

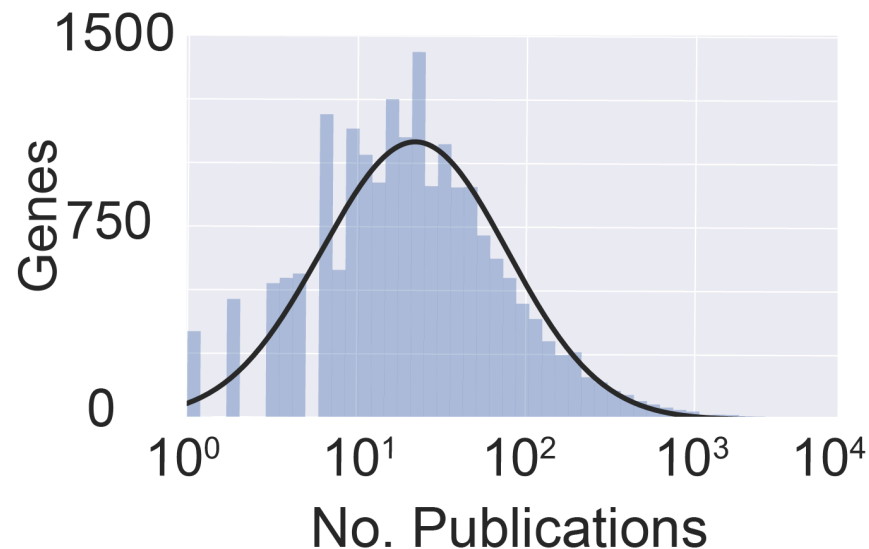
V9 – Functional annotation

Program for today:

- Have all genes been studied with the same **intensity**?
- **Functional annotation** of genes/gene products: Gene Ontology (GO)
- **significance** of annotations: hypergeometric test
- (mathematical) **semantic similarity** of GO-terms

High imbalance in intensity of research on individual genes

Frequency of the number of research publications associated with individual human protein-coding genes in MEDLINE.



The observed disparity could in principle reflect a lack of importance of many genes.

More likely it reflects

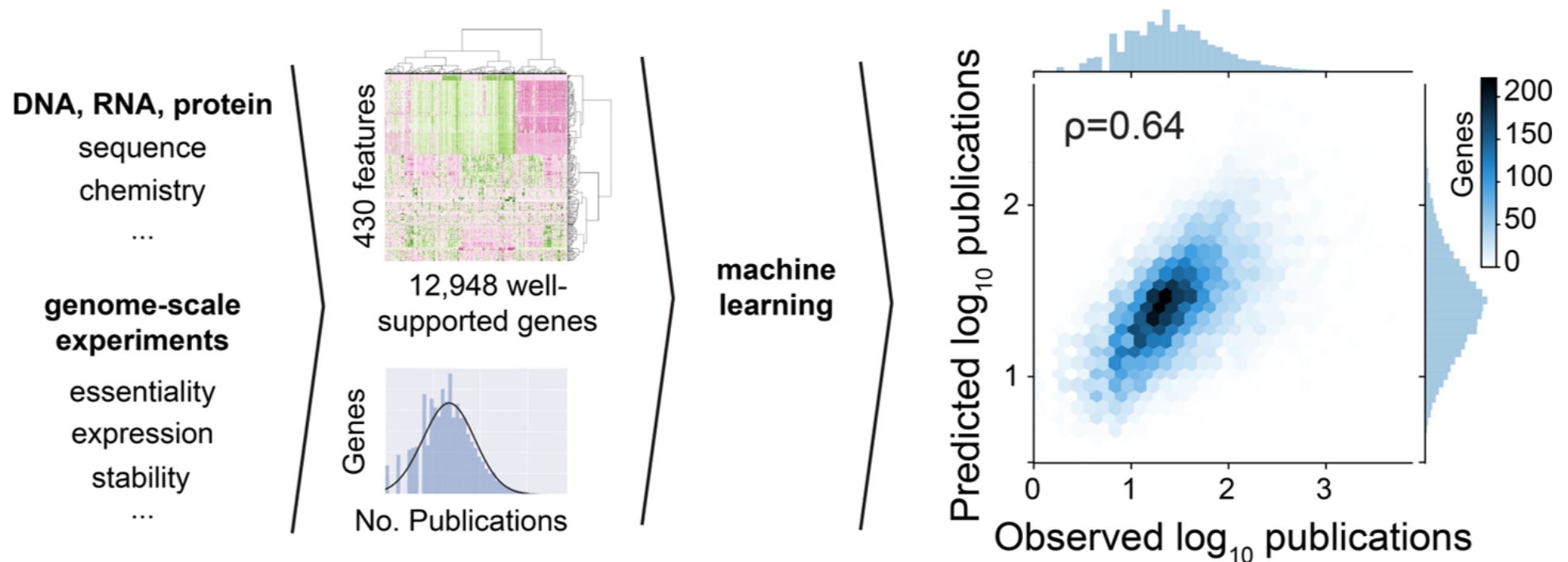
- existing social structures of research,
- scientific and economic reward systems,
- medical and societal relevance,
- preceding discoveries,
- the availability of technologies and reagents, etc.

Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

What determines the number of publications per gene?

Using information on 430 physical, chemical, and biological features of genes, one can predict the number of publications for single genes with 0.64 Spearman correlation.



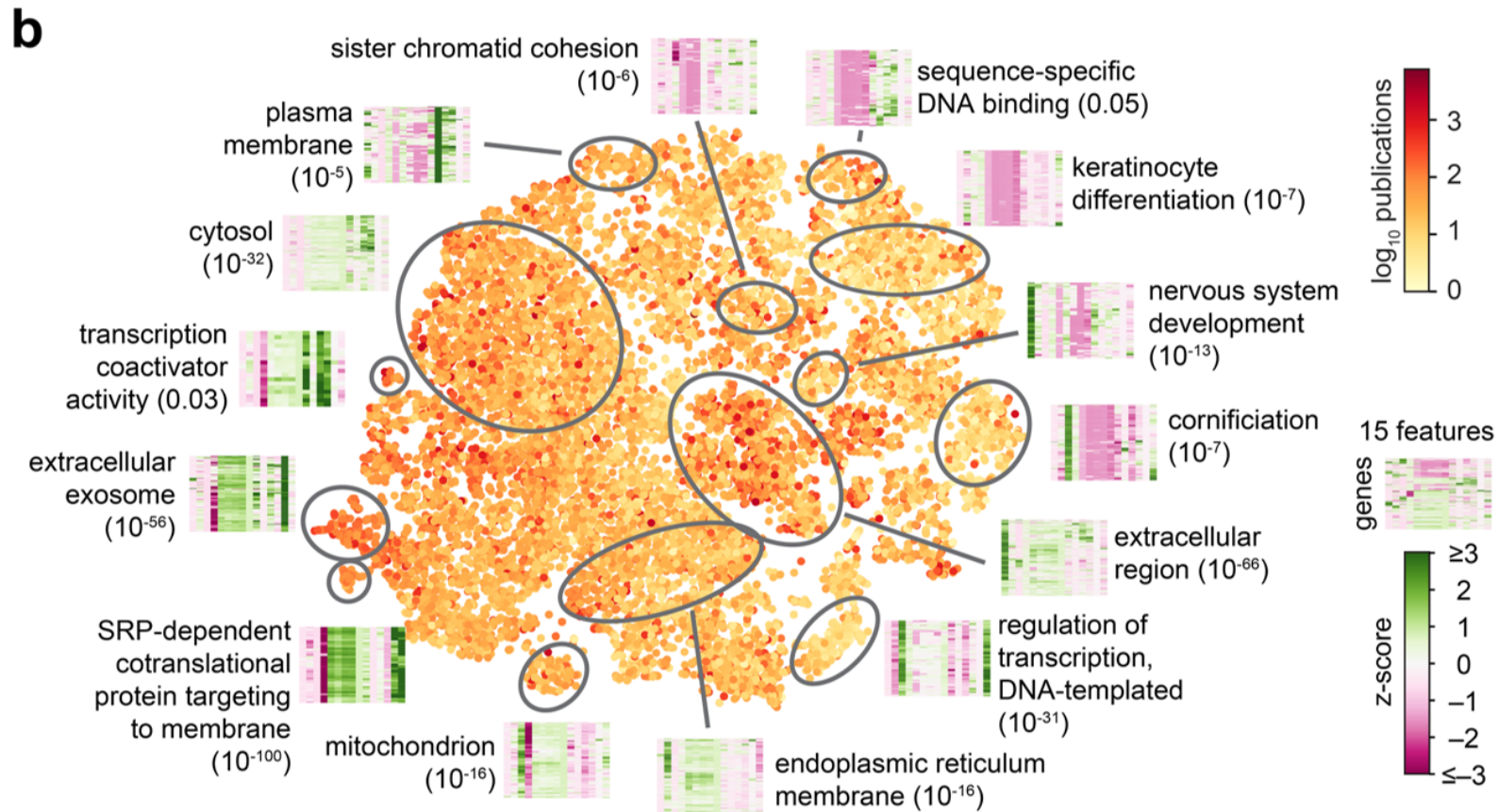
Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

What determines the number of publications per gene?

Individual genes grouped by the embedding technique “t-SNE visualization” using the 15 most informative features that determine #publications / gene.

Neighboring genes are most similar in these features.

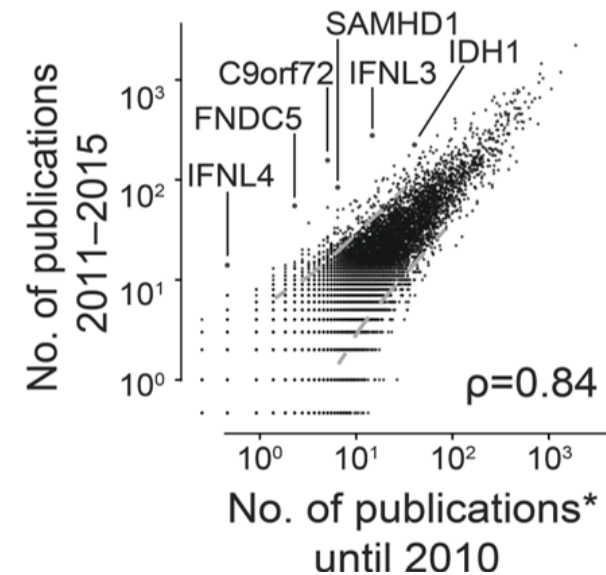


Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

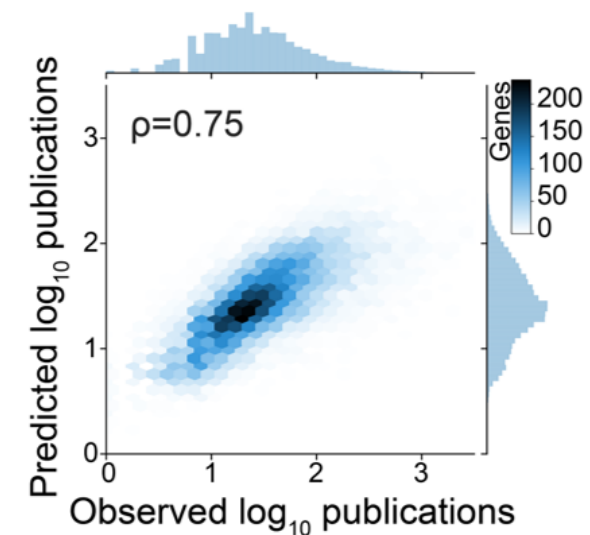
Earlier studied genes continue to be studied

The number of publications per gene is highly correlated between the current decade and preceding time periods of research (Spearman: 0.84).



- > Predict the number of research publications using the 430 features of the previous model AND the year of the first publication on the specific human gene.

Correlation improves from 0.64 to 0.75.



Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

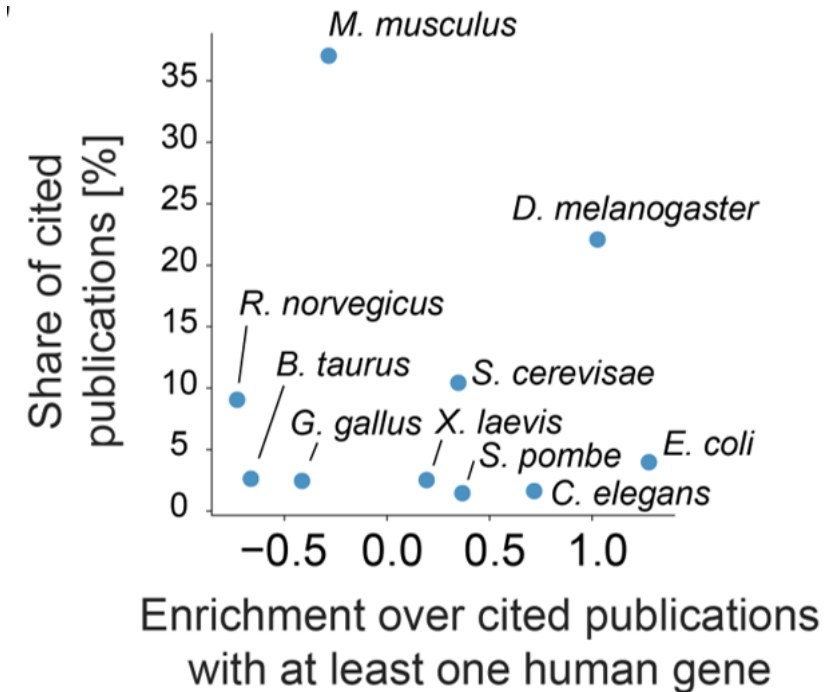
Studies on model organisms affect studies on human genes

Check whether publications reporting the discovery of new human genes also cite studies on (other) human or non-human genes.

(1) One group of papers preferentially cited studies on genes from *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, and *Gallus gallus* AND studies on (other) human genes.

(2) The second group preferentially cited genes from *Drosophila melanogaster*, *S. cerevisiae*, *E. coli*, *Xenopus laevis*, *C. elegans*, and *S. pombe* but DID NOT cite publications on (other) human genes,

-> initial reports on human genes have been particularly influenced by research in model organisms.



Fraction of nonhuman organisms cited by initial publications of human genes.

Enrichment represents log2 ratio of the fraction of nonhuman organisms among all initial publications on human genes over the fraction of nonhuman organisms among initial publications on human genes, which also cite publications on human genes.

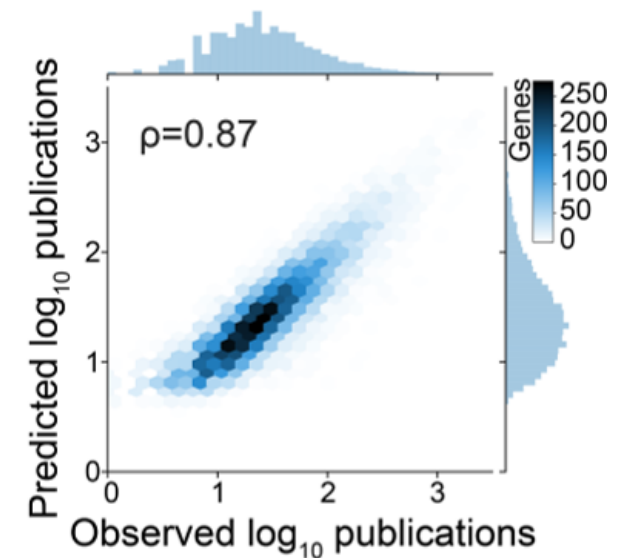
The 10 most cited organisms are shown

Human genes \leftrightarrow homologous genes

Including the years of the initial reports on homologous genes improved prediction accuracy of the number of publications to 0.87.

This is higher than when the year of the initial report on the human genes themselves is used (0.75).

- The number of publications on homologous genes yielded almost perfect predictions of the number of publications for individual human genes (Spearman: 0.87).
- Human-specific genes without homologous genes remain significantly less studied ($p\text{-value} < 10^{-32}$).
- The homologous genes of unstudied human genes are likewise unstudied in model organisms.



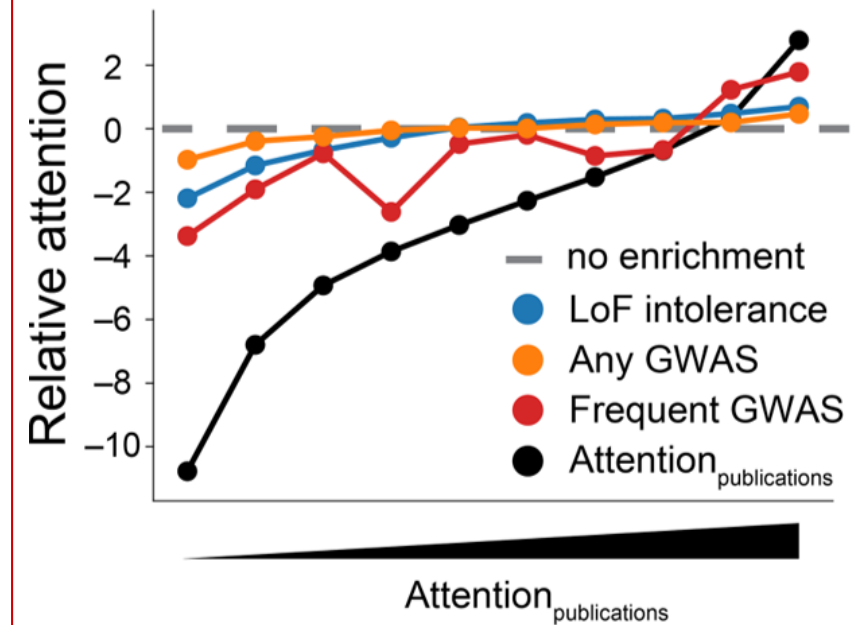
Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

Attention of genes

Attention = fractional counting of publications; rather than counting every publication as 1 towards every gene, the value of a publication towards a given gene is $1/(\text{number of genes considered in the publication})$.

Then, sum all the values of publications citing a particular gene.



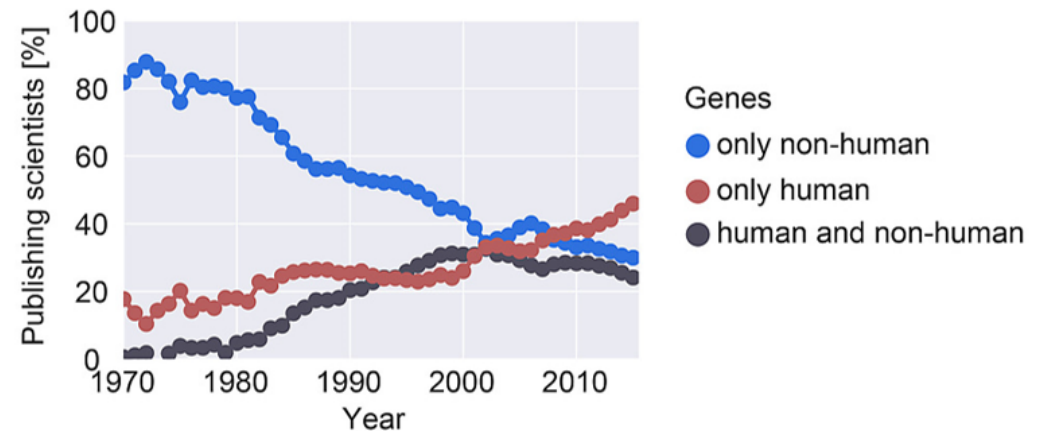
Genes that have received the most attention in publications are around 3 - 5 times more likely to be sensitive to loss-of-function (LoF) mutations or to have been identified in genome-wide association studies (GWAS).

If you visit many doctors, one of them will likely find something. If you study a gene in many ways, the effect of mutations will emerge more likely.

-> A disproportionately high amount of research effort concentrates on already well-studied genes.

Scientists working only on model organisms declining

-> Fraction of scientists who—within the indicated year—publish exclusively on nonhuman genes (or gene products) or exclusively on human genes (or gene products), or both.



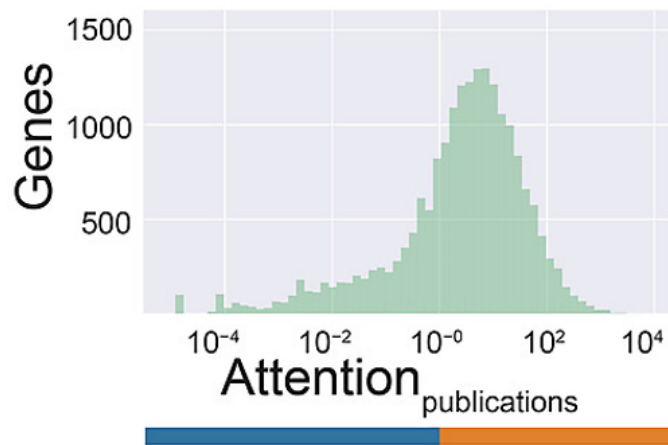
The fraction of scientists who exclusively published on human genes had been stable in the 1980s and 1990s, while the fraction of scientists working only on nonhuman genes has been steadily decreasing at the expense of scientists publishing exclusively on nonhuman genes.

Around 2000, the fraction of scientists working on human and nonhuman genes started to plateau, while the fraction of scientists working exclusively on human genes increased by approximately 10 percent points and has since been steadily increasing.

Stoeger et al. (2018)

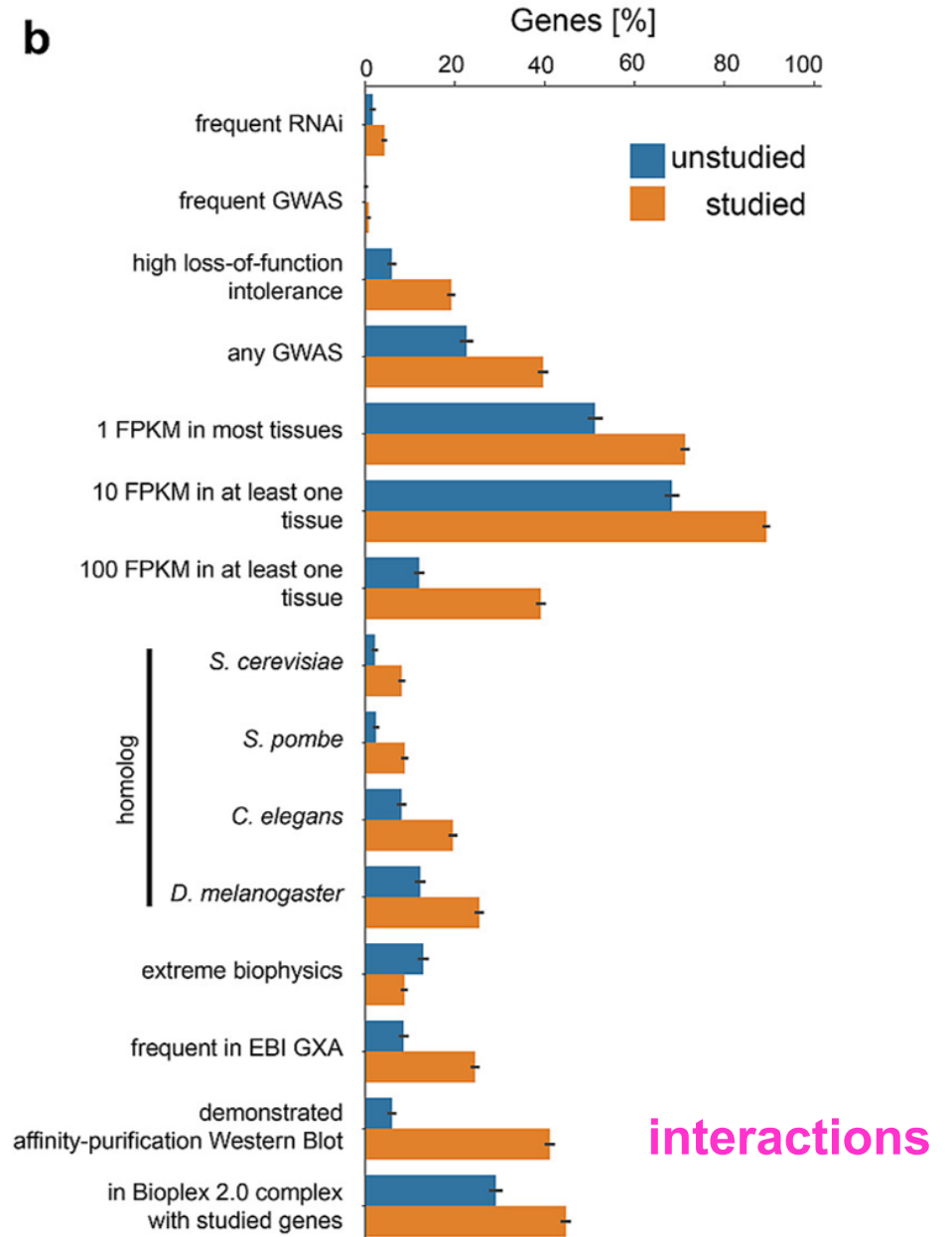
PLoS Biol 16(9): e2006643.

What do we know about genes



(A) Distribution of the attention in publications given to genes. Genes with attention levels below 1 are denoted **unstudied** (blue), whereas genes with attention levels above 1 are denoted **studied** (orange).

(B) Percentage of genes with indicated characteristic.



Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

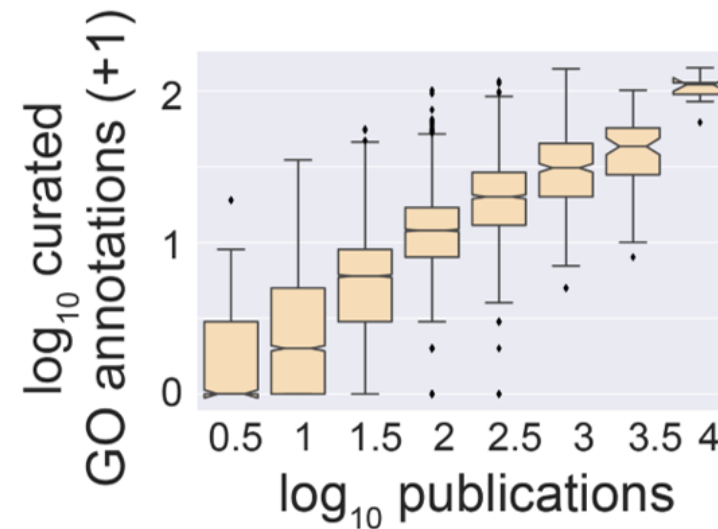
Processing of Biological Data

Summary

Using machine learning, we can predict the number of publications on individual genes, the year of the first publication about them, the extent of funding by the National Institutes of Health, and the existence of related medical drugs.

We find that biomedical research is primarily guided by a handful of generic chemical and biological characteristics of genes, which facilitated experimentation during the 1980s and 1990s, rather than the physiological importance of individual genes or their relevance to human disease.

of human-curated GO annotations for individual genes, binned by number of publications are also heavily biased!



Stoeger et al. (2018)

PLoS Biol 16(9): e2006643.

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.

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

The Gene Ontology (GO)

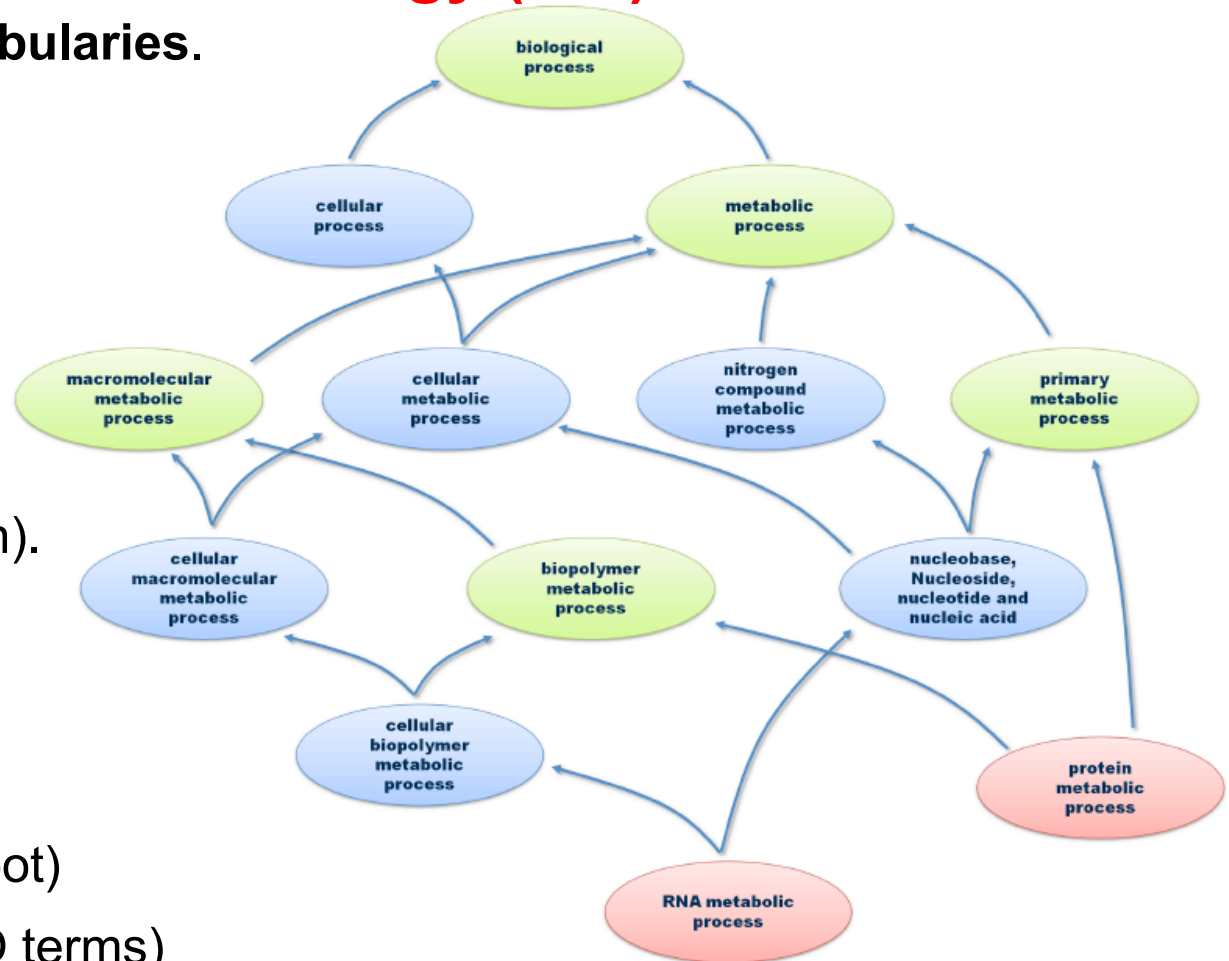
Ontologies are **structured vocabularies**.

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)

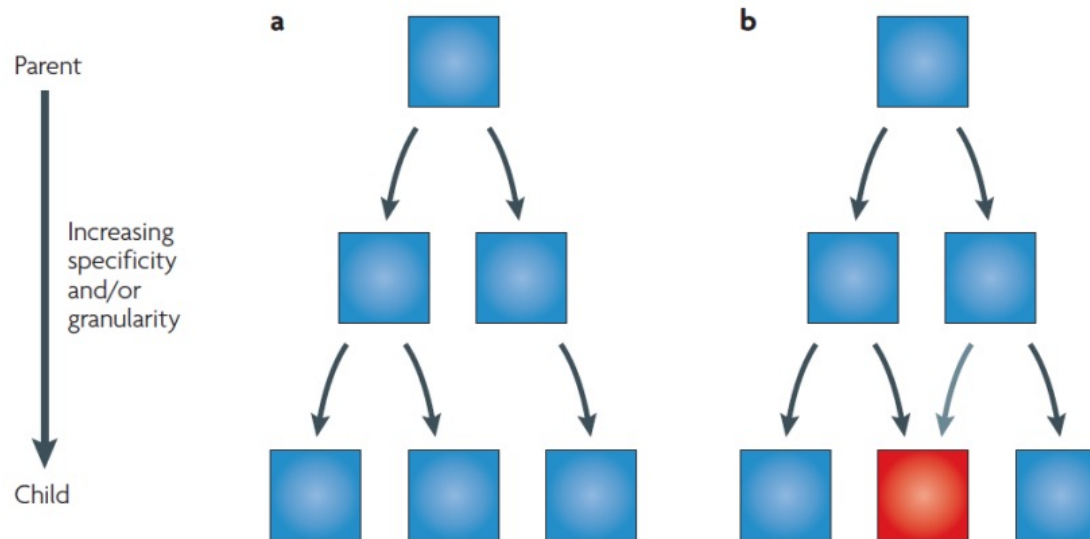
Green: common ancestor

Blue: other nodes.

Arcs: relations between parent and child nodes

PhD Dissertation Andreas Schlicker (UdS, 2010)

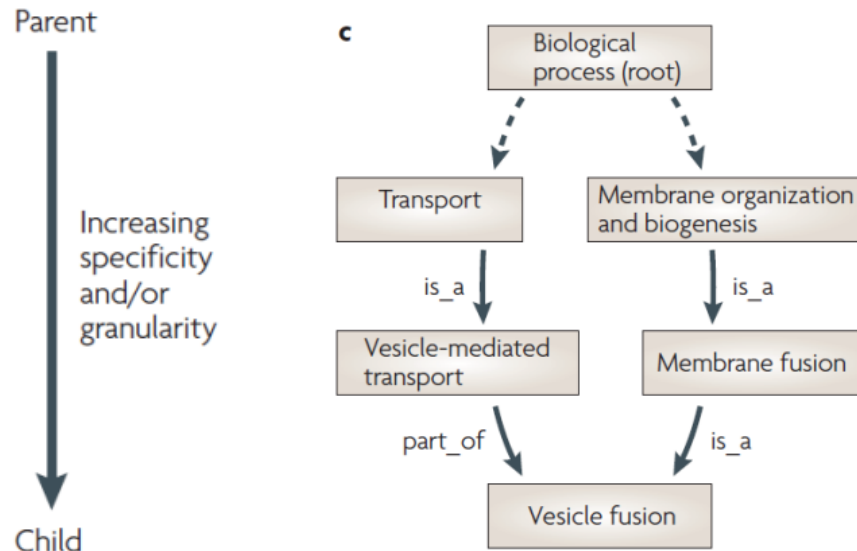
Simple tree vs. cyclic graphs



a | 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 either one or more parents. The node with multiple parents is colored red and the additional edge is colored grey.

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).

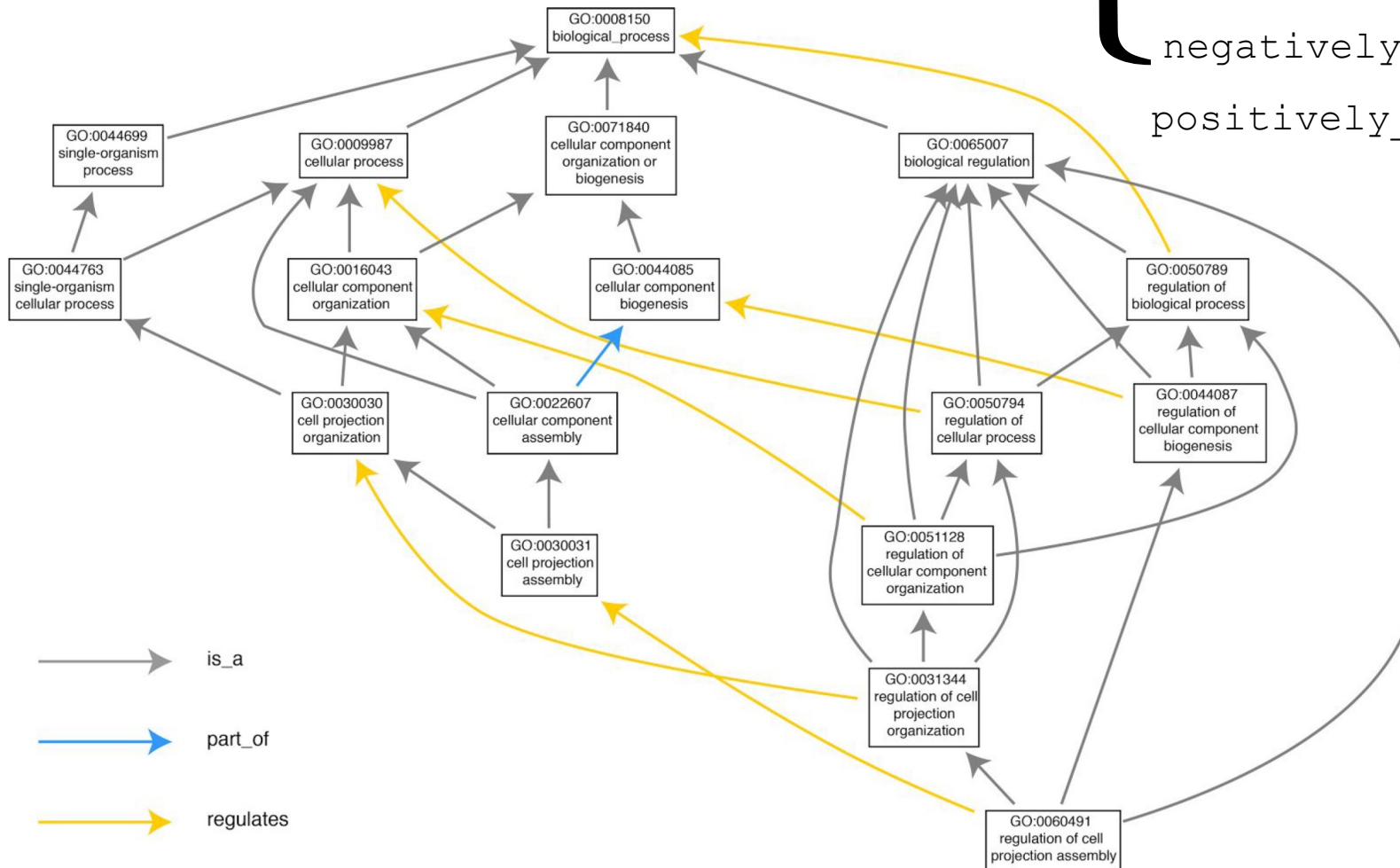
Similar to a simple tree, a DAG has directed edges and does not have cycles.

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>

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

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

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

The hypergeometric test is a statistical test.

It can be used to check e.g. whether a biological annotation π 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_{π} : number of genes in the genome with annotation π .
- k_{π} : number of genes in test set with annotation π .

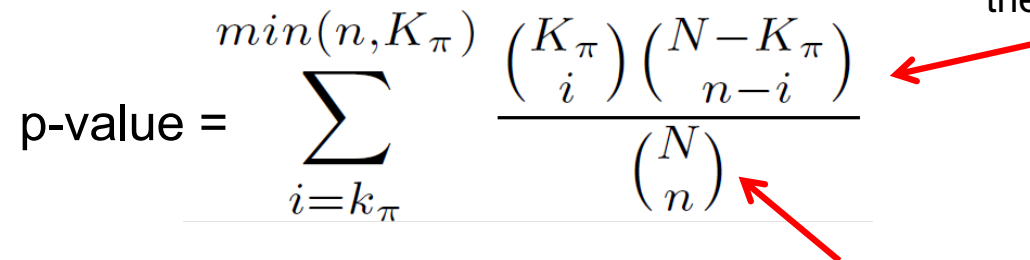
The hypergeometric test provides the **likelihood** that k_{π} or more genes that were **randomly selected** from the genome also have annotation π .

Hypergeometric test

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

There are K_π such genes.

The other $n - i$ genes in the test set do NOT have annotation π .
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_π elements to the maximal possible number of elements.

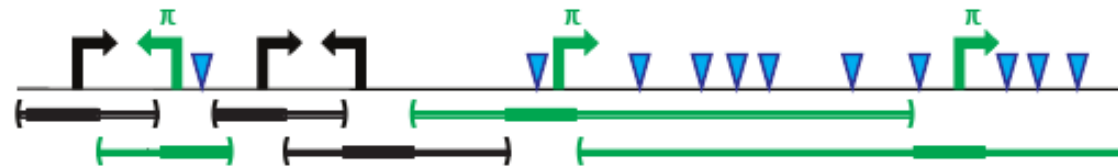
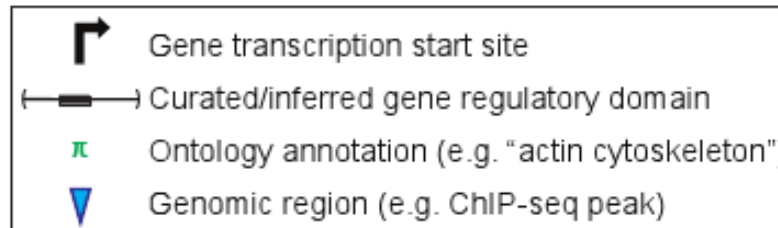
This is either the number of genes with annotation π in the genome (K_π) or the number of genes in the test set (n).

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

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 π significantly enriched in the test set of 3 genes?

Hypergeometric test over genes

N = 6 total genes

K_{π} = 3 genes annotated with π

n = 3 genes with an associated genomic region

k_{π} = 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 GO 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

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

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. This method saturates 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 leads to 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)

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**.

Gaudet, Škunca, Hu, Dessimoz

Primer on the Gene Ontology,

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

V9

Processing of Biological Data

Information content of GO terms

The **likelihood** of a node t is typically defined in the following way:

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

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

.

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

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

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

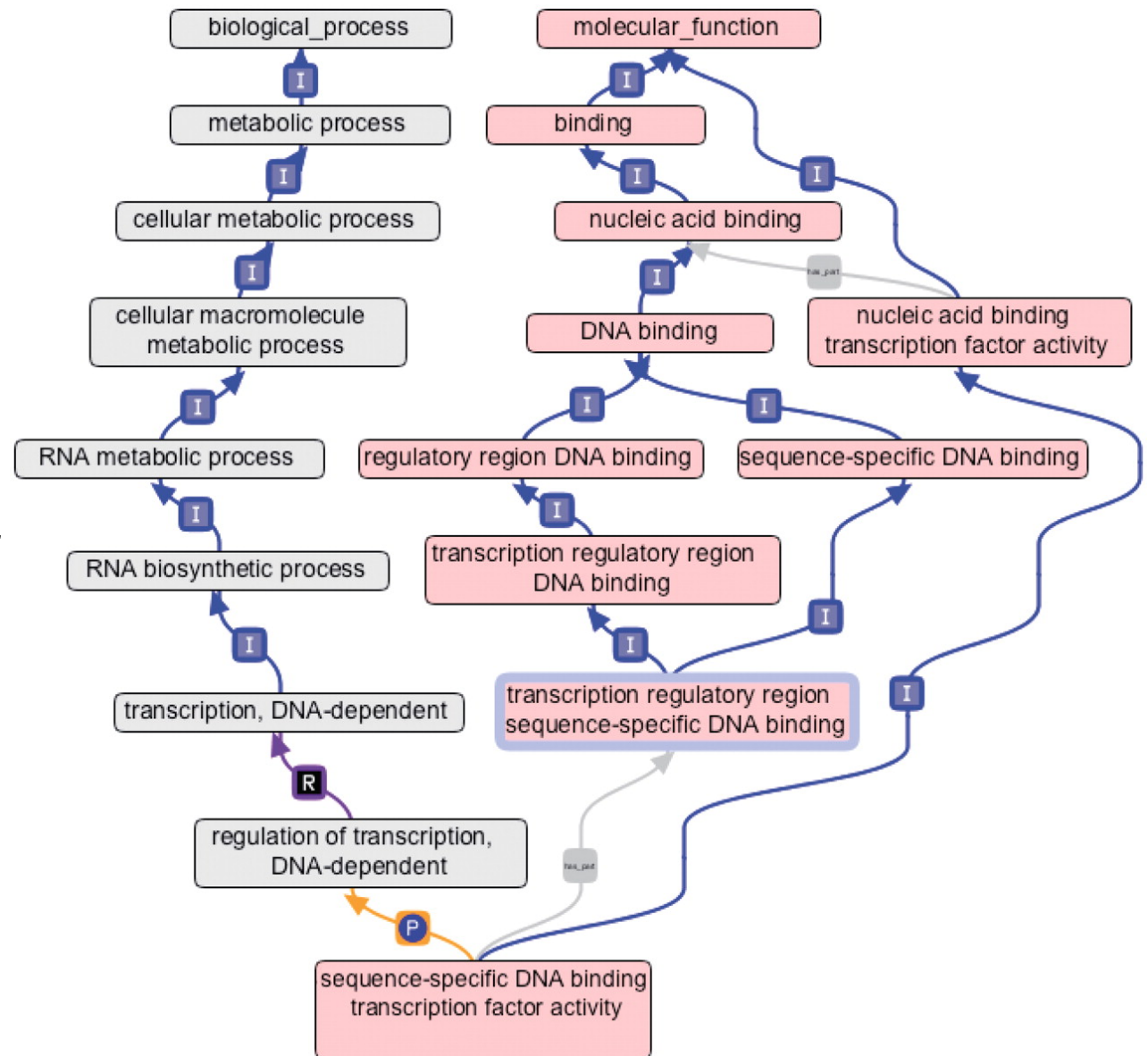
A rare node has high information content.

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)

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

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

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**

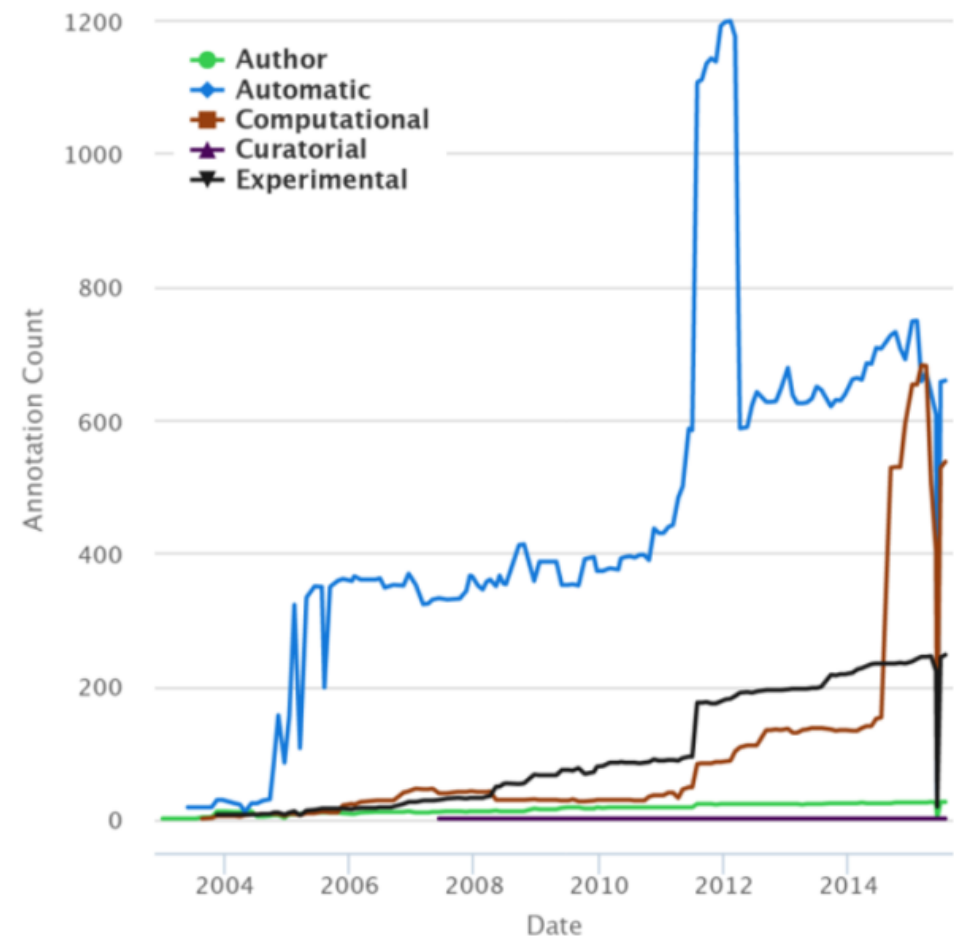
Gaudet, Dessimoz,
Gene Ontology: Pitfalls, Biases, Remedies
<https://arxiv.org/abs/1602.01876>

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 minimize 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,

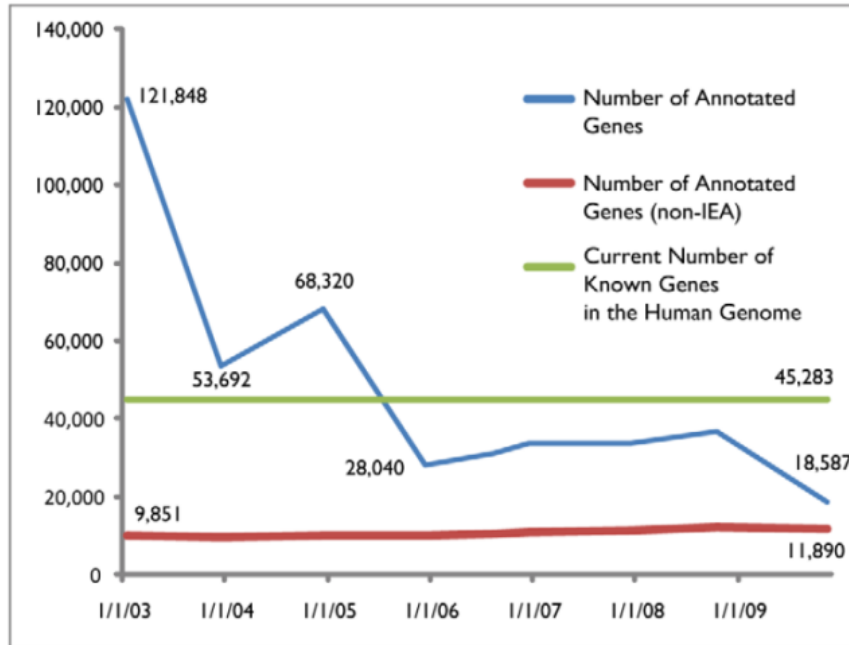
Gene Ontology: Pitfalls, Biases, Remedies

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

V9

Processing of Biological Data

Number of GO-annotated human genes



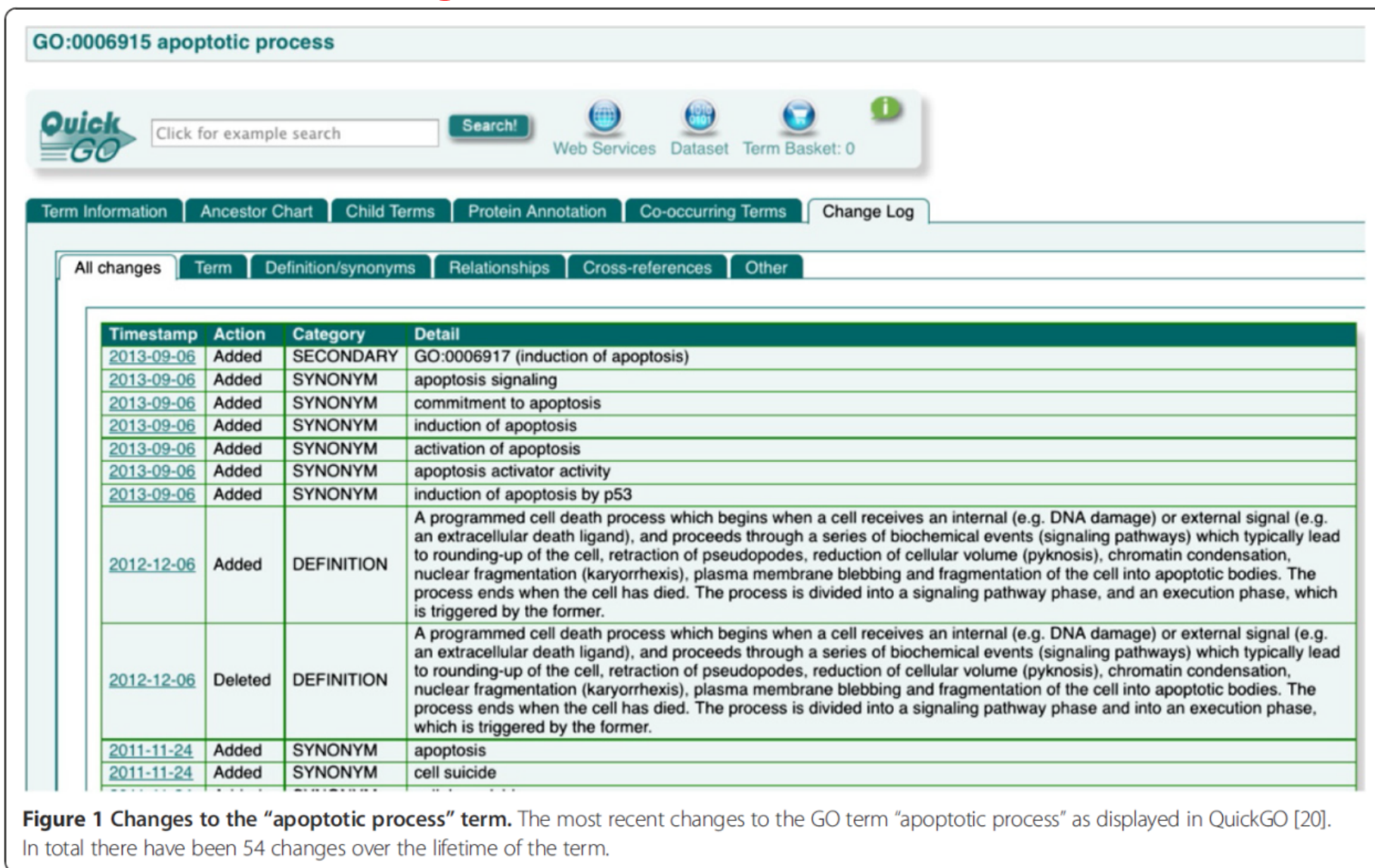
Between 01/2003 and 12/2003 the estimated number of known genes in the human genome was adjusted.

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**.

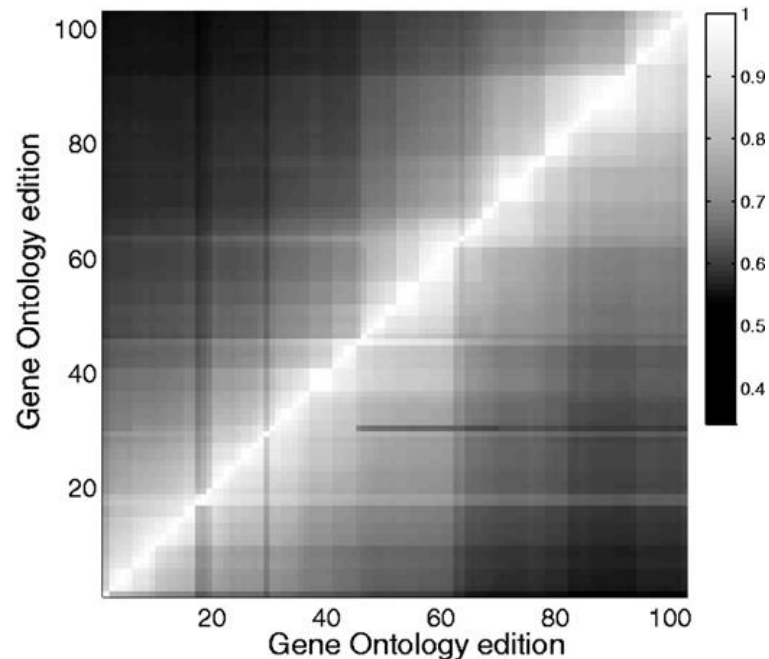
However, this increase in the number of genes with non-IEA annotations is very slow. Between 11/2003 and 11/2009, only 2,039 new genes received non-IEA annotations. At the same time, the number of non-IEA annotations increased from 35,925 to 65,741, indicating a strong **research bias** for a small number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

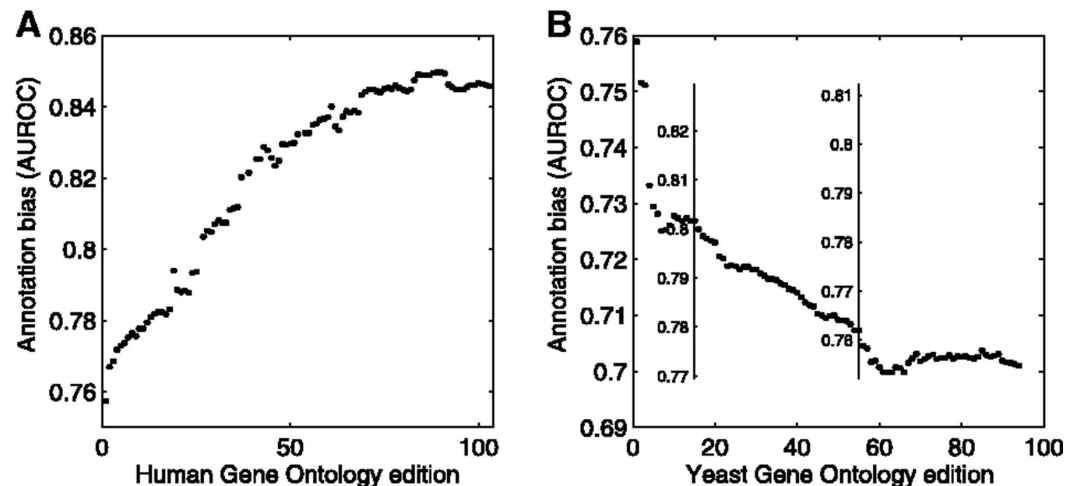
This means that, each month, the annotations of about 3% of the genes have changed so substantially that they are not functionally 'the same genes' anymore.

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

Annotation bias persists in the GO

Annotation bias: defined as area under ROC curve for ranking the genes by the number of GO terms.

If all genes had the same number of GO terms, the **annotation bias** would be 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 over time. Genes with many annotations have become more dominant within GO over time.

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

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.

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!**

Effect of high-throughput experiments

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

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

This information is further propagated by automated methods.

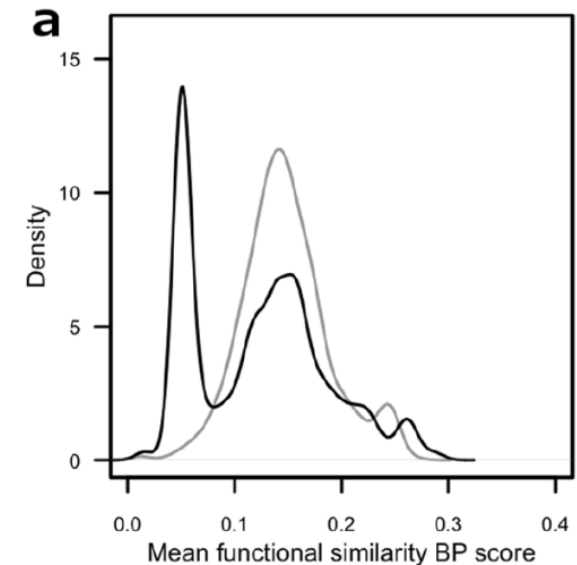
The huge body of electronic annotations (evidence code IEA) has therefore a strong influence on semantic similarity scores.

Influence of electronic annotations (IEA): BP scores

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

Shown are mean BP scores for different human proteins and in each case 1000 randomly selected mouse proteins.

- 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

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.

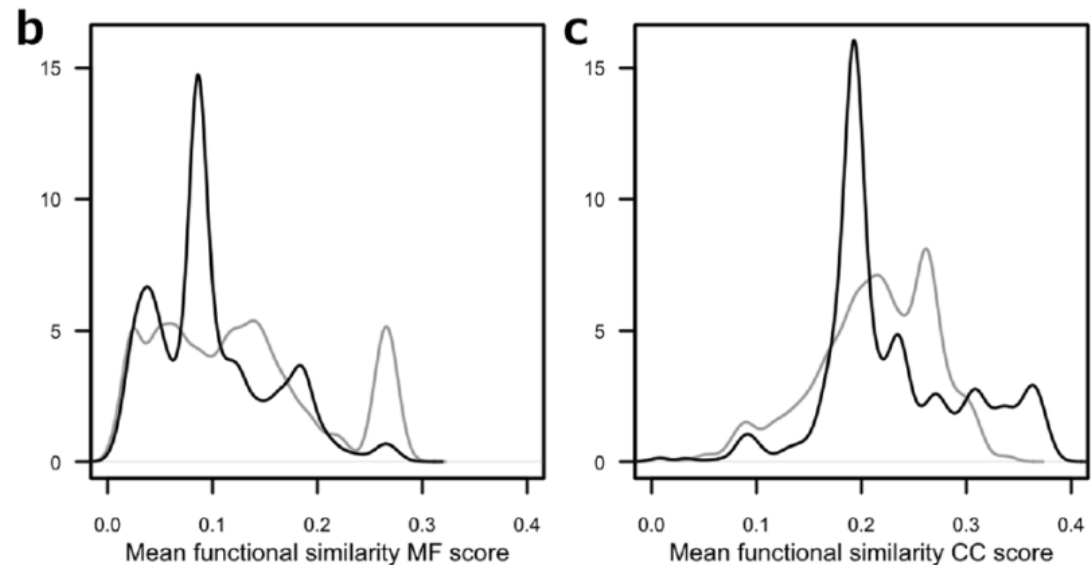
GO enrichment analysis of these proteins shows that they are significantly enriched in “protein binding” (GO:0005155, $p < 10^{-100}$).

This suggests that gene products annotated to this term generally yield much higher than average *simLin/fsAvg* MF scores.

Weichenberger et al. (2017)

Scientific Reports 7: 381

V9



(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.

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 semantic similarity (SS) values $s_{ij} = \text{sim}(t_i, r_j)$ with 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).

Several mixing strategies have been suggested:

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

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

Weichenberger et al. (2017)

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

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

A combined functional similarity F is computed by combining any of the semantic similarities for the different ontologies: biological process (F_{BP}), molecular function (F_{MF}), and cellular component (F_{CC}):

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

Optimal functional similarity score

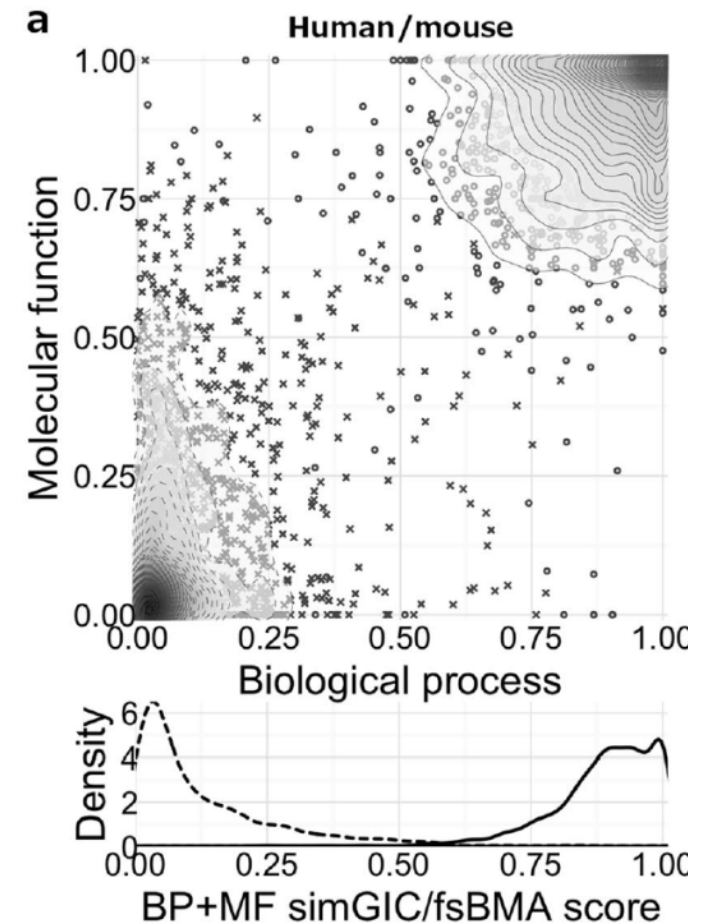
Test: see whether functional similarity score can distinguish true homologues from random gene pairs.

Top: scatter plot of BP (x-axis) and MF (y-axis) scores (IEA⁺ dataset) 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.



Weichenberger et al. (2017)

Scientific Reports 7: 381

V9

Processing of Biological Data

Optimal functional similarity score

Comment:

The human/mouse comparison is based on a cyclic argument:

- Orthologues are defined on the basis of sequence similarity
- Then we test whether their GO-annotations are more similar than for random protein pairs. BUT many GO annotations are made based on sequence similarity.

Thus, this is more a test for consistency rather than a real proof.

Weichenberger et al. (2017)

Scientific Reports 7: 381

V9

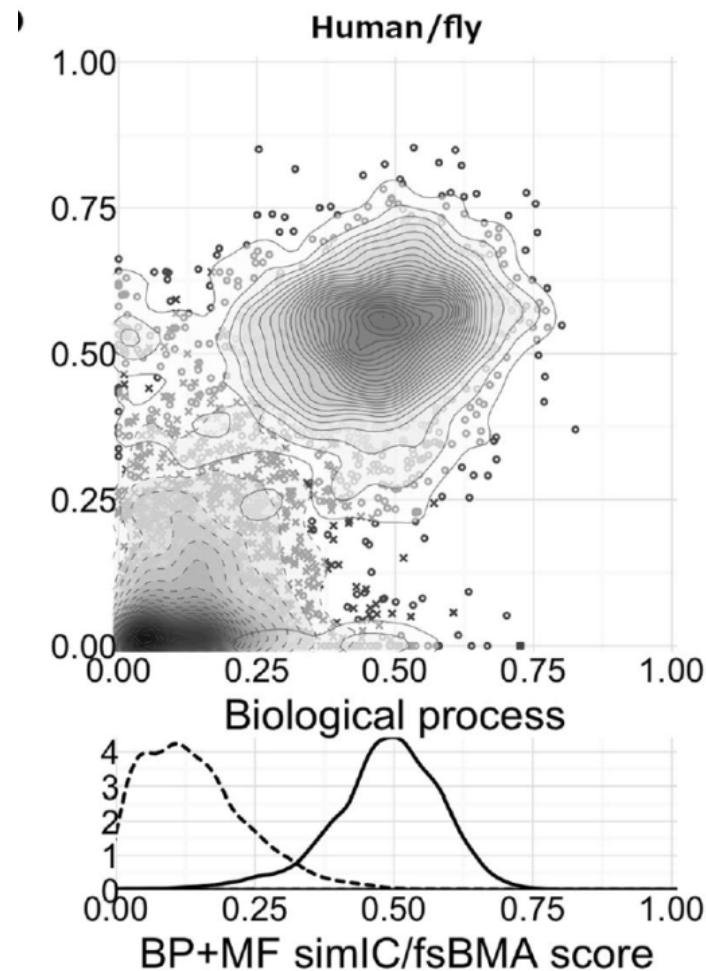
Processing of Biological Data

Optimal functional similarity score

(b) Human/fly

orthologues and controls with their associated *simIC/fsBMA* scores.

-> Slightly larger overlap than for human/mouse.



Weichenberger et al. (2017)

Scientific Reports 7: 381

V9

Processing of Biological Data

Summary

- The GO is the **gold-standard** for **computational annotation of gene function**. It is continuously updated and refined.
- **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
- **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.