Uniting the Time Scales of Genealogy and Genetics Using Correlation Techniques to Explore Y-DNA

-- William E. Howard III --

Version 24: 25 February 2014

Abstract:

We have investigated the errors and uncertainties involved in the RCC correlation approach to the analysis of Y-DNA haplotypes using extensive mutation models with weighted DYS values. We conclude that: (1) there is no evidence that the observed RCC time scale, properly calibrated, cannot be used over time periods of at least 100,000 years; (2) that the ratio of the standard deviation to the RCC value at each mutation averages about 43 percent (SD = 4%) over times of genetic interest; (3) a correction (F>1) needs to be applied to observed values of RCC over genetic time intervals to account for backward mutations that are known to occur but cannot be observed; (4) the distribution of marker values as well as the Chandler average mutation rate over 37 markers is consistent with Poisson statistics, confirming that the distribution of mutations take place randomly; and (5) using an extrapolation of junction points on a Y-DNA STR phylogenetic tree, we are able to estimate the date of the progenitor of Haplotype A to about 100,000 years ago (est. SD ~ 30%) as well as the dates for other haplotypes and SNPs. We suggest error assignments that should be applied to pairs and groups of haplotypes and we present methods to estimate the time when the progenitor of a group lived. We investigate the error assignments to be made in the analysis of groups of haplotypes (e.g., in surname clusters and junction points on a dated Y-STR phylogenetic tree). These errors directly affect time relationships in the ancestral lines of testees and the uncertainties in the determination of the dates when progenitors of SNPs or other ancestral lines lived. Using correlation techniques we are able to unify the genealogical and genetic time scales so that a single time scale can be used over times as long as 100,000 years ago.

Introduction and Background:

The RCC correlation approach was originally developed to analyze genealogical relationships (Howard 2009). We quickly recognized that the RCC approach could be useful over the longer times involved in genetics research. When two haplotypes are correlated, the correlation coefficient (cc) lies between 1.00 and 0.98. After calibrating this relation with pedigrees, we found that this cc range referred to the time interval between the present and 8800 years ago, respectively[1]. We simplified the relationship by converting the cc to a revised correlation coefficient (RCC). The conversion used is: RCC = 10^4(1/cc-1). We recognized that cc would not be linear with time, but in the interval between the present to 8800 years ago, it departed from linearity by less than 3 percent. A time scale error of only 0.2 percent would result within time periods less than 1000 years ago (RCC <20) when surnames were adopted. We decided to use a set of models to investigate these errors more

extensively. The models also offer valuable insight into mutation-driven errors that might be expected over the much longer time intervals appropriate to all Y-DNA haplogroups.

The First Model:

Our first approach is patterned on the model used in Howard (2009). We began with a starting 37-marker haplotype string whose DYS values fell at or near the midpoint of the range of values that have been observed for each marker. We used the mutation rates for each marker reported by Chandler (2006). Those values are given in Table 1.

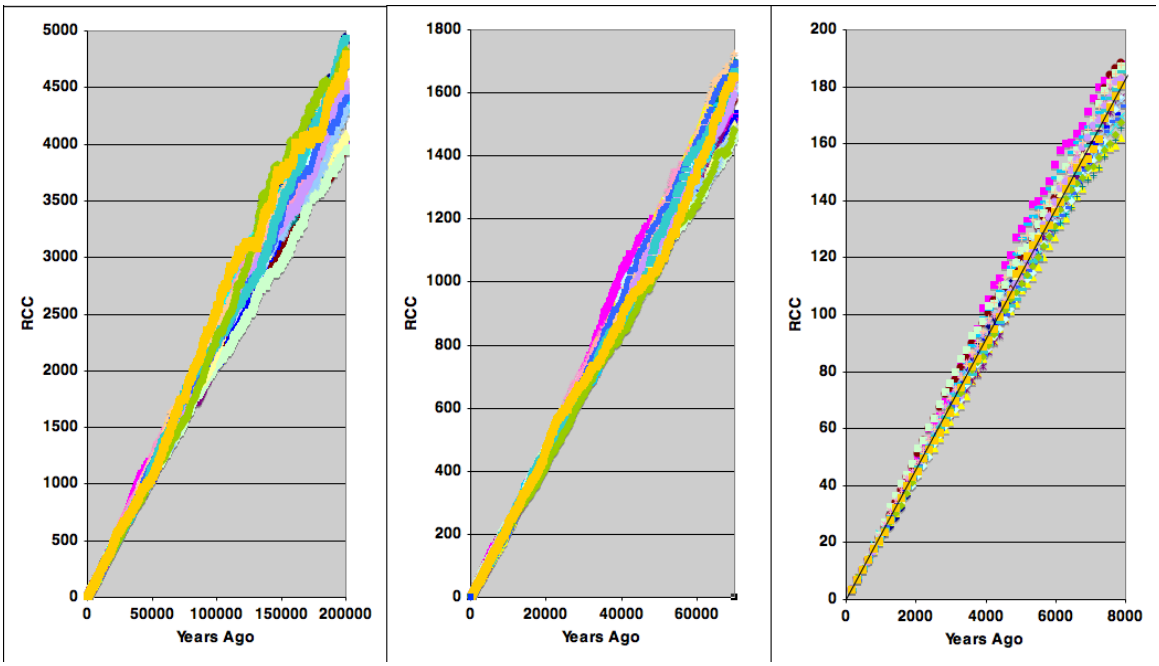
Table 1: The Values of the Beginning Model DYS Markers and their Mutation Rates

Starting Marker Values in FTDNA's DYS order Sequence[2]: 13, 22, 14, 10, 16, 16, 12, 12, 12, 14, 13, 30, 17, 9, 8, 10, 14, 24, 19, 18, 30, 15, 15, 15, 15, 10, 10, 18, 22, 20, 20, 20, 20, 23, 33, 20, 15. Chandler Mutation Rates in Same DYS order (The Chandler rate is 10^-5 times the values listed below): 76, 311, 151, 265, 226, 226, 9, 22, 477, 186, 52, 242, 814, 132, 132, 16, 16, 264, 99, 135, 838, 566, 566, 566, 566, 402, 208, 123, 123, 735, 411, 1022, 790, 3531, 3531, 324, 55.

The Mathematica codes (Wolfram 2010) were written by Fred Schwab. Starting with the marker values in Table 1, we selected one of the marker values and changed it by one unit, up or down. Each time step was one mutation. We then changed the new haplotype at one marker at the next step. At each step the RCC was computed by comparing the haplotype at that step with the starting haplotype. Random number generators within the program were used at each step (1) to select the marker to be mutated and (2) to determine the direction of marker mutation. The higher the mutation rate, the more often that marker was chosen to be changed.

We continued this process through each of 1460 time steps (229,600 years) and traced the effect of random mutations through time down one line of descent. From a study of the mutation rates found by others and from the results of the models, below, we adopted a value of 157 years per mutation, the length of a time step[3]. The results were presented in both graphical and tabular form. Figure 1 shows the results of running the model computation twenty independent times over 1460 time steps (229600 years; each step is 157 years, with one mutation at each step).

Figure 1: Mutation Model Results Showing 20 Runs of RCC vs. Years in the Past (Each of the 20 runs contains the average of 100 runs)

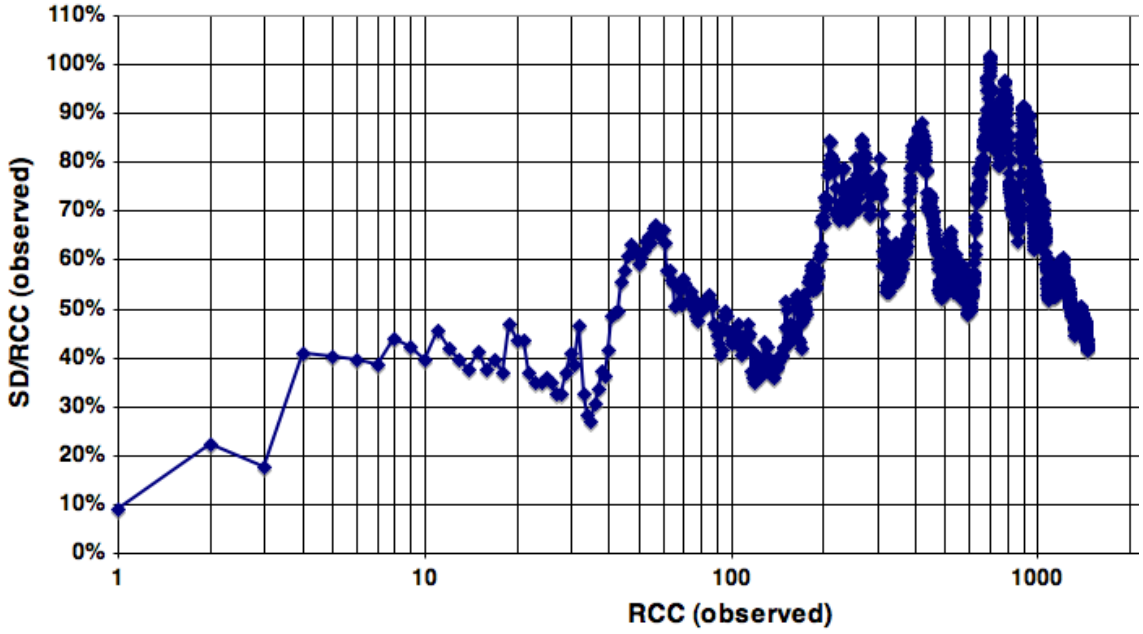


The left graph of Figure 1 shows results of each of the 20 runs, going back an equivalent of 200,000 years. It shows that the relation of the model-derived RCC with time is linear and it shows the divergence of the model runs over those long time periods. The middle graph shows results out to 70,000 years, the time scale of interest to geneticists in the study of Y-DNA haplotypes and haplogroups. The right graph shows results within the time scale of interest to genealogists, the period from ancient history to the present[4].

The Standard Deviation (SD) of an RCC value – Errors in Estimating Time:

The model also produced 50 individual runs that were used over 1460 steps to compute values of RCC we derive at each step. The standard deviation (SD) of RCC at each of the 1460 steps was computed. Since we want to find the percentage error expected when we compute an RCC value, we show that percentage error in Figure 2 as a function of the model-derived value of RCC.

Figure 2: The percentage error of the model-derived ratio SD/RCC over the 1460 mutations in the model, the interval of interest to Y-DNA haplogroup investigations.



By inspection we see that the percentage error of a model-derived RCC determination lies between 30-50 percent over years of interest to Y-DNA investigations appropriate to genealogy (RCC<40). Beyond that point, the ratio in the model runs fluctuates more widely, sometimes reaching 100 percent into the distant past.[5] The first two mutation time steps (out to RCC~3-4) appear to have lower-than-average values of SD, because mutations have hardly begun to take place from the starting progenitor's haplotype. The SD rapidly climbs to the region near 40 percent where it remains, fluctuating around that value until RCC ~ 40.

Rapidly mutating markers tend to drive differences in marker values throughout more recent time periods. Marker value differences begin to average out as time progresses. As a consequence, the slowly mutating markers begin to show their presence over longer time periods[6]. As pairs of haplotypes evolve with time, their marker values diverge and the model-derived RCC values increase as shown in Figures 1-2.

The Assignment of Error Bars to RCC and Date Determinations

Single Pairs of Haplotypes

Figures 1 and 2 can be used to estimate errors associated with a single pair of haplotypes, whether they appear as pairs of testees in surname clusters, or haplogroup clusters of any kind. 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 during times where we investigate pedigrees and surname projects.

Figure 2 shows the results of the detailed model that can be used to estimate the errors

expected over time intervals of genealogical interest. Table 2 summarizes these suggested error assignments.

Table 2: Suggested Standard Deviations to be used for RCC and Time Determinations.
(The estimated error of the SD percentage is 4% at RCC < 2300 and 9% at RCC > 2300)

RCC (observed)[12]	Standard Deviation
4 - 30	40%
40 to 200	50%
200 to 1000	70%

Since Y-DNA data of genealogical interest fall into the lower time interval, 40% (4% SD[7]) is the error estimate to be used for genealogy. The effects of RCC quantization are very small compared to the effect of mutations. They can be ignored.

If we compare only two haplotypes, and if the errors are distributed randomly, an error that exceeds one, two, and three SD is expected to occur about 32, 5 and 0.27 percent of the time. For example, if a pair of haplotypes has an RCC of 10 (433 years), it will have an SD of about 40 percent (170 years), but error analysis indicates that about five percent of the time it will be in error by 70 percent (300 years) or more[8].

Groups of Haplotypes

More than one pair of haplotypes usually appears in surname clusters, interclusters, and haplogroup clusters. The addition of more haplotypes in a group will reduce errors that are associated with a common ancestor of the group as a whole. In those cases, the distribution of observed RCC values will generally allow the SD of the average of a group to be calculated using Gaussian statistics. The average of the RCC values in the group is computed first. The SD of that distribution can be estimated using Table 2. 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).

The Second Model:

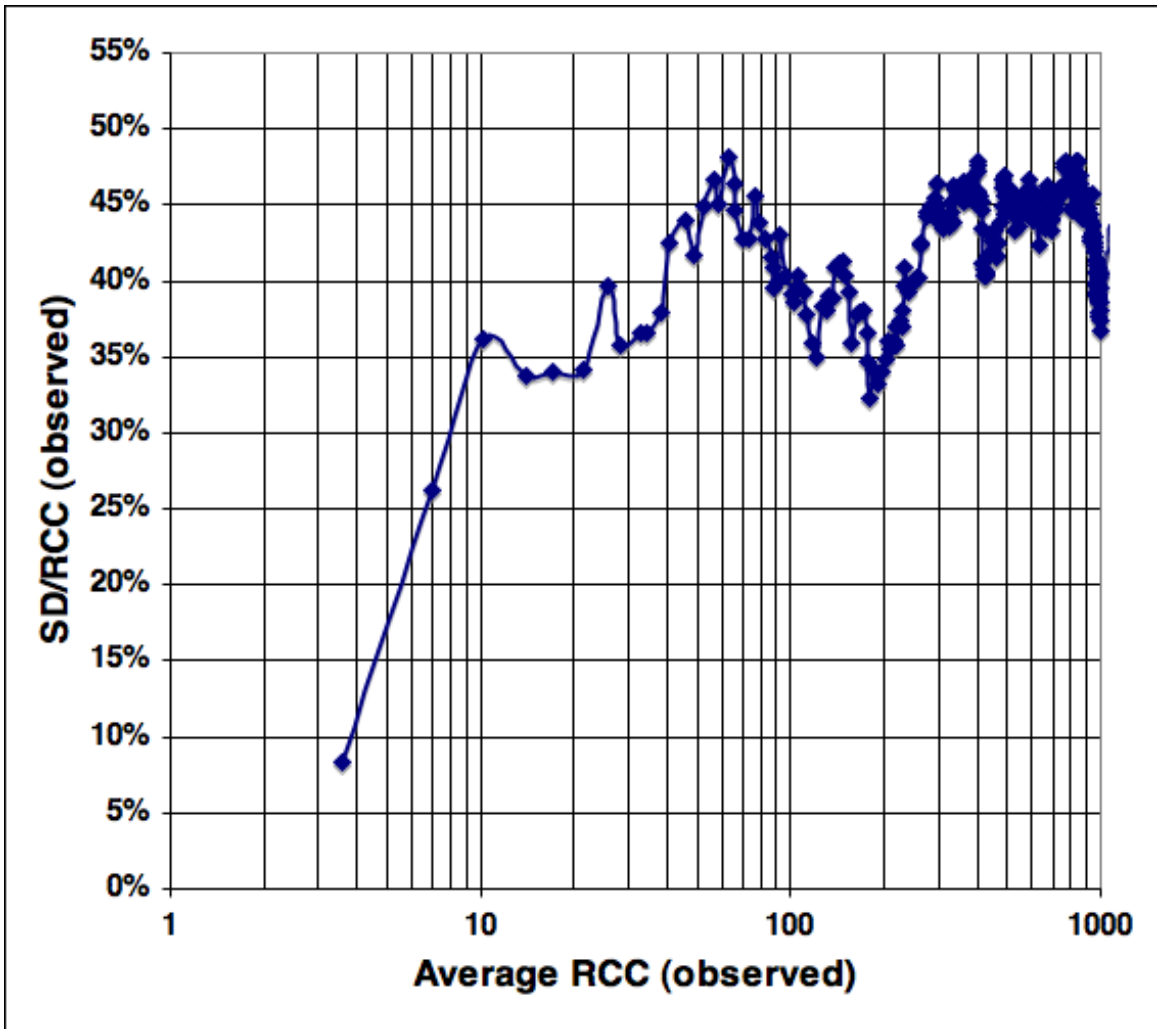
In the previous model, we took the beginning haplotype and computed RCC at each time step of a random process out to a large number of steps down one line of descent. In a second model, we consider two lines of descent and compare the model-derived RCCs of the two marker strings at each time (i.e., mutation) step rather than to compare them to the beginning haplotype.

We found that when the model-derived RCC values between two haplotypes are cross-compared at the same mutation step level, the ratio SD/Mean RCC is relatively constant at about 41+/- 2% (SD) averaged over 200 horizontal model runs. Figure 3

shows the ratio of the standard deviation to the mean value of RCC at each mutation point in the sequence of 20 time steps.

This ratio appears to remain relatively constant with mutation step level, except for the first 1-8 mutations (viz., to 1100-1300 years ago, or RCC ~ 25-29), shortly before surnames were assigned.

Figure 3: The Ratio of the Standard Deviation of the model-derived RCC to the Mean model-derived RCC as a Function of Average Value of RCC.



As expected, the average values of RCC found when we compare two haplotypes with each other at the same time step level, were double the values when a haplotype at that level was compared to the progenitor. This is because the number of mutations down the two lines totals twice the number down one line. The results in Figure 3 are consistent with those in Figure 2 for this range of mutations. They are remarkably similar. The errors expected from investigations in the time interval of interest to genealogists are derived from these results.

There are additional uncertainties that include: (1) the definition of a generation (e.g., 25 to 31 years); and (2) the interpretation of multi-copy markers[6]. When we apply the RCC correlation technique in our analysis, we first separate the FTDNA reporting into separate marker columns, keeping their same order. The multi-copy markers (385a, 385b; 459a, 459b; 464a, 464b, 464c, 464d; YCA II a, YCA II b; and CDY a, CDY b) are kept in the same order, and that order was the one we used to calibrate the RCC time scale using over 360 pairs of testees of pedigrees in four surname projects (Howard 2009). In a separate investigation of multi-copy markers, we found that permuting the order of the markers in DYS 464 results in a change in RCC of only 2.8 +/- 25% (SD) per marker change. Since our time calibration and our analysis are consistent with the same order of the multi-copy markers, the errors introduced through the use of multi-copy markers will be minimal or non-existent.

When Did the Earliest Mutation of a Group of Haplotypes Occur?

During the course of this study we recognized that since the correlation approach uses observed mutation changes among the 37 markers of the haplotypes, it requires a correction factor, F, if there are backward mutations present in either pair that we do not observe. Since a second mutation is not generally expected for 4 to 5 generations, we may use the observed RCC, equated to about 43.3 years for calculating the TMRCA during most of genealogical time. We now explore two cases, first, where we derive the TMRCA of the progenitor of a surname cluster, and later, a second case where we derive the time of origin of a haplogroup or a SNP. In the latter case, we are in the genetic time interval where the correction factor must be used.

Case 1: The TMRCA of the Progenitor of a Surname Cluster

Estimating the time to a common ancestor of a surname cluster is difficult because we only sample its total membership when we determine the average RCC of its members. We may have missed some critical haplotypes whose inclusion would have led to an earlier time. In Howard (2009) we used a heuristic approach that suggested that the TMRCA would have lived at a time corresponding to an RCC of 52.7, rather than the time corresponding to an RCC of 43.3 determined from the calibration of pairs of haplotypes using pedigrees. This approach in Case 1 is a specific application of the more generalized situation in Case 2 where the TMRCA occurs in a time interval (RCC above about 15), where the correction factor, F, should be used.

Case 2: Time of Origin of a Haplogroup or a SNP

We encounter an even more difficult problem when we try to estimate the time when a mutation forms a new haplogroup or when the progenitor of a SNP lived[9]. All date estimates are derived from haplotype pairs that are already on a line of descent from the progenitor. At least four factors may lead to an underestimate of those dates:

1. The sample may be incomplete;
2. We can only estimate the TMRCA of the pair, not of the earlier time when the

single progenitor lived;
3. The chance of survival of a lineage from a single founder through 20 generations is only 9.6%, so lines have died out from the progenitor to the present over that time interval[10].
4. A recognition that mutations have occurred that we cannot observe due to back mutations or recLOH events[11].

When we want to find the TMRCA of testees whose MRCA dates earlier than the situation in Case 1, we are back into the genetic time frame where the number of expected mutations exceeds about 3, and the observed value of RCC is >10-15. The more pairs of testees that appear in the group, the more certain will be TMRCA of the entire group. Under the reasonable assumption that the distribution of RCC of the pairs in the sample is Gaussian, the errors, as before, should decrease as the square root of n-1, where n is the number in the sample.

In both cases, the inclusion of as large a number of haplotypes in the sample as possible should reduce uncertainty. The positions of the junction points in a dated STR phylogenetic tree will be more certain the more haplotypes are presented to Mathematica's optimization process[12].

Using RCC to Determine Genetic Time Scales:

Past success of the RCC correlation approach to group Y-DNA results of testees with similar surnames on a dated Y-DNA phylogenetic tree has been encouraging. While the RCC time scale is affected by the same mutation-driven errors and uncertainties as other approaches, it can group clusters on the tree quickly and more efficiently than most surname administrators can do, and it often indicates the clusters to which unassigned test results belong. But the junction points of the branches on the tree in the distant past were appearing at a times more recently than expected. The RCC-derived dates (10 RCC=433 years) in the distant past appeared to underestimate the expected times by factors of the order of 1.5 to 3.[13]

Sidney Sachs deserves credit for pointing out that the mutation model that we had been using counts all mutations, while, in reality, a DYS marker change is missed whenever the change takes place in one direction and then mutates back. In formulating the model that takes account of all mutations we made the following assumptions:

- Mutations take place randomly.
- When a mutation occurs, the marker number will change up or down by one unit with equal probability.
- Mutations occur singly, one at a time, but they sometimes may occur rapidly and be misinterpreted as two simultaneous mutations.
- Mutation rates do not change with time, over 100,000 years or more.
- The mutation rates at each marker site are those in Table 1 (Chandler 2006).

- The average mutation rate over 37 markers is the Chandler mutation rate, 0.00492 +/- 15% (Chandler 2006). This rate translates to 0.182054 mutations per generation. A mutation occurs on average once every 157 years among the 37-marker set. These relationships are consistent with an average generation of 28.58 years. This number is an average since markers mutate at different rates.
- Over a long period of time, mutation counts can be used as a time clock.
- Counting mutations will lead to a time scale, with one mutation equaling one time step in the model. A true count of mutations that average out over long time intervals can be used to derive a time scale in which a corrected value of RCC is proportional to the elapsed time to the MRCA of a pair of testees. This time scale can be calibrated using large numbers of pedigrees. Thus we need to multiply our observed value of RCC by a factor F, a function of time, which will convert our observed value to the corrected value.
- We equate the corrected value of RCC to be the model-derived RCC, which will be linear with time. Thus,

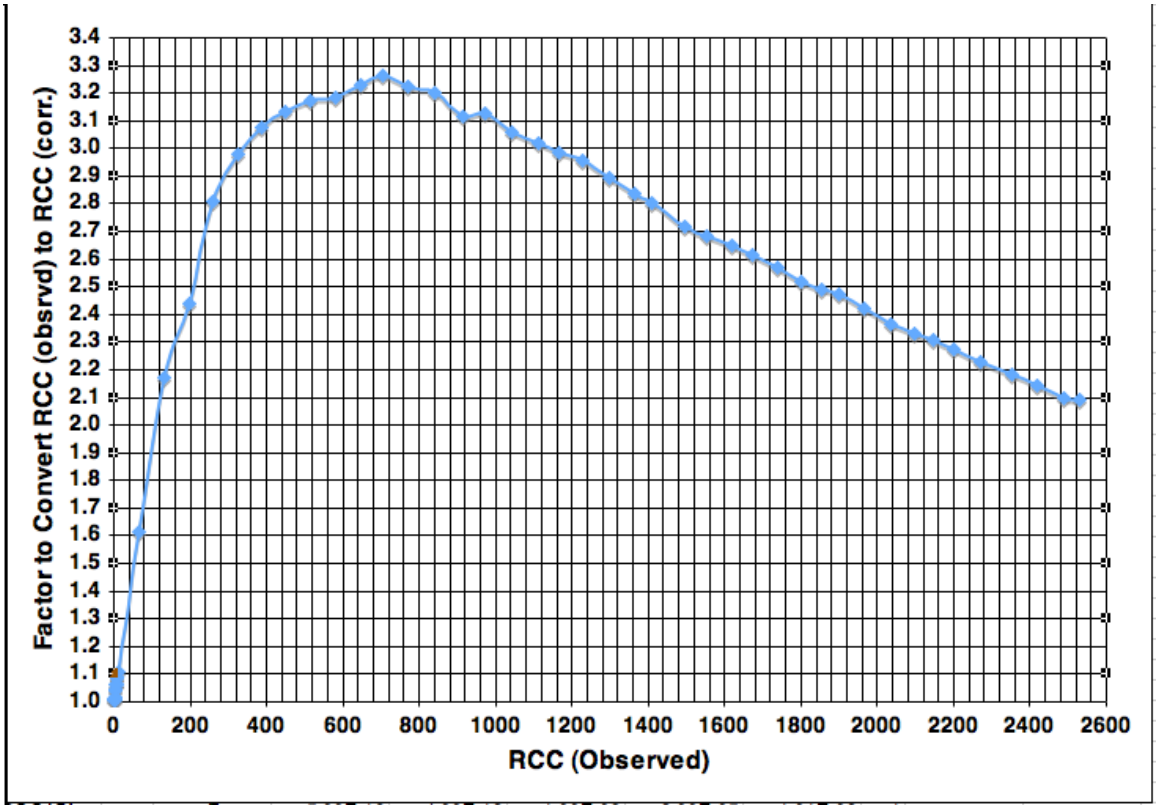
RCC (corrected, model-derived) = F x RCC (observed)

The Mathematica code that produced the model results in Figure 1 uses the mutation rates of individual marker sites to select a DYS site at random and then mutates that marker value randomly, either up or down. The model counts every mutation. In reality, however, if a marker changes in one direction and then changes back, the correlation approach misses the mutation, and the result will produce an apparent mutation rate that will be slower since the correlation technique cannot detect these backward mutations. The process nulls out some of the number of actual mutations. The RCC we observe will be smaller than it should be because the analysis does not count all mutations that we know must be occurring. These observed RCC values will indicate too recent a date, causing the junction points on a dated Y-DNA phylogenetic tree to appear nearer in time than they should. Thus the conversion factor (F) will be equal to, or greater than unity[14].

The conversion factor was derived by averaging the results of two slightly different codes: (1) a modification of Fred Schwab's Mathematica code we used to derive Figure 1 in this paper, and (2) a code produced by Sidney Sachs in Quick Basic that uses generations instead of mutations. The approach used was to change the marker values in only one direction, so the process of using a random number to determine the direction of the marker change was dropped. The RCC values produced by this one-way marker change model (RCC+) were then compared with the RCC values produced by the two-directional marker change model (RCC+/-). The ratio of RCC+ to RCC+/- is the correction factor, F. The two codes yielded virtually the same result shown in Figure 4.

Figure 4: The Correction Factor (F) By Which Observed Values of RCC Are

Converted to Corrected Values of RCC.



After multiplying RCC (observed) by F, we derive the relationships found in Figure 5.

Figure 5: Relationship Between the Observed and Corrected Values of RCC.

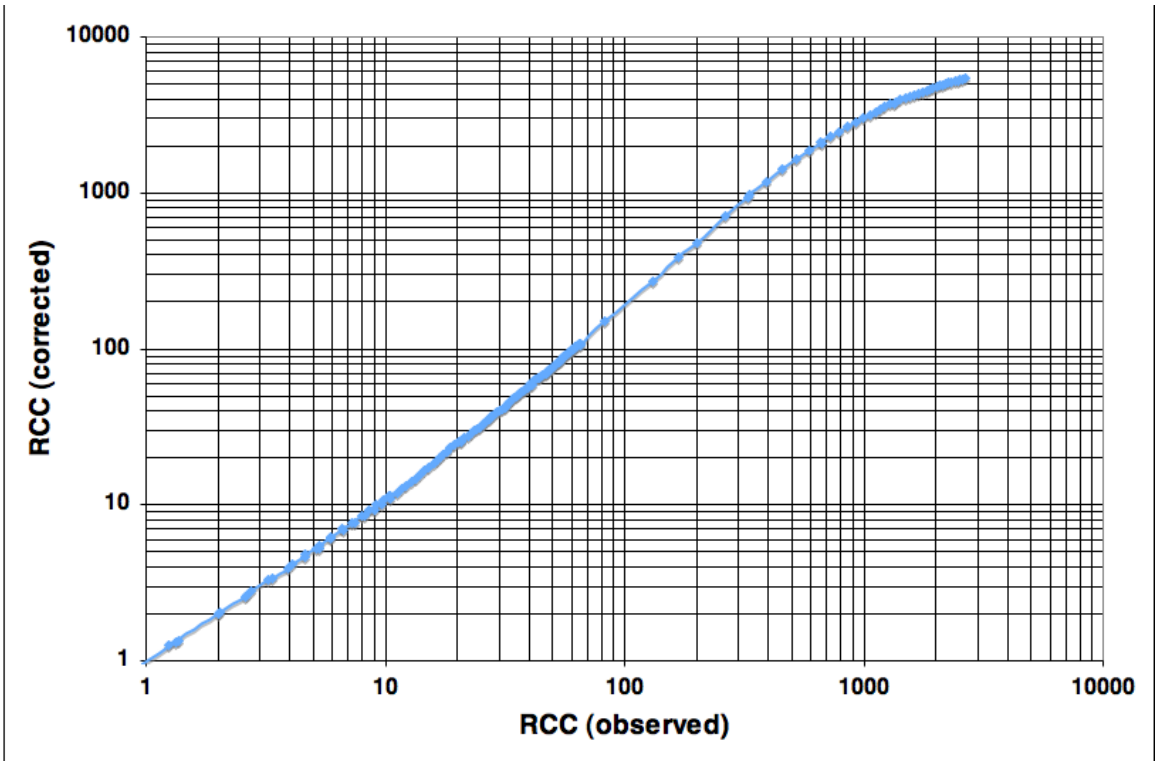


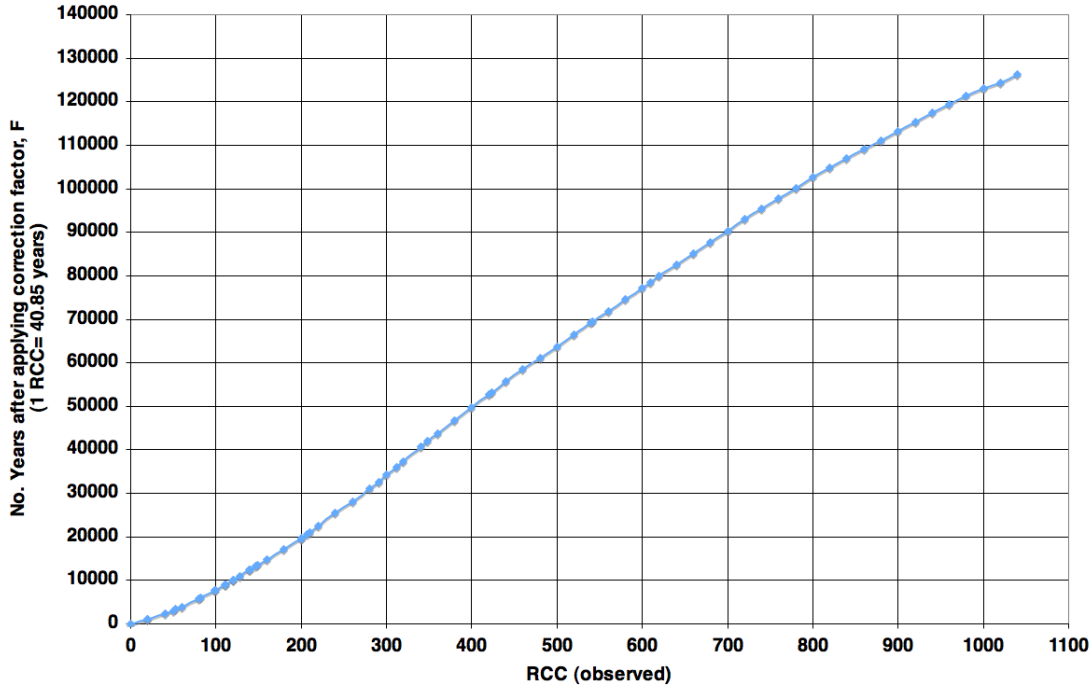
Figure 5 shows that the conversion between RCC (observed) and RCC (corrected) is not significantly different from unity for values of RCC less than about 15; for practical purposes the correction can be ignored within time intervals of interest to genealogists. The original calibration of the observed RCC time scale using pedigrees is well within the uncertainties due to mutations. However, for the higher observed values of RCC beyond the time when surnames were adopted, and in intercluster regions, the factor F needs to be applied. Figure 5 shows the corrections needed throughout both the genealogical and genetic time scales. These conversions are model-dependent, of course, but if the assumptions we made in applying the mutation model remain valid, the conversion should be valid beyond RCC (observed) ~ 1200 to 1300, the highest observed RCC values found in our studies[15]. The data from which the figures were made appear in Appendix A.

Once we have converted an observed RCC to a corrected RCC, we can use the latter to derive a time to a MRCA of a pair or group of testees. In previous papers we have used the relation 10 RCC = 433 years. In the calibration of the RCC scale using pedigrees we find that the average date to the most recent common ancestor of the 100+ pedigrees we used was about 1600 AD. This corresponds to 345 years for which the correction factor F would be 1.06. Therefore, a new derived calibration yields 433/1.06, or 10 RCC = 408.5 years. This new calibrator should be used in cases where genetic time scales are appropriate, at values of RCC (observed) > 10-15. Therefore:

Time (Years) to the MRCA = RCC (corrected)* 40.85 Equation 1

Figure 6 shows the number of years in the past that correspond to the observed value of RCC after the correction factor F has been applied.

Figure 6: The Number of Years in the Past That Correspond to the Observed Value of RCC



The relationships among the observed value of RCC, the correction factor F and the corresponding number of years from the first 37 markers reported by FTDNA are found in Appendix A.

The Application of the Corrected RCC Time Scale to the Origin of Haplogroups

These results show that the RCC correlation technique can be used for dating Y-DNA haplotypes throughout the very different times of interest to genealogists and geneticists. To show the validity of the approach, we investigated an extreme example, estimating the time when the progenitor of Haplogroup A lived.

We first selected a large sample of nearly 1000 haplotypes for which their different haplogroups had been identified. We divided them into groups by subclade. When there were three or more 37-marker examples in a subclade, we computed the modal haplotype of the group. We then computed the RCC matrix, a histogram of the matrix and a STR phylogenetic tree for the 115 modal haplotypes of the haplogroups that resulted from this process.

Next, we selected sets of haplotypes that shared a common haplogroup, and computed for each set its RCC matrix, the histogram of that matrix and its phylogenetic tree.

The Y-DNA phylogenetic tree is shown in Figure 7.

Figure 7: The Y-DNA Phylogenetic Tree of Modal Haplotypes of 115 Haplogroups and their Subclades. The RCC time scale (abscissa) is shown prior to correction.

The results of the positions and junction points on the STR tree agree in general, but not in detail, with the evolutionary time sequence determined from SNP studies by the International Society of Genetic Genealogy (ISOGG) and the European DNA site of Eupedia[16]. The overall time sequence, from the oldest A haplogroup through E, C,

F, and G is evident, but this Y-DNA tree of STR results does not show better agreement with the SNP evolutionary time sequence because our phylogenetic tree contains only a sample of STR modal haplotypes. The SNP sequence results from analyses that involve extensive, on-going studies of STRs, how they aggregate to form SNPs, the nesting of SNPs, the evolutionary relationships among SNPs and their associated time scales. If we inadvertently choose examples that have had an unusual number of mutations, their positions on our tree will not agree exactly with the time sequence of SNPs. We see this in Figure 7 where isolated modal haplotypes appear that are connected through a single line of descent or that appear as a lone entry among a cluster of other, closely-related haplogroups.

The junction point that joins the earliest modal haplotype A with the rest of the haplotypes does not represent the time when the progenitor of Haplotype A lived. The earliest connection on the tree is between pairs of Haplogroup A at an RCC of 610. When that RCC is converted to time, the junction of the two lines shown on the tree occurred about 78,500 years ago. This is the time of the earliest junction point of a pair of modal haplotypes, but this is not yet the earlier time when the progenitor lived. The first test is to explore that date of origin of that single progenitor, often referred to as Y-Adam in the literature.

For this, we will address and then discuss the question:

<u>When Did the Progenitor of the Oldest Haplogroup A Live?</u>

Earlier, we suggested the following three approaches for estimating the time when the single progenitor of a group lived (see papers by Howard in References). We now present the results of these approaches. We believe that the first approach is the most valid, with the second two as confirmatory.

<u>First Approach:</u> Extrapolating the Junction Points on the Dated Y-STR Phylogenetic Tree

In this approach, we count the number of times a line to a junction point on the tree crosses successive values of RCC on the time axis. We then plot the logarithm of the number of points against the date in the past that correspond to those successive numbers of RCC and extrapolate the graph to Log N = 0, the point at N=1 when the SNP originated or the progenitor of the line lived (see Figure 3 of Howard and McLaughlin 2011). Figure 7 presents the phylogenetic tree. Note that the RCC time scale on this tree shows the observed value of RCC before applying the correction, F. Figure 8 shows the results of the extrapolation.

Figure 8: A Log Log Plot Showing the Extrapolation of Lines of Descent to the Putative Date of Origin of the Progenitor of the 115 Modal Haplotypes of Haplogroups and their Subclades Derived from Counts of the Lines of Descent on the Y-DNA Phylogenetic Tree. The time scale is shown after the correction factor, F, was applied to the RCC values in Figure 7.

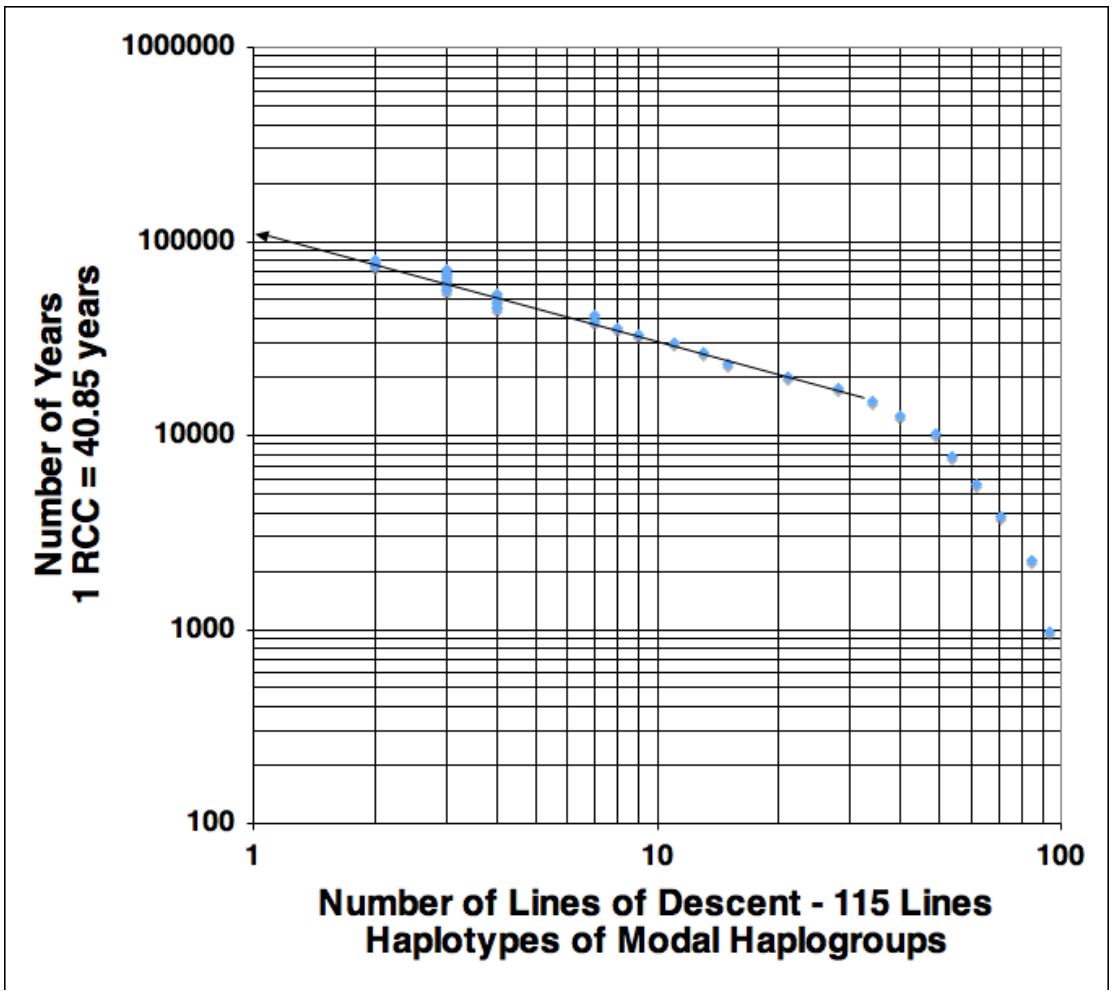


Figure 8 shows that when the number of lines of descent older than 20,000 years are extrapolated to when the progenitor lived at N = 1, we find that they fall along a line with very small scatter that points to the origin of the 115 modal haplotypes at about 103,000 years ago, but with an estimated standard deviation of about 30 percent.

The right hand side of the chart turns downward when more lines of descent, driven by an explosion in the world's population, appeared. At that time, mankind began to make a transition from hunter-gatherer to farmer in an environment that led to increased levels of procreation. More males led to more mutations that led to the appearance of new sets of haplogroups.

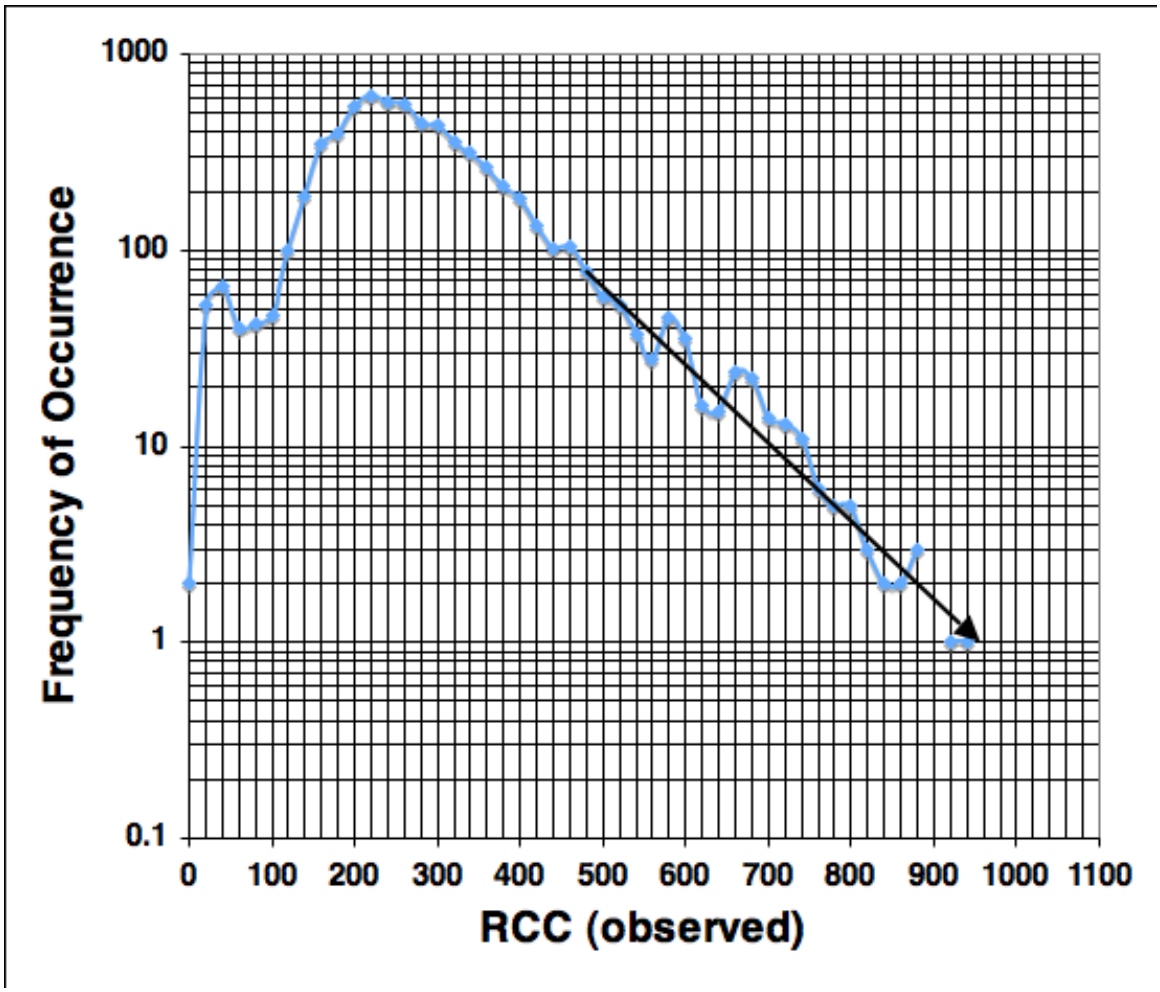
The progenitors, hence the points of origin of those haplogroups or SNPs, can be determined by similar extrapolations that involve only the members of that particular haplogroup or SNP[17].

Second Approach: Using a Histogram of the Observed RCC Matrix

After generating the observed RCC matrix from the group of haplotypes, we form a histogram of the frequency of occurrence of values of RCC over various intervals of

RCC and plot the logarithm of the frequency of occurrence against the appropriate value of RCC. The high RCC end of this distribution, when extrapolated to the RCC axis at Log Frequency = 0 (or N=1) will give an estimate of the time of origin after it is corrected by the appropriate factor F (see Figure 5 of the Howard 2012 submission to JoGG, now under review). Figure 9 shows the result.

Figure 9: A Log Plot of Values in a Histogram of the Observed RCC Matrix of 115 Haplogroups and their Subclades. The RCC time scale is shown prior to correction.



The extrapolation of the sequence of points at times greater than RCC (observed) = 500 points to an RCC of 950 when the frequency of occurrence is equal to unity. The correction appropriate to 950 is about 3.12. So, we find the date of origin at 950 x 3.12 x 40.85 years, or 121,000 years. This date may be an overestimate since it refers to an extrapolation toward the high end of the RCC matrix.

Third Approach: Using the Value of RCC (max) in the RCC matrix

In the first two approaches we use the entire distribution of points. However, in the

RCC matrix there will be a value, RCC (max), that is an estimate of the largest value of RCC found in all the pairs of the sample[18]. In the observed RCC matrix, RCC (max) is 926. When corrected by F= 3.12, RCC (max) corresponds to 118,000 years. We note that the second and third approaches not only use the same data, but use two different features of the observed RCC matrix. This date may also be an overestimate since it refers to the largest value of RCC in the RCC matrix. Considering the errors involved, the general agreement among all three dates is not unexpected.

Some Relationships Among the Variables in the Different Approaches to Finding the Progenitor of a Large Sample of Testees

There are four approaches that can be used to estimate the date of origin of a large sample of haplotypes, viz., (1) the observed RCC of the maximum junction point on the phylogenetic tree; (2) the observed RCC that results from the extrapolation of the junction points back in time to a single progenitor; (3) the observed RCC resulting from the extrapolation of the points at the high end of a histogram of RCC values to a single progenitor; and (4) the value of the maximum observed RCC in the RCC matrix. Each approach is expected to yield the best results when the sample is large. We first investigated the relationships among these principal parameters.

We selected representative haplotypes of the eight sets of haplogroups and SNPs shown in Table 3 and derived the observed RCC value of each parameter.

Table 3: Values of RCC (observed) at the Principal Points in the Phylogenetic Tree, the RCC Histogram and the RCC Matrix. The number of haplotypes in each sample is given in Column 2.

	Number in Sample	Maximum Junction on Tree	Extrapolation of Junction Points	Extrapolation of High End of Histogram	RCC Max in Matrix
Haplogroup A	50	638	700	1290	980
Modal Haplotypes	115	610	800	960	926
Haplogroup I	238	440	518	728	673
Haplogroup Q1a	177	281	315	575	595
E1b1a	79	235	270	480	478
L21a (1st half)	502	178	197	350	351
L21b (2nd half)	674	175	179	275	234
M222	684	74	84	88	81.3

We next plotted pairs of each of the four principal points and determined the slopes, zero intercept and variance of each relationship. The results are presented in Table 4.

Table 4: The Relationships Among the Slopes, Zero RCC Intercepts and Variance of the Principal Points in Table 3.

(y) 1st Member of Pair	(x) 2nd Member of Pair	(a) Slope	(b) RCC, Zero Intercept	Variance
Tree Junction Extrapolation	Max Junction on Tree	1.227	-21	0.9798
Histogram Extrapolation	RCC Max in Matrix	1.1984	-54	0.9515
Histogram Extrapolation	Max Junction on Tree	1.8078	-1.3	0.9478
RCC Max in Matrix	Max Junction on Tree	1.4704	+56	0.9464
RCC Max in Matrix	Tree Junction Extrapolation	1.1699	+92	0.9208
Histogram Extrapolation	Tree Junction Extrapolation	1.403	+56	0.8774

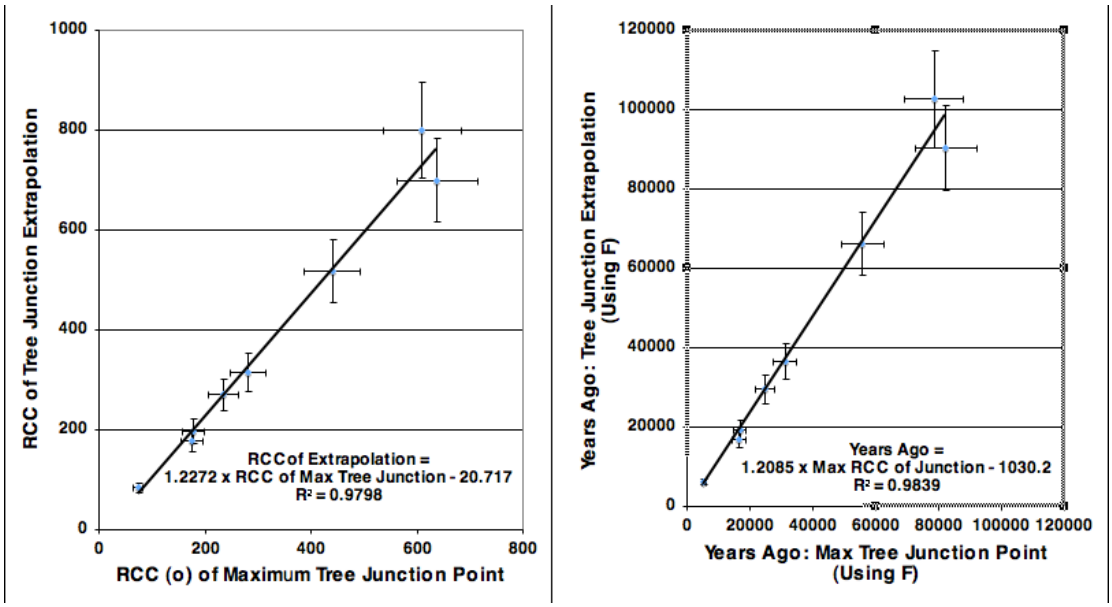
Note: The relationships are of the form: $y = ax + b$

The contents of Table 4 suggest the following conclusions:

- The highest correlation is found between the largest RCC junction point on the phylogenetic tree and the extrapolation of all the tree junction points to the RCC point where the progenitor lived. Both depend on the phylogenetic tree.
- The next highest correlation is found between the value of RCC (max) in the RCC matrix and the RCC of the point where the high end of the histogram is extrapolated to a frequency of occurrence of one, where the progenitor lived. Both depend on the RCC matrix.
- The RCC of the zero intercept is an indicator of the errors involved in the correlations. Relative to their average RCC, they are of the order of 0.2-18 percent, averaging 10 percent.
- The zero intercept-to-average RCC ratio was less than 5 percent for the top tree-to-tree comparison.
- Tree-to-tree comparisons and matrix-to-matrix comparisons had higher correlations than other, mixed comparisons.

These observations strongly suggest that the averaging approach employed by our Mathematica program to make our phylogenetic trees are more useful in determining the time of the progenitor than any other combination of principal points we have studied. Figure 10 shows that the maximum junction point and the extrapolation of the junction points to the progenitor are very highly correlated.

Figure 10. The Relationship Between the Oldest Junction Point on the Phylogenetic Tree and the Extrapolation of all the Junction Points to the Time of the Progenitor for each of the Groups shown in Table 4.



The graph on the left shows the observed values and extrapolated values of RCC on the tree. The graph on the right shows the corresponding years derived after applying the factor F and the Years per RCC to the observed values of RCC (see Table 5). Not only does the averaging approach of Mathematica appear to be preferable to the other options, the right hand graph has a higher variance after the factor F has been applied, indicating that using F has reduced the error in the determination. Consequently, we shall focus our attention on the time scale derived from the junction points on the tree, recognizing that mutations may overlap on the tree and that we do not observe all mutations that occur, necessitating the need for the correction factor, F.

Deriving the Time of the Progenitor Using the Extrapolation of the Junction Points on the STR Y-DNA Phylogenetic Tree.

The details of the first line in Table 4 are given in Table 5, which shows in the first column the haplogroup or SNP designation of the different haplogroup/SNP pairs in Table 3. The remaining columns show the appropriate observed RCC, the corresponding F and the number of years ago for (1) the oldest paired tree junction and (2) the extrapolated junction point. The number of years ago is determined from the relation:

Time (Years) to the MRCA = 40.85*F*RCC (observed) Equation 2

The last column in Table 5 gives the derived time when the progenitor of the group in the first column lived.

Table 5:

Sampled Group	Max Junction on Tree				Junction Extrapolation		
			Years				Years
Haplogroup/SNP	RCC	F	Ago		RCC	F	Ago
Haplogroup A	638	3.159	82331		700	3.16	90360
Modal Haplotypes	610	3.1525	78556		800	3.14	102615
Haplogroup I	440	3.1	55719		518	3.13	66232
Haplogroup Q1a	281	2.72	31222		315	2.845	36609
E1b1a	235	2.575	24719		270	2.68	29559
L21a (1st half)	178	2.315	16833		197	2.4	19314
L21b (2nd half)	175	2.31	16514		179	2.32	16964
M222	74	1.76	5320		84	1.806	6197

Note (1): The SNP group L21 was too large to be represented on a single tree, so the sample was divided in half. The differences in the values of RCC for the two groups is consistent with the expected errors.

Note (2): Details of the extrapolation of the junction points for the sample of modal haplotypes shown in the phylogenetic tree in Figure 7 are shown in Figure 8. The extrapolation led to 102615 years ago (estimated SD: 10-15%).

Recalibration of the RCC Time Scale for Different Haplotype Lengths and for Dating Older Haplotype Sets.

Up to this point, we have been determining time scales using 37-marker haplotype strings. During these investigations, Fred Schwab extended our ability to form dated Y-DNA phylogenetic trees using haplotypes of different lengths and to form corresponding RCC matrices in which the entries in the matrix are listed in the same order as their position on the tree. We noticed that trees formed from 37 or more marker sets were very similar in form whereas trees derived with fewer than 37 markers became increasingly dissimilar as decreasing numbers of markers were used. This observation strongly suggests that 37 or more markers are the minimum number required for detailed Y-DNA analysis.

Inspection of trees derived from 37 or more markers showed that the value of 43.3 years per RCC was approximately correct, but when trees are produced from other than 37-marker sets, we need to re-determine the number of years that correspond to a unit change in RCC. To do this, we selected a large sample of 209, 111-marker testees and, from the same sample we selected subsets at marker lengths of 12, 25, 37, 67, and 111 markers. Since the testees in the subsets were identical, the number of years represented by their RCC values must be identical. If we assume that the value of 43.3 years per RCC is correct at 37-markers where the pedigree calibration was made, then the conversion factors for results involving different marker lengths must be proportional to the average RCC value in each subset. To derive those average values, we used the RCC matrix for each subset and determined the average RCC throughout the 209 testees in each matrix. Table 6 shows for each marker length the average RCC derived from the 21736 RCC values in the matrix in Column 2, and the derived value of years per RCC in Column 3. The conversion factors in Column 3 can

be safely used for genealogical purposes, but, as discussed earlier, the corrected conversion factor in Column 4 should be used when genetic time scales are derived. Column 4 is the value of Column 3 divided by 1.06, the factor F for the average year in which the pedigree calibration was made.

Table 6:

Haplotype Length (No. Markers)	Average Matrix RCC	Derived Years/RCC	Corrected Years/RCC
12	181.5	33.6	31.73
25	202.2	30.2	28.49
37	141.0	43.3	40.85
67	151.4	40.3	38.05
111	166.2	36.7	34.65

Note (1): The standard deviations of the mean values in the second column were less than 0.75 (0.5%) for 37, 67, and 111 markers.

The results in Table 6 were derived using the first of the 12, 25, 37, 67, and 111 markers in the reporting sequence of FTDNA. If other marker sets are used, the results will be different because they will have different mutation rates. Intermediate values should not be derived using interpolation[19]. In estimating the number of years to the MRCA using haplotype lengths other than 37, the multiplier in the fourth column of Table 6 should be used in place of 40.85.

An additional uncertainty remains. The correction factor F was derived using the individual mutation rates for 37 markers derived by Chandler (2006). We have not derived F for marker lengths other than for 37. However, the RCC matrices used to derive Column 2 in Table 6 and the relative tightness of the conversion factors in Columns 3 and 4 show that using the same conversion factor for longer marker lengths will probably not result in errors in time due to differences in F, if they exist, by more than an estimated ~ 5-10%.

The results of Table 5 suggest the following observations and conclusions:

- The age sequence of the haplogroups and SNPs is in good agreement with the age sequence in the ISOGG and Eupedia.
- Haplogroup A is the oldest within this sample, but the group of modal haplotypes is still older, suggesting that the presence of a large variety of haplotypes or a larger sample of a single haplotype will lead to a larger age determination. The larger the set of junction points; the better the age determination will be.
- The L21 SNP is almost three times older than the M222 SNP.
- The derived age of the oldest progenitor in the study, often referred to as Y-Adam, is of the order of 100 Kyrs, but the intermediate ages from M222

through Haplogroup I are older than expected.

The process of junction point extrapolation and applying the correction factor has lead to somewhat older ages for the intermediate set of haplogroups and subclades than are given in the ISOGG. This new, different approach to age determination may be superior to older methods, but it needs further investigation. Other researchers determine ages by investigating variances. In our work we have found that the standard deviation (SD) of RCC divided by its average is approximately constant throughout most of the times of genetic interest in this study. This indicates that using SD, which is the square root of the variance, may not be the best approach to use for age determination[20].

The average RCC of the members of a group is very dependent on the samples that are chosen, many of which are often in family or surname clusters. This overweighting by members who have TMRCAs within the past few thousand years will cause the average of a group of haplotypes to be too small. Having identical haplotypes in the sample will lead to an age determination that is smaller than it should be. On the other hand, using outliers in the tree may lead to age determinations that are larger than they should be. By taking large samples and investigating their junction points on the phylogenetic tree, we can select an approach that minimizes cluster biases and uses the averaging approach of Mathematica to optimize the RCC values (viz., the time relationships) among all the sample members, not just those who may be more closely related.

Although the correlation approach can be used with any set of haplotypes, we need to be sure that all members belong to the particular group we choose to analyze. That is, to determine the TMRCA of a group, all samples must be members of that group. Correlation techniques can be applied to surname groups, intercluster groups, SNPs, haplogroups and subgroups, and better age determinations will be made from larger samples.

Complications Caused by Outliers and Identical Haplotypes in the Sample.

There are occasions when outliers appear on the phylogenetic tree. An outlier is a single testees whose junction point with the rest of the sample does not connect with others until much further back in time. It is important to decide whether to keep or discard an outlier. Inclusion of an outlier may indicate a line that has not died out, a rare testing error, an adoption or a non-paternal event, or a faulty decision to include it in the sample.

Whether an outlier should be included depends on the goal of the study. If the goal is only to define cluster memberships, then outliers are relatively unimportant. If the goal is to date clusters, interclusters or the sample, it is quite another matter since outliers become important to consider when sample dating is the goal. Time determinations within large clusters is more trustworthy when the outliers are not a part of the determination, but when the TMRCA of the whole group is the goal, the

reasons for including or excluding the outliers has to be carefully considered. Both the goal of the study and the rationale for including the samples to be studied are very important factors that should be considered in advance of any analysis. When larger numbers are included in the sample, better ages will be derived and the role of outliers will become better defined and understood.

One particular bias that can occur, particularly in surname samples, is the inclusion of numbers of identical haplotypes. To determine their effects on the rest of the tree we can collapse identical haplotypes to a single testee and compare the resulting tree with a tree that contains the identical haplotypes. Experience so far suggests that including identical haplotypes will not change dating determinations significantly when dealing with large samples and determining dates farther back in time, but they will affect the TMRCAs of individual clusters.

Summary:

This analysis spans a time interval that is more than 100 times the time interval within which pedigrees can be used in conjunction with RCC analysis. The model-derived values of RCC over this time interval is linear and the percentage of SD/RCC in both models is relatively constant at about 43 percent. Once a correction is made to account for backward mutations that are not observed, this correlation approach to the analysis of Y-DNA haplotypes offers a set of very powerful tools to explore time and evolutionary differences among the haplotypes of testees that are far back in time or that are in very different haplogroups. The results presented here agree with ages for the oldest Haplogroup A and a set of modal haplotypes of haplogroups, but this study predicts ages for intermediate clades and subclades that are somewhat higher than other researchers derive. We suggest that extrapolations of junction points on highly populated STR phylogenetic trees to where the progenitor lived is preferable to the use of variance to determine this age because genetic distance tends to saturate when applied to genetic time scales and the RCC correlation approach does not. This investigation suggests quantitative values for the errors, uncertainties and probabilities associated with the RCC correlation technique. We have shown that the distribution of markers on paired haplotypes can be explained by Poisson statistics, indicating that mutations do take place randomly.

This method of unifying the time scale of genealogy and genetics across a wide range of dates offers a quick and convenient way to derive times instead of using separate and often different approaches to date haplogroups or SNPs, etc. The recognition that only one RCC correlation approach, with its appropriate uncertainties, can be used for dating and analysis introduces a degree of simplicity to the analysis of a very complex, mutation-driven problem.

Acknowledgements:

Without the assistance and programming skill of both Fredric R. Schwab and Sidney A. Sachs the analysis that proved the linearity of the observed RCC time scale, the

results of all applications of Mathematica, and the corrections needed to convert RCC (observed) to a corrected value would not have been possible. My discussions with Sidney Sachs about the use of Poisson statistics in this analysis have been invaluable. We thank Eileen Krause Murphy for her explanation of the differences between an ISOGG (SNP-driven) phylogenetic tree and our RCC-dated phylogenetic trees that are based on STR sample haplotypes. Finally, we thank members of the Fairfax County (VA) Genetic Genealogy Special Interest Group, especially Jim Logan and Sidney Sachs, for discussions about SNP- and STR-oriented phylogenetic trees.

Appendix A: The Results of the Models That Led to Figures 8, 9, and 10 (Note: The numbers in this table are consistent with averages of 157 years per mutation, 28.583 years per generation, and 4085 years per 1000 RCC (corrected). They apply only to the use of the first 37 FTDNA markers.

RCC (observed)	F	Years		RCC (observed)	F	Years
0	1	0		520	3.13	66487
20	1.25	1021		540	3.136	69177
40	1.45	2369		560	3.14	71831
60	1.6	3922		580	3.145	74514
80	1.78	5817		600	3.15	77207
100	1.91	7802		620	3.155	79907
110	1.99	8942		640	3.16	82615
120	2.055	10074		660	3.16	85197
140	2.19	12525		680	3.16	87778
160	2.265	14804		700	3.16	90360
180	2.32	17059		720	3.16	92942
200	2.41	19690		740	3.154	95342
220	2.5	22468		760	3.15	97795
240	2.6	25490		780	3.144	100177
260	2.65	28146		800	3.14	102615
280	2.71	30997		820	3.13	104846
300	2.8	34314		840	3.118	106991
320	2.86	37386		860	3.105	109082
340	2.94	40834		880	3.092	111151
360	2.97	43677		900	3.08	113236
380	3.01	46724		920	3.07	115377
400	3.04	49674		940	3.06	117501
420	3.07	52672		960	3.045	119413
440	3.1	55719		980	3.03	121300
460	3.11	58440		1000	3.01	122959
480	3.12	61177		1020	2.985	124376
500	3.12	63726		1040	2.97	126177

Notes to Appendix A:
Columns 1 and 4 show RCC (observed). Columns 2 and 5 lists the correction factor, F. Columns 3 and 6 show the number of years in the past prior to 1945.

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, *A Comparative Analysis of the RCC Correlation and FamilyTreeDNA TiP™ Probability Approaches for Estimating the Time to the Most Recent Common Ancestor of a Pair of Y-DNA Haplotypes* (submitted to the Journal of Genetic Genealogy on 22 April 2013)). See: https://dl.dropbox.com/u/59120192/Genealogy/Papers/TipPaper.pdf

King, Turi E. and Jobling, Mark A., *Founders, Drift, and Infidelity: The Relationship between Y Chromosome Diversity and Patrilineal Surnames,* Mol Biol Evol. 2009 May; 26(5): 1093–1102).

Ma, Jian, Ratan, Aakrosh, Raney, Brian J., Suh, Bernard B., Miller, Webb, and Haussler, David, *The Infinite Sites Model of Genome Evolution*, Proceedings of the US National Academy of Sciences, 105, No. 38, 14254-14261, 2008.

AUTHOR POINT OF CONTACT:

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

END NOTES:

[1] 10 RCC ~ 433 years.

[2] Family Tree DNA (FTDNA) has listed the results of Y-DNA tests, giving public available Kit Numbers, haplogroup determinations and individual DYS marker values. This testing agency is located at 1445 North Loop West, Suite 820, Houston, Texas 77008 on their web sites.

[3] The average mutation rates over 37 markers derived by: (1) Howard (2009, Table 3; 0.00728 mutations per year); (2) Chandler (2009; 0.00492), and (3) Kerchner http://www.kerchner.com/dnamutationrates.htm; (0.0057); with the values we derived in this study (0.0064-0.0072), we adopted a value of 0.00637 mutations per year over 37 markers, which is equivalent to 157 years per mutation or 3.63 RCC per mutation. The estimated standard deviation of this result is of the order of 15 percent. In a separate, related study of 76 haplotypes of closely related fathers, sons and brothers, it was found that over the 2812 DYS sites the observed and predicted results agreed with a Poisson distribution to within 0.6 percent and that the Chandler average mutation rate was confirmed to within 8.4 percent.

[4] These three graphs, derived from model runs, illustrate departures from linearity that might be expected if RCC is used over long intervals of time, and should not be used to estimate the error associated with any one particular value of RCC. Due to the effects of back mutations, the observed value of RCC must be corrected to determine the true time scale.

[5] For specific descriptions of these time intervals, see Howard (2009).

[6] A statistical study has been made of the four components of DYS 464 to see how permuting them around would affect the correlation. It does affect it, but only to the extent of adding a very few RCCs of uncertainty to the result. That uncertainty is well within the error interval that might be expected from normal, random mutations elsewhere in the haplotype. That very jumpiness is valuable to Y-DNA oriented genealogists when they attempt to group testees into clusters using the rapidly varying markers like DYS 464 and CDY since they vary rapidly and add resolution that is valuable in separating the differences among family groups. Use of the 464 and the CDY markers also add valuable time resolution when the RCC time scale has been calibrated through the use of over 100 pedigrees whose TMRCAs were known. There is a difference between calculating intraclade and interclade ages. But here again, by choosing a combination of fast and slow moving markers, with a nice transition in mutation rates between the extremes, the geneticists have done us a favor. Correlation accents the differences in the near term markers because of their higher rates of mutation. The transition markers become more and more important as we study the transition from surname clusters, to the interclusters, and finally to the different haplogroups. As we transition from genealogical to genetic time intervals, the faster moving marker values average out and the effects of the slower moving markers become important to the correlation. This weighting is important to the RCC process and we would lose resolution if we were to ignore fast moving markers like the 464 and CDY groups in the RCC analysis. While genetic distance is easier to use, it does not yield the resolution that the correlation process does.

[7] Four percent is the standard deviation of the ~ 43% standard deviation.

[8] In a separate paper (Howard 2013) it is shown that the sum of the absolute values of changed markers (m) used in the RCC correlation approach is to be preferred to the use of the number of DYS marker sites (n) that have changed because n and genetic distance, become saturated as the time to the most recent common ancestor increases, and m does not.

[9] The Y chromosome contains two types of ancestral markers: The 37 marker values of individual haplotypes referred to in this paper are Short Tandem Repeats (STRs), the first type, that are most often used to trace recent ancestry. In this section we are investigating their use to trace genetic ancestry by investigating the errors in dating groups of haplotypes, called haplogroups. The second type of marker, the SNP (Single Nucleotide Polymorphism), consists of large numbers of haplotypes that share a small, rare genetic change, or variation, that can occur within a person's DNA sequence. The SNP is a shared, common characteristic that is increasingly used to define a haplogroup or a subhaplogroup. Y-SNP markers are used to sort human Y chromosomes into the various haplogroups. SNPs change on average at a rate of about one mutation every few hundred generations. The most recent common male ancestor of two people who share the same Y-SNP test haplogroup may have lived tens of thousands of years ago.

[10] In the paper by King et al (2009), twenty, 32-year generations correspond to 640 years, or an RCC ~ 14.8. Based on this average survival rate at the 20th generation, about 11% of a male line will not persist from one generation to the next.

[11] RecLOH stands for Recombinational Loss of Heterozygosity. In DNA it occurs when recombination takes place in a pair of slightly different genes, producing a pair of identical genes. The genetic code exchange between the chromosomes is not reciprocal and genetic information is lost in the process.

[12] In the final sections of this paper we will show that the observed values of RCC require a correction factor F (F>1) that must be used to correct for mutations that are occurring but we do not see. We will refer to that value as RCC(corrected). Over time intervals of interest to genealogists, the differences between the observed and corrected values of RCC are dwarfed by the extent of random mutations, but to avoid systematic errors, the correction factor should be applied.

[13] This inconsistency in dates derived through correlations of haplotypes was particularly noticeable when junctions of the TMRCAs of different haplogroups on the phylogenetic tree were compared to the times expected from the ISOGG sequence.

[14] Ma et. al (2008) explores the effect of mutations (speciation, deletion, insertion, duplication, and rearrangement of segments of bases) and reconstructs the history of the X chromosome in human, chimp, macaque, mouse, rat, and dog over time scales

longer than those in this paper. See also http://www.pnas.org/content/105/38/14254.full

[15] After increasing from F=1 to F= 3.2, the correction factor reaches a maximum and begins to decline as more and more markers become uncorrelated with the markers of earlier progenitors. The turnover happens between observed RCCs of 600-800, or 80,000 and 110,000 years ago.

[16] The International Society of Genetic Genealogy (ISOGG) is a non-commercial, non-profit organization in which a committee of genealogists and geneticists are continuing to study the evolution of haplogroups based on their analysis of SNPs and STDs. Their phylogenetic tree is frequently updated and can be found as part of their website at http://www.isogg.org. The Eupedia site gives descriptions of the origins, age, spread, and ethnic association of selected haplogroups and subclades. It can be found at http://www.eupedia.com/europe/origins_haplogroups_europe.shtml

[17] Within a phylogenetic tree there often appears two or more distinct, well-separated subclusters which are joined at an earlier date. In this case, the RCC of the progenitor of the pair of subclusters can be derived by inspection of that junction point on the tree. Since the two subclusters are independent of each other, we can derive the progenitor of each subcluster using the same extrapolation process we use to derive the progenitor of all the haplotypes in the sample. In the cases we studied, the date of the progenitor of each of the pair of subclusters agrees with the RCC of their junction point to within the errors expected. This date will be the time when a mutation occurs at the progenitor of each subcluster, causing the split at the junction point that results in the two separate lines of descent.

[18] There are arguments for and against the use of the second and third approaches, which are included here only to show that the results of their use gives answers relatively close to the first approach, which is simpler in concept and more straightforward. The histogram of the RCC matrix is not Gaussian (mean=272; mode=259; skewness=1.13; kurtosis=2.42). The mode is lower than the mean because the TMRCAs of recent haplogroups have low values of RCC (see the spike at RCC~ 40 in Figure 12). The second and third approaches may still yield a reasonable, albeit less reliable, estimate unless there has been a (rare) testing error.

[19] For example, if the last 44 markers are used (sites 68 to 111), the average RCC was 177 (SD ~ 0.5%)

[20] Howard (2013)