Molecular Genetics and Genomics
https://doi.org/10.1007/s00438-018-1469-7

METHODS PAPER

A new self-learning computational method for footprints of early human migration processes

Z. Juhász[1] · E. Dudás[2] · Horolma Pamjav[2]

Received: 13 February 2018 / Accepted: 28 June 2018

Abstract

We present a new self-learning computational method searching for footprints of early migration processes determining the genetic compositions of recent human populations. The data being analysed are 26- and 18-dimensional mitochondrial and Y-chromosomal haplogroup distributions representing 50 recent and 34 ancient populations in Eurasia and America. The algorithms search for associations of haplogroups jointly propagating in a significant subset of these populations. Joint propagations of Hgs are detected directly by similar ranking lists of populations derived from Hg frequencies of the 50 Hg distributions. The method provides us the most characteristic associations of mitochondrial and Y-chromosomal haplogroups, and the set of populations where these associations propagate jointly. In addition, the typical ranking lists characterizing these Hg associations show the geographical distribution, the probable place of origin and the paths of their protection. Comparison to ancient data verifies that these recent geographical distributions refer to the most important prehistoric migrations supported by archaeological evidences.

Keywords Y-chromosomal and mtDNA haplogroups · Archaeogenetics · Artificial intelligence · Self-learning algorithm · Clustering · Rank correlation

Introduction

Starting from the beginning of the millennium, paternal and maternal lineages based on Y-chromosomal MSY and mtDNA have been studied for population migration history in a chain reaction (Jobling and Tyler-Smith 2003; Underhill and Kivisild 2007; Karafet et al. 2008; Cinnioglu et al. 2004; Tambets et al. 2004; Semino et al. 2000; Yao et al. 2004; Pakendorf et al. 2007; Bermisheva et al. 2004; Simoni et al. 2000; Quintana-Murci et al. 2004).

The mtDNA and Y-chromosomal lineages seem to support the hypothesis that reconstructing the demographic history of human migration histories, which highlights a recent increase in effective population size, is compatible with admixture of both lineages between continents and geographic regions.

The last 10 years have witnessed a revolution in ancient DNA (aDNA) research. Genetic studies of ancient and modern populations significantly contributed to the picture drawn previously by archaeologists about prehistoric processes resulting in the contacts between different ancient cultures and populations by the next generation sequencing (NGS) methods (Skoglund et al. 2012; Lazaridis et al. 2014, 2016; Fu et al. 2016; Haak et al. 2015; Allentoft et al. 2015).

The sequencing focus was previously limited to hypervariable regions of mitochondrial DNA. Nowadays whole genome sequences are connected to the massive sequence throughput of next generation sequencing platforms able to target short and degraded DNA. Many ancient specimens being previously unsuitable for DNA analyses because of degradation can now successfully be used as templates for sequencing. At present, not only mitochondrial but also nuclear whole genomes have been sequenced from archaic

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-018-1469-7) contains supplementary material, which is available to authorized users.

✉ Horolma Pamjav
phorolma@hotmail.com

[1] Centre for Energy Research, Institute of Technical Physics and Materials Science, PO Box. 216, Budapest 1536, Hungary

[2] National Centre of Experts and Research, Institute of Forensic Genetics, Budapest, Hungary

hominins, ancient anatomically modern humans, and present-day populations (Lazaridis et al. 2016; Fu et al. 2016; Haak et al. 2015; Allentoft et al. 2015; Batini et al. 2017; Lopopolo et al. 2016; Ilyas et al. 2015; Ermini et al. 2015; Der Sarkissian et al. 2015). Ancient DNA analysis of autosome can provide detailed scenario of admixture. However, populations in different geographic locations tend to have their own special sub-lineages of Y-chromosome and mtDNA. Therefore, the studies of Y-chromosome and mtDNA have potential to yield better resolution than that of autosome when studying the origin and migration of human population.

A powerful method based on PCA of the Fst distance matrix of 101 ancient individuals arising from the period of 3400–200 BC indicated genetic transitions well corresponding to archaeological findings in Eurasia. The comparison to recent Fst data showed the connection between contemporary and Bronze-Age populations (Allentoft et al. 2015).

A significantly different approach is based on clustering of haplogroup (Hg) frequency distributions of recent and ancient populations, instead of analysis of pairwise Fst distances of individuals (Juhász et al. 2015, 2016). In this case, ancient and recent populations belonging to a common cluster directly point to the genetic connection of complete populations.

Representing populations by their Hg distributions proposes the following assumption: recent populations are products of prehistoric and historic interactions, disjunctions and junctions of certain ancient source populations [for instance, the admixture of indigenous European hunter-gatherers with Neolithic farmers arising from the Middle East resulted in a new population with an Hg distribution containing both European and Middle Eastern components (Skoglund et al. 2012)]. It has been verified by stepping stone simulation that such admixture processes starting from a few source areas result in strong correlations between Hgs arising from a common starting population, because they propagate necessarily jointly for a long time. Consequently, a search for strongly correlated Hgs in recent populations can reveal the Hg content of these ancient source populations (Juhász et al. 2016; Neparáczki et al. 2017). However, iterative rank correlation search applied in these studies finds pairwise correlations, so larger subsets of the correlating Hgs are hardly detectable. In addition, that method did not utilize an important advantage of rank correlation technique, namely that the ranks attributed to the populations may refer to the emitting source populations of jointly propagating "Hg associations".

In this paper, we present a new computational method aiming to reveal such groups of mitochondrial and Y-chromosomal haplogroups jointly propagating in a significant subset of contemporary Eurasian and American indigenous populations. In order to verify our starting assumption, namely that strong correlation of Hgs in recent populations may refer to ancient source populations, we compare the results to ancient mitochondrial Hg distributions.

Materials and methods

Materials

We analysed 50 populations for mtDNA, as well as 50 populations for Y-chromosomal haplogroups. The frequencies for mtDNA haplogroups and Y-chromosomal haplogroups together with publication sources and a three-letter code was used to label each population as presented in Online Resource 1 (ESM_1) and Online Resource 2 (ESM_2). Furthermore, 34 ancient mtDNA haplogroup distributions were used for this study that is also included in Online Resource 1 (ESM_1). The total sums of individuals represented by mitochondrial, Y-chromosomal and ancient mitochondrial Hg distributions are 13,631; 6746 and 1266, respectively.

Mitochondrial and Y-chromosomal data do not perfectly coincide in three cases, when Tuscany–Sicilian, Serbian–Croatian, as well as Karachay–Balkar data are combined, marked by the abbreviations TUS, SRB and KRC. Based on the close geographical, linguistic and historical contacts, we suppose that these couplings do not interfere the analysis in a significant manner.

The populations and the corresponding abbreviations of the modern data are shown in Table 1.

The ancient population mtDNA data, sample sizes, abbreviations, places and times of origin are included in Table 2.

The aim of this study was to test the new method for the genetic results accepted by scientific community, so we did not focus on the resolution of haplogroups, therefore we used mainly the distribution of the basic haplogroups to compare as many populations as possible.

Methods

Here, we present a new computational method aiming to reveal all groups of Hgs jointly propagating in a significant subset of 50 contemporary Eurasian and American indigenous populations. In rank correlation calculation of two Hgs, ranks are attributed to the 50 populations studied, according to the frequencies of the given Hgs in their Hg distributions. After that, the rank correlation coefficient is defined as the well-known linear correlation coefficient of the resulting two rank lists. Obviously, these 50-element rank lists of strongly correlating Hgs are necessarily similar therefore the whole set of the corresponding 50-dimensional vectors constitutes a clustered point system in its vector space. Thus, we reduce the search for groups of strongly correlating Hgs to clustering of their rank lists (vectors), instead of analysing the totally puzzling network of pairwise correlations.

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

Table 1 The populations and the corresponding abbreviations of the modern data

Population name	Abbreviation	Population name	Abbreviation
Han Chinese	CHN	Mongolian	MNG
Kyrgyz	KYG	Chuvash	CHU
Tuscany	TUS	Bulgarian	BLG
Azeri	AZR	Turkish	TUR
Karachay	KRC	Hungarian	HUN
Slovak	SLK	Czech	CZH
Romanian	ROM	Kashubian Poles	PLK
Finnish	FIN	Norwegian	NOR
North German	GEN	South German	GES
French	FRA	Netherlands	DUT
Scottish	CO	Galicia	GAL
Northwest Amerindian	NAW	Komi Zyryan	KOZ
Khanty	KHA	Serbian	SRB
Kurdish	KUR	Russian	RUS
Central Amerindian	NAC	Warmian	War
Poles	POL	Southern Amerindian	NAS
Greek	GRE	Estonian	EST
Saami	SAA	Karelian Finn	KRL
Ukrainian	UKR	Uyghur	UYG
Kazakh	KAZ	Mari	MRI
Tatar	TAT	Udmurt	UDM
Japanese	JPN	Székely	SEK
Altai Kazakh	ALK	Hui Chinese	HUI
Macedonian	MAC	Lithuanian	LIT
Tuvan	TUV	–	

We describe the determination of jointly propagating groups of Hgs using self organizing cloud (SOC) clustering of the inverse rank vectors as follows.

The basic assumption of this work is that there may exist certain "Hg-associations" with characteristic compounds of mitochondrial and Y-chromosomal Hgs. It also seems a realistic assumption that the members of these associations of Hgs were jointly emitted from certain "source populations" for a long period, therefore their correlation subsets and rank sequences became similar as a result of the migrations and admixtures in historic and prehistoric times. If this is true, the rank lists of the correlation subsets belonging to a given Hg-association may form different separable clusters.

In principle, the problem of finding of characteristic Hg-associations could be reduced to a clustering of the 44×44-dimensional symmetric matrix containing the rank correlation values of the 26+18=44 mitochondrial and Y-chromosomal Hgs. Rank correlation is itself a similarity measure, therefore distance-based clustering algorithms like *k-medoids*, nearest neighbour, *k* nearest neighbours, maximal relation probability could be applied for this purpose. However, the first experiments have shown that rank correlations show a rather fuzzy structure with hardly identifiable clusters. Therefore, we developed another clustering method based on training the so-called SOC algorithm by the "inverse rank vectors" derived from the iterative rank correlation algorithm. This method allowed us to simultaneously identify both the Hg-associations propagating regularly together as genetic components of certain propagating source-populations, and the groups of these propagating populations themselves.

The "inverse rank vector" (IRV) of a Hg is defined as follows.

Firstly, we execute the iterative rank correlation search for each pair of our 26+18=44 Hgs. Due to the iterations eliminating the populations causing the largest decrease of the correlation, the algorithm correlating the *i*th Hg (Hg*i*) to 44 other Hgs (including Hg*i* itself) results in 44 different rank lists for Hg*i*. We select the rank lists belonging to a correlation value higher than a critical value (0.7) from this set of rank lists. After finishing the whole process, we obtain a set of rank lists, each of them belonging to a strong correlation, while all other couplings of Hgs having no detectable correlation are eliminated.

Let $r_k(i)$ denote the rank of the *i*th population in the rank list of the *k*th member of the above rank list set. The corresponding "inverse rank" value is defined as

$$\bar{r}_k(i) = 1 - r_k(i)/\max(r_k(i)) \quad (1)$$

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

Table 2 The ancient populations, mtDNA sample sizes, abbreviations, places and times of origin are in the table

Middle East Neolithic-BrA	28	MEN	Middle East	11,840–1402
Iberian Neolithic	45	IBN	Iberia	10,310–3160
Near Eastern Neolithic	67	NEO	TR, IRN, SYR, JOR	8300–4000
Cisbaikalian Neolithic (Serovo)	15	SER	East Siberia	8000–4000
Early-Middle Neolithic	53	EMN	Europe	6000–3000
Starcevo	44	STR	Balkans	5700–5500
Dniepr–Donets Neolithic	17	DDO	Eastern Europe	5300–4700
Neolithic Hungary	85	NHU	Hungary	5200–4800
Yamnaya, Afanasievo	49	YAM	Russia, Ukraine	5000–2700
Kurgans (Eneolithic/Catacomb)	35	KGC	UKR, MLD, BLG	4700–2000
Baraba (UT-ODI-EK)	33	BB1	West Siberia	4000–1800
Late Neolithic-EBA Europe	56	LNB	Europe	3000–1600
Altai Bronze Age	12	ABA	South Siberia	2700–900
Tarim Basin Xiaohe	73	XIA	China	2515–1829
Sintashta-Andronovo	41	SIA	Russia, Siberia	2300–1400
Baraba (LK-FYOD-LBB)	45	BB2	West Siberia	1800–1000
Srubnaya	14	SRU	Russia	1800–1200
Bronze Age Kurgans	13	KBK	Kazakhstan	1400–1000
Baraba (Iron transition)	14	BB3	West Siberia	1000–800 BC
Iron Age Kurgans	13	KIK	Kazakhstan	800–600 BC
Tagar–Tachtyk	15	TAG	Russia	800 BC–400 AD
Scythian Iron age	14	SCI	Russia	600–200 BC
Pazyryk Scytho-Siberian	25	PAZ	Mongolia, Russia	400–200 BC
Qin China aDNA	19	QIN	East Asia	221 BC–210 AD
Egyin Gol Xiongnu	46	XIO	Inner Asia	200 BC–200 AD
Lombard early medieval	40	LOM	Hungary, Italy	500–800 AD
Vikings	65	VIK	Norway	780–790 AD
Karos	90	KAR	Hungary	850–900 AD
Hungarians 900 AD	27	AH2	Central Europe	900–1000 AD
Ancient Hungarian (tenth century)	67	AH1	Central Europe	900–1000 AD
Pre-conquest Hungary	49	HPC	Hungary	500–900 AD
Medieval Slavic	19	SLV	Slovakia	900–1200 AD
Italian medieval	27	ITM	Italy	900–1500
Cumanian	11	CUM	Hungary	1200–1300 AD
Total sample size	1266			

for the populations contained by the *k*th correlation subset, while the populations missing from the correlation subset have the inverse rank value $\bar{r}_k(i) = 0$.

Using this definition, the population having the highest frequency—and consequently the lowest rank in the ranking list of Hg*i*—becomes the highest "inverse rank" value approaching 1, and the ranks of the other populations decrease with decreasing frequency of Hg*i*. Thus, our "inverse rank" is really the inverse of the common rank definition, which increases with decreasing frequency in Hg*i*. The inverse ranks $\bar{r}_k(i)$ are stored in the N-dimensional "inverse rank vectors (IRV)" $\underline{R}(k)$, where the *i*th vector element represents the inverse rank of the *i*th population in the *k*th correlation subset. The serial numbers of the pair of Hgs belonging to the *k*th correlation subset are also stored by the algorithm. Three examples of similar inverse rank vectors are illustrated in Fig. S2a (ESM_3). The horizontal axis contains the serial numbers of our 50 populations in an ad hoc order, while the inverse rank values are represented by the corresponding column heights. (The order of the populations has no significance in the calculations.) Populations eliminated by the iterative process have zero inverse rank values.

Due to the operation of the iterative rank correlation algorithm, typically 15–20 vector elements dominate in the 50-dimensional IRV $\underline{R}(k)$, whereas the remaining elements are negligible or zero. The differences of these small components are also small, reducing the Euclidean distances of the vectors, so they damp the essential differences in the calculation. This problem was solved by a more advanced version of the SOC with a weighted Euclidean distance

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

measure highlighting the important components of each inverse rank vector (Juhász et al. 2016). These 50-dimensional weight vectors were also learned automatically during the training process. The mathematical description of the algorithm is given in ESM_3.

The complete analysis based on our self-learning computer programs accomplishing iterative rank correlation search for strongly correlating pairs of Hgs and SOC-clustering of the resulting IRVs can be summarized in two steps:

1. Collecting all Hg pairs having strong rank correlation values for a significant set of populations, using the iterative rank correlation method. The algorithm results in two IRVs for all pairs of these correlating Hg-pairs after finishing the iteration. For instance, the leftmost black columns in Fig. S2a (ESM_3) show the inverse rank values of mitochondrial Hg M strongly correlating with Y-chromosomal Hg O. The strong correlation between these Hgs is verified by the visible similarity of the corresponding IRV of Y-chromosomal Hg O, represented by the neighbouring grey columns.
2. The resulting vector set is used as the training set of the SOC algorithm determining the condensation centres of the corresponding 50-dimensional point system. The resulting "inverse rank vector type" (IRVT) vectors are used for clustering the whole IRV set, and the mitochondrial and Y-chromosomal Hgs belonging to IRVs assigned to a common cluster are collected into a set called the "Hg-association" of the cluster.

A more detailed mathematical description of the iterative rank correlation and self organizing cloud (SOC) algorithms is found in ESM_3 and (Juhász et al. 2016).

The goodness of the result was characterized by a calculation based on the correlation of the distance and inference matrices: after the clustering process, the inference matrix (containing values 1 when a pair of training vectors belongs to identical cluster and 0 if not) is determined. The goodness of the clustering is characterized by the correlation coefficient of the lower triangles (without the diagonal elements) of the symmetric inference matrix and the distance matrix of the training vectors.

It proved to be very favourable to put the results into historical context by extending the analysis by ancient data. To do this, we completed our 50 recent mitochondrial data with 34 ancient mtDNA distributions, and accomplished the whole analysis with the resulting 84-dimensional inverse rank vectors. To compare the recent part of the resulting 84-dimensional IRVTs to the original 50-dimensional recent ones, we had to eliminate ancient components and re-calculate the rank values of the remaining recent populations within the resulting 50-dimensional modified IRVTs.

Results

In the first step of the study, we accomplished the iterative rank correlation for all pairs of the 26 + 18 = 44 mtDNA and Y Hgs (including self-correlations). Subtracting the correlation coefficients from unity we obtain distance-like values approaching 0 and 2 in cases of strong positive as well as negative correlation, while this "distance" approaches 1 for uncorrelated Hg-pairs. Thus, the relationships of Hgs, determined by their correlations can be visualized by MDS maps, as it is shown in Fig. 1a, b for Y-chromosomal and mitochondrial Hgs.

High correlations refer to systematic joint propagation of pairs of Hgs within a significant subset of populations. Therefore, we selected all pairs of IRVs of the pairs of strongly correlating Hgs into an IRV set, with the constraint that the iteration resulted in a correlation exceeding 0.8 for a subset of populations exceeding the size of 15. We illustrate the results in Fig. S2 (ESM_3).

Finally, we obtained a set of selected IRVs containing 393 elements, and we trained the self-learning cloud (SOC) algorithm to determine all the characteristic local condensation centres within the corresponding 50-dimensional point system simultaneously. SOC learning resulted in an IRVT set of 10 elements, and the t-probe showed that the distances of the closest neighbouring IRVTs is significant with probability at least 95%. The correlation of the inference and distance matrices was −0.52.

Ordering the 393 IRVs to the most similar IRVT, we obtained ten clusters. The subset of Hgs whose IRVs belong to a common cluster build the "Hg association" propagating within the populations having nonzero inverse rank values in the IRVs of the cluster.

The geographical distribution of the IRVT discussed here is shown in the map of Eurasia in Fig. 2. (As the SOC algorithm ordered the serial number of 10 to this IRVT, we sign it as IRVT-10.) The heights of the columns show the inverse ranks of the corresponding populations, so the map shows a propagation from Eastern- and Inner Asia (CHN, HUI, JPN, MNG, KYG, KAZ) to the native Americans (NAW, NAC, NAS) and to the Volga region in Eastern Europe (TAT, CHU, UDM). The Y-chromosomal and mitochondrial members of this Hg association are shown in maps mirroring the correlation conditions of the Hgs in the right upper part of the figure. The columns ordered to the Hgs are proportional to the number of other Hgs strongly correlating with it. For example, Y-chromosomal O and mitochondrial A haplogroups have the most correlating partners within the Hg association (Y: **O**, C, Q, N; Mt: Z, B, F, G, M, C, **D, A, N***). As this Hg association was derived from the whole correlation subset belonging to IRVT-10, certain Hgs may be absent from different

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

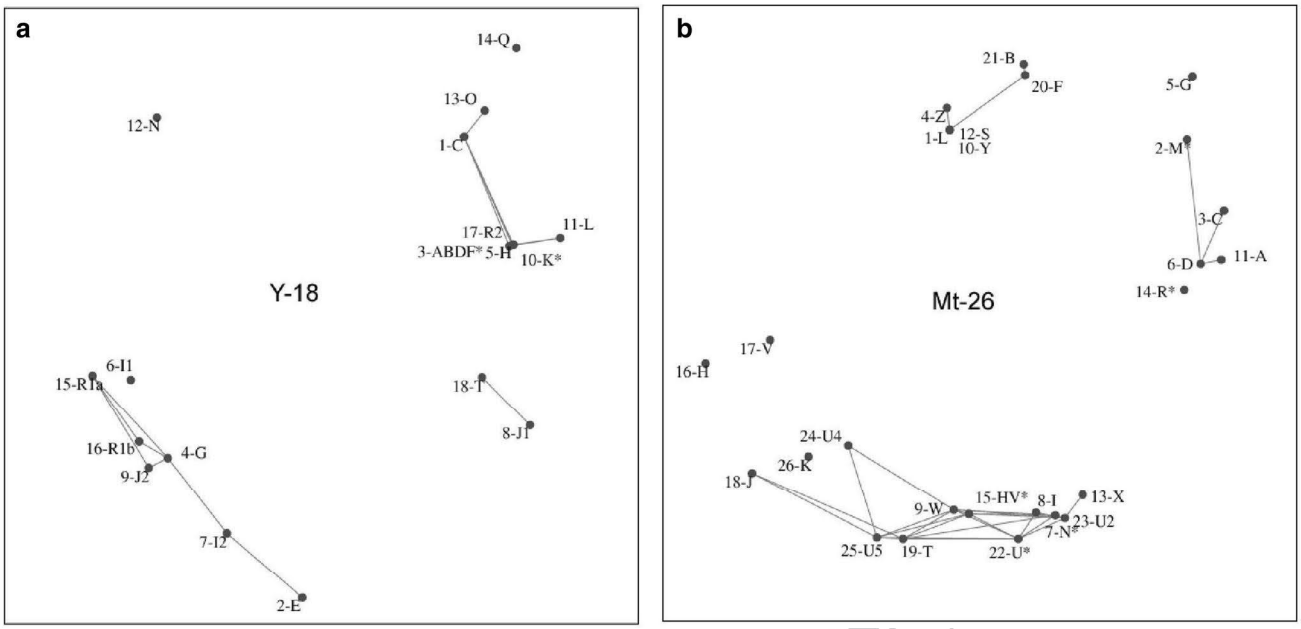

Fig. 1 MDS map of Y-chromosomal (**a**) and mitochondrial Hgs (**b**), determined from rank correlation data

individual populations. For example, Central Amerindian sample NAC contains only 3 Hgs of the Hg association (A, B and a low rate of D). This is the reason why NAC has a lower IR value than the neighbouring populations NAW and NAS in the geographical map of Fig. 2.

To understand the historical background of the propagation of the above Hg association, we accomplished the whole analysis on our mitochondrial database completed with 34 ancient mtDNA distributions. As this extended "**historical**" database contained 84 populations in sum, the IRVs in this analysis had 84 dimensions. We found that an appropriate significance of the clustering was reached for 44 IRVTs. This high number may be caused partly by the increased dimension, partly by the small sample sizes of the ancient mtDNA distributions resulting in a high noise level. To compare the recent parts of the resulting 84-dimensional and the originally recent 50-dimensional IRVTs, we eliminated the ancient components of the 84-dimensional IRVTs and re-calculated the modified inverse rank values of the remaining 50 recent components. After this, we selected the modified IRVTs having the less Euclidean distances from the original 50-dimensional IRVT-10. Finally, we turned back to the 84-dimensional original versions of the selected modified IRVTs and found that the highest inverse ranks of the complete versions are systematically attributed to the ancient samples arising from South Siberia, Inner Asia and China (ABA, Altai Bronze Age 2700–900 BC), SER (Serovo, East Siberia 8000–4000 BC), PAZ (Pazyryk, Scytho-Siberian, Mongolia, Russia 400–200 BC), XIA (Tarim Basin Xiaohe, China, 2515–1829 BC), XIO (Egyin Gol Xiongnu, Inner Asia, 200 BC–200 AD) and QIN (Qin China East Asia, 221 BC–210 AD). The historical migration transferring this Hg association to Eastern Europe is also verified by the not negligible inverse ranks in Iron-age Scythian (SCI) and ninth century Hungarian samples (KAR, AH1). In addition, the mitochondrial Hg content of the historical Hg associations (Mt: B, F, G, M, C, D, A, N*) is in a very good accordance with those of IRVT-10 (the recent IRVTs with their most similar historical pairs are shown in the maps of ESM_4 also showing the mitochondrial Hg content of the corresponding Hg associations. See Figs S1 and S2 in ESM_4). The accumulated rate of the Hg association in ancient Eastern samples (XIO, XIA, QIN) is in the range of 75–95%. The most Western appearance of this Hg association was detected in early Hungarian samples arising from 800 to 1000 AD, where it takes 15–27% of the whole distributions. These results verify that recent distribution of IRVT-10 and the corresponding Hg association is a consequence of the migrations of Scythians, Huns, Avars, Hungarian conquerors, Cumanians and other nomadic people on the Steppe.

A totally different geographical distribution characterizes the Hg association belonging to IRVT-1. This distribution represented in Fig. 3 shows a propagation from the Middle East and Asia Minor (KUR, AZE, TUR) to the Balkans and Central Europe. The Hg association also appears in Eastern Europe (TAT, CHU, RUS) and Inner Asia (UYG, KYG). The Y-chromosomal and mitochondrial correlation-based maps of the corresponding Hg association show the Hg-content (Y: **E, G, J1, J2**, L, R1b, T, Mt: N*, I, X, **HV***,**U***,**U2, K**, T, J). Comparing these maps to those shown in Fig. 2 shows

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

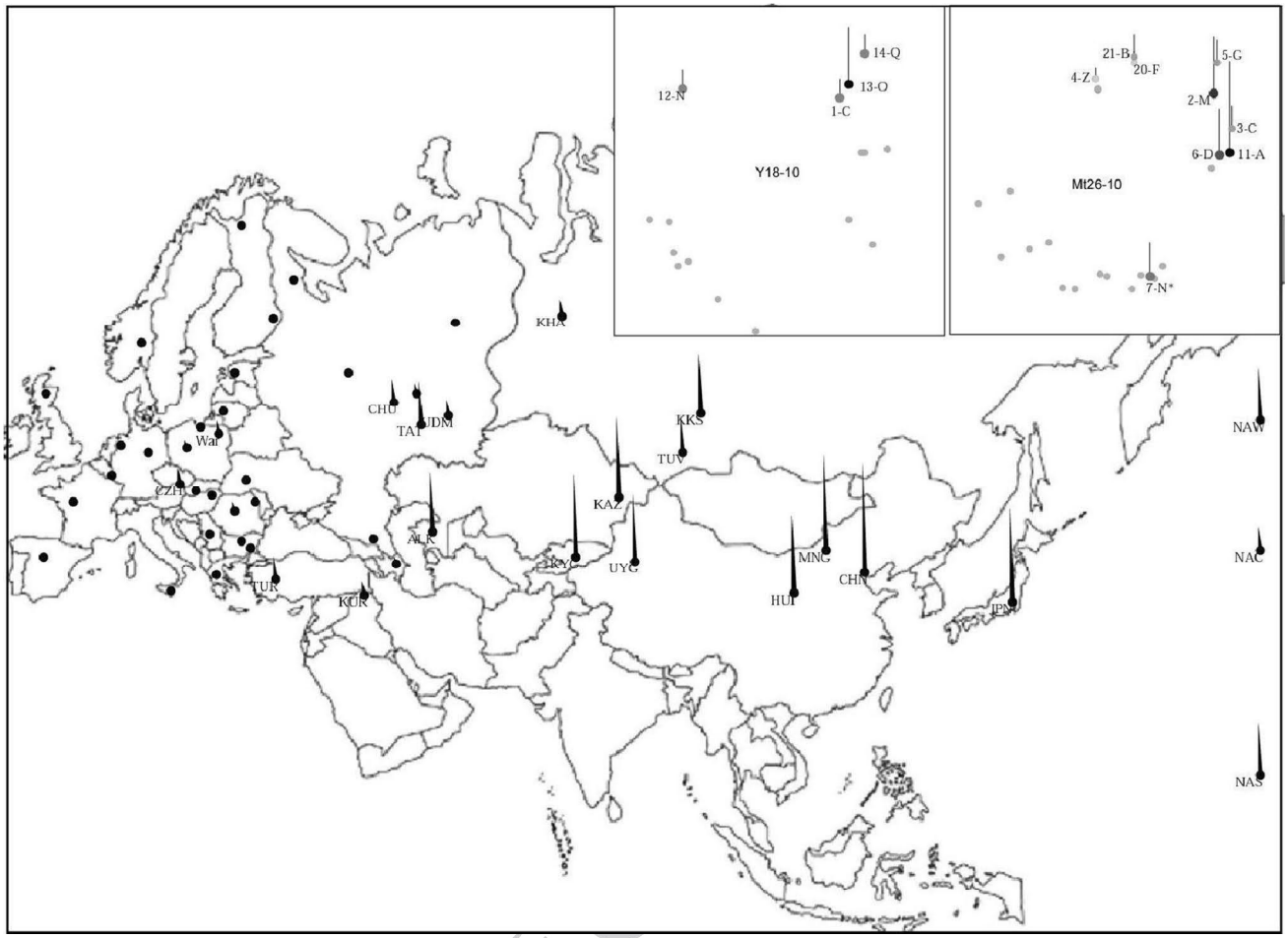

Fig. 2 Geographical distribution of the inverse ranks of IRVT-10. The Hg association propagating from East- and Inner Asia to Eastern Europe (CHU, TAT, UDM) and to the Native Americans (NAS, NAC, NAW) is shown in the right upper part. *Y18* Y-chromosomal, *Mt26* mitochondrial. The symbols referring to the Hgs contain a serial number and the name of the Hg (e.g. 12-N in the map "Y18-10" means Y-chromosomal Hg N having the serial number 12. 4-Z in the map "Mt26-10" means mitochondrial Hg Z having the serial number 4. The serial numbers were defined arbitrarily for the computation.) The heights of the grey columns in the Hg maps are proportional to the number of Hgs having strong correlation with the Hg belonging to the column. The black peaks in the geographical map are proportional to the IR value of the corresponding population in the given IRVT

that the Hg association of IRVT-10 and IRVT-1 have practically no overlap.

The geographical distribution of IRVT-1 makes one think that this recent distribution may originate in the ancient migration of Neolithic farmers from the Crescent Fertile to Europe. To verify this suspicion, we accomplished the whole analysis on our extended mitochondrial database completed with 34 ancient distributions. We found that the highest inverse ranks of the resulting historical IRVT versions are attributed exactly to Neolithic populations in the Fertile Crescent and Central Europe NEO (Near Eastern Neolithic, TR, IRN, SYR, JOR, 8300–4000 BC), MEN (Middle East Neolithic-BrA, Middle East, 11,840–1402 BC), STR (Starcevo, Balkans, 5700–5500), NHU (Neolithic Hungary, 5200–4800 BC), EMN (Early-Middle Neolithic, Europe, 6000–3000), while recent populations KUR, and TUR have also high inverse ranks. In addition, the mitochondrial Hg content of the historical Hg associations (Mt: N*, X, HV*, T, U*, U2, K) is in a very good accordance with those of IRVT-1 (see Figs S3 and S4 in ESM_4). The accumulated rate of the recent Hg association in ancient Neolithic samples MEN, NEO, NHU, STR is in the range of 68%-80%. These results verify that recent distribution of IRVT-1 and the corresponding Hg association is a consequence of the migration of Neolithic farmers containing essentially the same mitochondrial Hgs in the past as the recent populations living in the areas of the ancient migration.

According to Fig. 4, the geographical distribution of our next IRVT shows a Western European origin propagating to Central and Eastern Europe (IRVT-3). The corresponding Hg association (Y: **I1, R1b**, G, I2; Mt: **H, V, J, K, U5, T**) has some common elements with IRVT-1 (Y: R1b, G;

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

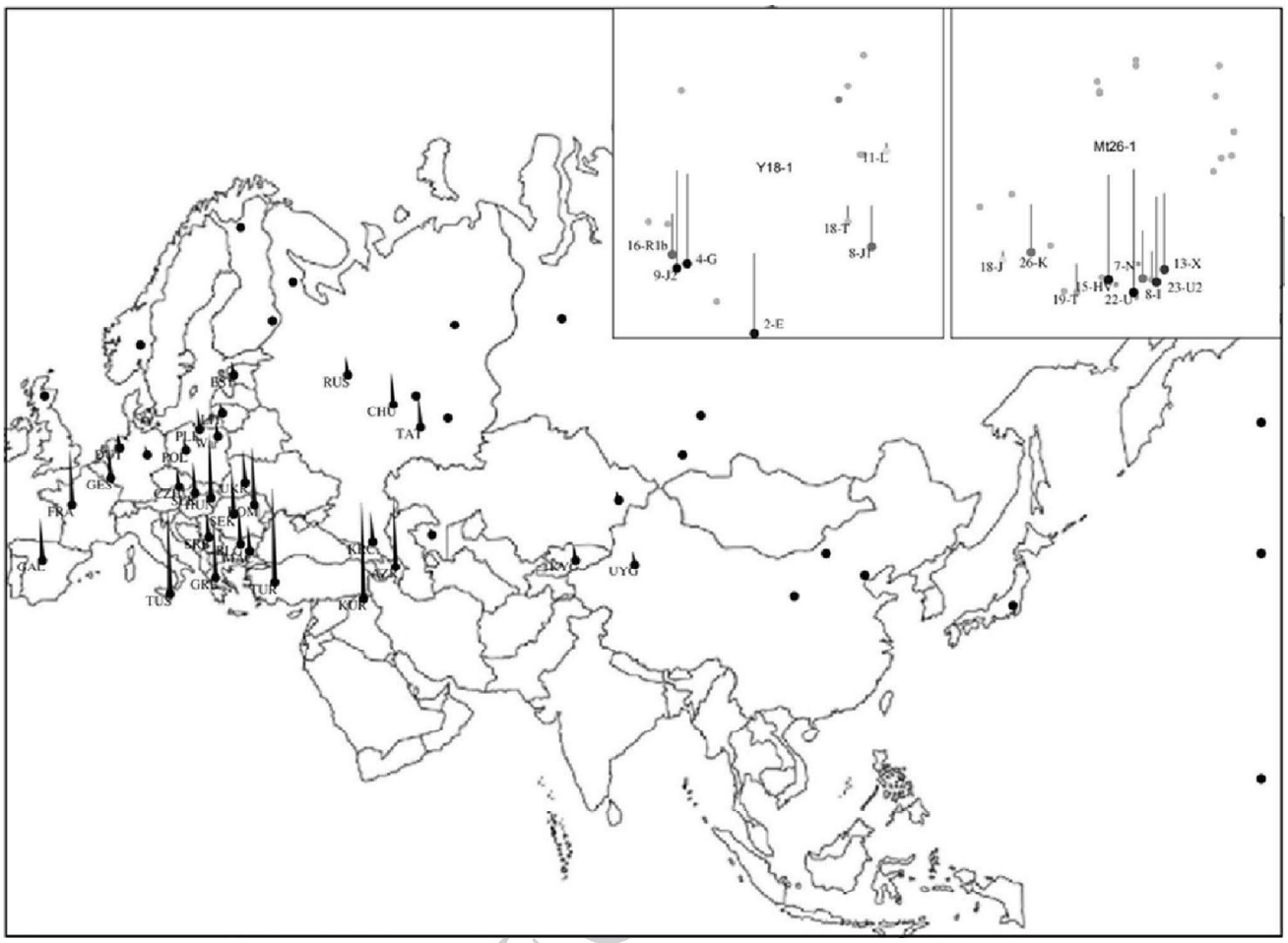

Fig. 3 Geographical distribution of the inverse ranks of IRVT-1. The Hg association propagating from the Middle East to the Balkans and Central Europe is shown in the right upper part. For description of the symbols, see Fig. 2

Mt: K), but Y: I1 and the total set of the mitochondrial Hgs except K are not found there. On the other hand, the set (Y: J2, J1, T, L, E; Mt: HV*, U*, U2, N*, X, I) in the Hg association of IRVT-1 is not found here. The comparison of the IRVTs of the extended ancient database indicated the highest inverse ranks for the Neolithic European samples STR (Starcevo, Balkans, 5700–5500) and EMN (Early-Middle Neolithic, Europe, 6000–3000), but NHU (Neolithic Hungary, 5200–4800 BC) and MEN (Middle East Neolithic, Middle East, 11,840–1402 BC) have also significant inverse ranks (see Figs S5 and S6 in ESM_4). Other ancient samples have zero inverse ranks for this Hg association. The accumulated rate of the recent Hg association in Neolithic European samples STR and EMN is 75%. Consequently, IRVT-3 may refer to the ancient European population preceding the migration of the farmers from the Fertile Crescent. This is supported by the fact that all the mitochondrial components of this Hg association are found in the Hg distribution of EMN, taking more than 75% in sum of the whole sample.

The highest inverse rank values indicate Central and Eastern Europe as the source area of the Hg association (Y: **R1a**, I1, **I2, J2**, E; Mt: **H, J, K**, W, **U***, **N***, **I**, X) belonging to IRVT-2 in Fig. 5. The mitochondrial components of this Hg association can be divided into 2 well-defined groups—Hgs H and J are connected to the Western European association of IRVT-3, whereas U*, N* I and X are common with IRVT-1 originating from the Fertile Crescent. The only common Hg between the two parts is Hg K. Y-chromosomal Hgs I1 and I2 are among the most important components of the Hg association of IRVT-3, whereas J2 and E are of high importance in the Hg association of IRVT-1. This suggests that this Hg association may be traced back to an admixture of ancient Europeans and farmers arising from the Fertile Crescent. The highest inverse rank values in the most similar historical IRVTs are assigned to neolithic samples MEN and NEO in the Fertile Crescent, as well as European IBN (Iberian Neolithic, Iberia, 10,310–3160 BC) and NHU (Neolithic Hungary, 5200–4800 BC). Six members of the relating historical (ancient) Hg distribution (Mt: **J, K**, U4, T, **U***, **I,**

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

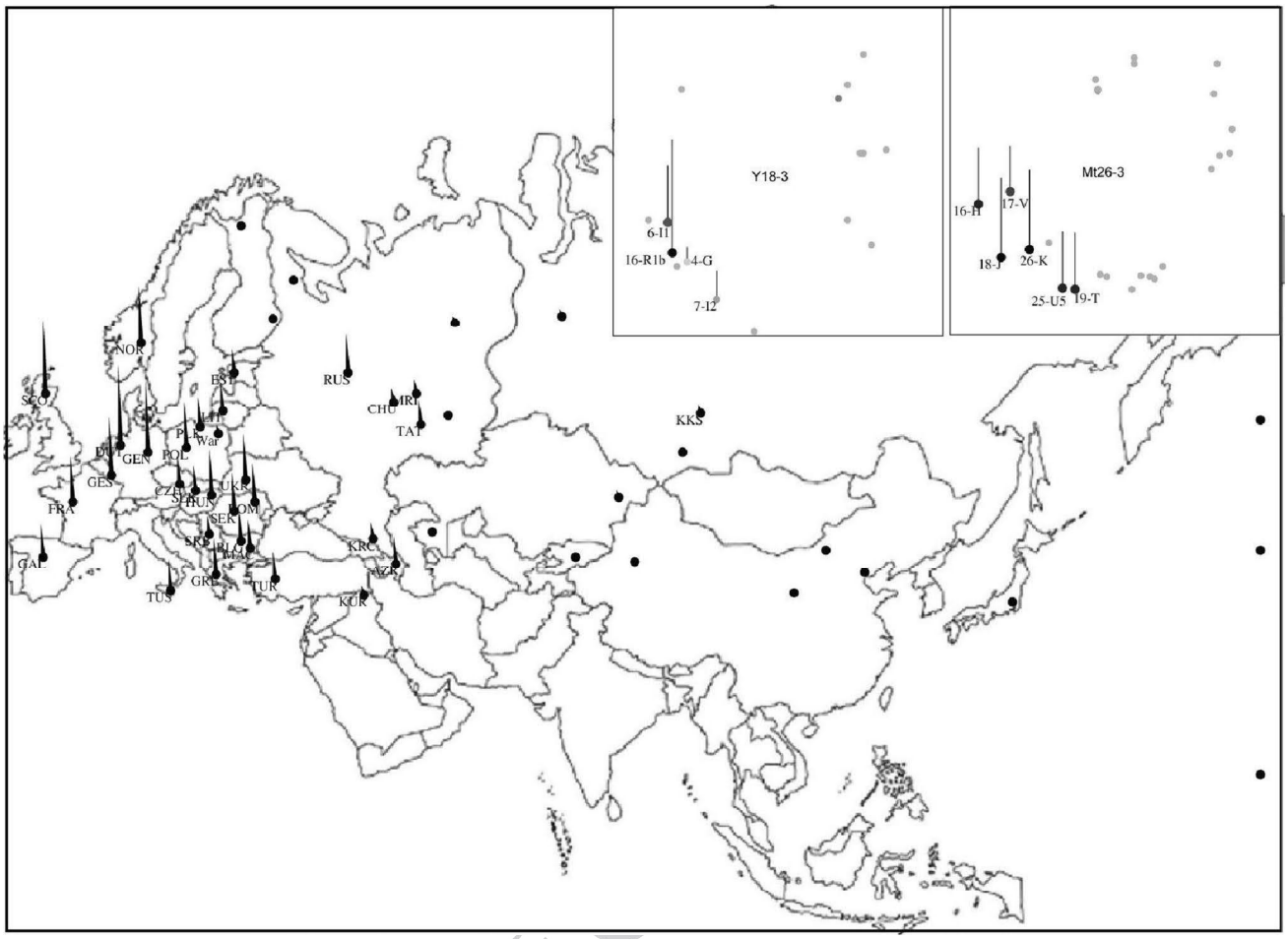

Fig. 4 Geographical distribution of the inverse ranks of IRVT-3. The corresponding Hg association is shown in the right upper part. For description of the symbols, see Fig. 2

N*, **X**) are common with those of IRVT-2 (see Figs S7 and S8 in ESM_4). The accumulated rates of the recent Hg association in the Neolithic samples are in the range of 68–82%. These results may really be explained by an admixture of early Europeans and farmers in the Balkans, the Carpathian Basin and Eastern Europe. The relatively high importance of the resulting Hg association in the Volga region (TAT, CHU), the ninth and tenth century Hungarian samples (AH2, Hungarians 900–1000 AD; AH1, Ancient Hungarian, 900–1000 AD) as well as HPC (pre-Conquest Hungary, 500–900 AD) may need a further explanation.

Figure 6 shows the geographic distribution of IRVT-4 having the largest inverse rank values in North-Eastern Europe. The corresponding Hg association is composed by (Y: **N, I1, R1a**, R1b, I2, E; Mt: **H, V, J, U4, U5**, T, W, U*, **I**, N*, U2, X). The overlaps between this Hg association with those of IRVT-3 and IRVT-2 are (Y: **I1, R1b**; Mt: **H, V, J, U5**, T) and (**I1, R1a**, I2; Mt: **H, J**, U*, N*, X). The Hgs being found in both overlaps (Y: **I1**; Mt: **H, J**) are more important components of IRVT-3 than IRVT-2 (see the column heights in the Hg-maps in Figs. 4, 5). The overlap with the Hg association of IRVT-1 (Y: R1b, E; Mt: T, U*, I, N*, U2) contains no important components in the Hg association of IRVT-4 (see the Hg-maps in Fig. 6).

The most similar historical IRVTs show high inverse rank values for Neolithic Western European sample LNB (Late Neolithic, Europe, 3000–1600 BC), as well as Copper-age Eastern European YAM (Yamnaya, Afanasievo, Russia, Ukraine, 5000–2700 BC) (see Figs S9-12 in ESM_4). Early AQ1 medieval Viking (Norway, 780–790 AD) sample has also high inverse rank of this Hg association. The complete set of the components of the historical IRVTs (Mt: **H**, J, U4, U5, T, W, U*, I, X) is very similar to that of IRVT-4. The accumulated rates of the Hg association in LNB, YAM and VIK are 81, 88, and 74%. These results imply an admixture of Neolithic European hunters as well as a population composed by Neolithic hunters and farmers arising from Eastern Europe. The geographical distribution of the inverse ranks shows that the resulting complex Hg association has the highest weight in Eastern and Northern Europe.

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

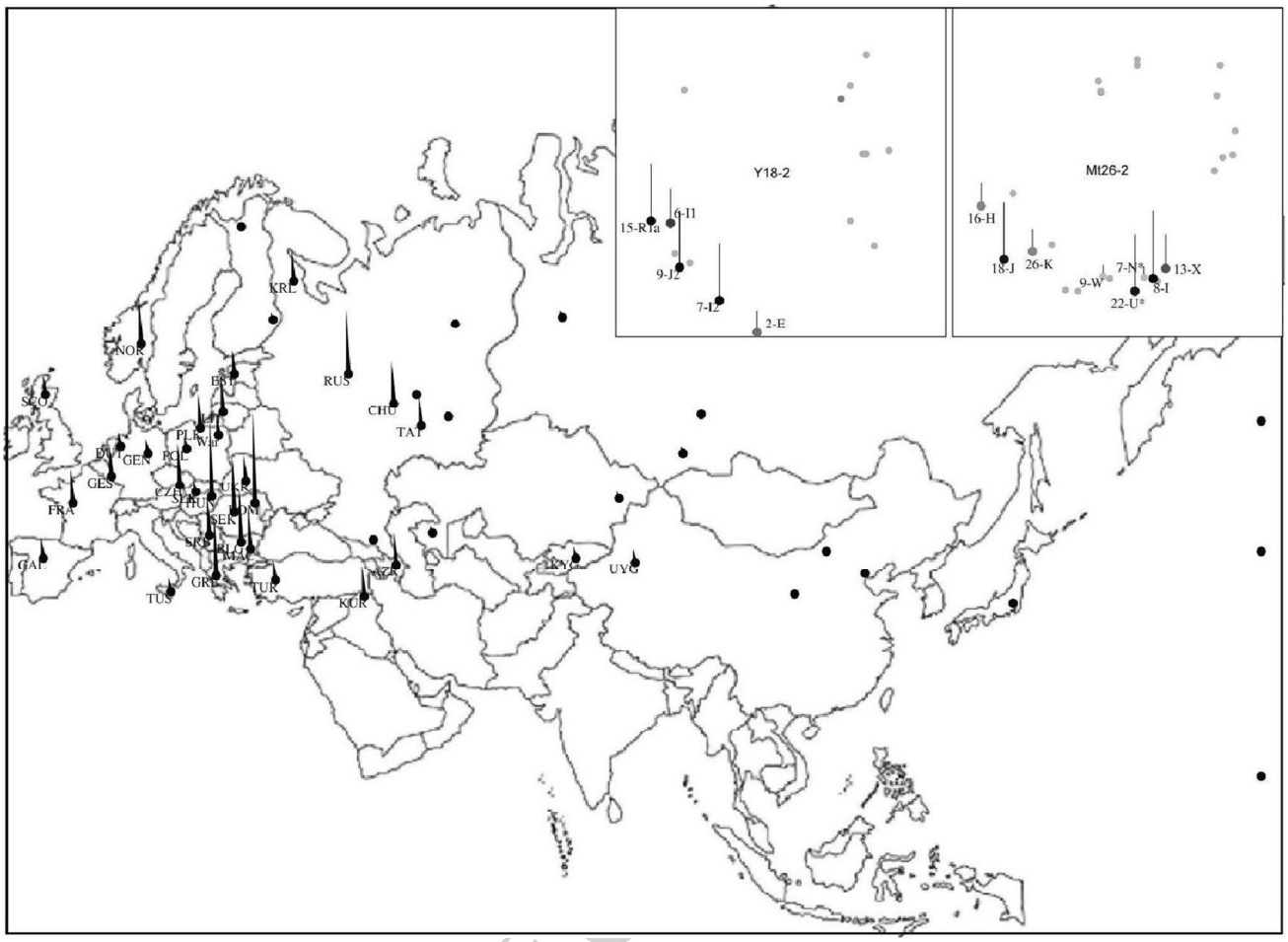

Fig. 5 Geographical distribution of the inverse ranks of IRVT-2. The corresponding Hg association is shown in the right upper part. For description of the symbols, see Fig. 2

The highest inverse rank values of our last example IRVT-5 show the Carpathian Basin and the Balkans as source area of the Hg association (Y: R1a, R1b, J2, **G, I2,** E; Mt: **H, U4**, J, K, T, **W, HV***, U*) (see Fig. 7). The largest overlaps (Y: R1a, R1b, J2, **I2**, E; Mt: **H**, J, K) and (Y: R1a, R1b, **I2**, E; Mt: **H, U4**, J, T, **W**) connect this Hg association to IRVT-2 and IRVT-4. The most important Y-chromosomal and mitochondrial Hgs G and I2 as well as H, U4 and W are of similar importance in IRVT-2 an IRVT-4, while HV* has similar importance in IRVT-1.

For the first sight, these results imply again an admixture of Neolithic farmers and European hunters in Central- and Eastern Europe, like IRVT-2. However, the most similar historical IRVTs show here the highest inverse rank values for the West Asian BB3 and KBK, as well as early Medieval Hungarian HPC (See Figs S13-S16 in ESM_4). The accumulated rates of the recent Hg association are 85, 77, and 78% in these ancient distributions, respectively. These results imply a more complex interpretation: first, the admixture of Neolithic European and Near-Eastern populations, detected in IRVT-2, migrated from Eastern Europe to Western Asia. (See the historical origin of the Andronovo culture.) This may be the reason of the high cumulated rates of the recent Hg association within Bronze-age samples BB3 (Baraba, Iron transition), West Siberia (1000–800 BC) and KBK (Bronze Age Kurgans, Kazakhstan, 1400–1000 BC) representing the populations of the late Andronovo culture and Western Asia. Later, the resulting West Asian people—Scythians, Sarmatians, Huns, Avars, Hungarians, Cumanians, etc.—invaded Eastern Europe and the Carpathian Basin. This is the reason, why early medieval Hungarian samples HPC and AH2 also fit into this IRVT with significant inverse weights.

We have found that the clusters belonging to the remaining four IRVTs are very small and the corresponding Hg sets contain only a few elements. This shows that these IRVTs do not represent realistic Hg associations, but they proved to be useful to separate outlier IRVs from the realistic clusters. The relationships between the 10 IRVTs are shown in the map constructed by the SOC algorithm in ESM_3 (Fig. S3).

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

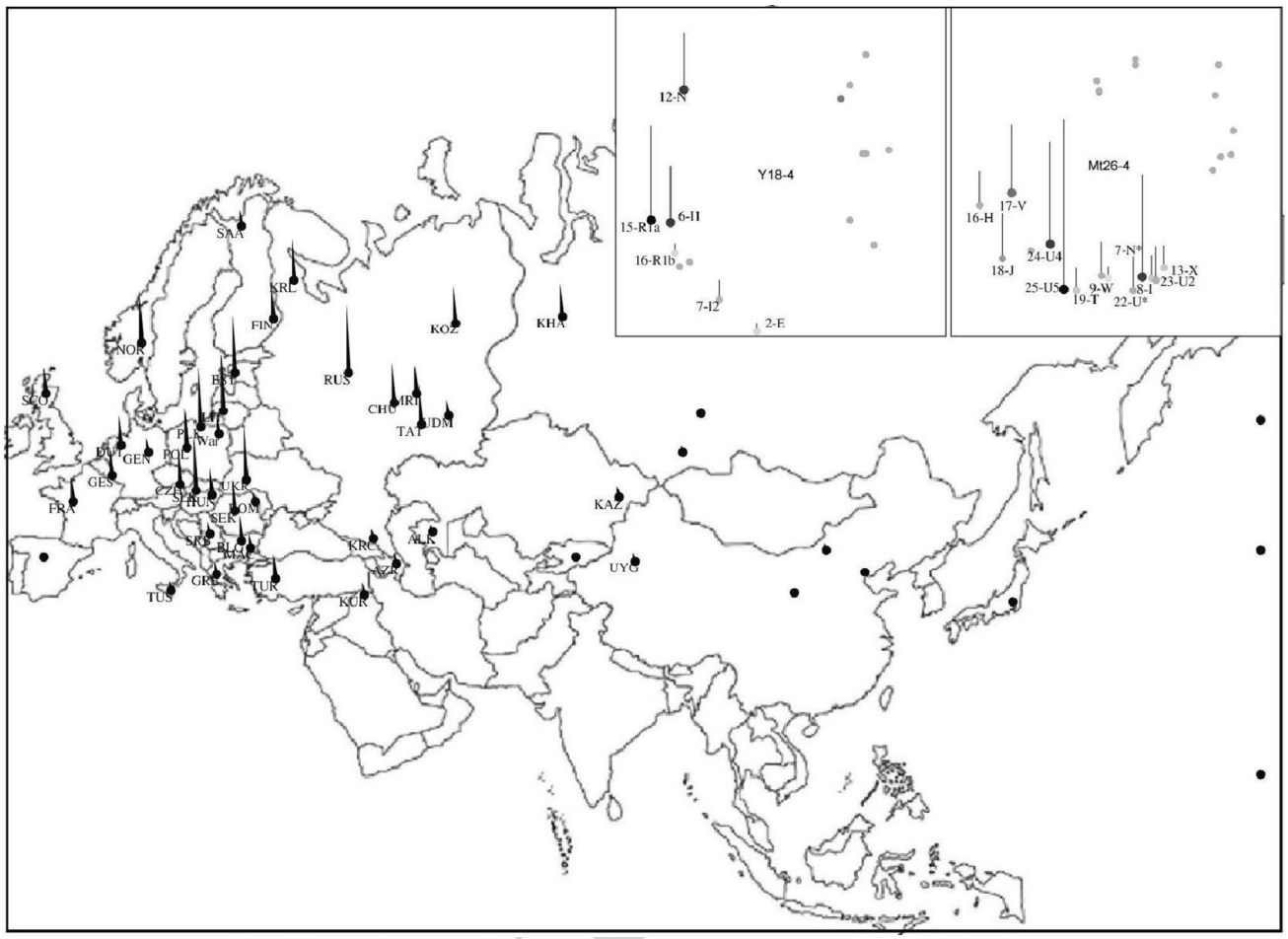

Fig. 6 Geographical distribution of the inverse ranks of IRVT-4. The corresponding Hg association is shown in the right upper part. For description of the symbols, see Fig. 2

Validation

The standard of the goodness of the inverse rank correlation method is the resulting correlation value itself obtained after finishing the iteration. The highest and lowest values 1 and −1 indicate totally correlating as well as totally anti-correlating pairs of Hgs, while total un-correlation is indicated by a correlation value of 0. Therefore, our constraints selecting pairs of Hgs showing a correlation higher than 0.8 for a correlation subset counting at least 15 populations clearly define the requirements regarding the goodness of the iterating rank correlation algorithm at the same time.

The goodness of the SOC-clustering of the IRV set fulfilling the above-mentioned constraints was measured by the correlation coefficients of the distance- and inference matrices of the IRV sets.

To validate the method with independent data, we accomplished the whole process with separated mitochondrial and Y-chromosomal Hg distributions. The results are found in ESM_3.

Discussion

Our first main contribution was to describe a new method based on a self-learning algorithm searching for systematically jointly propagating sets of Hgs in a significant subset of populations. The basic idea of the method is that the inverse rank vectors of jointly propagating Hgs are necessarily similar, so the complete set of the IRVs belonging to Hgs having at least one strongly correlated pair construct a clustered point system in their 50-dimensional vector space. The local condensation centres of these local condensations (IRVTs) were determined as the learning vectors of the self-learning SOC algorithm trained by the complete set of IRVs belonging to strongly correlating Hgs. In addition, clustering the training IRV set using these IRVT vectors determines the "Hg associations" as the corresponding subsets of Hgs. Thus, this method provides us the associations of jointly propagating Hgs and the paths of their propagations simultaneously and immediately.

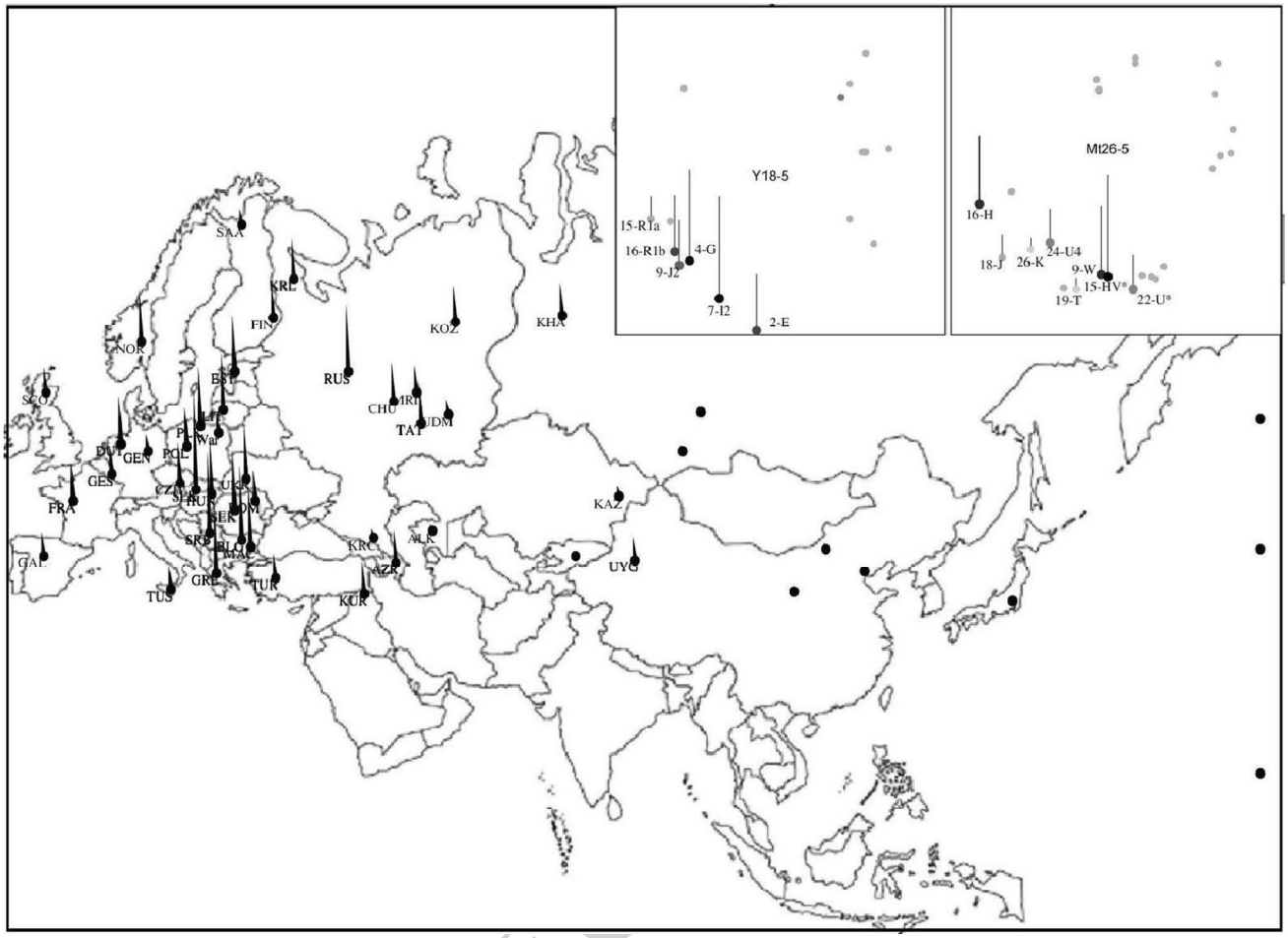

Fig. 7 Geographical distribution of the inverse ranks of IRVT-5. The corresponding Hg association is shown in the right upper part. For description of the symbols, see Fig. 2

It has been shown in previous works that a high frequency of a Hg does not necessarily indicate its source population, because bottleneck and founder effects may cause drastic changes in Hg frequencies (Cinnioglu et al. 2004; Bíró et al. 2009). However, these cases result in the loss of the correlation with other Hgs, so our iterative rank correlation method automatically eliminates them from the correlation subset of populations, while the remaining subset still indicates the real correlation.

Till now, Y-chromosomal and mitochondrial Hgs were studied usually separately in human population genetics. The novelty of our method lies in the possibility of studying jointly propagating associations of mitochondrial and Y-chromosomal Hgs. Moreover, joint propagation of genetic and/or cultural (e.g. linguistic or musical) characteristics could also be studied using IRV clustering.

The consideration that migrating human populations necessarily contain male and female components suggests the idea of studying all correlations including both mitochondrial and Y-chromosomal Hgs. Our first example (IRVT-10, Fig. 2) shows the geographical distribution of an IRVT having the highest values in Eastern and Inner Asia as well as American indigenous people, and shows a gradual reduction in Western direction. This clear correlation of IRVT-10 with the geographical conditions is itself an independent evidence of the goodness of our method, since the geographical conditions are totally ignored in the analysis. The Hg association derived from IRVT-10 clearly contains the set of male and female Hgs of well-known Eastern Asian origin (Yao et al. 2004; Derenko et al. 2007a, b, c; Forster et al. 1996; Kim et al. 2011; Zegura et al. 2004). As the possible areas of origin of the Hgs are also totally ignored from the analysis, this result is a further independent evidence supporting our method. In a good accordance with these results obtained from recent data, we also found high inverse ranks for the same mitochondrial Hg association in ancient Inner Asian populations.

Similarly, good accordance between geographical distributions of inverse ranks and places of origin of the corresponding Hg association was experienced for IRVT1 (Fig. 3)

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

showing the Fertile Crescent as source area of the corresponding Hg association propagating to Europe through Asia Minor, the Balkans and the Carpathian Basin. As the mitochondrial part of this Hg association was also detected with high inverse ranks in ancient populations in the Fertile Crescent, IRVT1 can be attributed to the well-known migration of Neolithic farmers starting from the Middle East (Bramanti et al. 2009; Malmström et al. 2009; Skoglund et al. 2012).

Also, very clear correlation between the geographical distribution and the place of origin of the corresponding Hg association was found for IRVT-3 (Fig. 4) playing the most important role in recent Western people. As the highest inverse ranks of the same mitochondrial Hg association are found in ancient European populations, IRVT-3 can be considered as the Hg association of Neolithic indigenous Europeans (Gamba et al. 2014; Szécsényi-Nagy 2015; Kivisild 2017; Wong et al. 2017).

The three Hg associations discussed above can be considered as "pure" descendants of early populations preceding the later admixture processes generated by the migrations in the Neolithic period, the Bronze- and Iron ages, the late Antiquity as well as early Middle Age.

The Hg associations derived from the remaining 3 IRVTs clearly mirror the admixture of the Hg associations of IRVT-3 and IRVT-1 representing indigenous Europeans and Neolithic Farmers. The geographical distribution of IRVT-2 (Fig. 5) shows that the most probable stages of this admixture were the Balkans and the Carpathian Basin. The geographic distribution also implies the propagation of the resulting population to Eastern Europe, in good accordance with earlier studies of the Yamnaya culture (Anthony 2007; Kristiansen and Larsson 2005; Kristiansen 2007; Wong et al. 2017). The high inverse ranks of the mitochondrial part of this Hg association in early Medieval Hungarian samples also support the Eastern European origin of this population.

The similarity of the geographical distribution and Hg association of IRVT-5 to IRVT-2 intimates that the population attributed above to IRVT-2 may play an important role in IRVT-5, too. However, the most similar IRVTs of ancient mitochondrial Hgs indicate the presence of the Hg association derived from IRVT-5 (Fig. 7) in Bronze-Age Western Asian samples, too. The explanation of this may be the known migration of Bronze-Age Eastern Europeans to Western Asia, and a further admixture with populations arising from the Middle East (Antony 2007; Hanks et al. 2007) (note that the Hg associations of both IRVT-2 and IRVT-5 contain numerous Hgs arising from the Middle East, but these sets are not identical). The resulting population is attributed to the Andronovo culture (Keyser et al. 2009; Allentoft et al. 2015). It is also supported by archaeological results that the descendants of the Andronovo culture were found in the Eurasian Steppe after 1700–1500 BC, so the high rate of the Hg distribution of IRVT-5 in early medieval Hungarians can be traced back to the expansions of Scythian, Sarmatian, Hun, Avar, etc., empires all of them reaching the Carpathian Basin, as well as the Hungarian conquest (Neparáczki et al. 2017; Cynarski and Maciejewska 2016; Szécsényi-Nagy 2015; Gamba et al. 2014; Korjakova and Epimakhov 2014).

The origin of the Hg association derived from IRVT-4 can also be traced back to the population attributed to IRVT-2 on the one hand, and early indigenous Europeans attributed to IRVT-3 on the other hand. The clear geographical distribution of IRVT-4 (Fig. 6) shows the propagation of the resulting complex Hg distribution of IRVT-4 from North-Eastern Europe into Southern and Western directions. The role of the Eastern European Yamnaya culture in the evolution of the Corded Ware culture in Northern Europe has also been shown previously (Harrison and Heyd 2007; Vandkilde 2007; Wong et al. 2017; Allentoft et al. 2015).

We have shown certain ancient populations where the cumulated rate of the mitochondrial Hg associations derived from recent data is extremely high. In principle, the frequency of an Hg in a population can be summed up by more Hg associations constructing the given population, because of the overlaps of their Hg contents (Zerjal et al. 2002; Sharma et al. 2009). Therefore, these cumulated rates refer merely to a possible maximal rate of the Hg associations, and the actual rates may be lower. However, an extremely high cumulated rate of a Hg association in an ancient population may refer to a situation antedating later admixtures.

Unfortunately, we could not collect all ancient Y-chromosomal data exactly corresponding to our ancient mitochondrial distributions. However, joint propagation of contemporary male and female haplogroups is itself a strong validation of past human population migrations. In addition, the validation of our method is also based on simultaneous search for both mitochondrial and Y-chromosomal IRVTs.

These considerations clearly show the importance of ancient Hg distributions in credible interpretation of the results. The mathematically correct estimation of the rates of the Hg associations in an actual Hg distribution and the completion of our ancient mitochondrial data by their Y-chromosomal counterparts would result in a much clearer insight into the early migration processes.

First and last, the above discussions illustrate that our method based on the clustering of the inverse rank vectors of Hgs provides a good insight into the most effective migration processes and the prehistory of the mankind. The accordance with prior knowledge regarding genetic and archaeological footprints of Neolithic and Bronze-age migrations validates our method in itself (Allentoft et al. 2015). The results also improve that correlations of jointly propagating Hgs in contemporary populations can be traced back to prehistoric migration processes. In addition, the method

Journal : Large 438 | Article No : 1469 | Pages : 16 | MS Code : MGAG-D-18-00097 | Dispatch : 2-7-2018

could be extended to study the correlations of cultural and genetic characteristics, to validate linguistic, archaeological, ethnomusicological, etc., theories and hypotheses by genetic evidences. Other biological applications, i.e. correlation analysis of frequencies of different species also could reveal joint propagations of different associations of plants and/or animals.

Acknowledgements The authors are grateful to Tibor Fehér for the collection of mitochondrial and Y-chromosomal Hg distributions from prior published results. No funding was received.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Research involving human participants and/or animals All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Allentoft ME, Sikora M, Sjögren K, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, Malaspinas A, Margaryan A, Higham T, Chivall D, Lynnerup N, Harvig L, Baron J, Della Casa P, Dabrowski P, Duffy PR, Ebel AV, Epimakhov A, Frei K, Furmanek M, Gralak T, Gromov A, Gronkiewicz S, Grupe G, Hajdu T, Jarysz R, Khartanovich V, Khokhlov A, Kiss V, Kolár J, Kriiska A, Lasak I, Longhi C, McGlynn G, Merkevicius A, Merkyte I, Metspalu M, Mkrtchyan R, Moiseyev V, Paja L, Pálfi GY, Pokutta D, Pospieszny L, Price D, Saag L, Sablin M, Shishlina N, Smrčka V, Soenov VI, Szeverényi V, Tóth G, Trifanova SV, Varul L, Vicze M, Yepiskoposyan L, Zhitenev V, Orlando L, Sicheritz-Ponté n T, Brunak S, Nielsen R, Kristiansen K, Eske Willerslev E (2015) Population genomics of bronze age eurasia. Nature 522(7555):167–172

Anthony D (2007) The horse, the wheel and language. How bronze-age riders from the Eurasian Steppes Shaped the Modern World. Princeton Univ. Press, Princeton

Batini C, Hallast P, Vågene AJ, Zadik D, Eriksen HA, Pamjav H, Sajantila A, Wetton JH, Jobling MA (2017) Population resequencing of European mitochondrial genomes highlights sex-bias in Bronze Age demographic expansions. Sci Rep 7(1):12086

Bermisheva MA, Kutuev IA, Korshunova TY, Dubova NA, Villems R, Khusnutdinova EK (2004) Phylogeographic analysis of mitochondrial dna in the nogays: a strong mixture of maternal lineages from Eastern and Western Eurasia. Mol Biol 38:516–523

Bíró AZ, Zalán A, Völgyi A, Pamjav H (2009) A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). Am J Phys Anthropol 139(3):305–310

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ, Lueth F, Terberger T, Hiller J, Matsumura S, Forster P, Burger J (2009) Genetic discontinuity between local hunter-gatherers and Central Europe's first farmers. Science 326:137–140

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome Haplotype Strata in Anatolia. Hum Genet 114:127–148

Cynarski WJ, Maciejewska A (2016) The proto-Slavic warrior in Europe: the scythians, sarmatians and lekhs. Ido movement for culture. J Martial Arts Anthropol 16(3):1–14

Der Sarkissian C, Allentoft ME, Ávila-Arcos MC, Barnett R, Campos PF, Cappellini E, Ermini L, Fernández R, da Fonseca R, Ginolhac A, Hansen AJ, Jónsson H, Korneliussen T, Margaryan A, Martin MD, Moreno-Mayar JV, Raghavan M, Rasmussen M, Velasco MS, Schroeder H, Schubert M, Seguin-Orlando A, Wales N, Gilbert MT, Willerslev E, Orlando L (2015) Ancient genomics. Philos Trans R Soc Lond B Biol Sci 370(1660):20130387

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee HK, Vanecek T, Villems R, Zakharov I (2007a) Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet 81(5):1025–1041

Derenko M, Malyarchuk B, Denisova G, Wozniak M, Grzybowski T, Dambueva I, Zakharov I (2007b) Y-chromosome haplogroup N dispersals from south Siberia to Europe. J Hum Genet 52(9):763–770

Derenko MV, Maliarchuk BA, Wozniak M, Denisova GA, Dambueva IK, Dorzhu CM, Grzybowski T, Zakharov IA (2007c) Distribution of the male lineages of Genghis Khan's descendants in northern Eurasian populations. Genetika 43(3):422–426

Ermini L, Der Sarkissian C, Willerslev E, Orlando L (2015) Major transitions in human evolution revisited: a tribute to ancient DNA. J Hum Evol 79:4–20

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59(4):935–945

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, Nickel B, Peltzer A, Rohland N, Slon V, Talamo S, Lazaridis I, Lipson M, Mathieson I, Schiffels S, Skoglund P, Derevianko AP, Drozdov N, Slavinsky V, Tsybankov A, Cremonesi RG, Mallegni F, Gély B, Vacca E, Morales MR, Straus LG, Neugebauer-Maresch C, Teschler-Nicola M, Constantin S, Moldovan OT, Benazzi S, Peresani M, Coppola D, Lari M, Ricci S, Ronchitelli A, Valentin F, Thevenet C, Wehrberger K, Grigorescu D, Rougier H, Crevecoeur I, Flas D, Semal P, Mannino MA, Cupillard C, Bocherens H, Conard NJ, Harvati K, Moiseyev V, Drucker DG, Svoboda J, Richards MP, Caramelli D, Pinhasi R, Kelso J, Patterson N, Krause J, Pääbo S, Reich D (2016) The genetic history of Ice Age Europe. Nature 534(7606):200–205

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kővári I, Pap I, Anders A, Whittle A, Dani J, Raczky P, Higham TFG, Hofreiter M, Bradley DG, Pinhasi R (2014) Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun 5:5257

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522(7555):207–211

Hanks BK, Epimakhov AV, Renfrew AC (2007) Towards a refined chronology for the Bronze Age of the southern Urals. Russ Antiquity 81:353–367

Harrison R, Heyd V (2007) The Transformation of Europe in the third millennium BC: the example of 'Le Petit-Chasseur I1III' (Sion, Valais, Switzerland). Praehistorische Zeitschrift 82:129–214

Ilyas M, Kim JS, Cooper J, Shin YA, Kim HM, Cho YS, Hwang S, Kim H, Moon J, Chung O, Jun J, Rastogi A, Song S, Ko J, Manica A, Rahman Z, Husnain T, Bhak J (2015) Whole genome sequencing of an ethnic Pathan (Pakhtun) from the north-west of Pakistan. BMC Genom 16:172

Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612

Juhász Z (2007) Analysis of melody roots in Hungarian folk music using self-organizing maps with adaptively weighted dynamic time warping. Appl Artif Intell 21(1):35–55

Juhász Z, Fehér T, Bárány G, Zalán A, Németh E, Pádár Z, Pamjav H (2015) New clustering methods for population comparison on paternal lineages. Mol Genet Genom 290(2):767–784

Juhász Z, Fehér T, Németh E, Pamjav H (2016) mtDNA analysis of 174 Eurasian populations using a new iterative rank correlation method. Mol Genet Genom 291(1):493–509

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838

Keyser C, Bouakaze C, Crubézy E, Nikolaev VG, Montagnon D, Reis T, Ludes B (2009) Ancient DNA provides new insights into the history of south Siberian Kurgan People. Hum Genet 126:395–410

Kim SH, Kim KC, Shin DJ, Jin HJ, Kwak KD, Han MS, Song JM, Kim W, Kim W (2011) High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. Investig Genet 2(1):10

Kivisild T (2017) The study of human Y chromosome variation through ancient DNA. Hum Genet 136(5):529–546

Korjakova L, Epimakhov AV (2014) The Urals and western Siberia in the Bronze and Iron ages. Cambridge University Press, Cambridge

Kristiansen K (2007) In the world system and the earth system. In: Hornborg B, Crumley C (eds) Global socioenvironmental change and sustainability since the neolithic. Left Coast Press, Walnut Creek

Kristiansen K, Larsson T (2005) The rise of bronze age society. travels, transmissions and transformations. Cambridge Univ. Press, Cambridge

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KI, Nordenfelt S, Li H, de Filippo C, Prüfer K, Sawyer S, Posth C, Haak W, Hallgren F, Fornander E, Rohland N, Delsate D, Francken M, Guinet JM, Wahl J, Ayodo G, Babiker HA, Bailliet G, Balanovska E, Balanovsky O, Barrantes R, Bedoya G, Ben-Ami H, Bene J, Berrada F, Bravi CM, Brisighelli F, Busby GB, Cali F, Churnosov M, Cole DE, Corach D, Damba L, van Driem G, Dryomov S, Dugoujon JM, Fedorova SA, Gallego Romero I, Gubina M, Hammer M, Henn BM, Hervig T, Hodoglugil U, Jha AR, Karachanak-Yankova S, Khusainova R, Khusnutdinova E, Kittles R, Kivisild T, Klitz W, Kučinskas V, Kushniarevich A, Laredj L, Litvinov S, Loukidis T, Mahley RW, Melegh B, Metspalu E, Molina J, Mountain J, Näkkäläjärvi K, Nesheva D, Nyambo T, Osipova L, Parik J, Platonov F, Posukh O, Romano V, Rothhammer F, Rudan I, Ruizbakiev R, Sahakyan H, Sajantila A, Salas A, Starikovskaya EB, Tarekegn A, Toncheva D, Turdikulova S, Uktveryte I, Utevska O, Vasquez R, Villena M, Voevoda M, Winkler CA, Yepiskoposyan L, Zalloua P, Zemunik T, Cooper A, Capelli C, Thomas MG, Ruiz-Linares A, Tishkoff SA, Singh L, Thangaraj K, Villems R, Comas D, Sukernik R, Metspalu M, Meyer M, Eichler EE, Burger J, Slatkin M, Pääbo S, Kelso J, Reich D, Krause J (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513(7518):409–413

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, Connell S, Stewardson K, Harney E, Fu Q, Gonzalez-Fortes G, Jones ER, Roodenberg SA, Lengyel G, Bocquentin F, Gasparian B, Monge JM, Gregg M, Eshed V, Mizrahi AS, Meiklejohn C, Gerritsen F, Bejenaru L, Blüher M, Campbell A, Cavalleri G, Comas D, Froguel P, Gilbert E, Kerr SM, Kovacs P, Krause J, McGettigan D, Merrigan M, Merriwether DA, O'Reilly S, Richards MB, Semino O, Shamoon-Pour M, Stefanescu G, Stumvoll M, Tönjes A, Torroni A, Wilson JF, Yengo L, Hovhannisyan NA, Patterson N, Pinhasi R, Reich D (2016) Genomic insights into the origin of farming in the ancient Near East. Nature 536(7617):419–424

Lopopolo M, Børsting C, Pereira V, Morling N (2016) A study of the peopling of Greenland using next generation sequencing of complete mitochondrial genomes. Am J Phys Anthropol 161(4):698–704

Malmström H, Gilbert MT, Thomas MG, Brandström M, Storå J, Molnar P, Andersen PK, Bendixen C, Holmlund G, Götherström A, Willerslev E (2009) Ancient DNA reveals lack of continuity between Neolithic hunter-gatherers and contemporary Scandinavians. Curr Biol 19:1758–1762

Neparáczki E, Juhász Z, Pamjav H, Fehér T, Csányi B, Zink A, Maixner F, Pálfi G, Molnár E, Pap I, Kustár Á, Révész L, Raskó I, Török T (2017) Genetic structure of the early Hungarian conquerors inferred from mtDNA haplotypes and Y-chromosome haplogroups in a small cemetery. Mol Genet Genom 292(1):201–214

Pakendorf B, Novgorodov IN, Osakovskij VL, Stoneking M (2007) Mating patterns amongst Siberian reindeer herders: inferences from mtDNA and Y-chromosomal analyses. Am J Phys Anthropol 133:1013–1027

Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K (2004) Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 74:827–845

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara- Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155–1159

Sharma S, Rai E, Sharma P, Jena M, Singh S, Darvishi K, Bhat AK, Bhanwer AJ, Tiwari PK, Bamezai RN (2009) The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. J Hum Genet 54:47–55

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262–278

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert MT, Götherström A, Jakobsson M (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336(6080):466–469

Szécsényi-Nagy A (2015) Molecular genetic investigation of the Neolithic population history in the western Carpathian Basin. PhD Dissertation

Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogvä̈li EL, Tolk HV, Reidla M, Metspalu E, Pliss L, Balanovsky O, Pshenichnov A, Balanovska E, Gubina M, Zhadanov S, Osipova L, Damba L, Voevoda M, Kutuev I, Bermisheva M, Khusnutdinova E, Gusar V, Grechanina E, Parik J, Pennarun E, Richard C, Chaventre A, Moisan JP, Bara´c L, Pericic´ M, Rudan P, Terzic´ R, Mikerezi I, Krumina A, Baumanis V, Koziel S, Rickards O, De Stefano GF, Anagnou N, Pappa KI, Michalodimitrakis E, Fera´k V, Fu¨redi S, Komel R, Beckman L, Villems R (2004) The western and eastern roots of the Saami–the story of genetic ''outliers'' told by mitochondrial DNA and Y chromosomes. Am J Hum Genet 74

Underhill PA, Kivisild T (2007) Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet 41:539–564

Vandkilde H (2007) Culture and change in the Central European Prehistory, 6th to 1st millennium BC. Aarhus Univ. Press, Aarhus

Wong EHM, Khrunin A, Nichols L, Pushkarev D, Khokhrin D, Verbenko D, Evgrafov O, Knowles J, Novembre J, Limborska S, Valouev A (2017) Reconstructing genetic history of Siberian and Northeastern European populations. Genome Res 27(1):1–14

Yao YG, Kong QP, Wang CY, Zhu CL, Zhang YP (2004) Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China. Mol Biol Evol 21(12):2265–2280

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. Mol Biol Evol 21(1):164–175

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. Am J Hum Genet 71(3):466–482

AQ3

Journal : Large 438	Article No : 1469	Pages : 16	MS Code : MGAG-D-18-00097	Dispatch : 2-7-2018

Journal:	438
Article:	1469

Springer
the language of science

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details Required	Author's Response
AQ1	Author: Since this article follows "UK English" kindly check and clarify if the word "medieval" can be changed to "mediaeval" globally.	
AQ2	Author: Kindly check and provide page range for Tambets et al. (2004).	
AQ3	Author: Fig1,Fig2,Fig3,Fig4,Fig5,Fig6,Fig7 - Figure is poor in quality as its labels are not readable. Please supply a new version of the said figure with legible labels preferably in .eps, .tiff or .jpeg format with 600 dpi resolution.	