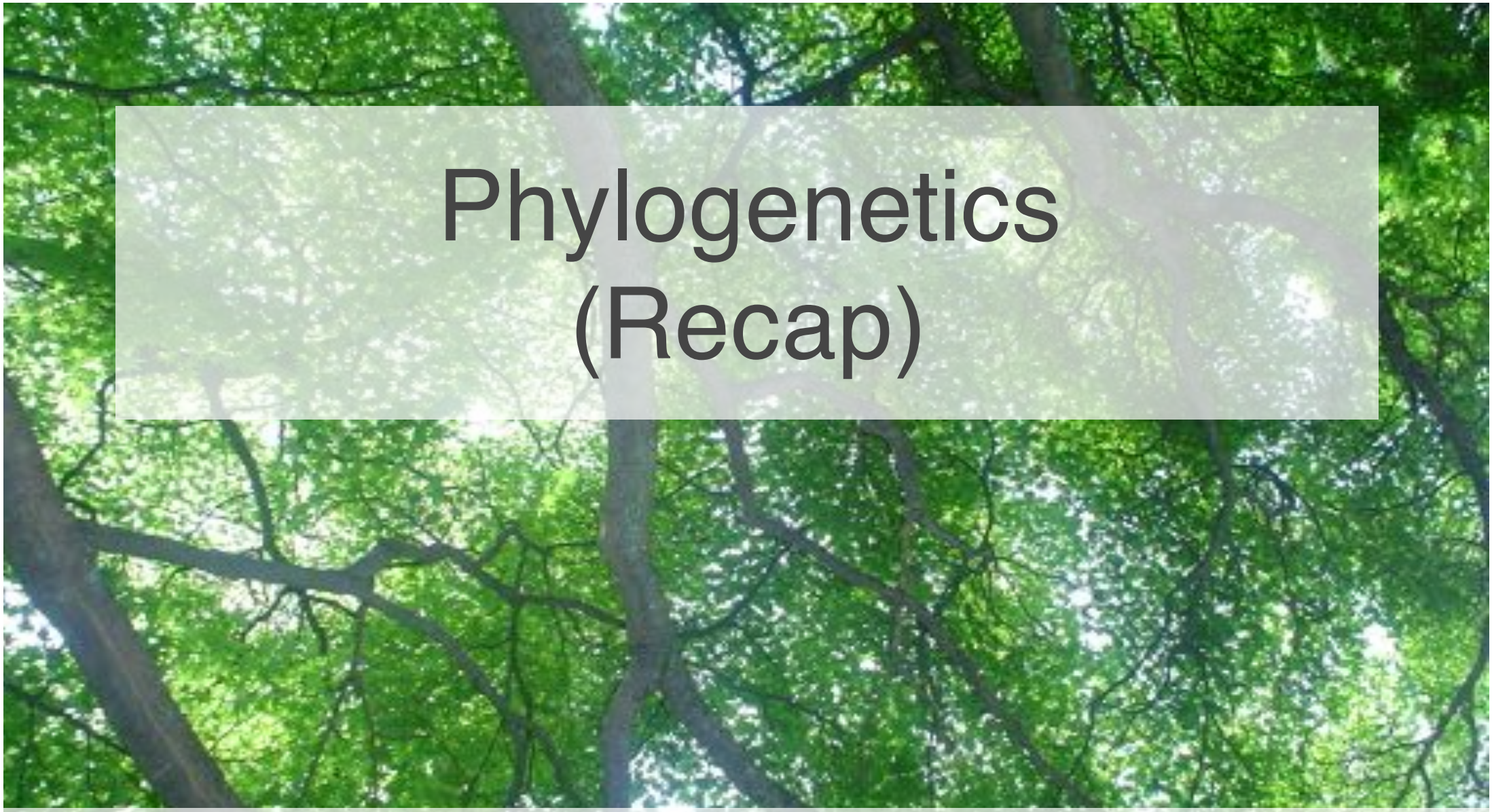
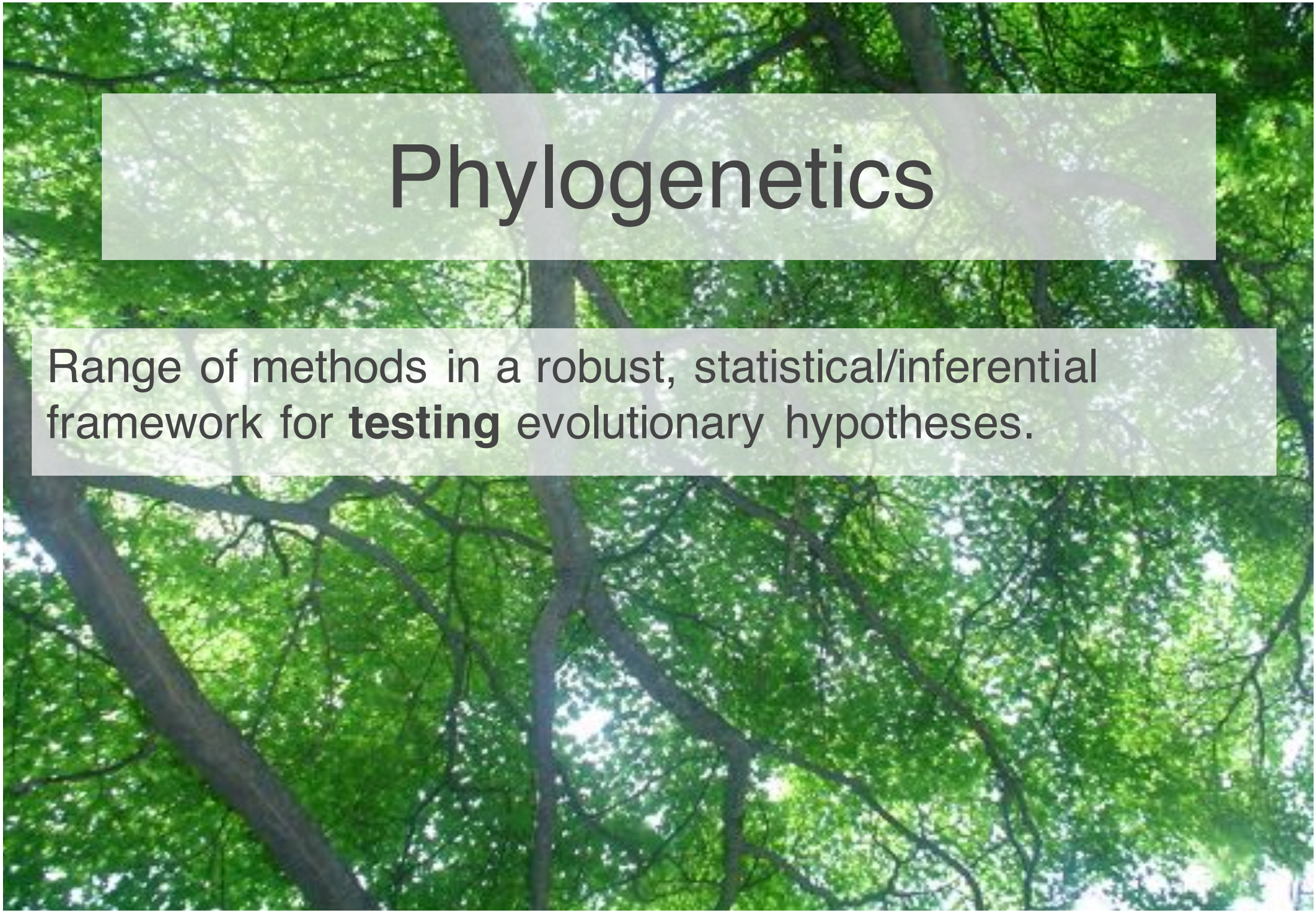


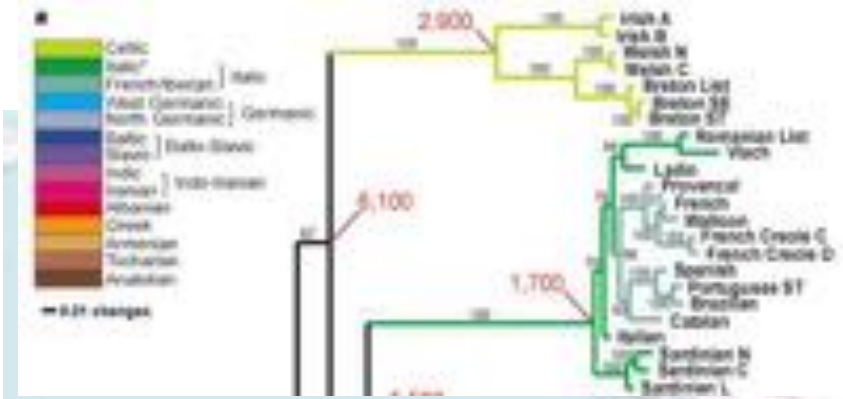
Phylogenetics (Recap)

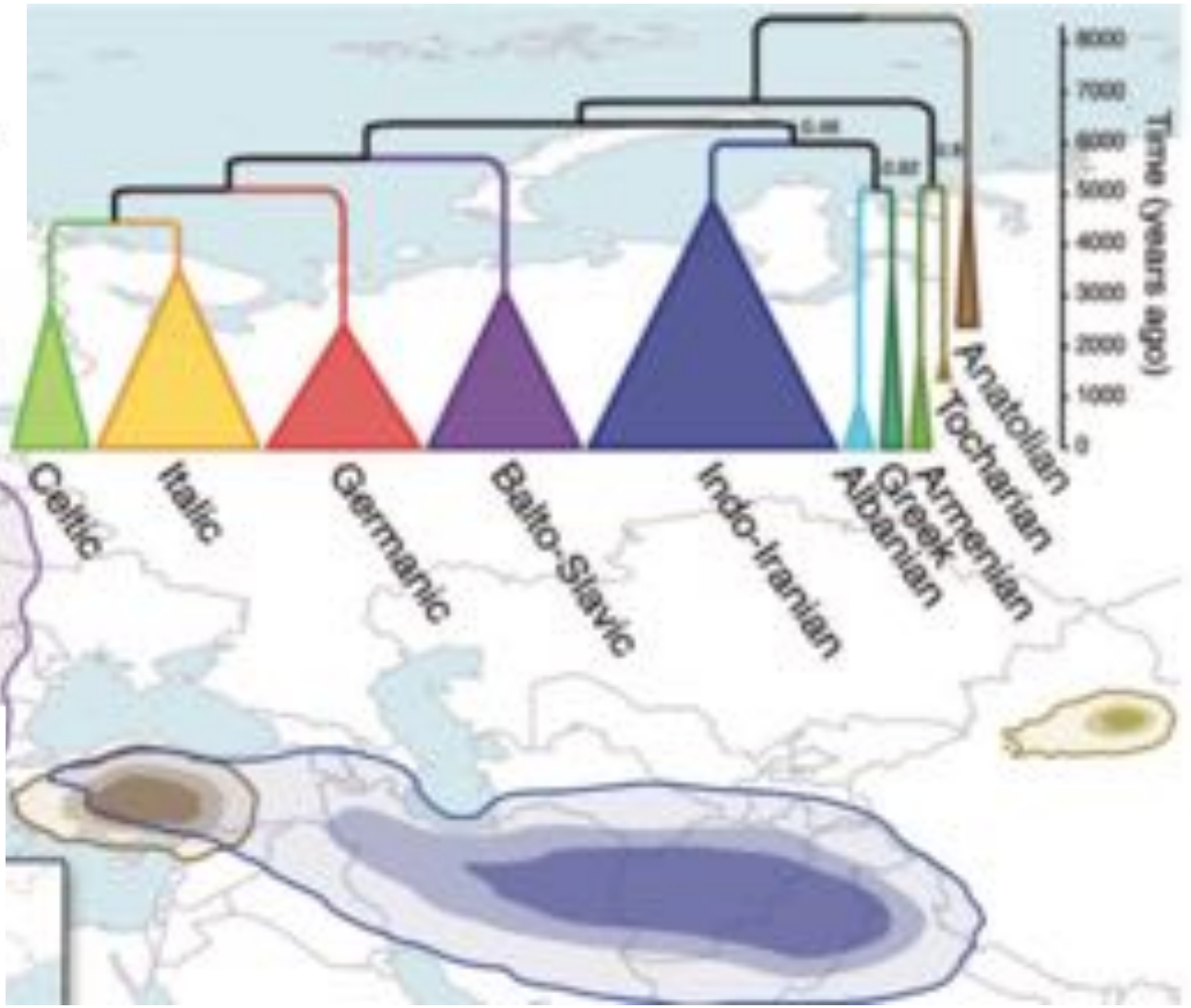
Simon J. Greenhill

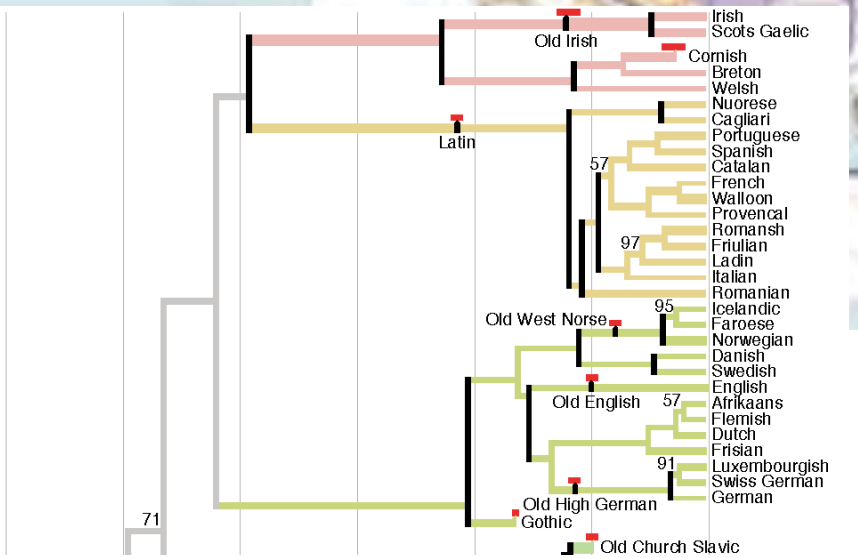
ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Max Planck Institute for the Science of Human History

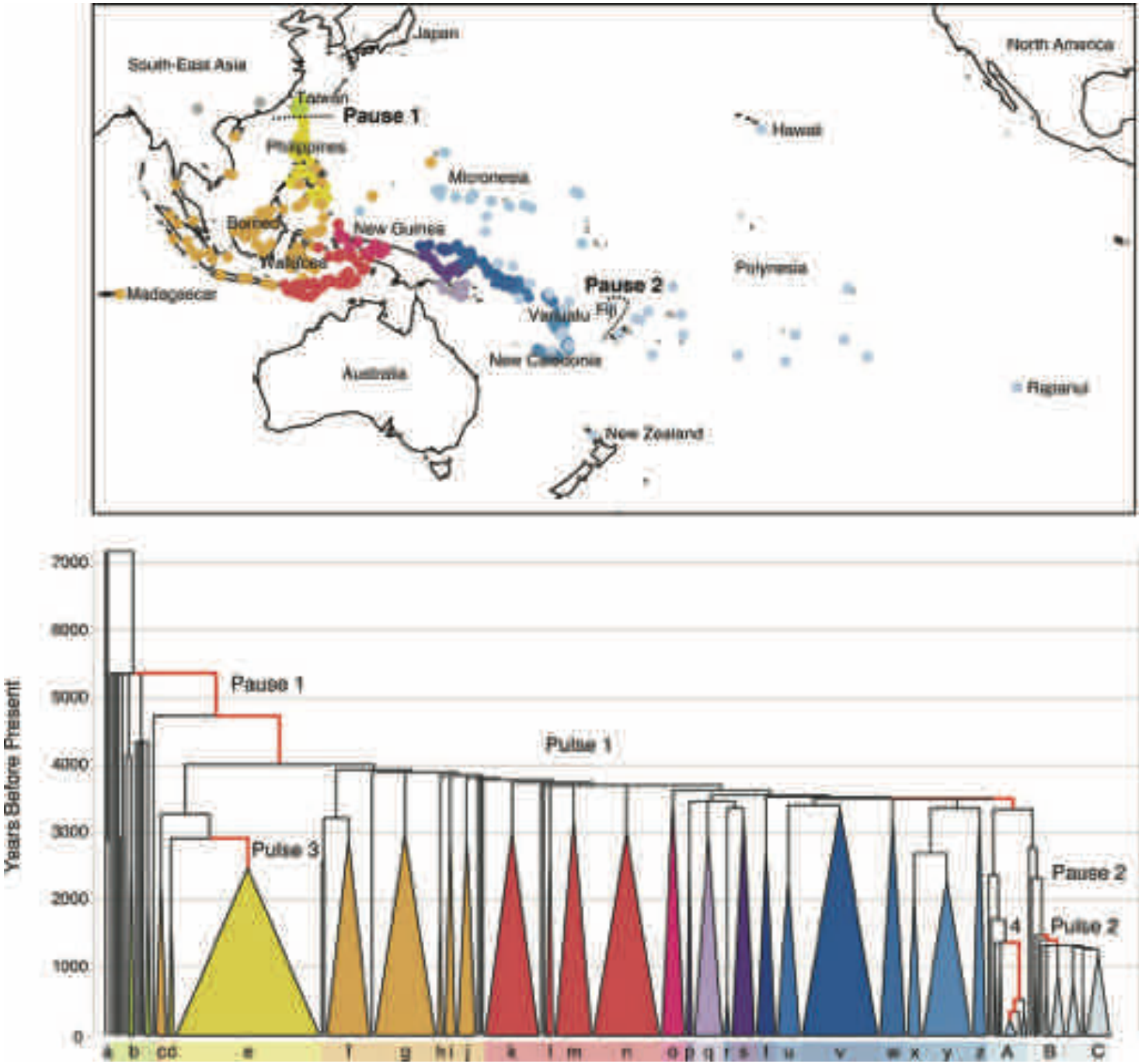

Phylogenetics
Range of methods in a robust, statistical/inferential framework for **testing** evolutionary hypotheses.



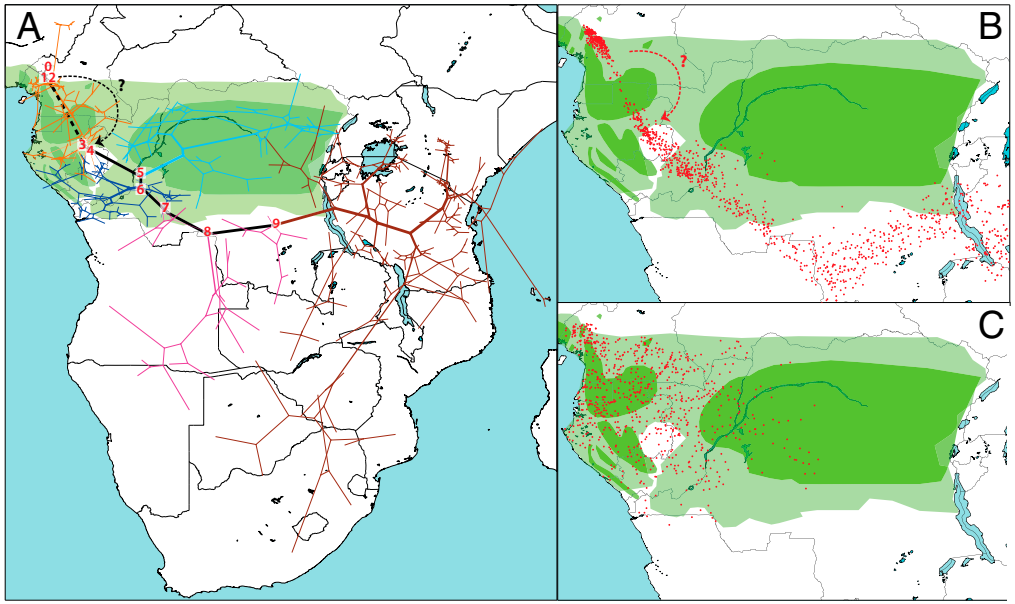


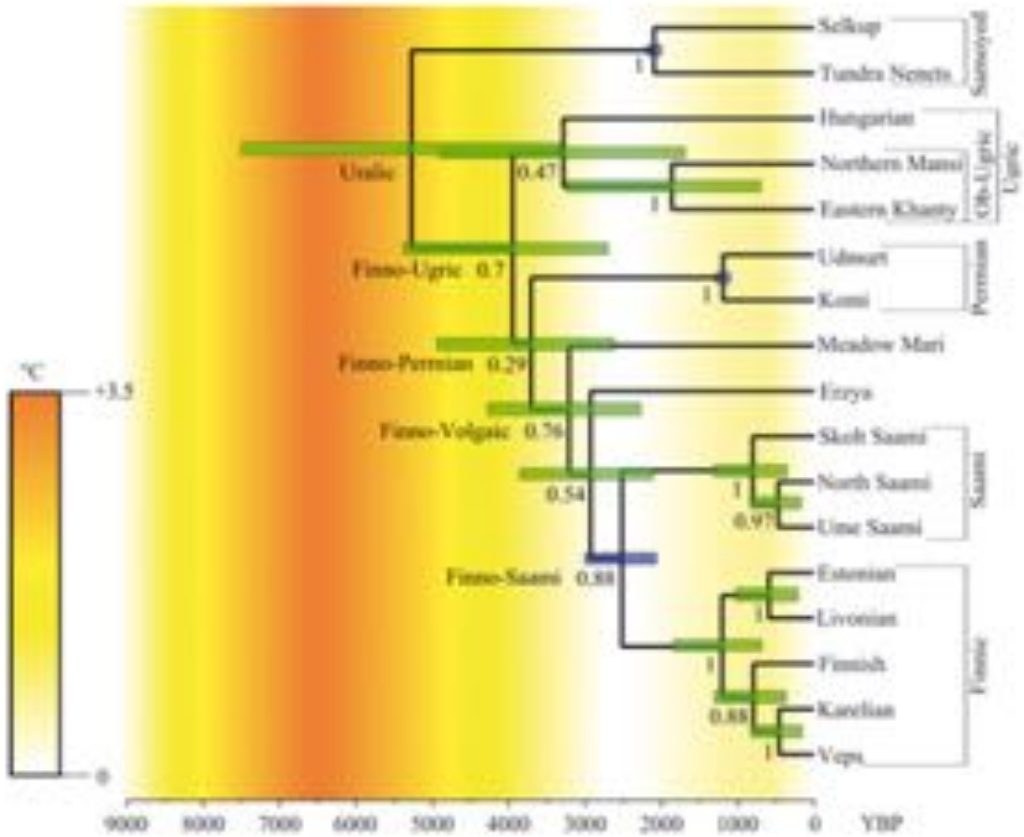


Gray & Atkinson ’03, Bouckaert et al’ 11, Chang et al ‘14.


Japan
South-East Asia
Taiwan
Pause 1
Philippines
Micronesia
Borneo
New Guinea
Wallacea
Madagascar
Pause 2
Vanuatu
Fiji
New Caledonia
Australia
North America
Hawaii
Polynesia
Rapanui
New Zealand
Years Before Present
7000
6000
5000
4000
3000
2000
1000
0
Pause 1
Pulse 1
Pulse 3
Pause 2
Pulse 2
4
a b cc e f g hi j a l m n opqrstu v wx y z A B C

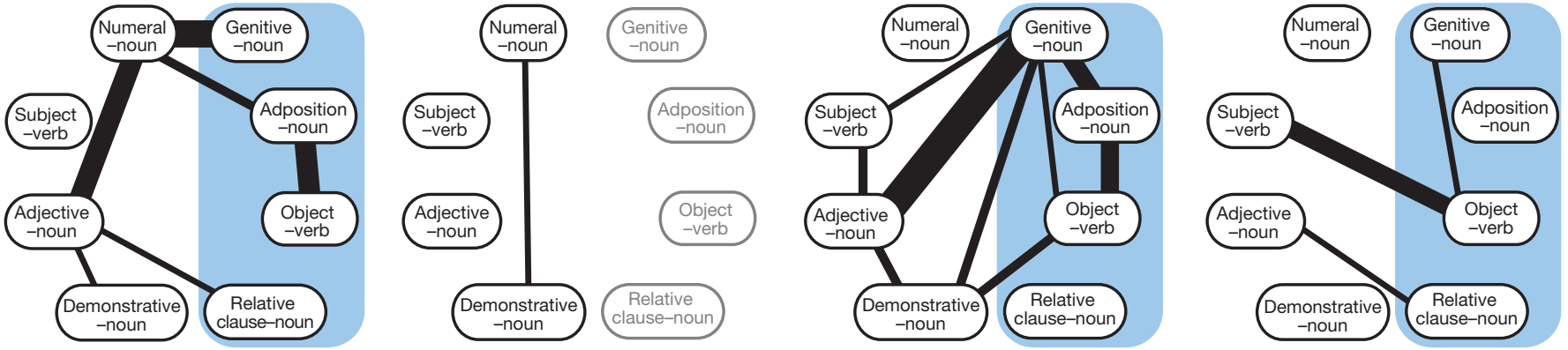
Gray et al ‘09



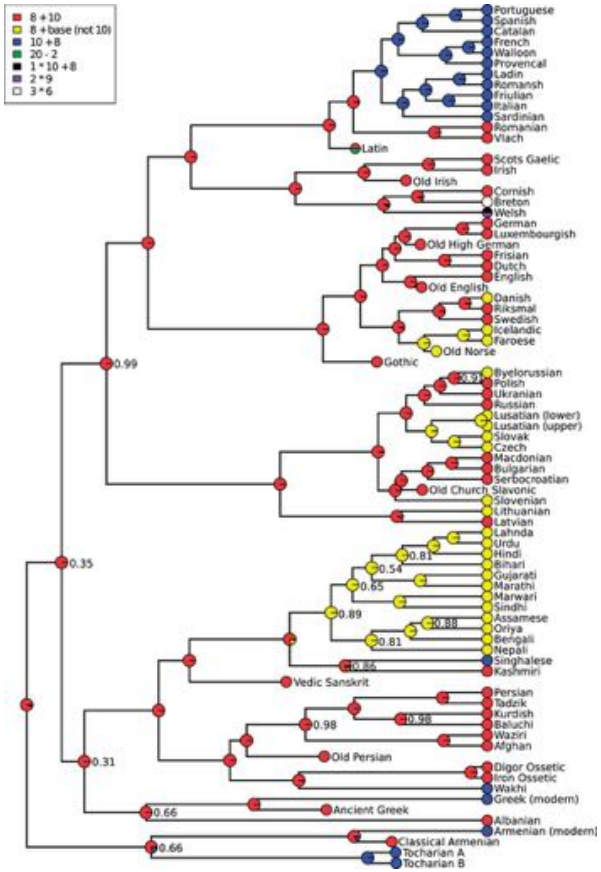


Grollemund et al. ‘15

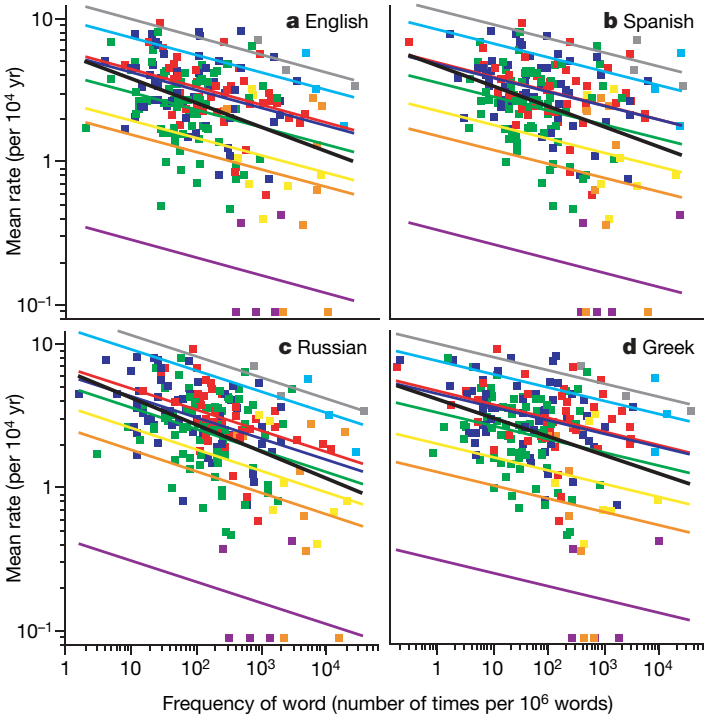
Honkola et al ‘13



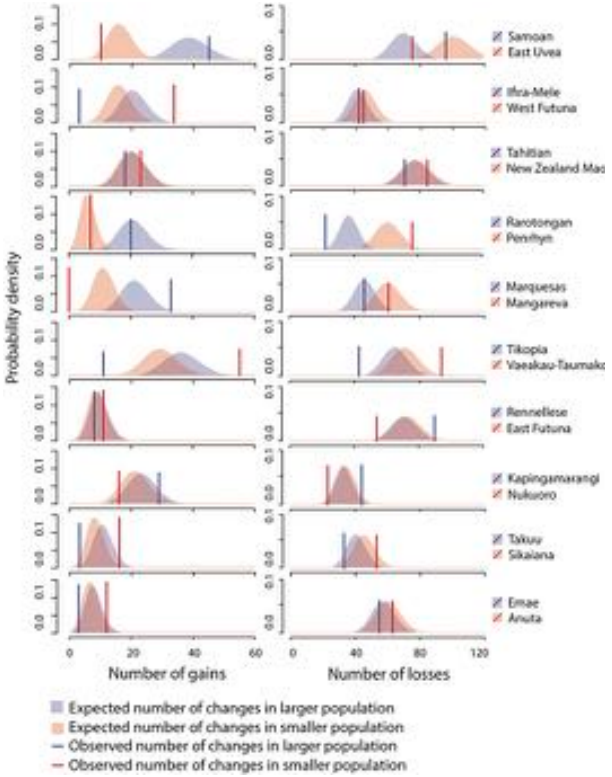
Dunn et al. ‘15



Calude and Verkerk ‘16

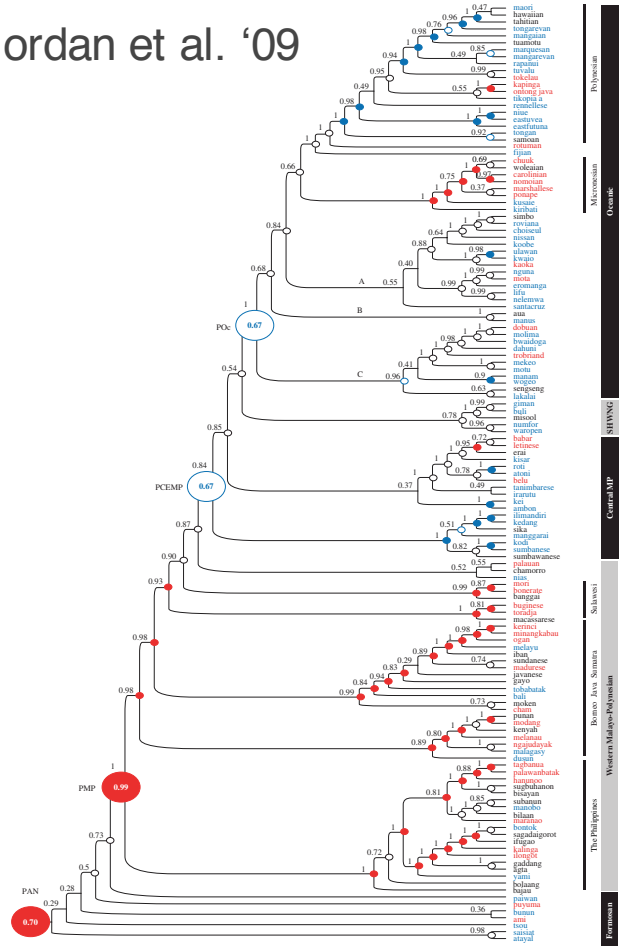


Pagel et al ’07

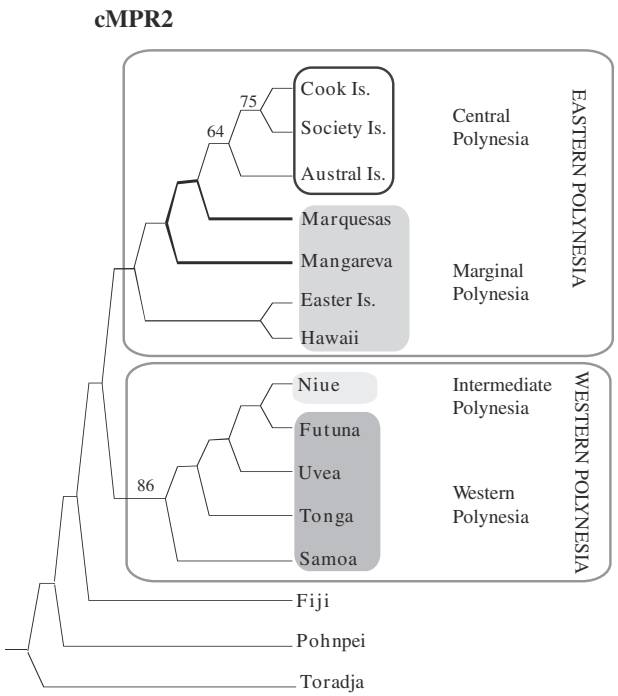


Bromham et al. ’15

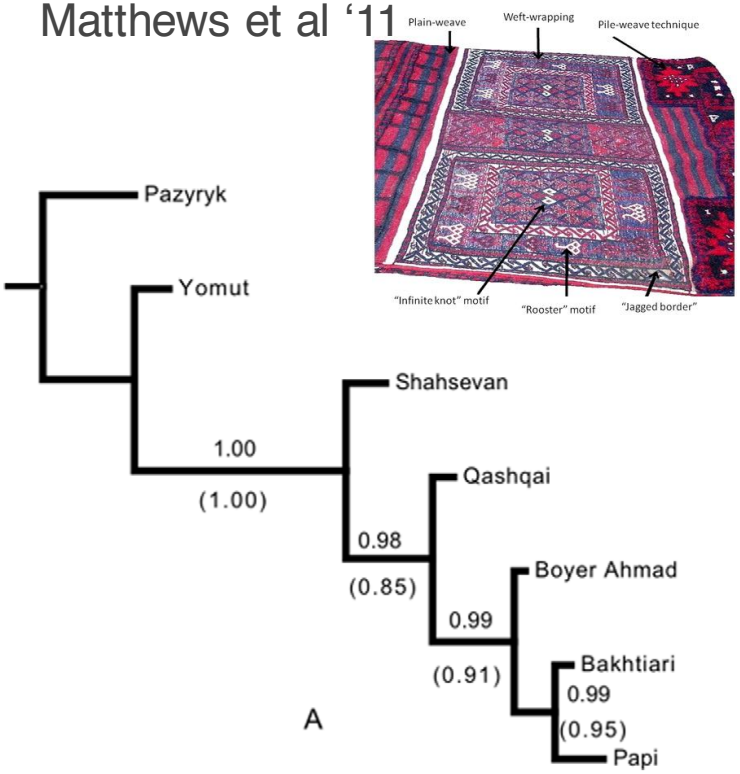
Jordan et al. ‘09

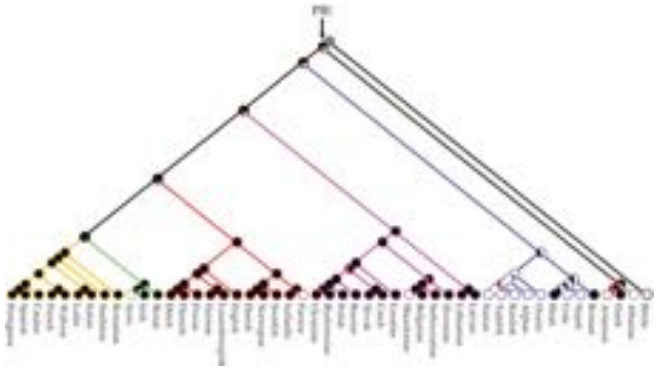


Larsen ‘11

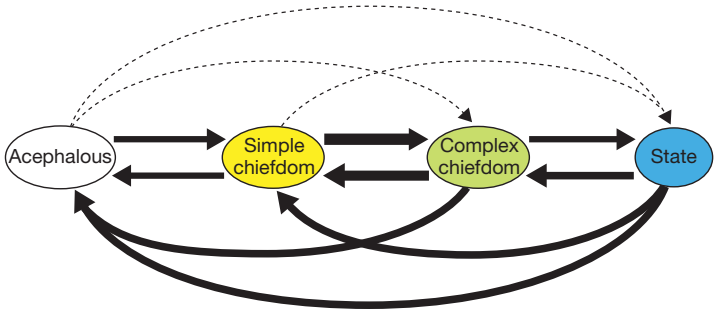


Matthews et al ‘11

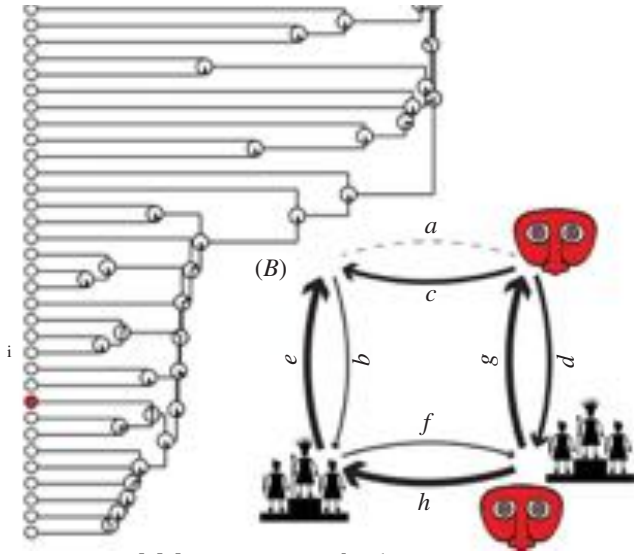




Da Silva & Tehrani ‘16



Currie et al. ‘10



Watts et al. ‘15

Controversial

"most vibrant stream of contemporary linguistics"

"Computational methodologies of this kind can only be helpful for historical linguistics"

"languages and biomolecular sequences evolve in very different ways"

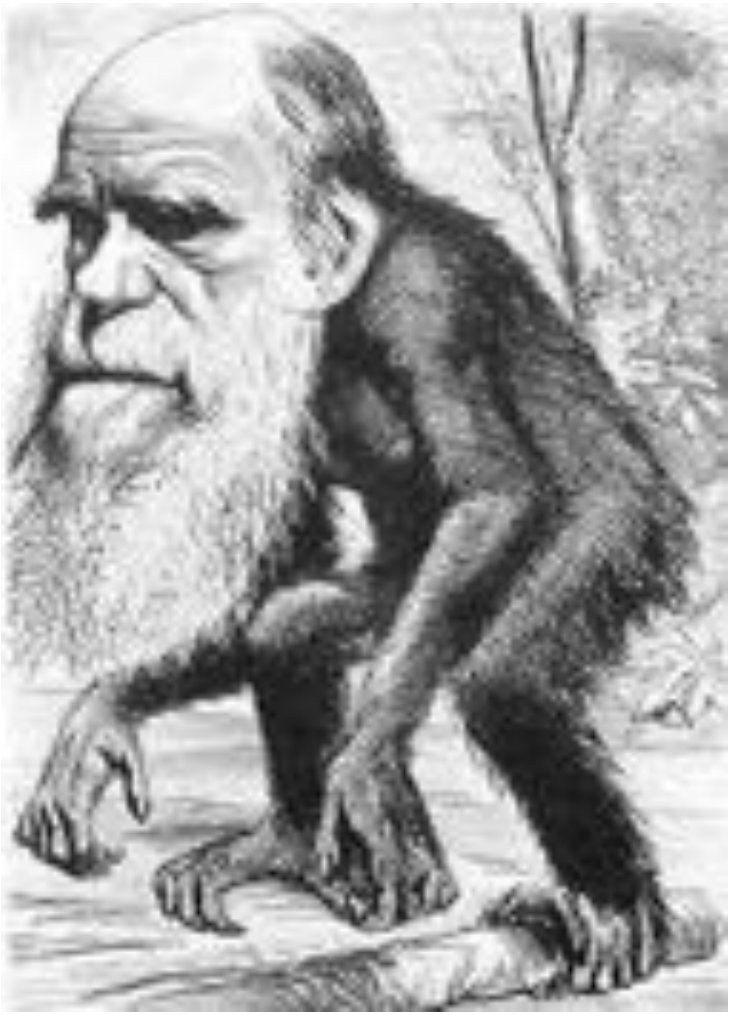
"more questions than answers"

"utter bollocks"

"biggest intellectual fraud since Chomsky"

"this isn't history, it's history put in nested boxes!"

What is **evolution**?



Variation

Heritability

Differential survival

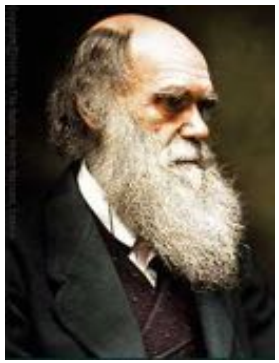
When and where did 🐨👱💬 **originate**?

What **differences** are there between 🐨👱💬's?

How are 🐨👱💬 **related** to other 🐨👱💬's?

What **processes** shaped 🐨👱💬?

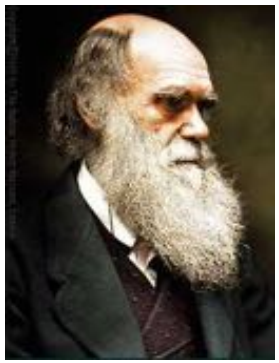
Can we infer what 🐨👱💬 were in the **past**?



Darwin (1871)

"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”



Darwin (1871)

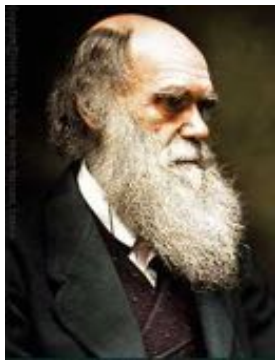
"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”

Schleicher 1863

Darwinism Tested by the Science of Language

“same process has long been generally assumed for linguistic organisms”



Darwin (1871)

"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”

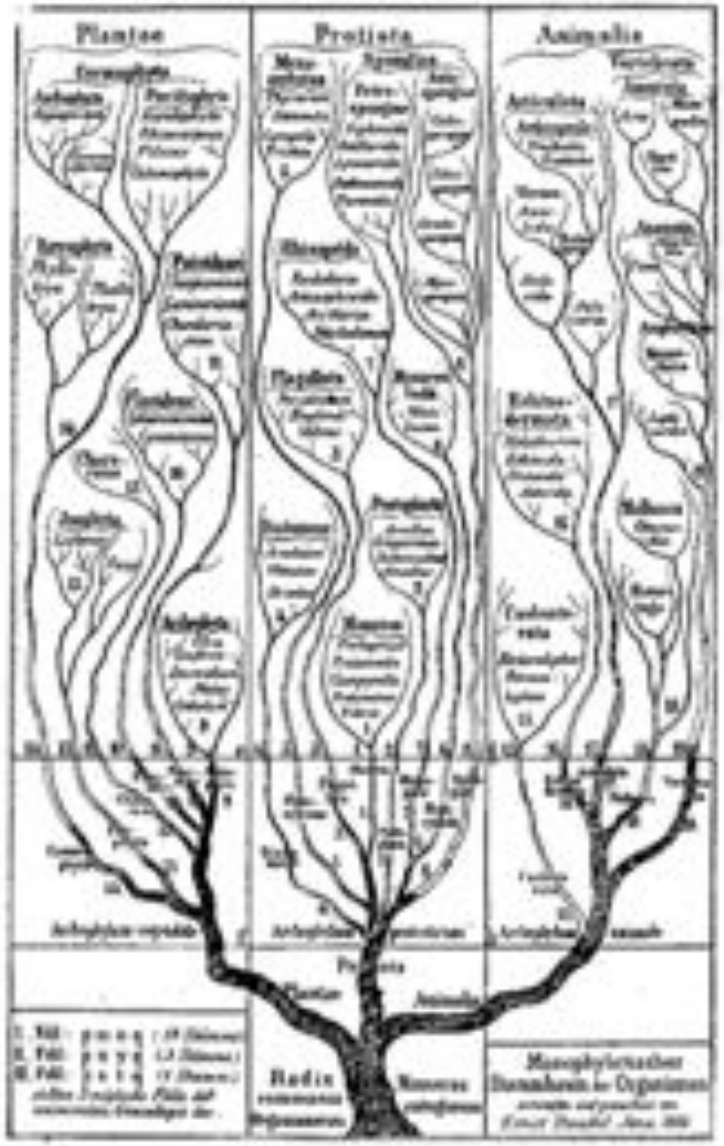
Schleicher 1863

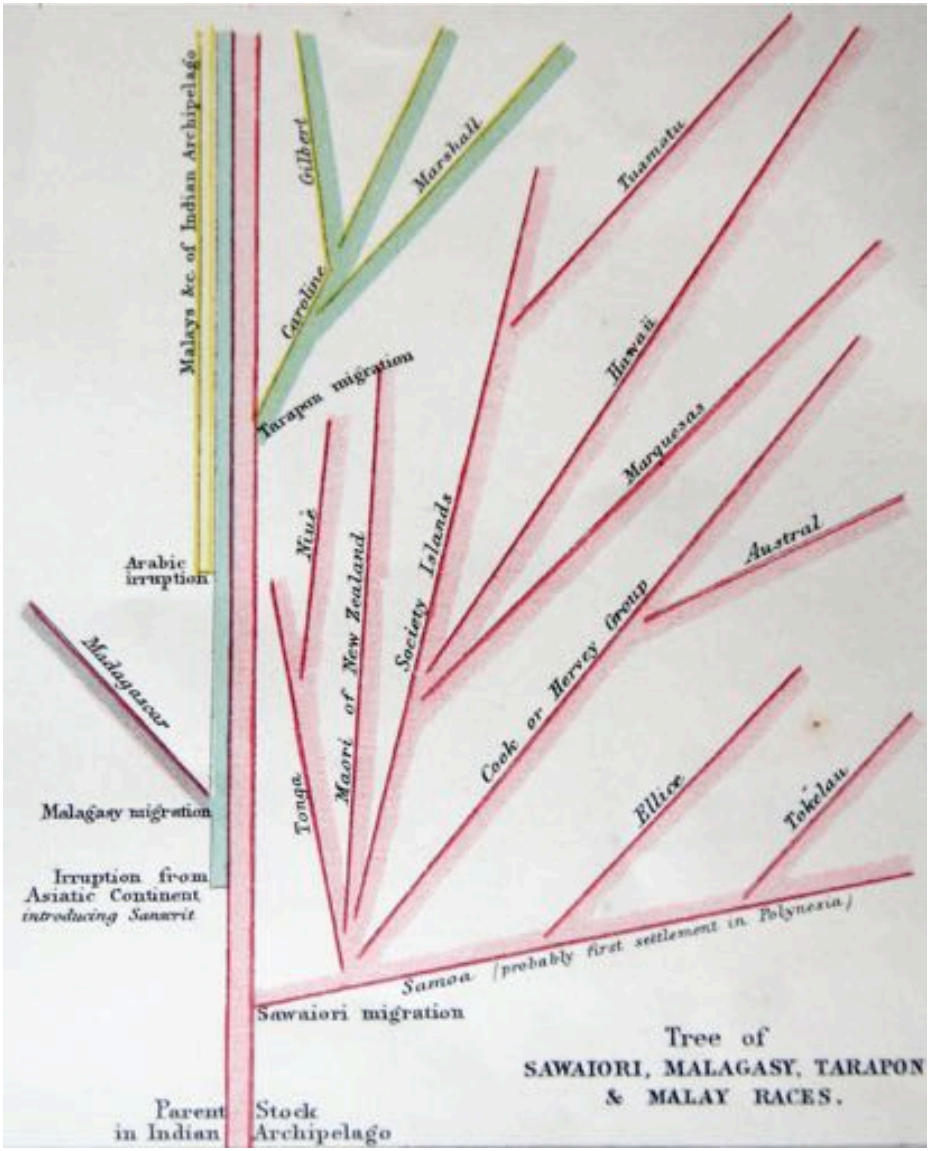
Darwinism Tested by the Science of Language

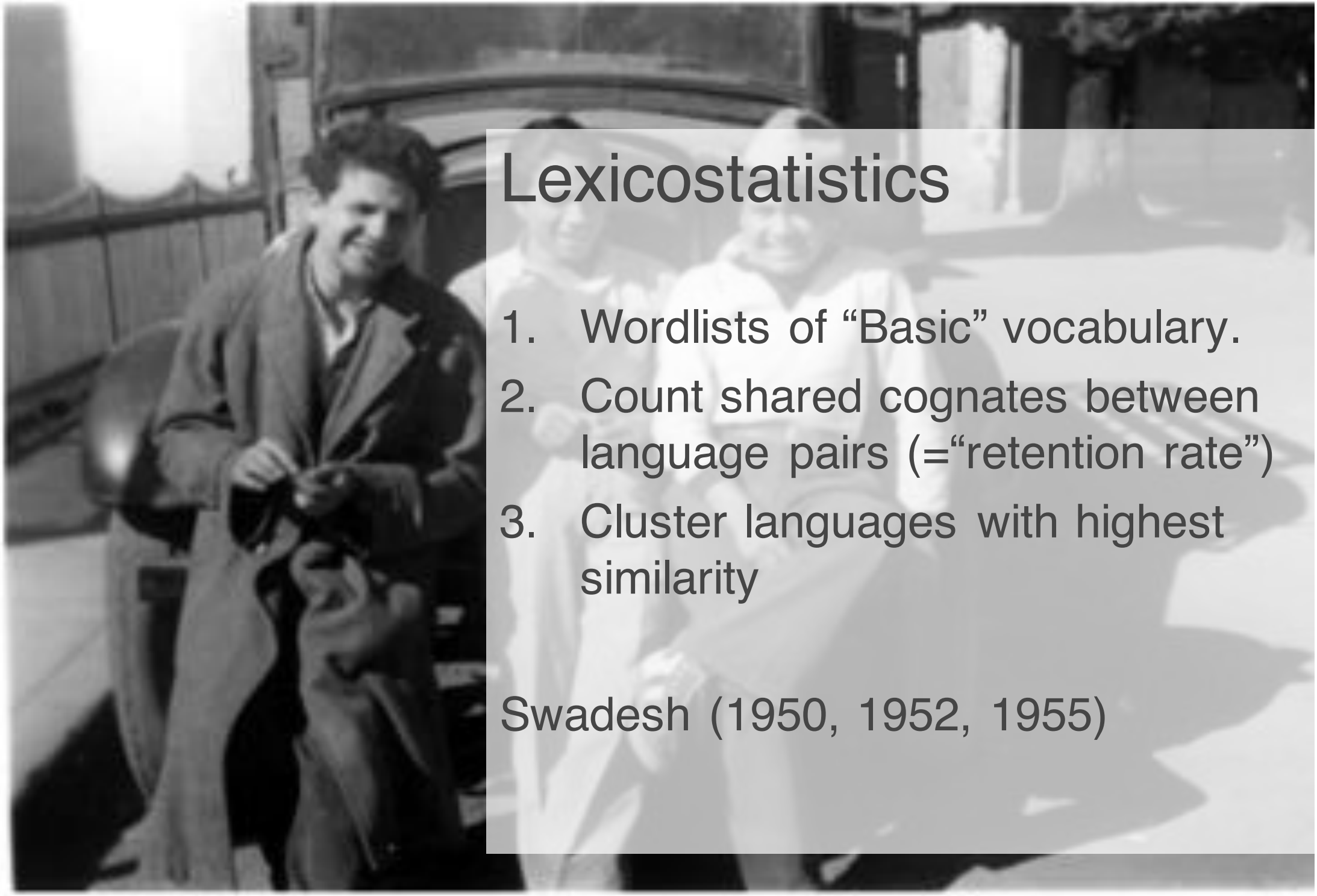
“same process has long been generally assumed for linguistic organisms”

Brugmann (1884)

Importance of using “shared innovations” to identify clades and not “shared retentions”


Plantae
Protista
Animalia
Radix


Malays &c. of Indian Archipelago
Gilbert
Marshall
Caroline
Tuamotu
Hawaii
Marquesas
Austral
Tarapon migration
Arabic irruption
Niué
Maori of New Zealand
Society Islands
Cook or Hervey Group
Madagascar
Tonga
Ellice
Tokelau
Malagasy migration
Irruption from Asiatic Continent introducing Sanscrit
Samoa (probably first settlement in Polynesia)
Sawaiori migration
Tree of
SAWAIORI, MALAGASY, TARAPON & MALAY RACES.
Parent Stock in Indian Archipelago



Lexicostatistics

1. Wordlists of “Basic” vocabulary.
2. Count shared cognates between language pairs (=“retention rate”)
3. Cluster languages with highest similarity

Swadesh (1950, 1952, 1955)

	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

Identified by Systematic Sound Correspondences
- e.g. Maori “t” = “k” in Hawaiian.

Elbert 1953

TABLE 2

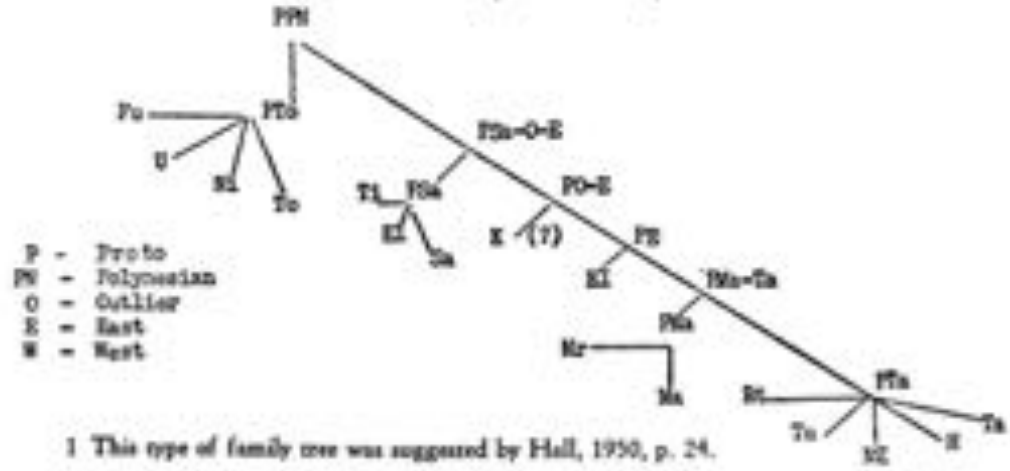
Polynesian cognate percentages

U	Ni	To	Fi	S	Sa	Si[1]	Fu[1]	Ou[1]	Nu[1]	E	Ei	Mr	Ma	Pa	Pu	Ni	H	Ta	
83	68	74	83	79	74	57	62	62	54	52	58	57	56	66	62	61	58	58	Pu
	72	86	78	74	70	53	59	58	52	51	53	53	51	62	62	61	55	59	U
		64	68	61	63	50	59	49	54	49	49	55	47	56	55	54	49	51	Ni
			70	64	60	46	55	55	49	45	48	49	45	58	53	54	49	52	To
				81	76	66	66	62	59	59	62	67	63	73	68	71	67	66	Fi
					78	66	61	62	59	58	61	62	63	66	66	67	68	66	E
						64	60	64	55	53	53	55	52	67	62	57	59	60	Sa
							55	60	52	55	52	55	56	60	58	60	60	59	Si
								57	52	54	58	59	53	55	61	60	55	54	Fu
									51	56	53	53	51	53	56	55	56	53	Ou
										53	49	65	66	53	50	52	49	48	Nu
											47	69	65	54	51	51	49	50	E
												64	63	64	62	63	64	62	Ei
													73	75	72	70	69	68	Mr
														73	69	67	70	67	Ma
															83	83	79	85	Pa
																79	77	83	Tu
																	71	73	Ni
																		76	H

1 Percentages based on incomplete data.

TABLE 4

A tentative family tree for Polynesia[1]



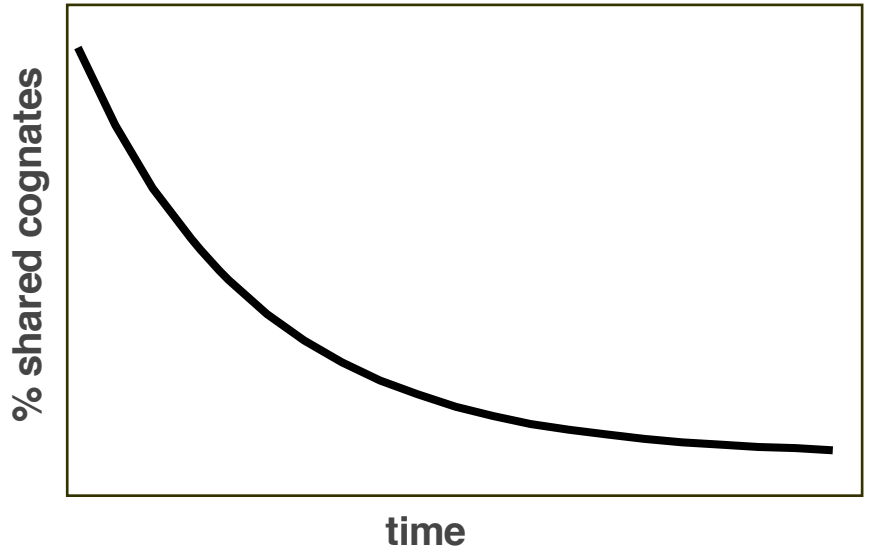
1 This type of family tree was suggested by Hall, 1950, p. 24.

Glottochronology

Loss of cognates happens at a constant rate
(=radioactive decay)

19% loss per 1000 years (Lees 1953)

$$time = \frac{\log(\%sharedcognates)}{2\log(retentionrate)}$$



The Rise of Lexicostatistics…

IN THE LAST DECADE glottochronology has excited international interest and acquired a literature of its own. To the anthropologist it promises a measure of time depth for language families without documented history, and yet another linguistic example of regularity in cultural phenomena.

Hymes (1960): "Lexicostatistics so far"

"... a significant work—one which may conceivably be as revolutionary for Oceanic linguistics and culture history as was the work of Greenberg (1949–54) for the interpretation of African languages and cultures"

Murdock (1964) p.117

…and the fall of Lexicostatistics

Major Criticism: Universality of Rates

Old Norse & Icelandic?

- Glottochronology: 200 years.
- Reality: 1000 years

Bergsland & Vogt 1962: “Our findings clearly disprove the basic assumption of glottochronology 'that fundamental vocabulary changes at a constant rate’ ”

Jungner, Hugo; Elisabeth Svärdström (1940-1971). Sveriges runinskrifter: V. Västergötlands runinskrifter. Stockholm: Kungl. Vitterhets Historie och Antikvitets Akademien. ISSN 0562-8016. p. 260

Fallout.

"a tradition of hostility towards probabilistic modelling in historical linguistics" (Sankoff '73)

"In summary, glottochronology is not accurate; all its basic assumptions have been severely criticized. It should not be accepted, it should be rejected" (Campbell '04)

"Linguists don't do dates" (McMahon & McMahon '03)

U.P.G.M.A

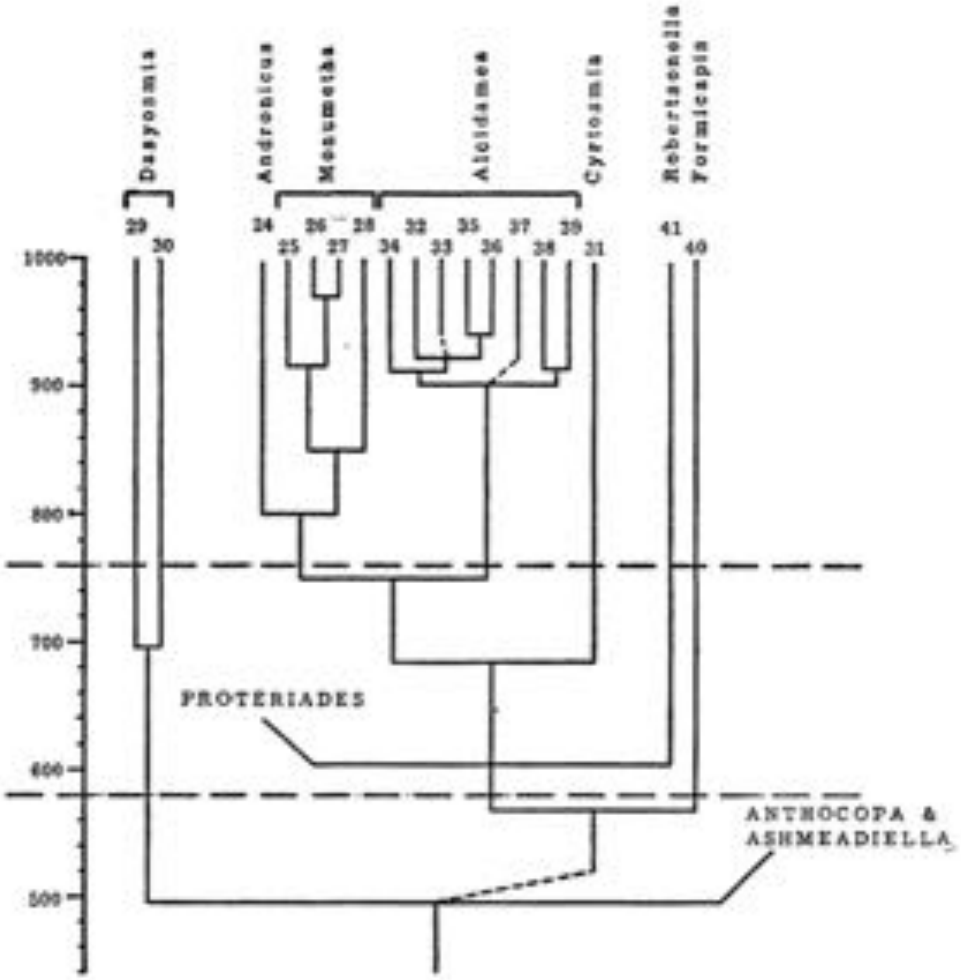


FIG. 6. Diagram of relationships for the genus *Hoplitis* obtained by the weighted variable group method.

A QUANTITATIVE APPROACH TO A PROBLEM IN CLASSIFICATION [1]

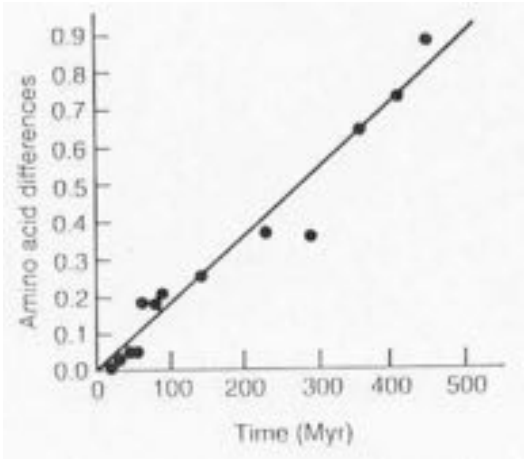
CHARLES D. MICHENER AND ROBERT R. SOKAL [2]

Department of Entomology, University of Kansas, Lawrence

Molecular Clock

Zuckerkandl & Pauling 1962

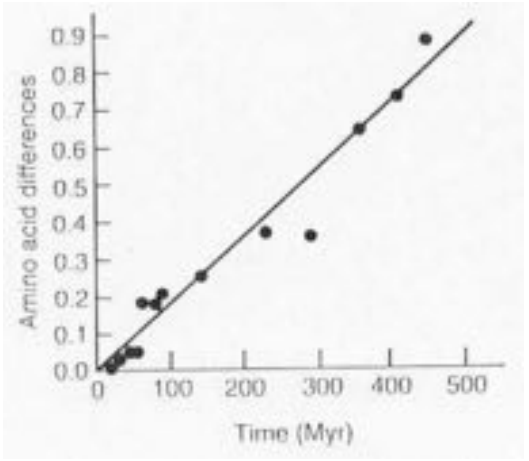
Number of AA differences were proportional to species divergence times.



Molecular Clock

Zuckerkandl & Pauling 1962

Number of AA differences were proportional to species divergence times.



Kimura 1968

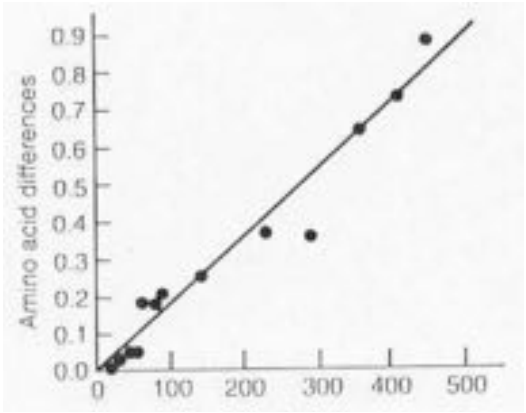
The average time taken for one base pair replacement within a genome is therefore

$$28 \times 10^6 \text{ yr} \div \left(\frac{4 \times 10^9}{300}\right) \div 1{\cdot}2 \doteqdot 1{\cdot}8 \text{ yr}$$

Molecular Clock

Zuckerkandl & Pauling 1962

Number of AA differences were proportional to species divergence times.



Kimura 1968

The average time taken for one base pair replacement within a genome is therefore

$$28 \times 10^6 \text{ yr} \div \left(\frac{4 \times 10^9}{300}\right) \div 1{\cdot}2 \doteqdot 1{\cdot}8 \text{ yr}$$

Sarich & Wilson 1967

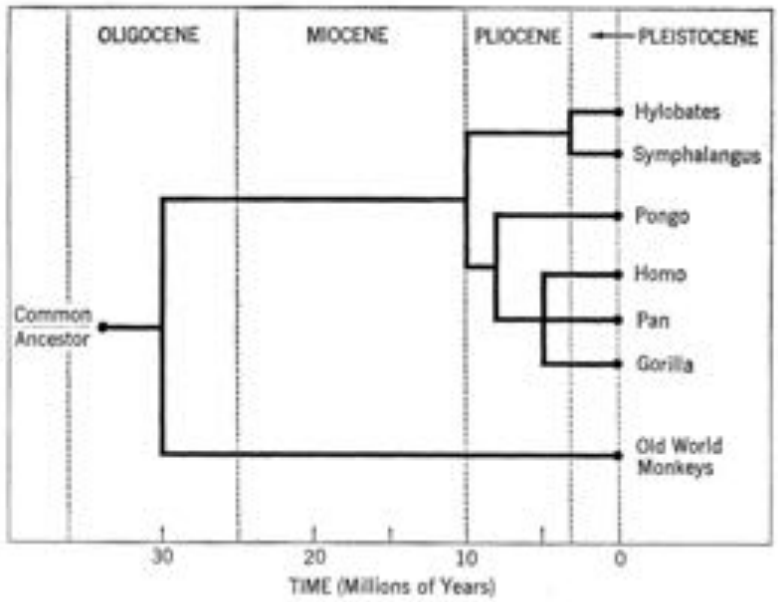


Fig. 1. Times of divergence between the various hominoids, as estimated from immunological data. The time of divergence of hominoids and Old World monkeys is assumed to be 30 million years.

Problems?

Kirsch 1969

SEROLOGICAL DATA AND PHYLOGENETIC INFERENCE:
THE PROBLEM OF RATES OF CHANGE

JOHN A. W. KIRSCH

Felsenstein 1978

CASES IN WHICH PARSIMONY OR COMPATIBILITY
METHODS WILL BE POSITIVELY MISLEADING[1]

JOSEPH FELSENSTEIN

Britten1986

Rates of DNA Sequence Evolution Differ
Between Taxonomic Groups

ROY J. BRITTEN

The Cladistics Wars

Solutions.

Cavalli-Sforza & Edwards 1967

Phylogenetic Analysis
Models and Estimation Procedures

L. L. CAVALLI-SFORZA AND A. W. F. EDWARDS*

Yang 1993

Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites[1]

Ziheng Yang

Sanderson 1997

A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy

Michael J. Sanderson

Drummond et al. 2006

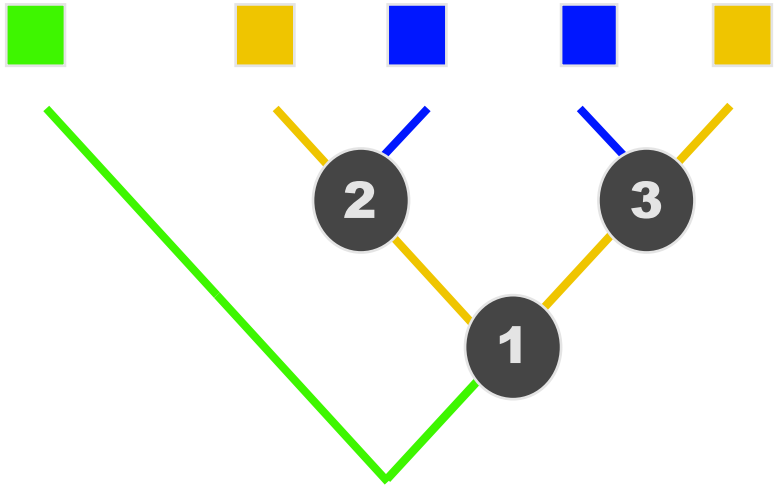
Relaxed Phylogenetics and Dating with Confidence

Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut

How do we **build** trees?

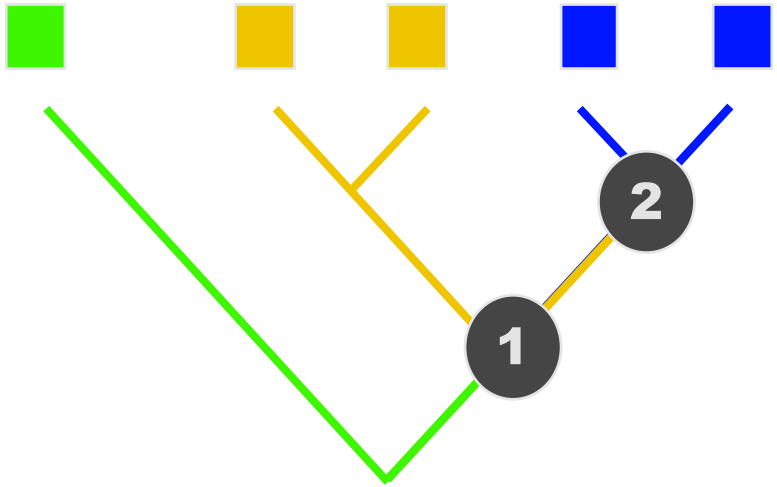
1. Maximum Parsimony.
2. Maximum Likelihood.
3. **Bayesian Phylogenetic Analyses.**

Maximum Parsimony



Aim to find “most parsimonious” tree

Smallest amount of evolution



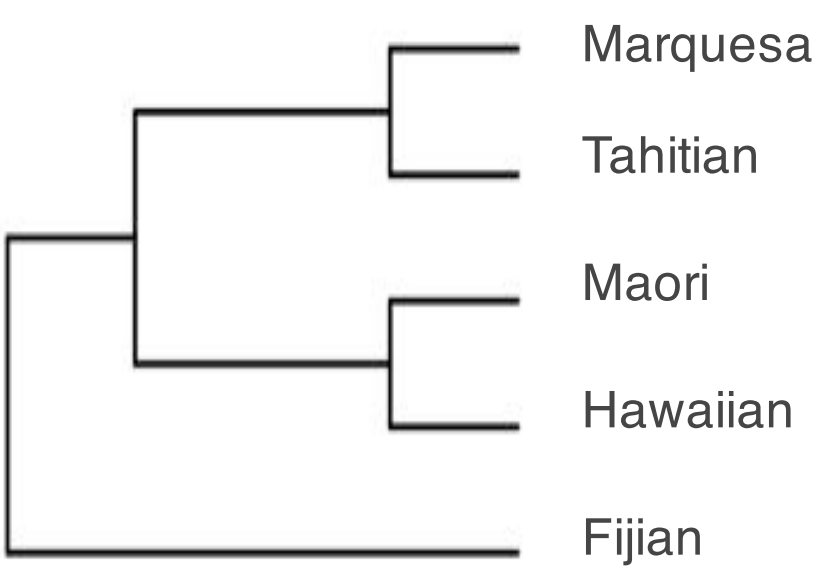
Unlikely that things should arise more than once

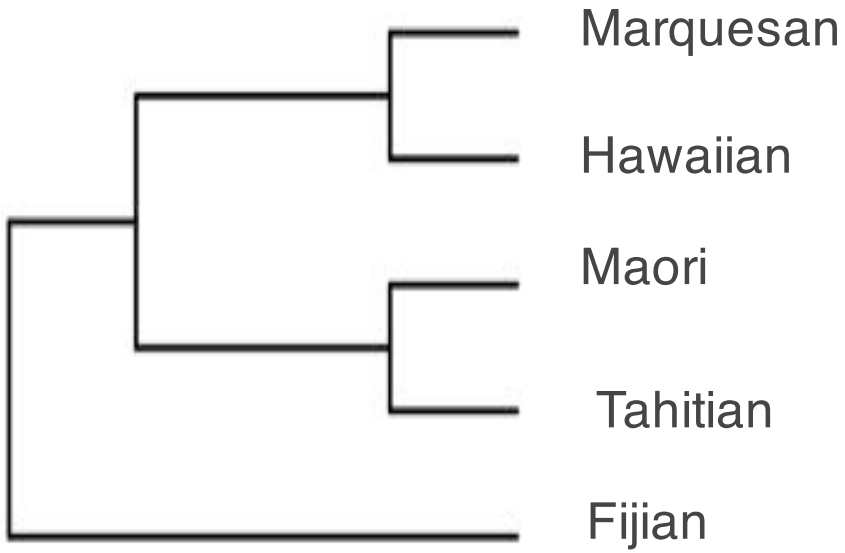
(i.e. cognates should innovate once)

	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

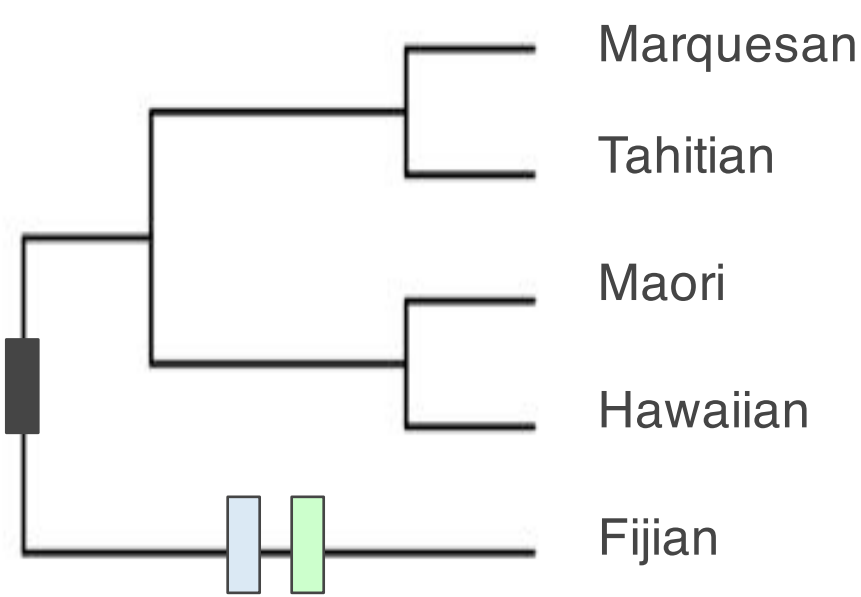
	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

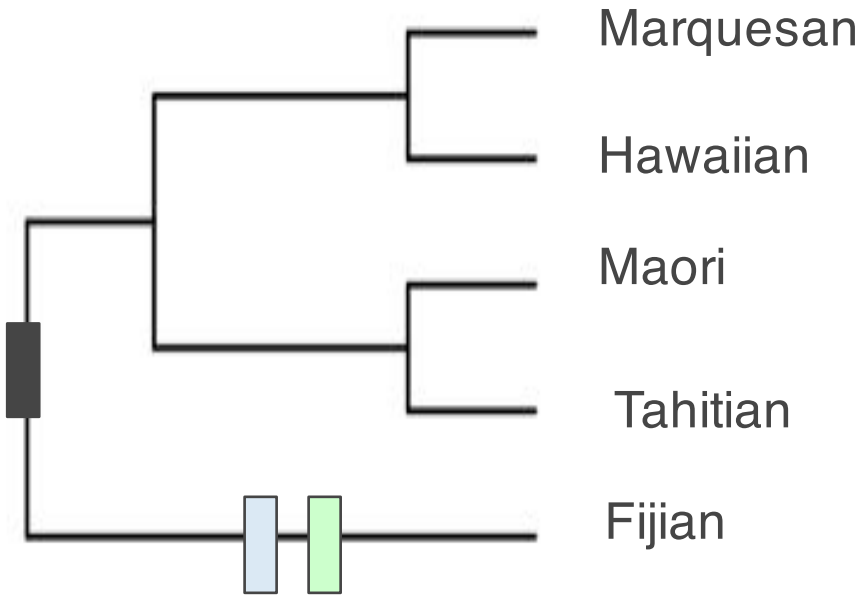




Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

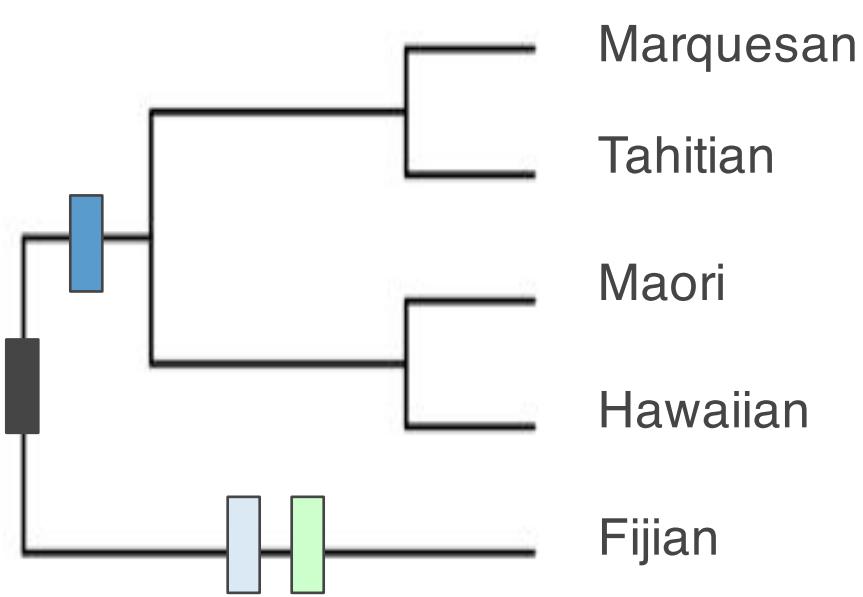


Length=3

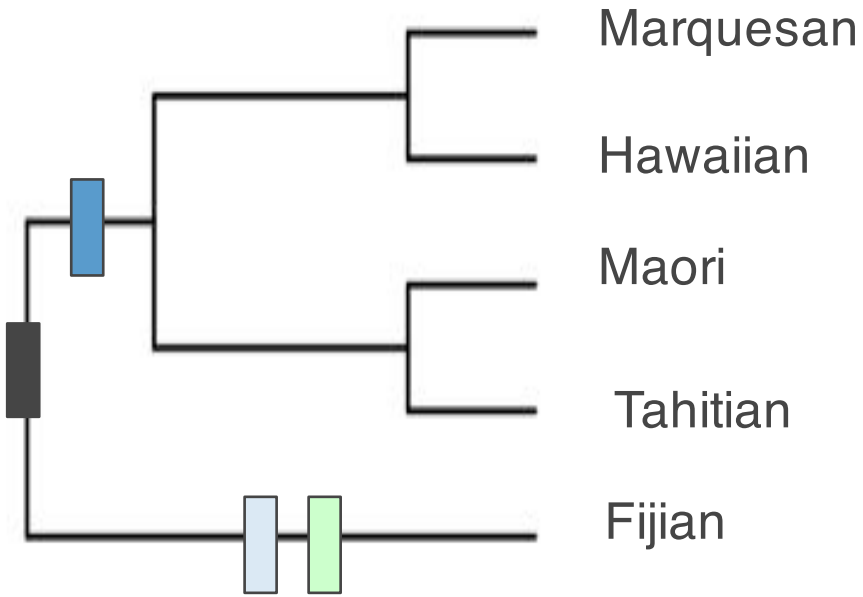


Length=3

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

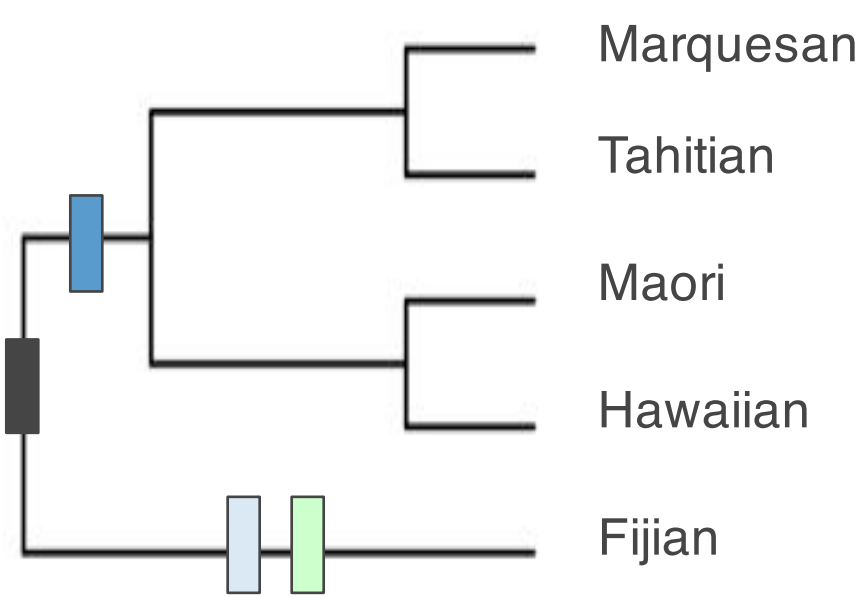


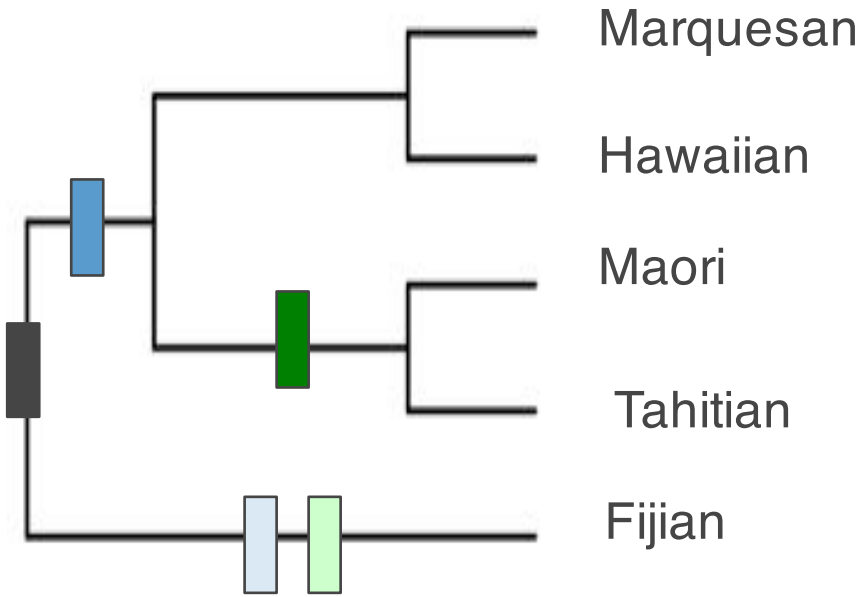
Length=4



Length=4

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

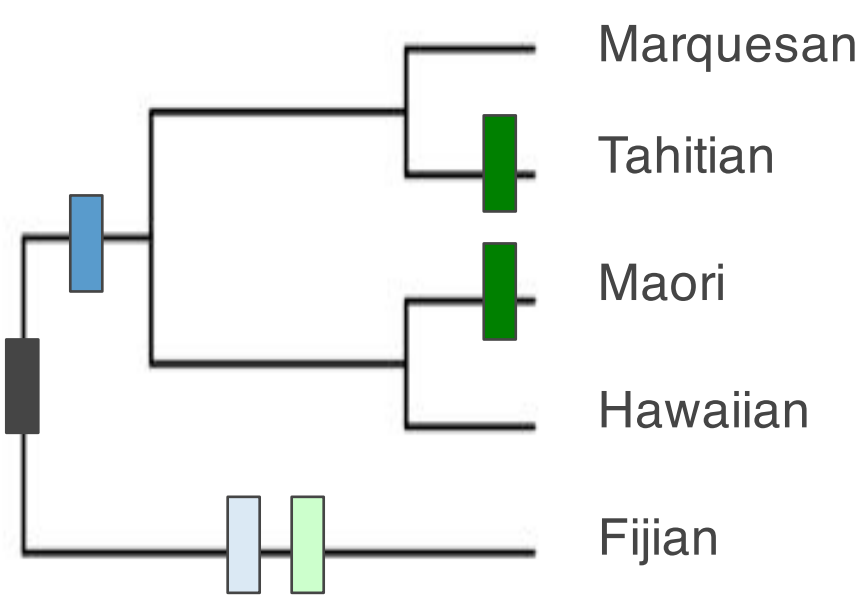


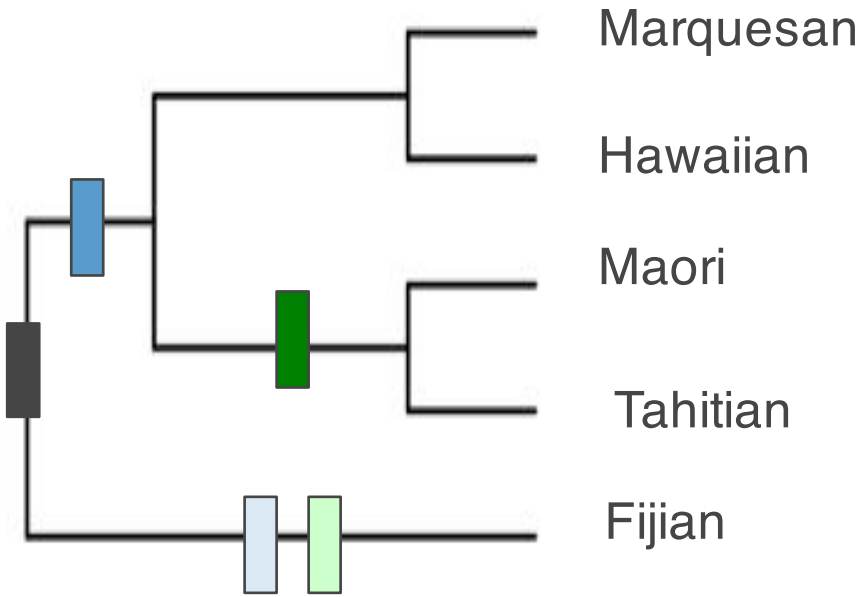


Length=4

Length=5

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

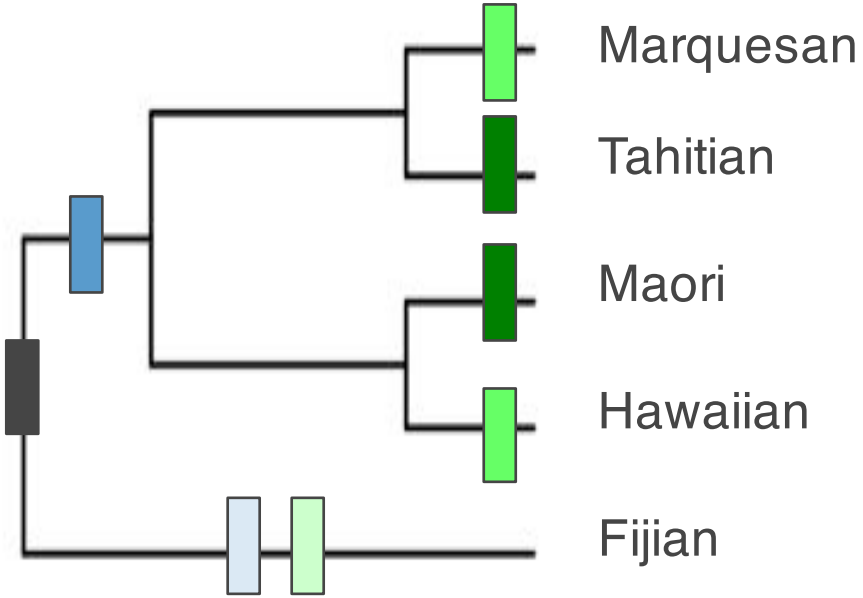


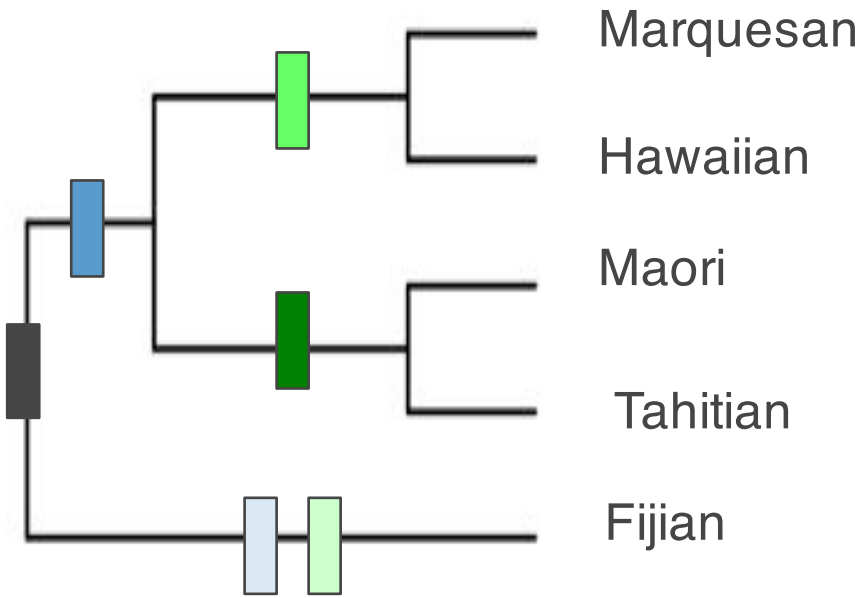


Length=6

Length=5

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

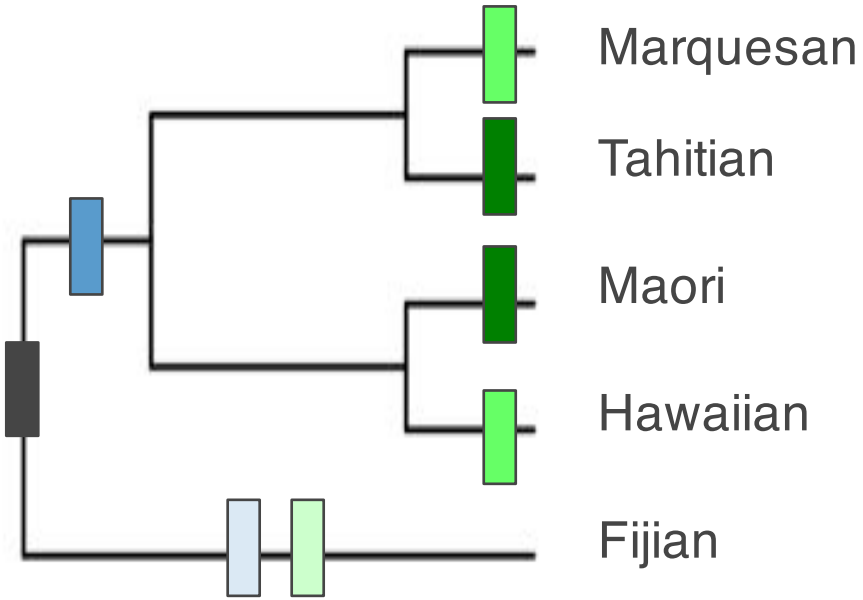


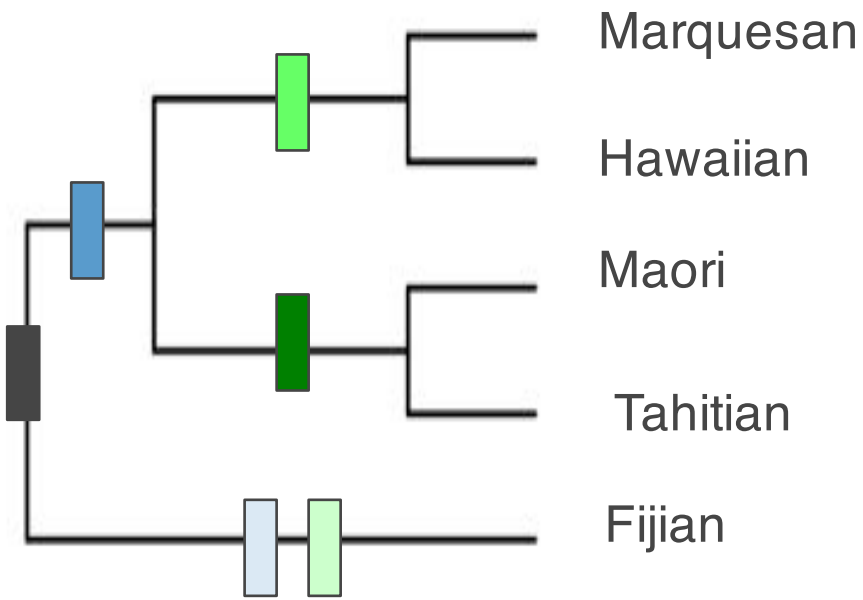


Length=8

Length=6

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**



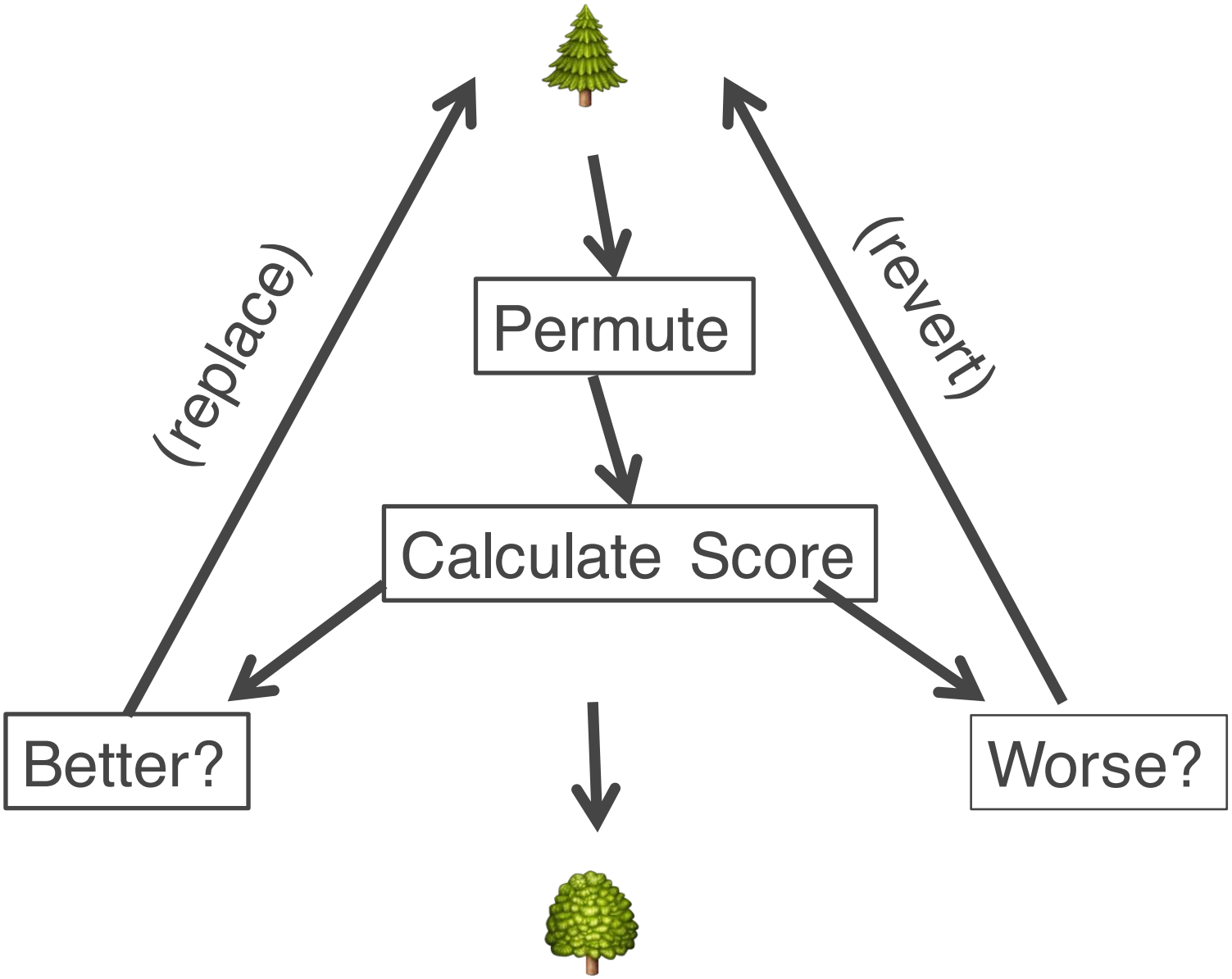


Length=8

Length=6

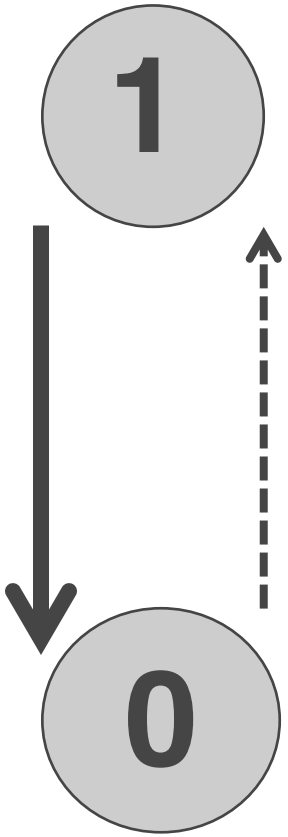
Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

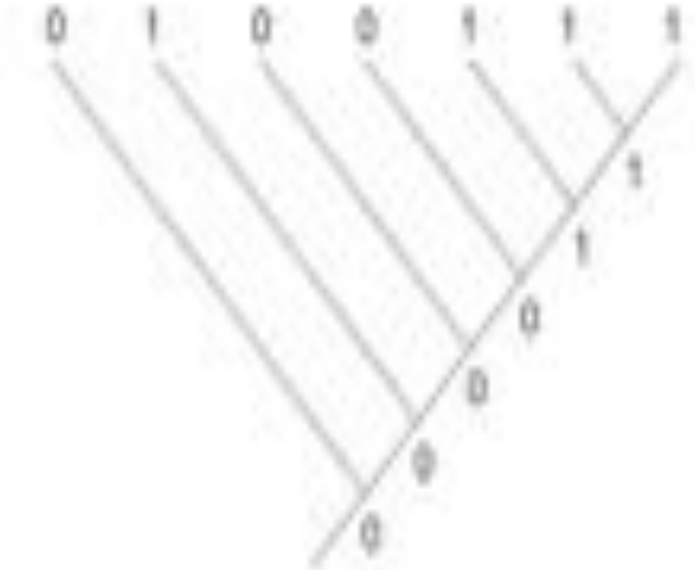
Algorithm



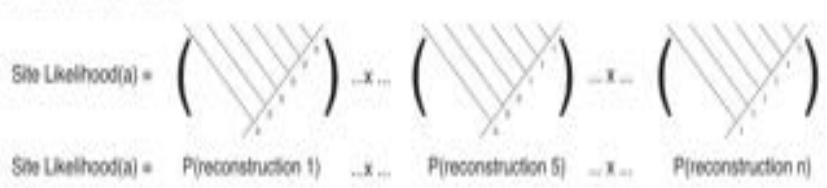
Maximum Likelihood

- Builds on Max. Parsimony
- Stochastic **model** of change.
- **Likelihood** = fit of data to tree under a model.
  - Very small number = log(Lh)
  - Closer to zero = better fit.





$$L(a) = P(0 \rightarrow 0|b1) \times P(0 \rightarrow 0|b2) \times P(1 \rightarrow 1|b3) \times P(1 \rightarrow 0|b4) \times P(0 \rightarrow 0|b5) \times P(0 \rightarrow 0|b6) \times P(0 \rightarrow 0|b7) \times P(0 \rightarrow 1|b8) \times P(1 \rightarrow 1|b9) \times P(1 \rightarrow 1|b10) \times P(1 \rightarrow 1|b11) \times P(1 \rightarrow 1|b12)$$



Lh =

P of being in state 0, and staying state 0 on branch 1.

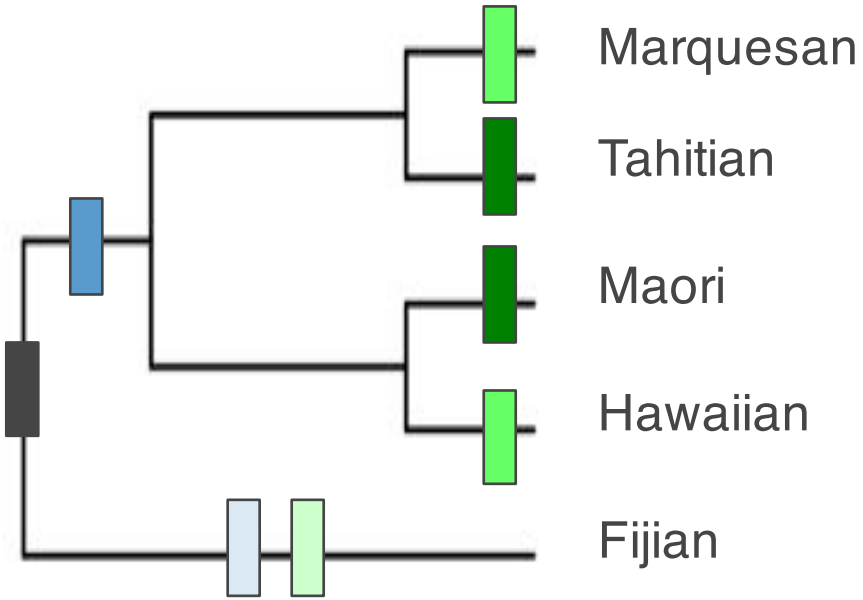
x

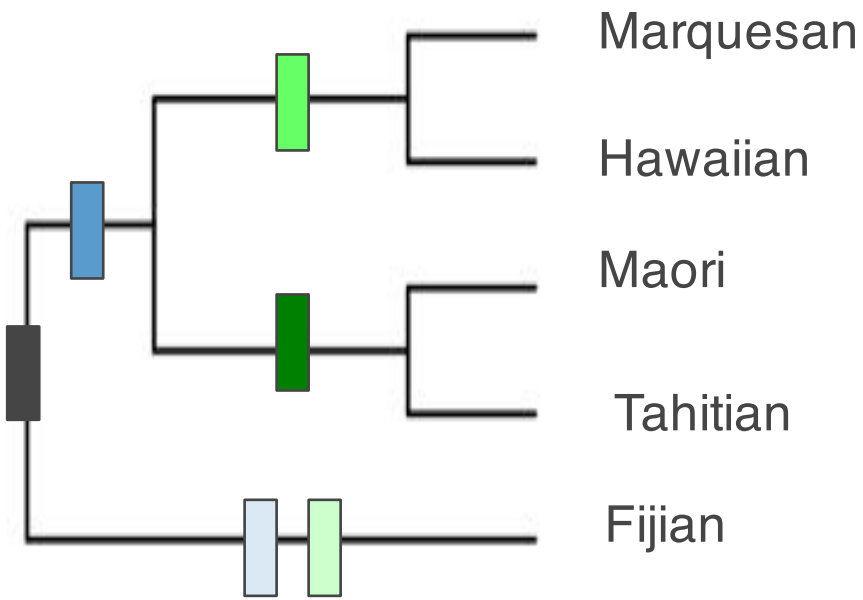
P of being in state 0, and staying state 0 on branch 2.

x

P of being in state 0, and staying state 0 on branch 2.

…. etc



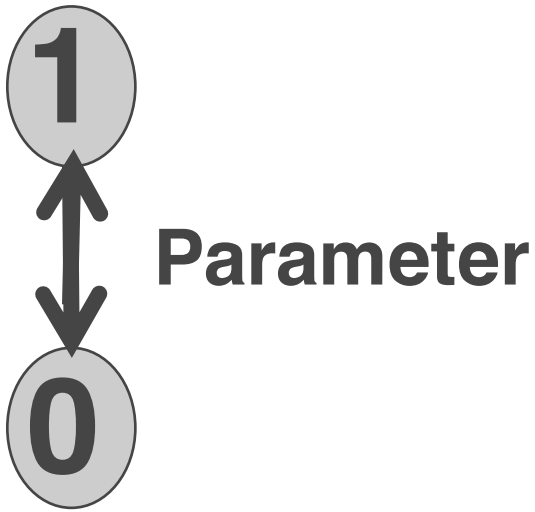


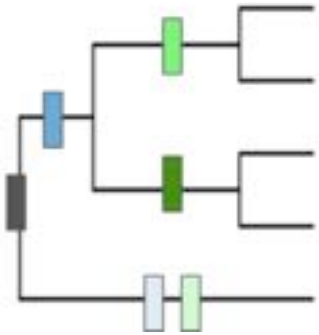
Ln(L)= -14.804

Ln(L)= -12.007

Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

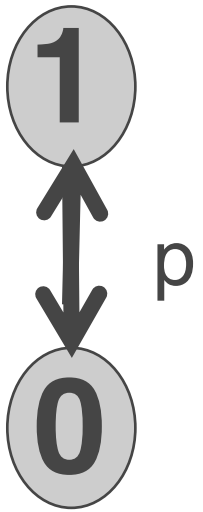
Models

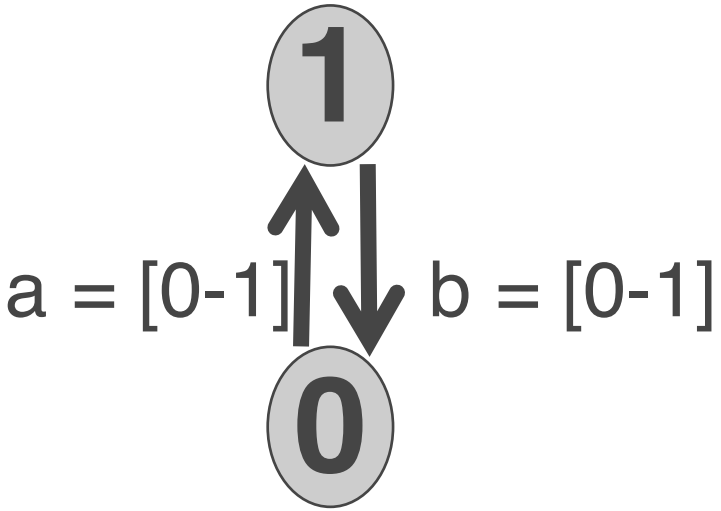


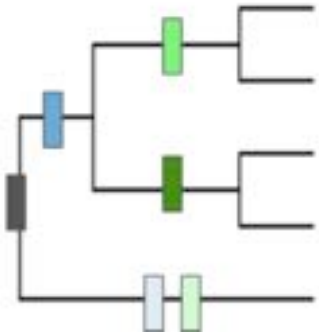


Ln(L)= -12.007

Models

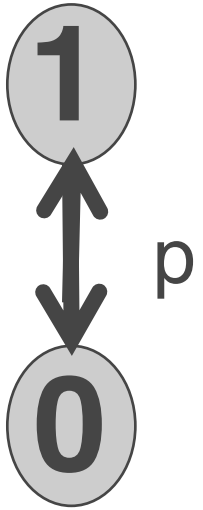


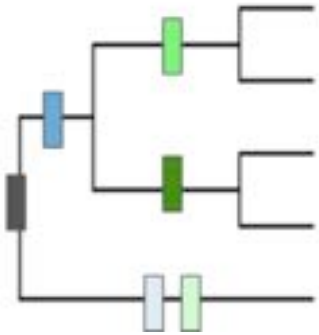




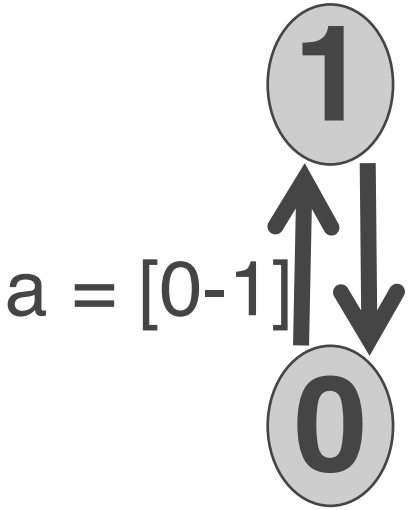
Ln(L)= -12.007

Models

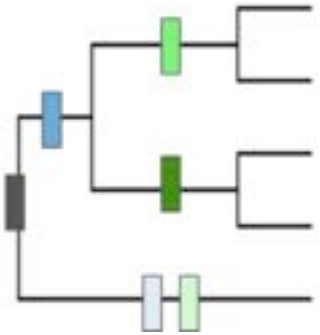




Ln(L)= -12.007



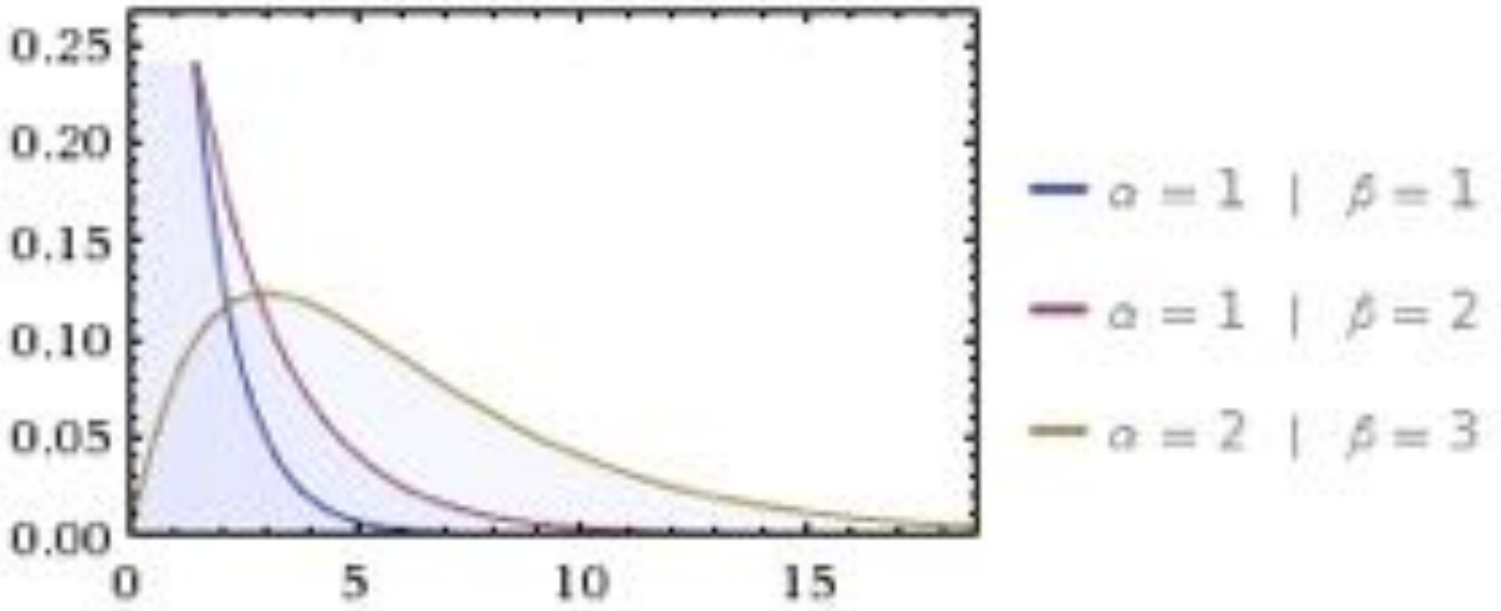
a = 0.92, b = 0.08

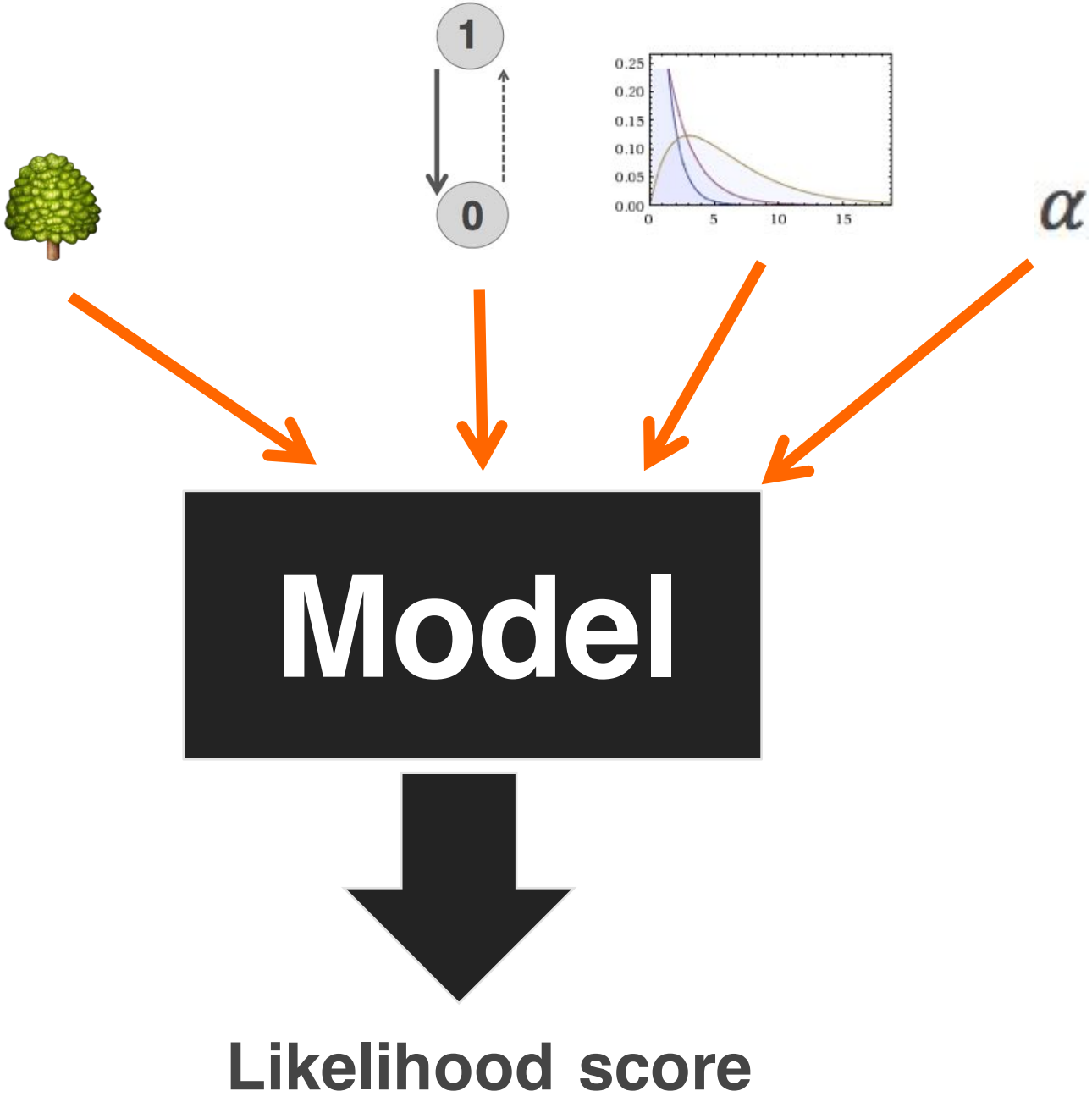


Ln(L)= -9.072

Rate Variation - Gamma

- Gamma Distribution
- One parameter, α, controls the shape
- Estimate the best value for α using Lh.




1
0
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
α
Model
Likelihood score

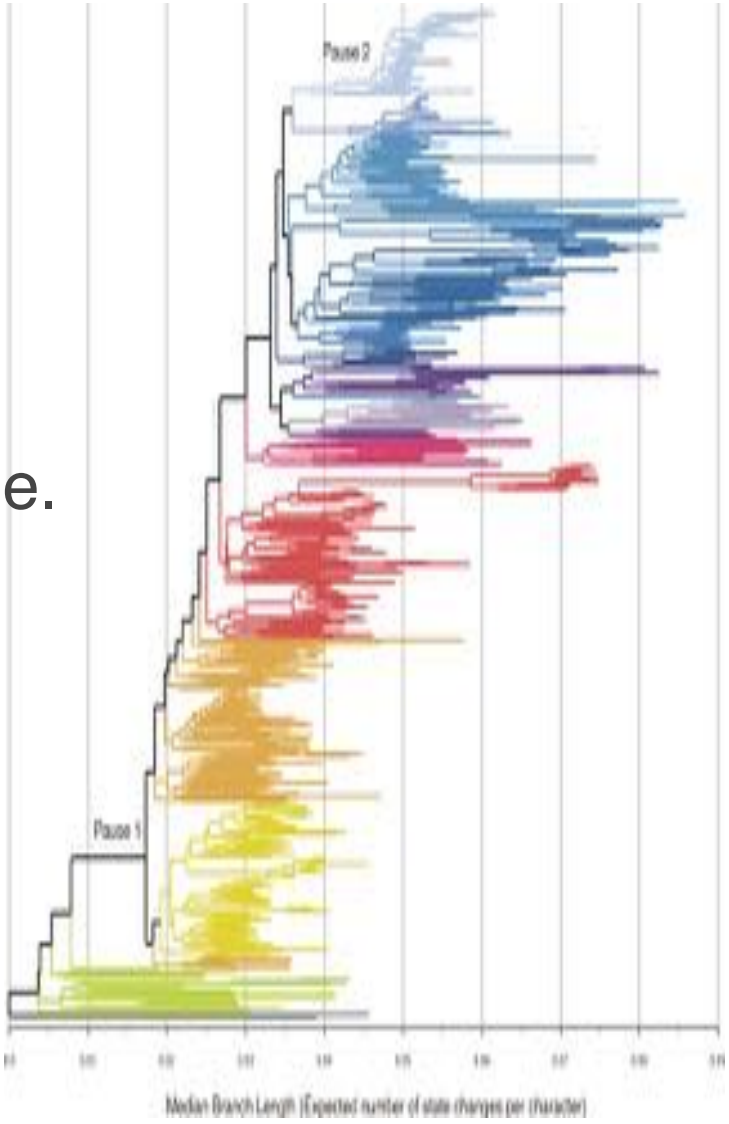
Linguists don’t do dates.

Phylogenetic Dating

Estimate how long branches are.

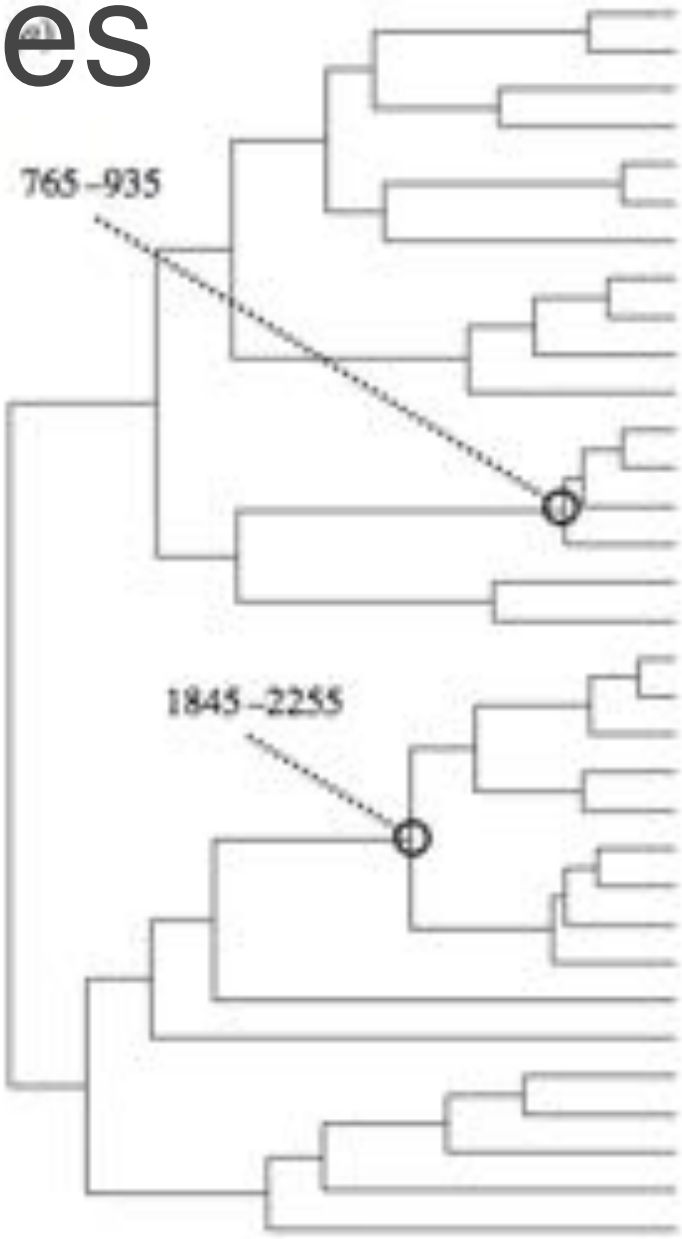
(number of changes per cognate set)

Awesome: Not a global retention rate (i.e. glottochronology) but a **per-language** estimate of the amount of change.



Convert Rates to Dates

- Use (pre)historical information to calibrate nodes
  - e.g. Archaeology suggests initial settlement was..
  - e.g. Historical evidence says that X and Y were separate by…
- Smooth rates over these calibrations

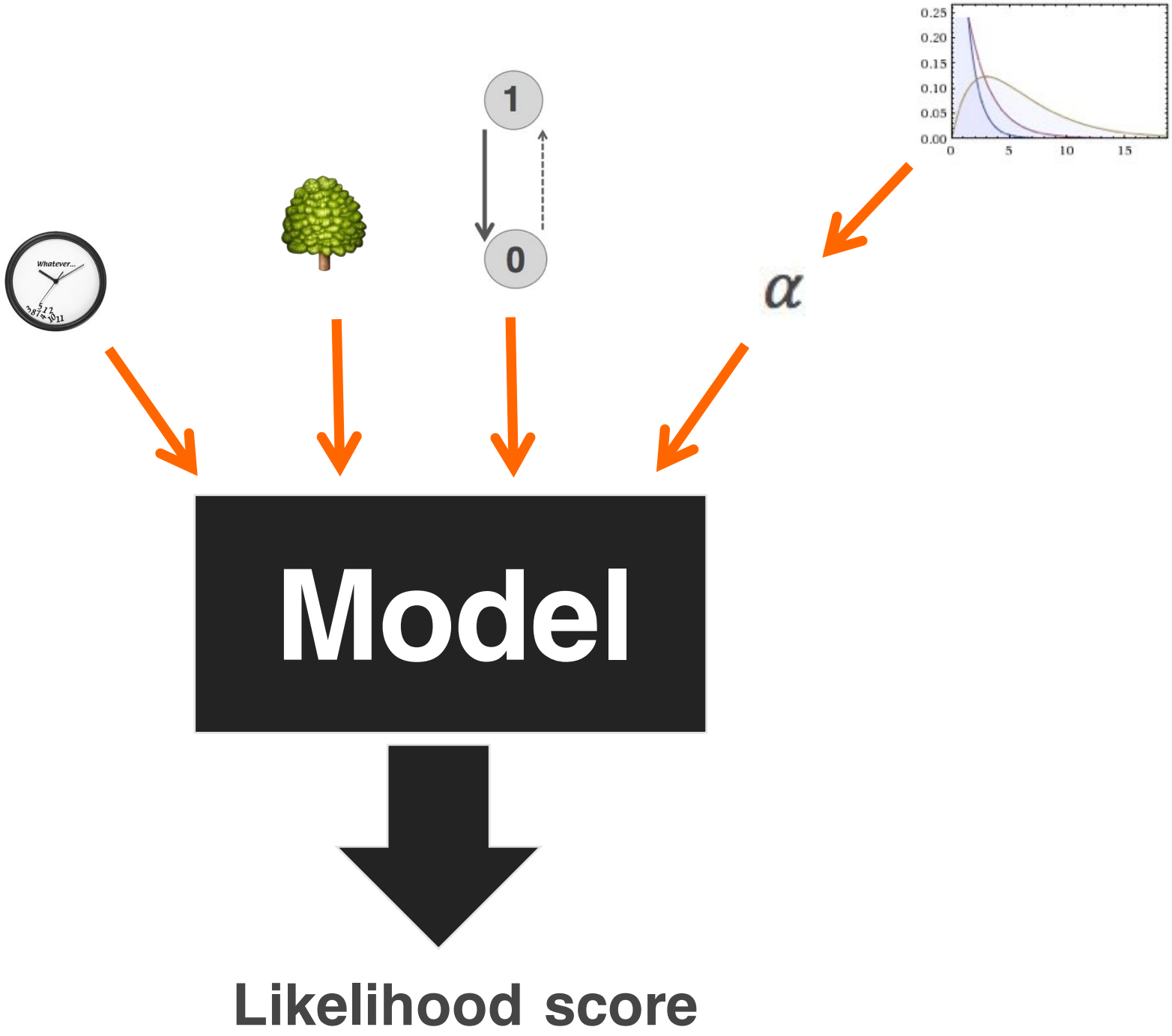


Strict Clock

- One rate for all languages.
- No variation
- = glottochronology

Relaxed Clock

- Allows rate to vary across branches.
- Rates are drawn from a parametric distribution with parameters estimated from the data


1
0
α
Model
Likelihood score

Problem: uncertainty

- ML and MP give **a point** estimate.
- But a single tree is not enough.
  - Reality is complicated.
  - Need to estimate uncertainty around that estimate.
  - “confidence intervals” = how confident can I be about my estimate.
- MP/ML Solution is “bootstrapping” = painfully labor intensive.

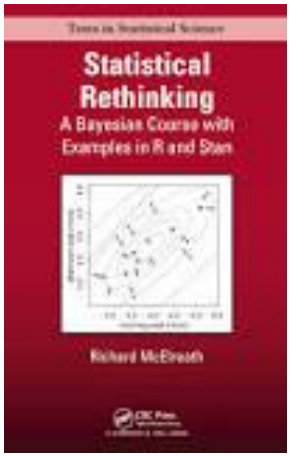
Bayesian methods

1. Explicitly account for uncertainty

2. Directly incorporate different information (*priors*)

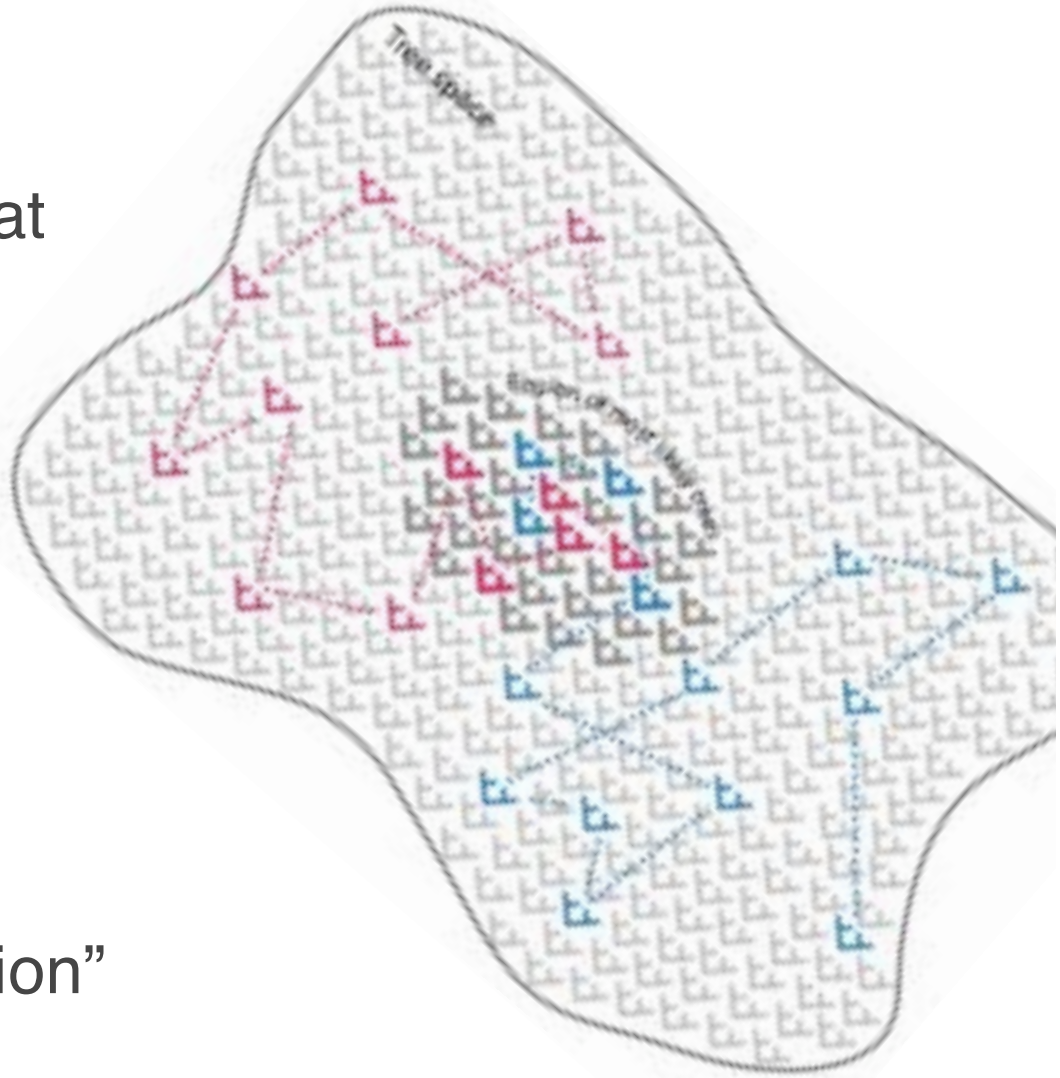
3. It's 2016. Everyone is Bayesian.

http://www.stat.wisc.edu/~ane/bot940/bayes.pdf
McElreath "Statistical Rethinking"
(http://xcelab.net/rm/statistical-rethinking/)



Bayesian Phylogenetics

1. Data & Model & Tree
2. Sample a Tree
3. Calculate the **Likelihood** of that tree
4. Modify the tree or a parameter
5. Repeat (MCMC walk through treespace)



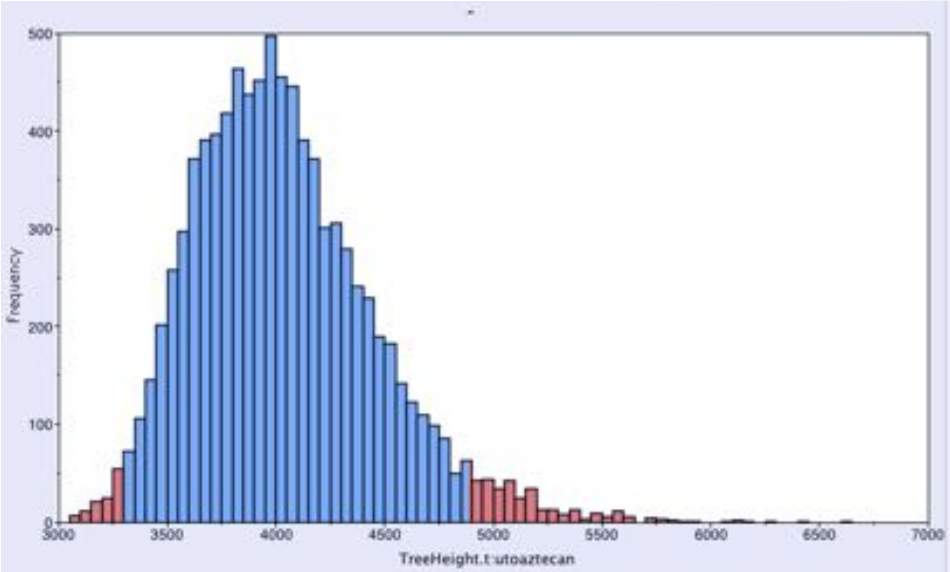
Samples best fitting trees from all possible trees.

- a distribution of trees
- = “Posterior Probability Distribution”

Posterior Probability Distribution:

Uto-Aztecan Languages

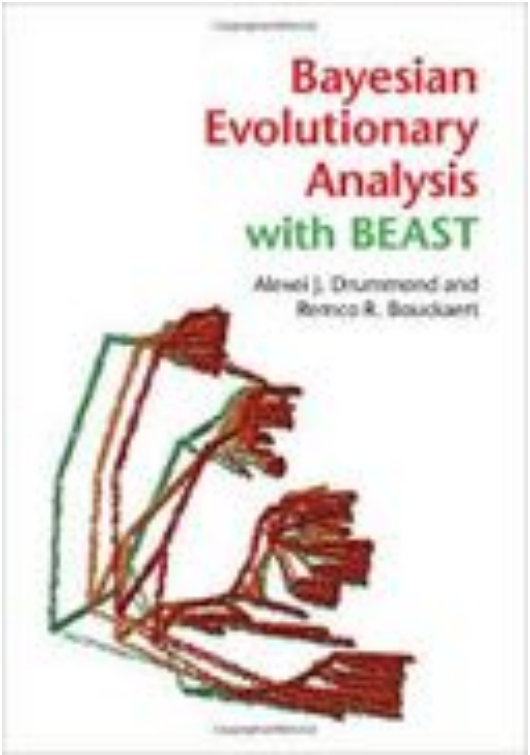
Sampled 10,000 trees.
Draw each one (Densitree, Bouckaert ‘10)
Some well supported regions
Some less…
Some conflict



95% HPD: 3302 – 4884 years

Summary

- Many conceptual parallels across disciplines.
- Grew out of many of the same concerns found in Linguistics/Anthropology (innovations vs retentions, rate variation, conflicting signal, etc).
- Wide range of techniques available, but Bayesian methods are clear winners to date.
- Practical: My BEAST tutorial in QMSS materials, BEAST Tutorial online, BEAST book.



Questions?