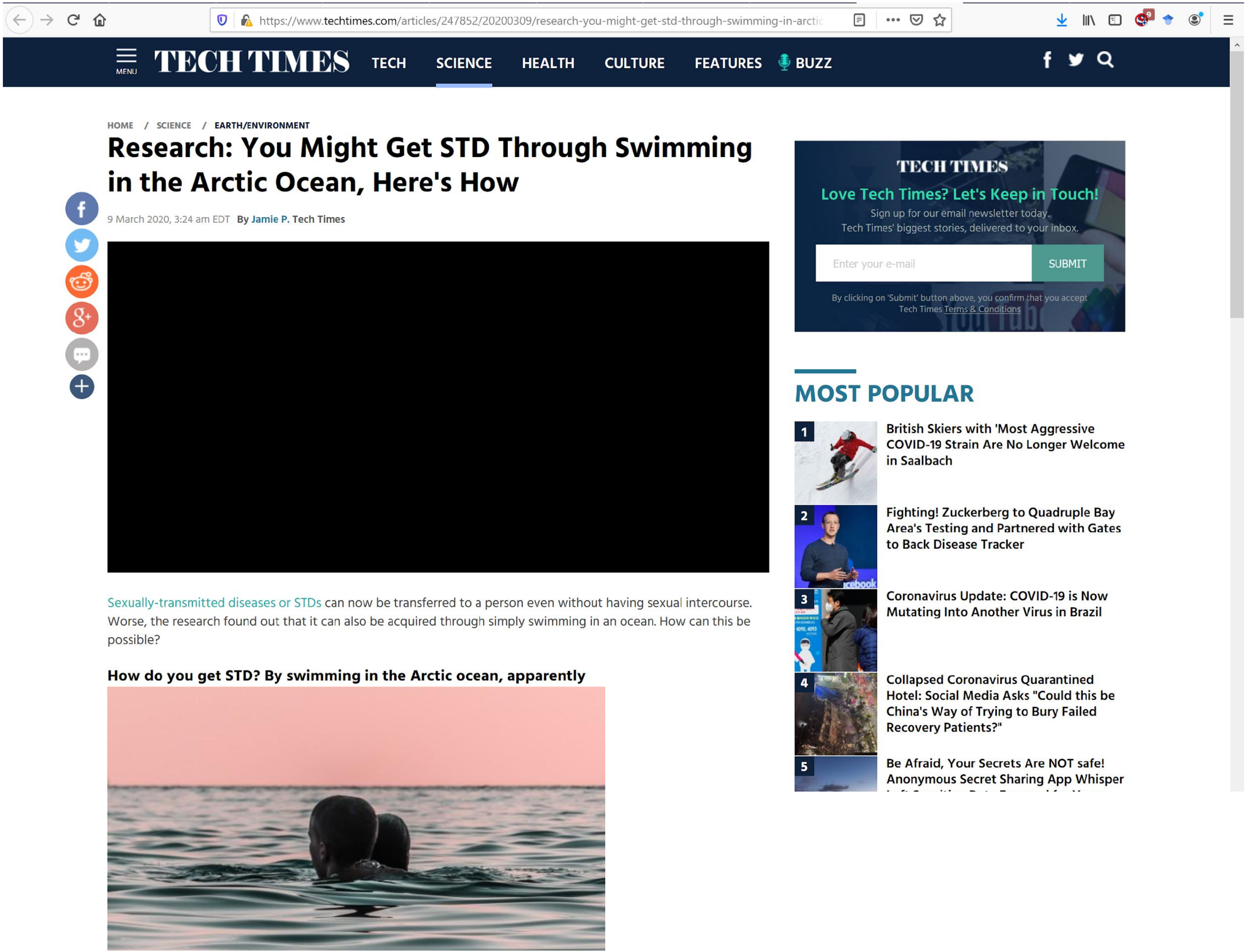


https://www.techtimes.com/articles/247852/20200309/research-you-might-get-std-through-swimming-in-arctic
MENU
TECH TIMES
TECH
SCIENCE
HEALTH
CULTURE
FEATURES
BUZZ

HOME / SCIENCE / EARTH/ENVIRONMENT

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

9 March 2020, 3:24 am EDT By Jamie P. Tech Times

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

MOST POPULAR

1 British Skiers with 'Most Aggressive COVID-19 Strain Are No Longer Welcome in Saalbach

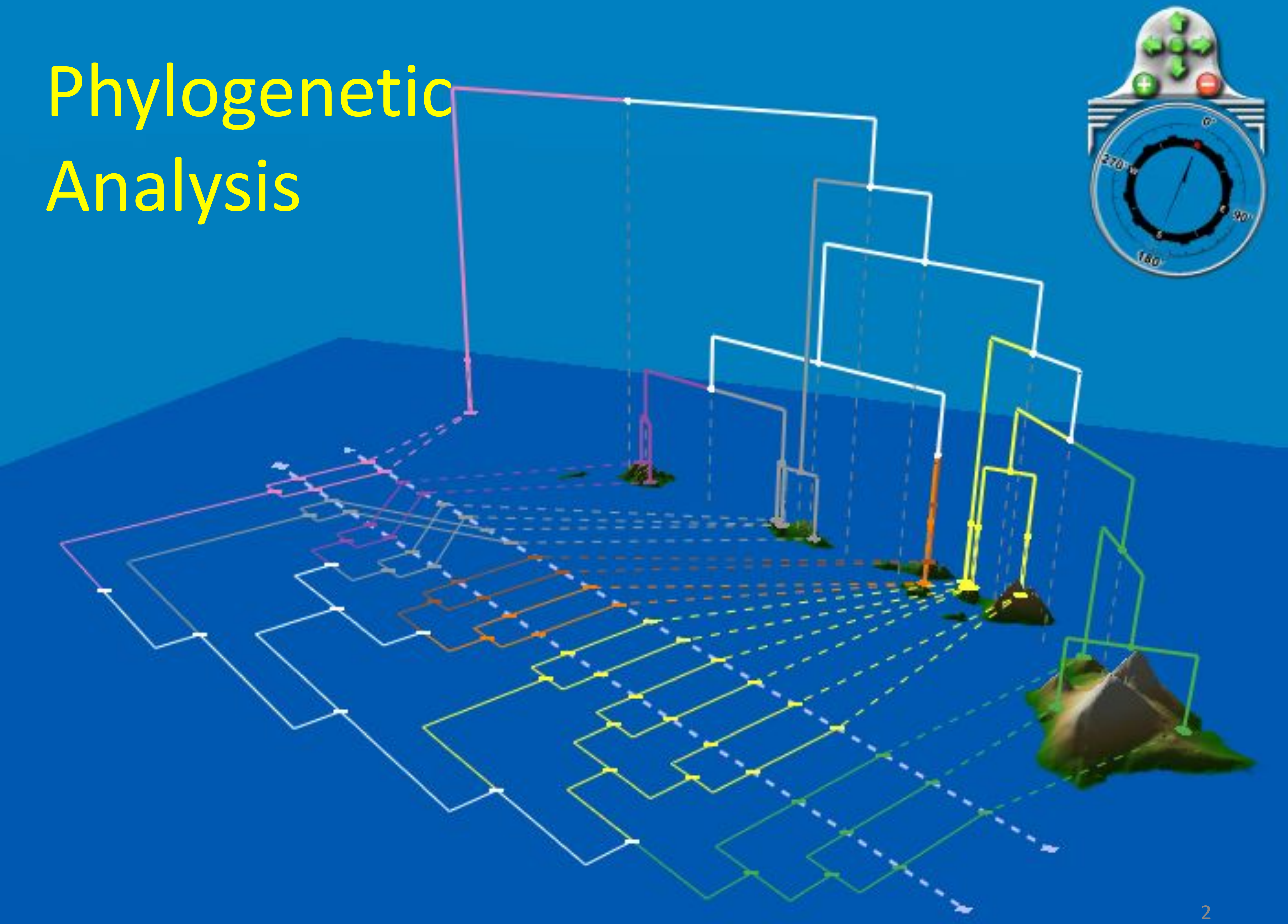
2 Fighting! Zuckerberg to Quadruple Bay Area's Testing and Partnered with Gates to Back Disease Tracker

3 Coronavirus Update: COVID-19 is Now Mutating Into Another Virus in Brazil

4 Collapsed Coronavirus Quarantined Hotel: Social Media Asks "Could this be China's Way of Trying to Bury Failed Recovery Patients?"

5 Be Afraid, Your Secrets Are NOT safe! Anonymous Secret Sharing App Whisper

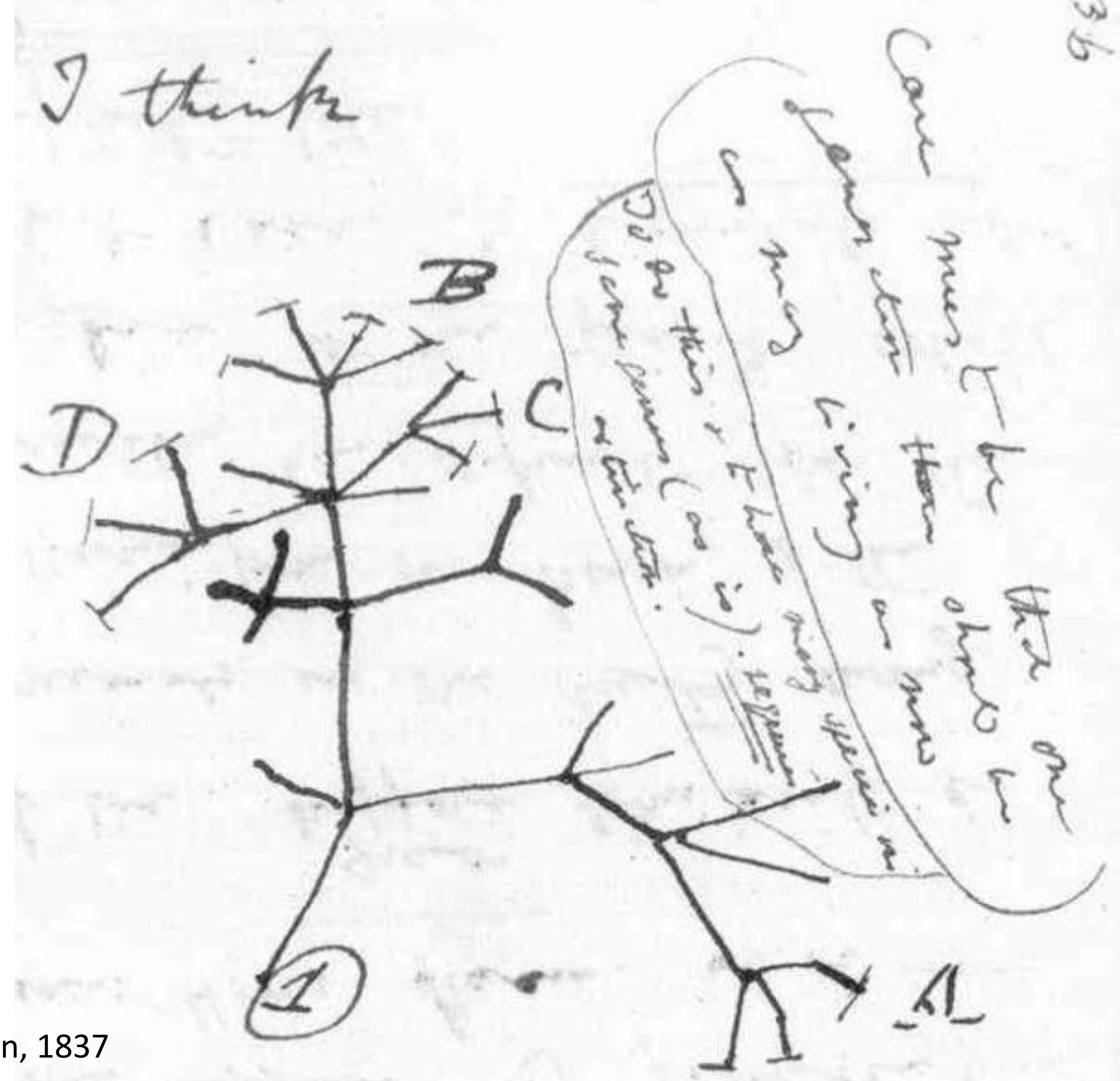
Phylogenetic Analysis



The Point

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

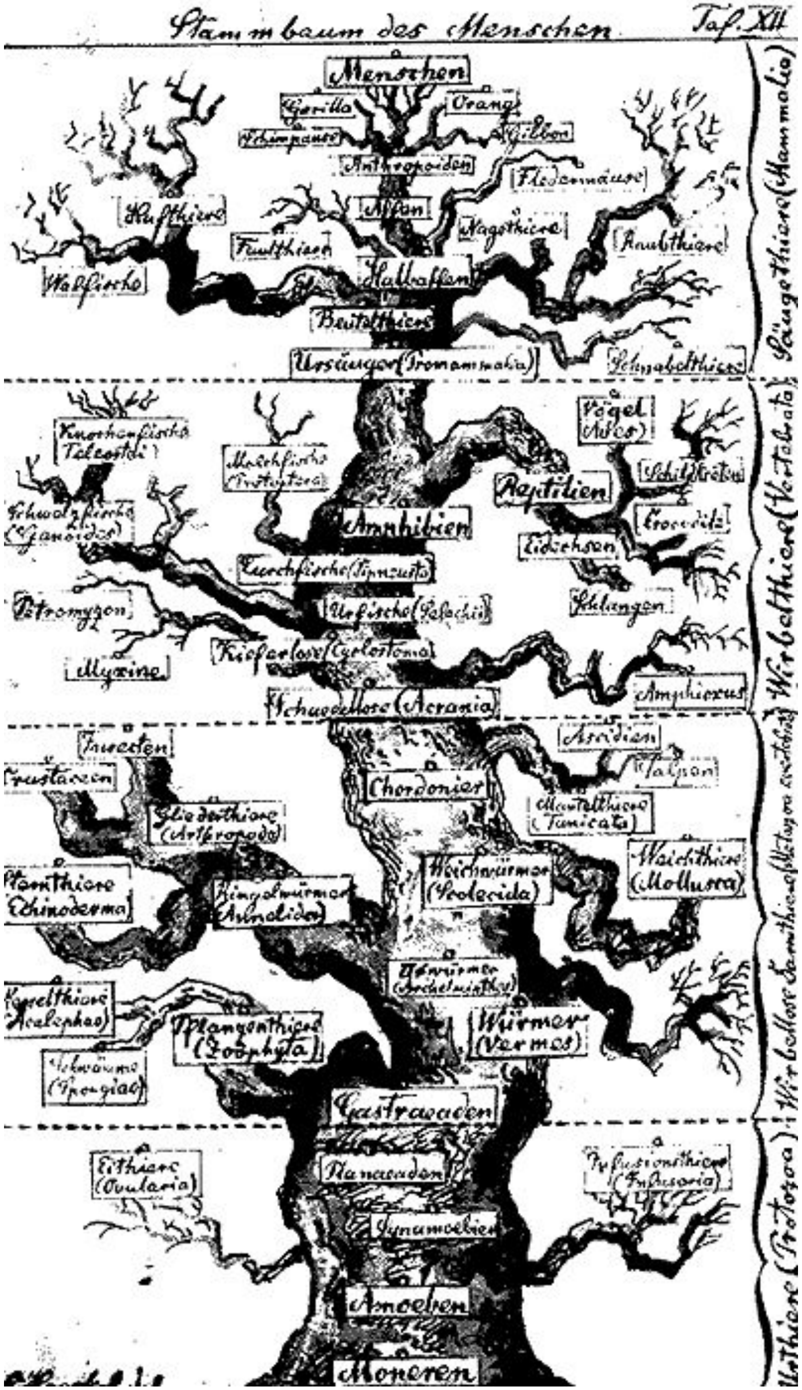
Often means aligned sequences (with homologous residues in columns)



I think

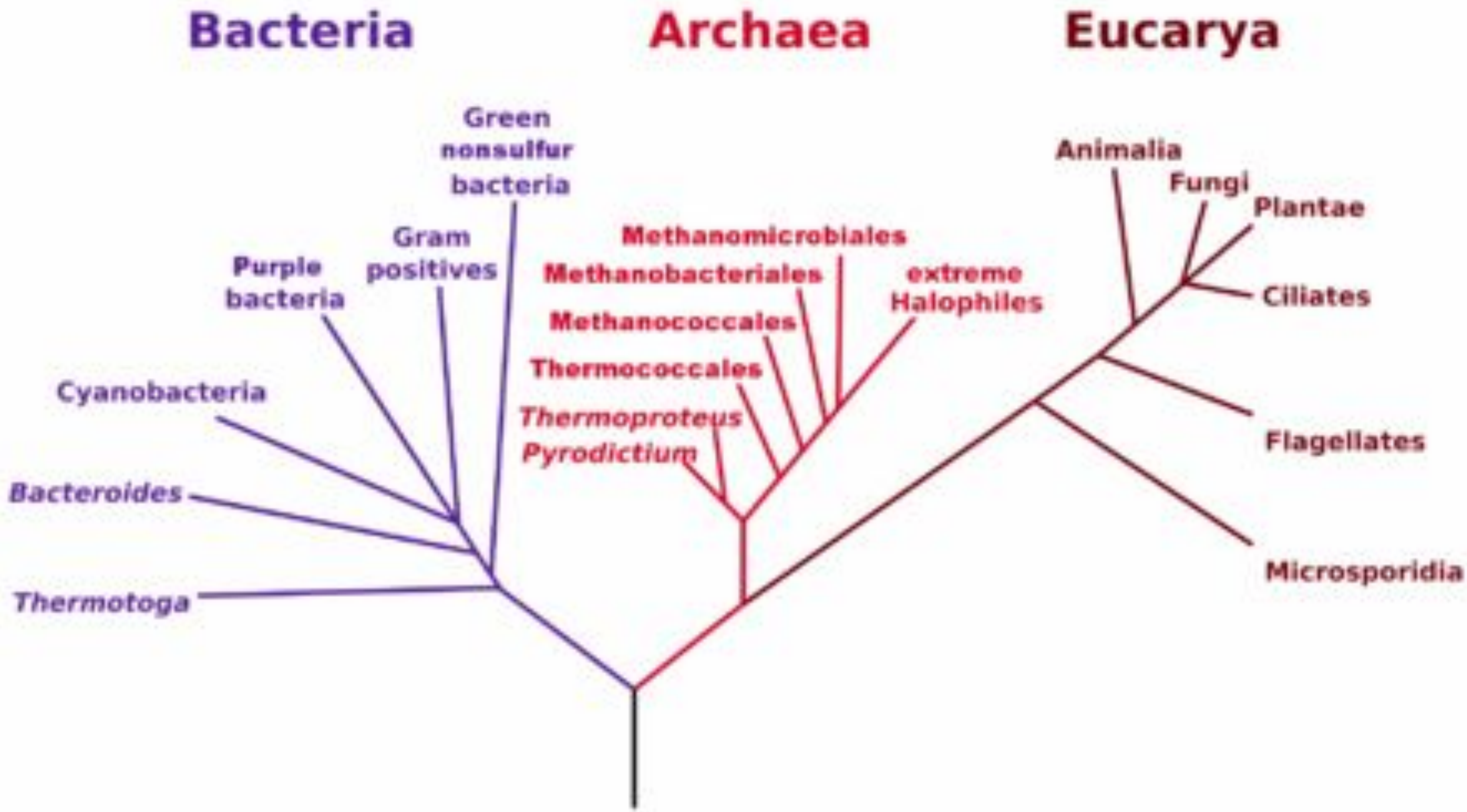
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Darwin, 1837

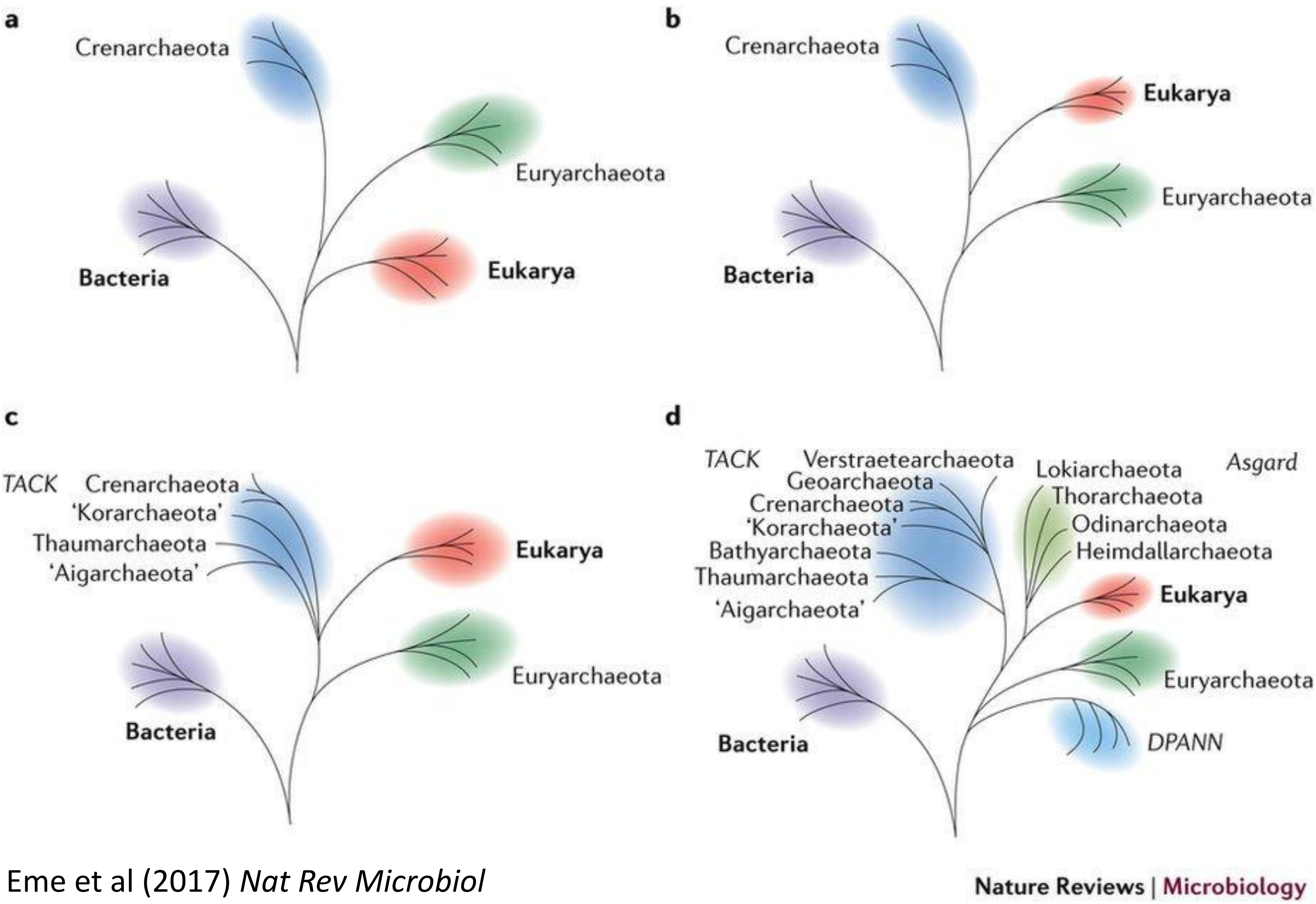


Haeckel, 1874
“Pedigree of Man”

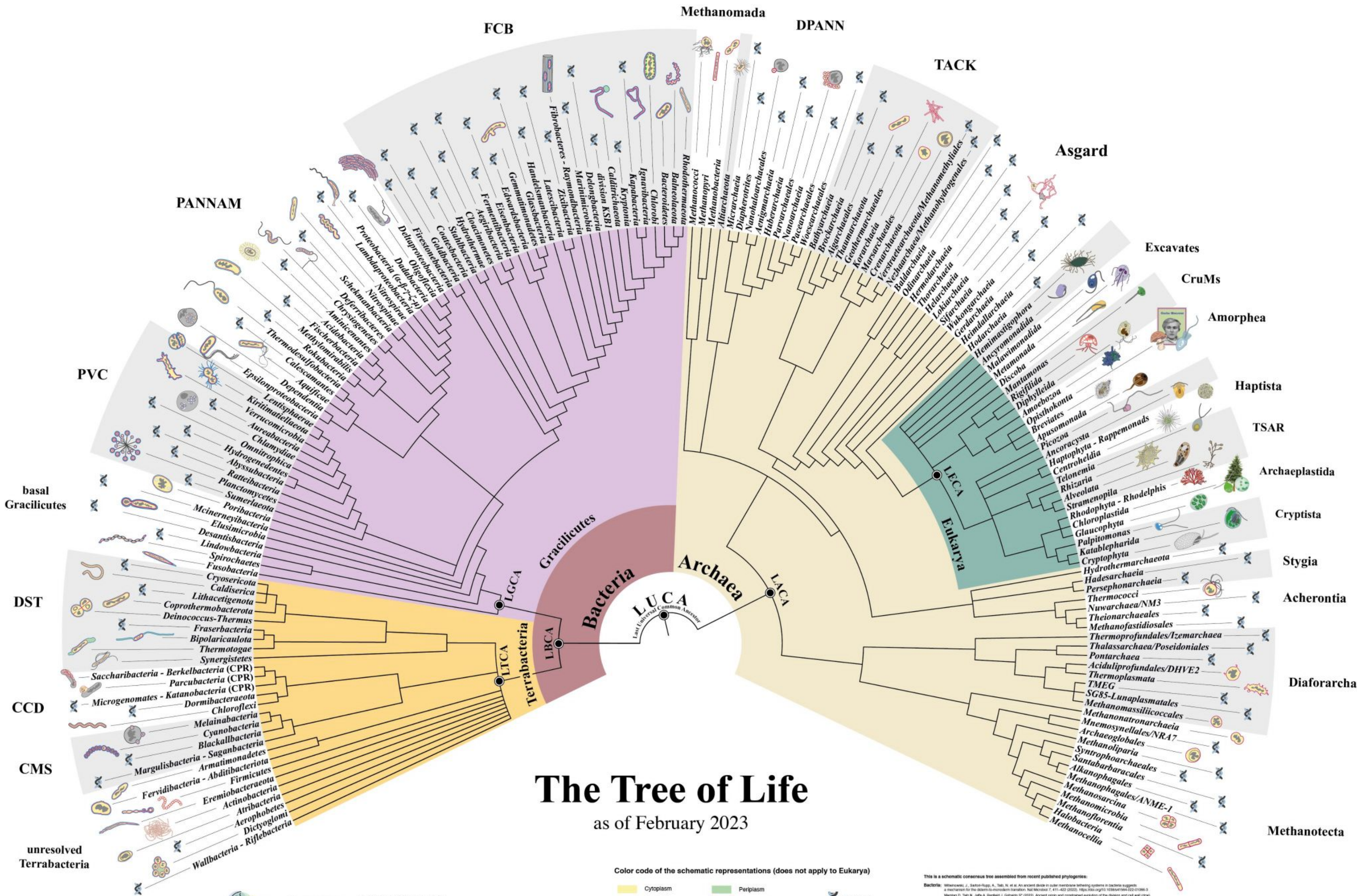
Phylogenetic Tree of Life



Fox and Woese (1977) and many, many others



Eme et al (2017) *Nat Rev Microbiol*



The Tree of Life

as of February 2023

Color code of the schematic representations (does not apply to Eukarya)

🟨	Cytoplasm	🟩	Periplasm	🧬	Uncultured
■	Nucleoid	🟦	Outer membrane		
🟪	Cytoplasmic membrane	■	Layer of unknown composition		

Evolutionary Biology of the Microbial Cell

INSTITUT PASTEUR

Author : Basile Beaud, PhD student
Director : Pr. Simonetta Gribaldo
Unit Evolutionary Biology of the Microbial Cell, Institut Pasteur, Paris 75015, France

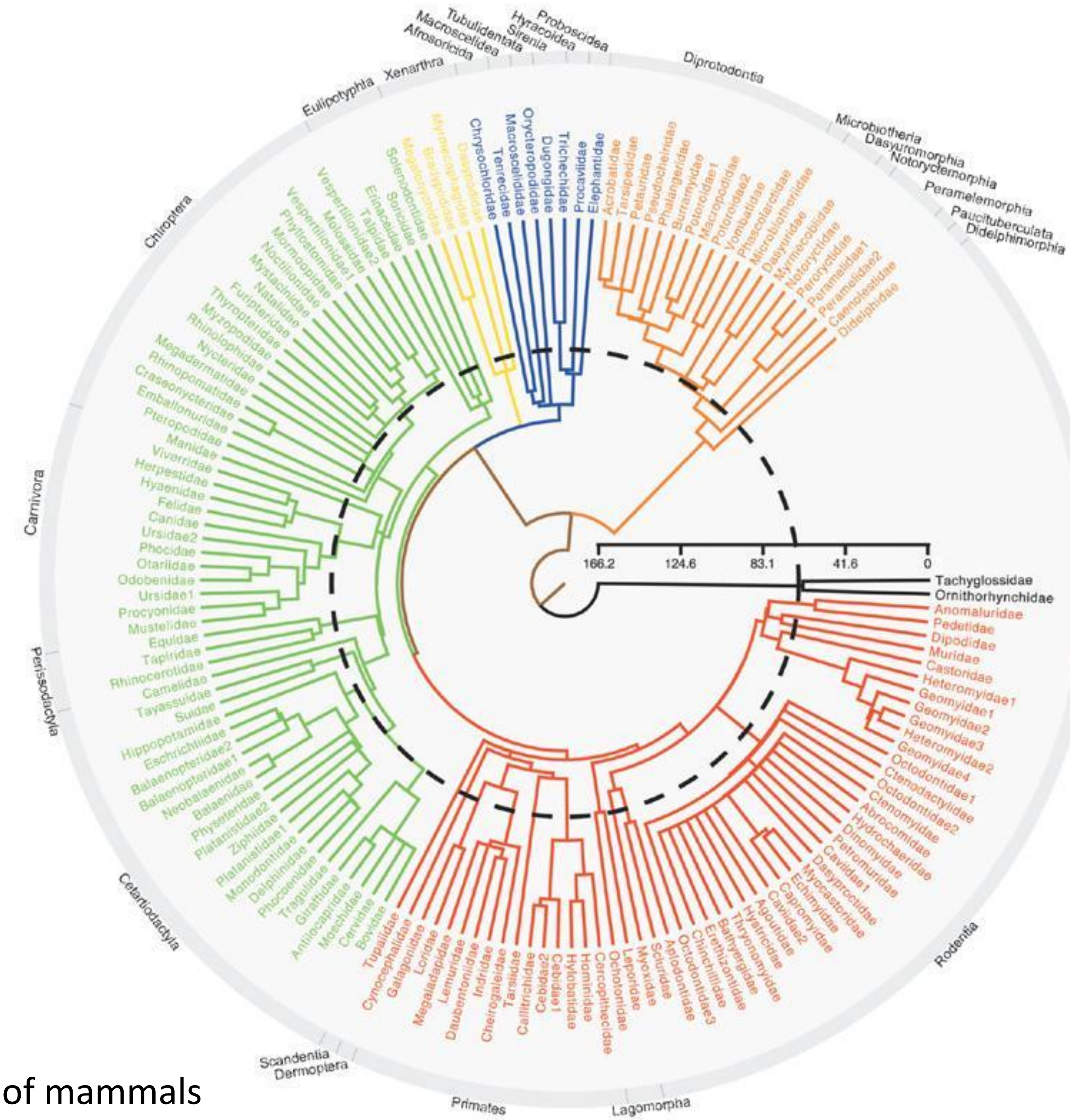
This is a schematic consensus tree assembled from recent published phylogenies:

Bacteria: Witwinowski, J., Sartori-Rupp, A., Taib, N. et al. An ancient divide in outer membrane tethering systems in bacteria suggests a mechanism for the diderm-to-monoderm transition. Nat Microbiol 7, 411–422 (2022). https://doi.org/10.1038/s41564-022-01066-3

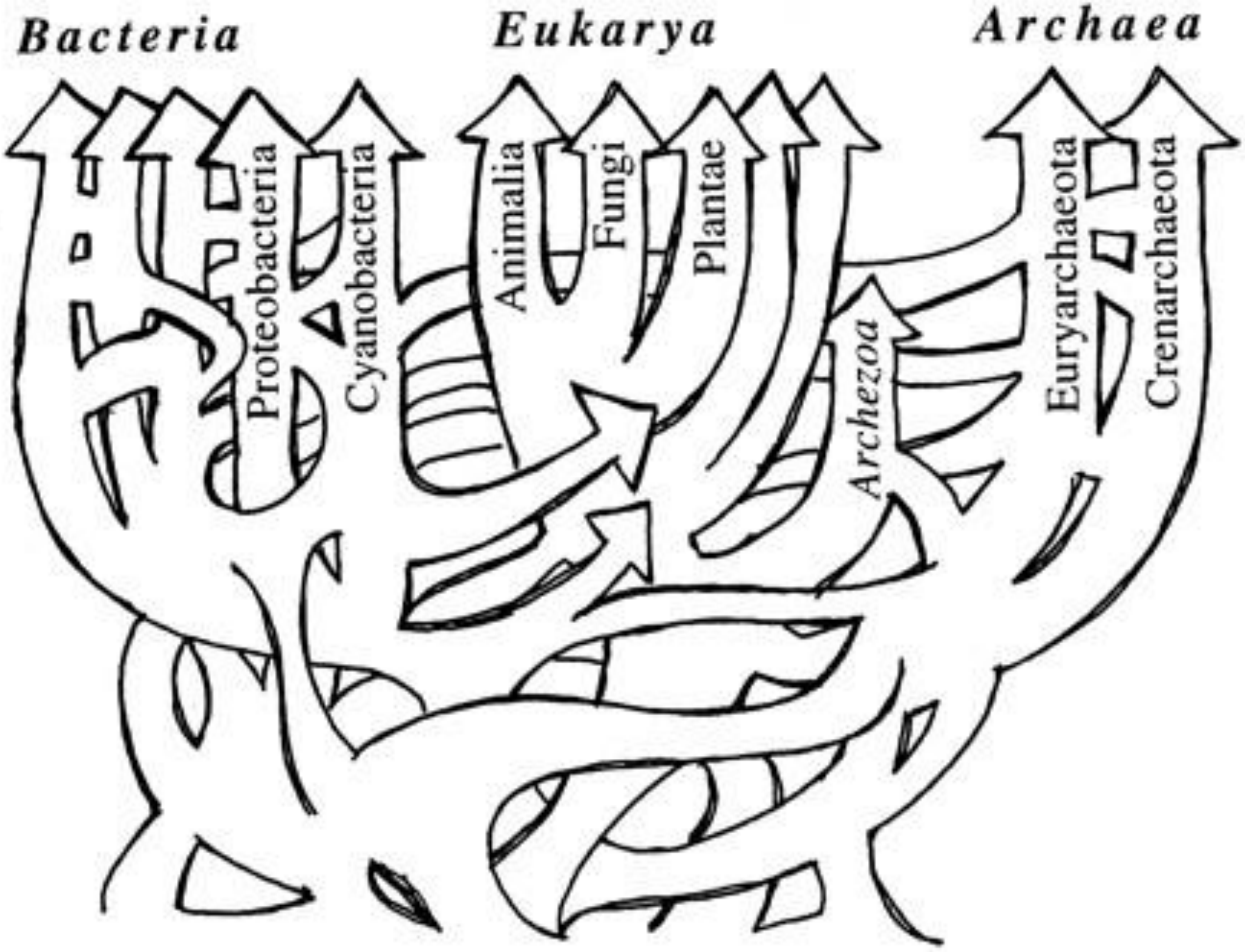
Megrian D, Taib N, Jaffe A, Banfield J, Gribaldo S* (2022). Ancient origin and constrained evolution of the division and cell wall (dcw) gene cluster across Bacteria. Nat Microbiol (in press)

Archaea: Garcia PS, Gribaldo S, Borrel G. Diversity and Evolution of Methane-Related Pathways in Archaea. Annu Rev Microbiol. 2022 Sep 8;76:727-755. doi: 10.1146/annurev-micro-041020-024935. Epub 2022 Jun 27. PMID: 35759872.

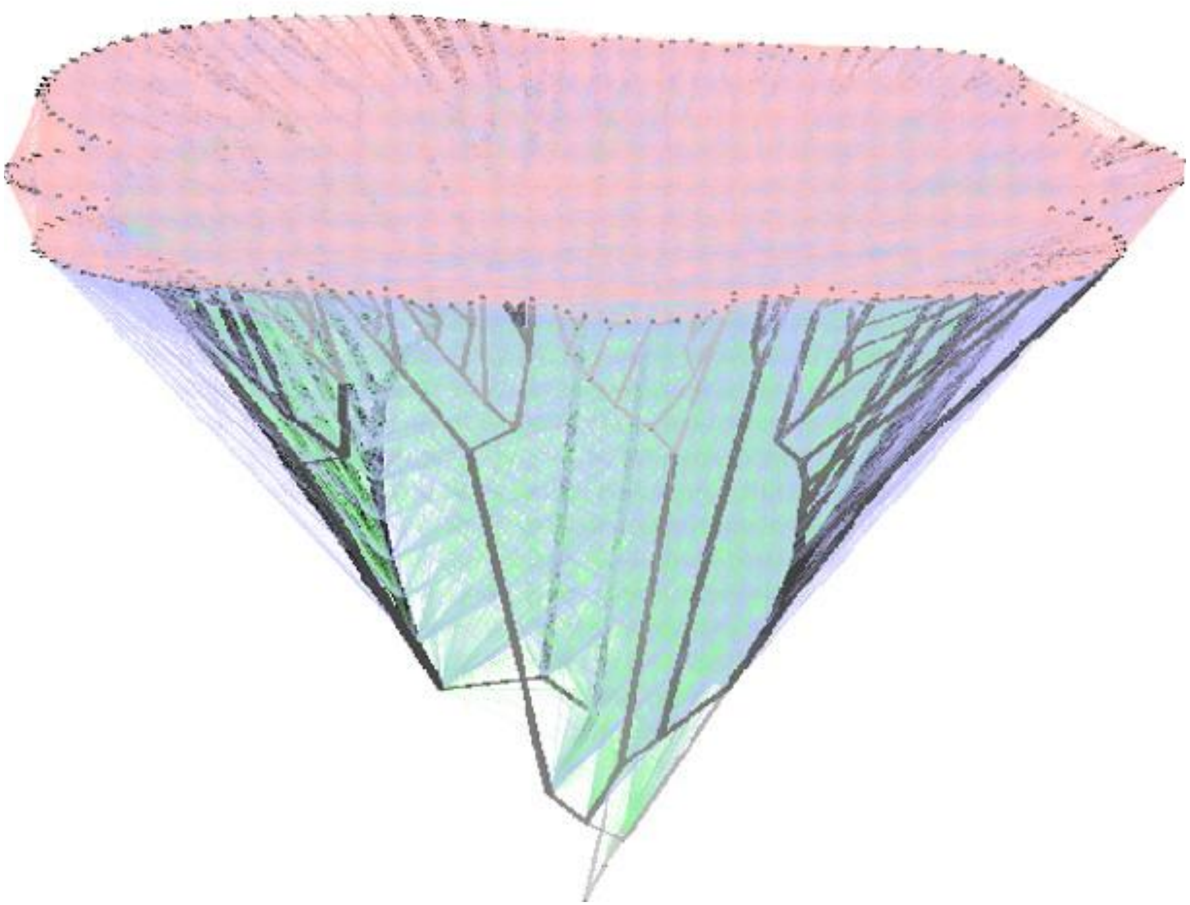
Eukarya: Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2019). The new tree of eukaryotes. Trends in Ecology & Evolution, https://doi.org/10.1016/j.tree.2019.08.008



(Super)tree of mammals
Bininda-Emonds et al., *Nature* (2007)



The tree of life is a network
Doolittle (1999) *Sci Am*



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

The problem

- **How** to build trees properly is not necessarily obvious, and depends on a large number of factors

- **Modeling** sequence similarity is challenging – evolution deals us a confusing hand

- **Searching** tree space can be a nightmare (again, exact vs. heuristic approaches)

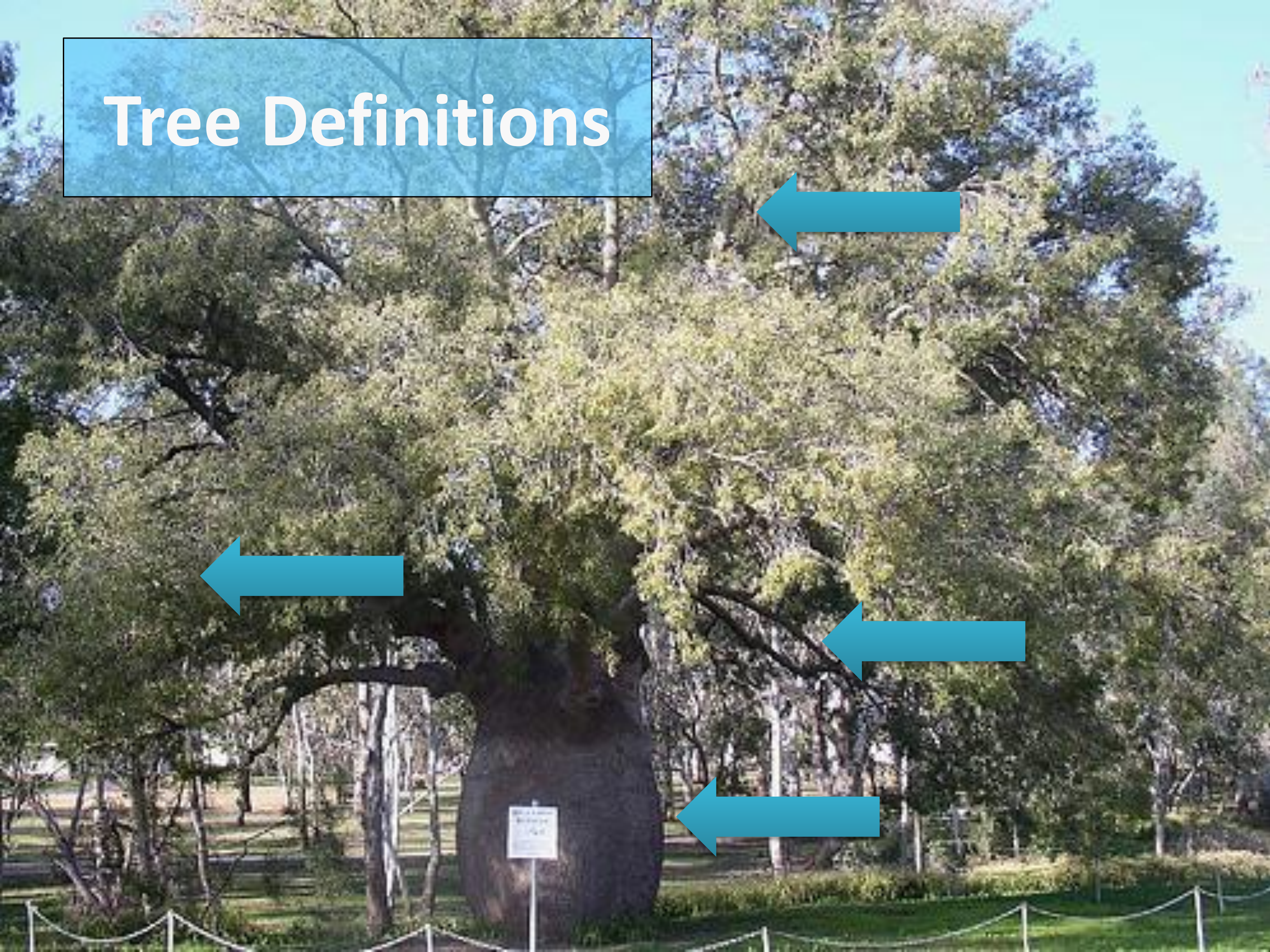
- Many problems in evolution **cannot be effectively represented using trees**

Phylogenetics is multidisciplinary

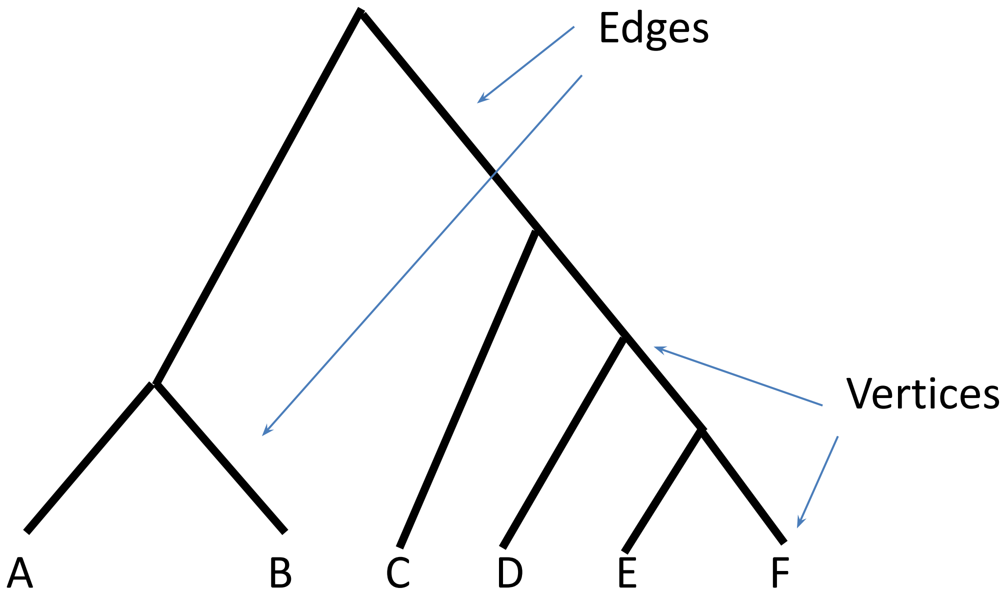
- First tree - Chemist (Linus Pauling)
- 1950s - Physicist (Margaret Dayhoff)
- 1960s - Statisticians
- 1970s - Computer Scientists
- Throughout - Biologists

- Lots of redundant terminology!

Tree Definitions

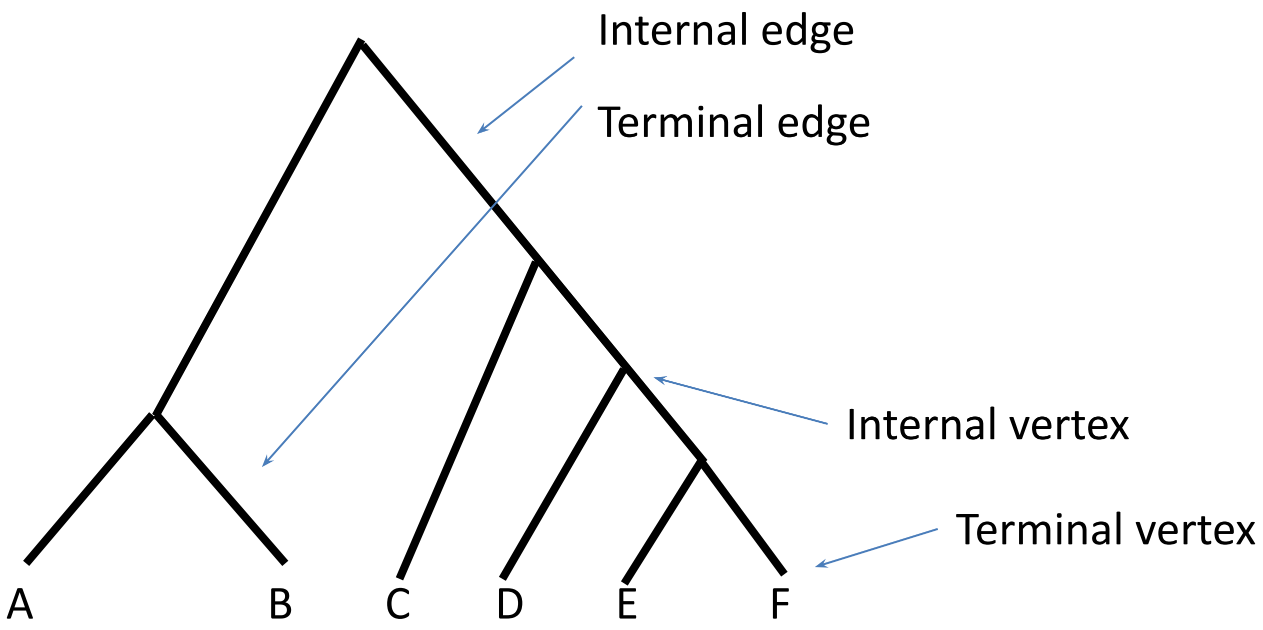


Tree Anatomy



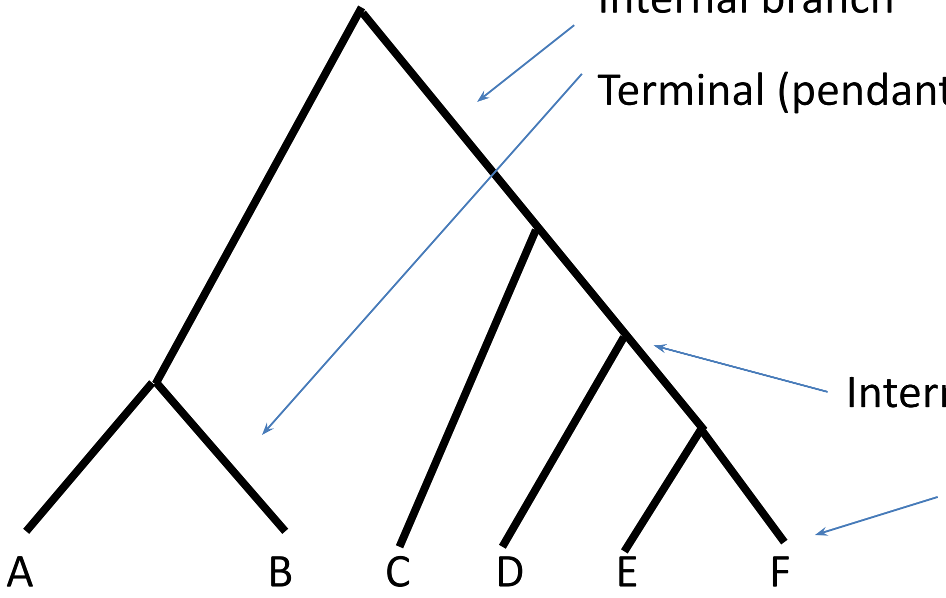
Trees can be described using the same terminology as graphs

Tree Anatomy



We distinguish between **internal** and **terminal** features

Tree Anatomy



Internal branch

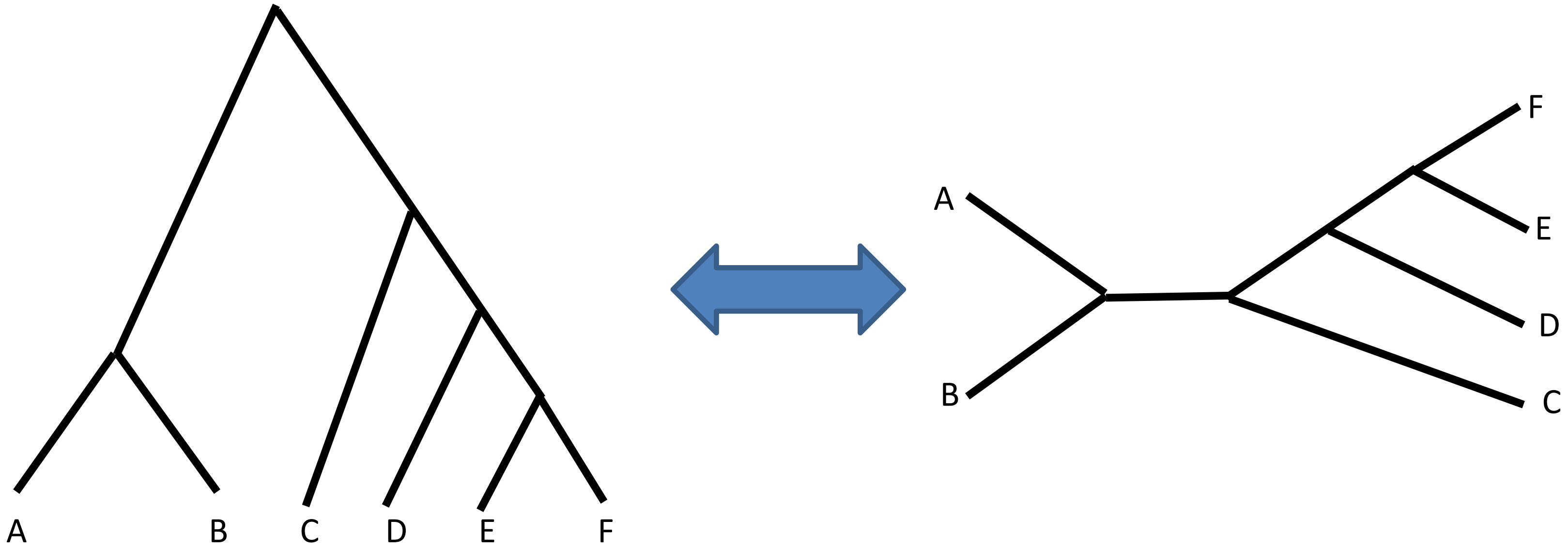
Terminal (pendant) branch

Internal **node** (last common ancestor)

Terminal (leaf) **node**

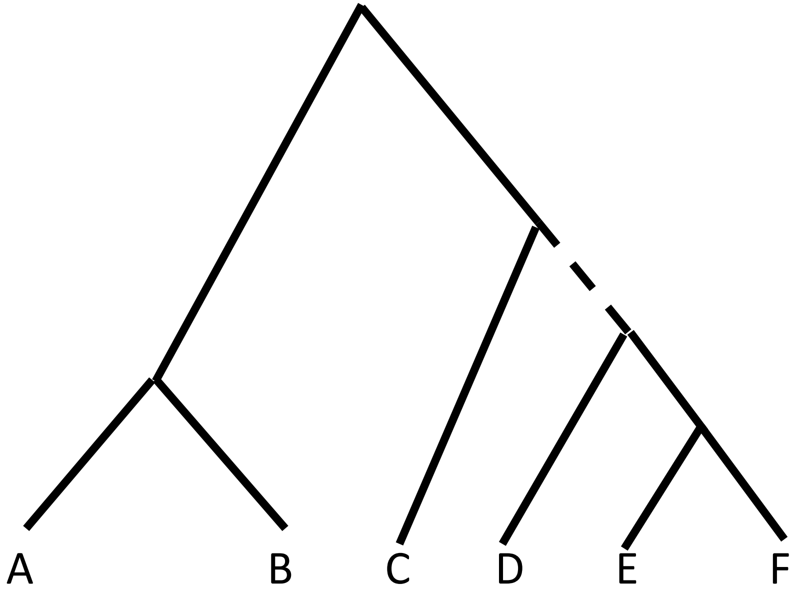
Some terms are used interchangeably

Rooted vs Unrooted Trees



Most methods (including parsimony) generate **unrooted** trees

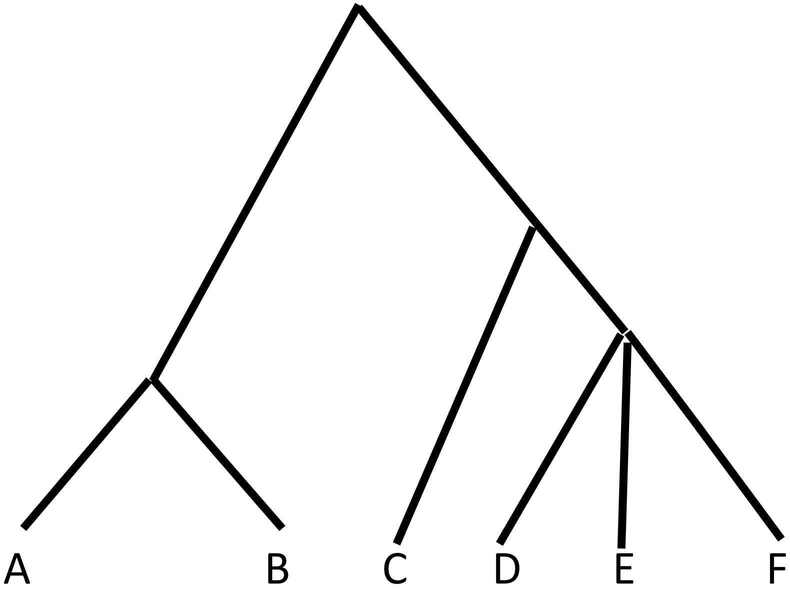
Tree splits (bipartitions)



(ABC | DEF)

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

Multifurcating Nodes

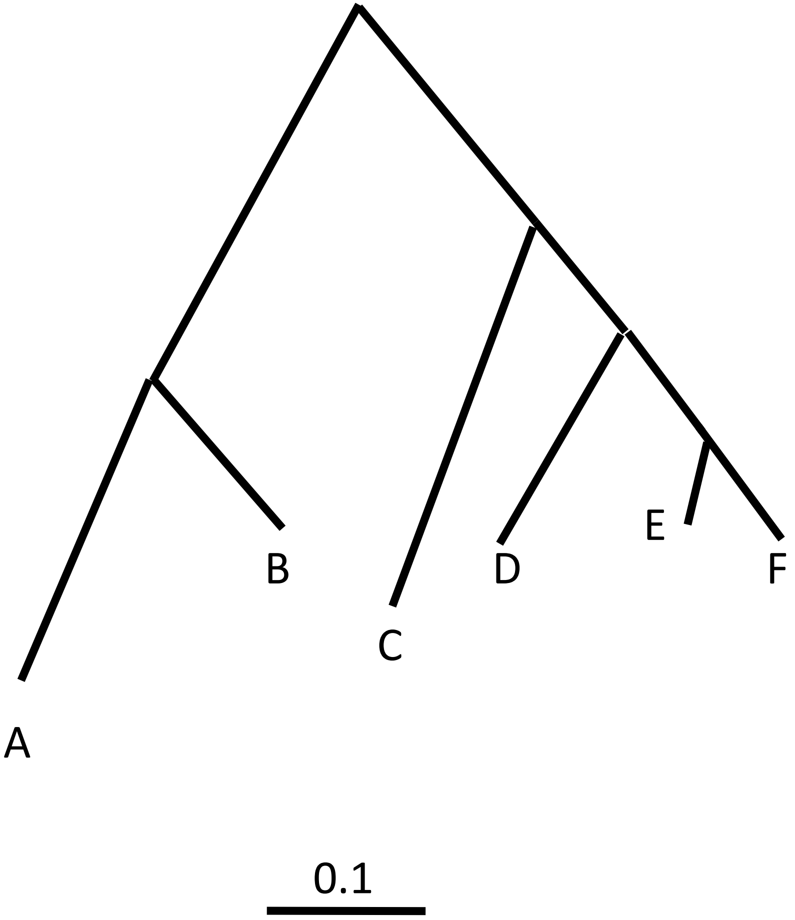


We may *collapse* a node in the tree for one of two reasons:

- 'Hard' polytomy (really a 3-way split)
-Lack of statistical support for any pairwise grouping

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

Branch lengths

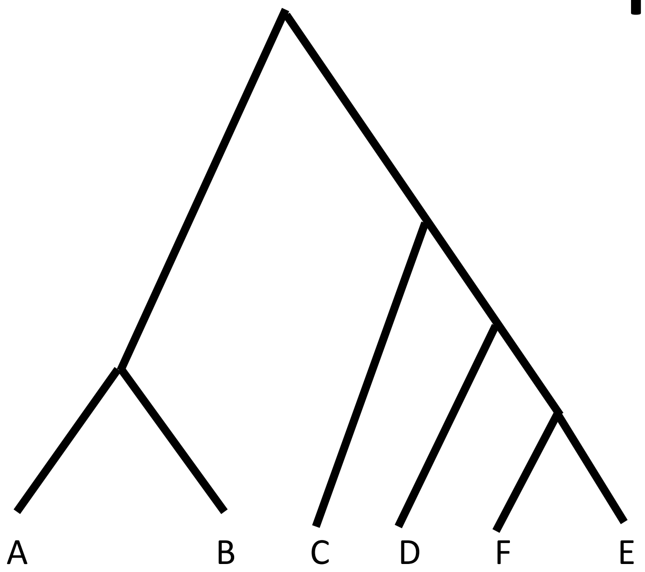


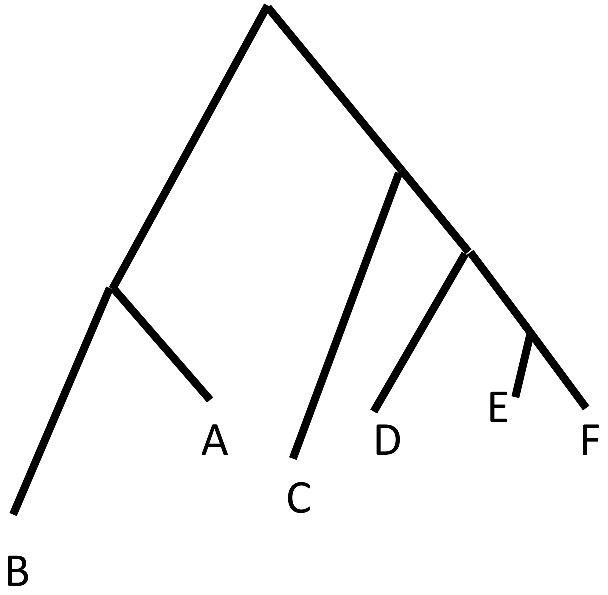
What (if anything) do branch lengths represent?

- Time?
- Sequence change?

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

Tree Shape



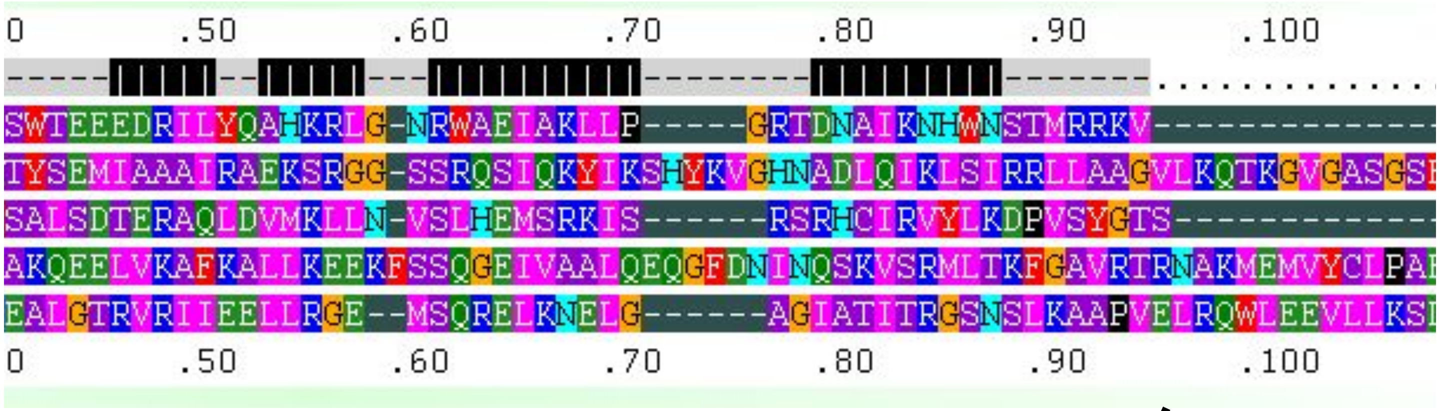


In general (and for the purposes of this course), the *shape* of a tree refers to its branching order, **not** to branch lengths

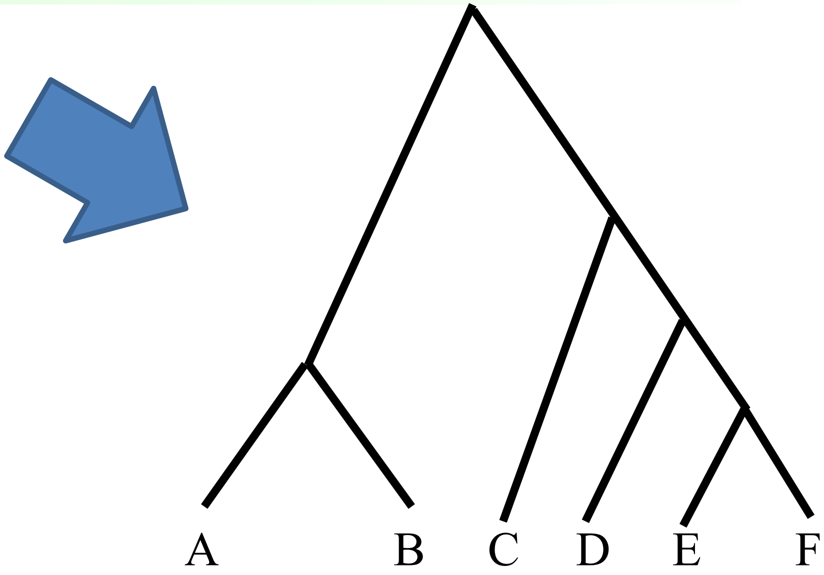
So the two trees on the left have the same shape

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

Nextstrain Intermission

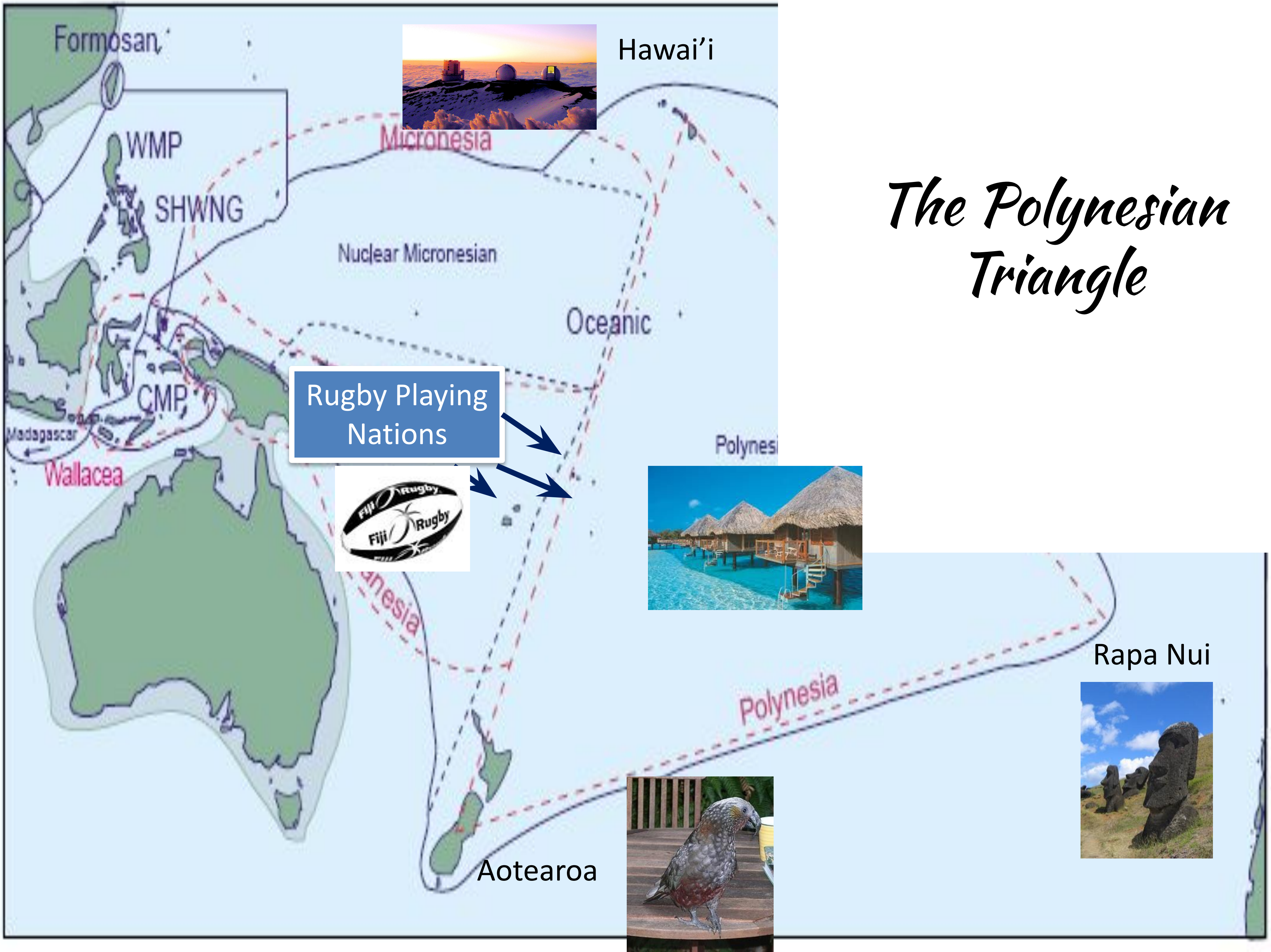


```
0       .50       .60       .70       .80       .90       .100
-----|||||--|||||---||||||||||--------|||||||||-------.............
SWTEEEDRILYQAHKRLG-NRWAEIAKLLP-----GRTDNAIKNHWNSTMRRKV-------------
TYSEMIAAAIRAEKSRGG-SSRQSIQKYIKSHYKVGHNADLQIKLSIRRLLAAGVLKQTKGVGASGSF
SALSDTERAQLDVMKLLN-VSLHEMSRKIS------RSRHCIRVYLKDPVSYGTS------------
AKQEELVKAFKALLKEEKFSSQGEIVAALQEQGFDNINQSKVSRMLTKFGAVRTRNAKMEMVYCLPAE
EALGTRVRIIEELLRGE--MSQRELKNELG------AGIATITRGSNSLKAAPVELRQWLEEVLLKSE
0       .50       .60       .70       .80       .90       .100
```



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

The Polynesian Triangle



Hawai’i

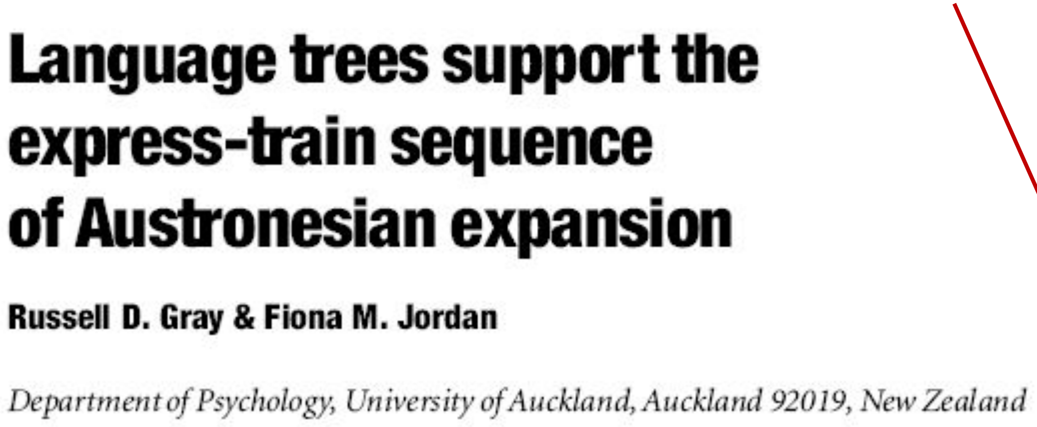
Rugby Playing Nations

Rapa Nui

Aotearoa

Words as homologous characters

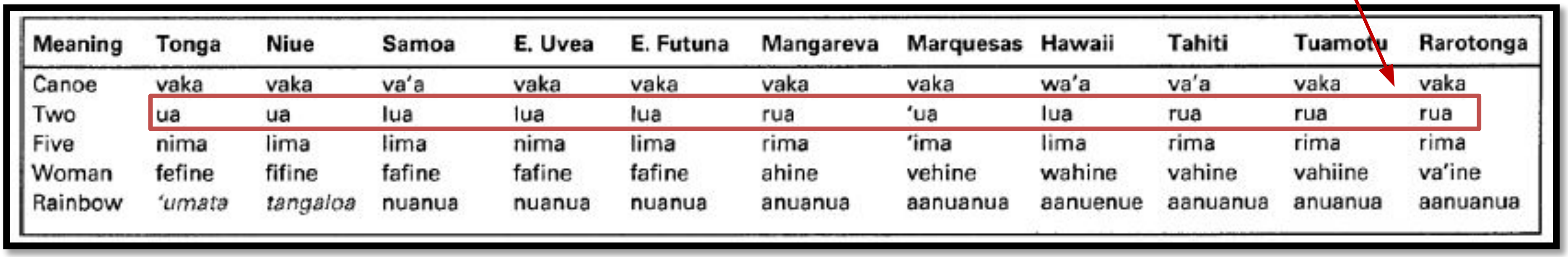
Nature, 2000



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

No collinearity constraint
(but who cares?)

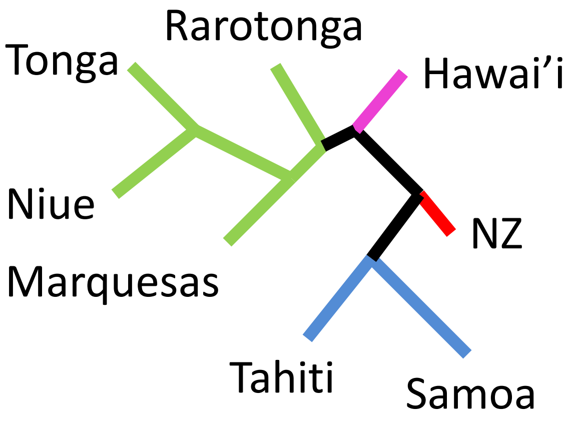
Character Convexity

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)

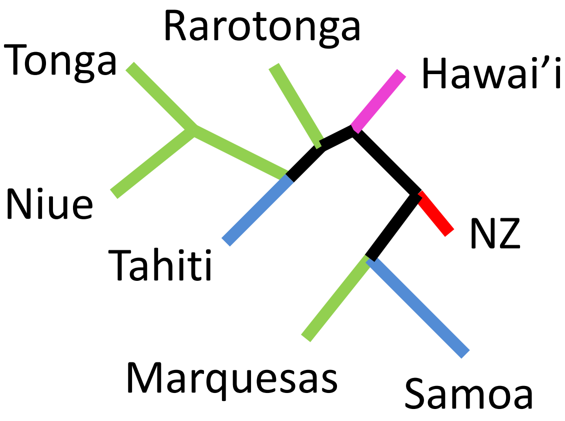
A character is *convex* on that tree if all states of that character can be partitioned to a separate 'region' of the tree

Think of it as a coloring problem!



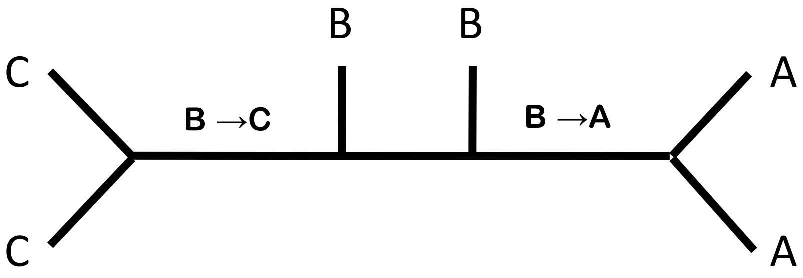
CONVEX

NOT CONVEX



What does convexity mean?

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

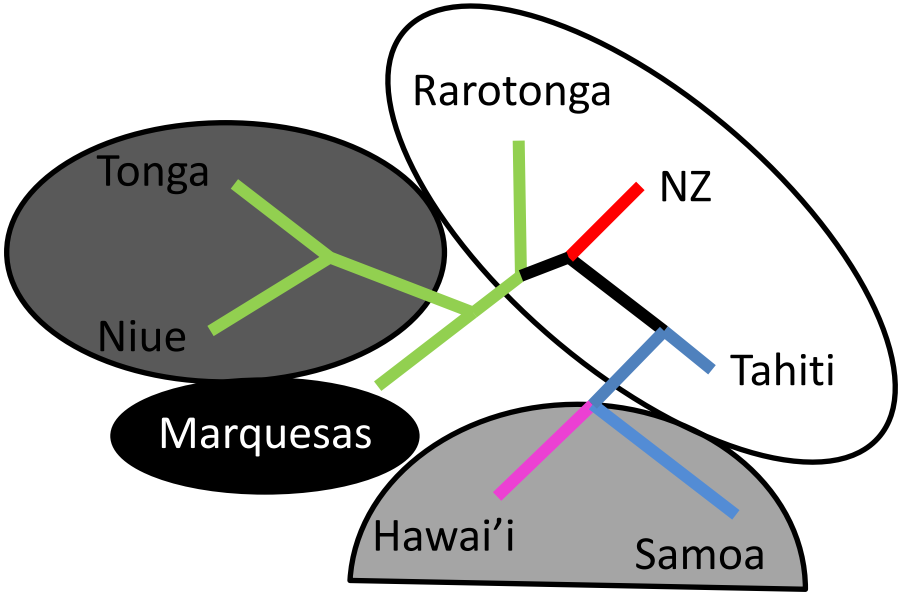


- The is the *most parsimonious* (simplest) situation

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

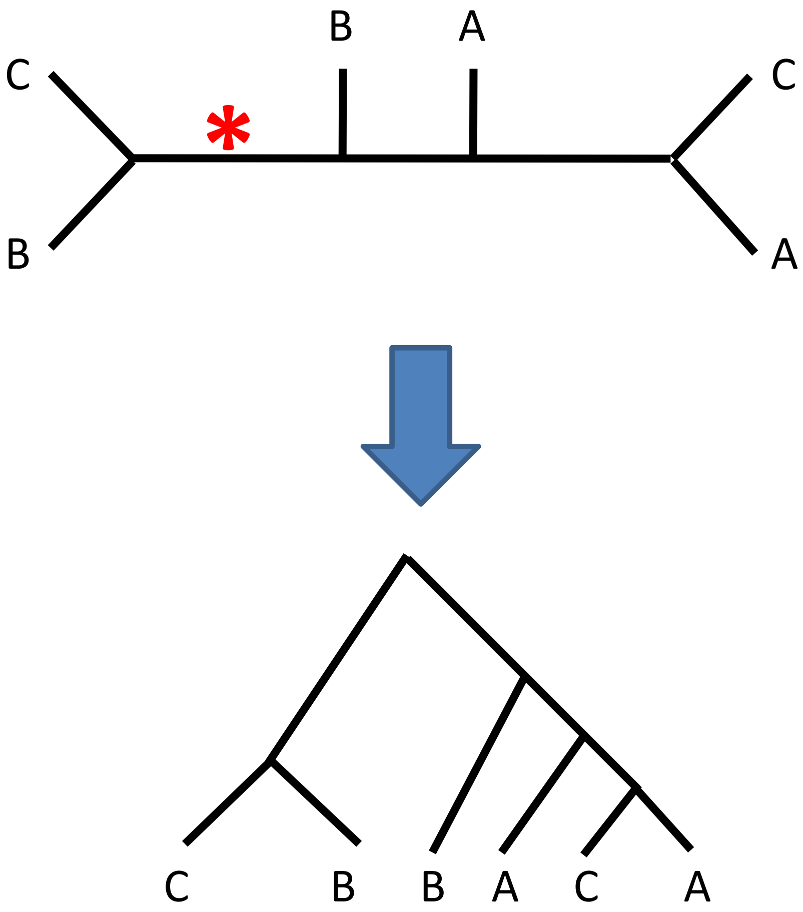
- What we typically want is the tree that minimizes the number of substitutions over *all* characters – this is the **maximum parsimony** tree

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

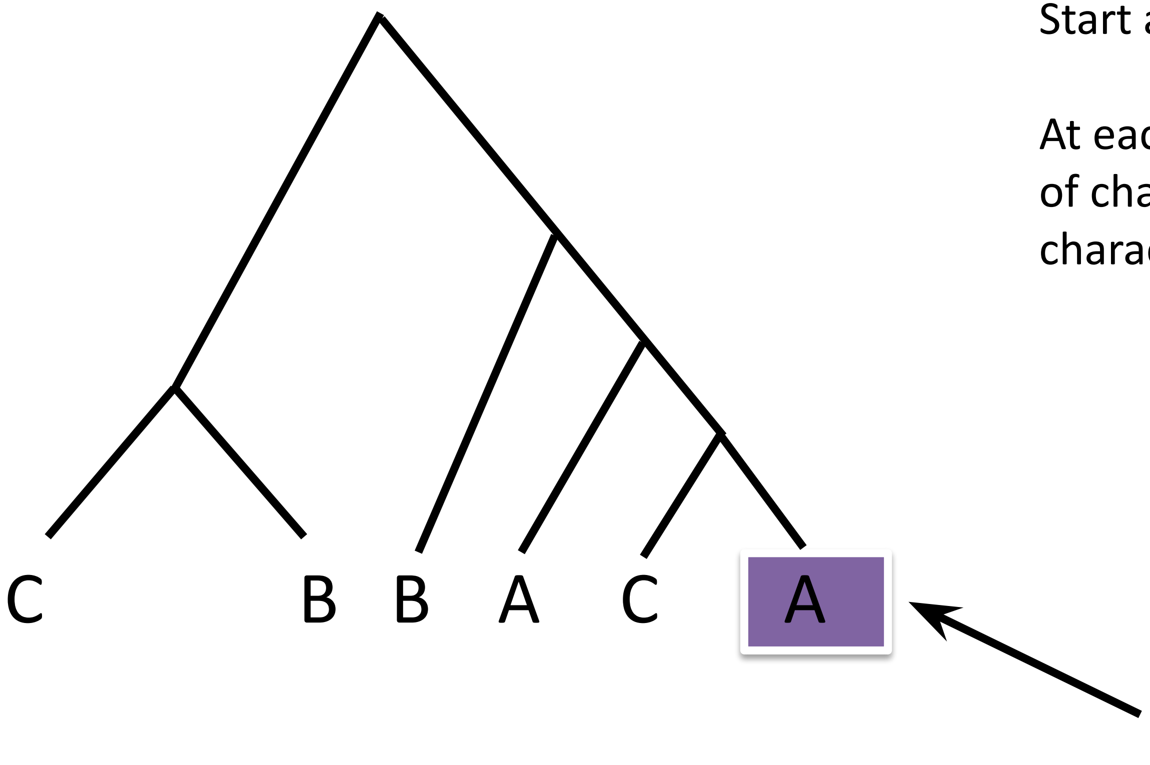
- How do we find this minimum for a single character?

Fitch-Hartigan algorithm



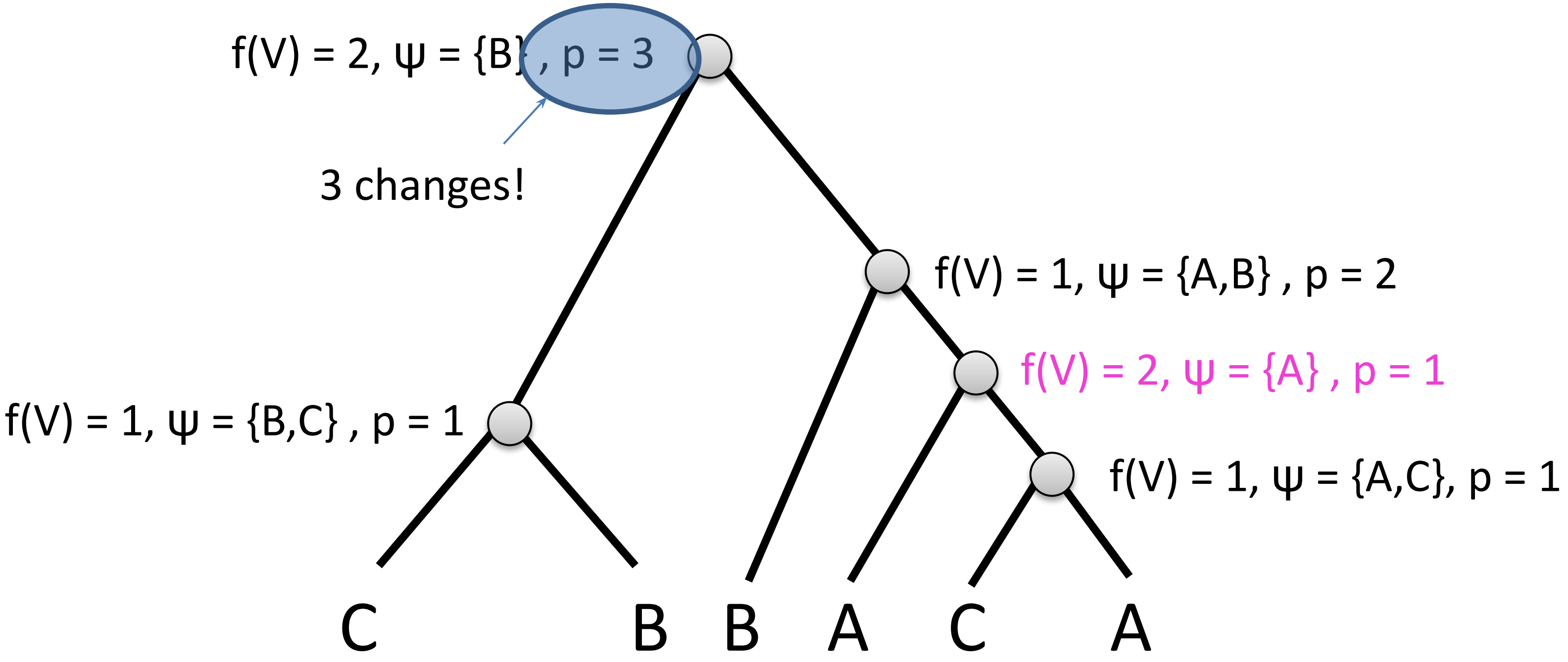
One character, three states

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 ψ = {X}


f(V) = 2, ψ = {B} , p = 3
3 changes!
f(V) = 1, ψ = {A,B} , p = 2
f(V) = 2, ψ = {A} , p = 1
f(V) = 1, ψ = {B,C} , p = 1
f(V) = 1, ψ = {A,C}, p = 1
C
B
B
A
C
A

Mapping to internal vertices V:

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)

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

(for all character columns)

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

How well do the characters fit the tree?

We can use the **consistency index**

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

Where m is the **minimum** number of steps (= number of character states – 1)

And s is the **actual** number of steps ($\geq m$), from the F-H algorithm

$$0.0 < CI \leq 1.0$$

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

- Search strategies coming up next class!

Tree Searching

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

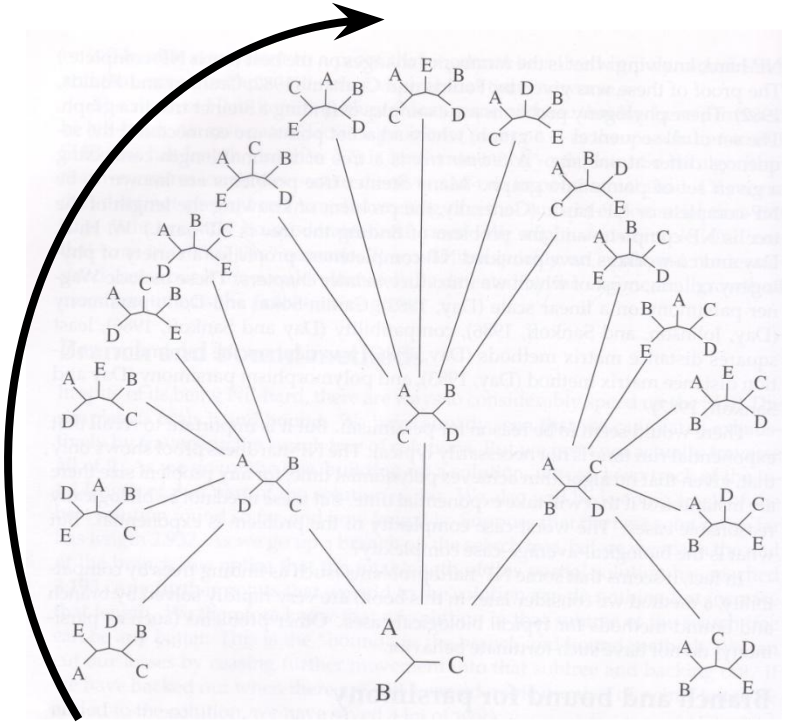
$$n_T = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

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

Branch-and-Bound

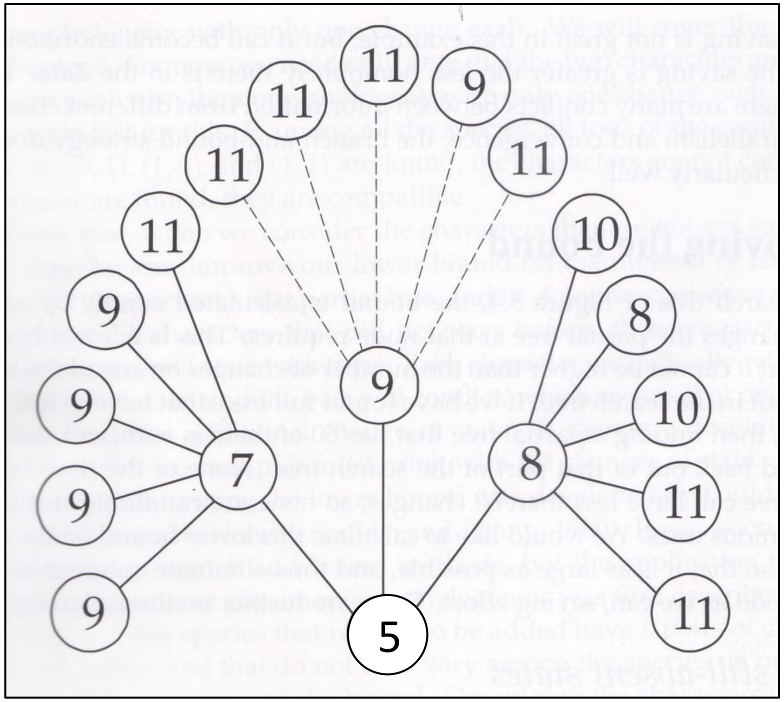
One way to restrict the search space is to explore it systematically, but identify and stop unproductive search paths

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

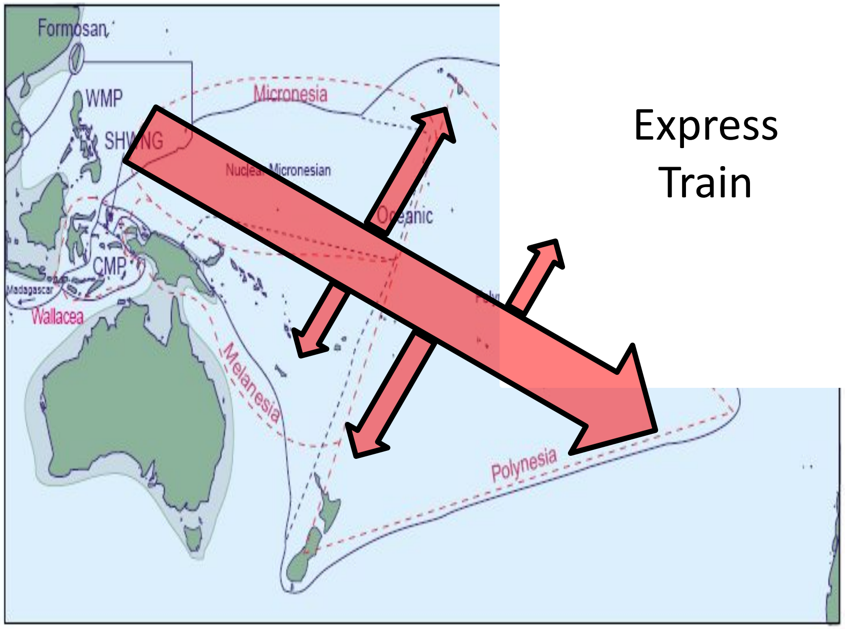
Felsenstein, 2004



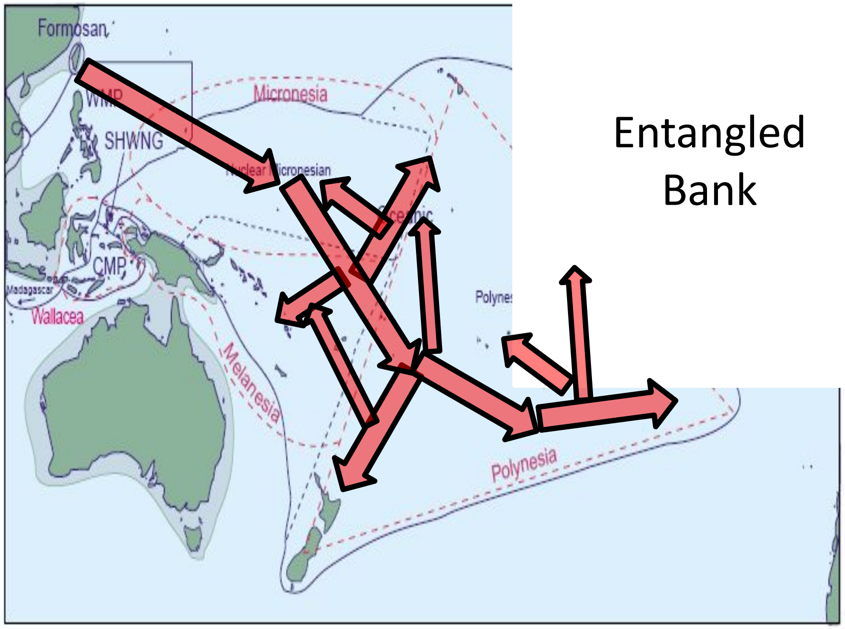
Number of substitutions required

Back to Polynesia

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

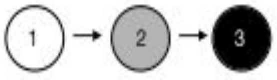


Express Train



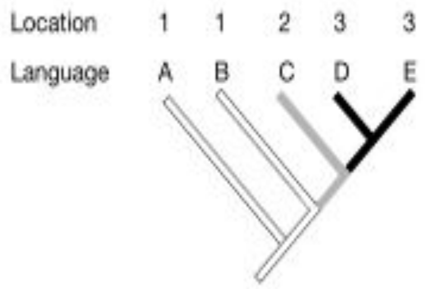
Entangled Bank

Predictions

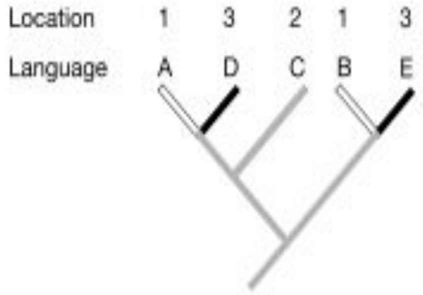


- Express train: strong tree-like signal, congruent with geography. **High CI**

  (assuming enough time for language to evolve)

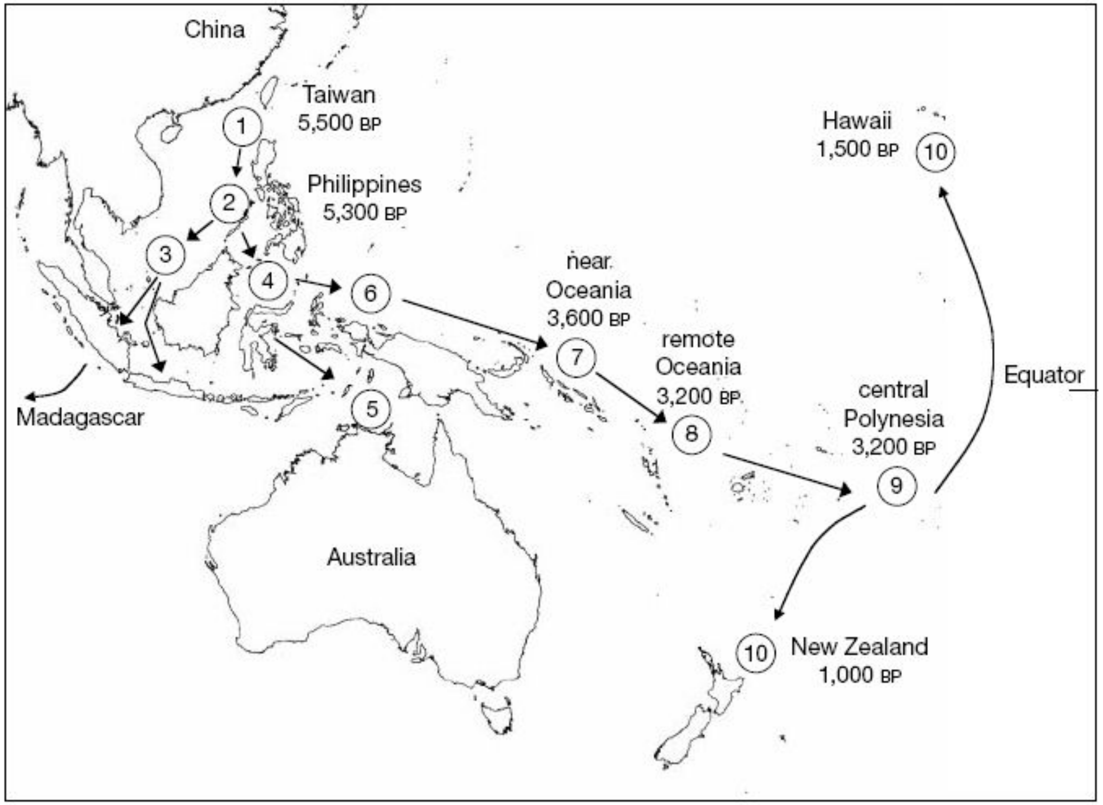


- Entangled bank: weaker signals, lots of sharing (travel / cultural exchange). **Low CI**

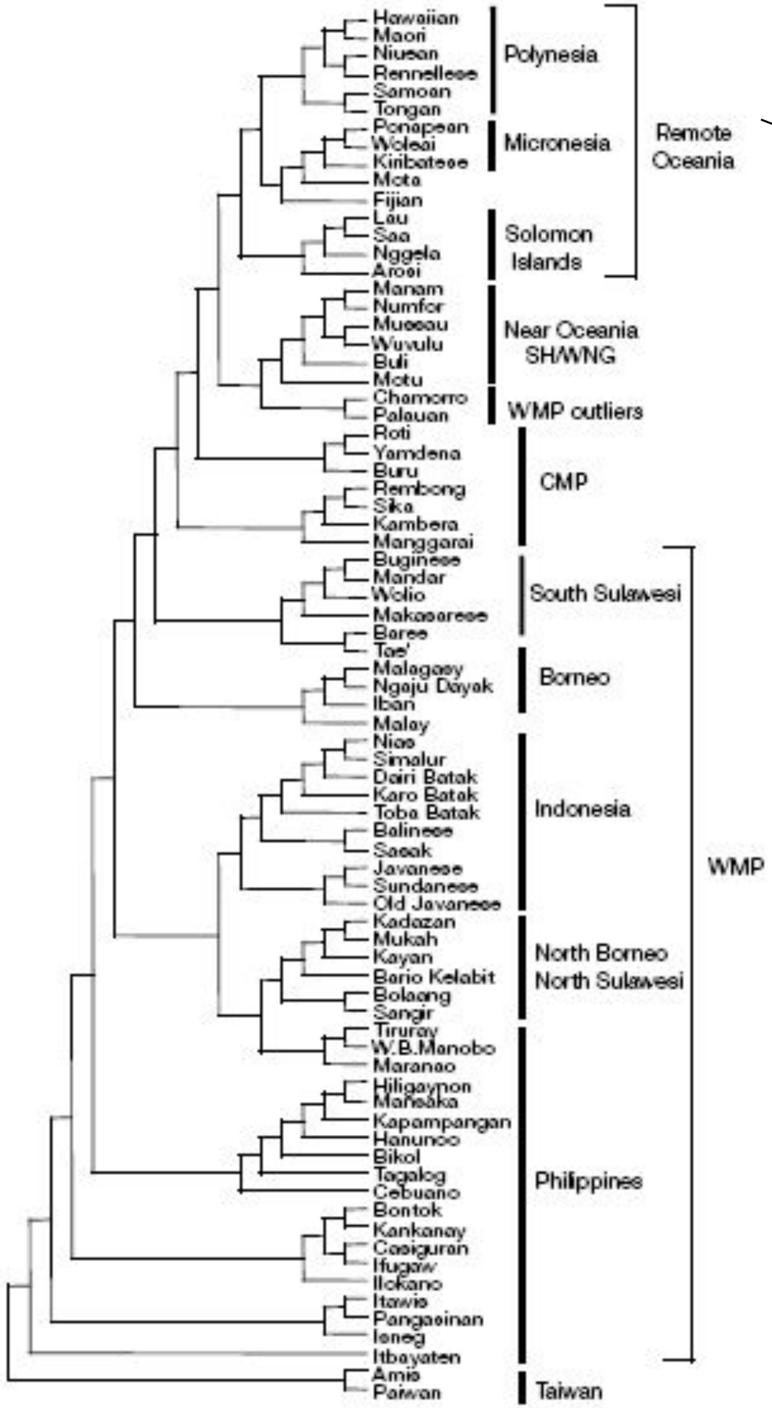


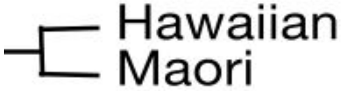
Analysis

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



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





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

Mininum number of transitions: 9 (= 10 – 1)

A total of **13** steps is needed to reconcile the 10 character states with the recovered tree (close to optimal)

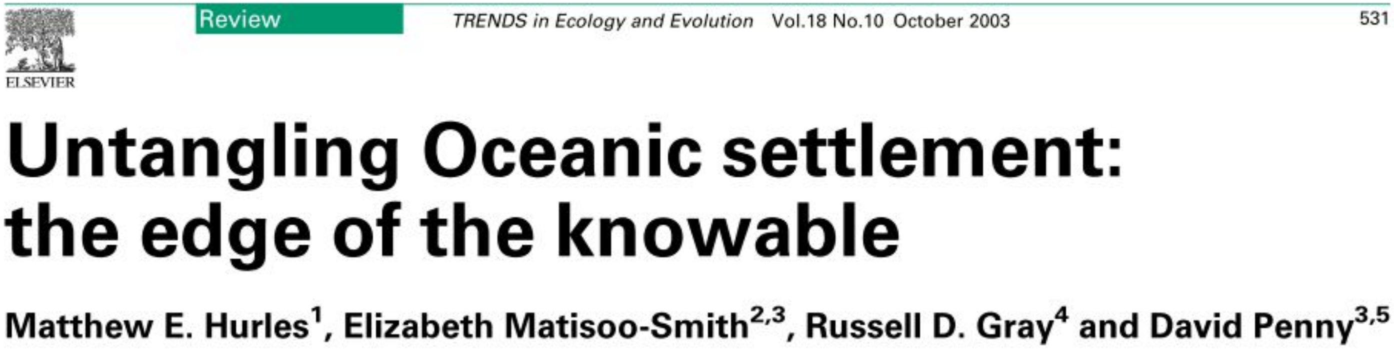
CI = 9/13 = 0.69

What does a CI of 0.69 mean?

We can compare it to the CI of **random** trees to see whether the fit is better than expected

Randomized trees: Average of **49** steps (CI = 9/49 = 0.053)

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



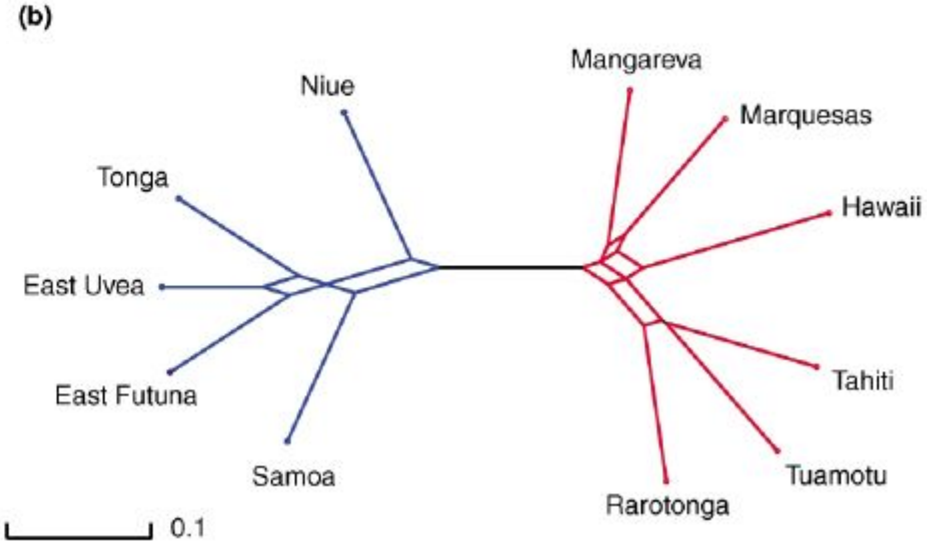
Review TRENDS in Ecology and Evolution Vol.18 No.10 October 2003 531

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

Problems with Parsimony

Not all alignment sites are informative

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

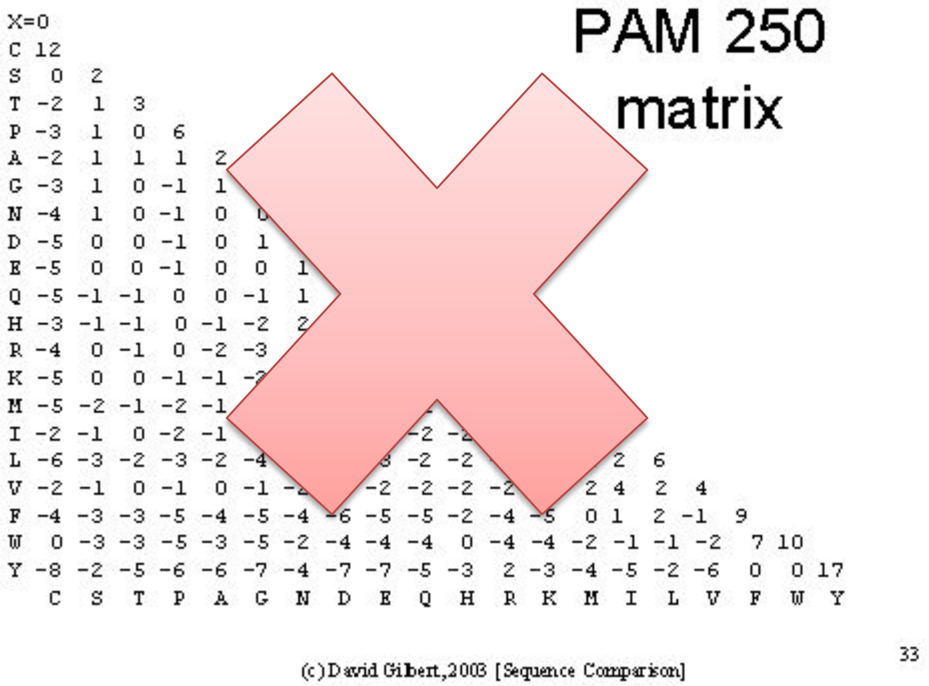
```
1 ACGTA
2 AGTGA
3 AGCCG
4 AGCAG
```

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

Other sites say nothing!

Parsimony treats all changes equally

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



PAM 250 matrix

X=0

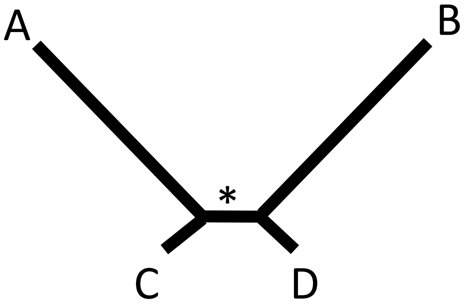
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	[illegible]														
N	-4	1	0	-1	0	0	[illegible]													
D	-5	0	0	-1	0	1	[illegible]	[illegible]												
E	-5	0	0	-1	0	0	1	[illegible]	[illegible]											
Q	-5	-1	-1	0	0	-1	1	[illegible]	[illegible]	[illegible]										
H	-3	-1	-1	0	-1	-2	2	[illegible]	[illegible]	[illegible]	[illegible]									
R	-4	0	-1	0	-2	-3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]								
K	-5	0	0	-1	-1	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]							
M	-5	-2	-1	-2	-1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]						
I	-2	-1	0	-2	-1	[illegible]	[illegible]	[illegible]	[illegible]	-2	-2	[illegible]	[illegible]	[illegible]	[illegible]					
L	-6	-3	-2	-3	-2	-4	[illegible]	[illegible]	-3	-2	-2	[illegible]	[illegible]	[illegible]	2	6				
V	-2	-1	0	-1	0	-1	-2	[illegible]	-2	-2	-2	-2	[illegible]	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
W	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Y	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	W	Y

(c)David Gilbert,2003 [Sequence Comparison]

33

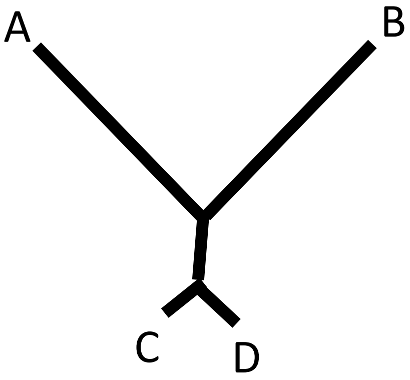
Long Branch Attraction

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



True tree

A	A**G**CAT**C**GAT
*	AATATACGA
B	C**G**TAC**C**CGA



Inferred tree

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*

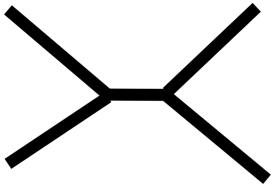
- Throwing away uninformative alignment columns is not necessarily ideal

Distance Methods

Overview

```
acca
gcca
gcct
tgca
```

	1	2	3	4
1				
2				
3				
4				



Step 1:
Construct distance matrix

Step 2:
Build tree

1: Sequences to Distances

Can use a model (e.g., PAM) to compute evolutionary distances

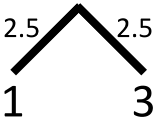
Distances to Trees

- Many different approaches:
  - Iterative/greedy (UPGMA, neighbour-joining)
  - Optimization (Fitch, minimum evolution)

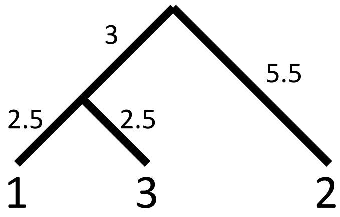
UPGMA again

Unweighted Pair Grouping with Arithmetic Mean

	1	2	3
1			
2	10		
3	5	12	



	1+3	2
1+3		
2	11	



Assumes a molecular clock
(distances from the root to all leaves will be EQUAL)

Neighbor-joining (Saitou and Nei 1987)

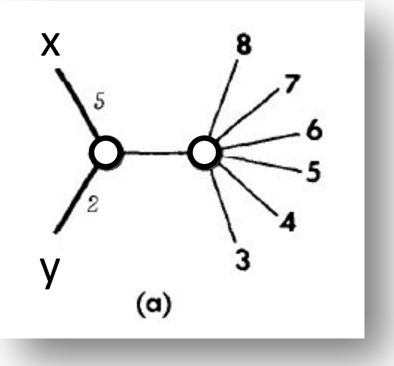
Start with a ‘star’ tree

At each iteration, split off the pair of taxa that minimizes the total sum of branch lengths in the tree

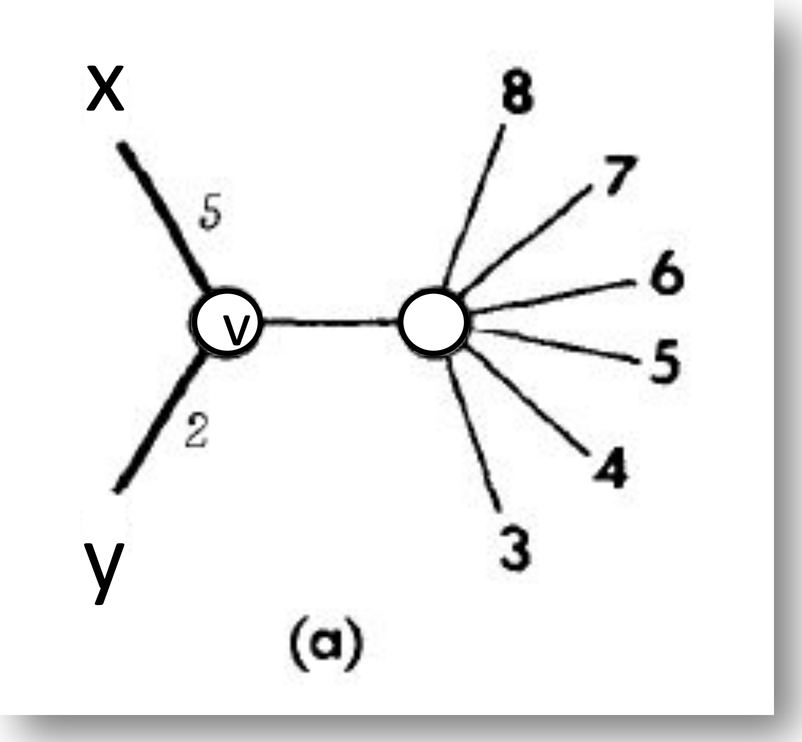
Choose groups x and y to minimize the **Q-criterion**:

$$\boxed{\delta(x,y)} - \underbrace{\frac{1}{(n-2)}\sum_{z}\delta(x,z) - \frac{1}{(n-2)}\sum_{z}\delta(y,z)}_{\text{Weighted distance to all leaves}}$$

Distance matrix entry for (x,y)



This splitting creates a new internal node, v, and assigns x and y as sisters in the growing tree



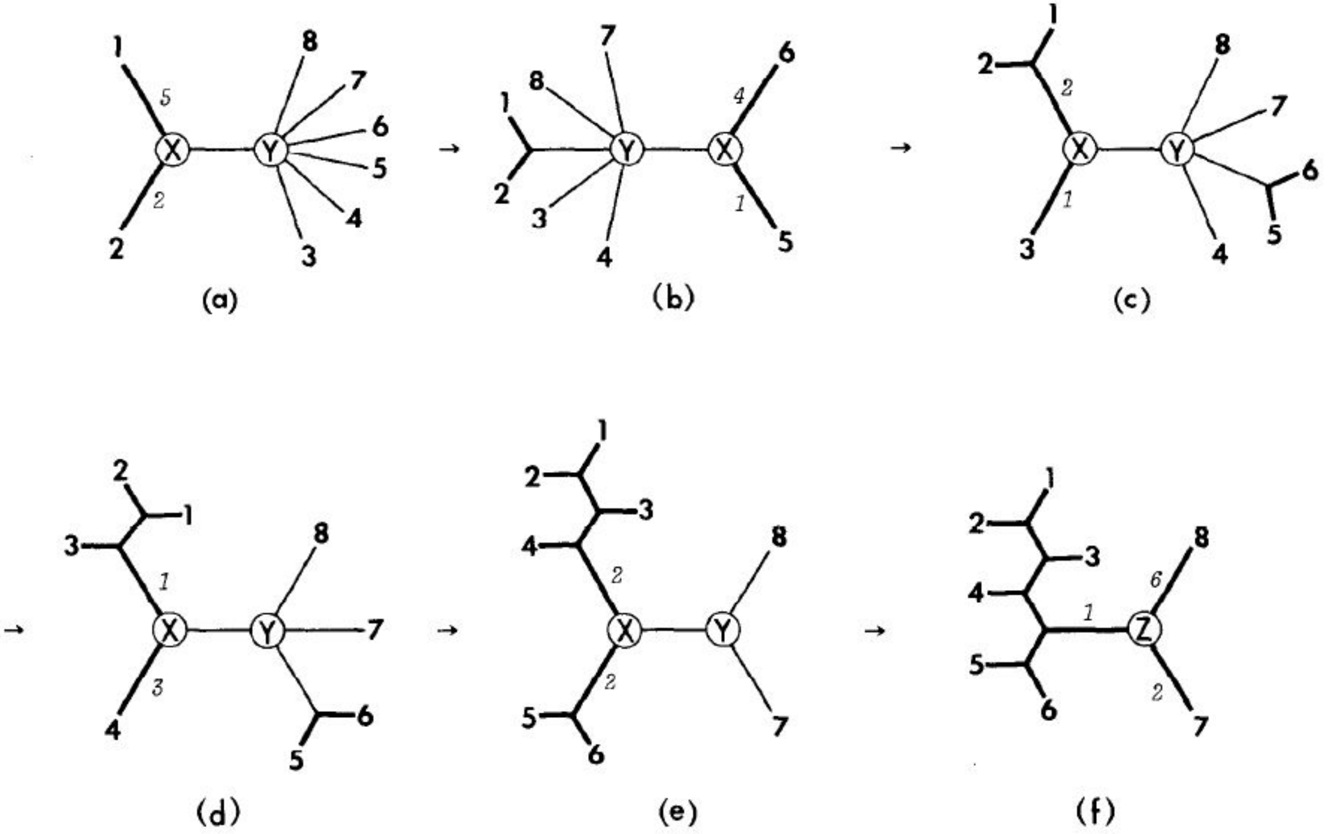
REDUCTION STEP: Recompute distances from all leaves to node v to allow subsequent computations of the Q criterion

$$\delta'(u, v_{xy}) = \tfrac{1}{2}(\delta(u,x) + \delta(u,y) - \delta(x,y))$$

And assign branch lengths x-v and y-v

$$b_x = \frac{1}{n-2} \sum_{z \neq x,y} (\delta(x,z) + \delta(x,y) - \delta(y,z))$$

Continue until binary tree is obtained



Figures from Saitou and Nei (1987)
Formulas from Bryant, *J Classific* (2005)

Neighbor-joining vs. UPGMA

- Neighbor-joining uses a somewhat less intuitive optimality criterion **Q**

- However, it is still iterative and still fast

- Another advantage is that it does not assume a molecular clock – branch lengths are assigned based on **all** distances in the matrix

Advantages of Distance Methods

- Explicit modelling of residue changes
- Can be very FAST – neighbour-joining can build trees with thousands of leaves

Disadvantages of Distance Methods

- A considerable amount of information is lost when sequence pairs are replaced with a single distance

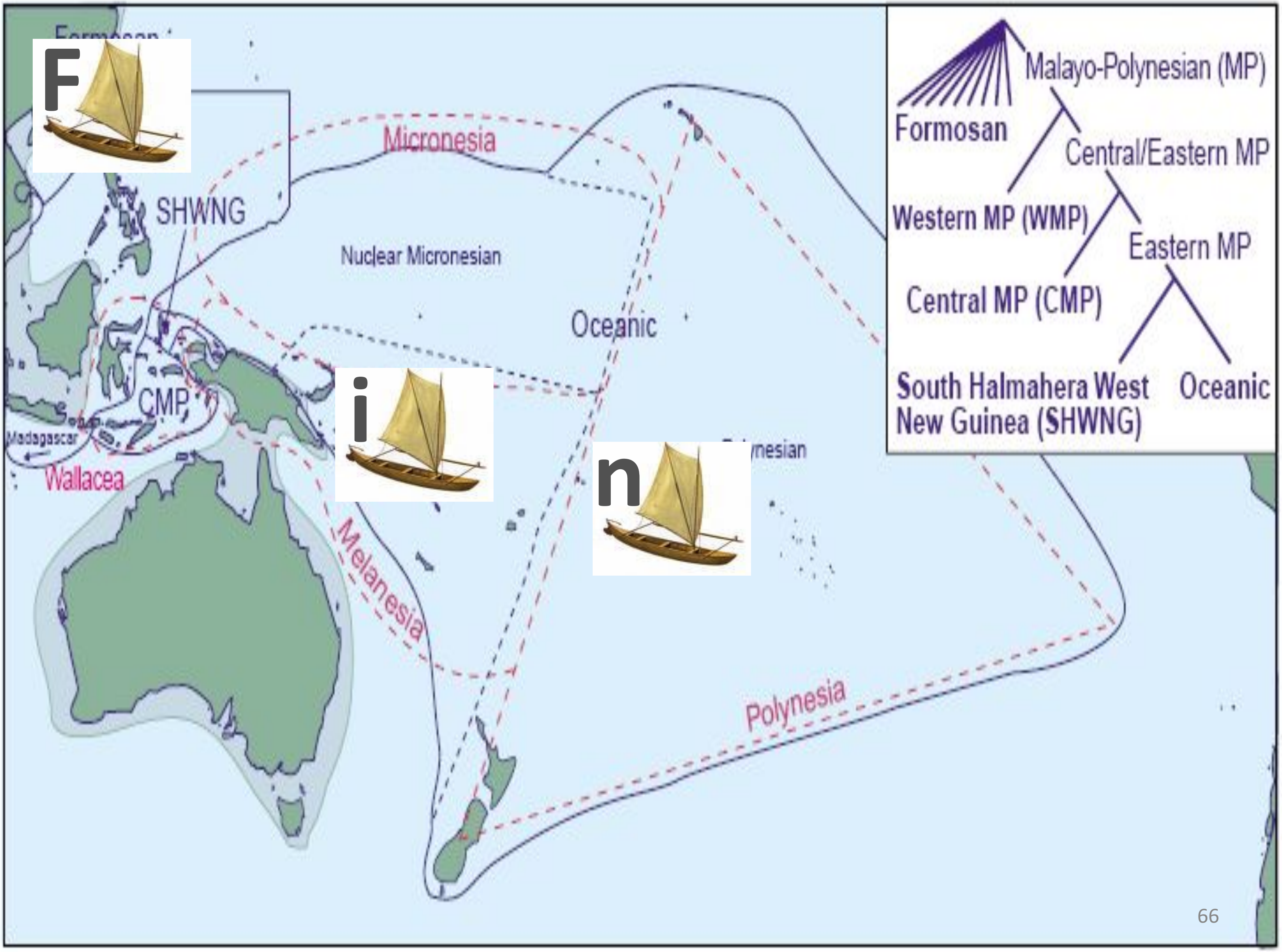
- Greedy methods may perform poorly for some problems

Conclusion

- **Parsimony**: Character-based, model-free
  - tree search required

- **Distance**: Pairwise distances, can use a model
  - Greedy approaches or iterative searches

- Is there a way to use models without collapsing each pair of sequences to a single distance value? yes


Formosan
F
SHWNG
Micronesia
Nuclear Micronesian
Oceanic
CMP
Madagascar
Wallacea
i
n
Polynesian
Melanesia
Polynesia
Malayo-Polynesian (MP)
Formosan
Central/Eastern MP
Western MP (WMP)
Eastern MP
Central MP (CMP)
South Halmahera West New Guinea (SHWNG)
Oceanic