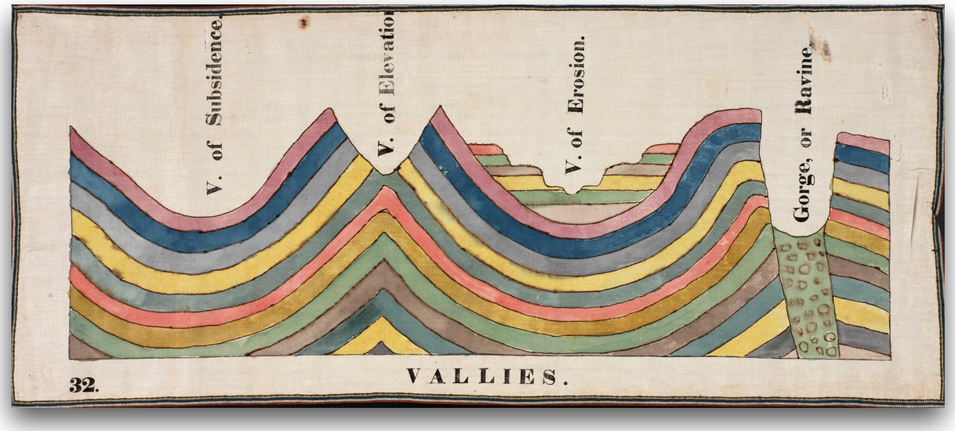
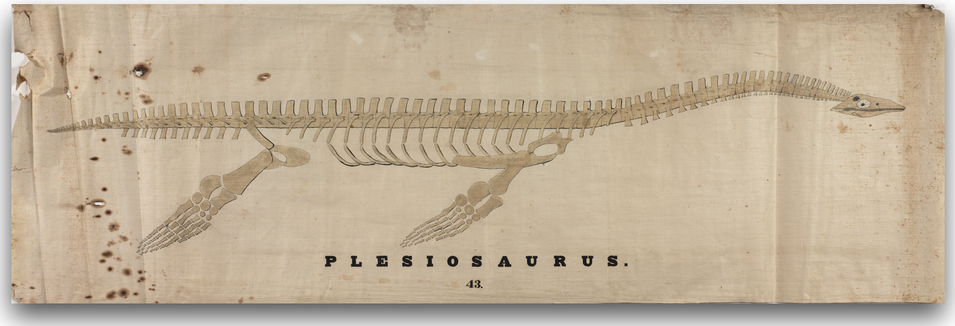
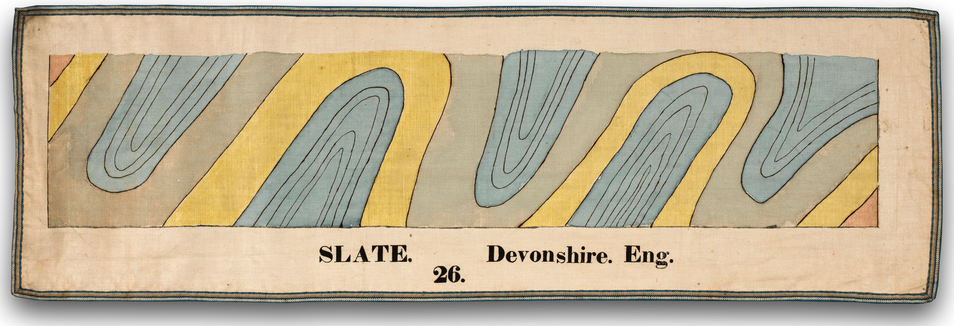


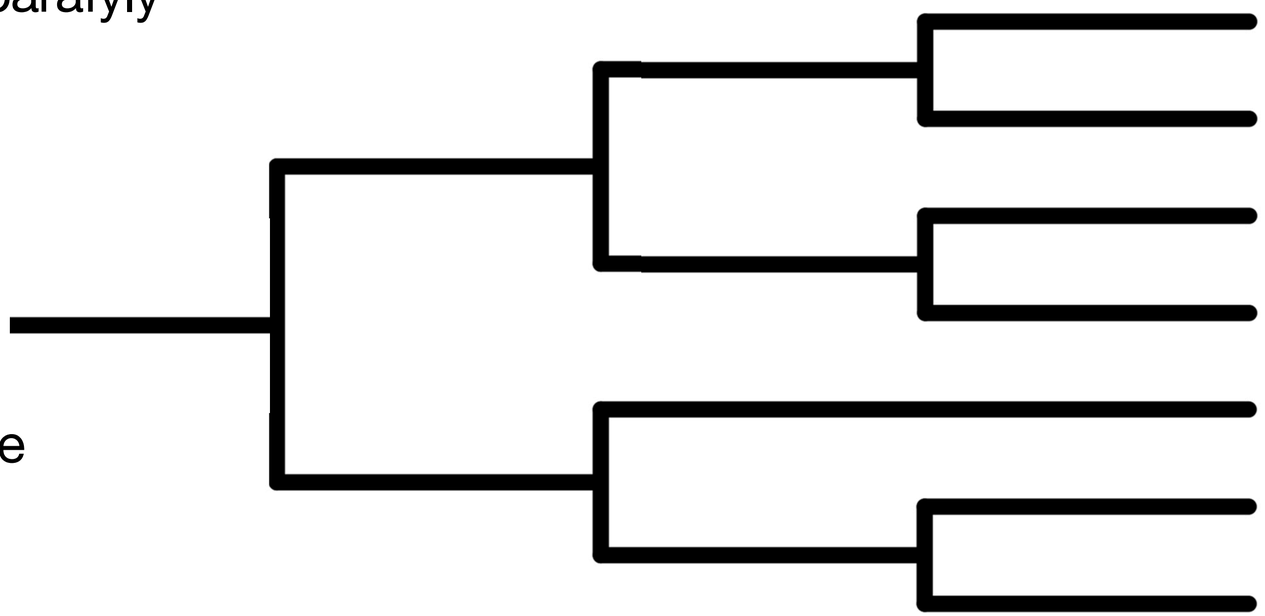


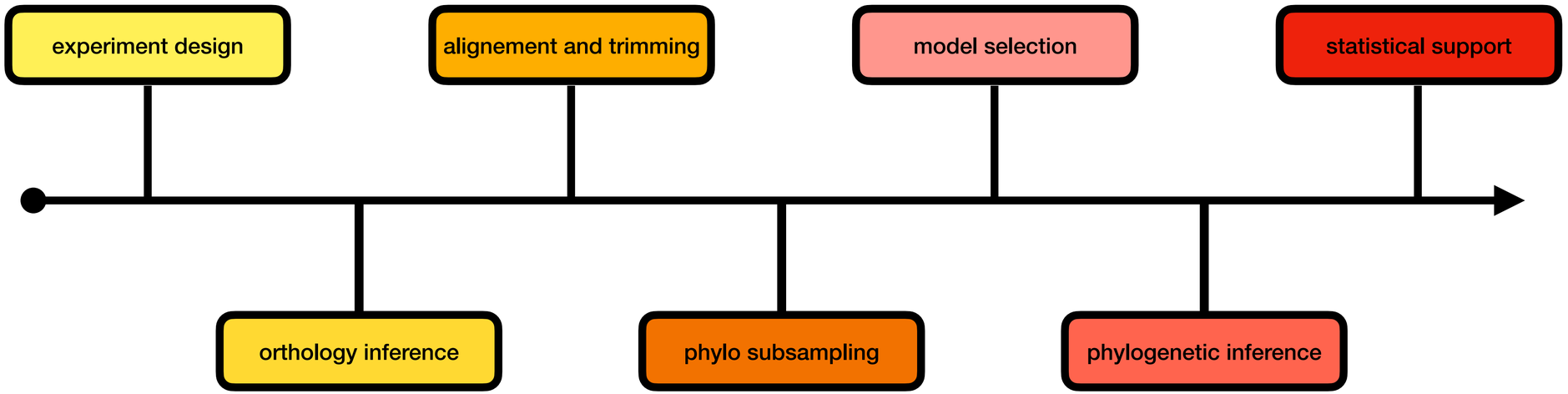


experimental
design
and
orthology
infererence

A QUICK RECAP:

rooted *versus* unrooted tree
monophily / poliphily / parafyly
internal node
terminal node
bipartitions
branch
cladogram
phylogram
chronogram or timetree
dicothomy
polytomy
clade
quartet
automorphy and sinapomorphy
ingroup and outgroup
mutations and substitutions




experiment design
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.
- **Misplaced Lineages & Rooting Errors** – Incomplete outgroup selection can misplace early-branching taxa and distort tree rooting.
- **Model Biases & Rate Heterogeneity** – 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.
- Models that account for rate variation and calibrate with fossil data.

EXP. DESING

=

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 and 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
- Fresh 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
- Fresh 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
- Discrepancies 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 markes outside the designed ones
- Potential misassembly (if probes are designed with a few species)
- Usefulness of markers known a posteriori
- 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:
direct information of seq data
pretty much shure they are orthologs (?)

Cons:
cometimes difficult to obtain (wet-lab misteries)
previous knowledge required to design primers
need maaany sequencing to have a decent amount of genes =many € …

A ROUGH ESTIMATE OF PRICE *PER* SAMPLE

- **Genome, Chromosome-Level (~3 Gb)**: €10,000 – €20,000
- **Genome, Draft (~3 Gb)**: €5,000 – €10,000
- **Transcriptome Sequencing (RNA-Seq):** €250 – €500
- **Mitochondrial Genome:** €150 – €500
- **UCEs:** €75 – €100
- **RAD-Seq**: €100 – €300
- **GBS**: €50 – €200
- **PCR sequencing**: €20 – €25 *per* single gene

by the way …

WHAT ARE WE SEQUENCING?

metagenomics and metatrascriptomics

ONE CELL / MULTIPLE ORGANISMS

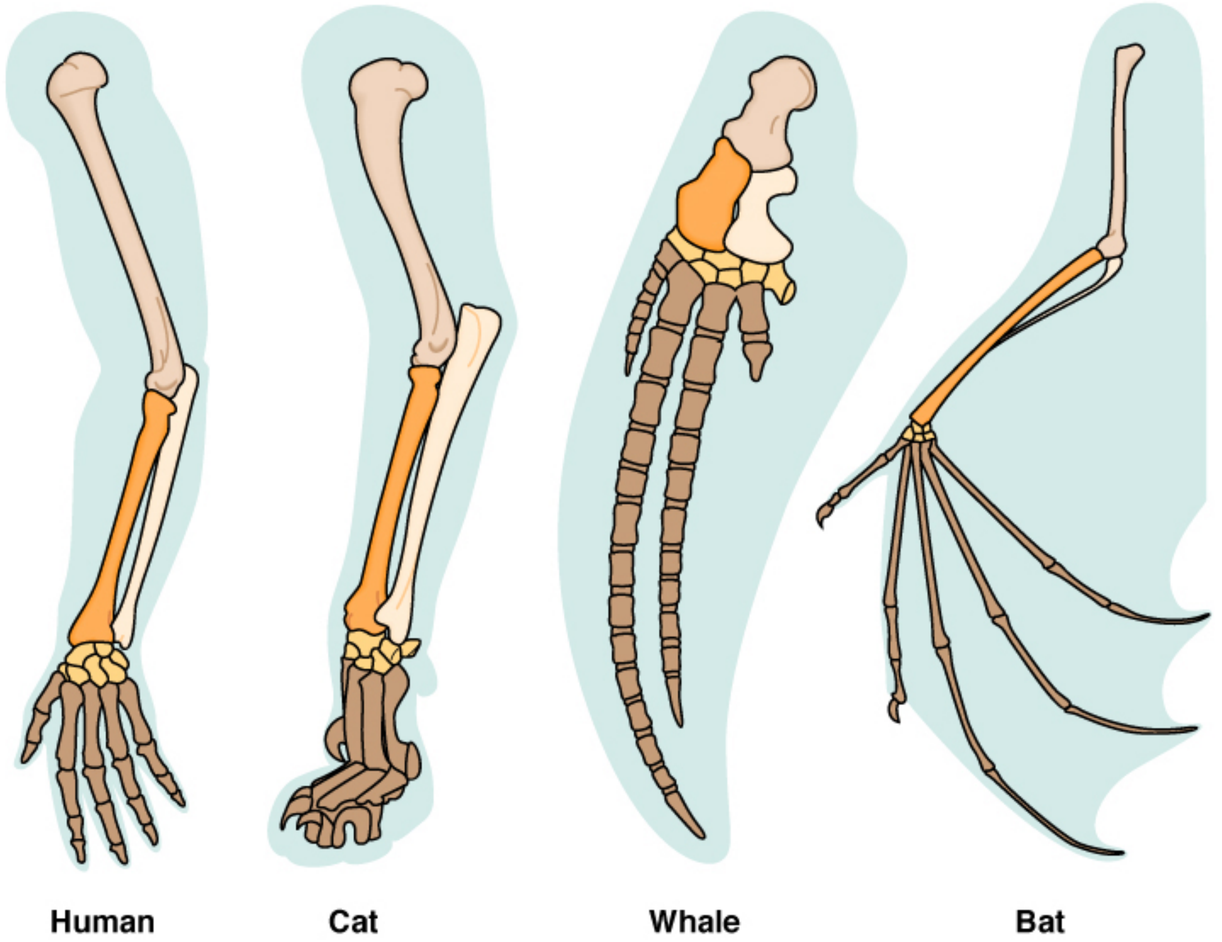
bulk sequencing

MULTIPLE CELL / ONE ORGANISMS

single cell genomics

ONE CELL / ONE ORGANISM

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

🤨

Compared characters **MUST** be homologous

… but within homology …

Orthology vs. Paralogy 🧬

- **Orthologs**

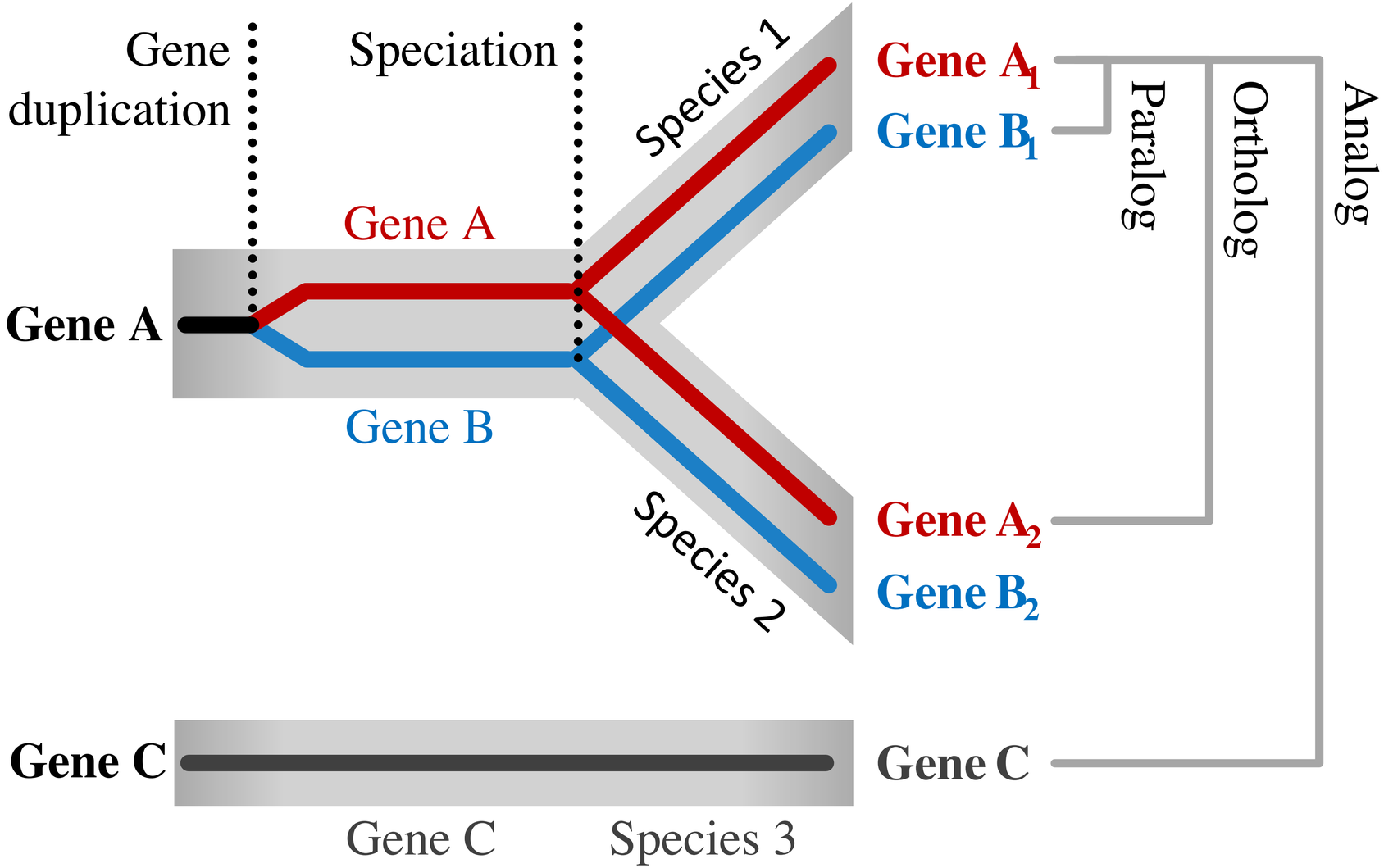
Genes that evolved from a common ancestral gene via **speciation** → **reflect species relationships** and may used for phylogenetic inference.

Two genes are orthologs if their MRCA is a **speciation** event!

- **Paralogs**:

Genes that originated from a common ancestral gene via **duplication** → may **confound phylogenies** as they do not reflect species relationships.

Two genes are paralogs if their MRCA is a **duplication** event!



wikipedia page on orthology where I took this figure 😛

"Phylogenies require orthologous, not paralogous genes"

Walter M. Fitch, 1970

Why in phylogenetics inference we are interested in strictly orhologs genes?

Since orthologs genes arise by speciation events,
they share the same evolutionary history of the underlying species.

the orthology conjecture

proposed in Nehrt et al. 2011

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

However, as usual in biology, be aware of this latest corollary, recently ortholog conjecture has been **largely questioned** … see Stamboulian et al. 2020, Lynch and Conery 2000, and Gout and Lynch 2015 for for some interesting hints on fate of duplicated genes 😛

Btw, I think the orthology conjecture is mostly false … 😑

The key concept … **in the context of phylogenetics do not consider function, just inheritance!**

Classification of ortologs and paralogs!

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

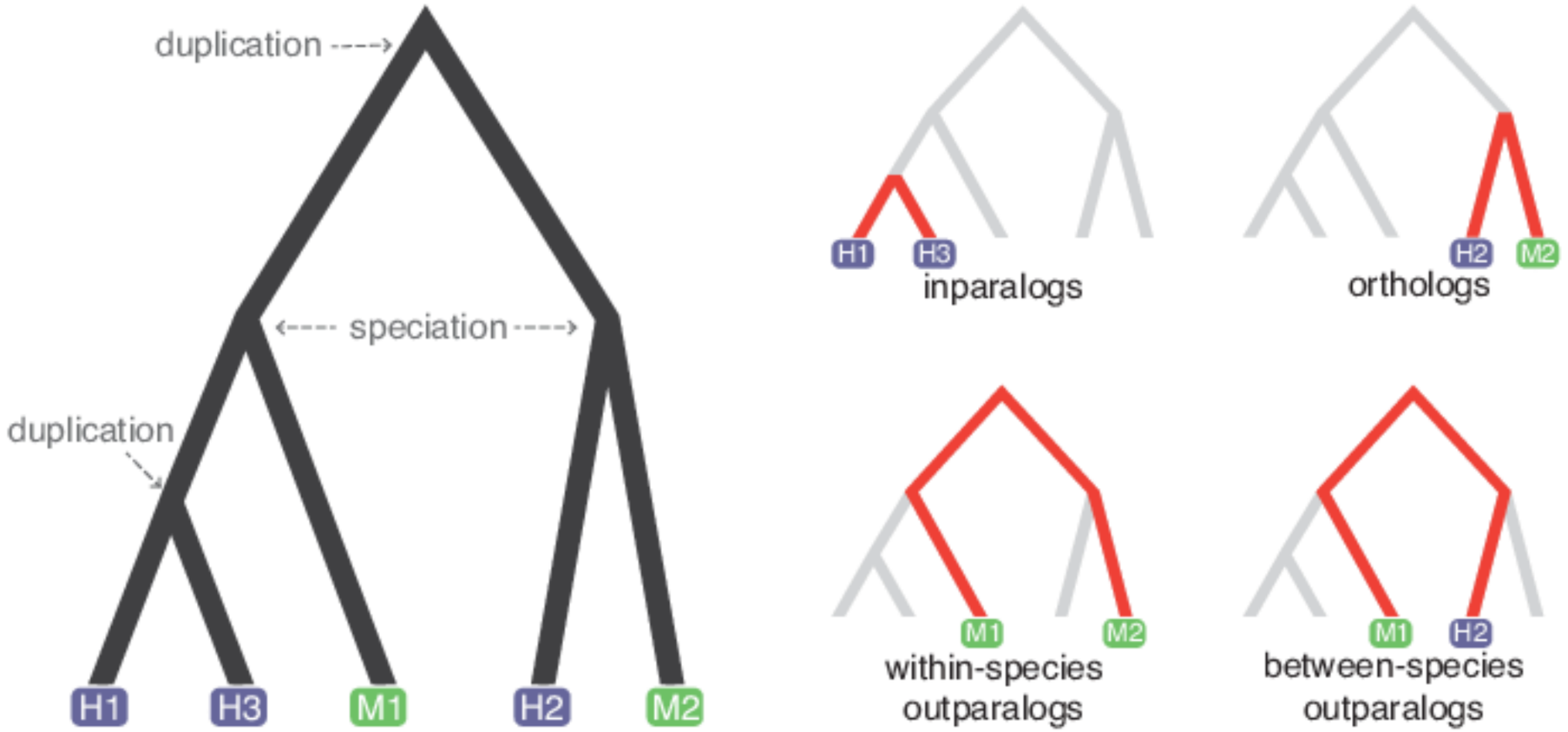
One-to-One Orthologs: A single copy of the gene is present in both species, originating from a common ancestor and retained without duplication after a speciation event.

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 inferences 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**

Reciprocal Best BLAST Hit or **RBH** is an easy way to infer orthologs between 2 genomes.

- **Perform All-vs-All BLAST Searches**

Each gene in 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 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.

- **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 Bi-Directional Best Hit Important?
Without a bi-directional comparison, a lost 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, and **RBH is strictly one-to-one**.

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.

In **phylogenetics** we either want:

- **1-to-1** or **single-copy orthogroups** (i.e. with only one copy for 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!

.. for further infos see *Inferring Orthology and Paralogy* by Anisimova 2019.

FINISH