A Comparative Analysis of the RCC Correlation and FamilyTreeDNA TiP^{TM} Probability Approaches for Estimating the Time to the Most Recent Common Ancestor of a Pair of Y-DNA Haplotypes

-- William E. Howard III --

Abstract:

We have investigated the relationship between the Y-DNA results of pairs of testees reported by the FamilyTreeDNA (FTDNA) TiP^{TM} predictions at 4, 8 and 12 generations and the RCC time scale prediction derived from the same groups of testees. We derive a relationship that, to a first approximation, provides a link between a TiP^{TM} probability and the RCC time scale[1]. We suggest that the RCC approach to dating haplotypes may be superior to the use of the TiP^{TM} probability in approximately 37-38 percent of cases where the absolute sum of haplotype marker differences (m) exceeds the number of markers (n) that have changed (viz., when m > n).

Introduction and Background:

FamilyTreeDNA uses a TiP™ prediction that incorporates STR marker-specific mutation rates in conjunction with STR marker values to estimate the number of generations to the most recent common ancestor (MRCA) of any given testee pair[2]. Since the correlation approach also estimates the time to the MRCA (TMRCA), we investigated the two approaches in order to compare the TiP™ result with the RCC result. The FTDNATiP™ process has a default mode that computes the probability that a pair of testees will share a MRCA at a certain number of generations. If a trusted pedigree indicates that a one of the pair of testees could not have lived within that number of generations, the TiP^{TM} prediction can be refined, using prior knowledge of the pedigree in a Bayesian type approach to recompute the probability. The details of the TiP^{TM} prediction are proprietary. In this analysis we only examine and compare the TiP^{TM} prediction probabilities prior to any recalculation.

Comparison of the FamilyTreeDNA TiP™ Predictor with the RCC Predictor

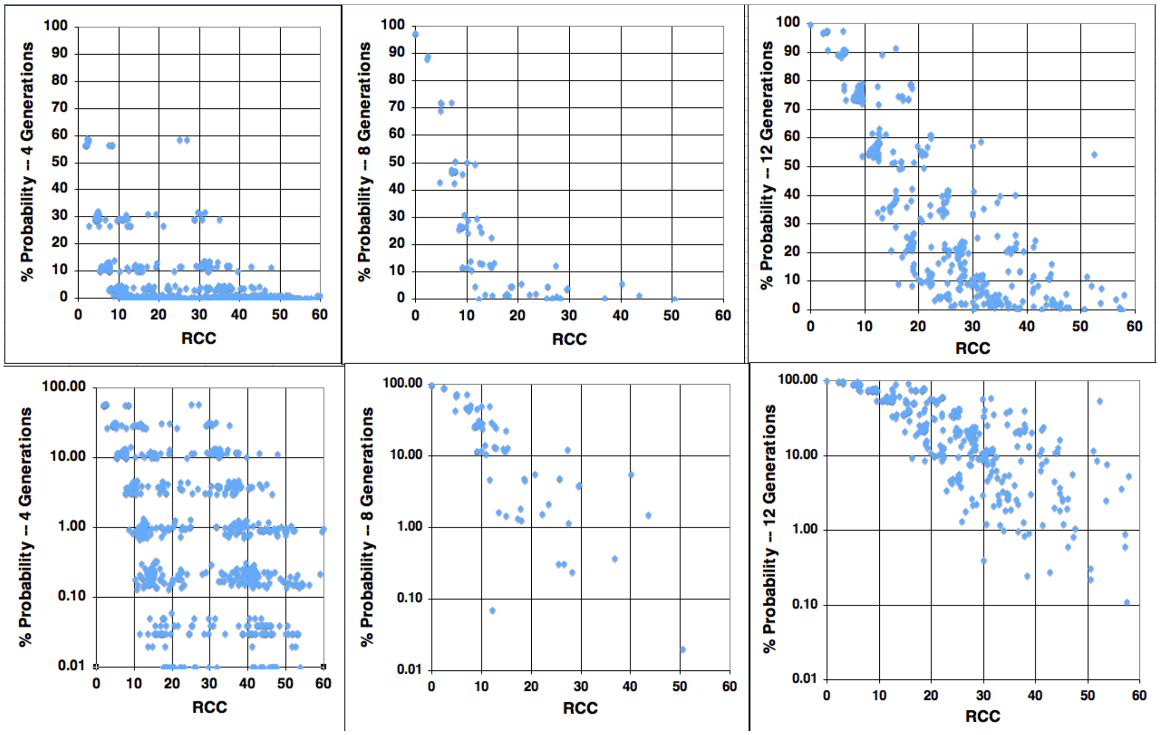
We investigated the relationship between the FamilyTreeDNA (FTDNA) TiP™ predictions for three different generations and the corresponding observed RCC predictions using representative samples of testee pairs within one haplogroup and the two surname projects in Table 1. Three different generations across three groups of haplotypes were chosen to illustrate that the same effects are seen in each sample.

Table 1: The Number of Pairs of Testees in Three Projects for which TMRCA Probabilities Over 4, 8, and 12 Generations and their Respective RCC Values Have Been Analyzed.

Project	Number of Testee Pairs	Number of Generations
Haplogroup J2b2e[3]	783	4
McKee	73	8
Pettigrews	378	12

The apparent success when a sample of seven McKee testees were paired with a sample of 27 other McKee haplotypes to produce 73 testee pairs led us to widen the investigation. Figure 1 shows the results of these three investigations. The top three charts in Figure 1 show the 4, 8, and 12-generation TiP™ probabilities along the ordinate and the RCC values appropriate to each pair along the abscissa. The bottom three charts show the same data on a Log$_{10}$ plot.

Figure 1: The TiP™-derived Probability that a Pair of Testees Will Have a MRCA Within 4, 8, and 12 Generations vs. the Observed RCC Value of Each Pair. The lower three charts show the same data in semilog format.



We can see by inspection of Figure 1 that there is a relatively tight relationship between the FTDNA's TiP™ 8 generation probability calculations and RCC down to a probability of ten percent and to an RCC of 15-20. This RCC interval extends from the present to dates near 1000-1200 AD when surnames began to come into use. At lower values of probability (and larger values of RCC) the relationship is not as well-defined and there are very large uncertainties beyond that point. Very roughly, the conversion between probability at 8 generations, P(8), and RCC is:

P(8) = 100 – (6.5 (+/- 5% est.) RCC)	for RCC< ~10

We then investigated the relationship between the TiP^{TM} probabilities at 12 generations vs. RCC. We purposely chose a larger, but different set of testees that, when paired, would yield a larger group of points since we suspected that the 12 generation probabilities would lead to a more complicated relationship than we found at 8 generations. We selected 28 haplotypes which, when paired, produced 378 testee pairs shown in Figure 1.

As more testee pairs are added to the analysis, it became evident that clusters of pairs tend to become grouped in both probability and RCC space. Moreover, these clusters are also aligned in two or more additional sequences that run parallel to the sequence at the far left of each chart. The reasons for these groups and sequences will be discussed in the next section. They appear in the P(4) and P(12) charts and would have appeared in the P(8) chart had more testee pairs been included.

We can see by inspection of Figure 1 that there is a relatively more complicated relationship between the FTDNA's TiP^{TM} 12 generation probability scale and RCC. There appear to be at least four sequences running from the upper left down the chart toward the lower right. At lower values of probability (and higher values of RCC) the 12 generation probabilities are not as well-delineated and there are very large uncertainties at probabilities below about 10 percent. Very roughly, the conversion between probability at 12 generations, P(12), and RCC for the pairs of testees is:

P(12) = 100 – (4.2 RCC)	for P(12)>10%

This relationship is derived from the sequence of points at the far left of the chart.

The points in the P(4) chart in Figure 1 show a very layered structure along the probability axis and a pronounced clustering along the RCC axis. Both the horizontal and vertical sequences and the clusters are most clearly shown in the Log$_{10}$ plot. There are at least four vertical sequences and seven horizontal sequences. Very roughly, the conversion between probability at 4 generations, P(4), and RCC is:

P(4) = 100 – (12.7(+/- 3% est.) RCC)	for P(4)>10%

This relationship is again derived from the sequence of points at the far left of the chart.

What Causes the Clustering and Sequences of Points in Figure 1?

To answer this question, we selected and identified 21 representative haplotype pairs in the P(12) group of Pettigrews. Each pair was near the centroid of all other pairs in each cluster that was located in a narrow probability group and in a narrow interval of RCC.

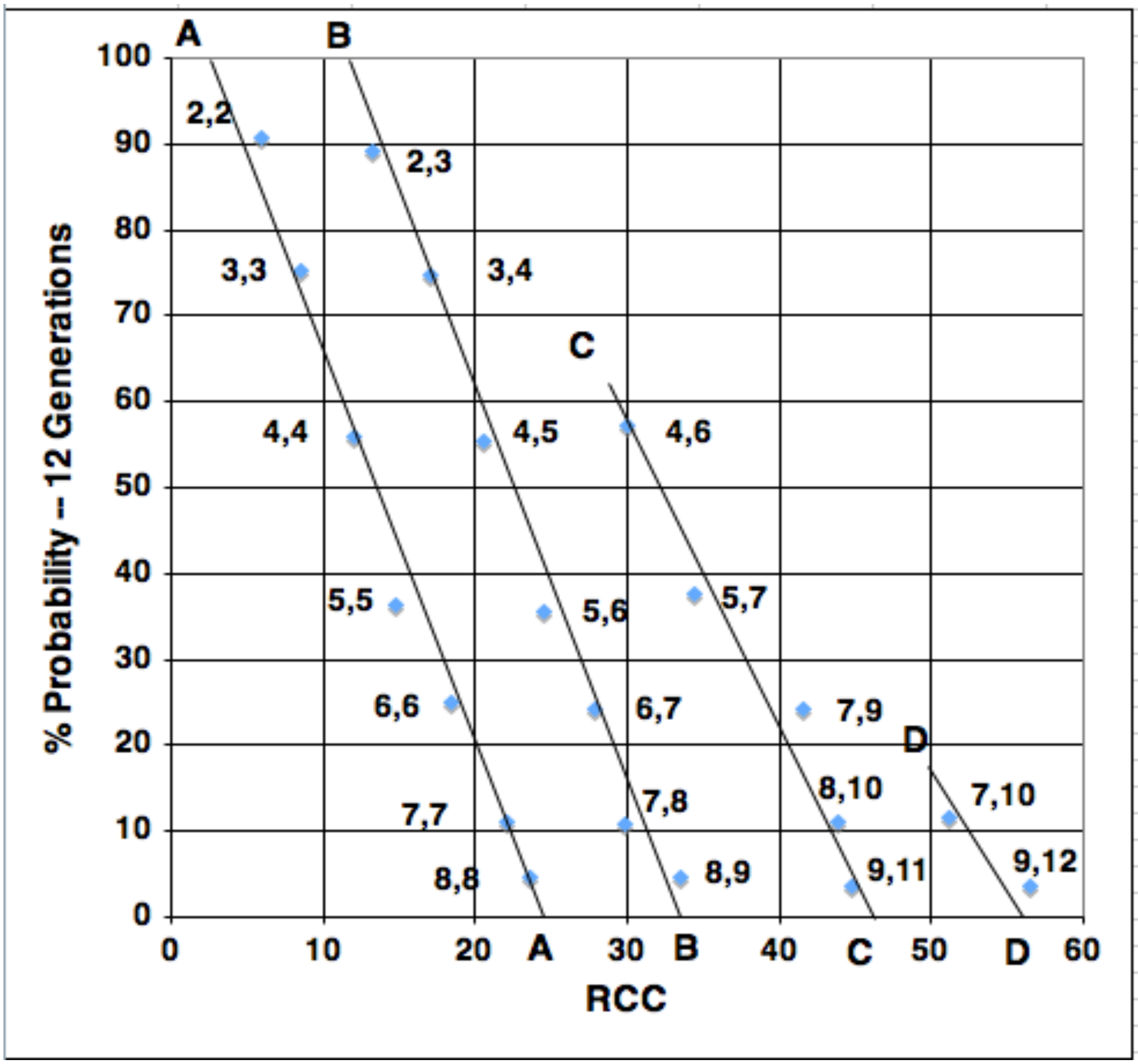
For each haplotype pair, we counted the number of DYS markers that had changed along each haplotype pair. We also counted the absolute marker value difference for each of the 37 sites, and summed those differences across the DYS sites. This process resulted in two numbers n, and m, where n was the number of specific DYS sites that had different marker values and m was the absolute sum of the differences in the marker values for sites that had changed.

Next, we plotted the representative points on the P(12) vs. RCC chart and identified each point by its (n,m) designation. The results are shown in Figure 2.

The first, leftmost sequence (n = m), designated as Line A in Figure 2, is the one that FTDNA's TiP^{TM} uses to determine the probability that a testee pair will have a MRCA – in this case within 12 generations. Thus the approximate P vs. RCC equations in the last section refer to the Line A sequence for each set of samples. The testing agency attributes small differences in the probability designation to differences in individual marker mutation rates. Throughout most of Line A, there is a relatively good, one-to-one relationship between P(12) and RCC.

The second sequence, Line B in Figure 2, is one where m = n+1. Since the absolute sum of the marker differences is one more than the number of different DYS sites, there will be one difference of 2 along the haplotype sequence. The FTDNA TiP^{TM} uses the Line A sequence, where n=m, to compute the probability. The RCC correlation approach uses m to compute RCC. Therefore, we expect that there will be sets of lines (viz., B, C, D, ….) when we compare the value of RCC to FTDNA's TiP^{TM} probability when a pair of haplotypes contains other than one marker difference in a marker site along the haplotype marker string. Thus Lines C and D represent pairs of haplotypes in which m = n+2 and m = n+3, respectively.

Figure 2: Schematic of P(12) in Figure 1 Showing (n,m), the Number of DYS Marker Sites that are Different in a Pair of Haplotypes (n) and the Sum of the Absolute Values of the Numerical Marker Differences Along the Haplotype Pair (m) vs. RCC (observed).



Although the TiP™ probabilities and the RCC time scale for pairs along Line A are in good agreement, the two approaches differ when m>n because the TiP™ uses the number of DYS sites that are different (viz., n ~ genetic distance (GD)) while RCC uses the numerical differences among the DYS marker numbers. For example, if there are four marker sites (n=4) that each differ by one between a pair of haplotypes, the TiP™ will predict a probability of about 55% and the RCC will indicate a date corresponding to RCC 12 (m=4). Both methods predict a reasonable chance of finding their MRCA. But if there are four marker sites (n=4) and one of the sites differs by two from the others (m=5), the TiP™ will still predict a probability of 55% but the RCC of 20 will indicate a date that is probably too long ago to permit a genealogist to find their MRCA. A study of the Pettigrew haplotypes indicates that (m=n) about 63% of the time, (m=n+1) 28% of the time, (m=n+2) 8% of the time, and (m=n+3) 1% of the time. We conclude that (1) the RCC determination will agree with the TiP^{TM} probability only about 2/3 of the time and (2) the TiP^{TM} probability may be not be valid for marker strings in which m>n.

FTDNA's approach in effect collapses Figure 2 along lines of equal probability so that all points merge with Line A. Unlike the RCC time scale, the TiP^{TM} does not recognize lines other than Line A where n=m.

Note that n, m, and marker values are all integers. When a marker mutation takes place, the position of the paired haplotypes in Figure 1 will jump from its original point on the diagram to another. Most of the time the jump will take place to an adjacent point because there has been only one mutation. This quantization explains the clumping in Figures 1 and 2; quantization also explains the horizontal vacant ridges in Figure 1 that are particularly noticeable in the P(4) log plot.

When a specific marker site shows a difference greater than two when comparing two haplotypes, that marker site has undergone two or more mutations during the time to their MRCA. In order to investigate which approach (i.e., TiP^{TM} or RCC) gives the most reliable result, we need to investigate whether the numbers of multiple mutations are consistent with what we expect from random mutations.

Are the Statistics of Multiple Mutations Consistent with Random Mutations?

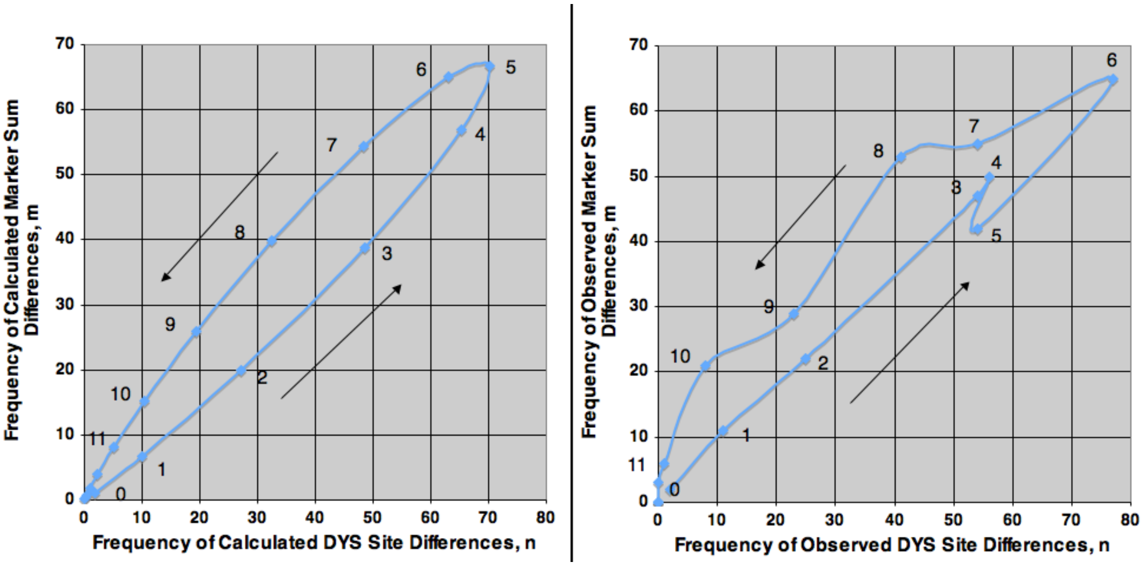
We first approach the answer by comparing our results with those predicted by a Poisson distribution[4]. We selected 406 pairs of Pettigrew haplotypes, derived the value of n and m for each pair and counted them. We then calculated the number of times that n and m had the values of 0, 1, 2,……16. The results are shown in Table 2.

Table 2: The Observed and Calculated Frequencies of Occurrence of n and m in a Representative Sample of 406 Pairs of Pettigrew Haplotypes. (Calculated values are predicted from Poisson statistics)

The Value of n or m	**Observed Frequency of Occurrence of DYS Site Differences (n, GD)**	**Calculated Frequency of Occurrence of DYS Site Differences (n, GD)**	**Observed Frequency of Occurrence of Sum of Marker Values (m)**	**Calculated Frequency of Occurrence of Sum of Marker Values (m)**
0	2	1.9	2	1.2
1	11	10.1	11	6.8
2	25	27.1	22	19.9
3	54	48.6	47	38.9
4	56	65.4	50	56.9
5	54	70.3	42	66.7
6	77	63.0	65	65.1
7	54	48.4	55	54.5
8	41	32.5	53	39.9
9	23	19.4	29	26.0
10	8	10.4	21	15.2
11	1	5.1	6	8.1
12	0	2.3	3	4.0
13	0	0.9	0	1.8
14	0	0.4	0	0.7
15	0	0.1	0	0.3
16	0	0	0	0.1

The results of Table 2 can be presented in graphical form (Figure 3) that shows a unique feature, the n-m Loop that is also characteristic of pairs of n and m values in other haplotypes.

Figure 3: The n-m Loop. A Comparison of the Number of Times that Calculated and Observed Marker Site Differences (n, GD) and DYS Marker Sum Difference (m, Sum of Marker Differences) Occurred Among 409 Haplotype Pairs in a Pettigrew Sample.



In the left hand chart of Figure 3, the calculated number of times that there were 8 differences between n and m (32.5 and 39.9, respectively) is shown in the sample of 409 pairs in the sample. In the right hand chart, the observed number of times that 8 differences occurred between n and m were 51 and 53, respectively. Note that the calculated value of n in Table 2 increases more rapidly than the calculated value of m in Table 2 (the upside of the figure) until the average of the distribution is reached (at 5-6); then m is greater than n (on the downside of the figure). The arrows show the direction of this progression. This situation also occurs in similar figures for other haplotype pairs and it is caused by faster mutations having a greater effect when n and m are low. When n and m become large, the faster mutations tend to average out and the slower mutations begin to have an effect. The right hand graph in Figure 3 shows the observed data. The n-m Loop is still apparent, but mutations that are slightly different than the Poisson expectation cause the shape of the n-m Loop to be less smooth.

The results in Table 2 are dependent on the particular pairs of haplogroups we selected. Nevertheless, they do conform to a Poisson distribution. To reduce any dependence on the selected haplotypes, we took each paired value of n and m and derived a table of the frequency of occurrence of (m-n). The statistics of that table are presented in Table 3.

Table 3: The Frequency of Occurrence of (m-n) for the 406 Haplotype Pairs in Table 2

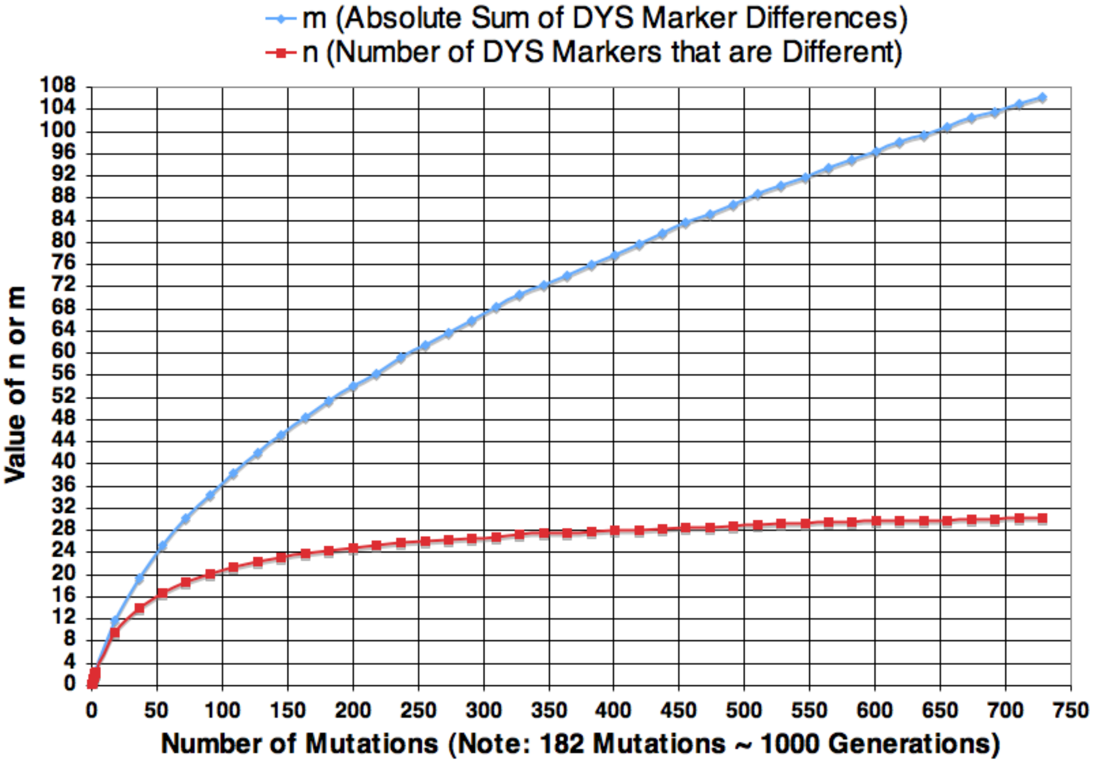
Value of (m-n)	Observed Frequency of Occurrence	Calculated Frequency of Occurrence
0	254	251.2
1	115	120.6
2	32	29.0
3	4	4.6
4	1	0.6
5	0	0.1

Table 3 demonstrates that the distribution of values of (m-n) conforms to Poisson statistics. We therefore conclude that:

- All mutations we observe are what we expect from random mutations.
- Marker strings in which m>n also appear to be random, with predictable frequencies of occurrence.
- While marker differences of two or more between DYS sites may occur quickly due to the random variations in mutation times, there is no evidence that they happen simultaneously or by other than random processes[5].
- The longer the time differences that separate two haplotypes from their progenitor, the higher the probability that n, m, and (m-n) will be larger.
- The different consequences of slow and rapidly mutating DYS sites are exhibited in diagrams that show the n-m Loop.

Since these distributions of n, m, and (m-n) are random, we conclude that the determination of the TMRCA of pairs of haplotypes using the RCC correlation approach appears to be preferred to the TiP^{TM} derived probability that only uses n (the genetic distance)[6]. Indeed, model codes produced by Sidney Sachs and Fred Schwab in support of the analysis in Howard (2013), show that n saturates while m continues to increase as the number of mutations increases. This effect is shown in Figure 4 where the TMRCA is more sensitive to changes in m in time intervals of genetic interest, but our analysis also shows that sensitivity begins over time intervals of genealogic interest. We conclude that the sum of the absolute values of the differences in DYS markers is the more sensitive indicator of the TMRCA than the number of DYS markers that are different.

Figure 4: The Saturation of n and the Growth of m as a Function of the Number of Mutations.



In cases where RCC indicates a TMRCA that lived long ago, yet trusted pedigree information shows a more recent TMRCA, the RCC value should be questioned because the number of mutations is not the average expected. If, in the same case, the TiP^{TM} result indicates a more recent TMRCA than RCC, the TiP^{TM} result should be questioned, particularly when a pedigree connection cannot be found. If enough of these cases can be studied, we will understand better the relationships between a testee's position on the dated Y-DNA STR phylogenetic tree, the pedigree, the RCC designation and the TiP^{TM} result.

The RCC time scale, calibrated by pedigrees, indicates the TMRCA of paired testees in years rather than by generations. By the positions of all testees on a dated STR phylogenetic tree, all testees can see their relationships with everyone else on the tree, giving a broader insight than individually paired probabilities. Nevertheless, we urge that the FTDNA TiP™ probabilities should be used in conjunction with the RCC correlation approach, particularly when n=m.

Summary:

The distribution of markers on paired haplotypes can be explained by Poisson statistics, indicating that mutations do take place randomly. We suggest that the RCC approach to dating haplotypes is superior to the use of the TiP^{TM} probability in about 37-38 percent of cases where the absolute sum of the marker differences exceeds the number of markers that have changed (viz., when m>n).

Acknowledgements:

We thank the three surname administrators, E. J. Hurley, Linda McKee and Sidney Sachs, for providing us with the TiP^{TM} data on the McKee and Pettigrew/Petticrew surnames and on the J haplotypes so that they could be compared with the RCC results. Discussions with Sidney A. Sachs have been particularly valuable.

REFERENCES:

Chandler, John F., *Estimating per-Locus Mutation Rates*, Journal of Genetic Genealogy 2, 27-33, 2006

Howard, William E. III, *The Use of Correlation Techniques for the Analysis of Pairs of Y-Chromosome DNA Haplotypes, Part I: Rationale, Methodology and Genealogy Time Scale*, Journal of Genetic Genealogy, 5, No. 2, Fall 2009, p. 256.

Howard, William E. III and McLaughlin, John D., *A Dated Phylogenetic Tree of M222 SNP Haplotypes: Exploring the DNA of Irish and Scottish Surnames and Possible Ties to Niall and the Uí Néill Kindred, Familia*, Ulster Genealogical Review No. 27, pp. 14-50, 2011. Ulster Genealogical & Historical Guild.

Howard, William E. III, *The Time of Origin and the Rate of Formation of Haplogroup I and its Subclades I1 and I2* (submitted to the Journal of Genetic Genealogy on 27 July 2012).

Howard, William E. III*, Uniting the Time Scales of Genealogy and Genetics: Estimates of the Errors, Uncertainties and Probabilities in Short- and Long-Term RCC Correlations* (in preparation, 2013).

AUTHOR POINT OF CONTACT:

William E. Howard III: McLean, Virginia wehoward@post.harvard.edu

END NOTES:

[1] The RCC correlation approach to estimating the time to the most recent common ancestor of a pair of haplotypes and some applications of the technique can be found in the various Howard papers listed in the references.

[2] Information on Family Tree DNA and its FTDNATiP™ process can be found at their web site: http://www.familytreedna.com/faq-tip.aspx. The FTDNATiP™ results are

based on the mutation rate study presented during the 1st International Conference on Genetic Genealogy, on Oct. 30, 2004. The probabilities take into consideration the specific 37-marker mutation rates for each individual marker being compared. Since each marker may have a different mutation rate, identical genetic distances will not necessarily yield the same probabilities. This paper uses the average mutation rate over all 37 markers to compute the value of RCC, and we use the individual marker rates given by Chandler (2006) when we explore the linearity of the observed RCC scale and its standard deviation as a function of time.

[3] The Family Tree DNA project designation is "J2b_455-8" for the STR of haplogroup J2b2e. Members all have a marker value of 8 at DYS 455. Credit is due to Whit Athey for finding this subclade haplogroup in JOGG. He stated that the mutation from 11 to 8 on DYS 455 happened less than 1000 years ago.

[4] The Poisson distribution can be used to predict the probability of a given number of events occurring in a fixed interval of time if these events are rare and occur at random with a known average rate. The Poisson distribution should be used when you have a small number of events that happen within a large population of possible events such as the expected annual number of deaths by horse kicks in the 19th century Prussian army — a small number of deaths within a large population of horses.

[5] Were a change in a DYS marker site occur by a non-random process (e.g., any situation related to the count that leads to a marker difference), the result would become more complicated to analyze. In this case we would be analyzing a random mutation process superimposed on marker changes that result from the test. We would then be dealing with two random processes with different origins (one natural, and one introduced by the counting process), and we would be analyzing either a random error plus a non-random error, or the superposition of two random errors. The presence of a second (counting) error would already have shown up in our statistics, and they have not done so. If we are dealing with the superposition of two random errors, both rare in nature, then we are dealing with a distribution that includes two Poisson distributions superimposed on each other. Again, the presence of such an error would have shown up in the work and there is no evidence of it. Thus, the application of Poisson statistics to a situation in which there were two causes of error would have yielded a non-Poisson distribution that we did not find. Since what we found is Poisson-distributed, then all the mutations we observe are consistent with randomness.

[6] In fact, when we study the effect of mutations over very long (genetic) time scales, n reaches a point of saturation while m continues to increase, showing that genetic distance is not as good a time indicator as RCC.

Submitted to the Journal of Genetic Genealogy on 22 April 2013
Revised 26 January 2014

Dr. William E. Howard III
1653 Quail Hollow Court
McLean, VA 22101-3234
703 532-8975; Email: wehoward@post.harvard.edu; Skype: wehowardiii