

In the format provided by the authors and unedited.

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*}

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

*e-mail: lijin@fudan.edu.cn

Supplementary Information

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

Menghan Zhang, Shi Yan, Wuyun Pan, Li Jin*

1. Supplementary Methods

- 1.1 The processing of lexical root-meaning data
- 1.2 Model settings in BEAST program
- 1.3 The reliability evaluation of the ST phylogenetic structure
- 1.4 Geographic homeland inference (*Urheimat* inference)
- 1.5 The data descriptions of archaeological sites

2. Supplementary Discussion

- 2.1 Model Selection
- 2.2 The phylogenetic topology of the Sino-Tibetan languages
- 2.3 The phylogenetic reliabilities of controversial branches in ST languages
- 2.4 The divergence time and tempo of the Sino-Tibetan languages
- 2.5 The Farming/Language Dispersal Hypothesis in the Sino-Tibetan languages
- 2.6 The homoplasy in the ST language phylogeny
- 2.7 *Urheimat* inference of the Sino-Tibetan languages

3. References in Supplementary Information:

Supplementary Table 1-4

1. Supplementary Methods

1.1 The processing of lexical root-meaning data

The lexical root-meanings used in this study came from the *Sino-Tibetan Etymological Dictionary and Thesaurus* (STEDT) project¹, which was developed by a number of experienced historical linguists led by James A. Matisoff over a 30-year period (URL: <http://stedt.berkeley.edu/>). For hundreds of Tibeto-Burman (TB) and Sinitic languages (and dialects), this project collected copious lexical data, annotated etymologies, reconstructed proto-forms, and inferred paths of semantic development. The data involve several lexical categories: *The Body*; *Animals and Animal Verbs*; *Natural Objects, Plants, Foods*; *Kinship Terms, Ethnonyms, Social Roles*; *Culture, Artifacts, Religion*; *Verbs of Motion, of Manipulation, and of Production*; *Adjectival Verbs*; *Abstract Nouns and Verbs*; *Psychological Verbs*; *Verbs of Utterance*; *Shape, Size, Color, Measure, Number, Time, Space*; and *Grammatical Words*. These cover the traditional items of basic vocabulary such as the Swadesh 100-word list². Based on these data, which originated from vast language surveys, experienced historical linguists reconstructed the roots of the proto-lexicon. Considering the semantic changes in the Sino-Tibetan (ST) languages, STEDT produced path diagrams that represent the paths of semantic association and change undergone by etyma. All fields are established by comparative and/or internal synchronic evidence in historical linguistics, see details in Matisoff's publication^{1,3}. Full references were given in the STEDT database. As an outcome, there is an ST lexical database available online, in which 4,024 etyma have been identified and labelled up to now.

In STEDT, when words in different languages are descended from a reconstructed proto-form of one ancestral language, these words are regarded belonging to an etymon (or cognate). For phylogenetic reconstruction, linguists tend to use those words that are resistant to borrowing. However, it is difficult to exclude that a common etymon in a group of language is due to borrowing, especially for the early borrowed words, when a systematic phonetic shift model has not been fully built for ST languages.

To minimize the word lateral transfers, in this study we chose only the words with meaning inside the Swadesh 100-word list, since they are relatively resistant to borrowing². Due to semantic drift, each etymon could correspond to several meanings in the ST languages, so we manually checked the word entries with a STEDT-labelled etymon and a meaning within the range of Swadesh 100 words. For example, the etymon #231 **krak* may have two meanings “blood” and “red”, we will

take #231- “blood” and #231- “red” as two separate root-meaning (RM) traits. An RM in this study has the similar meaning to a “cognate set” in the Indo-European paper ^{4,5}, rather than only considering an etymon, due to the fact that RM provides more information inside a language family. Also, due to semantic shift, different meanings of an etymon may be inside or outside the set of Swadesh wordlist, so that it is easier to define whether an RM, rather than an etymon, belongs to a wordlist. However, although the etyma were labelled for cognates in an STEDT entry, for a complex word, the meaning of a root was not unambiguously labelled. For example, an entry of a meaning “we two” with the form *ŋu:-ñi*, each root (syllable) of the word was labelled with an etymon respectively, however, it is not naturally clear which etymon correspond to “we” and which to “two”. Therefore, an additional manual labelling on basis of our linguistic knowledge was needed to correctly identify each RM. In total, we identified 949 RMs.

If an RM appeared in a language, we labelled this RM-language as “1” in the table. However, due to either inexhaustive investigation of a language or incomplete etymon annotation, an RM not shown in a language does not definitely mean an absence. We processed the data as following: if a meaning inside the Swadesh 100-word list was ever recorded in an entry of a language, and the RM was not recorded, we considered that the RM does not exist in this language and we would label a “0” in the table; but when a meaning was totally not recorded for a language in the database, since we thought that any of the 100 word meanings should be essential for a language, we would deem that the language was not fully investigated or annotated for this meaning and leave all the RMs with this meaning a “?” for this language in the table (Supplementary Table 1).

In order to make sure that all the languages were comparable to each other, we filtered only those languages with at least 90 lexical meanings of Swadesh 100-word list recorded (no matter whether an RM exists) and 30 – 120 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 (in our case, those with >120 RMs included Written Tibetan and Written Burmese), which would make the language not comparable to the others, especially when we calculate the divergence time and mutation rate.

Under the selection criteria mentioned above, we obtained 148 languages. We further removed those language that were possibly duplicated, or commonly believed to have suffered from strong

lateral transfers (or horizontal language influence) among the ST language family (such as Tujia, Bai, or Tangut-Xixia), or could not be located to a certain time point (like Middle Chinese). Finally, we retained 109 ST language samples with 949 binary-coded lexical RMs for further phylogenetic analyses. The language sample size reached one-fourth of all known ST languages. Note that these languages covered not only well-known linguistic branches such as Lolo-Burmese, but also small subgroups such as Tani and Digarish (see Extended Data Fig. 1).

1.2 Model settings in BEAST program

To reconstruct the phylogeny and estimate divergence time, we used two different clock models, two different mutation models, and two site heterogeneity models, as summarized in Supplementary Table 3. The first clock model was a strict clock which assumed a constant rate of RM change over time; the second clock model was a relaxed clock with log-normal distribution that allowed rates of RM change to vary freely across the branches, with the parameter estimation sampled from a log-normal distribution. The first RM change model we used here was a simple continuous time Markov chain (CTMC)^{6,7}, which allowed each RM to be gained or lost at the same rate; the second was the covarion model^{8,9}, which assumes that each RM could undergo with transition an arbitrary number of times between gain and loss, and also transition between fast evolution and slow evolution⁵. This model agrees with linguists' intuition that different lexicon have their own evolutionary process. According to a previous study¹⁰, we did not select a stochastic Dollo model^{11,12} because its assumption allows no semantic change, which does not intrinsically agree with the situation in STEDT database. In addition, gamma rate heterogeneity with four rate categories was investigated for the CTMC model.

1.3 The reliability evaluation of the ST phylogenetic structure

For any sub-tree structure of interest, we performed a four-point analysis to evaluate its reliability, given the inferred phylogenetic tree. This analysis reports the probability of obtaining a sub-tree in an inferred tree, and its rationale is similar with the previous studies, that is, to evaluate whether a two-to-two partition was supported by the data^{13,14}. The basic procedure of the analysis is to: (1) Select the sub-tree of interest including three language subgroups (named A, B, C), and a distantly related language as an outgroup (D); (2) Randomly choose one language from each

subgroup (named a, b, c), together with the out-group language (d), and construct a phylogenetic tree of these four languages using an informative trait, which has only three possibilities of tree topologies: (ab)-(cd), (ac)-(bd), and (ad)-(bc). An informative trait for the topology only refers to an RM with two '0's and two '1's in the four languages. (3) Repeat constructing the tree with all informative traits, and count the occurrence frequencies of each tree topology for these four languages. The most frequently occurring tree topology is ascertained as the optimal one. (4) Re-sample the languages from three language subgroups, and repeat the steps (2)-(3). (5) Calculate the percentage of optimal tree topologies with the maximum possibility in all combinations of languages exhaustively selected from these language subgroups. The percentage value refers to the reliability of the sub-tree topology for the language subgroup A, B, C and outgroup D. To compare the null hypothesis that the frequencies of each topology are equal (0.33 for each kind of sub-tree topology), the value of occurrence frequency of the sub-tree topology shows the weight of evidence supporting the sub-tree topology. A value between 0 and 0.33 indicates no support; one between 0.33 and 0.5, weak support; one between 0.5 and 0.7, moderate support; and over 0.7 strong support. The related Matlab codes were available in the Supplementary Data -'Matlab codes for Four-Point analysis.zip'.

1.4 Geographic homeland inference (*Urheimat* inference)

To infer the geographic origin of the ST language expansion/divergence, we applied the Monte Carlo Markov Chain (MCMC) approach using a geographical model, given a phylogenetic tree. The analysis was implemented in the *BayesTraits* package (V 3.0.1)¹⁵. According to the description in the manual, the geographic model in this package is able to model both longitude and latitude changes using Brownian motion within 3-D Cartesian coordinates system. We applied the default settings in the *BayesTraits* package using geographic coordinates of 109 ST language samples for one million iterations (Iterations = 1,000,000; Sample Period = 1,000). The first 25% of the iterations were burnt in. The average longitude and latitude values sampled via the MCMC approach refer to the estimated centre of the geographic origin of the ST expansion.

1.5 The data description of archaeological sites from Hosner *et al.*

The time ranges and geographic locations of the archaeological sites were from the study of

Hosner *et al.*¹⁶ (URL: <https://doi.pangaea.de/10.1594/PANGAEA.860072>). Hosner's data contained a total of 51,074 archaeological sites from the early Neolithic to the early Iron Age (about 10,000 – 2,500 years BP), with a spatial distribution covering most regions of China. Each site entry contained the cultural type (e.g. Yangshao culture or Majiayao culture), cultural time range (max dated time, min time, average time), and geographic location (province, longitude and latitude). Hosner *et al.* assembled the data of archaeological sites extracted from three major systematic archaeological surveys covering the whole country which were operated by the Chinese government in 1956, 1981 and 2007¹⁷. According to the description from Hosner *et al.*, “During these surveys, sites of all prehistoric and historic periods were registered and documented by teams of archaeologists representing central, provincial and city archaeological institutions and state bureaus of cultural relics of all administrative levels.”, such as the State Administration of Cultural Heritage (Chinese: Guojia Wenwuju) (see details in Hosner *et al.*¹⁶ and Wagner *et al.*¹⁷).

2. Supplementary Discussion

Summary of our study

There are challenges for the application of the historical linguistic comparative methods to systematically investigating the ST languages, due to that the time depth of the ST language is far beyond all the historical accounts and linguistic documentation. Therefore, it is very difficult to reconstruct the whole evolutionary history of the ST languages using the traditional comparative approaches. With the help of Bayesian phylogenetic methods, we examined the two competing hypotheses of the ST origin and dispersal, northern-origin hypothesis (NOH) and southwestern-origin hypothesis (SOH). We started with the phylogenetic reconstruction of the ST languages, which moderately supported the monophyly of the TB languages. The evolutionary tempo of the phylogenetic tree structure was also noted to identify the process of linguistic diversification in the ST languages. The reliable topology of the phylogenetic tree is the basis for further estimation of the time of ST language diversification. Therefore, we adopted four-point analysis to evaluate the reliability of controversial or weakly supported linguistic relationships in the inferred tree topology. Furthermore, we estimated that the initial divergence of the Proto-ST language into Sinitic and TB occurred at ~ 5,900 years BP, and identified the beginning time of the internal diversification of Proto-TB at ~ 4,700 years BP. According to the divergence time, we thus advocated the NOH that the

initial divergence of the ST languages should be connected to the two important Neolithic cultures, Yangshao and Majiayao, which were located in the upper and middle Yellow River basin. Our results of divergence time estimation for language subgroups were compatible with traditional linguistic evidence^{18,19} and genetic evidence²⁰. For example, historical linguists hypothesized that the divergence between Tibetan and Chinese could be beyond 3,000 years based on historical comparisons of numerals between Tibetan and Old Chinese¹⁸.

2.1 Model Selection

We reconstructed the ST language phylogeny without setting any monophyletic constraints (See details in Supplementary Data -‘BinaryCovariation.RelaxedClock.xml’). When the reconstruction is on the basis of the incorrect language classification, these settings have extremely destructive effects on phylogenetic reconstruction and divergence time estimation based on given phylogeny. In contrast, we here performed the phylogenetic analyses without any monophyletic constraints as priors, in order to avoid bias during the phylogenetic reconstruction.

The divergence times were estimated by constraining the ages of several internal nodes (for divergence) and tip date (for an ancient language) in accordance with historically attested events (See Supplementary Table 2). After comparing different combinations of lexical substitution models with various rate heterogeneities and clock models (see Extended Data Fig. 2), we identified the combination of covariation model with log-normal relaxed clock as the best-fitting one, which was with the highest marginal likelihood estimates in all models compared (see Extended Data Fig. 2 and Supplementary Table 3).

2.2 The phylogenetic topology of the ST languages

Based on the best-fitting combination of the covariation model with the log-normal relaxed clock model, the Bayesian phylogenetic tree of the 109 ST languages showed two distinct branches, the Sinitic and the TB, moderately supporting a Sinitic-TB dichotomy, with a posterior probability value of 0.68 (see Extended Data Figs. 3 and 4).

On the TB side, the topology illustrated several monophyletic clades in line with known linguistic classifications. Lolo-Burmese and Na-Qiangic languages formed an unambiguous monophyly, which was further divided into two monophyletic clades. Burmish, Loloish and Nusu

languages formed a monophyletic clade with a posterior probability of 1 (see Fig. 1 and Extended Data Fig. 4). This topology supported the traditionally acknowledged Lolo-Burmese branch in linguistics. Similarly, Qiangic, rGyalrongic, Ersuish and Naic languages constituted another monophyletic clade (named as Na-Qiangic, posterior probability = 0.99). Qiangic (including Muya) and rGyalrongic (including Daofu and Ergong) languages together formed an acceptable monophyly with a posterior probability of 0.68, even though these two language groups themselves were not monophyletic. Moreover, other clades in the ST phylogeny were consistent with known linguistic classifications such as Bodic (including Bodish and Gurung-Tamang, posterior probability = 0.89), Kuki-Chin-Naga (0.82), Sal (Bodo-Garo-Jingpo, 1). In contrast, the rest of TB languages constitute different independent monophyletic groups, e.g. Karenic, Tani, Digarish and Nungish, each with high posterior probabilities (>0.99). However, since the early diverged clades of TB were not strongly supported and showed instability in the trees based on different calculation models, we were not yet able to draw a clear conclusion about the early differentiation of these TB clades²¹.

2.3 The phylogenetic reliabilities of controversial branches in ST languages

To evaluate the reliability of several controversial ST subgroupings in the phylogeny, we carried out four-point analysis (see Supplementary Section 1.3). For example, the position of the Naic languages within the ST language is controversial²². Some linguistic studies classified the Naic languages as a sub-branch of the Lolo-Burmish languages^{23,24}, while others argued that it could be part of an undetermined Na-Qiangic clade²⁵. To address the question of the position of the Naic languages, we applied four-point analysis for the Naic, Lolo-Burmish (including Loloish, Burmish, Nusu) and Qiangic languages using Chinese as an outgroup. Based on the probabilities of occurrence of the tree topology (= 0.83), the result strongly supported that the Qiangic and Naic languages formed a monophyly. Thus, the Naic languages should be more closely related to the Qiangic languages than to the Lolo-Burmese, which was consistent with our Bayesian phylogenetic maximum clade credibility tree in Fig. 1 and Extended Data Fig. 4.

Likewise, the relationship between Tani and Digarish languages was also under debate. Sun T.-S. stated that these two languages were closest related with each other²⁶. In our study, the result of four-point analysis supported this statement (the probabilities of occurrence = 0.79).

Nusu languages were traditionally classified within the Loloish branch. However, Fu argued that the Nusu languages could be an inter-language in the Lolo-Burmish branch, but not explicitly grouped into either Burmish or Loloish languages (excluding Nusu)^{27,28}. We examined the position of the Nusu languages in the Lolo-Burmish branch, and found the probability of the grouping of the Nusu and Burmish languages was 0.49, while that of the Nusu and Loloish languages was 0.45. This finding agreed with Fu's argument that the Nusu languages seemed correlated with both Burmish and Loloish languages.

In addition, we evaluated the reliability values of the other groupings, that were not well supported (with posterior probability value < 0.5), and the results were shown in Extended Data Fig. 4.

2.4 The divergence time and tempo of the ST languages

Using several time calibrations of internal nodes and tip dates (see Supplementary Table 2), the divergence times of all nodes on the ST phylogenetic tree were estimated. The average time estimation for the divergence of the Sinitic and TB languages was 5,871 years BP (95% Highest posterior density (HPD): 4,183 – 7,817 years BP). In all the examined models, their estimations including average root ages and 95% HDP ranges were below 9,000 years (see Extended Data Fig. 5). Therefore, we could not support the SOH but advocate the NOH. Likewise, the divergence time estimation of the TB languages was 4,665 years BP (95% HPD: 3,716 – 5,688 years BP). The time estimations of the TB divergence in different model combinations were consistent with each other (see Extended Data Fig. 6).

2.5 The Farming/Language Dispersal Hypothesis in the ST languages

In our opinion, the ST language dispersal could be accompanied with the agricultural expansion. During the late Neolithic period (~6,000 BP), rice agriculture was predominantly distributed in the Yangtze basin, whereas the distribution of millet (foxtail millet, *Setaria italica* and broomcorn millet, *Panicum miliaceum*) agriculture was mainly in the northern China, covering the geographic scopes of the Yangshao and Majiayao cultures²⁹. After ~5,000 BP, the geographic distribution of the millet-cultivating agriculture apparently expanded^{29,30}, especially to the west and south, but not to the north, as the climatic condition made it inadequate for millet cultivation. For example, archaeological

evidence showed that the cultivation in southwest China was predominantly of millet associated with the Majiayao cultural expansion³¹. The migration routes of the millet-cultivating agriculture coincide with ST language divergence and dispersal, especially with Proto-TB language^{32,33}.

For example, the millet cultivation had spread along the eastern front of Tibetan Plateau into Sichuan Province before 5,000 years BP, which was earlier than the westward dispersal of rice cultivation from the middle Yangtze basin³¹. In our study, we considered the ancient people, who carried millet cultivation into Sichuan from the northwest China, spoke an ancient ST language. And then, the rice cultivation expanded from the middle Yangtze River basin into Sichuan Basin (not before 5,000 – 4,500 years BP) and Yunnan (not before 4,000 years BP)³⁴. Accordingly, we could not exclude the possible scenario that due to the expansion of rice cultivation, the ancient ST-speaking millet farmers could borrow the techniques of rice cultivation together with the vocabulary of the rice farmers. The farmers could speak Hmong-Mien, Austroasiatic etc. In this situation, it is challenging to identify lexical borrowings from shared vertical language inheritance during the evolution of ST languages³⁵.

2.6 The homoplasy in the ST language phylogeny

To measure the degree of homoplasy or horizontal influence (e.g. lexical borrowing) which is against the tree-like topology, we calculated the delta score³⁶ and Q-residual value³⁷ for each language sample and their average values for the entire dataset. We performed the calculation in SPLITSTREE v4.16.6 (<http://www.splitstree.org/>) using Gene Content distance³⁸.

Both the average delta score and the Q-residual value provide a quantification for the potentially conflicting signals in the lexical data (Supplementary Table 4). These two values of the ST languages were higher than those for the Indo-European and Austronesian languages, but lower than those for Polynesian³⁷ and Cariban languages³⁹. Higher values depict higher degree of reticulation, therefore the ST family did not fit as good a tree topology as the Indo-European family. Such reticulation of the ST languages could be resulted from the complex history of language contact and areal diffusion of linguistic features.

Population migrations and interactions among the ST languages were complex over a long period of time^{35,40}, such as between Chinese and Tibetan^{41,42}. As a consequence, contact-induced convergence in the ST languages such as lexical borrowing could occur at the early history of the ST

languages and continue till the present. These contacts occurred not only among the subgroups of the ST languages (i.e. the strong influence of Chinese to Bai, Tujia, and Tangut languages), but also between the ST languages and the surrounding language families, such as Austronesian⁴³, Tai-Kadai and Hmong-Mien^{44,45} during the ancient period.

2.7 *Urheimat* inference of the ST languages

To infer the original homeland of the ST languages, we modelled the latitude and longitude values of the 109 language samples in 3-D Cartesian coordinates system using Brownian motion given the language phylogenetic tree (See Supplementary Section 1.4). This model allows the latitudes and longitudes to change continuously through the time along the branches of the given tree and to reconstruct the ancestral locations at the internal nodes on the tree. As shown in Extended Data Fig. 7, our results suggested that the original homeland of the ST languages could be located in the eastern edge of Tibetan Plateau in China (latitude = 30.57 ± 1.57 N, longitude = 100.57 ± 1.83 E), and the estimated region covered the so-called Tibetan-Yi Ethnic Corridor^{32,46}. Although the ST homeland was not located in the Upper Yellow River basin or the geographic centre of the Majiayao Neolithic culture in Gansu and Qinghai provinces, the density estimation was also not compatible with SOH of Northeast India, but seemed compatible with SOH of southwest Sichuan.

According to the comments from Dunn *et al.*⁴⁷, the inferred homeland based on the phylogeographic model is most likely near the centre of phylogenetic diversity. In our study, the geographic distribution of the language samples was uneven, but the most language diversity appeared in southwest Sichuan and west Yunnan provinces. In addition, we argued that two important prerequisites should be satisfied by this *Urheimat* inference: (1) no strong regional language extinction occurred (violation of which could cause strong bias in language spatial diversity), and (2) no multiple languages migrated in the same direction (violation of which is against the assumption of Brownian motion).

However, none of the two were satisfied in the ST language history: (1) we clearly know from the historical records, oral legends, and the current status, that the language diversity in the Han Chinese region was heavily extinguished, and (2) a continuous and unidirectional migration (north to south) happened along the Tibetan-Yi Corridor at least throughout the recent 5,000 years. Therefore, the homeland based on the phylogeographic inference seemed strongly biased toward the southwest

direction. Although the SOH of southwest Sichuan Province could not be rejected by phylogeographic inference in our study, Stevens and Fuller²⁹ showed their strong attitude against the SOH, based on the recent archaeological evidence⁴⁸.

3. References in Supplementary Information

- 1 Matisoff, J. A. *Sino-Tibetan etymological dictionary and thesaurus (STEDT)*, <<http://stedt.berkeley.edu/search>> (2015).
- 2 Swadesh, M. Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Ling.* **21**, 121-137 (1955).
- 3 Matisoff, J. A. *The Tibeto-Burman reproductive system: Toward an etymological thesaurus*. Vol. 139 (Univ of California Press, 2008).
- 4 Bouckaert, R. *et al.* Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957-960 (2012).
- 5 Chang, W., Cathcart, C., Hall, D. & Garrett, A. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**, 194-244 (2015).
- 6 Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479-483 (2009).
- 7 Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57-86 (1986).
- 8 Yang, Z.-H. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396-1401 (1993).
- 9 Penny, D., McComish, B. J., Charleston, M. A. & Hendy, M. D. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* **53**, 711-723 (2001).
- 10 Kolipakam, V. *et al.* A Bayesian phylogenetic study of the Dravidian language family. *Royal Soc. Open Sci.* **5**, 171504 (2018).
- 11 Nicholls, G. K. & Gray, R. D. in *Phylogenetic methods and the prehistory of languages* (eds Peter Forster & Colin Renfrew) 161-171 (McDonald Institute for Archaeological Research, 2006).
- 12 Alekseyenko, A. V., Lee, C. J. & Suchard, M. A. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.* **57**, 772-784 (2008).
- 13 Iwabe, N., Kuma, K.-i., Hasegawa, M., Osawa, S. & Miyata, T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355-9359 (1989).
- 14 Katsura, Y., Stanley Jr, C. E., Kumar, S. & Nei, M. The reliability and stability of an inferred phylogenetic tree from empirical Data. *Mol. Biol. Evol.* **34**, 718-723 (2017).
- 15 Pagel, M. & Meade, A. *BayesTraits Version 3*, <<http://www.evolution.rdg.ac.uk/BayesTraits.html>> (2017).
- 16 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. *The Holocene* **26**, 1576-1593, doi:10.1177/0959683616641743 (2016).
- 17 Wagner, M. *et al.* Mapping of the spatial and temporal distribution of archaeological sites of northern China during the Neolithic and Bronze Age. *Quat. Int.* **290**, 344-357, doi:10.1016/j.quaint.2012.06.039 (2013).
- 18 Wang, W. S.-Y. *The ancestry of the Chinese language*. (Project on Linguistic Analysis, University of California, 1995).

- 19 Sagart, L., Blench, R. & Sanchez-Mazas, A. in *The peopling of East Asia: putting together archaeology, linguistics and genetics* (eds Laurent Sagart, Roger Blench, & Alicia Sanchez-Mazas) 1-14 (Routledge, 2005).
- 20 Wang, L.-X. *et al.* Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol. Genet. Genomics*, 1-8 (2018).
- 21 Van Driem, G. Languages of the Himalayas: An ethnolinguistic handbook of the greater Himalayan region with an introduction to the symbiotic theory of language (2 vols.). *Leiden: Brill* (2001).
- 22 Thurgood, G. & LaPolla, R. J. *The sino-tibetan languages*. Vol. 3 (Psychology Press, 2003).
- 23 Lama, Z. Q.-F. *Subgrouping of Nisoic (Yi) languages: A study from the perspectives of shared innovation and phylogenetic estimation* Doctor of Philosophy thesis, The University of Texas at Arlington, (2012).
- 24 Sun, H.-K. Nàxīyǔ zài Zàng-Miǎn yǔzú yǔyán zhōng de lìshǐ dìwèi [The position of Naxi among Tibeto-Burman languages]. *Yǔyán Yánjiù [Studies in Languages and Linguistics]*, 90-99 (2001).
- 25 Jacques, G. & Michaud, A. The reconstruction of Proto-Naish: a preliminary study of the historical phonology of highly eroded Sino-Tibetan languages. *Diachronica* **28** (2011).
- 26 Sun, T.-S. J. *A Historical-Comparative Study of the Tani (Mirish) Branch of Tibeto-Burman*. Doctor of Philosophy in Linguistics thesis, Berkeley: University of California, (1993).
- 27 Fu, A.-L. Nù yǔ (Nùsū) xīshǔ yánjiù - jiān lùn Miǎn-Yí yǔxì xīshǔ huàfēn wèntí [The study of the position of Nusu language: the debate on the classification of Lolo-Burmish]. *Yǔyán Yánjiù [Studies in Languages and Linguistics]* **16**, 131-149 (1989).
- 28 Fu, A.-L. Zài lùn Nù yǔ (Nùsū) de zhōngjièyǔ xìngzhì - jiān lùn zhōngjièyǔ zài Zàng-Miǎn yǔ yánjiù zhōng de dìwèi [Re-analysis of the interlingual characteristics of Nusu language: debate on the role of interlanguage in the study of Tibeto-Burman]. *Yǔyán Yánjiù [Studies in Languages and Linguistics]* **25**, 164-170 (1993).
- 29 Stevens, C. J. & Fuller, D. Q. The spread of agriculture in eastern Asia. *Language Dynamics and Change* **7**, 152-186, doi:10.1163/22105832-00702001 (2017).
- 30 Stevens, C. J. *et al.* Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *The Holocene* **26**, 1541-1555 (2016).
- 31 d'Alpoim Guedes, J. Millets, Rice, Social Complexity, and the Spread of Agriculture to the Chengdu Plain and Southwest China. *Rice* **4**, 104-113, doi:10.1007/s12284-011-9071-1 (2011).
- 32 Fei, X.-T. Guānyú wǒguó de mínzú shíbié wèntí [On the problem of distinguishing nationalities in China]. *Zhōngguó Shèhuì Kēxué [Social Sciences in China]* **1**, 158-174 (1980).
- 33 Hoffman, H. in *The Cambridge History of Early Inner Asia* (ed Denis Sinor) 371-399 (Cambridge University Press, 1990).
- 34 Fuller, D. Q. Pathways to Asian Civilizations: Tracing the Origins and Spread of Rice and Rice Cultures. *Rice* **4**, 78-92, doi:10.1007/s12284-011-9078-7 (2011).
- 35 LaPolla, R. J. in *Areal diffusion and genetic inheritance: Case studies in language change* (eds Robert M. W. Dixon & A. Y. Aikhenvald) 225-254 (Oxford University Press, 2001).
- 36 Holland, B. R., Huber, K. T., Dress, A. & Moulton, V. δ -plots: a tool for analyzing phylogenetic distance data. *Mol. Biol. Evol.* **19**, 2051-2059 (2002).
- 37 Gray, R. D., Bryant, D. & Greenhill, S. J. On the shape and fabric of human history. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3923-3933 (2010).
- 38 Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108 (1999).
- 39 Chousou-Polydouri, N., Birchall, J. & Meira, S. in *the Society for the Study of the Indigenous Languages of the Americas (SSILA)* (Austin, TX, 2015).
- 40 Handel, Z. J. *Old Chinese Medials and Their Sino-Tibetan Origins: A Comparative Study*. (Institute of Linguistics, Academia Sinica, 2009).
- 41 Wen, B. *et al.* Genetic evidence supports demic diffusion of Han culture. *Nature* **431**, 302-305 (2004).

- 42 Wen, B. *et al.* Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am. J. Hum. Genet.* **74**, 856-865 (2004).
- 43 Benedict, P. K. *Austro-Thai language and culture, with a glossary of roots.* (Human Relations Area Files, 1975).
- 44 Pan, W.-Y. in *The Ancestry of the Chinese Language* (ed William S.-Y. Wang) 113-144 (Berkeley: Project on Linguistic Analysis, 1995).
- 45 Zheng-Zhang, S.-F. in *The Ancestry of the Chinese Language* (ed William S.-Y. Wang) 269-282 (Berkeley: Project on Linguistic Analysis, 1995).
- 46 Sun, H.-K. Zài lùn xīnán mínzú zǒuláng dìqū de yǔyán jí qí xiāngguān wèntí [Re-analysis of the Languages of the Ethnic Corridor in Southwest China and Related Problems]. *Xīnán Mínzú Dàxué Xuébào: Rénwén Shèhuì Kēxué bǎn* [Journal of Southwest University for Nationalities: Humanity and Socieal Sciences] **6**, 29-39 (2013).
- 47 Dunn, M., Kruspe, N. & Burenhult, N. Time and place in the prehistory of the Aslian languages. *Hum. Biol.* **85**, 383-399 (2013).
- 48 Silva, F. *et al.* Modelling the Geographical Origin of Rice Cultivation in Asia Using the Rice Archaeological Database. *PLoS One* **10**, doi:10.1371/journal.pone.0137024 (2015).
- 49 Newton, M. A. & Raftery, A. E. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **56**, 3-48 (1994).
- 50 Suchard, M. A., Weiss, R. E. & Sinsheimer, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**, 1001-1013, doi:10.1093/oxfordjournals.molbev.a003872 (2001).

Supplementary Table 1. The information and lexical root-meanings of the ST languages in this paper selected from STEDT database according to the Swadesh 100-word list (See ‘Supplementary Table 1.’ spreadsheet’).

Supplementary Table 2. Node constraints used to calibrate the divergence time calculations, based on known historical information.

Language	Calibration setting in BEAST	Tips Date	Historical Information	The languages in our study	Source
Chinese	Normally distributed with mean 2700.0 and 150.0 standard deviation	2000 BP	The earliest known written records of the Chinese language dates from about 1250 BC. The old Chinese is the ancestor of modern Chinese Mandarin.	Chinese Mandarin, Old Chinese	Baxter, William H., Sagart, Laurent. 2014. Old Chinese: A New Reconstruction, Oxford University Press, ISBN 978-0-19-994537-5.
Old Chinese	Normally distributed with mean 2500.0 and 100.0 standard deviation. Tips Only.		From thirteenth century BCE to third century CE	Tips Only	LaPolla RJ, Thurgood G, editors. The Sino-Tibetan Languages. Taylor & Francis; 2016 Dec 8.
Tibetan dialects	Normally distributed with mean 1150.0 and 50.0 standard deviation		Tibetan dialects were formed in the mid-13th century. Tibetan dialects exclude the classical Tibetan.	Tibetan Amdo Blabrang, Tibetan_Khams, Tibetan_Lhasa, Tibetan_Batang, Tibetan_Jirel, Tibetan_Xiahe, Tibetan_Amdo_Zeku	Yi Xi. Zangyu Lishi De Fenqi yu Geqi Yuyin Tezheng. [The historical peorids of Tibetan languages and their speech characteristics]. Zangyu Yanjiu [Tibetan Studies]. 1993. 1, 101: 126-132
Burmese	Truncated between 400.0 and 1200.0		The language is not attested until early in the 12th century, when it begins to appear on tone inscriptions in the temples of Pagan.	Marma, Burmese Rangoon	LaPolla RJ, Thurgood G, editors. The Sino-Tibetan Languages. Taylor & Francis; 2016 Dec 8.
Pumi	Normally distributed with mean 750.0 and 50.0 standard deviation		Pumi people migarated into the regions of Ninglang, Lijiang, Weixi and Lanpin from 13th century.	Pumi_Qianghua, Pumi-Taoping	Sun, H., Hu, Z. and Huang, X., 2007. Zhongguo de yuyan [The languages of China]. Beijing: Shangwu Yinshuguan.
Yi	Normally distributed with mean 1500.0 and 100.0 standard deviation		The expansion of Yi people to northeast and south Yunnan, and northwest Guizhou dated to 3th century. About in the eighth and ninth century, a kindom called Nazhao was ruled by Yi (Lolo) speakers but the people also included Bai language.	Yi_Mojiang, Yi_Weishan, Yi_Mile, Yi_Wuding, Yi_Nanjian, Yi_Dafang, Lisu_Northern, Jinuo_Youle, Lahu_Yellow, Yi_Sani, Lisu_Central, Lahu_Lancang, Hani_Luchun, Mpi, Hani_Dazhai, Hani_Shuihui, Hani_Caiyuan, Hani_Mojiang, Hani_Gelanghe	Sun, H., Hu, Z. and Huang, X., 2007. Zhongguo de yuyan [The languages of China]. Beijing: Shangwu Yinshuguan.

Qiangic+rGaylrongic	Truncated between 1000.0 and 5000.0	The Qiang people in Gansu and Qinghai regions started southward migration after Qin and Han dynasties. And Southern Qiangic languages are formed at Tang and Song Dynasties.	Xumi, Queyu_Yajiang, Pumi_Qinghua, Ersu, Lyuzu, Pumi_Taoba, Qiang_Longxi, Namuyi, Muya_Min yak, Qiang_Mawo, Daofu, Ergong_Danba, rGyalrong_Maerkang, Caodeng	Sun, H., Hu, Z. and Huang, X., 2007. Zhongguo de yuyan [The languages of China]. Beijing: Shangwu Yinshuguan.
Karen	Normally distributed with mean 1150.0 and 100.0 standard deviation	About 8th century, Karen people migrated to Myanmar.	Pa_O, Karen, Kayan_Pekon	LaPolla, R.J., 2001. The role of migration and language contact in the development of the Sino-Tibetan language family. Areal diffusion and genetic inheritance: Case studies in language change: p. 225-254.

Supplementary Table 3. Model comparisons among different combinations of site models, clock models and rate heterogeneities. Each combination was run for 50 million generations, sampling in every 5,000 generations. The first 10% of the iterations were treated as burn-in and discarded. The Marginal likelihood estimated by the harmonic mean values were used to compare. Higher likelihood indicates better model fit, and the standard error values (S.E.) were estimated from 1,000 bootstrap replicates. The gray shadow highlights the best relative model fit of the row's model compared to the column's model.

No.	Model	ln P(model data)	S.E.	1	2	3	4	5	6
1	CTMC + Relaxed	-13167.914	+/- 0.077	-	-249.182	21.401	-211.054	-304.452	-269.649
2	CTMC + Relaxed + 4 γ	-12918.733	+/- 0.085	249.182	-	270.582	38.128	-55.271	-20.468
3	CTMC + Strict	-13189.315	+/- 0.431	-21.401	-270.582	-	-232.455	-325.853	-291.05
4	CTMC + Strict + 4 γ	-12956.86	+/- 0.169	211.054	-38.128	232.455	-	-93.398	-58.595
5	Covarion + Relaxed	-12863.462	+/- 0.085	304.452	55.271	325.853	93.398	-	34.803
6	Covarion + Strict	-12898.265	+/- 0.157	269.649	20.468	291.050	58.595	-34.803	-

Note: Marginal likelihood estimated using the method proposed by Newton and Raftery⁴⁹, with the modifications proposed by Suchard *et al.*⁵⁰.

Supplementary Table 4. The comparisons of delta scores and Q-residual values among different language families. The *p*-values for two-sided sign test were calculated by comparing delta score distribution of the 109 ST language samples with those values of other language families. The statistical test was implemented in MATLAB using the *signtest* function. NaN represents no report in corresponding language families. The references for language families are listed in the last column. The delta scores and Q-residual values of individual language sample are calculated by SPLITSTREE v4.16.6 using Gene Content distance (see Supplementary Section 2.6).

Language family	Sample size	Delta score	<i>p</i> -value	Q-residual score	<i>p</i> -value	Source
Sino-Tibetan	109	0.35	-	0.015	-	In this study
Indo-European	83	0.22	4.43E-25	0.002	4.43E-25	³⁷
Austronesian	20	0.33	2.73E-04	0.002	4.43E-25	³⁷
Polynesian	40	0.41	4.43E-25	0.02	3.75E-20	³⁷
Cariban	34	0.38	3.49E-17	NaN	NaN	³⁹