Methods for the automated dating of the world's language families - with illustrations from Sino-Tibetan

SØREN WICHMANN
LEIDEN UNIVERSITY, BEIJING LANGUAGE UNIVERSITY, KAZAN FEDERAL UNIVERSITY
- WITH CONTRIBUTIONS BY -
TARAKA RAMA
UNIVERSITY OF OSLO

[Symposium on Distance Calculation and Dating of Languages
College of Chinese Language and Culture,
Nankai University, Tianjin, China]

Preview of conclusions

- I will present a new method for dating language families and subgroups based on phylogenies inferred through Bayesian methods.
- The method is fully automated, since the trees are based on automated cognate recognition.
- So far the only other automated method is ASJP Chronology (Holman et al. 2011).
- The new method, which we called Generalized Bayesian Dating, performs as well as ASJP Chronology as far as can be judged, but the results are still different.
- One of the interesting results is a good match between our dates and the dates for Sino-Tibetan recently found by others (Zhang et al. 2019, Sagart et al. 2019).

Background on methods for dating language groups

- Three existing methods
  - Glottochronology
  - Bayesian dating
  - ASJP Chronology

Glottochronology

- Assumes a constant rate of lexical replacement
- Retention rate calibrated based on a small set of known linguistic diversification events, mostly from Indo-European languages

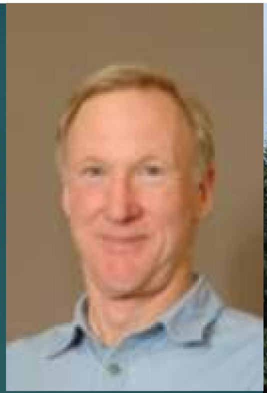
$$t = -\frac{\ln(c)}{2\ln(r)}$$

, where c = percent cognates
r = retention rate

Morris Swadesh and some friends in Mexico

Lees (1953), Swadesh (1955)

ASJP chronology



Eric Holman

Cecil Brown me

- Assumes a constant rate of phonological and lexical replacement
- The replacement rate calibrated based on 52 calibration points (r = -0.84)

$$t = \frac{\log s - \log s_0}{2 \log r}$$

, where s = similarity as measured by a modified edit distance
s_0 = similarity at time zero, a constant
r = retention rate, a constant

Holman et al. (2011)

Bayesian methods for classification, dating, and inference of homeland – two recent papers on Sino-Tibetan

- Submitted to *Nature*
  Jan. 13, 2019
- published
  April 24, 2019

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]*

- Submitted to *PNAS*
  Oct. 19, 2018
- published
  around
  May 6, 2019

PNAS

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)

Data

Zhang et al.

- 109 languages from the STEDT etymological database, 100 Swadesh items, 949 lexical root meanings (≈cognate classes)
- Removal of incompletely or excessively studied languages, and languages believed to have had a lot of borrowing

Sagart et al.

- 50 languages of which 22 from STEDT, 180 basic vocabulary items, 3,333 cognate classes
- "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."

Bayesian phylogenetic methods

Zhang et al.

- Used BEAST2.
- No ancestral or node constraints as priors.
- Methods tried:
  - Covarion + Relaxed LogNormal clock ← BEST FITTING
  - Covarion + strict clock
  - Continuous Time Markov Chains (CTMC) + Relaxed LogNormal clock
  - CTMC + Relaxed LogNormal clock + 4 gamma
  - CTMC + strict clock
  - CTMC + Strict clock + 4 gamma
- Coalescent skyline model as tree prior.

Sagart et al.

- Used BEAST2.
- Sinitic constrained.
- Methods tried:
  - Covarion + Relaxed clock ← BEST
  - Covarion + Strict clock
  - Stochastic Dollo
- Fossilized Birth-Death model as prior.

Methods (cont.)

Zhang et al.

- Give posterior probability values supporting the nodes; in addition, reliability values on internal nodes calculated from four-point analysis.
- Methods run for 50 million generations, sampling every 5,000 generations. The first 10% of the iterations were treated as burn-in, final sample of 10,000 trees.
- Deviation from a tree-like structure checked using delta and Q-residual scores.

Sagart et al.

- Posterior probability values supporting the nodes.
- Methods run for 100 million generations. The first 10% of the iterations treated as burn-on, final sample of 10,000 trees.
- Deviation from a tree-like structure checked by analyzing the data under the reanalyzing a subset of the data under the Lateral Transfer Stochastic Dollo model.

Methods (cont.)

Zhang et al.

- "We performed an *Urheimat* inference for the Sino-Tibetan languages (…). However, the prerequisites for *Urheimat* inference were not satisfied… ()". Too much language extinction in the Han Chinese region, continuous and unidirectional migration along the Tibetan-Yi Corridor throughout the recent 5000 years.

Sagart et al.

- No phylogeographical analysis reported, homeland inference based on 'Wörter und Sachen".

Methods for calibration

Zhang et al.

- Distributions given for all calibration dates, either normal distribution or uniform distribution.
    - Chinese: norm. distr., mean 2700, s.d. 150
    - Old Chinese: norm. distr., mean 2500, s.d. 100
    - Tibetan dialects: norm. distr., mean 1150, s.d. 50
    - Burmese: unif. distr., 400-1200
    - Pumi: norm. distr., mean 750, s.d. 50
    - Yi: norm. distr., mean 1500, s.d. 100
    - Qiangic+rGaylrongic, unif. distr., 1000-5000
    - Karen: norm. distr., mean 1150, s.d. 100

Sagart et al.

- Except for Old Chinese and Chinese it seems that all calibration dates represent a single date rather than a range.
    - Old Chinese: 2800-2300, unif. distr. [Main paper, p. 5]
    - Chinese dialects: 2200-2000, unif. distr. [Supplementary information, p. 17]
    - Old Tibetan: 1200 [Main paper, p. 5]
    - Old Burmese: 800 [Main paper, p. 5]
    - Tangut: 900 [Main paper, p. 5]

Methods from other disciplines

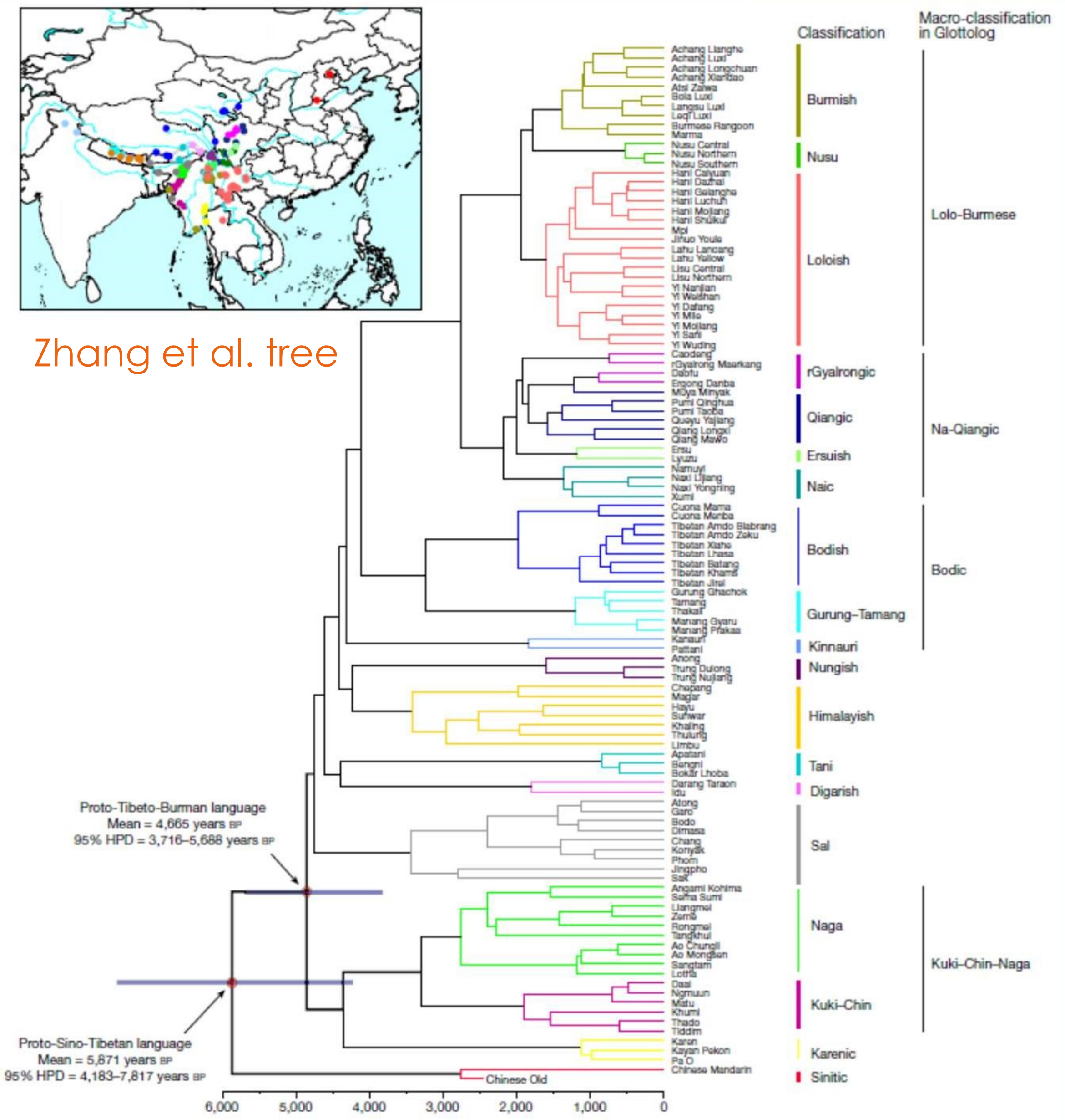
Zhang et al.

- Compare with genetic evidence (e.g. age for Tibeto-Burman from Y chromosome data)
- Compare with archaeological data on numbers of sites

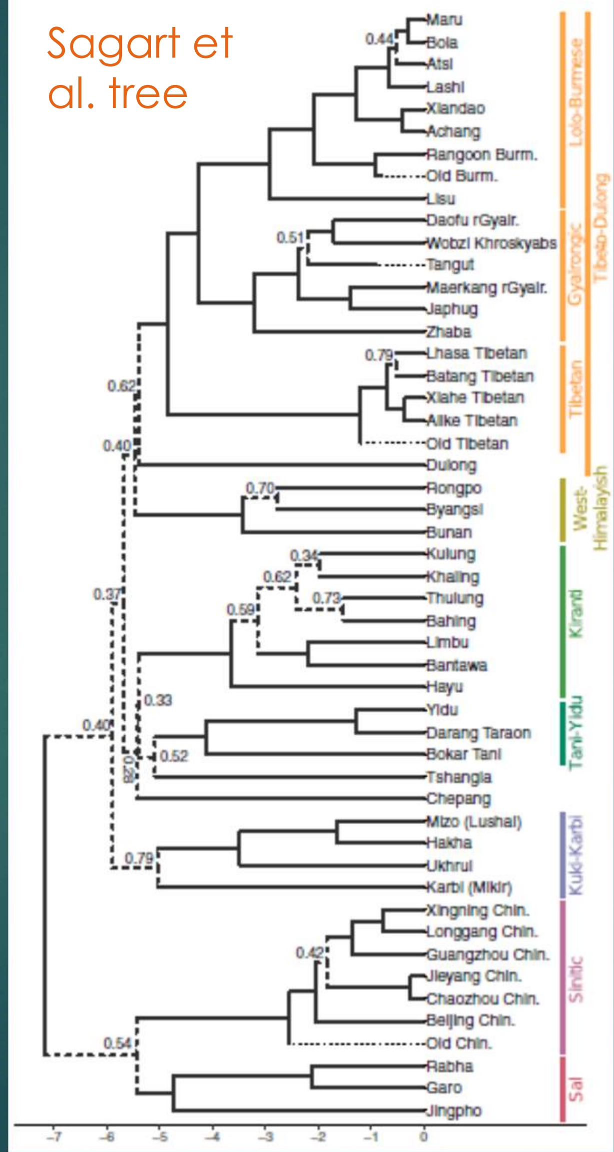
Sagart et al.

- Some discussion of genetic data supporting a north-to-south Sinitic demic diffusions
- Wörter-und-Sachen for domesticated plants and animals

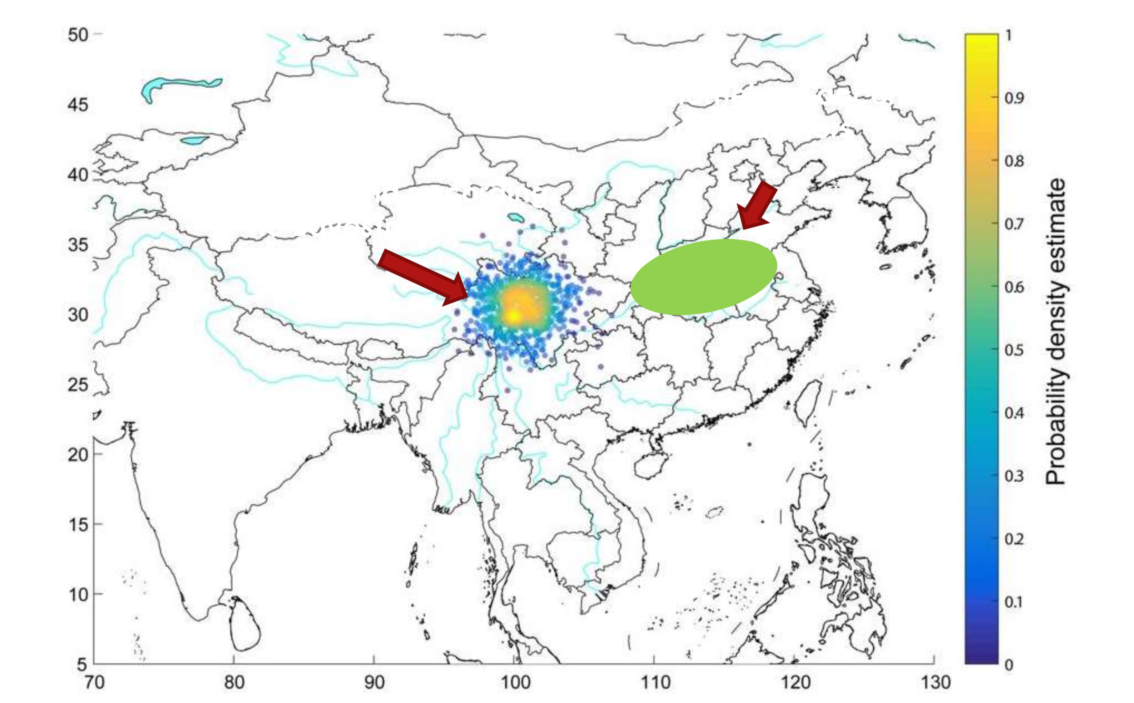
Zhang et al. tree



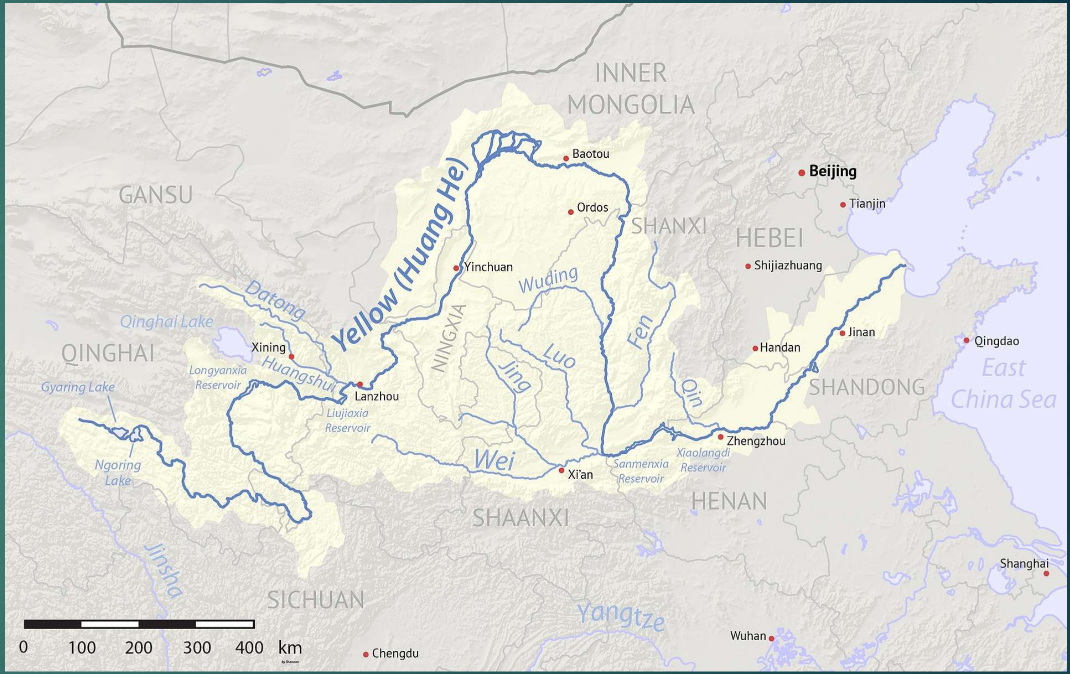
Sagart et al. tree



Homeland, Zhang et al.

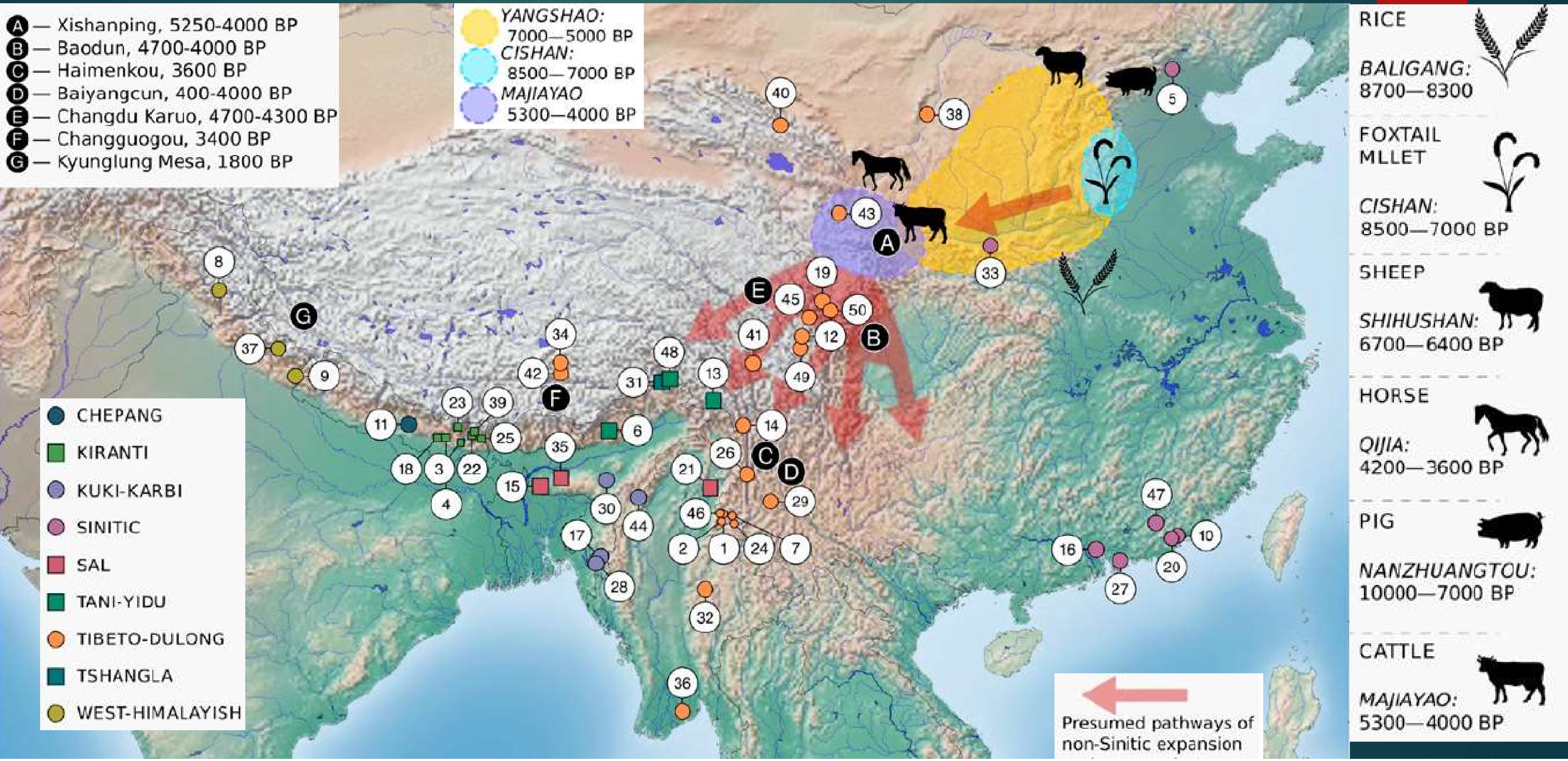


The probability density estimates for the original homeland of the Sino-Tibetan languages via the phylogeographical approach, implemented in BayesTraits package. The iterations in BayesTraits were set to 1,000,000. The sample period was set to 1,000. The first 25% of the iterations were treated as burn-in.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



The Yellow River Basin, supposed origin of Sino-Tibetan

Homeland, Sagart et al.



Summary of results

Zhang et al.

- S.T.: ~7800-~4200 BP, mean: 5871 BP
- Sinitic and Tibeto-Burman sisters
- Origin in Yellow River Basin, northern China, based on match between their date and Yangshao and/or Majiayao Neolithic cultures.

Sagart et al.

- S.T.: 9568-5093 BP, mean: 7184 BP
- Sinitic and Tibeto-Burman sisters (but only 33% probability)
- There are six domesticate names forming cognate sets in at least two of the branches: foxtail millet, pig, sheep, rice plant, cattle, and horse. Since all of these first appear in northern China it assumed that Sino-Tibetan originated in the Yellow River Basin.

Our new method: Generalized Bayesian Dating

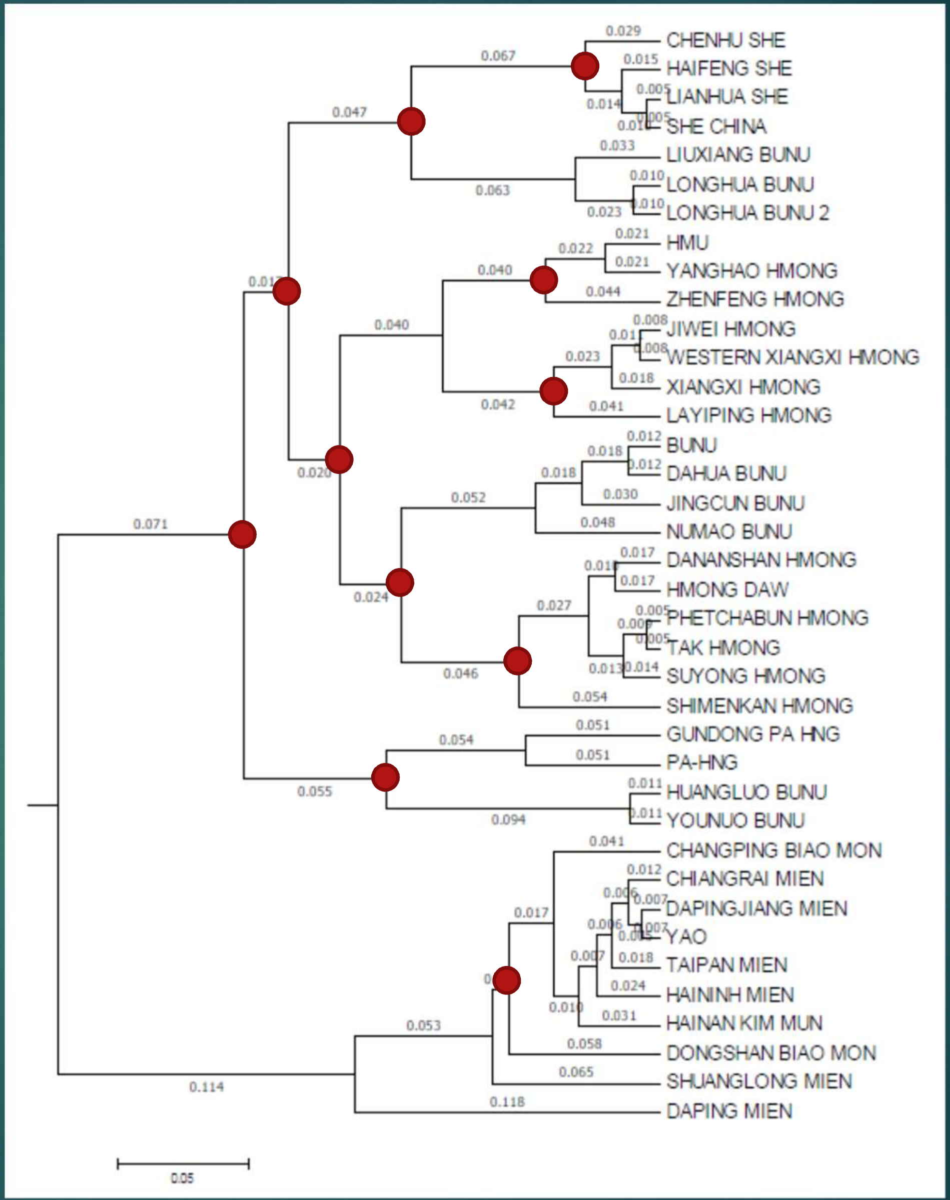
- Selection of doculects
  - should not be extinct
  - at least 28 items attested in the 40-item lists
  - for Indo-European, Austronesian, Atlantic-Congo, Nuclear Trans-New Guinea, and Afro-Asiatic doculects were selected so as to to cover as many branches as possible with a max of 200 doculects
- Selected 30 calibration points from the 52 of Holman et al. (2011)
  - as many as possible
  - no genealogically overlapping groups
- Extracted cognates automatically using a method developed by Taraka Rama
  - word distances, defined by a mixture of criteria, are computed
  - words are clustered according to a so-called Chinese Restaurant algorithm

me Taraka Rama

Our method (cont.)

- Infer Bayesian tree for 116 language families, following the Glottolog classification
  - Birth-death tree prior (same as Sagart et al.)
  - Relaxed clock (same as Sagart et al. and Zhang et al.), where the clock-rate is set to 1
  - Trees are uncalibrated (unlike all other attempts at using Bayesian trees for dating)
  - Constrain topologies using Glottolog; each constrained subgroup should contain at least 3 doculects (using constraints is common, but here we use many more than costumarily)

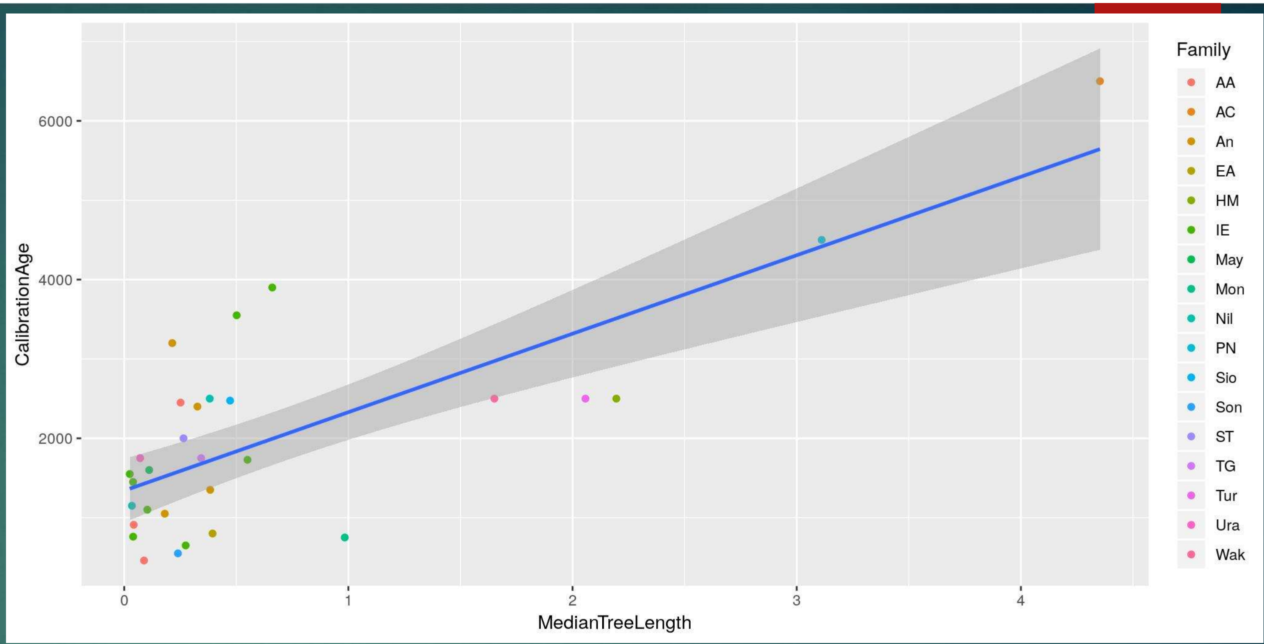
Example of a constrained tree: *Hmong-Mien*



Our method

- Construct trees and correlate tree lengths and dates for 30 language groups

$$age = MTL*989 + 1341$$
$$r = 0.75$$



Language Group	Family	Doculects	Median Tree Length	Calibration age
Oromo	Afro-Asiatic	5	0.0883	460
Maltese-Maghreb Arabic	Afro-Asiatic	3	0.0425	910
Ethiopian-Semitic	Afro-Asiatic	17	0.2514	2450
Benue-Congo	Atlantic-Congo	148	4.3527	6500
East Polynesian	Austronesian	18	0.1809	1050
Temotu	Austronesian	10	0.2143	3200
Inuit	Eskimo-Aleut	6	0.3941	800
Ma'anyan-Malagasy	Austronesian	44	0.3836	1350
Malayo-Chamic	Austronesian	32	0.3266	2400
Tupi-Guarani	Tupian	12	0.3431	1750
Dardic	Indo-European	28	0.5017	3550
Iranian	Indo-European	26	0.6604	3900
Romance	Indo-European	49	0.5498	1729
Brythonic	Indo-European	3	0.0396	1450
Scandinavian	Indo-European	10	0.1030	1100
English-Frisian	Indo-European	3	0.0246	1550
East Slavic	Indo-European	4	0.0398	760
Romani	Indo-European	26	0.2741	650
Cholan	Mayan	5	0.1113	1600
Southern-Nilotic	Nilotic	14	0.3816	2500
Ongamo-Maa	Nilotic	4	0.0344	1150
Chinese	Sino-Tibetan	16	0.2646	2000
Mississippi-Valley-Siouan	Siouan	8	0.4723	2475
Southern-Songhai	Songhay	5	0.2395	550
Saami	Uralic	6	0.0712	1750
Mongolic	Mongolic	8	0.9839	750
Hmong-Mien	Hmong-Mien	38	2.1953	2500
Turkic	Turkic	55	2.0575	2500
Wakashan	Wakashan	6	1.6512	2500
Pama-Nyungan	Pama-Nyungan	68	3.1111	4500

Our method (cont.)

- Calculate the tree length for all families based on the linear formula obtained from correlating tree lengths and median tree length for the calibration points (*age = MTL*989 + 1341*).
- Subgroup ages are calculated similarly obtaining tree lengths for subgroups.
- For young subgroups there is currently the problem that they cannot be younger than 1341 because that is the intercept for the linear formula.
- Taking the 30 calibration points and correlating median tree length and known ages gave r = 0.75, same as for ASJP Chronology.

Our dates compared to published dates based on calibrated Bayesian trees

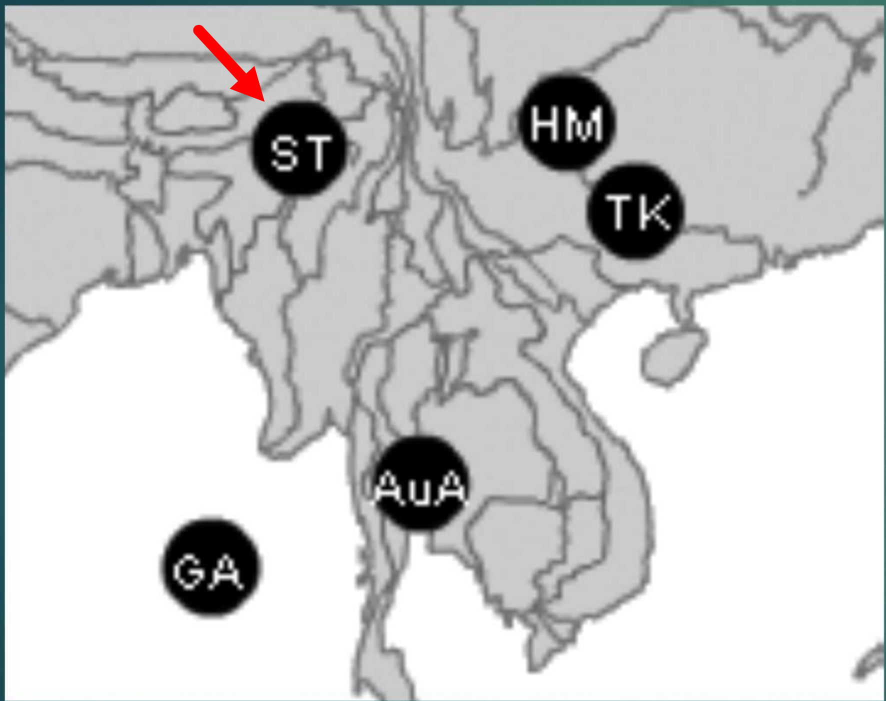
Underlined dates or date ranges fall within the range of dates in published Bayesian analyses

Group	Genera-lized Bayesian date	Other published dates based on Bayesian trees [mean and 95% HPD interval in brackets]	Reference	ASJP chronology (Holman et al. 2011) with interval of ± 29%]	ASJP (current database) [with interval of ± 29%]
Sino-Tibetan	**6301**	**5900 [7800-4200]** **7184 [9568-5093]**	Zhang et al. (2019) Sagart et al (2019)	**5261 [6787-3735]**	**5059 [6526-3592]**
Austronesian	**7035**	**5230 [5800-4750]**	Gray et al. (2009)	**3633 [4687-2579]**	**3706 [4781-2631]**
Dravidian	**3977**	**4500 [6500-3000]**	Kolipakam et al. (2018)	**2055 [2651-1459]**	**2358 [3042-1674]**
Bantu	**3274**	**4800 [4709-4985]**	Grollemund et al. (2015)	**4867 [6278-3456]**	**4644 [5991-3297]**
Core Indo-European	**5338**	**~4800**	Chang et al. (2015)	**4348 [5609-3087]**	**4134 [5333-2935]**
Turkic	**3376**	**2408 [3394-1279]**	Hruschka et al. (2015)	**3404 [4391-2417]**	**3406 [4394-2418]**
Pama-Nyungan	**4418**	**5671 [6966-4455]**	Bouckaert et al. (2018)	**4295 [5541-3049]**	**4369 [5636-3102]**

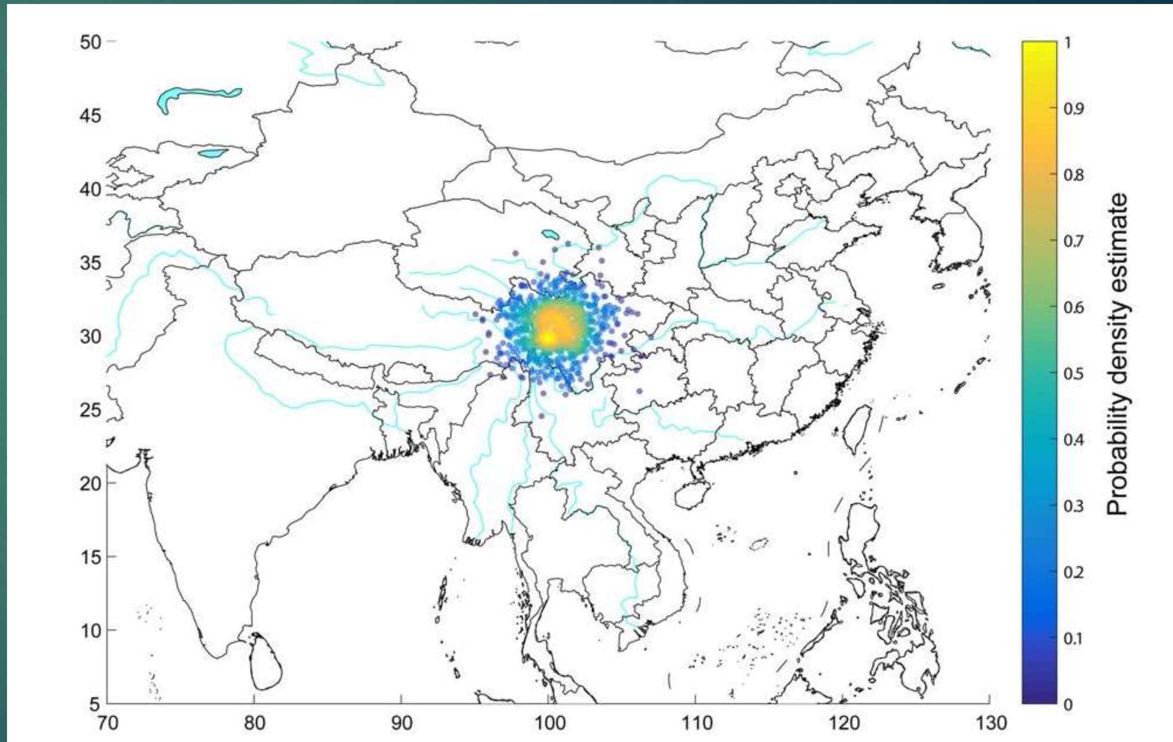
Advantages of automated methods like Generalized Bayesian Dating and ASJP Chronology

- Not dependent on internal calibration points.
- ASJP chronology has no limits on the number of languages that can be handled, Generalized Bayesian Dating more limited but generally fast because nodes are constrained and need not be inferred.
- The methods can be evaluated against known ages across many datapoints because the methods don't change settings from one family to the next.

Sino-Tibetan homeland according to Wichmann et al. (2010)



Wichmann et al. (2010)



Zhang et al. (2019)

Conclusions regarding the Sino-Tibetan case study

- Zhang et al. and Sagart et al. reach similar conclusions:
  - overlapping age range
  - weak evidence for a primary split Chinese vs. Tibeto-Burman
  - phylogeographical methods are either not used (Sagart et al.) or their results discarded (Zhang et al.), instead adopting evidence from archaeology.
- Our Generalized Bayesian dating gives results similar to Zhang et al. and Sagart et al.

Conclusions regarding the Sino-Tibetan case study (cont.)

- ASJP Chronology's ages also similar: within the range of Zhang et al. and Sagart et al.
- Inferring the homeland using the method of Wichmann et al. (2010) points to roughly the same region as Bayesian phylogeography. It is similarly not robust against directed migration and language extinction.

Future research

- Work on possible improvements of Generalized Bayesian dating.
- Expand the ASJP database to with more data, including more 100-item word lists.
- Make ASJP methods more user-friendly.

Topic for the following demo!

References

- Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52.6: 841-875.
- Lees, Robert. 1953. The basis of glottochronology. *Language* 29.2: 113–127.
- Rama, Taraka. 2018. Similarity dependent Chinese Restaurant Process for cognate Identification in multilingual wordlists. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 271–281.
- Sagart, Laurent, Guillaume Jacques, Yunfan Laib, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of sciences of the U.S.A.* doi: 10.1073/pnas.1817972116. [Epub ahead of print].
- Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121–137.
- Wichmann, Søren, André Müller, and Viveka Velupillai. 2010. Homelands of the world's language families: A quantitative approach. *Diachronica* 27.2: 247-276.
- Zhang, Menghan, Shi Yan, Wuyun Pan, Li Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* 569: 112–115.