Journal: www.jogg.info
Originally Published: Volume 5, Number 2 (Fall 2009)
Reference Number: 52.014

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: THE USE OF CORRELATION TECHNIQUES FOR THE ANALYSIS OF PAIRS OF Y-CHROMOSOME DNA HAPLOTYPES, PART II: APPLICATION TO SURNAME AND OTHER HAPLOTYPE CLUSTERS

Author(s): William E. Howard

Journal of Genetic Genealogy, 5(2):271-288, 2009

The Use of Correlation Techniques for the Analysis of Pairs of Y-STR Haplotypes, Part 2: Application to Surname and Other Haplotype Clusters

William E. Howard III

Abstract

The details and utility of using correlation techniques involving Y-chromosome DNA results to analyze differences in haplotypes are presented. A time scale is introduced that can determine the time to common ancestors of pairs of testees, the time of formation and evolution of clusters of testees, and times related to the evolution of haplotypes and haplogroups that are only distantly related. The technique can be applied to indicate (1) when surname groups evolved as separate entities in the past and (2) the time interval over which that evolution occurred. The time interval considered in this paper extends back far beyond 6000 years. The analysis to date shows no evidence that the derived time scale departs from linearity by more than a factor of 2 (internal error) over the entire time span of Y-DNA in mankind, back to the emergence of Homo sapiens from Africa. We find that the percentage uncertainties in time estimates over very long time scales are significantly less than those for epochs of genealogical interest. Thus, provided a link can be found to Y-DNA, this correlation approach may serve as a future vehicle to link more closely together the time scales of mitochondrial DNA, migration patterns, linguistic patterns, geology, anthropology, archeology and paleontology. The derivation of the RCC time scale and its applicability over distant times in the past is arguably the most important product of this study. Suggestions are made for future work that may extend the analysis and our understanding of these relationships.

Introduction

Part 1 of this two-part series of articles presented a new correlation method for analyzing Y-STR haplotypes (Howard, 2009). The method reduces pairs of haplotypes to a single number (RCC), shown to be proportional to time (Note 1). These differences in haplotypes correspond to genealogically interesting time scales and have application over longer time scales. In Part 1 the advantages and disadvantages of this new approach were compared with traditional methods. The RCC vs. time relationship was calibrated in Part 1. This, and the introduction of a testable time scale, is the power of this technique. In this paper we will assume an understanding of the contents of Part 1.

Here we apply the same correlation method to different kinds of 37-marker haplotypes to illustrate how this technique can be used in conjunction with traditional analytic methods to gain more information about the groupings of surnames, the time relationships of those groupings, and their evolution in time. It is not the intention in this paper to conduct full studies of individual surnames. Instead, we will pursue different types of analyses that can be applied to haplotypes and we will present some insights and conclusions that can be reached through various studies of the RCC matrix.

We will give illustrative examples from Hamilton, Cook, Logan and M222 haplotypes, showing how the correlation matrix can be used to analyze and date surname clusters. We will discuss the most recent common ancestor (MRCA) of pairs of surname haplotypes and the common ancestor of clusters and interclusters, including the phenomenon of surname clusters within clusters, and we will explore the dates of origin of surnames, concentrating on the Logan and Hamilton surnames (Logan, 2008; Hamilton, 2008). We will suggest future studies designed to compare RCC results with details within the ISOGG haplogroup tree (ISOGG, 2009).

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

Received: June 6, 2008; accepted: August 1, 2009

Introduction to the Surname Project Data Analysis:

As time passes, haplotype markers change at random, usually by +/-1 repeat unit. Each marker has a mutation rate, currently an area of active study by others. While many current techniques of analysis of DNA for genealogical time scales concentrate on individual marker mutation rates, we use an average mutation rate over the relatively large string of 37 markers because the effect of the uncertainties in individual marker mutation rates become less, the more markers we use in our analysis. If we knew more about the details of how each marker mutates, we might be able to reach a more satisfactory conclusion, but current discussions about individual marker mutation rates are full of speculation, uncertainty, argument and lack of sufficient agreement. Moreover, we will always be faced with uncertainties caused by mutation randomness. We therefore conclude that we may gain more by using average mutation rates across a large set of markers than by using individual marker rates. Because we know more about the average change over time of a string of markers than we do about each individual marker, the uncertainties of the resulting RCC for a pair of testees are dependent on only one number. The uncertainty in that number comes primarily from the quantization problem and the small number of mutations. This combination of effects leads to larger percentage uncertainties for smaller values of RCC, and conversely.

As we study the evolution of haplogroups over thousands of years, the use of an average mutation rate to determine a time scale becomes an even more powerful tool, particularly if the average mutation rate is constant in time. If the average mutation rate is found to be time-variable, we only need to find ways to recalibrate the RCC time scale and to change the average mutation rate accordingly. In this study, we assume that the relation between RCC and time is linear because there is no compelling evidence to do otherwise.

The RCC (Correlation) Matrix

A. *Schematic Matrix*

Figure 1 of Part I shows a portion of an RCC correlation matrix. Schematically, it has the components seen in **Figure 1** on the next page. The RCC values at the intersections of each row and column in the matrix are the results of the correlation-based calculation for each pair of haplotypes (**Figure 1a**). After they are grouped, values of RCC within the cluster region will be always be lower than in the intercluster regions.

Values of RCC within a cluster will be different because different pairs of cluster testees usually have different times to their most recent ancestors (TMRCAs). However, all the individual MRCAs in a cluster will have a common cluster ancestor (CA) who will have lived at least as far back in time as the earliest TMRCA found for any cluster pair. An estimate of the time to the common ancestor (TCA) of all the cluster members was discussed in the companion article (Part I).

We will call the part of the matrix that represents the RCC for pairs of participants from two different clusters, their intercluster region (**Figure 1b**). The intercluster region for Cluster 1 and Cluster 2 contains the RCCs of each member of Cluster 1 paired with each member of Cluster 2. Just as the individuals in a cluster have a CA, so will the individuals paired in the intercluster region have a CA. That CA will be the common ancestor of each cluster's common ancestor. This CA must be a single individual and all of the RCC values in the intercluster region must pertain to a single time back to this individual, in contrast to the times to a common ancestor for pairs within a cluster. This assumes that one cluster is not a subset of the other. Of course, even though all of the intercluster RCC values should indicate a single time, the effects of randomness will still be evident, but it is valid to average these RCC values and convert it to time using the factor of 43.3 years per RCC unit derived in Part I. Thus, an analysis of the intercluster region will indicate the TCA of the two clusters.

This process can be continued in the intercluster regions, pair by pair, so that an estimate can be made of the evolutionary sequence of all the clusters in the matrix through a study of their individual TCAs.

B. *Inferences from the Histogram of the RCC Values in the Correlation Matrix*

In developing this correlation approach, the following surnames have been studied: Athey, Bartlett, Barton, Bean, Campbell, Cooke, Doherty, Ewing, Fitzpatrick, Gordon, Hamilton, Howard, Logan, McLaughlin, Radcliffe, Richardson, and Thompson.

Once an RCC matrix has been developed, it is instructive to investigate its complexity using a histogram before analyzing the matrix further. As an example, we consider a 37-marker histogram composed of testees with the surname Hamilton (Hamilton, 2008), presented in **Figure 2**.

The Hamilton surname group consists of 168 testees who belong to two different haplogroups in more than a dozen clusters so we expect a large range of RCC results. The upper left histogram covers the complete set of haplotypes in the full matrix; it shows three prominent peaks. The first peak contains an overlap of all groups of Hamiltons who are closely related, regardless of the haplogroup to which they belong. There are two groups of Hamiltons with a large population: Hamilton Groups A and B. They appear in the histogram, grouped together, and dominate the population of the first peak. The second peak at RCC= 60 consists of members of

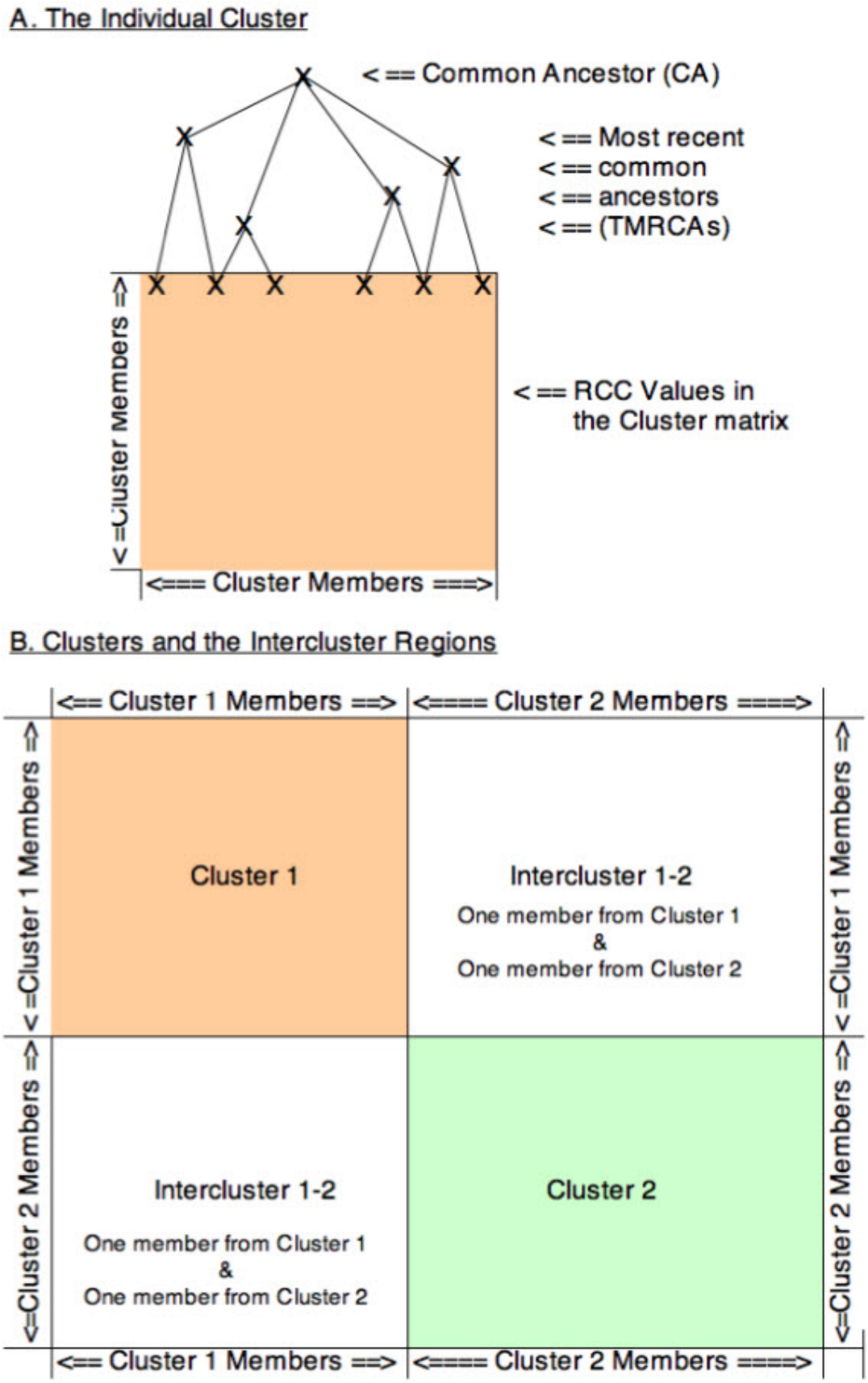


Figure 1. Cluster grouping and the intercluster regions in an RCC matrix.

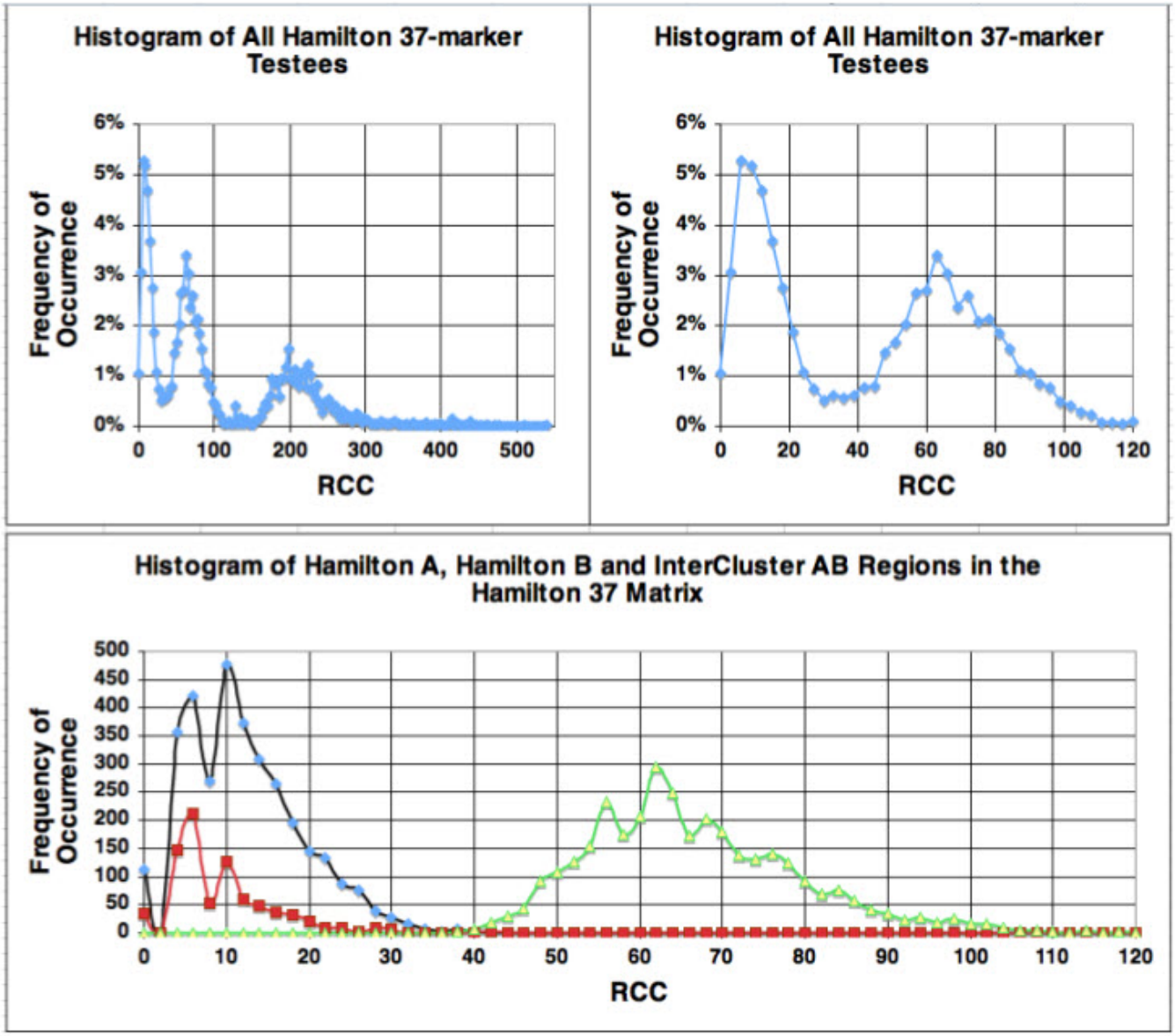


Figure 2. Histograms of the frequency of occurrence of values of RCC for the Hamilton A cluster (blue) and Hamilton B clusters (red), along with the histogram for the A-B intercluster region (green).

clusters who are slightly more distantly related. In this case, the second peak is dominated by testees from Hamilton A and B and consists mostly of the intercluster population of the Hamilton A and B groups.

A detail of the first two peaks is given in the upper right histogram. The lower histogram shows a further breakdown of Hamilton A, B and the intercluster region of Hamilton AB (Note 2). Both Hamilton A and Hamilton B are composed of a relatively large number of testees. Both these components in the lower histogram have RCCs below 30 but they both show indications of a double peak, especially Hamilton A. This is the first indication of subclusters forming within clusters, seen when RCC exceeds about 20.

If a matrix is composed of pairs of haplotypes that represent different testees in different haplogroups, their RCC values in the matrix will be large, indicating that their most recent common ancestor lived far back in time. When results are presented from pairs who share a surname or a deep clade, the RCC values will be smaller. Experience has shown that comparisons of haplogroups involve RCCs between 100-900; clade and

subhaplogroup comparisons between 50-100, surname intercluster region comparisons between 20-50, surname clusters between 0-20, with the best chance for discovering pedigree comparisons between 0-10. Pairs of all the Hamilton groups cause the third peak in the full histogram. There are more peaks above RCC= 300 that result from testee pairs from different haplogroups.

The peaks beyond the first peak have been termed pairwise mismatches (Fitzpatrick, 2005) and they also contain information about more distant ancestors who are shared by different groups of testees.

As an example of a surname that is composed of a very large group of very distantly related testees, we give the histogram for the surname Cook(e). Surname origins date from the 13th and 14th centuries in England when it became necessary to differentiate between growing numbers of individuals. Often people from particular professions would adopt the name of their profession. The surname Cook (Koch in German) and variations of its spelling appeared all over Europe as unrelated people from very different haplogroups chose the same name. Thus we should expect a Cook histogram to be composed of many haplogroups, which will cause a variety of RCC pairs. **Figure 3** shows the Cook histogram that illustrates that complexity.

Of the 90 Cooks in the sample, there were 14 clusters identified of which 36 pairs were in one cluster, 10 in another, 6 in two others, 3 in one other and only one pair in the 9 others (Note 3). In **Figure 3** they all appear at the extreme lower left. The figure consists of pairs of testees from Haplogroups E, G, I, J, and R. The most distant pair consists of one testee in E1b1b1 and the other in R1b1b2 (RCC= 676, corresponding to 29,000 years ago). Clearly groupings of this surname should be done by haplogroup, as the surname administrator has done (Cooke, 2009).

C. Use of the Correlation Matrix to Determine Common Ancestors and Evolutionary Sequences

Appendix A of Part I showed how to use a time slice algorithm to present in the correlation matrix only those values of RCC that fall between a specified high and low value of RCC. This process allows us to sample the matrix in various slices of time. Part I showed the details of two clusters and an intercluster region within a Logan surname matrix. A low and high value of RCC of 0 and 72 was sufficient to nearly fill that portion of the matrix, suggesting that a common ancestor for all pairs existed about 3100 years ago, or about 1200 BCE.
In **Figure 4** we have used the time slice algorithm to show four intervals of time in part of the Logan matrix

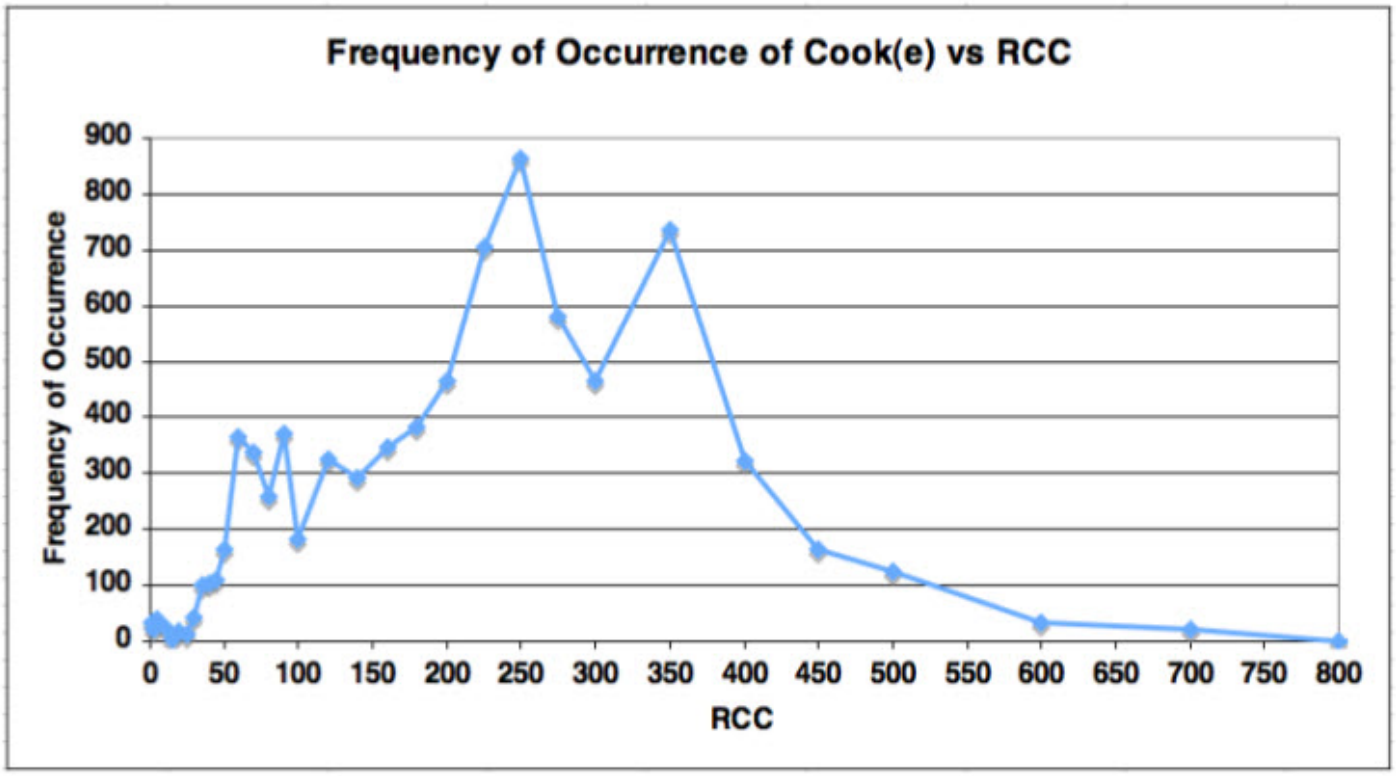


Figure 3. Histogram of the Cook (Cooke, Koch, etc.) Surname.

Journal of Genetic Genealogy, 5(2):271-288, 2009

that consists of three major and one minor Logan clusters. If the figure's resolution does not show the RCC values, their presence indicates how the identities of each cluster evolve from approximately 2200 years ago (RCC= 50), through RCC= 40, 20, and 10 to the present time.

Table 1 shows the RCC and corresponding time relationships among the testees in the three main Logan clusters, and those in the intercluster regions. The estimated TCA in Table 1a is determined using the methodology developed in Section 2 of Part 1. Table 1, followed by Figure 5, presents the evolutionary chronology.

The standard deviation (SD) of the average cluster RCC for Hamiltons, M222, Logans and Ewings was found to vary between 55 and 75 percent of its average value, but the correlation between SD and the average RCC is very high. The SD of the average intercluster RCC for Logans and Hamiltons was found to vary between 9 and 14 percent of its average value, indicating that the percentage uncertainties for the TCAs of interclusters are considerably less than for the TCAs of individual clusters.

INTRAcluster members have different pairs of MRCAs but all members have a CA who lived at a time that can be estimated (see Section 2 of Part 1). The INTERcluster members all point to the same intercluster CA who lived

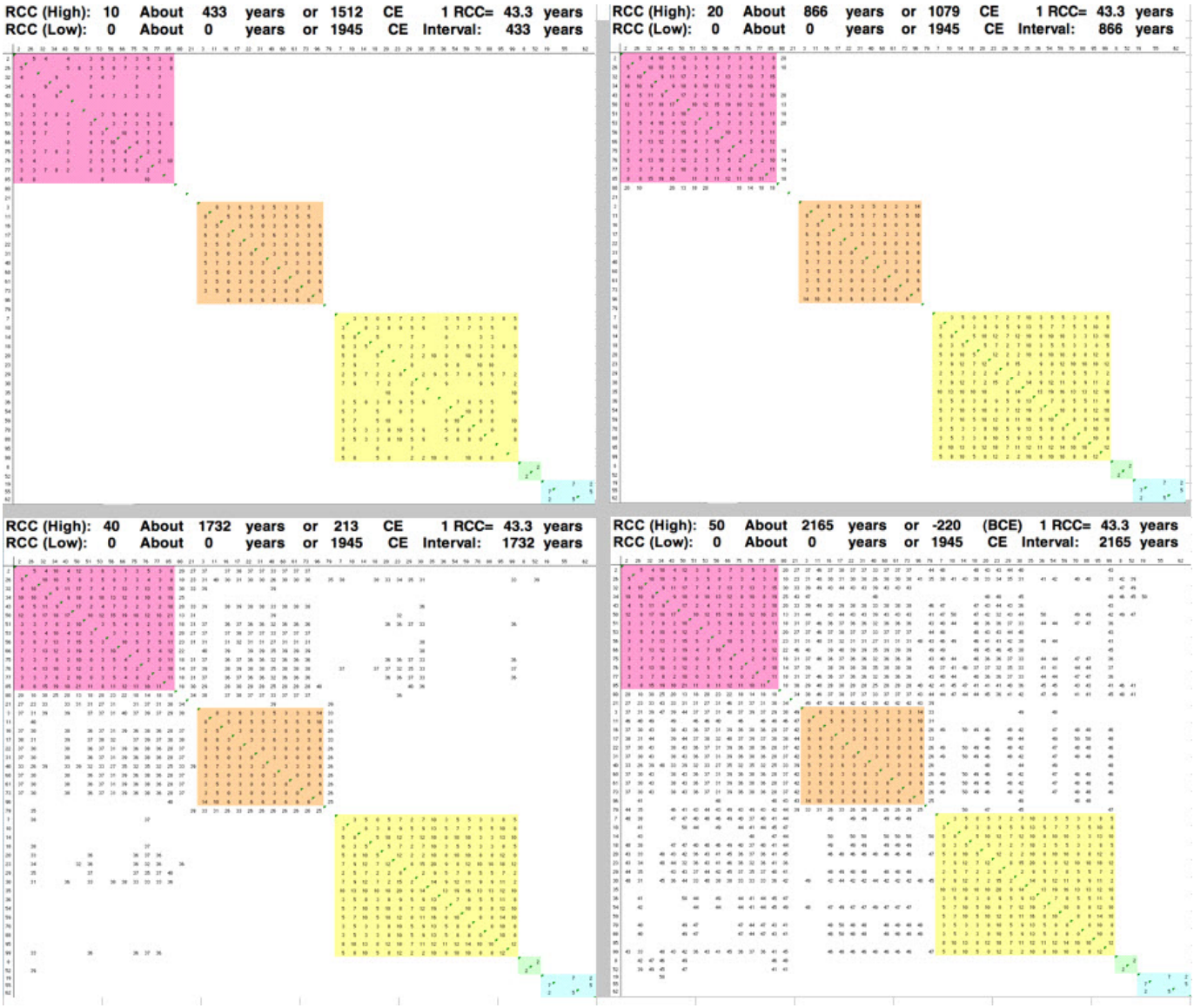


Figure 4. Filling the RCC matrix as the RCC/time interval threshold increases.

Table 1a
RCC and Corresponding Time Determinations for the Three Main Logan Clusters (RCC/Years Ago)

	Top Cluster 1	Middle Cluster 2	Lower Cluster 3
Average RCC	7.5 / 325 years	3.5 / 150 years	7.9 / 340 years
Std. Deviation	4.8 / 210 years	3.0 / 130 years	4.0 / 170 years
No. Members	14	11	16
Estimated TCA	400-490 years	180-310 years	400-420 years

Table 1b
RCC and Time Determinations for the Three Main Logan Intercluster Regions (RCC/Years ago, no correction needed))

	Intercluster Region 1-2	Intercluster Region 1-3	Intercluster Region 2-3
Average RCC	39.6 / 1700 years	49.1 / 2130 years	54.0 / 2340 years
Std. Deviation	8.2 / 360 years	10.4 / 450 years	7.4 / 320 years
No. Members	11	14	11
SD (Mean)	2.6 / 112 years	2.9 / 125 years	2.3 / 101 years

at a time that can be estimated by averaging the intercluster RCCs and converting that average to a time using the RCC-time relation.

Figure 5 summarizes graphically the three main Logan cluster and intercluster evolutionary relationships. Each cluster in the Logan example contains at least eleven members, so there are at least 55 pairs of testees who have individual TMRCAs that contribute to the average RCC of each cluster. We can trace the evolution of these three clusters down through time from a common Logan ancestor over 2500 years ago to the three current Clusters 1-3. Of course, that CA lived long before there were surnames .

It is important to note that the blue areas in each cluster indicate the spread of individual RCC values of the cluster testee pairs and the brown areas represent the SD of the intercluster CA. Since the RCCs in the brown areas of each intercluster region point to the same CA, the corresponding time to the CA will simply be the average RCC of the intercluster region times 43.3.

We emphasize one very important feature of Figure 5. When there are three clusters, two of the CAs of their interclusters MUST intersect at the same point—a shared CA. In Figure 5, the CA of Interclusters 1-3 and 2-3 will be identical, so we take the average of their individual CAs (Table 1b), obtaining a TCA of 2240 years ago. At that point, Cluster 3 splits and evolves separately. Meanwhile, from the split, the lines toward Clusters 1 and 2 evolve to the CA of those two clusters who lived about 1700 years ago. From that common ancestor, Clusters 1 and 2 evolved down different paths.

A study of the Hamilton surname resulted in the summary given in Table 2.

Of the Hamiltons in Haplogroup I1, Hamilton Groups A and B contain large numbers of testees. Their intercluster region points to a common ancestor at least 2850 years ago. Hamilton A is the oldest group; Hamilton D is the youngest group. The oldest intercluster age is about 3700 years old (the A-D intersection); the youngest intercluster age is about 1690 years old (the B-C intersection).

Hamiltons in Haplogroup R1b have fewer testees, but their clusters are well defined. Hamilton G is the oldest group; Hamilton E is the youngest group. The oldest intercluster age is about 5000 years old (the G-E intersection); the youngest intercluster age is about 1270 years old (the I-R intersection).

Journal of Genetic Genealogy, 5(2):271-288, 2009

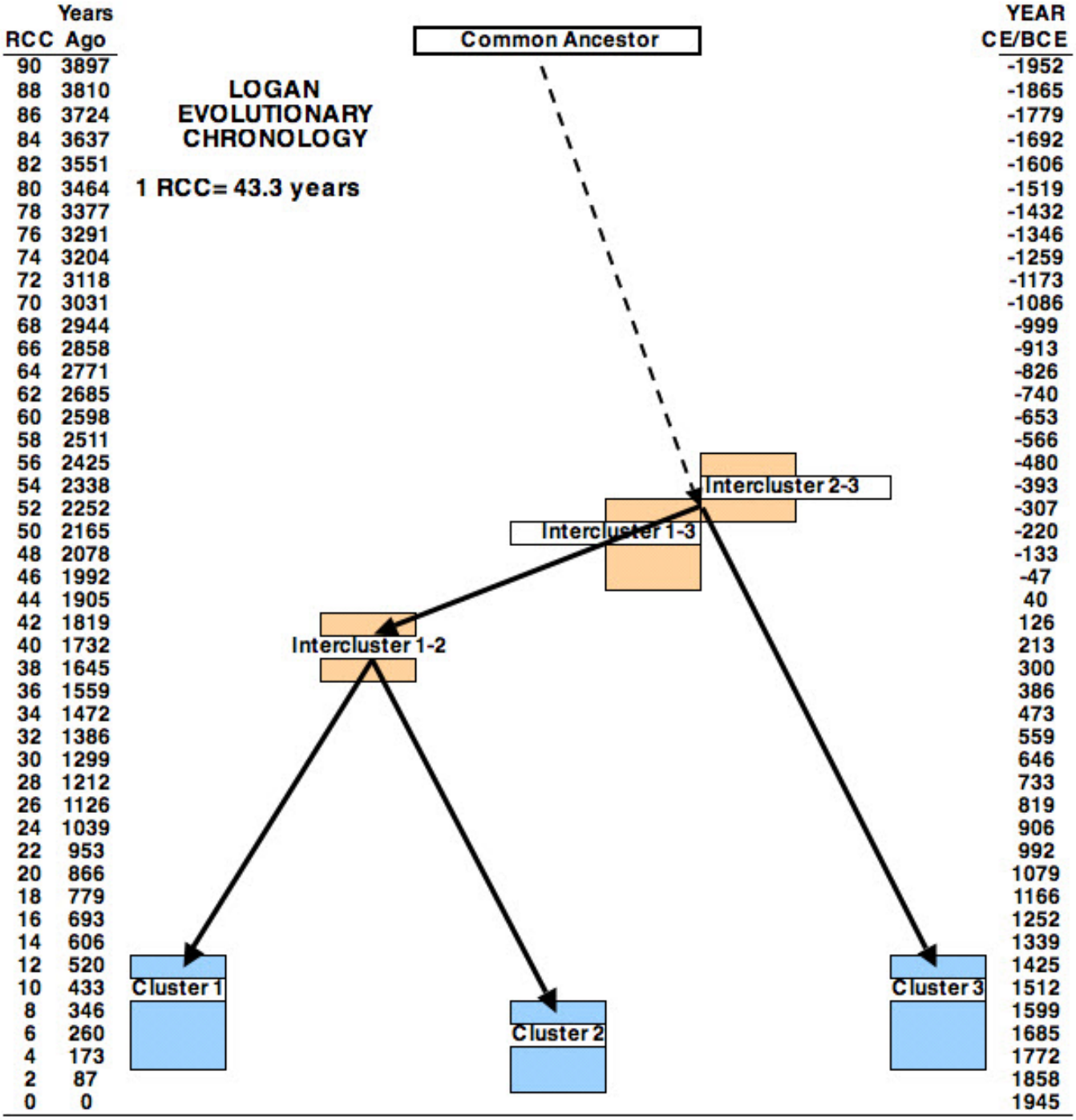


Figure 5. A Logan evolutionary chronology.

Table 2
Comparison of Hamilton Groups A-D (Haplogroup I1) and Groups E-R (Haplogroup R1b). See text for description.

1 RCC = 43.3 years

No. Testees =>		81	40	6	4	5	4	8	3	4	Yrs		Approx.
Hamilton Group:		A	B	C	D	E	G	I	M	R	Ago		Year (CE)
Average RCC =>	A	11.2	65.8	78.0	86.1	230.1	233.3	196.7	206.9	182.7	484	A	1461
St'd Deviation =>		6.9	13.1	14.1	9.3	14.3	20.9	15.2	12.5	17.3	299		1646
	B	2849	8.7	38.9	55.2	290.3	219.4	216.5	193.0	181.7	375	B	1570
		568	6.6	6.5	7.8	21.1	25.7	22.3	18.3	23.0	288		1657
	C	3376	1686	5.2	65.9	311.5	236.6	244.9	200.6	202.8	226	C	1719
		611	283	4.1	7.1	26.3	29.6	25.6	20.9	24.8	176		1769
	D	3728	2392	2855	1.0	338.2	222.1	215.7	198.6	184.3	45	D	1900
		403	340	308	1.3	3.9	21.2	17.7	10.4	15.4	57		1888
	E	9962	12568	13486	14643	3.2	115.8	74.6	102.5	92.8	138	E	1807
		620	912	1140	170	3.4	7.5	9.9	8.6	5.9	146		1799
	G	10102	9500	10245	9617	5013	9.7	43.4	45.3	41.2	419	G	1526
		903	1114	1280	920	323	8.2	7.9	6.3	5.1	356		1589
	I	8518	9376	10604	9341	3231	1881	8.1	48.5	29.4	350	I	1595
		656	965	1107	765	431	342	5.5	5.4	6.5	239		1706
	M	8957	8356	8684	8601	4438	1963	2101	8.1	45.1	353	M	1592
		542	791	905	450	372	274	232	6.5	2.1	279		1666
Avg years ago =>	R	7910	7866	8783	7981	4020	1783	1273	1955	6.6	285	R	1660
St'd Deviation =>		748	997	1076	668	254	220	283	90	5.6	242		1703

Notes to Table 2: The entries along the yellow diagonal give (1) the average RCC value of the testees in each Hamilton Group and (2) the standard deviation (SD) of that value. The number of testees in each Group is given in the row above the Group designation. The years corresponding to those entries are given in the last two yellow columns to the right of the Table. The average RCC values of the intercluster regions, together with their SD are listed at the intersections of different Hamilton Groups at the right of the yellow diagonal, and their corresponding times are located at the intersections to the left of the diagonal. The gray areas in the upper right and lower left show the average and SD values of RCC and years for the intersections of dissimilar haplogroups. Care must be exercised not to over interpret results that are derived from small numbers of testees.

The intercluster regions for Hamiltons in different haplogroups point toward a convergence of haplogroups I1 and R1b much farther back in time. The oldest comparison indicates an age of 14600 years (the D-E intersection); the most recent intersection indicates an age of 7900 years (the B-R intersection). Both ages were at or near the end of the most recent ice age.

Evolutionary Insight

The components in Tables 1 and 2 and the drawing in Figure 5 afford some insight into the evolution of a surname cluster. First, we note that we are sampling an evolutionary sequence at only one point in time—a snapshot of the evolution made at the present time. The

male ancestor of all the Logans who lived at least 2500 years ago is the progenitor of at least three lines of descent. Three lines end up as Logan Clusters 1, 2, and 3 that formed only within the last 600 years. Modern-day surname clusters are composed only of individuals who share a common ancestor within about 30 generations, almost always after the adoption of surnames. Pairs of testees located within these clusters have lines of ascent which intersect with members of other clusters farther back in time. For the Logans in these three clusters, that intersection occurred between 1700 and 2400 years ago, but all those lines of ascent lead to their common ancestor of 2500+ years ago.

Let us do a *gedanken experiment*. Let us suppose that we are future analysts in CE 4000-5000 looking at the results of an evolution that takes place between now and then. By that time members of each cluster will have evolved downward in **Figure 5** and, because of mutations, those evolutionary lines from each current cluster -- those which do not die out -- will fan out and produce new clusters, the members of which will again have a common ancestor within the past few thousand years. However, each cluster today will have evolved into more clusters over the next 2000-3000 years so that, looking ahead, today's clusters will become tomorrow's intercluster region. This thought experiment defines the roles of the subcluster, cluster and intercluster regions. The subclusters of today will evolve into the clusters of tomorrow unless their male lines die out; the common ancestor of the clusters of today can be identified through a study of the interclusters regions of which they are a part.

In a detailed study of the cluster Hamilton Group A we have found at least five internal subclusters (three are quite large) that represent future clusters from Hamilton Group A. Typically, these clusters-within-a-cluster have RCC values below 5 and exist among paired members whose largest RCCs may exceed 20. Hamilton A has an average RCC of 11. It is the oldest of the Hamilton clusters, so the subclusters have had time to form. The smaller subclusters will disappear if the male lines die out.

This insight into the process of the evolution of surname clusters allows us to draw the following conclusions from **Figure 5** and **Tables 1 and 2**.

1. The common ancestor of the present Logan clusters lived more than 2500 years ago.

2. All members of a cluster have a common ancestor whose TCA can be estimated.

3. We have traced surname cluster lines using their membership in interclusters as definable entities over time periods that have ranged from about 2000-2500 years (Logans) to even longer intervals of up to 5000 years (Hamiltons).

4. The distributions of RCC values in clusters and interclusters, as defined in a surname matrix, do not overlap in time.

5. The clusters of today began to be defined about 800-1000 years ago. It may not be a coincidence that a value of ~20 for RCC appears to be a practical starting point for forming a cluster. It corresponds to an epoch when surnames came into being, leading to a quicker cluster identification for testees who share a surname.

6. As clusters age, subclusters develop inside the clusters. Young clusters may not be old enough to have developed embryonic subclusters; older clusters, like Hamilton A, have members with RCCs in excess of 20 and contain subclusters.

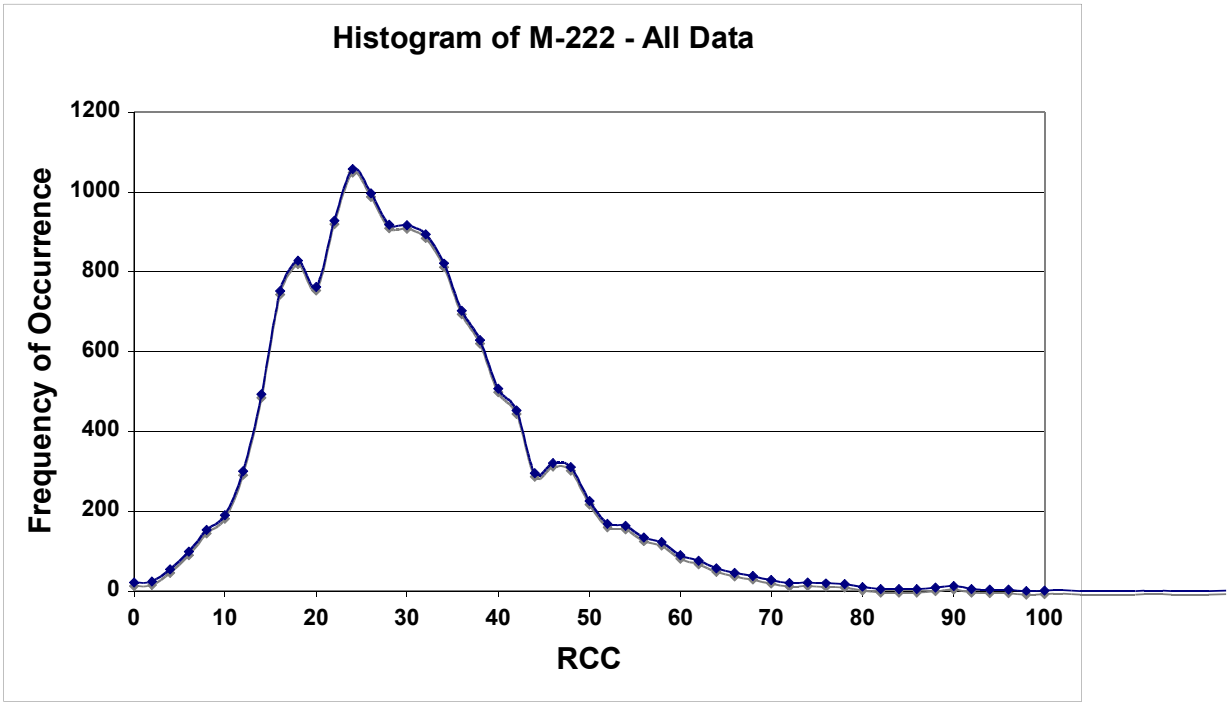
7. Analyses of other surname clusters have shown that cluster membership seldom contains pairs of testees with RCC greater than ~ 20-25 (800-1000 years). The average RCC of the three Logan clusters is 6.3, one-third of that RCC "limit," giving independent confirmation that the Logan clusters have two-thirds of their recognition time as independent clusters yet to go.

8. The clusters of today grow in the following way. When close members of a family get tested, they will show up in the matrix as a very young cluster–a subcluster, which will have an average RCC very close to zero. When more and more people get tested (many of whom will not know each other, but who share a recent common ancestor), their matrix cluster will grow and its average RCC will increase. This process will continue until the cluster fills up to a practical upper bound of RCC ~20. Thus recently formed subclusters of closely-related testees will either be recognized as separate entities within the matrix or they will occur within already existing clusters.

9. The times derived may need to be revised by a scale factor in the future as more data become available, but the time scale shows no evidence of non-linearity over at least several thousand years. In any event, the evolutionary sequences derived from these analyses appear to be well defined.

10. An analysis of surname clusters whose members belong to different haplogroups can lead to the epoch in the past when the common ancestor of these two lines existed. This will be the time at which the haplogroups merged as we go back in time. For example, the oldest age indicated by the

a



b

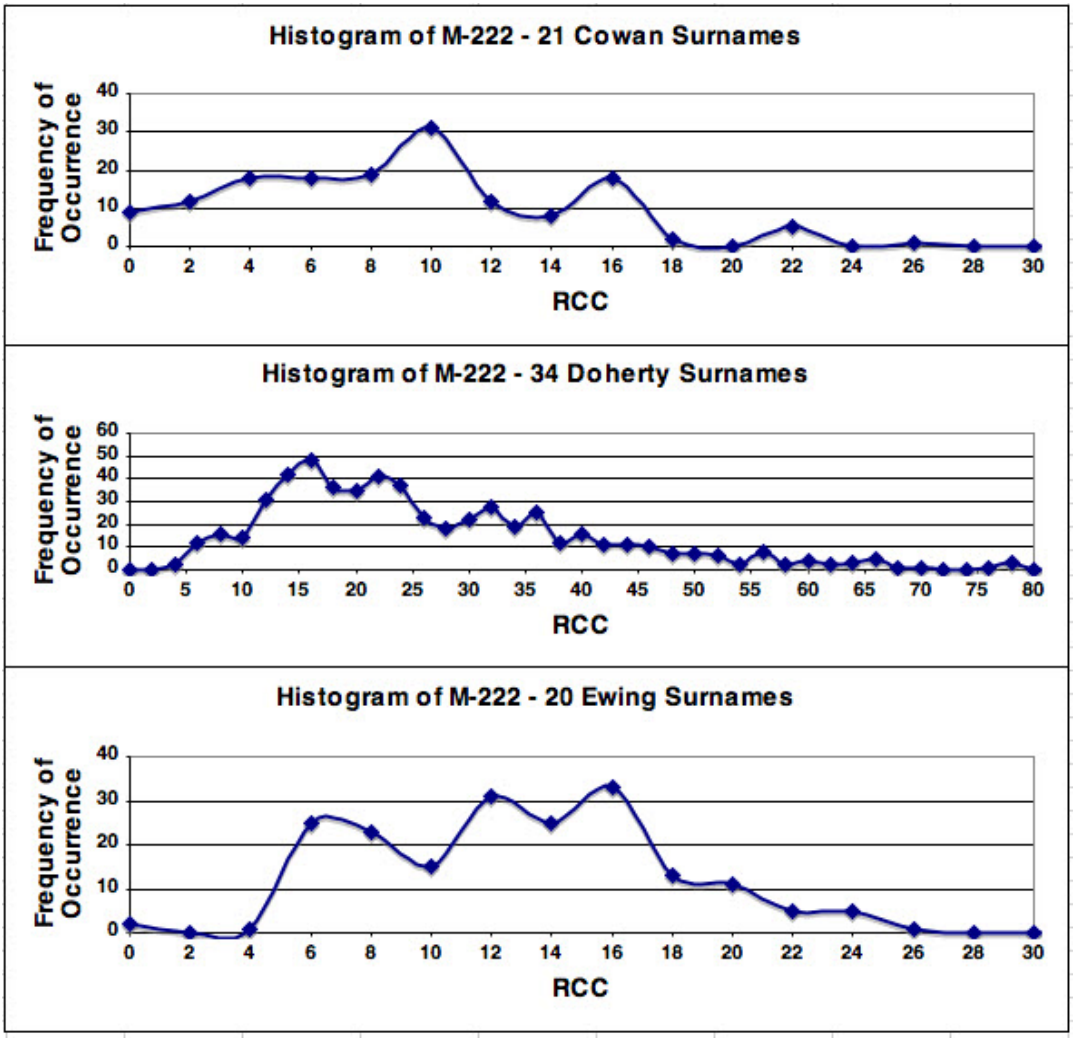


Figure 6. (a) Histogram of RCC values for the entire M222 matrix, (b) histogram of RCC values for three M222 surnames.

maximum RCC found in any Hamilton intercluster we studied, members of which are in the same haplogroup, is about 5000 years. The youngest age indicated by the minimum RCC found in a Hamilton intercluster, members of which are in the different haplogroups I1 and R1b, is about 7900 years. Therefore, the epoch at which these two haplogroups evolved away from their common ancestor was between 5000-7900 years ago (RCC between 115-182).

The Variation of The Standard Deviation of Matrix Groups as a Function of the Average RCC of the Group

During this analysis we have studied clusters where all members belong to the same haplogroup (values of RCC< 20-30) and interclusters where members belong to the same or to different haplogroups (values of RCC> 20-30). To investigate how the standard deviations of values of RCC vary among these matrix groups, we have plotted the standard deviation of the distribution of RCC values in each surname and group type in 100 groups (within five surnames) against the average RCC value of that group. The plots of all these relationships never depart significantly from linearity. We next explore the relationship between the slope of the each plot and the average RCC of the group.

There are two different types of variation at work in a cluster (viz., different MRCAs among the pairs of testees and the uncertainties caused by mutation randomness), whereas only one type of variation affects an intercluster (viz., the statistical variation of mutation randomness back to the single CA of the intercluster). Therefore, one might expect the ratio of SD to the average RCC of a group to be larger in clusters (RCC<20-30; two sources of variation) than that same ratio in interclusters (RCC>20-30; one source of variation).

Table 3 lists the derived slope (SD/Average RCC) of each cluster and intercluster. It is immediately apparent that the ratio of SD/Average RCC is high when RCC is low, and conversely.

Since RCCs for surname clusters are less than ~20, percentage uncertainties of 55 to 75 percent is consistent with the difficulty of determining TMRCAs for testees within a surname cluster who lived within time intervals of genealogical interest. The percentage uncertainty in time becomes significantly less when we try to estimate the TCAs of the intercluster regions or the times at which significant stages of evolution occurred among pairs of haplogroups. Analytically, a low percentage uncertainty works to our advantage when we investigate haplogroup evolution many thousands of years ago.

D. A Study of the SNP, M222

David Wilson, group administrator of the R-M222 Haplogroup Project, notes that this haplotype marker set is associated with many individuals whose roots lie in the counties of Northwest Ireland, Ulster and Lowland Scotland (Wilson, 2008). The marker set now known as the R-M222 group was first recognized in late 2004. It was noted that family names associated with the cluster were almost entirely Irish or Scottish. The original research team called this pattern the Irish Modal Haplotype and provocatively suggested that the haplotype was to be associated with the U' Neill kings of Northern Ireland who descended from the fifth century warlord, Niall of the Nine Hostages.

We undertook this study primarily because the M222 SNP indicates that there is a progenitor of the group who was the first to have this SNP. Thus we have a definite point of origin to which every male in his lines of descendant can be referenced, but we do not know *a priori* when he lived. Although family names came into

Table 3
The Slope of the SD/RCC Relation by Group Type

SURNAME	No. Clusters	Group Type	RCC Range	SD/Avg.RCC
Hamilton	9	Same Haplotype	1to12	76%
M-222	18	Same Haplotype	1 to 30	67%
Logan	6	Same Haplotype	2 to 10	62%
Ewing	12	Same Haplotype	1 to 10	56%
Cook	10	Diff. Haplogroups	188 to 298	22%
Logan	10	Interclusters	40 to 116	14%
Hamilton	35	Interclusters	29 to 312	9%

being many years after the progenitor lived, a study of those names, whose haplotypes are distinct from one another, can reveal the sequence of surname formation, and the approximate time when a surname originated, and perhaps when the progenitor lived.

Table 4 shows that the RCCs of the entire M222 matrix and the intercluster regions are indistinguishable from each other. Moreover, the oldest surname groups have RCC values that strongly suggest that their common ancestor lived close to the time that the M222 mutation appeared, about 1450 years ago, or about CE 500. Ken Nordtvedt has independently determined that the date of origin for M222 was about 1740 years ago, in good agreement with our determination, especially considering the date uncertainties in Column 6 of Table 4 (Nordtvedt, 2008). This RCC matrix is unique among all those studied because of its homogeneity, the uniqueness of the intercluster M222 time of origin, and the time of the oldest surnames in the group. The date of origin lends some credibility, but does not prove, that the appearance of the M222 clade can be traced to Niall of the Nine Hostages who lived close to where the concentration of the M222 descendants appears today. The dates associated with the surnames suggests a sequence of surname evolution that runs down the list of Table 4. The statistics of the surnames Ferguson, Daugherty-Doherty and Kelly indicates that they are older than the surnames Howle, Dunbar and McGonagill.

We anticipate that, since this group shares a set of common markers and has a narrow location in both space and time, the RCC matrix will be uncomplicated. Figure 6 shows histograms of the cluster and intercluster regions of the M222 matrix. As expected, both the cluster and intercluster regions are very simple, with no pairwise mismatches. The M222 matrix consisted of 172 testees, grouped by surname. Eighteen different surnames, but no surname that contained less than three individuals, were included in the matrix. Finally, statistical parameters were determined for the surname clusters that contained a minimum of five individuals. The results of the study are presented in Table 4 where the surnames are ranked by their average RCC. This ranking suggests the approximate order in which the surnames came into use.

In the case of M222, we are confronted with a unique situation that differs from the analysis of a typical surname cluster:

Table 4
Results of the M222 Surname Study: Surnames are Clusters (Time in Years Ago, 1 RCC = 43.3 Years)

SURNAME	NUMBER	Average RCC	SD	Years Ago 52.7*Avg RCC	Years Ago (From SD)	Avg Years Ago	Est Year (CE)
Ferguson	5	37.2	20.5	1962	2095	2029	-84
Daugherty	6	27.9	13.3	1471	1362	1416	529
Kelly	5	27.5	8.2	1449	838	1143	802
Doherty	34	25.6	14.2	1349	1448	1398	547
McCord	9	22.4	15.9	1179	1618	1398	547
McLaughlin	6	21.3	10.3	1121	1051	1086	859
Wilson	6	13.2	6.9	697	705	701	1244
Ewing	20	11.9	5.0	627	510	569	1376
Cowan	21	10.6	6.9	556	707	632	1313
Burns	6	10.1	5.9	534	599	567	1378
McGonagill	8	5.1	5.5	267	560	414	1531
Dunbar	6	3.8	2.0	200	204	202	1743
Howle	5	2.9	1.5	153	151	152	1793
Entire Matrix	172	29.1	13.5	1534	1377	1455	490
Intercluster Region	116 (equiv)	30.1	12.9	1586	1316	1451	494

Journal of Genetic Genealogy, 5(2):271-288, 2009

1. We are analyzing groups of surnames that all share a particular SNP that indicates a unique time of origin.

2. The RCC matrix is relatively plain but there is a hint of non-homogeneity because of the structure at RCCs near 20 and 48.

3. The average value of RCC for the entire matrix is equal to that of the intercluster region, and,

4. The earliest surnames (viz., Ferguson, Daugherty-Doherty and Kelly) appear to date back to very near the origin of the M222 SNP.

We can think of the M222 group as a "supercluster" that contains identifiable clusters (viz., areas in the RCC matrix where testees share a common surname within the larger group that shares the M222 SNP) and intercluster regions (viz., areas in the matrix that contain paired members of clusters that have been identified and regions of surname pairs who have not been clearly identified as members of identifiable clusters). Just as a cluster contains pairs of haplotypes with values of RCC of the order of 30 or less, we see that the M222 supercluster contains values of RCC of the order of 70 or less.

It is clear that the Nordtvedt date of origin of the M222 snip (1740 years ago, an RCC equivalent of about 40 where the matrix is 80 percent filled) falls within the distribution of RCCs in Figure 6. It is less than one standard deviation from the earliest surname date (Ferguson).

Topics for Future Study

There are at least four areas where further work using the correlation matrix technique may yield further insight and significant understanding. They are:

1. Investigating the evolutionary sequence of the ISOGG sequences to see if they converge into the sequence of evolution of the clusters in the correlation matrix, as we predict;

2. Calibrating (and refining) the steps in the ISOGG sequence in terms of the time distance among haplogroups as the ISOGG sequence becomes better defined;

3. Teasing out differences among the various marker mutation rates when the number of pairs of RCC values considerably exceeds the number of different markers that undergo mutations in a surname group; and,

4. Testing the linearity of the RCC time scale through more extensive studies of early haplogroups.

A. A Comparison of the RCC Results with the ISOGG Haplogroup Tree

Figure 5 shows an evolutionary sequence for three Logan clusters, starting from a common ancestor and progressing through an intercluster sequence to a present group of testees, many of whom belong to different clusters. The evolutionary sequence bears a strong resemblance to the evolutionary sequence in the Y-DNA Haplogroup Tree of the International Society of Genetic Genealogy (ISOGG) (Note 4). Recognizing this resemblance, we present the following assessment based on the schematic in Table 5.

The haplotype of a testee, much like a fingerprint, is a characteristic of that individual. Downward from the ancestor's male line, the haplotype changes slowly as markers mutate to their final present day configuration, providing a better identification. There is a striking

Table 5
A Comparison of the ISOGG Haplogroup Sequence with the Evolution of the Logan Clusters

I - The ISOGG Sequence (schematic only - R1b1b2 is an arbitrary starting point)							
R1b1b2 =>	R1b1b2a	=>	R1b1b2a1	=>	R1b1b2a1a		
		=>	R1b1b2b	=>	R1b1b2b1	=>	R1b1b2a1b
		=>	R1b1b2c	=>	R1b1b2c1	=>	R1b1b2a1c
II - The Sequence in the Figure (after Figure 5)							
Common Ancestor	=>	Intercluster 1-3	=>	Cluster 1	=>	Testee A	
	=>	Intercluster 2-3	=>	Cluster 3	=>	Testee B	
	=>	Intercluster 1-2	=>	Cluster 2	=>	Testee C	

parallel between the Logan sequence (II in Table 5) and the evolutionary sequence in the ISOGG tree (schematically shown in I, Table 5). Both sequences start with a haplotype that will also change slowly as it evolves toward the haplotype of a testee.

The ISOGG sequence is being continually improved by adding more detail to its evolutionary paths. As more Logans are tested, the Logan sequence will be better determined. We predict that soon, the more detailed designations of a haplotype in Sequence I will converge toward the haplotype in Sequence II until the two sequences merge into one. The time scale provided by the RCC approach holds promise at tying the two processes together, with future time scale refinements of either one serving to refine the time scale of the other.

Virtually all the Logans considered here are in Haplogroup R1b1b2 (That is why we chose that starting point in the ISOGG sequence, above, beginning with the 6th subdivision of R and ending with the 10th, drawing an analogy between the starting point and the three subdivisions shown in Sequence I and the three steps in Table 5's Sequence II). According to the 2009 version of the ISOGG types, the R haplogroup has now been subdivided so that some subclades consist of over 12 R subdivisions (e.g., R1b1b2a1a2d3a). If the prediction made above is valid, we are close to finding that the Logan Clusters 1, 2, and 3 are included in, or are very closely connected to, one or more of the detailed ISOGG subdivisions of the R haplogroup.

Both the RCC time scale and the associations of its cluster sequences with one or more deep haplogroup sequences are testable.

The real test of a new theory or hypothesis is the extent to which it can explain existing phenomena at least as simply as previous hypotheses can, and the degree to which it can make testable predictions. The correlation technique of analysis meets both criteria.

B. RCC Results with the ISOGG Haplogroup Tree

Since values of RCC are proportional to a time distance between haplotypes, RCC values should also be proportional to a time distance between pairs of haplogroups. The haplogroups in the ISOGG tree are identified by the letters, A through T. Haplogroups are subdivided into one or more levels, called subclades, forming a tree. The appropriate Y-chromosome haplogroup is assigned by performing a sequence of SNP tests. The presence of one or more particular SNPs defines the subclade. Let us define an ISOGG "step" as the addition or subtraction of one level to the clade. For example, to go from haplogroup I2b to J2 takes five steps; you go from I2b to I2 to I to IJ to J to J2. In this section we have used the approach of Karafet et al. (2008), to indicate how the correlation technique might be applied in the future to refine these relationships.

Since each evolutionary step takes time, it is reasonable to assume that the number of steps taken between clades will be correlated with values of RCC among separated pairs of clades. In a very preliminary investigation, we analyzed a typical mixture of haplogroups from the Cook surname project (Cooke 2009) and attempted to see how strongly the number of steps required to go from one haplogroup, through the ISOGG Haplogroup sequence (Note 4), to the other haplogroup correlated with the differences in RCC among the haplogroup pairs. The haplogroup pairs were in Haplogroups I, J, and R; the average RCC values ranged from 188 to 298; and the number of steps ranged from 2 to 12. The correlation found was 0.80, a reasonably strong correlation.

The reason why this is an important topic for future study is that the ISOGG Haplogroup sequence is being filled in and rearranged continuously as new SNPs are discovered and we predict that the correlation we found here will improve as we know more about the details of the ISOGG sequence. We caution that this result is based on only one surname (Cook), and a very limited range of haplogroups, so the correlation should be viewed only as suggestive. Nevertheless, if this sample of Cook haplotypes is typical of others, we have shown how the use of the RCC and its associated time scale may give us insight into the evolution of haplogroups through changes that have taken place in their subclades. The relationship needs further exploration and refinement as the ISOGG Haplogroup sequence becomes better defined. We anticipate that future work will lead to more definitive results.

C. RCC Values Associated Marker Locations - Average vs. Individual Mutation Rates

Jim Logan has studied the specific differences in marker values for individual testees, called "leafs", in what he calls Limb 3 of his Logan tree (Logan, 2008). This is Cluster 3 in Table 1 and Figure 5. In the group he has listed the specific marker locations that change as we go from testee to testee on the tree. For example, Kit number 54727 differs from Kit number 44163 through changes in DYS 458, DYS570 and CDYa (viz., three marker changes). For each pair of testees in this group we divided the RCC of the pair by the number of marker locations that changed between the pair. We then made a histogram of these ratios. Table 6 gives the results of that distribution.

The histogram shows a broad range of values, little skewness and remarkably good agreement among the average, median and mode of the distribution. Seventeen percent of the ratios were clustered within a 0.1

interval of RCC/mutation at 2.6. Column 3 indicates that an average mutation takes about 110 years (SD=34%), in reasonable agreement with the value indicated in Table 3 of Part 1.

There were 33 different markers involved in the study of this one limb, and over 100 pairs of RCC values. Therefore, it should be possible to tease out differences among the various marker mutation rates from the 33 parameters and over 100 equations using a least squares approach. Since there are at least two other Logan limbs for which similar information is available, this represents a challenge for a future study as well as an opportunity to compare the differences between limbs. We know the average mutation rate over all markers, but it should be possible to determine individual mutation rates, using the RCC time scale as intermediary points of comparison.

D. RCC Values Associated with Early Haplogroups - Is the RCC Time Scale Linear?

The RCC time scale has been calibrated using over a hundred pedigrees. It is useful over a few thousand years, but its applicability to haplogroups, whose origins were in the more distant past, needed to be assessed. In this study, high values of RCC (100-800) were found that indicated ages of haplotype pairs well beyond those of genealogical interest. But the groups we had studied had not contained any haplotypes within Haplogroups A or B. According to the ISOGG Haplogroup sequence, "Y-DNA haplogroup A represents the oldest branching of the human Y chromosome tree, thought to have begun about 60,000 years ago. Like Y-DNA haplogroup B, the A lineage is seen only in Africa and is scattered widely, but thinly across the continent" (Note 4).

Because of the interest of geneticists in earlier epochs, a very preliminary study was made in which RCC-derived ages of paired modal haplotypes of various haplogroups were compared with ages found in the ISOGG (Note 4). We met with only limited success, mainly because a sufficient number of ages have not yet been well defined. However, the study showed (1) a suggestive positive correlation between RCC and the age estimates that were available, and (2) an indication that an extrapolation of the time scale into regions of genetic and other scientific interests could be made without encountering major problems of non-linearity.

An attempt was then made to find haplotype pairs with very high RCCs that would indicate a very distant relationship between the oldest haplogroups and well-evolved haplotypes. The ISOGG Haplogroup sequence suggests combining haplotypes in Haplogroups A and B as the oldest of a pair. We took 37-marker haplotypes from ySearch and from FTDNA's Haplogroup projects for haplogroups A and B and matched them with the haplotypes of surname pairs for which high values of RCC had already been derived in this study. Preliminary results on 46 pairs with RCC >800 included the following observations:

1. All 46 highest-RCC pairs have one member in Haplogroup A.

2. The largest value of RCC found was 1203, pointing to a time about 52,000 years ago.

3. The oldest pair was a combination of Haplogroup A with Haplogroup C.

4. The oldest value of Haplogroup B paired with Haplogroup A had an RCC of 811, pointing to a time about 35,000 years ago.

Table 6
Statistics of Specific DYS Changes Between Pairs of Testees in Logan Limb 3 (Cluster 3)

Statistic	Value	Years (1 RCC = 43.3)
Average RCC/Marker Change	2.53	110
Median RCC/Marker Change	2.58	112
Mode (RCC/Marker Change)	2.58	112
Number of Testees	15	
SD of Distribution (RCC/Marker Change)	0.87	
Skewness	-0.01	
Kurtosis	-0.22	

5. The oldest value of Haplogroup B paired with a lower haplogroup E1b1b1)A had an RCC of 653, pointing to a time about 28,000 years ago.

6. This work suggests the times when Haplogroup B evolved from Haplogroup A and when other lines branched off from Haplogroup B.

7. There is no evidence of serious non-linearity in the RCC time scale.

Although the RCC vs. Time relation was calibrated using many pedigrees over the first millennium, the association of a member of Haplogroup A with an RCC-derived value that is only about 13 percent lower than the earliest reference to the presence of Y-DNA in the literature, 60,000 years ago, is in reasonable agreement with its ISOGG-derived time. This observation strongly suggests that the RCC-time scale can be used for epochs useful to geneticists and scientists in other fields whose research includes these epochs in time. Since mutations are implicit in this technique, there is indirect, but strong evidence that average 37-marker mutation rates have not changed significantly over these long periods of time.

While these arguments do not exclude the possibility of unknown systematic errors, the derivation of the RCC time scale and its applicability over distant times in the past is arguably the most important product of this study.

Final Remarks

The time scales derived here may contain uncertainties that are of the order of 20-30 percent, and could be higher if unrecognized systematic errors are present. Relative times in an RCC sequence should be more trustworthy. The time scales we use result from more than one approach which are internally consistent within those error bars. Hence, they may be better than the time scales used by genetic genealogists who employ more traditional techniques. Attention is still needed to improve the accuracy of the time scale and improve its precision.

The correlation approach uses an average mutation rate implicitly. In the future, when mutation rates become better known, the correlation coefficient could be determined by using weights appropriate to different mutation rates. Such an approach would be more mathematically intensive, but it lies well within the capability of most small computers.

After further research we may find the average mutation rates of strings of 37 markers are not the same in groups with different surnames, or in groups that are located in different parts of the world, or in groups who live under differing environmental conditions. They may have changed with time. If any of these conditions are found to be true, our approach will require a modification only to the average mutation rate. Such a modification will be less difficult to apply than having to modify many individual marker mutation rates, all of which may have changed with time. There is no evidence that such modifications are needed at this time.

If the number of markers is at least 37, if the time scale is calibrated using the same number of markers, if the marker DYS comparisons are the same for each haplotype analyzed, and if the comparisons are done on the same number of markers, one does not have to look at individual marker differences to reach the same conclusions described here. In fact, current ways to match markers not only are more time-consuming, but their conclusions regarding time scales are not as broad or as far-reaching as the ones that use the correlation approach. Moreover, the time scales suggested in this analysis should be testable through future work in related areas of research.

The correlation techniques developed here can be applied to any pair of haplotypes, sub-haplogroups or haplogroups regardless of differences in surnames or haplogroup designations. The RCC time scale bridges times of interest to genealogists and geneticists. If ties to Y-DNA can be found, this correlation technique may also be of practical use in fields that make time comparisons among events that occur in mitochondrial DNA, migration patterns, linguistic patterns, geology, anthropology, archeology and paleontology.

Acknowledgments

I wish to thank Gordon Hamilton, the Hamilton DNA surname project administrator, for our many conversations and for his valuable suggestions as my development of the correlation technique and the study of the Hamiltons progressed. I wish again to thank the many people mentioned in Part 1 for early discussions that encouraged me to pursue this new approach to Y-DNA analysis. Helpful comments and suggestions on this paper were gratefully received from others, including particularly David E. Hogg, Jim Logan, and Elizabeth B. Waltman.

Web Resources

Cook/Cooke/Koch & Variants DNA Project
http://www.familytreedna.com/public/cook/default.aspx?section=yresults

Logan DNA Project
http://www.familytreedna.com/public/LoganDNAProject/default.aspx

R-M222 Haplogroup Project
http://www.familytreedna.com/public/R1b1c7/default.aspx

References

Cooke J (2009) Cook/Cooke/Koch & Variants DNA Project. See Web Resources

Fitzpatrick C (2005) *Forensic Genealogy*, Rice Book Press, Huntington Beach, CA, pp. 214-215.

Hamilton G (2008) Hamilton Surname DNA Project. See Web Resources. The web site is continuously updated, but the data used in this study represents the Hamilton database on (date).

Howard WE (2009) The Use of Correlation Techniques for the Analysis of Pairs of Y-STR Haplotypes, Part 1: Rationale, Methodology and Genealogy Time Scale. *J Genet Geneal*, 5:256-270.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase the resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18:830-838. See also: http://genome.cshlp.org/content/18/5/830

Logan JJ (2008) private communication. See also his Logan DNA Project web site in Web Resources.

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

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

Notes

1. As we showed in Part I, an RCC from zero to 1200 corresponds approximately to the period from the present back in time to about 52,000 years ago. In his book, Deep Ancestry, Spencer Wells (National Geographic Society, 2006) points out that Y-Haplogroup A, the oldest of the Y-DNA haplogroups, dates back to about 60,000 years. If our RCC scale is linear, 60,000 years ago would correspond to an RCC of 1390.

2. For readers who are used to thinking in terms of marker differences, the sum of the absolute value of the marker differences between Hamilton Groups A and B are 3 at 12 markers, 8 at 25 markers, 18 at 37 markers, and 28 at 67 markers. The marker difference indicates that the observed marker distance definitely falls within the interval appropriate to both Hamilton groups being within a sub-haplogroup.

3. Caution is needed in the interpretation of clusters and groups that contain numbers of individuals of less than about four. Pairs of groups that contain small numbers are meant to be suggestive. More testees are needed in the group to increase the reliability of the results.

4. The International Society of Genetic Genealogy (ISOGG) publishes and periodically updates a phylogenetic chart that shows an evolutionary sequence of haplogroups, together with a list of one of more Single Nucleotide Polymorphisms (SNP (SNPs) that define the haplogroup. See http://www.isogg.org/tree/ISOGG_YDNATreeTrunk09.html