Journal: www.jogg.info
Originally Published: Volume 7, Number 1 (Fall 2011)
Reference Number: 71.003

ON THE STRUCTURE AND AGE OF MTDNA HAPLOGROUP JT – A PHYLOGENETIC TOUR

Author(s): J.J. Logan

On the Structure and Age of Haplogroup JT – A Phylogenetic Tour

J. J. (Jim) Logan

Abstract

A search of GenBank data produced 515 genotypes that satisfied mtDNA Haplogroup JT search criteria. This data was supplemented by data from an FTDNA project dedicated to Haplogroup T research, their markers were aligned, and then the markers and genotypes were organized to model a JT phylogeny. This process was guided by published phylogenies for J and T and then refined using a top down maximum parsimony approach and each of the genotypes was given a subclade designation. Distribution of pairwise differences within and between the sets of genotypes were analyzed to identify general patterns for the major clades. Finally, statistical properties of these distributions were used to develop an age tree for Haplogroup JT. The resulting tree indicates the relative sequence of the evolution of clades of Haplogroup JT and an order of magnitude estimate of ages.

Introduction

The majority of Haplogroup JT genotypes are found in European populations and thence today in the New World. It is generally accepted that these genotypes descended from a small group of early Homo sapiens that left Africa about 50-75 thousand years ago. (Modern European DNA from both the Y chromosome and the mitochondria show significant differences from the corresponding DNA drawn from modern African subjects.) Furthermore, as the small population grew and spread across the Eurasian continent the DNA of subpopulations became both diverse and intermixed. Recent research of modern mitochondrial DNA (mtDNA) has revealed patterns of diversity that are now being used to infer evolutionary patterns and give them names. Furthermore, based in part on where these patterns are found, very general inferences are being made about population migrations that occurred thousands and tens of thousands of years ago. One such hypothetical migration map is shown in Figure 1.

The focus of this paper is on the twin evolutionary patterns associated with Haplogroups J and T. These haplogroup are thought to have a common origin in the Near East, perhaps as much as 50 thousand years ago. Both Haplotypes J and T are now found throughout Europe but migration patterns that led to this dispersal are not well understood.

Using RFLP (restriction fragment length polymorphism) techniques, the first of these, Haplogroup J (Hg J) was isolated from several other haplogroups of European descent in 1994 (Torroni, et al, 1994). Using the same technique, Hg T was identified two years later (Richards, et al, 1996: Torroni, et al, 1996). Sequencing techniques had previously been developed and their use in developing branching patterns had already been demonstrated by Di Rienzo and Wilson (1991) using mtDNA data primarily from Sardinia and the Middle East. A retrospective analysis of the data presented in that paper shows that of the 88 sequences, ten were Hg J and eleven were Hg T. Once recognized, various researchers found Hg J and Hg T data among their

samples and began to specify classification motifs for the patterns observed. Initial classifications were made on the basis of limited RFLP techniques, but supplemented by sequencing of the mtDNA from a single hypervariable region (HVR1). A University of Tartu masters theses written by Piia Serk (2004) compiled existing data on about Hg J and developed a phylogeny and estimates of coalescence ages of the major clades. A synthesis for Hg T was published by Pike (2006).

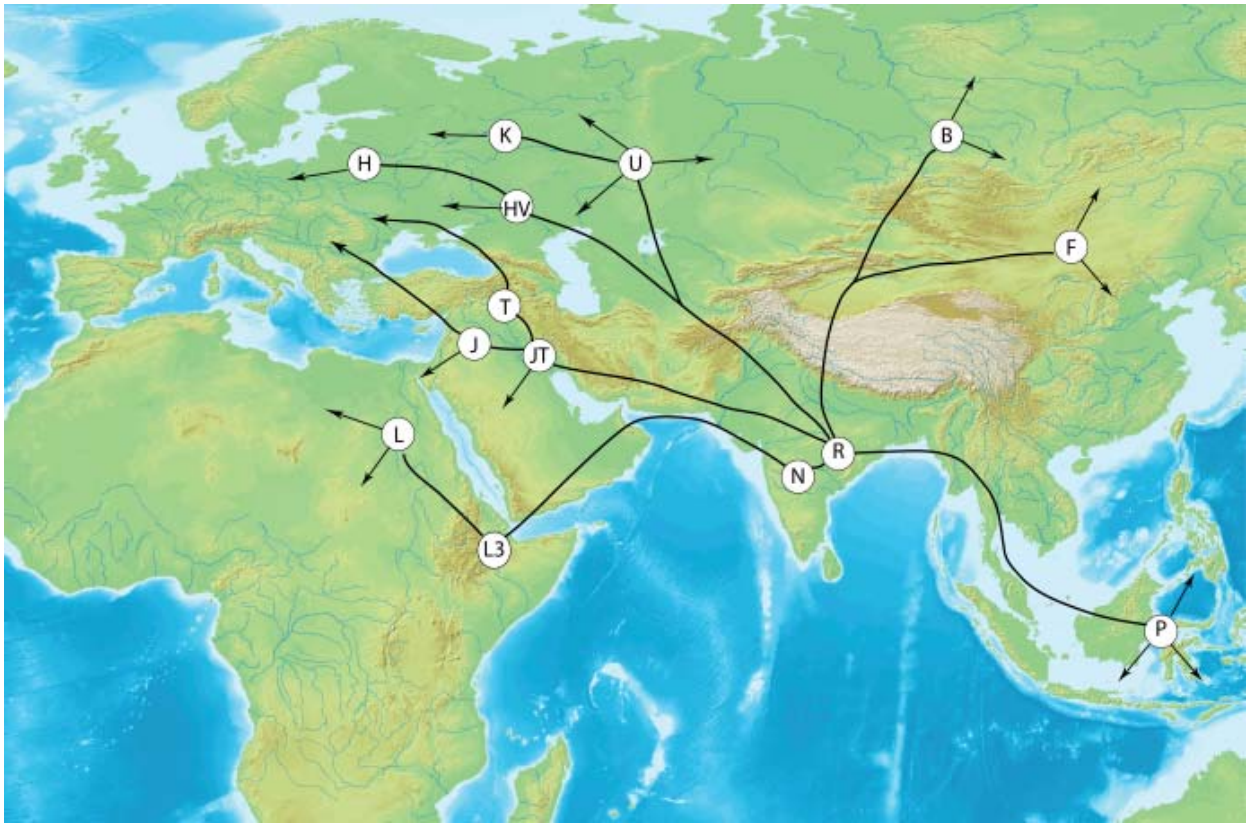

Figure 1. Hypothetical map of origins evolutionary diversity in European mitochondrial DNA. (Source: Logan 2010)

Once economic feasibility of full mitochondrial sequences (FMS) permitted more detailed analysis, it was found that neither the Hg J nor Hg T branches could be accurately specified using only hypervariable results. Logan (2008a) described this situation for Hg J. A more comprehensive analysis of Hg J, including a first look at the age of the clades, was presented by Logan (2008b) and an update of the J phylogeny was later presented by Logan (2009). A new development of the T phylogeny was recently presented. (Pike, et al, 2010).

PhyloTree (van Oven and Kayser, 2009) is the current de facto reference for a synthesis of the classification motifs and relationships for entire mitochondrial genetic tree as reported in the literature. Unfortunately, many of the relationships reported there were established using early RFLP and HVR1 results without realization that such limited data were not adequate for the purpose. For example, early research were misled when then did not realize that the same version of a polymorphism may occur in multiple clades that do not share a common descent; i.e., there multiple homoplasies occurring in the hypervariable region. These researchers used

markers at 16145 and 16261 as indicators for a clade that was called J1a until Palanichamy, et al (2004), demonstrated that the clade should actually be classified within the J2 branch of the mtDNA tree and it was renamed J2a. Starting with Build 3, PhyloTree has incorporated the revised relationships for Haplogroup J largely based on Logan (2008b) and subsequent revisions. The current release (Build 11, Feb 2011), does not reflect similar reorganization for Haplogroup T despite early recognition by Malyarchuk and Derenko (1999) that while subgroup T1 is well-resolved, "other HVS I sequences cannot be differentiated into subgroups due to possible homoplasies at nucleotide positions 16292, 16296, and 16304, leading to reticulations in the topology of phylogenetic networks." This report proposes significant restructuring of clades of Haplogroup T as currently presented in the literature.

Logan and Athey recently explored the super haplogroup N (which subsumes the major haplogroups found in Europe, including JT) and raised a number of issues about how clades of the mtDNA phylogeny should be defined (Logan and Athey, 2010). The current study incorporates guidelines inferred in that paper.

Development of the Haplogroup JT Phylogeny

The results in this paper are based on full-coding-region mtDNA sequences available in GenBank, supplemented by mtDNA test results obtained through the T-FGS project at Family Tree DNA as administered by Pike and his associates (Pike, et al, 2010). In conjunction with the introduction of geographic data about the samples and possibly genealogical data, it is hoped that the results of this study can become the starting point for both more detailed and broader studies and be correlated with archaeological and other anthropology studies to infer migration patterns of populations represented by the various branches of the evolutionary tree.

Using the same techniques described in earlier papers by the author, genotypes of 515 GenBank records were chosen as representative of Hg JT. The defining markers for JT are T4216C, A11251G and C18452A. Included within these 515 genotypes were 99 that were complete in the coding region (CRS positions 577 through 16023) but not complete in the major non-coding region. These genotypes were flagged within the working data set to permit special attention as appropriate. However, since the analysis reported here use only the coding region data, this presents no problem and their inclusion provides a larger sample and thus improves the statistical analysis. This process yielded 270 genotypes for Hg J and 245 for Hg T.

Pike, et al (2010) provided an additional 281 genotypes for Hg T obtained through the T-FGS project at Family Tree DNA. These were merged into the 515 GenBank genotypes for a total of 796 genotypes in the evolving working data set. It is realized that there may be a small bias introduced by inadvertent inclusion of duplicate records or records for closely related individuals. However, when pairwise comparisons were made within the composite of Hg T genotypes, only 0.53% did not show difference. Furthermore, much of the analysis below computes differences between disjoint populations rather than within a single clade, and thus such bias is even less significant. There is also bias associated with introducing the T-FGS project data since it was predominately from Americans of European descent. Although Hg T is of mid-eastern origin, it is now found through Europe but very sparsely in Asia; furthermore, it is old enough to participate in the Mesolithic repopulation of Europe and certainly any Neolithic

movements associated with the spread of agriculture. The number of genotypes from GenBank is large enough that this bias should not be seriously effects the basic structure and reduction of computed age is likely to be no larger than the effect of other uncertainties discussed below.

In the process of selection and merging, the genotypes were also aligned on the various markers and presented in matrix format where each genotype was described in a column with presence of each marker of the genotype indicated in the appropriate cell whose row corresponded to that marker. In addition to the name of the marker, each row contains additional data including the type of marker (e.g., whether it came from a gene that coded for a protein, ribosomal RNA, transfer RNA, or non-coding), and a count of the total occurrences of that marker in the dataset.

The set of markers were then reviewed for possible simplification. For example, every occurrence of the C522 deletion was accompanied by the A523 deletion. It is obvious that these deletions occur as a single event rather than two separate events. The two markers were thus combined and renamed as 522-CA (i.e., a CA deletion starting at position 522). As an extreme example of simplification, the matrix contained 13 genotypes with overlapping sequential deletions in the range of 8276 through 8289. By judiciously collapsing some of these sequences, the same data can be represented as just 4 markers and reducing the total occurrence of these markers within the dataset from 274 to 32. This makes a significant difference when analyzing the difference between two genotypes where one or both contain these markers. The net result is that the definition of the phylogeny is significantly simplified, and the reference dataset more realistically represents the actual process that occurred in the evolution of the population that produced these haplogroups.

The working dataset was then expanded by copying selective rows and inserting them in a new section and rearranging the columns so as to create a definition of the JT phylogeny in matrix format using the methods as previously reported and the Hg J portion of the matrix generally conforms to previous studies. The Hg T portion of the matrix used the work of (Pike, et al, 2010) as a starting reference but significant adjustments were made to produce what the author considers to be a more parsimonious structure. Pike used a progressive clustering technique which finds genotypes that are close together but can diverge significantly from the true phylogeny when small clusters are eventually clustered in a complete tree. The matrix approach is inherently from the trunk out and parsimonious alternative can be more easily seen and evaluated. The approach presented here used the best features of both. The portion of the matrix that was used to define the phylogeny is available in the supplementary material in two parts. Supplementary File A shows the markers used in the Hg J definition and how they are distributed across the genotypes. Supplementary Files B and C show the markers used in Hg T definitions.

Once an initial phylogeny was completed, issues outlined by Logan and Athey (2010) were considered and clade definitions were simplified by removing markers that were found to be indicative of the clade but not definitive because of being the same single nucleotide polymorphism (SNP) is found in multiple clades that do not share a common ancestor; that is the marker is homoplasic. Just because these markers are present in all or most of the genotypes assigned to the clade does not mean they should be used in the definition of that clade. Some of these actually added to the appearance of confusion as one stepped back from the details and viewed the larger regions of the matrix. Examples of confusing markers are C16145A and

C16261T in Hg J and T152C, T195C, and C16296T in Hg T. A graphic showing the top level results of this process is presented in Figure 2. The numbers along the line indicate the loci of the markers used to define clade and the number to the right of the boxes are the counts of genotypes in the clade.

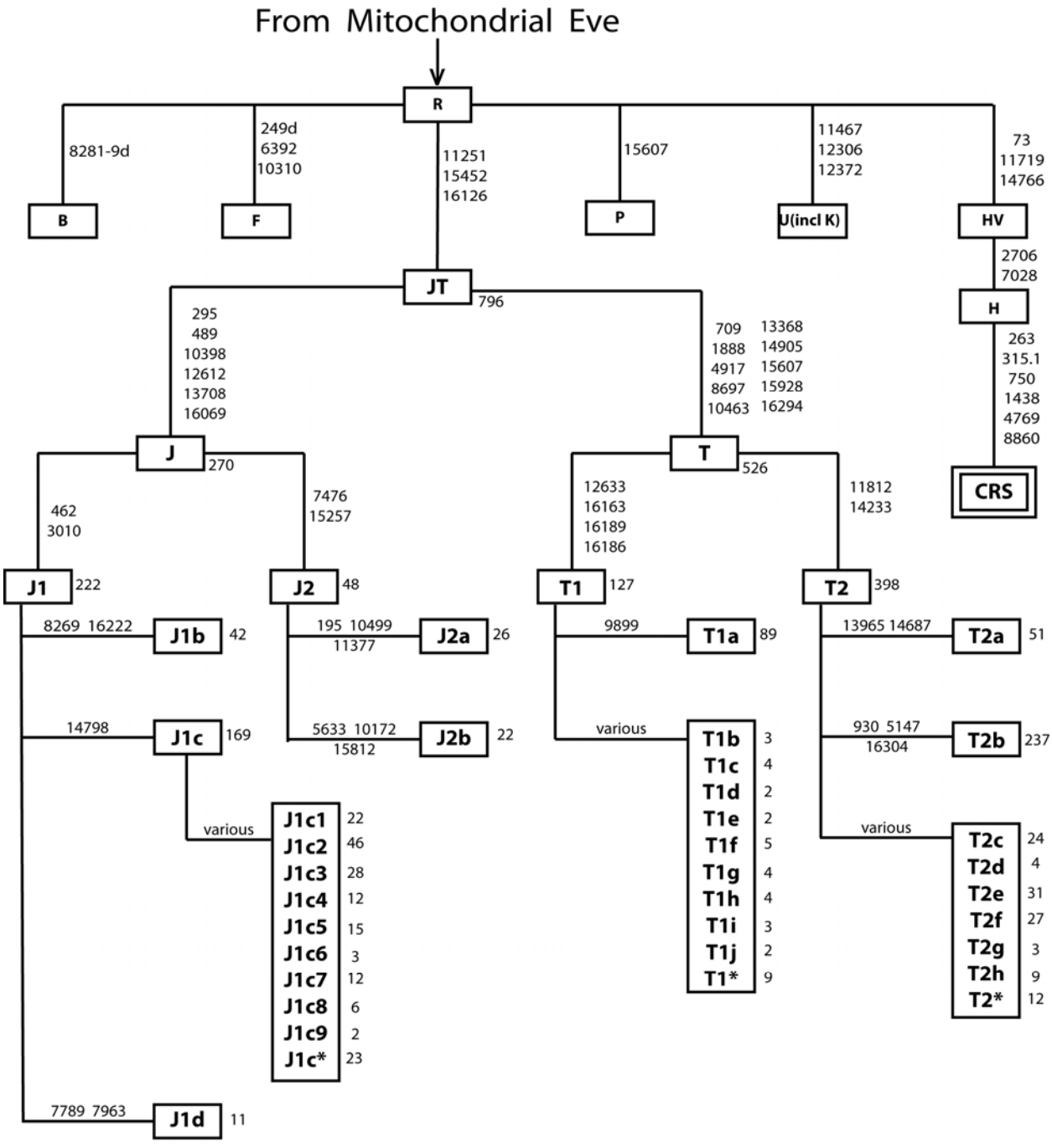

Figure 2. A segment of the human mitochondrial phylogeny showing positions of Haplogroup JT and top level branches of Haplogroups J and T.

Several observations can be made directly from this graphic.

Firstly, each of the 796 genotypes in the dataset were unambiguously assigned either to Hg J or to Hg T with none left over to be designated Hg JT*. A likely explanation of this phenomenon is that the original JT population became divided and isolated so that they underwent independent genetic drift. As a result of fixation, each developed defining markers. It is commonly thought that the origin of JT was actually in the Near East and their descendants moved into Europe concurrent with the advance of farming. The questions then become, what was the cause of this separation and isolation, exactly where were the two populations when defining drifts occurred, and when did these events occur. Statistical analysis of the diversity of the various clades may shed some light on relative dates, but other anthropological sources, such as linguistics and archaeology, are undoubtedly needed to infer location.

Secondly, the size of the T haplogroup is comparable to the J haplogroup. Remember that the original 515 genotypes from GenBank split 245 for T and 270 for J. However, since these genotypes are generally a byproduct of research not directed to JT or its clades, normal statistical inference cannot be assured. However, the size of the J1 clade is found to be almost five times as large as J2. Similarly, the size of T2 is found to be over three times the size of T1. The obvious question is, what is the source of this disparity. Although there is undoubtedly bias in the opportunistic dataset, it is does represent a wide spectrum of locations and thus bias is considered to be only a minor factor is his results. The question remains, is there something in the genetic makeup that results in different survival rate? With possible separation and isolation, is there something different about the environment that result in different growth rates? These are questions to be answered in future research.

Thirdly, within Haplogroup J1 the size of clade J1c is three times as large as J1b and J1d combined. The obvious question is why was J1c so advantaged. This graphic was also drawn to bring out the star-like characteristic of this clade. A likely interpretation of this is that there has been very rapid growth in this population and that there has not been enough time for the drift to result in fixation to fewer well defined clades. Perhaps the J1c clade was at the very front of the "wave of advance" associated with advancement of farming from the Near East into Europe (Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Excoffier and Ray, 2008). The details of the J1c clade were presented by Logan (2009) and are not significantly different from those in the current release of PhyloTree. They are not repeated here.

Fourthly, both T1 and T2 exhibit similar star-like structures, each with a dominant clade. Perhaps more significantly, once details are examined (see Figure 3), it is seen that both T1a, the dominant clade in T1, and T2b, the dominant clade in T2, themselves have very prominent star-like structures. Perhaps both T1a and T2b are also showing "wave of advance" phenomena from associated with the advancement of farming.

Figure 3 was derived directly from the matrix developed as described above and shows the next level of detail from Figure 2. The details of clade T1 is significantly different from PhyloTree in that the 18186 used to define T1a has been moved into the definition of T1; the subclades are correspondingly renamed; and the associated definition tree greatly expanded. This restructuring makes the star-like structure quite explicit. The lowest clades shown here are typically those

defined by Pike et al. However, there are significant differences in the relationships between these clades as determined by the matrix approach to identification and resolution of reticulations to achieve maximum parsimony.

Similarly, T2 is significantly different from the current release of PhyloTree. The most significant difference is the restructuring of T2b, where 16304 has been removed from its definition to form a new T2b1, and the remainder of T2b accordingly reorganized and renamed. This revealed the very prominent star-like structure. A comparable difference existed with the phylogeny of Pike et al. Although there are differences in both definition of clades and their relationships, results reported here are broadly comparable to those of Pike. The names from Pike have been used were possible.

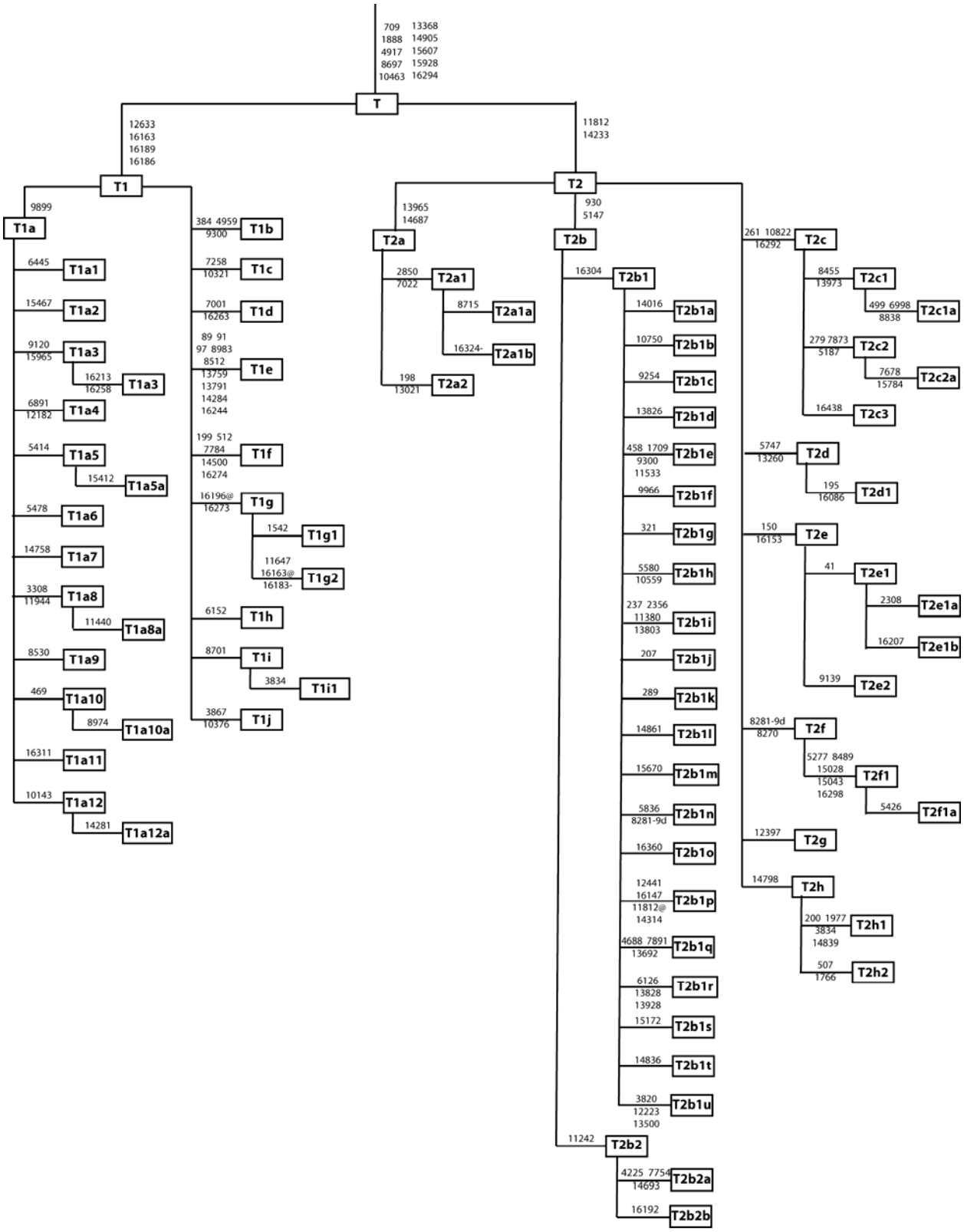

Figure 3. Details of Phylogeny for Haplogroup J.

Structural Analysis

As mtDNA is passed from one generation to the next, it accumulates random mutations. If these mutations can be identified and counted, these counts can then serve as an indicator of the number of generations between a distant ancestor and someone tested in a recent generation. Unfortunately, the DNA for ancestors generally cannot be tested. However, test results of members of the current population can be reported in terms of genetic markers relative to the rCRS reference and thus these results can be more easily compared to each other. This, in turn, permits the construction of a phylogeny as described above. For any given genealogical lineage, those markers that formed the definition of clades can be easily identified, but this does not identify which of the other mutations may have occurred in the lineage. Logan (2008b) took a subjective approach by counting all markers except for those that were considered to be at the root of the tree -- several of which were actually artifacts of the arbitrary reporting reference. The average resulting counts for each clade were simply multiplied by an assumed mutation rate to estimate a relative age. A more detailed analysis of these differences has been performed and is reported here.

In addition to average count, the distribution of differences from the reference can be informative. The left hand side of Figure 4 presents a set of histograms showing how counts of markers are distributed for the overall Hg JT and each of its two major components, J and T. Note that the mean count of JT is 25.8 which falls between the counts of 22.1 for J and 27.7 for T. The sample sizes are large enough that the central limit theorem of statistics should apply and thus a curve has been superimposed on each histogram to show the corresponding normal distribution for the same mean and standard deviation. From visual inspection, it is immediately obvious that there are significant differences, with distributions for J and T being better fits than the overall distribution for JT. This difference from a normal distribution is indicative of some factor beyond pure random mutations. That is, there is some type of structure present.

This presence of structure becomes quite clear in right hand side of Figure 4, where the same samples are used, but the counts are now pairwise differences, another measure of genetic distance. The most obvious feature is that the histogram for JT is now clearly bimodal, with the mode on the right centered at about 24 and a new mode centered at about 9. In this case the structural factor leading to this phenomenon is clear -- although both Hg J and Hg T have a common ancestry, they also have a number of identified differences as shown in the phylogeny. However, the two modes do not correspond to the two component haplogroups. Rather, the mode on the left is indicative of the average differences between the genotypes of Hg J and Hg T taken individually and the mode on the right corresponds to the average difference between these two haplogroups. Note that the mean pairwise counts of 10.1 for Hg J and 8.3 for Hg T encompass the value of approximately 9 for the center of the left hand mode of the total JT haplogroup.

Taking pairwise differences has removed the effect of counting the markers that are common for all genotypes and thus do not represent mutations. On the other hand, when pairwise differences are taken between haplogroups (or clades), all the differences, including those in the defining markers, and those defining markers closer to the root and the artifacts of the reference are

appropriately included. This is illustrated by the fact that the interclade test for J and T as shown in Figure 5 shows a mean value of 23.85.

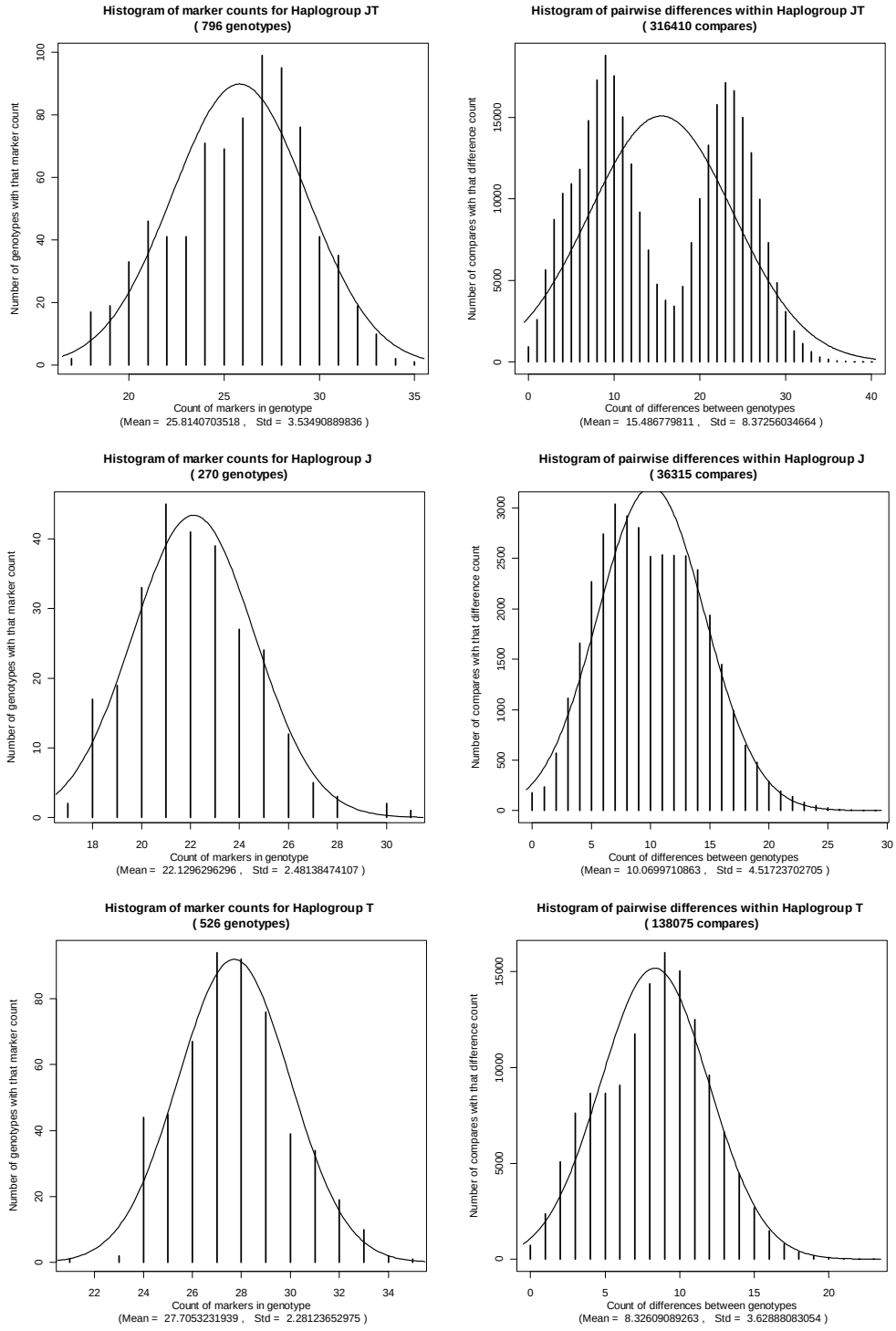

Figure 4. Distribution of marker counts found in Hg JT and its components J and T on the left and corresponding distribution of pairwise comparisons on the right.

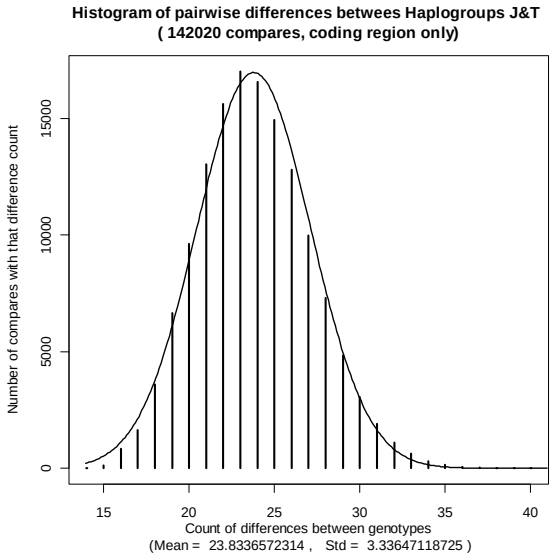

Figure 5. Pairwise differences betweeen Hg J and Hg T

Another feature of analysis with pairwise differences is that they are less sensitive to the choice of genotypes used in the respective sample. This is important when working with samples of opportunity where there has been little control in the selection process to guarantee that the sample is representative. One example of such selection bias would be samples that include a number of genotypes from the same lineage where the difference between them would be zero or small and thus the average marker count would be reduced. Bias would also be present in the test results from a single source, such as Family Tree DNA, since these are not randomly distributed geographically. On the other hand. when using pairwise differences across clades, i.e., interclade analysis, the effect of this bias is significantly less.

The use of pairwise differences (also called mismatch) in this type analysis has been justified in (Rogers, et al, 1996).

Histograms of pairwise differences based on interclade comparisons thus seem to be a reasonable basis for analysis of the phylogenetic structure and their statistical means appear to be a reasonable basis for estimating the number of mutations since the common ancestor.

Structural Analysis of Haplogroup J

Figure 6 shows the distribution of pairwise differences within several clades of Hg J. Clade J1 in the upper left, shows significant differences from normal, but nevertheless presents a smooth unimodal pattern in contrast to J2 (upper right) which is both multimodal and rather irregular. Smooth curves are typical of populations that have started small, perhaps after a bottleneck, and have experienced relatively rapid growth. An irregular pattern, conversely, is commonly found for populations that have become stagnant for some time (Slatkin, et al, 1991; Rogers and Harpending, 1992; and Excoffier and Ray, 2008)

A similar phenomenon is seen in the subclades of J1, shown in the first column, with J1c exhibiting characteristics of a growing population and J1b showing characteristics of a stagnant population.

Subclade J1d, not shown, is also very irregular, but with only eleven genotypes the sample is quite small for making statistical inferences. It also contains the same 152 homoplasic marker that is also present in J2a that caused considerable confusion in early attempts at classification using only HVR1/HVR2 results.

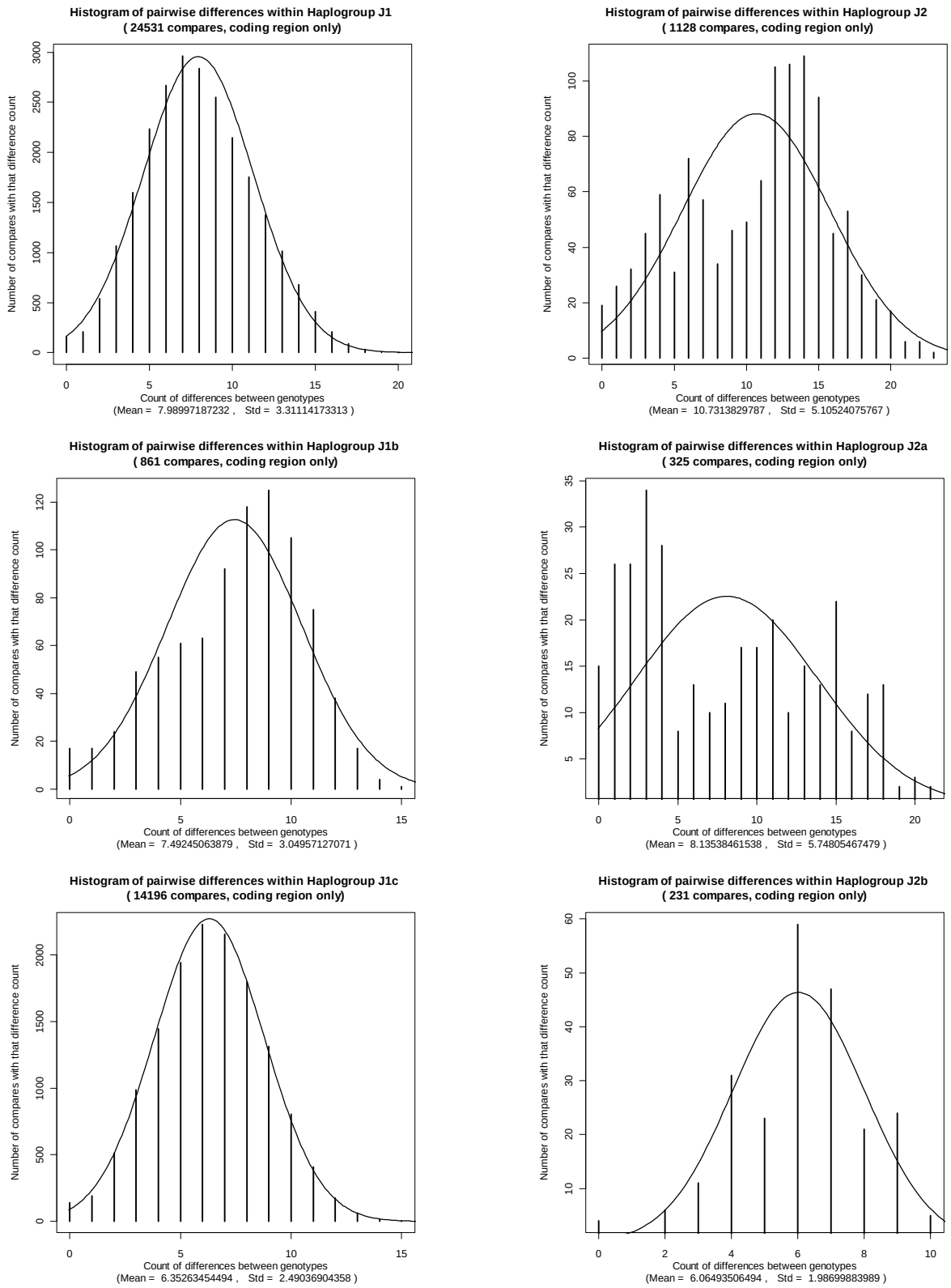

Figure 6. Distribution of pair-wise differences between genotypes within clades J1 and J2 and the corresponding distribution for their major subclades.

Irregularities in both J2 and its subclades, especially J2a, indicate that they are clearly neither young nor fast growing. Although not the primary cause of the non-regular profile, it is significant to point out that due to homoplasies of 16145 and 16261; earlier definitions of what is now called J2a were thought to be part of J1. When the error was discovered, J1a was renamed J2a and the J1a name retired (Palanichamy, et al, 2004).

As illustrated in Figure 4, histograms of both marker counts and pairwise comparisons within clades (i.e., intraclade compares) can be ragged or even multimodal. By contrast, relatively smooth profiles are produced by interclade analysis where each member of one clade is compared with each of the members of the other clade. This was illustrated in Figure 5 where members of Hg J were compared with members of Hg T. However, it may differ significantly from a normal distribution. As previously described, pairwise comparisons (whether within a population or between populations) are also better than simple marker counts since they remove the problem of determining which markers should be assigned to the clade and which ones should be considered as background. However, of the two, analysis between populations is less sensitive to ascertainment bias caused by using populations of opportunity rather than a scientifically designed random sample. In forming these subpopulations for analysis, it is desirable that closely related members be placed in one population or the other in order to minimize their impact. But how are these situations to be recognized and adjustments made before the actual counting process? The answer is found in using a phylogeny to define subpopulations for analysis corresponding to clades of the phylogeny. That is, the diversity of Hg JT is better represented by the interclade analysis between Hg J and Hg T rather than the marker counts or even the intraclade analysis of Hg JT. Similarly, the diversity of Hg J is represented by the interclade analysis of Hg J1 and Hg J2. Several such interclade results are shown in Figure 7.

But populations representing clades are not always easily bifurcated into clades. For example, Hg J1 has three relatively well defined subclades with J1c as the dominant one. J1c, in turn, has ten subclades including the J1c* which is a subpopulation of J1c without any further defining markers. Thus, included in our results are interclade analysis of J1b with J1c, J1c with J1d, and J1c with J1c. We have also included an 'interclade' analysis, designated J1c&J1(xJ1c), which compare members of J1c with all the other members of J1 excluding J1c. Note that all three present smooth profiles and that J1b&J1c is statistically very comparable to J1c&J1(xJ1c), whereas the eleven members of J1d produced a somewhat smaller average count when compared with J1c and a larger count when compared with J1b. (The latter two are not shown in the graphic but are included in the analysis where these results are used in inferring age of clades.)

The most irregular profile in Figure 7 is the results from the interclade comparison of J2a and J2b, indicating that Hg J2 is probably a stable population over many years with little growth.

On the other hand, it is significant to note that Hg J1c is by far the largest clade of Hg J. This is likely due to a high growth rate over the past few thousand years rather than being an old clade that has more years to expand.

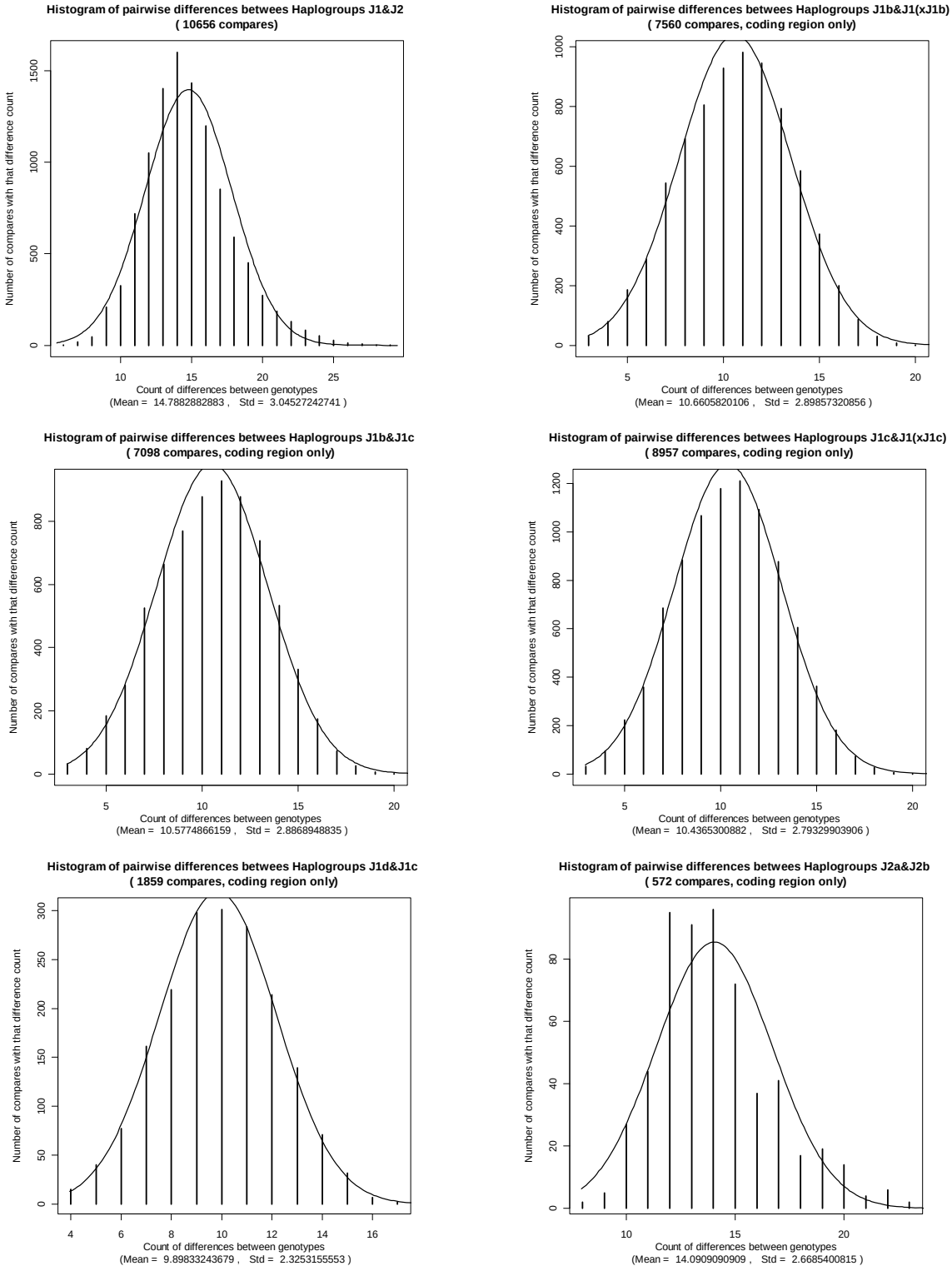

Figure 7. Interclade distributions for selected clades of Hg J

Structural Analysis of Haplogroup T

Interclade analyses for Hg T were performed as they were for Hg J. Since the T1 clade had only one major subclade (T1a), its T1 complement (all of T1 excluding T1a) was also analyzed. The resulting profiles (Figure 8) each show some irregularities from normal but not to the same extent as for J2 or J2a above. No general conclusions could be drawn.

The interclade profiles for Hg T (Figure 9) are also relatively unremarkable. A skew can be seen in the interclade comparison of T1a with its T1 complement. A large part of this skew is likely due to the fact that the populations are not clades and do not represent true independence. In fact, T1-complement includes nine different subclades that have been recognized, all of which are quite small and thus not useful for interclade analysis. The lack of independence of the two populations being compared is also reflected in the intraclade analysis of T1(xT1a) where the genetic distance shows a spike at 3 markers.

T2 has two major subclades and six minor subclades. As an approximation of the interclade profile for T2, a comparison was made between T2a and T2b, with relatively unremarkable results. The interclade technique was similarly applied to each of the largest subclades of T2 and the results for both T2a and T2b are also shown in the graphs. All of these profiles are relatively unremarkable.

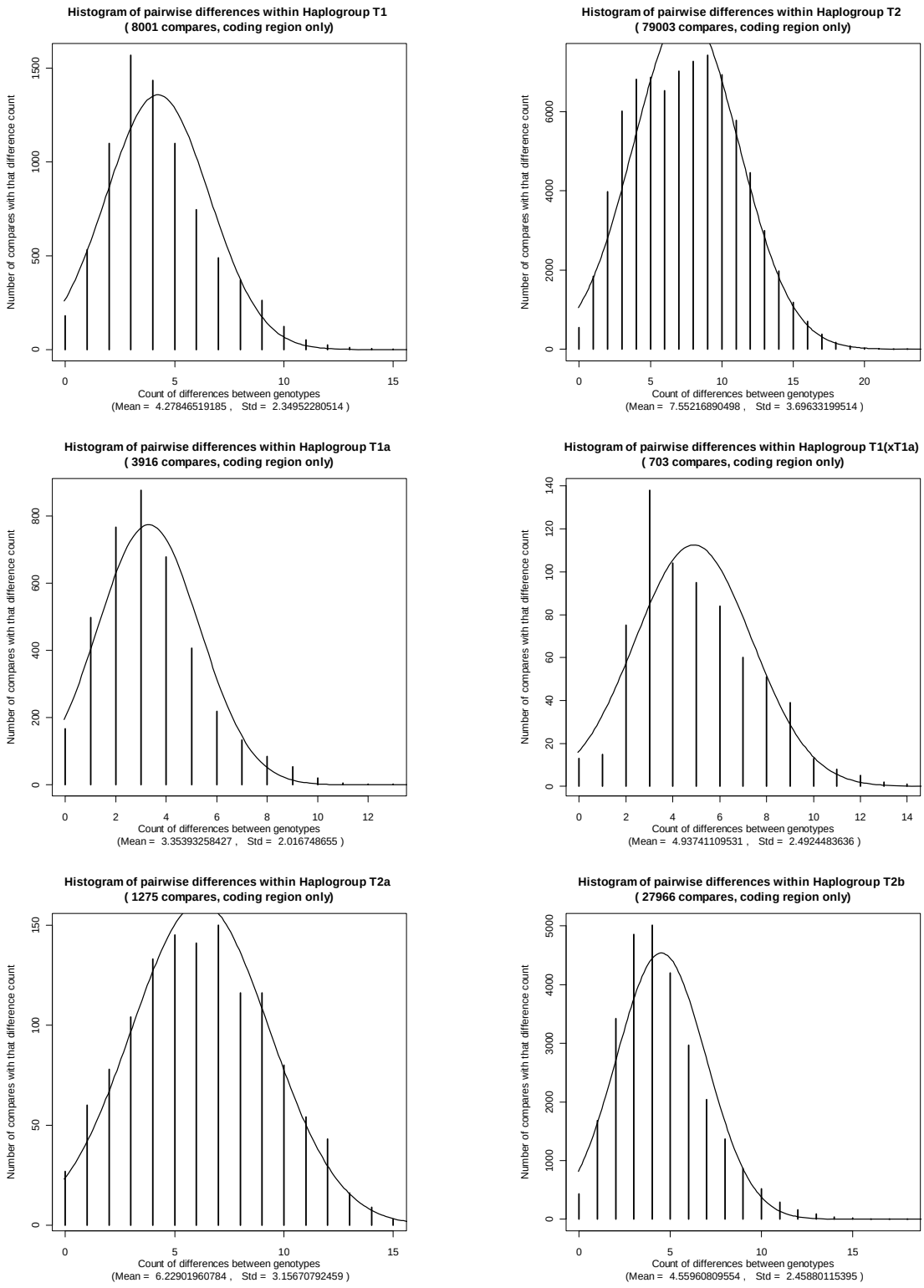

Figure 8. Distribution of pair-wise differences between genotypes within clades T1 and T2 and the corresponding distribution of their major subclades. (T1(xT1a) serves as substitute for nonexistent T2b).

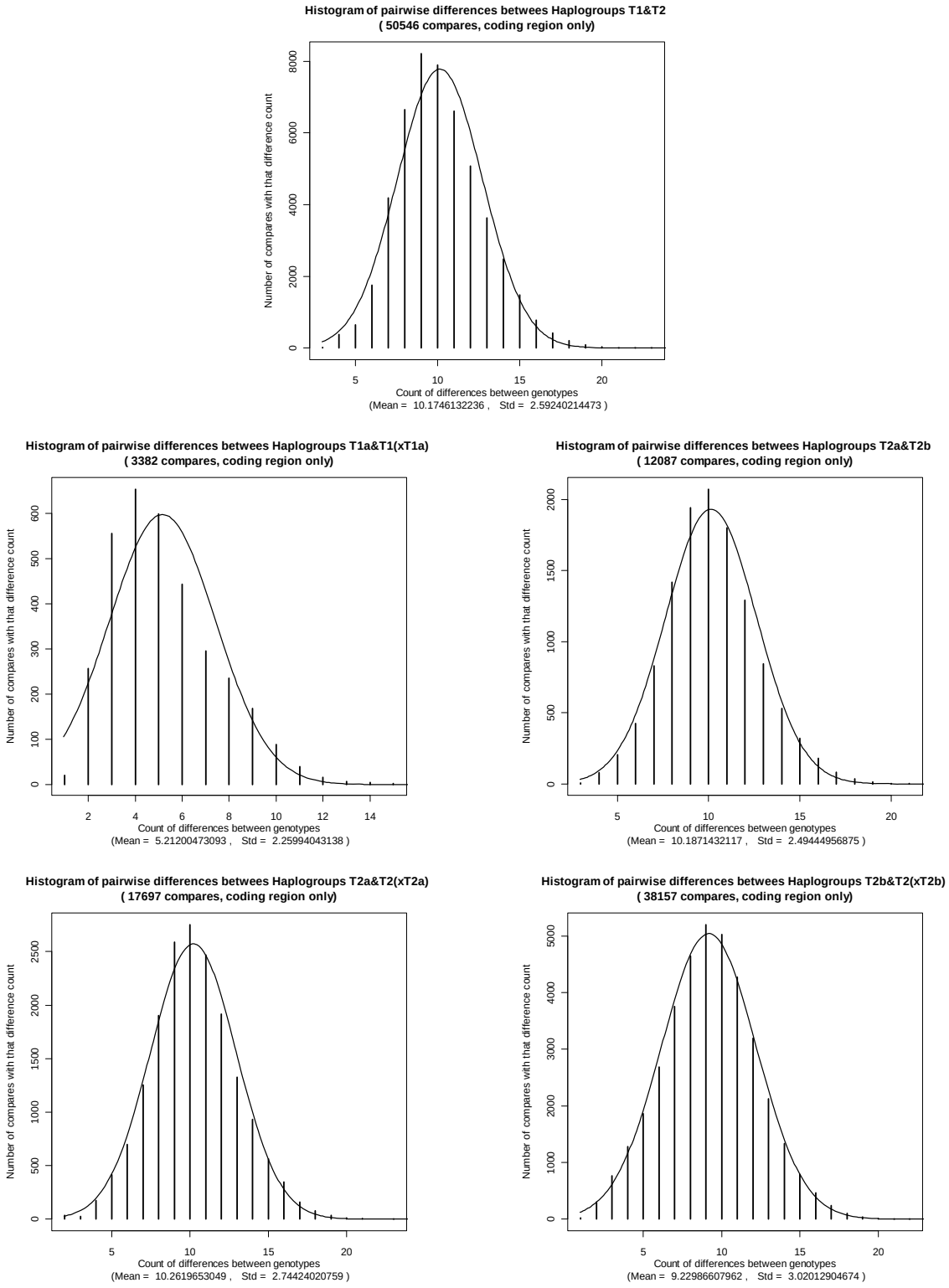

Figure 9. Interclade distribution for selected clades of Hg T.

Inferences about age of the clades.

For each application of the interclade technique, the mean, standard deviation, and standard error were computed. Summary of the results are shown in Table 1 where the first column indicates the clade being analyzed and the second column identifies the components used in that analysis. The next three columns provide the corresponding statistical data.

Table 1
Results of Interclade Analysis Used to Infer Age of Clades

Target	Components of	Interclade Differences			Inferred Age		
Clade	Interclade	Mean	Std Dev	SEM	Mean	Std Dev	SEM
JT	J & T	23.83	3.34	0.12	49590	6951	250
J	J1 & J2	14.79	3.05	0.18	30778	6347	375
J1	J1b&J1c	10.58	2.89	0.20	22017	6014	416
J1	J1d&J1b	11.94	2.78	0.38	24847	5785	791
J1	J1d&J1c	9.90	2.33	0.17	20602	4849	354
J1b	J1b&J1(xJ1b)	10.66	2.89	0.19	22183	6014	395
J1b1	J1b1&J1b(xJ1b1)	9.68	1.68	0.26	20144	3496	541
J1c	J1c&J1(xJ1c)	10.44	2.79	0.19	21726	5806	395
J1c1	J1c1&J1c(xJ1c1)	7.21	2.86	0.18	15004	5952	375
J1c2	J1c2&J1c(xJ1c2)	6.31	2.56	0.20	13131	5327	416
J1c3	J1c3&J1c(xJ1c3)	6.79	2.06	0.16	14130	4287	333
J1c4	J1c4&J1c(xJ1c4)	6.69	2.12	0.16	13922	4412	333
J1c5	J1c5&J1c(xJ1c5)	7.28	2.35	0.18	15150	4890	375
J1c7	J1c7&J1c(xJ1c7)	7.61	2.10	0.16	15836	4370	333
J1c*	J1c*&J1c(xJ1c*)	5.36	2.34	0.18	11154	4870	375
J1d	J1d&J1(xJ1d)	10.30	2.55	0.17	21434	5307	354
J2	J2a&J2b	14.09	2.67	0.39	29321	5556	812
J2a	J2a1&J2a(xJ2a1)	13.94	3.08	0.60	29009	6409	1249
J2b	J2b1&J2b(xJ1b1)	6.67	0.86	0.18	13880	1790	375
T	T1 & T2	10.18	2.59	0.11	21185	5390	229
T1	T1a&T1(xT1a)	5.21	2.26	0.20	10842	4703	416
T2	T2a&T2b	10.19	2.49	0.15	21205	5182	312
T2a	T2a&T2(xT2a)	10.27	2.74	0.14	21372	5702	291
T2b	T2b&(T1xT2b)	9.23	3.02	0.15	19208	6285	312
T2c	T2c&T2(xT2c)	10.98	2.60	0.13	22849	5411	271
T2e	T2e&T2(xT2e)	6.83	2.72	0.14	14213	5660	291
T2f	T2f&T2(xT2f)	10.14	2.94	0.15	21101	6118	312

Since J1b, J1c, and J1d form three mutually exclusive but exhaustive clades of Hg J1, each pair of clades was analyzed as an indicator of statistics of J1. Interestingly, the mean diversity as

indicated by the two largest clades is between the corresponding diversity computed for either combination that uses the small clade J1d.

For most of the clades, there is only one major subclade to be used in the interclade analysis, and the test of its properties must be obtained by comparing the population of that clade with its complement, that is, comparing it with all the members of its parent clade except those in the clade being analyzed. This indirect approach generally provides a greater diversity measure than the true diversity if the latter could be determined directly. Notwithstanding these limitations, computation of age from these diversity metrics helps in identifying relative age.

From the discussion in (Soares et al, 2009), we can infer a mutation rate of 1.56×10^{-8} mutations per year per nucleotide for the overall coding region. (These authors also analyzed variations for various kinds of markers, such as the differences in rates for the relative position on the codons found within the areas that code for protein. In this paper we take a straight forward approach, believing that such refinements are insignificant for our purposes.)

When multiplied by the coding region length of 15447 and a factor of two because of the use of pairwise comparisons rather than simple counts, this mutation rate becomes 4.82×10^{-4}, or 2075 years per count. Ages of the various clades have been computed by dividing the mean of the interclade differences by this mutation rate. These ages are shown in the last three columns of the table in Figure 10. A corresponding tree of ages is presented in Figure 11. Each of the largest clades of Hg JT as represented listed in the bottom of the chart. The order of their occurrence was chosen for clarity of presentation. Since clades were named as they were discovered and not based on sequence within the phylogeny, it is not possible to retain alphabetic order without excessive crossing of lines. The length of each vertical line represents the approximate computed age of the clade in thousands of years and the numeric value of that age is presented along the line. Each horizontal line leads from that clade to its parent clade. All clades computed in the table of Figure 10 are included except for T2c whose indicated age was 22.8 thousand years which was significantly older than the indicated age of 21.2 for its parent T2.

The coalescence ages for Hg J cannot be compared directly with those of Serk (2004) because there are major differences in her phylogenetic structure from what is now generally accepted. Furthermore, the opportunistic data set used here does not permit geographic segmentation as was done by her. Although geographic designation is available for many of the genotypes, much of it is non-specific, such as "Europe" or "European". Furthermore, even if the sample was taken in Europe, it could have been from a person with fairly recent descent from some other area such as the Middle East.

The Hg J ages can, however, be compared with (Logan, 2008b). One significant difference is that the ages presented here show greater structural integrity. For example, for the ages inferred here, none of the clade ages exceed that of their parent, whereas this was not the case for the previous study. To illustrate, the current age for J1 and J1b are 22.5 and 22.2 thousand years where as the corresponding data from the previous study was 17.6 and 19.8, an obvious impossibility. Similarly the ages for J2 and J2a here are 29.3 and 29.0 thousand but the previous computations showed 24.8 and 27.0 respectively. The second significant difference is that the

results of the earlier study are about 30% lower, probably due to the rather subjective way of partitioning observed markers between occurring time-wise after the clade was established vs. before.

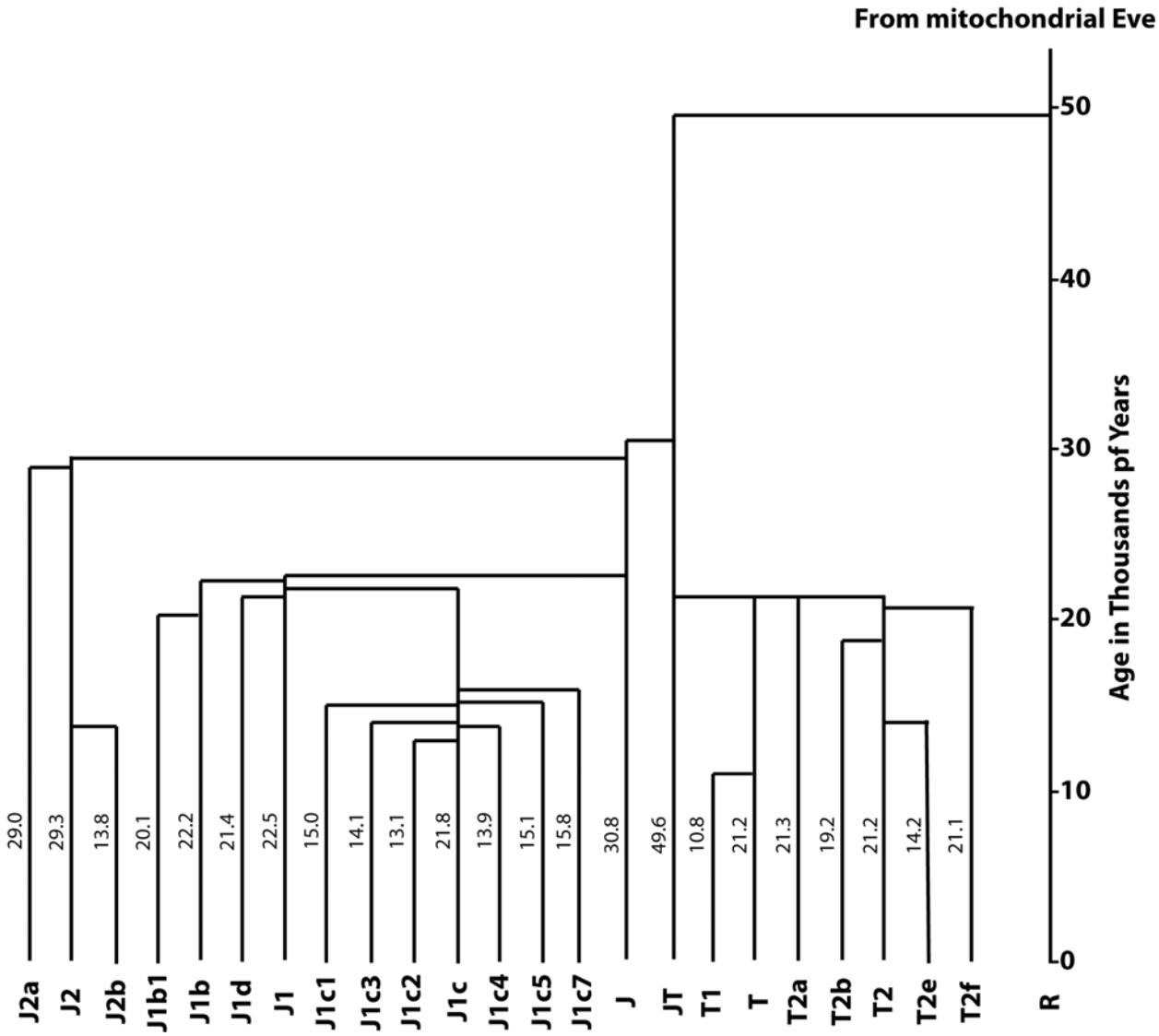

Figure 10. Computed age of Hg JT and its primary clades.

In developing the current phylogeny for Hg T, the naming criteria was to adopt the same names for comparable clades whenever possible. Although some differences were inevitable, they largely occur in the details and do not prohibit direct comparisons of ages for the more general clades. The most significant comparison here is the Pike ages are on average 91% higher (not lower) than those computed here. This is likely due to the use of extreme differences within the clade used there as distinct from averages distances between component subclades as used in the current analysis. As indicated above the latter approach should be more accurate.

It is instructive to compare these revised age estimates with those associated with the early colonization of Europe. It is generally agreed that a small group migrated from the Horn of Africa by a southern coastal route arriving in India by about 60-70 kya with segments proceeding in similar manner to South East Asia and even Australia. Migrations to the interior, however, were delayed by extreme desert conditions. Richards, et al (2006) suggest that members of this migration and thus all the major medina clades found in Europe today descend from a single mtDNA type (Hg L3) that originated about 85 thousand years ago (kya) in Eastern Africa. Furthermore, they state that the "Fertile Crescent was occupied by modern humans by ~50,000 years ago ... and the oldest western Eurasian clades (JT and U, within haplogroup R) date to this time and appear to have differentiated largely in this region." These conclusions were based in part on statements by Richards, et al (2000) where they compared frequency and ages of haplogroups in the Near East and Europe. They gave the age of Hg J in the Near East as approximately 48 kya and in Europe as 24.5 kya. The corresponding ages from Hg T were 47 kya and 37 kya. They also noted that both J or T were non-star-like but that clade T1 was star-like with a Near East age of approximately 22.5 kya and a European age of 9.4 kya.

The coalescence age of 49.6 kya as computed here for Haplogroups J and T is very comparable to the age of the differentiation that was said to have taken place in western India, in the region of the Fertile Crescent or in the migration path between them. It is important to emphasize that this is well before the appearance of modern humans in Europe.

The computed coalescence age of 30.8 kya for Hg J appears between the 48 and 24.5 kya respectively from the Near East and Europe as reported by Richards. This seems reasonable, considering that the data from GenBank includes both Near East and European genotypes. The coalescence age for Hg T reported here is significantly less than that for Richards. This is possibly due a significant bias toward western Europe sources resulting from use of data obtained from a special project where participants were primarily from the United States. In either case, the fact that the Richard's European ages are significantly smaller that their Near East ages supports the hypothesis of a Near Eastern origin for both J and T and a migration from there into Europe. These migrations are, no doubt, complex and need to be analyzed further within the context of several recognized migration episodes into and within Europe: a) the pioneer colonization into Europe replacing the Neanderthal, b) the re-colonization after the populations were displaced into refugia or decimated by the Last Global Maximum, c) population adjustments after the Younger Dryas glacial event, and d) the infusion of new Near Easterners with the wave of advancement of agriculture and new Neolithic technologies. Even though the specific migration structure is not known at this time, the fact that it occurred provided significant opportunity for the occurrence of successive founder effects and associated drifts. This would naturally lead to the star-like structure observed in the second and third level tiers of both J and T.

The age of J and T and their first order divisions suggest that migration into Europe was occurring well before introduction of agriculture. Currently available genetic alone data does not permit a definitive determination of the geographic location of the clades. Nevertheless, progress is being made under the name of archaeogenetics as exemplified by (Soares, et al, 2010).

Limitations and possible future work

Clustering and maximum parsimony techniques were used in the development of the phylogeny as reported here. Using maximum likelihood modeling or other such techniques could put this development on a more theoretical basis but are unlikely to make significant changes. Considerable experience from multiple researchers adds confidence in at least the top levels of this phylogeny. There is no doubt that as additional data becomes available, refinement will occur.

Use of pairwise differences as a basis of counting mutations within a clade is an approximation. Use of complements of clades as surrogates for clades is a further approximation. As additional data is gathered, more detailed analysis might produce a technique for allocating markers to clades vs. background and thus permit direct counting rather than using the approximation technique applied here. However, in the development of population sizes, migration paths of the clades, etc., the integration of DNA analysis with archaeology, paleoclimatology, and a spectrum of anthropology techniques, are not yet well enough developed to justify this process.

The single mutation rate used to cover the entire coding region is an average. Precision can be gained by separating the markers into categories that correspond to codon position at which the mutation occurred, or separating synonymous mutations for non-synonymous mutations, etc. (Soares, et al, 2009). However, other factors such as the difficulty of identifying which mutations should be considered occurring within the clade introduces uncertainty that obviates the utility of such precision.

Finally, it must be recognized that mutation rates are being estimated based on a theory of coalescence that uses data from modern human populations and is calibrated using other estimates such as the coalescence age of human with chimpanzee. Use of such data provides relative ages but leaves uncertainty associated with the absolute age. The fact that the standard error associated with current age estimates are generally less than one thousand years adds confidence to the structure of results presented here. It is the opinion of the author that relative ages of the clades are quite accurate. The inconsistency of the age computation for the T2c clade being greater than the age of its parent clade is not considered to be a serious flaw; the small size of the T2c clade (24 genotypes) is only about 6.5% of the size of the balance of the T2 clade to which it is compared (374 genotypes). Furthermore, there is currently no feasible way to know if this T2c sample is representative.

Future availability of test results could help remove some of these uncertainties. However, even more progress to an understanding of human population structure and its evolutionary migrations will come from the relatively new field of phylogeography -- the integration genetic analysis with archaeological and linguistic data.

Conclusions

This study developed and briefly analyzed the clade structure of Haplogroup JT and from it derived estimates for the ages of the various subclades. These results are believed to be suitable as a foundation for future research by genetic genealogists and anthropologists. Computer based

tools were developed to assist in the selection of GenBank data to represent haplogroups of interest, processing that data to produce a phylogeny, analysis of pairwise differences between genotypes, and the estimation of the age of subclades. Anyone wishing to engage in collaborative research using this tool set is invited to contact the author by email using JJLNV @ Comcast.net.

Glossary

Genetics and statistics each have their own terminology. Furthermore, much of the terminology of genetics is defined in terms of genes and chromosomes, whereas the same general concepts are applicable to describing mtDNA data but need some slight adaptation. This paper is intended to be a foundation of additional research in Haplogroup JT. It is intended to set a pattern for similar analyses in other mtDNA Haplogroups. It is thus appropriate to set down a few definitions as an aid in communicating in unambiguous terms. The terms are presented in the order of increasing dependence on terms previously defined.

Entity: Anything with relatively well defined characteristics or boundaries so that it can be easily distinguished from all other entities. Entities can be concrete, such as persons who have been mtDNA tested, or they may be abstract such as a set of integers.

Population: A set of entities. In genetics the entities are typically people (homo sapiens), or a set of results obtained from mtDNA testing of those persons. In statistics, entities can be anything as long as they are well defined or differentiated.

Sample: A set drawn from a population chosen for some specific purpose, including the possible set with a single entity or none at all. For example, a set of mtDNA results satisfying a given criteria.

Sample point: A sample containing a single entity.

Sequence: The results of mtDNA testing of an entity that describes some sequential pattern inherent in the DNA molecule or molecules (e.g., those in a mitochondrion) being tested. mtDNA sequences are typically aligned with the revised Cambridge Reference Sequence (rCRS) and only these differences are reported.

Locus: A specified part of a DNA molecule or the associated test results. The typical definition in the field of genetics refers to an entire gene on a chromosome, but for mtDNA a locus is typically a single point as defined with reference to the rCRS. Locus is a general term and need not refer to a specific DNA molecule. The plural of locus is loci.

Single nucleotide polymorphism (SNP): A locus and its associated value where the locus is restricted to the variation observed at a single nucleotide (as defined relative to the rCRS), a deletion of the nucleotide at the position or the insertion of a new nucleotide after that position.

Short tandem repeat (STR): A region of DNA where there has been observed to have a multiple tandem sequence of a short sequence such as AGCT or ATGC.

Allele: Any of two or more alternative forms that have been observed for a locus in a set of molecules. The alternative forms can be observed between multiple molecules in a single entity (a heteroplasmy) or between entities. The type of allele is dependent on what is tested and the purpose of the test. At the lowest level, the alleles are simply variations of a single specific nucleotide, a deletion, or an insertion, at a specific locus caused by a point mutation. These variations are frequently referred to as single nucleotide polymorphisms, of SNPs for short. In the chromosomes, allele may be the variation in length of short tandem repeats (STRs) found throughout the typical chromosome, such as the counts of STR markers for the Y chromosome now commonly used in genetic genealogy. Geneticists look at more complex variations such as the alternative forms of complete genes that contribute to variations in the traits that can be observed in the human population.

Genetic marker: A specific locus and its associated alleles chosen because it is thought to be useful in the analysis of DNA results. Generic marker is a general term. For example, in genetic genealogy of paternal ancestry the typical marker is the variation of a STR found on the Y chromosome. In mtDNA-based anthropology of humans, the marker is typically a SNP.

Haplotype: The values for a set of genetic markers, such as one used to compare samples or to define a sample set. More specifically, a set of alleles for a set of genetic markers. If two individuals are tested for a specific set of genetic markers and their results are found to be identical then they have the same haplotype.

Genotype: The haplotype for a specific entity. Although this term has a somewhat broader meaning in genetics, its use is restricted in this paper. For purposes of this paper, haplotype is a general term and can be spoken of in the abstract; genotype is the restrictive term and applies to a specific entity and thus is used in referring to a specific test result.

Phylogeny: A tree structure of alleles that represents an inferred evolutionary common ancestry of a set of entities where the most recent common ancestors are characterized by a set of alleles. This is a restrictive definition for use in this paper. Mathematics has a vocabulary for precisely characterizing this tree and its components. However, there has been considerable argument about applying these terms and thus they will be avoided in this paper. A top level phylogeny for Haplogroup JT is presented in Figure 2.

Clade: A part of a phylogeny that includes the most recent common ancestor of all of its genotypes and all of the genotypes for that common ancestor. Clades are strictly hierarchical and their names generally reflect that fact. For example, J1 is a clade of J and J1c is a clade of J1. This paper reports on the analysis of clades and pairs of clades; for example Haplogroup J is compared with Haplotype T and J1 is compared with J2. However, not all sets of genotypes used in the analysis reported here are clades. For example, J1c is of special interest within J1 and it is of interest to compare the genotypes of J1c with all of the genotypes of J1 that are not J1c. The second set will be designated with the special name J1(xJ1c) and by definition is not a clade.

Haplogroup: A member of a hierarchical classification system that groups related haplotypes. For mtDNA this is simply an upper level of the larger branches (the ones closest to the trunk of the tree) and are defined in terms of SNPs. See Figure 2 for an illustration of this hierarchy as it applies to Hg JT.

Sub-haplogroup: A portion of a Haplogroup. The clades are sub-haplogroups but not all sub-haplogroups are clades. For example, within Hg J, J1c is a sub-haplogroup, a clade, and also a haplogroup. On the other hand, the union of J1b and J1d is a sub-haplogroup but not a clade or a haplogroup. In the current working data set, this union of J1b and J1d is exactly the same as all genotypes of J except that part that is J1c, designated J(xJ1c).

Data Set: A set of data maintained for reference or analysis in a specific context. The working data set referred to throughout this paper is derived from the selection of genotypes for analysis and subsequently aligned in a matrix to show commonalities and differences in their markers. The columns of the matrix are arranged to place the genotypes in their phylogenetic sequence and rows of markers were copied and inserted in a new section to show the definition of the phylogeny.

Acknowledgement

The author gratefully acknowledges the constructive critiques provided by Whit Athey, Josh Weinstein, Blaine Bettinger, and an anonymous reviewer.

References

Di Rienzo A and Wilson A (1991). Branching pattern in the evolutionary tree for human mitochondrial DNA, Proc. Nat. Acad. Sci. USA, 88, 1597-1601.

Excoffier, Laurent, and Nicalas Ray (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecol Evol*. 23(7):347-351.

Logan, Jim (2008a). The Subclades of mtDNA Haplogroup J and Proposed Motifs for Assigning Control-Region Sequences into these Clades. *J Genet Geneol.* 4(1):12-26.

Logan, Jim (2008b) A comprehensive analysis of mtDNA Haplogroup J. *J Genet Geneol,* 4:104-124.

Logan, Jim (2009a) A Refined Phylogeny for mtDNA Haplogroup J. *J Genet Geneol*, 5:16-22.

Logan. Jim and Whit Athey (2010). A Reference Database to Support Analysis of mtDNA Haplogroup N, its Descendant Haplogroups, and Associated Clades. *J Genet Geneol*, 6(1).

Malyarchuk, B A, and M V Derenko (1999). Molecular Instability of the Mitochondrial Haplogorup T Sequences at Nucleotide Positions 16292 and 16296. *Ann Hum Genet* 63:489-497.

Palanichamy, Malliya grounder, Chang Sun, Surakesha Agrawal, Hans-Jurgen Bandelt, et al (2004), Phylogeny of Mitochondrial DNA Macrohaplogroup N in India, Based on Complete Sequencing: Implications for the Peopling of South Asia, *Am J Hum Genet*, 75(6):966-975.

Pike, David (2006). Phylogenetic Networks for the Human mtDNA Haplogorup T. *J. Genet Geneaol, 2:12-26*

Pike, David A., Terry J.Barton, Sjanra L. Bauer, and Elizabeth (Blake) Kipp (2010). mtDNA Haplogroup J Phylogeny based on Full Mitochondrial Sequences. *J. Genet Geneaol, 6(1)*

Richards, Martin, Helena Corte-Real, Peter Forster, Vincent Macaulay, Hilde Wilkinson-Herbots, Andrew Demaine, Surinda Papiha, Robert Hedges, Hans-Jurgen Bendelt, and Bryan Sykes (1996). Paleolithic and Neolithic Lineages in the European Mitochondrial Gene Pool. *Am. J. Hum. Genet*, 59:185-203.

Richards, Martin, Vincent Macaulay and 35 others (2000). Tracing European Founder Lineages in the Near Eastern mtDNA Pool. *Am. J. Hum. Genet*, 67:1251-1276.

Richards, Martin, Hans-Jurgen Bandelt, Thoomas Kivisild, and Stephen Oppenheimer (2006). A Model of Dispersal of Modern Humans out of Africa. In Hans-Jurgen Bandelt, Vincent Macaulay, and Martin Richards (Eds.), *Human Mitochondrial DNA and the Evolution of Homo sapiens*, No. 18 in Nucleic Acids and Molecular Biology series. Berlin, Springer Verlag.

Rogers, Alan R., and Henry Harpending (1992), Population growth makes waves in the distribution of pairwise genetic differences, Mol. BIol. Evol, 9(3), 552-569.

Roger, Alan R., Alexander E. Fraley, Michael J. Bamshad, W. Scott Watkins, and Lynn B. Jorde (1996). Mitochondrial mismatch analysis is insensitive to the mutation Process. *Mol. Biol. Evol*, 13(7):895-902.

Serk, Piia (2004). Human mitochondiral DNA Haplogroup J in Europe and Near East. M.Sc. Thesis. University of Tartu.

Slatkin, Montgomery, and Richard R. Hudson (1991), Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations, Genetics, 129, 555-562.

Soares, Pedro, Luca Ermini, Noel Thompson, Maru Mormina, Teressa Rito, Arne Rohl, Antonio Salas, Stephen Oppenheimer, Vincent Macaulay, and Marten B. Richards (2009). Correcting for Purifying Selection: An improved human mitochondrial molecular clock. *Am J Hum Genet*, 84(6):740-759.

Soares, Pedro, alessandro achilli, Omella Semino, William Davies, Vincent Macaulay, Hans-Jurgen Bandelt, Antonio Torroni, and Martin B Richards (2010). The Archaeogenetics of Europe. *Current Biology*, 20:R174-R183 (Feb 23, 2010).

Torroni, Antonio, M. T. Lott, M. F. Cabell, Y. S. Chen, L. Lavergne, and D. C. Wallace (1994), MtDNA and the origin of Caucasians: Identification of ancient Caucasian-specific haplogroups, one of which is prone to recurrent somatic duplication in the D-loop region, *Am J Hum Genet*, 55(4):760-776.

Torroni, Antonio, Kirsi Huoponen, Paolo Francalacci, Maurizio Petrozzi, Laura Morelli, Rosaria Scozzari, Domenica Obinu, Marja-Liisa Savontaus and Douglass C. Wallace (1996), Classification of European mtDNA's From an Analysis of Three European Populations, *Genetics*, 144:1835-1850.

van Oven M and M. Kayser (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30:E386-E394. (See also http://www.phylotree.org/)