

## V8 – Functional annotation

Program for today:

- Functional annotation of genes/gene products: Gene Ontology (GO)
- significance of annotations: hypergeometric test
- (mathematical) semantic similarity of GO-terms
- **Issues** in GO-analysis
  - protein annotation is biased and is influenced by different research interests:
  - model organisms of human disease are better annotated
  - promising gene products (e.g. disease associated genes) or specific gene families have a higher number of annotations
  - gene with early gene-bank entries have on average more annotations

# Primer on the Gene Ontology

The key motivation behind the Gene Ontology (GO) was the observation that similar genes often have conserved functions in different organisms.

A common vocabulary was needed to be able to compare the roles of **orthologous** (→ evolutionarily related) genes and their products across different species.

A **GO annotation** is the association of a gene product with a GO term

GO allows capturing isoform-specific data when appropriate. For example, UniProtKB accession numbers P00519-1 and P00519-2 are the isoform identifiers for isoform 1 and 2 of P00519.

# The Gene Ontology (GO)

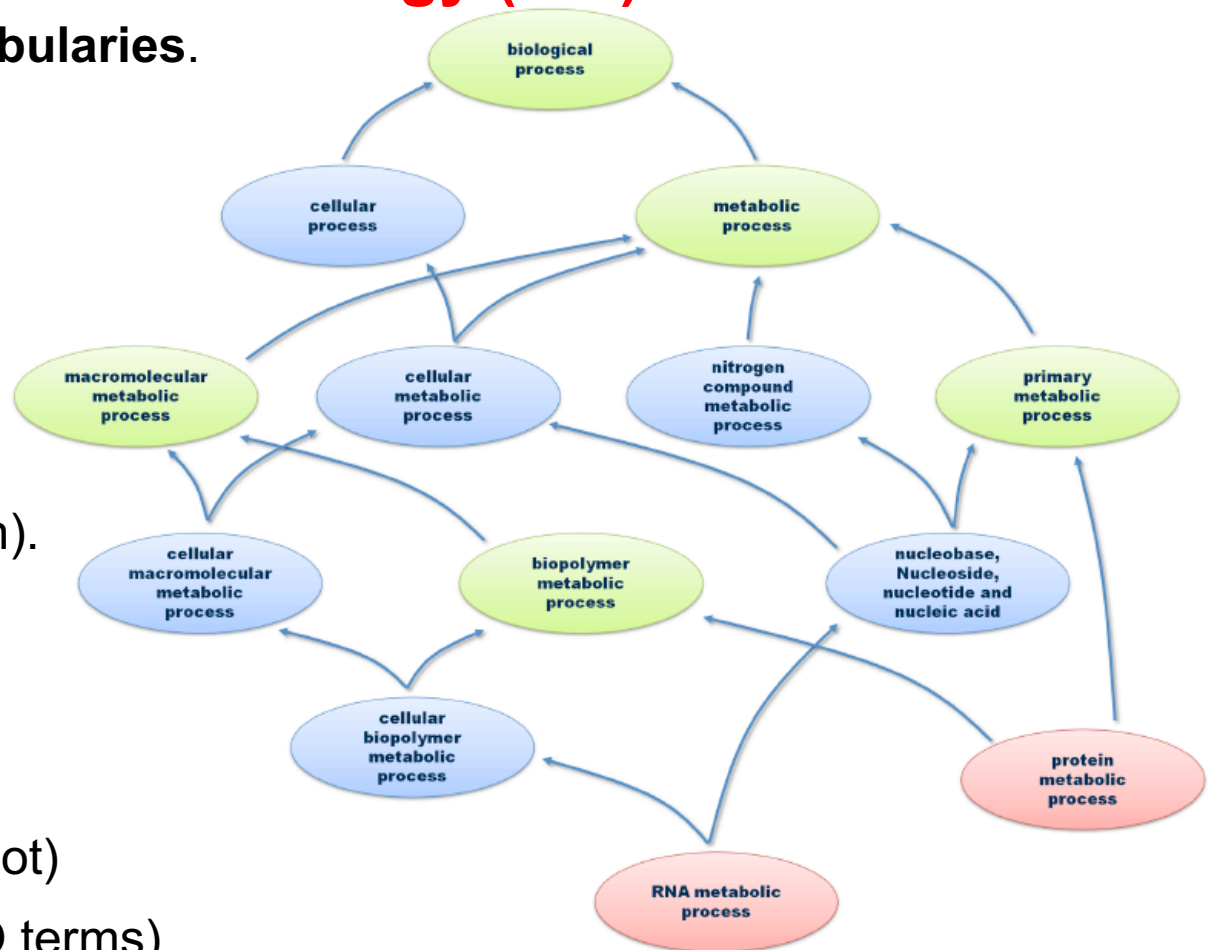
Ontologies are **structured vocabularies**.

The Gene Ontology consists of

3 non-redundant areas:

- Biological process (BP)
- molecular function (MF)
- cellular component (localisation).

Shown here is a part of the BP vocabulary.



At the top: most general term (root)

**Red**: tree leafs (very specific GO terms)

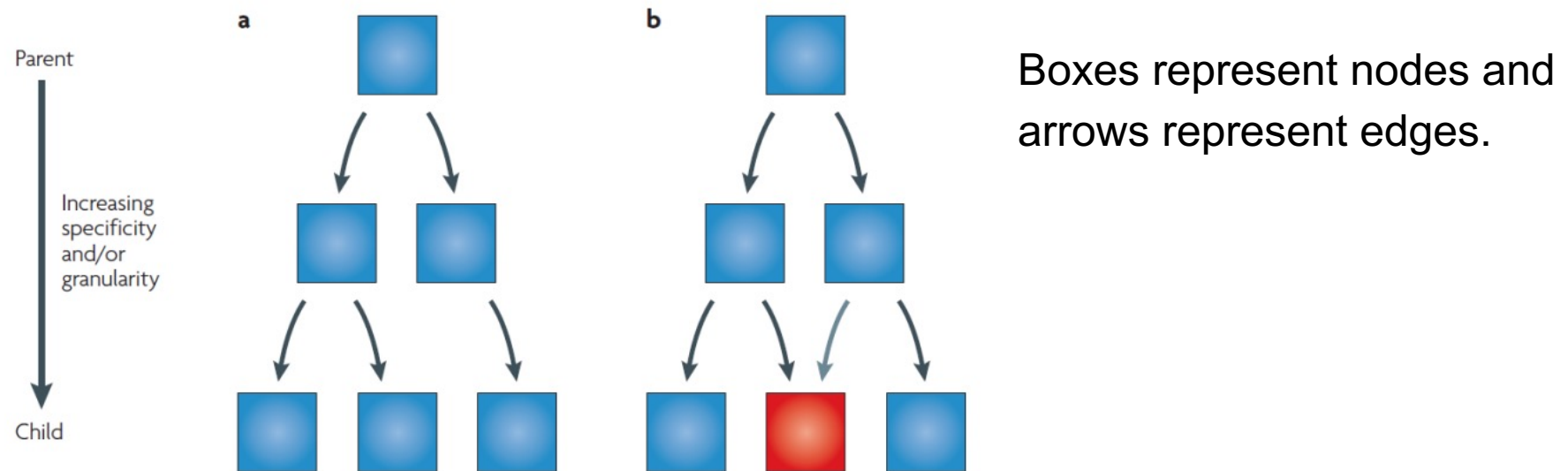
**Green**: common ancestor

**Blue**: other nodes.

Arcs: relations between parent and child nodes

PhD Dissertation Andreas Schlicker (UdS, 2010)

# Simple tree vs. cyclic graphs



**a** | An example of a simple **tree**, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge.

**b** | A **directed acyclic graph** (DAG), in which each child can have one or more parents. The node with multiple parents is coloured red and the additional edge is coloured grey.

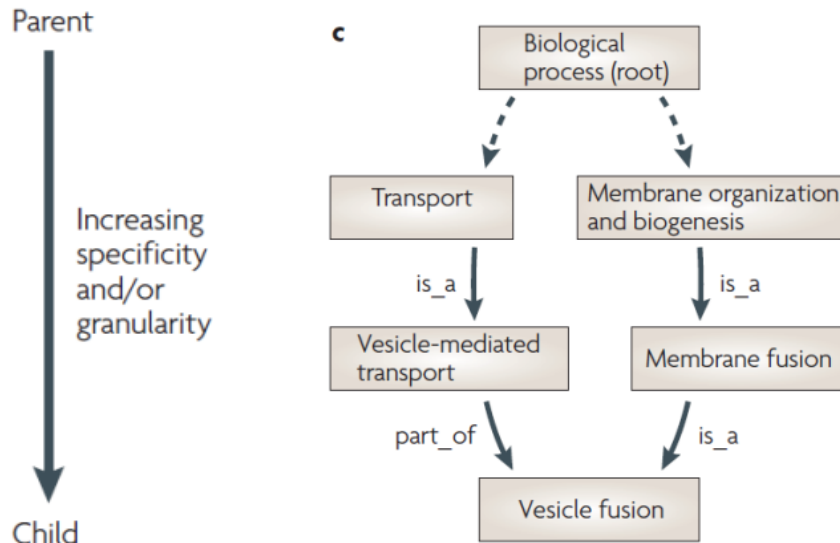
Rhee et al. (2008) Nature

Rev. Genet. 9: 509

V8

Processing of Biological Data

# Gene Ontology is a directed acyclic graph



An example of the node *vesicle fusion* in the BP ontology with multiple parentage.

**Dashed edges** : there are other nodes not shown between the nodes and the root node.

**Root** : node with no incoming edges, and at least one leaf.

**Leaf node** : a terminal node with no children (vesicle fusion).

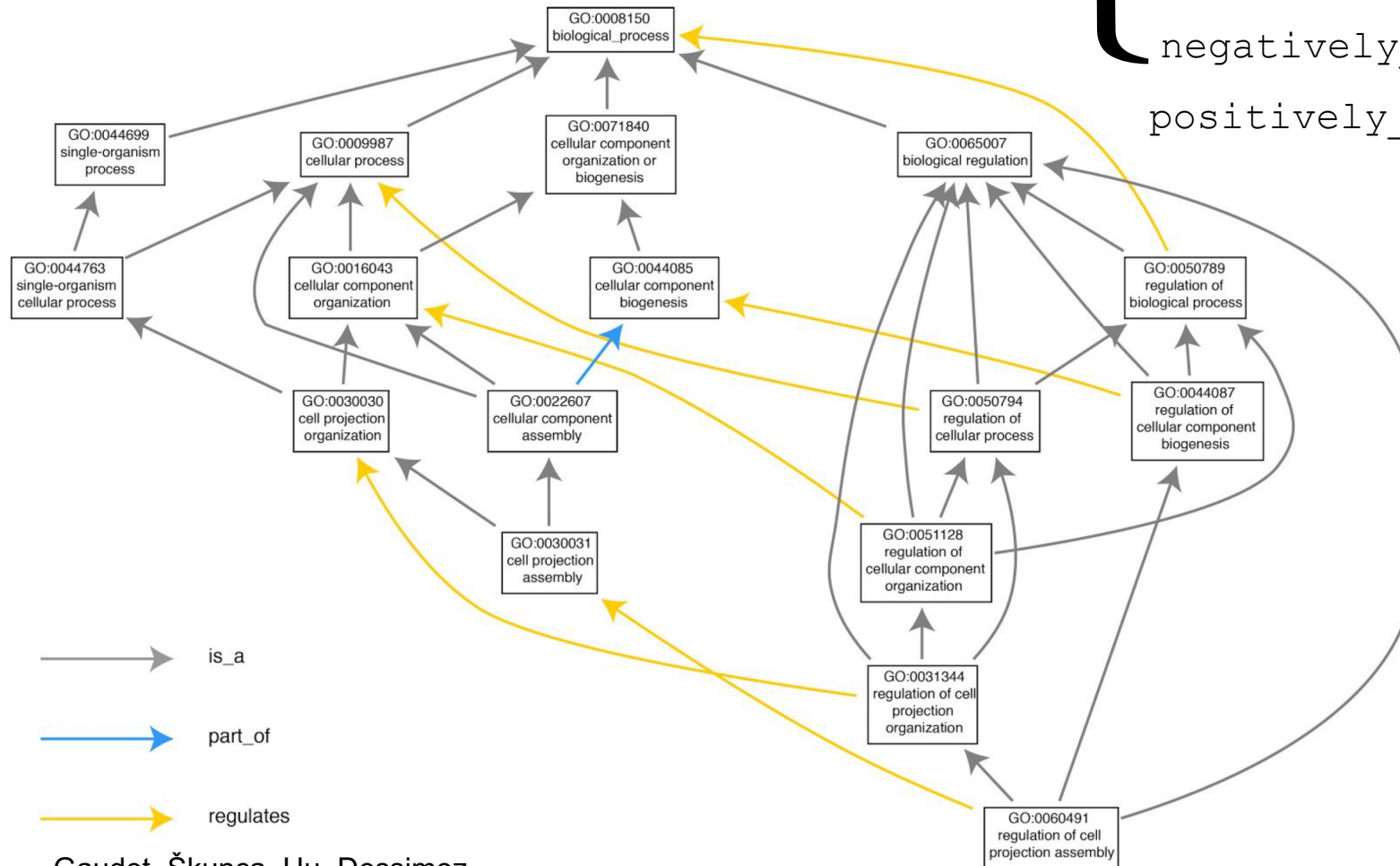
Similar to a simple tree, a DAG has directed edges and does not have cycles.

Depth of a node : length of the longest path from the root to that node.

Height of a node: length of the longest path from that node to a leaf.

# relationships in GO

Gene X {  
 is\_a  
 is a part\_of  
 regulates  
 negatively\_regulates  
 positively\_regulates  
 relationship



Gaudet, Škunca, Hu, Dessimoz  
 Primer on the Gene Ontology,  
<https://arxiv.org/abs/1602.01876>  
 V8

# Full GO vs. special subsets of GO

**GO slims** are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO.

They give a broad overview of the ontology content without the detail of the specific fine grained terms.

GO slims are created by users according to their needs, and may be specific to species or to particular areas of the ontologies.

**GO-fat** : GO subset constructed by DAVID @ NIH  
GO FAT filters out very broad GO terms

[www.geneontology.org](http://www.geneontology.org)

## Comparing GO terms

The hierarchical structure of the GO allows to compare proteins annotated to different terms in the ontology, as long as the terms have relationships to each other.

Terms located close together in the ontology graph (i.e., with a few intermediate terms between them) tend to be **semantically more similar** than those further apart.

One could simply count the **number of edges** between 2 nodes as a measure of their similarity.

However, this is problematic because not all regions of the GO have the same **term resolution**.



# Where do the Gene Ontology annotations come from?

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

\*October 2007 release

Rhee et al. Nature Reviews Genetics 9, 509-515 (2008)

## IEA: Inferred from Electronic Annotation

The evidence code IEA is used for all inferences made without human supervision, regardless of the method used.

The IEA evidence code is by far the most abundantly used evidence code.

Guiding idea behind computational function annotation:  
genes with similar sequences or structures are likely  
to be **evolutionarily related**.

Thus, assuming that they largely kept their ancestral function,  
they might still have **similar functional roles** today.

# Significance of GO annotations

Very **general GO terms** such as “cellular metabolic process” are annotated to many genes in the genome.

Very **specific terms** belong to a few genes only.

→ One needs to compare how **significant** the occurrence of a GO term is in a given set of genes compared to a randomly selected set of genes of the same size.

This is often done with the **hypergeometric test**.

## Hypergeometric test

$$\text{p-value} = \frac{\sum_{i=k_{\pi}}^{\min(n, K_{\pi})} \frac{\binom{K_{\pi}}{i} \binom{N-K_{\pi}}{n-i}}{\binom{N}{n}}}{1}$$

The hypergeometric test is a statistical test.

It can be used to check e.g. whether a biological annotation  $\pi$  is **statistically significant enriched** in a given test set of genes compared to the full genome.

- $N$  : number of genes in the genome
- $n$  : number of genes in the test set
- $K_{\pi}$  : number of genes in the genome with annotation  $\pi$ .
- $k_{\pi}$  : number of genes in test set with annotation  $\pi$ .

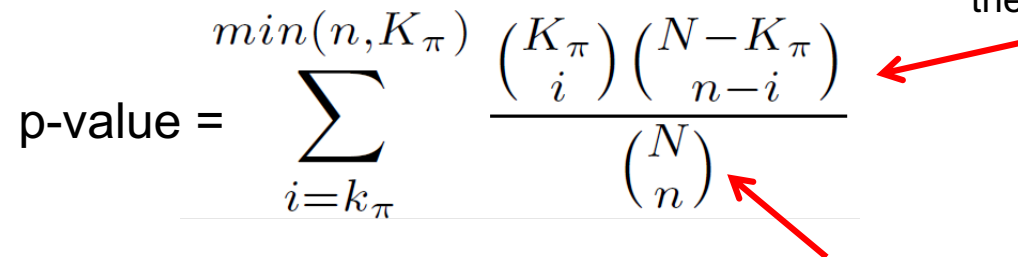
The hypergeometric test provides the **likelihood** that  $k_{\pi}$  or more genes that were **randomly selected** from the genome also have annotation  $\pi$ .

# Hypergeometric test

Select  $i \geq k_\pi$  genes with  
annotation  $\pi$  from the genome.

There are  $K_\pi$  such genes.

The other  $n - i$  genes in the test  
set do NOT have annotation  $\pi$ .  
There are  $N - K_\pi$  such genes in  
the genome.

$$\text{p-value} = \sum_{i=k_\pi}^{\min(n, K_\pi)} \frac{\binom{K_\pi}{i} \binom{N - K_\pi}{n - i}}{\binom{N}{n}}$$


The sum runs from  $k_\pi$   
elements to the maximal  
possible number of elements.

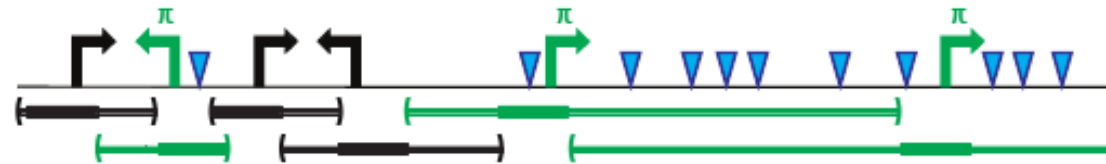
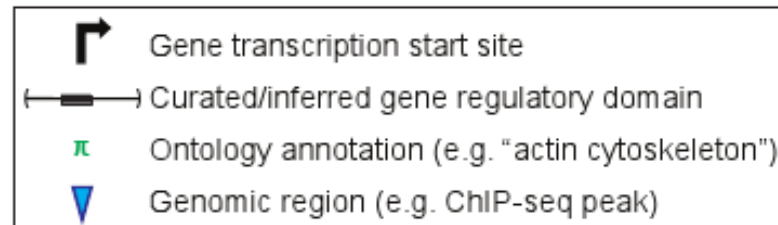
corrects for the number of  
possibilities for selecting  
 $n$  elements from a set of  
 $N$  elements.

This is either the number of  
genes with annotation  $\pi$  in the  
genome ( $K_\pi$ ) or the number of  
genes in the test set ( $n$ ).

This correction is applied if the  
sequence of drawing the  
elements is not important.

## Example

$$\text{p-Wert} = \sum_{i=k_{\pi}}^{\min(n, K_{\pi})} \frac{\binom{K_{\pi}}{i} \binom{N-K_{\pi}}{n-i}}{\binom{N}{n}}$$



Is annotation  $\pi$  significantly enriched in the test set of 3 genes?

Hypergeometric test over genes

$N$  = 6 total genes

$K_{\pi}$  = 3 genes annotated with  $\pi$

$n$  = 3 genes with an associated genomic region

$k_{\pi}$  = 3 genes annotated and with a genomic region

P-value = 0.05

Yes!  $p = 0.05$  is (just) significant.

# Multiple testing problem

In hypothesis-generating studies it is a priori not clear, which terms should be tested.

Therefore, one typically performs not only one hypothesis with a single term but many tests with many, often all terms that the Gene Ontology provides and to which at least one gene is annotated.

Result of the analysis: a list of terms that were found to be significant.

Given the large number of tests performed, this list will contain a large number of **false-positive** terms.

Sebastian Bauer, Gene Category Analysis  
Methods in Molecular Biology 1446, 175-188  
(2017)

## Multiple testing problem

For example, if one statistical test is performed at the 5% level and the corresponding null hypothesis is true, there is only a 5% chance of incorrectly rejecting the null hypothesis  
→ one expects 0.05 incorrect rejections.

However, if 100 tests are conducted and all corresponding null hypotheses are true, the expected number of incorrect rejections (also known as false positives) is 5.

If the tests are statistically independent from each other, the probability of at least one incorrect rejection is 99.4%.

[www.wikipedia.org](http://www.wikipedia.org)



# Bonferroni correction

Therefore, the result of a term enrichment analysis must be subjected to a **multiple testing correction**.

The most simple one is the **Bonferroni** correction. Here, each  $p$ -value is simply multiplied by the number of tests saturated at a value of 1.0.

Bonferroni controls the so-called **family-wise error rate**, which is the probability of making one or more false discoveries.

It is a very conservative approach because it handles all  $p$ -values as independent.

Note that this is not a typical case of gene-category analysis.

So this approach often goes along with a reduced statistical power.

Sebastian Bauer, Gene Category Analysis  
Methods in Molecular Biology 1446, 175-188  
(2017)

## Benjamini Hochberg: expected false discovery rate

The Benjamini–Hochberg approach controls the **expected false discovery rate** (FDR), which is the **proportion** of false discoveries among all rejected null hypotheses.

This has a positive effect on the statistical power at the expense of having less strict control over false discoveries.

Controlling the FDR is considered by the American Physiological Society as “the best practical solution to the problem of multiple comparisons”.

Note that less conservative corrections usually yield a higher amount of significant terms, which may be not desirable after all.

Sebastian Bauer, Gene Category Analysis  
Methods in Molecular Biology 1446, 175-188  
(2017)

# Information content of GO terms

The **likelihood** of a node  $t$  can be defined in 2 ways:

How many genes have annotation  $t$   
relative to the root node?

$$p_{anno}(t) = \frac{occur(t)}{occur(root)}$$

Number of GO terms in subtree below  $t$   
relative to number of GO terms in tree

$$p_{graph}(t) = \frac{D(t)}{D(root)}$$

The likelihood takes values between 0 and 1 and  
increases monotonic from the leaf nodes to the root.

Define **information content** of a node from its likelihood:

$$IC(t) = -\log p(t)$$

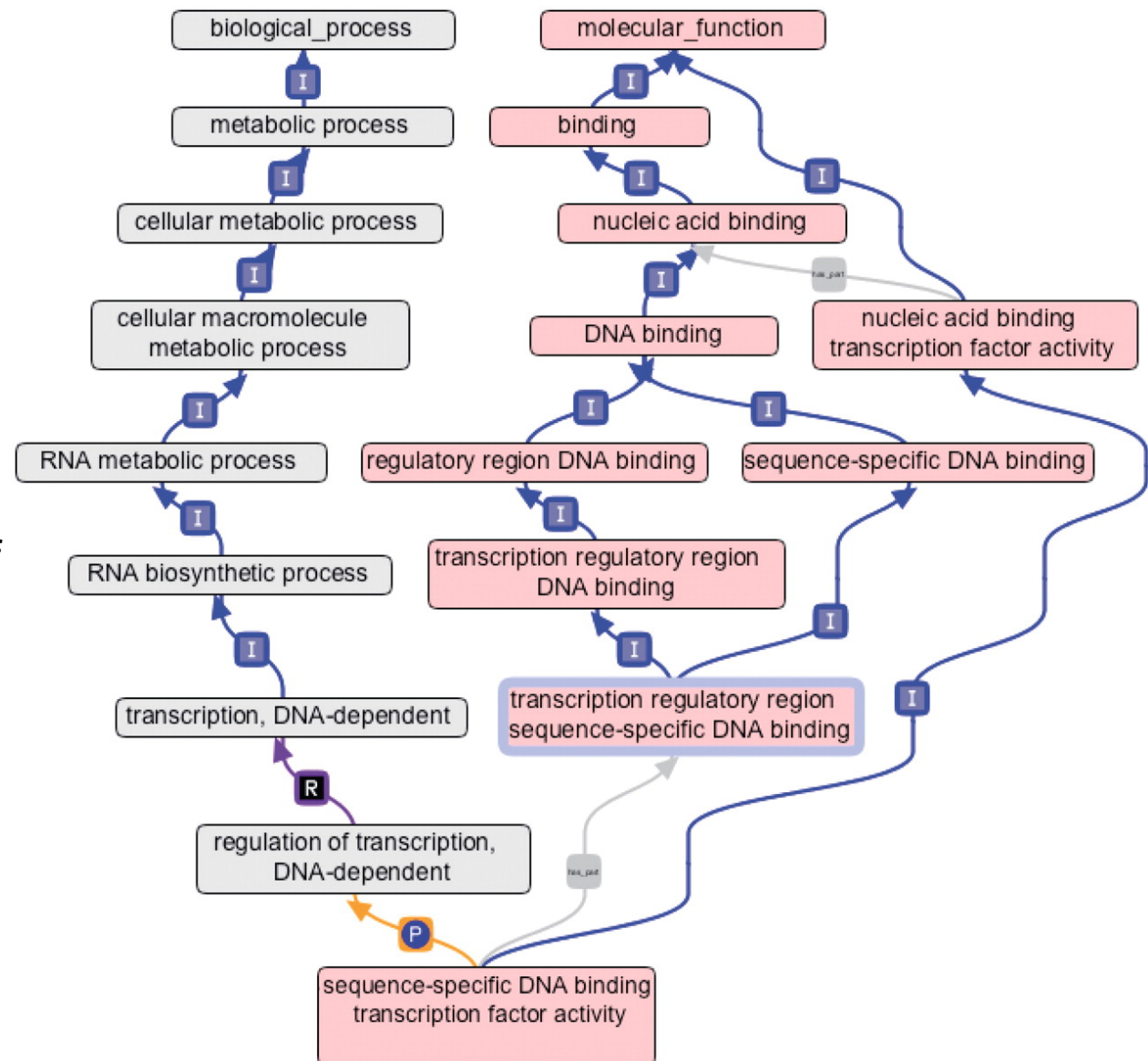
A rare node has high information content.

# Common ancestors of GO terms

Common ancestors of two nodes  $t_1$  and  $t_2$  :  
all nodes that are located on a path from  $t_1$  to root AND on a path from  $t_2$  to root.

The **most informative common ancestor** (MICA) of terms  $t_1$  and  $t_2$  is their common ancestor with highest information content.

Typically, this is the closest common ancestor.



*Nucl. Acids Res.* (2012) 40 (D1):  
D559-D564

## Measure functional similarity of GO terms

Lin *et al.* defined the **similarity** of two GO terms  $t_1$  and  $t_2$  based on the information content of the most informative common ancestor (MICA)

$$sim_{Rel}(t_1, t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1) + IC(t_2)}$$

MICAs that are close to their GO terms receive a higher score than those that are higher up in the GO graph

Schlicker *et al.* defined the following variant:

$$sim_{Rel}(t_1, t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1) + IC(t_2)} \cdot (1 - p(MICA))$$

where the term similarity is weighted with the counter-probability of the MICA. By this, shallow annotations receive less relevance than MICAs further away from the root.

## Measure functional similarity of two genes

Two genes or two sets of genes  $A$  and  $B$  typically have more than 1 GO annotation each. → Consider similarity of all terms  $i$  and  $j$ :

$$s_{ij} = \text{sim}(GO_i^A, GO_j^B), \forall i \in 1, \dots, N, \forall j \in 1, \dots, M.$$

and select the maxima in all rows and columns:

$$\text{rowScore}(A, B) = \frac{1}{N} \sum_{i=1}^N \max_{1 \leq j \leq M} s_{ij}, \quad \text{GOscore}_{\text{avg}}^{\text{BMA}}(A, B) = \frac{1}{2} \cdot (\text{rowScore}(A, B) + \text{columnScore}(A, B))$$

$$\text{columnScore}(A, B) = \frac{1}{M} \sum_{j=1}^M \max_{1 \leq i \leq N} s_{ij}. \quad \text{GOscore}_{\text{max}}^{\text{BMA}}(A, B) = \max(\text{rowScore}(A, B), \text{columnScore}(A, B))$$

Compute *funsim*-Score from scores for BP tree and MF tree:

$$\text{funsim}(A, B) = \frac{1}{2} \cdot \left[ \left( \frac{\text{BPscore}}{\max(\text{BPscore})} \right)^2 + \left( \frac{\text{MFscore}}{\max(\text{MFscore})} \right)^2 \right]$$

# IEA: Inferred from Electronic Annotation

The evidence code IEA is used for all inferences made without human supervision, regardless of the method used.

IEA evidence code is by far the most abundantly used evidence code.

The guiding idea behind computational function annotation is the notion that genes with similar sequences or structures are likely to be evolutionarily related. Thus, assuming they largely kept their ancestral function, they might still have similar functional roles today.

Gaudet, Škunca, Hu, Dessimoz

Primer on the Gene Ontology,

<https://arxiv.org/abs/1602.01876>.

Published in : Methods in Molecular Biology

Vol1446 (2017) – **open access!**

V8

Processing of Biological Data

## Heterogeneous nature of GO may introduce biases

GO data is **heterogeneous** in many respects — to a large extent simply because the body of knowledge underlying the GO is itself very heterogeneous.

This can introduce considerable **biases** when the data is used in other analysis, an effect that is magnified in large-scale comparisons.

Statisticians and epidemiologists make a clear distinction between

- *experimental data* — data from a controlled experiment, designed such that the case and control groups are as identical as possible in all respects other than a factor of interest — and
- *observational data* — data readily available, but with the potential presence of unknown or unmeasured factors that may **confound** the analysis.

GO annotations clearly falls into the second category.

Gaudet, Dessimoz,

Gene Ontology: Pitfalls, Biases, Remedies

<https://arxiv.org/abs/1602.01876>

V8

Processing of Biological Data



# Simpson's paradox: perils of data aggregation

Simpson's paradox is the counterintuitive observation that a statistical analysis of **aggregated data** (combining multiple individual datasets) can lead to **dramatically different conclusions** than if datasets are analyzed **individually**.

I.e. the whole appears to disagree with the parts.

Classic example from University of California at Berkeley:  
UC Berkeley was sued for **gender bias** against female applicants because in 1973, in total 44% of the male applicants were admitted to Berkeley but only 35% of the female applicants — an observational dataset.

Gaudet, Dessimoz,

Gene Ontology: Pitfalls, Biases, Remedies

<https://arxiv.org/abs/1602.01876>

V8

Processing of Biological Data

## Simpson's paradox: perils of data aggregation

However, when individually looking at the men vs. women admission rate for each department, the rate was in fact similar for both sexes (and even in favor of women in most departments).

→ The lower overall acceptance rate for women was **not** due to **gender bias**, but to the tendency of women to apply to more **competitive departments**, which have a lower admission rate in general.

The association between gender and admission rate in the aggregate data could almost entirely be explained through strong association of these two variables with a third, **confounding variable**, the department.

When controlling for the confounder, the association between the two first variables dramatically changes.

This type of phenomenon is referred to as **Simpson's paradox**.

Gaudet, Dessimoz,

Gene Ontology: Pitfalls, Biases, Remedies

<https://arxiv.org/abs/1602.01876>

26

# GO is inherently incomplete

The Gene Ontology is a representation of the **current state of knowledge**; thus, it is very **dynamic**.

The ontology itself is constantly being improved to more accurately represent biology across all organisms.

The ontology is augmented as new discoveries are made.

At the same time, the **creation of new annotations** occurs at a rapid pace, aiming to keep up with published work.

Despite these efforts, the information contained in the GO database is necessarily **incomplete**.

**Thus, absence of evidence of function does not imply absence of function.**

This is referred to as the **Open World Assumption**

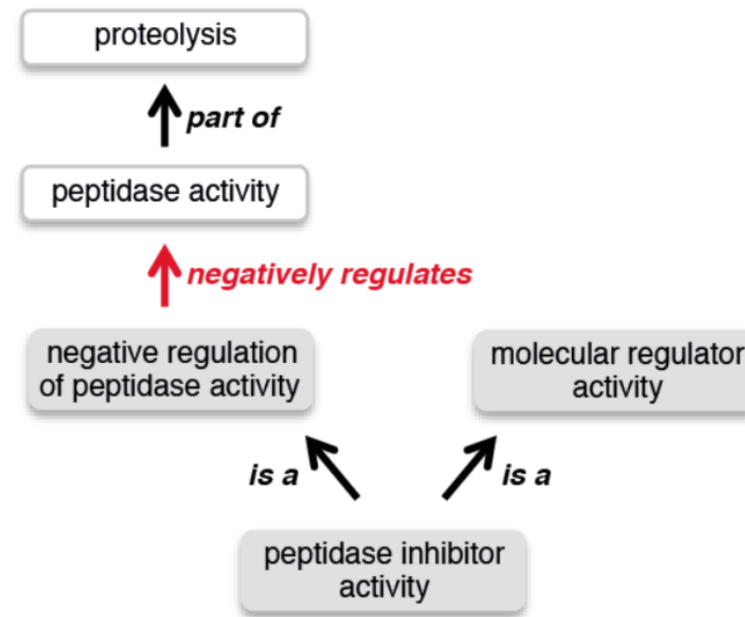
# Transitive vs. non-transitive relationships

Some relationships, such as “is a” and “part of”, are **transitive**.

→ any protein annotated to a specific term is also implicitly annotated to all of its parents

On the other hand, relations such as “regulates” are **non-transitive**.

→ the semantics of the association of a gene to a GO term is not the same for its parent: if A is *part of* B, and B *regulates* C, we cannot make any inferences about the relationship between C and A.



Example of transitive (black arrows) and non-transitive (red arrow) relationships between classes.

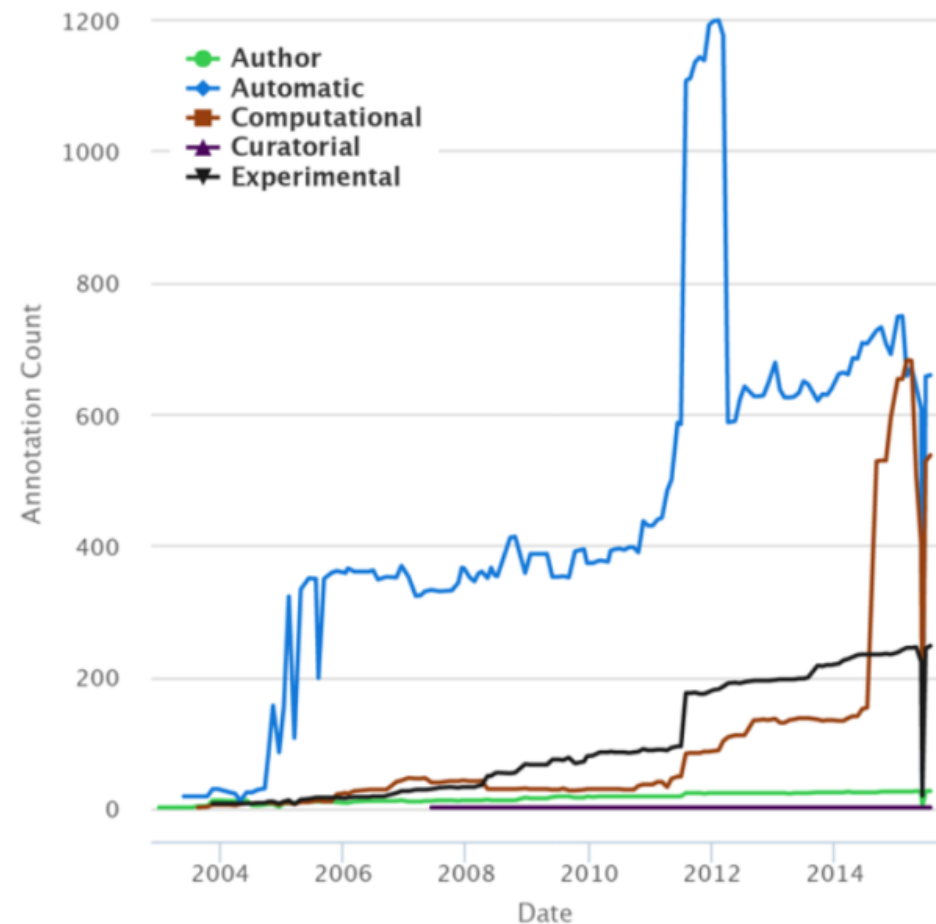
A protein annotated to “peptidase inhibitor activity” term does not imply it has a role in “proteolysis”, since the link is broken by the non-transitive relation *negatively regulates*.

# GO annotations are dynamic in time

Example: strong and sudden variation in the number of annotations with the GO term "ATPase activity" over time.

Such changes can heavily affect the estimation of the **background distribution** in enrichment analyses.

To minimise this problem, one should use an **up-to-date version** of the ontology/annotations and ensure that conclusions drawn hold across recent (earlier) releases.



Gaudet, Dessimoz,

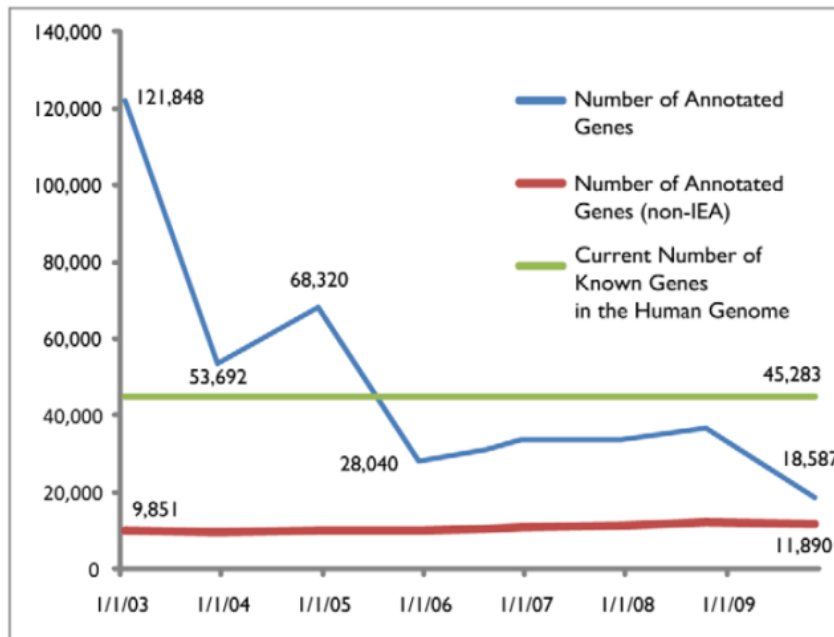
Gene Ontology: Pitfalls, Biases, Remedies

<https://arxiv.org/abs/1602.01876>

V8

Processing of Biological Data

## Number of GO-annotated human genes



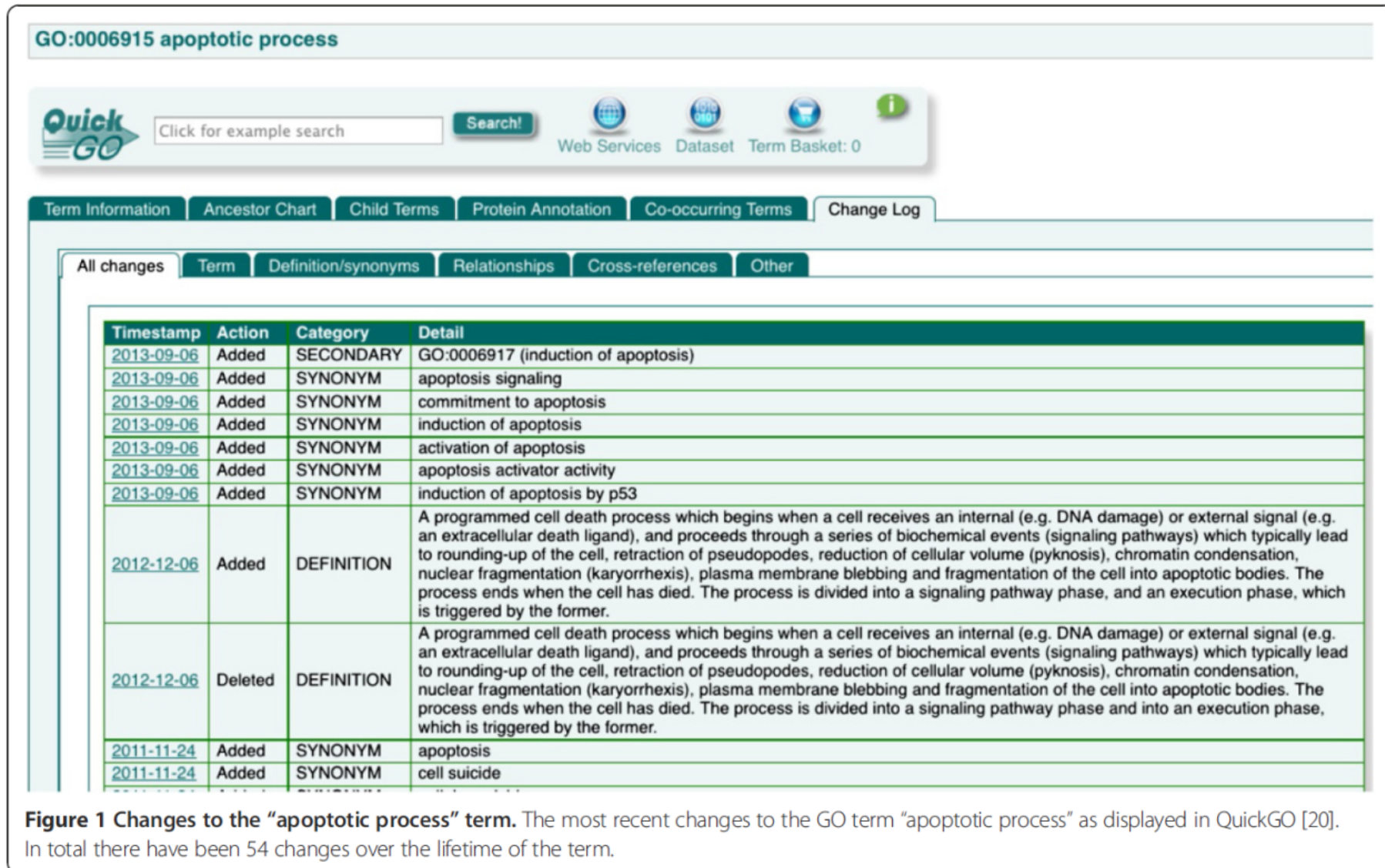
Between 01/2003 and 12/2003 the estimated number of known genes in the human genome was adjusted.

Between 12/2004 and 12/2005, and between 10/2008 and 11/2009 annotation practices were modified.

One can argue that, although the **number of annotated genes** decreased, the **quality of annotations** improved, see the steady increase in the number of **genes with non-IEA annotations**.

However, this increase in the number of genes with non-IEA annotations is very slow. Between 11/2003 and 11/2009, only 2,039 new genes received non-IEA annotations. At the same time, the number of non-IEA annotations increased from 35,925 to 65,741, indicating a strong **research bias** for a small number of genes.

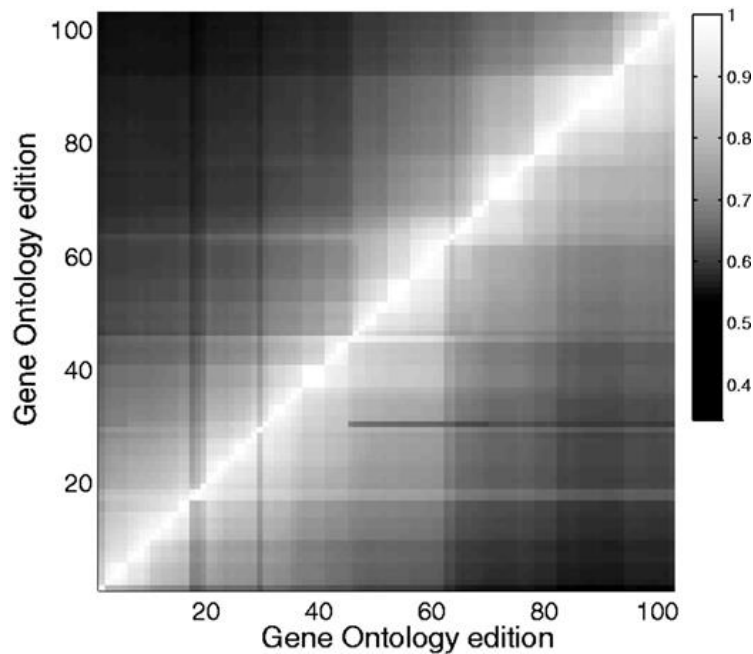
# Changes to GO terms are recorded



Huntley et al. GigaScience 2014, 3:4



# Gene functional identity changes over GO editions



Shading : fraction of genes that retain a functional identity between GO editions.

Semantic similarity is calculated and genes are matched between GO editions.

If a gene is most similar to itself between editions, it is said to **retain** its **identity**.

The average fraction of identity maintained in successive editions of GO is 0.971.

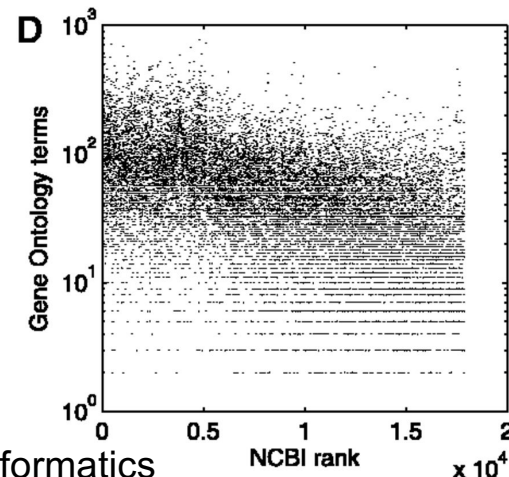
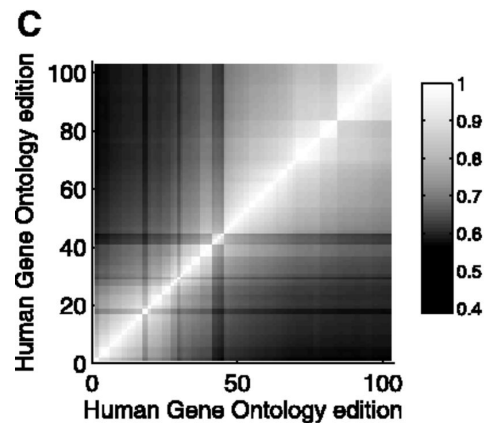
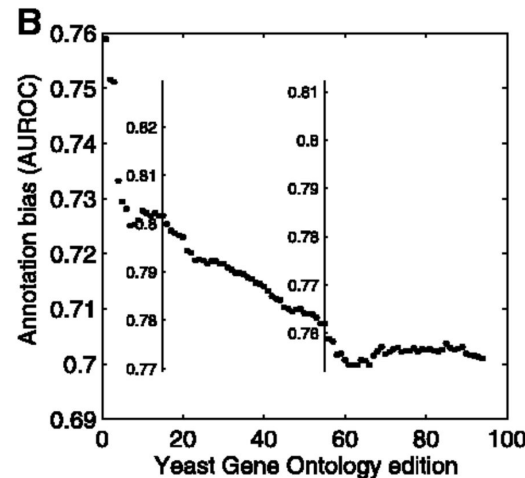
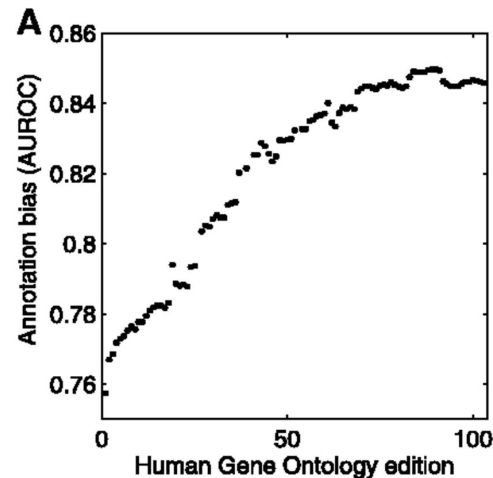
This means that, each month, the annotations of about 3% of the genes have changed so substantially that they are not functionally ‘the same genes’ anymore.

Gillis, Pavlidis, Bioinformatics  
(2013) 29: 476-482.



# Annotation bias persists in the GO

If all genes have the same number of GO terms, the **annotation bias** is 0.5. At the other extreme, if there are only a few GO terms used and they are all applied to the same set of genes, then the bias is 1.0.



(**A**) Annotation bias has risen among human genes. Genes with many annotations have become more dominant within GO over time.

(**B**) For yeast, annotation bias has generally fallen over time.

(**C**) The relative number of annotations per gene has remained fairly stable over time (shown is the correlation of the distributions).

(**D**) Number of GO terms per gene is correlated with the rank of the numerical ID of the gene in NCBI.  
→ early sequenced genes are better annotated = historical bias.

Gillis, Pavlidis, Bioinformatics  
(2013) 29: 476-482.

## Case study: network “modules” of PSP

Common use of GO: analysis of network ‘modules’ enriched for particular functions.

**Case study:** ‘post-synaptic proteome’ (PSP) = 620 proteins identified by MS in a structural component of the synapse that can be observed under the electron microscope beneath the postsynaptic membrane

Enrichment analysis on this list → 67 significantly enriched functions.

Based on a protein interaction data set for human (→ HIPPIE, 73324 interactions), Gillis and Pavlidis constructed a PSP subnetwork.

Using spectral partitioning, this was split into 6 subnetworks (modules) varying in size from 11 to 67 genes.

Gillis, Pavlidis, Bioinformatics  
(2013) 29: 476-482.

## Case study: network “modules” of PSP

4 modules had significantly enriched groups of GO terms, suggesting the PPI modules partly reflect different functions:

Cluster 1: glutamatergic activity and synaptic transmission

Cluster 3: cell junctions and adhesion

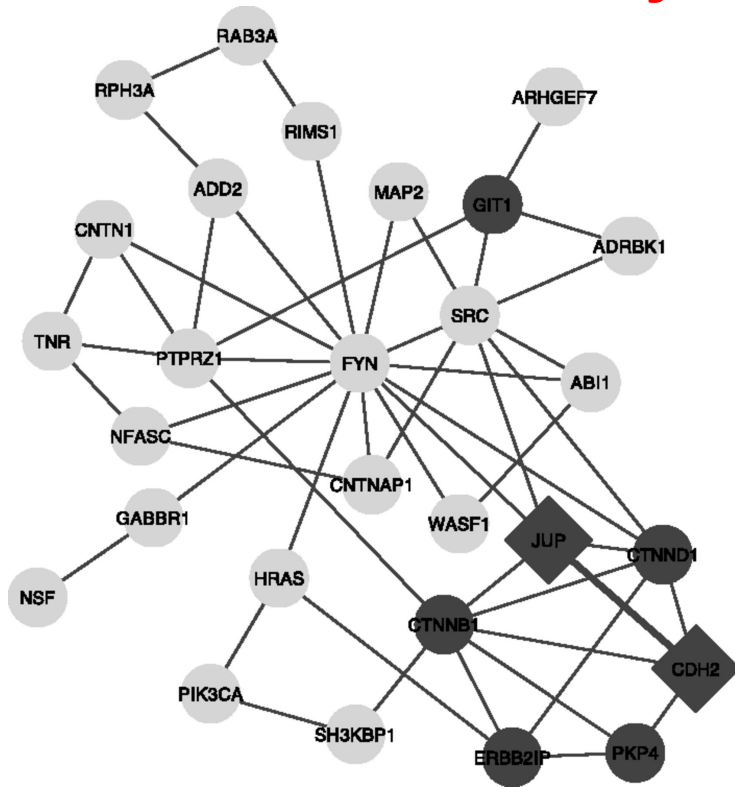
Cluster 4: ribosomal components

Cluster 6: endocytosis.

However, the authors speculated that separation of functions by PPI clustering does not indicate an orthogonal property, but simply be due that different articles reported both certain P-P interactions and certain protein functions.

Gillis, Pavlidis, Bioinformatics  
(2013) 29: 476-482.

## Case study: network “modules” of PSP



Module 3 from the PSP case study.  
Genes annotated with the enriched  
functions shown in dark gray.

Indeed, the 2 interacting proteins JUP and  
CDH3 in cluster 3 (diamonds) were  
confounded.

2 articles reported both their functional  
annotation and their interaction.

Removing the GO terms traced back to  
these 2 articles from the 2 genes  
reduced the functional enrichment for the  
module to the point that no functions met  
the  $FDR < 0.01$  threshold.

Gillis, Pavlidis, Bioinformatics  
(2013) 29: 476-482.

## Influence of electronic annotations (IEA)

High-throughput experiments are another source for **annotation bias**.

They contribute disproportionally large amounts of annotations by only few published studies.

This information is further propagated by automated methods.

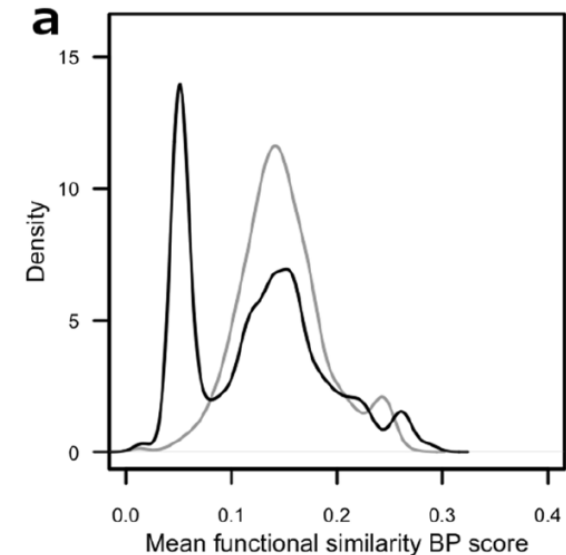
The huge body of electronic annotations (evidence code IEA) has therefore a strong influence on semantic similarity scores.

## Influence of electronic annotations (IEA): BP scores

Average *simLin/fsAvg* score distributions for BP ontology for human/mouse protein pairs.

For a human protein  $P$ , the score average is computed by forming pairs of proteins  $(P, R)$  over 1000 randomly selected mouse proteins  $R$  for

- the IEA(+) dataset (**black solid lines**, density computed from 93806 annotated proteins) and
- the IEA(-) dataset (**grey lines**, 21212 annotated proteins).



No random pair has  $SS > 0.4 \rightarrow$  good threshold to distinguish random / non-random

Manually annotated protein pairs (**grey**) show a clear peak at a score of 0.15.

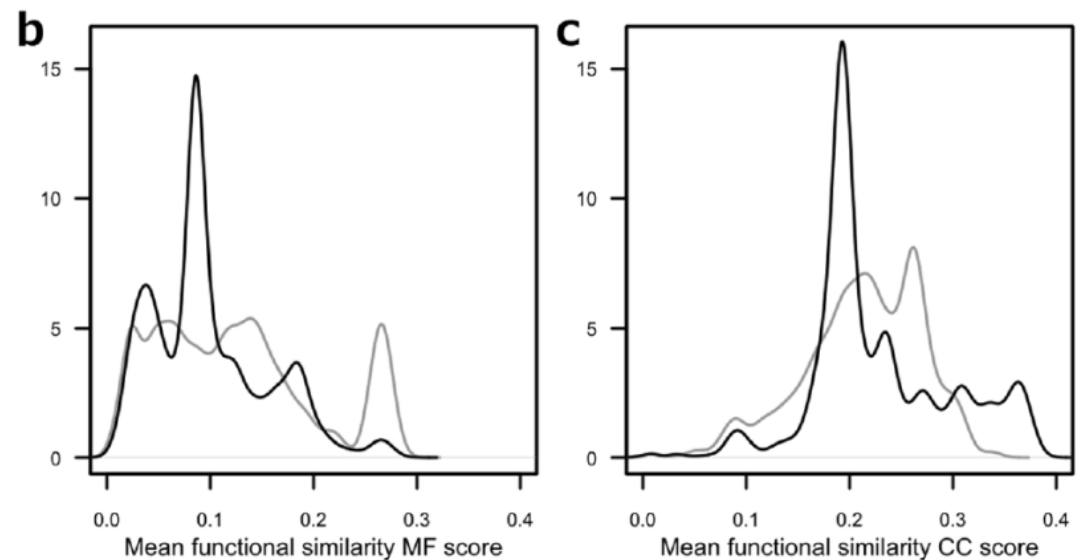
Including IEA evidence generates a second peak close to 0.0. A large portion of this peak can be attributed to the roughly 70000 human gene products, which are exclusively annotated with IEA evidence codes

Weichenberger et al. (2017)

## Influence of electronic annotations on MF + CC scores

**(b)** MF based score distribution. Unlike BP, this ontology is characterized by a more uniform distribution of scores, with a notable peak near 0.27, generated by ca. 1600 proteins.

GO enrichment analysis of these proteins shows that they are significantly enriched in “protein binding” (GO:0005155,  $p < 10^{-100}$ ), suggesting that gene products annotated to this term generally yield much higher than average *simLin/fsAvg* MF scores.



**(c)** CC score distribution. Here, both manual and electronic annotation peaks are closer to each other than in the other 2 ontologies. Electronic annotations have higher densities in the upper score range (>0.3), where the manual annotation scores have already tailed off.

Weichenberger et al. (2017)

Scientific Reports 7: 381

V8

Processing of Biological Data

## Using similarity z-scores

Another bias:

genes with a higher number of GO annotations tend to receive higher functional similarity scores.

Weichenberger et al. propose to improve the similarity scores of 2 proteins by taking into account their respective **score background distribution** and calculate a similarity z-score that is less affected by annotation biases of specific proteins.

Mean and standard deviation for each protein  $P$  are computed by evaluating functional similarity scores from protein pairs  $(P, Q)$ , where proteins  $Q$  are randomly sampled.

This mean score for protein  $P$  represents a baseline score that varies from protein to protein.

Together with a protein-specific standard deviation a normalized z-score is derived that adjusts for the annotation baseline of the particular proteins.

Weichenberger et al. (2017)

Scientific Reports 7: 381

V8

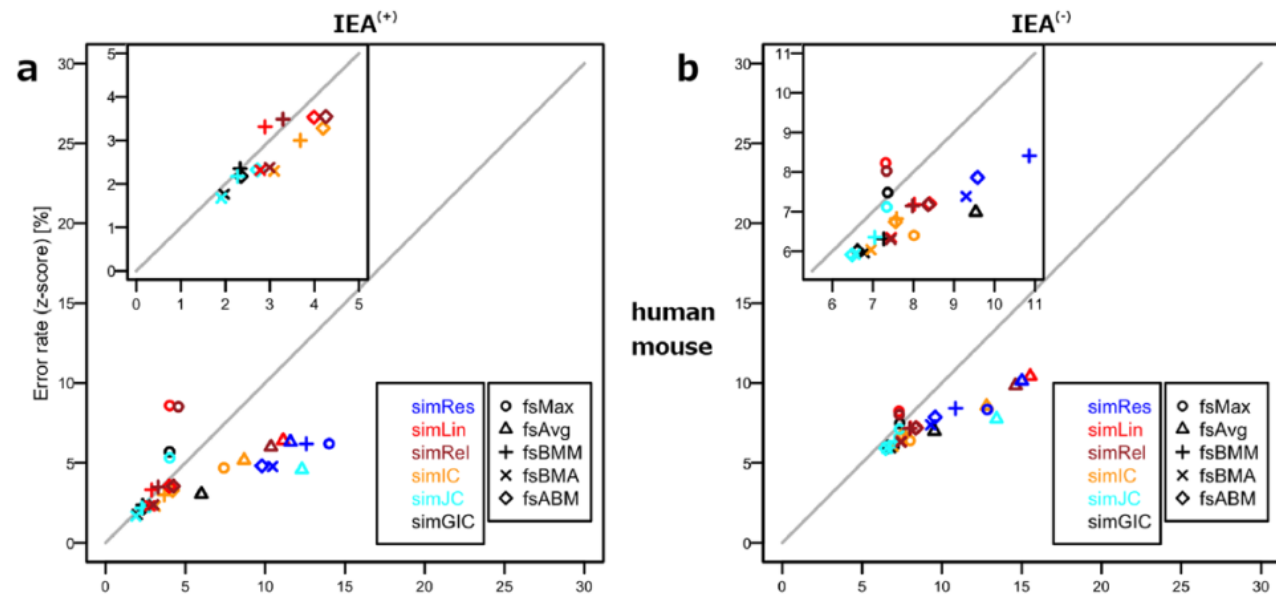
Processing of Biological Data



# Similarity z-scores detect orthologues better

error rates of raw functional similarity scores (x-axis) versus z-scores (y-axis) for BP ontology for pairs of orthologues and controls from selected organisms.

Small inlay panel: best scoring measures.



Points on the diagonal have the same error rate with raw and with z-scores.

Observed deviations from diagonal → lower error rate using z-scores.

(left) results from an annotation corpus including electronic annotations (IEA+),

(right) outcome where electronic annotations have been excluded (IEA-).

Weichenberger et al. (2017)

Scientific Reports 7: 381

V8

Processing of Biological Data

# Compare methods to measure functional similarity

$s$  and  $t$  : two GO terms that will be compared semantically

$S(s, t)$  : set of all common ancestors of  $s$  and  $t$ .

Resnik (*simRes*)

$$simRes(s, t) = \max_{c \in S(s, t)} I(c)$$

Lin (*simLin*)

$$simLin(s, t) = \max_{c \in S(s, t)} \frac{2 \cdot I(c)}{I(s) + I(t)}$$

Schlicker (*simRel*)

$$simRel(s, t) = \max_{c \in S(s, t)} \left( \frac{2 \cdot I(c)}{I(s) + I(t)} \cdot (1 - P(c)) \right)$$

information coefficient (*simIC*)

$$simIC(s, t) = \frac{2 \cdot \max_{c \in S(s, t)} I(c)}{I(s) + I(t)} \cdot \left( 1 - \frac{1}{1 - \max_{c \in S(s, t)} I(c)} \right)$$

Jiang and Conrath (*simJC*),

$$simJC(s, t) = \frac{1}{1 + I(s) + I(t) - 2 \cdot \max_{c \in S(s, t)} I(c)}$$

graph information content (*simGIC*).

$$simGIC(s, t) = \frac{\sum_{c \in \{S(s, s) \cap S(t, t)\}} I(c)}{\sum_{c \in \{S(s, s) \cup S(t, t)\}} I(c)}$$

Weichenberger et al. (2017)

## Mixing rules

Given:

protein  $P$  that is annotated with  $m$  GO terms  $t_1, t_2, \dots, t_m$  and  
protein  $R$  that is annotated with  $n$  GO terms  $r_1, r_2, \dots, r_n$ .

Then the matrix  $M$  is given by all possible pairwise SS values

$s_{ij} = \text{sim}(t_i, r_j)$  with  $\text{sim}$  being one of the SS measures introduced above,  
 $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ .

Functional similarity is computed from the SS entries of  $M$  according to a specific **mixing strategy** (MS). Here 5 different mixing strategies were investigated.

$fsMax$  uses the **maximum** value of the matrix,  $fsMax = \max_{i,j} s_{ij}$ ,

$fsAvg$  takes the **average** over all entries,  $fsAvg = \frac{1}{m \times n} \sum_{i,j} s_{ij}$ .

## Mixing rules

Using the maximum of averaged row and column best matches has been suggested for incomplete annotations,

$$fsBMM = \max\left(\frac{1}{m}\sum_i \max_j s_{ij}, \frac{1}{n}\sum_j \max_i s_{ij}\right)$$

Instead of taking the maximum, averaging gives the so-called **best match average**

$$fsBMA = \frac{1}{2}\left(\frac{1}{m}\sum_i \max_j s_{ij} + \frac{1}{n}\sum_j \max_i s_{ij}\right)$$

Conversely, the **averaged best match** is defined as

$$fsABM = \frac{1}{m+n}\left(\sum_i \max_j s_{ij} + \sum_j \max_i s_{ij}\right)$$

We additionally study the effect of combining multiple gene ontologies into a single score, as suggested by Schlicker et al..

A functional similarity  $F$  is computed by combining a SS measure with any mixing strategy defined above over any of the different ontologies: biological process ( $FBP$ ), molecular function ( $FMF$ ), and cellular component ( $FCC$ ). We compute the combined measures as the functions

$$F_{BP+MF} = \sqrt{\frac{1}{2}(F_{BP}^2 + F_{MF}^2)}$$

$$F_{BP+MF+CC} = \sqrt{\frac{1}{3}(F_{BP}^2 + F_{MF}^2 + F_{CC}^2)}$$

Weichenberger et al. (2017)

## Optimal functional similarity score

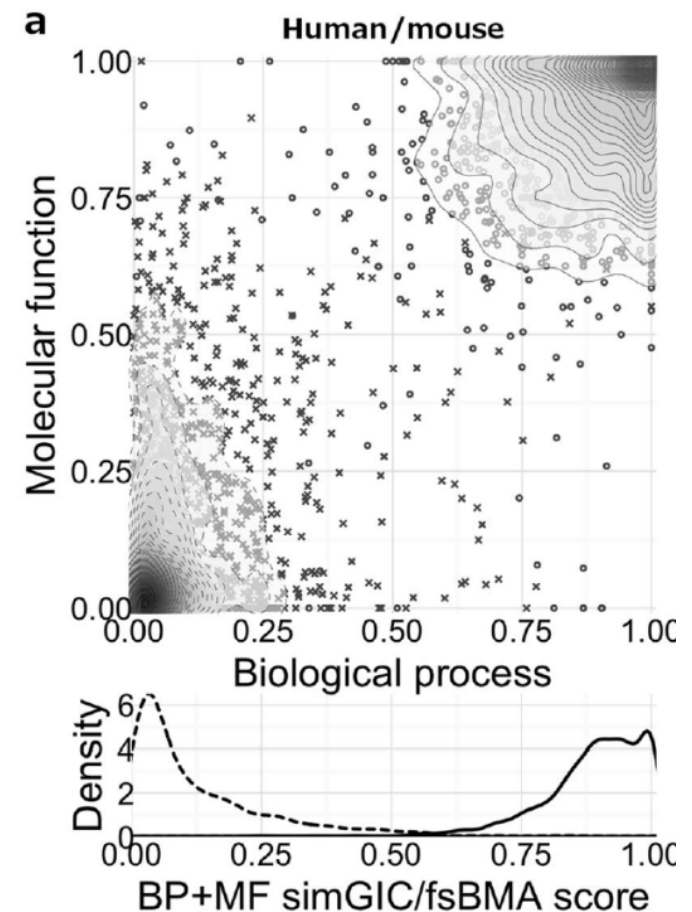
Top: scatter plot of BP (x-axis) and MF (y-axis) scores of orthologous gene pairs (circles) and randomly selected gene pairs (crosses) from human/mouse.

Solid/dashed iso-lines: 2D density function of the 2 distributions for cases and controls.

Bottom: 1D density function of the  $F^{BP+MF}$  scores for cases (solid line) and controls (dashed line).

Their crossing point defines the optimal threshold for minimizing the error rate.

The *simGIC* semantic similarity in conjunction with the  $F^{BP+MF}$  function separates cases from controls, with an error rate of only 1.24%.



Weichenberger et al. (2017)

Scientific Reports 7: 381

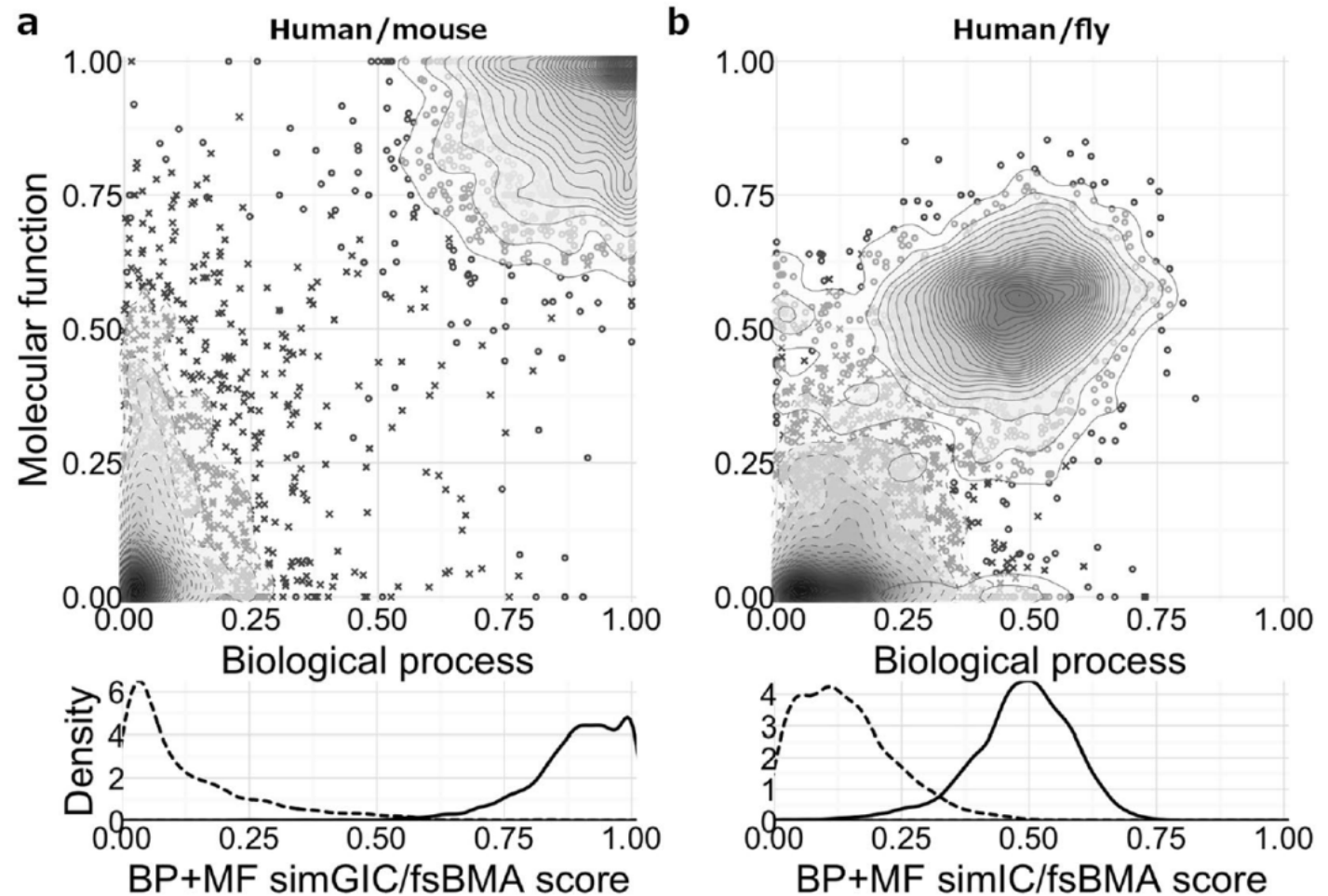
V8

Processing of Biological Data

## Optimal functional similarity score

(b) Human/fly orthologues and controls with their associated *simIC/fsBMA* scores.

On average, the error rate is slightly higher than for human / mouse: 5.55%.



Weichenberger et al. (2017)

Scientific Reports 7: 381

V8

Processing of Biological Data

## Summary

- The GO is the **gold-standard** for **computational annotation of gene function**.
- It is continuously updated and refined.
- **Hypergeometric test** is most often used to compute **enrichment** of GO terms in gene sets
- Semantic similarity concepts allow measuring the functional similarity of genes. Selecting an optimal definition for semantic similarity of 2 GO terms and for the mixing rule depends on what works best in practice.
- Functional gene annotation based on GO is affected by a number of **biases**.