

mtDNA Haplogroup T Phylogeny based on Full Mitochondrial Sequences

David A. Pike, Terry J. Barton, Sjana L. Bauer and Elizabeth (Blake) Kipp

Abstract

Using a collection of 445 full mitochondrial DNA sequences from members of human mtDNA haplogroup T, we develop a phylogeny for the haplogroup. We also calculate age estimates for several subgroups of haplogroup T.

Introduction

Human mitochondrial haplogroup T first surfaced in the academic literature in 1996, when Richards et al. associated it with a pair of hyper-variable region (HVR) mutations at nucleotides 16126 and 16294 within the mtDNA genome.¹ Very shortly afterwards, Torroni et al. identified a number of polymorphic restriction sites in the control region that were also associated with the haplogroup.²

It is now known that haplogroup T is one of two haplogroups within macrohaplogroup JT, which is characterised by coding region (CR) mutations at nucleotides 4216, 11251 and 15452 as well as an HVR mutation at 16126. Mutations at positions 709, 1888, 4917, 8697, 10463, 13368, 14905, 15607, 15928 and 16294 characterise haplogroup T and distinguish it from haplogroup J (which itself is characterised by mutations at 4216, 10398, 11251, 12612, 13708, 15452 and 16069).¹⁻⁵

As time passed and more genetic samples from members of haplogroup T were encountered, a subgroup structure began to emerge. For instance, Richards et al. delineated a collection of five subgroups based on HVR mutations.^{6,7} The T1 subgroup has since been further refined, leading to a revised subgroup definition as well as the introduction of motifs for the T1a, T1b and T1c subgroups as shown in Table 1.⁸⁻¹⁰

HVR mutations, however, are not always a reliable indicator of haplogroup membership, especially when used in isolation. To offer an illustrative

Subgroup	Associated HVR Mutations
T1	16189C
T1a	16163G, 16186T
T1b	16163G, 16243C
T1c	16182C, 16183C, 16298C
T2	16304C
T3	16292T
T4	16324C
T5	16153A

Table 1: HVR-Based Subgroups

example, there is an entry at GenBank (Accession Number AP008811) that has the following HVR mutations: 16126C, 16223T, 16263C, 16294T, 16362C and 16519C.¹¹ At first glance, it would appear that this individual may belong to haplogroup T, but when the coding region mutations are taken into consideration, it becomes evident that this person is in fact a member of haplogroup D.

Not surprisingly, there has been a shift towards the inclusion of coding region mutations when assessing haplogroup as well as subgroup determinations. As was the case with HVR mutations, a subgroup structure based on coding region mutations has also been developed for haplogroup T. In contrast to the five HVR-based subgroups, only two secondary-level CR-based subgroups have thus far been identified, each with a number of internal subgroups as summarised in Table 2.^{5,8,9,12-15}

Subgroup	Associated CR Mutations
T1	12633A
T1a	9899C
T1b	11647T
T2	11812G, 14233G
T2a	13965C, 14687G
T2b	930A, 5147A
T2c	6261A, 13973T
T2d	5747G, 13260C, 13708A
T2f	9181G, 13696G, 13803T, 13945G
T2g	3834A, 14798C, 14839G

Table 2: CR-Based Subgroups

Address for correspondence: David A. Pike, Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1C 5S7.
E-mail: dapike@mun.ca

Received: Sept. 16, 2010; Revised: Nov. 20, 2010; Accepted: Dec. 10, 2010.

Open Access article distributed under Creative Commons License Attribution License 3.0 (<http://creativecommons.org/licenses/by-nc-sa/3.0/>) which permits noncommercial sharing, reproduction, distribution, and adaptation provided there is proper attribution and that all derivative works are subject to the same license.

While the HVR-based and CR-based subgroups for haplogroup T have developed in parallel with one another, they have largely developed independently and with little correlation to one another. For instance, in 2001 Finnilä and Majamaa presented two sets of phylogenetic networks, one based on HVR mutations and the other on CR mutations.⁵ In 2002, Herrnstadt et al. focussed their attention exclusively on coding region mutations when they introduced the T2a and T2b subgroups.¹² In 2004, Kivisild et al., as well as Shen et al., incorporated HVR mutations into their analyses, but did not fully sequence the coding region when conducting their research.^{8,13} Also in 2004, Palanichamy et al. did make use of full mitochondrial sequences, but were limited to a set of ten samples from haplogroup T.⁹ One of these samples, which fell within their T2 subgroup, exhibited the 16153A mutation that is characteristic of the T5 HVR-based subgroup that had been introduced by Richards et al. in 2000.⁷ As we shall see in this paper, a number of other people who satisfy the criteria for the T3, T4 or T5 HVR-based subgroups are also classified as being within the T2 CR-based subgroup. So not only has the manner in which the phylogenetic refinement of haplogroup T taken place led to two distinct subgroup classification schemes (one based on just HVR mutations and the other primarily based on coding region mutations), but these schemes are not in agreement with one another.

This lack of consistency, and the corresponding confusion and frustration that it can generate, was the primary motivator for our present research. With a goal of developing a unifying and comprehensive phylogeny for haplogroup T, we enlisted a cohort of people who had previously been identified as members of the haplogroup. Full mitochondrial sequences were then determined for each participant in our study. Additional full mitochondrial sequences from GenBank were also taken into consideration. With a combined total of 445 full mitochondrial sequences at our disposal, we then developed a phylogenetic network for haplogroup T based on these sequences.

Subsequent to the commencement of our project, van Oven and Kayser began to maintain an online phylogenetic tree that is periodically updated as new mtDNA sequences are reported in the literature.^{15,16} Their tree uses a combination of HVR mutations and coding region mutations, but is nevertheless limited by the number of mtDNA genomes that have been publicly reported through the GenBank repository. Incidentally, in this tree the T2e haplogroup has no coding region mutations, but is based upon the control region mutations 150T and 16153A.

As we proceed to delineate subgroups of haplo-

group T, we do not propose to adhere to previously published phylogenies. Nevertheless, when possible we will endeavour to use motifs and nomenclature (i.e., defining mutations and names for subgroups) that are consistent with those already in use. In some cases the data that we have collected disagrees with some of the subgroup definitions that are in current use and in these cases we cannot help but to propose an alternative subgroup hierarchy.

Subjects and Methods

The Interdisciplinary Committee on Ethics in Human Research at Memorial University of Newfoundland granted its approval for this study on 15 January 2008. Participant enrollment commenced shortly afterwards. By 31 August 2010 a total of 301 members of mtDNA haplogroup T had joined our research project, obtained their full mitochondrial sequence (FMS) results from Family Tree DNA and granted informed consent to allow us to make use of their results.

In addition to the FMS data from direct participants, we supplemented our data with 164 haplo-

Source	No. of T Samples
Behar et al. ¹⁴	4
Coble et al. ¹⁷	39
Costa et al. ¹⁸	2
Detjen et al. ¹⁹	1
Family Tree DNA clients	32
Fendt et al. ²⁰	2
Finnilä et al., Moilanen et al., ^{21,22}	11
Fraumene et al. ²³	6
Gasparre et al. ²⁴	9
Ghelli et al. ²⁵	3
Hartmann et al. ²⁶	2
Ingman et al. ²⁷	1
Ingman and Gyllensten ²⁸	4
Kujanová et al. ²⁹	9
La Morgia et al. ³⁰	1
Maca-Meyer et al. ³¹	2
Malyarchuk et al. ³²	7
Mishmar et al. ³³	2
Palanichamy et al. ⁹	6
Pello et al. ³⁴	1
Pereira et al. ³⁵	7
Pichler et al. ³⁶	8
Rani et al. ³⁷	1
Rogaev et al. ³⁸	1
Tuo et al. ³⁹	1
Zaragoza et al. ⁴⁰	2

Table 3: Sources of FMS Data in GenBank

group T FMS entries that had previously been deposited with GenBank. These consist of 32 sequences submitted directly by customers of Family Tree DNA, plus an additional 132 sequences that have appeared in the scientific literature. Table 3 summarises the sources of the entries from GenBank that we used.

In 2009, Yao et al. published a list of mtDNA entries in GenBank that they suspect contain genotyping errors.⁴¹ For haplogroup T, nine sequences are in doubt: six of the twelve reported by Gasparre et al.²⁴, both of the two reported by Mishmar et al.³³ and the one reported by Tuo et al.³⁹ Of these nine questionable genotypes, we excluded three of those reported by Gasparre et al. from our FMS collection (namely those having Accession Numbers EF660972, EF660992 and EF661001) because they each have gaps in their data and cannot be considered to have been fully sequenced. The remaining six genotypes from haplogroup T that are questioned by Yao et al. were incorporated into the data set that we analysed.

Of the 32 sequences submitted to GenBank by Family Tree DNA clients, 19 had also enrolled as test subjects in our study. We accordingly redacted one copy of each of the corresponding 19 FMS results from our data set.

Two of the Family Tree DNA clients that had joined our project brought to our attention the fact that they share a recent maternal ancestor. Their FMS results were identical and so we took one of their two FMS sequences out of our data set in order to avoid mistaking mutations that might be exclusive to their immediately family as having phylogenetic significance.

In consequence of the foregoing considerations, we proceeded to work with a total collection of 445 FMS results from haplogroup T. This collection is available for download from <http://www.jogg.info/6/PikeData.txt>.

Additional pre-processing that we performed on our data set included the redaction of insertion mutations reported at nucleotide position 309.* Also, twenty-one samples contain a deletion of 9 base pairs starting at position 8281; we altered the representation of these deletions so that they would collectively be treated in our analysis as a single mutational event in each of the corresponding samples.

In conducting our analysis we divided our data set into three subsets as follows: the first subset consists of all 109 samples that contain the 12633A mutation that defines the CR-based T1 subgroup, the second subset consists of 205 samples that lack the 12633A mutation but have the 930A mutation which is part of the definition of CR-based T2b subgroup, and the third

subset consists of the remaining 131 samples that have neither 12633A nor 930A. The intent is for these data sets to respectively coincide with haplogroup T1, haplogroup T2b, and the remainder of the T2 haplogroup, although the third data set will also contain any genotypes that might belong to neither haplogroup T1 nor T2. We will refer to these three subsets as the T1, T2b and T2(-b) data sets respectively.

When the 109 genotypes in the T1 data set were initially inspected, aside from all of them containing the 12633A mutation, it was noted that the 16189C mutation occurs in 108 of the samples, 16163G in 107 samples, and 16186T in 104 samples.

Within the T2b data set, the 14233G and 11812G mutations that jointly define the CR-based T2 subgroup are respectively found in 204 and 203 of the 205 samples. The 5147A mutation that, together with 930A, defines the CR-based T2 subgroup is found in 204 samples. The 16304C mutation that defines the HVR-based T2 subgroup occurs in 185 sample genotypes.

Within the T2(-b) data set, the 14233G and 11812G mutations are respectively found in 131 and 130 of the 131 samples. The 5147A mutation is absent from all samples, although a heteroplasmic 5147R mutation is present in a single genotype.

Other mutations that are ubiquitous, or nearly so, in all three subsets include those that define macrohaplogroup JT, those that together with the JT mutations define haplogroup T, as well as highly recurrent mutations that are not considered to be part of the definition of the haplogroup:

JT: 4216C, 11251G, 15452A, 16126C

T: 709A, 1888A, 4917G, 8697A, 10463C, 13368A, 14905A, 15607G, 15928A, 16294T

Non-Defining: 73G, 263G, 315.1C, 750G, 1438G, 2706G, 4769G, 7028T, 8860G, 11719A, 14766T, 15326G, 16519C

On the premise that any mutation that is present in all samples (or all but one sample) of a subset does not represent an observable phylogenetic branch within the subset itself, we have simplified each data set of genotypes by redacting such mutations. However, any such mutations that are not present in parent haplogroups may represent defining mutations for the whole of the subset in consideration (e.g., 930A and 5147A in the T2b subset). We similarly treat any mutations that occur in at most one sample as being private rather than phylogenetic.

The resultant 109 simplified genotypes from the T1 subset yield the 59 distinct genotypes that are listed

* A polycytosine tract of DNA exists between nucleotides 303 and 309 inclusive. This region of the mtDNA genome is especially susceptible to insertions of additional cytosine nucleotides, so much so that it is commonplace for individuals to exhibit length heteroplasmy for this tract.^{42,43}

in Table 6. Likewise, the 205 (resp. 131) genotypes that comprise the simplified T2b (resp. T2(-b)) data set produce the 125 (resp. 98) distinct genotypes that are listed in Table 7 (resp. Table 8).

An integral component of the analysis of these three subsets of data consists of a visual representation in which each distinct genotype gives rise to a node within a graph.* Whenever two genotypes differ by a single mutation, then their corresponding nodes are joined by a solid blue edge. If a genotype has no others within a genetic distance of 1, then the nodes for any genotypes that are distance 2 away are joined to its node with dashed blue edges. Similarly, whenever a genotype has no others within a genetic distance of 2 (respectively 3), then any genotypes that are distance 3 (resp. 4) away are illustrated with dashed orange (resp. green) edges. Any nodes that remain isolated will therefore be a least 5 mutations away from any other genotype within the subset.

The intent is that the graphs that arise from the T1, T2b and T2(-b) data sets will approximate the phylogenetic networks for the corresponding portions of haplogroup T. These three graphs appear in Figures 1, 2 and 3, respectively. In each figure the nodes of the graph have been numbered in correspondence with the simplified genotypes that are listed in Tables 6, 7 and 8, respectively. Each node's size is proportional to the number of copies of the corresponding genotype that were present in our data set. As each solid blue edge represents a single mutational difference between two genotypes, we have labelled each such edge with the nucleotide position of the corresponding mutation.

We now proceed to focus on each data set, one at a time.

The T1 Data Set

As has already been observed, the mutations 12633A, 16163G, 16186T and 16189C are ubiquitous, or nearly so, within the T1 data set and so it is reasonable to put forward a FMS-based definition of the T1 subgroup based on these four mutations:

T1: 12633A, 16163G, 16186T, 16189C

To further justify this definition, we note that only one genotype in the T1 data set lacks the 16189C mutation; this genotype has additional mutations that are shared by another sample and thus place both of them into a subgroup of T1, thereby suggesting that the loss of 16189C in this single genotype is a comparatively recent back-mutation. Five of the T1 genotypes

lack the 16186T mutation: one of these can be similarly placed into a subgroup of T1, thereby indicating that its lack of 16186T is a recent back-mutation.

The other four genotypes that lack 16186T all share the 16243C mutation. Two of these four genotypes also happen to be the two genotypes in the T1 data set that lack the 16163G mutation. These four genotypes are slightly problematic in the sense that they pose an alternative theory regarding the possible evolution of the T1 haplogroup, namely that it could be that haplogroup T1 arose from just the three mutations 12633A, 16163G and 16189C, and that two subgroups thereafter developed: one which acquired the 16186T mutation (and which accounts for 105 of the 109 genotypes in our T1 data set, including one which we believe experienced a recent back-mutation), and the other which acquired the 16243C mutation (and which accounts for the remaining 4 samples).

A key reason for us to not favour this alternative phylogeny and its putative T1+16186T and T1+16243C sister subgroups is that its T1+16243C subgroup would pale in both size and genetic diversity when compared to its hypothetical T1+16186T sister. Moreover, the implied relative youth of the T1+16243C subgroup would in turn mean that a reservoir of pure T1 (i.e., having the three mutations 12633A, 16163G and 16189C but lacking 16186T) had to exist at least until the relatively recent birth of the T1+16243C subgroup. Such a reservoir would have had substantial opportunity to spawn additional subgroups as well, but we encountered no evidence of any.

In time it may be that additional genotypes that support this alternative will emerge. But for the time being the available data indicate that the most likely scenario to have occurred is that haplogroup T1 arose from the four stated mutations, and that the four T1 genotypes that have 16243C but lack 16186T belong together in a subgroup that includes a back-mutation as part of its motif (the specific subgroup in question is named T1f).

Referring now to Figure 1 and Table 6, we see that the genotype for node 58 has no additional mutations of relevance beyond the four defining mutations for haplogroup T1. Hence node 58 represents the root of the T1 subgroup.

Node 58 is connected to the large cluster that is spreading out from node 29 by a single path, along which the mutations, sequentially ordered, are 152C, 9899C and 195C. However, the 152C mutation is recurrent within the T1 data set (and, as we shall

*Within the mathematical field of Graph Theory, a graph consists of a set V of vertices (also called nodes) and an accompanying set E of edges such that each edge consists of a pair of two nodes. Graphs have numerous applications as network models.

see later, also in the other data sets), so rather than allow 152C to define any phylogenetic branching points, we instead include it within a subgroup motif only when it is accompanied by one or more other mutations. We thus define the T1a haplogroup to be based on the combination of 152C and 9899C, and T1a1 to be based on the additional presence of the 195C mutation.

As a general rule of thumb, we will not define subgroups based on only a single observed instance of a mutation, although we will occasionally define intermediate subgroups based on a single mutation that is part of a sequence of mutations. For example, when considering node 22, we will not define a subgroup based on its 12406A mutation since node 22 only represents a single observed genotype. Although mutation 12406A occurs twice in our data set, the other occurrence (in node 51) cannot be reconciled with that of node 22 (i.e., these two instances of 12406A are due to two independent mutations at 12406). The genotype for node 22 will therefore be classified as belonging to haplogroup T1a1 but not yet to any subgroup of T1a1. In contrast, we will define a subgroup based on the 469T mutation that leads toward node 7 and then the 8974T mutation that further leads to node 6. Node 5's mutation of 10915C will also define a nested subgroup on account of node 5 representing more than a single genotype of our data set.

The cluster of nodes surrounding node 29 now gives rise to several subgroups of T1a1:

T1a: 152C, 9899C
 T1a1: 195C
 T1a1a: 6445T
 T1a1b: 15467G
 T1a1c: 9120G, 15965G
 T1a1c1: 16213A, 16258G
 T1a1d: 6891G, 12182G, 16362C
 T1a1e: 5558G
 T1a1e1: 5414G
 T1a1e1a: 15412C
 T1a1f: 16304C
 T1a1g: 5478T
 T1a1h: 14758G
 T1a1i: 3308G, 11944C
 T1a1i1: 152(back-mutation), 524.1C, 524.2A, 11440A
 T1a1j: 8530G
 T1a1k: 469T
 T1a1k1: 8974T
 T1a1k1a: 10915C
 T1a1l: 16311C

Node 37 and several nearby nodes (i.e., nodes 20, 21, 31 and 32) appear as though they may belong to a sister subgroup of T1a1 that branches away from

node 38 via the 10143A mutation. However, nodes 20, 21, 31 and 32 all have the 195C mutation and therefore can be deemed to belong within haplogroup T1a1. What we propose is to define T1a1m to encompass nodes 20, 21, 31 and 32 as follows, and to leave node 37 (which represents only a single genotype) to fall within T1a but without any further specification.

T1a1m: 10143A, 14281T

We now turn our attention to the four small clusters in the top right quadrant of Figure 1. Since they all lack the 9899C mutation, they do not belong within T1a and hence now become named as sister haplogroups of T1a:

T1b: 152C, 384G, 4959A, 5558G, 9300A, 16261T
 T1c: 152C, 195C, 7258C, 10321C
 T1d: 152C, 7001G, 16263C
 T1e: 199C, 512G, 7784G, 14500G, 16274A

Nodes 47, 55 and 56 all share the 16243C mutation as well as a back-mutation at nucleotide 16186; this combination we put forward for the T1f subgroup. Within T1f we can now identify two subgroups: T1f1 based upon node 47 (which has the 1542C mutation) and T1f2 based upon nodes 55 and 56 and their shared motif of 11647T, 16163(back-mutation), 16183-. Additional subgroups of haplogroup T1 that can now be defined are included in the following list:

T1f: 16186(back-mutation), 16243C
 T1f1: 1542C
 T1f2: 11647T, 16163(back-mutation), 16183-
 T1g: 6152C
 T1h: 6656T
 T1i: 3834A, 8701G
 T1j: 16362C
 T1k: 3867T, 10376G
 T1l: 89C, 91T, 97A, 8083T, 8412C, 13759A, 13791T, 14284T, 16244A

Note that subgroups T1e, T1g and T1j each contain some nodes that have the 152C mutation and others that do not. We therefore chose to omit 152C from the definition of these groups.

The T1l subgroup is based upon node 59, which in turn represents two of our original 445 genotypes. Node 59 is isolated in the graph shown in Figure 1, indicating that no other genotypes within our data set were four or fewer mutations away. At this time we are unable to offer any insight regarding the phylogenetic order in which the nine mutations that correspond to this subgroup might have occurred.

The T2b Data Set

The T2b data set was chosen in such a way that all of its samples correspond to the T2b CR-based haplogroup, as they nearly all contain the two requisite mutations for haplogroup T2 (i.e., 11812G and 14233G) as well as the two requisite mutations for T2b (i.e., 930A and 5147A). To briefly comment on the four samples that do not have all four of these mutations, they each lack only one of these four mutations and so are still a good fit for haplogroup T2b.

Past studies of haplogroup T2 have noted an apparent instability with nucleotide 16296 and opted to disregard it when describing its phylogenetic structure.^{7,10} Referring to the T2b graph described in Figure 2 and Table 7, we observe that the 16296T mutation separates nodes 119 and 120 from each other as well as node pairs (9,10), (18, 19), (20, 21) and (75, 76). It also contributes to separating node 87 from node 109, node 31 from nodes 32 and 33, as well as node 82 from nodes 3, 78 and 83. Moreover, among the 205 samples that correspond to the T2b data set, one exhibits heteroplasmy at position 16296.

Nucleotide 152, which we saw to be recurrent within the T1 data set, is especially so within the T2b data set. It is responsible for the edges between node pairs (39, 95), (40, 100) and (45, 119) as well as the separation of node 5 from node 11, and nodes 13, 14, 15 and 16 from others that have the 151T and 10750G mutations (such as node 24). We also have a sample in our T2b data set that exhibits heteroplasmy at position 152.

This confounding behaviour on the part of nucleotides 16296 and 152 requires that we treat them differently than others. As was done with nucleotide 152 within the T1 haplogroup, we will adopt the convention that mutations at 16296 and 152 will only be included in haplogroup definitions if they are accompanied by a companion mutation. It should therefore be noted that if somebody possesses a genotype that would, were it not for their status at 16296 or 152, satisfy one of the subgroup definitions that we propose then it would not be unreasonable to nevertheless assume membership within the subgroup.

To begin to devise a FMS-based hierarchy for the T2b haplogroup and its subgroups, we must now decide upon the T2b motif. Mutations 930A and 5147A clearly belong as part of the T2b definition. But 16304C is also very common in the T2b data set (it is present in 185 of the 205 samples) and so we must decide whether to include or exclude 16304C

as part of the T2b haplogroup definition. Likewise, 16296T occurs in 151 of the T2b genotypes and warrants some consideration as well.

Looking at Figure 2, the only nodes in which the mutation 16304C is absent are nodes 30, 41, 42, 67, 81, 86, 102, 103, 104, 106, 121 and 122. Node 121 lacks additional mutations of relevance and is therefore a candidate that could be designated as the root node for haplogroup T2b; such an ancestral genotype would have some descendants that acquired the 16304C mutation while other descendants remained free of it. However, we argue that all current members of haplogroup T2b once included the 16304C mutation at some point in their past, even those that currently lack the mutation. For instance, node 81 is part of a cluster of three nodes that all contain the 634C mutation, and, moreover this trio is part of a larger cluster of nodes that all contain the 16362C mutation, so it is evident that node 81's genotype arose when a back-mutation at nucleotide 16304 occurred in an ancestral genotype (i.e., within the ancestral genotype represented by node 80).

The ten nodes other than 81 and 121 that lack 16304C all contain the 11242G mutation and can be argued to descend from the genotype that is represented by node 105. This hypothesis is supported not only by the single evolutionary pathway that is suggested by Figure 2 but also by conducting a comparison of the estimated age of this 11242G cluster to others in the graph.

Various research studies into human mtDNA mutation rates have calculated coding region divergence rates that vary from 0.019 to 0.041 (the divergence rate is twice the mutation rate, which is measured as the number of mutations per base pair per million years).^{27,33,44-47} Regardless of what the coding region mutation rate might actually be, we can perform a comparison of the relative ages of clusters to each other by directly measuring the divergence that can be observed within the coding region portion of their constituent genotypes.

When we consider the genetic distance that occurs between nucleotide positions 600 and 16000 (inclusive) for each pair of the 18 original genotypes* that gave rise to nodes 30, 41, 42, 67, 86, 102, 103, 104, 106 and 122, we find that the average genetic distance between the 153 pairs of genotypes is 2.516, which would indicate the cluster's age to be approximately 5400 years when using an estimated divergence rate of 0.030. However, when attempting to measure age it is more meaningful to compare the two most

*For these calculations we augmented the original genotypes slightly, to eliminate inflation of genetic distances that would have resulted from cases in which several mutations had occurred in tandem, such as base pair deletions starting at position 8281 and as well as an insertion of several bases subsequent to position 5899.

divergent genotypes; in this case the maximum genetic distance of 6 would suggest an age of approximately 13000 years. In contrast, the 185 genotypes that each have the 16304C mutation yield a maximum genetic distance of 14 (and an average of 4.308). As additional bases for comparison, the subset of 27 genotypes that have the 9254G mutation (but excluding the genotype that gave rise to node 79, since it belongs to a subclade having the 16362C and 634C mutations) has a maximum genetic distance of 8 (and an average of 1.766), and the 13 genotypes that have the 3826C mutation have a maximum genetic distance of 5 (and an average of 2.000).

These observations suggest that the 11242G cluster that consists of nearly every genotype that lacks 16304C is of a similar age to the 9254G and 3826C clusters, all three of which are measurably younger than the 16304C collection. Moreover, the age of the 16304C collection is barely distinguishable from the whole of the T2b data set (which has an average genetic distance of 4.317 and a maximum of 14). In particular, these observations do not support the competing hypothesis that the 11242G cluster evolved parallel to, and as an independent sister haplogroup of the 16304C cluster, but rather they are consistent with the theory that the 11242G cluster descends from the older and larger 16304C cluster by way of the back-mutation represented by the edge joining nodes 105 and 106. We conclude that there is little evidence to suggest that the T2b haplogroup arose from the mutations 930A and 5147A without also being accompanied by the 16304C mutation. We will therefore include 16304C within the FMS-based definition of haplogroup T2b.

We now consider whether it is reasonable to also include 16296T within the FMS-based definition of T2b, and in particular whether those nodes whose genotypes lack 16296T can be reasonably explained. Several nodes, such as 10, 19, 21, 31, 76, 109 and 120 are all easily attributed to a recent polymorphism that currently is not phylogenetically relevant (indeed, node 19 arises from a genotype that exhibits heteroplasmy at nucleotide 16296). The absence of 16296T from nodes 107, 116 and 121 may also be due to a recent back-mutation. Of the remaining nodes that lack 16296T, they do cluster with phylogenetical significance. For instance, nodes 27, 43 and 44 all share the 14836G mutation that we will soon use to define a subgroup (T2b21) for their small cluster. Every node in the 9254G cluster that we mentioned earlier lacks the 16296T mutation, and a phylogenetically identifiable portion of the 11242G cluster also lacks the 16296T mutation. It is there-

fore not unreasonable to conclude that the 16296T mutation should be included within the definition for the T2b haplogroup, and that back-mutations be attributed to each of the 14836G, 9254G and 11242G clusters.

We can now delineate several subgroups of T2b:

T2: 11812G, 14233G
 T2b: 930A, 5147A, 16304C, 16296T
 T2b1: 14016A
 T2b1a: 195C, 6530G, 11377A
 T2b2: 11242G, 16304(back-mutation)
 T2b2a: 4225G, 7754A, 14693G
 T2b2b: 16192T, 16296(back-mutation)
 T2b2b1: 12171G
 T2b2b1a: 15848G
 T2b2c: 152C, 16296(back-mutation)
 T2b3: 10750G
 T2b3a: 151T
 T2b3a1: 16187T
 T2b3a1a: 9809G, 11047A
 T2b3a2: 152C, 4561C
 T2b3a3: 152C, 5656G, 16292T
 T2b3a4: 6305A, 7150C, 7289G, 11887A, 12360T, 13623T
 T2b3b: 13722G
 T2b4: 9254G, 16296(back-mutation)
 T2b4a: 16172C
 T2b4a1: 114T
 T2b4b: 152C, 16104T
 T2b4c: 152C, 16239T
 T2b4d: 152C, 9653C, 16294(back-mutation)
 T2b4e: 6261A, 16192T, 16207G, 16274A
 T2b5: 3826C
 T2b5a: 5201C
 T2b5a1: 8504C
 T2b5a1a: 3C
 T2b6: 458T, 1709A, 9300A, 11533T
 T2b6a: 12007A
 T2b6b: 146C, 8730G, 14016A, 16218T, 16287T
 T2b7: 152C, 9180G, 9966A, 13768C, 16257–
 T2b8: 321C
 T2b9: 150T
 T2b9a: 5580C, 10559G
 T2b10: 237G, 2356G, 11380G, 13803G
 T2b11: 207A, 3398C
 T2b12: 189G, 16183–
 T2b13: 14861A
 T2b14: 15670C
 T2b15: 5836G, 8281-8289(9bp deletion)
 T2b16: 16362C
 T2b16a: 634C
 T2b17: 13692T
 T2b17a: 4688C, 7891T
 T2b18: 146C
 T2b19: 522–, 523–
 T2b19a: 13928C
 T2b19a1: 6126G
 T2b20: 15172A

T2b21: 152C, 14836G, 16296(back-mutation)
 T2b22: 152C, 3820T, 12223G, 13500C
 T2b23: 12441C, 16147T, 16224C, 16297C
 T2b23a: 5315G, 9224C, 11812(back-mutation),
 14314G, 16362C

Occasionally there are genotypes that do not fall into any subgroups of haplogroup T2b (or, to be more precise, not into any subgroups that can yet be identified). However, there are also a few genotypes that possess the motifs for multiple subgroups. For instance, the five genotypes that comprise node 6 all have the 14016A mutation that defines T2b1. However, they also all have the collection of four mutations that define T2b6. In this case we have deemed the genotypes to belong to T2b6 and have created the T2b6b subgroup based on mutations shared by these five genotypes (including 14016A). Similarly, node 5 is deemed to belong to haplogroup T2b9a rather than T2b18, nodes 38 and 88 respectively belong to T2b4 and T2b23 rather than T2b16, and nodes 66 and 67 respectively belong to T2b4a and T2b2b1 rather than T2b19.

The T2(-b) Data Set

The T2(-b) graph shown in Figure 3 and accompanied by Table 8 was constructed from the 131 genotypes that lack both the 12633A mutation that partially defines haplogroup T1 as well as the 930A mutation that partially defines T2b. It would include any genotypes that belong to neither T1 nor T2, but no such genotypes are present in our data set; all of the T2(-b) samples contain the 14233G mutation for T2 and all but one of them contain the companion 11812G mutation that also contributes to T2 (this single genotype that lacks 11812G has mutations that will soon be used to define the T2e4 subgroup and hence it is evident that it also belongs within haplogroup T2). Since all of the genotypes in the T2(-b) data set fall into the T2 haplogroup, we reiterate the defining motif for haplogroup T2:

T2: 11812G, 14233G

We will continue to give nucleotides 16296 and 152 special treatment, whereby they will not be afforded independent authority to define any subgroups.

Shown near the top right of Figure 3 is a sizeable cluster of nodes that all contain the 13965C mutation. This gives rise to the T2a haplogroup and several subgroups as follows:

T2a: 13965C
 T2a1: 14687G, 16296T
 T2a1a: 2850C, 7022C
 T2a1a1: 143A, 8715C, 8994A
 T2a1a1a: 13708A
 T2a1a2: 4688C
 T2a1a3: 4808T, 5498G
 T2a1a3a: 8435G
 T2a1a4: 4931T, 16296(back-mutation)
 T2a1b: 2141C, 9117C, 13966G
 T2a1b1: 16324C
 T2a1b1a: 12741T
 T2a1b1a1: 3350C
 T2a2: 195C, 198T, 13020C

In the bottom left region of the graph are several nodes that share the mutations 6261A, 10822T and 16292T. Two exceptions are node 76 which lacks 16292T and node 98 which lacks 6261A, but as they both contain two of these three mutations, we will incorporate all three together as the defining motif for haplogroup T2c:

T2c: 6261A, 10822T, 16292T
 T2c1: 8455T, 13973T
 T2c1a: 152C, 499A, 6998T, 8838A, 11914A,
 16296T
 T2c1a1: 15747C
 T2c2: 146C
 T2c2a: 279C, 5187T, 7873T
 T2c2a1: 152C, 7679C, 15784C
 T2c2a2: 11914A
 T2c2b: 522-, 523-
 T2c2b1: 16438A

The small cluster consisting of nodes 40, 41, 51 and 52 all share the mutations 152C, 5747G, 13260C and 13708A. Since 13260C is also present in the genotype for node 94, we define haplogroup T2d as follows:

T2d: 13260C, 16296T
 T2d1: 152C, 5747G, 13708A
 T2d1a: 194T
 T2d1b: 16086C

The genotypes for the cluster in the upper left corner of Figure 3 all share the 150T and 16153A mutations, which we use to define T2e:

T2e: 150T, 16153A
 T2e1: 41T, 16296T
 T2e1a: 2308G
 T2e1b: 200G
 T2e1c: 16092C
 T2e1d: 16207T
 T2e2: 9139A
 T2e3: 9947A
 T2e4: 16189C

Node 33 represents the single genotype in the T2(-b) data set that lacks the 11812G mutation. But given that this genotype has the 14233G mutation as well as the two mutations that define haplogroup T2e, it appears that this one sample has experienced a back-mutation at nucleotide 11812.

Except for node 60 (which lacks the 9 base pair deletion at nucleotide 8281), the genotypes comprising the cluster in the lower right corner all have this 9 base pair deletion as well as the 16189C mutation. This cluster gives rise to T2f:

T2f: 8281-8289(9bp deletion), 16189C, 16296T
 T2f1: 195C, 5277C, 5426C, 6489A, 15028A,
 15043A, 16298C
 T2f1a: 16182C, 16183C
 T2f1a1: 8270T
 T2f2: 8270T

Centred at the bottom of the figure is a cluster of nodes that all contain 14798C:

T2g: 14798C, 16296T
 T2g1: 1977C, 3834A, 14839G
 T2g1a: 200G
 T2g1a1: 16148T
 T2g2: 507C, 1766C, 7337A, 13834G

The only remaining haplogroups that can be identified at this point are the following two, one based on nodes 91 and 93 and the other on node 53:

T2h: 12397G, 16296T
 T2i: 152C, 10750G

Incidentally it is one of the two genotypes in the T2i haplogroup that is the genotype that was previously mentioned as having heteroplasmy at nucleotide 5147.

Discussion

At the outset of our analysis, we opted to include in our data set six genotypes that had been questioned by Yao et al.⁴¹ Looking to see what impact these genotypes may have had reveals that the mutations that Yao et al. had suspected to be in error did not contribute to the definition of any subgroups, and thus we are satisfied that our decision to include these genotypes did not have any adverse affects on our phylogenetic analysis. With the benefit of our phylogenetic analysis, it is interesting to now observe that the genotype having GenBank Accession Number EF660978 (reported by Gasparre et al.²⁴) belongs to subgroup T2b2b, which suggests that Yao et al. were mistaken when they reported EF660978's lack of 16304C as a probable error.

Each of the 445 genotypes that we had at our disposal was found to belong to one of the two FMS-defined subgroups T1 and T2. So not only did we have no cause to define subgroups with names such as T3, T4, etc., within our FMS-based phylogeny, but we found no examples of genotypes that were classified as T* (i.e., belonging to haplogroup T but not to either of its two known subgroups).

In 1998, Richards et al. estimated the age of haplogroup T to be at least 46500 years, the age of the T1 subgroup to be about 9000 years, and the remainder of T (i.e., all but T1) to be about 32000 years.⁶ We now present a collection of age approximations for haplogroup T and several of its subgroups based on the genotype diversity that we can observe in our data set. We use the same principles that we employed when delineating the phylogenetic structure of haplogroup T2b, which is to say that we consider the maximum genetic distances observed within the coding region portion of the genotypes (between nucleotide positions 600 and 16000), we use an estimated divergence rate of 0.030, and we treat cases of multiple deletions at 8281 or insertions at 5899 as single mutational events.

With 445 genotypes in our data set, there are 98790 pairs of genotypes. The maximum coding region genetic distance of these pairs is 23, which is found to occur twice. Both instances of this maximum involve the genotype having GenBank Accession Number DQ437577 (reported by Tuo et al.³⁹), which is one of the genotypes that Yao et al. had reason to question because it lacks the mutations 750G, 4917G, 8697A and 10463C that are expected to be present within haplogroup T genotypes. Moreover, DQ437577 is reported as having additional mutations not yet observed in other genotypes from haplogroup T. The respective absence or presence of these mutations was unique to DQ437577 in the T2(-b) data set and were therefore excluded from phylogenetic consideration when we assigned DQ437577 to subgroup T2d1a. Although the mutations reported for DQ437577 did not adversely affect our phylogenetic analysis, we now find that its mutations (particularly the four that are unexpectedly absent and truly appear to be in error) may cause genetic distances to be inflated.

We therefore ignore genetic distances involving DQ437577 and consequently find that the maximum coding region genetic distance in our data set is 22, which suggests that haplogroup T may be on the order of 47500 years old. This maximum genetic distance of 22 occurs only once; the full genotypes of the corresponding samples are shown in Table 4 along with their designated haplogroups.

<p>T1a1j sample: 73G, 152C, 195C, 263G, 309.1C, 315.1C, 709A, 750G, 1438G, 1888A, 2706G, 4216C, 4769G, 4917G, 5231A, 5585A, 5839T, 5899.1C, 7028T, 8530G, 8697A, 8860G, 9025R, 9899C, 10463C, 11251G, 11719A, 12633A, 13145A, 13368A, 14766T, 14905A, 15326G, 15452A, 15607G, 15928A, 16126C, 16163G, 16186T, 16189C, 16219G, 16294T, 16519C</p>
<p>T2a1b1a sample: 73G, 146C, 152C, 263G, 309.1C, 315.1C, 709A, 750G, 1420C, 1438G, 1888A, 2141C, 2706G, 4216C, 4769G, 4917G, 6249A, 6524C, 7028T, 8697A, 8860G, 9117C, 10463C, 11251G, 11719A, 11812G, 12741T, 13368A, 13965C, 13966G, 14233G, 14544A, 14687G, 14766T, 14905A, 15326G, 15452A, 15607G, 15884A, 15928A, 16126C, 16294T, 16296T, 16324C, 16519C</p>

Table 4: Genotypes with a CR GD of 22

By identifying the most divergent pair(s) of genotypes within each of several subgroups, we are able to calculate an approximation for the age of each. These estimates (rounded to the nearest multiple of 500 years) are listed in Table 5. These estimates continue to disregard genetic distances involving DQ437577. Of the other five genotypes that were questioned by Yao et al. and which we included in our data set, none were found to have inflated any of the genetic distances reported in Table 5.

It is interesting to observe that several upper-level haplogroups such as T2a, T2c and T2d have age estimates that are similar to those of their sister T2b, despite having far fewer samples than we have for T2b. It is also interesting to observe that even some very weakly sampled subgroups, such as T2b18, T2b21 and T2b23, have sufficient divergence within our data set to indicate that they too are quite old; in contrast, some subgroups (such as T1e and T2b15) with more samples have no coding region diversity whatsoever.

Out of concern that some subgroups, particularly those with little or no coding region diversity, might have inadvertently arisen from the participation of closely related individuals within our project (that is, relationships of which we were unaware), we performed a subsequent investigation into each subgroup. Most were easily seen to have sufficient diversity of geographical origins or variation in their genetic data (for instance, in the hyper-variable region) for us to be confident that the subgroups were not artificial.

A few, however, were not so easily dismissed and so we briefly comment on these now. Subgroup T2b2a corresponds to node 86 in Figure 2; although there is no apparent relationship between the three individuals represented by node 86, their FMS genotypes are identical and they have matrilineal origins

Haplogroup	No. of Samples	Max. GD	Approx. Age
T	445	22	47500
T1	109	15	32500
T1a	72	12	26000
T1b	3	5	11000
T1c	3	2	4500
T1d	2	5	11000
T1e	5	0	—
T1f	4	3	6500
T1g	3	9	19500
T1h	2	2	4500
T1i	2	0	—
T1j	2	3	6500
T1k	2	1	2000
T1l	2	2	4500
T2	336	21	45500
T2a	39	15	32500
T2b	205	14	30500
T2b1	8	6	13000
T2b2	18	6	13000
T2b3	21	12	26000
T2b4	27	8	17500
T2b5	13	5	11000
T2b6	10	6	13000
T2b7	3	2	4500
T2b8	2	3	6500
T2b9	3	5	11000
T2b10	2	0	—
T2b11	2	1	2000
T2b12	2	2	4500
T2b13	4	2	4500
T2b14	2	1	2000
T2b15	5	0	—
T2b16	9	6	13000
T2b17	3	3	6500
T2b18	3	9	19500
T2b19	9	4	8500
T2b20	2	0	—
T2b21	3	9	19500
T2b22	2	1	2000
T2b23	3	10	21500
T2c	24	14	30500
T2d	5	14	30500
T2e	30	12	26000
T2f	17	13	28000
T2g	9	8	17500
T2h	2	5	11000
T2i	2	4	8500

Table 5: Estimated Haplogroup Ages

in Québec and France. Subgroup T2b23a, which corresponds to node 88 in Figure 2, is based on two individuals whose identical FMS results were reported by Pichler et al. in a study of Hutterites.³⁶ For a few other subgroups (namely T1a1g, T1e, T2a1a1a and T2c2a1), the maximum FMS genetic distance was 1

(which suggests that they each had at least one pair of individuals who are not closely related) but they nevertheless exhibited some potential geographical and/or cultural confinement. For instance, the three genotypes in T1a1g all have Finnish origins, as reported by Finnilä et al. and Moilanen et al.^{21,22} The five genotypes comprising T1e were all collected in the El-Hayez oasis of Egypt, as reported by Kujanová et al.²⁹ The three genotypes in T2a1a1a are all from the Hutterite study conducted by Pichler et al.³⁶ And the three genotypes in T2c2a1 are all from a Sardinian study conducted by Fraumene et al.²³

One of the motivating factors for our research project was the disagreement in major subgroup names when comparing HVR-defined subgroups with CR-defined ones, and in particular, the change in nomenclature that affected those individuals within the T3, T4 and T5 HVR-based haplogroups when they ceased to remain in subgroups named T3, T4 or T5 upon having their mtDNA genomes fully sequenced. So we now devote some attention to HVR-based subgroups and their status in light of the FMS-based phylogeny that we have now delineated.

Recall that haplogroup T3 had been based upon the presence of the 16292T mutation. Within our FMS-based phylogeny, this mutation appears within the defining motifs of the T2b3a3 and T2c subgroups. Of the 28 samples in our overall data set that have the 16292T mutation, four of them now qualify as T2b3a3, 22 are T2c, and the remaining two are found in the T1a and T2b18 subgroups. While the presence of the 16292T mutation favours membership in T2c, it is evident that 16292T alone is not sufficient to be certain of one's subgroup placement.

Haplogroup T4 had been based upon the 16324C mutation, which we found in only eight of our 445 samples. The 16324C mutation appears in just one of our defined subgroups (namely haplogroup T2a1b1), and seven of the eight samples now find themselves in T2a1b1; the eighth sample is not far away, in T2a1. So while we find that the 16324C mutation does seem to imply membership in the T2a1 haplogroup, it does not guarantee placement within the deeper T2a1b1 subgroup.

Mutation 16153A, which was the defining mutation for haplogroup T5, occurs in 30 of the samples of our data set, all of which now find themselves within the FMS-based T2e haplogroup. In this case there is perfect correspondence between the presence of the 16153A mutation and membership in T2e or one of its subgroups.

Regarding other HVR-based haplogroups, a previous HVR-based study of haplogroup T concluded that the 16189C mutation ought to be the defining

mutation for HVR-based haplogroup T1.¹⁰ This same study put forward a definition for the HVR-based T1c subgroup consisting of the three additional mutations of 16182C, 16183C and 16298C. With the benefit of now being able to take FMS results into consideration, it is evident that several genotypes that have the 16189C mutation and which would therefore have appeared to belong to the HVR-based T1 haplogroup are in fact members of the FMS-based T2 haplogroup. In particular, our data set included seven genotypes that each have the four HVR mutations 16182C, 16183C, 16189C and 16298C that would imply membership in the HVR-based T1c haplogroup; these seven genotypes all belong to the FMS-based T2f1a haplogroup. Moreover, ten other genotypes that have the 16189C mutation make up the remainder of the T2f haplogroup. It was only by seeing these samples' FMS results and realising that they have the mutations 11812G and 14233G for haplogroup T2, rather than T1's coding region mutation of 12633A, that it became possible to determine their proper phylogenetic classification.

HVR-based haplogroups such as T1c and T3 serve as compelling examples to show that haplogroup prediction based on HVR data alone can be misleading. Moreover, since many of the FMS-based subgroups that we have been able to identify have definitions that are solely based on coding region mutations, it is imperative that anybody desiring to determine their place within the phylogenetic structure of haplogroup T with both accuracy and precision should have their mtDNA genome fully sequenced.

Closing Remarks

It warrants mention that the phylogeny that we have developed in this paper only represents the state of current knowledge. Even with 445 samples at our disposal, this is still only a small collection and we are sure to have been unable to observe some subgroups of haplogroup T. Likewise, even if a mutation that correctly defines a subgroup is present in our data set, we might have had too few instances of it to be able to predict its true phylogenetic role. Nevertheless, we are pleased to have been able to build a phylogeny for haplogroup T that we hope will serve as a solid foundation upon which the academic and genetic genealogy communities can build and make further refinements.

Acknowledgements

Thanks are extended to Bennett Greenspan, Eileen Krause Murphy and the rest of the staff at the Family

Tree DNA Genomics Research Center, for their support and cooperation throughout the course of this project.

The task of searching for Haplogroup T entries in GenBank was aided by the Human Mitochondrial Genome Database⁴⁸ as well as online resources provided by Ian Logan⁴⁹ and Ron Scott.⁵⁰

The task of drawing the graphs in Figures 1, 2 and 3 was aided by Pajek.⁵¹

David Pike acknowledges research support from CFI, IRIF and NSERC.

Last, but definitely not least, we wish to express our deeply felt gratitude to the 301 individuals who made this research possible by sharing their full mitochondrial sequences with us.

References

1. Richards, M., Côrte-Real, M., Forster, P., Macauley, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.-J., and Sykes, B. (1996). Paleolithic and Neolithic Lineages in the European Mitochondrial Gene Pool. *Am. J. Hum. Genet.*, 59, 185–203.
2. Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M. L., and Wallace, D. C. (1996). Classification of European mtDNAs From an Analysis of Three European Populations. *Genetics*, 144, 1835–1850.
3. Torroni, A., Lott, M. T., Cabell, M. F., Chen, Y.-S., Lavergne, L., and Wallace, D. C. (1994). mtDNA and the Origin of Caucasians: Identification of Ancient Caucasian-specific Haplogroups, One of Which is Prone to a Recurrent Somatic Duplication in the D-Loop Region. *Am. J. Hum. Genet.*, 55, 760–776.
4. Macauley, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonn -Tamir, B., Sykes, B., and Torroni, A. (1999). The Emerging Tree of West Eurasian mtDNAs: A Synthesis of Control-Region Sequences and RFLPs. *Am. J. Hum. Genet.*, 64, 232–249.
5. Finnil , S. and Majamaa, K. (2001). Phylogenetic analysis of mtDNA haplogroup TJ in a Finnish population. *J. Hum. Genet.*, 46, 64–69.
6. Richards, M. B., Macauley, V. A., Bandelt, H.-J., and Sykes, B. C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.*, 62, 241–260.
7. Richards, M., Macauley, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., et al. (2000). Tracing European Founder Lineages in the Near Eastern mtDNA Pool. *Am. J. Hum. Genet.*, 67, 1251–1276.
8. Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., and Villems, R. (2004). Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. *Am. J. Hum. Genet.*, 75, 752–770.
9. Palanichamy, M., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y., Chaudhuri, T. K., Palla, V., and Zhang, Y.-P. (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, 966–978.
10. Pike, D. A. (2006). Phylogenetic Networks for the Human mtDNA Haplogroup T. *J. Genet. Geneal.*, 2, 1–11.
11. Tanaka, M., Cabrera, V. M., Gonz lez, A. M., Laruga, J. M., Takeyasu, T., Fuku, N., Guo, L.-J., Hirose, R., Fujita, Y., Kurata, M., et al. (2004). Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan. *Genome Res.*, 14, 1832–1850.
12. Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M., Anderson, C., Ghosh, S. S., Olefsky, J. M., Beal, M. F., Davis, R. E., et al. (2002). Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups. *Am. J. Hum. Genet.*, 70, 1152–1171.
13. Shen, P., Lavi, T., Kivisild, T., Chou, V., Sengun, D., Gefel, D., Shpirer, I., Woolf, E., Hillel, J., Feldman, M. W., et al. (2004). Reconstruction of Patrilineages and Matrilineages of Samaritans and Other Israeli Populations From Y-Chromosome and Mitochondrial DNA Sequence Variation. *Hum. Mutat.*, 24, 248–260.
14. Behar, D. M., Metspalu, E., Kivisild, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., et al. (2008). Counting the Founders: The Matrilineal Genetic Ancestry of the Jewish Diaspora. *PLoS ONE*, 3, e2062.

15. van Oven, M. and Kayser, M. (2008). Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Hum. Mutat.*, *30*, E386–E394.
16. van Oven, M. and Kayser, M. <http://www.phylotree.org>.
17. Coble, M. D., Just, R. S., O’Callaghan, J. E., Letmanyi, I. H., Peterson, C. T., Irwin, J. A., and Parsons, T. J. (2004). Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int. J. Legal Med.*, *118*, 137–146.
18. Costa, M., Cherni, L., Fernandes, V., Freitas, F., Ammar el Gaaied, A., and Pereira, L. (2009). Data from complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association and the ‘golden mean’ to longevity. *Mech. Ageing Dev.*, *130*, 222–226.
19. Detjen, A., Tinschert, S., Kaufmann, D., Algermissen, B., Nürnberg, P., and Schuelke, M. (2007). Analysis of mitochondrial DNA in discordant monozygotic twins with neurofibromatosis type 1. *Twin Res. Hum. Genet.*, *10*, 486–495.
20. Fendt, L., Zimmermann, B., Daniaux, M., and Parson, W. (2009). Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. *BMC Genomics*, *10*, 139.
21. Finnilä, S., Lehtonen, M. S., and Majamaa, K. (2001). Phylogenetic Network for European mtDNA. *Am. J. Hum. Genet.*, *68*, 1475–1484.
22. Moilanen, J. S., Finnilä, S., and Majamaa, K. (2003). Lineage-Specific Selection in Human mtDNA: Lack of Polymorphisms in a Segment of MTND5 Gene in Haplogroup J. *Mol. Biol. Evol.*, *20*, 2132–2142.
23. Fraumene, C., Belle, E. M. S., Castri, L., Sanna, S., Mancosu, G., Cosso, M., Marras, F., Barbujani, G., Pirastu, M., and Angius, A. (2006). High Resolution Analysis and Phylogenetic Network Construction Using Complete mtDNA Sequences in Sardinian Genetic Isolates. *Mol. Biol. Evol.*, *23*, 2101–2111.
24. Gasparre, G., Porcelli, A. M., Bonora, E., Pennisi, L. F., Toller, M., Iommarini, L., Ghelli, A., Moretti, M., Betts, C. M., Martinelli, G. N., et al. (2007). Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. *Proc. Natl. Acad. Sci. USA*, *104*, 9001–9006.
25. Ghelli, A., Porcelli, A. M., Zanna, C., Vidoni, S., Mattioli, S., Barbieri, A., Iommarini, L., Pala, M., Achilli, A., Torroni, A., et al. (2009). The background of mitochondrial DNA haplogroup J increases the sensitivity of Leber’s hereditary optic neuropathy cells to 2,5-hexanedione toxicity. *PLoS ONE*, *4*, e7922.
26. Hartmann, A., Thieme, M., Nanduri, L. K., Stempfl, T., Moehle, C., Kivisild, T., and Oefner, P. J. (2009). Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum. Mutat.*, *30*, 115–122.
27. Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, *408*, 708–713.
28. Ingman, M. and Gyllensten, U. (2007). Rate variation between mitochondrial domains and adaptive evolution in humans. *Human Molecular Genetics*, *16*, 2281–2287.
29. Kujanová, M., Pereira, L., Fernandes, V., Pereira, J. B., and Černý, V. (2009). Near Eastern Neolithic Genetic Input in a Small Oasis of the Egyptian Western Desert. *Am. J. Phys. Anthropol.*, *140*, 336–346.
30. La Morgia, C., Achilli, A., Iommarini, L., Barboni, P., Pala, M., Olivieri, A., Zanna, C., Vidoni, S., Tonon, C., Lodi, R., et al. (2008). Rare mtDNA variants in Leber hereditary optic neuropathy families with recurrence of myoclonus. *Neurology*, *70*, 762–770.
31. Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C., and Cabrera, V. M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics*, *2*, 13.
32. Malyarchuk, B., Derenko, M., Denisova, G., and Kravtsova, O. (2010). Mitogenomic diversity in Tatars from the Volga-Ural region of Russia. *Mol. Biol. Evol.*, *27*, 2220–2226.
33. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M. D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA*, *100*, 171–176.
34. Pello, R., Martín, M. A., Carelli, V., Nijtmans, L. G., Achilli, A., Pala, M., Torroni, A., Gómez-Durán, A., Ruiz-Pesini, E., Martinuzzi, A., et al.

- (2008). Mitochondrial DNA background modulates the assembly kinetics of OXPHOS complexes in a cellular model of mitochondrial disease. *Human Molecular Genetics*, 17, 4001–4011.
35. Pereira, L., Gonçalves, J., Franco-Duarte, R., Silva, J., Rocha, T., Arnold, C., Richards, M., and Macaulay, V. (2007). No Evidence for an mtDNA Role in Sperm Motility: Data from Complete Sequencing of Asthenozoospermic Males. *Mol. Biol. Evol.*, 24, 868–874.
 36. Pichler, I., Fuchsberger, C., Platzer, C., Çalişkan, M., Marroni, F., Pramstaller, P. P., and Ober, C. (2010). Drawing the history of the Hutterite population on a genetic landscape: inference from Y-chromosome and mtDNA genotypes. *Eur. J. Hum. Genet.*, 18, 463–470.
 37. Rani, D. S., Dhandapany, P. S., Nallari, P., Govindaraj, P., Singh, L., and Thangaraj, K. (2010). Mitochondrial DNA haplogroup R is associated with Noonan syndrome of South India. *Mitochondrion*, 10, 166–173.
 38. Rogaev, E., Grigorenko, A., Moliaka, Y., Faskhutdinova, G., Goltsov, A., Lahti, A., Hildebrandt, C., Kittler, E., and Morozova, I. (2009). Genomic identification in historical case of the Nicholas II royal family. *Proc. Natl. Acad. Sci. USA*, 106, 5258–5263.
 39. Tuo, Y., Hou, Q., and Li, S. (2006). Complete sequence of Mongol mitochondrial DNA. Unpublished.
 40. Zaragoza, M. V., Fass, J., Diegoli, M., Lin, D., and Arbustini, E. (2010). Mitochondrial DNA Variant Discovery and Evaluation in Human Cardiomyopathies through Next-Generation Sequencing. *PLoS ONE*, 5, e12295.
 41. Yao, Y.-G., Salas, A., Logan, I., and Bandelt, H.-J. (2009). mtDNA Data Mining in GenBank Needs Surveying. *Am. J. Hum. Genet.*, 85, 929–933.
 42. Lutz, S., Weisser, H.-J., Heizmann, J., and Pollak, S. (1999). Mitochondrial heteroplasmy among maternally related individuals. *Int. J. Legal Med.*, 113, 155–161.
 43. Lee, H. Y., Chung, U., Yoo, J.-E., Park, M. J., and Shin, K.-J. (2004). Quantitative and qualitative profiling of mitochondrial DNA length heteroplasmy. *Electrophoresis*, 25, 28–34.
 44. Tang, H., Siegmund, D. O., Shen, P., Oefner, P. J., and Feldman, M. W. (2002). Frequentist Estimation of Coalescence Times From Nucleotide Sequence Data Using a Tree-Based Partition. *Genetics*, 161, 447–459.
 45. Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtDNA Variation Predicts Population Size in Humans and Reveals a Major Southern Asian Chapter in Human Prehistory. *Mol. Biol. Evol.*, 25, 468–474.
 46. Ho, S. Y. and Endicott, P. (2008). The Crucial Role of Calibration in Molecular Date Estimates for the Peopling of the Americas. *Am. J. Hum. Genet.*, 83, 127–146.
 47. Henn, B. M., Gignoux, C. R., Feldman, M. W., and Mountain, J. L. (2009). Characterizing the Time Dependency of Human Mitochondrial DNA Mutation Rate Estimates. *Mol. Biol. Evol.*, 26, 217–230.
 48. Ingman, M. and Gyllensten, U. (2006). mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.*, 34, D749–D751.
 49. Logan, I. <http://www.ianlogan.co.uk/mtDNA.htm>.
 50. Scott, R. <http://freepages.genealogy.rootsweb.com/~ncscotts/>.
 51. Batagelj, V. and Mrvar, A. Pajek – Program for Large Network Analysis. Home page: <http://vlado.fmf.uni-lj.si/pub/networks/pajek>.

Table 6: Simplified T1 Genotypes

Node No.	No. of Samples	Genotype
1	1	73G 146C 152C 195C 315.1C 469T 750G 9899C 11914A 16163G 16186T 16519C
2	1	73G 146C 152C 195C 315.1C 750G 7258C 10321C 16163G 16186T 16519C
3	1	73G 146C 152C 195C 315.1C 750G 9899C 16163G 16186T 16256T 16519C
4	1	73G 152C 195C 315.1C 384G 750G 4959A 5558G 9300A 16163G 16186T 16261T 16519C
5	2	73G 152C 195C 315.1C 469T 750G 8974T 9899C 10915C 16163G 16186T 16519C
6	1	73G 152C 195C 315.1C 469T 750G 8974T 9899C 16163G 16186T 16519C
7	1	73G 152C 195C 315.1C 469T 750G 9899C 16163G 16186T 16519C
8	1	73G 152C 195C 315.1C 750G 3308G 9899C 11944C 16163G 16186T 16519C
9	2	73G 152C 195C 315.1C 750G 5414G 5558G 9899C 15412C 16163G 16186T 16519C
10	1	73G 152C 195C 315.1C 750G 5414G 5558G 9899C 16163G 16186T 16519C
11	1	73G 152C 195C 315.1C 750G 5478T 9899C 16163G 16186T 16519C
12	1	73G 152C 195C 315.1C 750G 5558G 9899C 16163G 16186T 16519C
13	2	73G 152C 195C 315.1C 750G 6445T 9899C 16163G 16186T 16519C
14	2	73G 152C 195C 315.1C 750G 6891G 9899C 12182G 16163G 16186T 16362C 16519C
15	1	73G 152C 195C 315.1C 750G 7258C 10321C 16163G 16186T 16242T 16519C
16	1	73G 152C 195C 315.1C 750G 7258C 10321C 16163G 16186T 16519C
17	4	73G 152C 195C 315.1C 750G 8530G 9899C 16163G 16186T 16519C
18	3	73G 152C 195C 315.1C 750G 9120G 9899C 15965G 16163G 16186T 16213A 16258G 16519C
19	1	73G 152C 195C 315.1C 750G 9120G 9899C 15965G 16163G 16186T 16519C
20	1	73G 152C 195C 315.1C 750G 9899C 10143A 14281T 16163G 16186T 16244A 16519C
21	2	73G 152C 195C 315.1C 750G 9899C 10143A 14281T 16163G 16186T 16519C
22	1	73G 152C 195C 315.1C 750G 9899C 12406A 16163G 16186T 16519C
23	2	73G 152C 195C 315.1C 750G 9899C 14758G 16163G 16186T 16519C
24	2	73G 152C 195C 315.1C 750G 9899C 15467G 16163G 16186T 16519C
25	1	73G 152C 195C 315.1C 750G 9899C 16163G 16172C 16186T 16519C
26	1	73G 152C 195C 315.1C 750G 9899C 16163G 16186T 16242T 16519C
27	2	73G 152C 195C 315.1C 750G 9899C 16163G 16186T 16304C 16519C
28	2	73G 152C 195C 315.1C 750G 9899C 16163G 16186T 16311C 16519C
29	27	73G 152C 195C 315.1C 750G 9899C 16163G 16186T 16519C
30	2	73G 152C 195C 750G 5478T 9899C 16163G 16186T 16519C
31	1	73G 152C 195C 750G 9899C 10143A 14281T 16163G 16186T 16519C
32	1	73G 152C 195C 9899C 10143A 14281T 16163G 16186T 16519C
33	2	73G 152C 199C 315.1C 512G 750G 7784G 14500G 16163G 16186T 16274A 16519C
34	2	73G 152C 315.1C 384G 750G 4959A 5558G 7853A 9300A 16163G 16186T 16261T 16519C
35	1	73G 152C 315.1C 750G 6152C 11914A 15412C 16163G 16186T
36	1	73G 152C 315.1C 750G 7001G 16163G 16186T 16263C 16311C 16519C
37	1	73G 152C 315.1C 750G 9899C 10143A 16163G 16186T 16519C
38	2	73G 152C 315.1C 750G 9899C 16163G 16186T 16519C
39	1	73G 152C 315.1C 750G 16129A 16163G 16172C 16186T 16362C 16519C
40	1	73G 152C 315.1C 750G 16163G 16186T 16519C
41	1	73G 152C 315.1C 7001G 16163G 16186T 16263C 16519C
42	1	73G 195C 315.1C 523.1C 523.2A 750G 3308G 9899C 11440A 11944C 16163G 16186T
43	1	73G 195C 315.1C 523.1C 523.2A 750G 3308G 9899C 11440A 11944C 16163G 16186T 16519C
44	3	73G 195C 315.1C 750G 9899C 16163G 16186T 16519C
45	3	73G 199C 315.1C 512G 750G 7784G 14500G 16163G 16186T 16274A 16519C
46	1	73G 315.1C 523.1C 523.2A 750G 3834A 8701G 16163G 16186T 16519C
47	2	73G 315.1C 750G 1542C 16163G 16243C 16519C
48	1	73G 315.1C 750G 3834A 8701G 16163G 16186T 16519C
49	1	73G 315.1C 750G 3867T 10376G 16163G 16186T 16519C

continued on next page

Node No.	No. of Samples	Genotype
50	1	73G 315.1C 750G 3867T 10376G 16163G 16519C
51	1	73G 315.1C 750G 6152C 12406A 16163G 16186T 16519C
52	1	73G 315.1C 750G 6152C 16129A 16163G 16186T
53	1	73G 315.1C 750G 6656T 16163G 16186T 16311C 16519C
54	1	73G 315.1C 750G 6656T 16163G 16186T 16519C
55	1	73G 315.1C 750G 11647T 16183- 16193.1C 16243C 16256T 16519C
56	1	73G 315.1C 750G 11647T 16183- 16193.1C 16243C 16519C
57	1	73G 315.1C 750G 16163G 16186T 16362C 16519C
58	1	73G 315.1C 750G 16163G 16186T 16519C
59	2	89C 91T 97A 152C 315.1C 750G 8083T 8412C 13759A 13791T 14284T 16163G 16186T 16244A 16519C

Table 7: Simplified T2b Genotypes

Node No.	No. of Samples	Genotype
1	1	3C 263G 315.1C 522- 523- 709A 3826C 8697A 11812G 16294T 16296T 16304C 16519C
2	1	3C 263G 315.1C 523.1C 523.2A 709A 3826C 5201C 8504C 8697A 11812G 16294T 16296T 16304C 16519C
3	1	3C 263G 315.1C 709A 3826C 5201C 8504C 8697A 11812G 16294T 16296T 16304C 16519C
4	3	114T 263G 315.1C 709A 8697A 9254G 11812G 16172C 16294T 16304C 16519C
5	1	146C 150T 152C 263G 315.1C 709A 5580C 8697A 10559G 11812G 16294T 16296T 16304C
6	5	146C 263G 315.1C 458T 709A 1709A 8697A 8730G 9300A 11533T 11812G 14016A 16218T 16287T 16294T 16296T 16304C 16519C
7	1	146C 263G 315.1C 458T 709A 1709A 8697A 9300A 11533T 11812G 12007A 16294T 16296T 16304C 16519C
8	1	146C 263G 315.1C 709A 8697A 11812G 16292T 16296T 16304C 16311C 16519C
9	1	146C 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
10	1	146C 263G 315.1C 709A 8697A 11812G 16294T 16304C 16519C
11	1	150T 263G 315.1C 709A 5580C 8697A 10559G 11812G 16294T 16296T 16304C 16519C
12	1	150T 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
13	2	151T 152C 263G 315.1C 709A 4561C 8697A 10750G 11812G 16294T 16296T 16304C 16519C
14	1	151T 152C 263G 315.1C 709A 5656G 8697A 10750G 11812G 16292T 16294T 16296T 16304C 16519C
15	1	151T 152C 263G 315.1C 709A 8697A 10398G 10750G 11812G 16294T 16296T 16304C 16519C
16	3	151T 152C 263G 709A 5656G 8697A 10750G 11812G 16093C 16292T 16294T 16296T 16304C 16519C
17	1	151T 263G 315.1C 709A 3826C 8697A 11812G 16294T 16296T 16304C 16519C
18	1	151T 263G 315.1C 709A 6305A 7150C 7289G 8697A 10750G 11812G 11887A 12360T 13623T 16294T 16296T 16304C 16519C
19	1	151T 263G 315.1C 709A 6305A 7150C 7289G 8697A 10750G 11812G 11887A 12360T 13623T 16294T 16304C 16519C
20	1	151T 263G 315.1C 709A 8697A 9809G 10750G 11047A 11812G 16187T 16294T 16296T 16304C 16519C
21	1	151T 263G 315.1C 709A 8697A 9809G 10750G 11047A 11812G 16187T 16294T 16304C 16519C
22	1	151T 263G 315.1C 709A 8697A 10750G 11812G 14180C 15884A 16294T 16296T 16304C 16519C
23	1	151T 263G 315.1C 709A 8697A 10750G 11812G 16187T 16294T 16296T 16304C 16519C
24	1	151T 263G 315.1C 709A 8697A 10750G 11812G 16294T 16296T 16304C 16519C
25	1	151T 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
26	1	152C 195C 263G 315.1C 709A 8697A 11812G 14180C 16294T 16296T 16304C 16519C
27	1	152C 195C 263G 709A 11812G 14836G 16294T 16304C 16519C
28	1	152C 263G 315.1C 709A 1719A 3918A 8697A 9254G 10398G 11812G 16294T 16304C 16519C
29	2	152C 263G 315.1C 709A 3820T 8697A 11812G 12223G 13500C 16294T 16296T 16304C 16519C
30	1	152C 263G 315.1C 709A 3918A 8697A 11242G 11812G 16294T 16519C
31	1	152C 263G 315.1C 709A 8697A 9180G 9966A 11812G 13768C 16147T 16257- 16294T 16304C 16519C
32	1	152C 263G 315.1C 709A 8697A 9180G 9966A 11812G 13768C 16257- 16294T 16296T 16304C 16311C 16519C
33	1	152C 263G 315.1C 709A 8697A 9180G 9966A 11812G 13768C 16257- 16294T 16296T 16304C 16519C
34	1	152C 263G 315.1C 709A 8697A 9254G 9653C 11812G 16304C 16519C
35	4	152C 263G 315.1C 709A 8697A 9254G 11812G 16104T 16294T 16304C 16519C
36	2	152C 263G 315.1C 709A 8697A 9254G 11812G 16239T 16294T 16304C 16519C
37	1	152C 263G 315.1C 709A 8697A 9254G 11812G 16294T 16304C
38	1	152C 263G 315.1C 709A 8697A 9254G 11812G 16294T 16304C 16362C 16519C
39	5	152C 263G 315.1C 709A 8697A 9254G 11812G 16294T 16304C 16519C
40	3	152C 263G 315.1C 709A 8697A 10750G 11812G 13722G 16294T 16296T 16304C 16519C
41	1	152C 263G 315.1C 709A 8697A 11242G 11812G 16187T 16294T 16519C
42	5	152C 263G 315.1C 709A 8697A 11242G 11812G 16294T 16519C
43	1	152C 263G 315.1C 709A 8697A 11812G 14836G 16189C 16294T 16304C 16519C
44	1	152C 263G 315.1C 709A 8697A 11812G 14836G 16294T 16304C 16519C
45	2	152C 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
46	1	152C 263G 709A 8697A 9254G 9653C 11812G 16304C 16519C
47	1	189G 263G 315.1C 709A 8697A 11812G 14861A 16294T 16296T 16304C 16519C
48	2	189G 263G 315.1C 709A 8697A 11812G 16183- 16294T 16296T 16304C 16519C
49	1	189G 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C

continued on next page

Node No.	No. of Samples	Genotype
50	1	195C 263G 315.1C 321C 709A 8697A 11812G 16294T 16296T 16304C 16519C
51	2	195C 263G 315.1C 709A 6530G 8697A 11377A 11812G 14016A 16294T 16296T 16304C 16519C
52	1	195C 263G 315.1C 709A 8697A 9254G 11812G 16172C 16294T 16304C 16519C
53	1	195C 263G 315.1C 709A 8697A 10750G 11812G 13500C 16294T 16296T 16304C 16519C
54	1	195C 263G 315.1C 709A 8697A 11812G 14861A 16093C 16104T 16294T 16296T 16304C 16519C
55	1	195C 263G 315.1C 8697A 11812G 16294T 16296T 16304C 16519C
56	1	200G 263G 315.1C 709A 1719A 8697A 11812G 14861A 16296T 16304C 16519C
57	1	200G 263G 315.1C 709A 8697A 11812G 14016A 16294T 16296T 16304C 16519C
58	2	207A 263G 315.1C 709A 3398C 8697A 11812G 16294T 16296T 16304C 16519C
59	1	228A 263G 315.1C 522- 523- 709A 8697A 11812G 16294T 16296T 16304C 16519C
60	1	228A 263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
61	2	237G 263G 315.1C 709A 2356G 8697A 11380G 11812G 13803G 16294T 16296T 16304C 16519C
62	1	263G 315.1C 321C 709A 8697A 11812G 16294T 16296T 16304C 16519C
63	3	263G 315.1C 458T 709A 1709A 8697A 9300A 11533T 11812G 12007A 16294T 16296T 16304C 16519C
64	1	263G 315.1C 458T 709A 1709A 8697A 9300A 11533T 11812G 16294T 16296T 16304C 16519C
65	2	263G 315.1C 522- 523- 709A 6126G 8697A 11812G 13928C 16294T 16296T 16304C 16519C
66	1	263G 315.1C 522- 523- 709A 8697A 9254G 11812G 16172C 16294T 16304C 16519C
67	1	263G 315.1C 522- 523- 709A 8697A 11242G 11812G 12171G 16192T 16294T 16519C
68	1	263G 315.1C 522- 523- 709A 8697A 11812G 13928C 16248T 16294T 16296T 16304C 16519C
69	2	263G 315.1C 522- 523- 709A 8697A 11812G 13928C 16294T 16296T 16304C 16519C
70	2	263G 315.1C 522- 523- 709A 8697A 11812G 16294T 16296T 16304C 16519C
71	1	263G 315.1C 523.1C 523.2A 709A 3826C 8697A 11812G 16294T 16296T 16304C 16519C
72	4	263G 315.1C 523.1C 523.2A 709A 5836G 8281- 8697A 11812G 16294T 16296T 16304C 16519C
73	1	263G 315.1C 523.1C 523.2A 709A 8697A 11242G 11812G 16294T 16296T 16304C 16519C
74	1	263G 315.1C 523.1C 523.2A 709A 8697A 11812G 14016A 16294T 16296T 16304C 16519C
75	1	263G 315.1C 523.1C 523.2A 709A 8697A 11812G 16294T 16296T 16304C 16519C
76	1	263G 315.1C 523.1C 523.2A 709A 8697A 11812G 16294T 16304C 16519C
77	1	263G 315.1C 573.1C 709A 3826C 5201C 5836G 8504C 8697A 11812G 16294T 16296T 16304C 16519C
78	2	263G 315.1C 573.1C 709A 3826C 5201C 8504C 8697A 11812G 16294T 16296T 16304C 16519C
79	1	263G 315.1C 634C 709A 8697A 9254G 11812G 16294T 16296T 16304C 16362C 16519C
80	1	263G 315.1C 634C 709A 8697A 11812G 16294T 16296T 16304C 16362C 16519C
81	1	263G 315.1C 634C 709A 8697A 11812G 16294T 16296T 16362C 16519C
82	1	263G 315.1C 709A 3826C 5201C 8504C 8697A 11812G 16294T 16304C 16519C
83	1	263G 315.1C 709A 3826C 5201C 8697A 11812G 16294T 16296T 16304C 16519C
84	2	263G 315.1C 709A 3826C 8697A 11812G 16294T 16296T 16304C 16519C
85	1	263G 315.1C 709A 3826C 11812G 16294T 16296T 16304C 16519C
86	3	263G 315.1C 709A 4225G 7754A 8697A 11242G 11812G 14693G 16294T 16296T 16519C
87	2	263G 315.1C 709A 4688C 7891T 8697A 11812G 13692T 16294T 16296T 16304C 16519C
88	2	263G 315.1C 709A 5315G 8697A 9224C 12441C 14314G 16147T 16224C 16294T 16296T 16297C 16304C 16362C 16519C
89	1	263G 315.1C 709A 5836G 8281- 8697A 11812G 16294T 16296T 16304C 16519C
90	2	263G 315.1C 709A 6261A 8697A 9254G 11812G 16192T 16207G 16274A 16294T 16304C 16519C
91	1	263G 315.1C 709A 7289G 8697A 11812G 16294T 16296T 16304C 16519C
92	1	263G 315.1C 709A 8504C 8697A 11812G 16294T 16296T 16304C 16519C
93	1	263G 315.1C 709A 8697A 9254G 11812G 16172C 16294T 16304C 16519C
94	1	263G 315.1C 709A 8697A 9254G 11812G 16248T 16294T 16304C 16519C
95	2	263G 315.1C 709A 8697A 9254G 11812G 16294T 16304C 16519C
96	1	263G 315.1C 709A 8697A 9843G 11812G 16294T 16296T 16304C 16362C 16519C
97	1	263G 315.1C 709A 8697A 9843G 11812G 16294T 16296T 16304C 16519C
98	1	263G 315.1C 709A 8697A 9966A 11812G 16294T 16296T 16304C
99	1	263G 315.1C 709A 8697A 10750G 11812G 13722G 16294T 16296T 16304C
100	1	263G 315.1C 709A 8697A 10750G 11812G 13722G 16294T 16296T 16304C 16519C

continued on next page

Node No.	No. of Samples	Genotype
101	1	263G 315.1C 709A 8697A 10750G 11812G 16172C 16294T 16296T 16304C 16519C
102	1	263G 315.1C 709A 8697A 11242G 11812G 11914A 16294T 16296T 16519C
103	2	263G 315.1C 709A 8697A 11242G 11812G 12171G 15848G 16192T 16294T 16519C
104	2	263G 315.1C 709A 8697A 11242G 11812G 12171G 16192T 16294T 16519C
105	1	263G 315.1C 709A 8697A 11242G 11812G 16294T 16296T 16304C 16519C
106	1	263G 315.1C 709A 8697A 11242G 11812G 16294T 16296T 16519C
107	1	263G 315.1C 709A 8697A 11812G 11914A 16294T 16304C 16519C
108	1	263G 315.1C 709A 8697A 11812G 12441C 16147T 16224C 16294T 16296T 16297C 16304C 16519C
109	1	263G 315.1C 709A 8697A 11812G 13692T 16294T 16304C 16519C
110	4	263G 315.1C 709A 8697A 11812G 14016A 16294T 16296T 16304C 16519C
111	1	263G 315.1C 709A 8697A 11812G 14861A 16294T 16296T 16304C 16519C
112	2	263G 315.1C 709A 8697A 11812G 15670C 16294T 16296T 16304C 16519C
113	1	263G 315.1C 709A 8697A 11812G 15884A 16294T 16296T 16304C 16519C
114	1	263G 315.1C 709A 8697A 11812G 16189C 16294T 16296T 16304C 16519C
115	1	263G 315.1C 709A 8697A 11812G 16189Y 16294T 16296T 16304C 16362C 16519C
116	1	263G 315.1C 709A 8697A 11812G 16274A 16294T 16304C 16519C
117	2	263G 315.1C 709A 8697A 11812G 16292Y 16294T 16296T 16304C 16519C
118	4	263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16362C 16519C
119	26	263G 315.1C 709A 8697A 11812G 16294T 16296T 16304C 16519C
120	1	263G 315.1C 709A 8697A 11812G 16294T 16304C 16519C
121	1	263G 315.1C 709A 8697A 11812G 16294T 16519C
122	1	263G 315.1C 8697A 11242G 11812G 16192T 16294T 16519C
123	1	315.1C 523.1C 523.2A 709A 8697A 11812G 14180C 16294T 16296T 16304C 16519C
124	1	315.1C 709A 8697A 11812G 15172A 16189Y 16294T 16296T 16304C 16519C
125	1	315.1C 709A 8697A 11812G 15172A 16294T 16296T 16304C 16519C

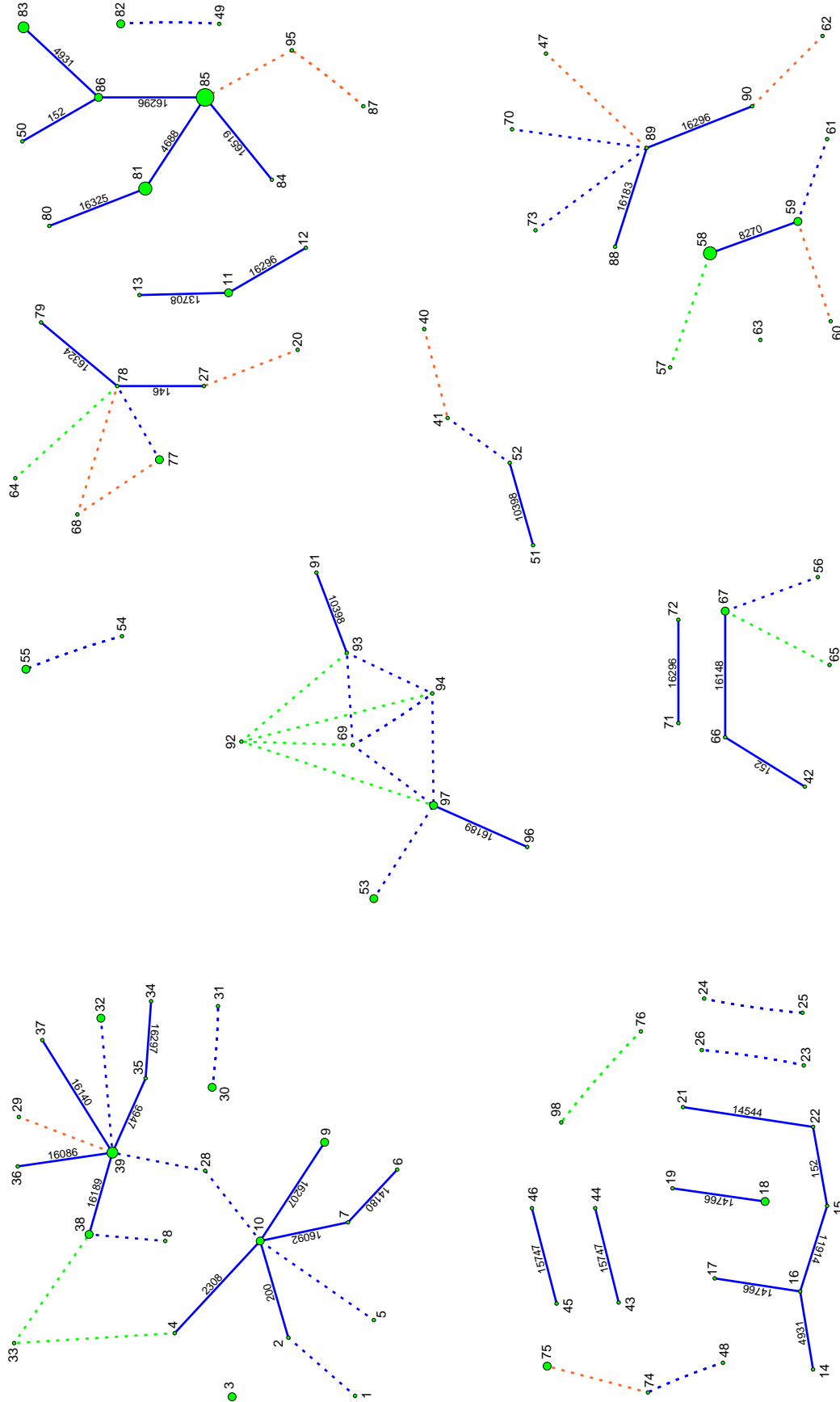


Figure 3: T2(-b) Graph

Table 8: Simplified T2(-b) Genotypes

Node No.	No. of Samples	Genotype
1	1	41T 150T 200G 315.1C 522- 523- 14766T 16153A 16519C
2	1	41T 150T 200G 315.1C 14766T 16153A 16519C
3	2	41T 150T 247A 315.1C 2308G 4924A 8864C 10682T 14766T 15499T 16153A 16183C 16189C 16519C
4	1	41T 150T 315.1C 2308G 14766T 16153A 16519C
5	1	41T 150T 315.1C 3107N 10398G 14766T 16153A 16519C
6	1	41T 150T 315.1C 14180C 14766T 16092C 16153A 16519C
7	1	41T 150T 315.1C 14766T 16092C 16153A 16519C
8	1	41T 150T 315.1C 14766T 16153A 16189C 16260T 16296T 16519C
9	2	41T 150T 315.1C 14766T 16153A 16207T 16519C
10	2	41T 150T 315.1C 14766T 16153A 16519C
11	2	143A 315.1C 523.1C 523.2A 2850C 7022C 8715C 8994A 13708A 13965C 14687G 14766T 16296T 16519C
12	1	143A 315.1C 523.1C 523.2A 2850C 7022C 8715C 8994A 13708A 13965C 14687G 14766T 16519C
13	1	143A 315.1C 523.1C 523.2A 2850C 7022C 8715C 8994A 13965C 14687G 14766T 16296T 16519C
14	1	146C 152C 279C 315.1C 4931T 5187T 6261A 7873T 10822T 14766T 16292T 16519C
15	1	146C 152C 279C 315.1C 5187T 6261A 7873T 10822T 11914A 14766T 16292T 16519C
16	1	146C 152C 279C 315.1C 5187T 6261A 7873T 10822T 14766T 16292T 16519C
17	1	146C 152C 279C 315.1C 5187T 6261A 7873T 10822T 16292T 16519C
18	2	146C 152C 279C 5187T 6261A 7679C 7873T 10822T 14766T 15784C 16292T 16519C
19	1	146C 152C 279C 5187T 6261A 7679C 7873T 10822T 15784C 16292T 16519C
20	1	146C 152C 315.1C 2141C 9117C 12741T 13965C 13966G 14544A 14687G 14766T 16296T 16324C 16519C
21	1	146C 279C 315.1C 5187T 6261A 7873T 10822T 11914A 14544A 14766T 16292T 16519C
22	1	146C 279C 315.1C 5187T 6261A 7873T 10822T 11914A 14766T 16292T 16519C
23	1	146C 315.1C 522- 523- 3834A 6261A 10822T 14766T 16292T 16296T 16438A 16519C
24	1	146C 315.1C 522- 523- 4823C 6261A 10822T 14766T 16292T 16296T 16519C
25	1	146C 315.1C 522- 523- 6261A 10822T 14766T 16292T 16296T
26	1	146C 315.1C 522- 523- 6261A 10822T 14766T 16292T 16438A 16519C
27	1	146C 315.1C 2141C 9117C 13965C 13966G 14687G 14766T 16296T 16324C 16519C
28	1	150T 152C 315.1C 14766T 16153A 16519C
29	1	150T 199C 315.1C 3107N 3882A 14766T 16153A 16296T 16519C
30	2	150T 315.1C 522- 523- 9139A 14766T 16153A 16296T 16519C
31	1	150T 315.1C 522- 523- 14766T 16136C 16153A 16296T 16519C
32	2	150T 315.1C 523.1C 523.2A 14766T 16153A 16296T 16519C
33	1	150T 315.1C 736T 2308G 14180C 14766T 16153A 16189C 16519C
34	1	150T 315.1C 9947A 14766T 16153A 16296T 16297C 16519C
35	1	150T 315.1C 9947A 14766T 16153A 16296T 16519C
36	1	150T 315.1C 14766T 16086C 16153A 16296T 16519C
37	1	150T 315.1C 14766T 16140C 16153A 16296T 16519C
38	2	150T 315.1C 14766T 16153A 16189C 16296T 16519C
39	3	150T 315.1C 14766T 16153A 16296T 16519C
40	1	152C 194T 200G 315.1C 5747G 13260C 13708A 14180C 14766T 16519C
41	1	152C 194T 315.1C 5747G 13260C 13708A 14766T 16296T 16519C
42	1	152C 200G 315.1C 1977C 3834A 14766T 14798C 14839G 16148T 16296T 16519C
43	1	152C 315.1C 499A 573.1C 573.2C 573.3C 6261A 6998T 8455T 8838A 10822T 11914A 13973T 14766T 15747C 16292T 16296T 16519C
44	1	152C 315.1C 499A 573.1C 573.2C 573.3C 6261A 6998T 8455T 8838A 10822T 11914A 13973T 14766T 16292T 16296T 16519C
45	1	152C 315.1C 499A 6261A 6998T 8455T 8838A 10822T 11914A 13973T 14766T 15747C 16292T 16296T 16519C
46	1	152C 315.1C 499A 6261A 6998T 8455T 8838A 10822T 11914A 13973T 14766T 16292T 16296T 16519C
47	1	152C 315.1C 523.1C 523.2A 8270T 8281- 14766T 16189C 16296T 16519C
48	1	152C 315.1C 573.1C 573.2C 6261A 8455T 10822T 13973T 14766T 16136C 16292T 16519C
49	1	152C 315.1C 2850C 4808T 5498G 7022C 13965C 14687G 14766T 16296T 16519C

continued on next page

Node No.	No. of Samples	Genotype
50	1	152C 315.1C 2850C 7022C 13965C 14687G 14766T 16519C
51	1	152C 315.1C 5747G 10398G 13260C 13708A 14766T 16086C 16296T 16519C
52	1	152C 315.1C 5747G 13260C 13708A 14766T 16086C 16296T 16519C
53	2	152C 315.1C 10750G 14766T 16519C
54	1	195C 198T 215G 315.1C 13020C 13965C 14766T 16519C
55	2	195C 198T 315.1C 13020C 13965C 14766T 16295T 16519C
56	1	195C 315.1C 1977C 3834A 14766T 14798C 14839G 16296T 16519C
57	1	195C 315.1C 5277C 5426C 6489A 8270T 8281- 14766T 15028A 15043A 16182- 16183- 16189C 16193.1C 16193.2C 16296T 16298C 16519C
58	4	195C 315.1C 5277C 5426C 6489A 8270T 8281- 14766T 15028A 15043A 16182C 16183C 16189C 16296T 16298C 16519C
59	2	195C 315.1C 5277C 5426C 6489A 8281- 14766T 15028A 15043A 16182C 16183C 16189C 16296T 16298C 16519C
60	1	195C 315.1C 5277C 5426C 6489A 14766T 15043A 16183C 16189C 16296T 16298C 16519C
61	1	195C 315.1C 5277C 6489A 8281- 14766T 15028A 15043A 16182C 16183C 16189C 16296T 16297C 16298C 16519C
62	1	195C 315.1C 8270T 8281- 14766T 16140C 16189C 16311C 16519C
63	1	195C 523.1C 523.2A 5277C 5426C 8281- 14766T 15028A 15043A 16182- 16183- 16189C 16193.1C 16193.2C 16296T 16298C 16519C
64	1	199C 315.1C 522- 523- 2141C 9117C 13965C 13966G 14687G 14766T 16296T 16324C 16362C 16519C
65	1	200G 315.1C 522- 523- 1977C 3834A 14766T 14798C 14839G 16362C 16519C
66	1	200G 315.1C 1977C 3834A 14766T 14798C 14839G 16148T 16296T 16519C
67	2	200G 315.1C 1977C 3834A 14766T 14798C 14839G 16296T 16519C
68	1	215G 315.1C 2141C 9117C 12741T 13965C 13966G 14687G 14766T 16324C 16519C
69	1	215G 315.1C 14766T 16296T 16519C
70	1	315.1C 499A 8281- 14766T 16189C 16296T 16519C
71	1	315.1C 507C 1766C 7337A 13834G 14766T 14798C 16296T 16519C
72	1	315.1C 507C 1766C 7337A 13834G 14766T 14798C 16519C
73	1	315.1C 522- 523- 8270T 8281- 14766T 16189C 16296T 16519C
74	1	315.1C 573.1C 573.2C 6261A 8455T 10822T 13973T 14766T 16292T 16519C
75	2	315.1C 574C 6261A 8455T 10822T 13973T 14766T 16292T 16519C
76	1	315.1C 736T 4823C 6261A 10822T 14766T 16296T 16519C
77	2	315.1C 2141C 3350C 9117C 12741T 13965C 13966G 14687G 14766T 16296T 16324C 16519C
78	1	315.1C 2141C 9117C 13965C 13966G 14687G 14766T 16296T 16324C 16519C
79	1	315.1C 2141C 9117C 13965C 13966G 14687G 14766T 16296T 16519C
80	1	315.1C 2850C 4688C 7022C 13965C 14687G 14766T 16296T 16325C 16519C
81	4	315.1C 2850C 4688C 7022C 13965C 14687G 14766T 16296T 16519C
82	2	315.1C 2850C 4808T 5498G 7022C 8435G 13965C 14687G 14766T 16296T 16519C
83	3	315.1C 2850C 4931T 7022C 13965C 14687G 14766T 16519C
84	1	315.1C 2850C 7022C 13965C 14687G 14766T 16296T
85	7	315.1C 2850C 7022C 13965C 14687G 14766T 16296T 16519C
86	2	315.1C 2850C 7022C 13965C 14687G 14766T 16519C
87	1	315.1C 3882A 13965C 14687G 14766T 16296T 16324C 16519C
88	1	315.1C 8270T 8281- 14766T 16183C 16189C 16296T 16519C
89	1	315.1C 8270T 8281- 14766T 16189C 16296T 16519C
90	1	315.1C 8270T 8281- 14766T 16189C 16519C
91	1	315.1C 10398G 12397G 14766T 16296T 16519C
92	1	315.1C 11914A 14766T 16260T 16296T 16325C 16519C
93	1	315.1C 12397G 14766T 16296T 16519C
94	1	315.1C 13260C 14766T 16296T 16519C
95	1	315.1C 13965C 14687G 14766T 16296T 16311C 16519C
96	1	315.1C 14766T 16189C 16519C
97	2	315.1C 14766T 16519C
98	1	4823C 10822T 14766T 16292T 16296T 16519C