Frequently Asked Questions

The following FAQ results from questions I have been asked about the RCC approach to analyzing Y-DNA results.

QUESTION: What prompted you to write your first two papers on DNA analysis?

ANSWER: First, I became interested in taking the Y-DNA test because I wanted to see if I connected with Howards in England (no, so far!). I used FamilytreeDNA and became aware that they also made an estimate of the time to the most recent common ancestor (TMRCA) of any two testees that had the same surname. When I looked further into it, I found that the methodology by which people with the same surname were grouped by the surname administrators seemed to vary and that the method for determining the TMRCA was proprietary to the testing company. It was then that I began to explore an alternative method for determining surname matches. I found that my method resulted in a single number between pairs of testees (the RCC) that was correlated with the TMRCA. The first paper in the Journal of Genetic Genealogy (JoGG) derived the RCC vs. time relationship using four different methods, all of which appear to give similar answers that were then combined into one time scale. Further study led to the realization that that time scale could be applied much further back in time -- tens of thousands of years. It may have direct relevance to the time scales derived independently in other sciences.

QUESTION: Other methods of the measurement of time distance back to the most recent common ancestor from pairs of Y-DNA results exist (e.g., genetic distance and variance methods). Why choose another approach? For example, no mathematical justification of the appropriateness of the correlation metric as a substitute for genetic distance is given.

ANSWER: This field is still only a decade or so old. We should be looking for either the best method of determining surname groups or the TMRCA or seeking different methods that can be used together to achieve results that are more meaningful than any single method yields alone. Genetic distance suffers from the uncertainty of how to use increases or decreases in genetic distance that are a function of the mutation process that is occurring at each individual marker that has been tested. The definition of genetic distance has been variously calculated as the sum of the absolute values of the haplotype allele differences, or as the sum of the squares of these differences, or as simply a count of the number of loci that differ, or various hybrids that treat some loci one way and some another. The reason there are several competing algorithms for genetic distance is that mutations are statistically infrequent events, so that when one looks at even a large number of haplotypes with known pedigrees, no one model is particularly better than any other. However, the correlation technique attempts to get around these difficulties by selecting a large number of marker results (in this case 37, but it could be more) and by adopting an average mutation rate for the entire marker sequence. This average mutation rate is implicit in the process. When we use the marker string to be analyzed, the determination of the correlation coefficient between each pair of testees is a very simple process and avoids any intermediate set of decisions having to do with the details of genetic distance. Variance methods can be applied to the end result as the standard deviations of the distributions and their histograms are determined. Using correlation techniques to analyze differences between pairs of long haplotype strings, calibrated through pedigrees, we get around the complications of different marker mutation rates, although they remain the major uncertainties in the analysis of haplotypes.

QUESTION: Why not simply use the McGee Utility to set the time scale?

ANSWER: I am aware that administrators have been using the McGee utility, largely because it is easily accessible. Like the correlation method, it results in a matrix but it is based solely on genetic distance. While one can often spot related families clumped together by genetic distance, it still suffers from a lack of transparency that is afforded by the correlation approach and, as I point out elsewhere in this FAQ, the correlation technique incorporates individual marker differences in ways that genetic distance cannot do. It is hard to understand the details of how the McGee Utility works and to critique and understand the nuances that are built into the program. There is evidence that it is slowly being displaced by ASD methods which unfortunately many surname administrators cannot reproduce and do not understand.

QUESTION: Isn't the ASD method your main competitor? Why not just use it or Ken Nordtvedt's method?

ANSWER: My papers present an approach that will be used with the conventional one because it approaches the time scale differently and puts all the analyses on the same time scale that can be applied to all types of haplotypes - all at once. The time scale can be modified as a whole instead of calculating times haplogroup by haplogroup, as others appear to be doing. Note that I am not saying they are wrong -- just that I have another approach that adds a new dimension that might well be better. Until it is proven so, I am suggesting that it should be used in conjunction with traditional methods. Ken Nordtvedt's approach uses variance analysis that is difficult for a novice to follow; mine uses the standard deviation of a cluster matrix which is more straightforward and is both understandable and reproducible. They are different approaches, but both rely on variance and diversity as measures of elapsed time.

QUESTION: While the correlation approach can incorporate modal haplotypes into the results for comparison purposes, you don't use the modal as a reference point as other approaches do. Why not use it as a reference point?

ANSWER: Modal haplotypes can be useful if one wants to see how far a given haplotype is from a modal and we can compare every haplotype to the modal, which is assumed to be the haplotype of the 'founder/progenitor'. However, there is no general agreement about how to construct that modal, particularly if it occurs far back in time. Our analysis suggests that the modal haplotype is not that of the progenitor of a family line. You cannot easily identify a 'founder' if the testees you compare are unrelated except in the distant past. Using the mean value for genetic distances against the modal and using that to determine the age of the cluster is fraught with error, particularly when the definition of the modal is in question. The correlation approach finesses the modal issue through its ability to compare every haplotype in a matrix or surname cluster with every other haplotype. You cannot employ the modal approach to analyze haplotypes far back in time, and that's just what I am setting out to do using the correlation technique. There is no sense embedding the modal approach in the technique if you know it will lose its utility back in time.

QUESTION: The correlation technique uses a built in Excel function. However, in Excel it is equally simple to implement genetic distance as an array function. For example,

```
={SUM(ABS(B3:H3-B4:H4))}
```

would do it. [Note: the {} are added by Excel after pressing Ctrl-Shift-Enter.]. Why not use that approach?

ANSWER: The genetic distance between two pairs of testees that are in very different surname groupings or have different haplogroups or subhaplogroups will result from individual marker changes that can randomly change upward or downward and that there is a random walk from marker to marker away from the TMRCA, downward in time. The correlation technique essentially lumps this process into an average mutation rate, initially taken to be time invariant (but is testable), and reduces the effect by averaging over a large number of individual marker results. Moreover, the correlation process saves at least one step in the analysis without losing the valuable information that might be lost if only genetic distance was used. In addition, my work shows that the average mutation rate over the 37 markers tested by FTDNA probably has not varied by more than a factor of two over 70,000 years. This gives us a powerful license to use that rate over stretches of time that are far longer than those of interest to genealogists since it has genetic implications.

QUESTION: Do you not approve or distrust the genetic distance or variance methods?

ANSWER: Not at all. We are still at an early enough stage in the analysis of Y-DNA results that we ought to keep our eye open for any way that might improve how individuals are assigned within surname groups and how TMRCAs are determined, or that might add value or information to those existing methods. So far, genetic distance and variance methods have been used in studies that are more restricted. There is no evidence to suggest that the correlation approach is not applicable for determining the time where progenitors of surnames, different surnames, and different haplogroups lived, even many thousands of years ago.

QUESTION: Does the correlation approach avoid pitfalls in the more well-known methods?

ANSWER: It is clear both approaches depend on mutation rates which, at this time, are more uncertain when individual marker mutation rates are considered than they are when an average mutation rate over many markers are used. As improvements in individual marker mutation rates are made, the TMRCA will probably be more accurately determined by the more well-known methods but we are far from accomplishing that determination. However, improvements can then be made in the correlation approach by assigning weights to each marker. One can never conquer the uncertainties caused by the randomness of the mutation process, of course. If two testees have a TMRCA estimated to lie within 1000 years, the errors of the correlation technique are of the same order as the errors derived by using the better-known method. Errors in RCC due to number quantization have standard deviations of the order of one to four (from 50 to 250 years), for example, but those errors still exist in the traditional methods (see my first JoGG paper). Mutation uncertainties add to that uncertainty by about double those numbers of years, or perhaps even more. This works against the precision with which we can determine the TMRCA in the era over which we have pedigrees as the results of FamilytreeDNA have shown. Thus, errors may amount to as much as 50 percent (2-3 SD) in the near term - e.g., 500 years in 1000 years or so. Nevertheless, the RCC time scale can be used over time intervals of tens of thousands of years but mutations will still cause standard deviation errors of the order of 20 percent over these longer time intervals. My work is showing that we can make TMRCA estimates over a number of millennia and they will all be on a common time scale that has one assumed – and implicit -- average mutation rate. If future work shows a need to change it, it will be simple to revise the time scale to reflect those changes.

QUESTION: Is the correlation methodology derived from known principles governing the mutation of haplotypes?

ANSWER: Yes. The first JoGG paper presents the details and the second JoGG paper presents the application to different groups of surnames and the M-222 clade, many of which have distributions in more than one haplogroup. In an experiment, I assumed a haplotype of a progenitor, gave him three sons and, used a pair of random number generators (one for determining the marker that undergoes a change, and the other to determine whether the marker number increases or decreases), to follow marker changes down in time through 50 mutations. I determined the RCC values of each of the three haplotype pairs as they changed. When I used Chandler's mutation rate for an average over 37 markers, I could reproduce the same results that I got from a calibration of the RCC scale from over 100 pedigrees used for the RCC calibration.

QUESTION: What insights can the RCC matrix reveal about the evolution of haplotypes that the traditional approaches cannot do?

ANSWER: We can only test haplotype distributions at our current epoch. We must analyze that snapshot now in order to gain insight into the evolution of haplotypes and surnames in the past, through the present, and into the future. We assume, because there are no indications to the contrary, that the evolutionary process, via mutation, will continue as it has in the past. From the current snapshot, the RCC matrix, we are able to identify (1) embryonic subclusters within (2) existing surname clusters and (3) intercluster regions that were the precursors of the current clusters. Testees who carry the M222 subclade, can be considered to be members of "superclusters". We can infer how the subclusters of today will evolve into clusters, which will, in turn, evolve into future intercluster regions on the RCC matrix diagram over a few thousand years. We can infer from examples like the M222 subclade, when surnames originated and how they evolved. The correlation methodology enables us to discern how past evolution proceeded and approximately when it occurred. The traditional methodology, if it can do it all, is more ponderous and the approach is being assembled by more of a piecemeal approach, one haplogroup at a time. The correlation methodology permits us to do it in one single process.

QUESTION: What is the relation, if any, between the evolution of clusters and the ISOGG sequence?

ANSWER: Using the RCC matrix we show that the evolutionary sequence involving interclusters, clusters and subclusters in the RCC matrix bears a strong resemblance to the evolutionary sequence in the ISOGG that subdivides haplogroups and haplotypes in more and more detail, sometimes approaching a dozen subdivisions of a haplotype. In the second JoGG paper we predict that soon, the more detailed designations of a haplotype in the ISOGG sequence will converge toward the unique subclusters that are shown in the RCC matrix until the two sequences merge into one. The time scale provided by the RCC approach holds promise to tie the two processes together, with future time scale refinements of either one serving to refine the time scale of the other. In the beginning of 2011 we began developing an approach that takes haplotype strings of up to 500 testees and forms a phylogenetic tree that shows where each testee is located on the tree. It also shows points in time when the common ancestor of any cluster joins the common ancestor of another cluster. In addition, it shows the evolutionary path that every testee's DNA has evolved from the earliest progenitor of the entire database. When we produced a tree from modal haplotypes of various haplogroups, we basically could reproduce the ISOGG sequence. The RCC time scale was not applicable because we used modal values, however, but

the success of using modals to reproduce the ISOGG sequence was very encouraging.

QUESTION: Aren't several fundamental problems swept under the rug by using the correlation approach?

ANSWER: The boiling down of strings of numbers to one, the RCC, may appear to oversimplify the situation, especially if one depends on specific marker mutations as traditional methods do, to make surname associations. But, the power of the correlation approach is to use averages over large strings to make broader statements about relationships, particularly for epochs farther back in time than the traditional methods try to cover. In the second JoGG paper I have made studies of over 14 surnames that include different haplotypes and different haplogroups. They all show the power of applying average mutation rates to large strings of marker results. The correlation technique can not only reproduce the groupings that other surname administrators have assigned to the testees, but it has suggested members of some groups that probably do not belong (e.g., in the Howard and McLaughlin groups), and others that should belong. In the case of McLaughlins that are associated with the U'Neill descendancy, the RCC time scale for the presumed formation time is consistent with the historical record dating back over 1000 years.

QUESTION: A new metric is proposed and used, RCC, which is not directly or obviously related to Genetic Distance. It is based on the Pearson Correlation function in Excel, CORREL. Since all haplotypes are very similar, the correlation coefficient (CC) for any pair is >0.95. The Revised Correlation Coefficient (RCC) is 0 for exact agreement between pairs of haplotypes, and increases as the correlation decreases and as the TMRCA increases. Isn't the RCC, except for a scale factor, almost identical to the Pearson distance, (1 – CC), between the two haplotypes.

$$RCC = ((1/CC)-1)*10000 = ((1-CC)/CC)*10000 \sim (1-CC)*10000,$$

where the last approximation assumes CC is almost 1.

ANSWER: Genetic distance is only one way of expressing differences in strings of haplotypes. It suffers from the problems already mentioned. As the genetic distance increases, the RCC increases as shown both graphically and in the text of the first JoGG paper. Rather than using the formal result of the correlation coefficient such as 0.995322758777402 it is much simpler to map it, by the process outlined in that first paper, to another unique number, 46.99220611. That number can be carried along in future computations or can be rounded off at 47, since the errors in the number are of the order of RCC~3-4 depending on how it has been handled.

QUESTION: Why not use a simpler redefinition of the correlation matrix, for example, RC = 10^4(1-cc)?

ANSWER: A previous collaborator, Fred Schwab, suggested that, instead of RCC, we might use RC = 10^4(1-cc). As we began to apply the correlation technique to older haplotypes, we investigated these two definitions using a rather elaborate mutation model. We first investigated the behavior of the RCC and the RC definitions and how they related to the correlation coefficient (cc) over time. We drew the following conclusions:

- The relation between time and cc, RC, and RCC is linear to within 2 percent between the present and cc = 0.98, RC = 100, and RCC = 101. This means that we can use either definition back to about 2000 BC and for estimating dates of interest to genealogists.
- The curve of RC vs. cc is linear. At cc = 0.7 (over 100,000 years ago), RC= 3000.

- The curve of RCC vs. cc is non-linear. At cc = 0.7, RCC= 4286.
- The curve of the observed value of RCC with time is linear, and 10 RCC ~ 433 years.
- A correction factor, F, must be applied to the observed value of RCC to account for mutations that are know to occur, but, due to backward mutations, cannot be observed. The relationship between this corrected value of RCC with time is nonlinear.

QUESTION: The calculation includes multicopy markers. Because of the ambiguity in the information available from test results on such markers, shouldn't they either be omitted or treated conservatively? For example, shouldn't DYS464 probably be omitted?

ANSWER: I have investigated this situation and find that the differences between using it or omitting it are minor compared to the other sources of error that affect both this and the traditional methods. Moreover, since the selection of markers to be tested are somewhat based on the desire to investigate both fast moving and slow moving markers, the inclusion of all the markers are preferable to excluding selected ones. However, because of the inclusion of markers with a variety of mutation rates, it is very important to be sure that the same 37 markers are used throughout the RCC matrix analysis.

QUESTION: FTDNA finds the values of DYS 464 a, b, c and d and reports them in increasing order. Since you not know which value belongs to which marker, why do you include them? Doesn't that adversely affect the correlation result?

ANSWER: Since I used these markers in my very laborious calibration via pedigrees, I was really concerned that this criticism was a valid one. So, I took a couple of typical marker strings and permuted them around, recomputing the average changes and their standard deviation (SD). The result was that the average RCC of one marker change was
3.5 (with SD of 1.1) in one study (185 years)*. A 2d result was that the average RCC of one marker change was 3.2 (with SD of 1.0) in the 2d study (170 years)*
* Computed with 1 RCC = 52.7 years, the value of the TMRCA for a cluster, not 43.3 years, the value for a pair) I have run models where changes were made with and without such sorting and have found that the correlation result is well within the errors expected when any other marker change happens. Those differences amount to about 2-3 in RCC. The changes are actually lower when the markers are sorted in increasing marker number.

QUESTION: The first JoGG paper states, "While it (i.e., the correlation approach) works on average mutation rates for each string of markers, it treats the marker differences automatically, without the need for human decision-making." Isn't some of that decision-making unavoidable if meaningful results are desired?

ANSWER: I have found that the groupings that individual surname administrators make cannot be adequately defined. They tell me that " I will know a group exists when I see it." In the correlation approach, when one plots the distribution of RCC values in a histogram, the RCC limit where a group's boundary occurs (i.e., when a value of RCC belongs to the group or not), is generally well-defined by that distribution. It may vary slightly from surname to surname, but it can be better defined as the break point in the distribution of RCC values. In early 2011 when we were able to present the phylogenetic tree and an associated RCC time scale, we found that the clusters formed by the application program, Mathematica, is done at least as quickly, objectively and effectively than a project administrator can do it.

QUESTION: Isn't there serious systematic weighting bias against markers with values that are far from the haplotype average? For example, the 37-marker haplotype for the Hamilton project (Modal A) has an average allele value of 15.7. A change of 1 mutation at DYS576 (from 17 to 18) changes RCC by 3.01. But, a change of 1 mutation at CDYb (from 36 to 37) changes RCC by only 2.17. And, a change of 1 mutation at DYS 455 (from 8 to 7) changes RCC by only 2.88. This is not random; markers with allele values close to 15.7 make a larger contribution to RCC than markers that are either much larger or much smaller. The effect is somewhat masked at 37 markers, but is fatally serious at 12 markers (see the table in Ref. 2 of the ms). Shouldn't this be fixed by the arbitrary pre-subtraction or -addition of an appropriate constant, different for each marker, to precondition the data before calculating RCC? It should make only a small numerical difference, but it's important to reduce bias if possible.

ANSWER: This is a good point. Up to now my concentration has been to see what the correlation approach can do. The numerics cited in the question are correct. Such effects are noted in the first JoGG paper. The questioner rightly points out that the effects of a unit marker change is different depending on the initial marker value, but that marker change can be both upward and downward and the effect is well-masked by the time we consider 37 markers. The first paper points out that the value of RCC appears not to change with the number of markers tested, except that its uncertainty is far lower at 37 markers. That is why 37 and not 12 or 25 markers are chosen for the analysis. It is indeed fatally serious at 12 markers. The small numerical difference that would result from a small pre-subtraction or -addition of a (small) appropriate constant, different for each marker is probably not warranted because (1) the effect would be very small and masked by other uncertainties, and (2) such a fix would be quite arbitrary as the question points out. Besides, we are searching for bigger issues than this one. As the approach is further refined, perhaps with the addition of weights when the individual mutation rates become better refined, this issue can be addressed in order to reduce a bias which, right now, is very minor.

QUESTION: RCC appears to be a reasonable (although biased) surrogate for genetic distance for RCC < 10, but the uncertainties rapidly build up. Of course, for small changes, everything looks linear. Isn't a linear regression calculation doomed to get a slope and intercept out of any data you give it?

ANSWER: Not at all. First, I know of no argument or any indication that the calculation of RCC is biased. The second JoGG paper points out that there are no indications that the RCC vs. time relation is significantly nonlinear out to times that are very large compared to the number of years within which conventional pedigrees are available or even out to several tens of thousands of years. Any uncertainty present should not exceed the same percentage uncertainty that is present at the shortest time intervals. Linear fits are made only to data that appear to be linear.

QUESTION: The large uncertainties in the attempted calibration of RCC vs. time propagate rapidly. Isn't it inappropriate to extrapolate beyond the time span of the oldest pedigree? It would appear that this problem is not peculiar to the use of RCC but it is shared by the genetic distance approach whose use beyond a few hundred years is subject to debate. Don't the uncertainties of genetic drift and population substructure in recent times render the calculations impossible for now?

ANSWER: The question assumes, without any reason, that uncertainties of RCC vs. time propagate rapidly, and rapidly is not defined. The results in the second JoGG paper show that

there is no extrapolation. In fact arguments are presented that define approximately the shape of the RCC vs. Time relation and the small extent of its non-linearity. This argument presents the probable end point where the value of RCC appropriate to 40,000 to 70,000 years represents only a very mild, if any, departure from non-linearity from the present back to that era. This is certainly not an extrapolation of the data. It spans a time interval that amounts of more than 40 times the time interval appropriate to pedigrees and this makes it a very powerful tool to explore haplotype differences between testees that are in very different haplogroups. I know of no attempts to take genetic distance back that far, but the correlation technique seems to do it, and with reasonable assumptions. In fact our current study of the Gordon surname shows that the formation of interclusters and clusters occur at the times when the glaciers receded and when surnames originated, respectively.

QUESTION: For the calibration plots, shouldn't the regression package (Tools| Data Analysis| Regression) in Excel be used, rather than the Charting routines? This will give you standard errors for the slope and intercept, which are directly relevant to the error/uncertainty analysis. More recently, updated versions of Excel in Microsoft Office have not included the Data Analysis Tool Kit that includes the correlation and other useful tools. How can the correlation be done with more modern editions of Excel?

ANSWER: This question is confusing and indicates a lack of knowledge of either Excel or what was done to get the results in my first two papers. The correlation technique in Excel leads to the RCC values of the pairs of testees. Any time a plot is presented from data, the Chart routine in Excel is used to produce the graphs. Then, when appropriate, trend lines are derived from the data, including the uncertainties in the plotted points, and the equations for the fits are derived. Along with the fits, the variance of the plotted points from the equation of the fit is calculated. Here, the variance is R^2. The standard deviation of the points is the square root of the variance. That is how it was done. Microsoft Office eliminated the Data Tool Kit in recent versions. But Microsoft has suggested that a free download is available that will do these types of analyses. I have downloaded the third-party Tool Kit, but I am still using the 2004 Excel version that still works with my Macintosh computer because I have found that when problems exist that I cannot solve, the three parties, Apple, Microsoft, and StatPlus-Mac pass the blame, saying that the problem is not their fault.

QUESTION: In attempting to achieve simplicity for a complex problem, don't the papers go too far in ignoring important features of the mutation process?

ANSWER: No, but without further definition of "important features" this may not be a full answer to the question. Even if some features have been ignored, the value added in such areas as (1) the time scale, (2) the sequence of evolution of surname subgroups, (3) the time these subgroups take to evolve from its parent group, (4) the fact that only one RCC vs. time relation is derived instead of separate times for separate haplogroups, etc. introduces a degree of simplicity to a very complex problem. But the RCC process does this by using an average mutation rate over a large number of markers. If that average mutation rate is in error, one has only to introduce a scale factor in the RCC vs. Time relationship. For this reason, when I present the results in a phylogenetic tree, I present the time scale in units of RCC, letting the viewer make the time conversion which, right now appears to be 10 RCC= 433 years.

QUESTION: Since RCC is not particularly easier to calculate than genetic distance, why not use that? Genetic distance at least has a theoretical underpinning and acceptance.

ANSWER: The question is somewhat naive. In fact, RCC is easy to calculate. One has only to set up a marker matrix composed of a 37 marker string of numbers and any number of participants. The Excel correlation routine can be applied in seconds, even to hundreds of participants. Conversion into the RCC scale takes less than a minute. I do not doubt that genetic distance has a theoretical underpinning. But the RCC approach gives results that agree with results obtained by traditional techniques, and they can be applied to longer time scales with the ease mentioned earlier.

QUESTION: What happens when you group the small numbers together in the RCC Matrix? I can't picture the clustering process. Can you explain how it is done?

ANSWER: This is the most labor-intensive part of the clustering operation.
You form a matrix that consists only of numerical values of RCC, using the correlation function and the approach outlined in my first JoGG paper.
Be sure that the sequence of testees down the page is identical to the sequence across the page. Then color in the diagonal.
The intersections of a testee with himself should be zero, but remove the zeros because they will appear elsewhere in the matrix different pairs of testees have identical haplotype markers.
Then, by inspection, identify the low RCC values and group the rows as tightly as possible by cutting and pasting the haplotype strings so that they are adjacent to each other.
Once that is done, cut and paste the columns so that the new sequence of entries is the same in the rows and the columns. You can tell if you have done it correctly because the colored diagonal will again line up. Once this has been done, the clustering will jump out of the matrix at you. An example can be found in Figure 1 of my first JoGG paper where I present the results for a group of two Logan clusters. In that Figure 1 of that paper, the clusters in the boxes contain low RCC entries which delineate the presence of a cluster. Then, in the intercluster shaded area, the RCC values are higher. The TMRCA of Cluster A is the average RCC of the entries in Cluster A and the TMRCA of Cluster B is the average RCC of the entries in Cluster B. Note that the average of Cluster B is lower than that of A, so Cluster B is younger. Since the RCCs in the shaded intercluster region are composed of the RCC of one member of Cluster A and one member of Cluster B, the TMRCA of the intercluster will be the common ancestor of the TMRCA of Cluster A AND the TMRCA of Cluster B. That, in a nutshell, is one of the powers of using a matrix prior to the formation of a tree. You can also set up a 'time-slice' matrix approach that will show only pairs of RCC that are located within various cuts/intervals of time. You can show this by inserting the high and low points of RCC and the program does the rest. In Figure 1 of the first JoGG paper, the high value of RCC is set at 75, high enough to show all the entries in the matrix. The use of the time-slice approach adds even more power to the analysis. You can even make a movie of it, showing how the values of RCC appear in the matrix as a function of time. My JoGG papers present further details.

THIS SECTION WAS ADDED AFTER PHYLOGENETIC TREES BECAME AVAILABLE IN EARLY 2010, WITH AN RCC TIME SCALE INDICATED ON THE TREE.

QUESTION: Neither the genetic difference nor the correlation approach takes into account changes that occur in particular markers. Isn't it important to identify and study the actual markers that are changing from haplotype to haplotype and from cluster to cluster?

ANSWER: Yes, and much work needs to be done in this area. Because the RCC method only looks at the whole marker string, it ignores which of the markers are the culprits of change. We

know the effect of changing small and large value markers and by how much, but we have not concentrated on which markers are changing. Even though Mathematica makes a 'call' based on only the degree to which the haplotypes differ, we recognize a few things – (1) that clusters defined by Mathematica are based only on the DNA changes in the marker strings, independent on which markers do the changing; (2) that the markers that change tend to be characteristic of a new DNA line of descent; (3) that if you compare which markers have changed during the evolution of surname clusters, you may be able to follow those changes within the individual haplotypes; and (4) if you have additional information in which you have trust, like a pedigree, you should put reliance on the pedigree unless it would do violence with Mathematica's position of that cluster on the tree. While we have spent some time analyzing these internal marker changes of clusters defined by the RCC correlation technique, we have reached no conclusions about it or how to best do the analysis. This is an area very ripe for further exploration.

QUESTION: Then what is the role of pedigrees? How does knowledge of pedigrees interact with a testee's position in an RCC matrix or on a phylogenetic tree? FTDNA's Tip™ process gives probabilities of an association with another testee. How does the RCC approach do that?

ANSWER: You can use the presence of a good pedigree as an additional piece of valuable information in both the Tip™ and the RCC analytic process. You can use this information as part of a Bayesian approach. If you know other information than just the RCC output, it improves your chances of making the right decision. A knowledge of pedigree information increases the probability that your grouping is better than it would be without that knowledge. With Tip™, the probability of finding a TMRCA at certain specified generations changes if you know you are NOT connected with another testee in the near term. With RCC, you can conclude that the testee should be nearer another testee or nearer a cluster than without this knowledge. In my mind, I don't really pay much attention to the 1 in 37, 2 in 37, 3 in 37 matches, etc., but prefer to go with Mathematica's call or to improve it with a pedigree. Having said that, I think there is much to say about looking at the specific markers once we know that a cluster or cluster+pedigree looks good. This is territory that needs a lot of new attention because I think that specific markers define the cluster and can separate it from others nearby. That's the first step. The second step is to use the tree and try to see which markers morph as once goes from one cluster to another. Finally, Mathematica does the best it can with the haplotype strings, but it knows nothing about pedigrees, and pedigrees certainly add to our confidence in making and defining clusters. Nevertheless, mutations lurk and Mathematica knows nothing about them. So, this is a 'caveat emptor' for us.

QUESTION: What is the relationship between FTDNA's Tip™ probability and the RCC time scale?

ANSWER: We have considered 72 pairs of haplotypes. For each of them FTDNA has determined (1) their number of marker mismatches and (2) their probability of sharing a most recent common ancestor within 8 generations at 37 markers. For each pair we have determined their RCC values, also using 37 markers. When the probability is plotted against RCC, we find that the relation is linear (variance = 0.77) within a range of RCC between 0 and 18. The approximate equation that links the two is:

$$P(8 \text{ generations}) = 100-(20*N75)/3$$

For probabilities less than 10 percent, the relationship flairs out and becomes much less trustworthy compared to the RCC scale. At that point, however, we are at the limit of interest to genealogists. We conclude that FTDNA's probability is in reasonable agreement with the RCC time scale above P = 10% and RCC less than 15-20

QUESTION: The applications program you use to form the dated tree from STRs, Mathematica, has seven options that can be used. You use the average option. Why choose that one and not the others.

ANSWER: Mathematica has seven options that can be used to optimize the tree positions using hierarchical clustering approaches with input from Y-DNA STR haplotypes. They are: (1) Single (uses the smallest intercluster values); (2) Average (uses the average intercluster values); (3) Complete (uses the largest intercluster values; (4) Weighted average (uses a weighted average cluster dissimilarity); (5) Centroid (uses the distance from cluster centroids); (6) Median (uses the median intercluster values; and (7) Ward (uses Ward's minimum variance dissimilarity. Since we used averaging when we considered the intercluster regions in the RCC matrix, we saw that the Average option in Mathematica best agreed with the results in the intercluster. When we compared the tree positions with the trees produced in the Gordon-Howard paper in the JoGG http://www.jogg.info/72/files/Gordon.htm we found that the time scale of the groupings we made in the figures agree nicely with the time scale of the tree.

QUESTION: When the Y-DNA results of testees who share a surname are positioned in clusters on the tree, they often trace to a progenitor who lived before the beginnings of surnames. Can you explain the nuances of this evolutionary process? What happens as the evolution progresses from that progenitor to the recent sets of testees.

ANSWER: Y-DNA traces ancestral lines independently of surname. As the progenitor's line descends to the present, it will experience mutations along the way and the number of lines of descent will increase, with each line containing distinctly different marker values. Near the year 1000 AD surnames were chosen. Some males along these lines chose different surnames although their Y-DNA remained quite similar. If a progenitor lived significantly before the appearance of surnames, not all his descendants will carry that surname because recent descendants will be located along different branches and will have chosen other surnames. When we confine our attention, and our tree, to only one group of testees, our testee clusters will contain only the surnames we chose, so the tree will not be fully sampled since it is a biased sample due to the way we choose our selection of testees for study. The other branches of descent, containing other surnames, are part of the downward probability distribution. Their extent is not known because we do not know how many lines were generated, but their presence and probability distribution must be recognized when we consider surname clusters that are not fully sampled. The tree is generated only from a restricted set of data that we have chosen to analyze. It is probably more reliable to consider the intercluster regions on the tree in order to understand surname positions on a tree.

QUESTION: What factors determine a testee's position on the tree? Why does a testee not appear in a cluster when his pedigree links him with others who do cluster on the tree? What effect does a marker change have on a testee's position on the tree?

ANSWER: At least two things determine a testee's position on the tree. First, mutations may be

different for one cluster member than for the others. That will separate him from the others on the tree. Second, the program optimizes the entire 37-marker string, not just the low values of RCC that will indicate his cluster membership with the rest. If you look at the testee's result in his RCC matrix string where his RCC values are shown for all other testees, you may find a sequence of low values of RCC within that part of the matrix that would indicate cluster membership. But remember that the program looks at the whole string, not just those at low RCC values, so the program may position the testee as an outlier on the tree. If you believe the pedigree, use it and not the position on the tree. I have done a fairly extensive study of the effect of marker changes on the position on the tree, and here are my rough conclusions => The logic goes this way: We start at a point in time that corresponds to when surnames were chosen (in order not to have the data too messy with respect to pedigree lines since they hopefully follow the surname, say at RCC = 20. Then models I have run indicate that a 37 marker set will have a mutation about every 137 years, or about 5.5 generations, as you go down the line from progenitor to testee, i.e., from RCC = 20 to zero. (See my first JoGG paper) There will be, on average, about 6.3 mutations along the line, but that can vary a great deal, perhaps by +/- 2.5 mutations. In any one line from the progenitor at RCC 20 down to a testee, there will be between 3.5 and 8.5 mutations along any one line (1 SD= a probability about 68% in this interval). As you progress down the evolutionary lines from a common ancestor to the present, the testee lines will 'fan out' due to mutations. The more mutations, the greater the distance apart they will be in RCC and on the tree. In general, testees who are in a tight cluster may have mutated less than normal and testees that fall outside a cluster may have mutated more than normal. One mutation corresponds to a 2.6 to 3.3 change in RCC depending on what marker value does the mutating, and a change of one RCC is equivalent to about 50 years. So one mutation is equivalent to about 130-150 years. You can immediately see the problem — the error in RCC is often close to the value of RCC, itself for testees who are closely related on a pedigree. Since errors in RCC may average 2-3, they can go to 6-9 occasionally (three SD), so I would be cautious about trusting relationships that differ by 9 or less. For that reason, positions on the tree could be quite different from where you would put them via a pedigree. Put differently, one should not over-interpret the position of a testee on the tree.

QUESTION: Can you identify the progenitor of a particular line on the phylogenetic tree, i.e., the time of origin of a particular line?

ANSWER: The appearance of a son on a pedigree is an identifiable event, but in the case of DNA, a branching on the tree indicates a mutation, not a birth event. I have found that many beginners don't understand the difference. In the case of a haplotype, the marker values contain little of value to an analysis until they are compared with something else, generally another haplotype. This means that most comparisons take place by pairing. When you look at a phylogenetic tree, the largest RCC you will see results from a pair of testees. All you know is that at the earliest junction point, the RCC at that point indicates the first time you see his DNA on the tree. If this is a junction between two haplogroups, it is a strong indication that the common progenitor had a mutation that separates the two haplogroups on the tree, and the downward descent from there on represents different lines in different haplogroups. But that is not the full story. There is a history at times earlier on the tree that we know nothing about. This means that the time of origin we see is a more recent limit on the actual origin. Put another way, it is a lower limit. Moreover, since male lines often die out, this is another reason why the real progenitor may have lived earlier. That uncertainty diminishes when more and more testees become available, because it may lead to more lines that go back even further in time, but you cannot beat the fact that male lines will die out along the way. I have felt for sometime that it

would be very important to estimate the probability that lines have died out along the way, but there are great uncertainties in each epoch of history that include how many males were produced in a family, how many families actually existed, whether plagues or other events seriously reduced a population, the effects that location and climate have on a population and the overall growth rate of the population. These are serious impediments to this type of estimate. Another way to put it is that lines dying out and incomplete sampling causes almost insurmountable problems. In an ideal world, everyone would be tested and would appear somewhere on the tree, BUT, some of the lines that died out may extend back further than the MRCAs that we can identify on the DNA tree. This means that we derive a lower limit on the TMRCA and that's bothersome to the folks who try to date origins of SNPs or haplogroups. One can only strive for a big sample and hope that our analysis comes close.

QUESTION: The RCC values that appear in the RCC matrix do not correspond exactly to the RCC values at the junction points on the phylogenetic tree. Why is this?

ANSWER: The original RCC values must be thought of as the raw input to Mathematica and each of those values contains the usual uncertainty belonging to any pair of testees. Each pair of entries suffers from an error that I estimate to be of the order of about 23 percent (one SD). You don't know *a priori* which are the good ones and which are bad. But when those data are inputted to Mathematica, the program works on them in ways that will minimize the SDs of the information on groups that it produces. When Mathematica uses those values of RCC, the program computes their optimum distance from each other on the tree. We may think of the process as having many RCCs attached to each other by virtual rubber bands that Mathematica rearranges so that the tree is optimized. This means that the distance between pairs of RCC in the input matrix will differ from those in the tree because Mathematica is optimizing their placing relative to others in each database that is used. This means that if the same pairs of testees are in one database, their placing on the tree will be different than their placing using a subset of that database. Thus the output of Mathematica, now in the form of a tree, displays the best that the program can accomplish with the data at hand. The tree is optimized and the original matrix is not. Again, you cannot beat mutations, but when one looks at a lot of testees at once (as we do in optimizing both the matrix and the tree via the program), we get an average result that may be nearer to reality than when we just look at the components. We can show this statistically. If there are 10 surnames in a cluster formed by the tree, the errors involved will be of the order of 23%/sqrt(10-1), or 7.6 percent instead of 23. The errors will go down by roughly the square root of the numbers of items that are in a group we are analyzing. We must keep these considerations in mind when we are making conclusions about where testees or groups appear on the phylogenetic tree. Singleton entries will have larger time and position errors than larger groups. In the course of this particular investigation which ranged over RCC values of up to 600, we found (1) no evidence of a departure from linearity in the time scale, and (2) the means of the intercluster regions on the original RCC matrix were the same as their junction points on the tree to within the margins of error.

QUESTION: Can you be more specific? For example,
(a) how do the intercluster RCC matrix averages compare with their corresponding values found at the junction points on the phylogenetic tree?
(b) we know the standard deviations of intercluster values in the RCC matrix and we know the average RCC values of the interclusters as well as the RCC values of where the junction points appear on the phylogenetic tree. How to they compare?

ANSWERS:

(a) When the average RCC value of an intercluster in the RCC matrix is plotted against the corresponding RCC value on the phylogenetic tree, we get the following relation using 21 interclusters: Matrix RCC = 0.9574 Tree RCC, with a variance of 0.51.

The large variance reflects the fact that we are dealing with mutations that Mathematica attempts to minimize, but the relationship shows that the application preserves the RCC during optimization to within about 4-5 percent. Thus the two time scales of the matrix and the tree are preserved.

(b) When the standard deviation of each RCC intercluster in the matrix is plotted against the average RCC of the corresponding matrix intercluster, we get the following relation:

SD of RCC matrix intercluster = 0.2083 x average RCC of the intercluster matrix, with variance equal to 0.0745.

When the standard deviation of each RCC intercluster in the matrix is plotted against the corresponding junction point on the phylogenetic tree, we get the following relation:

SD of RCC matrix intercluster = 0.2073 x RCC junction point on the tree, with variance equal to 0.5996.

The slopes of these two relations are virtually indentical, but the larger variance of the second relation shows the very significant improvement achieved by Mathematica when the junction points are identified through the optimization process.

QUESTION: You have calibrated the RCC time scale using 37 markers. Why not do the calibration using 67 markers?

ANSWER: At the time the calibration was done the product of (number of available markers times the number of people who had good pedigrees) was much larger at 37 markers than at 67. This was also the case regardless of the number of pedigrees that were available. Once more 67-marker results are available for testees with good pedigrees, is would be advisable to redo the RCC calibration using 67 markers.

QUESTION: In what ways does the phylogenetic tree using 67-marker haplotypes differ from the tree using 37-marker haplotypes?

ANSWER: The position of a testee on the 67-marker tree may differ from the position of the same testee on the 37-marker tree, just as the position may be different when different members of the same database are inputted to the program. Both explanations are the same – the program optimizes the ensemble of haplotypes. For the 67-marker result there are more markers that require optimization; for the 37-marker result, different markers are being optimized. While this might be viewed by some as a criticism, to a first approximation the same clusters are preserved and the general shape of the tree is also preserved, although some testees might be moved from one edge of the tree to the other edge. We analyzed 209 haplotypes of testees that had 67-markers and produced a 37marker and a 67-marker tree. We selected 23 pairs of testees, determined their MRCA junction points on both trees, plotted the pairs of junction points and found that the slope was 1.03, very close to unity ($R^2=0.74$). This indicated that the time scales, on average, were the same for the 67-marker tree. The scatter in the data indicated an SD of about 27 percent. This standard deviation is large, but expected because different DYS values were being optimized. There is no indication that the RCC time scale derived from the 37marker calibration cannot be used on a 67-marker tree, but that individual pairs of testees will have results that differ by ~ 27 percent (SD), so caution is advised when drawing conclusions based on small groups of data.

QUESTION: Many people produce cladograms that show the distance (usually marker differences) between people on the tree and the modal. Why don't you use modals?

ANSWER: Let's do the following 'thought experiment' -- Compute the modals for a group of testees and include that modal as a separate inputted haplotype when forming the phylogenetic tree. You will find that the modal is located near the middle of the phylogenetic tree, not near the ends. So, its location on the tree is not meaningful. It is not representative of a progenitor. Nor are the distances from the modal meaningful for any particular testee. The cladogram uses the modal as a distance reference, while the RCC approach computes the time using the matrix RCC values. There is a big difference in these two approaches. I am not saying that a cladogram is wrong. It probably gives useful information, but I am will concede that the RCC-derived tree, used with a cladogram, may be more powerful than using either alone. I have not looked more deeply into cladograms.

QUESTION: What is needed to produce an RCC matrix and a phylogenetic tree?

ANSWER: First you need an Excel spreadsheet or one that has a tool to do correlation. The spreadsheet must be configured this way:
They must be the raw marker numbers.
They should be in the order that FTDNA reports them on their web sites.
They should not include testees with zero values in any marker.
They must be in an Excel spreadsheet.
They must show identifiers in Columns 1 & 2. People usually choose Kit Numbers in Column 1.
Remaining Columns 3-39 must show marker values separately, not grouped as FTDNA now is presenting it.
Individual testee results should be in rows.
The marker numbers must be numerical, not text entries.
Columns after Column 40 can also show identifiers, but those data will not show up on the tree.

In summary, as Y-DNA haplotype results accumulate, it becomes increasingly difficult to analyze the results for patterns using traditional techniques. What is needed is a metric, a single number that can be calculated for any pair of haplotypes, which will reliably indicate their probable degree of relatedness. The RCC approach has the following attributes, only some of which are shared by approaches that use analyses of genetic distance:

- it is fast and easy to apply to very large marker matrices and numbers of haplotypes,
- the results account for marker mutations that have taken place over tens of thousands of years,
- it presents a time scale that can be tested as the results of future research become available,
- the time scale that results is scalable over all haplotypes,
- there is no evidence as yet that the RCC time scale can not be used when the number of markers exceeds 37 (but this needs further exploration),
- it leads to the formation of surname groups, and often of groups within groups,
- it permits a determination of which members belong or don't belong to surname groups derived by traditional methods.
- it indicates through a study of pairs of individuals, when each member of the pair belongs to a different subgroup, when those subgroups evolved from their parent group, how long it took to evolve and how future evolution will probably take place.
- the results can be presented in the form of a phylogenetic tree that will indicate an evolutionary history of a testee's Y-DNA from earlier progenitors, down in time to the

present, showing points at which mutations took place and through times when surnames were selected.

QUESTION: Can you summarize for us the errors involved in the analysis?

ANSWER: A DRAFT paper concerning errors in the RCC approach and the positions on a dated STR Y-DNA Phylogenetic tree is in preparation. A short summary of observations and conclusions is given here.
My colleague, Fred Schwab, wrote a Mathematica code that mutated each allele in a 37-marker beginning haplotype, using Chandler's 2006 mutation rates. The program choses one marker at random at each time step of 157 years and mutates the current value up or down at random. The new haplotype is presented at the next step and its RCC value is computed by comparing it to the starting marker haplotype.
We continued this process through each of 1460 time steps (229,600 years) and traced the effect of random mutations down one line of descent. We made 20 averages of 100 individual runs. We drew the following conclusions for times of interest to genealogists:

- The relation between time, the correlation coefficient and RCC is linear to within 2 percent between the present and cc=0.98 and RCC= 101. This means that we can safely use the RCC approach definition out to about 2000 BC to estimate dates.
- The curve of RCC vs. cc is non-linear. At cc = 0.7, RCC= 4286. However,
- The relation between RCC and time is linear.

We next investigated times of interest to geneticists.
The model also produced 50 individual runs that were used over 1460 steps to compute values of RCC and its standard deviation (SD) at each step. We found that the relation is linear out to more than 200,000 years although the percentage error of an RCC determination of a single haplotype pair lies between 30-50 percent from the present to 70,000 years ago. Those errors are large. They arise from minor effects like number quantization, which causes SDs of the order of one to four in RCC (50-250 years), to the larger mutation errors discussed above. Quantization errors at low values of RCC work against the precision with which we can determine the TMRCA in an era where we investigate pedigrees and surname projects. If we only compare two haplotypes, an SD error of 40 percent can give rise to a total error of two times that uncertainty (two sigma, or two SD if the distribution of errors is Gaussian). As an example, if a pair of haplotypes has an RCC of 10 (433 years), it will have an SD of about 35 percent (150 years), but error analysis indicates that about five percent of the time it will be in error by 70 percent (300 years) or more. The error situation improves when more than one pair of haplotypes are involved, a situation that occurs in surname clusters, interclusters, and haplogroup clusters. In those cases, the distribution of RCC values will generally allow the SD of the average of a group to be calculated assuming we can use Gaussian statistics. The average of the RCC values in the group is computed first. The SD of that distribution can be estimated using the table, below. Then, if there are n testees in the group, the SD of the average RCC of the entire group will be the SD of the distribution divided by the square root of (n-1).

RCC interval	Time Interval (Years ago)	Percentage (SD/RCC) or SD/Time ago)	Estimated error of the Percentage
0 to 2300	0 to 100,000	43%	4%
2300 to 5300	100000 to 200000	52%	9%

There is little statistical difference between these two time intervals, and since Y-DNA data fall into the lowest time interval, 43% +/- 4% is the figure to be used. Sources of error such as

departures from linearity or RCC quantization are very small compared to the effect of mutations. They can be ignored.
Our model spans a time interval that is more than 40 times the time interval within which pedigrees can be used in conjunction with RCC analysis. The linearity of the RCC time scale over this period of time and the relative constancy of the percentage of SD/RCC make a very powerful tool to explore haplotype differences between testees that have common ancestors far back in time or that are in very different haplogroups. The recognition that only one RCC vs. time relation needs to be used instead of deriving separate times for separate haplogroups, etc. introduces a degree of simplicity to the analysis of a very complex problem.

QUESTION: The sequence of positions on a dated phylogenetic tree derived from STRs is often not in the same sequence of haplogroup subclades, determined by SNPs, that appear on FTDNA's tree and on the ISOGG tree? Shouldn't they be in the same time sequence?

ANSWER: (I thank FTDNA's helpdesk for answering most of this question)
The processes by which time and evolutionary sequences are derived using SNPs and STRs are not the same. SNPs and STRs are located in different parts of the genetic code, so a sequence derived from only SNPs should not have a one-to-one relationship with a sequence derived from only STRs. Ages are estimated by STR variance, and in general this will line up with the SNP order, but it may not always. For example, if SNP A evolved very soon after SNP B (his 3rd son, say), and then SNP A went on to have many sons, who had many sons, etc, compared to SNP B's much more modest population growth, then SNP A's population quickly grows much larger than SNP B's population. Population size doesn't dictate age, but mutations have about as good a chance of happening with one birth as with another, regardless of what SNPs the father has and how old that SNP is. If more members of a haplogroup have sons, there will be more opportunity for mutations to take place. If this happens very early on, then today, after thousands of years have passed, the accumulated genetic variance makes the haplogroup look much older than another haplogroup that did not experience an early population explosion. In the SNP A vs. SNP B example, SNP A has an early population explosion and SNP B does not. Ten thousand years later when we test members of SNP A, we see a lot more variation in the STRs because they have had more opportunity to mutate than we see in SNP B. Of course, the accumulated mutations per lineage will probably be similar because they are of the same age, but population size may make a difference in the calculations of age, depending on how the calculations are done. Additionally, some haplogroup branches seem to experience mutations more frequently than others. Whether that is due to chance or some other mechanism, we do not know, but this too can make the age calculations of STRs mismatch the relative ages determined only by SNPs. Although a SNP sequence may be correct, the boundaries of what defines the STRs of a sample of haplogroup subclades like Q1a3a are probably quite broad. Their distribution on an STR-derived tree may sometimes impinge on the tree boundaries of an adjacent, earlier subclade, making a testee of Q1a3 appear younger than an adjacent testee who is in subclade Q1a3a. Research devoted to determining the ages of SNP sequences is still in considerable flux, with additional subclade symbols often being revised yearly.

ACKNOWLEDGEMENTS: I wish to thank my many correspondents who have read my papers and have had questions about the correlation technique. Their advice and counsel have lead to improvements in our papers as well as improvements in the RCC technique and in the understanding and interpretation of our results.

- Dr. William E. Howard III. McLean, VA
wehoward at post.harvard.edu
Revised January 28, 2013