Journal: www.jogg.info
Originally Published: Volume 5, Number 2 (Fall 2009)
Reference Number: 52.011

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: DNA GENEALOGY, MUTATION RATES, AND SOME HISTORICAL EVIDENCE WRITTEN IN Y-CHROMOSOME, PART I: BASIC PRINCIPLES AND THE METHOD

Author(s): Anatole A. Klyosov

Journal of Genetic Genealogy, 5(2):186-216, 2009

DNA Genealogy, Mutation Rates, and Some Historical Evidence Written in the Y-Chromosome: I. Basic Principles and the Method

Anatole Klyosov

Abstract

The task of genetic genealogy with respect to a group of potentially related people is (1) the assignment of them to a particular "tribe," all members of which belong to a certain haplogroup or its sub-group (a lineage), initiated in a genealogical sense by a common ancestor, and (2) an estimation of a time span between the common ancestor and his current descendants. At least two stumbling blocks in this regard are as follows: (1) sorting out haplotypes from a random series in order to assign them to their proper common ancestors, and (2) an estimation of a "calibrated" time span from a common ancestor. The respective obstacles are (1) a random series of haplotypes are often descended from a number of common intermediate ancestors, which, as a result of non-critical approaches in calculations, coalesce to a some "phantom common ancestor", and (2) mutation rates, which depend on the set of markers employed, sometimes "evolutionary" and sometimes "pedigree-based," but often lacking a clear explanation when either of them can and should be employed. We have developed a convenient approach to kinetics of haplotype mutations and calculating the time span to the common ancestor (TSCA) using both established and modified theoretical methods (Part I) and illustrated it with a number of haplotype series related to various populations (Part II). The approach involves both the "logarithmic" (count of haplotypes unmutated from the ancestral state) and the "linear" or "mutation-count" approaches, as complementary to each other, along with a separation of genealogical lineages as branches on a haplotype tree, each having its base (ancestral) haplotype. Besides the "linear" approach, we employ a correction of dating using the degree of asymmetry of mutations in the given haplotype series, and a correction for reverse mutations, using either a mathematical formula or a reference table. The latter approach is compared with the ASD (average square distance) method, using both base haplotypes and a permutational ASD method (no ancestral haplotypes employed), showing that the "linear" method has a lower error margin compared to the ASD.

Introduction

The principles of DNA genealogy have been developed over the last decade and volumes can be written on each approach. The main principles (Nei, 1995; Karafet et al., 1999; Underhill et al., 2000; Semino et al., 2000; Weale, et al., 2001; Goldstein et al., 1995a; Goldstein et al., 1995b; Zhivotovsky & Feldman, 1995; Jobling & Tyler-Smith, 1995; Takezaki & Nei, 1996; Heyer et al., 1997; Skorecki et al., 1997; Thomas et al., 1998; Thomas et al., 2000; Nebel et al., 2000; Kayser et al., 2000; Hammer et al., 2000; Nebel et al., 2001) are summarized briefly below.

Address for correspondence: Anatole Klyosov, aklyosov@comcast.net

Received: January 8, 2009; accepted: June 20, 2009

First- The DNA marker values considered in this study have nothing to do with genes. In rare cases the absence of a particular marker may be associated with an abnormal health condition, but these arguable associations are irrelevant in the context of this study.

Second- Copying of the Y chromosome from father to son sometimes results in mutations and these can be of two kinds, (1) single nucleotide polymorphisms (SNP), which are certain changes in single bases, or insertions or deletions, at particular points on the Y chromosome, and (2) mutations in short tandem repeats (STR), which make them shorter or longer by certain blocks of nucleotides. A DNA Y-chromosome segment (DYS) containing an STR is called a locus, or a marker. A combination of certain markers is called a haplotype.

Third- All human males have a single common patrilineal ancestor who lived by various estimates between

50,000 and 90,000 years ago. This time is required to explain variations of haplotypes in all tested males.

Fourth- Haplotypes can be practically of any length. Typically, the shortest haplotype considered in DNA genealogy is a 6-marker haplotype (though an example of a rather obsolete 5-marker haplotype is given in Table 1 below). The six-marker haplotype used to be the most common in peer-reviewed publications on DNA genealogy several years ago, then it was gradually replaced with 9-, 10-, and 11-marker haplotypes, and lately with 17-, 19- and 20-marker haplotypes, see Table 1. Twelve-marker haplotypes are also often considered in DNA genealogy; however, they are rather seldom presented in academic publications. For example, a common 12-marker haplotype is the "Atlantic Modal Haplotype" (in Haplogroup R1b1b2 and its subclades):

13-24-14-11-11-14-12-12-12-13-13-29

In this case, the order of markers is different when compared with the 6-marker haplotypes (typically DYS 19, 388, 390, 391, 392, 393), and it corresponds to the so-called FTDNA standard order: DYS 393, 390, 19, 391, 385a, 385b, 426, 388, 439, 389-1, 392, 389-2.

In a similar manner, 17-, 19-, 25-, 37-, 43- and 67-marker haplotypes have been used in genetic genealogy. On average, when large haplotype series are employed, containing thousands and tens of thousands of alleles, one mutation occurs once in about: 2,840 years in 6-marker haplotypes, shown above, 1,140 years in 12-marker haplotypes, 740 years in 17-marker haplotypes (Y-filer), 880 years in 19-marker haplotypes, 540 years in 25-marker haplotypes, 280 years in 37-marker haplotypes, and 170 years in 67-marker haplotypes, using the mutation rates that we will derive later. This gives a general idea of a time scale in DNA genealogy. Specific examples are given below for large and small series of haplotypes.

Fifth- The above times generally apply on average only to a group of haplotypes, whereas a pair of individuals may have large differences from these values. One cannot calculate an accurate time to a common ancestor based upon just a pair of haplotypes, particularly short haplotypes. As it is shown below in this paper, one mutation between two 12-marker haplotypes (of the same haplogroup or a subclade) places their most recent common ancestor any time between 1,140 ybp and the present time (the 68% confidence interval) or between 1725 ybp and the present time (the 95% confidence interval). Even with four mutations between two 12-marker haplotypes, their common ancestor can only be placed – with 95% confidence – between 4575 ybp and the present time, even when the mutation rate is determined with 5% accuracy. On the other hand, as it will be shown below, with as many as 1527 of 25-marker haplotypes, collectively having almost 40 thousand alleles, the standard deviation (SD) of the average number of mutations per marker is as low as ±1.1% at 3500 years to the common ancestor, and the uncertainty of the time is determined only by uncertainty in the mutation rate employed for the calculation. Similarly, with 750 of the 19-marker haplotypes, collectively having 14,250 alleles, the SD for the average number of mutations per marker equals to ±2.0% at 3600 years to the common ancestor.

As one can see, mutations are ruled by statistics and can best be analyzed statistically, using a large number of haplotypes and particularly when a large number of mutations in them. The smaller the number of haplotypes in a set and the smaller the number of mutations, the less reliable the result. A rule of thumb, supported by mathematical statistics (see below) tells us that for 250 alleles (such as in ten 25-marker haplotypes, 40 of the 6-marker haplotypes, or four 67-marker haplotypes), randomly selected, a standard deviation of an average number of mutations per marker in the haplotype series is around 15% (actually, between 11 and 22%), when its common ancestor lived 1,000 – 4,000 years before present. The smaller the amount of markers, the higher the margin of error.

Sixth- An average number of STR mutations per haplotype can serve to calculate the time span lapse from the common ancestor for all haplotypes in the set, assuming they all derived from the same common ancestor and all belong to the same clade. That ancestor had a so-called base, or ancestor (founder) haplotype. However, very often haplotypes in a given set are derived not from one common ancestor from the same clade, but represent a mix from ancestors from different clades.

Since this concept is very important for the following theoretical and practical considerations in this work, it should be emphasized that by a "common ancestor for a series of haplotypes" we mean haplotypes directly discernable from the most recent common ancestor. Such series of haplotypes are called sometimes "a cluster," or "a branch," or "a lineage." Each of them should have a founding haplotype motif, and the founding haplotype is called the base haplotype. Each "cluster," or "branch," or a "uniform" series of haplotypes typically belong to the same haplogroup, marked by the respective SNP (Single Nucleotide Polymorphism) tag, and/or to its downstream SNP's, or clades.

Granted, any given set of haplotypes has its common ancestor, down to the "Chromosomal Adam." However, when one tries to calculate a time span to a [most recent] "common ancestor" for an assorted series of haplotypes, which belong to different clades within one designated haplogroup, or to different haplogroups, he comes up with a "phantom common ancestor." This "phantom common ancestor" can have practically any time span separating it from the present time, and that

"phantom time span" would depend on the particular composition of the given haplotype set.

Since some descendants retain the base haplotype, which is passed down along the lineage from father to son, and mutations in haplotypes occur on average once in centuries or even millennia, then even after 5000 years, some descendants will still have the unchanged base haplotype—for example 23% will retain the base haplotype after 5000 years in the case of 6-marker haplotypes. In 12-marker haplotypes 23% of the descendants of the founder will still have the base haplotype after 1,800 years. These times are calculated from mutation rates that are developed later in this article.

Seventh- The chronological unit employed in DNA genealogy is commonly a generation. The definition of a generation in this study is an event that occurs four times per century. A "common" generation cannot be defined precisely in years and floats in its duration in real life and it depends on time in the past, on culture of the given population, and on many other factors. Generally, a "common" generation in typical male lineages occurs about three times per century in recent times, but may be up to four times (or more) per century in the prehistoric era. Furthermore, generation times in specific lineages may vary.

However, there is a reasonable escape from such a conundrum. A convenient factor in DNA genealogy is the number of mutations in a given set of haplotypes per marker. In turn, this ratio equals to a product of the mutation rate constant and a number of generations passed since the common ancestor times. Therefore, this product can be determined experimentally. Clearly, the mutation rate constant is fixed per generation (as it should be), but directly depends on a generation time.

This gives us two important opportunities: first, by "arbitrarily" setting the generation time, we respectively adjust the mutation rate constant; second, by using a known timespan to a common ancestor, we can "calibrate" the mutation rate constant.

Of course, we can restrict ourselves with using generations only, however, historical events are normally described in years, not in generations passed since then. Indeed, the basic quantity in DNA genealogy is generation, and the basic quantity in history is year, century, and millennium. In order to make the two disciplines compatible, we have to firmly set – for the sake of that compatibility – a number of years per generation. In this study 25 years per generation was employed

Since this issue is a very important for presentations of results in DNA genealogy and their interpretation, I will reiterate it in different terms, and give specific examples. Again, many argue that a generation often is longer than 25 years, and point at 33-35 years. However, it is irrelevant in the presented context. What actually matters in the calculations is the product (n·k), that is a number of generations by the mutation rate. If the mutation rate is, say, 0.0020 mutations per 25 years (a generation), then it is 0.0028 per 35 years (a generation), or 0.0080 per 100 years (a "generation"). The final results in years will be the same. A different amount of years per generation would just require a recalculation of the mutation rate constant for calibrated data. The "years in a generation" is a non-issue in this context, if the mutation rates are calibrated as will be shown later in this article.

Eighth- Particular haplotypes are often common in certain territories. In ancient times, people commonly migrated by tribes. A tribe was a group of people typically related to each other. Their males shared the same or similar haplotypes. Sometimes a tribe population was reduced to a few, or even to just one individual, passing though a so-called population bottleneck. If the tribe survived, the remaining individual or group of individuals having certain mutations in their haplotypes passed their mutations to the offspring. Many members left the tribe voluntarily or by force as prisoners, escapees, through journeys, or military expeditions. Survivors continued and perhaps initiated a new tribe in a new territory or group. As a result, a world DNA genealogy map is rather spotty, with each spot demonstrating its own prevailing haplotype, sometimes a mutated haplotype, which deviated from the initial, base, ancestral haplotype. The most frequently occurring haplotype in a territory is called a modal haplotype. It often, but not necessarily, represents the founder's ancestral haplotype.

Ninth- The Y-chromosome lineages of human males can be assigned to a family of Y chromosomes based on their SNP's, which in turn lead to their haplogroups and sub-haplogroups—so-called clades. SNP mutations are practically permanent. Once they appear, they remain, and they are passed on to all of the descendants of the carrier. Theoretically, some other mutations can happen at the same spot, in the same nucleotide, changing the first one. However, such an event is very unlikely. There are more than three million known chromosomal SNP's in the human genome (The International HapMap Consortium, 2007), and DNA genealogists have employed a few hundreds of them.

Examples include Haplogroups A and B (African, the oldest ones), Haplogroups C (Asian, as well as a significant part of Native Americans, descendants of Asians), Haplogroups J (Middle Eastern) with J1 (mainly Semitic, including both Jews and Arabs), and J2 (predominantly Mediterranean, including also many Turks, Armenians, Jews). Others include Haplogroup N (represented in many Siberian peoples and Chinese, as well as in many Northern Europeans) and Haplogroup R1b and its subgroups (observed primarily, but not

exclusively, in Europe, Western Asia, and Northern Africa). Haplogroup R1a1 dominates in Eastern Europe and Western Asia, with a minute percentage along the Atlantic coast. R1a1 represents close to 50% (and higher) of the population in Russia, Ukraine, Poland, and the rest of Eastern Europe, and 16% of the population in India. Haplogroup R1a1 also occurs in some areas in Central Asia particularly in Kirgyzstan and Tadzhikistan.

In other words, each male has a SNP from a certain set, which assigns his patrilineal lineage to a certain ancient family.

Tenth- It is unnecessary to have hundreds or thousands of different haplotypes in order to determine an ancestral (base) haplotype for a large population and calculate a time span from its common ancestor to the present time. Alleles in haplotypes do not have random values. Rather, they are typically restricted in rather narrow ranges. Then, after thousands of years descendants of common ancestors for whole populations of the same haplogroup have typically migrated far and wide. In Europe, for example, one can hardly find an enclave in which people have stayed put in isolation for thousands of years. Last but not least, wherever bearers of haplotypes are hiding, their mutations are "ticking" with the same frequency as the mutations of anyone else.

For example, an ancestral (base) haplotype of the Basques of Haplogroup R1b1b2, deduced from only 17 of their 25-marker haplotypes (see below) follows (in the FTDNA order):

13-24-14-11-11-14-12-12-12-13-13-29-17-9-10-11-11-25-14-18-29-15-15-17-17

This base haplotype is very close to a deduced haplotype (Klyosov, 2008a) of a common ancestor of 184 individuals, who belong to Haplogroup R1b1b2, subclade U152:

13-24-14-11-11-14-12-12-12-13-13-29-17-9-10-11-11-25-**15-19**-29-15-15-17-17

Two alleles (in bold) differ between the two base haplotypes and have average values of 14.53 and 18.35 in the Basques, while in subclade U152 the average values are 14.86 and 18.91, respectively. Adding the two differences we get 0.89 mutations total between the 25-marker haplotypes. This amount of difference between two founding haplotypes would suggest only approximately ten generations between them, using a method to be presented shortly. However, a margin of error will be significantly higher when, e.g., 425 alleles are considered (17 of 25-marker haplotypes) compared to 14,250 alleles (750 of 19-marker haplotypes), as in the following example.

The 750 Iberian R1b1 19-marker haplotypes as published in (Adams et al, 2008) apparently all descended from the following base (ancestral) haplotype, shown here in the same format as the above:

13-24-14-11-11-14-X-12-12-13-13-29

The base haplotype on the first 12 markers is exactly the same, plus the only marker, DYS437, from the second FTDNA panel, determined in the 19-marker haplotype series, is also "15" in the base Iberian haplotype in both 25- and 19-marker formats. As it will be shown below, an average number of mutations in these two series of Basque haplotypes, seventeen 25- and seven hundred fifty 19-markers ones, is also practically the same: 100 mutations in the first series and 2796 mutations in the second series give, respectively, 0.257 and 0.262 after normalizing for their average mutation rates (see **Table 1**). However, the margin of error is much lower in the second case. It will be considered in detail below.

This kind of a comparison would, however, be misleading when comparing haplotypes of two individuals on or near the modal values of a haplogroup (Nordtvedt, 2008). As it was stated above (section Fifth), mutations are ruled by statistics and can best be analyzed statistically, using a number of haplotypes, not just two, as it was demonstrated above using 17 Basque, 184 subclade U152, and 750 Basque haplotypes from three different series.

To further illustrate the example, consider 12,090 of 25-marker R1b haplotypes (including subclades) from the YSearch database. When combined, they have the following modal (base) haplotype:

13-24-14-11-11-14-12-12-12-13-13-29-17-9-10-11-11-25-15-19-29-15-15-17-17

This is exactly the same base haplotype as shown above for R1b1b2-U152, and practically the same for that for the Basques of Haplogroup R1b1b2. Furthermore, as it is shown below, common ancestors of the 17 Basques, 750 Basques, 184 bearers of U152 subclade, and 12,090 bearers of R1b haplogroup lived in about the same time period, within less than a thousand years.

The power of DNA genealogy is not in large numbers, though they are always welcomed and greatly reduce the standard deviation of the TSCA, but in randomness of haplotype selections. Again, that power can be significantly reduced when small haplotype series (with less than 250 alleles collectively, see above) are employed.

Eleventh- Unlike languages, religion, cultural traditions, anthropological features, which are often assimilated over centuries and millennia by other languages, cultures, or peoples, haplotypes and haplogroups cannot be assimilated. They can be physically exterminated,

though, and haplotype trees very often point at extinct lineages. This non-assimilation makes haplogroups and haplotypes practically priceless for archaeologists, linguists, and historians, as well as geneticists. They not only stubbornly transcend other assimilations across millennia, but also provide means for calculations of when, and sometimes where, their common ancestors lived.

Methods

We will discuss several methods for calculating time spans to common ancestors (TSCA) for a given series of haplotypes. The underlying principles of the methods are well established, and all are based on a degree of microsatellite variability. We will consider two main classes of methods, (1) counting of mutations or "genetic distances" and conducting a statistical evaluation, and (2) using the fraction of a series of haplotypes that still match the haplotype of the common ancestor (CA)--those that have no mutations.

In principle, any of the methods may be used, and–theoretically–they should yield approximately the same result. In reality, they do not, and results vary greatly, often by factors of two or more, when used by different researchers, even when practically the same populations are under study.

The main reasons of such a discrepancy are typically as follows: (a) different mutation rates employed by researchers, (b) lack of calibration of mutation rates using known genealogies or known historical events, or when a time depth for known genealogies was insufficient to get all principal loci involved, (c) mixed series of haplotypes, which are often derived from different clades, and in different proportions between those series, which directly affect a number of mutations in the series, (d) lack of corrections for reverse mutations (ASD-based calculations [see below] do not need such a correction), (e) lack of corrections for asymmetry of mutations in the given series of haplotypes – in some cases. All these issues are addressed in this study.

In the present study we will make use of both of the main classes of methods for calculating the TMRCA for a series of haplotypes. We will use both approaches and then demand reasonable agreement between them. We will discuss first the "no-mutations" approach, and then follow with a discussion of the mutation-counting approaches.

Brief Introduction to Mutation Rates

At this point we will simply refer to Table 1 for the mutation rates for different haplotypes. The average rate per marker is provided, along with the rate for the whole haplotype. Later in this article, we will explain the origin of these rates. It should be noted that the reader is free to use a different rate--the method may still be applied.

Base Haplotypes Remaining After t Generations

We first consider the approach to calculating the "age" of a common ancestor that is based not on counting mutations, but on base haplotypes remaining unmutated in a series of haplotypes. This method does not suffer from "asymmetry" of mutations, or from multiple mutations of the same marker, and does not consider which mutations to include and which to neglect in a total count of mutations, because it only considers unmutated haplotypes. Its principal limitation is that it requires an appreciable number of base haplotypes in the series, preferably more than ten.

Naturally, the longer the haplotypes (more markers) in the series, the less of the base haplotypes the series will retain. However, for large haplotype series this not a concern. For example, the 19-marker haplotypes in the 750-haplotype Iberian R1b1 series contains 16 base haplotypes, identical to that shown above. A series of 857 English 12-marker haplotypes contains 79 base haplotypes (Adamov and Klyosov, 2009b). While a series of 325 of Scandinavian I1 25-marker haplotypes contained only two base haplotypes, the same series of 12-marker haplotypes contained as many as 26 base haplotypes.

Probabilities of mutations, or mutation rates in haplotypes can be considered from quite different angles, or starting from different paradigms. One of them assumes that a discrete probability distribution of mutations in a locus (or an average number of mutations in a multi-loci haplotype), that is a probability of a number of independent mutations occurring with a known average rate and in a given period of time, is described by the Poisson distribution

$$P(m) = \frac{(kt)^m}{m!} e^{-kt}$$

where:

P(m) = probability of appearance of "m" mutations in a marker (or haplotype),

m = a number of mutations in a marker (or haplotype),

k = average number of mutations per marker (or haplotype) per generation (or year),

t = time in generations (or years).

As an example, for k = 0.00088 mutations per 12-marker haplotype per year, a 100-haplotype series will contain on average 80 base (unchanged, identical)

Table 1
Average Mutation Rates for Haplotypes Consisting of Various Combinations of Markers

Haplotypes in the FTDNA Order	Average mutation rate per generation		Notes
	Per Haplotype	Per Marker	
393-390-X-391-X-X-X-X-X-389$_1$-X-389$_2$	0.0108	0.00216	5-marker haplotype, e.g., in Cordaux et al. (2004)
393-X-19-X-X-X-388-X-389$_1$-X-389$_2$	0.0068	0.00135	5-marker haplotype, e.g., in Bittles et al, (2007)
393-390-19-391-X-X-X-388-X-X-392-X	0.0088	0.00147	6-marker haplotypes in the "old scientific" format: 19-388-390-391-392-393
393-390-19-391-X-X-X-X-X-389$_1$-X-389$_2$	0.0123	0.00205	6-marker haplotype, e.g., in Thanseem et al. (2006)
393-390-19-391-X-X-X-X-X-389$_1$-392-389$_2$	0.013	0.00186	7-marker haplotypes, with missing markers 385a, 385b, 426, 388, 439
393-390-19-391-X-X-X-388-X-389$_1$-392-389$_2$	0.013	0.00163	8-marker haplotypes, with missing markers 385a, 385b, 426, 439
393-390-19-391-385a-385b-X-X-X-389$_1$-X-389$_2$	0.0168	0.0021	8-marker haplotype, with missing markers 426, 439, 388, 392, e.g. in Contu et al. (2008)
393-390-19-391-385a-385b-X-Y-Z-389$_1$-392-389$_2$	0.017	0.00189	9-marker haplotypes, with missing markers 426, 388, 439
393-390-19-391-X-Y-Z-388-439-389$_1$-392-389$_2$	0.018	0.002	9-marker haplotypes, with missing markers 385a, 385b, 426
393-390-19-391-385a-385b-X-388-Y-389$_1$-392-389$_2$	0.018	0.0018	10-marker haplotypes, with missing markers 426, 439
393-390-19-391-385a-385b-X-Y-439-389$_1$-392-389$_2$	0.022	0.0022	10-marker haplotypes, with missing markers 426, 388
393-390-19-391-X-Y-426-388-439-389$_1$-392-389$_2$	0.018	0.0018	10-marker haplotypes, with missing markers 385a, 385b
393-390-19-391-X-X-X-388-439-389$_1$-392-389$_2$-(…)- 461	0.018	0.0018	10-marker haplotype, e.g. in Sengupta et al. (2006)
393-X-19-391-X-X-X-X-439-X-X-X-(…)- 413a-413b-460-461-GATAA10-YCAIIa-YCAIIb	0.020	0.00182	11-marker haplotype, e.g. in Cruciani et al. (2007)
393-390-19-391-X-X-X-388-439-389$_1$-392-389$_2$-(…)- 437-438	0.019	0.00176	11-marker haplotype, e.g. in Zalloua et al. (2008)
393-390-19-391-385a-385b-426-388-439-389$_1$-392-389$_2$	0.022	0.00183	12-marker haplotype in the FTDNA order
393-390-19-391-385a-385b-X-Y-439-389$_1$-392-389$_2$-437-438	0.024	0.00197	12-marker haplotype, e.g. in Mertens (2007)
393-390-19-391-X-X-X-388-439-389$_1$-392-389$_2$-(…)-YCAIIa-YCAIIb-460	0.024	0.00203	12-marker haplotype, e.g. in Fornarino et al. (2009)
393-390-19-391-X-X-X-388-439-389$_1$-392-389$_2$-(…)-YCAIIa-YCAIIb-461	0.021	0.00178	12- marker haplotype, e.g. in Chiaroni et al. (2009)
393-390-19-391-385a-385b-X-X-439-389$_1$-392-389$_2$-458-(…)-437-448-GATAH4-456-438-635	0.034	0.002	17-marker haplotype (Yfiler, FBI/National Standards) in Mulero et al. (2006)
393-390-19-391-X-X-X-388-439-389$_1$-392-389$_2$-(...)-434-435-436-437-438-460-451-462	0.024	0.00141	17- marker haplotype, e.g. in King et al. (2007)
393-390-19-391-426-388-439-3891-392-3892-458-455-454- 447-437-448-438	0.032	0.00188	17- marker haplotype, e.g. in Hammer et al. (2009)
393-390-19-391-385a-385b-X-388-439-389$_1$-392-389$_2$-(…)-434-435-436-437-438-460-461-462	0.0285	0.0015	19-marker haplotype, e.g. in Adams et al. (2008)
393-390-19-391-385a-385b-388-439-389$_1$-392-389$_2$-458-(…)-437-448-GATAH4-YCAIIa-YCAIIb-456-438-635	0.050	0.0025	20-marker haplotype, e.g. in Tofanelli et al. (2009)
393-390-19-391-385a-385b-426-388-439-389$_1$-392-389$_2$-458- 459a-459b-455-454- 447-437-448-449-438	0.047	0.00214	22- marker haplotype, e.g. in Hammer et al. (2009)
393-390-19-391-385a-385b-426-388-439-389$_1$-392-389$_2$-458-459a-459b-455-454-447-437-448-449-464a-464b-464c-464d	0.046	0.00184	25-marker haplotype in the FTDNA order
Standard 37-marker haplotype	0.090	0.00243	37-marker haplotype in the FTDNA order
Standard 67-marker haplotype	0.145	0.00216	67-marker haplotype in the FTDNA order

haplotypes (m=0) after 250 years, since kt = 0.22 and $e^{-0.22}$ = 0.8. One could also use an equivalent mutation rate of .022 mutations per 12-marker haplotype per generation, and an equivalent number of 10 generations (both quantities assuming 25 years per generation), and kt remains 0.22.

Another approach employs the binomial theorem, according to which a fraction of haplotypes with a certain number of mutations in a series equals

$$P(m) = \frac{t!\,p^{(t-m)}}{(t-m)!m!}q^m$$

where:
m = a number of mutations,
q = probability of a mutation in the haplotype each year,
t = time in years (or generations),
p = 1-q

Similarly with the above example, for q = 0.00088 mutations per 12-marker haplotype per year, a 100-haplotype series will contain 80 base (unchanged) haplotypes (m=0) after 250 years, since 0.99912^{250} = 0.8.

The third approach, which I prefer to employ in this work due to its simplicity and directness, is the "logarithmic" approach. It states that a transition of the base haplotypes into mutated ones is described by the first-order kinetics:

$$B = Ae^{kt,} \qquad (1)$$

that is,

$$\ln(B/A) = kt \qquad (2)$$

where:

B = a total number of haplotypes in a set,
A = a number of unchanged (identical, not mutated) base haplotypes in the set
k = an average mutation rate

For the example given above it shows that for a series of 100 12-marker haplotypes (the average mutation rate of 0.00088 mutations per haplotype per year),

$$\ln(100/80)/0.00088 = 250 \text{ years.}$$

It is exactly the same number as those obtained by the Poisson distribution and binomial theorem above.

Needless to say, all the above three approaches to the unmutated haplotypes have the same mathematical basis, and, as it was said above, are simply presented from three different points of view.

Note that the uncertainty in Equation (2) depends upon the uncertainty in B (the number of unmutated haplotypes), but after propagation of the uncertainty through the logarithmic function, the uncertainty in the time is much smaller.

Brief Introduction to Mutation-Counting Methods

Mutation-counting methods are all based on accumulation of mutations in haplotypes over time, starting from a base haplotype in a founder. They can be subdivided into three approaches:

(a) The "linear" method, in which a total number of mutations in a set of haplotypes is counted, an average number of mutations per marker is calculated, a correction for back mutations is introduced, either numerically (see below) or using a handy table, such as Table A (see Appendix A), and a time span to a common ancestor is calculated, either using the same Table A, or applying the respective mutation rate, taken from Table 1. In other words, it is described by the following equation

$$n/N/\mu = t \qquad (3)$$

where n is a number of mutations in all N haplotypes in the given series of haplotypes, μ is an average mutation rate per haplotype (Table 1), and t is the time back to the common ancestor (in generations or years), which can be corrected for back mutations (Table A) .

(b) The "quadratic" method (ASD, see below), after a base haplotype is identified in the haplotype set.

(c) The "permutation" method, in which a base haplotype is not directly considered (see below).

All the three approaches are based on the same principles, that is, taking a cumulative mutation distance (or a square of it) between each allele in each haplotype and the "base" (presumably ancestral) allele, or between each allele in the haplotype set (permutation method), and requires separation of a haplotype set into "branches," or separate genealogical lineages, each with its own nearest common ancestor.

Following the introduction of these methods, along with dissection of haplotype trees into branches, or lineages, and their separate analysis, Table A below is provided that will make it possible to avoid most of the math that is involved to make corrections for reverse mutations.

Applying the "Logarithmic" and Linear Methods Together for Calculating a Time to the Common Ancestor

Either of the two methods – the logarithmic and the mutation-counting methods for calculating a time to the common ancestor may be used, but with one condition:

they both should give approximately the same result. This is important, since both of them are based on quite different methodology. If the two methods yield significantly different results, for example, different by a factor of 1.5, 2 or more, then the haplotype series probably represents a mixed population, that is haplotypes of different clades, clusters, lineages. Or it might indicate some other details of the genealogy or population dynamics, which is inconsistent with one lineage, and will result in a "phantom common ancestor." In this case it will be necessary to divide the group appropriately into two or more subgroups and to treat them separately. Constructing a haplotype tree is proven to be very effective in identifying separate lineages, as will be introduced below and shown in more detail in the second part of this two-part series of articles (Klyosov, 2009).

Brief Introduction to Mutation Rates

At this point we will simply refer to Table 1 for the mutation rates for different haplotypes. In this table, for each haplotype, the average rate per marker is provided, along with the rate for the whole haplotype. Later in this article, we will explain the origin of these rates. It should be noted that the reader is free to use a different rate--the method may still be applied.

Examples of Using the Logarithmic and Linear Methods to Check the Integrity of a Set of Haplotypes

An important part of the method is to assess a haplotype set with both the logarithmic and linear methods to make sure they give consistent results, which would mean that the collection of haplotypes was an appropriate one for analysis as a cluster (descended from a common ancestor). We present here two brief example to illustrate this important principle.

Let us consider two sets with 10 haplotypes in each:

Set 1	Set 2
14-16-24-10-11-12	14-16-24-10-11-12
14-16-24-10-11-12	14-16-24-10-11-12
14-16-24-10-11-12	14-16-24-10-11-12
14-16-24-10-11-12	14-16-24-10-11-12
14-16-24-10-11-12	14-16-24-10-11-12
14-16-24-10-11-12	14-16-24-10-11-12
14-17-24-10-11-12	14-16-25-09-11-13
15-16-24-10-11-12	14-16-25-10-12-13
14-15-24-10-11-12	14-17-23-10-10-13
15-17-24-10-11-12	16-16-24-10-11-12

The first six haplotypes in each set are base (ancestral) haplotypes. They are identical to each other and are presumed to represent the unmutated base haplotypes. The other four are mutated base haplotypes or possibly represent admixtures from descendant haplotypes of a different common ancestor. A number of mutations in the two sets with respect to the base haplotypes are 5 and 12, respectively. If to operate only with observed mutations (the linear approach), the apparent number of generations to a common ancestor in the two sets is, according to Equation (3), equal to 5/10/0.0088 = 57 generations and 12/10/0.0088 = 136 generations, respectively (without a correction for back mutations), which represents a large difference. However, in both cases a ratio of base haplotypes in the logarithmic approach gives us a number of generations equal to ln(10/6)/0.0088 = 58 generations. Hence, only the first set of haplotypes gave closely matching numbers of generations (57 and 58) and represents a "clean" set, having formally one common ancestor. The second set is "distorted," or "mixed," as it certainly includes descendant haplotypes from apparently more than one common ancestor. Hence, this set cannot be used for calculations of a number of generations to a common ancestor.

As a second example Figure 1 shows a haplotype tree of the Clan Donald 25-marker haplotypes. There are 84 haplotypes in the series, and 21 of them are identical to each other and presumably the common ancestor. We find from Table 1 that the mutation rate for 25-marker haplotypes is .046, hence, ln(84/21)/0.046 = 30 generations to a common ancestor. All those 84 haplotypes contain 109 mutations, so the linear method gives 109/84/0.046 = 28 generations to a common ancestor. Since these results match closely, this haplotype series is appropriate for analysis as a single cluster.

In this initial check on a set of haplotypes, it does not matter which mutation rate is employed, as long as the same rate is used for both parts of the check.

Hence, concerning the Donald haplotypes, the above calculations give three pieces of evidence: (1) the consistency of the calculations, (2) a proof of a single common ancestor in the series of 84 haplotypes, (3) approximately 29 ± 4 generations to a common ancestor, using the mutation rate of 0.046 mutations per 25-marker haplotype per generation. This example will be considered in more detail below. However, it should be noticed here that the uncertainty in the value 29±4 generations is based on a standard deviation of 5.8% for the average number of mutations per marker, and 5% for the standard deviation for the mutation rate, which when combined by the square root of the sum of the squares of the two standard deviations, we get 7.6% for the overall standard deviation, or a result of 29 ± 4. The theory behind it is considered below.

Lately, four more mutated haplotypes were added to the Donald Clan series. 21 base haplotypes stay the same, and all 88 haplotypes contain 123 mutations. This gives ln(88/21)/0.046 = 31 generations, and 123/88/0.046 = 30 generations to a common ancestor. It still holds the preceding value of 29 ± 2 generations to a common

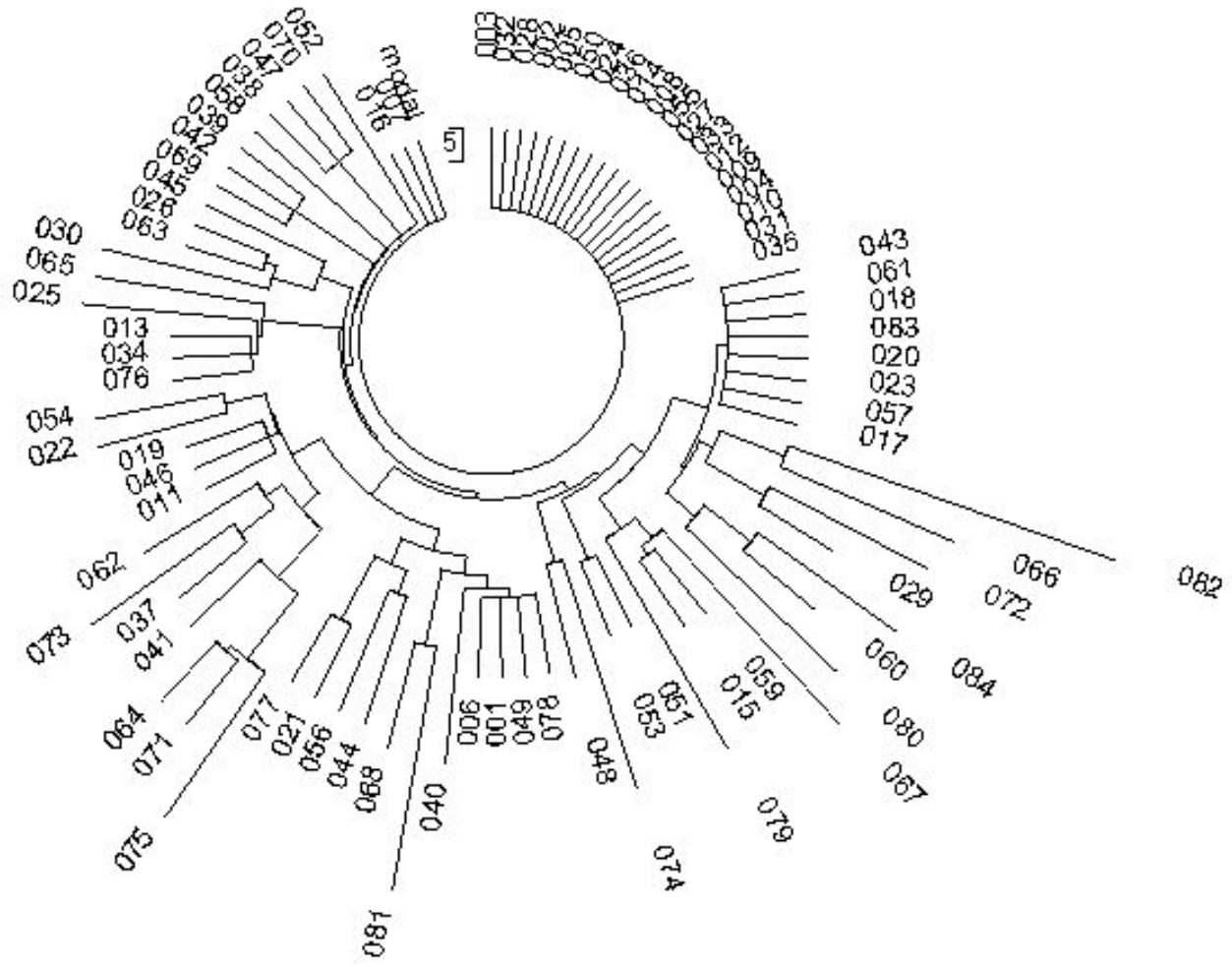

Figure 1. The 84-haplotype 25-marker tree for R1a1 Donald haplotypes. The tree was composed according to data of the DNA Project Clan Donald (DNA-Project.Clan-Donald, 2008). The tree shows 21 identical "base" haplotypes sitting at the top of the tree.

ancestor without considering the "experimental" standard deviation, and 30 ± 4 generations with that consideration. In the last case, with the inclusion of four additional haplotypes, the two standard deviations described above became 9.0% and 13.5%, respectively. It will be explained below.

Mutation Rates and Tabular Calculation Aides

The application of the proposed method is to clusters of haplotypes that all descend from a common ancestor. However, in any such cluster, if the father-son mutation rates are used, the TMRCA will tend to be underestimated. What is needed are "effective rates" that will provide the correct TMRCA, and these are generally smaller than the father-son rates. If we calibrate the 12-marker and 25-marker panels to the corresponding Clan Donald haplotypes, we get effective rates very close to 0.0216 for the 12-marker panel and .045 for the 25-marker panel. When the 17-marker, 37-marker, and 67-marker panels are calibrated to the Clan Donald dataset, values of .034, .090, and .145 are obtained. These rates are reported directly in Table 1. For the different kinds of haplotypes that have been used in research studies, representing many different combinations of the common markers, the relative rates of Chandler (2006) were used along with a convenient overall panel rate to arrive, first, at estimates of the needed individual marker rates,

and finally, at the overall haplotype mutation rate as a sum of individual rates (along with the average marker rates for the haplotype). A 25-year generation time was assumed in these calculations.

Consider for a moment that we had assumed a different number of years per generation when calibrating the mutation rates to the Clan Donald data. For example, if there were actually 33.3 years per generation, then the mutation rates obtained would have been higher by a factor of 4/3 and then using this rate with any TMRCA calculation would give a TSCA result lower by that factor. Therefore, a lower number of generations would be obtained in Table A, which when multiplied by 33.3 years per generation, would provide exactly the same time in years that was obtained by the 25 years per generation assumption. So, there is no loss of generality in using the 25-year generation assumption as we have here.

The cluster examples we present in Part 1 (this article) and in Part 2 will use the effective mutation rates we have determined here and in previous studies, which have been quite successful, but any user is free to substitute a different mutation rate and still carry out our method. In fact, our approach always involves a scaling to a different mutation rate (from that assumed in constructing Table A), and this different mutation rate can be one of those we suggest in Table 1 or one of the user's own choosing.

Table A provides a computation aid for determining the TMRCA. For a given number of mutations per marker per haplotype in the left-most column, the table provides the number of generations to the common ancestor. A column of values that includes correction for back mutations is also provided. The table is normalized to a mutation rate of 0.0020 mutations per generation per marker so that only one such table will be needed. If a mutation rate different from 0.0020 is desired, due to the kind of haplotype or the user's own preferences, the number of generations (or time) obtained from Table A should simply be multiplied by the ratio 0.0020 / (the appropriate average rate per marker).

Detailed Procedure for a Calculation of a Timespan to a Common Ancestor of a Series of Haplotypes

Our approach consists of rather simple steps that will simplify a calculation of a timespan to a common ancestor for a given series of haplotypes. Here are suggested steps to follow:

Step 1: Make sure that the series of haplotypes under consideration is derived each from a single common ancestor, not from two or more common ancestors. The cluster began with the ancestral haplotype, which served as a base for subsequent branching via mutations. These branchings have led to a series of haplotypes under consideration. In order to make sure that the series is derived from a single "common ancestor," we can employ a few criteria.

The first criterion is to analyze a haplotype tree. In case of one common ancestor, the tree will ascend to one "root" at the trunk of the tree. If two or more separate roots are present, each with separate branches, the construction would point to separate "common ancestors." All of them, if within one haplogroup, have their "common ancestor." This may occur within several haplogroups as well. However, a given haplotype series should be treated separately, with one common ancestor at a time. Otherwise some "phantom common ancestor" will be numerically created, typically as a superposition of several of them.

A "base" haplotype can be equivalent to the ancestral one, or it can be its approximation, particularly when it is not present in multiple copies in the series of haplotypes under consideration. Hence, two different terms, "ancestral haplotype" and "base haplotype" can be utilized. The simplest and the most reliable way to identify an ancestral (base) haplotype is to find the most frequently repeated copy in a given series of haplotypes.

The integrity of the series of haplotypes should be verified by using the "linear" and "logarithmic" models as discussed above. According to the linear model:

$$n/N/\mu = t \qquad \text{(repeating Equation 3)}$$

where n is a number of mutations in all N haplotypes in the given series of haplotype, μ is an average mutation rate per haplotype per generation (Table 1), and t is a number of generations to a common ancestor. Unlike the "linear" model, the "logarithmic" one, as it was described above, considers the number of base haplotypes in the given series, and does not count mutations. It employs Equation (2):

$$\ln(N/m)/\mu = t_{ln} \qquad \text{(repeating Equation 2)}$$

where m is the number of base (identical) haplotypes in the given series of N haplotypes, t_{ln} is a number of generations to a common ancestor. If $t = t_{ln}$ (within a reasonable range, for example, within 10%), then the series of haplotypes is derived from the same common ancestor. If t and t_{ln} are significantly different (for example, by 50% or greater), the haplotype series is certainly heterogeneous. Table A can be applied only after any necessary separation of haplotypes into appropriate groups, each deriving from its own common ancestor. For that separation, the respective haplotype tree can be used (Klyosov, 2008a, 2008b, 2008c; Adamov & Klyosov, 2008a).

In other words, for a "homogeneous" series of haplotypes that are derived from a single common ancestor, a

number of mutations should be compatible with a number of base haplotypes remaining in the same series.

Step 2: Count a number of mutations in the "homogeneous" series of haplotypes. This number should be counted with respect to the base (ancestral) haplotype identified in the preceding step. All mutations should be counted, considering them as independent ones. This is justified below.

Step 3: Calculate an average number of mutations per marker from all haplotypes in the "homogeneous" series, as described in the preceding step. For example, if in 20 haplotypes, each with seven markers, there are 65 mutations, then the average number of mutations equals to 65/20/7 = 0.464±0.057 mutations per marker per haplotype, accumulated during a timespan from a common ancestor. This value is obtained with the assumption of a full symmetry of the mutations (see below).

Step 4: Carry out the following simple calculations: Find the average per marker mutation rate for your haplotypes in **Table 1**, or supply an average marker rate of your own choosing. Calculate the following ratio:

$$\text{Ratio} = \frac{0.0020}{(\text{appropriate average marker rate})}$$

We can use this ratio to scale values obtained in Table A from one average marker mutation rate to another, allowing just one version of Table A to be compiled, rather than 28 different tables, one for each rate.

Step 5: Using the value obtained in Step 3 above for average number of mutations per marker per haplotype, find the corresponding entry in the first column of Table A. Then note the corresponding number of generations in column 3 of Table A. This column contains a correction for back mutations (the second column does not).

Step 6: Multiply the Ratio obtained in Step 4 times the number of generations found in Step 5. This step will scale your result to the proper mutation rate for your haplotypes.

For example, consider the series of 20 seven-marker haplotypes discussed above. In this series we calculate an average number of accumulated mutations of 0.464±0.057 mutations per marker (Step 3 above). We then determine the Ratio from Step 4. We first find the average marker mutation rate for the seven-marker haplotype in **Table 1** and it is 0.00186. Then the Ratio = .002/.00186 = 1.075. Next we look down column 1 of **Table A** until we find the row for 0.464, then look across to column 3 and find the TMRCA (column 3 is corrected for back mutations) in generations as 300 (or the time as 7500 years). Finally, the TMRCA is scaled by the Ratio from Step 4, resulting in TMRCA = 1.075 x300 = 322.5 generations or 8060 years.

We can apply these factors to the uncertainties as well. We have a basic quantity, .464 ±0.057, so we can add and subtract the 0.057 value from .464 and follow those two values through the same procedure, we get 376 and 274, or 322 ± 53 generations or 8060 ± 1,325 years to a common ancestor. The timespan in years is calculated by assigning 25 years to a generation, as explained above.

This process can be illustrated using the above Basques and Iberian series of haplotypes. In all 44 of the 12-marker Basques haplotypes there are 122 mutations from the base haplotype

13-24-14-11-11-14-12-12-12-13-13-29

Any other haplotype mistakenly assumed as the ancestral one ("base") would give more mutations in the haplotype series. Hence, a "base" haplotype is defined as that which produces a minimal number of mutations in the haplotype series. A search for the base haplotype can be called a minimization of mutations in the given haplotype series. The average number of mutations per marker in this Basques haplotype series is 0.231. In fact, it is 0.231±0.021, however, the calculation of error is explained in the following section.

There are at least three ways for conversion of this number to the number of generations or years:

(a) Using the mutation rate (0.00183 in this particular case, **Table 1**), followed by correction for back mutation employing **Table A**. This gives 0.231/0.00183 = 126 generations, that is (see **Table A**, third column) 145 generations, corrected for back mutations, or 3625 ybp. In fact, it is 3625 ± 490 ybp (see below).

(b) Using the ratio of 0.00183 to the standard reference value of 0.002 for the average mutation rate, based on which the first column of **Table A** was created. Practically, in this particular case 0.231x2/1.83 = 0.252 mutations per marker, which is a normalized value for **Table A** (the first column). 0.252 in **Table A** gives 126 generations without a correction for back mutations (the second column), or 145 generations with the correction (the third column), or 3625 ybp (the last column). It is the same result as that in (a).

(c) Using a numerical method, explained in the sixth step below, employing Equation (4), which includes asymmetry of mutations in the haplotype series, or its simplified form, Equation (6), which in this particular case gives us 0.261, which is the actual (not

observed, as 0.231) average number of mutations per marker in this particular haplotype series (44 of 12-marker haplotypes). "Actual" in this context means corrected for back mutations. Therefore, 0.261/0.00183 = 143 generations to a common ancestor, or 3575 ybp. In fact, it is 3575±480 ybp, or well within the margin of error with the above values.

Essentially the same values are obtained with the other two haplotype series for the Basques and the Iberian R1b1, as described below.

For the sake of consistency of the explanation, we will continue with considerations of other aspects of the calculations.

If only the linear model is employed, without a consideration for reverse mutations, then 65 mutations in 20 of 7-marker haplotypes would lead one to an erroneous conclusion. The erred "result" would show only 65/20/0.013 = 250±40 generations (6,250±995 years) to a common ancestor, versus the more correct 331±53 generations (8,275±1,320 years) to a common ancestor. Formally, these two results are overlapping within their margins of error, however, the lower one is still incorrect.

The same Table A considers contributions of reverse mutations into results of the logarithmic model in which reverse mutations are not included. For example, if the logarithmic model results in 250 generations to a common ancestor, Table A shows that it corresponds to 331 generations, corrected for reverse mutations.

In this study, haplotype trees were constructed using PHYLIP, the Phylogeny Inference Package program (Felsenstein, 2005). A "comb" around the wheel, a "trunk," in haplotype trees identifies base haplotypes, identical to each other and carrying no mutations compared to their ancestral haplotypes (see Fig. 2 below). The farther the haplotypes lies from the wheel, the more mutations they carry compared to the base haplotype and the older the respective branch.

For more sophisticated researchers, three more steps in haplogroup analysis are suggested below.

Sixth: Calculate a degree of asymmetry of the haplotype series under consideration. A degree of asymmetry, when significant, affects a calculated time span to a common ancestor at the same number of mutations in the haplotype series. Generally, the more asymmetrical is the haplotype series (that is, mutations are predominantly one-sided, either "up" or "down" from the base haplotype), the more overestimated is the TSCA. Specific examples are considered in the subsequent section.

The degree of asymmetry is calculated as a number of +1 or -1 mutations (whichever is higher) from the base haplotype divided by a combined number of +1 and -1 mutations. For a symmetrical haplotype series the degree of asymmetry is equal to 0.5, as in the East European Slav R1a1 12- and 25-marker marker haplotypes (see Part II). For a moderately asymmetrical series the degree of asymmetry is equal to about 0.65, as in the R1b1b2 Basque 12-marker haplotype series, though for the 19-marker extended haplotype series of 750 haplotypes it is equal to 0.56 (see below). For a significantly asymmetrical series it is equal to 0.86, as in the N1c1 Yakut haplotype series (Adamov and Klyosov, 2008b), or to 0.87, as with the English I1 extended haplotype series (see below), and in extreme cases approaches to 1.0.

The degree of asymmetry (a) is useful for a correction of an average number of mutations per marker (λ), which in turn is used for calculations of a TSCA for the given population (given series of haplotypes), using the following three formulae (Adamov and Klyosov, 2009a):

$$\lambda = \frac{\lambda_{obs}}{2}(1 + \exp(a_1 \lambda_{obs}))$$

$$a_1 = 1 - a^{0.8} \qquad (4)$$

$$a = (2\varepsilon - 1)^2$$

where:

λ_{obs} = observed average number of mutations per marker,

λ = average number of mutations per marker corrected for reverse mutations,

ε = degree of asymmetry (= 0.5 for complete symmetry, = 1.0 for complete asymmetry)

a = normalized degree of asymmetry (a = 0 for complete symmetry, a = 1.0 for complete asymmetry)

For a completely asymetrical series of haplotypes, ε = 1, a = 1, a_1 = 0.

For a completely symmetrical series of haplotypes, ε = 0.5, a = 0, a_1 = 1.

$$\lambda = \frac{\lambda_{obs}}{2}(1 + \exp(\lambda_{obs})) \qquad (6)$$

Equations (4) - (6) can be used for calculations of average number of mutations per marker corrected for back mutations and for an asymmetrical haplotype series

using Equations (4) and (5), and for symmetrical series using Equation (6).

For a case of fully asymmetrical haplotype series (with respect to mutations) a "linear" and a "quadratic" (ASD) calculation procedures give the same time span to a common ancestor.

A degree of asymmetry of haplotype series also affects a standard deviation for a calculated TSCA, as discussed in the following paragraph.

Seventh: Calculate a standard deviation for an average number of mutations per marker. The following formula may be employed (Adamov and Klyosov, 2008):

$$\frac{\sigma(\lambda)}{\lambda} \approx \frac{1}{\sqrt{N\lambda}}\left(1+\frac{a_1\lambda}{2}\right) \qquad (7)$$

Where,

$\lambda = \lambda_{obs}$

N = a number of markers in the haplotype series under consideration,

a = was defined above as a normalized degree of asymmetry.

Specific examples of calculated standard deviations are given in the subsequent section.

Equation (7) does not include a standard deviation for the average mutation rate in haplotypes, but this will be added separately. As an example of the use of Equation (7), consider ten 25-marker haplotypes (N = 250) and λ_{obs} = 0.276 (4,000 years to a common ancestor), and for a symmetrical series of haplotypes, a standard deviation equals 14% (13.7%, to be exact). For 100 haplotypes of the same kind a standard deviation will be 4% (4.3%, to be exact). For the set of 750 Iberian 19-marker haplotypes the SD equals 2.0%. Again, these standard deviations do not include standard deviations for mutations rates. This is a subject for the subsequent paragraph.

Eighth: Calculate the overall standard deviation for the obtained time span to the common ancestor (TSCA), including the uncertainty in the mutation rate. Generally, margins of errors for average mutation rates are more guesswork than science, at least in reality. They probably vary between 5% and 15-20%. For the most of mutation rates employed in this work I estimated the standard deviation as 5%, so that the 95% confidence interval ("two sigma") was ±10%. This estimate looks very reasonable and is verified by many practical examples in this (Part I) and the subsequent Part II of the article. This margin of error means that at 3-5 thousand years to a common ancestor it cannot be lower than plus-minus 300-500 years, even for haplotype series of thousands of haplotypes (see below and Part II).

The standard deviation (SD) for the time span to the common ancestor is based on the standard deviations for each of the two components, that is the SD for the average number of mutations per marker (see above, Item 7) and the SD for the average mutation rate for the given series of haplotypes. For example, for R1b1 Iberian haplotypes (see above) the SD for the TSCA would be equal to the root-mean-square of the two errors. If we take the SD for the mutation rates in this case (the 750 Iberian 19-marker haplotypes) to be ±5% and the SD for the average mutations per marker to be 2%, then the overall SD equal to 5.4%. In other words, for such large series of haplotypes, the standard deviation for a time span to a common ancestor is more affected by the SD for the employed average mutation rate than by the SD in the average number of mutations per marker. For a smaller series, both uncertainties would contribute to the overall uncertainty.

The 95% confidence interval ("two sigma") for TSCA for the same R1b1 Iberian haplotypes would be equal to 10.8%, or 3,625±390 years before present (see below).

Practical Examples

Let us consider other examples, the first one is the Basque R1b1b2 haplotype series in 12- and 25-marker format, the second one is the Iberian R1b1 19-marker haplotypes, and the third one is the British Isles I1 12- and 25-marker haplotype series. The Iberian and the Isles haplotype series include hundreds of extended haplotypes.

Basques, Haplogroup R1b1b2. 12- and 25-marker haplotypes

The Basque DNA Project (Cervantes, 2008) lists 76 haplotypes which belong to Haplogroups E1b1a, E3b1a, E3b1b2, G2, I, I1b, I2a, J1, J2, R1a and R1b1, and their downstream haplogroups and subclades. Of this grouping of haplotypes, 44 haplotypes (or 58% of total), belong to subclades R1b1 (one haplotype), R1b1b2a (three haplotypes), and R1b1b2 (40 haplotypes, or 91%). The last one is often considered to be of Western European origin, though it is more conjecture than proven fact. The origin of R1b1b2 will be the subject of a forthcoming study (to be published).

Only 17 of those R1b1 Basque haplotypes were available in the 25-marker format (numbering is according to the 44 12-marker haplotypes; the haplotypes are presented in the FTDNA order) in **Table 2**. The respective haplotype tree is given in **Figure 2**.

One can see from Figure 2 that the tree stems from a single mutation coming from a presumably common ancestral haplotype for all 17 individuals in the haplotype set. The base (ancestral) haplotype can be identified as follows:

13-24-14-11-11-14-12-12-12-13-13-29-17-9-10-11-11-25-14-18-29-15-15-17-17

In fact, this is the haplotype 021 on the tree (Figure 2) and in the list of haplotypes above. However, one base haplotype is not enough to use the logarithmic approach. A rule of thumb tells that there should be at least 3-4 base haplotypes in a series in order to consider the logarithmic method.

For the 12-marker haplotypes the Basque ancestral haplotype is also identical to the so-called Atlantic Modal Haplotype (Klyosov, 2008a, 2008b):

13-24-14-11-11-14-12-12-12-13-13-29

The "linear" method. All 17 of 25-marker haplotypes have 100 mutations from the above base haplotype (DYS389-1 was subtracted from DYS389-2 and the result used in place of DYS389-2), which gives 0.235 mutations per marker on average (the statistical treatment of the data is given below). Using Table 1 and Table A, one can calculate a time span to a common ancestor of the Basques presented in the haplotype set, which is equal to 147 generations, or 3,675 years.

Using the same approach for all 44 of 12-marker Basque R1b1b2 haplotypes, one finds that all of them contain 122 mutations from the base haplotype

13-24-14-11-11-14-12-12-12-13-13-29

which corresponds to 0.231 mutations per marker on average, resulting in 145 generations or 3,625 years to a common ancestor. It is practically equal to the findings above of 3,675 years obtained from the 25-marker set of haplotypes. Note that it is only a statistical accident that the two results are so close, as we will show below that the uncertainties are much larger than this small difference.

However, these calculations are applicable only for symmetrical mutations over the whole haplotype series, which does not exactly apply in the considered case since the mutations were asymmetrical: 65 of single mutations were "up" and only 36 "down," all three double mutations were up, and all five triple mutations were down. The degree of asymmetry for 12-marker haplotypes equals to 0.64, hence, α = 0.0784, α_1 = 0.869, and an average number of mutations per marker, corrected for

Table 2
Basque 25-marker Haplotypes

ID	Basque 25-Marker Haplotypes (FTDNA order)
009	13 23 14 10 11 11 12 12 12 14 13 30 18 9 10 11 11 25 15 19 29 15 17 17 18
009	13 23 14 11 11 14 12 12 13 14 13 30 18 9 10 11 11 24 15 19 29 15 16 17 19
013	13 24 14 10 11 14 12 12 13 13 13 29 18 9 10 11 11 25 15 19 28 15 15 17 18
014	13 24 14 10 11 15 12 12 12 13 13 29 17 9 10 11 11 25 14 18 29 15 15 17 17
015	13 24 14 10 11 15 12 12 12 13 13 29 17 9 10 11 11 25 14 18 29 15 16 17 17
015	13 24 14 11 11 14 12 12 11 14 14 31 17 9 9 11 11 25 14 18 29 15 15 15 15
021	13 24 14 11 11 14 12 12 12 13 13 29 17 9 10 11 11 25 14 18 29 15 15 17 17
024	13 24 14 11 11 14 12 12 12 14 13 30 17 9 10 11 11 25 14 18 29 15 15 16 17
027	13 24 14 11 11 15 12 12 12 13 13 29 16 9 10 11 11 25 15 19 28 15 15 17 17
029	13 24 14 11 11 15 12 12 12 13 13 30 16 9 10 11 11 25 15 19 28 15 15 17 17
030	13 24 14 11 11 15 12 12 13 13 13 29 17 9 10 11 11 25 14 18 31 15 15 17 17
032	13 24 14 11 12 14 12 12 12 14 13 30 17 9 10 11 11 25 14 18 30 15 15 17 17
034	13 24 15 11 11 14 12 12 12 13 13 29 19 9 10 11 11 24 15 19 30 15 16 17 17
035	13 25 14 10 11 15 12 12 12 13 13 29 18 9 10 11 11 25 15 19 29 15 15 17 19
035	13 25 14 11 11 11 12 12 12 12 13 28 18 9 10 11 11 25 16 17 28 15 15 17 17
036	13 25 14 11 11 14 12 12 11 14 13 30 17 9 10 11 11 25 14 18 30 15 15 16 17
038	13 25 14 11 11 14 12 12 12 14 13 30 18 9 9 11 11 25 14 18 29 15 16 16 17

Figure 2. The 25-marker haplotype tree for Basque R1b1 (mainly R1b1b2) haplotypes. The 17-haplotype tree was composed according to data of the Basque DNA Project (Basque DNA Project, 2008).

reverse mutations, calculated by using Equation (4), is equal to 0.257.

Thus, for the 12-marker haplotypes, the "linear" method gave us a result of $\lambda_{obs} = 0.231$; the same method, corrected for back mutations and assuming a symmetrical pattern of mutations and using Table A gave us $\lambda = 0.265$; and corrected for back mutations and the asymmetry of the mutations we got $\lambda = 0.257$; respectively. For the two corrected values, this results in 145 and 140 generations, or 3625 and 3500 years to a common ancestor. Here, the degree of asymmetry of 0.64, that is about two thirds of the mutations were "one-sided," resulted in a slightly decreased TSCA compared to the value assuming symmetry. The difference in this particular case was 5 generations, or 3.6% of the total. The TSCA is progressively overestimated without this correction for asymmetry, as the "age" of the common ancestor increases.

The standard deviation must also be corrected for the asymmetry in the 12-marker haplotypes, calculated by using Equation (7):

$$\frac{\sigma(\lambda)}{\lambda} \approx \frac{1}{\sqrt{528 \cdot 0.257}}(1 + 0.869\frac{0.257}{2}) = 0.095$$

that is, 9.5%. Assuming again an SD of 5% in the mutation rate, we get the overall SD by root-mean-square of 10.7%. The 95% confidence interval would then be ± 21% or ± 735 years.

Similar results are obtained for the 25-marker haplotypes.

We can also illustrate the difference one will obtain if the father-son mutation rates are blindly used instead of effective rates. The father-son rates of Kerchner (2008), who also found that the 12-marker and 25-marker rates were essentially the same, are .0025 and .0028 per marker per generation. Using these rates would result in 0.231/0.0025 = 92 generations and 0.235/0.0028 = 84 generations, for the 12- and 25-marker Basque series, respectively, or 102 and 92 generations after correction for back mutations. For 25 years per generation this would give 2550 - 2300 years to the Basque common ancestor, which is an unbelievably recent time period.

ASD methods: We can also compare the obtained timespan to a common ancestor using our methods, with that provided by the average square distance (ASD) method in its two principal variants– (a) employing a base (ancestral) haplotype, and (b) without a base haplotype, but employing permutations of all alleles (Adamov & Klyosov, 2008b). Both of the ASD methods already correct for back mutations, but they are more tedious otherwise, when used manually. Besides, the variant (a) is sensitive to asymmetry of mutations in the series (Adamov & Klyosov, 2008b, 2009a), and particularly to even a small amount of extraneous haplotypes. Both the ASD methods typically give a higher error margin compared with the "linear" method, commonly as a result of multiple (multi-step) mutations and accidental admixtures of haplotypes from a different common ancestor (Adamov & Klyosov, 2009a).

The ASD method, using the base haplotype: Since all 44 of 12-marker Basque haplotypes contain 101 single-step mutations, three double mutations, and five triple mutations, the sum of the squares of the differences from the base haplotype ("actual" number of mutations estimated to have occurred) in the 44-haplotype set is 101 + $3x2^2$ + $5x3^2$ = 158. The observed or apparent number of mutations was 122 (see above), or 77% of the actual, as the calculations showed. Hence, an average number of actual mutations per marker is 0.299±0.027 (compared to the observed 0.231±0.021, see above), which (using our effective rates) corresponds to 163 generations or 4,075 years to a common ancestor. It has its own uncertainty, but it is still within the 3500 ± 735 ybp obtained with our methods. It would be expected that the ASD approach would be slightly higher due to a higher sensitivity of the "quadratic" method to admixtures as well as to double and triple mutations (which are counted as two or three successive single mutations) in the haplotype series. Note that if we had used the father-son rates of Kerchner, we would have gotten 119 generations, and would again have an underestimate of the number of generations.

The permutational ASD method, no base haplotype: We will illustrate this method using the 25-marker haplotypes. There are 17 alleles for each marker in the haplotypes, and the method considers permutations between each one of them, with squares of all the differences summed up for all the 25 markers. For the 17 Basques haplotypes this value equals to 3728. It should be divided by 17^2 (all haplotypes squared), then by 25 (the number of markers in a haplotype) and by 2 (since all permutations are doubled by virtue of the procedure). This gives 0.258 as an average number of "actual" mutations per marker, which corresponds to 141 generations to a common ancestor--quite close to the value obtained with our method.

Iberian R1b1 19-Marker Haplotypes

In order to further verify the approach, 750 of 19-marker Iberian R1b1 haplotypes were considered. The haplotype tree, based on the published data (Adams et al, 2008) is shown in Figure 3, not to show the details, but with a purpose to show that the tree is quite uniform, reasonably symmetrical, and does not contain ancient, distinct branches. All branches are of about the same length. This all indicates that the tree, with its most or all of the 750 haplotypes, is derived from a relatively recent common ancestor, who lived no more than four or five thousand years ago. It would be impossible for the tree to be derived from a common ancestor who lived some 10-15 thousand years ago, much less 30 thousand years ago. But, let us verify it.

First, the base haplotype for all the 750 entries, obtained by a minimization of mutations, in the presented in the same order as employed by the authors (Adams et al, 2008), DYS 19-388-389.-389.-390-391-392-393-434-435-436-437-438-439-460-461-462-385a-385b, is as follows:

14-12-13-16-24-11-13-13-11-11-12-15-12-12-11-12-11-11-14

In this format the Atlantic Modal Haplotype (AMH) is as follows (Klyosov, 2008a):

14-12-13-16-24-11-13-13-X-X-Y-15-12-12-11-X-X-11-14

in which X replaces the alleles which are not part of the 37-marker FTDNA panel, and Y stands for DYS436 which is uncertain for the AMH. The same haplotype is the base one for the subclade R1b-M269 and R1b-U152 (Klyosov, 2008a). Hence, the Iberian R1b1 haplotypes are likely to have a rather recent origin.

All 750 haplotypes showed 2796 mutations with respect to the above base haplotype, with a degree of asymmetry of 0.56. Therefore, the mutations are fairly symmetrical, and a correction for the asymmetry would be a minimal one. The whole haplotype set contains 16 base haplotypes.

An average mutation rate for the 19-marker haplotypes is not available in the literature, as far as I am aware of, and cannot be calculated using the Chandler's, Kerchner's, or other similar data. However, the Donald Clan latest edition of 88 haplotypes contains 63 mutations in the above 19 markers. Taking into account the 26 generations to the Clan founder (see above), this results in the mutation rate of 0.0015 mut/marker/gen and 0.0285 mut/haplotype/gen, listed in Table 1.

The logarithmic method gives ln(750/16)/0.0285 = 135 generations, and a correction for reverse mutations results in 156 generations (Table A), that is 3900 years to a common ancestor of all the 750 Iberian 19-marker haplotypes. It corresponds well with 3500±480 ybp value, obtained above for 12- and 25-marker Basque haplotype series. The "mutation count" method gives 2796/750/19 = 0.196±0.004 mutations per marker (without a correction for back mutations, that is λ_{obs} = 0.196±0.004), or after the correction it is 0.218 ± 0.004 mutations per marker, or 0.218/0.0015 = 145±15 generations, that is 3625±370 years to a common ancestor of all 750 Iberian R1b1 haplotypes.

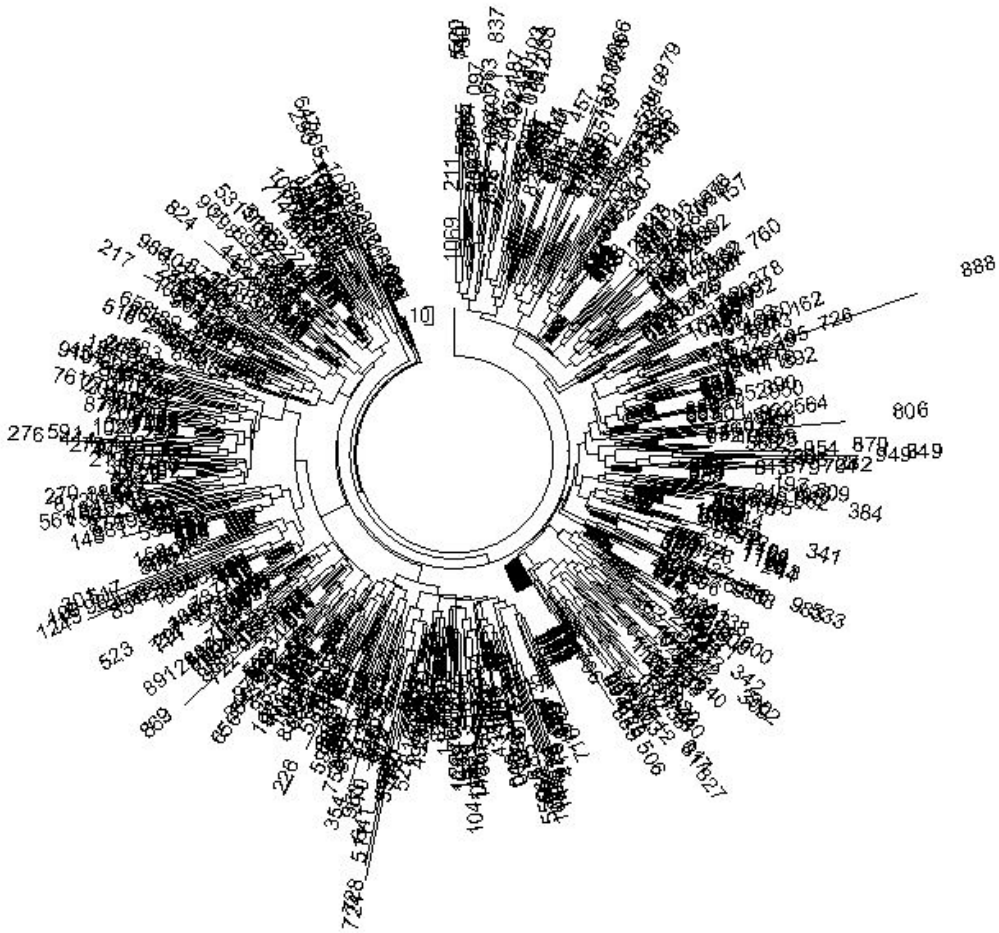

Figure 3. The 19-marker haplotype tree for 750 Iberian R1b1 haplotypes. The tree was composed according to data published (Adams et al, 2008), but here is not intended to be completely legible.

We can also calculate this number of generations by using Table A. If we begin with the value of 0.196 for the number of mutations per haplotype per marker, then adjust for the fact we need a mutation rate of 0.0015, while Table A is normalized to a mutation rate of .0020, we get an adjusted value of 0.196 x .0020/.0015 = 0.261. The table entry for this value under the column that accounts for back mutations is 149.

Considering the degree of asymmetry of 0.56, and using Equation (4) we obtain:

$$\lambda = \frac{0.196}{2}(1 + \exp(0.965 \cdot 0.196)) = 0.217$$

In other words, at the degree of asymmetry of 0.56 the average number of mutations per marker, 0.217 ± 0.004, is practically equal to 0.218 ± 0.004 for the fully symmetrical (å = 0.50) pattern of mutations in the haplotype series. It gives 145 ± 15 generations, that is 3625 ± 370 years to a common ancestor. Equation (4) gives the standard deviation

$$\frac{\sigma(\lambda)}{\lambda} \approx \frac{1}{\sqrt{14250 \bullet 0.196}}(1 + \frac{0.965 \bullet 0.196}{2}) = 2\%$$

that is 0.217±0.004 mut/marker, and for the 5% standard deviation for the mutation rate, the 95% confidence interval for the time span to a common ancestor with be equal to 10.2% (root-mean-square of the two components). This is how the above value of 3625 ± 370 ybp for the 750 Iberian 19-marker haplotypes was calculated. This value is practically equal to 3,500 ± 480 ybp for 12- and 25-marker Basque haplotypes.

English, Irish, and Scotish I1 12- and 25-marker Haplotypes

857 of English 12-marker I1 haplotypes were considered in (Adamov and Klyosov, 2009b). They all contain 79 base haplotypes and 2171 mutations. This gives ln(857/79)/0.022 = 108 generations, or 121 generations with a correction for back mutations, that is 3025 years to a common ancestor (the statistical considerations are given below). By mutations, it gives 2171/857/12 = 0.211 ± 0.005 mut/marker (without a correction), or 0.238 ± 0.005 mut/marker (corrected for back mutations), or 0.220 ± 0.005 (corrected for back mutations and the asymmetry of mutations, which equal to 0.87 in this particular case). This results in 0.220/0.00183 = 120±12 generations to a common ancestor. This is practically equal to the 121 generations, obtained by the logarithmic method. Obviously, the logarithmic method, being irrelevant to asymmetry of mutations (since only base, non-mutated haplotypes are considered), can be preferred method in cases of high asymmetry of mutations. This results in 3,000 ± 300 years to a common ancestor for all the 857 English 12-marker haplotypes.

The asymmetry of mutations, which is rather high in this particular case (0.87) adds 10 generations (250 years) to the TSCA, corrected for reverse mutations. This addition was properly corrected back in this particular case.

The same haplotypes, but in the 25-marker format, contain 4863 mutations, which gives λ_{obs} = 0.227 ± .003 and λ = .260 ± .004 (corrected for back mutations) and λ = .251 ± .004 (corrected for both back mutations and asymmetry). This gives 0.251/0.0018 = 137±14 generations, that is 3,425±350 years to a common ancestor.

If we combine the above 857 English I1 haplotypes with 366 Irish and 304 Scottish I1 haplotypes, the combined set has 1527 haplotypes, each with 25 markers. All of the haplotypes contain 8785 mutations, so λ_{obs} = 0.230 ± .002 and λ = .265 ± .003 (corrected for back mutations) and λ = .255 ± .003 (corrected for both back mutations and asymmetry). This gives 0.255/0.0018 = 139±14 generations, that is 3,475±350 years to a common ancestor. Again, in this combined haplotype set the degree of asymmetry was higher for 12-marker haplotypes: 0.85, compared to 0.65 for 25-marker haplotypes. The result for the combined set is essentially the same as for just the English set.

The standard deviations for the 1527 25-marker haplotype series were calculated using Equation (7) which gives "two sigma" in this particular case as 1.1%, that is the average number of mutations to be 0.255±0.003. For the 5% standard deviation for the mutation rate, the 95% confidence interval for time span to a common ancestor with be equal to 10.1%. This gives 3475±350 ybp for the 1527 Isles 25-marker I1 haplotypes.

Discussion

There are a number of typical questions and issues addressed when the very basis of quantitative DNA genealogy is considered. Among them are the following ones:

1) Which mutations should be counted and which should be not?

The underlying reason is that there is a potential problem of counting the same mutation multiple times. For example, in Figure 2 one can see three branches, each stemming from a supposedly one ancestral (base) haplotype. The branch at the bottom of the figure contains eight haplotypes with the distinct common DYS437 = 15, while all other branches contain 14 in that locus.

Hence, the typical argument is that the common ancestor of this branch had DYS437 = 15, and this very mutation (one step from the AMH) was counted eight times (in fact, nine times, since 036 contains there 16). Therefore, as opponents argue, one probably over-counts the mutations, and obtains an erroneously high number of generations to the common ancestor of the entire haplotype set.

Generally, this consideration may be valid (see below), but not in this particular case. First, as it was shown above, the same number of generations to the common ancestor was obtained from both 12-marker haplotypes (which do not include DYS437), using both the logarithmic and the "linear" methods, also from the 25-marker haplotypes, and from a large series of 19-marker haplotypes. However, there is another way to examine the obtained value, namely, to consider all the three branches in Figure 2. The branch at the bottom contains eight haplotypes, all contain 46 mutations from its common ancestor of the branch, which gives 144±26 generations from the common ancestor of the branch. The five-haplotype upper-right branch contains 16 mutations, which gives 75±19 generations to its common ancestor. The three-haplotype upper-left branch contains only 5 mutations from its base haplotype, which gives 39±17 generations to its common ancestor. An average number of generations for all three branches is 86±21. Certainly, these operations are very approximate ones, and they aim at the semi-quantitative verification of the concept.

Then we apply the same approach to those three base haplotypes as described above. They all have 8 mutations between them from the base haplotype for the entire haplotype series, which results in summarily 62±6 generations from the common ancestor for the whole series to the "averaged" common ancestor of the separate branches. This gives (86±21)+(62±6) = 148±27 generations from the initial common ancestor to the present time, .

This value is close to 147±21 generations obtained by the "linear" method (see above). Indeed, mutations in haplotypes can be considered as practically independent ones, and we can count them either for the entire haplotype series, provided that all haplotypes are derived from one common ancestor, or analyze branches separately, as it was shown above.

In many cases one indeed can over-count mutations, particularly when they belong to different branches and to different common ancestors. For example, English and Irish R1a1 haplotype series contain many DYS388 = 10 (in one particular haplotype series of 57 English haplotypes there are 10 of them, and in 52 Irish haplotypes there are 12 of them, that is 18% and 23%, respectively). There are practically no such DYS388=10 alleles in Polish, Czech, Slovak, Hungarian, Russian, Jewish and Indian R1a1 haplotypes, and very few among Swedish and German haplotypes. Incidentally, there was not a single case of DYS388=10 in R1b1 haplotype series considered in the subsequent paper (Part II), containing 750 Iberian, and 983 and 218 Irish haplotypes. DYS388 is an extremely slow marker, and it is likely that all R1a1 haplotypes with DYS388=10 descended from the same, just one common ancestor.

When mutations in R1a1 haplotypes are counted assuming that DYS388=10 is a common, random mutation, without a consideration that those haplotypes are derived from a different common ancestor, each DYS388 = 10 haplotype adds a double mutation, which is particularly damaging when the ASD method is employed. On a haplotype tree of English and Irish R1a1 haplotypes the DYS388 = 10 branches stand out quite distinctly (the trees are shown in the subsequent paper, Part II). If to count all those double mutations, without separation the branches, the Irish "phantom" common ancestor comes out as of 5000 ybp. However, a separate consideration of the branches results in 3575±450 ybp for the DYS388 = 10 common ancestor, and in 3850±460 ybp for the DYS388 = 12 common ancestor. However, their 25-marker base haplotypes differ by six mutations, which places their common ancestor at 5,700±600 years before present (see Part II).

Another remarkable example of a potential over-count of mutations is related to haplotypes with DYS426 = 10 in Haplogroup J1. It is known that DYS426 is an extremely slow marker. Those mutations are so infrequent that they are practically irreversible. In haplogroups of an earlier origin, including C through O, a great majority of people have DYS426 = 11. Only in "younger" Haplogroups, Q and R, a great majority of people have DYS426 = 12. For example, among all 343 haplotypes of Haplogroup J1 in YSearch, collected in 2008, only 23 had DYS426 = 10 or 12. It turned out that all of them derived from one common ancestor each, and in fact the same mutation was carried through practically all the generations in the respective lineage over many thousand years.

Figure 4 shows the 12-marker J1 haplotype tree with mutated DYS426. Haplotypes of all said 23 individuals are shown there. Of all the 23 haplotypes, eleven are located in the vicinity of the "trunk" of the tree, and eight of their bearers have Jewish surnames (haplotypes 002 through 006, 008, 010 and 011 in Figure 4). They have five base haplotypes and four mutations among those eight, which gives ln(8/5)/0.022 = 21 generations, and 4/8/0.022 = 23±12 generations, that is 550±200 years to their common ancestor, who lived, apparently, around the 15th century, and had the following haplotype:

12-24-13-10-12-19-10-15-13-12-11-29

Another eight individuals with DYS426 = 10 did not have typical Jewish surnames. They had the following base haplotype:

12-24-13-10-12-19-10-15-12-12-11-29

All those 8 haplotypes had 24 mutations, which brings their common ancestor to 3950±1450 ybp.

Haplotypes of the remaining seven individuals had DYS426 = 12, and form a distinct, obviously very ancient branch on the right-hand side in Figure 4. Most of them have rather typical European surnames, along with one Palestinian individual among them. Their deduced base haplotype:

12-24-14-10-12-14-12-13-12-13-11-29

All the seven haplotypes contained 46 mutations from this base haplotype, which indicates that their common ancestor lived 10600±1900 years ago.

The above and the "Jewish" J1 haplotypes differ by 12 mutations on 12-markers, which brings their common ancestor to 760±135 generations ago, that is 19,000±3,400 ybp. This is as close to the "bottom" of J1 haplogroup as one gets.

These examples show that in order to avoid over-count of mutations one should consider a haplotype tree, separate branches, and calculate them separately.

2) Which mutation rates to use?

Zhivotovsky "evolutionary" mutation rate of 0.00069 mutations per marker per generation was empirically derived from three different populations and three dif-

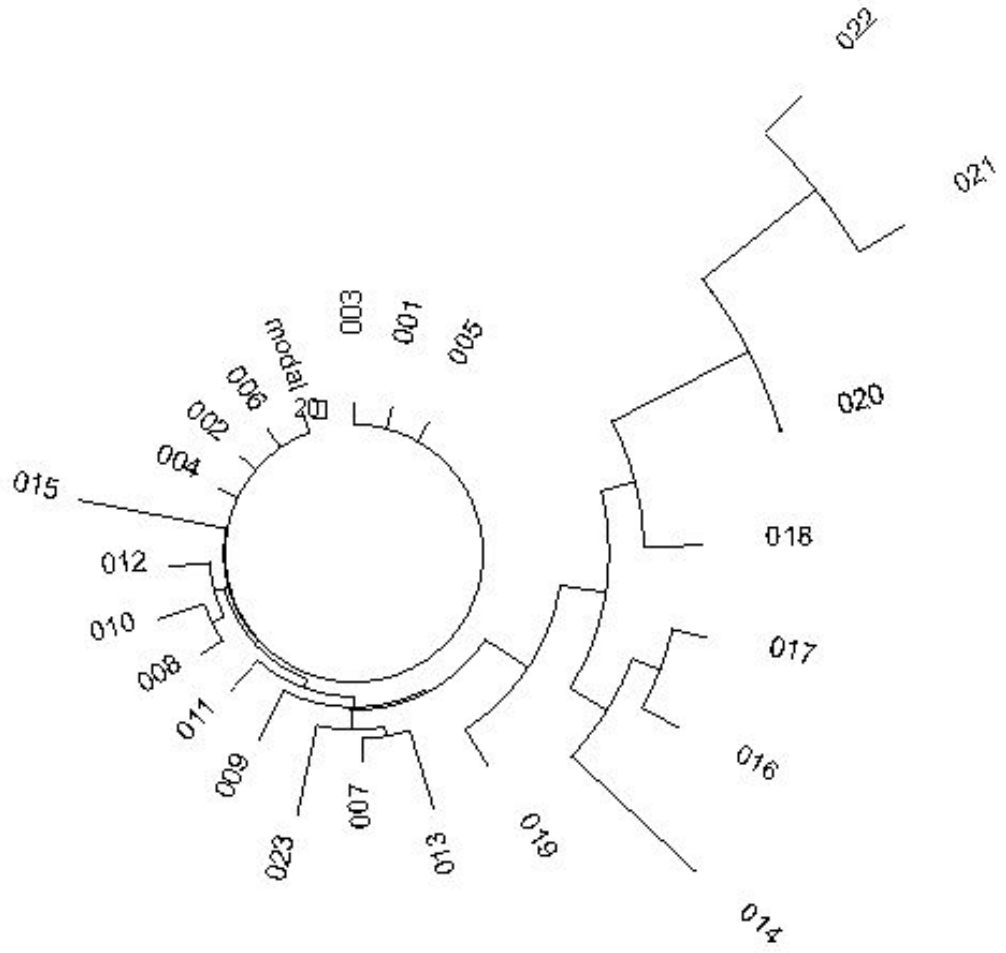

Figure 4. The 12-marker haplotype tree for J1 haplotypes with mutated DYS426. The tree was composed from haplotypes collected in YSearch data base. Bearers of haplotypes 002, 003, 004, 005, 006, 008, 010 and 011 have Jewish surnames.

ferent kinds of haplotypes using a number of questionable assumptions, and it was recommended for use not in "genealogical" or "pedigree-based" studies, but in "population" studies. Criteria for determining when a series of haplotypes represents a "population" and when it represents a "genealogical" situation were not provided. As a result, this mutation rate has been widely used in the academic literature quite indiscriminately, often (or always) resulting in time spans to common ancestors some 200-300% greater compared to those obtained with "genealogical" mutation rates.

In fact, it is easy to calculate from Table A that the 0.00069 mut/marker/gen mutation rate is applicable for a time span equal to 2560 generations, that is 64,000 years ago. For >64,000 ybp the actual mutation rate will be lower than 0.00069, for <64,000 ybp the actual mutation rate will be higher than 0.00069. This is valid, of course, if to suggest that the only factor which effectively reduces the apparent mutation rate in progressively ancient times is progressively accumulated reverse mutations. In reality there may be many factors in "population dynamics," such as genetic drift, extinctions of lineages, etc., however, they would necessarily result in appearance of branches on a haplotype tree. Each of these branches should be considered and analyzed separately. "Weights" of these branches should be necessarily taken into consideration, since some branches outweigh others, resulting in some phantom "common ancestors." All of this make the "population mutation rates" even less meaningful than those just neglecting reverse mutations. In other words, results of calculations of timespans to "common ancestors" using "population mutation rates" applied to a mass of haplotypes without sorting them out into branches can be substituted with a qualitative in kind expression: "it was a long time ago."

3) Are the same mutation rates applicable to the cases where the time depth is a few hundred years, and where it is over a thousand or more years?

As it was shown in this study, the same mutation rates are well applicable in the both cases, from as recent times as a couple of centuries to 3625±370 ybp (the Basques and R1b Iberian haplotypes), and to 16,300±3,300 years (Native Americans of Q-M3 haplogroup, see the subsequent paper) and there is no reason to believe that they cannot be applicable to a much deeper time spans. The most important requirement for that is that the mutations are independent, hence, are governed by statistics. There has not been any proof to the contrary. All apparent deviations observed and reported from time to time are easily explained by multiple counting of inherited mutations, without separations of the respective branches (see Figure 4 with the explanations above), by mixing separate genealogical lineages (branches of the tree), and similar missteps in data analysis.

An illustration can be provided from a recent publication (Tofanelli et al, 2009), in which the authors listed 282 of 20-marker haplotypes of Haplogroup J1-M267. The authors gave an overall estimate of the "median TMRCA" between 6643 and 47439 ybp. Their list of 282 haplotypes showed a base haplotype (in the format DYS 19-389-389-390-391-392-393-385a-385b-437-438-439-456-458-635-GATAH4-YCAIIa-YCAIIb)

14-23-13-17-10-11-12-13-19-17-14-10-11-20-15-18-21-11-22-22

From all the 282 haplotypes we have 2746 mutations from that base haplotype, which gives the average number of mutations per marker of 0.487±0.009, and the TSCA of 6,025±610 years bp. In order to verify this value, the haplotype tree, shown in Figure 5 was subdivided to seven major branches, and the TSCAs were calculated to each of them. Surprisingly, except the "oldest" branch in the upper right area with the TSCA of 5,300±600 years, all other branches are relatively young, with the youngest one having TSCA of 1800±230 years. Overall, the analysis of the branches showed that the common ancestor of all of them lived 5,400±800 ybp. Again, there is no reason to believe that calculations of the TSCA work only at depths of no more than a few hundred of years ago, even with rather complicated haplotype trees.

Another support to this statement is provided with a calculation of a time span to the common ancestor of a few dozen haplotypes of Haplogroup A, which came out as about 37,000 ybp (to be published). There is nothing unexpected in this result. Clearly, in order to handle such distant time spans the tree should be dissected to separate branches, and corrections for back mutations should be applied.

4) How can one take the mutational analysis seriously, since the standard deviations must be so high?

Such a general misconception is propagated mainly as a result of a non-critical analysis of the well known paper by Walsh (2001). In his excellent paper Walsh considered pairs of haplotypes, since the main goal was to provide a basis for forensic analysis. Naturally, with only two haplotypes an error margin would be huge, as follows from Equation (7) above. For example, for two of 12-marker haplotypes having a common ancestor 600 years ago the formula in its simplified form (for a fully asymmetrical mutations and fully symmetrical mutations, respectively)

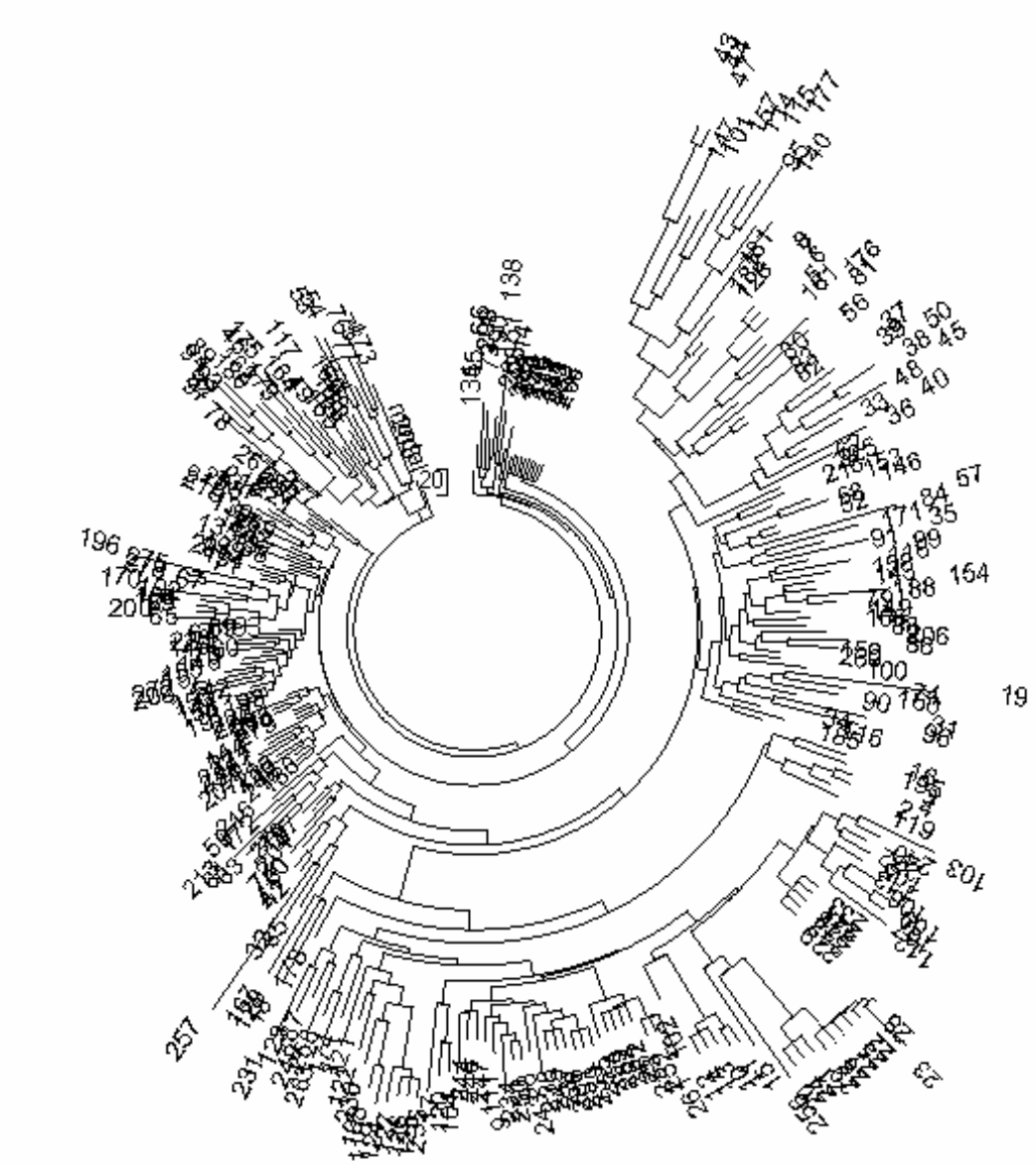

Figure 5. The 20-marker haplotype tree for 282 J1-M267 haplotypes, composed according to data published by Tofanelli et al (2009).

$$\frac{\sigma(\lambda)}{\lambda} \approx \frac{1}{\sqrt{24 \bullet 0.042}}$$

$$\frac{\sigma(\lambda)}{\lambda} \approx \frac{1}{\sqrt{24 \bullet 0.042}}(1+\frac{0.042}{2})$$

gives the margin of errors of 100%. The standard deviation of the mutation rate will be added on top of it, however, it will not add much.

The same results will be observed if two 12-marker haplotypes (which belong to the same haplogroup or a subclade) contain one mutation between them. Then for the 68% confidence interval the common ancestor of these two haplotypes lived between 1,140 ybp and the present time, and for the 95% confidence interval he lived between 1,725 ybp and the present time.

A similar in kind situation will be observed for two, three or four mutations between two 12-marker haplotypes in the FTDNA format. At the 95% confidence level the common ancestor of the two individuals would

have lived between 2,900 ybp, or 3625 ybp, or 4575 ybp, respectively, and the present time, even when the mutation rate is determined with the 5% accuracy.

Only with five mutations between two 12-marker haplotypes it is unlikely – with the 95% confidence level – that the common ancestor lived within 12 generations, that is 300 years before present. He rather lived between 5500 ybp and 300 ybp.

Obviously, we have considered quite different situations, typically with multiple haplotype series, which progressively reduce the standard deviations of a number of average mutations per marker, in some cases with hundreds or even well over a thousand of 25-marker haplotypes, down to 2.0% (with 750 of 19-marker haplotypes) and 1.1% (with 1527 of 25-marker haplotypes, collectively having almost 40 thousand alleles). In those cases the standard deviation of a time span to a common ancestor is dominated by the standard deviation of the mutation rate. However, relative values of TSCA's will stay with a reasonably good accuracy.

5) What are limitations of the logarithmic method of determining of a time span to a common ancestor?

The logarithmic method has a firm basis, as it is shown above. In this work it is applied to the number of non-mutated, or base haplotypes, which disappear in accord with the mutation rate. The faster the mutation rate, the faster base haplotypes disappear from the haplotype series. For 12-, 25-, 37- and 67-marker haplotypes, half of base haplotypes will disappear (become mutated) after 32, 15, 8, and 5 generations, respectively. Clearly, for 37- and 67-marker haplotypes the logarithmic method is hardly applicable. For large series of 25-marker haplotypes it can be quick and convenient. For example, in a series of 200 of 25-marker haplotypes, even after 65 generations, that is 1625 years, as many as 10 base haplotypes will still be present. This can easily be seen from $\ln(200/10)/0.046 = 65$ generations. For 12-marker haplotypes $\ln(200/10)/0.022 = 136$, that is 10 base haplotypes in the series will still stay after about 3400 years.

To use the logarithmic method is not recommended when less than 4-5 base haplotypes present in the haplotype series because of the large uncertainties.

A concern that a considerable data is discarded in order to focus on unmutated haplotypes is a non-issue, since this method is recommended to be applied along with the traditional method of mutation counting. Only when the two methods give similar results (in terms of a number of generations or years to the common ancestor), the results are justified. If the results are significantly different, such as by 1.4-2 times or higher, neither of the results can be accepted. A difference of 1.3-1.4 times is conditionally acceptable, however, results will have a high margin of error.

6) Are a few dozen haplotypes enough to characterize a population of thousands, hundreds of thousands, or millions of people?

This is a typical question which always arises – sooner or later – in discussions on DNA genealogy. An indirect answer is – mixing of haplotypes in the population is much more important than a number of haplotypes.

One example is given above. It turned out that 17 of 25-marker haplotypes of the Basques (Haplogroup R1b1) gave the same results as those for a series of 750 of 19-marker Iberian haplotypes. The results were practically the same in terms of the base (ancestral) haplotype and the timespan to a common ancestor of the 17 or 750 individuals - 3675±520 years bp and 3625±370 years bp, respectively. 44 of 12-marker haplotypes gave 3625±490 years bp to a common ancestor.

Table 3
A Short History of Results from Collecting Available 25-Marker R1a1 Haplotypes

Date	Number of haplotypes	Total number of mutations	Timespan to a common ancestor (years)	Reference
June 2008	26	178	4400 ± 550	Klyosov, 2008e
November 2008	44	326	4825 ± 550	Klyosov, 2008f
January 2009	58	423	4725 ± 520	This article, Part 2
March 2009	98	711	4700 ± 500	
June 2009	110	804	4750 ± 500	This article

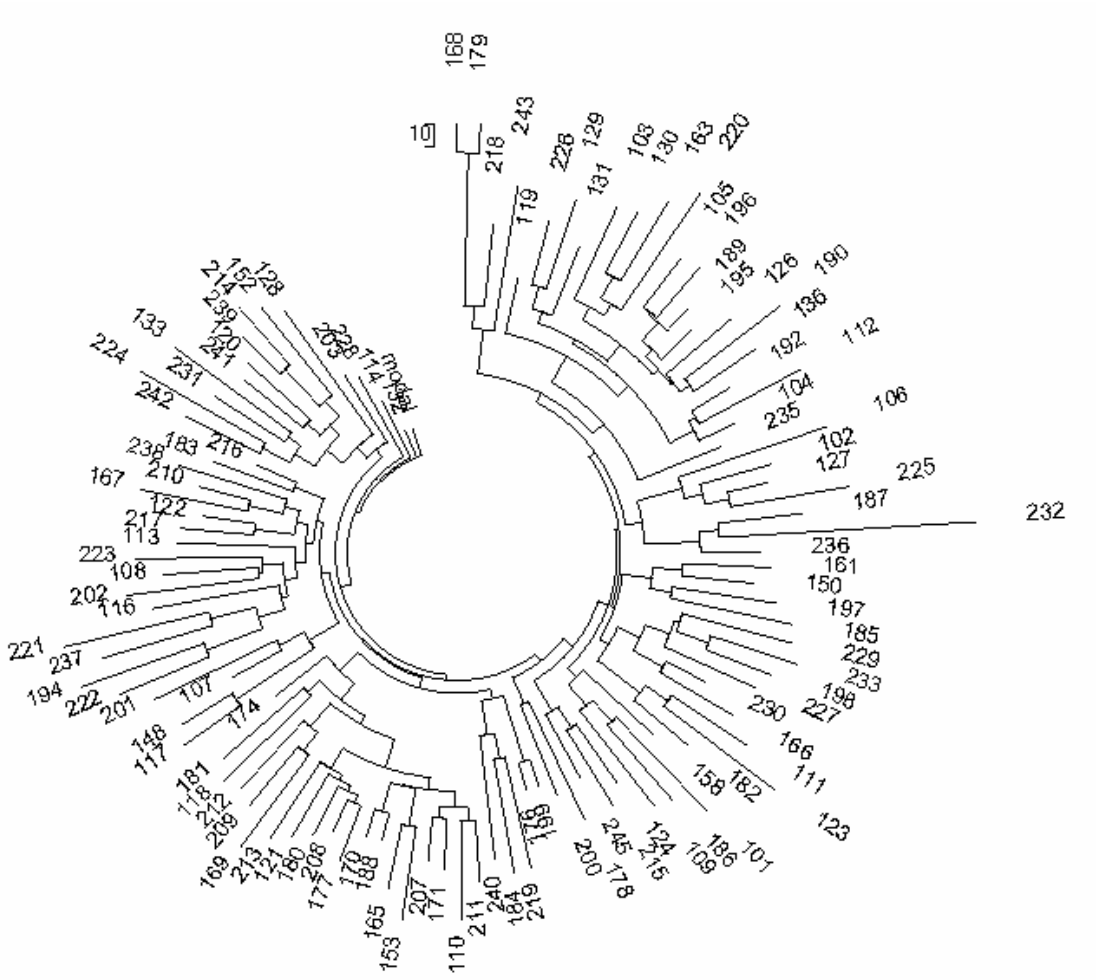

Figure 6. The 25-marker haplotype tree for Russia and Ukraine, haplogroup R1a1. The 110-haplotype tree was composed from data of YSearch database and provided by the individuals.

One can see that the number of haplotypes in this case merely decreases a margin or error, since it depends on a number of mutations in a given series of haplotypes.

Another example can be given with the Slavic haplotypes (Haplogroup R1a1) in Russia and Ukraine, shown in Figure 6. In Table 3 is a short history of collection of available 25-marker haplotypes from said region, from YSearch database and directly from the individuals, along with the results of analysis.

One can see that all five haplotype sets gave the same dating within a margin of error. All the five sets coalescent to the same base (ancestral) haplotype

13 25 16 10(11) 11 14 12 12 10 13 11 30 15 9 10 11 11 24 14 20 32 12 15 15 16

in which only the fourth allele (DYS391) fluctuates between 10 and 11, and, actually, from 10.46 to 10.53, on average.

The reason for such reproducibility is simple: all those haplotypes were collected across all Russia and Ukraine, from the Carpathian Mountains in the West to the Pacific ocean in the East, and from the frozen tundra North to the Iranian border South. The selection is certainly a representative one.

Asymmetry of Mutations

This phenomenon was considered in this study using a number of specific examples. It was shown that when mutations are fairly symmetrical, that is both-sided (the degree of asymmetry is around 0.5), no corrections to the TSCA are needed. The TSCA is typically calculated as an average number of mutations per marker, divided

by the appropriate average mutation rate (Table 1) and corrected for back mutations using Table A. Alternatively, Equation (1) can be employed in its simplified version (with $a_i = 1$ for a symmetrical pattern of mutations). Even when the degree of mutations reaches about 0.66 (two-thirds of mutations in the haplotype series are one-sided), the respective correction is not significant and is typically within a corresponding margin of error (for the Basque R1b1b2 haplotypes it was of 5 generations, that was 3.6% of total). For the degree of asymmetry around 0.85 (as in the case of the Isles I1 haplotypes, considered above) the necessary correction can be around 10-20 generations (250 years) on the 120-generation span, that is reach 8-16%, and further increase with an "age" of the TSCA. At a fully one-sided mutation pattern (the degree of asymmetry equal to 1), it completely nullifies the correction for reverse mutations, hence, can increase the calculated TSCA by 750 years at 4000 years and by 1200 years at 5000 years to the common ancestor, respectively, and continue to grow. This is, of course, the extreme case of asymmetry, however, it should be taking into consideration.

Overall, this section has essentially shown how to make calculations and interpret data extracted from a number of mutations and a number of base haplotypes in haplotype sets. The subsequent paper (Part II) will follow with principal illustrations and conclusions, without repeating the methodology.

Acknowledgements

I am indebted to Theresa M. Wubben and Dmitry Adamov for valuable discussions.

Web Resources

Ashina Project
http://www.familytreedna.com/public/AshinaRoyalDynasty/

Haplogroup Q Project
http://m242.haplogroup.org

ISOGG 2009 Y-DNA Haplogroup Q and its Subclades
http://www.isogg.org/tree/ISOGG_HapgrpQ09.html

PHYLIP: Phylogeny Inference Package
http://bioweb.uwlax.edu/GenWeb/Evol_Pop/Phylogenetics/Phylip/phylip.htm

References

Adamov DS, Klyosov AA (2008a) Theoretical and practical estimates of reverse mutations in haplotypes of Y chromosome (in Russian). *Proc Russian Academy of DNA Genealogy* (ISSN 1942-7484), 1:631-645.

Adamov DS, Klyosov AA (2008b) Evaluation of an "age" of populations from Y chromosome using methods of average square distance (in Russian). *Proc Russian Academy of DNA Genealogy* (ISSN 1942-7484), 1:855-905.

Adamov DS, Klyosov AA (2009a) Evaluation of an "age" of populations from Y chromosome. Part I. Theory (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 2:81-92.

Adamov DS, Klyosov AA (2009b) Practical methods for determining "age" of common ancestors for large haplotype series (in Russian). *Proc Russian Academy of DNA Genealogy* (ISSN 1942-7484), 2:422-442.

Adamov DS, Klyosov AA (2009c) Evaluation of an "age" of populations from Y chromosome. Part II. Statistical considerations (in Russian). *Proc Russian Academy of DNA Genealogy* (ISSN 1942-7484), 2:93-103.

Adams SM, Bosch E, Balaresque, PL, Ballereau SJ, Lee AC, Arroyo E, López-Parra AM, Aler M, Gisbert-Grifo MS, Brion M, et al. (2008) The Genetic Legacy of Religious Diversity and Intolerance: Paternal Lineages of Christians, Jews, and Muslims in the Iberian Peninsula. *Am J Hum Genet*, 83:725-736.

Athey W (2007) Mutation rates - who's got the right values? *J Genet Geneal*, 3(2):i-iii.

Cervantes A (2008) Basque DNA Project.

Bittles AH, Black ML, Wang W (2007) Physical anthropology and ethnicity in Asia: the transition from anthropology to genome-based studies. *J Physiol Anthropol*, 26:77-82.

Chandler JF (2006) Estimating per-locus mutation rates. *J Genet Geneal*, 2:27-33.

Cordaux R, Bentley G, Aunger R, Sirajuddin SM, Stoneking M (2004) Y-STR haplotypes from eight South Indian groups based on five loci. *J Foren Sci*, 49: 1-2.

Cruciani F, LaFratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon J-M, Crivellaro F, Benincase T, Pascone R, et al (2007) Tracing past human male movement in Northern/Eastern Africa and Western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol*, 24:1300-1311.

Clan Donald USA Genetic Genealogy Project (2008)

Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.. See Web Resources.

Fornarino S, Pala M, Battaglia V, Maranta R, Achilli A, Modiano G, Torroni A, Semino O, Santachiara-Benerecetti SA (2009) Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evolutionary Biol*, 9:154.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995a) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci US*, 92, 6723-6727.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995b) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139: 463-471.

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjabazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, et al (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci US*, 97:6769-6774.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet*, 6:799-803.

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. TIG, 11:449-456.

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, et al (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *Am J Hum Genet*, 64:817-831.

Kayser M, Roewer L, Hedman M, Henke L, Hemke J, Braue, S, Kruger C, Krawczak M, Nagy M, Dobosz T, et al (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580-1588.

Kerchner C (2008) Y-STR haplotype observed mutation rates in surname projects study and log, http://kerchner.com/cgi-kerchner/ystrmutationrate.cgi

Klyosov AA (2008a) The features of the "West European" R1b haplogroup (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:568-629.

Klyosov AA (2008b) Origin of the Jews via DNA genealogy. *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:54-232.

Klyosov AA (2008c) Basic rules of DNA genealogy (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:252-348.

Klyosov AA (2008d) Calculations of time spans to common ancestors for haplotypes of Y chromosome (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:812-835.

Klyosov AA (2008e) Where Slavs and Indo-Europeans came from? (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:400-477.

Klyosov AA (2008f) Haplotypes of haplogroup R1a1 in the post-Soviet region (in Russian). *Proc Russian Academy DNA Geneal* (ISSN 1942-7484), 1:947-957.

Mertens G (2007) Y-Haplogroup frequencies in the Flemish population. *J Genet Geneal*, 3:19-25.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006) Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *J Foren Sci*, 51:64-75.

Nebel A, Filon D, Weiss DA, Weale M, Faerman M, Oppenheim A, Thomas M (2000) High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. *Hum Genet,* 107:630-641.

Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M, Oppenheim A (2001) The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am J Hum Genet*, 69:1095-1112.

Nei M (1995) Genetic support for the out-of Africa theory of human evolution. *Proc Natl Acad Sci US*, 92:6720-6722.

Nordtvedt KN (2008) More realistic TMRCA calculations. *J Genet Geneal*, 4:96-103.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science,* 290:1155-1159.

Skorecki K, Selig S, Blazer S, Bradman R, Bradman N, Warburton PJ, Ismajlowicz M, Hammer MF (1997) Y chromosomes of Jewish Priests. *Nature*, 385:32.

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenic trees from microsatellite DNA. *Genetics*, 144:389-399.

Thanseem I., Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy MB, Reddy AG, Singh L (2006) Genetic affinities among the lower castes and tribal groups of India: Inference from Y chromosome and mitochondrial DNA. *BMC Genet*, 7(1):42.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851-862.

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. *Nature*, 394:138-140.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y Chromosomes traveling South: the Cohen Modal Haplotype and the origin of the Lemba - the "Black Jews of Southern Africa." *Am J Hum Genet*, 66:674-686.

Tofanelli S, Ferri G, Bulayeva K, Caciagli L, Onofri V, Taglioli L, Bulayev O, Boschi I, Alù M, Berti A, Rapone C, Beduschi G, Luiselli D, Cadenas AM, Awadelkarim KD, Mariani-Costantini R, Elwali NE, Verginelli F, Pilli E, Herrera RJ, Gusmão L, Paoli G, Capelli C (2009) J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur J Hum Genet*, 17:1520-1524.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, et al (2000) Y chromosome sequence variation and the history of human populations. *Nature Genetics*, 26:358-361.

Walsh B (2001) Estimating the time to the most common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*, 158:897-912.

Weale ME, Yepiskoposyan L, Jager RF, Hovhannisyan N, Khudoyan A, Burbage-Hall O, Bradman N, Thomas M (2001) Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. *Hum Genet,* 109:659-674.

Zhivotovsky LA, Feldman MW (1995) Microsatellite variability and genetic distances. *Proc Natl Acad Sci US*, 92:11549-11552.

Appendix A
Tabular Calculation Aides for TSCA

Table A
A number of generations (at 25 years per generation, calibrated), calculated for average mutation rates 0.002 mutations per marker per generation. For haplotypes with an average mutation rates different from 0.002 per marker (see Table 1, third column), an average number of mutations per marker should be re-calculated (see the main text).

An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation		An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per genera-
	Without correction for back mutations	With correction for back mutations				Without correction for back mutations	With correction for back mutations	
0.002	1	1	25		0.082	41	43	1075
0.004	2	2	50		0.084	42	44	1100
0.006	3	3	75		0.086	43	45	1125
0.008	4	4	100		0.088	44	46	1150
0.010	5	5	125		0.09	45	47	1175
0.012	6	6	150		0.092	46	48	1200
0.014	7	7	175		0.094	47	50	1250
0.016	8	8	200		0.096	48	51	1275
0.018	9	9	225		0.098	49	52	1300
0.020	10	10	250		0.1	50	53	1325
0.022	11	11	275		0.102	51	54	1350
0.024	12	12	300		0.104	52	55	1375
0.026	13	13	325		0.106	53	56	1400
0.028	14	14	350		0.108	54	57	1425
0.030	15	15	375		0.11	55	58	1450
0.032	16	16	400		0.112	56	60	1500
0.034	17	17	425		0.114	57	61	1525
0.036	18	18	450		0.116	58	62	1550
0.038	19	19	475		0.118	59	63	1575
0.040	20	20	500		0.12	60	64	1600
0.042	21	21	525		0.122	61	65	1625
0.044	22	22	550		0.124	62	66	1650
0.046	23	23	575		0.126	63	67	1675
0.048	24	25	625		0.128	64	68	1700
0.050	25	26	650		0.13	65	69	1725
0.052	26	27	675		0.132	66	71	1775
0.054	27	28	700		0.134	67	72	1800
0.056	28	29	725		0.136	68	73	1825
0.058	29	30	750		0.138	69	74	1850
0.060	30	31	775		0.14	70	75	1875
0.062	31	32	800		0.142	71	77	1925
0.064	32	33	825		0.144	72	78	1950
0.066	33	34	850		0.146	73	79	1975
0.068	34	35	875		0.148	74	80	2000
0.070	35	36	900		0.15	75	81	2025
0.072	36	37	925		0.152	76	83	2075
0.074	37	38	950		0.154	77	84	2100
0.076	38	40	1000		0.156	78	85	2125
0.078	39	41	1025		0.158	79	86	2150
0.080	40	42	1050		0.16	80	87	2175

Table A
(continued)

An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation		An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation
	Without correction for back mutations	With correction for back mutations				Without correction for back mutations	With correction for back mutations	
0.162	81	89	2225		0.252	126	145	3625
0.164	82	90	2250		0.254	127	146	3650
0.166	83	91	2275		0.256	128	147	3675
0.168	84	92	2300		0.258	129	148	3700
0.170	85	93	2325		0.260	130	149	3725
0.172	86	95	2375		0.262	131	150	3750
0.174	87	96	2400		0.264	132	152	3800
0.176	88	97	2425		0.266	133	154	3850
0.178	89	98	2450		0.268	134	155	3875
0.180	90	99	2475		0.270	135	156	3900
0.182	91	100	2500		0.272	136	158	3950
0.184	92	102	2550		0.274	137	159	3975
0.186	93	103	2575		0.276	138	161	4025
0.188	94	104	2600		0.278	139	162	4050
0.190	95	105	2625		0.280	140	163	4075
0.192	96	107	2675		0.282	141	164	4100
0.194	97	108	2700		0.284	142	166	4150
0.196	98	109	2725		0.286	143	167	4175
0.198	99	110	2750		0.288	144	168	4200
0.200	100	111	2775		0.290	145	169	4225
0.202	101	112	2800		0.292	146	170	4250
0.204	102	114	2850		0.294	147	172	4300
0.206	103	115	2875		0.296	148	174	4350
0.208	104	116	2900		0.298	149	175	4375
0.210	105	117	2925		0.300	150	176	4400
0.212	106	118	2950		0.302	151	178	4450
0.214	107	120	3000		0.304	152	179	4475
0.216	108	121	3025		0.306	153	180	4500
0.218	109	122	3050		0.308	154	182	4550
0.220	110	123	3075		0.310	155	183	4575
0.222	111	124	3100		0.312	156	184	4600
0.224	112	126	3150		0.314	157	186	4650
0.226	113	128	3200		0.316	158	187	4675
0.228	114	129	3225		0.318	159	188	4700
0.230	115	130	3250		0.320	160	190	4750
0.232	116	132	3300		0.322	161	192	4800
0.234	117	133	3325		0.324	162	193	4825
0.236	118	134	3350		0.326	162	195	4875
0.238	119	135	3375		0.328	164	196	4900
0.240	120	136	3400		0.330	165	197	4925
0.242	121	138	3450		0.332	166	198	4950
0.244	122	140	3500		0.334	167	200	5000
0.246	123	141	3525		0.336	168	202	5050
0.248	124	142	3550		0.338	169	203	5075
0.250	125	143	3575		0.340	170	204	5100

Table A
(continued)

An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation		An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation
	Without correction for back mutations	With correction for back mutations				Without correction for back mutations	With correction for back mutations	
0.342	171	206	5150		0.432	216	274	6850
0.344	172	208	5200		0.434	217	276	6900
0.346	173	210	5250		0.436	218	278	6950
0.348	174	211	5275		0.438	219	280	7000
0.35	175	212	5300		0.44	220	281	7025
0.352	176	214	5350		0.442	221	282	7050
0.354	177	216	5400		0.444	222	284	7100
0.356	178	217	5425		0.446	223	286	7150
0.358	179	218	5450		0.448	224	288	7200
0.36	180	219	5475		0.45	225	289	7225
0.362	181	220	5500		0.452	226	290	7250
0.364	182	222	5550		0.454	227	292	7300
0.366	183	224	5600		0.456	228	294	7350
0.368	184	225	5625		0.458	229	296	7400
0.37	185	226	5650		0.46	230	297	7425
0.372	186	228	5700		0.462	231	298	7450
0.374	187	229	5725		0.464	232	300	7500
0.376	188	230	5750		0.466	233	302	7550
0.378	189	232	5800		0.468	234	304	7600
0.38	190	234	5850		0.47	235	306	7650
0.382	191	236	5900		0.472	236	308	7700
0.384	192	238	5950		0.474	237	310	7750
0.386	193	239	5975		0.476	238	311	7775
0.388	194	240	6000		0.478	239	313	7825
0.39	195	241	6025		0.48	240	314	7850
0.392	196	242	6050		0.482	241	316	7900
0.394	197	244	6100		0.484	242	318	7950
0.396	198	246	6150		0.486	242	320	8000
0.398	199	248	6200		0.488	244	322	8050
0.4	200	249	6225		0.49	245	323	8075
0.402	201	250	6250		0.492	246	324	8100
0.404	202	252	6300		0.494	247	326	8150
0.406	203	254	6350		0.496	248	328	8200
0.408	204	256	6400		0.498	249	330	8250
0.41	205	257	6425		0.5	250	331	8275
0.412	206	258	6450		0.502	251	332	8300
0.414	207	260	6500		0.504	252	334	8350
0.416	208	262	6550		0.506	253	336	8400
0.418	209	264	6600		0.508	254	338	8450
0.42	210	265	6625		0.51	255	340	8500
0.422	211	266	6650		0.512	256	342	8550
0.424	212	268	6700		0.514	257	344	8600
0.426	213	270	6750		0.516	258	346	8650
0.428	214	272	6800		0.518	259	348	8700
0.43	215	273	6825		0.52	260	349	8725

Table A
(continued)

An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation		An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation
	Without correction for back mutations	With correction for back mutations				Without correction for back mutations	With correction for back mutations	
0.522	261	350	8750		0.612	306	436	10900
0.524	262	352	8800		0.614	307	438	10950
0.526	263	354	8850		0.616	308	440	11000
0.528	264	356	8900		0.618	309	442	11050
0.530	265	358	8950		0.620	310	444	11100
0.532	266	360	9000		0.622	311	446	11150
0.534	267	362	9050		0.624	312	448	11200
0.536	268	364	9100		0.626	313	450	11250
0.538	269	366	9150		0.628	314	452	11300
0.540	270	367	9175		0.630	315	454	11350
0.542	271	368	9200		0.632	316	456	11400
0.544	272	370	9250		0.634	317	458	11450
0.546	273	372	9300		0.636	318	460	11500
0.548	274	374	9350		0.638	319	462	11550
0.550	275	376	9400		0.640	320	464	11600
0.552	276	378	9450		0.642	321	466	11650
0.554	277	380	9500		0.644	322	468	11700
0.556	278	382	9550		0.646	323	470	11750
0.558	279	384	9600		0.648	324	472	11800
0.560	280	385	9625		0.650	325	474	11850
0.562	281	387	9675		0.652	326	476	11900
0.564	282	389	9725		0.654	327	478	11950
0.566	283	391	9775		0.656	328	480	12000
0.568	284	393	9825		0.658	329	482	12050
0.570	285	395	9875		0.660	330	485	12125
0.572	286	396	9900		0.662	331	487	12175
0.574	287	398	9950		0.664	332	490	12250
0.576	288	400	10000		0.666	333	492	12300
0.578	289	402	10050		0.668	334	494	12350
0.580	290	404	10100		0.670	335	496	12400
0.582	291	406	10150		0.672	336	498	12450
0.584	292	408	10200		0.674	337	500	12500
0.586	293	410	10250		0.676	338	502	12550
0.588	294	412	10300		0.678	339	504	12600
0.590	295	414	10350		0.680	340	506	12650
0.592	296	416	10400		0.682	341	508	12700
0.594	297	418	10450		0.684	342	510	12750
0.596	298	420	10500		0.686	343	512	12800
0.598	299	422	10550		0.688	344	514	12850
0.600	300	424	10600		0.690	345	517	12925
0.602	301	426	10650		0.692	346	519	12975
0.604	302	428	10700		0.694	347	522	13050
0.606	303	430	10750		0.696	348	524	13100
0.608	304	432	10800		0.698	349	526	13150
0.610	305	434	10850		0.70	350	528	13200

Table A
(continued)

An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation		An average number of mutations per marker per haplotype at mutation rate of 0.002 mutations/ marker/gen	A number of generations to a common ancestor for a given set of haplotypes		Years to a common ancestor, with correction for back mutations, assuming 25 years per generation
	Without correction for back mutations	With correction for back mutations				Without correction for back mutations	With correction for back mutations	
0.704	352	533	13325		0.94	470	835	20875
0.708	354	537	13425		0.95	475	850	21250
0.712	356	542	13550		0.96	480	864	21600
0.716	358	546	13650		0.97	485	879	21975
0.720	360	551	13775		0.98	490	894	22350
0.724	362	556	13900		0.99	495	910	22750
0.728	364	560	14000		1.00	500	925	23125
0.732	366	565	14125		1.01	505	940	23500
0.736	368	570	14250		1.02	510	956	23900
0.740	370	574	14350		1.03	515	972	24300
0.744	372	578	14450		1.04	520	988	24700
0.748	374	582	14550		1.05	525	1004	25100
0.752	376	588	14700		1.06	530	1020	25500
0.756	378	592	14800		1.07	535	1037	25925
0.760	380	597	14925		1.08	540	1054	26350
0.764	382	602	15050		1.09	545	1070	26750
0.768	384	606	15150		1.10	550	1087	27175
0.772	386	611	15275		1.11	555	1104	27600
0.776	388	616	15400		1.12	560	1122	28050
0.780	391	624	15600		1.13	565	1139	28475
0.784	393	629	15725		1.14	570	1157	28925
0.788	395	634	15850		1.15	575	1174	29350
0.792	397	640	16000		1.16	580	1192	29800
0.796	399	644	16100		1.17	585	1210	30250
0.80	401	649	16225		1.18	590	1229	30725
0.815	403	655	16375		1.19	595	1247	31175
0.810	406	662	16550		1.20	600	1266	31650
0.815	408	668	16700		1.30	650	1460	36500
0.820	411	674	16850		1.40	700	1672	41800
0.825	413	680	17000		1.50	750	1900	47500
0.830	416	687	17175		1.60	800	2140	53500
0.835	418	693	17325		1.70	850	2400	60000
0.840	421	700	17500		1.80	900	2674	66850
0.845	423	706	17650		1.90	950	2960	74000
0.850	426	713	17825		2.00	1000	3280	82000
0.855	428	720	18000		2.10	1050	3600	90000
0.860	431	727	18100		2.20	1100	3920	98000
0.865	433	733	18835		2.30	1150	4280	107000
0.870	436	740	18500		2.40	1200	4640	116000
0.880	441	753	18825		2.50	1250	5040	126000
0.890	446	767	19100		2.60	1300	5440	136000
0.900	455	792	19800		2.70	1350	5840	146000
0.910	460	806	20150		2.80	1400	6280	157000
0.920	465	821	20525		2.90	1450	6720	168000
0.930	470	835	20875		3.00	1500	7200	180000