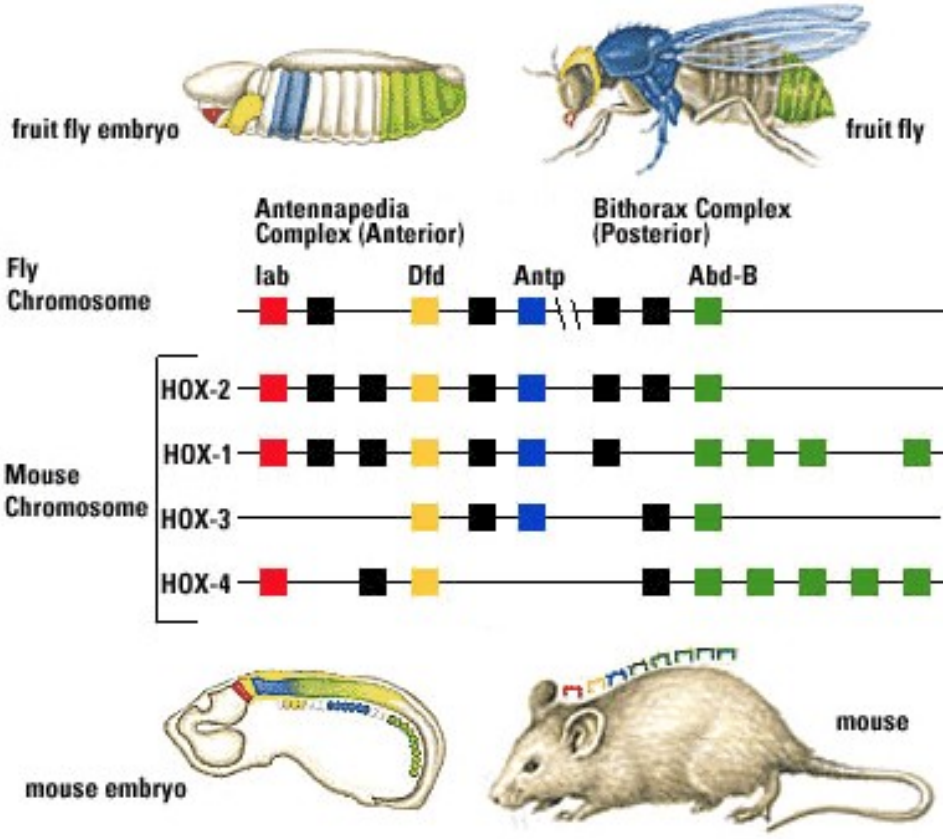
WBO 10

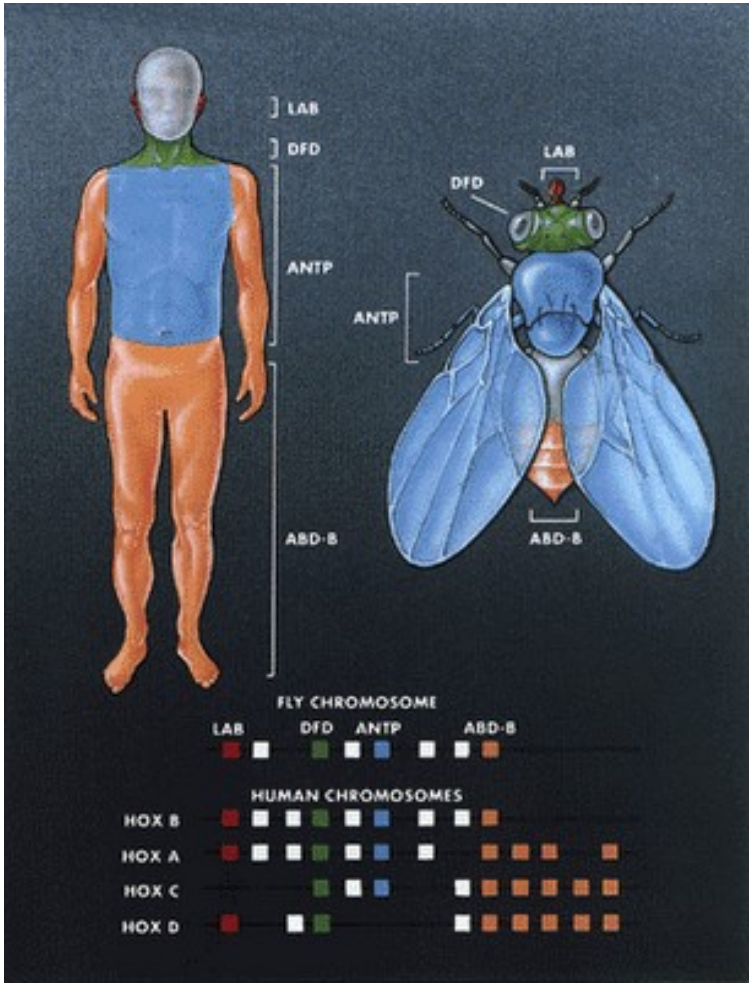
Gene families and
Gene function annotation

May, 5th 2020

Bartek Wilczyński

Many genes have conserved sequence and function over milions of years of evolution





How did the gene families originate

- As we discussed earlier, we cannot answer questions about the pre-cellular origins of life
- We know that the simplest autnomous cellular life forms have ~1000 protein coding gene and some parasitic life forms can function with ~500 genes
- How do we arrive at a certain number of genes and do they evolve intependently?
- What processess lead to evolution of organisms with tens of thousands of genes with hundreds of thousands gene products.

Duplication and speciation

- As we discussed earlier, some events during DNA replication (e.g. non-homologous recombination) can create duplicates of genes or gene clusters
- If the duplication of a gene did not lead to deleterious effects and the resulting organism is viable, there is a lesser selective pressure on maintaining the function of each of the gene copies, i.e. one of them can randomly mutate to potentially perform a different function, while the other remains to perform the original function

Specialization after duplication

- Since one copy of a gene was enough before duplication, there is now an additional gene copy, that can acquire mutations leading to potential for a more specialized function (**subfunctionalization**) or a new function (**neofunctionalization**)
- If the mutations are deleterious, the gene becomes non-functional **pseudogene**
- If the mutated gene provides some new function, this can contribute to a speciation event – organisms with a single copy might not be able to cross

Evolution of duplicated genes

- Since selection of individuals is not happening on the genotype, but on the phenotype level, many genes tolerate changing the number of gene copies
- However, there are some genes, called **dosage sensitive genes**, that have their function dependent on the exact dosage of the gene product, and introducing another copy (or deleting one) changes the amount of gene product, influencing the phenotype

What is a gene's ***function***?

- We are using the word *function* here in a very informal way
- It basically means any ablility of a gene product (RNA or protein) to change the behavior of the cell containing this gene versus the one without it
- Usually this may mean taking part in a chemical reaction or some other molecular interaction
- These are usually non-unique (i.e. many functions for a single gene) and somewhat arbitrary, as it is difficult to prove exactly what the function is

Gene homologs

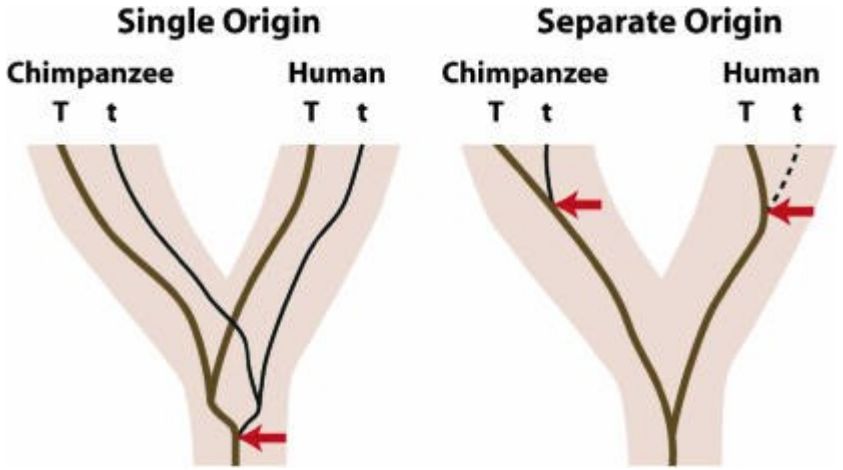
- **Genes with a common ancestral sequence are called homologous, or homologs**. Theoretically, this *would be an equivalence relation*, however, the number of classes of this relation depends on how far back we look.
- In practive, **genes with a certain level of sequence similarity,** making the probability of their independent evolution extremely small, to be homologs. It is not an equivalence relationship, as the sequence similarity is not transitive.
- Sometimes people distinguish this second raltion as „sequence homology” as opposed to „true homology”

Types of homologs

- Homologous genes originated from gene duplication are called **paralogs**
- Homologous genes originated from speciation are called **orthologs**
- Lesser frequently used types of homologs are *xenologs* – originating from horizontal gene transfers, and *ohnologs* – originating from whole genome duplication

The timing of evolutionary events makes a difference

- We have very different expectations from paralogs and orthologs
  - Evolving orthologs usually maintain their function
  - Paralogs are much more prone to neofunctionalization

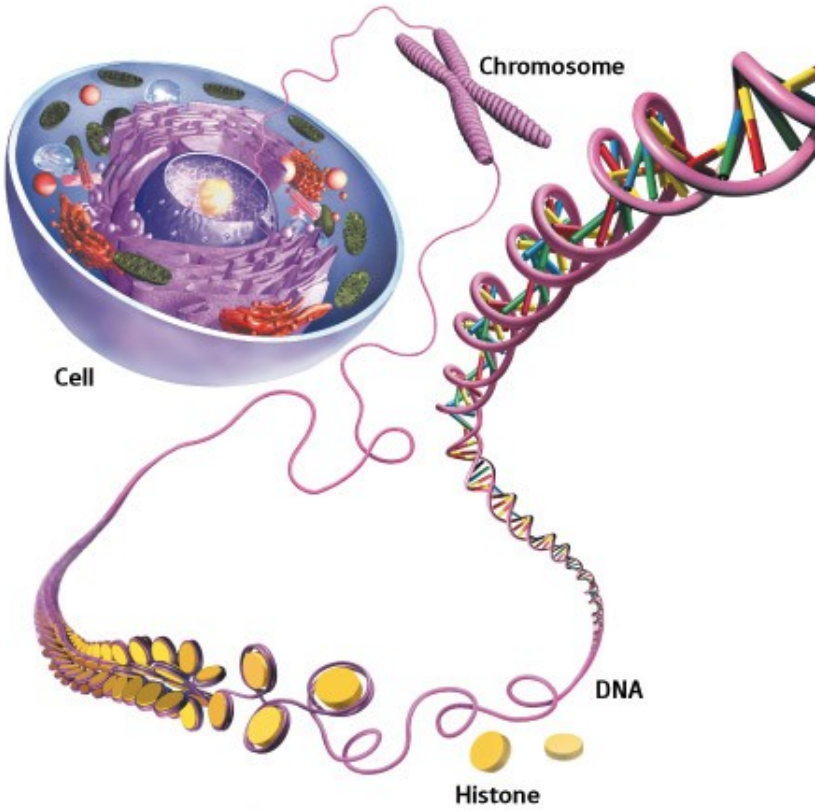


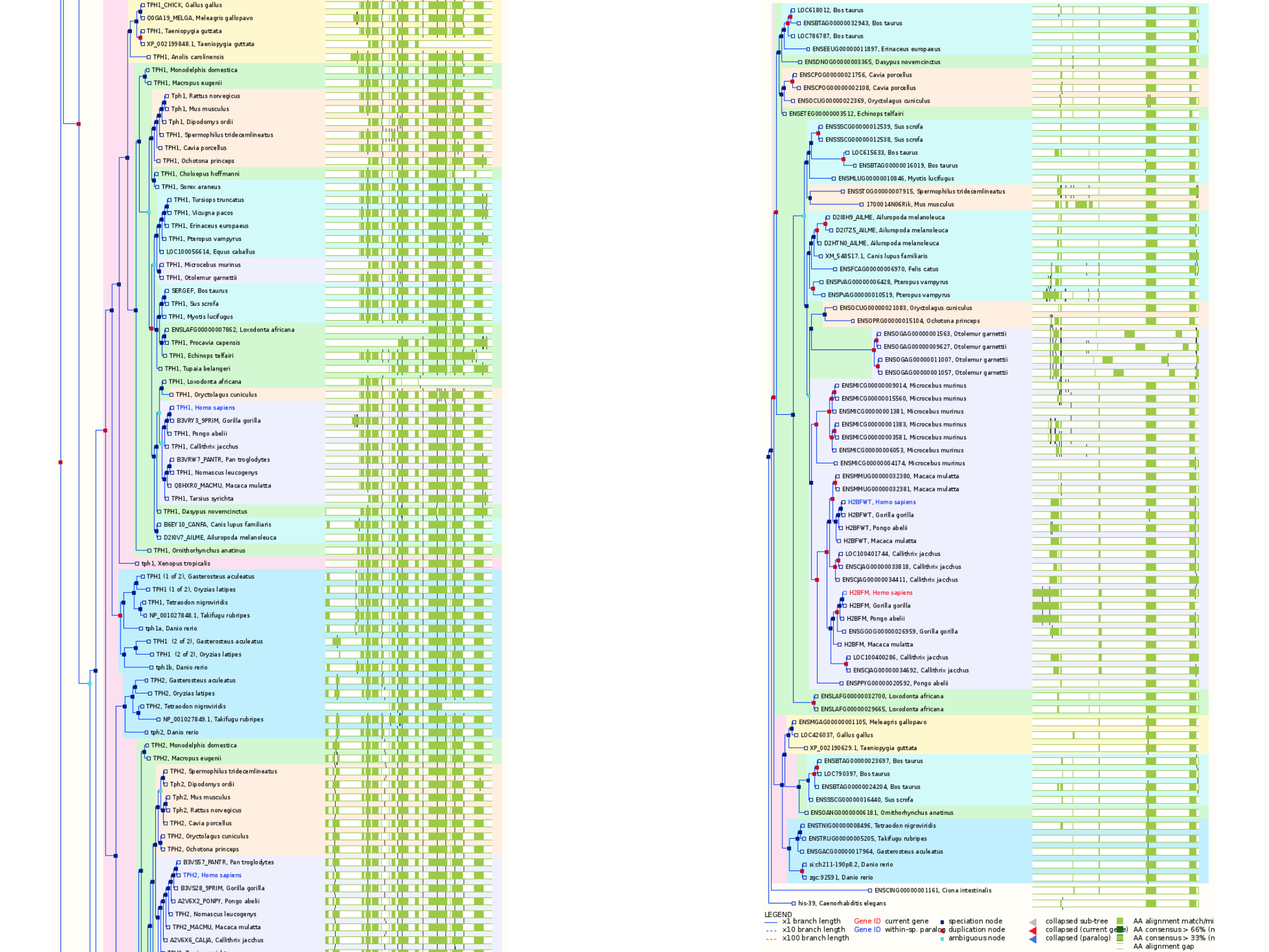
Two examples – PAH and TPH

- PAH – phenylalanine hydroxylase, a very well known and old enzyme, that is almost always existing in a single copy - a gene that has a very well defined function,
- mutations in it cause a genetic disease called phenylketonuria
- A paralog of PAH is TPH – tryptophan hydroxylase, not a dosage sensitive gene, exists in multiple copies in many genomes

Example 2 - histones

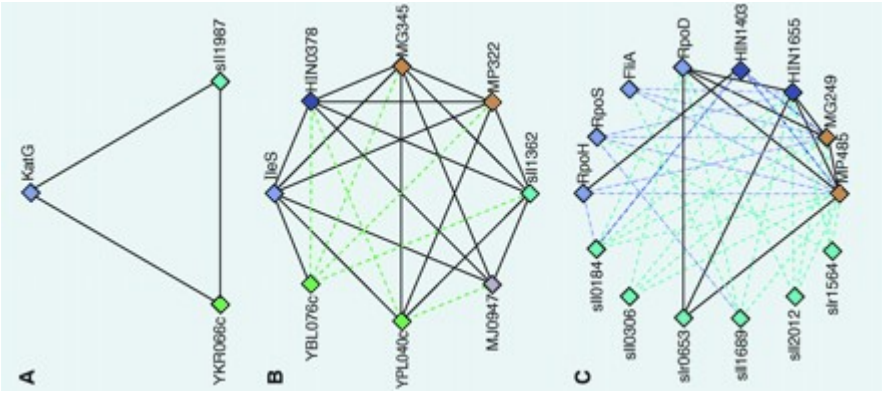
- Histones are the proteins that are combined to create nucleosomes that are used to organized chromatin.
- Every eukaryotic cell needs milions of nucleosomes, so they exist in many copies in the genome, with some functionalization (H2A, H2B, H3, H4 and linker histone H1)




TPH1_CHICK, Gallus gallus
Q0GA19_MELGA, Meleagris gallopavo
TPH1, Taeniopygia guttata
XP_002199848.1, Taeniopygia guttata
TPH1, Anolis carolinensis
TPH1, Monodelphis domestica
TPH1, Macropus eugenii
Tph1, Rattus norvegicus
Tph1, Mus musculus
Tph1, Dipodomys ordii
TPH1, Spermophilus tridecemlineatus
TPH1, Cavia porcellus
TPH1, Ochotona princeps
TPH1, Choloepus hoffmanni
TPH1, Sorex araneus
TPH1, Tursiops truncatus
TPH1, Vicugna pacos
TPH1, Erinaceus europaeus
TPH1, Pteropus vampyrus
LOC100056614, Equus caballus
TPH1, Microcebus murinus
TPH1, Otolemur garnettii
SERGEF, Bos taurus
TPH1, Sus scrofa
TPH1, Myotis lucifugus
ENSLAFG00000007862, Loxodonta africana
TPH1, Procavia capensis
TPH1, Echinops telfairi
TPH1, Tupaia belangeri
TPH1, Loxodonta africana
TPH1, Oryctolagus cuniculus
TPH1, Homo sapiens
B3VRY3_9PRIM, Gorilla gorilla
TPH1, Pongo abelii
TPH1, Callithrix jacchus
B3VRW7_PANTR, Pan troglodytes
TPH1, Nomascus leucogenys
Q8HXR0_MACMU, Macaca mulatta
TPH1, Tarsius syrichta
TPH1, Dasypus novemcinctus
B6EY10_CANFA, Canis lupus familiaris
D2I0V7_AILME, Ailuropoda melanoleuca
TPH1, Ornithorhynchus anatinus
tph1, Xenopus tropicalis
TPH1 (1 of 2), Gasterosteus aculeatus
TPH1 (1 of 2), Oryzias latipes
TPH1, Tetraodon nigroviridis
NP_001027848.1, Takifugu rubripes
tph1a, Danio rerio
TPH1 (2 of 2), Gasterosteus aculeatus
TPH1 (2 of 2), Oryzias latipes
tph1b, Danio rerio
TPH2, Gasterosteus aculeatus
TPH2, Oryzias latipes
TPH2, Tetraodon nigroviridis
NP_001027849.1, Takifugu rubripes
tph2, Danio rerio
TPH2, Monodelphis domestica
TPH2, Macropus eugenii
TPH2, Spermophilus tridecemlineatus
Tph2, Dipodomys ordii
Tph2, Mus musculus
Tph2, Rattus norvegicus
TPH2, Cavia porcellus
TPH2, Oryctolagus cuniculus
TPH2, Ochotona princeps
B3VS57_PANTR, Pan troglodytes
TPH2, Homo sapiens
B3VS28_9PRIM, Gorilla gorilla
A2V6X2_PONPY, Pongo abelii
TPH2, Nomascus leucogenys
TPH2_MACMU, Macaca mulatta
A2V6X6_CALJA, Callithrix jacchus
LOC618012, Bos taurus
ENSBTAG00000032943, Bos taurus
LOC786787, Bos taurus
ENSEEUG00000011897, Erinaceus europaeus
ENSDNOG00000003365, Dasypus novemcinctus
ENSCPOG00000021756, Cavia porcellus
ENSCPOG00000002108, Cavia porcellus
ENSOCUG00000022369, Oryctolagus cuniculus
ENSETEG00000003512, Echinops telfairi
ENSSSCG00000012539, Sus scrofa
ENSSSCG00000012538, Sus scrofa
LOC615633, Bos taurus
ENSBTAG00000016019, Bos taurus
ENSMLUG00000010846, Myotis lucifugus
ENSSTOG00000007915, Spermophilus tridecemlineatus
1700014N06Rik, Mus musculus
D2I8H9_AILME, Ailuropoda melanoleuca
D2I7Z5_AILME, Ailuropoda melanoleuca
D2HTN0_AILME, Ailuropoda melanoleuca
XM_548517.1, Canis lupus familiaris
ENSFCAG00000006970, Felis catus
ENSPVAG00000006428, Pteropus vampyrus
ENSPVAG00000010519, Pteropus vampyrus
ENSOCUG00000021083, Oryctolagus cuniculus
ENSOPRG00000015104, Ochotona princeps
ENSOGAG00000001563, Otolemur garnettii
ENSOGAG00000009627, Otolemur garnettii
ENSOGAG00000011007, Otolemur garnettii
ENSOGAG00000001057, Otolemur garnettii
ENSMICG00000009014, Microcebus murinus
ENSMICG00000015560, Microcebus murinus
ENSMICG00000001381, Microcebus murinus
ENSMICG00000001383, Microcebus murinus
ENSMICG00000003581, Microcebus murinus
ENSMICG00000006053, Microcebus murinus
ENSMICG00000004174, Microcebus murinus
ENSMMUG00000032380, Macaca mulatta
ENSMMUG00000032381, Macaca mulatta
H2BFWT, Homo sapiens
H2BFWT, Gorilla gorilla
H2BFWT, Pongo abelii
H2BFWT, Macaca mulatta
LOC100401744, Callithrix jacchus
ENSCJAG00000033818, Callithrix jacchus
ENSCJAG00000034411, Callithrix jacchus
H2BFM, Homo sapiens
H2BFM, Gorilla gorilla
H2BFM, Pongo abelii
ENSGGOG00000026959, Gorilla gorilla
H2BFM, Macaca mulatta
LOC100400286, Callithrix jacchus
ENSCJAG00000034692, Callithrix jacchus
ENSPPYG00000020592, Pongo abelii
ENSLAFG00000032700, Loxodonta africana
ENSLAFG00000029665, Loxodonta africana
ENSMGAG00000001105, Meleagris gallopavo
LOC426037, Gallus gallus
XP_002190629.1, Taeniopygia guttata
ENSBTAG00000023697, Bos taurus
LOC790397, Bos taurus
ENSBTAG00000024204, Bos taurus
ENSSSCG00000016440, Sus scrofa
ENSOANG00000006181, Ornithorhynchus anatinus
ENSTNIG00000008496, Tetraodon nigroviridis
ENSTRUG00000005205, Takifugu rubripes
ENSGACG00000017964, Gasterosteus aculeatus
si:ch211-190p8.2, Danio rerio
zgc:92591, Danio rerio
ENSCING00000001161, Ciona intestinalis
his-39, Caenorhabditis elegans
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/mi
AA consensus > 66% (n
AA consensus > 33% (n
AA alignment gap

Using BLAST to find homologs

- Bidirectional BLAST Best hits – proteins that are mutally most similar to each other in two genomes
- We can construct a graph where any two genes from different genoms are connected if they are BBB hits
- This graph can be searched for dense structures, similar to cliques, to identify Clusters of Orthologous Genes, or COGs (Koonin 1997)



How to identify gene function?

- It is not easy to test protein functions experimentally
- It is costly, takes time and effort and is never 100% complete, and usually not 100% reliable
- For majority of genes, even in humans, we still don’t know their function
- However, We have a lot of data (in very diverse form) published on gene function of thousands of genes and need to systematize it

Different approaches to gene function annotation

- Gene Ontology – a systematic effort of manual curation of both an **ontology of function terms** (A directed acyclic graph with terms in nodes, and is_a and part_of relationships) and **annotations of genes** to this graph
- Wikigenes, genecards, etc. – many commercial or open approaches to start a knowledge-base using text-mining tools and published articles and open it to human annotation at some point in „wikipedia” style

Gene annotations - summary

- Ontology related approaches are relatively costly and slow to curate, but they are very quick and easy to automatically process
- They usually require some software to process, however, a lot of software is already available
- Wiki-style text-mining based approaches are the opposite: much cheaper to create, more difficult to automatically process, more readable for humans and less reliable factually

the Gene Ontology

Search | Browse | BLAST | Homolog Annotations | Tools & Resources | Help

Search GO [ ] (•) terms ( ) genes or proteins [ ] exact match [Submit Query]

PAH

Phenylalanine-4-hydroxylase

protein from *Homo sapiens* (human)

Term associations ↓ | Gene product information → | Peptide Sequence → | Sequence information →

Term Associations

Download all association information in: gene association format | RDF-XML

Filter associations displayed

Filter Associations

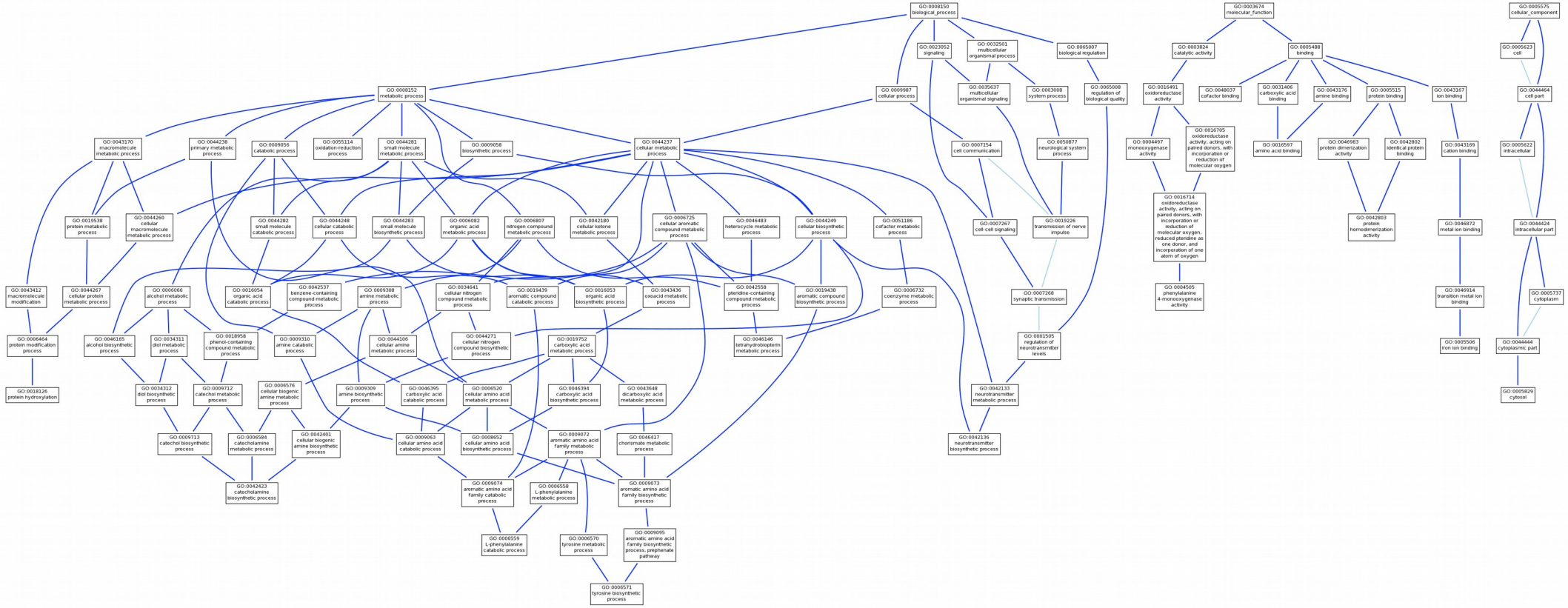
Ontology	Evidence Code
All	All
biological process	IC
cellular component	IDA
molecular function	IEA

[Set filters] [Remove all filters]

[Select all] [Clear all] [Perform an action with this page's selected terms...] [Go!]

Accession, Term		Ontology	Qualifier	Evidence
GO:0042423 : catecholamine biosynthetic process	71 gene products view in tree	biological process		NAS
GO:0008652 : cellular amino acid biosynthetic process	6303 gene products view in tree	biological process		TAS

Gene ontology graph



[edit this page]

Gene Review

PAH - phenylalanine hydroxylase

Homo sapiens

Synonyms: PH, PKU, PKU1, Phe-4-monooxygenase, Phenylalanine-4-hydroxylase

Green, E.K. et al., **Chao, H.M.** et al., **Maass, A.** et al., **Teigen, K.** et al., **Ramus, S.J.** et al., et al.

Welcome! If you are familiar with the subject of this article, you can contribute to this open access knowledge base by deleting incorrect information, restructuring or completely rewriting any text.

Ideally this entry shall become one comprehensive and continuous article. Bulleted lists, for instance, were only used because it is

Disease relevance of PAH

- We present the case of a girl affected by classical phenylketonuria who has been screened for mutations in the PAH gene [1].
- Benzo[a]pyrene-induced DNA damage and p53 modulation in human hepatoma HepG2 cells for the identification of potential biomarkers for PAH monitoring and risk assessment [2].

Testing for annotation enrichment

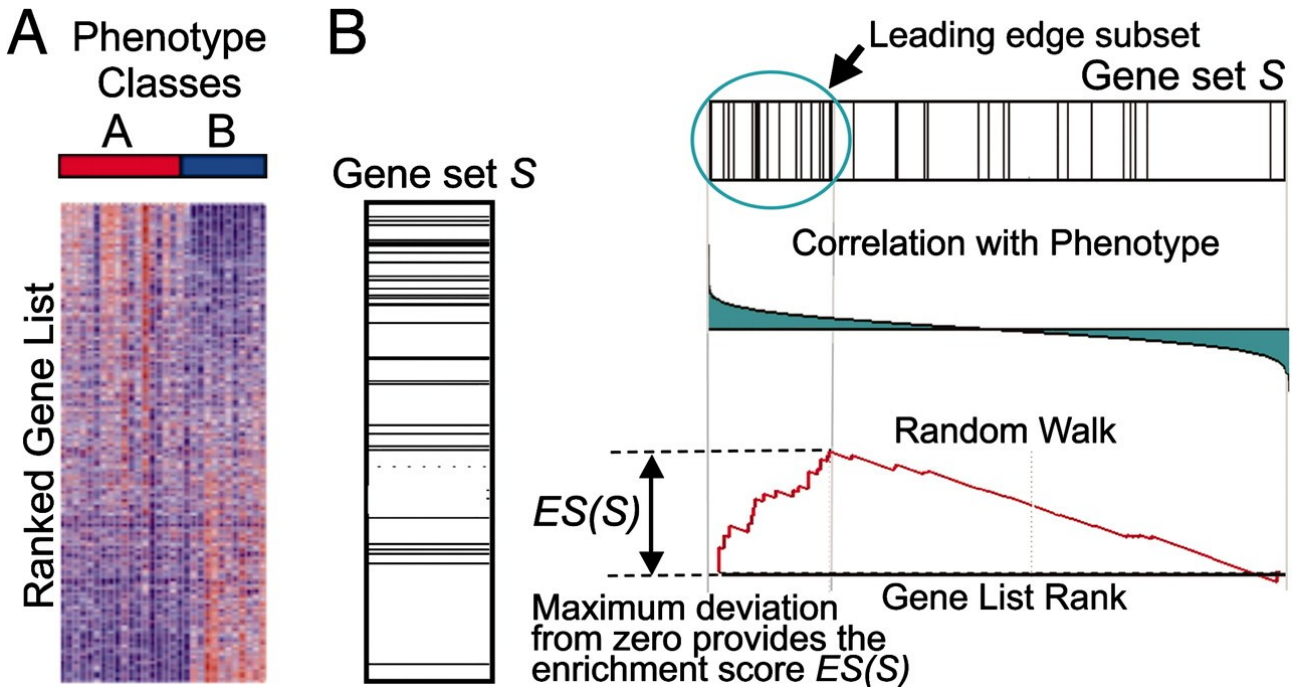
- For many sets of genes (usually originating from some experiment) we might be interested if there is a common function to them.
- Usually this comes from a set of genes changed in a disease state, developmental stage or stress condition
- Sometimes, this can be a ranking of genes from most to least „affected”, and we can ask if there is anything non random in functions of genes at both ends of the ranking.
- How to answer these questions statistically?

Significant overlap of gene sets

- The most straightforward approach is to ask if the overlap of the set of genes in question with a set of genes annotated with a given function is greater than expected by chance
- This can be answered by Fisher's exact test, that uses the hypergeometric distribution to describe the overlap distribution under the null hypothesis
- This is similar to an urn scenario: What is the probability that I draw more than x black balls while taking N balls from an urn originally containing M balls of which k were black

Gene set enrichment analysis

- Assuming that we have a ranking of genes, we can ask, similarly to the Kolmogorov-Smirnov test, if there is a significant difference in the enrichment at any point in the ranking



Mulitple hypothesis testing corrections

- Multiple hypothesis testing corrections are very important in gene function enrichment testing as they are thousands of known functions and we usually test all of them
- The simplest correction is to multiply the p-value by the number of hypotheses – the Bonferroni correction
- Usually, more useful results are obatined with False Discovery Rate – FDR, that indicates what is the percentage of false positive predictions under a given p-value threshold.