# Improving archaeological metadata reporting in human paleogenomic studies

Robert Staniuk[1,6]* (0000-0002-9941-1875), Victor Yan Kin Lee[2,6]* (0009-0001-5598-9632), Adrian Timpson[3] (0000-0003-0292-8729), Kristian Kristiansen[4] (0000-0003-2423-308X), Stephen Shennan[5] (0000-0001-6605-064X), Fernando Racimo[2] (0000-0002-5025-2607), Mark G. Thomas[3]* (0000-0002-2452-981X)

*Corresponding authors

[1] Faculty of Archaeology, Adam Mickiewicz University, ul. Uniwersytetu Poznańskiego 7/2.69, 61-614 Poznań, Poland

[2] Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark

[3] Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom

[4] Department of Historical Studies, University of Gothenburg, Renströmsgatan 6, 412 55 Göteborg, Sweden

[5] Institute of Archaeology, University College London, WC1H 0PY, London, United Kingdom

[6] These authors contributed equally

Abstract

While dramatically increasing our understanding of past demographic and adaptive processes, paleogeneomic research has also been criticized for a perceived disconnect between geneticists and other parties involved in the study of the past. For interdisciplinary research to be productive, contextual metadata associated with paleogenomic samples should be accessible in the same publication. Here, we examined the extent of archaeological metadata reporting in 30 peer-reviewed ancient DNA studies, based on genetic sequences from 3911 ancient humans, published between 2013 and 2024. Our findings show that the reporting of archaeological data is inconsistent across studies, and we have sought to identify the driving factors. Overall, we found no strong explanatory variables, though some metadata fields - like *Geolocation* - have improved in completeness over time. These inconsistencies mean that metadata reporting is often insufficient to directly investigate the relationship between patterns of cultural and genetic change on the basis of published data alone. We propose a minimum information checklist as a best-practice guideline for collecting and reporting archaeological data in the future.

Keywords

paleogenomics; archaeology; metadata; publishing standards; cultural evolution

## Introduction

The last 15 years of paleogenomic research in archaeology have revolutionized our understanding of the past. Improvements in ancient genome extraction and sequencing, as well as new bioinformatic and inferential methods, have led to a dramatic increase in ancient DNA data from around the world. This has allowed researchers to study biological relatedness[1–3], identify signatures of natural selection[4–7], gain insight into past demographic processes[8–14], track domestication processes[15], and help reconstruct language spread[16] in the past. This rapid pace of development has been facilitated by the public data archiving requirements that are *de rigueur* in the field of genetics, and provide researchers with access to previously generated ancient DNA (aDNA) sequences[17–19]. Although ancient DNA analysis as a methodological approach has been welcomed by many in the archaeological community, criticisms have been raised about the way it has been deployed, interpreted, and presented[20–26]. The issues raised include, but are not limited to, inferred relationships between genetically derived clusters and archaeological cultures[22–24,27–29], sampling biases[20–22,25], epistemological incongruence between the questions posed and processes inferred[20,22,25,27], and finally a tendency to generalize rather than provide nuanced[30] interpretations of findings[20,21,23,24,27].

The more recent focus of paleogeneomic research on site-specific investigations signals the emergence of a more collaborative framework suitable for incorporating high-resolution archaeological and genetic data[1,31–34], and suggests macroscale studies can benefit from explicitly incorporating high-resolution cultural data[35,36]. If successful, this approach would stimulate research into the mechanisms responsible for the long-term patterns of human evolutionary change documented in archaeology[37,38]. However, achieving this goal requires datasets incorporating aDNA sequences and detailed cultural, contextual and environmental information associated with the sampled individuals. Presently, only one of these datasets is widely available - aDNA - as archaeology continues to struggle with generating unified open-access datasets with standardized information on excavated human remains[39,40]. This is precisely where archaeology needs to draw from the data archiving developments of DNA research, like INSDC[41], and propose not only a general standard of reporting archaeological data but also encourage its adoption into a database, where this information becomes recorded and easily available to other researchers[40].

With regard to humans, aDNA and archaeological information are attributes of the individual. This individual is the unique identifier and the bridge between the two types of information, which provides an additional advantage of preventing data duplication or misattribution. The latter is especially relevant as the number of available genomes continues to increase rapidly, requiring more time and effort investment into aggregation and harmonization of datasets[18,42,43].

Paleogenomic studies routinely report archaeological information about sequenced individuals in the Supplementary Information (SI) using a two-tiered system. General

cultural data is reported in the descriptive supplementary, while additional information such as geolocation or radiocarbon dating is typically published in a tabulated form. Together they represent the most up-to-date information on the spatiotemporal positioning of cultural traits linked to the genetic data, and the most accessible complementary cultural dataset for the published genetic data. However, even brief examination of the contents of relevant SI files indicates that reporting standards for archaeological data vary within a single study, as well as across different papers. We examine the standards of published archaeological metadata, which variables are provided, what factors explain their presence, how they have improved through time, and whether further improvements are necessary.

To do this, we collated archaeological metadata reported in 30 ancient or historical DNA peer-reviewed articles published between 2013 and 2024 (Table 1). We recorded which of 16 important metadata variables were available, and recorded 5 plausible predictors (Table 2). We then explored variation between publications, looked for improvements in availability of metadata over time, and tested how well the explanatory variables explained metadata availability.

We find that archaeological metadata in paleogenomic studies has been reported haphazardly. We conclude that, over the last 11 years, the most visible improvement has occurred in the reporting of a sample's *Geolocation* but that variation in reporting for the other variables is generally not well explained by the factors we tested. In light of our findings, we compiled a suggested data collection checklist for archaeologists to implement as a best-practice set of guidelines to help organize the structure of reporting archaeological data in future studies.

| | **CitationID** | **Title** | **Author(s)** | **Year of publication** | **Reference** |
|---|---|---|---|---|---|
| 1 | dersa2013 | Ancient DNA Reveals Prehistoric Gene-Flow from Siberia in the Complex Human Population History of North East Europe | Der Sarkissian et al. | 2013 | [9] |
| 2 | prüfe2013 | The complete genome sequence of a Neanderthal from the Altai Mountains | Prüfer et al. | 2013 | [44] |
| 3 | gamba2014 | Genome flux and stasis in a five millennium transect of European prehistory | Gamba et al. | 2014 | [45] |
| 4 | lazar2014 | Ancient human genomes suggest three ancestral | Lazaridis et al. | 2014 | [10] |

| | | | | | |
|---|---|---|---|---|---|
| | | populations for present-day Europeans | | | |
| 5 | skogl2014 | Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers | Skoglund et al. | 2014 | 11 |
| 6 | allen2015 | Population genomics of Bronze Age Eurasia | Allentoft et al. | 2015 | 46 |
| 7 | haak2015 | Massive migration from the steppe was a source for Indo-European languages in Europe | Haak et al. | 2015 | 12 |
| 8 | mathi2015 | Genome-wide patterns of selection in 230 ancient Eurasians | Mathieson et al. | 2015 | 4 |
| 9 | fu2016 | The genetic history of Ice Age Europe | Fu et al. | 2016 | 47 |
| 10 | lazar2016 | Genomic insights into the origin of farming in the ancient Near East | Lazaridis et al. | 2017 | 48 |
| 11 | jones2017 | The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers | Jones et al. | 2017 | 49 |
| 12 | lazar2017 | Genetic origins of the Minoans and Mycenaeans | Lazaridis et al. | 2017 | 50 |
| 13 | lipso2017 | Parallel palaeogenomic transects reveal complex genetic history of early European farmers | Lipson et al. | 2017 | 51 |
| 14 | mathi2018 | The genomic history of southeastern Europe | Mathieson et al. | 2018 | 52 |
| 15 | mittn2018 | The genetic prehistory of the Baltic Sea region | Mittnik et al. | 2018 | 53 |
| 16 | olald2018 | The Beaker phenomenon and the genomic transformation of northwest Europe | Olalde et al. | 2018 | 54 |
| 17 | günth2018 | Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high- | Günther et al. | 2018 | 55 |

| | | | | | |
|---|---|---|---|---|---|
| | | latitude adaptation | | | |
| 18 | mittn2019 | Kinship-based social inequality in Bronze Age Europe | Mittnik et al. | 2019 | [1] |
| 19 | naras2019 | The formation of human populations in South and Central Asia | Narasimhan et al. | 2019 | [56] |
| 20 | olald2019 | The genomic history of the Iberian Peninsula over the past 8000 years | Olalde et al. | 2019 | [57] |
| 21 | brune2020 | Ancient genomes from present-day France unveil 7,000 years of its demographic history | Brunel et al. | 2020 | [58] |
| 22 | linde2020 | Corded Ware cultural complexity uncovered using genomic and isotopic analysis from south-eastern Poland | Linderholm et al. | 2020 | [59] |
| 23 | rivol2020 | Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers | Rivollat et al. | 2020 | [60] |
| 24 | furtw2020 | Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland | Furtwängler et al. | 2020 | [61] |
| 25 | papac2021 | Dynamic changes in genomic and social structures in third millennium BCE central Europe | Papac et al. | 2021 | [62] |
| 26 | saag2021 | Genetic ancestry changes in Stone to Bronze Age transition in the East European plain | Saag et al. | 2021 | [63] |
| 27 | liu2022 | Ancient DNA reveals five streams of migration into Micronesia and matrilocality in early Pacific seafarers | Liu et al. | 2022 | [64] |
| 28 | marót2022 | The genetic origin of Huns, Avars, and conquering Hungarians | Maróti et al. | 2022 | [65] |
| 29 | pensk2023 | Early contact between late farming and pastoralist societies in southeastern Europe | Penske et al. | 2023 | [66] |

| | | | | | |
|---|---|---|---|---|---|
| 30 | allen2024 | Population genomics of post-glacial western Eurasia | Allentoft et al. | 2024 | [67] |

Table 1: Human paleogenomic studies sampled in this study

Results

Our dataset comprises 30 publications, covering 3911 newly published or re-sequenced ancient human individuals from 49 countries across the globe but with the majority from Western Eurasia. The number of samples per publication varied from 2 to 522 with a mean of 130.37.

For each sample, we recorded the availability of 16 metadata variables typically expected in an archaeological publication (Table 2). These variables fall into either ‘core’ or ‘extended’ tiers based on their functions. The core tier comprises metadata that serve basic identification purposes; they describe the ”who”, ”when” and ”where” of a sample. The extended tier provides further information on cultural traits specific to the sample.

| **Data type** | **Tier** | **Variable** | **Description** |
|---|---|---|---|
| Archaeological metadata | Core | Geolocation | Precise site location reported as coordinates |
| | | Individual ID | Unique identifier of the archaeological individual in the source publication |
| | | Multiple names | Other names the site was reported as, including local languages and alphabets |
| | | Overall chronology | The entire documented chronological sequence of the site |
| | | References | References to previous publications of the site, findings, or analyses |
| | | Sample chronology | Dating of the sampled individual |
| | | Site description | General summary of findings from the excavated area |
| | Extended | Anthropological age | Biological age based on physical anthropology |
| | | Anthropological sex | Biological sex based on physical anthropology |

| | | | |
|---|---|---|---|
| | | Body arrangement | Arrangement of the individual according to the cardinal directions |
| | | Body positioning | How the body was placed inside the grave |
| | | C14 | Reported radiocarbon dating |
| | | Culture | Cultural affiliation of the sampled individual |
| | | Grave | Information about the immediate find circumstances of the individual |
| | | Grave goods | Finds discovered with the individual |
| | | Images | Images showing the grave, individual, and/or grave goods |
| Explanatory variables | | Archaeology Index | Average disciplinary affiliation profile of all authors |
| | | Number of authors | Total number of authors |
| | | Number of countries | Total number of countries from which the samples were reported |
| | | Number samples | Total number of newly published and re-sequenced sampled individuals |
| | | Year of publication | Year the study published in the journal |

Table 2: Archaeological metadata and candidate explanatory variables used in the study.

For each of the 3911 sequenced humans, we recorded the presence or absence of each metadata variable. In cases where presence could not be determined from the publication, ”not applicable” (NA) was assigned (see Methods). These represent only 0.52% of observations overall. We present this archaeological metadata reporting dataset in Figure 1, illustrating the presence/absence status for each sample published in the examined studies.

Figure 1: An overview of archaeological metadata reporting for all sampled individuals of the 30 peer-reviewed publications from 2013 to 2024.

To quantify the variation in reporting, we aggregated the presence/absence observation per sample, i.e. 1 (presence), 0 (absence) or 0.5 (NA), into a percentage presence (completeness) for each publication, i.e. [0,1], for each metadata variable (Fig. S1). This allowed us to treat every publication with equal statistical weight, and therefore retain an effective sample size of 30.

We found significant differences in the level of completeness across metadata variables (Kruskal-Wallis test $\chi^2(15, 480) = 199.76$; $p < 2.2e{-}16$). Core variables tended to have greater completeness than the extended variables (post-hoc Dunn tests Fig. 2A; Table S6), with *Site description* being the most complete. Our dataset also showed that core variables generally had a lower completeness variance than the extended variables, except for *Geolocation* and *Individual ID*, which had relatively high variances among all metadata (Fig. 2A). In addition, we found that the variation in completeness is more influenced by differences in the type of metadata (Kruskal-Wallis $\eta^2 = 0.398$, 95% CI [0.34 − 0.50]) than by the differences across publications ($\eta^2 = 0.080$, 95% CI [0.07 − 0.21]).

Next, we analyzed correlations between metadata variables, in their level of completeness. We first calculated the Spearman's rank correlation ($\rho$) between each pair of the 16 metadata variables. Generally, correlations were weak (non-significant accounting for multiple testing). However, we found significant positive correlations

between demographic and burial-related categories respectively (Fig. 2B).

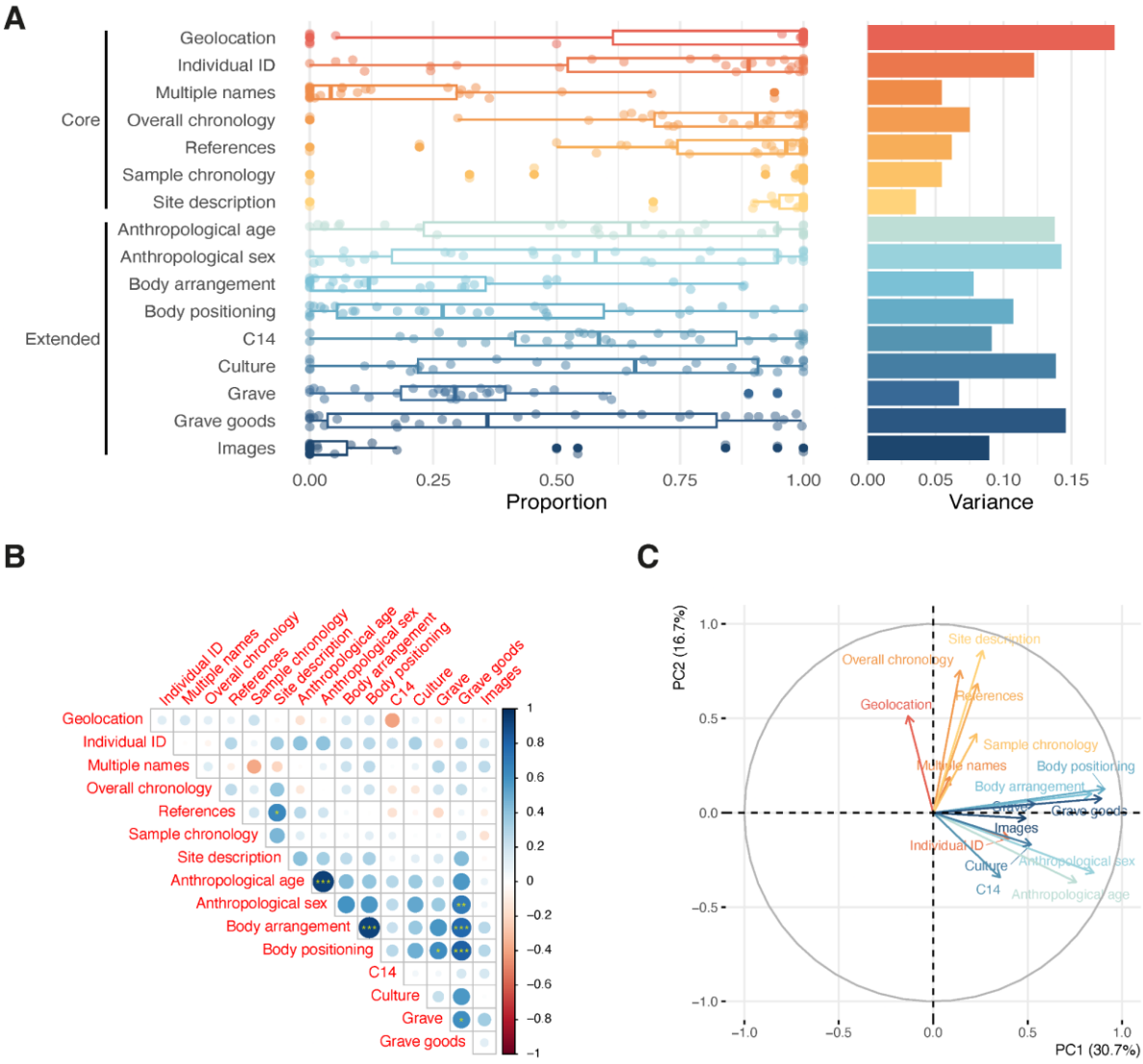


Figure 2: Variation and correlation in metadata completeness. (A) The box plot shows the distribution of the data which are represented by the points. Each point corresponds to the proportion of a metadata variable for a given publication. The bar plot illustrates sample variance of the proportions. (B) Spearman correlations. Spearman's $\rho$ is indicated by the colour and the size of the circle whereas the significance level after Bonferroni correction is indicated by the asterisks ("*": $p_{adj} < 0.05$; "**": $p_{adj} < 0.01$; "***": $p_{adj} < 0.001$). (C) The loading plot of PCA. The projected vectors correspond to the contributions of metadata variables to the PCs as measured by correlations.

To better understand the overall variance we performed a principal component analysis (PCA) to identify if there was any relationship structure in the completeness of variables (Fig. 2C). Briefly, a $m \times n$ matrix of completeness was transformed into a new coordinate system of $m \times p$, where $m$ denotes the publications, $n$ the metadata fields and $p$ the principal components. Specifically, we were interested in how different metadata contributed to the principal components. We found that PC1 and PC2 explained 30.7% and 16.7% of the total variation, respectively, and the clustering in

the loading vectors supported and more clearly represented the pairwise correlations previously found. Additionally, we found that the first two PCs broadly correspond to the extended and core metadata categories, respectively (Fig. S8G-H).

To identify factors influencing the completeness of the archaeological metadata variables, we considered five plausible explanatory variables: (1) the proportion of authors with a primarily archaeological background; (2) the total number of authors; (3) the number of sequenced samples; (4) the number of countries from which the samples originate; and (5) the year of publication (Fig. 3). As an estimate for (1), we devised an affiliation-based metric we call the “Archaeology Index” that aims to approximate the contribution of archaeologically-trained researchers in a given study (see Methods). We recognize that representing author backgrounds as a binary categorical variable (archaeology, not archaeology) does not capture the full range of their expertise, but as a first order of approximation, institutional affiliation should give a good proxy for expertise. In addition, we log-transformed the observations of candidate explanatory variables (2)-(4) as they appeared to be skewed (Fig. S4).

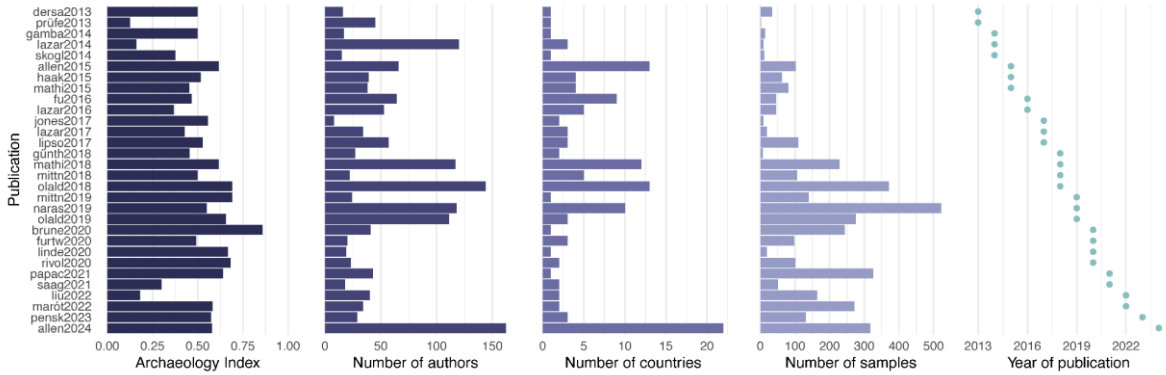


Figure 3: An overview of the candidate explanatory variables.

To determine whether these candidate explanatory variables can explain variation in the completeness of the metadata response variables, we used fractional logistic regression (Fig. 4), which is suited to working with bounded fractional response variables (see Methods). Generally, the explanatory variables provided very little explanatory power (Fig. 5). Across all models, we only identified a statistically significant positive effect for the year of publication on *Geolocation* after adjusting for multiple testing ($\hat{\beta} = 1.56$, $p_{adj} = 0.018$), suggesting improved reporting of geographical coordinates over time. We also identified negative associations between *Images* and the *log(number of samples)* ($\hat{\beta} = -0.92$, $p_{adj} = 0.017$) and between *Site description* and the *log(number of authors)* ($\hat{\beta} = -0.65$, $p_{adj} = 0.024$), suggesting these two metadata variables have lower completeness in the larger publications that sequenced more ancient humans and had a longer author list.

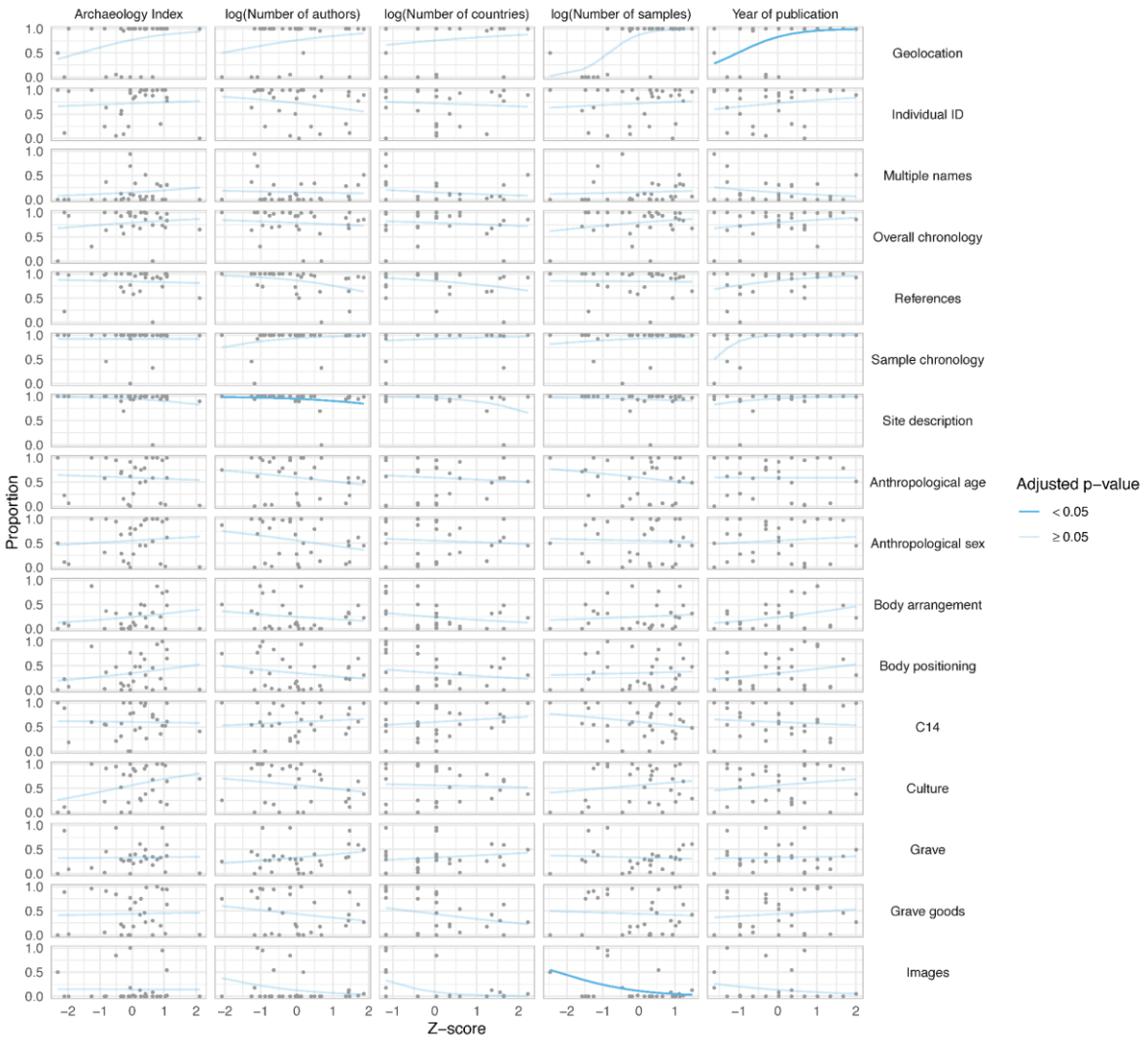


Figure 4: Model fit of fractional logistic regression model. An opaque line indicates the coefficient estimate ($\hat{\beta}$) is significant after Bonferroni correction.

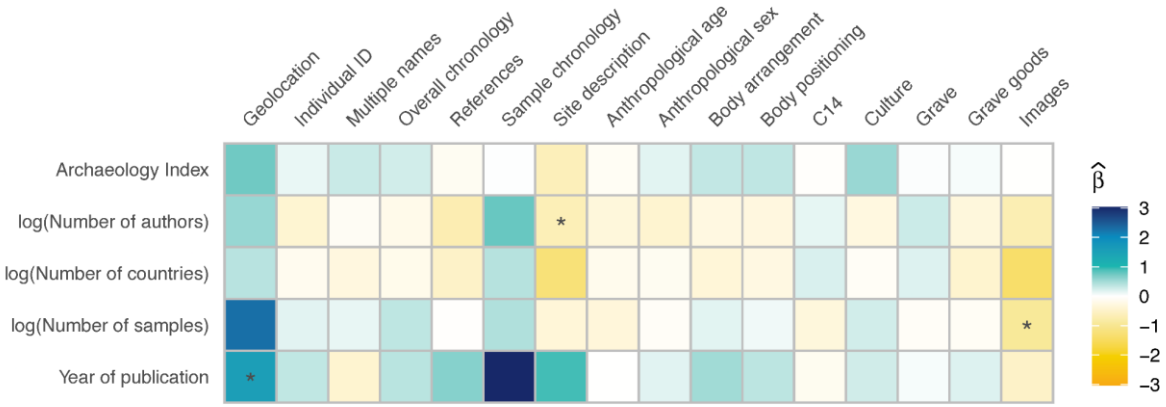


Figure 5: The effects of candidate explanatory variables on metadata completeness. The coefficient estimate ($\beta$) is indicated by the color, whereas the significance level is indicated by the asterisks ("*": $p_{adj} < 0.05$; "**": $p_{adj} < 0.01$; "***": $p_{adj} < 0.001$).

In an attempt to identify latent explanatory variables — not well represented by our individual variables, but perhaps captured by a combination of them — we carried out a PCA of the five explanatory variables, and then performed a multiple regression analysis using the first three PCs (explaining 90% of variation in all five explanatory variables (see Methods). Across all metadata, only *Geolocation* ($\chi^2(3, 26) = 8.51$, $p = 0.037$), *Sample chronology* ($\chi^2(3, 26) = 14.12$, $p = 0.003$), and *Site description* ($\chi^2(3, 26) = 13.21$, $p = 0.004$) were statistically significant based on the Wald test. In these models, PC1 appears to have significant positive associations with *Geolocation* ($\hat{\beta} = 1.19$, $p_{adj} = 0.019$) and *Sample chronology* ($\hat{\beta} = 1.51$, $p_{adj} = 0.002$); PC2 has a positive association with *Sample chronology* ($\hat{\beta} = 1.37$, $p_{adj} = 0.01$), while PC3 has a negative association with *Site description* ($\hat{\beta} = -4.23$, $p_{adj} = 0.001$) (Fig. 6). We note, however, that here PCs are linear combinations of the explanatory variables, and that the regression is simply a test for (linear) association (Fig. S11). Overall, *Geolocation* and *Site description* were the only two metadata variables found to be impacted by at least one of the explanatory variables in both simple and multiple regression analyses.

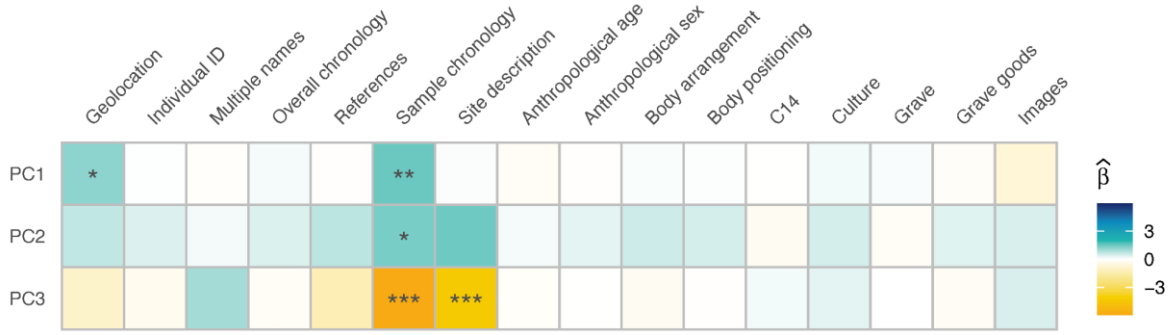


Figure 6: The effects of candidate explanatory variables on metadata completeness. The coefficient estimate ($\hat{\beta}$) is indicated by the color, whereas the significance level is indicated by the asterisks ("*": $p_{adj} < 0.05$; "**": $p_{adj} < 0.01$; "***": $p_{adj} < 0.001$). Only the $\hat{\beta}$ significance levels from the models significant against the null model based on the Wald test are shown.

## Discussion

Our study found considerable variation in the completeness of metadata reporting, across variables and studies. We identified improvement in the completeness of *Geolocation* over time, but our candidate explanatory variables poorly explained differences in the other 14 variables. It is likely that the lack of a universally accepted standard for reporting cultural data in archaeology is a significant contributing factor for this[40]. In the absence of clear guidelines, archaeologists make individual decisions shaped by country of origin, period, school, and experience, which collectively generate an impression of inconsistent data reporting.

The observed improvement through time in *Geolocation* reporting may be due to the

increased use of explicit spatial modeling with genetic data[68–70], as well as the increasing importance of spatial positioning of new sequences given the ever-growing number of available samples across the world. The latter is likely associated with the creation of aDNA data repositories such as AADR[18] and Poseidon[19], where geolocation became a core metadata requirement[71]. The creation of large datasets has sparked the emergence of aDNA research communities such as SPAAM (https://www.spaam-community.org/), HAAM (https://haam-community.github.io/), or MIxS-MInAS (https://www.mixs-minas.org/) which spearhead the examination of metadata reporting standards and the development of community standards that ensure consistency and longevity of available data[71,72]. Together, these factors have likely resulted in the widely established aDNA metadata fields, which include geolocation, sample age, and sample type/material (e.g. petrous bone and tooth)[43,71].

We identified a decrease in the completeness of *Images* reported as the number of samples increased, which likely represents the difficulties in acquiring permits for reproducing existing images of burials and finds in papers with a large number of new sequences. As paleogenomic studies generally use samples from already reported discoveries, this problem is usually resolved by providing references to original publications. We found a relatively high completeness level for *References*, suggesting that this information may be obtainable from the source publications. However, the process of acquiring and examining each publication can be challenging, time-consuming, and the source publications may not contain relevant images. Since images are as close as possible to raw archaeological data, indicating whether they can be obtained in published resources can streamline data collection for other researchers.

We identified a decrease in *Site description* reporting as the number of authors increased. This information allows correct attribution of samples to their origins and enables differentiating between locations with similar names that are found in immediate proximity. As the scope of papers increases and requires larger collaborative work, it becomes more challenging to assemble all the necessary metadata in the supplementary information. Currently these supplementaries are not a mandatory part of the peer-review process, so the omission of these relevant data are likely to increase in the future.

Apart from these widely established aDNA metadata fields, *Sample chronology*, *Overall chronology*, and *References*, show consistently high completeness across publications. This trend only partially overlaps with the metadata fields of AADR[18], Poseidon[19], or recommendations from the different metadata consortia. The relatively high variance of *Individual ID* reporting could result from assigning sample IDs to aDNA data, deeming *Individual ID* from the original archaeological literature optional. *Multiple names*, which records inclusion of original site names from countries using non-Latin alphabets, was consistently missing from the core metadata. This could pose severe challenges for tracing originally published contexts.

Unlike core metadata, the extended tier is not well scrutinized at the peer-review stage, and efforts to collect large archaeological datasets remain specific to particular research projects[40]. At the time of writing, large databases that incorporate archaeological data from different studies or metadata consortia addressing the data reporting standards are in their infancy. This in turn makes it challenging to openly propose which data carries the most importance and which form of reporting is most effective. However, the increasing rate at which archaeological data is collected and reported in paleogenomic papers raises the question of whether it is time these discussions emerge to counteract the reporting variability we document here. Its importance cannot be overstated as the increasing number of published paleogenomes is already causing problems with correct attribution of samples to particular individuals, distinguishing between sequenced genomes originating from the same site, and distinguishing duplicates from twins. This is precisely the avenue where archaeological data will prove most important as the combination of variables should reduce misattribution or duplication of results. The currently utilized standard that relies on geolocation, sample date, and sample type is insufficient as often more than a single sequenced genome originates from the same archaeological site, same grave, or even the same individual.

Beyond challenges with data attribution, higher variability in metadata reporting has broader impacts on the interdisciplinary nature of human paleogenetic and archaeological research. Without explicit reporting standards, research on links between patterns of genetic and cultural variation will require additional efforts for data collection. The time-consuming and error-prone process of individual tracing of every relevant publication to obtain information that is not necessarily the most up-to-date can be streamlined if clear guidelines are provided to data curators responsible for providing samples for paleogenomic research. As researchers with access to the most detailed descriptions of archaeological finds, either through working in museums or excavating the individuals themselves, they are the most competent to report the fine-grained details of the samples they have acquired. If provided with a clear outline of what should be considered the minimum required information, they can create standardized descriptions to be re-evaluated in the future. Last but not least, it would address the ever-persistent challenge of heterogeneous cultural data reporting standards allowing archaeologists to more effectively utilize the data reported in aDNA papers in their work[40,73]. This is crucial for a more detailed exploration of the relationship between genetic and cultural patterns. Booth et al. recently demonstrated the relevance of high-quality reporting standards by critically examining supplementary information to provide a broader context for the genetic turnover patterns in the British Isles during the Bell Beaker period[74].

Changing the existing metadata reporting standards has to be adopted by both paleogenomics as well as archaeology. As paleogeneticists already create consortia directed towards developing better metadata standards, initiatives similar to MIxS-MInAS directed towards the structuring of archaeological metadata are needed[40]. Our

results indicate that although archaeologists were always part of the research groups and discussions throughout the paleogenomics revolution, insufficient attention was paid to the archaeological data ending up in the final publications and supplementary materials.

## Checklist for reporting archaeological data

Drawing on the work of communities involved in evaluating and improving standards in publishing aDNA data (SPAAM, HAAM, MIxS-MInAS), from the results of our analysis, and from conversations with researchers, we suggest a minimum information standard for reporting archaeological data in paleogenetic research. We acknowledge that the archaeological record is complex, shaped by the history of discovery as well as skills, knowledge, and access to technology archaeologists had at the time of discovery. However, we believe that some key data should always be reported. The key requirement is that the archaeological individual—the specific, unique, discovery associated with a particular human being—who was discovered, documented, and used in scientific research, is adopted as the primary data reporting unit. This standard of data reporting allows all the relevant archaeological and genetic data to be linked to a single, unique individual, providing a more robust way of associating different strands of relevant data. This standard was adopted by the COREX project and its Big Interdisciplinary Archaeological Database (BIAD)[40] to enable a more streamlined data collection protocol and limit the possibility of data duplication or misattribution. We believe other researchers will benefit from it and can improve upon it by an informed consideration of how to best report specific cultural data. With the right kind of momentum and increasing data availability, we hope that initiatives like SPAAM, HAAM, or MIxS-MInAS, will emerge around this data standard and provide discussions to further improve data reporting. The responsibility for collecting cultural data falls on the archaeological community, as archaeologists discover, document, extract, report, data-mine, aggregate and store past material remains. However, ensuring that this information is provided in a comprehensive form will be of benefit to archaeologists and geneticists alike.

Presently, following our experience with associating paleogenomic data with archaeological individuals, we propose the following checklist for reporting archaeological data from aDNA-sampled individuals in peer-reviewed journals, and crucially we recommend this is checked at peer review:

### Site information

Specify the site name, including the historical site names, alternative reporting conventions, names according to the local script, and heritage board identification numbers. If the name has been transliterated make sure to include the original name. Be mindful of diacritic signs. This will prevent site duplication based on small transcription differences, whilst enabling the search for other types of data for other research purposes.

Always include latitude and longitude coordinates and report if there are any uncertainties regarding site location in cases of legacy data, preferably in degrees decimal. If the geolocation cannot be disclosed for heritage protection reasons, state that information explicitly. If the location is only known to a broad region, a shape file (such as a .kml) of the region is useful.

Specify dating information, including how it was generated. If the dating was established using radiocarbon dating, include lab codes, material type, and raw uncalibrated data. If archaeological dating was provided, include the method and criteria used for its estimation.

Include general information on site chronology, specifying its periodization and associated finds. If possible, include quantification of the documented findings. Report the presence of identified cultural groups. If archaeological cultures are not reported in the area, make that information explicit. Always include references and always report if the samples originate from an on-going archaeological project.

Individual information

Specify the individual name as originally reported, and include alternative reporting conventions. Always include the lab ID used for genetic sequencing.

Specify the dating of the individual. If the dating was established using archaeometry, include lab codes, material type, and raw (uncalibrated) data. If archaeological dating was used, report the method and criteria used.

Report the find circumstances: if the sample comes from a grave, report details of the grave as well such as type, shape, and dimensions.

Report the cultural group and the classification method. If the individual originates from a region or is assigned a time period without cultural groups, be explicit about this.

Report the burial rite (inhumation, cremation, partial cremation, etc.).

Report the body positioning using at least the conventional labels (crouched, extended, supine, sitting). If the position was indeterminate, state this explicitly. If possible include the image of the individual in situ.

Report the body orientation using angles or cardinal directions of the head of the deceased.

Report the anatomical completeness of the individual. Specify if such an evaluation was not performed or other factors prevented such an analysis.

Report the osteological sexing of the sampled individual if any determination was made. If the analyses yielded an uncertain classification or proved indeterminate, include this information. Uncertainty can be reported as confidence probability, where 0 represents 100% confidence of male determination, while 1 represents 100%

confidence of female determination[40].

Report the presence of grave goods and provide a list of finds and, if possible, images. If finds have been classified following an existing typological scheme include the reference. If the finds have been published in detail elsewhere provide a reference. If no grave goods were found, report it.

Report health markers if such data was collected. Include an explicit statement about absence of documented health markers. Report if such information is being prepared for a separate publication.

We encourage authors to include images, or hyperlinks to images, with find circumstances, body positioning, body orientation, and grave goods, as they provide the possibility of verifying and re-classifying data. We acknowledge the objective difficulties in acquiring permissions for re-publishing images from existing sources, especially for previously published individuals, and recommend that in these instances a source reference is provided in the Supplementary Information. We recognize that in some countries the presentation or publication of images of the dead is banned or prevented by publishing outlets. We encourage authors to explicitly state these conditions in the SI, so that the researchers attempting the incorporate such data in their work can account for the cultural biases and legal regulations.

In all instances absence of information about any of the aforementioned variables should be explicitly stated to prevent subjective data input. While such a standard has been already encountered in one of the analysed papers[62], it should become the default for reporting archaeological data.

We are confident that applying this minimum information checklist as a template for reporting archaeological data will streamline publication of ancient DNA research using prehistoric and historic individuals. Firstly, the checklist provides an explicit list of requirements that paleogeneticists should expect from their collaborators. Secondly, irrespective of their temporal or regional interests, archaeologists can now assess which information should be included in such a report. Thirdly, reviewers will know which information they should seek in the SI files and what to mark as incomplete or missing entries.

## Conclusions

The paleogenomic revolution in archaeology is here to stay; whether it will spearhead truly interdisciplinary research on the human past will depend on our ability to effectively integrate rich archaeological data with genetic data. Collaboration and communication are the necessary first steps. Adopting strict practices of academic rigor in data reporting and curation can be challenging, but will enable more complex research questions to be addressed, further blurring the line between humanities and sciences[75].

As cultural information has been reported in an unstructured form since 2013, our

recommendations do not require creating a new workflow as much as modifying the already existing one and allowing geneticists, archaeologists, and reviewers, to address the quality of information they use in their research. This is why the key structural component is to ensure that the sample metadata reporting undergoes evaluation as part of the default manuscript review process. We are confident that, in our examination of the first 15 years of paleogenomics, we have identified a critical issue in an otherwise cutting-edge research field. In order to make the paleogenomics revolution in archaeology truly revolutionary, this issue needs to be tackled ahead of time for the benefit of all parties involved.

## Methods

### Sampling of papers for metadata collection

We sampled 30 peer-reviewed articles (cf. Table 1) published between 2013 and 2024, as this period represents the most intensive period of aDNA analysis in archaeology. No regional or chronological preference was given, although the prevalence of studies on Eurasian prehistory is observable, as it is the area and period most frequently chosen in evolutionary genetics[18,19]. However, given the high density of sampling of this space and time it should be considered representative for determining the quality of the reported archaeological data and estimating the overall reporting standards. Our goal was to select at least 2 papers per year published since 2013 and initially included pre-prints. The data collection was finalized in early 2023 with a single paper[66]. While preparing the paper one of the pre-prints was published[67], so the paper was re-examined to reflect the peer-reviewed version and bibliographic reference was modified to reflect the final version of the study.

### Protocol for archaeological metadata collection - variable selection

The protocol derives from an on-going collection of archaeological data for the COREX project, specifically the cultural data associated with individuals sampled for aDNA research. The variables selected for the purpose of the project correspond to the generally accepted standards of publishing archaeological individuals in the scientific literature[40,76–79]. It should be stressed that some degree of idiosyncrasy caused by the evolution of the discipline, divergent archaeological schools, discovery circumstances and preservation of finds, last but not least, culture-specific depositional behavior is common. However, it is because of the developed data collection protocol that problems with reporting archaeological data were encountered and after extensive documentation over a period of 2 years, the initial protocol was re-worked to prepare a template for the evaluation of archaeological metadata reporting standards in paleogenomics. The data collection protocol followed the original hierarchical structure, where the highest order of information is represented by information related to the archaeological site, followed by data directly linked to the sampled individual.

### Data collection

The descriptive Supplementary Information (SI) was the primary source of data since it is the designated section for reporting details about the sampled individuals. The tabulated SI files were examined to extract information about the coordinates, radiocarbon dating, or sexing of the individuals. However, cross-referencing the two sources has often revealed the discrepancies between the genetic identification numbers reported between supplementaries of the same paper. Most commonly encountered problems were aggregations of IDs reported in two or more columns, re-ordering of the ID numbers, or replacement of symbols, such as hyphens or backslashes. For convenience sake, it was decided to rely on the ID numbers reported in the descriptive SI as they correspond directly to the archaeological descriptions. The varying standards and qualities of reporting archaeological data in the descriptive SI meant that the data collection had to be handled manually.

Data preparation

The reporting of archaeological metadata was classified as ”presence”, ”absence” or ”not applicable” which were encoded numerically as 1, 0, and 0.5. The latter applied to metadata of previously published individuals re-sequenced in the publications examined or those of new individuals from sites that were reported in previously published studies. As only the publications listed above were examined, we considered 0.5 fair since we had no information on the presence or absence of the metadata. We also provide a more nuanced classification of the data in the Supplementary, which for the final analysis was reduced to presence or absence, irrespective of the overall quality. This data was transformed into proportions per publication for downstream quantitative analyzes.

Archaeology Index

The Archaeology Index was used to measure the contribution of archaeology-affiliated authors. To minimize the subjectivity of assigning author’s affiliation due to limited acces to academic histories of every author we decided to score each author by their affiliation(s) as reported in the publications, as this would be relatively consistent and thus fairer to all authors in terms of information availability. An archaeology affiliation would have score 1 and a non-archaeology one would have 0. In cases where the affiliation could not be easily established, the website of the institution was examined to look for data on the main research areas, details of the academic position held by the author, or any additional information which could help assign the disciplinary affiliation. In cases where no information could be obtained we erred on the side of caution and assigned a score of 0.5. With these scores, we first calculated an author’s average and used that to obtain the publication’s average which we termed the Archaeology Index.

Testing the differences in the level of completeness among metadata and its effect size

The Kruskal-Wallis test, a non-parametric test, was used to assess whether there were significant differences in the level of completeness among metadata variables. The

post-hoc Dunn test with Bonferroni correction was then used for pairwise comparisons with the R package dunn.test v1.3.6[80]. We computed the effect size of the Kruskal-Wallis tests on the differences between metadata and between publications, respectively, for comparison using the R package rstatix 0.7.2[81].

Correlation analysis and PCA of archaeological metadata

The Spearman correlation was used to measure the strength and direction of the relationships between the metadata variables. The p-values of Spearman's $\rho$ were adjusted for multiple testing using Bonferroni correction. To further explore and visualize the variation in reporting, PCA was performed using the R package FactoMineR v2.11[82] on the same data (Fig. S6-S8).

Modeling the relationships between metadata completeness and candidate explanatory variables

To model the relationships between metadata completeness and candidate explanatory variables, we applied fractional logistic regression (as called in Stata)[83] which uses a quasi-likelihood approach to handle fractional response variables. This was done via fitting a logistic regression model to estimate the coefficients and estimating robust standard errors using the R package sandwich v3.1.1[84] with the "HC1" estimator[85]. For simple regressions where candidate explanatory variables were modeled separately, the p-values of coefficient estimates were adjusted for multiple testing for each metadata variable using Bonferroni correction. Multiple regressions were performed to address potential confounding between candidate explanatory variables. In these models, the first three PCs instead of the original variables were used due to multicollinearity detected. Finally, the Wald test was used to assess model fit which tests whether at least one coefficient estimate is significantly different from 0.

We also performed extended-support beta (XBX) regression[86,87], which is fully parametric in comparison to the quasi-likelihood based fractional logistic regression. Log-likelihood (log-pseudolikelihood in the case of fractional logistic regression), AIC, BIC and pseudo $R^2$ (squared correlation between the predicted and observed values) were calculated to assess goodness of fit (Fig. S20 & S21). As their pseudo $R^2$ were almost identical, we report in the main text the results from the fractional logistic regression model, which is a simpler model that generally resulted in a lower AIC or BIC.

Assessing multicollinearity in multiple regression analysis

We found significant positive correlations between several pairs of candidate explanatory variables (Fig. S5). This suggests potential multicollinearity in multiple regression which would bias coefficient and standard error estimation. To assess multicollinearity, we computed the variance inflation factor (VIF) for each explanatory variable in each model (Fig. S13-S16). A VIF that exceeds 5 or 10 was considered problematic[88].

All analyses were carried out in R v4.4.0[89]. The figures were generated using packages ggh4x v0.2.8[90], MetBrewer v0.2.0[91], patchwork v1.3.0[92], RColorBrewer v1.1-3[93] and tidyverse v2.0.0[94].

## Data Availability

All data used in this study can be found in the Supplementary tables.

## Code Availability

All code used in this study is openly available on GitHub at https://github.com/vicyklee/arch_adna_metadata, under the MIT License.

## References


1. Mittnik, A. *et al.* Kinship-based social inequality in Bronze Age Europe. *Science* **366**, 731–734 (2019).
2. Popli, D., Peyrégne, S. & Peter, B. M. KIN: a method to infer relatedness from low-coverage ancient DNA. *Genome Biology* **24**, 10 (2023).
3. Cassidy, L. M. *et al.* Continental influx and pervasive matrilocality in Iron Age Britain. *Nature* **637**, 1136–1142 (2025).
4. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
5. Evershed, R. P. *et al.* Dairying, diseases and the evolution of lactase persistence in Europe. *Nature* **608**, 336–345 (2022).
6. Barrie, W. *et al.* Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* **625**, 321–328 (2024).
7. Irving-Pease, E. K. *et al.* The selection landscape and genetic legacy of ancient Eurasians. *Nature* **625**, 312–320 (2024).
8. Bramanti, B. *et al.* Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. *Science* **326**, 137–140 (2009).
9. Sarkissian, C. D. *et al.* Ancient DNA Reveals Prehistoric Gene-Flow from Siberia in the Complex Human Population History of North East Europe. *PLOS Genetics* **9**, e1003296 (2013).
10. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
11. Skoglund, P. *et al.* Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. *Science* **344**, 747–750 (2014).
12. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
13. Brace, S. *et al.* Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat Ecol Evol* **3**, 765–771 (2019).
14. Allentoft, M. E. *et al.* 100 ancient genomes show repeated population turnovers in Neolithic Denmark. *Nature* **625**, 329–337 (2024).
15. Librado, P. *et al.* Widespread horse-based mobility arose around 2200 bce in Eurasia. *Nature* 1–7 (2024) doi:10.1038/s41586-024-07597-5.
16. Lazaridis, I. *et al.* The genetic history of the Southern Arc: A bridge between West Asia and Europe. *Science* **377**, eabm4247 (2022).
17. Ehler, E. *et al.* AmtDB: a database of ancient human mitochondrial genomes. *Nucleic*

*Acids Research* **47**, D29–D32 (2019).
18. Mallick, S. *et al.* The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci Data* **11**, 182 (2024).
19. Schmid, C. *et al.* Poseidon – A framework for archaeogenetic human genotype data management. *eLife* **13**, (2024).
20. Hofmann, D. What Have Genetics Ever Done for Us? The Implications of aDNA Data for Interpreting Identity in Early Neolithic Central Europe. *European Journal of Archaeology* **18**, 454–476 (2015).
21. Vander Linden, M. Population history in third-millennium-BC Europe: assessing the contribution of genetics. *World Archaeology* **48**, 714–728 (2016).
22. Ion, A. How Interdisciplinary is Interdisciplinarity? Revisiting the Impact of aDNA Research for the Archaeology of Human Remains. *Current Swedish Archaeology* **25**, 177–198 (2017).
23. Johannsen, N. N., Larson, G., Meltzer, D. J. & Vander Linden, M. A composite window into human history. *Science* **356**, 1118–1120 (2017).
24. Frieman, C. J. & Hofmann, D. Present pasts in the archaeology of genetics, identity, and migration in Europe: a critical essay. *World Archaeology* **51**, 528–545 (2019).
25. Sykes, N., Spriggs, M. & Evin, A. Beyond curse or blessing: the opportunities and challenges of aDNA analysis. *World Archaeology* **51**, 503–516 (2019).
26. Booth, T. J. A stranger in a strange land: a perspective on archaeological responses to the palaeogenetic revolution from an archaeologist working amongst palaeogeneticists. *World Archaeology* **51**, 586–601 (2019).
27. Furholt, M. De-contaminating the aDNA – archaeology dialogue on mobility and migration: discussing the culture-historical legacy. *Current Swedish Archaeology* **27**, 53–68 (2019).
28. Furholt, M. Massive Migrations? The Impact of Recent aDNA Studies on our View of Third Millennium Europe. *European Journal of Archaeology* **21**, 159–191 (2018).
29. Heyd, V. Kossinna's smile. *Antiquity* **91**, 348–359 (2017).
30. Healy, K. Fuck Nuance. *Sociological Theory* **35**, 118–127 (2017).
31. Cassidy, L. M. *et al.* A dynastic elite in monumental Neolithic society. *Nature* **582**, 384–388 (2020).
32. Schroeder, H. *et al.* Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proceedings of the National Academy of Sciences* **116**, 10705–10710 (2019).
33. Sjögren, K.-G. *et al.* Kinship and social organization in Copper Age Europe. A cross-disciplinary analysis of archaeology, DNA, isotopes, and anthropology from two Bell Beaker cemeteries. *PLOS ONE* **15**, e0241278 (2020).
34. Brami, M. *et al.* Was the Fishing Village of Lepenski Vir Built by Europe's First Farmers? *J World Prehist* **35**, 109–133 (2022).
35. Kristiansen, K. *Archaeology and the Genetic Revolution in European Prehistory*. (Cambridge University Press, Cambridge, 2022). doi:10.1017/9781009228701.
36. Shennan, S. Population, culture history, and the dynamics of change in European prehistory. *Journal of the Royal Anthropological Institute* **30**, 1085–1101 (2024).
37. Perreault, C. *The Quality of the Archaeological Record*. (University of Chicago Press, Chicago, IL, 2019).
38. Orlando, L. *et al.* Ancient DNA analysis. *Nat Rev Methods Primers* **1**, 1–26 (2021).
39. Anagnostou, P. *et al.* When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement. *PLOS ONE* **10**, e0121409 (2015).

40. Reiter, S. S. *et al.* The BIAD Standards: Recommendations for Archaeological Data Publication and Insights From the Big Interdisciplinary Archaeological Database. *Open Archaeology* **10**, (2024).
41. Karsch-Mizrachi, I. *et al.* The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Research* **53**, D62–D66 (2025).
42. Charlton, S., Booth, T. & Barnes, I. The problem with petrous? A consideration of the potential biases in the utilization of pars petrosa for ancient DNA analysis. *World Archaeology* **51**, 574–585 (2019).
43. Bergström, A. Improving data archiving practices in ancient genomics. *Sci Data* **11**, 754 (2024).
44. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
45. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun* **5**, 5257 (2014).
46. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
47. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
48. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
49. Jones, E. R. *et al.* The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Current Biology* **27**, 576–582 (2017).
50. Lazaridis, I. *et al.* Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017).
51. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**, 368–372 (2017).
52. Mathieson, I. *et al.* The genomic history of southeastern Europe. *Nature* **555**, 197–203 (2018).
53. Mittnik, A. *et al.* The genetic prehistory of the Baltic Sea region. *Nature Communications* **9**, 1–11 (2018).
54. Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018).
55. Günther, T. *et al.* Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biology* **16**, e2003703 (2018).
56. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019).
57. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
58. Brunel, S. *et al.* Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proceedings of the National Academy of Sciences* **117**, 12791–12798 (2020).
59. Linderholm, A. *et al.* Corded Ware cultural complexity uncovered using genomic and isotopic analysis from south-eastern Poland. *Sci Rep* **10**, 6885 (2020).
60. Rivollat, M. *et al.* Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers. *Science Advances* **6**, eaaz5344 (2020).
61. Furtwängler, A. *et al.* Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland. *Nat Commun* **11**, 1915 (2020).

62. Papac, L. *et al.* Dynamic changes in genomic and social structures in third millennium BCE central Europe. *Science Advances* **7**, eabi6941 (2021).
63. Saag, L. *et al.* Genetic ancestry changes in Stone to Bronze Age transition in the East European plain. *Science Advances* **7**, eabd6535 (2021).
64. Liu, Y.-C. *et al.* Ancient DNA reveals five streams of migration into Micronesia and matrilocality in early Pacific seafarers. *Science* **377**, 72–79 (2022).
65. Maróti, Z. *et al.* The genetic origin of Huns, Avars, and conquering Hungarians. *Current Biology* **32**, 2858-2870.e7 (2022).
66. Penske, S. *et al.* Early contact between late farming and pastoralist societies in southeastern Europe. *Nature* **620**, 358–365 (2023).
67. Allentoft, M. E. *et al.* Population genomics of post-glacial western Eurasia. *Nature* **625**, 301–311 (2024).
68. Loog, L. *et al.* Estimating mobility using sparse data: Application to human genetic variation. *Proceedings of the National Academy of Sciences* **114**, 12213–12218 (2017).
69. Racimo, F. *et al.* The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences* **117**, 8989–9000 (2020).
70. Schmid, C. & Schiffels, S. Estimating human mobility in Holocene Western Eurasia with large-scale ancient genomic data. *Proceedings of the National Academy of Sciences* **120**, e2218375120 (2023).
71. Bergström, A., Fellows Yates, J. A. & Warinner, C. Ancient DNA data hold insights into past organisms and ecosystems — handle them with more care. *Nature* **636**, 296–298 (2024).
72. Fellows Yates, J. A. *et al.* Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci Data* **8**, 31 (2021).
73. Alpaslan-Roodenberg, S. *et al.* Ethics of DNA research on human remains: five globally applicable guidelines. *Nature* **599**, 41–46 (2021).
74. Booth, T. J., Brück, J., Brace, S. & Barnes, I. Tales from the Supplementary Information: Ancestry Change in Chalcolithic–Early Bronze Age Britain Was Gradual with Varied Kinship Organization. *Cambridge Archaeological Journal* **31**, 379–400 (2021).
75. Racimo, F., Sikora, M., Vander Linden, M., Schroeder, H. & Lalueza-Fox, C. Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nat Rev Genet* **21**, 355–366 (2020).
76. Parker Pearson, M. *The Archaeology of Death and Burial*. (Sutton Publishing, Stroud, 1999).
77. Knüsel, C. J. Introduction. Archaeothanatology, funerary archaeology and bioarchaeology: perspectives on the long view of death and the dead. in *The Routledge Handbook of Archaeothanatology, Bioarchaeology of Mortuary Behaviour* (eds. Knüsel, C. J. & Schotsmans, E. M. J.) 1–20 (Routledge, London, 2022).
78. Tõrv, M. *Persistent Practices. A Multi-Disciplinary Study of Hunter-Gatherer Mortuary Remains from c. 6500-2600 Cal. BC, Estonia*. (University of Tartu Press, Tartu, 2016).
79. Nilsson Stutz, L. *Embodied Rituals & Ritualized Bodies. Tracing Ritual Practices in Late Mesolithic Burials*. vol. 46 (Wallin & Dalholm Bokrytckeri AB, Lund, 2003).
80. Dinno, A. dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. (2024).
81. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2023).
82. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25**, 1–18 (2008).
83. Papke, L. E. & Wooldridge, J. M. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* **11**,

619–632 (1996).
84. Zeileis, A., Köll, S. & Graham, N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software* **95**, 1–36 (2020).
85. Zeileis, A. Object-oriented Computation of Sandwich Estimators. *Journal of Statistical Software* **16**, 1–16 (2006).
86. Kosmidis, I. & Zeileis, A. Extended-support beta regression for $[0, 1]$ responses. Preprint at https://doi.org/10.48550/arXiv.2409.07233 (2025).
87. Cribari-Neto, F. & Zeileis, A. Beta Regression in R. *Journal of Statistical Software* **34**, 1–24 (2010).
88. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning*. (Springer, New York, NY, 2021).
89. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, 2018).
90. Brand, T. van den. ggh4x: Hacks for 'ggplot2'. (2024).
91. Mills, B. R. MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art. (2022).
92. Pedersen, T. L. patchwork: The Composer of Plots. (2024).
93. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2022).
94. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

## Author Contributions

All authors contributed to the study conception and design. Data collection was conducted by R.S. with input from V.Y.K.L., S.S., M.G.T., A.T. Data analysis was conducted by V.Y.K.L with input from F.R., R.S., M.G.T., A.T. All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript.

## Competing Interests

The authors declare no competing interests.

## Acknowledgements

This research is the result of the synergy project 'COREX: From Correlations to Explanations: towards a new European prehistory', funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 951385). R.S. was additionally supported by the NAWA Polish Returns 2023 project (BPN/PPO/2023/1/00013/U/00001) and the Polish National Science Center Research Component (2024/03/1/HS3/00008). F.R. was additionally supported by a Novo Nordisk Fonden Data Science Ascending Investigator Award (NNF22OC0076816) and by the European Research Council (ERC) under the European Union's Horizon Europe programme (Grant Agreement No. 101077592).