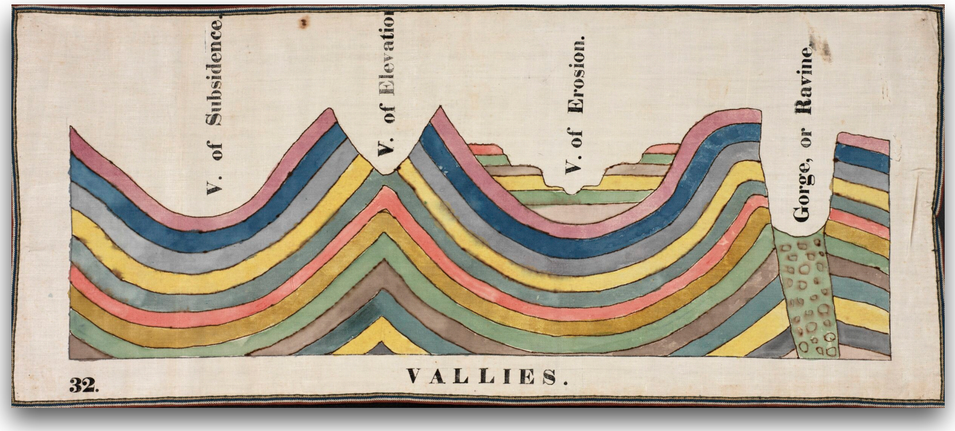
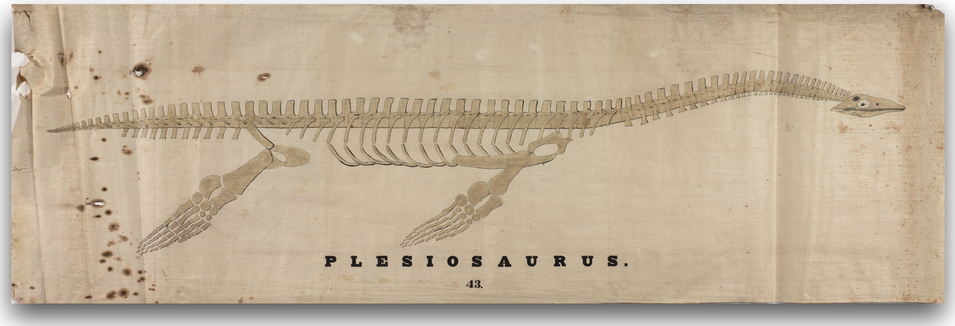
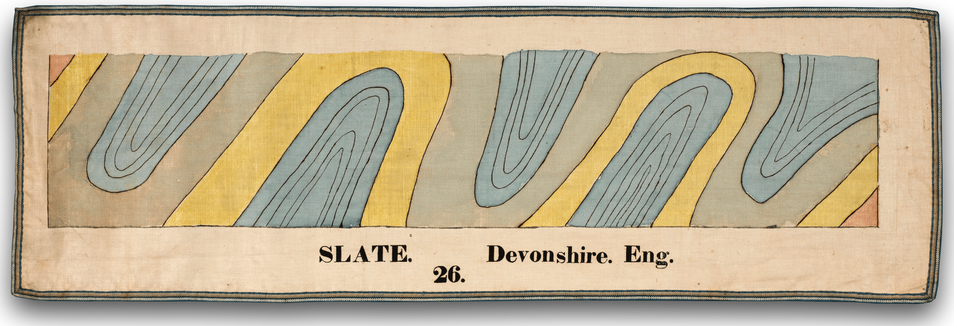


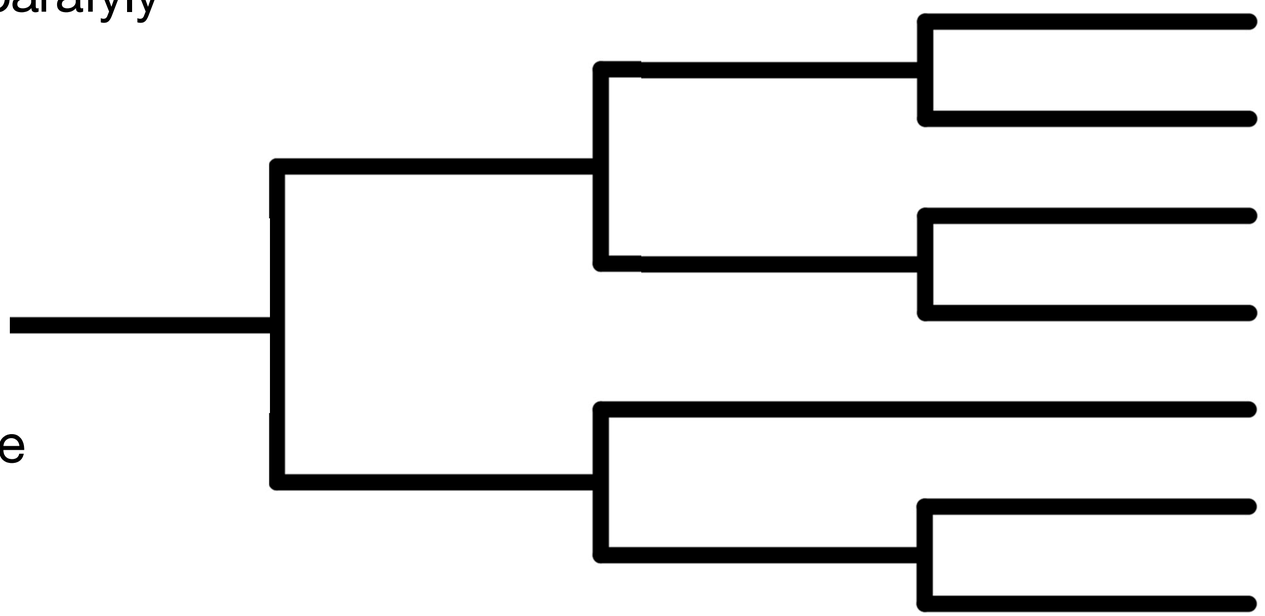


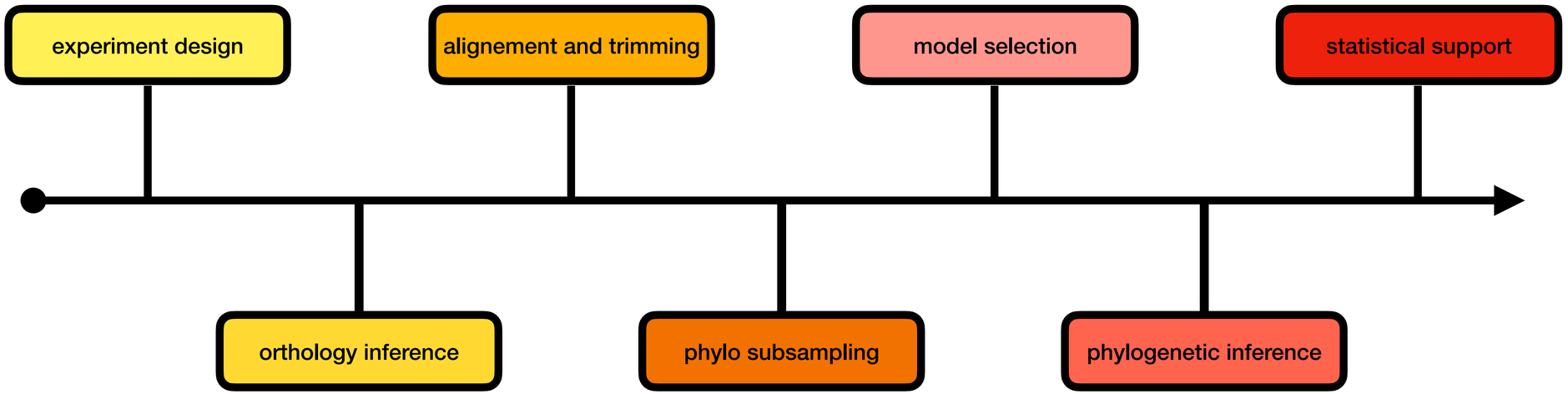


experimental
design
and
orthology
infererence

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 taxon sampling to break up long branches.
- Use appropriate outgroups and intermediate taxa to improve tree rooting and lineage placement.
- Apply models that account for rate variation and calibrate with fossil data when possible.

EXP. DESING

=

WHICH SEQUENCING

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 tissue 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)

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

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:

- Medium-large set of genetic markers
- Cheaper than sequencing genomes and easier to have many species
- Not dependent upon a reference genome
- Tissues fixed in EtOH or museum specimens are OK

Cons:

- Limited availability of markes outside the designed ones.
- Potential misassembly (if probes are designed with a few species)
- Retrieval success dependent on DNA quality
- 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
- Samples fixed in ethanol OK
- Markers distributed evenly across the genome

Cons:

- Li 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:

difficult to obtain (require previous knowledge)

need many sequencing (=many €…)

relatively limited variability

If we are setting up an experiment involving the Sanger sequencing of a marker, we should know *a priori* that all fragments are orthologs between each other (choose the markers and primers based on bibliography knowledge). A common way is to use mtDNA sequences and nuclear ribosomial RNA (*e.g.* 28s).

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

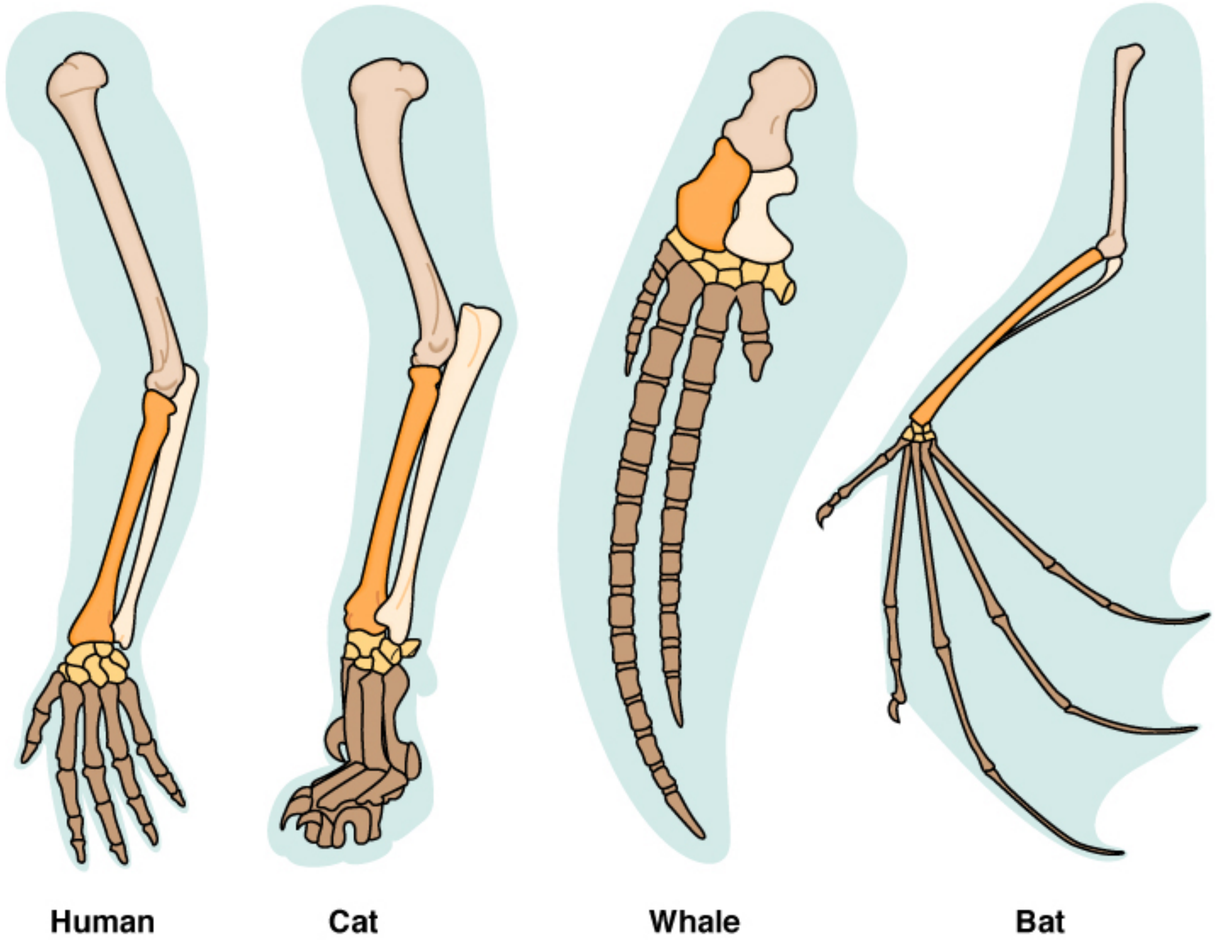
bulk sequencing

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 🧬

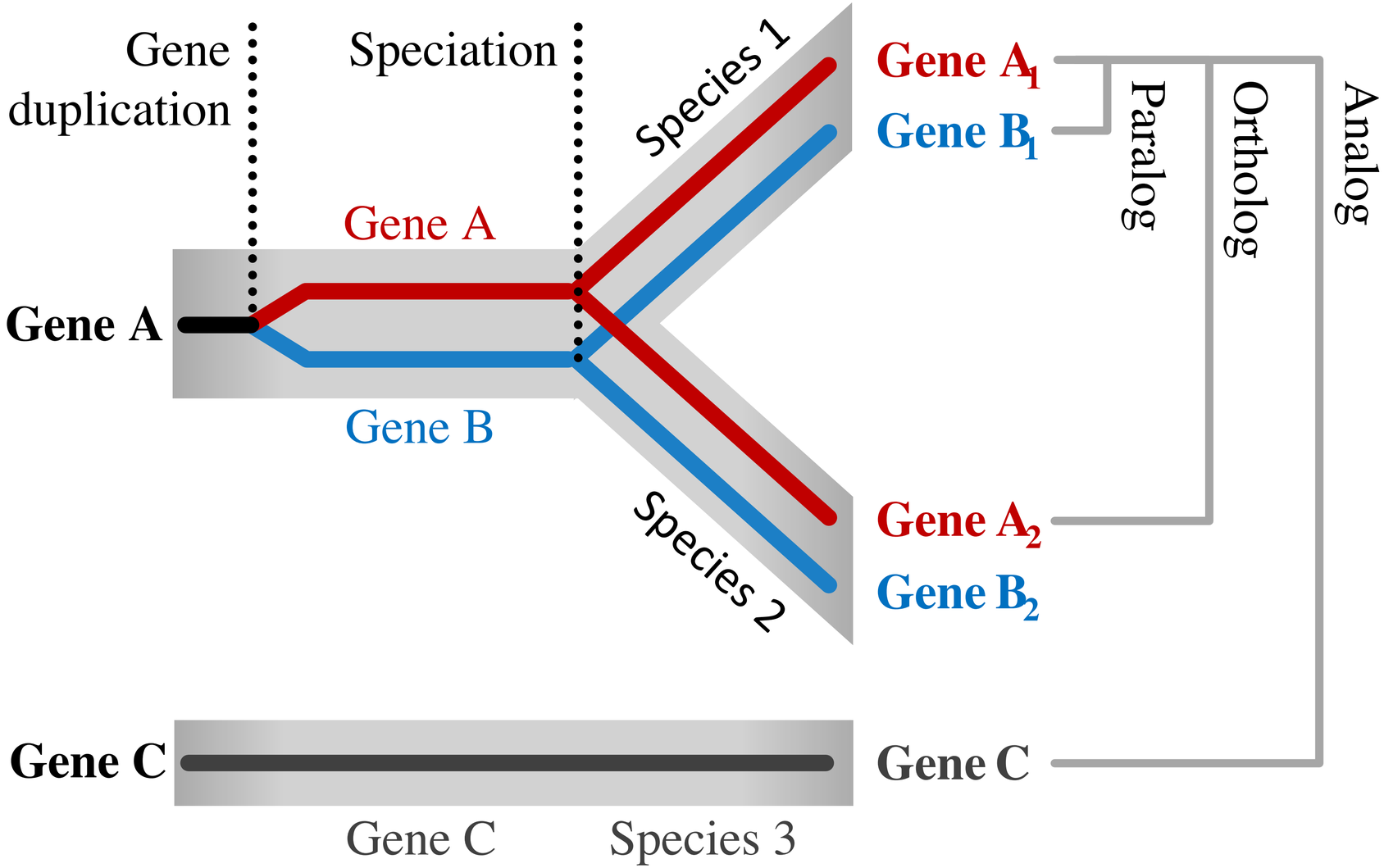
- **Orthologs**

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 if not properly distinguished as they do not reflect species relationships.



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.

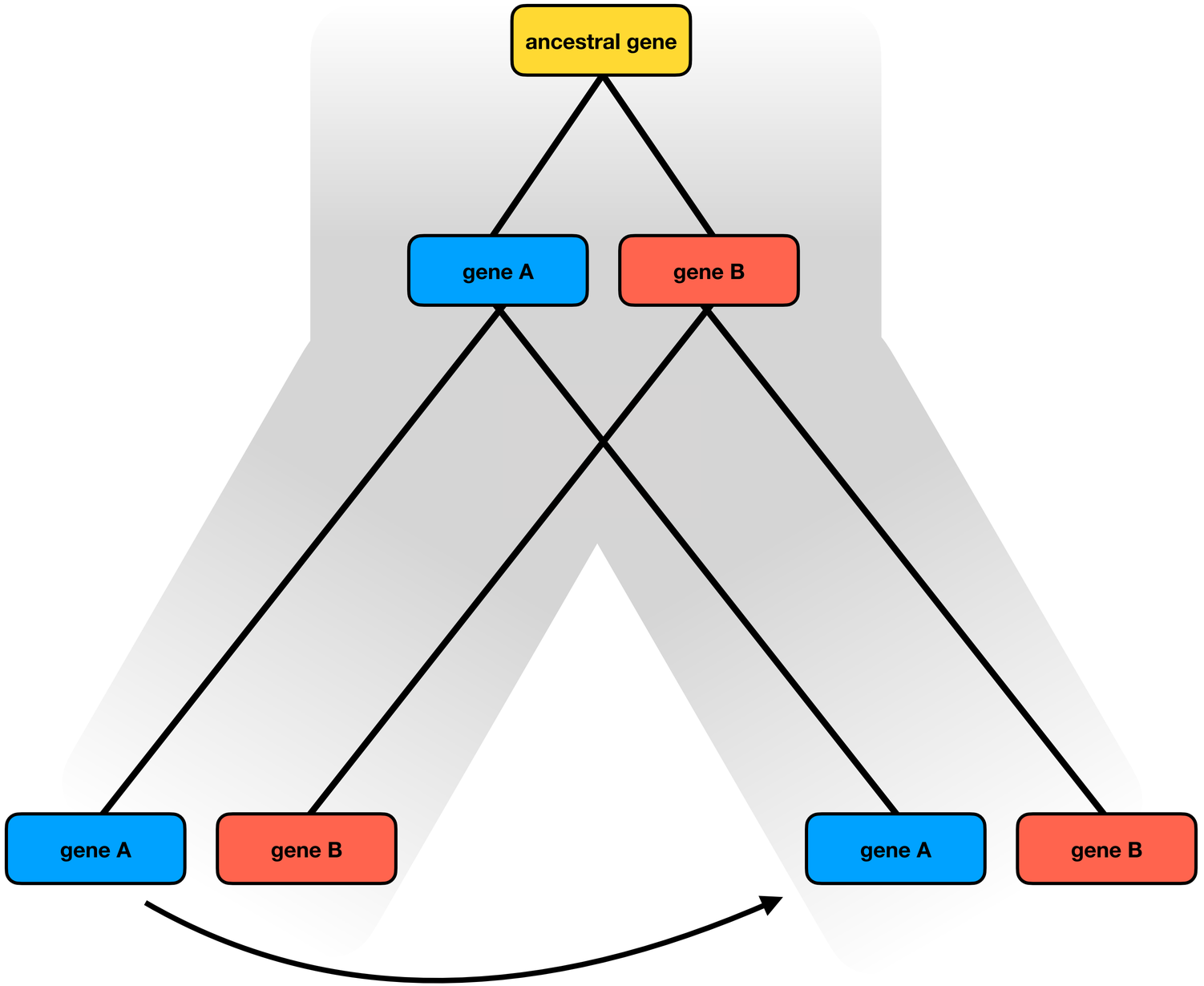
the orthology conjecture

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 … 😑


ancestral gene
gene A
gene B
gene A
gene B
gene A
gene B

Classification of ortologs and paralogs!!

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

One-to-one ortologs: in both species is present only one copy of the gene, arise after speciation event.

Many-to-one and **one-to-many ortologs:** after specieation event, one or more duplication events occured in one of the two lineages, as a result we have three or more ortholog genes!

Many-to-many ortologs: after speciation event, one or more duplications events occured in both lineages, as a result we have multiple copies of ortologs genes.

In-paralogs: is definied over a triplet. It involves a pair of genes and a speciation event of reference. A gene pair is an in-paralog if they are paralogs and duplicated after the speciation event of reference.

Out-paralogs: is also a relation defined over a pair of genes and a speciation event of reference. This pair is out-paralogs if the duplication event through which they are related to each other predates the speciation event of reference.

...and others (see chapter "Inferring Orthology and Paralogy" [Anisimova, 2019](https://core.ac.uk/download/pdf/289121767.pdf)).

Orthology inferences are inferred pairwise
when we have multiple species we should contemplate the concept of Orthogroups

An orthogroups is a group of orthologs genes descending from the ****last common ancestor**** (LCA) of a group of species. (**i.e.** extension of concept of orthology to multiple species). An orthogroup is always defined by a reference speciation event!

****In phylogenomics one of the most common things is to use only 1-to-1 ortologs**** (Orthogroups with only one copy for each specie)
****However the study of gene families (paralogs + orthologs genes) evolutionary dynamics is getting more and more attention in the latest years****

An orthogroups is a group of orthologs genes descending from the **last common ancestor** (LCA) of a group of species (*i.e.* the extension of the concept of orthology to multiple species).

An orthogroup is always defined by a reference speciation event!

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)

Just take in mind the extremely importance of the **bi-directional best hit**, indeed is the only way to take into account possible gene loss in one of the two lineages. (Think about what can happen in orthology inference if the blue human gene in figure 1A would be lost...)

FINISH