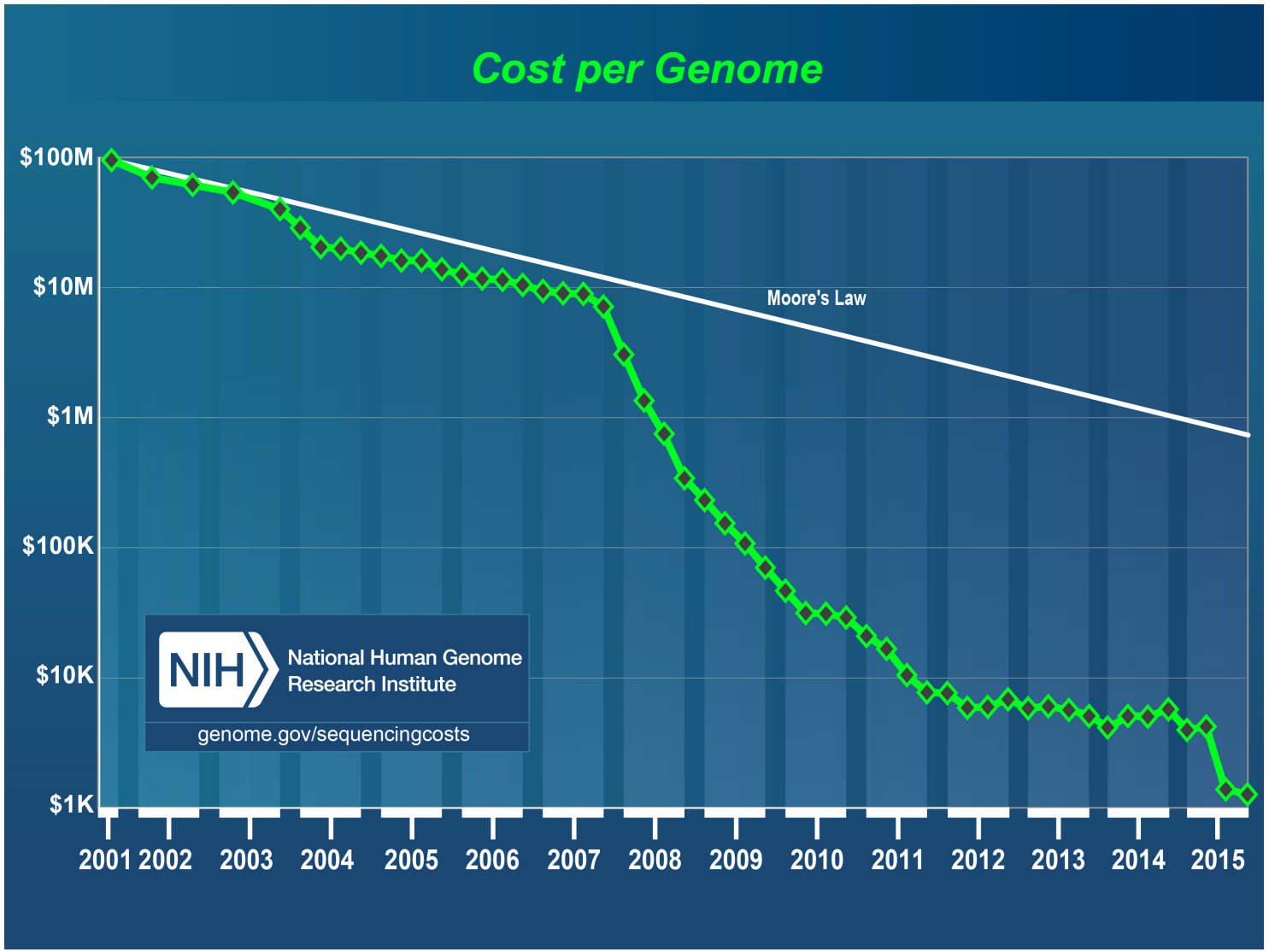
Australian National University

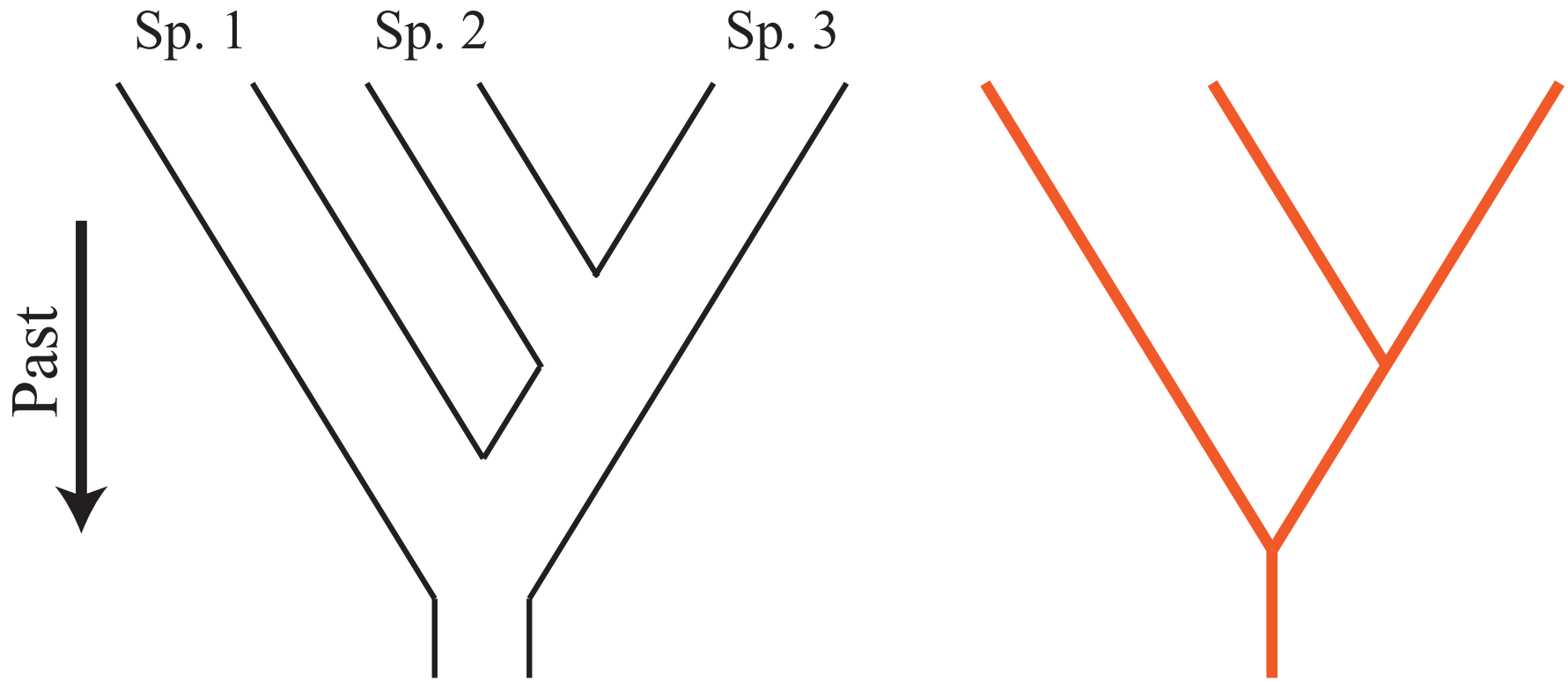
Species trees and Gene trees

David A. Duchêne

Genome sequencing is increasingly within reach

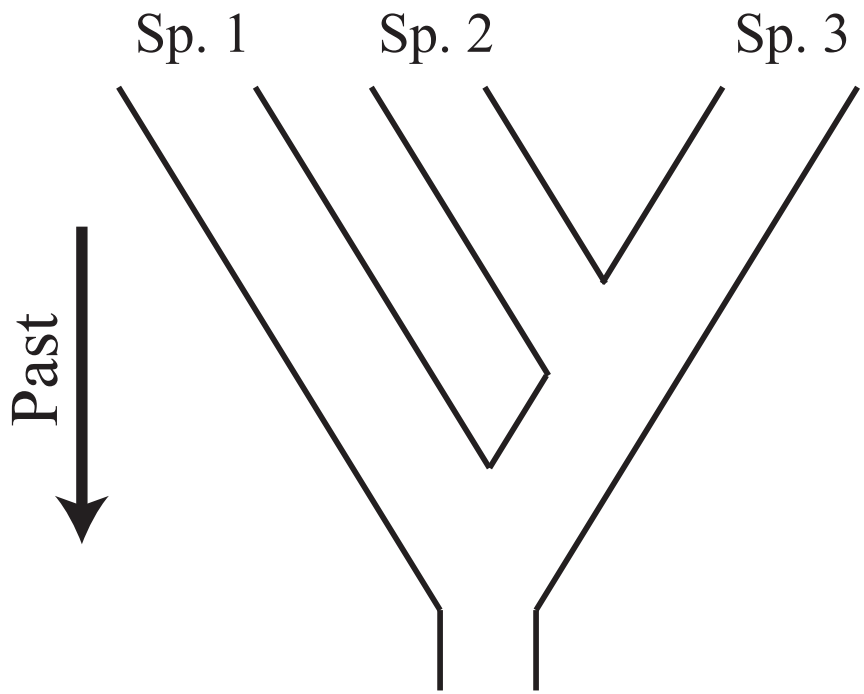


Species trees and gene trees



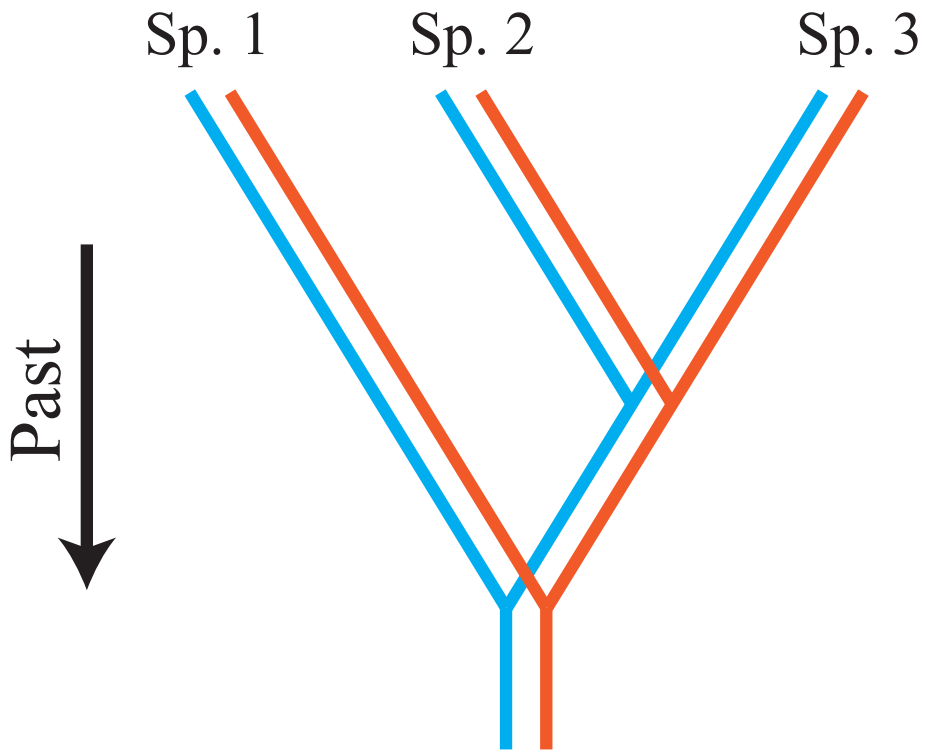
Species trees

- "Species" not necessarily a taxonomic rank.
- Species trees define barriers to gene flow.
- Each species is a single branch.



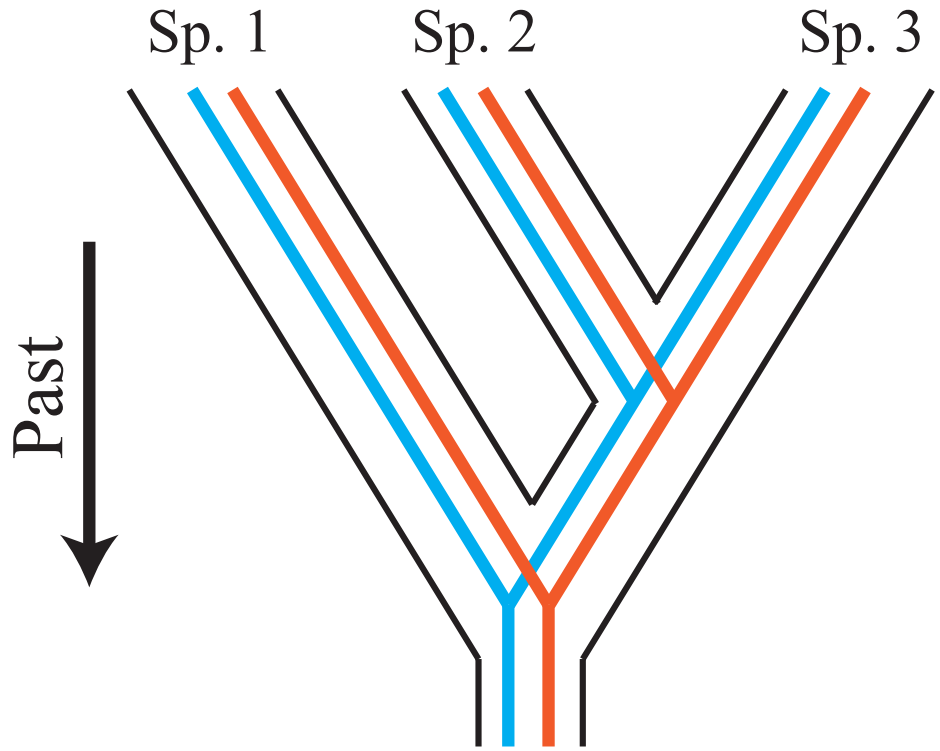
Gene trees

- Histories of loci within genomes of species.

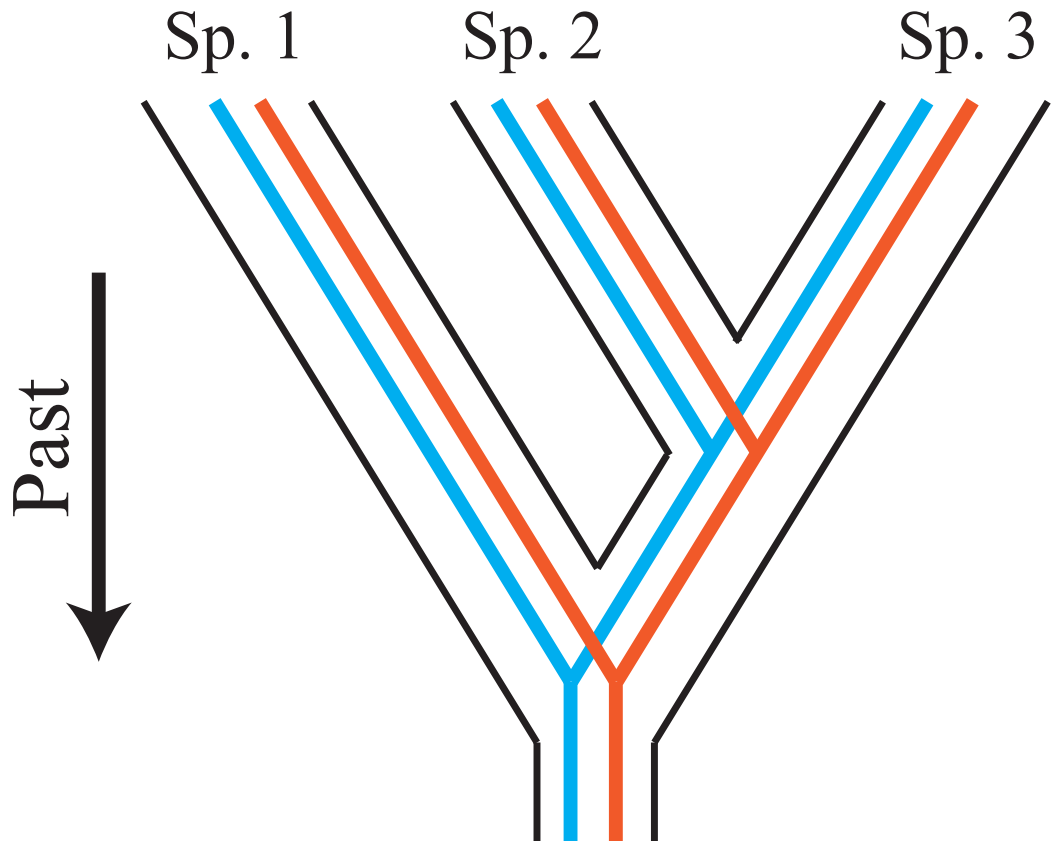


Gene trees

- Histories of loci within genomes of species.
- Embedded in species trees.

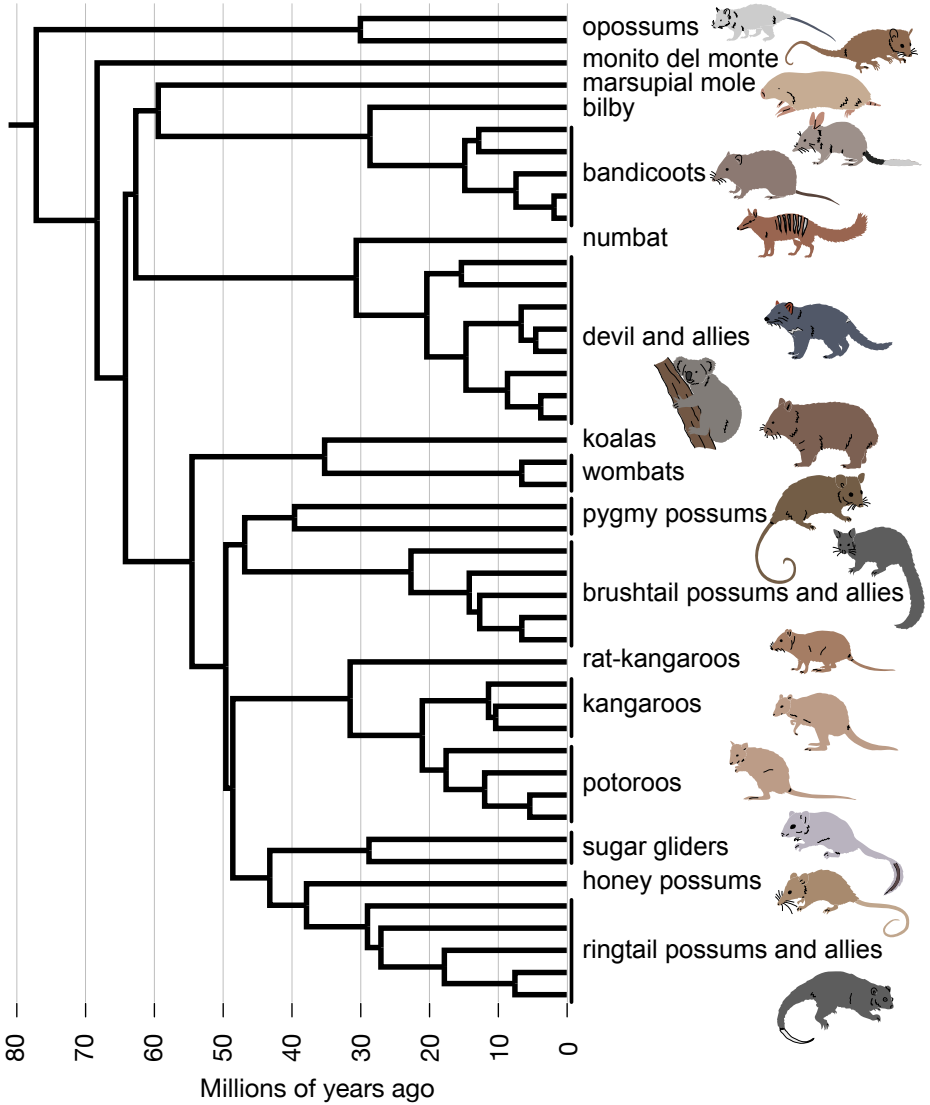


Gene tree **incongruence** is ubiquitous



Gene tree **incongruence** is ubiquitous

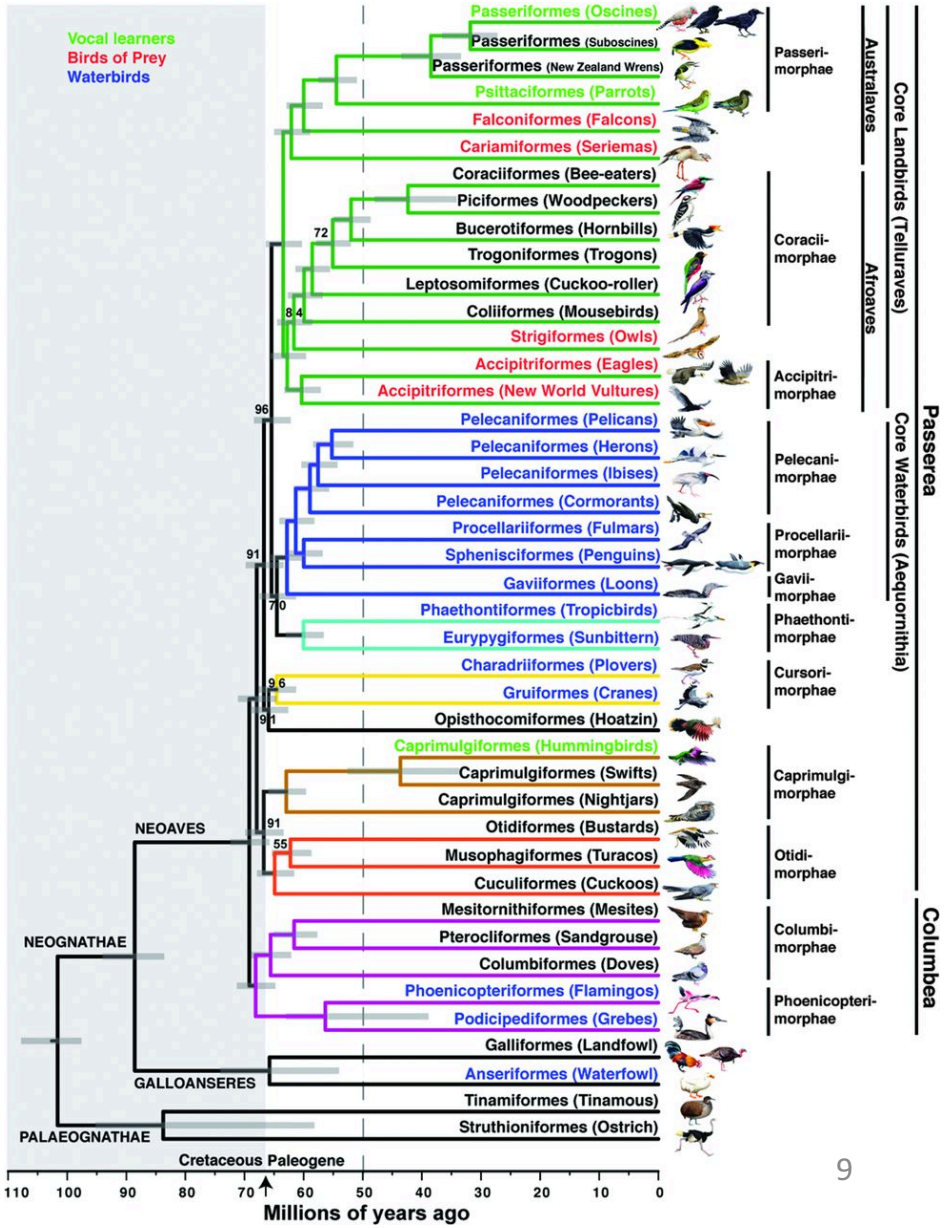
- Data set including 18 of 22 extant marsupial families.
- 1550 exon loci.
- Every locus leads to a different topology.



Duchene et al. 2018, *Syst. Biol.*

Gene tree **incongruence** is ubiquitous

- Data set including all avian orders.
- >10k loci including exons, introns, and UCEs.
- Every locus leads to a different topology to the inferred species trees.



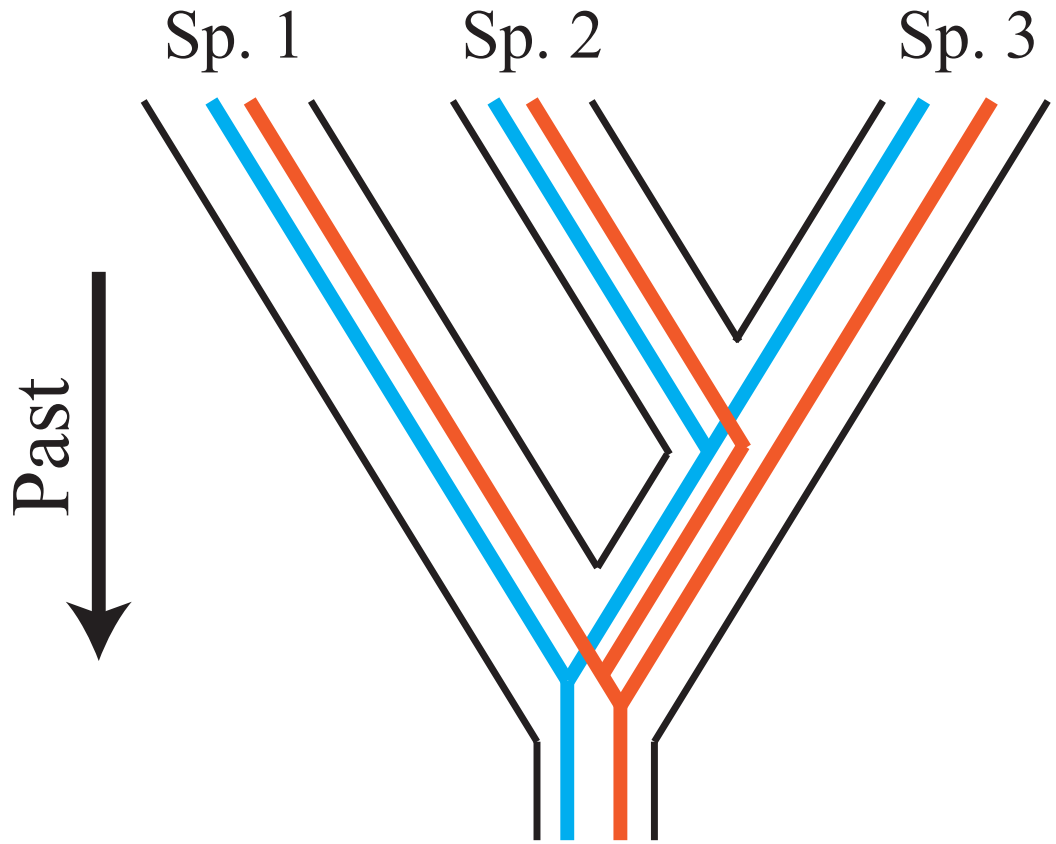
Jarvis et al. 2014, *Science*

Gene tree incongruence explained

- Deep coalescence (incomplete lineage sorting; ILS).
- Gene duplication or loss.
- Horizontal/lateral gene transfer (asexual) or gene flow/introgression (sexual).
- Hybridization.

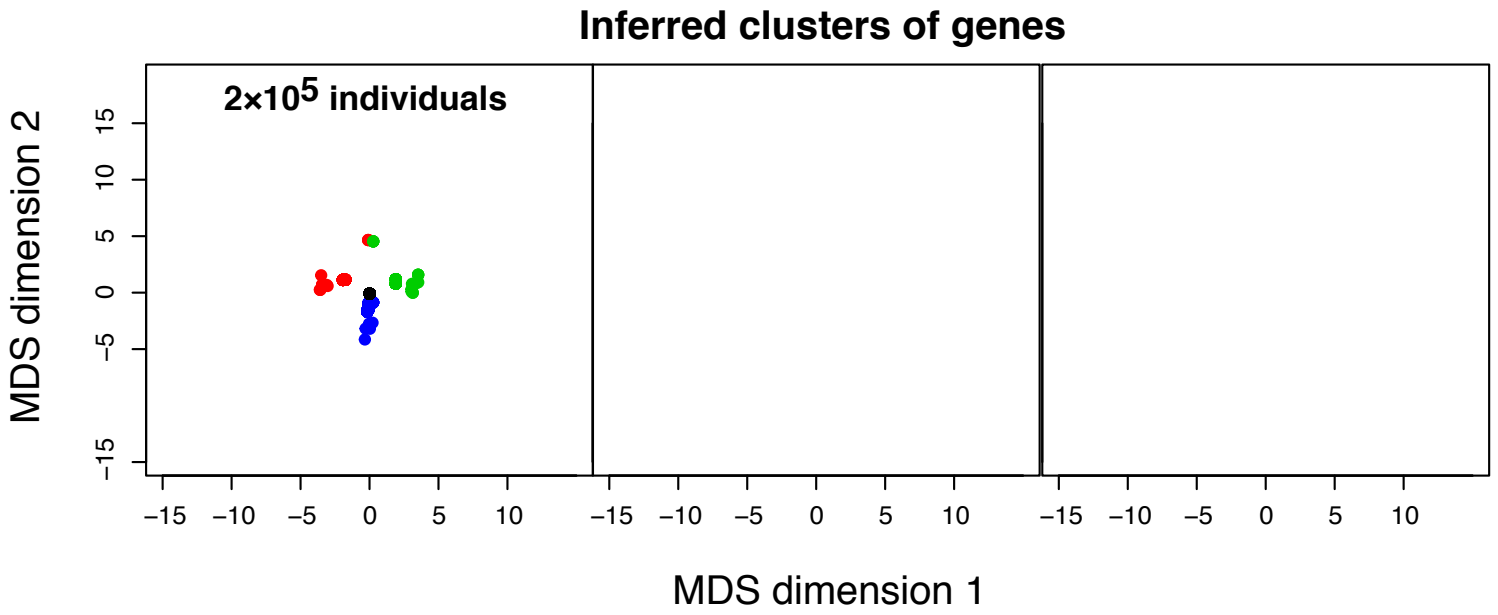
Gene tree incongruence explained

- Deep coalescence (incomplete lineage sorting; ILS)



Gene tree incongruence explained

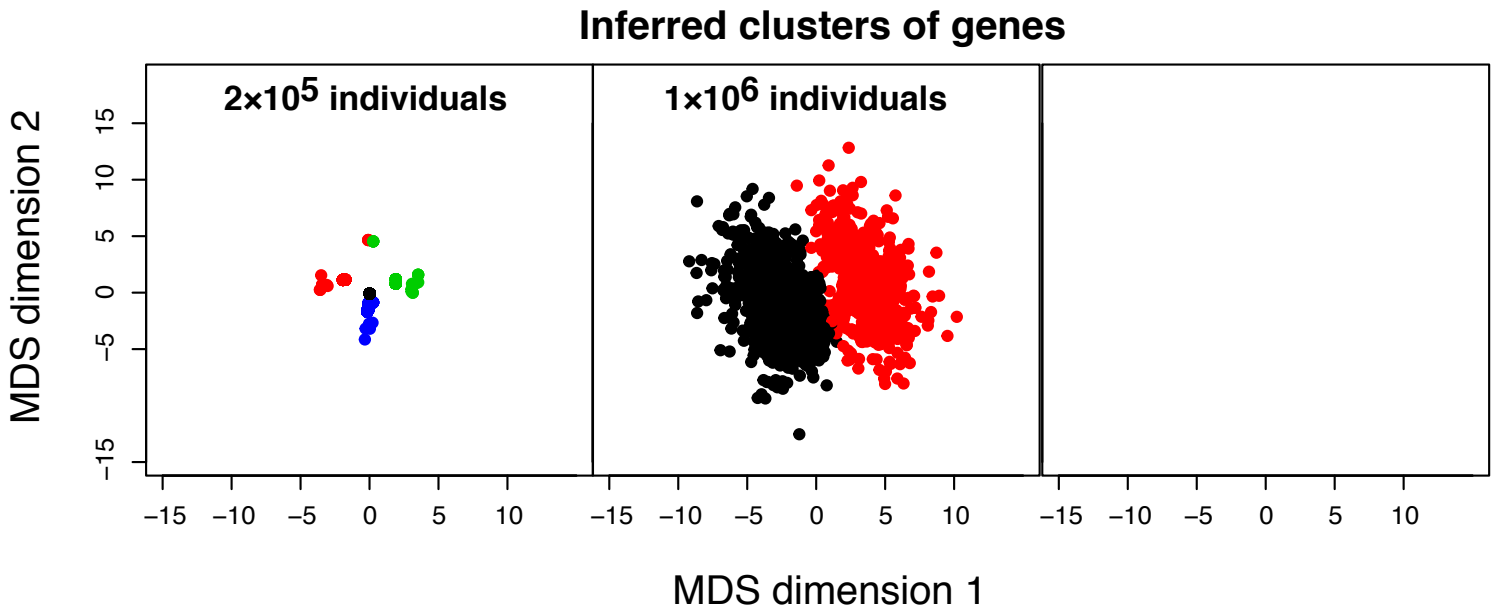
- Deep coalescence (ILS)



Duchene et al. 2018, *Syst. Biol.*

Gene tree incongruence explained

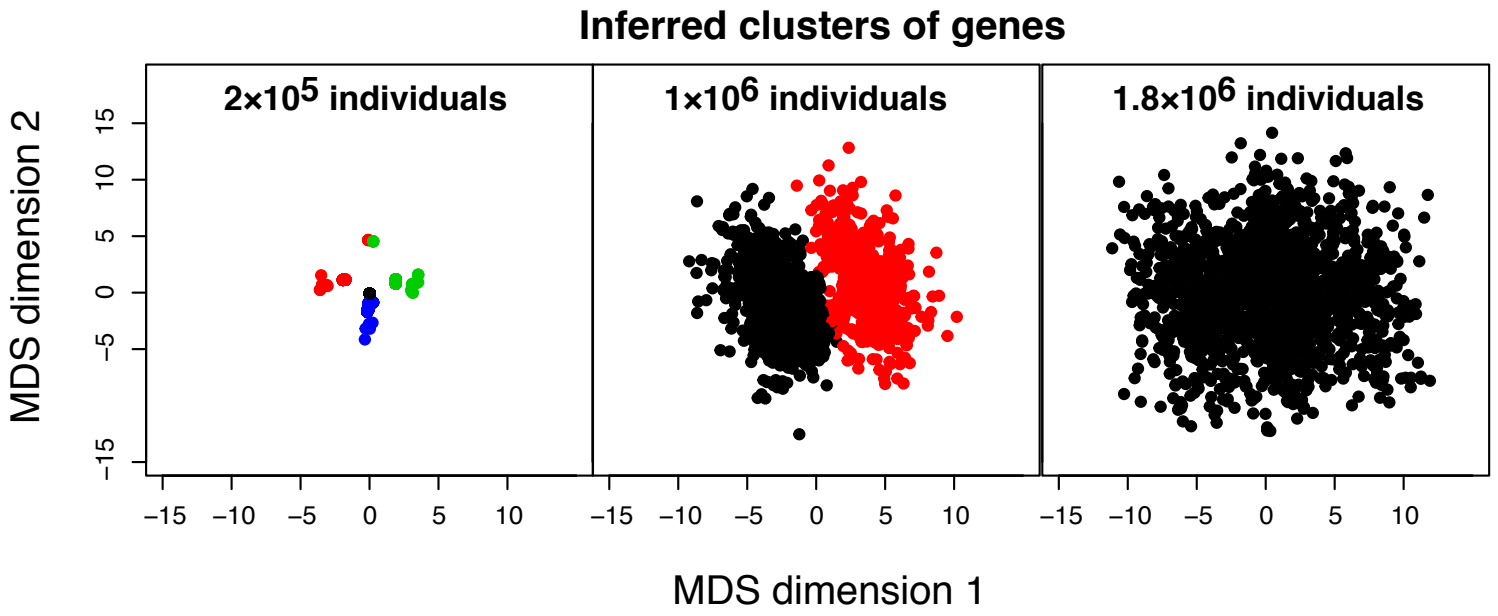
- Deep coalescence (ILS)



Duchene et al. 2018, *Syst. Biol.*

Gene tree incongruence explained

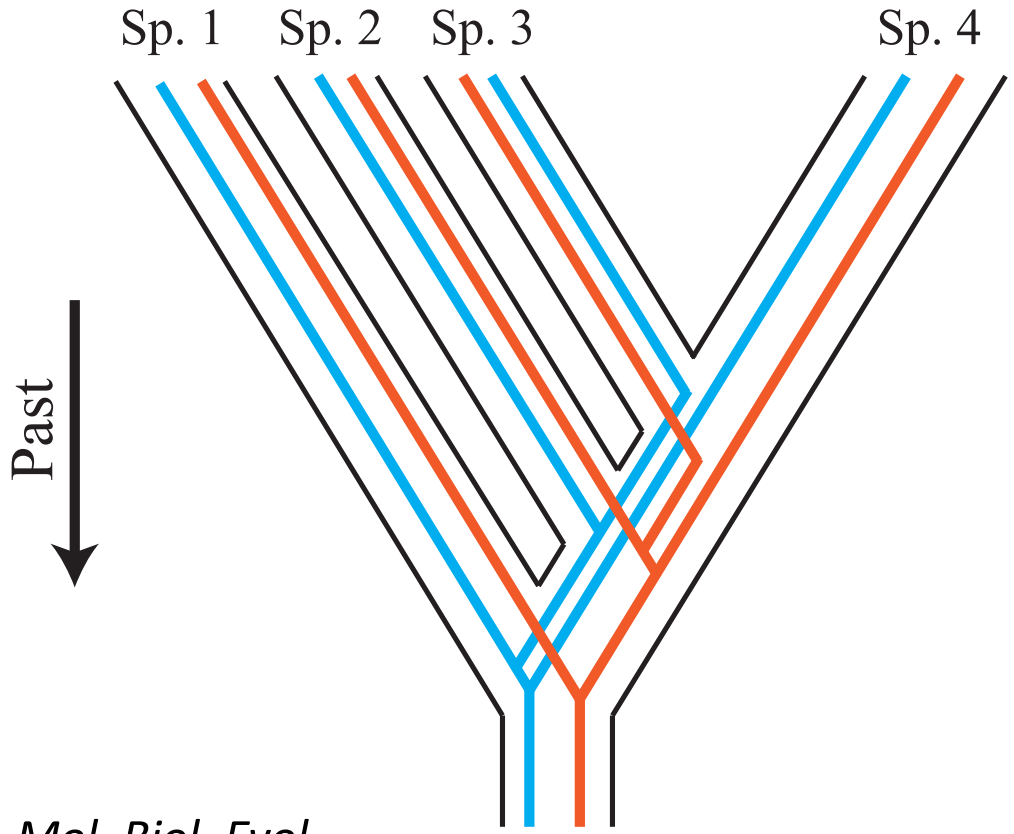
- Deep coalescence (ILS)



Duchene et al. 2018, *Syst. Biol.*

Anomalous gene trees under ILS

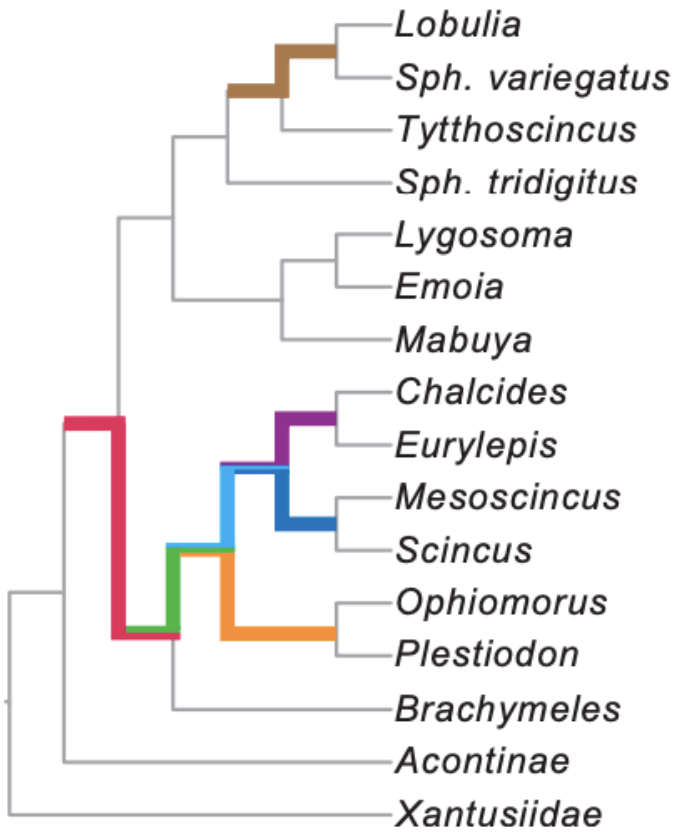
- Trees other than the species tree that are **more** probable.



Rosenberg 2013, *Mol. Biol. Evol.*

Anomalous gene trees under ILS

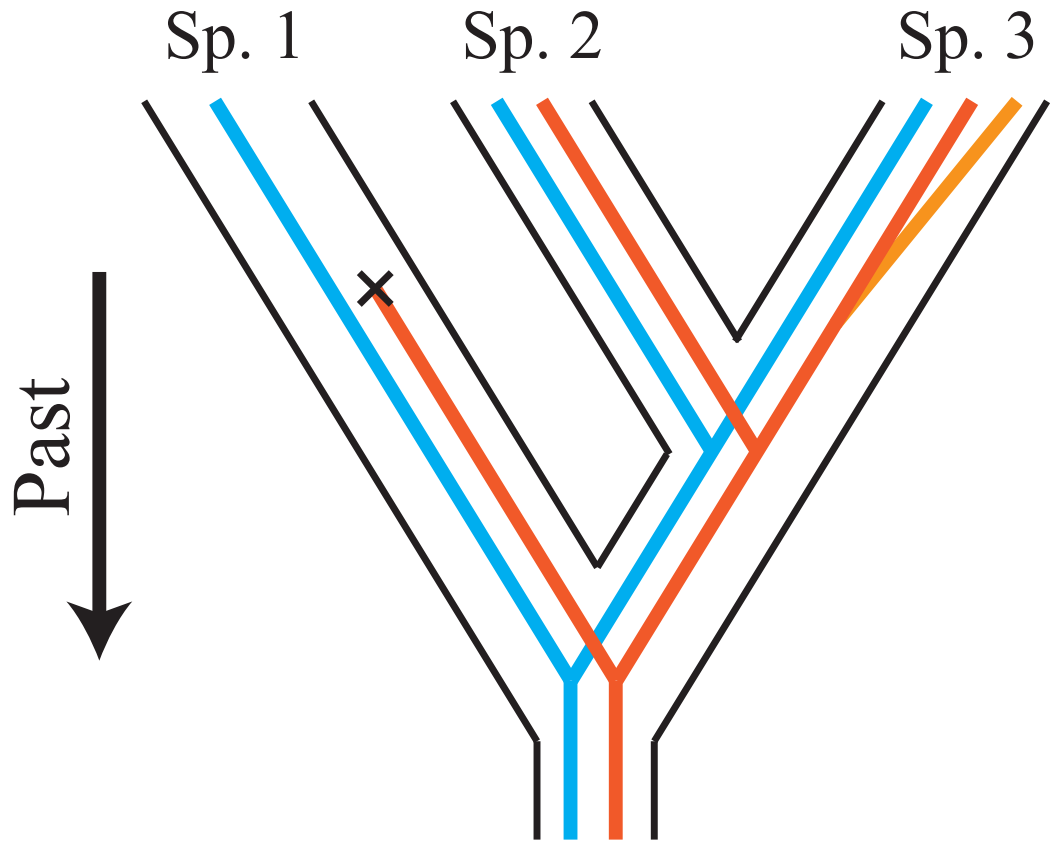
- Trees other than the species tree that are **more** probable.



Linkem et al. 2016, *Syst. Biol.*

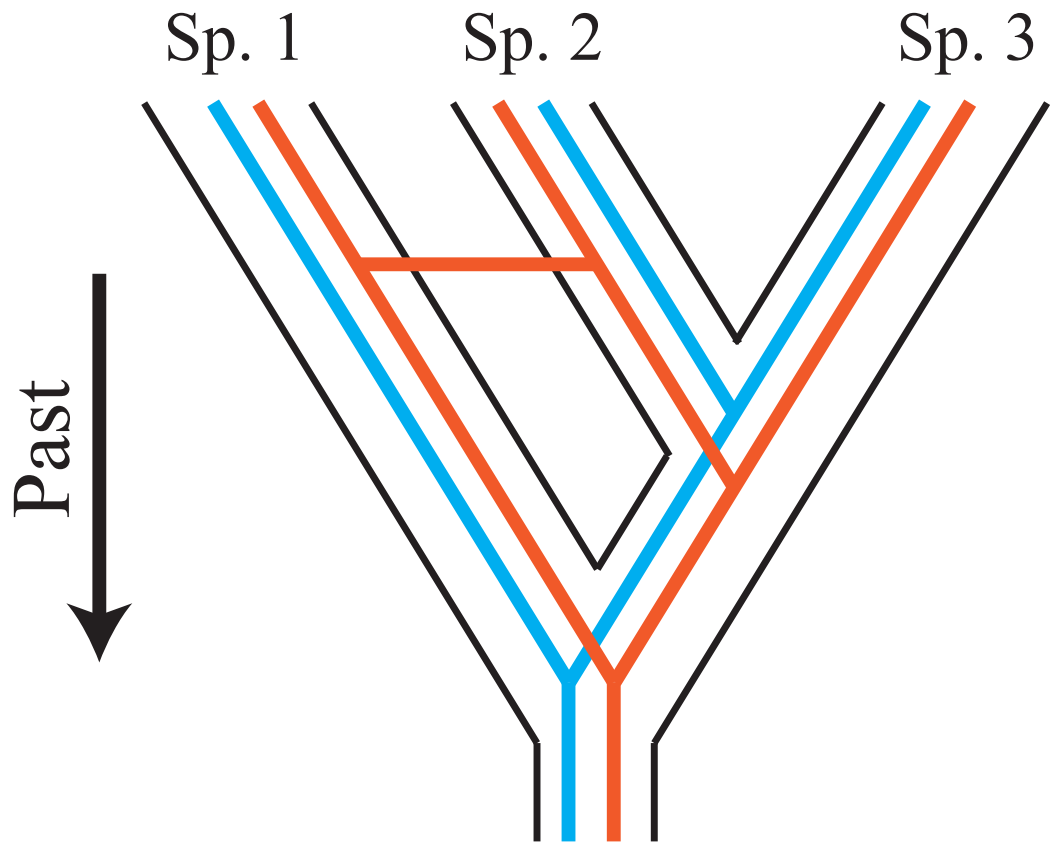
Gene tree incongruence explained

- Gene duplication/loss



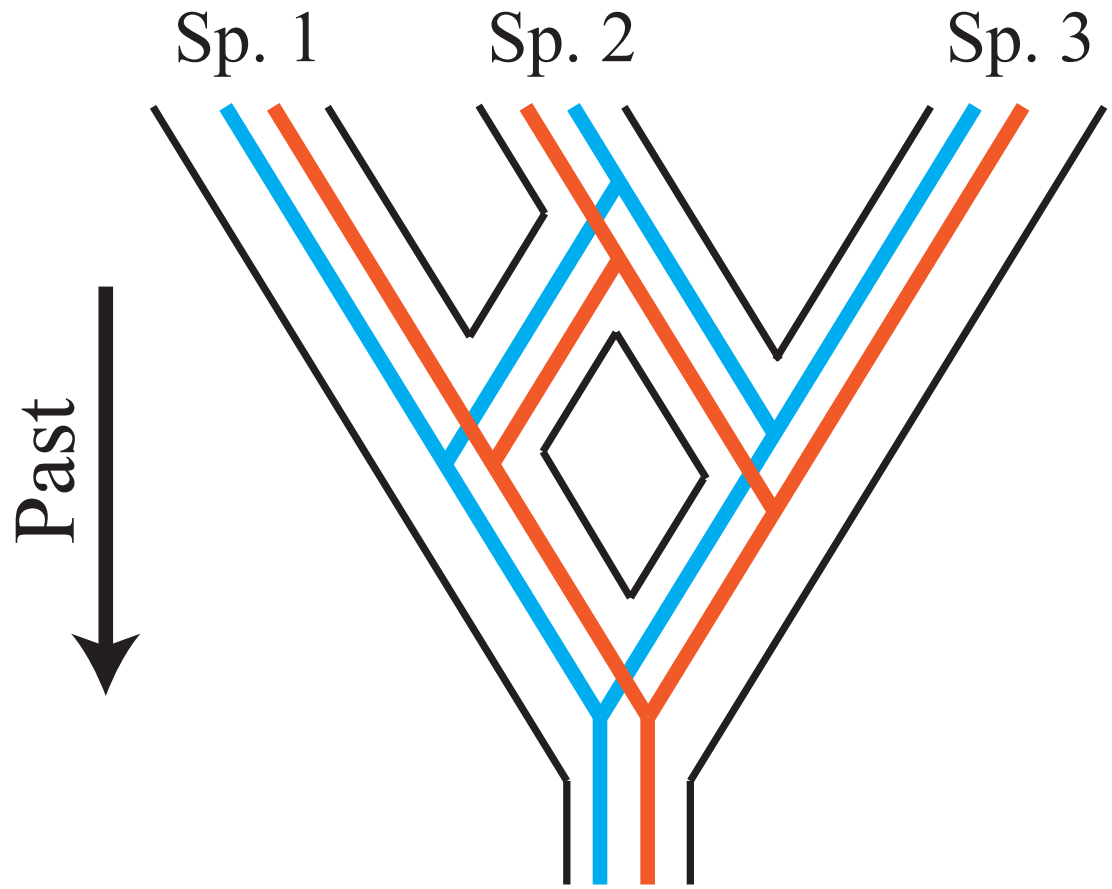
Gene tree incongruence explained

- Horizontal/lateral gene transfer (asexual) or gene flow/introgression (sexual)

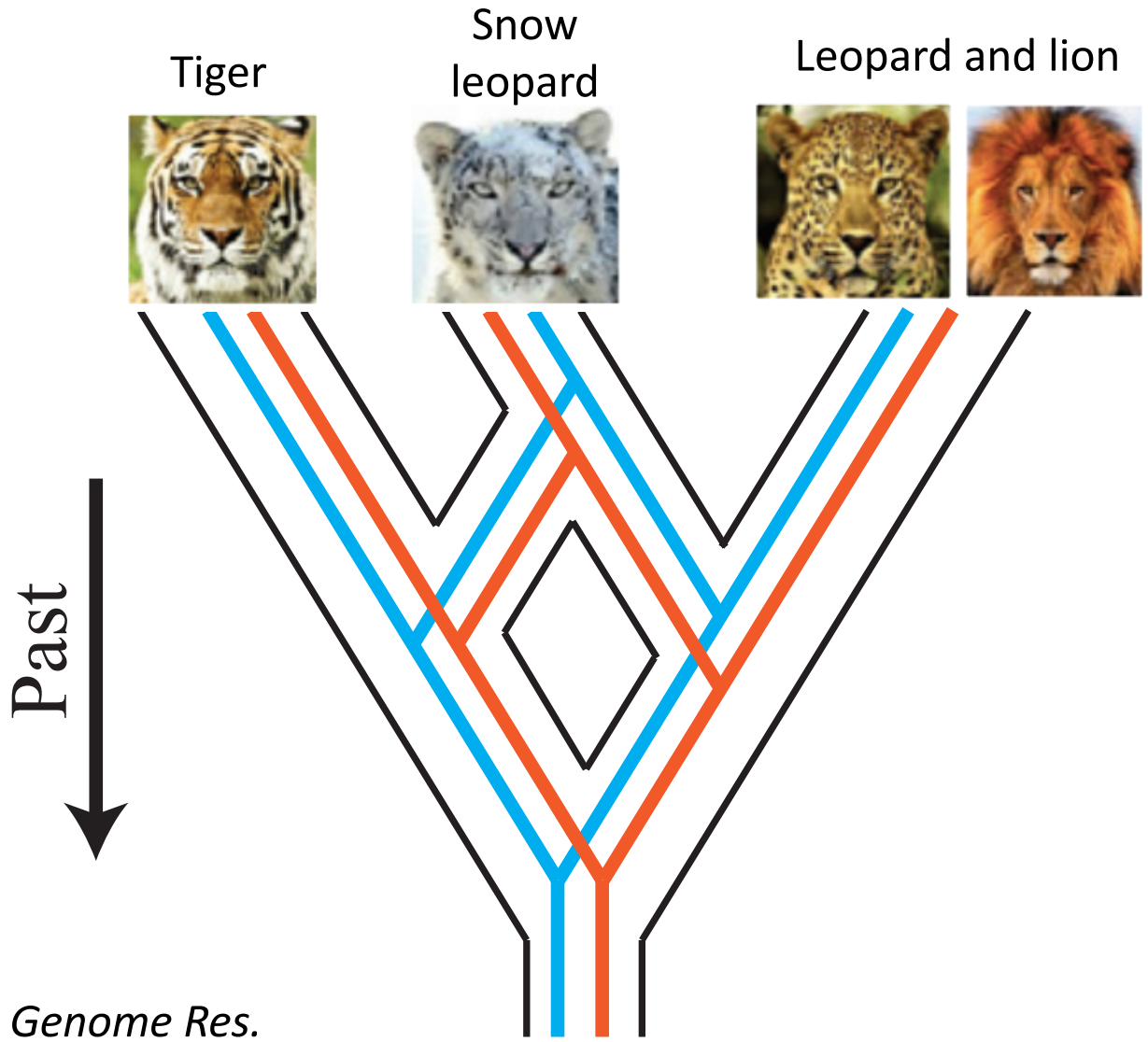


Gene tree incongruence explained

- Hybridizing speciation



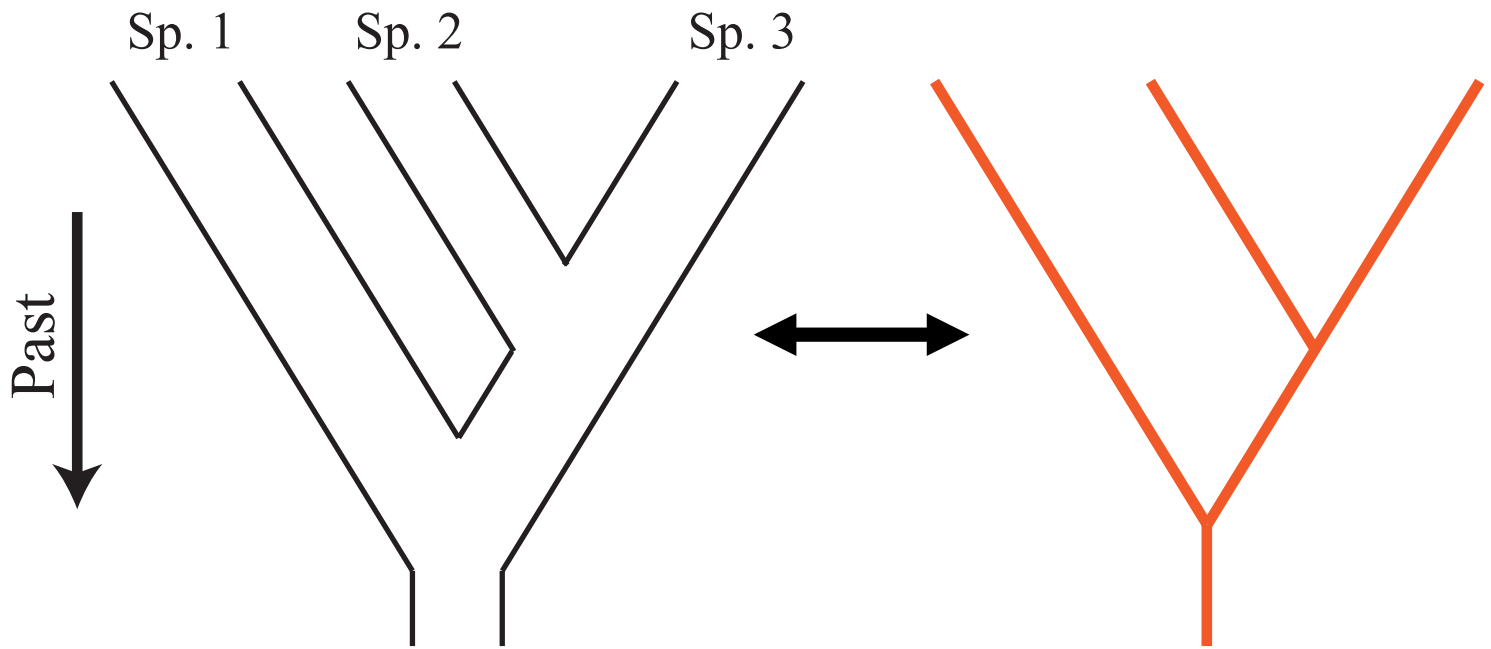
Gene tree incongruence explained



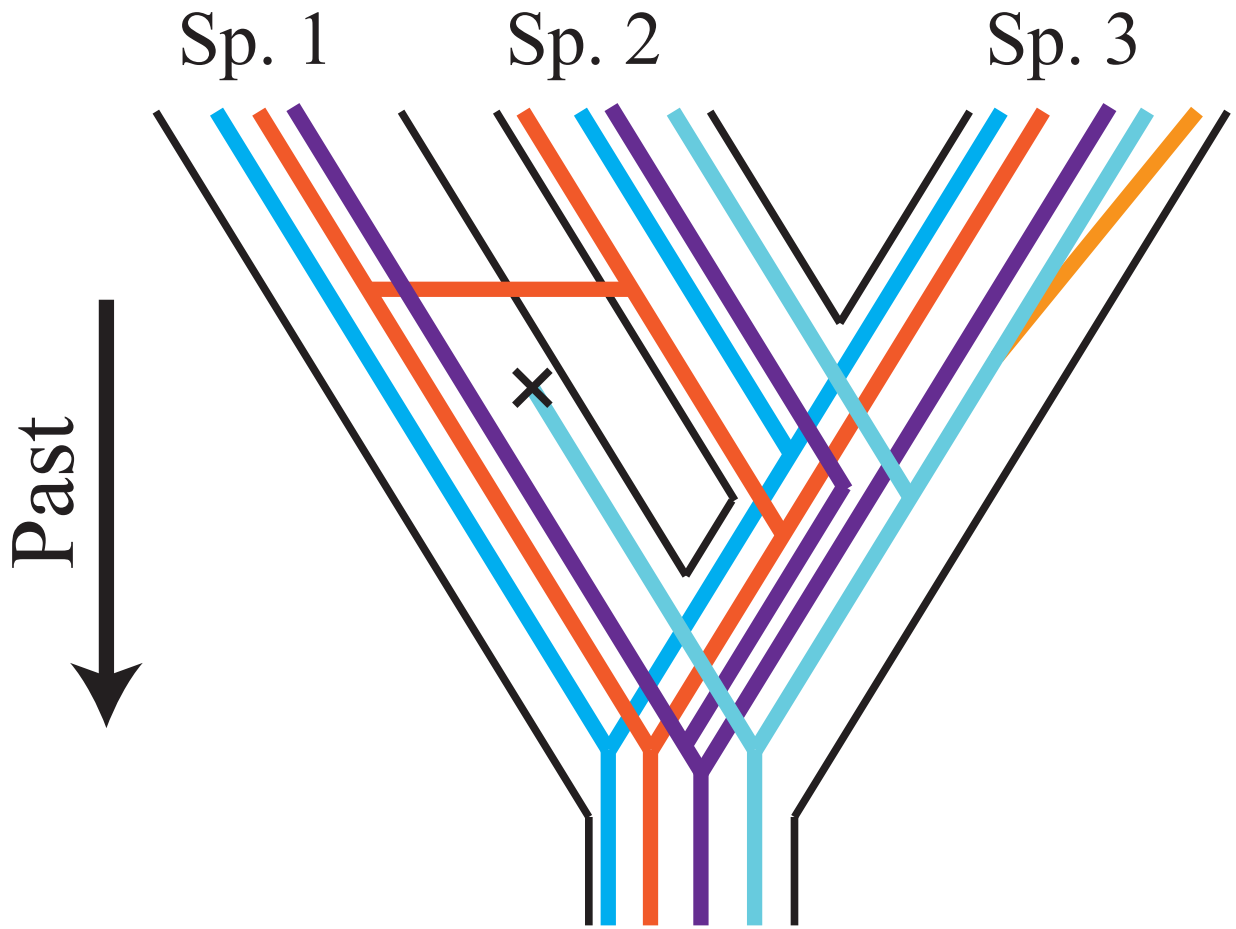
Li et al. 2016, *Genome Res.*

Species and gene trees inform each other

- Gene trees inform species tree inference.
- Species trees inform gene evolutionary processes.



Handling gene tree incongruence



Concatenation

- Combining of loci into a single data matrix for inferring the species tree.
- Assumes that gene tree incongruence is negligible or can be "averaged out".
- Statistically inconsistent in the presence of ILS.

Locus 1	Locus 2	Locus 3	Locus 4
AAACCCAACA	TGGGAACT	CCAAACCAA	GGGCCCAACC
AATCGGTTCA	TGCCTTCT	CGATTGGTC	GGCAGGTTCC
AATCGGAATT	TGCCAATA	CTCGGAAGG	GGCAGGAATA
AATCGGTTCA	TGCCTTCT	CTCGGTTTC	GGCAGGTTCC
AATCGGAATT	TGCCAATA	CTCTTTTGG	GGCAGGAATT

Concatenation

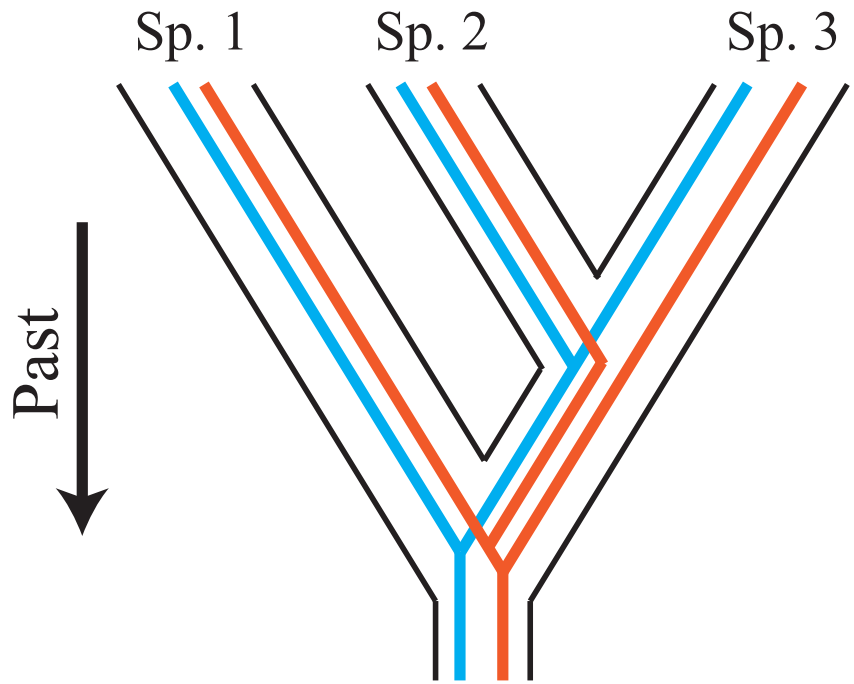
- Combining of loci into a single data matrix for inferring the species tree.
- Assumes that gene tree incongruence is negligible or can be "averaged out".
- Statistically inconsistent in the presence of ILS.

Concatenated data

```
AAACCCAACATGGGAACTCCAAACCAAGGGCCCAACC
AATCGGTTCATGCCTTCTCGATTGGTCGGCAGGTTCC
AATCGGAATTTGCCAATACTCGGAAGGGGCAGGAATA
AATCGGTTCATGCCTTCTCTCGGTTTCGGCAGGTTCC
AATCGGAATTTGCCAATACTCTTTTGGGGCAGGAATT
```

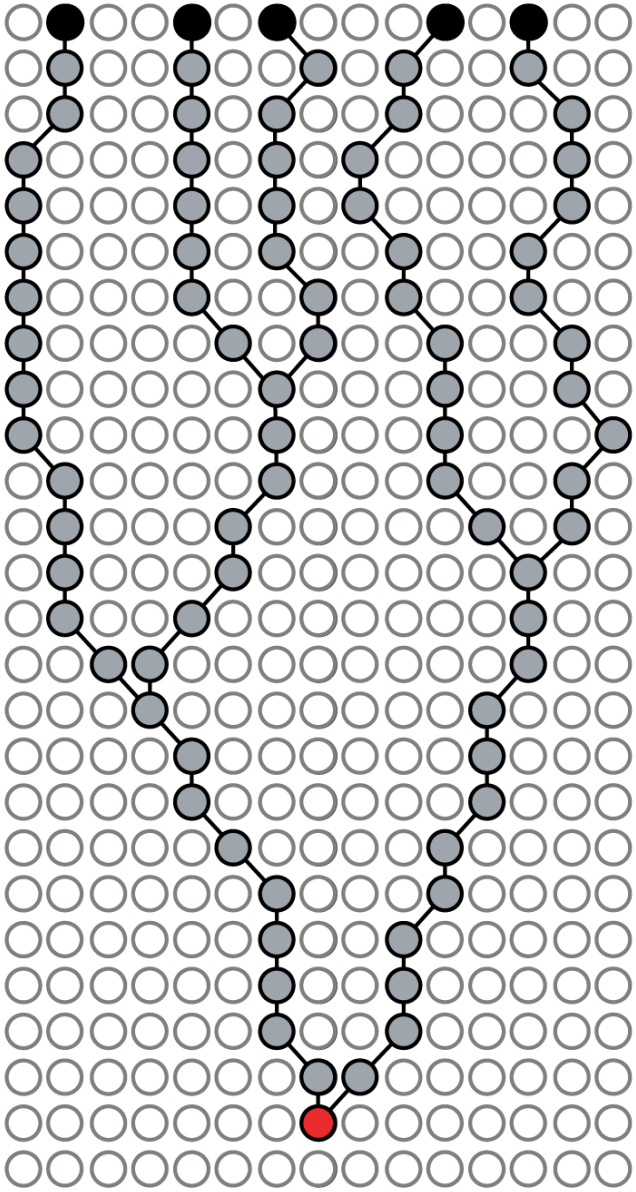
Accounting for incomplete lineage sorting using coalescent theory

- ILS is the most discussed form of incongruence.
- Substantial literature on theory and inference frameworks.



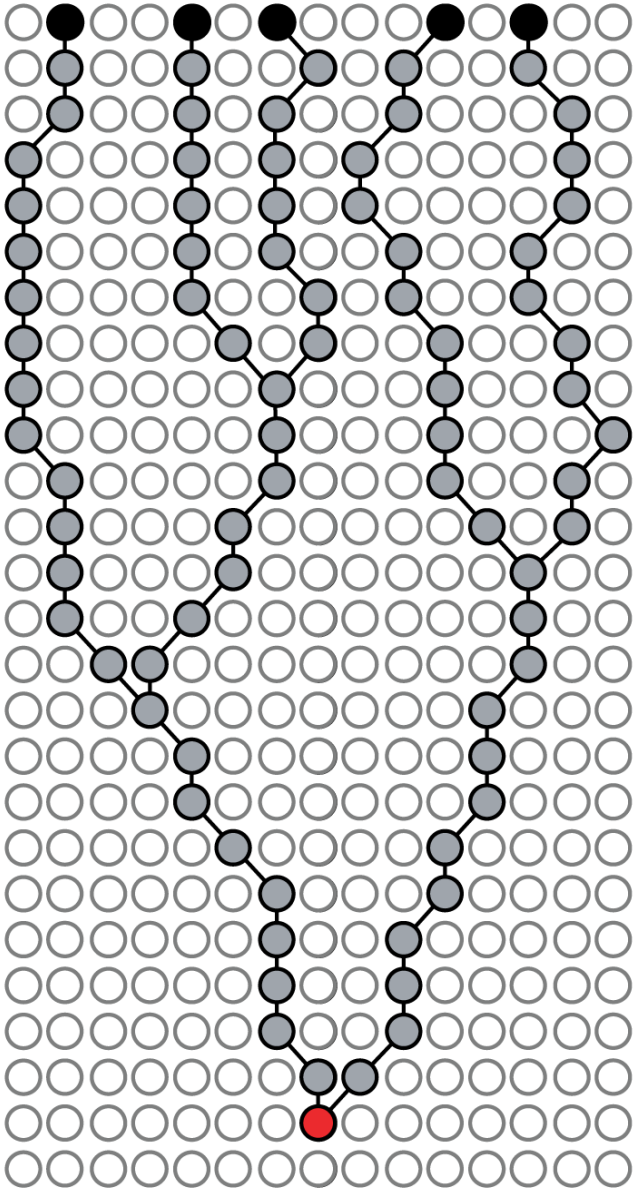
Coalescent theory

- Links population size with lineage history.



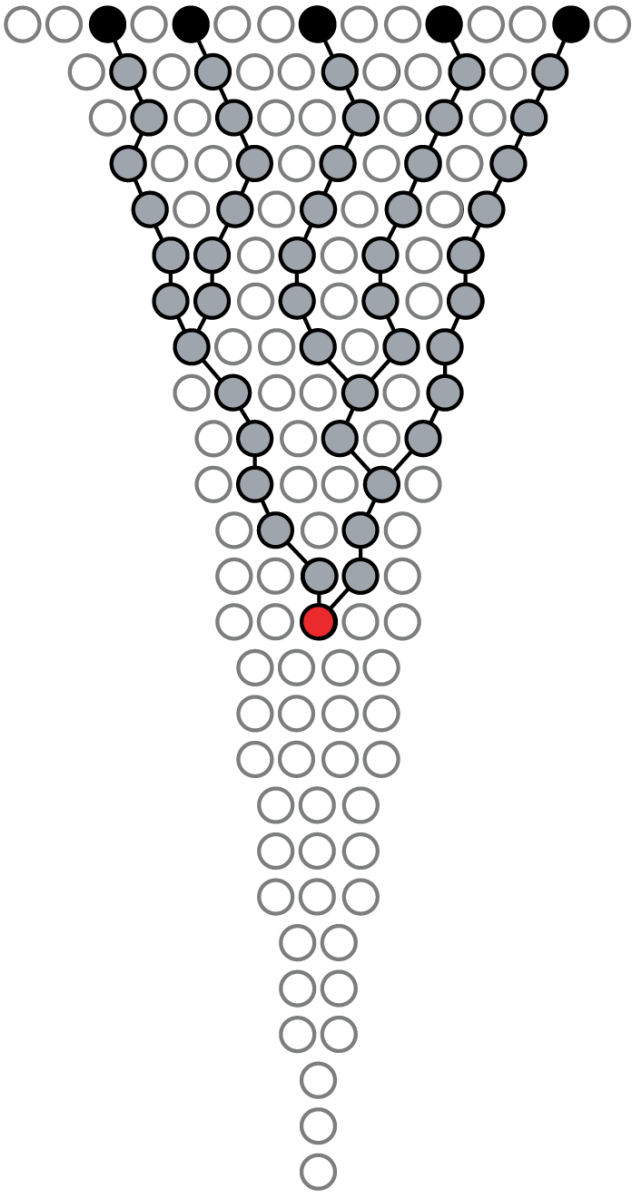
Coalescent theory

- Links population size with lineage history.
- Simple demographic models include:
  - Constant population.



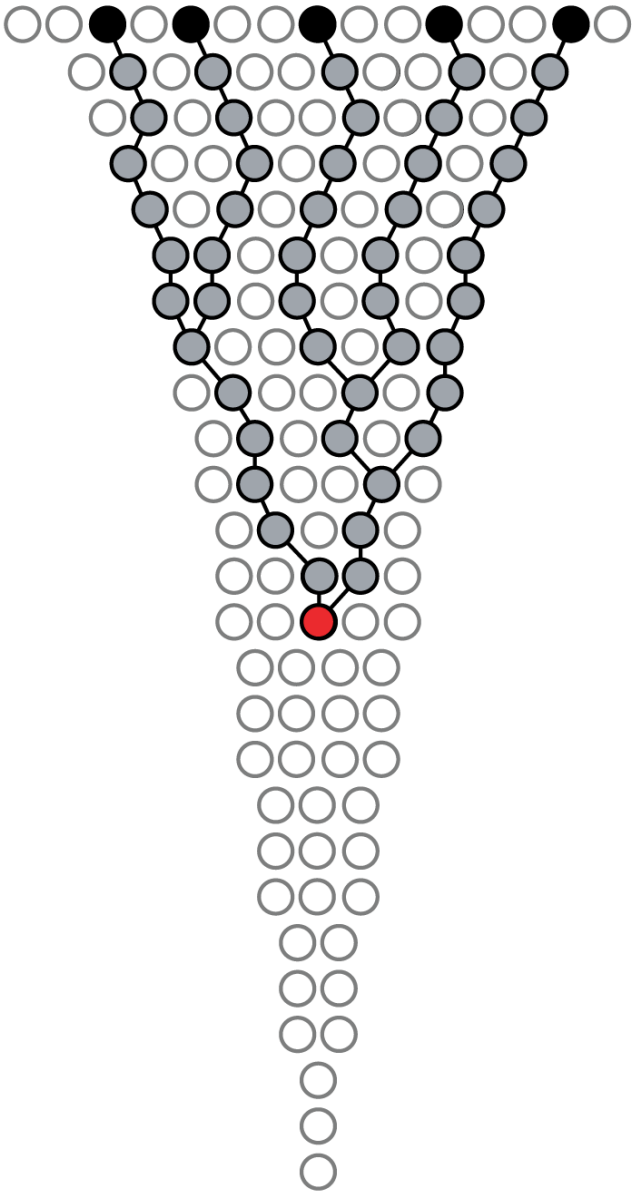
Coalescent theory

- Links population size with lineage history.
- Simple demographic models include:
  - Constant population.
  - Exponential growth.
  - Logistic growth.



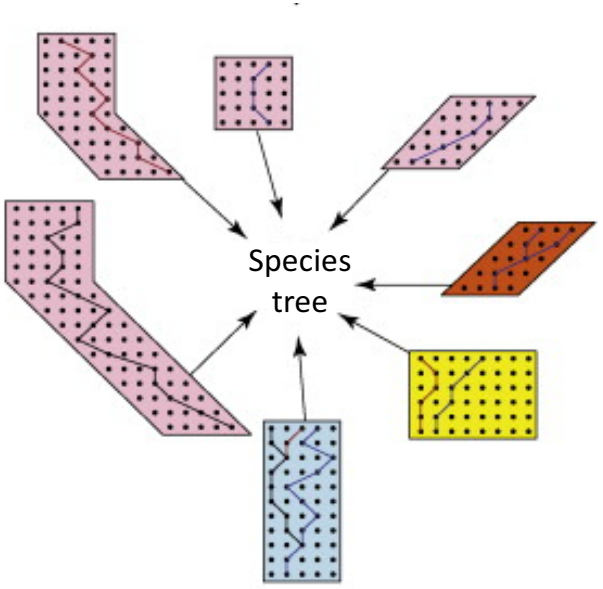
Coalescent theory

- Links population size with lineage history.
- Simple demographic models include:
  - Constant population.
  - Exponential growth.
  - Logistic growth.
- Assumes:
  - No within-gene recombination.
  - Non-overlapping generations.
  - Random mixing.



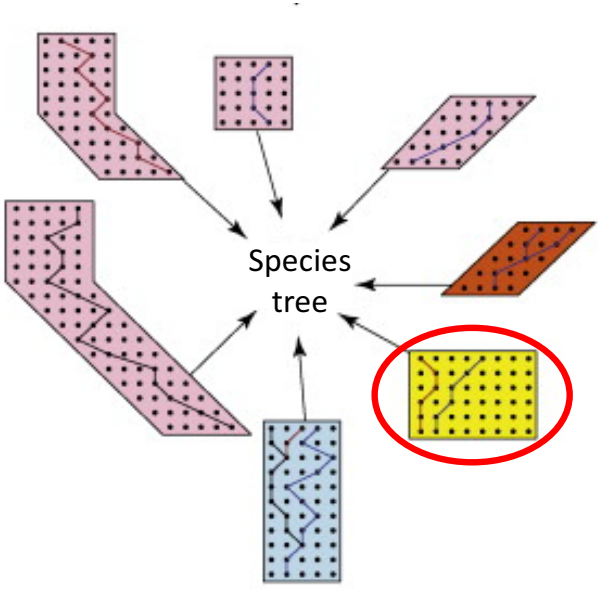
The multispecies coalescent (MSC)

- Taken to be composed by populations.



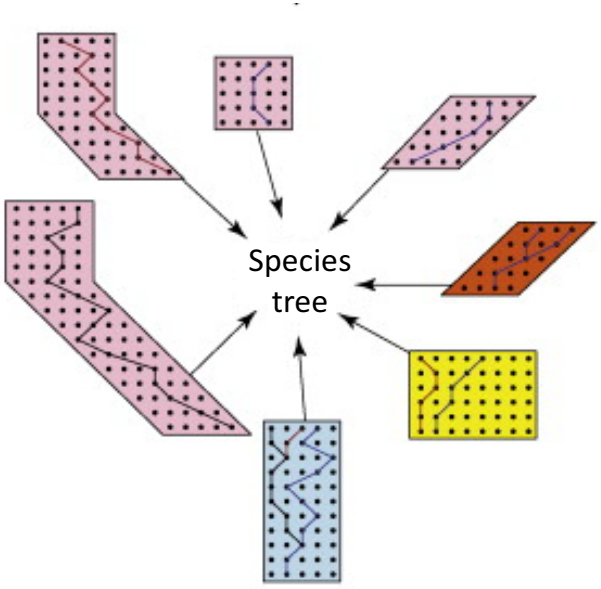
The multispecies coalescent (MSC)

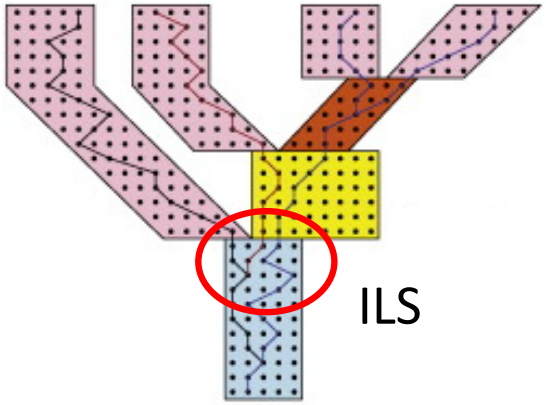
- Taken to be composed by populations.
- Lineages do not necessarily coalesce within a population.



The multispecies coalescent (MSC)

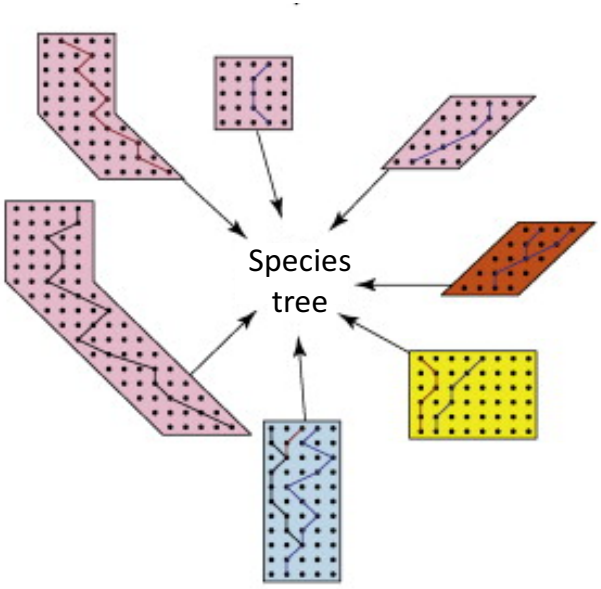
- Taken to be composed by populations.
- Lineages do not necessarily coalesce within a population.
- Allows for ILS given sufficient generations/*Ne*

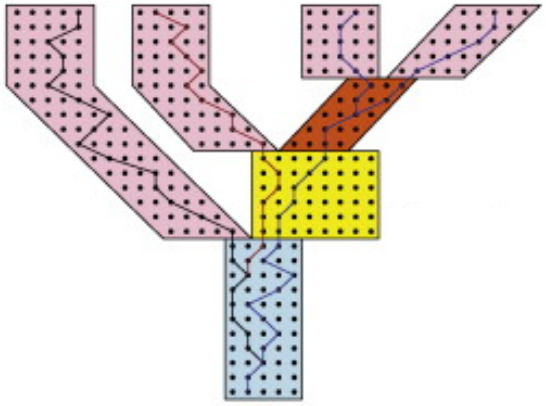


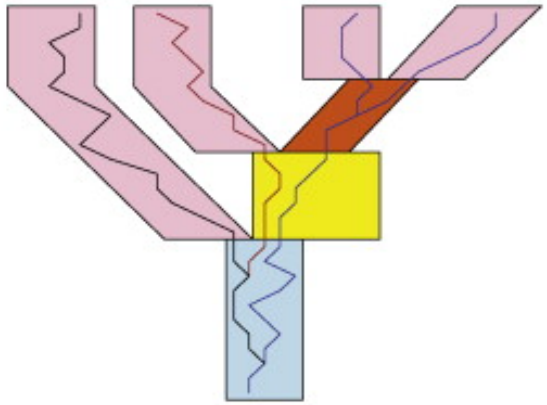


The multispecies coalescent (MSC)

- Taken to be composed by populations.
- Lineages do not necessarily coalesce within a population.
- Allows for ILS given sufficient generations/*Ne*







Multiple methods are consistent under the MSC

- Maximum pseudo-likelihood (e.g., MP-EST)
- Maximum quartet support (e.g., ASTRAL)
- Maximum likelihood (e.g., PhyloNet)
- Bayesian Inference (e.g., starBEAST2)

Bayesian inference naturally…

- Incorporates prior knowledge
- Allows the expression of realistic (complex) hierarchical models
- Co-estimates parameters
- Accommodates for uncertainty

and

- Has a vibrant community of developers and users

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G.$$

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G.$$

Felsenstein's likelihood of a gene tree given the alignment

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

Gene genealogies along the branches of the species tree

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

Population sizes along the branches of the species tree

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S)\, dG.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

$$P(S) = f_{BD}(S) P_N(S).$$

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S)\, dG.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

$$P(S) = f_{BD}(S) P_N(S).$$

Prior on species divergence times

starBEAST2

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S)\, dG.$$

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)).$$

$$P(S) = f_{BD}(S) P_N(S).$$

Prior on species population sizes

Mixing ILS with gene flow

- Supernetworks and combinatorics (Holland et al. 2008, *BMC Evol. Biol.*)
- Tests of gene flow using simulations (e.g., Joly et al. 2009, *Am. Nat.*).
- Parsimony (Stolzer et al. 2012, *Bioinf.*)
- Approximate Bayesian Computation (Woodhams et al. 2016, *Syst. Biol.*)
- Extending starBEAST2 to include gene flow (Müller et al. 2018, *bioRxiv*)
- Multispecies network coalescent (e.g., Wen et al. 2016, *PLOS Gen.*, Degnan 2018, *Syst. Biol.*)

Other considerations: Sampling of taxa, loci, and populations

- Sampling requirements (multiple loci and individuals per species).

Other considerations: Sampling of taxa, loci, and populations

- Sampling requirements (multiple loci and individuals per species).
- Great computational burden with hundreds of loci.

Other considerations: Sampling of taxa, loci, and populations

- Sampling requirements (multiple loci and individuals per species).
- Great computational burden with hundreds of loci.
- Number of loci and individuals sampled per species leads to markedly improved estimates.

Other considerations: Sampling of taxa, loci, and populations

- Sampling requirements (multiple loci and individuals per species).
- Great computational burden with hundreds of loci.
- Number of loci and individuals sampled per species leads to markedly improved estimates.
- More species lead to substitution detected BUT shorter times for coalescence.

Other considerations: The hierarchical model

- Complex models can lead to non-identifiability of parameters.
- Gene tree discordance compounded by model misspecification.

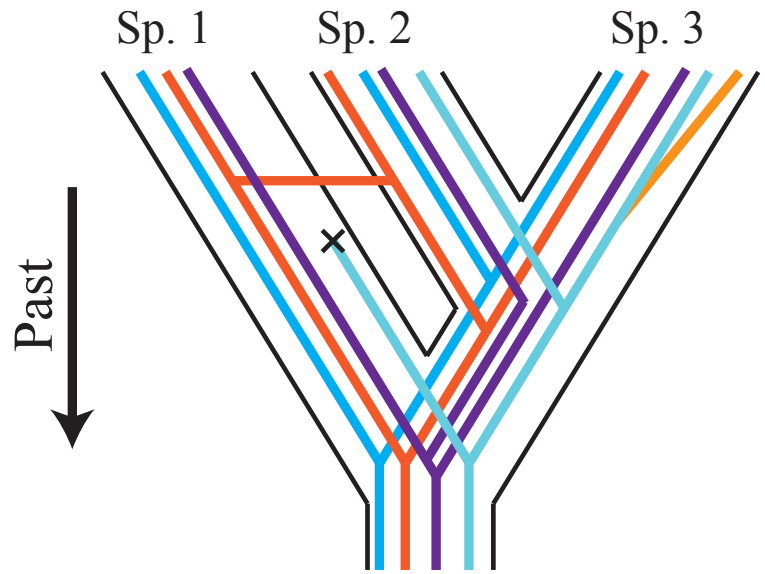
PhyloMAd

`github.com/duchene/phylomad`

Tree Model Adequacy

Model adequacy using tree summary statistics in BEAST2

Take home



- Multiple causes for gene tree discordance.
- The MSC accommodates ILS by linking population and speciation processes.
- Accounting for multiple sources of discordance is not straightforward.
- Multiple approaches available for inference (watch this space).

Further reading

- Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, *24*(6), 332-340.
- Heled, J., & Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular biology and evolution*, *27*(3), 570-580.
- Mallo, D., & Posada, D. (2016). Multilocus inference of species trees and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150335.
- Mueller, N. F., Ogilvie, H., Zhang, C., Drummond, A., & Stadler, T. (2018). Inference of species histories in the presence of gene flow. *bioRxiv*, 348391.
- Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Systematic biology*, *67*(5), 786-799.