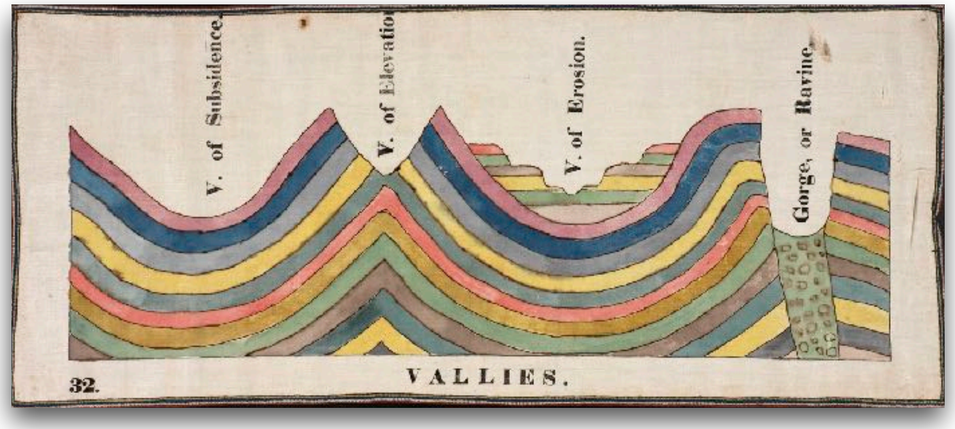
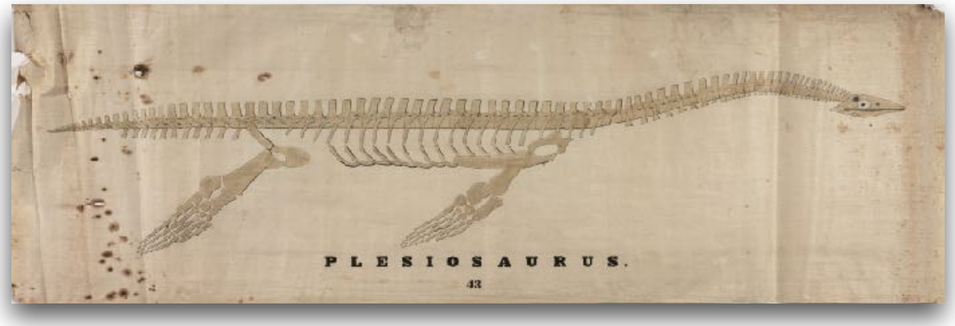




biases in phylogenetics

STOCHASTIC BIAS

STOCHASTIC BIAS:

- random noise introduced by limited data can outweigh the true phylogenetic signal
- … this can happen even when the model is correct 😱

CAUSES:

- insufficient sequence length or number of genes
- low phylogenetic signal (few informative sites)
- high variance in substitution processes

MITIGATING STOCHASTIC BIAS:

- use longer alignments and more characters to average out noise
- focus on genes with high phylogenetic signal - we will see that in the practicals 👀
- filter hyper variable / poorly aligned and conserved / uninformative portions of the alignment
- use branch support metrics to assess confidence in inferred relationships

some guidelines / ballpark numbers from literature

- **dozens of genes** can resolve shallow divergences - recent speciation events - but deeper or more complex trees require hundreds of loci …
- **100–500 genes** significantly mitigate stochastic errors in resolving phylogenies at intermediate evolutionary depths
- **thousands of genes** are recommended for genome-scale studies and are recommend to recover accurate phylogenetic signals across a range of divergence times

SYSTEMATIC BIAS

MODEL ASSUMPTIONS AND VIOLATIONS IN PHYLOGENETICS

Most models of sequence evolution assume that the process is:

- **Stationary** – base or amino acid frequencies remain constant over time
- **Reversible** – the process looks the same forward and backward in time
- **Homogeneous** – the same model applies across all branches of the tree

These are collectively known as SRH ASSUMPTIONS.

SRH assumptions simplify inference, but may not reflect biological reality:

- violations of SRH assumptions are common in real datasets
- such violations can lead to systematic bias and incorrect topologies

MITIGATING MODEL VIOLATIONS:

- test for model violations prior to tree reconstruction.
- do phylogenetic subsampling and exclude problematic partitions
- apply non-SRH models such as non-reversible models - btw we are skipping them 😛
- use partitioned analyses or profile mixture models to better fit real data

SATURATION:

- sequences accumulate multiple substitutions at the same sites
- divergence underestimated - multiple substitutions along a branch interpreted as one or few
- homoplasy increases - identical states arise independently in different lineages
- Leads to a loss of phylogenetic signal, especially at deeper divergences.

CAUSES:

- high substitution rates relative to divergence time in the characters used
- long evolutionary timescales without model correction

MITIGATING SATURATION:

- choose less saturated and slowly evolving genes or regions
- use amino acid alignments over nucleotides for deep phylogenies
- apply models that account for multiple hits and rate heterogeneity (e.g. gamma)
- use amino acid (e.g., Dayhoff6, SR4) or nucleotides (RY) recoding

RATE HETEROGENEITY - LONG BRANCH ATTRACTION:

- in long branches, the probability of multiple substitutions at the same site increases
- excess of subs. leads to some parallel ones that falsely inflate similarity of unrelated taxa

CAUSES:

- uneven substitution rates across taxa
- … coupled with long evolutionary timescales

MITIGATING LBA:

- use better-fitting models (like CAT, GTR+Γ) that better account for rate variation
- avoid parsimony and use ML or BI methods
- if LBA is suspected and unsolvable, consider excluding the long-branched taxa
- add intermediate taxa to break long branches reduces the chance of convergent substitutions being interpreted as shared ancestry

COMPOSITION HETEROGENEITY

- phylogenetics assume that nt and aa composition is homogeneous across lineages
- taxa that evolve with distinct frequencies cluster based on composition rather than ancestry 🥵

CAUSES:

- varying GC content from mutation bias
- lifestyle shifts (thermophily, parasitism, endosymbiosis)
- different metabolic constraints across lineages

MITIGATING COMPOSITION HETEROGENEITY:

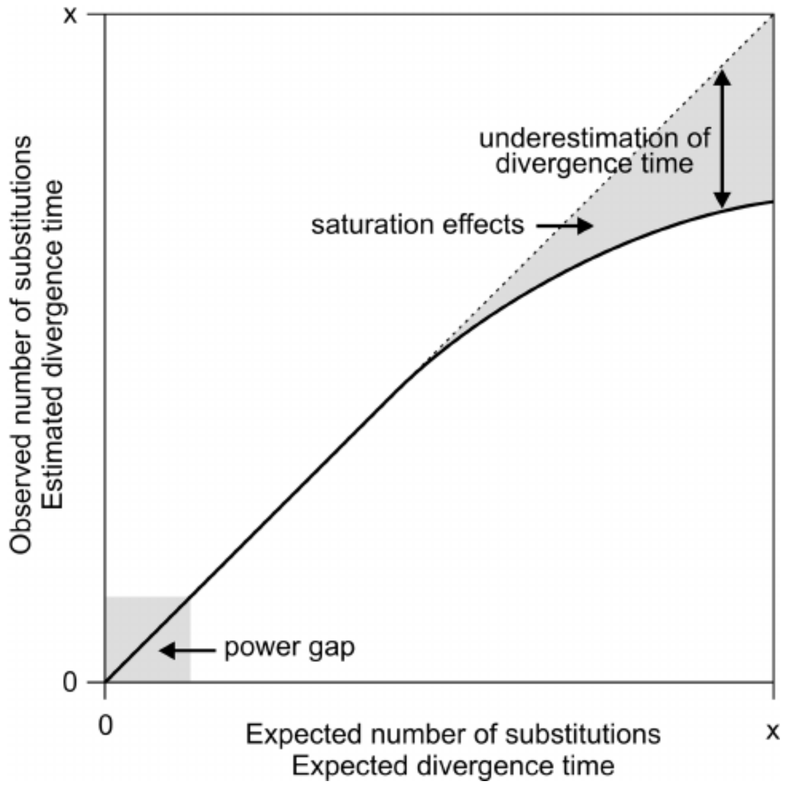
- add intermediate taxa to break long branches
- use better-fitting models - more specifically profile mixture models
- use amino acid (e.g., Dayhoff6, SR4) or nucleotides (RY) recoding
- If certain taxa strongly deviate in composition and distort topology, exclude them.

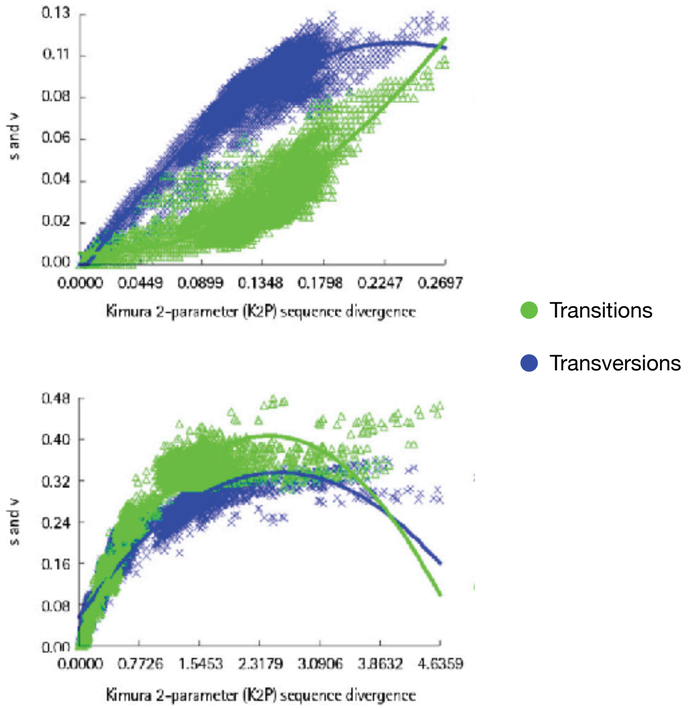
bias	stochastic	systematic
cause	random error due to limited data	incorrect model and assumptions
consistency	random; decreases with more data	persistent; remains despite more data
mitigation	increase data quantity (genes, taxa)	appropriate models and assumptions

INTERPRETING A SATURATION PLOT:

- Early divergence (left side): substitutions increase linearly; good phylogenetic signal.
- Later divergence (right side): substitution rate slows and plateaus—this indicates saturation.

Typically, **transitions** saturate faster than **transversions** because they're biochemically more frequent (purine ↔ purine, pyrimidine ↔ pyrimidine).





FINISH