Language evolution

Brian O'Meara

EEB464 Fall 2019

Could languages evolve?

1: Yes

2: No

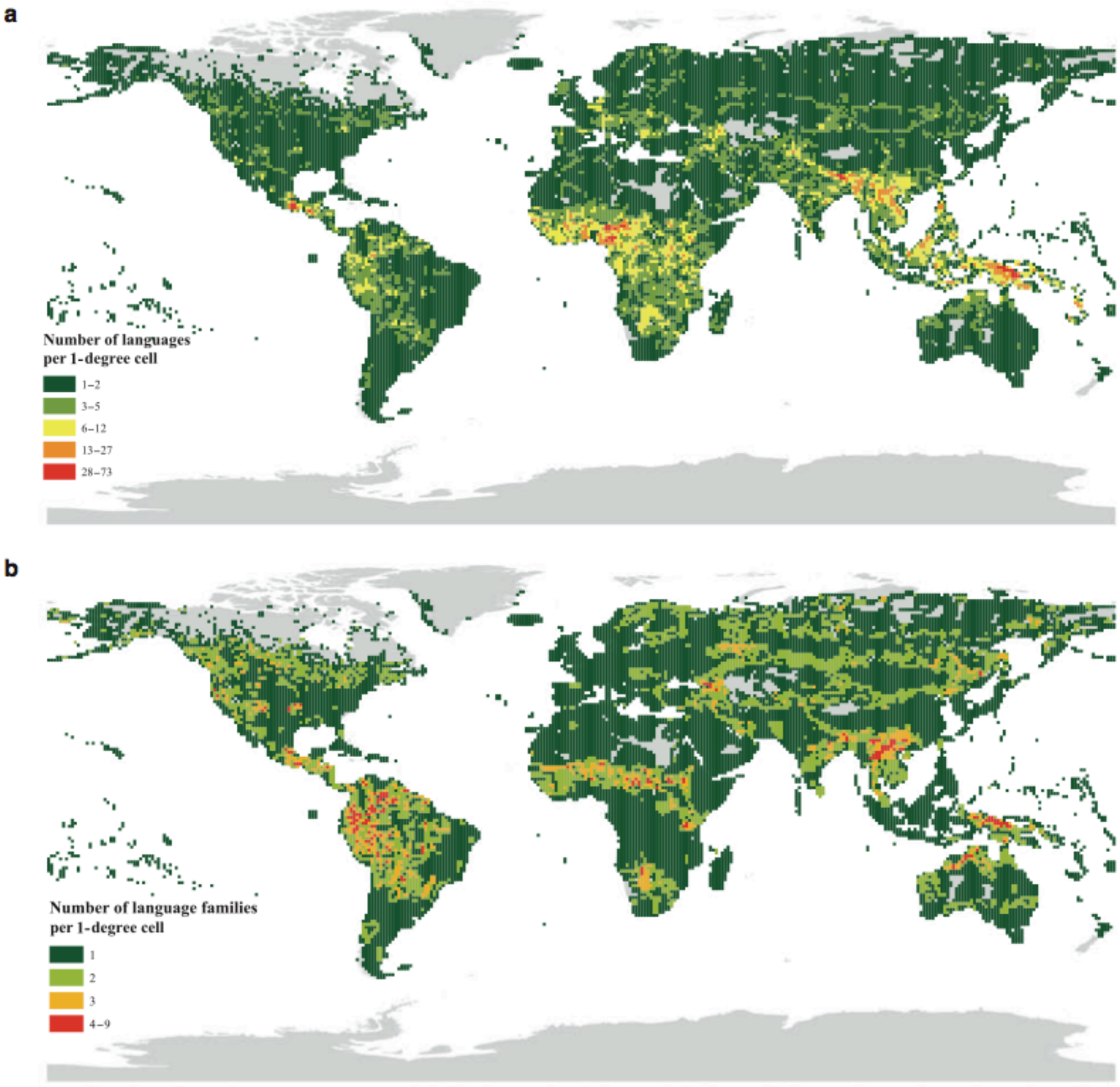
Could languages evolve through natural selection?

1: Yes

2: No

Learning objectives

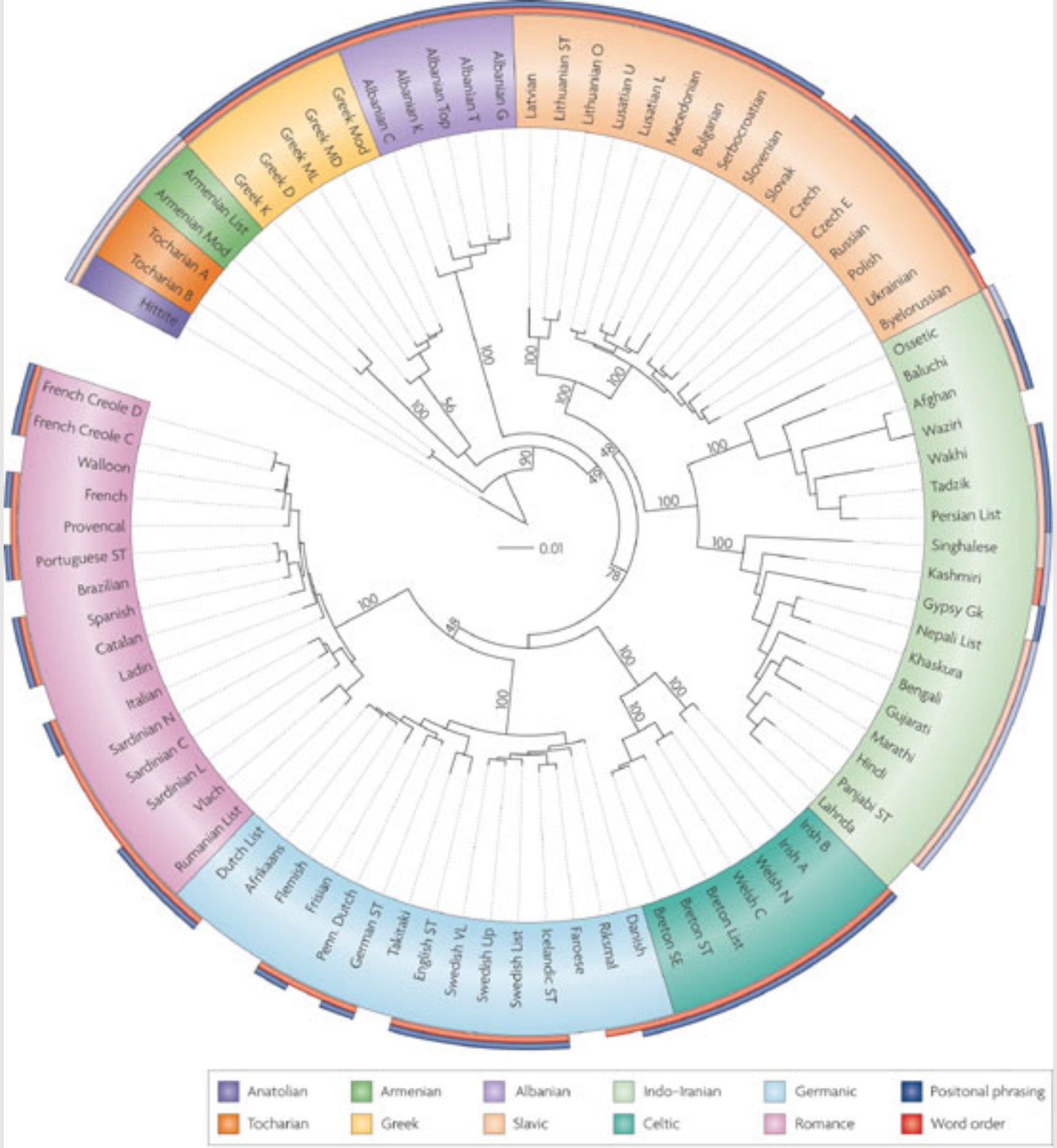
- Understand how languages change
- History of language evolution



Gavin et al. 2013

***Figure 1. Global patterns of linguistic diversity as expressed by (a) the number of languages per 1-degree cell and (b) the number of language families per 1-degree cell. The size of cells used in the analysis for these studies was measured in degrees of latitude and longitude, such that a 1-degree cell spanned 1-degree latitude and 1-degree longitude. The data are from the World Language Mapping System* (www.worldgeodatasets.com/language).**

Words for mother in different languages


Albanian G
Albanian T
Albanian Top
Albanian K
Albanian C
Greek Mod
Greek MD
Greek ML
Greek D
Greek K
Armenian List
Armenian Mod
Tocharian A
Tocharian B
Hittite
Latvian
Lithuanian ST
Lithuanian O
Lusatian U
Lusatian L
Macedonian
Bulgarian
Serbocroatian
Slovenian
Slovak
Czech
Czech E
Russian
Polish
Ukrainian
Byelorussian
Ossetic
Baluchi
Afghan
Wazir
Wakhi
Tadzik
Persian List
Singhalese
Kashmiri
Gypsy Gk
Nepali List
Khaskura
Bengali
Gujarati
Marathi
Hindi
Panjabi ST
Lahnda
Irish B
Irish A
Welsh N
Welsh C
Breton List
Breton ST
Breton SE
Danish
Riksmal
Faroese
Icelandic ST
Swedish List
Swedish Up
Swedish VL
English ST
Takitaki
German ST
Penn. Dutch
Frisian
Flemish
Afrikaans
Dutch List
Rumanian List
Vlach
Sardinian L
Sardinian C
Sardinian N
Italian
Ladin
Catalan
Spanish
Brazilian
Portuguese ST
Provencal
French
Walloon
French Creole C
French Creole D
0.01
Anatolian
Tocharian
Armenian
Greek
Albanian
Slavic
Indo-Iranian
Celtic
Germanic
Romance
Positonal phrasing
Word order

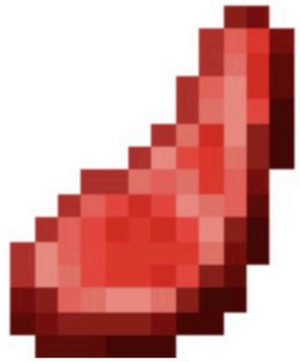
Pagel 2009

In English, think of words for live and cooked animals

Sheep

Health points	8 (♥♥♥♥)
Size	**Adult:** Height: 1.3 Blocks Width: 0.9 Blocks **Baby:** Height: 0.65 Blocks Width: 0.45 Blocks
Spawn	Opaque blocks with at least two block space above them.

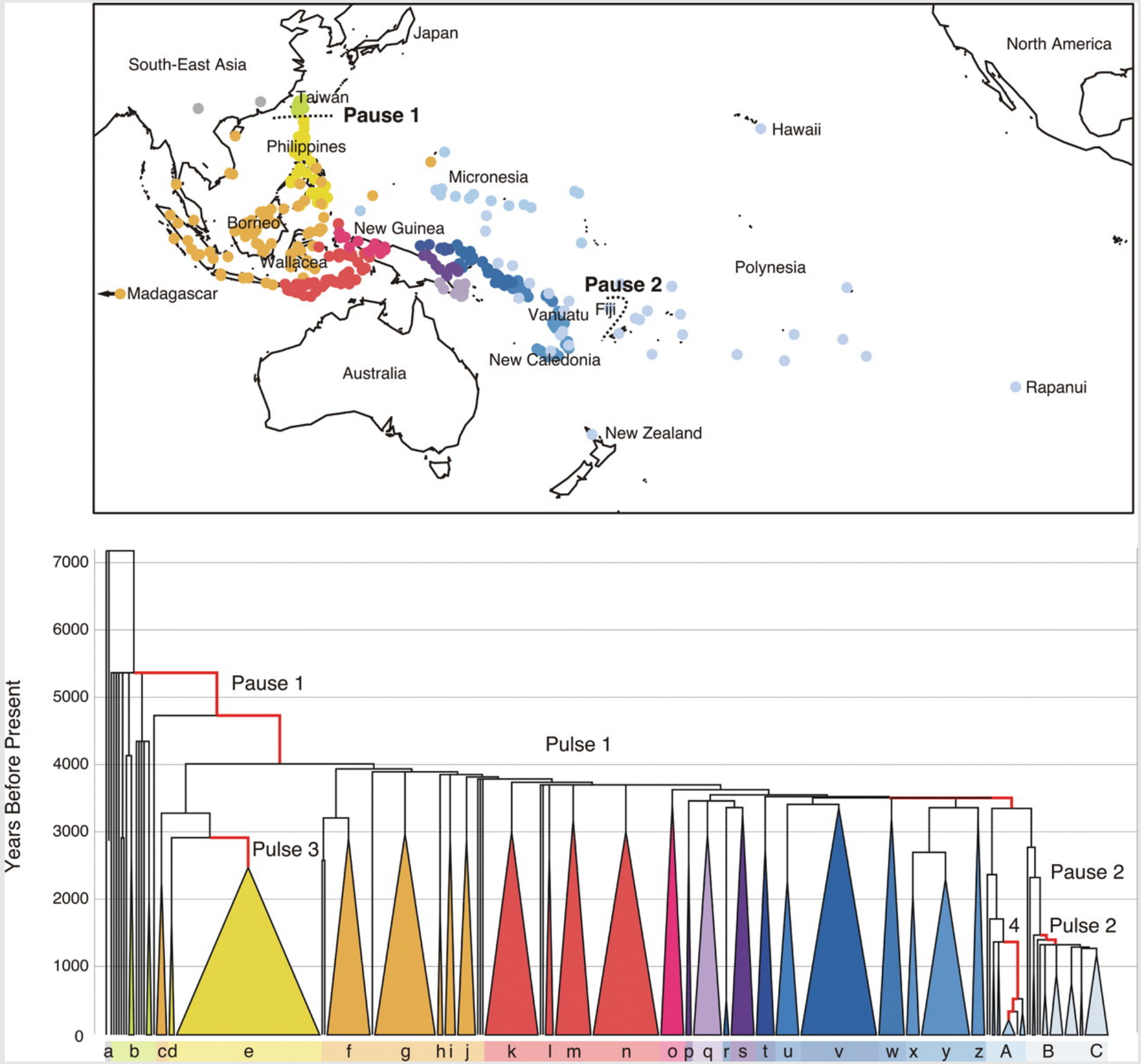
Raw Mutton



Type	Food
Restores	2 (🍖)
Cookable	Yes
Renewable	Yes
Stackable	Yes (64)
Experience	Cooked: 0.35

The Austronesian language family is the one of the largest in the world, with around 1200 languages spread from Taiwan to New Zealand and Madagascar to Easter Island. We have constructed a large database of Austronesian basic vocabulary (23, 26), which stores 210 items of basic vocabulary from each language, including words for animals, kinship terms, simple verbs, colors, and numbers. Basic vocabulary is both relatively stable over time and generally less likely to be borrowed between languages (27). From this database, a team of linguists identified the sets of homologous words ("cognates") following the linguistic comparative method (28). We extracted the cognate sets for 400 well-attested languages for analysis. These languages comprise a third of the entire family and include a representative sample of each recognized Austronesian subgroup. We included two non-Austronesian languages as outgroups to "root" the trees: an archaic variant of the Sino-Tibetan language Chinese that was spoken between 2300 and 2900 years B.P. and the Tai-Kadai language Buyang (28). These languages are not traditionally part of the Austronesian family, but a number of cognates have been identified (29). The cognate sets for all 210 meanings across these 400 languages were encoded into a binary matrix. Identified "borrowings" between languages were removed from further analyses. Simulation studies have shown that the amount of undetected borrowing needs to be very substantial (>20%) to substantially bias either the tree topology or the date estimates (30). The resulting matrix contained a total of 34,440 characters (twice the length of whole mitochondrial genomes), and 6436 of these characters were parsimony informative.

Gray et al. 2009



Gray et al. 2009

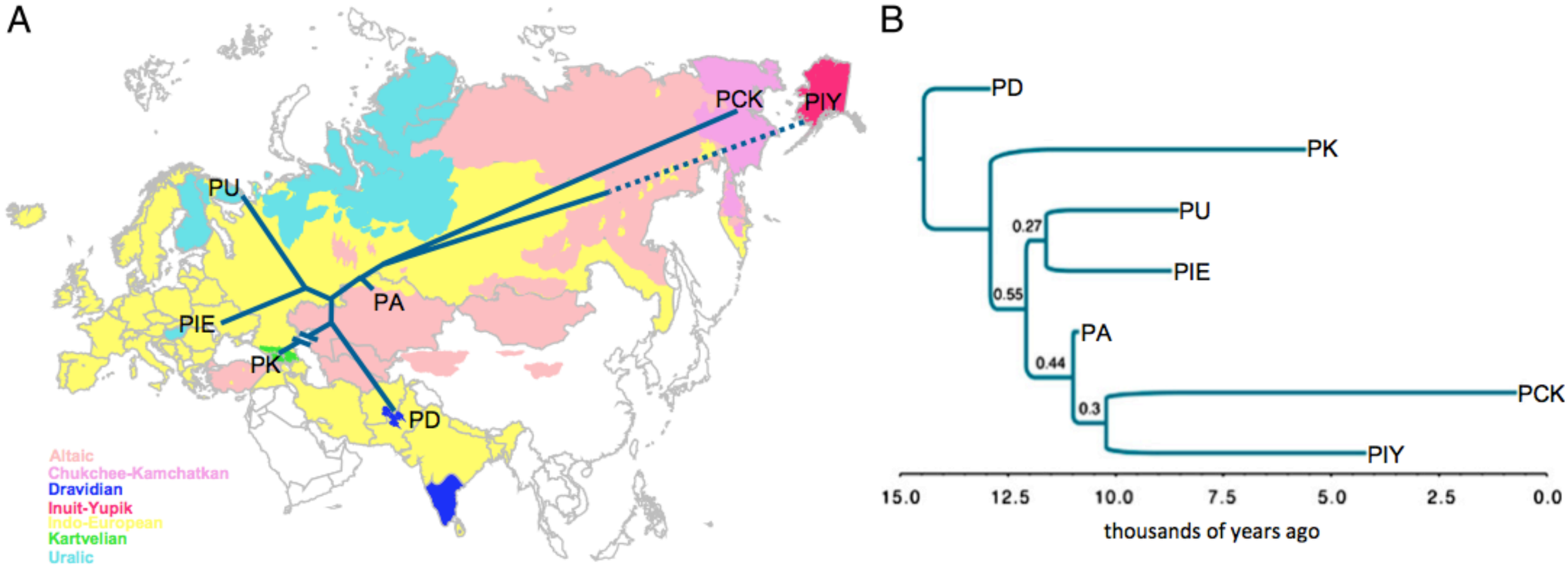


Fig. 4. Consensus phylogenetic tree of Eurasiatic superfamily (*A*) superimposed on Eurasia and (*B*) rooted tree with estimated dates of origin of families and of superfamily. (*A*) Unrooted consensus tree with branch lengths (solid lines) shown to scale and illustrating the correspondence between the tree and the contemporary north-south and east-west geographical positions of these language families. Abbreviations: P (proto) followed by initials of language family: PD, proto-Dravidian; PK, proto-Kartvelian; PU, proto-Uralic; PIE, proto–Indo-European; PA, proto-Altaic; PCK, proto–Chukchi-Kamchatkan; PIY, proto–Inuit-Yupik. The dotted line to PIY extends the inferred branch length into the area in which Inuit-Yupik languages are currently spoken: it is not a measure of divergence. The cross-hatched line to PK indicates that branch has been shortened (compare with *B*). The branch to proto-Dravidian ends in an area that Dravidian populations are thought to have occupied before the arrival of Indo-Europeans (see main text). (*B*) Consensus tree rooted using proto-Dravidian as the outgroup. The age at the root is 14.45 ± 1.75 kya (95% CI = 11.72–18.38 kya) or a slightly older 15.61 ± 2.29 kya (95% CI = 11.72–20.40 kya) if the tree is rooted with proto-Kartvelian. The age assumes midpoint rooting along the branch leading to proto-Dravidian (rooting closer to PD would produce an older root, and vice versa), and takes into account uncertainty around proto–Indo-European date of 8,700 ± 544 (SD) y following ref. 35 and the PCK date of 692 ± 67 (SD) y ago (*SI Text*).

Pagel et al. 2013

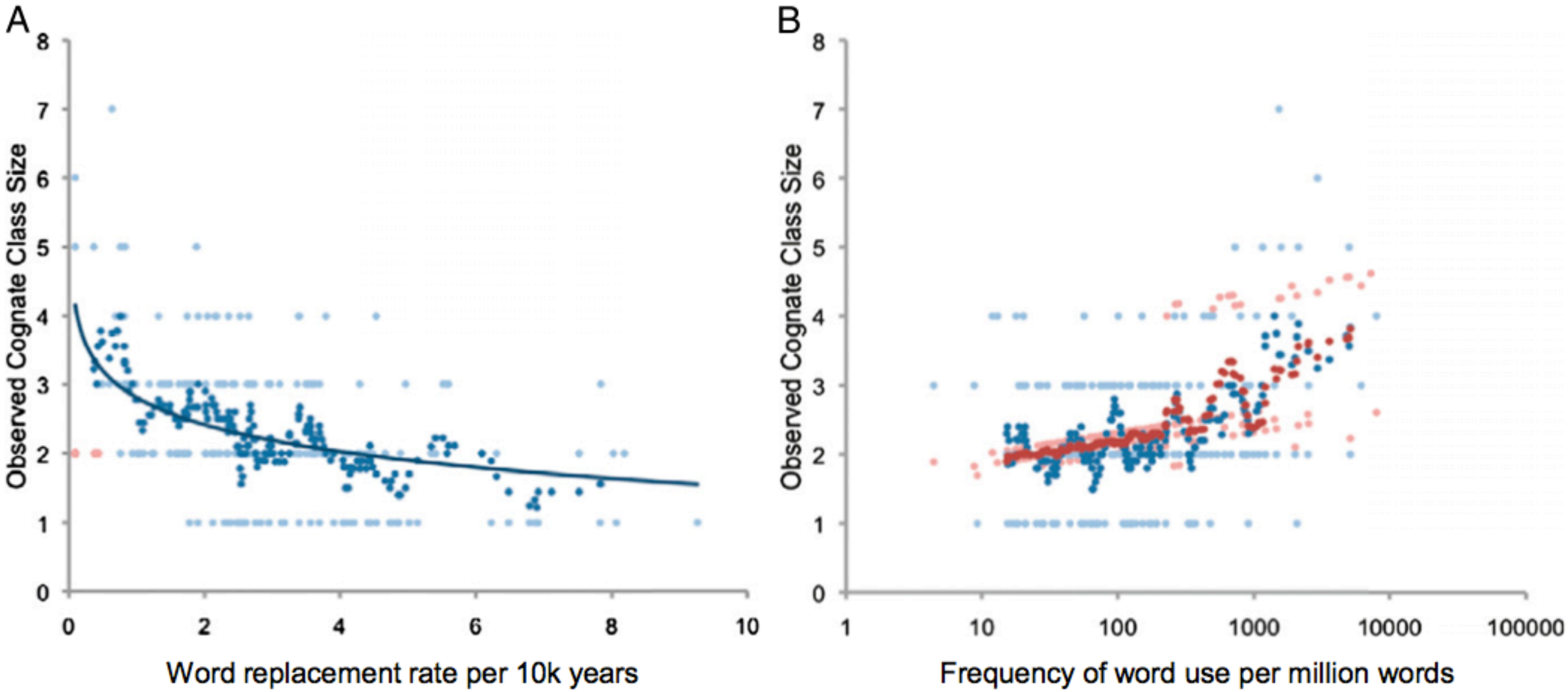


Fig. 3. Rates of lexical replacement (*A*) and word-use frequencies (*B*) predict cognate class sizes among the seven Eurasiatic language families. (*A*) Cognate class size predicted from the rate of lexical replacement as measured in the Indo-European languages [$n = 188$ lexical items, $r = -0.43$, $P < 0.001$; excluding number terms (see text) the correlation increases to $r = -0.55$]. Rates record the expected number of replacements by a new unrelated word per 1,000 years (14); cognate class size is the number out of the seven families for which the proto-word of a given meaning is ancestrally shared; the correlation is fitted to the raw data; smoothed data (darker symbols) are based on a running mean with a window width of 10. (*B*) Cognate class size predicted from a regression model combining frequency of word use and part of speech (see text). The regression is calculated on raw data (blue), smoothed data as in *A* ($r = 0.48$, $P < 0.001$). The trend in *B* is unchanged if we use the principal component factor scores (*Materials and Methods*) in place of mean frequencies.

Pagel et al. 2013

Table 1. Twenty-three words with cognate class sizes of four or more among the Eurasiatic language families

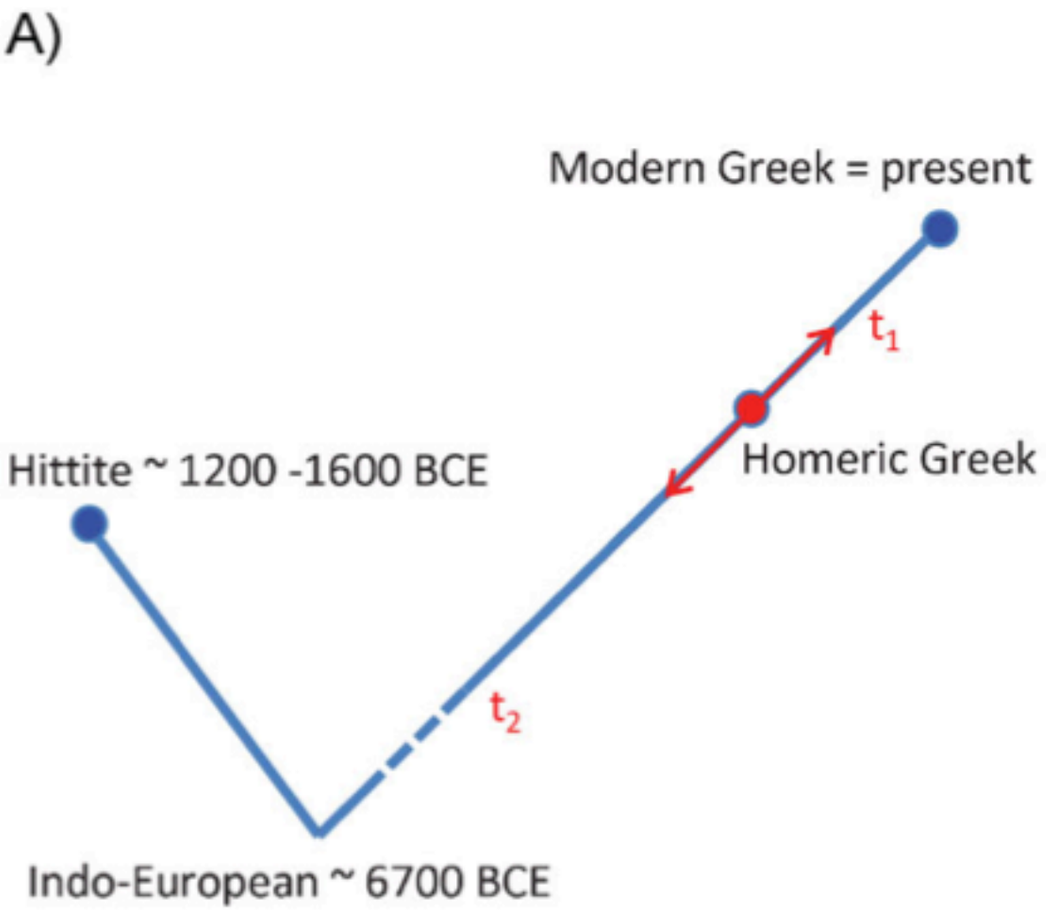
Meaning	Cognate class size*	I-E rate†	Half-life 1,000s of years	Frequency of use‡	Part of speech
Thou	7	0.064	10.83	2,524	Pronoun
I	6	0.009	77	4,332	Pronoun
Not	5	0.082	8.45	7,602	Adverb
That	5	0.188	3.69	5,846	Adjective
We	5	0.037	18.73	2,956	Pronoun
To give	5	0.076	9.12	1,606	Verb
Who	5	0.009	77	1,172	Pronoun
This	4	0.218	3.18	11,185	Adjective
What	4	0.069	10.04	3,058	Adverb
Man/male	4	0.338	2.05	2,800	Noun
Ye	4	0.132	5.25	1,459	Pronoun
Old	4	0.253	2.74	746	Adjective
Mother	4	0.236	2.94	717	Noun
To hear	4	0.235	2.95	680	Verb
Hand	4	0.082	8.45	658	Noun
Fire	4	0.175	3.96	398	Noun
To pull	4	0.453	1.71	279	Verb
Black	4	0.191	3.62	135	Adjective
To flow	4	0.34	2.04	91	Verb
Bark	4	0.379	1.82	49	Noun
Ashes	4	0.265	2.62	23	Noun
To spit	4	0.204	3.38	23	Verb
Worm	4	0.216	3.19	21	Noun

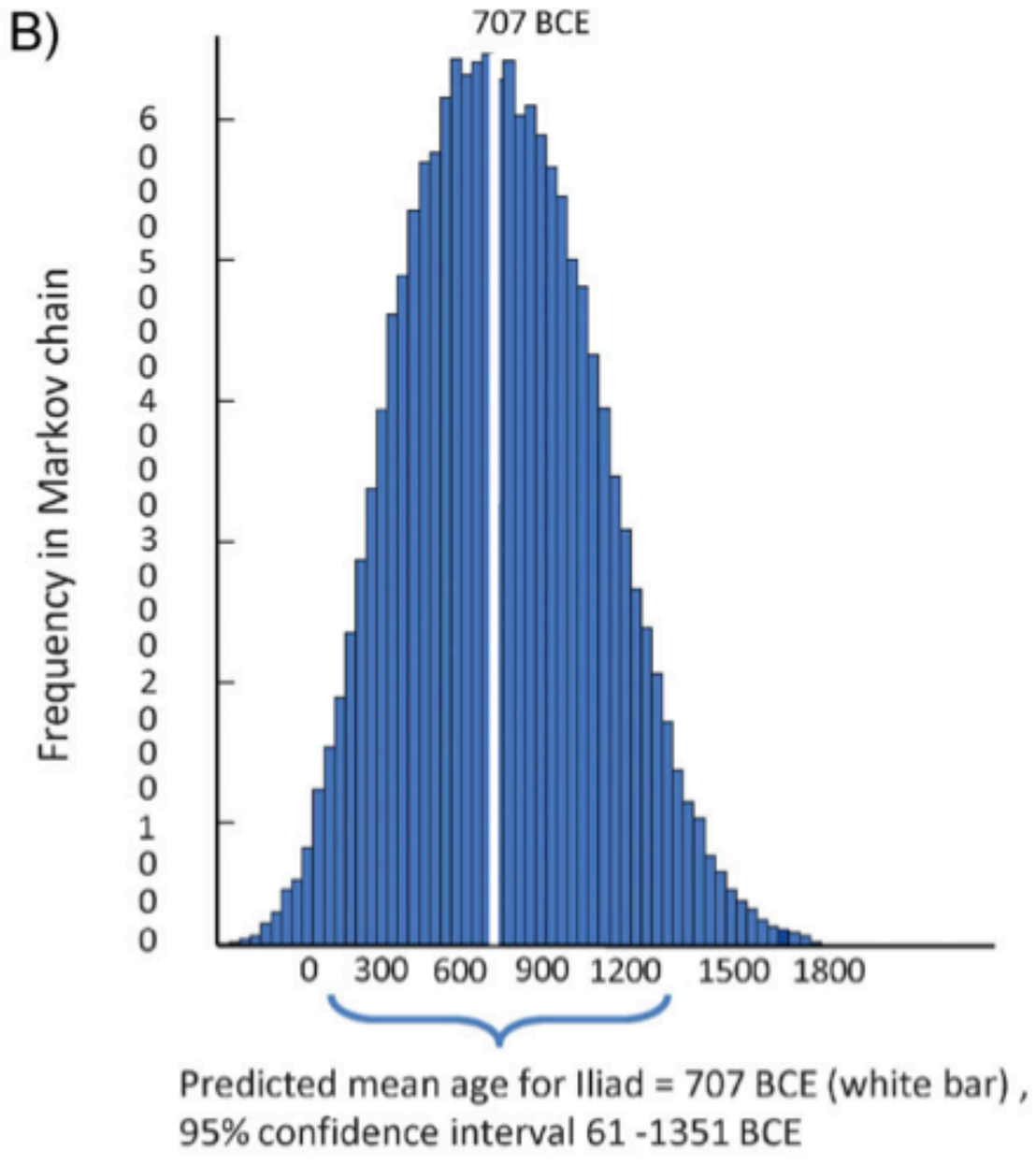
*Defined as the number (of seven) of Eurasiatic language families that are reconstructed as cognate for the word used to convey the meaning shown.
†The rate of lexical replacement measured in number of expected new or unrelated words per 1,000 y and rates of replacement expressed as "half-lives" or the expected time until a word has a 50% chance of being replaced by a new noncognate word (14).
‡The frequency of use per million based on mean of 17 languages from six language families and the two isolates (16).

Pagel et al. 2013

The Homeric epics are among the greatest masterpieces of literature, but when they were produced is not known with certainty. Here we apply evolutionary-linguistic phylogenetic statistical methods to differences in Homeric, Modern Greek and ancient Hittite vocabulary items to estimate a date of approximately 710–760 BCE for these great works. Our analysis compared a common set of vocabulary items among the three pairs of languages, recording for each item whether the words in the two languages were *cognate* – derived from a shared ancestral word – or not. We then used a likelihood-based Markov chain Monte Carlo procedure to estimate the most probable times in years separating these languages given the percentage of words they shared, combined with knowledge of the rates at which different words change. Our date for the epics is in close agreement with historians' and classicists' beliefs derived from historical and archaeological sources.

Altschuler et al. 2013





Altschuler et al. 2013