V8 – 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

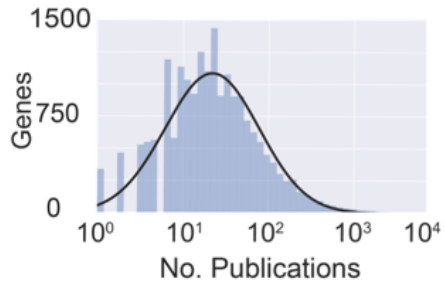
In lecture 8, we will deal with the downstream functional analysis of raw experimental transcriptomics data.

A typical transcriptomics or proteomic experiment may yield a set of upregulated or downregulated genes. Functional annotation then deals with extracting the biological meaning from these findings.

Often, this is done using the hypergeometric test based on functional terms from the Gene Ontology or based on biochemical pathways from KEGG or Reactome.

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.

V8 Processing of Biological Data WS 2021/22 2

Link to this paper:
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2006643

Importantly, the amount of knowledge about individual genes is largely different. This figure shows how many papers were published about individual human protein-coding genes up to 2018.

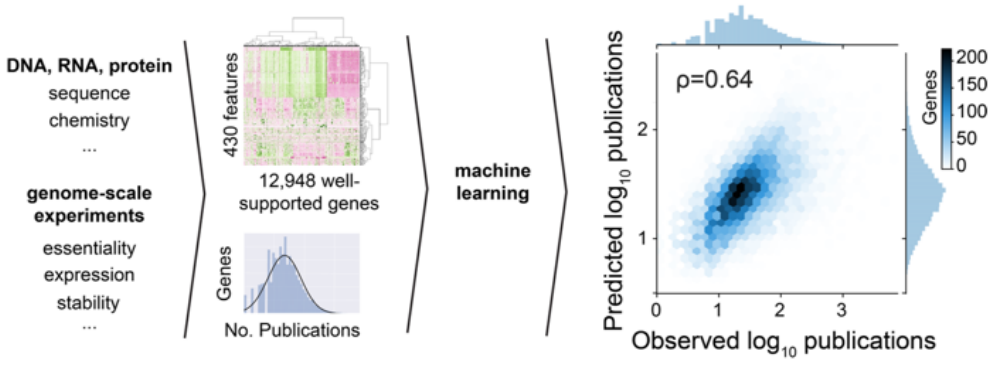
Some genes (right tail of the distribution) were studied by more than 1000 publications. On the other hand, some genes were only addressed by a handful of publications. What is responsible for this imbalance?

Possibly the **most studied genes** are the **most important genes** in terms of their function. But who should decide what functions are important?

Often, the research directions of individual scientists are the result of many coincidences: How did they pick their PhD supervisor and post-doc advisor? What were the bosses working on? Which ones of the many grant applications that scientists write got funded?

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.

V8 Processing of Biological Data WS 2021/22 3

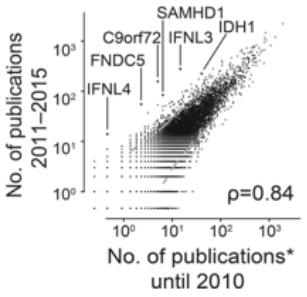
Here, the authors tried to find out which features determine what genes are well studied.

Obviously, genes that can be robustly expressed and proteins that can be easily synthesized have an advantage.

The reason is that many scientists don‘t like to work on „difficult“ things that only work once in a while.

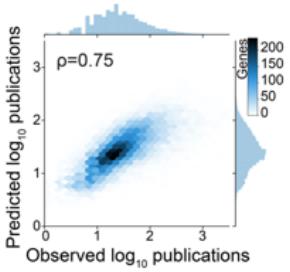
Earlier studied genes continue to be studied

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.

V8 Processing of Biological Data WS 2021/22 4

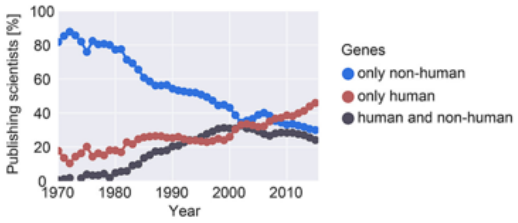
The upper figure shows that the number of publications for a gene in the period 2011-2015 is strongly correlated to the number of publications until 2010.

This shows that scientists continue to study research questions around certain genes that they and others have already studied before.

If one includes the year of the first publication, the prediction accuracy improves considerably, which emphasizes the importance of this feature relative to the other 430 features.

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.



In the 1980s and 1990s, the **fraction of scientists who exclusively published on human genes** had been stable. But there were two opposite trends during this time: the **fraction of scientists working on human and nonhuman genes** has been steadily increasing in parallel to a decrease 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.

There has been a continuous decrease in the scientific activities on model organisms. This negative trend accelerated around the year 2000 in favor of an increased fraction of scientists that exclusively work on human genes.

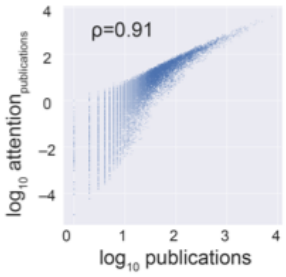
One can speculate whether this is related to the ability of obtaining funding for research projects. Also, this may be due to the availability of the human genome sequence after 2001.

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/(number of genes considered in the publication).

Then, all the values of publications citing a particular gene are **summed**.



Plotted here are the ranking of **fractional counting** versus **normal counting** of publications with multiple genes.
In normal counting, the occurrence of a gene in a publication counts as 1.

Stoeger et al. (2018)
PLoS Biol 16(9): e2006643.

For genes addressed by many publications with log10 > 2, there is a good linear correlation of both counting measures.

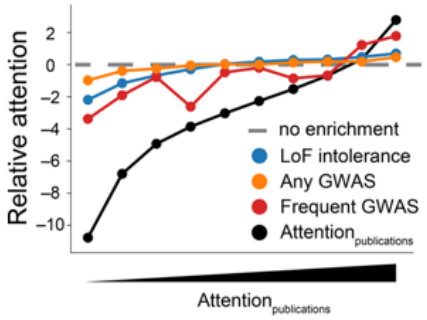
For genes addressed in only few publications, the attention scores based on fractional counting are downward shifted = the attention values of such genes are reduced with respect to normal counting.

Attention of genes

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).

This means that scientists are focusing on „important“ or „relevant“ genes.

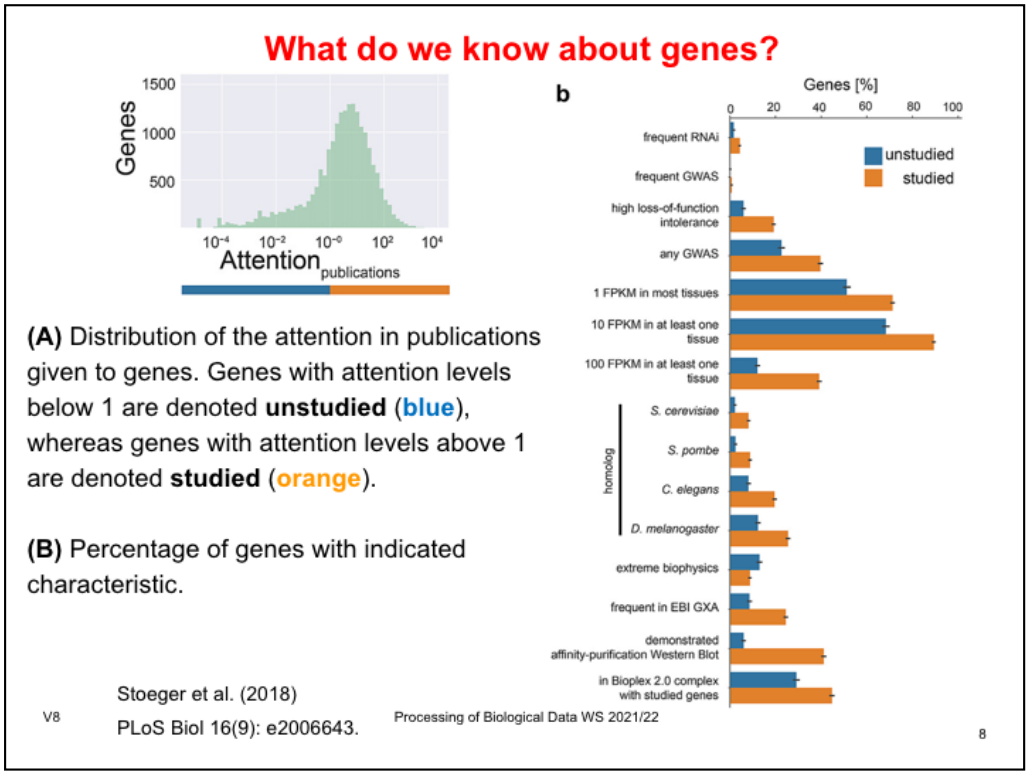
But a disproportionally high amount of research effort concentrates on already well-studied genes.



Stoeger et al. (2018)
PLoS Biol 16(9): e2006643.

Given the observed historic continuity of scientific endeavors, Stoeger et al. wondered whether biomedical research has already identified all particularly important human genes and hence allocates the production of publications accordingly. Inspite of the simplifying assumption made for fractional counting (see previous slide), the authors reassuringly observed that genes that have received the most attention in publications are around three to five times more likely to be sensitive to loss-of-function mutations or to have been identified in genome-wide association studies (GWAS). This enrichment is greatest for genes that have been repeatedly identified by several independent studies **(“frequent GWAS”)** on the most frequently studied human phenotypic traits.

However, one notices an extraordinarily more extreme 13-fold enrichment in the **average attention** (from -10 to more than +2) when comparing the genes that have received the least attention to those genes that have received the highest attention. Hence, while biomedical research does focus on important genes, a disproportionally high amount of research effort concentrates on already well-studied genes.



(Top left) Attention_publication levels. Genes with values below 1 („unstudied genes“) were only addressed in publications addressing several or many genes.

(Right) Statistics whether certain types of experiments have been performed, or whether homologs exist in model organisms.

For some experiments (e.g. Western Blots), there is a drastic difference between „studied“ genes (> 40%) and „unstudied“ genes (< 10%).

Also, „unstudied“ genes are only about half as likely to have a homolog in model organisms.

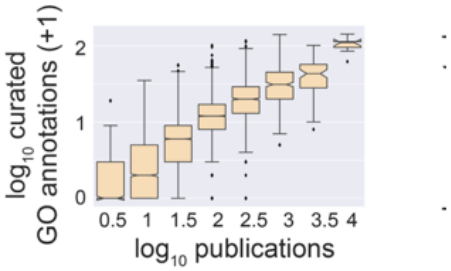
Thus, the „old-fashioned“ scientists who worked and are working on a gene-by-gene basis on model organisms had no chance to detect these genes.

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.

The authors suggest that an insufficient understanding of the biology of many disease genes has prevented the successful development of further medical therapies and that current preclinical research is biased towards experimentally well-accessible genes

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

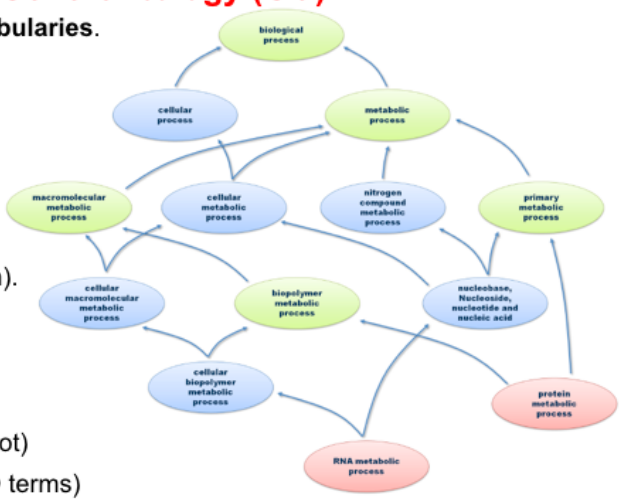
For those of you who are not closely familiar with the Gene Ontology, here is some introduction or review.

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.

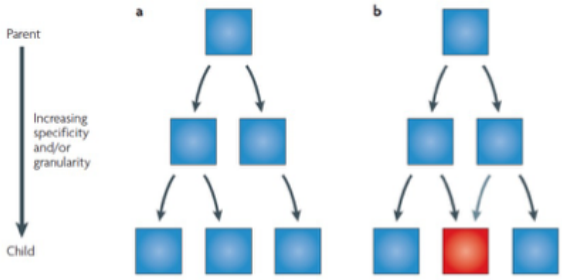
Arcs: relations between parent and child nodes

PhD Dissertation Andreas Schlicker (UdS, 2010)

The Gene Ontology consists of 3 branches: **biological process, molecular function** (chemical details), and the **cellular component** that the encoded protein localizes to.

Each branch starts with a root node on top and subsequent child nodes with more and more specific functions that inherit the functions of all their parents and grand-parents.

Simple tree vs. cyclic graphs



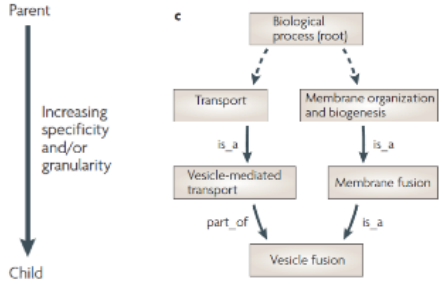
a | A simple **tree**, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge.

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.

Rhee et al. (2008) Nature Rev. Genet. 9: *509*

The Gene Ontology has the topology of a **directed acyclic graph** where child nodes can have multiple parent nodes.

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.

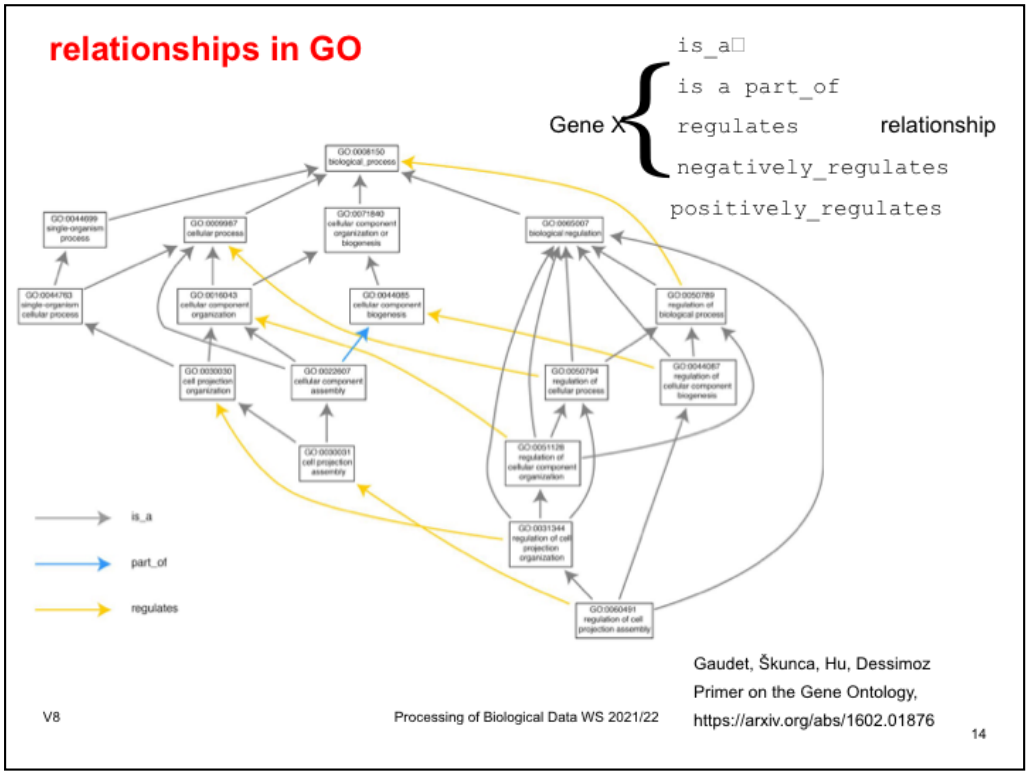
Depth of a node : length of the longest path from the root to that node.
Height of a node: length of the longest path from that node to a leaf.

V8 Rhee et al. (2008) Nature Rev. Genet. 9: *509* Processing of Biological Data WS 2021/22 13

This example shows that the leaf node „vesicle fusion“ (found e.g. in endocytosis and exocytosis and in vesicular transport between different compartments) has two branches of parent nodes.

The left branch focuses on the vesicles, the right branch on the membrane processes.

Although the arrows are directed downwards in this figure, they should be read in the opposite direction. E.g. „vesicle fusion“ is a „part_of“ „vesicle-mediated transport“, not the other way around.



Here, the arrows are oriented in the correct upward direction.

There exist five different types of relationships shown on the top right.

All terms (except from the root terms representing each aspect) have an "**is a**" sub-class relationship to another term; e.g. GO:1904659:glucose transport is a GO:0015749:monosaccharide transport.

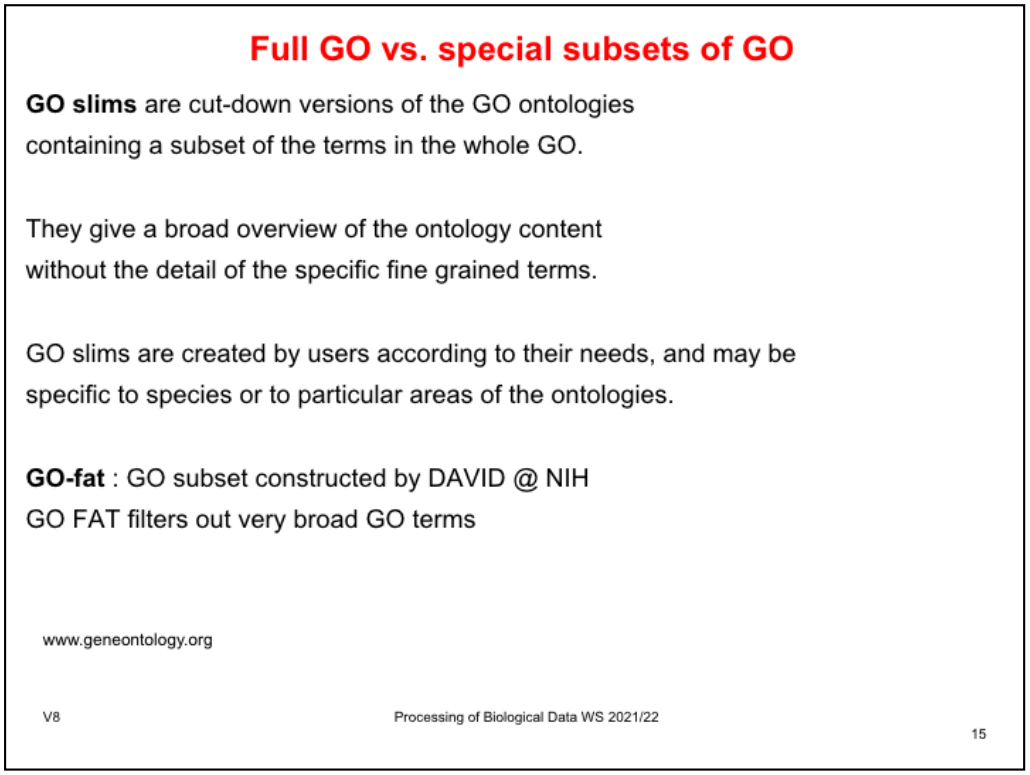
The Gene Ontology employs a number of other relations, including "**part of**", e.g. GO:0031966:mitochondrial membrane is part of GO:0005740:mitochondrial envelope

and "**regulates**", e.g: GO:0006916:anti-apoptosis regulates GO:0012501:programmed cell death

As shown in the figure, „regulating" arrows may connect different branches or reach directly to upper levels.

Obviously, „negatively_regulates" and „positively_regulates" are specifications of „regulates". Sometimes, the direction of regulation (up/down) may not be known – then one would assign „regulates".

Also, in some cases, the direction of regulation may be in both directions depending on the particular condition. Also then, one would assign „regulates".



The gene ontology terms are of different nature ranging from very general terms that are annotated to thousands of genes to very specialized terms that are annotated only to few genes.

Depending on the application, scientists may consider using either only subsets of general terms (**GO slim**) or subsets of specific terms (**GO fat**).

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**.

PhD Dissertation Andreas Schlicker (UdS, 2010)

Often, one wants to annotate biological meaning e.g. to the results of a differential expression analysis. It may not be helpful to know that half of the upregulated genes carry out „metabolic processes".

But it would be very helpful to know if several among them are e.g. annotated with „**purine nucleotide biosynthetic process** ", which is a much more specific GO term (0006164).

Hence, one needs to determine the statistical significance of the fact that out of 393 human genes in total that are annotated with this GO term, e.g. 100 are up-regulated.

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.

V8 Processing of Biological Data WS 2021/22 http://great.stanford.edu/ 17

Often, one uses the hypergeometric test to compute a p-value for the statistical significance of GO terms.

The formula needs to be interpreted in the following way:

In the denominator (Dt. Nenner), we consider the combinatorial number of drawning n genes out of a large set of N genes.

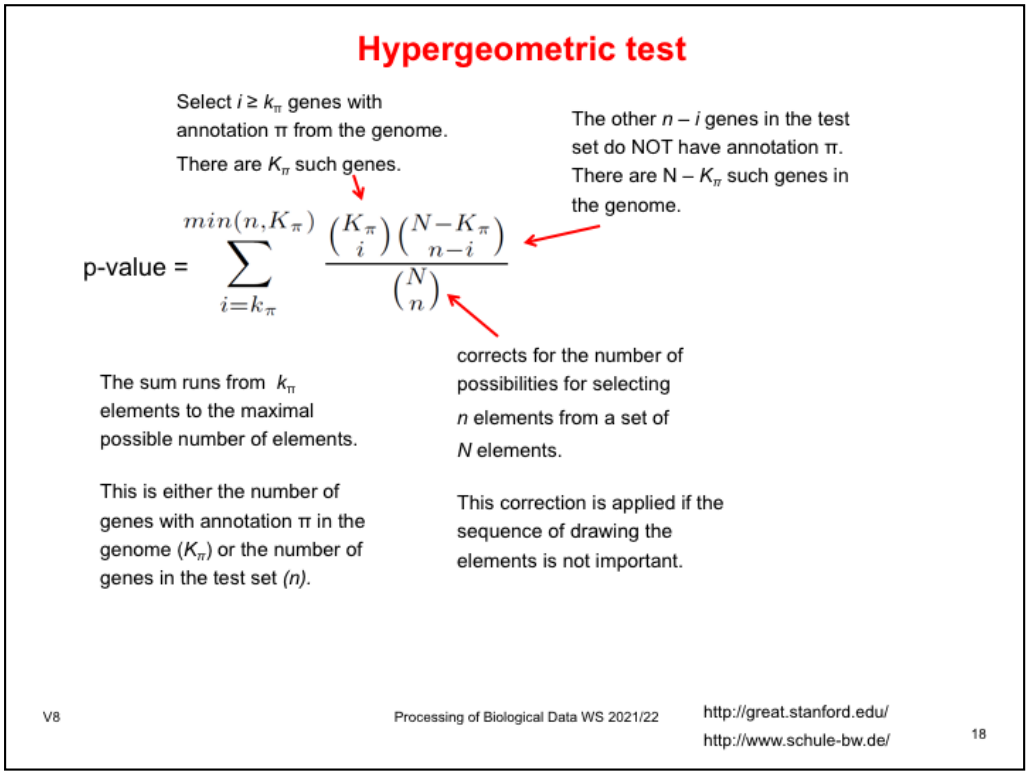
In the numerator (Dt. Zähler), we enter the current situation: the first term is the number of i genes having a particular GO term (out of K_π genes in the full set of N genes).

The second term considers the remaining n-i genes that do not have this GO term assigned (here, we assume that they then actually do not have this function – which may be incorrect due to partial knowledge).

These n-i genes can be drawn from the remaining N-K_π genes in the full set of N genes that do not have this GO term assigned.

By computing this ratio, we compute the number of cases where we could generate such a scenario by chance.

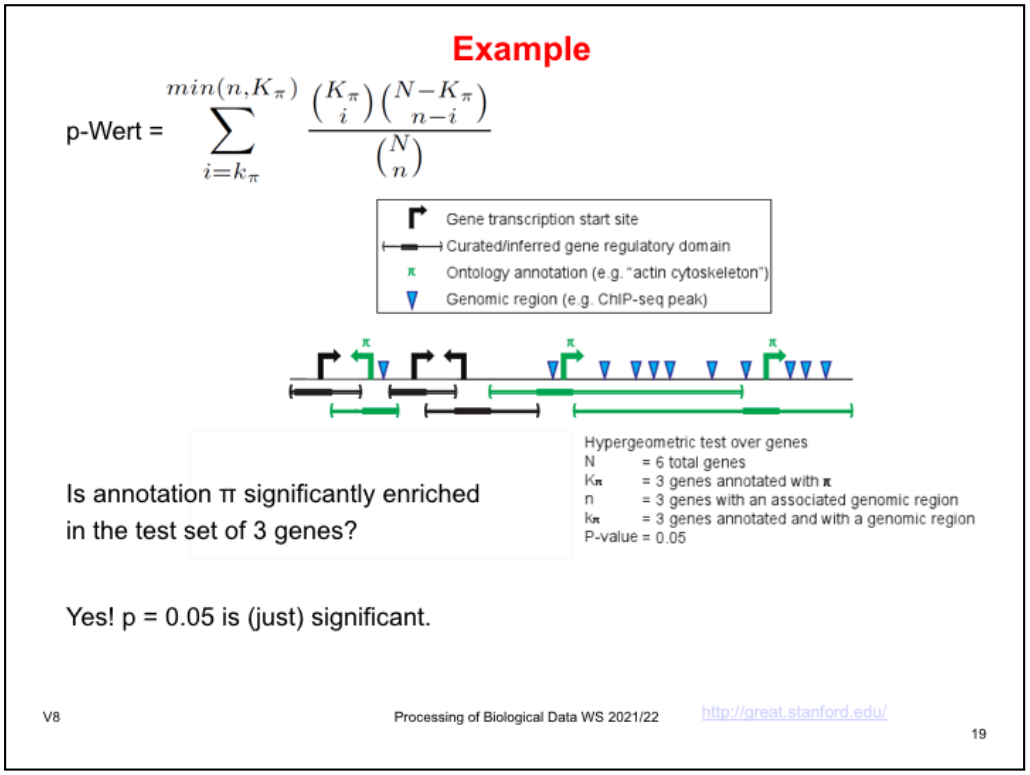
If there exist many such cases, then the p_value would be quite high, and hence the statistical significance low.



The p-value is the probability that a scenario at least as extreme as observed could occur by chance.

Therefore, we also consider cases where more than k_π genes in the small set of n genes are annotated with this GO term. This is the reason why we need to sum over all these more extreme cases.

At least k_π genes should have the GO term. At most all n genes could be annotated with this GO term.



This is a small-scale example, where we can evaluate the hypergeometric test by hand. We assume a case where a genome contains only N = 6 genes (linear bars between brackets below the line, the arrows indicate the position of transcriptional start sites and the direction of transcription). Further, we assume that the $K_\pi = 3$ genes colored green possess a property (GO annotation) pi.

Now we perform an experiment, e.g. differential expression analysis, and find that n = 3 genes are upregulated in condition 2 vs. condition 1. Interestingly, all these 3 genes have property π -> $k_\pi = 3$.

Is this reason enough to get superexcited about this finding? What is the chance of obtaining a similar result by chance, i.e. blindly picking the 3 white balls out of a box with 3 white balls and 3 black balls.

In total, there are 6 over 3 possibilities of selecting 3 genes out of 6 genes. In this example k_π , K_π and n are all equal to 3. Therefore, we only need to consider the case $i = 3$ and can omit the summation.

In the numerator, the first term is 3 over 3, which is equal to 1 by definition. The second term is 3 over 0, which is also equal to 1 by definition.

The denominator is 6 over 3, which is (6 x 5 x 4) / (1 x 2 x 3) = 20. So the observed result of this experiment is just significant (p-value = 0.05).

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)

In the example just discussed, we had considered only 1 property named π.

However, in a typical differential expression analysis, we consider a large number of GO terms.

This leads to a severe problem, the so-called multiple testing problem, because we subject the same experimental outcome (which genes are up/down-regulated for a given number of samples?) to many statistical tests for the various GO terms. Each hypergeometric test applies to a particular GO term.

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

Now we will discuss the so-called multiple testing problem.

This typically leads to the application of the **False Discovery Rate** (FDR) correction of the obtained p-values and yields „adjusted p-values“.

First, we need to understand what the problem is.

There is no problem if we only perform one statistical test where we test one null hypothesis.

The problem arises if we conduct a lot of statistical tests on the same data.

For example, we could have a cohort of 100 tumor patients and 100 healthy individuals. The first test could be to see if gene 1 is differentially expressed between both groups.

The second test would be the same for gene 2 and so on. In the end, we would have conducted 20.000 statistical tests.

The chance that some of these genes will in fact show a significant difference between both groups is very high.

Bonferroni correction

Carlo Bonferroni (1892-1960) did not invent the „Bonferroni" correction, but it uses his inequalities.

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); wikipedia.org

Let us consider the same example (100 healthy, 100 tumor patients, 20000 genes and assume that the smallest (not adjusted) p-value is 10^{-5}.

The Bonferroni correction simply multiplies all p-values by the number of statistical tests (20000). This yields 2×10^{-1} as smallest adjusted p-value, which would not be considered significant.

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)

Let us consider an example where 500 genes were determined as differentially expressed.

With a "false discovery rate" set to 0.1, this actually means you expect 50 of them to be false positives, so they are actually NOT differentially expressed.

This is a nice video that motivates the BH method:
https://www.youtube.com/watch?v=K8LQSvtjcEo

Benjamini Hochberg correction: how to recipe

0. Select a FDR threshold Q (this is a percentage, chosen by you). Depending on the specific project, FDR may be set to values between 1% and 25%.

1. Put the individual p-values in ascending order.

2. Assign ranks to the p-values. For example, the smallest has a rank of 1, the second smallest has a rank of 2 etc

3. Calculate each individual p-value's Benjamini-Hochberg critical value, using the formula $(i/m)Q$, where:

   i = the individual p-value's rank,

   m = total number of tests,

   Q = the false discovery rate

4. Compare your original p-values to the critical B-H from Step 3; find the largest p value that is smaller than the critical value.

https://www.statisticshowto.com/benjamini-hochberg-procedure/

Steps 1 – 4 are the main steps of the Benjamini Hochberg procedure.

I have added step 0 to this because the FDR threshold should be determined first, not after seeing what results are obtained.

Benjamini Hochberg correction: how to recipe

As an example, the following list of data shows a **partial list of results from 25 tests** with their p-values in column 2.

The list of p-values was ordered (Step 1) and then ranked (Step 2) in column 3. Column 4 shows the calculation for the critical value with a false discovery rate of 25% (Step 3). For instance, column 4 for item 1 is calculated as (1/25) * .25 = 0.01:

The bolded p-value (for Children) is the highest p-value that is also smaller than the critical value: .042 < .050. **All** values above it (i.e. those with lower p-values) are highlighted and considered significant, even if those p-values are not lower than the critical values.

Variable	P Value	Rank	(I/m)Q
Depression	0.001	1	0.01
Family History	0.008	2	0.02
Obesity	0.039	3	0.03
Other health	0.041	4	0.04
Children	**0.042**	**5**	**0.05**
Divorce	0.060	6	0.06
Death of Spouse	0.074	7	0.07
Limited income	0.205	8	0.08

E.g. Obesity and Other Health are individually not significant when you compare the result to the final column (e.g. .039 > .03). However, with the B-H correction, they are considered significant; i.e. you would reject the null hypothesis for those values.

https://www.statisticshowto.com/benjamini-hochberg-procedure/

This is an example how FDR-adjusted p-values are computed in practice.

Column 2 contains the p-values obtained by applying a statistical test to the data, e.g. a t-test.

Then, for a particular FDR-threshold, one determines the critical value (I/m)xQ.

Interestingly, the magnitude of the p-values itself does not enter here.

If the p-values are very small, they have a better chance of being smaller than the critical value. Note that p-values tend to become smaller and smaller the more data points are available.

On the other, the critical values decrease inversely with the number of tests performed (m). This penalizes against doing many tests on the same data.

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**

Now, we will discuss an important aspects of the Gene Ontology: its incompleteness.

(1) The functional annotations in GO try to follow the expansion of the scientific knowledge, but can only do this with a significant time delay. Also, it is impossible to completely cover all scientific discoveries.

Sometimes, there may be even contradictory scientific reports in the literature about the function of one gene.

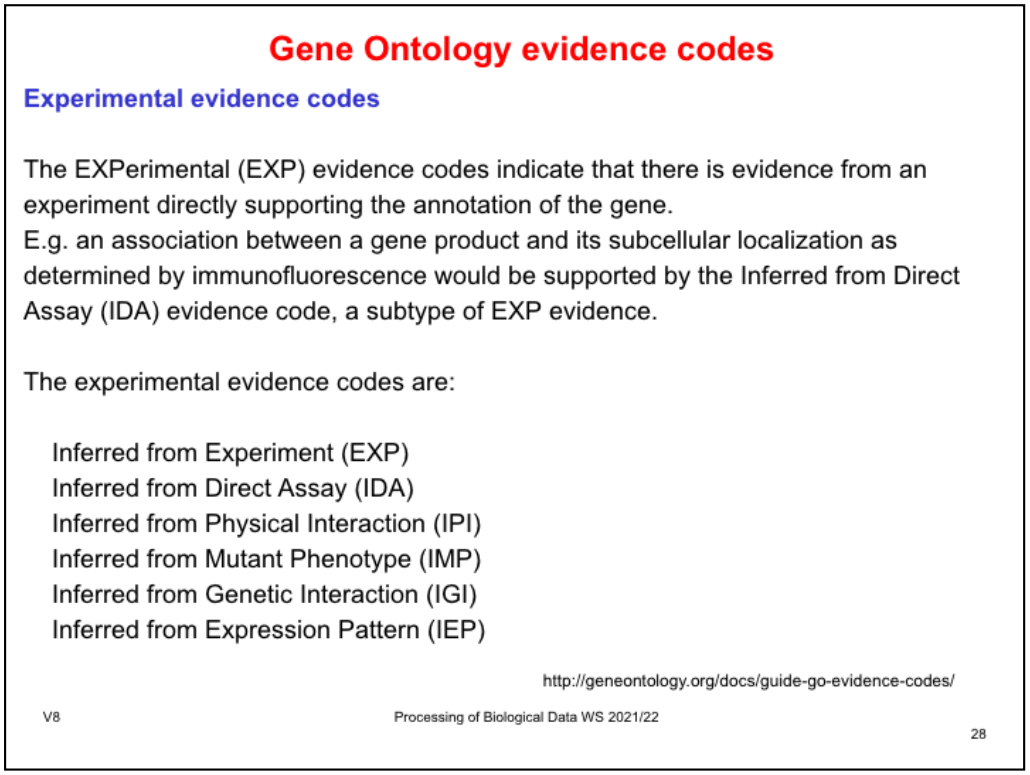
Statistics of Gene Ontology terms

Ontology	Property	Value
	Valid terms	44411 (Δ = -97)
	Obsoleted terms	2947 (Δ = 23)
	Merged terms	2056 (Δ = 91)
	Biological process terms	29112
	Molecular function terms	11118
	Cellular component terms	4181
Annotations	Property	Value
	Number of annotations	7,975,639
	Annotations for biological process	3,069,526
	Annotations for molecular function	2,455,089
	Annotations for cellular component	2,451,024
	Annotations for evidence PHYLO	4,163,423
	Annotations for evidence IEA	1,978,576
	Annotations for evidence EXP	759,654
	Annotations for evidence OTHER	791,743
	Annotations for evidence ND	241,978
	Annotations for evidence HTP	40,265
	Number of annotated scientific publications	159,963

V8 Processing of Biological Data WS 2021/22 27

http://geneontology.org/stats.html

This statistics was taken from the Gene Ontology website and refers to the release of June 2020.



The link

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

provides detailed further information about each „inferred from“ code.

Experimental evidence codes are the strongest informations because the evidence is taken from direct experimental assays of this particular gene in this organism.

Gene Ontology: Phylogenetically-inferred annotations

Phylogenetic principles, reconstructing evolutionary events to infer relationships among genes, provide a powerful way to gain insight into gene function.

Phylogenetically-based annotations are derived from an explicit model of gain and loss of gene function at specific branches in a phylogenetic tree.

Each inferred annotation can be traced to the direct experimental annotations that were used as the basis for that assertion.

Inferred from Biological aspect of Ancestor (IBA)
Inferred from Biological aspect of Descendant (IBD)
Inferred from Key Residues (IKR)
Inferred from Rapid Divergence (IRD)

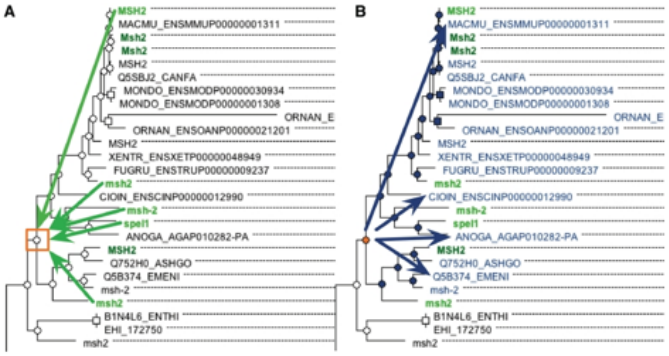
A curation tool, Phylogenetic Annotation and INference Tool (**PAINT**) helps curators to infer annotations among members of a protein family.

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

Phylogeny-based annotations make up an important part of all GO annotations.

On the next slides, we will discuss a few examples how how the PAINT tool is used to decide on phylogeny-based annotations.

Gene Ontology: PAINT

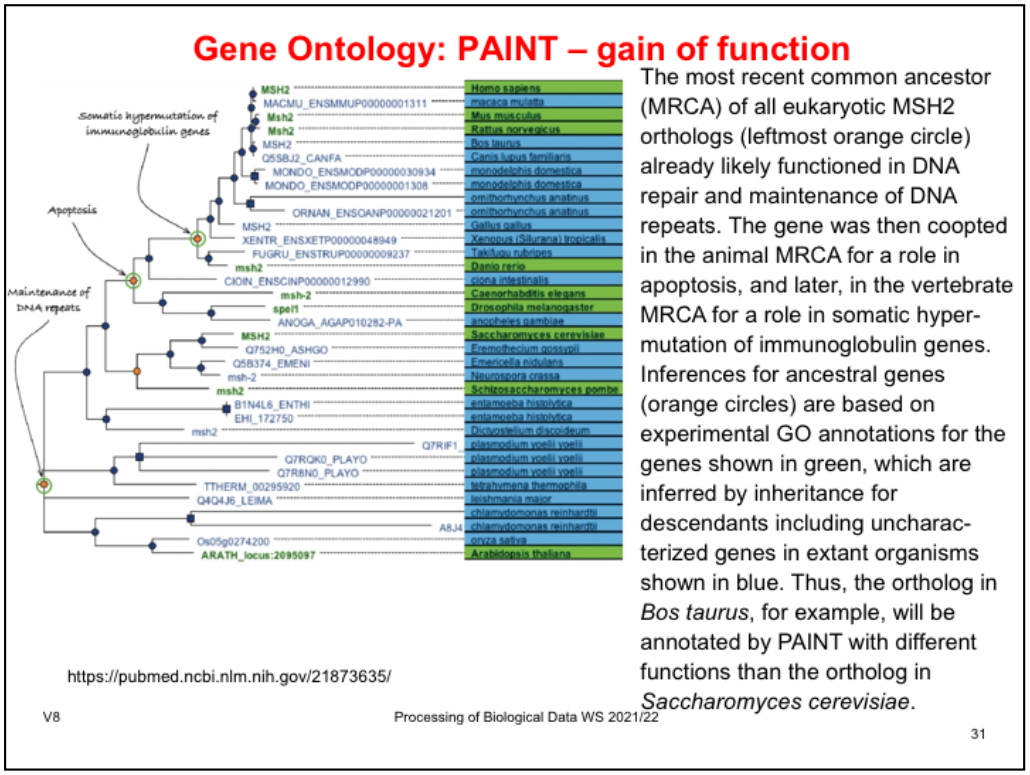


This example presents a MutS homolog family showing experimental evidence for 'GO term'. (**A**) Primary experimentally based annotations to one term or any of its ancestors (light green labels) are used to infer that the most recent common ancestor (CA) of the all those proteins also had that function. The curator notes this by dragging the term onto the node of the MCRA (orange box).
(**B**) Subsequently, PAINT propagated this annotation forward to other descendant leaves (blue labels).

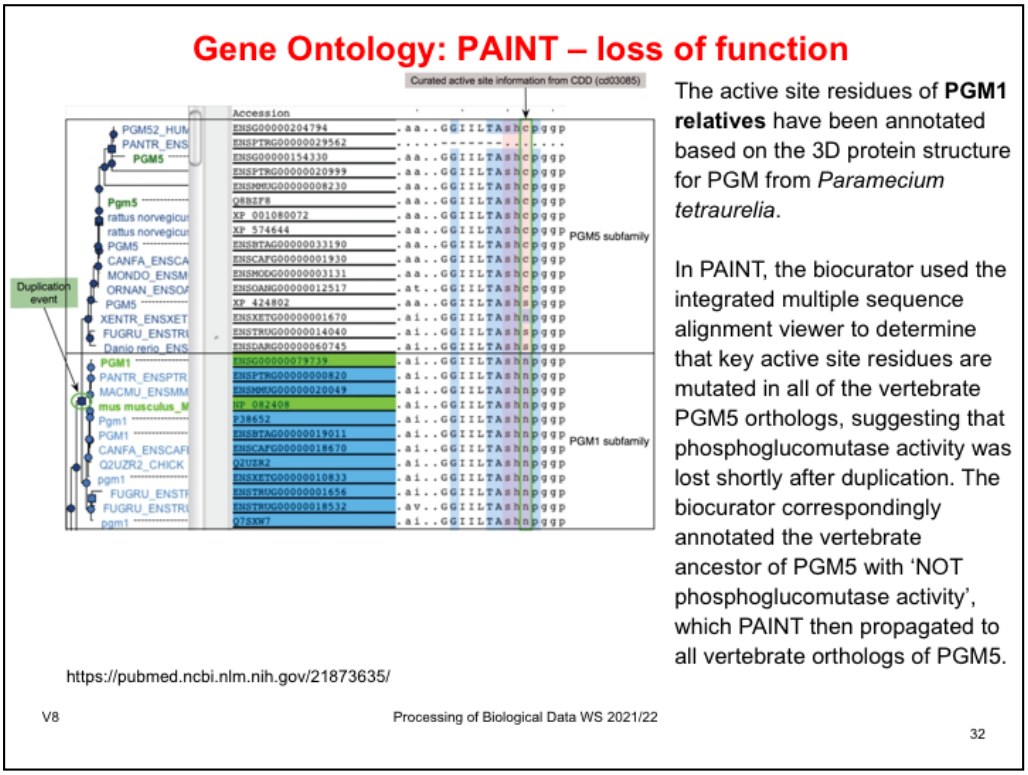
https://pubmed.ncbi.nlm.nih.gov/21873635/

Publication on PAINT: https://pubmed.ncbi.nlm.nih.gov/21873635/

The first element necessary for PAINT curation is the generation of phylogenetic trees to be annotated with functional evolution events. PAINT presents the biocurator with a phylogenetic tree and a multiple sequence alignment dynamically retrieved from the PANTHER database, and auxiliary information such as gene and protein names and identifiers. In addition it displays all the experimentally based annotations dynamically retrieved from the live GO database.

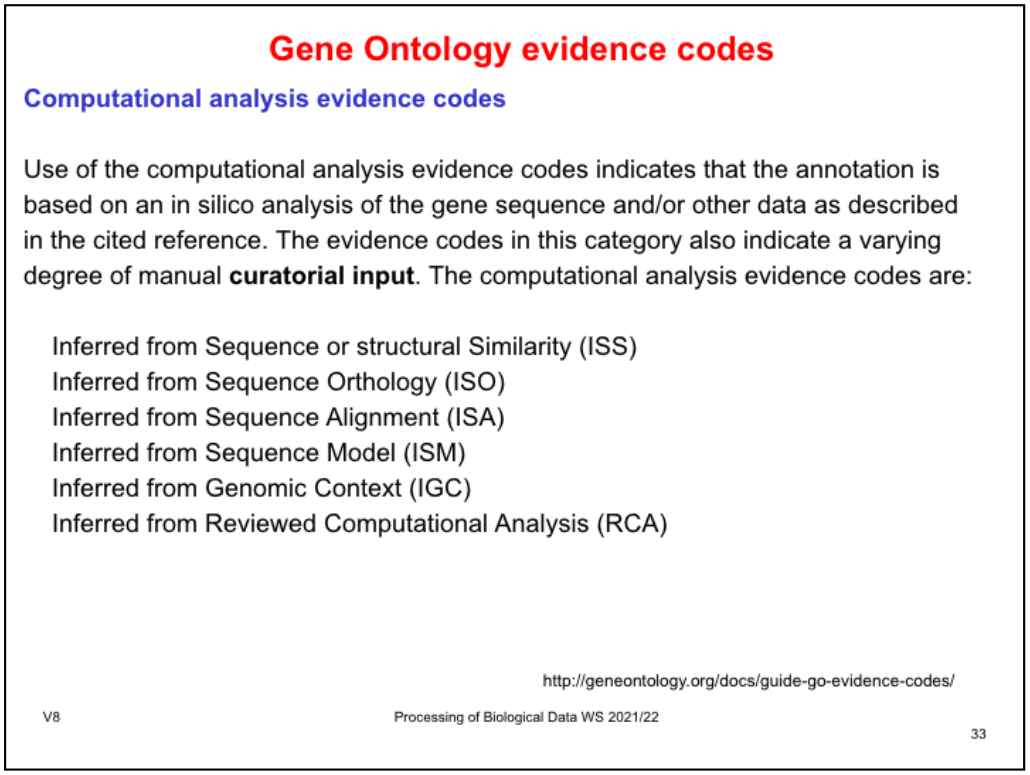


A gain of function is the addition of a function to a protein, while retaining its other existing functions. In PAINT, a biocurator is presented with all of the experiment-based GO annotations for the genes in a given family. For each annotation, the curator infers when in the evolutionary history of the family a given function was most likely to have first evolved, i.e. which ancestor 'gained' the function. This is recorded as an annotation of a gene at an internal node in the phylogenetic tree and means that the function is inferred to have evolved along the branch leading to that gene. The location of the inferred annotation determines the possible 'phylogenetic span' of the inferred annotations, since only direct descendants of the annotated ancestral gene can inherit that annotation. Gain of function may occur after a speciation event, meaning that orthologous genes will not share all functions in common. One example occurs in the MSH2 subfamily of PTHR11361, where a gene originally involved in recognizing DNA mismatches and recruiting the DNA repair machinery was co-opted in animals to regulate apoptosis and in vertebrates to mediate somatic hypermutation of immunoglobulin genes



When a biological characteristic was lost during evolution, GO annotates an ancestral (or extant) gene with the 'NOT' qualifier prefixed to the relevant annotation. 'NOT' annotations are inherited by descendants just like other GO annotations, in addition to preventing the inheritance of the corresponding positive annotation. 'NOT' annotations of ancestral genes must be supported by evidence, either: (i) an experiment-based annotation of a descendant sequence indicating it lacks this function; or (ii) absence of specific residues in the sequence, e.g. a missing active site residue.

In this example, loss of function can be observed in the phosphoglucomutase (PGM) family. Based on the phylogeny and experimental annotations, phosphoglucomutase activity most likely evolved prior to the last universal common ancestor and is found in most eubacteria and eukaryotes. A gene duplication event in the vertebrate ancestor in this family resulted in two genes that would become PGM1 and PGM5 in humans. Both mouse and human PGM5 have been demonstrated experimentally to have lost phosphoglucomutase activity. These experimental annotations strongly suggest that the loss occurred before the mouse–human common ancestor, but how long before? Based on active site mutations present in almost all of the vertebrate PGM5 proteins, the biocurator determined that the loss of function occurred in the vertebrate common ancestor. Obviously, curators must go deeply into the specific biology of this gene, its function, and its phylogeny.



Also „computational“ analysis requires the manual activity of a curator. An ISS annotation is often based on more than just one type of sequence-based evidence and may involve searches with **BLAST, profile HMMs, TMHMM, SignalP, PROSITE, InterPro**, etc. Evaluation of output from these search tools leads an annotator to a particular ISS annotation for a particular protein.

E.g., a BLAST search might reveal that a query protein matches an experimentally characterized protein from another species at 50% identity over the full lengths of both proteins. After reading literature about the match protein, the curator sees that the match protein is known to contain a domain located in the plasma membrane and another domain that extends into the cytoplasm. It is also known from the literature that the experimentally characterized match protein requires the binding of ATP to function. TMHMM analysis of the query protein predicts several membrane spanning regions in one half of the protein. In addition there are PROSITE and Pfam results which reveal the presence of an ATP-binding domain in the other half of the protein which TMHMM predicts to be cytoplasmic. These four search results taken together point to a probable identification of the query protein as having the function of the match protein.

Gene Ontology evidence codes: electronic annotations

'Electronic' (IEA) annotation are not manually reviewed. IEA-supported annotations are ultimately based on either homology and/or other experimental or sequence information, but cannot generally be traced to an experimental source.

Three methods make up the bulk of these annotations.

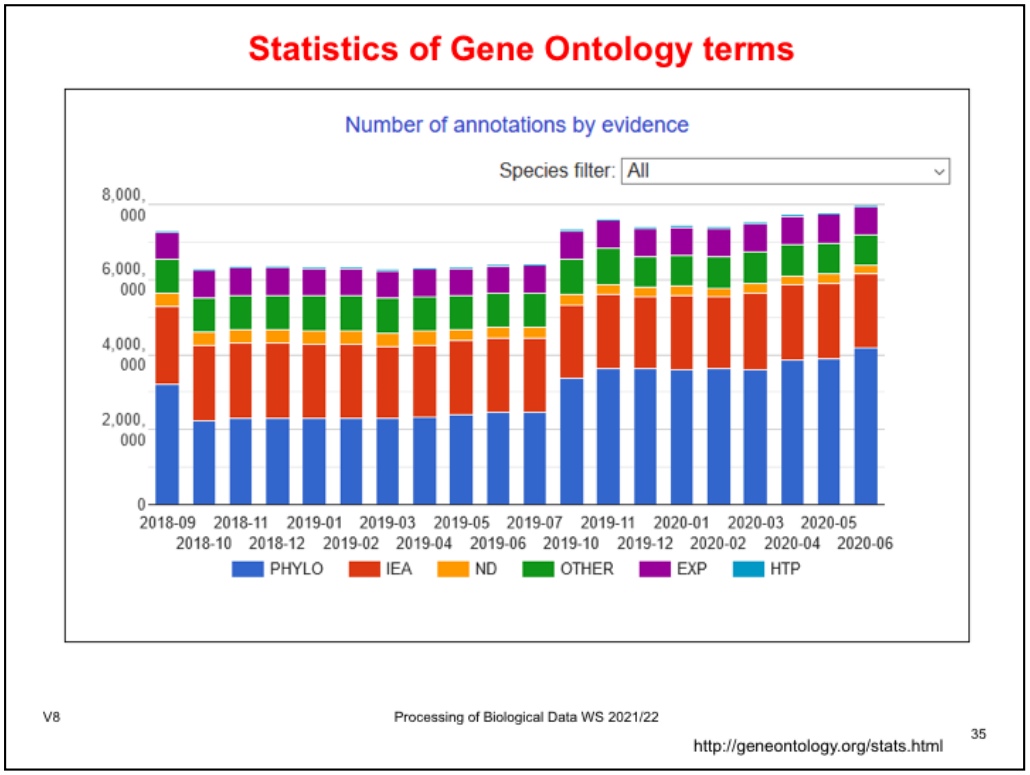
(1) InterPro2GO is based on the curated association of a GO term with a generalized sequence model ('**signature**') of a group of homologous proteins. Protein sequences with a statistically significant match to a signature are assigned the GO terms associated with the signature, a form of homology inference.

(2) computational conversion of **UniProt controlled vocabulary** terms (including Enzyme Commission numbers describing enzymatic activities, and UniProt keywords describing subcellular locations), to associated GO terms.

(3) annotations are made based on 1:1 orthologs inferred from **Ensembl gene trees**, an approach which automatically transfers annotations found experimentally in one gene, to its 1:1 orthologs in the same taxonomic clade (e.g. those within the vertebrate clade, and separately, those within the plant clade).

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

Electronically inferred annotations are the „weakest" functional annotations in GO. Still, they are based on careful methodological considerations.



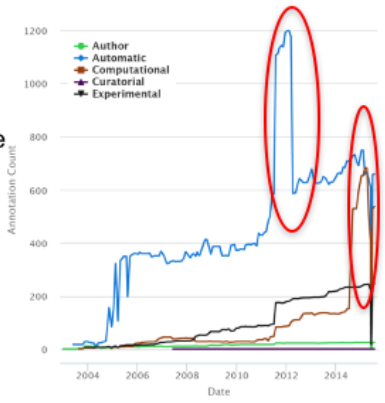
Statistics of the number of GO terms over the past 2 years taken from the listed GO website. The number of experimental annotations is growing very slowly. The largest changes are due to modifications in the PHYLO algorithm (blue).

GO annotations are dynamic in time

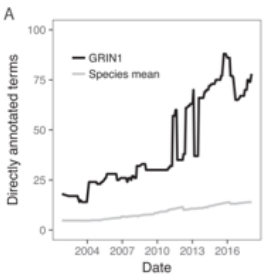
Example: strong and sudden variation in the number of annotations with the GO term "ATPase activity" over time.

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.



Bottom: Number of terms directly annotated to the human gene GRIN1. Large drops and rises are observed superimposed over a general gradual increase in annotation since 2002 (black).



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6113503/
Gaudet, Dessimoz,
Gene Ontology: Pitfalls, Biases, Remedies
https://link.springer.com/protocol/10.1007%2F978-1-4939-3743-1_14

First of all, the number of genes with annotation „ATPase activity" increases constantly over time.

There are 2 problematic cases of up/down jumps: in the blue curve and in the brown curve.

The blue curve suddenly jumped up near 2012. The reason for this is unclear – maybe a change of the underlying algorithm was made, that was later corrected – and then the curve jumped back.

A similar case is visible in the brown curve for „computational" annotations.

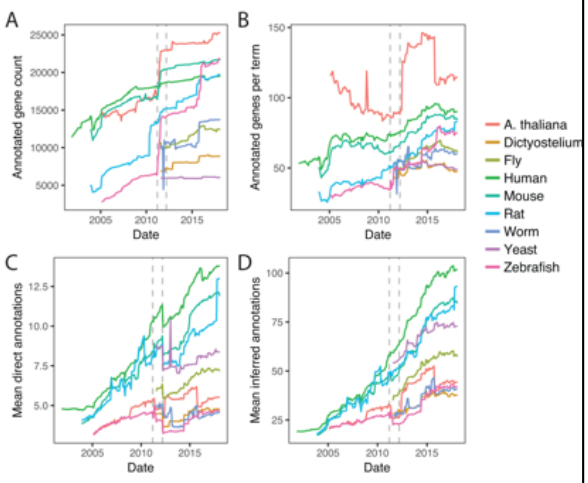
Taxon-wide GO annotation statistics

(A) Number of annotated genes.

(B) Mean annotations per term (inferred + direct).

(C) Mean number of direct annotations per gene.

(D) Mean number of inferred (including direct) annotations per gene. Times of prominent discontinuities affecting multiple species in A and C are marked by dashed gray lines in all four panels.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6113503/

Large jumps and drops are sometimes simultaneously observed in multiple, or even all, species. E.g. a rapid increase in the number of annotated genes started in March 2011 for *Arabidopsis*, mouse, and zebrafish (A). Another dramatic event was a large drop in the mean number of direct annotations per gene in March 2012 for all species (C). The jump is not visible in the plots for indirect annotations (D). This would be consistent with a large-scale purging of redundant annotations (rejecting higher-level terms that are inferable from more specific terms).

Changes to GO terms are recorded: GO:0006915 apoptotic process

Change Log

All changes | Term | Definition / Synonyms | Relationships | Cross-references | Other

Timestamp	Action	Category	Detail
2020-02-28	Deleted	XREF	MIPS_funcat:40.10.02
2019-05-04	Added	XREF	MIPS_funcat:40.10.02
2017-07-28	Added	SLIM	goslim_pombe
2016-03-05	Added	SYNONYM	caspase-dependent programmed cell death
2015-12-09	Added	CONSTRAINT	only_in_taxon NCBITaxon:33154 (Opisthokonta)
2014-04-12	Deleted	DEFINITION	A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signaling pathway phase, and an execution phase, which is triggered by the former.
2014-04-12	Added	DEFINITION	A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathway phase) which trigger an execution phase. The execution phase is the last step of an apoptotic process, and is typically characterized by rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. When the execution phase is completed, the cell has died.
2013-09-06	Added	SECONDARY	GO:0006917 (induction of apoptosis)
2013-09-06	Added	SYNONYM	induction of apoptosis
2013-09-06	Added	SYNONYM	commitment to apoptosis

https://www.ebi.ac.uk/QuickGO/term/GO:0006915

GO carefully logs all changes made to GO terms over time at the end of each QuickGO entry.

QuickGO is a web-based browser of the Gene Ontology and Gene Ontology annotation data.

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

Before, we introduced the structure of the Gene Ontology and how one can identify significantly enriched GO terms. Sofar, we dealt with individual GO terms.

Now, we will discuss how one can compare different GO terms by a numerical measure.

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?

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

Here, one counts all genes annotated with t and their child nodes.

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.

PhD Dissertation Andreas Schlicker (UdS, 2010)
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712090/

Term **information content** (IC) approaches can be divided into two families: annotation and topology-based IC approaches. The definition of p_{anno} shown here belongs to the annotation-based approaches.

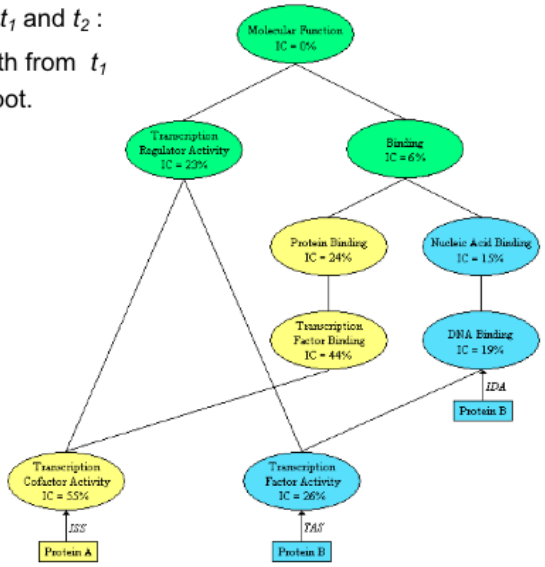
Common ancestors of GO terms

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 und t_2 is their common ancestor with highest information content.

Typically, this is also the closest common ancestor.

In this example, the MICA of the terms 'Transcription Factor Activity' and 'Transcription Cofactor Activity' is the term 'Transcription Regulator Activity', since it has a higher IC than all other common ancestors (terms in green).



One way of assigning semantic similarity between GO terms is to consider the common ancestors of 2 GO terms. Intuitively, the „closest" common ancestor would be most meaningful.

Due to the DAG-nature of the Gene Ontology, there may be multiple „closest" common ancestors either on the same hierarchical GO level or with the same path length to them.

Instead, one often selects the common ancestor with the highest information content (IC). This is called the **most informative common ancestor**.

Measure functional similarity of GO terms

Lin *et al.* defined the **similarity** of two GO terms t_1 und 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

PhD Dissertation Andreas Schlicker (UdS, 2010)

One normalizes the IC of the MICA by the sum of the ICs of the two GO terms.

Because one is taking the ratio of 1 node attribute over 2 node attributes, one multiplies this ratio by 2 to bring numerator and denominator on the same level.

At most, this ratio can reach a value of 1 if IC(MICA) = IC (t_1) = IC (t_2).

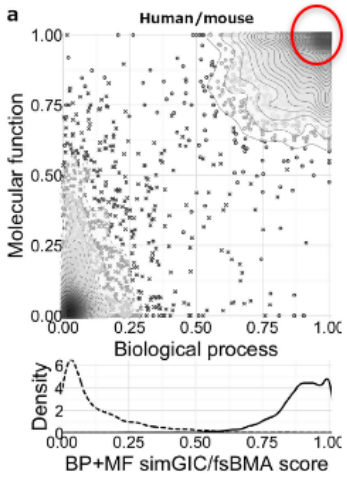
Optimal functional similarity score

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*

Link to the paper: https://www.nature.com/articles/s41598-017-00465-5

Any two genes will have a certain **semantic similarity**, even if they „have nothing to do with eachother". What is a good threshold to distinguish „real" functional similarity from the similarity of random gene pairs?

Here, the authors did a large-scale comparison of gene pairs from human and mouse. Orthologous gene pairs (circles) have high BP and MF functional similarity and are placed in the upper right quadrant.

Random gene pairs are in the bottom left quadrant. Shown in the bottom panel is a combined BP + MF similarity score. Here, the best separation point would be around 0.55 or so.

Optimal functional similarity score

Comment:

The human/mouse comparison is based somehow 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.

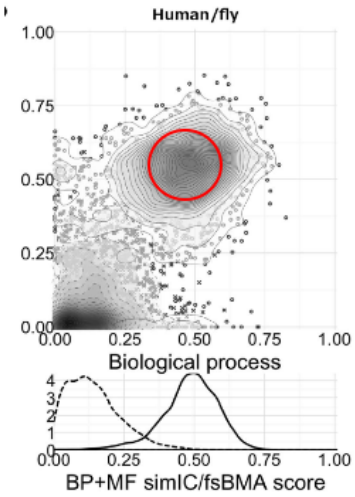
Weichenberger et al. (2017)
Scientific Reports 7: *381*

We should not forget where GO terms come from. This may sometimes lead to circular arguments.

Optimal functional similarity score

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

-> More overlap than for human/mouse because real orthologues have smaller similarity (red circle centered at 0.5, not at 1.0).



Weichenberger et al. (2017)
Scientific Reports 7: *381*

For the more remotely related organism pair human/fly, the densities for cases and controls calculated with the *simIC/fsBMA* measures overlap to some extent. Notably, there is a smaller fraction of orthologues that do not share any similarity in the MF ontology, but do have considerable high BP scores

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.

Additional slides (not used)

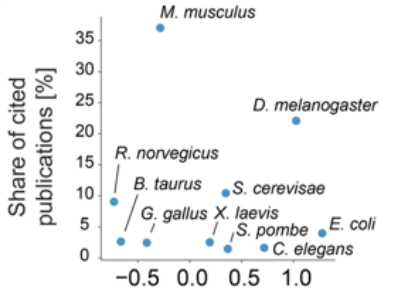
Studies on model organisms affect studies on human genes

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.

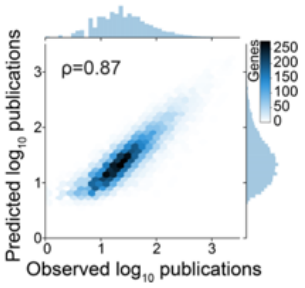


Fraction of nonhuman organisms cited by initial publications of human genes. Enrichment represents log2 ratio of the fraction of nonhuman organisms among all initial publications on human genes over the fraction of nonhuman organisms among initial publications on human genes, which also cite publications on human genes.
The 10 most cited organisms are shown

V8 Stoeger et al. (2018) PLoS Biol 16(9): e2006643.

Human genes <–> 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).



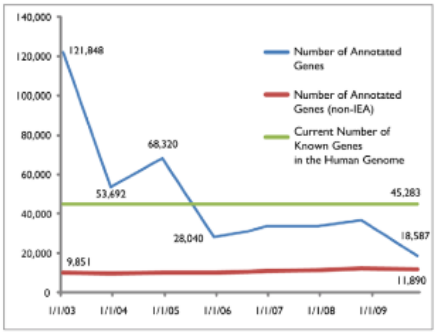
- The number of publications on homologous genes yielded almost perfect predictions of the number of publications for individual human genes (Spearman: 0.87).

- Human-specific genes without homologous genes remain significantly less studied (p-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.

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.

V8 Khatri et al. (2012) PLoS Comput Biol 8: *e1002375* Processing of Biological Data WS 2021/22 50

IEA stands for „inferred by electronic annotation“.

Non-IEA annotations are considered more trustworthy than IEA.

Mixing rules

Given:

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

Then the matrix M is given by all possible pairwise semantic similarity (SS) values $s_{ij} = sim(t_i, r_j)$ with *sim* being one of the SS measures introduced above, $i = 1, 2, .., m$ and $j = 1, 2, .., n$.

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)
Scientific Reports 7: *381*

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)
Scientific Reports 7: *381*

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}):

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

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

Weichenberger et al. (2017)
Scientific Reports 7: *381*

Effect of high-throughput experiments

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

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

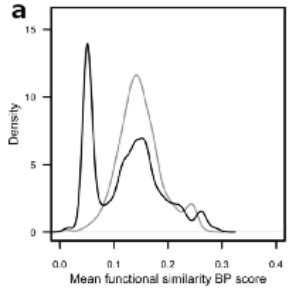
This information is further propagated by automated methods.

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

Weichenberger et al. (2017)
Scientific Reports 7: *381*

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 → 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

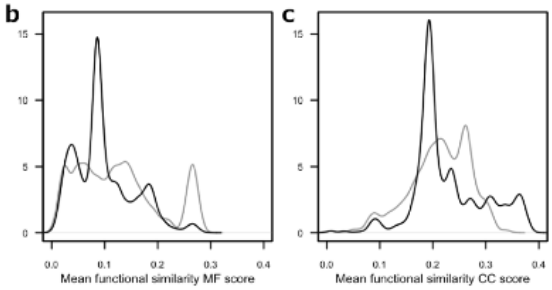
Weichenberger et al. (2017)
Scientific Reports 7: *381*

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, $p < 10^{-100}$).

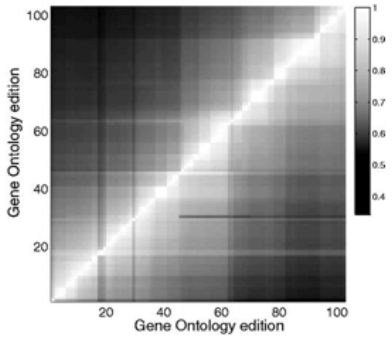
This suggests that gene products annotated to this term generally yield much higher than average *simLin*/*fsAvg* MF scores.



(**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*

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.

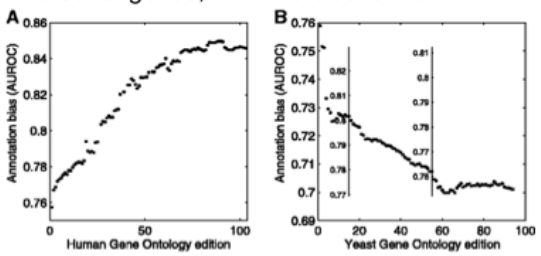
This means that, each month, the annotations of about 3% of the genes have changed so substantially that they are not functionally 'the same genes' anymore.

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.

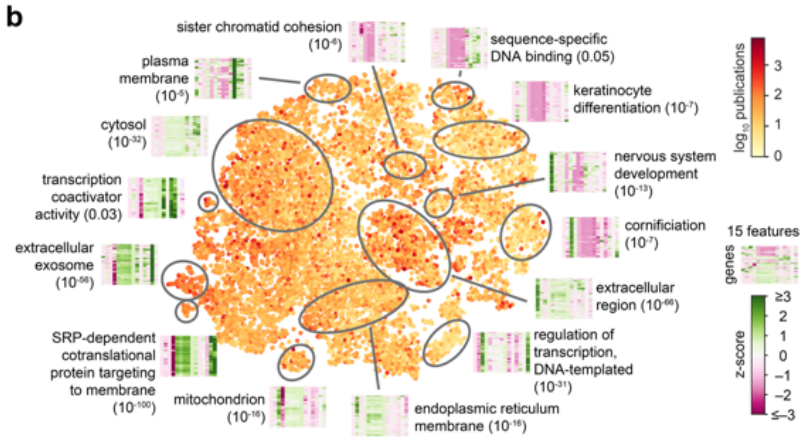
(**B**) For yeast, annotation bias has generally fallen over time.

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

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.

Red dots are genes studied in many publications, yellow dots are little studied.