Estimating species trees using fast methods

David L. Swofford
Florida Museum of Natural History
University of Florida

Collaborators:
Laura Kubatko (Ohio State University)
Julia Chifman (American University)
Colby Long (Ohio State University)

Why estimate population/species trees?

- May be interested in a phylogeographic question, and need a tree upon which to infer ancestral locations
- Hypothesis generation (e.g., isolation with migration models)
- Need a tree for ABBA/BABA statistics
- Interest in the phylogeny/systematics of a group of organisms

Invariants methods for phylogenetic inference

Phylogenetic invariants work by examining patterns in the data in order to determine whether they show relationships that are predicted by particular tree topologies.

Typically, invariants are linear or polynomial expressions that evaluate to specific expected values for each possible tree topology. By tabulating site pattern frequencies, we can calculate the values of the invariants and ask whether they approximate these expected values. This information allows selection of a preferred tree, and rejection of the other possibilities.

Early work on phylogenetic invariants

Journal of Classification 4:57-71 (1987)

Journal of Classification

Invariants of Phylogenies in a Simple Case with Discrete States

James A. Cavender
Martin Marietta Denver Aerospace

Joseph Felsenstein
University of Washington

Abstract: Under a simple model of transition between two states, we can work out the probabilities of different data outcomes in four species with any given phylogeny. For a given tree topology, if all characters are evolving under the same probabilistic model, there are two quadratic forms in the frequencies of outcomes that must be zero. It may be possible to test the null hypothesis that the tree is of a particular topology by testing whether these quadratic forms are zero. One of the tests is a test for independence in a simple 2 × 2 contingency table. If there are differences of evolutionary rate among characters, these quadratic forms will no longer necessarily be zero.

Keywords: Phylogenies; Statistical tests.

A Rate-independent Technique for Analysis of Nucleic Acid Sequences: Evolutionary Parsimony[1]

James A. Lake
Molecular Biology Institute and Department of Biology,
University of California, Los Angeles

The method of evolutionary parsimony—or operator invariants—is a technique of nucleic acid sequence analysis related to parsimony analysis and explicitly designed for determining evolutionary relationships among four distantly related taxa. The method is independent of substitution rates because it is derived from consideration of the group properties of substitution operators rather than from an analysis of the probabilities of substitution in branches of a tree. In both parsimony and evolutionary parsimony, three patterns of nucleotide substitution are associated one-to-one with the three topologically linked trees for four taxa. In evolutionary parsimony, the three quantities are operator invariants. These invariants are the remnants of substitutions that have occurred in the interior branch of the tree and are analogous to the substitutions assigned to the central branch by parsimony. The two invariants associated with the incorrect trees must equal zero (statistically), whereas only the correct tree can have a nonzero invariant. The χ^2-test is used to ascertain the nonzero invariant and the statistically favored tree. Examples, obtained using data calculated with evolutionary rates and branchings designed to camouflage the true tree, show that the method accurately predicts the tree, even when substitution rates differ greatly in neighboring peripheral branches (conditions under which parsimony will consistently fail). As the number of substitutions in peripheral branches becomes fewer, the parsimony and the evolutionary-parsimony solutions converge. The method is robust and easy to use.

Cavender and Felsenstein (1987)

Quadratic invariants for 2-state data

Lake (1987)

Linear invariants for 4-state nucleotide data

Much subsequent work by mathematicians

The Annals of Statistics
1993, Vol. 21, No. 1, 355–377

INVARIANTS OF SOME PROBABILITY MODELS USED

Comparative Genomics via Phylogenetic Invaria

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Mathematical Biosciences

Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees

M. Casanellas and J. Fernández-Sánchez
Department of Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain

An attempt to use phylogenetic invariants for tree reconstruction was made at the end of the 80s and the beginning of the 90s by several researchers (the initial idea due to Lake [1987] and Cavender and Felsenstein [1987]). However, the efficiency of methods based on invariants is still in doubt (Huelsenbeck 1995; Jin and Nei 1990). Probably because these methods only used few generators of the set of phylogenetic invariants. The method studied in this paper was first introduced

Toric Ideals of Phylogenetic Invariants

Bernd Sturmfels and Seth Sullivant
Department of Mathematics, University of California, Berkeley

Abstract

Statistical models of evolution are algebraic varieties in the space of joint probability distributions on the leaf colorations of a phylogenetic tree. The phylogenetic invariants of a model are the polynomials which vanish on the variety. Several widely used models for biological sequences have transition matrices that can be diagonalized by means of the Fourier transform of an abelian group. Their phylogenetic invariants form a toric ideal in the Fourier coordinates. We determine minimal generators and Gröbner bases for these toric ideals. For the Jukes-Cantor and Kimura models on a binary tree, our Gröbner basis consists of quadrics, cubics and quartics.

... and many more

Much subsequent work by mathematicians

$$\mathbb{E}\langle Y_1, \theta\rangle\mathbb{E}\langle Y_2, \phi\rangle = \mathbb{E}\langle Y_1, \phi\rangle\mathbb{E}\langle Y_2, \theta\rangle.$$

Before closing this discussion of the invariants of the Kimura two-parameter model for the two-leaf tree, it is both of independent interest and convenient for later examples to relate our notation and results to those of Cavender (1989, 1991). Following Cavender, we let A (resp. G, C, T) double as the *function* on $\mathbb{G} = \{A, G, C, T\}$ which takes the value 1 on A (resp., G, C, T) and 0 elsewhere. We then see that $A - G = (1/2)(\phi + \theta)$ and $C - T = (1/2)(\phi - \theta)$; and, letting $\otimes$ denote the tensor product of functions on $\mathbb{G}$, we see that

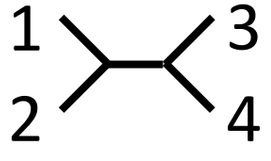
$$\begin{aligned}(A - G) \otimes (C - T) &= \tfrac{1}{4}(\phi + \theta) \otimes (\phi - \theta)\\ &= \tfrac{1}{4}(\phi \otimes \phi - \phi \otimes \theta + \theta \otimes \phi - \theta \otimes \theta)\end{aligned}$$

and

$$\begin{aligned}(C - T) \otimes (A - G) &= \tfrac{1}{4}(\phi - \theta) \otimes (\phi + \theta)\\ &= \tfrac{1}{4}(\phi \otimes \phi - \theta \otimes \phi + \phi \otimes \theta - \theta \otimes \theta).\end{aligned}$$

Despite providing a great deal of recreational opportunities for mathematicians/algebraic statisticians, these methods were almost completely ignored by empirical evolutionary biologists!

Lake's linear invariants ("Evolutionary Parsimony")

For each set of 4 sequences (quartet): 

For the tree ((1,2),(3,4)), tabulate

$$E = f_{AA|CC} + f_{AA|TT} + f_{CC|AA} + f_{CC|GG} + f_{GG|CC} + f_{GG|TT} + f_{TT|AA} + f_{TT|GG}$$

$$u = f_{AG|CT} + f_{AG|TC} + f_{CT|AG} + f_{CT|GA} + f_{GA|CT} + f_{GA|TC} + f_{TC|AG} + f_{TC|GA}$$

$$H = f_{AG|CC} + f_{AG|TT} + f_{CT|AA} + f_{CT|GG} + f_{GA|CC} + f_{GA|TT} + f_{TC|AA} + f_{TC|GG}$$

$$J = f_{AA|CT} + f_{AA|TC} + f_{CC|AG} + f_{CC|GA} + f_{GG|CT} + f_{GG|TC} + f_{TT|AG} + f_{TT|GA}$$

Then:

$$X = E + u - H - J$$

Calculate similar terms Y and Z for the trees ((1,3),(2,4)) and ((1,4),(2,3)).

$$Y = F + v - L - N \qquad Z = G + w - Q - S$$

(patterns not involving 2 pyrimidines and 2 purines are ignored)

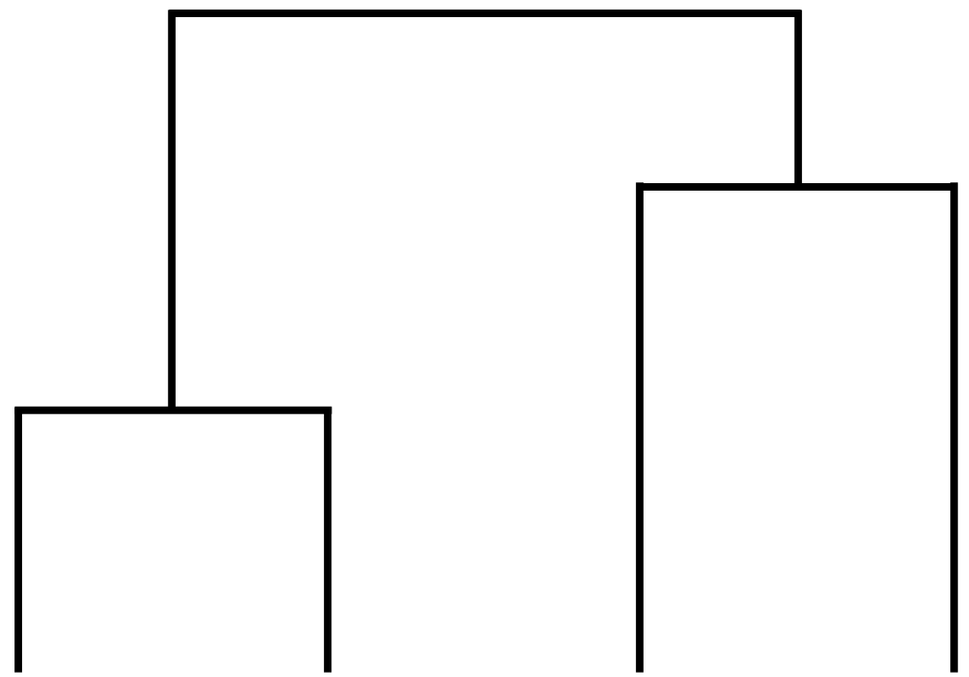
The true tree is expected to take a positive value for one of *X*, *Y*, and *Z*; the other two are expected to be zero (i.e., the invariants).

Inferring evolutionary trees using matrix rank and the Singular Value Decomposition(SVD)

- Allman and Rhodes (2003, 2004)
- Eriksson(2005)
- Chifman and Kubatko (2014, 2015)
- Fernández-Sánchez and Casanellas (2016)

Exploit the fact that there are linear dependencies in site-pattern frequencies that are tree-topology specific (the “invariants”)

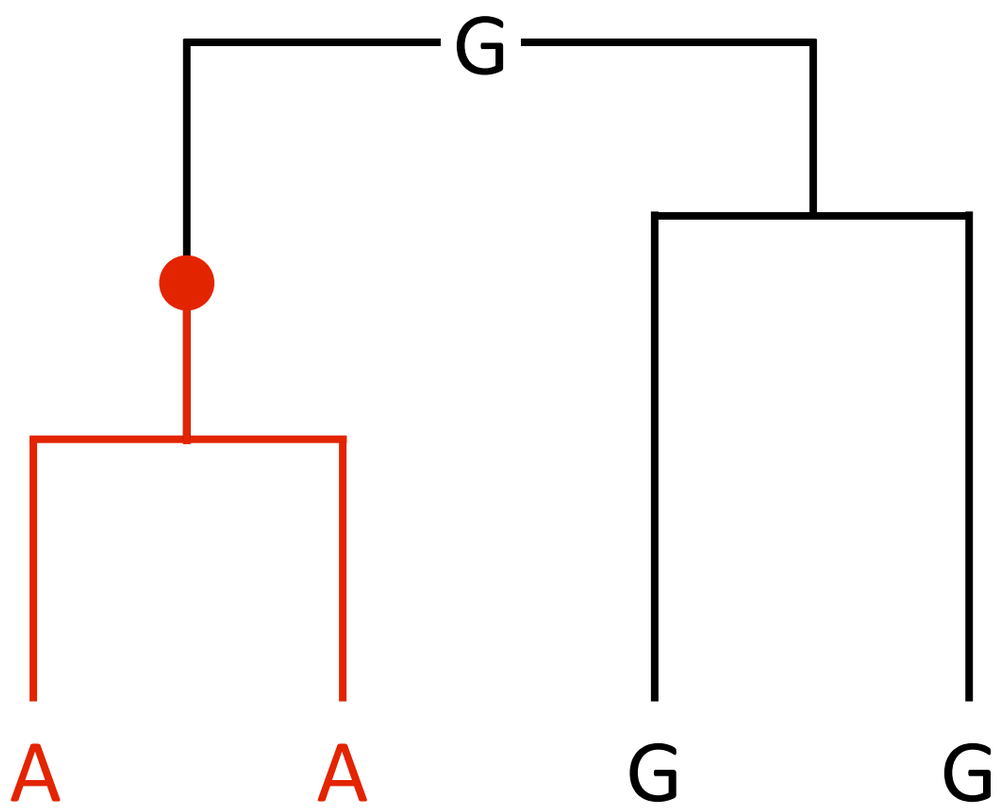
Site pattern frequencies



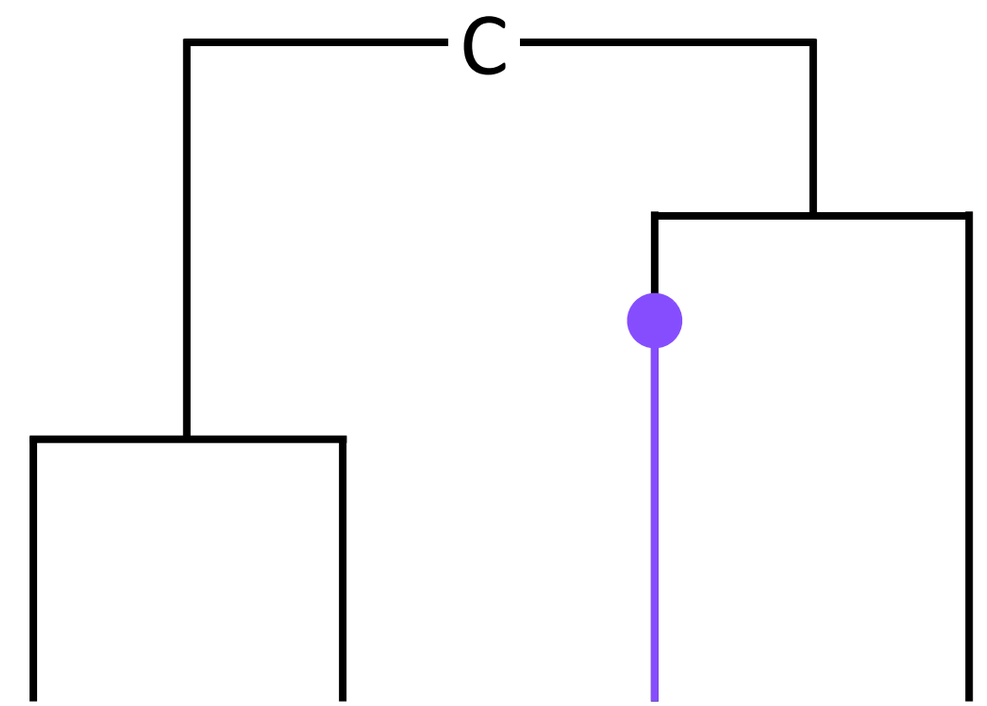
1 3 5 6

A tree for 4 taxa, which may be a subtree of a larger tree

Site pattern frequencies

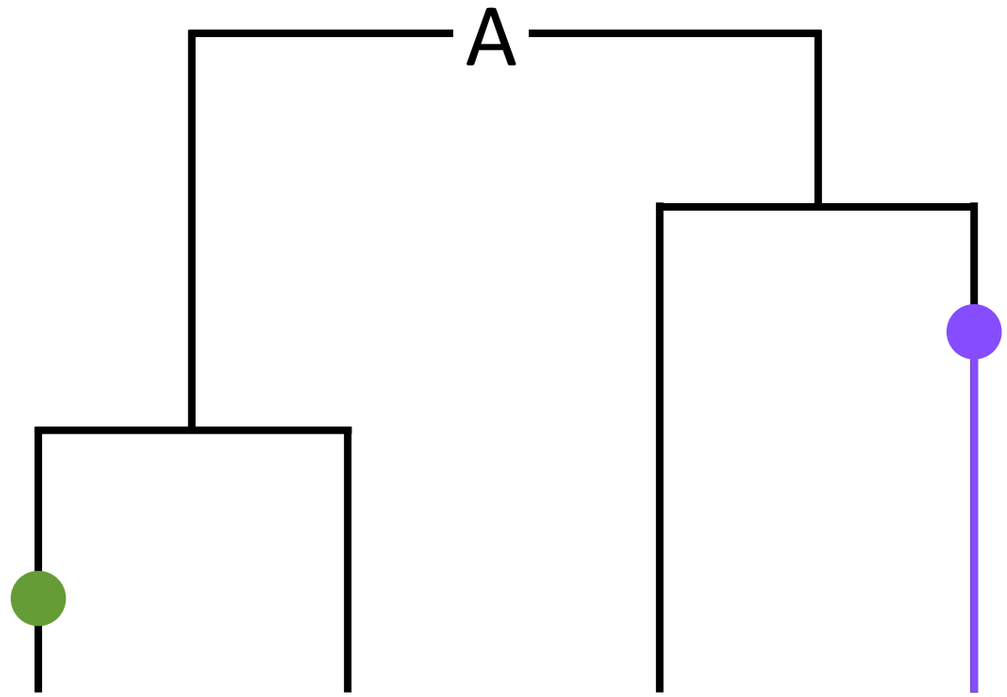


Site pattern frequencies



C	C	G	C
A	A	G	G

Site pattern frequencies



C	A	A	G
C	C	G	C
A	A	G	G
.	.	.	.
.	.	.	.
.	.	.	.

Allman-Rhodes-Erikkson invariants from Singular Value Decomposition

For each set of 4 sequences (quartet), we can count the relative frequencies of the 256 possible site patterns

p_{ijkl}	Taxon A	Taxon B	Taxon C	Taxon D	Frequency
1	A	A	A	A	p_{AAAA}
2	A	A	A	C	p_{AAAC}
3	A	A	A	G	p_{AAAG}
4	A	A	A	T	p_{AAAT}
. . .	. . .	. . .	. . .	. . .	. . .
129	G	G	G	A	p_{GGGA}
130	G	G	G	C	p_{GGGC}
. . .	. . .	. . .	. . .	. . .	. . .
255	T	T	T	G	p_{TTTG}
256	T	T	T	T	p_{TTTT}

Flattening matrices

For each set of 4 sequences (quartet):

Represent the pattern frequencies by three "flattening matrices" (one for each resolution of the quartet):

$$\mathrm{Flat}_{\{1,3\},\{2,4\}}(P) = \begin{array}{c} \\ \mathtt{AA} \\ \mathtt{AC} \\ \mathtt{AG} \\ \mathtt{AT} \\ \mathtt{CA} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} \mathtt{AA} & \mathtt{AC} & \mathtt{AG} & \mathtt{AT} & \mathtt{CA} & \mathtt{CC} & \ldots \end{array} \\ \begin{pmatrix} p_{\mathtt{AAAA}} & p_{\mathtt{AAAC}} & p_{\mathtt{AAAG}} & p_{\mathtt{AAAT}} & p_{\mathtt{ACAA}} & p_{\mathtt{ACAC}} & \cdots \\ p_{\mathtt{AACA}} & p_{\mathtt{AACC}} & p_{\mathtt{AACG}} & p_{\mathtt{AACT}} & p_{\mathtt{ACCA}} & p_{\mathtt{ACCC}} & \cdots \\ p_{\mathtt{AAGA}} & p_{\mathtt{AAGC}} & p_{\mathtt{AAGG}} & p_{\mathtt{AAGT}} & p_{\mathtt{ACGA}} & p_{\mathtt{ACGC}} & \cdots \\ p_{\mathtt{AATA}} & p_{\mathtt{AATC}} & p_{\mathtt{AATG}} & p_{\mathtt{AATT}} & p_{\mathtt{ACTA}} & p_{\mathtt{ACTC}} & \cdots \\ p_{\mathtt{CAAA}} & p_{\mathtt{CAAC}} & p_{\mathtt{CAAG}} & p_{\mathtt{CAAT}} & p_{\mathtt{CCAA}} & p_{\mathtt{CCAC}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{array}.$$

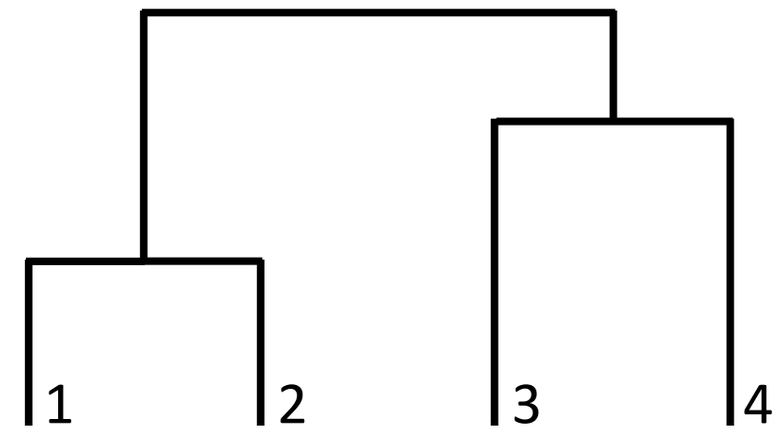
Allman-Rhodes and Eriksson main result:

Under very general Markov assumptions, the flattening matrices are full rank (16) for the two incorrect trees, but the rank of the matrix corresponding to the true is tree expected to be 4.

("rank" = number of linearly independent rows and columns)

Intuition on reduced rank/linear dependencies

True tree

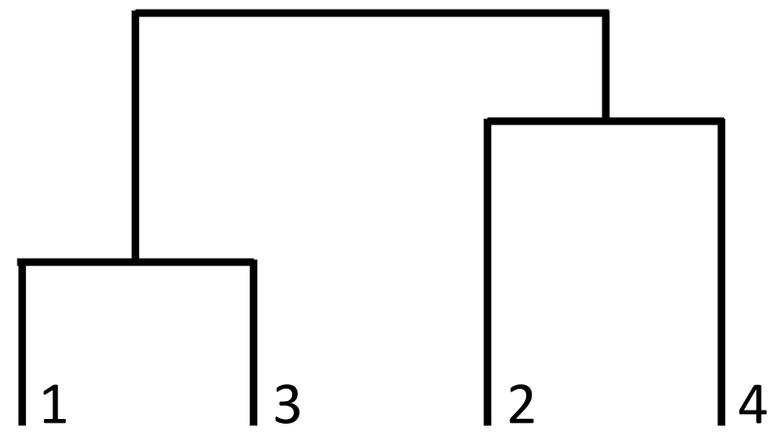


A	C	A	C
A	C	C	A
C	A	A	C
C	A	C	A

E.g., all 4 of these site patterns have the same expected frequency

f(AC|AC)=f(AC|CA)=f(CA|AC)=f(CA|CA)

Incorrect tree

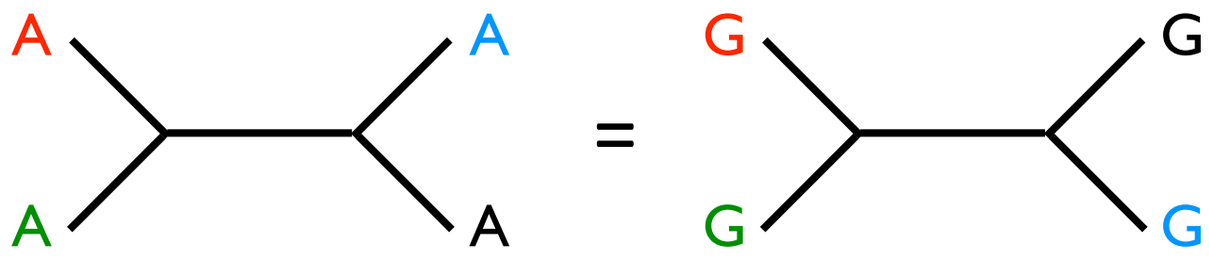


A	A	C	C
A	C	C	A
C	A	A	C
C	C	A	A

These patterns are **not** all expected to have the same expected frequency *if they evolved on the other tree*

f(AA|CC)≠f(AC|CA)≠f(CA|AC)≠f(CC|AA)

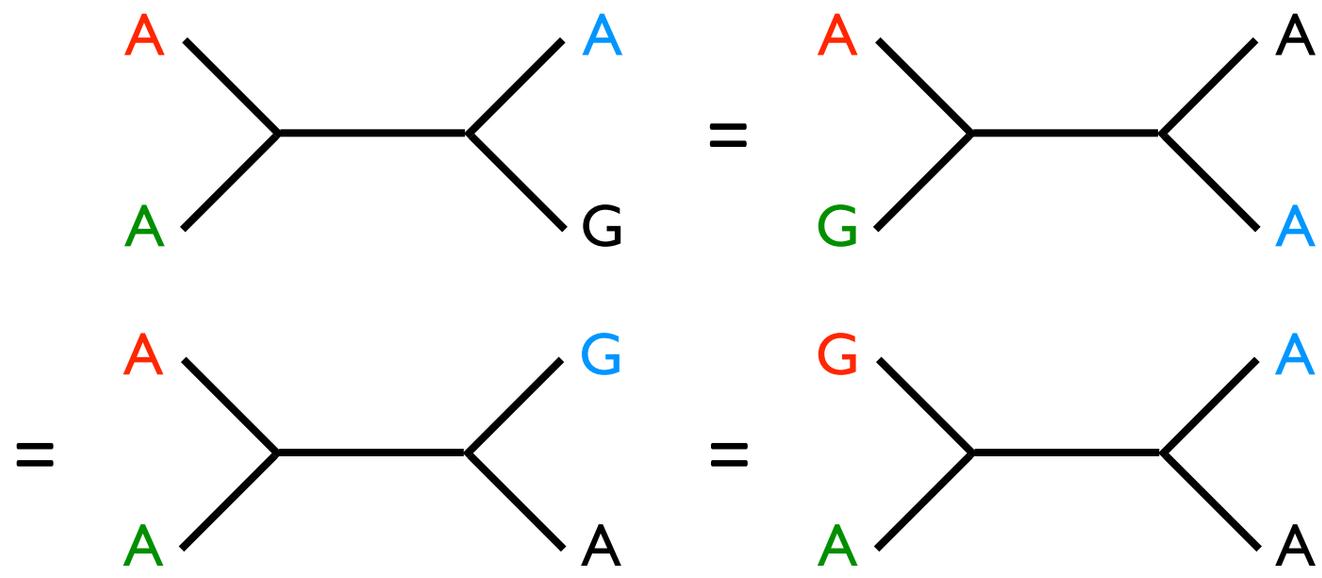
Flattenings for a 2-State Jukes-Cantor model



Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*			
AG				
GA				
GG				*a*

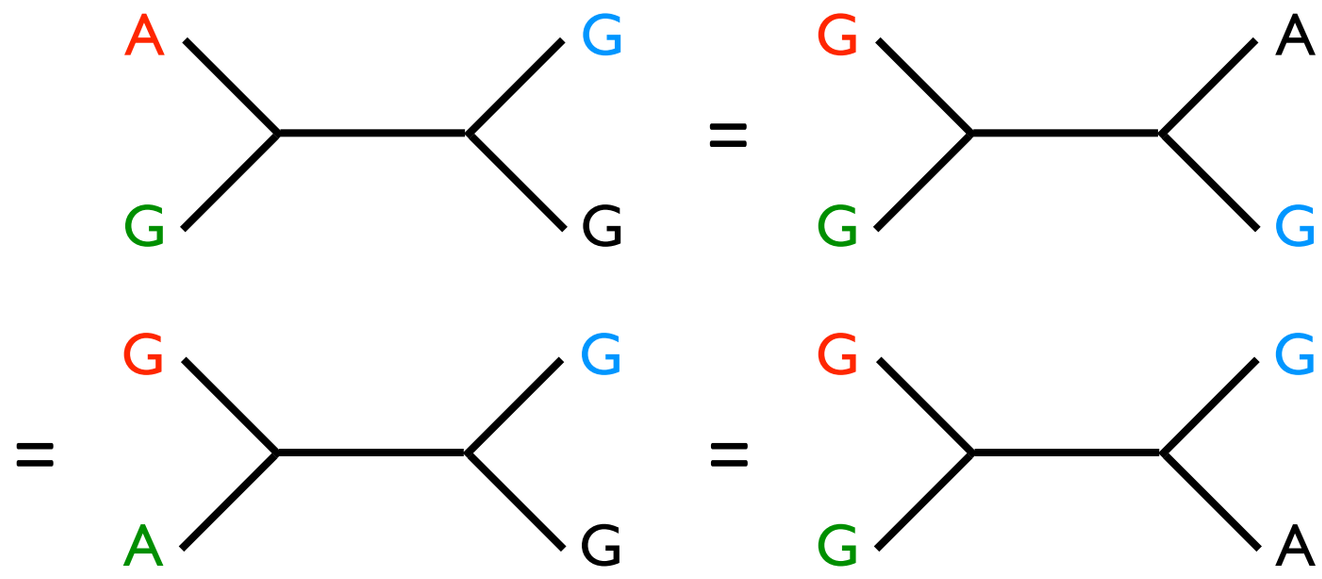
Flattenings for a 2-State Jukes-Cantor model



Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	
AG	*b*			
GA	*b*			
GG				*a*

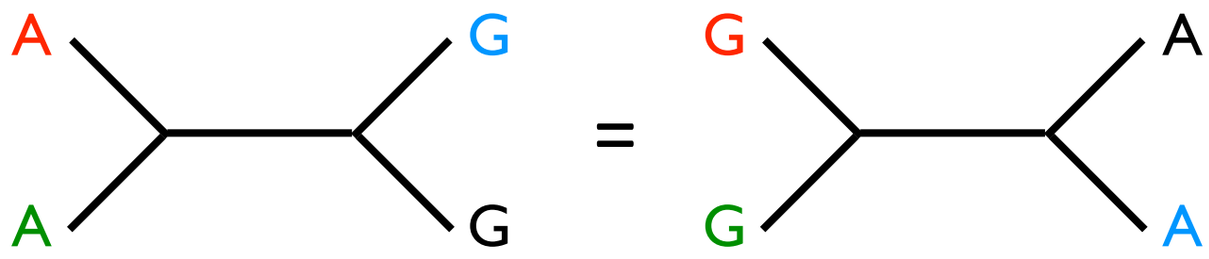
Flattenings for a 2-State Jukes-Cantor model



Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	
AG	*b*			*b*
GA	*b*			*b*
GG		*b*	*b*	*a*

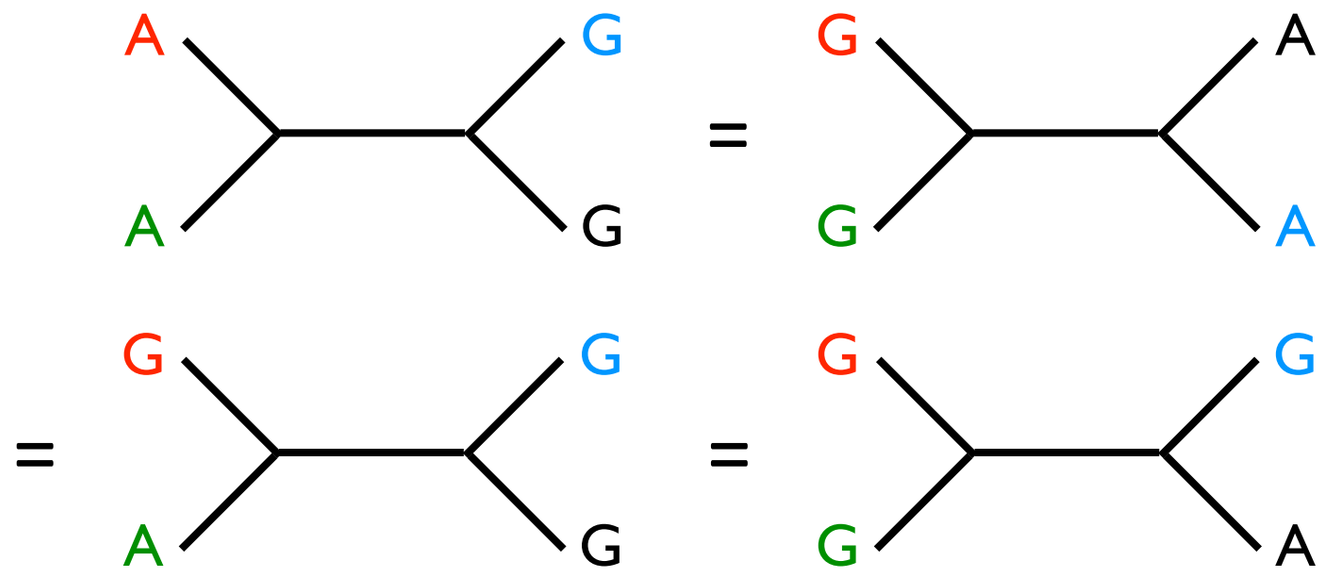
Flattenings for a 2-State Jukes-Cantor model



Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*c*
AG	*b*			*b*
GA	*b*			*b*
GG	*c*	*b*	*b*	*a*

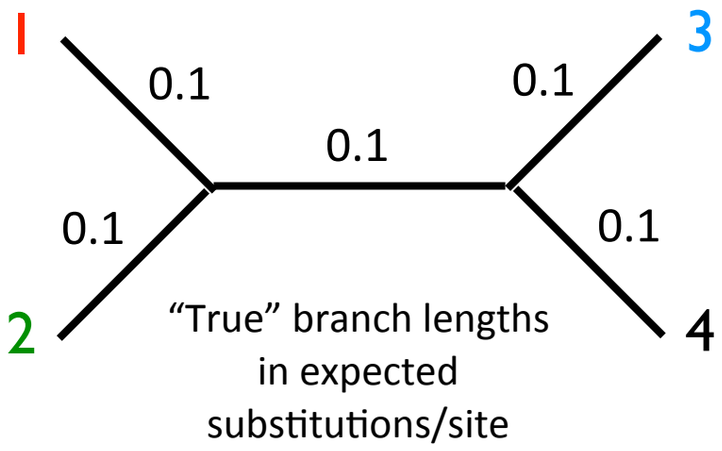
Flattenings for a 2-State Jukes-Cantor model



Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*c*
AG	*b*	*d*	*d*	*b*
GA	*b*	*d*	*d*	*b*
GG	*c*	*b*	*b*	*a*

Some numbers



Expected flattening matrix for 1,2|3,4

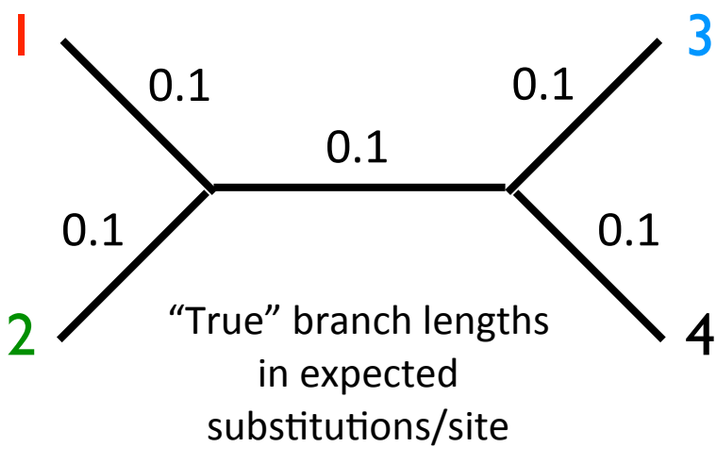
	AA	AG	GA	GG
AA	0.093008	0.061355	0.061355	0.068115
AG	0.061355	0.046728	0.046728	0.061355
GA	0.061355	0.046728	0.046728	0.061355
GG	0.068115	0.061355	0.061355	0.093008

Expected site-pattern frequencies

p_{AAAA}	0.09300841
p_{AAAG}	0.06135527
p_{AAGA}	0.06135527
p_{AAGG}	0.06811487
p_{AGAA}	0.06135527
p_{AGAG}	0.04672782
p_{AGGA}	0.04672782
p_{AGGG}	0.06135527
p_{GAAA}	0.06135527
p_{GAAG}	0.04672782
p_{GAGA}	0.04672782
p_{GAGG}	0.06135527
p_{GGAA}	0.06811487
p_{GGAG}	0.06135527
p_{GGGA}	0.06135527
p_{GGGG}	0.09300841

etc.

Some numbers



Expected site-pattern frequencies

p_{AAAA}	0.09300841
p_{AAAG}	0.06135527
p_{AAGA}	0.06135527
p_{AAGG}	0.06811487
p_{AGAA}	0.06135527
p_{AGAG}	0.04672782
p_{AGGA}	0.04672782
p_{AGGG}	0.06135527
p_{GAAA}	0.06135527
p_{GAAG}	0.04672782
p_{GAGA}	0.04672782
p_{GAGG}	0.06135527
p_{GGAA}	0.06811487
p_{GGAG}	0.06135527
p_{GGGA}	0.06135527
p_{GGGG}	0.09300841

Expected flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	0.093008	0.061355	0.061355	0.068115
AG	0.061355	0.046728	0.046728	0.061355
GA	0.061355	0.046728	0.046728	0.061355
GG	0.068115	0.061355	0.061355	0.093008

Delete redundant 3rd row and column...

	AA	AG	GG
AA	0.093008	0.061355	0.068115
AG	0.061355	0.046728	0.061355
GG	0.068115	0.061355	0.093008

Note that we can now obtain the last column of the above matrix as a linear combination of the first two columns:

$$f_{AA,GG} = -f_{AA,AA} + 2.62617\, f_{AA,AG} = 0.068115$$
$$f_{AG,GG} = -f_{AG,AA} + 2.62617\, f_{AG,AG} = 0.061355$$
$$f_{GG,GG} = -f_{GG,AA} + 2.62617\, f_{GG,AG} = 0.093008$$

$\therefore$ ***matrix* has *only two linearly independent rows and columns; rank is 2***

Estimating the rank

To estimate the rank, compute the singular value decomposition (SVD) of each matrix:

$$\boldsymbol{s} = [s_1, s_2, s_3, \ldots s_{16}].$$

For the true tree, $\sum_{i=5}^{16} s_i \approx 0$. Otherwise, $\sum_{i=5}^{16} s_i > 0$

Thus, for each of the three trees for four taxa, we can compute the Frobenius distance from each to the nearest rank-4 matrix:

$$score = \sqrt{\sum_{i=5}^{16} s_i^2}$$

where the s_i are the 16 singular values resulting from the SVD

Then choose the tree with the lowest score.

The Singular Value Decomposition (SVD)

Decompose an initial matrix into 3 new ones, such that multiplying the new matrices as shown below returns the original matrix exactly

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

The Singular Value Decomposition (SVD)

$$f_{12,34} = \begin{bmatrix} 0.093008 & 0.061355 & 0.061355 & 0.068115 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.068115 & 0.061355 & 0.061355 & 0.093008 \end{bmatrix}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T$$

$$\mathbf{U} = \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.254578 & 0 & 0 & 0 \\ 0 & 0.024893 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Last two singular values are zero; rank = 2

Check:

$$f_{12,34} = \mathbf{U}\Sigma\mathbf{V}^T$$

$$= \begin{bmatrix} -0.562539 & 0.707107 & 0.428427 & 0 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.428427 & 0 & -0.562539 & -0.707107 \\ -0.562539 & 0.707107 & 0.428427 & 0 \end{bmatrix} \begin{bmatrix} 0.254578 & 0 & 0 & 0 \\ 0 & 0.024893 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -0.562539 & -0.428427 & -0.428427 & -0.562539 \\ 0.707107 & 0 & 0 & 0.707107 \\ 0.428427 & -0.562539 & -0.562539 & 0.428427 \\ 0 & -0.707107 & -0.707107 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.093008 & 0.061355 & 0.061355 & 0.068115 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.061355 & 0.046728 & 0.046728 & 0.061355 \\ 0.068115 & 0.061355 & 0.061355 & 0.093008 \end{bmatrix}$$

✔

When the flattening corresponds to a tree that did not generate the data

Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*c*
AG	*b*	*d*	*d*	*b*
GA	*b*	*d*	*d*	*b*
GG	*c*	*b*	*b*	*a*

Flattening matrix for 1,3|2,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	
AG				
GA				
GG				

When the flattening corresponds to a tree that did not generate the data

Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*c*
AG	*b*	*d*	*d*	*b*
GA	*b*	*d*	*d*	*b*
GG	*c*	*b*	*b*	*a*

Flattening matrix for 1,3|2,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*d*
AG		*c*		
GA				
GG				

When the flattening corresponds to a tree that did not generate the data

Flattening matrix for 1,2|3,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*c*
AG	*b*	*d*	*d*	*b*
GA	*b*	*d*	*d*	*b*
GG	*c*	*b*	*b*	*a*

Flattening matrix for 1,3|2,4

	AA	AG	GA	GG
AA	*a*	*b*	*b*	*d*
AG	*b*	*c*	*d*	*b*
GA	*b*	*d*	*c*	*b*
GG	*d*	*b*	*b*	*a*

No redundant rows;
matrix is full rank (=4)

When the flattening corresponds to a tree that did not generate the data

Demonstration that matrix is full rank:

$$f_{13,24} = \begin{bmatrix} 0.093008 & 0.061355 & 0.061355 & 0.046728 \\ 0.061355 & 0.068115 & 0.046728 & 0.061355 \\ 0.061355 & 0.046728 & 0.068115 & 0.061355 \\ 0.046728 & 0.061355 & 0.061355 & 0.093008 \end{bmatrix}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T$$

$$\mathbf{U} = \begin{bmatrix} -0.524622 & 0.707107 & 0 & 0.474101 \\ -0.474101 & 0 & -0.707107 & -0.524622 \\ -0.474101 & 0 & 0.707107 & -0.524622 \\ -0.524622 & -0.707107 & 0 & 0.474101 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} -0.524622 & 0.707107 & 0 & 0.474101 \\ -0.474101 & 0 & -0.707107 & -0.524622 \\ -0.474101 & 0 & 0.707107 & -0.524622 \\ -0.524622 & -0.707107 & 0 & 0.474101 \end{bmatrix}$$

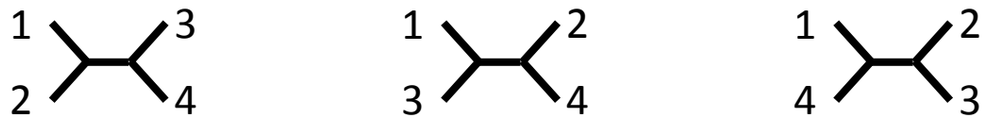
$$\Sigma = \begin{bmatrix} 0.250629 & 0 & 0 & 0 \\ 0 & 0.046280 & 0 & 0 \\ 0 & 0 & 0.021387 & 0 \\ 0 & 0 & 0 & 0.003950 \end{bmatrix}$$

All singular values are nonzero;
matrix is full rank (= 4)

Calculation of SVD Scores (4-state data)

$$score = \sqrt{\sum_{i=5}^{16} s_i^2}$$

= "Frobenius distance" to nearest rank 4 matrix



SV (*s*)	1,2 I 3,4	1,3 I 2,4	1,4 I 2,3
1	0.279686	0.278714	0.278716
2	0.218990	0.219191	0.219191
3	0.109020	0.110392	0.110389
4	0.056873	0.057090	0.057090
5	8.00E-05	0.006875	0.006886
6	6.14E-05	0.006315	0.006305
7	4.93E-05	0.003286	0.003286
8	3.80E-05	0.003244	0.003246
9	3.26E-05	0.002905	0.002903
10	3.09E-05	0.002499	0.002499
11	2.69E-05	0.001471	0.001472
12	2.23E-05	0.001182	0.001181
13	1.30E-05	0.001009	0.001008
14	1.03E-05	0.000937	0.000937
15	6.19E-06	0.000382	0.000384
16	1.56E-06	0.000377	0.000375
score	0.000133	0.011353	0.011354

Simulation conditions:

- tree = (((1:0.05,2:0.05):0.05,3:0.1):0.05,4:0.15)
- 1,000,000 sites
- HKY model: κ=4 π=(0.1, 0.2, 0.3, 0.4)
- all sites share same history (no incomplete lineage sorting, horizontal transfer, gene duplication and loss, etc.)

Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

$12|34$
$12|35$
$12|45$
$14|35$
$23|45$

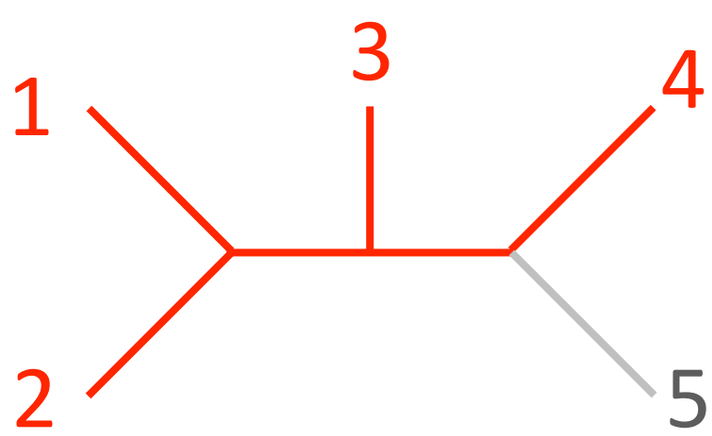
Suppose we infer these quartet relationships for 5 taxa

Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

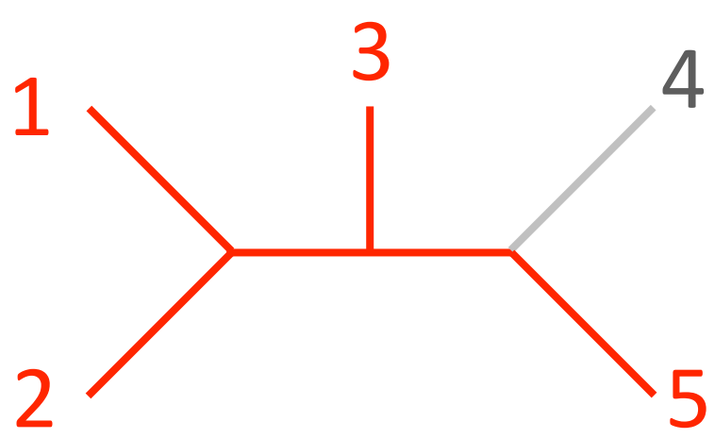


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

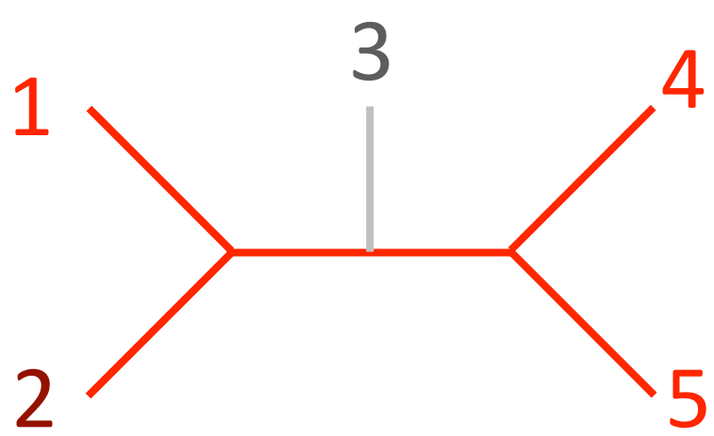


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

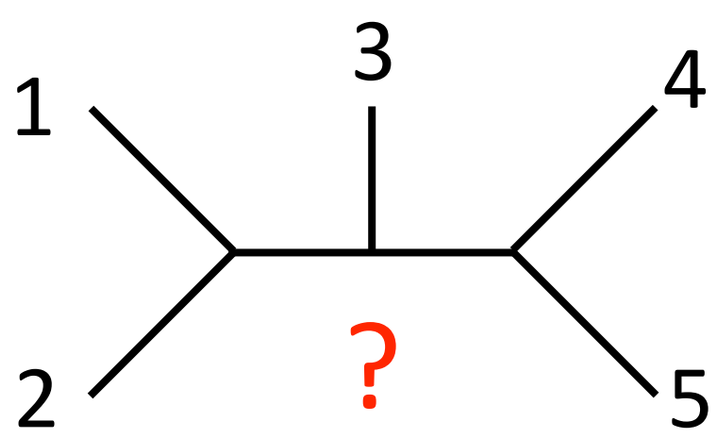


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

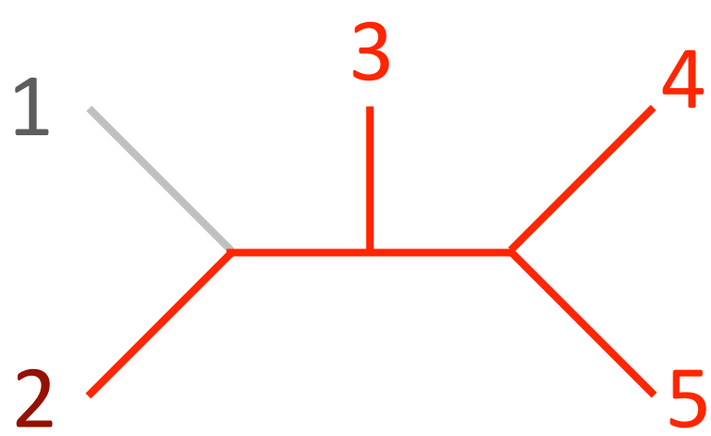


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

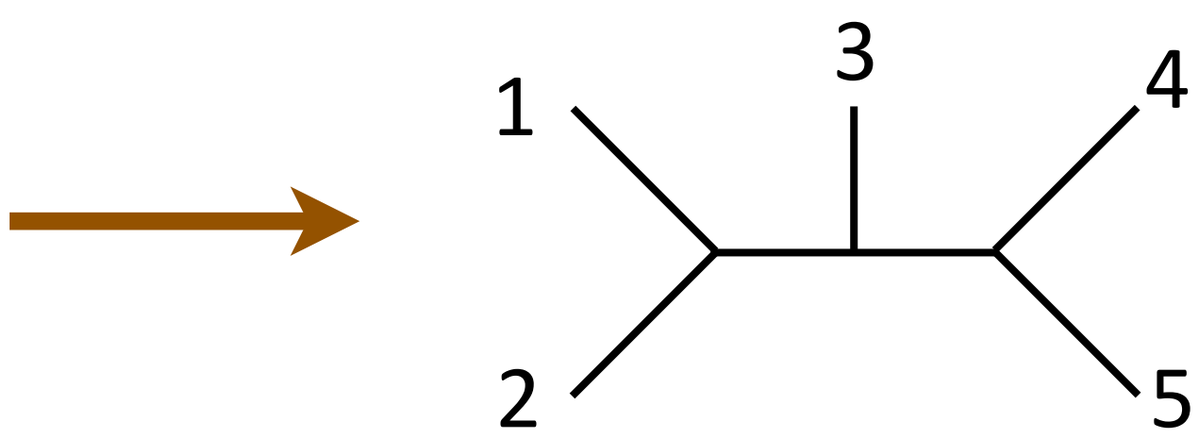


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45



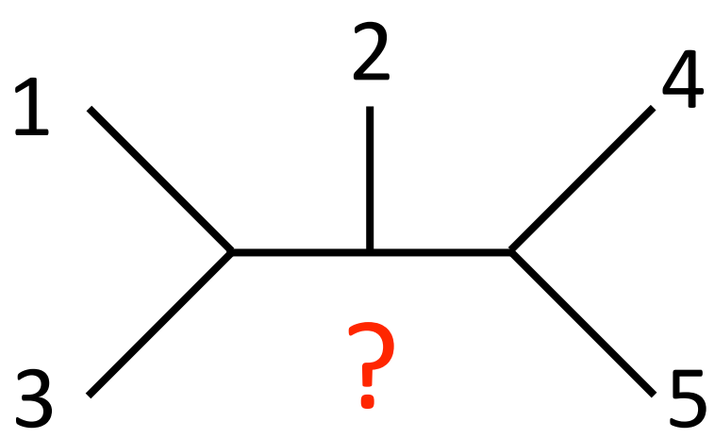
4 consistent quartets, 1 inconsistent quartet

Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

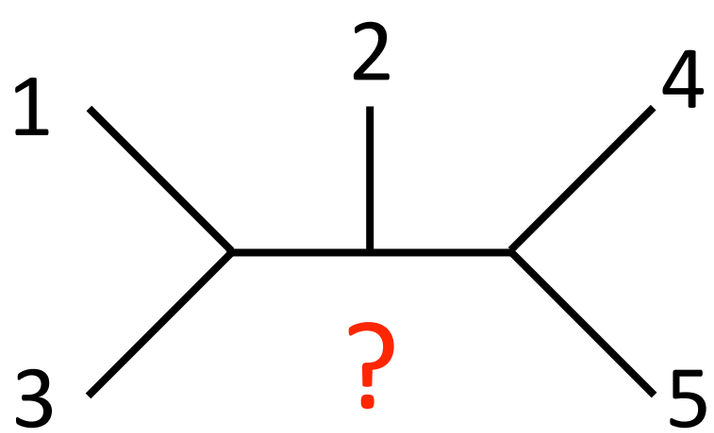


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

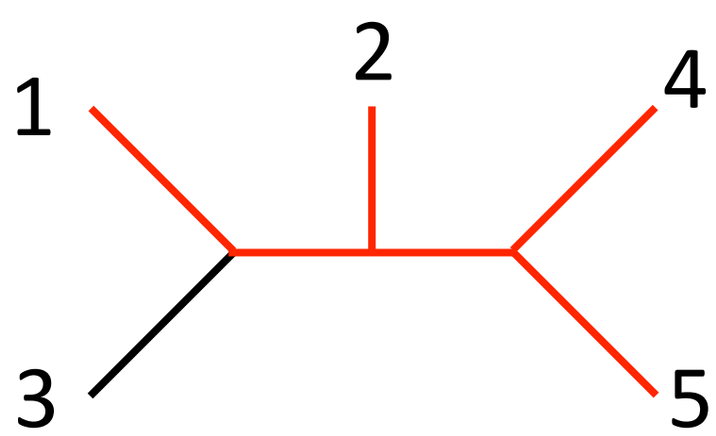


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

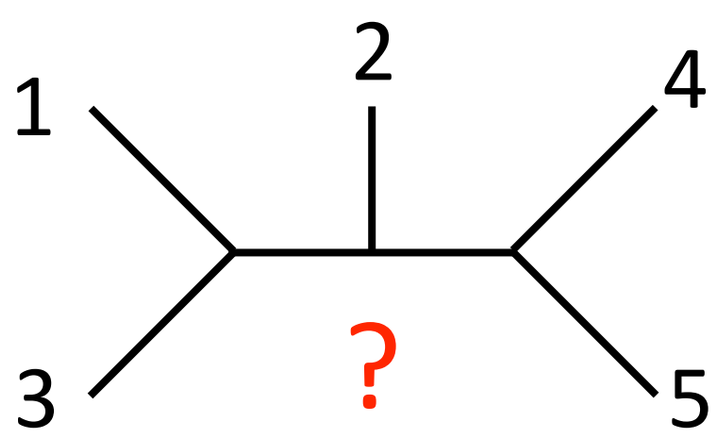


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

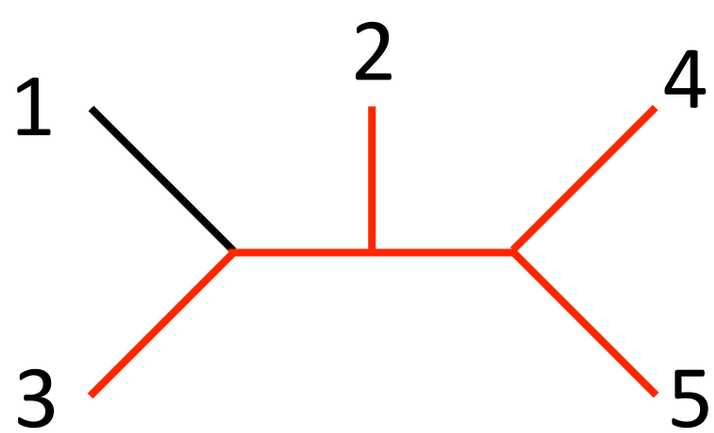


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45

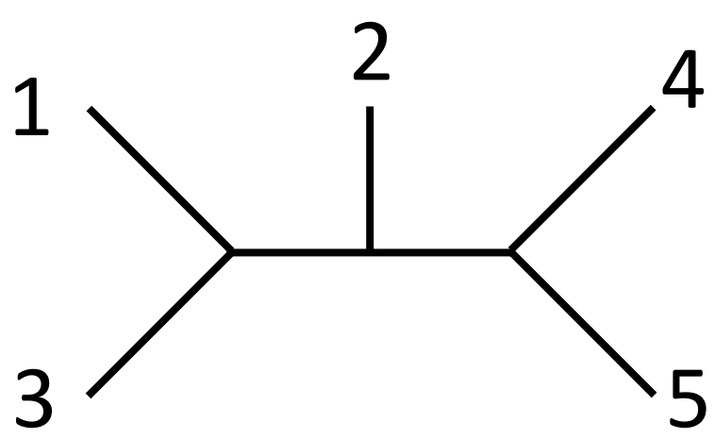


Handling >4 taxa

Compute invariant scores for all quartets, choosing the best resolution for each one.

Search for a tree that minimizes the number of *inconsistent quartets* (i.e., seek a solution to the Maximum Quartet Consistency problem).

12|34
12|35
12|45
14|35
23|45



2 consistent quartets, 3 inconsistent quartet

Now evaluate the remaining 13 trees and choose the one that maximizes the number of consistent quartets

Handling >4 taxa

While evaluation of each possible tree might work well for 5-tip trees, the number of possible trees for *n* tips grows too quickly to make it a general strategy.

Must use a heuristic algorithm to search for the best tree:

- The default in PAUP* is a heavily modified version of "QFM" (Reaz et al., 2014)
- Other algorithms are available in PAUP* and elsewhere
- Unfortunately, the MQC problem is NP-hard (i.e., exact solution will be slow for large numbers of tips)

Allman-Rhodes-Eriksson method (ErikSVD)

Work for extremely general models:

Does assume that all sites in the alignment are independently and identically distributed according to a general Markov model.

But:

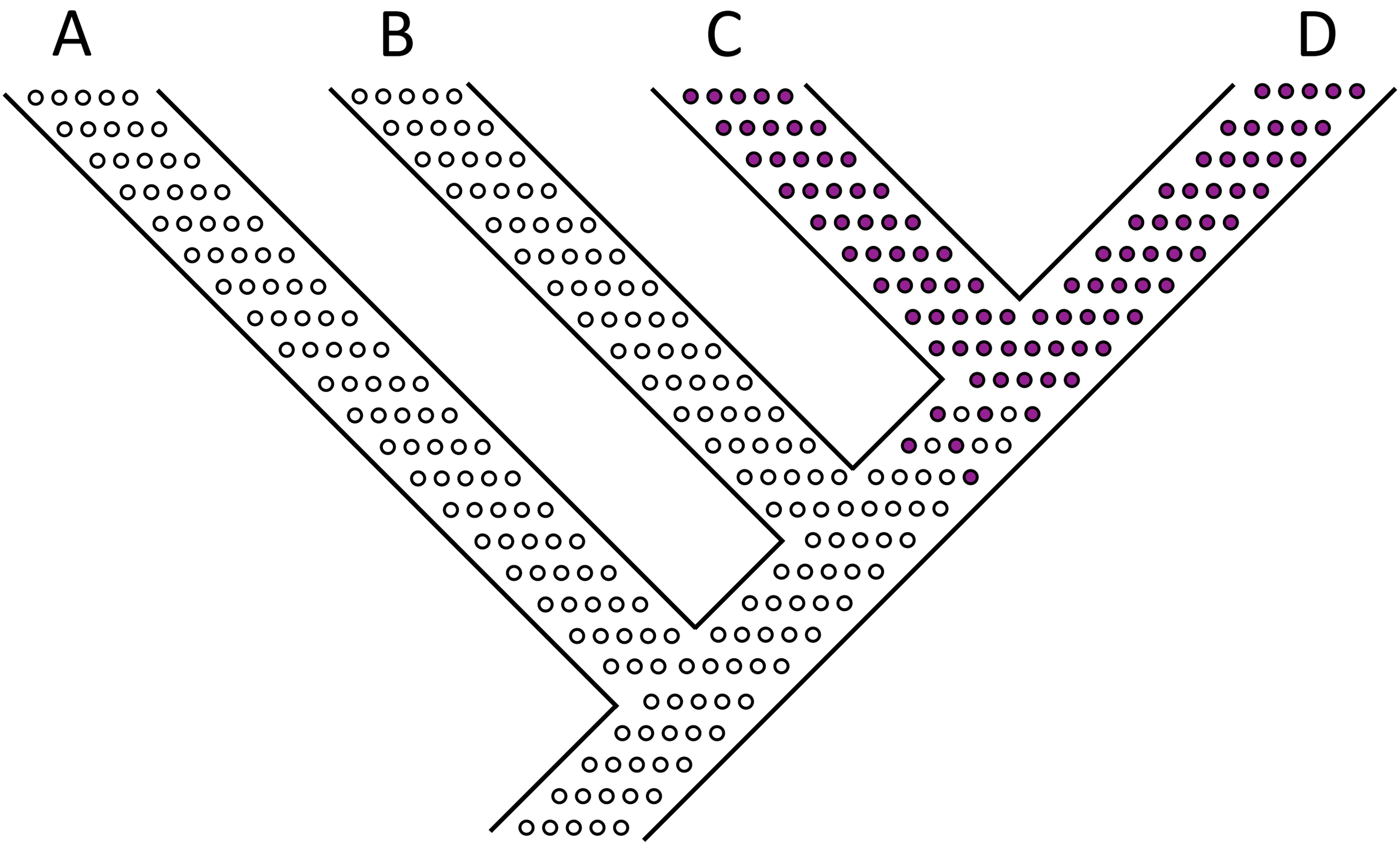
- No assumption of stationarity/time-reversibility!
- No assumption of homogeneity over the tree!

  Each branch may have its own transition matrix, or even multiple transition matrices along the same branch.

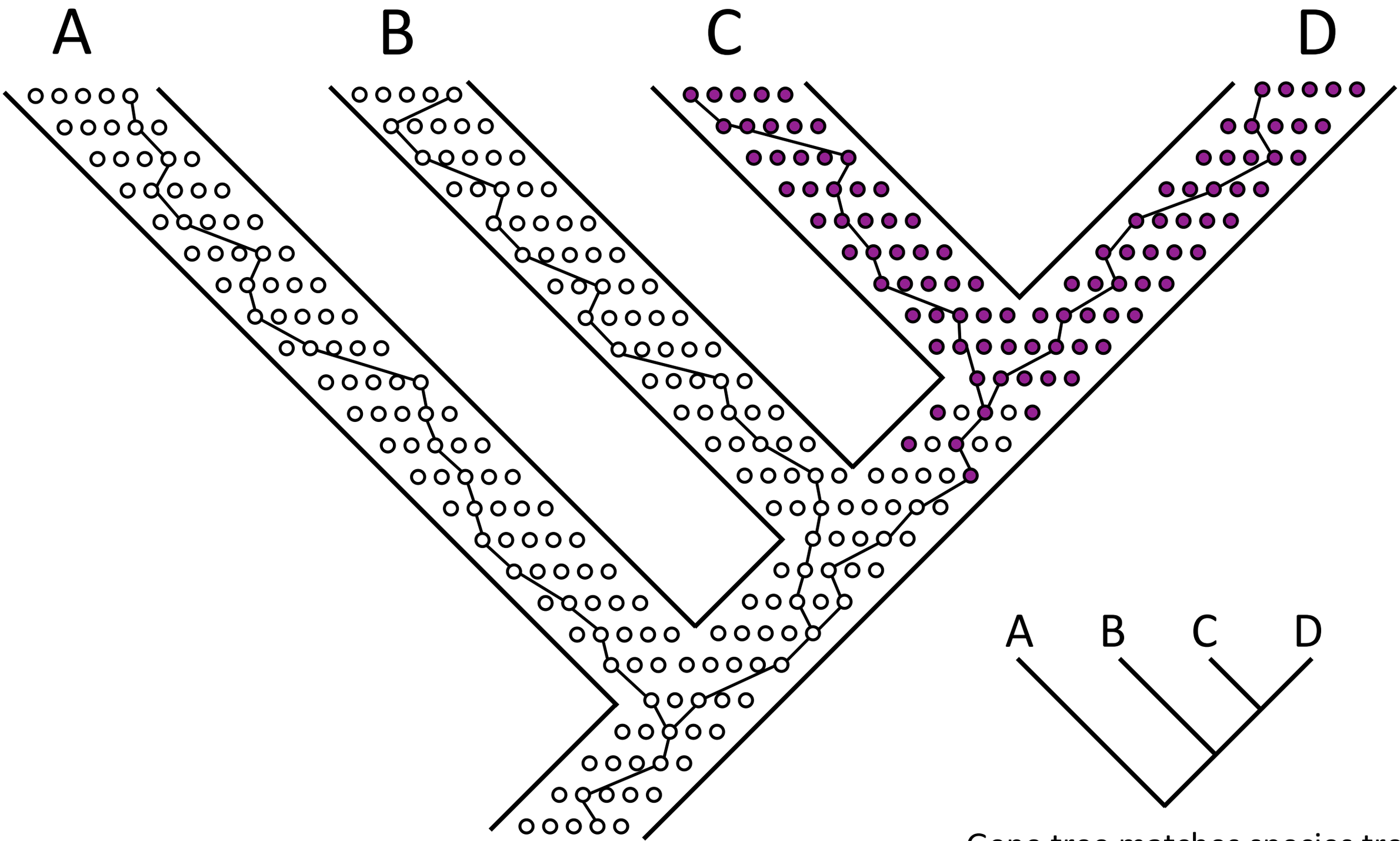
Enables inferences about the tree topology to be made without having to estimate the parameters of the underlying model(s).

A recent improvement by Jesús Fernández-Sánches and Marta Casanellas (Syst. Biol. 2016): Do an additional row and column normalization to reduce the error associated with low counts for certain entries of the flattening matrices ("Erik+2")

Ancestral polymorphism and species trees

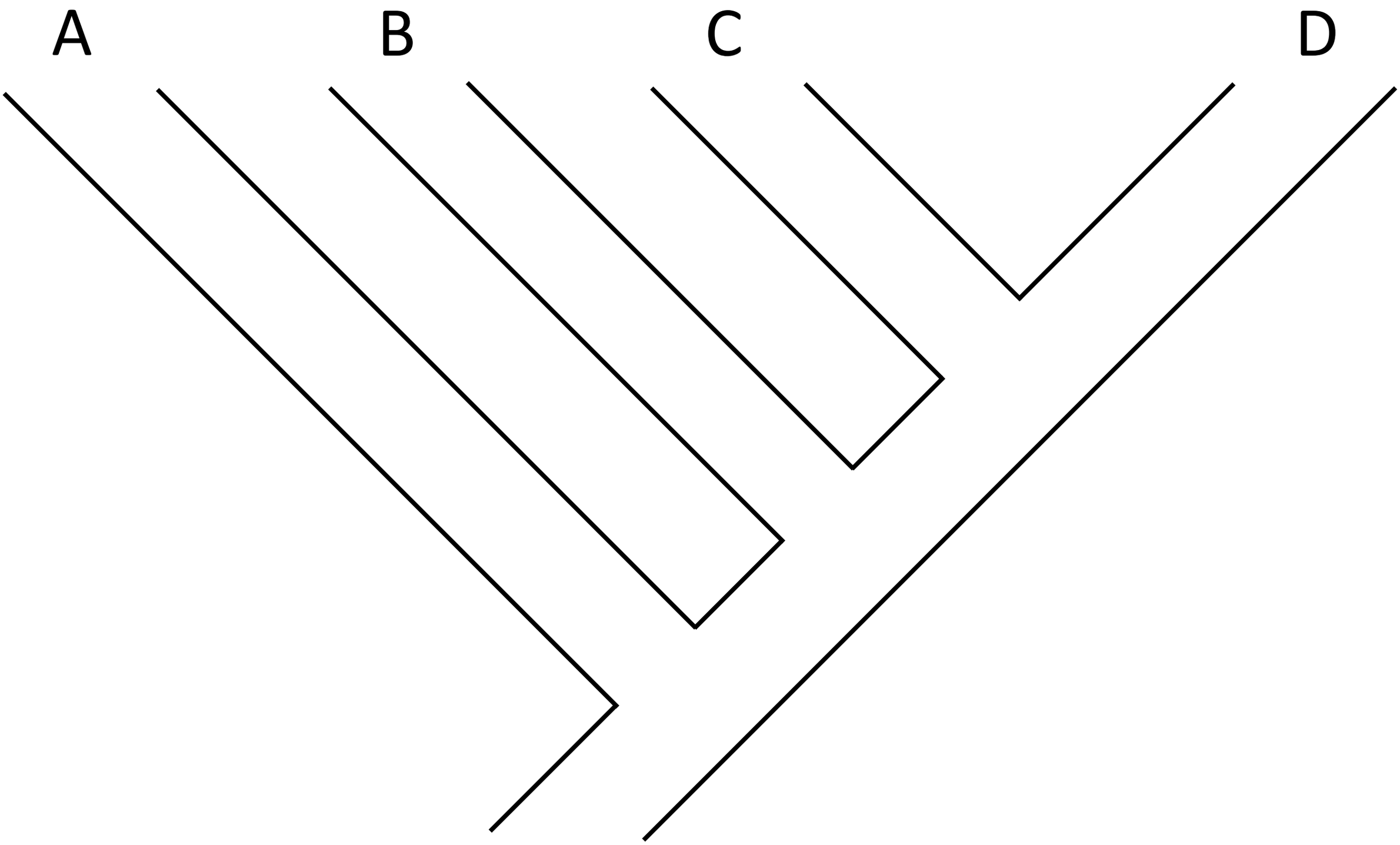


Ancestral polymorphism and species trees

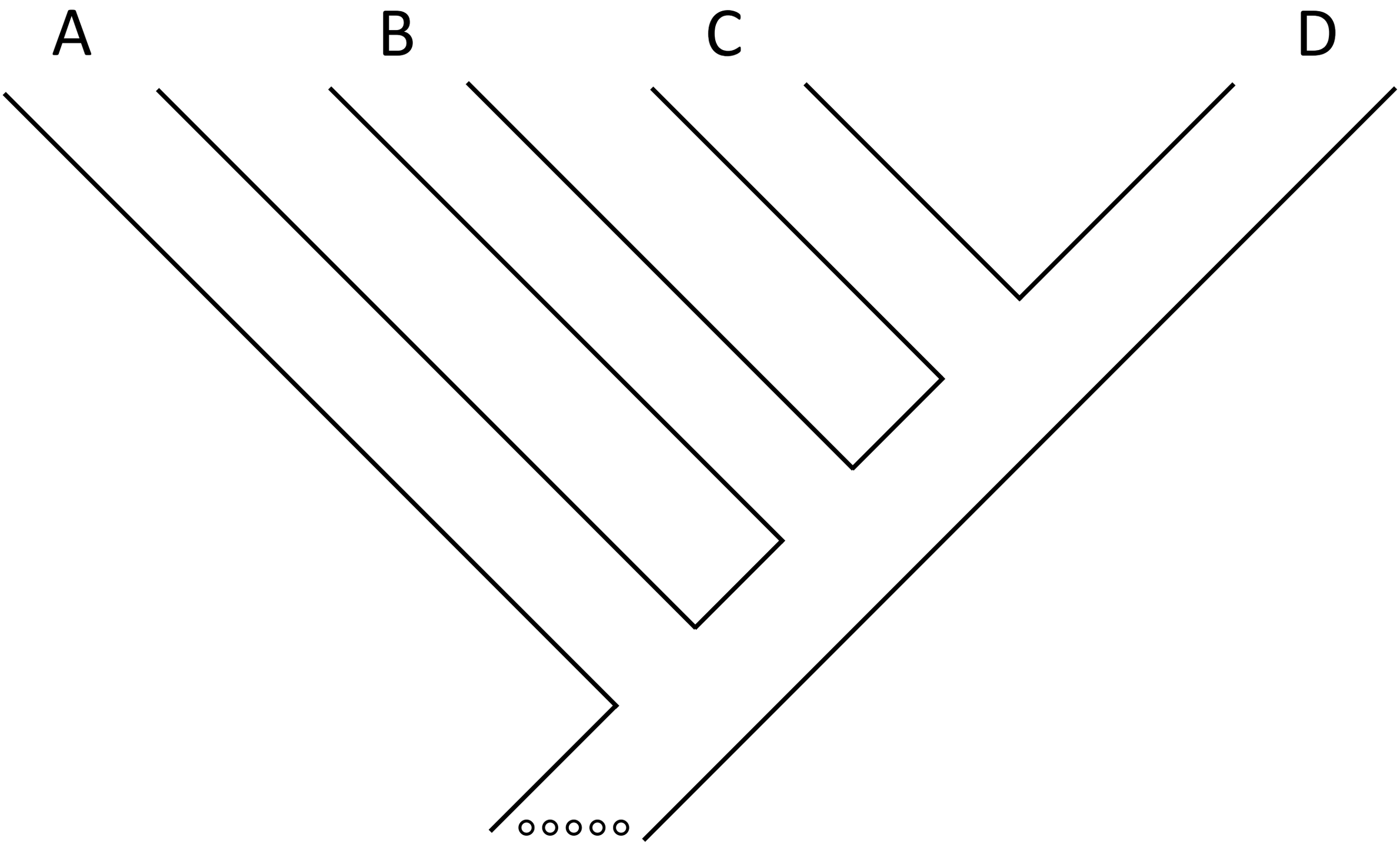


Gene tree matches species tree

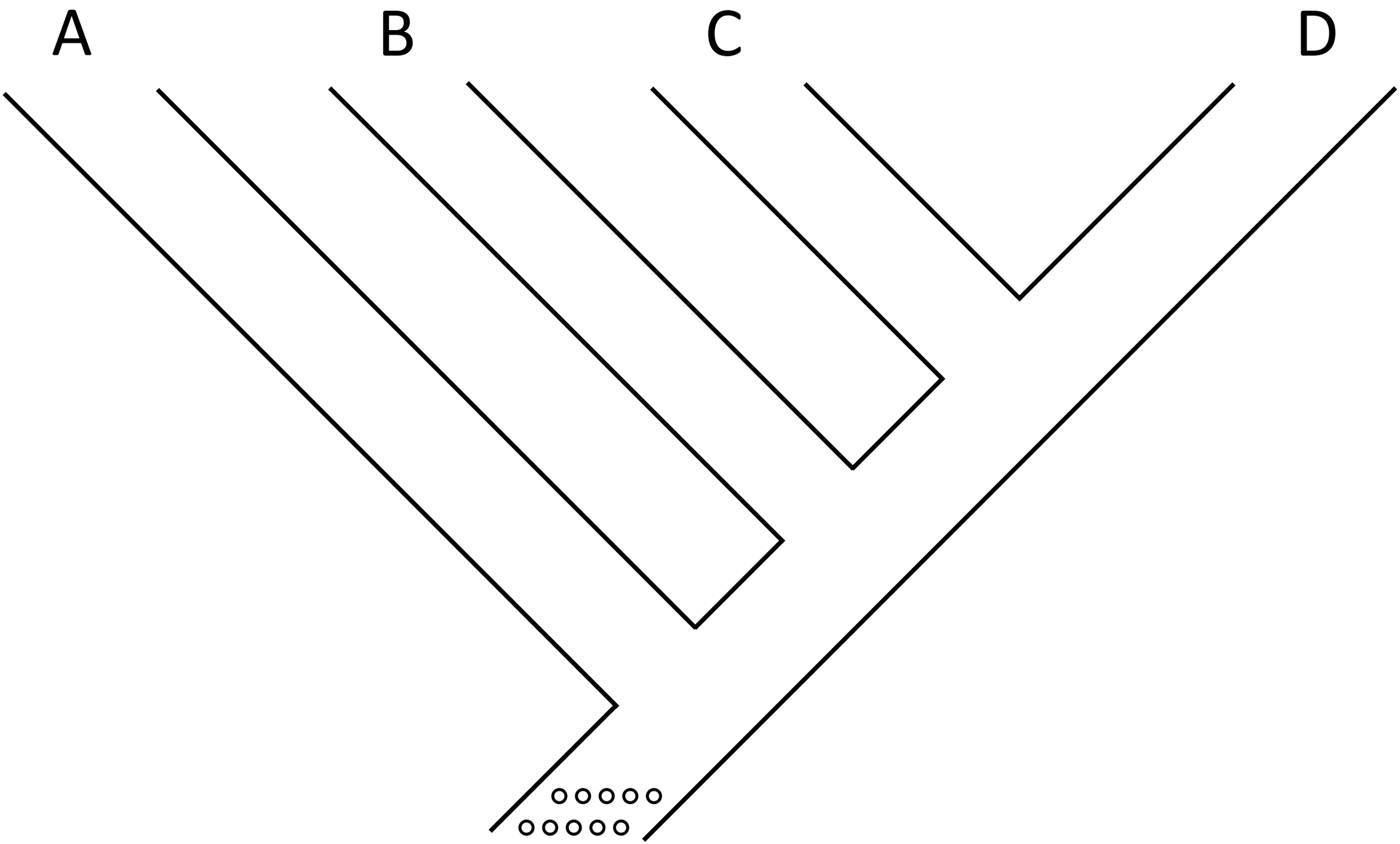
Ancestral polymorphism and species trees



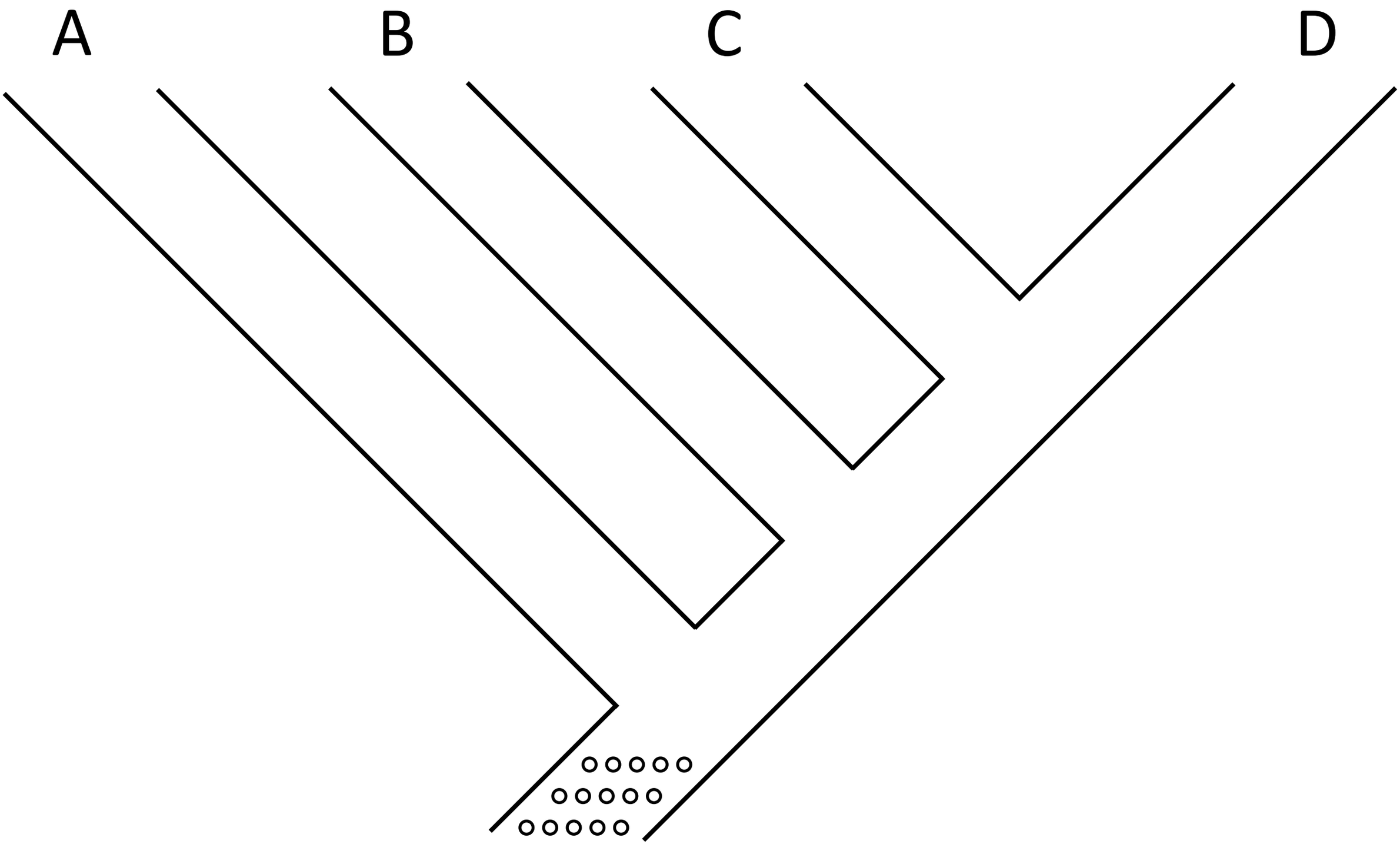
Ancestral polymorphism and species trees



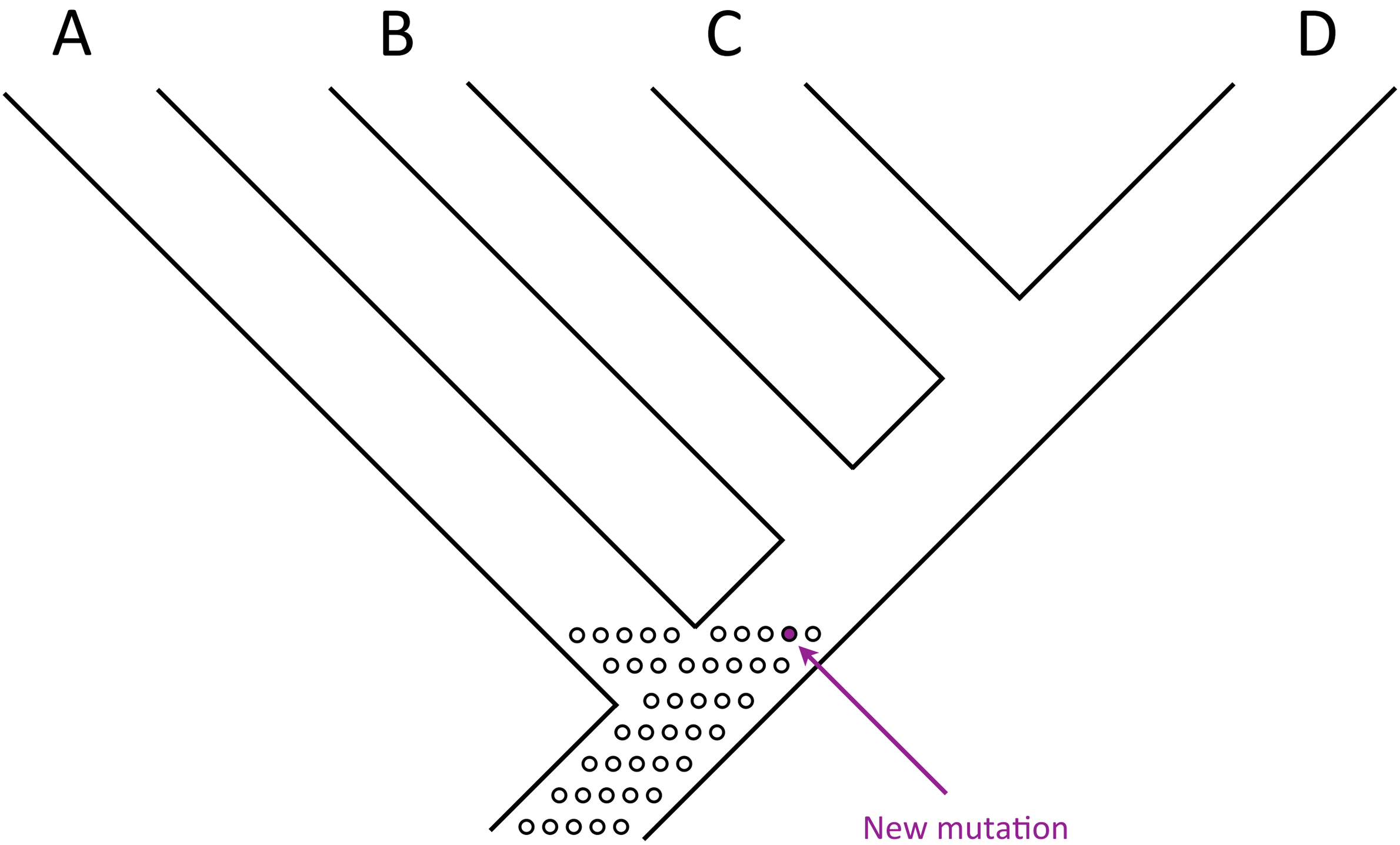
Ancestral polymorphism and species trees



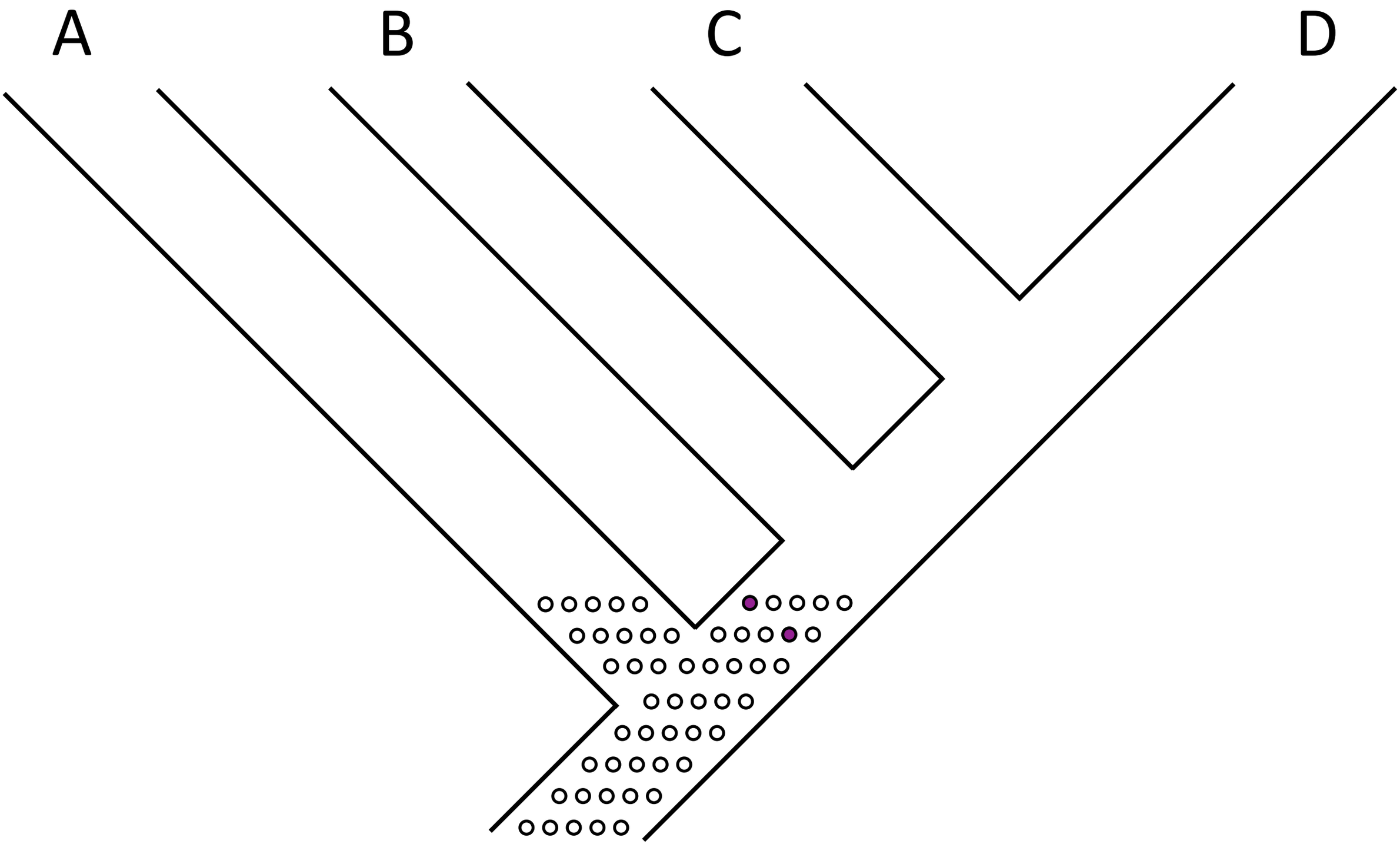
Ancestral polymorphism and species trees



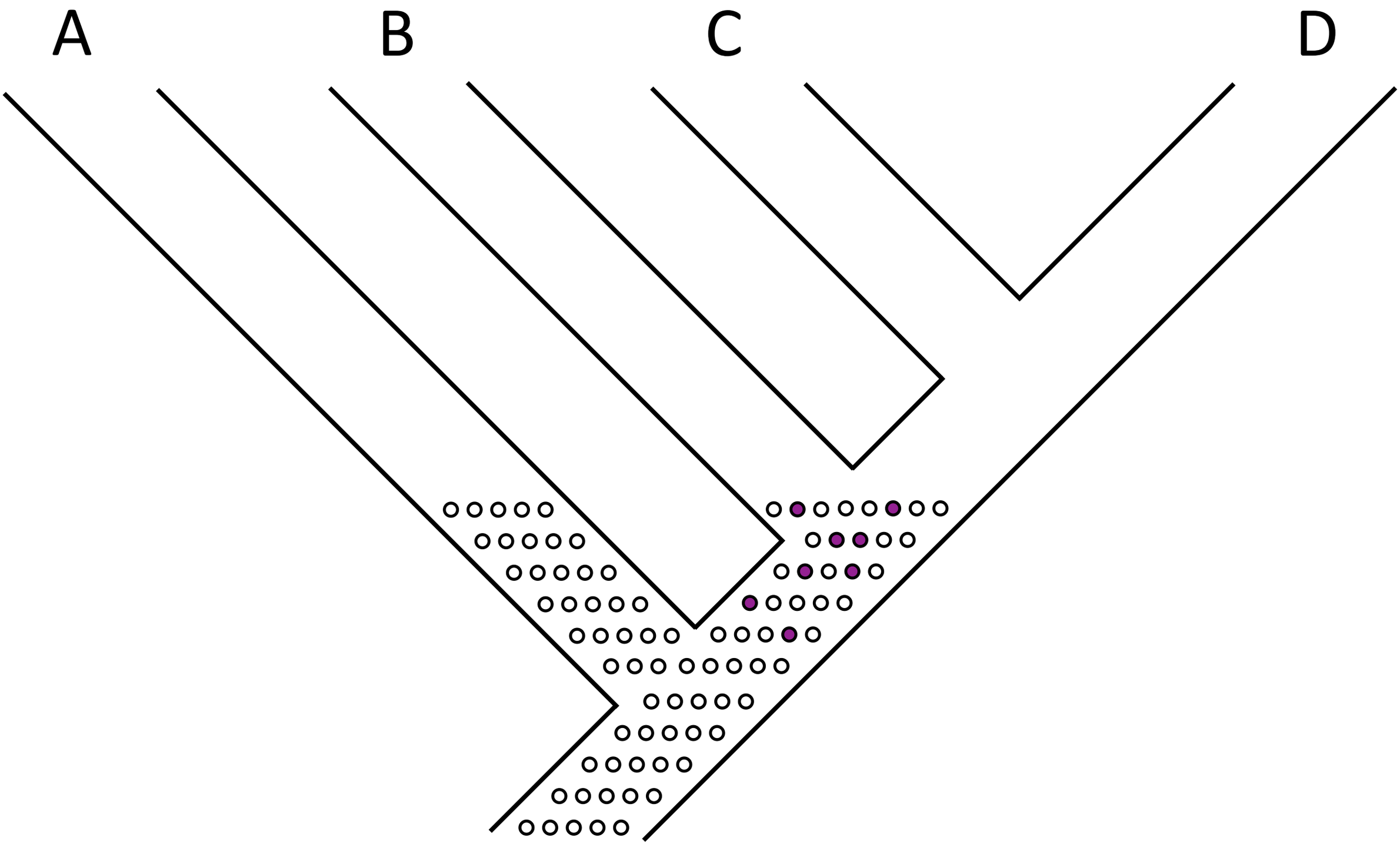
Ancestral polymorphism and species trees



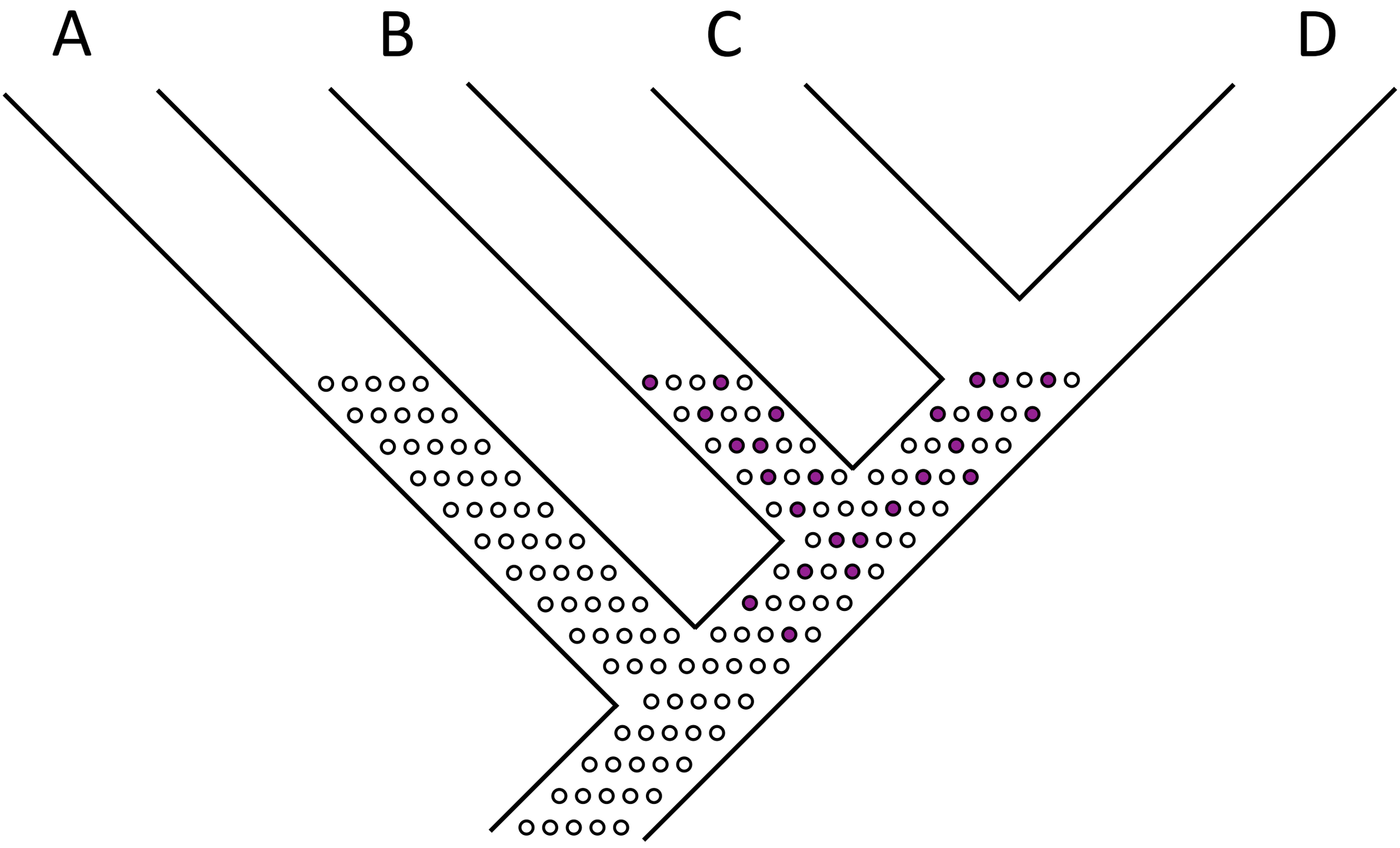
Ancestral polymorphism and species trees



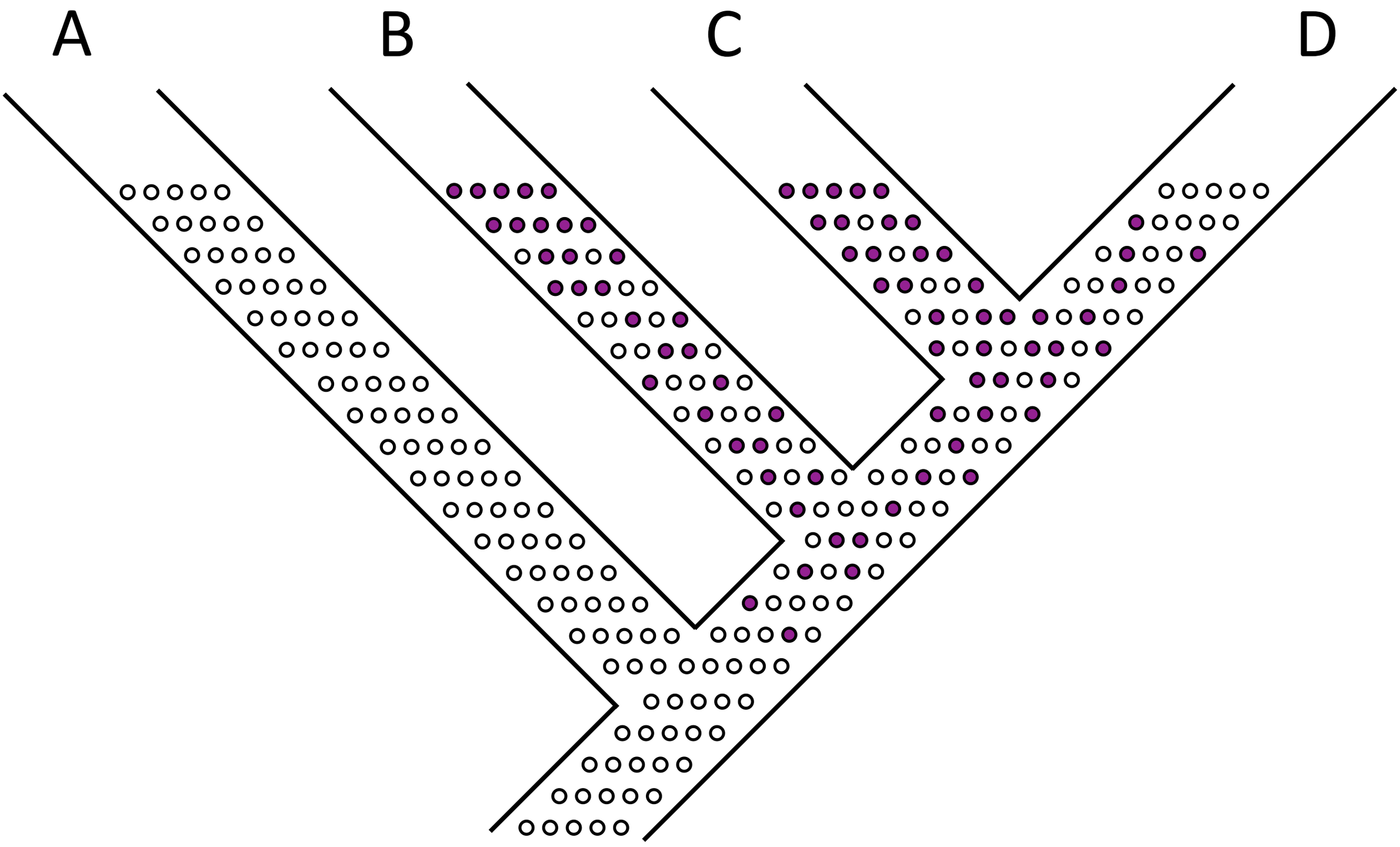
Ancestral polymorphism and species trees



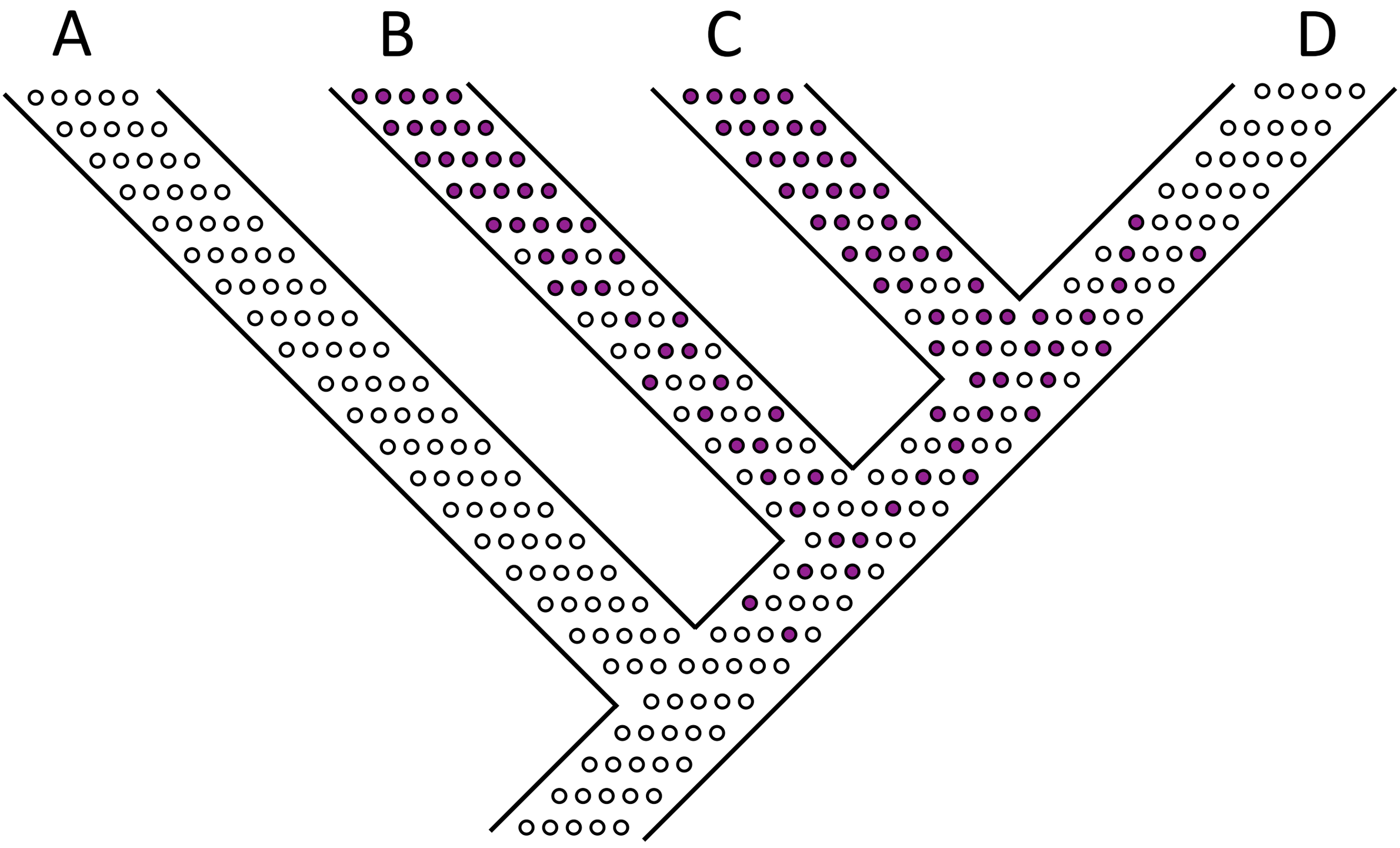
Ancestral polymorphism and species trees



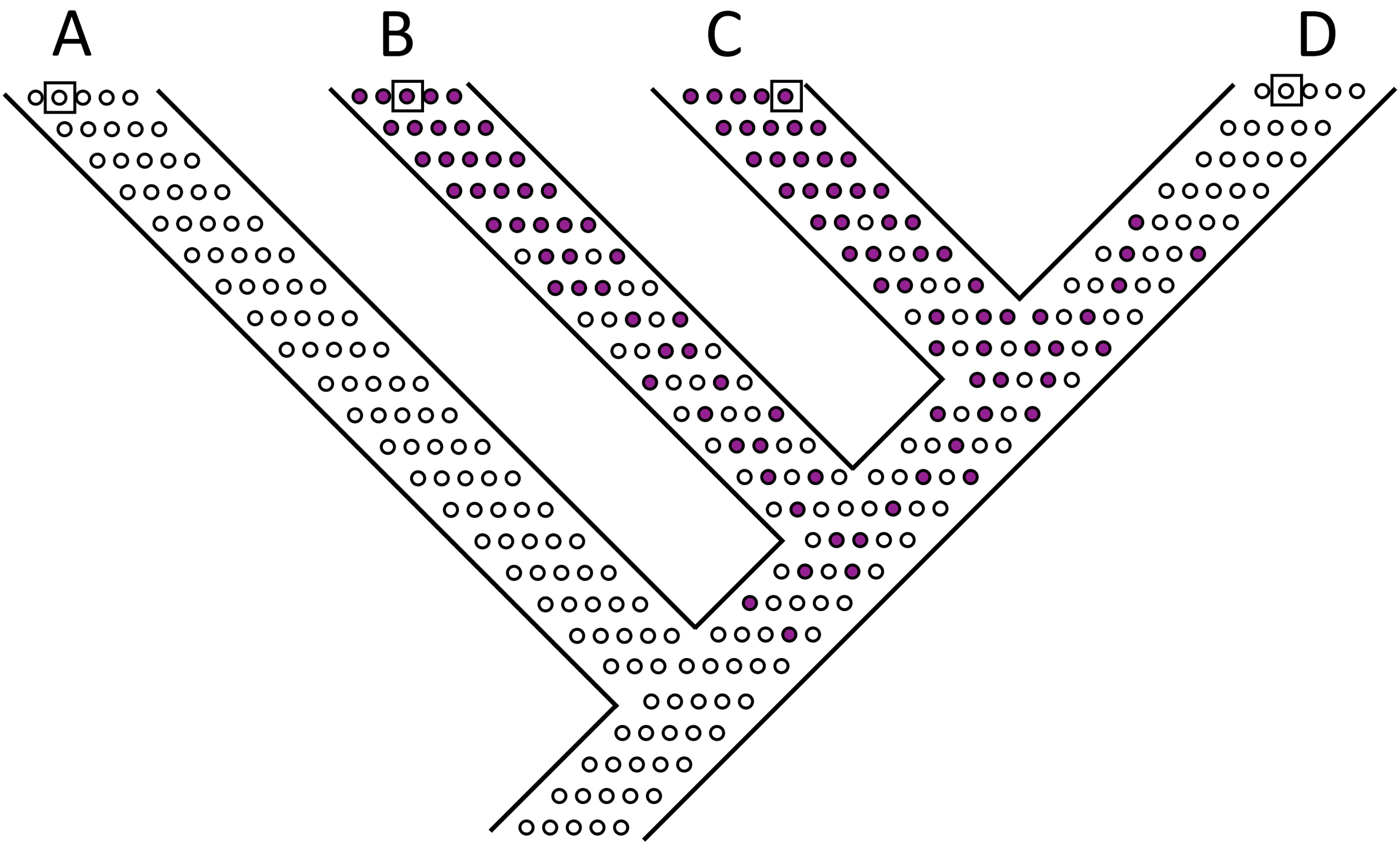
Ancestral polymorphism and species trees



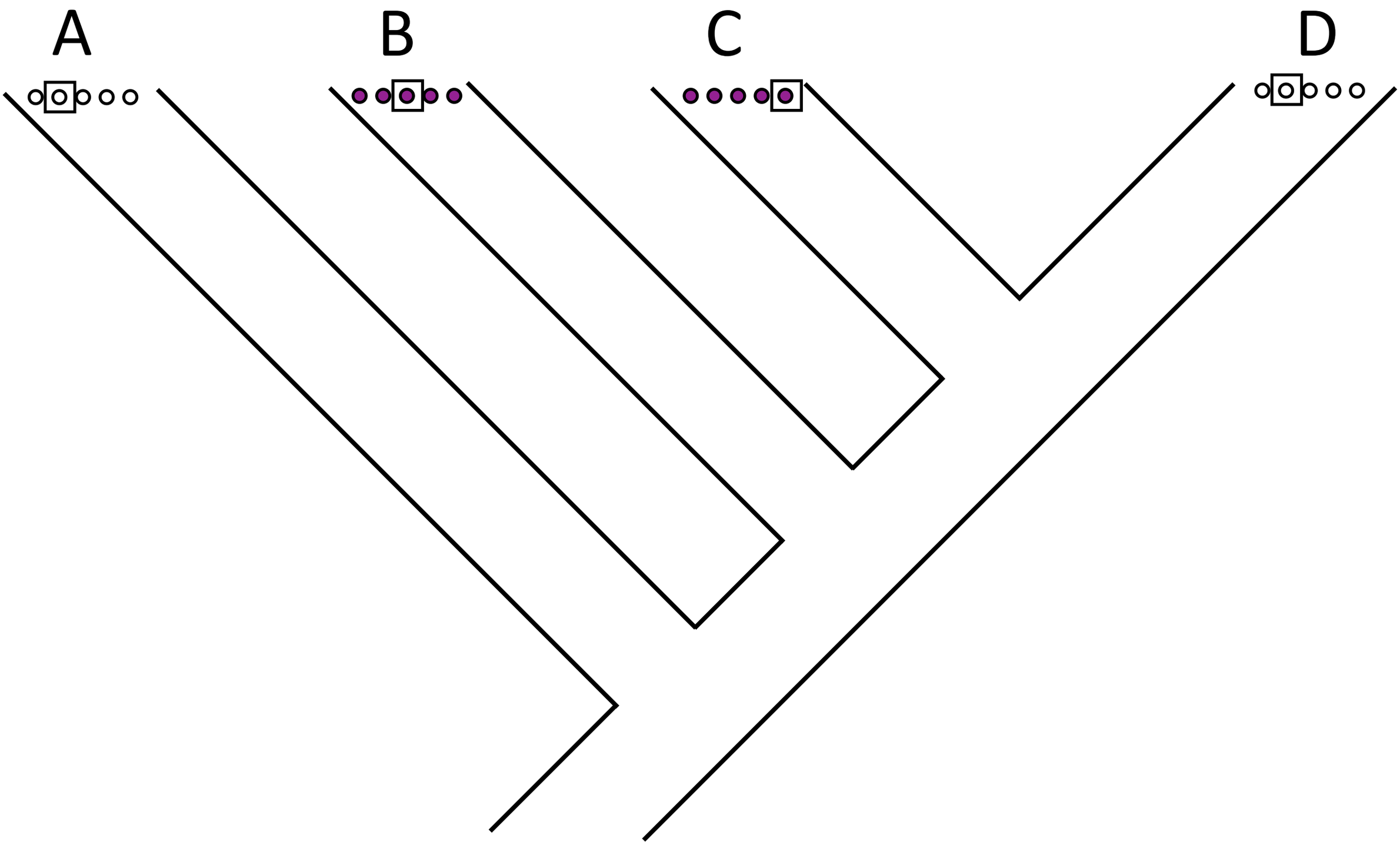
Ancestral polymorphism and species trees



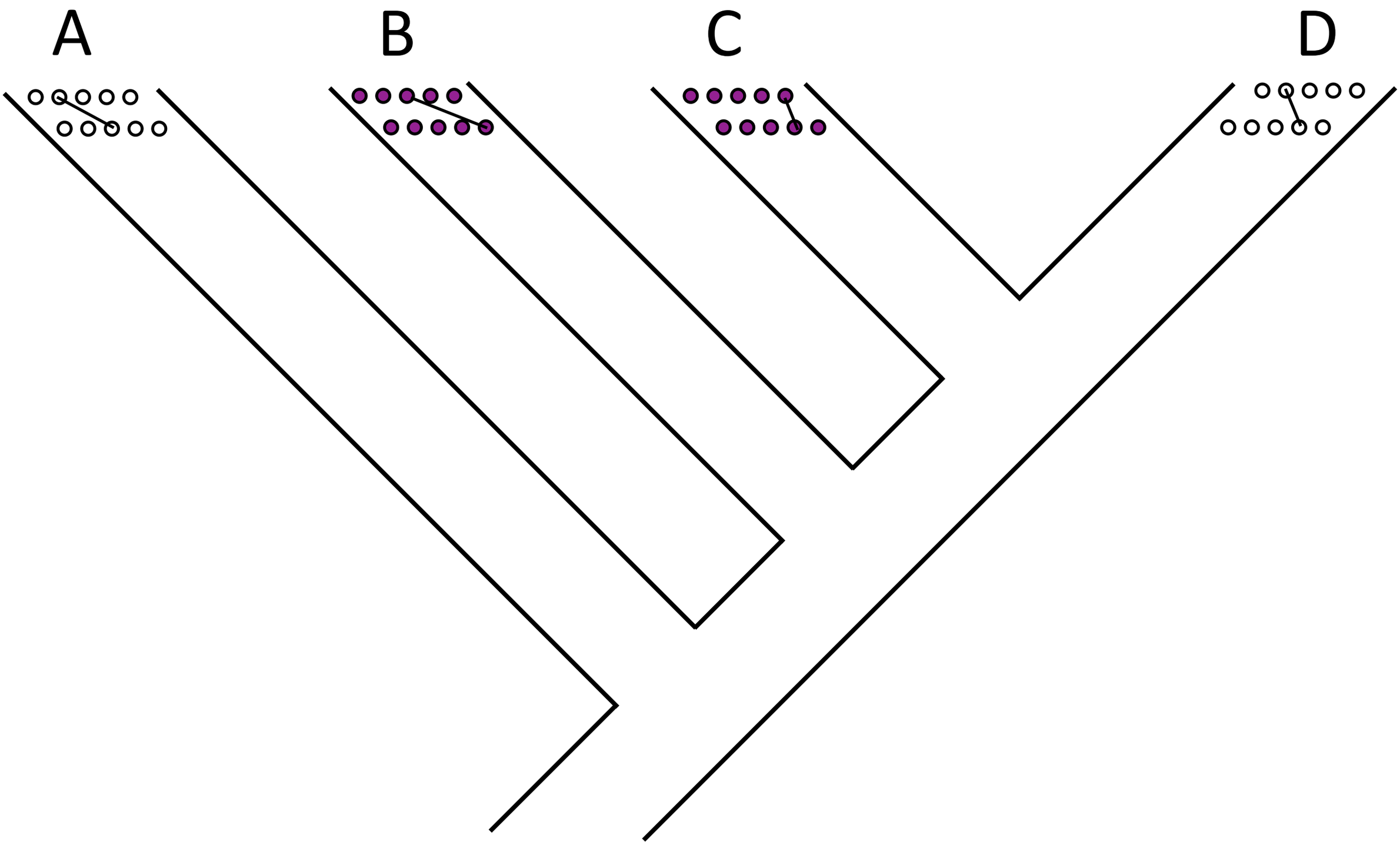
Ancestral polymorphism and species trees



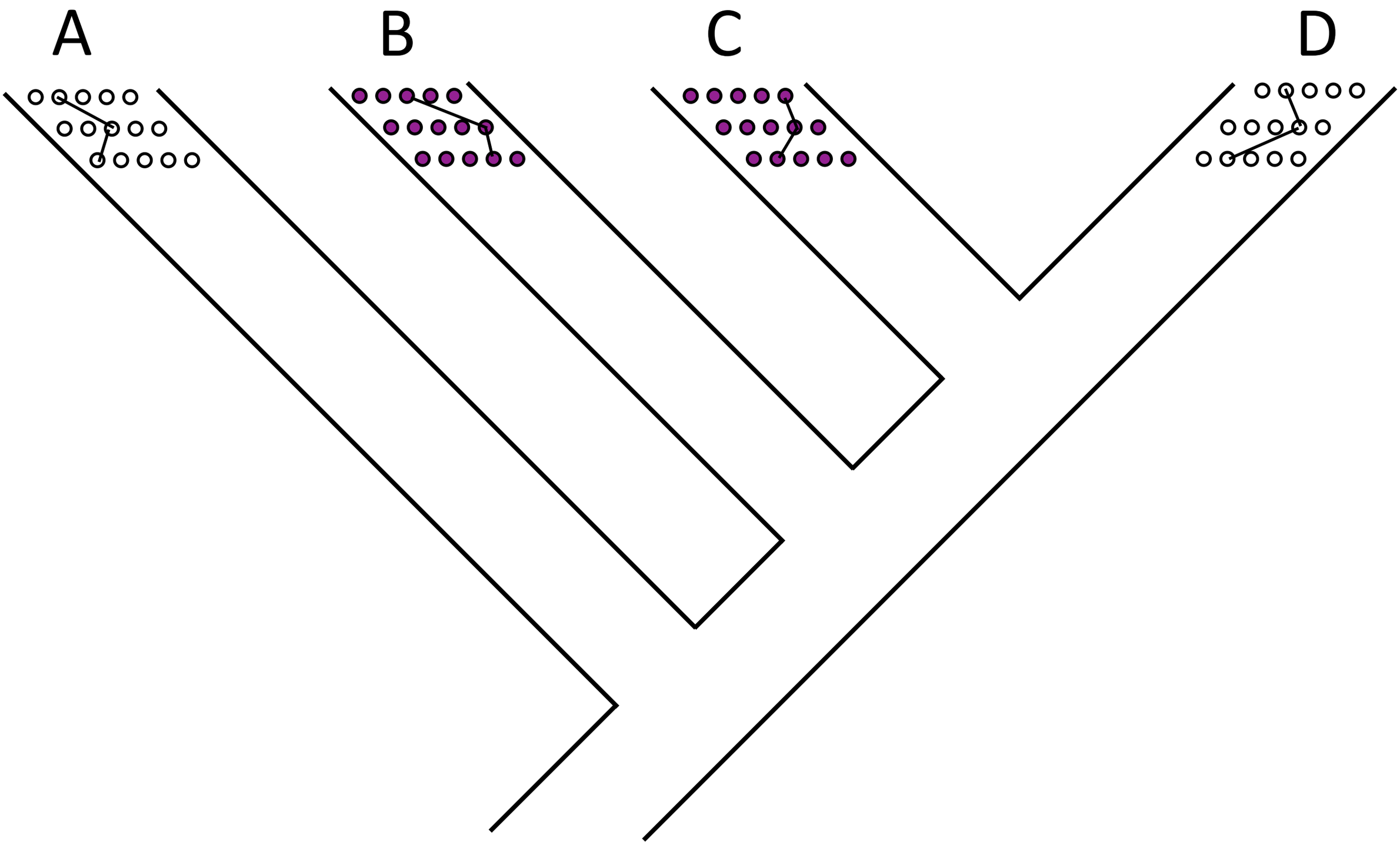
Ancestral polymorphism and species trees



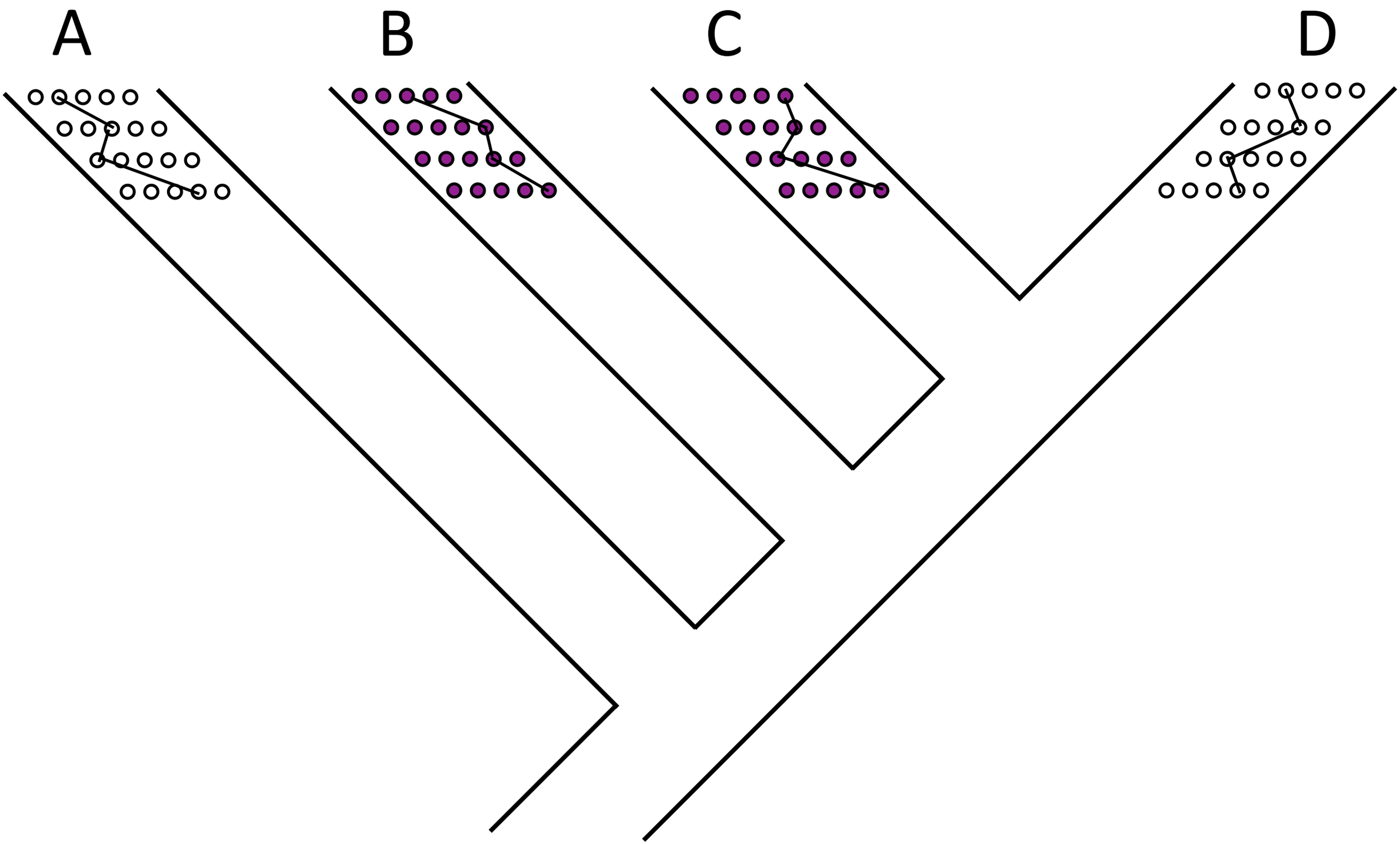
Ancestral polymorphism and species trees



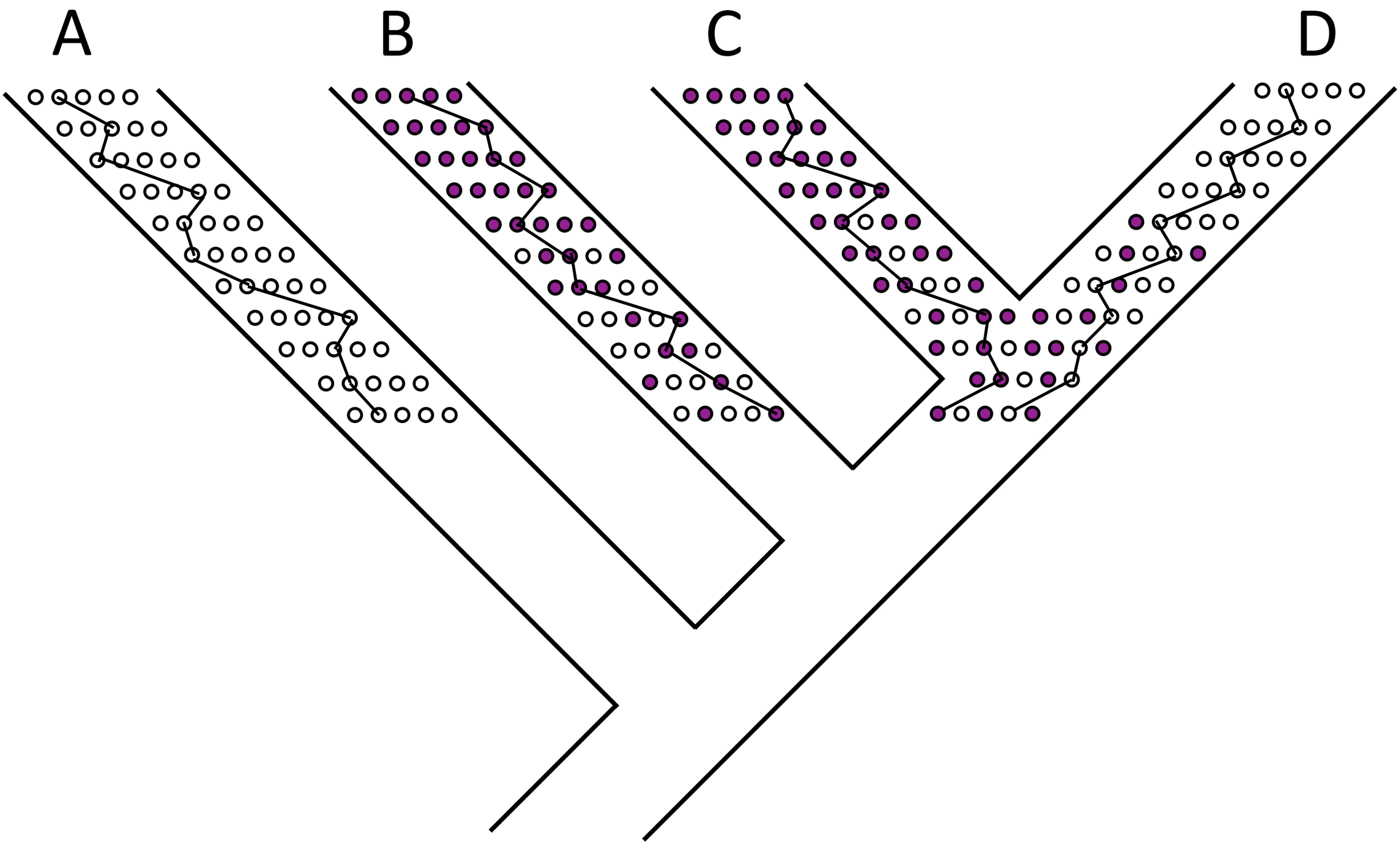
Ancestral polymorphism and species trees



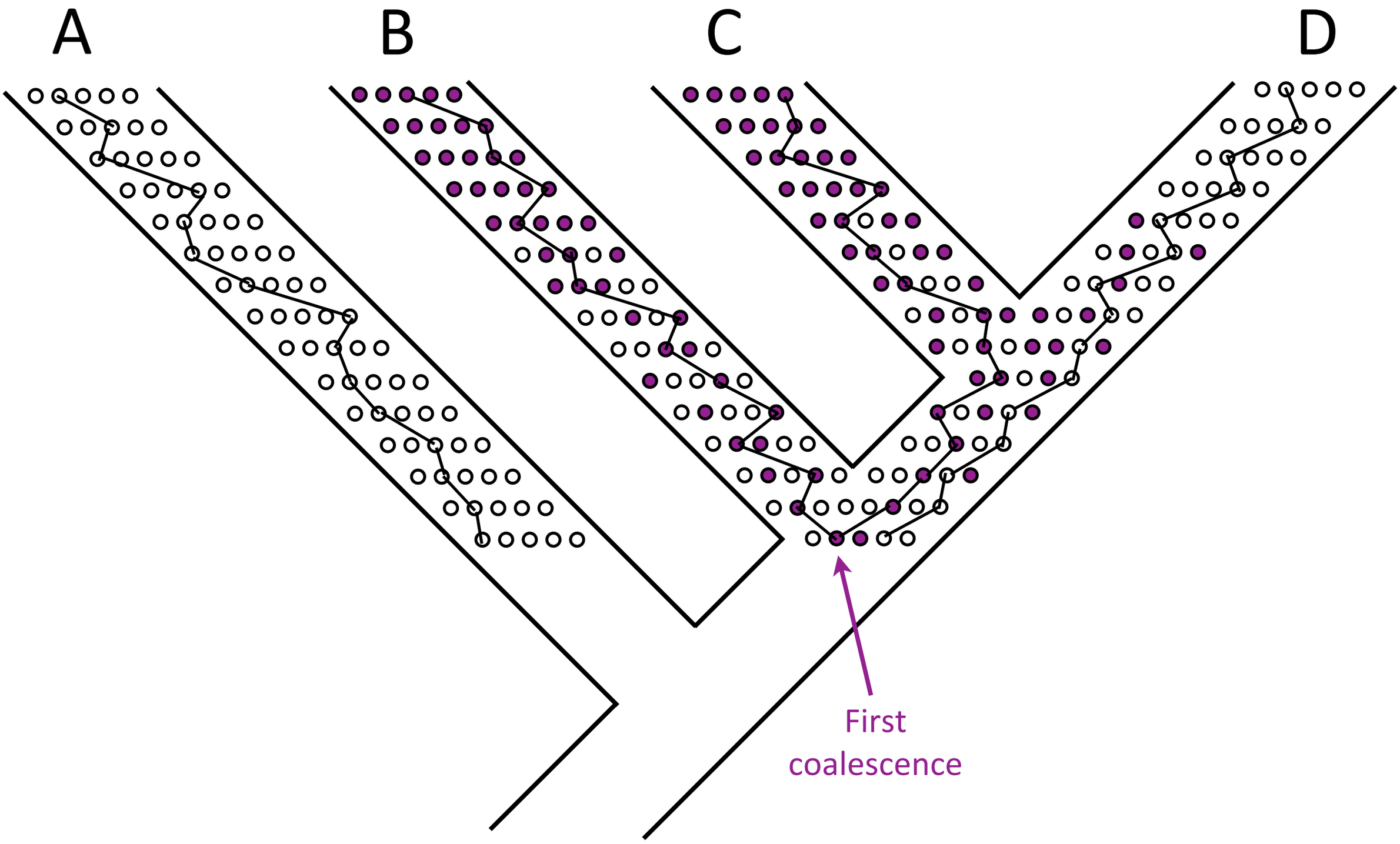
Ancestral polymorphism and species trees



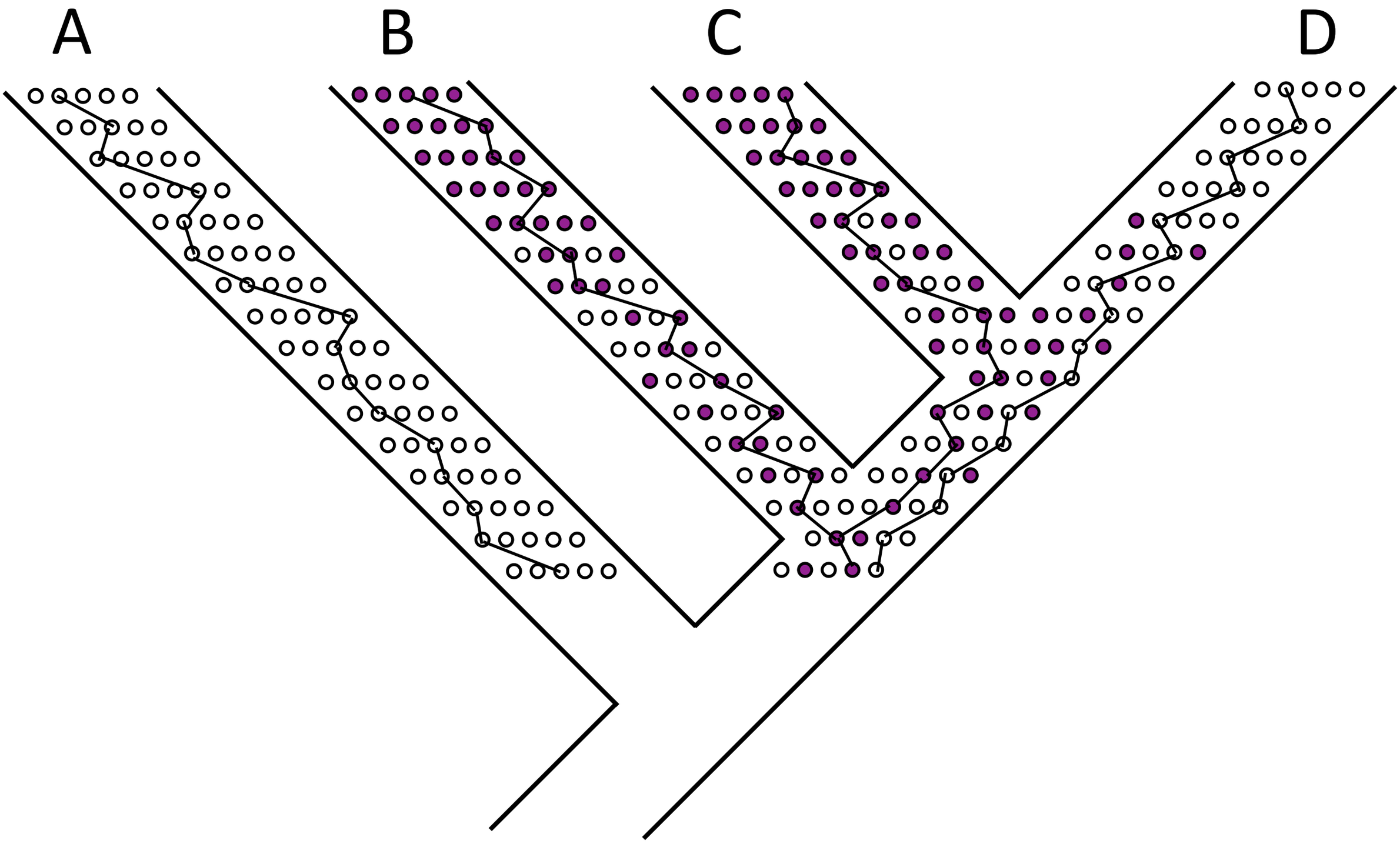
Ancestral polymorphism and species trees



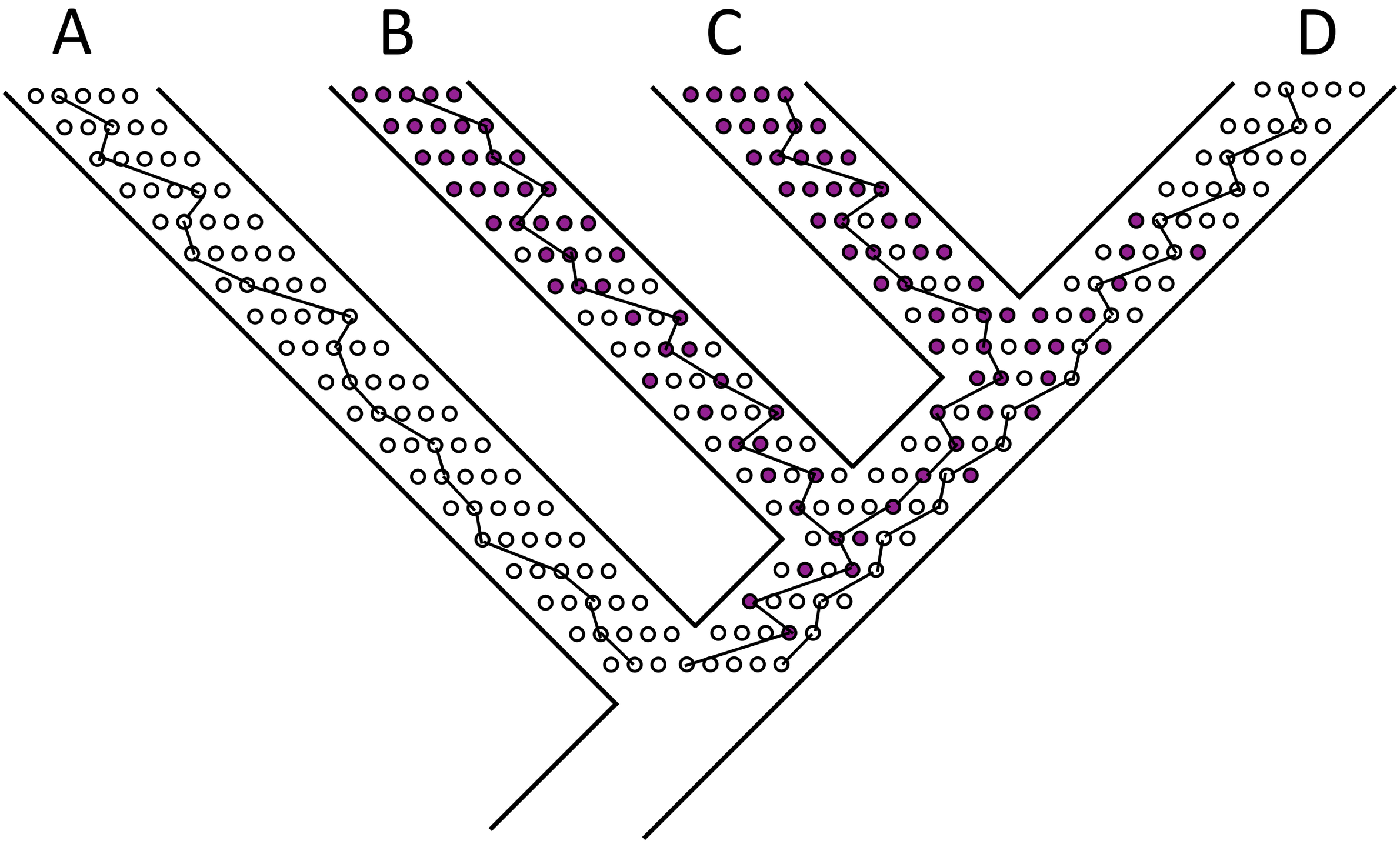
Ancestral polymorphism and species trees



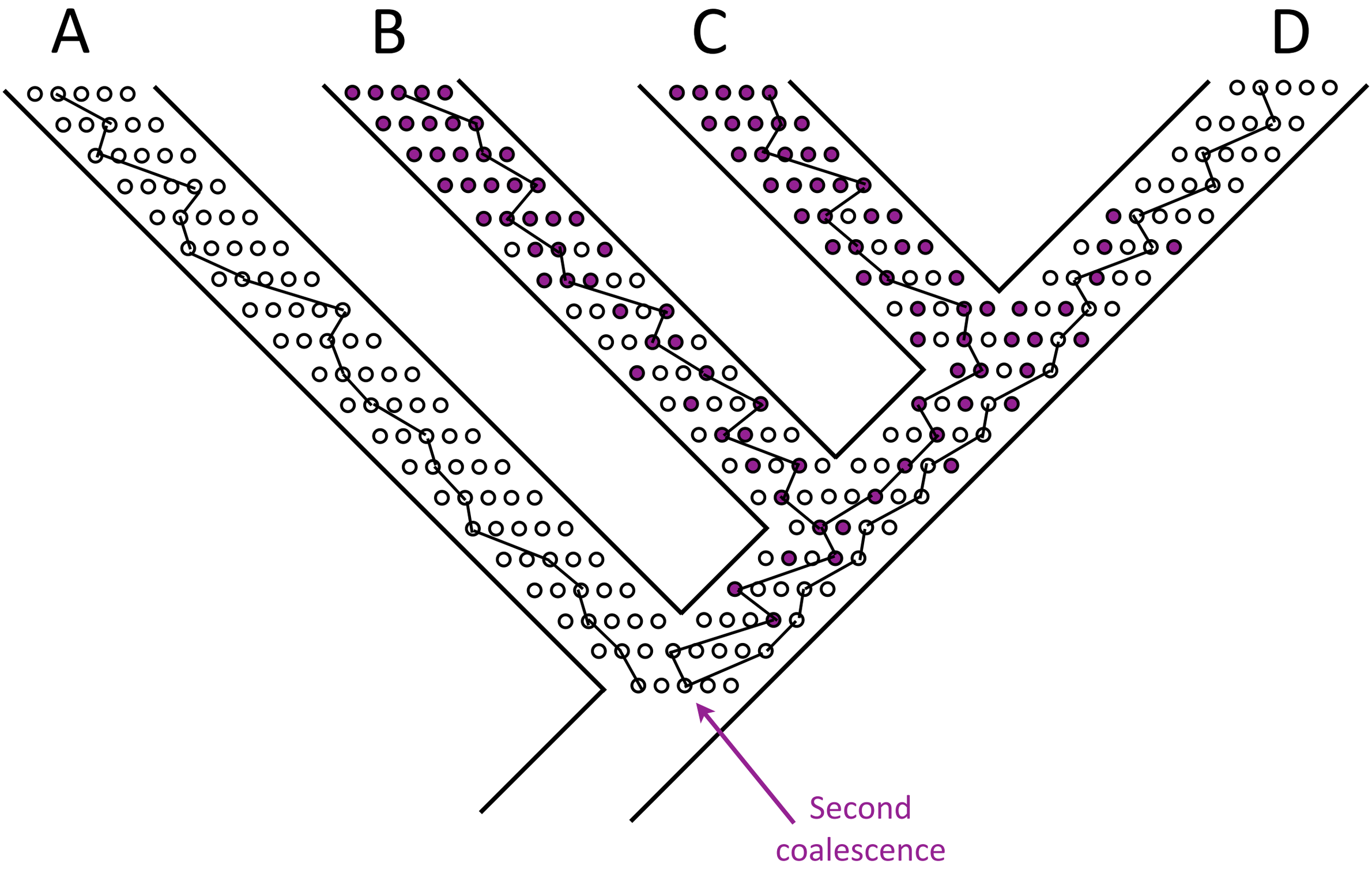
Ancestral polymorphism and species trees



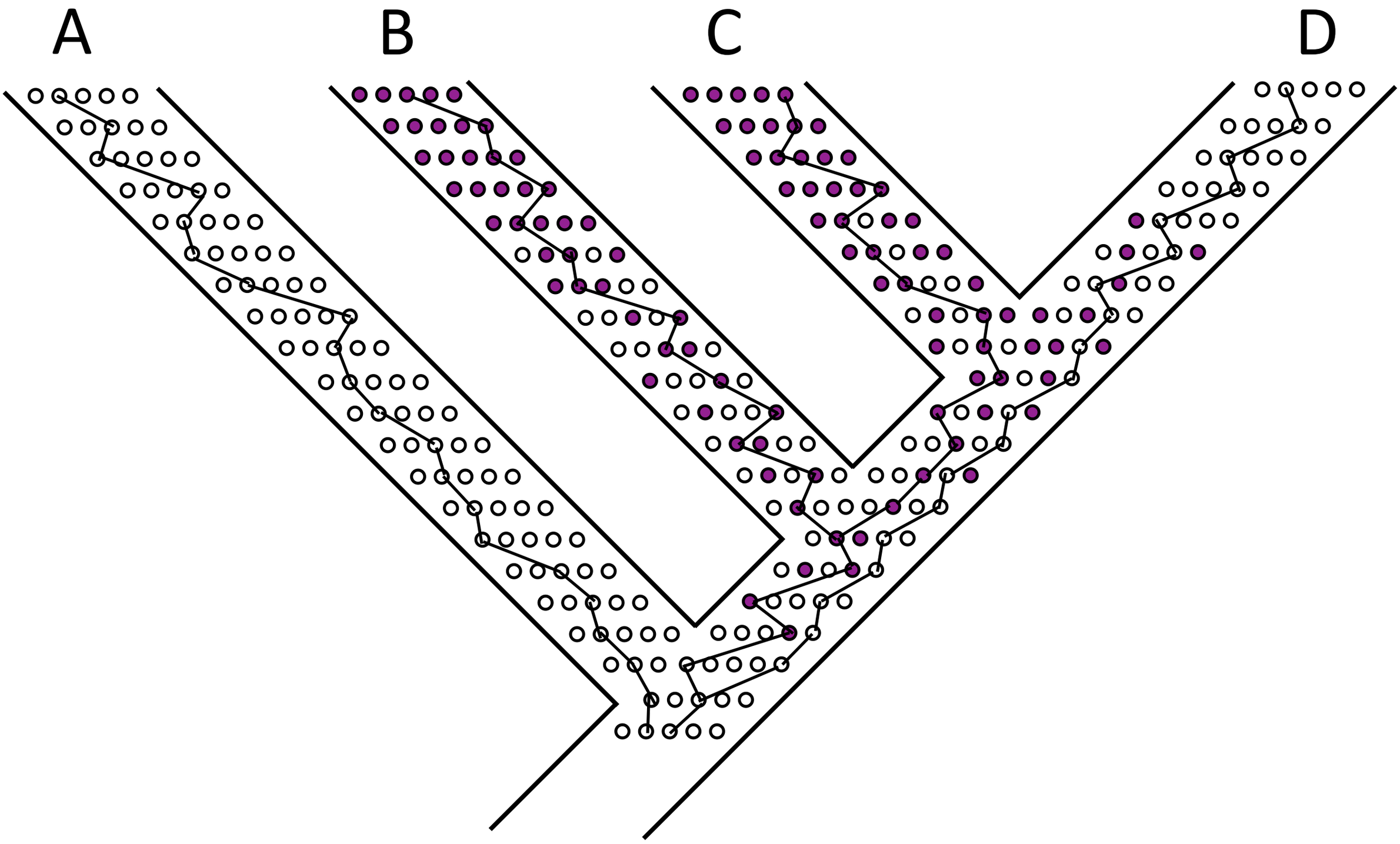
Ancestral polymorphism and species trees



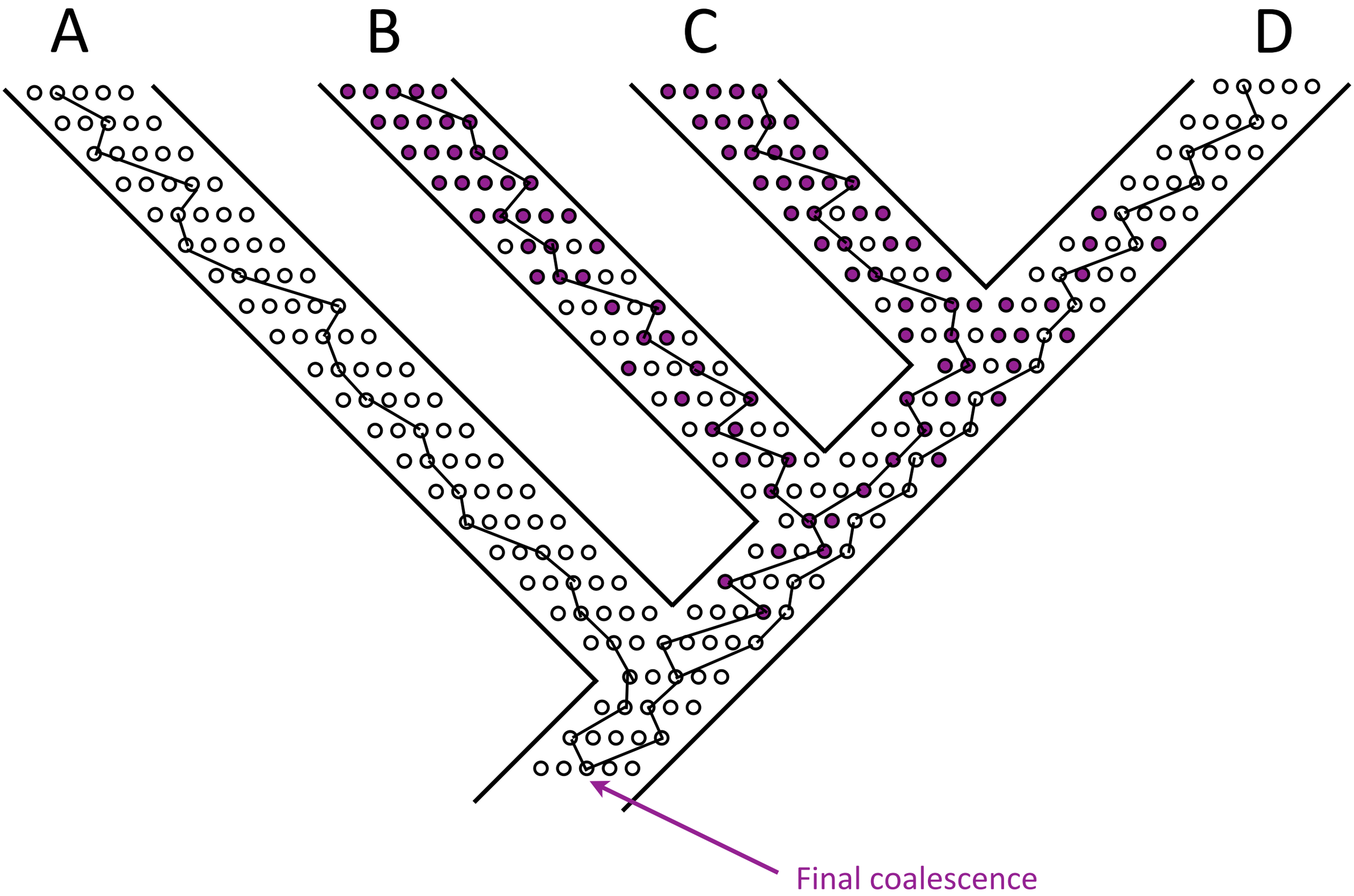
Ancestral polymorphism and species trees



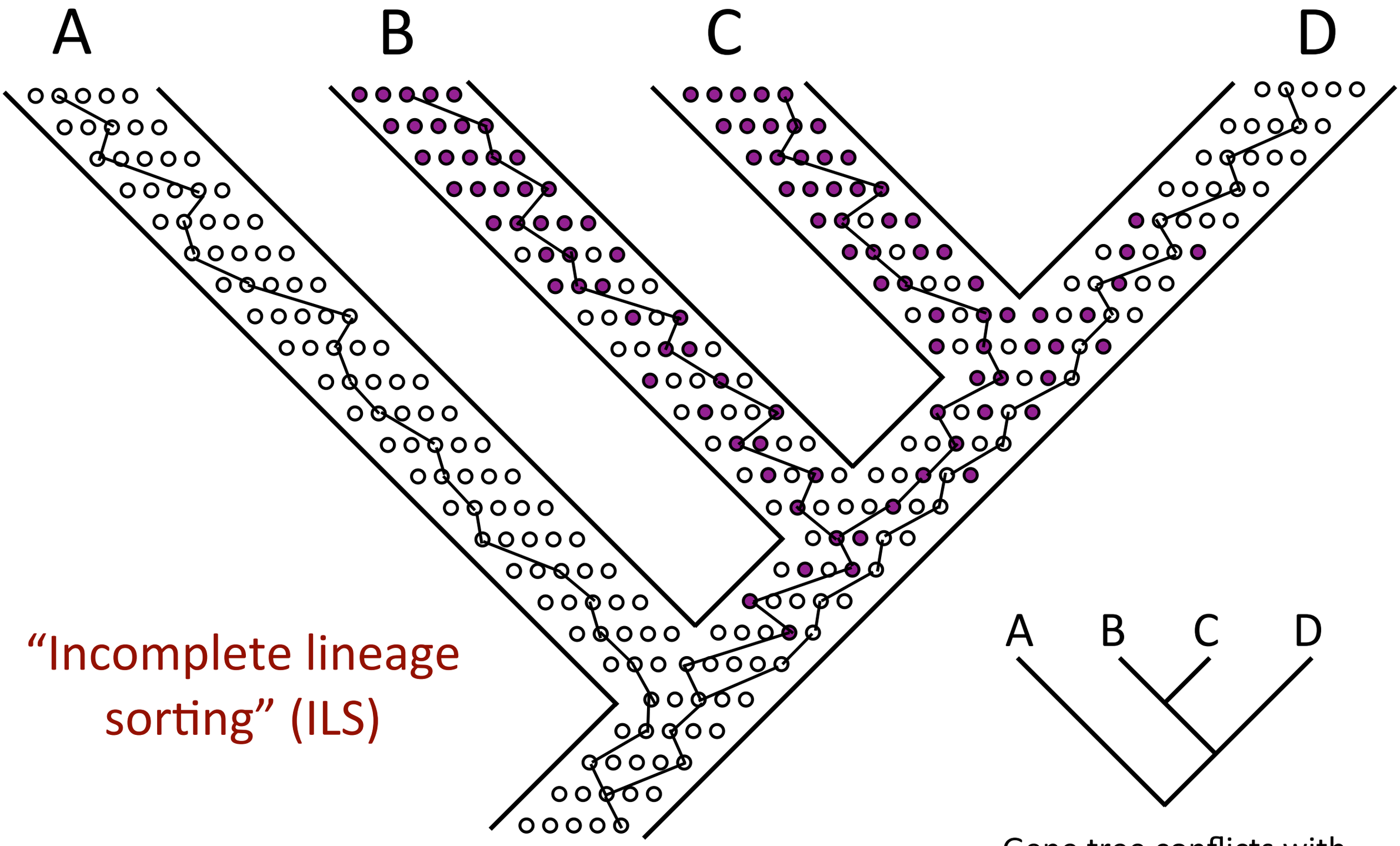
Ancestral polymorphism and species trees



Ancestral polymorphism and species trees



Ancestral polymorphism and species trees



"Incomplete lineage sorting" (ILS)

Gene tree conflicts with species tree

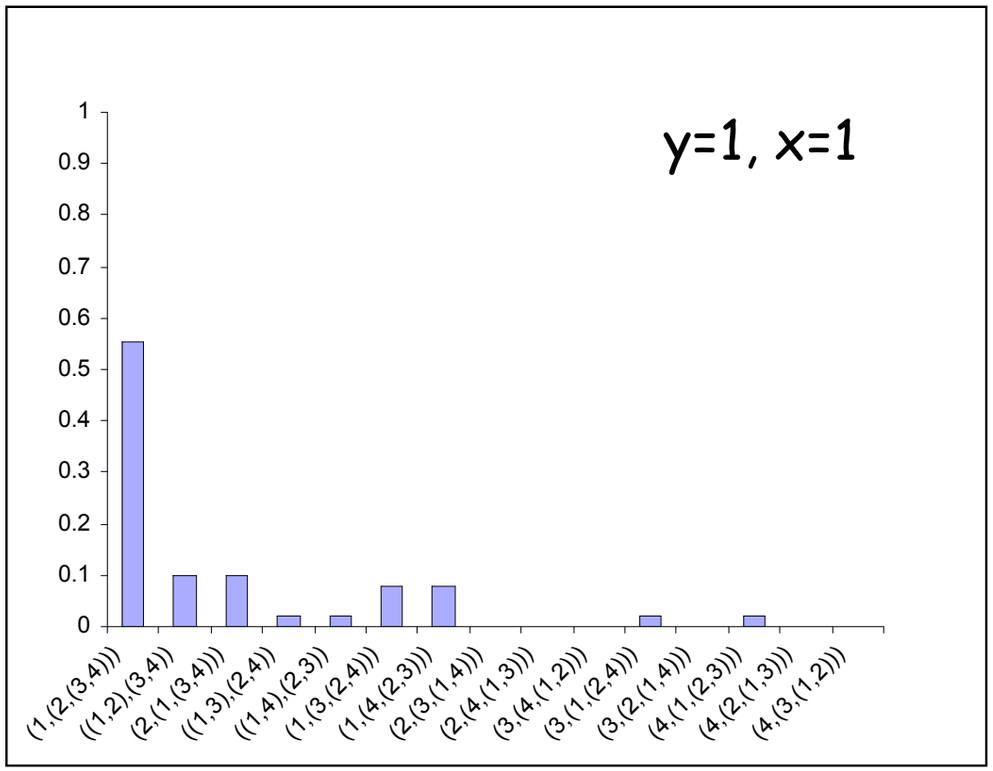
The multispecies coalescent model

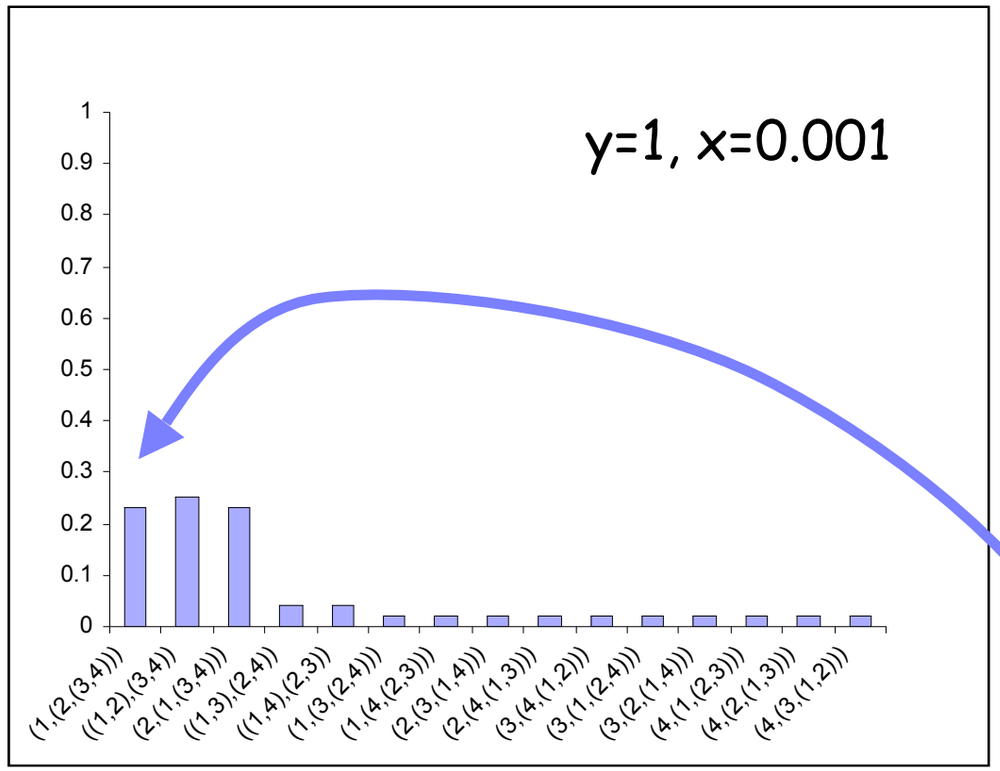
- Independence between branches—coalescent events that occur in one population are independent of what happens in other populations within the phylogeny.
- Panmixia—within a population, all pairs of lineages are equally likely to coalesce.
- Divergence is instantaneous and complete—no gene flow occurs after speciation
- ILS only—no other evolutionary processes (e.g., horizontal transfer, duplication and loss, . . .) have led to incongruence between gene trees and the species tree.
- No recombination within genes; free recombination between genes

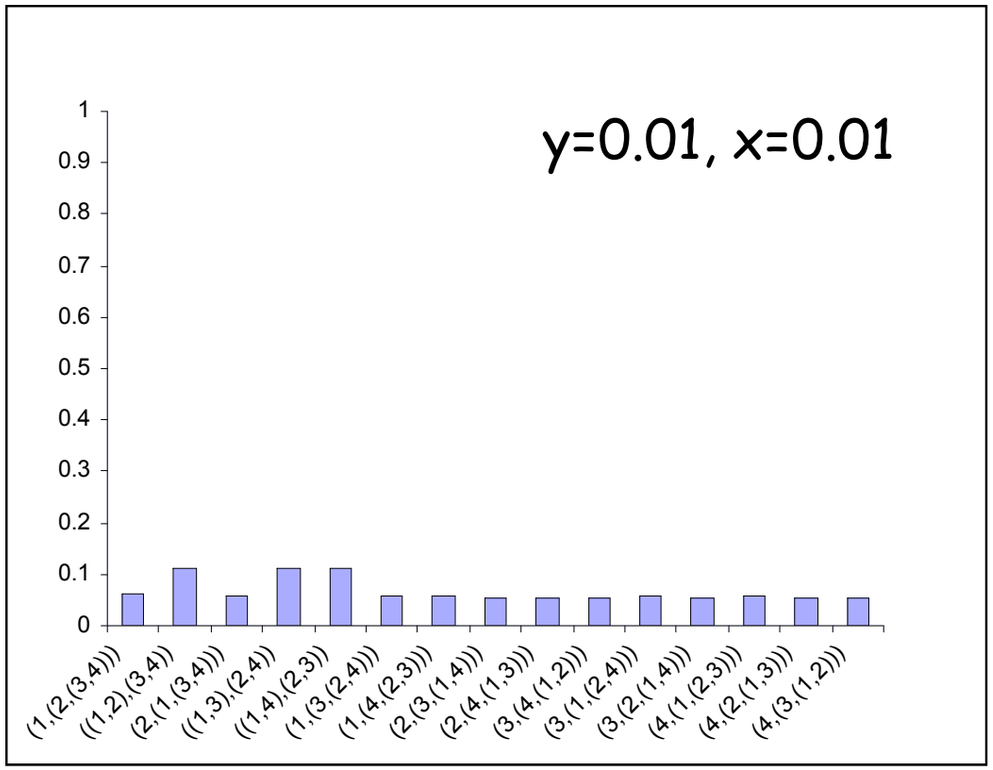
Some species tree methods that assume the multispecies coalescent process

- Full data methods
  - Fully Bayesian (integrate over gene trees within species trees, estimate posterior distribution of population sizes, branch lengths, and other model parameters in addition to the species tree)
    BEST (Liu and Pearl, 2007; Liu, 2008)
    *BEAST (Heled and Drummond, 2010)
    SNAPP (Bryant et al., 2012)
    BPP (Yang and Rannala, 2010)
  - SVDQuartets (Chifman and Kubatko, 2014, 2015)
- Summary methods (start with estimated gene trees)
  - Methods that use branch lengths:
    STEM (Kubatko et al., 2009)
    STEAC (Liu et al. 2009)
  - Methods that only use topology information
    STAR (Liu et al. 2009)
    Minimize deep coalescences ((PhyloNet; Than & Nakhleh 2009)
    MP-EST (Liu et al. 2010)
    ST-ABC (Fan and Kubatko 2011)
    STELLS (Wu 2011)
    ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015)

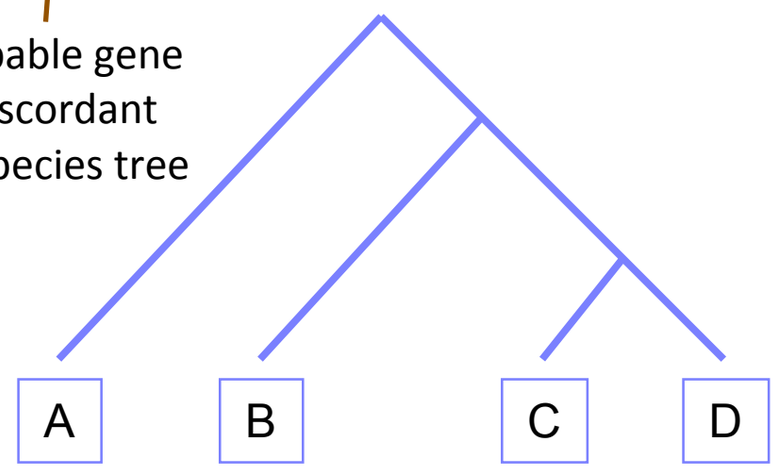
“Anomaly zones”







Most probable gene tree is discordant with the species tree

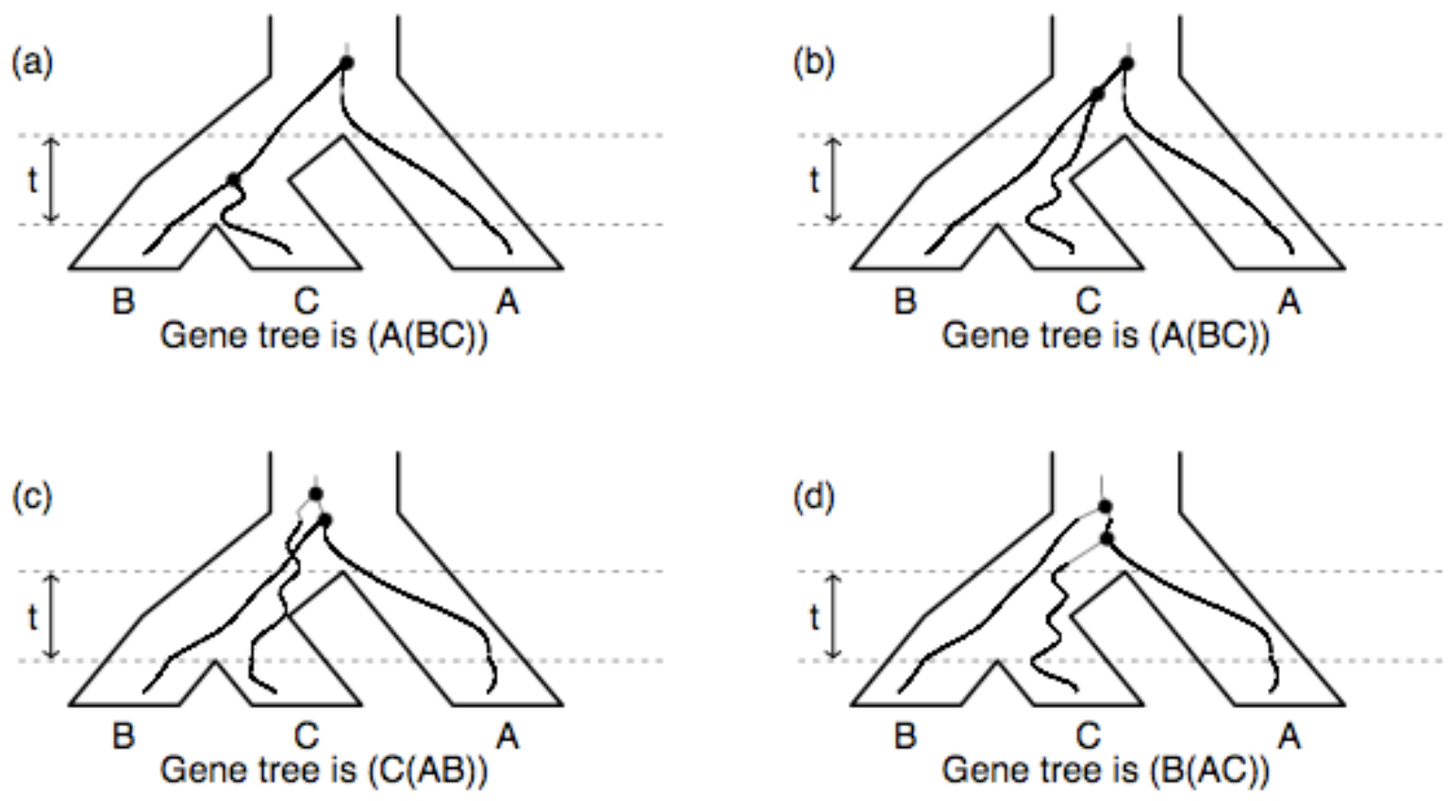


https://www.asc.ohio-state.edu/kubatko.2/SpeciesTreeEstimation2017.pdf

Handling incomplete lineage sorting

Chifman and Kubatko (2014, 2015)

When times between species divergences are short (or when population sizes are large), the history of individual genes may be discordant from the species tree topology in several (or many) ways:

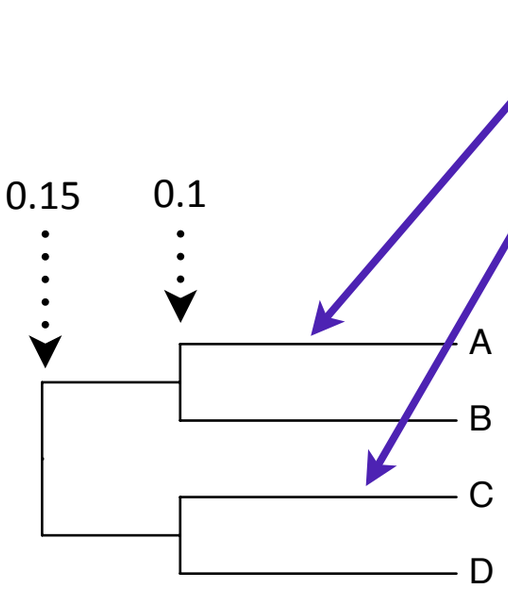


SVDQuartets: Expected rank of flattening matrix is **10** for the true species tree and **16** for the other two trees, under GTR+I+G or any of its submodels!

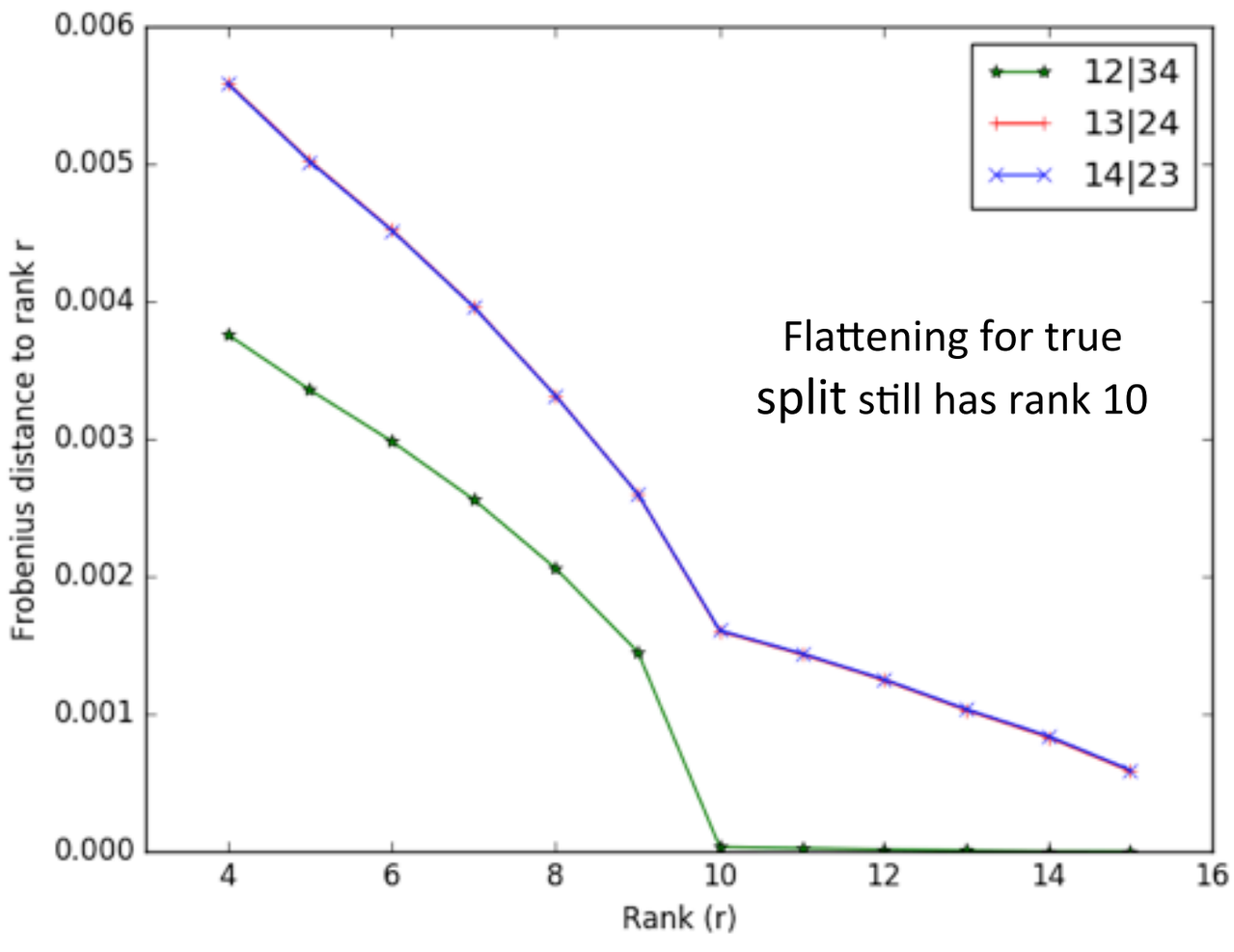
SVDQuartets

Does not need to assume a molecular clock
or constant population size

(Long and Kubatko https://arxiv.org/abs/1701.06871)

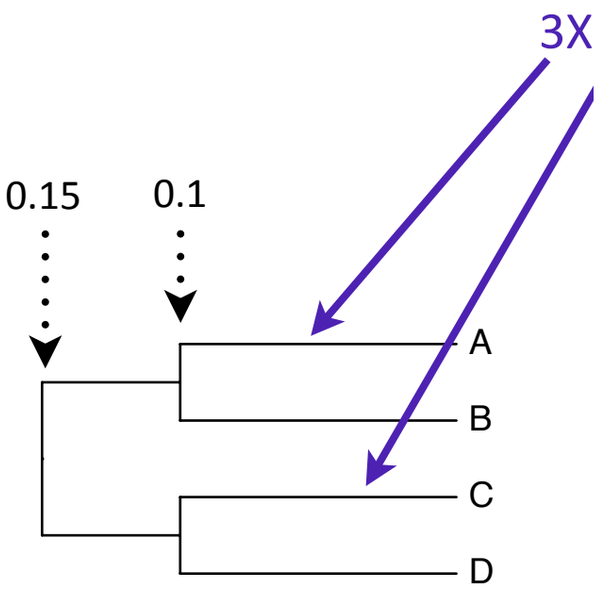


Θ=0.1
100 sites/locus
10,000,000 loci
(1,000,000,000 sites)

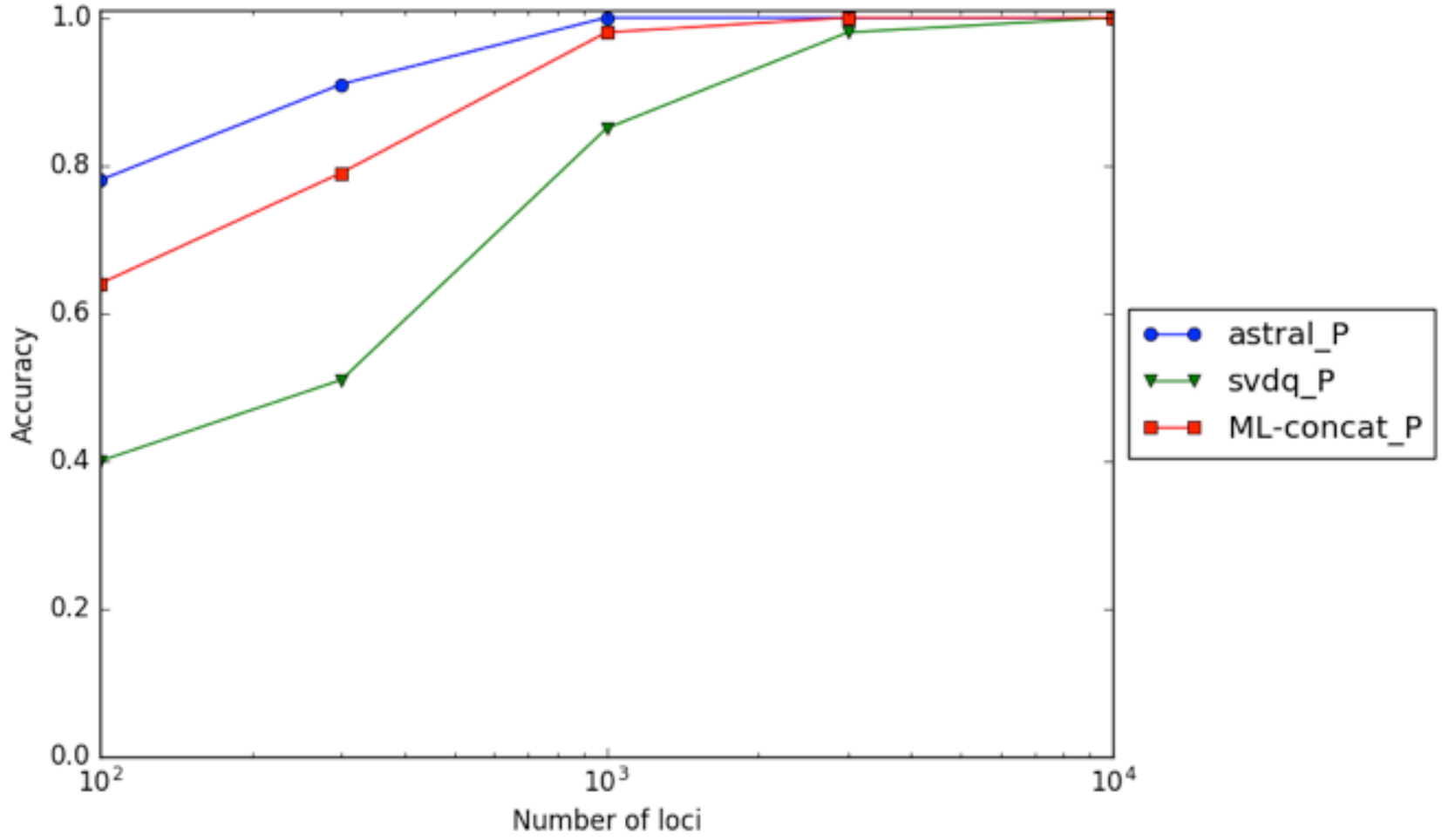


SVDQuartets

But more data may be needed to achieve the same level of accuracy (the price of generality)



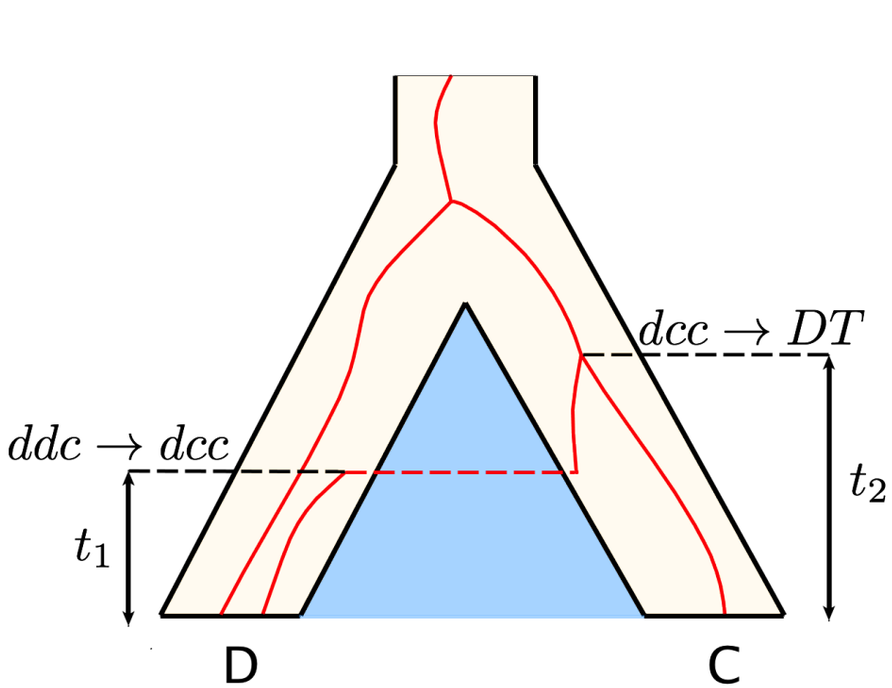
Θ=0.1
100 sites/locus
10,000,000 loci
(1,000,000,000 sites)

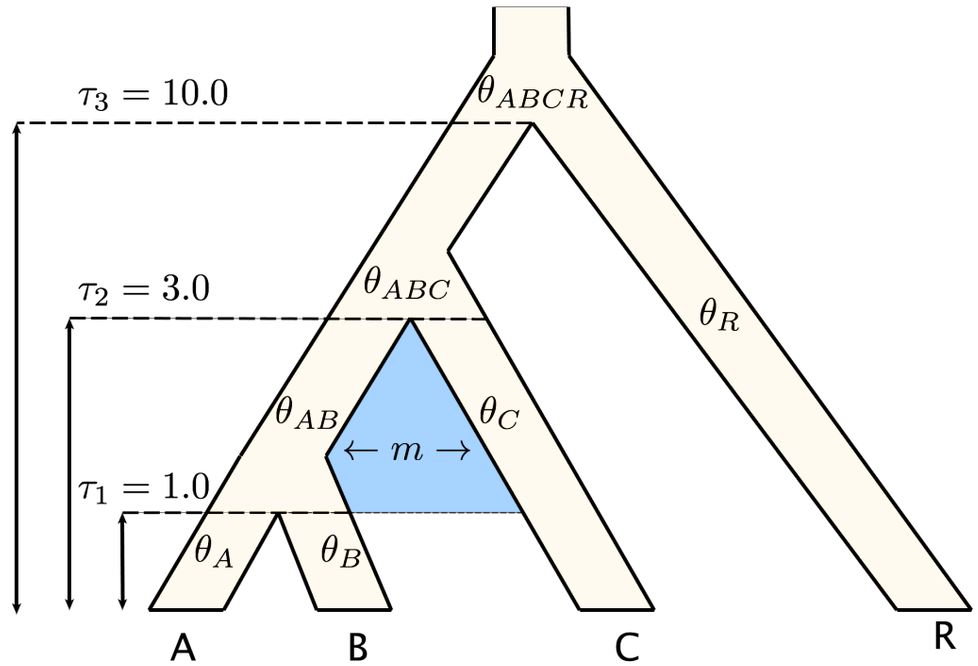


SVDQuartets

Can also handle migration between pairs of sister lineages (IM model)

Long and Kubatko 2017 https://arxiv.org/abs/1710.03806





SVDQuartets

Chifman and Kubatko (2014, 2015)

A disadvantage:

- No estimates of node ages (branch lengths) or theta parameter
- But we're working on that...

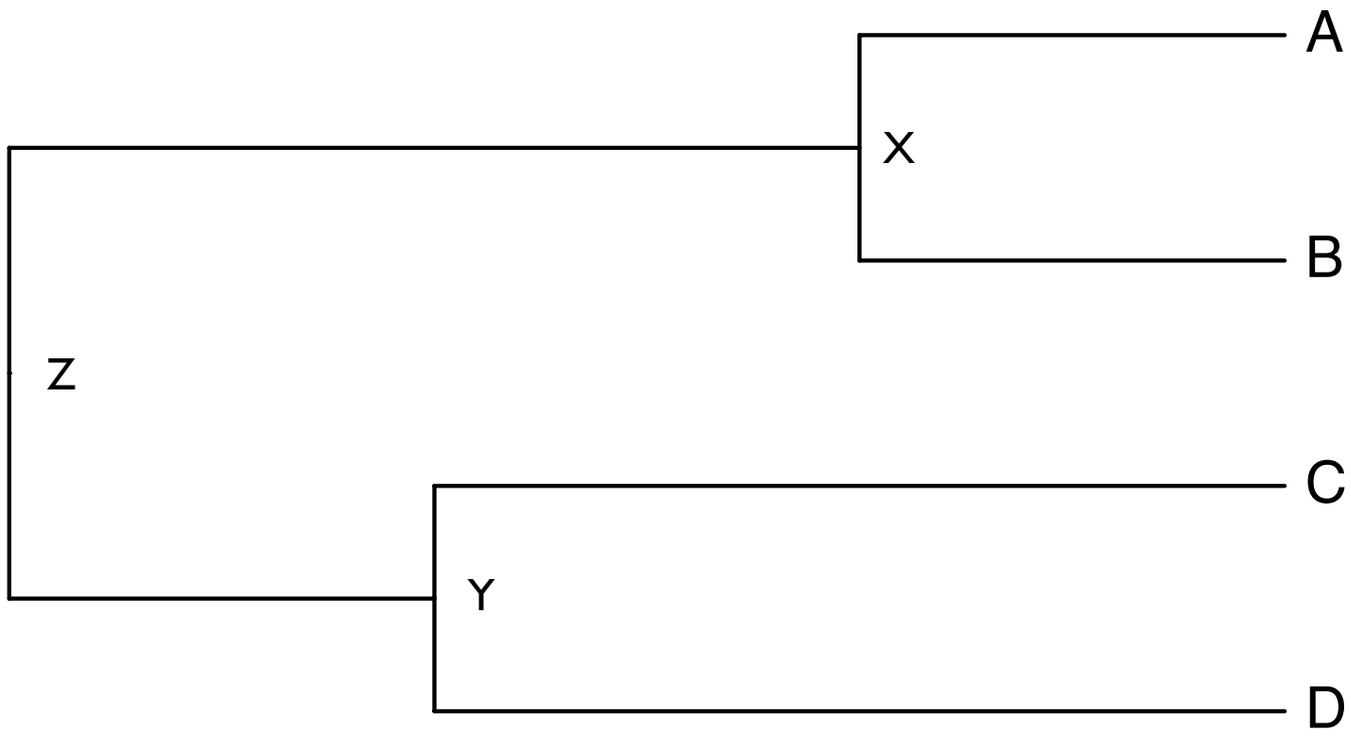
  - Can calculate (analytically) expected site pattern probabilities as a function of the species tree, node age, and theta parameters by integrating over the coalescent times and summing over all labeled histories (Chifman and Kubatko, 2015)

$$
\begin{aligned}
\sum_{x \in [\kappa]} p^*_{i_1 i_2 x i_4 i_5 \,|\, (S_5,\boldsymbol{\tau}_\star)} &= \sum_{(G,\mathbf{t}) \in \mathcal{G}_4(G,\mathbf{t})\,|\,(G_5,\mathbf{t}_\star) \in \mathcal{G}_5} \int_{\mathbf{t}_\star} p_{\sigma(i_1 i_2 i_4 i_5)\,|\,(G,\mathbf{t})} \\
&\quad \times f((G_5,\mathbf{t}_\star)\,|\,(S_5,\boldsymbol{\tau}_\star))\, d\mathbf{t}_\star \\
&= \sum_{(G,\mathbf{t}) \in \mathcal{G}_4(G,\mathbf{t})\,|\,(G_5,\mathbf{t}_\star) \in \mathcal{G}_5} \int_{\mathbf{t}} \int_{t^*} p_{\sigma(i_1 i_2 i_4 i_5)\,|\,(G,\mathbf{t})} \\
&\quad \times \left(\prod_{b \in B \setminus B^*} f_{P_b}(\mathbf{t}) \right) f_{P_{B^*}}(\mathbf{t}, t^*)\, dt^* d\mathbf{t} \\
&= \sum_{(G,\mathbf{t}) \in \mathcal{G}_4} \int_{\mathbf{t}} p_{\sigma(i_1 i_2 i_4 i_5)\,|\,(G,\mathbf{t})} \prod_{b \in B \setminus B^*} f_{P_b}(\mathbf{t}) \\
&\quad \times \left(\sum_{(G,\mathbf{t})\,|\,(G_5,\mathbf{t}_\star) \in \mathcal{G}_5} \int_{t^*} f_{P_{B^*}}(\mathbf{t}, t^*)\, dt^* \right) d\mathbf{t} \\
&= \sum_{(G,\mathbf{t}) \in \mathcal{G}_4} \int_{\mathbf{t}} p_{\sigma(i_1 i_2 i_4 i_5)\,|\,(G,\mathbf{t})} f((G,\mathbf{t})\,|\,(S,\boldsymbol{\tau}))\, d\mathbf{t} \\
&= p^*_{i_1 i_2 i_4 i_5 \,|\, (S,\boldsymbol{\tau})}.
\end{aligned}
$$

- Perform a ML optimization of node ages and theta, maximizing fit of observed to expected pattern frequencies under a multinomial model
- Recently, we have derived equations for computing the first and second order derivatives for node ages and theta, allowing estimation of sampling variance via the Fisher Information Matrix.

SVDQuartets

Chifman and Kubatko (2014, 2015)



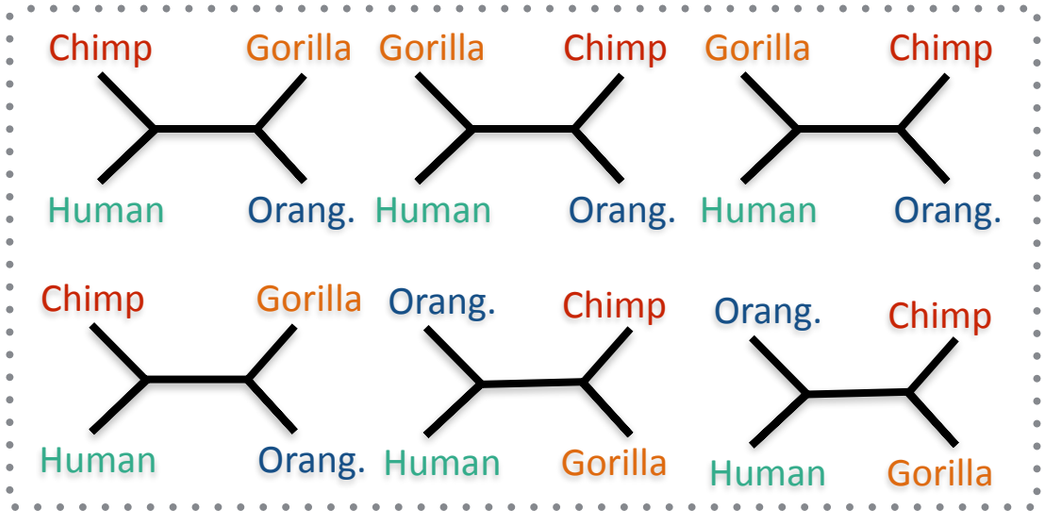
1,000 loci
sites/locus=50
theta=0.1

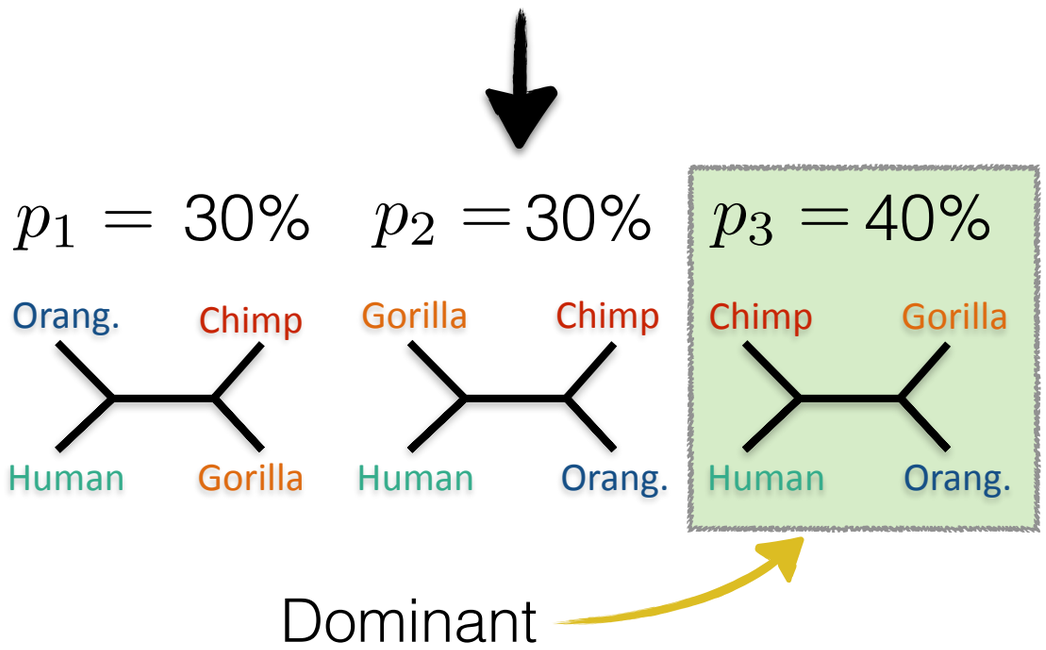
Node	True age	Estimated Age	Standard Error
X	0.5	0.5206	0.0623
Y	1.0	0.9714	0.0822
Z	1.5	1.4980	0.1024

estimated Θ = 0.101 (0.0044)

(need to test bigger trees)

ASTRAL





- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]

1. Break up input each gene tree into $\binom{n}{4}$ trees on 4 taxa (quartet trees)
2. Find all $\binom{n}{4}$ dominant quartet topologies
3. Combine dominant quartet trees

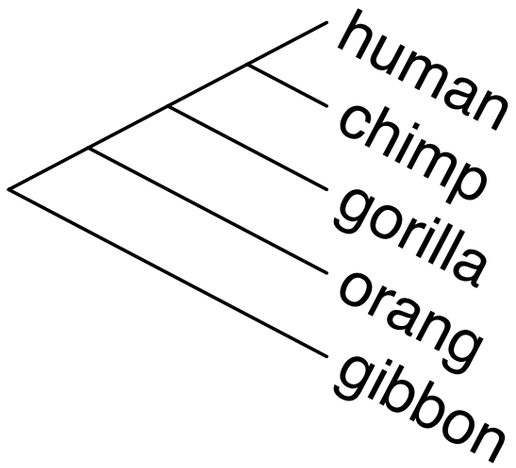
http://tandy.cs.illinois.edu/siavash-astral2-ISMB.pdf

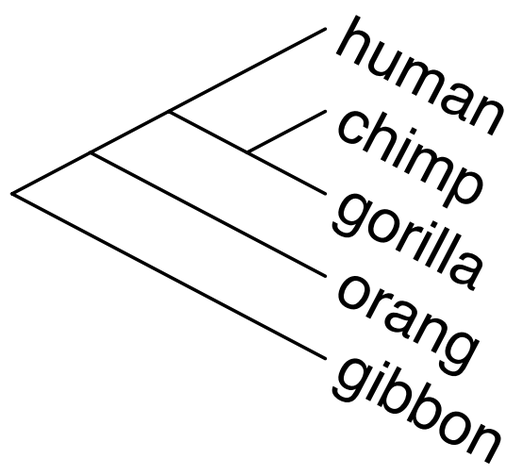
ASTRAL with >4 species

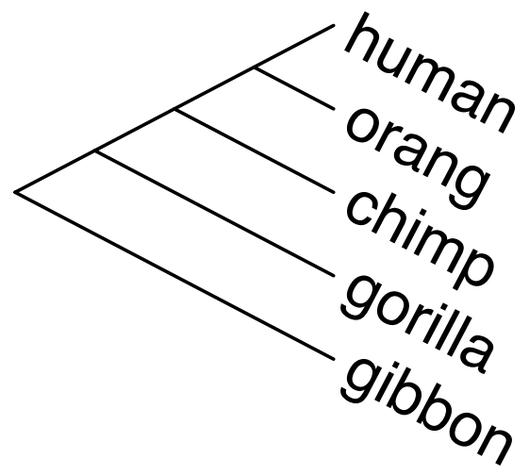
Find the species tree with the maximum number of induced quartet trees shared with a collection of gene trees

Three example gene trees and their induced quartets:

(trees are considered to be unrooted)







Number of loci with this tree	10	5	1
	(human,chimp),(gorilla,orang)	(chimp,gorilla),(human,orang)	(human,orang),(chimp,gorilla)
	(human,chimp),(gorilla,gibbon)	(human,gibbon),(chimp,gorilla)	(human,chimp),(gorilla,gibbon)
	(human,chimp),(orang,gibbon)	(human,chimp),(orang,gibbon)	(human,orang),(chimp,gibbon)
	(human,gorilla),(orang,gibbon)	(human,gorilla),(orang,gibbon)	(human,orang),(gorilla,gibbon)
	(chimp,gorilla),(orang,gibbon)	(chimp,gorilla),(orang,gibbon)	(chimp,orang),(gorilla,gibbon)

ASTRAL with >4 species

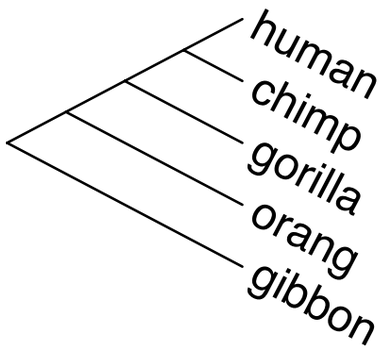
Compute the total number of times each resolved quartet was found over all input gene trees

Resolved quartet	Induced by tree(s)	Weight
(human,chimp),(gorilla,orang)	1	10
(human,chimp),(gorilla,gibbon)	1,3	10+1=11
(human,chimp),(orang,gibbon)	1,2	10+5=15
(human,gorilla),(orang,gibbon)	1,2	10+5=15
(chimp,gorilla),(orang,gibbon)	1,2	10+5=15
(chimp,gorilla),(human,orang)	2	5
(human,gibbon),(chimp,gorilla)	2	5
(human,orang),(chimp,gorilla)	3	1
(human,orang),(chimp,gibbon)	3	1
(human,orang),(gorilla,gibbon)	3	1
(chimp,orang),(gorilla,gibbon)	3	1

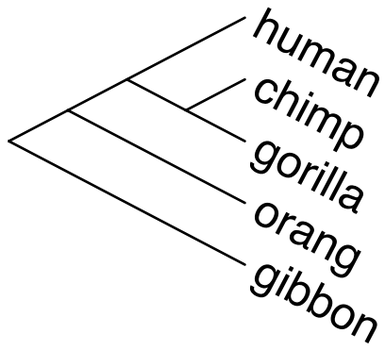
ASTRAL with >4 species

Find the species tree that maximizes the number of consistent quartets (*a la* SVDQuartets)

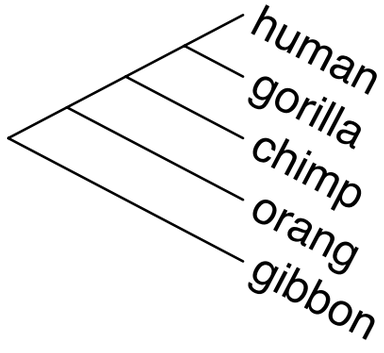
For each species tree evaluated, we sum the weights of all satisfied quartets



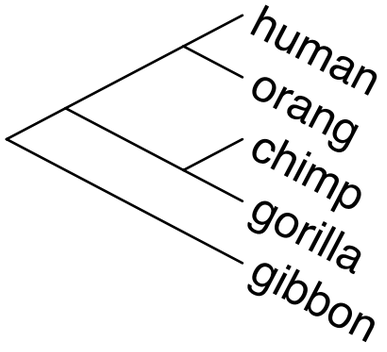
score=10+11+15+15+15=**66**
(best tree)



score=1+5+15+15+15=**51**



score=0+0+15+15+15=**45**



score=1+1+1+5+15=**23**

Could also evaluate 11 more species trees (have to search over all unrooted species trees, including trees that never appeared as a gene tree)

ASTRAL

- Where do the gene trees come from?

  That's your problem! Typically, people run RAxML or IQ-TREE to estimate gene trees. ASTRAL is a very fast method once you have the gene trees, but the gene-tree estimation typically dominates the total run time.

- ASTRAL makes a consistent estimate of the species tree, as long as the input gene trees themselves are estimated using a consistent method. If the gene trees estimates are biased, there is no guarantee of consistency.

- Astral provides exact and heuristic algorithms for the MQC tree search. The exact method will be too slow if there are very many tips.

- Download at: https://github.com/smirarab/ASTRAL

Selected References

Allman E.S., Rhodes J.A. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. Math Biosci. 186:113–144.

Allman E.S., Rhodes J.A. 2004. Quartets and parameter recovery for the general Markov model of sequence mutation. Appl Math Res Express. 2004:107–131.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. MBE. 29:1917–1932.

Chifman J., Kubatko L.S. 2014. Quartet inference from SNP data under the coalescent model. 30:3317–3324.

Chifman J., Kubatko L.S. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J Theor Biol. 374:35–47.

Eriksson N. 2005. Tree Construction using Singular Value Decomposition. In: Pachter L., Sturmfels B., editors. Algebraic Statistics for Computational Biology. Cambridge University Press. p. 347–358.

Fernández-Sánchez J., Casanellas M. 2016. Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. Syst Biol. 65:280–291.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. MBE. 27:570–580.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56:504–514.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T.J. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. 30:i541–8

Mirarab S., Warnow T.J. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. 31:i44–52.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. PNAS. 107:9264–9269.

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. In: Meidanis J., Nakhleh L. (eds) Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science, vol 10562. Springer.