Journal: www.jogg.info
Originally Published: Volume 5, Number 2 (Fall 2009)
Reference Number: 52.013

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: THE USE OF CORRELATION TECHNIQUES FOR THE ANALYSIS OF PAIRS OF Y-CHROMOSOME DNA HAPLOTYPES, PART I: RATIONALE, METHODOLOGY, AND GENEALOGY TIME SCALE

Author(s): William E. Howard

The Use of Correlation Techniques for the Analysis of Pairs of Y-STR Haplotypes, Part 1: Rationale, Methodology and Genealogy Time Scale

William E. Howard III

Abstract

A new method for analyzing Y-chromosome haplotypes is presented that uses correlation techniques to reduce differences in pairs of haplotypes to a single number, which we call the RCC. Used in conjunction with traditional ways to analyze haplotypes, this technique can add new dimensions to the interpretation of Y-DNA results. By using pedigrees to divide the time to the most recent common ancestor of any pair of testees by its corresponding RCC, we can calibrate RCC as a time indicator. The RCC for a set of individual haplotypes can be presented in a two-dimensional matrix where the value of RCC for each pair of testees appears at the intersections of the appropriate rows and columns. When testees who have the lowest values of RCC are grouped together in the matrix, they are seen to form clusters that are virtually identical to the testee groupings that surname administrators make from haplotypes using traditional methods. We can estimate the time to the common ancestor of cluster members, the most recent common ancestor (TMRCA) of all members of each cluster and from the matrix intersection of pairs in different clusters (the intercluster region), we can estimate the TMRCA of any pair of clusters. Analysis shows that the RCC time scale can be used over tens of thousands of years without significant evidence of non-linearity, making it also potentially useful for tying together the time frames of events involving mitochondrial DNA, migration patterns, linguistic patterns, geology, anthropology, paleontology and archeology. The pros and cons of using traditional techniques versus the RCC correlation technique are presented.

Introduction

This study presents a new correlation method for organizing Y-chromosome haplotypes and calculating the time to the most recent common ancestor (TMRCA). We suggest that the technique be used in conjunction with traditional methods of analysis. It is simple, straightforward, reproducible and non-proprietary. It presents an easily available adjunct to proprietary tools now in use. Moreover, it utilizes an easily accessible software program, Excel, that permits the analysis to be done quickly using small personal computers. The technique produces matched pairs of Y-DNA testees from which groups of people who are more closely related can be determined. It can be applied to any pair of haplotypes, from closely related testees in surname groups to haplotypes in remotely related haplogroups.

The process correlates haplotypes--long strings of numbers (called alleles or marker values) from each pair of testees. It reduces each pair of strings to a single number (RCC). It can be applied simultaneously to pairs of very large numbers of testees. The only restriction is that for each run, the same number of markers, in the same order, must be used in the analysis.

This study demonstrates that RCC is a time indicator, and that it indicates an approximate time to the most recent common ancestor (MRCA) of the pair of testees. Validated pedigrees are used to calibrate the RCC time scale. The time scale is corroborated by models and related analytic studies. The analysis suggests time scales over which all testees may have been more closely associated further back in time. By reducing two strings of haplotypes to a single number we ignore the individual marker numbers that have traditionally been used to make associations. But this approach can provide quick checks on those associations, and it can suggest other members of a group. It can be used to decide whether a more targeted genealogical pedigree might help determine the MRCA between two testees -- something that the traditional approach cannot as easily do.

Address for correspondence: William E. Howard, wehoward@post.harvard.edu

Received: June 6, 2008; accepted: August 26, 2009

We make no claim that the calculation of RCC between pairs of testees will result in significantly better matches within groups of testees, but it extends the analysis by investigating the time to the most recent common ancestor (TMRCA) of any pair of testees.

At the conclusion of this paper, the pros and cons of both the traditional and correlation methods will be summarized. We recommend that both methods be used together to obtain a greater degree of analytic insight than either approach can yield alone.

This correlation approach also gives insight into relationships that may have occurred during time periods beyond which pedigrees and genealogical information are either unavailable or cannot be used. The RCC time scale can be shown to apply back in time to epochs at which separation between sub-haplogroups or haplogroups occurred. The correlation approach may provide a means to tie together the time scales of mitochondrial DNA, migration patterns, linguistic patterns, geology, anthropology, archeology, and paleontology.

In Part 2 of this two-part series of articles, we will show the application of the correlation technique to investigate the construction and dating of surname groups, to set a time frame for the common ancestor of clusters, and to explore the dates of origin and evolution times of haplotype groups.

Rationale for Introducing a New Analytic Approach for Y-Chromosome Analysis

The analysis of Y-chromosome haplotypes is still very young. We must continue to look for quick, simple methods to group haplotypes and to determine the TMRCA -- exploring different methods that can be used together to achieve more meaningful results.

The traditional process of analyzing testee results often involves minimizing the sums of the arithmetic or absolute marker differences among the testees. There is no general consensus about how to treat marker differences. Moreover, different analysts do not always group testee results using the same criteria; the concept of an optimum grouping is not defined.

A correlation analysis provides a rapid, reproducible, and easily understood way to make initial groupings or to validate them. It is simple because it reduces each pair of haplotype sequences to a single number. It treats the marker differences automatically, without the need for human decision-making. It reduces the problem of individual mutation rates to a calibration problem using pedigrees and other indicators. The power and flexibility of the technique allows comparisons to be made that would be much more difficult if very many long strings of haplotypes must be grouped together by inspection. Some companies that process DNA suggest probabilities that the most recent common ancestor (MRCA) will be located a specific number of generations ago. The techniques they use are proprietary and details of the method are not easily available. The approach presented here is successful at determining TMRCAs by using pedigrees and correlation techniques. The errors in determining the time to the MRCA are still the result of random mutations. They are comparable to the errors that are inherent in more traditional matching techniques, both of which may be quite large.

Methods--Part I: Forming the RCC Matrix

Here is the approach we use:

- Assemble the haplotypes of individual testees in a spreadsheet (e.g., Excel).

- Separate them into groups that have the same markers and numbers of markers tested. Our approach uses results that consist of at least 37 tested markers in a haplotype string. Results from 67 marker strings can also be used; they are virtually identical to the results using 37 markers. We use the 37-marker set of FamilyTreeDNA because of the larger set of pairs who have been tested.

- Determine a correlation coefficient between the marker strings of each pair of testees. The Microsoft Excel data analysis tool kit does this with ease. The result is presented as a one-sided matrix (Note 1).

- Cut and paste the one-sided matrix using the transpose feature of Excel to form a transposed one-sided matrix. This intermediate step is needed in order to form a two-sided matrix, which will have many uses.

- Cut and paste the second matrix and use an algorithm to produce a third matrix that is two-sided. This matrix contains correlation coefficients (CC) that vary from unity downward in value. They are awkward numbers like 0.9995, which may be simplified for convenience as described next.

- Simplify (scale) the result in each row and column by taking the reciprocal of the number, subtract unity from it and multiply the result by 10,000. This is the Revised Correlation Coefficient, called RCC. Thus a CC of 0.995 becomes an RCC of 50. We find values of RCC in this analysis vary from 0 about 1000—much easier to analyze. In this conversion, the number of significant figures and the linearity of the scales before and after the conversion are not affected.

- In the original matrix, pairs of testees who have results near unity are more closely related. When the result is presented as RCC, pairs of testees who have results near zero are more closely related. They have a TMRCA nearer in time than the others.

Methods - Part II: Organizing the Matrix

Grouping the RCC Matrix into Clusters by Using Time Slices of RCC Values

If a surname administrator has not grouped the haplotype results, we must group them into clusters. **Appendix A** shows how to develop an algorithm that will produce a time slice of the matrix and **Appendix B** shows how to use the time slicing algorithm to form clusters. Once clusters have been formed we can use both the RCC time scale and the time slice algorithm to investigate how clusters evolve and find the approximate epoch at which the ancestors of cluster members lived.

Testing Different Numbers of Markers: Influence on RCC Results and Uncertainties

We settled on a 37-marker analysis for this study because the product of the number of people tested and the numbers of markers tested was greatest at 37 markers. A number of 37-marker testees had been tested at 67 markers. We compared the RCC results for these testees, using 12, 25, 37 and 67 markers. No statistically significant change was found in the average results of the samples but the uncertainty in that average value increased markedly when less than 37 markers were used.

The results of an investigation of how the quantization of markers affects the determination of RCC is given in Note 2. RCC differences of the order of 3 result from one marker change when 37 markers are tested. If this RCC difference of 3 (genetic distance of +/-1) is used as the uncertainty, then this uncertainty corresponds approximately to 130 years (see **Table 1**). While quantization errors of this kind cause large percentage errors in the results for recent time periods, their effects become proportionally smaller as longer time periods are considered. The comparison with time scale and marker differences are based on the work of Walsh (2001).

In a separate study we took 69 participants in the Hamilton surname project, all of whom had been tested at 67 markers and compared their derived values of RCC, pair by pair, with the same pairs at the 37, 25, and 12 marker level. If we use the 67 marker value as a reference (1.0), we found that the median RCC marker difference (RCC derived at 67 markers minus RCC derived at 37 markers) between the 67 marker value and the median values at 37, 25, and 12 markers varied from about +2 (Standard Deviation [SD]≅5), +5 (SD≅14), and +8 (SD≅15), respectively. While this may be an indication of a small systematic error between the RCC values derived from pairs of testees, it shows that any scale error is small between the 37-marker level and the 67-marker level. Thus, RCC values derived with 67 markers should fall well within the other errors inherent in the RCC derivation. The SDs of the medians for 25 and 12 markers are a factor of about three larger than the SD of the medians for 37- and 67-marker haplotypes. This result shows why the testing agencies and surname project administrators urge that at least 37 markers be tested (See Note 3).

The Interpretation of the RCC Matrix

The RCC matrix is derived from the 37-marker values displayed in columns and the individual testees in rows. If the testees have already been sorted into groups, the matrix analysis is straightforward. If not, the matrix needs to be sorted by placing groups of testees together when they share low values of RCC. Sorting should be done simultaneously on rows and columns. **Appendices A and B** give methods for sorting.

The most common type of RCC matrix is one from a surname project, derived from haplotypes of individuals with the same surname. Clusters of testees will appear in the overall project matrix. Different clusters will contain different groups of testees although everyone in the matrix might share the same surname.

Within a cluster, each pair of testees will have a MRCA that will differ from other MRCAs of other pairs of members. The members in a cluster will all share a common ancestor (CA) who will have been born at an earlier time than most or all of the individual MRCAs of the various cluster pairs. Similarly, pairs of testees who are members of a different cluster will also have their individual MRCAs. They will share a CA other than the one in the first cluster. The two CAs will, in turn, have an older CA, thus starting a hierarchy of CAs reaching back in time. Insight into this hierarchy can be seen in the area of intersection between the members of any two clusters. The entries in this intercluster region are composed of the RCCs of pairs where one member of the pair belongs to one cluster and the other member of the pair belongs to the other cluster. The distribution of RCCs in this intersecting region will correspond to the single MRCA of the two clusters and the average of the RCC set will indicate the epoch when the CA of the two clusters lived. Different clusters will have different hierarchical CAs, back in time. A more detailed description of how the CAs are determined can be found in Sections 1 and 2 below.

This procedure can be extended to individual testees in clades, in subhaplogroups and haplogroups; it can be used to place haplotypes into an evolutionary sequence.

If we were to fill an RCC matrix using a random group of haplotypes, we would find values of RCC that ranged from zero to the maximum RCC we have found for 37-marker haplotypes, about 1200 (see Note 4). The

Table 1
Relationships Among RCC, the Corresponding Time Scale, Date in the Past, Genealogy Match and Approximate Marker Difference. A full description leading to Table 1 may be found later in this article.

RCC	Years (Note 1)	Date in Past (Note 2)	Genealogy Match	Approximate Marker Difference (Note 3)
0	0	CE 1945	Exact	0
1	43	CE 1902	Very Tightly Related	0.3
2	87	CE 1858	Very Tightly Related	0.7
3	130	CE 1815	Very Tightly Related	1
4	173	CE 1772	Very Tightly Related	1.4
5	217	CE 1728	Very Tightly Related	1.7
6	260	CE 1685	Very Tightly Related	2.1
7	303	CE 1642	Very Tightly Related	2.4
8	346	CE 1599	Tightly Related	2.7
9	390	CE 1555	Tightly Related	3.1
10	433	CE 1512	Tightly Related	3.4
12	520	CE 1425	Tightly Related	4
14	606	CE 1339	Related	4.7
16	693	CE 1252	Related	5.3
18	779	CE 1166	Probably Related	5.9
20	866	CE 1079	Probably Related	6.6
25	1083	CE 862	Probably Related	8.1
30	1299	CE 646	Possibly Related	9.5
35	1516	CE 430	Possibly Related	11
40	1732	CE 213	Probably Not Related	12.3
45	1949	3 BCE	Probably Not Related	13.7
50	2165	220 BCE	Probably Not Related	15
55	2382	436 BCE	Probably Not Related	16.2
60	2598	653 BCE	Too Distantly Related	17.4

Notes:
(1) Derived from pedigrees of three surname groups where MRCAs are known.
(2) Derived assuming the average birth year of testees is 1945.
(3) See the analysis by Walsh (2001). Values here are for comparisons only.

lower values of RCC would point to close relatives while higher values of RCC would include pairs whose TMRCA would be located further into the distant past.

Experience has shown that the groups and cluster associations correspond approximately to the intervals of RCC shown in **Table 2**.

Using the RCC Matrix

Having set up the RCC matrix there are several uses to which it can be put. They include (1) providing a quick check on pair associations made by surname administrators and identifying pairs of testees who have been missed; (2) using the matrix to assign relationships among pairs of testees; (3) doing fast sorting of matrix entries to investigate relationships among matrix pairs within chosen intervals of time; (4) making a histogram of portions of the matrix to show how groups of testees are related; (5) investigating the time span over which the testee relationships are distributed; (6) investigating the individual MRCAs within surname clusters; and (7) deriving the evolutionary time sequences of groups within the matrix.

Figure 1 shows a partial view of a Logan surname matrix (Logan, 2008). Testee identifications, all of whom belong to Haplogroup R1b, appear in the top row and left column. RCC values of the pairs appear in the matrix. Logan RCC values range between zero and 75. Two surname clusters are apparent, marked A and B. They contain lower values of RCC indicating that their members have shorter TMRCAs than pairs who appear outside the cluster boundaries. Each entire cluster also has a time to the common ancestor (TCA) that will be less than the TMRCA of pairs who appear

Table 2
Representative Values of RCC for Various Associations of Paired Testees

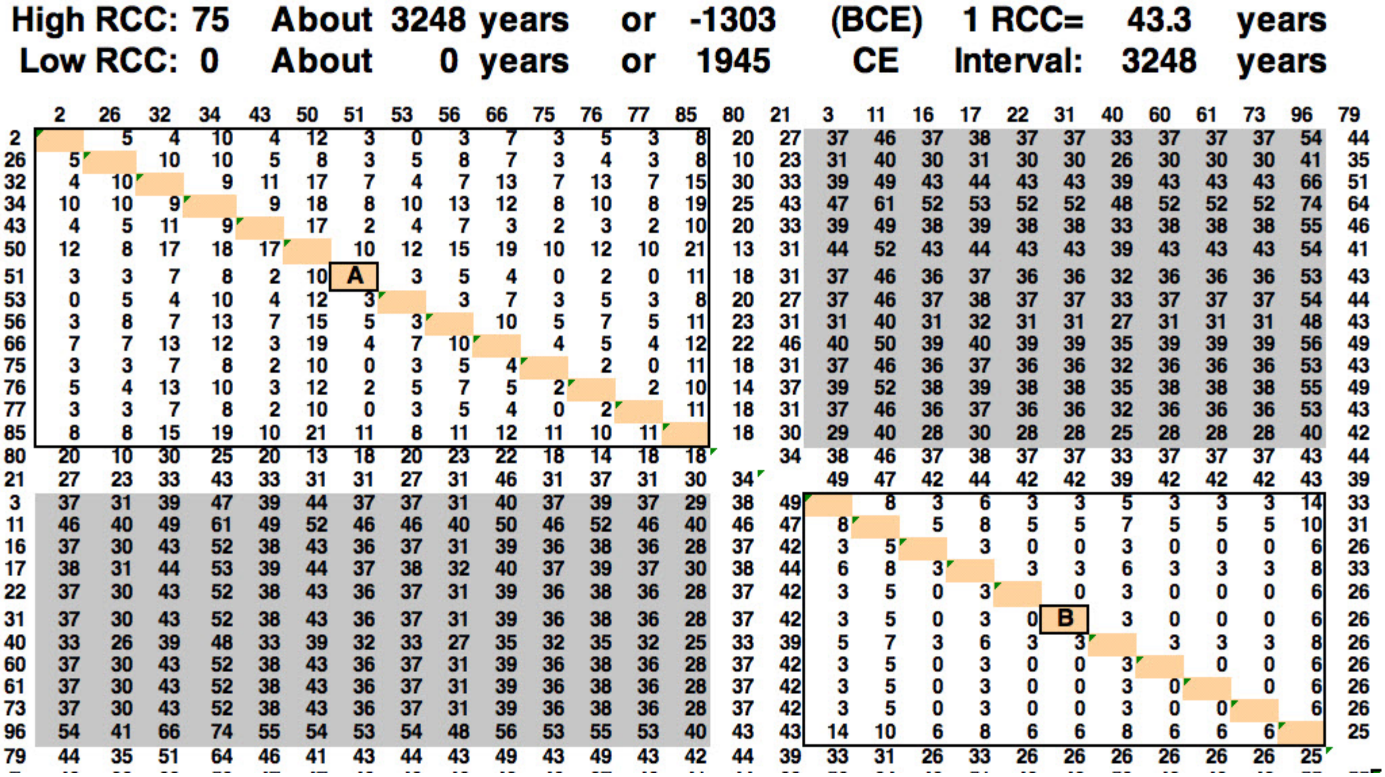
Group and Cluster Associations	Approximate RCC Interval
All human males	Up to 1000
Haplogroups	Mid to High 100s
Sub haplogroups	Low 100s
Clade groupings	Less than 100
Interclusters (pairs in different clusters)	25- 100
Clusters (e.g., surname groups)	0-25
Lines back to earliest pedigrees	0-15
Close relatives (e.g., Father-son-uncle)	0-5
Identical Twins	0

outside the cluster. The intercluster region contains pairs of testees, one of whom is in Cluster A and the other is in Cluster B. Thus, Intercluster AB consists of all RCC values that appear in the gray areas of **Figure 1**. An analysis of the intercluster region will indicate the approximate time when the common ancestor of Clusters A and B lived. A comparison of the distribution of RCCs in Clusters A and B show that Cluster B is younger (lower average RCCs) and both clusters are younger than the average RCCs in the Intercluster AB region; the latter have higher values of RCC. Both clusters were formed relatively recently; the intercluster region indicates TCA for the two clusters of about 3250 years ago.

Methods - Part III: Calibrating Values of RCC as a Time Indicator

The application of time scales that result from this analysis may be divided into three parts:

A. Time scales of genealogical interest -- RCCs in the range from 0 to about 25, especially those that are from 0-10. We will calibrate this interval as a time sequence using pedigrees, and we will test the consistency of the result.



High RCC: 75 About 3248 years or -1303 (BCE) 1 RCC= 43.3 years
Low RCC: 0 About 0 years or 1945 CE Interval: 3248 years

	2	26	32	34	43	50	51	53	56	66	75	76	77	85	80	21	3	11	16	17	22	31	40	60	61	73	96	79
2		5	4	10	4	12	3	0	3	7	3	5	3	8	20	27	37	46	37	38	37	37	33	37	37	37	54	44
26	5		10	10	5	8	3	5	8	7	3	4	3	8	10	23	31	40	30	31	30	30	26	30	30	30	41	35
32	4	10		9	11	17	7	4	7	13	7	13	7	15	30	33	39	49	43	44	43	43	39	43	43	43	66	51
34	10	10	9		9	18	8	10	13	12	8	10	8	19	25	43	47	61	52	53	52	52	48	52	52	52	74	64
43	4	5	11	9		17	2	4	7	3	2	3	2	10	20	33	39	49	38	39	38	38	33	38	38	38	55	46
50	12	8	17	18	17		10	12	15	19	10	12	10	21	13	31	44	52	43	44	43	43	39	43	43	43	54	41
51	3	3	7	8	2	10	A	3	5	4	0	2	0	11	18	31	37	46	36	37	36	36	32	36	36	36	53	43
53	0	5	4	10	4	12	3		3	7	3	5	3	8	20	27	37	46	37	38	37	37	33	37	37	37	54	44
56	3	8	7	13	7	15	5	3		10	5	7	5	11	23	31	31	40	31	32	31	31	27	31	31	31	48	43
66	7	7	13	12	3	19	4	7	10		4	5	4	12	22	46	40	50	39	40	39	39	35	39	39	39	56	49
75	3	3	7	8	2	10	0	3	5	4		2	0	11	18	31	37	46	36	37	36	36	32	36	36	36	53	43
76	5	4	13	10	3	12	2	5	7	5	2		2	10	14	37	39	52	38	39	38	38	35	38	38	38	55	49
77	3	3	7	8	2	10	0	3	5	4	0	2		11	18	31	37	46	36	37	36	36	32	36	36	36	53	43
85	8	8	15	19	10	21	11	8	11	12	11	10	11		18	30	29	40	28	30	28	28	25	28	28	28	40	42
80	20	10	30	25	20	13	18	20	23	22	18	14	18	18		34	38	46	37	38	37	37	33	37	37	37	43	44
21	27	23	33	43	33	31	31	27	31	46	31	37	31	30	34		49	47	42	44	42	42	39	42	42	42	43	39
3	37	31	39	47	39	44	37	37	31	40	37	39	37	29	38	49		8	3	6	3	3	5	3	3	3	14	33
11	46	40	49	61	49	52	46	46	40	50	46	52	46	40	46	47	8		5	8	5	5	7	5	5	5	10	31
16	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5		3	0	0	3	0	0	0	6	26
17	38	31	44	53	39	44	37	38	32	40	37	39	37	30	38	44	6	8	3		3	3	6	3	3	3	8	33
22	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5	0	3		0	3	0	0	0	6	26
31	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5	0	3	0	B	3	0	0	0	6	26
40	33	26	39	48	33	39	32	33	27	35	32	35	32	25	33	39	5	7	3	6	3	3		3	3	3	8	26
60	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5	0	3	0	0	3		0	0	6	26
61	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5	0	3	0	0	3	0		0	6	26
73	37	30	43	52	38	43	36	37	31	39	36	38	36	28	37	42	3	5	0	3	0	0	3	0	0		6	26
96	54	41	66	74	55	54	53	54	48	56	53	55	53	40	43	43	14	10	6	8	6	6	8	6	6	6		25
79	44	35	51	64	46	41	43	44	43	49	43	49	43	42	44	39	33	31	26	33	26	26	26	26	26	26	25	

Figure 1. A partial Logan surname matrix showing two clusters and the inter-cluster region

B. Time scales appropriate to surname groups and more distant haplotype pairs and haplogroups -- mainly RCCs in the range from 25-100. This range covers surnames, subhaplogroups and recently formed haplogroups.

C. Time scales in the RCC range from 100 to 1000 that are representative of more distant paired relationships. RCC results in his time interval may relate to studies of mitochondrial DNA, migration patterns, linguistic patterns, geology, anthropology, paleontology and archeology.

The present article will focus on the first of these three time ranges. In Part 2 of this two-part series of articles, we will investigate the deeper RCC ranges of the second and third of these time ranges.

Time scales of genealogical interest

1. Calibrate the RCC Time Scale Using Pedigrees

The best way to calibrate the RCC time scale is to use pedigrees of pairs of testees, each of whom can trace his ancestry to the same MRCA. Because Y-DNA testing is so new, and so few people with reliable, long-term pedigrees have been tested, it is difficult to find large numbers of pedigrees that meet these criteria. Nevertheless, the Athey, Ewing, Logan, and Hamilton surname projects have well-documented pedigrees and TMRCA pairs from which the RCC time scale can be determined. The TMRCAs and RCCs of the 363 pairs serve to calibrate the number of years that correspond to a unit change in RCC.

We note that the distribution of the ratios of TMRCA/RCC is not Gaussian. It is skewed toward high values of the ratio. Moreover, the SD of the histogram of the ratios is large and the kurtosis shows that the distribution is more peaked than a Gaussian. Averaging the ratios gives a significantly different result than summing the values of the TMRCAs and dividing by the sum of the RCCs. In cases like this a robust estimator of the number of years that corresponds to a unit change in RCC is given by using a statistical method called The Hodges-Lehmann estimator. This method is preferred in any situation where the degree of contamination (i.e., effects of mutations) and the type of distribution is not known with great precision (Hodges and Lehmann 1963; Saleh 1976). If the dataset contains n data points, it is possible to define n(n + 1) / 2 pairs within the data set, including the pairs formed by each item with itself. The average value is calculated for each pair and the final estimate is the median of the n(n + 1) / 2 averages. One advantage of using the Hodges-Lehmann estimator is that it minimizes the effect of the extreme values of TMRCA/RCC while still using them in the calculation. When the Hodges-Lehmann estimator is applied to these data, we find that 1 RCC = 43.3 years. We estimate that the SD for this determination is about 8%. This calibration for the RCC time scale will be checked for consistency by applying it to other available data.

2. Estimating the Time of the Common Ancestor of Cluster Members (TCA) when the Common Ancestor (CA) is Not Known.

While the main thrust of this article and the companion article is to use haplotype pairs in a new way to derive, calibrate and apply an RCC-derived time scale with an eye toward the evolution of clusters, interclusters, surname groups and haplogroups, it is important to investigate its applicability for determining the time of the common ancestor of all the members within a cluster.

Initially it was thought that the TCA of a surname cluster would be that point on a histogram of RCC values where RCC was near the maximum among all cluster pairs. However, in practice, it is difficult to choose that point, especially when the histogram is non-Gaussian and contains a long tail toward high RCC values. The possibility of the choice being biased by the presence of unrecognized pairwise mismatches further complicates using this approach to estimate the TCA.

Because of the importance of determining the TCA in the genealogy of surname groups, efforts have been directed toward investigating the following three approaches, all of which were applied to pedigrees or clusters with known CAs. But all have their unique uncertainties:

a. The determination and application of a genealogical structure factor (GSF), suggested by Athey (2009), that uses a pedigree, the TCA of the group, and the values of TMRCA and RCC among pairs of its members to tie that investigation to the structure of the pedigree. From the statistics of members of a cluster with an unknown CA, it was hoped that the GSF could be instrumental in determining the TCA. However, different groups of surnames have different pedigree structures, making the determination of an unknown TCA very difficult to predict. This approach was not pursued further because of this problem.

b. The identification of a point in the histogram of the RCC cluster matrix that would lead to the TCA. Virtually all histograms of the members of a cluster show pronounced skewness toward large values of RCC, often accompanied by a long tail of the distribution that has much noise. Nevertheless, it was hoped that by choosing an RCC at which the distribution first encountered base noise at the high side of the distribution, it would lead to the TCA.

c. The determination of a factor by which one or more statistical parameters of a cluster might be combined to indicate the TCA. This has been selected as the best approach to determining the TCA of a cluster.

To select the most important statistical parameters, we took the following approach:

- Select groups, each of which have a good combination of known TCA, a good pedigree and reasonable numbers of RCC values. The group Hamilton B, four Ewing groups, the M222 group and the Athey group met these criteria. They consisted of 273 pairs of RCC and 7 different known TCAs.

- Determine the most important parameters by looking for high correlations between the known TCA and the other statistical parameters of the groups. We looked at the average and median RCC, SD, the RCC (Point P) at which the downward slope in the histogram first encountered base noise on the high side of the distribution, the percent that the matrix was filled at the RCC of Point P, the percentage down from the peak at which the histogram encounters noise on its high side, and the skewness and the kurtosis of the distribution.

- Derive equations that make the most optimum TCA predictions from those parameters.

- Use those equations to estimate the unknown TCA in other clusters.

The parameters to be used are those that have a high correlation with the known TCAs of the calibration set, and the factors in the equations are those that minimize the difference between the computed and known TCAs among the seven groups. Three such high correlations were found. The best correlation involved the average RCC of the cluster members; next best was the correlation involving the Point P; then followed a correlation involving the SD of the cluster. We proceeded by fitting a linear relationship to the data for each of them. The results follow:

- The RCC of the CA = 1.285 times the average RCC of the cluster members.

- The RCC of the CA = 0.61 times the RCC value at Point P in the histogram of the cluster.

- The RCC of the CA = 2.356 times the SD of the cluster members.

The correlation coefficients of these three relations are 0.977, 0.927 and 0.899, respectively, indicating that they are reasonable parameters to use in TCA determinations where the CA and TCA are unknown. These relationships derive the TCA from its appropriate RCC of the CA, using 1 RCC= 43.3 years.

We have approached the relationship between the average RCC of the cluster members and the RCC of the CA by another route. We recognized from one of the correlations that SD is highly correlated (0.905) with the average RCC of the cluster. We can use that relation to provide a best fit to the data by minimizing the difference between the known value of CA and the computed value of CA for all seven calibration surname groups using the relationship:

Computed value of CA=
Average RCC + (F times the SD of the cluster).

The factor F that minimizes the difference between the observed and computed values of the RCC at CA was found to be 0.2857. Therefore the RCC of the CA should be located at the average RCC plus 0.2857 times SD. But, since SD= 0.5146 times the average RCC, we derive:

RCC of the CA = 1.147 x (average RCC for cluster)

Thus, we can estimate the TCA from these two different approaches. The first and second approaches lead to TCA= 55.6 and 49.7 times the average RCC of the cluster, respectively. Averaging these results leads to the relation:

TCA = 52.7 x (the average RCC of the cluster)

We tested the robustness of the second approach by varying the number of years corresponding to a unit change in RCC. The factor did not change by more than 8 percent over a range of 20 percent. We suggest using the average value of 52.7 first, then comparing the result with the two other relationships involving Point P and the SD. Experience with the uncertainties involved suggest that the result may have errors as high as 25 percent.

How does this result compare with traditional methods of determining TCA? Setting aside the M222 TCA from the discussion because it extends far beyond the most genealogical interesting times, the TCAs in the calibration range from 215 to 550 years. An error of the order of 25 percent translates to uncertainties in that range upward to 150 years, but those uncertainties will certainly be larger when the calibration is turned around and applied to clusters whose CAs and TCAs are unknown. Thus, our result for TCAs may not be significantly better than those reported by the testing companies of individual pairs of TMRCAs, but these results apply to clusters, not to pairs of testees. This application to clusters significantly broadens application for Y-DNA analysis from the TMRCA of pairs to the TCA of clusters.

3. Observations about the Surname Groups Used to Calibrate the RCC Time Scale

Although the Athey, Ewing, Logan, and Hamilton surname projects have well-documented pedigrees and TMRCA pairs, and contribute to the RCC vs. Time relation, they tend to have unique differences that could lead to uncertainties in the calibration.

Both the Athey and Ewing groups have well-researched pedigrees and sets of TMRCAs and both have TCAs that range from 215 to 300 years ago–relatively recently in times of genealogical interest. The Logan group has good TMRCA pairs but the pedigrees have been presented in generations, not years, leading to uncertainties in converting from generations to years. Two major Hamilton groups were considered. Hamilton B has a known founder, providing a unique CA and TCA, and was used in the time calibration. Hamilton A, while larger, was not used since its CA is uncertain. The M222 group was not used in the calibration of the RCC time scale, but since the oldest of the surnames within this group appear to be very near the TCA derived from the overall supercluster, it was used in the determination of the equations to be used to estimate the TCA of a cluster whose CA is unknown. Its major contribution to the results of Section 2 was to provide RCC and SD calibration points at the extreme high end of the year interval of interest to genealogists. More detailed comments about the Hamilton Groups and M222 Clade follow.

3a. The Hamilton B Group

A large, very reliable set of testees has been found in a cluster called Hamilton B in which 39 males have at least 37 markers tested (Hamilton, 2008). The pedigrees of many group members, combined with their Y-DNA results, point to a CA, James Hamilton, 4th Baron of Cadzow, who married Janet Livingston. From the Hamilton B data, we determine that the average RCC of the Hamilton B Group is 8.9 with a SD of the distribution equal to 6.6. With 39 testees, the SD of the average RCC is (6.6/Sqrt (39-1), or 1.0, which is 12% of 8.9). From this average RCC, we can calculate the preliminary estimate for the TCA, 8.9*52.7 = 468 (SD 13%) years ago. Using the SD of the cluster, we derive a CA of 2.356*6.6= 15.4 or a TCA = 671 years ago. Using the Point P approach, we derive a CA of 0.61*18= 11.0, or a TCA of 475 years ago

J. Leslie Hamilton, in his history of the Maymore Hamiltons (Hamilton, 2000), gives this James Hamilton's birth year as 1396-1398. We estimate that the time interval between James' birth and the average year when his descendants were tested is: about 2005 (average year of the test), minus 60 years (the average age of the testees when tested), minus James' birth year (about 1397), or 548 years ago with an estimated uncertainty of about 30 years. This 'observed' value is in good agreement with the computed values in the previous paragraph.

3b. The Hamilton A Group

There is a larger group of 80 Hamilton testees, called Hamilton A, that have pedigrees that go back to or through Sir Walter Fitzgilbert de Hamilton, 1st Laird of Cadzow (Hamilton, 2008). From the Hamilton A data, we determine that their average RCC is 11.3 with a SD of the distribution of 6.8. From this average RCC, we can calculate the preliminary estimate for the TCA, 11.3*52.7 = 596 (SD 13%) years ago. Using the SD of the cluster, we derive a CA of 2.356*6.9= 16.1 or a TCA = 697 years ago. Using the Point P approach, we derive a CA of 0.61*27.5= 16.8, or a TCA of 726 years ago.

Sir Walter Fitzgilbert de Hamilton first appears in the records as a witness to a charter of James Stewart, 5th High Steward of Scotland, granting land to the monks of Paisley Abbey in the year 1294, and he died about 1346 (Wikipedia, 2008). These dates suggest that he was born about 1274, consistent with other sources that give his birthplace as Blackball, Renfrewshire, Scotland. If we take 1945 as the birth year of the average testee and 1274 as the birth year of Sir Walter, the difference, 671 years shows that Sir Walter lived very close to the time of the progenitor of the Hamilton A group, and he may have been the progenitor of the group, himself.

3c. The M222 Clade

Haplotypes from the R-M222 project (Wilson, 2008), covering a wide variety of surnames, each derived for the SNP M222, have been analyzed. The average RCC for this matrix of pairs of 172 testees was 30.1. We can estimate the time of origin of the M222 SNP from the three methods as (1) 30.1 x 52.7 = 1590 years, (2) 2.356*13.35*43.3= 1360 years, and (3), 0.61*60*43.3= 1580 years. Using traditional techniques, Nordtvedt (2008) has estimated the time of origin for this SNP as about 1740 years ago, in reasonable agreement with the correlation prediction.

4. Investigating the RCC Time Scale Using Average Mutation Rates

We can use Chandler's (2006) average 37-marker mutation rate of 0.00492 (SD of 15 percent) mutations per locus per generation to derive the relationships in **Table 3**. All values refer to a 37-marker haplotype.

The resulting value for one RCC unit, about 46 years, is consistent with the value of 43.3 derived from pedigrees. Note that in the calibration of the RCC time scale based on pedigrees, we did not use an average

Table 3
Comparison of RCC and Chandler Rates

Number of mutations per generation	*0.182*
Number of years per generation assumed	25
Number of mutations per year	0.00728
Average number of years before one mutation occurs	137
Average number of generations before one mutation occurs	5.5
Average number of units of RCC corresponding to one mutation change (from the model in this paper, below. Est. SD is ~ 15%)	3
Number of years corresponding to one unit change of RCC	45.8

mutation rate and we did not assume any number of years for a generation, so our calibration is independent of those quantities.

5. Testing the Linearity of the RCC Time Scale by Using a Model

In this section we report on an investigation of the relationships between RCC, the RCC time scale, the mutation number, and the average absolute marker distance (genetic distance) using a model. We show that the RCC time scale is approximately linear. The model uses an average mutation rate over the 37 markers.

We started with the 37-marker modal haplotype of the Hamilton Group A as the hypothetical MRCA of a cluster (Hamilton, 2008), and synthesized three lines of descent, with each row and line entry experiencing one mutation change in each line of descent through 50 mutations. For each mutation change we used a random number table to choose the marker that will undergo a one-marker change and another random number table to choose whether the marker should be increased or decreased. Using this model we have investigated the relationship between mutation number, the absolute value of the marker distance and their associated RCC value for each of the three lines. We used the relationships in **Table 1** in the investigation and worked with the average values of the three lines of descent. The model covered 50 mutations. Since about 5.5 generations must elapse for one mutation to occur (Chandler, 2006), the model covers 50 x 5.49 x 25 years, or 6900 years. **Figures 2-4** show the results of these calculations.

As expected, as more mutations occur, the value of the absolute marker distance (genetic distance) increases, but not linearly because mutations can change upward or downward. The well known average-squared-distance (ASD) approach is used to model this effect.

This same non-linearity is present in the relation between the average absolute marker distance and the average value of RCC where the distance goes up at a slower rate than RCC.

While **Figures 3 and 4** show a considerable deviation from linearity, the relationship between RCC and mutation number in **Figure 2** is much more linear.

The relationships in **Table 3** were derived from studies of father-son mutations that have occurred near the present time, so the average number of years before one mutation occurs in 37 markers (viz., 137 years, assuming 25 years per generation) may be valid only for the present era. However, the model and **Figure 2** strongly suggest that over the time period of genealogical interest (viz., out to 2000 years in the past, or to values of RCC of 40-50), we may assume that RCC and time are linearly related. If they are not, the errors in assuming a linear relationship are small compared to the other errors inherent in our analysis.

The comparisons of the marker differences with time between each testee and the MRCA are consistent, and are in substantial agreement, with the work done by Walsh (2001; see also Kershner, 2009). They are shown in **Table 1**.

Any pair of individuals will have a MRCA back somewhere in time, but if we use the RCC value as a guide to the TMRCA, we will rarely expect to find an MRCA in a surname cluster earlier in time than about 900-1150 years ago (RCC > 20-25). This is the epoch just before the advent of surnames. There are few genealogical pedigrees that extend earlier than that epoch.

Discussion of Errors in the Method

We have investigated the effects on RCC caused by the quantization of marker changes. We have determined the effect this quantization error has on the value of RCC. The quantization error has a progressively larger effect when fewer than 37 markers are used. RCC differences of the order of 3 result from one marker change when about 37 markers are tested. An RCC difference of 3 corresponds approximately to an uncertainty of about 130 years (see **Table 1**).

Quantization uncertainty will occur in any method of assigning a time scale to haplotype differences. In fact, we have found that the standard deviation of the RCC

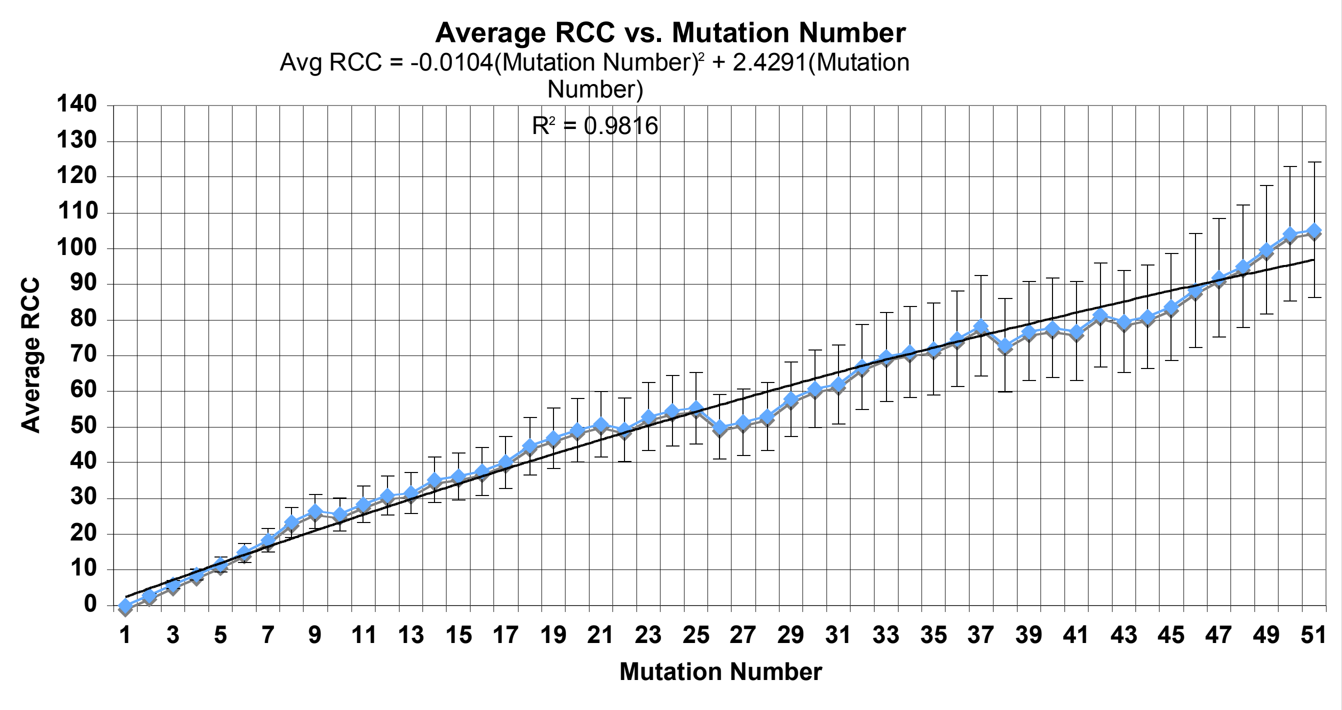


Figure 2. The relation between the average RCC (over three lines of descent) and the corresponding mutation number in the model. The 18% error bars are the average SD of the three lines of descent. Individual values range from a few percent to about 30 percent near Mutation Number 50.

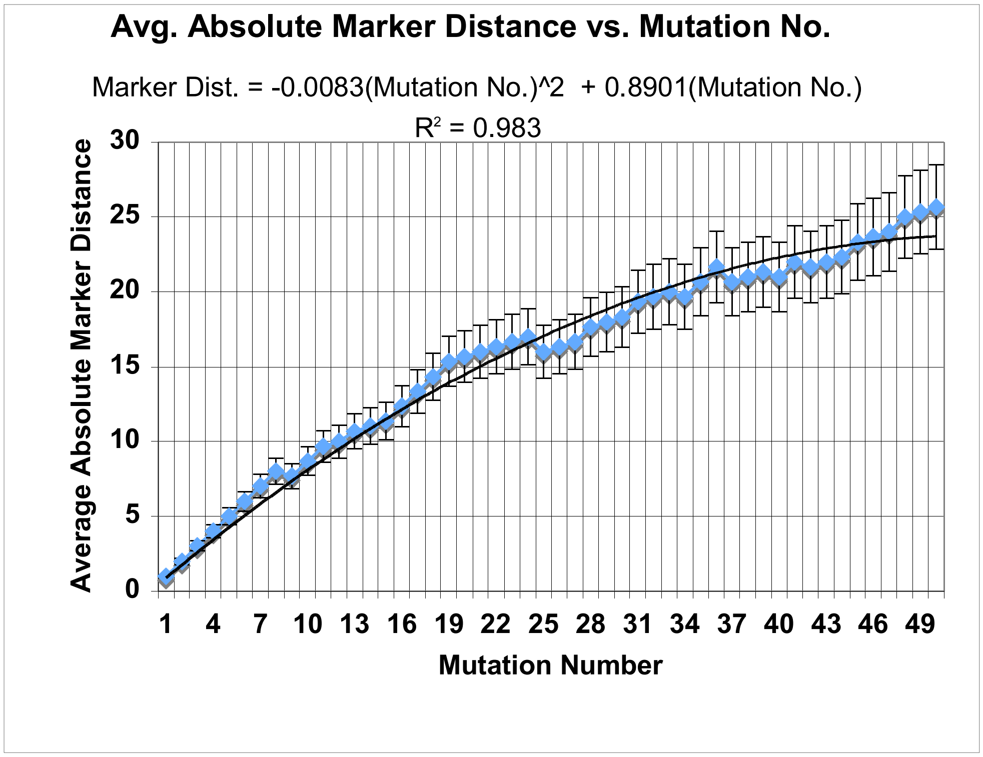


Figure 3. The relation between the absolute marker distance (over three lines of descent) and the corresponding mutation number in the model. The 11% error bars are the average SD of the three lines of descent. Individual values range from a few percent to about 20 percent near Marker Distance 26.

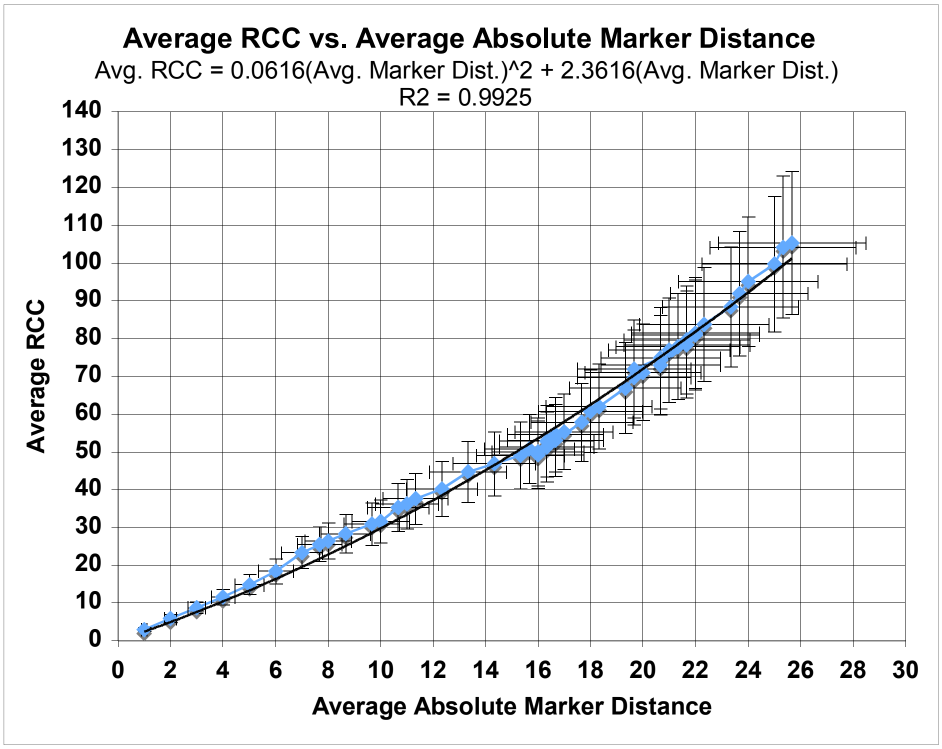


Figure 4. The relation between the average RCC and the average absolute value of marker distance (over three lines of descent) and the corresponding number of mutations in the model. The values of Average RCC and the values of the Average Marker Distance have average SD error bars of 18 and 11% over the three lines of descent in the model.

is about 5.8 (~250 years), which translates to about the same degree of uncertainty that has been ascribed to traditional methods. These percentage errors become increasingly smaller as we go back in time and more mutations occur.

Throughout this paper the errors cited are based on the statistics of the analysis and they represent internal errors. Unknown errors, particularly if they are systematic errors, will add to the uncertainty of our conclusions. It is doubtful that they will be any worse than the ones that are also inherent in the more traditional ways of finding MRCAs or determining time scales.

The haplotype of a testee can be viewed as the accumulation of random mutations over many generations. Errors due to mutation randomness are present in our estimates of time. The correlation analysis is based on the assumption that a haplotype evolves smoothly over time, but its evolution actually has random walk characteristics and does not proceed smoothly in time. Therefore, it is necessary to consider random mutation errors in haplotypes when we characterize that evolution as taking place smoothly in time. In an average of a sufficiently large number of testees, these random mutation errors will average just as any random errors will average, but especially in a small collection of testees, one cannot ignore the mutation randomness errors. This is not an error in testing the haplotype; rather it is a result of a random mutation process.

While there is no evidence that a linear relation between the values of RCC and its time scale cannot be used within time intervals that are of interest to genealogists, there are indications that it may become non-linear farther back in time. **Figure 2** suggests that as the total mutation number increases (roughly linearly over long periods of time), the average RCC tends to increase at a slower rate, which will introduce nonlinearity in the RCC time scale. However, the effect is small, particularly in comparison with other uncertainties that are inherent in the analysis.

To a first approximation we can put a limit on the magnitude of non-linearity using the following reasoning. The earliest time in the past that can be associated with Y-DNA results is the time back to "Y Adam," the most recent common patrilineal ancestor of all human males. This time has been estimated to be 70-80 kya.

Our study of Y-DNA haplotypes from different haplogroups has shown no larger RCC value than about RCC= 1200. Therefore, if we divide the largest time interval, say 75,000 years by the largest RCC value we have observed, we obtain about 62 years per RCC unit. This suggests that the assumption of long-term linearity can be safely used to perhaps tens of thousands of years ago. Even if later evidence shows that the time of Y-Adam is of the order of 90 kya, the RCC time scale appropriate for the most distant past will have an upper limit of about 75 years per RCC unit. Thus there are indications that any nonlinearity will not result in an error greater than a factor of two throughout the human time scale.

The Pros and Cons of the Traditional Approach and the Correlation Approach

We finally list the pros and cons of the traditional approach to analysis of Y-STR clusters and of the RCC approach. Pros and cons that are shared by both approaches are not considered; only the differences in the two approaches will be summarized.

Pros of the Traditional Approach:

- You can analyze and compare individual marker differences.
- You can consider different lengths of haplotypes for comparison purposes, but TMRCA calculations must still use haplotypes of the same length.
- The presence or absence of particular markers may lead to better pair associations and to better time estimates.
- TMRCA calculations, both for pairs of haplotypes and for clusters of haplotypes have a firm theoretical basis.

Cons of the Traditional Approach:

- Methods of matching testees are vague, hard to define, and vary among project administrators.
- Methods for determining the TMRCA are proprietary and/or depend on sophisticated statistical variance techniques.
- Methods based on genetic distance lose information inherent in the original marker values.
- Different algorithms based on genetic distance (viz., numerical sum vs. absolute values) exist.
- If mutation rates are revised, the influence of each marker must be reevaluated.
- Assignment of haplogroup time scales is done on a case-by-case basis. They are often mutually inconsistent, especially when done by different investigators.
- Does not directly lead into an overall, uniform time scale as the correlation method does.
- Does not directly lead to a sequence of evolution of surname and other subgroups.
- As test results accumulate, it is increasingly difficult to match and analyze the results.
- The matching of haplotypes and the TMRCA analysis must be done separately.

Pros of the Correlation Approach:

- Pair differences expressed by one number allow quick, deeper comparisons to be made.
- The approach can be applied simultaneously to very large numbers of marker pairs.
- The derived value of RCC is a single number that directly correlates with the TMRCA.
- A methodology exists that will identify the approximate time when the common ancestor of a surname cluster lived.
- The technique implicitly accounts for marker mutations that have taken place over many thousands of years, permitting the genealogical time scale to be extended far beyond the time horizon of pedigrees.
- An RCC vs. time relation is derived that need only be modified by a scale factor if future research requires changes.
- The RCC time scale is based on a direct time calibration from pairs of well-researched pedigrees. It is consistent with other calibration approaches.
- Individual markers in the correlation program can be weighted in the event more dependable mutation rates are derived that are individually better than the average of all markers.
- Assignment of haplogroup time scales is done on a uniform basis. It is scalable over all haplotypes.
- Averages over large strings permit the investigation of relationships farther back in time than the traditional methods attempt to cover.

- The association of testees done by the traditional approach can be reevaluated for group membership. The analysis permits the identification of subgroups.

- Directly leads to sequences of evolution of surname and other subgroups.

- Directly leads to the time that subgroups take to evolve from their parent group.

Cons of the Correlation Approach:

- Must use the same length of marker strings.

- Two strings of markers result in a single number. Information from individual marker values is lost.

- Administrators must be minimally adept at using Excel-type spreadsheets.

Consideration of the pros and cons highlights differences in the methods and provides the rationale for using both the traditional and the correlation approaches together. Part 2 of this article will discuss specific applications of this technique to selected surname groups, and haplogroups showing how they have evolved with time.

Acknowledgements

I wish to thank the following people for providing pedigree information which, with haplotypes, permitted the RCC time calibration: T. Whit Athey, David N. Ewing, Gordon Hamilton, and J. J. "Jim" Logan. The early encouragement of Linda MacLachlan to pursue this line of analysis and to investigate the M222 clade was stimulating. Discussions with David E. Hogg and Elizabeth B. Waltman improved both the analysis process and the presentation. The advice by Frederic R. Schwab to use the Hodges-Lehmann estimator was invaluable in indicating the best methodology to pursue the RCC time calibration.

References

Chandler J (2006) Estimating Per-Locus Mutation Rates. *J Genet Geneal*, 2:27-33, 2006.

Logan JJ, Falls SL (2008) Logan DNA Project.

Hamilton G (2008) Hamilton Surname DNA Project. See

Hamilton JL (2000) *The Maymore Hamiltons and Related Families.* Privately published. Copies have been deposited in the Library of Congress and in the Family History Library, 35 North West Temple Street, Salt Lake City, Utah, 84150.

Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34:598-611.

Nordtvedt K (2008) Note to Rootsweb's Genealogy-DNA-L@rootsweb.com of 20 May 2008.

Saleh AK, Ehsanes MD (1976) Hodges-Lehmann estimate of the location parameter in censored samples. *Ann Inst Stat Math*, 28:235-247.

Walsh B (2001) Estimating the time to the MRCA for the Y chromosome or mtDNA for a pair of individuals. *Genetics*, 158:897-912.

Wikipedia (2008) Walter fitz Gilbert of Cadzow.

Wilson D, McLaughlin JD (2008) R-M222 Haplogroup Project..

Notes

1. We define the correlation coefficient used in this study the same way that the Microsoft Excel program defines it, namely: The correlation coefficient between two strings of markers (X and Y) is:

$$\frac{\sum (x_i - x_m)(y_i - y_m)}{\sqrt{(\sum (x_i - x_m)^2 \sum (y_i - y_m)^2)}}$$

where x_m and y_m are the means of the array strings X and Y, respectively.

The Microsoft Excel program has a data analysis tool that can be used to compute many statistical results, of which the correlation and histogram routine is used in the present article. Both routines can be used on literally hundreds of strings and matrices of data.

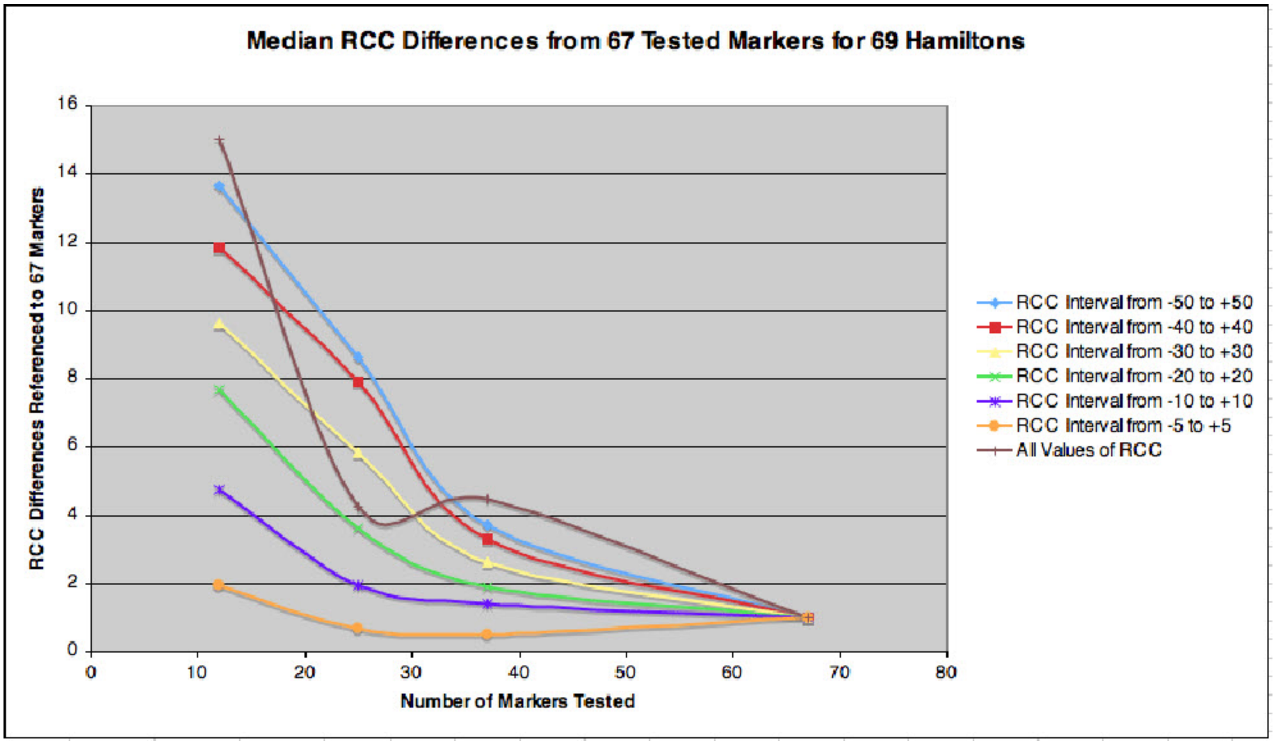
2. We took a representative string pair of 12, 25, and 37 markers and made one-marker changes in one of the pairs of the string. The following changes in RCC were obtained:

Number of Markers	Min. RCC change	Average RCC	Max. RCC change
12	4.6	22.1	32.9
25	5.1	8.1	11.3
37	2	3.1	4.5

While an analysis of 67 markers would have reduced the quantization error, there are more markers available at the 37-marker level, an advantage that outweighs the quantization error.

3. The more detailed relationships from which these conclusions were reached appear in the figure, below.

4. The largest value of RCC yet found has been 1202, between Kit Nos. 25210 (Haplogroup A*) and 27507 (Haplogroup C3).



APPENDIX A

An Algorithm that will Produce a Particular Time Slice of the Matrix

The following procedure, using an Excel spreadsheet, builds an algorithm that will show a particular time slice within the RCC matrix:

1. Set up a completely filled, two-sided RCC matrix.

2. Duplicate that matrix below the above matrix.

3. On two lines above the second matrix enter a value of High RCC on the first line and a value of Low RCC on the second line. As an example, in Step 5, the high and low RCC values are located in C171 and C172, respectively, and the upper left hand corner of the first matrix is located at B88.

4. Insert a formula, patterned on the example in Step 5, in the upper left hand corner of the second matrix.

5. IF(B88="","",IF(B88=0,0,IF(AND(B88< $C$171,B88>$C$172),B88,"")))

6. Copy that formula to all entries of the second matrix.

This procedure will show in the lower matrix only the RCC values between the high and low values selected. It will insert blanks along the diagonal of the matrix and it will retain any zero values that appear in the first matrix. This is the end product that allows us to sample the matrix in various slices of time.

APPENDIX B

Using the time slicing algorithm to group low RCCs together and form clusters

The following approach, although labor-intensive, allows us to form good clusters using the time slice algorithm:

1. Use the algorithm in **Appendix A** to sort the full matrix so that only RCCs between 0 - 5 are shown.

2. At the bottom of the first testee in column 1, note the row identification of testees who share RCCs in that RCC column 1 interval.

3. Label this first group A.

4. Go to the second testee in column 2 and note the row identification of testees who share RCCs in that same RCC column 2 interval.

5. Label that group B unless members of Group A are present; otherwise, label the second testee as belonging to Group A.

6. Do this, one by one, for all the columns, adding more groups if the testee is not already in a previously named group.

7. Use the same algorithm to sort the full matrix so that only RCCs from 5 - 10 are shown.

8. Repeat steps 2-6 and repeat the process through RCC intervals 10-15 and 15-20.

9. Sort the full matrix in rows and columns so that the results are grouped into clusters.

10. The end product will be an RCC matrix in which groups with low values of RCC are gathered within one or more clusters. Those clusters will be centered along a diagonal in the RCC matrix.

Alternatively,

1. Form the full matrix in Step 2 of **Appendix A.**

2. Identify areas in the matrix that have low values of RCC (e.g., values between 0 and about 20).

3. Cut and paste the rows so that the adjacent, low values of RCC are grouped together.

4. Cut and paste the columns so that their sequence matches the new row order.