Dating Y-DNA Haplotypes on a Phylogenetic Tree: Tying the Genealogy of Pedigrees and Surname Clusters into Genetic Time Scales

-- William E. Howard III and Frederic R. Schwab --

ABSTRACT:

An RCC matrix (Howard 2009a), resulting from a new correlation approach to analyze Y-DNA haplotypes, is used in conjunction with a standard Mathematica application program to produce a dated phylogenetic tree. The program displays the evolutionary relationships among all haplotypes in the matrix; it groups closely related surnames into family clusters that correlate well with genealogical pedigrees. The time scale assigned to the tree is monotonic, linear, and dates the evolutionary relationships of Y-DNA testees that may go back tens of thousands of years. This study is arguably the first to investigate the time relationships between surname Y-DNA haplotypes, pedigree- and RCC matrix-derived surname clusters and their associated phylogenetic tree. It offers a straightforward methodology and a uniform time scale that can also be used to estimate the evolutionary relationships among Y-DNA haplogroups.

INTRODUCTION:

A new approach to analyze Y-DNA haplotypes has been introduced (Howard 2009a, Howard 2009b). An RCC time scale, calibrated with over 100 pedigrees, has been developed that can be applied over tens of thousands of years to investigate the evolutionary relationships that tie genealogy and genetics together by analyzing surname clusters and haplogroups.

More recently these correlation techniques and genealogical pedigrees have provided additional insight into the history and evolution of the Gordon surname (Gordon and Howard 2012, hereafter called the Gordon paper). That paper presented relationships among pedigrees, surname cluster and subcluster membership, and possibly geographic location. Available pedigrees are associated with membership in subclusters and clusters. The ISOGG time estimates[1], the RCC time scale, and the Y-DNA evidence can be used to suggest places where surnames originated. The times derived for early Gordon surname migration are consistent with the history of those places derived from archaeological excavations, genetics and anthropologic studies. A comparison of the ISOGG dates and the RCC time scale shows good agreement and no inconsistency between the RCC- and ISOGG-derived estimates. Moreover, there is no evidence of non-linearity, and the time scale appears to be valid for use over time intervals out to over 50,000 years, making it potentially valuable when genetic considerations are important to

[1] The ISOGG's Y-DNA Haplogroup Tree can be found at <http://www.isogg.org/tree/index.html>. It is being continually updated.

studies of mitochondrial DNA, migration and linguistic patterns, geology, anthropology, paleontology and archeology.

The traditional process of analyzing family groups and clusters from Y-DNA haplotypes is challenging. If the number of testees is small, the results are statistically ill-determined and therefore ambiguous; if the number is large, the analysis becomes arduous. While the description of the evolution of surname clusters appears to be correct in the above three papers, the process lacks a quick ordering of the cluster groupings as well as a good presentation of the time scale of the evolution. To provide that confirmation, we looked into the availability of a computer program that would have one or more of the following attributes:

1. It would group entries into family surname clusters and subclusters.
2. It would derive evolutionary relationships among the clusters and subclusters.
3. It would derive a phylogenetic tree from the evolutionary relationships.
4. It would attach a time scale to the phylogenetic tree.
5. It would be unambiguous in its presentation.
6. It would execute quickly.
7. It would be available at reasonable cost.
8. It would be easy to use.

One of us (FRS) uses the computational software program, Mathematica (Wolfram 2010) in support of research related to radio astronomy and suggested that its built-in package for hierarchical clustering and dendrogram generation would address some of the problems we wished to solve. This paper describes the use of this package, including preparation of the input data, and the success we have had with it in confirming and extending previous results.

THE MATHEMATICA PROGRAM – DATA PREPARATION:

The result of a Y-DNA test is expressed as a haplotype or marker string, which, in this study, is 37 markers long. When the haplotypes of *n* testees are selected for study, we use the correlation function (e.g., Excel for Mac 2004) to derive an *n* by *n* matrix of correlation coefficients that is then converted to an RCC matrix (Howard 2009a, Howard 2009b). That *n* by *n* matrix is inserted into Mathematica which produces a phylogenetic tree.[2]

APPLICATION OF THE MATHEMATICA PROGRAM:

The so-called agglomerative hierarchical clustering methods, as implemented in *Mathematica,* begin with a square *n* by *n* distance, or dissimilarity, matrix, *D*, the *i-j* element of which contains a measure of the distance *d(i,j)* between objects *i* and *j* (i.e., in our case, the *i*th and *j*th set of marker data for the *n* individuals) according to some appropriately defined metric. The Euclidean distance might be chosen, or the correlation distance, $d(i,j)=1-\rho(i,j)$ where ρ is the Pearson correlation coefficient. In our case we

[2] Note in proof: We now have a code that will produce a phylogenetic tree directly from haplotypes.

choose the RCC matrix, with elements $10^4(1/\rho(i,j)-1)$, as defined in the earlier references (Howard 2009a, Howard 2009b), and for which we have an associated time scale.

Next, for splitting the data into clusters, a measure of intercluster distance needs to be defined; this is specified via the so-called linkage option. There are a number of standard choices for the linkage option, seven of which are built into the Mathematica clustering package[3]; it can also be user-defined. We have selected Mathematica's *linkage->``average"* method, which coincides with what is known in the literature as the *unweighted pair group method using arithmetic averages* (UPGMA). See Press et al. 2007. The hierarchical clustering proceeds in a bottom-up fashion, beginning with a list consisting of *n* singleton clusters. Then the following steps are repeated *n*-2 times:

(1) find the two nearest clusters, according to the intercluster distance measure associated with the chosen linkage option;
(2) create a new cluster that agglomerates the two;
(3) update the active cluster list accordingly; and
(4) go back to (1).

One ends up with a phylogenetic tree, the structure of which can be displayed in a so-called dendrogram plot.

A good textbook reference, in addition to Press *et al. (*2007), is Everitt (2001). And full algorithmic details, including analyses of computational complexity, are given in Day & Edelsbrunner (1984).

We first used the full 187 by 187 RCC matrix that contained all 37-marker Y-DNA Gordon surname test results, grouped by haplotype, as input to Mathematica[4]. The initial output, using the program's default setting, presented a well-formed phylogenetic tree. The program grouped the testees into exactly the same clusters derived in the Gordon paper. The output contained the Kit Numbers of each testee, many of which are not members of the major clusters[5]. Some were in minor clusters; some in no clusters at all.

The program output using the full matrix did not initially contain a time scale, but it clearly fulfilled the first three attributes we were seeking. To investigate the time axis of the tree, we compared the average value of each Gordon cluster with its height along the

[3] The seven options for the linkage method are: single, complete, average, centroid, median, weighted average, and Ward. The precise definitions of these options are not provided in the Mathematica documentation. Clear definitions are given by Day and Edelsbrunner (1984), as well as in the documentation for the Matlab hierarchical cluster package, available at <http://www.mathworks.com/help/toolbox/stats/linkage.html>.
The plots with the selection of average, centroid, or weighted average were considered to have time, or RCC axes closer to what we might have expected.

[4] In April 2010, the cutoff date of the Gordon paper analysis, 242 Y-DNA results were available in the Gordon surname project (Note 4). We selected only those results where testees had been tested at 37 or more markers and we use only the first 37 markers to form the correlation matrix and then the RCC matrix (Howard 2009a).
This process narrowed the analysis to 187 individuals from which we were able to group 119 testees (64%) into well-defined Gordon clusters and subclusters (viz., clusters within a cluster) in the RCC matrix.

[5] See the Gordon DNA Project at <www.TheGordonDNAproject.com>

ordinate[6]. We found that the scale was linear ($R^2 = 0.9997$), so values of RCC could be safely assigned to the ordinate, presenting a time axis in which 10 RCC ~ 433 years (Howard 2009a).

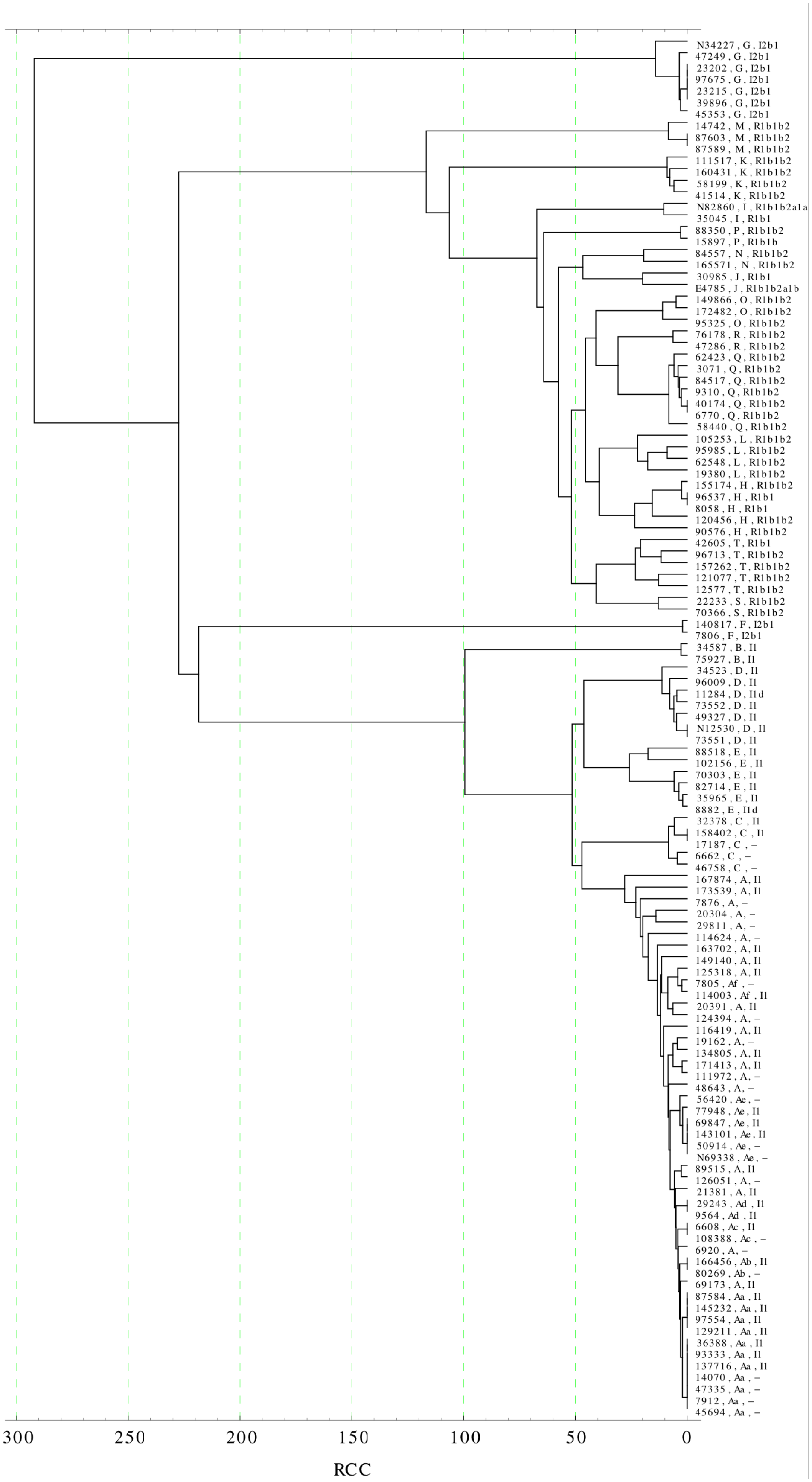
A comparison of each of the seven trees with the evolutionary diagrams in Figures 4A and B of the Gordon paper showed that the average linkage option gave the best agreement. This result was not unexpected since averages were used in the Gordon paper to derive the intercluster relationships. To investigate how well the time axis of the tree in the average linkage option fit the data in the RCC matrix, we measured the lengths along the RCC axis of the intercluster junctions and compared them with the average values of each Gordon intercluster at their RCC junction points found in the Gordon paper[6]. We found that the relation was:
RCC (tree)= 0.962 x (RCC, matrix), with $R^2 = 0.826$.

This agreement between the RCC values for the tree and the matrix RCC values for the interclusters gave us confidence that the average linkage was the option to be used. Thus we conclude that the RCC matrix, using Mathematica with the average linkage option, will yield a phylogenetic tree that will show the evolution of the modern Gordon surname at least ten times further back in time than pedigrees and cluster designations can do. Moreover, this application of the Mathematica cluster analysis program will show the evolution of any surname matrix for which a sufficient number of testees is available.

We next generated a tree using the average link option only for the 119 specific clusters described in the Gordon paper. Figure 1 shows the results.

Figure 1: The Phylogenetic Tree Produced by Using the 119 by 119 RCC Matrix with Mathematica's Average Linkage Option for all Gordon Y-DNA 37-marker Haplotypes that formed clusters in the Gordon paper[4]. The ordinate shows for each testee his Kit Number, Gordon Cluster assignment, and Haplotype. The abscissa shows the RCC time scale (10 RCC ~ 433 years).

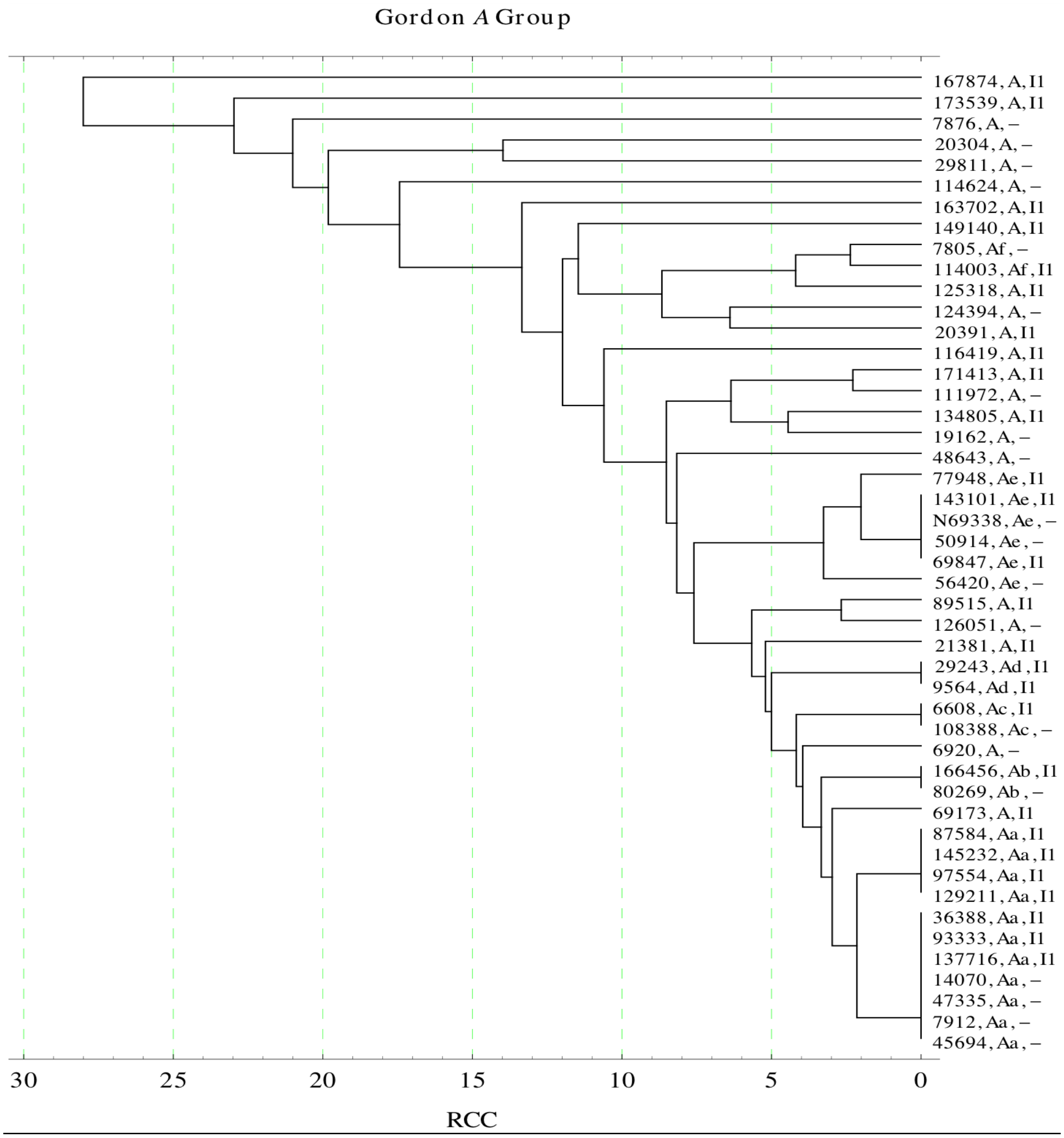
[6] We compared the RCC values from the matrix in the Gordon paper with measured values of RCC from the phylogenetic tree and derived this relationship. One might expect that there should be an exact comparison of these values, but the presence of small measuring errors on the graphs combined with the difference in the way Mathematica derived the tree led to the small discrepancies noted here. Those discrepancies are small compared to the effect of unknown mutations, especially at low values of RCC present in the time interval of genealogy.


N34227 , G , I2b1
47249 , G , I2b1
23202 , G , I2b1
97675 , G , I2b1
23215 , G , I2b1
39896 , G , I2b1
45353 , G , I2b1
14742 , M , R1b1b2
87603 , M , R1b1b2
87589 , M , R1b1b2
111517 , K , R1b1b2
160431 , K , R1b1b2
58199 , K , R1b1b2
41514 , K , R1b1b2
N82860 , I , R1b1b2a1a
35045 , I , R1b1
88350 , P , R1b1b2
15897 , P , R1b1b
84557 , N , R1b1b2
165571 , N , R1b1b2
30985 , J , R1b1
E4785 , J , R1b1b2a1b
149866 , O , R1b1b2
172482 , O , R1b1b2
95325 , O , R1b1b2
76178 , R , R1b1b2
47286 , R , R1b1b2
62423 , Q , R1b1b2
3071 , Q , R1b1b2
84517 , Q , R1b1b2
9310 , Q , R1b1b2
40174 , Q , R1b1b2
6770 , Q , R1b1b2
58440 , Q , R1b1b2
105253 , L , R1b1b2
95985 , L , R1b1b2
62548 , L , R1b1b2
19380 , L , R1b1b2
155174 , H , R1b1b2
96537 , H , R1b1
8058 , H , R1b1
120456 , H , R1b1b2
90576 , H , R1b1b2
42605 , T , R1b1
96713 , T , R1b1b2
157262 , T , R1b1b2
121077 , T , R1b1b2
12577 , T , R1b1b2
22233 , S , R1b1b2
70366 , S , R1b1b2
140817 , F , I2b1
7806 , F , I2b1
34587 , B , I1
75927 , B , I1
34523 , D , I1
96009 , D , I1
11284 , D , I1d
73552 , D , I1
49327 , D , I1
N12530 , D , I1
73551 , D , I1
88518 , E , I1
102156 , E , I1
70303 , E , I1
82714 , E , I1
35965 , E , I1
8882 , E , I1d
32378 , C , I1
158402 , C , I1
17187 , C , –
6662 , C , –
46758 , C , –
167874 , A, I1
173539 , A, I1
7876 , A, –
20304 , A, –
29811 , A, –
114624 , A, –
163702 , A, I1
149140 , A, I1
125318 , A, I1
7805 , Af , –
114003 , Af , I1
20391 , A, I1
124394 , A, –
116419 , A, I1
19162 , A, –
134805 , A, I1
171413 , A, I1
111972 , A, –
48643 , A, –
56420 , Ae , –
77948 , Ae , I1
69847 , Ae , I1
143101 , Ae , I1
50914 , Ae , –
N69338 , Ae , –
89515 , A, I1
126051 , A, –
21381 , A, I1
29243 , Ad , I1
9564 , Ad , I1
6608 , Ac , I1
108388 , Ac , –
6920 , A, –
166456 , Ab , I1
80269 , Ab , –
69173 , A, I1
87584 , Aa , I1
145232 , Aa , I1
97554 , Aa , I1
129211 , Aa , I1
36388 , Aa , I1
93333 , Aa , I1
137716 , Aa , I1
14070 , Aa , –
47335 , Aa , –
7912 , Aa , –
45694 , Aa , –
300
250
200
150
100
50
0
RCC

A subset of Figure 1 can be used to present the subclusters of the Gordon Cluster A. This result is presented in Figure 2. It is particularly interesting for several reasons:

1. It contains evolutionary information over time intervals of genealogical interest.
2. It contains a large number of testees within the single Gordon Cluster A.
3. It contains subclusters that show evolutionary branching, and,
4. It contains the best pedigree information outlined in the Gordon paper.

Figure 2: A Detail of the Gordon Cluster A in Figure 1[4]. The abscissa shows the RCC time scale (10 RCC ~ 433 years). The ordinate shows for each testee his Kit Number identification, his Gordon Cluster and Subcluster assignment and his haplotype.

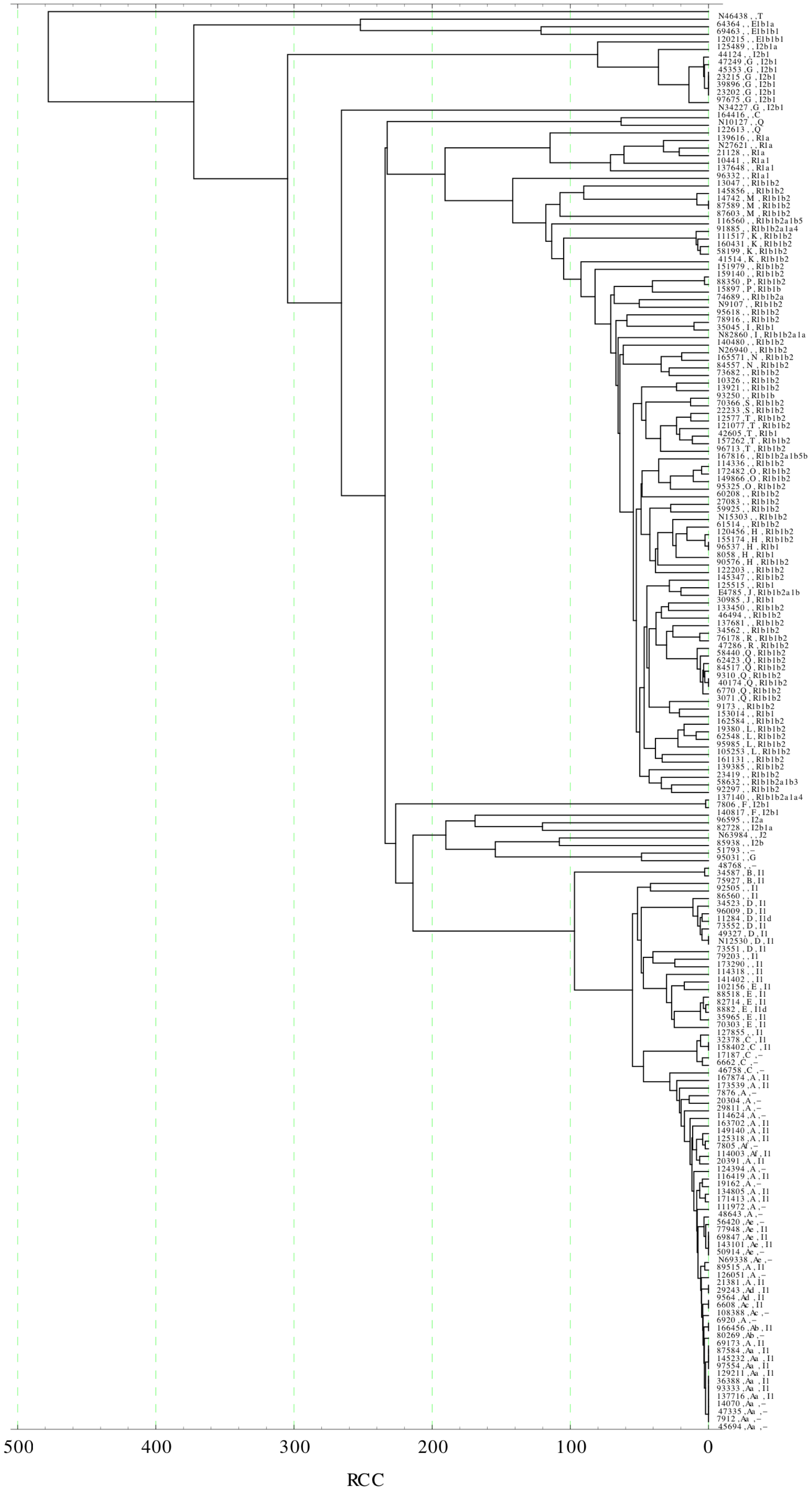


At the bottom right of Figure 2 we find that Gordon subcluster Aa actually consists of two sub-sub clusters (Kit numbers 45694 to 36388 and Kit numbers 129211 to 87584). The existence of the Gordon A subclusters was not recognized originally on the Gordon FTDNA web site and the existence of the two sub-sub clusters was not originally recognized in the RCC matrix (Gordon and Howard 2011). The remarkable ability to recognize and resolve surname clusters with such high resolution shows the power of the hierarchical clustering methods as implemented in Mathematica.

To investigate how well the time axis of the Gordon Cluster A tree in the average linkage option fit the data in the RCC matrix, we measured the lengths along the RCC axis of the intersubcluster junctions and compared them with the average values of each Gordon intersubcluster using values of the RCC junction points found in the Gordon paper[6]. We found that the relation was:
RCC (tree)= 1.0213 x RCC (matrix), with $R^2 = 0.48$.
Again, we find that the time relationship is linear, but the scatter is larger than for the full matrix due to mutation uncertainties over this short time interval of about 1100 years. For the pedigree relationships in the subclusters Gordon Cluster A, see the Gordon paper.

The 68 testees who were not designated as cluster members can still be presented in a phylogenetic tree. Although not in clusters, their positions on the tree with other testees gives valuable insight into points in the tree where they share common ancestry with cluster members and all other testees. With few exceptions, their TMRCAs lie at RCC > 20-25, as expected. Figure 3 presents the phylogenetic tree for the full 187 x 187 RCC matrix.

Figure 3: The Phylogenetic Tree Using the full 187 by 187 RCC Matrix and Mathematica's Average Linkage Option for all Gordon Y-DNA 37-marker Haplotypes[4]. The ordinate shows for each testee his Kit Number, Gordon Cluster assignment, and Haplotype. The abscissa shows the RCC time scale (100 RCC ~ 433 years).


N46438 , ,T
64364 , , E1b1a
69463 , , E1b1b1
120215 , , E1b1b1
125489 , , I2b1a
44124 , , I2b1
47249 ,G , I2b1
45353 ,G , I2b1
23215 ,G , I2b1
39896 ,G , I2b1
23202 ,G , I2b1
97675 ,G , I2b1
N34227 ,G , I2b1
164416 , ,C
N10127 , ,Q
122613 , ,Q
139616 , , R1a
N27621 , , R1a
21128 , , R1a
10441 , , R1a1
137648 , , R1a1
96332 , , R1a1
13047 , , R1b1b2
145856 , , R1b1b2
14742 , M , R1b1b2
87589 , M , R1b1b2
87603 , M , R1b1b2
116560 , , R1b1b2a1b5
91885 , , R1b1b2a1a4
111517 , K , R1b1b2
160431 , K , R1b1b2
58199 , K , R1b1b2
41514 , K , R1b1b2
151979 , , R1b1b2
159140 , , R1b1b2
88350 , P , R1b1b2
15897 , P , R1b1b
74689 , , R1b1b2a
N9107 , , R1b1b2
95618 , , R1b1b2
78916 , , R1b1b2
35045 , I , R1b1
N82860 , I , R1b1b2a1a
140480 , , R1b1b2
N26940 , , R1b1b2
165571 , N , R1b1b2
84557 , N , R1b1b2
73682 , , R1b1b2
10326 , , R1b1b2
13921 , , R1b1b2
93250 , , R1b1b
70366 ,S , R1b1b2
22233 ,S , R1b1b2
12577 ,T , R1b1b2
121077 ,T , R1b1b2
42605 ,T , R1b1
157262 ,T , R1b1b2
96713 ,T , R1b1b2
167816 , , R1b1b2a1b5b
114336 , , R1b1b2
172482 ,O , R1b1b2
149866 ,O , R1b1b2
95325 ,O , R1b1b2
60208 , , R1b1b2
27083 , , R1b1b2
59925 , , R1b1b2
N15303 , , R1b1b2
61514 , , R1b1b2
120456 , H , R1b1b2
155174 , H , R1b1b2
96537 , H , R1b1
8058 , H , R1b1
90576 , H , R1b1b2
122203 , , R1b1b2
145347 , , R1b1b2
125515 , , R1b1
E4785 , J , R1b1b2a1b
30985 , J , R1b1
133450 , , R1b1b2
46494 , , R1b1b2
137681 , , R1b1b2
34562 , , R1b1b2
76178 , R , R1b1b2
47286 , R , R1b1b2
58440 ,Q , R1b1b2
62423 ,Q , R1b1b2
84517 ,Q , R1b1b2
9310 ,Q , R1b1b2
40174 ,Q , R1b1b2
6770 ,Q , R1b1b2
3071 ,Q , R1b1b2
9173 , , R1b1b2
153014 , , R1b1
162584 , , R1b1b2
19380 , L , R1b1b2
62548 , L , R1b1b2
95985 , L , R1b1b2
105253 , L , R1b1b2
161131 , , R1b1b2
139385 , , R1b1b2
23419 , , R1b1b2
58632 , , R1b1b2a1b3
92297 , , R1b1b2
137140 , , R1b1b2a1a4
7806 , F , I2b1
140817 , F , I2b1
96595 , , I2a
82728 , , I2b1a
N63984 , , J2
85938 , , I2b
51793 , , –
95031 , ,G
48768 , , –
34587 , B , I1
75927 , B , I1
92505 , , I1
86560 , , I1
34523 , D , I1
96009 , D , I1
11284 , D , I1d
73552 , D , I1
49327 , D , I1
N12530 , D , I1
73551 , D , I1
79203 , , I1
173290 , , I1
114318 , , I1
141402 , , I1
102156 , E , I1
88518 , E , I1
82714 , E , I1
8882 , E , I1d
35965 , E , I1
70303 , E , I1
127855 , , I1
32378 ,C , I1
158402 ,C , I1
17187 ,C , –
6662 ,C , –
46758 ,C , –
167874 ,A , I1
173539 ,A , I1
7876 ,A , –
20304 ,A , –
29811 ,A , –
114624 ,A , –
163702 ,A , I1
149140 ,A , I1
125318 ,A , I1
7805 ,Af , –
114003 ,Af , I1
20391 ,A , I1
124394 ,A , –
116419 ,A , I1
19162 ,A , –
134805 ,A , I1
171413 ,A , I1
111972 ,A , –
48643 ,A , –
56420 ,Ae , –
77948 ,Ae , I1
69847 ,Ae , I1
143101 ,Ae , I1
50914 ,Ae , –
N69338 ,Ae , –
89515 ,A , I1
126051 ,A , –
21381 ,A , I1
29243 ,Ad , I1
9564 ,Ad , I1
6608 ,Ac , I1
108388 ,Ac , –
6920 ,A , –
166456 ,Ab , I1
80269 ,Ab , –
69173 ,A , I1
87584 ,Aa , I1
145232 ,Aa , I1
97554 ,Aa , I1
129211 ,Aa , I1
36388 ,Aa , I1
93333 ,Aa , I1
137716 ,Aa , I1
14070 ,Aa , –
47335 ,Aa , –
7912 ,Aa , –
45694 ,Aa , –
500
400
300
200
100
0
RCC

DISCUSSION OF RESULTS:

This investigation is arguably the first to study the time relationships among surname Y-DNA haplotypes, pedigree and RCC matrix-derived surname clusters shown in a dated phylogenetic tree. Our conclusions are based only on the 37-marker haplotype data that met the selection criteria (Gordon 2012)[5]. More data on Y-DNA haplotypes and family pedigrees will lead to more certain ancestral relationships.

Some of the Gordon clusters in Figure 2 have pedigree assignments (Gordon 2012). Since Mathematica's hierarchical clustering algorithm grouped those clusters together just as they were grouped in the Gordon paper, we conclude that as more pedigrees become available, each major cluster will be representative of a particular pedigree line[7]. The TMRCAs of pairs of testees in Figure 2 agree well with those in Figure 4 of the Gordon paper. This result clearly meets the first six attributes we were seeking.

Figure 2 of the Gordon paper showed three peaks in the histogram of all Gordon surname pairs. The first peak has a maximum at RCC ~ 8 and a minimum at RCC ~ 24, corresponding to dates of about 1600 and 900 CE, respectively. These are typical RCC limits for major surname clusters. The phylogenetic trees clearly show surname clusters that each have TMRCAs in that same time interval.

Each cluster has its own TMRCA, and the tree shows estimates of the epochs when the TMRCAs of each pair of clusters lived. These are the intercluster TMRCAs (using the terminology of the Gordon paper). Pedigrees rarely trace to epochs before 1000 CE and this is about the earliest time when the presence of surnames can be traced. However, the inclusion of testees that have RCC connections >20-25 (i.e., those that were not assigned to clusters) shows when these non-cluster Gordons shared TMRCAs with all other Gordons.

Earlier than 1000 CE information on pedigrees is nonexistent, but the Y-DNA record in the phylogenetic tree indicates clearly the approximate dates when these family groups and clusters joined at a progenitor ancestor. Those dates coincide with the second peak in Figure 2 of the Gordon paper, which has a maximum at RCC ~ 48, corresponding to about 2000 years ago (~ 100 BCE). The phylogenetic tree shows the joining of clusters in the time interval between 950 and 12,500 years ago, leading up to the estimated time when the representatives of Haplotypes I and R had their most recent joint ancestor.

These results strongly indicate that when all Gordon testees are included in a phylogenetic tree, the testees that were ungrouped by pedigree or RCC value can still be presented in an evolutionary sequence, except that their TMRCAs with other testees may

[7] We define a major cluster as an RCC grouping that must contain at least four different testees so that at least 6 pairs ((4 x 3)/2 = 6) will be available for comparison. This process led to the identification of 10 major Gordon Clusters, A, C, D, and E (in Haplogroup I1), H, K, L, Q and T (in Haplogroup R1b1b2) and Cluster G (in Haplogroup I2b1). The members of each major Gordon cluster are given in Appendix B of the Gordon paper.

be located further back in time. As more testee results become available, more clusters, more pedigree information and more insight into the evolution of the Gordon surname will be gained.

It is important to realize that, while the surname of these testees is Gordon, the figures trace the relationships of their Y-DNA, through the mutation process, back in time before surnames were chosen. Figure 3 suggests that the progenitor Gordon Haplogroups I and R probably lived about 10-11,000 years ago. Mutations at that time caused these two haplogroups to split, with descendants of Haplogroup I going to Scandinavia, then to the UK, and descendants of Haplogroup R going to Western Europe, then to the UK. Many of these descendants, carrying the mutated markers of the progenitor, began living in Scottish clans or groups without knowing their Y-DNA heritage. The names of those clans converged to the Gordon surname less than 1000 years ago, and their descendants now live in Scotland and all over the world.

The analysis that led to these results for Gordon testees can now be applied with confidence to the analysis of other surnames, other haplotypes, and the haplotypes of different haplogroups. The RCC time scale can be applied to all these associations but mutations will still cause standard deviation errors of the order of 20-40 percent over time intervals of tens of thousands of years (Howard 2009a).

CONCLUSIONS:

Mathematica, together with its hierarchical clustering package, has met the first four of the six desirable goals described in this paper's Introduction. The costs of the program in 2010 vary upwards of $295 USD. Other programs with phylogenetic clustering capabilities are free, but may have ease-of-use issues. The use of Mathematica described here is only a small subset of its program capability. We hope that the description of the process we used to produce the figures in this paper will enable other researchers and surname administrators to use Excel-type programs to form the RCC matrix. The use of an application like Mathematica, in conjunction with the RCC matrix, opens several lines of analysis that have not been available to surname administrators and to the testing agencies in the past[2]. The application:

- Matches haplotypes with clusters much more efficiently than can be done by hand.
- Automatically forms a phylogenetic tree showing evolutionary relationships among all testees, including the TMRCA of individual pairs of testees, the TMRCAs of pairs of clusters, and members in different haplogroups.
- Automatically applies a calibrated time scale to the phylogenetic tree.
- Indicates the times when haplogroups undergo major mutations that spin off new haplogroups.
  - It puts them on a uniform time scale whereas previous efforts tended to estimate haplogroup ages through studies of variance, one by one.
- Encourages comparisons between the genetic chronology of Y-DNA testees and events in human history.

- Allows a better understanding of surname histories and haplotype evolution.

When this series of papers began, we suggested that the correlation approach should be used in conjunction with the traditional approach when surname administrators attempt to group markers by eye. We now have evidence that the correlation approach can produce better, more uniform results. Moreover, when the correlation approach is used in conjunction with an application program like the Mathematica clustering package, it produces not only a time scale, but yields even better results because of its ability to show a higher degree of resolution into clusters, subclusters and, now, even sub-sub clusters. The approach appears to be valid and can be applied uniformly over long time intervals of tens of thousands of years, using an extrapolation derived from over 100 pedigrees. The time scale is linear, and we expect that it can be used to study the evolution of haplotypes in the distant past[8].

FamilyTreeDNA uses a TIP process involving STR marker-specific mutation rates to estimate the time to the most recent common ancestor of any designated testee pair within a specific number of generations based on their STR marker strings[9]. To convert this to time, FTDNA suggests using 25 years per generation. The FTDNA method contains possible errors in each of the 37 individual mutation rates and additional uncertainties in the number of years per generation. Moreover, since the results of a pair of testees, even from very different haplotypes, will have approximate dates connecting their ancestors in the phylogenetic tree, that information on many pairs is missing from FTDNA's phylogenetic tree.

In contrast, the correlation approach adopts an average mutation rate over all 37 markers. Uncertainties in the average mutation rate over all markers will have less influence on the resulting time estimates than in the TIP process. It enables the calculation of time relationships between any testee and all other testees in the RCC matrix or on the phylogenetic tree. The applications program places the testee relative, not to just another testee, but to all testees in the matrix or tree. This process of showing how a testee relates in time and evolutionary position to every other testee on the phylogenetic tree is a decided advantage to this new approach. It estimates time, not probabilities[10,11].

[8] Since this paper was submitted, we recognized that the junction points on the tree corresponding to dates 20,000 or more years ago may be underestimated and that the tree did not always indicate the same sequence of phylogeny that the ISOGG tree reports. Since the dates in this paper refer to more recent times, the dated positions on the tree out to 20,000 years should be approximately correct. Work to investigate trees within haplogroups that have TMRCAs located at earlier dates is in progress.

[9] For a description of the TIP process used, see the FTDNA web site at https://www.familytreedna.com/faq/answers/default.aspx?faqid=9#913

[10] Mutations are used explicitly in the FTDNA TIP process and they are implicit, but still present in the correlation approach. The TIP process is proprietary, but the probabilities included in their estimates are probably of the same order as the uncertainties in the correlation method expressed in years.

[11] On page 871 of Press et al, 2010 it is stated that an *ultrametric* tree has the property that the time distance of all testees in a cluster from the MRCA of the members of a cluster is the same for all the cluster testees. This is clearly the case here where the path length on a phylogenetic tree denotes evolutionary time. Further, it was proposed in the early 1960s that accepted mutation rates might be close enough to constant that, at the molecular level, actual evolutionary data might be close to ultrametric, i.e., that there was a "molecular clock." In Rammal et. al 1986, it is stated " An additional superiority of molecular phylogeny

We hope that the description of the process we used to produce the figures in this paper will enable other researchers and surname administrators to use Excel-type programs to form the RCC matrix and use a program such as Mathematica to perform the same types of analyses from the matrix or the tree that we describe in this paper.

FUTURE WORK:

The current status of this investigation suggests the following areas that might be explored:

1. The correlation approach should be used to study the haplotypes of different haplogroups. The time scale on a phylogenetic tree of haplogroups generated by Mathematica should then be compared to the times assigned to the different haplogroups on the ISOGG phylogenetic tree[8].
2. Now that the Mathematica hierarchical clustering program has been shown to generate surname clusters that can be ordered in an evolutionary sequence and can be identified by pedigree line, a more detailed study of individual marker changes should give us additional insight as individual marker differences are found to be associated with differences in cluster membership and their evolution.
3. Work is needed to investigate:
    a. The time relationships between SNPs and STRs (the marker values)[12].
    b. The optimum method to determine the age of a haplogroup using the RCC approach.
    c. The degree of agreement between the phylogeny of the ISOGG and the Mathematica-generated phylogenetic trees.

ACKNOWLEDGEMENTS:

We thank Tei A. Gordon, the FTDNA Gordon surname administrator, who emphasized the need to objectively analyze Y-DNA data and to minimize issues of integrity that involve individual pedigree-grouping relationships. We thank J.J. (Jim) Logan for discussions involving differences in phylogeny seen between the ISOGG tree and the

comes from the existence of molecular clocks (i.e., constancy in time of the rate of change of a given molecule), for which there is considerable empirical evidence. Indeed, at the molecular level, the evolution of a gene is to a large extent a random walk in sequence space, with a well-defined clock….." We believe that such a clock is present and can be used to date haplotype differences, particularly over long intervals of time when mutations tend to average out. We use the average mutations over a relatively large number of markers (37) and we have calibrated the time scale using different families and over 100 pedigrees. Thus we believe that our method is relatively well-shielded from criticism. Moreover, since there is no indication of non-linearity in the time scale, and since the largest RCC (~1300) found in our analyses is close to the time when Y-Adam left Africa, we believe that our RCC time scale is a reasonable "molecular clock".

[12] A SNP is a Single Nucleotide Polymorphism that is used to determine the phylogeny of an evolutionary sequence. See http://www.isogg.org/wiki/Single-nucleotide_polymorphism.

Mathematica-generated tree. We are grateful for discussions about the difficulties that surname administrators encounter when they use traditional methods to group testee results into family clusters.

REFERENCES:

Day, H. E. and Edelsbrunner, Herbert, Efficient *Algorithms for Hierarchical Clustering Methods*, Journal of Classification, 1, 1984, pp. 7-24.

Everitt, Brian S., Landau, Sabine, and Leese, Morven, *Cluster Analysis,* Fourth Ed., Wiley, NY, 2001.

Gordon, Tei A. and Howard, William E. III, *The Evolution of the Gordon Surname: New Insight From Y-DNA Correlations and Genealogical Pedigrees,* Journal of Genetic Genealogy, 2012, this issue.

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

Howard (2009b): Howard, William E. III, *The Use of Correlation Techniques for the Analysis of Pairs of Y-Chromosome DNA Haplotypes, Part II: Application to Surname and Other Haplotype Clusters*, Journal of Genetic Genealogy, 5, No. 2, Fall 2009, p. 271.

Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing,* Third Ed., Cambridge Univ. Press, Cambridge, 2007; see Section 16.4, Hierarchical Clustering by Phylogenetic Trees, pp. 868-883.

Rammal, R., Toulouse, G., Virasoro, M.A., *Ultrametricity for Physicists*, Reviews of Modern Physics, 58, No. 3, July 1986, pp. 765-788.

Wolfram (2010): Wolfram Research, Inc., *Mathematica*, Version 8.0, Champaign, IL, 2010.

AUTHORS AND POINTS OF CONTACT:

William E. Howard III: McLean, VA wehoward@post.harvard.edu
Fredric R. Schwab: National Radio Astronomy Observatory, Charlottesville, VA
fschwab@nrao.edu

Submitted to the Journal of Genetic Genealogy: 6 February 2011
Resubmitted to JoGG, (With Referee Response): 21 August 2011