Journal: www.jogg.info
Originally Published: Volume 5, Number 2 (Fall 2009)
Reference Number: 52.008

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: INTRODUCTION

Author(s): T. Whit Athey

Introduction to a Special Section on Alternative Methods of Analysis for Y-STR Clusters and the Determination of the Time to the Most Recent Common Ancestor

T. Whit Athey, Editor

Introduction

There is something about the idea of a "ticking" clock in our genes that intrigues us. In principle, if we have a set of Y-STR haplotypes that all derive from a common ancestor, we could calculate, from the observed mutations and the mutation rates, the approximate number of years or generations ago that the common ancestor lived, the Time to the Most Recent Common Ancestor (TMRCA). But, there are complications to this simple-sounding idea.

In this special section we present several new or alternative approaches to the grouping of haplotypes from possibly related individuals, and for the calculation of the TMRCA. In the calculation of the TMRCA, both new and traditional approaches are used, and it would be fair to say that the new approaches are not generally accepted at present--they are proposed approaches that will need confirmation in other hands to gain acceptance. However, because of the substantial interest in this subject, we are presenting these articles in the hope that one or more of them will ultimately be validated and will come to be an accepted way of analyzing Y-STR clusters. If that doesn't happen, then at least we may have stimulated some more discussion on the subject.

If we have a set of haplotypes, our analysis will only be valid if all of the haplotypes really descend from a common ancestor and if they are representative of the descendants of that ancestor. Two of the articles in this special section address the issue of how to form clusters and confirm their integrity as a clade (Gwozdz, 2009a; Howard, 2009a). The approaches are quite different, with that of Howard being more empirical and that of Gwozdz involving a complex formalism.

The other main focus in the articles in this section is the calculation of the TMRCA and the proper mutation rates and correction factors to use in this process.

Mutation rates represent the rate at which the genetic clock ticks. There are two basic types of mutation rates for STR markers: (1) father-son rates--those derived from father-son transmissions (Gusmão, 2005), and (2) effective rates--those derived from descendants of a historical figure whose birth date is known, or from descendants of a founding population where the founding date is known, at least approximately (Zhivotovsky, 2004). Father-son rates do not need to be derived only from living father-son pairs, but can also be derived from a genealogical tree (e.g., Kerchner, 2008), where the common ancestor's haplotype may be reconstructed unambiguously and the genealogy of each living descendant available for testing is known. The important thing to know about father-son rates is that they be determined from a genealogy only when the whole genealogy is known so that a mutation that has occurred just once in the genealogy, but perhaps appears in two or more descendants, is only counted once in calculating the mutation rate. The genealogical structure is also important for interpreting the results, as we will show below.

Effective mutation rates may be calculated when the genealogy is unknown, but the time of birth of the common ancestor is known. If the same mutation is showing in two or more subjects who are tested, then it will be unclear whether both subjects inherited the mutation from an intermediate ancestor, or whether the mutation occurred independently in both lines, but the whole issue is side-stepped in the calibration process.

It is also important to identify precisely what it is we are looking for in a TMRCA and what we are actually calculating. I believe that if we are considering a haplotype series from a group of people who are all descended from the same common ancestor, most people want to know the average lineage length back to the common ancestor. That is, we want to be able to say, "The common ancestor of this set of subjects is 12 generations back, on average." We must add the part about "on average" because the common ancestor will be slightly different numbers of generations back for different subjects. The average lineage length is what we are

Address for correspondence: T. Whit Athey, wathey@hprg.com

seeking in a TMRCA, but is that what we really calculate with traditional methods?

For example, when a genealogy is used for calculating father-son mutation rates as in the Kerchner project, we take the total number of independent mutations and divide by the total number of father-son transmissions in the genealogy, and we get the mutation rate as the average number of mutations per transmission per haplotype. If desired, we can also determine the average marker mutation rate by dividing by the number of markers in the haplotype.

Once we have the mutation rate we can apply it to clusters where the TMRCA is unknown--we turn the operation around and start with the number of mutations in the new series of haplotypes and divide by the mutation rate and we get the number of transmissions. The number of transmissions??? That's not what most of us think of when we think of the TMRCA! However, we usually go further and divide the transmissions by the number of haplotypes, and this gives us what most people are calling the TMRCA in terms of the average number of transmissions per subject. However, this quantity, the average number of transmissions per subject, is different from the average lineage length because many of the transmissions will appear in more than one lineage. The average number of transmissions per subject will always be smaller than the average lineage length, usually by 10-20%.

Consider a simple example of a grandfather with two sons, each of whom have two sons, as shown in **Figure 1**:

In this simple example there have been six father-to-son transmissions and we end up with four third-generation descendants of the grandfather. This results in an average of 1.5 transmissions per third-generation subject. This value of 1.5 is somewhat smaller than the average lineage length, which is 2.0 for the four grandsons. Let's assume for a moment that we have a million markers available to test in these four grandsons (this is just to give us good statistics with this small number of generations and subjects), and assume that we know that the father-son mutation rate for each marker was 1/1000 mutations per generation. In the first generation each son would get about 1000 mutations. Each son's sons would get his 1000 mutations plus about 1000 more. So each grandson would be different from the grandfather on about 2000 markers, though between grandsons, we would see some shared mutations. If we carefully noted the number of *independent* mutations, which is about 6000, and divided that number by the father-son rates, we would get a result of about 1.5, because it's the average transmissions per subject that the father-son rates give us. We would not get what we wanted--the average lineage length of 2.0. However, if we ignore the possibility of counting the same mutation more than once, and simply counted the total mutations in the grandsons, then divided by the rates, we would get 8000/(4 x 0.001 x 1000000) = 2, which is the answer we wanted. However, this seems fortuitous and perhaps results from the simplicity and regularity of this particular genealogy.

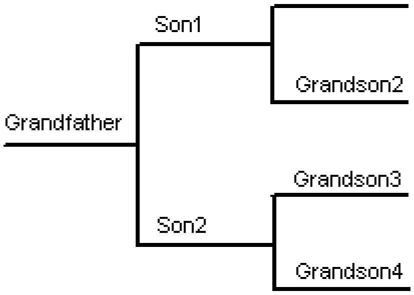

Figure 1. Simple three-generation desendancy model.

Consider a real, rather than synthetic, example. In my own Athey surname project, at 37 markers, we have 19 participants who all descend from the same common ancestor born in 1642. Since we know the whole descendancy tree for these 19 participants, we can see that they are 8, 9, or 10 generations removed from the common ancestor, and the average lineage length for all participants is 8.8. There have been 132 father-son transmissions in the tree, for an average number of transmissions per participant of 6.9. The value of 8.8 is what I believe most people think they are calculating when they carry out a TMRCA calculation, but if father-son rates are used, then it is not what they get--they get instead, as already pointed out, 6,9, the average number of transmissions per participant--a value that is 21% lower.

The (father-son) mutation rate for this set of 19 subjects on a 37-marker haplotype may be found by dividing the number of mutations in the genealogy by the number of haplotype transmissions, or 23/132 = 0.174. One can go further and calculate the average individual marker mutation rate by dividing by 37, and we get 0.174/37 = 0.0047, which is close to the average in the Kerchner project (0.0042) and to Chandler's calibrated average (0.0049), so that means that our set of Athey clocks has been running very close to the average rate. This is simply a lucky result, however, because we don't have a large enough series of haplotypes, to assure that we would be that close.

Usually, when one has a cluster of closely matching participants of the same surname, implying that they share a common ancestor, one does not know the genealogy. If the genealogy were known, there would be no point in calculating the TMRCA--it would be obvious from the genealogy. However, if one just counts mutations in living participants, there is the risk of counting a mutation that has occurred only once (but has been passed down to multiple participants) multiple times, because the only practical approach is usually to just compare each participant with the reconstructed ancestral haplotype. When we count the same mutation multiple times, we will overestimate the TMRCA.

Most commonly, the father-son mutation rates are used to calculate the TMRCA in a cluster with unknown genealogy. Therefore, it appears that we will have partially offsetting errors occurring in the TMRCA calculations. When we have multiple counting of mutations we will be overestimate the TMRCA as a result. However, the result of the calculation if we use father-son mutation rates is actually the average number of transmissions per participant, which is a lower number than the desired average lineage length. These two errors partially compensate and result in a value for TMRCA that is fairly close to what we wanted, or exactly the right answer in the case of a very regular genealogy like that of Figure 1.

Using the Athey example again, the first error from over-counting mutations results in 27 mutations instead of the actual 23 in the genealogy. This would result in an apparent average number of mutations per lineage, the so-called *rho factor*, of 27/19 = 1.42 which is 17% larger than the actual value of rho = 23/19 = 1.21. If we then divide the apparent average number of mutations per lineage (apparent rho) by the father-son mutation rate of .174 mutations per 37-marker haplotype. The rate is derived from the Athey data (just to remove this as an additional variable), we get a result of 1.42/.174 = 8.16, whereas using actual rho, 1.21/.174 = 6.95 (average number of transmissions per participant). However, if we call this value of 8.16 the TMRCA, then we are actually much closer to the true average lineage length of 8.8 than we would have been otherwise. Our TMRCA value was overestimated by a factor of 27/23 = 1.17 because of over-counting of mutations, while it was underestimated by a factor of 0.79 from applying father-son rates to a genealogical tree (the genealogical structure factor), with the final result being off by only by about 8% because of the compensating errors.

I have not seen a discussion of this particular phenomenon before, and possibly it is because the compensating errors bring us close to the right answer that we have overlooked it. However, it would seem that there may be cases where one should take these factors into account when attempting to determine the TMRCA for a set of haplotypes.

Another method often used in population studies is the average-square-distance (ASD) method. This method is needed if the time scale for the cluster is long enough that more than one mutation on the same marker in the same lineage becomes likely. The second (and any further) mutation on the same marker in the same lineage may be up or down, and could erase the evidence for any previous mutation. However, the second mutation could add to the first in the same direction, and the ASD method provides an average correction for these “random walk” effects. In the ASD method, the ancestral haplotype is reconstructed, often using the modal haplotype for the cluster, and the ASD is calculated for each of the living participants with respect to the ancestral haplotype. The results for each participant are averaged.

There is another ASD approach for the case when the ancestral or founder haplotype cannot be reconstructed, called the permutation method. Both of these are discussed in the article by Klyosov (2009a). If is sometimes claimed that the ASD methods avoid the genealogical structure problem, but a simple example will show that this is not true. Consider the case where a mutation in a particular marker occurs only once in the history of the cluster. If that mutation occurred in the transmission from the common ancestor to one of his two sons, then on the order of half of the present day descendants of that son would show his value on that marker. However, if that same mutation only occurred very late in the history of the cluster, it might show up in only one participant. The contribution to the ASD from half the participants having that mutation will be different from the case of only one participant having that mutation. It is true that the ASD for one participant with respect to the ancestral haplotype should represent an unbiased estimate of the number of generations between that participant and the ancestor, but when several participants have the same inherited mutation that has actually occurred only once, the average of the ASD for that group of participants will no longer be unbiased. The ASD approach assumes that mutations on a particular marker in one subject will be independent of the same mutation in a different participant. The ASD method is likely to be used only in applications where mutations have occurred multiple times on each marker, so in practice it will be difficult or impossible to determine if some mutations are not independent. One can only assume that all are independent, but this will not be true in general, and the difference that this effect makes, appears to be exactly the same as the mutation-over-counting effect discussed earlier.

The ASD method commonly uses father-son mutation rates, so the same genealogical structure factor will apply to this method as well. One can easily see that this is true because the ASD approach becomes the simpler linear approach in the limit that no mutation is more than one unit from the ancestral value--the two approaches must yield the same result in that limit.

Is there any way around these difficulties? Following are some approaches to take these effects into account.

Method 0 – We could ignore the bothersome details. This seems to be the most popular approach.

Method 1a -- In cases where we could estimate the true number of mutations that has occurred in the genealogical tree, or when it was clear from the pattern of mutations that all mutations had occurred independently (e.g., when no two participants showed the same mutation), then we could use only the independent mutations in the calculation. In this method we would go ahead and use the father-son mutation rates, but we would apply a correction factor to our TMRCA to correct for the genealogical structure. We can call this factor a *genealogical structure factor* or *population structure factor.* This genealogical structure factor is a property of the descendancy tree alone--it is equal to the ratio of the number of transmissions in the genealogy to the sum of all the individual lineage lengths. This factor would probably range from about 0.75 to 0.85 for typical surname clusters or genealogical trees. We would need to calculate this factor for a large number of trees and average them to get the best factor to apply to a cluster with an unknown genealogy. For example, the correction factor for the Athey example would be 0.79. This is the only correction we would need to apply since we are assuming no over-counting of mutations.

Method 1b -- In this case we would typically have a large cluster with several cases of the same mutation showing in multiple participants. We would then need to correct for both the genealogical structure and the over-counting of mutations. These factors work in opposite directions (compensating errors) as discussed earlier For the Athey example, the two factors combined would be

$$(132/167)(27/24) = 0.927$$

We could then use the combined correction factor on new clusters. Note that while the genealogical structure factor will probably vary only within a narrow range, there is potentially a larger range for the correction for over-counting of mutations, and there will often be no good way of estimating the degree of over-counting. The best situation would be a set of haplotypes where it appears that no over-counting can occur (e.g., the same mutation does not appear in more than one participant), taking us back to Method 1a.

Method 2 -- In this method we would essentially "hide" the genealogical structure factor (and any over-counting of mutations) in the mutation rates, producing "effective" mutation rates. We could either (a) take the father-son rates and correct with an average genealogical structure factor and a factor to correct for average over-counting, or (b) we could calculate our effective mutation rates directly from a group of haplotypes with a known time to the common ancestor. Again, we would need to calculate our rates from many clusters to get a good average. Note that it appears that the problems have gone away in this approach, but the effects are still there, only incorporated into the mutation rates. One will simply be accepting whatever genealogical structure and mutation over-counting exists in the calibration datasets as being the factors that will be used in applying the effective rates to unknown clusters. A further discussion of this latter approach is discussed further below.

Method 3 -- Is there any way to get around the problem of unknown genealogical structure? Actually, there is a way. At least there is a way to calculate the *most likely* genealogical structure or more accurately, the structure requiring the smallest number of mutations to produce the observed set of haplotypes. We can simply use the Network program to calculate the structure. This method finesses the over-counting issue at the same time. A slight additional complication is that the actual structure might have had slightly more mutations than the minimum, but this minimum will likely be closer to the right number than just counting mutations in the series of haplotypes. Another complication is that there are often a large number of different trees that have close to the minimum number of mutations, and these can have quite different structures. A bonus of using the Network program is that the resulting value of the rho factor is provided automatically. The rho factor can be directly converted to generations from the father-son mutation rates with no adjustments.

With so many advantages, why doesn't everyone just use Network? This approach is actually becoming more widely used as more people recognize its usefulness. However, if the method is applied over a time scale where the generation time may not be constant at present values, or if various forms of population dynamics have been in play, then the rho/network approach can give results that have substantial errors. A recent article by Murray P. Cox discusses the limitations of the method (Cox, 2008). Cox creates synthetic descendancy trees using a variety of demographic models and then uses Network to calculate the expected value and variance for the rho statistic. Cox shows that the 95% confidence intervals for rho are usually rather large. In Cox's lead example, he creates a tree with a 9300-year age, using a commonly accepted mutation rate. He then uses Network to calculate the 95% confidence interval on rho, which when converted to years yields an age between 3250 and 43,500 years. While this confidence interval definitely contains the actual value of 9300 years used in the simulation, and his sample size was small, the exercise doesn't inspire a lot of confidence in the method, at least as far as this example is concerned. Cox uses a number of demographic models and sample sizes in his simulations, and many of the models indicate that the result derived from the rho statistic are not within the predicted 95% confidence intervals. Part of the problem is that with repeated simulations, all of the different ways a tree could grow from a common ancestor become possible, each of which may produce a different effect on the result. When we are faced with just one actual tree to analyze, we generally do not know the history of the population, so we need to accept the very wide confidence intervals. These simulations were of mutations in mitochondrial DNA, but the same principles would seem to apply to Y chromosome dating studies. Cox did not even investigate the uncertainties in mutation rates, which would add to the uncertainties in the TMRCA. When the network/rho approach is applied to the more shallow time depth where extreme population dynamics would be unlikely to have occurred, the results would hopefully be somewhat better, but still the method should be used with caution.

Note that if we are comparing just two haplotypes, there are none of the complications of a branching genealogy, and none of the effects discussed above will apply, other than the issue of generation time. The lineage lengths and the transmissions are the same for a pair of descendants of the same common ancestor. However, if we do a pairwise comparison of every pair of haplotypes in a haplotype set and try to average them to get a better estimate of the TMRCA, then we again have the complications of the structure in the genealogy--the average of the TMRCA for each pair will not give us the TMRCA for the whole set--we will get an underestimate of the TMRCA by the genealogical structure factor. Dividing by this factor will increase the average pairwise value, an increase of 10-20%, to get the TMRCA for the group. One could also take a similar approach for the median of the pairwise set in cases where there are likely to be some outliers in the set. One could also use a set of known trees to examine the distribution of pairwise TMRCA values and find the point (percentile) on the distribution where it crosses the actual value. This percentile will probably be around 65 to 75 in general, and could be averaged over several known clusters to find a good percentile to apply to unknown clusters.

In regard to the Method #2b for correcting for genealogical structure, the complications can be skirted and we can calculate an effective mutation rate in mutations per haplotype per generation (or year) that subsume the different effects discussed above. Such approaches depend for their calibration on the determination of the effective mutation rates from a known time to an ancestor and corresponding number of generations (using some assumption for years per generation) and a set of haplotypes from a set of known descendants of that ancestor. Generally, one would not know the genealogy, and it wouldn't be needed anyway. One would reconstruct the ancestral haplotype from the set, and simply add up all the mutational differences from that ancestral haplotype, without regard to the possibility of multiple counting of mutations (that factor would be included in the effective rate also). If we then divide the total mutations by the number of haplotypes, we get the average number of mutations per haplotype. If we take my Athey example again, we would find 27/19 = 1.42 mutations per haplotype. The common ancestor for the Athey group was born on average approximately 305 years before the set of 19 participants, so 305 years would correspond to 1.42 mutations per haplotype. That is, we would have one mutation for every 215 years. We could convert this to a rate and obtain 1/215 = .00466 mutations per 37-marker haplotype per year per participant. Alternatively, if we didn't know that the Athey common ancestor was 8.8 generations back on average, we could estimate the number of generations using, for example, 33.3 years per generation. We would estimate 305/33.3 = 9.1 generations, which is not much different from the actual value of 8.8. Then we could calculate the effective mutation rate in terms of generations instead of time.

The mutation rate just obtained for the Athey cluster is only for one small cluster, so we shouldn't expect that it would necessarily be the best rate to use in general. Ideally, we would repeat the calculation on a large number of clusters and average the results. However, for purposes of illustration of how this would be applied to a cluster with unknown genealogy, let's apply it back to the known Athey cluster from which it was derived. In this group of 19 participants we observe 27 mutations, so the TMRCA would be 27/(.00466 x 19) = 305 years, which, of course, matches the known time that we started with.

This effective-rate procedure leads to a rather simple approach with all the complications comfortably out of sight. However, if we apply the approach to a new set of haplotypes, any deviation from the calibration-set averages in the new set for either the years per generation value or the genealogical structure factor, will affect the accuracy of the result. These factors have not gone away just because we can't see them. The advantage of this of this approach may also be its principal defect--when the complications are comfortably out of sight we may tend to put too much credence in the results.

This approach has the advantage that it would eliminate the systematic errors discussed above. All the competing effects would still be there, but they would be averaged out over the large number of datasets that should be used to calculate the effective rate, leaving only the random errors to contend with. One would also not need to worry about the number of years per generation either, but in effect, the average generation time present in the all the calibration datasets is silently chosen by default. If we then attempt to apply an effective mutation rate to a cluster that goes back beyond the present era of 33-35 years per generation (from which era most of the calibration datasets come), the TMRCA result may be in error.

The whole idea of effective mutation rates is frequently disparaged in the on-line discussion groups on genetic genealogy. One reads of "fudge factors" being used to modify mutation rates, as though there is any way around such adjustments. In clusters that are only a few centuries old, the correction factor for the genealogical structure, accounting for the fact that the genealogy will have some intermediate ancestors for some pairs of subjects, will typically be between 0.75 and 0.9. That is, any TMRCA might need to be adjusted upward by a factor of 1.1-1.3. For populations going back much further in time, there will be the effects of population dynamics also--the extinction of many lines and marked expansion of others, plus a possible reduction in the generation time. All of these effects will require correction factors that collectively may reach the factor of two used by Zhivotovsky (2004) in his often-cited article. Note: The factor of two is derived from the mutation rate Zhivotovsky determined, .00069 per marker per 25 years, or equivalently .00092 per 33.3 years (to put it on approximately the same basis as the father-son studies), and the father-son rate for the same markers of about .00184 per marker per generation (of about 33.3 years), and these two values, .00092 and .00184 differ by a factor of two.

An example of the effective-rate approach is included in this section in the article by Klyosov (2009a). Since his effective rates were determined from a large cluster ranging back about 800 years, they would presumably be most applicable to other datasets with a similar age. However, his article sometimes applies the method to older clusters (correcting only for additional back mutations and asymmetry of mutation distributions) (Klyosov, 2009b), and, indeed, he makes the point that he believes no additional correction is needed. Klyosov's approach is interesting in that he uses a very simple simple "linear" approach to TMRCA, but then he provides correction factors for back mutations and asymmetry where needed.

Howard (2009a) uses a correlation-based approach to compare pairs of haplotypes, which results in a pairwise measure that is proportional to time. He calibrates his time scale using several surname clusters with known genealogies, and his rates are also directly in terms of years. The generation time is subsumed in his calibration, but he must still use a genealogical structure factor to correct the average (or median) pairwise time for the cluster to arrive at overall TMRCA for the cluster.

The two articles by Gwozdz (2009a, 2009b) focus primarily on evaluating the integrity of clusters, but in addition he reviews the ASD and other methods for calculating the TMRCA. His cluster analysis tools will provide a way to assess when a cluster has a good chance of representing a clade. In his second article he applies his methods to several apparent clusters that have been found in datasets from Poland.

Again, this special section is intended to stimulate discussion and comment. We will welcome any formal responses in the form of further articles on the subject, or just a letter to the editor.

References

Cox MP (2008) Accuracy of molecular dating with the rho statistic: Deviations from coalescent expectations under a range of demographic models. *Hum Genet*, 80:335-357.

Dupuy, BM, et al (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum Mutat*, 23:117-124.

Gusmão, P, et al (2005) Mutation rates at Y chromosome specific microsatellites. *Hum Mutat*, 26:520-528.

Gwozdz P (2009a) Y-STR mountains in haplospace, Part 1: Methods. *J Genet Geneal*, 5:000-000.

Gwozdz P (2009b) Y-STR mountains in haplospace, Part 2: Common Polish Y-DNA clades. *J Genet Geneal*, 5:000-000.

Howard W (2009a) The use of correlation techniques for the analysis of pairs of Y-STR haplotypes, Part 1: Rationale, methodology and genealogy time scale. *J Genet Geneal*, 5:000-000.

Howard W (2009b) The use of correlation techniques for the analysis of pairs of Y-STR haplotypes, Part II: The application to surnames and other haplotypes. *J Genet Geneal*, 5:000-000.

Klyosov AA (2009a) DNA genealogy, mutation rates, and historical evidence written in Y-chromosome, Part I: Basic principles and the method. *J Genet Geneal*, 5:000-000.

Klyosov AA (2009b) DNA genealogy, mutation rates, and historical evidence written in Y-chromosome, Part II: Walking the map *J Genet Geneal*, 5:000-000.

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G,. Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am. J Hum Genet, 74:50–61.