Abstract— This paper describes archaeogenetic data mining results based on a novel mitochondrial and y-chromosome haplogroup distance metric. The analysis shows that the Minoan genes are composed of two originally distinct groups. One group was the descendant of Mesolithic European hunter-gatherers, while the other group was from Neolithic farmer populations from Anatolia. These groups intermingled in the Danube Basin and the western Black Sea littoral area for thousands of years as indicated by the presence of mutations that appear first in those areas. A group from that population migrated southward to Crete to form the Minoan culture.

Keywords— Archaeogenetics, data mining, distance metric, haplogroup, Minoan, mtDNA, y-DNA.

I. INTRODUCTION

For a long time, archaeologists, geneticists and linguists thought that the Minoans came either from the Near East or Africa (Bernal [2], Campbell-Dunn [3], Evans [9], Gimbutas [10], Gordon [12], Hughey et al. [14], Lazaridis et al. [15], and Marinatos [16]).

However, careful data mining of the archaeogenetical data led to a revolution in our view of Minoan origins. In particular, we show in this paper that the Minoans came from the Danube Basin and the western Black Sea littoral area.

That surprising result is possible due to data mining and data science that use generally applicable methods. These methods, which are also suitable to archaeogenetical data, are different from the common statistical methods used in Hughey et al. [14] and Lazaridis et al. [15]. These statistical methods have several limitations in dealing with deep-clade data. Instead we propose in Section 3 a new similarity measure on a set of haplogroups.

The rest of this paper is organized as follows. Section 2 describes the Minoan mitochondrial DNA (mtDNA) and y-chromosomal DNA (y-DNA) data sources and our reclassifications of the mtDNA haplogroups.

Section 3 describes similarity measures and distance metrics on sets of haplogroups.

Section 4 presents some experimental results. In particular, we compare the Minoan archaeogenetical data with those of several ancient European cultures in the Danube Basin and the western Black Sea littoral area.

Section 5 describes how the present archaeogenetic data mining results on Minoan origins are supported by cultural similarities between Neolithic and Bronze Age cultures in the Danube Basin and the Minoan culture in Crete. These cultural similarities include art motifs, metallurgy, linguistics based on the decipherment of Minoan texts, the potter’s wheel technology and writing scripts.

Section 6 reviews earlier archaeogenetic results on the relationships and origins of the Minoans.

Finally, Section 7 gives some conclusions and directions for future work.

II. DATA SOURCES AND RECLASSIFICATIONS

We collected all the available Minoan genetic data from the European Nucleotide Archive (ENA) and the original articles by Hughey [14] and Lazaridis [15]. Table 1 shows the ENA ID in the first column and the alias IDs that were used in the original articles in the second column. The third column shows the mtDNA haplogroups and y-DNA haplogroups as reported in the original articles and/or the mtDB. The fourth column shows our classification of these haplogroups. We corrected the classifications of eleven haplogroups based on the human mitochondrial PhyloTree Build 17 (February 18, 2016), which is the latest available version.

The fifth and the sixth columns list the set of reported mutations with respect to the revised Cambridge Reference Sequence (rCRS). In particular, the fifth columns lists mutations below 16,000 and the sixth column lists the mutations above 16,000 in the efficient format of giving only the last three digits. These reported mutations help explain the reclassifications. For example, consider the Minoan mitochondrial genome M13, which had only a partial read. Fig. 1 shows a part of the successfully read genome.
This genome can be reclassified as belonging to the I5a1b mtDNA haplogroup. The classification of I5a is warranted by the reported mutation 16148T as shown in Fig. 2. In addition, Hughey et al. [14] reported mutations at positions 199 and 250 but not at position 204. It is likely that the DNA sequencing machine read correctly some small fragment that included every position from 199 to 250, including position 204. Hence we can assume 204 to be a position without a mutation with respect to both rCRS and RSRS, which is used by PhyloTree Build 17. Both of these reference sequences have 204T and reporting no mutation at 204 is equivalent to the back mutation 204T shown in the classification tree in Fig. 2. That explains the I5a1b reclassification. The explanations of the other reclassifications are shown in Figs. 5 and 6 in the Appendix.

In Table 1, we highlighted in green those haplogroups for which there were no other haplogroups with a deeper level classification. These green shaded haplogroups were used in our deep-clade analysis in the next section.

For each Minoan haplogroup, we also collected other ancient mtDNA and y-DNA that were closely related from the amtDB database of Ehler et al. [8] and from Allentoft et al. [1] Csáký et al. [5], Neparáczki et al. [17-18] and Šebest et al. [33]. In Table 2, the origin of the genes is indicated by a yellow highlighting in case of a Fertile Crescent or Anatolian origin and a blue highlighting in case of a Mesolithic European origin. The y-DNA J2a haplogroup, which occurs in Mesolithic Georgia is also highlighted in blue. It likely also entered Europe in the Mesolithic given the Neolithic presence of y-DNA J2a1 in Croatia.

---

>gi|293791582|gb|HM022316.1| Homo sapiens isolate M13; mitochondrial

16056 ccaagtattg actcaccct caacaaccgc tatgtatctc gtacattact gccagccacc
16116 atgaatattg tacagtacca taaatacctg aactacgta gtacataaaa acccaatcca
16176 catcaaaaa ccctcccccccct gccttcaccgc aagtacagca atcaaccttc aacctacaca
16236 catcaactgc aacctccaaag cccacctctca cccactagga taccaacaaa cctacccacc
16296 ctttaacagta catagtacat aaagccatts accgtacata gcacattaca gtcaaatccc
16356 ttctcgtccc

---

Fig. 1. A part of the mitochondrial control region of Minoan sample M13.

Fig. 2. The reclassification of Minoan sample M13 as mtDNA haplogroup I5a1b.
Table 1. The currently known Minoan mtDNA (black) and y-DNA (red) haplotypes from Moni Odigitria, Heraklion, Crete (gray) [15] and from the Hagios Charalambos Cave, Lasithi, Crete [14,15].

<table>
<thead>
<tr>
<th>ENA ID</th>
<th>Alias ID</th>
<th>mtDNA or y-DNA</th>
<th>Reported mutations with respect to rCRS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reported</td>
<td>Corrected</td>
</tr>
<tr>
<td>HM022275</td>
<td>6AH</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022276</td>
<td>7AH</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022285</td>
<td>16AH</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022286</td>
<td>17AH</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022288</td>
<td>19AH</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022302</td>
<td>20H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022303</td>
<td>7H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022294</td>
<td>8H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022299</td>
<td>16H</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>HM022314</td>
<td>M11</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>ERS1770862</td>
<td>10073</td>
<td>H1bm</td>
<td></td>
</tr>
<tr>
<td>ERS1770864</td>
<td>19005</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>ERS1770869</td>
<td>19129</td>
<td>H+163</td>
<td></td>
</tr>
<tr>
<td>HM022278</td>
<td>9AH</td>
<td>H5</td>
<td></td>
</tr>
<tr>
<td>ERS1770863</td>
<td>10074</td>
<td>H51g</td>
<td>2706A</td>
</tr>
<tr>
<td>HM022281</td>
<td>12AH</td>
<td>H7</td>
<td></td>
</tr>
<tr>
<td>ERS1770861</td>
<td>10070</td>
<td>H13a1</td>
<td></td>
</tr>
<tr>
<td>HM022271</td>
<td>2AH</td>
<td>H13a1a</td>
<td></td>
</tr>
<tr>
<td>HM022310</td>
<td>M6</td>
<td>HV</td>
<td>HV+16311</td>
</tr>
<tr>
<td>HM022304</td>
<td>24H</td>
<td>HV</td>
<td></td>
</tr>
<tr>
<td>HM022270</td>
<td>1AH</td>
<td>HV</td>
<td></td>
</tr>
<tr>
<td>HM022316</td>
<td>M13</td>
<td>I5</td>
<td></td>
</tr>
<tr>
<td>HM022318</td>
<td>M17</td>
<td>I5</td>
<td></td>
</tr>
<tr>
<td>HM022315</td>
<td>M12</td>
<td>I5</td>
<td></td>
</tr>
<tr>
<td>ERS1770868</td>
<td>19128</td>
<td>I5</td>
<td></td>
</tr>
<tr>
<td>HM022305</td>
<td>26H</td>
<td>J2</td>
<td></td>
</tr>
<tr>
<td>ERS1770867</td>
<td>19127</td>
<td>J2b1a1</td>
<td></td>
</tr>
<tr>
<td>HM022322</td>
<td>M22</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>HM022317</td>
<td>M14</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>HM022321</td>
<td>M21</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>HM022284</td>
<td>15AH</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>ERS1770871</td>
<td>19131</td>
<td>K1a2</td>
<td></td>
</tr>
<tr>
<td>HM022311</td>
<td>M7</td>
<td>R0</td>
<td></td>
</tr>
<tr>
<td>HM022287</td>
<td>18AH</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>HM022295</td>
<td>9H</td>
<td>T1</td>
<td>T1a4</td>
</tr>
<tr>
<td>HM022319</td>
<td>M19</td>
<td>T2</td>
<td>T2b</td>
</tr>
<tr>
<td>HM022272</td>
<td>3AH</td>
<td>T2</td>
<td>T2b</td>
</tr>
<tr>
<td>HM022323</td>
<td>M23</td>
<td>T3</td>
<td>T2c1</td>
</tr>
<tr>
<td>ERS1770870</td>
<td>19130</td>
<td>U3b3</td>
<td></td>
</tr>
<tr>
<td>HM022291</td>
<td>4H</td>
<td>U5a</td>
<td>U5a1f1</td>
</tr>
<tr>
<td>ERS1770860</td>
<td>10071</td>
<td>U5a1</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>M4</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>HM022292</td>
<td>6H</td>
<td>W</td>
<td>W1</td>
</tr>
<tr>
<td>HM022312</td>
<td>M8</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
III. HAPLOGROUP DISTANCE METRICS

Let \( L(r,a,b) \) be the number of shared classification levels between haplogroups \( a \) and \( b \) beyond a reference \( r \) haplogroup. For example, if the reference is the RSRS, whose haplogroup description is the empty string, then we have:

\[
L(\epsilon, I, I5a1a) = 1
\]

and

\[
L(\epsilon, I5a1a, I5a1b) = 4
\]

If we take the reference to be mtDNA haplogroup \( I \), then we have:

\[
L(I, I, I5a1a) = 0
\]

and

\[
L(I, I5a1a, I5a1b) = 3
\]

Further, let \( M(r,a,b) \) be the number of shared mutations between \( a \) and \( b \) with respect to a reference \( r \) haplogroup.

For example, suppose that we take mtDNA haplogroup \( I \) as a reference. As Fig. 2 shows, the extra number of mutations from \( I \) to \( I5 \) is one, while the extra number of mutations from \( I \) to \( I5a \) is three. Each deletion also can be counted as one mutation. Since 8287-8289d indicates nine deletions, the extra number of mutations from \( I5a1a \) to \( I5a1b \) is thirteen. Since \( I5a1a \) and \( I5a1b \) share all of these mutations beyond the reference \( I \), we have:

\[
M(I, I, I5a1a) = 0
\]

and

\[
M(I, I5a1a, I5a1b) = 13
\]

Let \( S_1, S_2, \) and \( S_3 \) be three arrays of haplogroups, each with \( n \) elements such that \( S_i[i] \) is a common precedent of \( S_j[i] \) and \( S_k[i] \) for any \( 1 \leq i \leq n \). Then the average added levels function \( Al(S_1, S_2, S_3) \) is defined as follows:

\[
Al(S_1, S_2, S_3) = \frac{\sum_{i=1}^{n} L(a_i, b_i, c_i)}{n} \tag{1}
\]

Further, the average added mutations function \( Am(S_1, S_2, S_3) \) is defined as follows:

\[
Am(S_1, S_2, S_3) = \frac{\sum_{i=1}^{n} M(a_i, b_i, c_i)}{n} \tag{2}
\]

Example 1. Let \( n = 1 \) and suppose that \( S_1, S_2, \) and \( S_3 \) contain, respectively, only the haplogroups \( a = I, b = I5a1a, \) and \( c = I5a1b \). Then, by Equations (1) and (2) we have:

\[
Al(S_1, S_2, S_3) = L(I, I5a1a, I5a1b) = 3
\]

and

\[
Am(S_1, S_2, S_3) = M(I, I5a1a, I5a1b) = 13
\]

Intuitively, \( Al \) and \( Am \) are similarity functions for sets of haplogroups \( S_2, \) and \( S_3 \) that have a common ancestor \( S_1 \). Here the larger values that \( Al \) and \( Am \) give, the more similar \( S_2 \) and \( S_3 \) are to each other.

For a distance metric, we need the opposite. The larger values should indicate less similarity. Let us define the length function as:

\[
Ln(a) = L(\epsilon, a, a)
\]

and the mutations function as:

\[
Mut(a) = M(RSRS, a, a)
\]

Let us also define \( Pre(a,b) \) to be the common prefix of two strings \( a \) and \( b \), which in our examples will be haplogroup descriptions. For example,

\[
P(I5a1a, I5a1b) = I5a1
\]

Now we define the distance metric as follows.

**Definition 1.** Given any fixed ancestral population, the distance between any two of its descendant populations characterized by haplogroup arrays \( A=[a_1, \ldots, a_n] \) and \( B=[b_1, \ldots, b_n] \) where for each \( 1 \leq i \leq n \) the haplogroups \( a_i \) and \( b_i \) correspond to each other, is defined as:

\[
D(A,B) = \frac{1}{2n} \sum_{i=1}^{n} \ln(a_i) + \ln(b_i) - 2\ln(P(a_i, b_i))
\]

**Example 2.** Let \( A=[I5a1a] \) and \( B=[I5a1b] \). Then we have:

\[
D(A,B) = \frac{\ln(I5a1a) + \ln(I5a1b) - 2\ln(I5a1)}{2}
\]

The above is intuitive because the two haplogroups differ from each other only at the lowest grouping. The lowest grouping of \( I5a1a \) is \( a \), while the lowest grouping of \( I5a1b \) is \( b \).

**Definition 2.** A mathematical metric \( D \) on pairs of haplogroup arrays satisfies the following conditions [21]:

\[
D(A,B) = \frac{\ln(a_i) + \ln(b_i) - 2\ln(P(a_i, b_i))}{2}
\]
\( D(A, B) \geq 0 \) \hspace{1cm} (3)
\( D(A, B) = 0 \iff A = B \) \hspace{1cm} (4)
\( D(A, B) = D(B, A) \) \hspace{1cm} (5)
\( D(A, B) \leq D(A, C) + D(C, B) \) \hspace{1cm} (6)

Next we prove the following theorem.

**Theorem 1.** Function \( D \) is a mathematical metric.

**Proof:** We have to show that Equations (3-6) are all true.

*Show Equation (3):* Equation (3) can be shown as follows. First note that the length of the common prefix of two strings \( a \) and \( b \) is always less than or equal to the minimum of the lengths of the two strings, that is:

\[
\ln(P(a, b)) \leq \min(\ln(a), \ln(b))
\]

Further, twice the minimum of two non-negative values is always less than or equal to their sum. Hence,

\[
2 \min(\ln(a), \ln(b)) \leq \ln(a) + \ln(b)
\]

It follows from the above two inequalities that:

\[
\ln(a) + \ln(b) - 2\ln(P(a, b)) \geq 0
\]

From the previous condition follows that in the summation of Definition 1, for each \( 1 \leq i \leq n \) we add a non-negative value to the sum. Hence the entire sum must be non-negative.

*Show Equation (4):* Let us assume that we have \( D(A, B) = 0 \). We can argue as before that we only add non-negative values in the summation of Definition 1. Therefore, \( D(A, B) = 0 \) can only happen if for each \( 1 \leq i \leq n \) we have the following:

\[
\ln(a_i) + \ln(b_i) - 2\ln(P(a_i, b_i)) = 0 \hspace{1cm} (7)
\]

Now we reason by showing a contradiction to an assumption. Suppose that \( a_i \neq b_i \) for some \( 1 \leq i \leq n \). Then without loss of generality \( a_i = de \) and \( b_i = df \) where \( d \) and \( e \) may be empty strings but \( f \) is a non-empty string and \( \ln(P(e, f)) = 0 \). Then the left hand side of Equation (7) can be rewritten as follows:

\[
\ln(de) + \ln(df) - 2\ln(d) = \ln(e) + \ln(f)
\]

Since \( f \) is non-empty, the above is greater than zero, which clearly contradicts Equation (7). Hence the assumption was incorrect. Hence \( a_i = b_i \) must be true for all \( 1 \leq i \leq n \). Hence, if \( D(A, B) = 0 \) then \( A = B \).

For the reverse direction, if we have \( A = B \), then \( D(A, B) = 0 \) follows directly from Definition (1).

*Show Equation (5):* This follows from the fact that the prefix function is symmetric. Hence \( D \) is also a symmetric function.

*Show Equation (6):* This is called the triangle inequality. To prove the triangle inequality it is enough to show that for each triplet of strings \( a, b \) and \( c \), we have:

\[
\ln(a) + \ln(b) - 2\ln(P(a, b)) \leq (\ln(a) + \ln(c) - 2\ln(P(a, c))) + (\ln(c) + \ln(b) - 2\ln(P(c, b)))
\]

The above can be simplified as follows:

\[
\ln(P(a, c)) + \ln(P(c, b)) \leq \ln(P(a, b)) + \ln(c) \hspace{1cm} (8)
\]

There are two cases.

**Case I:** \( \ln(P(a, c)) < \ln(P(a, b)) \). In this case Equation (8) follows from the fact that \( \ln(P(c, b) \leq \ln(c) \) must be true.

**Case II:** \( \ln(P(a, b)) \leq \ln(P(a, c)) \). In this case, the strings \( a, b \), and \( c \) must have the forms:

\[
a = de
\]
\[
b = dg
\]
\[
c = deh
\]

In terms of concatenation of some possibly empty substrings \( d, e, f, g, \) and \( h \) where \( \ln(P(e, g)) = 0 \). Hence Equation (8) can be written as:

\[
\ln(de) + \ln(d) \leq \ln(d) + \ln(deh)
\]

The above inequality is equivalent to:

\[
\ln(d) + \ln(e) \leq \ln(d) + \ln(e) + \ln(h)
\]

Simplifying we get:

\[
0 \leq \ln(h)
\]

The above is clearly true because the length of the strings \( d \) and \( h \) are non-negative.

Since both cases are true, Equation (6) also has to be true.

Next we define and prove an alternative distance metric.

**Definition 2.** Given any fixed ancestral population, the distance between any two of its descendant populations characterized by haplogroup arrays \( A = [a_1, \ldots, a_n] \) and \( B = [b_1, \ldots, b_n] \) where for each \( 1 \leq i \leq n \) the haplogroups \( a_i \) and \( b_i \) correspond to each other, is defined as:
\[ d(A,B) = \frac{1}{2n} \sum_{i=1}^{n} \text{Mut}(a_i) + \text{Mut}(b_i) - 2\text{Mut}(a_i, b_i) \]

**Theorem 2.** Function \( d \) is a mathematical metric.

**Proof:** We have to show that Equations (3-6) are all true for function \( d \).

*Show Equation (3):* For any pair of haplogroups \( a \) and \( b \), the common prefix of \( a \) and \( b \) is a haplogroup that is shorter or equal in length to \( a \). Hence it requires fewer or equal number of mutations than \( a \) requires according to the phylogenetic tree. Hence

\[ \text{Mut}(a) \geq \text{Mut}(P(a, b)) \]

By symmetry we also have:

\[ \text{Mut}(b) \geq \text{Mut}(P(a, b)) \]

The above two inequalities imply that:

\[ \text{Mut}(a) + \text{Mut}(b) - 2\text{Mut}(P(a, b)) \geq 0 \quad (9) \]

The above inequality implies that each of the \( n \) terms of the summation is non-negative. Hence the value of the entire sum in Definition (2) is also non-negative. That shows that \( d(A,B) \geq 0 \).

*Show Equation (4):* Let us assume that we have \( d(A,B) = 0 \). Then by Equation (9), each of the \( n \) terms of the summation in Definition (2) has to be equal to zero. That can happen only if for each \( 1 \leq i \leq n \) the \( a_i = b_i \). Therefore \( A=B \).

For the other direction, if \( A=B \), then it is straightforward to show that \( d(A,B) = 0 \) by using Definition (2).

*Show Equation (5):* Since the prefix function is symmetric, the \( a_i \) and \( b_i \) in Definition (2) are interchangeable. Hence \( d \) is also symmetric.

*Show Equation (6):* To prove the triangle inequality, it is enough to show that for any triplet of haplogroups \( a, b, \) and \( c \) the following holds:

\[ d(a,b) \leq d(a,c) + d(c,b) \]

By Definition (2), the above is equivalent to:

\[ \text{Mut}(a) + \text{Mut}(b) - 2\text{Mut}(P(a,b)) \leq (\text{Mut}(a) + \text{Mut}(c) - 2\text{Mut}(P(a,c))) + (\text{Mut}(c) + \text{Mut}(b) - 2\text{Mut}(P(c,b))) \]

Simplifying the above, we get the following:

\[ \text{Mut}(P(c,b)) + \text{Mut}(P(c,b)) \leq \text{Mut}(P(a,b)) + \text{Mut}(c) \quad (10) \]

We argue by two mutually exclusive cases.

**Case I:** \( \text{Ln}(P(a,c)) < \text{Ln}(P(a,b)) \). In this case \( \text{Mut}(P(a,c)) < \text{Mut}(P(a,b)) \) and \( \text{Mut}(P(c,b)) \leq \text{Mut}(c) \) both trivially hold. Hence Equation (10) is true.

**Case II:** \( \text{Ln}(P(a,b)) \leq \text{Ln}(P(a,c)) \). In this case, the strings \( a, b, \) and \( c \) must have the forms \( a = def, b = dg, \) and \( c = deh \) in terms of concatenation of some possibly empty substrings \( d, e, f, g, \) and \( h \) where \( \text{Ln}(P(e,g)) = 0 \). Then Equation (10) can be written as follows:

\[ \text{Mut}(de) \leq \text{Mut}(deh) \]

The above is true because the right hand side may have some extra mutations over the left hand side if \( h \) is a non-empty string.

Since both cases are true, Equation (6) also has to be true. ■

**IV. EXPERIMENTAL RESULTS**

In Table 3, we calculated the average added mutations using Equation (3). Here we took \( S_i \) and \( S_i \) to be the haplogroups in the first and last columns of Table 2, respectively. We also took \( S_{i} \) to be an array that consists always of the haplogroup from the intermediate columns that best fits to the Minoan haplogroup in each row. We did not divide the \( S_i \) into separate groups of European time periods from Neolithic, Creded Ware etc. cultures because these cultures are genetically continuous in spite of the fact that there are many missing pieces of data. These missing pieces are likely due to the still too few samples that we have available from many periods of time. Therefore, it made more sense to treat all the European periods as one group. If more data become available to fill in the missing entries, then it will make sense to separate them.

When the mutational change was only one extra mutation that was not indicated by a subhaplogroup, then we rounded that up by one level. For example, this had to be done in the case of mutation HV+16311.

According to the calculations in Table 3, in the Minoan and ancient European samples the average added levels is exactly one over the twenty different haplogroups that already existed in Mesolithic Europe or the Fertile Crescent/Anatolia. The distance is 0.525 levels. Instead of levels we could calculate the number of mutations, but it would be more complex. We chose for simplicity to count only the number of levels.

In both the mtDNA and the y-DNA phylogenetic trees, one classification level is equal to many individual mutations. The interesting question is what time one classification level means. Clearly, it means several thousand years, but we may know the precise value only when more data is available.
<table>
<thead>
<tr>
<th>Mesolithic or Fertile Crescent</th>
<th>Neolithic Old Europe</th>
<th>Corded Ware Cult.</th>
<th>Copper/BA Cent. Europe</th>
<th>Medieval Cent. Europe</th>
<th>Neolithic/BA Pontic Steppe</th>
<th>Minoan Crete</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>6450-6380</td>
<td>Catalhöyük, Turkey</td>
<td>H1 5976-5751 Zemunica Cave</td>
<td>H1c 2800-2300 Hunieck Poland</td>
<td>H1b1 2500-2200 Szigetszentmikl os, Hungary</td>
<td>H1 4949-4799 Derovka, Ukraine</td>
</tr>
<tr>
<td>H5</td>
<td>6500-6200</td>
<td>Barcin, Turkey</td>
<td>H5a 5800-5400 Malak Prelsavets, Bu</td>
<td>H5a1 2625-2400 Jäbara, Estonia</td>
<td>H5a1g 2500-2200 Prague, Czech R.</td>
<td>H5a 3931-3640 Vertebr C., U.</td>
</tr>
<tr>
<td>H</td>
<td>6450-6380</td>
<td>Čatalhöyük, Turkey</td>
<td>H7c 5641-5560 Kardagur, Cr.</td>
<td>H7d 2900-2200 Brandysek, CR.</td>
<td>H7c 4455-4359 Yunasite, Bul.</td>
<td>H7 2000-1700 Charalambos</td>
</tr>
<tr>
<td>J2a</td>
<td>7940-7600</td>
<td>Ktias Kilde, Georgia</td>
<td>J2a1 4692-4546 Osierek, Croatia</td>
<td>J2a1a 1270-1110 Lujas-Varjia, H.</td>
<td>J2a1a 950-1000 Sarrattudvar, H.</td>
<td>J2a1d 2000-1700 Charalambos</td>
</tr>
<tr>
<td>H1a</td>
<td>8300-7400</td>
<td>Lepenski Vir, Serbia</td>
<td>H13a1a2c 2457-2142 Quedlinburg, G.</td>
<td>H13a1a2 412-604 Szolad, Hu.</td>
<td>H13a1a 3300-2700 Samara, Rus.</td>
<td>H13a1a 2000-1700 Charalambos</td>
</tr>
<tr>
<td>H</td>
<td>5836-5632</td>
<td>Lepenski Vir, Serbia</td>
<td>HV+16311 5500-5300 Mezökövesd, H.</td>
<td>HV0e Saxony-Anhalt, Ger.</td>
<td>HV+16311 3600-3000 Budakalász, Hu.</td>
<td>HV 3758-3636 Vertebr C., U.</td>
</tr>
<tr>
<td>K1a2</td>
<td>6400-5600</td>
<td>Mentese, Turkey</td>
<td>Kla2a 4230-3995 Kletios, Gr.</td>
<td>Kla2a 2621-2472 Víby, SE</td>
<td>Kla2a 500-600 Szolad, Hu.</td>
<td>Kla2 5715-5626 Odigitria, Gr.</td>
</tr>
<tr>
<td>R0a</td>
<td>7722-7541</td>
<td>'Ain Ghazal, Jordan</td>
<td>R0a1 3095-2915 Ozeria, Ukraine</td>
<td>R0a1 2500-2200 Vertebr C., U.</td>
<td>R0a1 2000-1700 Charalambos</td>
<td></td>
</tr>
<tr>
<td>T1a</td>
<td>7446-7058</td>
<td>'Ain Ghazal, Jordan</td>
<td>T1a 5500-3700 Diros, Greece</td>
<td>T1a1 2570-2471 Kardosfor, G.</td>
<td>T1a1 895-950 Karos, Hu.</td>
<td>T1a1 1598-1398 Kapova C., Ra.</td>
</tr>
<tr>
<td>T2a</td>
<td>6500-6200</td>
<td>Barcin, Turkey</td>
<td>T2b 5800-5400 Malak Prelsavets, Bu</td>
<td>T2b 2457-2201 Szigetszentmikl os, Hungary</td>
<td>T2b 950-1000 Sarrattudvar, Hungary</td>
<td>T2b 2000-1700 Charalambos</td>
</tr>
<tr>
<td>T2c</td>
<td>8204-7755</td>
<td>T. Abdul Hosein, Iran</td>
<td>T2c 5500-3700 Diros, Greece</td>
<td>T2c 2500-2200 Szigetszentmikl os, Hungary</td>
<td>T2c 500-600 Fonyód, Hungary</td>
<td>T2c1a 3339-2918 Samara, Russia</td>
</tr>
<tr>
<td>T2</td>
<td>6500-4000</td>
<td>Ukraine</td>
<td>T2e 5800-5400 Malak Pr. Bu</td>
<td>T2e 2500-2300 Maláyece, Pl.</td>
<td>T2e 412-604 Szolad, Hu.</td>
<td>T2c 2000-1700 Charalambos</td>
</tr>
<tr>
<td>U3b</td>
<td>6450-6380</td>
<td>Catalhöyük, Boncuklu</td>
<td>G2a2b2a 5800-5400 Malak Prelsavets, Bu</td>
<td>G2a2b2a 3910-3650 Abony, Hu.</td>
<td>G2a2b2a 895-950 Karos, Hu.</td>
<td>U3b 2210-1680 Odigitria, Gr.</td>
</tr>
<tr>
<td>U5a1c</td>
<td>9140-8570</td>
<td>Padina, Serbia</td>
<td>U5a1a 5207-4945 Osierek, Croatia</td>
<td>U5a1a 4250-2050 Esperstedt, G.</td>
<td>U5a1a 895-950 Karos, Hu.</td>
<td>U5a1c 2000-1700 Charalambos</td>
</tr>
<tr>
<td>U5b1b</td>
<td>7940-7585</td>
<td>Vlasac, Serbia</td>
<td>U5b1b 16290 5300-4900 Hejöjkúrt, Hu.</td>
<td>U5b1b 2500-2050 Quedlinburg, G.</td>
<td>U5b1b 4000-3600 Vertebr C., U.</td>
<td>U5b1b 2210-1680 Odigitria, Gr.</td>
</tr>
<tr>
<td>W1+119</td>
<td>6500-6200</td>
<td>Barcin, Turkey</td>
<td>W1 5196-5019 Veszprém, Hungary</td>
<td>W1c 2500-2050 Heimbuch, G.</td>
<td>W1c 2000-1700 Charalambos</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>8179-7613</td>
<td>Gajn Dareh, Iran</td>
<td>X2 5196-5019 Veszprém, Hungary</td>
<td>X2b 2500-2050 Heimbuch, G.</td>
<td>X2b 2000-1700 Charalambos</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Calculation of the average added levels and distance. The * indicates hypothetical haplogroups that are assumed to be present based on the known presence of successors. The > indicates the successor relation between two haplogroups.

<table>
<thead>
<tr>
<th>#</th>
<th>Mesolithic or Fertile Crescent</th>
<th>Ancient European closest to Minoan</th>
<th>Minoan</th>
<th>L(a, b, c)</th>
<th>Ln(b) + Ln(c) - 2Ln(P(b, c))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>H1</td>
<td>H1bm</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>H5</td>
<td>H5a1g</td>
<td>H5a1g</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>H</td>
<td>H7</td>
<td>H7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>J2a</td>
<td>J2a1</td>
<td>J2a1d</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>H13</td>
<td>H13a1a2c</td>
<td>H13a1a</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>HV</td>
<td>HV+16311</td>
<td>HV+16311</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>I</td>
<td>I5a1a</td>
<td>I5a1b</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>J2b1</td>
<td>J2b1a</td>
<td>J2b1a1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Kla2</td>
<td>Kla2</td>
<td>Kla2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>R0a</td>
<td>R0a1</td>
<td>R0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>T1a</td>
<td>T1a1</td>
<td>T1a4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>T2b</td>
<td>T2b</td>
<td>T2b</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>T2c</td>
<td>T2c1a2</td>
<td>T2c1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>T2</td>
<td>T2e</td>
<td>T2e6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>U3b</td>
<td>U3b1b</td>
<td>U3b3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>G2a2b2b</td>
<td>G2a2b2a</td>
<td>G2a2b2a</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>*U5a1 &gt; U5a1c</td>
<td>U5a1f2</td>
<td>U5a1f1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>U8b1b</td>
<td>U8b1b+16290</td>
<td>U8b1b+16290</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>*W1 &gt; W1+119</td>
<td>W1c1</td>
<td>W1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>X2</td>
<td>X2+16179</td>
<td>X2+16179</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

S1 = [a1, ..., a20], S2 = [b1, ..., b20], S3 = [c1, ..., c20], Al(S1, S2, S3) = 1
D(S2, S1) = 0.525

Fig. 3. Three examples of Neolithic Settlements in the Danube Basin: Malak Preslavets, Bulgaria (bottom, with 10 samples), Cotatcu, Romania (middle with the I2532 sample) and Verteba Cave, Ukraine (top, with 4 samples). This map was generated based on the amtDB database [8].
V. DISCUSSION OF THE RESULTS

When the Early European Farmers (EEFs) entered Europe from Anatolia, they were foremost looking for land near the rivers because a steady supply of fresh water was essential for farming. EEFs were likely less interested in exploring islands, because those rarely provide good agricultural areas or major rivers. The EEFs likely found the banks of the Danube and its tributaries preferred places to settle. Fig. 3 shows three Neolithic settlements in the Danube Basin are Malak Preslavets, Bulgaria; Cotacatu, Romania and Verteba Cave, Ukraine. All these settlements have similar genetic profiles, indicating a common source. In particular, none of them has the mtDNA haplogroup N1a1a1, which is found at the archaeological site Boncuklu but not at Çatalhöyük, Turkey. Hence these two Neolithic sites were inhabited by different groups of people. The Boncuklu group seems to have traveled westward and may be the originator of the Cardium Pottery culture on the northern Mediterranean shore, while the Çatalhöyük group migrated into the Danube Basin.

Fig. 4 shows a map of the hypothetical migration of EEFs. As indicated by the yellow arrow, farmers from northern Anatolia carrying at least the mtDNA and y-DNA genes shown in the yellow box moved along the Black Sea coast into the western Black Sea littoral and the Danube Basin. Almost all the data from the second to sixth columns in Table 2 come from the western Black Sea littoral or the Danube Basin. After thousands of years, the EEFs spread upstream along the Danube and its tributaries, reaching the Carpathian Basin and present southern Germany. From there they could easily reach the Elbe and spread northward.

During this process of spreading, the EEFs mixed with the local hunter-gatherer population that was carrying Mesolithic European genes as shown in the blue box. After thousands of years, there were many genetic mutations in the mixed population. These are characterized by the deep-clade classifications in the green box of Fig. 4.

During these millennia in the Danube Basin, the population learned the use of copper and bronze. They also likely developed greater shipping and fishing skills. Around five thousand years ago, something happened that prompted a large scale migration southward from the Danube Basin and the western Black Sea littoral towards the islands of the Aegean Sea and brought Minoan civilization to Crete as indicated by the green arrow in Fig. 4.

The migration may have been caused by climate change making agriculture harder and fishing more attractive. It may have been caused by overpopulation and consequent civil war as various groups were fighting each other for scarce resources. It may have been caused by a conflict between the Neolithic groups and new settlers into the area. Perhaps people were fleeing pandemic diseases that swept through the European continent.

VI. REVIEW OF PREVIOUS WORK

The story of migration revealed by achaeogenetics as shown in Fig. 4, fits well with other data. In particular, the following additional connections were noted earlier between the Neolithic and Bronze Age Danubian Basin and Minoan Crete:

Art Motifs: Revesz [31] shows a strong connection between the art motifs of EEFs, their Bronze Age successors and the Minoan culture, especially the Middle Minoan culture that existed between 2200 and 1500 BC according to Evans [9]. Moreover, many of the art motifs survived in Hungarian folk art.

Linguistics: The Minoan language is classified as a close relative of Hungarian with both belonging to the Ugric group of Finno-Ugric languages [25-30]. The classification is based on translations of dozens of Cretan Hieroglyph and Linear A texts from the Minoan culture. Archaeologists identified the Mezhovskaya culture as a common Bronze Age Ugric homeland [20]. The southwestern Ural Mountains’ Kapova Cave that is mentioned several times in the “Neolithic/BA Pontic Steppe” column of Table 2 also belongs to the Mezhovskaya culture.

Metallurgy: According to Childe [4], bronze metallurgy spread from the Danube Basin to Crete during the Middle Minoan period.

Potter’s Wheel: According to Haarmann [13], the potter’s wheel was also used in the Cucuteni culture that existed in the late Neolithic. The Cucuteni migrants may have introduced the potter’s wheel to Crete in the Middle Minoan period.

Writing Scripts: The Danubian script [13], the Linear A script [11, 19, 32], which was used by the Minoans during the Middle Minoan period, the Carian alphabet and the Old Hungarian alphabet have many similarities as shown by both neural networks [6, 7] and feature-based similarity analyses [22, 28].
mtDNA haplogroup I. Later, Neparáczki et al. [18] resampled the aforementioned Sárrétudvar remains and published a classification of it as I5a1a, as shown in Table 2.

The other classifications in the “Medieval Central Europe” column of Table 2 are based on Neparáczki et al. [17,18] and Csáky [5]. The rest of the data comes form the amtDB database [8].

VII. CONCLUSIONS AND FUTURE WORK

The archaeogenetic data gives the clearest evidence of the Danube Basin and western Black Sea littoral origin of the Minoans. This result gives a strong support to the earlier proposals of art motifs, linguistic, metallurgical, and writing script similarities between the Danube Basin and western Black Sea littoral and Minoan Crete.

The combination of these data points to a multifaceted relationship between the Minoans and Hungarians. While Hungarian belongs to the Ugric branch of the Finno-Ugric and Uralic languages, it is possible that Proto-Uralic people lived also somewhere on the western Black Sea littoral and the Pontic Steppe in ancient times. The Samoyedic and the Finno-Permic groups separated from that entity leaving the Ugric branch behind. Later the Ugric branch also dissolved with the Minoans migrating southward to Crete. This possible scenario needs to be further explored in the future.

ACKNOWLEDGEMENT

The author would like to thank the J. William Fulbright Program for supporting him on two Fulbright Scholarships. First, on a visit to the University of Athens, Greece, in 2008, and second, to Budapest, Hungary, in spring 2017 to the Aquincum Institute of Technology, an affiliate of the Budapest University of Technology and Economics. The author would like to thank the referees for helpful comments.

REFERENCES


Peter Z. Revesz (Ph.D. ’91) holds a Ph.D. degree in Computer Science from Brown University and was a postdoctoral fellow at the University of Toronto. He is an expert in bioinformatics, computational linguistics, databases and data mining. He is the author of Introduction to Databases: From Biological to Spatio-Temporal (Springer, 2010) and Introduction to Constraint Databases (Springer, 2002). He is currently a professor in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln, Lincoln, NE 6815, USA.

Dr. Revesz held visiting appointments at IBM T. J. Watson Research Center, INRIA, Max Planck Institute for Computer Science, University of Athens, University of Hasselt, University of Helsinki, U.S. Air Force Office of Scientific Research and U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright
Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award.

APPENDIX

Fig. 5. Details for the reclassifications of the Minoan mtDNA haplogroups.
(f) W to W1 (W² is unlikely given the Anatolian and Old European presence of W1.)

(g) X to X²+16179 (This may be a predecessor of X²j.)

Fig. 6. Further details for the reclassifications of the Minoan mtDNA haplogroups.