Minoan Archaeogenetic Data Mining Reveals Danube Basin and Western Black Sea Littoral Origin

Peter Z. Revesz

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 archaeogentic 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 amtDB. 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.

Peter Z. Revesz is a professor in the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588 USA (phone: 402-472-3488; fax: 402-472-7767; email: revesz@cse.unl.edu). A preliminary version of this research was presented at the conference [24].

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

16056 ccaagtattg actcacccat caacaaccgc tatgtatttc gtacattact gccagccacc
16116 atgaatattg tacagtacca taaatacttg actacctgta gtacataaaa acccaatcca
16176 catcaaaacc ccctccccat gcttacaagc aagtacagca atcaaccttc aactatcaca
16236 catcaactgc aactccaaag ccacccctca cccactagga taccaacaaa cctacccacc
16296 cttaacagta catagtacat aaagccattt accgtacata gcacattaca gtcaaatccc
16356 ttctcgtccc
```

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

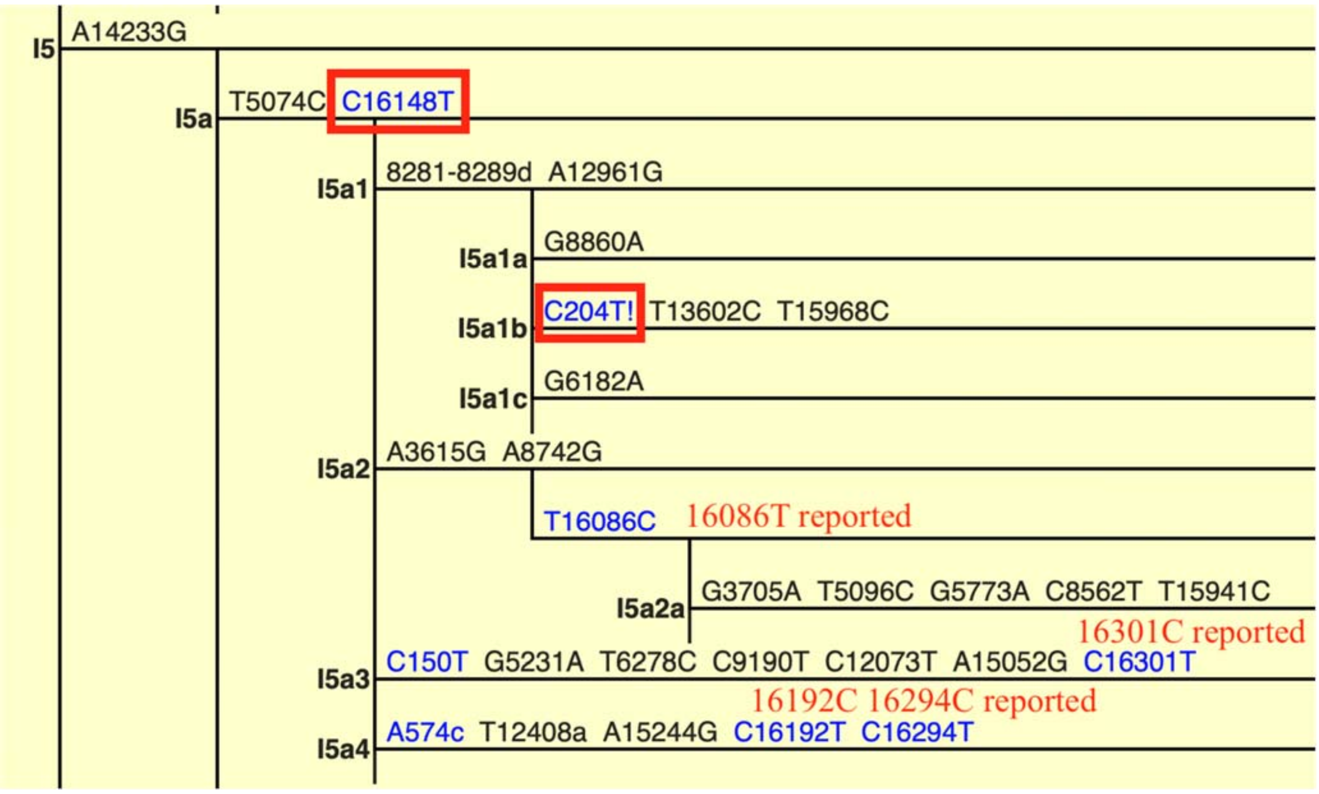


Fig. 2. The reclassification of Minoan sample M13 as mtDNA haplogroup *I5a1b*.

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] Csáky 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.

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].

ENA ID	**Alias ID**	**mtDNA or y-DNA**		**Reported mutations with respect to rCRS**	
		Reported	**Corrected**	**< 16000**	**16000 + *n***
HM022275	6AH	H			
HM022276	7AH	H			
HM022285	16AH	H			
HM022286	17AH	H			
HM022288	19AH	H			
HM022302	20H	H			
HM022293	7H	H			
HM022294	8H	H			
HM022299	16H	H			
HM022314	M11	H			
ERS1770862	I0073	H1bm			
		J2a1			
ERS1770864	I9005	H			
ERS1770869	I9129	H+163			
HM022278	9AH	H5	H5a1g	2706A	304C
ERS1770863	I0074	H5			
HM022281	12AH	H7			
ERS1770861	I0070	H13a1			
		J2a1d			
HM022271	2AH	H13a1a			
HM022310	M6	HV	HV+16311	2706G, 7028T, 11719A, 14766C	311C
HM022304	24H	HV			
HM022270	1AH	HV			
HM022316	M13	I5	I5a1b	199T, 250C	129A, 148T, 223T, 391A
HM022318	M17	I5			
HM022315	M12	I5			
ERS1770868	I9128	I5			
HM022305	26H	J2			
ERS1770867	I9127	J2b1a1			
HM022322	M22	K			
HM022317	M14	K			
HM022321	M21	K			
HM022284	15AH	K			
ERS1770871	I9131	K1a2			
HM022311	M7	R0			
HM022287	18AH	T			
HM022295	9H	T1	T1a4	10463C, 15928A	126C, 163G, 186T, 189C, 263C, 294T
HM022319	M19	T2	T2b	13368A	089A, 126C, 294T, 296T, 304C
HM022272	3AH	T2	T2b	4917G, 10463C	126C, 294T, 296T, 304C
HM022323	M23	T3	T2c1	4917G, 11251G, 15607G, 15928A	126C, 292T, 294T, 339T
HM022303	21H	T5	T2e6	10463C, 13368A, 15607G, 15298A	126C, 153A, 240C, 294T, 296T
ERS1770870	I9130	U3b3			
		G2a2b2a			
HM022291	4H	U5a	U5a1f1	12308G, 12372A	192T, 256T, 270T, 311C
ERS1770860	I0071	U5a1			
-	M4	U	U8b1b+16290		182C, 183C, 189C, 234T, 290T, 324C
HM022292	6H	W	W1	204C, 207A, 15884C	093C, 223T, 243C, 292T, 243C
HM022312	M8	X	X2+16179	153G, 14470C	145A, 179T, 189C, 223T, 278T, 332T, 344T, 362C

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 *8281-8289d* indicates nine deletions, the extra number of mutations from *I* to *I5a1* 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 $S_1[i]$ is a common precedent of $S_2[i]$ and $S_3[i]$ for any $1 \leq i \leq n$. Then the *average added levels* function $Al(S_1, S_2, S_3)$ is defined as follows:

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

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

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

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:

$$Al(S_1, S_2, S_3) = \frac{L(I, I5a1a, I5a1b)}{1} = 3$$

and

$$Am(S_1, S_2, S_3) = \frac{M(I, I5a1a, I5a1b)}{1} = 13$$

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:

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

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

$$D(A,B) = \frac{Ln(I5a1a) + Ln(I5a1b) - 2Ln(I5a1)}{2} = 1$$

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]:

$$D(A,B) \geq 0 \qquad (3)$$
$$D(A,B) = 0 \leftrightarrow A = B \qquad (4)$$
$$D(A,B) = D(B,A) \qquad (5)$$
$$D(A,B) \leq D(A,C) + D(C,B) \qquad (6)$$

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:

$$Ln(a) + Ln(b) - 2Ln(P(a,b)) \geq 0$$

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:

$$Ln(a_i) + Ln(b_i) - 2Ln(P(a_i, b_i)) = 0 \qquad (7)$$

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:

$$Ln(de) + Ln(df) - 2Ln(d) = Ln(e) + Ln(f)$$

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:

$$\begin{aligned} Ln(a) + Ln(b) - 2Ln(P(a,b)) \\ \leq (Ln(a) + Ln(c) - 2Ln(P(a,c))) \\ + (Ln(c) + Ln(b) - 2Ln(P(c,b))) \end{aligned}$$

The above can be simplified as follows:

$$Ln\big(P(a,c)\big) + Ln(P(c,b)) \leq Ln(P(a,b)) + Ln(c) \quad (8)$$

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:

$$a = def$$
$$b = dg$$
$$c = deh$$

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:

$$Ln(d) + Ln(e) \leq Ln(d) + Ln(d) + Ln(e) + Ln(h)$$

Simplifying we get:

$$0 \leq Ln(d) + Ln(h)$$

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:

$$d(A,B) = \frac{1}{2n}\sum_{i=1}^{n} Mut(a_i) + Mut(b_i) - 2Mut(P(a_i,b_i))$$

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

$$Mut(a) \geq Mut(P(a, b))$$

By symmetry we also have:

$$Mut(b) \geq Mut(P(a, b))$$

The above two inequalities imply that:

$$Mut(a) + Mut(b) - 2\,Mut(P(a, b)) \geq 0 \quad (9)$$

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:

$$\begin{aligned} Mut(a) + Mut(b) - 2Mut(P(a,b)) &\\ \leq (Mut(a) + Mut(c) - 2Mut(P(a,c))) &\\ + (Mut(c) + Mut(b) - 2Mut(P(c,b))) & \end{aligned}$$

Simplifying the above, we get the following:

$$\begin{aligned} Mut(P(a,c)) + Mut(P(c,b)) &\\ \leq Mut(P(a,b)) + Mut(c) \quad (10) & \end{aligned}$$

We argue by two mutually exclusive cases.

Case I: $Ln(P(a,c)) < Ln(P(a,b))$. In this case $Mut(P(a,c)) < Mut(P(a,b))$ and $Mut(P(c,b)) \leq Mut(c)$ both trivially hold. Hence Equation (10) is 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 = def, b = dg,* and *c = deh* in terms of concatenation of some possibly empty substrings *d, e, f, g*, and *h* where $Ln(P(e,g)) = 0$. Then Equation (10) can be written as follows:

$$Mut(de) \leq Mut(deh)$$

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_1 and S_3 to be the haplogroups in the first and last columns of Table 2, respectively. We also took S_2 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_2 into separate groups of European time periods from Neolithic, Corded 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 2. Mesolithic (blue) and Fertile Crescent (yellow) origins of Minoan mtDNA (black) and y-DNA (red) haplogroups.

Mesolithic or Fertile Crescent	**Neolithic Old Europe**	**Corded Ware Cult.**	**Copper/BA Cent. Europe**	**Medieval Cent. Europe**	**Neolithic/BA Pontic Steppe**	**Minoan Crete**
H 6450-6380 Çatalhöyük Turkey	H1 5976-5751 Zemunica Cave Croatia	H1e 2800-2300 Hubinek Poland	H1b1 2500-2200 Szigetszentmiklós, Hungary	H1b2 895-950 Karos Hungary	H1 4949-4799 Dereivka Ukraine	H1bm 2000-1700 Charalambos Greece
H5 6500-6200 Barcin, Turkey	H5a2 5800-5400 Malak Preslavets, Bu	H5a1 2625-2400 Jäbara, Estonia	H5a1g 2500-2200 Prague, Czech R.	H5a2 7th cent. Szarvas, Hu.	H5a 3931-3640 Verteba C., U.	H5a1g 2000-1700 Charalambos
H 6450-6380 Çatalhöyük, Turkey	H7c 5641-5560 Kargadur, Cr.	H7d 2900-2200 Brandýsek, CR.	H7 4455-4359 Yunatsite, Bul.	H7c 8th cent. Cífer-Pác, SK		H7 2000-1700 Charalambos
J2a 7940-7600 Kotias Klde, Georgia	J2a1 4692-4546 Osijek, Croatia		J2a1 1270-1110 Ludas-Varjú, H.	J2a1a 950-1000 Sárrétudvar, H.		J2a1d 2000-1700 Charalambos
H13 8300-7400 Lepenski Vir, Serbia			H13a1a2c 2457-2142 Quedlinburg, G.	H13a1a2 412-604 Szólád, Hu.	H13a1a1 3300-2700 Samara, Rus.	H13a1a 2000-1700 Charalambos
HV 5836-5632 Lepenski Vir, Serbia	HV+16311 5500-5300 Mezőkövesd, H.	HV0e Saxony-Anhalt, Ger.	HV+16311 3600-3000 Budakalász, Hu.	HV+16311 975-1025, Fajd-Jegeshegy	HV 3758-3636 Verteba C., U.	HV+16311 2000-1700 Charalambos
I 6900-6800 'Ain Ghazal, Jordan				I5a1a 950-1000 Sárrétudvar, H.	I5a2 or I5c 1598-1398 Kapova C., Ru.	I5a1b 2000-1700 Charalambos
J2b1 6205-5907, Lepenski Vir, S.	J2b1 5800-5400 Malak Pr., Bu	J2b1a Switzerland	J2b1a 2136-1977 Königsbrunn, G..	J2b1 412-604 Szólád, Hu.	J2b1a 1598-1398 Kapova, Rus.	J2b1a1 2210-1680 Odigitria, Gr.
K1a2 6400-5600 Mentese, Turkey	K1a2a 4230-3995 Kletios, Gr.	K1a2a 2621-2472 Viby, SE	K1a2a 4452-4354 Dzhulyunitsa, B.	K1a 500-600 Szólád, Hu.	K1a2 5715-5626 Cotatcu, Ro.	K1a2 2210-1680 Odigitria, Gr
R0a 7722-7541 'Ain Ghazal, Jordan					R0a1 3095-2915 Ozera, Ukraine	R0 2000-1700 Charalambos
T1a 7446-7058 'Ain Ghazal, Jordan	T1a 5500-3700 Diros, Greece	T1a1 2570-2471 Karsdorf, G.	T1a1 2000-1500 Érd, Hungary	T1a1 895-950 Karos, Hu.	T1a1 3300-2700 Shevchenko, U	T1a4 2000-1700 Charalambos
T2b 6500-6200 Barcin Turkey	T2b 5800-5400 Malak Preslavets, Bu	T2b Switzerland	T2b 2457-2201 Szigetszentmiklós, Hungary	T2b 950-1000 Sárrétudvar Hungary	T2b 3911-3659 Verteba Cave Ukraine	T2b 2000-1700 Charalambos Greece
T2c 8204-7755 T. Abdul Hosein Iran		T2c Saxony-Anhalt, Germany	T2c1d+152 2500-2200 Szigetszentmiklós, Hungary	T2c1c 500-600 Fonyód, Hungary	T2c1a2 3339-2918 Samara Russia	T2c1 2000-1700 Charalambos Greece
T2 6500-4000 Ukraine	T2e 5800-5400 Malak Pr., Bu	T2e 2800-2300 Malżyce, Pl.	T2e Brandýsek Czech Rep.	T2e 412-604 Szólád, Hu.	T2e 2890-2696 Dereivka, Uk	T2e6 2000-1700 Charalambos
U3b 6450-6380 G2a2b2b 8300-7952 Çatalhöyük, Boncuklu	G2a2b2a 5800-5400 Malak Preslavets, Bu		G2a2b2a 3910-3650 Abony, Hu.	U3b1b G2a2b 895-950 Karos, Hu.	G2a2b2a 3931-3640 Verteba C., Uk.	U3b3 G2a2b2a 2210-1680 Odigitria, Gr.
U5a1c 9140-8570 Padina, Serbia	U5a1a2 5207-4945 Osijek, Croatia	U5a1 2500-2050 Esperstedt, G.	U5a1f2 2050-1800 Quedlinburg	U5a1a2a 895-950 Karos, Hu.	U5a1f2 1850-1200 Samara, Rus	U5a1f1 2000-1700 Charalambos
U8b1b 7940-7585 Vlasac, Serbia	U8b1b+16290 5300-4900 Hejőkürt, Hu.				U8b1b 4000-3600 Verteba Cave	U8b1b+16290 2000-1700 Charalambos
W1+119 6500-6200 Barcin, Turkey	W1+119 Veszprém, Hungary		W1c1 2500-2050 Heimburg, G.			W1 2000-1700 Charalambos
X2 8179-7613 Ganj Dareh, Iran	X2+16179 4625-4250 Oberweiderstedt, G.	X2b4 2464-2210 Esperstedt, G.		X2l 895-950 Karos, Hu.		X2+16179 2000-1700 Charalambos

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.

#	**Mesolithic or Fertile Crescent a**	**Ancient European closest to Minoan b**	**Minoan c**	**$L(a_i, b_i, c_i)$**	**$Ln(b_i) + Ln(c_i) - 2Ln(P(b_i, c_i))$**
1	H	H1	H1bm	1	1
2	H5	H5a1g	H5a1g	3	0
3	H	H7	H7	1	0
4	J2a	J2a1	J2a1d	1	1
5	H13	H13a1a2c	H13a1a	3	2
6	HV	HV+16311	HV+16311	1	0
7	I	I5a1a	I5a1b	3	2
8	J2b1	J2b1a	J2b1a1	1	1
9	K1a2	K1a2	K1a2	0	0
10	R0a	R0a1	R0	0	2
11	T1a	T1a1	T1a4	0	2
12	T2b	T2b	T2b	0	0
13	T2c	T2c1a2	T2c1	1	2
14	T2	T2e	T2e6	1	1
15	U3b	U3b1b	U3b3	0	3
16	G2a2b2b	G2a2b2a	G2a2b2a	1	0
17	*U5a1 > U5a1c	U5a1f2	U5a1f1	1	2
18	U8b1b	U8b1b+16290	U8b1b+16290	1	0
19	*W1 > W1+119	W1c1	W1	0	2
20	X2	X2+16179	X2+16179	1	0
	$S_1 = [a_1, \ldots, a_{20}]$	$S_2 = [b_1, \ldots, b_{20}]$	$S_3 = [c_1, \ldots, c_{20}]$	$Al(S_1, S_2, S_3) = 1$	$D(S_2, S_3) = 0.525$

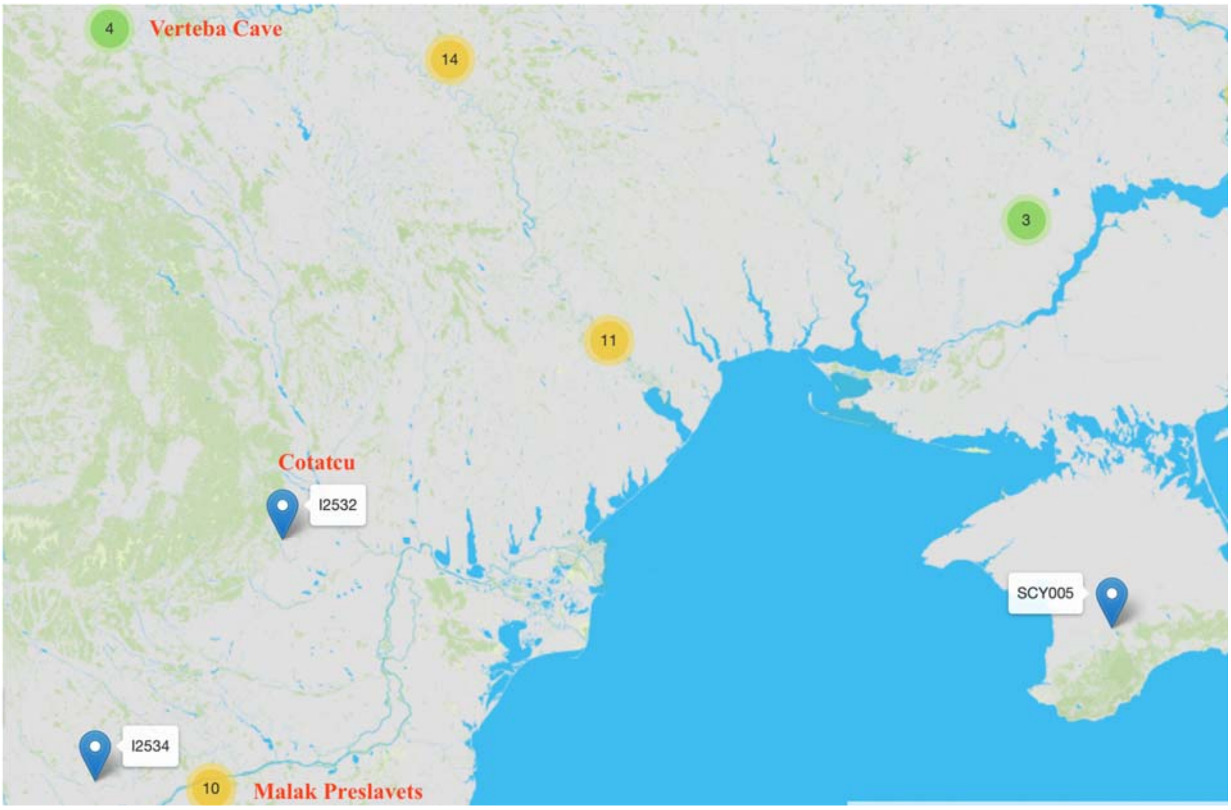


Fig. 3. Three examples of Neolithic Settlements in the Danube Basin: Malak Preslavets, Bulgaria (bottom, with 10 samples), Cotatcu, Romania (middle with the I2532 sample) and Verteba Cave, Ukraine (top, with 4 samples). This map was generated based on the amtDB database [8].

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, Cotatcu, 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 upriver 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.

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].

VI. REVIEW OF PREVIOUS WORK

The two previous archaeogenetic studies that published Minoan data from the Hagios Charalambos Cave (2000-1700 BC) and the Moni Odigitria Cave (2210-1680 BC) in Crete concluded that the Minoans came from the east and are closely related to the Western Anatolian Neolithic and Bronze Age populations [14, 15].

We have seen that the relationship with Anatolia is only part of the story of Minoan origins because the Minoan population also contains many other genes that are of Mesolithic European origin.

Revesz [23] already called attention to several mtDNA relationships among Minoans and other ancient populations, such as the Andronovo, Bell Beaker, Rössen and Únětice cultures. In addition, Revesz [24] identified the mtDNA *I5a* relationship between Minoans and Hungarians at Sárrétudvar. That needed a reclassification of the Hungarian sample to *I5a* from Tömöry et al. [34]'s earlier classification as simply

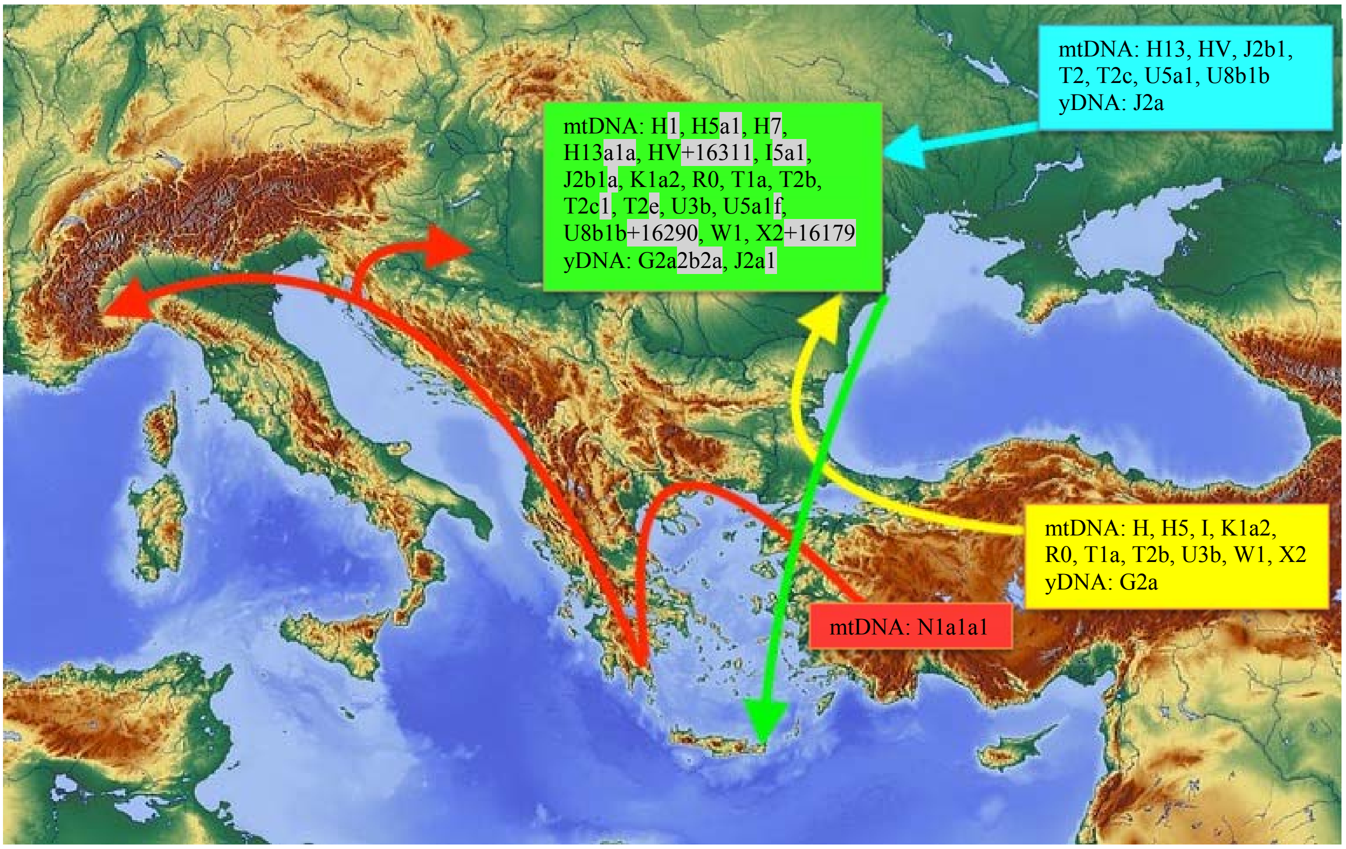


Fig. 4. The above map shows the spread of mtDNAs and y-DNAs from northern Anatolia (yellow) and southern Anatolia (red) into different parts of Europe. The first mixed with native European mtDNAs and y-DNAs (blue) on the western littoral of the Black Sea and in settlements along the Danube and its tributaries. The mixed set of genes underwent considerable mutations (gray) before the Proto-(Middle)-Minoan people migrated to Crete (green).

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 archaeogentic 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

[1] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. Della Casa, P. Dąbrowski, P. R. Duffy, A. V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolár, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtchyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, Ł. Pospieszny, T. D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V. I. Soenov, V. Szeverényi, G. Tóth, S. V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S.

Brunak, R. Nielsen, K. Kristiansen, and E. Willerslev, Population genomics of Bronze Age Eurasia, *Nature***,** Vol. 522, 167–172, 2015.

[2] M. Bernal, *Black Athena: The Afroasiatic Roots of Classical Civilization: The Linguistic Evidence*, Vols. III, Rutgers University Press, 2006.

[3] G. Campbell-Dunn, *Who were the Minoans?: An African Answer*, BookWhirl Publishing: Green Bay, WI, 2014.

[4] V. C. Childe, The East-European relations of the Dimini Culture, The Journal of Hellenic Studies, Vol. 42, No. 2, 254-275, 1922.

[5] V. Csáky, D. Gerber, I. Koncz, G. Csiky, B. G. Mende, A. Marcsik, E. Molnár, G. Pálfy, A. Gulyás, B. Kovacsóczy, G. M. Lezsák, G. Lőrinzcy, A. Szécsényi-Nagy, T. Vida, Inner Asian maternal genetic origin of the Avar period nomadic elite in the 7th century AD Carpathian Basin, *bioRxiv*, Avail: Sept. 13, 2018.

[6] S. Daggumati, P. Z. Revesz, Data mining ancient script image data using convolutional neural networks, *Proc. 22nd International Database Engineering and Applications Symposium,* New York, ACM Press, 2018, pp. 267-272.

[7] S. Daggumati and P. Z. Revesz, Data mining ancient scripts to investigate their relationships and origins, *Proceedings 23rd International Database Engineering and Applications Symposium*, New York, ACM Press, 2019, pp. 209-218.

[8] E. Ehler, J. Novotný, A. Juras, M. Chyleński, O. Moravčík, and J. Pačes, AmtDB: A database of ancient human mitochondrial genomes, *Nucleic Acids Research*, vol. 47, issue D1, p. D29–D32, 2019.

[9] A. Evans, *The Palace of Minos at Knossos,* Vols. I-IV*,* London: MacMillan and Co., 1921-1935.

[10] M. Gimbutas, *The Language of the Goddess*, Thames & Hudson Inc., New York, NY, 1989.

[11] L. Godart and J.-P. Olivier, *Recueil des inscriptions en Linéaire A* (*Études Crétoises 21)*, De Boccard, 1976.

[12] C. H. Gordon, *Evidence for the Minoan Language*, Ventnor Publ., Ventnor, NJ, 1966.

[13] H. Haarmann, *Roots of Ancient Greek Civilization: The Influence of Old Europe*, McFarland, 2014.

[14] J. R. Hughey, P. Paschou, P. Drineas, D. Mastropaolo, D. M. Lotakis, P. A. Navas, M. Michalodimitrakis, J. A. Stamatoyannopoulos, and G. Stamatoyannopoulos, A European population in Minoan Bronze Age Crete, *Nature Communications*, Vol. 4, 1861, 2013.

[15] J. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwängler, A. Peltzer, C. Posth, A. Vasilakis, P. J. P. McGeorge, E. Konsolaki-Yannopoulou, G. Korres, H. Martlew, M. Michalodimitrakis, M. Özsait, N. Özsait, A. Papathanasiou, M. Richards, S. A. Roodenberg, Y. Tzedakis, R. Arnott, D. M. Fernandes, J. R. Hughey, D. M. Lotakis, P. A. Navas, Y. Maniatis, J. A. Stamatoyannopoulos, K. Stewardson, P. Stockhammer, R. Pinhasi, D. Reich, J. Krause, G. Stamatoyannopoulos, Genetic origins of the Minoans and Mycenaeans**,** *Nature*, Vol. 548, pp. 214-218, 2017.

[16] N. Marinatos, *Minoan Kingship and the Solar Goddess: A Near Eastern Koine*, University of Illinois Press, 2010.

[17] E. Neparáczki, Z. Maróti, T. Kalmár, K. Maár, I. Nagy, D. Latinovics, Á. Kustár, G. Pálfi, E. Molnár, A. Marcsik, C. Balogh, G. Lőrinczy, S. S. Gál, P. Tomka, B. Kovacsóczy, L. Kovács, I. Raskó, and T. Török, Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin, *Nature Scientific Reports***,** Vol. 9, Article No. 16569, 2019.

[18] E. Neparáczki, Z. Juhász, H. Pamjav, T. Fehér, B. Csányi, A. Zink, F. Maixner, G. Pálfi, E. Molnár, I. Pap, Á. Kustár, L. Révész, I. Raskó, and T. Török, Genetic structure of the early Hungarian conquerors inferred from mtDNA haplotypes and Y-chromosome haplogroups in a small cemetery, *Molecular Genetics and Genomics*, Vol. 292, p. 201-214, 2017.

[19] J.-P. Olivier, Cretan writing in the second millennium B.C., *World Archaeology*, Vol. 17, no. 3, 1986, pp. 377–389.

[20] В. Т Петрин, Т. И. Нохрина, А. Ф. Шорин, Археологические памятники Агразинского водохранилища (эпохи камня и бронзы). [Archaeological sites of the Agrazinsky reservoir (Neolithic and Bronze Age)], Novosibirsk, 1993.

[21] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer: New York, 2010.

[22] P. Z. Revesz, Bioinformatics evolutionary tree algorithms reveal the history of the Cretan Script Family, *International Journal of Applied Mathematics and Informatics*, Vol. 10, No. 1, 2016, pp. 67-76. Available: http://www.naun.org/main/UPress/ami/2016/a182013-133.pdf

[23] P. Z. Revesz, A mitochondrial DNA-based model of the spread of human populations, *International Journal of Biology and Biomedical Engineering*, Vol. 10, p. 124-133, 2016.

[24] P. Z. Revesz, A spatio-temporal analysis of mitochondrial DNA haplogroup I, *Proc. 20th Int. Conf. on Circuits, Systems, Comm. and Computers – MATEC Web of Conferences*, Vol. 76, No. 04048, 2016.

[25] P. Z. Revesz, A computer-aided translation of the Phaistos Disk, *International Journal of Computers*, Vol. 10, No. 1, 2016, pp. 94-100. Available:http://www.naun.org/main/NAUN/computers/2016/a282001-455.pdf

[26] P. Z. Revesz, A computer-aided translation of the Cretan Hieroglyph Script, *International Journal of Signal Processing*, Vol. 1, No. 1, 2016, pp. 127-133. Available: iaras.org/iaras/filedownloads/ijsp/2016/003-0017.pdf

[27] P. Z. Revesz, A translation of the Arkalochori Axe and the Malia Altar Stone, *WSEAS Trans. on Information Science and Applications*, Vol. 14, No. 1, 2017, pp. 124-133. Available: http://www.wseas.org/multimedia/journals/information/2017/a285909-084.pdf

[28] P. Z. Revesz, Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A, *WSEAS Transactions on Information Science and Applications*, Vol. 14, No. 1, 2017, 306-335. Available: http://www.wseas.org/multimedia/journals/information/2017/a605909-068.pdf

[29] P. Z. Revesz, Spatio-temporal data mining of major European river and mountain names reveals their Near Eastern and African origins, *Proceedings 23rd International Database Engineering and Applications Symposium*, New York, ACM Press, 2018, pp. 20-32. Available: https://cse.unl.edu/~revesz/papers/ADBIS18.pdf

[30] P. Z. Revesz, Sumerian contains Dravidian and Uralic substrates associated with the Emegir and Emesal dialects, *WSEAS Transactions on Information Science and Applications*, Vol. 16, No. 1, pp. 8-30, 2019. Available: http://www.wseas.org/multimedia/journals/information/2019/a045109-930.pdf

[31] P. Z. Revesz, Art motif similarity measure analysis: Fertile Crescent, Old European, Scythian and Hungarian elements in Minoan culture, *WSEAS Transactions on Mathematics*, Vol. 18, No. 1, p. 264-287, 2019.

[32] P. Z. Revesz, M. P. Rashid, and Y. Tuyishime, The design and implementation of AIDA: Ancient Inscription Database and Analytics system, *Proceedings 23rd International Database Engineering and Applications Symposium*, New York, ACM Press, 2019, pp. 292-297. Available: https://dl.acm.org/citation.cfm?id=3331117

[33] L. Šebest, M. Baldovič, A. Frtús, C. Bognár, K. Kyselicová, L. Kádasi, R. Beňuš, Detection of mitochondrial haplogroups in a small Avar-Slavic population from the eigth–ninth century AD, *Am. J. of Physical Anthropology*, pp. 1-8, 2018. DOI: 10.1002/ajpa.23380

[34] G. Tömöry, B. Csányi, E. Bogácsi-Szabó, T. Kalmár, Á. Czibula, A. Csősz, K. Priskin, B. Mende, P. Langó, C.S. Downes, and I. Raskó, Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. *American journal of physical anthropology*, Vol. 134, No. 3, pp. 354-368, 2007.

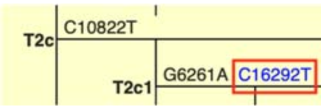
Peter Z. Revesz (Ph.D.'91) holds a Ph.D. degree in Computer Science from Brown University and was a postdoctoral fellow at the University of Toronto.

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

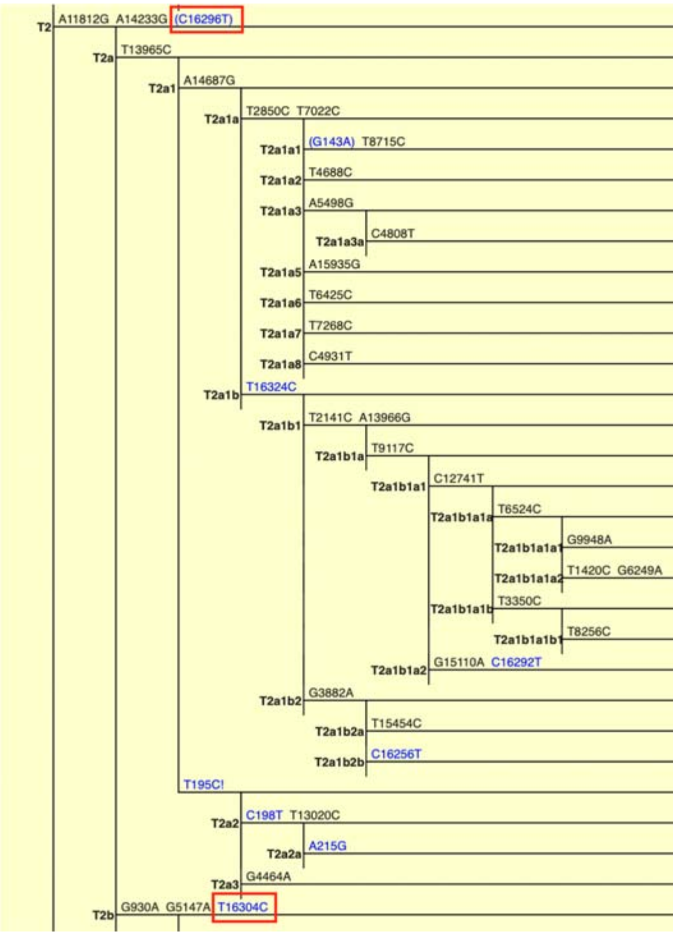
Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award.

APPENDIX

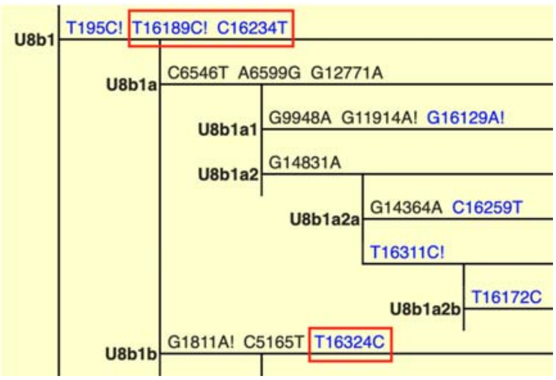


(b) T3 to T2c1

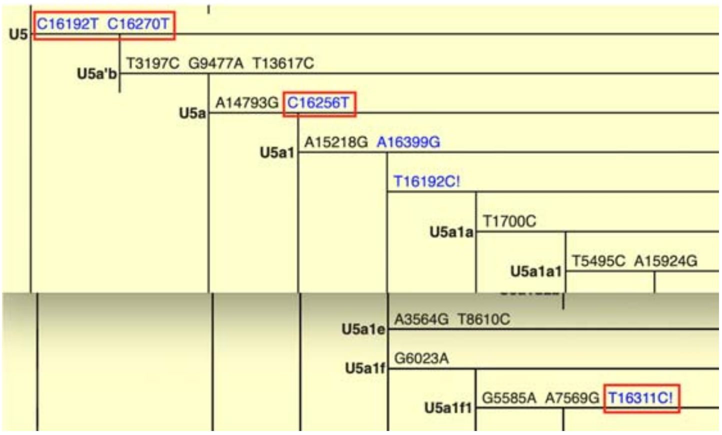
(a) T1 to T1a4



(c) T2 to T2b

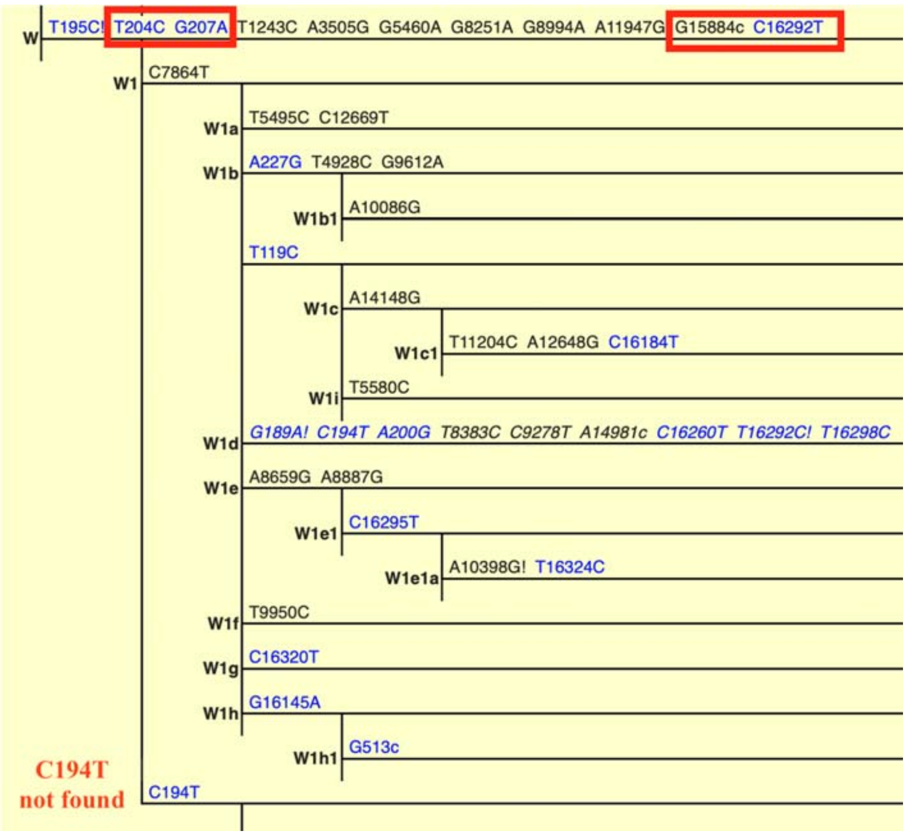


(d) U to U8b1b + 16290

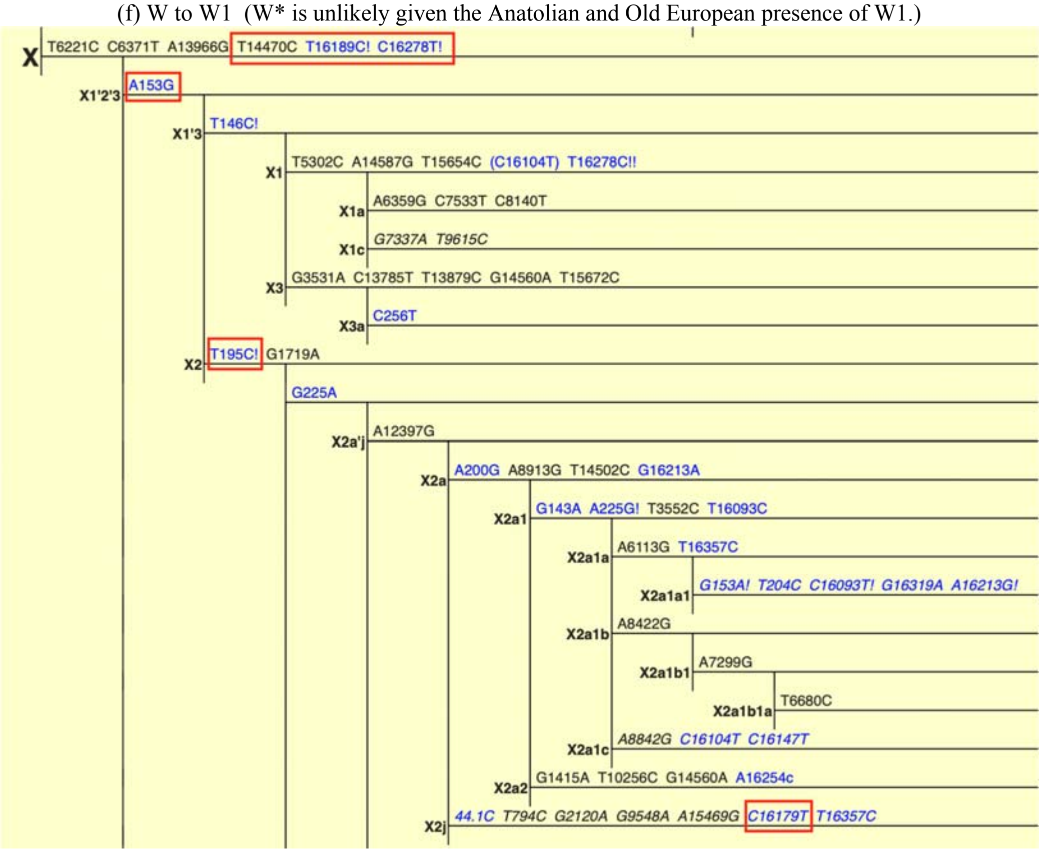


(e) U5a to U5a1f1

Fig. 5. Details for the reclassifications of the Minoan mtDNA haplogroups.



(f) W to W1 (W* 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.