LETTER

https://doi.org/10.1038/s41586-019-1153-z

Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic

Menghan Zhang[1,2,8], Shi Yan[3,4,8], Wuyun Pan[5,6] & Li Jin[1,3,7]*

The study of language origin and divergence is important for understanding the history of human populations and their cultures. The Sino-Tibetan language family is the second largest in the world after Indo-European, and there is a long-running debate about its phylogeny and the time depth of its original divergence[1]. Here we perform a Bayesian phylogenetic analysis to examine two competing hypotheses of the origin of the Sino-Tibetan language family: the 'northern-origin hypothesis' and the 'southwestern-origin hypothesis'. The northern-origin hypothesis states that the initial expansion of Sino-Tibetan languages occurred approximately 4,000–6,000 years before present (BP; taken as AD 1950) in the Yellow River basin of northern China[2–4], and that this expansion is associated with the development of the Yangshao and/or Majiayao Neolithic cultures. The southwestern-origin hypothesis states that an early expansion of Sino-Tibetan languages occurred before 9,000 years BP from a region in southwest Sichuan province in China[5] or in northeast India[6], where a high diversity of Tibeto-Burman languages exists today. Consistent with the northern-origin hypothesis, our Bayesian phylogenetic analysis of 109 languages with 949 lexical root-meanings produced an estimated time depth for the divergence of Sino-Tibetan languages of approximately 4,200–7,800 years BP, with an average value of approximately 5,900 years BP. In addition, the phylogeny supported a dichotomy between Sinitic and Tibeto-Burman languages. Our results are compatible with the archaeological records, and with the farming and language dispersal hypothesis[7] of agricultural expansion in China. Our findings provide a linguistic foothold for further interdisciplinary studies of prehistoric human activity in East Asia.

Knowledge of prehistoric human populations is founded upon three disciplines: archaeology, genetics and linguistics. The similarities between genetics and linguistics reflect comparable underlying processes of historical population activities[8,9]. Because language carries cultural information, the evolution of language provides insight into prehistoric human culture.

The Sino-Tibetan language family is the second-largest language family in the world, and consists of more than 400 languages and dialects that collectively are spoken by approximately 1.5 billion native speakers[10]. The Sino-Tibetan language family is geographically distributed across East Asia, peninsular Southeast Asia and the northern part of South Asia[3], and includes well-documented languages such as Chinese, Burmese, and Tibetan. Understanding the history of the Sino-Tibetan language family will enable us to delve into the relationships among its member languages, and their interactions with the neighbouring language families such as the Altaic, Austroasiatic, Hmong–Mien, Tai–Kadai and Austronesian families. Additionally, this knowledge is crucial for resolving questions about the source, formation and history of migration of human populations throughout eastern Eurasia.

Although linguistic studies of Sino-Tibetan languages have recently flourished[1], three fundamental issues remain unresolved for reconstructing the early history of the Sino-Tibetan language family. The first issue concerns the primary classification of the Sino-Tibetan languages—especially the position of the Sinitic languages. There are three hypotheses for the primary classification of the Sino-Tibetan languages, the most widely accepted of which is one that proposes a dichotomy between the Sinitic and Tibeto-Burman languages[1]: that is, the Sinitic languages (primarily Chinese and its dialects) are thought to form a primary branch of the Sino-Tibetan family, and all Tibeto-Burman languages form a monophyletic group[2]. An opposing hypothesis considers Sinitic to be a sub-branch within a primary clade of the Sino-Tibetan family[6,11], and the third hypothesis proposes that several parallel clades exist at the root of the Sino-Tibetan family (and that Sinitic is one of these clades)[12]. In addition to these hypotheses, there is the 'fallen leaves' model, which suggests that there are no clearly discernible internal relationships among the primary subgroups of the Sino-Tibetan language family[5].

Other controversial issues—which have their basis in this disputed classification of the Sino-Tibetan languages—include the timing of, and area of origin (*Urheimat*) for, the divergence and expansion of the Sino-Tibetan languages. These controversies can be grouped into two primary hypotheses—the northern-origin and southwestern-origin hypotheses (Table 1). The northern-origin hypothesis argues that people from the upper and/or middle Yellow River basin, who spoke languages ancestral to the Sino-Tibetan family, were divided into two groups[1,2] at 4,000–6,000 years BP: one group then migrated west into Tibet and south into Myanmar (becoming the main ancestors of contemporary populations speaking Tibeto-Burman languages), whereas the other group (speaking an ancestor of the Sinitic languages) moved east- and southward, and ultimately became the Han Chinese[13]. The majority of historical linguists prefer this hypothesis, and suggest that the expansion of the Sino-Tibetan languages is reasonably associated with the development of the Yangshao culture (about 7,000–5,000 years BP) and/or the Majiayao culture (about 5,500–4,000 years BP) in the Neolithic period[14]. By contrast, the southwestern-origin hypothesis suggests that the expansion of the Sino-Tibetan languages occurred

Table 1 | Three hypotheses for the time depth of, and homeland for, the origins and expansion of the Sino-Tibetan language family

	Hypothesis	Date (years BP)	Homeland
I (refs. [2–4])	Northern origin	About 4,000–6,000	The upper and middle Yellow River plains, China
II (ref. [5])	Southwestern origin	Over 9,000	Southwest Sichuan province, China
III (ref. [6])		About 9,000–10,000	Northeast India

[1]State Key Laboratory of Genetic Engineering, and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. [2]Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China. [3]Human Phenome Institute, Fudan University, Shanghai, China. [4]Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China. [5]Institute for Humanities and Social Science Data, School of Data Science, Fudan University, Shanghai, China. [6]Institute of Linguistics, College of Humanities and Communications, Shanghai Normal University, Shanghai, China. [7]Chinese Academy of Sciences Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai, China. [8]These authors contributed equally: Menghan Zhang, Shi Yan.
*e-mail: lijin@fudan.edu.cn

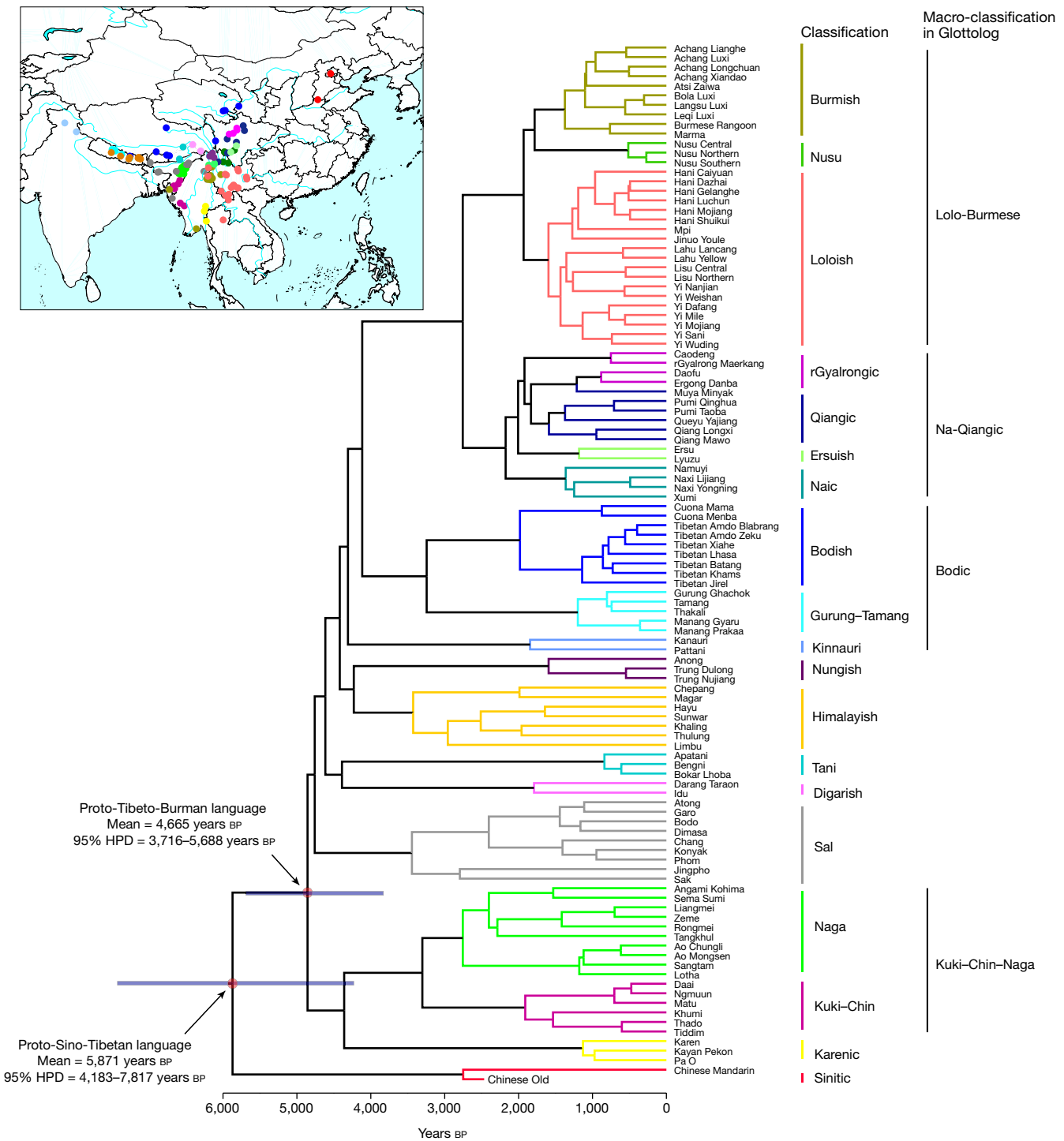


Fig. 1 | The maximum clade credibility tree and the geographical distribution of the Sino-Tibetan languages. Language sample size, $n = 109$. The maximum clade credibility tree shows the divergence time estimates for proto-Sino-Tibetan and proto-Tibeto-Burman, the emergence of the major branches, and the subsequent diversification of these branches. The colours of the leaf nodes and branches represent language sub-groups or languages in linguistic documents (Supplementary Table 1). The two blue bars at the red nodes show the 95% highest posterior density (HPD) of the respective divergence times for proto-Sino-Tibetan and proto-Tibeto-Burman. The iterations for tree reconstruction in BEAST2 software were set to 50 million generations, sampling trees in every 5,000 generations and resulting in a sample of 10,000 trees. The first 10% of the iterations were treated as burn-in. The maximum clade credibility tree was established from 9,000 trees. This tree is available as Supplementary Data (filename '109SinoTibetanlanguage.MCC.tree'). The geographical distribution of the 109 languages is shown in the top left (see Extended Data Fig. 1). The map is based on vector map data from https://www.naturalearthdata.com.

at approximately 9,000–10,000 years BP from the southwest region of East Asia. The southwestern-origin hypothesis exists in two main forms (see Table 1), here named hypothesis II and hypothesis III. Hypothesis II argues that populations speaking Sino-Tibetan languages have their origins in southwest Sichuan province as early as 13,500 years BP, and divided into two groups at approximately 9,000 years BP: one group travelled into northeast India and the other travelled north to the Yellow River basin; the latter group is considered to be the predecessor of contemporary Chinese and the Tibetan populations[5]. Hypothesis III supposes that the Sino-Tibetan languages had their origins in northeast

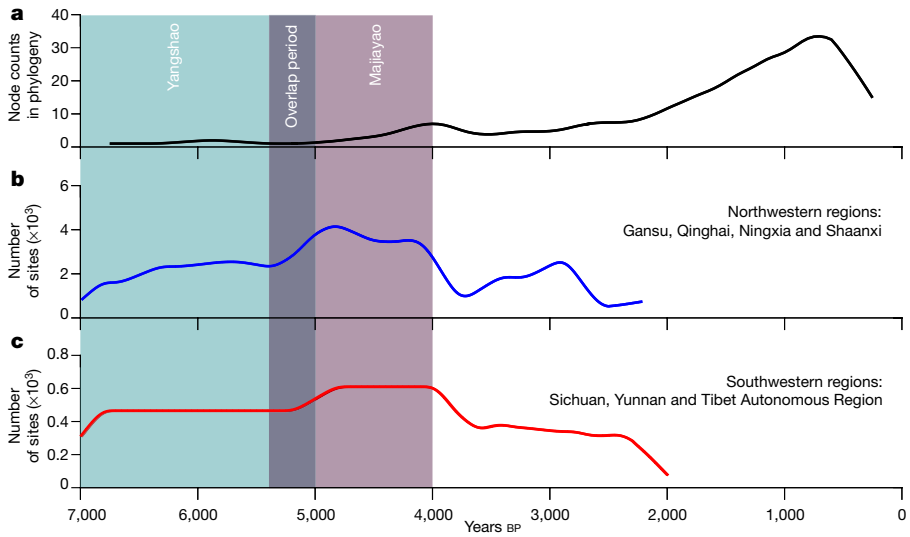


Fig. 2 | The tempo of divergence of the Sino-Tibetan languages and the changes in the number of archaeological sites in China. a, The black curve shows the tempo of the divergence of Sino-Tibetan languages (based on 109 language) in the Sino-Tibetan Bayesian phylogeny. **b**, **c**, Changes in the number of archaeological sites mapped in the northwest ($n = 12{,}356$ sites) and southwest regions of China ($n = 1{,}169$ sites), dating to between about 7,000 and 2,500 years BP. The number of sites was taken from a previous publication[30].The coloured regions show the time periods that correspond to the Yangshao (about 7,000–5,000 years BP) and Majiayao (about 5,300–4,000 years BP) Neolithic cultures, and the grey region refers to the period that these two cultures overlapped.

India at approximately 9,000 years ago, and the earliest speakers of Sino-Tibetan languages are seen as highly diverse foragers rather than as agriculturists[6].

Disputations on *Urheimat* of the Sino-Tibetan languages are interwoven with phylogenetic uncertainty and disagreements over time depths for divergences. All linguists that have discussed the relationship of language and archaeological cultures in East Asia have linked the Yangshao and Majiayao cultures to the Sino-Tibetan language family, owing to the clear archaeological links between these archaeological cultures and the Zhou dynasty of China and at least some modern Tibeto-Burman-speaking populations[2,5]. The question is whether the geographical regions associated with Yangshao and Majiayao Neolithic cultures are the primary divergence location for the Sino-Tibetan language family. An explicit language phylogenetic reconstruction and reliable divergence-time estimation are essential for extrapolating the *Urheimat* of the Sino-Tibetan languages. In historical linguistics, the comparative method—using abundant contemporary materials and historical documents—is an approach that has widely been used to establish language relationships[15]. Glottochronology, which is an extension of the comparative method, uses lexical data to estimate absolute divergence times[16]. However, glottochronology has considerable limitations (such as assuming a constant rate of language evolution[17]) and does not account for different evolutionary rates of languages owing to contact, environmental change or varied rates of substitution among different categories of words[18]. These issues also frequently appear in linguistic studies of the Sino-Tibetan languages. For example, language contact between the languages of the Sino-Tibetan family and surrounding non-Sino-Tibetan language families—such as Hmong–Mien and Tai–Kadai—was prevalent in East Asia. Furthermore, the lack of well-documented historical accounts and comprehensive linguistic surveys of the Sino-Tibetan languages is challenging for linguistic comparative approaches. However, recent advances in Bayesian phylogenetic methods from evolutionary biology provide alternative opportunities to circumvent these limitations. These approaches permit flexible evolutionary models, and are a powerful tool for inferring evolutionary tempo and mode of change in language families worldwide[19–21].

To examine the northern-origin and southwestern-origin hypotheses, we performed Bayesian phylogenetic analyses on 949 binary-coded lexical root-meanings of 109 languages of the Sino-Tibetan family that were geographically distributed across China, Southeast Asia (for example, Myanmar) and South Asia (for example, India, Nepal and Bhutan) (Fig. 1, Extended Data Fig. 1). By matching the lexical meanings in the Swadesh 100-word list[22], the root-meanings were identified and collated manually under multiple selection criteria (Supplementary Information, section 1.1). Several time calibrations of known historical events were used to estimate the divergence time for Sino-Tibetan languages (Supplementary Table 2), and several model combinations were compared (Extended Data Fig. 2). Unlike previous studies[20], we performed the analyses without any ancestral or monophyletic constraints as priors, to avoid artificial bias during the phylogenetic reconstruction.

The results of the Bayesian phylogenetic analyses showed a dichotomy between Sinitic and Tibeto-Burman languages, in which the traditionally categorized Tibeto-Burman clade was confirmed as a monophyly that was supported by a posterior value of 0.68 (Fig. 1, Extended Data Fig. 3). The reliability of several controversial subgroupings of the Sino-Tibetan language family were also evaluated (Supplementary Information, section 2.3, Extended Data Fig. 4). The average time estimate for the initial divergence of Sino-Tibetan languages (5,871 years BP) (Fig. 1, Extended Data Fig. 5) occurred in the time of the Yangshao Neolithic culture, and the initial Tibeto-Burman divergence time (4,665 years BP) (Fig. 1, Extended Data Fig. 6) was estimated to have occurred in the middle period of the Majiayao culture (which derived from the Yangshao culture)[23]. Accordingly, the phylogeny and divergence time were consistent with the northern-origin hypothesis; that is, the primary divergence of the Sino-Tibetan languages could be associated with these two Neolithic cultures in northern China (Fig. 2a, Supplementary Information, section 2.4). The estimated divergence time of Tibeto-Burman languages was compatible with genetic evidence from Y chromosome data[24].

Archaeological evidence[25] suggests that the association of the diversification of Sino-Tibetan languages and the development of the two Neolithic cultures might be attributable to rapid demographic growth and the spread of agriculture. A rapidly increasing number of archaeological sites and sustained deforestation in the Yellow River basin indicate two rapid population increases that began at approximately 6,000 years BP and intensified at 5,000–4,500 years BP[25] (Fig. 2b, c). This timeline coincides with the time estimates for the divergence of Sino-Tibetan and Tibeto-Burman languages in this study. Moreover, the Sino-Tibetan language dispersal could be related to the geographical spread of millet agriculture after 6,000 years BP, which is consistent with

the farming and language dispersal hypothesis[26]. A range of archaeological evidence—such as architectural forms, and patterns and shapes of pottery—also shows consecutive waves of southward dispersal along the Tibetan–Yi ethnic corridor into the western Sichuan and western Yunnan provinces; this dispersal could trace back to the Yangshao, Majiayao and Qijia (a Neolithic-to-Bronze Age culture that succeeded the Majiayao culture at approximately 4,300–3,500 years BP) cultures[27]. In particular, the spread of millet agriculture occurred primarily from what is now northern China (especially from the Yellow River basin) to the west and south, along the edge of Tibetan plateau, after 5,000 years BP (Supplementary Information, section 2.5). In addition, following previous work[20], we performed an *Urheimat* inference for the Sino-Tibetan languages (Extended Data Fig. 7). However, the prerequisites for *Urheimat* inference were not satisfied in the case of the Sino-Tibetan languages (Supplementary Information, section 2.7).

Although we adopt a family-tree model to demonstrate the lineages among the Sino-Tibetan languages, we do not claim that the cultural history of Sino-Tibetan speakers is indeed tree-like. Population migrations and interactions among speakers of Sino-Tibetan languages were complex and occurred over a long period of time[28]. As a consequence, substantial language contact among the Sino-Tibetan languages could have occurred at an early stage of the diversification of these languages, and could continue into the present. These contacts have previously been known to have occurred among the Sino-Tibetan languages, as well as with the surrounding Austronesian, Tai–Kadai and Hmong–Mien languages[29] (Supplementary Information, sections 2.5, 2.6). Unfortunately, we cannot yet provide concrete data for identifying these influences, and we cannot, therefore, reconstruct the explicit genetic relationships between the Sino-Tibetan languages and other language families. Many Sino-Tibetan languages remain poorly described, which makes it challenging to perform definite comparisons in historical linguistics. Thus, the study of the evolution of the Sino-Tibetan languages is at an early stage and requires additional interdisciplinary data. To explicitly demonstrate the evolution of the Sino-Tibetan languages, we need comprehensive archaeological surveys and sufficient evidence from studies of ancient DNA.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1153-z.

Received: 13 January 2019; Accepted: 28 March 2019;
Published online 24 April 2019.

1. Handel, Z. What is Sino-Tibetan? Snapshot of a field and a language family in flux. *Lang. Linguist. Compass* **2**, 422–441 (2008).
2. LaPolla, R. J. in *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics* (eds Dixon, R. M. W. & Aikhenvald, A. Y.) 225–254 (Oxford Univ. Press, Oxford, 2001).
3. Matisoff, J. A. Sino-Tibetan linguistics: present state and future prospects. *Annu. Rev. Anthropol.* **20**, 469–504 (1991).
4. LaPolla, R. J. & Thurgood, G. *Sino-Tibetan Languages* (Routledge, London, 2016).
5. van Driem, G. in *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics* (eds Sagart, L. et al.) 81–106 (Routledge, London, 2005).
6. Blench, R. & Post, M. in *Trans-Himalayan Linguistics: Historical and Descriptive Linguistics of the Himalayan Area* (eds Hill, N. & Owen-Smith, T.) 71–104 (De Gruyter Mouton, Berlin, 2013).
7. Bellwood, P. in *The Peopling of East Asia* (eds Sagart, L. et al.) 41–54 (Routledge, London, 2005).
8. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl Acad. Sci. USA* **85**, 6002–6006 (1988).
9. Longobardi, G. et al. Across language families: genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.* **157**, 630–640 (2015).
10. Simons, G. F. & Fennig, C. D. *Ethnologue: Languages of the World* 21st edn (SIL International, Dallas, 2018).
11. Peiros, I. & Starostin, S. *A Comparative Vocabulary of Five Sino-Tibetan Languages* (Univ. of Melbourne, Department of Linguistics and Applied Linguistics, Melbourne, 1996).
12. Shafer, R. Classification of the Sino-Tibetan languages. *Word* **11**, 94–111 (1955).
13. Fei, X.-T. On the problem of distinguishing nationalities in China. *Soc. Sci. China* **1**, 158–174 (1980).
14. Janhunen, J. *Manchuria: an Ethnic History* (Finno-Ugrian Society, Helsinki, 1996).
15. Campbell, L. *Historical Linguistics* (Edinburgh Univ. Press, Edinburgh, 2013).
16. Lees, R. B. The basis of glottochronology. *Language* **29**, 113–127 (1953).
17. Greenhill, S. J., Atkinson, Q. D., Meade, A. & Gray, R. D. The shape and tempo of language evolution. *Proc. R. Soc. Lond. B* **277**, 2443–2450 (2010).
18. Aikhenvald, A. Y. & Dixon, R. M. W. in *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics* (eds Dixon, R. M. W. & Aikhenvald, A. Y.) 1–26 (Oxford Univ. Press, Oxford, 2001).
19. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
20. Bouckaert, R. et al. Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960 (2012).
21. Kolipakam, V. et al. A Bayesian phylogenetic study of the Dravidian language family. *R. Soc. Open Sci.* **5**, 171504 (2018).
22. Swadesh, M. Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137 (1955).
23. Liu, L. & Chen, X. *The Archaeology of China: From the Late Paleolithic to the Early Bronze Age* (Cambridge Univ. Press, Cambridge, 2012).
24. Su, B. et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* **107**, 582–590 (2000).
25. Stevens, C. J. & Fuller, D. Q. The spread of agriculture in eastern Asia. *Lang. Dyn. Chang.* **7**, 152–186 (2017).
26. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).
27. Li, K. S. *Yunnan kaoguxue lunji (in Chinese)* (Yunnan People's Publishing House, Kunming, 1998).
28. Handel, Z. J. *Old Chinese Medials and Their Sino-Tibetan Origins: A Comparative Study* (Institute of Linguistics, Academia Sinica, 2009).
29. Sagart, L., Blench, R. & Sanchez-Mazas, A. in *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics* (eds Sagart, L. et al.) 1–14 (Routledge, London, 2005).
30. Hosner, D., Wagner, M., Tarasov, P. E., Chen, X. & Leipe, C. Spatiotemporal distribution patterns of archaeological sites in China during the Neolithic and Bronze Age: an overview. *Holocene* **26**, 1576–1593 (2016).

Acknowledgements This study is supported by projects at the National Natural Science Foundation of China (31521003, 31501010 and 31401060), the Postdoctoral Science Foundation of China (2015M570316 and 2015T80394), the Special Program for Key Basic Research of the Ministry of Science and Technology of the People's Republic of China (2015FY111700), the Science and Technology Commission of Shanghai Municipality (16JC1400500 and 2017SHZDZX01) and the National Social Science Fund of China (13&ZD132 and 18ZDA296).

Reviewer information *Nature* thanks Joshua B. Plotkin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.Z., S.Y., W.P. and L.J. designed the research; M.Z., S.Y. and W.P. collated the linguistic and geographical data; M.Z. and S.Y. performed the research; M.Z., S.Y., W.P. and L.J. analysed the results; and M.Z., S.Y. and L.J. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1153-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1153-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.J.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Lexical root-meaning data. Based on a series of selection criteria (see Supplementary Information, section 1.1), we chose 109 Sino-Tibetan language samples that were well-attested in low-level subgroups[31]. The 109 languages involved 108 contemporary languages and one ancient language (Old Chinese), and their language branches were named according to Glottolog[32] and Ethnologue[10] (Supplementary Table 1). According to the items in the Swadesh 100-word list[22], we collated 949 lexical root-meanings across 109 Sino-Tibetan languages from the STEDT database, under multiple strict selection criteria (Supplementary Information, section 1.1). These root-meanings were coded as discrete binary characters (Supplementary Table 1; available in Nexus format as Supplementary Data (filename '109SinoTibetanLanguage_Swadesh100.nex')).

Phylogenetic reconstruction and divergence time estimates. We used BEAST2 software (v.2.4.8)[33] with the Babel package to establish the phylogenetic trees of the Sino-Tibetan languages, and to estimate the time depth of the root. The Babel package is used for performing linguistic analysis (https://github.com/rbouckaert/Babel). No outgroup is required a priori in BEAST; instead, BEAST samples the root position along with the rest of the nodes in the tree. We evaluated six combinations of two lexical root-meaning models (continuous time Markov chains (CTMC) and covarion), clock models (strict and relaxed log normal clock) and the gamma rate heterogeneity with one or four rate categories for the CTMC model (Supplementary Information, section 1.2). Because not all language samples in our analysis were contemporary, we selected the coalescent Bayesian skyline model as the tree prior.

To estimate the root age of the Sino-Tibetan language family, we scaled the trees by incorporating calibrations based on known anthropological and demographic evidence and historical accounts (Supplementary Table 2, Supplementary Data (filename 'BinaryCovarion.RelaxedClock.xml')). These calibrations were expressed in terms of probability distributions. According to prior settings of calibrations in previous work[20,34], we selected two kinds of distributions: the normal distribution for situations that involve a probable date with an evenly distributed estimated-error factor, and the uniform distribution for assessing the potential ranges of divergence times. Our goal was to obtain an inferred phylogeny of the Sino-Tibetan languages. Therefore, we did not set any monophyletic constraints as priors, even if they were well-attested language branches.

The lexical root-meaning dataset was analysed with each of the six combinations of clock models, site models and rate heterogeneity. We ran each of the 6 combinations for 50 million generations, sampling trees in every 5,000 generations, which resulted in a sample of 10,000 trees. The first 10% of the iterations were treated as burn-in, and discarded. Tracer (v.1.6)[35] was used to check autocorrelation and convergence status, and to test the best-fitting model combination by $\log_{10}$ Bayes factors (Supplementary Information, section 2.1, Supplementary Table 3). The maximum clade credibility tree was generated using the TREEANNOTATOR package in BEAST v.2.4.8.

Estimates of the tempo of language divergence and the changes in the number of archaeological sites. To estimate the divergence tempo, we applied a sliding window to the language phylogeny, which shifted from 7,000 years BP to the present. In each window, we counted the numbers of internal nodes. The mean height of each internal node was obtained from the given Sino-Tibetan language phylogeny (see height data in Supplementary Data (filename '109SinoTibetanlanguage.MCC.tree')). The length of the sliding window was set to 500 years, and the shift step was set to 50 years. Using this estimation, we obtained the curve for node counts of the whole phylogeny. The curve was smoothed by the locally weighted scatter plot smoothing algorithm[36] implemented as the smooth function in MATLAB 2015b, in which the parameter 'smooth span' was set to 0.1.

Similarly, we applied the same approach to count the changes in the number of archaeological sites, shown in Fig. 2b, c. Specifically, we investigated the number of archaeological sites in two geographical regions in China: northwest (including Gansu, Qinghai, Ningxia and Shaanxi), and southwest (including Sichuan, Yunnan and the Tibet Autonomous Region). The data from the archaeological sites were collected from a previous publication[30] (Supplementary Information, section 1.5). At each time point between 7,000 to 2,500 years BP (the moving step was set to 1), we counted the number of sites that had a time range that covered the time point (Supplementary Data (filename 'Matlab codes for estimation of Sino-Tibetan evolutionary tempo.zip')).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

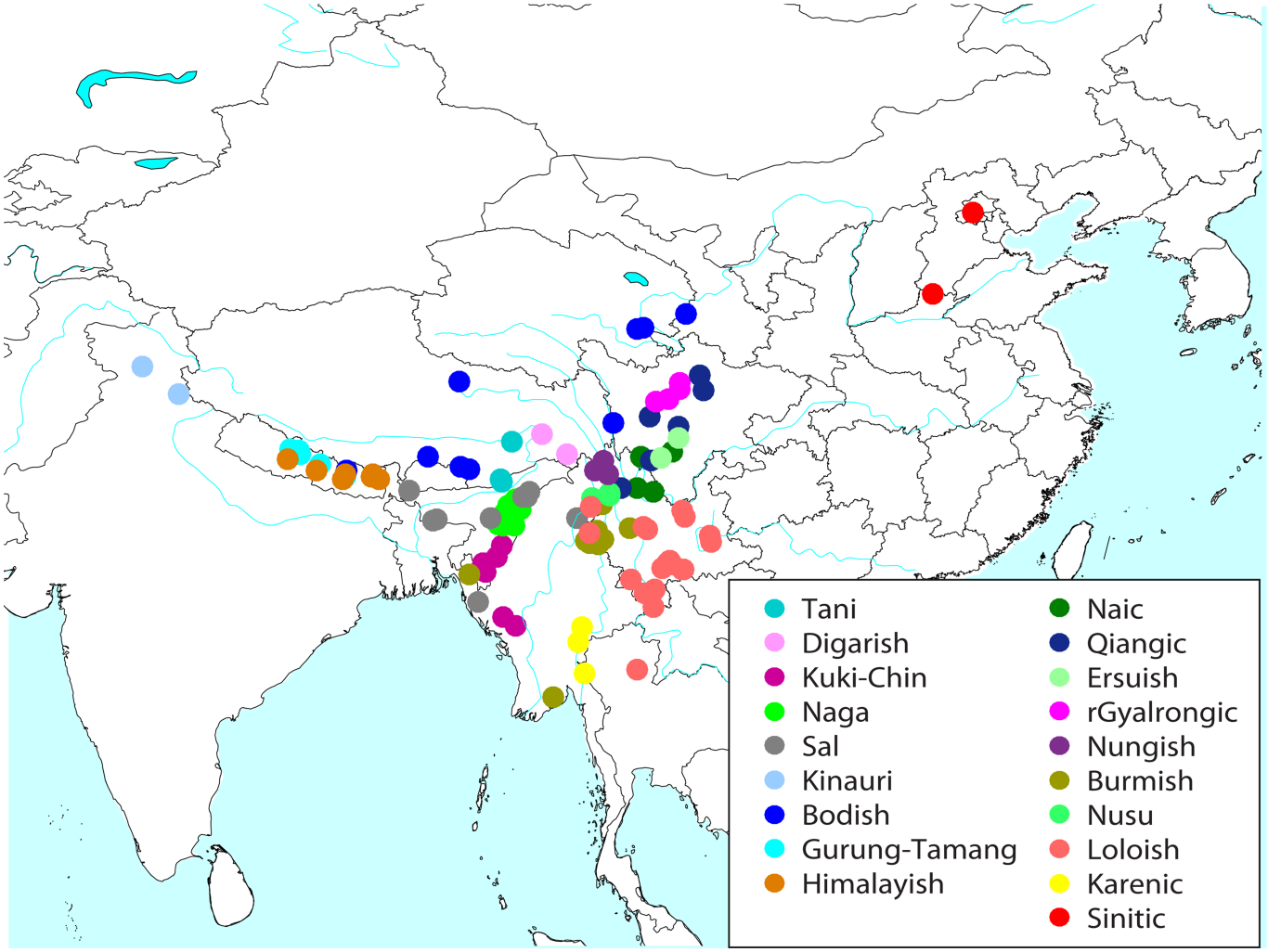
Data availability

The data supporting the findings of this study are available in the Supplementary Information. Any other relevant data are available from the corresponding author upon reasonable request.

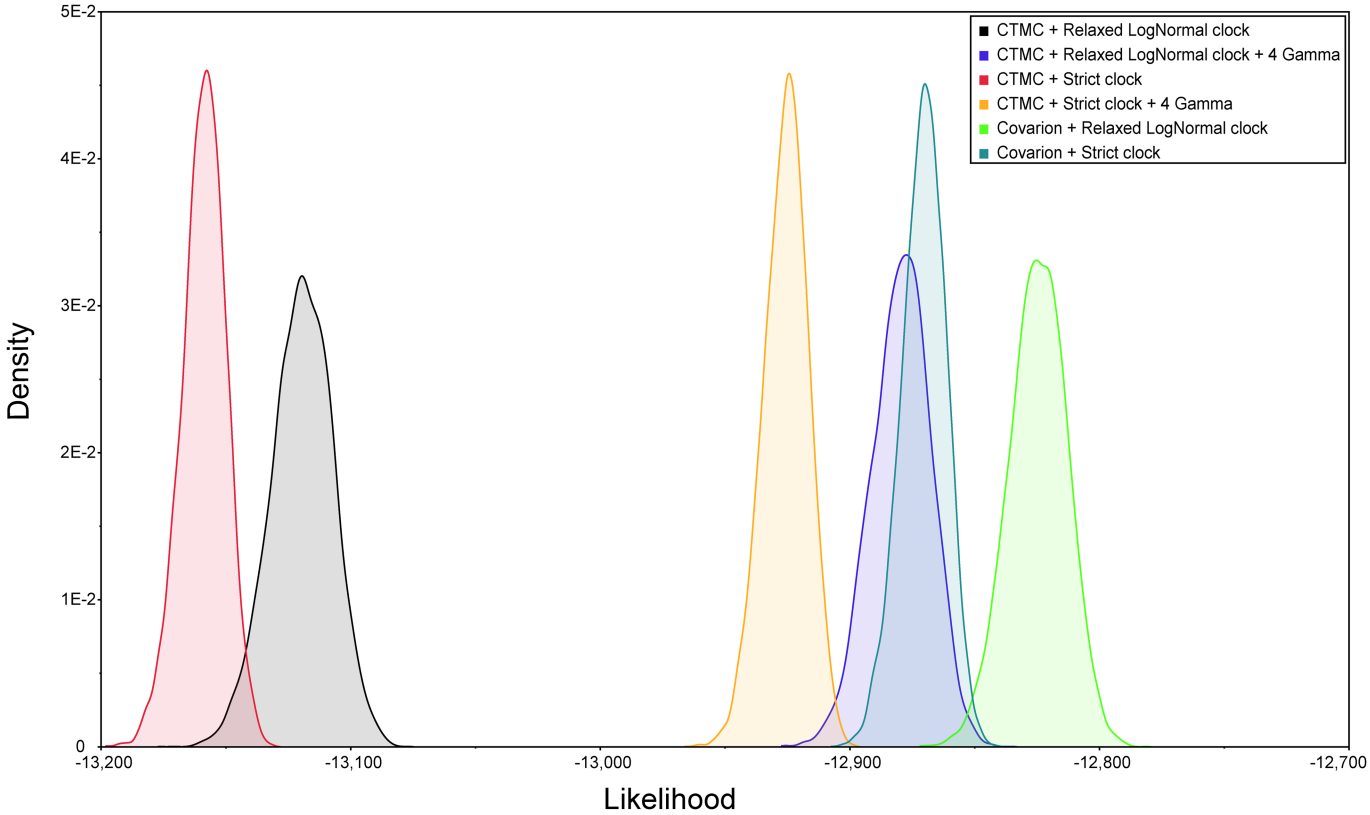
Code availability

The codes supporting the findings of this study are available in the Supplementary Information.

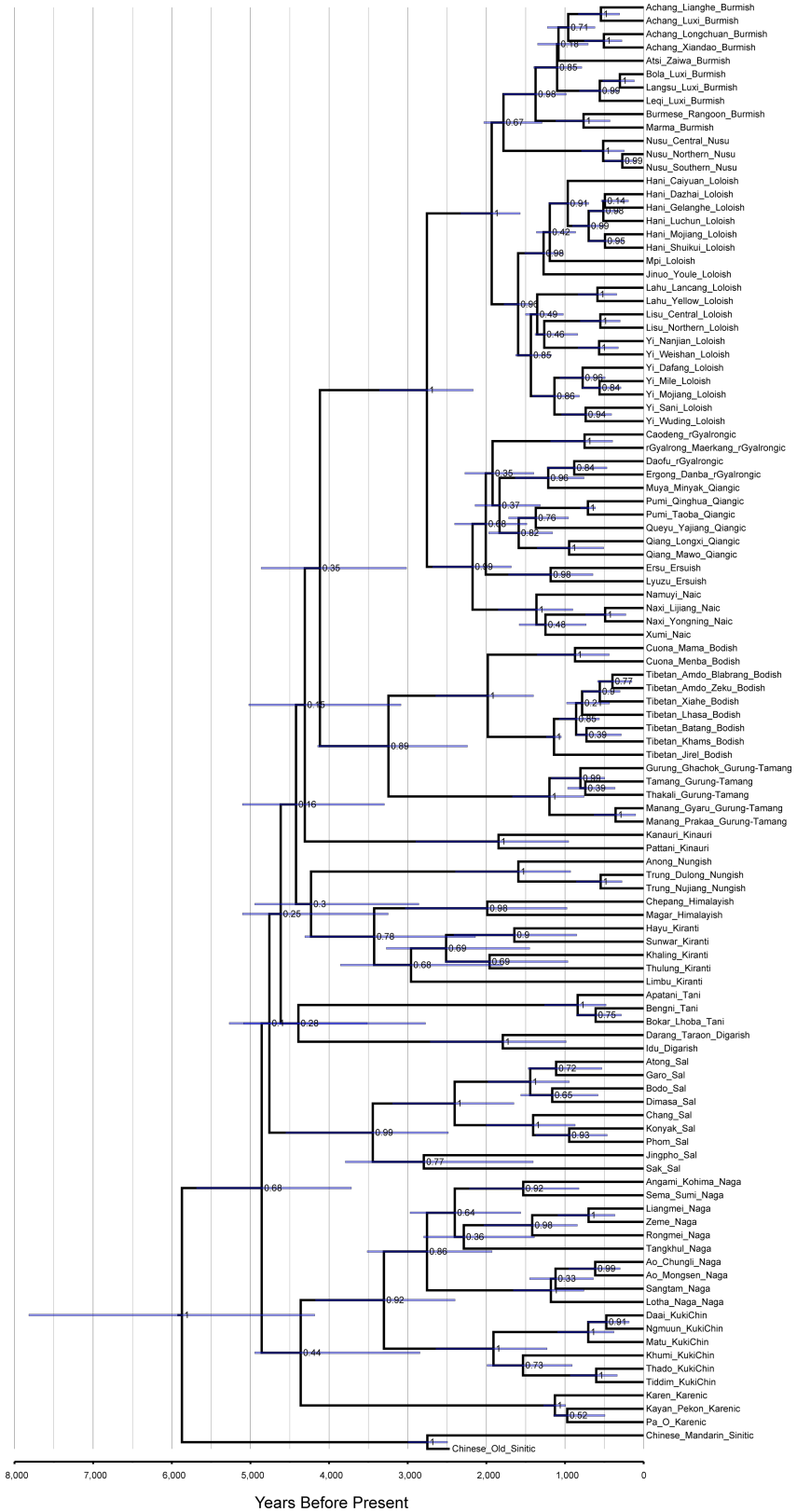
31. Dryer, M. S. in *The Sino-Tibetan Languages* (eds Thurgood, G & LaPolla, R. J.) 43–55 (Routledge, London, 2003).
32. Hammarström, H., Forkel, R. & Haspelmath, M. *Glottolog 3.1* http://glottolog.org (Max Planck Institute for the Science of Human History, Jena, 2017).
33. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).
34. Dunn, M., Kruspe, N. & Burenhult, N. Time and place in the prehistory of the Aslian languages. *Hum. Biol.* **85**, 383–400 (2013).
35. Rambaut, A. et al. Tracer v.1. 6 http://beast.bio.ed.ac.uk (2014).
36. Cleveland, W. S. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* **35**, 54 (1981).



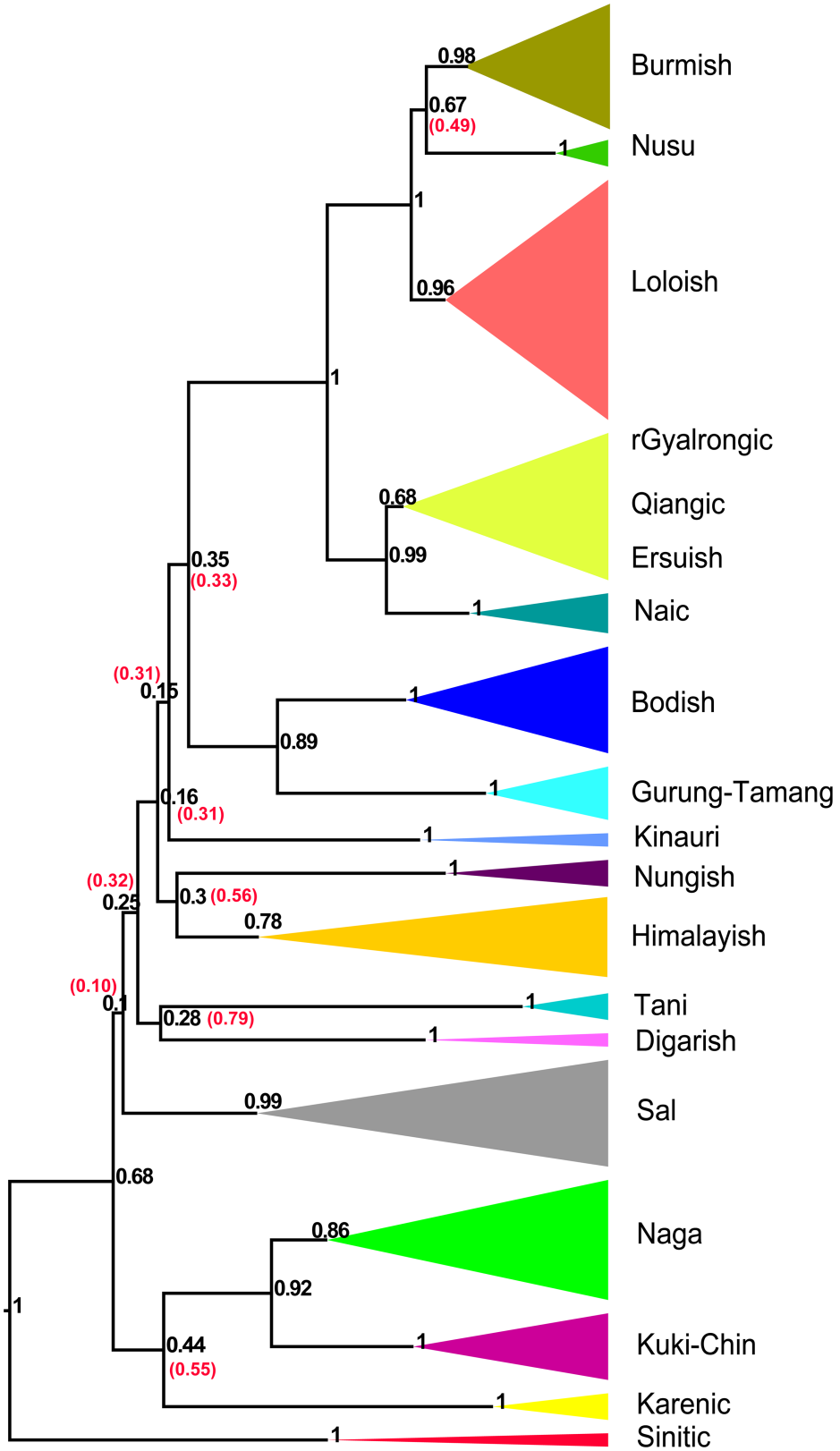
Extended Data Fig. 1 | The geographical distribution of samples of 109 Sino-Tibetan languages in the East Asia. The colours show affiliations to linguistic clades. The map is based on vector map data from https://www.naturalearthdata.com.



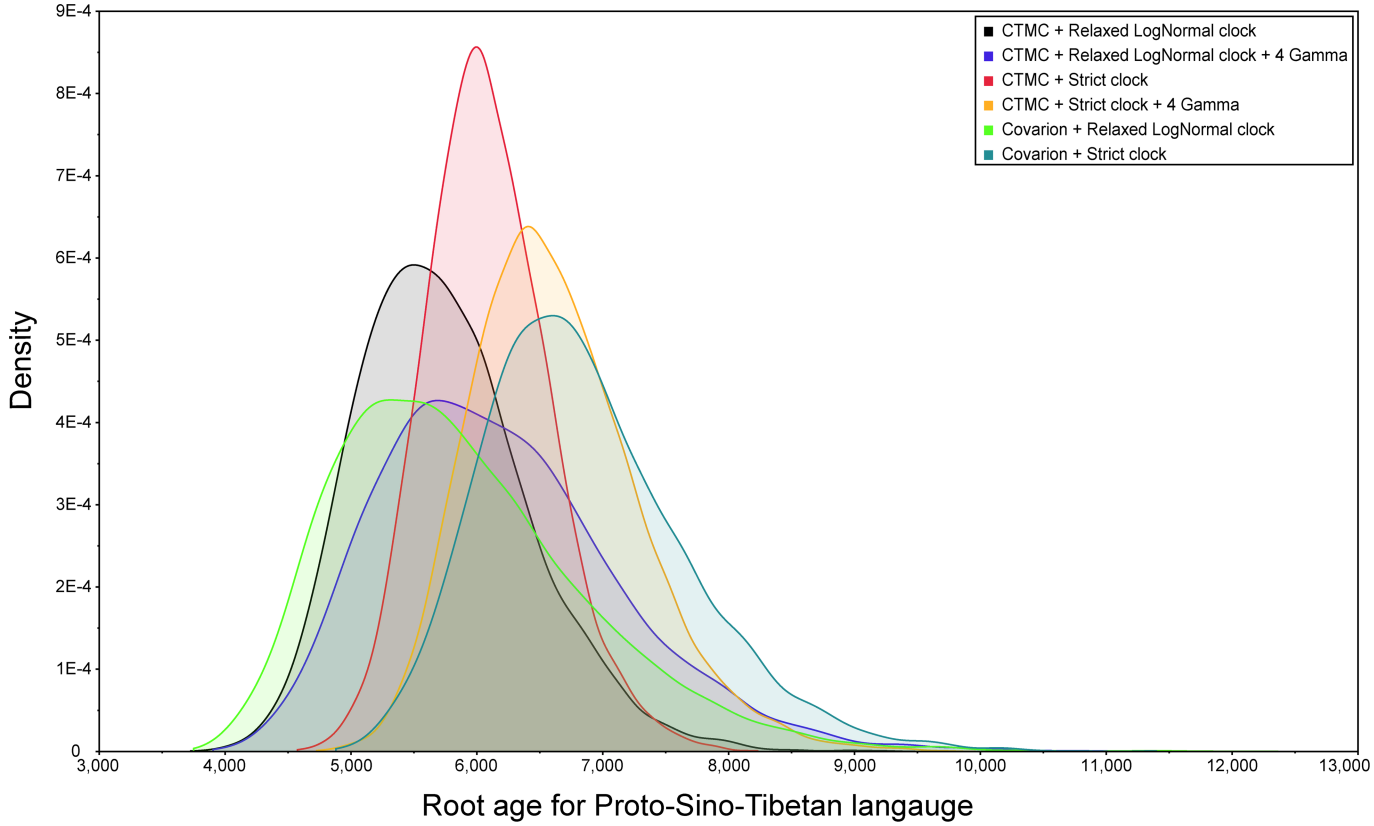
Extended Data Fig. 2 | The distributions of the likelihood values. The likelihood values for different combinations of mutation models, clock models and rate heterogeneity models. Each combination was run for 50 million generations, sampling every 5,000 generations. The first 10% of the iterations were treated as burn-in.



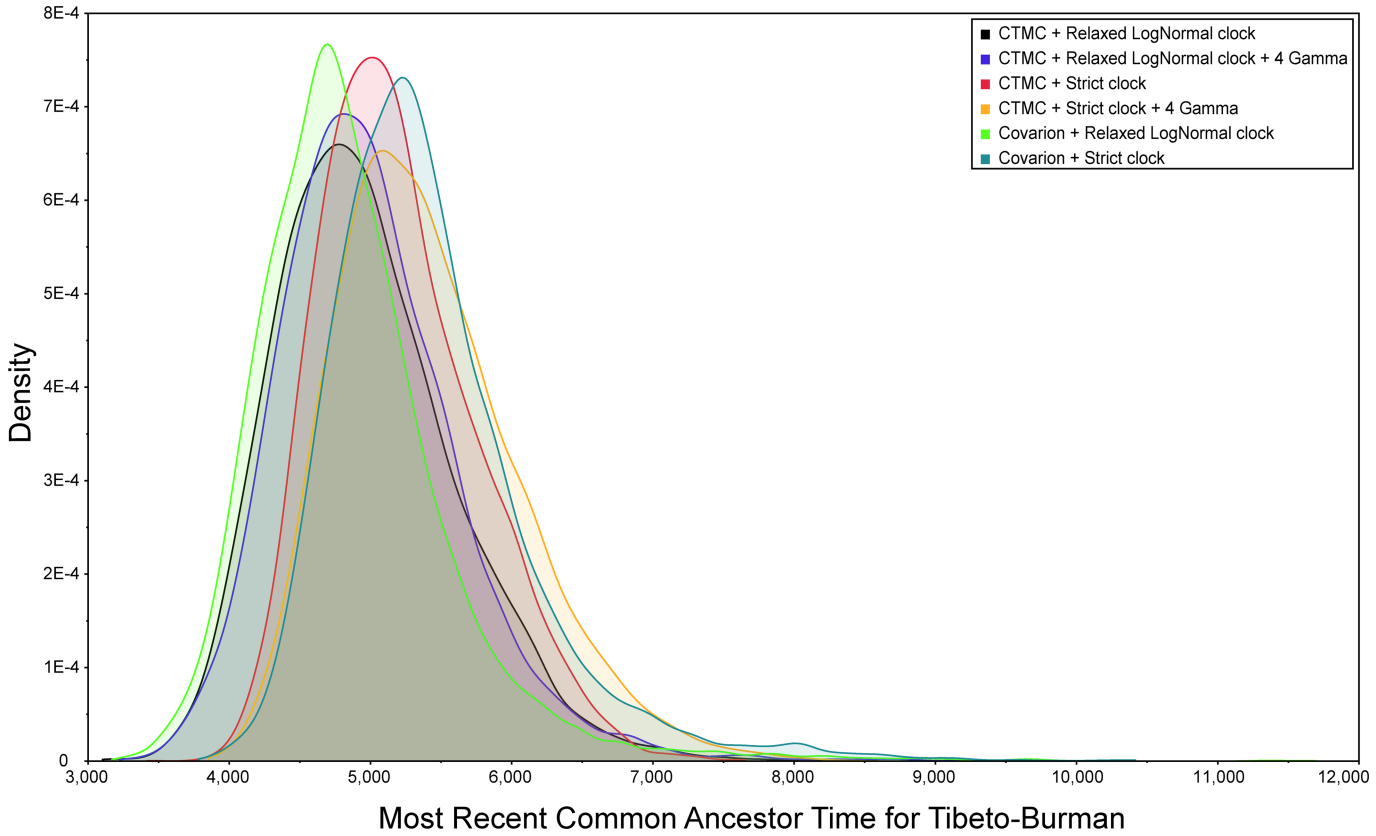
Extended Data Fig. 3 | The maximum clade credibility tree of 109 Sino-Tibetan languages with node bars of ages and posterior probability values. The iterations for tree reconstruction were set to 50 million generations, sampling trees every 5,000 generations and resulting in a sample of 10,000 trees. The first 10% of the iterations were treated as burn-in. The maximum clade credibility tree was established from 9,000 trees.



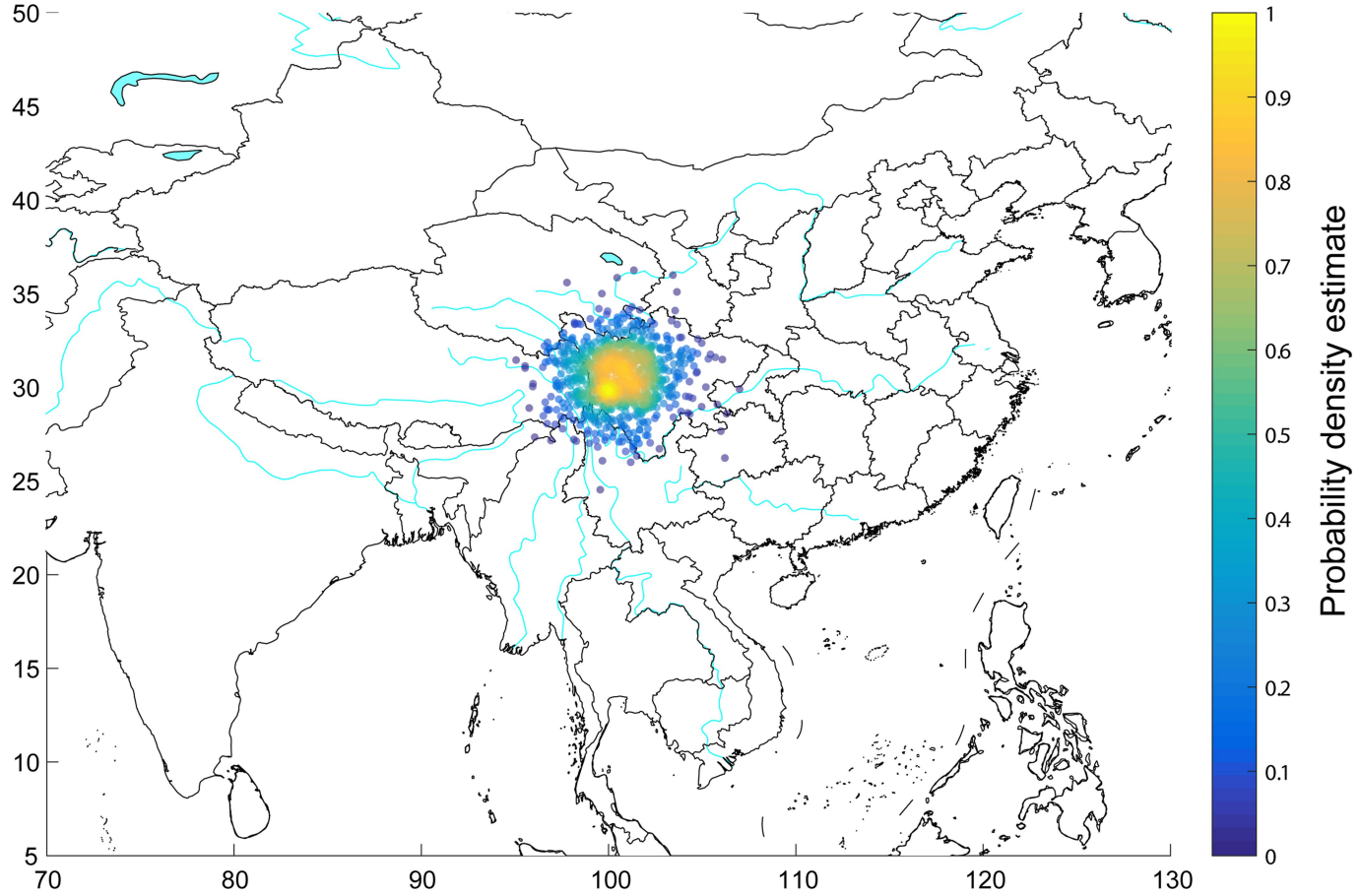
Extended Data Fig. 4 | The results of four-point analysis for the maximum clade credibility tree of 109 Sino-Tibetan languages. The maximum clade credibility tree was constructed on the basis of the best-fitting model combination, which was run for 50 million generations, sampling every 5,000 generations and treating the first 10% of the iterations as burn-in. The 109 Sino-Tibetan languages were grouped into major linguistic clades, which are labelled with the same colours and names as those shown in Fig. 1. The black numbers show the posterior probability values supporting the descendent clade. The red numbers in the parentheses show the reliability values on the internal nodes, calculated from four-point analysis.



Extended Data Fig. 5 | The distribution of the root time of 109 Sino-Tibetan languages. The root time estimates for the 109 Sino-Tibetan languages with different combinations of mutation models, clock models and rate heterogeneity models. Each combination was run for 50 million generations, sampling every 5,000 generations. The first 10% of the iterations were treated as burn-in.



Extended Data Fig. 6 | The distribution of the root time of 107 Tibeto-Burman languages. The root time estimates for the 107 Tibeto-Burman languages (that is, excluding 'Chinese Mandarin' and 'Chinese Old' from the Sino-Tibetan sample set) with different combinations of mutation models, clock models and rate heterogeneity models. Each combination was run for 50 million generations, sampling every 5,000 generations. The first 10% of the iterations were treated as burn-in.



Extended Data Fig. 7 | The geographical plot of *Urheimat* inference of 109 Sino-Tibetan languages. The probability density estimates for the original homeland of the Sino-Tibetan languages via the phylogeographical approach, implemented in BayesTraits package. The iterations in BayesTraits were set to 1,000,000. The sample period was set to 1,000. The first 25% of the iterations were treated as burn-in. The map is based on vector map data from https://www.naturalearthdata.com.

nature research

Corresponding author(s): Li Jin

Last updated by author(s): Mar 14, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No specific software was used in data collection.
Data analysis	The BEAST v2.4.8 software can be downloaded from https://www.beast2.org/. The Babel package is a BEAST package for performing linguistic analysis (URL: https://github.com/rbouckaert/Babel). The BayesTrait program can be downloaded from http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html. The SPLITSTREE v4.16.6 was used which can be downloaded from http://www.splitstree.org/ The Matlab codes used in this study are available in the supplementary files ('Matlab codes for estimation of ST evolutionary tempo.zip' and 'Matlab codes for Four-Point analysis.zip').

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the findings of this study are available in the supplementary information files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used 109 Sino-Tibetan language samples with 949 binary-coded lexical Root-meanings in this study. To the extent that it is possible to increase this sample size, we do not expect this will improve the accuracy or precision of our date and geographic homeland estimates, because our sample already represents all the available languages in the database.
Data exclusions	The languages and lexical root-meanings used in this study came from the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) project (http://stedt.berkeley.edu/). 1. We manually checked the word entries with a STEDT-labelled etymon and a meaning within the range of Swadesh 100 words, and excluded the incomplete etymon annotation and the inexhaustive investigation of a language. 2. We filtered only those languages with at least 90 lexical meanings of Swadesh 100 word-list recorded and 30 – 120 Root-Meanings (RMs). The reason for this filtering process was that if too few meanings or RMs existed for a language, we considered the language was either not well investigated in the literatures or not well labelled for etyma in the database. And too many RMs might indicate that the language was excessively recorded. 3. Furthermore, we removed those language that were possibly duplicated, or commonly believed to have suffered from strong lateral transfers (or horizontal language influence) among the Sino-Tibetan language family, or could not be located to a certain time point. 4. Finally, we retained 109 ST language samples with 949 binary-coded lexical RMs from STEDT database.
Replication	Findings were reproduced under a range of different model assumptions as reported in the manuscript.
Randomization	This is not relevant to our study because it does not include an experimental treatment.
Blinding	This is not relevant to our study because it does not include an experimental treatment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging