Using Y-DNA Haplotypes to Estimate Their Dates of Origin
--- Pitfalls and Prospects ---

William E. Howard III

(Version 14b– August 3, 2014)

As part of a dating investigation I set out to use the RCC correlation approach to estimate the date of origin of a group of 67-marker haplotypes from the R1b-L21 project. Anatole Klyosov sent me a SNP-tested set of 2000 L21 haplotypes that he had dated. The data were identified only by a running number, and my goal was to compare my estimated date of origin with his. The conclusions reached here for this SNP generally apply to any set of well-defined haplotypes. Because three of the marker strings had blank or zero entries, I discarded them (Nos. 930, 1120 and 1939) instead of substituting a value for them. The tree can be viewed at the following website:
https://dl.dropboxusercontent.com/u/59120192/Genealogy/Trees/InitialL21.pdf
(Copy and paste in your browser and keep it open for the following discussion). We will refer to this tree as the "Initial L21 tree" and a short description of the tree and the RCC time scale is found in an endnote.[1]

We each dated this sample using our respective methods, which are significantly different. I used the RCC correlation approach; Klyosov used his Kilin-Klyosov TMRCA calculator 111 ver 1 approach. I found that the initial date for the origin of the SNP using the data that led to the initial tree was significantly older than Klyosov's date. It was wrong, and further study showed why. The crucial difference is how outliers on the tree are treated. Those outliers can be clearly seen at the top of the initial tree. Once I determined which outliers should be included and which should be excluded in the dating procedure, the disagreement was narrowed but still persisted. This paper explores in detail why a dating difference can occur and what must be done to correct it. The rest of this paper explains why our different approaches gave different answers. The paper raises a question about whether the origin of a SNP is determined to be at the time of a first observed mutation or whether the progenitor that first carried the SNP lived further back in time. We will highlight the importance of preparing samples prior to analysis, ending in a summary of the lessons learned.

I used the RCC correlation approach to do the dating and Klyosov used his approach that involves counting mutations and inserting the marker data into a spreadsheet that will automatically calculate the age. Although Klyosov and I did not come to an agreement on the cause of the difference, I learned much from our dialogue that might be useful to those of our community who are beginning to date SNPs being found from the Big-Y results of FTDNA. I will now describe why the initial dating difference occurred.

Let's look at the initial L21 tree. A quick inspection of the tree shows that there are three additional "outliers" at the top. The RCC method of dating the origin of the SNP or haplotype string depends critically on whether or not the outliers should be included in the sample. The three testees at the top of the initial tree have a junction point with all the

rest of the testees at RCC ~290. If some of these outliers were excluded, the junction point would be lower, at a more recent time. As we prepare our sample for an age determination, we have to develop criteria that will help us decide whether to include these outliers. Regardless of the dating method used, be it correlation or another approach, I submit that the date depends critically on how one treats the outliers.

After a number of false starts, I settled on a heuristic method that is both simple and reasonable. After eliminating the three haplotypes with zero or blank markers, I took the remaining 1997, 67 marker haplotype strings and investigated the absolute difference between each marker value and the modal marker value for the DYS site in which that marker resides, often defined as its genetic distance (GD). In this sample, Table 1 shows the resulting frequency distribution.

Table 1:

Genetic Distance (GD)	**Number of Markers in Sample**
0	106163
1	22940
2	3487
3	976
4	189
5	16
6	1
7	24
8	1
9	2
10 and above	0
TOTAL:	133799 markers

By inspection of Table 1 it was decided to exclude haplotypes from the sample that contained a GD of 6 or more. The run of points from GD 0 to GD 5 can be very closely fit to an exponential function with R^2= 0.98, showing that markers beyond GD 5 are undoubtedly outliers that should not be a part of the sample. The additional four discarded haplotypes had a GD of 6 or more. Three of them were on DYS 413 a&b (Nos 530, 1769 and 421); No. 1467 was probably a transcription or testing error on DYS 570. The first three exclusions are a different type of mutation called a RecLOH which is caused by one copy overwriting another copy [2].

Let us specify as Type A an outlier that should be included and Type B as an outlier that should be excluded from the study. The four additional discarded entries are definitely Type B outliers. All but one are RecLOHs. Other outliers in the sample are SNP-tested, and they appear among the markers in the well-behaved exponential distribution with GD below 6. Those outliers are members of Type A and should be included.

After careful study in preparing the sample to be dated, one should first see if the outlier looks like a RecLOH. If it does and if it is an outlier, it is a Type B and should be excluded. IF it is a RecLOH and IF it is NOT an outlier, it would still be a Type A and

we should keep it in the sample. Of all the outliers, only those of Type A appear to affect the date determination; any RecLOH well inside the tree will not significantly affect the date we derive.

The initial tree leads to a faulty date because it included Type B outliers. After we eliminated the Type B outliers using the above criteria, we produced the tree that can be found at:
https://dl.dropboxusercontent.com/u/59120192/Genealogy/Trees/L21Final.pdf
We will refer to this tree later as the "Final L21 tree". See also NOTE ADDED at the end of this paper.

Now, look at distribution of testees on the final tree. Virtually all the many clusters that occur at RCC<20 (viz., groups whose MRCAs lived in times of interest to genealogists) are components of larger and larger groups of clusters whose TMRCAs occur at junction points that appear higher and higher on the RCC time scale. The number of junction points that connect the MRCAs of these larger, older clusters become fewer as we proceed to TMRCAs that converge at the high end of the RCC time scale. This distribution of junction points on the tree is a good illustration of the theory of coalescence[3].

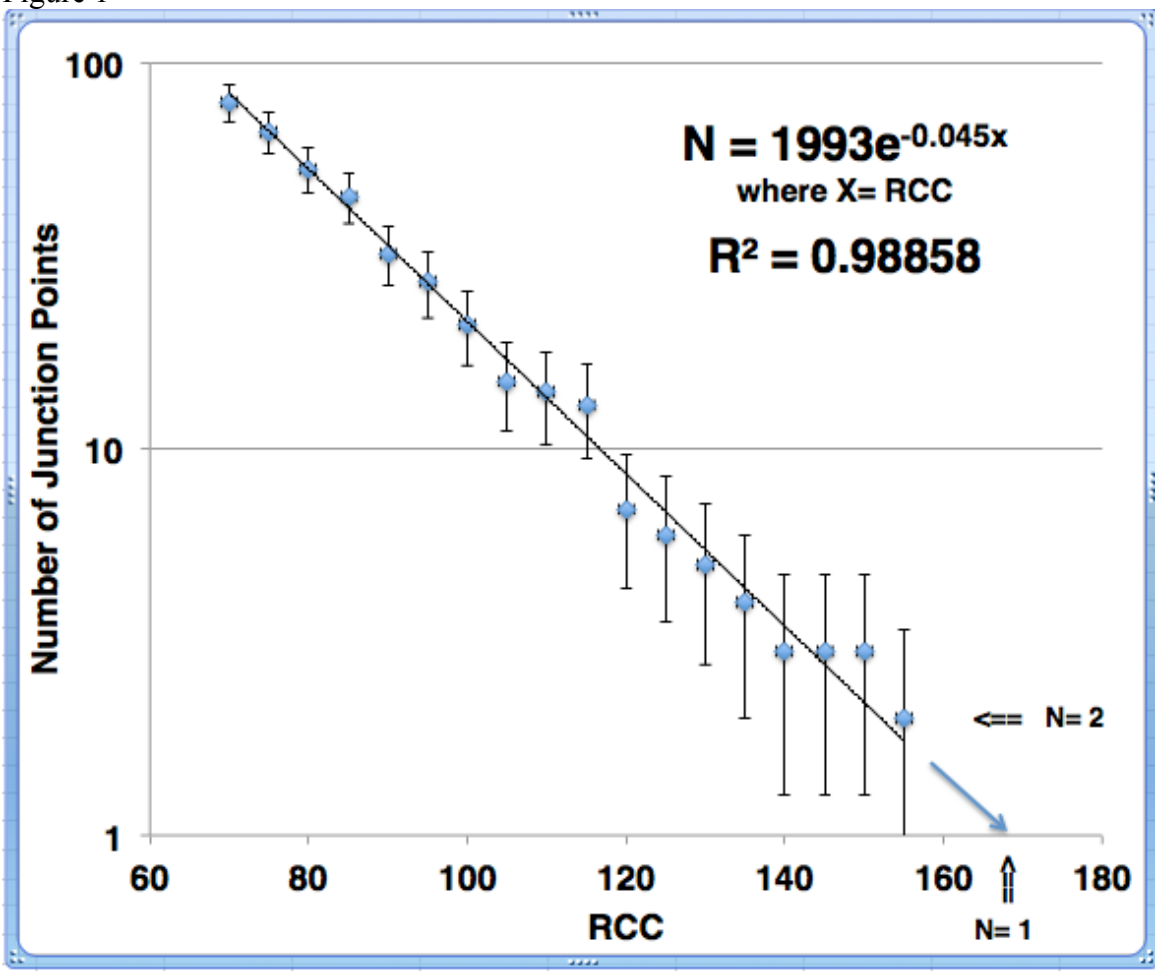
Not only are there major groups of clusters on the tree, there are many subclusters. Those subclusters undoubtedly contain members that are in identifiable subclades that can be separately dated using the same approach used for the complete sample. We pursue this thought further in an endnote.[4]

Our analysis of the junction points on the tree allows us to predict the time when mutations started the branching on the tree. The oldest junction point in the tree occurs at the time when the first mutation we observe caused the earliest branching on the tree. This is the point where the STRs have coalesced. We can now use the distribution of branching points on the tree to estimate the age of the sample. If the sample is large, it should yield a good date estimate. The L21 sample used here is the largest sample I have studied.

There are 1993 testees in the sample. They are all listed along the ordinate at RCC 0 on the tree. If you count the number of horizontal crossings in the tree (N) at RCC intervals of 5, 10, 15, … (i.e., parallels of RCC), a plot of the number of counts (N) vs. RCC will show a convergence toward older times on the tree.

Figure 1 is what we found for this large, final sample of L21 SNPs. The theory of coalescence predicts that if the sample is large, as it is in this case, the run of points can be approximated by an exponential. That is exactly what we observe. The exponential equation that fits this semi-log plot is shown in the figure in which the error bars represent the square root of the number of points counted; its coefficient of determination (R^2) is exceptionally high – a remarkably good fit. For purposes of simplification, only the number count of junction points over RCC 60 are shown.

Figure 1



Now we come to a controversy: Is the point of origin of the SNP at N=2 or is it at N=1? We know that N=2 is the point where we see the oldest mutation in the sample. An extrapolation of the run of points to the value of RCC where N=1 can also be to be the time when the progenitor of the SNP lived. Here are the two conflicting arguments:

ARGUMENT FAVORING N=2:

One side of the argument says that the date of origin of the SNP will be at the point where N=2. Let's look at the coalescence tree construction. Consider any coalescence point. There is a branch segment that goes back toward the root and two branch segments that represent a split in the tree. By analogy, the branch segment that goes back to the root is a father. The branch segment that represents the split going forward in time from the past represents two sons. One of these sons did not receive any mutation that we know of -- his haplotype was presumably identical to his father. On the other hand, one of the sons did receive a mutation from his father. This is the mutation that caused the coalescence split. The mutation occurred in the reproductive process of the father and was not carried by him and could not have been identified in his somatic cells. The mutation occurred at exactly the coalescence point -- not some point back in time that requires extrapolation. Using the exponential equation in the figure, the value of RCC at

the earliest junction point (i.e., the time of the oldest observed mutation at N=2) can be estimated to be at approximately RCC 153.4 (SD estimated at about 7%) or about 5840 (SD ~ 410) years before the average year of birth of the average testee (~ 1945 CE).

ARGUMENT FAVORING N=1:

The other side of the argument says that the progenitor who originated the SNP must have lived somewhat earlier than the first split we see on the tree. While we see the earliest mutation on the tree at N=2, the origin of the split occurred in the reproduction process of someone along the line earlier than the split we see at N=2. We see the first split at N=2 because we cannot see evidence of the SNP before that time. The progenitor who was the originator of the SNP is not necessarily the father of the sons where the mutation we see occurred. The origin of the SNP will be along the male line of descent farther back in time from N=2, but we need some way to estimate how far back it happened. We can estimate the degree of extrapolation needed by noting the tightness of the run of points on the graph, indicated by the fact that R^2 is very close to unity. This argument states that the time of origin will correspond to the value of RCC near where N=1, the original first male that carried the SNP. We can extrapolate the sequence of junction points to that point (or use the equation to determine it). Thus, the origin of the SNP will be the time when the progenitor who first carried the SNP lived, near N=1. From the chart, that point will be at an RCC about 168.8 (SD estimated at about 10%) or about 6424 (SD ~ 640) years before the average year of birth of the average testee (~ 1945 CE)

The difference in dates (6424 years vs. 5838 years) is within the errors of the data, but since it represents a systematic difference (10%) that will occur every time one dates a haplotype or SNP, the question needs to be resolved. In this particular case the difference in time divided by the number of elapsed generations indicates the approximate number of generations between the progenitor and the first split we observe. In this case it is about 590 years or about 20 generations between the progenitor and the first mutation we see.

In an attempt to settle the disagreement over the date of origin, I wrote three letters of inquiry, one to FTDNA, one to the J. Craig Venter Institute in La Jolla and one to a colleague at 23andme. The question posed and their answers are in an endnote.[5] Most of the issues they cited in their answers are already addressed elsewhere in this paper. My conclusion is that the origin of the SNP probably lies in or around the time derived from the value of RCC when the run of run of junction points in the figure is extrapolated to N=1. That date is probably about 6424 (SD ~640) years before the date of birth of the average testee, taken to be in 1945. Obviously, it could be any time between dates that correspond to the RCC values at N=2 and N=1.

It is not the intention of this paper to try to justify one dating method over the other because Klyosov and I use markedly different approaches. He derives a date of 4325 years before the present (SD ~ 8%), lower than my estimated date. The difference in dates is mainly caused by our different treatment of the outliers. If I average the RCC values in the RCC matrix, I derive a date that is closer to the one derived by Klyosov

which indicates that the outliers are indeed treated differently between the two dating methods. Those outliers need special attention.

What is the rationale for excluding or including the outliers in the sample? Why are they important, no matter which dating approach is used? We can summarize the situation as follows:

STATISTICAL REASONS FOR INCLUDING OR EXCLUDING THE OUTLIERS:

1. While outliers are more easily distinguished when you use the distribution on the tree together with the statistics of the numbers of junction points as we do in the RCC method, they may still be present in the data and must be addressed properly by analysts who use other dating methods. How they may be identified is beyond the scope of this paper, but they will probably exhibit themselves through statistics that involve genetic distances like we have done in Table 1, or Poisson statistics. Once identified as Type B outliers, they must be excluded.
2. There could be testing errors in a fraction of the sample that would appear as outliers. But the percentage of errors needed to produce the effects we see from the outliers would have to be over ~ 5% of the sample. I doubt that testing errors are that high. We found only one suspicious example of such an error (0.05% of the sample). Only if a testing error were responsible for an outlier high on the tree would this be a problem in dating.
3. There could be RecLOHs in the sample. This is the best reason to exclude a haplotype if its GD lies outside of the exponential fit to the rest of the sample.

NON-STATISTICAL REASONS FOR INCLUDING THE TYPE A OUTLIERS:

1. If the origin of a subclade takes place in the distant past, many of the lines of descent since that time will have died out, but a few of them may have survived as the Type A outliers. Their numbers will be considerably less than the number of testees who have more recent MRCAs. In the case of the L21 sample, we found three out of 1997 outliers that were excluded as Type B (0.15% of the sample). The rest may result from lines that have NOT died out.
2. If ancestors of a subclade have migrated from one region to another, and if the majority of males who have taken the test and who carry the subclade now reside overwhelmingly in those new areas of the world rather than where the subclade originated. The outliers result from the few testees whose haplotypes are concentrated at the original place of origin. This is a reasonable argument for inclusion of what may be an older set of haplotypes.
3. Of those testees who may still reside near the place of origin of the subclade, fewer have been tested either due to affordability considerations or to a lack of interest. This is a reasonable argument for inclusion.

The run of junction points in the tree is an exponential and the run of points up to the oldest junction point is smooth. Those observations provide a strong argument that random mutations are responsible for the general form of the tree. The earliest junction point is where one descendant line from the progenitor splits in two. The origin of the

sample is before the RCC that corresponds to the time when N=2, probably at or nearer the point where N=1.

The lessons learned from this exercise include the following:

- The inclusion or exclusion of outliers is critical to date determinations
  - Outliers need special attention in the dating process
  - They should not be included in processes where they are simply averaged
- Great emphasis should be given to the purity and eliminating biases in the sample
  - If dating a subclade is the goal, do not include older subclades.
  - Be sure that all haplotypes in the sample have been SNP-tested
  - If testees are chosen who are already known to be related, a bias results
    - Inclusion will drive the average age of the sample downward
    - It will lead to a more recent date for the sample
    - It will not affect the way the RCC correlation method dates the tree
    - But we must include outliers of Type A
      - Their inclusion will affect the sample date
    - This bias may adversely affect other methods that date a sample
  - If a family group is replaced with a single representative haplotype,
    - It will cause some insignificant restructuring on the tree
    - It will not affect the deep structure of the tree
    - It will lead to a significant age difference if other methods are used
- The RCC process of extrapolating junction points on the tree is unique
  - It results from the fact that the run of junction points is exponential
  - The date of origin can be estimated to be before the date of the oldest pair
  - The coefficient of determination (R^2) characterizes the goodness of fit
- The more the number of testees in the sample, the better the date estimate will be
  - But, bottlenecks, difficult to identify, may affect a date estimate due to:
    - A reduced variation in the gene pool
    - Smaller genetic diversity
    - Reduced robustness of the population
    - Lowered ability to survive radical changes in the environment
- Subhaplogroups may be easier to date on the tree than the SNP, itself
  - Their respective earliest junction points will indicate their age
  - Use them to determine an age order of subhaplogroups
- Some analysts elect to change the data prior to analysis. Proceed with caution.
  - Some substitute markers in place of blanks and zero marker values
  - Some eliminate fast mutating markers
  - Some change marker values within a RecLOH, and use the result
  - Some count a RecLOH as a GD of one.
  - Some eliminate the haplotype that includes a RecLOH.
  - This RCC dating method resists these modifications of data input
- Other types of mutations may be in the sample and must be considered:
  - Different marker lengths within the sample
  - The presence of exceptional marker changes within SNP-tested samples
    - On rare occasions they will occur within single valued markers
      - E.g., DYS 455 values may change from 11 to 8

        - In Hap I1a & J2b at different epochs
    - Some testing companies report fractional marker values
        - Other companies may not report a fractional value they find

UNCERTAINTIES IN DATING NEED FURTHER ATTENTION

- Various methods for dating yield significantly different results even when using the same sample input. It is not clear which method is best and more work is needed to explore those methods and to investigate the uncertainties in each one. Some analysts compute variances of each marker, combine the results for all markers and divide by the mutation rates. The RCC correlation approach uses the Mathematica application to optimize the inter-haplotype distances and plots the resulting testee locations on the tree. Then the RCC scale is converted to time using many pedigrees to determine the time scale. The average mutation rate of the haplotype string is responsible for determining the RCC time distance. So far, extensions of the RCC method from the genealogical times of interest into the genetic times of interest have yielded promising results, but more work is needed.
- There are different methods for identifying clusters or clades that are amenable to dating by these various methods. Some use utilities like the ones by McGee, Fluxus and Phylip software sets, including the Kitsch version. Ours uses Mathematica's hierarchical clustering routine. Some use different packages like NJPlot to display a tree. Some use polar displays like FigTree. Our RCC tree displays a time scale; few others do.
- Some analysts maintain that taking into account individual marker mutation rates rather than using an average across the entire haplotype set will make a difference in the date determination. No evidence has been seen, so this issue must be explored further.
- There is some disagreement whether the earliest branch point on the RCC-derived tree at N=2 is a genetic one or a point that results only from a statistical coalescence. We have maintained in this paper that it is a genetic one, caused by the first mutation in the sample and that an extrapolation of the exponential distribution of N as a function of time (RCC) will lead to a better estimate of the time of origin of the SNP. This must be explored further.
- It is evident from this paper that it will be a lot easier to put SNPs in date order than to assign a specific date to them. Samples can be chosen to include SNP-tested subclades and the relative position of their junction points on the tree can be used to order them by date.

Further work is needed to settle on the best dating procedure before we can fully trust the dates that are derived. While the RCC correlation method appears to be a good one, it might not be the best one.

ACKNOWLEDGEMENTS
I have had significant discussions with a number of colleagues about this issue of dating SNPs. I wish to thank J.J (Jim) Logan for his very thorough and extensive critiques of the methodology as well as the concept of coalescence and the biology of SNPs; Anatole

Klyosov for sending me the haplotypes we both analyzed and for the considerable discussions we have had about dating; Sidney Sachs for his deep insights into statistics, particularly the application of Poisson statistics to the sample used here; and Paul Burns for his insights in assessing comparisons of pedigree information and the position of testees on the tree whose pedigrees are known. I also appreciate the response of the staff at the J. Craig Venter Institute and 23andme.
Any errors in this paper or in the interpretation of results are mine alone.
==========

[1] The testee identifications are listed on the ordinate at the right of the tree. The RCC scale along the abscissa is a time scale where 10 RCC is approximately 380 years (SD~10%). The junction points on the tree indicate MRCA connections between and among the testees. Clusters whose TMRCA is within an RCC of 10 will all generally have an identifiable MRCA within that time span; an RCC of 20 is approximately the time at which surnames were adopted. You can see many junction points at RCC 20 or below where similar surnames are expected to cluster.
The trees in this paper are derived from haplotypes that have a length of 67 markers. The RCC-to-time conversion for a haplotype string of 37 markers has been derived from pedigrees. Using a large number of entries in the RCC matrix it is possible to determine the conversion for other marker lengths. The following table shows the result:

FTDNA Sequence of Markers	Years Per RCC (previous conversion)	Years Per RCC (current conversion)
25 Marker Length	30.2	28.49
37 Marker Length	43.3	40.85
67 Marker Length	40.3	38.05
111 Marker Length	36.7	34.65

In previous papers we have used 43.3 years as the conversion factor for 37 markers. Model results recently suggest that a factor of 40.8 is slightly better. These conversion factors cannot be interpolated because haplotypes with different marker lengths consist of different sets of markers. The RCC correlation analysis was introduced in the following paper: http://www.jogg.info/52/files/Howard1.pdf

[2] See http://www.isogg.org/wiki/RecLOH where the term RecLOH is defined by the International Society of Genetic Genealogy. It happens on markers with multiple alleles like DYS 413 and DYS 464.

[3] See the following references that are relevant to the analysis in this paper:
http://en.wikipedia.org/wiki/Coalescent_theory
http://en.wikipedia.org/wiki/Dendrogram
http://en.wikipedia.org/wiki/Hierarchical_clustering
http://en.wikipedia.org/wiki/Mathematica
http://dictionary.sensagent.com/RecLOH/en-en/

[4] The table below, kindly sent by J.J (Jim) Logan, gives a top-to-bottom list of L21 subclade designations as of the summer of 2014. We would expect that the subclades below L21 would be represented as tighter subclusters on our tree. If all haplotype

samples are SNP-tested at L21, then there should be no contamination of the derived date since older subclades are not included in the sample. We observe tighter subclusters on the tree that might correspond to whose in the figure under the L21 subclade. We expect that these subclusters would be represented as subclades somewhere on the tree and could be individually dated or at least placed in order of date. We note that our sample would be biased if it contained haplotypes that were in SNPs with designations above L21 on our tree. As we extend the SNP sequence to more recent dates, I believe that we will find that we will slowly converge on a DNA description of a testee as an individual, much as the description of loops and whorls in a fingerprint converge on the description of an individual.

FTDNA HG (SNP)	ISOGG Haplogroup		ISOGG Haplogroup	FTDNA HG (SNP)
R-CTS11722	R1b1a2a1a2c1k1		R1b1a2a1a2c	R-L21
R-CTS1202	R1b1a2a1a2c1f2c1		R1b1a2a1a2c1a1	R-DF23
R-CTS2501	R1b1a2a1a2c1i		R1b1a2a1a2c1a1a1	R-M222
R-CTS3087	R1b1a2a1a2c1b4		R1b1a2a1a2c1b	R-L513
R-CTS3655	R1b1a2a1a2c1g2a1b		R1b1a2a1a2c1b2	R-L193
R-CTS4466	R1b1a2a1a2c1l		R1b1a2a1a2c1b4	R-CTS3087
R-CTS6838	R1b1a2a1a2c1k1		R1b1a2a1a2c1c	R-L96
R-CTS6919	R1b1a2a1a2c2a		R1b1a2a1a2c1d	R-L144
R-DF21	R1b1a2a1a2c1g		R1b1a2a1a2c1d	R-L195
R-DF23	R1b1a2a1a2c1a1		R1b1a2a1a2c1e	R-Z255
R-DF25	R1b1a2a1a2c1g2a		R1b1a2a1a2c1e1	R-L159
R-DF5	R1b1a2a1a2c1g2a1		R1b1a2a1a2c1f	R-Z253
R-L130	R1b1a2a1a2c1g5		R1b1a2a1a2c1f1	R-L554
R-L144	R1b1a2a1a2c1d		R1b1a2a1a2c1f2a	R-L226
R-L159	R1b1a2a1a2c1e1		R1b1a2a1a2c1f2b	R-L643
R-L193	R1b1a2a1a2c1b2		R1b1a2a1a2c1f2c1	R-CTS1202
R-L195	R1b1a2a1a2c1d		R1b1a2a1a2c1g	R-DF21
R-L21	R1b1a2a1a2c		R1b1a2a1a2c1g2a	R-DF25
R-L226	R1b1a2a1a2c1f2a		R1b1a2a1a2c1g2a1	R-DF5
R-L270	R1b1a2a1a2c1l1		R1b1a2a1a2c1g2a1b	R-CTS3655
R-L371	R1b1a2a1a2c1h		R1b1a2a1a2c1g2a1b1	R-L627
R-L513	R1b1a2a1a2c1b		R1b1a2a1a2c1g5	R-L130
R-L554	R1b1a2a1a2c1f1		R1b1a2a1a2c1h	R-L371
R-L627	R1b1a2a1a2c1g2a1b1		R1b1a2a1a2c1i	R-CTS2501
R-L643	R1b1a2a1a2c1f2b		R1b1a2a1a2c1k1	R-CTS6838
R-L679	R1b1a2a1a2c1n		R1b1a2a1a2c1k1	R-CTS11722
R-L96	R1b1a2a1a2c1c		R1b1a2a1a2c1l	R-CTS4466
R-M222	R1b1a2a1a2c1a1a1		R1b1a2a1a2c1l1	R-L270
R-Z253	R1b1a2a1a2c1f		R1b1a2a1a2c1n	R-L679
R-Z255	R1b1a2a1a2c13		R1b1a2a1a2c2a	R-CTS6919

[5] The question posed was: "If you place a number of testees on a phylogenetic tree, all of whom carry a particular SNP, you find that the number of points where mutations have

taken place along the various lines of descent decreases as you go back in time. The earliest pair on the tree corresponds to the earliest mutation we can identify on the tree, but is that where the SNP originated, OR is it further back in time when an even earlier male ancestor was the progenitor who first carried the SNP? The answer to this question is very important to those of us who are engaged in dating SNPs."
The first answer: A professor of the J. Craig Venter Institute wrote: "'It depends...'One is not likely to know from a single sample if there were not other individuals in generations prior to the individual for which your tree converges that carried that SNP and ultimately transmitted that SNP to the individual for which your tree converges. Thus, there is sampling variation associated with the individuals you have identified that carried the SNP who you are using to recreate the history or short-term evolution of the SNP. Even if you had exhaustively identified ALL the carriers of the SNP in the current, contemporary population, you would not know with certainty how that SNP was ultimately transmitted to them from generation to generation (i.e., the SNP could have occurred multiple times via recurrent de novo mutation or there could have been homoplasy). Bruce Weir has commented on this type of phenomenon as have many of the people who have developed coalescent theory over the years, like Richard Hudson. There are other ways of dating DNA sequence variants that leverage linkage disequilibrium that might be of interest. However, all of the different methods provide an estimate of the date of origin of the variant +/- some error, where that error is associated with not only sampling variation but also the potential routes that could have led to the variant cropping up in the individuals you have sampled, genotyped and identified as the carriers of the SNP."
The second answer: 23andMe wrote: "I believe the origin of a SNP would generally be in an ancestor that predates everyone on the tree of testees."
The third answer: FTDNA's Chief Scientist wrote: "This is a good question that is somewhat complicated by the issue of sampling. In principle, typing STRs on the background of a single SNP should help to date when the SNP originated. However, this assumes good sampling of individuals with the SNP such that most of the diversity on that SNP background has been captured by the sample. For example, in the extreme case of a sample that only included members of one surname-- or a even paternally relative males within a single family—one can imagine that the age of the SNP will be underestimated. It would be wise to try to sample broadly (ethnically and/or geographically) to include diverse individuals with the same SNP.

NOTE ADDED:
Table 1 showed the absolute genetic distances of all 1993 pairs of L21 testees. The run of GDs for 56 of the 67 sites (84%) looked normal, running from higher numbers at GD 0 to lower numbers at larger values of GD. However, there were 11 sites that showed anomalies in that distribution and they are shown as shaded entries in the following table. The eleven DYS sites are listed across the top, and the distribution of GDs from 1 to 10 are shown in rows. As an example, DYS YCAIIb has 42 entries at GD 4. The haplotype number of each of those 42 entries is given in the column below that DYS site. Once the haplotype numbers were identified, they were then listed on the final tree. By inspection of the tree, most of these groups appear as distinct subclusters on that tree. In an earlier endnote, we speculated that tighter subclusters on the tree might correspond to L21

subclades, but we cannot test that association until the haplotype numbers on the tree have been identified with subclades.

GD	DYS389i	DYS464d	YCAIIa	YCAIIb	DYS456	DYS570	DYS413a	DYS413b	DYS490	DYS568	DYS565		SUMS
0	1632	1677	1937	1703	726	1288	1012	1845	1976	1941	1767		17504
1	358	300	33	238	1164	556	301	107	17	43	223		3340
2	6	17	8	13	100	123	644	40	3	3	3		960
3	0	0	2	1	3	23	13	2	0	10	4		58
4	1	3	17	42	4	5	3	0	1	0	0		76
5	0	0	0	0	0	1	0	0	0	0	0		1
6	0	0	0	0	0	0	0	1	0	0	0		1
7	0	0	0	0	0	0	23	1	0	0	0		24
8	0	0	0	0	0	1	0	0	0	0	0		1
9	0	0	0	0	0	0	1	1	0	0	0		2
10	0	0	0	0	0	0	0	0	0	0	0		0
Haplotype Number (see tree)												SUM:	21967
	552	1050	42	21	137	1467	94	421	982	248	459	Average:	1.06737379
		1051	255	27	562		145			312	477		
		1938	333	81	1159		164	1769		558	513		
			376	117	1306		195			715	1783		
			485	194			197	530		857			
			516	268			243			888			
			632	273			246			1068			
			654	294			287			1157			
			773	436			299			1167			
			890	521			377			1168			
			935	538			421						
			1080	543			707						
			1354	564			1079						
			1491	595			1111						
			1506	635			1265						
			1632	656			1443						
			1908	677			1470						
				678			1543						
				688			1644						
				693			1678						
				785			1765						
				833			1769						
				888			1788						
				944									
				1014			530						
				1135									
				1244									
				1357									
				1389									
				1404									
				1439									
				1477									
				1527									
				1564									
				1580									
				1617									
				1637									
				1659									
				1813									
				1824									
				1962									
				1983									