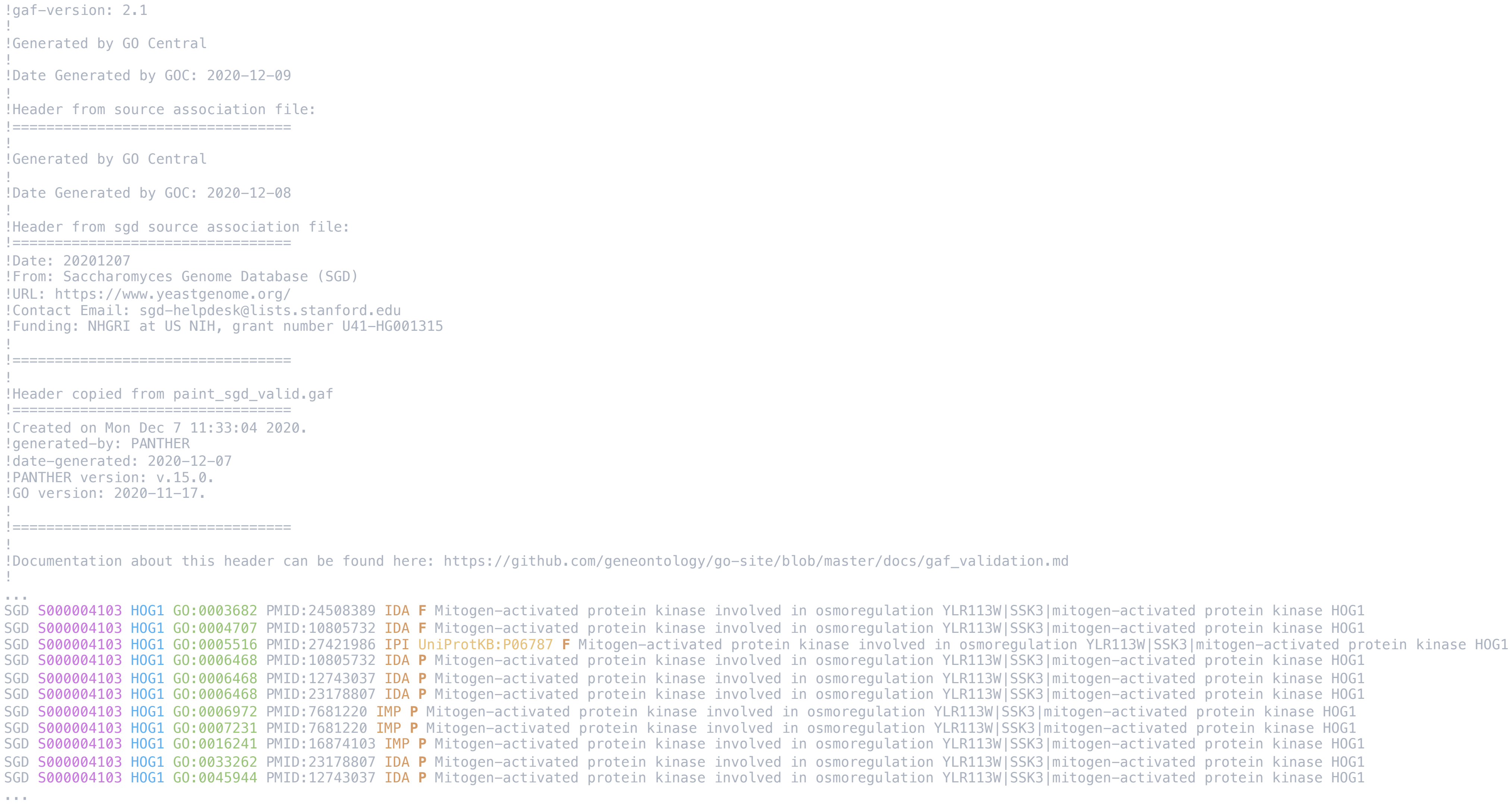
Downloading GO annotations

GAF format:

It’s in the name: **G**O **Annotation** **F**ormat

Downloading GO annotations

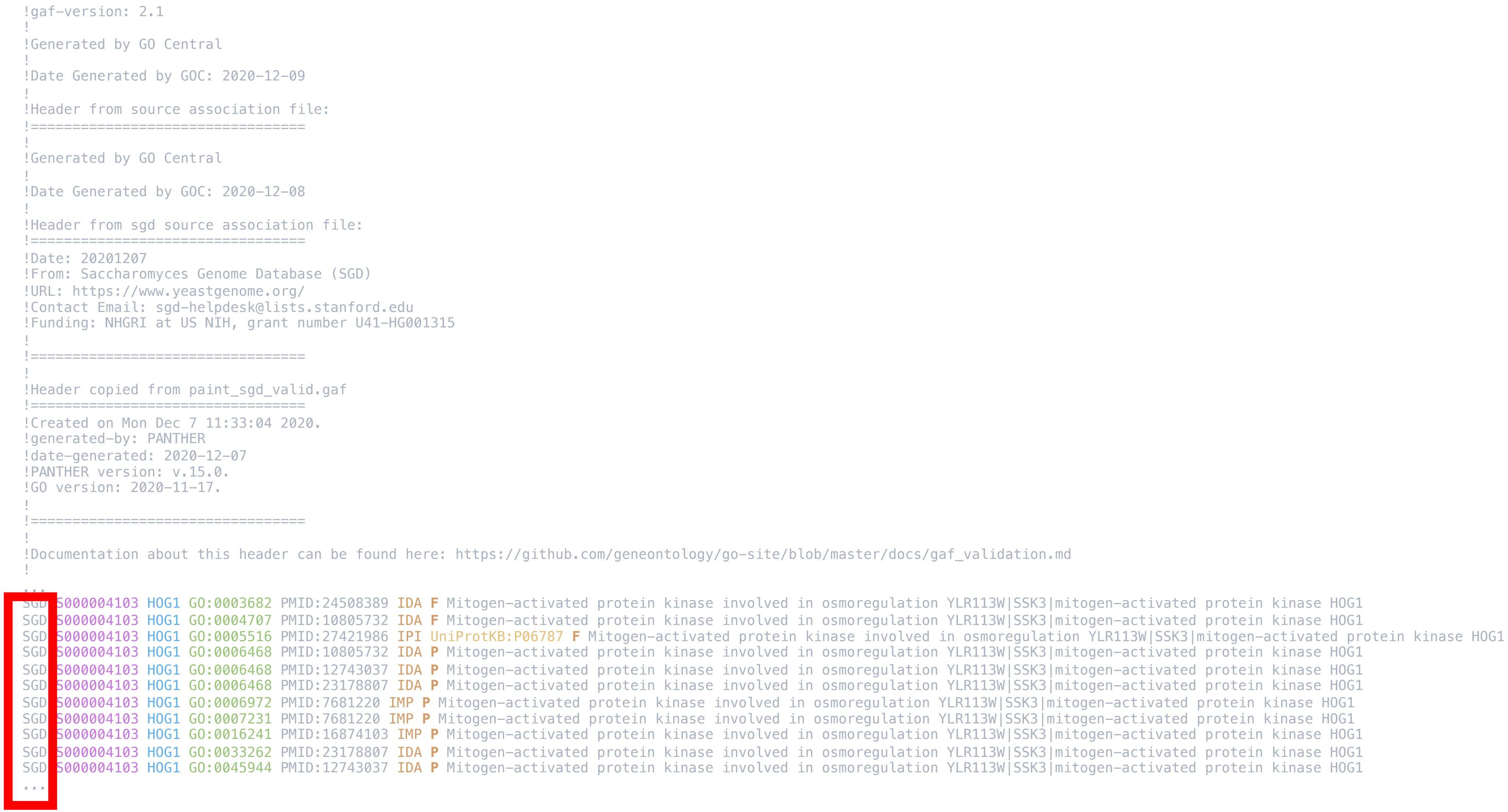
GAF format:



```
!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

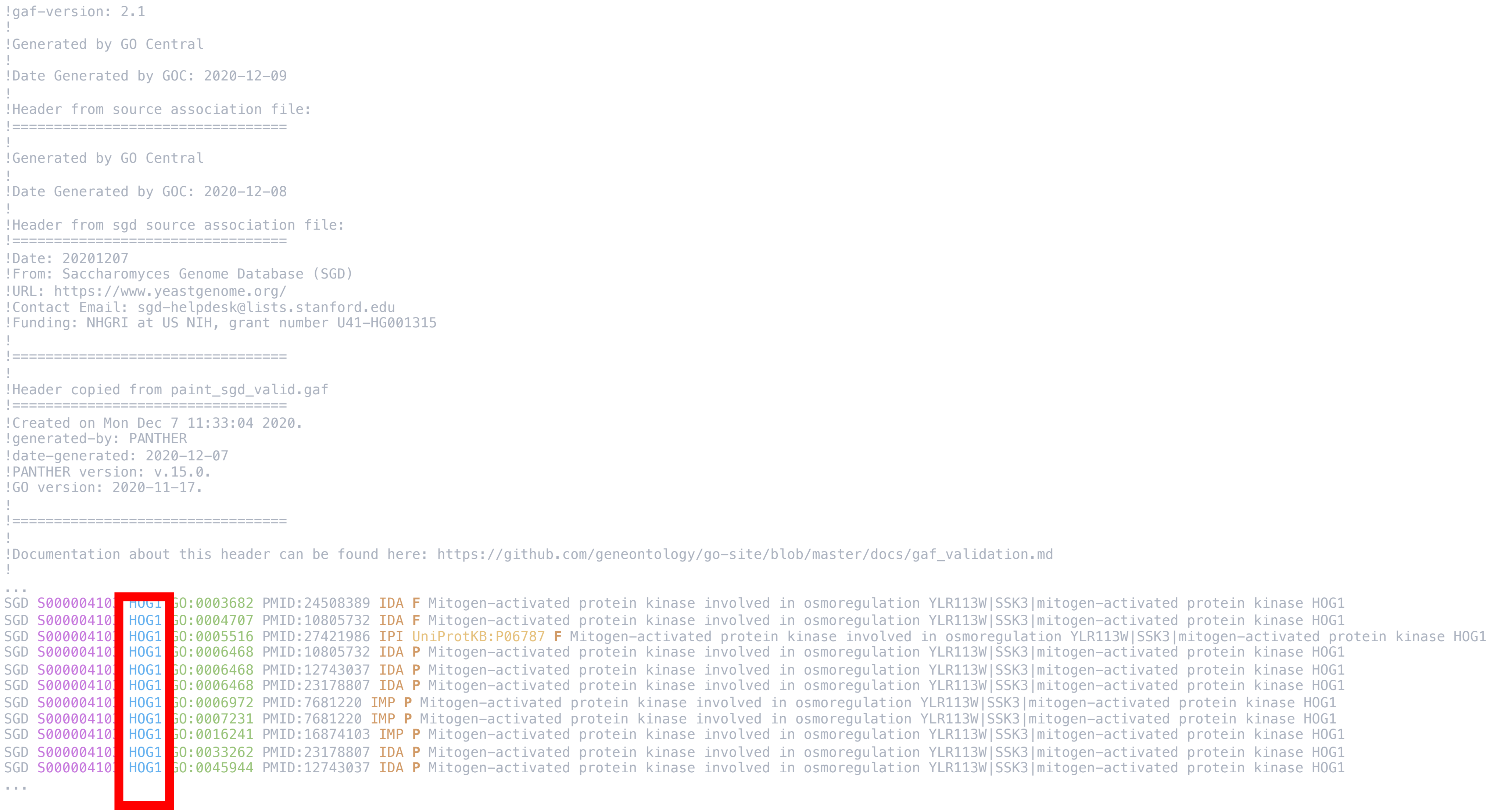


```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=================================
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=================================
!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

Downloading GO annotations

GAF format

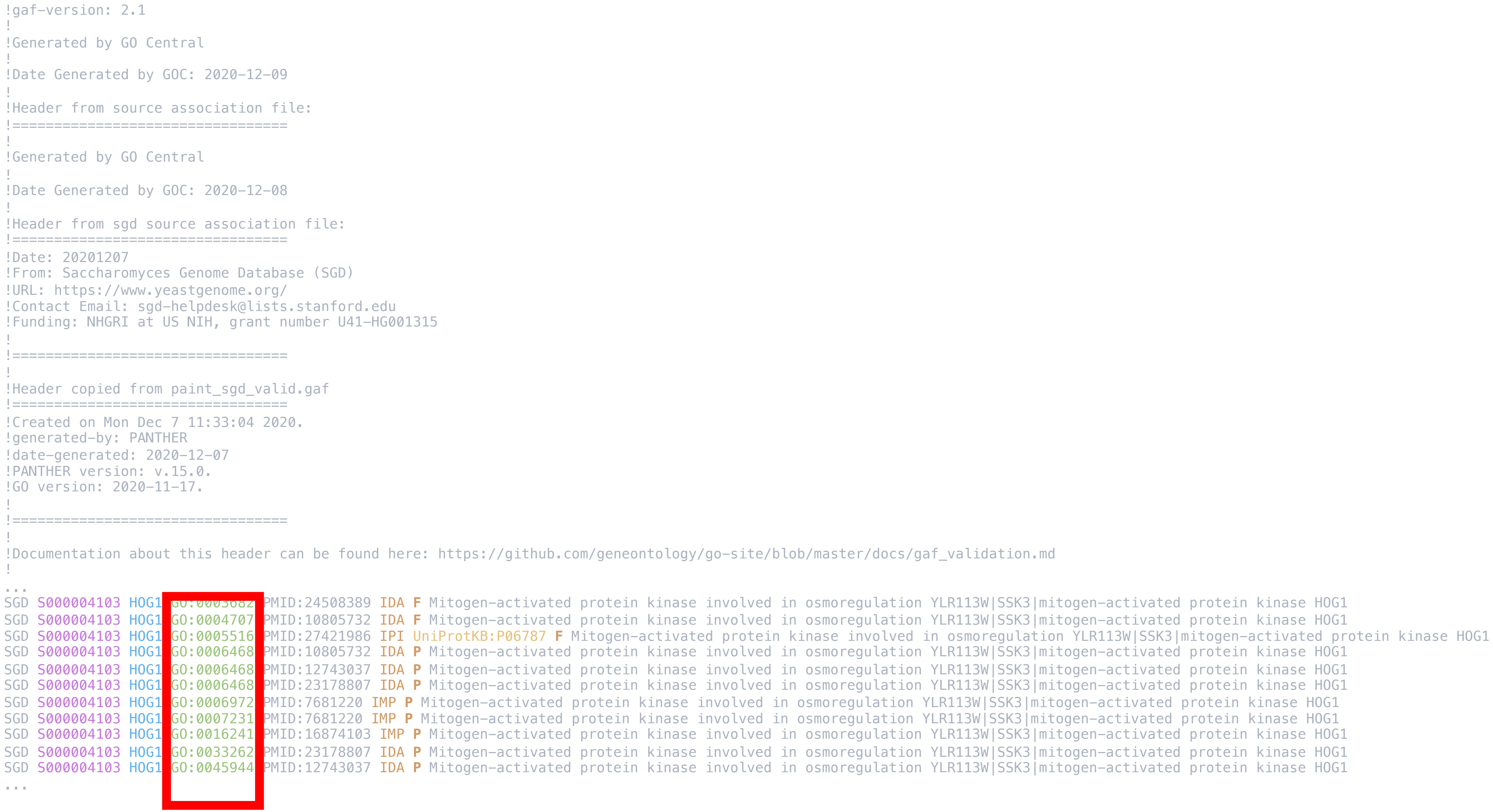


```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=================================
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=================================
!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 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

Downloading GO annotations

GAF format

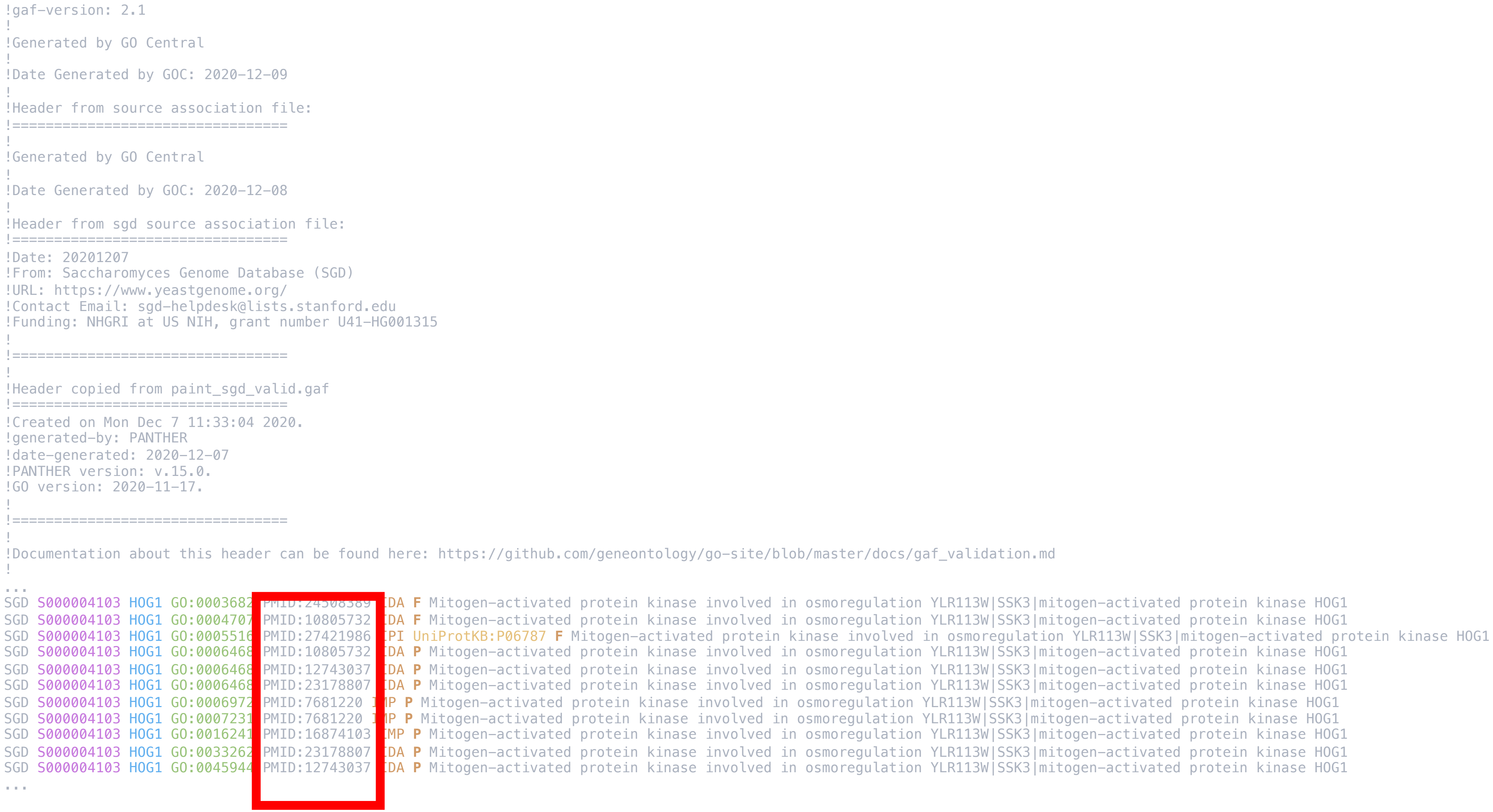


```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=================================
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=================================
!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
...
```

GO term

Downloading GO annotations

GAF format

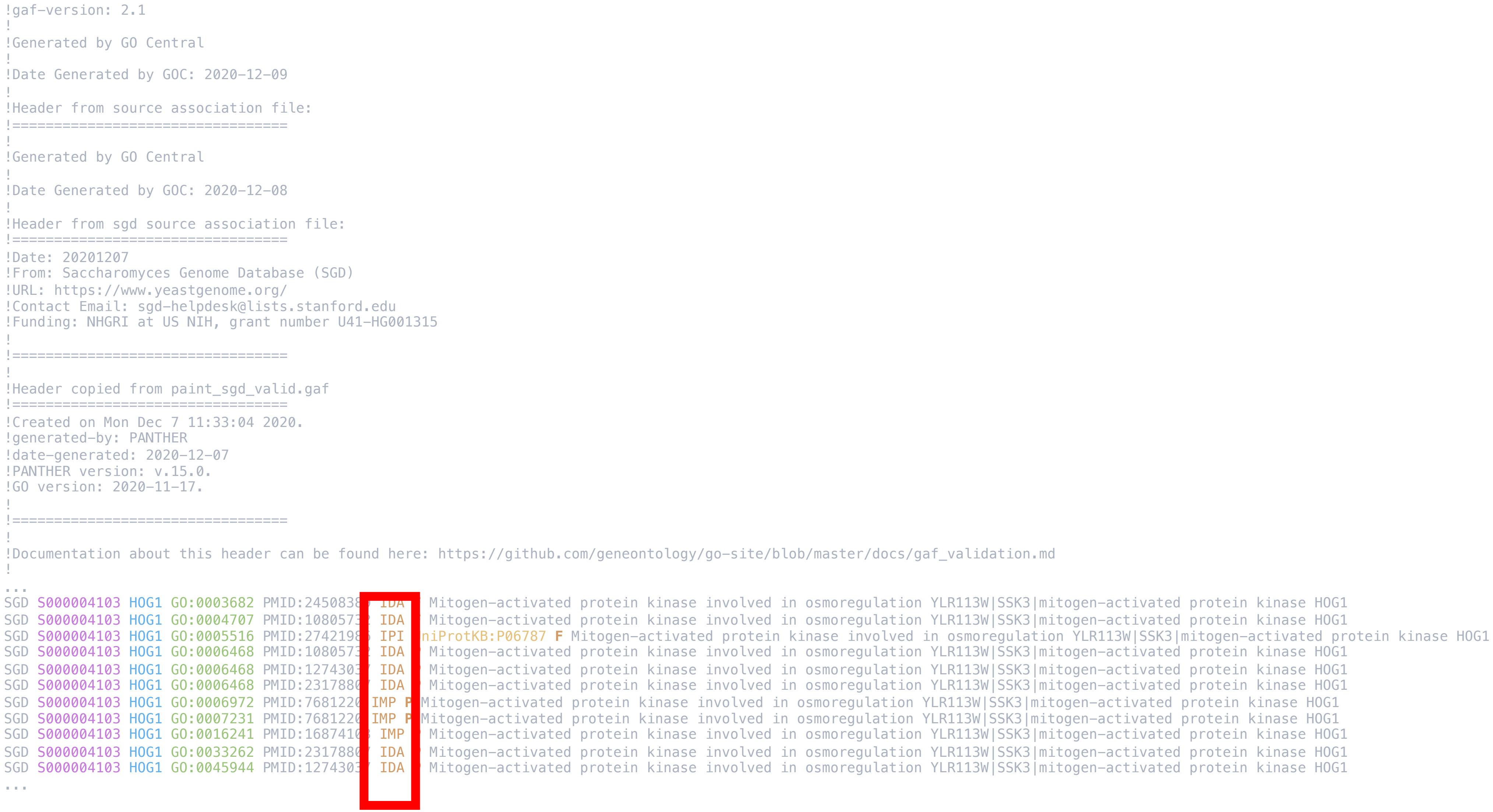


```
!gaf-version: 2.1
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-09
!
!Header from source association file:
!=================================
!
!Generated by GO Central
!
!Date Generated by GOC: 2020-12-08
!
!Header from sgd source association file:
!=================================
!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
...
```

Ref. for evidence

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:2450838  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  niProtKB: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 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: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

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

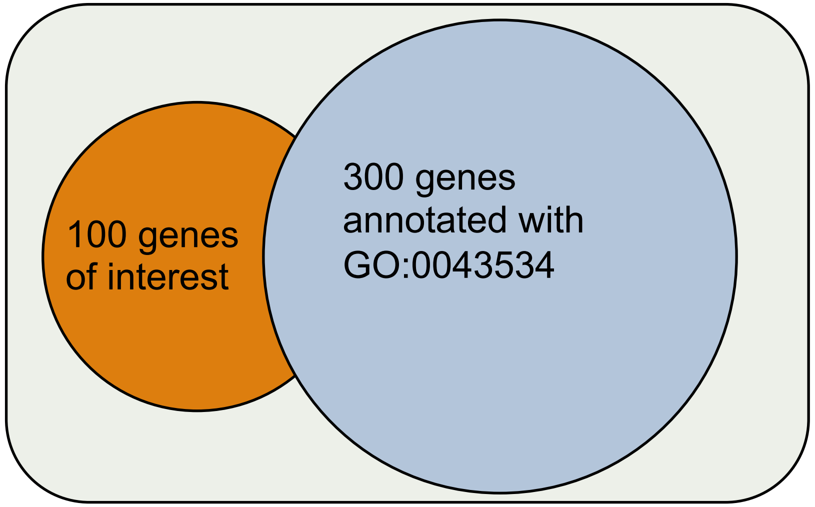
Guy who has drunk a ton of tea	Milk first	Tea added first
Cup is good	23	52
Cup is bad	43	12

Milk first / Tea first ?
Good / bad ?

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



Universe: ~5,000 genes with at least 1 GO annotation

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** ***in an assay*** ***(152)***	$\overline{\textbf{Genes over-expressed } \textit{in an assay}}$
Genes annotated ***in GO:0006836*** ***(243)***	89	154
$\overline{\textbf{Genes annotated } \textit{in GO:0006836}}$	63	5067 – 89 – 154 - 63

→ Sum = number of genes in GO:0006836 (243)

↓

↓ 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** ***in an assay*** ***(152)***	$\overline{\textbf{Genes over-expressed } \textbf{\textit{in an assay}}}$
Genes annotated ***in GO:0006836*** ***(243)***	89	154
$\overline{\textbf{Genes annotated } \textbf{\textit{in GO:0006836}}}$	63	5067 – 89 – 154 - 63

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

GO over-representation analyses: how

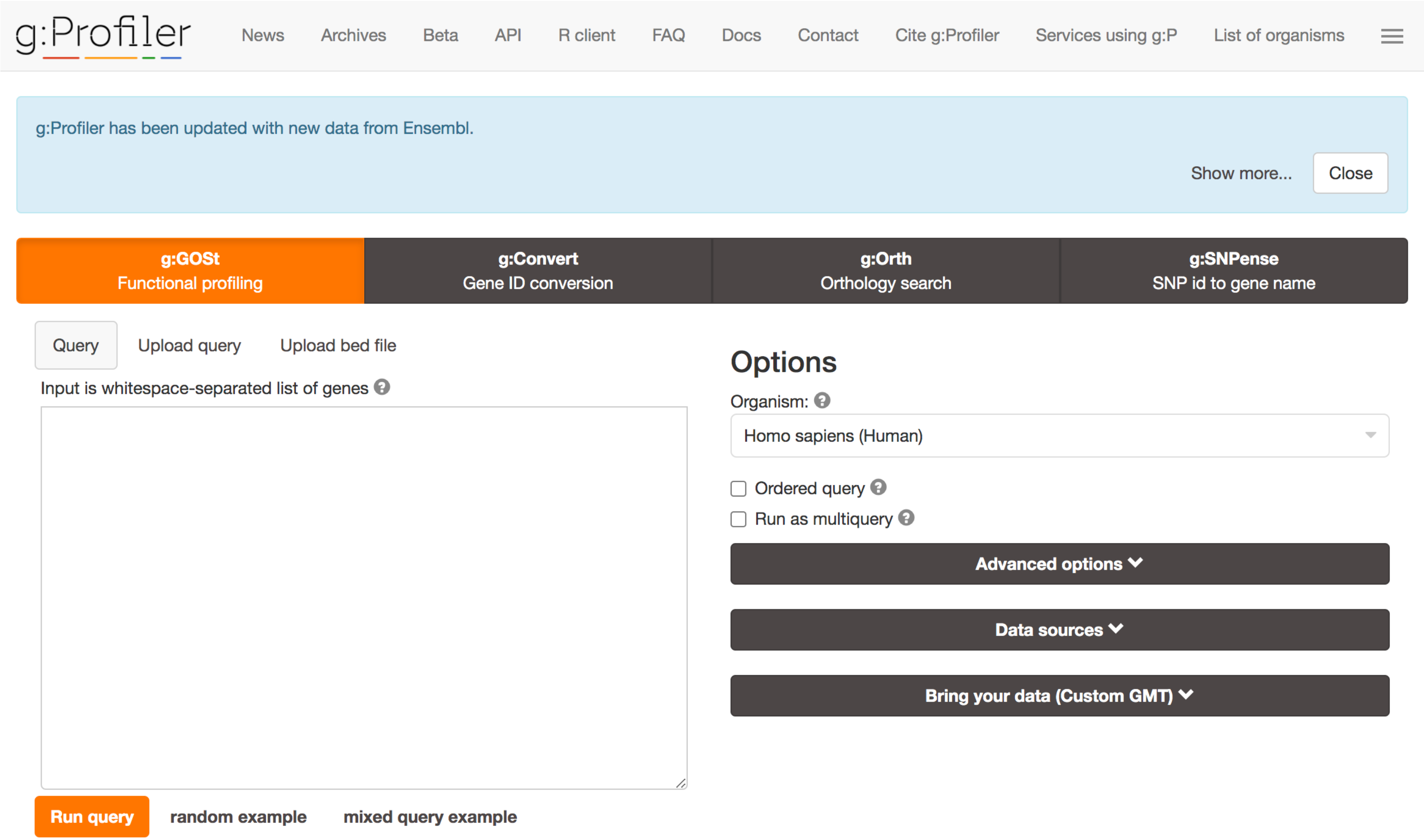
UNIVERSE = **All Yeast genes annotated in the Biological Processes (5067 genes)**	**Genes over-expressed** ***in an assay*** ***(152)***	$\overline{\textbf{Genes over-expressed } \textit{in an assay}}$
Genes annotated ***in GO:0006836*** ***(243)***	89	154
$\overline{\textbf{Genes annotated } \textit{in GO:0006836}}$	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 has been updated with new data from Ensembl.

Input is whitespace-separated list of genes

Options

Organism: Homo sapiens (Human)

□ Ordered query

□ Run as multiquery

Programs to run GO over-representation analyses: gProfiler

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

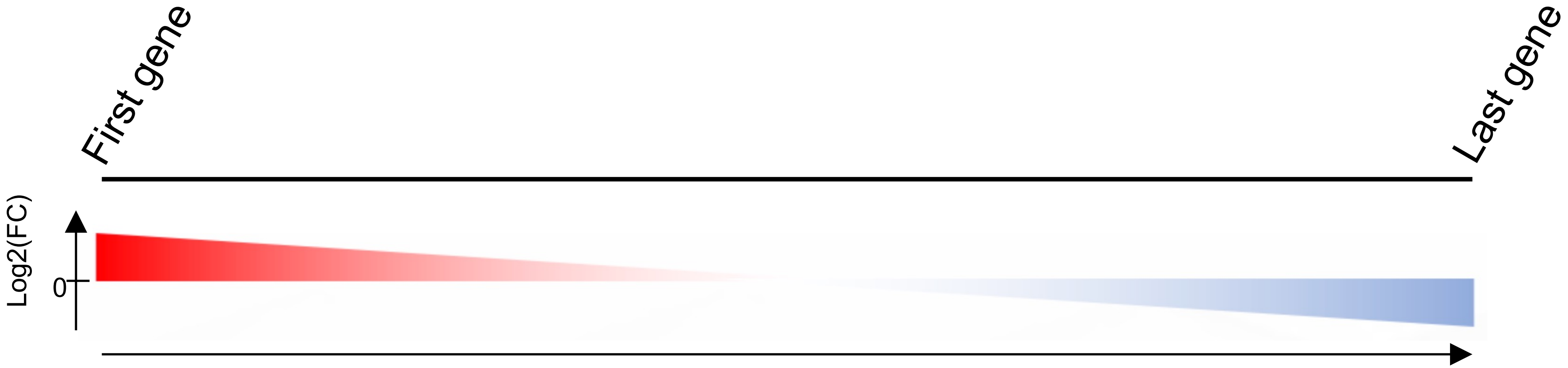
```
gprofiler2::gost(
    geneList,
    organism = 'scerevisiae’
)
```

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

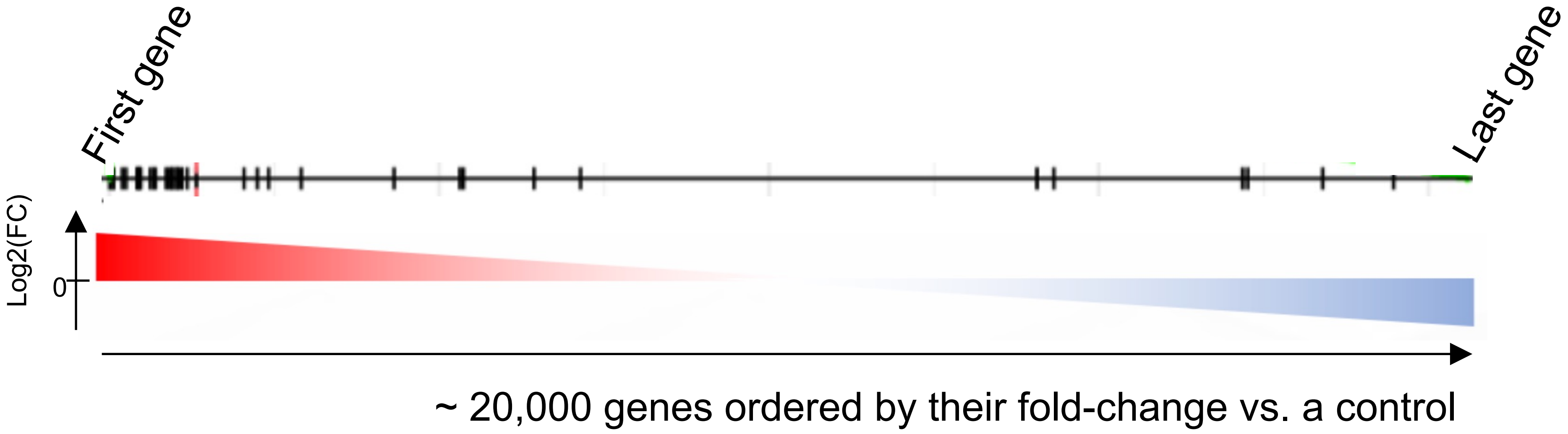
GSEA (Gene Set Enrichment Analysis) uses a ranked list of genes as input



~ 20,000 genes ordered by their fold-change vs. a control

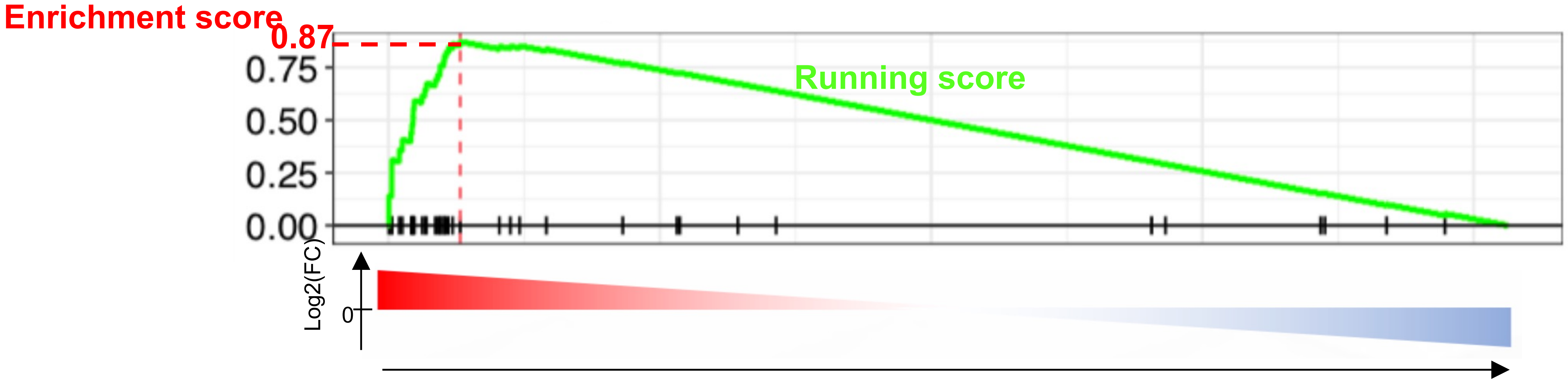
Gene Set Enrichment Analysis: cutoff-free functional analysis of ranked lists of genes

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

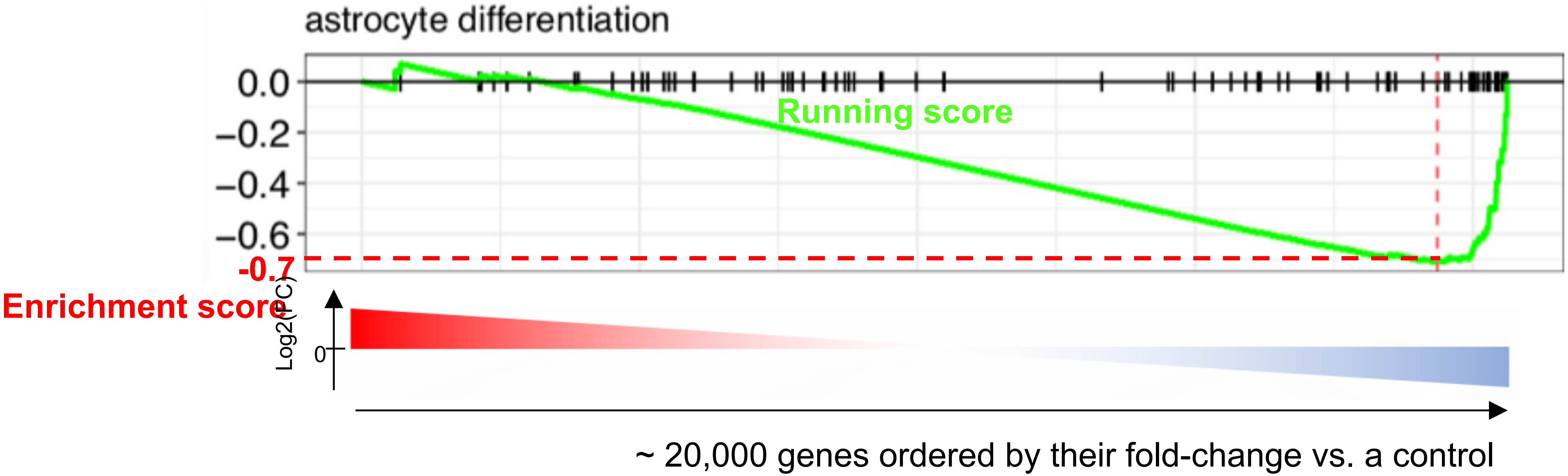
Based on the distribution of the flagged genes, it computes a “running score” and an “enrichment score”



~ 20,000 genes ordered by their fold-change vs. a control

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

It can also find negative enrichment scores (indicated 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

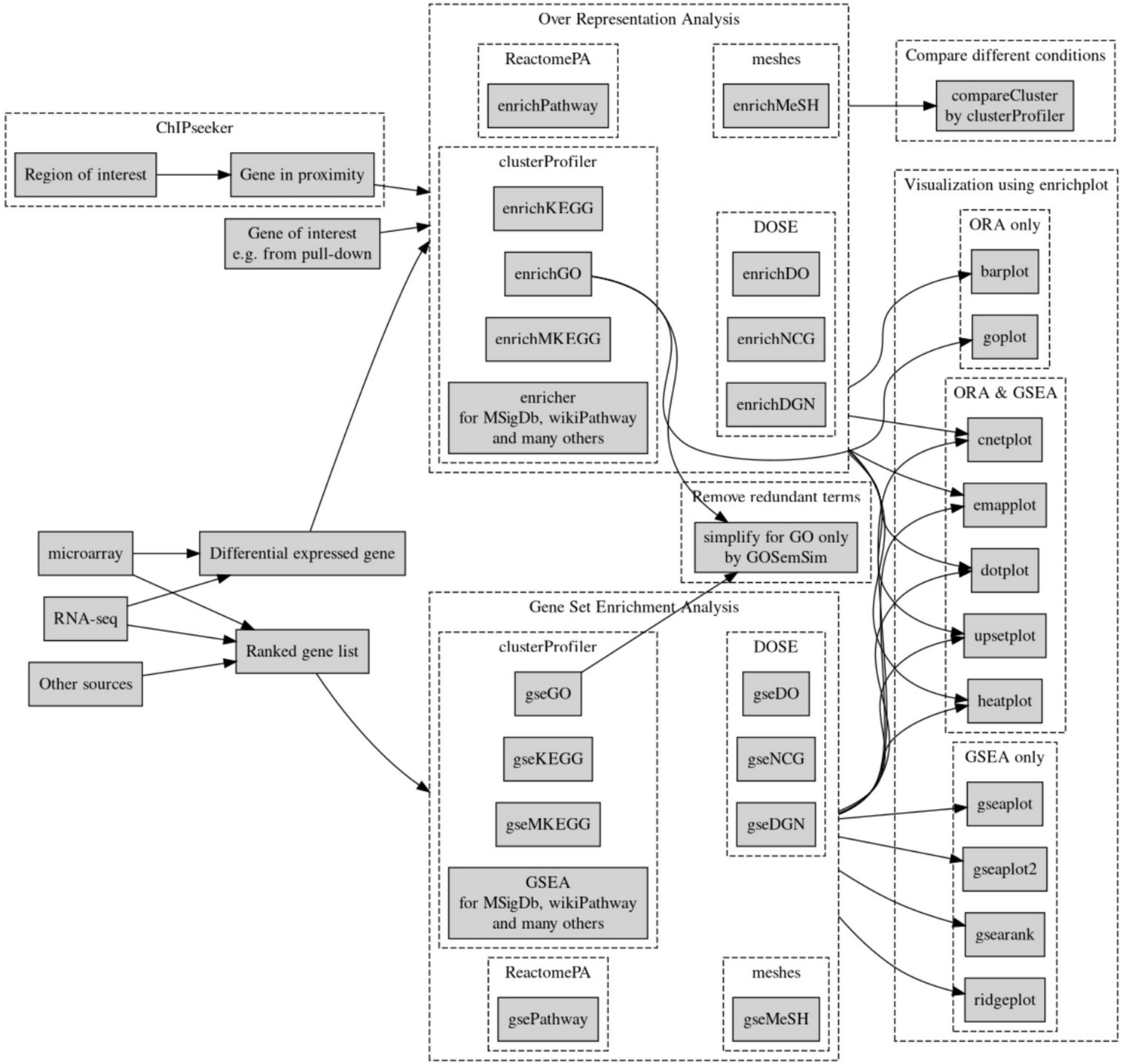
- I never managed to use it…

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, …)

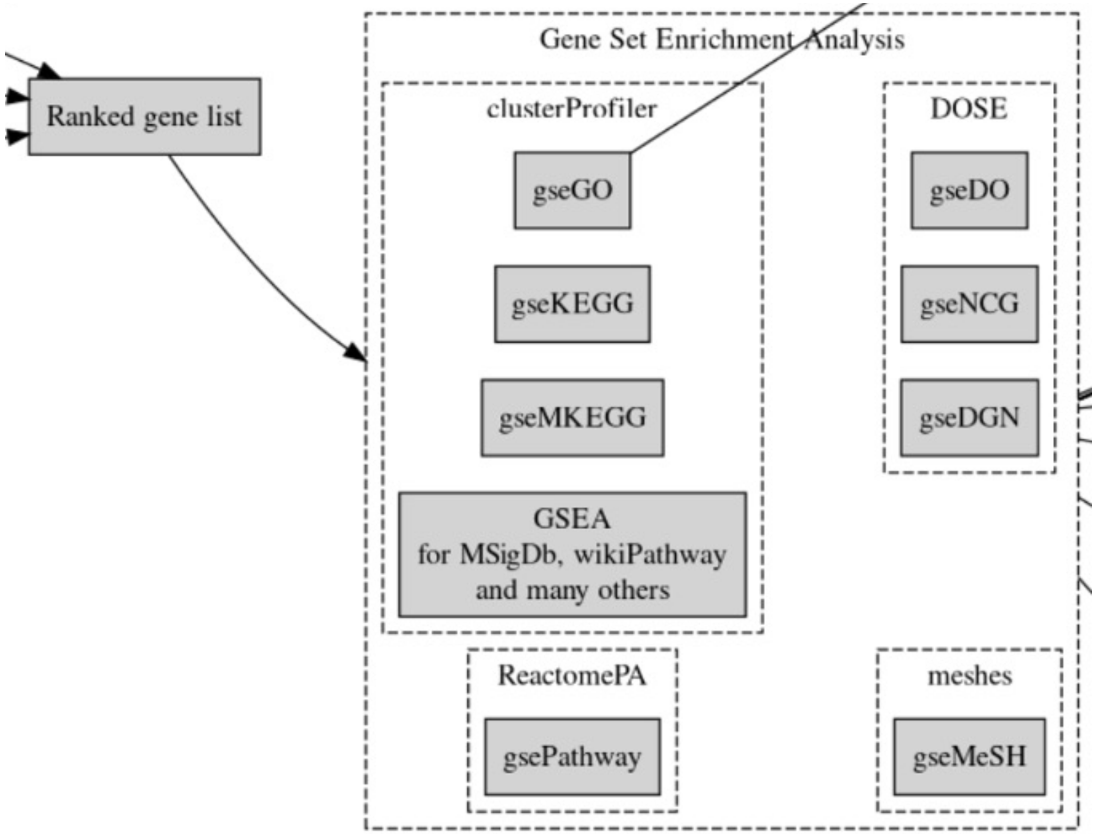
clusterProfiler

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

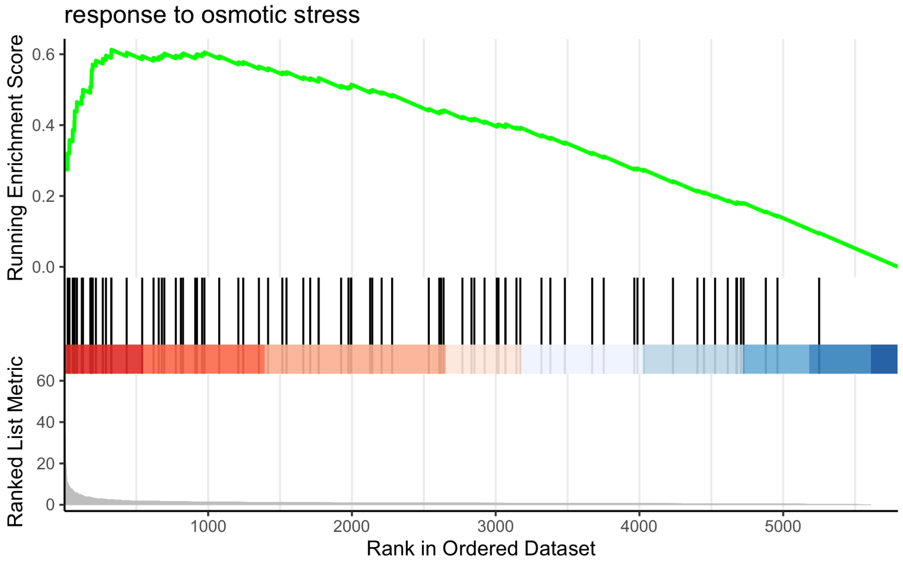
- gseGO() compares your ranked list of genes to the reference up-to-date GO annotations

```
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)
)
```



Resources

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

Tip 1a: Know the Source of the GO Annotations You Use
Tip 1b: Know the Version of the GO Annotations You Use

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