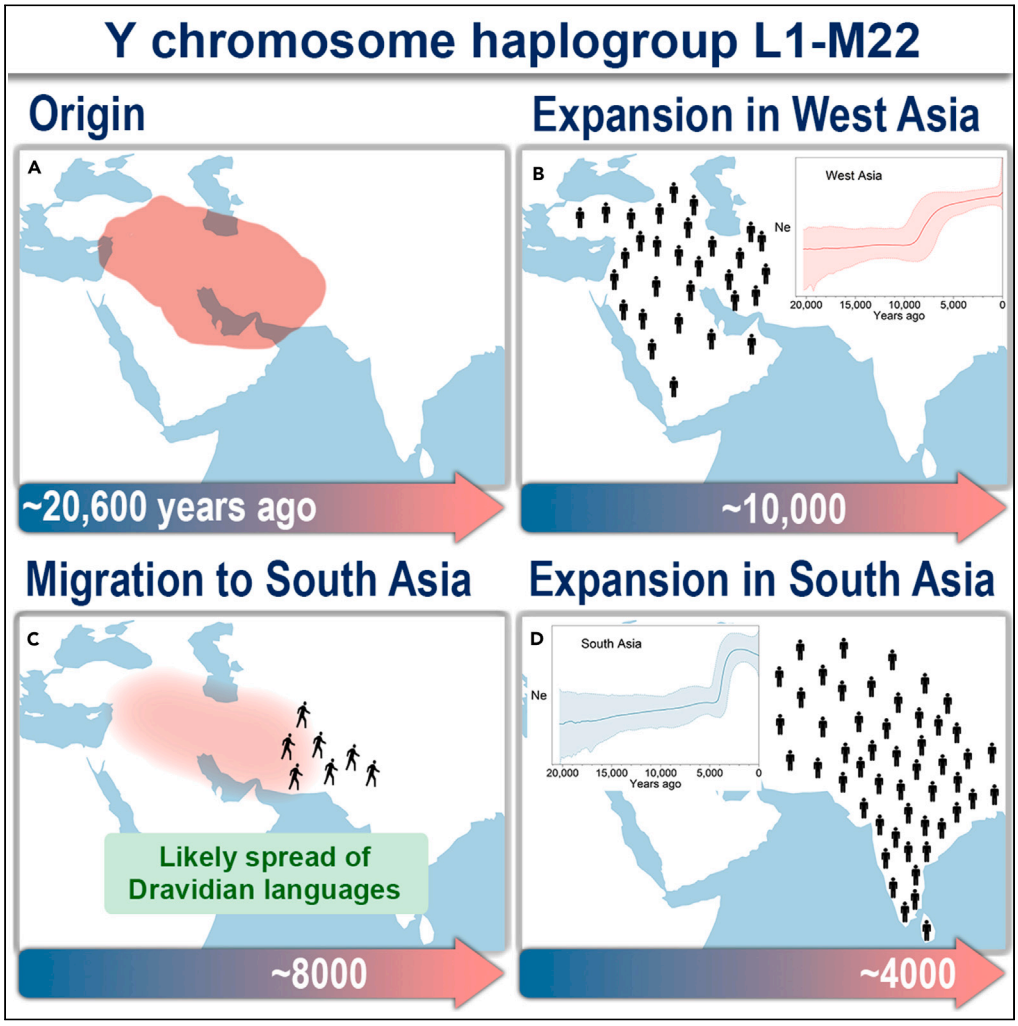
iScience

CellPress
OPEN ACCESS

Article

Human Y chromosome haplogroup L1-M22 traces Neolithic expansion in West Asia and supports the Elamite and Dravidian connection



Ajai Kumar Pathak, Hovann Simonian, Ibrahim Abdel Aziz Ibrahim, ..., Phillip Endicott, Richard Villems, Hovhannes Sahakyan

pathak@ut.ee (A.K.P.)
hovhannes.sahakyan@ut.ee (H.Sa.)

Highlights

The Y chromosome haplogroup L1-M22 originated in West Asia around 20,600 years ago

A group of L1-M22 harboring population expanded with West Asian Neolithic transition

Another one moved to South Asia, likely participating in Dravidian languages' spread

Their descendants expanded in South Asia around 4,000 to 3,000 years ago

Pathak et al., iScience 27, 110016
June 21, 2024
https://doi.org/10.1016/j.isci.2024.110016

iScience

CellPress
OPEN ACCESS

Article

Human Y chromosome haplogroup L1-M22 traces Neolithic expansion in West Asia and supports the Elamite and Dravidian connection

Ajai Kumar Pathak,[1,2,*] Hovann Simonian,[3] Ibrahim Abdel Aziz Ibrahim,[4] Peter Hrechdakian,[3] Doron M. Behar,[1] Qasim Ayub,[5] Pakhrudin Arsanov,[6] Ene Metspalu,[1] Levon Yepiskoposyan,[7] Siiri Rootsi,[1] Phillip Endicott,[1,8,9,10] Richard Villems,[1] and Hovhannes Sahakyan[1,7,11,12,*]

SUMMARY

West and South Asian populations profoundly influenced Eurasian genetic and cultural diversity. We investigate the genetic history of the Y chromosome haplogroup L1-M22, which, while prevalent in these regions, lacks in-depth study. Robust Bayesian analyses of 165 high-coverage Y chromosomes favor a West Asian origin for L1-M22 ~20.6 thousand years ago (kya). Moreover, this haplogroup parallels the genome-wide genetic ancestry of hunter-gatherers from the Iranian Plateau and the Caucasus. We characterized two L1-M22 harboring population groups during the Early Holocene. One expanded with the West Asian Neolithic transition. The other moved to South Asia ~8-6 kya but showed no expansion. This group likely participated in the spread of Dravidian languages. These South Asian L1-M22 lineages expanded ~4-3 kya, coinciding with the Steppe ancestry introduction. Our findings advance the current understanding of Eurasian historical dynamics, emphasizing L1-M22's West Asian origin, associated population movements, and possible linguistic impacts.

INTRODUCTION

Today's world bears profound imprints of past human communities in West and South Asia. After the major out-of-Africa migration, modern humans first appeared in West Asia,[1] then shortly after, in South Asia.[2] The Y chromosome haplogroups J-M304 and G-M201 inform us about the genetic legacy of these early migrations in West Asia, while the H-M69 in South Asia.[3–5] Similarly, mitochondrial DNA haplogroups U, J, and T echo the early exodus in West Asia, while many haplogroups within the macro-haplogroups M, N, and R represent South Asia.[6–10] The genome-wide variation also shows distinct spatiotemporal patterns in these regions.[11–13] After these early events, the regions underwent the Last Glacial Maximum (LGM), a challenging period around 26.5 to 19 kya.[14] The LGM appears to have had a more pronounced impact on West Asian human populations.[7,8,15,16] Unfortunately, no ancient DNA (aDNA) study has been conducted with samples originating from such a deep time of the regions.

The Neolithic demographic transition, a crucial shift in human history,[17] involved the domestication of plants and animals, leading to increased sedentism, population growth, and the development of complex social structures.[18,19] This transition started in the Fertile Crescent in West Asia at ~12 kya, representing the earliest known instance globally.[20–22] By this time, hunter-gatherer populations from present-day Iran and the Caucasus, Anatolia, and the southern Levant accumulated substantial genetic differences and adopted the new lifestyle probably independently.[23–31] Populations from present-day Iran and the Caucasus started mixing with Anatolian populations earlier during the Neolithic period, while gene flows with Levantine populations occurred later.[27,28,31] Agricultural practices were gradually reaching other regions. In Europe and Central Asia, the transmission was driven by population expansions from West Asia.[30,32–35] In South Asia, the nature of the Neolithic transition remains elusive. Uncertainties persist regarding whether it transpired through population expansion from neighboring regions[27,36–41] or proceeded through cultural adoption or local developments without substantial population movement.[42–46] The

[1]Estonian Biocentre, Institute of Genomics, University of Tartu, 51010 Tartu, Estonia
[2]Department of Human Genetics, KU Leuven, 3000 Leuven, Belgium
[3]Armenian DNA Project at Family Tree DNA, Houston, TX 77008, USA
[4]Department of Pharmacology and Toxicology, Faculty of Medicine, Umm Al-Qura University, Makkah 21955, Saudi Arabia
[5]Monash University Malaysia Genomics Platform, School of Science, Monash University, Bandar Sunway, Selangor Darul Ehsan 47500, Malaysia
[6]Chechen-Noahcho DNA Project at Family Tree DNA, Kostanay 110008, Kazakhstan
[7]Laboratory of Evolutionary Genomics, Institute of Molecular Biology of National Academy of Sciences of the Republic of Armenia, Yerevan 0014, Armenia
[8]Department of Archaeology and Anthropology, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK
[9]Department of Linguistics, University of Hawai'i at Mānoa, Honolulu, Hawai'i 96822, USA
[10]DFG Center for Advanced Studies, University of Tübingen, 72074 Tübingen, Germany
[11]Senior author
[12]Lead contact
*Correspondence: pathak@ut.ee (A.K.P.), hovhannes.sahakyan@ut.ee (H.Sa.)
https://doi.org/10.1016/j.isci.2024.110016

Neolithic started later in South Asia than in West Asia. The earliest Neolithic site in South Asia, Mehrgarh, dating to ~9 kya, exhibits similarities to West Asian Neolithic cultures and is located near West Asia, specifically in Balochistan, present-day southwestern Pakistan.[44,45,47] This region later hosted the beginnings of the sophisticated Indus Valley civilization (IVC). In other South Asian regions, the Neolithic transition unfolded even later, yet featuring distinctive developments indicative of alternative pathways of independent evolution.[44,45,47] It is noteworthy, however, that despite the importance of understanding the South Asian Neolithic, there is a notable lack of aDNA studies. Reconstructions with younger ones find no shared ancestry with Anatolian Neolithic (AN) farmers but rather the presence of ancestry shared with Caucasus/Iranian hunter-gatherers (CIHG).[30,48] The South Asian Neolithic is primarily studied in conjunction with the Chalcolithic and Bronze Age periods and in the context of language spread, a topic explored further in the next paragraph.

In the Chalcolithic and Bronze Ages, West and South Asia underwent large-scale population movements,[27,30,49–51] coinciding with the time when the large language families were spread. In West Eurasia, these were Afro-Asiatic[52] and Indo-European,[27,30,50,51,53,54] among others. Currently, South Asians predominantly speak Indo-European or Dravidian languages and adhere to the hierarchical caste system and Hinduism.[55–57] Indo-European languages, prevalently spoken in northern regions, likely arrived from West/Central Asia during the Middle-Late Bronze Age and have been associated with the origin of the caste system.[30,41,50,58] On the contrary, Dravidian languages are predominant in southern India and Sri Lanka. Some isolated groups of speakers also live in southwestern Pakistan (Brahui) and northern India. The origin of Dravidian languages remains highly debated. They have been argued to have either an indigenous origin[46,59] or linked to the Neolithic dispersals from West Asia, as summarized in the Elamo-Dravidian hypothesis.[36–38,40–42,60] The hypothesis postulates linguistic and cultural connections between the extinct Elamite language spoken in ancient Elam (present-day southwestern Iran) and Dravidian languages. Frequently, the Indus Valley or Harappan civilization is suggested as an essential step for the spread of Dravidian languages, meaning that the language of this civilization was a related one. Certain linguistic studies have scrutinized the proposed connection, acknowledging Elamite as a language isolate.[61] A recent aDNA study suggests that the Iranian ancestral component in the IVC people came from individuals related to, but distinct from, Iranian farmers.[30] The contributing group lacked AN-related ancestry, common in Iranian farmers after ~8 kya. These CIHG-related individuals may have arrived in the Indus Valley before the advent of farming there and at ~7.4–5.7 kya, before the mature IVC mixed with people related to Indian hunter-gatherers (AASI), making a population called the "Indus Periphery Cline" (IPC). A population from IPC later mixed once more with AASI, giving birth to the ancient South Indian (ASI) ancestry,[62] which is currently widespread in southern Indian populations. However, a question remains: Which population of the two (CIHG-related or AASI) initially spoke a Dravidian language? Another aDNA study directly analyzed a Harappan genome, pushing the split between Iranian farmers and the CIHG-related group back to ~12 kya.[48] Unfortunately, these conclusions are based on a single individual, and direct dating was impossible. Moreover, the study was criticized for improperly modeling population history.[63]

The human Y chromosome haplogroup L-M20 holds significant potential as an avenue for unraveling the complex dynamics of ancient population interactions. M20, M11, M61, and other bi-allelic markers define this haplogroup.[64,65] It occurs more in South Asia but also in West Asia, Central Asia, and Europe (Figure 1). The haplogroup splits from its sister branch T ~45 kya.[3–5] Commercial Y chromosome whole sequencing efforts, as documented by phylogenetic trees of the Y Full (YFull) (v12.00.00, https://www.yfull.com/tree/L) and Family Tree DNA Discover (FTDNA) (accessed on 20-03-2024, https://discover.familytreedna.com/y-dna/L-M20/tree), found that at ~23.5 kya, the haplogroup L-M20 diverged into haplogroups L1 and L2 defined by M22[64] and L595 markers, respectively. The L1-M22 is the major branch so far, while L2-L595 is rare. Whole high-coverage Y chromosome studies included a small number of L1-M22 samples.[3–5,68] They have estimated the time to the most recent common ancestor (TMRCA) of L1-M22 to be around the LGM. Nevertheless, limited attention has been given to this haplogroup's origin, population dynamics, and migration patterns. Genotyping studies found three main branches defined by M317,[46] M27 or M76,[65] and M357[46] bi-allelic markers. All these branches fall within the L1-M22. A small number of individuals fall within paragroup L*. Studies conclude that the haplogroup L-M20 may represent early modern human populations in South Asia.[43,69–71] Others suggest its later migration to South Asia from West Asia with the Neolithic demographic expansion.[72,73] The earliest ancient individuals affiliated with this haplogroup substantially postdate its age. They are found in the late-Neolithic/Chalcolithic (~6.6 kya) in present-day Turkmenistan in a site bordering present-day Iran and in the Chalcolithic (~6.1 kya) in present-day Armenia within the Caucasus region.[27,74] Other individuals living before the common era (BCE) are found in the North Caucasus, present-day Iran, Greece, Turkey, Uzbekistan, Israel, and Pakistan.

In this study, we applied the Bayesian approach to unravel the unfolded dimensions of the haplogroup L1-M22 within South and West Asia. Specifically, we analyzed 165 high-coverage whole Y chromosome sequences to reconstruct the haplogroup's detailed phylogeny and demographic history. We conducted a statistically robust Bayesian phylogeographic analysis to infer the region where the haplogroup L1-M22 and its various lineages might have originated. Furthermore, we placed published ancient haplogroup L1-M22 genomes in the reconstructed phylogeny. Lastly, we compared our results to the published genetic, archaeological, and linguistic studies relevant to the conclusions about the origin and spread of the haplogroup L1-M22 and its branches.

RESULTS

Human Y chromosome haplogroup L-M20 and its major branch L1-M22 exhibit a predominant distribution in South and West Asia (Figure 1 and Table S1). In Central Asia, its prevalence is higher in the southern slopes adjacent to South Asia and Iran. This haplogroup remains rare in Europe, primarily concentrated in the central and eastern Mediterranean regions. Although this haplogroup's branches occur in both South and West Asia, the L1a1-M27/M76 and L1a2-M357/L1307 tend to have a higher frequency in South Asia. In contrast, the L1b-M317/PH982 branch is more prevalent in West Asia.

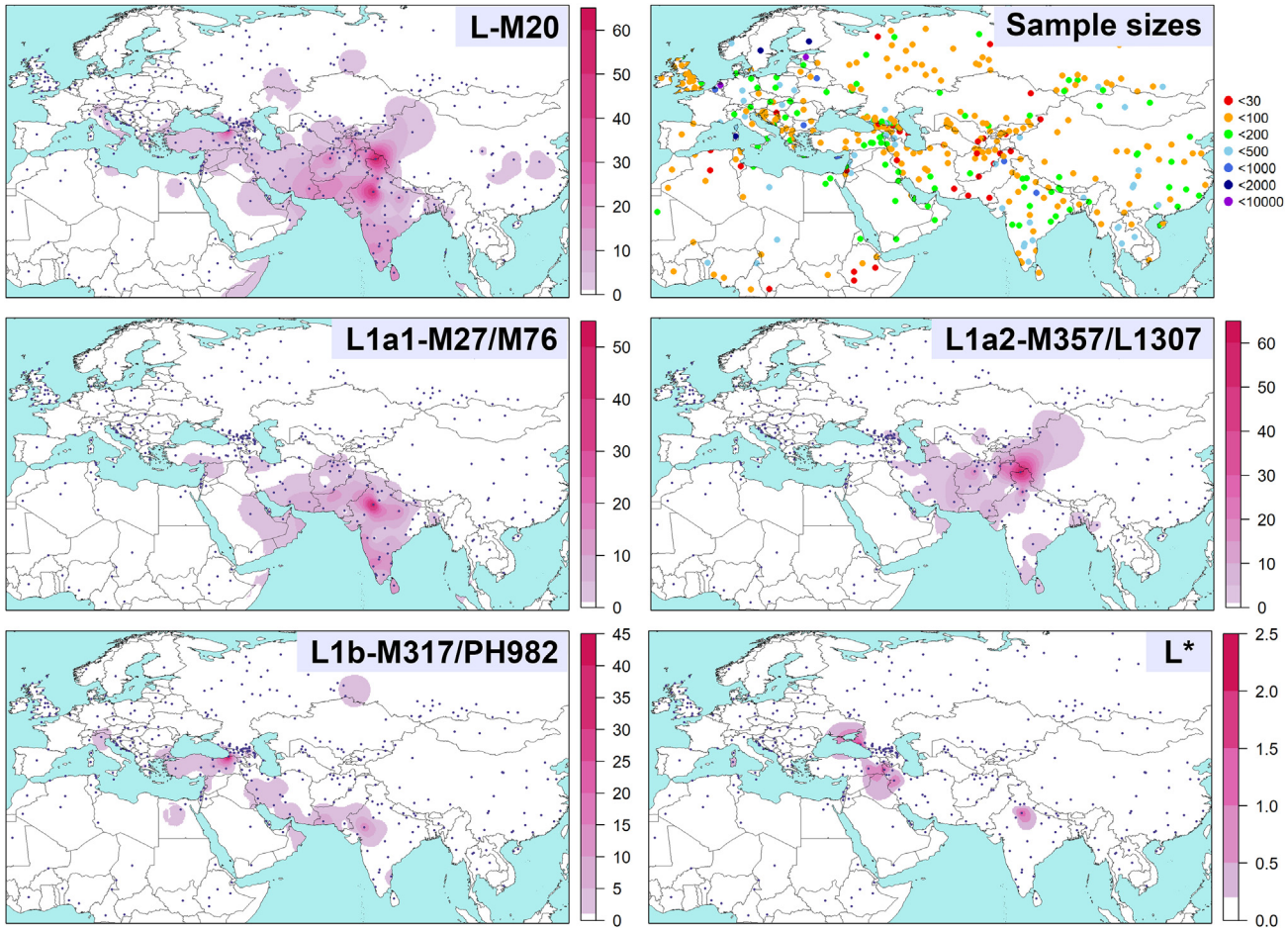


Figure 1. Spatial distribution maps of the haplogroup L-M20 and its branches

The frequency maps presented here are generated using data from Table S1. The datasets used for the frequency plots in these maps may differ. Blue-filled circles indicate sampling locations. Scale bars represent frequencies in percentages. Please be aware of the difference in scales between the maps. The figures were created using RStudio software.[66,67] The contour map was obtained from http://tapiquen-sig.jimdo.com.

By using 165 high-coverage whole Y chromosome sequences, we reconstructed the phylogeny of the haplogroup L1-M22. To improve our reconstruction, we included eleven sequences from other haplogroups. We called 4384 high-quality SNPs in the L1-M22 section of the phylogeny. Out of these SNPs, we observed that 0.4% (17) define more than one branch. The proportion of recurring SNPs aligns with the findings of other studies.[3,16,75]

Utilizing a published calibration point,[3] we estimated the TMRCA of haplogroup L1-M22 to be ~20.6 kya, with 95% HPD interval of 17.9–23.1 kya (Figure 2 and File S1). Our estimate aligns with those suggested previously[5,68] and overlaps with the Last Glacial Maximum (LGM) (26.5–19.0 kya).[14] The resulting Y chromosome mutation rate corresponds to $6.78e^{-10}$ mutations bp^{-1} $year^{-1}$ (95% HPD = $6.04e^{-10}$–$7.57e^{-10}$), which is comparable with earlier estimates.[3,16,76]

Despite our best efforts, we were unable to differentiate the SNPs that define the L1-M22 branch from those defining the L-M20 branch because we needed more data from other L-M20 branches. While most haplogroup L-M20 members known to date belong to the L1-M22 branch, some sources indicate the existence of at least one L-M20(xM22) lineage. The YFull includes three present-day individuals in the L2-L595 lineage, with residences in Lebanon, Turkey, and the Italian island of Sardinia. Additionally, the FTDNA features more individuals from Sardinia and one from the USA. aDNA studies have revealed additional individuals in this lineage, most of whom trace their origins to West Asia. All individuals exhibit CIHG ancestry; many also have AN and/or Levantine Neolithic (LN) ancestry. One individual of non-West Asian origin is found in Europe. Genotyping studies have also reported individuals that belong to the haplogroup L-M20 but do not fall in branches defined by M27/M76, M357, or M317 markers (Figure 1 and Table S1). This suggests a potential association with the L2-L595 lineage or even an unknown L-M20(xM22) lineage. Alternatively, they could belong to the rare L1a2-Y215410 sub-branch within the L1-M22 (Figures 2 and S1).

Our phylogenetic analysis unveils a detailed framework, as Figures 2 and S1 illustrate. Initially, L1-M22 bifurcates into two primary branches, L1a-M2481 and L1b-M317/PH982. The L1a-M2481 further diverges into L1a1-M27/M76 and L1a2-M2533, which, in turn, splits into L1a2-Y215410 and L1a2-M357/L1307. We will comprehensively explore these major branches in the subsequent paragraphs.

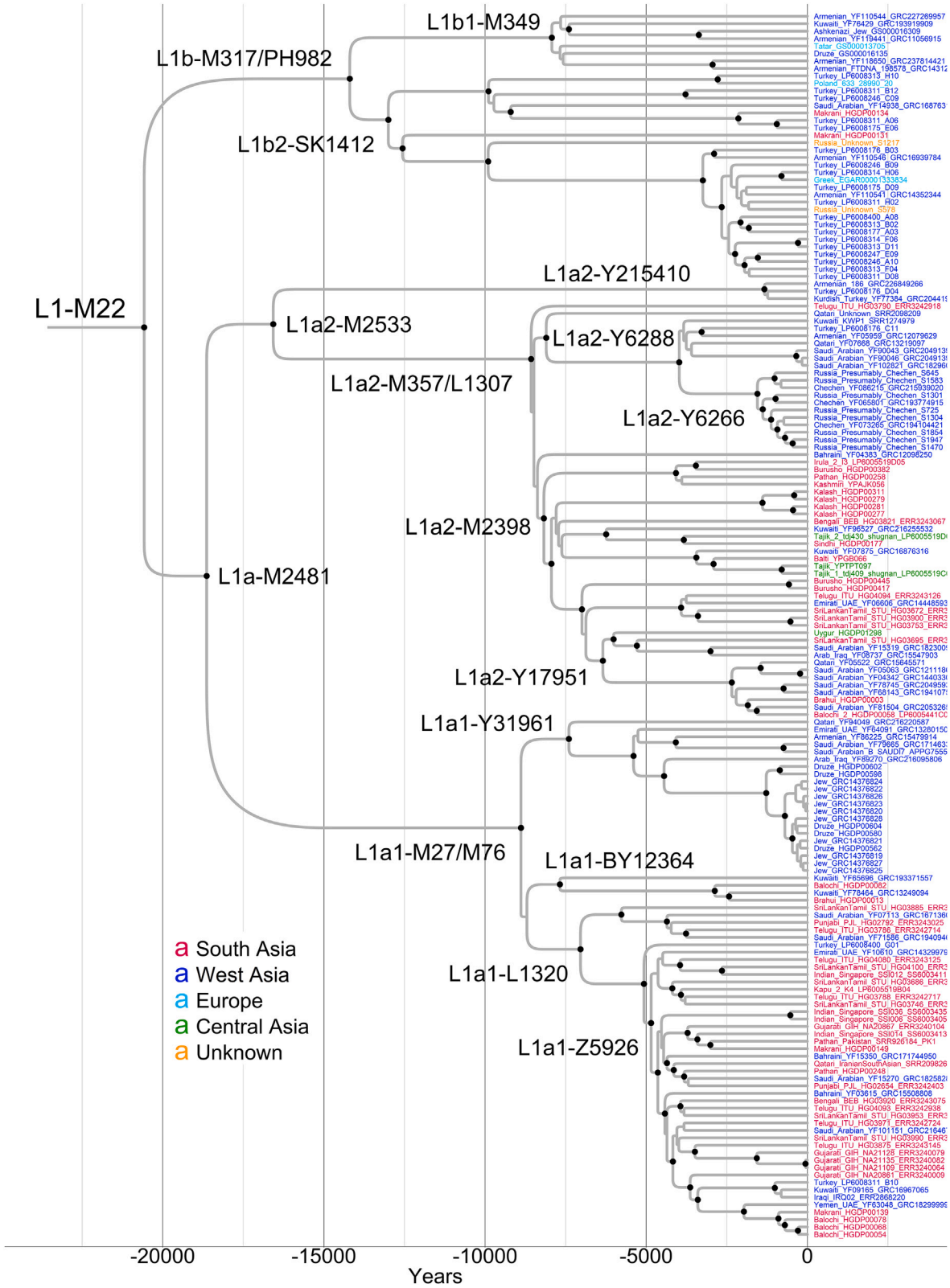


Figure 2. Bayesian phylogenetic tree of the haplogroup L1-M22

Black-filled circles mark the nodes with posterior probability ≥0.95. Sample IDs indicate population affiliation if the original publication explicitly states this; otherwise, the collection country is provided. See also Figure S1, Tables S2 and S4, and File S1.

The L1a1-M27/M76 splits starting at ~8.9 kya, giving rise to three sub-branches (Figures 2 and S1). The L1a1-Y31961 is strictly West Asian and coalesces at ~7.4 kya, including individuals from the Armenian Highland, the Levant, and the Arabian Peninsula. The L1a1-L1320 sub-branch is widespread in southern (more) and northern (less) India, Pakistan, and Bangladesh, with a small number of samples also present in West Asia. The L1a1-BY12364 sub-branch is minor, represented by only two individuals from Kuwait and two from Pakistan. Genotyping analyses confirm that L1a1-M27/M76 occurs widely in South and West Asia. As mentioned above, South Asian populations exhibit higher frequencies than West Asian ones (Figure 1 and Table S1).

The L1a2-M357/L1307 splits starting at ~8.6 kya, similar to the L1a1-M27/M76 (Figures 2 and S1). It yields two singleton lineages and two sub-branches. The singleton lineages originate from West and South Asia. The L1a2-Y6288 sub-branch is strictly West Asian and coalesces at ~8.1 kya. It includes individuals from the Arabian Peninsula, the Armenian Highland, and Anatolia. It also contains a recently diversified lineage (L1a2-Y6266, ~1.6 kya) with members from the Russian Federation. We need population information on many of these individuals, but we know three are Chechens from the Northeast Caucasus. Genotyping studies have shown that L1a2-M357/L1307 is found mainly among Chechens and Ingushes from the Northeast Caucasian populations and is rather frequent there.[77,78] Furthermore, in the YFull, the L1a2-Y6266 lineage predominantly comprises individuals from the Chechen and Ingush populations. It corresponds to the L-Y6248 in the FTDNA and includes individuals from the same populations. Therefore, the L1a2-Y6266 lineage is specific to these Nakh-Dagestanian-speaking populations. Importantly, this lineage coalesces in the West Asian part of the L1a2-M357/L1307, contradicting its migration from South Asia.[78] Unfortunately, it is hard to say anything conclusive about the more specific origin of this lineage as it splits at ~4 kya from other lineages. The L1a2-M2398 sub-branch coalesces at ~8.2 kya and occurs mainly in South Asia. It includes many individuals from southern and northern India, Pakistan, and Bangladesh. This sub-branch also encompasses all four sequenced Central Asian individuals of haplogroup L1-M22 in our study, as well as the most from a recent study about Central Asia.[79] Their lineages are scattered in South Asian ones, indicating a probable origin. Genotyping analyses found that L1a2-M357/L1307 more frequently occurs among South Asian populations than in West Asian ones (Figure 1 and Table S1). Within South Asia, it is more frequent among populations from Pakistan and northern and northwestern India than in southern ones. The L1a2-Y215410 is a minor recent (~1.3 kya) sub-branch consisting of an Armenian, a Kurdish, and an individual from Turkey. In addition, the YFull and FTDNA include one or two individuals each from Iraq, Syria, and Kazakhstan.

It is crucial to highlight that the L1a1-M27/M76 and L1a2-M357/L1307 branches exhibit strikingly similar temporal and spatial patterns. Both display sub-branches in both West Asia and South Asia, and notably, they share a coalescence time frame in the Early Holocene (Figures 2 and S1). This significant observation suggests a shared historical context or demographic events that impacted populations across these regions during this critical period.

The L1b-M317/PH982 branch presents a notable contrast (Figures 2 and S1). It is primarily distributed in West Asia (27 out of 34), more in its northern latitudes (23 out of 34). Coalescing at ~14.2 kya, this is the oldest region-specific branch of the L1-M22. Non-West Asian members in this branch are two Makranis from Pakistan, two individuals with unknown ancestry, a Tatar from Russia, one from Poland, and one from Greece. Notably, no one from 32 L1-M22 individuals from India and Sri Lanka belongs to this branch. Genotyping analyses largely corroborate the northern West Asian concentration of this branch, adding Iranian populations to delineate the core distribution area (Figure 1 and Table S1). With few populations having L1b-M317/PH982 individuals, Pakistan, northwestern India, and Central Asia represent the periphery of the distribution. The distinct concentration of this early branch in West Asia and its age add credence to the hypothesis of L1-M22's potential origin in this region.

We conducted a robust statistical framework to infer the place of origin of the haplogroup L1-M22 and its spread explicitly using Bayesian continuous phylogeographic analysis (Figure 3). The credible (80% HPD) area of the L1-M22 locations includes southeastern Anatolia, the Armenian Highland, the South Caucasus, the Iranian Plateau, Mesopotamia, the Levant, northern and eastern Arabian Peninsula, southwestern Pakistan, western Afghanistan, and southern Turkmenistan. At ~10 kya, the overall area extends west toward central Anatolia and Cyprus, north to the North Caucasus, and slightly east, more in Afghanistan and Pakistan, while also covering a tiny pinch of western Gujarat in India. Importantly, more intensive diversification was occurring in and around the Fertile Crescent. At ~8 kya, the extension continues toward Anatolia, covering almost all of it, and toward India, covering slightly more in the western region. While the earlier diversification was still ongoing in and around the Fertile Crescent, at this time, two new ones appeared in the junction area of West and South Asia. These events correspond to the L1a1-M27/M76 and L1a2-M357/L1307 branches. Both branches direct two opposing vectors to West and East. At ~3.5 kya, the overall haplogroup L1-M22 area extends to northern, western, central, and southern India and covers Pakistan and Afghanistan almost wholly. In the west, it extends to southeastern Europe. In the south, it diffuses more in the Arabian Peninsula and covers the northeastern Sinai Peninsula. In other regions, the diffusion extended after ~3.5 kya. To summarize, our Bayesian continuous phylogeographic analysis provides essential insights into the origin of the haplogroup L1-M22 and its diversification, which initially occurred in West Asia and then in South Asia.

We performed Bayesian skyline analysis to examine the dynamics of the haplogroup L1-M22's Ne (Figure 4, upper plot). The analysis revealed that the population remained constant from the beginning until ~10 kya. Between roughly 10 and 7 kya, Ne experienced a significant increase, indicating a potential correlation with the demographic shift during the Neolithic period. Despite a bottleneck signal from ~7 to ~4 kya, we consider it non-significant due to overlapping HPD intervals. From ~4 kya onwards, the population underwent a remarkable expansion within only ~1 ky. The lowest HPD bound at ~3 kya surpasses the highest HPD bound of all prior epochs. Lastly, Ne shows no change over the past ~3 kya, which can also result from the limited sample size.

To investigate whether changes in population dynamics characterize one population or are shared by both, we ran separate Bayesian skyline analyses using genetic data of individuals from only West Asia or South Asia (Figure 4, lower plot). Our results show distinct population trajectories for the two regions, with marked differences emerging ~10 kya. In West Asia, Ne of the haplogroup L1-M22 increases starting

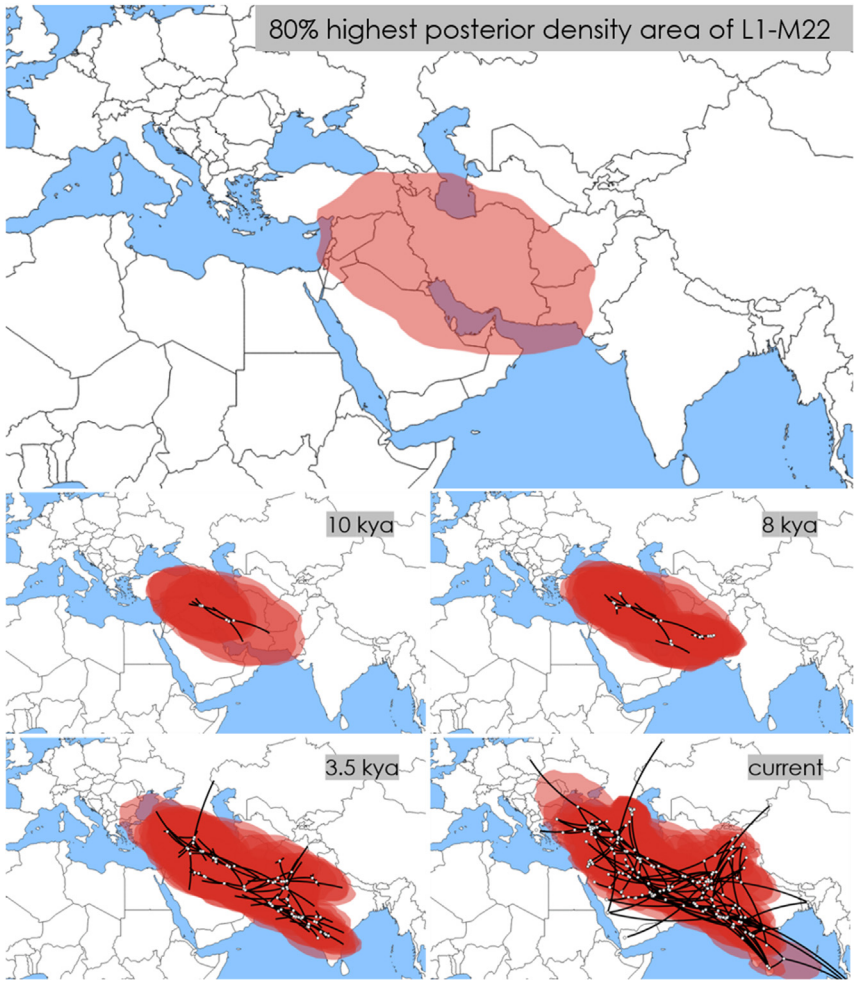


Figure 3. Inferred locations of the haplogroup L1-M22's root and nodes in different time points

Shaded in pink are the 80% HPD areas of the node locations inferred by Bayesian continuous phylogeographic analysis in Beast v1.10.4 software.[80] Open circles show median estimates, while black lines indicate the branches of the maximum clade credibility tree. Maps were generated in spreaD3 software (v0.9.7.1rc).[81] The base map was downloaded from https://github.com/johan/world.geo.json/blob/master/countries.geo.json.

around this time and continuing until ~7 kya. This pattern reproduces the pattern seen during this time in the whole L1-M22 haplogroup's Bayesian skyline analysis. It is also consistent with the intensive diversification in and around the Fertile Crescent revealed by the Bayesian continuous phylogeographic analysis described above. After ~7 kya, Ne continues to increase, but only slightly as the HPD bounds overlap before and after this epoch.

In contrast, Ne of South Asian L1-M22 shows only a slight fluctuation between ~10 and ~7 kya and remains relatively stable from the beginning until ~4 kya. Interestingly, at ~4 kya, Ne of South Asian L1-M22 undergoes a rapid expansion within only ~1 ky, a trend also observed in the entire haplogroup analysis and Bayesian phylogeographic analysis. Earlier research[4] also detected an expansion of haplogroup L1-M22 in South Asia ~4.4 kya, which coincided with that of haplogroup R1a-Z93. However, they consider this expansion weak, a conclusion likely owing to the small sample size of their study.

Ancient DNA studies extend our knowledge about the human past. However, we want to draw attention to two critical limitations that, together with low DNA preservation, affect the utility of aDNA for the haplogroup L-M20's research. Firstly, the earliest known ancient individuals who belong to this haplogroup lived ~6.6[74] and ~6.1[27] kya (Figure S3), a time frame much later than the haplogroup's age (>20.6 kya). Consequently, while aDNA samples offer valuable insights into more recent periods, they have limited power to inform about this haplogroup's origin and early diversification. Secondly, our analysis of the ancient genomic data is constrained to the available haplogroup L-M20 SNPs in the "1240k capture" technology, as all the ancient genomes (Table S5), except for one, were genotyped using this approach.

The earliest so far found ancient (~6.6 kya) individual of haplogroup L1-M22 lived in present-day Turkmenistan, close to the present-day border with Iran[74] (Figure S3). He belongs to the L1a2-M357 branch and shares most of his autosomal ancestry with CIHG-related individuals while sharing no ancestry with AN individuals. Almost the same-age (~6.1 kya) individuals lived in the Areni-1 cave in the South Caucasus.[27]

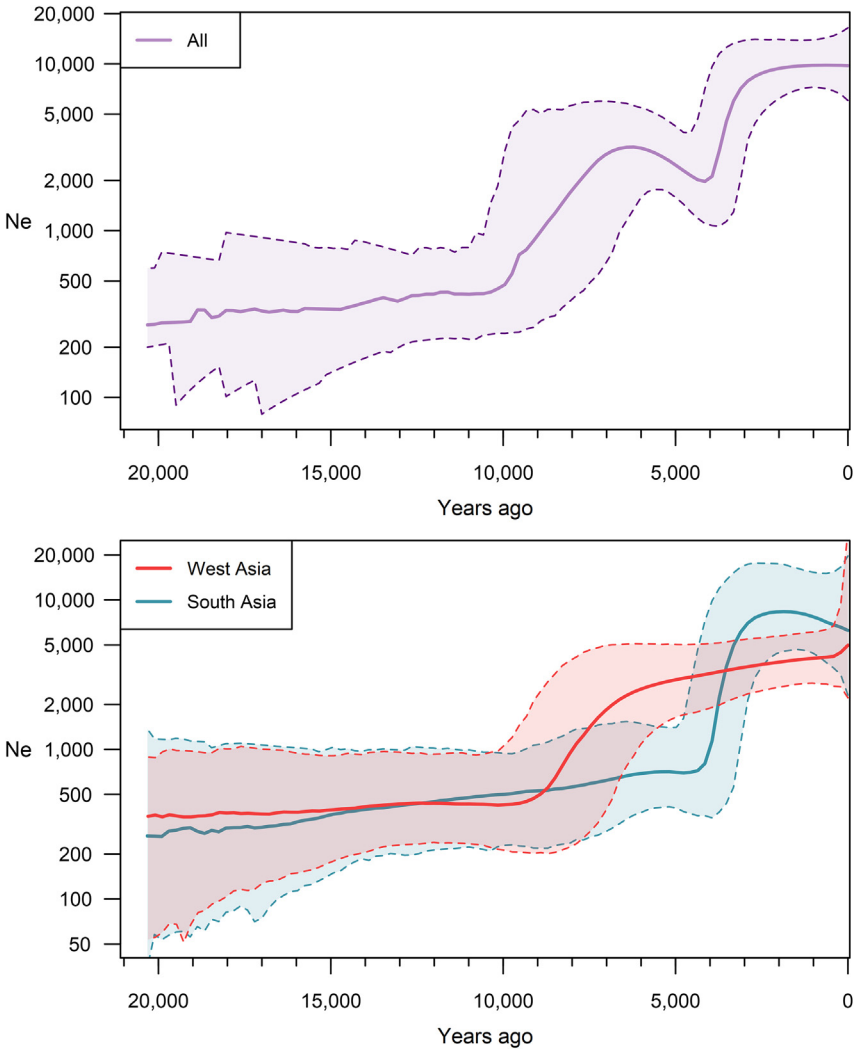


Figure 4. Effective population size dynamics of the haplogroup L1-M22

The solid line is the median estimate, while the dashed lines show the 95% HPD limits. Ne - effective population size.

Interestingly, two of three individuals belong to the L1a1-Y31961 branch (Figure S2), the West Asian sub-branch of L1a1-M27/M76 in our phylogeny (Figures 2 and S1). The other individual belongs to the L1a1-M27/M76 branch and lacks reads covering SNPs that define the known downstream branches, including the L1a1-Y31961. These individuals represent a Chalcolithic population of the South Caucasus who share ancestry with CIHG (~50%), AN (~30%), East European hunter-gatherer (EHG) (~10%), and LN (~9%) populations.[51] The next earliest (~4.9 kya) individual of the haplogroup L1-M22 is found in the Shahr-i-Sokhta site in present-day southeastern Iran.[30] This Bronze Age individual belongs to the L1a2-M357/L1307 (Figure S2) branch and shares all his ancestry with CIHG populations,[51] although others in the cluster share on average ~19% of ancestry with AN, ~4% with Andamanese hunter-gatherer (AHG), and ~12% with West Siberian hunter-gatherer (WSHG) populations.[30] In this population, there are some individuals who, compared to the so-called main cluster, share more ancestry with AHG populations and lack a detectable AN-related ancestry. Similar composition is also found in some individuals from Gonur, the Bactria Margiana Archaeological Complex (BMAC) site in Central Asia. These individuals from Shahr-i-Sokhta and Gonur represent the IPC formed between ~5400 and 3700 BCE.[30] Current South Asian populations descend from admixture between a yet unsampled population from this cline and a population from the Central Steppe Middle Bronze Age (CSMBA). Two younger (~3.8 kya) individuals of the haplogroup L1-M22 lived in Central Asia (Uzbekistan).[30] Both belong to the L1a2-M357/L1307 branch. One has the derived allele for a downstream L1a2-M2398 lineage-defining SNP, the lineage to which all four present-day individuals from Central Asia affiliate. The other genome lacks reads for positions defining the downstream lineages. Although postdating, these individuals resemble individuals from the majority group of the BMAC mentioned above. That is, they share ~59% ancestry with CIHG, ~26% with AN, ~2% with AHG, and ~12% with WSHG populations.[30] The haplogroup L1-M22 is also found among 14 ancient (~3.0–2.3 kya) individuals from present-day Swat Valley, Pakistan.[30] They all belong to the L1a2-M357/L1307 branch. Three were further assigned to the L1a2-Y6288, L1a2-Y17951, and Balti_YPGB066 lineages. These individuals are

those that originate from admixture between IPC-related and CSMBA-related populations. An ancient (~2.8 kya) individual from Hassanlu, present-day Iran, and another one (~2.3 kya) from present-day Armenia[51] share derived alleles of three and two SNPs, respectively, with a Qatari individual (SRR2098209) down to the L1a2-Y6288 lineage. They look similar in autosomal variation and share ~56% of their ancestry with CIHG, ~24% with LN, ~16% with AN, and ~3% with EHG populations. Interestingly, another ancient individual (~1.9 kya) from present-day Armenia[82] also belongs to the L1a2-Y6288 lineage. However, he is in a distinct sub-lineage, which includes currently living individuals from Armenia and Turkey. A Hunnic individual (~1.7 kya) from the Tian Shan region of present-day Kyrgyzstan[83] belongs to the L1a1-Y31961 branch. Another Central Asian individual of the same age is found in present-day Kazakhstan.[84] He belongs to the L1a2-M357/L1307 branch and shares derived alleles for five SNPs with a Telugu individual (HG03790) from India. The earliest L1-M22 individual from Europe lived in the Medieval time (~0.85 kya) in present-day Italy.[85] He belongs to the L1a1-Z5926 branch.

The earliest representatives (~5.2 kya) of the L2-L595 lineage were uncovered in the Late Maykop culture from both the Northwest and Northeast Caucasus[86] (Figure S3). Their ancestral profile resembles that of Chalcolithic populations from present-day Armenia and Iran, as well as Kura-Araxes individuals from present-day Armenia and the Northeast Caucasus.[86] A possible Iranian Chalcolithic representative (~4.8 kya) is found in Tepe Hissar, a Copper Age-to-Bronze Age urban settlement in the central Iranian Plateau.[30] However, the assignment to L2-L595 is based on only 3 C-to-T and 2 A-to-G transitions, which can also result from postmortem damage. The genetic makeup of Chalcolithic populations from present-day Armenia and Iran suggests a mixture of ancestries shared with CIHG, AN, and LN populations and a minor amount of ancestry shared with EHG/WHG population,[51,86] likely inherited alongside the AN-related ancestry.[27] Another ancient (~3.9 kya) member of the L2-L595 lineage is found in the Alalakh population from the Middle-Late Bronze Age northern Levant.[31] This population combines ancestries shared with AN, CIHG, and LN populations. Another ancient (~3.2 kya) member is found in the Iron Age southern Levant. He shares the same three ancestry components.[87] One more ancient (~2.5 kya) member is from present-day Turkey's Batman region.[88] This individual shares ancestry overwhelmingly with CIHG and LN populations. This lineage's most recent ancient individual is found among Vikings (~1.1 kya) from Sweden.[85]

For two ancient L-M20 individuals, owing to low data quality, it was impossible to assign them to the downstream branches (Figure S3). One (~4.7 kya) is from the Early Bronze Age time of present-day Greece.[89] This population displays excessive allele sharing with earlier and contemporaneous West Eurasian groups from present-day Iran and the Caucasus. Another ancient individual (~2.4 kya) is from the Aegean of present-day Turkey.[88]

While aware of the inherent limitations in sample ages and data availability, our analysis of ancient DNA reveals a striking consistency: the sole genetic component universally shared among all individuals within haplogroup L-M20 related to CIHG. This finding underscores a strong genetic affinity between haplogroup L-M20 and these ancient populations, at least since ~6.6 kya.

DISCUSSION

This study presents extensive research on human Y chromosome haplogroup L1-M22 - the major branch of the haplogroup L-M20. Our research draws from an analysis of 165 high-coverage whole Y chromosome sequences. It employs a robust statistical framework, shedding more light on this haplogroup's origin, distribution, and diversification. Consistent with the previously suggested estimates,[4,13,68] it coalesces ~20.6 kya, shortly after the L-M20 (the YFull and FTDNA). Many West Asian Y chromosome and mitochondrial DNA lineages display a pattern similar to the haplogroup L-M20: an extended bottleneck before and diversification during or shortly after the LGM. Other examples are Y chromosome haplogroups G2a-P15, J1-M267, J2a-M410, J2b-M12,[3,16] and mitochondrial DNA haplogroups U7 and J1,[9,15] among others. The reduced genetic diversity can be attributed to the harsh glacial conditions, which may have markedly restricted the number of founding individuals. Nevertheless, the divergence began right during the LGM, which, although surprising, aligns with growing archaeological evidence regarding the habitation of northern West Asia during the same period.[90]

Our Bayesian continuous phylogeographic analysis supports the Y chromosome haplogroup L1-M22's origin in West Asia more than in South Asia (Figure 3). The 80% HPD area overwhelmingly encompasses West Asian regions such as the Iranian Plateau, Mesopotamia, the Armenian Highland, the Caucasus, the Arabian Peninsula, and the Levant. Although some overlap of the 80% HPD area with northwestern South Asia may suggest the origin there, two additional lines of evidence add more weight to the origin of the haplogroup L1-M22 in West Asia. Firstly, the sister branch of L1-M22, L2-L595, is found exclusively in West Asia and Europe[30,31,85–88] (the YFull, FTDNA, Table S5, and Figures S2 and S3). Second, as detailed in the previous paragraph, the distinct pattern of a long bottleneck in the haplogroup L-M20 after splitting from haplogroup T at ~45 kya resembles what is observed in other West Asian lineages[3,4,15,16] and differs from the earlier divergence seen in South Asian lineages like the Y chromosome haplogroup H and mitochondrial DNA haplogroup U2, as well as others native to South Asia.[3–8]

Our results indicate that the haplogroup L1-M22 population in West Asia began to expand ~10 kya (Figure 4). This coincides with the Neolithic demographic transition, a crucial period of human history when the world's population started transitioning from a hunter-gatherer lifestyle to a more settled way of life based on agriculture and animal husbandry.[17] This transition occurred earliest in the Fertile Crescent region of West Asia ~12 kya.[20–22] Centers with at least three genetically distinct population groups were revealed.[23,26,27,29] Intriguingly, our Bayesian continuous phylogeographic analysis infers that starting at ~10 kya, L1-M22 expands, more intensively, around the northern Fertile Crescent region (Figure 3), probably inhabited by populations with the CIHG autosomal heritage. Therefore, we propose that these populations may have contributed to the dissemination of the haplogroup L1-M22 in West Asia. This conclusion finds further support in aDNA studies, which reveal that the genome-wide ancestral component universally shared among all individuals within the haplogroup L-M20 corresponds to that shared with the CIHG populations. This observation may initially seem counterintuitive given the absence of haplogroup

L-M20 in Bronze Age Pontic-Caspian steppe populations, despite these groups having half of their genetic heritage originating from a West Asian population.[53,54] Additionally, contemporary populations with a prominent steppe ancestry, like those in northeastern Europe,[53,54] exhibit only minimal, if any, the presence of haplogroup L-M20.[91,92] This inconsistency can be elucidated by considering that, unlike the autosomal ancestry, the predominant paternal lineages in Bronze Age Pontic-Caspian steppe populations were inherited from earlier local populations.[30] This marked scarcity of representation also extends to other West Asian Y chromosome haplogroups.

We did not observe an expansion in South Asian L1-M22 Y chromosomes during the Neolithic time (Figure 4). This suggests that while the West Asian L1-M22 population was expanding, likely linked to the changes during the Neolithic demographic transition, the ancestral population of the present-day South Asian L1-M22 lineages was already distinct from their West Asian Neolithic fellows. Thus, at least two populations genetically connected to the CIHG have probably emerged in West Asia around the Holocene. Groups of one population then remained in or close to the center of the Neolithic demographic transition with a lack of expansion to South Asia. Groups of another one migrated to South Asia.[30,48,50,73] These non-Neolithic CIHG-related groups probably also involve the ancestors of the South Asian L1a1-L1320 and L1a2-M2398 branches. Individuals representing these branches migrated to South Asia starting at ~8 to ~6 kya. Notably, no aDNA individuals from West Asia belong to these branches, further supporting their association with South Asia. Concurrently, the L1a1-Y31961 and L1a2-Y6288 branches remained in West Asia with the L1b-M317/PH982. Interestingly, ancient individuals from the L1a1-Y31961 branch lived already at the Chalcolithic time (~6.1 kya) in present-day Armenia. It is also interesting that an individual from the Early Bronze Age (~4.9 kya) Shahr-i-Sokhta site in present-day southeastern Iran belongs to the L1a2-M357/L1307 branch and shares 100% autosomal ancestry with CIHG-related populations. Collectively, these findings provide further evidence supporting the association of haplogroup L1-M22 with autosomal CIHG-related ancestry, potentially indicating their role in spreading the haplogroup not only in West Asia but also into South Asia. In the latter case, though, the population had not been mixed with AN ones[30] and shows no sign of expansion. It is important to note that the precise geographic locations of the CIHG-related populations may not align with current state borders due to historical factors.

The finding of a distinct ancient group likely closely related to CIHG during the Holocene is a crucial discovery, potentially shedding light on the origins of contemporary South Asian L1-M22 lineages. This finding corroborates the previously proposed hypothesis of a unique CIHG-related population in the eastern regions of present-day Iran and bordering regions of Central Asia.[30] The recent discovery of a ~6.6 kya individual with an L1-M22 lineage in present-day Turkmenistan[74] provides additional evidence for the association of this haplogroup with a CIHG-related ancestral group to that region, which served as a junction between West and South Asia. The IPC population originated due to the admixture of this kind of population and an AHG-related population at around 5400 to 3700 BCE,[30] a time frame overlapping closely with our age estimates of South Asian L1-M22 branches. L1-M22 lineages in a subset of ancient individuals from the IPC and later individuals from the Swat Valley[30] reinforce this hypothesis. Our results also align with the formal modeling of that study, demonstrating that the substantial shared ancestry between present-day South Asians and early Holocene populations in present-day Iran[27] arose as a result of a genetic influx from IPC people into later South Asians rather than a substantial westward gene flow of South Asian ancestry onto the Iranian Plateau.

Such an ancient group closely related to the IPC holds profound implications for understanding the origin and dissemination of the Dravidian family of languages, the second-largest language family in South Asia. Studies suggest that the migration of agriculture and herding from West to South Asia may have led to the introduction of proto-forms of Dravidian languages with the Neolithic migration.[36–38,40–42] However, the absence of AN ancestry in the IPC[30] narrows down the origin of such movement to a population from West Asia lacking AN ancestry. A recent study found notable AN ancestry in a few contemporary populations of northern India.[93] Nevertheless, this can be linked with a more recent event: the Steppe migration. Our finding about the lack of expansion of South Asian L1-M22 lineages before ~4 kya is consistent with the absence of a large-scale Neolithic population movement from West to South Asia. Instead, it is plausible that a population bearing only CIHG-related ancestry and L1-M22 lineages moved from West to South Asia, potentially also during the Neolithic time, but without a substantial expansion.

Within South Asia, the dispersals of Dravidian languages were most probably conducted by people with ASI ancestry,[30,62] which was likely formed after the collapse of IVC when its people migrated eastward and southward and mixed with populations carrying higher AASI ancestry.[30] Hence, one of the two ancestral population groups, i) the IPC or ii) the ancient eastern and southern South Asian populations with higher AASI ancestry, might be the speakers of Dravidian languages.[30] Linguistic studies show that already IVC people may have spoken a Dravidian language.[94–96]

The potential Elamo-Dravidian linguistic connection[60,97] is critically important in unraveling the origins of the Dravidian language family and its possible association with the IPC. If substantiated, a population with a CIHG-related genetic heritage would be the best candidate for disseminating both Elamite and Dravidian languages. Our study supports this hypothesis by suggesting a connection between the roots of all L1-M22 lineages and CIHG-related genetic ancestry while also delineating temporal boundaries with the reconstructed Y chromosome haplogroup L1-M22 tree. Arguably, West Asian L1a lineages likely contributed to the development of the Elamite language. In contrast, South Asian L1a lineages after ~8 kya probably migrated from the Iranian plateau, potentially contributing to the spread of Dravidian languages to South Asia. Our results do not support the suggestion that the geographical expansion started from ancient Elam, present-day southwestern Iran. Instead, both these regions could be the final areas of migration started from yet another region inhabited by a population of CIHG-related ancestry.

The expansion of Dravidian languages into southern India aligns with the population expansion that began ~4 kya, as observed in our Bayesian analysis of South Asian lineages of the haplogroup L1-M22. The expansion coincided with the arrival of Steppe ancestry in South Asia during the Middle and Late Bronze Age.[30,50] Notably, Steppe ancestry-rich individuals in Central Asia belong to paternal haplogroups

other than L1-M22. Consequently, local South Asian populations bearing haplogroup L1-M22 underwent population expansion simultaneously with incoming Steppe individuals as Middle and Late Bronze Age Steppe-related paternal haplogroup R1a[24,30,54] also expanded during this time in South Asia.[4] A similar pattern is seen also with South Asian maternal lineages.[73] Interestingly, the beginning of population expansions followed the megadrought event that transformed several complex societies of the Bronze Age, including the IVC.[98]

This study illuminates the genetic history of Y chromosome haplogroup L1-M22. Our research emphasizes West Asia's pivotal role in this haplogroup's emergence and its possible association with CIHG-related genome-wide ancestry. We characterized at least two distinct population groups bearing this ancestry during the Early Holocene. One was expanding in West Asia during the Neolithic demographic transition, and the other migrated without expansion to South Asia (~8-6 kya), possibly contributing to the spread of Dravidian languages. Importantly, our findings support the connection of Dravidian languages with ancient Elamite language spoken in present-day southwestern Iran, possibly linked to migration from West to South Asia. Nevertheless, our inferences challenge earlier claims about the dispersal of Dravidian languages in connection to the expansive spread of farming and align with the recent study about the lack of AN legacy in the ancestors of South Asians. The local South Asian L1-M22 lineages expanded between ~4 and ~3 kya, coinciding with the introduction of the Steppe ancestry. Hence, our research offers valuable insights into the confluence of genetic and linguistic developments during this pivotal period in South Asian history. Further interdisciplinary research is needed to fully understand these intricate patterns of the human past, integrating genetics, linguistics, and archaeology to provide a more comprehensive narrative of our shared heritage.

Limitations of the study

Several cautions and limitations must be acknowledged when interpreting our findings. The limited sample size and potential temporal and spatial gaps in genetic data may not fully capture the diversity of West and South Asian populations, impacting the depth of insights into L1-M22's history. It is also essential to caution against overemphasizing direct associations or causal relationships between haplogroup L1-M22 and the CIHG autosomal ancestry. Additionally, caution is advised against using a single genetic lineage as the sole representative of a language group, as populations comprise multiple lineages. Furthermore, linking genetic data with cultural and historical events, like disseminating Elamite and Dravidian languages, calls for careful consideration, as the process is complex. We hope this study serves as a foundation for further research. A comprehensive understanding of genetic and cultural dynamics in West and South Asia requires future investigations using additional genetic markers and interdisciplinary approaches. Ultimately, we emphasize that our findings cannot and must not be used to promote divisive or exclusionary narratives or to undermine or criticize individuals or groups associated with different haplogroups. Genetic data are valuable tools for understanding the human past; however, they are incompatible with being misused to oversimplify or misrepresent the rich diversity of human 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 STUDY PARTICIPANT DETAILS
  - Whole high-coverage Y chromosome sequences
  - Published whole high-coverage Y chromosomes
  - Genotyping of SNP markers
- METHOD DETAILS
  - Reads mapping and multi-sample variants calling
  - Variant filtering
  - Tree reconstructions and coalescence analysis
  - Bayesian phylogeography
  - The demographic history reconstruction
  - Annotation
  - Phylogenetically informative SNP genotyping
  - Spatial frequency analyses
  - Ancient L-M20 representatives' affiliation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110016.

ACKNOWLEDGMENTS

We acknowledge the individuals who provided their complete Y chromosome data for this study and the DNA donors whose contributions were essential to our research. We also express our gratitude for the various sources of support that enabled this study, including the Estonian Research Council (Grant Number PRG1071, supporting D.M.B., E.M., S.R., and R.V., H.Sa.; and Grant Number PUTJD1186, for supporting A.K.P.); and the Science Committee of the Armenian Ministry of Education and Science (Grant Number 21AG-1F025, for supporting L.Y.); and the European Union's Horizon 2020 research and innovation program (OCSEAN, Grant Number 873207, for supporting P.E.). We thank Professor Toomas Kivisild (KU Leuven, Belgium) for helpful comments on this article. Many analyses were performed on the High-Performance Computing cluster at the University of Tartu (UT Rocket, 2018).[99] We also acknowledge the availability of twenty-seven high-coverage complete Y chromosomes from the 1000 Genomes Project panel, which were generated at the New York Genome Center. This endeavor was made possible through funding provided by the National Human Genome Research Institute (Grant Number 3UM1HG008901-03S1) and the National Institute of General Medical Sciences at the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Conceptualization, A.K.P. and H.Sa.; sequencing, E.M., S.R., R.V., and H.Sa.; SNP genotyping and frequency data, H.Sa.; data analysis, H.Sa.; interpretation, A.K.P., P.E., R.V., and H.Sa.; provided samples and high-coverage Y chromosome sequences, H.Si., I.A.A.I., P.H., D.M.B., P.A., L.Y., and H.Sa.; wrote the manuscript: A.K.P., and H.Sa., with inputs from all co-authors. All authors reviewed the manuscript. Correspondence: A.K.P., Email: pathak@ut.ee; H.Sa., Email: hovhannes.sahakyan@ut.ee

DECLARATION OF INTERESTS

D.M.B. declares stock ownership at Gene by Gene, Ltd. All other authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, H.Sa. used ChatGPT v3.5 to improve the text's readability and language. After using it, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

Received: January 12, 2024
Revised: April 6, 2024
Accepted: May 14, 2024
Published: May 17, 2024

REFERENCES

1. Stringer, C.B., and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. Science *239*, 1263–1268. https://doi.org/10.1126/science.3125610.
2. Mellars, P. (2006). Going East: New genetic and archaeological perspectives on the modern human colonization of Eurasia. Science *313*, 796–800. https://doi.org/10.1126/science.1128402.
3. Karmin, M., Saag, L., Vicente, M., Wilson Sayres, M.A., Järve, M., Talas, U.G., Rootsi, S., Ilumäe, A.-M., Mägi, R., Mitt, M., et al. (2015). A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. *25*, 459–466. https://doi.org/10.1101/gr.186684.114.
4. Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., Wilson Sayres, M.A., Ayub, Q., McCarthy, S.A., Narechania, A., Kashin, S., et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat. Genet. *48*, 593–599. https://doi.org/10.1038/ng.3559.
5. Hallast, P., Agdzhoyan, A., Balanovsky, O., Xue, Y., and Tyler-Smith, C. (2021). A Southeast Asian origin for present-day non-African human Y chromosomes. Hum. Genet. *140*, 299–307. https://doi.org/10.1007/s00439-020-02204-9.
6. Thangaraj, K., Chaubey, G., Singh, V.K., Vanniarajan, A., Thanseem, I., Reddy, A.G., and Singh, L. (2006). In situ origin of deep rooting lineages of mitochondrial macrohaplogroup "M" in India. BMC Genom. *7*, 151. https://doi.org/10.1186/1471-2164-7-151.
7. Atkinson, Q.D., Gray, R.D., and Drummond, A.J. (2008). mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. Mol. Biol. Evol. *25*, 468–474. https://doi.org/10.1093/molbev/msm277.
8. Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B.N., Urade, B.P., Mallick, S., Bandopadhyay, S.S., Barua, P., Barik, S.S., Basu, D., et al. (2009). Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: Dispersal of modern human in South Asian corridor. PLoS One *4*, e7447. https://doi.org/10.1371/journal.pone.0007447.
9. Pala, M., Olivieri, A., Achilli, A., Accetturo, M., Metspalu, E., Reidla, M., Tamm, E., Karmin, M., Reisberg, T., Hooshiar Kashani, B., et al. (2012). Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. Am. J. Hum. Genet. *90*, 915–924. https://doi.org/10.1016/j.ajhg.2012.04.003.
10. Richards, M.B., Soares, P., and Torroni, A. (2016). Palaeogenomics: Mitogenomes and migrations in Europe's past. Curr. Biol. *26*, R243–R246. https://doi.org/10.1016/j.cub.2016.01.044.
11. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206. https://doi.org/10.1038/nature18964.
12. Pagani, L., Lawson, D.J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., et al. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. Nature *538*, 238–242. https://doi.org/10.1038/nature19792.
13. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012. https://doi.org/10.1126/science.aay5012.
14. Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., and McCabe, A.M. (2009). The Last Glacial Maximum. Science *325*, 710–714. https://doi.org/10.1126/science.1172873.
15. Sahakyan, H., Hooshiar Kashani, B., Tamang, R., Kushniarevich, A., Francis, A., Costa, M.D., Pathak, A.K., Khachatryan, Z., Sharma, I., van Oven, M., et al. (2017). Origin and spread of human mitochondrial DNA

haplogroup U7. Sci. Rep. *7*, 46044. https://doi.org/10.1038/srep46044.
16. Sahakyan, H., Margaryan, A., Saag, L., Karmin, M., Flores, R., Haber, M., Kushniarevich, A., Khachatryan, Z., Bahmanimehr, A., Parik, J., et al. (2021). Origin and diffusion of human Y chromosome haplogroup J1-M267. Sci. Rep. *11*, 6659. https://doi.org/10.1038/s41598-021-85883-2.
17. Barker, G., and Goucher, C.L. (2015). A World with Agriculture, 12,000 BCE–500 CE. In the Cambridge World History (Cambridge University Press).
18. Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. Nature *418*, 700–707. https://doi.org/10.1038/nature01019.
19. Bocquet-Appel, J.-P. (2011). When the world's population took off: the springboard of the Neolithic demographic transition. Science *333*, 560–561. https://doi.org/10.1126/science.1208880.
20. Cavalli-Sforza, L.L. (1996). The spread of agriculture and nomadic pastoralism: insights from the genetics, linguistics and archaeology. In The origins and spread of agriculture and pastoralism in Eurasia, D.R. Harris, ed. (University College London Press), pp. 51–69.
21. Zeder, M.A. (2011). The origins of agriculture in the Near East. Curr. Anthropol. *52*, S221–S235. https://doi.org/10.1086/659307.
22. Riehl, S., Zeidi, M., and Conard, N.J. (2013). Emergence of Agriculture in the Foothills of the Zagros Mountains of Iran. Science *341*, 65–67. https://doi.org/10.1126/science.1236743.
23. Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat. Commun. *6*, 8912. https://doi.org/10.1038/ncomms9912.
24. 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.
25. Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., et al. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. Proc. Natl. Acad. Sci. USA *113*, 6886–6891. https://doi.org/10.1073/pnas.1523951113.
26. Broushaki, F., Thomas, M.G., Link, V., López, S., van Dorp, L., Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-del-Molino, D., et al. (2016). Early Neolithic genomes from the eastern Fertile Crescent. Science *353*, 499–503. https://doi.org/10.1126/science.aaf7943.
27. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424. https://doi.org/10.1038/nature19310.
28. Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G., Antonović, D., et al. (2022). Ancient DNA from Mesopotamia suggests distinct pre-pottery and pottery Neolithic migrations into Anatolia. Science *377*, 982–987. https://doi.org/10.1126/science.abq0762.
29. Feldman, M., Fernández-Domínguez, E., Reynolds, L., Baird, D., Pearson, J., Hershkovitz, I., May, H., Goring-Morris, N., Benz, M., Gresky, J., et al. (2019). Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. Nat. Commun. *10*, 1218. https://doi.org/10.1038/s41467-019-09209-7.
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. https://doi.org/10.1126/science.aat7487.
31. Skourtanioti, E., Erdal, Y.S., Frangipane, M., Balossi Restelli, F., Yener, K.A., Pinnock, F., Matthiae, P., Özbal, R., Schoop, U.-D., Guliyev, F., et al. (2020). Genomic history of Neolithic to Bronze Age Anatolia, northern Levant, and southern Caucasus. Cell *181*, 1158–1175.e28. https://doi.org/10.1016/j.cell.2020.04.044.
32. Price, T.D. (2000). Europe's First Farmers (Cambridge University press).
33. Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P.J., et al. (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am. J. Hum. Genet. *74*, 1023–1034. https://doi.org/10.1086/386295.
34. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413. https://doi.org/10.1038/nature13673.
35. 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.
36. Cavalli-Sforza, L.L. (1988). The basque population and ancient migrations in Europe. Munibe *6*, 129–137.
37. Renfrew, C. (1989). The Origins of Indo-European Languages. Sci. Am. *261*, 106–114.
38. Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S.H., Hammer, M.F., Mehdi, S.Q., Ayub, Q., Qamar, R., Mohyuddin, A., Radhakrishna, U., et al. (2001). Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. Am. J. Hum. Genet. *68*, 537–542. https://doi.org/10.1086/318200.
39. Gangal, K., Sarson, G.R., and Shukurov, A. (2014). The Near-Eastern roots of the Neolithic in South Asia. PLoS One *9*, e95714. https://doi.org/10.1371/journal.pone.0095714.
40. Palanichamy, M.G., Mitra, B., Zhang, C.-L., Debnath, M., Li, G.-M., Wang, H.-W., Agrawal, S., Chaudhuri, T.K., and Zhang, Y.-P. (2015). West Eurasian mtDNA lineages in India: an insight into the spread of the Dravidian language and the origins of the caste system. Hum. Genet. *134*, 637–647. https://doi.org/10.1007/s00439-015-1547-4.
41. Parpola, A. (2015). The Roots of Hinduism: The Early Aryans and the Indus Civilization (Oxford University Press), p. 1. https://doi.org/10.1093/acprof:oso/9780190226909.001.0001.
42. Kivisild, T., Bamshad, M.J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W.S., Dixon, M.E., et al. (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr. Biol. *9*, 1331–1334. https://doi.org/10.1016/S0960-9822(00)80057-3.
43. Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.-V., Stepanov, V., et al. (2003). The genetic heritage of the earliest settlers persists both in Indian Tribal and Caste populations. Am. J. Hum. Genet. *72*, 313–332. https://doi.org/10.1086/346068.
44. Fuller, D. (2003). An agricultural perspective on Dravidian historical linguistics: Archaeological crop packages, livestock and Dravidian crop vocabulary. In Examining the farming/language dispersal hypothesis, P. Bellwood and C. Renfrew, eds. (McDONALD INSTITUTE MONOGRAPHS), pp. 191–213.
45. Fuller, D.Q. (2006). Agricultural origins and frontiers in South Asia: A working synthesis. J. World Prehist. *20*, 1–86. https://doi.org/10.1007/s10963-006-9006-8.
46. Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.-E.T., Lin, A.A., Mitra, M., Sil, S.K., Ramesh, A., et al. (2006). Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am. J. Hum. Genet. *78*, 202–221. https://doi.org/10.1086/499411.
47. Kivisild, T., Rootsi, S., Metspalu, M., Metspalu, E., Parik, J., Kaldma, K., Usanga, E., Mastana, S., Papiha, S.S., and Villems, R. (2003). The genetics of language and farming spread in India. In Examining the farming/language dispersal hypothesis, P. Bellwood and C. Renfrew, eds. (MCDONALD INSTITUTE MONOGRAPHS), pp. 215–222.
48. Shinde, V., Narasimhan, V.M., Rohland, N., Mallick, S., Mah, M., Lipson, M., Nakatsuka, N., Adamski, N., Broomandkhoshbacht, N., Ferry, M., et al. (2019). An ancient Harappan genome lacks ancestry from Steppe pastoralists or Iranian farmers. Cell *179*, 729–735.e10. https://doi.org/10.1016/j.cell.2019.08.048.
49. 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. https://doi.org/10.1038/nature23310.
50. 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. https://doi.org/10.1126/science.aar7711.
51. Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G.,

Antonović, D., et al. (2022). The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science *377*, eabm4247. https://doi.org/10.1126/science.abm4247.
52. Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C.J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc. Biol. Sci. *276*, 2703–2710. https://doi.org/10.1098/rspb.2009.0408.
53. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211. https://doi.org/10.1038/nature14317.
54. 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.
55. Chaubey, G., Metspalu, M., Kivisild, T., and Villems, R. (2007). Peopling of South Asia: investigating the caste-tribe continuum in India. Bioessays *29*, 91–100. https://doi.org/10.1002/bies.20525.
56. Majumder, P.P., and Basu, A. (2014). A genomic view of the peopling and population structure of India. Cold Spring Harb. Perspect. Biol. *7*, a008540. https://doi.org/10.1101/cshperspect.a008540.
57. Pathak, A.K. (2021). Delineating Genetic Ancestries of People of the Indus Valley, Parsis, Indian Jews and Tharu Tribe (PhD thesis). Institute of Molecular and Cell Biology (Tartu, Estonia: University of Tartu).
58. Witzel, M. (2005). Central Asian roots and acculturation in South Asia: Linguistic and archaeological evidence from western Central Asia, the Hindukush and northwestern South Asia for early Indo-Aryan language and religion. In Linguistics, Archaeology and the Himan Past: Occasional Paper 1 Indus Project, T. Osada, ed. (Research Institute for Humanity and Nature), pp. 87–211.
59. Fuller, D.Q. (2007). Non-human genetics, agricultural origins and historical linguistics in South Asia. In The Evolution and History of Human Populations in South Asia, M.D. Petraglia and B. Allchin, eds. (Springer), pp. 393–443.
60. McAlpin, D.W. (1981). Proto-Elamo-Dravidian: The Evidence and Its Implications. Trans. Am. Phil. Soc. *71*, 1. https://doi.org/10.2307/1006352.
61. Stolper, M.W. (2008). Elamite. In The Ancient Languages of Mesopotamia, Egypt, and Aksum, R.D. Woodard, ed. (Cambridge University Press), pp. 47–82.
62. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489–494. https://doi.org/10.1038/nature08365.
63. Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. (2023). On the limits of fitting complex models of population history to f-statistics. Elife *12*, e85492. https://doi.org/10.7554/eLife.85492.
64. Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L., and Oefner, P.J. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. *7*, 996–1005. https://doi.org/10.1101/gr.7.10.996.
65. Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., et al. (2000). Y chromosome sequence variation and the history of human populations. Nat. Genet. *26*, 358–361. https://doi.org/10.1038/81685.
66. R CoreTeam (2024). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
67. Posit team (2024). RStudio: Integrated Development Environment for R (Posit Software, PBC)).
68. Almarri, M.A., Haber, M., Lootah, R.A., Hallast, P., Al Turki, S., Martin, H.C., Xue, Y., and Tyler-Smith, C. (2021). The genomic history of the Middle East. Cell *184*, 4612–4625.e14. https://doi.org/10.1016/j.cell.2021.07.013.
69. Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., Metspalu, M., et al. (2006). A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. Proc. Natl. Acad. Sci. USA *103*, 843–848. https://doi.org/10.1073/pnas.0507714103.
70. Arunkumar, G., Soria-Hernanz, D.F., Kavitha, V.J., Arun, V.S., Syama, A., Ashokan, K.S., Gandhirajan, K.T., Vijayakumar, K., Narayanan, M., Jayalakshmi, M., et al. (2012). Population differentiation of southern Indian male lineages correlates with agricultural expansions predating the caste system. PLoS One *7*, e50269. https://doi.org/10.1371/journal.pone.0050269.
71. Tariq, M., Ahmad, H., Hemphill, B.E., Farooq, U., and Schurr, T.G. (2022). Contrasting maternal and paternal genetic histories among five ethnic groups from Khyber Pakhtunkhwa, Pakistan. Sci. Rep. *12*, 1027. https://doi.org/10.1038/s41598-022-05076-3.
72. Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., and Mehdi, S.Q. (2002). Y-chromosomal DNA variation in Pakistan. Am. J. Hum. Genet. *70*, 1107–1124. https://doi.org/10.1086/339929.
73. Silva, M., Oliveira, M., Vieira, D., Brandão, A., Rito, T., Pereira, J.B., Fraser, R.M., Hudson, B., Gandini, F., Edwards, C., et al. (2017). A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. BMC Evol. Biol. *17*, 88. https://doi.org/10.1186/s12862-017-0936-9.
74. Allentoft, M.E., Sikora, M., Refoyo-Martínez, A., Irving-Pease, E.K., Fischer, A., Barrie, W., Ingason, A., Stenderup, J., Sjögren, K.-G., Pearson, A., et al. (2022). Population genomics of Stone Age Eurasia. bioRxiv. https://doi.org/10.1101/2022.05.04.490594.
75. Hallast, P., Batini, C., Zadik, D., Maisano Delser, P., Wetton, J.H., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Destro Bisol, G., Dupuy, B.M., et al. (2015). The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. Mol. Biol. Evol. *32*, 661–673. https://doi.org/10.1093/molbev/msu327.
76. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445–449. https://doi.org/10.1038/nature13810.
77. Balanovsky, O., Dibirova, K., Dybo, A., Mudrak, O., Frolova, S., Pocheshkhova, E., Haber, M., Platt, D., Schurr, T., Haak, W., et al. (2011). Parallel evolution of genes and languages in the Caucasus region. Mol. Biol. Evol. *28*, 2905–2920. https://doi.org/10.1093/molbev/msr126.
78. Yunusbayev, B., Metspalu, M., Järve, M., Kutuev, I., Rootsi, S., Metspalu, E., Behar, D.M., Varendi, K., Sahakyan, H., Khusainova, R., et al. (2012). The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol. Biol. Evol. *29*, 359–365. https://doi.org/10.1093/molbev/msr221.
79. Zhabagin, M., Wei, L.-H., Sabitov, Z., Ma, P.-C., Sun, J., Dyussenova, Z., Balanovska, E., Li, H., and Ramankulov, Y. (2022). Ancient components and recent expansion in the Eurasian heartland: Insights into the revised phylogeny of Y-chromosomes from Central Asia. Genes *13*, 1776. https://doi.org/10.3390/genes13101776.
80. Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. *4*, vey016. https://doi.org/10.1093/ve/vey016.
81. Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., and Lemey, P. (2016). SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol. Biol. Evol. *33*, 2167–2169. https://doi.org/10.1093/molbev/msw082.
82. Antonio, M.L., Weiß, C.L., Gao, Z., Sawyer, S., Oberreiter, V., Moots, H.M., Spence, J.P., Cheronet, O., Zagorc, B., Praxmarer, E., et al. (2023). Stable population structure in Europe since the Iron Age, despite high mobility. bioRxiv. https://doi.org/10.1101/2022.05.15.491973.
83. Damgaard, P.d.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.
84. 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.
85. Margaryan, A., Lawson, D.J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., Cassidy, L.M., Jørsboe, E., Ingason, A., Pedersen, M.W., et al. (2020). Population genomics of the Viking world. Nature *585*, 390–396. https://doi.org/10.1038/s41586-020-2688-8.
86. Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., Cheronet, O., Ferry, M., Harney, E., Keating, D., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nat. Commun. *10*, 590. https://doi.org/10.1038/s41467-018-08220-8.
87. Feldman, M., Master, D.M., Bianco, R.A., Burri, M., Stockhammer, P.W., Mittnik, A., Aja, A.J., Jeong, C., and Krause, J. (2019).

Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. Sci. Adv. *5*, eaax0061. https://doi.org/10.1126/sciadv.aax0061.

88. Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G., Antonović, D., et al. (2022). A genetic probe into the ancient and medieval history of Southern Europe and West Asia. Science *377*, 940–951. https://doi.org/10.1126/science.abq0755.
89. Skourtanioti, E., Ringbauer, H., Gnecchi Ruscone, G.A., Bianco, R.A., Burri, M., Freund, C., Furtwängler, A., Gomes Martins, N.F., Knolle, F., Neumann, G.U., et al. (2023). Ancient DNA reveals admixture history and endogamy in the prehistoric Aegean. Nat. Ecol. Evol. *7*, 290–303. https://doi.org/10.1038/s41559-022-01952-3.
90. Asouti, E., Kabukcu, C., Swinson, K., and Martin, L. (2023). Environment and subsistence in the Zagros epipalaeolithic. In The Epipalaeolithic and Neolithic in the Eastern Fertile Crescent (Routledge), pp. 106–118. https://doi.org/10.4324/9781003335504-9.
91. Ilumäe, A.-M., Post, H., Flores, R., Karmin, M., Sahakyan, H., Mondal, M., Montinaro, F., Saag, L., Bormans, C., Sanchez, L.F., et al. (2021). Phylogenetic history of patrilineages rare in northern and eastern Europe from large-scale re-sequencing of human Y-chromosomes. Eur. J. Hum. Genet. *29*, 1510–1519. https://doi.org/10.1038/s41431-021-00897-8.
92. Kaja, E., Lejman, A., Sielski, D., Sypniewski, M., Gambin, T., Dawidziuk, M., Suchocki, T., Golik, P., Wojtaszewska, M., Mroczek, M., et al. (2022). The Thousand Polish Genomes—A Database of Polish Variant Allele Frequencies. Int. J. Mol. Sci. *23*, 4532. https://doi.org/10.3390/ijms23094532.
93. Pathak, A.K., Kadian, A., Kushniarevich, A., Montinaro, F., Mondal, M., Ongaro, L., Singh, M., Kumar, P., Rai, N., Parik, J., et al. (2018). The genetic ancestry of modern Indus Valley populations from Northwest India. Am. J. Hum. Genet. *103*, 918–929. https://doi.org/10.1016/j.ajhg.2018.10.022.
94. Parpola, A. (1994). Deciphering the Indus Script (Cambridge University Press).
95. Krishnamurti, B. (2003). The Dravidian Languages (Cambridge University Press).
96. Ansumali Mukhopadhyay, B. (2021). Ancestral Dravidian languages in Indus Civilization: ultraconserved Dravidian tooth-word reveals deep linguistic ancestry and supports genetics. Humanit. Soc. Sci. Commun. *8*, 193. https://doi.org/10.1057/s41599-021-00868-w.
97. Ruhlen, M. (1991). A Guide to the World's Languages: Classification (Stanford University Press).
98. Giesche, A., Hodell, D.A., Petrie, C.A., Haug, G.H., Adkins, J.F., Plessen, B., Marwan, N., Bradbury, H.J., Hartland, A., French, A.D., and Breitenbach, S.F.M. (2023). Recurring summer and winter droughts from 4.2-3.97 thousand years ago in north India. Commun. Earth Environ. *4*, 103. https://doi.org/10.1038/s43247-023-00763-z.
99. University of Tartu (2018). UT Rocket. Preprint at share.neic.no. https://doi.org/10.23673/PH6N-0144.
100. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
101. Kars, M.E., Başak, A.N., Onat, O.E., Bilguvar, K., Choi, J., Itan, Y., Çağlar, C., Palvadeau, R., Casanova, J.-L., Cooper, D.N., et al. (2021). The genetic structure of the Turkish population reveals high levels of variation and admixture. Proc. Natl. Acad. Sci. USA *118*. e2026076118. https://doi.org/10.1073/pnas.2026076118.
102. Wong, L.-P., Lai, J.K.-H., Saw, W.-Y., Ong, R.T.-H., Cheng, A.Y., Pillai, N.E., Liu, X., Xu, W., Chen, P., Foo, J.-N., et al. (2014). Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. PLoS Genet. *10*, e1004377. https://doi.org/10.1371/journal.pgen.1004377.
103. Yang, X.-Y., Rakha, A., Chen, W., Hou, J., Qi, X.-B., Shen, Q.-K., Dai, S.-S., Sulaiman, X., Abdulloevich, N.T., Afanasevna, M.E., et al. (2021). Tracing the genetic legacy of the Tibetan Empire in the Balti. Mol. Biol. Evol. *38*, 1529–1536. https://doi.org/10.1093/molbev/msaa313.
104. Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., et al. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Res. *26*, 151–162. https://doi.org/10.1101/gr.191478.115.
105. Carmi, S., Hui, K.Y., Kochav, E., Liu, X., Xue, J., Grady, F., Guha, S., Upadhyay, K., Ben-Avraham, D., Mukherjee, S., et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat. Commun. *5*, 4835. https://doi.org/10.1038/ncomms5835.
106. Haber, M., Mezzavilla, M., Bergström, A., Prado-Martinez, J., Hallast, P., Saif-Ali, R., Al-Habori, M., Dedoussis, G., Zeggini, E., Blue-Smith, J., et al. (2016). Chad genetic diversity reveals an African history marked by multiple Holocene Eurasian migrations. Am. J. Hum. Genet. *99*, 1316–1324. https://doi.org/10.1016/j.ajhg.2016.10.012.
107. Gilly, A., Suveges, D., Kuchenbaecker, K., Pollard, M., Southam, L., Hatzikotoulas, K., Farmaki, A.-E., Bjornland, T., Waples, R., Appel, E.V.R., et al. (2018). Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. Nat. Commun. *9*, 4674. https://doi.org/10.1038/s41467-018-07070-8.
108. Serra-Vidal, G., Lucas-Sanchez, M., Fadhlaoui-Zid, K., Bekada, A., Zalloua, P., and Comas, D. (2019). Heterogeneity in Palaeolithic population continuity and Neolithic expansion in North Africa. Curr. Biol. *29*, 3953–3959.e4. https://doi.org/10.1016/j.cub.2019.09.050.
109. Alsmadi, O., John, S.E., Thareja, G., Hebbar, P., Antony, D., Behbehani, K., and Thanaraj, T.A. (2014). Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from Kuwaiti population subgroup of inferred Saudi Arabian tribe ancestry. PLoS One *9*, e99069. https://doi.org/10.1371/journal.pone.0099069.
110. Ilyas, M., Kim, J.-S., Cooper, J., Shin, Y.-A., Kim, H.-M., Cho, Y.S., Hwang, S., Kim, H., Moon, J., Chung, O., et al. (2015). Whole genome sequencing of an ethnic Pathan (Pakhtun) from the north-west of Pakistan. BMC Genom. *16*, 172. https://doi.org/10.1186/s12864-015-1290-1.
111. 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.
112. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
113. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.
114. Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.
115. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.
116. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
117. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
118. Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P., et al. (2012). BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. Syst. Biol. *61*, 170–173. https://doi.org/10.1093/sysbio/syr100.
119. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.
120. Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214. https://doi.org/10.1186/1471-2148-7-214.
121. Relethford, J.H. (2008). Geostatistics and spatial analysis in biological anthropology. Am. J. Phys. Anthropol. *136*, 1–10. https://doi.org/10.1002/ajpa.20789.
122. Martiniano, R., De Sanctis, B., Hallast, P., and Durbin, R. (2022). Placing ancient DNA sequences into reference phylogenies. Mol. Biol. Evol. *39*, msac017. https://doi.org/10.1093/molbev/msac017.
123. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual (CreateSpace Press).
124. Behar, D.M., Saag, L., Karmin, M., Gover, M.G., Wexler, J.D., Sanchez, L.F.,

Greenspan, E., Kushniarevich, A., Davydenko, O., Sahakyan, H., et al. (2017). The genetic variation in the R1a clade among the Ashkenazi Levites' Y chromosome. Sci. Rep. *7*, 14969. https://doi.org/10.1038/s41598-017-14761-7.

125. Poznik, G.D., Henn, B.M., Yee, M.-C., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A., and Bustamante, C.D. (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science *341*, 562–565. https://doi.org/10.1126/science.1237619.
126. Drummond, A.J., Ho, S.Y.W., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biol. *4*, e88. https://doi.org/10.1371/journal.pbio.0040088.
127. Tavaré, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences (American Mathematical Society: Lectures on Mathematics in the Life Sciences (Amer Mathematical Society)), pp. 57–86.
128. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306–314. https://doi.org/10.1007/BF00160154.
129. Drummond, A.J., Rambaut, A., Shapiro, B., and Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. *22*, 1185–1192. https://doi.org/10.1093/molbev/msi103.
130. Heled, J., and Drummond, A.J. (2008). Bayesian inference of population size history from multiple loci. BMC Evol. Biol. *8*, 289. https://doi.org/10.1186/1471-2148-8-289.
131. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. *8*, 28–36. https://doi.org/10.1111/2041-210X.12628.
132. Yu, G., Lam, T.T.-Y., Zhu, H., and Guan, Y. (2018). Two Methods for mapping and visualizing associated data on phylogeny using ggtree. Mol. Biol. Evol. *35*, 3041–3043. https://doi.org/10.1093/molbev/msy194.
133. Wang, L.-G., Lam, T.T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C.W., Jones, B.R., Bradley, T., et al. (2020). Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol. Biol. Evol. *37*, 599–603. https://doi.org/10.1093/molbev/msz240.
134. Wickham, H. (2007). Reshaping data with the reshape package. J. Stat. Softw. *21*, 1–20. https://doi.org/10.18637/jss.v021.i12.
135. Wickham, H. (2009). Ggplot2: Elegant Graphics for Data Analysis (Springer).
136. Henry, L., Wickham, H., and Chang, W. (2019). ggstance: Horizontal "ggplot2" Components.
137. Lemey, P., Rambaut, A., Welch, J.J., and Suchard, M.A. (2010). Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time. Mol. Biol. Evol. *27*, 1877–1885. https://doi.org/10.1093/molbev/msq067.
138. Pybus, O.G., Suchard, M.A., Lemey, P., Bernardin, F.J., Rambaut, A., Crawford, F.W., Gray, R.R., Arinaminpathy, N., Stramer, S.L., Busch, M.P., and Delwart, E.L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc. Natl. Acad. Sci. USA *109*, 15066–15071. https://doi.org/10.1073/pnas.1206598109.
139. Faria, N.R., Kraemer, M.U.G., Hill, S.C., Goes de Jesus, J., Aguiar, R.S., Iani, F.C.M., Xavier, J., Quick, J., du Plessis, L., Dellicour, S., et al. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science *361*, 894–899. https://doi.org/10.1126/science.aat7115.
140. Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Philos. Mag. A *4*, 161–173. https://doi.org/10.1080/14786442808674769.
141. Fenner, J.N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. *128*, 415–423. https://doi.org/10.1002/ajpa.20188.
142. Wang, R.J., Al-Saffar, S.I., Rogers, J., and Hahn, M.W. (2023). Human generation times across the past 250,000 years. Sci. Adv. *9*, eabm7047. https://doi.org/10.1126/sciadv.abm7047.
143. Y Chromosome Consortium (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. *12*, 339–348. https://doi.org/10.1101/gr.217602.
144. Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., and Hammer, M.F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. *18*, 830–838. https://doi.org/10.1101/gr.7172008.
145. Miller, S.A., Dykes, D.D., and Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. *16*, 1215. https://doi.org/10.1093/nar/16.3.1215.
146. Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R (Springer-Verlag).
147. Bivand, R.S., Pebesma, E.J., and Virgilio, G.-R. (2013). Applied Spatial Data Analysis with R, Second edition (Springer).
148. Pebesma, E.J., and Bivand, R.S. (2005). Classes and methods for spatial data in R. R. News *5*, 9–13.
149. Hijmans, R.J. (2023). Raster: Geographic data analysis and modeling. R project.
150. Bivand, R.S., Keitt, T., and Rowlingson, B. (2019). Rgdal: Bindings for the "Geospatial" Data Abstraction Library. R project.
151. Bivand, R.S., and Rundel, C. (2019). Rgeos: Interface to Geometry Engine - Open Source ('GEOS'). R project.
152. Bivand, R.S. (2023). ClassInt: Choose Univariate Class Intervals. R project.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw (FASTQ) and reference mapped (BAM) reads	This study	ENA: PRJEB71943
Supplemental Figures, Tables, and Files	This study	Mendeley Data: https://doi.org/10.17632/ts4vc55rzp.1
Human reference genome NCBI build 37, GRCh37, decoy version	The 1000 Genomes Project Consortium et al.[100]	https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
Raw (FASTQ) or (BAM/CRAM) reads	The 1000 Genomes Project Consortium et al.[100]	https://www.ebi.ac.uk/ena/browser/view/PRJEB31736
Raw (FASTQ) or (BAM/CRAM) reads	Bergström et al.[13]	https://www.ebi.ac.uk/ena/browser/view/PRJEB6463
Raw (FASTQ) or (BAM/CRAM) reads	Kars et al.[101]	https://www.ebi.ac.uk/ena/browser/view/PRJNA674530
Raw (FASTQ) or (BAM/CRAM) reads	Mallick et al.[11]	https://www.ebi.ac.uk/ena/browser/view/PRJEB9586
Raw (FASTQ) or (BAM/CRAM) reads	Wong et al.[102]	https://www.ncbi.nlm.nih.gov/biosample/?term=SS6003405; https://www.ncbi.nlm.nih.gov/biosample/?term=SS6003411; https://www.ncbi.nlm.nih.gov/biosample/?term=SS6003413; https://www.ncbi.nlm.nih.gov/biosample/?term=SS6003435
Raw (FASTQ) or (BAM/CRAM) reads	Yang et al.[103]	https://ngdc.cncb.ac.cn/gsa/search?searchTerm=PRJCA000457
Raw (FASTQ) or (BAM/CRAM) reads	Karmin et al.[3]	https://www.ebi.ac.uk/ena/browser/view/PRJEB8108
Raw (FASTQ) or (BAM/CRAM) reads	Rodriguez-Flores et al.[104]	https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR2098261&display=download
Raw (FASTQ) or (BAM/CRAM) reads	Carmi et al.[105]	https://ega-archive.org/search/egad00001000781
Raw (FASTQ) or (BAM/CRAM) reads	Haber et al.[106]; Gilly et al.[107]	https://ega-archive.org/datasets/EGAD00001001440
Raw (FASTQ) or (BAM/CRAM) reads	Serra-Vidal et al.[108]	https://www.ebi.ac.uk/ena/browser/view/PRJEB29142
Raw (FASTQ) or (BAM/CRAM) reads	Alsmadi et al.[109]	https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR1274979&display=download
Raw (FASTQ) or (BAM/CRAM) reads	Ilyas et al.[110]	https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR926184&display=download
Raw (FASTQ) or (BAM/CRAM) reads	Kaja et al.[92]	The Thousand Polish Genomes: 633_28990_20
Raw (FASTQ) or (BAM/CRAM) reads	Almarri et al.[68]	https://www.ebi.ac.uk/ena/browser/view/PRJEB28504
Software and algorithms		
SAMtools v1.9	Li et al.[111]	http://www.htslib.org
BEDtools v2.24	Quinlan and Hall[112]	https://bedtools.readthedocs.io/en/latest
BWA-MEM v0.7.17	Li and Durbin[113]; Li[114]	http://bio-bwa.sourceforge.net
Picard-tools-2.0.1	N/A	http://broadinstitute.github.io/picard
GATK-3.5	McKenna et al.[115]	https://gatk.broadinstitute.org/hc/en-us
BCFtools v1.6	Danecek et al.[116]	https://www.htslib.org
R v4.3.3	R CoreTeam[66]	https://cran.r-project.org/bin/windows/base/
RStudio Build 402	RStudio Team[67]	
BEAST v1.10.4	Suchard et al.[80]	https://beast.community
spreaD3 v0.9.7.1rc	Bielejec et al.[81]	https://rega.kuleuven.be/cev/ecv/software/SpreaD3
RAxML v7.3.2	Stamatakis[117]	https://github.com/stamatak/standard-RAxML
BEAGLE library v3.1.2	Ayres et al.[118]	https://beast.community/beagle
Tracer v1.7	Rambaut et al.[119]	https://beast.community/tracer
LogCombiner v1.10.4	Drummond and Rambaut[120]	https://beast.community/logcombiner

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TreeAnnotator v1.10.4	Drummond and Rambaut[120]	https://beast.community/treeannotator
Surfer v8	Relethford[121]	https://www.goldensoftware.com/products/surfer/
pathPhynder	Martiniano et al.[122]	https://github.com/ruidlpm/pathPhynder
Python v3.8.0	Van Rossum and Drake[123]	https://www.python.org
FigTree v1.4.4	N/A	http://tree.bio.ed.ac.uk/software/figtree/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Hovhannes Sahakyan (hovhannes.sahakyan@ut.ee).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The whole Y chromosome high-coverage sequence data shared by individuals (forty-five) or generated in the current study (one) are deposited in the European Nucleotide Archive at EMBL-EBI (https://www.ebi.ac.uk/ena/browser/view) under accession ENA: PRJEB71943. They are publicly available as of the date of publication. This paper also analyzes existing, publicly available data. Accession numbers for the datasets are listed in the key resources table. Following the consent form signed by the customers of Gene by Gene commercial genetic testing company, the sequencing data included in this study is used for the sole purpose of scientific inquiry. Restrictions apply to the availability of these data, so they are not publicly available. It is reported here on an aggregate level as phylogenetic trees. In addition, the original MCC tree is provided as File S1, and the original ML tree is available as File S2. These text files can be opened using FigTree software (http://tree.bio.ed.ac.uk/software/figtree/), enabling researchers to explore the phylogenetic relationships in a user-friendly manner. Raw data from Supplemental Figures, Tables, and Files were deposited on Mendeley at Mendeley Data: https://doi.org/10.17632/ts4vc55rzp.1.

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 STUDY PARTICIPANT DETAILS

Whole high-coverage Y chromosome sequences

In this study, all subjects included were adult males. In the phylogenetic reconstructions, we included 64 high-coverage whole Y chromosome sequences that had not been reported in academic publications. They have been all sequenced with the Illumina HiSeq 2500 platform following Y chromosome capture using a proprietary capture protocol available at Gene by Gene (Family Tree DNA) using the commercially available "BigY" service (https://learn.familytreedna.com/wp-content/uploads/2014/08/BIG_Y_WhitePager.pdf). Its targeted enrichment design utilizes 67,000 capture probes for sequencing more than 10 Mbp in the non-recombining male-specific parts of the Y chromosome at > 60× coverage.

All participants were informed about the study's purpose and provided informed consent for their data to be used in scientific inquiry. Many of these genomes were found with the help of the Yfull and FTDNA. Ten of the high-coverage whole Y chromosome genomes belonged to individuals of haplogroup L1-M22 out of a total of 2018 male donors with self-reported ancestry from various countries, including Finland, Germany, Latvia, Lithuania, Poland, the Russian Federation, Sweden, and Ukraine. Notably, the haplogroup L1-M22 individuals in this study were from the Russian Federation. Additionally, ten genomes were obtained from individuals with Jewish ancestry. These twenty sequences were provided in scientific collaboration with the commercial genetic testing company Gene by Gene, based in Houston, Texas, USA. One more sample from our laboratory collection was sequenced using the aforementioned "BigY" service. We followed the approved guidelines by the Research Ethics Committee of the University of Tartu, and all experimental protocols were approved by the Research Ethics Committee of the University of Tartu (252/M-17). Table S2 contains detailed information on the samples analyzed in this study.

Published whole high-coverage Y chromosomes

From the published sources, we collected 99 high-coverage Y chromosome genomes sequenced with next-generation sequencing technologies targeting over 9 Mb regions of the chromosome (Table S2).[3,11,13,68,92,100–110] If fastq reads were unavailable in the public repositories, these were extracted from BAM or CRAM genome files using SAMtools v1.9[111] and BEDtools v2.24.[112]

Genotyping of SNP markers

We collected blood specimens from 707 healthy unrelated adult males who belong to three distinct West Asian populations. Patrilineal ancestors of the individuals for at least two generations belong to the populations reported here. Prior to the study, informed consent was obtained from all participants. All experimental procedures were carried out following the approved guidelines by the Research Ethics Committee of the University of Tartu. All experimental protocols were approved by the Research Ethics Committee of the University of Tartu (252/M-17).

METHOD DETAILS

This study defines West Asia as including the Iranian Plateau, Mesopotamia, the Armenian Highland, the Caucasus, Anatolia, the Levant, and the Arabian Peninsula. In Table S1, we listed the populations from the Caucasus and other West Asian regions separately for ease of reading. Next, we use Anatolia with its geographic definition, marking the area west of the Anatolian diagonal.

Reads mapping and multi-sample variants calling

These were performed on 97 genomes sequenced with Illumina technology following the previously described protocols.[124] The reads were mapped to the GRCh37 human reference assembly, decoy version (obtained from the 1000 Genomes Project[100]) using BWA-MEM.[113,114] Duplicate reads were removed with Picard-tools-2.0.1 (http://broadinstitute.github.io/picard), and indel realignment was performed with GATK-3.5.[115] The multi-sample base calling was carried out using SAMtools[111] and BCFtools v1.6.[116] All 97 Y chromosome genomes were mapped and called, starting with the raw fastq reads using the same script. These variants were merged[16,124] with those called from two high-coverage haplogroup L1-M22 Y chromosomes sequenced using Complete Genomics technology (Mountain View, California) (Table S2).

Variant filtering

The region mask we used is detailed in our earlier study.[16] It is based on the published regions[125] and supplemented with high-quality regions.[3] The latter approach minimizes platform bias after datasets are merged. We have excluded the regions (i) containing two or more subsequent singletons within 50 base pairs, (ii) having discrepant SNPs between genomes of the same individuals sequenced more than once or between paternally related individuals, (iii) having recurrent SNPs in three or more branches in the phylogeny composed of samples sequenced with the same platform, (iv) with missing data in more than 10% of samples. Moreover, we have also excluded the regions between two > 10% no-call sites if they are placed nearby and lack variants. In the end, we recovered 9,514,762 bases of the male-specific region of the Y chromosome (Table S3). We did not perform imputation because we expect platform-specific differences to be negligible with this region mask. Consistent with this, we observe a low variation of mutation rates among the branches in our initial Bayesian phylogenetic analysis with an uncorrelated relaxed clock model[126] (Coefficient of variation = 0.0355, 95% highest posterior density (HPD) = 0.00003–0.0847).

Tree reconstructions and coalescence analysis

The phylogeny was reconstructed using maximum-likelihood (ML) and Bayesian Markov Chain Monte Carlo (MCMC) approaches. The ML reconstruction was performed using RAxML software version 7.3.2,[117] with the generalized time-reversible (GTR) substitution matrix[127] and rapid bootstrapping (n=200), followed by a subsequent ML search. Eleven members from different Y chromosome haplogroups were included to ensure proper rooting of the haplogroup L1-M22's most recent common ancestor. All identified variants were annotated based on this phylogeny using in-house scripts and subsequently curated manually. Figure S1 displays the ML tree, while Table S4 lists the polymorphic positions and their corresponding annotations.

Coalescence time estimates were determined in the tree reconstruction with the Bayesian MCMC approach implemented in BEAST v1.10.4 software.[80] To ensure proper rooting, we included nine members from other Y chromosome haplogroups. Three parallel analyses were run with different random number seeds, each with 50 million chains. Coalescence time estimates were inferred by providing a normal prior with a mean of 45610 and a standard deviation of 2300 to the LT node, based on a previously published estimate.[3] We used a non-informative, uniformly distributed ($1.0e^{-20}$ – 1.0) prior for mutation rate. For the site model, we used the GTR substitution model[127] and the Gamma site heterogeneity model[128] with four categories. Bayesian skyline model[129] was used as the tree model with group sizes of 5. Population dynamics were smoothed using a piecewise-linear approach.[130] Uniform distribution bounded by 1.0 and $1.0e^{15}$ was given as the skyline.popSize prior. The uncorrelated relaxed log-normal clock model[126] was initially used (data not shown). However, all subsequent analyses were run with the Strict clock model, as variation in mutation rates between branches was negligible (0.0355, 95% HPD = 0.00003–0.0847). In the analyses, we used BEAGLE library v3.1.2[118] for accelerated, parallel likelihood evaluation.

The results were manually inspected using Tracer v1.7 software,[119] with satisfactory convergence being achieved as evidenced by effective sample size (ESS) values exceeding 200 for all parameters. The results of the parallel chains were combined using LogCombiner software,[120] with a burn-in of the first 10% of records discarded. The maximum clade credibility (MCC) tree was generated with TreeAnnotator,[120] with node heights being summarized using posterior median values. Lastly, the MCC tree was visualized in RStudio software,[66,67] including "ggtree",[131,132] "ape", "treeio",[133] "reshape2",[134] "ggplot2",[135] and "ggstance",[136] in addition to basic R packages. These procedures ensured the reliability of the results and facilitated the production of a clear and reproducible visualization of the MCC tree.

Bayesian phylogeography

We conducted a Bayesian phylogeographic analysis in continuous space, following established methods.[137,138] This approach was originally designed to uncover the spatial dynamics and ancestral locations of viruses in continuous space[138,139] but has also been applied to human Y chromosome studies.[16,91]

We selected 163 out of 165 whole high-coverage Y chromosome sequences. Two excluded genomes from the Russian Federation lacked population or geographic information and did not belong to the L1a2-M357/L1307 branch composed of Nakh-Dagestanian-speaking Northeast Caucasians.

We used the BEAST v1.10.4 software[80] and applied molecular clock, site, and tree models and priors similar to those used in our Bayesian phylogenetic analysis described above. To infer coalescence time information, we provided a normally distributed prior with a mean of 20,600 years and a standard deviation of 100 years to the root node. This mean value corresponds to the age we estimated for haplogroup L1-M22 in this study. For the diffusion model in continuous space, we used the Brownian random walk (BRW) model.[137,140] We ran three sets of 500,000,000 chains and ensured that ESS values were well above 200. The MCC tree was generated using a -hpd2D 0.8 flag to summarize 80% HPD area for the tree nodes. We visualized the uncertainties of the MCC tree node locations using the spreaD3_v0.9.7.1rc software,[81] with the base world map in "geojson" format downloaded from https://github.com/Stefie/geojson-world.

The demographic history reconstruction

This was performed using the Bayesian skyline analysis framework.[129] We employed a similar analysis setup to the Bayesian phylogenetic reconstruction analysis, with the exception that no outgroups were included. The dynamics of effective population size (Ne) over time were estimated using Tracer v1.7 software,[119] and we assumed five population groups. A Bayesian skyline plot was generated in RStudio software with basic R packages.[66,67] We used an average per-generation time of 31 years for human males.[141,142]

Annotation

Labeling the clades of the Y chromosome tree presents a challenge due to many available whole Y chromosome sequences and their rapid generation. Consequently, clade labels tend to be lengthy and subject to frequent changes. We follow our ML phylogenetic tree to label the clades. To simplify the labeling process, we prefer using one of the defining markers' names for clade labels instead of long alphanumeric ones. We utilized a comprehensive source of marker information from https://ybrowse.org/gbrowse2/gff, which was downloaded on 01/Dec/2022. For widely known clades, we retained the marker names previously defined in academic publications or utilized by commercial Y chromosome trees, occasionally incorporating two of these names. For other clades, we selected the shortest name among the defining markers. To assist readers, in both the main text and Figure 2, we maintain the known alphanumeric labels up to the four-symbol levels for deeper clades while also including the known marker names. For shallower clades, we preserve the four-symbol alphanumeric part of their upstream clades and add the respective marker names. Although this approach may not be ideal, we found it optimal for this study. It avoids burdening readers with long alphanumeric labels and refers to established labels of major branches while giving a reference for all clades if necessary. The annotations can be found in Table S4, which includes not only our designated labels for the clades but also references to the labels provided by the Y Chromosome Consortium[143,144] and the International Society of Genetic Genealogy v15.73 (ISOGG) (https://isogg.org/tree/index.html) when available. Additionally, SNP rs IDs corresponding to the dbSNP v156 database (https://ftp.ncbi.nih.gov/snp/) are provided for further identification.

Phylogenetically informative SNP genotyping

DNA was extracted with the published "salting out" method.[145] We genotyped M11 or M20,[64] M317,[46] M27,[65] and M357[46] SNP markers. The alleles were identified by direct Sanger sequencing or restriction fragment length polymorphism (RFLP) analysis. The results are presented in Table S1, indicating the number of individuals associated with each lineage.

Spatial frequency analyses

Spatial frequency analyses were conducted with the Surfer program (version 8, Golden Software, Inc., Golden, CO, USA), following the Kriging procedure.[121] The input data are represented in Table S1. The maps were generated using the RStudio software[66,67] with the following packages – "lattice",[146] "sp",[147,148] "raster",[149] "rgdal",[150] "rgeos",[151] and "classInt"[152] in addition to the basic packages.

Ancient L-M20 representatives' affiliation

We scanned published ancient DNA studies and found individuals who belong to the Y chromosome haplogroup L (Figure S3). The information on the samples is provided in Table S5. Their affiliation to the reconstructed phylogeny was performed by the software pathPhynder.[122] With a likelihood-based workflow in R[66] and Python3,[123] it takes advantage of all the polymorphic sites in the target sequence and effectively evaluates the number of ancestral and derived alleles present on each branch, then reports the most likely placement of an ancient sample in the phylogeny, together with alternatives and supporting evidence. Every informative position at critical branches was also manually looked at in the genome files.