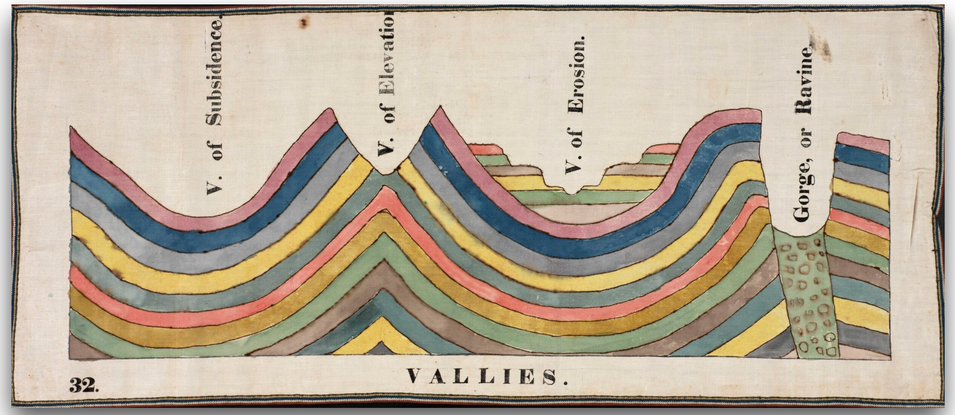
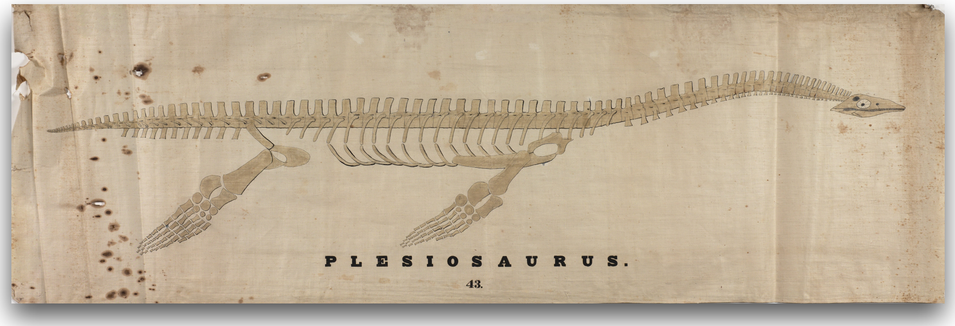
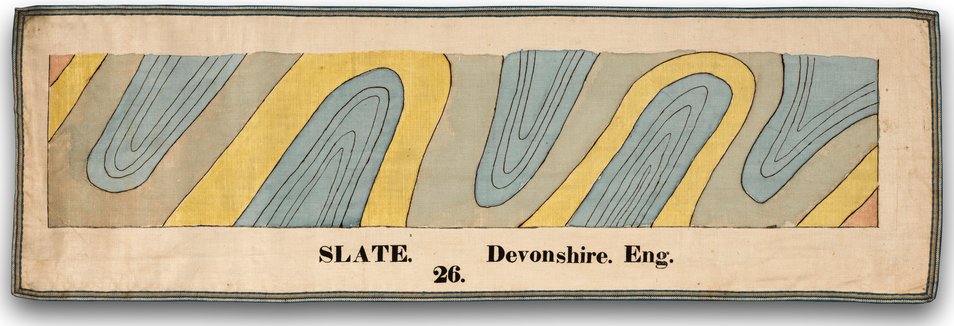






biases in phylogenetics

stochastic bias

How stochastic bias happens:

- Random noise introduced by limited data can outweigh the true phylogenetic signal.
- This can happen even when the model is correct.

Common 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 loci to average out noise - phylogenomics!
- Focus on genes with high phylogenetic signal - we will see that in the practicals!
- Filter hyper variable / poorly aligned and conserved / uninformative portions of alignement.
- Use branch support metrics to assess confidence in inferred relationships.
- Adding more taxa to break up long branches can reduce stochastic noise.

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 require 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 violation:

- Test for model violations prior to tree reconstruction.
- Do phylogenetic subsampling and exclude problematic partitions - will see in the practicals!
- Apply non-SRH models when appropriate such as non-reversible models.
- Use partitioned analyses or profile mxiture models to better fit real data.

- unmodelled **heterogeneity in rates** - also known as Long Branch Attraction or LBA.

  Can mislead tree reconstruction methods - especially parsimony - so that fast-evolving lineages may be incorrectly inferred as closely related.

- unmodelled **heterogeneity in composition**

  Differences in base or amino acid composition across taxa can cause unrelated lineages with similar compositions to cluster together.

How LBA happens:

- 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.

Common causes:

- of course uneven rates of evolution across taxa
- poor taxon sampling - missing intermediate branches
- deep divergence times

Mitigating LBA:

- Improved Taxon Sampling and add intermediate taxa to break long branches Reduces the chance of convergent substitutions being interpreted as shared ancestr

- Use Better-Fitting Models Employ e.g., site-heterogeneous models like CAT, GTR+Γ) These better account for rate variation and reduce LBA effects

- Avoid or Supplement Parsimony Use likelihood- or Bayesian-based methods with model testing Parsimony is more sensitive to LBA, especially with high divergence

- Remove Problematic Taxa (with caution!) If LBA is suspected and unsolvable, consider excluding the long-branched taxa.

How heterogenity in compositions happens:

- Phylogenetics assume that nt and aa composition is homogeneous across lineages.
- Taxa that evolve with distinct frequencies cluster based on composition rather than ancestry.

Common causes:

- of course uneven rates of evolution across taxa
- poor taxon sampling - missing intermediate branches
- deep divergence times

Mitigating heterogenity in compositions:

- Add intermediate taxa to break long branches Reduces the chance of convergent substitutions being interpreted as shared ancestr

- Use better-fitting models - more specifically profile mixture models - these better account for rate variation and reduce LBA effects

- Use amino acid recoding (e.g., Dayhoff6, SR4) or RY-coding for nucleotides to reduce noise from compositional differences.

- If certain taxa strongly deviate in composition and distort topology, consider excluding them after sensitivity analyses.

+ unmodelled saturation

- Sequences accumulate multiple substitutions at the same sites over time.
- Divergence underestimation: multiple substitutions along a branch interpreted as one or few.
- Homoplasy increase: identical states arise independently in different lineages.
- Leads to a loss of phylogenetic signal, especially at deeper divergences.
- Results in underestimation of true evolutionary distances and produces misleading topologies

Common causes:

- High substitution rates relative to divergence time
- Use of highly variable or fast-evolving markers
- Long evolutionary timescales without model correction
- Relying on nucleotide data for deep splits

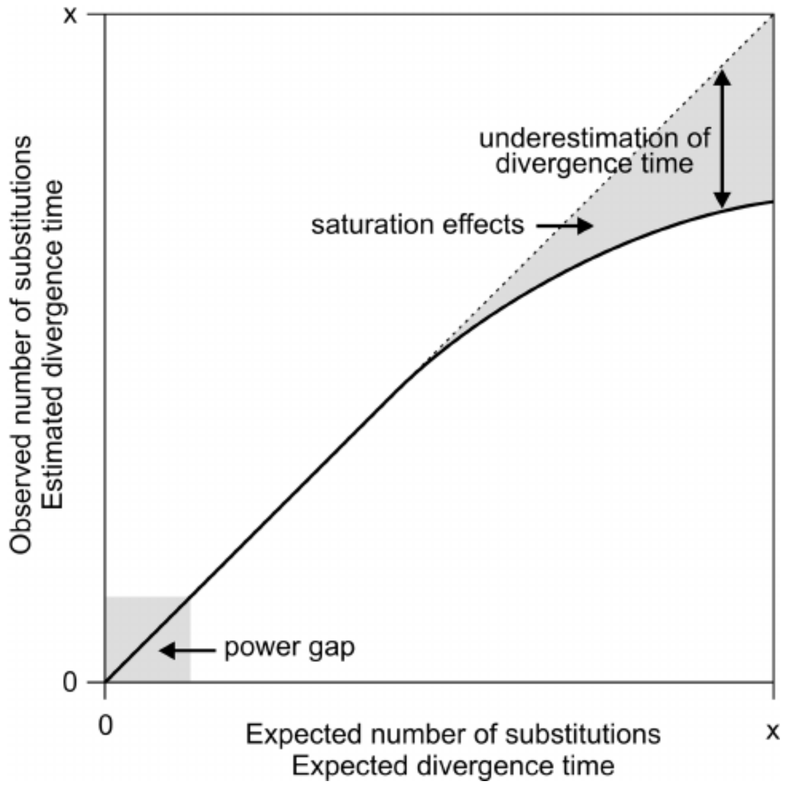
Mitigating Saturation:

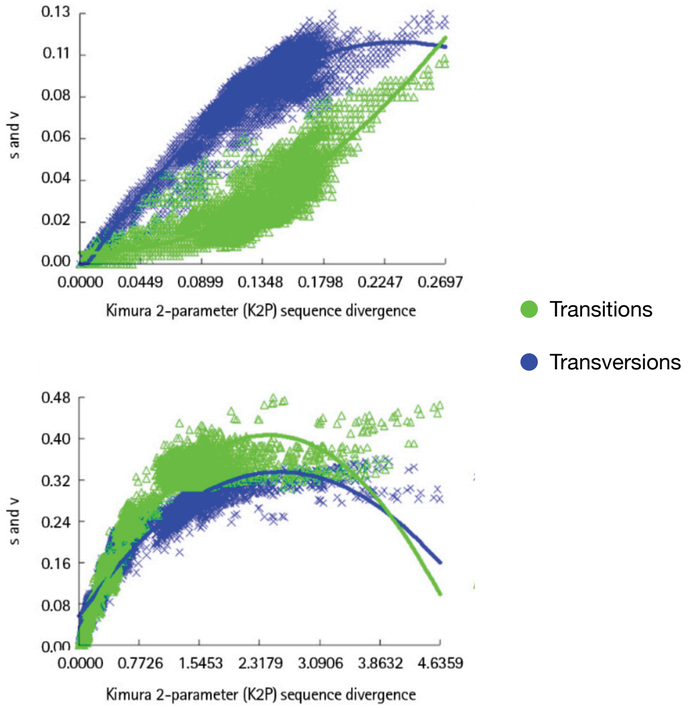
- Choose less saturated and slowly evolving genes or regions
- Use amino acid alignments over nucleotides for deep phylogenies.
- Apply models that account for multiple hits and rate heterogeneity (gamma).
- Recoding strategies as RY coding (nucleotides) or Dayhoff-type recoding (proteins).

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).





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

FINISH