Research: You Might Get STD Through Swimming in the Arctic Ocean, Here's How

By Jamie P.

Updated: Mar 09 2020, 03:24 AM EDT

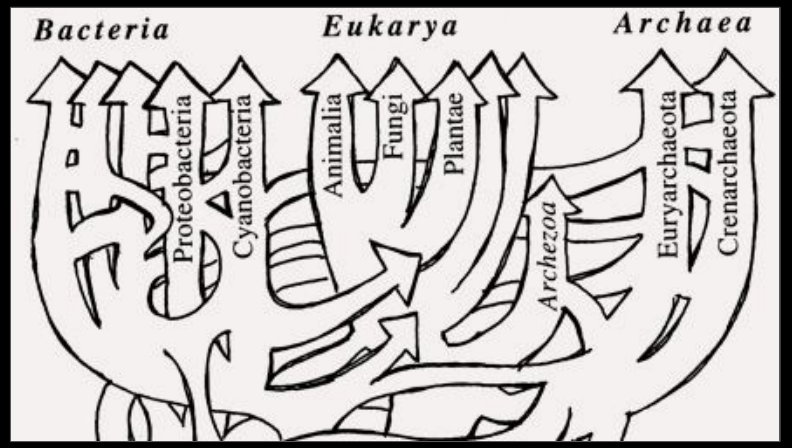
Sexually-transmitted diseases or STDs can now be transferred to a person even without having sexual intercourse. Worse, the research found out that it can also be acquired through simply swimming in an ocean. How can this be possible?

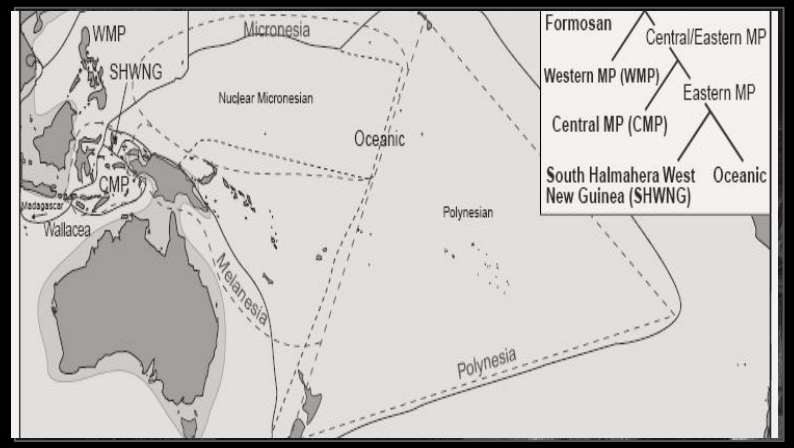
How do you get STD? By swimming in the Arctic ocean, apparently

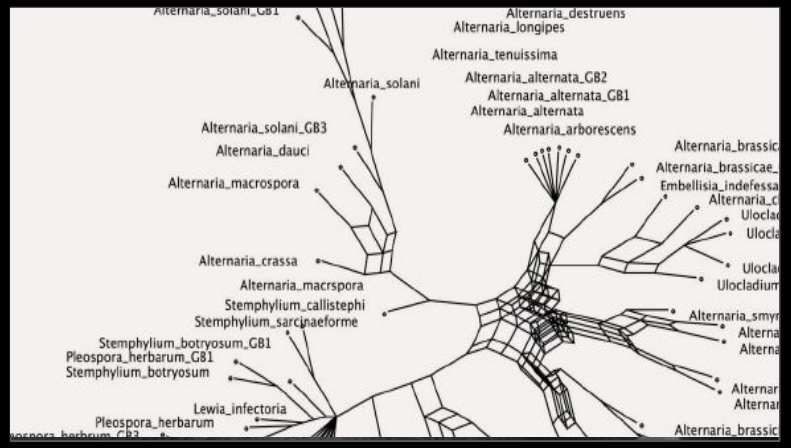
Scientists discover new chlamydia cousin deep under Arctic Ocean

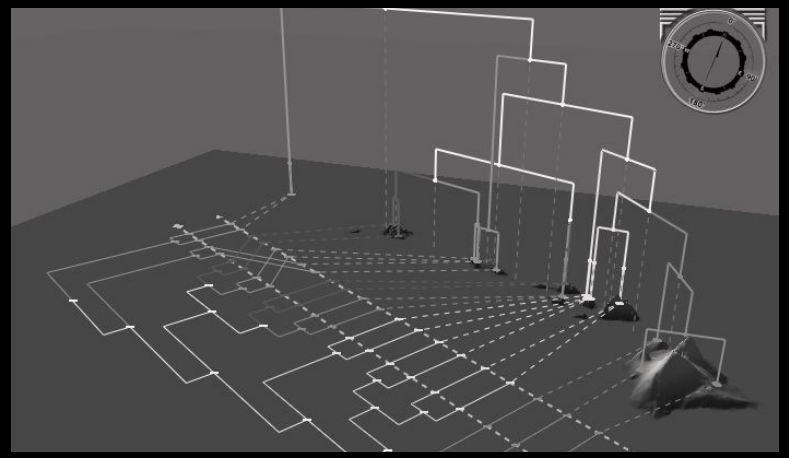
'You don't have to worry about swimming in the ocean,' researcher says

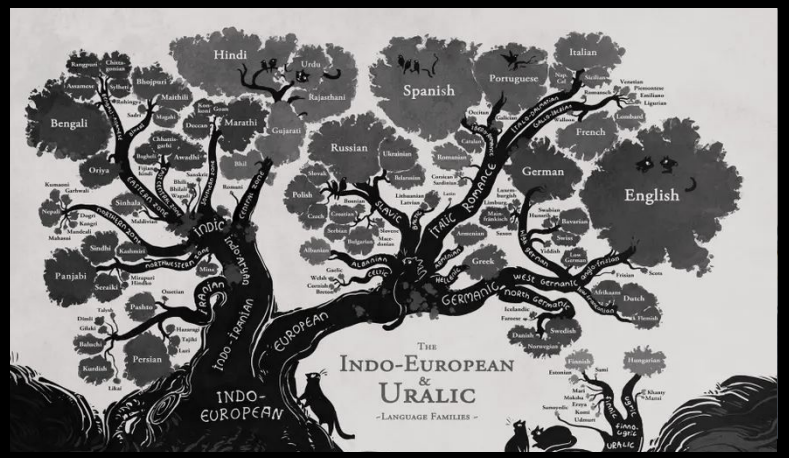
Emily Blake · CBC News · Posted: Mar 19, 2020 7:00 AM ADT | Last Updated: March 19, 2020

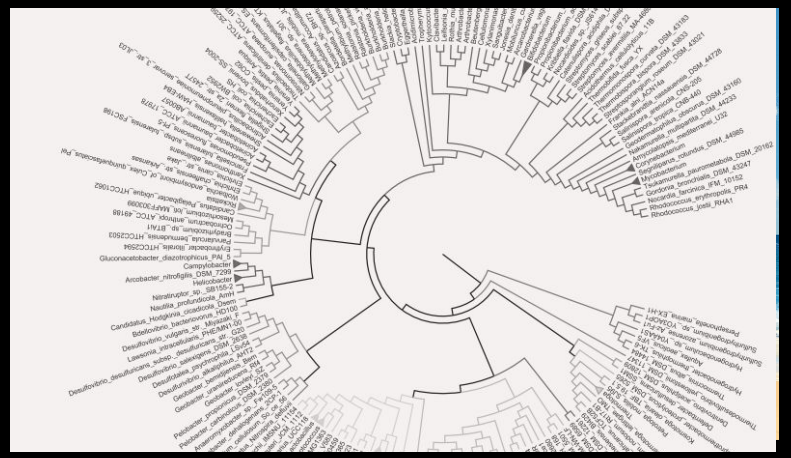












Module 3

Phylogenetics

What is Phylogenetics?

- The study of evolutionary history and relationships among living (and extinct) biological entities
- We use trees (which are ultimately just special kinds of directed / undirected graphs) and similar structures to represent this
- Phylogenetics and other comparative approaches are key to understanding evolution, function, ecology, as well as the next pandemic

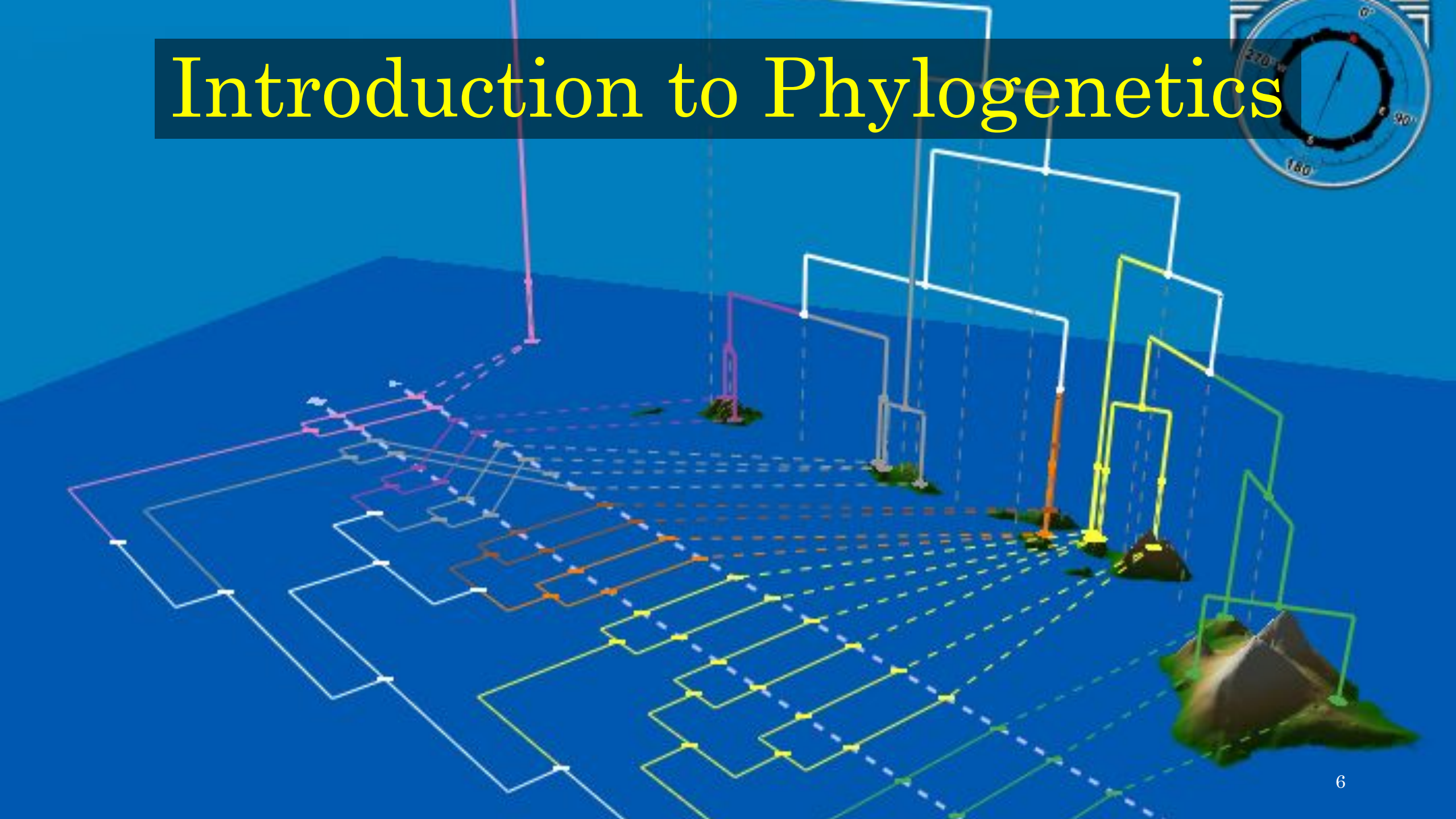
Key Questions

- Processes of evolution:
  - What assumptions do we make about how evolution works?
  - How do we implement these in practice?
- Inferring histories:
  - What algorithms do we use to find plausible answers in finite time?
- Interpretation:
  - How do we interpret the outputs?

Overview

- The basics: trees and how we think about them
- A zoo of methods: parsimony, distance, likelihood, Bayesian
- What do we do when evolution is not tree-like?

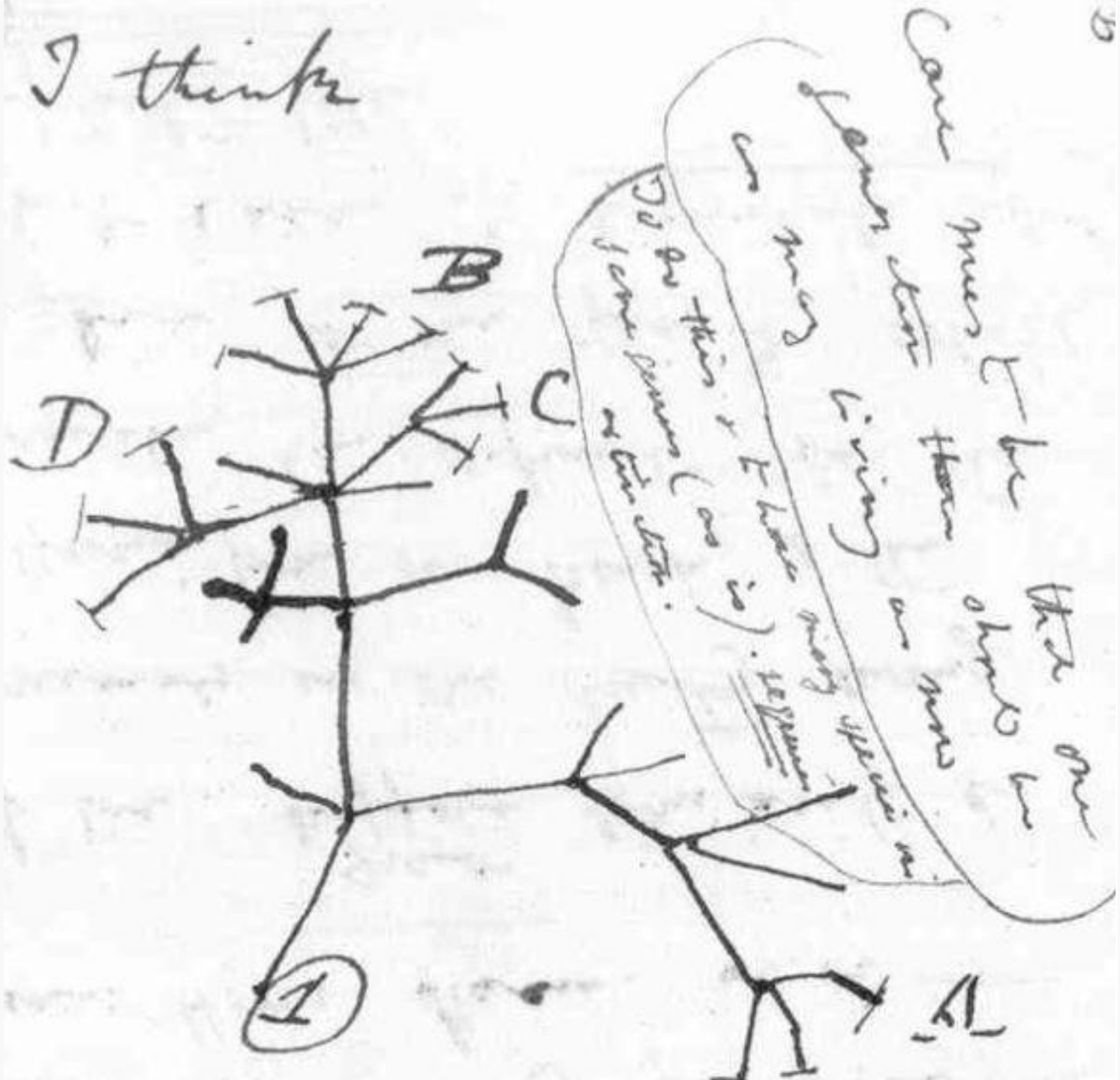
Introduction to Phylogenetics



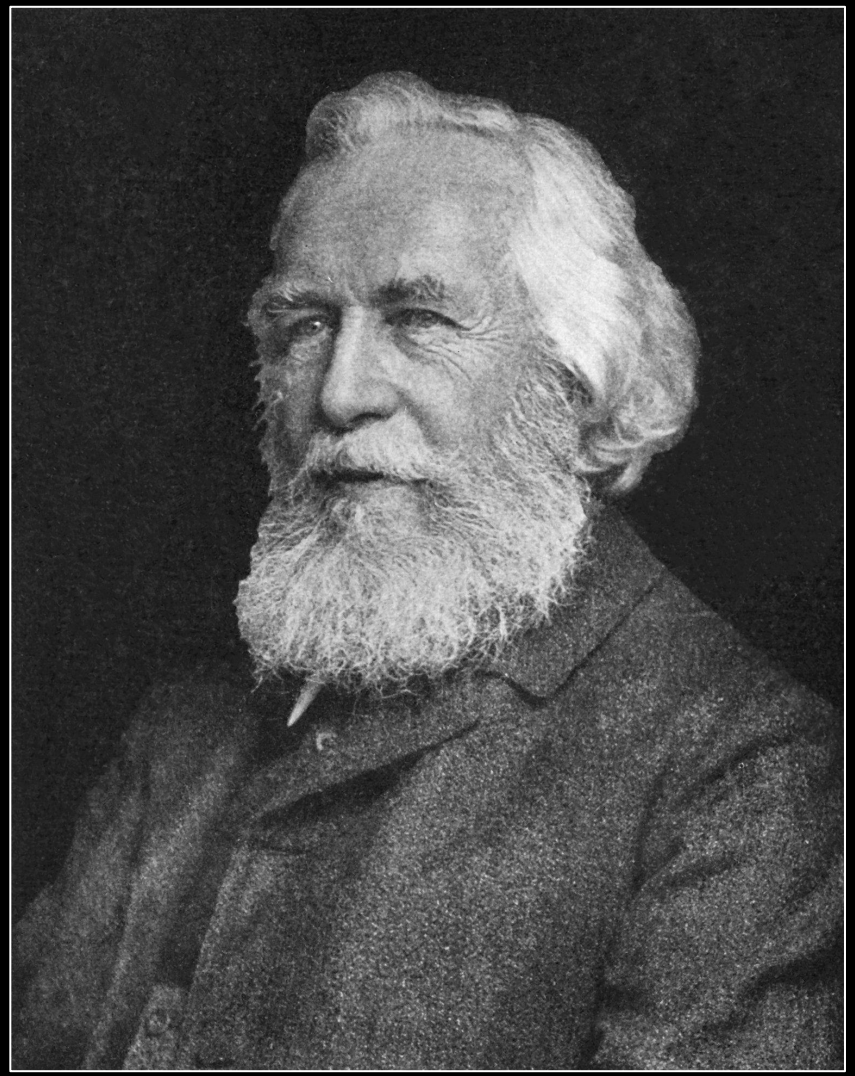
The Point

Use the relationships among one or (ideally) many homologous characters to reconstruct an evolutionary tree

Often means aligned sequences (with homologous residues in columns)



Darwin
Notebook B
(1837)

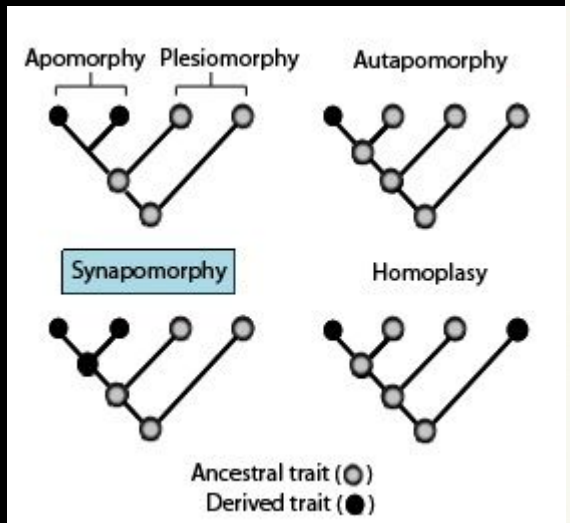


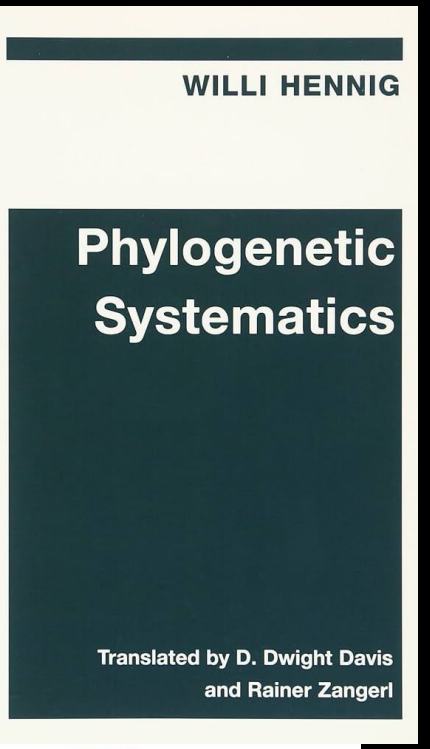
Haeckel (1874) “Pedigree of Man”

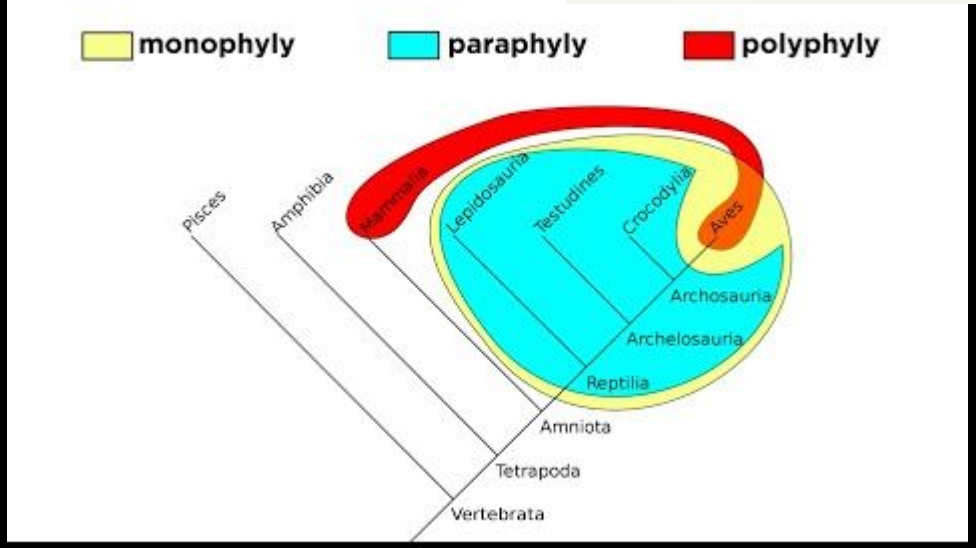
Hennig's Phylogenetic Systematics

- Formalising systematics
- Move away from pure phenetics (grouping by similarity)

- Introduction of key terms
  - **Synapomorphy** - *shared derived character state*
  - **Homoplasy** - *similarity not due to shared ancestry*
  - **Monophyletic** - *common ancestor and all descendants*
  - **Paraphyletic** - *common ancestor and some descendants*
  - **Polyphyletic** - *group not including a unique common ancestor*







The "Discovery" of the Archaea

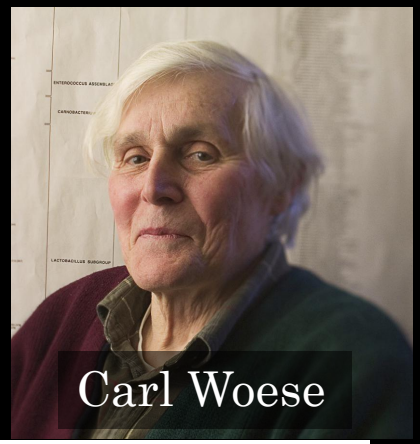
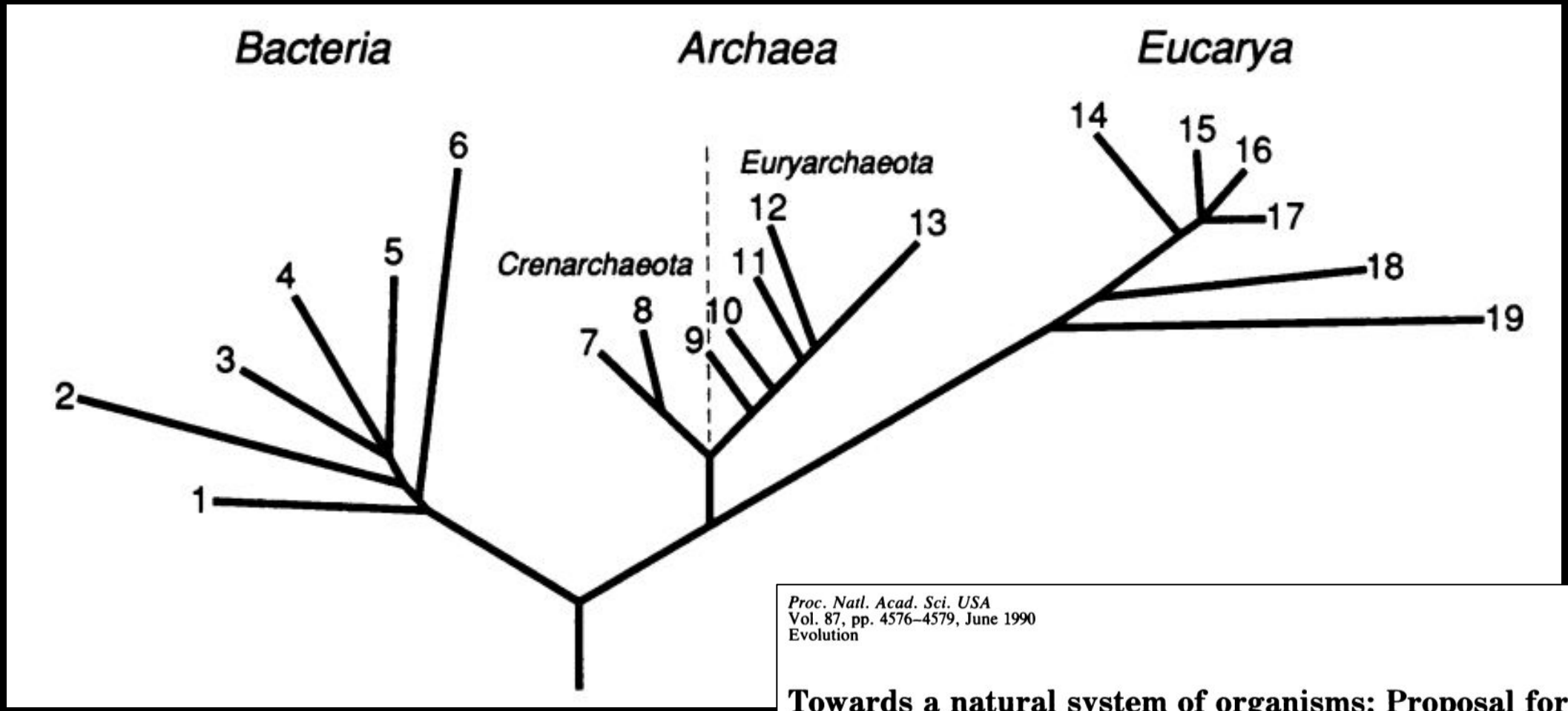


Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *Saccharomyces cerevisiae*, 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
2. *Lemna minor*, 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
4. *Escherichia coli*	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
5. *Chlorobium vibrioforme*	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
6. *Bacillus firmus*	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
7. *Corynebacterium diphtheriae*	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
8. *Aphanocapsa* 6714	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
9. Chloroplast (*Lemna*)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
10. *Methanobacterium thermoautotrophicum*	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
11. *M. ruminantium* strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
12. *Methanobacterium* sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
13. *Methanosarcina barkeri*	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

Fox and Woese (1977) *Proc Natl Acad Sci*



Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 4576–4579, June 1990
Evolution

Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya

(Euryarchaeota/Crenarchaeota/kingdom/evolution)

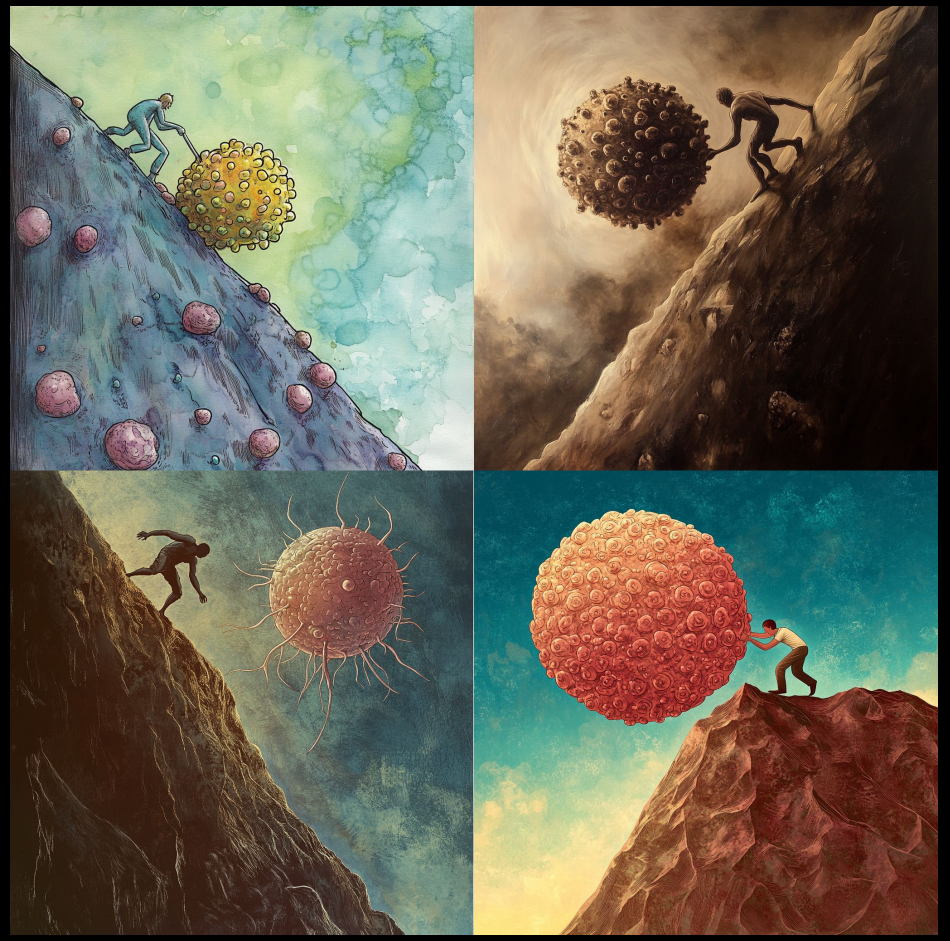
CARL R. WOESE*†, OTTO KANDLER‡, AND MARK L. WHEELIS§

*Department of Microbiology, University of Illinois, 131 Burrill Hall, Urbana, IL 61801; ‡Botanisches Institut der Universität München, Menzinger Strasse 67, 8000 Munich 19, Federal Republic of Germany; and §Department of Microbiology, University of California, Davis, CA 95616

Contributed by Carl R. Woese, March 26, 1990

Woese et al. (1990) *Proc Natl Acad Sci*

"A few notable researchers denounced the proposal of a phylogenetic classification of life into three kingdoms as a misguided move to impose a new order, citing microbiologists' long unsuccessful attempts at classifying prokaryotes, an endeavor written off as Sisyphean."

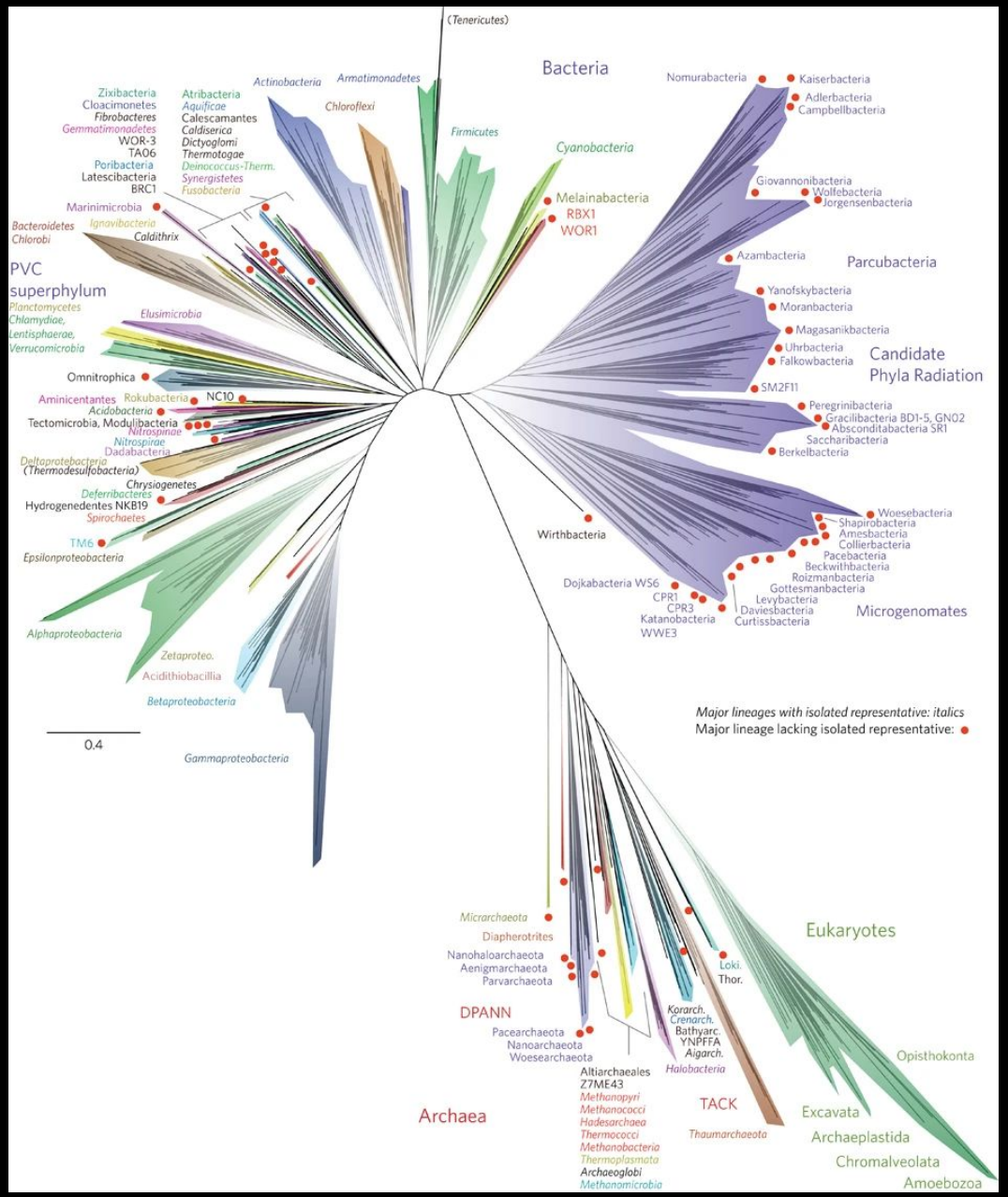


Nair (2012) *Proc Natl Acad Sci*

Refining Our View of the Tree of Life

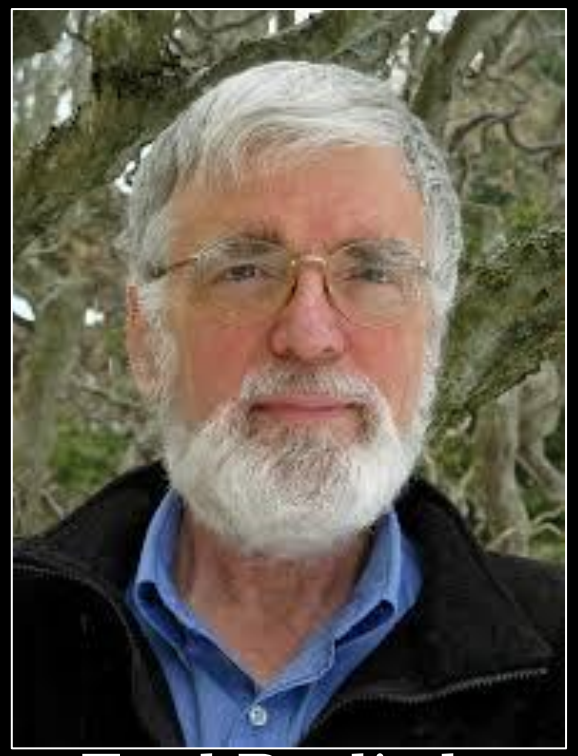
Laura Hug

Jill Banfield

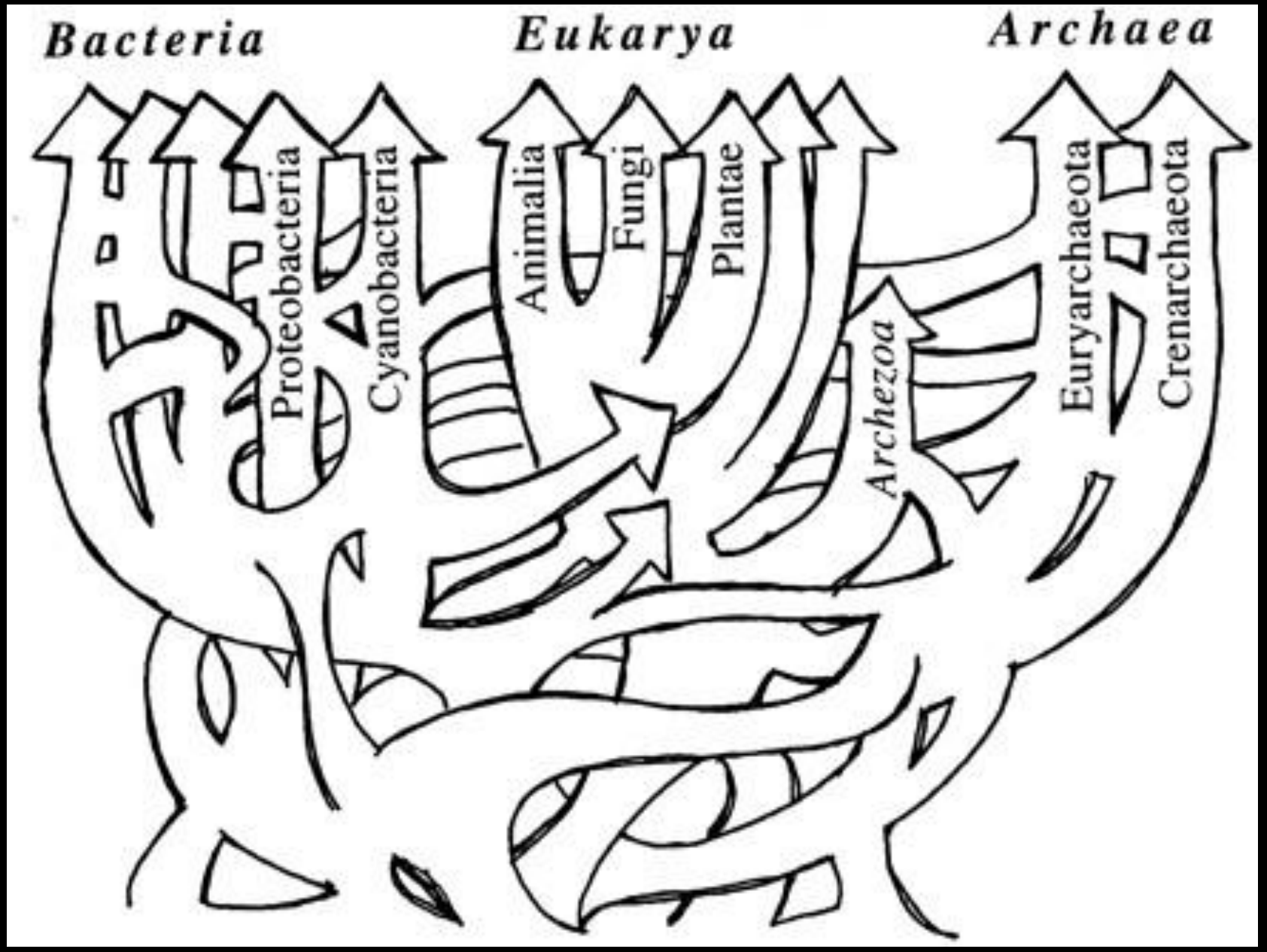


Hug et al. (2016) *Nat Microbiol*

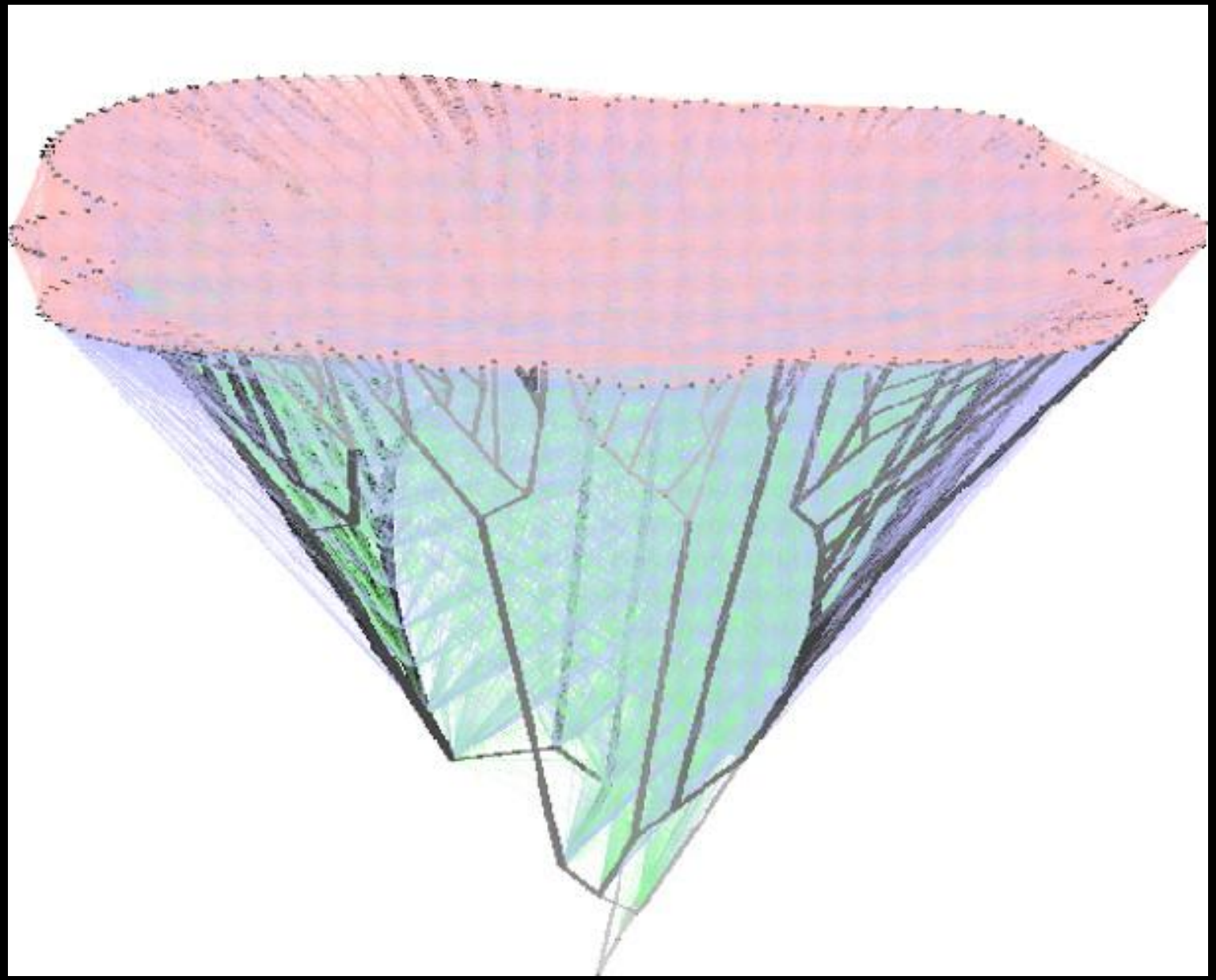
Wait a Minute…



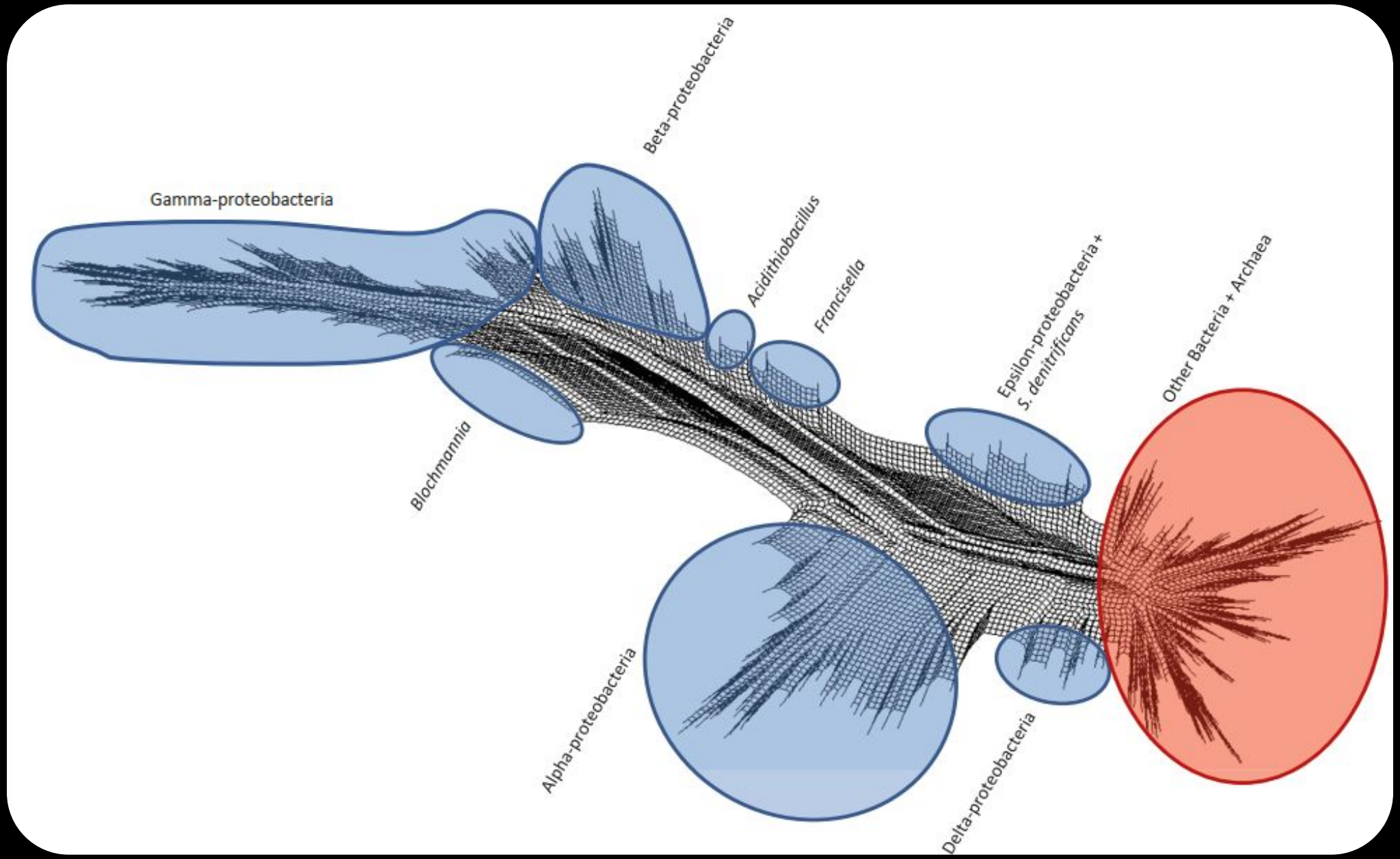
Ford Doolittle



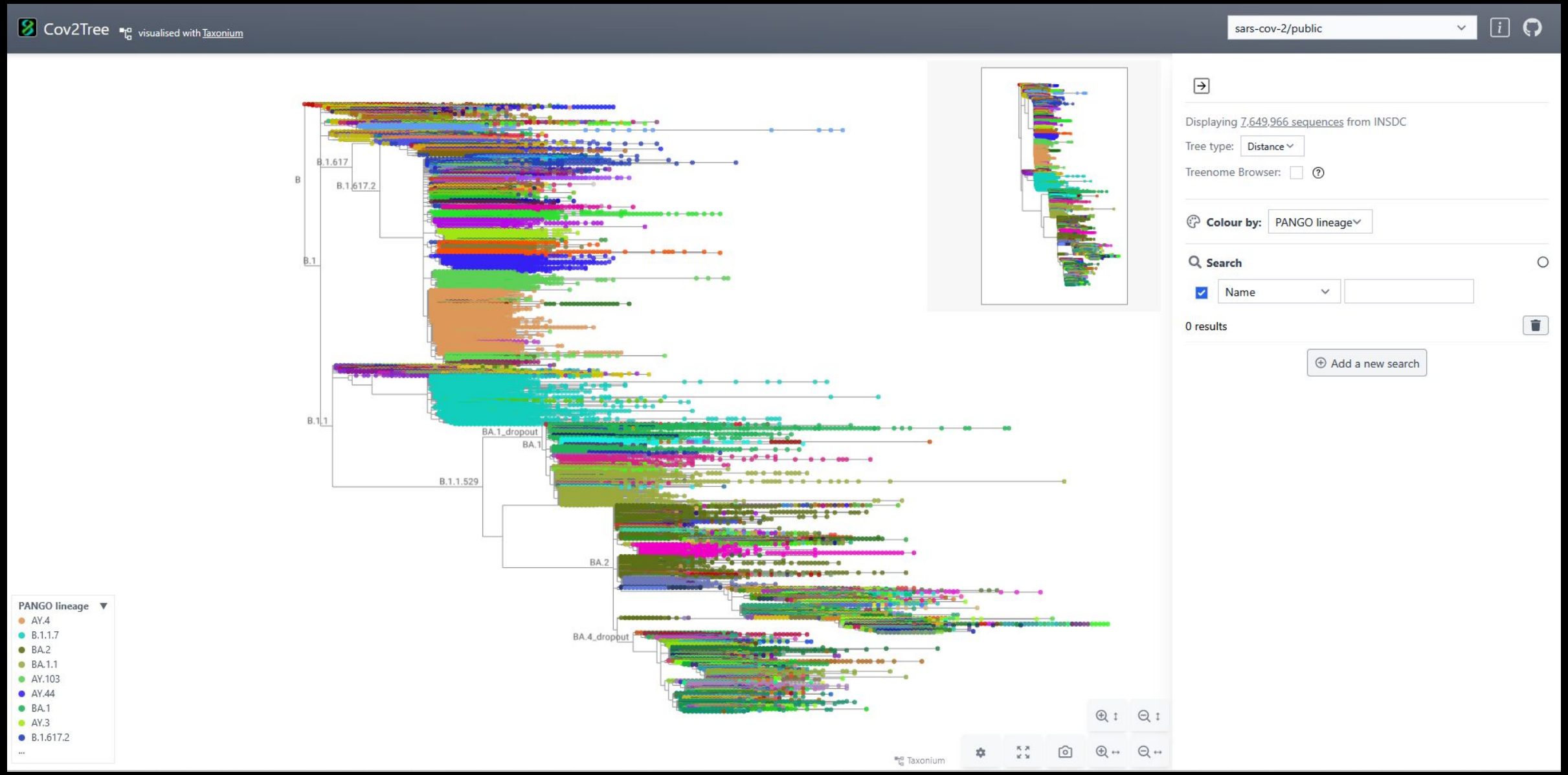
Doolittle (1999) *Sci Am*



Bacterial evolution is a mess of a network
Dagan et al.(2008) *PNAS*



Beiko (2011) *Biol Direct*

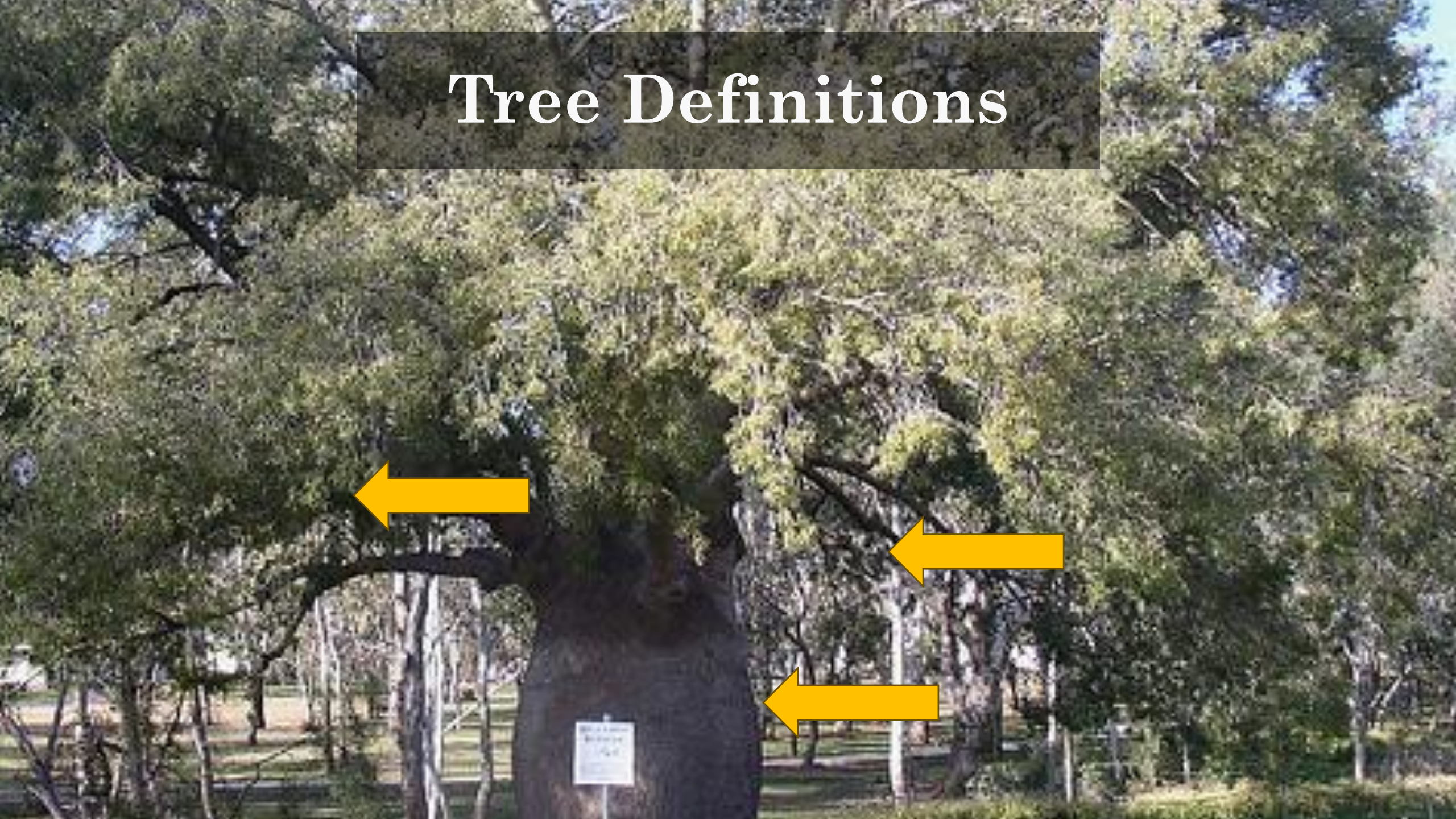


7,649,966 SARS-CoV-2 sequences
https://taxonium.org/sars-cov-2/public?xType=x_dist

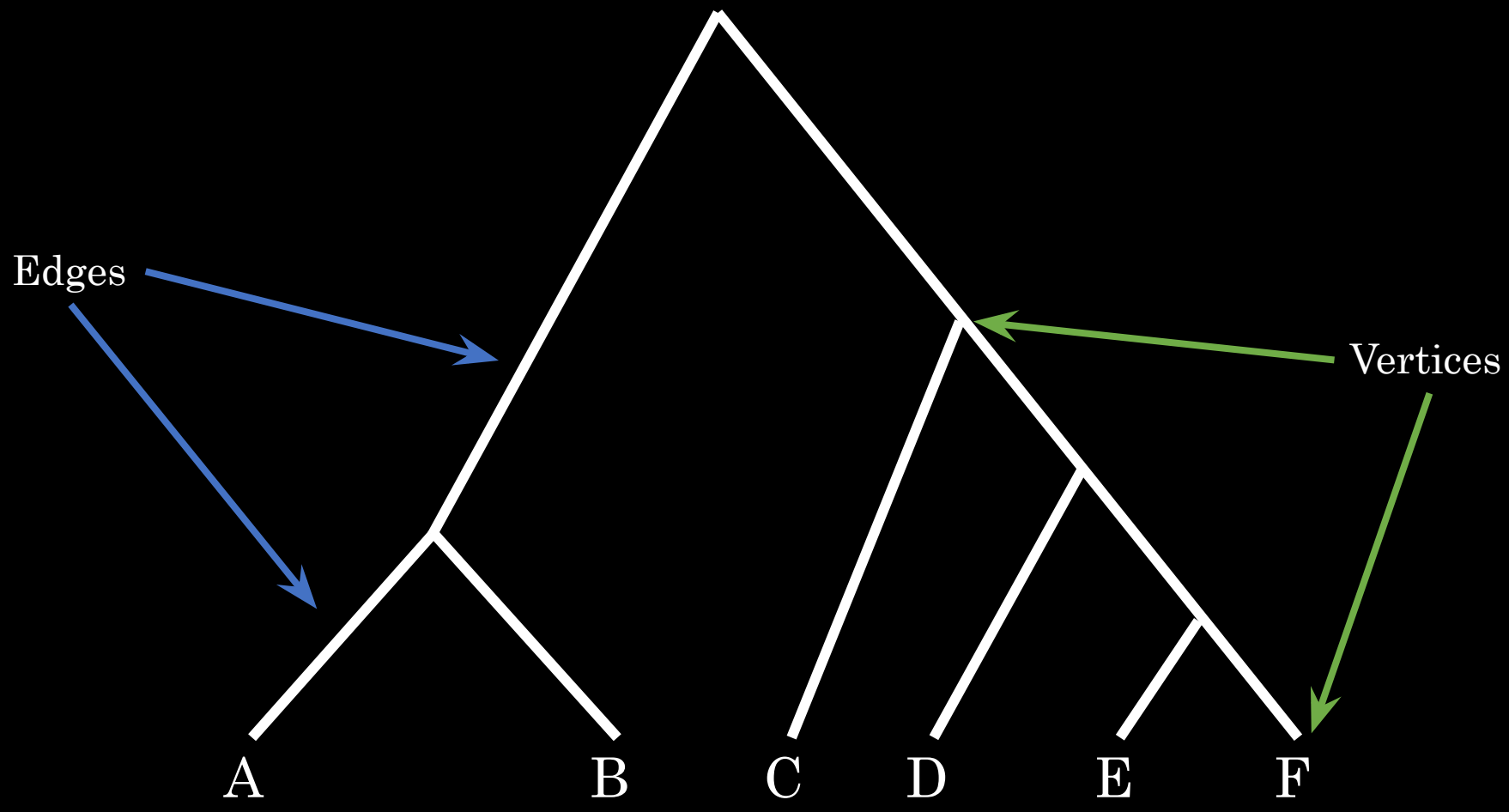
Phylogenetics is Hard
(in both the formal and informal sense)

- There are limits to what phylogenetics can tell us
- Finding the optimal tree is a non-trivial exercise
- Modeling evolution is very, very complicated

Tree Definitions

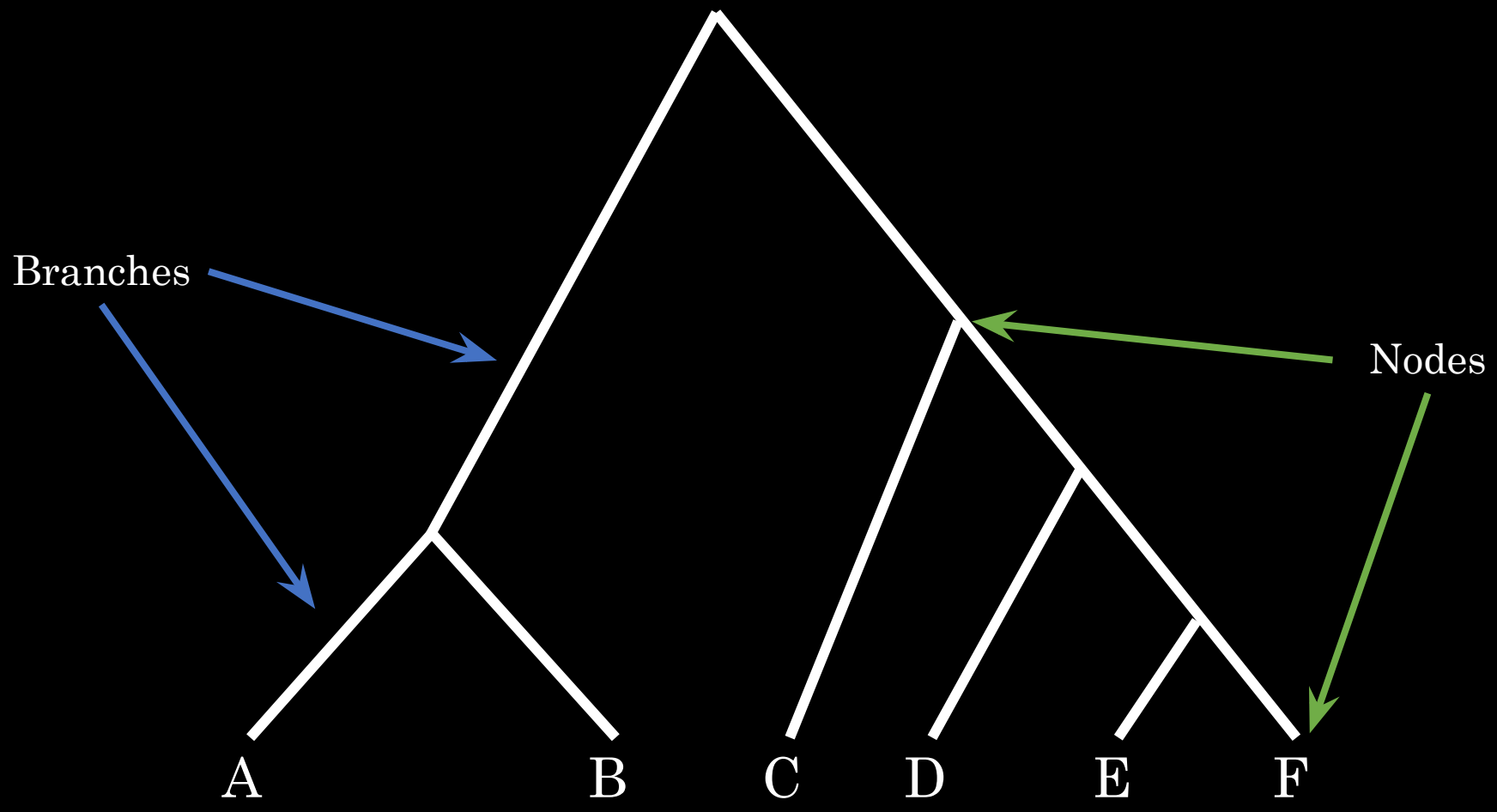


Tree Anatomy



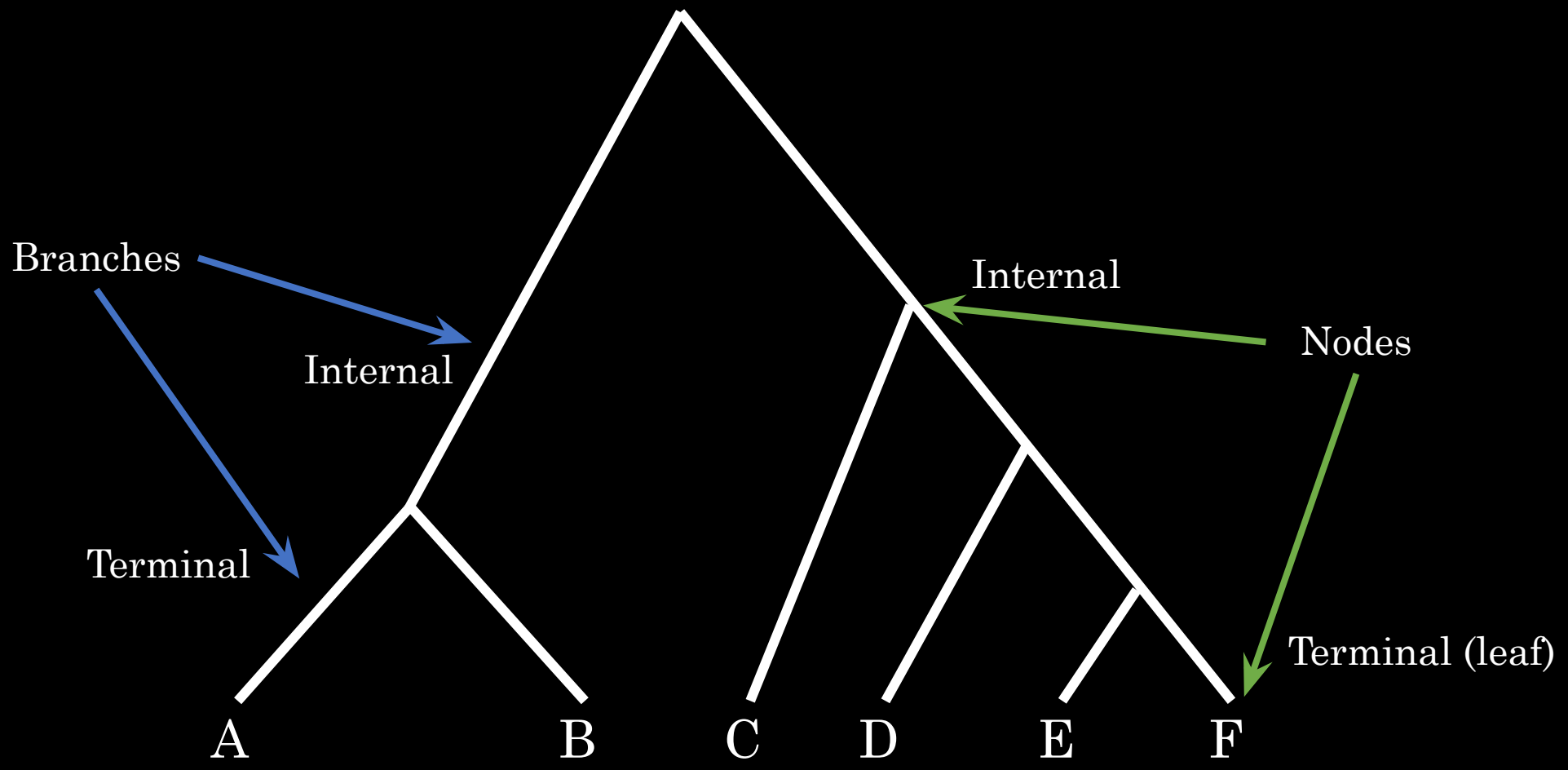
Trees can be described using the same terminology as graphs

Tree Anatomy



Or leaning into the tree metaphor

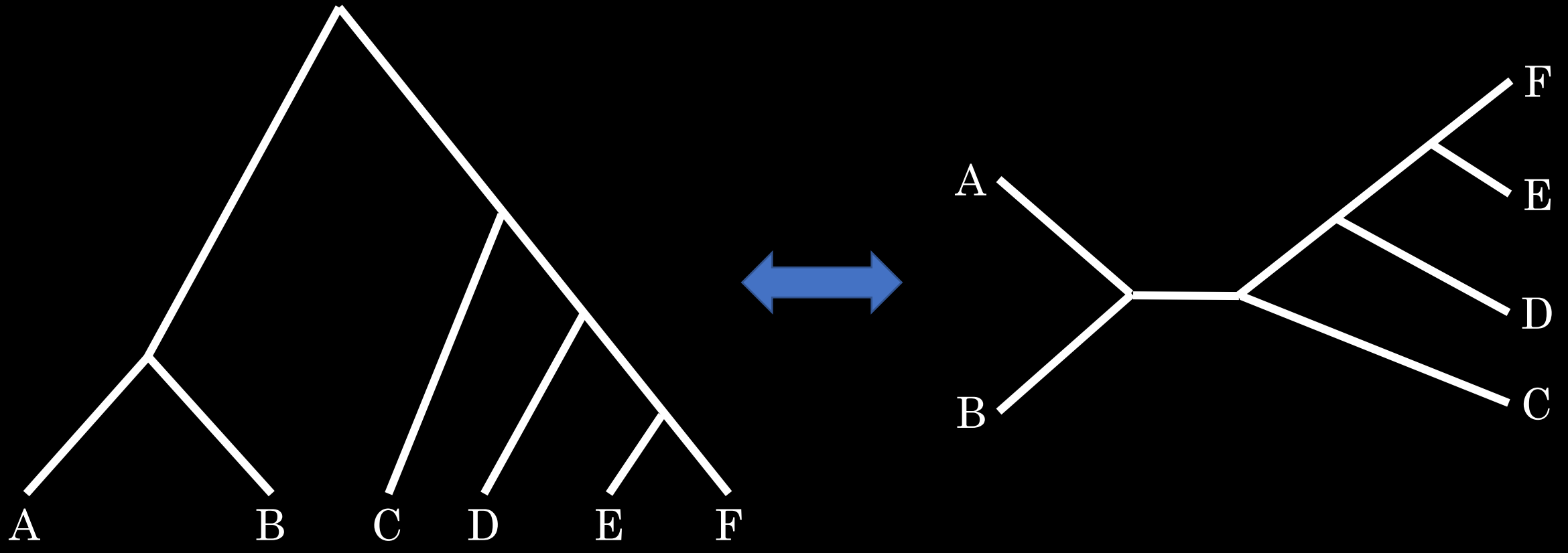
Tree Anatomy



We distinguish between internal and terminal features

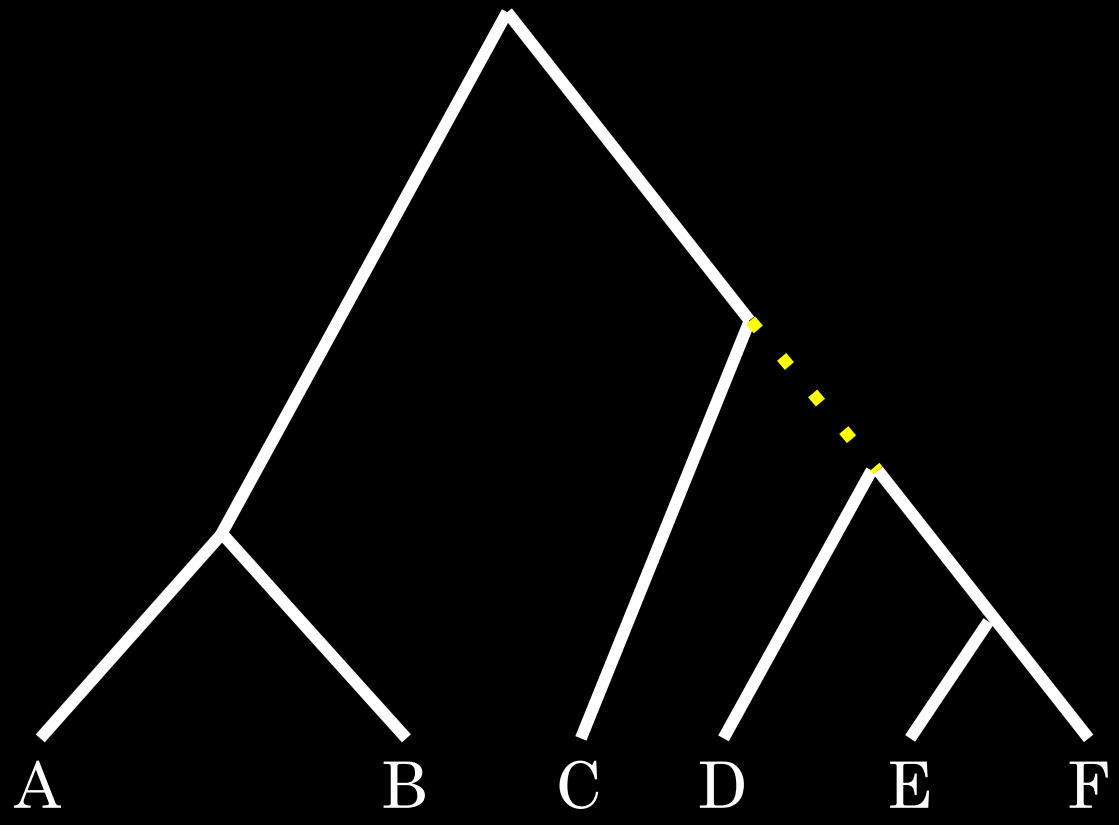
Rooted vs Unrooted Trees

Do we know the common ancestor?



Most methods generate unrooted trees

Tree splits (bipartitions)



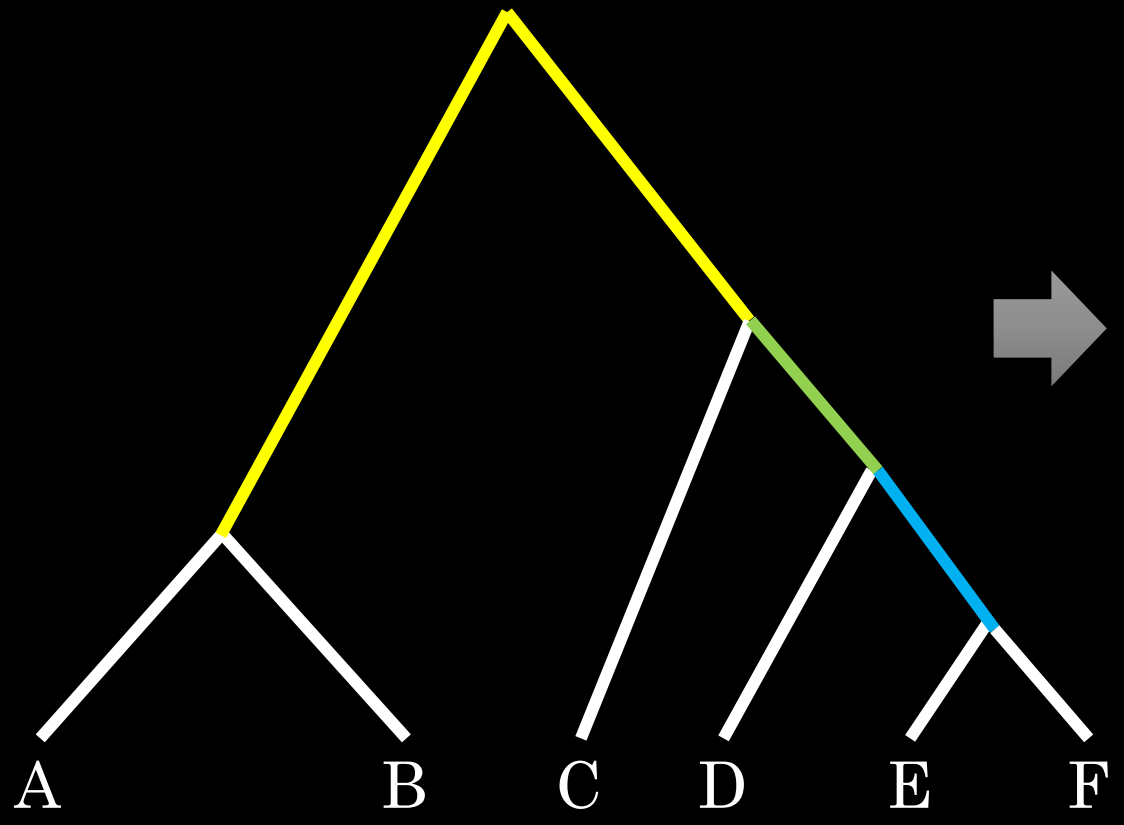
Cutting an edge in the tree yields a split or bipartition

ABC | DEF

Splits are *compatible* if they can appear in the same tree

ABC | DEF is not compatible with ABD | CEF

Split Decomposition



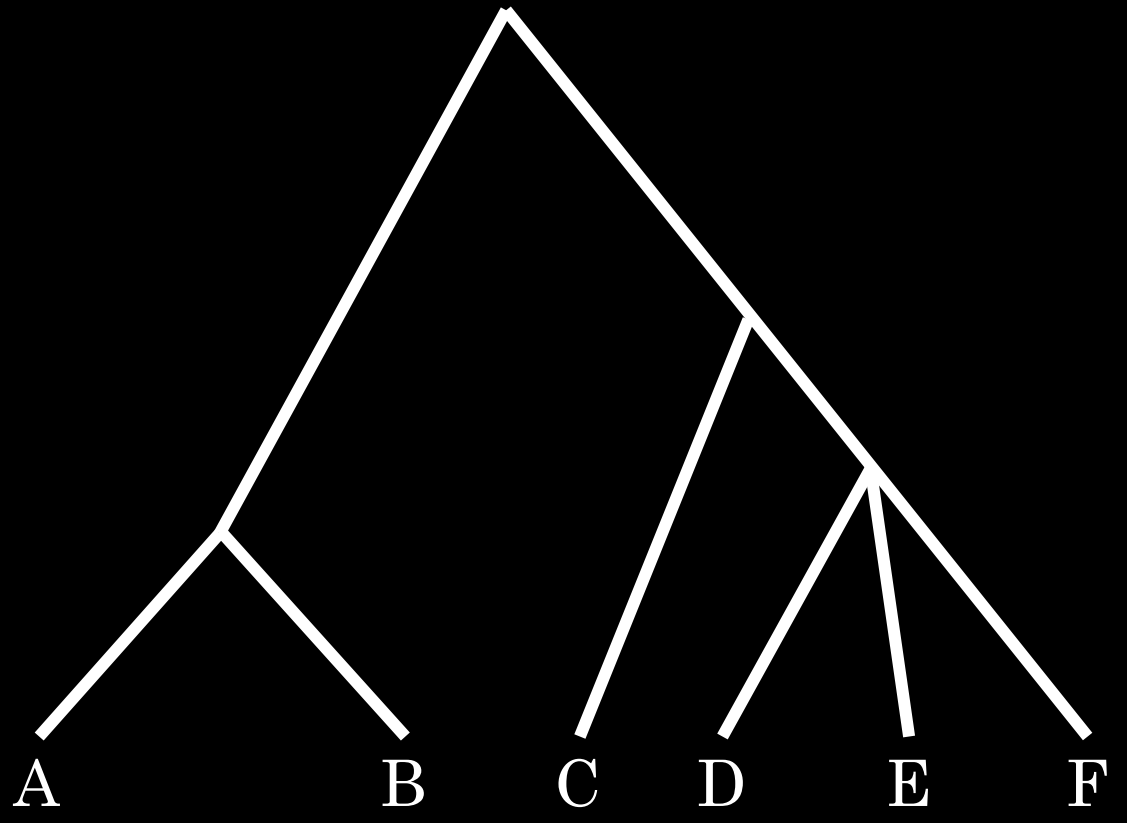
AB	CDEF
ABC	DEF
ABCD	EF

Nontrivial splits

A	BCDEF
B	ACDEF
C	ABDEF
D	ABCEF
E	ABCDF
F	ABCDE

Trivial splits

Multifurcating Trees



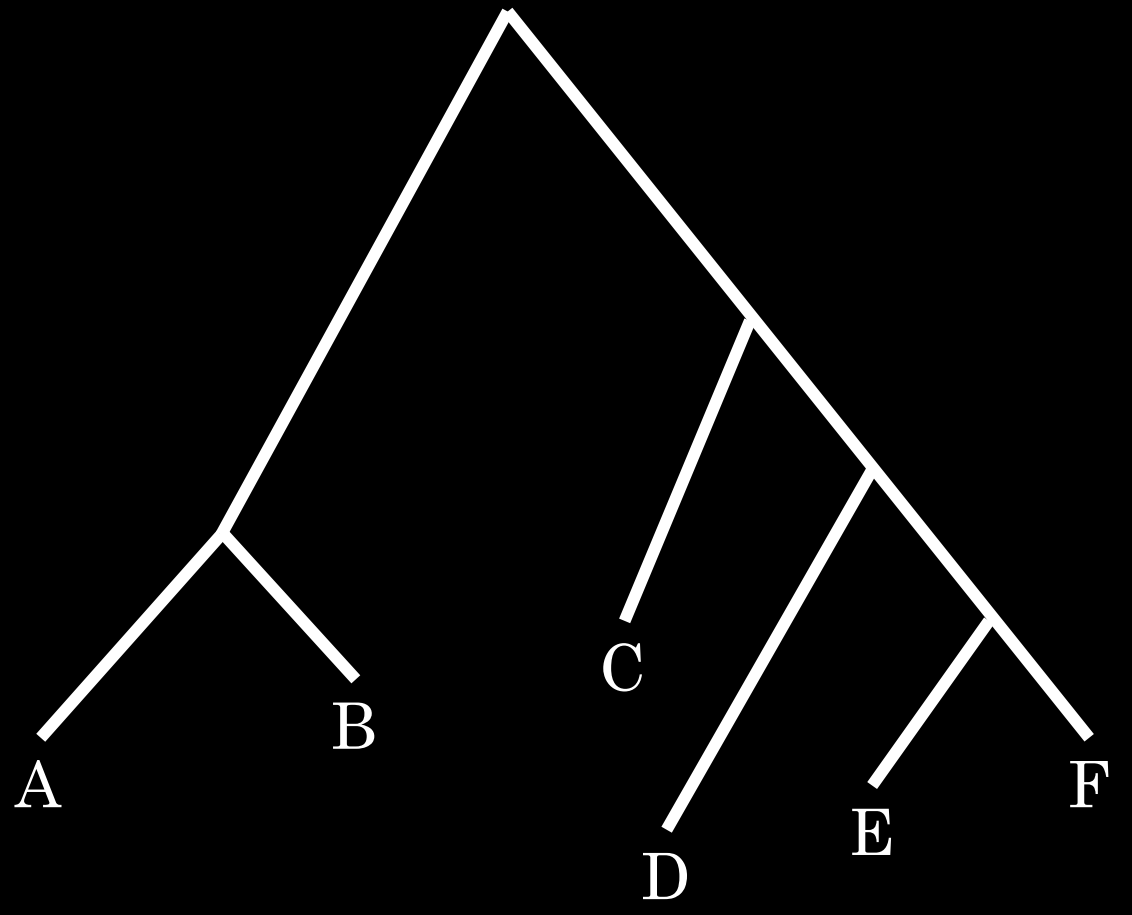
We may *collapse* a node in the tree for one of two reasons:

- Genuine 3-way split (rare)
- Lack of statistical support for any specific pairing of taxa (painfully common)

What is the correct branching order of (D,E,F)?

Most phylogenetic methods produce only binary trees
(but you can roll back relationships that lack support)

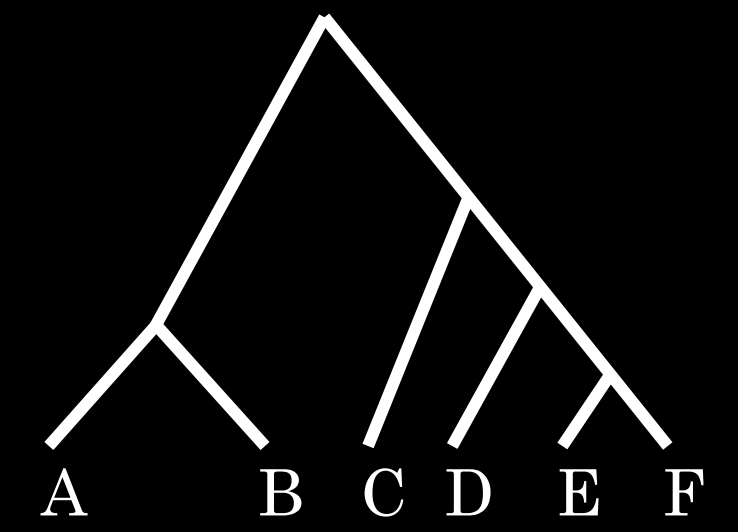
Informative Branch Lengths

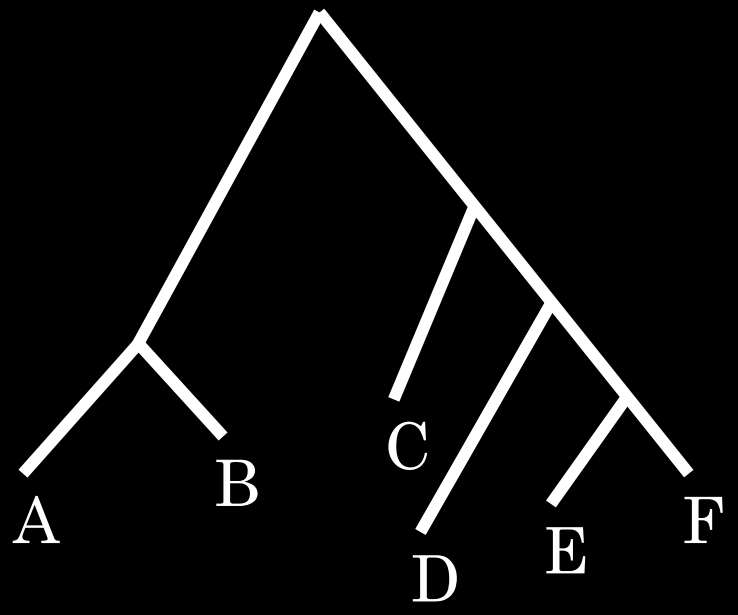


What (if anything) do branch lengths represent?

- Time?
- Sequence change?

Some methods (notably parsimony) do not produce meaningful branch lengths



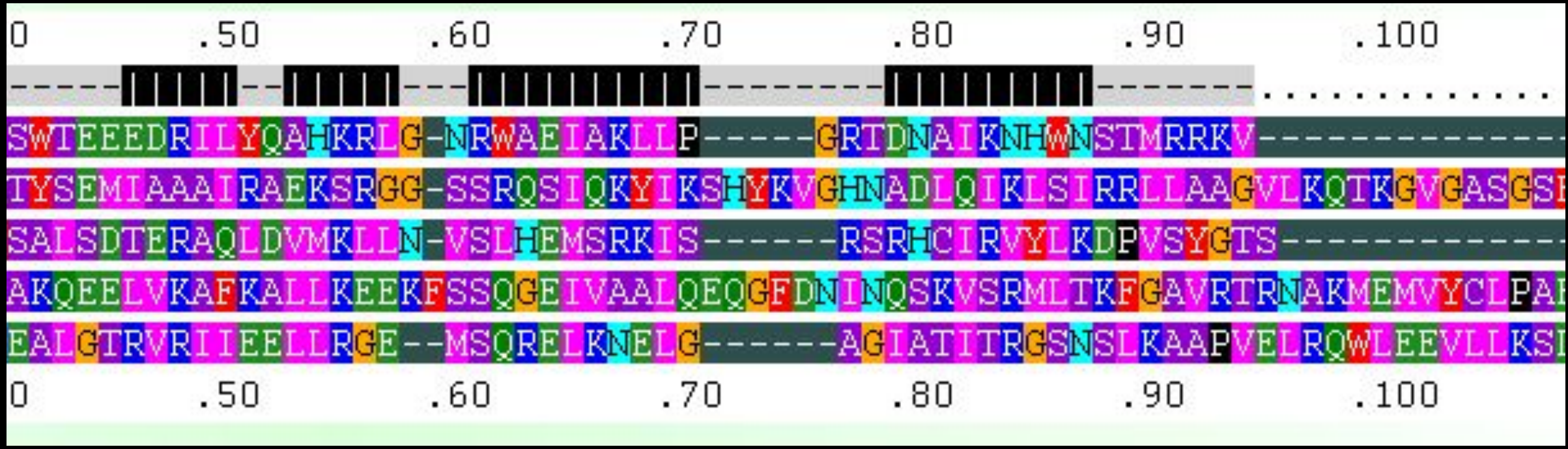


Tree Shape

The shape of a tree refers to its branching order, not to branch lengths

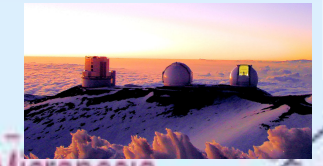
So the two trees on the left have the same shape

Shape can be described completely using a split decomposition of the tree



But nucleotides and amino acids are not the only type of character that can be compared!

The Polynesian Triangle



Hawai'i

Formosan

WMP

SHWNG

Micronesia

Nuclear Micronesian

Oceanic

CMP

Madagascar

Wallacea

Rugby Playing Nations

Polynesian

Polynesia

Rapa Nui

Aotearoa

Words as homologous characters

Language trees support the express-train sequence of Austronesian expansion

Russell D. Gray & Fiona M. Jordan

Department of Psychology, University of Auckland, Auckland 92019, New Zealand

Meaning	Tonga	Niue	Samoa	E. Uvea	E. Futuna	Mangareva	Marquesas	Hawaii	Tahiti	Tuamotu	Rarotonga
Canoe	vaka	vaka	va'a	vaka	vaka	vaka	vaka	wa'a	va'a	vaka	vaka
Two	ua	ua	lua	lua	lua	rua	'ua	lua	rua	rua	rua
Five	nima	lima	lima	nima	lima	rima	'ima	lima	rima	rima	rima
Woman	fefine	fifine	fafine	fafine	fafine	ahine	vehine	wahine	vahine	vahiine	va'ine
Rainbow	*'umata*	*tangaloa*	nuanua	nuanua	nuanua	anuanua	aanuanua	aanuenue	aanuanua	anuanua	aanuanua

Like a sequence alignment – homologous characters identified

No collinearity constraint

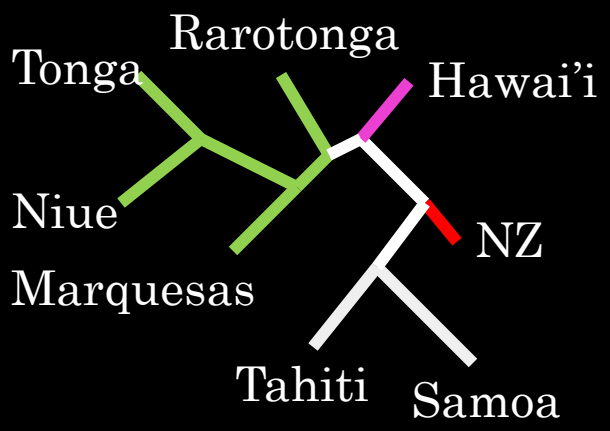
(but who cares?)

Gray and Jordan (2000) *Nature*

Character Convexity

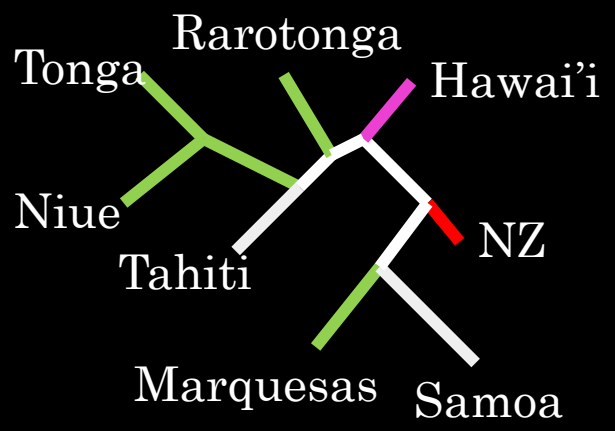
Island	Canoe
Tonga	Vaka
Niue	Vaka
Rarotonga	Vaka
Marquesas	Vaka
Hawai'i	Wa'a
Tahiti	Va'a
Samoa	Va'a
NZ	Waka

- Choose a tree at random (for now)
- A character is convex on that tree if all states of that character can be partitioned to a separate 'region' of the tree
- Think of it as a coloring problem!



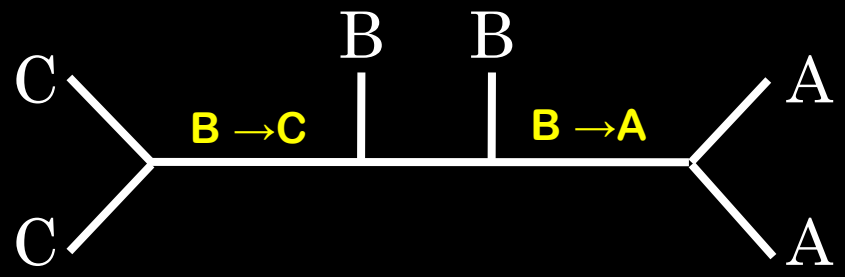
CONVEX

NOT CONVEX



What does convexity mean?

- If we have n states (waka, vaka, etc.) for a given character, then we only need at minimum $n - 1$ state changes within the tree

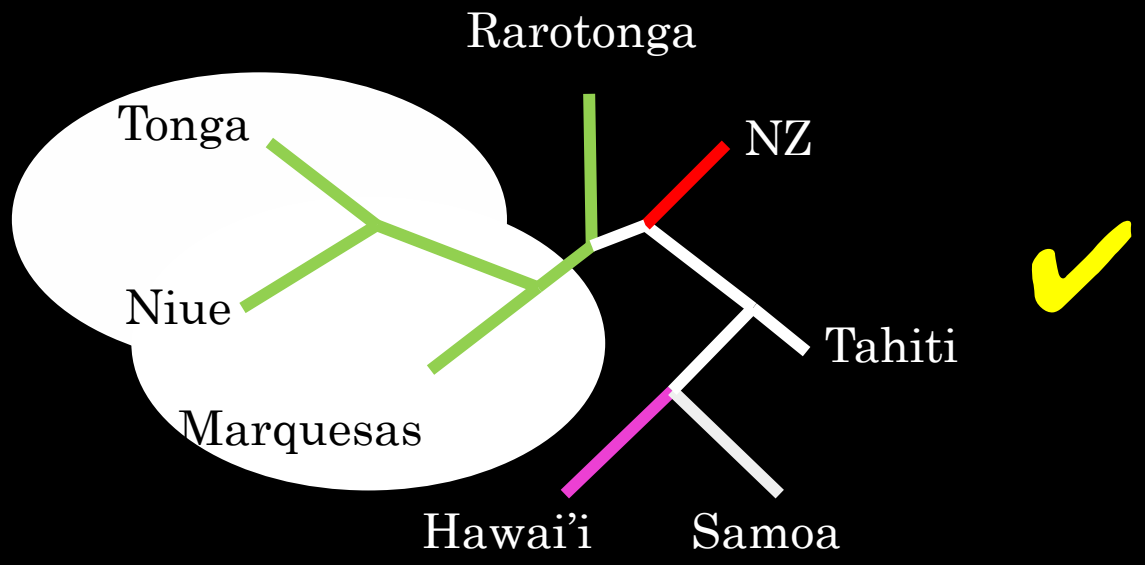


- The is the most parsimonious (simplest) situation

Character Compatibility

Island	Canoe	Two
Tonga	Vaka	Ua
Niue	Vaka	Ua
Rarotonga	Vaka	Rua
Marquesas	Vaka	‘ua
Hawai’i	Wa’a	Lua
Tahiti	Va’a	Rua
Samoa	Va’a	Lua
NZ	Waka	Rua

Two characters (words, alignment columns, etc.) are *compatible* if there exists at least one tree where both characters are convex



What is the “best” tree?

- Is it the maximum compatibility tree that maximizes the number of convex characters from the set C of characters?

maybe…but usually not

- What we typically want is the tree that minimizes the number of substitutions over *all* characters – this is the maximum parsimony tree

What is the “best” tree?

- Is it the maximum compatibility tree that maximizes the number of convex characters from the set C of characters?

Meaning	Tonga	Niue	Samoa	E. Uvea	E. Futuna	Mangareva	Marquesas	Hawaii	Tahiti	Tuamotu	Rarotonga
Canoe	vaka	vaka	va'a	vaka	vaka	vaka	vaka	wa'a	va'a	vaka	vaka
Two	ua	ua	lua	lua	lua	rua	'ua	lua	rua	rua	rua
Five	nima	lima	lima	nima	lima	rima	'ima	lima	rima	rima	rima
Woman	fefine	fifine	fafine	fafine	fafine	ahine	vehine	wahine	vahine	vahiine	va'ine
Rainbow	*'umata*	*tangaloa*	nuanua	nuanua	nuanua	anuanua	aanuanua	aanuenue	aanuanua	anuanua	aanuanua

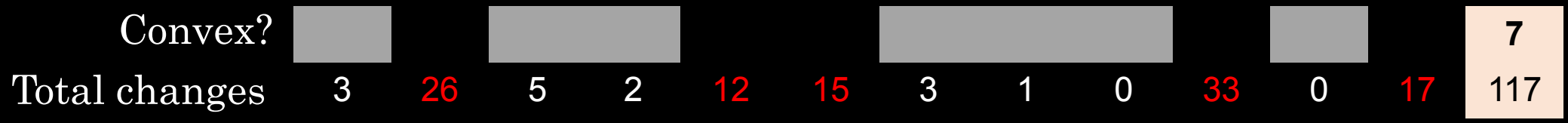
Convex? 

What is the “best” tree?

What if the sum of scores for a given tree are awful?

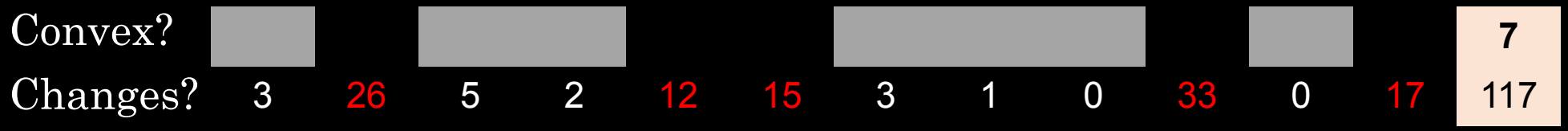
Meaning	Tonga	Niue	Samoa	E. Uvea	E. Futuna	Mangareva	Marquesas	Hawaii	Tahiti	Tuamotu	Rarotonga
Canoe	vaka	vaka	va'a	vaka	vaka	vaka	vaka	wa'a	va'a	vaka	vaka
Two	ua	ua	lua	lua	lua	rua	'ua	lua	rua	rua	rua
Five	nima	lima	lima	nima	lima	rima	'ima	lima	rima	rima	rima
Woman	fefine	fifine	fafine	fafine	fafine	ahine	vehine	wahine	vahine	vahiine	va'ine
Rainbow	*'umata*	*tangaloa*	nuanua	nuanua	nuanua	anuanua	aanuanua	aanuenue	aanuanua	anuanua	aanuanua

(imagine a much larger table…)



What is the “best” tree?

Tree T_1:



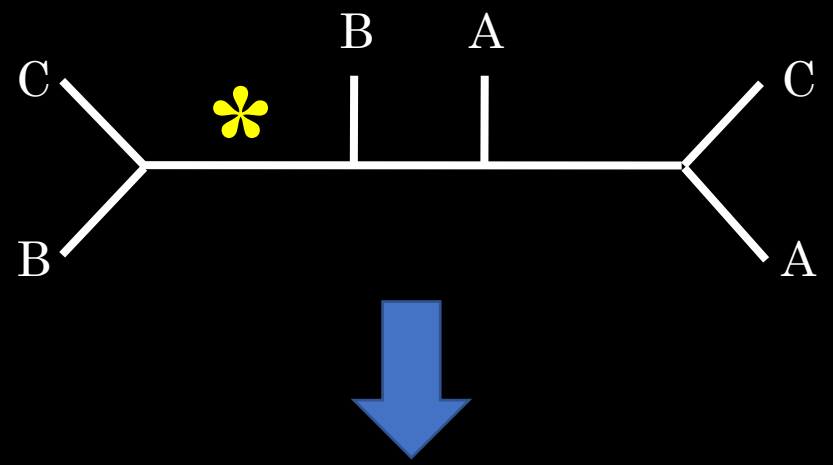
Tree T_2:

Parsimony Score

- The parsimony score (p) for a given character on a given tree T is the minimum number of changes needed to map character states onto leaves of the tree

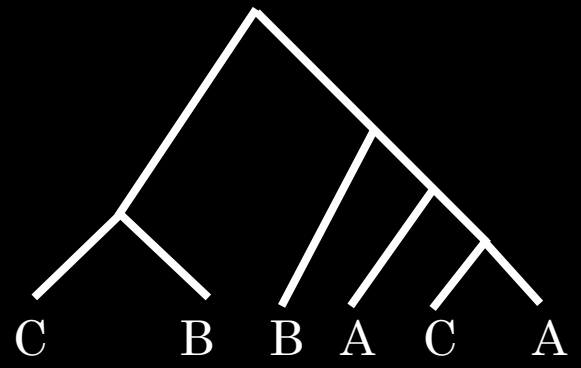
- How do we find this minimum for a single character?

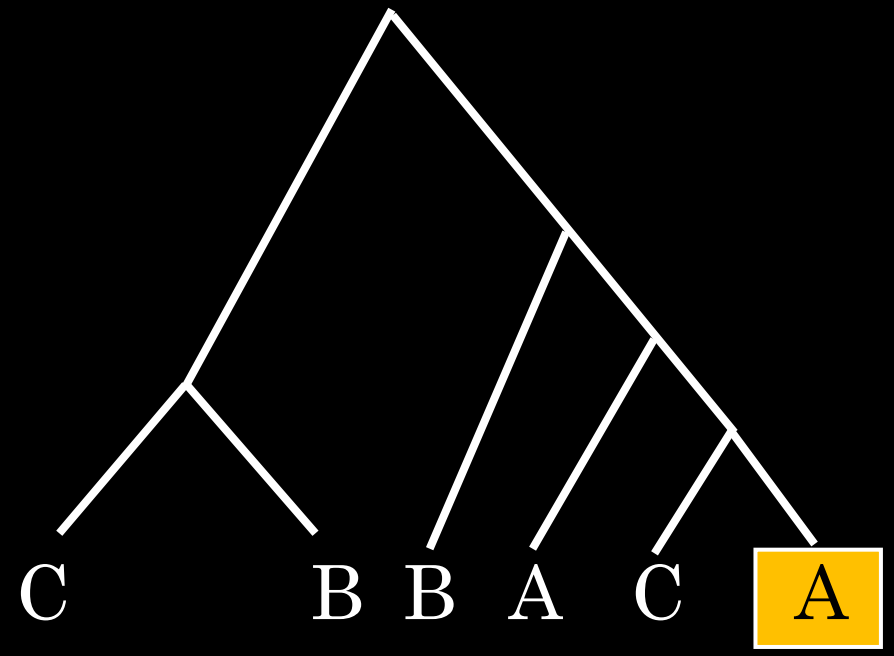
Fitch-Hartigan algorithm



One character, three states

Introduce an arbitrary root to the tree if unrooted

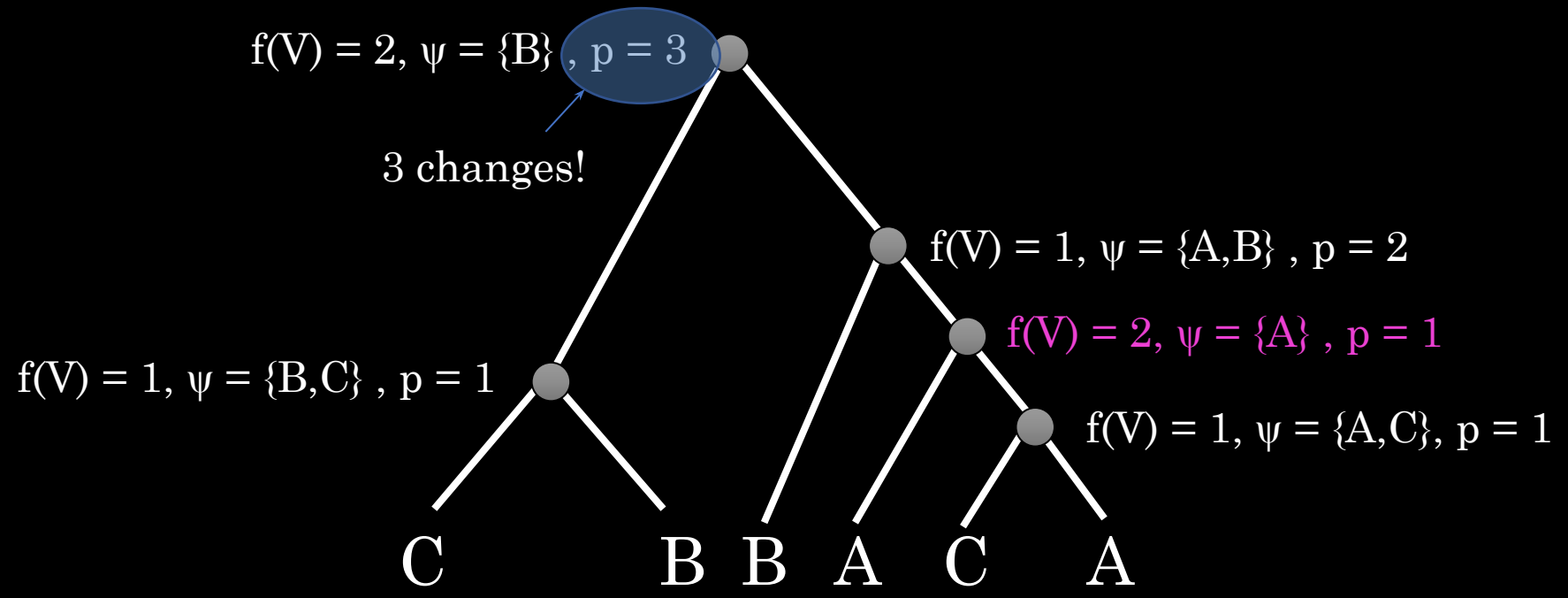




Start at the LEAF vertices

At each leaf vertex, the count of changes $p = 0$ and the set of characters $\psi = \{X\}$

p = 0
ψ = {A}



Mapping to internal vertices V:

O(N)

f(V) is the maximum number of immediate children that contain any **particular** character state

→ *best guess for internal states*

ψ is the character or characters that cover f(V) children

→ *equally good internal state guesses*

p is equal to (p of all children) + (number of children) – f(V)

→ *number of required changes so far*

Total Parsimony Score

(for a given tree)

$$p_T = \sum_{c \in C} p_T(c)$$

The maximum parsimony *tree* is the tree that minimizes p_T

Note that it does not explicitly count convex characters!
They simply contribute the minimum possible changes given the number of states they contain

Maximum Parsimony

- There is no closed-form solution to find T such that p_T is minimal

- We must carry out a search through *tree space* – typically use a random starting tree T_0 and explore by permuting this tree

Tree Searching

1. Choose a random starting tree T_0
2. n ← 0 *(this is the iteration number)*
3. Compute p_{T0}
4. While (patience remains)
    1. Permute T_n
    2. $T_{n+1} = \text{argmin}_p(T_n, \text{permuted } T_n)$
    3. n ← n+1
5. Output T_n

Problem

- There are a lot of trees!

- For n leaves, there are

1 x 3 x 5 x … x (2n - 3) rooted, bifurcating trees

$$n_T = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

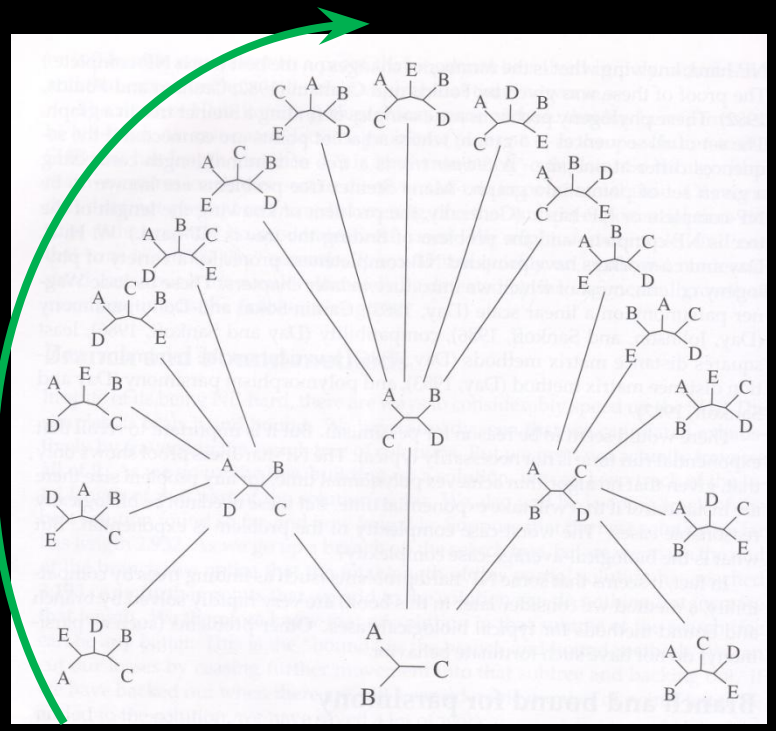
20 leaves → 8,200,794,532,637,891,559,375 trees

Branch-and-Bound

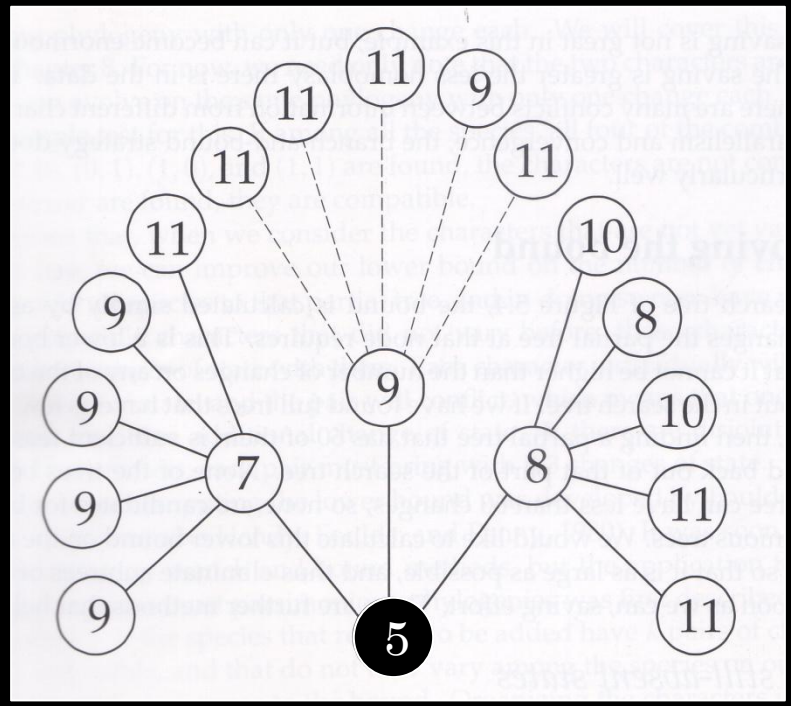
One way to restrict the search space is to explore it systematically, but identify and stop unproductive search paths

Branch-and-Bound

Species	Character					
	1	2	3	4	5	6
A	1	0	0	1	1	0
B	0	0	1	0	0	0
C	1	1	0	0	0	0
D	1	1	0	1	1	1
E	0	0	1	1	1	0



Tree building procedure

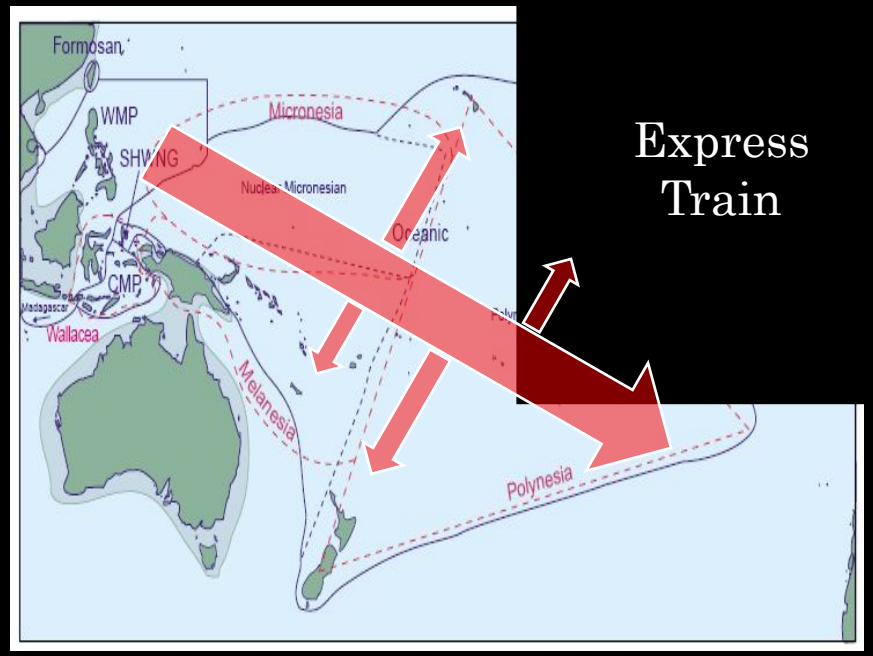


Number of substitutions required

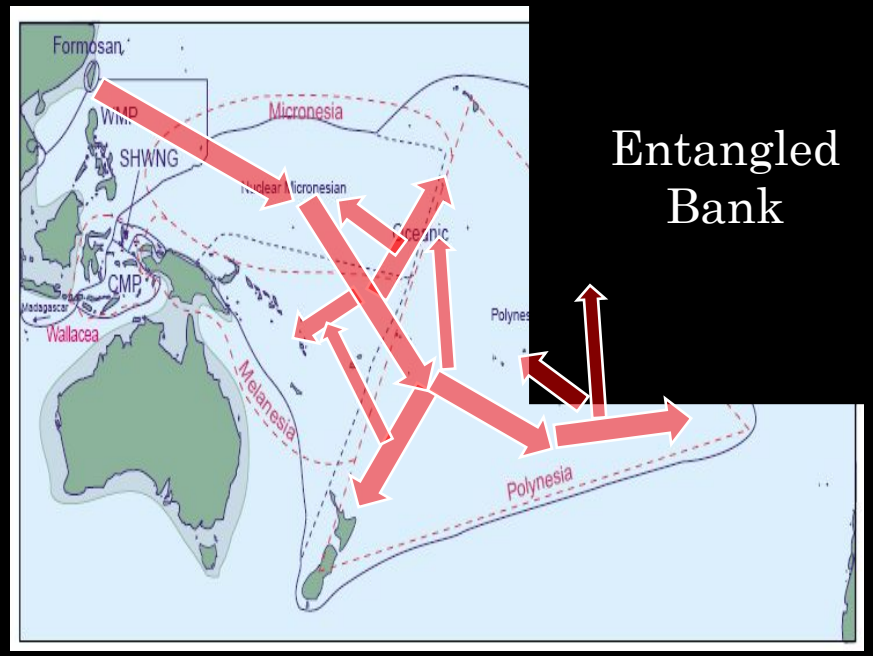
Felsenstein (2004) *Inferring Phylogenies*

Back to Polynesia

- Hypotheses about Polynesian expansion
- What are the predictions of these two models?

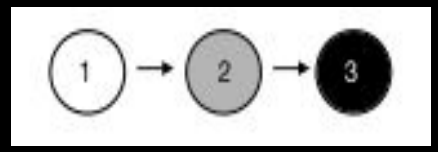


Express Train

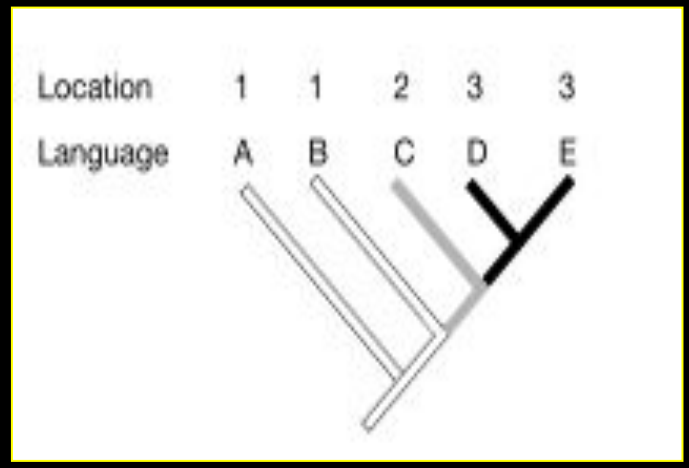


Entangled Bank

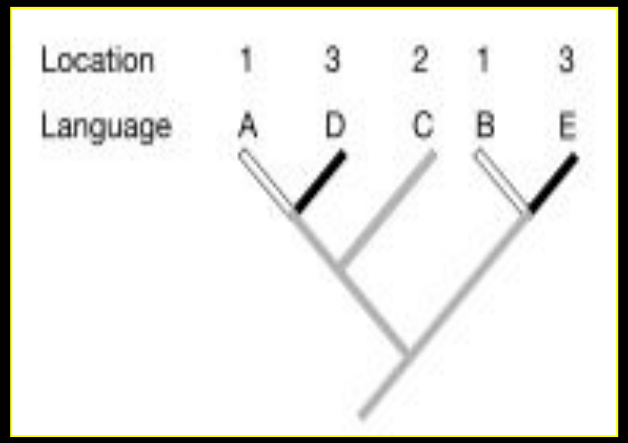
Predictions



- Express train: strong tree-like signal, congruent with geography. **High CI**
  (assuming enough time for language to evolve)

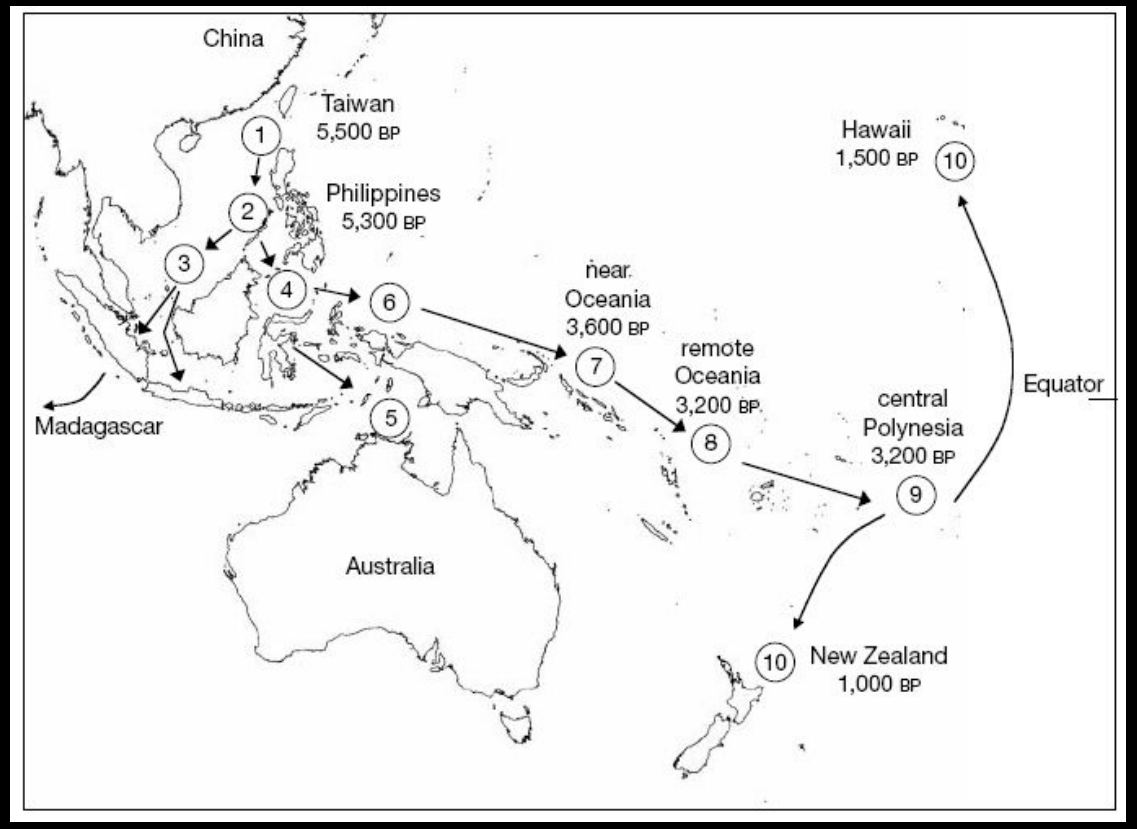


- Entangled bank: weaker signals, lots of sharing (travel / cultural exchange). **Low CI**

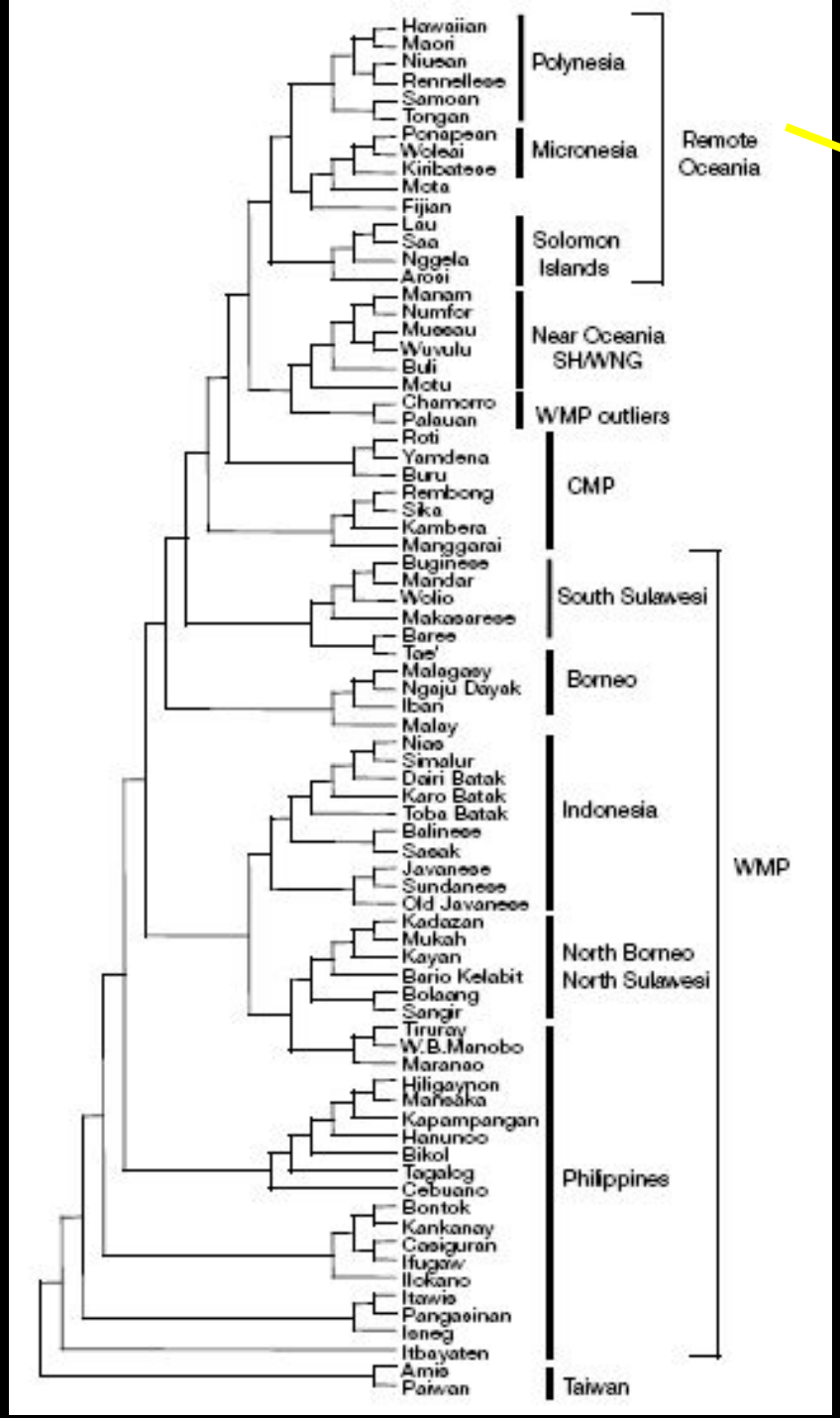


Analysis

- 77 Austronesian languages
- 5185 characters (no equivalent to NCBI!)



Express train model – 77 languages grouped into 10 categories (Jared Diamond's archaeological 'stations')



(NZ)

Relationships in the language tree are driven more by express-train predictions than by geographic proximity

Minimum number of transitions with respect to stations: 9 (= 10 – 1)

A total of 18 steps is needed to reconcile the 10 character states with the recovered tree (close to optimal)

We can compare the fit to random trees to see whether the fit is better than expected

Randomized trees: Average of 95 steps

So there is significant tree-like signal, and the *shape* of the tree is consistent with express-train predictions

How well do the characters fit the tree?

We can use the consistency index

$$CI_{character} = m / s$$

Where m is the minimum number of steps across all characters
(= number of character states – 1)
And s is the actual number of steps (≥ m), from the F-H algorithm

$$0.0 < CI \leq 1.0$$

Now Let's Apply the Consistency Index

- We need to add up the total number of steps required for every word
- Tree with fewest changes: 52,129 steps across all words
- Reported CI: 0.25

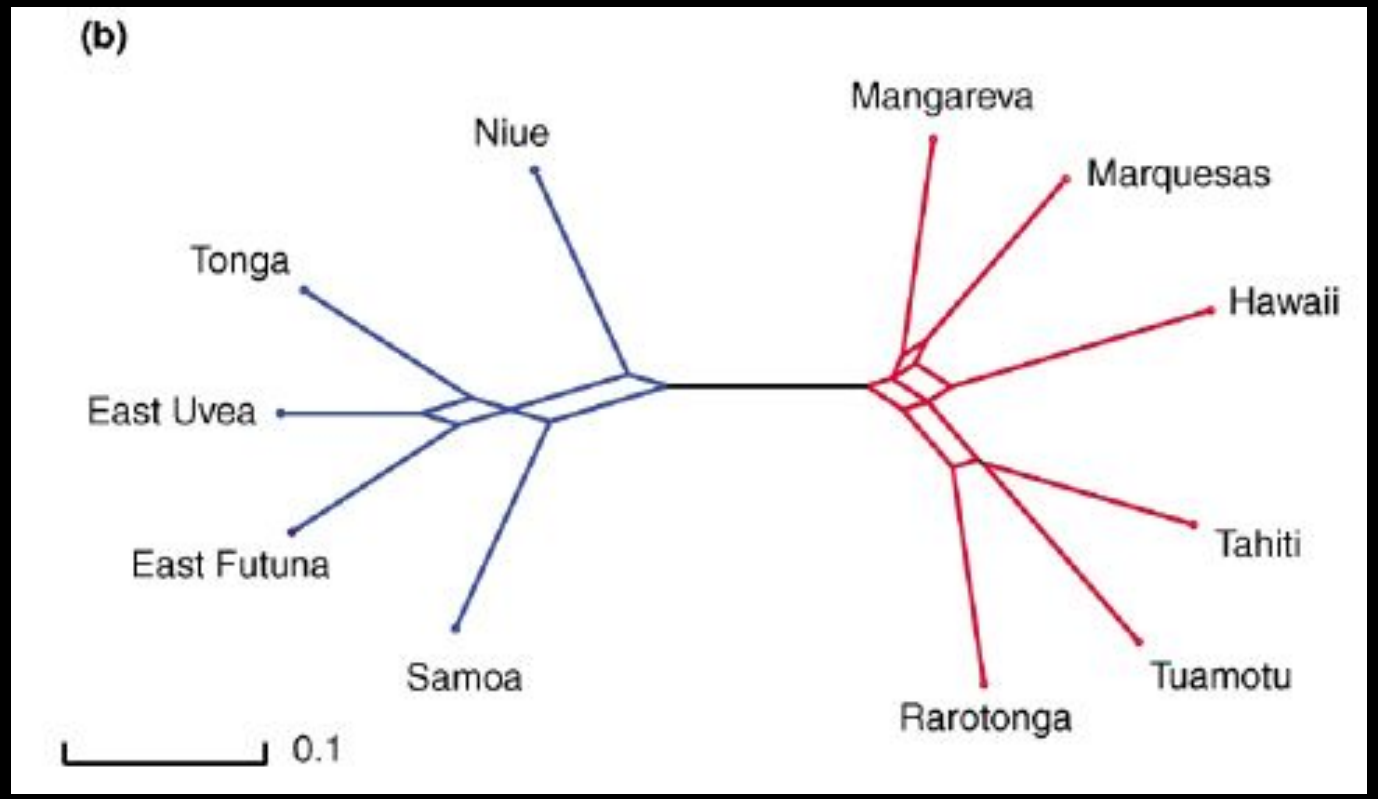
- Therefore, optimal number of steps (not reported) should be about 13,000

Review TRENDS in Ecology and Evolution Vol.18 No.10 October 2003 531

Untangling Oceanic settlement: the edge of the knowable

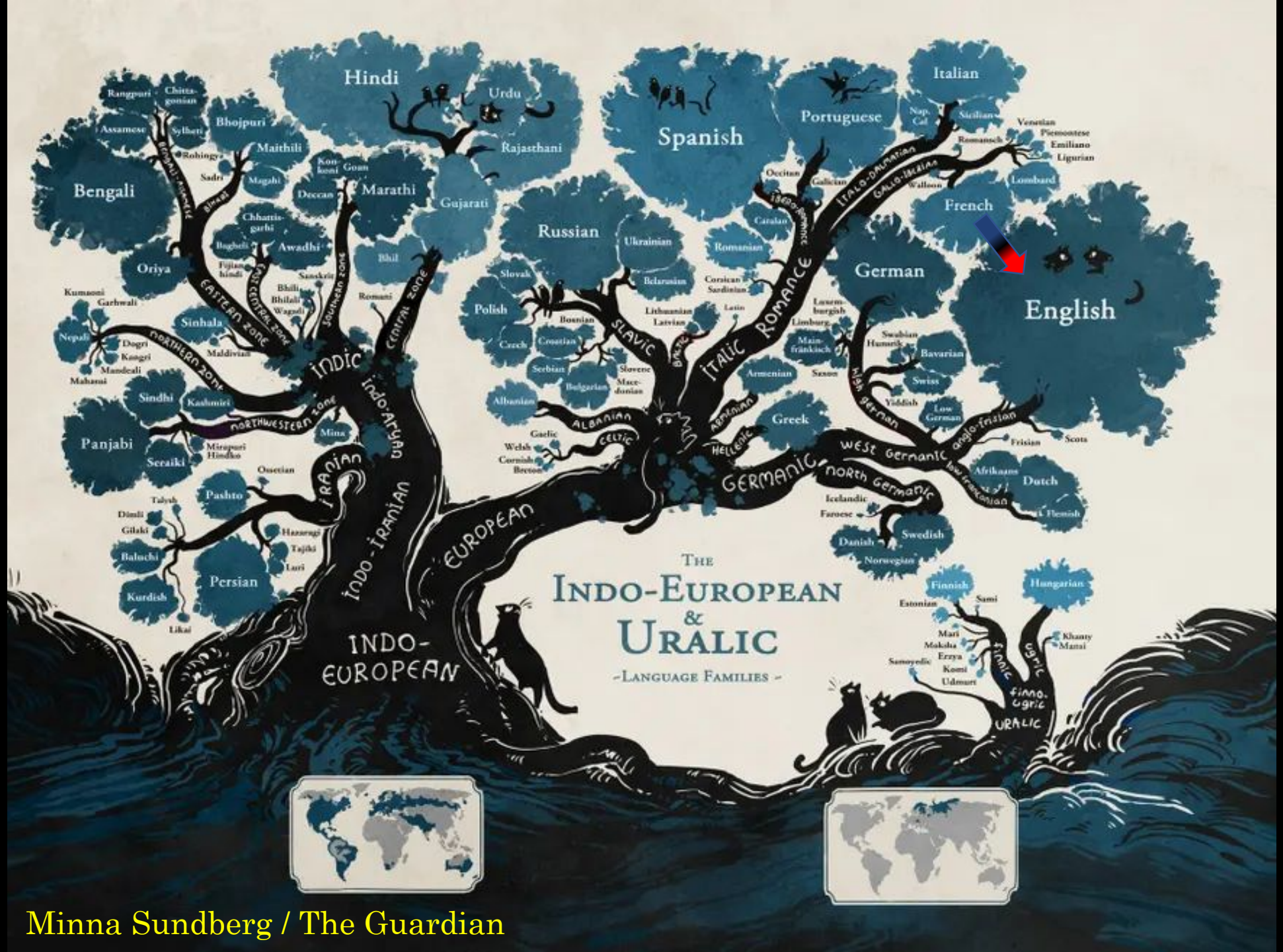
Matthew E. Hurles[1], Elizabeth Matisoo-Smith[2,3], Russell D. Gray[4] and David Penny[3,5]

West Polynesia
East Polynesia



Splits graph

Significant signals that conflict with the canonical tree

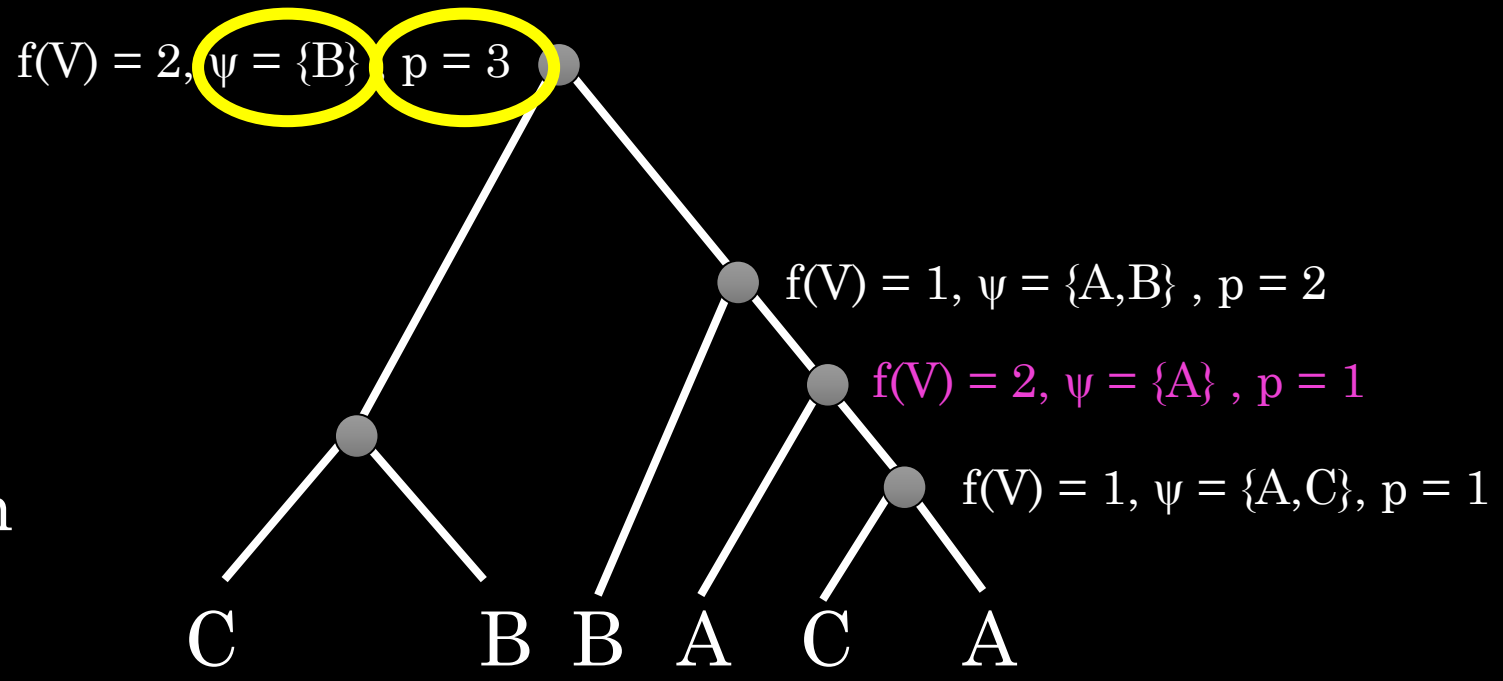


Minna Sundberg / The Guardian

Remember: Parsimony

- Find the tree that minimizes the number of changes

- We use Fitch's algorithm to compute the minimum number of changes necessary to explain the distribution of characters on a given tree, and tell us what the most-plausible characters are



Problems with Parsimony

(1) Not all alignment sites are informative

Unless it can assign different scores to different trees, a given alignment column is not parsimoniously informative

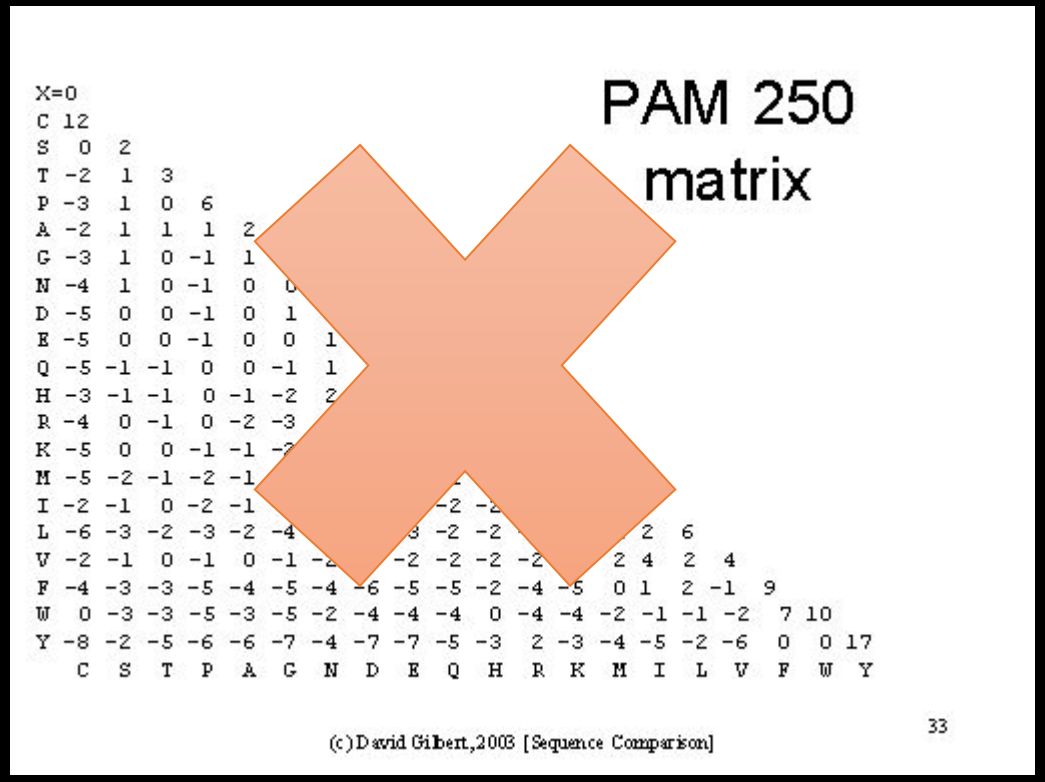
```
1 ACGTA
2 AGTGA
3 AGCCG
4 AGCAG
```

Favours ((1,2),(3,4))
over ((1,3),(2,4))
and ((1,4),(2,3))

Other sites say nothing!

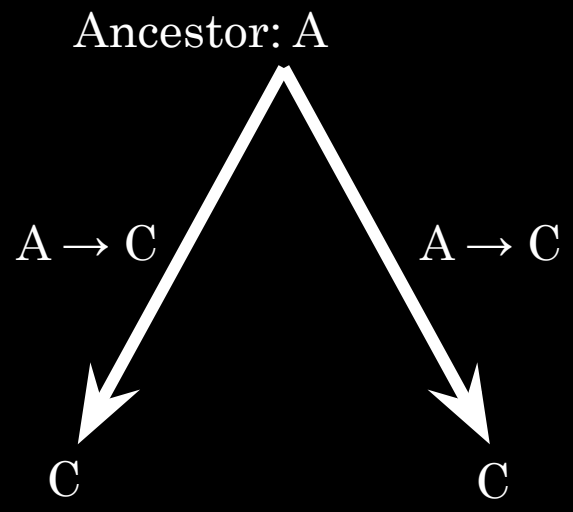
(2) Parsimony treats all changes equally

Parsimony is “model-free”, so there is no distinction between frequent and infrequent changes



(3) Homoplasies

- Sequences mutate over time
- But sometimes sequences on separate branches will independently mutate to the same nucleotide or amino acid



(3 cont'd) Long Branch Attraction

- Branches that accumulate many changes (e.g. parasites, mice) will share many homoplasies, and appear to be more similar than they really are

Parsimony: Summary

- Relatively easy (though potentially time-consuming) to use and understand
- The basic principle (the simplest explanation is the best) is attractive but not necessarily correct
- The lack of an explicit model can be an *advantage* or a serious *disadvantage*
- Throwing away uninformative alignment columns is not necessarily ideal