Journal of Genetic Genealogy

Journal: www.jogg.info
Originally Published: Volume 12, Number 2 (Summer 2024)
Reference Number: 122.002

TMRCA FOR A MATCHING Y-STR CLUSTER BY FITTING A BINOMIAL DISTRIBUTION

Author(s): T. Whit Athey, PhD

Journal of Genetic Genealogy

TMRCA FOR A MATCHING Y-STR CLUSTER BY FITTING A BINOMIAL DISTRIBUTION

By T. Whit Athey

Abstract

There exist standard tools for calculating the time-to-the-most-recent-common-ancestor, or TMRCA, for the Y-STR values for a pair of individuals, for example, there is the TIP tool at Family Tree DNA (FTDNA). There are fairly large uncertainties in the resulting values. The existence of projects with numerous participants forming a Y-STR cluster provides the opportunity to calculate the TMRCA to much better accuracy, provided there are suitable analysis tools. One such approach is presented here for a group of Y-STR haplotypes. The number of cases of 0, 1, 2, 3, . . . mutations or differences from the ancestral haplotype of the TMRCA forms a histogram that theoretically should approximate a binomial distribution. This report shows how to apply this idea to a cluster of Y-STR results and obtain the TMRCA in generations.

Introduction

In a set of closely matching Y-STR haplotypes, there will be various numbers of differences of each individual's haplotype (the set of Y-STR values) from the ancestral haplotype of the TMRCA. The ancestral haplotype, without significantly affecting generality, can be assumed to be the modal values of the cluster being considered. In this development it will be assumed that each haplotype has the same number of markers and that the Y-STR markers of each haplotype are one of the standard panels offered by FTDNA. That is, each haplotype should contain one of the standard 12-, 25-, 37-, 67-, or 111-marker sets, and each haplotype should have the same number of markers. For the cluster of Y-STR markers, if the number of cases of 0, 1, 2, 3, . . . mutations from the ancestral haplotype is calculated and graphed as a histogram, the distribution should theoretically approximate a binomial distribution. This report will show how the best (binomial) fit to the observed distribution can estimate the TMRCA.

The present method would be most suitable for cases where the TMRCA lived within the last 1000 years, so that back mutations would be minimized. For deeper ancestry, such as determining the TMRCA for haplogroups, methods such as those based upon Nordvedt's Interclade Estimation method would be more appropriate.[1,2]

[1] Ken Nordvedt has developed a number of utilities and descriptive files that are linked at the ISOGG web site. See also the Reference Section and Footnote 2.

[2] Vance JD (2020) SAPP Toolset (based upon Nordvedt's Interclade Estimation method). http://www.jdvsite.com/

Naturally, clusters with larger numbers of Y-STR haplotypes will yield a TMRCA with a smaller uncertainty. In practice, at least ten haplotypes should be used, and ideally more than 20. There will be a trade-off between selecting a larger number of markers for more "marker transmissions" and a lesser number of markers, which will usually mean more haplotypes to consider in the cluster. Often in practice choosing the 37-marker panel may be optimal, but the process may be repeated for 25 markers and 67 or 111 markers where the data are available. This issue will be discussed again later in this report.

Methods and Data

For each of the panels of Y-STR markers from FTDNA, 1-12, 1-25, 1-37, 1-67, and 1-111 markers. various approaches have been used to determine the average mutation rate For the purposes of this report the average mutation rate for each of the five panels shown in Table 1 will be used[3,4].

Table 1 Average Mutation Rates: Five Y-STR Panels

Panel of Markers	Average Mutation Rate (mutations per marker per generation)
1-12	.0025
1-25	.0028
1-37	.0042
1-67	.0031
1-111	.00258

If more accurate rates are available, they may be substituted in the program, which is easily implemented in an Excel spreadsheet (see Reference Section for an example of an implementation.

The Binomial Distribution

When events are expected to occur randomly, independently, and at a constant rate, the probability of the event occurring on the xth trial out of n trials, follows a binomial distribution. That is, the probability of the number of events, n, occurring when the rate is p is given by the binomial distribution:[5]

$$B(x, n, p) = \frac{n! p^x (1-p)^{(n-x)}}{x!(n-x)!} \qquad (1)$$

Where x = 0, 1, 2, 3, . . . , (in our case, this will be the genetic distance from the ancestral haplotype)

n = number of trials (in our case, this will be the number of marker transmissions after k generations, which will be k times the number of markers). n! = n x (n-1) x (n-2) x . . .x 1 (n factorial)

p = probability of an occurrence (in our case, the average probability of a mutation in a marker)

A table of the distribution of the probability for the 111-marker case (p = .00258) can be generated in Excel using the BINOM.DIST function. The function's syntax is:

[3] An example of the determination of average mutation rates for the first three panels, which are slightly different from those used here, may be found in Chandler J (2006) Estimating per locus mutation rates. *Journal of Genetic Genealogy*, 2:27-33.

[4] A discussion of mutation rates from different sources may be found in Athey TW (2007) (Editorial) Mutation rates—who has the right values? *Journal of Genetic Genealogy*, 3(2):i-iii. The values used here represent an average from different sources.

[5] https://en.wikipedia.org/wiki/Binomial_distribution

BINOM.DIST(x,n,p,FALSE)

The "FALSE" value is necessary to distinguish the present case from a cumulative distribution. In some versions of Excel, the function may be written as "BINOMDIST" without the "dot." Table 2 was calculated using the Excel function:

BINOM.DIST(x,n,0.00258,FALSE)

n will be the product of 111 and the generation number G (that is, after each generation, 111 markers will have been transmitted to the next generation in the line by each participant).

In the following example we will consider 26 haplotypes, each with 111 markers. Each column of the BINOM.DIST distribution will represent the theoretical distribution of the fraction of the 26 haplotypes that are a genetic distance of x from the ancestral haplotype. The third column of Table 2 has the corresponding observed distribution, and we will be seeking the BINOM.DIST column that best fits the observed distribution. The interpretation of the column under Generation 1 is that after one generation, we would expect .759 of the haplotypes to have the unmutated ancestral values, .209 of the haplotypes to have one mutation from the ancestral values, .028 of the haplotypes to have two mutations from the ancestral values, etc.

Table 2 Observed Distribution and Candidate Bionomial Distributions (for 111 markers)

Distribution for Example			Binomial Distribution							
Genetic Distance from the Ancestral Values (x = 0, 1, 2, . . .)	**Observed number of individuals with x genetic distance from the ancestral values**	**Actual Distri-bution Divide By No. of haplo-types (26)**	**Gen 1 n=111**	**Gen 2 n=222**	**Gen 3 n=333**	**Gen 4 n=444**	**Gen 5 n=555**	**Gen 6 n=666**	**Gen 7 n=777**	**Gen 8 n=888**
0	**5**	**0.1923**	0.75910	0.57623	0.43742	0.33204	0.25205	0.19133	0.14524	0.11025
1	**10**	**0.3846**	0.20948	0.31804	0.36214	0.36653	0.34779	0.31681	0.28057	0.24341
2	**9**	**0.3461**	0.02864	0.08737	0.14945	0.20184	0.23951	0.26189	0.27065	0.26838
3	**2**	**0.0769**	0.00259	0.01593	0.04100	0.07393	0.10976	0.14411	0.17383	0.19706
4	**0**	**0.0000**	0.00017	0.00217	0.00841	0.02027	0.03766	0.05939	0.08362	0.10840
5	**0**	**0.0000**	0.00001	0.00024	0.00138	0.00443	0.01032	0.01955	0.03214	0.04765
6	**0**	**0.0000**	0.00000	0.00002	0.00019	0.00081	0.00235	0.00535	0.01028	0.01743
7	**0**	**0.0000**	0.00000	0.00000	0.00002	0.00013	0.00046	0.00126	0.00282	0.00546
8	**0**	**0.0000**	0.00000	0.00000	0.00000	0.00002	0.00008	0.00026	0.00067	0.00150
9	**0**	**0.0000**	0.00000	0.00000	0.00000	0.00000	0.00001	0.00005	0.00014	0.00036

10	**0**	**0.0000**	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00003	0.00008

In the second and third columns of Table 2 are shown the actual distribution of the genetic distances from the ancestral haplotype of an example cluster of 26 haplotypes with 111 markers. Note that the columns labeled Gen 5 and Gen 6 appear closest to the observed values, as shown in Table 3. However, we can calculate a better value for the TMRCA if we use all of the information available.

Table 3 The Three Columns That Are Best Fits to the Observed Distribution

Genetic Distance x from the Ancestral Values	Number of Individuals with this Distance from Ancestral Haplotype (out of 26)	The resulting observed fraction of the cluster at this distance = #mut/26	Column Bracketing Gen 5 on the left (Column from Table 2 for Gen 4)	Column for Gen 5 (possibly the best fit, though Gen 6 is also a possibility)	Column Bracketing Gen 5 to the right (Column from Table 2 for Gen 6)
0	5	0.1923	0.3320	0.2520	0.1913
1	10	0.3846	0.3665	0.3478	0.3168
2	9	0.3461	0.2018	0.2395	0.2618
3	2	0.0769	0.0739	0.1098	0.1441

We see that our observed distribution lies close to the theoretical (binomial) distributions for Gen 5 or 6, or that our cluster TMRCA is approximately 5 or 6 generations back from the present test takers. Of course, we have assumed that everyone in the cluster is the same number of generations removed from the common ancestor. When that is not the case, the final TMRCA will be an average of the number of generations back to the common ancestor.

In terms of the present example, we can calculate the best fit to the theoretical distribution by using the method of least squares. That is, we can automate the above "bracketing" by calculating the sum of squares for the differences between observed and binomial distributions for the most likely generation, along with the two bracketing generations—those columns on either side of the one with lowest sum of squared differences. This can be followed by determining a second-degree polynomial fit to the sum-of-squares values and finding the generation value G that represents the minimum of the fitted polynomial.

Table 4 Sum-of-Squares of Differences Between Observed and Binomial Distributions

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8
Sum of Squares of Differences	0.4582	0.2225	0.1006	0.0411	0.0189	0.0201	0.0362	0.0615

Our observed distribution is very unlikely to be near the generation 1 binomial distribution, so naturally the sum-of-squares for Gen 1 is the largest (of those showing) in Table 4. The sum-of-squares values

decrease for a few generations as we approach more likely possibilities, and finally start rising again as we move beyond the most likely generation. As an example of the Table 4 calculation, consider the sum-of-squares for the column for Gen 5:

$$\text{Sum} = (.1923-.2520)^2 + (.3846-.3478)^2 + (.3461-.2395)^2 + (.0769-.1098)^2 + (.0769-.0739)^2 + (0-.0377)^2 + (0-.0103)^2 + \ldots$$

$$= .00356 + .00135 + .01136 + .00108 + .00001 + .00142 + .00011 + \ldots = 0.0189$$

We can see that the best fit—the column that minimizes the sum-of-squares--is near Generation 5 or 6. If we let s_1, s_2, and s_3 represent the sum-of-squares values from the columns for Generations 4, 5, and 6, and let d be the generation number of the middle of the three columns with the lowest sum-of-squares value (column 5), then the generation number that minimizes the second-degree polynomial fit to those values is given by:[6]

$$G = d-1+(1.5s_1-2s_2+0.5s_3)/(s_1-2s_2+s_3) \qquad (2)$$

Substituting the values from Table 4:

$$G = 5 - 1 + [1.5(0.0411)-2(0.0189)+0.5(0.0201)]/[0.0411-2(0.0189)+0.0201]$$
$$= 4 + .0338/.0233$$
$$= 5.45$$

Therefore, the best fit to the given data is TMRCA = 5.45 generations, which represents an average value for the group of haplotypes. For example, this average for the 26 haplotypes might result from 14 of the 26 being 5 generations from the common ancestor and 12 being 6 generations.

The above procedure may be set up in an Excel spreadsheet with inputs: number of markers, number of haplotypes, the observed distribution of distances from the ancestral haplotype, and an array with the mutation rates for the five possible panels of Y-STR values. A table like Table 2 can be set up, followed by a table of squared differences for Gen 1, Gen 2, etc, with a summation at the bottom of each column similar to Table 4 above. Finally, the Generation number with the lowest sum-of-squares value can be used, along with values from the two adjacent columns, and the generation number may be calculated from Equation 2.

When selecting the best-fit column, there may be cases like the one above where there is only a small difference in sum-of-squares for two columns, it is not really critical as to which one is chosen as the best one. If we had chosen Gen 6 as the "central column" in the above calculation, we would get G = 5.41, which is very close to our previous value.

Therefore, the best fit to the given data is TMRCA = 5.4 generations, which is an average value for the group of participants.

A Special Case

Consider Equation 1 for the case of x = 0. That is, consider only the fraction of participants who are unchanged from the ancestral value after n marker transmissions. In our example the observed fraction value would be 5/26 = .1923. When we substitute x = 0 into Equation 1, we get:

[6] The process for determining Equation 2 is as follows. We fit a quadratic (second-degree) polynomial of the form $G = ax^2 + bx + c$ to the s_1, s_2, and s_3 values for x = 4, x = 5, and x = 6, then the minimum value for G is found by setting dG/dx = 0, and solving the resulting equation for x. This results in the minimum value given by $x_{min} = -b/2a$.

$B(0,n,p) = n!p^0(1-p)^{(n-0)}/[0!(n-0)!]$

But, $n!/(n-0)! = 1$, $p^0 = 1$, and $0! = 1$ (by definition), so we get a much simpler form:

$B(0,n,p) = (1-p)^n$

If let g_0 represent the observed value corresponding to the theoretical B(0,n,p), let G be the number of generations to the TMRCA, and let j be the number of markers, then substitute into this equation we get

$g_0 = (1-p)^n = (1 – p)^{jG}$

If we take the logarithm of both sides of this equation, we get

$\log(g_0) = jG \log(1 – p)$

Or $G = \log(g_0)/[j \log(1 - p)]$ (3)

For example, if we had used this simpler approach for the example above of the 26 haplotypes with j = 111 markers, 5 of which still had exactly the ancestral Y-STR values, we would have:

$G = \log(5/26)/[111 \log(1 - .00258)]$

$= - 0.7160/(-0.1245)$

$= 5.75$

This compares to the value of 5.45 that we found when using equation 2 on the entire observed distribution. Note, however, that there is more uncertainty in the result when using this simpler formula. If there were just one more mutation (leaving 4 of the 26 unmutated), or just one less mutation, our Equation 3 would have given a result of 6.5 or 5.1.

Discussion

In planning an analysis of a Y-STR cluster, it is likely that the best approach in choosing the number of markers and number of haplotypes to be analyzed will be a choice that maximizes the number of marker transmissions per generation in the whole cluster. There will be a trade-off since the larger the choice of markers in the panel to be considered, the smaller will be the number of participants in the cluster who have that many markers. So, one would usually seek to maximize the product jk where j is the number of markers and k is the number of participants with at least that number of markers.

For example, in our case study above, we used 111 markers, which gave us 26 participants with that number of markers, so the product gives us 111 x 26 = 2886 marker transmissions per generation for the group. If we considered 67 markers and assumed (for example) that we would have 32 participants available with results on that number of markers, then the product would be 2144. If we assumed the participants with at least 37 markers (for example) to be 73, then the product would be 2701, which is almost as large as in the 111-marker case. The best choice will, of course, depend on the particular cluster under consideration, but one should not always assume that choosing 111-marker haplotypes will be the best one, though it was in this example.

Another approach to choosing the best number of markers for a cluster is simply to use the one that results in the broadest distribution of mutations.

References

Nordvedt K (2008) See links to his utilities and informational files under the Y-DNA tools at the ISOGG website:
https://isogg.org/wiki/Y-DNA_tools

Chandler J (2006) Estimating per locus mutation rates. *Journal of Genetic Genealogy*, **2**:27-33.

Athey TW (2007) (Editorial) Mutation rates—who's got the right values? *Journal of Genetic Genealogy,* **3**(2):i-iii.

Athey TW. An Excel spreadsheet that shows an implementation of the program described above is available for download at:
http://www.hprg.com/storage/binomial3.xlsx

Conflicts of Interest

The author declares no conflicts of interest and no commercial interests in the subjects covered or companies mentioned in this report. Family Tree DNA (FTDNA) is mentioned above only because currently, it is the primary company still offering testing services for Y-STR markers, it is the main source of surname projects that potentially have sufficient Y-STR clusters, and whose panel definitions have become widely used.