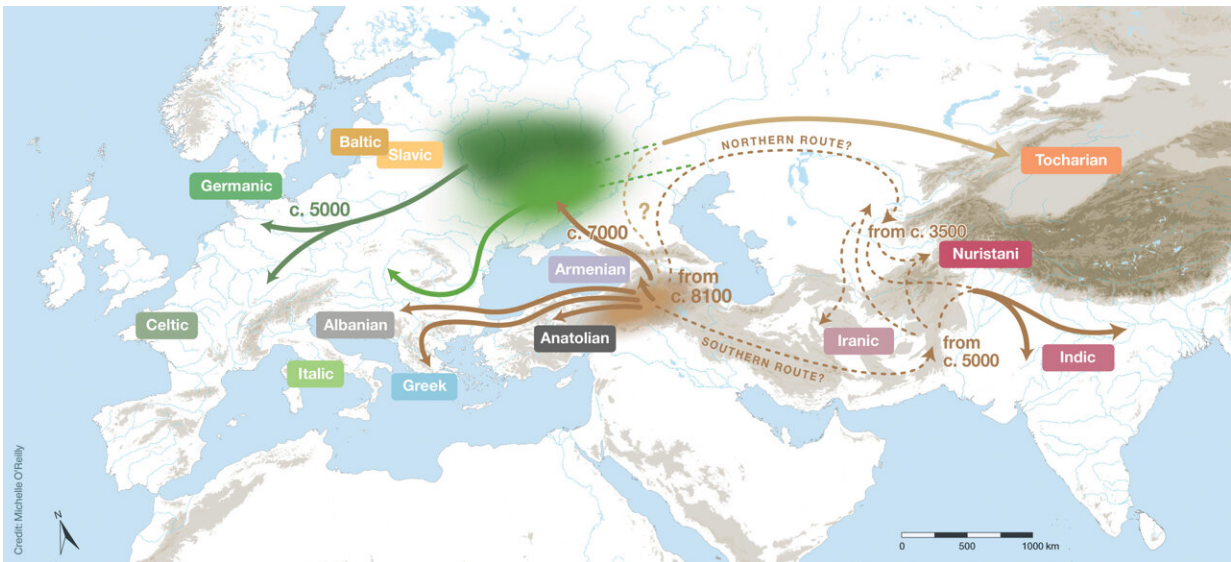# New insights into the origin of the Indo-European languages

July 27 2023

The language family began to diverge from around 8,100 years ago, out of a homeland immediately south of the Caucasus. One migration reached the Pontic-Caspian and Forest Steppe around 7,000 years ago, and from there subsequent migrations spread into parts of Europe around 5,000 years ago. Credit: P. Heggarty et al., *Science* (2023)

An international team of linguists and geneticists led by researchers from the Max Planck Institute for Evolutionary Anthropology in Leipzig has achieved a significant breakthrough in our understanding of the origins of Indo-European, a family of languages spoken by nearly half of the world's population. The work is published in the journal *Science*.

For over two hundred years, the origin of the Indo-European languages has been disputed. Two main theories have recently dominated this debate: the "Steppe" hypothesis, which proposes an origin in the Pontic-Caspian Steppe around 6,000 years ago, and the "Anatolian" or "farming" hypothesis, suggesting an older origin tied to early agriculture around 9,000 years ago.

Previous phylogenetic analyses of Indo-European languages have come to conflicting conclusions about the age of the family, due to the combined effects of inaccuracies and inconsistencies in the datasets they used and limitations in the way that phylogenetic methods analyzed ancient languages.

To solve these problems, researchers from the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology assembled an international team of over 80 language specialists to construct a new dataset of core vocabulary from 161 Indo-European languages, including 52 ancient or historical languages. This more comprehensive and balanced sampling, combined with rigorous protocols for coding lexical data, rectified the problems in the datasets used by previous studies.

Indo-European estimated to be around 8,100 years old

The team used recently developed ancestry-enabled Bayesian phylogenetic analysis to test whether ancient written languages, such as Classical Latin and Vedic Sanskrit, were the direct ancestors of modern Romance and Indic languages, respectively.

Russell Gray, Head of the Department of Linguistic and Cultural Evolution and senior author of the study, emphasized the care they had taken to ensure that their inferences were robust. "Our chronology is robust across a wide range of alternative phylogenetic models and

sensitivity analyses," he stated. These analyses estimate the Indo-European family to be approximately 8,100 years old, with five main branches already split off by around 7,000 years ago.

These results are not entirely consistent with either the Steppe or the farming hypotheses. The first author of the study, Paul Heggarty, observed that, "Recent ancient DNA data suggest that the Anatolian branch of Indo-European did not emerge from the Steppe, but from further south, in or near the northern arc of the Fertile Crescent—as the earliest source of the Indo-European family. Our language family tree topology, and our lineage split dates, point to other early branches that may also have spread directly from there, not through the Steppe."

New insights from genetics and linguistics

The authors of the study therefore proposed a new hybrid hypothesis for the origin of the Indo-European languages, with an ultimate homeland south of the Caucasus and a subsequent branch northwards onto the Steppe, as a secondary homeland for some branches of Indo-European entering Europe with the later Yamnaya and Corded Ware-associated expansions.

"Ancient DNA and language phylogenetics thus combine to suggest that the resolution to the 200-year-old Indo-European enigma lies in a hybrid of the farming and Steppe hypotheses," said Gray.

Wolfgang Haak, a Group Leader in the Department of Archaeogenetics at the Max Planck Institute for Evolutionary Anthropology, says, "Aside from a refined time estimate for the overall language tree, the tree topology and branching order are most critical for the alignment with key archaeological events and shifting ancestry patterns seen in the ancient human genome data. This is a huge step forward from the mutually exclusive, previous scenarios, towards a more plausible model

that integrates archaeological, anthropological and genetic findings."

More information: Paul Heggarty et al, Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages, *Science* (2023). DOI: 10.1126/science.abg0818. www.science.org/doi/10.1126/science.abg0818

Provided by Max Planck Society