Dated language phylogenies shed light on the ancestry of Sino-Tibetan

Laurent Sagart[a,1], Guillaume Jacques[a,1], Yunfan Lai[b], Robin J. Ryder[c], Valentin Thouzeau[c], Simon J. Greenhill[b,d], and Johann-Mattis List[b,2]

[a]Centre de Recherches Linguistiques sur l'Asie Orientale, CNRS, Institut National des Langues et Civilisations Orientales, Ecole des Hautes Etudes en Sciences Sociales, 75006 Paris, France; [b]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena 07743, Germany; [c]Centre de Recherches en Mathématiques de la Décision, CNRS, Université Paris-Dauphine, PSL University, 75775 Paris, France; and [d]Australian Research Council Center of Excellence for the Dynamics of Language, Australian National University, Canberra, ACT 0200, Australia

Edited by Balthasar Bickel, University of Zurich, Zurich, Switzerland, and accepted by Editorial Board Member Richard G. Klein April 8, 2019 (received for review October 19, 2018)

The Sino-Tibetan language family is one of the world's largest and most prominent families, spoken by nearly 1.4 billion people. Despite the importance of the Sino-Tibetan languages, their prehistory remains controversial, with ongoing debate about when and where they originated. To shed light on this debate we develop a database of comparative linguistic data, and apply the linguistic comparative method to identify sound correspondences and establish cognates. We then use phylogenetic methods to infer the relationships among these languages and estimate the age of their origin and homeland. Our findings point to Sino-Tibetan originating with north Chinese millet farmers around 7200 B.P. and suggest a link to the late Cishan and the early Yangshao cultures.

Sino-Tibetan languages | human prehistory | East Asia | peopling | computer-assisted language comparison

The past 10,000 y have seen the rise, at the western and eastern extremities of Eurasia, of the world's two largest language families. Together, these families account for nearly 60% of the world's population: Indo-European (3.2 billion speakers) and Sino-Tibetan (1.4 billion). The Sino-Tibetan family comprises about 500 languages (1) spoken across a wide geographic range, from the west coast of the Pacific Ocean, across China, and extending to countries beyond the Himalayas, such as Nepal, India, Bangladesh, and Pakistan (map, *SI Appendix*, section 2). Speakers of these languages have played a major role in human prehistory, giving rise to several of the world's great cultures in China, Tibet, Burma, and Nepal. However, while the debate on Indo-European origins has recently been renewed by archaeogeneticists, phylogeneticists, and linguists (2–5), the circumstances of the formation of Sino-Tibetan remain shrouded in obscurity.

While Sino-Tibetan languages have been studied from the beginning of the 19th century (6), our knowledge of the history of this family is still severely limited, since it is structurally one of the most diverse families in the world, including all of the gradation of morphological complexity from isolating (Lolo-Burmese, Tujia) to polysynthetic (Gyalrongic, Kiranti) languages (7, 8). Knowledge of Sino-Tibetan sound correspondences is improving (*SI Appendix*, section 2), yet important aspects of its phonological and grammatical history remain poorly understood, e.g., the voicing and aspiration of modern stops, or the correspondences between tones and nontonal elements. These difficulties place some uncertainty on cognate identification and, in turn, affect our ability to identify shared innovations. This complexity has led to claims that Sino-Tibetan is one of the greatest challenges that comparative-historical linguistics currently faces (ref. 9, p. 422).

Where did these languages originate and when? The vast majority of Sino-Tibetan speakers speak a Chinese, or Sinitic, language. The Sinitic languages, whose ancestor was spoken about 2,000 y ago, form a homogeneous group in the eastern half of the Sino-Tibetan area. The earliest paleographical inscriptions in Chinese date to before 1400 BCE, and Chinese has an abundant and well-studied literature dating back to the early first millennium BCE. The Shāng Kingdom, the Chinese polity associated with these inscriptions, was centered on the lower Yellow River valley. Gradual annexation of neighboring regions and shift of their peoples to the Chinese language led to the striking numerical predominance of Chinese speakers today, and, consequently, to the lack of linguistic diversity in the eastern part of the Sino-Tibetan domain. Tibetan, Tangut, Newar, and Burmese, the family's other early literary languages, were reduced to script considerably more recently: The oldest texts in these languages date from 764 CE, 1070 CE, 1114 CE, and 1113 CE, respectively. The area with the most diverse Sino-Tibetan languages is in northeastern India and Nepal. This has suggested to some authors that the family's homeland was located there (10). However, Sino-Tibetan diversity in India and Nepal may have been boosted by intimate contact with very divergent and mostly extinct non–Sino-Tibetan languages, in much the same way that Austronesian diversity in northwest Melanesia was boosted by contact with Papuan languages (11) despite their homeland in Taiwan (12). Due to these difficulties, no consensus exists about the phylogenetic relationships within the family. The position of Chinese, in particular, is in dispute. A first group of proposals recognizes a two-branch structure: One branch leads to Chinese, and the other leads to a node labeled "Tibeto-Karen" or "Tibeto-Burman," out of which all other languages proceed (13, 14). A second group presents Sino-Tibetan basal topology as a rake, with Chinese being one of several primary branches (10). A third

Significance

Given its size and geographical extension, Sino-Tibetan is of the highest importance for understanding the prehistory of East Asia, and of neighboring language families. Based on a dataset of 50 Sino-Tibetan languages, we infer phylogenies that date the origin of the language family to around 7200 B.P., linking the origin of the language family with the late Cishan and the early Yangshao cultures.

Author contributions: L.S., G.J., and J.-M.L. initiated the study, later joined by S.J.G., who provided initial phylogenetic analyses; L.S., G.J., Y.L., and J.-M.L. assembled the data; L.S. and G.J. coded the cognates; R.J.R., V.T., and S.J.G. made phylogenetic analyses; J.-M.L. developed the system for data curation and organized the project; V.T., S.J.G., and J.-M.L. provided the figures; L.S., G.J., Y.L., R.J.R., V.T., S.J.G., and J.-M.L wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. B.B. is a guest editor invited by the Editorial Board.

[1]L.S. and G.J. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: list@shh.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1817972116/-/DCSupplemental.

Published online May 6, 2019.

group places Chinese in a lower-level subgroup with Tibetan (15, 16). Apart from the second group, which relies on lexicostatistic methodology, the tree topologies in these proposals are based on an investigator's perception of relative proximities between branches, with no quantification of uncertainty. A search for linguistic innovations uniting several branches of the family is ongoing; the limited results so far are consistent with the first group of hypotheses (9, 17). *SI Appendix*, section 2 summarizes different proposals.

Here we combine classical historical linguistics with cutting-edge computational methods and domestication studies. First, we develop a lexical database of 180 basic vocabulary concepts from 50 languages. The data were either directly collected in the field by ourselves or gathered from the literature with verification by external specialists whenever possible. The list of most appropriate concepts was established through careful evaluation of concept lists used in similar studies (*SI Appendix*, section 3), and lexical cognates were identified by experts in Sino-Tibetan historical linguistics using the comparative method supported by state-of-the-art annotation techniques. Second, we apply Bayesian phylogenetic methods to these data to estimate the most probable tree, outgroup, and timing of Sino-Tibetan under a range of models of cognate evolution; similar methods have been applied to several other families of languages, including Indo-European (18–20), Austronesian (12), Semitic (21), and Bantu (22). Third, we examine Sino-Tibetan expansion under the two most probable phylogenetic scenarios through a consideration of the family's plant and animal domesticates, the regions where they are earliest attested archaeologically, and the distribution of the corresponding cognate sets across the family's branches.

Results

Cognate Set Distribution. Of the 3,333 cognate sets distributed over 9,160 lexical items, 90% are shared by fewer than five languages. The majority of these low-frequency sets are singletons for which no related word in any other language was found (2,189, 66%). Four cognate sets are found in all or almost all languages in our sample, reflecting well-known Sino-Tibetan cognates ("three," "four," "dream," and "name"). These numbers compare well with the data obtained for other challenging language families (see *SI Appendix*, section 3 for details).

Tree Topologies and Dating. We present tree topologies and ages inferred using a relaxed-clock covarion model with BEAST, a phylogenetic software package performing Bayesian evolutionary analysis; see Fig. 2 for summarization. Apart from the Sinitic group which was constrained in the priors, the posterior distribution provides strong evidence (>95% probability) for six subgroups: (*i*) Tibeto-Gyalrongic (possibly including Dulong in a Tibeto-Dulong clade), (*ii*) Kiranti, (*iii*) West-Himalayish, (*iv*) Tani-Yidu, (*v*) Kuki-Tangkhul (possibly including Karbi), and (*vi*) Sal. Tshangla and Chepang are isolated branches. Within the Tibeto-Gyalrongic group, there is also support for Tibetan, Lolo-Burmese, Gyalrongic, and Burmo-Gyalrongic. The more recent part of the tree is thus well resolved.

On the other hand, the results do not allow us to unambiguously resolve the root of the tree. The most plausible outgroups, judging from posterior probabilities, are Sinitic (33%), West-Himalayish (15%), Tani-Yidu (9%), a Sinitic-Sal group (8%), and Sal (6%). The mean root age estimated with the relaxed-clock model is at 7184 B.P., with 95% highest posterior density interval (HPD) [5093–9568] B.P.

In addition to this main analysis, we also analyzed the data using two more constrained models: a strict-clock covarion model and a Stochastic Dollo model. These more constrained models lead to less uncertainty in the deep tree topology. In particular, under the strict-clock model, Sinitic is the only possible outgroup; the Stochastic Dollo model gives outgroups probabilities similar to the relaxed-clock model. The differences are discussed further in *SI Appendix*, section 4. Repeating the analyses on a smaller sample representing each of the major subgroups yielded similar results, further discussed in *SI Appendix*, section 4. Tests of the adequacy of the tree model are further discussed in *Adequacy of the Tree Model*.

Discussion

Tree Topology and Subgrouping Hypotheses. Despite the preliminary character of our study, until further key languages of the family like Newar are sufficiently analyzed and added, our results consistently support two nontrivial subgrouping hypotheses previously proposed by historical linguists on the basis of lexical innovations: The clade comprising Garo, Rabha, and Jinghpo in the sample is compatible with the Sal subgroup (23), and the clade including Burmish languages, Lisu, Gyalrongic (Japhug, Situ, Tangut, Stau, and Khroskyabs), and Zhaba corresponds to the Eastern Tibeto-Burman or Burmo-Gyalrongic subgroup (24, 25). Our results also indicate that the Burmo-Gyalrongic group belongs to a larger Tibeto-Gyalrongic clade comprising Tibetan and also possibly Dulong, a hypothesis that had not been explicitly proposed before.

The results are inconsistent with a certain number of subgrouping proposals, in particular, Sino-Bodic [grouping together Chinese, Tibetan, and Kiranti, excluding Lolo-Burmese (26)]; Post and Blench's hypothesis that subgroups in northeastern India such as Tani (Bokar) and Mishmi (Yidu and Deng) are among the first branches of the family, while Sinitic is closer to Lolo-Burmese and Tibetan (16); the Central Trans-Himalayan hypothesis [a clade comprising Sal and Kuki-Chin (27)]; and the Rungic hypothesis (28), according to which the morphology-rich subgroups (Gyalrongic, Kiranti, and Dulong) constitute a clade to the exclusion of Lolo-Burmese and Tibetan (*SI Appendix*, section 4). The last two hypotheses are exclusively based on verbal morphology. The fact that these subgroups are not confirmed by our results suggests that the commonalities in verbal morphology adduced by these authors to support these subgroups are more likely to reflect a combination of retentions from a common ancestor and parallel innovations.

Since the common origin of person agreement morphology among Gyalrongic, Dulong, and Kiranti is not controversial (28, 29), and since Kiranti is outside of the Tibeto-Dulong clade, phylogenetic inference supports the idea that the absence of person inflexion in Lolo-Burmese and Tibetan is due to a massive loss of morphology (7), a hypothesis also supported by potential traces of these inflexions in Tibetan (30). The proximity of a set of isolating (Lolo-Burmese) and polysynthetic (Japhug and Situ) languages in our results supports the idea that the rate of change of structural features can be much more volatile than that of basic vocabulary (31), and provides an additional example of abrupt loss of inflectional morphology, comparable to the case of Goemai in Chadic (32).

Although the likely Urheimat of the family lies in northern China and Sinitic may be the first group to branch off, the diversity of the subgroups of Sino-Tibetan is highly skewed toward northern India and Nepal. Of the nine subgroups supported by our results (Sinitic, Tibeto-Dulong, Sal, Kiranti, Kuki-Karbi, Tani-Yidu, West-Himalayish, and the isolated languages Chepang and Tshangla), only two groups (Sinitic and Tibeto-Dulong) are well represented in China. Three other branches (Sal, Tani-Yidu, and Tshangla) are mainly spoken in Burma, India, and Bhutan but straddle the border with China. This geographical distribution suggests that the historical success of a few subgroups (Sinitic, Tibetan, and Lolo-Burmese, the latter two belonging to the Tibeto-Dulong group) has eroded linguistic diversity in China, including on the Tibetan plateau, whereas the

Himalayan area has served as a refuge zone (27, 33, 34), resulting in a higher diversity.

Homeland, Archaeology, and Agriculture. It is claimed that language families arise through demographic processes driven by favorable changes in food procurement (35); thus, any account of the origins of a language family should pay attention to its domesticates. We identify six domesticate names forming cognate sets with regular sound correspondences in at least two of the branches identified in our phylogenetic analysis: foxtail millet, pig, sheep, rice plant, cattle, and horse (*SI Appendix*, section 5). The fact that, archaeologically, all of these first appear in northern China, even those with cognate sets lacking a Chinese member, is a strong indication that, early in its expansion phase, the Sino-Tibetan family was located in that broad area (Fig. 1), now occupied by Sinitic languages. In particular, under our root date of *ca*. 7,200 yBP, broomcorn millet, foxtail millet, rice, pigs, and sheep are early enough to have played a demography-boosting role in the early stages of Sino-Tibetan expansion, although we do not think rice was known to ancestral Sino-Tibetan speakers. The northeastern part of the Sino-Tibetan domain is thus the family's most likely homeland. Alternative proposals such as Sichuan (26), eastern India (16), and the Tibetan plateau (36) lack an archaeologically and demographically supported account of the family's expansion.

Our recognition of the family's most likely outgroups (*SI Appendix*, section 4) suggests two possible expansion scenarios of Sino-Tibetan languages: a Chinese outgroup scenario and a West-Himalayish outgroup scenario. Under the first (33% probability, median root date: 7400 B.P.), the Sino-Tibetan homeland was located in the eastern half of the north Chinese loess plateau, during the final stages of Cishan culture or the initial stages of Yangshao culture (Fig. 1). The Sinitic homeland is located just to the south: No significant Sinitic migration needs to be assumed. The non-Sinitic group would have individualized as a result of its expansion to the western half of the plateau, inside the northern bend of the Yellow River.

Both the Cishan and Yangshao cultures derived their subsistence principally from broomcorn and foxtail millet, and pigs; in addition, domesticated sheep have been identified at the northern edge of early Yangshao culture (37). These four domesticates were still widely relied upon in early China and are important to Sino-Tibetan speakers elsewhere. Rice, horses, cattle, wheat, and barley are absent archaeologically in Cishan and Yangshao: Under the Sinitic out-group scenario, a Sino-Tibetan homeland in the Cishan–Yangshao region predicts that cognate sets, reflected in and outside of Sinitic, should exist for the two millets, pig, and sheep, but not for rice, horses, cattle, wheat, or barley. The individual cognate set distributions, including two identified sets for rice and pig, support these predictions (*SI Appendix*, section 5).

The secondary Sino-Tibetan domesticates (rice, cattle, horses, wheat, and barley) would have entered the non-Sinitic branch of Sino-Tibetan in the early times of its westward and southward expansion, through contact with neighboring groups. First, *Japonica* rice, presumably spread from the southeast (Henan), was adopted: Indirect evidence of rice cultivation (phytoliths) appears in the late Yangshao culture *ca.* 5690 BP, in the

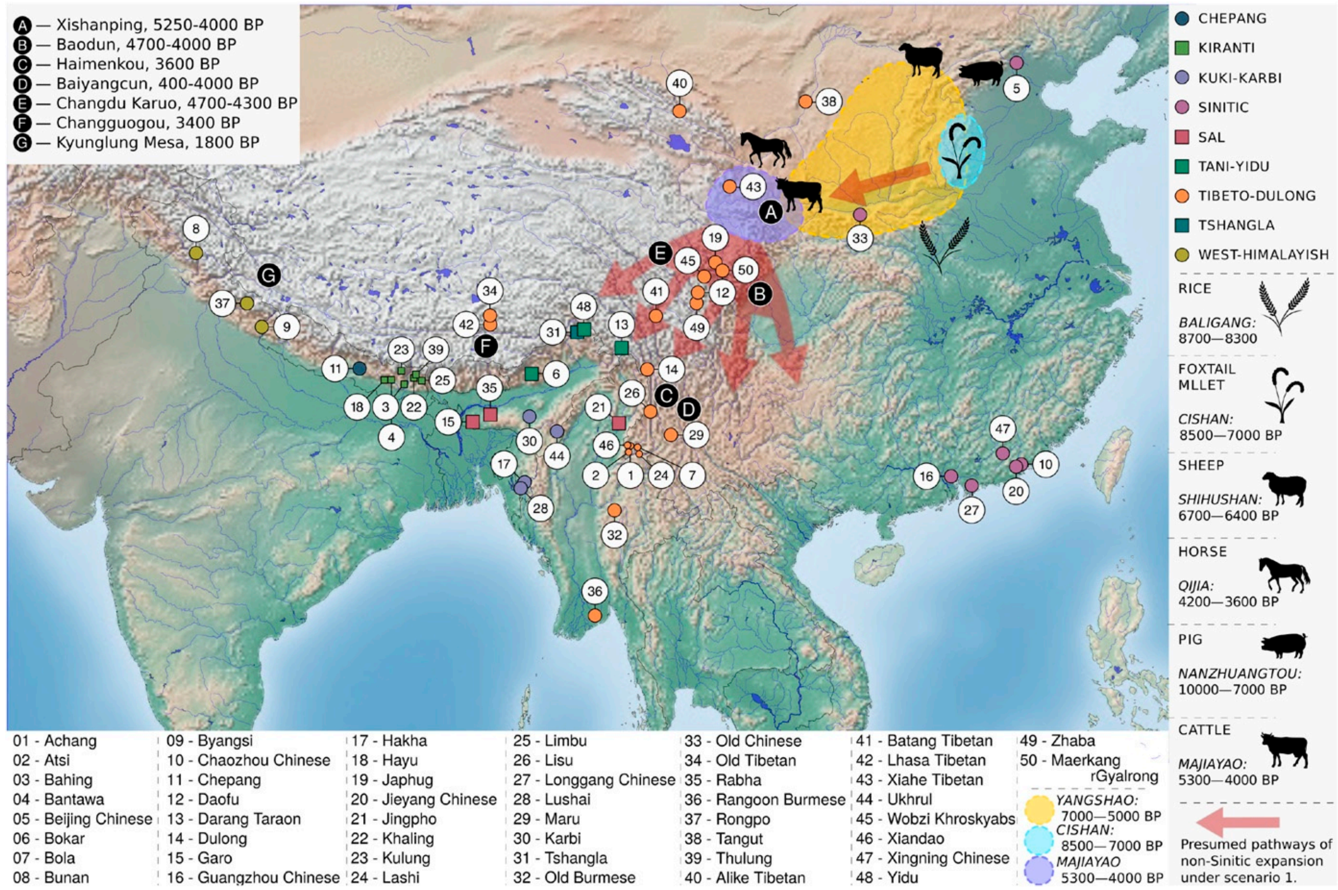

Fig. 1. Languages in our sample, contrasted with ancient sites reflecting early stages of domestication and the estimated spread of non-Sinitic languages (see *Homeland, Archaeology, and Agriculture*). Ancient language locations reflect supposed political and cultural epicenters of the varieties.

southwest part of the Yangshao area (Wei River valley) (38). Rice, by then outside of its wild habitat, is well established in Xishanping, at the far western end of the loess plateau in southeastern Gansu (Majiayao culture), where pigs and the two millets are also found, by 5070 BP. Millet-and-rice agriculture then expands south along the eastern edge of the Tibetan plateau, entering Sichuan (Baodun, 4700 BP), although d'Alpoim Guedes (39) sees Baodun rice as an extension of Yangtze irrigated cultivation rice. This complex subsistence strategy further expands south into Yunnan (e.g., Baiyangcun, 4500 BP; Haimenkou, 3600 BP). The earliest archaeological evidence for cattle and horses, domesticated west of China, comes from Gansu, at the eastern edge of the Tibetan plateau, in the period 5400 BP to 4200 BP; we assume that the horse entered the non-Sinitic branch in that region, and that its name was later transferred to Sinitic; for the name of cattle, see (*SI Appendix*, section 5).

The archaeology of Sino-Tibetan–speaking regions in Burma, northeast India, and the Himalayan area is very limited, but radiation from Yunnan along the main rivers which flow out of the Himalayas would bring non-Sinitic speakers and their domesticates to many of their current locations. Genetics support the idea (40, 41) that a second route carried Sino-Tibetan speakers southwest from Gansu across the Tibetan plateau: foxtail millet—but not rice—was cultivated at Changdu Karuo on the Mekong River in eastern Tibet between 4700 BP and 4300 BP, later at Changguogou along the Yarlung Tsangpo in southern Tibet beginning *ca.* 3400 BP, and, by *ca.* 1700 BP, farther west, near the Indus River, at Kyung-lung Mesa (39). This route would bring Sino-Tibetan speakers to the area occupied by modern West-Himalayish speakers, the western-most Sino-Tibetan group. Modern speakers of these languages have replaced the millets and rice by the more frost-tolerant cereals barley and wheat (42), indicative of an adaptation to the Tibetan plateau.

Alternatively, under the West Himalayan outgroup scenario (15%, 7200 BP) and assuming a homeland in the eastern loess plateau, the two groups West-Himalayish and non–West-Himalayish expanded westward in parallel, reaching the western end of the loess plateau at similar places and times, in the late sixth millennium BP, speaking distinct languages. One group then moved southwestward across the plateau, while the other expanded south following the plateau's foothills. Depending on the timing of subsequent splits, the group ancestral to Sinitic might face a lengthy eastward back-migration.

Taken together, these two scenarios account for nearly half of the tree output behind our phylogeny. However, with 33% posterior probability and a more straightforward pattern of expansion, the Chinese outgroup scenario is the better supported of the two.

Materials and Methods

Lexical Data. Sino-Tibetan languages differ considerably as to their syllabic structure. Some languages only allow consonant–vowel-type syllables, while others have complex clusters and final consonants. The former type of languages can be shown to be highly innovative, and a series of specific sound changes generally make cognate judgments very difficult, except for a few well-investigated cases. For this reason, it was decided to exclude from the sample all languages having lost the final stops -p, -t, and -k, unless published sources on the sound laws necessary to recover the lost segments were available (as in the case of Lisu, for which ref. 43 was used). This resulted in a collection of 50 Sino-Tibetan languages (see *SI Appendix*, section 3), which reflects the major particularly well-studied subgroups of the language family, including modern Chinese dialects. The present concept list is based on a larger set of 250 words (see *SI Appendix*, section 3), reduced to 180 on the basis of the following criteria: (*i*) availability of data, (*ii*) avoiding pairs of concepts with high polysemy ("hand" vs. "arm"; in such cases, only one concept was chosen), and (*iii*) avoiding words prone to have nursery forms ("father" and "mother").

Particular importance was attributed to assembling a dataset of high average mutual coverage, defined as the proportion of the overlap of concepts for which a translation exists in each language pair (44). With low coverage, the uncertainty of phylogenetic analyses increases, as well as the difficulty for experts to identify reliable cognates. To avoid this problem, we made sure that all languages have translations for at least 85% of the concepts in our questionnaire. Due to our strict procedure, some potentially important languages, such as Newar, are missing from our sample (*SI Appendix*, section 3). This means that our results are preliminary, but we think they are interesting and important enough to be shared.

When preparing the language data, we selected the translations for our concept list semiautomatically and used publicly available software libraries (45) to convert the language-specific orthographies into standard phonetic transcriptions (46), to ease the task of cognate assignment. At all stages of data preparation, we maintained a computer-assisted as opposed to a solely computer-based or solely manual workflow: We made use of automatic tools and custom scripts for data preprocessing, but we made sure that all data were always checked again by an expert to avoid errors.

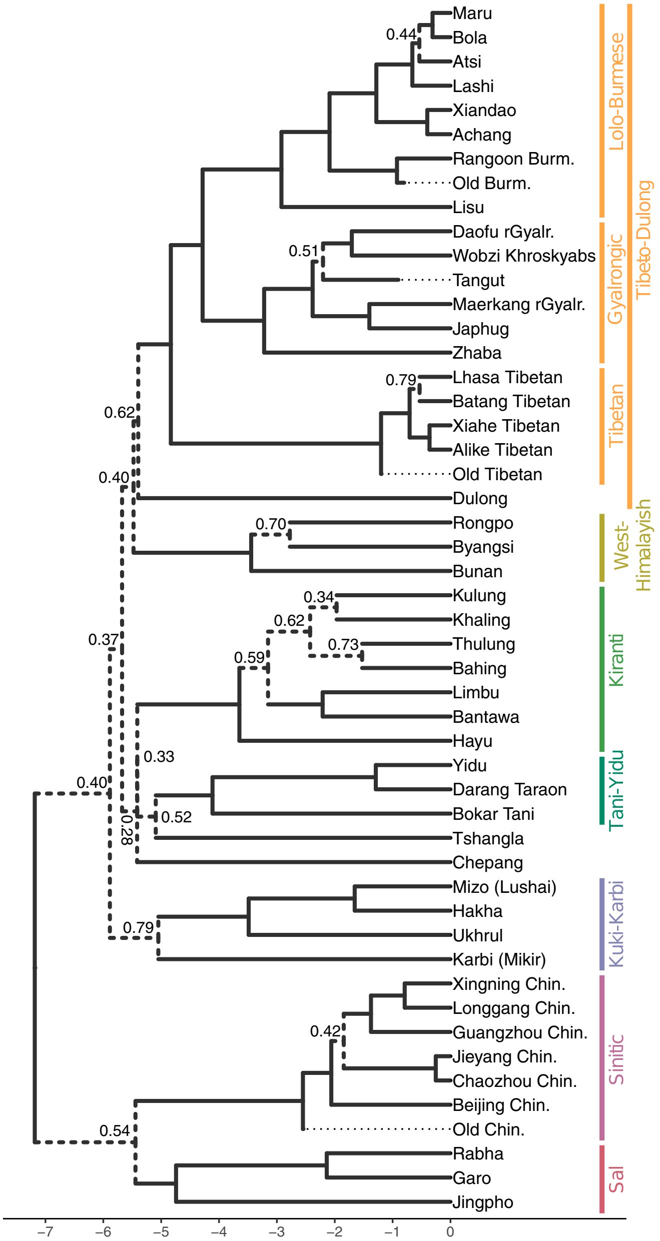

Fig. 2. The Maximum Clade Credibility tree from the best-fitting model (relaxed clock with covarion). Branches with less than 0.8 posterior probability are dashed; other branches have posterior probabilities >0.8. For densitrees of the data, see *SI Appendix*, section 4.

Cognate Judgments. During cognate coding, particular care was devoted to the identification of loanwords in the dataset. Loanwords from Chinese, Tibetan, and Indic languages (in particular, Hindi and Nepali) are widespread in languages of the sample. They can be identified using the methodology described in refs. 47 and 48 and discriminated from inherited cognates. More difficult to detect are borrowings among closely related languages. In the case of Sinitic and Tibetic languages, knowledge of sound laws is precise enough, and allows recognition of a considerable number of borrowings. For nonliterary language families such as Gyalrongic and Kiranti, identification of such borrowings, although possible in exceptional cases, cannot be systematically undertaken. Identified loanwords were coded as cognates in the database, with a specific tag, allowing them to be disregarded when performing analyses.

Since the Sino-Tibetan languages exhibit rich patterns of compounding (14) and derivation (49), a particular problem of cognate assessment in Southeast Asian languages is to deal with words that are only related in some of their morphemes, but not entirely (50). Words for plural personal pronouns, for example, are often compounds of the word for the singular pronoun and a plural suffix (compare Chinese *wǒ* "I" vs. *wǒ-men* "we"). If words are only cognate in some of their parts, it may be difficult to make a judgment on their overall cognacy (50). We tried to circumvent this problem by (*i*) excluding items of high compoundhood, (*ii*) using transparent annotation techniques to identify the main meaning-bearing unit, and (*iii*) using multiple alignments to show which parts of words are cognate (*SI Appendix*, section 3).

The cognate judgments were carried out with help of a new annotation framework (51), reflecting expert decisions with high transparency. Where possible, we also link to the original sources, especially those taken from the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) database (36). To facilitate reuse of our data, we further provide the data in the form suggested by the Cross-Linguistic Data Formats initiative (52), which facilitates data reuse by linking to public reference catalogs, like Glottolog for languages (1) and Concepticon for concepts (53).

Phylogenetic Reconstruction and Dating. The cognate sets can be coded as a binary matrix with missing values. We used three models of cognate change along a tree to reconstruct the tree topology and internal node ages in a Bayesian framework. For each model, we performed Markov Chain Monte Carlo (MCMC) sampling to obtain a sample of plausible trees from the posterior distribution. Each model defines how cognates may switch from being present (1) to absent (0) in a language, and vice versa. Under the binary covarion model (54), there is a latent variable indicating whether cognates switch between 0 and 1 at a "fast" or "slow" rate on each branch, with transitions from 0 to 1 and 1 to 0 being equally probable. In addition to the slow and fast rates, this model also includes a parameter governing the transition of the latent parameter. We used two different covarion models: The strict-clock model forces cognates to change at the same rate over all branches on the tree; the relaxed-clock model allows each branch of the tree to have its own clock rate, drawn in a lognormal distribution. Under the Stochastic Dollo model (55), cognates appear and disappear at constant rates, with the additional constraint that each cognate may only be born once on the tree; parallel innovations are thus not allowed under this model. We specified calibrations as follows: Old Chinese, [2,800 to 2,300] yBP, in a uniform prior; Old Burmese, 800 yBP; Old Tibetan, 1,200 yBP; and Tangut, 900 yBP (the date range for Old Chinese corresponds to the period of the great Classical Chinese texts; the other dates correspond to the date of the earliest text rounded to the nearest century; see also *SI Appendix*, section 4). In addition to the main consensus tree in Fig. 2, we also present densitrees in *SI Appendix* to show which aspects of the reconstructions are uncertain. To make sure that our sampling of language varieties does not influence the results, we also analyze a subset of the data (see *SI Appendix*, section 4).

Analyses Under the Covarion Models. We constructed trees using the software BEAST2 v2.4.7 (56). We fitted a Fossilized Birth-Death model (57) which allows us to include extinct languages, with both a strict-clock and a relaxed-clock model. We performed 10^8 MCMC iterations, with a burn-in of 10^7 iterations, reaching convergence with effective sample size (ESS) $>$ 300 for each parameter. We thinned trees every 10^4 generations to remove autocorrelation. We sampled 10^4 trees from the posterior tree distribution to produce a maximum clade credibility tree using the software TreeAnotator v2.4.7. Verification of convergence and ESS computation were produced using Tracer v.1.6 (58). We then used a Nested Sampling algorithm (59) to compare the marginal likelihood of a strict-clock model and a relaxed-clock model (60). In each case, we used 40 particles to compute the marginal likelihood. The log Bayes factor was estimated at 85 (SD = 23) in favor of the relaxed-clock model, indicating decisive evidence against the strict-clock model. The results presented in Fig. 2 therefore correspond to the relaxed clock covarion model analysis with BEAST, and were plotted using the ggtree R package (61).

Analyses Under the Stochastic Dollo Model. We also analyzed the data under the Stochastic Dollo model implemented in TraitLab (62). The results we present in *SI Appendix* correspond to a run of 10^8 iterations thinning every 10^4 and discarding the first 10% as burn-in. Visual checks indicated that the Markov Chain had reached stationarity and mixed well. We included catastrophic rate heterogeneity as proposed by ref. 19, but the results are the same whether we include or exclude rate heterogeneity.

Adequacy of the Tree Model. We verified the adequacy of a tree model by reanalyzing a subset of the data under the Lateral Transfer Stochastic Dollo model of ref. 63, which allows for non–tree-like evolution. For computational reasons, this was done on a subset of 15 randomly chosen leaves representing all of the major subfamilies. Following ref. 63, we assessed model fit by estimating the model on a randomly selected subset of 75% of the traits in the data, then computing the posterior predictive probability for the remaining 25% of the data. This gave a Bayes factor of $\log_{10} BF = 1.8$ in favor of the network model; although this favors the network model, the evidence is not decisive, and using a tree model should not lead to issues in the inference. Furthermore, the estimated level of lateral transfer in the network model is $\hat{\beta}/\mu = 0.104$ (95% HPD [0.058 to 0.162]), which is within the range shown by section 4.2.1 of ref. 64 to not be liable to systematic bias when estimating the root age and topology under a tree model.

ACKNOWLEDGMENTS. We thank L. Konnerth, M. Widmer, N. W. Hill, K. Ma, H. Hammarström, and R. LaPolla for help and information on languages and data; H. Sell for the geographic map; M.-S. Wu for comments on the map; and R. D. Gray for comments on the draft. J.-M.L. and Y.L. are supported by European Research Council Starting Grant 715618. S.J.G. is supported by Australian Research Council (ARC) Discovery Project Grant DE 120101954 and ARC Center of Excellence Grant CE140100041. V.T. is supported by a Paris Sciences et Lettres Data Science, Science Data scholarship.

1. Hammarström H, Forkel R, Haspelmath M (2018) *Glottolog* (MPI-SHH, Jena, Germany).
2. Haak W, et al. (2015) Massive migration from the steppe is a source for Indo-European languages in Europe. *Nature* 522:207–211.
3. Allentoft ME, et al. (2015) Population genomics of Bronze Age Eurasia. *Nature* 522:167–172.
4. Bouckaert R, et al. (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337:957–960.
5. Garnier R, Sagart L, Sagot B (2017) Milk and the Indo-Europeans. *Language Dispersal Beyond Farming* (John Benjamins, Amsterdam), pp 291–311.
6. Leyden J (1808) On the languages and literature of the Indo-Chinese nations. *Asiatick Res* 3:158–289.
7. DeLancey S (2015) The historical dynamics of morphological complexity in Trans-Himalayan. *Ling Discov* 13:37–56.
8. Jacques G (2016) Tangut, Gyalrongic, Kiranti and the nature of person indexation in Sino-Tibetan/Trans-Himalayan. *Ling Vanguard* 2:1–13.
9. Handel Z (2008) What is Sino-Tibetan? Snapshot of a field and a language family in flux. *Lang Linguist Compass* 2/3:422–441.
10. Peiros I (1998) *Comparative Linguistics in Southeast Asia* (Australian National Univ, Canberra, ACT, Australia).
11. Pawley A (2006) Explaining the aberrant Austronesian languages of southeast Melanesia: 150 years of debate. *J Polyn Soc* 115:215–257.
12. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.
13. Benedict PK (1972) *Sino-Tibetan: A Conspectus* (Cambridge Univ Press, Cambridge).
14. Matisoff JA (2003) *Handbook of Proto-Tibeto-Burman*, University of California Publications in Linguistics (Univ California Press, Berkeley), Vol 135.
15. van Driem G (2003) Review of Thurgood and LaPolla 2003. *Bull Sch Orient Afr Stud* 66:282–284.
16. Blench R, Post MW (2014) Rethinking Sino-Tibetan phylogeny from the perspective of north east Indian languages. *Trans-Himalayan Linguistics*, eds Hill NW, Owen-Smith T (Mouton de Gruyter, Berlin), pp 71–104.
17. Sagart L (2017) A candidate for a Tibeto-Burman innovation. *Cah Linguist Asie Orient* 46:101–119.
18. Gray RD, Atkinson QD (2003) Language-tree divergence times support the anatolian theory of Indo-European origin. *Nature* 426:435–439.
19. Ryder RJ, Nicholls GK (2011) Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *J R Stat Soc B* 60: 71–92.

20. Chang W, Cathcart C, Hall D, Garrett A (2015) Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91:194–244.
21. Nicholls G, Ryder R (2011) Phylogenetic models for Semitic vocabulary. *Proceedings of the 26th International Workshop on Statistical Modelling* (Copiformes, Valencia, Spain), p 26.
22. Currie TE, Meade A, Guillon M, Mace R (2013) Cultural phylogeography of the Bantu languages of Sub-Saharan Africa. *Proc R Soc. B* 280:20130695.
23. Burling R (1983) The Sal languages. *Linguist Tibeto-Burman Area* 7:1–32.
24. Bradley D (1997) Tibeto-Burman languages and classification. *Papers in Southeast Asian Linguistics*, ed Bradley D (Pacific Linguistics, Canberra, ACT, Australia), pp 1–72.
25. Jacques G, Michaud A (2011) Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze. *Diachronica* 28:468–498.
26. van Driem G (1997) Sino-bodic. *Bull Sch Orient Afr Stud* 60:455–488.
27. DeLancey S (2015) Morphological evidence for a central branch of Trans-Himalayan (Sino-Tibetan). *Cah Linguist Asie Orient* 44:122–149.
28. Thurgood G (2017) Sino-Tibetan: Genetic and areal subgroups. *The Sino-Tibetan Languages*, eds Thurgood G, LaPolla R (Routledge, London), pp 3–39.
29. van Driem G (1993) The Proto-Tibeto-Burman verbal agreement system. *Bull Sch Orient Afr Stud* 61:292–334.
30. Jacques G (2010) A possible trace of verbal agreement in Tibetan. *Himalayan Linguist* 9:41–49.
31. Greenhill SJ, et al. (2017) Evolutionary dynamics of language systems. *Proc Natl Acad Sci USA* 114:E8822–E8829.
32. Hellwig B (2011) *A Grammar of Goemai* (Mouton de Gruyter, Berlin).
33. Nichols J (1992) *Language Diversity in Space and Time* (Univ Chicago Press, Chicago).
34. Bickel B, Nichols J (2005) Inclusive/exclusive as person vs. number categories worldwide. *Clusivity*, eds Haspelmath M, Dryer MS, Gil D, Comrie B (Oxford Univ Press, Oxford), pp 94–97.
35. Diamond J, Bellwood P (2003) Farmers and their languages: The first expansions. *Science* 300:597–603.
36. Matisoff JA (2015) *The Sino-Tibetan Etymological Dictionary and Thesaurus Project* (Univ California, Berkeley).
37. Dodson J, et al. (2014) Oldest directly dated remains of sheep in China. *Sci Rep* 4: 7170.
38. Zhang J, et al. (2010) Phytolith evidence for rice cultivation and spread in Mid-Late Neolithic archaeological sites in central North China. *Boreas* 39:592–602.
39. d'Alpoim Guedes J, et al. (2013) Moving agriculture onto the Tibetan plateau: The archaeobotanical evidence. *Archaeol Anthropol Sci* 6:255–269.
40. Kang L, et al. (2012) Y-chromosome O3 haplogroup diversity in Sino-Tibetan populations reveals two migration routes into the eastern Himalayas. *Ann Hum Genet* 76:92–99.
41. Wang LX, et al. (2018) Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol Genet Genomics* 293:1293–1300.
42. d'Alpoim Guedes JA, Lu H, Hein AM, Schmidt AH (2015) Early evidence for the use of wheat and barley as staple crops on the margins of the Tibetan Plateau. *Proc Natl Acad Sci USA* 112:5625–5630.
43. Bradley D (1979) *Proto-Loloish* (Curzon, London).
44. Rama T, List JM, Wahle J, Jäger G (2018) Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?*Proceedings of the North American Chapter of the ACL* (N Am Assoc Computational Linguistics, Stroudsburg, PA), pp 393–400.
45. List JM, Greenhill S, Forkel R (2017) *LingPy. A Python Library for Quantitative Tasks in Historical Linguistics* (Max Planck Inst Sci Human History, Jena).
46. List JM, et al. (2019) *Cross-Linguistic Transcription Systems* (Max Planck Institute for the Science of Human History, Jena, Germany).
47. Sagart L, Xu S (2001) History through loanwords: The loan correspondences between Hani and Chinese. *Cah Linguist Asie Orient* 30:3–54.
48. Jacques G (2004) Phonologie et Morphology du Japhug (rGyalrong). Ph.D. thesis (Univ Paris VII–Denis Diderot, Paris).
49. Jacques G (2017) A reconstruction of Proto-Kiranti verb roots. *Folia Linguist Hist* 38:177–215.
50. List JM (2016) Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *J Lang Evol* 1:119–136.
51. List JM (2017) A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. *Proceedings of the EACL* (European Assoc Computational Linguistics, Barcelona), pp 9–12.
52. Forkel R, et al. (2018) Cross-Linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Sci Data* 5:1–10.
53. List JM, Cysouw M, Greenhill S, Forkel R (2018) *Concepticon. A Resource for the Linking of Concept List* (Max Planck Inst Sci Human History, Jena).
54. Huelsenbeck JP (2002) Testing a covariotide model of DNA substitution. *Mol Biol Evol* 19:698–707.
55. Nicholls GK, Gray RD (2008) Dated ancestral trees from binary trait data and their application to the diversification of languages. *J R Stat Soc B* 70:545–566.
56. Bouckaert R, et al. (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10:e1003537.
57. Gavryushkina A, Welch D, Stadler T, Drummond AJ (2014) Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput Biol* 10:e1003919.
58. Rambaut A, Drummond A, Suchard M (2014) Tracer (v. 1.6). Available at beast.community/. Accessed April 5, 2018.
59. Maturana P, Brewer BJ, Klaere S, Bouckaert R (2017) Model selection and parameter inference in phylogenetics using nested sampling. arXiv:1703.05471. Preprint, posted March 16, 2017.
60. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
61. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY (2017) ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8:28–36.
62. Nicholls GK, Ryder RJ, Welch D (2013) TraitLab: A MatLab package for fitting and simulating binary tree-like data.
63. Kelly LJ, Nicholls GK (2017) Lateral transfer in stochastic Dollo models. *Ann Appl Stat* 11:1146–1168.
64. Ryder R (2010) Phylogenetic Models of Language Diversification. DPhil dissertation (Univ Oxford, Oxford).