



## Using Y-chromosome capture enrichment to resolve haplogroup H2 shows new evidence for a two-path Neolithic expansion to Western Europe

Adam B. Rohrlach, Luka Papac, Ainash Childebayeva, Maïté Rivollat, Vanessa Villalba-Mouco, Gunnar U. Neumann, Sandra Penske, Eirini Skourtanioti, Marieke van de Loosdrecht, Murat Akar, et al.

### ► To cite this version:

Adam B. Rohrlach, Luka Papac, Ainash Childebayeva, Maïté Rivollat, Vanessa Villalba-Mouco, et al.. Using Y-chromosome capture enrichment to resolve haplogroup H2 shows new evidence for a two-path Neolithic expansion to Western Europe. Scientific Reports, 2021, 11 (1), pp.15005. 10.1038/s41598-021-94491-z . hal-03331786

**HAL Id: hal-03331786**

**<https://hal.science/hal-03331786>**

Submitted on 25 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



OPEN

# Using Y-chromosome capture enrichment to resolve haplogroup H2 shows new evidence for a two-path Neolithic expansion to Western Europe

Adam B. Rohrlach<sup>1,2✉</sup>, Luka Papac<sup>1</sup>, Ainash Childebayeva<sup>1</sup>, Maïté Rivollat<sup>1,3</sup>, Vanessa Villalba-Mouco<sup>1,4</sup>, Gunnar U. Neumann<sup>1</sup>, Sandra Penske<sup>1</sup>, Eirini Skourtanioti<sup>1</sup>, Marieke van de Loosdrecht<sup>1</sup>, Murat Akar<sup>5</sup>, Kamen Boyadzhiev<sup>6</sup>, Yavor Boyadzhiev<sup>6</sup>, Marie-France Deguilloux<sup>3</sup>, Miroslav Dobeš<sup>7</sup>, Yilmaz S. Erdal<sup>8</sup>, Michal Ernée<sup>7</sup>, Marcella Frangipane<sup>9</sup>, Mirosław Furmanek<sup>10</sup>, Susanne Friederich<sup>11</sup>, Emmanuel Ghesquière<sup>12,13</sup>, Agata Hałuszko<sup>10,14</sup>, Svend Hansen<sup>15</sup>, Mario Küßner<sup>16</sup>, Marcello Mannino<sup>17</sup>, Rana Özbal<sup>18</sup>, Sabine Reinhold<sup>15</sup>, Stéphane Rottier<sup>3</sup>, Domingo Carlos Salazar-García<sup>19,20,21</sup>, Jorge Soler Diaz<sup>22</sup>, Philipp W. Stockhammer<sup>1,23</sup>, Consuelo Roca de Togores Muñoz<sup>22</sup>, K. Aslihan Yener<sup>24</sup>, Cosimo Posth<sup>1,25</sup>, Johannes Krause<sup>1</sup>, Alexander Herbig<sup>1</sup> & Wolfgang Haak<sup>1,26✉</sup>

Uniparentally-inherited markers on mitochondrial DNA (mtDNA) and the non-recombining regions of the Y chromosome (NRY), have been used for the past 30 years to investigate the history of humans from a maternal and paternal perspective. Researchers have preferred mtDNA due to its abundance in the cells, and comparatively high substitution rate. Conversely, the NRY is less susceptible to back mutations and saturation, and is potentially more informative than mtDNA owing to its longer sequence length. However, due to comparatively poor NRY coverage via shotgun sequencing, and the relatively low and biased representation of Y-chromosome variants on capture assays such as the

<sup>1</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany. <sup>2</sup>ARC Centre of Excellence for Mathematical and Statistical Frontiers, School of Mathematical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia. <sup>3</sup>Université de Bordeaux, CNRS, PACEA-UMR 5199, 33615 Pessac, France. <sup>4</sup>Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, Spain. <sup>5</sup>Department of Archaeology, Mustafa Kemal University, 31060 Alahan-Antakya, Hatay, Turkey. <sup>6</sup>National Institute of Archaeology with Museum, Bulgarian Academy of Sciences, 1000 Sofia, Bulgaria. <sup>7</sup>Department of Prehistory, Institute of Archaeology CAS, Prague, Czech Republic. <sup>8</sup>Department of Anthropology, Hacettepe University, 06800 Ankara, Turkey. <sup>9</sup>Department of Classics, Sapienza University of Rome, 00185 Rome, Italy. <sup>10</sup>Institute of Archaeology, University of Wrocław, Wrocław, Poland. <sup>11</sup>State Office for Heritage Management and Archaeology Saxony-Anhalt and State Museum of Prehistory, Halle, Germany. <sup>12</sup>Inrap Grand Ouest, Bourguébus, France. <sup>13</sup>Université de Rennes 1, CNRS, CReAAH-UMR, 6566 Rennes, France. <sup>14</sup>Archeolodzy.org Foundation, Wrocław, Poland. <sup>15</sup>Eurasia Department, German Archaeological Institute, Berlin, Germany. <sup>16</sup>Thuringian State Office for Heritage Management and Archeology, Weimar, Germany. <sup>17</sup>Department of Archaeology, School of Culture and Society, Aarhus University, 8270 Højbjerg, Denmark. <sup>18</sup>Department of Archaeology and History of Art, Koç University, 34450 Istanbul, Turkey. <sup>19</sup>Grupo de Investigación en Prehistoria IT-1223-19 (UPV-EHU)/IKERBASQUE-Basque Foundation for Science, Vitoria, Spain. <sup>20</sup>Departament de Prehistòria, Arqueologia i Història Antiga, Universitat de València, Valencia, Spain. <sup>21</sup>Department of Geological Sciences, University of Cape Town, Cape Town, South Africa. <sup>22</sup>MARQ Archaeological Museum of Alicante, Alicante, Spain. <sup>23</sup>Ludwig Maximilian University Munich, 80799 Munich, Germany. <sup>24</sup>Institute for the Study of the Ancient World (ISAW), New York University, New York, NY 10028, USA. <sup>25</sup>Archaeo- and Palaeogenetics Group, Institute for Archaeological Sciences Eberhard Karls University Tübingen, 72070 Tübingen, Germany. <sup>26</sup>School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia. ✉email: rohrlach@shh.mpg.de; haak@shh.mpg.de

**1240 k, ancient DNA studies often fail to utilize the unique perspective that the NRY can yield. Here we introduce a new DNA enrichment assay, coined YMCA (Y-mappable capture assay), that targets the "mappable" regions of the NRY. We show that compared to low-coverage shotgun sequencing and 1240 k capture, YMCA significantly improves the mean coverage and number of sites covered on the NRY, increasing the number of Y-haplogroup informative SNPs, and allowing for the identification of previously undiscovered variants. To illustrate the power of YMCA, we show that the analysis of ancient Y-chromosome lineages can help to resolve Y-chromosomal haplogroups. As a case study, we focus on H2, a haplogroup associated with a critical event in European human history: the Neolithic transition. By disentangling the evolutionary history of this haplogroup, we further elucidate the two separate paths by which early farmers expanded from Anatolia and the Near East to western Europe.**

Uniparentally inherited markers such as mtDNA and the NRY are an attractive source of information about the demographic history of a population due to the fact that their history can be represented by a simple evolutionary tree<sup>1,2</sup>. Since the seminal studies of the 1980s<sup>3,4</sup>, and prior to the genomic era, much of the genetic history of humankind and the peopling of the world was inferred from uniparentally inherited mtDNA and NRY<sup>5–7</sup>.

Due to the high copy number of mtDNA in the cells (Ingman and Gyllenstein 2001), the short genome length (< 17 kb), and the relatively high substitution rate<sup>8</sup>, mtDNA has been particularly well-studied, yielding an inexpensive and yet reliable source of information about the genetic variability of a population<sup>4,9,10</sup>.

Conversely, the mappable portion (the regions for which short reads, such as in ancient DNA studies, have been reliably mapped) of the NRY is much longer (~ 10,445 kb) and presents only as single-copy in the cells of male individuals. The evolutionary substitution rate (in substitutions per site per year) was estimated to be up to two orders of magnitude lower for the NRY<sup>11</sup>, e.g.  $7.77 \times 10^{-10}$  –  $8.93 \times 10^{-10}$  than for the entire mitogenome<sup>8</sup>, e.g.  $1.36 \times 10^{-8}$  –  $1.95 \times 10^{-8}$ , though much debate surrounds estimating substitution rates<sup>12</sup>. However, the greater genome length of the NRY, compared to the mtDNA, means that from these substitution rates, and for a single lineage, we still expect to observe a point mutation approximately every ~ 108 to ~ 123 years for the NRY, compared to between ~ 3094 and 4440 years for the entire mitogenome. Consequently, the NRY can contain more information about the paternal demographic history of a population and can be informative about male-biased population demographic changes, such as through male-driven migration<sup>13–15</sup> or patrilocality<sup>16</sup>, so seeking insights into the paternal history of a population can be of critical importance.

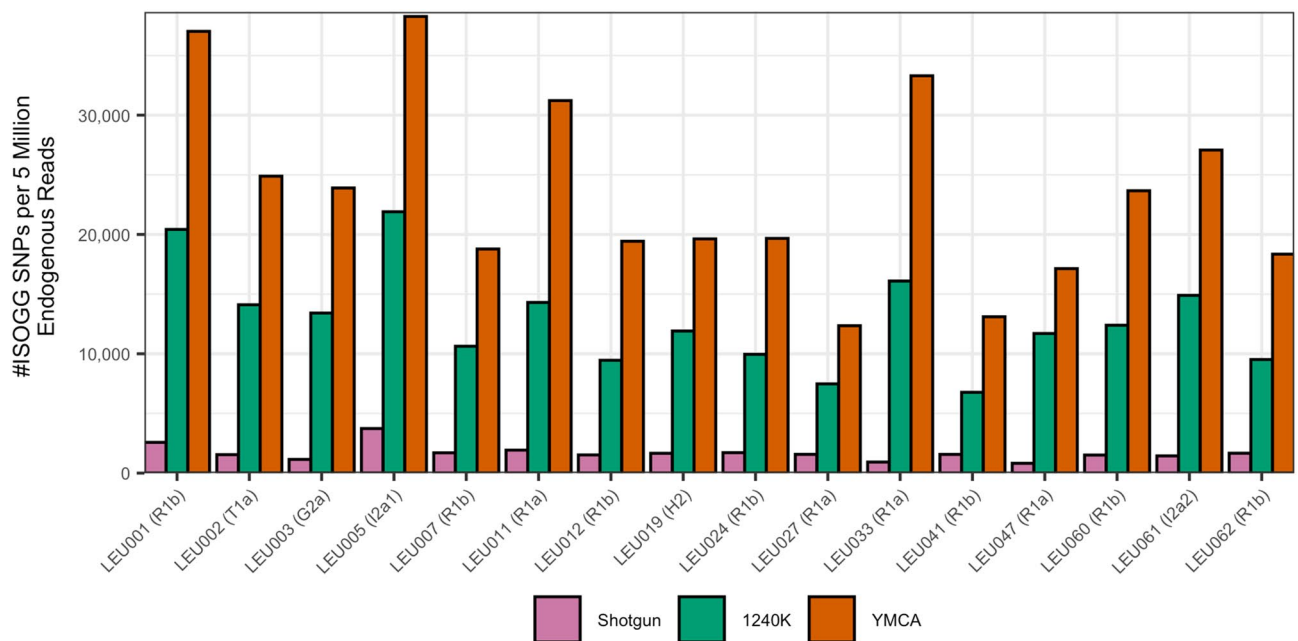
When studying the demographic history of humans, aDNA has been shown to be an irreplaceable source of information. aDNA studies have revealed large-scale population movements and genetic turnover events in Western Eurasia<sup>17–20</sup> that were otherwise impossible to recover from human genetic data of modern-day populations. Studies of the uniparentally inherited markers of ancient individuals have also yielded otherwise undetectable results, e.g. the loss of European mtDNA diversity following the re-peopling of Europe after the last glacial maximum<sup>10</sup>, or the decrease in and partial replacement of diversity of hunter-gatherer Y-chromosome lineages in eastern and central Europe following the Neolithic expansion<sup>21–24</sup>, followed by the loss of diversity of Neolithic Y-chromosomes lineages with the arrival of Steppe-like ancestry at the beginning of the 3<sup>rd</sup> millennium BCE<sup>15,18–20</sup>.

Researchers using aDNA data usually encounter problems related to sample quality, specifically a decrease of endogenous human DNA due to post-mortem DNA decay and environmental contamination<sup>25,26</sup>. The Y chromosome makes up < 2% of the total DNA in male cells, meaning that if researchers wish to use shotgun (SG) sequencing to adequately cover enough informative sites on the single-copy NRY, then, even for samples with good DNA preservation, a substantial sequencing effort is required.

The development of targeted capture assays has allowed aDNA researchers to enrich specific sites and regions of the genome for sequencing, vastly improving the yield of human endogenous DNA from ancient samples<sup>27,28</sup>. One such popular assay is the 1240 k assay, which targets ~ 1.24 M ancestry-informative sites on the human genome, of which ~ 32 k represent a selection of known variants on the Y chromosome (based on an ISOGG list of informative Y-chromosomal SNPs as of 2013/14)<sup>28</sup>. Of note, the commercially available version (myBaits Expert Human Affinities, Daicel Arbor Biosciences) contains an additional 46 k Y-chromosomal SNPs identified by ISOGG to be variable in extant males.

This relatively low number of targeted Y-SNPs, compared to the number of currently known, informative Y-SNPs (as defined by ISOGG,  $n = 73,163$ , or Yfull,  $n = 173,801$ , <https://isogg.org/tree>; <https://www.yfull.com>), allows for basic Y haplogroup (YHG) assignments, but is heavily biased towards modern-day diversity and certain geographic regions. As a consequence, depending on the representation of particular Y-SNPs on the 1240 k assay, the resulting YHG assignments can be of low and uneven resolution, while the targeted approach does not allow for the detection of hidden and/or potentially extinct lineages in the human past.

To better study and understand the male history of human populations, we saw a need for a targeted assay that specifically enriches sequence data for sites on the NRY, without targeting only already well-known SNPs. To achieve this, we designed and implemented YMCA (Y-mappable capture assay), a tiled capture-assay for NRY sequence data that targets regions of the NRY for which short reads, typical in ancient DNA samples, are reliably mapped to the human genome, as defined by Poznik et al.<sup>29</sup>. A similar approach has been explored by two previous studies<sup>30,31</sup>. However, we avoid in-depth comparison with the probe set presented by Petr et al. which targets ~ 6.9 Mb was designed to substantially older samples, such as Denisovan and Neanderthal individuals, and hence the definition of "mappable" was far more conservative and stricter. Conversely, Cruz-Dávalos et al. also present a capture-enrichment approach designed for ancient human samples with low endogenous DNA. The reported ~ 8.9 Mb regions are almost completely included in our target regions (99.97%), and we show that the remaining ~ 1.5 Mb in our target regions still yield reliably mapped sites (see Supplementary Table S1.4).



**Figure 1.** The number of ISOGG SNPs covered per five million quality-filtered mapped reads (y-axis) for the same libraries (x-axis) for shotgun, 1240 k and YMCA sequencing (colours).

Here we show that YMCA significantly improves the relative coverage of NRY sites when compared to shotgun sequencing, allowing for the enrichment of NRY sites for the same sequencing effort. We also show that YMCA significantly outperforms 1240 k SNP assay sequencing in two ways. Empirically, we show that YMCA improves the number of NRY sites that are covered. We also show, by considering the targeted NRY sites as defined by the associated bed files, and if we were to sequence a sample with high complexity to exhaustion, that YMCA has an improved potential resolution for Y-haplogroup assignment and the discovery of new diagnostic SNPs when compared to 1240 k assay sequencing.

We highlight the improved performance obtained via YMCA by analysing the Y-chromosomal haplogroup H2 (H-P96), a low-frequency YHG that is associated with early farmers during the Neolithic transition in Western Eurasia. We curated a data set of 46 previously published individuals (45 ancient and 1 modern), and 49 newly YMCA-sequenced individuals (all ancient). We show that our current understanding of H2, which is based largely on modern H2 samples ( $n=20$ ), is inconsistent with the ancient diversity of our H2 individuals. In resolving this ancient haplogroup, we can show two distinct migration paths along the Mediterranean and Danube for Neolithic groups from Anatolia to Western Europe, ultimately resulting in the Mediterranean-derived groups also reaching Britain and Ireland.

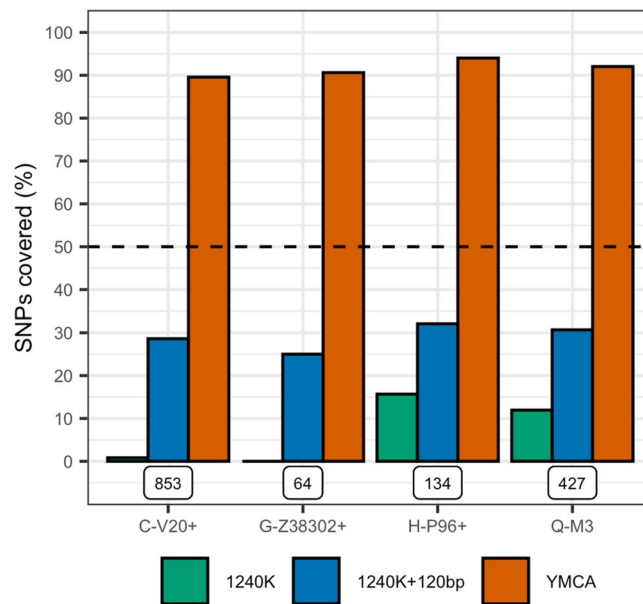
## Results and discussion

**Validating the performance of YMCA.** To evaluate the performance of our new NRY-Capture assay (YMCA), we calculated the empirical fold-increase in endogenous human DNA for a range of samples with varying levels of preservation. We chose samples from the same site (Leubingen, Germany) to avoid the effects of too many environmental variables, for which we had shotgun, 1240 k capture and YMCA sequence data (see Table S3). We then compared the empirical performance of YMCA against standard shotgun sequencing and 1240 k capture on the same libraries by inspecting the number of NRY sites covered, as well as the number of ISOGG SNPs covered at least once for each library type. We account for sample quality and input sequencing effort by filtering for only human endogenous reads, and then normalising the number of sites/SNPs covered per five million endogenous reads.

We observed a significant fold-increase in the amount of endogenous human DNA when comparing shotgun sequencing to YMCA (see Figure S1), which we refer to as “enrichment” from here on. We found that enrichment diminished as the preservation of the sample increased, i.e. for samples with higher starting endogenous DNA % the effect of the enrichment was reduced, but still significant.

We observed a significant mean fold increase of  $\sim 15.2\times$  in the number of NRY sites covered by YMCA captured libraries when compared to shotgun sequencing ( $p=5.5 \times 10^{-7}$ ), and  $\sim 1.84\times$  when compared to 1240 k sequencing ( $p=8.8 \times 10^{-12}$ ), showing that YMCA covers on average more NRY sites than both shotgun and 1240 k sequencing (see Fig. 1). This also indicated that, since we covered on average 15.2 times as many ISOGG SNPs per five million reads for SG sequencing, we would need to sequence  $\sim 76$  million reads to cover the same number of NRY sites for shotgun sequencing compared to only five million reads for YMCA.

Interestingly, we also found mean fold increase of  $\sim 4.36\times$  in the number of ISOGG SNPs covered at least once with YMCA captured libraries when compared to 1240 k sequencing ( $p=9.0 \times 10^{-14}$ ). This indicated that, for the same sequencing effort, YMCA also covers more informative SNPs.



**Figure 2.** The percentage of SNPs (y-axis) covered (up to three branches downstream) for four Y-chromosome haplogroups (x-axis) associated with ancient populations. Colours indicate assay SNPs targeted for 1240 k (green), 1240 k with a 120 bp window (blue) and our YMCA (orange). The dashed black line indicates at least half of the SNPs are represented, and the total number of targetable SNPs is given below each group.

We also found that the fold-increase in the number of NRY sites that we covered, and the endogenous DNA percentage for shotgun and 1240 k sequencing were uncorrelated ( $p = 0.976$  and  $p = 0.617$ ), and that the number of ISOGG SNPs covered and the endogenous DNA percentage for 1240 k sequencing were uncorrelated ( $p = 0.1825$ ) indicating that our results are not dependent on the relative abundance of retrievable human DNA in the sample. Hence, we found that, although the SNPs covered on the Y chromosome are an added bonus when using the 1240 k assay, as it is primarily used for analysing the autosomal genome of male and female individuals, YMCA is clearly a significant improvement if researchers wish to efficiently and thoroughly investigate the non-recombining portion of the Y chromosome.

We then compared the percentage of haplogroup-informative SNPs on the ISOGG SNP list v14.8, that are also included in the 1240 k assay, and YMCA, according to their respective bed files. This comparison will be particularly powerful as YMCA and the 1240 k assay are based on the same technology, and captured via identical lab protocols. The 1240 k assay targets 24.44% of the currently listed ISOGG SNPs, whereas the YMCA targets 90.01% (Fig. 2). Note that the remaining 9.99% of ISOGG SNPs exist in regions of the NRY which are considered “unmappable” for short reads common in ancient DNA. Since each of the sites in the 1240 k assay is targeted by two probes (allele and alternate allele) and two 52 bp probes on either side of the variant, additional sites flanking the “targeted” sites can also be recovered from the mapped reads. Hence, we also allow a window of 120 bp (60 bp on either side) for each SNP on the 1240 k assay, which is a reasonable average read length for aDNA. For this 1240 k + 120 bp list of sites the percentage of targeted ISOGG SNPs increases to 45.34%, but this also illustrates that the 1240 k SNP assay is fundamentally limited by the total number of informative Y chromosome SNPs included. This significant increase in ISOGG targeted SNPs would also explain why, for the same sequencing effort, YMCA covers more ISOGG SNPs.

Additionally, recovering as much of the NRY as possible is of critical importance, especially when researchers are interested in looking for new variants on the Y chromosome, or uncovering past diversity that might no longer exist. When comparing the raw number of sites targeted by the 1240 k assay to YMCA, we observed that the 1240 k capture assay potentially targets a total of 32,670 sites, which is approximately 0.31% of the number of sites targeted by YMCA (~10,445 k). However, if one is to include a window of 120 bp around each SNP again, then the 1240 k assay potentially targets ~3,953 k sites or 37.82% of the number of sites one can potentially analyse using our YMCA. Hence, YMCA is a predictably better tool for exploring the NRY for new ancestry informative SNPs.

We were also interested in comparing the potential resolution to which YHG assignments can be made, given the available ISOGG SNPs targeted by YMCA and 1240 k. We also found that the resolution of a YHG call cannot be improved, even when including a 120 bp tiling window around the ~32 k Y-SNPs of the 1240 k assay, according to the ISOGG SNPs occurring in the respective bed files. This holds true both for dominant YHGs today and in particular for those that are associated with known ancient populations, but that have significantly reduced in frequency in modern populations, and which are not well covered for diagnostic SNPs on the 1240 k assay.

We often observe low resolution in haplogroups such as the early hunter-gatherer haplogroup C-V20<sup>17</sup>, and the Neolithic expansion-associated G-Z38202<sup>32</sup> and H-P96, for which the Y-SNPs of the 1240 k assay target 0.8%, 0% and 13% of the associated ISOGG SNPs, respectively (we include SNPs within three branches downstream of each terminal SNP). If we include a 120 bp window, then these percentages increase to a more respectable 32.5%,



31.2% and 36.2%, which are still much lower than the 89.6%, 90.6% and 95.2% of SNPs targeted by YMCA (see Fig. 2). In addition, poor theoretical coverage for YHGs which are thought to be present in early human population movements, but which remain relatively prevalent in modern populations, can still be an issue for sites on the 1240 k assay. For example, Q-M3, which is associated with the initial peopling of the Americas<sup>33</sup> has only 11.9% (33.5% if a 120 bp window is included) of the relevant diagnostic SNPs covered, compared to 92% for YMCA.

To summarize, YMCA enriches the relative proportion of reads mapping to the NRY, when compared to shotgun sequencing and the 1240 k assay. YMCA also targets more than 2.5 times as many sites on the NRY than the 1240 k assay, allowing for the detection of new diagnostic SNPs. Critically, however, YMCA targets SNPs which are already known to be informative, but the 1240 k assay cannot target.

**Application of YMCA to YHG H2 as a case study.** Through routine application of SG sequencing for sample screening, followed by 1240 k capture for suitable samples in our lab, we were able to explore the performance of the new YMCA on a range of YHG in ancient male individuals. Here, we showcase an example of YHG H-P96, for which the resolution of the evolutionary tree is still unclear due to the scarcity of data and low frequency in modern-day populations. Judging from our current ancient DNA record, it appears that YHG H was more common in the past, in particular among males that were associated with the spread of farming across Western Eurasia during the Neolithisation. As a result, we can show that aDNA research, and in particular high-resolution typing of YHG, can help elucidate the evolutionary relationship of Y chromosome lineages past and present.

The YHG H (H-L901) is thought to have formed in South Asia approximately ~ 48 kya<sup>34</sup>. Three subsequent sub-haplogroups, H1 (H-M69), H2 (H-P96) and H3 (H-Z5857), appear to have quickly formed over the following 4000 years. H1 and H3 have estimated formation times of ~ 44.3 kya, however, H2 is estimated to have formed slightly earlier at ~ 45.6 kya [<https://www.yfull.com>].

Haplogroups H1 and H3 are still found in frequencies as high as 20% in South Asia<sup>35</sup>, but in extremely low frequencies in Europe, with H1 only being found associated with the spread of the Romani people ~ 900 ya. Conversely, H2 has been present in Western Eurasia since at least 10 kya<sup>36</sup>, and is strongly linked with the spread of agriculture<sup>37,38</sup>, but is found at no higher than 0.2% frequency in modern-day western European populations. In contrast, H2 was more common in Neolithic groups, and has been found to have constituted between 1.5 and 9% of the observed Y haplogroups, with the exception of the highly related individuals from Rivollat et al. 2020, for which H2 was ~ 30%<sup>15,19,24,39,28,36,38–41</sup>.

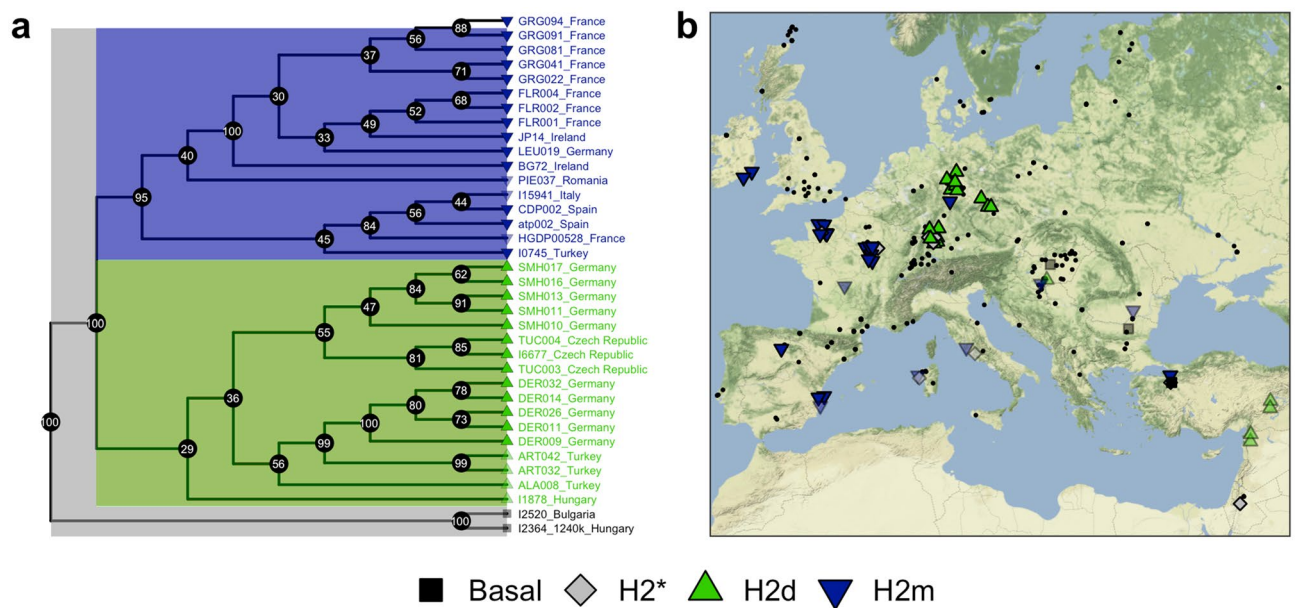
With the arrival of Steppe-related ancestry ~ 5 kya, incoming YHGs such as R1a and R1b would largely replace many of the older, “Neolithic” YHGs, such as G2, T1a, and H2<sup>19</sup>, and although H2 was never found in particularly high frequencies among Neolithic individuals, we expect that its diversity was also greatly reduced, and many sub-lineages were potentially lost altogether.

To test whether our YMCA could improve the haplotyping quality to a point which would allow us to also draw phylogeographic inferences, we made use of newly collated collection of prehistoric ancient human DNA data and selected individuals, who were tentatively assigned to YHG H2. We generated new data for n = 49 individuals, and merged this with n = 46 published Y-chromosomal genomes (see Tables S1 and S2). While H2 is commonly found alongside the more dominant Neolithic YHG G2a (G-Z38302)<sup>24,37</sup>, it is precisely the low frequency of H2 which is of interest here. The relative scarcity of H2 individuals, especially compared to the relatively high frequency of the accompanying G2a individuals, allows us to better track the ‘genealogical history’ and thus potential dispersal routes as we would expect a stronger effect of lineage sorting and therefore a higher chance of observing geographic patterns. In this particular case, we could trace expanding Neolithic farmers from Anatolia to Western Europe through the use of unique markers associated with H2 individuals and test whether we can genetically discern the proposed so-called “Danubian or inland” and “Mediterranean” routes of the Neolithic expansion<sup>42</sup>, which had recently also found support by genomic signals from the nuclear genome<sup>38</sup>.

Unfortunately, we found that the H2 subsection of the evolutionary tree for the Y chromosome is currently poorly understood (due to the scarcity of modern samples of H2 individuals and the relative rarity of ancient H2 individuals), and, in many cases, inconsistent with a tree-like history for almost all of the published and unpublished ancient samples. In all but one case that we found that H2 individuals carry a mixture of derived SNPs from two bifurcated clades in the current ISSOG topology, such as from H2a1 and H2b1. Encouraged by the performance of the YMCA presented above, we thus analysed further H2 individuals in an attempt to resolve the branching pattern of this lineage.

For a non-recombining portion of DNA, the evolutionary history is expected to follow a tree-like structure, and therefore hybrids of sibling haplogroups (such as between ISSOG H2a, H2b1 and H2c1a) are impossible. To try and find a better resolved evolutionary history for these individuals, we constructed a maximum likelihood (ML) phylogenetic tree using IQ-TREE (see “Materials and methods”). We identified two major clades from the ML tree (see Fig. 3A), and denoted these two tentatively named clades H2m (blue clade) and H2d (green clade). With respect to the current ISSOG nomenclature, we note that H2m appears to be defined by a mix of H2, H2a, H2a1 ~ and H2c1a ~ SNPs (see Table S6). H2d appears to be defined by two H2b1 SNPs, and four additional SNPs which were previously undetected (see Table S7). Hence, it could be that H2d is simply derived from the basal H2 group, but with a few private mutations. However, we also note that H2d contained a sub-clade containing individuals from Turkey (ART) and Germany (DER) which were uniquely defined by a further ten SNPs associated with H2b1, potentially indicating further sub-structure (see Figure S4).

Based on our extended set of diagnostic SNPs, we were able to assign n = 58 of our individuals to either one of these two sub-clades, or (basal) H2\* (due to low coverage), even for those who did not meet minimum coverage requirements to be included in the ML tree, which also provides bootstrap support for individual clades (Fig. 3A).



**Figure 3.** (a) Phylogenetic relationships (no branch length units) and (b) H2 sample locations (map created using ggmap, Kahle and Wickham 2013). Shapes and colours indicate the two major clades inferred from the phylogenetic tree. Symbol shading indicates early to late Neolithic (solid) or post-Neolithic (transparent). Black dots indicate all non-H2 Neolithic individuals from Freeman2020 to indicate H2 sampling prevalence<sup>53</sup>. Stars in haplogroup assignments in (a) indicate a lack of resolution to assign samples (not used in the ML tree) to downstream sub-haplogroups.

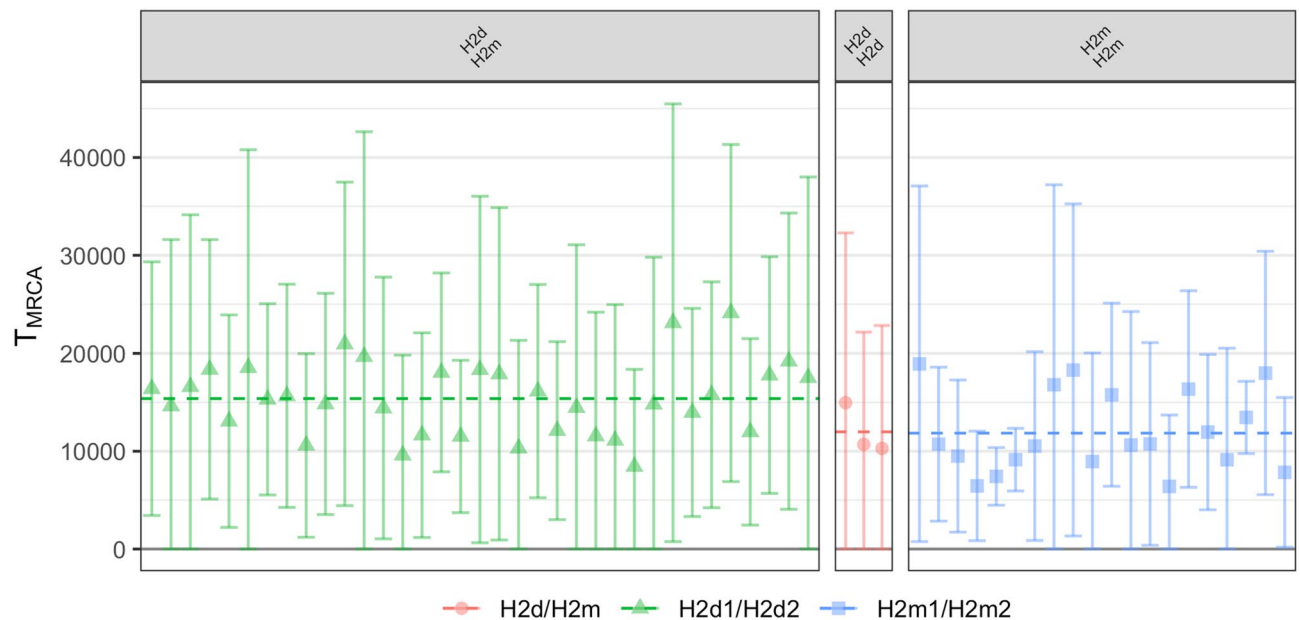
Finally, we also had three individuals who were not derived for any of these additional SNPs, and were ancestral for many of the H2 SNPs (denoted basal,  $n = 3$ ).

When we plotted all of the samples in our study on a map of Europe, a phylogeographic pattern clearly emerged (Fig. 3B). The H2d individuals are all found along the so-called inland/Danubian route into central Europe, and all but one of the H2m individuals are found along the so-called Mediterranean route into Western Europe, the Iberian Peninsula and ultimately, Ireland. The solitary H2m individual (LEU019) found in central Germany is dated to the Late Neolithic/Early Bronze Age context, postdating the Neolithic expansion by 2000–3000 years. Archaeological and mtDNA evidence of an eastward expansion of Middle/Late Neolithic groups such as Michelsberg<sup>43–45</sup> could potentially explain this single geographically outlying observation.

Due to the incomplete and varying coverage of our ancient samples, we were unable to produce a reliable, calibrated tree for divergence time estimates using the radiocarbon dates of ancient samples as tip dates. Instead we estimated the time since the most recent common ancestor (TMRCA) for each pair of individuals to investigate the split times for our newly identified H2 clades (see “Materials and methods”, Figure S2). First, we calibrated our relative substitution rate so that we estimated a mean TMRCA of ~161.3 kya for haplogroup A0 with all other haplogroups (see Figure S3). Using this calibrated substitution rate, we estimated a TMRCA for haplogroup A1 of ~133.2 kya, and a TMRCA of ~48 kya for haplogroup HIJK, which are extremely close to the current estimates of ~133.4 kya and ~48.5 kya respectively (<https://www.yfull.com>). Our estimated TMRCA for H2 was ~24.1 kya, which is slightly older than the current estimate of ~17.1 kya, and which could be explained by our extremely limited access (only one) to high-coverage modern H2 samples, as well as our increased number of ancient samples (<https://www.yfull.com>).

We found that the estimated TMRCA for H2d and H2m was ~15.4 kya. We also found that H2m and H2d had estimated TMRCA of ~11.8 and ~11.9 kya (see Fig. 4). We note, however, that even though the associated error bars are wider due to fewer overlapping SNPs, the mean estimates are still relatively consistent. These estimates, plus the fact that H2d and H2m individuals are found in Anatolia and the Levant, show that H2 diversity most likely existed in Near-Eastern hunter-gatherers before the establishment of agriculture and animal husbandry and likely also in early farmers, and subsequently spread via the Neolithic expansion into Central and Western Europe.

**Identifying diagnostic SNPs for improved YHG H2 resolution.** Having used YMCA to identify two novel subclades of H2, we also aimed to identify which SNPs are diagnostic for these subclades, when compared to the human genome reference hs37d5 (see Tables S5–S7). To do this we looked for segregating sites with the following properties: (1) no individual from the ingroup is ancestral at the site, (2) more than one individual from the ingroup is covered at the site, (3) no individual in the outgroup is derived at the site, and (4) more than one individual in the outgroup is covered at the site. We also restricted the search for “new” SNPs to substitutions that are not C to T or G to A, and thus are less likely to be the result of ancient DNA damage, however we included variants that are C to T or G to A in our results if they are previously-discovered SNPs in the ISOGG or YFull databases. Note, that when we report that x/N samples in a group are derived for some SNP, this means



**Figure 4.** Estimated time since most recent common ancestor ( $T_{MRCA}$ ) (y-axis) for each H2 subgroup with each other (facets), calibrated by the split time of ~163 kya of haplogroup A0 with all other Y haplogroups. All pairwise calculations are filtered to exclude individuals from the same sampling site. The dashed line indicates the mean estimate, and error bars indicate 95% confidence intervals for individual observations.

that the remaining N-x samples are not covered at this position, and we have simply recorded a “missing” base read (Supplementary Tables S5–S7).

We identified 312 potential diagnostic SNPs for the sub-haplogroups/branches in Fig. 3 defined as H2 (all samples), H2d (green) and H2m (blue). Encouragingly, we found that out of the 312 diagnostic SNPs that we identified, 258 (80.1%) are already found to be H2 associated (H2-P96 or more derived) on either the ISOGG list, or on the YFull SNP list. We found only two previously discovered SNPs (0.31%), which were not associated with H2: a C → T SNP associated with R1a1 ~ and R1a1a ~ (found in 17/31 H2 samples). It is unlikely that the C to T substitution is due to damage since 17/31 individuals have this substitution. Furthermore, we also found that for our ancient H2 individuals (except for one modern French H2 individual), we were able to find 110 of the 134 known, basal H2 SNPs.

The remaining 62 newly discovered SNPs for the varying sub-haplogroups listed above represent either undiscovered diagnostic SNPs, or potentially lost H2 diversity (Supplementary Tables S5–S7). However, for several of our newly discovered SNPs, such as an A → G substitution at site 8611196 (found in 20/31 of our H2 individuals), we find overwhelming evidence for new, true diagnostic SNPs (see Table S5).

Our ability to detect these distinct H2 sub-haplogroups, and hence our ability to further elucidate and estimate the divergence times for an informative Y-haplogroup during the Neolithic expansion, is made possible only due to the increased coverage, and the increased number of sites we were able to target with YMCA (when compared to SG or 1240 k sequencing).

## Discussion

The analysis of the Y-chromosomal history of populations can be of significant importance to the understanding of population histories. To this end we advocate for the adoption of targeted sequencing strategies for ancient Y-chromosomal sequence data. Our focused study highlighted the improved coverage and number of SNPs that are attainable when using YMCA, when compared to SG or 1240 k sequencing, for the same amount of sequencing effort, accounting for endogenous human DNA content.

Targeted endogenous human DNA enrichment is of critical importance to overcome poor sample preservation in ancient DNA studies. We have shown that the Y-SNPs of the 1240 k assay ascertained from modern-day males simply do not cover enough of the diagnostic SNPs on the NRY for reliable Y-haplogroup assignment, especially in the case of haplogroups that predate modern diversity, highlighting a need for targeting contiguous regions in favour of an updated “Y-chromosome SNP panel”. YMCA can be applied to the same libraries that are used for other captures and require no additional extractions or library preparations. While it is certainly possible to combine YMCA with other captures assays (which we have not attempted), we argue that a bespoke YMCA of selected male samples in a directed study might outweigh that of a routine combined application (to male and female samples) with additional sequencing effort.

We were also able to show, through a deeper analysis of H2 (H-P96), that the current understanding of ancient H2 diversity is incompatible with a tree-like history (which must be true for NRY history), and that a resolution of this diversity leads to further support for the two paths of the Neolithic expansion from the Near East into Europe; an observation that would not have been possible without the improved resolution offered by YMCA.



We foresee future applications in the study of Y-chromosomal sub-structure in Eurasian hunter-gatherers (within I2a, I2b or C1a diversity) or to better characterise the R1a/R1b diversification of Bronze Age Western Eurasia, Central and South Asia.

## Materials and methods

**Data.** Note that for Y-haplogroup assignment, tree building and SNP identification, we use a mix of shotgun, 1240 k, and YMCA capture sequencing runs. However, to estimate the TMRCA, we use only shotgun and YMCA data as they do not target known segregating sites (which would upwardly bias the substitution rate for samples with 1240 k capture compared to those with shotgun and YMCA data only). Previously published samples were selected from published data with “H2” designated for Y haplogroups<sup>15,20,24,28,36,38–41,46–49</sup>.

**Contamination quality filtering.** To screen our samples for contamination, we consider the heterozygosity for sites on the NRY as our in-house samples are all merged and filtered for sites on the Y-chromosome only. We measured heterozygosity (the proportion of sites with more one than one type of base read per site) for a pileup of the quality filtered reads. We found that 47 of our 49 samples had less than 0.1%, with the remaining 2 samples 1% and 1.85% heterozygous sites. However, we were also confident in the quality of our samples as H2 is a very rare modern haplogroup, with only 19 individuals being downstream of H2-P96 on YFull at the time of this publication. Hence, if any of our samples had been contaminated by a *male* source, it would be readily noticeable in bam pileups as derived alleles for another haplogroup, which means these samples would not have been identified as H2, and hence would not be in the study.

**Method of Y haplogroup assignment.** To assign Y haplogroups to samples we follow a partially-automated process. We begin by taking trimmed, merged, deduplicated, quality-filtered bam files, and, for each bam file, creating a pileup of every site that was covered using the *pileup* function in the Rsamtools library for the R statistical software package. We then filter this pileup of SNPs found on the ISOGG list (<https://isogg.org/tree>), and then for each SNP we calculate and record the number of derived and ancestral SNP calls, the form of the ancestral and derived SNPs, and the difference (defined as the number of derived minus the number of ancestral SNP calls). Note that a positive difference indicates evidence for the ancestral form of the associated ISOGG SNP, and a negative difference indicates the converse. Recording the form of the called SNPs (i.e. C to T or G to A transitions), allows us to identify where DNA damage could have caused us to infer false calls.

We return two CSVs: one CSV of only ISOGG SNPs with positive differences (for ease of reading the easiest path from root to terminal SNP), and a second CSV of all SNPs (negative or positive) allowing us to double check potentially spurious SNPs (say to check to see if more basal branches from our terminal branch are not just missing, i.e. not covered, but also not associated with negative differences). This second CSV also allows us to discover when some SNPs are derived and some are ancestral for the same branch, indicating a transitional form of the basal haplogroup.

Finally, in cases where we are uncertain of the dependability of a call (say a C to T transition with only one read), we also manually inspect where the site falls on the associated read(s), placing increased trust in SNP calls originating further from the terminal ends of a read.

**Comparing the performance of our Y-capture assay.** When comparing the empirical performance of our Y-capture assay to both shotgun and 1240 k sequencing, we took libraries for which shotgun, 1240 k and Y-capture sequencing had all been performed. All samples were prepared and analysed using the same methods and parameters values as for the main data set.

To compare the theoretical performance of YMCA against the 1240 k assay, we downloaded the ISOGG SNP list v15.64. We then took the bed files for the NRY and 1240 k assay, and found which sites overlapped with the SNP son the ISOGG SNP list.

When comparing empirical data performance for shotgun, 1240 k and YMCA data, we included only samples that had shotgun endogenous DNA greater than 0.1 %, had sequencing results for shotgun, 1240 k and YMCA sequencing, and then normalized the number of SNPs covered by the number of reads mapping to the human reference (hs37d5) after quality filtering. We did this to avoid any potential bias from sample quality or sequencing effort.

## Phylogenetic tree reconstruction

We began by taking pileups of each bam file, and performing the following quality filters for calling a consensus alignment; for each sequence we considered only sites for which we had at least two reads, with a minimum allele frequency less than 10%, and called the majority allele. We then took the aligned consensus sequences, and kept only samples for which at least 1100 segregating sites were covered, and then filtered sites for which more than at least one sample was covered. We selected a lower bound of 1100 segregating sites by varying this value, and inspecting bootstrap node support values. A minimum bootstrap support for major cladal splits of 80% was required.

We also included high-coverage samples from the 1000 Genomes Project<sup>50</sup> from Y-haplogroups A, H1, H2 and H3, as well as one ancient H1 sample<sup>46</sup> as outgroups.

We performed DNA substitution model selection using ModelFinder<sup>51</sup> and selected the transversion model (TVM) as it had the minimum Bayesian information criterion value. We found a maximum likelihood tree using IQ-TREE v1.5.5<sup>52</sup>.

## Data availability

Prehistoric human skeletal material for this study was collected with permission of the respective archaeological or state heritage organisations. All sample providers and collaboration partners are also co-authors. All unpublished data used in this study was produced by our team in-house and thus does not require additional permission from third parties. Data generated for this study can be found at the European Nucleotide Archive under the study accession number PRJEB45741.

## Code availability

All R scripts can be obtained by contacting the corresponding authors.

Received: 26 February 2021; Accepted: 9 July 2021

Published online: 22 July 2021

## References

1. Brown, W. M. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. USA* **77**(61), 3605–3609. <https://doi.org/10.1073/pnas.77.6.3605> (1980).
2. Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: An evolutionary marker comes of age. *Nat. Rev. Genet.* **4**(8), 598–612. <https://doi.org/10.1038/nrg1124> (2003).
3. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**(6099), 31–36. <https://doi.org/10.1038/325031a0> (1987).
4. Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H. J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* **22**(6), 339–345. <https://doi.org/10.1016/j.tig.2006.04.001> (2006).
5. Pakendorf, B. & Stoneking, M. Mitochondrial DNA and human evolution. *Annu. Rev. Genom. Hum. Genet.* **6**(1), 165–183. <https://doi.org/10.1146/annurev.genom.6.080604.162249> (2005).
6. Underhill, P. A. & Kivisild, T. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41**, 539–564. <https://doi.org/10.1146/annurev.genet.41.110306.130407> (2007).
7. Kivisild, T. The study of Y chromosome variation through ancient DNA. *Hum. Genet.* **136**(5), 529–546. <https://doi.org/10.1007/s00439-017-1773-sz> (2017).
8. Soares, P. *et al.* Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84**(6), 740–759. <https://doi.org/10.1016/j.ajhg.2009.05.001> (2009).
9. Finnäll, S., Lehtonen, M. S. & Majamaa, K. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.* **68**(6), 1475–1484. <https://doi.org/10.1086/320591> (2001).
10. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr. Biol.* **26**(6), 827–833. <https://doi.org/10.1016/j.cub.2016.01.037> (2016).
11. Helgason, A. *et al.* The Y-chromosome point mutation rate in humans. *Nat. Genet.* **47**(5), 453–457. <https://doi.org/10.1038/ng.3171> (2015).
12. Scally, A. & Durbin, R. Revising the human mutation rate: Implications for understanding human evolution. *Nat. Rev. Genet.* **13**(10), 745–753. <https://doi.org/10.1038/nrg3295> (2012).
13. Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* **25**(4), 459–466. <https://doi.org/10.1101/gr.186684.114> (2015).
14. Zeng, T. C., Aw, A. J. & Feldman, M. W. Cultural hitchhiking and competition between patrilineal kin groups explain the post-Neolithic Y-chromosome bottleneck. *Nat. Commun.* **9**(1), 1–2. <https://doi.org/10.1038/s41467-018-04375-6> (2018).
15. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**(6432), 1230–1234. <https://doi.org/10.1126/science.aav4040> (2019).
16. Deguilloux, M.-F. & Mendisco, F. Ancient DNA: A window to the past of Europe. *Hum. Hered.* **76**(3–4), 121–132. <https://doi.org/10.1159/000356933> (2013).
17. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**(7606), 200–205. <https://doi.org/10.1038/nature17993> (2016).
18. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**(7555), 167–172. <https://doi.org/10.1038/nature14507> (2015).
19. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**(7555), 207–211. <https://doi.org/10.1038/nature14317> (2015).
20. Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**(7695), 190–196. <https://doi.org/10.1038/nature26164> (2018).
21. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**(5949), 137–140. <https://doi.org/10.1126/science.1176869> (2009).
22. Haak, W. *et al.* Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biol.* **8**, 11. <https://doi.org/10.1371/journal.pbio.1000536> (2010).
23. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of Central European mitochondrial genetic diversity. *Science* **342**(6155), 257–261. <https://doi.org/10.1126/science.1241844> (2013).
24. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**(7680), 368–372. <https://doi.org/10.1038/nature24476> (2017).
25. Higgins, D., Rohrlach, A. B., Kaidonis, J., Townsend, G. & Austin, J. J. Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies. *PLoS One* **10**, 5. <https://doi.org/10.1371/journal.pone.0126935> (2015).
26. Llamas, B. *et al.* From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *Sci. Technol. Archaeol. Res.* **3**(1), 1–14. <https://doi.org/10.1080/20548923.2016.1258824> (2017).
27. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**(7523), 445–449. <https://doi.org/10.1073/pnas.1221359110> (2014).
28. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**(7583), 499–503. <https://doi.org/10.1038/nature25778> (2015).
29. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**(6145), 562–565. <https://doi.org/10.1126/science.1237619> (2013).
30. Cruz-Dávalos, D. I. *et al.* In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genom.* **19**(1), 1–16. <https://doi.org/10.1186/s12864-018-4945-x> (2018).
31. Petr, M., Hajdinjak, M., Fu, Q., Essel, E. & Rougier, H. The evolutionary history of Neanderthal and Denisovan Y chromosomes. *Science* **369**(6511), 1653–1656. <https://doi.org/10.1126/science.abb6460> (2020).
32. Lacan, M. *et al.* Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proc. Natl. Acad. Sci.* **108**(24), 9788–9791. <https://doi.org/10.1073/pnas.1100723108> (2011).

33. Ruiz-linares, A. *et al.* Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. *Proc. Natl. Acad. Sci.* **96**, 6312–6317. <https://doi.org/10.1073/pnas.96.11.6312> (1999).
34. Sengupta, S. *et al.* 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**(2), 202–221. <https://doi.org/10.1086/499411> (2006).
35. Rai, N. *et al.* The phylogeography of Y-chromosome haplogroup H1a1a-M82 reveals the likely Indian origin of the European Romani populations. *PLoS One* **7**, 11. <https://doi.org/10.1371/journal.pone.0048477> (2012).
36. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**(7617), 419–424. <https://doi.org/10.1038/nature19310> (2016).
37. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. USA* **113**(25), 6886–6891. <https://doi.org/10.1073/pnas.1523951113> (2016).
38. Rivollat, M. *et al.* Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers. *Sci. Adv.* **6**, 22. <https://doi.org/10.1126/sciadv.aaz5344> (2020).
39. Mathieson, I. *et al.* The genomic history of southeastern Europe. *Nature* **555**(7695), 197–203. <https://doi.org/10.1038/nature16152> (2018).
40. Brunel, S. *et al.* Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proc. Natl. Acad. Sci. USA* **117**(23), 12791–12798. <https://doi.org/10.1073/pnas.1918034117> (2020).
41. Skourtanioti, E. *et al.* Genomic history of neolithic to bronze age Anatolia, Northern Levant, and Southern Caucasus. *Cell* **181**(5), 1158–1175. <https://doi.org/10.1016/j.cell.2020.04.044> (2020).
42. Price, T. D. The introduction of farming in northern Europe. In *Europe's First Farmers* 260–300 (Cambridge University Press, 2009). <https://doi.org/10.1017/cbo9780511607851.011>.
43. Jeunesse, C. Pour une origine occidentale de la culture de Michelsberg?. *Mater. Archäol. Baden-württemb.* **43**, 29–45 (1998).
44. Küßner, M. Thüringen in der ersten Hälfte des 4. vorchristlichen Jahrtausends. *Alteuropäische Forschungen* **9**, 75–82 (2016).
45. Beau, A. *et al.* Multi-scale ancient DNA analyses confirm the western origin of Michelsberg farmers and document probable practices of human sacrifice. *PLoS One* **12**, 7. <https://doi.org/10.1371/journal.pone.0179742> (2017).
46. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, 6457. <https://doi.org/10.1126/science.aat7487> (2019).
47. Antonio, M., Gao, Z., Moots, H. & Lucci, M. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science* **366**(6466), 708–714. <https://doi.org/10.1126/science.aay6826> (2019).
48. Cassidy, L. M. *et al.* A dynastic elite in monumental Neolithic society. *Nature* **582**(7812), 384–388. <https://doi.org/10.1038/s41586-020-2378-6> (2020).
49. Fernandes, D. M. *et al.* The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. *Nat. Ecol. Evol.* **4**(3), 334–345. <https://doi.org/10.1038/s41559-020-1102-0> (2020).
50. Genomes Project. A global reference for human genetic variation. *Nature* **526**(7571), 68–74. <https://doi.org/10.1038/nature15393> (2015).
51. Kalyanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**(6), 587–589. <https://doi.org/10.1038/nmeth.4285> (2017).
52. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**(1), 268–274. <https://doi.org/10.1093/molbev/msu300> (2015).
53. Freeman, L., Brimacombe, C. S., & Elhaik E. aYChr-DB: a database of ancient human Y haplogroups. *NAR Genom. Bioinform.* **2**(4), <https://doi.org/10.1093/nargab/lqaa081> (2020).

## Acknowledgements

We thank Nigel Bean, Jonathan Tuke, Vincent Braunack-Mayer and Gary Glonek for important discussions regarding the manuscript. We thank Chris Tyler-Smith and Yali Xue for critical advice and support during the design of the study and capture assay.

## Author contributions

A.B.R., W.H., J.K. and A.H. initiated and led the study. L.P., S.P., A.C., M.R., V.V.M., G.U.N., E.S., M.L., M.A., K.B., Y.B., M.F.D., M.D., Y.S.E., M.E., M.F., M.F., S.F., E.G., A.H., S.H., M.K., M.M., R.Ö., S.R., S.R., D.C.S.-G., J.S.D., P.W.S., C.R.-d.-T.M., K.A.Y., C.P. participated in the laboratory work, sample management, gathering of contextual information and organized the sample collection for genetic analyses. A.B.R. performed the statistical analyses. All authors contributed to the interpretation of the data. A.B.R. wrote the first version of the manuscript which was edited by W.H., L.P., S.P., A.H. and C.P. and all authors contributed to its improvement.

## Funding

Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Max Planck Society, the French (ANR) and German (DFG) Research Foundations under the INTERACT project (ANR-17-FRAL-0010, DFG-HA-5407/4-1, 2018-2021) to M.R. and W.H., the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under Grant agreement no. 771234-PALEORIDER to W.H., the award Praemium Academiae of the Czech Academy of Sciences to M.E. and the project RVO 67985912 of the Institute of Archaeology of the Czech Academy of Sciences, Prague to M.S.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-94491-z>.

**Correspondence** and requests for materials should be addressed to A.B.R. or W.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021