Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Orthology
Part I
concepts and implications

Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Toni Gabaldón

Contact: tgabaldon@crg.es
Group website: http://gabaldonlab.crg.es
Science blog: http://treevolution.blogspot.com
Twitter: @gabaldonlab, @Toni_Gabaldon

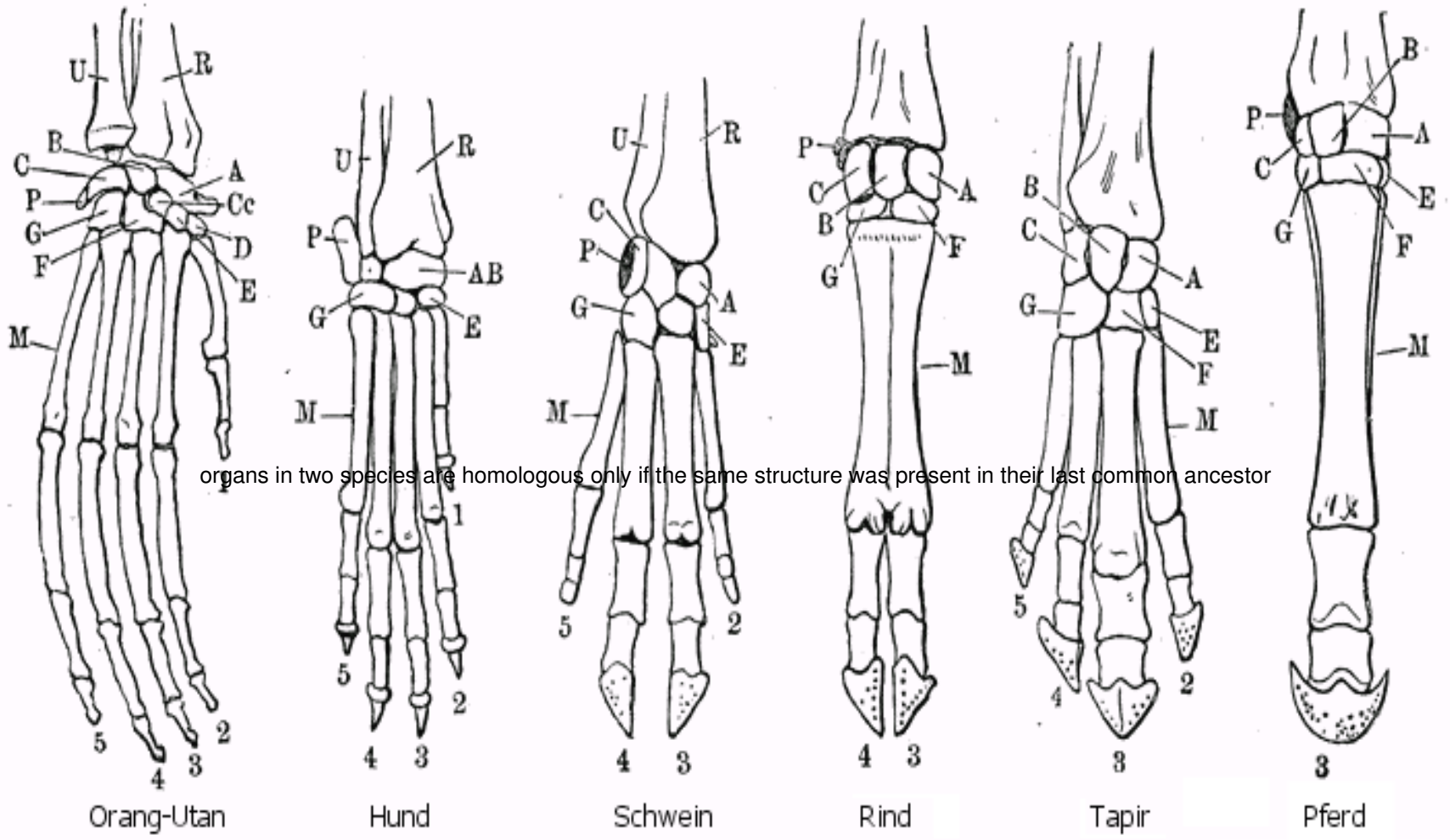
Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Orthology

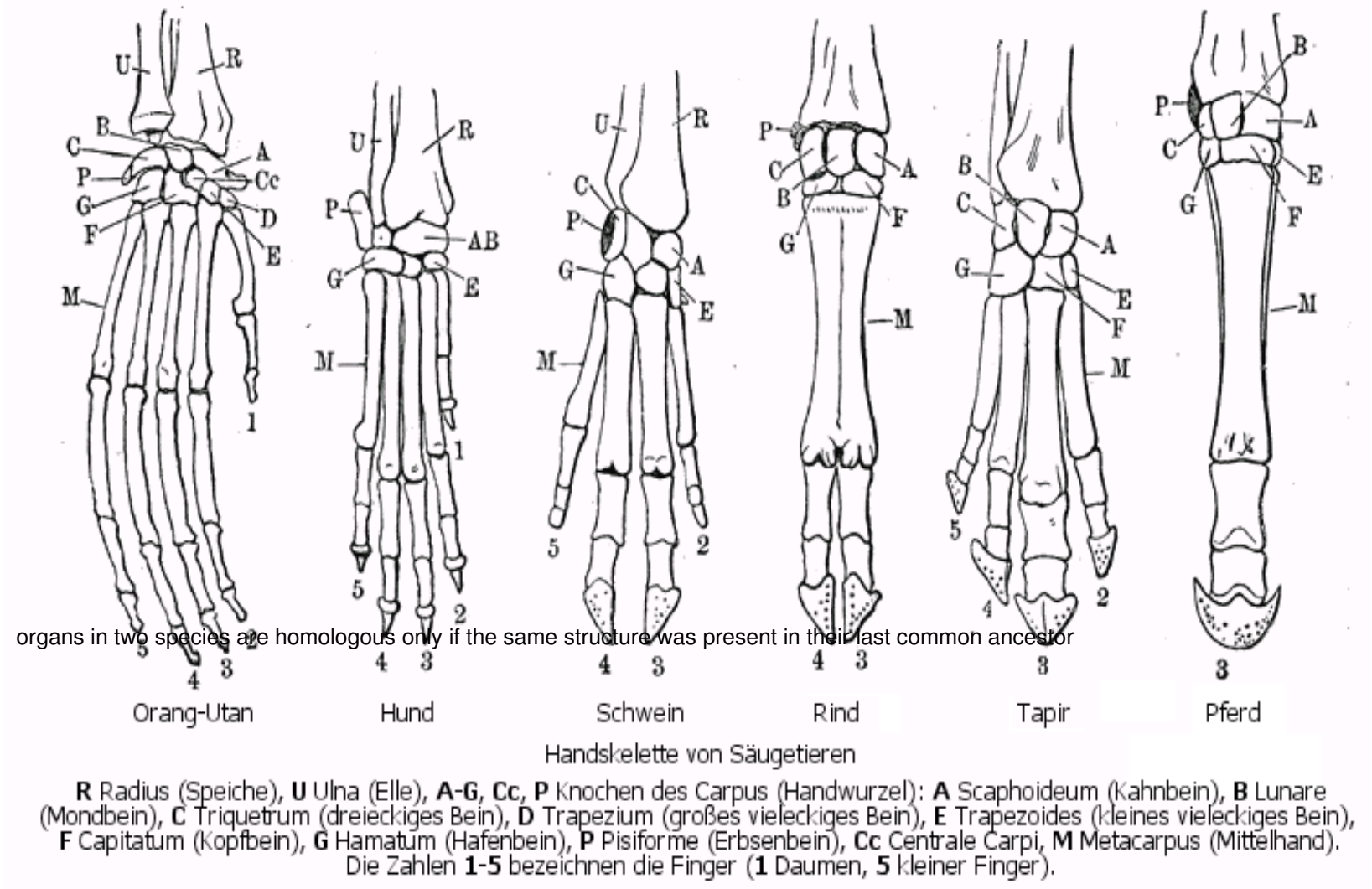
“*concepts and implications*”

organs in two species are homologous only if the same structure was present in their last common ancestor



Handskelette von Säugetieren

R Radius (Speiche), **U** Ulna (Elle), **A-G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hafenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1-5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).



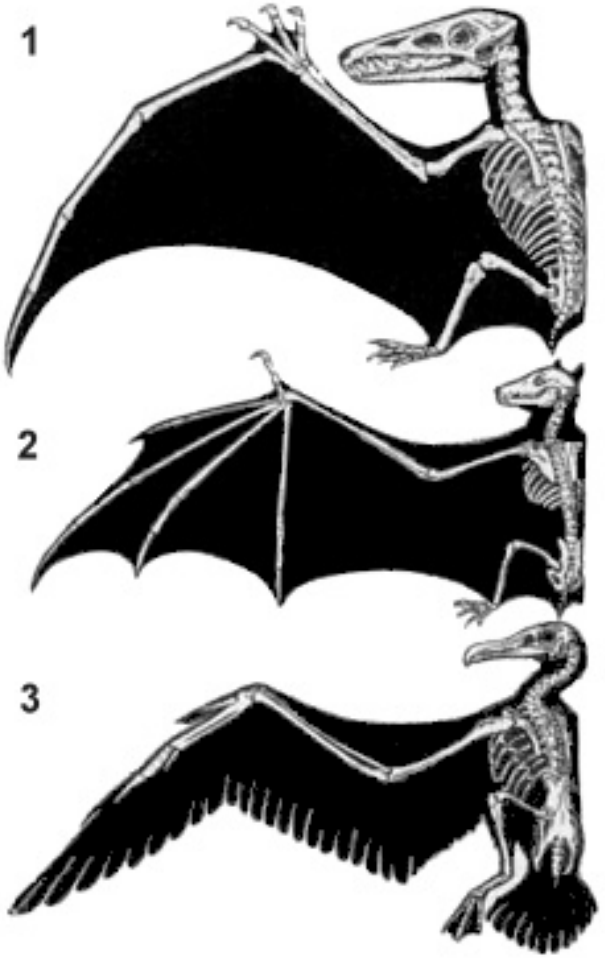
Handskelette von Säugetieren

R Radius (Speiche), **U** Ulna (Elle), **A-G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hafenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1-5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).

"the same organ in different animals under every variety of form and function" **R. Owan**

→ organs in two species are **homologous** only if the same structure was present in their last common ancestor

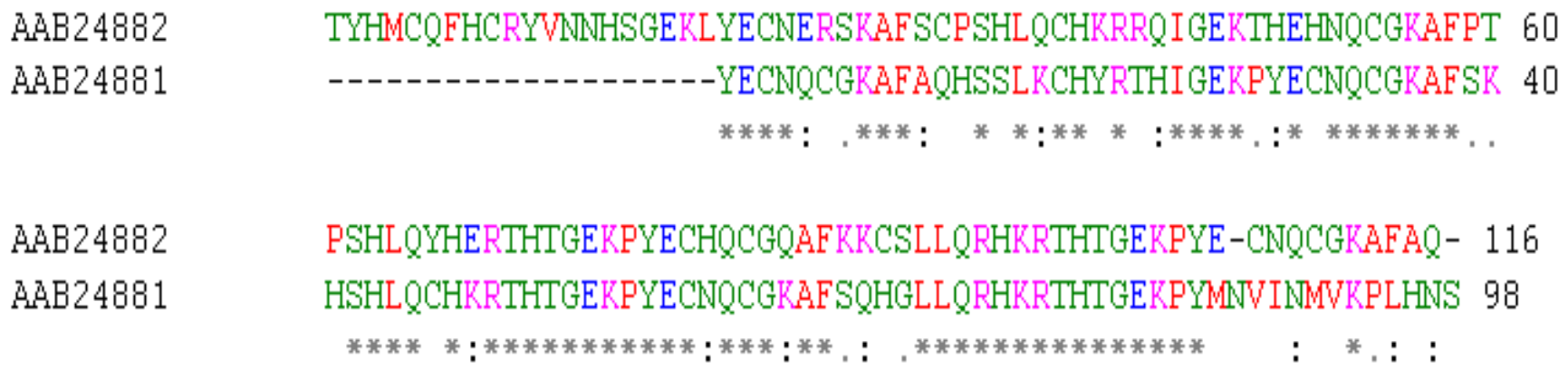
Analogous structures:
Similar function but independet origin.



Homologous as forelimbs
But
Analogous as wings

Extension of the concept of homology to sequences:

Two sequences are homologous if they share common ancestry



Important: Similarity and Homology

Similarity and homology are often confused. e.g. “the sequences are 50% homologous”, “these two sequences are highly homologous”

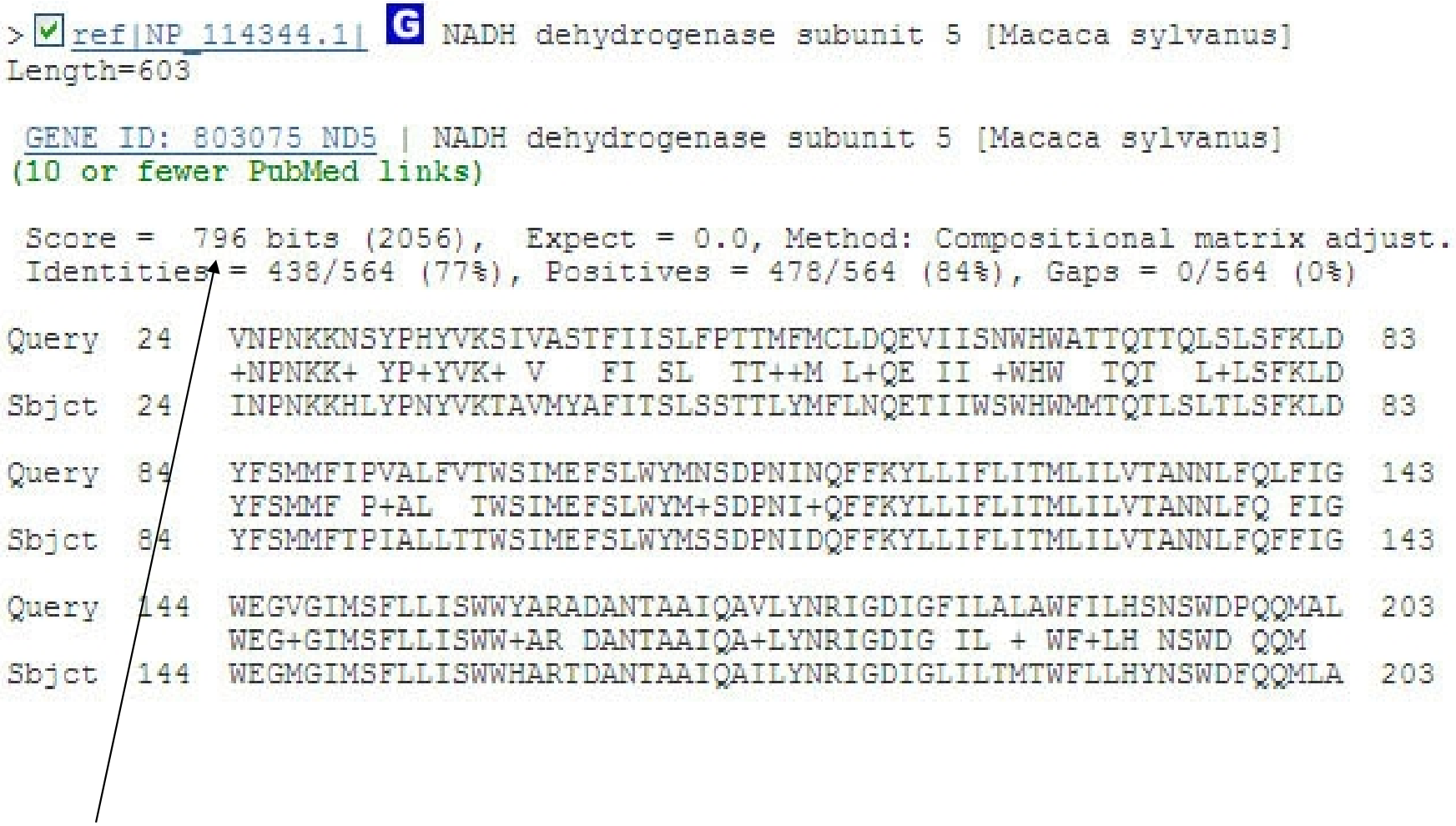
Why is this incorrect? Where does the confusion comes from?

Detour

Sequence similarity, homology detection and blast database queries

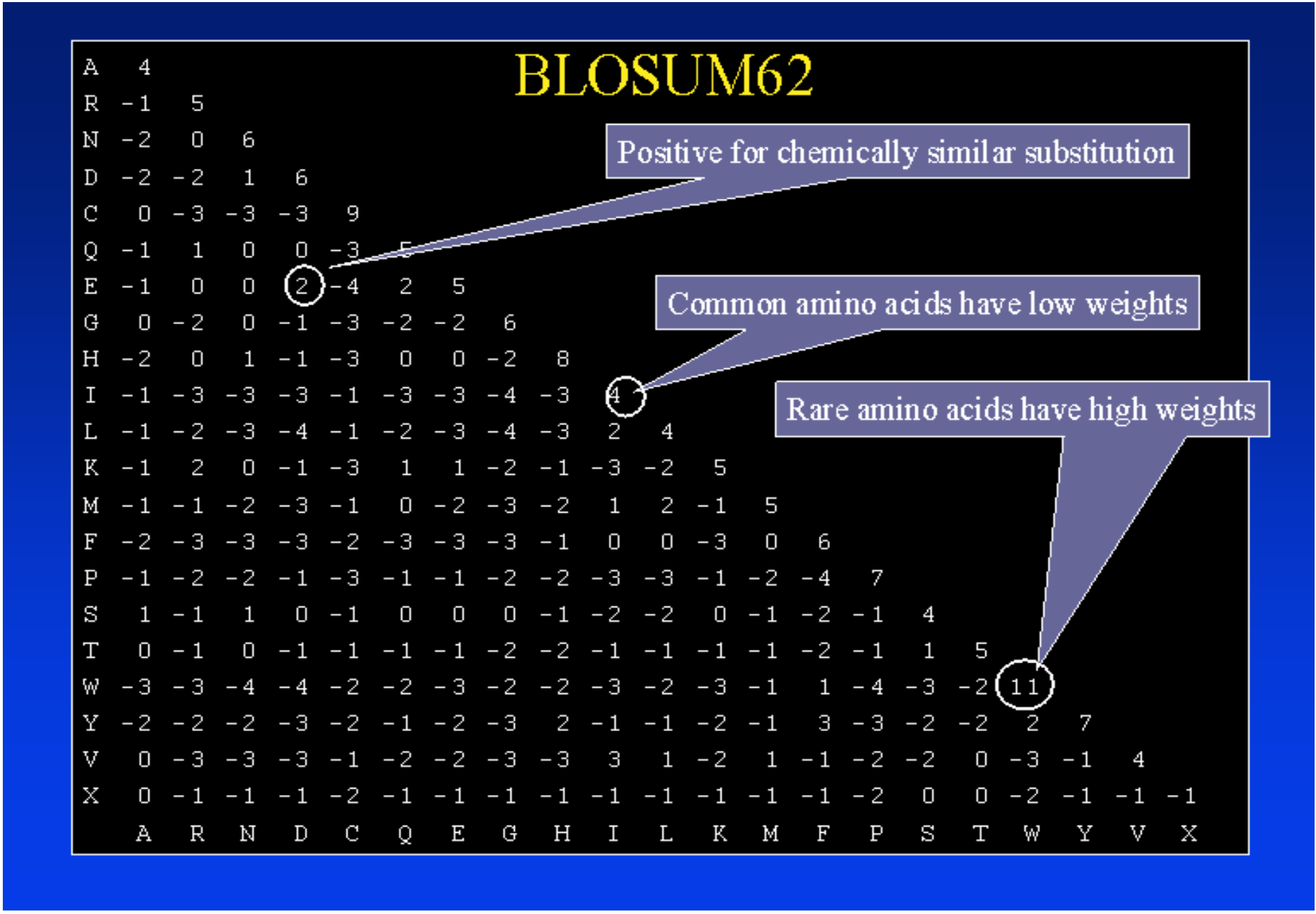
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```

Are this two sequences **significantly** similar?
(i.e how likely is that such an alignment is the result of chance)



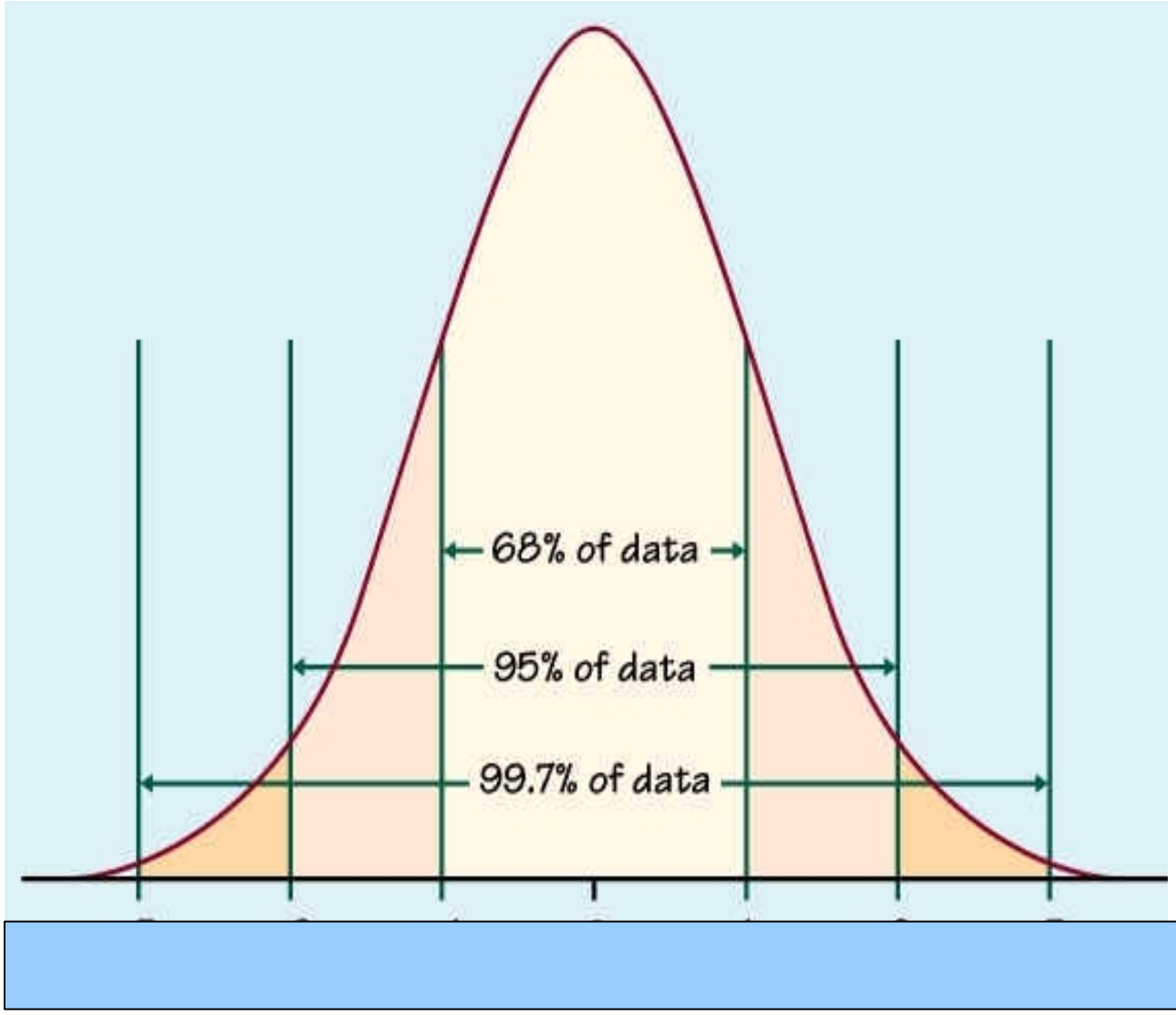
Score of a High Scoring Pair (HSP)

Alignment scores are sums of residue-pairing scores according to a Scoring Matrix



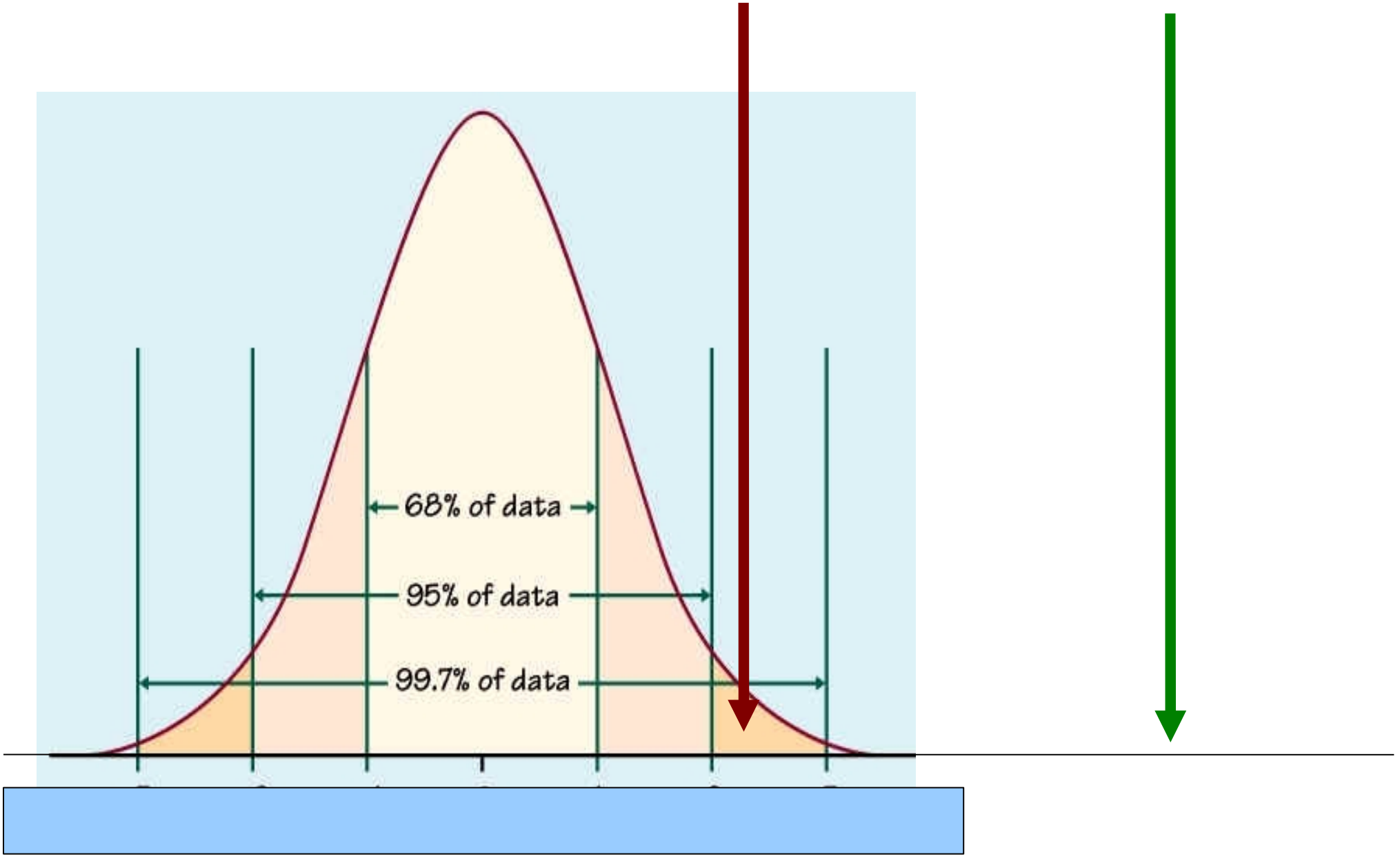
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1

Distribution of scores in comparisons of **random***-sequences



* considering the representation of the different amino acids (nucleotides) in a DataBase

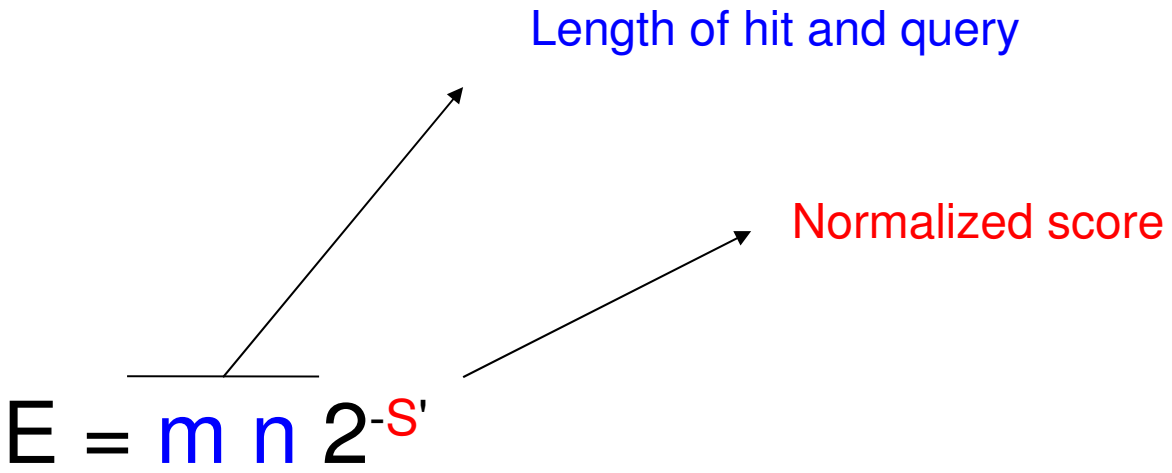
Your score



The significance of each alignment is computed as a P value or an E value

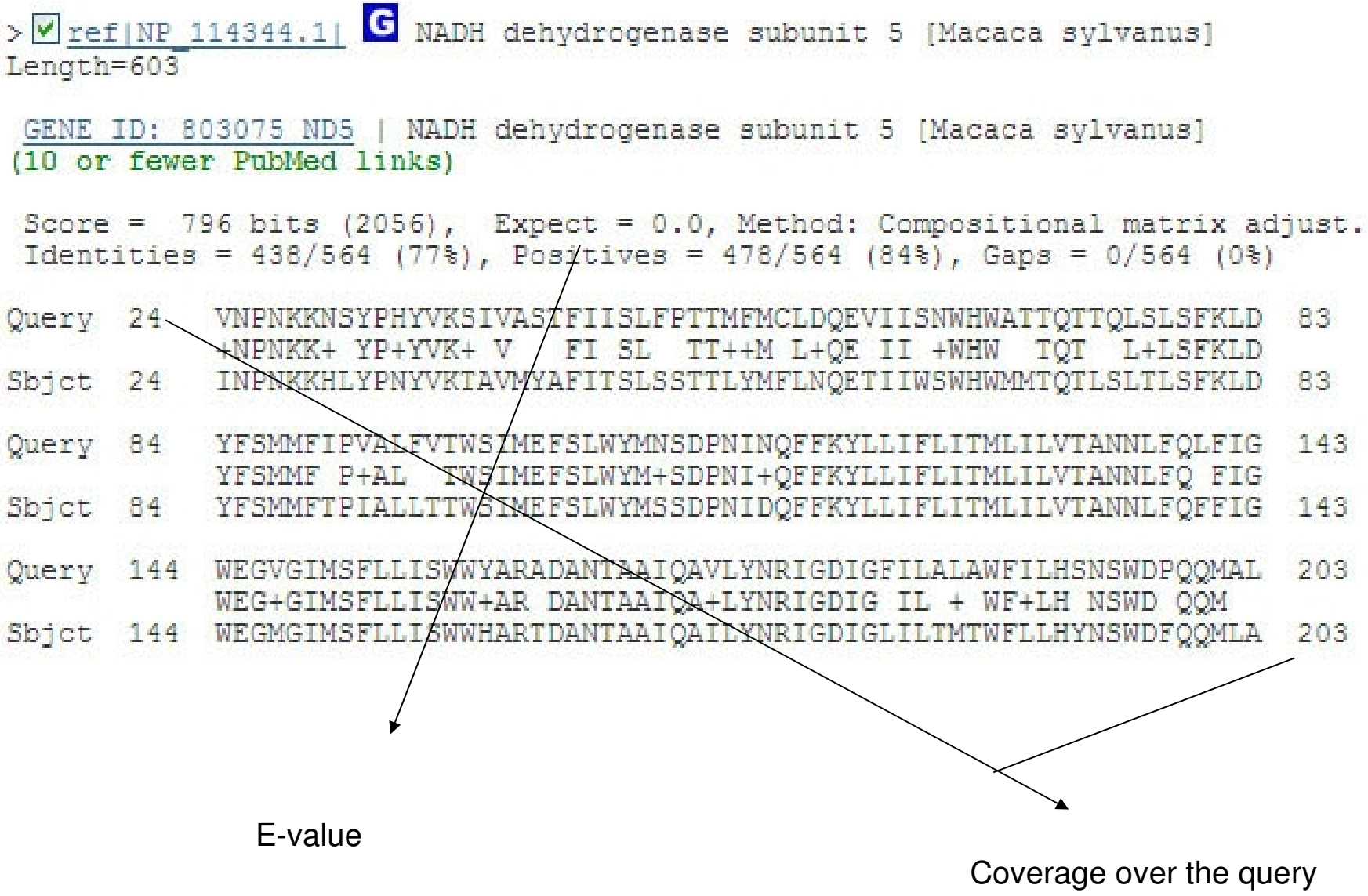
E value: Expectation value. The number of different alignents with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

P value :The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.



E-value (Expectation value)= the number of sequences that would be expected to have that **score** (or higher) if the query sequence were compared against a **database** containing unrelated sequences

E-value= ranges from 0 to the number of sequences in the DB, **and depends on the Database!!!**


> ref|NP_114344.1| NADH dehydrogenase subunit 5 [Macaca sylvanus]
Length=603
GENE ID: 803075 ND5 | NADH dehydrogenase subunit 5 [Macaca sylvanus]
(10 or fewer PubMed links)
Score = 796 bits (2056), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 438/564 (77%), Positives = 478/564 (84%), Gaps = 0/564 (0%)
Query 24 VNPNKKNSYPHYVKSIVASTFIISLFPTTMFMCLDQEVIISNWHWATTQTTQLSLSFKLD 83
+NPNKK+ YP+YVK+ V FI SL TT++M L+QE II +WHW TQT L+LSFKLD
Sbjct 24 INPNKKHLYPNYVKTAVMYAFITSLSSTTLYMFLNQETIIWSWHWMMTQTLSLTLSFKLD 83
Query 84 YFSMMFIPVALFVTWSIMEFSLWYMNSDPNINQFFKYLLIFLITMLILVTANNLFQLFIG 143
YFSMMF P+AL TWSIMEFSLWYM+SDPNI+QFFKYLLIFLITMLILVTANNLFQ FIG
Sbjct 84 YFSMMFTPIALLTTWSIMEFSLWYMSSDPNIDQFFKYLLIFLITMLILVTANNLFQFFIG 143
Query 144 WEGVGIMSFLLISWWYARADANTAAIQAVLYNRIGDIGFILALAWFILHSNSWDPQQMAL 203
WEG+GIMSFLLISWW+AR DANTAAIQA+LYNRIGDIG IL + WF+LH NSWD QQM
Sbjct 144 WEGMGIMSFLLISWWHARTDANTAAIQAILYNRIGDIGLILTMTWFLLHYNSWDFQQMLA 203
E-value
Coverage over the query

Other aspects in Blast searches

- E-value depends on database (specially important when locally searching in small databases)
- Use of Low complexity filtering
- Why multiple HSPs in a hit
- PSI-Blast, HMMER searches

End of the detour

From homology to orthology

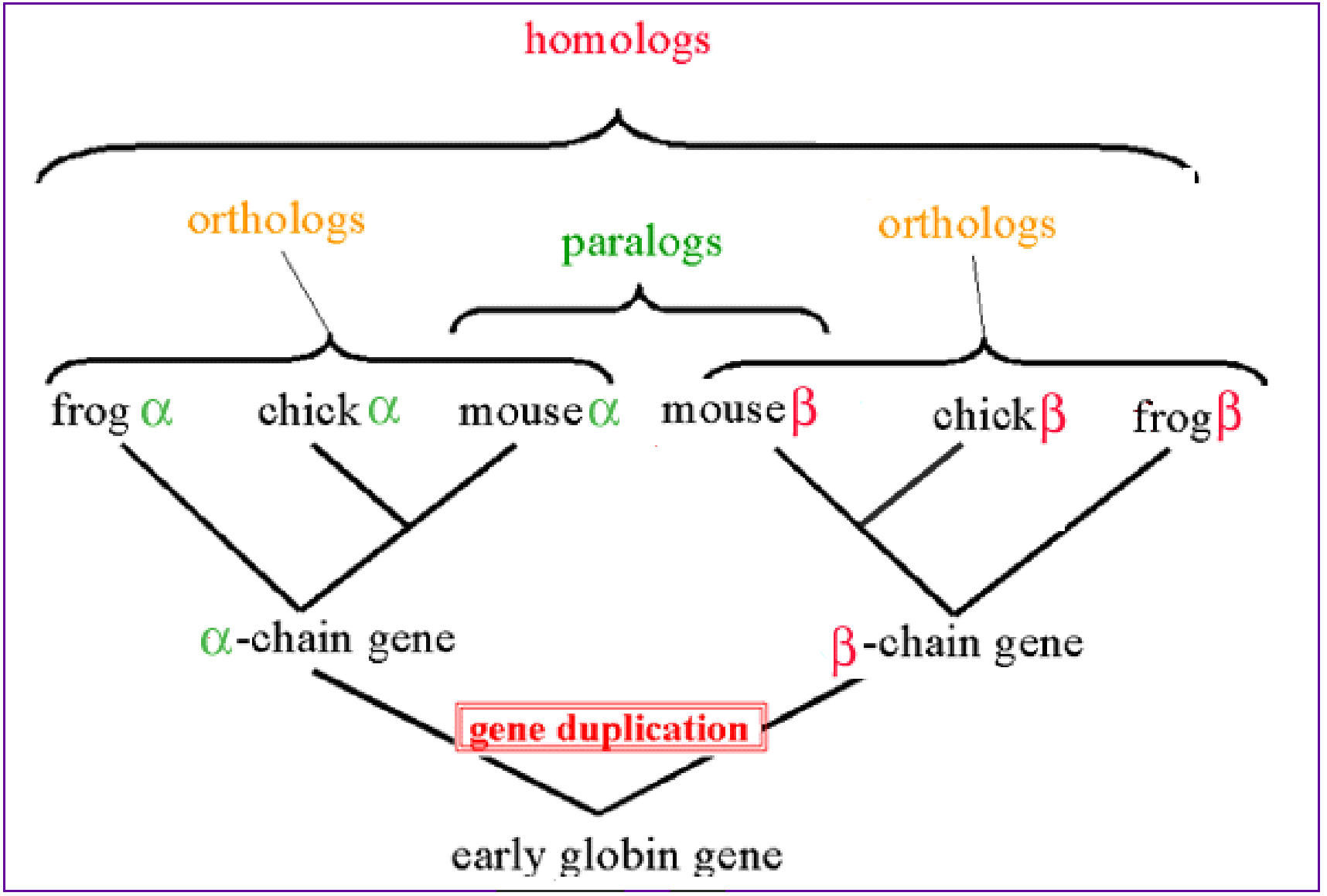
- Homologues are sequences derived from a common ancestor...

- What are then orthologues?.... and paralogues?

Original definition of orthology and paralogy by Walter Fitch (1970, Systematic Zoology 19:99-113):

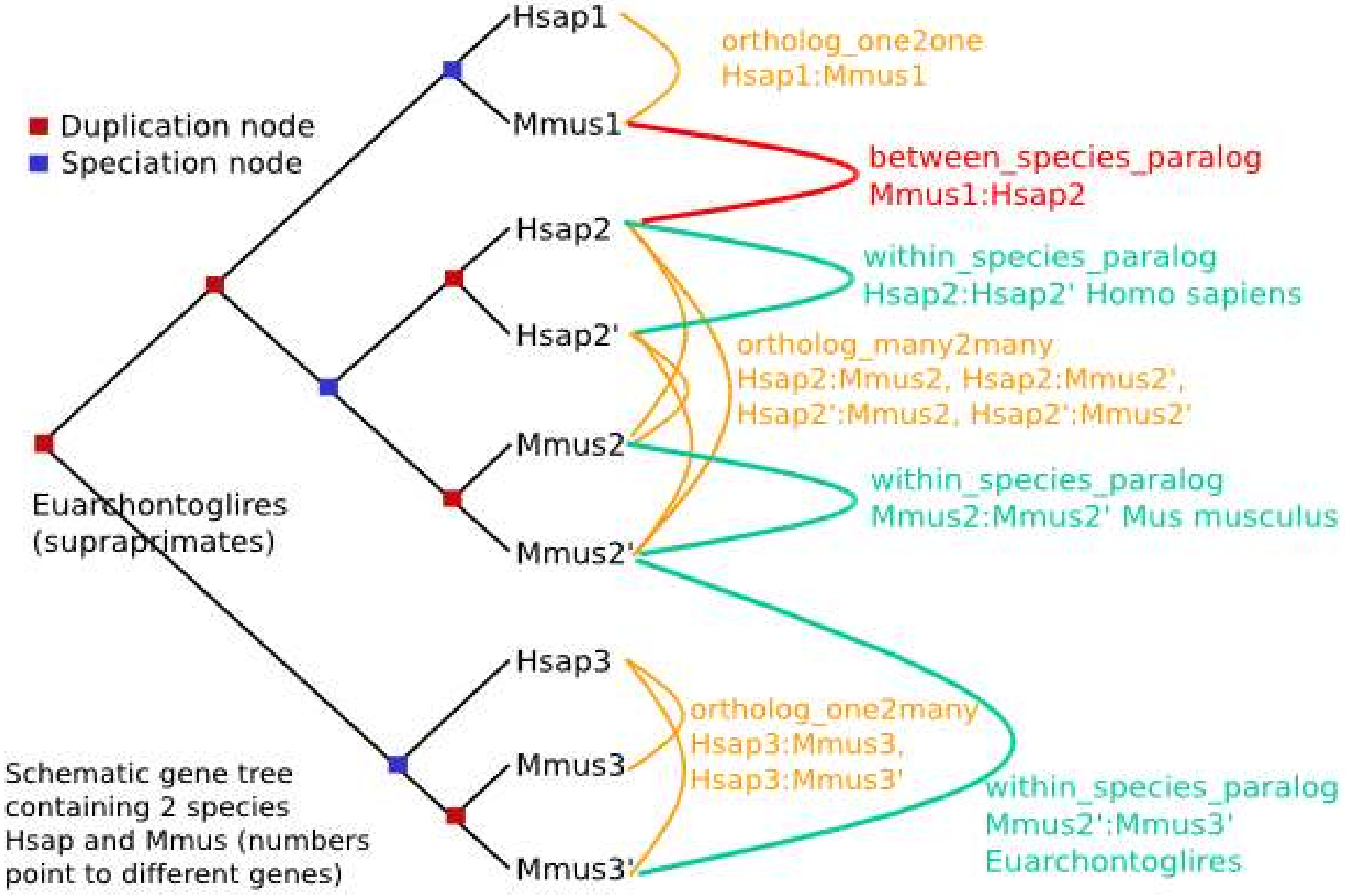
*"Where the homology is **the result of gene duplication** so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called **paralogous** (para = in parallel).*

*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact)."*


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

Corollary:

- Orthology definition is purely on evolutionary terms (not functional, not synteny…)
- Orthology/paralogy defines a pair-wise relationship between two genes
- There is no limit on the number of orthologs or paralogs that a given gene can have (when more than one ortholog exist, there is nothing such as "*the true ortholog*",)
- Many-to-Many orthology relationships do exist (co-orthology)
- No limit on how ancient/recent is the ancestral relationship of orthologs and paralogs
- Orthology is non-transitive (as opposed to homology)

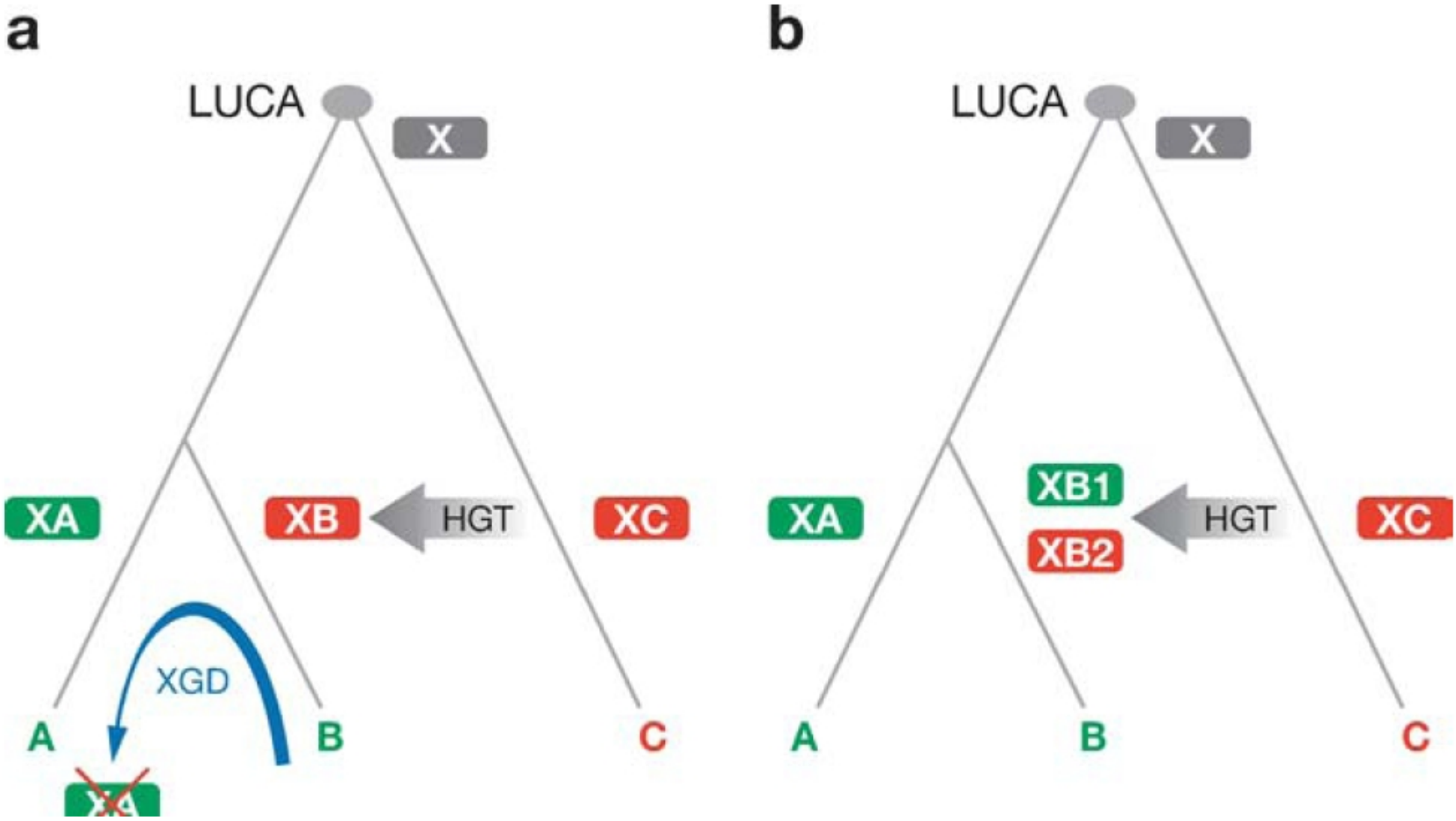

Duplication node
Speciation node
Hsap1
Mmus1
Hsap2
Hsap2'
Mmus2
Mmus2'
Hsap3
Mmus3
Mmus3'
Euarchontoglires
(supraprimates)
ortholog_one2one
Hsap1:Mmus1
between_species_paralog
Mmus1:Hsap2
within_species_paralog
Hsap2:Hsap2' Homo sapiens
ortholog_many2many
Hsap2:Mmus2, Hsap2:Mmus2',
Hsap2':Mmus2, Hsap2':Mmus2'
within_species_paralog
Mmus2:Mmus2' Mus musculus
ortholog_one2many
Hsap3:Mmus3,
Hsap3:Mmus3'
within_species_paralog
Mmus2':Mmus3'
Euarchontoglires
Schematic gene tree
containing 2 species
Hsap and Mmus (numbers
point to different genes)

Additional useful definitions

- **In-paralogs and out-paralogs** (Sohnhammer and koonin): It is defined relative to a given speciation event. In-paralogs are derived from duplications occurred subsequent to the speciation event and are therefore specific of one lineage. Out-paralogs are paralogs emerged from duplications occurred before the speciation. (Important: if you change the speciation events these relationships change)

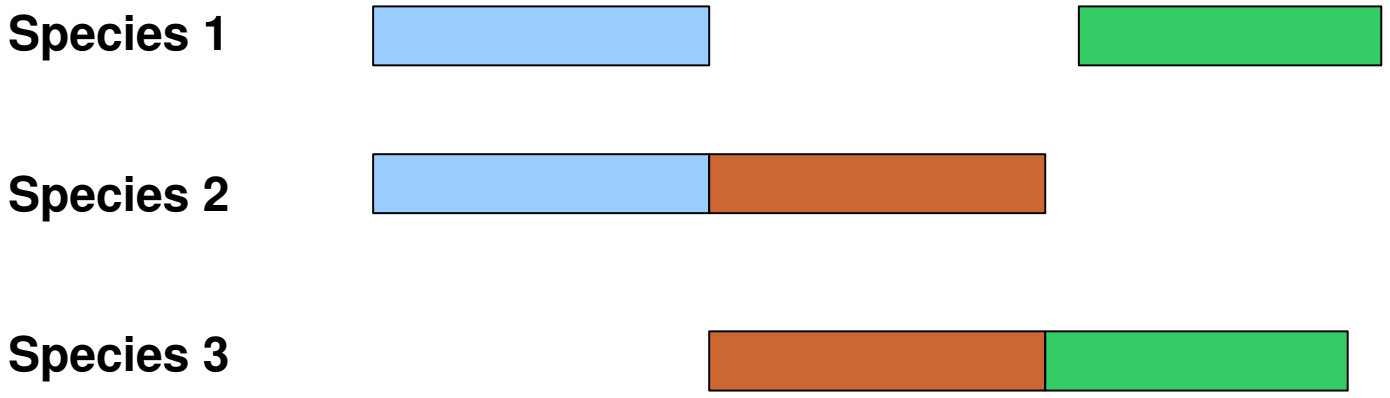
- **Orthologous group (~Orthogroup):** Also defined relative to a speciation event. It is the complete set of genes in one of the lineages formed by a speciation event. (it includes orthologs and in-paralogs, so not all the genes in an orthologous group are orthologs to each other)

The effect of HGT: Xenology and pseudoparalogy



Orthology and multi-domain proteins

- Orthology was defined at the level of genes, but this is not always the smallest level of evolution: domains do constitute smaller units of evolution, due to gene fusion/fission and recombination.



Why predicting orthology is important?

- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

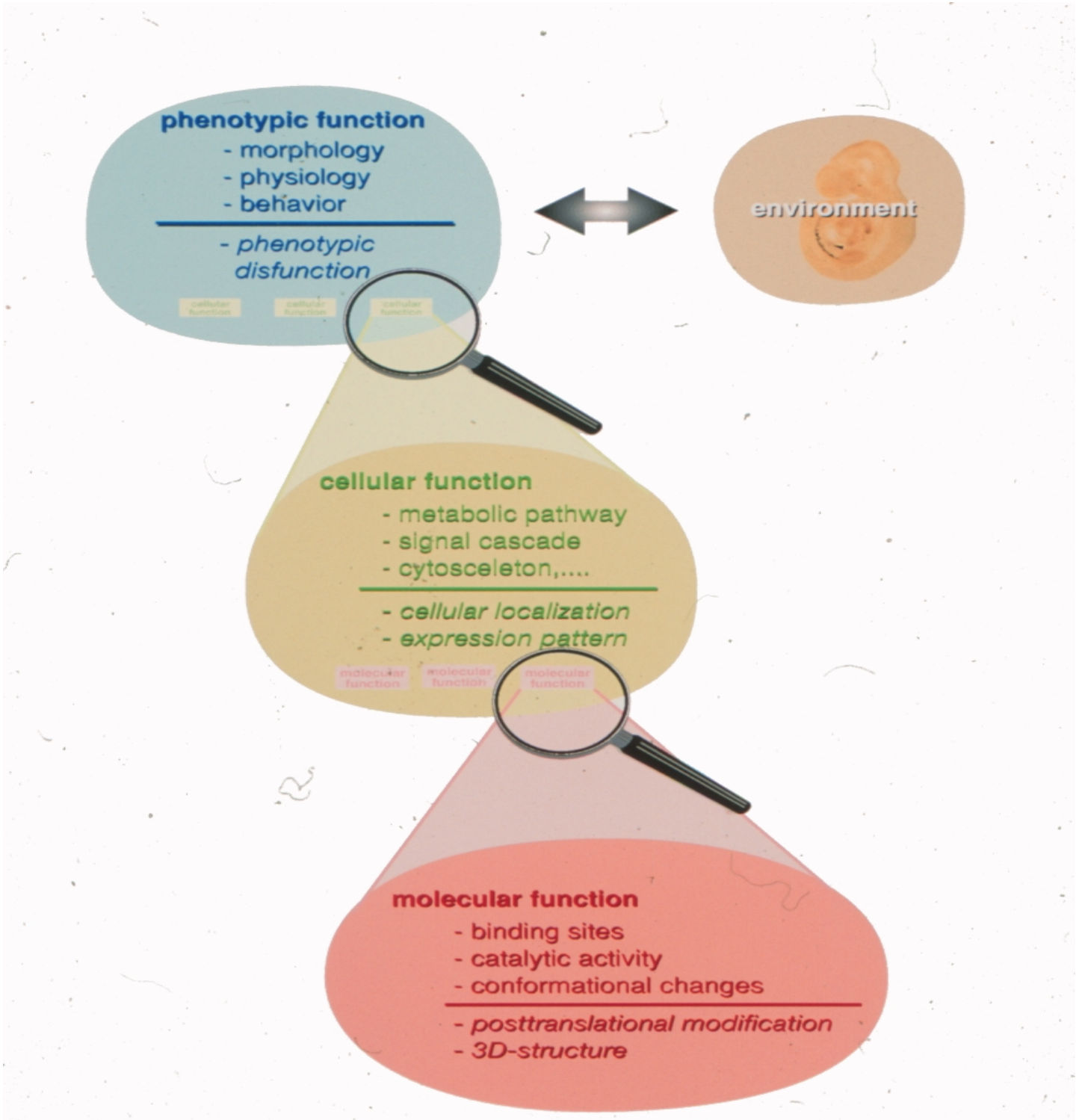
Why predicting orthology is important?

- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

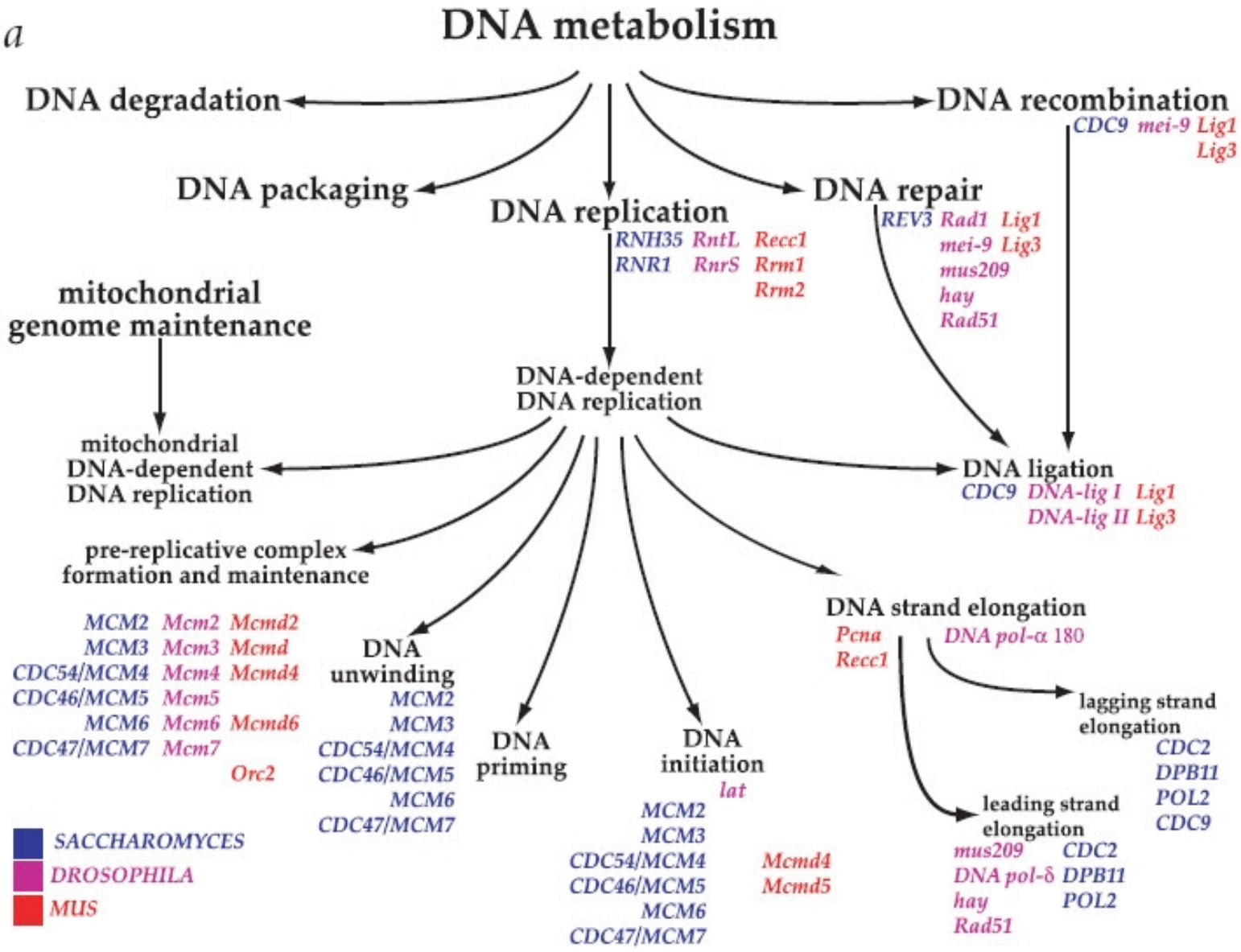
REALLY???, IS THIS TRUE IF SO, WHY IS THAT?


phenotypic function
- morphology
- physiology
- behavior
- phenotypic disfunction
environment
cellular function
- metabolic pathway
- signal cascade
- cytosceleton,....
- cellular localization
- expression pattern
molecular function
- binding sites
- catalytic activity
- conformational changes
- posttranslational modification
- 3D-structure

Accession, Term		Ontology	Qualifier	Evidence
GO:0006915 : apoptotic process	10407 gene products view in tree	biological process		ISS With UniProtKB:P04637
GO:0002326 : B cell lineage commitment	34 gene products view in tree	biological process		IEA With Ensembl:ENSMUSP0000
GO:0007569 : cell aging	878 gene products view in tree	biological process		ISS With UniProtKB:P04637
GO:0035690 : cellular response to drug	1521 gene products view in tree	biological process		IEA With Ensembl:ENSP0000

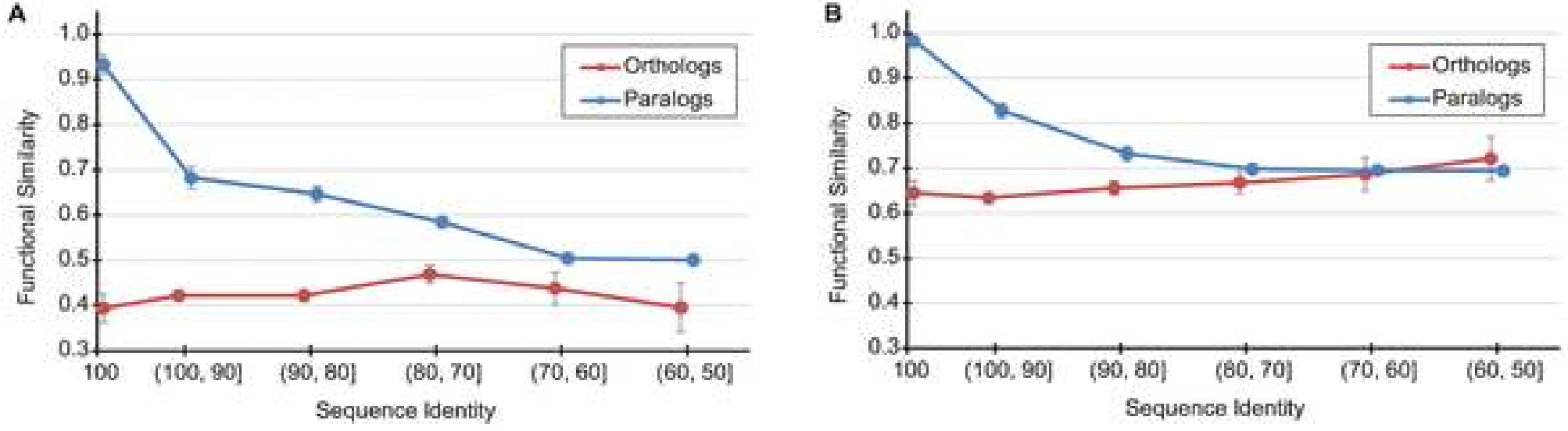
GO:0006915 (Apoptotic process)

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.


a
DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNH35 RntL Recc1
RNR1 RnrS Rrm1
Rrm2
DNA repair
REV3 Rad1 Lig1
mei-9 Lig3
mus209
hay
Rad51
DNA recombination
CDC9 mei-9 Lig1
Lig3
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
pre-replicative complex formation and maintenance
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
DNA initiation
lat
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
DNA ligation
CDC9 DNA-lig I Lig1
DNA-lig II Lig3
DNA strand elongation
Pcna
Recc1
DNA pol-α 180
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
mus209 CDC2
DNA pol-δ DPB11
hay POL2
Rad51
SACCHAROMYCES
DROSOPHILA
MUS

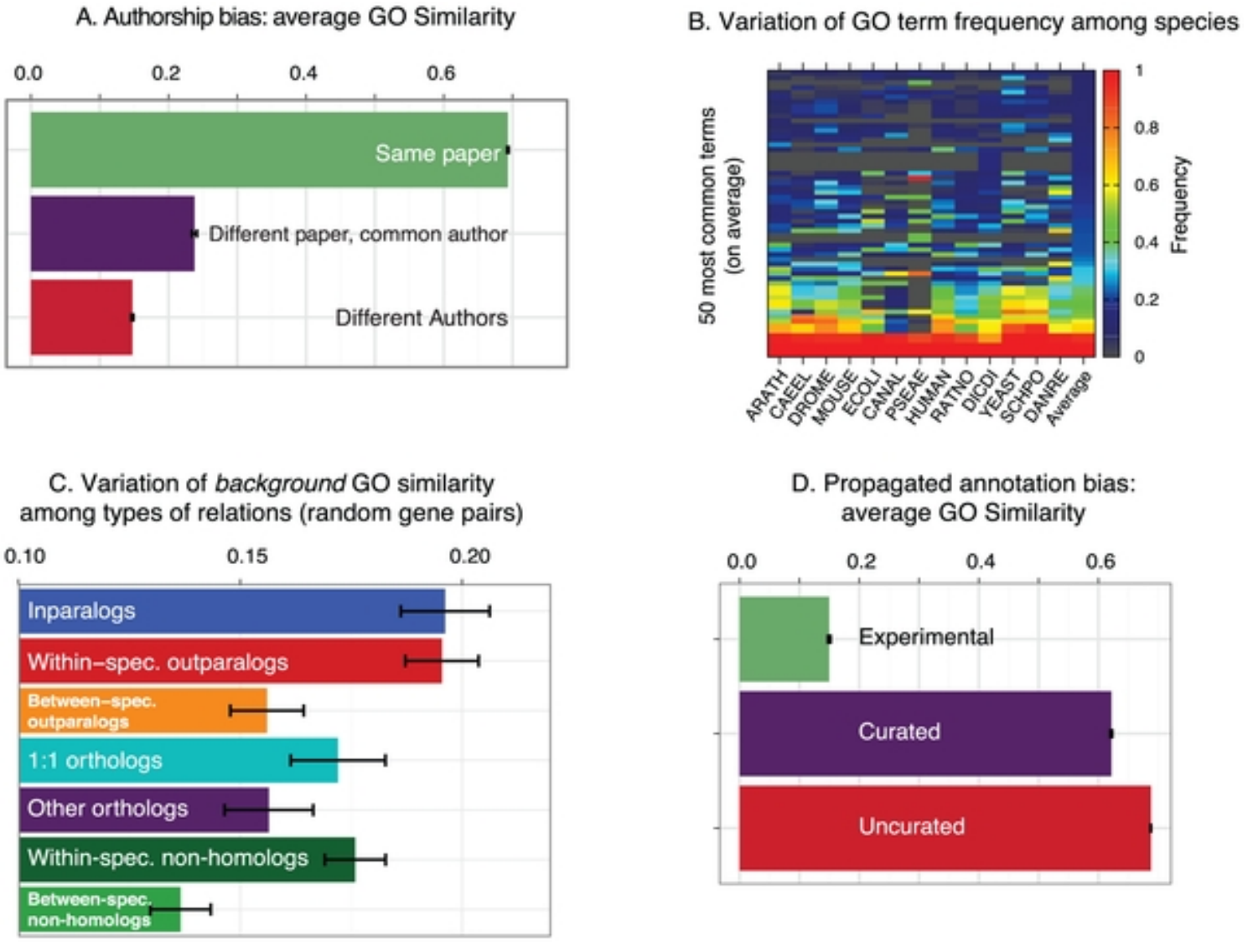
Do orthologs have more similar GO terms than paralogs?

Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue).



Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. PLoS Comput Biol 7(6): e1002073. doi:10.1371/journal.pcbi.1002073
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002073

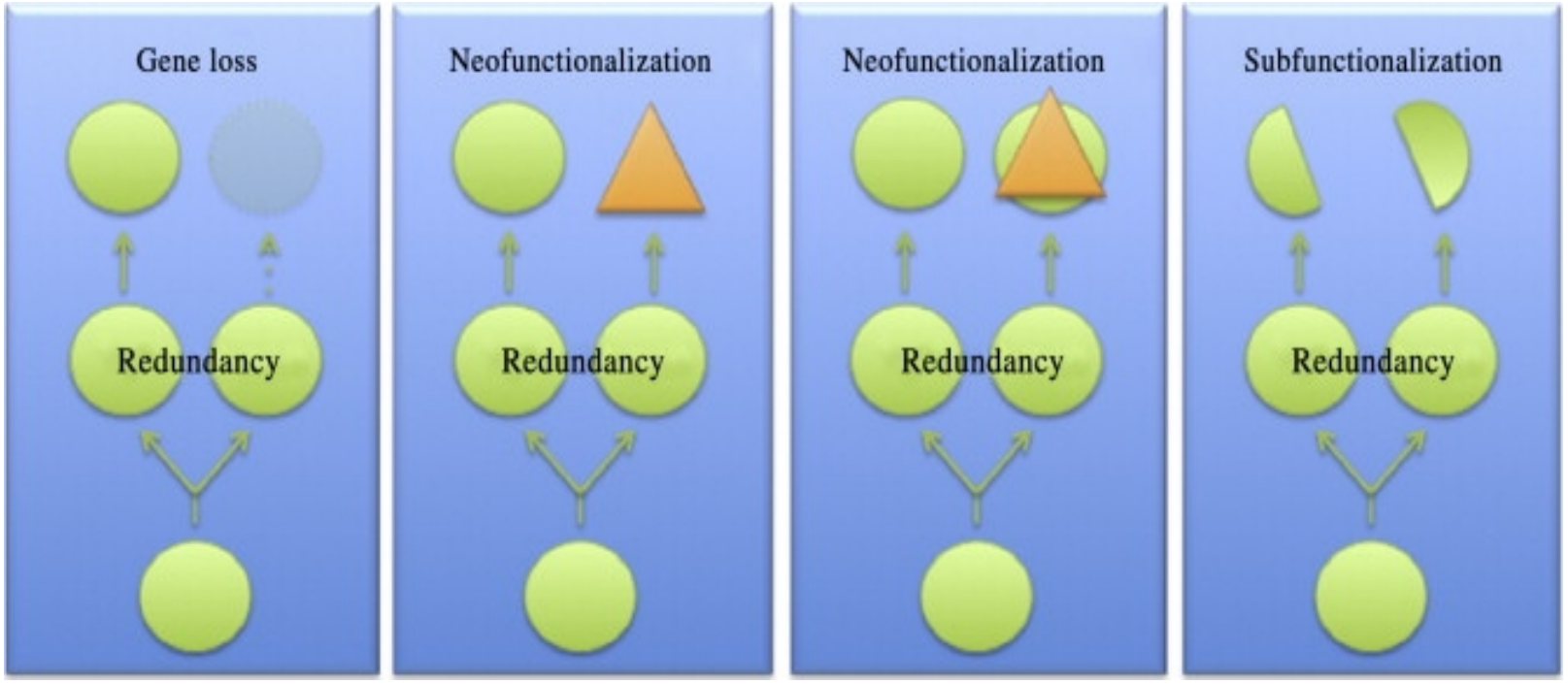
Figure 1. Potential confounding factors in GO analyses.

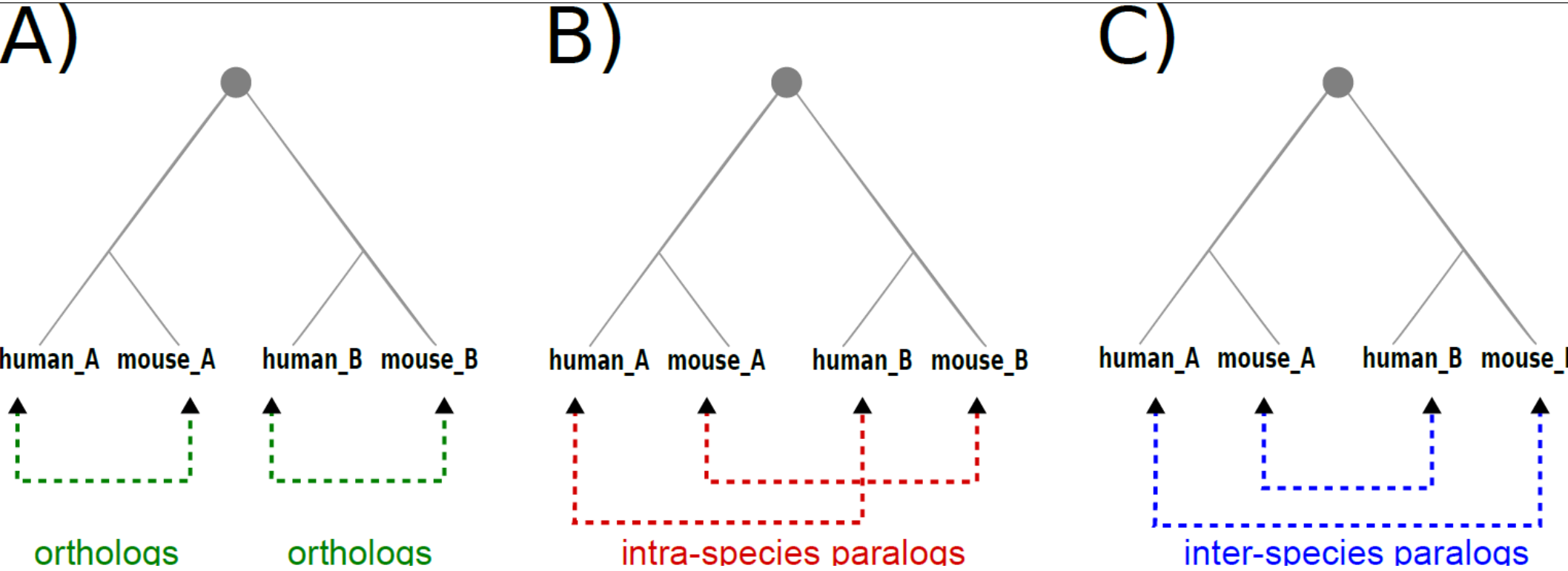


Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Comput Biol 8(5): e1002514. doi:10.1371/journal.pcbi.1002514
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002514

Orthologs do **tend** to have a more similar function because duplications promote functional divergence.

However, orthologs do also may vary their functions with time.



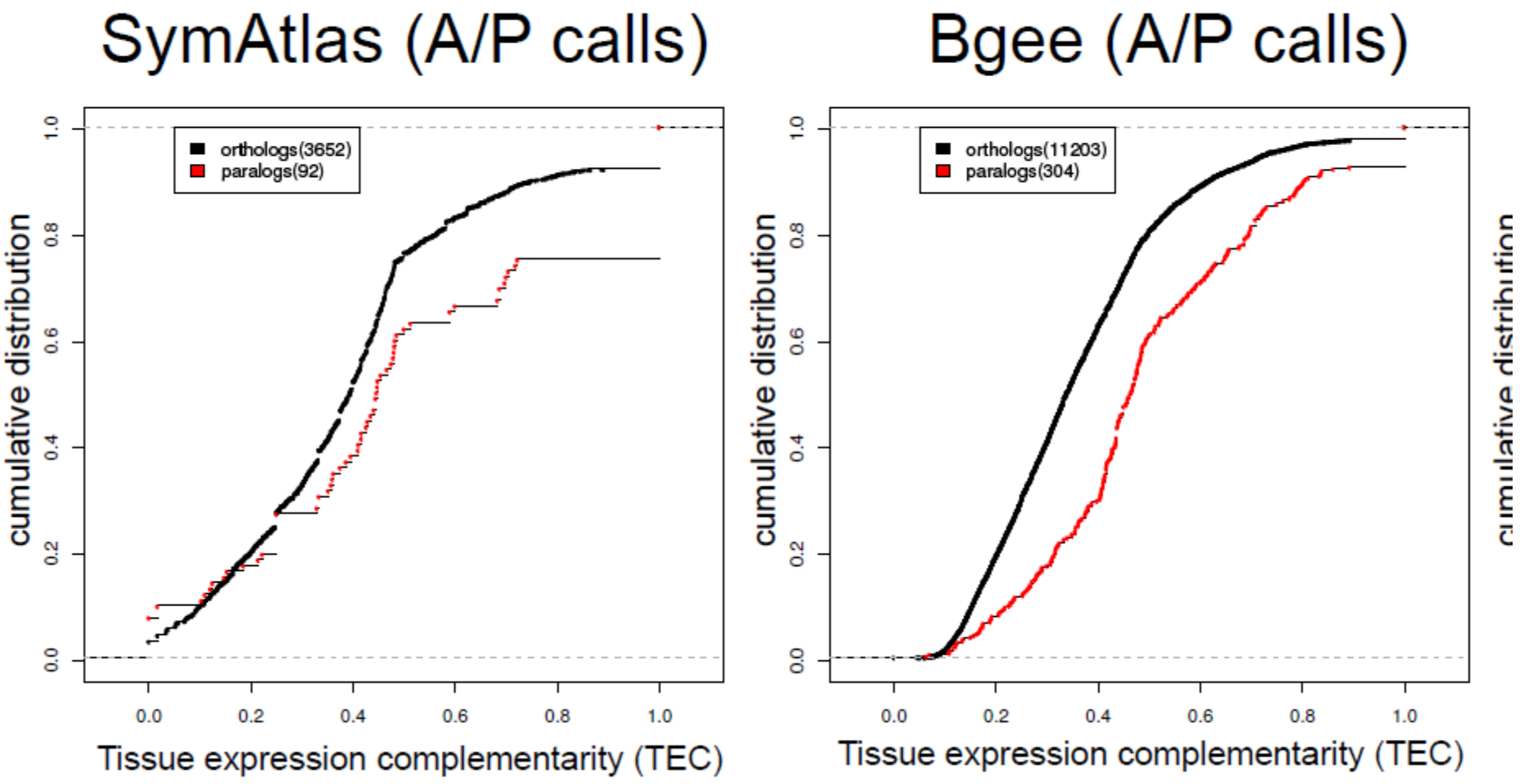


Comparison of differences in tissue-specific patterns of expression across orthologs and paralogs.

Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T.

Brief Bioinform. 2011 Sep;12(5):442-8. doi: 10.1093/bib/bbr022


SymAtlas (A/P calls)
orthologs(3652)
paralogs(92)
cumulative distribution
Tissue expression complementarity (TEC)
Bgee (A/P calls)
orthologs(11203)
paralogs(304)
cumulative distribution
Tissue expression complementarity (TEC)