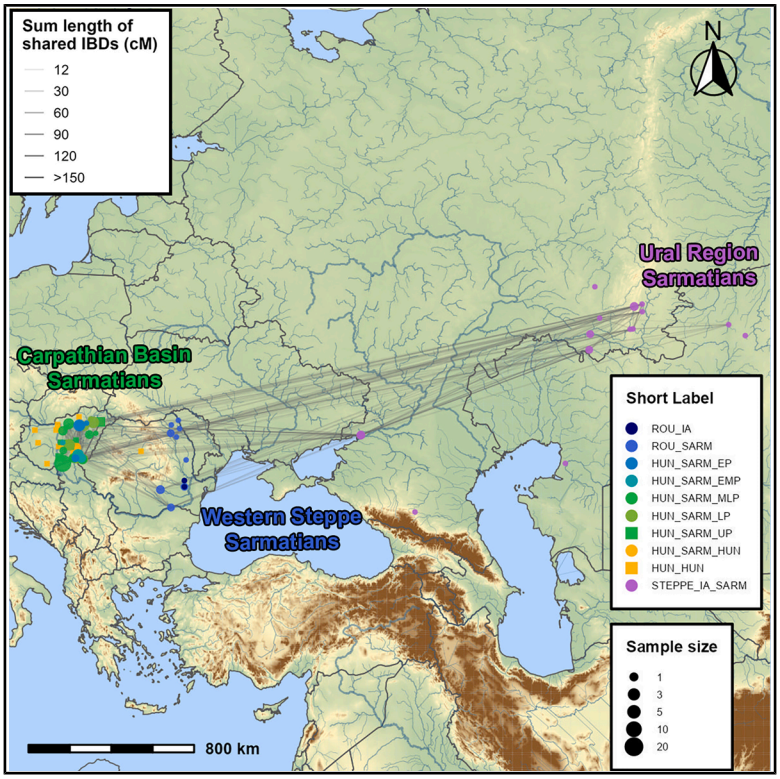
Cell

Article

Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin

Graphical abstract



Highlights

- 156 new ancient genomes unravel the origin of Sarmatians in central Europe
- Direct IBD connections reveal the Uralic roots of the different Sarmatian groups
- Y chromosome turnover suggests male-driven migration
- Continuity of the Sarmatian population into the Hun period

Authors

Oszkár Schütz, Zoltán Maróti, Balázs Tihanyi, ..., Sándor Varga, Endre Neparáczki, Tibor Török

Correspondence

torokt@bio.u-szeged.hu

In brief

Analysis of ancient genomes reveals that the Sarmatians in the Carpathian Basin had mainly European genomic composition, with minor Asian ancestry setting them apart from the locals. Apart from this, they shared direct genealogical links with other Sarmatian groups from the Pontic Steppe and the Ural region.

Schütz et al., 2025, Cell *188*, 1–17
July 24, 2025
https://doi.org/10.1016/j.cell.2025.05.009

CellPress

Cell

CellPress
OPEN ACCESS

Article

Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin

Oszkár Schütz,[1,2,19] Zoltán Maróti,[2,3,19] Balázs Tihanyi,[2,4] Attila P. Kiss,[5] Emil Nyerki,[3] Alexandra Gînguță,[2] Petra Kiss,[1] Gergely I.B. Varga,[1,2] Bence Kovács,[1,2] Kitti Maár,[1,2] Bernadett Ny. Kovacsóczy,[6] Nikoletta Lukács,[7] István Major,[8] Antónia Marcsik,[4] Eszter Patyi,[6] Anna Szigeti,[8,9,10] Zoltán Tóth,[11] Dorottya Walter,[10] Gábor Wilhelm,[6] Réka Cs. Andrási,[12] Zsolt Bernert,[13] Luca Kis,[2,4] Liana Oța,[14] György Pálfi,[4] Gábor Pintye,[7] Dániel Pópity,[15] Angela Simalcsik,[16] Andrei Dorian Soficaru,[17] Olga Spekker,[4,18] Sándor Varga,[15] Endre Neparáczki,[1,2,18] and Tibor Török[1,2,18,20,*]

[1]Department of Genetics, University of Szeged, 6726 Szeged, Hungary
[2]Department of Archaeogenetics, Institute of Hungarian Research, 1041 Budapest, Hungary
[3]Department of Pediatrics and Pediatric Health Center, University of Szeged, 6725 Szeged, Hungary
[4]Department of Biological Anthropology, University of Szeged, 6726 Szeged, Hungary
[5]Faculty of Humanities and Social Sciences, Institute of Archaeology, Pázmány Péter Catholic University, 1088 Budapest, Hungary
[6]Katona József Museum, 6000 Kecskemet, Hungary
[7]Hungarian National Museum, Department of Archaeology, 1088 Budapest, Hungary
[8]International Radiocarbon AMS Competence and Training Center (INTERACT), HUN-REN Institute for Nuclear Research, 4026 Debrecen, Hungary
[9]Isotoptech Zrt., 4026 Debrecen, Hungary
[10]Department of Archaeology, University of Szeged, 6722 Szeged, Hungary
[11]István Dobó Castle Museum, 3300 Eger, Hungary
[12]Türr István Museum, 6500 Baja, Hungary
[13]Department of Anthropology, Hungarian Natural History Museum, 1083 Budapest, Hungary
[14]"Vasile Pârvan" Institute of Archaeology, 010667 Bucharest, Romania
[15]Móra Ferenc Museum, 6720 Szeged, Hungary
[16]"Olga Necrașov" Center of Anthropological Research, Romanian Academy - Iași Branch, 700481 Iași, Romania
[17]Department of Paleoanthropology, "Francisc J. Rainer" Institute of Anthropology, Romanian Academy, 050711 Bucharest, Romania
[18]Ancient and Modern Human Genomics Competence Centre, University of Szeged, 6726 Szeged, Hungary
[19]These authors contributed equally
[20]Lead contact
*Correspondence: torokt@bio.u-szeged.hu
https://doi.org/10.1016/j.cell.2025.05.009

SUMMARY

The nomadic Sarmatians dominated the Pontic Steppe from the 3rd century BCE and the Great Hungarian Plain from 50 CE until the Huns' 4th-century expansion. In this study, we present a large-scale genetic analysis of 156 genomes from 1st- to 5th-century Hungary and the Carpathian foothills. Our findings reveal minor East Asian ancestry in the Carpathian Basin (CB) Sarmatians, distinguishing them from other regional populations. Using F4 statistics, qpAdm, and identity-by-descent (IBD) analysis, we show that CB Sarmatians descended from Steppe Sarmatians originating in the Ural and Kazakhstan regions, with Romanian Sarmatians serving as a possible genetic bridge between the two groups. We also identify two previously unknown migration waves during the Sarmatian era and a notable continuity of the Sarmatian population into the Hunnic period despite a smaller influx of Asian-origin individuals. These results shed new light on Sarmatian migrations and the genetic history of a key population neighboring the Roman Empire.

INTRODUCTION

The Sarmatians were a group of nomadic people who likely originated from the southern Ural region during the 4th and 2nd centuries BCE, identified archaeologically with the Prokhorovka culture.[1] In the subsequent centuries, they gradually expanded into the Pontic Steppe territories, displacing the culturally related Scythians.[2,3] During the Iron Age, they established the first significant political formations in the area between the Don, Volga, North Caucasus, and Ural Mountains. Based on names preserved in ancient sources, they are believed to have been part of a group of northern Iranian-speaking people.[4]

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

By the 1st century CE, Sarmatian groups had settled in the area between the eastern foothills of the Carpathians and the Lower Danube region (modern Romania).[5] In the early decades CE, the first Sarmatian tribes, known as the Iazyges, entered the Carpathian Basin (CB), occupying the northern and central areas of the Danube-Tisza interfluve. They then gradually expanded into the Trans-Tisza region, eventually occupying the entire Great Hungarian Plain and likely extending their rule over the local Celtic and Scythian groups.

By the end of the 1st and the beginning of the 2nd century CE, despite initially good relations, they gradually became a formidable enemy of the Roman Empire in the Danube region.[4] After the Marcomannic-Sarmatian wars (166–180 AD), their material culture became a peripheral extension of the Roman Empire.[6–10] The dense settlement network in the CB within a century of their arrival indicates that the nomadic herders also adopted farming and achieved a large population size.[11–13] Despite this, many steppe traditions persisted in daily life, culture, and warfare. Sarmatians in the eastern CB maintained close contact with other Sarmatian groups in the steppe, and archaeological finds show several instances of eastern groups moving in during the late 2nd and 4th centuries.[4,14,15] They also formed close contacts and military alliances with Germanic tribes (Quadi, Marcomanni, Vandals), which significantly influenced their material culture, particularly in the surrounding border areas.[16–18]

It is intriguing that this once-dominant people, who ruled over a vast region and significantly influenced the ancient and early medieval world (military innovations, relations with the Roman Empire, and even ties to the Arthurian legend) are not claimed as ancestors by any modern European state-forming nations and remain a group of ancient, now forgotten, people.[4]

Previous archaeological research has classified Sarmatian archaeological remains in the CB into three chronological periods, based on the main historical events of Roman-Sarmatian relations and changes in material culture.[9,19] These periods are as follows: (1) Early Sarmatian Period, from the arrival of the Sarmatians in the Great Hungarian Plain (~50 CE) to the 2nd half of the 2nd century CE; (2) Middle Phase, From the period of the Marcomannic wars to the end of the 3rd century CE abandonment of Dacia; (3) Late Period, from the end of the 3rd century CE to the last third of the 5th century CE. Unfortunately, the Sarmatian archaeological chronology of the CB is not fully aligned with the chronological systems used for the central European Germanic and eastern European regions.[20]

Between the late 4th and mid-5th centuries, a major migration initiated by the Huns brought diverse communities, including eastern Germanic tribes, Huns, and other eastern Sarmatian groups, into the Great Hungarian Plain.[21,22] After the Hunnic Empire moved its center to the CB, many Sarmatians remained in their original homeland. Their cemeteries were used until the early 5th century, and their settlements continued until the mid-5th-century collapse of the Hun Empire.[12,22,23] According to written sources, the Sarmatians may have maintained an independent political organization until the 470s.[5] After the fall of the Hun Empire, they were assimilated into the population of the Gepid Kingdom.[24,25]

To date, 45 published Sarmatian genomes are available across seven different studies, from the Ural region and the Central Steppe.[26–32] Among these studies, only two articles provide a detailed discussion of the Sarmatians in context,[28,32] while the others address the topic only marginally. Key characteristics identified in the Uralic Sarmatians and Eastern Steppe Sarmatians (collectively referred to as Steppe Sarmatians) include the following: (1) their genomes exhibit the admixture of three main ancestral components, i.e., 70% Steppe Middle-Late Bronze Age (steppe_MLBA), 18% Bactria-Margiana Archaeological Complex (BMAC)-related, and 12% Baikal Early Bronze Age (Baikal_EBA)-Khovsgol-related; (2) their lower Khovsgol-related East Asian component, compared with the eastern Scythians, suggests they might have originated from distinct, independent Late Bronze Age populations in the Ural area; and (3) despite their extensive geographical distribution and relatively high genetic diversity, they remained genetically very homogeneous for over 500 years.

From the CB, 17 Sarmatian-period individuals were published in Gnecchi-Ruscone et al.[33] While detailed analyses were not provided, these genomes display a distinct shift toward European genetic profiles, compared with Steppe Sarmatians, raising questions about the potential relationships between these populations.

To clarify the origins and genetic relationships of the CB Sarmatians and to explore their connections to other populations from the Eurasian Steppe, as well as to local groups from preceding and succeeding periods, we sequenced 156 genomes from the CB and surrounding regions, spanning the Sarmatian and Hun periods (Figure 1A). We have shown that the CB Sarmatians are descendants of the Sarmatians from the Ural and Kazakhstan regions, who migrated from the Carpathian foothills in present-day Romania. The descendants of the substantial CB Sarmatian population formed a significant portion of the population during the subsequent Hun era.

RESULTS

Samples

Out of the 156 samples, 118 were collected from the Great Hungarian Plain, spanning the 1st to 4th centuries, representing the Sarmatian-period population of the CB Barbaricum, termed HUN_SARM. Two cemeteries used during the Sarmatian period clearly extended into the Hun period, leading to their classification as HUN_SARM_HUN. To gain insights into the arrival of the Sarmatians, we also sampled 17 individuals from the Romanian Plains, likely representing the incoming Sarmatian population (ROU_SARM). Additionally, we generated 21 whole-genome sequences from the 4th to 5th centuries (HUN_HUN) to assess the potential long-term impact of the Sarmatian population on the region. The 156 shotgun-sequenced whole genomes have a mean coverage of 1.42× (ranging from 0.24× to 3.75×) with negligible contamination (Table S1A).

The newly sequenced genomes were co-analyzed with 17 Sarmatian-period individuals published in Gnecchi-Ruscone et al.[33] and 9 Hun period individuals from Maróti et al.,[34] creating the most comprehensive genomic database of the region for these periods (Figure 1A; Table S1B).

The samples underwent a thorough review for accurate archaeological classification (for further details; see Data S1).

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

CellPress
OPEN ACCESS

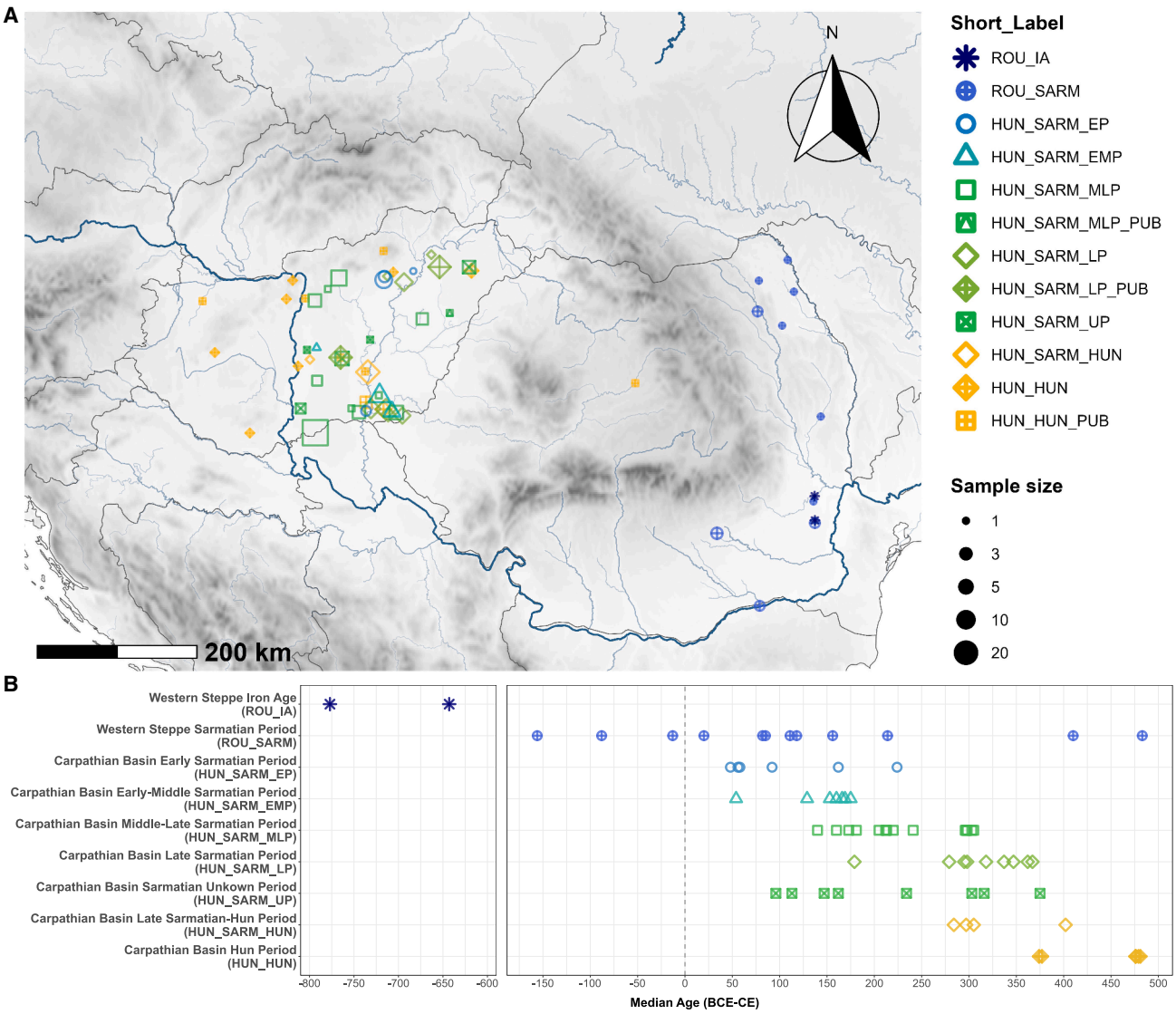


Figure 1. Archaeological sites and chronology of the studied samples

(A) Map showing the locations of the cemeteries analyzed in this study. Reanalyzed published samples are marked with the suffix "PUB" and are labeled as HUN_SARM_MLP_PUB (n = 1), HUN_SARM_LP_PUB (n = 16), and HUN_HUN_PUB (n = 9).

(B) Chart depicting the median age of the samples analyzed by radiocarbon dating. A gap is included to account for the significantly earlier radiocarbon dates of the two Iron Age individuals.

We also conducted radiocarbon measurements on 68 samples to anchor and validate the archaeological dating approach (Figure 1B; Table S1C). The two approaches yielded concordant dates in most cases; however, some samples showed indications of a possible reservoir effect (see Data S1). For this reason, we did not rely solely on radiocarbon results for sample grouping but instead integrated verified archeological data with radiocarbon dating. This led to the creation of a separate group for individuals with uncertain periodization, termed HUN_SARM_UP. Radiocarbon dating of two Romanian samples, LMO-8 and RAK-7, revealed that these individuals were significantly older, dating to the Early Iron Age, aligning with their sparse archeological descriptions. We include these individuals in the publication as Romania Iron Age (ROU_IA).

Based on the integrated dating procedure, we classified our samples into nine groups representing progressive archeological periods: ROU_IA, ROU_SARM, HUN_SARM_EP, HUN_SARM_EMP, HUN_SARM_MLP, HUN_SARM_LP, HUN_SARM_UP, HUN_SARM_HUN, and HUN_HUN (short labels are referred to in Tables S1A and S1B and other supplemental tables and figures).

Genetic kinship was determined using correctKin.[35] We identified 15 kin groups, some of which connects the Sarmatian and Hun periods directly to the Avar period (Table S1D).

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

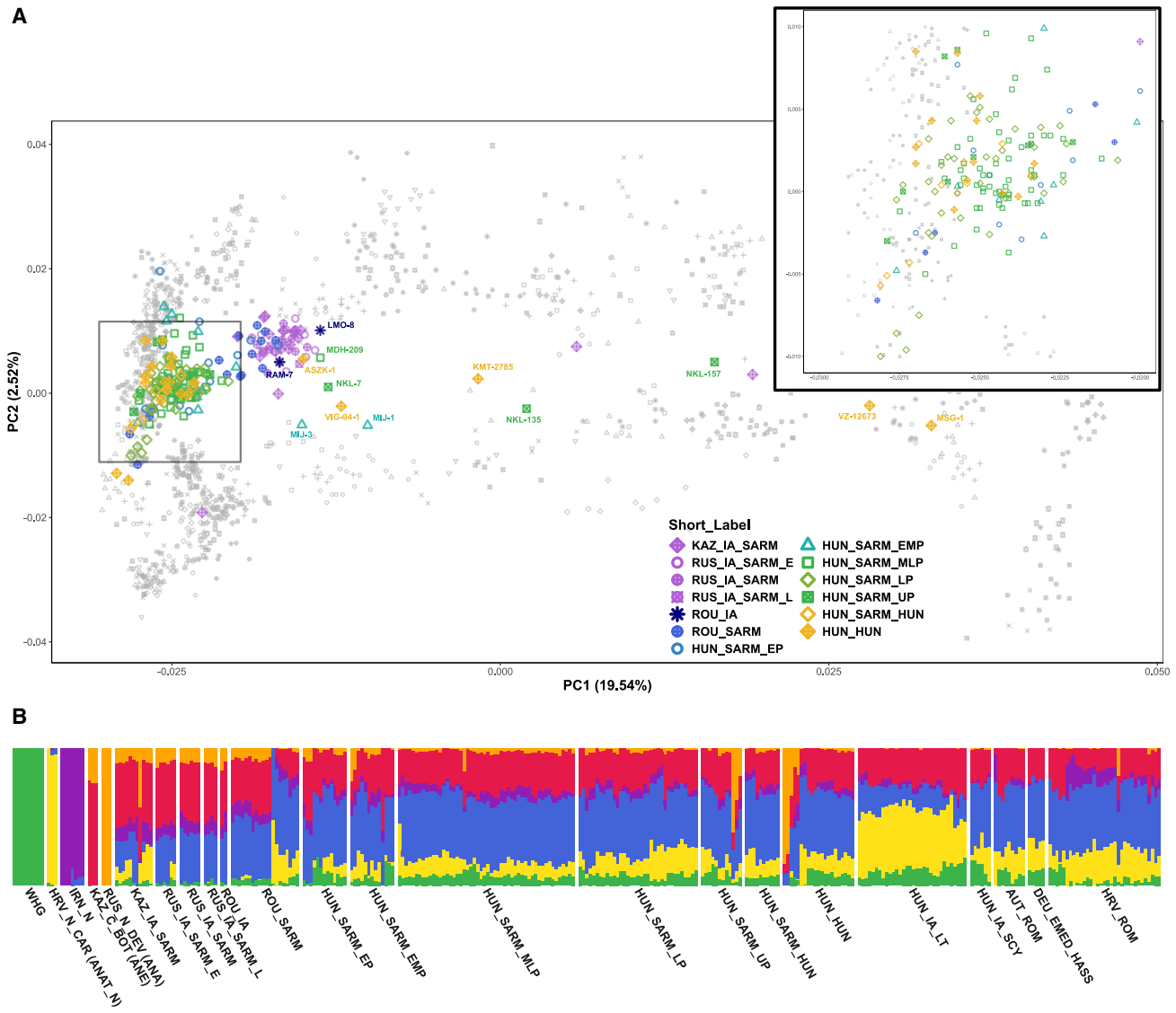


Figure 2. PCA and ADMIXTURE

(A) PCA of the studied individuals and all available Sarmatians from the literature projected onto a modern Eurasian background (gray symbols). Overlapping samples on the European side are enlarged for improved visibility in the inset.

(B) Unsupervised ADMIXTURE analysis results (K = 6) of the same samples arranged in chronological order. On the right, we also show selected contemporary populations from the CB for comparison. Groups best representing the main ancestral components are shown on the left side. Abbreviations are listed in Table S3A.

Genetic diversity of the Sarmatian- and Hun-period samples

To determine the underlying genetic structure among our samples. we first applied principal-component analysis (PCA) and ADMIXTURE.[36,37] We calculated primary PC axes from a present-day Eurasian population set (Table S2A) described in details in Maróti et al.,[34] and projected the studied genomes onto these axes (Figure 2A; Table S2B). In these analyses we also included the published 45 Steppe Sarmatians grouped by their geographic origin and age.

In line with previous findings, the Steppe Sarmatians are closely clustered together on the PCA, with a few outliers, and are distinctly separated from European populations. By contrast, most CB Sarmatians cluster near present-day central European populations, and only a few individuals, mainly from the early and early-middle periods, show a strong affinity toward the Steppe Sarmatian cluster. Within the European cluster, some individuals map more closely to modern northern European populations, while others align more with southern Europeans. Additionally, several genetic outliers are shifted toward Asian populations, suggesting connections to other nomadic groups beyond the Steppe Sarmatians. This ancestry, however, is also seen among some Steppe Sarmatian outliers, raising the possibility that these genetic outliers might have arrived alongside Steppe

Sarmatians. Notably, a significant portion of the Romanian Sarmatians is positioned among the Steppe Sarmatians, while others form a cline between the Steppe and CB Sarmatians.

The unsupervised ADMIXTURE results clearly indicate the genomic components responsible for the differing PCA positions of the individuals (Table S3A). At K = 6, we identified the typical macroregional ancestry components: Western Hunter-Gatherer (WHG), Anatolia Neolithic (ANAT_N), Iran Neolithic (IRAN_N), Ancient North Eurasian (ANE), and Ancient Northeast Asian (ANA) (Figure 2B; Table S3A). Additionally, a distinctive component (blue in Figure 2B) appeared, maximized in CB individuals, likely due to the overrepresentation of samples from this region in the analysis.

Most Sarmatian- and Hun-period individuals from the CB closely resemble contemporaneous populations from the region (e.g., Austria and Croatia Roman period—AUT_ROM, HRV_ROM) or populations from the immediately preceding period (e.g., Hungary Scythians—HUN_IA_SCY and Hungary Celtic—HUN_IA_LT). However, a distinctive feature of our studied groups is the presence of a small but significant fraction of the ANA component. This component is highest in the earliest groups (HUN_SARM_EP and EMP) and appears to decline over successive periods. The ANA component is significantly more pronounced in the Steppe Sarmatians and Romanian Sarmatians, who have very similar genome compositions, corresponding to their clustering on PCA and their significant shift from European groups (Figure 2B). The PCA Asian outlier individuals also stand out in the ADMIXTURE analysis due to their substantial ANA component, which is much larger than that of the Steppe Sarmatians.

The two Iron Age individuals from Romania show very surprising ADMIXTURE patterns, displaying identical component ratios with the RUS_IA_SARM group (Figure 2B; Table S3A), which are also reflected by their close PCA positions. This sharply distinguishes them from the contemporaneous Scythian and Celtic populations (HUN_IA_SCY or HUN_IA_LT).

CB Sarmatians show genetic affinity with Steppe Sarmatians

The East Asian ancestry found in the CB Sarmatians sets them apart from their contemporary neighbors and suggests that they may have descended from Steppe Sarmatians, as supported by historical and archeological sources. To test the potential genetic affinity between the two Sarmatian populations, we first reanalyzed the published Steppe Sarmatian individuals with sufficient coverage (Data S1). We were able to cluster them into two homogeneous yet similar groups, termed STEPPE_IA_SARM_URAL, STEPPE_IA_SARM_STEPPE, which we used as proxies in our ancestry analysis.

We applied F4 statistics in which we co-analyzed the Sarmatian- and Hun-period samples with the populations that previously inhabited the CB. First, we measured the direct affinity of the samples toward the Steppe Sarmatians against a Late Neolithic CB population (Lengyel culture), possibly representing local elements, with the statistics: F4(Ethiopia_4500BP, Test; HUN_LN_LGY, STEPPE_IA_SARM_URAL). Positive values in this statistic indicate a major affinity toward the Steppe Sarmatian proxy, while negative values show more shared drift toward the local proxy. This was plotted together with another combination, F4(Ethiopia_4500BP, STEPPE_IA_SARM_URAL; HUN_LN_LGY, Test), which uses the same references but actually measures the samples' affinity toward the Sarmatian proxy if we exclude their shared drift with the local proxy (Figure 3A; Table S4A).

Figure 3A illustrates that most samples, including those from the Scythian and Celtic groups in the Iron Age CB, share the majority of their markers with the local proxy. However, a few individuals, particularly from the ROU_SARM and HUN_SARM_EP groups, are positioned on the positive side of the *x* axis. This trend is further supported by the *y* axis, where nearly all individuals appear on the positive side, indicating at least a limited affinity toward the Steppe Sarmatian proxy in addition to their local affinity. The genomes located in the upper right quadrant of Figure 3A suggest a strong Steppe Sarmatian affinity. Notably, these samples also exhibit overlapping PCA positions and similar ADMIXTURE patterns to those of Steppe Sarmatians.

A few individuals in the lower right quadrant show strong negative F scores on the *y* axis but retain Steppe Sarmatian affinity on the *x* axis. They also have an elevated East Asian genomic component, as seen in PCA and ADMIXTURE, distinguishing them from the Steppe Sarmatian proxy. On the other hand, these samples share significant drift with Steppe Sarmatians on the *x* axis, likely due to their East Asian ancestry. This raises the possibility that other samples with Steppe Sarmatian affinity on the *y* axis might show similar patterns because of their elevated East Asian ancestry.

To address this uncertainty, we conducted two additional F4 analyses: (1) F4(Ethiopia_4500BP, MNG_EIA_SG; HUN_LN_LGY, Test) that measures the East Asian affinity of the samples using Mongolia_EIA_SlabGrave[38] as proxy, while excluding markers shared with the local proxy; and (2) F4(Ethiopia_4500BP, MNG_EIA_SG; STEPPE_IA_SARM_URAL, Test) that assesses the same East Asian affinity but excludes markers shared with the Steppe Sarmatian proxy. In Figure 3B, we plotted *Z* scores instead of F values, as the significance level of the statistics provides a more precise answer to our question.

The *x* axis of Figure 3B demonstrates that all individuals on the right side of Figure 3A show significant shared drift with the East Asian proxy beyond their local ancestry. However, the *y* axis reveals that the East Eurasian affinity of the samples in the upper right quadrant of Figure 3A is equivalent to the Uralic Sarmatians, as this affinity falls around zero when only considering markers not shared with the Uralic Sarmatian. In contrast, individuals in the lower right quadrant of Figure 3A show significant *Z* scores on both axes, indicating a higher level of East Asian genetic affinity that cannot be attributed to the Steppe Sarmatian proxy alone. The presence of these individuals suggests the existence of a source distinct from the Sarmatians.

While the two Iron Age groups from the CB, HUN_IA_SCY and HUN_IA_LT, show some affinity toward the Steppe Sarmatian proxy in Figure 3A (especially HUN_IA_SCY), this affinity likely stems from a different component. This is supported by Figure 3B, where these groups fall well below any significance line, indicating no detectable East Asian genetic affinity.

Next, we devised a qpAdm analysis framework to determine whether Steppe Sarmatians are essential for modeling the CB

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

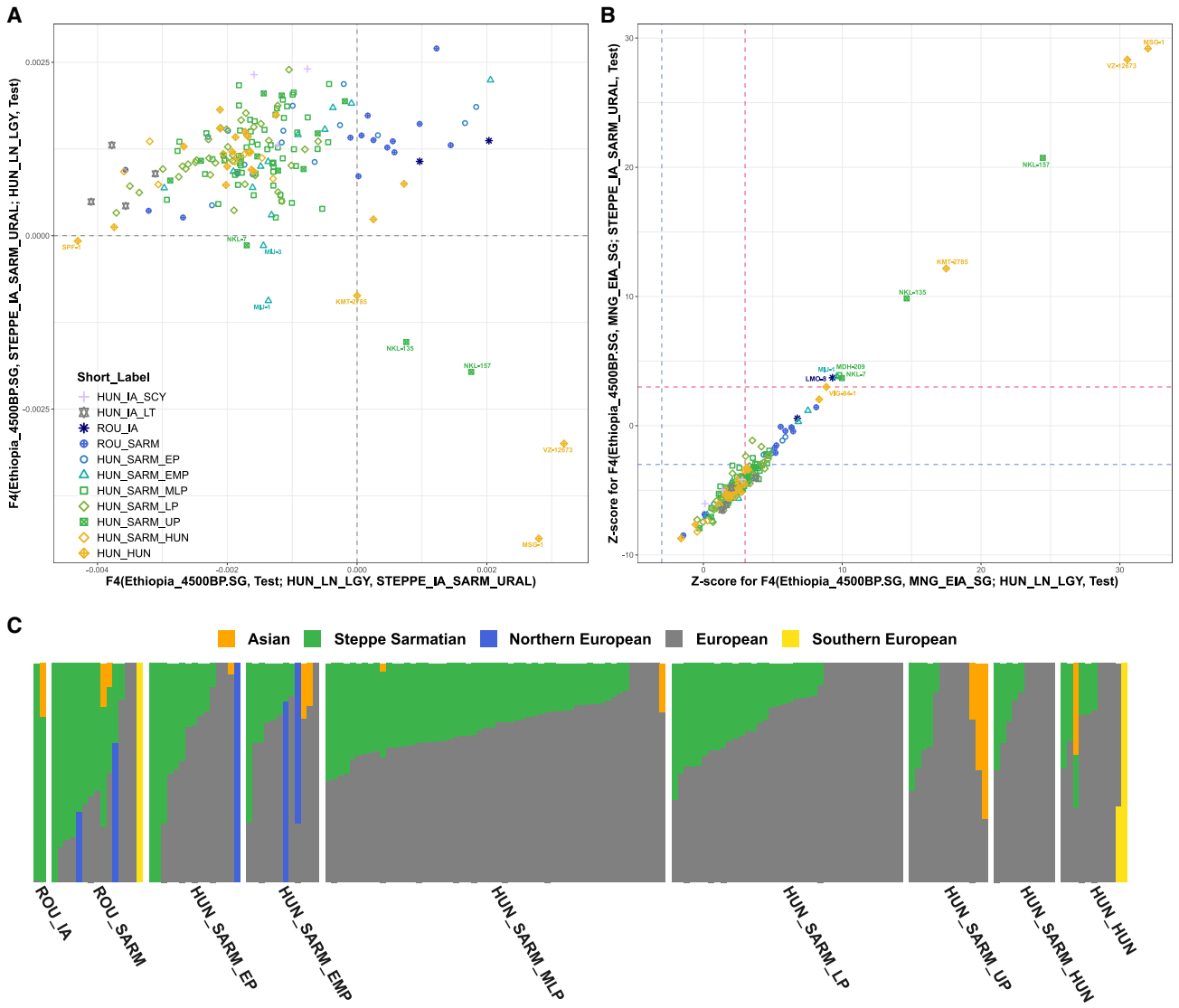


Figure 3. F4 statistics and summary of qpAdm results

(A) F scores of Steppe Sarmatian affinity measured in our samples and other Iron Age groups against local CB ancestry. The *x* axis shows the two-way affinity of the samples to either the local or the Sarmatian proxy, while the *y* axis measures the affinity toward the Steppe Sarmatian proxy after excluding their local marker set.

(B) *Z* scores of East Asian affinity of the samples measured by excluding their local marker set (*x* axis) or their marker set potentially shared with the Steppe Sarmatians (*y* axis). Blue segmented lines indicate the negative significance threshold (−3), red lines indicate the positive significance threshold (3). Samples with exceptionally high East Asian affinity are labeled.

(C) Simplified summary of the most plausible qpAdm models from Tables S5B–S5D, where selected models are indicated in the "used on qpAdm plot" column. Color code is given above the plot; models were arranged according to similarity of components.

Sarmatians. Based on preliminary qpAdm runs, we assembled a comprehensive set of 15 source populations to best represent the potential local CB elements in most individuals. The two Steppe Sarmatian groups were used as representative sources for the proposed Sarmatian immigrants. Additionally, we included eight other sources, consisting of central and Inner Asian populations with known connections to the CB or those that could represent further Asian immigrants independent of the Sarmatians. For further details on the analysis framework and the exact composition of the LEFT and RIGHT populations see the STAR Methods and Table S5.

Two individuals, DZS-41 and FKD-150, who were among the earliest Sarmatians in the CB according to archeological data and radiocarbon dating, along with one Romanian Sarmatian (OSU-1), produced unambiguous models, in which they formed a genetic clade with the Steppe Sarmatians. We successfully obtained appropriate two-source or three-source models for all other individuals using the local European and Steppe Sarmatian

sources (see Tables S5B and S5C), except for 12 samples. To model these remaining samples, we incorporated additional source candidates from a broader spatiotemporal range and successfully developed valid models for all but one individual, HVF-10 (Table S5D).

Among the Romanian Sarmatians, ~2/3 (9 out of 15) had over 50% Steppe Sarmatian ancestry, while only 12% of CB Sarmatians (14 of 118) showed this level.

We need to note that nearly half of the samples had very low East Eurasian components, making it impossible for qpAdm to identify the exact source. These samples included both Sarmatian and other central-East Asian sources in their feasible models, raising doubts about the precise origin of the East Asian ancestry. Nevertheless, many explicit models clarify some doubts. For instance, several models clearly show Steppe Sarmatian ancestry as a minor component (e.g., MDH-265, A181015). These often appear in the same cemetery as uncertain models, suggesting that Steppe Sarmatian ancestry is the most plausible source.

In summary, the hypothesis tests confirmed the F4 results, showing that most CB Sarmatians are best modeled with Steppe Sarmatian sources. Furthermore, the qpAdm models align with the periodization of our samples. Romanian Sarmatians have the highest Steppe Sarmatian component, consistent with their PCA positions. Among the CB Sarmatians, those with a predominant Steppe Sarmatian ancestry are primarily from the early and middle periods (HUN_SARM_EP and MLP groups). Only one individual from the subsequent Hun period (ASZK-1) shows a majority Steppe Sarmatian ancestry. However, this individual is most likely a Hun period immigrant from the same Steppe Sarmatian population, with a minor East Asian genetic component. Overall, all analyses suggest that Romanian Sarmatians may represent a genetic link between the Steppe and CB Sarmatians.

We obtained unambiguous models for the PCA Asian outlier Sarmatians (e.g., MIJ-1, MIJ-3, MDH-209, NKL-135). Their Eastern ancestry was modeled from Hun, Xiongnu, or Avar elite sources but never from Steppe Sarmatians, consistent with their substantial ANA ancestry components. However, most individuals with majority East Asian ancestry are found in the HUN_HUN group (NKL-157, KMT-2785, MSG-1, VZ-12673), indicating 4th-century population changes reported in historical sources. Despite the appearance of new Eastern immigrants, most HUN_SARM_LP, HUN_SARM_HUN, and HUN_HUN individuals still exhibit at least marginal Steppe Sarmatian ancestry, highlighting substantial population continuity.

IBD analysis links all Sarmatian groups

In order to explore the genealogical links across different geographic regions and time periods in central Europe and the Central Steppe, we conducted identity-by-descent (IBD) analyses. For this purpose, we selected and imputed 504 individuals spanning from the Iron Age to the early Middle Ages, additionally to the 158 genomes presented in this study.

The main criteria for selecting individuals were their spatiotemporal origin and the library preparation method, excluding capture-enriched genomes to avoid erroneous genotype inferences during imputation. We applied a stringent entry threshold for imputation recommended by Rubinacci et al.,[39] requiring a minimum of 0.5-fold coverage and low contamination, resulting in the exclusion of seven new samples. Additionally, we excluded the 17 Sarmatian individuals published in Gnecchi-Ruscone et al.[33] because of their capture sequencing method. The final list of samples used for IBD analyses is provided in Table S6B.

We identified IBD genomic segments of at least 8 centimorgans (cM) in length, using the ancIBD software[40] with optimizations described in the STAR Methods section. IBD connection networks were visualized as graphs using the Fruchterman-Reingold (FR) weight-directed algorithm.[41] In these graphs, individual samples are represented as points (vertices), and IBD connections between points are shown as lines (edges).

First, we investigated the genealogical links between different populations across different time periods. For this reason, we grouped the samples by archeological period and culture. This allowed us to examine intergroup connections among various European and CB groups, including relevant groups from the Central Steppe and Asia. In these analyses the clouds of groups were handled as single points and a group relation layout was calculated with the FR algorithm, where the weights were the number of IBD connections between each group (see details in STAR Methods). This way, the distribution of the clouds themselves actually reflects the connectedness between the groups (Figure 4).

The CB Sarmatians occupy a central position in this plot, alongside the Hun period and Avar period individuals from the CB published in Maróti et al.[34] This central position is not surprising, as these groups are populous and occupy a central temporal location, providing them with the greatest opportunities to produce detectable genealogies that connect earlier and later populations. However, this does not undermine the importance of their numerous connections with seemingly distant groups and especially with one another.

Interestingly, steppe-related groups—Scythians, Steppe Sarmatians, and Romanian Sarmatians—all cluster near the CB Sarmatians. These groups share the most IBD segments with the CB Sarmatians and with one another. It is especially notable that within these groups, the western Scythians (HUN_IA_SCY, MDA_IA_SCY, UKR_IA_CIMM, and UKR_IA_SCY) exhibit the fewest connections to any other group, including the nearby and contemporary CB Sarmatians, who have significantly more connections with the geographically more distant Steppe Sarmatians.

Also interesting is the Hun period group, which barely shows any pattern of intragroup relatedness and instead has a disproportionately high ratio of intergroup connections, especially with individuals from the preceding Sarmatian period and the subsequent Avar period. This suggests that the Hun period group may not represent a distinctly separate population.

These observations are further illustrated in Figure 4B, where we used another FR algorithm with a 100-iteration limit, allowing points to shift between groups based on their primary attraction forces. As can be seen, the Roman provincial samples along the borders of the plot barely moved, reflecting their sparse connectedness to the CB individuals. In contrast, the Sarmatians excavated in Romania and on the Central Steppe moved rapidly inward, clustering with the CB Sarmatians. As expected, the Hun period individuals also quickly lost their group coherence and shifted toward the Sarmatian- and Avar-period individuals.

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

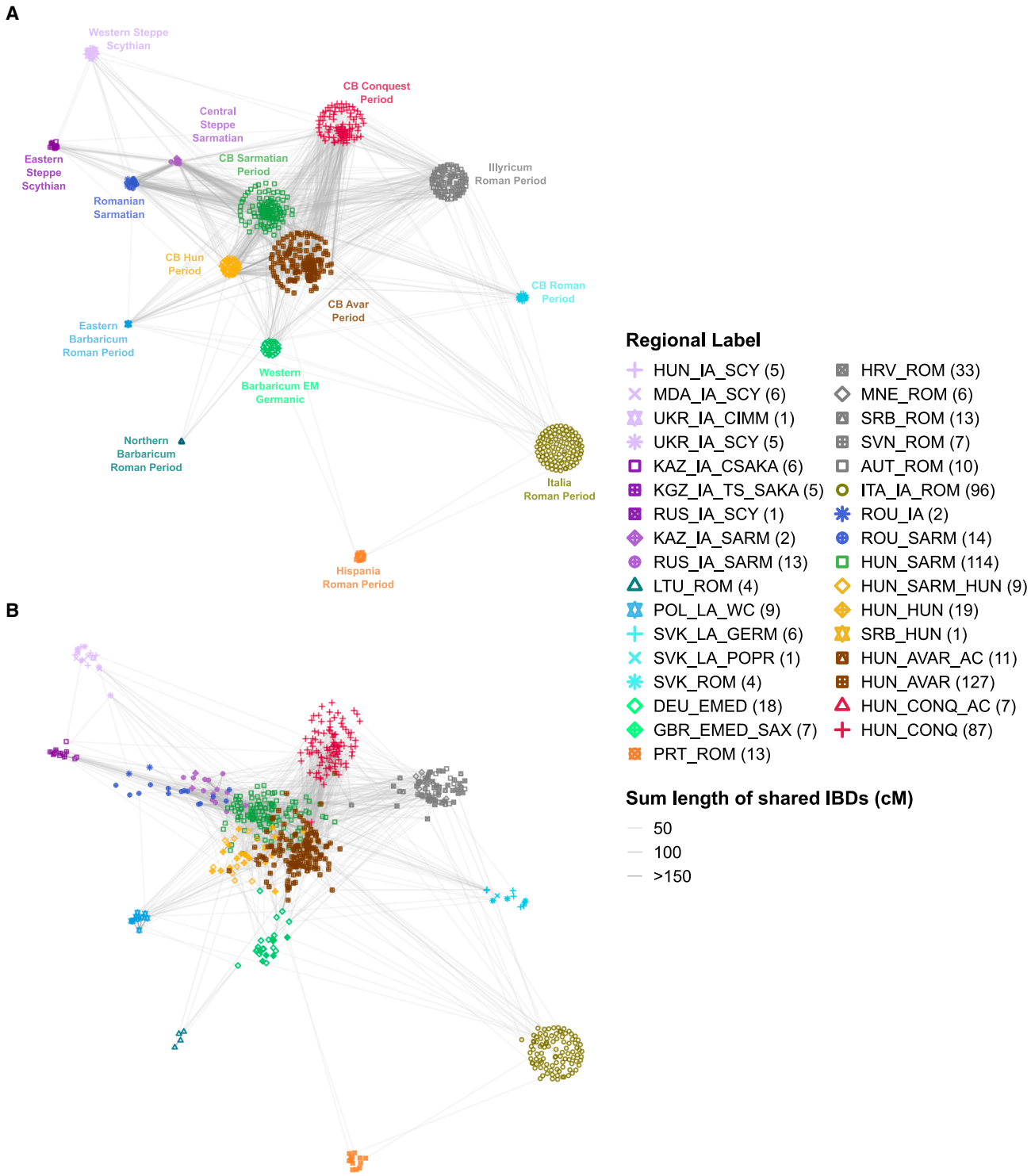


Figure 4. IBD sharing of groups potentially related to Sarmatians

(A) Intergroup IBD sharing graph of 662 ancient shotgun genomes, including the samples presented in this publication. Individuals were grouped according to geographical region and archaeological period (for further data see Table S6B).

(B) Vertices were allowed to reposition, driven by the FR weight-directed algorithm for 100 iterations. The maximum displacement of the points was reciprocally toned down by the number of outgoing connections (edges pointing outside the respective groups). Only edges representing intergroup connections are plotted. See also Figure S1.

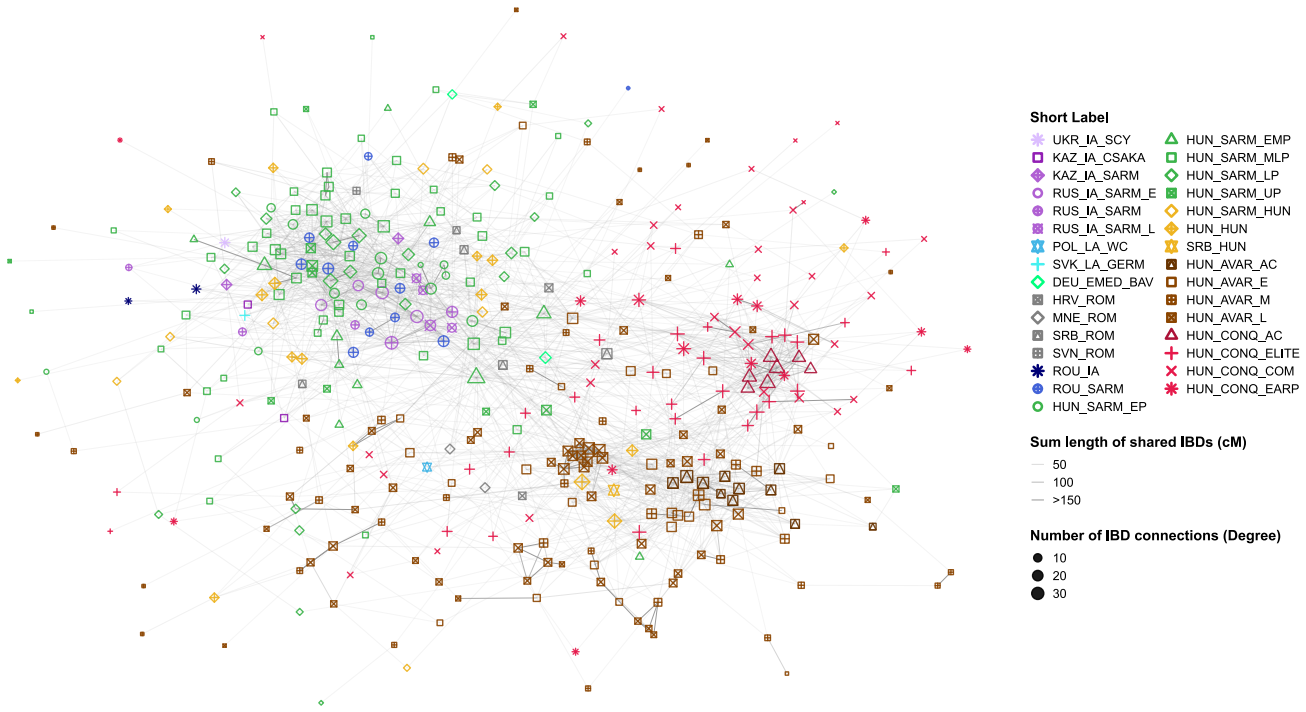


Figure 5. IBD sharing

IBD sharing graph of 423 selected individuals. Points represent individual samples. The size of a point is proportional to the number of connections (degree) a point has. Edges are shaded according to the total lengths of IBDs shared along them in cM. The individuals included in this plot are highlighted in Table S6B. See also Figures S1 and S3.

To better illustrate the IBD sharing patterns on an individual level, we prepared a graph with no grouping (Figure 5). Here, we included all studied groups from the CB (HUN_SARM, HUN_SARM_HUN, HUN_HUN, HUN_AVAR, and HUN_CONQ) and its vicinity (ROU_IA and ROU_SARM) as well as Steppe Sarmatians (KAZ_IA_SARM and RUS_IA_SARM). From the other groups shown in Figure 4A, we included only 17 samples that had at least 5 IBD connections to the individuals studied. This threshold was chosen to reduce the number of additional individuals to ≤10%, thereby reducing the clutter of the figure and improving its clarity. This resulted in a collection of 423 individuals, who were plotted with the same original coordinate positions as seen in Figure 4A, but the FR algorithm was allowed to freely reposition the points for 1,000 iterations. Individual IBD sharing data can be inspected in Table S6C.

In Figure 5, the three main groups, Sarmatian-, Avar-, and Conquest-period samples from the CB, form distinct clusters, reflecting high intraperiod connectivity. The first-generation immigrant "core" individuals of each medieval group appear to occupy central positions, particularly during the Conquest period, where they exhibit very high levels of both intragroup and intergroup connectedness. During the Sarmatian period, this central position is seemingly delegated to the Steppe Sarmatians, within the cluster of ROU_SARM and HUN_SARM_EP individuals. Conversely, the Hun period individuals do not form isolated groupings but instead seamlessly blend into the "Sarmatian cloud." Notably, some HUN_SARM_UP (e.g., NKL-157) and HUN_HUN individuals (e.g., MSG-1, VZ-12673, and KMT-2785) have the majority of their shared IBDs with later Avar period individuals (Figure S1). This pattern provides further evidence that Eastern immigrants distinct from the Steppe Sarmatians also appeared during these periods.

IBD connections between the periods suggest new migrations

Next, we analyzed pairwise combinations of group connectedness across subsequent time frames (Figure 6). We carefully normalized the degree centrality data by dividing the detected connections by the total number of possible connections, resulting in the ratio of fulfilled connections as described in the STAR Methods section. The *x* axis in Figure 6 displays the short label of each group (see Table S6B), while the columns represent the ratio of fulfilled IBD connections between the indicated group and every other group.

Figure 6 illustrates that the Steppe Sarmatians (STEPPE_IA_SARM) exhibit a consistently decreasing sharing pattern across the progressive time periods of the CB. This trend is consistent with a possible founding effect, where the genomic contribution of the earliest group naturally diminishes over subsequent generations. Additionally, the steadily declining pattern of intergroup connections suggests a continuous chain of generational transmission, without any abrupt population turnovers throughout the successive archeological periods.

A similar trend is observed in the connectedness of the ROU_SARM and HUN_SARM_EP groups with subsequent periods. However, there is a notable sharp decline in their

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

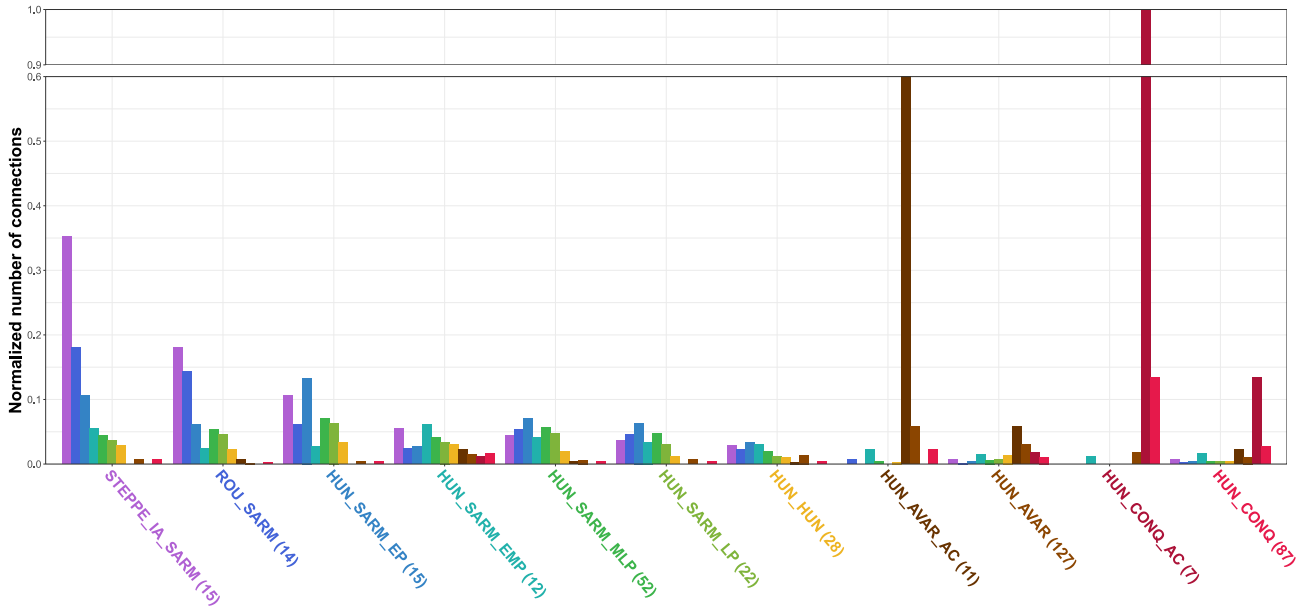


Figure 6. Normalized number of IBD connections across the different archaeological periods of the CB and its vicinity

The colors of the columns correspond to the colors of the group labels, with the height of each column representing the strength of the IBD connections for the group labeled with the letter code. The groups are arranged from left to right in chronological order. The plot has been truncated above the 0.6 line to accommodate the unusually high (100%) intragroup sharing of the HUN_CONQ_AC group. Normalized values were calculated as described in the STAR Methods section.

See also Figures S2 and S3.

connectedness with the HUN_SARM_EMP group, indicating a significant gap in IBD transmission during the early-middle Sarmatian period. The HUN_SARM_EMP group again shows a declining pattern of IBD connections with subsequent periods but also reveals extensive connections with the post-Sarmatian-, Avar-, and Conquest-period groups (HUN_AVAR and HUN_CONQ), which were negligible in the earlier Sarmatians.

This phenomenon is likely attributed to a second wave of immigration during the EMP period, involving new groups. The HUN_SARM_EMP group is represented by two large cemeteries, Makó-Igási Járandó (MIJ) and Hódmezővásárhely-Fehértó (HVF). F4 and qpAdm analyses identified at least two individuals from MIJ (MIJ-1 and MIJ-3) as potential migrants from eastern or central Asia. In contrast, four individuals from HVF (HVF-4, HVF-8, HVF-10, and HVF-21) showed significant northern European-related ancestry (Tables S5B and S5D), with three of these individuals being genetic outliers, and we were unable to model HVF-10 accurately (see Table S5A). This suggests that the HVF population is likely representing new migration from northern Europe. Nevertheless, the MIJ and HVF cemeteries do not fully represent the entire population of the HUN_SARM_EMP period, as the populations from the HUN_SARM_MLP and LP periods show a much stronger connection with the HUN_SARM_EP group.

The HUN_HUN group has the lowest IBD sharing within itself but shows high connections to Sarmatian and Avar periods, bridging the HUN_SARM_LP and HUN_AVAR groups. However, many individuals from this group were from solitary graves or single representatives, which may underestimate their true intragroup connectedness.

The so-called "immigrant cores" of the later Avar and Conquest periods (HUN_AVAR_AC and HUN_CONQ_AC) described by Maróti et al.[34] exhibit distinctive IBD sharing patterns, compared with other groups, with the exception of the STEPPE_IA_SARM. Their prominent intragroup IBD sharing and relatively low sharing with contemporary neighbors indicate a distinct, endogamous population. In contrast, the majority of sequenced individuals from the Avar and Conquest periods (HUN_AVAR and HUN_CONQ) exhibit a more regular sharing pattern, suggesting they likely represent a broader segment of the population from that era. Despite a reduction in connections, links to Sarmatian-period individuals are still observable in these later groups, indicating that at least a portion of the pre-Hun period population persisted in the CB. We also visualized the normalized intragroup and intergroup connectivity of each individual from each period in Figure S2, which further highlights that individuals from the three steppe groups, STEPPE_IA_SARM, HUN_AVAR_AC, and HUN_CONQ_AC, display an unusually high ratio of intragroup IBD sharing, compared with the other groups.

Plotting intergroup IBD sharing by cemetery (Figure S3) reveals a significant depletion of IBD connections in about half of the late Sarmatian- (HUN_SARM_LP) and Hun-period (HUN_HUN) cemeteries, compared with earlier periods. This phenomenon can be well explained for the Hun period, where significant new immigrants with elevated East Asian genomic components were detected. Most of their IBD connections are with later Avar and Conquest period groups (Figure S1), rather than with preceding Sarmatians.

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

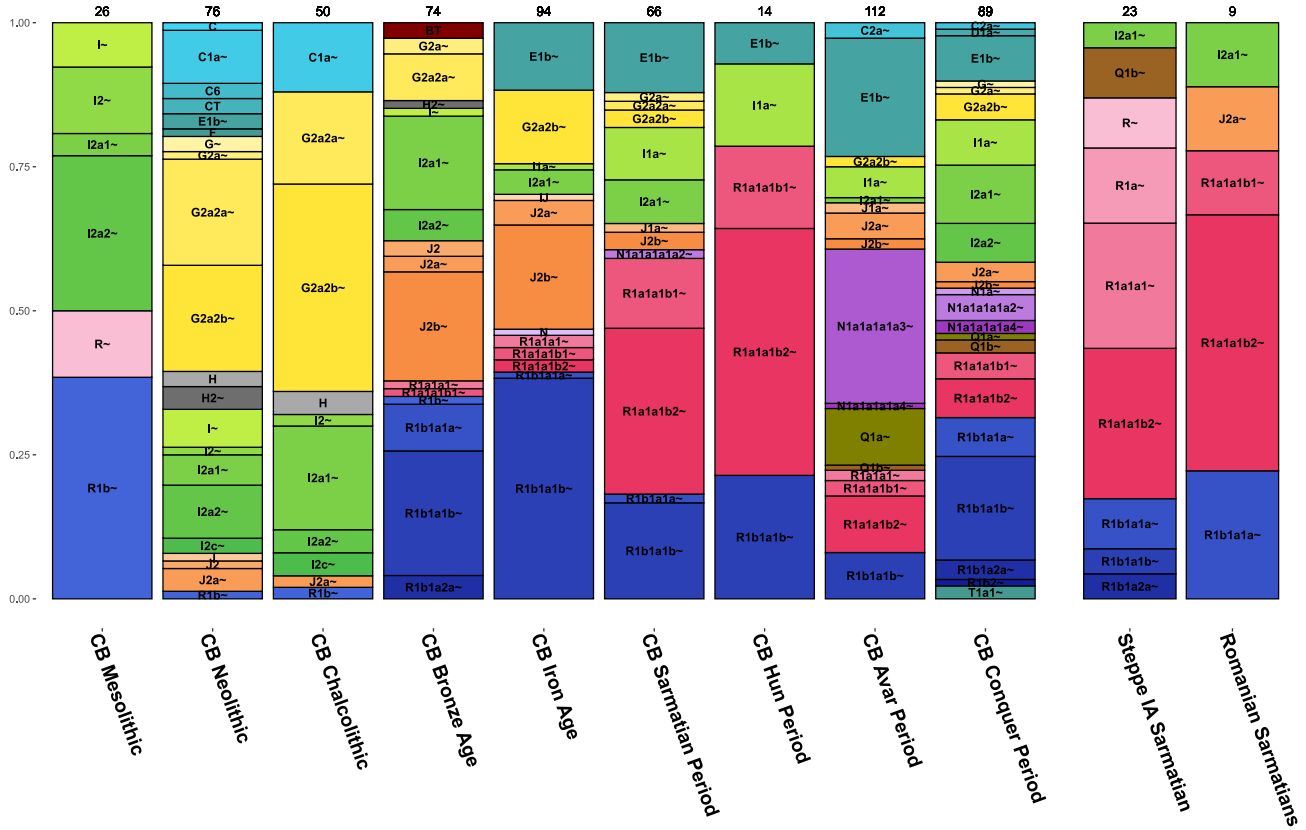


Figure 7. Y-haplogroup distribution of the progressive CB populations

Data were collected as described in the STAR Methods; further details are provided in Table S7.
See also Figure S4.

A similar pattern observed in the late Sarmatian period is also likely due to new immigration rather than sampling error, as it appears in half of the cemeteries from this period. This suggests that these cemeteries (OFU, TD, CSO, and SPT; see Table S1A for abbreviations) likely belong to a new wave of immigrants.

Male lineage turnover in the Sarmatian period

When examining the distribution of Y chromosome haplogroups (Y-Hg) across different periods in the CB (Figure 7), we observe significant turnover in male lineages over time. The Neolithic turnover linked to the migration of Anatolian farmers[42–44] and the Bronze Age turnover associated with the Yamnaya migrations[42,44,45] are well documented. However, during the Sarmatian period, we observe a new and previously unreported shift: the R1a∼ haplogroups R-Z283 and R-Z93 appear in abundance and remain prevalent into the Hun period. Particularly notable is the subclade R1a1a1b2∼ (R-Z93), which is characteristic of Middle-Late Bronze Age Steppe populations, such as the Sintashta and Srubnaya cultures,[31] and their Iron Age descendants, the eastern Scythians.[27,32,46] This haplogroup is also the most prevalent among the Steppe Sarmatians and Romanian Sarmatians, highlighting a direct link between all Sarmatian groups and reinforcing the conclusions drawn from the autosomal data. In contrast to the paternal lineages, the maternal lineages remained largely unchanged in the CB during the Sarmatian period (Figure S4).

Finally, the influx of Far Eastern Y-Hgs is observed during the Avar period, with similar changes noted during the Conquest period, as previously reported.[47,48]

DISCUSSION

Relation to Steppe Sarmatians

To study the origin and genetic relations of CB Sarmatians, we have compiled the most representative database of this era and region to date. We have shown that the CB Sarmatians differed significantly from Steppe Sarmatians and more closely resembled the earlier local populations, except for their small but significant ANA component. In contrast, most Sarmatians from outside the Carpathians in Romania were more similar to the Steppe Sarmatians and appeared to form a genetic link between the two groups. Nevertheless, we cannot entirely rule out the possibility of an island model-type migration event, in which the ROU_SARM and HUN_SARM groups both originated from the same original population.

It is quite striking that the Romanian and CB Sarmatians shared negligible IBDs with the geographically closest and immediately preceding western Scythians from Hungary,

Moldova, and Ukraine. Instead, their immediate source populations, the Steppe Sarmatians, evidently derived from the more remote Ural and Kazakh regions.

We demonstrated that most studied Sarmatians require a Steppe Sarmatian source in their genome modeling, with this component gradually diminishing over time. This pattern suggests a founder effect from a single migration, indicating that the newly arriving migrants were close descendants of Steppe Sarmatians who mixed with the local population after migrating to the CB. IBD data support this, showing strong connections between Steppe, Romanian, and CB Sarmatians. Notably, the FKD-150 CB Sarmatian female shares 4 IBD segments totaling 48.5 cM with the DA139 Steppe Sarmatian female from the Pontic Steppe. DA139 in turn shares 88 cM IBD with the chy001 Uralic Sarmatian and 64 cM with the POG-10 Romanian Sarmatian.

During the Sarmatian period, the male lineage composition in the CB underwent significant changes, highlighted by the rapid spread of R1a lineages. Notably, the Asian R1a-Z93 subclade gained prominence, clearly originating from the Steppe and Romanian Sarmatians, where this lineage is particularly common. It is worth noting that the available Steppe Sarmatian genomes generally have low coverage, so a large proportion of those classified under the broader R1a1a1 haplogroup likely belong to the R1a1a1b2-Z93 subclade. In contrast, the maternal lineages did not undergo significant changes (Figure S4), which suggests that the westward migration of the Sarmatians may have been primarily driven by male participants. Nevertheless, it is noteworthy that in early Sarmatian-period cemeteries, female burials such as those in the "Golden Horizon" graves were especially prominent.

These findings align with historical accounts suggesting that the migrating Sarmatians initially targeted the Roman Empire along its northern borders at the Lower Danube.

We observed unexpectedly high levels of intragroup IBD sharing among the analyzed Steppe Sarmatians as well as other groups recently arriving from the steppe (HUN_AVAR_AC and HUN_CONQ_AC). In contrast, the later Sarmatians who settled in the CB and transitioned to a more sedentary lifestyle display a decreasing trend of intragroup connectedness. We consider it plausible that the high connectedness observed in the steppe groups is primarily driven by their mobile nomadic lifestyle, while the declining pattern among the later CB Sarmatians is attributed to their lifestyle shift and increasing population size noted in Istvánovits and Kulcsár.[4]

Multiple new migration waves

Based on changes observed in Sarmatian archeological material, archaeologists hypothesize multiple waves of Sarmatian migration, suggesting the arrival of new populations during the late 2nd and 4th centuries. Our findings support this, indicating that new groups likely arrived during both the early-middle and late Sarmatian periods, roughly aligning with the archeological timeline.

In the EM period, the populations of the HVF and MIJ cemeteries show significant differences from both the Sarmatians and earlier local populations, and they share substantial IBD connections with Avar and Conqueror populations.

The HVF individuals show a genetic shift toward Northern European populations, with qpAdm analysis confirming the presence of Northern European ancestry in this cemetery. For example, HVF-4 and HVF-21 can be exclusively modeled from Scandinavian genomes, while HVF-8 forms a clade with the Poland Wielbark population. The local component of the remaining HVF individuals was typically modeled from Germany_EMedieval_Alemanic_SEurope.[49]

The MIJ individuals, on the other hand, appear to carry Northern European genomes admixed with East Asian ones, with most of them significantly shifted toward Asia in PCA. Their local components were typically modeled from Germany_EMedieval_Alemanic_SEurope, while MIJ-1 and MIJ-3 also carry 25% and 20% Xiongnu/Hun-Elite ancestry, respectively. Additionally, MIJ-7 and HVF-2, women with Sarmatian genetic affinity, are closely related, sharing 6 IBD segments with a total length of 142 cM.

These findings suggest that during the Sarmatian_EM period, there may have been two distinct migration waves: one from northwestern Europe, possibly related to the Germanic tribes of the Marcomannic Wars, and another from the Eastern Steppe, consisting of a population of East Asian origin distinct from the Sarmatians.

A likely second wave of migration detected in the late Sarmatian period is evident from individuals in the OFU, TD, CSO, and SPT cemeteries (Figure S3). These individuals have sparse IBD connections with the Sarmatians but show significant ties to the Avar period. In PCA, they align with the local European population, with three individuals showing a clear shift toward modern southern Europeans. In qpAdm analysis, all these individuals exhibited the most significant affinity to sources related to the Roman Empire (e.g., Germany_Roman.SG, Italy_IA_Republic.SG, Austria_Ovilava_Roman.SG, and Italy_Imperial.SG), with some local and Sarmatian admixture (Table S5B). Therefore, in the late Sarmatian period, the new migration likely came from neighboring Roman provinces rather than from the steppes.

It is important to note that, genetically, we cannot detect the possible new influx of groups with a similar composition to the first Sarmatian wave, especially if these migrations followed a stepping-stone pattern. For instance, while archeological data suggest possible elite migrations from the east between the late 2nd century and early 3rd century,[15] these have not been detected genetically.

All analyses identified five outlier individuals with elevated ANA ancestry from the Sarmatian period, which cannot be attributed to Steppe Sarmatians. These individuals were excavated from the MIJ, MDH, and NKL cemeteries. The two individuals from the MIJ site have already been discussed above. MDH-209, dating to the middle-late Sarmatian period, shares IBD segments with multiple Avar period samples, as well as with individuals from the early Sarmatian, Hun, and Roman periods.

The NKL cemetery is divided archeologically into two sections, one from the Sarmatian period and the other from the Hun period. However, all four Sarmatian NKL samples were grouped into the uncertain category (HUN_SARM_UP) because their radiocarbon dates were spread across an unrealistically wide time range. Despite this, the Eastern components of the NKL-7, NKL-135, and NKL-157 outliers are consistently modeled from Xiongnu/

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

Hun-elite ancestry, and they share IBD segments almost exclusively with Avar samples, including Avar elites (Table S6C). These findings strongly suggest that these individuals are more likely associated with migrations during the Hun period, although they may have arrived somewhat earlier.

The Hun period samples are distinctly separated into two IBD-sharing clusters (see Figure S1). Nearly all the newly sequenced HUN_HUN genomes carry local ancestry and are associated with the Sarmatian cluster, including Steppe Sarmatians, with only marginal connections to Roman, Avar, and Conquest period individuals. In contrast, the previously published Hun era genomes[34] contain significant Asian components and align with the Avar cluster, including a few Conqueror elites. These results clearly indicate that during the Hun period, most of the CB population represented the existing local population, which survived well into the Conquest period, while new Hun-era immigrants with Asian roots were in the minority. This is entirely consistent with historical data, which indicates that Sarmatian cemeteries were used until the early 5th century and settlements until the mid-5th century.[4]

Particularly noteworthy is the genome of the ASZK-1 individual from the Hun period, which forms a clade with Steppe Sarmatians in most qpAdm models, while also exhibiting some East Asian admixture in other valid models.[34] This solitary and rich Hun burial is well dated and bears numerous parallels to similar finds in the Kazakh Steppe. The burial customs and the entirety of the findings suggest an eastern individual from an eastern environment, making it likely that this individual arrived with the Huns.[50] This genome suggests that the descendants of the Steppe Sarmatians were also present among the incoming Huns.

The IBD connections separate the populations of the three successive migration waves (Sarmatian, Avar, and Hungarian Conquest) into three distinct clusters, in Figure 5, suggesting that the three migration waves are largely associated with different populations.

Iron Age samples

The two Early Iron Age samples LMO-8 and RAM-7 from the Carpathian foothills were contemporary neighbors of the European Scythians, radiocarbon dated 2.6–2.7 kyears BP and predating the first Steppe Sarmatians (Table S1C). The LMO-8 male had a Scythian-style bronze arrowhead within his ribcage, which may have caused his death, or perhaps he wore it as a necklace (Data S1). Surprisingly, despite the large temporal and geographical distances, the ADMIXTURE patterns of LMO-8 and RAM-7 are identical to those of the Steppe Sarmatians. In qpAdm models, RAM-7 nearly forms a clade with the Steppe Sarmatians, while LMO-8 appears to be approximately 75%–90% Steppe Sarmatian with about 10%–25% East or central Asian admixture. Additionally, the R1a1a1b2a2a∼ (R1a-Z2124) Y-Hg of LMO-8 aligns with the typical haplogroups found among the Steppe Sarmatians (RAM-7 being female).

Their potential connection to the Steppe Sarmatians is further supported by IBD data. LMO-8 shares IBDs with three Steppe Sarmatians, including one segment that is 22.5 cM long. Additionally, LMO-8 shares IBDs with a central Saka and a Tien Shan Saka, as well as with FKD-150, an early CB Sarmatian. Similarly, RAM-7 shares IBD with two early Sarmatians from the Ural region and with a Scythian from Moldova.

These data demonstrate that the two Iron Age individuals can indeed be considered genetically Sarmatian. The discovery of Sarmatian-like genomes in the two ROU_IA samples is particularly unexpected, given the 500–600-year gap between Steppe Sarmatians and the ROU_SARM samples. Although contemporary Iron Age samples from Romania are unavailable, neighboring regions, such as Moldova, Ukraine, and Hungary, do not exhibit the presence of Sarmatian-like ancestry during this period.[27,28,30]

Our new data suggest that migrations from the Ural region westward may have already occurred during the Early Iron Age, at least sporadically. It is worth noting that the age of these two individuals is much closer to the European appearance of the Cimmerians, and one well-covered genome identified as Cimmerian (MDA_IA_CIMM, Table S3A) does indeed show a similar ADMIXTURE pattern. Thus, it cannot be ruled out that their appearance may be connected to the Cimmerian migrations.

The ROU_IA and Steppe Sarmatian genomes are equally effective sources for qpAdm modeling of ROU_SARM (Table S5E), indicating potential continuity with Iron Age migrants. However, ROU_IA individuals are genetic outliers in the region during the Iron Age. Furthermore, historical and IBD evidence, such as the close genealogical link of FKD-150 (48.5 cM) to DA139 Steppe Sarmatian, argue against this interpretation.

Limitations of the study

The primary limitation of this study is the lack of comparative samples from the Carpathian foothills in Romania, dating to the period preceding the 1st century CE arrival of the Romanian Sarmatians. Additionally, a significant portion of the available Iron Age genomes from neighboring regions (Moldova, Ukraine) were either not sequenced using the shotgun method or have low coverage, making them unsuitable for imputation and IBD analysis. To some extent, this issue also applies to the Great Hungarian Plain, where Iron Age samples remain limited. These constraints hinder a detailed understanding of population dynamics both beyond the Carpathians and within the CB in the period preceding the Sarmatian migration into the region. Furthermore, the high genetic similarity among European genomes from different sub-regions during this period hinders the precise identification of "local" genome components, using statistical methods such as F statistics and qpAdm.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tibor Török (torokt@bio.u-szeged.hu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Aligned sequence data have been deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number European Nucleotide Archive: PRJEB80732 and are publicly available as of

the date of publication. Accession numbers are listed in the key resources table.

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at (www.github.com/zmaroti/scoreFilterIBD) and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We are grateful to our archaeologist colleagues—Csilla Balogh, János Dani, Csilla Farkas, Péter Gróf, Gyöngyi Gulyás, Eszter Istvánovits, Mónika Merczi, Boglárka Mészáros, Margit Nagy, Katalin Ottományi, Ágota S. Perémi, Enea Sergiu, Kornél Sóskúti, and Csaba Szalontai—for providing archaeological material. We are thankful to all the anthropologists who provided bone material for this study—Ágota Buzár, Sándor Évinger, Ana Ştefan, and János Rovó. This research was funded by grants from the National Research, Development and Innovation Office (TUDFO/5157-629 1/2019-ITM and TKP2020-NKA-23) to E. Neparáczki and a grant from the Ministry of Culture and Innovation (MCI-670-19/2023/FÁFIN) to T.T. and E. Neparáczki This research was partially funded by the Competence Centre of the Life Sciences Cluster of the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged to E. Neparáczki, Z.M., and T.T. (the authors are members of the "Ancient and modern human genomics competence center" research group). I.M. was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00710/23/10). The publication was supported by the University of Szeged Open Access Fund, grant ID: 7769 to T.T.

AUTHOR CONTRIBUTIONS

Conceptualization, T.T., O. Schütz, A.P.K., and B.T.; supervision, T.T.; project administration and funding acquisition, T.T. and E. Neparáczki; data curation and software, Z.M., O. Schütz, and E. Nyerki; formal analysis, validation, and methodology, Z.M. and O. Schütz; investigation, O. Schütz, Z.M., A.G., P.K., B.K., K.M., and G.I.B.V.; resources, B.N.K., N.L., I.M., A.M., E.P., A. Szigeti, Z.T., D.W., G.W., R.C.A., Z.B., L.K., L.O., G. Pálfi, G. Pintye, D.P., A. Simalcsik, A.D.S., O. Spekker, and S.V.; visualization, O. Schütz and Z.M.; writing – original draft, O. Schütz, T.T., B.T., A.P.K., Z.M., B.N.K., N.L., I.M., A.M., E.P., A. Szigeti, Z.T., D.W., and G.W.; writing – review & editing, all authors contributed.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ancient samples
- METHOD DETAILS
  - Ancient DNA preparation
  - NGS library construction
  - DNA sequencing
  - Radiocarbon dating
  - Map
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data processing and quality control
  - Sex determination
  - Haplogroup assignment and uniparental analysis
  - PCA
  - Unsupervised ADMIXTURE
  - qpAdm analysis and F-statistics
  - Imputation
  - Kinship estimation and IBD sharing analysis
  - Plotting

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.05.009.

Received: October 5, 2024
Revised: February 3, 2025
Accepted: May 8, 2025

REFERENCES

1. Balakhvantsev, A., and Yablonskii, L. (2009). A Silver Bowl from the New Excavations of the Early Sarmatian Burial-Ground Near the Village of Prokhorovka. Anc. Civil. Scythia Siberia *15*, 167–182. https://doi.org/10.1163/092907709X12474657004809.
2. Koryakova, L. (2018). Europe to Asia. In The Oxford Handbook of the European Iron Age, C. Haselgrove, K. Rebay-Salisbury, and P.S. Wells, eds. (Oxford University Press), pp. 1–41.
3. Mordvintseva, V. (2013). The Sarmatians: The Creation of Archaeological Evidence. Oxford J. Archaeology *32*, 203–219. https://doi.org/10.1111/ojoa.12010.
4. Istvánovits, E., and Kulcsár, V. (2017). Sarmatians History and Archaeology of a Forgotten People (Verlag des Römisch-Germanischen Zentralmuseums).
5. Kovács, P. (2023). Fontes Sarmatarum in Hungaria Habitantium – A Magyarországi Szarmatákra Vonatkozó Antik Forrásaink (Magyarságkutató Intézet).
6. Bene, Z., Istvánovits, E., and Kulcsár, V. (2016). Some characteristic types of Roman imports in Sarmatian Barbaricum in the Carpathian Basin (caskets decorated with metal mounts, bronze vessels, mirrors). In Archäologie zwischen Römern und Barbaren. Zur Datierung und Verbreitung römischer Metallarbeiten des 2. und 3. Jahrhunderts n. Chr. im Reich und im Barbaricum - ausgewählte Beispiele (Gefäße, Fibeln, Bestandteile militärischer Ausrüstung, Kleingerät, Münzen). Internationales Kolloquium Frankfurt am Main, N. Müller-Scheeßel and V. Hans-Ulrich, eds. (Dr. Rudolf Habelt GmbH), pp. 743–760.
7. Vaday, H., A., and Horváth, F. (2005). Corpus der Römischen Funde Im Europäischen Barbaricum. Ungarn. 1, Komitat Szolnok (Archäologisches Institut der UAW).
8. Istvánovits, E., and Kulcsár, V. (2020). Sarmatians on the Borders of the Roman Empire. Steppe Traditions and Imported Cultural Phenomena. Anc. Civil. Scythia Siberia *26*, 391–402. https://doi.org/10.1163/15700577-12341381.
9. Vaday, A.H. (1988). Die sarmatischen Denkmäler des Komitats Szolnok. Ein Beitrag zur Archäologie und Geschichte des sarmatischen Barbaricums (Archäologisches Institut der UAW).
10. Kovács, P. (2009). Marcus Aurelius' Rain Miracle and the Marcomannic Wars. . Mnemosyne, Supplements, *308* (Brill).
11. Istvánovits, E., and Kulcsár, V. (2015). Animals of the Sarmatians in the Carpathian Basin: Archaeozoology through the eyes of archaeologists. Materialy po Arheologii Istorii I Etnografii Tavrii *20*, 49–78.
12. Masek, Z. (2021). A római kori és kora középkori őséghajlati és környezettörténeti kutatások régészeti vonatkozásai. In A Kárpát-medence környezettörténete a középkorban és a kora újkorban, E. Benkő and C. Zatykó, eds. (Archaeolingua Alapítvány), pp. 111–146.
13. Pető, Á., Kenéz, Á., and Tóth, Z. (2017). Régészeti növénytani adatok a szarmaták mezőgazdaság- és gazdaság történeti kutatásához Hatvan–Baj-puszta és Apc–Farkas-major lelőhelyek alapján/Archaeobotanical

data on the economy of the sarmatians: the case study of Hatvan–Bajpuszta and Apc–Farkas-major (Heves county, Hungary). Archeometriai Műhely *14*, 117–128.

14. Khrapunov, I.N. (2001). On the contacts between the populations of the Crimea and the Carpathian Basin in the Late Roman Period. In International connections of the Barbarians of the Carpathian Basin in the 1st–5th centuries A. D. In Proceedings of the International Conference held in 1999 in Aszód and Nyíregyháza, E. Istvánovits and V. Kulcsár, eds. (Jósa András Múzeum), pp. 267–274.
15. Kulcsár, V. (1998). Újabb szempontok a hévízgyörki szarmata sírok etnikai meghatározásához. In Egy múzeum szolgálatában. Tanulmányok Asztalos István tiszteletére, T. Asztalos, ed. (Osváth Gedeon Múzeumi Alapítvány), pp. 75–84.
16. Istvánovits, E., and Kulcsár, V. (2003). Some traces of Sarmatian–Germanic contacts in the Great Hungarian Plain. In Kontakt – Kooperation – Konflikt. Germanen und Sarmaten zwischen dem 1. und dem 4. Jahrhundert nach Christus, C.V. Carnap-Bornheim, ed. (Wachholtz Verlag), pp. 227–238.
17. Istvánovits, E., and Kulcsár, V. (2000). Iranian-Germanic contacts in the Sarmatian Barbaricum of the Carpathian basin. In Die spätrömische Kaiserzeit und die frühe Völkerwanderungszeit in Mittel- und Osteuropa, M. Mączyńska and T. Grabarczyk, eds. (Wydawnictwo Uniwersytetu Łódzkiego), pp. 237–260.
18. Vaday, A.H. (2001). Military system of the Sarmatians. In International Connections of the Barbarians of the Carpathian Basin in the 1st-5th centuries A. D.: Proceedings of the international conference held in 1999 in Aszód and Nyíregyháza., E. Istvánovits and V. Kulcsár, eds. (Jósa András Múzeum), pp. 171–194.
19. Grumeza, L. (2014). Sarmatian Cemeteries from Banat (Late 1 st – Early 5 th Centuries AD) (Mega Publishing House).
20. Bârcă, V. (2014). Sarmatian Vestiges Discovered South of the Lower Mureș River. The Graves from Hunedoara Timișană and Arad (Mega Publishing House).
21. Soós, E. (2019). Békés együttélés vagy erőszakos hódítás?: Adatok a kontinuitás kérdéséhez a hun korban az északkelet-kárpát-medencei települések alapján = Peaceful Coexistence or Violent Conquest?: Data on the Question of Settlement Continuity in the North-eastern Part of the Carpathian Basin in the Hun Age. Pontes *2*, 123–158.
22. Tejral, J. (2011). Einheimische und Fremde. Das norddanubische Gebiet zur Zeit der Völkerwanderung (Archäologisches Institut der Akademie der Wissenschaften der Tschechischen Republik Brno).
23. Vaday, H., and A.. (1994). Late Sarmatian graves and their connections within the Great Hungarian Plain [Neskorosarmatské hroby a ich vzt'ahy v rámci Vel'kej uhorskej nížiny.]. Slov. Archeológia *42*, 105–124.
24. Kiss, A.P. (2015). ''...ut strenui viri...'' A gepidák Kárpát-medencei története (University of Szeged), PhD Thesis.
25. Kiss, P., and A.. (2021). Which came first, the chicken or the egg? The ethnic interpretations of the hoards of Șimleu Silvaniei / Szilágysomlyó: A case study in mixed argumentation. In Attila's Europe? Structural Transformation and Strategies of Success in the European Hun Period, Z. Rácz and G. Szenthe, eds. (Hungarian National Museum, Eötvös Lóránd University), pp. 477–500.
26. Unterländer, M., Palstra, F., Lazaridis, I., Pilipenko, A., Hofmanová, Z., Groß, M., Sell, C., Blöcher, J., Kirsanow, K., Rohland, N., et al. (2017). Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. Nat. Commun. *8*, 14615. https://doi.org/10.1038/ncomms14615.
27. Damgaard, P.B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J.V., Pedersen, M.W., Goldberg, A., Usmanova, E., et al. (2018). 137 ancient human genomes from across the Eurasian steppes. Nature *557*, 369–374. https://doi.org/10.1038/s41586-018-0094-2.
28. Krzewińska, M., Kılınç, G.M., Juras, A., Koptekin, D., Chyleński, M., Nikitin, A.G., Shcherbakov, N., Shuteleva, I., Leonova, T., Kraeva, L., et al. (2018). Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. Sci. Adv. *4*, eaat4457. https://doi.org/10.1126/sciadv.aat4457.
29. Veeramah, K.R., Rott, A., Groß, M., Van Dorp, L., López, S., Kirsanow, K., Sell, C., Blöcher, J., Wegmann, D., Link, V., et al. (2018). Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. Proc. Natl. Acad. Sci. USA *115*, 3494–3499. https://doi.org/10.1073/pnas.1719880115.
30. Järve, M., Saag, L., Scheib, C.L., Pathak, A.K., Montinaro, F., Pagani, L., Flores, R., Guellil, M., Saag, L., Tambets, K., et al. (2019). Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance. Curr. Biol. *29*, 2430–2441.e10. https://doi.org/10.1016/j.cub.2019.06.019.
31. Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. Science *365*, eaat7487. https://doi.org/10.1126/science.aat7487.
32. Gnecchi-Ruscone, G.A., Khussainova, E., Kahbatkyzy, N., Musralina, L., Spyrou, M.A., Bianco, R.A., Radzeviciute, R., Martins, N.F.G., Freund, C., Iksan, O., et al. (2021). Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. Sci. Adv. *7*, eabe4414. https://doi.org/10.1126/sciadv.abe4414.
33. Gnecchi-Ruscone, G.A., Szécsényi-Nagy, A., Koncz, I., Csiky, G., Rácz, Z., Rohrlach, A.B., Brandt, G., Rohland, N., Csáky, V., Cheronet, O., et al. (2022). Ancient genomes reveal origin and rapid trans-Eurasian migration of 7th century A var elites. Cell *185*, 1402–1413.e21. https://doi.org/10.1016/j.cell.2022.03.007.
34. Maróti, Z., Neparáczki, E., Schütz, O., Maár, K., Varga, G.I.B., Kovács, B., Kalmár, T., Nyerki, E., Nagy, I., Latinovics, D., et al. (2022). The genetic origin of Huns, Avars, and conquering Hungarians. Curr. Biol. *32*, 2858–2870.e7. https://doi.org/10.1016/j.cub.2022.04.093.
35. Nyerki, E., Kalmár, T., Schütz, O., Lima, R.M., Neparáczki, E., Török, T., and Maróti, Z. (2023). correctKin: an optimized method to infer relatedness up to the 4th degree from low-coverage ancient human genomes. Genome Biol. *24*, 38. https://doi.org/10.1186/s13059-023-02882-4.
36. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.
37. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. PLOS Genet. *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.
38. Wang, C.-C., Yeh, H.-Y., Popov, A.N., Zhang, H.-Q., Matsumura, H., Sirak, K., Cheronet, O., Kovalev, A., Rohland, N., Kim, A.M., et al. (2021). Genomic insights into the formation of human populations in East Asia. Nature *591*, 413–419. https://doi.org/10.1038/s41586-021-03336-2.
39. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. *53*, 120–126. https://doi.org/10.1038/s41588-020-00756-0.
40. Ringbauer, H., Huang, Y., Akbari, A., Mallick, S., Olalde, I., Patterson, N., and Reich, D. (2024). Accurate detection of identity-by-descent segments in human ancient DNA. Nat. Genet. *56*, 143–151. https://doi.org/10.1038/s41588-023-01582-w.
41. Fruchterman, T.M.J., and Reingold, E.M. (1991). Graph drawing by force-directed placement. Softw. Pract. Exp. *21*, 1129–1164. https://doi.org/10.1002/spe.4380211102.
42. Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., et al. (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature *551*, 368–372. https://doi.org/10.1038/nature24476.

43. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503. https://doi.org/10.1038/nature16152.

44. Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., et al. (2018). The genomic history of southeastern Europe. Nature *555*, 197–203. https://doi.org/10.1038/nature25778.

45. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172. https://doi.org/10.1038/nature14507.

46. Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., Wu, X., Cao, P., Liu, Y., Yang, R., et al. (2021). Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell *184*, 3829–3841. e21. https://doi.org/10.1016/j.cell.2021.05.018.

47. Neparáczki, E., Maróti, Z., Kalmár, T., Maár, K., Nagy, I., Latinovics, D., Kustár, Á., Pálfi, G., Molnár, E., Marcsik, A., et al. (2019). Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin. Sci. Rep. *9*, 16569. https://doi.org/10.1038/s41598-019-53105-5.

48. Csáky, V., Gerber, D., Koncz, I., Csiky, G., Mende, B.G., Szeifert, B., Egyed, B., Pamjav, H., Marcsik, A., Molnár, E., et al. (2020). Genetic insights into the social organisation of the Avar period elite in the 7th century AD Carpathian Basin. Sci. Rep. *10*, 948. https://doi.org/10.1038/s41598-019-57378-8.

49. O'Sullivan, N., Posth, C., Coia, V., Schuenemann, V.J., Price, T.D., Wahl, J., Pinhasi, R., Zink, A., Krause, J., and Maixner, F. (2018). Ancient genome-wide analyses infer kinship structure in an Early Medieval Alemannic graveyard. Sci. Adv. *4*, eaao1262. https://doi.org/10.1126/sciadv.aao1262.

50. Tomka, P. (2001). The Grave of Árpás from the 5th century. Arrabona *39*, 165.

51. Maár, K., Varga, G.I.B., Kovács, B., Schütz, O., Maróti, Z., Kalmár, T., Nyerki, E., Nagy, I., Latinovics, D., Tihanyi, B., et al. (2021). Maternal Lineages from 10–11th Century Commoner Cemeteries of the Carpathian Basin. Genes *12*, 460. https://doi.org/10.3390/genes12030460.

52. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

53. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

54. Pebesma, E., and Bivand, R. (2023). Spatial Data Science: With Applications in R (Chapman and Hall/CRC) https://doi.org/10.1201/9780429459016.

55. Dunnington, D. (2023). ggspatial: Spatial Data Framework for ggplot2.

56. Massicotte, P., and South, A. (2023). rnaturalearth: World Map Data from Natural Earth.

57. Hollister, J., Shah, T., Nowosad, J., Robitaille, A.L., Beck, M.W., and Johnson, M. (2023). jhollist/elevatr: CRAN Release v0.99.0 https://doi.org/10.5281/zenodo.8335450.

58. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

59. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

60. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

61. Broad Institute (2019). Picard Toolkit. Broad Institute, GitHub Repository.

62. Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., and Wegmann, D. (2017). ATLAS: Analysis Tools for Low depth and Ancient Samples. Preprint at bioRxiv. https://doi.org/10.1101/105346.

63. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356. https://doi.org/10.1186/s12859-014-0356-4.

64. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193.

65. Renaud, G., Slon, V., Duggan, A.T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. *16*, 224. https://doi.org/10.1186/s13059-015-0776-0.

66. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

67. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58–W63. https://doi.org/10.1093/nar/gkw233.

68. Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. Mol. Biol. Evol. *35*, 1291–1294. https://doi.org/10.1093/molbev/msy032.

69. Harney, É., Patterson, N., Reich, D., and Wakeley, J. (2021). Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics *217*, iyaa045. https://doi.org/10.1093/genetics/iyaa045.

70. Sousa da Mota, B., Rubinacci, S., Cruz Dávalos, D.I., G Amorim, C.E., Sikora, M., Johannsen, N.N., Szmyt, M.H., Włodarczak, P., Szczepanek, A., Przybyła, M.M., et al. (2023). Imputation of ancient human genomes. Nat. Commun. *14*, 3660. https://doi.org/10.1038/s41467-023-39202-0.

71. Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., and Noom, D. (2024). igraph for R: R interface of the igraph library for graph theory and network analysis. https://doi.org/10.5281/zenodo.7682609.

72. Varga, G.I.B., Kristóf, L.A., Maár, K., Kis, L., Schütz, O., Váradi, O., Kovács, B., Gînguță, A., Tihanyi, B., Nagy, P.L., et al. (2023). The archaeogenomic validation of Saint Ladislaus' relic provides insights into the Árpád dynasty's genealogy. J. Genet. Genomics *50*, 58–61. https://doi.org/10.1016/j.jgg.2022.06.008.

73. Harney, É., Cheronet, O., Fernandes, D.M., Sirak, K., Mah, M., Bernardos, R., Adamski, N., Broomandkhoshbacht, N., Callan, K., Lawson, A.M., et al. (2021). A minimally destructive protocol for DNA extraction from ancient teeth. Genome Res. *31*, 472–483. https://doi.org/10.1101/gr.267534.120.

74. Meyer, M., and Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor Protoc. *2010*, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448.

75. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil – DNA – glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624. https://doi.org/10.1098/rstb.2013.0624.

76. Molnár, M., Rinyu, L., Veres, M., Seiler, M., Wacker, L., and Synal, H.-A. (2013). EnvironMICADAS: A Mini 14C AMS with Enhanced Gas Ion Source Interface in the Hertelendi Laboratory of Environmental Studies (HEKAL), Hungary. Radiocarbon *55*, 338–344. https://doi.org/10.1017/S0033822200057453.

77. Molnár, M., Janovics, R., Major, I., Orsovszki, J., Gönczi, R., Veres, M., Leonard, A.G., Castle, S.M., Lange, T.E., Wacker, L., et al. (2013). Status Report of the New AMS 14C Sample Preparation Lab of the Hertelendi Laboratory of Environmental Studies (Debrecen, Hungary). Radiocarbon *55*, 665–676. https://doi.org/10.1017/S0033822200057829.

78. Reimer, P.J., Austin, W.E.N., Bard, E., Bayliss, A., Blackwell, P.G., Bronk Ramsey, C., Butzin, M., Cheng, H., Edwards, R.L., Friedrich, M., et al. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). Radiocarbon *62*, 725–757. https://doi.org/10.1017/RDC.2020.41.

79. Yan, D., Li, C., Zhang, X., Wang, J., Feng, J., Dong, B., Fan, J., Wang, K., Zhang, C., Wang, H., et al. (2022). A data set of global river networks and corresponding water resources zones divisions v2. Sci. Data *9*, 770. https://doi.org/10.1038/s41597-022-01888-0.

80. Andrews, S. (2023). FastQC: A quality control tool for high throughput sequence data Version 0.12.0. .

81. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., et al. (2011). An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science *334*, 94–98. https://doi.org/10.1126/science.1211177.

82. Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. J. Archaeol. Sci. *40*, 4477–4482. https://doi.org/10.1016/j.jas.2013.07.004.

83. Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., and Reich, D. (2024). The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci. Data *11*, 182. https://doi.org/10.1038/s41597-024-03031-7.

84. Freeman, L., Brimacombe, C.S., and Elhaik, E. (2020). aYChr-DB: a database of ancient human Y haplogroups. NAR Genom. Bioinform. *2*, lqaa081. https://doi.org/10.1093/nargab/lqaa081.

85. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.

86. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

87. Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D., and Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. J. Stat. Soft. *48*, 1–18. https://doi.org/10.18637/jss.v048.i04.

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human archaeological remains	This paper	N/A
Critical commercial assays		
MinElute PCR Purification Kit	QIAGEN	Cat. No.: 28006
Accuprime Pfx Supermix	ThermoFisher Scientific	Cat. No.: 12344040
TapeStation 2200 System	Agilent	G2964AA
iSeq 100 i1 Reagent v2 (cartridge + flow cell)	Illumina	Cat. No.: 20031374
Deposited data		
Human reference genome NCBI build 37,GRCh37	Genome Reference; Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Modern comparison dataset	Allen Ancient DNA Resource (V54.1.p1)[51]	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
Ancient comparison dataset	Allen Ancient DNA Resource (V54.1.p1)[51]	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
Ancient comparison dataset (DamgaardNature2018)	Damgaard et al.[27]	https://www.ebi.ac.uk/ena/browser/home
Ancient comparison dataset (KrzewinskaScienceAdvances2018)	Krzewińska et al.[28]	https://www.ebi.ac.uk/ena/browser/view/PRJEB27628
Ancient comparison dataset (JarveCurrentBiology2019)	Järve et al.[30]	https://www.ebi.ac.uk/ena/browser/view/PRJEB32764
Ancient comparison dataset (AntoniobioRxiv2022)	Reference in Table S6B	https://www.ebi.ac.uk/ena/browser/view/PRJEB53565
Ancient comparison dataset (VeeramahPNAS2018)	Veeramah et al.[29]	https://www.ebi.ac.uk/ena/browser/view/PRJEB23079
Ancient comparison dataset (SchiffelsNatureCommunications2016)	Reference in Table S6B	https://www.ebi.ac.uk/ena/browser/view/PRJEB4604 https://www.ebi.ac.uk/ena/browser/view/PRJEB6915
Ancient comparison dataset (AntonioGaoMootsScience2019)	Reference in Table S6B	https://www.ebi.ac.uk/ena/browser/view/PRJEB32566
Ancient comparison dataset (MarotiTorokCurrBio2022)	Maróti et al.[34]	https://www.ebi.ac.uk/ena/browser/view/PRJEB49971
Newly published ancient genomes	This paper	https://www.ebi.ac.uk/ena/browser/view/PRJEB80732
Oligonucleotides		
Illumina specific adapters	Custom synthetized	https://www.sigmaaldrich.com/HU/en/product/sigma/oligo?lang=en&region=US&gclid=CjwKCAiAgvKQBhBbEiwAaPQw3FDDFnRPc3WV75qapsXvcTxxzBXy48atqyb6Xi5f8e6Df2EJI0NNhoCmzIQAvD_BwE
Software and algorithms		
R 4.1.0	R Core Team[52]	https://cran.r-project.org/bin/windows/base/old/4.1.0/
ggplot2 (3.4.2) package	Wickham[53]	https://cran.r-project.org/web/packages/ggplot2/index.html
Sf	Pebesma and Bivand[54]	https://r-spatial.github.io/sf/
Ggspatial	Dunnington[55]	https://cran.r-project.org/web/packages/ggspatial/index.html

(*Continued on next page*)

Please cite this article in press as: Schütz et al., Unveiling the origins and genetic makeup of the "forgotten people": A study of the Sarmatian-period population in the Carpathian Basin, Cell (2025), https://doi.org/10.1016/j.cell.2025.05.009

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Naturalearth	Massicotte and South[56]	https://www.rdocumentation.org/packages/rnaturalearth/versions/1.0.1
Elevatr	Hollister et al.[57]	https://cran.r-project.org/web/packages/elevatr/index.html
Cutadapt	Martin[58]	https://cutadapt.readthedocs.io/en/stable/#
Burrow-Wheels-Aligner	Li and Durbin[59]	http://bio-bwa.sourceforge.net/
Samtools	Li et al.[60]	http://www.htslib.org/
PICARD tools	Picard toolkit[61]	https://github.com/broadinstitute/picard
ATLAS software package	Link et al.[62]	https://bitbucket.org/wegmannlab/atlas/wiki/Home
ANGSD software package	Korneliussen et al.[63]	https://github.com/ANGSD/angsd
MapDamage 2.0	Jónsson et al.[64]	https://ginolhac.github.io/mapDamage/
Schmutzi software package	Renaud et al.[65]	https://github.com/grenaud/schmutzi
Mosdepth software	Pedersen and Quinlan[66]	https://github.com/brentp/mosdepth
HaploGrep 2	Weissensteiner et al.[67]	https://haplogrep.i-med.ac.at/category/haplogrep2/
Yleaf software tool	Ralf et al.[68]	https://github.com/genid/Yleaf
Smartpca	Patterson et al.[37]	https://github.com/chrchang/eigensoft/blob/master/POPGEN/README
ADMIXTURE software	Alexander et al.[36]	https://dalexander.github.io/admixture/
ADMIXTOOLS software package	Harney et al.[69]	https://github.com/DReichLab/AdmixTools
GLIMPSE2	Rubinacci et al.[39]	https://odelaneau.github.io/GLIMPSE/
correctKin	Nyerki et al.[35]	https://github.com/zmaroti/correctKin
ancIBD (version 0.5)	Ringbauer et al.[40]	https://ancibd.readthedocs.io/en/latest/index.html
scoreFilterIBD	integrated into ancIBD	www.github.com/zmaroti/scoreFilterIBD
igraph package	Sousa da Mota et al.[70]	https://igraph.org
qgraph package	Csárdi et al.[71]	https://cran.r-project.org/web/packages/qgraph/index.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ancient samples

To uncover the genetic composition of the Sarmatian population living in the Carpathian Basin Barbaricum we collected bone samples from a wide spatio-temporal range (from the I-Vth c. CE). From the sampled 244 individuals we successfully generated whole genome shotgun sequences in 156 cases (Table S1A). The unsuccessful attempts were mainly the cause of low endogenous DNA content. To ensure transparency we publish the Master ID and cemetery label of the unsuccessfully sampled individuals in Table S1E.

The human bone material used for ancient DNA analysis in this study were obtained from anthropological collections or museums, with the permission of the custodians in each case. In addition, we also contacted the archaeologists who excavated and described the samples, as well as the anthropologists who published anthropological details. In most cases these experts became co-authors of the paper, who provided the archaeological background, which is detailed in Data S1.

METHOD DETAILS

Ancient DNA preparation

All steps of sampling, DNA extraction and library preparation were carried out as described in the Supplementary materials of G. I. B. Varga et al.,[72] in the joint, dedicated ancient DNA laboratory of the Department of Archaeogenetics, Institute of Hungarian Research and the Department of Genetics, University of Szeged.

The bone samples were prepared with a minimally invasive extraction method described in Harney et al.[73] We primarily collected teeth samples, where it was feasible and their connection to other bones of the studied individual could be securely determined. In every other case we sampled the petrous bone. Bone pieces were prepared with care using a Dremel® 3000 multifunctional hand drill and powdered with a VWR™ Star-Beater ball grinder.

DNA extraction was carried out by cleaning and soaking the whole teeth or 200 mg bone powder in digestion buffer (0.45 M EDTA, 250 μg/ml Proteinase-K, 0.1% Triton X-100) for 72 hours on 48 °C, than binding the DNA on Qiagen™ MinElute DNA purification columns with freshly prepared binding buffer (5 M Guanidine hydrochloride, 90 mM Sodium acetate, 40% Isopropanol and 0.05%

Tween-20). Elution was carried out with a standard TE buffer (1 mM EDTA, 10 mM TRIS-HCl). The advantage of the teeth extraction method lies in the fact, that the tooth samples could be retrieved after the DNA extraction process and safely deposited back into their original position in the sampled skulls.

NGS library construction

We prepared double stranded DNA libraries according to the protocol described in Meyer and Kircher[74] with minor modifications. We applied partial UDG treatment to counteract the effects of extensive postmortem damage (PMD). The reaction mix was prepared as described in Rohland et al.[75] containing 1X Tango Buffer (Thermo Scientific™), 100 μM dNTPs, 1 mM ATP and 0.03 U/μl USER enzyme with 30 μl sample DNA. We incubated the samples for 30 minutes than stopped the reaction with Uracil Glycosylase Inhibitor (UGI). The samples were prepared for adapter ligation with blunt-end repair using a mixture of T4 polynucleotide kinase (0.5 U/μl) and T4 DNA polymerase (0.1 U/μl) and incubated for 20 minutes on 25 and 15 °C. Following this, the samples were purified on Qaigen™ MinElute columns and eluted in 20 μl Elution Buffer (EB). Adapter ligation was carried out according to.[74] We used universal P5 and P7 adapter molecules in a mixture of 1X T4 DNA ligase buffer (Thermo Scientific™), 5% PEG-4000, 1.25 μM adapter mix and 0.125 U/μl T4 DNA ligase with 20 μl purified sample DNA. We incubated the samples for 30 minutes on 22 °C followed by another round of DNA purification on MinElute columns. Finally, we carried out an adapter fill-in reaction with 1X ThermoPol® reaction buffer (NEB®), 250 μM dNTPs and 0.3 U/μl Bst polymerase large fragment with 20 μl sample DNA to fill out the partially single stranded adapters and correct any nucleotide errors remaining on one of the strands. We omitted preamplification and directly double indexed our libraries in a single PCR step with Accuprime™ Pfx Supermix (Invitrogen™), containing 10 mg/ml BSA and 200 nM indexing P5 and P7 primers, in the following cycles: 95 °C 5 minutes, 12 times 95 °C 15 sec, 60 °C 30 sec and 68 °C 3 sec, followed by 5-minute extension at 68 °C. The indexed libraries were purified on MinElute columns and eluted in 20 μl EB.

DNA sequencing

Quantity measurements of the DNA extracts and libraries were performed with the Qubit fluorometric quantification system. The library fragment distribution was checked on TapeStation 2200 (Agilent). During the library preparation step, we used partial-UDG treatment to counteract extensive post-mortem damage. The double stranded libraries were than shallow sequenced on Illumina iSeq platform to monitor their human DNA content. Selected libraries were deep sequenced on Illumina NovaSeq platform to an average genome coverage of 1.42-fold (0.24x-3.75x, Table S1A).

Radiocarbon dating

Radiocarbon analysis was mainly performed on skeletal bone fragments of the sampled individuals to confirm the archaeological dating of the remains. When it was viable we instead used part of the remaining petrous powder to conduct the analysis, thus minimising the destruction of the human remains (Tables S1A and S1C). The measurements were done by accelerator mass spectrometry (AMS) in the AMS laboratory of the Institute for Nuclear Research, Hungarian Academy of Sciences, Debrecen, Hungary (AMS Lab ID: DeA-37107; technical details concerning the sample preparation and measurement in Molnár et al.[76,77]).

Bone samples were ultrasonicated in distilled water, dried, surface cleaned and grinded. Samples were sieved to get the appropriate sized sample fraction (0.5–1 mm) out of which 500–1000 mg was measured, depending on the preservation state of the bone. A continuous-flow bone sample preparation equipment - similar to the method used at the Oxford Radiocarbon Accelerator Unit (ORAU) - has been developed in the lab. In this unit, Omnifit® columns are used as flow cells to automate the ABA cleaning system. From 3 types of reagents, each one is injected via a 4-way valve and inert plastic tubing to an Ismatech® IPC 12 channel peristaltic pump to ensure a constant flow rate. Reagents are selectively pumped to the reaction cells containing small-grained bone samples, with a sequence of 0.5 M HCl and 0.1 M NaOH solution, interspersed with flushing with distilled water. At the end of the process, the reagents together with the contaminants are collected using a collection bottle for each cell. During the 16-hr-long process, reagents follow a well-defined sequence that is controlled by a computer program and a special electronic driver device.

The cleaned samples were inserted into a test tube containing 5 ml, pH 3 aqueous solution, and were placed into a heating block at 75 °C for 24 hr. Dissolved collagen/gelatin was filtered via a 0.45-μm glass fiber filter (Whatman® AUTOVIAL 5) into a clean vial, and after freezing, it was freeze-dried, which takes about 1–2 days.

For radiocarbon dating by AMS, graphite targets from the purified CO2 samples were prepared using a customized sealed tube graphitization method. The overall measurement uncertainty for modern samples is < 3.0‰, including normalization, background subtraction, and counting statistics.

The conventional radiocarbon date was calibrated with the OxCal 4.4.4 software (https://c14.arch.ox.ac.uk/oxcal/OxCal.html, date of calibration: 20.02.2024) with IntCal 20 settings.[78]

Map

Maps were created in R 4.1.0[52] with the help of ''ggplot2'', ''sf'', ''ggspatial'', ''rnaturalearth'' and ''elevatr'' packages.[53–57] First, a data table was called containing the spatial coordinates for the selected geographical area using ''rnaturalearth''. Waterways were added as a separate layer using the spatial data published in Yan et al.[79] Finally, we obtained elevation data with ''elevatr'' than superimposed it on the spatial coordinates as a ''geom_tile'' using ''ggplot2''.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing and quality control

The adapters of paired-end reads were trimmed with the Cutadapt software,[58] and sequences shorter than 25 nucleotides were removed. Read quality was assessed with FastQC.[80] The raw reads were aligned to GRCh37 (hs37d5) reference genome using the Burrows-Wheeler-Aligner (v 0.7.17) software, with the MEM command in paired mode, with default parameters and disabled re-seeding.[59] Only properly paired primary alignments with $\geq$ 90% identity to reference were considered in all downstream analyses to remove high mapping quality exogenous DNA containing non-aligned overhangs as detailed in Maróti et al.[34] Samtools v1.1 was used for merging the sequences from different lanes and also for sorting, and indexing binary alignment map (BAM) files.[60] PCR duplicates were marked using Picard Tools MarkDuplicates v 2.21.3.[61] To randomly exclude overlapping portions of paired-end reads and to mitigate potential random pseudo haploidization bias, we applied the mergeReads task with the options ''updateQuality mergingMethod=keepRandomRead'' from the ATLAS package.[62] Single nucleotide polymorphisms (SNPs) were called using the ANGSD software package (version: 0.931–10-g09a0fc5)[63] with the ''-doHaploCall 1 -doCounts 1'' options and restricting the genotyping with the ''-sites'' option to the genomic positions of the 1240K panel.

Ancient DNA damage patterns were assessed using MapDamage 2.0[64] (Table S1F). Mitochondrial genome contamination was estimated using the Schmutzi algorithm.[65] Contamination for the male samples was also assessed by the ANGSD X chromosome contamination estimation method,[81] with the ''-r X:5000000-154900000 -doCounts 1 -iCounts 1 -minMapQ 30 -minQ 20 -setMinDepth 2'' options (Table S1A).

The raw nucleotide sequence data of the samples were deposited to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number European Nucleotide Archive: PRJEB80732.

Sex determination

Biological sex was determined with the method described in Skoglund et al.[82] Fragment length of paired-end data and average genome coverages (all, X, Y, mitochondrial) were assessed by the ATLAS software package[62] using the BAMDiagnostics task. Detailed coverage distribution of autosomal, X, Y, mitochondrial chromosomes was calculated by the mosdepth software[66] (Table S1A).

Haplogroup assignment and uniparental analysis

Mitochondrial haplogroups (Mt Hg) were determined using the HaploGrep 2 (version 2.1.25) software,[67] using the consensus endogen fasta files resulting from the Schmutzi Bayesian algorithm. The Y Hg assessment was performed with the Yleaf software tool,[68] updated with the ISOGG2020 Y tree dataset (Table S1A). In one case (MDH-405) the Y Hg could not be determined due to low coverage, this was marked as ''inconclusive'' in Table S1A.

To shed light on the most feasible origin of the uniparental lineages of our samples, we assembled a comprehensive uniparental database of the Carpathian Basin. We have chosen samples from the Allen Ancien DNA Resource (AADR)[83] database based on their country of origin to cover this region. This included samples from Hungary, Slovakia, Romania, Serbia, Croatia, Slovenia and Austria. The haplogroups were collected from the original publications and cross-referenced with the publicly available databases published in[51,84] as well as our own classifications in the cases where the genomes were already downloaded for other analyses.

PCA

To uncover the underlying structure of our studied individuals in a hypothesis independent manner, we conducted PCA analysis. We used the same modern Eurasian genome dataset as our previous publication[34] confined to the HO SNP set, to draw a modern PCA background on which ancient samples could be projected. This consisted of a generalized set of 1397 modern individuals from 179 modern Eurasian populations (Table S2A). PCA Eigen vectors were calculated from these pseudo-haploidized modern genomes with smartpca (EIGENSOFT version 7.2.1).[37]

All ancient genomes were projected on the modern background with the ''lsqproject: YES and inbreed: YES'' options. Since the ancient samples were projected, we used a more relaxed genotying threshold (>50k genotyped markers) to exclude samples only where the results could be questionable due to the low coverage.

Unsupervised ADMIXTURE

We used ADMIXTURE analysis to model our genomes as compositions of hypothetical ancestral populations.[36]

We performed unsupervised ADMIXTURE analysis on 2578 ancient individuals using the 1240K SNP set.[83] This included **149** newly sequenced individuals and **2429** ancient individuals from the AADR[83] (Table S3A). As we did not include modern individuals from the HO dataset we obtained a significantly higher number (391,178) of overlapping SNP sights. We excluded individuals with less than 200K SNPs covered and samples with >4% contamination, we also taken out close relatives to avoid the appearance of undesired ancestral components. Cross-validation error calculation showed that modeling K-6 ancestral components yielded the most consistent results (Table S3B).

qpAdm analysis and F-statistics

F4-statistics allow us to infer allele frequency covariance among four populations, providing a hypothesis test for specific population tree structures. In our analysis, we employed two types of rooted trees based on the D-statistics approach of.[85] First, we tested for simple directional gene flow or genetic affinity using the formula F4(Outgroup, Test; Ref1, Ref2). Here, the Test represents the studied sample, while Ref1 and Ref2 are reference populations. A negative F4 value indicates a stronger affinity between the Test and Ref1, while a positive value suggests a closer relationship to Ref2.

Second, we leveraged the exclusive nature of F4-statistics to our advantage. The test F4(Outgroup, Ref1; Ref2, Test) only produces a meaningful value at loci where both the Outgroup-Ref1 and Ref2-Test pairs exhibit different alleles. This approach is particularly useful because F4-statistics measure net genetic similarities, making them less effective in detecting complex, overlapping affinities. If the Test has a strong genetic relationship with one reference population, it may obscure weaker affinities with the other. By subtracting the major affinity, we can reveal minor affinities that would otherwise remain undetectable using the more conventional F4(Outgroup, Test; Ref1, Ref2) approach.

We tested multiple F4-statistic frameworks. The statistics were calculated using the 1240K SNP set with the qpF4ratio algorithm from the ADMIXTOOLS software package.[86] The results of the F4 analyses were visualized in a 2D framework.

We used qpAdm[69] from ADMIXTOOLS[86] for modelling our genomes as admixtures of two or three source populations and estimating ancestry proportions. The qpAdm analysis was done with the HO dataset, as in many cases suitable RIGHT or LEFT populations were only available in this dataset.

Our main goal of the study was to investigate the relationship between the Sarmatian individuals found in the Carpathian Basin and Sarmatians of the Central Steppe. Thus, in our qpAdm analysis framework we wanted to evaluate whether the available steppe Sarmatian individuals are a necessary source population for modelling the studied individuals. We considered three types of possible modelling sources (LEFT populations); a) a population set representing the supposed local inhabitants of the region, b) a population set representing the proposed Sarmatian ancestry, and c) a population set representing other possible central and East Asian sources as our ADMIXTURE analyses indicated at least a marginal appearance of these.

The set of 15 local sources (highlighted with grey in Tables S5B–S5E) were selected from an extensive preliminary qpAdm run, among 162 possible candidates from the AADR.[83] The selection was based on population size, number of markers, PCA and ADMIXTURE clustering and goodness of fit in the acquired models. We assembled this optimal source population subset for the explicit purpose to model the highest number of our test subjects in a single qpAdm run and not to acquire their true ancestral composition. To represent the arriving Sarmatian population we assembled two genetically homogeneous source population from the available Steppe Sarmatians published from Russia and Kazakhstan[83] (see Data S1). The remaining sources represent other possible central or eastern Asian immigrants. We also used populations from our previously published article,[34] which had been shown to have extensive connections to the Carpathian Basin.

The reference population set (RIGHT populations) contained Ethiopia_4500BP (fixed), Iran_GanjDareh_N, Turkey_N, Latvia_HG, Baikal_EN (Russia_Shamanka_Eneolithic.SG and Russia_Lokomotiv_Eneolithic.SG), WSHG (Russia_Tyumen_HG and Russia_Sosnoviy_HG), Russia_Steppe_Maikop, Karitiana and Poland_Koszyce_GlobularAmphora.SG. For a detailed list of LEFT and RIGHT populations see Table S5F.

During the runs we set the details:YES parameter to evaluate Z-scores for the goodness of the fit of the model (estimated with a Block Jackknife). As qpWave is integrated in qpAdm, the nested p values in the log files indicate the optimal rank of the model. This means that if p value for the nested model is above 0.05, the Rank-1 model should be considered.[69]

As our extensive source population set (LEFT populations) portended a great number of alternate models we applied the model competition framework explicitly discussed in Narasimhan et al.[31] and Maróti et al.[34] In this setup we test each resulting qpAdm model with a feasible *p*-value, by iteratively rerunning it with moving each of the LEFT populations in the RIGHT population set. This results in a bilateral improvement. On one hand the true source population – when included in the RIGHT population set – should consistently exclude any suboptimal models, as the test will (by design) have their highest shared drift with their true sources. On the other hand we will have a distribution of p-values for each resulting model which enables us to better quantify their goodness of fit, as demonstrated in.[69] As we run each model multiple times, we can obtain further useful information concerning the feasibility of the individual models. Thus, based on the output of the qpAdm algorithm we included further quality measurements into our analysis framework. The meaning of the columns in Tables S5B–S5E is as follows. **Test**: the name of the individual/population modelled. **SourceX**: the name of the designated sources in descending order of contribution. **SourceX ratio**: the obtained mixture coefficient for the source population/individual in question. **Valid models**: the number of cases a given model passed the quality criteria. **Excluded models**: the number of cases a model has been excluded by one of the reference populations. **BadFit models**: quality criterion, number of cases the obtained mixture coefficient differs from the jackknife calculated mixture coefficient by at least 10%. **Negative models**: quality criterion, number of cases where one of the source coefficients is lower than 0. **Non-significant nested p-value models**: the number of models where the obtained nested p-value is higher than 0.05. This indicates that the k-1 model is more plausible. **Average nested p-value**: the average of the obtained p-values for the k-1 model. **Minimum p-value**: the lowest p-value obtained from the model competitions. **Maximum p-value**: the highest p-value obtained from the model competitions. **Average p-value**: the average of the obtained p-values of the valid model repetitions (passing the quality criteria). **Average p-value summary**: either the average p-value or the average nested p-value if the number of non-significant nested p-value models is higher than the half of the valid models. This was used to order the qpAdm results. **Minimum p-value reference**: the name of the

LEFT population/individual which was moved to the RIGHT population set when the lowest p-value was obtained for the model in question. **Maximum p-value reference**: the name of the LEFT population/individual which was moved to the RIGHT population set when the highest p-value was obtained for the model in question. **Excluding reference**: the name of the LEFT population/individual which was moved to the RIGHT population set when the model was excluded. **Used on qpAdm plot**: an "x" signifies the specific model that was used to compile Figure 3C.

We ran comprehensive 2-way modelling runs for all studied TEST individuals with the above-described RIGHT population set and freely combined LEFT population set (Table S5B). Subsequent 3-way modelling was only conducted on a selected subset of the TEST individuals with unsatisfactory 2-way models (Table S5C). We selected these individuals based on preliminary qpAdm analyses where a sufficient increase in their p-value was reasonably expected. After 3-way modeling, 12 individuals still remained with no feasible models (p-value <0.05 or all models excluded by the model-competition). As the ADMXITURE and PCA profile of these individuals showed a very similar composition as outlier individuals in our previous article,[34] we included some further sources representing possible northern and southern European populations of the time as some of the outliers seemed clearly deriving a portion of their ancestry from these regions. We swapped our Steppe Sarmatians sources for some of our already modelled individuals and individual Russian and Kazakh Sarmatians, as they may contain some minor components that were not sufficiently represented in the grouped Sarmatian reference populations (Table S5D). Finally, since the two Iron Age individuals could be adequately modeled using Steppe Sarmatian sources and could be considered potential sources due to their age and geographic proximity, we included them in a simple two-way modeling run. For this analysis, we focused exclusively on individuals with substantial (>40%) Steppe Sarmatian ancestry, as determined by previous analyses (Table S5E). In the end we successfully modelled all of our studied individuals except for a single individual (HVF-10) with high average p-values (Table S5A). The single unmodeled individual seems to be an outlier which has no sufficient source in the database yet.

Imputation

We imputed our studied genomes together with other shotgun sequenced ancient genomes from a similar spatio-temporal distribution (Table S6) with the GLIMPSE2 framework (version 2.0.0)[39] according to the recommendations of.[70] Approximately 78 million biallelic common markers from the 1KG dataset were imputed with GLIMPSE2, utilizing the 1KG phase III data as a reference. The reference dataset was normalized, and multi-allelic sites were split using bcftools (version 1.16-63-gc021478 with htslib 1.16-24-ge88e343), applying the "norm –m –any" subcommand. Biallelic SNPs were filtered using the "view –m 2 –M 2 –v snps" subcommand. The autosomal chromosomes of the human reference genome were divided into 580 genomic chunks using the GLIMPSE2_chunk tool with the "-sequential" option. Following the GLIMPSE2 guidelines, we generated the binary reference data using the GLIMPSE2_split_reference tool, based on the 580 genomic regions and 1KG biallelic SNP variants. For the imputation process, we included only samples with shotgun WGS data exceeding 0.25x mean genome coverage as recommended by Ringbauer et al.[40] and a contamination level below 0.04, a stricter criterion than the 0.05 threshold used in the AADR database.[83] Due to these criteria, we excluded 17 Sarmatian samples published in Gnecchi-Ruscone et al.,[33] as they were obtained through capture enrichment sequencing, along with 7 of our newly sequenced genomes that exceeded the contamination threshold.

Kinship estimation and IBD sharing analysis

Kinship analysis was performed with correctKin[35] (Table S1D). As reference population we applied the same database as in Varga et al.[72]

The shared IBD segments were identified using the ancIBD framework (version 0.5)[40] according to the recommended workflow outlined in the official ancIBD documentation (https://ancibd.readthedocs.io/en/latest/run_ancIBD.html). Phased and imputed variants were post-filtered to include only the positions of the 1240K AADR marker set and lifted to the hdf5 data format using the 'vcf_to_1240K_hdf' method with the default parameters. IBD fragments were identified using the 'hapBLOCK_chroms' method applying the standard five-state HMM with the haploid_gl2 emission model, which is appropriate for the GLIMPSE2 posterior likelihoods, to ascertain the raw IBD segments >=8cM.

During the subsequent filtration of raw IBD segments, we deviated from the marker density threshold ($\geq$ 220 SNPs/cM) used in the original ancIBD framework.[40] We found that applying this global threshold led to a significant number of false positive and false negative IBD segment identification as detailed in Data S1, under the title: "Novel approach for IBD segment filtration of raw ancIBD output". Instead, we implemented a novel method that uses marker informativity scores to dynamically mask and exclude genomic regions lacking sufficient power to detect true IBD segments. A detailed description of our algorithm, along with a comparison of the two methods, can be found in Data S1. We developed this filtration method in Python for use with the raw IBD output generated by ancIBD. Our tool integrates seamlessly into the ancIBD framework and is available on GitHub [www.github.com/zmaroti/scoreFilterIBD].

Shared IBD network was generated in R 4.1.0,[52] with the application of packages ggplot2 3.4.2,[53] igraph[71] and qgraph.[87]

Plotting

To generate Figure 4A, we partitioned our dataset into wider spatio-temporal groups based on the archaeological and geographical data (Regional Group in Table S6B). We first contracted our whole IBD network into single points (vertices) representing each group, then calculated the distribution of these points using the Fruchterman-Reingold weight directed algorithm implemented in qgraph,[87]

where weights were the number of connections (edges) between each group. We centralized and expanded this distribution to create sufficient space, then we calculated graph distributions for each groups separately with the same weight directed algorithm, but now the weights were given as the sum total length of shared IBDs between each individual. We added the coordinates obtained from the single vertex calculation to each corresponding groups' coordinates to arrange the separate group distributions according to the single vertex distribution. Finally we ran a new weight directed algorithm, with the recalculated coordinates as initial coordinates and with *niter = 1* and *max.delta = 0* parameters to not allow points deviating from their initial coordinates. We only plotted edges corresponding to intergroup connections to make the plot more straightforward. Intragroup connections are represented in this cause by the distance of the points from each other.

For Figure 4B we allowed the algorithm to run for 100 iterations and we used *max.delta = (0.1 + % between-group connection)* which allows for points to move along their edges but proportional to the number of intergroup connections. This approach emphasizes the net direction of intergroup attractions.

For Figure 5 we defined a core set of populations that we were primarily interested in. This included individuals with the Regional Group labels: Western_Steppe_Sarmatian, Carpathian_Basin_Sarmatian_Period and Carpathian_Basin_Hun_Period, which included the individuals published in this article, and Central_Steppe_Sarmatian, Carpathian_Basin_Avar_Period, Carpathian_Basin_Conquer_Period as main candidates for references (Table S6B). We included a further 17 individuals from other Regional Groups that had at least five connections to any of the core populations as to not inflate the plot with non-informative data. This partition was made to find an equilibrium between a strict constrain that also preserves only the most informative individuals. After defining the desired individual set, we ran a simple Fruchterman-Reingold weight directed algorithm with no constrains for 1000 iterations, where weights were given as the sum total length of IBD fragments shared between individuals.

An important metric for graph-based calculations is the number of connections an individual has (degree or degree centrality, d). When comparing degree centrality among groups or individuals it's important to consider that a connection always exists between two endpoints, thus we carefully avoided counting individual connections multiple times. Another pitfall to consider is comparing groups of different sizes, where the sample size disproportionately affects the chance of finding IBD connections (linearly increasing sample sizes quadratically increase the possibility of uncovering connections) thus, to properly compare groups we must normalize the obtained degree centrality values by dividing them with the number of possibly available connection thus producing the ratio of fulfilled connections. For Figures 6 and S2 the number of connections (degree centrality, d) was normalized by the product of the sizes of the two groups in case of intergroup connections ($d_i' = d_i / [n_i \times n_j]$), while intragroup connections were normalized by the equation: $d_i' = d_i / ([n_i \times \{n_i\text{-}1\}]/2)$, where n symbolizes the group size. To produce Figure S3, we normalized the degree centrality of the cemetery groups by dividing it with the product of the size of the cemetery groups (n_i) and the number of all remaining individuals ($n\text{-}n_i$): $d_i' = d_i / (n_i \times [n\text{-}n_i])$.

CellPress
OPEN ACCESS

Supplemental figures

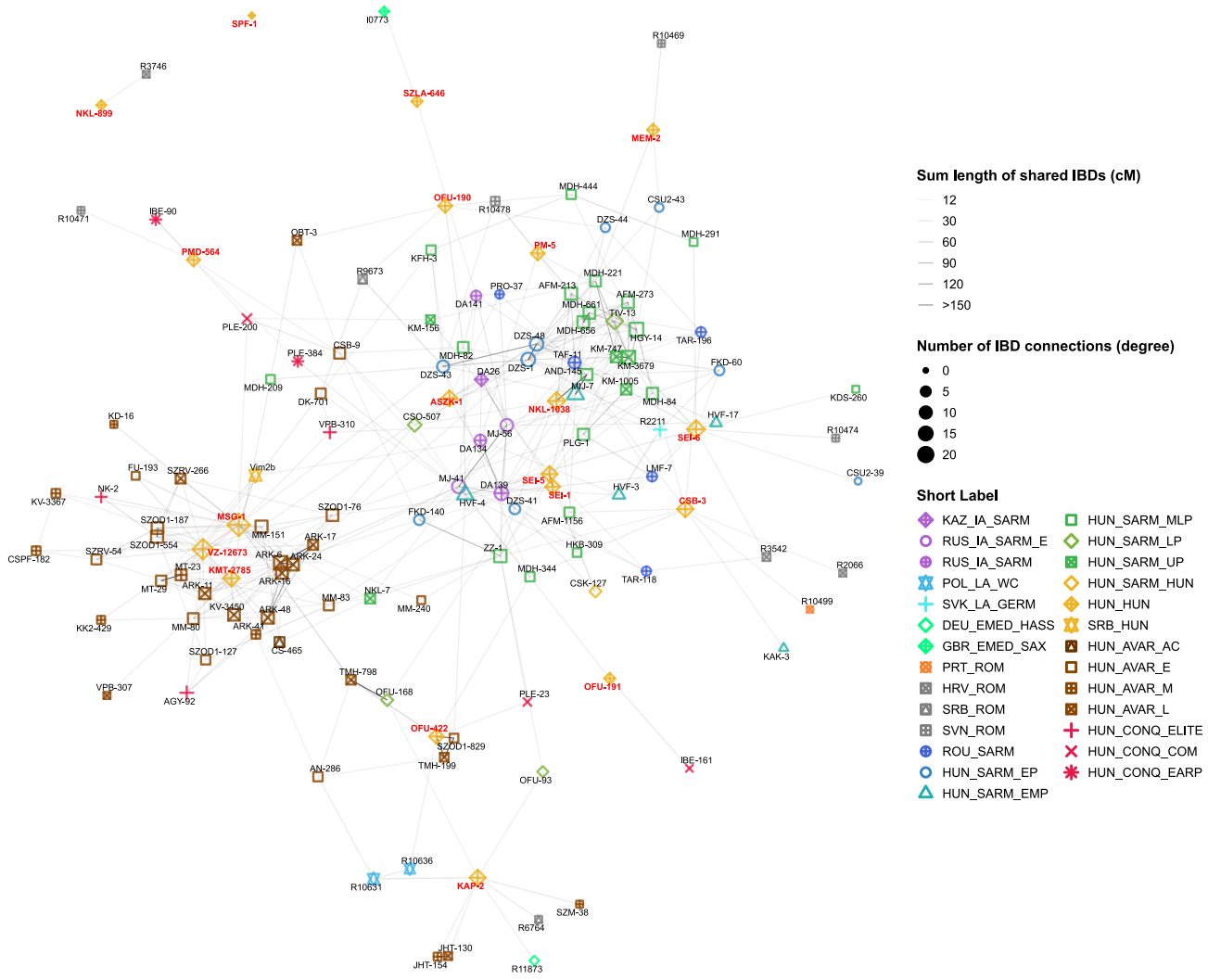


Figure S1. IBD sharing graph of the HUN_HUN individuals, related to Figures 4 and 5 and the kinship estimation and IBD sharing analysis section of the STAR Methods

All individuals from the HUN_HUN group are plotted alongside others with a minimum of 12 cM sum IBD shared with any member of the group. The samples from the HUN_HUN group are highlighted in yellow, with red labels indicating their sample names. This plot was prepared in the same manner as the separate cemetery plots (see Data S1).

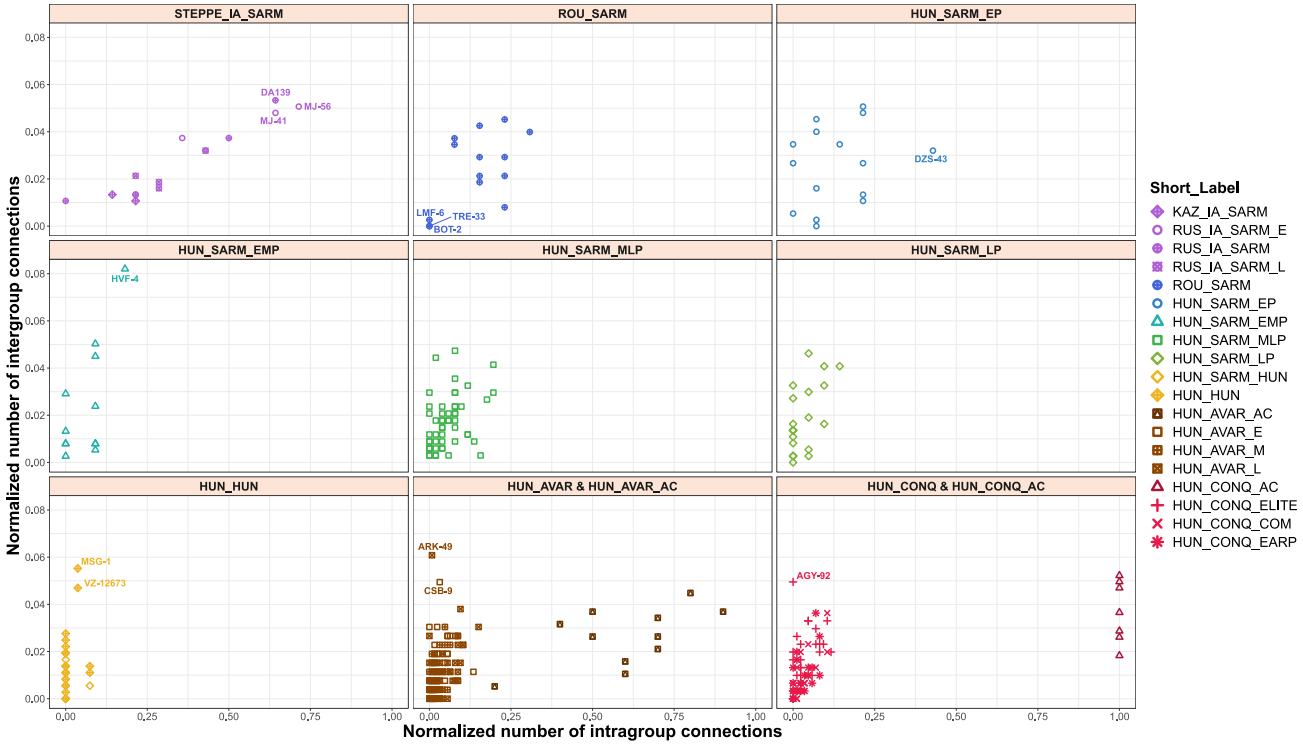


Figure S2. Normalized counts of intragroup and intergroup connections of individuals across various archaeological periods in the CB and its vicinity, related to Figure 6 and the kinship estimation and IBD sharing analysis section of the STAR Methods

Group names are written in the strip text above each subplot. For simplicity, the HUN_AVAR-HUN_AVAR_AC and HUN_CONQ-HUN_CONQ_AC groups are plotted together, although calculations were performed using the original groupings (see column label in Table S6B). The *y* axis represents the normalized number of intergroup connections (number of detected outgroup connections/[total number of all individuals − group size]). The *x* axis represents the normalized number of intragroup connections (number of detected intragroup connections/[group size − 1]). Each connection was considered equal, independent of IBD number and size.

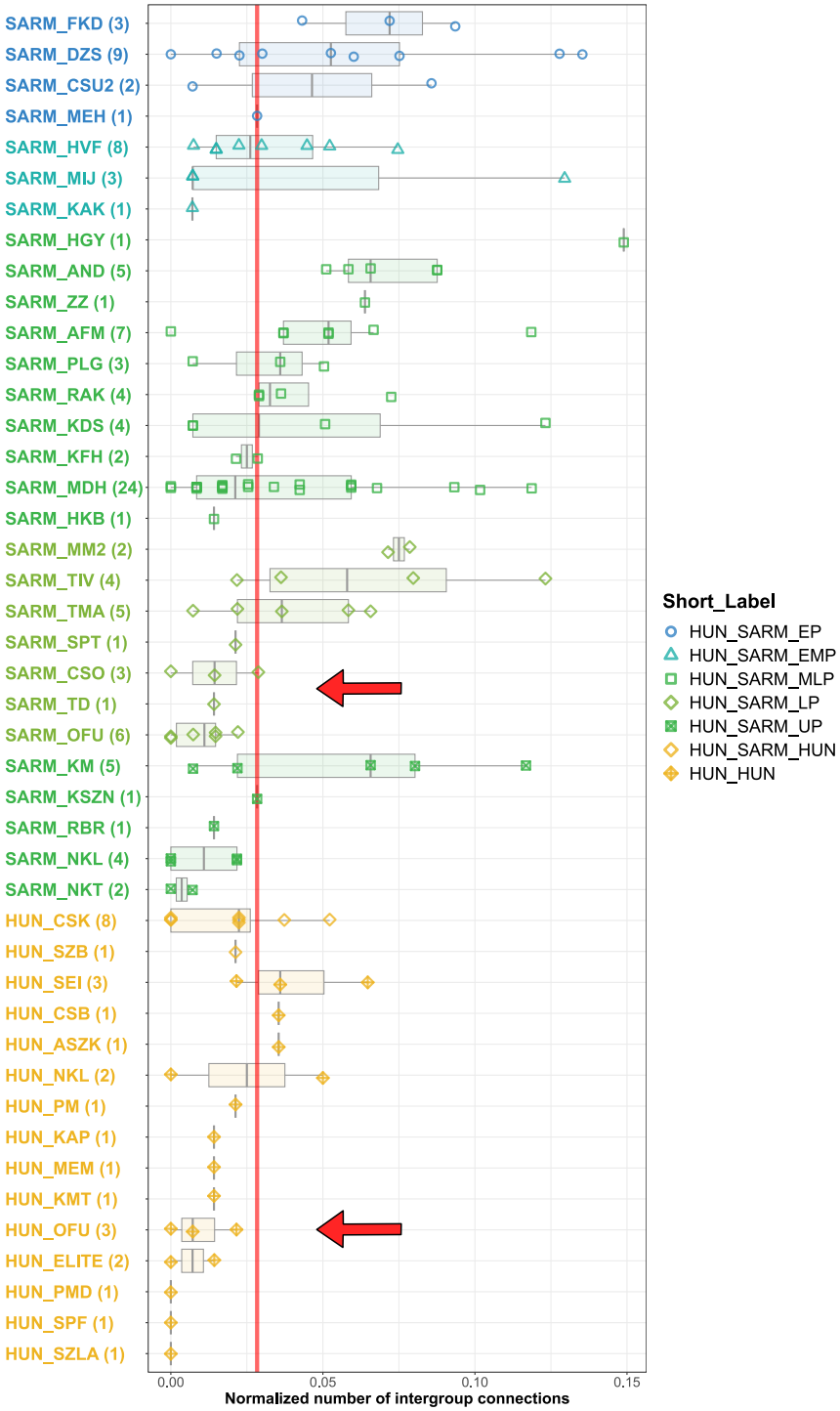


Figure S3. Intergroup sharing among cemeteries of the Sarmatian and Hun periods of the CB, related to Figures 5 and 6 and the kinship estimation and IBD sharing analysis section of the STAR Methods

Boxplots display the normalized number of intergroup IBD sharing among Sarmatian- and Hun-period cemeteries. Abbreviations for the cemeteries are listed in Table S6B. The red line indicates the median of the intergroup IBD connections across the plotted individuals.

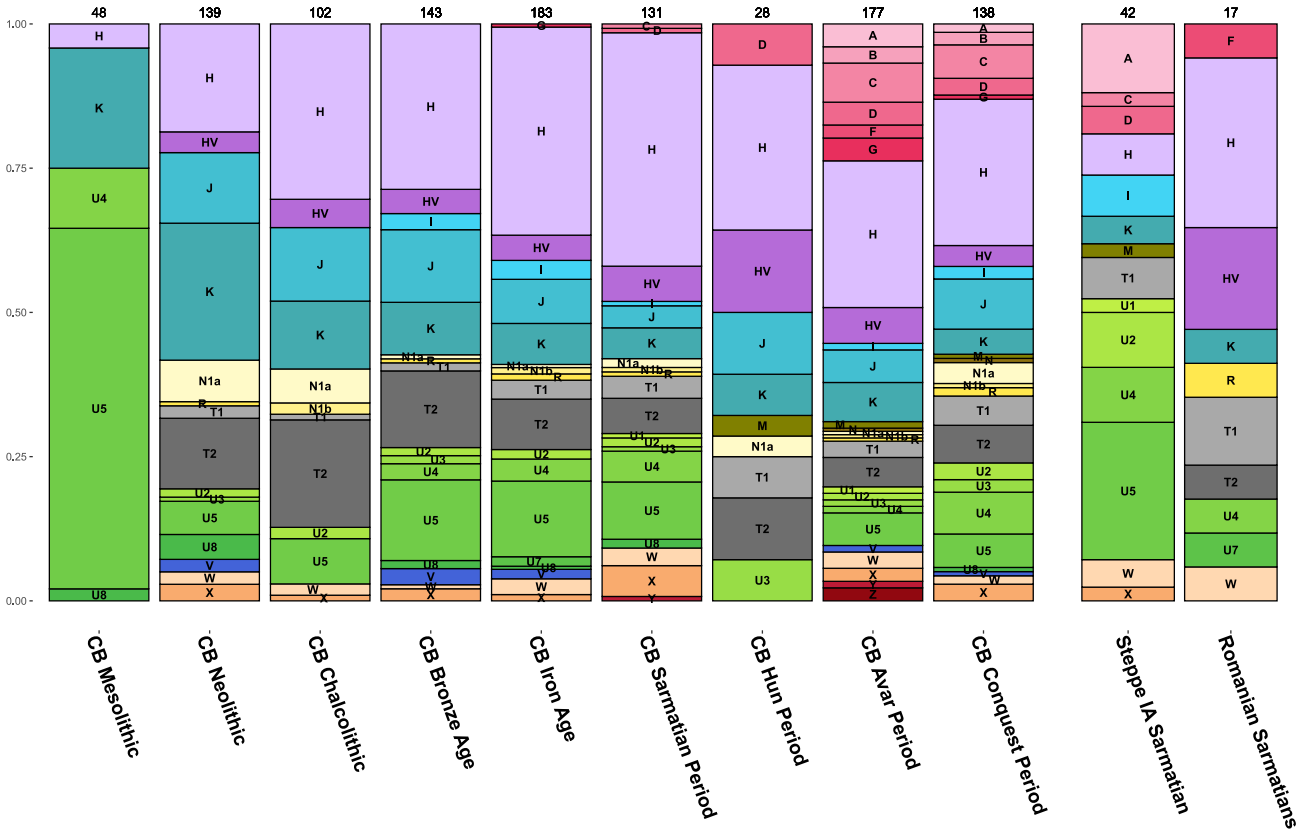


Figure S4. Mitochondrial haplogroup distribution across successive periods in CB populations, related to Figure 7 and the haplogroup assignment and uniparental analysis section of the STAR Methods

Article

Current Biology

The genetic origin of Huns, Avars, and conquering Hungarians

Highlights

- 265 new ancient genomes help to unravel the origin of migration-period populations
- Genetic continuity is detected between Xiongnus and European Huns
- European Avars most likely originated from Mongolia and were related to Huns
- Conquering Hungarians had Ugric ancestry and later admixed with Sarmatians and Huns

Authors

Zoltán Maróti, Endre Neparáczki, Oszkár Schütz, ..., Szilárd Sándor Gál, Péter Tomka, Tibor Török

Correspondence

torokt@bio.u-szeged.hu

In brief

Maróti et al. show that immigrants were in minority compared with the locals of the Carpathian Basin in each period. Several Hun period immigrants had Asian Hun (Xiongnu) ancestry. The Avar immigrant elite had ancient Mongolian origin. Conquering Hungarians and Mansis had common ancestors, but proto-Hungarians further admixed with Sarmatians and Huns.

Maróti et al., 2022, Current Biology *32*, 1–13
July 11, 2022
https://doi.org/10.1016/j.cub.2022.04.093

CellPress

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

Current Biology

CellPress
OPEN ACCESS

Article

The genetic origin of Huns, Avars, and conquering Hungarians

Zoltán Maróti,[1,2] Endre Neparáczki,[1,3] Oszkár Schütz,[3] Kitti Maár,[3] Gergely I.B. Varga,[1] Bence Kovács,[1,3] Tibor Kalmár,[2] Emil Nyerki,[1,2] István Nagy,[4,5] Dóra Latinovics,[4] Balázs Tihanyi,[1,6] Antónia Marcsik,[6] György Pálfi,[6] Zsolt Bernert,[7] Zsolt Gallina,[8,9] Ciprián Horváth,[9] Sándor Varga,[10] László Költő,[11] István Raskó,[12] Péter L. Nagy,[13] Csilla Balogh,[14] Albert Zink,[15] Frank Maixner,[15] Anders Götherström,[16] Robert George,[16] Csaba Szalontai,[17] Gergely Szenthe,[17] Erwin Gáll,[18] Attila P. Kiss,[19] Bence Gulyás,[20] Bernadett Ny. Kovacsóczy,[21] Szilárd Sándor Gál,[22] Péter Tomka,[23] and Tibor Török[1,3,24,*]

[1]Department of Archaeogenetics, Institute of Hungarian Research, 1041 Budapest, Hungary
[2]Department of Pediatrics and Pediatric Health Center, University of Szeged, 6725 Szeged, Hungary
[3]Department of Genetics, University of Szeged, 6726 Szeged, Hungary
[4]SeqOmics Biotechnology Ltd., 6782 Mórahalom, Hungary
[5]Institute of Biochemistry, Biological Research Centre, 6726 Szeged, Hungary
[6]Department of Biological Anthropology, University of Szeged, 6726 Szeged, Hungary
[7]Department of Anthropology, Hungarian Natural History Museum, 1083 Budapest, Hungary
[8]Ásatárs Ltd., 6000 Kecskemét, Hungary
[9]Department of Archaeology, Institute of Hungarian Research, 1041 Budapest, Hungary
[10]Móra Ferenc Museum, 6720 Szeged, Hungary
[11]Rippl-Rónai Municipal Museum with Country Scope, 7400 Kaposvár, Hungary
[12]Institute of Genetics, Biological Research Centre, 6726 Szeged, Hungary
[13]Praxis Genomics LLC, Atlanta, GA 30328, USA
[14]Department of Art History, Istanbul Medeniyet University, 34720 Istanbul, Turkey
[15]Institute for Mummy Studies, EURAC Research, 39100 Bolzano, Italy
[16]Department of Archaeology and Classical Studies, Stockholm University, 11418 Stockholm, Sweden
[17]Hungarian National Museum, Department of Archaeology, 1088 Budapest, Hungary
[18]"Vasile Pârvan" Institute of Archaeology, 010667 Bucharest, Romania
[19]Faculty of Humanities and Social Sciences, Institute of Archaeology, Pázmány Péter Catholic University, 1088 Budapest, Hungary
[20]Institute of Archaeological Sciences, Eötvös Loránd University, 1088 Budapest, Hungary
[21]Katona József Museum, 6000 Kecskemét, Hungary
[22]Mureş County Museum, 540088 Târgu Mureş, Romania
[23]Department of Archaeology, Rómer Flóris Museum of Art and History, 9021 Győr, Hungary
[24]Lead contact
*Correspondence: torokt@bio.u-szeged.hu
https://doi.org/10.1016/j.cub.2022.04.093

SUMMARY

Huns, Avars, and conquering Hungarians were migration-period nomadic tribal confederations that arrived in three successive waves in the Carpathian Basin between the 5th and 9th centuries. Based on the historical data, each of these groups are thought to have arrived from Asia, although their exact origin and relation to other ancient and modern populations have been debated. Recently, hundreds of ancient genomes were analyzed from Central Asia, Mongolia, and China, from which we aimed to identify putative source populations for the above-mentioned groups. In this study, we have sequenced 9 Hun, 143 Avar, and 113 Hungarian conquest period samples and identified three core populations, representing immigrants from each period with no recent European ancestry. Our results reveal that this "immigrant core" of both Huns and Avars likely originated in present day Mongolia, and their origin can be traced back to Xiongnus (Asian Huns), as suggested by several historians. On the other hand, the "immigrant core" of the conquering Hungarians derived from an earlier admixture of Mansis, early Sarmatians, and descendants of late Xiongnus. We have also shown that a common "proto-Ugric" gene pool appeared in the Bronze Age from the admixture of Mezhovskaya and Nganasan people, supporting genetic and linguistic data. In addition, we detected shared Hun-related ancestry in numerous Avar and Hungarian conquest period genetic outliers, indicating a genetic link between these successive nomadic groups. Aside from the immigrant core groups, we identified that the majority of the individuals from each period were local residents harboring "native European" ancestry.

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

INTRODUCTION

Successive waves of population migrations associated with the Huns, Avars, and Hungarians or Magyars from Asia to Europe had an enduring impact on the population of the Carpathian Basin.[1] This is most conspicuous in the unique language and ethnocultural traditions of the Hungarians, the closest parallels of which are found in populations east of the Urals. According to present scientific consensus, these eastern links are solely attributed to the last migrating wave of conquering Hungarians (henceforth shortened as Conquerors), who arrived in the Carpathian Basin at the end of the 9th century CE. On the other hand, medieval Hungarian chronicles, foreign written sources, and Hungarian folk traditions maintain that the origin of Hungarians can be traced back to the European Huns, with subsequent waves of Avars and Conquerors considered kinfolk of the Huns.[2,3]

Both Huns and Avars founded a multiethnic empire in Eastern Europe centered on the Carpathian Basin. The appearance of Huns in European written sources ca 370 CE was preceded by the disappearance of Xiongnus from Chinese sources.[4] Likewise, the appearance of Avars in Europe in the sixth century broadly correlates with the collapse of the Rouran Empire.[5] However, the possible relations between Xiongnus and Huns as well as Rourans and Avars remain largely controversial due to the scarcity of sources.[6]

From the 19th century onward, linguists reached a consensus that the Hungarian language is a member of the Uralic language family, belonging to the Ugric branch with its closest relatives the Mansi and Khanty languages.[7,8] On this linguistic basis, the Hungarian prehistory was rewritten, and the Conquerors were regarded as descendants of a hypothetical Proto-Ugric people. At the same time, the formerly accepted Hun-Hungarian relations were called into question by source criticism of the medieval chronicles.[9,10]

Due to the scarceness of bridging literary evidence and the complex archaeological record, an archaeogenetic approach is best suited to provide insights into the origin and biological relationship of ancient populations. To this end, we performed whole genome analysis of European Hun, Avar, and Conquest period individuals from the Carpathian Basin to shed light on the long-debated origin of these groups. The majority of our 271 ancient samples (Data S1A) were collected from the Great Hungarian Plain (Alföld), the westernmost extension of the Eurasian steppe, which provided a favorable environment for the arriving waves of nomadic groups. The overview of archaeological sites and time periods of the studied samples is shown in Figure 1, and a detailed archaeological description of the periods, cemeteries, and individual samples is given in Methods S1. From the studied samples, we report 73 direct accelerator mass spectrometry (AMS) radiocarbon dates, of which 50 are first reported in this paper (Data S2).

In this work, we identified immigrant and local groups from each period and revealed the most likely genetic origin of all individuals.

RESULTS

Genome-wide data were generated for 271 ancient individuals using shotgun sequencing. We obtained genome coverages ranging from 0.15- to 7.09-fold, with an average of 1.97-fold, and quality control with MapDamage 2.0,[11] Schmutzi,[12] and ANGSD[13] estimated negligible contamination in nearly all samples (Data S1A and S1B). All data were pseudohaploidized by randomly calling SNPs at all positions of the human origins (HOs, 600K SNPs) and the 1240K dataset. Most of the analysis was done with the HO dataset, as relevant modern genomes were available only in this format. We identified 43 kinship relations in our samples with the PCAngsd software[14] (Table S1), and just one of the relatives was included in the subsequent analysis. The new data were merged and coanalyzed with 2,364 ancient (Data S3) and 1,397 modern Eurasian genomes (Table S2). As the studied samples represent three archaeologically distinguishable periods from three consecutive historically documented major migration waves into the Carpathian Basin, we evaluated Hun, Avar, and Conquest period samples separately. In order to group the most similar genomes for population genetic analysis, we clustered our samples together with all published ancient Eurasian genomes, according to their pairwise genetic distances obtained from the first 50 principal component analysis (PCA) dimensions (PC50 clustering; STAR Methods; Data S3).

Most individuals in the study had local European ancestry

We performed PCA by projecting our ancient genomes onto the axes computed from modern Eurasian individuals (Figures 2A and S1). In Figure 2A, many samples from nearly each period project onto modern European populations; moreover, these samples form a South-North cline along the PC2 axis, which we termed the Eur-cline. PC50 clustering identified five genetic clusters within the Eur-cline (Figures 2 and S1B; Data S3), well sequestered along the PC2 axis. We selected representative samples from each cluster based on individual distal qpAdm and grouped them as Eur_Core1 to Eur_Core5, respectively. Eur_Core groups include samples from multiple periods, and they are not considered distinct populations, rather they represent distinct local genome types suitable for subsequent modeling. We also showed that each Eur-cline member can be modeled from the Eur_Core groups (Data S4; summarized in Data S1C).

Eur_Core1 clusters with Langobards from Hungary;[15] Iron Age, Imperial, and Medieval individuals from Italy;[16] and Minoans and Mycenaeans from Greece[17] (Data S3). Eur_Core2, 3, and 4 cluster among others with Langobards[15] and Bronze Age samples from Hungary,[18,19] the Czech Republic, and Germany,[19] whereas Eur_Core5 clusters with Hungarian Scythians.[20]

Unsupervised ADMIXTURE analysis revealed a gradient-like shift of genomic components along the Eur-cline (Figure 2B) with increasing "Ancient North Eurasian" (ANE) and "Western Hunter-Gatherer" (WHG) and decreasing "early Iranian farmer" (Iran_N) and "early European farmer" (EU_N) components from South to North. It is also apparent that Eur-cline samples contain negligible Asian ("Nganasan" and "Han") components. ADMIXTURE also confirms that similar genomes had been present in Europe and the Carpathian Basin before the Migration Period, as Eur_Core1 and 5 have comparable patterns with

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

CellPress
OPEN ACCESS

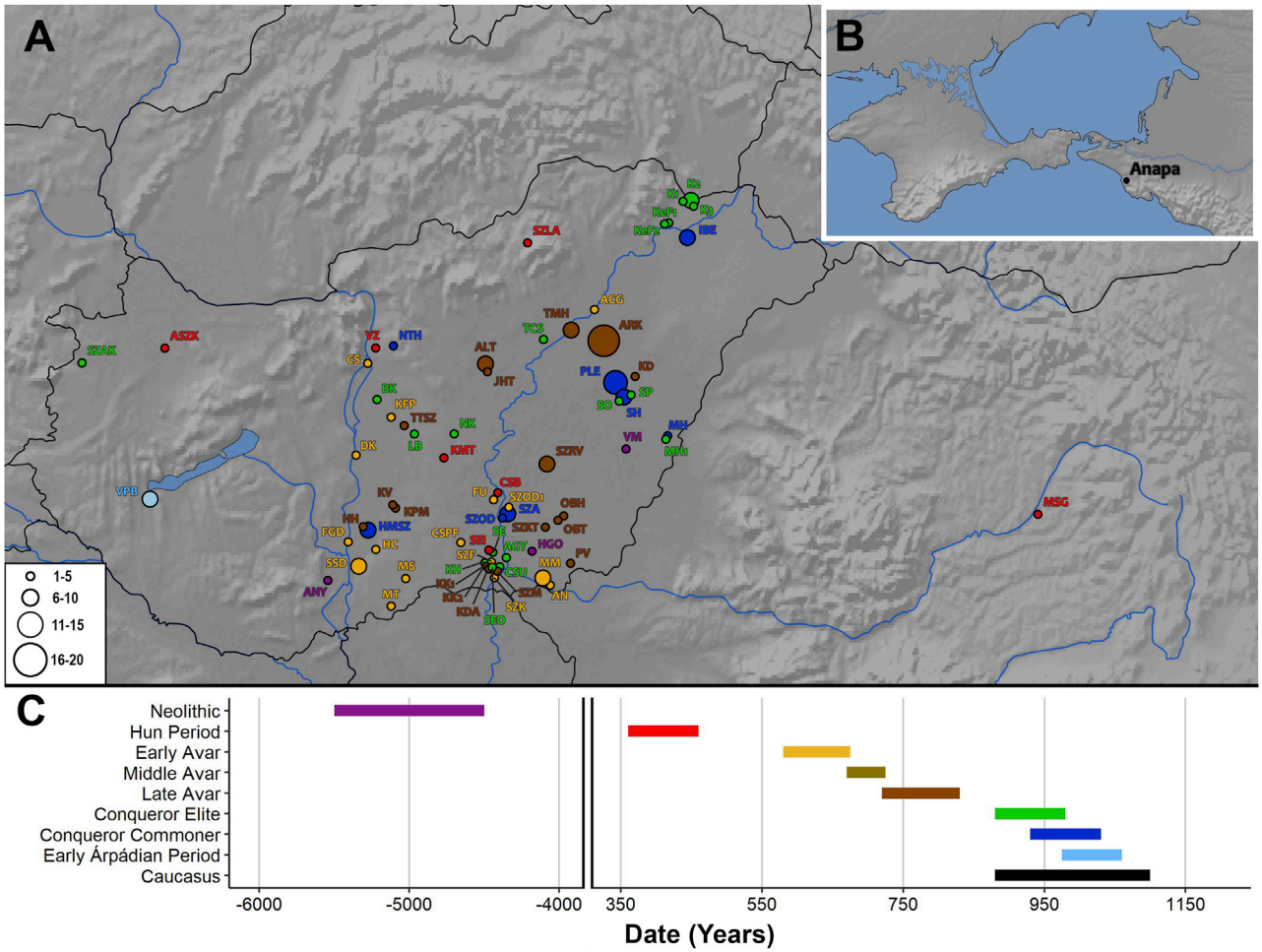


Figure 1. Archaeological sites and time periods of the studied samples

Map of Hungary with neighboring regions and the north Black Sea region (inset) is shown.

(A) Distribution of sites with their associated culture and time period are indicated by color. Color coding corresponds to the time periods labeled with the same color in (C); circle size is proportional to sample size, as indicated.

(B) Inset map of part of the north Black Sea region, from which 3 samples were studied.

(C) Timeline in years of the studied historical periods or archaeological cultures. Negative values indicate BCE dates.

See also Methods S1.

Imperial Period individuals from Italy and Iron Age Scythians from Hungary, respectively.

The diversity of the medieval Hungarian population in the Eurcline is conspicuous. We considered these groups local residents, although similar populations could hypothetically have been present on the Medieval Pontic Steppe too.

The European Huns had Xiongnu ancestry

Despite the paucity of Hun period samples, we can discern a "Hun-cline" along the PC1 axis (Figure 2A) extending east to west. Two individuals, MSG-1 and VZ-12673 (the same sample as HUN001,[21] resequenced with higher coverage), project to the extreme eastern pole of the cline, close to modern Kalmyks and Mongols. Both graves belonged to warriors and contained partial horse remains. On PC50 clustering, they tightly cluster together with two other Hun period samples: Kurayly_Hun_380CE (KRY001)[21] and a Tian Shan Hun outlier (DA127)[20] (Data S3). As latter samples also form a genetic clade with VZ-12673 in qpAdm (see below), we grouped these four samples under the name of Hun_Asia_Core (Figure 2), although analyzed the new samples separately. The Hun_Asia_Core also clusters with numerous Xiongnu, Medieval Mongol, Turk,[22] and Xianbei[21] genomes from Mongolia as well as several Avar samples from this study. ADMIXTURE confirmed the similarity of Hun_Asia_Core individuals and showed prevailing east Eurasian Nganasan and Han components with no traces of WHG (Figure 2B; Data S5), implying that these individuals represent immigrants with no European background.

Outgroup f3-statistics indicated common ancestry of MSG-1 and VZ-12673 (Data S6A), as both individuals shared highest drift with Mongolia-related populations like earlyXiongnu_rest, Ulaanzukh, and SlabGrave.[22] Although the error bars overlap with populations related to surrounding regions too, like Russia_MN_Boisman[23] and Tasmola_ Korgantas,[21] f3 results

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

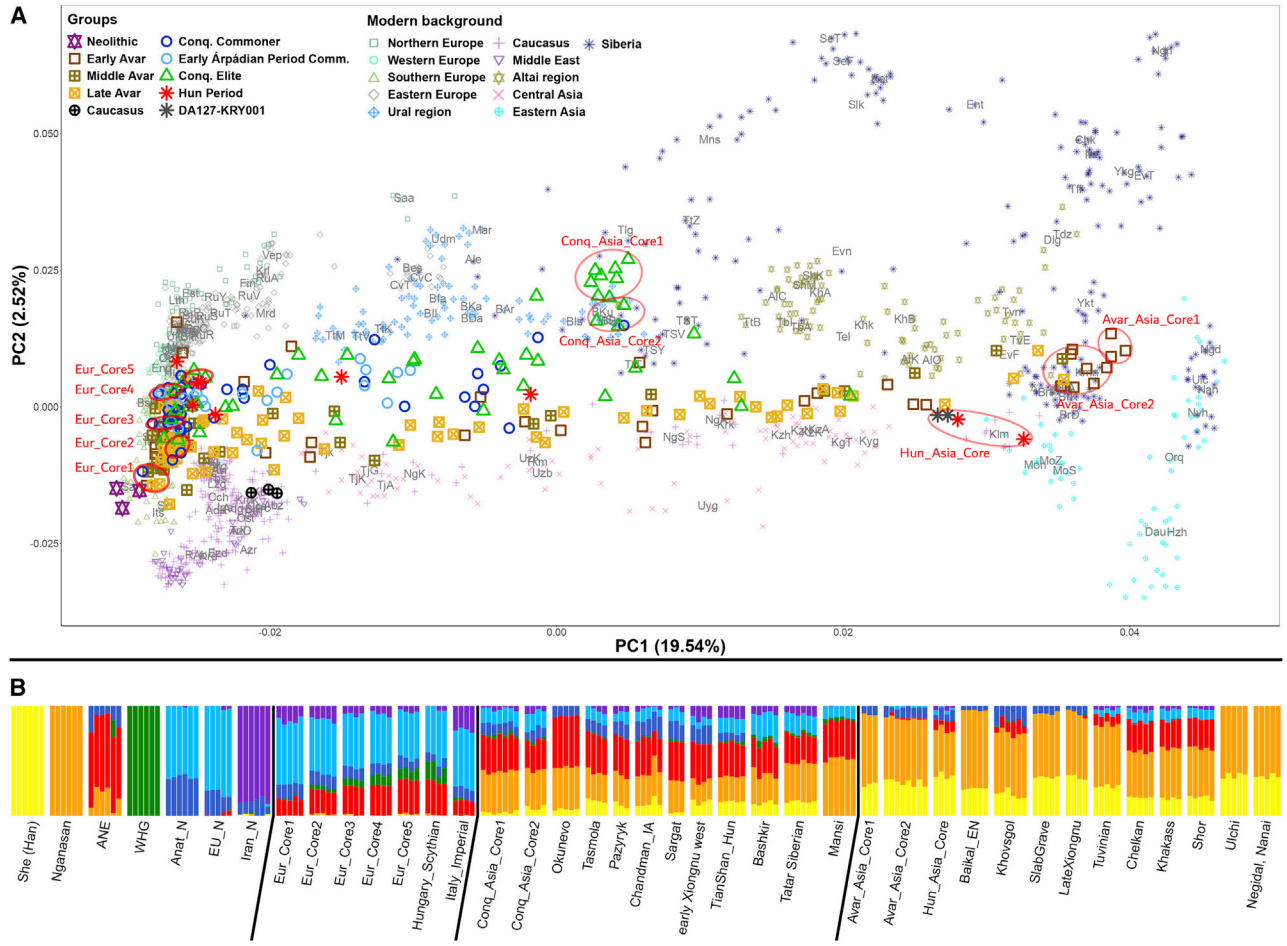


Figure 2. PCA and ADMIXTURE analysis

(A) PCA of 271 ancient individuals projected onto contemporary Eurasians (gray letter codes, defined in Table S2, printed on the median of each population). Labels of modern populations correspond to geographical regions as indicated. Conquest and Avar period samples form two separable genetic clines. Genetically homogeneous groups are encircled with red.

(B) Unsupervised ADMIXTURE (K = 7) results of the red circled core groups and the populations with most similar ADMIXTURE composition to them. The 7 populations representing each ADMIXTURE component are shown at the left.

See also Figure S1, Table S2, and Data S4, S5, S7, S8, and S9.

point to a likely Mongolian origin and early Xiongnu affinity of these individuals.

Distal qpAdm modeling from pre-Iron Age sources indicated major Khovsgol_outlier (DSKC)[22] and minor West Liao River Neolithic/Yellow River Late Bronze Age[24] ancestries in MSG-1 and VZ-12673 (Figure 3A; Data S7A) predicting Han Chinese admixture in these individuals, whereas proximal modeling from post-Bronze Age sources gave two types of alternative models representing two different time periods (Data S7B). The best p value models showed major late Xiongnu (with Han admixture) and minor Scytho-Siberian/Xianbei ancestries, whereas alternative models indicated major Kazakhstan_OutTianShanHun or Kurayly_Hun_380CE and minor Xiongnu/Xianbei/Han ancestries (Figure 3A). In latter models, VZ-12673 formed a clade with both published Hun_Asia_Core samples. In conclusion, our Hun_Asia_Core individuals could be equally modeled from earlier Xiongnu and later Hun age genomes.

The two other Hun period samples, KMT-2785 and ASZK-1, were located in the middle of the Hun-cline (Figures 2A and S1A), and accordingly, they could be modeled from European and Asian ancestors. The best passing models for KMT-2785 predicted major Late Xiongnu and minor local Eur_Core, whereas alternative model showed major Sarmatian[20] and minor Xiongnu ancestries (Data S7C). Both models implicate Sarmatians as in the Late Xiongnus of the first model, and 46%–52% Sarmatian and 48%–54% Ulaanzuukh_SlabGrave components had been predicted.[22] The ASZK-1 genome formed a clade with Sarmatians in nearly all models. The rest of the Hun period samples map to the northern half of the Eur-cline; nevertheless, two of these (SEI-1 and SEI-5) could be modeled from major Eur_Core and minor Sarmatian components (Data S7D). The prevalent Sarmatian ancestry in 4 Hun period samples implies significant Sarmatian influence on European Huns (Figure 3B).

CSB-3 was modeled from major Eur_Core and minor Scytho-Siberian ancestries, whereas SEI-6 formed a clade with the

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

CellPress
OPEN ACCESS

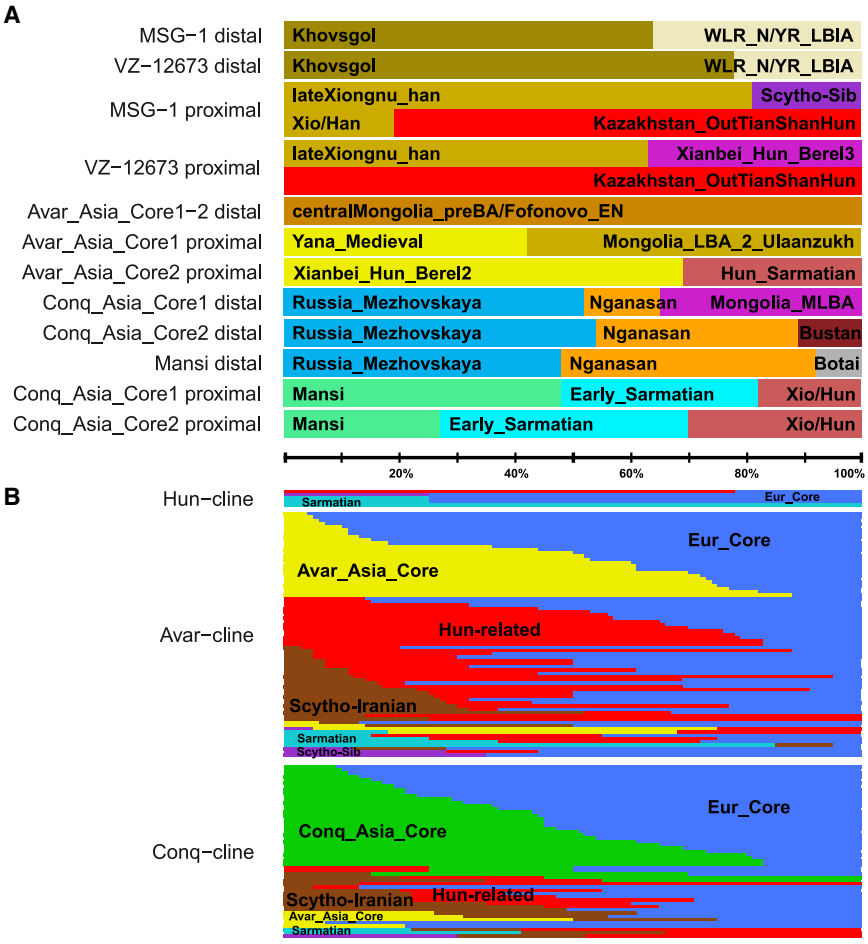


Figure 3. Summary of the qpAdm models

(A) Distal and proximal qpAdm models for Hun_Asia_Core individuals and Avar_Asia_Core and Conq_Asia_Core groups. Distal model of modern Mansi is also shown for comparison.

(B) Proximal qpAdm models for 5 Hun-cline, 53 Conqueror-cline, and 75 Avar-cline individuals. Each individual is represented by a thin column, which are grouped according to the similarity of components. Data are summarized from Data S1C, S7, S8, and S9. Identical or similar genome compositions are shown with similar colors.

Ukraine_Chernyakhiv[25] (Eastern Germanic/Goth) genomes (Data S7D). The SZLA-646 outlier individual at the top of the Eur-cline formed a clade with Lithuania_Late_Antiquity[20] and England_Saxon[26] individuals (Data S7E). The last two individuals presumably belonged to Germanic groups allied with the Huns.

Huns and Avars had related ancestry

Our Avar period samples also form a characteristic PCA "Avar-cline" on Figure 2A, extending from Europe to Asia. PC50 clustering, at level 50, identified a single genetic cluster at the Asian extreme of the cline with 12 samples, derived from 8 different cemeteries, that we termed Avar_Asia_Core (Figure 2; Data S3). In total, 10 of 12 samples of Avar_Asia_Core were assigned to the early Avar period, 4 of them belonging to the elite and 9 of 12 males. Elite status is indicated by richly furnished burials, e.g., swords and sabers with precious metal fittings, gold earrings, gilded belt fittings, etc., as previously described.[27]

Avar_Asia_Core clusters together with Shamanka_Eneolithic and Lokomotiv_Eneolithic[28] samples from the Baikal region, as well as with Mongolia_N_East, Mongolia_N_North,[23] Fofonovo_EN, Ulaanzuukh_SlabGrave, and Xiongnu[22] from Mongolia (Data S3). This result is recapitulated in ADMIXTURE (Figure 2B), which also shows that Nganasan and Han components are predominant in Avar_Asia_Core with traces of Anat_N and ANE, whereas Iranian and WHG constituents are entirely missing. It follows that Avar_Asia_Core was derived from East Asia, most likely from present day Mongolia.

We performed two-dimensional f4-statistics to detect minor genetic differences within the Avar_Asia_Core group. Avar_Asia_Core individuals could be separated based on their affinity to Bactria-Margiana Archaeological Complex (BMAC) and Steppe Middle-Late Bronze Age (Steppe_MLBA) populations (Figure 4), with 3 individuals bearing negligible proportion of these ancestries. The Steppe_MLBA-ANE f4-statistics gave similar results. As the 3 individuals with the smallest Iranian and Steppe affinities also visibly separated on PCA, we set these apart under the name of Avar_Asia_Core1, whereas the other 9 samples were regrouped as Avar_Asia_Core2 (Figure 2).

According to outgroup f3-statistics, both Avar_Asia_Core groups had highest shared drift with genomes with predominantly Ancient North-East Asian (ANA) ancestry (Data S6B), like earlyXiongnu_rest, Ulaanzuukh, and Slab Grave.[22] It is notable that from the populations with top 50 f3 values, 41 are shared with Hun_Asia_Core; moreover, Avar_Asia_Core1 is in the top 50 populations for both Hun_Asia_Core samples, signifying common deep ancestry of European Huns and Avars.

According to distal qpAdm models, Avar_Asia_Core formed a clade with the Fofonovo_EN and centralMongolia_preBA genomes (Figure 3A; Data S8A), both of which had been modeled from 83%–87% ANA and 12%–17% ANE.[22] All data consistently show that Avar_Asia_Core preserved very ancient Mongolian pre-Bronze Age genomes, with ca 90% ANA ancestry.

Most proximodistal qpAdm models (defined in STAR Methods) retained distal sources, as Avar_Asia_Core1 was modeled from 95% UstBelaya_N[29] plus 5% Steppe Iron Age (Steppe_IA) and Avar_Asia_Core2 from 80%–92% UstBelaya_N plus 8%–20% Steppe_IA (Data S8B). The exceptional proximal model for Avar_Asia_Core1 indicated Yana_Medieval[29] plus Ulaanzukh, whereas for Avar_Asia_Core2, Xianbei_Hun_Berel[21] plus Kazakhstan_Nomad_Hun_Sarmatian[20] ancestries (Figure 3A).

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

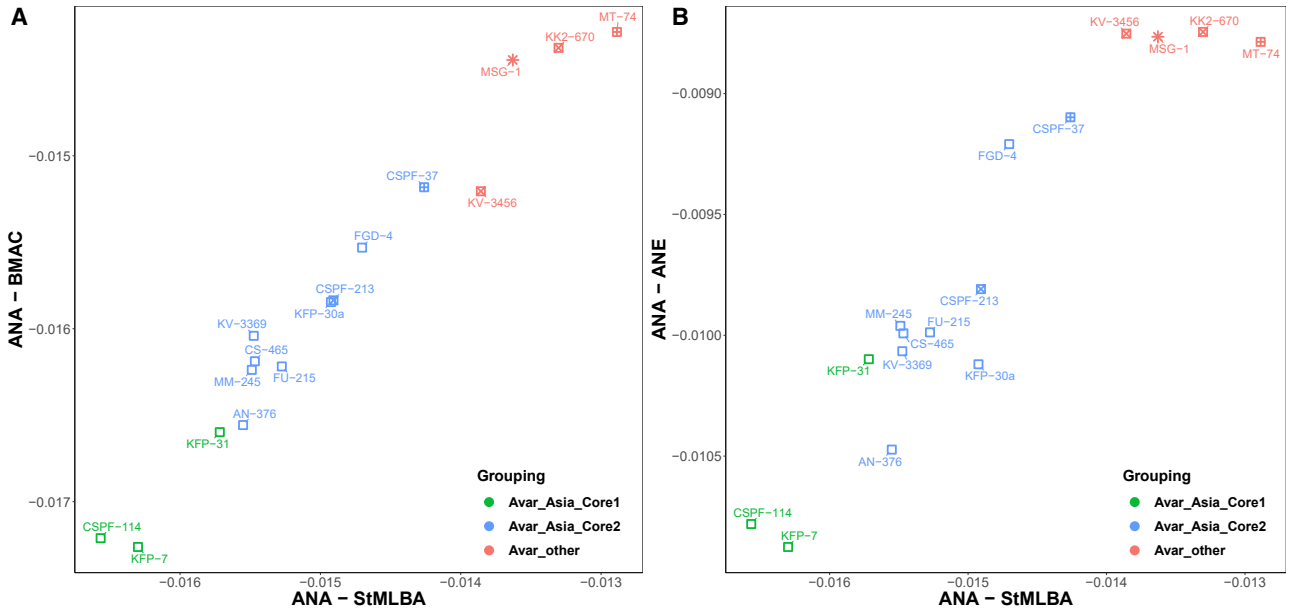


Figure 4. Two-dimensional f4-statistics of Avar_Asia_Core individuals

(A) f4 values from the statistics *f4(Ethiopia_4500BP, Test; Ulaanzuukh_SlabGrave, MLBA_Sintashta)* versus *f4(Ethiopia_4500BP, Test; Ulaanzuukh_SlabGrave, Uzbekistan_BA_Bustan)* measuring the relative affinities of Test individuals with Steppe_MLBA and BMAC.

(B) f4 values form the statistics *f4(Ethiopia_4500BP, Test; Ulaanzuukh_SlabGrave, MLBA_Sintashta)* versus *f4(Ethiopia_4500BP, Test; Ulaanzuukh_SlabGrave, Kazakhstan_Eneolithic_Botai)* measuring the relative affinities of test individuals with Steppe_MLBA and ANE.

The latter model also points to shared ancestries between Huns and Avars.

From the 76 samples in the Avar-cline, 26 could be modeled as a simple 2-way admixture of Avar_Asia_Core and Eur_Core (Figure 3B; Data S8C), indicating that these were admixed descendants of locals and immigrants, whereas further 9 samples required additional Hun- and/or Iranian-related sources. In the remaining 41 models, Hun_Asia_Core and/or Xiongnu sources replaced Avar_Asia_Core (Figure 3B; Data S8D; summarized in Data S1C). Scythian-related sources with significant Iranian ancestries, like Alan, Tian Shan Hun, Tian Shan Saka,[20] or Anapa (this study), were ubiquitous in the Avar-cline, but given their low proportion, qpAdm was unable to identify the exact source.

Xiongnu/Hun-related ancestries were more common in certain cemeteries, for example, it was detected in most samples from Hortobágy-Árkus (ARK), Szegvár-Oromdűlő (SZOD), Makó-Mikócsa-halom (MM), and Szarvas-Grexa (SZRV) (Data S8D).

The Conquerors had Ugric, Sarmatian, and Hun ancestries

The Conquest period samples also form a characteristic genetic "Conq-cline" on PCA (Figure 2A). It is positioned north of the Avar-cline, although only reaching the midpoint of the PC1 axis. PC50 clustering identified a single genetic cluster at the Asian extreme of the cline (Data S3) with 12 samples, derived from 9 different cemeteries, that we termed Conq_Asia_Core. This genetic group consists of 6 males and 6 females, and 11 of the 12 individuals belonged to the Conqueror elite according to archaeological evaluation.

The PCA position of Conq_Asia_Core corresponds to modern Bashkirs and Volga Tatars (Figure 2A), and they cluster together with a wide range of eastern Scythians, western Xiongnus, and Tian Shan Huns,[20] which is also supported by ADMIXTURE (Figure 2B).

Two-dimensional f4-statistics detected slight genetic differences between Conq_Asia_Core individuals (Figure 5), obtained via multiple gene flow events, as they had different affinity related to Miao (a modern Chinese group) and Ulaanzuukh_SlabGrave (ANA).[22] Individuals were arranged linearly along the Miao-ANA cline, suggesting that these ancestries covary in the Conqueror group and thus could have arrived together, most likely from present day Mongolia. As four individuals with highest Miao and ANA affinities also had shifted PCA locations, we set these apart under the name of Conq_Asia_Core2, whereas the rest were regrouped as Conq_Asia_Core1 (Figure 2). Along the ANE and BMAC axes, the samples showed a more scattered arrangement, although Conq_Asia_Core2 individuals showed somewhat higher BMAC ancestry.

Admixture f3-statistics indicated that the main admixture sources of Conq_Asia_Core1 were ancient European populations and ancestors of modern Nganasans (Data S6C). The most likely direct source of the European genomes could be Steppe_MLBA populations, as these distributed European ancestry throughout of the Steppe.[30]

Outgroup f3-statistics revealed that Conq_Asia_Core1 shared highest drift with modern Siberian populations speaking Uralic languages, Nganasan (Samoyedic), Mansi (Ugric), Selkup (Samoyedic), and Enets (Samoyedic) (Data S6E), implicating that Conq_Asia_Core shared evolutionary past with language relatives of modern Hungarians. We also performed f4-statistics to test whether the shared evolutionary past was restricted to language relatives. The f4-statistics showed that Conq_Asia_Core1 indeed

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

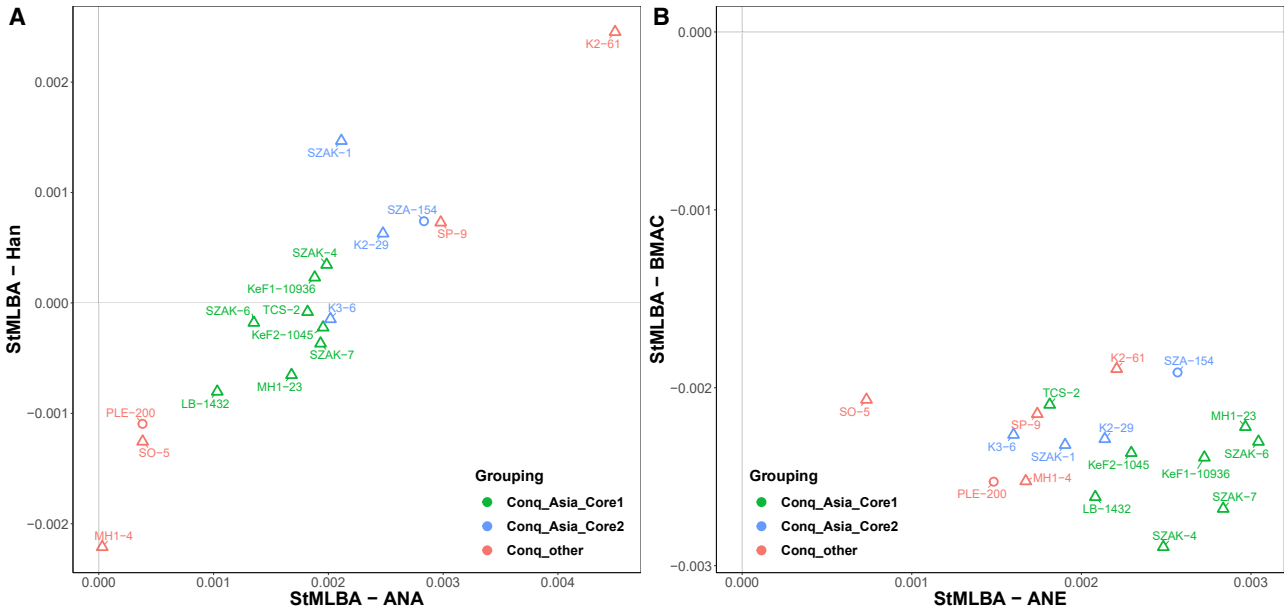


Figure 5. Two-dimensional f4-statistics of Conq_Asia_Core individuals

(A) f4 values from the statistics *f4(Ethiopia_4500BP, Test; MLBA_Sintashta, Ulaanzuukh_SlabGrave)* versus *f4(Ethiopia_4500BP, Test; MLBA_Sintashta, Miao_-modern)* measuring the relative affinities of Test individuals with ANA and Han.

(B) f4 values from the statistics *f4(Ethiopia_4500BP, Test; Sintashta, Kazakhstan_Eneolithic_Botai)* versus *f4(Ethiopia_4500BP, Test; Sintashta, Uzbekistan_BA_Bustan)* measuring the relative affinities of Test individuals with ANE and BMAC.

had highest affinity to Mansis, the closest language relatives of Hungarians, but its affinity to Samoyedic-speaking groups was comparable with that of Yeniseian-speaking Kets and Chukotko-Kamchatkan-speaking Koryaks (Data S6G). For this reason, we coanalyzed Mansis with Conq_Asia_Core.

From pre-Iron Age sources, Mansis could be qpAdm modeled from Mezhovskaya,[18] Nganasan, and Botai,[28] and Conq_Asia_Core1 from Mezhovskaya, Nganasan, Altai_MLBA_o,[21] and Mongolia_LBA_CenterWest_4D[23] (Figure 3A; Data S9A and S9B), confirming shared late Bronze Age ancestries of these groups but also signifying that the Nganasan-like ancestry was largely replaced in Conq_Asia_Core by a Scytho-Siberian-like ancestry including BMAC[21,23] derived from the Altai-Mongolia region. The same analysis did not give passing models for Kets and Koryaks, confirming that they had different genome histories.

From proximal sources, Conq_Asia_Core1 could be consistently modeled from 50% Mansi, 35% Early/Late Sarmatian, and 15% Scytho-Siberian-outlier/Xiongnu/Hun ancestries, and Conq_Asia_Core2 had comparable models with shifted proportions (Figure 3A; Data S9C). As the source populations in these models defined inconsistent time periods, we performed DATES analysis[30] to clarify admixture time.

DATES revealed that the Mansi-Sarmatian admixture happened ~53 generations before death of the Conqueror individuals, around 643–431 BCE, apparently corresponding to the Sauromatian/early Sarmatian period. The Mansi-Scythian/Hun-related admixture was dated ~24 generations before death, or 217–315 CE, consistent with the post-Xiongnu, pre-Hun period rather than the Iron Age (Figure 6).

Most individuals of the Conqueror cline proved to be admixed descendants of the immigrants and locals: from the 42 samples in the Conq-cline, 31 could be modeled as two-way admixtures of Conq_Asia_Core and Eur_Core (Figure 3B; Data S9D; summarized in Data S1C). The remaining samples mostly belonged to the elite, many projecting with the Avar-cline (Figure 2A); of these 5 samples could be modeled from Conq_Asia_Core requiring Hun- and Iranian-associated additional sources. In total, 17 outlier individuals lacked Conq_Asia_Core ancestry, which was replaced with Avar_Asia_Core or Xiongnu/Hun-related sources, accompanied by Iranian-associated 3rd sources (Figure 3B; Data S9E). It seems from our data that Conqueror elite individuals with Hun-related genomes were clustered in certain cemeteries; for example, each sample from Szeged-Öthalom (SEO), Algyő 258-as kútkörzet (AGY), Nagykőrös-Fekete-dűlő (NK), Sándorfalva-Eperjes (SE), and Sárrétudvari-Poroshalom (SP) had this ancestry.

Neolithic and Caucasus samples

We have sequenced three new Neolithic genomes from Hungary from three different cemeteries. Two individuals represented the Hungary_Tisza Neolithic culture and one individual the Hungary_Starcevo early Neolithic culture (Data S1). Other genomes had been published previously from each site,[31] and our genomes have the same ADMIXTURE profile (Data S5) and PCA location (Figure 2) as the previously published samples and also cluster together with Anatolian and European farmers on our PC50 clustering (Data S3).

We have also sequenced 3 samples derived from the Caucasus region (Figure 1) with possible archaeological affinity to the Conquerors. On PCA, Anapa individuals project onto modern samples from the Caucasus (Figure 2), and they cluster together with ancient samples from the Caucasus and BMAC regions

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

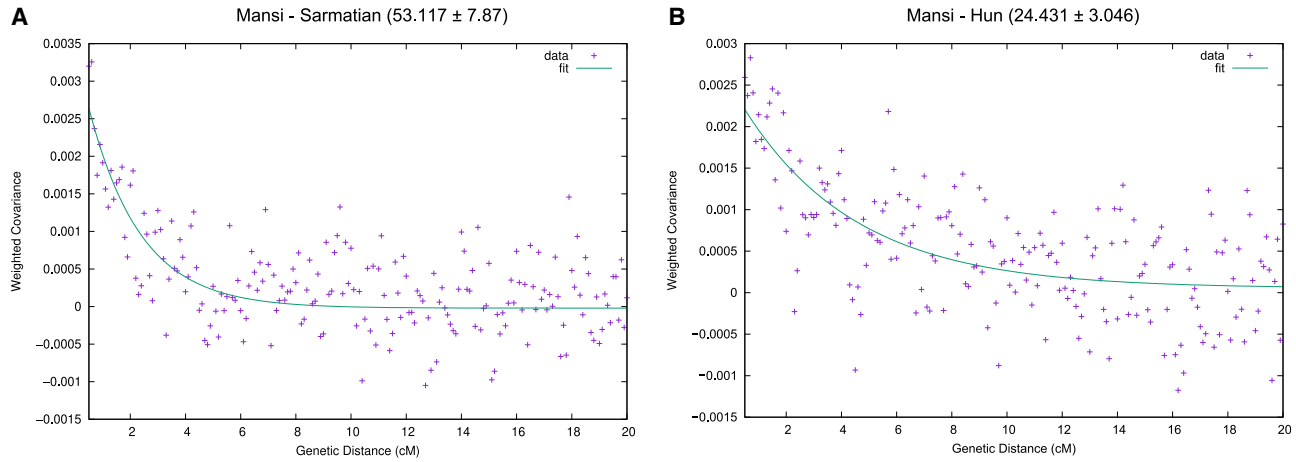


Figure 6. DATES analysis to estimate admixture time

Figures show the weighted ancestry covariant decays for the indicated two-way admixture sources. Curves show the fitted exponential functions, from which the number of generations since admixture are calculated by the program. (A) Mansi-Sarmatian and (B) Mansi-Hun admixture time in generations, in the Conq_Asia_Core target.

(Data S3). The ADMIXTURE profile of the Anapa individuals indicated 33% Iranian, 32% EU_N, 19% ANE, 8% Anat_N, 6% Han 3% Nganasan, and no WHG components (Data S5). These data implied European, Iranian, and Steppe admixtures, the latter including ANA ancestry.

qpAdm revealed that the Anapa individuals carried ancient genomes derived from the Caucasus region, as Armenia_MBA was the majority (76%–96%) source in each model (Data S7F and S7G). In a single lower p value model, Russia_SaltovoMayaki[20] formed a clade with our Anapa samples, but this has to be interpreted with caution, as the three available Saltovo-Mayaki genomes have very low coverage (0.029×, 0.04×, and 0.072×). The close proximity to Saltovo-Mayaki samples on PCA as well as their clustering together suggest that proximal sources of Anapa could be similar to that of Saltovo-Mayaki, as these were contemporary individuals and both could be part of the Khazar Khaganate.

Y-chromosome and mtDNA results

The distribution of uniparental markers along the PCA genetic clines shows a general pattern: at the Asian side of the cline, we find individuals with Asian haplogroups (Hgs), whereas at the European side, individuals carry European Hgs. Along the cline from Asia toward Europe, the same trend prevails, decreasing frequency of Asian and increasing of European Hgs. The few exceptions from this rule are nearly always detected in admixed individuals (Data S1A); nevertheless, several individuals in the Eur-cline carried Asian Hgs, testifying distant Asian forefathers. This is especially prominent in the Jánoshida-Tótkérpuszta (JHT) Late Avar graveyard, where all three males carried R1a1a1b2a (R1a-Z94) Asian Y-Hg, in spite of their European genomes. The Middle Avar MT-17 individual from Madaras-Téglavető in the Eur-cline also carried R1a-Z94, although in this cemetery, all three other males carried N1a1a1a1a3a with Asian genomes.

Both Hun_Asia_Core individuals (VZ-12673 and MSG-1) carry R1a-Z94 as well as ASZK-1 in the Hun-cline. The other two published genomes united in Hun_Asia_Core, Kurayly_Hun_380CE[21] and Kazakhstan_OutTianShanHun,[20] carry Hgs R1a-Z94 and Q, respectively, suggesting that these Hgs could be common among the Huns. Considering all published post-Xiongnu Hun era genomes (Hun period nomad, Hun-Sarmatian, Tian Shan Hun,[20] and Xianbei-Hun Berel[21]), we counted 10/23 R1a-Z93 and 9/23 Q Hgs, supporting the above observation. These Y-Hgs were most likely inherited from Xiongnus, as these Hgs were frequent among them[22,32] but were rare in Europe before the Hun period. The rest of our Hun period samples with European genomes carried derivatives of R1a1a1b1, an Hg typical in North-Western Europe, in line with the Germanic affinity of many of these samples shown above.

From the 9 Avar_Asia_Core males, 7 carried the N1a1a1a1a3a (N1a-F4205) Y-Hg, one C2a1a1b1b, and one R1a1a1b∼ (very likely R1a-Z94). This confirms that N1a-F4205, most prevalent in modern Chukchis and Buryats,[33] was also prevailing among the Avar elite as shown before.[34,35] This Hg was also common in members of the Avar-cline and seems to cluster in certain cemeteries. In the Ároktő (ACG), Felgyő (FU), SZOD, Csepel-Kavicsbánya (CS), Kiskőrös-Vágóhídi dűlő (KV), Kunpeszér-Felsőpeszér (KFP), Csólyospálos-Felsőpálos (CSPF), Kiskundorozsma-Kettőshatár II (KK2), Tatárszentgyörgy (TTSZ), and Madaras-Téglavető (MT) Avar graveyards, all or the majority of males carried the N1a-F4205 Hg, mostly accompanied with Asian maternal lineages. These cemeteries must have belonged to the immigrant Avar population, whereas the local population seems to have separated, as many Avar period cemeteries show no sign of Asian ancestry. The latter include Mélykút-Sáncdűlő (MS), Szeged-Fehértó A (SZF), Szeged-Kundomb (SZK), Szeged-Makkoserdő (SZM), Kiskundorozsma-Kettőshatár I (KK1), Kiskundorozsma-Daruhalom (KDA), Oroshá-za-Bónum Téglagyár (OBT), Székkutas-Kápolnadűlő (SZKT), Homokmégy-Halom (HH), Alattyán-Tulát (ALT), Kiskőrös-Pohibuj Mackó dűlő (KPM), and Sükösd-Ságod (SSD), in which Asian lineages barely occur. In the SZK, ALT, KK1, OBT, SZKT, HH, and SZM cemeteries, most males belonged to the

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

CellPress
OPEN ACCESS

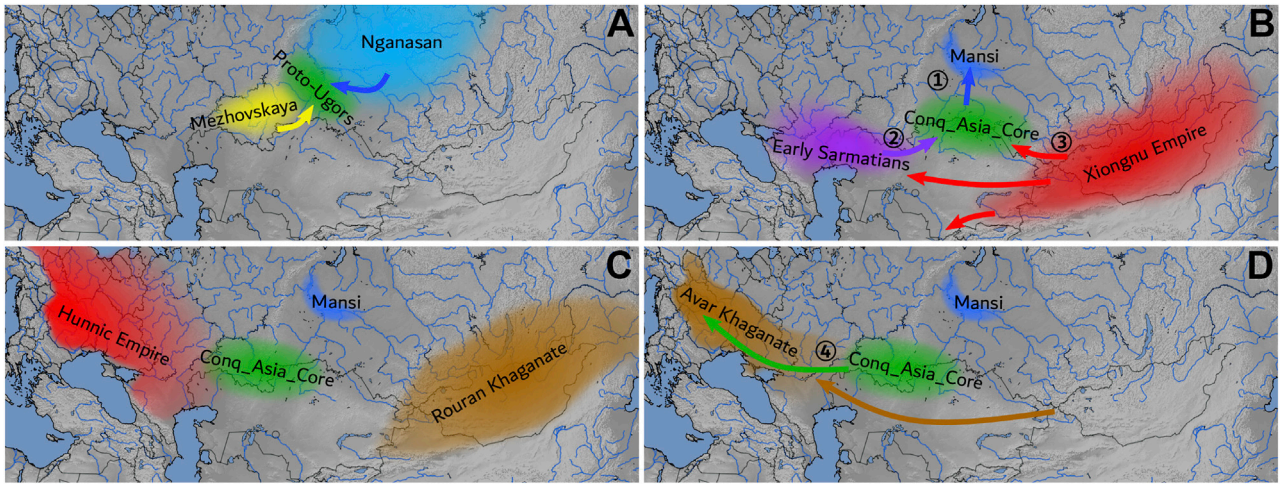


Figure 7. Summary map

(A) Proto-Ugric peoples emerged from the admixture of Mezhovskaya and Nganasan populations in the late Bronze Age.

(B) (1) During the Iron Age Mansis separated and (2) Proto-Conquerors admixed with Early Sarmatians ca 643–431 BCE and (3) with pre-Huns ca 217–315 CE.

(C) By the 5th century, the Xiongnu-derived Hun Empire occupied Eastern Europe, incorporating its population, and the Rouran Khaganate emerged on the former Xiongnu territory.

(D) By the middle 6th century, the Avar Khaganate occupied the territory of the former Hun Empire, incorporating its populations. (4) By the 10th century, Conquerors associated with the remnants of both empires during their migration and within the Carpathian Basin.

E1b1b1a1b1 (E-V13) Hg, which is most prevalent in the Balkan,[36] and accordingly, many of the samples from these cemeteries fell in Eur_Core1 or its vicinity, with typical southern European genomes.

There is a third group of Avar period cemeteries representing immigrants from Asia, but with a different genetic background. In males from MM, Dunavecse-Kovacsos dűlő (DK), Árkus Homokbánya (ARK), and SZRV Y-Hgs R1a-Z94 and Q1a2a1 dominated, which seem typical in European Huns, and were mostly accompanied by Asian maternal lineages. These Avar period people could have represented Hun remnants that joined the Avars but isolated in separate communities. These inferences are perfectly in line with genomic data, as most qpAdm models from these cemeteries indicated the presence of Hun_Asia_Core or Xiongnu ancestries (Data S1C). As mentioned above, Hun ancestry was also present in several other cemeteries, just like Hg R1a-Z94, but in those cemeteries, the population was genetically less uniform.

The Conqueror population had a more heterogeneous Hg composition compared with the Huns and Avars. In the 6 Conq_Asia_Core males, we detected three N1a1a1a1a2∼, one D1a1a1a1b, one C2a1a1b1b, and one Q1a1a1 Y-Hg, generally accompanied by Asian maternal lineages. Two other N1a1a1a1a2a1c∼ Y-Hgs were detected in the SO-5 Conqueror elite and the PLE-95 commoner individuals; thus, this Hg seems specific for the Conqueror group. Obviously, this Hg links Conquerors with Mansis, as had been shown before.[37] Another related Y-Hg, N1a1a1a1a4 (M2128), was detected in two Conqueror elite samples from present study as well as from another two Conqueror elite samples in our previous study.[34] This Hg is typical for modern Yakuts and occurs with lower frequency among Khantys, Mansis, and Kazakhs,[33] and thus may also link Conquerors with Mansis, although it was also present in one Middle Avar individual. It is notable that the European Y-Hg I2a1a2b1a1a was also specific for the Conqueror group, especially for the elite as also shown before,[34] very often accompanied by Asian maternal lineages, indicating that I2a1a2b1a1a could be more typical for the immigrants than to the local population. Additionally, two other Y-Hgs appeared with notable frequencies among the Conquerors: R1a-Z94 was present in 3 elite and 2 commoner individuals, whereas Hg Q was carried by 3 elite individuals, which may be sign of Hun relations, also detected at the genome level. This result is again in line with genome data, as nearly all Conquest period males with R1a-Z94 or Q Hgs carried Hun-related ancestry.

DISCUSSION

The genomic history of Huns, Avars, and Conquerors revealed in this study is compatible with historical, archaeological, anthropological, and linguistic sources (summarized in Figure 7). Our data show that at least part of the military and social leader strata of both European Huns and Avars likely originated from the area of the former Xiongnu Empire, from present day Mongolia, and both groups can be traced back to early Xiongnu ancestors. Northern Xiongnus were expelled from Mongolia in the second century CE, and during their westward migration, Sarmatians were one of the largest groups they confronted. Sergey Botalov presumed the formation of a Hun-Sarmatian mixed culture in the Ural region before the appearance of Huns in Europe,[38] which fits the significant Sarmatian ancestry detected in our Hun samples, although this ancestry had been present in late Xiongnus as well.[22] Thus our data are in accordance with the Xiongnu ancestry of European Huns, claimed by several historians.[39,40] We also detected Goth- or other Germanic-type genomes[25] among our Hun period samples, again consistent with historical sources.[39]

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

Most of our Avar_Asia_Core individuals represented the early Avar period and half of the "elite" samples belonged to Avar_Asia_Core (Data S1). The other elite samples also contained a high proportion of this ancestry, suggesting that this ancestry could be prevalent among the elite, although also present in common people. The elite preserved very ancient east Asian genomes with well-defined origin, as had been also inferred from Y-Hg data.[34,35] Our data are compatible with the Rouran origin of the Avar elite,[5] although the single low-coverage Rouran genome[41] provided a poor fit in the qpAdm models (Data S8B). However, less than half of the Avar-cline individuals had Avar_Asia_Core ancestry, indicating the diverse origin of the Avar population. Our models indicate that the Avars incorporated groups with Xiongnu/Hun_Asia_Core and Iranian-related ancestries, presumably the remnants of the European Huns and Alans or other Iranian peoples on the Pontic Steppe, as suggested by Kim.[39] People with different genetic ancestries were seemingly distinguished, as samples with Hun-related genomes were buried in separate cemeteries.

The Conquerors, who arrived in the Carpathian Basin after the Avars, had a distinct genomic background with elevated levels of western Eurasian admixture. Their core population carried very similar genomes to modern Bashkirs and Tatars, in agreement with our previous results from uniparental markers.[34,42] Their genomes were shaped by several admixture events, of which the most fundamental was the Mezhovskaya-Nganasan admixture around the late Bronze Age, leading to the formation of a "proto-Ugric" gene pool. This was part of a general demographic process, when most Steppe_MLBA populations received an eastern Khovsgol-related Siberian influx together with a BMAC influx,[21] and ANA-related admixture became ubiquitous on the eastern Steppe,[30] establishing the Scytho-Siberian gene pool. Consequently proto-Ugric groups could be part of the early Scytho-Siberian societies of the late Bronze Age to early Iron Age steppe-forest zone in the northern Kazakhstan region, in the proximity of the Mezhovskaya territory.

Our data support linguistic models, which predicted that Conquerors and Mansis had a common early history.[7,43] Then Mansis migrated northward, probably during the Iron Age, and in isolation, they preserved their Bronze-Age genomes. In contrast, the Conquerors stayed at the steppe-forest zone and admixed with Iranian-speaking early Sarmatians, also attested by the presence of Iranian loanwords in the Hungarian language.[43] This admixture likely happened when Sarmatians rose to power and started to integrate their neighboring tribes before they occupied the Pontic-Caspian Steppe.[44]

All analysis consistently indicated that the ancestors of Conquerors further admixed with a group from Mongolia, carrying Han-ANA-related ancestry, which could be identified with ancestors of European Huns. This admixture likely happened before the Huns arrived in the Volga region (370 CE) and integrated local tribes east of the Urals, including Sarmatians and the ancestors of Conquerors. These data are compatible with a Conqueror homeland around the Ural region, in the vicinity of early Sarmatians, along the migration route of the Huns, as had been surmised from the phylogenetic connections between the Conquerors and individuals of the Kushnarenkovo-Karayakupovo culture in the Trans-Uralic Uyelgi cemetery.[45] Recently, a Nganasan-like shared Siberian genetic ancestry was detected in all Uralic-speaking populations, Hungarians being an exception.[46] Our data resolve this paradox by showing that the core population of conquering Hungarians had high Nganasan ancestry. The fact that this is negligible in modern Hungarians is likely due to the substantially smaller number of immigrants compared with the local population.

The large number of genetic outliers with Hun_Asia_Core ancestry in both Avars and Conquerors testifies that these successive nomadic groups were indeed assembled from overlapping populations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Ancient samples
- METHOD DETAILS
  - Accelerator mass spectrometry radiocarbon dating
  - Ancient DNA laboratory work
  - NGS library construction
  - DNA sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Bioinformatical processing
  - Quality assessment of ancient sequences
  - Uniparental haplogroup assignment
  - Genetic sex determination
  - Estimation of genetic relatedness
  - Population genetic analysis
  - Principal Component Analysis (PCA)
  - Unsupervised Admixture
  - Hierarchical Ward clustering
  - Admixture modeling using qpAdm
  - f3-statistics
  - Two dimensional f4-statistics
  - Dating admixture time with DATES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.04.093.

ACKNOWLEDGMENTS

We are grateful to our archaeologist colleagues Gabriella M. Lezsák and Andrej Novicsihin for providing us with the Anapa samples, Gábor Lőrinczy for his help regarding the Avar material, and Zsófia Rácz for her help with the Hun period samples. We are thankful to all the museum curators and archaeologists who provided bone material for this study: Herman Ottó Museum Miskolc, Laczkó Dezső Museum Veszprém, Budapest History Museum, Ferenczy Museum Szentendre, Dobó István Castle Museum Eger, Jósa András Museum Nyíregyháza, Katona József Museum Kecskemét, and Janus Pannonius Museum Pécs. This research was funded by grants from the National Research, Development and Innovation Office (K-124350 to T.T. and TUDFO/5157-1/2019-ITM and TKP2020-NKA-23 to E.N.), The House of Árpád

Programme (2018–2023) Scientific Subproject: V.1. Anthropological-Genetic portrayal of Hungarians in the Árpadian Age to T.T. and no. VI/1878/2020. certificate number grants to E.N.

AUTHOR CONTRIBUTIONS

Conceptualization, supervision, project administration, and funding acquisition, T.T. and E. Neparáczki; data curation and software, Z.M., T.K., and E. Nyerki; formal analysis, validation, and methodology, Z.M., O.S., and T.T.; resources; Z.G., C.H., S.V., L.K., C.B., C.S., G.S., E.G., A.P.K., B.G., B.N.K., S.S.G., P.T., B.T., A.M., G.P., Z.B., A.Z., and F.M.; investigation, K.M., G.I.B.V., B.K., E. Neparáczki, O.S., P.L.N., I.N., D.L., A.G., and R.G.; visualization, Z.M. and O.S.; writing – original draft, T.T. with considerable input from I.R.; writing – review & editing, all authors.

DECLARATION OF INTERESTS

P.L.N. from Praxis Genomics LLC, I.N. and D.L. from SeqOmics Biotechnology Ltd., and Z.G. from Ásatárs Ltd. were not directly involved in the design of the experiments, data analysis, and evaluation. These affiliations do not alter our adherence to *Current Biology*'s policies on sharing data and materials.

Received: January 28, 2022
Revised: March 10, 2022
Accepted: April 28, 2022
Published: May 25, 2022

REFERENCES

1. Schmauder, M. (2015). Huns, Avars, Hungarians – reflections on the interaction between steppe empires in southeast Europe and the late roman to early byzantine empires. In Complexity of Interaction along the Eurasian Steppe Zone in the First C.E. Millennium, J. Bemmann, and M. Schmauder, eds. (Bonn Contributions to Asian Archaeology), pp. 671–692.
2. Szentpétery, I. (1937). Scriptores Rerum Hungaricarum Tempore Ducum Regnumque Stirpis Arpadianae Gestarum (Acad. Litter. Hungarica).
3. Hóman, B. (1925). A magyar hún-hagyomány és hún monda (Studium).
4. De La Vaissiére, É. (2014). The Steppe World and the Rise of the Huns (Cambridge University Press), pp. 175–192.
5. Golden, P.B. (2013). Some notes on the Avars and Rouran (Editura Universitatii "Alexandru Ioan cuza"). In The Steppe Lands and the World Beyond Them: Studies in Honor of Victor Spinei on His 70th Birthday, F. Curta, and B.-P. Maleon, eds. (Editura Universitatii "Alexandru Ioan Cuza"), pp. 43–66.
6. Golden, P.B. (1992). An Introduction to the History of the Turkic Peoples: Ethnogenesis and State Formation in Medieval and Early Modern Eurasia and the Middle East (O. Harrassowitz).
7. Róna-Tas, A. (1999). Hungarians and Europe in the Early Middle Ages: An Introduction to Early Hungarian History (Central European University Press).
8. Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjänen, K., and Wahlberg, N. (2013). Cultural and climatic changes shape the evolutionary history of the Uralic languages. J. Evol. Biol. *26*, 1244–1253.
9. Fodor, I. (2010). A hun–magyar rokonság elmélete. In Egyezzünk ki a múltal! Műhelybeszélgetések történelmi mítoszainkról, tévhiteinkről, L. Lőrinc, ed. (Történelemtanárok Egylete), pp. 165–168.
10. Rady, M. (2018). Attila and the hun tradition in Hungarian medieval texts. In Project MUSE - Studies on the Illuminated Chronicle, J.M. Bak, and L. Veszprémy, eds. (Central European Medieval Texts), pp. 127–138.
11. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.
12. Renaud, G., Slon, V., Duggan, A.T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. *16*, 224.
13. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinform. *15*, 1–13.
14. Nyerki, E., Kalmár, T., Schütz, O., Lima, R.M., Neparáczki, E., Török, T., and Maróti, Z. (2022). An optimized method to infer relatedness up to the 5th degree from low coverage ancient human genomes. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.11.480116.
15. Amorim, C.E.G., Vai, S., Posth, C., Modi, A., Koncz, I., Hakenbeck, S., La Rocca, M.C., Mende, B., Bobo, D., Pohl, W., et al. (2018). Understanding 6th-century barbarian social organization and migration through paleogenomics. Nat. Commun. *91*, 1–11.
16. Antonio, M.L., Gao, Z., Moots, H.M., Lucci, M., Candilio, F., Sawyer, S., Oberreiter, V., Calderon, D., Devitofranceschi, K., Aikens, R.C., et al. (2019). Ancient Rome: a genetic crossroads of Europe and the Mediterranean. Science *366*, 708–714.
17. Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., Furtwängler, A., Peltzer, A., Posth, C., Vasilakis, A., et al. (2017). Genetic origins of the Minoans and Mycenaeans. Nature *548*, 214–218.
18. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.
19. Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. Nature *555*, 190–196.
20. Damgaard, P.B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J.V., Pedersen, M.W., Goldberg, A., Usmanova, E., et al. (2018). 137 ancient human genomes from across the Eurasian steppes. Nature *557*, 369–374.
21. Gnecchi-Ruscone, G.A., Khussainova, E., Kahbatkyzy, N., Musralina, L., Spyrou, M.A., Bianco, R.A., Radzeviciute, R., Martins, N.F.G., Freund, C., Iksan, O., et al. (2021). Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. Sci. Adv. *7*, 4414–4440.
22. Jeong, C., Wang, K., Wilkin, S., Taylor, W.T.T., Miller, B.K., Bemmann, J.H., Stahl, R., Chiovelli, C., Knolle, F., Ulziibayar, S., et al. (2020). A dynamic 6,000-year genetic history of Eurasia's eastern steppe. Cell *183*, 890–904.e29.
23. Wang, C.C., Yeh, H.Y., Popov, A.N., Zhang, H.Q., Matsumura, H., Sirak, K., Cheronet, O., Kovalev, A., Rohland, N., Kim, A.M., et al. (2021). Genomic insights into the formation of human populations in East Asia. Nature *591*, 413–419.
24. Ning, C., Li, T., Wang, K., Zhang, F., Li, T., Wu, X., Gao, S., Zhang, Q., Zhang, H., Hudson, M.J., et al. (2020). Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat. Commun. *11*, 1–9.
25. Järve, M., Saag, L., Scheib, C.L., Pathak, A.K., Montinaro, F., Pagani, L., Flores, R., Guellil, M., Saag, L., Tambets, K., et al. (2019). Shifts in the genetic landscape of the western Eurasian steppe associated with the beginning and end of the Scythian dominance. Curr. Biol. *29*, 2430–2441.e10.
26. Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., et al. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat. Commun. *71*, 1–9.
27. Balogh, C. (2019). Új szempont a kora Avar hatalmi központ továbbélésének kérdéséhez – az Avar fegyveres réteg temetkezései a Duna–Tisza közén — new light on the possible survival of the early Avar power centre – burials of Avar warriors int he Danube–Tisza interfluve. In Hatalmi Központok Az Avar Kaganátusban – Power Centres of the Avar Khaganate, C. Balogh, J. Szentpétery, and E. Wicker, eds. (Katona József Múzeum Kecskemét), pp. 115–138.
28. De Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J.V., Kroonen, G., Peyrot, M., Barjamovic, G., Rasmussen, S., Zacho, C., Baimukhanov, N., et al. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science *360*, eaar7711.

29. Sikora, M., Pitulko, V.V., Sousa, V.C., Allentoft, M.E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., et al. (2019). The population history of northeastern Siberia since the Pleistocene. Nature *570*, 182–188.

30. Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I., Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South and Central Asia. Science *365*, eaat7487.

31. Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., et al. (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature *551*, 368–372.

32. Keyser, C., Zvénigorosky, V., Gonzalez, A., Fausser, J.L., Jagorel, F., Gérard, P., Tsagaan, T., Duchesne, S., Crubézy, E., and Ludes, B. (2021). Genetic evidence suggests a sense of family, parity and conquest in the Xiongnu Iron Age nomads of Mongolia. Hum. Genet. *140*, 349–359.

33. Ilumäe, A.M., Reidla, M., Chukhryaeva, M., Järve, M., Post, H., Karmin, M., Saag, L., Agdzhoyan, A., Kushniarevich, A., Litvinov, S., et al. (2016). Human Y chromosome haplogroup N: a non-trivial time-resolved phylogeography that cuts across language families. Am. J. Hum. Genet. *99*, 163–173.

34. Neparáczki, E., Maróti, Z., Kalmár, T., Maár, K., Nagy, I., Latinovics, D., Kustár, Á., Pálfi, G., Molnár, E., Marcsik, A., et al. (2019). Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin. Sci. Rep. *9*, 16569.

35. Csáky, V., Gerber, D., Koncz, I., Csiky, G., Mende, B.G., Szeifert, B., Egyed, B., Pamjav, H., Marcsik, A., Molnár, E., et al. (2020). Genetic insights into the social organisation of the Avar period elite in the 7th century AD Carpathian Basin. Sci. Rep. *10*, 1–14.

36. Cruciani, F., La Fratta, R., Trombetta, B., Santolamazza, P., Sellitto, D., Colomb, E.B., Dugoujon, J.M., Crivellaro, F., Benincasa, T., Pascone, R., et al. (2007). Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol. Biol. Evol. *24*, 1300–1311.

37. Post, H., Németh, E., Klima, L., Flores, R., Fehér, T., Türk, A., Székely, G., Sahakyan, H., Mondal, M., Montinaro, F., et al. (2019). Y-chromosomal connection between Hungarians and geographically distant populations of the Ural Mountain region and West Siberia. Sci. Rep. *91*, 1–10.

38. Botalov, S.G., and Gutsalov, S. (2000). Hunno-Sarmatians of the Ural-Kazakh Steppes (Рифей).

39. Kim, H.J. (2013). The Huns, Rome and the Birth of Europe (Cambridge University Press).

40. De La Vaissiére, É. (2014). The steppe world and the rise of the Huns. In Cambridge Companion to Age Attila, M. Maas, ed. (Cambridge University Press), pp. 175–192.

41. Li, J., Zhang, Y., Zhao, Y., Chen, Y., Ochir, A., Sarenbilige, Z., Zhu, H., and Zhou, H. (2018). The genome of an ancient Rouran individual reveals an important paternal lineage in the Donghu population. Am. J. Phys. Anthropol. *166*, 895–905.

42. Neparáczki, E., Maróti, Z., Kalmár, T., Kocsy, K., Maár, K., Bihari, P., Nagy, I., Fóthi, E., Pap, I., Kustár, Á., et al. (2018). Mitogenomic data indicate admixture components of Central-Inner Asian and Srubnaya origin in the conquering Hungarians. PLoS One *13*, e0205920.

43. Abondolo, D.M. (1998). The Uralic Languages (Routledge).

44. Istvánovits, E., and Kulcsár, V. (2017). Sarmatians: History and Archaeology of a Forgotten People (Römisch-Germanisches Zentralmuseum).

45. Csáky, V., Gerber, D., Szeifert, B., Egyed, B., Stégmár, B., Botalov, S.G., Grudochko, I.V., Matveeva, N.P., Zelenkov, A.S., Sleptsova, A.V., et al. (2020). Early medieval genetic data from Ural region evaluated in the light of archaeological evidence of ancient Hungarians. Sci. Rep. *10*, 1–14.

46. Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A.M., Rootsi, S., Honkola, T., Vesakoski, O., Atkinson, Q., Skoglund, P., Kushniarevich, A., et al. (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biol. *19*, 139.

47. Reich Lab, David (2020). Allen ancient DNA resource. https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data.

48. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12.

49. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

50. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

51. Broad Institute (2016). Picard tools. https://broadinstitute.github.io/picard/.

52. Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., and Wegmann, D. (2017). ATLAS: analysis tools for low-depth and ancient samples. Preprint at bioRxiv. https://doi.org/10.1101/105346.

53. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58–W63.

54. Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. Mol. Biol. Evol. *35*, 1291–1294.

55. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

56. Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. Genetics *210*, 719–731.

57. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

58. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

59. R Development Core Team. (2015). R: a language and environment for statistical computing (R Foundation for Statistical Computing).

60. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

61. Kuhn, J.M.M., Jakobsson, M., and Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. PLoS One *13*, e0195491.

62. Molnár, M., Janovics, R., Major, I., Orsovszki, J., Gönczi, R., Veres, M., Leonard, A.G., Castle, S.M., Lange, T.E., Wacker, L., et al. (2013). Status report of the new AMS 14C sample preparation Lab of the Hertelendi Laboratory of Environmental Studies (Debrecen, Hungary). Radiocarbon *55*, 665–676.

63. Reimer, P.J., Austin, W.E.N., Bard, E., Bayliss, A., Blackwell, P.G., Bronk Ramsey, C., Butzin, M., Cheng, H., Edwards, R.L., Friedrich, M., et al. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). Radiocarbon *62*, 725–757.

64. Maár, K., Varga, G.I.B., Kovács, B., Schütz, O., Maróti, Z., Kalmár, T., Nyerki, E., Nagy, I., Latinovics, D., Tihanyi, B., et al. (2021). Maternal lineages from 10–11th century commoner cemeteries of the Carpathian Basin. Genes *12*, 460.

65. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*. pdb.prot5448.

66. Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

67. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624.

68. Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. J. Archaeol. Sci. *40*, 4477–4482.

69. Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyzy, N., Flegontov, P., Zaporozhchenko, V., Immel, A., Wang, C.C., Ixan, O., Khussainova, E., et al. (2019). The genetic history of admixture across inner Eurasia. Nat. Ecol. Evol. *3*, 966–976.

70. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., and Kozlowski, K. (2018). "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. https://svn.r-project.org/R-packages/trunk/cluster/.

71. Unterländer, M., Palstra, F., Lazaridis, I., Pilipenko, A., Hofmanová, Z., Groß, M., Sell, C., Blöcher, J., Kirsanow, K., Rohland, N., et al. (2017). Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. Nat. Commun. *8*, 14615.

72. Krzewińska, M., Kılınç, G.M., Juras, A., Koptekin, D., Chyleński, M., Nikitin, A.G., Shcherbakov, N., Shuteleva, I., Leonova, T., Kraeva, L., et al. (2018). Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. Sci. Adv. *4*, eaat4457.

73. Harney, É., Patterson, N., Reich, D., and Wakeley, J. (2021). Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics *217*, iyaa045.

74. Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W., Orlando, L., Metspalu, E., et al. (2014). Upper palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature *505*, 87–91.

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human archaeological remains	This paper	N/A
Critical commercial assays		
MinElute PCR Purification Kit	QIAGEN	Cat No./ID: 28006
Accuprime Pfx Supermix	ThermoFisher Scientific	Cat. No: 12344040
Deposited data		
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Modern comparison dataset	Allen Ancient DNA Resource (Version v42.4)[47]	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
Ancient comparison dataset	Allen Ancient DNA Resource (Version v42.4)[47]	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
Ancient comparison dataset	Gnecchi-Ruscone et al.[21]	https://www.ebi.ac.uk/ena/browser/view/PRJEB42930
Ancient comparison dataset	Jeong et al.[22]	https://www.ebi.ac.uk/ena/browser/view/PRJEB35748
Ancient comparison dataset	Wang et al.[23]	https://www.ebi.ac.uk/ena/browser/view/PRJEB42781
Ancient comparison dataset	Ning et al.[24]	https://www.ebi.ac.uk/ena/browser/view/PRJEB36297
Newly published ancient genomes	This paper	https://www.ebi.ac.uk/ena/browser/view/PRJEB499771
Oligonucleotides		
Illumina specific adapters	Custom synthetized	https://www.sigmaaldrich.com/HU/en/product/sigma/oligo?lang=en&region=US&gclid=CjwKCAiAgvKQBhBbEiwAaPQw3FDDFnRPc3WV75qapsXvcTxxzBXy48atqyb6Xi5f_8e6Df2EJI0NNhoCmzIQAvD_BwE
Software and algorithms		
Cutadapt	Martin[48]	https://cutadapt.readthedocs.io/en/stable/#
Burrow-Wheels-Aligner	Li and Durbin[49]	http://bio-bwa.sourceforge.net/
samtools	Li et al.[50]	http://www.htslib.org/
PICARD tools	Broad Institute[51]	https://github.com/broadinstitute/picard
ATLAS software package	Link et al.[52]	https://bitbucket.org/wegmannlab/atlas/wiki/Home
MapDamage 2.0	Jónsson et al.[11]	https://ginolhac.github.io/mapDamage/
Schmutzi software package	Renaud et al.[12]	https://github.com/grenaud/schmutzi
ANGSD software package	Korneliussen et al.[13]	https://github.com/ANGSD/angsd
HaploGrep 2	Weissensteiner et al.[53]	https://haplogrep.i-med.ac.at/category/haplogrep2/
Yleaf software tool	Ralf et al.[54]	https://github.com/genid/Yleaf
mosdepth software	Pedersen and Quinlan[55]	https://github.com/brentp/mosdepth
PCAangsd software	Meisner and Albrechtsen[56]	https://github.com/Rosemeis/pcangsd
RcppCNPy R package	N/A	https://rdocumentation.org/packages/RcppCNPy/versions/0.2.10
smartpca	Patterson et al.[57]	https://github.com/chrchang/eigensoft/blob/master/POPGEN/README

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ADMIXTURE software	Alexander et al.[58]	https://dalexander.github.io/admixture/
R 3.6.3	R core development team[59]	https://cran.r-project.org/bin/windows/base/old/3.6.3/
ADMIXTOOLS software package	Patterson et al.[60]	https://github.com/DReichLab/AdmixTools
DATES algorithm	Narasimhan et al.[30]	https://github.com/priyamoorjani/DATES/tree/v753
READ algorithm	Kuhn et al.[61]	https://doi.org/10.1371/journal.pone.0195491.s007

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tibor Török (torokt@bio.u-szeged.hu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Aligned sequence data have been deposited at European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB49971 and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ancient samples

We present genome-wide data of 265 ancient individuals from the Migration Period of the Carpathian Basin between the 5th and 11th centuries and 3 individuals from the 9-10th century Caucasus with archaeological affinity to the Conquering Hungarians. We also sequenced 3 Neolithic individuals from the Carpathian Basin. The 265 Migration Period samples represent the following time range: 9 samples from the Hun Period (5th century), 40 from the early Avar Period (7th century), 33 from the middle Avar Period (8th century), 70 from the late Avar Period (8-9th century), 48 from 10th century Conquering Hungarian elite cemeteries, 65 from commoner cemeteries of the Hungarian conquer-early Árpádian Period (10-11th centuries).

The majority of samples were collected from the Great Hungarian Plain (Alföld), the westernmost extension of the Eurasian steppe, which provided favorable ground for the arriving waves of nomadic groups. Cemeteries and individual samples were chosen on archaeological, anthropological and regional basis. We made an effort to assemble a sample collection from each period representing a) all possible geographical sub-regions, b) all archeological types, c) all anthropological types. From large cemeteries we selected individuals with the same criterions and possibly from all part of the cemetery with the following bias: We preferably choose samples with good bone preservation, archaeologically well described ones (with grave goods) and males (for Y-chromosomal data). Nevertheless, we took care to also include females and samples without grave goods, though these are definitely underrepresented in our collection.

The human bone material used for ancient DNA analysis in this study were obtained from anthropological collections or museums, with the permission of the custodians in each case. In addition, we also contacted the archaeologists who excavated and described the samples, as well as the anthropologists who published anthropological details. In most cases these experts became co-authors of the paper, who provided the archaeological background, which is detailed in Methods S1.

METHOD DETAILS

Accelerator mass spectrometry radiocarbon dating

Here we report 73 radiocarbon dates, of which 50 are first reported in this paper. The sampled bone fragments were measured by accelerator mass spectrometry (AMS) in the AMS laboratory of the Institute for Nuclear Research, Hungarian Academy of Sciences, Debrecen, Hungary. Technical details concerning the sample preparation and measurement are given in Kuhn et al.[62] Several

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

radiocarbon measurements were done in the Radiocarbon AMS facility of the Center for Applied Isotope Studies, University of Georgia (n = 6;), technical details concerning the sample preparation and measurement are available here: https://cais.uga.edu/facilities/radiocarbon-ams-facility/). The conventional radiocarbon data were calibrated with the OxCal 4.4 software (https://c14.arch.ox.ac.uk/oxcal/OxCal.html, date of calibration: 4th of August 2021) with IntCal 20 settings.[63] Besides, we collected all previously published radiocarbon data related to the samples of our study.

Ancient DNA laboratory work

All pre-PCR steps were carried out in the dedicated ancient DNA facilities of the Department of Genetics, University of Szeged and Department of Archaeogenetics, Institute of Hungarian Research, Hungary. Mitogenome or Y-chromosome data had been published from many of the samples used in this study,[42,64] and we sequenced whole genomes from the same libraries, whose preparations had been described in the above papers.

For the rest of the samples we used the following modified protocol. DNA was extracted from bone powder collected from petrous bone or tooth cementum. 100 mg bone powder was predigested in 3 ml 0.5 M EDTA 100 μg/ml Proteinase K for 30 minutes at 48°C, to increase the proportion of endogenous DNA. After pelleting, the powder was solubilized for 72 hours at 48°C, in extraction buffer containing 0.45 M EDTA, 250 μg/ml Proteinase K and 1% Triton X-100. Then 12 ml binding buffer was added to the extract, containing 5 M GuHCl, 90 mM NaOAc, 40% isopropanol and 0.05% Tween-20, and DNA was purified on Qiagen MinElute columns.

NGS library construction

Partial UDG treated libraries were prepared as described in Neparáczkiet al.[42] In short we used the double stranded library protocol of Meyer and Kircher[65] with double indexing,[66] except that all purifications were done with MinElute columns. We also applied partial UDG treatment of Rohland et al.,[67] but decreased the recommended USER and UGI concentrations to half (0.03 U/μL) and at the same time increased the incubation time from 30 to 40 minutes. The reaction was incubated at 37°C for 40 minutes in PCR machine, with 40°C lid temperature.

Then 1,8 μL UGI (Uracil Glycosylase Inhibitor, 2U/μL NEB) was added to the reaction, which was further incubated at 37°C for 40 minutes. Next blunt-end repair was done by adding 3 μL T4 polynucleotide kinase (10 U/μL) and 1,2 μL T4 DNA polymerase-t (5 U/μL) to each reaction followed by incubation in PCR machine at 25°C for 15 minutes, and another incubation at 12°C for 5 minutes and cooling to 4°C. After adding 350 μl MinElute PB buffer (QIAGEN) to the reaction, it was purified on MinElute columns, and the DNA was eluted in 20 μl EB prewarmed to 55°C. Adapter ligation and adapter fill-in was done as in Meyer and Kircher.[65]

The library preamplification step was omitted, and libraries were directly double indexed in one PCR-step after the adapter fill with Accuprime Pfx Supermix, containing 10mg/ml BSA and 200nM indexing P5 and P7 primers, in the following cycles: 95°C 5 minutes, 12 times 95°C 15 sec, 60°C 30 sec and 68°C 3 sec, followed by 5 minute extension at 68°C. The indexed libraries were purified on MinElute columns and eluted in 20μL EB buffer (Qiagen).

DNA sequencing

Quantity measurements of the DNA extracts and libraries were performed with the Qubit fluorometric quantification system. The library fragment distribution was checked on TapeStation 2200 (Agilent). We estimated the endogenous human DNA content of each library with low coverage shotgun sequencing generated on iSeq 100 (Illumina) platform. Whole genome sequencing was performed on HiSeqX or NovaSeq 6000 Systems (Illumina) using paired-end sequencing method (2x150bp) following the manufacturer's recommendations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Bioinformatical processing

Sequencing adapters were trimmed with the Cutadapt software[48] [https://doi.org/10.14806/ej.17.1.200] and sequences shorter than 25 nucleotides were removed. The raw reads were aligned to the GRCh37 (hs37d5) reference genome by Burrow-Wheels-Aligner (v 0.7.17),[49] using the MEM command with reseeding disabled.

The current high throughput aligners were designed to work on modern uncontaminated DNA, thus they will find the best fit for all DNA fragments that partially match the reference genome. Since the human genome is very complex, the majority of the exogenous, non human reads will be aligned to it with partial fits. The mapping quality (MAPQ) is a metric that describes only the alignment quality of the matching part of the sequence, regardless of any overhanging (soft, hard clipped) parts. As a consequence, non human reads will receive high MAPQ values, when the length of the aligned part is above 29-30 bases and the mismatches are not excessive, despite of the overall bad matching due to the overhanging ends. For this reason, MAPQ based filtering of exogenous DNA may lead to excessive bias. To avoid this bias we filtered our alignments, based on the proportion of the full length reads matching the human reference genome. To remove exogenous DNA only the primary alignments with >= 90% identity to reference were considered in all downstream analysis. This way we could eliminate most exogenous DNA with partial matching, while still keeping human reads with PMD and real SNP variations. Although this method also excludes reads corresponding to true human structural variants, but these are negligible. This strategy results in much more reliable variant calls compared to the simple MAPQ based filtration in case of ancient DNA.

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

From paired-end sequencing data we only kept properly paired primary alignments. Sequences from different lanes with their unique read groups were merged by samtools.[50] PICARD tools[51] were used to mark duplicates. In case of paired-end reads we used the ATLAS software package[52] mergeReads task with the options "updateQuality mergingMethod=keepRandomRead" to randomly exclude overlapping portions of paired-end reads, to mitigate potential random pseudo haploidization bias.

Quality assessment of ancient sequences

Ancient DNA damage patterns were assessed using MapDamage 2.0[11] (Data S1B), and read quality scores were modified with the Rescale option to account for post-mortem damage. Mitochondrial contamination was estimated with the Schmutzi software package.[12] Contamination for the male samples was also assessed by the ANGSD X chromosome contamination method,[13] with the "*-r X:5000000-154900000 -doCounts 1 -iCounts 1 -minMapQ 30 -minQ 20 -setMinDepth 2*" options.

Contamination estimations are detailed in Data S1A. The estimated contaminations ranged from 0 to 0.42 with an average of 0.017 in the case of Schmutzi and from 0 to 0.13 with an average of 0.014 in the case of ANGSD. We detedcted negligible contamination in all but 7 samples. Schmutzi estimated significant contamination in 6 samples, but in 3 male samples of these ANGSD measured low X-contamination. These contradictory estimates may be explained by the UDG overtreatment of these libraries, as Schmutzi reckons Uracil free molecules as derived from contamination. In another male sample ANGSD measured significant contamination, which was not confirmed by Schmutzi.

Uniparental haplogroup assignment

Mitochondrial haplogroup determination was performed with the HaploGrep 2 (version 2.1.25) software,[53] using the consensus endogen fasta files resulting from the Schmutzi Bayesian algorithm. The Y haplogroup assessment was performed with the Yleaf software tool,[54] updated with the ISOGG2020 Y tree dataset.

Genetic sex determination

Biological sex was assessed with the method described in Skoglund et al.[68] Fragment length of paired-end data and average genome coverages (all, X, Y, mitochondrial) was assessed by the ATLAS software package[52] using the BAMDiagnostics task. Detailed coverage distribution of autosomal, X, Y, mitochondrial chromosomes was calculated by the mosdepth software.[55]

Estimation of genetic relatedness

Presence of close relatives in the dataset interferes with unsupervised ADMIXTURE and population genetic analysis, therefore we identified close kins and just one of them was left in the dataset (Table S1). We performed kinship analysis using the 1240K dataset and the PCAngsd software (version 0.931)[56] from the ANGSD package with the "-inbreed 1 -kinship" options. We used the R (version 4.1.2); the RcppCNPy R package (version 0.2.10) to import the Numpy output files of PCAngsd. Though the PCAangsd software in the ANGSD package has not been a standard method to infer kinship, nevertheless it outperforms all other methods, as we have shown it in our manuscript.[14] Close kinship relations identified with PCAangsd were also confirmed with the READ (Relationship Estimation from Ancient DNA) method[61] (Table S1).

Population genetic analysis

The newly sequenced genomes were merged and co-analyzed with 2364 ancient (Data S3) and 1397 modern Eurasian genomes (Table S2), most of which were downloaded from the Allen Ancient DNA Resource (Version v42.4).[47] We also downloaded the Human Origins dataset (HO, 600K SNP-s) and/or the 1240K datasets published in Gnecchi-Rusconeet et al.,[21] Jeong et al.,[22] and Wang et al.[23] As the HO SNPs are fully contained in the larger 1240K set we filtered out the HO data when only the 1240K dataset was published. In case the Reich data set contained preprint data of the same individual published later, we always used the published genotypes. Since some dataset contained diploid, and mixed call variants we performed random pseudo haploidization of all data prior to downstream analysis. To avoid bias of PE sequencing in random haploid calling, we used the ATLAS software "mergeReads" task with parameters "updateQuality mergingMethod=keepRandom" to set BASE QUALITY to 0 of overlapping portion of a random mate of PE reads. Then used ANGSD (software package version: 0.931-10-g09a0fc5)[13] for random allele calling with the options "--doHaploCall 1 -doCounts 1", with the default BAM quality filters (including -minQ 20; Discard bases with base quality below) on these mate pair merged BAMs. Since every DNA fragment is represented by only one copy (with BASE QUALITY > 0), this method is equal to random allele calling.

Most of the analysis was done with the HO dataset, as most modern genomes are confined to this dataset, however, we run some of the f-statistics with the 1240K data if these were available.

Principal Component Analysis (PCA)

We used the modern Eurasian genome data published in Jeong et al.,[69] confined to the HO dataset, to draw a modern PCA background on which ancient samples could be projected. However, in order to obtain the best separation of our samples in the PC1-PC2 dimensions, South-East Asian and Near Eastern populations were left out, and generally just 10 individuals were selected from each of the remaining populations, leaving 1397 modern individuals from 179 modern populations in the analysis (Table S2).

PCA Eigen vectors were calculated from 1397 pseudo-haploidized modern genomes with smartpca (EIGENSOFT version 7.2.1).[57] Before projecting pseudo-haploidized ancient genomes, we excluded all relatives, and used the individuals with best genome

coverage. All ancient genomes were projected on the modern background with the "lsqproject: YES and inbreed: YES" options. Since the ancient samples were projected, we used a more relaxed genotying threshold (>50k genotyped markers) to exclude samples only where the results could be questionable due to the low coverage.

Unsupervised Admixture

We carried out unsupervised admixture with 3277 genomes including 1010 modern and 2027 ancient published genomes plus 240 ones from present study, excluding all published relatives from each dataset. For this analysis we used the autosomal variants of the HO dataset as many relevant modern populations are missing from the 1240K set. We set strict criteria for the selection of individual samples to minimize bias and maximize the information content of our dataset. We excluded all samples with QUESTIONABLE flag or Ignore tag based on the annotation file of the dataset to remove possibly contaminated samples and population outliers. To compose a balanced high quality dataset furthermore we restricted the selection to maximum 10 individuals per populations and excluded all poorly genotyped samples (<150K genotyped markers).

To prepare the final marker set for ADMIXTURE analysis we removed variants with very low frequency (MAF <0.005) leaving 471,625 autosomal variants. We pruned 116,237 variants in linkage disequilibrium using PLINK with the options "*–indep-pairwise 200 10 0.25*" leaving a final 355388 markers for the 3277 individuals. The total genotyping rate of this high quality dataset was 0.811831. We performed the unsupervised ADMIXTURE analysis for K=3-12 in 30 parallel runs with the ADMIXTURE software (version 1.3.0)[58] and selected the lowest cross validation error model (K=7) with the highest log-likelihood run for visualization.

Hierarchical Ward clustering

As qpAdm is sensitive both to genome components and their proportions present in the source and Test populations, it works best if genetically homogenous populations are used in the analysis. Similar Admixture composition and small PC1-PC2 distances may indicate population relations derived from shared genetic ancestry, but these data do not have enough resolution to identify homogenous genome subsets which can be merged into populations. Though the first two PCA Eigenvalues capture the highest levels of variation in the data, in our analysis subsequent Eigenvalues had comparable magnitude to the second one, indicating that lower Eigenvectors still harbored significant additional genetic information, which must be taken into account. Individuals with most similar genomes should be closest in the entire PC space, which can be measured mathematically with the combined Euclidean distances along multiple PC-s, therefore we introduced a novel approach, called PC50 clustering, to identify the most similar genomes. We clustered our genomes according to the pairwise weighed Euclidean distances of the first 50 PCA Eigenvalues and Eigenvectors (PC50 distances), where distances were calculated as follows: $\sqrt[2]{W1 * P1^2 + W2 * P2^2 + ...Wn * Pn^2}$, where W=Eigenvalue and P=Eigenvector.

As our PCA was confined to the HO dataset, this analysis was also done with the HO dataset. According to the Tracy-Widom statistics the first 110 PC vectors contained significant variation in our data set while the first 50 PCs accounted for 61.26% of the total variation. For Hierarchical Clustering we applied (ward.D2)[70] implemented in R 3.6.3.[59]

Published genomes are commonly grouped according to their major genetic ancestry components, but members of the same group often vary in the proportion of these. In order to obtain genetically homogenous source populations for qpAdm, we also regrouped relevant published samples with the same method. For this end we projected 2364 relevant published Eurasian ancient samples on the same modern PCA background described above and performed the same PC50 clustering.

As clustering splits down to the individual level, for obtaining groups it makes sense to cut the smaller branches. We cut the tree arbitrarily at a relatively great depth, considering the 50 deepest branches, where published "homogenous" groups were already divided into subclusters. If samples from published genetic groups fell into different clusters, we subdivided the original group according to the clusters. To create as homogenous groups as possible, in some cases we distinguished subgroups even within the same clusters, if they were separated by relatively large PC1 or PC2 distances. This did not bias the results, as theoretically subgrouping can be performed down to the individual level, if qpAdm could handle thousands of sources. This way we regrouped Late Bronze Age, Iron Age and Medieval samples published in Allentoft et al.,[18] Damgaard et al.,[20] Gnecchi-Ruscone et al.,[21] Jeong et al.,[22] Wang et al.,[23] Järve et al.,[25] Unterländer et al.,[71] and Krzewińska et al.[72] (Data S3). We generally merged all samples within each group, but when many samples were available from a certain group, we left out the ones below 50 thousand SNPs corresponding to the HO dataset, and merged the remaining genomes with PASS quality assessment (described in the Allen Ancient DNA Resource datatable).

The hierarchical clusters based on the first PC50 distances are shown in Data S3. PC50 clustering identified the Avar_Asia_Core and Conq_Asia_Core populations as homogenous groups, and even Core1 and Core2 subgroups are separated to finer branches. Though representatives of the Eur_Core groups were identified in a preliminary distal qpAdm analysis, all five Eur_Core groups fall on separate clusters, with the exception of two samples (TMH-388 and TMH-199), which fall into the neighboring cluster. We found that the sensitivity of PC-50 clustering for grouping genomes according to similarity is comparable to the resolution of individual qpAdm.

Admixture modeling using qpAdm

We used qpAdm[73] from the ADMIXTOOLS software package[60] for modelling our genomes as admixtures of 2 or 3 source populations and estimating ancestry proportions. The qpAdm analysis was done with the HO dataset, as in many cases suitable Right or Left populations were only available in this dataset. We set the details:YES parameter to evaluate Z-scores for the goodness of fit of the model (estimated with a Block Jackknife).

As we tested a large number of source populations, testing every possible combination of sources (Left populations) and outgroups (Right populations) was impossible. Instead we ran the analysis just with source combinations of 2 and 3 (rank 2 and 3). As qpWave is integrated in qpAdm, the nested p values in the log files indicate the optimal rank of the model. This means that if p value for the nested model is above 0.05, the Rank-1 model should be considered.[73]

To reveal past population history of the Test populations from different time periods, we run two separate qpAdm analysis. In the so-called "distal analysis" pre-Bronze Age and Bronze Age populations were included as sources. Next we run a so-called "proximodistal analysis", in which just the most relevant distal sources were included in the Left population list, supplemented with a large number of post-Bronze Age populations. In latter runs potentially more relevant proximal sources competed with distant Bronze Age sources, and plausible models with distal sources indicated the lack of relevant proximal sources. In some cases, we used modern populations as sources, because the more relevant ancient sources were seemingly unavailable.

We performed a preliminary distal qpAdm analysis for each Eur-cline sample (data not shown), which supported PC50 clustering. As a result, we established five Eur-cline groups representing five different genome compositions. Next, we selected a few representatives from each group, with the most equivalent qpAdm models, and merged these genomes under the name of Eur_Core1 to Eur_Core5 respectively.

qpAdm analysis strategy

We must note that our Middle Age samples are highly admixed (Figure 2B; Data S5), likewise most of the Iron Age populations from which they possibly derived contain similar genomic components obtained from various admixture histories of related populations. In such cases qpAdm expectedly results in multiple feasible alternative models, most of which is very difficult to exclude. As we wished to include all relevant potential source populations in the analysis, despite their clearly similar genome histories, excluding suboptimal models was the largest challenge of the analysis. Thus, we set out to optimize our qpAdm strategy in multiple steps. Since our Test populations had largely different genome structures, we optimized the Right populations for each analysis.

In the first step we performed an iterative optimization of our Right populations, to exclude redundant, non discriminative Right populations, based on the log analysis of qpAdm for each run. Based on PCA, outgroup f3-statistics and unsupervised ADMIXTURE data, we began by assembling a large set of plausible pre-Bronze Age and early Bronze Age Right populations containing different ancestry components present in our Test populations. After several initial runs with a diverse set of Right populations, we collected the models with at least one significant Z-score from the detailed "gendstat" lines of the log files of all qpAdm models. We also counted how many models we would not reject if we excluded the f4-statistics with significant Z-scores of a given Right population. Based on this information we could test if all the Right populations were needed to reject the models. Then we repeated the qpAdm analysis with the optimally reduced Right populations until most Right populations were needed to reject models. As an important exception, we always kept Right populations that measured the main genetic components of our test population. Since all our Test populations were Eurasian samples we used a suitable outgroup, (Ethiopia_4500BP_published.SG) as Right Base throughout our analyses. As a result of Right optimization, a unique Right population set was used in each analysis, which are listed in details in Data S4, S7, S8, and S9, where qpAdm results are presented.

In order to further exclude suboptimal models, finally we applied the "model-competition" approach described in Narasimhan et al.[30] the following way: From the Left populations present in the plausible models we moved one at a time to the Right set and rerun qpAdm for each model. This was repeated with each Left population which appeared in any of the plausible models. As the best sources have the highest shared drift with the Test population, including these in the Right list is expected to exclude all models with similar suboptimal sources. This way we were able to filter out most suboptimal models and identify the most plausible ones, which were not excluded by any of the Right combinations. As each model-competition run gave a different p value with different standard deviations, we deemed it more informative to provide the maximum, minimum and average p values for the best final models, instead of the p values and standard deviations of the original models.

In some cases, model-competition excluded all 3 source models, indicating that additional sources are required for optimal modeling. However, running 4 source models with numerous sources is not feasible, thus we run 4 source models with a reduced set of Left populations. To select the best candidate subset of sources we evaluated the gendstat data from the log files of the 3 source models and identified populations which excluded the best models. Latter populations presumably had additional shared drift with the Test, not shared by any of the sources. Thus, in the rank 4 models we included just the best sources from the rank 3 models, plus their excluding populations.

Many of our samples were part of genetic clines between East and West Eurasia. In order to reveal the genetic ancestry of individual samples within genetic clines the identified genetic groups at the eastern and western extremes were also added to the Left-populations as sources. Many of the samples within clines could be modelled as simple two-way admixture of these two populations, or three-way admixtures with a third source. The remaining individuals were considered genetic outliers, which were modelled from different sources. As three source modeling of large number of samples from large number of sources is not feasible, we decreased computation time by decreasing the sources (removing the most redundant ones), and running constrained 3 source models by fixing one source at a time, and swapping only the rest of the Left populations. We repeated these constrained runs in 3 combinations, once fixing Eur_Core, second fixing Conq_Asia_Core and third fixing Avar_Asia_Core. Finally, we compared the 3 different models for each sample and selected the ones with best p values. Obtaining the same high p value models from two different constrained runs, one with Eur_Core fixed, the other with Asia_Core fixed, rendered it very likely that the optimal model was found. In cases when Eur_Core fixed, and Asia_Core fixed runs gave different models, but all Asia_Core fixed models indicated the presence of

Please cite this article in press as: Maróti et al., The genetic origin of Huns, Avars, and conquering Hungarians, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.04.093

Eur_Core, we accepted the best Eur_Core fixed models. Finally, we also tested whether samples at the Asian side of the cline require entirely different sources by running unconstrained 3 source models for these few samples.

We noticed that model-competition cannot exclude alternative qpAdm models with similar minor components, therefore a given source is best identified from samples which carry it in large fraction. We had several within cemetery PCA clines with varying fractions of the same components, as nearly each member could be modelled from individuals at the extremes of the cemetery cline (Data S8E). In these cases the best Asian source could be accurately identified from the unconstrained 3 source models, which also applies to all members of the cemetery cline.

We present qpAdm results in the following format in Data S4, S7, S8, and S9: All passing models are shown which received significant p values after model competition runs, or the best models if none of the models passed (unless indicated otherwise). Column LEFT lists the source populations which were tested in each 2-way or 3-way combinations for the given Test populations. Column RIGHT lists the reference populations used in the given qpAdm. The "total" column indicates the number of competition runs repeated for each plausible model. Column "valid" shows how many times the given model passed, while "excluded" shows how many times it was excluded. In latter case the excluding extra Right population is shown in the "excluding pop" column. "neg" column indicates the number of models in which any of the sources produced a negative fraction. "nested P" column indicates the number of models with significant (>0.05) nested p values, when the Rank-1 model is more plausible. "max p" indicates the maximum p value obtained from the competition runs, while "max P right pop" shows the extra Right population in this model. "min P" and "min P Right pop" means the same for the lowest p value model, while "avg P" means the average p value of the valid models. Models are arranged according to max p and min p values.

f3-statistics

Outgroup f3-statistics is suitable to measure shared drift between two test populations after their divergence from an outgroup[74] thus providing a similarity measure between populations. We measured the shared drift between the identified homogeneous new genetic groups in our sample set and all published modern and ancient populations, to identify populations with shared evolutionary past. As an outgroup we used African Mbuti genomes and applied ADMIXTOOLS[60] to calculate f3 statistics. This analysis was also done with the HO dataset, and we removed populations below 40 thousand overlapping markers.

Admixture f3-statistics in the form f3(Test; X, Y) can be used to identify potential admixture sources of the Test population,[60] and most negative f3 values indicate the major admixture sources. We used the qp3Pop program of ADMIXTOOLS with the *inbreed: YES* parameter. All population combinations are listed in Data S6C and S6D.

Two dimensional f4-statistics

To measure different levels of bi-directional gene flow into populations with shared genomic history Narasimhan et al.[30] applied a so called ''Pre-Copper Age affinity f4-statistics'', with a 2-dimensional representation of the f4 values from two related statistics. This way populations or individuals with significantly different proportions of ancestry related to the two sources can be visualized. We applied this 2-dimensional f4-statistics to measure gene flow from multiple sources into members of the same population. This way we could explore the fine substructure of populations and identify potential sources of the gene flow. To calculate f4-statistics we used the qpF4ratio from ADMIXTOOLS.[60] We run two-dimensional f4-statistics with the 1240K marker set.

Dating admixture time with DATES

The DATES algorithm[30] was developed to infer the date of admixture, and this software was optimized to work with ancient DNA and single genomes. As qpAdm often revealed that a two- or three-way admixture well explained the genome history of the studied population, we used DATES to determine admixture time. As we used modern populations with DATES, this analysis was also confined to the HO dataset.