Supplementary information for:

Tracing the genetic origin of Europe's first farmers reveals insights into their social organization

Anna Szécsényi-Nagy[1,2,*], Guido Brandt[1], Wolfgang Haak[3], Victoria Keerl[1], János Jakucs[4], Sabine Möller-Rieker[1], Kitti Köhler[4], Balázs Gusztáv Mende[2], Krisztián Oross[4], Tibor Marton[4], Anett Osztás[4], Viktória Kiss[4], Marc Fecher[1], György Pálfi[5], Erika Molnár[5], Katalin Sebők[6], András Czene[7], Tibor Paluch[8], Mario Šlaus[9], Mario Novak[10], Nives Pećina-Šlaus[11], Brigitta Ősz[12,‡], Vanda Voicsek[12,‡], Krisztina Somogyi[6,‡], Gábor Tóth[13], Bernd Kromer[14], Eszter Bánffy[4,15], Kurt W. Alt[1,16,17]

[1] Institute of Anthropology, Johannes Gutenberg University of Mainz, D-55128 Mainz, Germany

[2] Laboratory of Archaeogenetics, Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, H-1014 Budapest, Hungary

[3] Australian Centre for Ancient DNA, School of Earth and Environmental Sciences, AUS-5005 University of Adelaide, Australia

[4] Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, H-1014 Budapest, Hungary

[5] Department of Biological Anthropology, University of Szeged, H-6726 Szeged, Hungary

[6] Institute of Archaeological Sciences, Eötvös Loránd University, H-1088 Budapest, Hungary

[7] Salisbury Archaeological Ltd., 2040 Budaörs, Hungary

[8] Móra Ferenc Museum, H-6720 Szeged, Hungary

[9] Anthropological Center, Croatian Academy of Sciences and Arts, 10000 Zagreb, Croatia

[10] School of Archaeology, University College Dublin, 4 Dublin, Ireland

[11] Department of Biology, School of Medicine, University of Zagreb, 10000 Zagreb, Croatia,

[12] National Heritage Protection Centre of the Hungarian National Museum, Department Pécs, H-7621 Pécs, Hungary

[13] Biology Department, University of West Hungary, H-9700 Szombathely, Hungary

[14] Curt-Engelhorn-Centre for Archaeometry, D-68159 Mannheim, Germany

[15] German Archaeological Institute, Roman-Germanic Commission, D-60325 Frankfurt am Main, Germany

[16] Institute for Prehistory and Archaeological Science, University of Basel, 4003 Basel, Switzerland

[17] Danube Private University, A-3500 Krems, Austria.

[#] e-mail: szecsenyi-nagy.anna@btk.mta.hu

[‡] currently out of profession

Index

Material and Methods 3

Archaeological sites and dating 3

Samples and sampling 3

Ancient DNA work and authenticity of the results 4

Analyses of mitochondrial DNA 4

Analyses of Y chromosomal DNA 5

Comparative data for population genetic analyses 6

Haplotype diversity 8

Fisher's exact test 8

Genetic distances 8

Test of population continuity (TPC) 8

Principal component analysis (PCA) 9

Multidimensional scaling (MDS) 10

Analysis of molecular variance (AMOVA) 10

Ancestral shared haplotype analysis (ASHA) 10

Genetic distance maps 11

References 12

Supplementary Table 16

Supplementary Figures 18

Supplementary Datasets 28

Material and Methods

Archaeological sites and dating

For the genetic investigations, we selected one Mesolithic site from south Croatia, six Starčevo (STA) sites from western Hungary, three STA sites from northern Croatia, and eight Hungarian sites of the Linearbandkeramik in Transdanubia (LBKT) (Figure S8, Dataset S1). Our aim was to cover the major distribution area of both cultures and sample the most important Early and Middle Neolithic sites in western Hungary [1,2]. The LBKT site of Harta-Gátőrház geographically does not belong to today's Transdanubia, but rather to the Danube-Tisza Interfluve, although culturally this site was connected to the Transdanubian distribution of the LBK.

Since graveyards are absent in the STA and LBKT, all investigated samples are from settlement burials. Most of the studied sites were inhabited throughout several archaeological periods (Dataset S1). Each analysed individual was dated and assigned to a particular archaeological culture by characteristic grave goods, archaeological context and stratigraphic position (Dataset S1-S2). In 26 instances, where characteristic grave goods were absent or the archaeological context was insufficient for unambiguous attribution of the individuals to a certain culture, we dated the human skeletal remains by radiocarbon analyses (Dataset S2).

Samples and sampling

The skeletons from western Hungary were uncovered during the last decade in the course of rescue excavations preceding motorway constructions or other infrastructure projects, while the sites from northern Croatia were excavated between 1977 and 1999. Samples from three STA and two LBKT sites (Dataset S2) were obtained in their untreated and unwashed state after the excavation. To monitor potential sources of contamination, we took swab samples from all genetic and anthropological investigators as well as most of the archaeological investigators who came into contact with the skeletons or the samples (Dataset S19). The sampling was carried out by A.S-N., J.J., M.F., and V.Ke. with the anthropological assistance of BG.M. and K.K.. Human remains of multiple burials were individualized by anthropological experts listed in Dataset S2. To avoid contamination of the skeletal remains with modern DNA, the samples were taken with all possible precautions through the use of gloves, face masks and disposable oversleeves. All instruments and materials used were cleaned extensively with bleach before, between, and after sampling. We took two to five samples per individual, from different skeletal elements. Whenever

possible, teeth were favoured for ancient DNA analyses, otherwise pieces of long bone compacta or the petrous bones (*pars petrosa ossis temporalis)* were taken, which were sawn out using a cleaned diamond drill. The samples were then directly transferred to the Institute of Anthropology at the University of Mainz and stored at -20 °C.

Ancient DNA work and authenticity of the results

The ancient DNA work was carried out in specialized facilities of the Bioarchaeometry Group, Institute of Anthropology, at the Johannes Gutenberg University of Mainz following well-established protocols to prevent and minimize contamination with modern DNA [3–6]. These facilities are composed of pre-PCR labs for sample preparation, DNA extraction, and PCR set-up and post-PCR labs for amplification, sequencing, and cloning. The precautions against contamination, the sample preparation and DNA-extraction followed our standard protocols as described previously [6] with the following modifications: during the grinding step, every tenth sample was a grinding blank, consisting of DNA free hydroxyl apatite powder (Roth, Karlsruhe, Germany). For the extraction, 0.2-1 g bone or tooth powder was used. 8-22 samples were processed at once, with one or two extraction blanks and a grinding blank at each extraction event. In previous publications, we have discussed several criteria compiling a chain of evidence for the authentication of our ancient data [3–6], which we have applied analogously in this study.

Analyses of mitochondrial DNA

Mitochondrial DNA (mtDNA) diversity was investigated by the analyses of multiple independent and informative loci of the mitochondrial genome, including the HVS-I and II of the control region and 22 haplogroup-defining SNPs of the coding region (Dataset S3-S4). Depending on the state of DNA preservation of each sample, HVS-I was amplified using one of three different primer systems consisting of two, four or six overlapping primer pairs with decreasing amplicon length (Dataset S18). These primer systems produced contiguous HVS-I sequences of 356 (np 16046-16401), 413 (np 15997-16409), and 383 base pairs (np 16019-16401), respectively. HVS-I sequences were replicated by at least three independent amplifications from a minimum of two samples per DNA extracts, producing 6-18 independent and overlapping amplicons (depending on the primer system used). In addition, selected PCR products with ambiguous nucleotide positions were cloned, and an average of 5 clones per amplicon were sequenced to monitor possible background contaminations and DNA damage. Poorly preserved samples with numerous

ambiguous nucleotide positions were cloned entirely. Individuals with inconsistent HVS-I results were either discarded or replicated using an independent third sample. HVS-II sequences were obtained from individuals from the same archaeological site that showed consistent and identical HVS-I motifs in order to detect potential maternal kinship. HVS-II sequences were amplified at least twice from two extracts by using four overlapping primer pairs (Dataset S18) that produced a contiguous sequence of 364 bp (np 34-397). Amplicons with numerous ambiguous nucleotide positions were additionally cloned. HVS-I and HVS-II were amplified, purified, sequenced and cloned according to our standard protocols as described previously [3,6]. Coding region information was achieved using the GenoCoRe22 SNP multiplex assay [5], which was amplified once per DNA extract.

Mitochondrial sequence polymorphisms were reported relative to the revised Cambridge Reference Sequence (rCRS) [7] as well as the Reconstructed Sapiens Reference Sequence (RSRS, www.mtdnacommunity.org) [8]. Haplogroup determination was carried out according to the mtDNA phylogeny of PhyloTree build 14, accessed 05 April 2012 (www.phylotree.com) [9]. The HVS-I sequence data were deposited in GenBank (www.ncbi.nlm.nih.gov/genbank) under the accession numbers: KP828071 - KP828154.

Analyses of Y chromosomal DNA

Y chromosome diversity was obtained through the analyses of 33 haplogroup-defining SNPs of the non-recombining part of the Y chromosome (NRY), which enabled us to distinguish the most frequent Eurasian haplogroups (Dataset S5). Most of these informative SNPs (25) were retrieved using the GenoY25 SNP multiplex assay [5]. In addition, we designed singleplex PCR for 8 additional SNP markers to increase the sub-haplogroup resolution of particular branches (I, G and F*) of the Y chromosome phylogeny (Dataset S18), which were detected in our ancient samples. Multiplex assays were replicated at least 4 times, twice from each extract. The singleplex PCRs were performed once from each DNA extract. Due to the incomplete morphological sex data, all individuals with reproduced mtDNA results were tested for NRY markers. The GenoY25 SNP multiplex was carried out according to the protocol published previously [5] with the following modifications: PCR was set up in a volume of 16 µl consisting of 1× Buffer Gold, 8 mM $MgCl_2$ (Applied Biosystems), 0.7 mM dNTPs (Qiagen), ≤0.2 µM of each primer, 13.4 µg BSA (Roche), 1.25 U of Amplitaq Gold Polymerase (Applied Biosystems), and 4 µl target DNA. Thermocycling conditions consisted of an initial denaturation at 95°C for 6 min, 37–45 cycles of 10 sec at 95°C, 1

min at 60°C, and 30 sec at 65°C, followed by a final extension at 65°C for 6 min. PCR products were purified by incubating 2.5 µl PCR product with 1.5 U FastAP, 0.6 U ExoI, 1 µl 10x FastAP buffer (Fermentas, Thermo Fischer Scientific) and 1 µl HPLC water at 37°C for 10 min, followed by heat inactivation at 75°C for 5 min. SBE reaction was performed with 30 cycles, and the amplicons were purified with 1U FastAP (Fermentas) using cycling conditions consistent to the purification of the PCR products. The singleplex PCR were carried out according to the protocol used for mtDNA. Amplicons were purified with ExoI and FastAP and subsequently sequenced in line with the protocols used for mtDNA.

Y chromosomal haplogroups were determined using the phylogeny of the International Society of Genetic Genealogy (2014), Y-DNA Haplogroup Tree 2014, Version: 9.52, Date: 5 April 2014 (http://www.isogg.org/tree/).

Comparative data for population genetic analyses

For comparative analyses of the two investigated populations, we used prehistoric and present-day mtDNA and NRY data from published sources.

The mtDNA data of the STA and LBKT samples were compared with 533 published prehistoric data across Europe, which were pooled into 14 groups according to cultural, chronological, and geographic aspects (Dataset S6). These groups comprise a hunter-gatherer metapopulations from Central/North Europe [10–13], which is complemented by an Early Neolithic period sample from eastern Hungary whose genome shows clear western European hunter-gatherer ancestry [13,14]. Furthermore, hunter-gatherers from southwestern Europe [15–17], the LBK population from Central Europe [3,5,6,18], a temporal succession of four cultures' people from the 5^{th}/4^{th} millennia BC of central Germany [6], four populations from the 3^{rd}/2^{nd} millennia BC of central Germany [4,6,19], and three southwest European populations from the 6^{th}-4^{th} millennia BC of Portugal [15], Basque County and Navarra [16], and Catalonia [20,21], representing the Neolithic of the Iberian Peninsula were incorporated to the analysis. Recent mtDNA data of early farmers from eastern Hungary [3,13,22] were used in different combinations (Dataset S7-10). MtDNA data of the Guba et al. study [23] from Hungary were omitted from the population genetic analysis since serious doubts have been raised concerning the accurate dating and cultural assignment of some of these samples [24].

In order to identify affinities of our prehistoric sample sets in the maternal and paternal gene pool of present-day Eurasian and African populations, we gathered 67,996 mitochondrial HVS-I sequences and 49,516 NRY SNP profiles from the literature. We generated different mtDNA and

NRY datasets, which were used for PCA and genetic distance maps and we pooled the modern-day data into different populations according to geography or ethnicity, as described in the original publications.

For PCA with mtDNA data, the present-day samples were grouped into 73 populations. This dataset was composed of 50,688 sequences with an average sample size of 694 samples per population (Dataset S14). Mitochondrial genetic distance maps were generated from HVS-I sequences of 130 modern-day populations. Whenever possible, the administrative subdivisions of a country were considered in order to increase the phylogeographic resolution. In this dataset, we only included population data with a minimum sequence range of np 16068-16365 to exclude biases by varying sequence ranges. Each population is represented by a maximum of 140 randomly selected individuals, which resulted in a total amount of 17,074 sequences used in the analysis (Dataset S15).

For population genetic analysis of NRY data, we combined our results with eight published LBK and two eastern Hungarian LBK data [5,13,14] to enlarge the prehistoric dataset up to 19 individuals. Present-day population data were only considered when the Y chromosome sub-haplogroups I, G, F, K, and R1a were differentiated, the selection of which includes the most frequent haplogroups observed from the prehistoric data. PCA was carried out with 24,464 samples from 80 present-day populations with an average sample size of 305 individuals per population (Dataset S16). The Y chromosome genetic distance map consists of 100 modern-day populations with 215 samples per population on average, using of 21,478 individuals from our database (Dataset S17).

Recent ancient DNA study from 8,000 BC Near Eastern farmers raises the question whether modern Near Eastern mtDNA can be used as a proxy for the Near Eastern Neolithic variability [25]. In our opinion, these newly described seven different incomplete HVS-I haplotypes (np 16095-16369) only provide a limited basis for comparative ancient DNA analyses. Future next generation genomic data would probably provide a basis of prehistoric comparison. We thus still consider modern-day Near Eastern genetic data as the best available proxies, when tracing the origin of the first European farmers.

Haplotype diversity

Haplotype diversity [26] of the two western Carpathian Basin populations and the Central European LBK was computed in DnaSP Version 5.10.01 [27], using HVS-I sequences (np 16056-16400).

Fisher's exact test

We used the mtDNA haplogroup frequencies in order to identify significant variation between the haplogroup composition of the HGCN, STA, LBKT, and LBK (Table S1) using Fisher's exact test [28]. Overall, 16 mtDNA (H, HV, V, J, K, N1a, T1, T2, U, U2, U3, U4, U5a, U5b, W, and X) haplogroups were distinguished. The Fisher test was carried out in R 3.0.2 (The R Foundation for Statistical Computing 2011, http://www.r-project.org) by using the implemented *fisher.test* function. Significant variation in haplogroup compositions was assessed by 10,000 permutations.

Genetic distances

F_{st} values were computed in Arlequin 3.5.1 [29] based on HVS-I sequences (np 16056-16400) of the Central/North European hunter-gatherers, the two Carpathian Basin populations, and the Central European LBK (Table S1). We used the Tamura & Nei substitution model [30] and an associated gamma value of 0.177, which were inferred from the software FindModel based on PAML likelihoods (www.hcv.lanl.gov/content/sequence/findmodel/findmodel.html) and tested significant variations in F_{st}-values by 10,000 permutations. The p values were adjusted post hoc to correct for multiple comparisons with the Benjamin and Hochberg method, using the function *p.adjust* in R 3.0.2 [31].

Test of population continuity (TPC)

We performed tests of population continuity as described by Brandt and his colleagues [6] using the absolute haplogroup frequencies of the hunter-gatherer Central/North, STA, LBKT and LBK datasets (Dataset S7). In order to apply conservative parameters, i.e. maximizing the chances of genetic drift, we used the terminal dates of the Mesolithic in the Carpathian Basin (6,000 cal BC) and of each Neolithic culture's timespan to define the difference in time between populations in n generations of 25 years. We also ran each of the pairwise tests of all possible group combinations

with three different effective population sizes (Ne=500, 5,000 and 30,000) (Dataset S13). The TPC script is available at https://github.com/joepickrell/tpc.

Principal component analysis (PCA)

PCA was carried out based on mtDNA and NRY haplogroup frequencies of prehistoric and modern-day populations. On the prehistoric level, the mtDNA haplogroup composition of the STA and LBKT samples were compared at first to 15 ancient populations. In this analysis, we considered 22 mtDNA haplogroups (H, H5, HV, HV0, V, I, J, K, N, N1a, R, T1, T2, U, U2, U3, U4, U5a, U5b, U8, W, and X), which were observed in the ancient samples (Dataset S7, Figure 2). The same haplogroups were differentiated in a second PCA, where the Transdanubian data were pooled with the contemporaneous eastern Hungarian datasets [3,13,22] (Dataset S8, Figure S1). In order to exclude biases induced by potential maternal kinship within the prehistoric datasets, which could have led to an overestimation of haplogroup frequencies and genetic affinities, we included a reduced dataset (marked with the symbol *) in the analysis. Redundant haplotypes with identical HVS-I and II sequences from the same site, were counted once.

Mitochondrial haplogroup frequencies of the two Neolithic Carpathian Basin cultures and 73 populations were used for PCA with present-day comparative data. The following 21 haplogroups were differentiated that cover the most frequent haplogroups of modern-day Eurasian populations: H, HV, HV0/V, I, J, K, N1a, T1, T2, U, U2, U3, U4, U5a, U5b, U8, W, and X, African haplogroups (L), Asian haplogroups (A, B, C, D, E, F, G, Q, Y, and Z) and other (all remaining haplogroups) (Dataset S14, Figure S4).

The Y chromosomal data of the STA, LBKT, and LBK were pooled and compared to 80 modern-day populations. Y chromosome haplogroup frequencies were condensed to 13 groups (AB, E, DHC, F, G, I, J, KTS, L, NO, N, PR, and R1a) (Dataset S16, Figure S6), based on the Y chromosome phylogeny and phylogeography. This haplogroup classification was conditioned by the varying resolution of the published comparative data, which were in many cases insufficiently resolved to distinguish further subgroups such as I1 and I2 or G1 and G2.

All PCA were carried out using the *prcomp* function for categorical PCA, implemented in the R 3.0.2 [31] and plotted in a two-dimensional (prehistoric PCAs) or three-dimensional space (present-day population PCAs), displaying the first two or three principal components, respectively.

Multidimensional scaling (MDS)

HVS-I sequences (np 16056-16400) of the two western Carpathian Basin populations and 15 comparative prehistoric populations were used for genetic distance computation in Arlequin 3.5.1 [29], with the same substitution model as at the genetic distance calculation. Analogous to the PCA, we also integrated the reduced datasets (*) to exclude biases by potential maternal kinship. In a second MDS the Transdanubian data were pooled with the contemporaneous eastern Hungarian mtDNA data [3,13,22]. The F statistic was calculated based on 10,000 permutations. MDS was applied on the matrix of linearised Slatkin F_{st} values [32] and visualized in a two-dimensional space using the *metaMDS* function based on Euclidean distances implemented in the *vegan* library of R 3.0.2. [31] (Dataset S8-9, Figure S1, S3).

Analysis of molecular variance (AMOVA)

We arranged the HVS-I sequences (np 16056-16400) of the STA and LBKT together with the genetic data from nine archaeological cultures from Central Europe, ranging from the LBK to the Early Bronze Age into varying groups. According to our previous publication [6], we pooled the Central European cultures into 6^{th}-4^{th} millennia BC and 3^{rd}/2^{nd} millennium BC groups and subsequently transferred one or more cultures to the STA and LBKT until each group was subsumed with our samples. Overall we tested 82 different arrangements to find out the best combination indicated by the greatest among-groups and the least within-group variance (Dataset S11). Variance, F_{st}, and significant values (p) were computed with the standard AMOVA function implemented in Arlequin 3.5.1 [29] by using the Tamura & Nei substitution model [30] and a gamma value of 0.177. F_{st} values were tested on significance by 10,000 permutations.

Ancestral shared haplotype analysis (ASHA)

We used shared haplotype analysis [29] and modified this approach by accounting for the temporal succession of cultures in order to ascribe mtDNA haplotypes to particular cultures or time periods, and to identify the amount of ancestral lineages in each culture. Therefore, hunter-gatherers from Central/North Europe, the two investigated Carpathian Basin cultures, and nine cultures from Central Europe ranging from the LBK to the Early Bronze Age were placed into a chronological order. Each lineage within a given cultural dataset was traced back to its earliest match in the chronology and regarded as ancestral lineage that arose in this culture for the first

time (Dataset S12, Figure 3). This approach enabled us to estimate the amount of mtDNA lineages that were prevalent in Central European cultures since i) the Mesolithic, ii) the STA, iii) the LBKT, iv) the LBK, or v) that emerged in Central Europe in later periods. We counted the exact HVS-I haplotype matches between the prehistoric populations. However, haplotypes that were detected only among the farmers could also have potential hunter-gatherer origin (e.g. further U haplotypes). Therefore certain amount of ancestral hunter-gatherer lineages might remain undetected in this analysis.

Genetic distance maps

Genetic distance maps were generated from mtDNA and NRY data. HVS-I sequences (np 16068-16365) of the two investigated Neolithic Carpathian Basin populations were compared to 130 present-day Eurasian populations (Dataset S15, Figure S5ab). Genetic distances were calculated in Arlequin 3.5.1 [29] using the Tamura and Nei substitution model [30] and a gamma value of 0.177.

The Y chromosomal data of the STA, LBKT, ALBK [13] and LBK [5,14] were pooled and compared to 100 modern-day populations (Dataset S17, Figure S7). Haplogroup frequencies were differentiated into 16 groups (AB, D, E, F, G, H, C, I, J, KTS, L, NO, N, PR, R1, and R1a) and pairwise F_{st} values were computed in Arlequin 3.5.1 using the conventional F statistic.

Mitochondrial and Y chromosome genetic distances between the cultural datasets and modern-day populations were combined with longitudes and latitudes according to the sampling information in the literature and interpolated with the Natural Neighbor method implemented in ArcGis version 10.2 (Arcmap, Environmental Systems Research Institute [Esri] Inc, Redlands, USA).

References

1. Kalicz, N. 2010 An Grenze „zweier Welten" -Transdanubien (Ungarn) im Frühneolithikum. In *Die Neolithisierung Mitteleuropas. The Spread of Neolithic to Central Europe. International Conference, 24-26th June 2005. Mainz RGZM* (eds D. Gronenborn & J. Petrasch), pp. 235–254. Mainz: RGZM.

2. Oross, K. & Marton, T. 2012 Neolithic burials of the Linearbandkeramik settlement at Balatonszárszó and their European context. *Acta Archaeol. Acad. Sci. Hungaricae* **63**, 257–299. (doi:10.1556/AArch.63.2012.2.1)

3. Haak, W. et al. 2005 Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**, 1016–8. (doi:10.1126/science.1118725)

4. Haak, W. et al. 2008 Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. *Proc Natl Acad Sci U S A.* **105**, 18226–31. (doi:10.1073/pnas.0807592105)

5. Haak, W. et al. 2010 Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biol.* **8**, e1000536. (doi:10.1371/journal.pbio.1000536)

6. Brandt, G. et al. 2013 Ancient DNA reveals key stages in the formation of Central European mitochondrial genetic diversity. *Science* **342**, 257–261. (doi:10.1126/science.1241844)

7. Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. 1999 Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147. (doi:10.1038/13779)

8. Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A. & Villems, R. 2012 A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–84. (doi:10.1016/j.ajhg.2012.03.002)

9. Van Oven, M. & Kayser, M. 2009 Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–94. (doi:10.1002/humu.20921)

10. Bramanti, B. et al. 2009 Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**, 137–40. (doi:10.1126/science.1176869)

11. Bollongino, R., Nehlich, O., Richards, M. P., Orschiedt, J., Thomas, M. G., Sell, C., Fajkosová, Z., Powell, A. & Burger, J. 2013 2000 years of parallel societies in Stone Age Central Europe. *Science* **342**, 479–81. (doi:10.1126/science.1245049)

12. Fu, Q. et al. 2013 A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 1–7. (doi:10.1016/j.cub.2013.02.044)

13. Gamba, C. et al. 2014 Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257. (doi:10.1038/ncomms6257)

14. Haak, W. et al. 2015 Massive migration from the steppe is a source for Indo-European languages in Europe. *Nature* in Press, available on BioRxiv. (doi:http://dx.doi.org/10.1101/013433)

15. Chandler, H., Sykes, B. & Zilhão, J. 2005 Using ancient DNA to examine genetic continuity at the Mesolithic-Neolithic transition in Portugal. In *Actas dell III Congreso del Neolítico en la Península Ibérica, Santander, Monografías del Instituto internacional de Investigaciones Prehistóricas de Cantabria 1* (eds P. Arias R. Ontañón & C. García-Moncó), pp. 781–786.

16. Hervella, M., Izagirre, N., Alonso, S., Fregel, R., Alonso, A., Cabrera, V. M. & de la Rúa, C. 2012 Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PLoS One* **7**, e34417. (doi:10.1371/journal.pone.0034417)

17. Sánchez-Quinto, F. et al. 2012 Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Curr. Biol.* **22**, 1494–9. (doi:10.1016/j.cub.2012.06.005)

18. Bramanti, B. 2008 Ancient DNA: genetic analysis of aDNA from sixteen skeletons of the Vedrovice. *Anthropol.* **46**, 153–160.

19. Lee, E. J. et al. 2012 Emerging genetic patterns of the European Neolithic: perspectives from a late Neolithic Bell Beaker burial site in Germany. *Am. J. Phys. Anthropol.* **148**, 571–9. (doi:10.1002/ajpa.22074)

20. Lacan, M., Keyser, C., Ricaut, F.-X., Brucato, N., Tarrús, J., Bosch, A., Guilaine, J., Crubézy, E. & Ludes, B. 2011 Ancient DNA suggests the leading role played by men in the Neolithic dissemination. *Proc Natl Acad Sci U S A.* **108**, 18255–9. (doi:10.1073/pnas.1113061108)

21. Gamba, C. et al. 2012 Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. *Mol. Ecol.* **21**, 45–56. (doi:10.1111/j.1365-294X.2011.05361.x)

22. Szécsényi-Nagy, A., Keerl, V., Jakucs, J., Brandt, G., Bánffy, E. & Alt, K. W. 2014 Ancient DNA Evidence for a Homogeneous Maternal Gene Pool in sixth Millennium cal BC Hungary and the Central European LBK. In *Early farmers. The View from Archaeology and Science Proceedings of the British Academy 198.* (eds A. Whittle & P. Bickle), pp. 71–93. Oxford: Oxford University Press/British Academy.

23. Guba, Z., Hadadi, É., Major, Á., Furka, T., Juhász, E., Koós, J., Nagy, K. & Zeke, T. 2011 HVS-I polymorphism screening of ancient human mitochondrial DNA provides evidence for N9a discontinuity and East Asian haplogroups in the Neolithic Hungary. *J. Hum. Genet.* **56**, 784–96. (doi:10.1038/jhg.2011.103)

24. Bánffy, E., Brandt, G. & Alt, K. W. 2012 "Early Neolithic" graves of the Carpathian Basin are in fact 6000 years younger-appeal for real interdisciplinarity between archaeology and ancient DNA research. *J. Hum. Genet.* **57**, 467–9; author reply 470–1. (doi:10.1038/jhg.2012.36)

25. Fernández, E., Pérez-Pérez, A., Gamba, C., Prats, E., Cuesta, P., Anfruns, J., Molist, M., Arroyo-Pardo, E. & Turbón, D. 2014 Ancient DNA Analysis of 8000 B.C. Near Eastern Farmers Supports an Early Neolithic Pioneer Maritime Colonization of Mainland Europe

through Cyprus and the Aegean Islands. *PLoS Genet.* **10**, e1004401. (doi:10.1371/journal.pgen.1004401)

26. Nei, M. 1987 *Molecular Evolutionary Genetics*. New York: Columbia Univ. Press.

27. Librado, P. & Rozas, J. 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–2. (doi:10.1093/bioinformatics/btp187)

28. Fisher, R. A. 1922 On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**, 87–94.

29. Excoffier, L. & Lischer, H. E. L. 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–7. (doi:10.1111/j.1755-0998.2010.02847.x)

30. Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–26.

31. R Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available:http://www.R-project.org/.

32. Slatkin, M. 1995 A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–62.

33. Gronenborn, D. 1999 A variation on a basic theme : The transition to farming in southern Central Europe. *J. World Prehistory* **13**, 123–210.

34. Gronenborn, D. 2007 Beyond the models: "Neolithisation" in Central Europe. In *Going over: the Mesolithic-Neolithic transition in North-West Europe* (eds A. Whittle & V. Cummings), pp. 73–98. Oxford: Oxford University Press/British Academy.

35. Price, T. D., editor 2000 *Europe's first farmers*. Cambridge: Cambridge University Press.

36. Oross, K. & Bánffy, E. 2009 Three successive waves of Neolithisation: LBK development in Transdanubia. *Doc. Praehist.* **36**, 175–189. (doi:10.4312/dp.36.11)

1

Supplementary Table

Table S1. Mt haplogroup frequencies, Fisher's exact test and genetic distances of four prehistoric metapopulations and archaeological cultures.

Culture name		**Hunter-gatherers in Central and North Europe**	**Population of the Starčevo culture**	**Linearbandkeramik culture's population in Transdanubia**	**Linearbandkeramik culture's population in Central Europe**
Culture abbreviation		**HGCN**	**STA**	**LBKT**	**LBK**
n		**23**	**44**	**39**	**108**
mtDNA haplogroup frequencies	**H**	0	6.82	30.77	16.67
	HV	0	2.27	2.56	4.63
	V	0	6.82	2.56	4.63
	J	0	11.36	7.69	12.04
	K	0	27.27	12.82	20.37
	N1a	0	6.82	10.26	12.04
	R	4.35	0	0	0
	T1	0	2.27	2.56	0.00
	T2	0	20.45	25.64	22.22
	U	8.70	0	0	0
	U2	4.35	0	2.56	0
	U3	0	2.27	0	0.93
	U4	8.70	2.27	0	0
	U5a	17.39	0	2.56	1.85
	U5b	52.17	0	0	0.93
	U8	4.35	0	0	0
	W	0	4.55	0	2.78
	X	0	6.82	0	0.93
Fisher's test p values	**HGCN**	*			
	STA	0.00009999	*		
	LBKT	0.00009999	0.06389	*	

	LBK	0.00009999	0.2707	0.5518	*
Fst values	**HGCN**	*			
	STA	0.18700	*		
	LBKT	0.18552	0.01345	*	
	LBK	0.17642	0.00143	-0.00521	*
Fst p values / *adjusted p values*	**HGCN**	*	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
	STA	*0.00000*	*	0.15527+-0.0099	0.33984+-0.0138
	LBKT	*0.00000*	*0.46581*	*	0.62012+-0.0156
	LBK	*0.00000*	*0.50976*	*0.62012*	*
References		[10–13], this study	this study	this study	[3,5,6,18]

Fisher's exact test was based on mtDNA haplogroup frequencies. Genetic distances or Fst values were calculated from HVS-I sequences (np 16056-16400). Genetic distance p values were post hoc adjusted to correct for multiple comparison by Benjamin and Hochberg method (italicized). Culture and population information are presented in Dataset S6.

Supplementary Figures

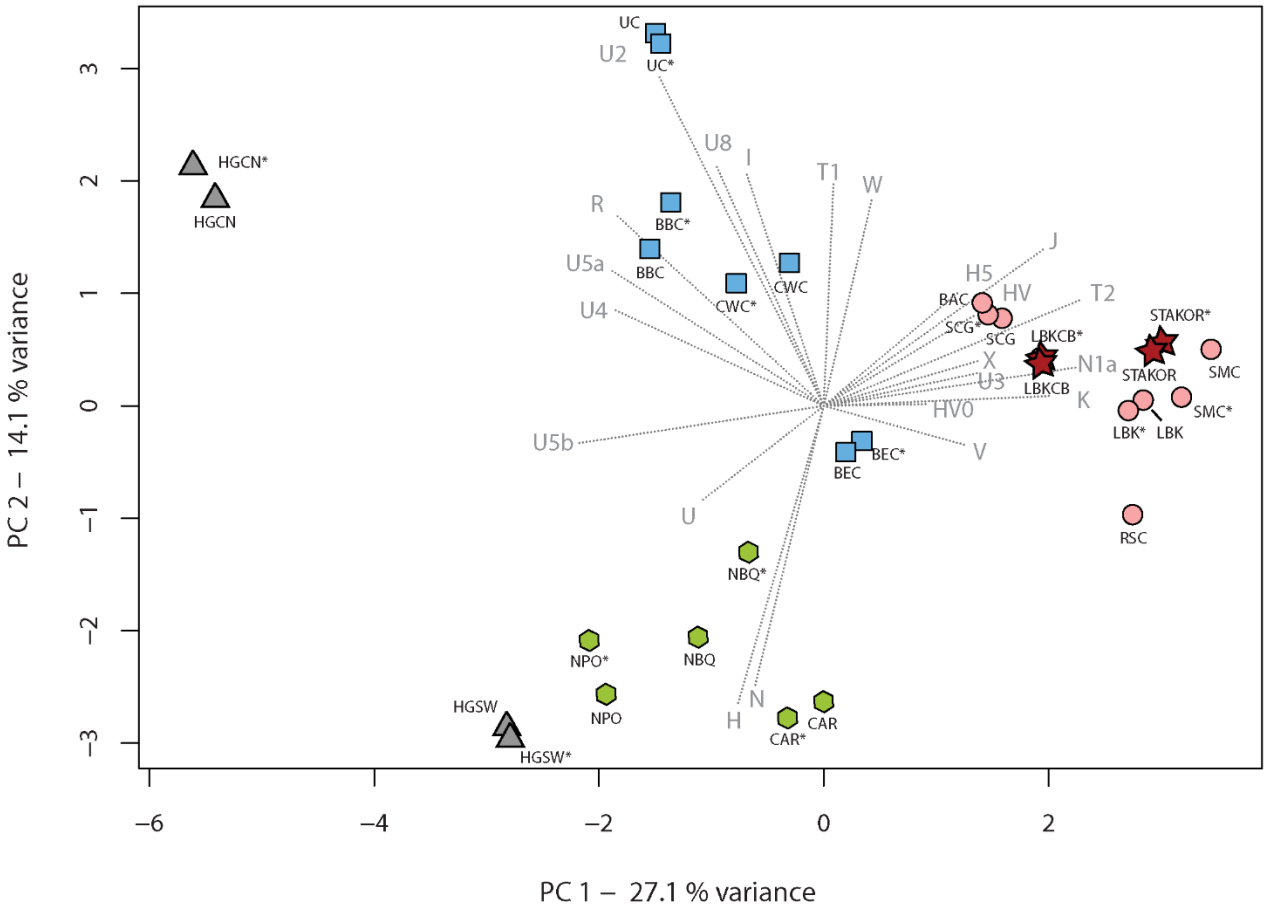

Figure S1. PCA with the pooled contemporaneous Neolithic western and eastern Carpathian Basin datasets.

PCA is based on the frequencies of 22 mtDNA haplogroups in the Neolithic Carpathian Basin and 14 comparative prehistoric populations. Colour shadings and symbols denote cultures of different periods or European regions: hunter-gatherers (grey triangles), 6th millennium BC Carpathian Basin populations (red stars), LBK and 5th/4th millennia BC populations in Central Europe (rose circles), Central European 3rd/2nd millennia BC populations (blue rectangles), 6th/5th millennia BC populations of the Iberian Peninsula (green hexagons). The reduced version of each dataset is marked by an asterisk (*). Detailed information about the comparative data and haplogroups frequencies are listed in Dataset S8. Population abbreviations: hunter-gatherers in Central and North Europe (HGCN), hunter-gatherers in Southwestern Europe (HGSW), Starčevo and Körös (STAKOR), LBK in Transdanubia and Alföld (LBKCB), LBK in Central Europe (LBK), Rössen (RSC), Schöningen (SCG), Baalberge (BAC), Salzmünde (SMC), Bernburg (BEC), Corded Ware (CWC), Bell Beaker (BBC), Únětice (UC), Cardial and Epicardial (CAR), Neolithic population in Basque Country and Navarre (NBQ), Neolithic in Portugal (NPO).

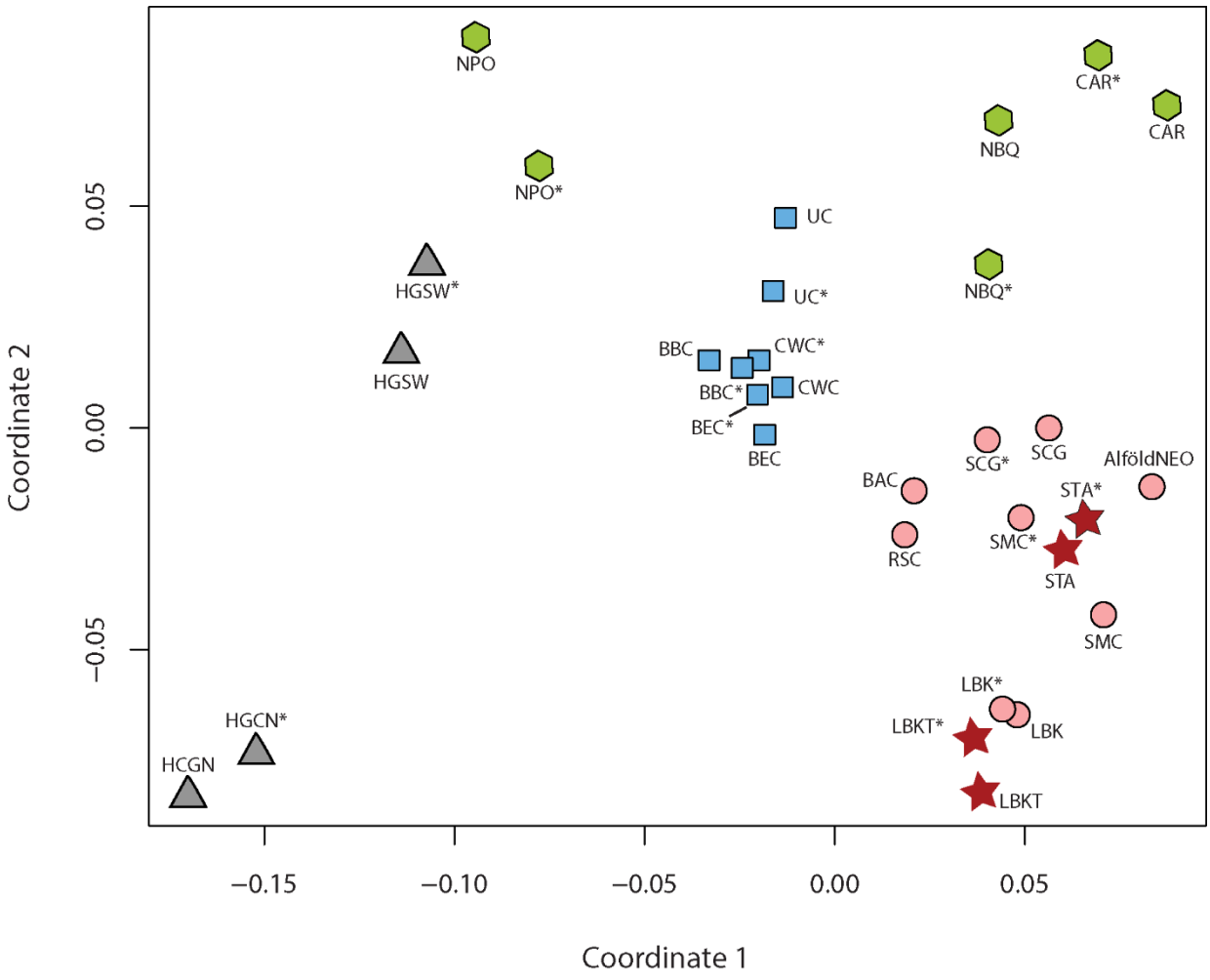

Figure S2. Multidimensional scaling of 17 prehistoric populations.

Genetic distances (F_{st}) were computed between the STA, LBKT and 15 prehistoric populations, and visualized by multidimensional scaling with a stress value of 0.1195. Colour shadings and symbols denote populations of different periods or European regions according to Figure S1. The reduced version of each dataset is marked by an asterisk (*). Detailed information about the comparative data and F_{st} values are listed in Dataset S9.

New abbreviation on this plot: Neolithic Alföld (eastern Hungary) (AlföldNEO), Starčevo (STA), LBK in Transdanubia (LBKT). For further abbreviations, see legend of Figure S1.

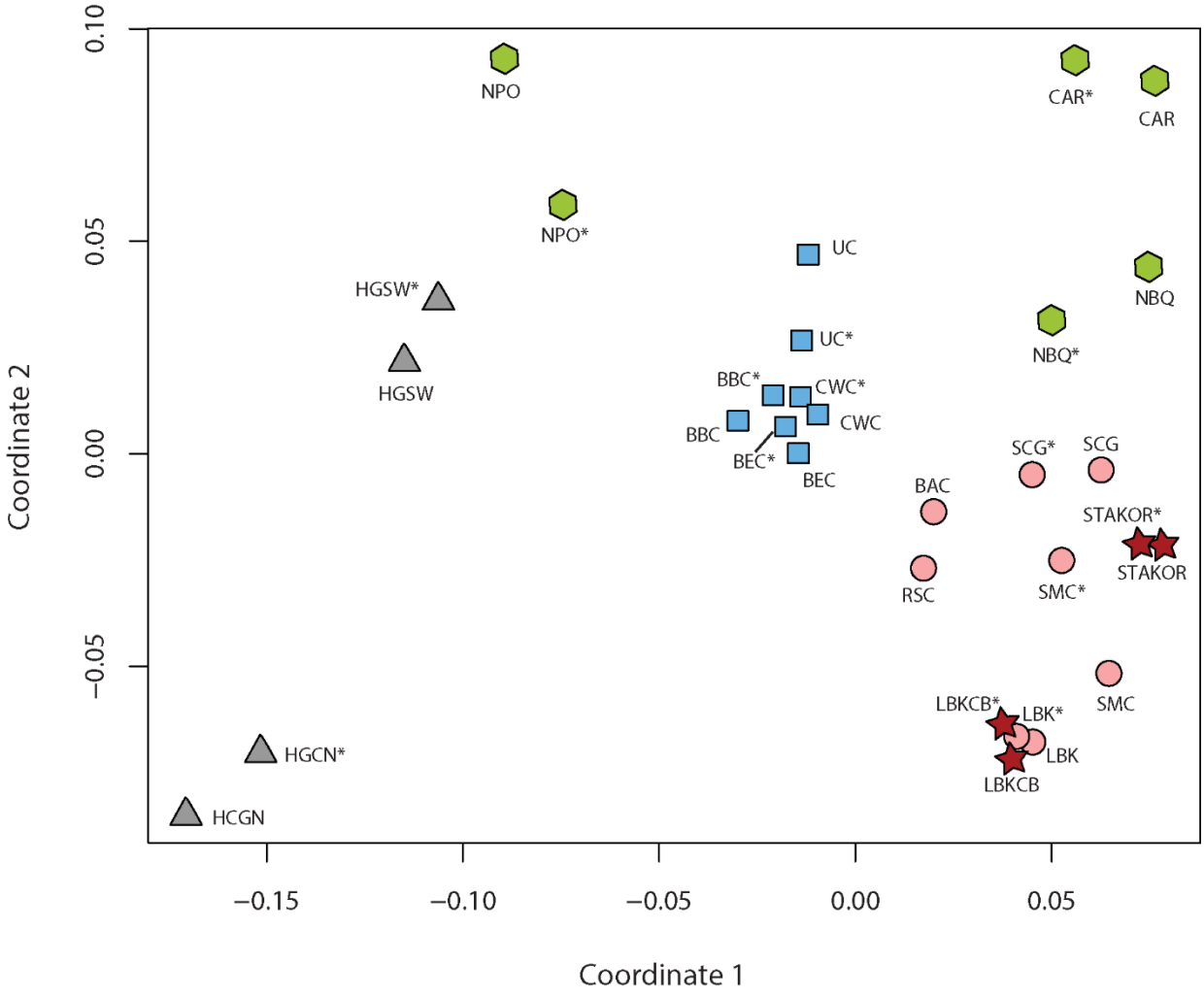

Figure S3. MDS with the pooled Neolithic contemporaneous western and eastern Carpathian Basin datasets.

The published contemporaneous Neolithic mtDNA data from eastern Hungary were pooled on this MDS with the Transdanubian datasets. Stress value is 0.1049. Abbreviations are according to Figure S1. Colour shadings and symbols denote populations of different periods or European regions according to Figure 2. The reduced version of each dataset is marked by an asterisk (*). Detailed information about the comparative data and F_{st} values are listed in Dataset S10.

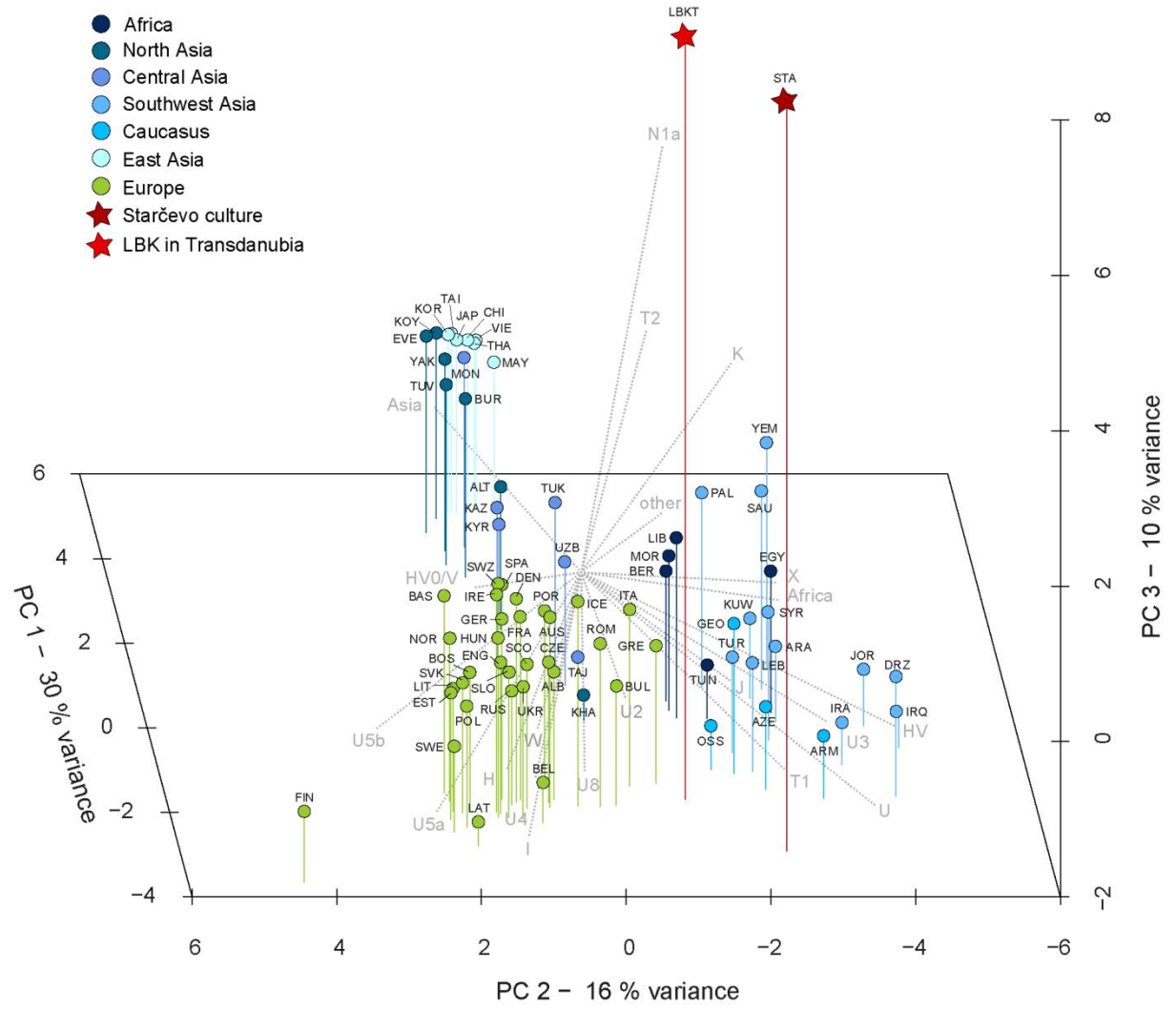

Figure S4. PCA comparing the Starčevo, LBKT and modern-day mtDNA data.

PCA of the STA, LBKT and 73 present-day populations of Eurasia, and North Africa plotted in a three dimensional space. Colours indicate populations from different Eurasian and African regions. The contribution of each haplogroup is superimposed as grey component loading vector. The first three principal components of the PCA display 56% of the total genetic variation. Population information and haplogroup frequencies are listed in Dataset S14.

Population codes: Linearbandkeramik culture in Transdanubia (LBKT), Starčevo culture (STA), Albanians, Macedonians (ALB), Altai (ALT), Arabs in UAE, Oman, Qatar (ARA), Armenians (ARM), Austrians (AUS), Azeris (AZE), Basques (BAS), Belarusians (BEL), Berber (BER), Bosnians, Croatians, Serbians (BOS), British (ENG), Bulgarians (BUL), Buryats (BUR), Chinese, Tibetan (CHI), Czechs (CZE), Danes (DEN), Druze (DRZ), Egyptians (EGY), Estonians (EST), Evenks (EVE), Finns (FIN), French (FRA), Georgians (GEO), Germans (GER), Greeks (GRE), Hungarians (HUN), Icelanders (ICE), Iranians (IRA), Iraqi (IRQ), Irish (IRE), Italians (ITA), Japanese (JAP), Jordanians (JOR), Kazakhs (KAZ), Koreans (KOR), Koryaks (KOY), Kuwaiti (KUW), Kyrgyz (KYR), Latvians (LAT), Lebanese (LEB),

Libyans (LIB), Lithuanians (LIT), Malays (MAY), Khants, Mansi (KHA), Mongolians (MON), Moroccans (MOR), Norwegians (NOR), Ossetians (OSS), Palestinians (PAL), Poles (POL), Portuguese (POR), Romanians (ROM), Russians (RUS), Saudi Arabians (SAU), Scots (SCO), Slovaks (SVK), Slovenians (SLO), Spaniards (SPA), Swedes (SWE), Swiss (SWZ), Syrians (SYR), Taiwanese (TAI), Tajiks (TAJ), Thai (THA), Tunisians (TUN), Turkmens (TUK), Turks (TUR), Tuvinians (TUV), Ukrainians (UKR), Uzbeks (UZB), Vietnamese (VIE), Yakuts, Yukaghir (YAK), Yemenis (YEM).

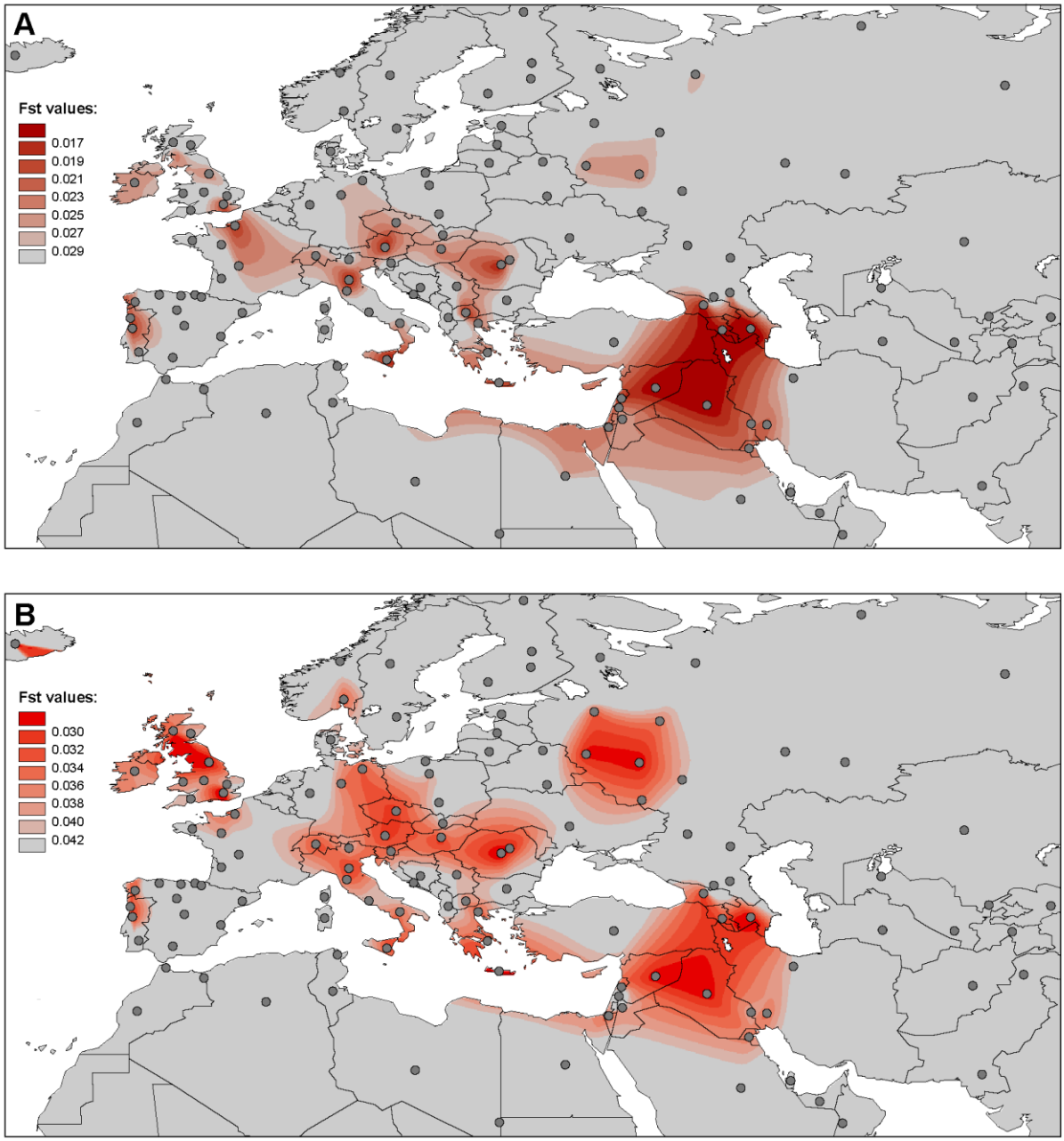

Figure S5a-b. MtDNA genetic distance maps of the STA, LBKT and 130 present-day populations.

Genetic distances (F_{st}) between the STA (A) and LBKT (B) and 130 present-day populations of Eurasia and North Africa were computed based on HVS-I sequences and visualized on a geographic map. Grey dots denote the location of modern populations. Colour shadings indicate the degree of similarity or dissimilarity of the Neolithic cultures to these populations. Short distances and great similarities are marked by dark red (STA) and red (LBKT) areas. F_{st} values are scaled by an interval range of 0.002. F_{st} values higher than 0.029 (STA), 0.042 (LBKT) were not differentiated (grey areas). Population information and F_{st} values are listed in Dataset S15.

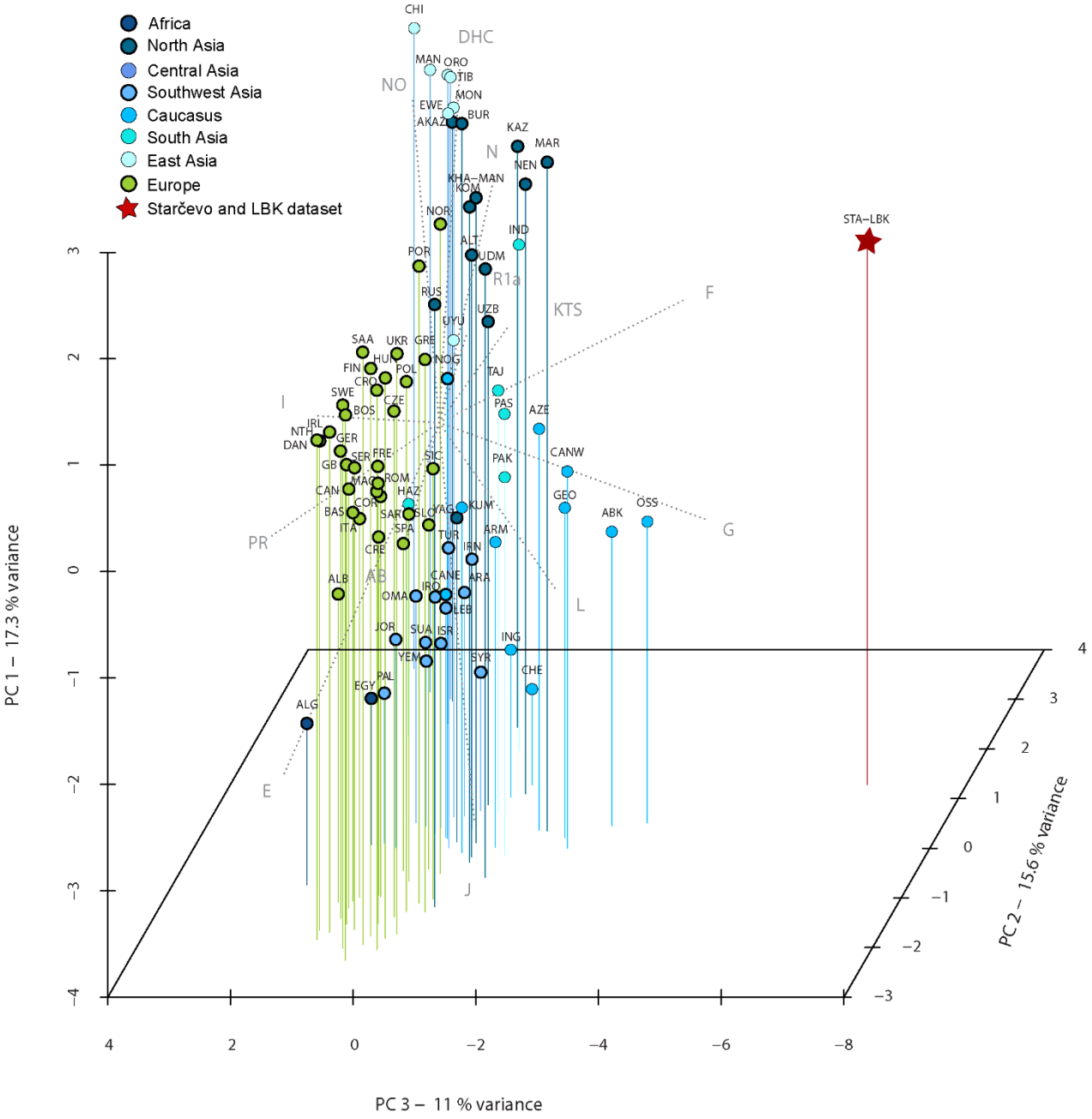

Figure S6. PCA from Y chromosomal data of the STA-LBK sample set and 80 modern-day populations.

The PCA, based on the frequencies of 13 Y chromosomal haplogroup in the STA-LBK sample set (n=19) and 80 present-day populations from Eurasia and North Africa, were plotted in a three dimensional space. Colours of data points indicate populations from different Eurasian and African regions. The contribution of each haplogroup is superimposed as grey component loading vector. The first three principal

components of the PCA display 43.9% of the total genetic variation. Population information and haplogroup frequencies are listed in Dataset S16.

Population codes: Neolithic Starčevo and Linearbandkeramik population (STA-LBK), Abkhazians (ABK), Albanians (ALB), Algerians (ALG), Altai (ALT), Altai Kazakhs (AKAZ), Arabs in UAE, Qatar, Kuwait (ARA), Armenians (ARM), Azeri (AZE), Basques (BAS), Bosnians (BOS), British (GB), Buryats (BUR), north-east Caucasus (CANE), north-west Caucasus (CANW), Chechens (CHE), Chinese (CHI), Corsicans (COR), Crete (CRE), Croatians (CRO), Czech-Slovakians (CZE), Danish (DAN), Dutch (NTH), Egyptians (EGY), Ewenki (EWE), Finns (FIN), French (FRE), Georgians (GEO), Germans (GER), Greek (GRE), Hazara (HAZ), Hungarians (HUN), Indian (IND), Ingush (ING), Iranians (IRN), Iraqis (IRQ), Irish (IRL), Israeli (ISR), Italians (ITA), Jordanians (JOR), Kazakhs (KAZ), Komi (KOM), Kumyks (KUM), Lebanese (LEB), Macedonians (MAC), Manchu (MAN), Mansi & Khanti (KHA-MAN), Mari (MAR), Mongolians (MON), Nenets (NEN), Nogays & Kara Nogays (NOG), Norwegians (NOR), Omani (OMA), Oroqen (ORO), south-north Ossetians (OSS), Pakistani (PAK), Palestinian (PAL), Pashtun (PAS), Poles (POL), Portuguese (POR), Romanians (ROM), Russians (RUS), Sami (SAA), Sardinians (SAR), Saudi Arabians (SAU), Serbians (SER), Sicilians (SIC), Slovenians (SLO), Spaniards (SPA), Spaniards in Canary Islands (CAN), Swedes (SWE), Syrians (SYR), Tajik (TAJ), Tibetans (TIB), Turks (TUR), Udmurts (UDM), Ukrainians (UKR), Uyghurs (UYU), Uzbeks (UZB), Yagnobi (YAG), Yemeni (YEM).

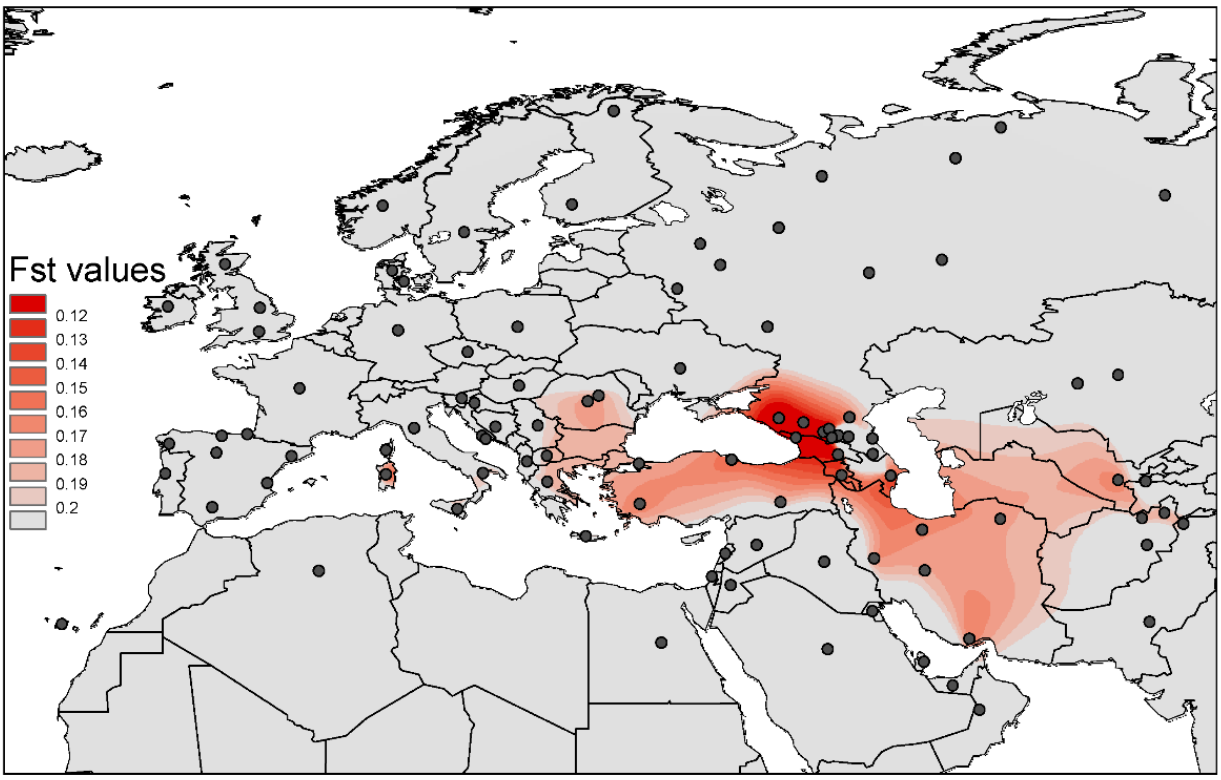

Figure S7. Genetic distance map of the STA-LBK Y chromosomal data.

Y chromosomal genetic distances (F_{st}) were computed between the STA-LBK samples and 100 present-day populations of Eurasia and North Africa and visualized on a geographic map. Grey dots denote the location of present-day populations. Colour shadings indicate the degree of similarity or dissimilarity of Neolithic samples to the modern-day populations. Short distances and great similarities to present-day populations are marked by red areas. F_{st} values were scaled by an interval range of 0.01. F_{st} values higher than 0.21 were not differentiated (grey areas). Population information and F_{st} values are listed in Dataset S17.

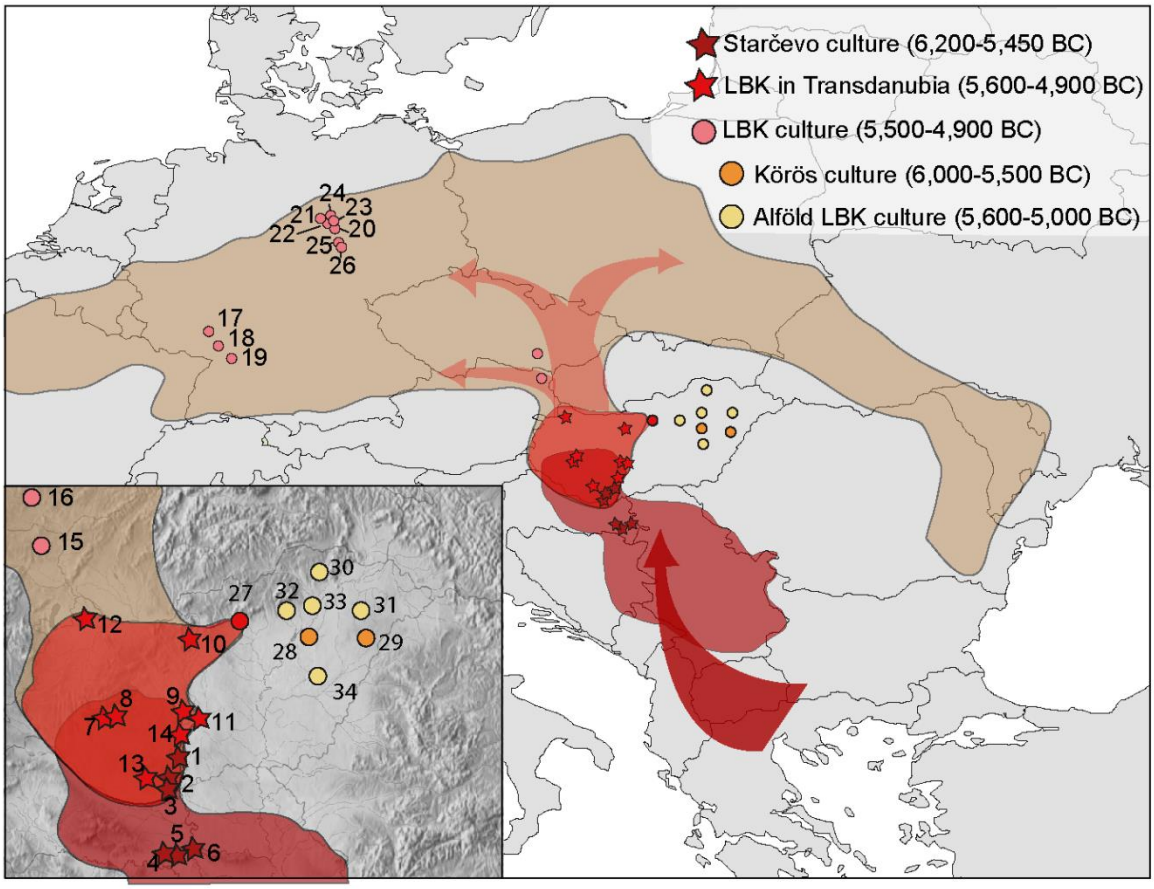

Figure S8. Geographic distribution of the Starčevo, LBK cultures and locations of the studied sites.

The shaded areas of the maps show the distribution of the Starčevo culture (STA) and the LBK in Transdanubia (LBKT) and Central Europe (LBK) [1,33–36]. The absolute chronological frames refer to the whole dissemination area of each culture. The arrows show the direction of the farmers' expansion into Central Europe, suggested by the above cited archaeological records. Coloured circles indicate the studied sites of the STA (dark red) and LBKT (red) in western Hungary and northern Croatia. Central European LBK sites (rose), Körös sites (orange) and Alföld LBK sites (yellow) that were included in the comparative analyses are presented as well. 1: Alsónyék-Bátaszék, Mérnöki telep, 2: Lánycsók-Csata-alja, 3: Lánycsók-Gata-Csotola, 4: Vinkovci-Nama, 5: Vinkovci-Jugobanka, 6: Vukovar-Gimnazija, 7: Balatonszárszó-Kis-erdei-dűlő, 8: Balatonszemes-Bagódomb, 9: Bölcske-Gyűrűsvölgy, 10: Budakeszi-Szőlőskert, 11: Harta-Gátőrház, 12: Kóny II. Proletár-dűlő 85., 13: Szemely-Hegyes, 14: Tolna-Mözs, 15: Asparn Schletz 2, 16: Vedrovice, 17: Flomborn, 18 Schwetzingen, 19: Vaihingen, 20: Seehausen, 21: Derenburg, Merenstieg II, 22: Halberstadt, Sonntagsfeld, 23 Oberwiederstedt I, 24: Eilsleben, 25: Karsdorf, 26: Naumburg [3,5,6,18]., 27: Apc-Berekalja-I, 28: Tiszaszőlős-Domaháza, 29: Berettyóújfalu-Morotva-liget, 30: Garadna, 31: Debrecen Tócópart Erdőalja, 32: Kompolt-Kígyósér, 33: Polgár-Ferenci-hát,[13] 34: Ecsegfalva 23 [3].

Supplementary Datasets

Dataset S1: Basic information of the archaeological sites.

Dataset S2: Anthropological data and ^{14}C dating of the sampled skeletons.

Dataset S3: Summary of mtDNA and Y chromosomal results from western Hungary and Croatia.

Dataset S4: Detailed results of the GenoCoRe22 SNP multiplex assay.

Dataset S5: Detailed results of the GenoY25 SNP multiplex and singleplex PCR analyses.

Dataset S6: Summary of published prehistoric mtDNA data used for population genetic analyses.

Dataset S7: MtDNA haplogroup frequencies of 17 hunter-gatherer, Neolithic, and Early Bronze Age populations, used for PCA.

Dataset S8: MtDNA haplogroup frequencies of 16 hunter-gatherer, Neolithic, and Early Bronze Age populations, used for PCA.

Dataset S9: Fst values, their p values and Slatkin matrix of 17 prehistoric populations.

Dataset S10: Fst values, their p values and Slatkin matrix of 16 prehistoric populations.

Dataset S11: Results of AMOVA.

Dataset S12: Results of ASHA of the Central/North European hunter-gatherers, Neolithic and Early Bronze Age populations.

Dataset S13: Results of TPC with STA, LBKT, LBK and the hunter-gatherer metapopulation of Central/North Europe.

Dataset S14: Population information and mtDNA haplogroup frequencies used for PCA with Neolithic Starčevo, LBKT populations and 73 present-day populations.

Dataset S15: Population information and Fst values used for mapping the mitochondrial genetic distances between the Neolithic Carpathian Basin populations and 130 present-day populations.

Dataset S16: NRY haplogroup frequencies of 80 modern populations and the combined Starčevo-LBK dataset.

Dataset S17: Information on 100 modern-day populations and their Fst values from the STA-LBK dataset, and used for the Y chromosomal genetic distance map.

Dataset S18: List of primers, used for the mtDNA and NRY singleplex amplifications.

Dataset S19: Genetic profiles of the co-workers.

Dataset S20: Summary of published prehistoric Y chromosomal data from Eurasia.