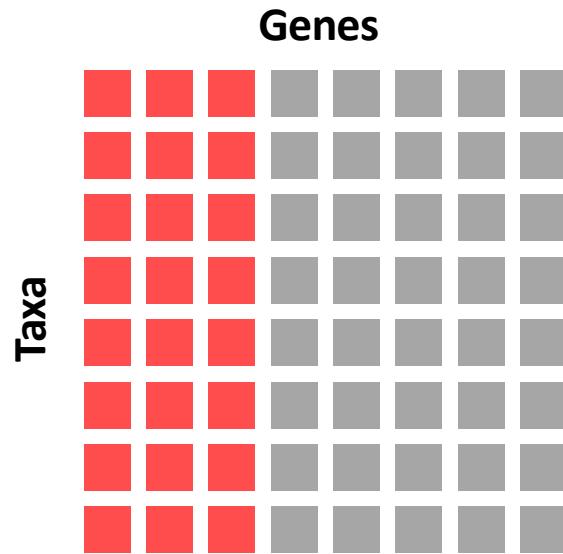

Lecture 2.1

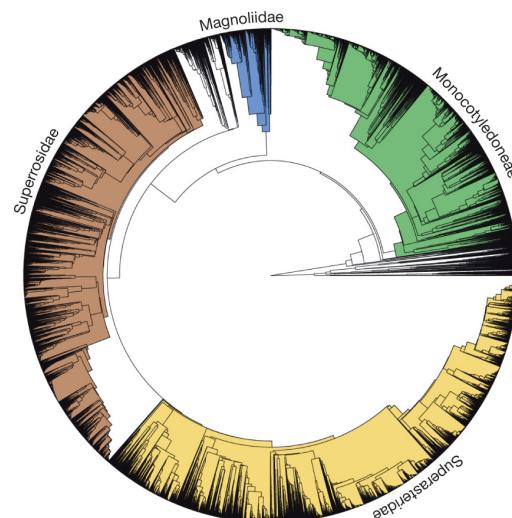
Phylogenomics

Analysing Large Data Sets

Large data sets

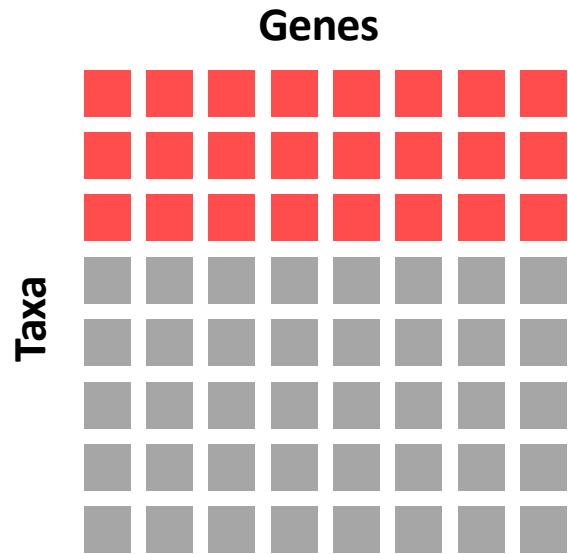


- Tree-space is extremely large
 - Efficient tree-searching heuristics

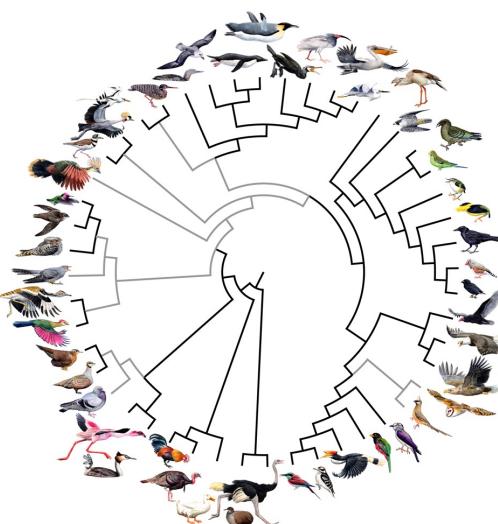


32,223 taxa
7 genes

Large data sets

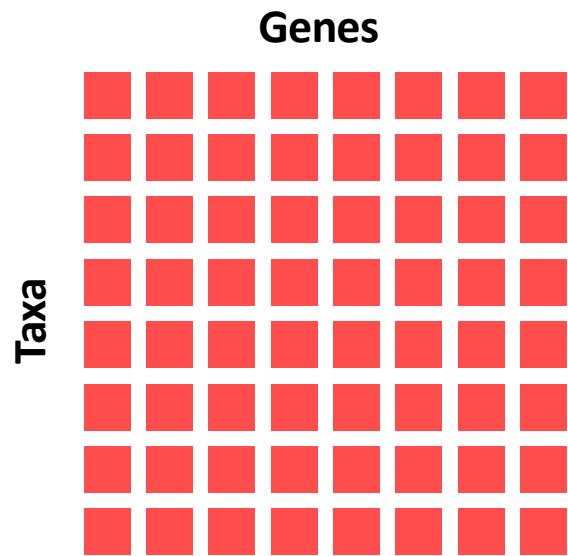


- Calculation of likelihood is expensive
 - Approximate likelihood calculation
 - Multithreading/parallelisation



48 taxa
8,295 genes

Large data sets

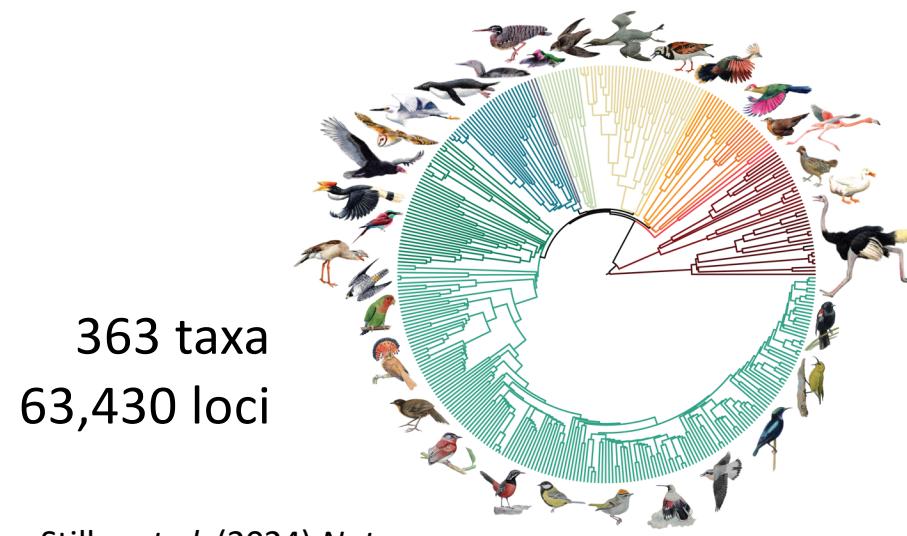


- Analysis is computationally expensive



9506 taxa
353 genes

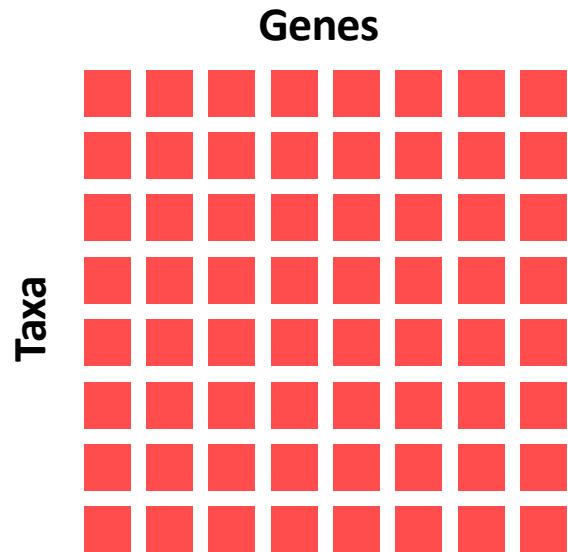
Zuntini et al. (2024) *Nature*



363 taxa
63,430 loci

Stiller et al. (2024) *Nature*

Large data sets

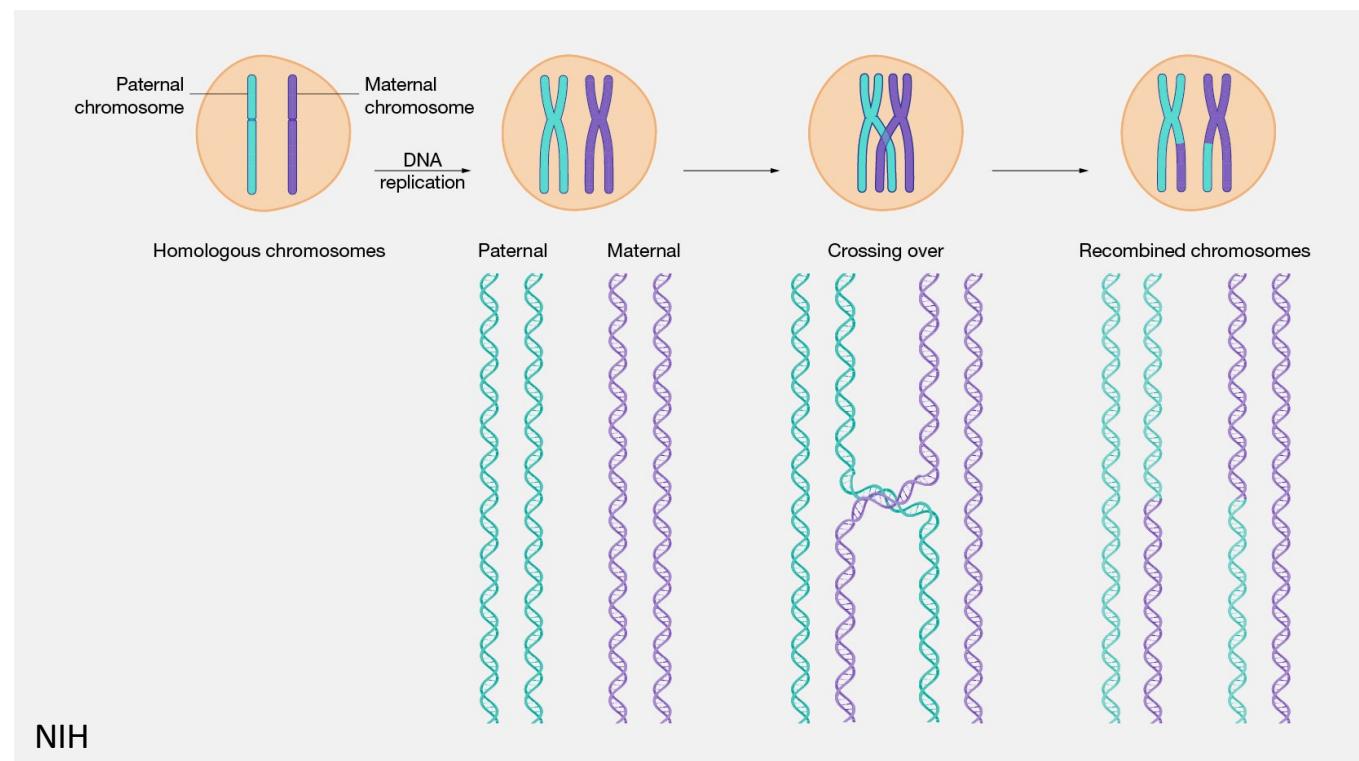


- Analysis is computationally expensive
- Consider filtering the data
 - Phylogenetic signal
 - Substitution saturation
 - Missing data
 - Coding or non-coding sequences
 - Random subsample
- Dividing the tree into subtrees

Gene Trees

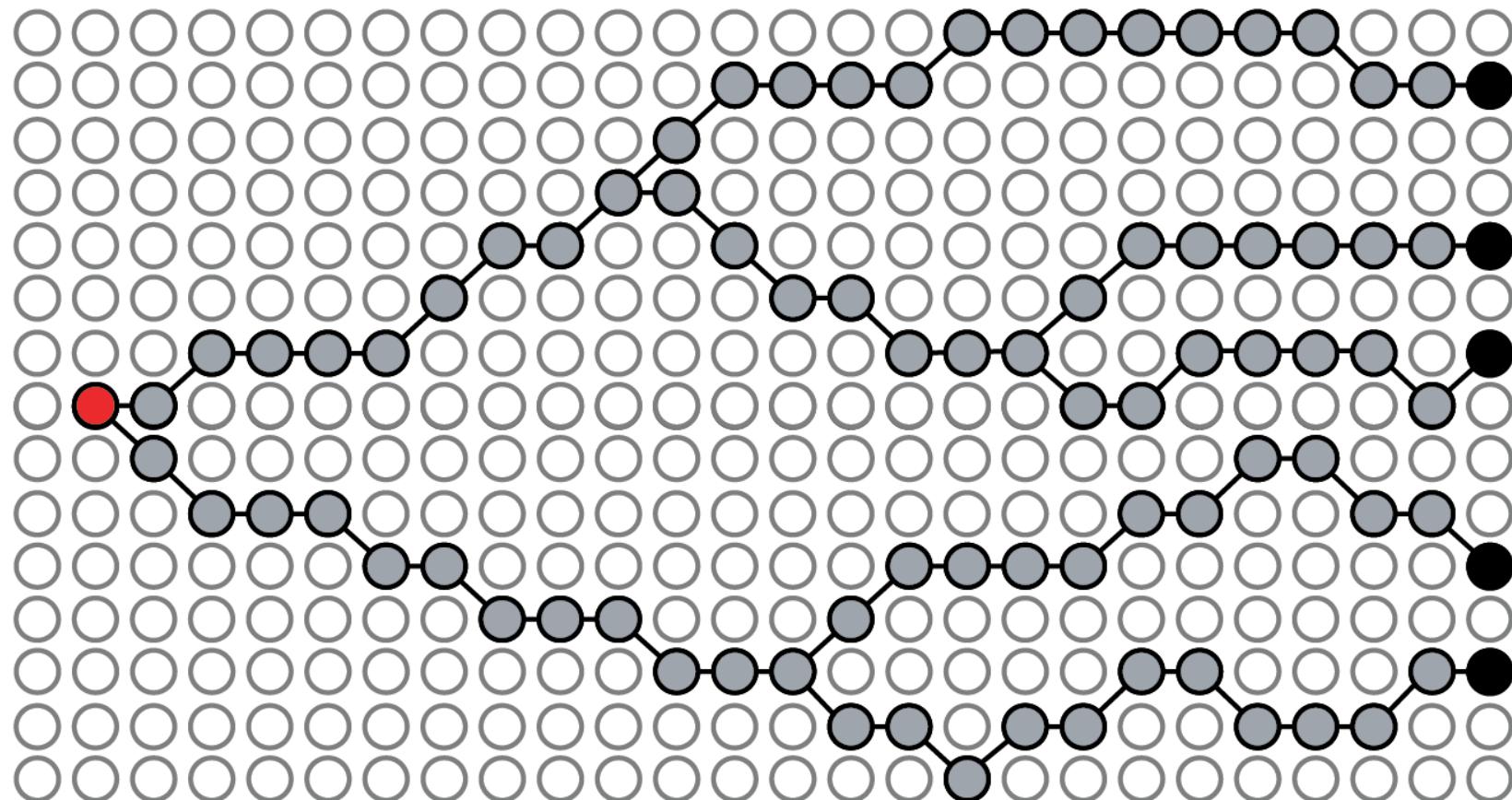
Gene trees

- Many phylogenetic methods assume that there is a single tree that describes the evolution of the whole data set
- But recombination complicates this



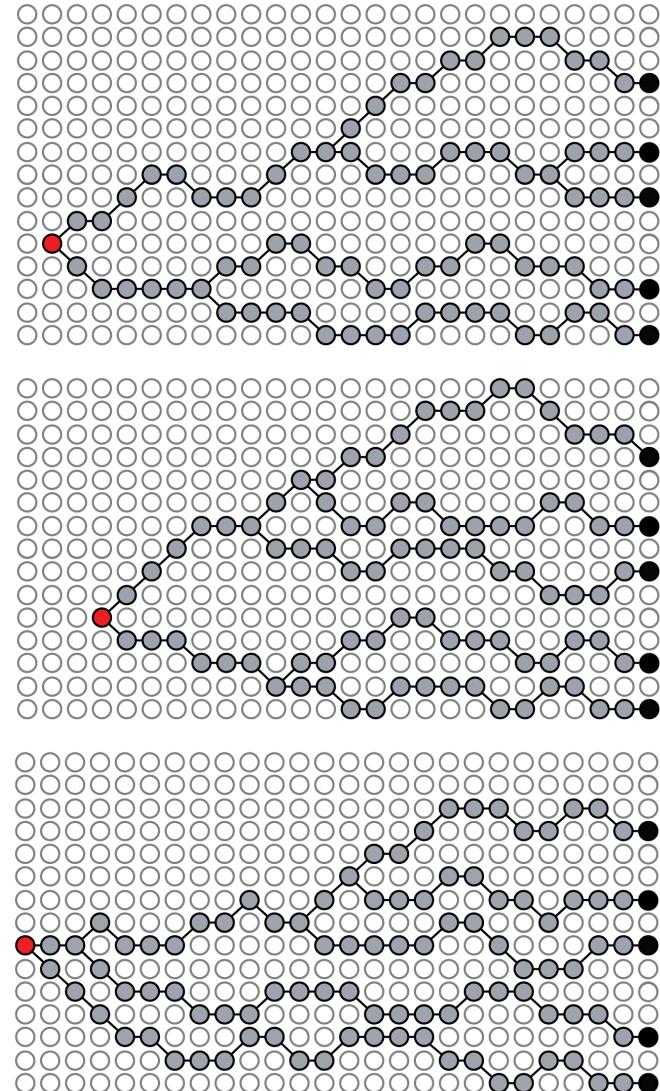
Coalescent theory

Past ← Generations → Present

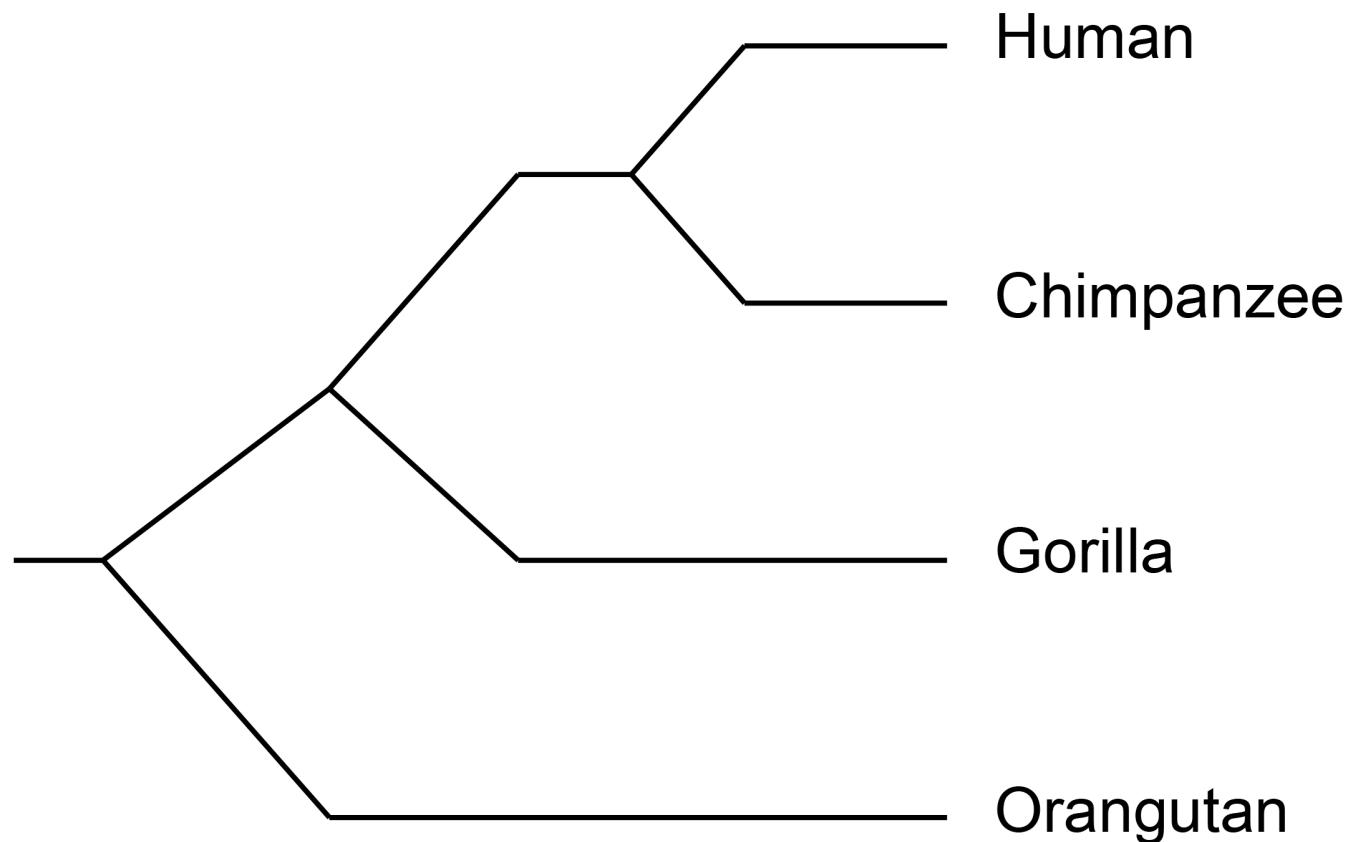


Gene trees in a species

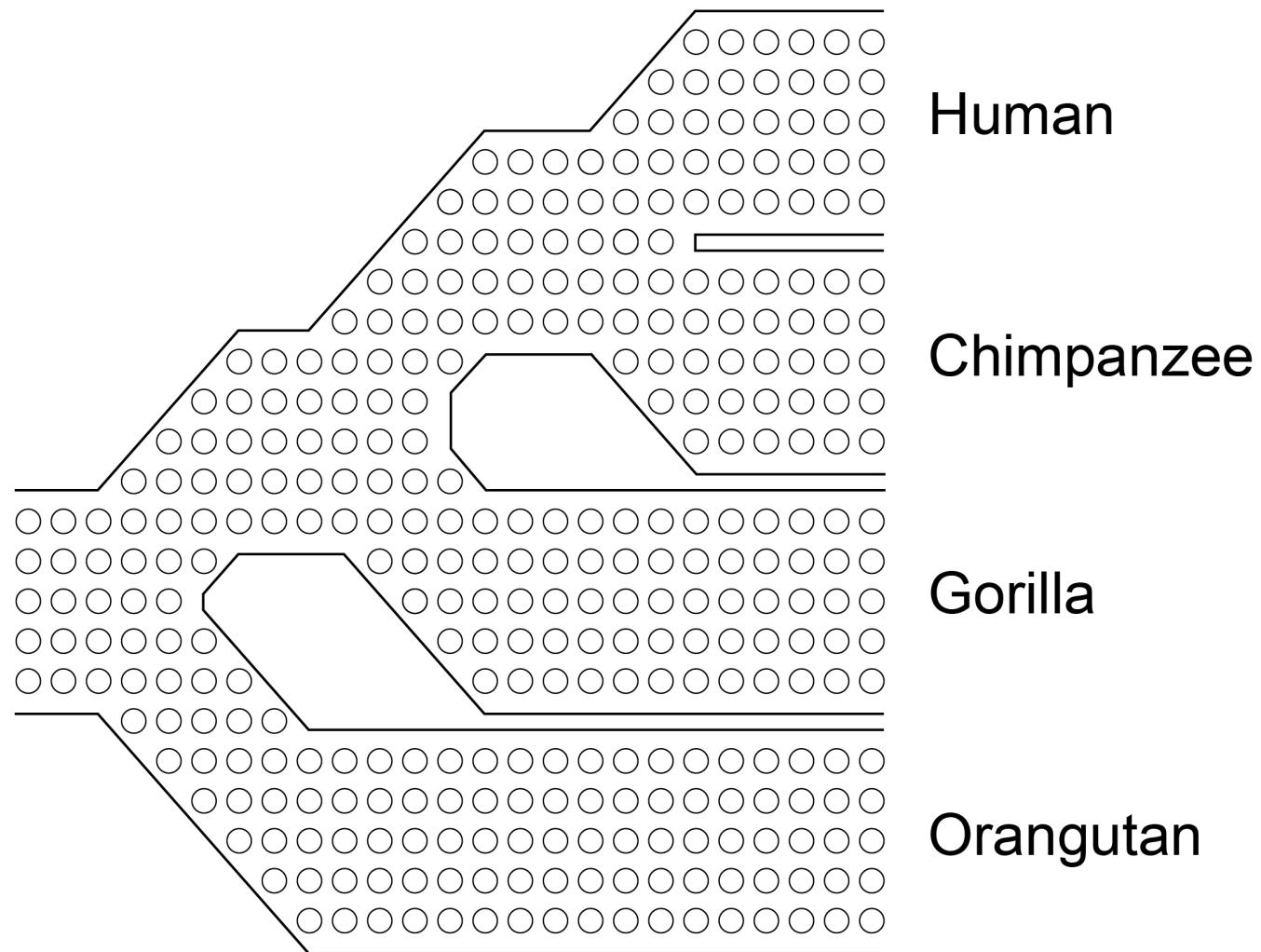
- Gene trees vary by chance among unlinked genes
 - Different trees
 - Different timescales



Species tree

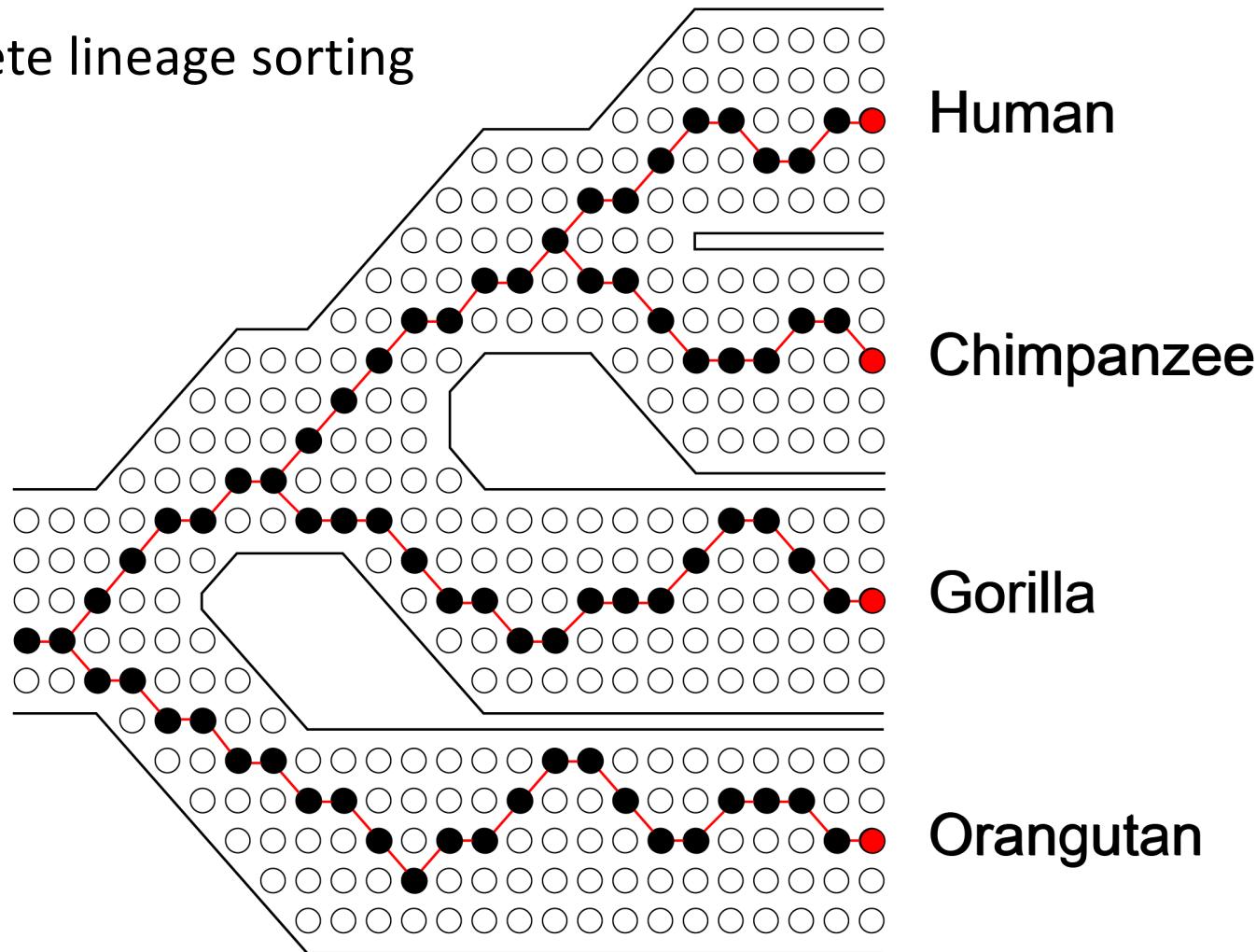


Multispecies coalescent



Gene tree (concordant)

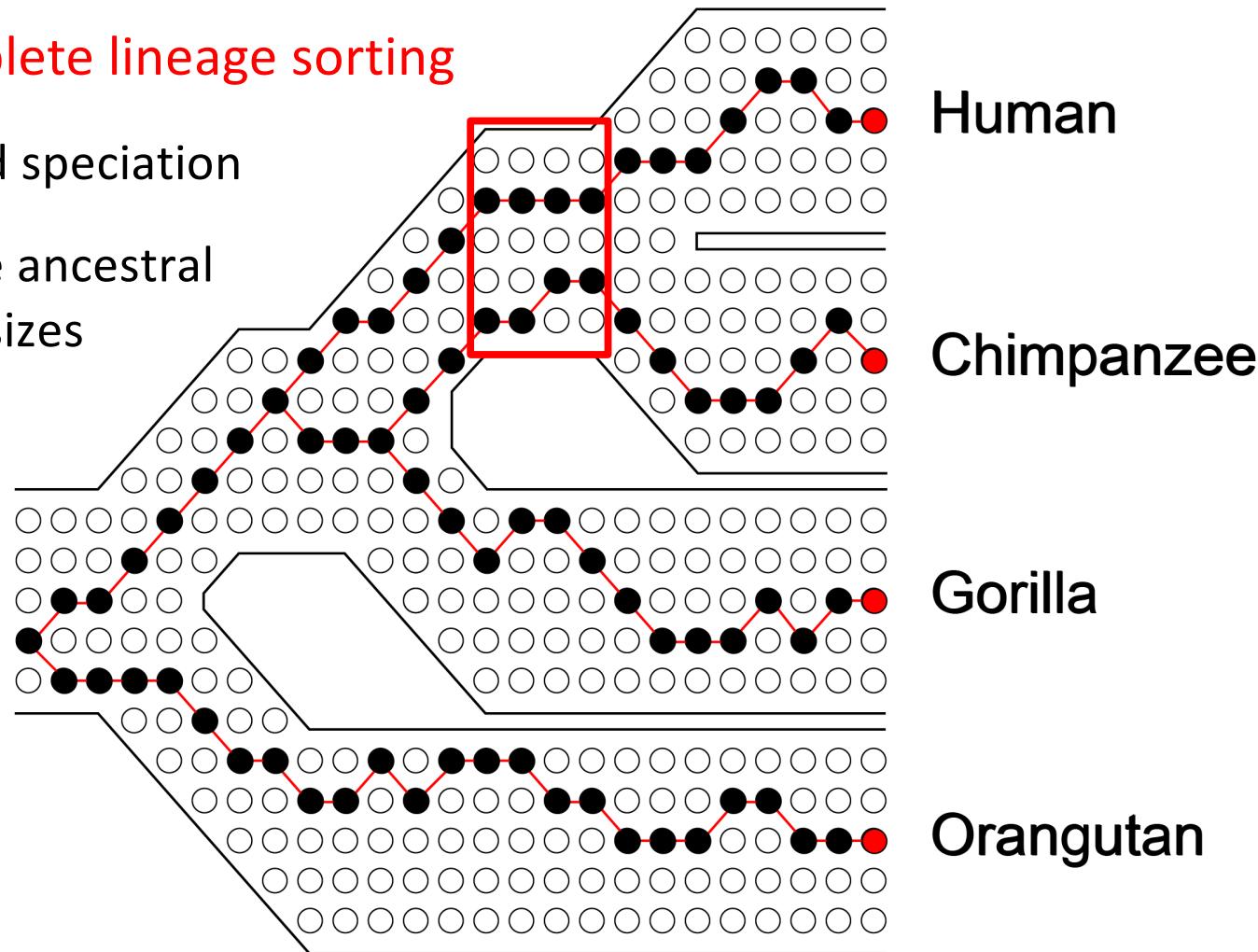
Complete lineage sorting



Gene tree (discordant)

Incomplete lineage sorting

- Rapid speciation
- Large ancestral pop sizes



Incongruence among gene trees

- Phylogenetic analyses of genome-scale data sets must deal with incongruence among gene trees
 - **Incomplete lineage sorting**
 - Different direction and strength of selection
 - Stochastic variation in the mutational process
 - Biases in nucleotide composition

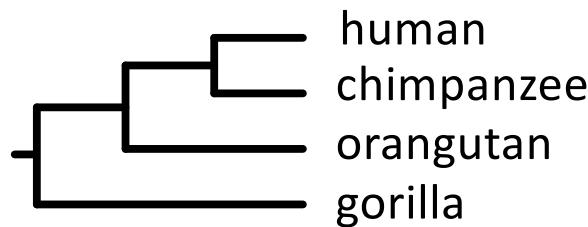
Inferring the species tree

- **Incomplete lineage sorting** can lead to gene trees that do not match the species tree
- We can infer the species tree from multiple gene trees even when they are incongruent
- Three approaches
 1. Consensus
 2. Concatenation
 3. Coalescent

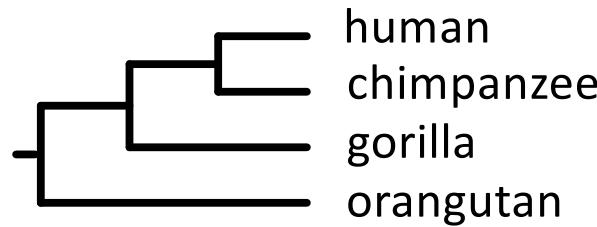
Species tree

1. Consensus

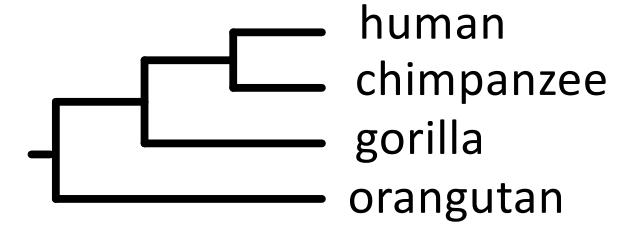
Estimate genealogy from each gene and find the consensus



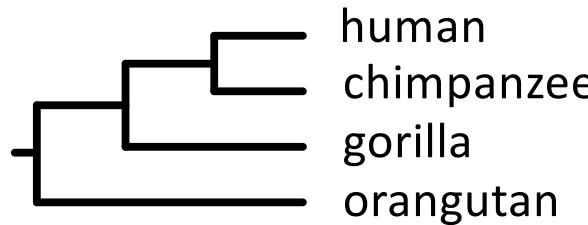
Gene 1



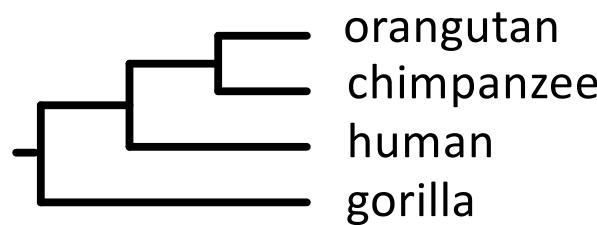
Gene 2



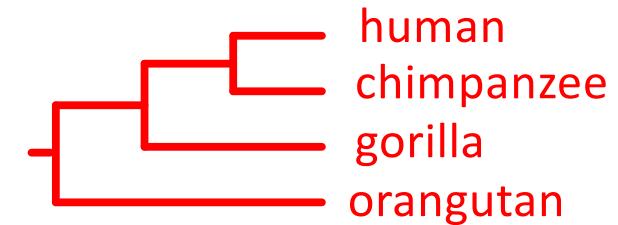
Gene 5



Gene 4



Gene 5



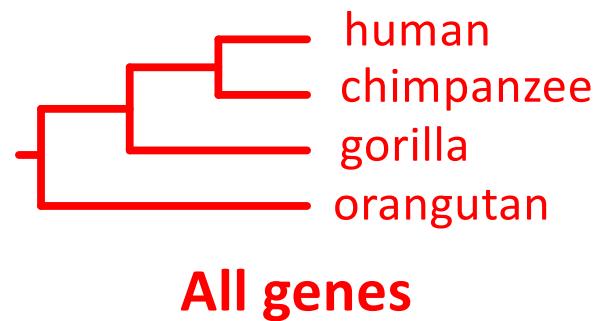
Consensus

But the most frequent gene tree does not always
match the true species tree (“anomaly zone”)

Analysing multiple loci

2. Concatenation

Assume that all genes share the same evolutionary history



But this ignores the occurrence of different gene trees

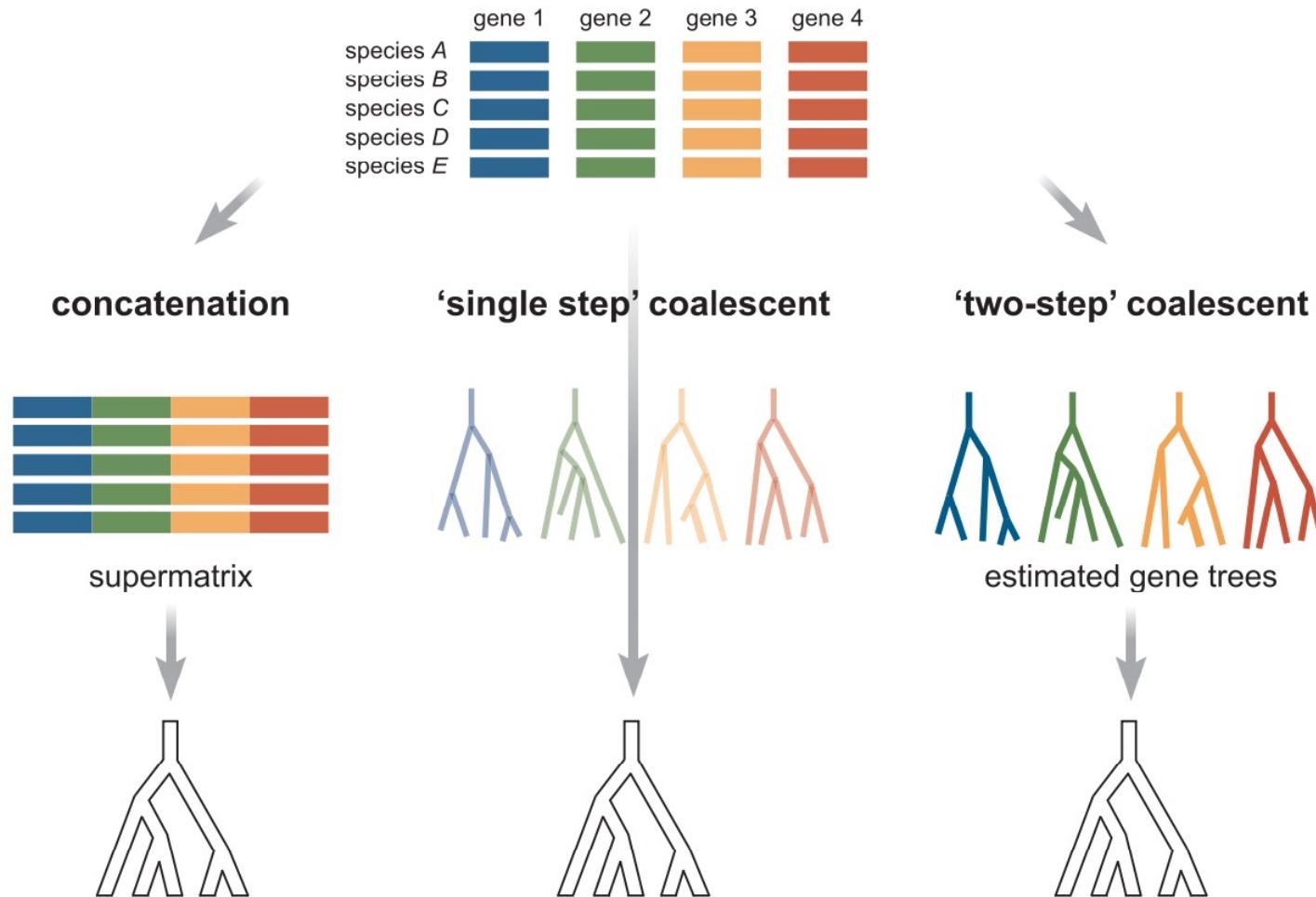
Species tree

3. Coalescent-based methods

Estimate the species tree based on gene trees

- Gene trees are independent realisations of a stochastic process (the coalescent) on the same species tree

Species tree



ASTRAL

RESEARCH

ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees

Chao Zhang³, Maryam Rabiee², Erfan Sayyari¹ and Siavash Mirarab^{1*}

- Accurate Species TRee ALgorithm
- Finds the species tree with the highest agreement with ‘quartets’ among the gene trees
- Can rapidly analyse a genome-scale data set

Inferring the species tree

- Choosing between concatenation and coalescent-based approaches
- **Shallower timescales:** gene trees inferred accurately but incomplete lineage sorting is important
- **Deeper timescales:** gene trees are inferred less accurately but incomplete lineage sorting is less important
- The choice of methods to use should be informed by the largest sources of error

Useful references

- **Estimating phylogenetic trees from genome-scale data**
Liu *et al.* (2015) *Annals New York Acad Sci*, 1360: 36–53.
- **The concatenation question**
Bryant & Hahn (2020) In: *Phylogenetics in the Genomic Era*.
- **Phylogenetic tree building in the genomic age**
Kapli *et al.* (2020) *Nat Rev Genet*, 21: 428–444.