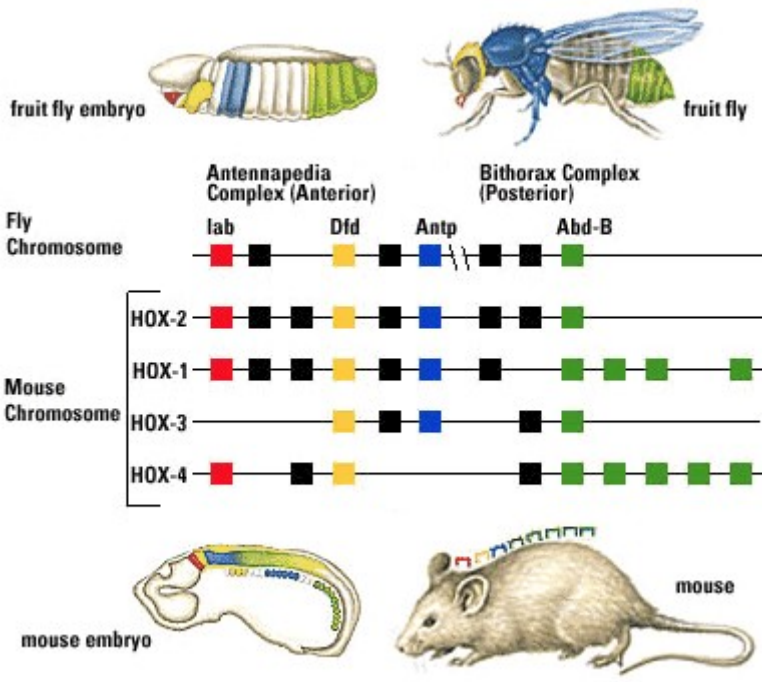
WBO 9

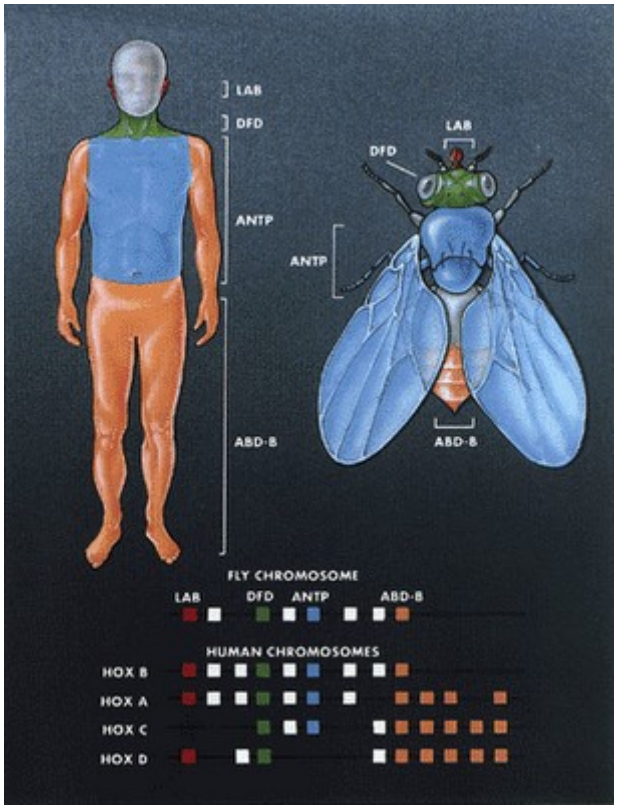
Rodziny genów
i określanie funkcji genów

8. maja 2018

Bartek Wilczyński

Wiele genów zachowuje podobną sekwencję i funkcję poprzez miliony lat ewolucji





Jak powstały rodziny genów ?

- Nie znamy dokładnie początków życia, w szczególności mamy różne hipotezy nt. Życia przed powstaniem komórek
- Wiemy, że najprostsze organizmy komórkowe mają około tysiąca genów (Mycoplasma genitalium), a niektóre pasożyty mogą funkcjonować mając nawet około 500 genów
- Podejrzewa się, że pierwsze formy komórkowe prawdopodobnie stanowią pra-przodka całego żywego świata jaki znamy
- Jakie procesy stoją za powstaniem organizmów o dużo większej liczbie genów?

Duplikacja i specjacja

- W procesie replikacji chromosomów, duże fragmenty mogą zostać zduplikowane w tej samej komórce
- Następnie procesy mutacji następują niezależnie w różnych kopiach
- Jeśli duplikacja genu nie spowodowała negatywnych efektów ubocznych, presja selekcyjna na zduplikowane geny jest słabsza

Specjalizacja po duplikacji

- Jeśli wystarczy aby tylko jedna z kopii zduplikowanego genu zachowała funkcje przodka, drugi gen może nabywać nowych funkcji w procesie mutacji
- Jeśli mutacje nie doprowadzą do powstania użytecznej funkcji, często gen przestaje być używany i pozostaje pseudogenem
- Jeśli nowy gen zyskuje nową funkcję, może przyłożyć się do procesu specjacji – powstania nowego gatunku z podpopulacji posiadającej zduplikowany gen

Zachowanie genów duplikowanych

- Selekcja osobników przebiega na podstawie fentoypu, a nie genotypu, więc często posiadanie dwóch kopii tego samego genu nie wpływa bezpośrednio na fenotyp
- Chyba, że mamy do czynienia z genem, który musi mieć ściśle regulowaną liczbę produktów (transkryptów, białek), bo zwiększenie liczby kopii DNA genu prowadzi często do zwiększonej ekspresji genu

Homologi i ich rodzaje

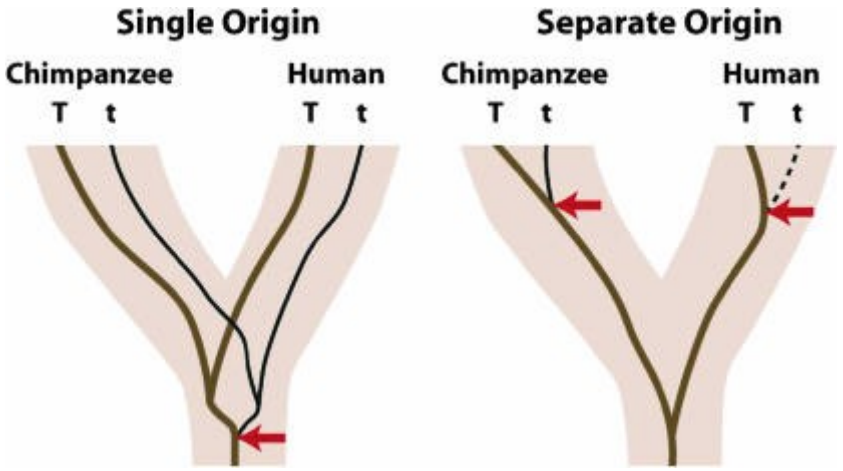
- Geny pochodzące od wspólnego przodka nazywamy homologami. Teoretycznie jest to relacja równoważności, choć nie mamy możliwości weryfikacji homologii ze 100% pewnością
- W praktyce, za homologi uznaje się geny o dużym (arbitralnie) podobieństwie sekwencyjnym, względem modeli losowych ewolucji sekwencji. To nie jest relacja równoważności, ze względu na brak przechodniości

Rodzaje homologów

- Pary homologów, pochodzące z duplikacji, nazywamy paralogami
- Pary homologów pochodzące ze specjacji, nazywamy ortologami
- Są jeszcze ksenologi – pochodzące z horyzontalnego transferu genów i ohnologi – pochodzące z duplikacji całych genomów, ale te rozważamy rzadziej.

Kolejność wydarzeń ewolucyjnych ma znaczenie

- Mamy zupełnie różne oczekiwania wobec para i ortologów jeśli chodzi o ich funkcje:
  - Ortologi powinny mieć niemal niezmienioną funkcję
  - Paralogi mogą dość istotnie odbiegać w swym działaniu

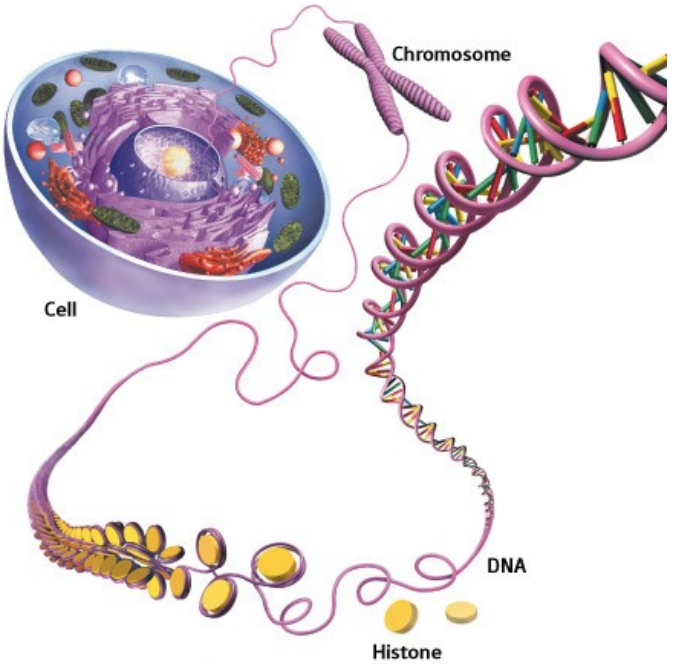


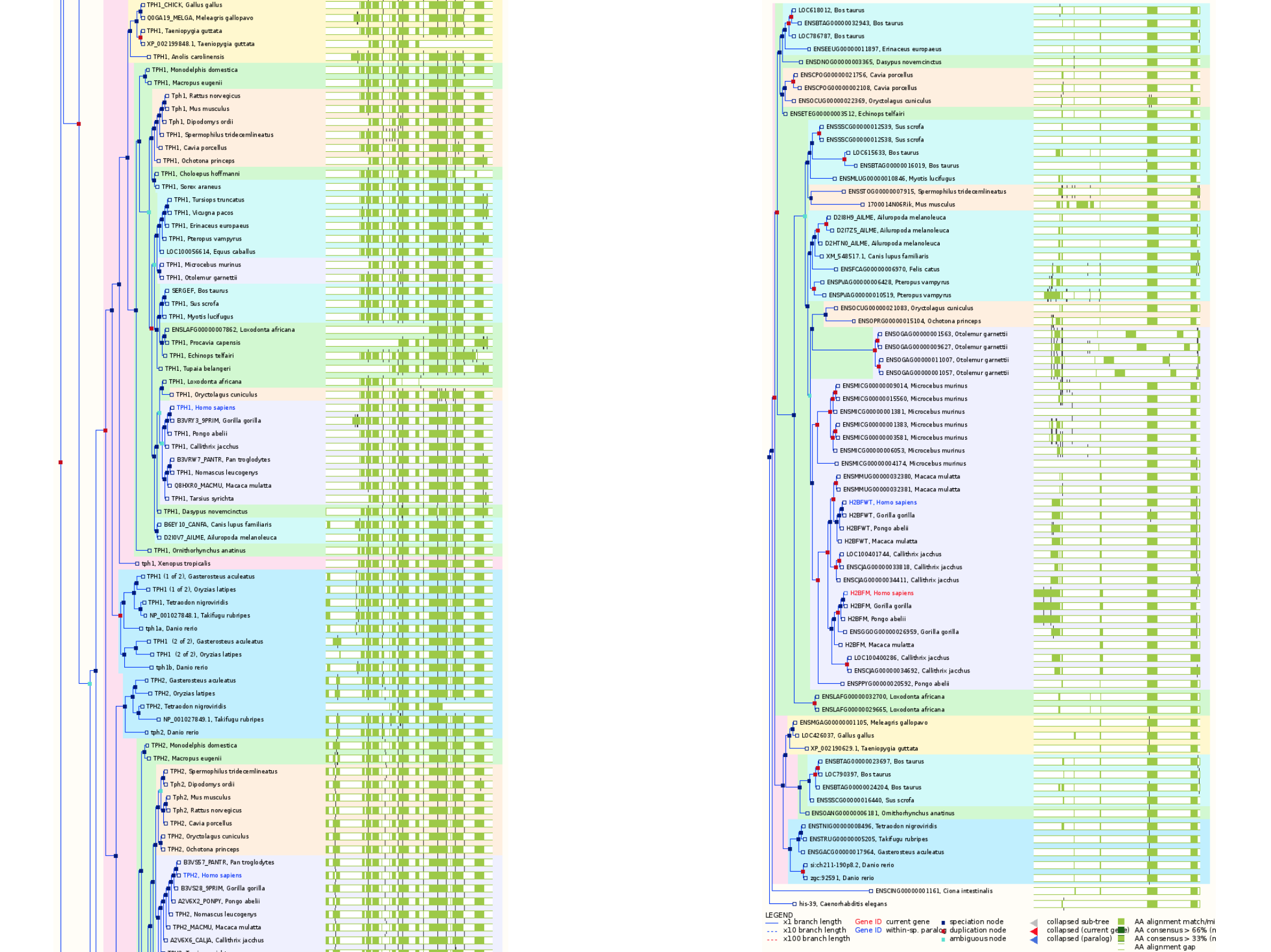
Przykład PAH i TPH

- PAH – hydroksylaza fenyloalaniny, to znany enzym, który występuje w pojedynczej kopii w u niemal wszystkich zwierząt
- Mutacje w tym genie prowadzą do recesywnej choroby genetycznej – fenyloketonurii
- Paralog PAH – TPH – hydroksylaza tryptofanu, występuje w bardzo różnej liczbie kopii u różnych organizmów

Przykład 2. histony

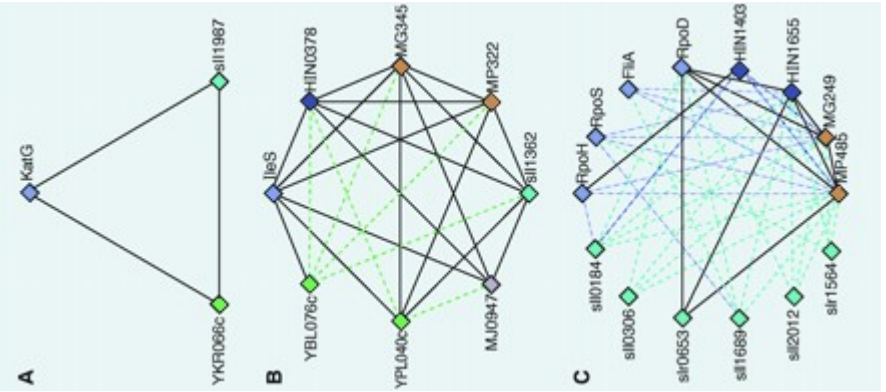
- Histony, które służą jako element strukturalny w komórkach eukariontycznych, mają zupełnie inne wymagania – wiele niezależnych kopii w genomie jest potrzebnych dla produkcji milionów białek




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
L0C618012, 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

Wykorzystanie BLASTa do znajdowania homologów

- BLAST Best hits – białka, które są dla siebie wzajemnie „trafieniami” BLASTa
- Możemy skonstruować graf, gdzie wierzchołkami są wszystkie geny, a krawędziami trafienia BLASTem
- Szukając „prawie-klik” w tym grafie możemy skonstruować COGs – klastry genów ortologicznych



Jak określić funkcje genów?

- Większość białek ma więcej niż jedną „funkcję”, i nie jest łatwe ich testowanie
- Zwykle jest to wykonywane eksperymentalnie, i bardzo czasochłonne
- Dla większości białek nadal nie wiemy jaka jest ich funkcja
- Od lat trwają wysiłki w systematyzacji tej wiedzy

Różne podejścia do opisu funkcji

- Gene Ontology – baza danych tworzona ręcznie przez wielu kuratorów, składająca się z kontrolowanego słownika funkcji w postaci DAGu (grafu skierowanego acyklicznego) i przypisań do niego
- Wikigenes – oparte na metodach text miningu automatycznie tworzone wiki na podstawie tekstów naukowych

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

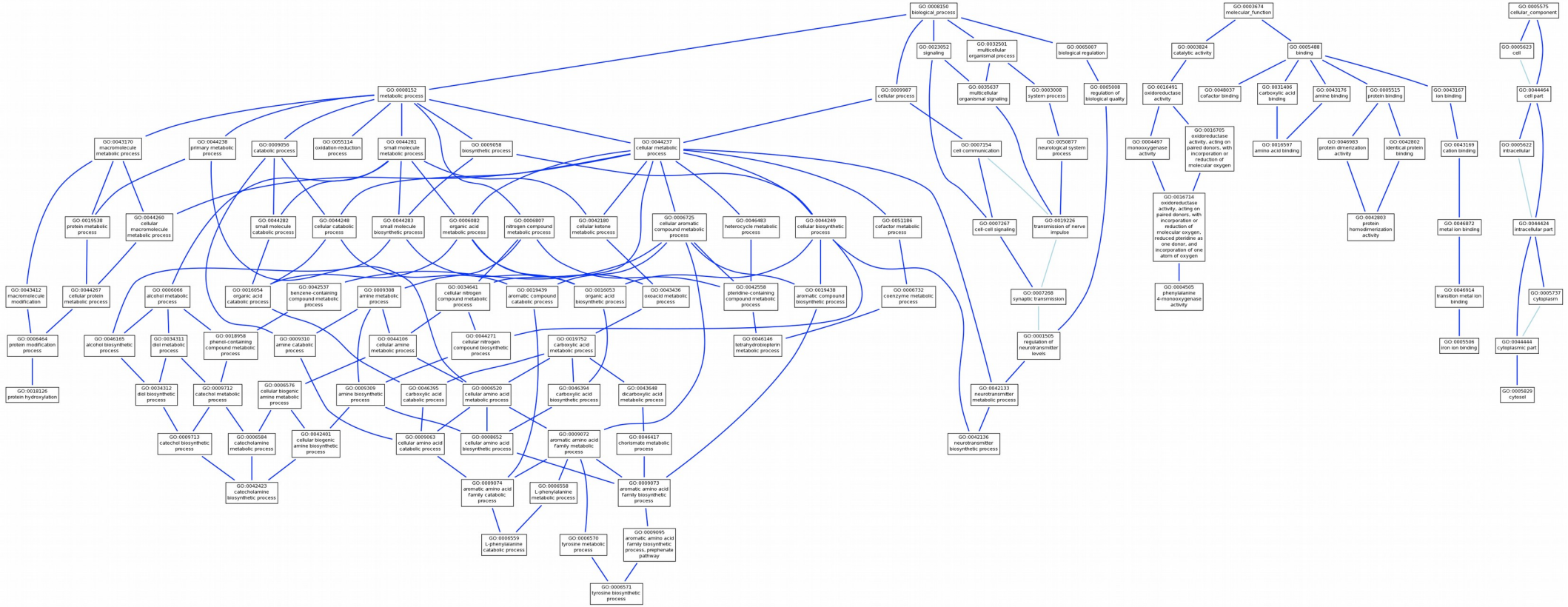
Filter Associations

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


GO:0008150
biological_process
GO:0003674
molecular_function
GO:0005575
cellular_component
GO:0023052
signaling
GO:0032501
multicellular
organismal process
GO:0065007
biological regulation
GO:0003824
catalytic activity
GO:0005488
binding
GO:0005623
cell
GO:0008152
metabolic process
GO:0009987
cellular process
GO:0035637
multicellular
organismal signaling
GO:0003008
system process
GO:0065008
regulation of
biological quality
GO:0016491
oxidoreductase
activity
GO:0048037
cofactor binding
GO:0031406
carboxylic acid
binding
GO:0043176
amine binding
GO:0005515
protein binding
GO:0043167
ion binding
GO:0044464
cell part
GO:0043170
macromolecule
metabolic process
GO:0044238
primary metabolic
process
GO:0009056
catabolic process
GO:0055114
oxidation-reduction
process
GO:0044281
small molecule
metabolic process
GO:0009058
biosynthetic process
GO:0044237
cellular metabolic
process
GO:0007154
cell communication
GO:0050877
neurological system
process
GO:0004497
monooxygenase
activity
GO:0016705
oxidoreductase
activity, acting on
paired donors, with
incorporation or
reduction of
molecular oxygen
GO:0016597
amino acid binding
GO:0046983
protein dimerization
activity
GO:0042802
identical protein
binding
GO:0043169
cation binding
GO:0005622
intracellular
GO:0019538
protein metabolic
process
GO:0044260
cellular
macromolecule
metabolic process
GO:0044282
small molecule
catabolic process
GO:0044248
cellular catabolic
process
GO:0044283
small molecule
biosynthetic process
GO:0006082
organic acid
metabolic process
GO:0006807
nitrogen compound
metabolic process
GO:0042180
cellular ketone
metabolic process
GO:0006725
cellular aromatic
compound metabolic
process
GO:0046483
heterocycle metabolic
process
GO:0044249
cellular biosynthetic
process
GO:0051186
cofactor metabolic
process
GO:0007267
cell-cell signaling
GO:0019226
transmission of nerve
impulse
GO:0016714
oxidoreductase
activity, acting on
paired donors, with
incorporation or
reduction of
molecular oxygen,
reduced pteridine as
one donor, and
incorporation of one
atom of oxygen
GO:0042803
protein
homodimerization
activity
GO:0046872
metal ion binding
GO:0044424
intracellular part
GO:0043412
macromolecule
modification
GO:0044267
cellular protein
metabolic process
GO:0006066
alcohol metabolic
process
GO:0016054
organic acid
catabolic process
GO:0042537
benzene-containing
compound metabolic
process
GO:0009308
amine metabolic
process
GO:0034641
cellular nitrogen
compound metabolic
process
GO:0019439
aromatic compound
catabolic process
GO:0016053
organic acid
biosynthetic process
GO:0043436
oxoacid metabolic
process
GO:0042558
pteridine-containing
compound metabolic
process
GO:0019438
aromatic compound
biosynthetic process
GO:0006732
coenzyme metabolic
process
GO:0007268
synaptic transmission
GO:0004505
phenylalanine
4-monooxygenase
activity
GO:0046914
transition metal ion
binding
GO:0005737
cytoplasm
GO:0006464
protein modification
process
GO:0046165
alcohol biosynthetic
process
GO:0034311
diol metabolic
process
GO:0018958
phenol-containing
compound metabolic
process
GO:0009310
amine catabolic
process
GO:0044106
cellular amine
metabolic process
GO:0044271
cellular nitrogen
compound biosynthetic
process
GO:0019752
carboxylic acid
metabolic process
GO:0046146
tetrahydrobiopterin
metabolic process
GO:0001505
regulation of
neurotransmitter
levels
GO:0005506
iron ion binding
GO:0044444
cytoplasmic part
GO:0018126
protein hydroxylation
GO:0034312
diol biosynthetic
process
GO:0009712
catechol metabolic
process
GO:0006576
cellular biogenic
amine metabolic
process
GO:0009309
amine biosynthetic
process
GO:0046395
carboxylic acid
catabolic process
GO:0006520
cellular amino acid
metabolic process
GO:0046394
carboxylic acid
biosynthetic process
GO:0043648
dicarboxylic acid
metabolic process
GO:0042133
neurotransmitter
metabolic process
GO:0005829
cytosol
GO:0009713
catechol biosynthetic
process
GO:0006584
catecholamine
metabolic process
GO:0042401
cellular biogenic
amine biosynthetic
process
GO:0009063
cellular amino acid
catabolic process
GO:0008652
cellular amino acid
biosynthetic process
GO:0009072
aromatic amino acid
family metabolic
process
GO:0046417
chorismate metabolic
process
GO:0042136
neurotransmitter
biosynthetic process
GO:0042423
catecholamine
biosynthetic process
GO:0009074
aromatic amino acid
family catabolic
process
GO:0006558
L-phenylalanine
metabolic process
GO:0009073
aromatic amino acid
family biosynthetic
process
GO:0006559
L-phenylalanine
catabolic process
GO:0006570
tyrosine metabolic
process
GO:0009095
aromatic amino acid
family biosynthetic
process, prephenate
pathway
GO:0006571
tyrosine biosynthetic
process

[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].

Testowanie istotności funkcji

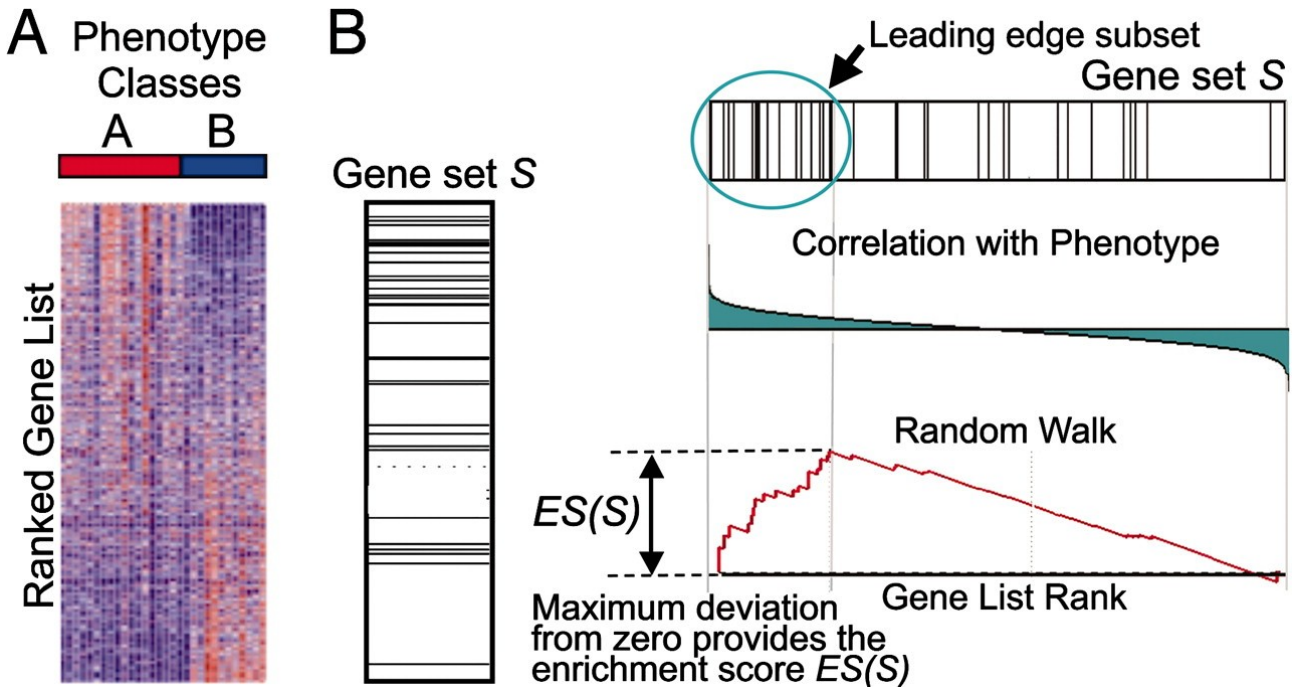
- Dla wielu zbiorów genów możemy interesować się ich funkcją (np. genów o zmienionej ekspresji, genów z mutacjami w określonej chorobie, genów wpływających na jakąś ilościową cechę
- Często mamy dany zbiór genów, lub ranking genów (według wpływu na jakąś cechę lub ekspresji) i możemy zapytać, czy wśród genów w danym zbiorze nadreprezentowana jest jakaś funkcja. Jak odpowiedzieć na takie pytanie?

Istotne przecięcie zbiorów genów

- Najprościej jest zbadać, czy przecięcie zbioru genów posiadających daną funkcję z genami ze zbioru „zapytania” jest statystycznie istotne biorąc pod uwagę rozmiary tych zbiorów
- Odpowiada na to pytanie test Fishera, który korzysta z rozkładu hipergeometrycznego
- Standardowe pytanie: jaka jest szansa, że wyciągnę k lub więcej czarnych kul, jeśli losuję n kul spośród N kul w urnie, jeśli wiem, że kul czarnych jest M

Gene set enrichment analysis

- Jeśli mamy ranking genów to możemy zbadać, czy w górnej części rankingu mamy wzbogacenie którejś z interesujących nas funkcji wykorzystując miejsce największego wzbogacenia i empiryczne obliczanie p-wartości



Poprawki na testowanie wielu hipotez

- Ponieważ w przypadku zbiorów i rankingów genów testujemy zwykle tysiące hipotez związanych z różnymi funkcjami potencjalnie wzbogaconymi, musimy stosować poprawki na testowanie wielu hipotez
- Często stosuje się poprawkę Bonferroniego, czyli mnożenie p-wartości przez liczbę testowanych funkcji
- Lepszym sposobem jest wyliczanie false Discovery rate – FDR, który mierzy ile wyników fałszywie pozytywnych spodziewamy się w zadanej grupie predykcji