Generating a Dated STR-based Phylogenetic Tree From Y-DNA Haplotypes

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

ABSTRACT:

We describe a computer code, designed to be run as an application program within the Mathematica[1] programming environment, which can be used to derive a phylogenetic tree directly from marker groups of Y-DNA haplotypes. We illustrate its use by deriving a phylogenetic tree from a group of 18, 37-marker haplotypes. The tree illustrates graphically the evolutionary relationships of each haplotype to every other haplotype in the group. Each entry on the tree can be identified by FTDNA Kit number[2], and by other identifiers. The tree also contains a time scale, derived from pedigrees, which is built into the program code. Having such a tree provides surname project administrators and testees with an easily applied tool that produces a comprehensible graphic overview of all the genealogical and genetic relationships and their associated timescales among haplotypes.

INTRODUCTION:

A new approach to analyze Y-DNA haplotypes has been introduced and has been used to analyze surname relationships (Howard 2009a&b, Gordon and Howard 2011). An RCC time scale, calibrated with over 100 pedigrees[3], has been developed that can be applied to investigate the evolutionary relationships that tie genealogy and genetics together, using the same time scale, over tens of thousands of years by analyzing clusters of haplotypes. Examples of such investigations have been provided in Gordon and Howard (2011) and Howard and McLaughlin (2011). We then used the same RCC matrix, in conjunction with Mathematica, an application program, to derive an STR-based phylogenetic tree with its associated RCC time scale. This result confirmed and extended the time and evolutionary relationships among all Gordon Y-DNA testees (Howard and Schwab 2011).

The code accepts as input data an Excel file listing the marker sequences in tabular form, one row of data for each of *n* haplotype sequences, which are assumed to be of equal length (generally either 25, 37, or 67). The program proceeds by calling up three built-in functions from the "Hierarchical Clustering Package" (one of the standard packages within Mathematica). The first call, to the "DistanceMatrix" function, generates the RCC matrix. The second call, to "DirectAgglomerate", generates the hierarchical clustering. The final call, to "DendrogramPlot" generates the phylogenetic tree diagram, in the form of a dendrogram plot.

The Gordon-Howard paper uses the same data set to derive the phylogenetic tree directly from the Y-DNA marker sequence. The RCC time scale is built into the code.

1 RCC ~ 43.3 years for the time to the most recent common ancestor (TMRCA) for pairs of testees and for the time scale of intercluster TMRCAs;

1 RCC ~ 52.7 years for the TMRCA of a surname cluster; the multiplier is higher in order to account for incomplete cluster membership (Howard 2009a).

THE MATHEMATICA PROGRAM – PURPOSE AND DATA PREPARATION:

The success at analyzing the Gordon surname using this new approach indicated that Mathematica can be used to illustrate the genetic relationships and related timescales within any group of Y-haplotypes. Since Mathematica requires a depth of programming expertise that some investigators may not have, it is the purpose of this paper to present a simple code that other investigators can use or modify if they have access to Mathematica (see Appendix). We have found success at following the procedure below, although other applications can surely be used to show a similar result.

The process involves preparing a single spreadsheet (e.g., Excel) with the following columns in the case of a 37 marker data set (also see Appendix):

Column 1: A number that identifies the testee (e.g., Kit Number)
Column 2: A short secondary identifier such as a group, ancestor or other designation (These two parameters will be shown in along the vertical axis of the phylogenetic tree)
Columns 3-39: The marker numbers (values, not text) associated with each DYS location. (The code requires markers to be in separate columns)
Columns beyond 39: Any additional information pertaining to that haplotype

Each row in the spreadsheet will contain this information for a single testee. Thus the salient information actually used by the program constitutes an *n* by 39 input matrix. There should be no zero entries in any marker string and no entries should be made in rows below the last row of haplotypes.

Appendix A gives further details; (1) Figure A -- a shortened, but typical spreadsheet, (2) the code, (3) the resulting phylogenetic tree, and (4) a brief note about data entry.

DISCUSSION OF RESULTS:

When Mathematica is used to derive a dated STR phylogenetic tree directly from marker strings, it forms family groups, or clusters, unambiguously, quickly and automatically, without using an RCC matrix as an intermediate step. Entering the haplotypes in a different order does not noticeably change the appearance of the tree, although entering identical haplotypes in a different order may change the order of presentation on the tree.[4]

It is a distinct advantage for a testee to be able to see where his Y-DNA test result is located on a phylogenetic tree. Showing the approximate time when he shares a most recent common ancestor (MRCA) with every other testee is a distinct advantage of this approach. Arguably, this is the first time that a uniform time scale can be applied to this type of a phylogenetic tree. From the tree he can see his time and evolutionary relationships back several tens of thousands of years when he compares his result with other testees whose haplotypes differ significantly from his. Care must be exercised to prevent a null marker from inclusion in the haplotypes because Mathematica will interpret it as a zero entry and its position on the tree will significantly bias its relationships with other entries.

Errors of Position of Testees on the Dated STR Phylogenetic Tree

The presence of mutations causes large percentage uncertainties in time that average an estimated 43 percent (Standard Deviation (SD) ~ 4%) between two haplotypes over spans of time of interest to Y-DNA geneticists. Within a genealogically interesting time scale mutation models have shown that uncertainties expressed as 1 SD are of the order of 2.8-3.5 in RCC, or about 130-160 years, equivalent to about 4-6 generations. Times from the present out to about 400 years appear to have lower-than-average values of SD, because mutations have hardly begun to take place from the starting haplotype. The SD rapidly climbs to the region near 40 percent where it remains, fluctuating around that value for many millennia (Howard and Schwab, in preparation).

Although Mathematica optimizes the groups of haplotypes prior to placing testees on the tree, the position assigned to the tree could be in error by an RCC of about 3 (1 SD), but ten percent of the time, the location error could amount to an RCC of about 6. This situation may be sufficient to move a testee into a surname cluster where he is more distantly related than the others in the cluster or, a testee might show up outside a cluster in which he should otherwise belong. We conclude that if you believe that your pedigree is good, put more value in the pedigree relationship than in the position on the tree. Pedigrees are not subject to mutations. However, if you are uncertain of your pedigree, there is a high probability that you will share a pedigree with other testees in your shared surname cluster, particularly when the members of the tree share a common ancestor at RCC less than about 20.

FTDNA's TIP process reports on a testee's result by giving the probability that he has a MRCA within a certain number of generations with another testee over a time period that covers genealogical pedigrees. Some testees with TIP results will have no genealogically significant matches. In contrast, all testees on our tree have results that will show them their connection with everyone else on the tree, even back to his deep ancestry – a decided advantage for testees who otherwise would have no results. Although a testee may have no pedigree connections at first, he will immediately see his tie to the genetic evolution of his surname and he will gain insight into other testees with whom he shares a MRCA, and approximately how far back in time his pedigree must be pursued to match his MRCA with another testee.

Occasionally a testee may find that his Y-DNA result matches best with a surname that is not his. For instance, there are a group of at least 17 testees who carry the surname Robertson who, in actuality, match one group of 140 Hamiltons called Hamilton A at the FTDNA web site of Hamiltons. A dated phylogenetic tree for these testees indicates that the TMRCA for the Robertson-Hamilton progenitor event occurred at about RCC 15, or 650-800 years ago (1150-1300 CE). This is near the edge of where pedigrees are valid and when surnames were chosen, but the relatively large number of testees among the

Robertson and Hamilton A groups may yield a more definitive identification and date of the TMRCA. In this case, Robertsons are Robertsons and Hamiltons are Hamiltons, but they share a common Y-DNA line back in time.

CONCLUSION:

We have presented a methodology and a code that will produce a dated STR phylogenetic tree and its associated evolutionary time scale from Y-DNA haplotype test results. At the same time, we have established a hierarchical clustering process. The Mathematica program can now be used with minimal training and experience to produce results that could be routinely made available to those tested.

POSSIBLE FUTURE APPLICATIONS:

Two initiatives can be explored. The first would substantially broaden what project administrators and testees can learn about their Y-DNA results; the second would add to our knowledge of how specific mutations are correlated with the marker differences that, like fingerprints, define specific family clusters:

1. Since surname administrators have different levels of expertise at programming and analysis, we suggest that a web site might be developed that would contain the analysis program. A haplotype would be submitted by the testing agency or by a surname administrator. The program would process the data and return an output that would show where that haplotype is located on the phylogenetic tree derived from an appropriate set of haplotypes. This process would provide ease-of-access to a remote computing site on the internet and would not require end-user knowledge of the program or the application that delivers the service. Although we believe the technical problems associated with this concept can be easily solved, we also recognize that a number of administrative and financial issues would have to be addressed.

2. Now that we can place surname clusters in a dated time sequence that is highly correlated with, and calibrated by pedigrees, a more detailed study of individual marker changes should give us additional insight into marker differences that are found to be associated with differences in cluster membership and their evolution in time. This fresh approach offers the possibility of exploring in much finer detail the relationships between haplotypes, haplogroups, and their associated SNPs and subclades back through tens of thousands of years.

As we stressed in previous papers in this series, the methodology we have developed is intended to complement rather than to replace existing tools for analyzing Y-DNA haplotype test results.

ACKNOWLEDGEMENTS: We would like to thank the following people for discussions that have contributed significantly to this work. They are Tei A. Gordon, David E. Hogg, James M. Irvine, Linda MacLachlan, and Sidney Sachs. We particularly thank Eileen Krause Murphy at FTDNA for clarifying why STR-based trees are not always in the same sequence as on a SNP-based tree.

APPENDIX:

In this section we present a sample code that will take a STR haplotype string 37 markers long and produce a dated phylogenetic tree. Mathematica optimizes the entire set of values when it correlates pairs of markers using the input haplotype matrix. In this example, we produce a relatively simple tree, but variations on the code will allow the program to produce trees with 500-600 different haplotypes. Identifiers appear in the first two columns of input and they reappear positioned on the tree through Mathematica's optimization process.

To install and run the code, follow these steps:

(0) Place a copy of the code into the directory (or "folder") where your input files are stored, e.g., "/Users/myusername/what/have/you/", under the name "treecode.m".

(1) Start up Mathematica and create a new notebook.

(2) Type

```
SetDirectory["/Users/myusername/what/have/you/"]
```

followed by <Shift+Enter> . (If you are a Unix user and have started Mathematica from a command-line window in the desired working directory, you may omit this step).

(3) Type either

```
<<treecode.m
```

or

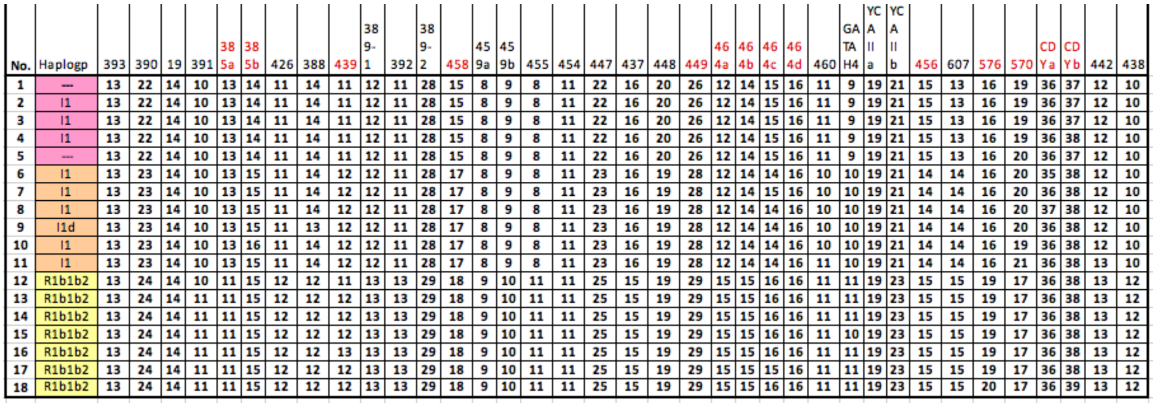
```
Get[treecode.m]
```

followed by <Shift+Enter> . You will then be prompted to enter the name of the desired input file (e.g., "myfile.xls").

Omit Step 0 on subsequent runs.

Figure A1 is an example of an 18 by 39 matrix of haplotypes representative of n = 18 testees. The 18 haplotypes chosen are a representative group of members of three Gordon clusters, two of which are in the same haplogroup. They are subcluster Ae, and Clusters E and Q of Gordon and Howard (2011).

Figure A1

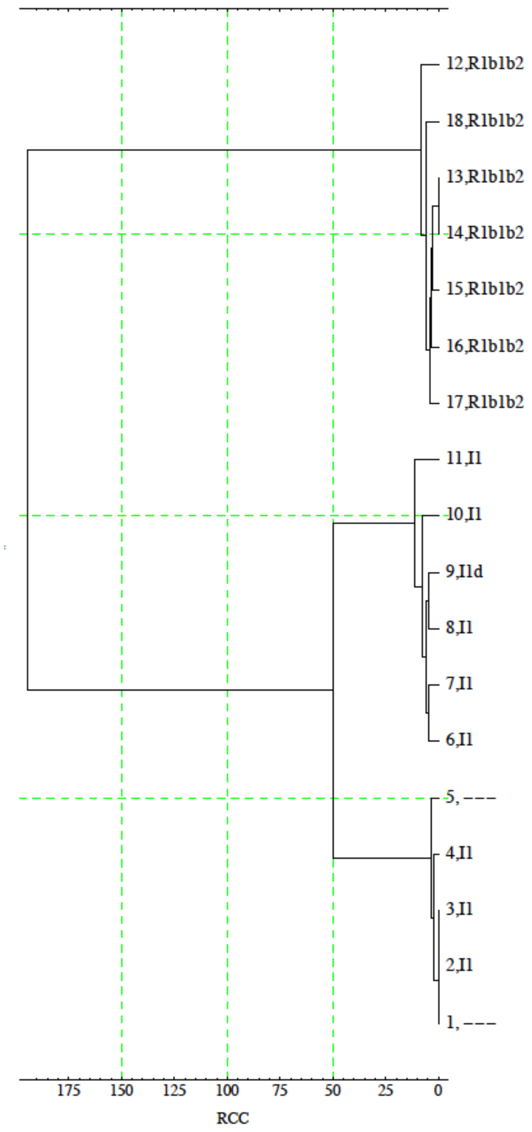


No.	Haplogp	393	390	19	391	385a	385b	426	388	439	389-1	392	389-2	458	459a	459b	455	454	447	437	448	449	464a	464b	464c	464d	460	GATAH4	YCAIIa	YCAIIb	456	607	576	570	CDYa	CDYb	442	438
1	---	13	22	14	10	13	14	11	14	11	12	11	28	15	8	9	8	11	22	16	20	26	12	14	15	16	11	9	19	21	15	13	16	19	36	37	12	10
2	I1	13	22	14	10	13	14	11	14	11	12	11	28	15	8	9	8	11	22	16	20	26	12	14	15	16	11	9	19	21	15	13	16	19	36	37	12	10
3	I1	13	22	14	10	13	14	11	14	11	12	11	28	15	8	9	8	11	22	16	20	26	12	14	15	16	11	9	19	21	15	13	16	19	36	37	12	10
4	I1	13	22	14	10	13	14	11	14	11	12	11	28	15	8	9	8	11	22	16	20	26	12	14	15	16	11	9	19	21	15	13	16	19	36	38	12	10
5	---	13	22	14	10	13	14	11	14	11	12	11	28	15	8	9	8	11	22	16	20	26	12	14	15	16	11	9	19	21	15	13	16	20	36	37	12	10
6	I1	13	23	14	10	13	15	11	14	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	14	16	10	10	19	21	14	14	16	20	35	38	12	10
7	I1	13	23	14	10	13	15	11	14	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	15	16	10	10	19	21	14	14	16	20	36	38	12	10
8	I1	13	23	14	10	13	15	11	14	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	14	16	10	10	19	21	14	14	16	20	37	38	12	10
9	I1d	13	23	14	10	13	15	11	13	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	14	16	10	10	19	21	14	14	16	20	36	38	12	10
10	I1	13	23	14	10	13	16	11	14	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	14	16	10	10	19	21	14	14	16	19	36	38	12	10
11	I1	13	23	14	10	13	15	11	14	12	12	11	28	17	8	9	8	11	23	16	19	28	12	14	14	16	11	10	19	21	14	14	16	21	36	38	13	10
12	R1b1b2	13	24	14	10	11	15	12	12	11	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	11	19	23	15	15	19	17	36	38	13	12
13	R1b1b2	13	24	14	11	11	15	12	12	12	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	11	19	23	15	15	19	17	36	38	13	12
14	R1b1b2	13	24	14	11	11	15	12	12	12	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	11	19	23	15	15	19	17	36	38	13	12
15	R1b1b2	13	24	14	11	11	15	12	12	12	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	10	19	23	15	15	19	17	36	38	13	12
16	R1b1b2	13	24	14	11	11	15	12	12	13	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	11	19	23	15	15	19	17	36	38	13	12
17	R1b1b2	13	24	14	11	11	15	12	12	12	13	13	29	18	9	10	11	11	25	15	19	29	15	15	15	16	11	11	19	23	15	15	19	17	36	38	13	12
18	R1b1b2	13	24	14	11	11	15	12	12	12	13	13	29	18	9	10	11	11	25	15	19	29	15	15	16	16	11	11	19	23	15	15	20	17	36	39	13	12

The Mathematica code ("treecode.m") applied to this matrix follows:

```
Needs["HierarchicalClustering‘"]


(*  This defines a function, "analyze", which - given a file name, desired
    number of markers, and plot label - will perform the analysis: *)
analyze[infile_,nm_,plotlabel_:""]:=Module[{n,spreadsheetdata,fontsize,
      kitnos,col2info,md,RCCMatrix,da,plt,ft,f},

   spreadsheetdata=Import[infile][[1]];
   (* Select from the spreadsheet all the rows whose length is at least nm+2
      and all of whose elements in columns 3 through nm+2 are numeric values;
      Rows not meeting these criteria are assumed to be labeling information
      or incomplete marker data: *)
   data=Select[spreadsheetdata,
      (Length[#]>=nm+2&&Union[Map[NumberQ,Take[#,{3,nm+2}]]]=={True})&];

   (* n is the number of data points: *)
   n=Length[data]; Print["Found marker data for ",n," testees in file: ",infile];

   fontsize=10; (* If printing, the following choice may be better: *)
   (* fontsize=Which[n>500,4,n>180,6,n>90,7,n>45,8,True,10];*)

   (* The kit numbers are in column 1. *)
   kitnos=data[[All,1]];
   (* This command converts all the numeric kit numbers from character
      strings to integers (it just gets rid of the decimal points that
      would appear in the labels otherwise): *)
   kitnos=Map[If[NumberQ[#],IntegerPart[#],#]&,kitnos];

   (* The secondary IDs or other info are in column 2; and the marker data
      are in columns 3 through nm+2, where nm is the number of markers: *)
   col2info=data[[All,2]];
   md=data[[All,3;;nm+2]];

   (* Define the RCC distance function: *)
   RCCDistance[x_,y_]:=10^4 (1/Correlation[x,y]-1)//Chop;

   (* These three commands generate the RCC matrix, perform the clustering
      analysis, and generate the dendrogram plot: *)
   RCCMatrix=DistanceMatrix[md,DistanceFunction->RCCDistance];
   da=DirectAgglomerate[RCCMatrix,Linkage->"Average"];
   plt=DendrogramPlot[da,Orientation->Left,(*AspectRatio->Automatic,*)
      PlotRange->{All,{0,n+1}},Frame->{True,False},FrameLabel->{"RCC",""},
      GridLines->Automatic,GridLinesStyle->Directive[Dashed,Green],
      LeafLabels->(Rotate[ToString[kitnos[[#]]]<>","<>col2info[[#]],0,
         BaseStyle->Directive[FontSize->fontsize]]&)];

   (* We would be finished now, except that with "Left" orientation the RCC
      axis isn’t labeled properly. The tricky business below will fix it: *)
   ft=FrameTicks/.FullOptions[plt];
   f=If[#1=!="",Rationalize[Abs[#1]],#1]&; ft[[1,All,2]]=Map[f,ft[[1,All,2]]];
   Show[plt,FrameTicks->ft,PlotLabel->plotlabel,
      AspectRatio->(1.6*n*fontsize)/864,ImageSize->{864,Automatic}]]


infile=InputString[ "Please type the name of the input file (e.g.,
   mydata.xls): "];
myplot=analyze[infile,37,infile<>"  Data Set"]; Print[myplot]
(* One can write an output file (pdf, gif, html, etc.) via, e.g.: *)
Export["myplot.pdf",myplot]
```

The resulting dated Y-DNA phylogenetic tree is given in Figure A2, below.



< == Gordon Cluster Q
(Haplogroup R1b1b2)

< == Gordon Cluster E
(Haplogroup I1)

< == Gordon Subcluster Ae
(Haplogroup I1)

Comments about Figure A2:

Members of Gordon Subcluster Ae (Nos. 1-5), Cluster E (Nos. 6-11), and Cluster Q (Nos. 12-18) each have MRCAs within epochs of genealogical interest (RCC <~ 17, or 700-900 years ago, with SD~ 30%). The MRCA of Subcluster Ae and Cluster E lived about RCC ~ 50, or ~ 2200 years ago (~220 BCE). The MRCA of Clusters A, E and Cluster Q, which are in different haplogroups, lived about RCC ~ 200 (about 8700 years ago). Differences between the time estimates of Figure A2 and Table 1 of Gordon and Howard (2011) are due to small differences among selected testees.

REFERENCES:

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 2009a, p. 256.

Howard, William E. III and Schwab, Frederic R., *Dating Y-DNA Haplotypes on a Phylogenetic Tree: Tying the Genealogy of Pedigrees and Surname Clusters into Genetic Time Scales*, Journal of Genetic Genealogy, 7, Fall 2011.

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 2009b, p. 271.

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, 7, Fall 2011.

Howard, William E. III and McLaughlin, John D., *A Dated Phylogenetic Tree of M222 SNP Haplotypes: Exploring the DNA of Irish and Scottish Surnames and Possible Ties to Niall and the Uí Néill Kindred,* Familia, Ulster Genealogical Review No. 27, pp. 14-50, 2011. Ulster Genealogical & Historical Guild.

AUTHORS AND POINTS OF CONTACT:

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

Communications regarding this paper should be addressed to William E. Howard III

END NOTES

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

[2] Family Tree DNA, Genealogy by Genetics, Ltd., 1445 North Loop West, Suite 820, Houston, TX 77008, USA. See also < https://www.familytreedna.com>

[3] The RCC time scale was calibrated using the 37 FTDNA markers. Trees can also be produced with any number of markers. No differences in time scale have been noted if markers other than those 37 are used.

[4] The tree we produce is an STR tree because the program optimizes the string of STR markers using only the haplotypes in the sample. If haplogroup designations are shown in an STR tree, the evolutionary sequence may not be exactly in the same time order as in the Y-DNA haplogroup tree shown on the ISOGG (International Society of Genetic

Genealogy) website at http://www.isogg.org/tree/. Those ISOGG designations are derived from SNPs (Single Nucleotide Polymorphisms). There is a high correlation between the evolutionary sequence of STRs and SNPs, however. If an early ancestor in an STR line of descent has more sons, there will be more opportunity for mutations to take place. Some of those more frequent mutations, if they occur in earlier parts of the evolution, will lead to what appears to be a mismatch on an STR-derived tree relative to the ISOGG sequence. Although a SNP sequence may be correct, the boundaries of what defines the STRs of a sample of haplogroup subclades like Q1a3a are probably quite broad. Their distribution on an STR-derived tree may sometimes impinge on the tree boundaries of an adjacent, earlier subclade, making a testee of Q1a3 appear younger than an adjacent testee who is in subclade Q1a3a. Research devoted to determining the ages of SNP sequences is still in considerable flux, with additional subclade symbols often being added or revised yearly.

=====================

Submitted to the Journal of Genetic Genealogy in July 2012