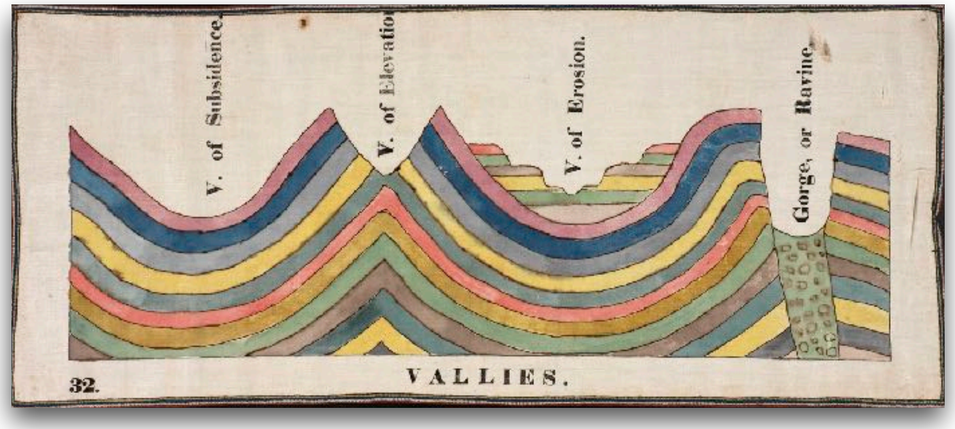
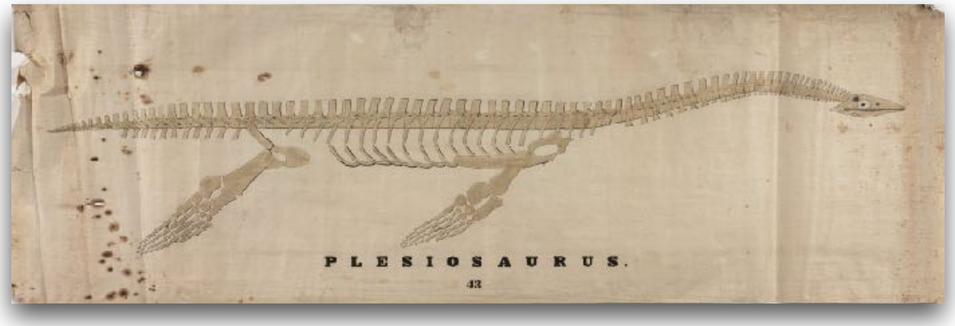
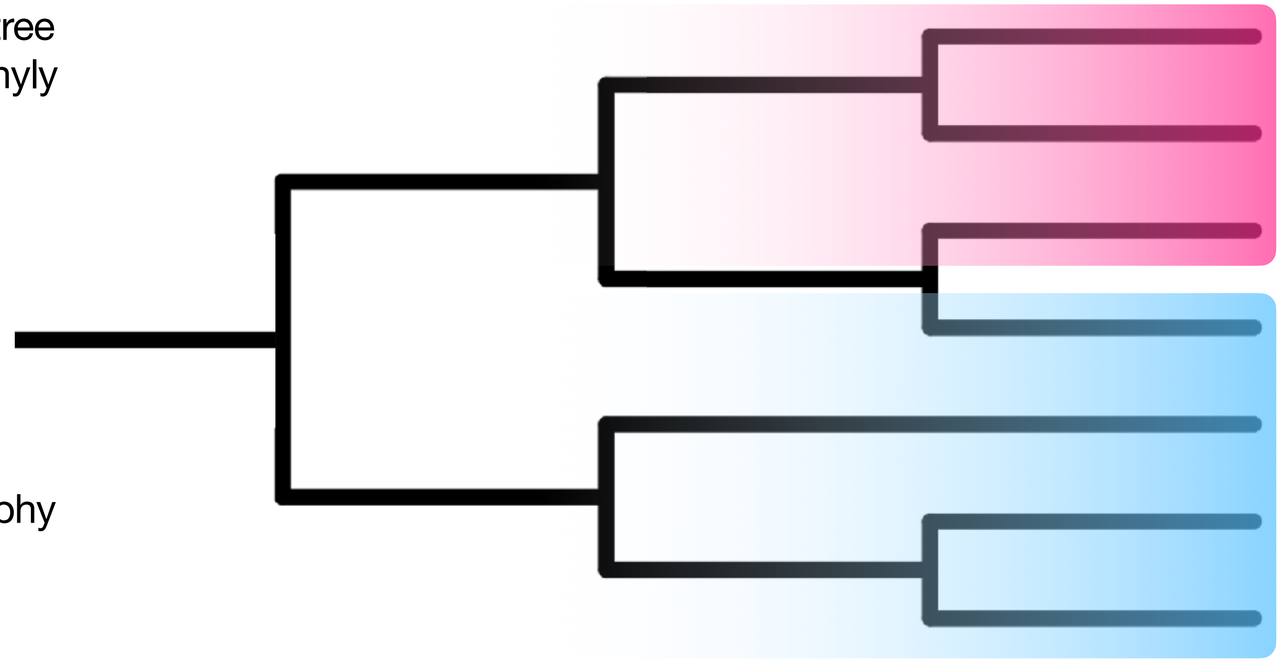


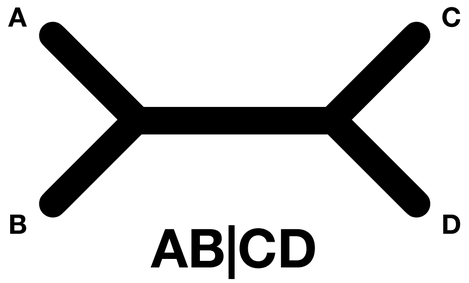


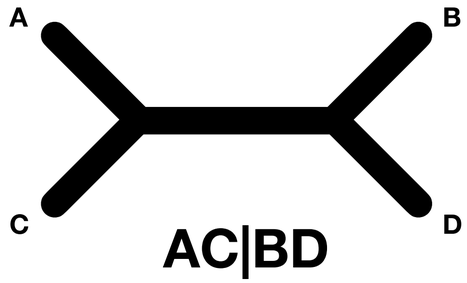
orthology infererence and taxon sampling

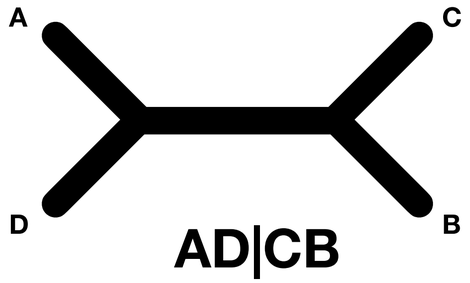
A QUICK RECAP:

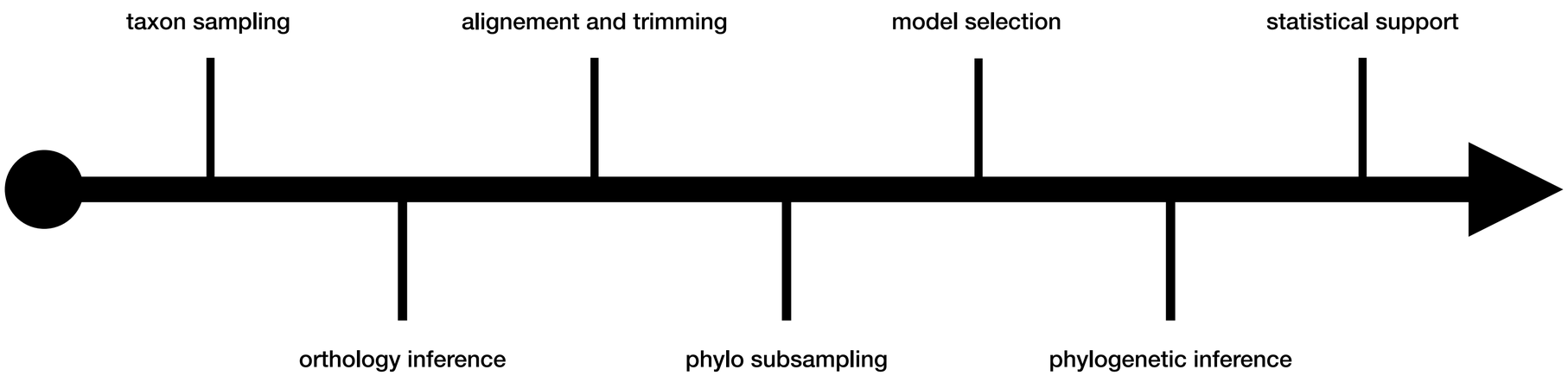
root / rooted *versus* unrooted tree
monophily / polyphyly / paraphyly
internal / terminal node
bipartitions
topology / branchlengths
cladogram
phylogram
chronogram (or timetree)
dichotomy / polytomy
clade
synapomorphy - sinplesiomorphy
ingroup and outgroup
mutations and substitutions
quartet
homoplasy / homology
molecoular clock hypothesis










taxon sampling
alignement and trimming
model selection
statistical support
orthology inference
phylo subsampling
phylogenetic inference

EXP. DESING

=

WHICH SPECIES?

Impact of incomplete and/or biased sampling on phylogenetics:

- **long branch attraction** – distantly related taxa with long branches may cluster erroneously due to missing intermediate taxa.
- **artificial clade resolution** – missing taxa can create misleading monophyletic groups that do not reflect true evolutionary history.
- **loss of phylogenetic signal** – sparse taxon sampling reduces informative sites, leading to unresolved or ambiguous trees.
- **model inadequacy** – missing taxa can skew evolutionary rate estimations, affecting divergence time inferences

Best Practices:

- ensure broad and representative sampling to break up long branches
- appropriate outgroups and intermediate taxa to improve tree rooting

EXP. DESIGN

=

WHICH SEQUENCING?

Transcriptomes

Pros:

- very large set of genetic markers
- much cheaper than sequencing genomes -> high number of species
- not dependent upon a reference genome
- good for shallow & deep evolutionary distances

Cons:

- incomplete identification of full-length genes and single-copy transcripts
- potential misassembly of transcripts (duplicates chimerism)
- missing data as transcriptome representing a snapshot of expression
- Ffesh tissue needed

Genomes

Pros:

- very large set of genetic markers
- good identification of full-length genes, less chimeras
- good for shallow and deep evolutionary distances
- ethanol-fixed specimens are OK (for draft genomes)

Cons:

- annotation may may not be comparable between species (software, etc)
- expensive (money and computing time)
- more difficult to have a high number of species
- Ffesh tissue needed (for chromosome-level genomes)

Mitochondrial genomes

Pros:

- high copy number, easy to sequence and assemble low quality samples
- compact and conserved genome structure facilitating annotation (often)
- useful for shallow and moderate distances due to fast mutation rates
- cost-effective, low sequencing depth and computational resources

Cons:

- limited number of genetic markers compared to nuclear genomes
- high substitution rates may lead to saturation and homoplasy
- possible heteroplasmy (within-individual variation in sequences)
- maternal inheritance just tells a part of the story
- dscrepancies between software may still occur

REDUCED REPRESENTATION

Ultraconserved elements (UCEs)

Pros:

- a medium-to-large set of genetic markers
- much cheaper than sequencing genomes -> a large amount of species
- not dependent upon a reference genome
- ethanol-fixed and museum specimens are OK

Cons:

- no markers outside the ones designed *a priori*
- potential misassembly (if probes are designed with a few species)
- no proper orthology inference

RADseq and GBS

Restriction site-Associated DNA Sequencing and Genotyping-By-Sequencing

Pros:

- the cheapest of the methods! 💸
- not dependent upon a reference genome
- ethanol-fixed specimens are OK
- markers distributed evenly across the genome

Cons:

- no full genes, only SNPs
- only for population genomics or phylogeny of closely-related species
- no proper orthology inference

PCR amplified gene fragments

Pros:

- …

Cons:

- previous knowledge required to design amplification primers
- sometimes difficult to obtain results (also related to primers)
- no proper orthology inference

A ROUGH ESTIMATE …

data type	cost	markers
Chromosome-Level (~3 Gb):	€10,000 – €20,000	thousand of genes
Draft Genome (~3 Gb):	€5,000 – €10,000	thousand of genes
Transcriptome (RNAseq)	€250 – €500	thousand of genes
Mitochondrial Genome:	€150 – €500	fifteen genes usually
UCEs:	€75 – €100	thousand (short)
RAD-Seq:	€100 – €300	thousand (short)
GBS:	€50 – €200	thousand (short)
PCR sequencing:	€20 – €25 *per* single gene	one

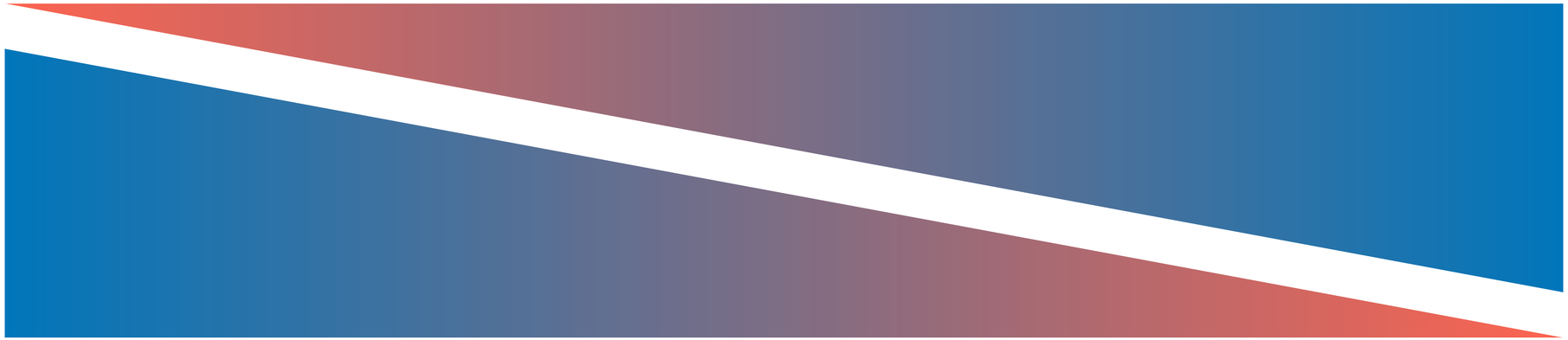
by the way …

WHAT ARE WE SEQUENCING?

one cell / one organism

multiple cells / one organism

one cell / multiple organism

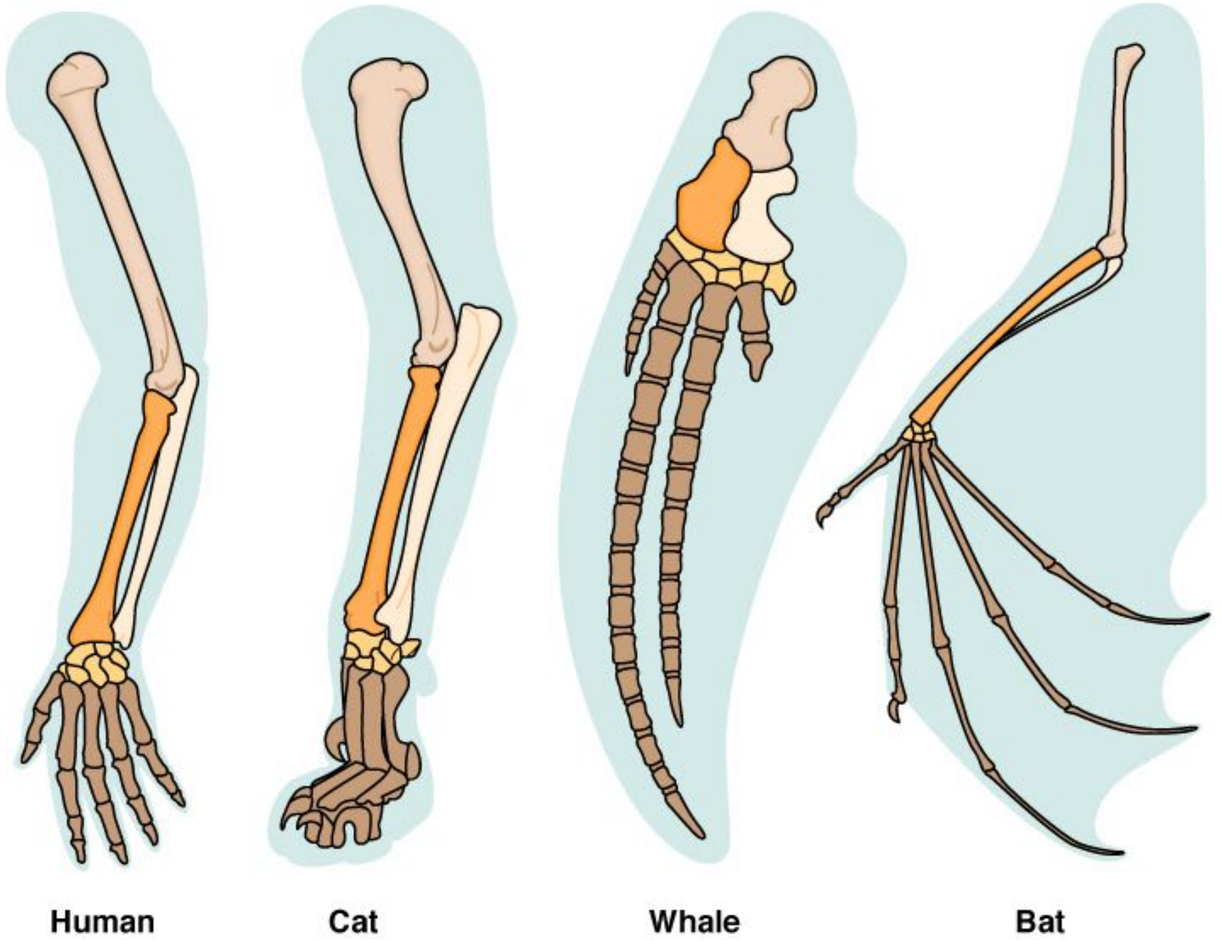


single cell genomics

bulk sequencing

metagenomics

Phylogenetics is done using homologus characters! 😎



… but what does it mean when we think about genes? 😱

HOMOLOGY

characters sharing ancestry

ANALOGY

characters sharing function but not ancestry

PS: 😵‍💫

Homoplasy is the broader concept — it refers to any similarity between organisms that was *not* inherited from a common ancestor. This includes convergent evolution (unrelated species independently evolving similar traits), parallel evolution (related species evolving similar traits independently), and evolutionary reversals (a trait reverting to an ancestral state). **Analogy** is a specific type of homoplasy. It refers to structures in different organisms that perform the same function but have different evolutionary origins, either parallel or convergent. The emphasis is on functional similarity.

… but within homology …

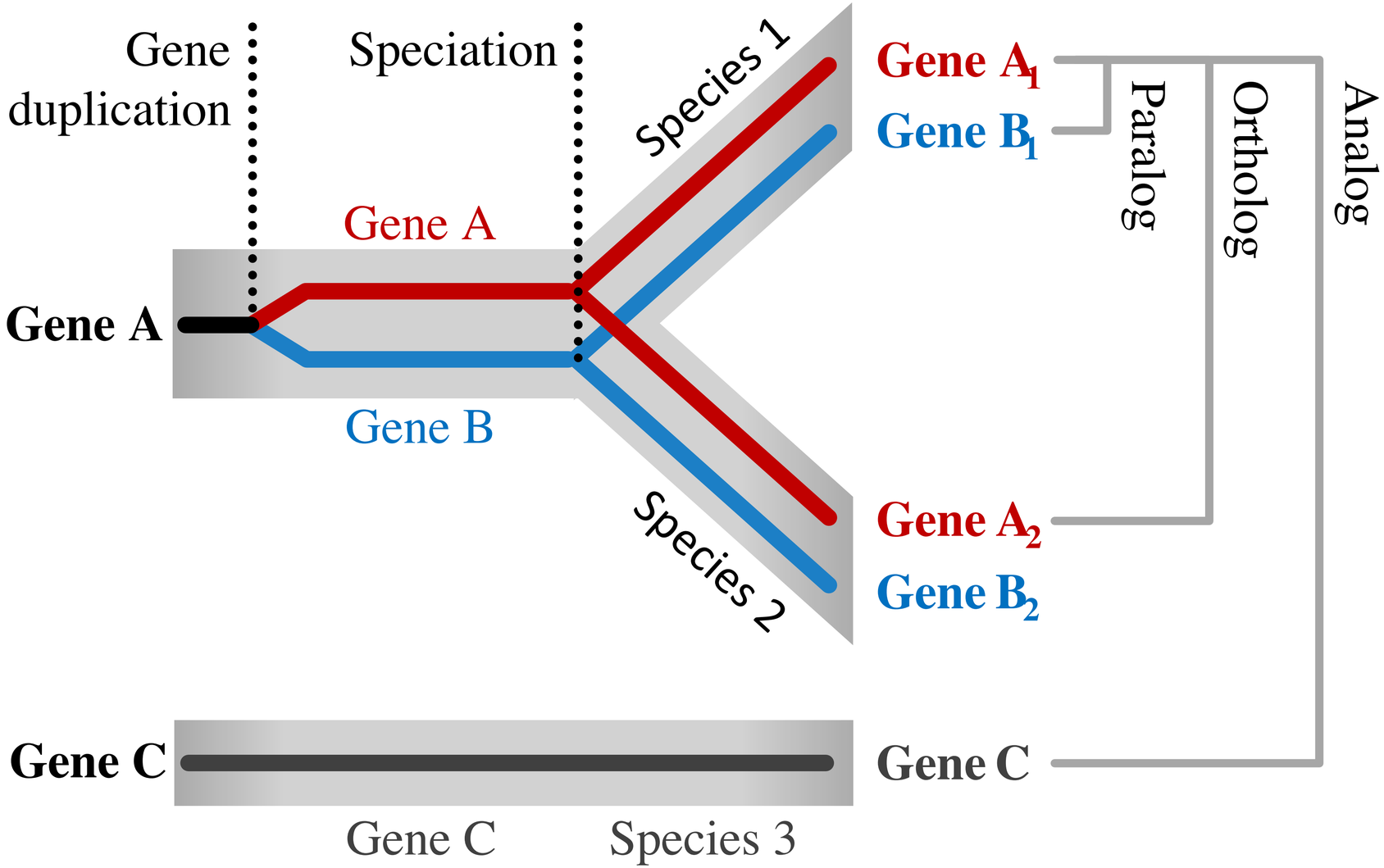
Orthology *vs* Paralogy *vs* Xenology 🧬

- **Orthologs** - genes in diff. species whose most recent common ancestor is a **speciation event**. Because they diverged when species split, their gene tree is expected to track the species tree. These are the primary markers for phylogenetic inference.

- **Paralogs** - genes whose most recent common ancestor is a gene **duplication event**. They can exist within the same genome or across species. As their divergence predates (out-paralogs) or postdates (in-paralogs) a given speciation, they do not necessarily reflect species relationships.

- **Xenologs** - genes whose most recent common ancestor is a **transfer event**. Xenologs are especially common in prokaryotes but also occur in eukaryotes. Like paralogs, they can confound species tree estimation because their history reflects transfer rather than descent.

The unifying principle: what distinguishes these categories is not the sequences themselves, but the nature of the **evolutionary mechanism at their most recent common ancestor** - speciation, duplication, or transfer.



wikipedia page on orthology where I took this figure 😛

the orthology conjecture

proposed in Nehrt et al. 2011

Orthologs genes are expected share the same **biological function**, while paralgs genes are believed to differ in function.

However, as usual in biology, be aware of this latest corollary … 😑

This consept has been **largely questioned** … see Stamboulian et al. 2020, Lynch and Conery 2000, and Gout and Lynch 2015 😛

The key concept …

do not consider function, just the mechanism of orign!

Classification of orthologs and paralogs!

Orthology is **always** defined by phylogenetics and unit of comparison.

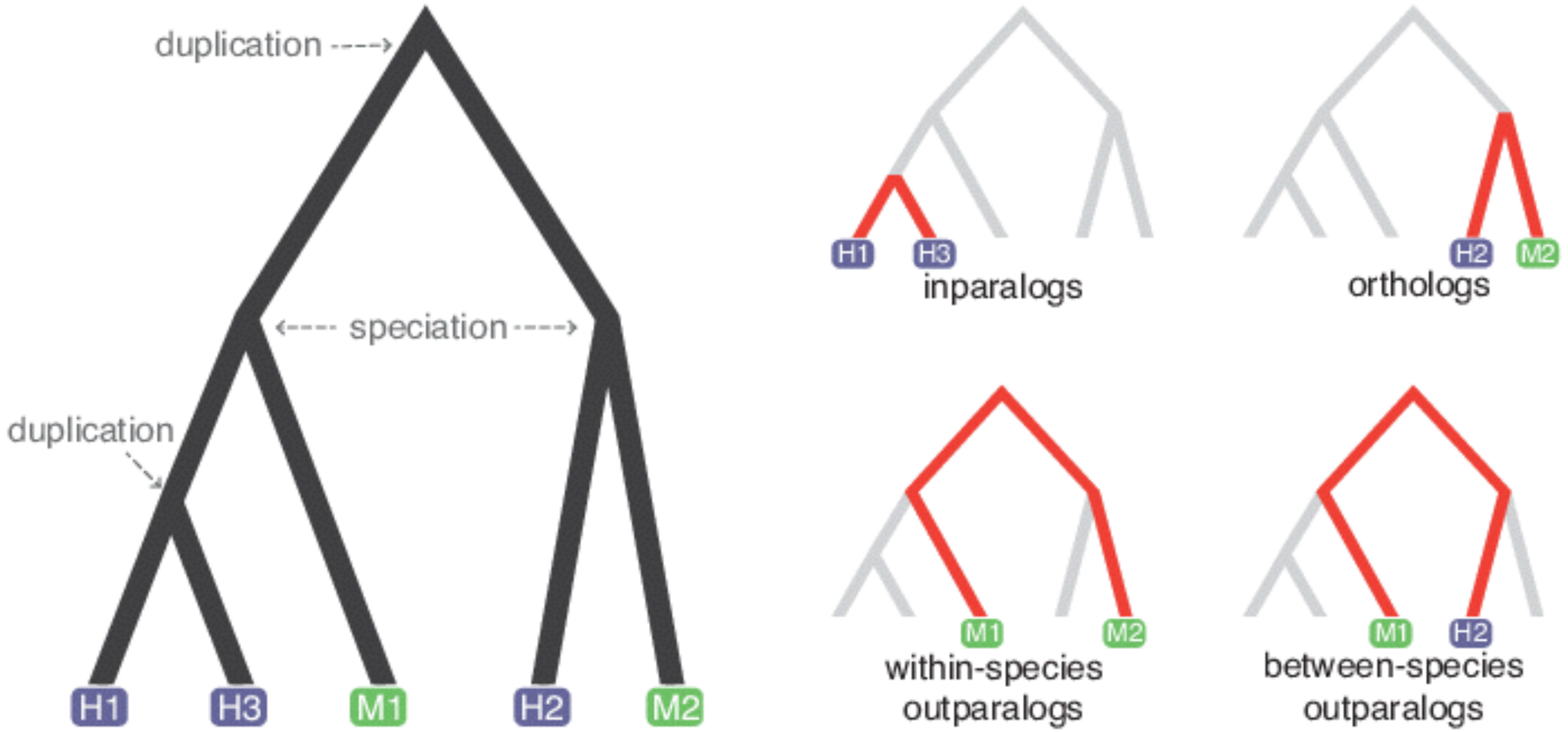
One-to-One Orthologs: A single copy of the gene is present in both species, originating from a common ancestor and retained without duplication after a speciation event.

One-to-Many and Many-to-One Orthologs: After a speciation event, gene duplication occurs in one of the two species. This results in a single copy in one species and multiple copies in the other.

Many-to-Many Orthologs: After a speciation event, gene duplications occur in both species, leading to multiple orthologous copies in each lineage.

In-Paralogs: Paralogous genes that arise after a given speciation event. These duplicates are restricted to the lineage that experienced the duplication.

Out-Paralogs: Paralogous genes that originated before a given speciation event. These genes are related by a duplication event that occurred in the common ancestor of both species.

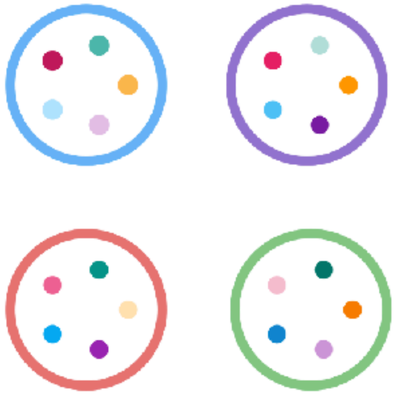

duplication
speciation
duplication
H1
H3
M1
H2
M2
H1
H3
inparalogs
H2
M2
orthologs
M1
M2
within-species
outparalogs
M1
H2
between-species
outparalogs

orthology relationsips are inferred **pairwise**

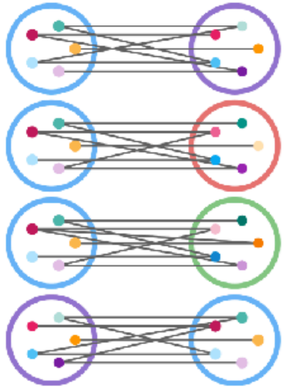
when considering **multiple species**,
we should move to the concept of **orthogroup**

an orthogroup is a group of **homologous** genes descending from the MRCA of a group of species
→ extending the concept of orthology to multiple species

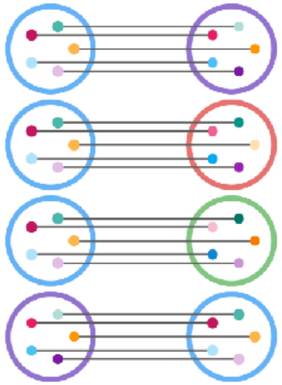
an **orthogroup** is always defined by
a **reference speciation event**



protein sequences
from annotated
genomes



homology searches
between pairs



extract best
reciprocal hits

cluster network
of hits
into orthogroups

Reciprocal Best Hit (**RBH**) is an easy way to infer orthologs between two genomes.

Start by performing an all-vs-all BLAST searches …

Each gene in genome S1 is used as a query to search against the entire genome of S2.
The top hit by e-value or bit score is recorded.

Each gene in genome S2 is used as a query to search against the entire genome of S1.
The top hit by e-value or bit score is recorded.

… then identify reciprocal best hits

Sp1 gene	Best Hit in Sp2	Sp2 gene	Best Hit in Sp1	RBH?
GeneA	GeneX	GeneX	GeneA	✅ Yes
GeneB	GeneY	GeneY	GeneC	❌ No
GeneC	GeneZ	GeneZ	GeneC	✅ Yes

Why is the reciprocal best hit important?
Without a bi-directional comparison, a lost (or not sequenced) gene in one species could lead to incorrect inferences of orthology.

What is RBH biggest limitation?
Fails to Detect One-to-Many or Many-to-Many Orthologs. Cases where a single gene in Sp1 corresponds to multiple genes in Sp2 are ignored.

Beyond RBH:

- **Clustering orthologs into orthogroups:** groups multiple related genes across species, allowing many-to-many relationships.

- **Gene tree inference:** uses phylogenetic reconstruction to distinguish true orthologs vs. paralogs, improving accuracy.

"Phylogenies require orthologous, not paralogous genes"
Walter M. Fitch, 1970

Why in phylogenetics we are interested in strictly orhologs genes?
Because orthologs genes arise by speciation events,
they share the same evolutionary history of the species they are found in.

Phylogenetics traditionally leverages:

- **1-to-1** or **single-copy orthogroups** (i.e. with only one copy per specie)
- **trimmed orthogroups** (without genes derived from duplication events)

However the use of gene families (**paralogs** + **orthologs**) to infer species **phylogenetic relationship** is getting more and more attention lately!

FINISH