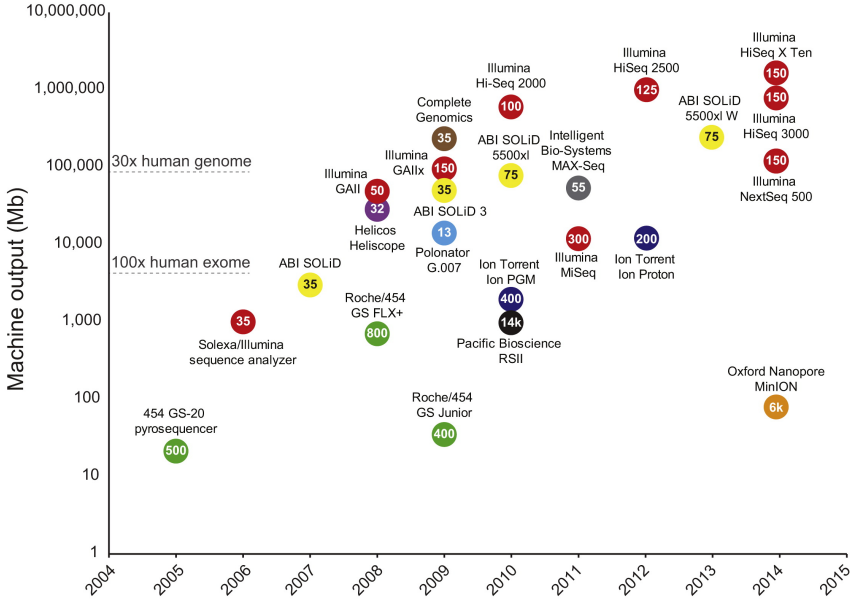
Lecture 3.1

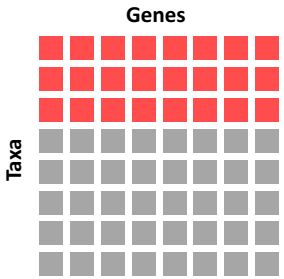
Phylogenomics

Simon Ho

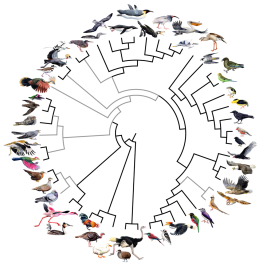


Reuter *et al.* (2015) *Molecular Cell*

Large data sets

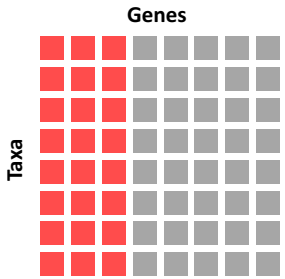


- Calculation of likelihood is expensive
  - Speed up by grouping sites with identical patterns
  - Approximate likelihood calculation
  - Multithreading/parallelisation

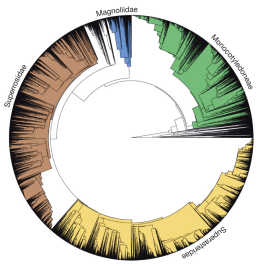


48 taxa
8,295 genes
Jarvis *et al*. (2014) *Science*

Large data sets



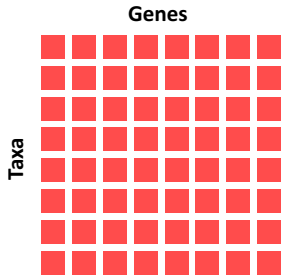
- Tree-space is extremely large
  - Efficient tree-searching heuristics



32,223 taxa
7 genes

Zanne *et al*. (2014) *Nature*

Large data sets

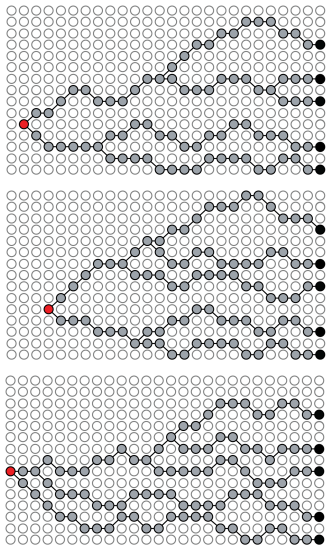


- Analysis is computationally expensive
- Consider filtering the data
  - Phylogenetic signal
  - Mutational saturation
  - Missing data
  - Model fit

Genome-Tree Incongruence

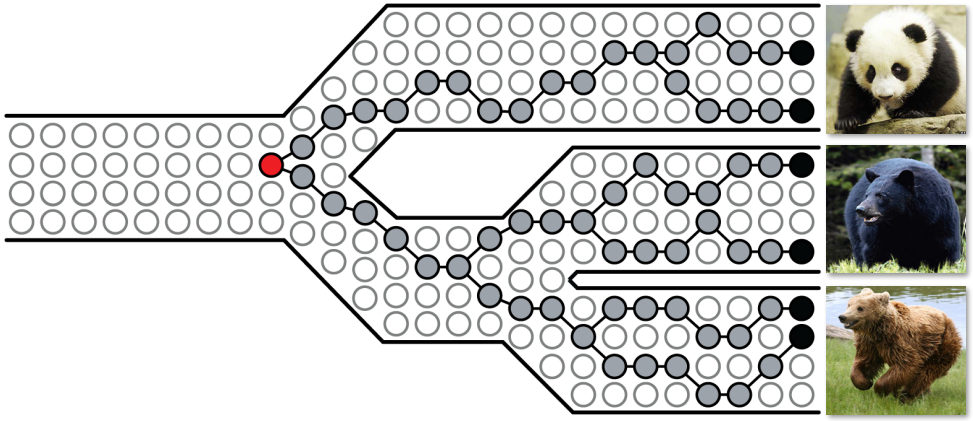
Gene trees in a species

- Genealogies vary stochastically among unlinked loci
- Should not concatenate independent loci in a phylogenetic analysis of intraspecific data
  - Different trees
  - Different coalescence times



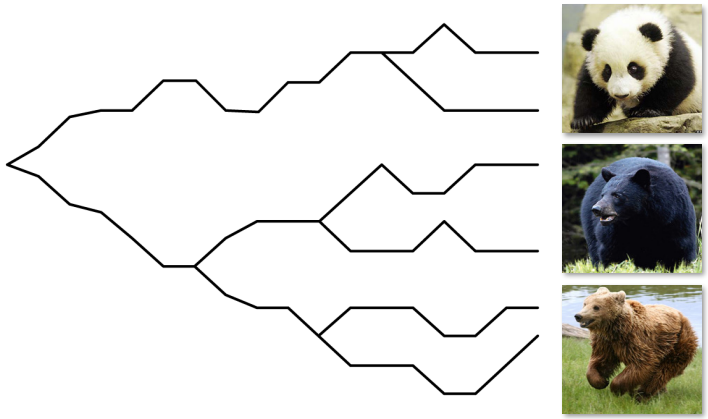
Gene trees in multiple species

- Gene trees are embedded in the species tree



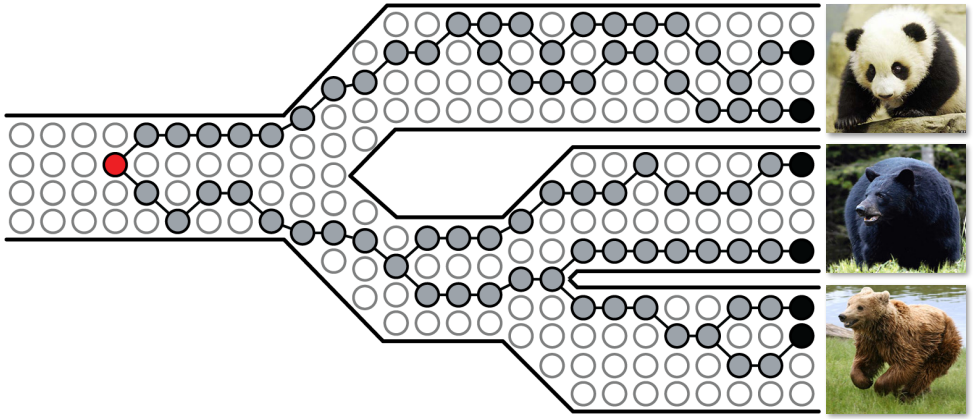
Gene trees in multiple species

- Gene trees are embedded in the species tree



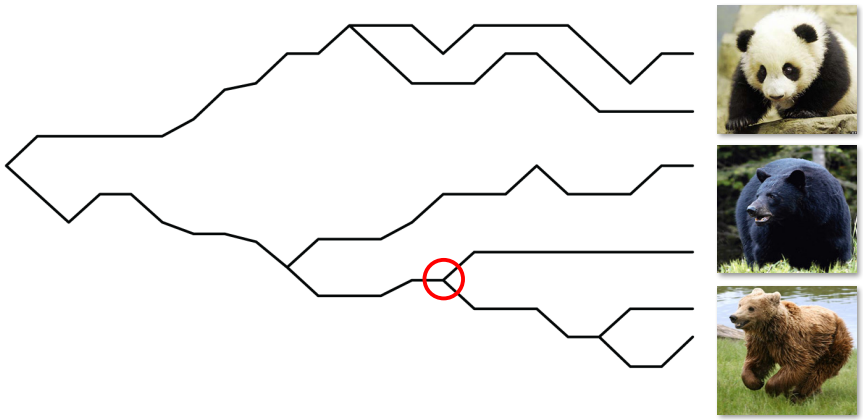
Gene trees in multiple species

- Incomplete lineage sorting



Gene trees in multiple species

- Incomplete lineage sorting

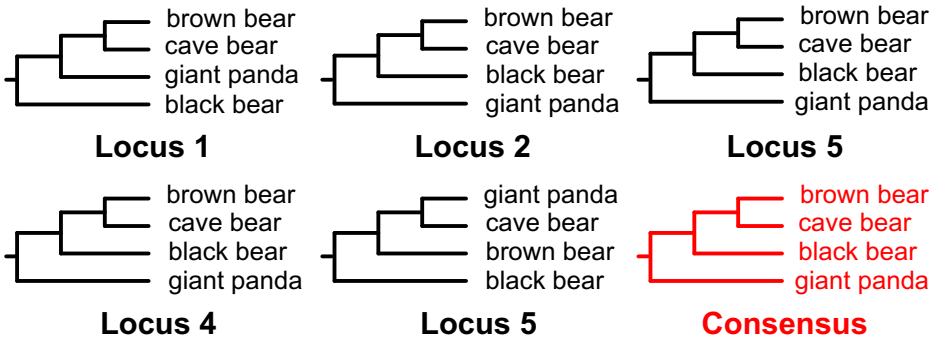


Species tree

- **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
- Three approaches
  1. Consensus
  2. Concatenation
  3. Species-tree methods

Species tree

1. **Consensus**
   Estimate genealogy from each locus and find the 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 loci share the same evolutionary history

But this ignores the occurrence of different gene trees

Species tree

3. **Species-tree 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
- Various methods
  - **BEAST* co-estimates gene trees and the species tree

BIOINFORMATICS

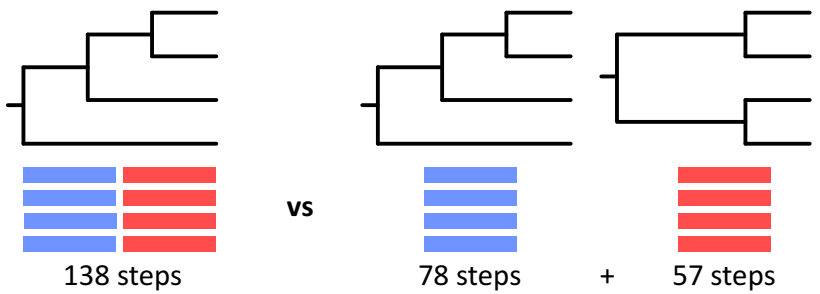
Vol. 30 ECCB 2014, pages i541–i548
doi:10.1093/bioinformatics/btu462

ASTRAL: genome-scale coalescent-based species tree estimation

S. Mirarab[1], R. Reaz[1], Md. S. Bayzid[1], T. Zimmermann[1,2], M. S. Swenson[3] and T. Warnow[1,*]

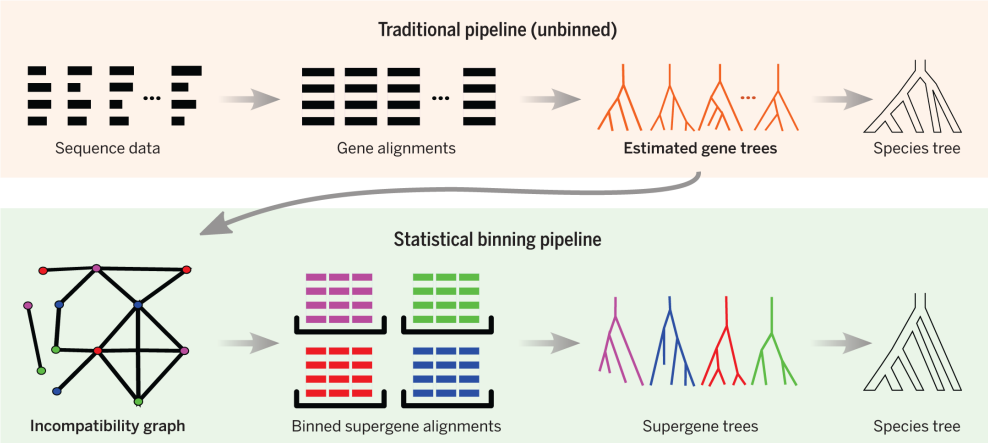
Partition-homogeneity test

- Test for phylogenetic congruence across markers
- Partition-homogeneity (incongruence length difference) test



Species tree

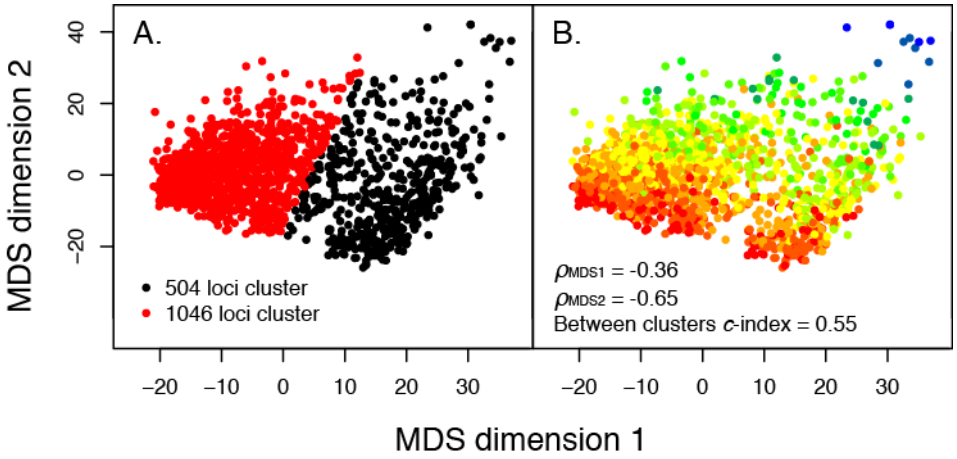
- Statistical binning



Mirarab *et al*. (2014) *Science*

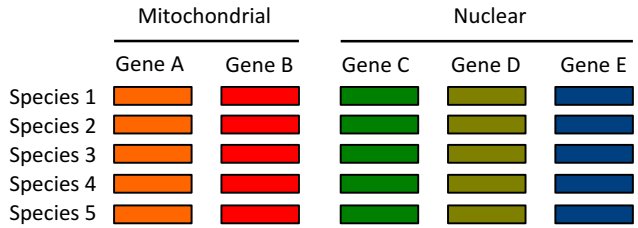
Species tree

- Topology clustering



Duchene *et al*. (submitted)

Data partitioning



Lanfear *et al. BMC Evolutionary Biology* 2014, **14**:82
http://www.biomedcentral.com/1471-2148/14/82

BMC Evolutionary Biology

METHODOLOGY ARTICLE **Open Access**

Selecting optimal partitioning schemes for phylogenomic datasets

Robert Lanfear[1,2*†], Brett Calcott[3†], David Kainer[1], Christoph Mayer[4] and Alexandros Stamatakis[5,6]

Genome-Scale Dating

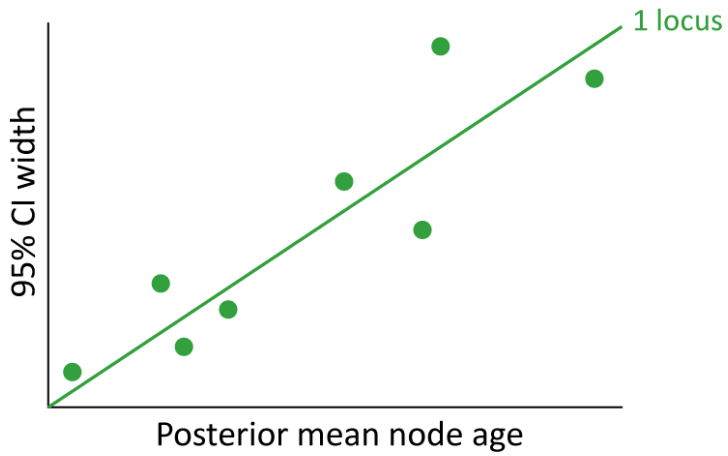
Infinite-sites theory

- In Bayesian dating, the calibration priors provide the only information about absolute times
- With increasing data, there is no guarantee that posterior node ages will converge to the true values

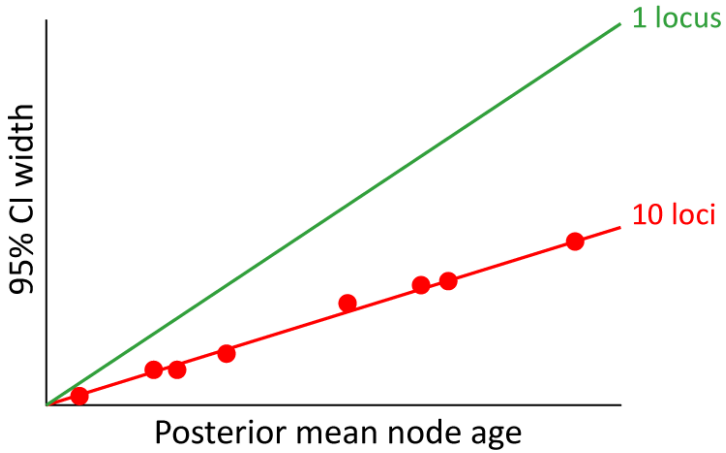
Infinite-sites plot
Plot of 95% credibility interval widths *vs* posterior means of node ages

Yang & Rannala (2006) *Mol Biol Evol*

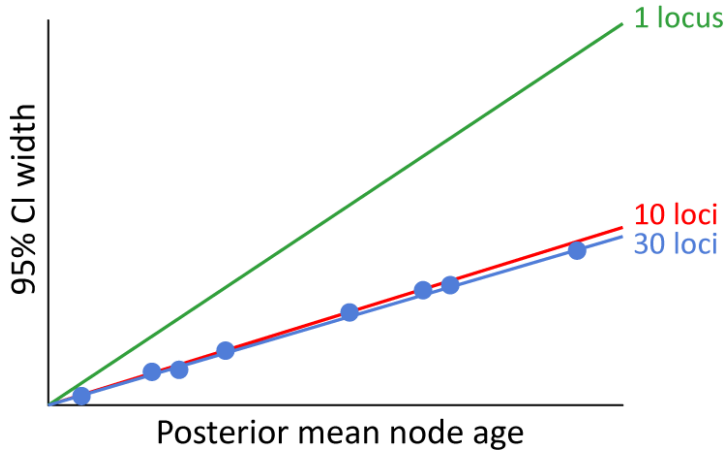
Infinite-sites theory



Infinite-sites theory



Infinite-sites theory

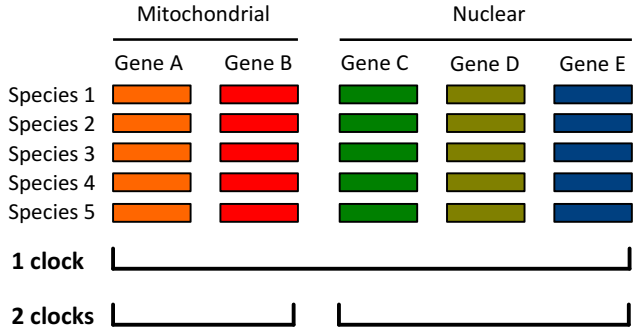


Infinite-sites theory

- As the amount of data increases:
  - Approaches a straight line
  - Decline in the slope
- Moderate amounts of data behave similarly to infinite data
- Note that the slope does not reach zero with infinite data

Even with infinite data
Uncertainty in date estimates depends on uncertainty in calibrations

Data partitioning



BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 7 2014, pages 1017–1019
doi:10.1093/bioinformatics/btt665

Phylogenetics
Advance Access publication November 14, 2013

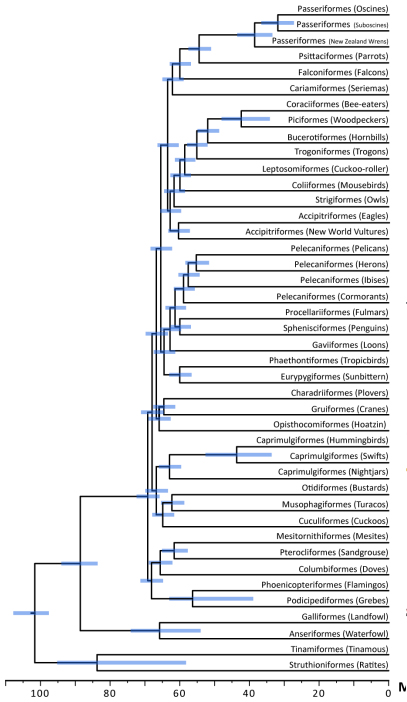
ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis

Sebastián Duchêne*,†, Martyna Molak and Simon Y. W. Ho†

Whole-genome analyses resolve early branches in the tree of life of modern birds

- 8295 genes from 52 taxa
- Identified 1156 clocklike genes
- 20 fossil-based calibrations
- Bayesian dating in *MCMCtree* with approximate likelihood calculation
- Able to test impact of choices of priors and calibrations

Jarvis *et al.* (2014) *Science*



Data filtering

- Genome-scale projects have produced effectively infinite data
- We can be selective about the data that we use
- **Filter the data** according to some criterion
  - Sequence quality
  - Minimise missing data
  - Clocklike evolution
  - Fewest clock models
  - Model adequacy

Useful references

- **Phylogenomic subsampling: a brief review**
  Edwards (2016) *Zoologica Scripta*, 45: 63–74.
- **Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model**
  Song *et al*. (2012) *Proceedings of the National Academy of the USA*, 109: 14942–14947.
- **The changing face of the molecular evolutionary clock**
  Ho (2014) *Trends in Ecology and Evolution*, 29: 496–503.
- **Reconstructing evolutionary timescales using phylogenomics**
  Tong, Lo, & Ho (2016) *Zoological Systematics*, 41: 343–351.