Köln 2018-04-27

Phylogenetics: Why and How

T. Mark Ellison

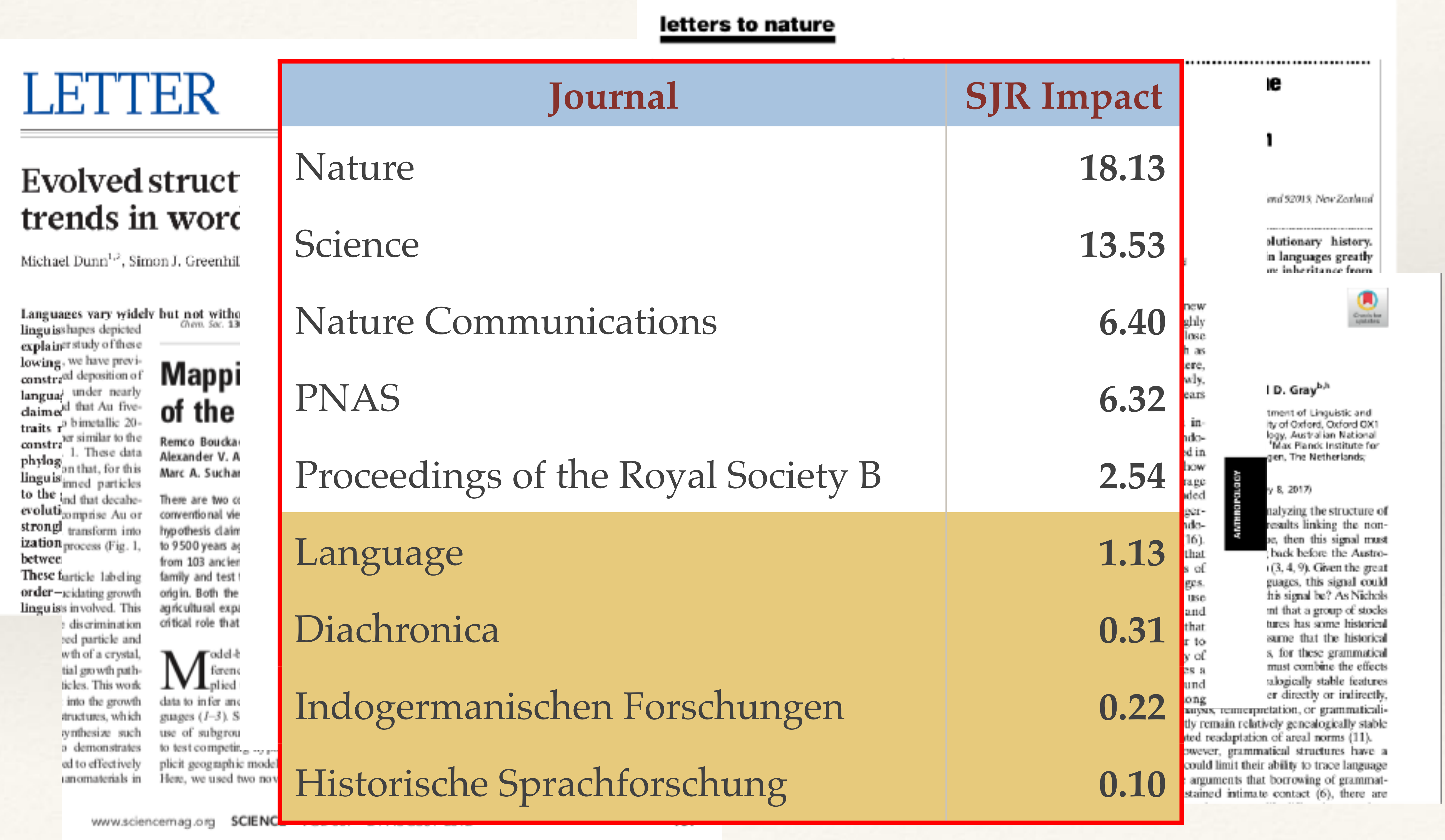
http://phylogenet.ics.pw

What is Coming Up

- Why the biological phylogenetic approach to the history of languages get funding and media attention
- How biological phylogenetics works
- Limitations of this approach

Why Phylogenetics gets Attention

The Credibility Gap



letters to nature

LETTER

Evolved struct
trends in worc

Michael Dunn, Simon J. Greenhil

Mappi
of the

Journal	SJR Impact
Nature	18.13
Science	13.53
Nature Communications	6.40
PNAS	6.32
Proceedings of the Royal Society B	2.54
Language	1.13
Diachronica	0.31
Indogermanischen Forschungen	0.22
Historische Sprachforschung	0.10

The Credibility Gap

Publications in these generalist journals are picked up in the popular press around the world.

Here is a paper from PNAS:

(20x SJR *impact factor* of Diachronica)

PNAS

Ultraconserved words point to deep language ancestry across Eurasia

Mark Pagel[a,b,1], **Quentin D. Atkinson**[c], **Andreea S. Calude**[d], **and Andrew Meade**[a]

[a]School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AS, United Kingdom; [b]Santa Fe Institute, Santa Fe, NM 87501; [c]School of Psychology, University of Auckland, Auckland 1142, New Zealand; and [d]Linguistics Programme, University of Waikato, Hamilton 3240, New Zealand

Edited* by Colin Renfrew, University of Cambridge, Cambridge, United Kingdom, and approved April 15, 2013 (received for review October 31, 2012)

The search for ever deeper relationships among the World's languages is bedeviled by the fact that most words evolve too rapidly to preserve evidence of their ancestry beyond 5,000 to 9,000 y. On the other hand, quantitative modeling indicates that some "ultraconserved" words exist that might be used to find evidence for deep linguistic relationships beyond that time barrier. Here we use a statistical model, which takes into account the frequency with which words are used in common everyday speech, to predict the existence of a set of such highly conserved words among seven language families of Eurasia postulated to form a linguistic superfamily that evolved from a common ancestor around 15,000 y ago. We derive a dated phylogenetic tree of this proposed superfamily with a time-depth of ∼14,450 y, implying that some frequently used words have been retained in related forms since the end of the last Ice age. Words used more than once per 1,000 in everyday speech were 7- to 10-times more likely to show deep ancestry on this tree. Our results suggest a remarkable fidelity in the transmission of some words and give theoretical justification to the search for features of language that might be preserved across wide spans of time and geography.

cultural evolution | phylogeny | historical linguistics

The English word *brother* and the French *frère* are related to the Sanskrit *bhrātar* and the Latin *frāter*, suggesting that words as mere sounds can remain associated with the same meaning for millennia. But how far back in time can traces of a word's genealogical history persist, and can we predict which words are

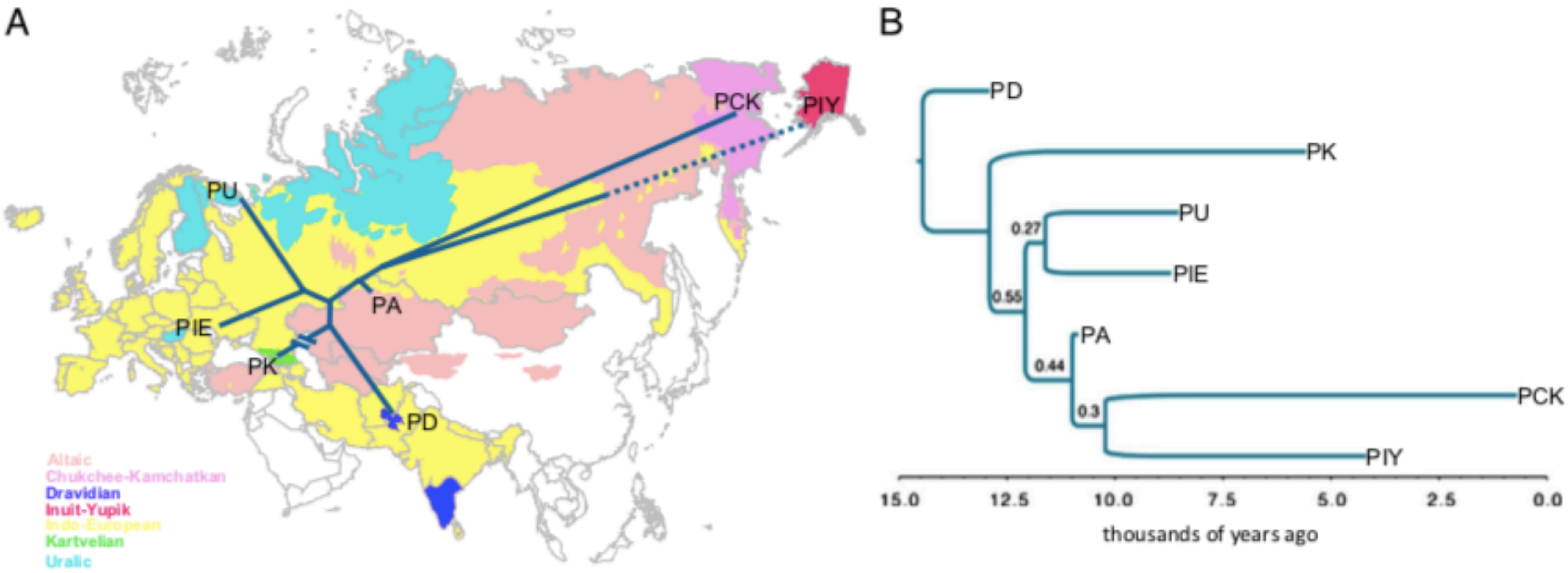
words have about a 50% chance of being replaced by a new noncognate word [a word's linguistic half-life (14, 15)], roughly every 2,000–4,000 y, consistent with the belief that words lose traces of their ancestry quickly. However, some words, such as the numerals, pronouns, and special adverbs (e.g., I, you, here, how, not, there, what, two, five) are replaced far more slowly, with half-lives of once every 10,000, 20,000 or even more years (14, 15).

Usefully, these words can be predicted from information independent of their sounds. We showed in a sample of Indo-European languages that the frequency with which a word is used in everyday speech, along with its part of speech, can predict how rapidly words evolve, with frequently used words on average retained for longer periods of time (14). We have recently extended this result to include speakers from the Uralic, Sino-Tibetan, Niger-Congo, Altaic, and Austronesian families, in addition to Indo-European, plus the isolate Basque and the Creole Tok Pisin (16). Even in languages as widely divergent as these, we found that a measure of the average frequency of use predicted rates of lexical replacement as estimated in the Indo-European languages.

Taken together, these findings suggest that the way we use a core set of vocabulary words in everyday speech is a stable and shared feature of human discourse, and raises the possibility that words will evolve in other language families at rates similar to those found in the Indo-European languages, with frequency of word-use acting as the common causal factor. This provides a statistical framework for predicting—without recourse to sound correspondences—words likely to show deep ancestry among

ANTHROPOLOGY

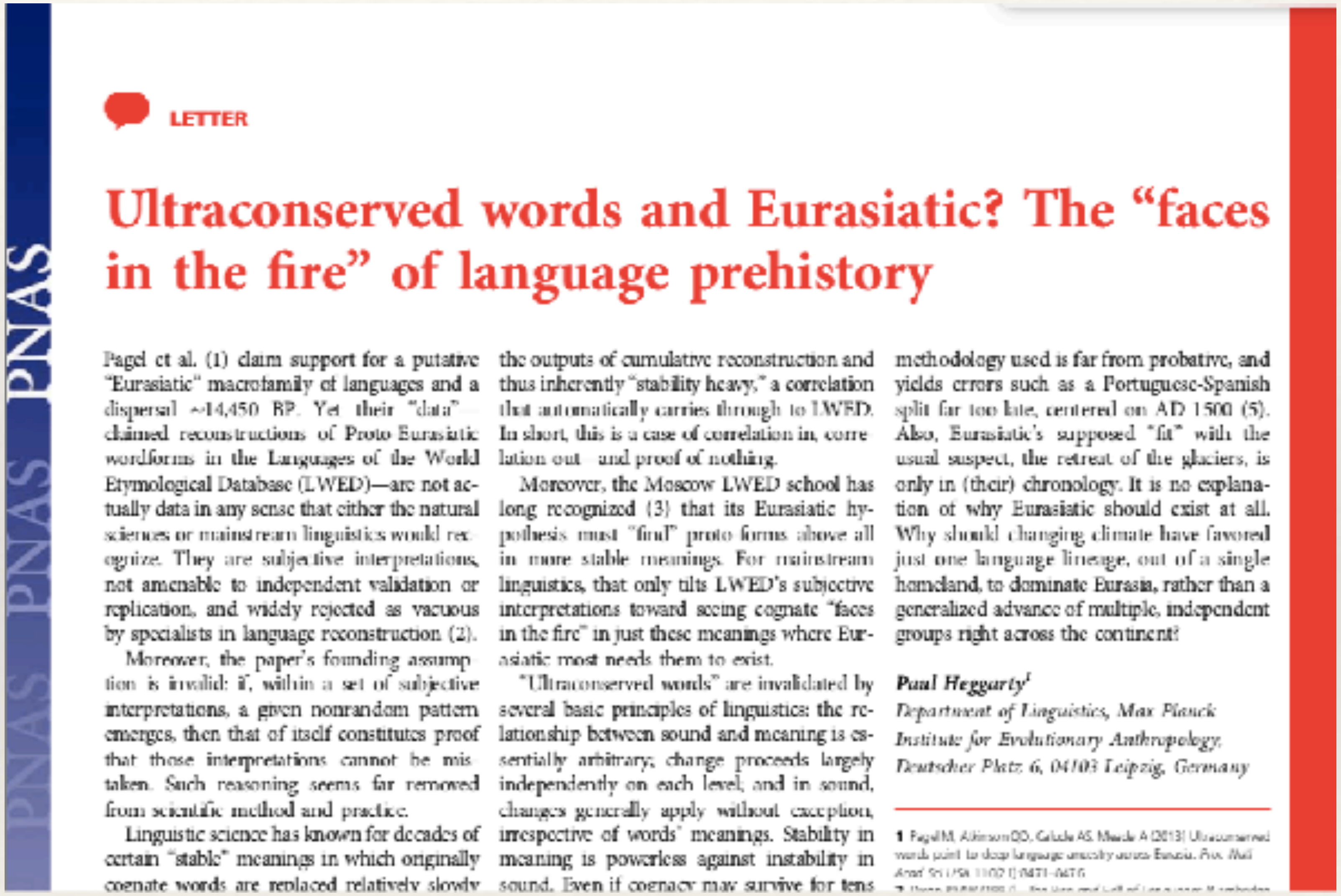
The Credibility Gap



The Credibility Gap

Heggarty (2013)

Yet their "data"… are not actually data in any sense that either the natural sciences or mainstream linguistics would recognize. They are subjective interpretations, not amenable to independent validation or replication, and widely rejected as vacuous by specialists in language reconstruction (2).



LETTER

Ultraconserved words and Eurasiatic? The "faces in the fire" of language prehistory

Pagel et al. (1) claim support for a putative "Eurasiatic" macrofamily of languages and a dispersal ~14,450 BP. Yet their "data"—claimed reconstructions of Proto Eurasiatic wordforms in the Languages of the World Etymological Database (LWED)—are not actually data in any sense that either the natural sciences or mainstream linguistics would recognize. They are subjective interpretations, not amenable to independent validation or replication, and widely rejected as vacuous by specialists in language reconstruction (2).

Moreover, the paper's founding assumption is invalid: if, within a set of subjective interpretations, a given nonrandom pattern emerges, then that of itself constitutes proof that those interpretations cannot be mistaken. Such reasoning seems far removed from scientific method and practice.

Linguistic science has known for decades of certain "stable" meanings in which originally cognate words are replaced relatively slowly

the outputs of cumulative reconstruction and thus inherently "stability heavy," a correlation that automatically carries through to LWED. In short, this is a case of correlation in, correlation out—and proof of nothing.

Moreover, the Moscow LWED school has long recognized (3) that its Eurasiatic hypothesis must "find" proto forms above all in more stable meanings. For mainstream linguistics, that only tilts LWED's subjective interpretations toward seeing cognate "faces in the fire" in just these meanings where Eurasiatic most needs them to exist.

"Ultraconserved words" are invalidated by several basic principles of linguistics: the relationship between sound and meaning is essentially arbitrary; change proceeds largely independently on each level; and in sound, changes generally apply without exception, irrespective of words' meanings. Stability in meaning is powerless against instability in sound. Even if cognacy may survive for tens

methodology used is far from probative, and yields errors such as a Portuguese-Spanish split far too late, centered on AD 1500 (5). Also, Eurasiatic's supposed "fit" with the usual suspect, the retreat of the glaciers, is only in (their) chronology. It is no explanation of why Eurasiatic should exist at all. Why should changing climate have favored just one language lineage, out of a single homeland, to dominate Eurasia, rather than a generalized advance of multiple, independent groups right across the continent?

Paul Heggarty[1]
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

---

1 Pagel M, Atkinson QD, Calude AS, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proc Natl Acad Sci USA* 110(21):8471–8476.

Paget response: http://www.pnas.org/content/pnas/early/2013/08/01/1309838110.full.pdf

The Credibility Gap

In the press in the:

Washington Post →

The Economist

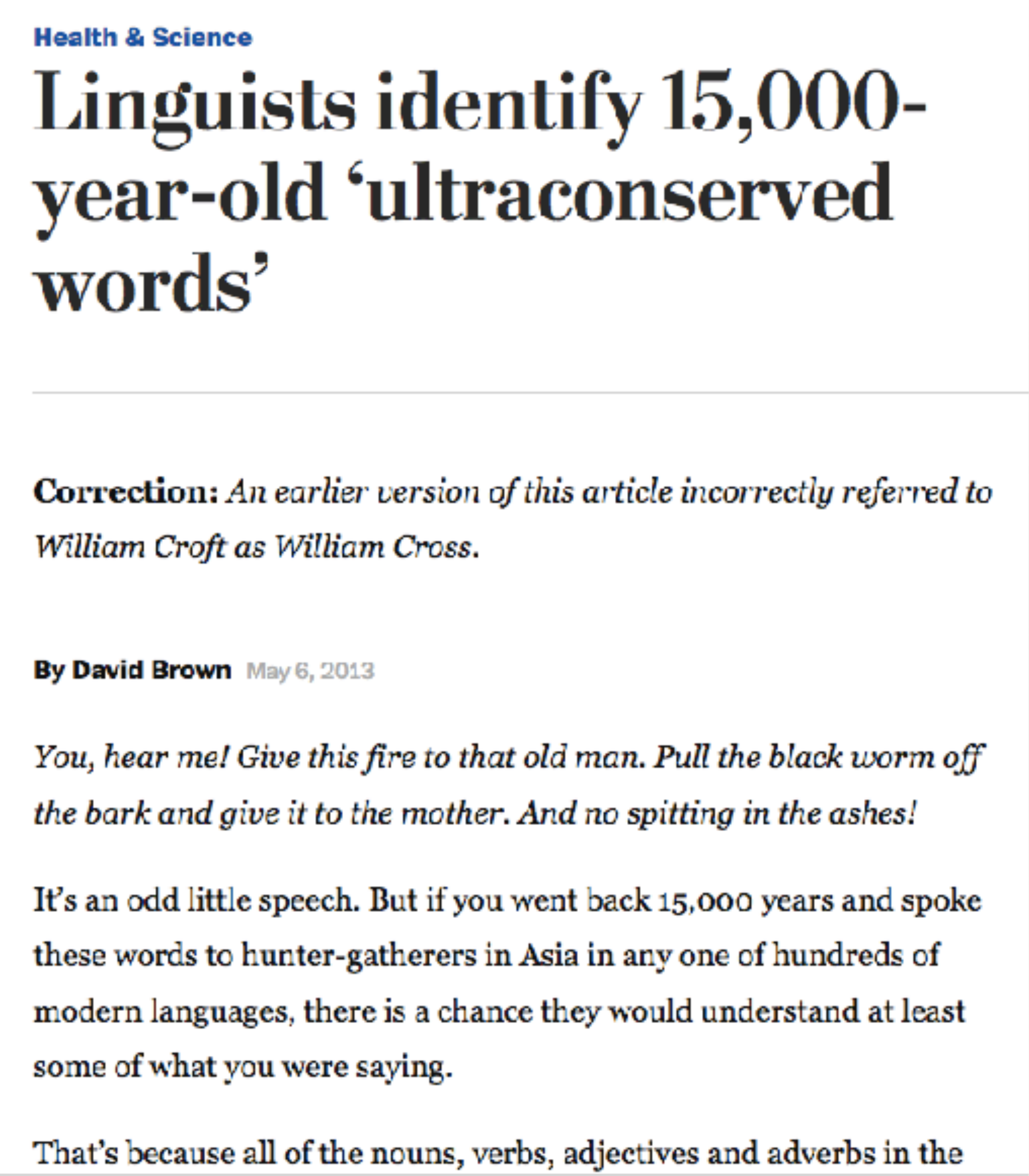
The New York Times

The New Scientist

The Conversation

It should not be dismissed out of hand just because it does not respect the limitations of traditional historical linguistics.

https://theconversation.com/thinking-the-unthinkable-tracing-language-back-15-000-years-14251



Health & Science

Linguists identify 15,000-year-old 'ultraconserved words'

Correction: *An earlier version of this article incorrectly referred to William Croft as William Cross.*

By David Brown May 6, 2013

You, hear me! Give this fire to that old man. Pull the black worm off the bark and give it to the mother. And no spitting in the ashes!

It's an odd little speech. But if you went back 15,000 years and spoke these words to hunter-gatherers in Asia in any one of hundreds of modern languages, there is a chance they would understand at least some of what you were saying.

That's because all of the nouns, verbs, adjectives and adverbs in the

The Credibility Gap

- Public perceptions < mass media < high-prestige ‘generalist’ journals
    - not specialist journals
- high-impact-factor journals portrayed as *modern*
    - by strong association with hard sciences
    - contrast specialist journals as old-fashioned
- historical linguists, computational or otherwise, should understand this type of work enough to be able to comment appropriately

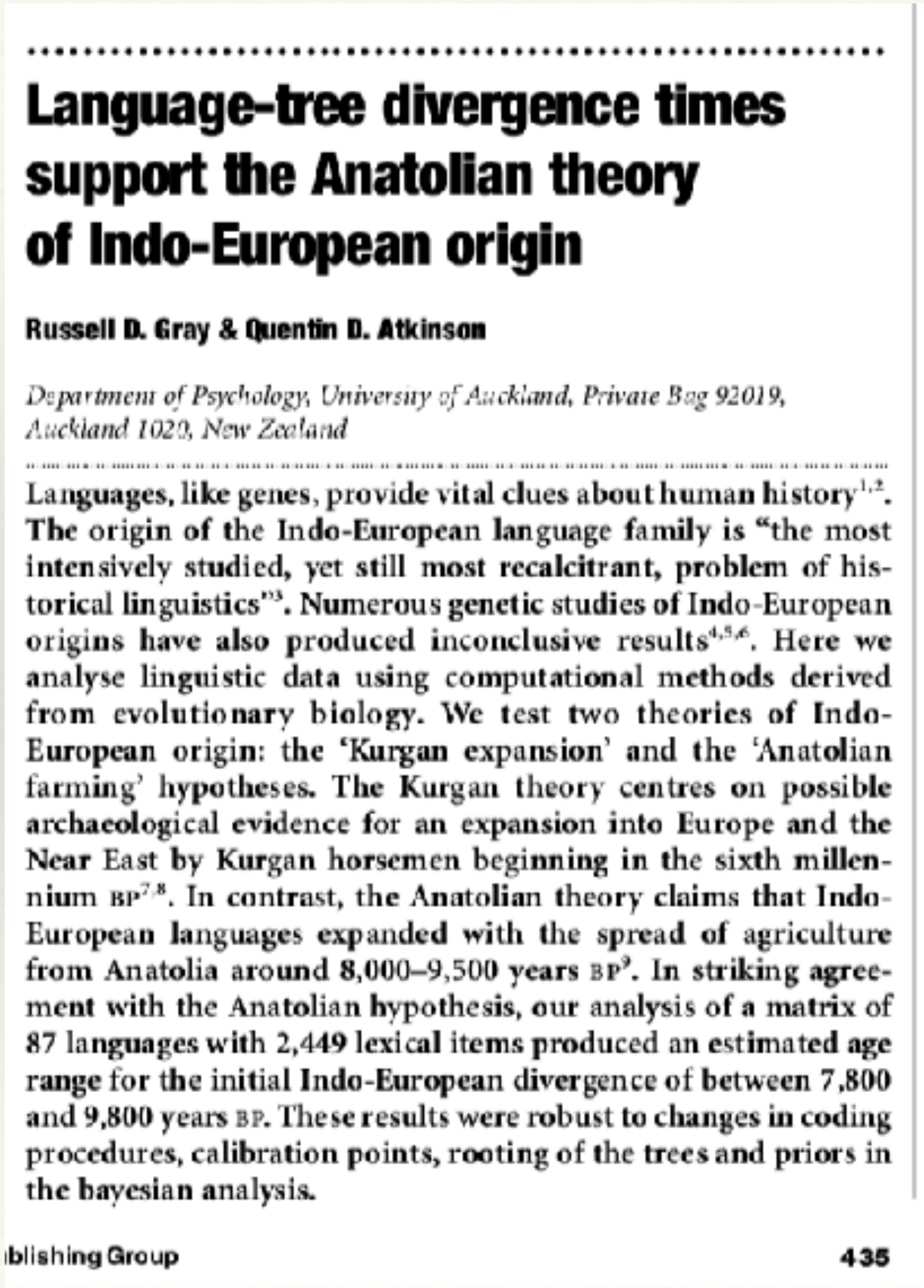
How Phylogenetics Works

Gray & Atkinson on Indo-European

published in *Nature* 2003

Motivating question: which account of I-E origins is better:

- Kurgan hypothesis: 7-8 kya from the Steppes
- Anatolian hypothesis: 8-9.5 kya from Anatolia (with agriculture)



Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

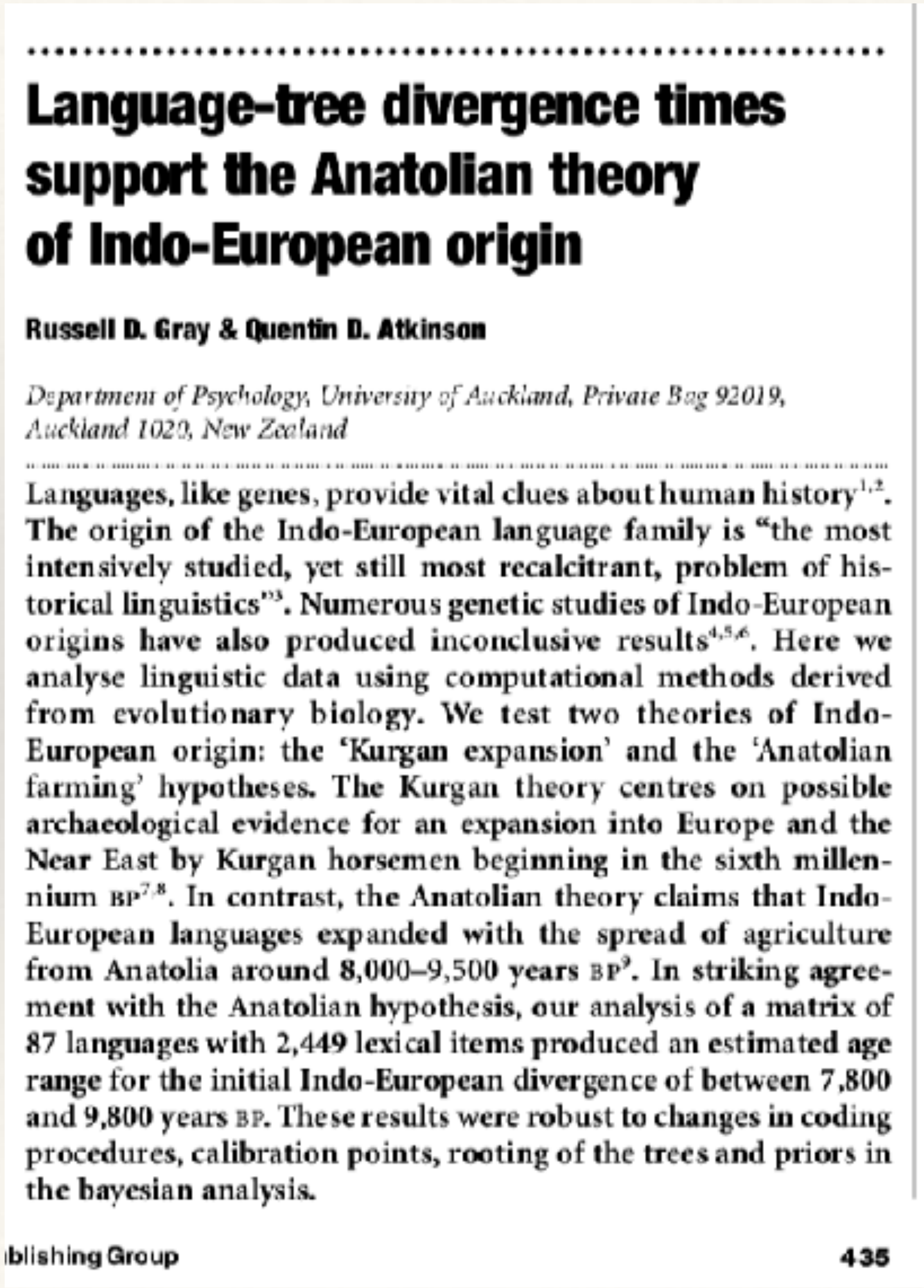
blishing Group 435

Gray & Atkinson on Indo-European

in summary:

they give a date for Proto-Indo-European split of 7800-9800BP (before present)

based on cognate retention/replacement, and dates for historical languages



Language-tree divergence times support the Anatolian theory of Indo-European origin

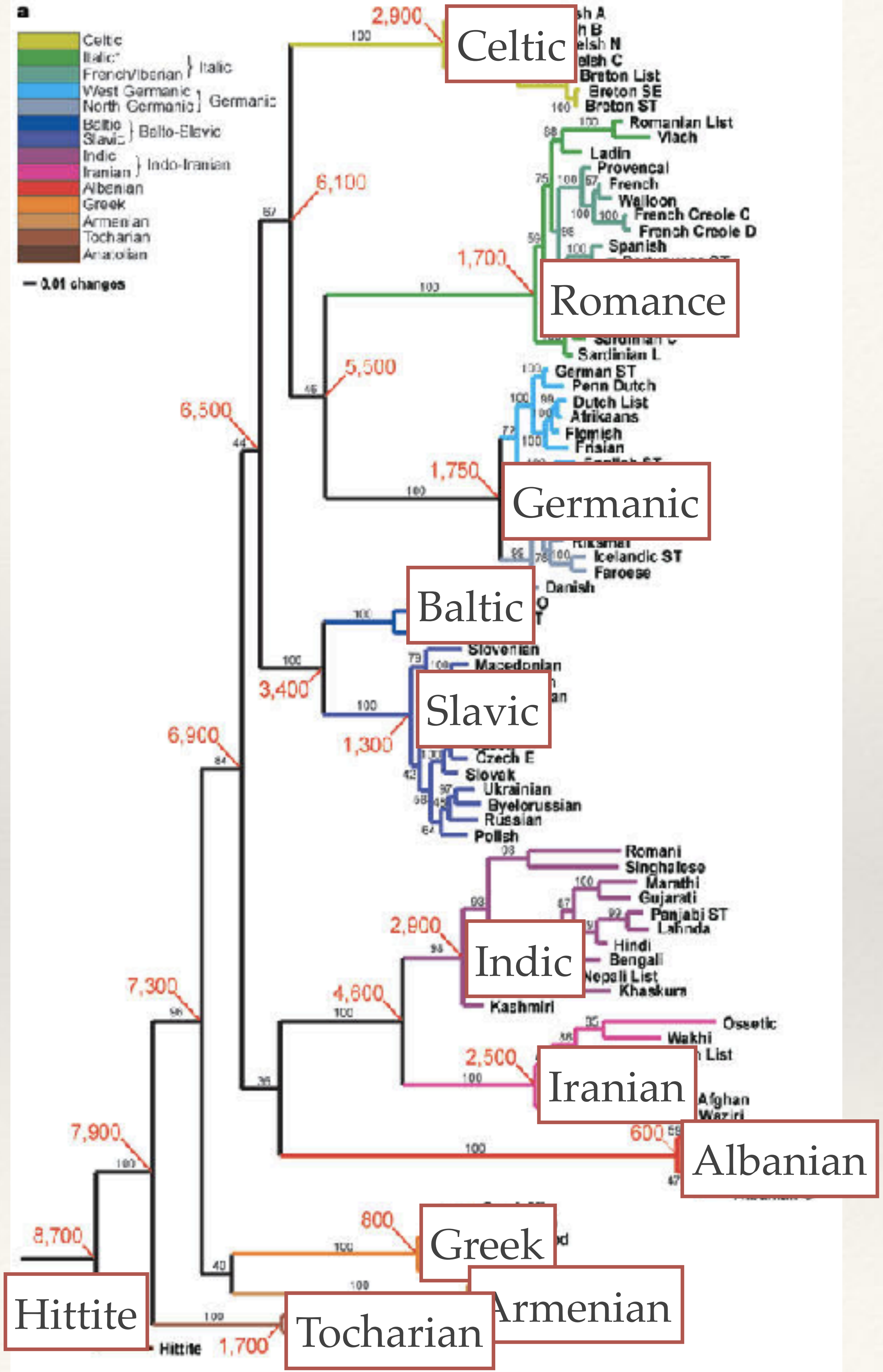
Russell D. Gray & Quentin D. Atkinson

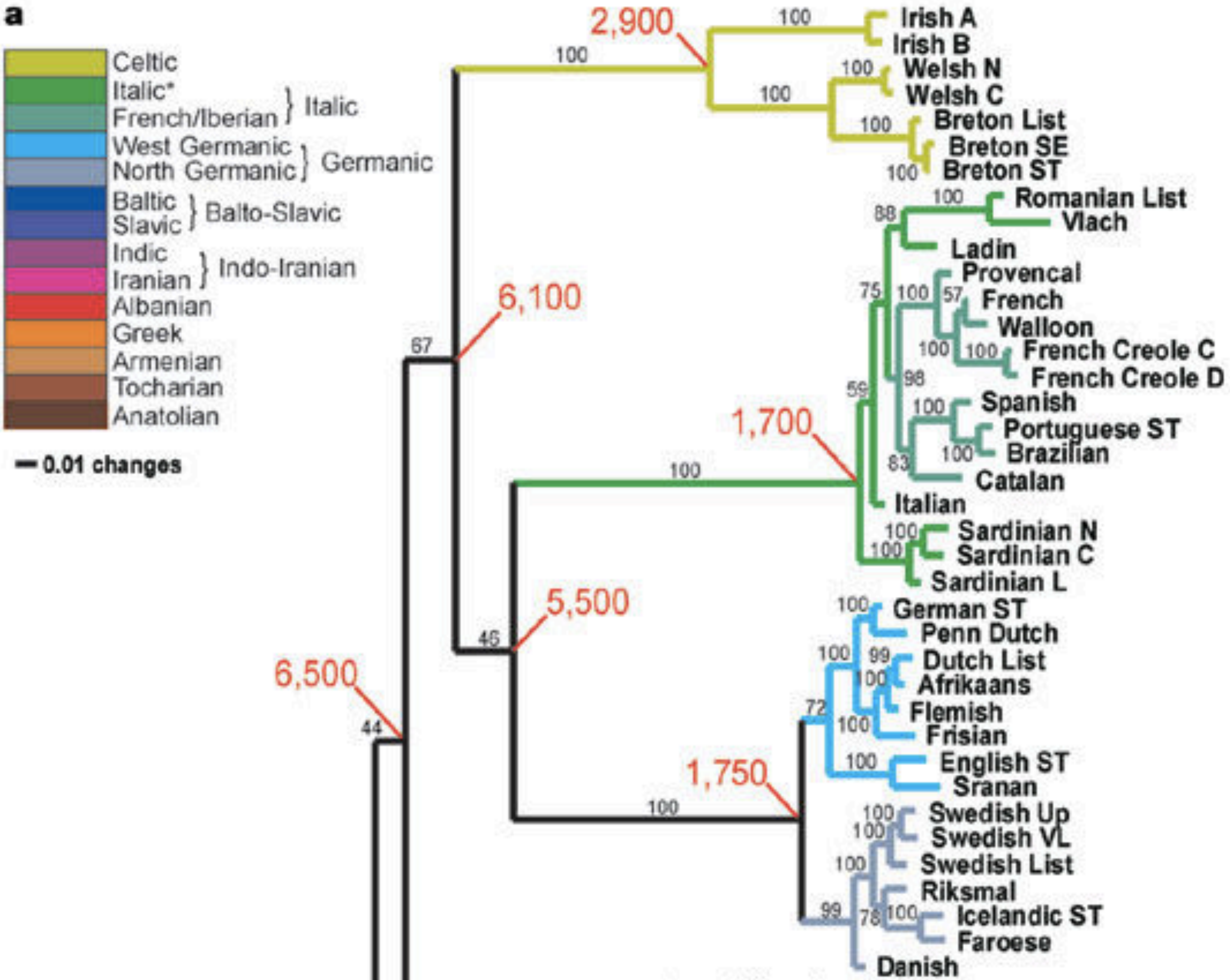
Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

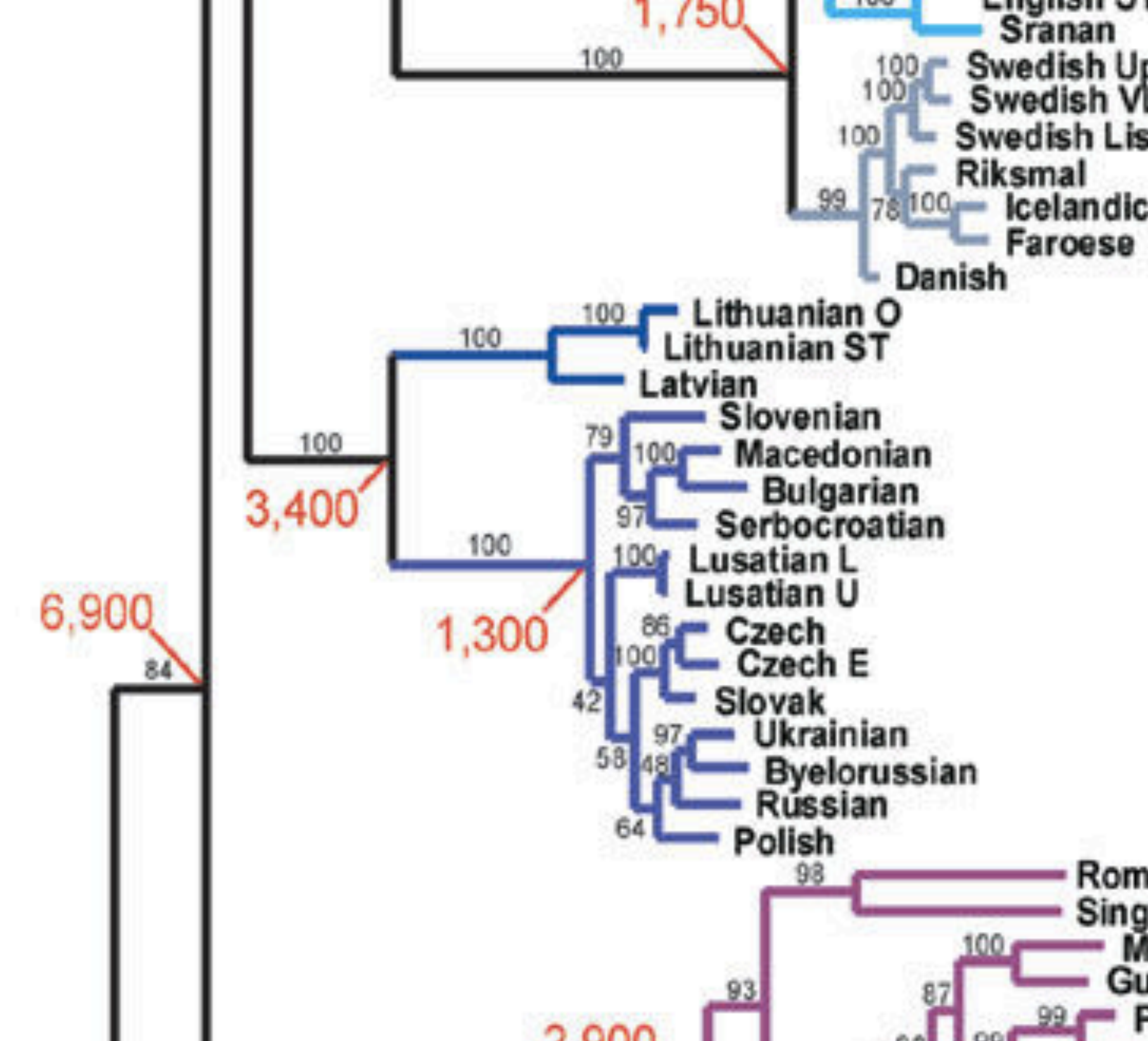
Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

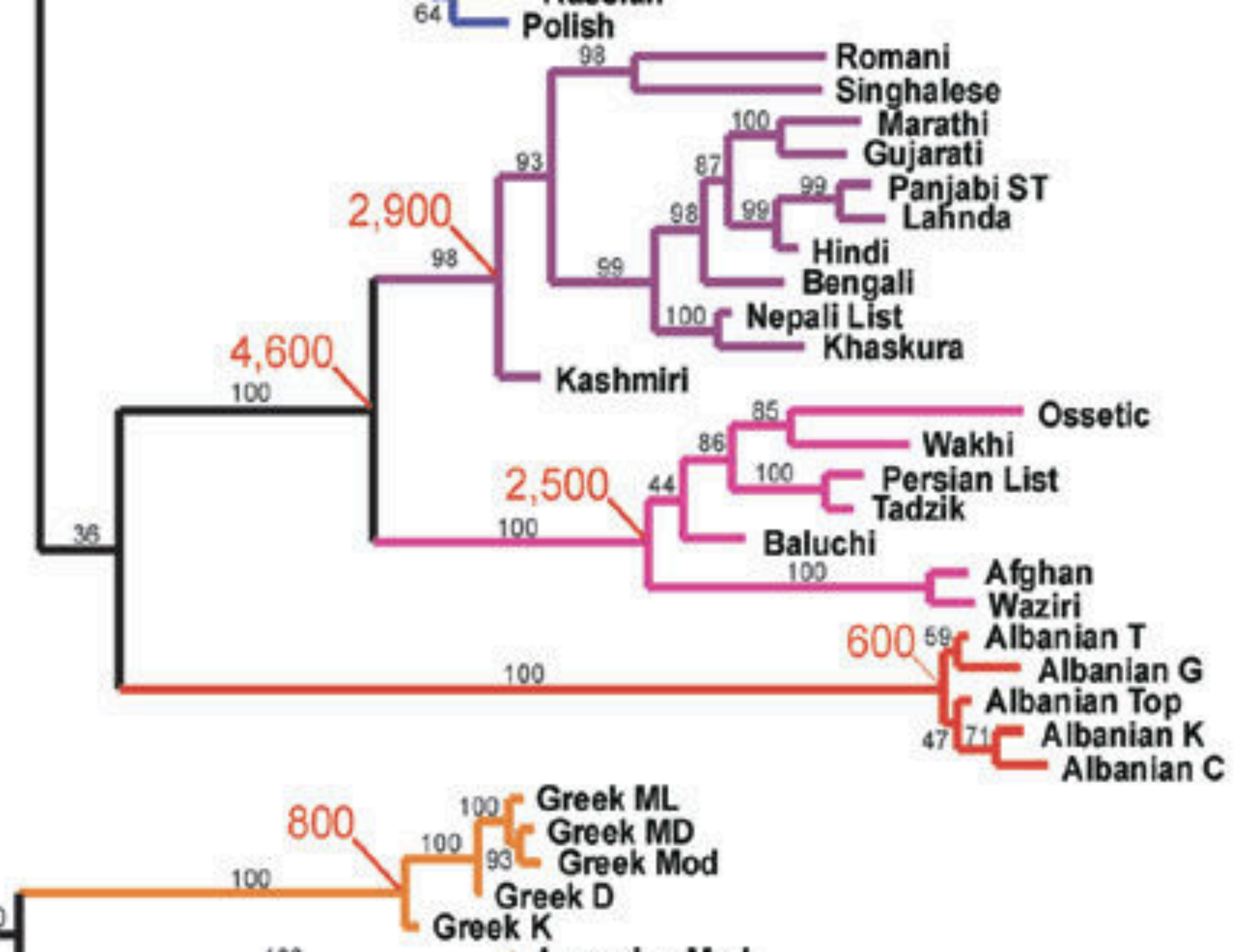
Glottochronology

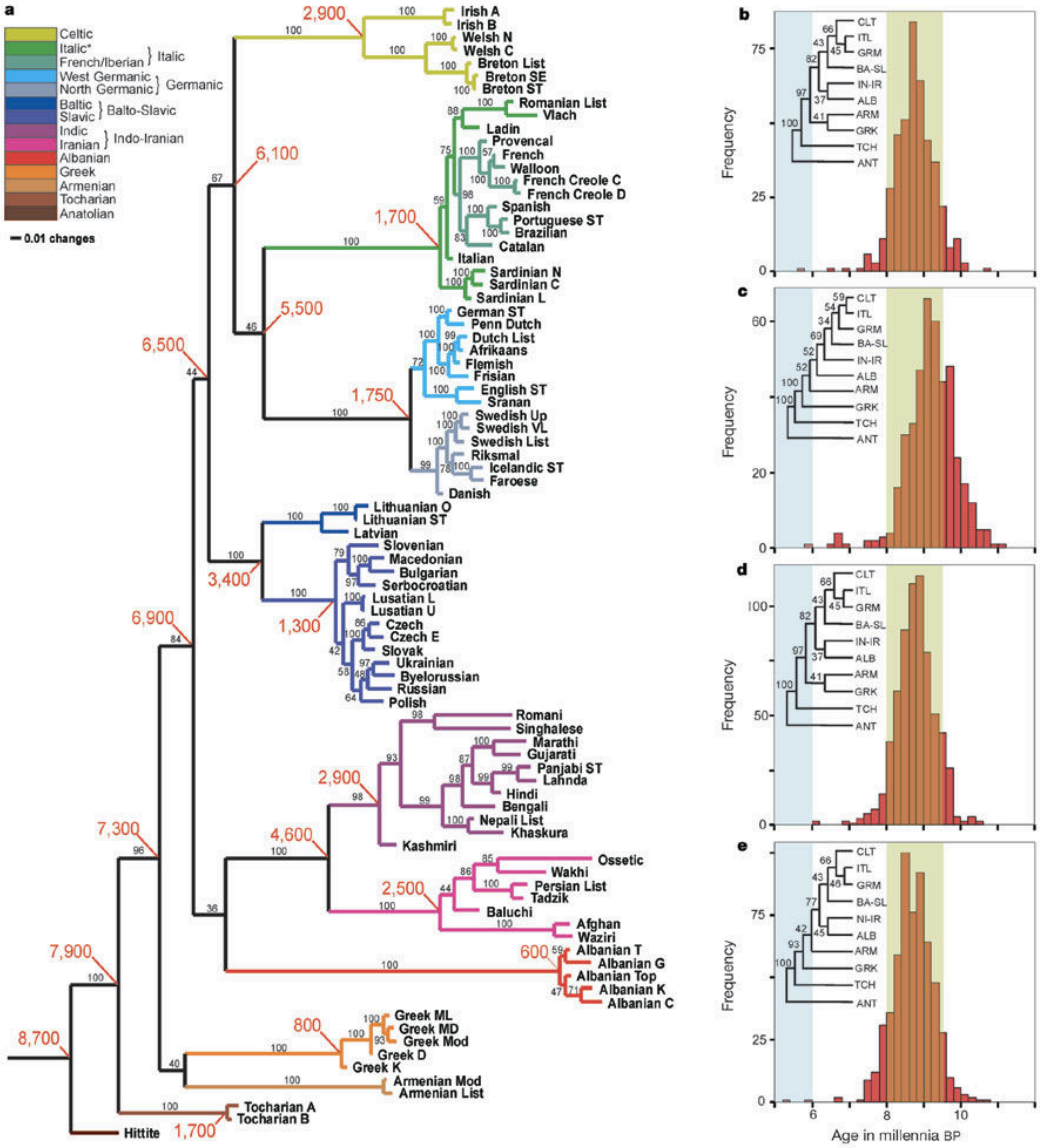
- aware of the failure of glottochronology
- argue that the biological methods are fundamentally different


a
Celtic
Italic
French/Iberian
Italic
West Germanic
North Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
2,900
Celtic
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
6,100
1,700
Romance
Sardinian L
5,500
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
6,500
1,750
Germanic
Icelandic ST
Faroese
Danish
Baltic
Slovenian
Macedonian
3,400
Slavic
1,300
Czech E
Slovak
Ukrainian
Byelorussian
Russian
Polish
6,900
Romani
Singhalese
Marathi
Gujarati
Panjabi ST
Lahnda
2,800
Indic
Hindi
Bengali
Nepali List
Khaskura
Kashmiri
7,300
4,600
Ossetic
Wakhi
2,500
Iranian
Afghan
Waziri
600
Albanian
7,900
800
Greek
Armenian
8,700
Hittite
Tocharian
1,700
Hittite


a
Celtic
Italic*
French/Iberian
Italic
West Germanic
North Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
2,900
6,100
1,700
5,500
6,500
1,750
67
46
44
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish

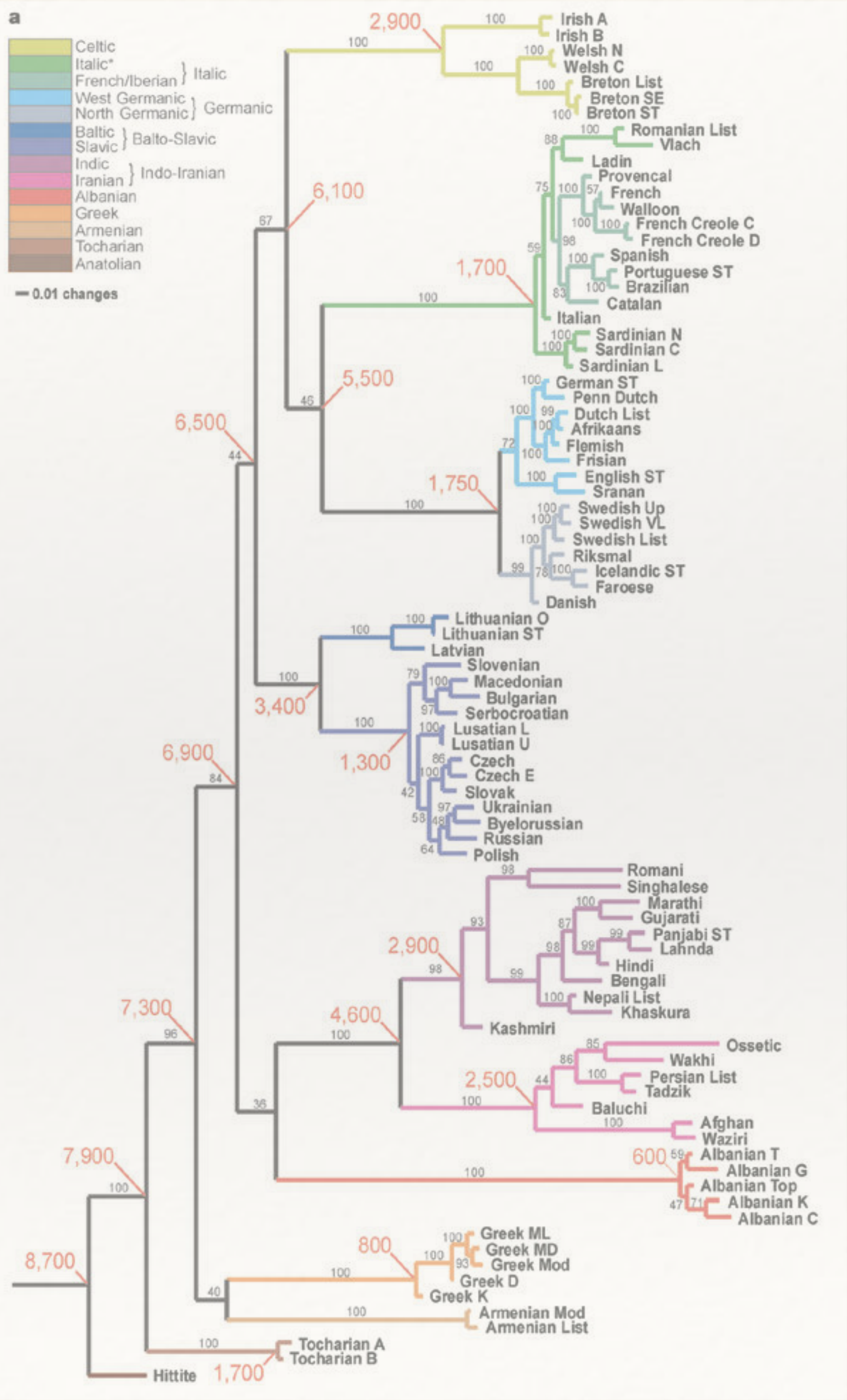

1,750
100
Sranan
100
100
Swedish Up
Swedish Vl
100
Swedish Lis
Riksmal
99
78
100
Icelandic
Faroese
Danish
100
100
Lithuanian O
Lithuanian ST
Latvian
100
3,400
Slovenian
79
100
Macedonian
Bulgarian
97
Serbocroatian
100
100
Lusatian L
Lusatian U
1,300
86
Czech
100
Czech E
Slovak
42
97
Ukrainian
58
48
Byelorussian
Russian
64
Polish
6,900
84
98
Rom
Sing
100
93
87
99
3,900


64
Polish
98
Romani
Singhalese
100
Marathi
Gujarati
93
87
99
Panjabi ST
99
Lahnda
2,900
98
Hindi
98
Bengali
99
100
Nepali List
Khaskura
4,600
Kashmiri
100
85
Ossetic
Wakhi
86
100
Persian List
2,500
44
Tadzik
100
Baluchi
36
100
Afghan
Waziri
600
59
Albanian T
Albanian G
100
Albanian Top
47
71
Albanian K
Albanian C
800
100
Greek ML
Greek MD
100
93
Greek Mod
100
Greek D
Greek K

a
Celtic
Italic*
French/Iberian
Italic
West Germanic
North Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
2,900
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
6,100
1,700
5,500
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
1,750
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish
6,500
Lithuanian O
Lithuanian ST
Latvian
Slovenian
Macedonian
Bulgarian
Serbocroatian
3,400
Lusatian L
Lusatian U
1,300
Czech
Czech E
Slovak
Ukrainian
Byelorussian
Russian
Polish
6,900
Romani
Singhalese
Marathi
Gujarati
Panjabi ST
Lahnda
Hindi
Bengali
2,900
Nepali List
Khaskura
Kashmiri
7,300
4,600
Ossetic
Wakhi
Persian List
Tadzik
2,500
Baluchi
Afghan
Waziri
600
Albanian T
Albanian G
Albanian Top
Albanian K
Albanian C
7,900
8,700
800
Greek ML
Greek MD
Greek Mod
Greek D
Greek K
Armenian Mod
Armenian List
Tocharian A
Tocharian B
Hittite
1,700
b
c
d
e
CLT
ITL
GRM
BA-SL
IN-IR
NI-IR
ALB
ARM
GRK
TCH
ANT
Frequency
Age in millennia BP
6
8
10
0
25
75
20
60
50
100

Understanding the Phylo Method

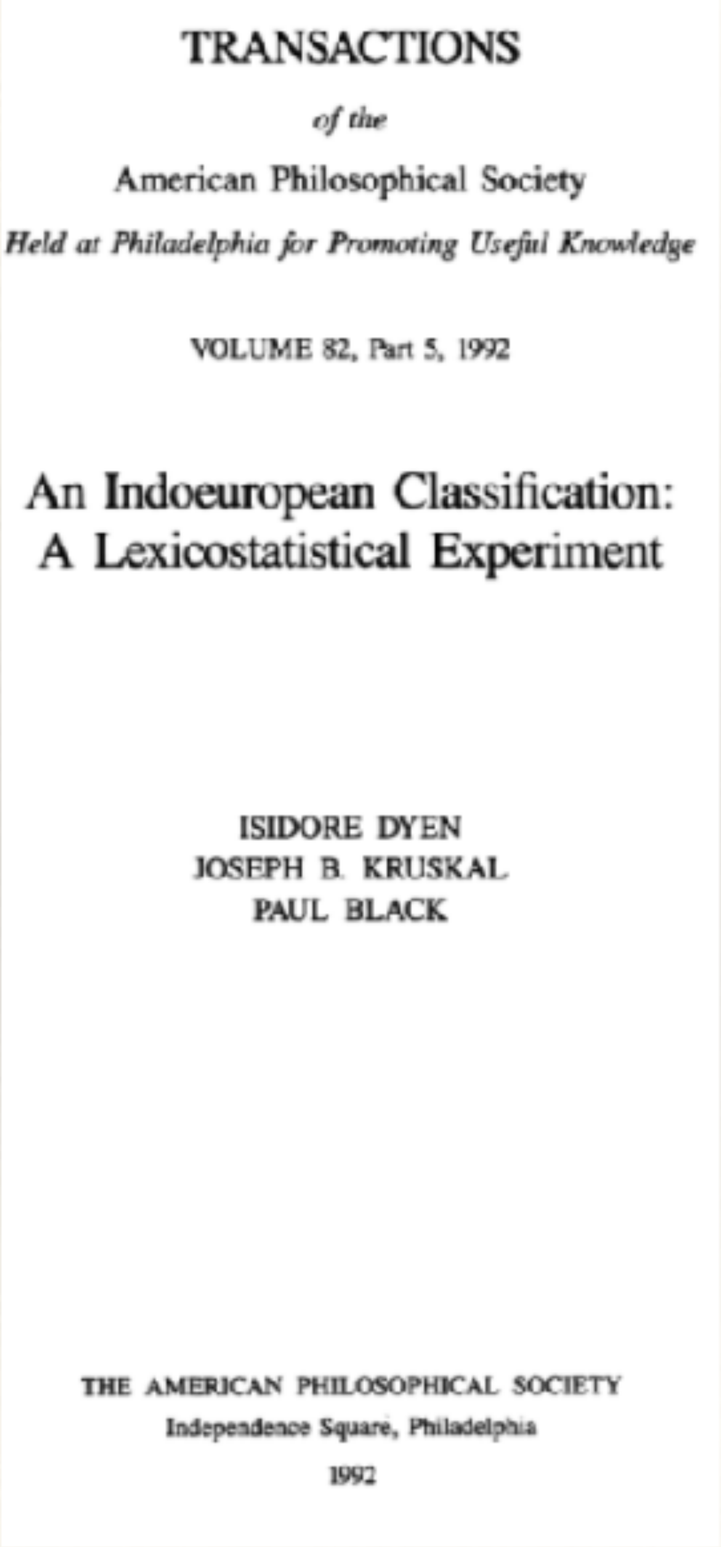
1. Indo-European cognate data
2. Prepare data into *characters*
3. The probability of *unit change*
4. The probability of change on a branch
5. The probability of a tree
6. Sampling the tree space
7. Consensus trees



The Data

Dyen, Kruskal & Black (1992) reports work on lexicostatistics using data collected at Charles Darwin University (Australia) in the 1960s.

Used lexicostatistics to get groupings of subfamilies.



TRANSACTIONS

of the

American Philosophical Society

Held at Philadelphia for Promoting Useful Knowledge

VOLUME 82, Part 5, 1992

An Indoeuropean Classification:
A Lexicostatistical Experiment

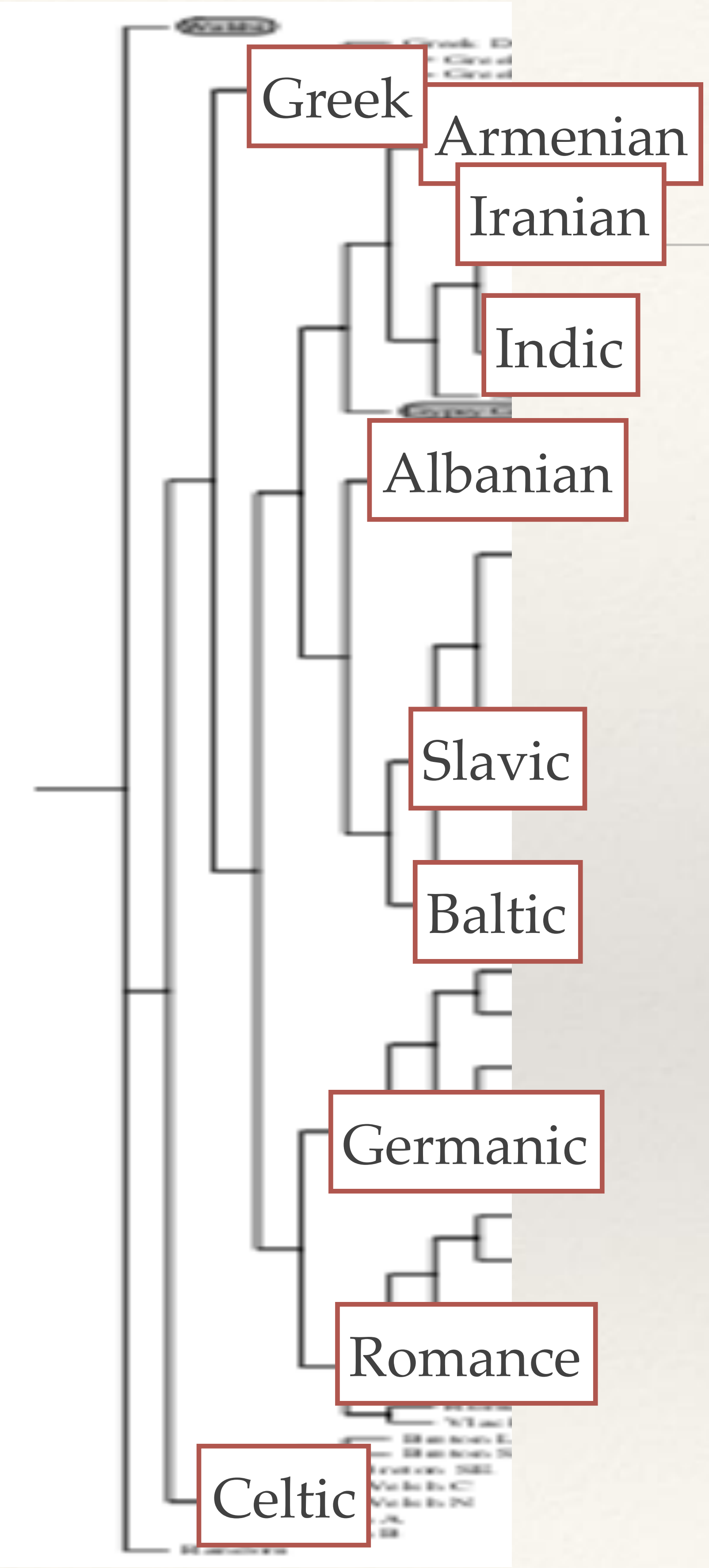
ISIDORE DYEN
JOSEPH B. KRUSKAL
PAUL BLACK

THE AMERICAN PHILOSOPHICAL SOCIETY
Independence Square, Philadelphia
1992

Cognates as Characters

- The Dyen data combines 200 meanings (extended Swadesh list), 87 languages, 1 word per language per meaning
- For each meaning, the authors determined cognate sets across the languages

Signal in the Data



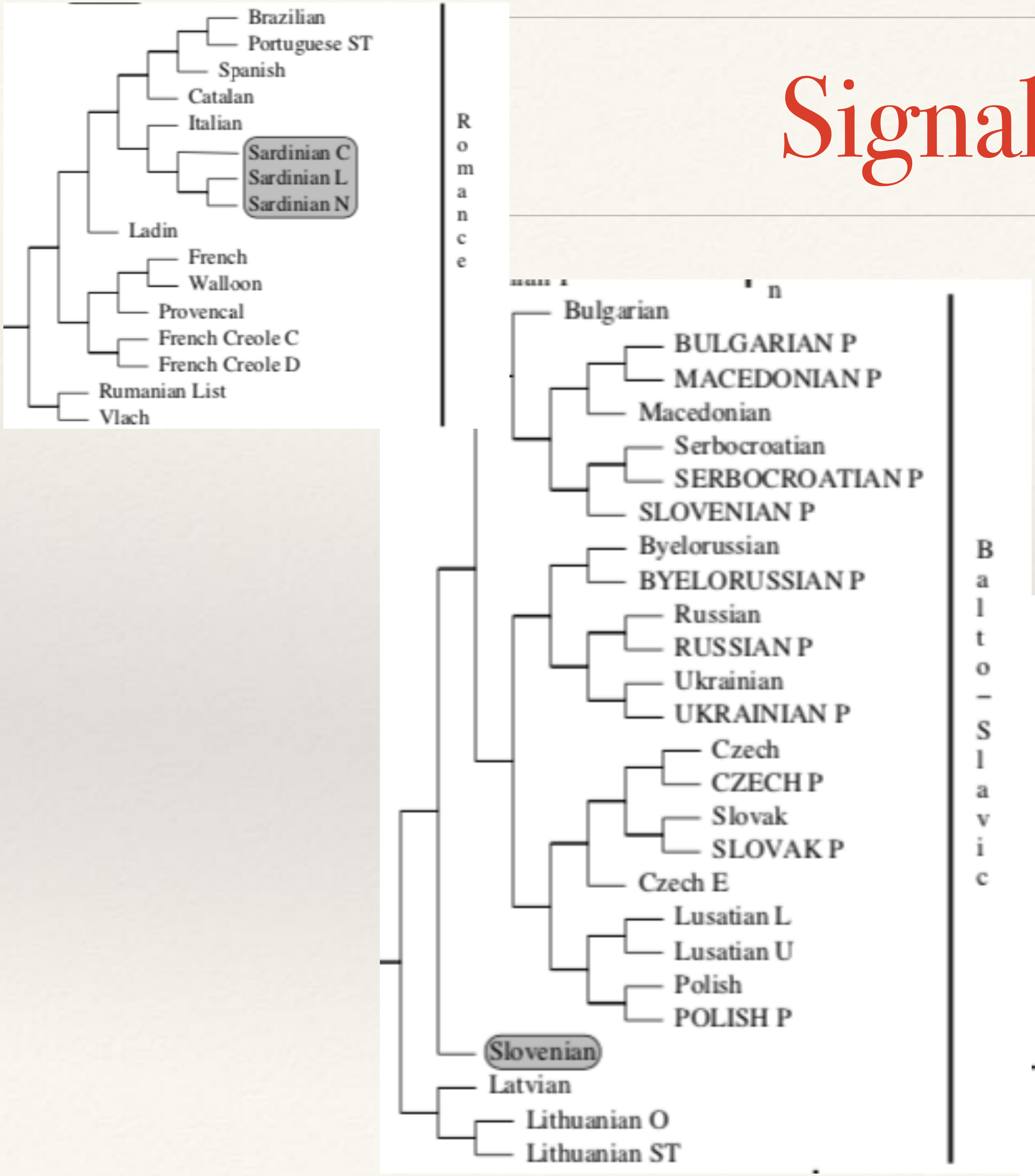
The data collected by Dyen et al. has a strong family tree signal.

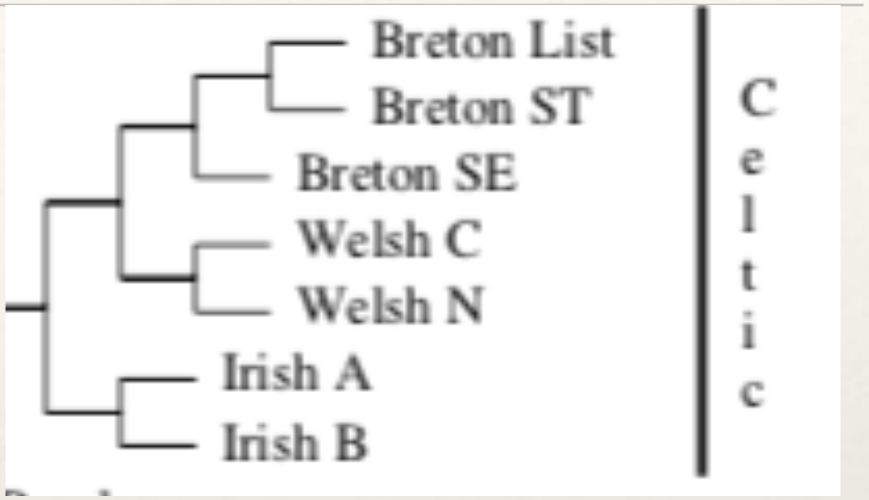
There has been a cottage industry in Computational Linguistics building trees from this data using a wide range of methods.

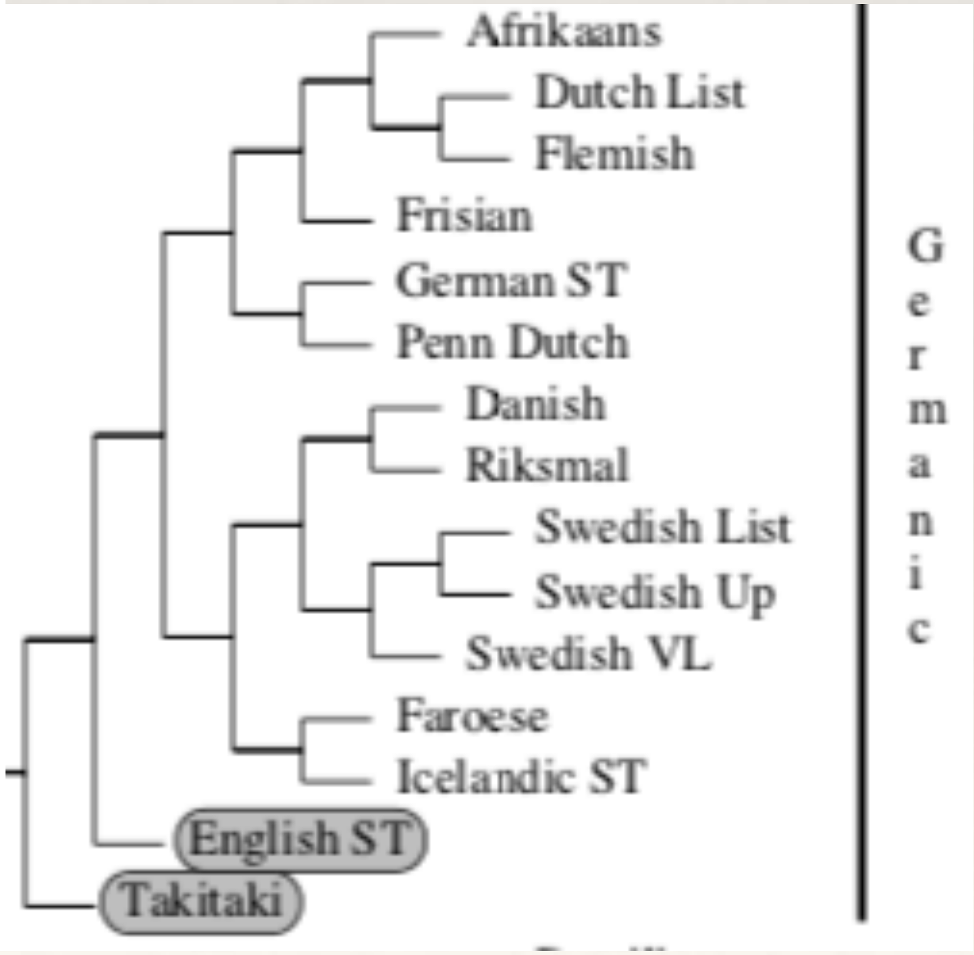
Any reasonable method should be able to get close to the right tree.

E.g. Ellison & Kirby (2006)

Signal in the Data





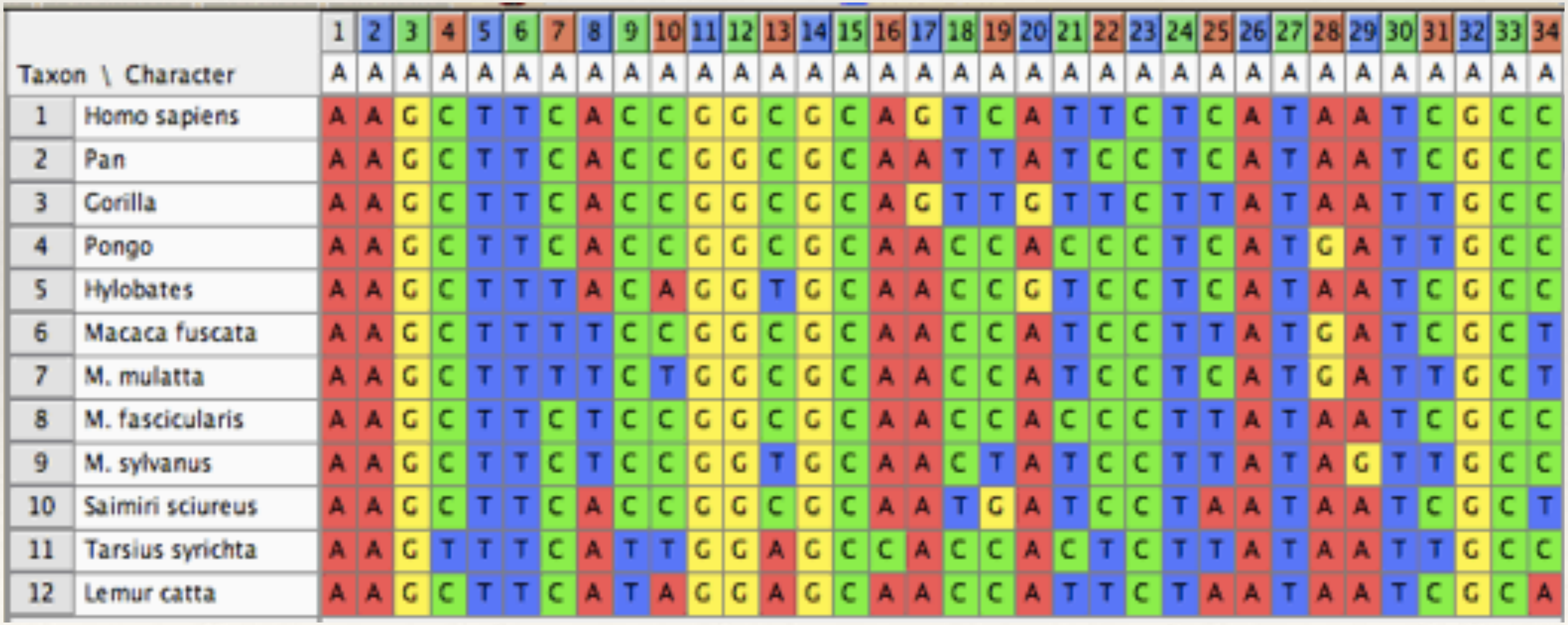


The Data

MEANING	Sanskrit	Latin	Greek	English
I	aham	ego	egṓ	i
THREE	trí	trēs	treîs	three
MOTHER	mā́tṛ	māter	mḗtēr	mother
FIRE	agní	ignis	pûr	fire
SNAKE	sarpá	serpens	óphis	snake
TOOTH	dánta	dens	odoús	tooth
TO FLOAT	plávate	fluctuāre	néō	float
FULL	pūrṇá	plēnus	plḗrēs	full
...	...	...	...	...

Cognates as Characters

DNA uses 4 *bases* to encode its information: A C G T
DNA sequences from different species can be aligned
with the corresponding points (in columns) being called *characters*



Taxon \ Character		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
2	Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
3	Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
4	Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
5	Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
6	Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
7	M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
8	M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
9	M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
10	Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
11	Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
12	Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

Cognates as Characters

CHARACTER		
	MEANING	COGNATE SET
C1	I	i
C2	THREE	three
C3	MOTHER	mother
C4	FIRE	ignis
C5	FIRE	fire
C6	SNAKE	serpens
C7	SNAKE	óphis
C8	SNAKE	snake
..	..	..

	C1	C2	C3	C4	C5	C6	C7	C8	..
Sanskrit	1	1	1	1	0	1	0	0	..
Latin	1	1	1	1	0	1	0	0	..
Greek	1	1	1	0	1	0	1	0	..
English	1	1	1	0	1	0	0	1	..

Cognates as Characters

- The Dyen data uses 200 meanings, 87 languages, 1 word per language per meaning
- Each language has its ‘lexicon` defined by a vector of 2449 binary numbers, each one indicating that a particular ***meaning-cognateset*** combination is present or absent
- PROBLEM: as a language can only has one word per meaning, characters are not independent: FIRE/*fire* only when not FIRE/*ignis* (from language.log)

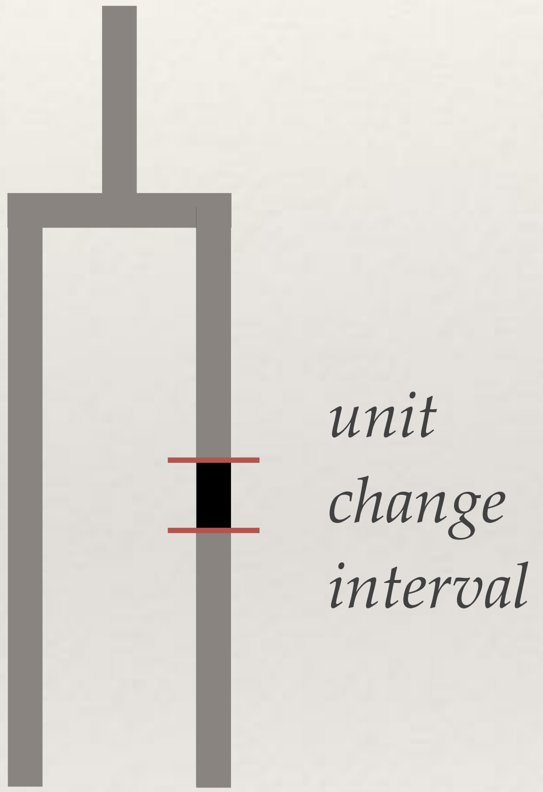
Normal Transmission

What happens in a *unit change interval* - is now a question of character change.

In the course of a unit interval what is the probability of:

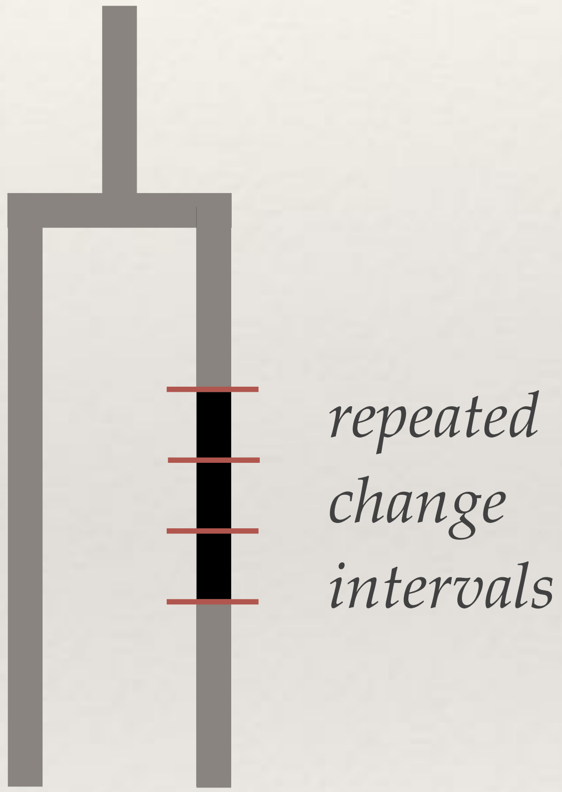
a) a character changing 1 ↦ 0

b) a character changing 0 ↦ 1



Normal Transmission

- over multiple change intervals, the same probabilities of transition apply
- if we know the values at the start and end of a stretch of change intervals, we can work out how likely that change was (given our model)
- assuming the characters vary independently



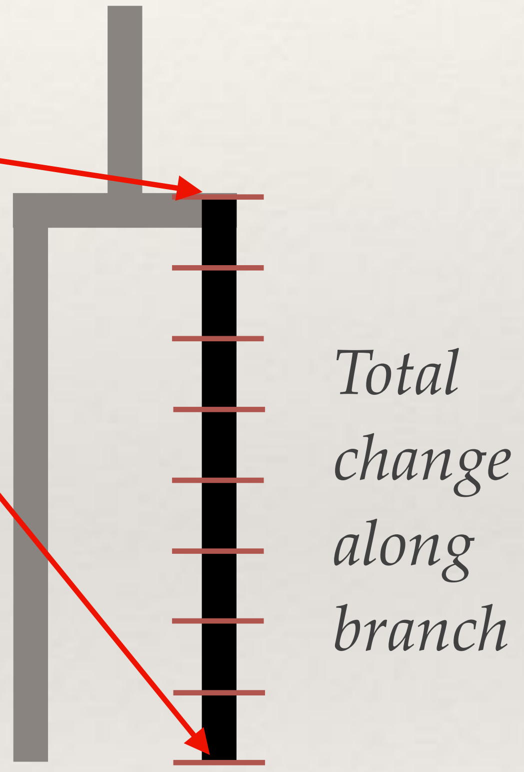
MCMC SIMULATION

Normal Transmission

if we know

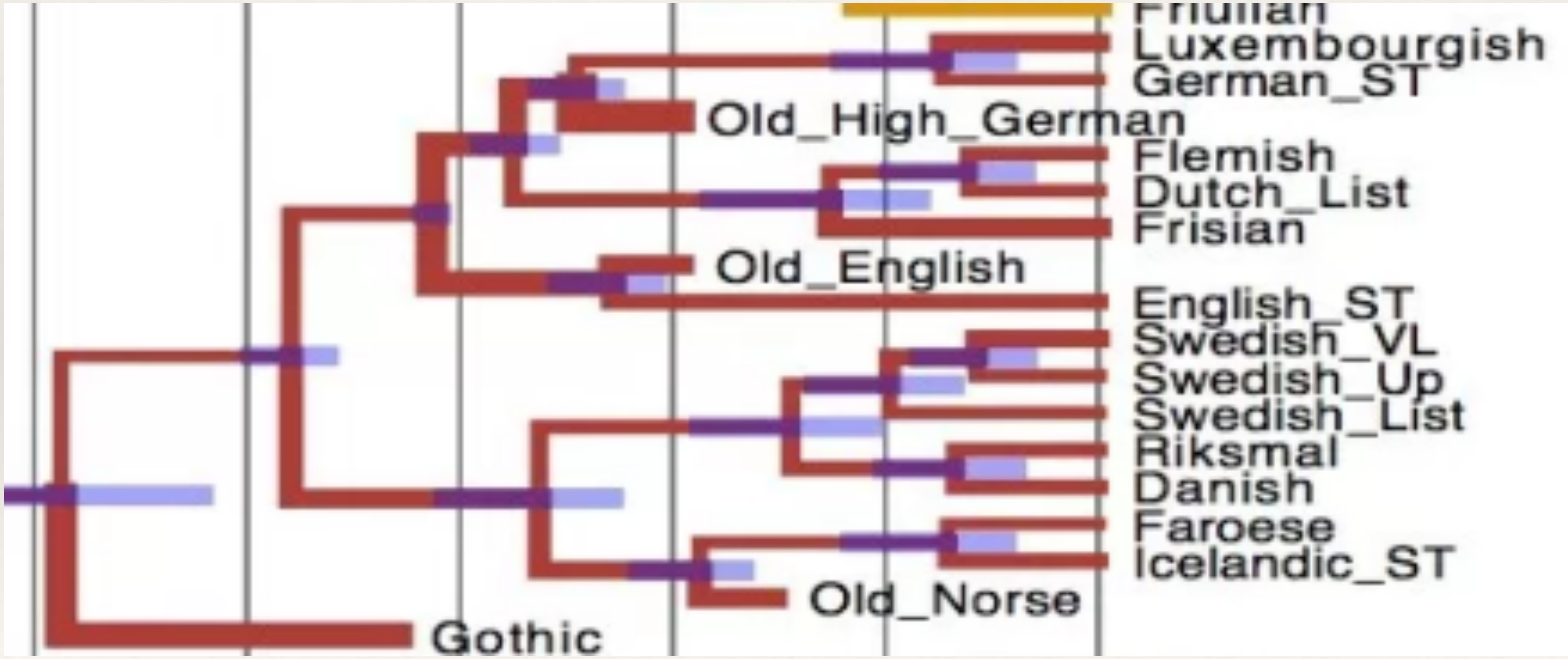
- character vector at parent
- character vector at child
- number of change intervals per year 0.01
- time-length along branch 800 (8 intervals)
- transition probabilities

we can work out the probability of the change seen in that branch



Normal Transmission

Bouckaert & al (2012) **Mapping the Origins and Expansion of the Indo-European Language Family** *Science*



Each Possible Phylogeny Specifies:

Probabilities of transition: $1 \mapsto 0$ and $0 \mapsto 1$

At each branch-point or leaf:

- a character vector

On each branch

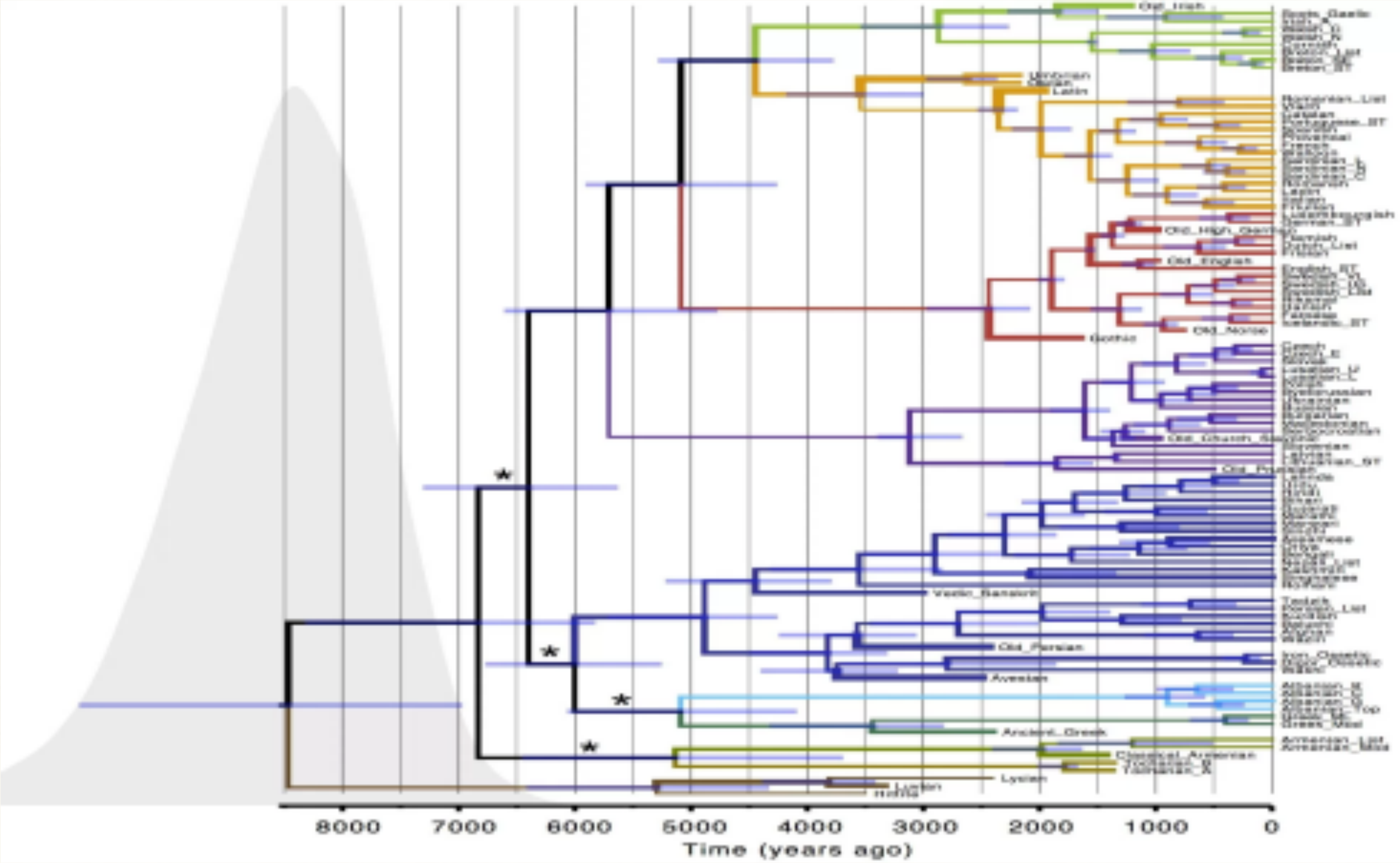
- a change rate, and
- a duration

Work out probability of each branch

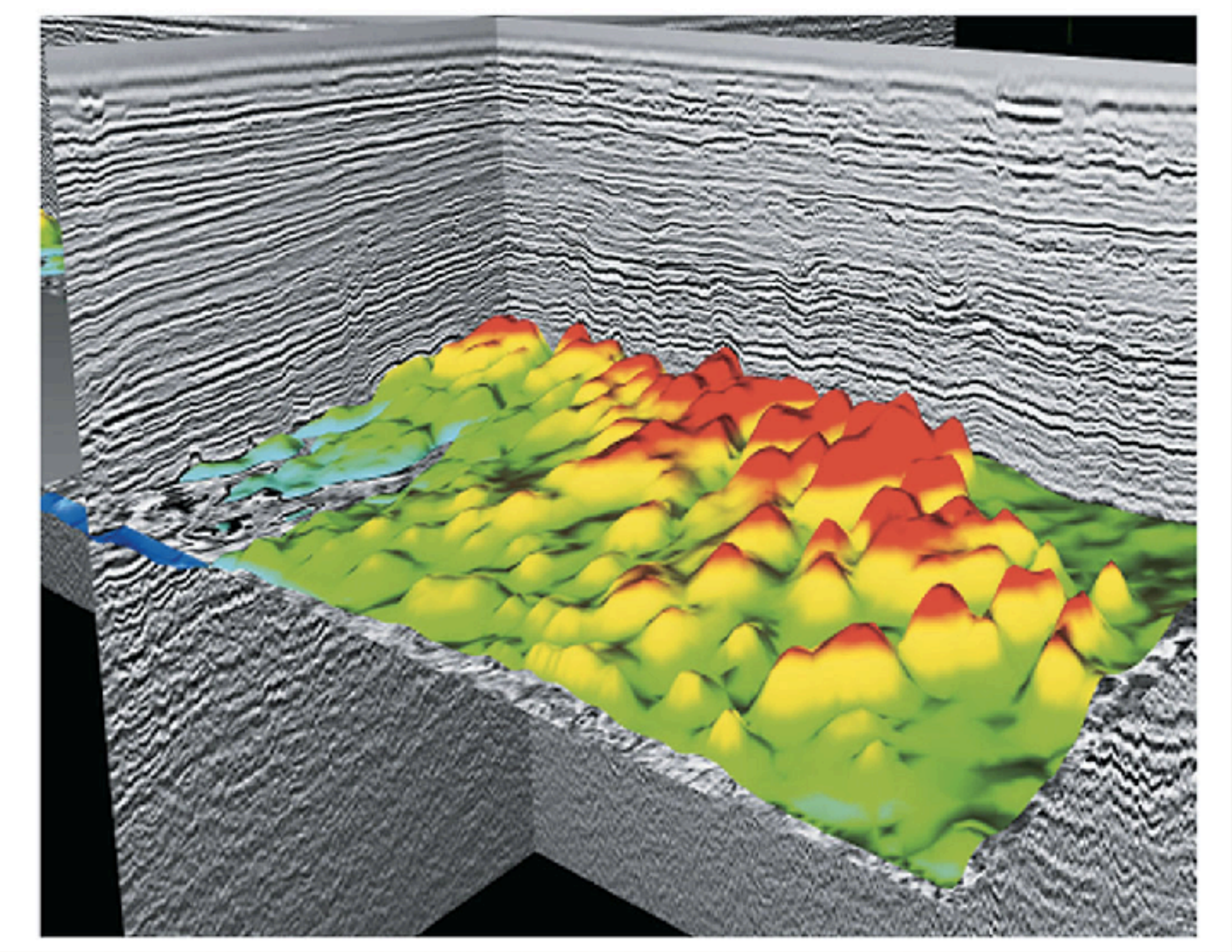
Product these for probability on total tree

Normal Transmission

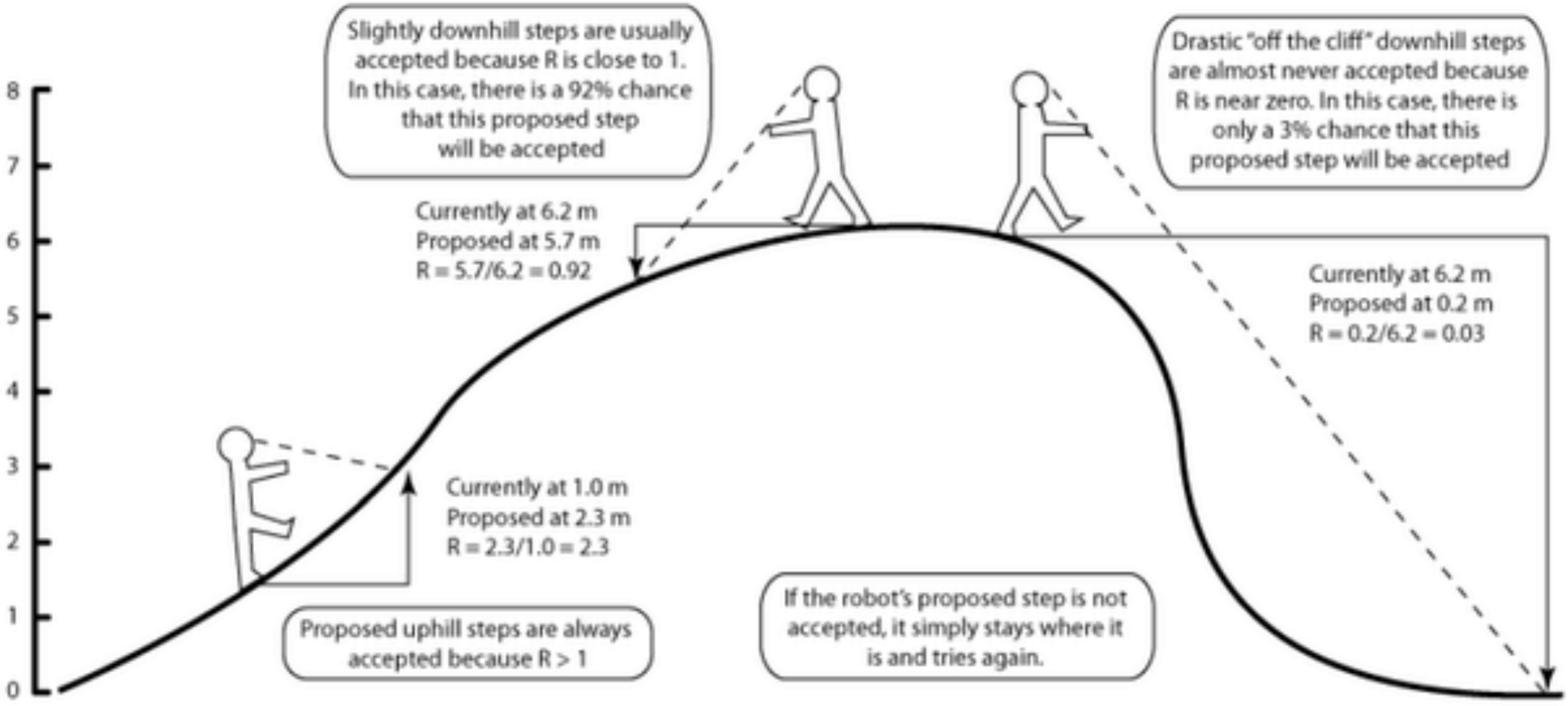
Bouckaert & al (2012) **Mapping the Origins and Expansion of the Indo-European Language Family** *Science*



A Distribution on Trees



Finding Good Trees



Slightly downhill steps are usually accepted because R is close to 1. In this case, there is a 92% chance that this proposed step will be accepted

Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. In this case, there is only a 3% chance that this proposed step will be accepted

Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 = 0.92

Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03

Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3

Proposed uphill steps are always accepted because R > 1

If the robot's proposed step is not accepted, it simply stays where it is and tries again.

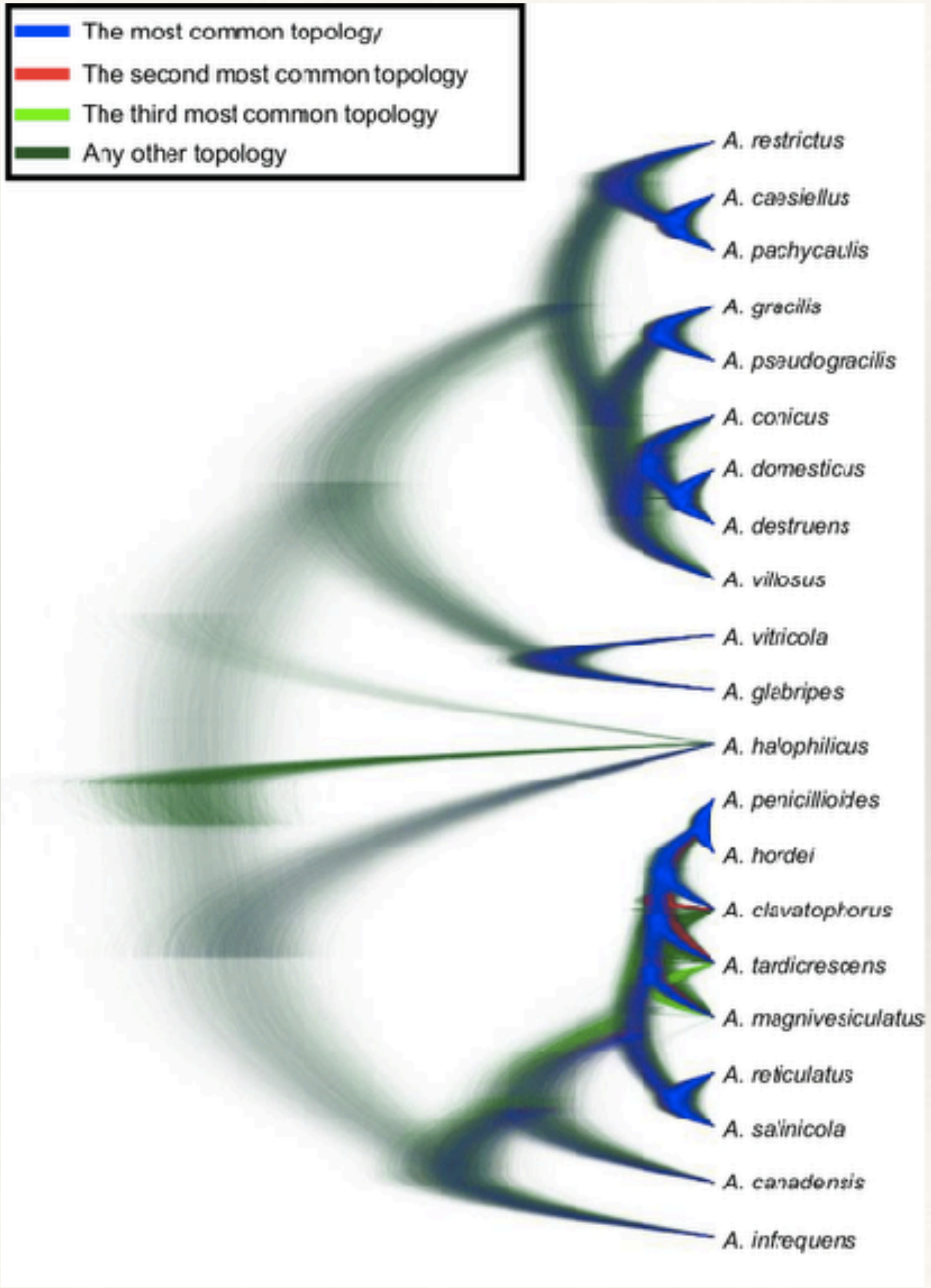
Illustration of MCMC method process (Lewis, 2011)

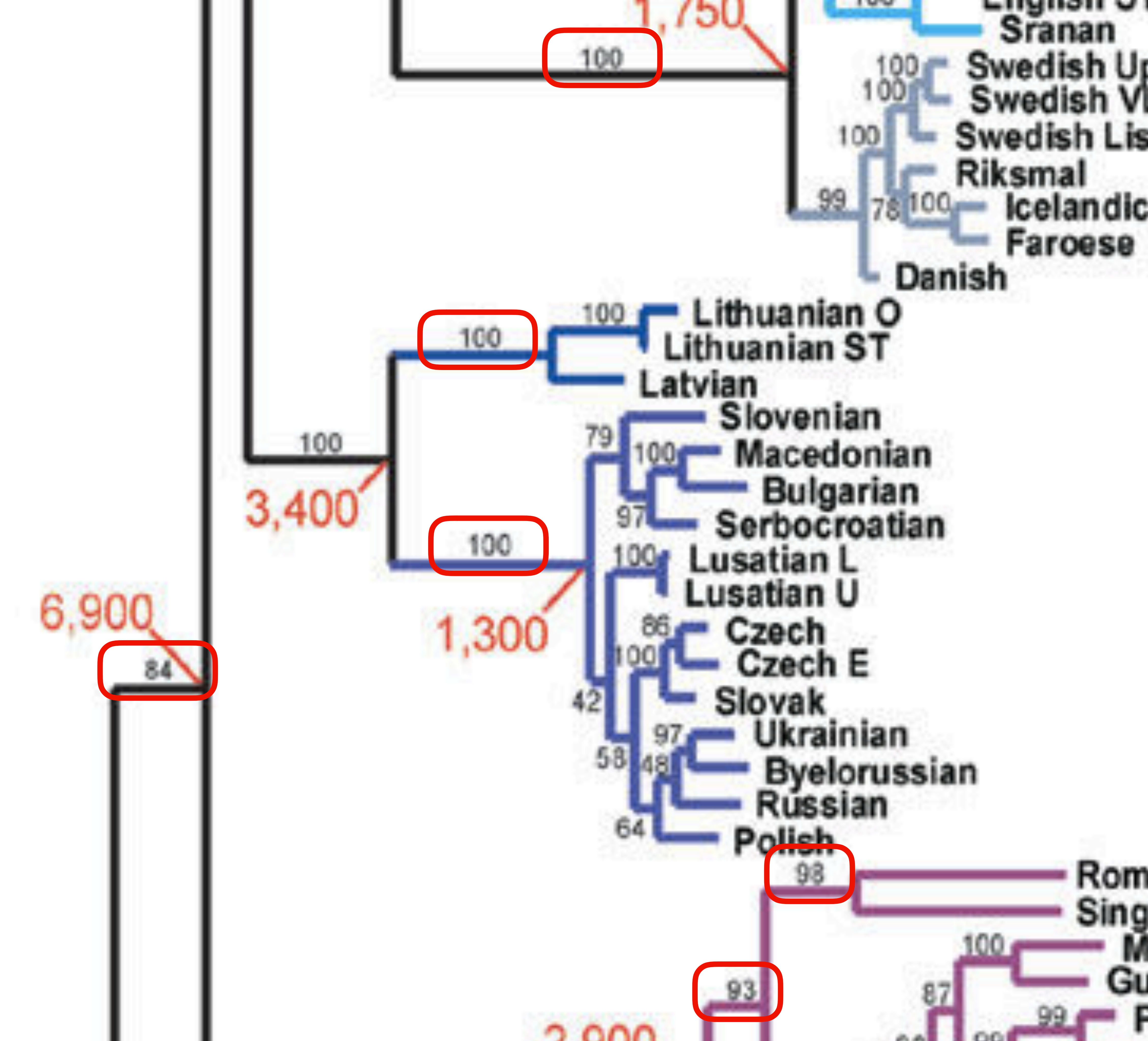
Building a Consensus Tree

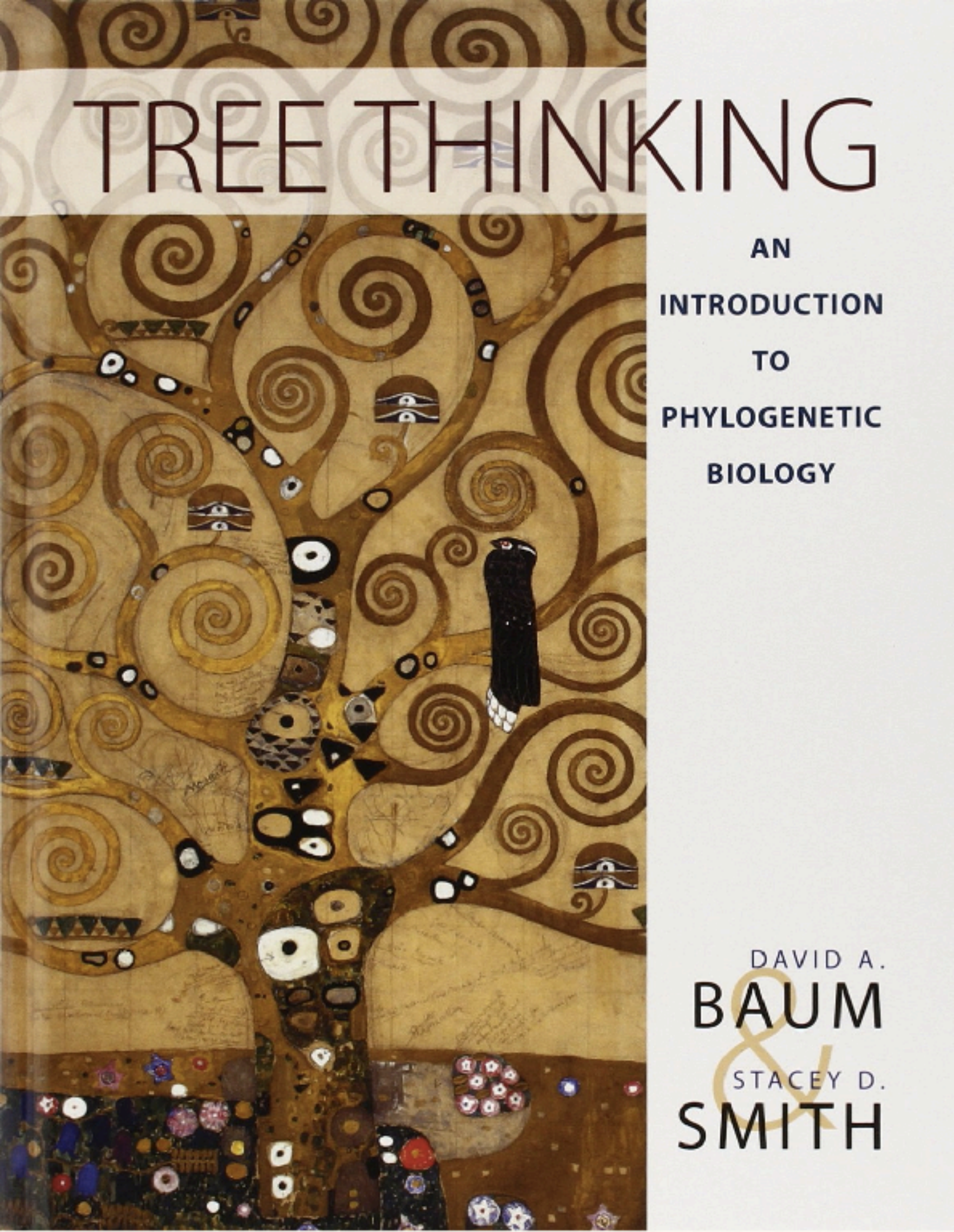
Collect 1000 relatively probable trees

Create a consensus tree: retain subgroups that they agree on

Building a Consensus Tree




1,750
100
Sranan
Swedish Up
100
Swedish Vl
100
Swedish Lis
100
Riksmal
99
78
100
Icelandic
Faroese
Danish
100
Lithuanian O
100
Lithuanian ST
Latvian
100
3,400
Slovenian
79
100
Macedonian
Bulgarian
97
Serbocroatian
100
100
Lusatian L
Lusatian U
6,900
84
1,300
86
Czech
100
Czech E
42
Slovak
97
Ukrainian
58
48
Byelorussian
Russian
64
Polish
98
Rom
Sing
100
M
Gu
93
87
99
2,900



Recommended introduction to how biological phylogenetics builds trees from data.

Language phylo people use the same techniques.

Software

Tutorials helpful for BEAUti and BEAST etc.

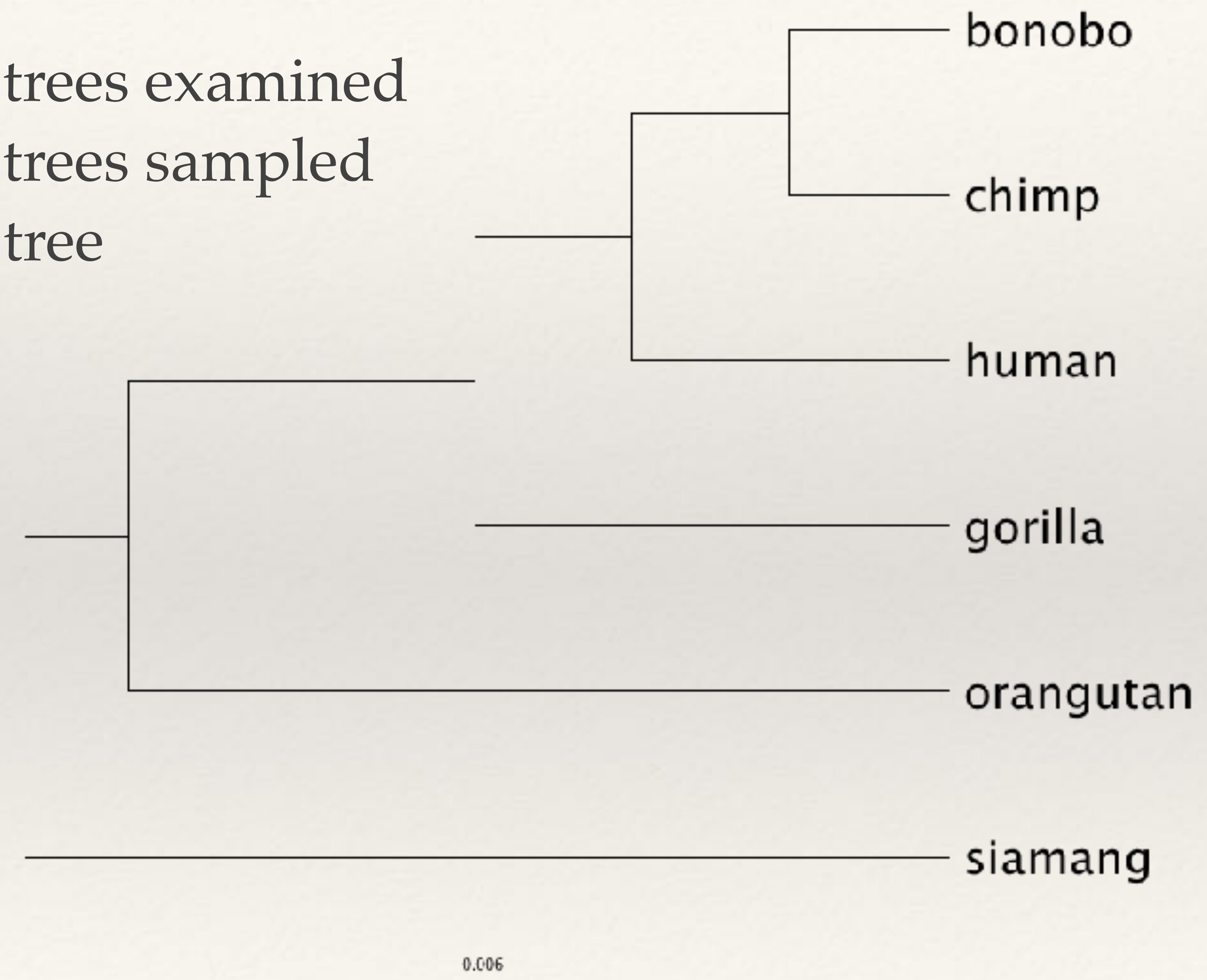
- **BEAUti** - sets up the parameters for doing a phylogenetic analysis
- **BEAST** - does the sampling of the tree distribution
- **TreeAnnotator** - pulls together information from this sample
- **FigTree** - view trees in various formats

Apes

10 000 000 trees examined

10 000 trees sampled

consensus tree



Limitations

Loss of Information

- phylo is informationally shallow: each word is no more than its cognacy affiliation
- detailed linguistic investigation is important for understanding not just change in the language, but society at large

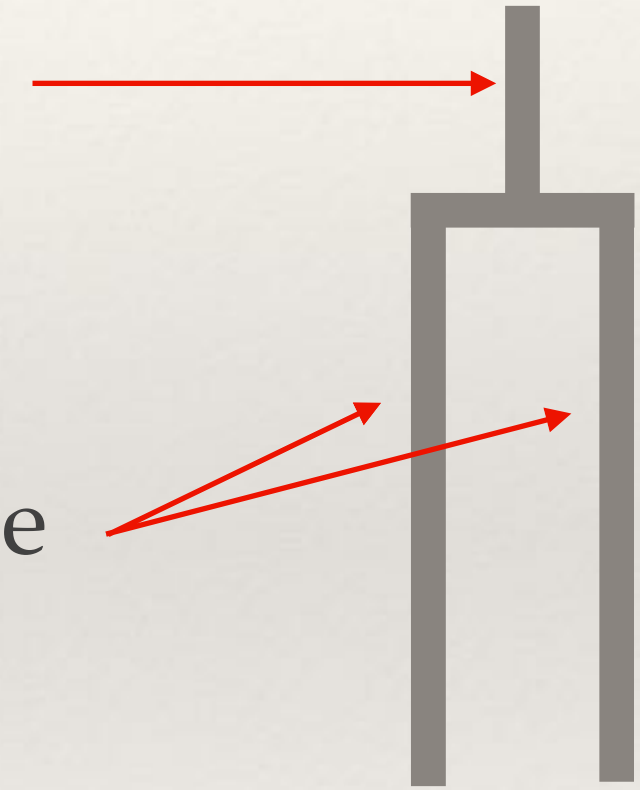
Calibration

- *dates out* rely on *dates in*
    - dated data is needed (e.g. dated written languages)
    - the more you allow rate change, the more calibration data you need

Too Much Consistency

- same rate of change for all lexemes
  - addressed (at least partially) by *covarion* model of lexical togetherness
- same rate of change for all branches
  - addressed by allowing rates to vary but encouraging similar rates in similar languages

Dendrophilia

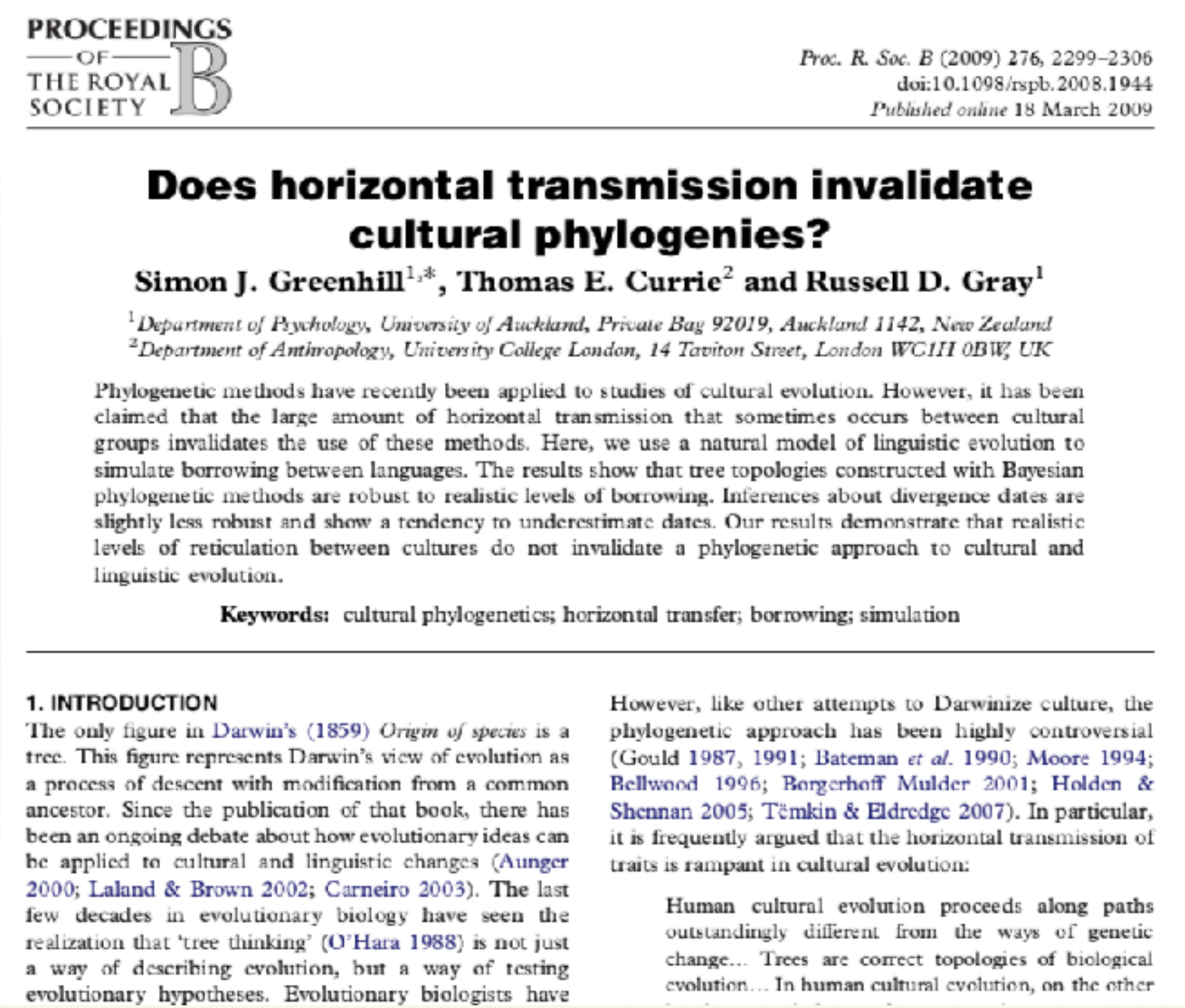


Before branching: 0% independence
between communities:
total togetherness

After branching: 100% independence
of communities:
total isolation

Communities in contact lie in the middle
diffusion, borrowing, anti-doppel effects

Dendrophilia



PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2009) 276, 2299–2306
doi:10.1098/rspb.2008.1944

Does horizontal transmission invalidate cultural phylogenies?

Simon J. Greenhill[1,*], **Thomas E. Currie**[2] **and Russell D. Gray**[1]

[1]*Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand*
[2]*Department of Anthropology, University College London, 14 Taviton Street, London WC1H 0BW, UK*

Phylogenetic methods have recently been applied to studies of cultural evolution. However, it has been claimed that the large amount of horizontal transmission that sometimes occurs between cultural groups invalidates the use of these methods. Here, we use a natural model of linguistic evolution to simulate borrowing between languages. The results show that tree topologies constructed with Bayesian phylogenetic methods are robust to realistic levels of borrowing. Inferences about divergence dates are slightly less robust and show a tendency to underestimate dates. Our results demonstrate that realistic levels of reticulation between cultures do not invalidate a phylogenetic approach to cultural and linguistic evolution.

Keywords: cultural phylogenetics; horizontal transfer; borrowing; simulation

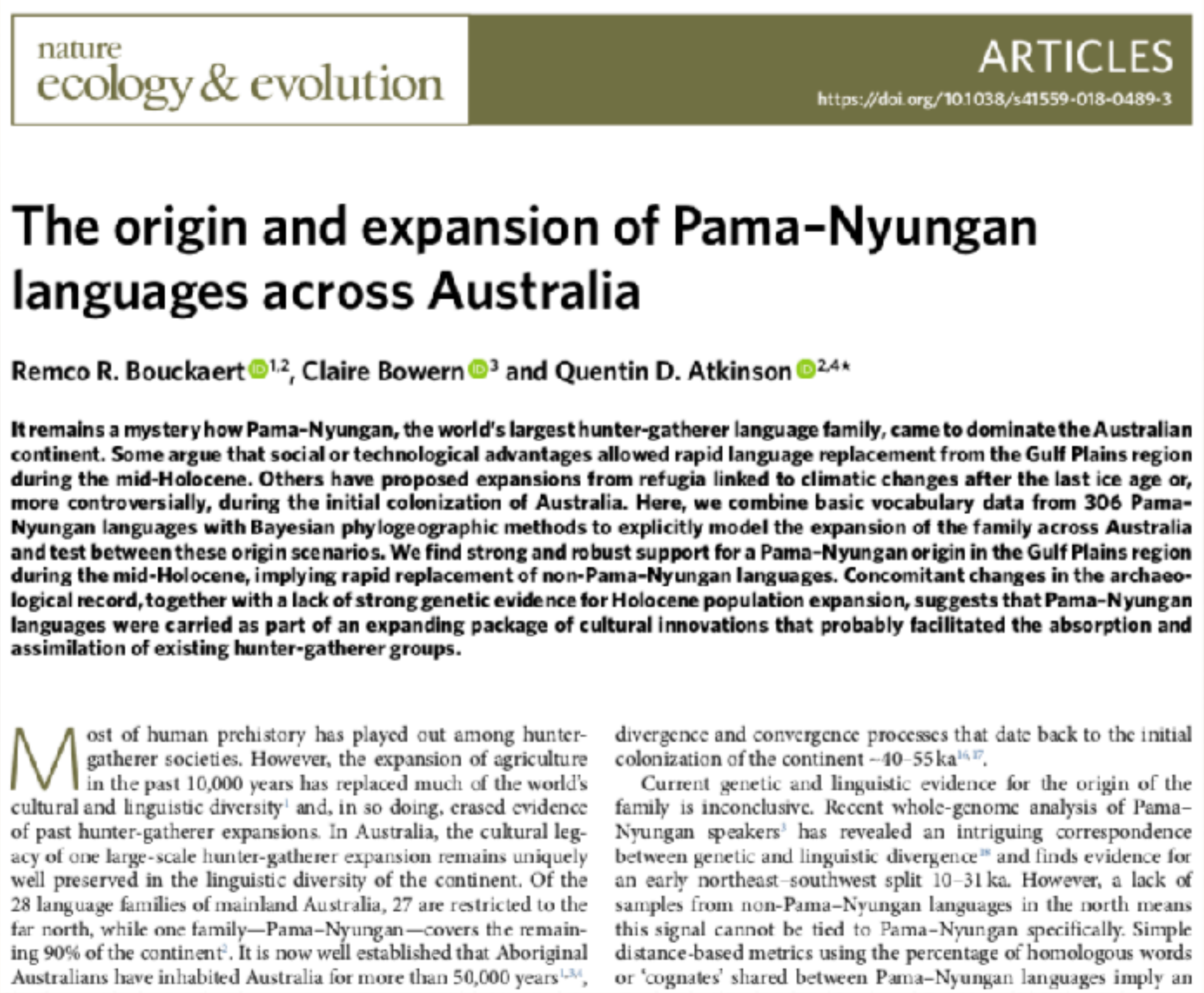
1. INTRODUCTION

The only figure in Darwin's (1859) *Origin of species* is a tree. This figure represents Darwin's view of evolution as a process of descent with modification from a common ancestor. Since the publication of that book, there has been an ongoing debate about how evolutionary ideas can be applied to cultural and linguistic changes (Aunger 2000; Laland & Brown 2002; Carneiro 2003). The last few decades in evolutionary biology have seen the realization that 'tree thinking' (O'Hara 1988) is not just a way of describing evolution, but a way of testing evolutionary hypotheses. Evolutionary biologists have

However, like other attempts to Darwinize culture, the phylogenetic approach has been highly controversial (Gould 1987, 1991; Bateman *et al.* 1990; Moore 1994; Bellwood 1996; Borgerhoff Mulder 2001; Holden & Shennan 2005; Tëmkin & Eldredge 2007). In particular, it is frequently argued that the horizontal transmission of traits is rampant in cultural evolution:

> Human cultural evolution proceeds along paths outstandingly different from the ways of genetic change… Trees are correct topologies of biological evolution… In human cultural evolution, on the other

Dendrophilia



nature
ecology & evolution

ARTICLES
https://doi.org/10.1038/s41559-018-0489-3

The origin and expansion of Pama–Nyungan languages across Australia

Remco R. Bouckaert[1,2], Claire Bowern[3] and Quentin D. Atkinson[2,4*]

It remains a mystery how Pama–Nyungan, the world's largest hunter-gatherer language family, came to dominate the Australian continent. Some argue that social or technological advantages allowed rapid language replacement from the Gulf Plains region during the mid-Holocene. Others have proposed expansions from refugia linked to climatic changes after the last ice age or, more controversially, during the initial colonization of Australia. Here, we combine basic vocabulary data from 306 Pama-Nyungan languages with Bayesian phylogeographic methods to explicitly model the expansion of the family across Australia and test between these origin scenarios. We find strong and robust support for a Pama–Nyungan origin in the Gulf Plains region during the mid-Holocene, implying rapid replacement of non-Pama–Nyungan languages. Concomitant changes in the archaeological record, together with a lack of strong genetic evidence for Holocene population expansion, suggests that Pama–Nyungan languages were carried as part of an expanding package of cultural innovations that probably facilitated the absorption and assimilation of existing hunter-gatherer groups.

Most of human prehistory has played out among hunter-gatherer societies. However, the expansion of agriculture in the past 10,000 years has replaced much of the world's cultural and linguistic diversity[1] and, in so doing, erased evidence of past hunter-gatherer expansions. In Australia, the cultural legacy of one large-scale hunter-gatherer expansion remains uniquely well preserved in the linguistic diversity of the continent. Of the 28 language families of mainland Australia, 27 are restricted to the far north, while one family—Pama–Nyungan—covers the remaining 90% of the continent[2]. It is now well established that Aboriginal Australians have inhabited Australia for more than 50,000 years[1,3,4],

divergence and convergence processes that date back to the initial colonization of the continent ~40–55 ka[16,17].

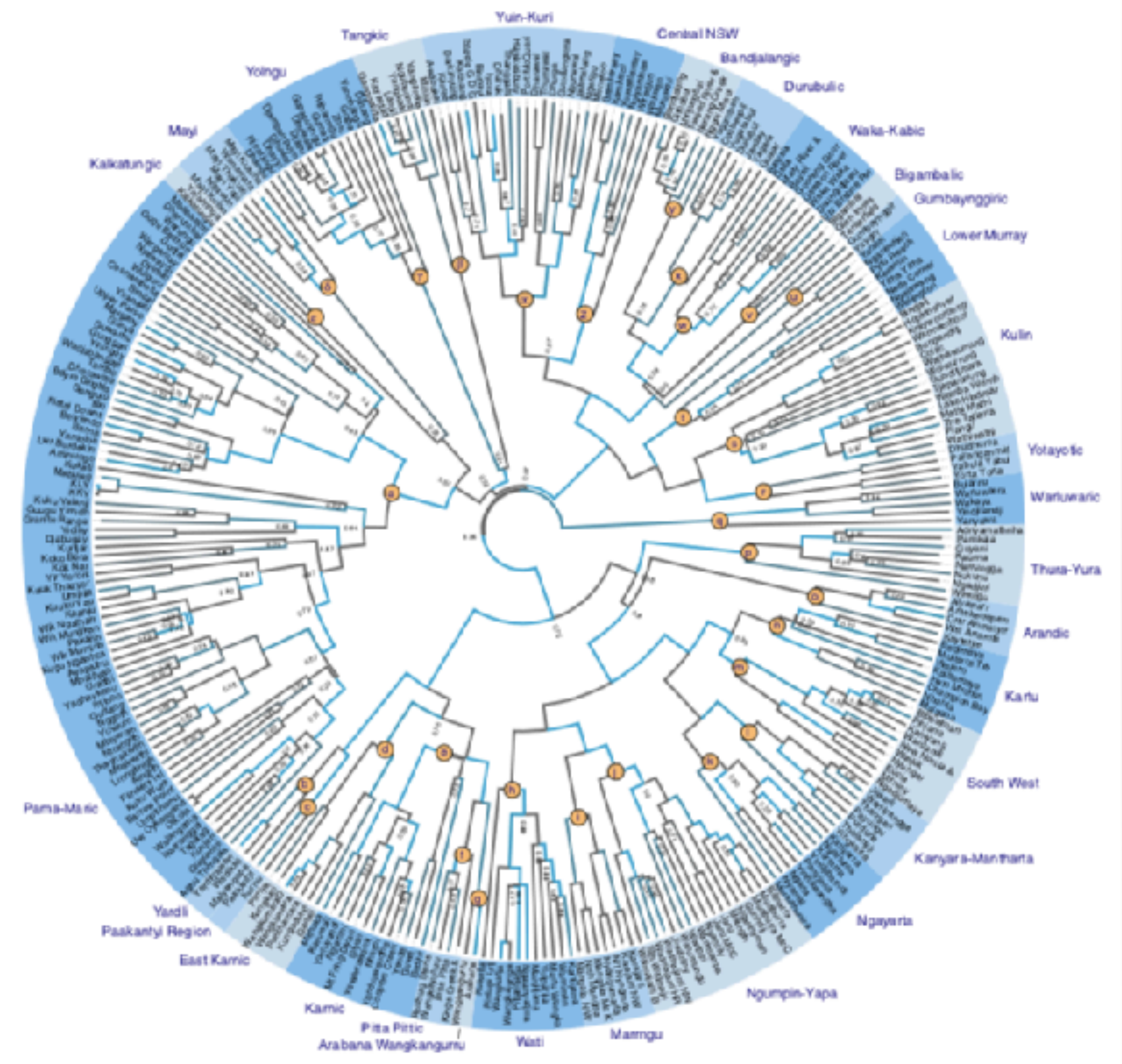
Current genetic and linguistic evidence for the origin of the family is inconclusive. Recent whole-genome analysis of Pama–Nyungan speakers[3] has revealed an intriguing correspondence between genetic and linguistic divergence[18] and finds evidence for an early northeast–southwest split 10–31 ka. However, a lack of samples from non-Pama–Nyungan languages in the north means this signal cannot be tied to Pama–Nyungan specifically. Simple distance-based metrics using the percentage of homologous words or 'cognates' shared between Pama–Nyungan languages imply an

Dendrophilia

Pama-Nyungan 4.5-7kya

genetics suggest no major migration since since settlement ~60kya

borrowings lead to phylogenetic underestimation of age - factor of 10?



What Was

- Phylogenetic approach to the history of languages gets published in the ‘top’ non-specialist journals - reaching a wide audience
    - would be good for specialists to publish there too
- Phylo works by evaluating the probability of a tree+change model, and then builds most likely trees
- It throws away a lot of detail; while making assumptions of its own


Thank You