

Triangulation reduces the polygon of error for the history of Transeurasian

Martine Robbeets^{1,2*}, Mark Hudson^{1,3*}, Chao Ning^{1,4*}, Remco Bouckaert^{1,5}, Alexander Savelyev^{1,6}, Geonyoung Kim⁷, Tao Li^{1,8,9}, Sofia Oskolskaya^{1,10}, Ilya Gruntov^{6,11}, Olga Mazo^{6,11}, Seongha Rhee^{12,13}, Kyou-Dong Ahn¹², Ricardo Fernandes^{1,14,15}, Ken-ichi Shinoda¹⁶, Hideaki Kanzawa-Kiriyama¹⁶, Rasmus Bjørn¹, Bingcong Deng¹, Deog-im An¹⁷, John Bentley¹⁸, Takamune Kawashima¹⁹, Joanna Dolińska¹

1 Max Planck Institute for Geoanthropology, Jena, Germany

2 Johannes Gutenberg Universität, Mainz, Germany

3 Institut d'Asie Orientale, ENS de Lyon, France

4 School of Archaeology and Museology, Peking University, Beijing, China

5 Centre of Computational Evolution, University of Auckland, Auckland, New Zealand

6 Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

7 Department of Archaeology and Art History, Seoul National University, Seoul, South Korea

8 Department of Archaeology, Wuhan University, Wuhan, China

9 Archaeological Institute for Yangtze Civilization, Wuhan University, China.

10 Institute for Linguistic Studies, Russian Academy of Sciences, Saint Petersburg, Russia

11 National Research University Higher School of Economics, Moscow, Russia

12 Hankuk University of Foreign Studies, Seoul, South Korea

13 Faculty of Liberal Arts Mahidol University, Thailand

14 Faculty of Arts, Masaryk University, Czechoslovakia

15 School of Archaeology, University of Oxford, UK

16 National Museum of Nature and Science, Department of Anthropology, Tsukuba, Japan

17 Department of Conservation of Cultural Heritage, Hanseo University, Seosan, Korea

18 Department of World Languages and Cultures, Northern Illinois University, USA

19 Hiroshima University Museum, Higashi-Hiroshima, Japan

*corresponding authors: robbeets@shh.mpg.de; hudson@shh.mpg.de; cning@pku.edu.cn

Abstract

In a recent study we used an interdisciplinary approach combining linguistics, archaeology and genetics to analyse the Transeurasian languages¹. Our analysis concluded that the early dispersals of these languages were driven by agriculture. A preprint published on this server presents objections to the Transeurasian hypothesis and its association with farming dispersals². However, close inspection of that text reveals numerous misinterpretations and inconsistencies. In the interest of furthering scientific debate over Transeurasian language and population history, we address the critiques, revising datasets and fine-tuning approaches. The linguistic critique questions the quantity and quality of our datasets. Here we show that the number of surviving cognate sets for Transeurasian is in line with that for well-established language families. In addition, we find that Tian et al.'s failure to reject a core of regularly corresponding cognates in the basic vocabulary creates ground for a consensus about the genealogical relatability of the Transeurasian languages. The archaeological critique attempts a re-analysis of one Bayesian test using re-scored data only for northern China. Over half of the suggested re-scorings contain inconsistencies and it is not explained why the re-analysis retains the original data for sites outside northern China, comprising almost 60% of the total. More importantly, the sweeping claim that there is no evidence supporting the prehistoric migrations analysed in our study is not backed by any discussion of the archaeological record. With respect to genetics, the preprint claims a re-analysis showing that the data 'do not conclusively support the farming-driven dispersal of Turkic, Mongolian, and Tungusic, nor the two-wave spread of farming to Korea.' In fact, the only genetic re-analysis presented is limited to samples from Korea and Japan and does not contradict our original conservative modelling of Neolithic individuals with Hongshan and our Bronze Age ones with Upper Xiajiadian. In sum, in bringing multiple lines of evidence together through triangulation, we gained a more balanced and richer understanding of Transeurasian dispersals than each discipline could provide individually. Our research doubtless leaves

¹ Robbeets, M. et al. Triangulation supports agricultural spread of the Transeurasian languages. *Nature* **599**, 616-621 (2021).

² Tian, Z. et al. Comment. "Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative" (2022) BioRxiv preprint posted June 12, 2022, <https://doi.org/10.1101/2022.06.09.495471>

room for improvement but we remain confident that triangulation did not ‘fail’, but rather brought us a step closer to understanding the history of the Transeurasian languages.

Linguistics

Our Article used the classical historical comparative linguistic method to propose the genealogical relatedness of the Transeurasian languages. We then combined Bayesian phylogenetics with classical approaches to draw inferences about tree branching, dating and homelands. The critique that ‘the number of cognate sets proposed for Transeurasian is exceptionally low’ is based on comparison of our cognate set distributions with those of long-established language families, such as Sino-Tibetan and Indo-European. However, the authors ignored the full body of evidence in our traditional linguistic approach indicated in red in Table 1 and counted only the portion of semantically equivalent cognates used in our Bayesian analysis. In fact, the distribution of cognates is completely in line with expectations, considering that Transeurasian has an older age, fewer surviving daughter languages and early attestations of a relatively shallow time-depth, so that traces of cognancy are more blurred than in the other families.

Table 1: Distribution of cognates across a collection of datasets. Adapted from [2].

	Transeurasian Robbeets et al. 2021	Sino-Tibetan Sagart et al. 2019	Sino-Tibetan Zhang et al. 2019	Indo-European Dunn & Tresoldi 2021
Major subgroups	5	7	10	10
Concepts	254	250	100	207
Cognates in 60% of subgroups	38 (0.15) 39 (0.15)	47 (0.19)	19 (0.19)	52 (0.25)
Cognates in 80% of subgroups	5 (0.02) 14 (0.05)	23 (0.09)	11 (0.11)	32 (0.15)
Cognates in 100% of subgroups	2 (0.01) 5 (0.02)	19 (0.08)	7 (0.07)	12 (0.06)
Age	9200 BP	7200 BP	5900 BP	7500 BP
Surviving languages	Ca. 75	Ca. 500	Ca. 500	Ca. 450
Earliest Historical attestation	Ca. 1300 BP	Ca. 3000 BP	Ca. 3000 BP	Ca. 3000 BP

Furthermore, the criticism that ‘Out of the 3139 cognate sets listed, only 317 are found in more than one language family’ visualised in their Figure 1a is misleading because it is common in Bayesian cognate alignments for the bulk of cognates to be singletons; Sino-Tibetan and Indo-European would yield a similar graphic.

The authors claim we ‘ignore [our] own principles of historical language comparison and bend the evidence to [our] needs.’ However, in line with the requirements of the historical comparative method, all compared sounds correspond regularly, except for some root-final vowels (see SI for correspondences in each cognate set). Our item-by-item treatment of Tian et al.’s alleged ‘irregularities in the sound correspondences’ reveals that only 2 basic etymologies involve flaws, while other comments are open to interpretation.

Moreover, it is generally known³ that computer-assisted cognate detection methods, such as the one applied by the authors, still perform worse than expert analysis and that the algorithm fails when the datasets are too small. Therefore, it is misleading to apply these methods to only a small fraction of our evidence (46 out of 160 etymologies; see SI).

As with any work of massive scope, differences of interpretation are inevitable. Pokorny’s corpus of Indo-European data has been continuously reworked since the 1920s^{4,5,6}.

Nevertheless, the availability of a publicly available dataset that can be further improved aims to push the field forward.

Archaeology

The archaeological critique focuses on one quantitative Bayesian test, ignoring our qualitative analyses which provide key support for the overall conclusions. Tian et al. attach a file which re-scores ‘missing data’ for selected categories. However, examination of their data shows inconsistencies for over half of the suggested re-scorings (SI). Moreover, data are re-scored only for northern China and not for all regions in our database. These issues lend little confidence to the new tree found in the Comment.

Although the authors justly point out that ‘different subsets of the data have different histories’, taken together, the entire body of cultural features can nevertheless contribute to an

³ Rama, T. et al. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?. *Proceedings of NAACL-HLT* **2018**, 393–400 (2018).

⁴ Walde, A. & Pokorny, J. *Vergleichendes Wörterbuch der indogermanischen Sprachen*, 3 vols. (De Gruyter, 1927–1932).

⁵ Pokorny, J. *Indogermanisches etymologisches Wörterbuch*, 2 vols. (Francke, 1948–69)

⁶ Lubotsky, A. (ed.) *Leiden Indo-European Etymological Dictionary Series*, 12 vols. (Brill, 1991–2021).

overall history.⁷ Their suggestion of inferring trees inside trees is appealing but computationally not yet feasible at this time.⁸

The second issue about the appropriateness of phylogenetic trees to represent archaeological data challenges research — including that by a member of Tian’s team — finding that certain levels of horizontal transmission do not invalidate phylogenetic approaches to cultural evolution.^{7,9,10} Recognizing that the phylogenetic analysis of our archaeological dataset does not only capture a signal of descent but also one of interaction, our Article used clustering in the Bayesian tree as an indication of cultural similarity between different archaeological assemblages. This result was broadly consistent with our qualitative analysis and with the large body of archaeological literature which models agricultural expansions from northeast China into Manchuria^{11,12}, the Primorye^{13,14,15,16,17}, Korea^{18,19,20}, and Japan^{21,22,23,24}. While

⁷ Greenhill, S.J., Currie, T.E., Gray, R.D. Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299-2306 (2009).

⁸ Evans, C.L., Greenhill, S.J., Watts, J., List, J.-M., Botero, C.A., Gray, R.D., Kirby, K.R. The uses and abuses of tree thinking in cultural evolution. *Phil. Trans. R. Soc. B* **376**, 20200056 (2021).

⁹ Gray, R.D. & Jordan, F.M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055 (2000).

¹⁰ Currie, T.E., Greenhill, S.J., Mace, R. Is horizontal transmission really a problem for phylogenetic comparative methods? A simulation study using continuous cultural traits. *Phil. Trans. R. Soc. B* **365**, 3903-3912 (2010).

¹¹ Stevens, C. & Fuller, D.Q. The spread of agriculture in eastern Asia: archaeological bases for hypothetical farmer/language dispersals. *Lang. Dyn. Chang.* **7**, 152-186 (2017).

¹² Leipe, C. et al. Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics. *Sci. Adv.* **5**, eaax6225 (2019).

¹³ Yan, W. Dongbeiya nongye de fasheng yu chuanbo [Beginning and dispersal of agriculture in Northeast Asia]. *Nongye Kaogu* **3**, 37-44 (1993).

¹⁴ Vostretsov, Y.E. The interaction of marine and agricultural adaptations in the basin of the Sea of Japan, in *The Russian Far East in Prehistory and the Middle Ages* (ed Andreeva, Z.V.) 159-186 (Vladivostok: Dalnauka, 2005) (in Russian).

¹⁵ Kuzmin, Y.V. The beginnings of prehistoric agriculture in the Russian Far East: current evidence and concepts. *Documenta Praehist.* **40**, 1-12 (2013).

¹⁶ Miyamoto, K. The initial spread of early agriculture into Northeast Asia. *Asian Archaeol.* **3**, 1-12 (2014).

¹⁷ Li, T. et al. Millet agriculture, dispersed from Northeast China to the Russian Far East: integrating archaeology, genetics and linguistics. *Archaeol. Res. Asia* **22**, 100177 (2020).

¹⁸ Nelson, S.M. *The Archaeology of Korea* (Cambridge Univ. Press, 1993).

¹⁹ Lee, G.A. The spread of domesticated plant resources in prehistoric northeast Asia, in *Routledge Handbook of Archaeology and Globalization* (ed Hodos, T.) 394-412 (Routledge, 2017).

²⁰ Hudson, M.J. & Robbeets, M. Archaeolinguistic evidence for the farming/language dispersal of Koreanic. *Evol. Hum. Sci.* **2**, e52 (2020).

²¹ Miyamoto, K. *Nōkō no kigen o saguru: Ine no kita michi* [In search of the origins of agriculture: the northern route of rice]. (Yoshikawa kōbunkan, 2009).

²² Miyamoto, K. The emergence of ‘Transeurasian’ language families in Northeast Asia as viewed from archaeological evidence. *Evol. Hum. Sci.* **4**, e3 (2022).

²³ Ahn, S.M. The emergence of rice agriculture in Korea: archaeobotanical perspectives. *Archaeol. Anthropol. Sci.* **2**, 89-98 (2010).

²⁴ Leipe, C. et al. The spread of rice to Japan: insights from Bayesian analysis of direct radiocarbon dates and population dynamics in East Asia. *Quat. Sci. Rev.* **244**, 106507 (2020).

debate has revolved around scale, timing and routes, these spreads themselves are widely accepted. Tian et al.'s sweeping claim that there is no archaeological evidence supporting these migrations goes completely against the existing literature, yet no archaeological evidence is provided in support.

Genetics

Against our suggestion that the first farmers in the West Liao River area were of Amur-like genetic profile, Tian et al. object that there was long-term genetic continuity of Amur-related hunter-gatherer ancestry in Northeast Asia, which predated agriculture. However, our Article does not question this continuity but instead shows that while the genetic profile of Early Neolithic millet farmers in the West Liao Basin was originally Amur-like, it received gradual gene flow from the Yellow River Basin. The authors assert that '[t]his conjecture is contradicted by their claim that both the spread of Japonic and that of Koreanic were induced by West Liao River farmer ancestry with Yellow River farmer admixture.' However, there is no contradiction because the originally Amur-like farmers in the West Liao Basin had become admixed with 60 % Yellow River-related ancestry by 5500 BP, the time associated with the dispersal of millet agriculture to Korea. That explains why Neolithic Koreans and Bronze Age Japanese can be modelled as an admixture of Amur, Yellow River and Jomon genomes, a finding confirmed in new research.²⁵

Finally, the authors object to our distinction between Hongshan and Upper Xiajiadian as a genetic model for our ancient Korean and Japanese populations. However, both our Article (SI 13) and extended analyses (SI: Fig. 1-6) confirm that the precise mainland East Asian ancestry source for ancient Koreans and Japanese cannot be well distinguished by current genetic data. If models cannot distinguish which population provides the best ancestry source for the target populations, in population genetics the principle is to select a population which is close to the target population both in time and space. We therefore modelled our Neolithic individuals with Hongshan and our Bronze Age ones with Upper Xiajiadian. Even if our data lack resolution to identify the precise source, they still allow us to associate the spread of farming to Korea with different waves of eastward gene flow from mainland East Asia. The

²⁵ Cooke, N. et al. Ancient genomics reveals tripartite origins of Japanese populations. *Science Advances* 7: eabh2419 (2021).

increasing Yellow river component in our Bronze Age samples can be associated with rice farming, an observation supported by archaeological and linguistic findings (see SI).

Triangulation

Tian et al. insist that our methods ‘fail to replicate’ but concentrate on replicating a single Bayesian test, ignoring the multiple other approaches in our paper. Replication of a single test may lead to repeating the same methodological error and thus provide an unfounded sense of certainty.²⁶ We believe that an essential protection against such flaws is triangulation. In contrast to the authors’ claim that we ‘neither define nor describe the method of “triangulation”’, the concept is well described in the methodology section and illustrated in the supplementary information of our Article. Triangulation is the strategic use of multiple disciplines, methods and datasets to address one question in order to enhance the validity and credibility of research findings. Bringing the multiple lines of evidence together, we gained a more balanced and richer understanding of Transeurasian dispersals than each of them could provide individually. Our research doubtless leaves room for improvement but we remain confident that triangulation did not ‘fail’.

²⁶ Munafò, M.R. & Smith, G.D. Repeating experiments is not enough. *Nature* **553**, 399-401. (2018)