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)



		-	🗖 Q0GA19_MELGA, Meleagris gallopavo		
		ιT	TPH1, Taeniopygia guttata		
	XP_002199848.1, Taeniopygia guttata		VD 0001000401 Terrismusis antista		1.1.1.1
			L XP_002135846.1, Taenopygia guccata		
		TPH1, Anolis carolinensis			
			TPH1. Macropus eugenii		
		Tph 1, Rattus norvegicus			
			Tph 1, Mus musculus		
			Tph 1, Dipodomys ordii		
			TTPU1 Searman billus trida combine attus	i lini li	
		1	a fill in the spennophilas cide certainea aus		
			D TPH1, Cavia porcellus		
ſ	1	'	D TPH1. Ochotona princeps		
			TPH1, Choloepus hoffmanni		
			TOUL Comercentration		
			12 I'PH1, Sorex araneus		
			TPH1, Tursiops truncatus		
		1	🛏 📫 TPH1, Vicugna pacos		
			TPH1, Erin aceus europaeus		
			La mini, recipius vampyrus		
			LOC100056614, Equus caballus		
			P TPH1, Microcebus murinus		
			TPH1, Otolemur garnettii		
	1		C SERGEE Bostaurus		
	1		E IPH1, Sus scrota		
		T	TPH1, Myotis lucifugus		
			📫 🗖 ENSLAFG0000007862, Loxodonta africana		
			TPH1 Procavia capensis		
					1 1 1 1 1 1 1
			C TPH1, Echinops telfairi		
			🗂 TPH1, Tupaia belangeri		
			TPH1, Loxodonta africana		
			TEH1 Opertalianus cuniculus		
			a mit, orycolagos cunculos		
			IPH1, Homo sapiens		
			B3VRY 3_9PRIM, Gorilla gorilla		
			TPH1, Pongo abelii		
			TPH1 Callithrix jacchus		
			B3VRW /_PANIR, Pan troglodytes		
			TPH1, Nomascus leu cogenys		
			🛛 Q8HXR0_MACMU, Macaca mulatta		
			T TPH1 Tarsius syrichta		
			TOUL Deserve environmentes		
			U IFHI, Dasypus novemcinctus		
			B6EY 10_CANFA, Canis lupus familiaris		
			D210V7 ALIME Ailuropoda melanoleuca		
		L			
			TPH1, Ornithorhyn chus an atinus		
			TPH1, Ornithorhynchus anatinus		
		-0	TPH1, Ornithorhynchus anatinus the provide straight of the provided straight of the pro		
		-0	□ TPH1, Ornithorhynchus anatinus □ tph1, Xenopus tropicalis □ TPH1 (1 of 2), Gasterosteus aculeatus		
	_	•	TPH1, Ornithorhynchus anatinus Tph1, Xenopus tropialis TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oryzias latipes		
	F		TPH1, Ornithorhynchus anatinus TPH1, Crnithorhynchus anatinus TPH1 (c of 2), Gasterosteus aculeatus TPH1 (c of 2), Oyzias latipes TPH1, Tetraodon nigroviridis		
[TPH1, Ornithorthynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Ogzisa latipes TO TPH1, Tetraodon nigrovinidis TO TPH1, Tetraoda 1, Takifugu rubinoes		
[TPH1, Ornithorhynchus anatinus TpH1, Kenopus tropicalis TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oryzias latipes TPH1, Tetraadon nigroviridis Neo NP, On1027848, 1, Takifugu rubripes TPH1, Tetra		
			TPH1, Ornithorhynchus anatinus To TPH1, Steadon nigrovintis To TPH1 (1 of 2), Oxyzias latipes TPH1 (1 of 2), Oxyzias latipes D TPH1, Teradon nigrovintis D NP_001027848, 1, Takifugu rubripes D th1a, Danio renio		
			TPH1. Ornithorhynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TO TPH1 (1 of 2), Oyzias latipes TO TPH1. Tetra odon nigroviridis TO NP_001027848.1, Takifugu rubripes To thh.1, Danio reno TPH1 (2 of 2), Gasterosteus aculeatus		
			TPH1, Ornithorhynchus anatinus TPH1 (1 of 2), Gastarostaus aculeatus TPH1 (1 of 2), Oryzias latipes TPH1 (1 of 2), Oryzias latipes To TPH1, Tetraodon nigrovindis To TPH0, Danio renio TPH1 (2 of 2), Gasterosteus aculeatus To TPH1 (2 of 2), Oryzias latipes		
			TPH1. Ornithorhynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TO TPH1. Tetraodon nigrovinidis TO PP01027848. Takifugu rubripes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH5, Danio rerio TPH5, Caster atus aculeatus		
			TPH1, Ornithorhynchus anatinus TPH1 (1 of 2), Gastarosteus aculeatus TPH1 (1 of 2), Oryzias latipes TPH1 (1 of 2), Oryzias latipes To TPH1, Tetraodon nigroviridis To TPH0, Onio127848,1, Takifugu rubripes To tha, Danio rerio TPH1 (2 of 2), Gasterosteus aculeatus To TPH1 (2 of 2), Oryzias latipes To thib, Danio rerio TPH2, Gasterosteus aculeatus		
			TPH1. Orrithortynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes To TPH1. Tetraodon nigroviridis To NP_001027848.1, Takifugu rubripes TPH1. (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1. (2 of 2), Oyzias latipes TPH1. (2 of 2), Oyzias latipes TPH2. Gasterosteus aculeatus TPH2. Gasterosteus aculeatus TPH2. Tetraodon nigroviridis		
			TPH1. Ornithorhynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes To NP_001027848.1, Takifugu rubripes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH2, Tetraodon nigroviridis TPH2, Tetraodon nigroviridis TPH2, Tetraodon nigroviridis TPH2, Tetraodon nigroviridis TPH2, Danio rerio TPH2, Danio rerio		
			TH1. Orrithorhynchus anatinus TH1. Orrithorhynchus anatinus TH1. Control of the second of the se		
			TPH1. Orinithorthynchus anatinus TPH1. (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2, Coyzias latipes TPH2, Tetraedon nigrovindis TPH2, Coyzias latipes T		
			TPH1. Onrich orkponchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2 (2 of 2), Oyzias latipes TPH2 (2 of 2), Oyzias latipes TPH2 (2 of 2), Oyzias latipes TPH2, Casterosteus aculeatus TPH2, Danio rerio TPH2, Manodelphis domestica TPH2, Manodelphis consection TPH2, TPH2		
			TPH1. Orrithorthynchus anatinus TPH1. Control of the second of		
			TPH1, Onrichorlynchus anatinus TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2 (2 of 2), Oyzias latipes TPH2, Gasterosteus aculeatus TPH2, Casterosteus aculeatus TPH2, Danio rerio TPH2, Marcodelphis domestica TPH2, Marcodelphis tridecemlineatus TPH2, Dipodomys ordii TPH2, Dipodomys ordii TPT2, Dipodomys ordii TPT2, Dipodomys ordii TPT2, Nermophilus tridecemlineatus		
			TPH1, Orridborthynchus anatinus TPH1, Control or TPH2, Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2, Casterosteus aculeatus TPH2, Casterosteus aculeatus TPH2, Oyzias latipes TPH2, Casterosteus aculeatus TPH2, Oyzias latipes TPH2, Tetraedon nigrovindis TPH2, Macropus eugenii TPH2, These musculus TPH2, These musculus TPH2, These musculus TPH2, These more picus		
			TPH1. Un of 2). Gasterosteus aculeatus TPH1 (1 of 2). Gasterosteus aculeatus TPH1 (1 of 2). Oyzias latipes To TPH1. (2 of 2). Gasterosteus aculeatus TPH1 (2 of 2). Gasterosteus aculeatus TPH2. Gasterosteus aculeatus TPH2. Gasterosteus aculeatus TPH2. Tetraadon nigrovindis TPH2. Tetraadon nigrovindis TPH2. Gasterosteus aculeatus TPH2. Tetraadon nigrovindis TPH2. Monodelphis domestica TPH2. Spermophilus tride camlineatus TPH2. Dipadomys ordii TPH2. Cavia porcellus		
			TPH1, Ornithorthynchus anatinus TPH1, Corpus tropicalis TPH1 (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2, Mancpus eugenii TPH2, Oyzias latipus encelus		
			TPH1. Orrithorthynchus anatinus TPH1. Orrithorthynchus anatinus TPH1. U of 2). Gasteroteus aculeatus TPH1 (1 of 2). Oyzias latipes To TPH1. Tetraodon nigroviridis TPH1. (2 of 2). Gasteroteus aculeatus TPH1. (2 of 2). Oyzias latipes TPH2. Dipodomys ordit TPH2. Austrusculus TPH2. Cavia porcellus TPH2. Cavia porcellus TPH2. Cavia porcellus TPH2. Oyzias latipes TPH2. Oyzias		
			TPH1. Orrithorthynchus anatinus TPH1. (1 of 2). Gasterosteus aculeatus TPH1 (1 of 2). Gasterosteus aculeatus TPH1 (1 of 2). Oyzias latipes TPH1 (1 of 2). Oyzias latipes TPH1 (2 of 2). Gasterosteus aculeatus TPH2. Gryzias latipes TPH2. Tetraodon nigrovindis TPH2. Tetraodon nigrovindis TPH2. Tetraodon nigrovindis TPH2. Macropus sugenti TPH2. Spermophilus tride cemlineatus TPH2. Dipodomys ordii TPH2. Cavia porcellus TPH2. Gryctolagus cuniculus TPH2. Ochotona princeps		
			TPH1, Ornithorthynchus anatinus TPH1, Caroladon nigrovindis TPH1, 11 of 2), Gasterosteus aculeatus TPH1 (11 of 2), Oyzias latipes TPH1 (21 of 2), Oszterosteus aculeatus TPH1 (21 of 2), Oszterosteus aculeatus TPH2, Caroladon nigrovindis TPH2, Organis latipes TPH2, Macropus eugenii TPH2, Macropus eugenii TPH2, Carola porcellus TPH2, Carola porcellus TPH2, Oryctolagus cuniculus		
			TPH1. Orrithorhynchus anatinus TPH1. Control of the second of t		
			TFH1, Onricherhynchus anatinus TFH1, U of 2), Gasterosteus aculeatus TFH1 (1 of 2), Oyzias latipes TFH1 (1 of 2), Oyzias latipes TFH1 (1 of 2), Oyzias latipes TFH1 (2 of 2), Oyzias latipes TFH2, Tetradon nigroviridis TFH2, Tetradon nigroviridis TFH2, Ceaterosteus aculeatus TFH2, Oyzias latipes TFH2, Ceaterosteus aculeatus TFH2, Nacropus eugenii TFH2, Ceaterosteus TFH2, Ceaterosteus TFH2, Oyzias latipes TFH2, Ceaterosteus TFH2, Oyzias latipes TFH2, Oyzias latipes TFH2, Oyzias latipes TFH2, Oyzias latipes TFH2, Ceaterosteus TFH2, Oyzias latipes TFH2, Oyzias latipes TFH2, Ceaterosteus TFH2, Oyzias latipes TFH2		
			TPH1. Orrithorthynchus anatinus TPH1. U of 2). Gasterosteus aculeatus TPH1 (1 of 2). Oyzias latipes TO TPH1. Tetraodon nigrovinids To the 1. Tetraodon nigrovinids To the 0.01027848.1. Takifugu rubripes TPH2. Tetraodon nigrovinids TPH1. (2 of 2). Gasterosteus aculeatus TPH1. (2 of 2). Oyzias latipes TPH1. (2 of 2). Oyzias latipes TPH2. Tetraodon nigrovinids TPH2. Gasterosteus aculeatus TPH2. Oyzias latipes TPH2. Tetraodon nigrovinids TPH2. Monodelphis domestica TPH2. Danio rerio TPH2. Monodelphis domestica TPH2. Dipodomys ordii TPH2. Dipodomys ordii TPH2. Cavia porcelus TPH2. Tetra objets TPH2. Tetra objets TPH2. Cavia porcelus TPH2. Tetra objets TPH2. Cavia porcelus TPH2. Tetra objets TPH2. Cavia porcelus TPH2. Cavia porcelus TPH2. Cavia porcelus TPH2. Tetra objets TPH2. Orycoblagus cuniculus TPH2. Cavia porcelus TPH2. Tetra objets TPH2. Orycoblagus cuniculus TPH2. Cavia porcelus TTPH2. Cavia porcelus TPH2. Cavia porcelus TTPH2. Tetra traolodytes		
			TPH1. Orridborthynchus anatinus TPH1. Orridborthynchus anatinus TPH1. (1 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Gasterosteus aculeatus TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2, Gasterosteus aculeatus TPH2, Gasterosteus aculeatus TPH2, Oyzias latipes TPH2, Macropus sugenti TPH2, Norosedius uniculus TPH2, Oyzias latipes DTPH2, Ataus norvegicus TPH2, Oyzias latipes DTPH2, Oyzias latipes DTPH2, Oyzias latipes DTPH2, Control princeps DB3/S57_PMIFR, Pan troglo dytes DTPH2, Control princeps DB3/S57_PMIFR, Pan troglo dytes DTPH2, Norosedigus cuncilus DTPH2, Macropus sugens DTPH2, Macropus sugens DTPH2, Macropus sugens DTPH2, Macropus sugens DTPH2, Norosedigus cuncilus DTPH2, Macropus sugens DTPH2, Norosedigus cuncilus DTPH2, Norosedig		
			TPH1, Onridborthynchus anatinus TPH1, 10 of 2), Gasterosteus aculeatus TPH1 (1 of 2), Oyzias latipes TPH1 (1 of 2), Oyzias latipes TPH1 (2 of 2), Oyzias latipes TPH2, Casterosteus aculeatus TPH2, Manodelphis domestica TPH2, Macropus eugenii TPH2, Marcopus eugenii TPH2, Casterosteus latis TPH2, Casterosteus TPH2, Normaculus TPH2,		

гŧ

LUC618012, Bostaurus	
ENSBTAG00000032943, Bos taurus	
LOC786787, Bostaurus	
ENSEEUG00000011897. Erina ceus europaeus	
LI ENSDING GOOD 00003365, Dasypus novem cinclus	
ENSCP0 G0 00 00 02 1756, Cavia porcellus	
ENSCP0 G0 00 00 00 2108, Cavia porcellus	
ENSO CUG 00 000 02 23 69. Orvetol agus cuniculus	<u>`</u>
ENSELEGOUDUUUUS 12, EChinops teitairi	
ENSSSCG0000012539, Sus scrofa	
ENSSSCG00000012538, Sus scrofa	
JUC615633 Bostaurus	
ENSBIAG00000016019, Bos taurus	
ENSMLUG0000010846, Myotis lucifugus	
ENSST 0G 0000007 915, Spermophilus tride cemlineatus	
T 17000 14N06 Rik Mus musculus	
D218H9_AILME, Ailuropoda melanoleuca	
D217Z5_AILME, Ailuropoda melanoleuca	
D2HT N0_AILME, Ailuropoda melanoleuca	
M 548517 1 Canis Junus familiaris	
ENSECAGO000006970, Felis catus	
ENSPVAG0000006428, Pteropus vampyrus	
ENSPVAG0000010519, Pteropus vampyrus	
ENSO CUGODODO2 1092 On stal anus quaisulus	
ENSOPRG0000015104, Ochotona princeps	
🗆 ENSO GAG 00 000 00 15 63, Otolemur garnettii	
ENSO GAG0000009627. Otolemur garnettii	
	N
ENSOGAGODODOTIOU7, Otolemur garneto	
🗉 ENSO GAG 00 000 00 1057, Otolemur garnettii	
ENSMICG0000009014, Microcebus murinus	
ENSMICG00000015560 Microcebus mutinus	
	- * I
ENSMICG0000001381, Microcebus munnus	
ENSMICG0000001383, Microcebus murinus	
ENSMICG 000 00 0358 1, Microcebus murinus	
ENSMICG0000006053 Microcebus mutinus	
ENSMICG00000041/4, Microcebus munnus	
ENSMMUG0000032380, Macaca mulatta	
ENSMMUG00000032381, Macaca mulatta	
T H2REWT Homo saniens	
H2BFWT, Gonlla gonlla	
🖬 H2BFWT, Pongo abelii	
🖸 H2BFWT, Macaca mulatta	
T LOCIO0401744 Callithrix ia crbus	
ENSCIAGODODO033818, Califfrin× Jacobus	
🖿 🖿 ENSCJAG 00 00 034411, Callithri× ja cchus	
H2BFM, Homo sapiens	
H2BEM, Gorilla gorilla	
L HZOFM, FONGO ADEIN	
ENSGGOG0000026959, Gorilla gorilla	
H2BFM, Macaca mulatta	
L0C100400286 Callithrix jacthus	
ENISCIA CODEREGACION CHIER IN THE	
Li ENGENGOU DU SA652, Californix ja corus	
ENSPPYG0000020592, Pongo abelii	
ENSLAFG00000032700, Loxodonta africana	
ENSLAFG 000 00 02 9665, Loxodonta africana	
En Ensilo Addiction of 105, meleagins ganopavio	
LOC426037, Gallus gallus	
XP_002 19062 9. 1, Taeniopygia guttata	
ENSBTAG 0 000 00 23 69 7. Bos taurus	
LOC700307 Destaurus	
LUC/90397, Bostaurus	
ENSBTAG 0 000 00 2420 4, Bos taurus	
ENSSSCG0000016440, Sus scrofa	
ENSOANG0000006 181, Omithorhynchus anatinus	
TO ENSTRUCTOR OF COMPANY AND A STATE OF	
Enstraisou ou ou 8496, letrao don nigrovindis	
- ENSTRUGODODODE20E Tabélugu pubrinor	
E ENSTRODODODOS205, Takit ugu Tubripes	
ENSGACG00000017964, Gasterosteus aculeatus	
ENSGACG000001250, raking rubines ENSGACG0000017964, Gasterosteus aculeatus disch211-19082.2. Danio refn	
ENSIGACG0000000220, initiality funities ENSGACG0000017964, Gasterosteus aculeatus isich211-190p8.2, Danio rerio	
Existro 00000000202, instruger tuninges GENSGACG 00000017954, Gastreosteus a culeatus gi sich 211-190p8.2, Danio rerio ge: 92591, Danio rerio	
a Sich211-190p8.2, Danio rerio zgc:92591, Danio rerio c ENSCING0000001161, Ciona intestinalis	
ENSIGACG0000003203, iainings tumpes ENSIGACG00000017964, Gasterosteus aculeatus p sich211-190p8.2, Danio rerio zgc:92391, Danio rerio ENSCING00000001161, Ciona intestinalis Dis:39, Caenorhabditis elegans	
D EXISTO 000000000202, instruge trainings D SiSGACC00000001954, Gasterosteus a culeatus D SiSCA211-190p8.2, Danio rerio D agc: 9259 1, Danio rerio D ENSCING00000001161, Ciona intestinalis D His-39, Caenorhabditis elegans	
D ELSINGACG0000003202, lanting tumpes D ELSINGACG0000017964, Gasterosteus aculeatus D sich211-190p8.2, Danio rerio D sig: 92591, Danio rerio D sis-39, Caenonhabditis elegans LEGEND Xb pranch length Gene ID current gene speciation node	collapsed sub-tree
LEGEND The Scarch length Gene Down of the second	collapsed sub-tree collapsed (surrent g

T.

Т

ТŅ.

in 1

1

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 o	r proteins exact match Sub	mit Query	

PAH

Phenylalanine-4-hydroxylase

protein from Homo sapiens (human)

	Term associations Gene product information	on 🌩 Peptide Sequence 🖣	 Sequence information • 		
Term Associations					
Download all association information in: U gene association format	RDF-XML				
 ▼ Filter associations displayed Filter Associations Ontology Evidence Code All Diological process Cellular component DA EA 					
Select all Clear all Perform an action with this page's selected	ed terms				
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





Hoffmann, R. A wiki for the life sciences where authorship matters. Nature Genetics

[search][advanced]

Editor

Edit this page Discussion History

Permanent link

Print Export^{new}

Share

Send to a friend - 🖂 Share @

Personal info

Your homepage Your articles Create article

View

Font size

Gene Review
PAH - phenylalanine hydroxylase
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 <u>classical phenylketonuria</u> who has been screened for mutations in the <u>PAH</u> gene [1].
- Benzo[a]pyrene-induced <u>DNA damage</u> and <u>p53</u> modulation in human <u>hepatoma</u> HepG2 cells for the identification of potential biomarkers for PAH monitoring and <u>risk assessment [2]</u>.

[Sign in / Create account]

POWERED WITH

[edit this page]

Homo sapiens

Simply click in the text to find out who wrote what. Fair credit for authors. Always know your sources.

AUTHORSHIP TRACKING TECHNOLOGY





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.