How to Interpret Entries on an RCC Tree

-- William E. Howard III --

INTRODUCTION:

Once a male has taken a Y-DNA test and after he gives his permission to share, his test results will be listed on a web site along with the results of other testees who are in the same project (e.g., groups of surnames, a haplogroup, SNP dating and ordering, geographical area of the world, etc.). The Y-DNA result consists of a series of numbers (STR markers), each of which is associated with a particular marker site on the Y-chromosome. Those marker numbers can be thought of as "fingerprints" that tend to genetically identify the testee. The more markers tested, the more precise that identification will be. Generally a male will test for 37, 67, and 111 markers. Based on a sample of markers, an RCC tree[1] can be produced that shows the time relationships among the participating testees. This paper shows how an RCC tree can be interpreted.

At the reporting site for the Bean surname[2], you can see the Kit number that uniquely identifies the testee, his earliest ancestor in his father's pedigree line, his Y-haplogroup designation and SNPs[3] that have been found by the testing agency, and the string of markers, called STRs, that resulted from the test. A particular marker string is meaningless until it is compared to the test results of others. That comparison may lead to an estimated date when the most recent common ancestor (MRCA) of a pair, or of a group of testees lived.

THE RCC TIME SCALE:

The Time to the Most Recent Common Ancestor of a pair of testees (TMRCA) is closely related to the correlation coefficient that is found between the marker string values of any two testees. This process is explained in a seminal paper by Howard in 2009.[4] A time relationship was found between the TMRCA and the correlation coefficient (cc) between pairs of marker strings. Over 100 testees who had validated pedigrees in three different surname groups were identified and their TMRCAs and ccs were known. These pedigrees were used to calibrate the time scale by dividing the TMRCA by cc. The correlation coefficient was converted to a new Revised Correlation Coefficient (RCC) so that the RCC in the new time scale between an identical pair of haplotypes is zero and, as RCC increases, its equivalent time increases.[5] The number of years that correspond values of RCC for 37, 67, and 111 marker lengths are summarized in Table 1.

Table 1:

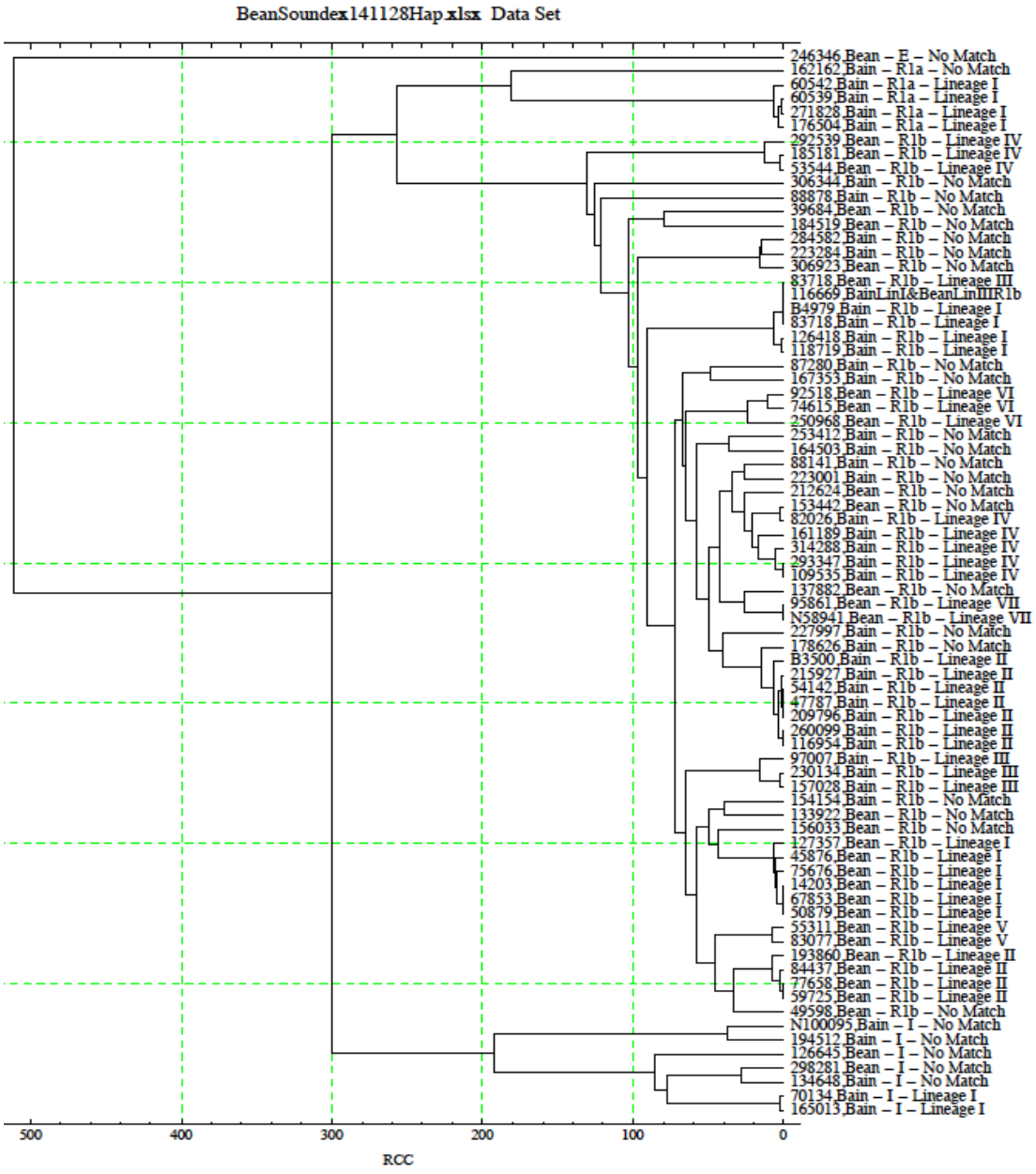
Marker Length	Number of Years per RCC[6]
37	**40.85**
67	**38.05**
111	**34.65**

The standard deviation of the times in the second column is estimated to be about 8 percent.

THE RCC TREE:

A special code developed for the application Mathematica[7] optimizes the average distances of the haplotype matrix of marker strings from each other and places them on an RCC tree. A description of the RCC tree for some Bean and Bain testees follows. Figure 1a shows the RCC tree for 76 Bean and Bain testees available at the time this paper was written.

Figure 1a:



The Kit number identifier is listed to the right of the vertical line at RCC=0.The surname of the testee is then given, followed by his haplogroup, and then the lineage (cluster designation), which is automatically detected and assigned by an administrator function on the Worldfamilies.net web site[8]. Only the surnames Bean and Bain are among this

sample of testees. Note that Figure 1a includes two separate surname listings, one for Bean and one for Bain. There are separate lineage designations for each surname listing (i.e., Lineages I-VII for Bean are separate from Lineages I-IV for Bain).

The inclusion of Haplogroups I, E and R1a introduce TMRCAs that have junction points on the RCC tree that go far back in time beyond RCC ~ 500, over 20,000 years ago. The majority of the testees in the RCC tree belong to Haplogroup R1b.

DESIGNATIONS AND ESTIMATED AGES OF HAPLOGROUPS ON THE RCC TREE:

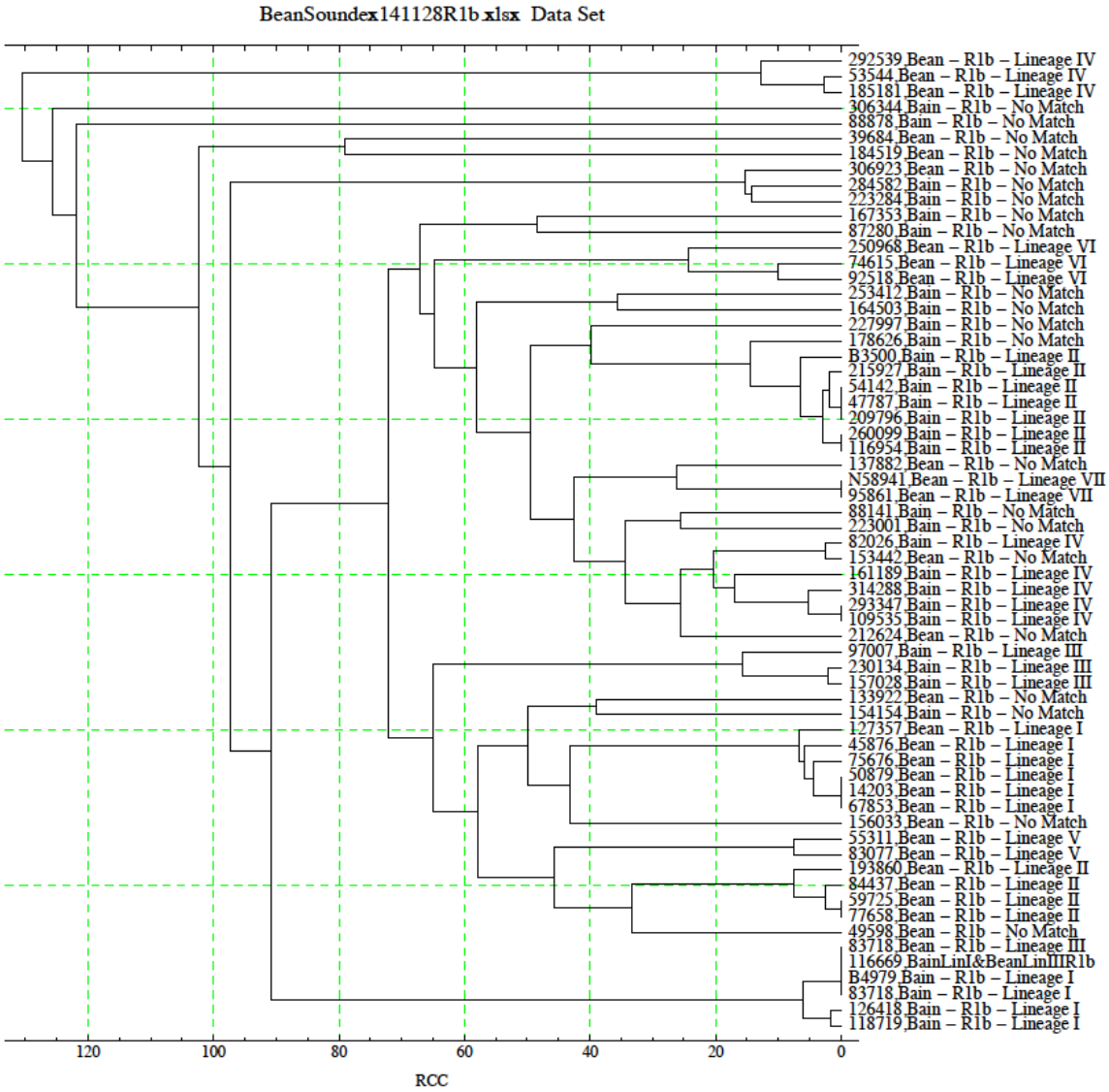
Geneticists have determined the evolutionary sequence of haplogroups, and the ones that appear in Figure 1a are, in order downward on the tree: Haplogroups E (one example at the top of the RCC tree, R1a and R1b (both subgroups of Haplogroup R) and Haplogroup I (seven examples at the bottom of the RCC tree). The last entry is the lineage assigned by the surname administrator (I through VII and No Match).

ISOGG postings estimate the ages at which these haplogroups evolved -- Haplogroup E is the oldest (50-55 thousand years ago), then I (25-30 thousand years ago), and R (~27 thousand years ago) evolved at about the same time. While there is only one example of a Bean in Haplogroup E at the top of the tree, his junction point on the tree with all the others is clearly the oldest on the RCC tree, at RCC over 500. At the bottom of the RCC tree there are seven examples of Beans and Bains in Haplogroup I and they all have a TMRCA at about RCC ~ 197, more recent than the evolution of the Bean in Haplogroup E.

Haplogroup R split into two lines, R1a and R1b. There are only five examples in Haplogroup R1a at the top of the first tree, all Bains. Four of them are in a well-defined cluster, but they all have a MRCA at RCC ~ 180. All remaining testees in this RCC tree are in Haplogroup R1b. Their junction point on the tree suggests they have a MRCA who lived earlier than about RCC ~ 130.

If we limit the entries on the tree only to Haplotypes R1b, we derive the following tree.

Figure 1b:



REGIONS IN THE TREE OF INTEREST TO GENEALOGISTS:

Genealogists who are tracing their ancestry are interested in events that happened from the present back to a time beyond which family records or sources of information in public records have not been found. That end point where the genealogy trail ends is sometimes called a "brick wall." Y-DNA test results can be very useful in breaching a brick wall because a male descendant's individual haplotype string does not change significantly from generation to generation. Individuals who have similar pedigrees will also have similar Y-DNA. They will be grouped in tightly knit clusters on the RCC tree. Because similar marker strings, or haplotypes, indicate family relationships, an RCC tree is valuable because it shows not only the clusters within which a testee appears, but it indicates time relationships among all other testees on the tree.

Most European surnames were adopted between CE 1100 and CE 1400, first by the nobility, then by others. These dates correspond to RCC dates of between 21 and 13, respectively. Some pedigrees date to those times, so we would expect many surname clusters to have MRCAs whose junction points on the RCC tree appear at RCCs at or below RCC ~ 21 when surnames came into use. However, since the Y-DNA of males dates back much further in time, we should expect to see some clusters whose junction points will date even longer ago (i.e., to RCCs beyond ~ 21, and that is what we see in Figures 1a and 1b.

A DETAILED EXPLANATION OF A SELECTED ZONE ON THE RCC TREE:

Let us now look at the groupings of the 14 Beans near the bottom of Figure 1a, between Kit Numbers 127357 down to 49598. Twelve of them are shown Bean Lineages I, II, and V on the tree and members of each lineage are tightly grouped, having junction points less than RCC ~ 8. There are two testees among these 14 designated as "No Match", but their positions in the RCC tree strongly suggest that they belong to specific close-by lineages. The process by which the individual RCC values of all testees are optimized gives us additional valuable information that cannot readily shown by inspection of the marker values. We now look into these RCC tree positions in more detail.

Bean Lineage I consists of the six positions on the tree that lie between 127357 and 67853. The six testees in this cluster are all tightly related with a MRCA at RCC ~ 7 and three of them have identical 37-marker haplotypes. By comparing traditional pedigrees, they probably know how closely they are related. Since the Kit Numbers are assigned chronologically, we may conclude that they were tested at different times[9]. If all the Beans in Lineage I did not know the others prior to taking the test, their tight clustering in the tree provides a powerful incentive to contact the others so that they all can compare their pedigrees.

Below Lineage I there are only two testees in Bean Lineage V (55311 and 83077) and they share a MRCA at RCC ~ 7-8. They can benefit from sharing pedigrees with each other to determine their MRCA.

Below Lineage V there are four members of Lineage II. They share a MRCA at about the same time as Lineages V. Again and V. Again, members of this cluster can benefit by sharing their pedigrees.

Where do the two "No Match" entries belong? The first "No Match" (156033) is more closely related to the cluster Lineage I at RCC ~ 43 (or CE 190), too early for a pedigree comparison. The second "No Match" (49598) while close to Lineage II on the tree, only shares a MRCA with that lineage at RCC ~ 33, again too early for a pedigree comparison.

Since RCC is a time scale, we can infer from these 14 positions on the RCC tree that at RCC ~ 58 (425 BC), the Lineages I, V and II split, with the upper branch evolving at RCC ~ 50 (100 BC) into Lineage I, and the lower branch splitting at RCC ~ 46 (CE 70) into Lineages V and II that then evolved independently[10].

Although only a portion of the RCC tree was selected for this description, the analytic approach can be extended to other portions of the tree.

HOW MUCH CAN WE TRUST THE POSITIONS ON THE RCC TREE?

In determining cluster membership and time relationships of testees on the tree, we assume that average mutation rates apply as haplotype strings evolve. But since we are dealing with mutations that occur at random, we expect that some anomalous marker sites will have evolved more quickly or more slowly than average. The way that Mathematica groups the testees using its hierarchical clustering routine depends on differences in all pairs of entire marker strings and any anomalous difference in mutations will lead to positions and time determinations that are not what we expect from average mutation rates. Thus, mutations will lead to positions on the tree that appear correct but are actually in error. Unanticipated mutations are the major cause of false positions on the tree.[11]

Two important questions need to be addressed: (1) How many unexpected mutations in a Y-DNA line of descent will move a testee out of a cluster on the RCC tree to which he should belong, and (2) how many unexpected mutations will move a testee into a cluster on the RCC tree to which he should not belong? The answers to both these questions will lead to a better understanding of how much credibility should be placed on membership in the clusters that appear in the tree. Since each of the two questions is an inverse of the other, finding the answer to either question is sufficient to finding the answer to the other.

To address these questions, we first selected a group of STR haplotypes of testees whose numbers were large enough so that the Mathematica hierarchical clustering routine would not significantly disturb the positions on the tree when one testee's markers were deliberately changed to see what happens to the positions of that testee on the tree after different changes had been made to selected markers. We chose a sample of 106, 37-marker Y-DNA testee results of the surname Howard. Second, we next selected a well-defined cluster of markers in which one haplotype was chosen to be analyzed further by changing selected markers. Third, the selected Kit Number (101420) had a neighbor in the cluster with an identical haplotype so that we could easily identify the relative effects of marker changes between the selected testee and his neighbor in the cluster. Fourth, we changed marker values and considered the effects of:

(a) Changing one of two markers (DYS 458 and 460) upward by 1, 2, 3 & 4 units,
(b) Changing the same two markers by 2, 3, 4, 5, 6 & 8 units, and
(c) Changing an additional third marker by 3, 4, & 5 units upward in value.

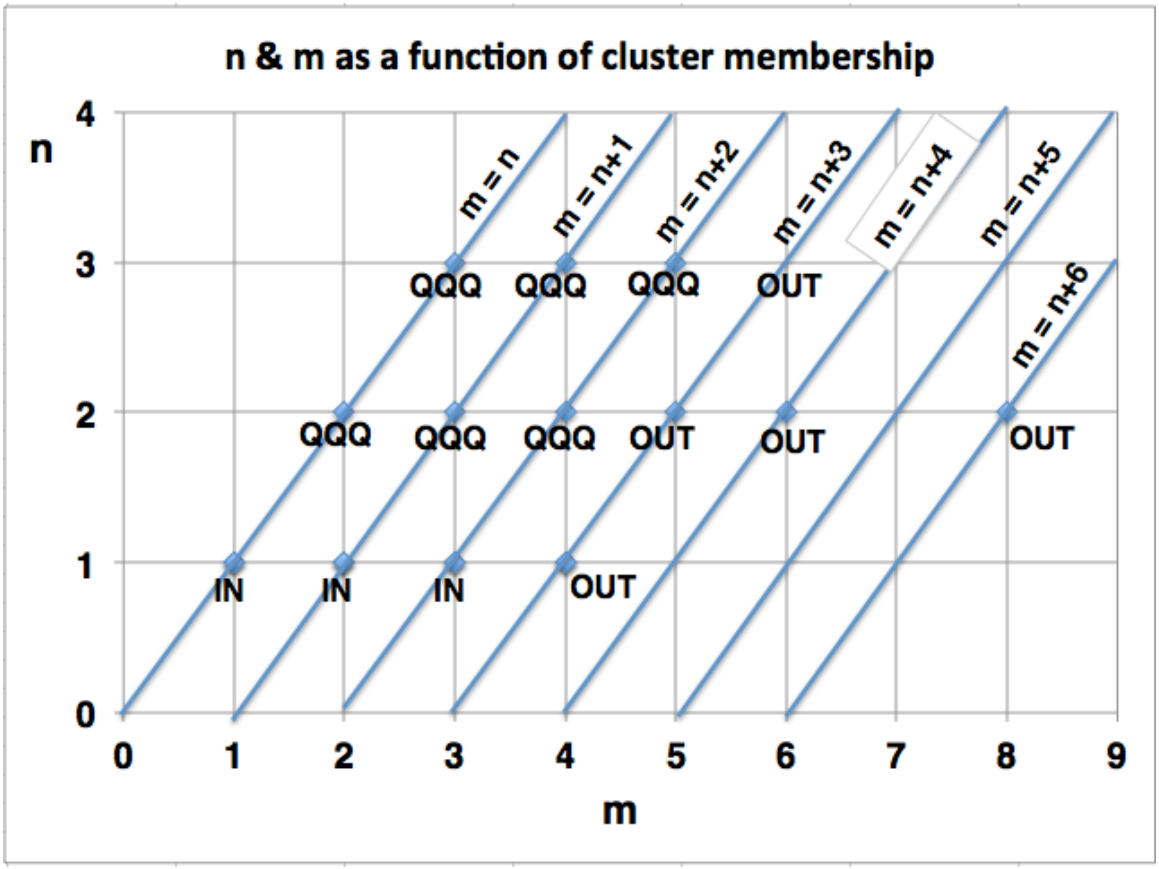
The tables in the Appendix shows the result for cases in which n is the number of markers that were different in a pair of markers, and m is the sum of the absolute values of all maker differences.[12]

Observations:

- When one DYS site experiences a marker change of one, the testee remains in the cluster. The RCC between the changed marker and the original position is very low, well within the time interval in which a pedigree may be found.
- When one DYS site has a marker change of two, the testee remains in the cluster, but is positioned nearer the edge of the cluster. The RCC of about 12 between the changed marker and the original position is about 500 years (or 16 generations) ago.
- When one DYS site has a marker change of three, the testee tends to appear at the edge of the cluster. The RCC between the changed marker and the original position is located at a value of RCC just beyond the time when surnames were adopted. Depending on the validity and extent of the traditional pedigree research, there is a decreasing chance that his MRCA with the rest of the testes in the original cluster can be found.
- When one DYS site has a marker change of four, the testee appears at a very different part of the RCC tree. The RCC between the changed marker and the original position is located at a value of RCC far beyond the time of interest to genealogists. There is no chance that his MRCA with the rest of the testees in the original cluster can be found.
- These observations suggest that:
  - The moved position of the haplotype will remain in the original cluster when n is one and m is three or less. But as m increases, the moved haplotype moves toward the edge of the original cluster.
  - The moved position of the haplotype will be out of the original cluster when m is equal to or greater than n+4.
  - The moved position of the haplotype will be near or at the edge the original cluster when: (1) n is two and m is four or less; (2) n is three and m is 5 or less.

Figure 2 summarizes in chart form the relationships between n, m, and whether or not the moved haplotype remains in its original cluster. n is plotted against m. In the chart there is an indication whether the modified haplotype remains in the cluster, out of the cluster or at a difficult-to-assess (QQQ) position.

Figure 2:



Fifty model runs were used to estimate the expected value of RCC for one mutation. One mutation in a 37-marker haplotype string occurs on average about every 130 to 160 years and it will cause an average change in RCC of about 3.2 to 3.4 (Standard Deviation ~ 15%). The set of 106 haplotypes contained nearly 4000 markers. For each marker, the number of mutations that differ from the average marker value can be calculated and then summed over the 37 markers. Table 2 shows the percentage observed and its comparison with the percentage expected by a Poisson calculation. The two percentages are nearly the same, confirming that random mutations are responsible for the observed results.

Table 2: The number of mutations found in the 109, 37-marker haplotypes.

Number of Mutations (m)	Observed (%)	Poisson Prediction (%)
0	71.0	70.6
1	23.9	24.6
2	4.3	4.3
3	0.6	0.5
4	0.1	0.0
5	0.0	0.0

The entries in Figure 1 and Table 2 suggest the following conclusion:

1. Since 71 percent of the 37 marker haplotypes experienced no apparent mutations (m=0) from the average marker value of that marker, and since 24 percent of the 37 marker haplotypes experienced only one mutation (m=1), about 95 percent of the sample should be placed in relatively correct positions on the RCC tree. The clusters on the RCC tree should be representative of testees who share a most recent common ancestor. If a testee under investigation is placed within a reasonably defined cluster, the probability of his placement being correct is very high.
2. If the number of mutations in the sample is higher (m=2, 3, or 4), the position of a testee under investigation will probably remain near its non-mutated position, most often at the edge of the correct cluster.
3. The probability of mutations greater than 4 will be very rare, so any change in the testee's position on the RCC tree will also be very unlikely.
4. This result should be typical of larger samples of haplotypes being investigated within time periods of genealogical interest.

While this investigation looked at mutations that might take a testee out of a cluster to which he really belongs, the conclusions should apply to the converse of the study (i.e., mutations that might drive a testee into a cluster on the RCC tree to which he does not belong).

USING THE RCC TREE TO DATE GROUPS OF RELATED HAPLOTYPES

We now use the RCC tree in Figure 1b to present the steps necessary to estimate the date when the progenitor of all testees in the sample lived. We take advantage of the fact that convergence theory predicts that if you count the run of junction points in the RCC tree, their numbers can be fitted to an exponential function that will yield that date.

In Figure 1b there are 63 testees on the RCC tree, 18 of which shared one or more identical haplotypes. We use the following steps to estimate the date in Figure 3 when the progenitor of the group lived:

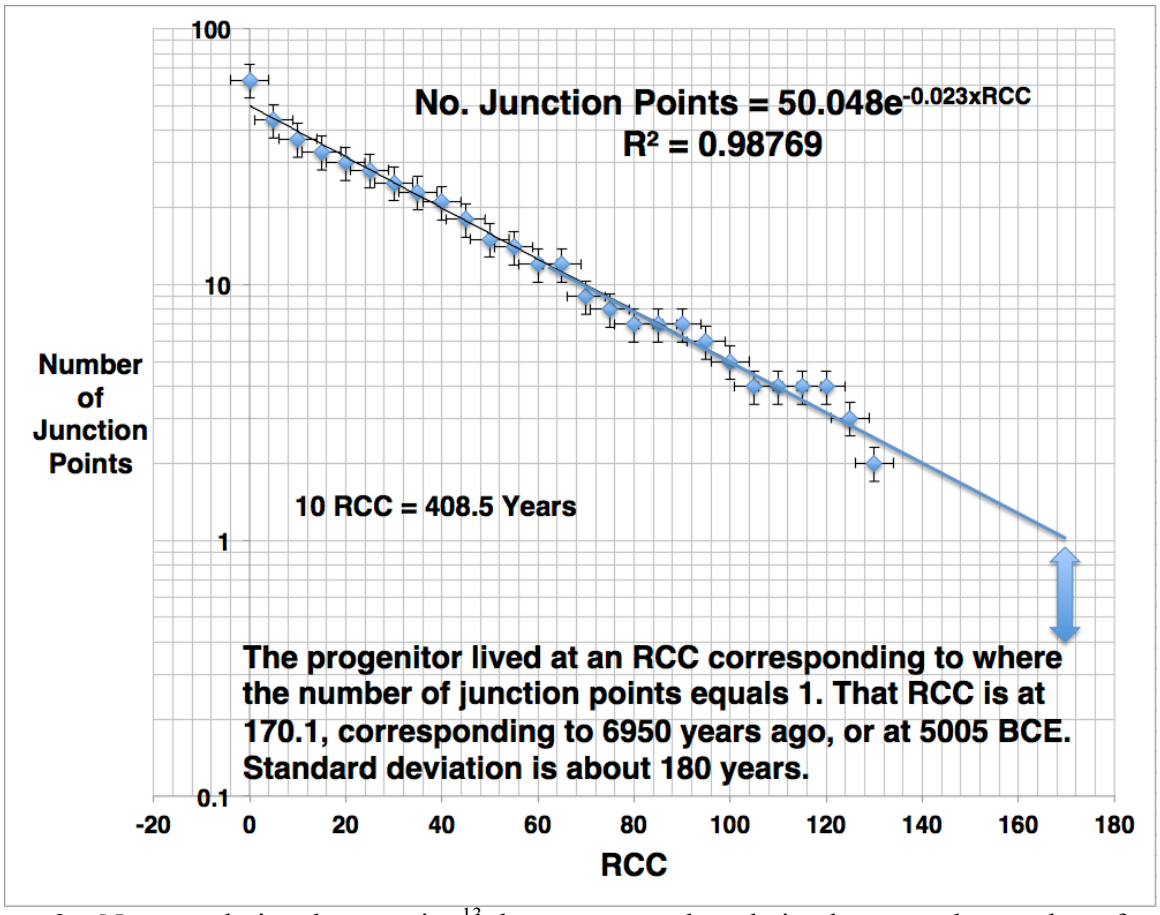
1. In Figure 1b, place a vertical straightedge on the tree at each value of RCC at the bottom of the tree. Proceeding from right to left, at each value of RCC, count the number of junction points each time the straightedge crosses a horizontal line. At RCC 0 the count will be 63, the number of testees. We perform the count at intervals of RCC = 5, from RCC 0 to RCC 145. Table 3 gives the result.

Table 3: The number of junction points (N) in Figure 1b as a function of RCC.

RCC	No. of Junction Points	RCC	No. of Junction Points	RCC	No. of Junction Points	RCC	No. of Junction Points	RCC	No. of Junction Points
0	63	30	25	60	12	90	9	120	4
5	44	35	23	65	12	95	6	125	3
10	37	40	21	70	9	100	5	130	2
15	33	45	18	75	8	105	4	135	0
20	30	50	15	80	7	110	4	140	0
25	28	55	14	85	7	115	4	145	0

2. We then plot the number of junction points as a function of RCC. The result is shown in Figure 3.

Figure 3:



3. Next we derive the equation[13] that expresses the relation between the number of junction points (N) and RCC. As expected, the best fitting equation is an exponential function whose parameters are presented in Figure 3. The fit to the equation is very good, having a variance of 0.98769.
4. Table 2 shows the oldest junction point is found at N = 2. It is the RCC of the oldest pair, but the progenitor would have lived shortly before that time. We need to extrapolate the run of points to estimate the time at N = 1 where a single male, the progenitor, would have lived.
5. We estimate that date by solving the equation for the RCC at which N = 1. That point is at RCC 170.1. Using the relation 10 RCC = 408.5 years, we estimate that the progenitor lived about 7000 years ago, or about 5000 BCE.
6. We note that the point at RCC 0 is high because identical haplotypes are in the sample. The point at N=2 is low, due to statistical fluctuations, but it is within the margin of error.
7. Critique: The run of points in Figure 3 is so tight that error bars may not be needed, but error bars have been added. The error for one point in the ordinate is of the order of 15% (mentioned earlier); the error in the abscissa value of RCC is of the order of 4. However, the error bars assigned to a single point are not the

issue. It is the error bar ensemble for the whole string of points that is important. Since there are about 27 points in Figure 3, the error in the extrapolation to N=1 is of the order of 15/(sqrt(27-1)) or about 3%. But errors in the RCC time scale calibration will be larger than 3%, so the dominant uncertainty will be caused by systematic errors, which are inherently unknown.

ACKNOWLEDGMENTS:
The author thanks E.J. Hurley and Sidney Sachs for their comments and suggestions on this paper, which have improved the presentation.

APPENDIX:

Table A: The effect of changes in DYS sites on positions on the RCC tree. (No entry indicates No Change. n is the number of DYS sites that have changed to produce the result. m is the sum of the absolute differences of DYS sites in marker pairs. The RCC of the junction point with the unaltered Kit No. 101420 on the RCC tree is shown in the fourth column)

Table A:

DYS 458 from 17 to:	DYS 460 from 10 to:	DYS CDYa from 36 to:	RCC of Junction Point w/ Kit 101420	n	m	Result of Move of Original Marker From RCC=0 to its New Position on the RCC Tree
18			4	1	1	To RCC 4. Remained in cluster
19			12	1	2	To RCC 12. At edge of cluster
20			27	1	3	To RCC 27. At edge of cluster
21			93	1	4	To RCC 44. Out of original cluster
	11		4	1	1	To RCC 2. Remained in cluster
	12		12	1	2	To RCC 12. Near edge of cluster
	13		26	1	3	To RCC 25. At edge of cluster
	14		60	1	4	To RCC 59. Out of original cluster to extreme edge of a larger cluster
18	11		14	2	2	n=2; m=2,3,&4 appear together in an adjacent cluster, with m=2&3 paired at RCC 3. m=4 joins with m=2&3 slightly higher at RCC 5
18	12		14	2	3	n=2; m=2,3,&4 appear together in an adjacent cluster, with m=2&3 paired at RCC 3. m=4 joins with m=2&3 slightly higher at RCC 5
19	12		14	2	4	n=2; m=2,3,&4 appear together in an adjacent cluster, with m=2&3 paired at RCC 3. m=4 joins with m=2&3 slightly higher at RCC 5
19	13		26	2	5	To RCC 4. Paired with n=3, m=5, below. A separate cluster within larger cluster at original position
20	13		105	2	6	To RCC 5 paired with position n=2; m=8; far out of original cluster to a distant position near the top of the tree
21	14		105	2	8	To RCC 5 paired with position n=2; m=6; far out of original cluster to a distant position near the top of the tree
18	11	37	11	3	3	To RCC 2. Paired with position n=3; m=4; just outside tight original cluster
18	11	38	11	3	4	To RCC 2. Paired with position n=3; m=3; just outside tight original cluster
19	12	37	26	3	5	To RCC 4. Paired with n=2, m=5, above. A separate cluster within larger cluster at original position
19	13	37	93	3	6	To RCC 33 paired another testee; far out of original cluster to a distant position near the top of the tree

Table B: The results in Table A can be summarized in Table B

RCC of the Junction Point with Kit No. 101420	**n**	**m**	**Result of the Move: In or Out of the Reference Cluster?**
4	1	1	In the Cluster
12	1	2	In the Cluster
27	1	3	In the Cluster
4	1	1	In the Cluster
12	1	2	In the Cluster
26	1	3	In the Cluster
60	1	4	Out of the Cluster
93	1	4	Out of the Cluster
14	2	2	Questionable
14	2	3	Questionable
14	2	4	Questionable
33	2	5	Out of the Cluster
105	2	6	Out of the Cluster
105	2	8	Out of the Cluster
11	3	3	Questionable
11	3	4	Questionable
26	3	5	Questionable
93	3	6	Out of the Cluster

[1] The RCC approach to dating Y-DNA results is described in a series of papers that can be found at: https://dl.dropboxusercontent.com/u/59120192/Genealogy/Papers%26TreesIndex.pdf

[2] http://www.worldfamilies.net/surnames/bean/results.

[3] For description of these terms, see the web site of the International Society of Genetic Genealogy (ISOGG) at http://www.isogg.org where definitions, age estimates and importance of haplogroups, Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) are given.

[4] http://www.jogg.info/52/files/Howard1.pdf

[5] The conversion used is RCC= ((1/cc-1)*10^4

[6] Model calculations show that one mutation in 37 markers causes an average change in RCC of 3.185 (SD=21%; SD of the mean =0.4%)

[7] Mathematica is a product of Wolfram Research and is described at: http://www.wolfram.com/mathematica-home-edition/?src=google&129&gclid=Cj0KEQiA5K-kBRDZ9r71gOvlxOMBEiQAwkK52N7NbdS69kMSABCdnrFlomOCkXrKeuPNVNNffhNRIFMaAlLd8P8HAQ. Fred Schwab developed the code that produces the RCC tree from sets of haplotype marker string values that result from the Y-DNA test.

[8] See http://www.worldfamilies.net/surnames/bean/results and http://www.worldfamilies.net/surnames/bain/results

[9] If these testees had had closely grouped Kit Numbers, they might be a biased sample caused by a closely related group of males who all decided to be tested at nearly the same time/

[10] As RCC junction points are seen further back in time, there is an increasing chance that the TMRCA of common surnames might be coincidental because mutations among the 37 marker haplotypes may produce RCC values that are similar to those observed among the sample of testees studied. The chances of anomalous mutations leading to false TMRCA conclusions diminish when haplotypes with longer marker lengths are used in the analysis.

[11] Different combinations of haplotypes and different lengths of marker strings will not yield the same position on the tree relative to the others. The trees will look very similar, but the details will not be identical. This situation can be seen when the individual positions of testees with R1b haplotypes are compared on the two RCC trees shown in Figures 1a and 1b.

[12] A separate study has been made of how marker string pairs having different values of n and m affect the RCC time scale and FTDNA's Tip predictions. See: https://dl.dropboxusercontent.com/u/59120192/Genealogy/Papers/TipPaper.pdf.

[13] The Excel application in Microsoft Office can be used to fit the points to an exponential function.

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

April 3, 2015