aYChr-DB: a database of ancient human Y haplogroups

Laurence Freeman[1], Conrad Brimacombe[1,2], Eran Elhaik[1,3*]

[1] University of Sheffield, Department of Animal and Plant Sciences, Sheffield, UK.
[1,2] University of Bristol, Department of Archaeology and Anthropology, Bristol, UK.
[1,3] Lund University, Department of Biology, Lund, Sweden, 22362

* Please address all correspondence to Eran Elhaik at eran.elhaik@bio.lu.se

Keywords: Paleogenomics, Y chromosome, Y haplogroups, Ancient DNA, aYChr-DB

Abstract

Ancient Y-Chromosomal DNA is an invaluable tool for dating and discerning the origins of migration routes and demographic processes that occurred thousands of years ago. Driven by the adoption of high-throughput sequencing and capture enrichment methods in paleogenomics, the number of published ancient genomes has nearly quadrupled within the last three years (2018-2020). Whereas ancient mtDNA haplogroup repositories are available, no similar resource exists for ancient Y-Chromosomal haplogroups. Here, we present aYChr-DB – a comprehensive collection of 1797 ancient Eurasian human Y-Chromosome haplogroups ranging from 44,930 BC to 1945 AD. We include descriptors of age, location, genomic coverage, and associated archaeological cultures. We also produced a visualisation of ancient Y haplogroup distribution over time. The aYChr-DB database is a valuable resource for population genomic and paleogenomic studies.

Introduction

The genomic history of populations is a tapestry of undirected changes as no population remains immutable over time. Whereas coalescent and other reconstruction methods that rely on modern populations are inaccurate and carry a high risk of misinterpretation (1), analysing the DNA of ancient human populations allows capturing past events as they were. Combining this evidence with environmental, cultural, and other genomic information enables a more accurate representation of the past (2).

The Y-Chromosome contains the largest nonrecombining block in the human genome (3). Using both traditional methods (e.g., PCR) and high throughput sequencing, haplogroups of ancient individuals are identifiable, allowing the study of past genetic diversity (2). Combining Y-DNA with radiocarbon dating also provides a means to map Y-chromosomes onto a phylogenetic tree, which can be used to assess whether previous reports of ancestral variation based on modern DNA are supported by ancient samples and if we can find representatives of ancient clades which no longer exist (4).

Over the past two years, ancient Y-chromosomal data have begun to accumulate rapidly. Published data from the period 2007-2017 (480 Y chromosomes) was nearly quadrupled within the next three years 2018-2020 (1797 Y chromosomes) (Table S1). In concert with mitochondrial DNA, Y-Chromosomal DNA has been used to study the origins of present-day and ancient Eurasians along with their languages (5-8) and disease prevalence (2).

Only a handful of ancient DNA databases have been compiled to date, such as the Online Ancient Genome Repository (https://www.oagr.org.au) – which primarily stores samples sequenced by the Australian Centre for Ancient DNA – and the AmtDB (9), which predominantly features ancient mtDNA. The lack of a dedicated database focusing on the collection of ancient Y-Chromosomal data has impeded research in the field and prompted us to develop aYChr-DB.

aYChr-DB collates a large proportion of the published Eurasian ancient Y-DNA data over the past 13 years (2007-2020) into an easily accessible archive. The manually curated database not only standardizes the reporting of data and makes haplogroup comparison feasible but also offers socio-cultural annotation. The genomic sequences are available through the source studies.

Materials & Methods

Relevant papers were identified by querying PubMed and Google Scholar with the key words “ancient Y”, “ancient haplogroup” and “ancient DNA” + “Y chromosome”. Both reviews and research articles were selected, with no restrictions on date of publication or journal of publication. Records were then manually curated to remove duplications. Maps were drawn using the ggmap R package (10).
aYChr-DB (Table S1) is publicly and freely accessible at https://github.com/eelhaik/aYDB.

Results

aYChr-DB contains 1797 samples (Table S1). Multiple descriptors are available for each sample, which are named according to the official/published ID, such as country and location. The age of the sample, where applicable, is provided in both BC and BP calibrated from 1950. Carbon-dated samples are shown as calBC/BP. For samples without published coordinate data, we provide coordinates based on location names and descriptions. The archaeological period of each sample has been assigned based on age and location. Where given, average genomic coverage has been included. The comments section clarifies additional information on the samples which may be pertinent to database users.

We produced a visualisation of the aYChr-DB – for a total of 1723 samples after removing 74 undated samples (Figures 1 and S1). The full 1797 samples were included in the main 'all time periods' map. For coherency, haplogroups were trimmed to 3 letters at most, (i.e., - R1a1a1 is shown as R1a). Samples were classified into one of six periods, spanning the range of published dates, using the age or average age of the sample. Several trends are noteworthy. A large proportion (65.5%) of collected ancient samples are dated between 0 and 4999 BC. R1b is the modal haplogroup in the ancient Eurasian samples, accounting for 22.3% of the data. I2a is the second most common at 13.9%, followed by G2a at 11.3%, and R1a at 7.1%. That the majority of the samples are located in Europe is likely due to the availability of large depositories and history of archaeological research in this region and its propensity for cool, temperate conditions suitable for the preservation of ancient DNA (11). Over 40% of the samples were found in four countries: Spain (11.4%), Russia (10.4%), Hungary (9.7%) and Italy (9.6%).

The major challenge in our efforts to provide coherent and useful annotation was in ascribing meaningful cultural information to the samples. Delineating between archaeological cultural phases is inherently made ambiguous by the plurality of geographic boundaries and significantly different chronologies within each region. Major Eurasian regions vary in their definitions in the archaeological literature, yet these regions are often interconnected. In West Asia, particularly areas such as the Levant and Anatolia, there are shared origins for several European cultural phrases, and West Asian societies undergo the same general transitions, albeit earlier (12). European prehistoric periods are conventionally defined by technological innovations, excepting the Paleolithic-Mesolithic transition, which is a climate transition. The primary European cultural phases are the Neolithic, Copper Age, Bronze Age, and Iron Age, followed by historic periods such as the Romans and Medieval times. Up to the Bronze Age within Europe and West Asia, this technological framework is useful for geneticists as it often corresponds well with major shifts in population structure because these technologies enabled certain groups to move into adjacent regions. The Iron Age and beyond are characterised by advanced civilizations across Europe and West Asia while in the colder and less fertile regions of Central and Northeastern Asia, nomadic, and hunter-gatherer

lifestyles persisted in a scattering of small populations across a broad expanse of territory (13). These people often possessed iron and bronze technologies but with no sedentary agricultural base and demonstrating high mobility. Their cultures have been challenging to classify archaeologically in terms of any overarching technological or historical framework.

In East Asia, we can observe a parallel, although typically not synchronous, development of agriculture, copper/bronze technology, and eventually iron (14). The transition to agriculture does correspond with population movement (15,16) and is a pattern demonstrated throughout the region. However, subsequent archaeological transitions are usually referred to through dynastic change rather than technological change (17). This is particularly true within China and adjacent regions, despite migration associated with these technological shifts proven at a genetic level (16).

To account for the different regions where these transitions occurred at slightly different times and patterns, we have broken down the regions into Europe, European Steppe, Central Asia, West Asia, and China. We left China as a single 'region' because it is the only representative of East Asian societies in our database and because of the significant technology of East Asia usually started in China. We distinguished Europe and European Steppe because from a cultural and genetic standpoint; these represent different phases of the Bronze Age and prior Neolithic cultures. We noted the movement of Iranian people in and out of the steppe regions, although we note that little is known about many Iranian populations (e.g., the Xiongnu people). Historical periods are indicated as "Historic." Where no culture is given, and only a date exists, we used the culture that was in the location at that time.

Several trends are noteworthy within the database (Table S1). More than 75% of the samples are dated between 4,999 – 0 BC. R1b is the modal haplogroup in the ancient Eurasian samples, accounting for 29.2%. I2a is the second most common at 15.6%, followed by G2a at 8.1%. A large majority of the samples are located in Europe – in particular Great Britain, Hungary, and Spain – likely due to the availability of large depositories and history of archeological research in this region. Europe is also characterised by cool temperate conditions suitable for the preservation of ancient DNA (11). However, subsequent archaeological transitions are usually referred to through dynastic change rather than technological change (17). This is particularly true within China and adjacent regions, despite migration associated with these technological shifts proven at a genetic level (16).

Discussion

We developed a database of ancient Eurasian Y-Chromosomal haplogroups, collating published data from the last twelve years. We assigned missing descriptors to many samples and provided a socio-cultural annotation, which contributes to the uniqueness and usefulness of this resource. Finally, a geographical visualisation of the data provides a convenient review of the samples at discrete intervals.

Version 1.0 of the database includes samples from across Eurasia due to the rarity of ancient Y haplogroups from elsewhere. The database will be updated periodically with recently published Y-Chromosome data. We expect that later updates will provide a denser and more extensive global coverage of published data. We hope that the aYChr-DB will increase the accessibility and availability of ancient Y-DNA data.

Competing interests

EE is a consultant to DNA Diagnostic Centre.

Authors' contributions

EE initiated the study. LF carried our the analyses. LF and CB annotated the data. All the authors wrote the paper.

Acknowledgment

This work was partially supported by the MRC (MR/R025126/1) to EE.

REFERENCES

1. Brandt, G., Haak, W., Adler, C.J., Roth, C., Szecsenyi-Nagy, A., Karimnia, S., Moller-Rieker, S., Meller, H., Ganslmeier, R., Friederich, S. *et al.* (2013) Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science*, **342**, 257-261.
2. Prohaska, A., Racimo, F., Schork, A.J., Sikora, M., Stern, A.J., Ilardo, M., Allentoft, M.E., Folkersen, L., Buil, A., Moreno-Mayar, J.V. *et al.* (2019) Human Disease Variation in the Light of Population Genomics. *Cell*, **177**, 115-131.
3. Underhill, P.A. and Kivisild, T. (2007) Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.*, **41**, 539-564.
4. Kivisild, T. (2017) The study of human Y chromosome variation through ancient DNA. *Hum. Genet.*, **136**, 529-546.
5. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E. and Stewardson, K. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, **522**, 207-211.
6. Szécsényi-Nagy, A., Brandt, G., Haak, W., Keerl, V., Jakucs, J., Möller-Rieker, S., Köhler, K., Mende, B.G., Oross, K. and Marton, T. (2015) Tracing the genetic origin of Europe's first farmers reveals insights into their social organization. *Proc. R. Soc. B.*, **282**, 20150339.
7. Haak, W., Balanovsky, O., Sanchez, J.J., Koshel, S., Zaporozhchenko, V., Adler, C.J., Der Sarkissian, C.S., Brandt, G., Schwarz, C., Nicklisch, N. *et al.* (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biol.*, **8**, e1000536.
8. Zhao, Y.-B., Zhang, Y., Li, H.-J., Cui, Y.-Q., Zhu, H. and Zhou, H. (2014) Ancient DNA evidence reveals that the Y chromosome haplogroup Q1a1 admixed into the Han Chinese 3,000 years ago. *Am. J. Hum. Biol.*, **26**, 813-821.
9. Ehler, E., Novotny, J., Juras, A., Chylenski, M., Moravcik, O. and Paces, J. (2019) AmtDB: a database of ancient human mitochondrial genomes. *Nucleic Acids Res.*, **47**, D29-D32.
10. Kahle, D. and Wickham, H. (2013) ggmap: Spatial Visualization with ggplot2. *R. J.*, **5**, 144-161.
11. Brandt, G., Szécsényi-Nagy, A., Roth, C., Alt, K.W. and Haak, W. (2015) Human paleogenetics of Europe – The known knowns and the known unknowns. *J. Hum. Evol.*, **79**, 73-92.
12. Amzallag, N. (2009) From metallurgy to Bronze Age civilizations: the synthetic theory. *Am. J. Archael.*, **113**, 497-519.
13. Koryakova, L. and Epimakhov, A.V. (2014) *The Urals and western Siberia in the Bronze and Iron ages*. Cambridge University Press, Cambridge.
14. Roberts, B.W., Thornton, C.P. and Pigott, V.C. (2009) Development of metallurgy in Eurasia. *Antiquity*, **83**, 1012-1022.

15. Fuller, D.Q. (2011) Pathways to Asian civilizations: Tracing the origins and spread of rice and rice cultures. *Rice*, **4**, 78-92.
16. Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T.O., Willis, A., Matsumura, H., Buckley, H. *et al.* (2018) Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*, **361**, 92-95.
17. Mei, J., Wang, P., Chen, K., Wang, L., Wang, Y. and Liu, Y. (2015) Archaeometallurgical studies in China: some recent developments and challenging issues. *J. Archaeol. Sci.*, **56**, 221-232.

Figure legend

Figure 1

The geographical distribution of 1723 ancient Eurasian haplogroups over time. The location of each archaeological site is marked as a dot. Coloured shapes denote the different haplogroups found on the site. A small random variation was used in the plotting to avoid cluttering. Low-frequency haplogroups (<3% in interval maps, <1% in 'all time periods' map) are represented as black wedges in the pie charts and their corresponding locations marked as black crosses on the maps.