Gergely Csaba
LFE Bioinformatik
Institut für Informatik
Ludwig-Maximilians-Universität München

Hand-on :: Gene Set Enrichment

GoBi WS 2019/20

Motivation

Most comprehensive functional annotation of genes are provided by the Gene Ontology (GO).

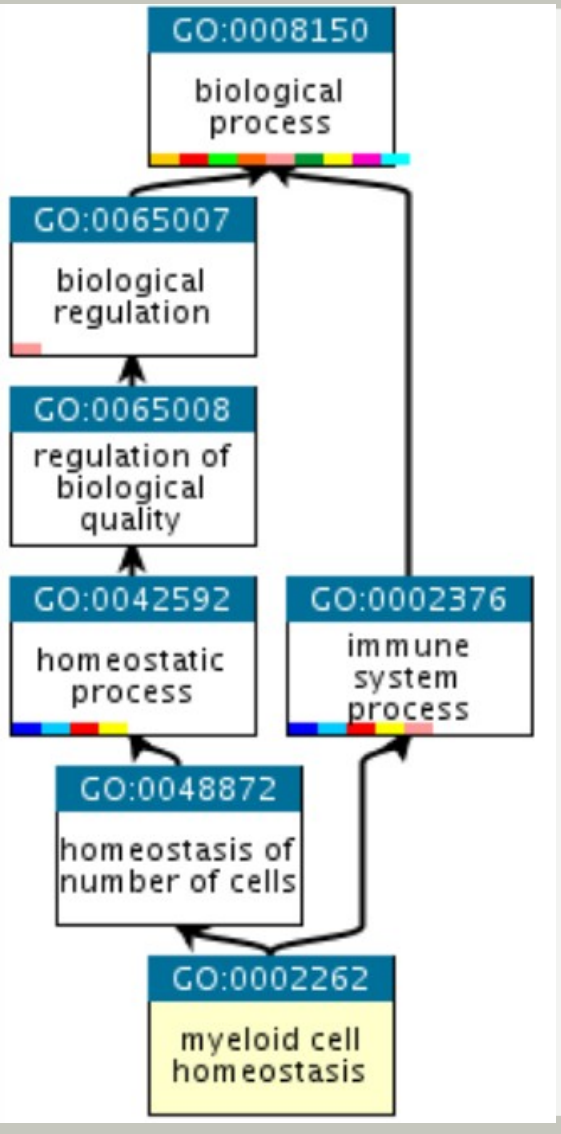
GO (and generally ontologies) is a directed acyclic graph (DAG) structure.

The DAG entries in GO along with a mapping (the associated genes to the GO entries of a given organism) enable simple enrichment analysis (for example on differential expression results) leading to insights of the involved molecular functions or biological processes.

but:

- genes are associated with multiple functions, in some cases where different ones – this is reflected by the complex overlapping nature of the DAG entries. These overlaps may lead to many false positive results in the statistical tests for set enrichment.
- there is no gold standard available for real experiments → to analyze the results of different enrichment strategies / the effect of overlaps we have to use simulations

GO is a DAG



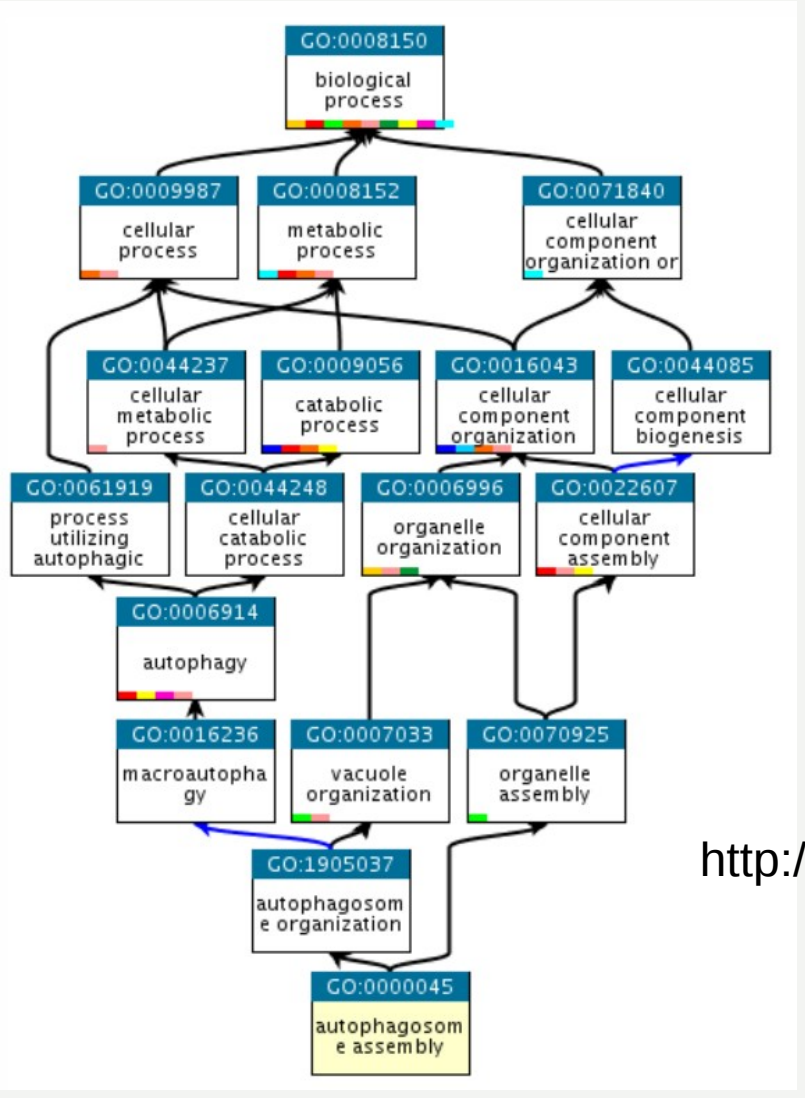
DAG (directed acyclic graph) → the entries may have multiple parents

obo file format:

```
[Term]
id: GO:0002262
name: myeloid cell homeostasis
namespace: biological_process
def: "The process of regulating the proliferation a
cells such that the total number of myeloid cells w
organism is stable over time in the absence of an e
[CL:0000763, GOC:add]
is_a: GO:0002376 ! immune system process
is_a: GO:0048872 ! homeostasis of number of cells
```

- record based (util BlockIterator?)
- **is_a** fields define the parents
- multiple DAG-s may be defined (defined by **namespace**)

mappings



genes are associated with different GO classes, based on different evidences

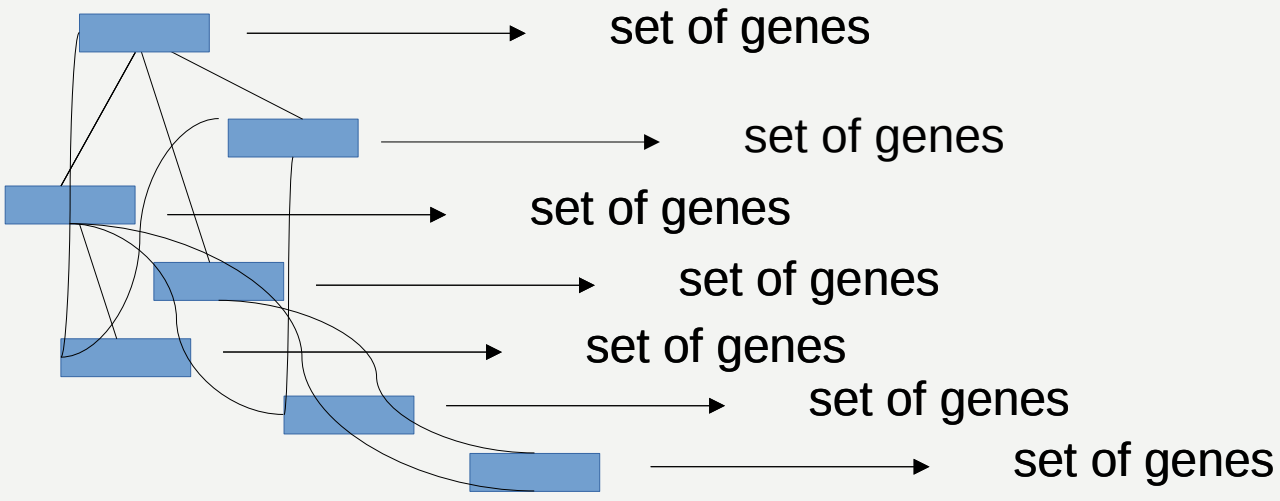
RAB1A	GO:0000045	IMP
RAB1A	GO:0000139	IBA
RAB1A	GO:0000139	TAS
RAB1A	GO:0000139	TAS
RAB1A	GO:0003924	IDA
RAB1A	GO:0003924	IDA
RAB1A	GO:0005515	IPI
RAB1A	GO:0005515	IPI

http://geneontology.org/page/guide-go-evidence-codes

association of gene g to class x implies associations to all x ascendents!

overlaps, LCA-s

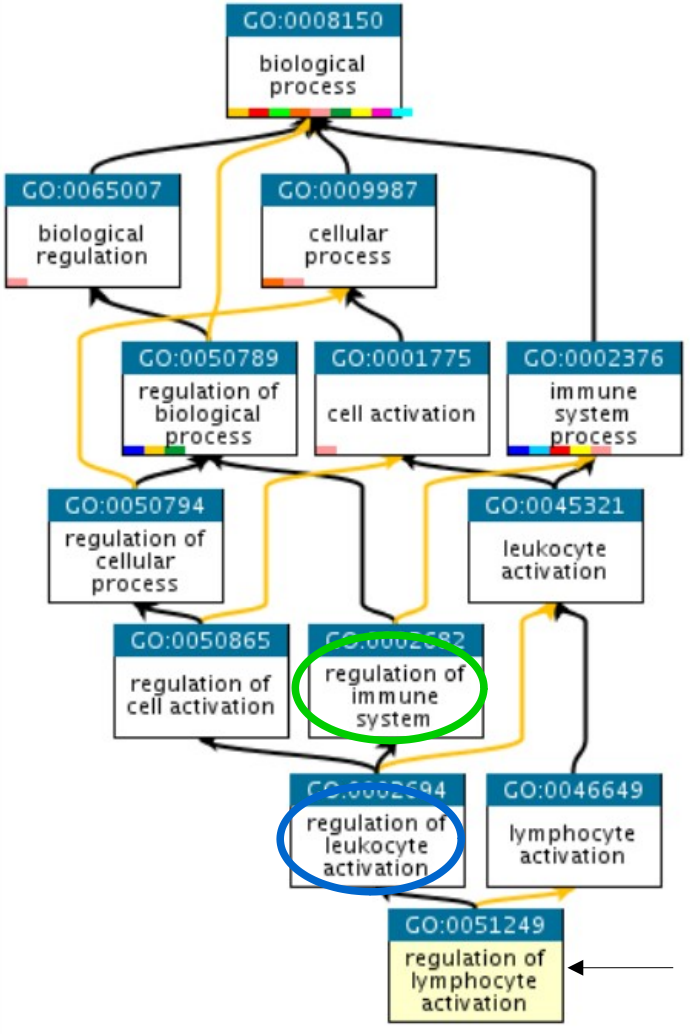
Given GO + mapping we have a structure:



overlap between two DAG entries is defined by the intersection size of their sets of genes

(take care ~ 30.000 in GO:BP → ~ 450 Mio DAG entry pairs →
comparing all pairs takes too long, and is not needed)

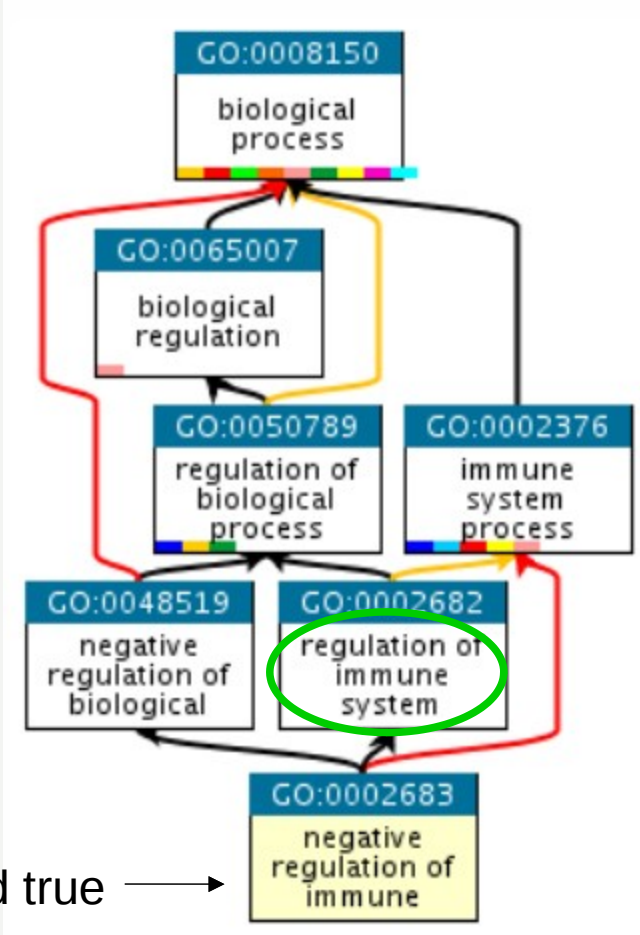
LCA-s



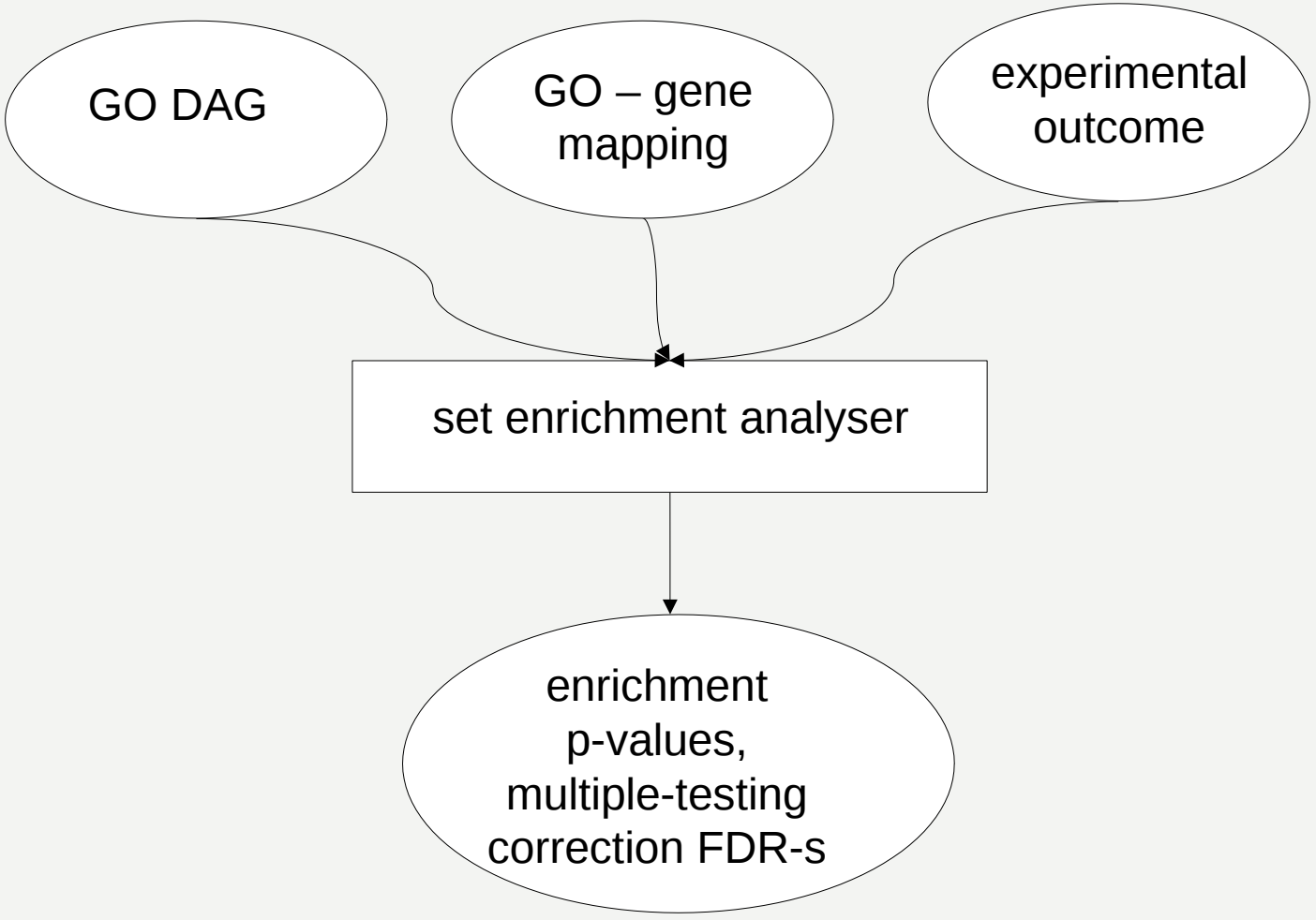
For the DAG structure consider only the black arrows (**is_a** relationship)

LCA

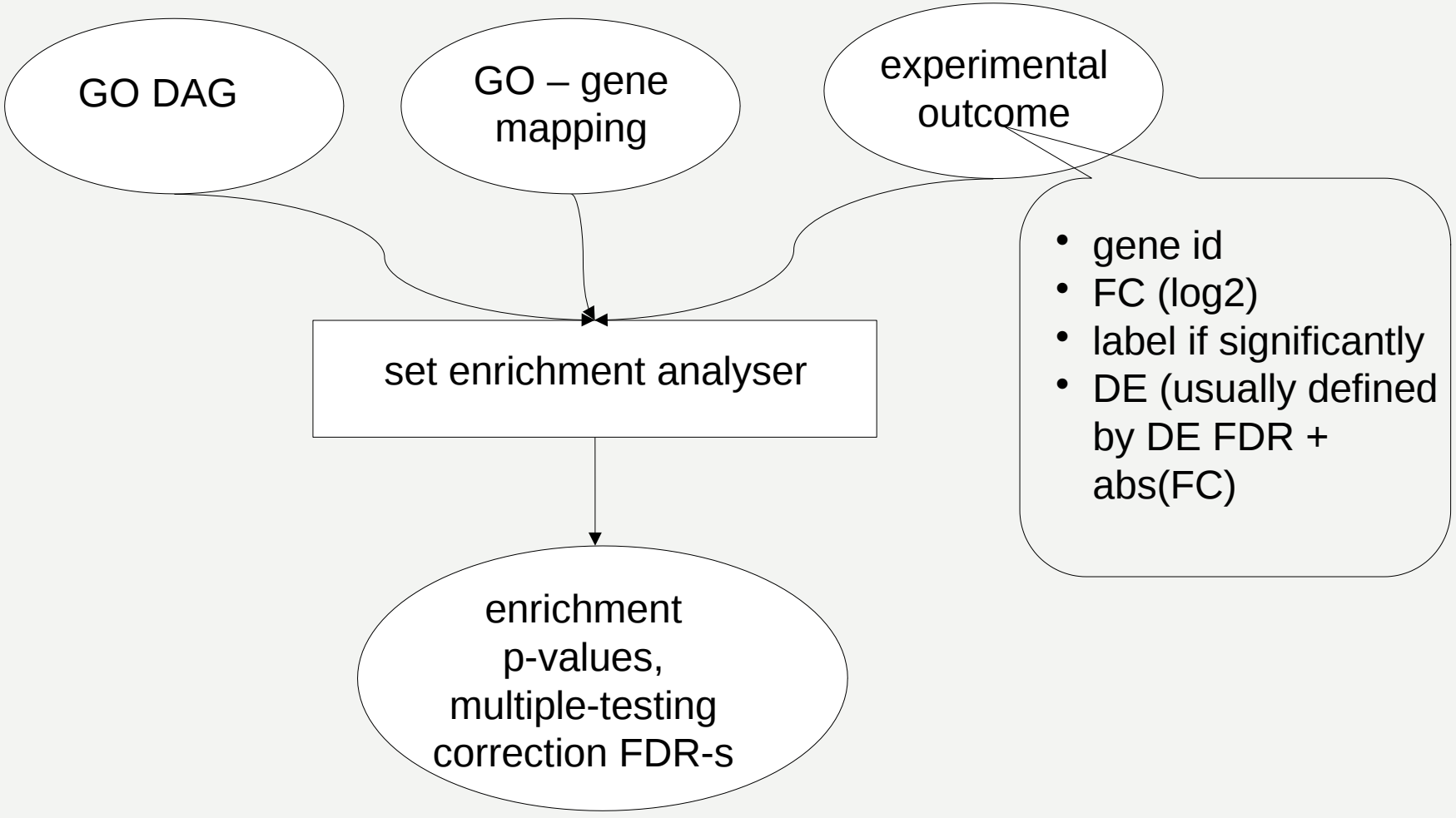
used path



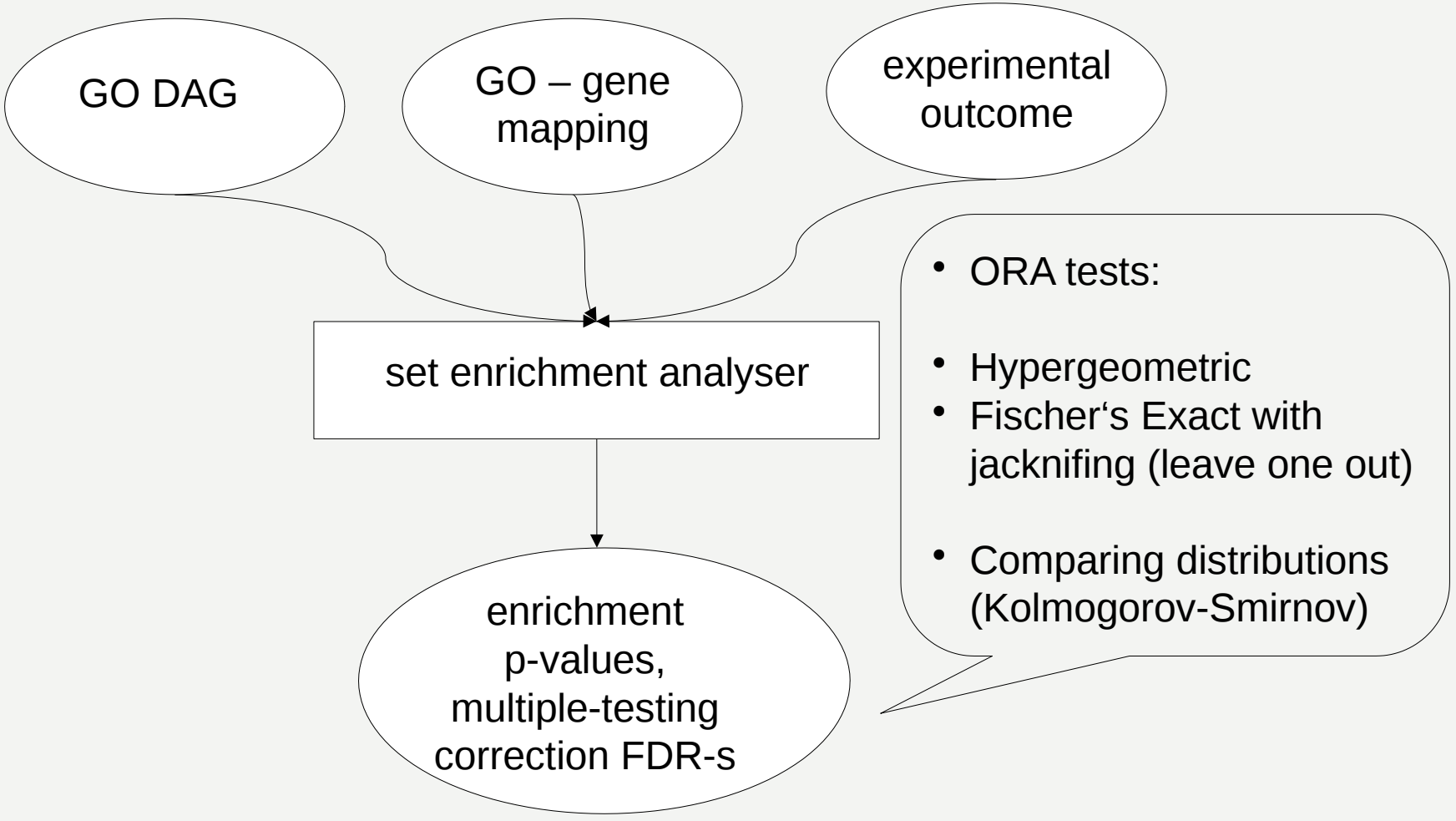
Workflow


GO DAG
GO – gene mapping
experimental outcome
set enrichment analyser
enrichment p-values, multiple-testing correction FDR-s

Workflow


GO DAG
GO – gene mapping
experimental outcome
set enrichment analyser
enrichment p-values, multiple-testing correction FDR-s
gene id
FC (log2)
label if significantly DE (usually defined by DE FDR + abs(FC)

Workflow


GO DAG
GO – gene mapping
experimental outcome
set enrichment analyser
enrichment p-values, multiple-testing correction FDR-s
ORA tests:
Hypergeometric
Fischer‘s Exact with jacknifing (leave one out)
Comparing distributions (Kolmogorov-Smirnov)

Statistical tests

```
import org.apache.commons.math3.distribution.HypergeometricDistribution;
```

```
HypergeometricDistribution hg = new HypergeometricDistribution(...);
return hg.upperCumulativeProbability(...);
```

```
import org.apache.commons.math3.stat.inference.KolmogorovSmirnovTest;
```

```
KolmogorovSmirnovTest ks = new KolmogorovSmirnovTest();
double[] in_set_distrib;
double[] bg_distrib;
ks.kolmogorovSmirnovTest(in_set_distrib, bg_distrib);
ks.kolmogorovSmirnovStatistic(in_set_distrib, bg_set_distrib);
```

Fischer's Exact

Fischer Exact

	In set	Not in set	Row total
Significant DE	a	b	a+b
Non-significant DE	c	d	c+d
Column Total	a+c	b+d	a + b + c + d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Hypergeometric

N: total genes
K : DE genes
n: set size
k: overlap(DE genes, set genes)

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

Fischer exact = Hypergeometric with:

N = a+b+c+d
K = a+b
n = a+c
K = a

Hypergeometric = Fischer with:

a = k
b = K – k
c = N – k
d = N – n – K + k

Questions?