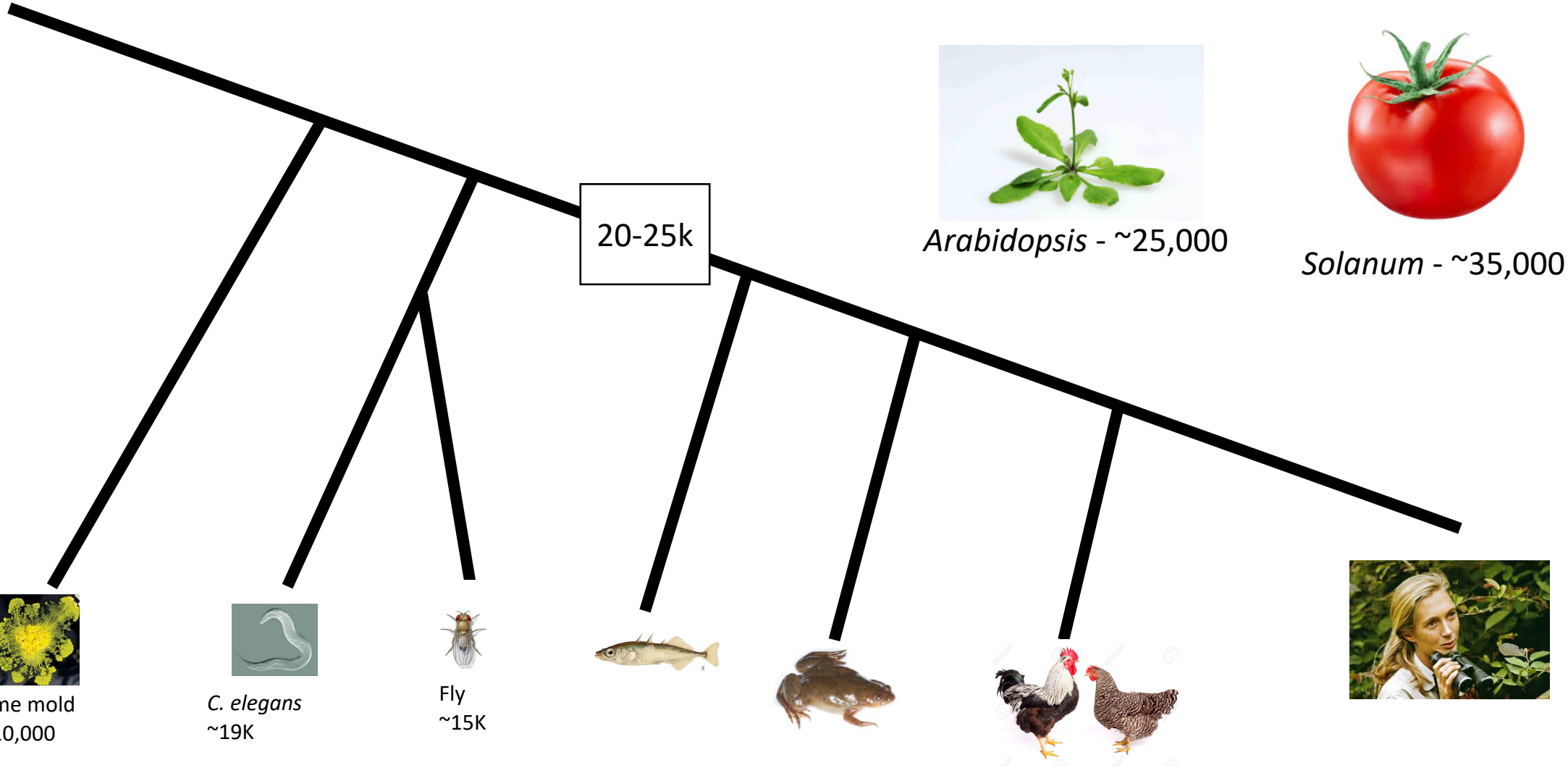
Orthology

Outline

1. ***Basic concepts***
2. ***BLAST-based approaches to orthology***
3. ***Phylogenetic-based approaches to orthology***

The number of genes in select genomes



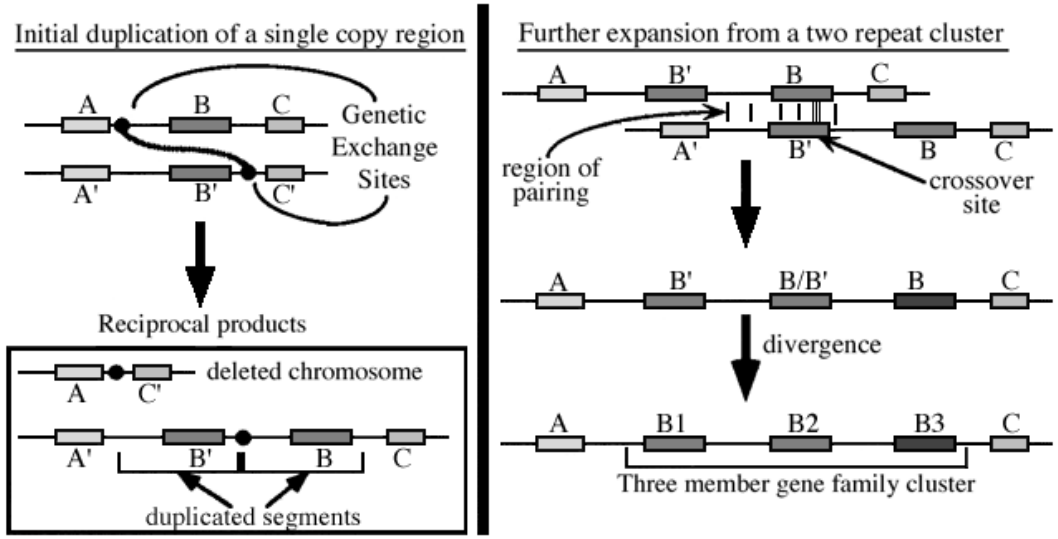
Where do new genes come from?

Gene duplication

Where do new genes come from?

Gene duplication

- mechanisms of duplication
  - **Tandem duplication** – from unequal crossing over of sister chromatids in mitosis

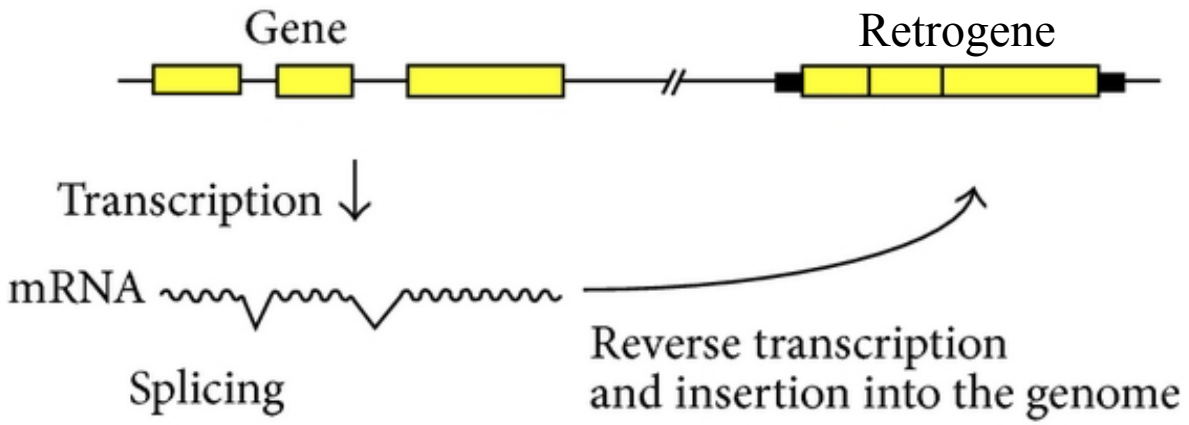


Silver. 1995 Mouse Genetics

Where do new genes come from?

Gene duplication

- mechanisms of duplication
  - Tandem duplication
  - **Retrotransposition** - ”retrogenes” arise through reverse transcription
    - results in a gene that lacks introns

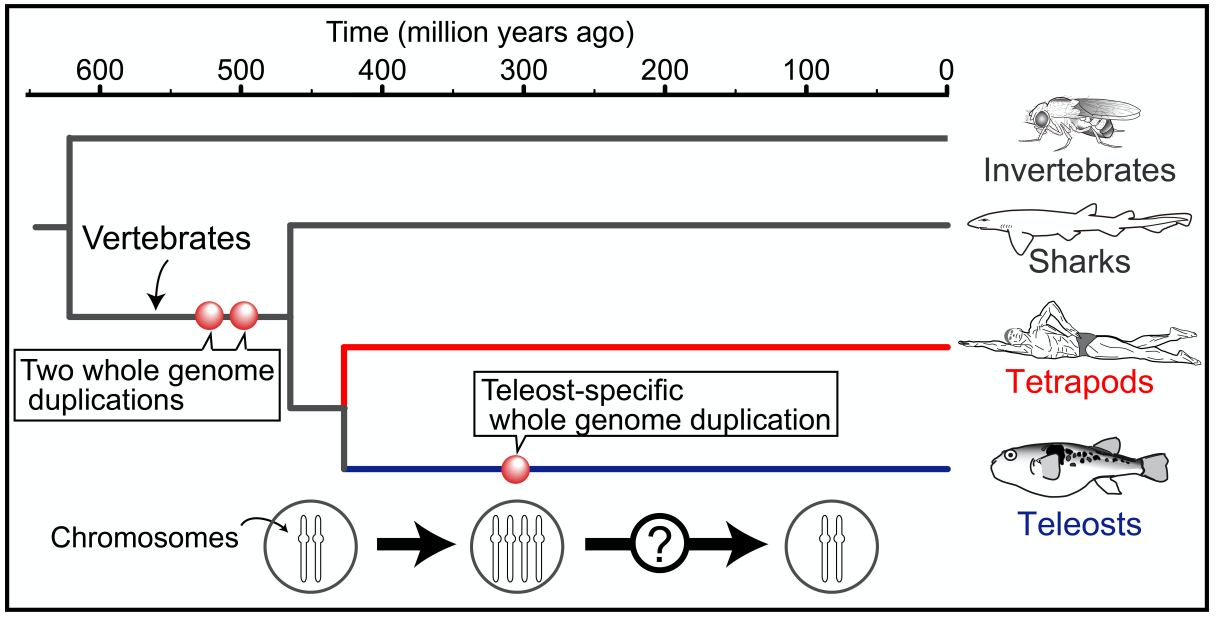


Ohshima. 2013. *Intl Journal of Evol Biol*

Where do new genes come from?

Gene duplication

- mechanisms of duplication
  - Tandem duplication
  - Retrotransposition
  - **Whole genome duplication**



Okinawa Institute of Science and Technology (2015)

Where do new genes come from?

Gene duplication

- mechanisms of duplication
  - Tandem duplication
  - Retrotransposition
  - Whole genome duplication

Formation of mosaic genes by duplication

Horizontal gene transfer

Where do new genes come from?

Gene duplication

- mechanisms of duplication
  - Tandem duplication
  - Retrotransposition
  - Whole genome duplication
- Fates of duplicated genes
  - non-functionalization: one copy becomes mutated “pseudogene”
  - redundancy: both copies retain same function
  - neofunctionalization: both copies are retained and one evolves a new function
  - subfunctionalization: both copies are retained and the original function is partitioned

Homologs

Genes with a common origin

- May be genes in the same or different organisms
- May be genes that are duplicates of one another
- Does not say that function is identical
- Can only be true or false, not a percentage

Homologs

Genes with a common origin

- May be genes in the same or different organisms
- May be genes that are duplicates of one another
- Does not say that function is identical
- Can only be true or false, not a percentage

Homologs can be:

- ***Orthologous*** – the shared evolutionary history is due to speciation
- ***Paralogous*** – the shared evolutionary history is due to gene duplication

Homologs

Genes with a common origin

- May be genes in the same or different organisms
- May be genes that are duplicates of one another
- Does not say that function is identical
- Can only be true or false, not a percentage

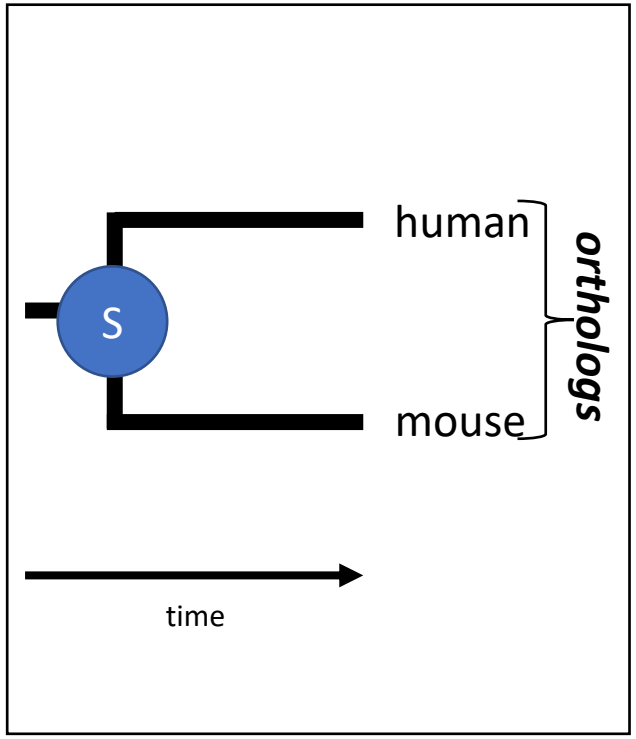
Homologs can be:

- ***Orthologous*** – the shared evolutionary history is due to speciation
- ***Paralogous*** – the shared evolutionary history is due to gene duplication

Does species A have gene X?

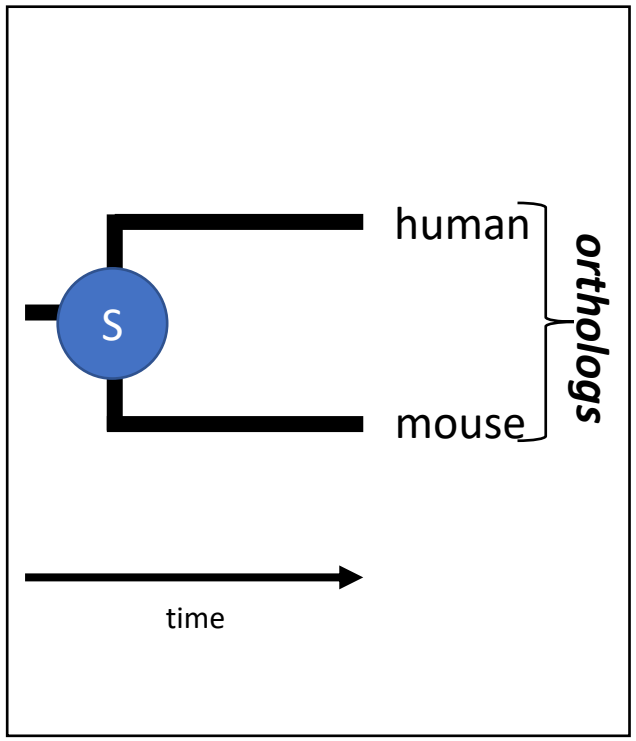
- What is really being asked is: does A have an ortholog of X

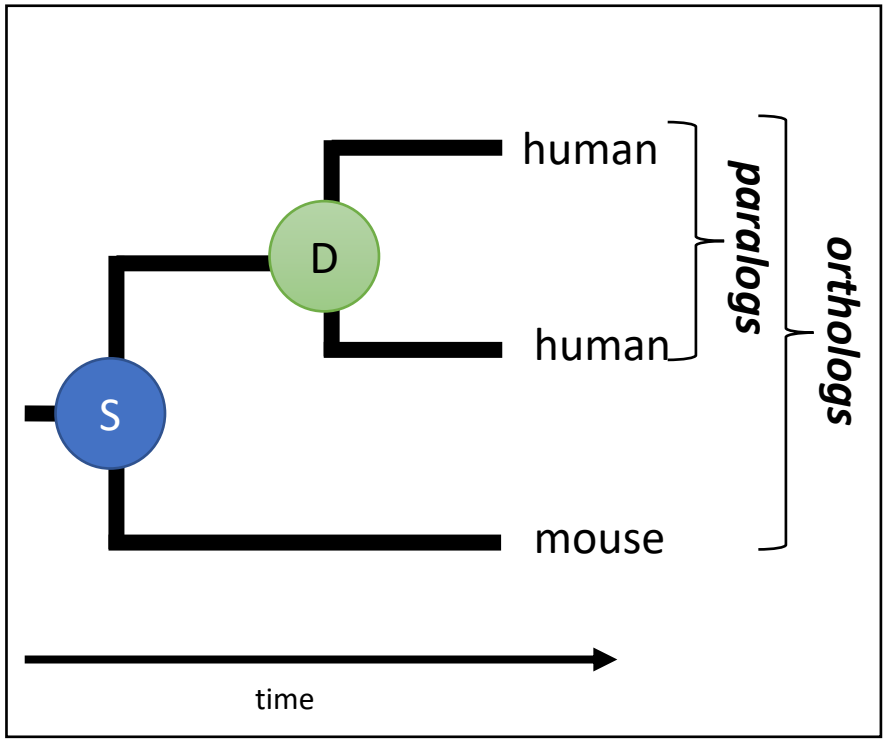
Homology Types



S = speciation D = duplication

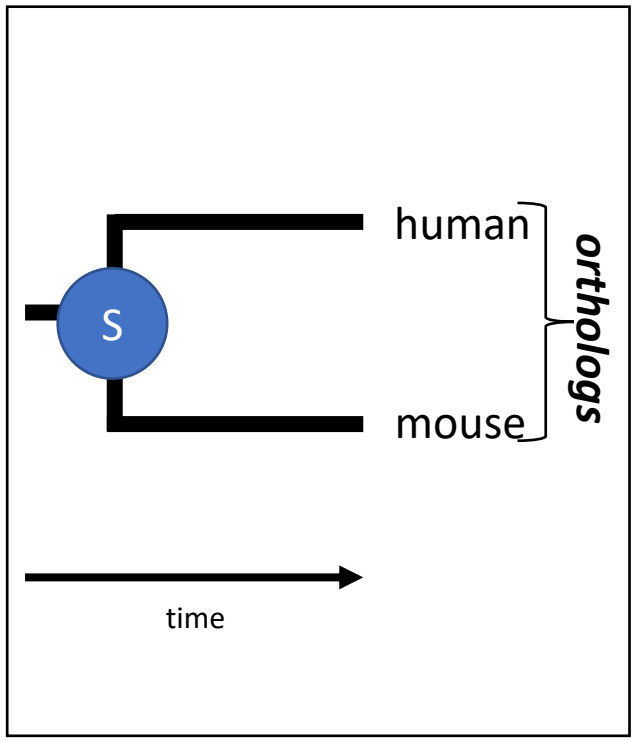
Homology Types

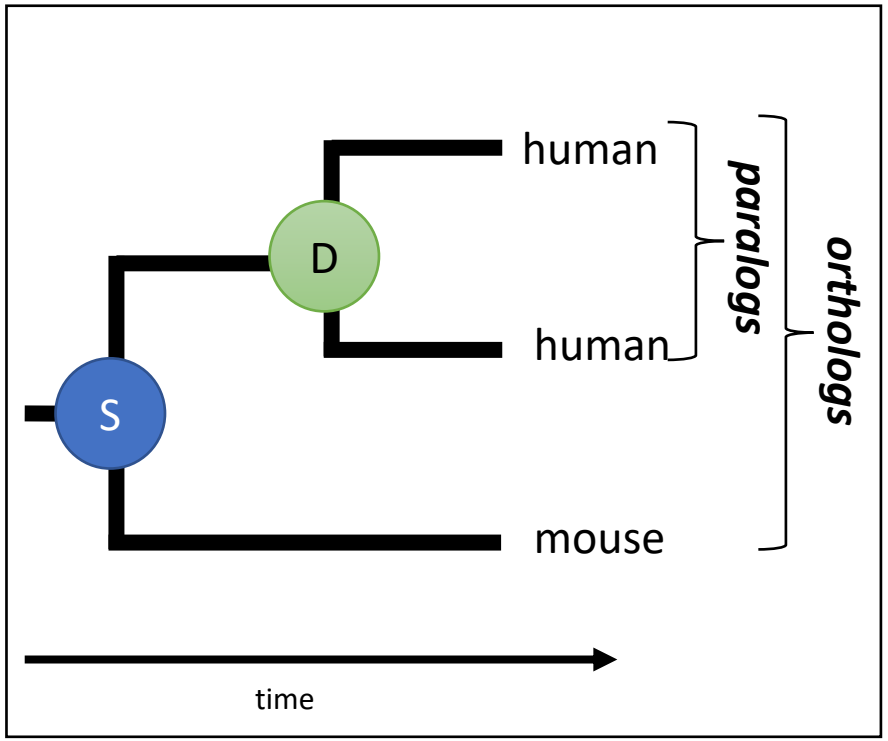


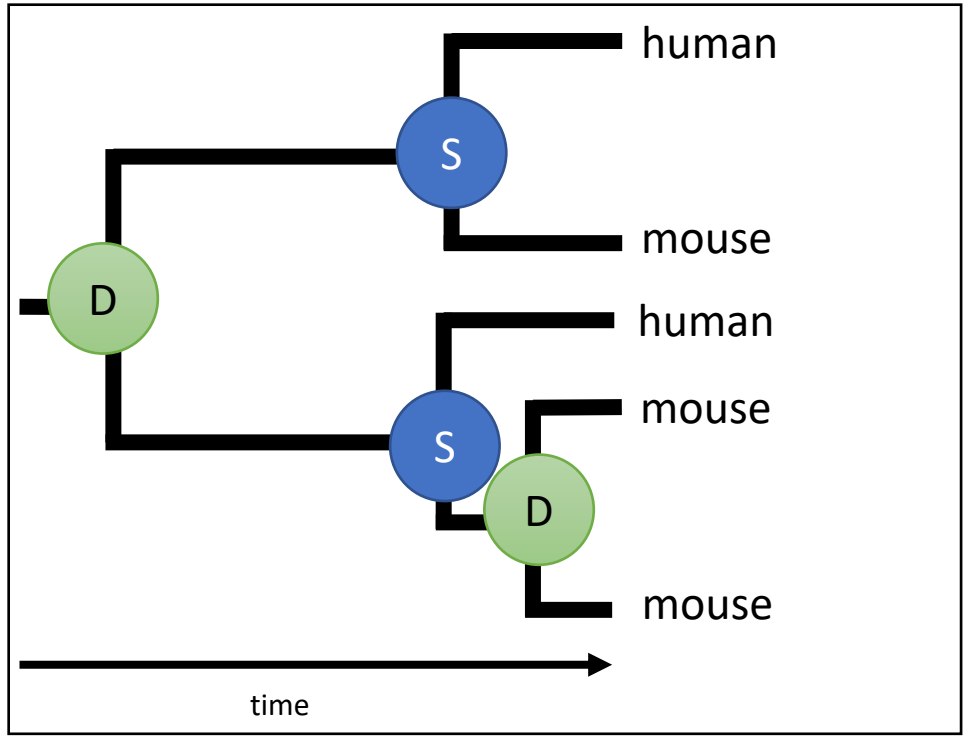


S = speciation D = duplication

Homology Types

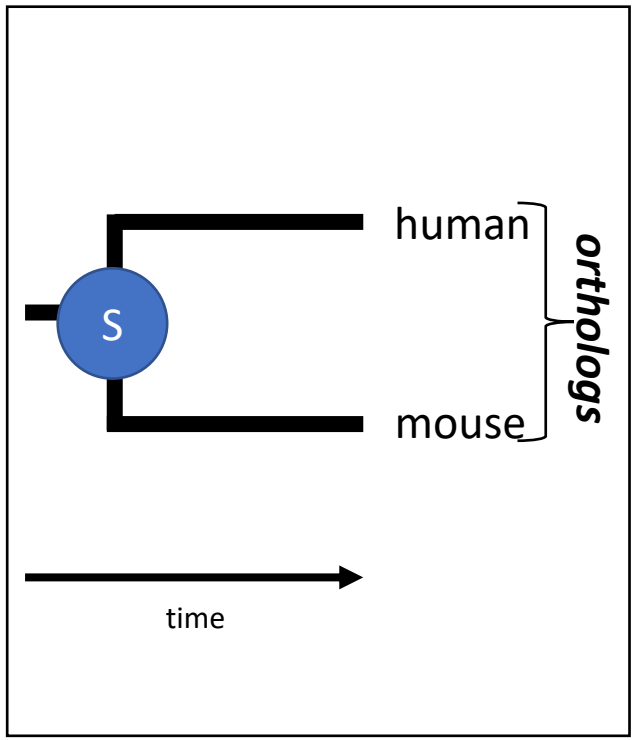


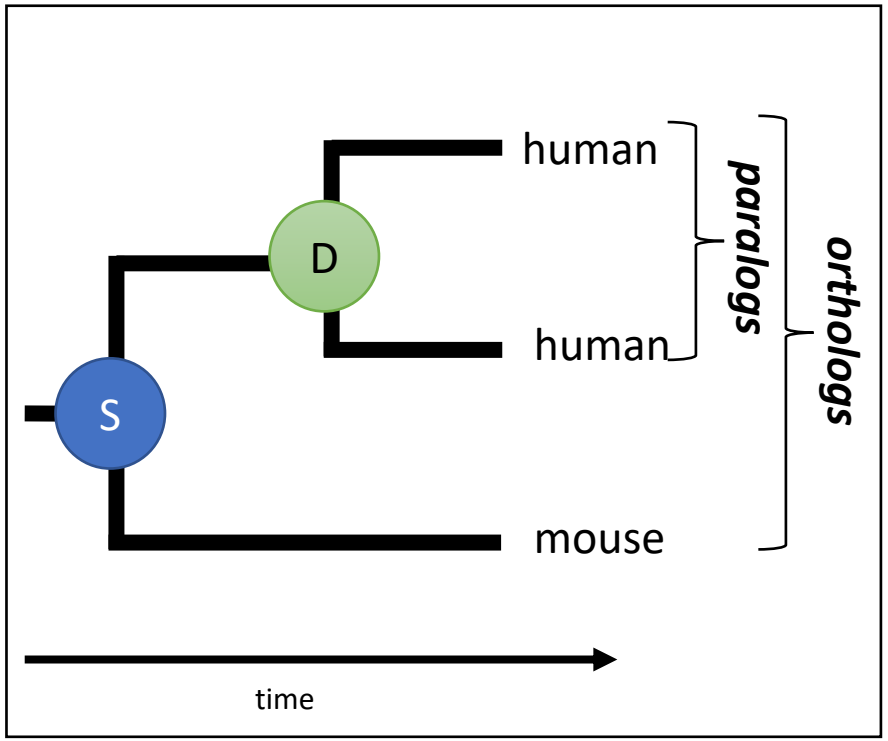


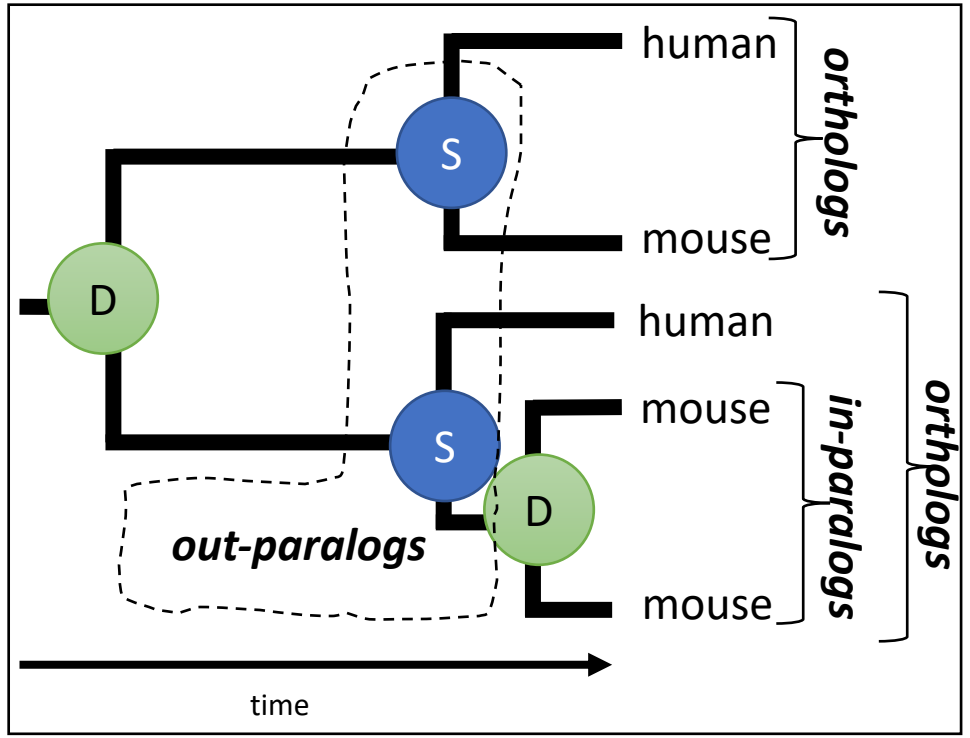


S = speciation D = duplication

Homology Types







S = speciation D = duplication

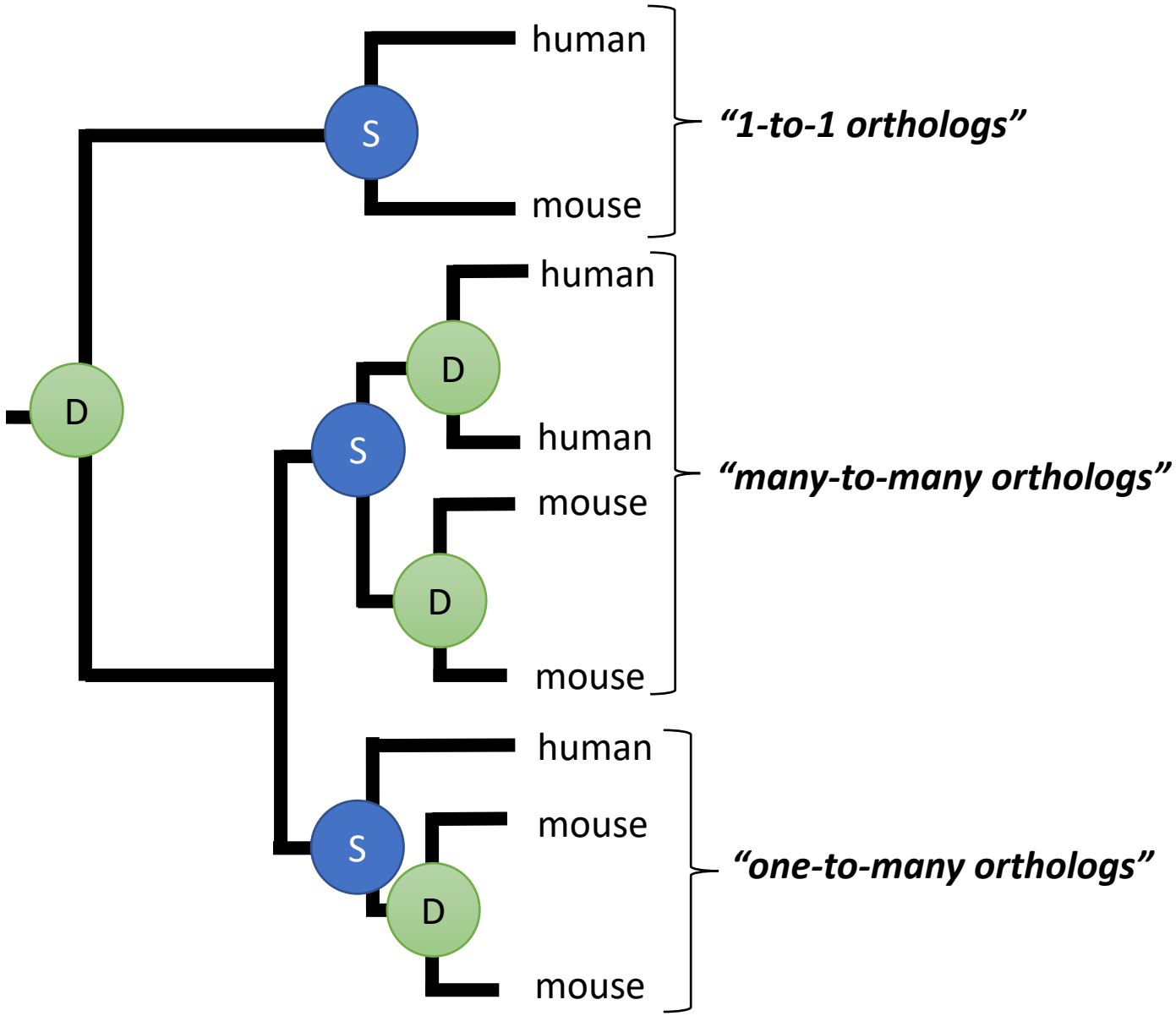
In-paralogs

- duplicated **after** the speciation event
- "extra" copies in the same genome
- are orthologs to a cluster in the other species

Out-paralogs

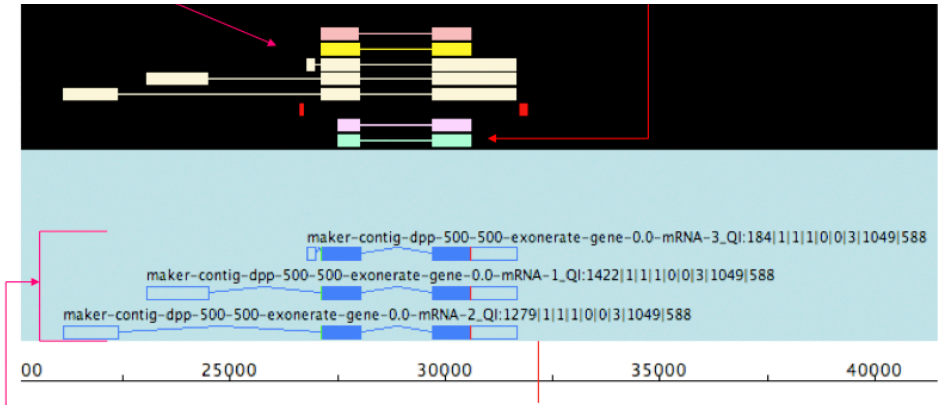
- duplicated **before** the speciation event
- not in the same genome

Homology Types: Orthologs are Separated by Speciation Events



D = duplication

S = speciation



chr7	127471196	127472363	Pos1	0	+	127471196	127472363	255,0,0
chr7	127472363	127473530	Pos2	0	+	127472363	127473530	255,0,0
chr7	127473530	127474697	Pos3	0	+	127473530	127474697	255,0,0
chr7	127474697	127475864	Pos4	0	+	127474697	127475864	255,0,0

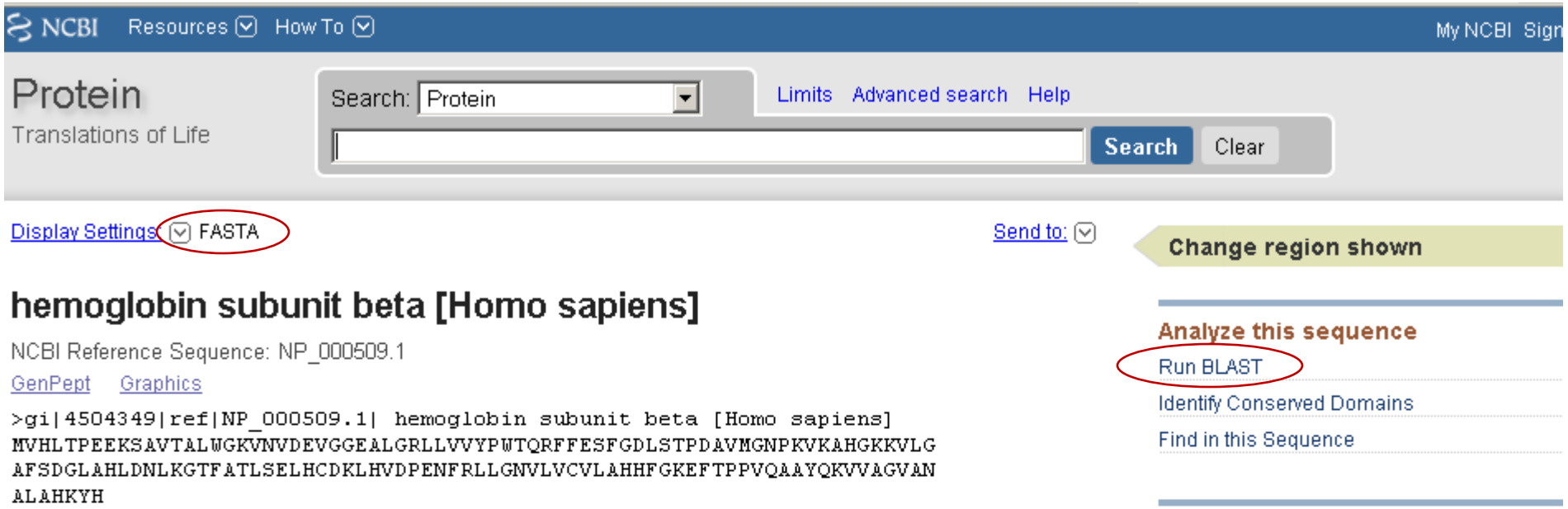
How do I know what gene this is?

Or:

How does one 'establish' orthology?

ACGGGTTCGCTACAGATGAACTGAATTTATACACGGACAACTCATCGCCCATTTGGGCGTGGGCACCGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTCATAGATATTTCTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG
ATCTGGATCCGAGTCAGAAATACGAGTTAATGCAAATTTACGTAGACCGGTGAAAACACGTGCCATGGGTTGCGT
AGACCGTAGTCAGAAGTGTGGCGCGCTATTCGTACCGAACCGGTGGAGTATACAGAATTGCTCTTCTACGACGTA
AGGAGCTCGGTCCCCAATGCACGCCAAAAAAGGAATAAAGTATTCAAACTGCGCATGGTCCCTCCGCCGGTGGCA
CTATTATCCATCCGAACGTTGAACCTACTTCCTCGGCTTATGCTGTCCTCAACAGTATCGCTTATGAATCGCATG
CGGCTGTGGATCTTAACGGCCACATTCTTAATTCCGACCGATCACCGATCGCCTTTCCTCGCTGGTACAATGAGT
ACTAAGTTATCCAGATCAAGGTTTGAACGGACTCGTATGACATGTGTGACTGAACCCGGGAGGAAATGCAGAGAA
CTGTTTCAAGGCCTCTGCTTTGGTATCACTCAATATATTCAGACCAGACAAGTGGCAAAATTTCGTGCGCCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAATTGCGGCGAAAATAGAGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGCCGATTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG

You can use your annotated genes as queries in BLAST searches….



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Choose your BLAST adventure:

blastn (nucleotide BLAST)

blastp (protein BLAST)

blastx (translated BLAST)

tblastn (translated BLAST)

tblastx (translated BLAST)

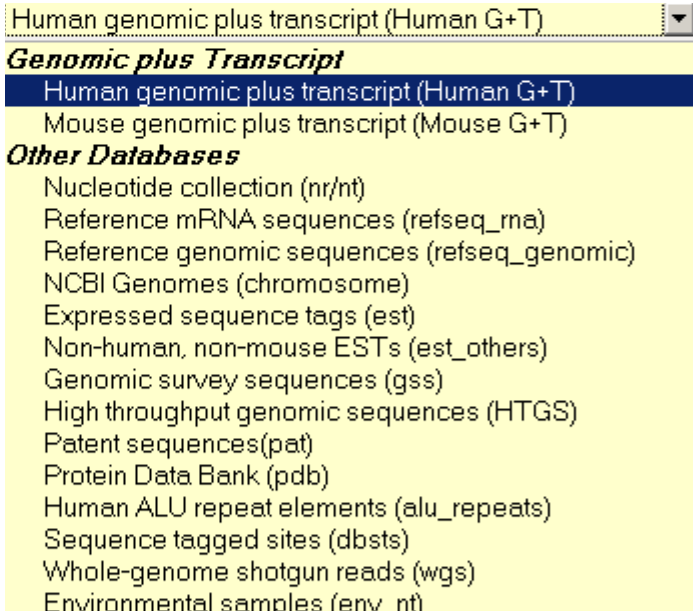
Choose the database :

nr = non-redundant (most general database)

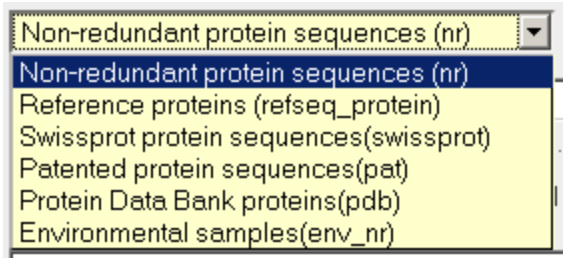
dbest = database of expressed sequence tags

dbsts = database of sequence tag sites

gss = genomic survey sequences



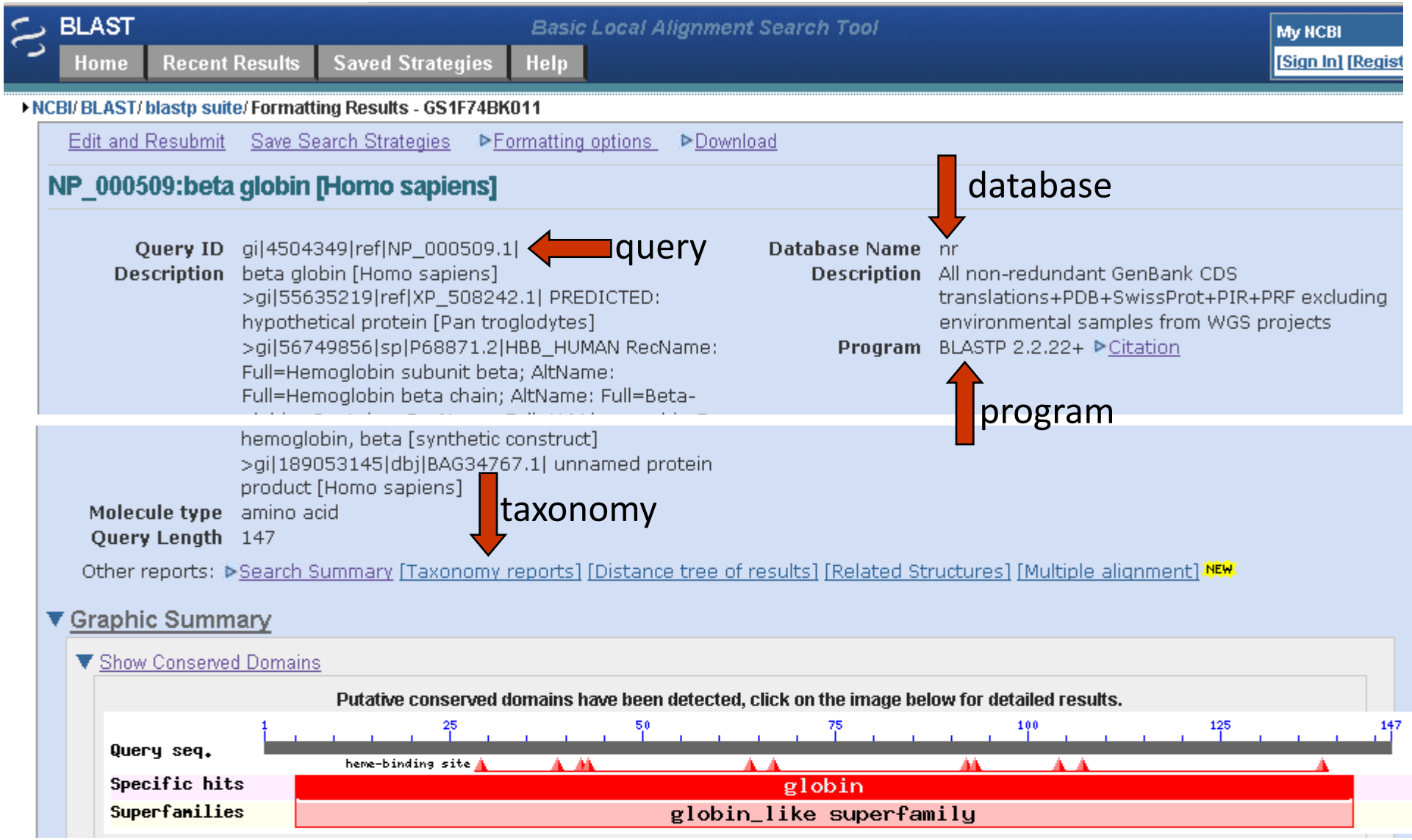
nucleotide databases



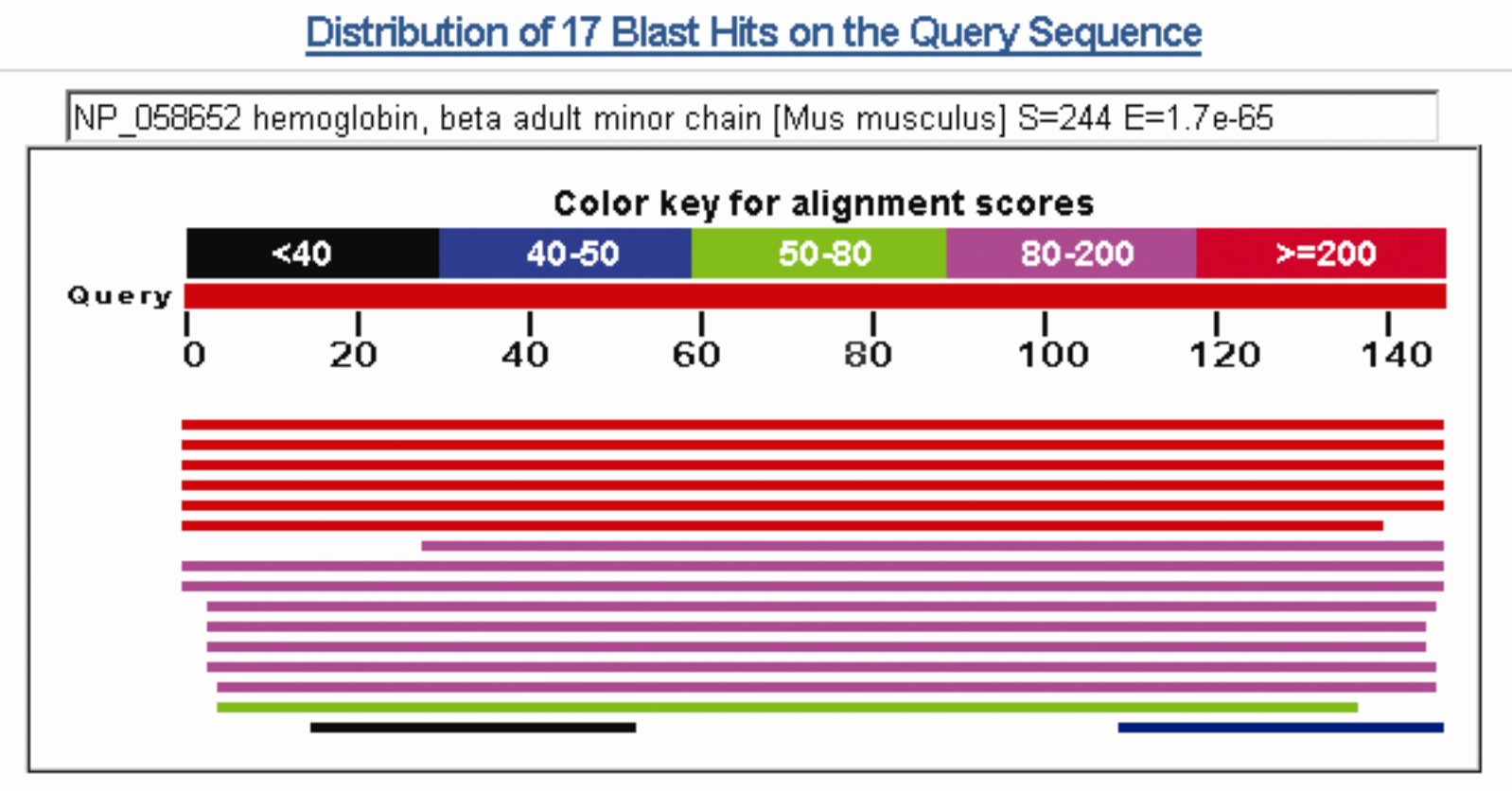
protein databases

Proteins are more conserved than nucleotides over evolution.

BLAST search output: top portion



BLAST search output: graphical output



BLAST search output: tabular output

Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value
ref|NP_058652.1|	hemoglobin, beta adult minor chain [Mus musculu	244	2e-65
ref|NP_032246.2|	hemoglobin, beta adult major chain [Mus musculu	228	2e-60
ref|XP_978992.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	226	3e-60
ref|NP_032247.1|	hemoglobin Y, beta-like embryonic chain [Mus mu	223	4e-59
ref|NP_032245.1|	hemoglobin Z, beta-like embryonic chain [Mus mu	223	6e-59
ref|XP_998314.1|	PREDICTED: similar to Hemoglobin beta-H1 sub...	203	4e-53
ref|XP_978924.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	187	2e-48
ref|XP_912634.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	161	2e-40
ref|XP_488069.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	154	3e-38
ref|NP_032244.1|	hemoglobin alpha 1 chain [Mus musculus]	105	1e-23
ref|XP_994669.1|	PREDICTED: similar to Hemoglobin alpha subun...	101	3e-22
ref|XP_356935.3|	PREDICTED: similar to Hemoglobin alpha subun...	100	4e-22
ref|NP_034535.1|	hemoglobin X, alpha-like embryonic chain in ...	94.0	4e-20
ref|NP_001029153.1|	similar to hemoglobin, theta 1 [Mus musculus	88.2	2e-18
ref|NP_778165.1|	hemoglobin, theta 1 [Mus musculus]	73.9	5e-14
ref|XP_978150.1|	PREDICTED: similar to hemoglobin, beta adult...	41.6	2e-04
ref|NP_795942.2|	5'-nucleotidase, cytosolic II-like 1 protein [M	28.9	1.5

High scores
low E values

Cut-off:
.05?
10^{-10}?

E values and Bit Scores in BLAST Results

***E (expect) value*:** The number of chance alignments with equivalent or better raw scores that are expected to occur in a database search by chance.

The lower the E value, the more significant the score.

- The E value decreases exponentially as the Score (S) that is assigned to a match between two sequences increases.
- The E value depends on the size of database and the scoring system in use.
- When the Expect value threshold is increased from the default value of 10, more hits can be reported.

Bit score: The bit score is calculated from the raw score by ***normalizing*** with the statistical variables that define a given scoring system.

- bit scores from different alignments can be compared.

BLASTing your gene to NCBI databases is a good method of finding orthologs for your gene.

- It is a form of ***FUNCTIONAL ANNOTATION.***

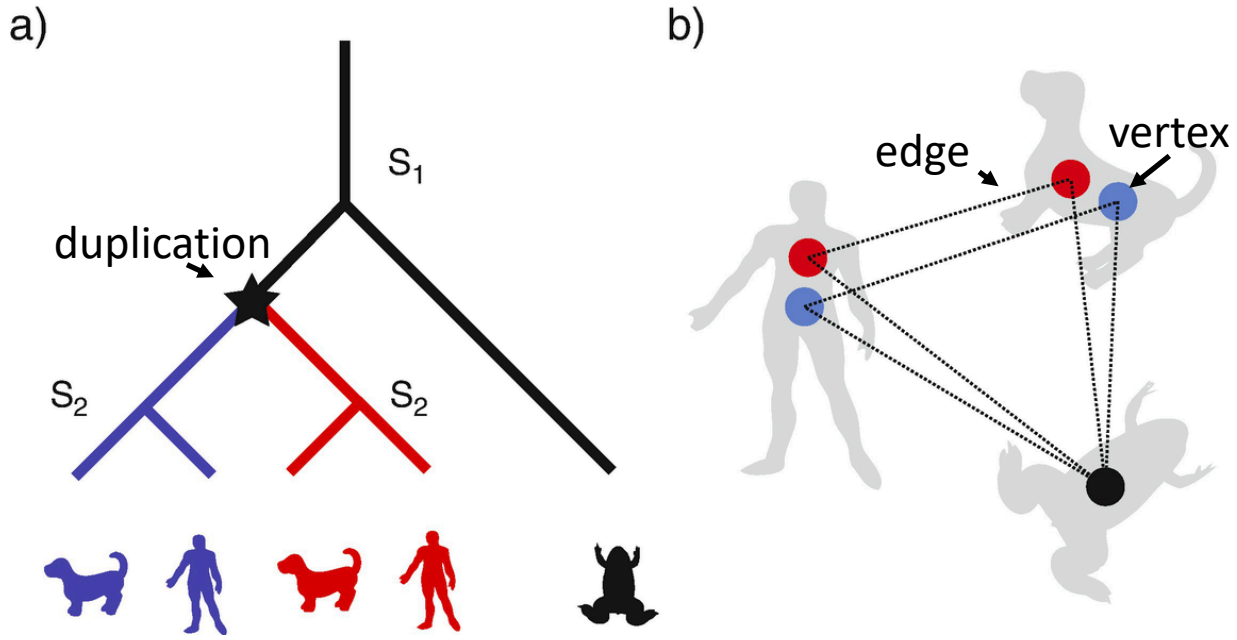
But what if you want to construct orthologous relationships between many, or all of the genes in a newly annotated genome?

We will discuss ***graph-based*** and ***tree-based*** methods.

Graph-based Methods

Two phases:

1. Graph construction – pairs of orthologous genes (vertices) are inferred and connected by edges (using alignment scores)
2. Clustering phase – groups of orthologous genes are constructed based on graph structure.



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Graph Construction

Based on the concept that between any two genomes, orthologs tend to be homologs that are the least divergent.

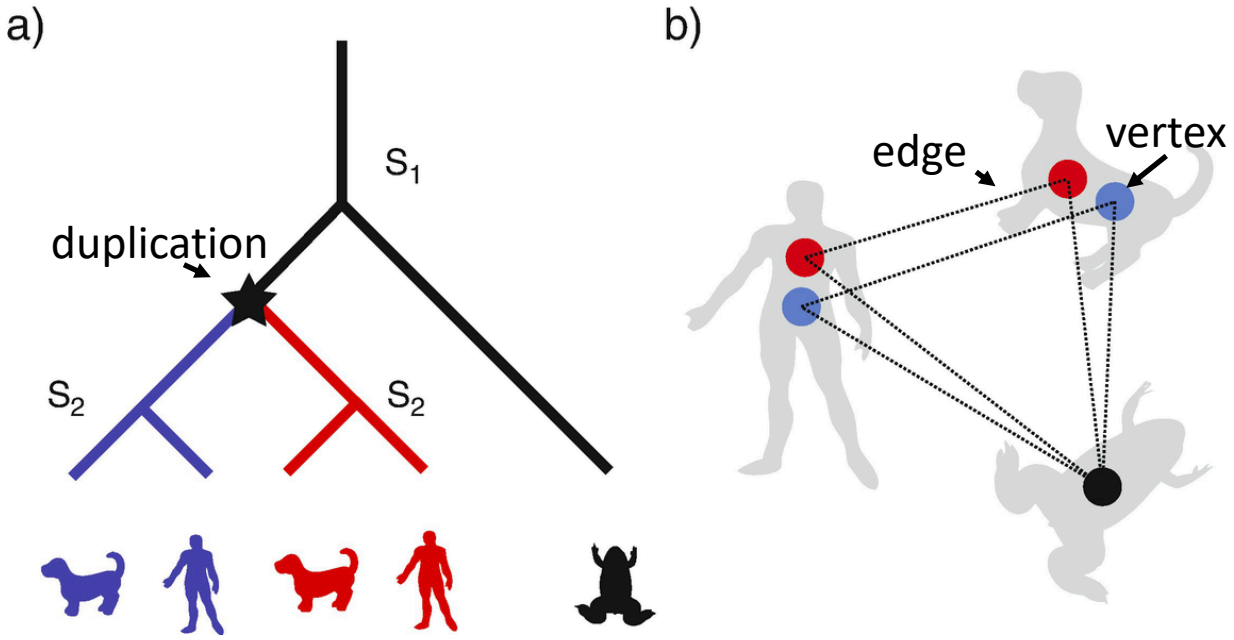
Uses sequence similarity scores as a measure of closeness.

First finds genome-wide best hit of a gene in another genome*, then determines the* reciprocal hit *in the original genome.*

Computationally efficient, because genome pairs can be processed independently and high-scoring alignments can be computed efficiently using BLAST.

Usually considered reciprocal-best BLAST hit (RBH).

Time complexity scales quadratically in terms of the number of genes.



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Clustering

From pairs to groups! This is more useful (and more accurate for establishing orthology).

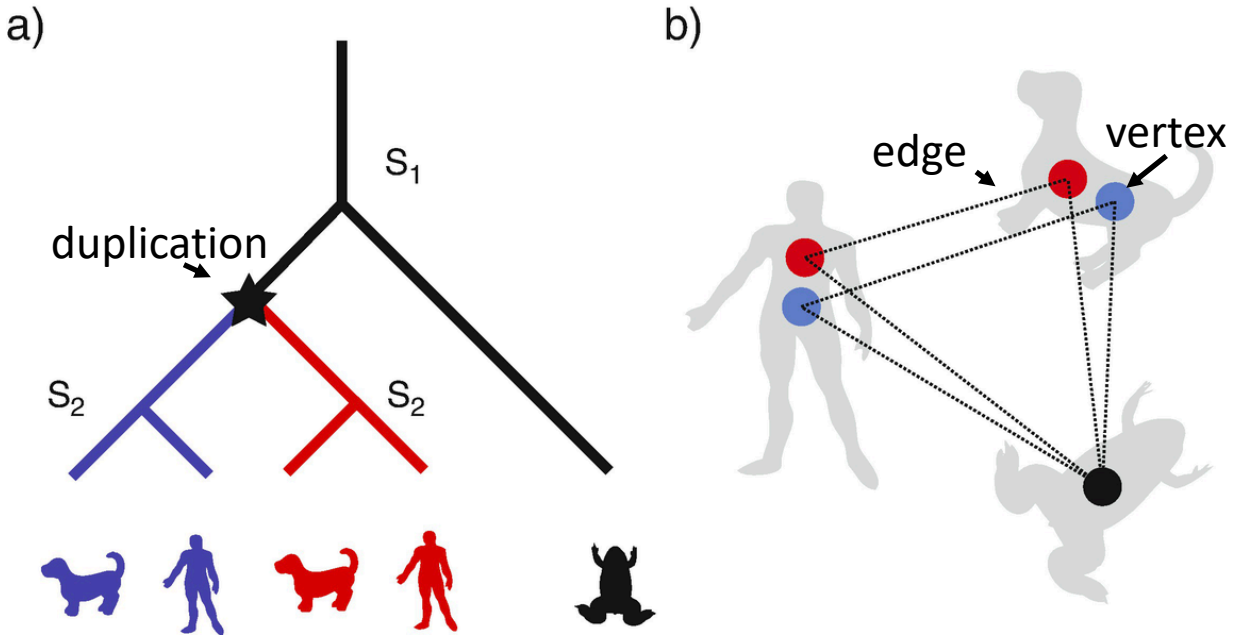
*Clusters of orthologous groups (**COGs**) – uses triangles as seeds followed by merging (Tatusov et al. 1997).*

***OrthoMCL** – Li et al. 2003 –Markov clustering simulates a random walk along the graph where edges are weighted by similarity scores.*

- *Generates a probability that two genes belong to the same cluster*
- *Graph is partitioned by probabilities to form orthologous groups*

*Hierarchical clustering – such as **OrthoDB** – creates groups with reference to speciation and common ancestry*

- *Better for determining paralogy*



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Example

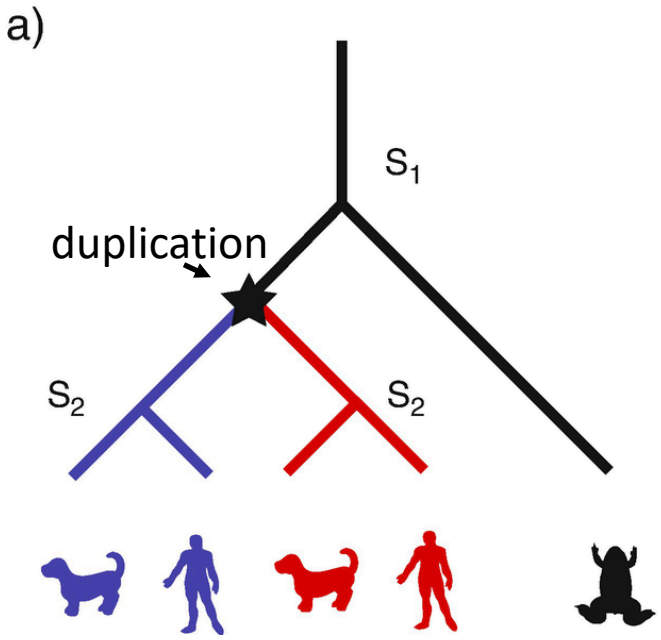
In the example, a speciation event is followed by a duplication and then another speciation event.

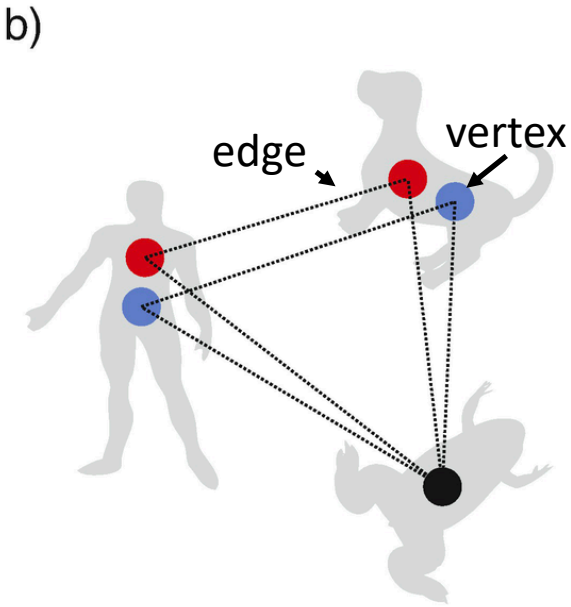
Frog gene is orthologous to all other genes. (why?)

Blue genes are orthologous to each other. (why?)

Red genes are orthologous to each other. (why?)

Red and blue genes are paralogous (why?)



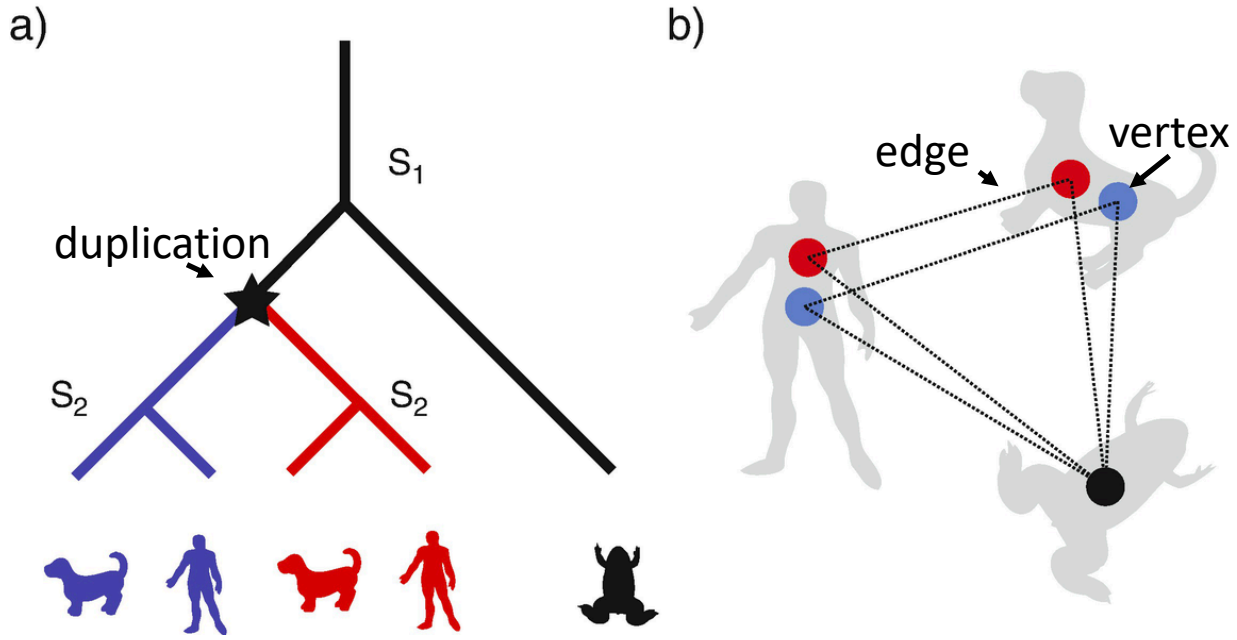


Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Example

In the corresponding graph, genes are vertices and the orthologous relationships are edges.

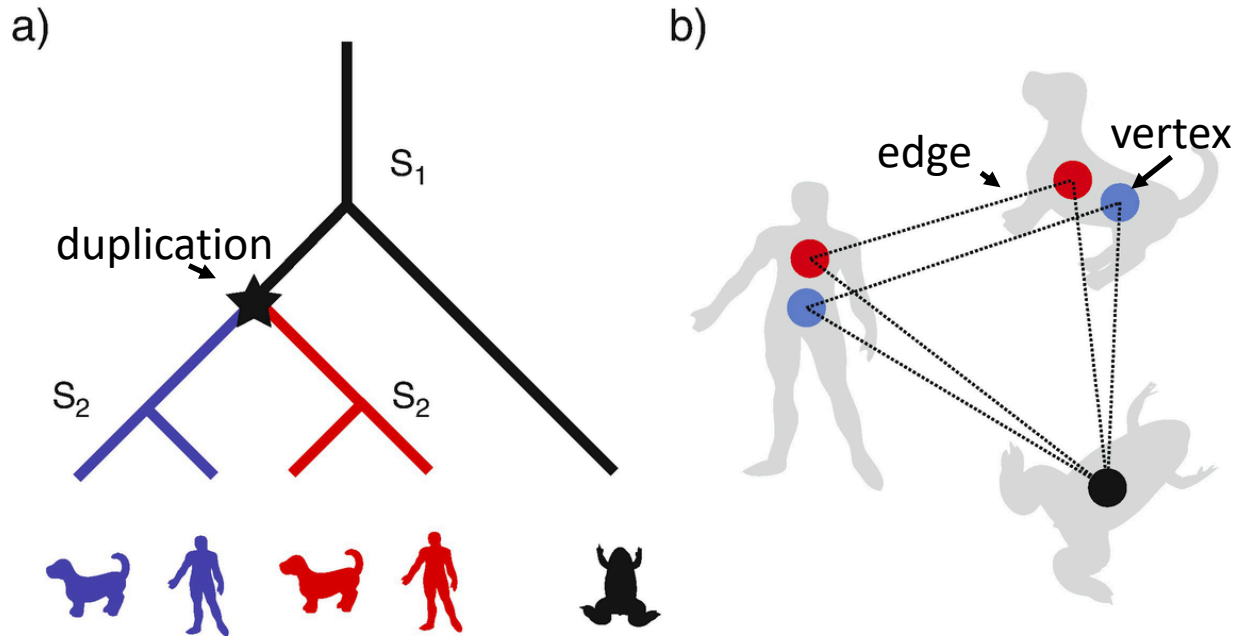
The frog gene has a 'one-to-many' relationship with the human and dog genes.



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Graph-based Methods

Method	Detects in-paralogs?	Based on	Uses Trees?
RBH	No	BLAST scores	No
COG	Yes	BLAST scores	No
EggNOG	Yes	Smith Waterman scores	Yes
InParanoid	Yes	BLAST scores	No
OrthoDB	Yes	Smith Waterman scores	Yes
OrthoMCL	Yes	BLAST scores	No



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Tree-Based Methods

*Use gene trees whose internal nodes are labeled as **speciation** or **duplication** nodes.*

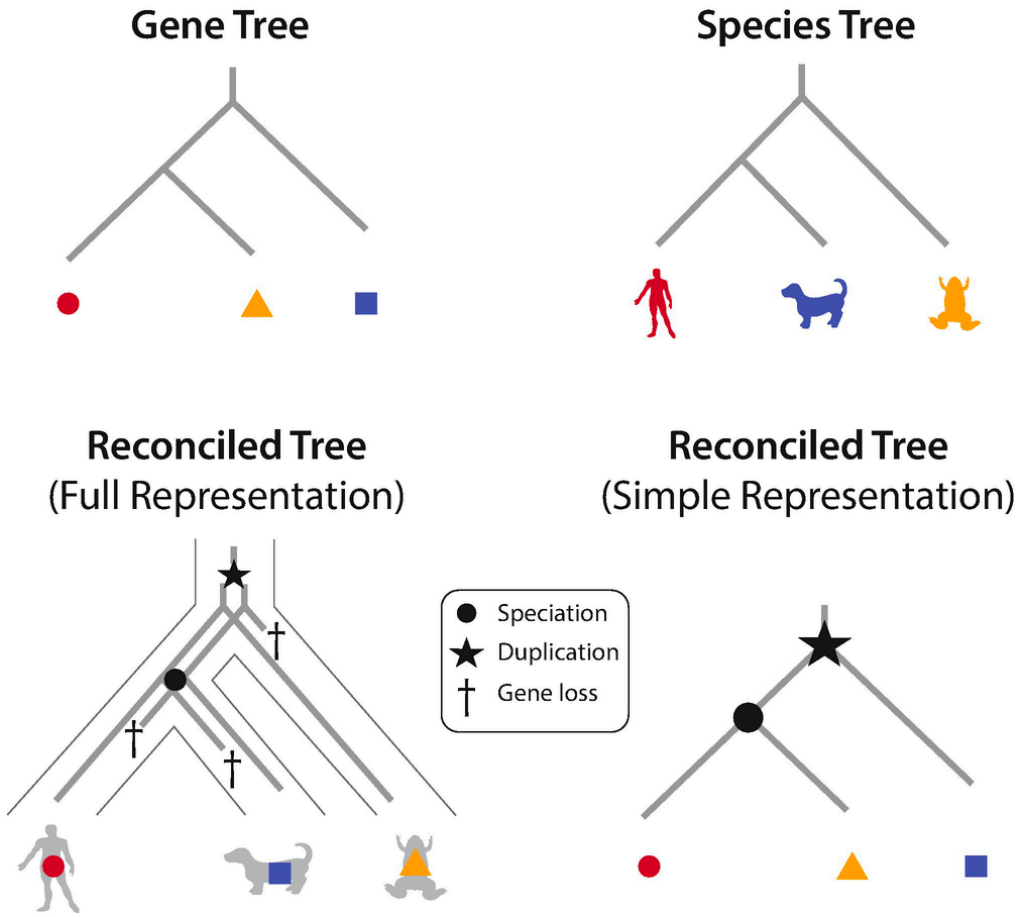
Also called 'reconciling' gene and species trees

In the example, the gene and species tree are not compatible.

The inferred history is one where the frog and human genes are orthologs and are paralogous to the dog.

- *The ancient duplication event was followed by **losses** of the paralogs in frog and human, and of the ortholog in the dog.*

*Most methods rely on **maximum parsimony.***



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Tree-Based Methods

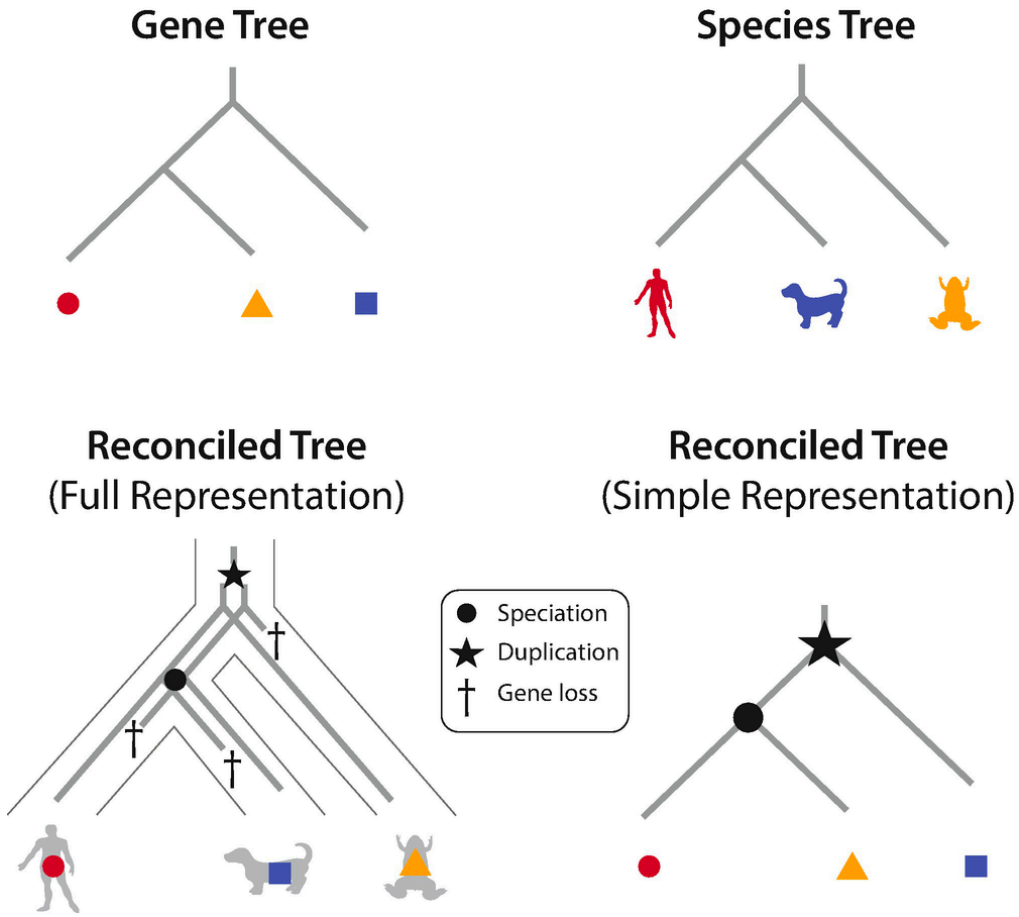
Use of maximum parsimony makes most tree-based methods computationally efficient.

Limitations include:

- *Unresolved species trees*
- *Unrooted trees*
  - *although an outgroup can simplify*
- *Gene tree uncertainty*
  - *Bootstrapping helps but at efficiency cost*

Probabilistic models using MCMC can provide posterior probabilities (and highest posterior densities)

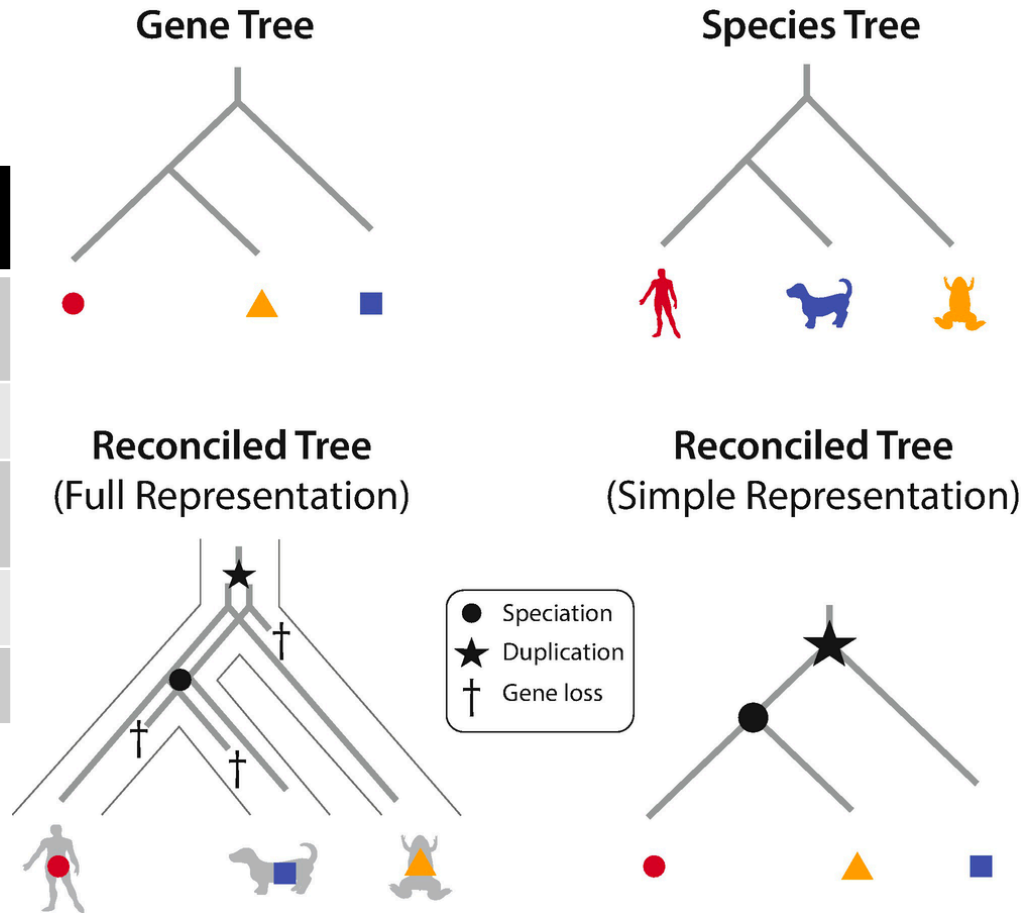
- *But there are issues of over-parameterization*
- *And, the MP results is often also the most probable one anyway!*



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Tree-Based Methods

Method	Framework	Species tree	Rooting	Gene tree uncertainty?
BranchClust	n.a.	Species overlap	Min. no. of clusters	none
DLRSOrthology	Probabilistic	Fully resolved	n.a.	none
Ensembl/TreeBeST	Parsimony	Partially resolved	Min. dupl + min. loss	none
Orthostrapper	Parsimony	Fully resolved	Min. dupl	Bootstrap
PhylomeDB	Parsimony	Species overlap	Outgroup	none



Altenhoff, Glover and Dessimoz. 2019. "Inferring Orthology and Paralogy". ***Methods in Molecular Biology.***

Conclusions

1. *Distinguishing between orthologs and paralogs is crucial for successful genome annotation*
2. *Classifying orthologs and paralogs reflects the interplay between gene duplication and speciation*
3. *Methods for identifying orthologs and paralogs rely on sequence similarity as well as phylogenetics (which we will explore further next lecture)*

Next

I am looking forward to your proposals but I’ve received few questions

- please reach out to me before you hand in your proposal! ☺