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

Z. Juhász1 · E. Dudás2 · Horolma Pamjav2

Received: 13 February 2018 / Accepted: 28 June 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 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 54 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 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 propagation. 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; Circiu et al. 2004; Tamm et al. 2004; Sein et al. 2000; Yao et al. 2004; Pakendorf et al. 2005; Bernshcheva et al. 2004; Simon 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
hominins, ancient anatomically modern humans, and present-day populations (Lazaridis et al. 2016; Fu et al. 2016; Haak et al. 2015; Alenstoft 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 autosomes can provide detailed scenarios 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 autosomes when studying the origin and migration of human populations.

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 (Alenstoft 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 poses the following assumption: recent populations are products of prehistoric and historic interactions, disjunctions and junctions of certain ancient source populations (for instance, 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; Nepärázki 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 entities of 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 sum of individuals represented by mitochondrial, Y-chromosomal and ancient mitochondrial Hg distributions are 3,637, 6746 and 1266, respectively.

Mitochondrial and Y-chromosomal data do not perfectly coincide in three cases, when Tuscany–Sicilian, Serb–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 contexts, 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.
<table>
<thead>
<tr>
<th>Population name</th>
<th>Abbreviation</th>
<th>Population name</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Han Chinese</td>
<td>CHN</td>
<td>Mongolian</td>
<td>MNG</td>
</tr>
<tr>
<td>Kyrgyz</td>
<td>KYG</td>
<td>Chuvash</td>
<td>CHU</td>
</tr>
<tr>
<td>Tuscany</td>
<td>TUS</td>
<td>Bulgarian</td>
<td>BLG</td>
</tr>
<tr>
<td>Azeri</td>
<td>AZR</td>
<td>Turkish</td>
<td>TUR</td>
</tr>
<tr>
<td>Karacay</td>
<td>KRC</td>
<td>Hungarian</td>
<td>HUN</td>
</tr>
<tr>
<td>Slovak</td>
<td>SLK</td>
<td>Czech</td>
<td>CZH</td>
</tr>
<tr>
<td>Romanian</td>
<td>ROM</td>
<td>Kashubian Poles</td>
<td>PLK</td>
</tr>
<tr>
<td>Finnish</td>
<td>FIN</td>
<td>Norwegian</td>
<td>NOR</td>
</tr>
<tr>
<td>North German</td>
<td>GEN</td>
<td>South German</td>
<td>GES</td>
</tr>
<tr>
<td>French</td>
<td>FRA</td>
<td>Netherlands</td>
<td>DUT</td>
</tr>
<tr>
<td>Scottish</td>
<td>CO</td>
<td>Galicia</td>
<td>GAL</td>
</tr>
<tr>
<td>Northwest Amerindian</td>
<td>NAW</td>
<td>Komi Zeyan</td>
<td>KOZ</td>
</tr>
<tr>
<td>Khanty</td>
<td>KHA</td>
<td>Serbian</td>
<td>SRB</td>
</tr>
<tr>
<td>Kurdish</td>
<td>KUR</td>
<td>Russian</td>
<td>RUS</td>
</tr>
<tr>
<td>Central Amerindian</td>
<td>NAC</td>
<td>Warpath</td>
<td>War</td>
</tr>
<tr>
<td>Poles</td>
<td>POL</td>
<td>Southern Amerindian</td>
<td>NAS</td>
</tr>
<tr>
<td>Greek</td>
<td>GRE</td>
<td>Estonian</td>
<td>EST</td>
</tr>
<tr>
<td>Sami</td>
<td>SAA</td>
<td>Karelian Finn</td>
<td>KRL</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>UKR</td>
<td>Uyghur</td>
<td>UYG</td>
</tr>
<tr>
<td>Kazakh</td>
<td>KAZ</td>
<td>Mari</td>
<td>MRI</td>
</tr>
<tr>
<td>Tatar</td>
<td>TAT</td>
<td>Udmatr</td>
<td>UDM</td>
</tr>
<tr>
<td>Japanese</td>
<td>JPN</td>
<td>Szekely</td>
<td>SEK</td>
</tr>
<tr>
<td>Altai Kazakh</td>
<td>AFK</td>
<td>Hui Chinese</td>
<td>HUI</td>
</tr>
<tr>
<td>Macedonian</td>
<td>MAC</td>
<td>Lithuanian</td>
<td>LIT</td>
</tr>
<tr>
<td>Tuvan</td>
<td>TUV</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

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 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 4th Hg (H4) to 44 other Hgs (including Hg itself) results in 44 different rank lists for H4. 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 $r(i)$ denote the rank of the $i$th population in the rank list of the $i$th member of the above rank list set. The corresponding “inverse rank” value is defined as

$$r(i) = 1 - r(i)/\max(r(i))$$  \hspace{1cm} (1)
<table>
<thead>
<tr>
<th>Population</th>
<th>Sample Size</th>
<th>Location</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle East Neolithic - BrA</td>
<td>28</td>
<td>MEN</td>
<td>Middle East</td>
</tr>
<tr>
<td>Iberian Neolithic</td>
<td>45</td>
<td>IBN</td>
<td>Iberia</td>
</tr>
<tr>
<td>Near Eastern Neolithic</td>
<td>67</td>
<td>NEG</td>
<td>TR, IRN, SYR, JOR</td>
</tr>
<tr>
<td>Central Asian Neolithic (Serovo)</td>
<td>15</td>
<td>SER</td>
<td>East Siberia</td>
</tr>
<tr>
<td>Early-Middle Neolithic</td>
<td>53</td>
<td>EMN</td>
<td>Europe</td>
</tr>
<tr>
<td>Starčevo</td>
<td>44</td>
<td>STR</td>
<td>Balkans</td>
</tr>
<tr>
<td>Dniepr-Donets Neolithic</td>
<td>17</td>
<td>DDG</td>
<td>Eastern Europe</td>
</tr>
<tr>
<td>Neolithic - Hungary</td>
<td>85</td>
<td>NHI</td>
<td>Hungary</td>
</tr>
<tr>
<td>Yamnaya, Alansky</td>
<td>49</td>
<td>YAM</td>
<td>Russia, Ukraine</td>
</tr>
<tr>
<td>Kurgans (Eneolithic/Catacomb)</td>
<td>35</td>
<td>KGC</td>
<td>UKR, MLD, BLY</td>
</tr>
<tr>
<td>Baraba (UT-OVI-EK)</td>
<td>33</td>
<td>BB1</td>
<td>West Siberia</td>
</tr>
<tr>
<td>Late Neolithic - EBA Europe</td>
<td>56</td>
<td>LNB</td>
<td>Europe</td>
</tr>
<tr>
<td>Altai Bronze Age</td>
<td>12</td>
<td>ABA</td>
<td>South Siberia</td>
</tr>
<tr>
<td>Tarim Basin Xiaohe</td>
<td>73</td>
<td>XIA</td>
<td>China</td>
</tr>
<tr>
<td>Sintashta-Andronovo</td>
<td>41</td>
<td>SIA</td>
<td>Russia, Siberia</td>
</tr>
<tr>
<td>Baraba (L-K, FOD-LBB)</td>
<td>45</td>
<td>BB2</td>
<td>West Siberia</td>
</tr>
<tr>
<td>Srubnya</td>
<td>14</td>
<td>SRU</td>
<td>Russia</td>
</tr>
<tr>
<td>Bronze Age Kurgans</td>
<td>13</td>
<td>KBER</td>
<td>Kazakhstan</td>
</tr>
<tr>
<td>Baraba (Iron transition)</td>
<td>14</td>
<td>BB3</td>
<td>West Siberia</td>
</tr>
<tr>
<td>Iron Age Kurgans</td>
<td>13</td>
<td>KIK</td>
<td>Kazakhstan</td>
</tr>
<tr>
<td>Tagar - Tachyt</td>
<td>15</td>
<td>TAG</td>
<td>Russia</td>
</tr>
<tr>
<td>Seychuanan Iron age</td>
<td>14</td>
<td>SCI</td>
<td>Russia</td>
</tr>
<tr>
<td>Pazyryk Scytho-Siberian</td>
<td>25</td>
<td>PAZ</td>
<td>Mongolia, Russia</td>
</tr>
<tr>
<td>Qin China aDNA</td>
<td>19</td>
<td>QIN</td>
<td>East Asia</td>
</tr>
<tr>
<td>Egyin Gol Xiongnu</td>
<td>46</td>
<td>XO</td>
<td>Inner Asia</td>
</tr>
<tr>
<td>Lombard early medieval</td>
<td>40</td>
<td>LOM</td>
<td>Hungary, Italy</td>
</tr>
<tr>
<td>Vikings</td>
<td>65</td>
<td>VIK</td>
<td>Norway</td>
</tr>
<tr>
<td>Karos</td>
<td>90</td>
<td>KAR</td>
<td>Hungary</td>
</tr>
<tr>
<td>Hungarians 900 AD</td>
<td>27</td>
<td>AH2</td>
<td>Central Europe</td>
</tr>
<tr>
<td>Ancient Hungarian (10th century)</td>
<td>67</td>
<td>AH1</td>
<td>Central Europe</td>
</tr>
<tr>
<td>Pre-conquest Hungary</td>
<td>49</td>
<td>HPC</td>
<td>Hungary</td>
</tr>
<tr>
<td>Medieval Slovenian</td>
<td>19</td>
<td>SLV</td>
<td>Slovakia</td>
</tr>
<tr>
<td>Italian medieval</td>
<td>27</td>
<td>ITM</td>
<td>Italy</td>
</tr>
<tr>
<td>Cumarian</td>
<td>11</td>
<td>CUM</td>
<td>Hungary</td>
</tr>
</tbody>
</table>

The populations contained by the top correlation subset, while the populations missing from the correlation subset have the inverse rank value $\hat{\mu}(i) = 0$.

Using this definition, the population having the highest frequency—and consequently the lowest rank in the ranking list of $H_{gi}$—becomes the highest “inverse rank” value approaching 1, and the ranks of the other populations decrease with decreasing frequency of $H_{gi}$. Thus, our “inverse rank” is really the inverse of the common rank definition, which increases with decreasing frequency in $H_{gi}$. The inverse ranks $\hat{\mu}(i)$ are stored in the N-dimensional “inverse rank vectors (IRV)” $R_{ii}$, where the ith vector element represents the inverse rank of the ith population in the correlation subset. The serial numbers of the pair of $H_{gi}$ belonging to the $\mu$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 $R_{ij}$, 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.
measure highlighting the important components of each
inverse rank vector (Juhász et al. 2016). These 50-dimen-
sional weight vectors were also learned automatically dur-
ing the training process. The mathematical description of
the algorithm is given in ESM 3.

The complete analysis based on our self-learning com-
puter 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 iter-
ative 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 cor-
responding 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 mito-
chondrial 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) algorithm
is found in ESM 3 and (Juhász et al. 2016).

The goodness of the result was characterized by a calcu-
lated 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 coeffi-
cient 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 his-
torical 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-cal-
culate 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 posi-
tive as well as negative correlation, while this “distance”
approaches 1 for uncorrelated Hg-pairs. Thus, the rela-
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 con-
straint 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-dimen-
sional 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 signif-
cantly with probability at least 95%. The correlation of the
inference and distance matrices was −0.52.

Ordering the 393 IRVTs to the most similar IRVT, we
obtained ten clusters. The subset of Hgs whose IRV's
belong to a common cluster build the “Hg association”
propagating within the populations having nonzero inverse
rank values in the IRVTs 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, HUL, JPN, MNG, KYG, KAZ) to the native Amer-
icans (NAW, NAC, NAS) and to the Volga region in East-
ern 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 corre-
ating partners within the Hg association (Y: O, C, Q, N;
M: Z, B, F, G, M, C, D, A, N³). As this Hg association
was derived from the whole correlation subset belonging
to IRVT-10, certain Hgs may be absent from different
Fig. 1 MDS map of Y-chromosomal (a) and mitochondrial Hgs (b), determined from case 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 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 clustering was reached for 44 IRVs.
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 com-
pare the recent parts of the resulting 84-dimensional and the
originally recent 50-dimensional IRVs, we eliminated the
ancient components of the 84-dimensional IRVs and re-
calculated the modified inverse rank values of the remaining
50 recent components. After this, we selected the modified
IRVs 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
IRVs and found that the highest inverse ranks of the com-
plete 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 Xianhe,
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
eggible 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 IRVs with their most
similar historical pairs are shown in the maps of ESM_4
also showing the mitochondrial Hg content of the corre-
sponding Hg associations, See Figs S1 and S2 in ESM_4).

The accumulated rate of the Hg association in ancient East-
ern samples (XIO, XIA, QIN) is in the range of 75–95%.
The most Western appearance of this Hg association was
detected in early Hunan samples arising from 800 to
1000 AD, where it takes 15–25% 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 conquer-
or, 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, I, R1b, T, M1, N1, I, X, HV*, U*1, U2, K,
T, J). Comparing these maps to those shown in Fig. 2 shows

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

The geographical distribution of IRVT-1 means 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 (Starčevo, 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, 11, R1b, G, 12; Mt, H1, V, J, K, U5, T) has some common elements with IRVT-1 (Y; R1b, G; R1a, H1b).
Mt: K1, but Y: H1 and the total set of the mitochondrial Hgs except K are not found there. On the other hand, the set (Y: J2, J1, T, I, E; Mt: HV*, U1*, 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 rank values for the Neolithic European samples STR (Starevo, 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 5S and 6S 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, H1, J2, J1, F; Mt: H, K, W, U*, N*, X, I) 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 H1 and J2 are among the most important components of the Hg association of IRVT-3, whereas J2 and F 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, U*)
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 from 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, I, R1a, R1b, 12, 12, 12, Mt: H, I, V, U5, T) and (11, R1a, 12; Mt: H, J, U3, N5, X). The overlap between this Hg association with those of IRVT-3 and IRVT-2 are (Y: I, R1b, Mt: H, V, J, U5, T) and (I1, R1a, 12; Mt: H, J, U3, N5, X). The Hgs being found in both overlaps (Y: I, 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; Mt: T, U3, I, N5, 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 (Yamay, Altai, Bashkortostan, Russia, Ukraine, 5000–2700 BC) (see Figs S9-12 in ESM_4). Early medieval Viking (Norway, 780–790 AD) sample has also a high inverse rank of this Hg association. The complete set of the components of the historical IRVTs (Mt: H, I, U4, U5, T, W, U3, 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.
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*; (see Fig. 7). The largest overlaps (Y: R1a, R1b, J2, I2, E; Mt: H, U4, J, K) and (Y: R1a, R1b, I2, E; Mt: H, U4, J, K, 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 and 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—Seythians, Sarmatiats, Huns, Avars, Hungarians, Cumans, etc.—invaded Eastern Europe and the Carpathian Basin.

This is the reason, why early medieval Hungarian samples HPC and AH2 also fall 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 IRVTs 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).
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 interference 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.
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 (Chimpanzees et al. 2004; Biró 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; Derenzo 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).
showing the Fertile Crescent as source area of the corresponding \( H_g \) association propagating to Europe through Asia Minor, the Balkans and the Carpathian Basin. As the mitochondrial part of this \( H_g \) association was also detected with high inverse ranks in ancient populations in the Fertile Crescent, IRVT can be attributed to the well-known migration of Neolithic farmers starting from the Middle East (Bramanti et al. 2008; Malmström et al. 2009; Skoglund et al. 2012).

Also, very clear correlation between the geographical distribution and the place of origin of the corresponding \( H_g \) 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 \( H_g \) association are found in ancient European populations, IRVT-3 can be considered as the \( H_g \) association of Neolithic indigenous Europeans (Gamba et al. 2014; Szécsényi-Nagy 2015; Krivsild 2017; Wong et al. 2017).

The three \( H_g \) 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 \( H_g \) associations derived from the remaining 3 IRVTs clearly mirror the admixture of the \( H_g \) 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 (Antony 2007; Kristiansen and Larsson 2005; Kristiansen 2007; Wong et al. 2017). The high inverse ranks of the mitochondrial part of this \( H_g \) association in early Medieval Hungarian samples also support the Eastern European origin of this population.

The similarity of the geographical distribution and \( H_g \) 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 \( H_g \) 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 \( H_g \) 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 \( H_g \) 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 (Neparat et al. 2017; Czynarski and Maciejewska 2016; Szécsényi-Nagy 2015; Gamba et al. 2014; Korfák and Epimakhov 2014).

The origin of the \( H_g \) 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 \( H_g \) 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 \( H_g \) associations derived from recent data is extremely high. In principle, the frequency of an \( H_g \) in a population can be summed up by more \( H_g \) associations constructing the given population, because of the overlaps of their \( H_g \) contents (Zerjal et al. 2002; Sharma et al. 2009). Therefore, these cumulated rates refer merely to a possible maximal rate of the \( H_g \) associations, and the actual rates may be lower. However, an extremely high cumulated rate of a \( H_g \) association in an ancient population may refer to a situation anteceding 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 \( H_g \) distributions in credible interpretation of the results. The mathematically correct estimation of the rates of the \( H_g \) associations in an actual \( H_g \) 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
could be extended to study the correlations of cultural and
genetic characteristics, to validate linguistic, archaeologi-
cal, ethnographic, etc., theories and hypotheses by
are the same species could also
reveal joint propagations of different associations of plants
and/or animals.

Acknowledgements The authors are grateful to Tiber 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 amend-
ments or comparable ethical standards.

References

Allentoft ME, Sikora M, Stigter C, Rasmussen S, Rasmussen M,
Veerkamp J, Damgaard PK, Schroeder H, Ahlstrom T, Vinther J,
Malaspinas AS, Margaryan A, Higuchi T, Carnell D, Lynnerup N,
Haesig L, Baran J, Delisa Can P, Danilov P, Daffertshofer P, Helg 
Av, Epimakhov A, Frei K, Fumagali M, Gratia T, Grøning A,
Grönkvist S, Grube G, Højbjerg T, Jærys R, Katsaragou A,
Moiseyev V, Pali L, Paliyi G, Pouto D, Dubovyk I, Pace D,
Naag L, Sabin M, Strehaia S, Strehaia S, Strehaia S, Strehaia S,
Thoth G, Triftskaia SV, Vazir I, Vosseli M, Yegorovskaia I,
Zheiten E, Orlandi L, Schirazer-Pontes T, Berman S, Nielsen R,
Kristiansen K, Iske Willerslev E (2015) Population genetics of
bronze age europe. Nature 529:757-762

Anthony D (2007) The horse, the wheel and language. How bronze-
age riders from the Eurasian Steppe shaped the modern world.

Batini C, Hallast P, Votyte M, Zutik J, Dikovski I, Padjam H, Sojer-
tile A, Wesolow JH, Deboulet P (2017) Population reconceiv-
ing of European archaeological genomes highlights sex-bias in Bronze Age demographic expansions. Sci Rep 7(1):12085

Bermisheva MA, Kater IA, Kotoshchina TY, Dabova NA, Vilems R,
DNA from the poggetti: a strong mixture of maternal lineages from
Eastern and Western Europe. Mol Biol 38:516-523

nosition of the Madars (Kurzhauten) and the Magyars (Hungary).
Ann J Phys Anthropol 139(3):405-410

Brantam B, Thomas MG, Haak W, Unterlaender M, Jones P, Tamberi
Genetic discontinuity between local hunter-gatherers and Central
Europe’s first farmers. Science 326:137-140

Cromb C, King R, Kivisild T, Kallneg T, Asante F, Cavalli GL,
Lilie AS, Rosenman CC, Lin A, Prince P, Oefner P, Shul P,
Y-chromosomal Haplotypic Strata in Anatolia. Hum Genet 114:127-148

Cyranowski W, Maciejewski A (2016) The proto-Slavic warrior in
Europe: the scythians, samaritans and leksis. Ide movement for culture. J Martial Arts Antrhopol 16(6):14


group N dispersals from south Siberia to Europe. J Hum Genet 52(9):376-370


Forster P, Hurto R, Torroni A, Bandelt HH (1996) Origin and evolu-
tion of the native American mtDNA: a reappraisal. Am J Hum Genet 59:491-495

Fuk Qi, Rose C, Haidmik M, Pet M, Malik J, Fernandes D, Fum-


© Springer

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


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.

<table>
<thead>
<tr>
<th>Query</th>
<th>Details Required</th>
<th>Author’s Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQ1</td>
<td>Author: Since this article follows &quot;UK English&quot; kindly check and clarify if the word &quot;medieval&quot; can be changed to &quot;mediaeval&quot; globally.</td>
<td></td>
</tr>
<tr>
<td>AQ2</td>
<td>Author: Kindly check and provide page range for Tambets et al. (2004).</td>
<td></td>
</tr>
<tr>
<td>AQ3</td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>