In order to use yDNA results to do more than sort male haplotypes into patrilineages, it is necessary to analyze the results in detail to try to reconstruct the mutational history of the patrilineage and correlate this with the genealogical history as worked out by research. Many have proposed relying on Fluxus or other cladogram software to do this, but I argue here that cladistic analysis is way too simplistic an approach to bear useful and reliable fruit, leaving out as it does a significant part of the evidence: what we know of the genealogy. The principles of this sort of analysis are useful so far as they go, but given the metastable nature of the hypothetical results, I find it far better that we work out even the basic architecture of the tree by hand, so as to remain alert to the alternative possibilities that will suggest themselves in light of our family historical knowledge. However, in order to be alert to the pitfalls of any sort of statistics-based approach, and more specifically, in order to manage the reconstruction ourselves, it is necessary that we be explicit about the variables involved—the mutation rates, the depths of the patrilineage, and the number of haplotypes. While Fluxus, and perhaps other cladogram software seems to have some ability to manage these other kinds of variables, I for one, have found that program to be hopelessly cumbersome and confusing to use in anything other than the default mode, and in the absence of an adequate explication of its assumptions and algorithms, I do not trust it to do the right things.

At any rate, based on one extended example that I discuss more fully below, I do not find its results to be at all satisfactory, nor its diagrams either very useful or meaningful.

I propose instead, a more subjectively guided procedure aimed at constructing what I call a "mutation history tree", which is capable of incorporating both the DNA evidence and the genealogical evidence. I will discuss my methodology for constructing these trees below, but first I would like to try to put Fluxus through it's paces on the following haplotypes of a particular patrlineage: ROBB Patrilineage 2 :

Test Subject Information			FTDNA 37-Marker Panel																																				
Proj #	Principal Researcher	Earliest Known Ancestor	393	390	19/394	391	385a	385b	426	388	439	389I	392	389II	458	459a	459b	455	454	447	437	448	449	464a	464b	464c	464d	460	YGATAH4	YCAIIa	YCAIIb	456	607	576	570	CDYa	CDYb	442	438
R-18	John B. Robb	Joseph Robb b.s1735,Sco	15	23	15	10	15	17	11	13	12	13	12	31	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	37	37	11	10
R-02	-----	James Robb d.1814,IN	16	23	15	10	15	17	11	13	12	14	12	32	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	37	37	11	10
	MRCAncestral	RPH	16	23	15	10	15	16	11	13	12	14	12	32	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	37	37	11	10
R-05	-----	Wm Robe b.s1690,Sco?	16	23	15	10	15	16	11	13	12	14	12	32	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	36	37	11	10
R-24	-----	John Robb b.s1695,Sco?	16	23	15	10	15	16	11	13	12	14	12	32	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	36	37	11	10
R-14	-----	Wm Robb b.1750/1,Ire	16	23	15	10	15	16	11	13	13	14	12	33	15	8	10	11	11	25	14	20	27	11	14	15	16	11	11	19	21	15	14	16	19	37	37	11	10
R-11	-----	Wm Rab b.s1700,Ire/Sco	16	23	15	10	15	16	11	13	11	14	12	32	15	8	10	11	11	25	14	21	28	11	14	15	15	11	11	19	21	15	14	16	19	36	37	11	10
R-12	-----	Johnston Rabb b.1790,PA	16	23	15	10	16	16	11	13	11	14	12	32	15	8	10	11	11	25	14	20	28	11	14	15	15	11	11	19	21	15	14	16	19	37	37	11	10

The test subject whose "Earliest Known Ancestor" is colored red is the one who is most closely related to all the others, and is therefore the one whose haplotype is most likely to approximate that of the common ancestor. The haplotype of this person is therefore adopted as the Root Prototype Haplotype (RPH) of the patrilineage—the reference haplotype to which all the others are compared. Marker results which deviate from the RPH are highlighted in lime green, and are presumed to be mutations.

With fewer than 4 test subjects, there's insufficient data to guess at the RPH, yet it's still desirable to indicate marker values which vary from one haplotype to another; for this purpose, the first listed haplotype will be treated as the reference haplotype, and markers which deviate from it will be shown in yellow. Variations in haplotypes with markers beyond the 37th will also be treated this way, unless they include the RPH. The RPH can be expected to change as more tested members are added to the patrilineage. Occasionally. as here, it may seem desirable to postulate an RPH slightly different from, and prior to that of the member with the least genetic distance from all the others, in this case is R-05 (& R-24).

The Data

In my version of the ROBB Patrilineage 2 haplotype set, which consists of 9 members, I have pruned clusters of descendants related no more distantly than second cousins to a single representative haplotype, thereby reducing the set to 7, to minimize self-selection bias in the sample. I have also excluded any haplotypes that do not include all the markers found in the FTDNA 37-marker set.

The RPH (Root Prototype Haplotype) in my haplotype charts is normally the member of the set with the least sum of the genetic distances between each member and every other. I then mark deviations from the RPH haplotype as mutations, by color coding them lime green. The RPH calculated by this method[1] is that of Robb project member R-05, but I have instead posited a slightly different haplotype as the RPH for reasons that will be more fully explained below in the second part of this paper, Constructing a Mutation History Tree the Old-fashioned Way.

Accounting for the Mutation History with Fluxus Diagrams

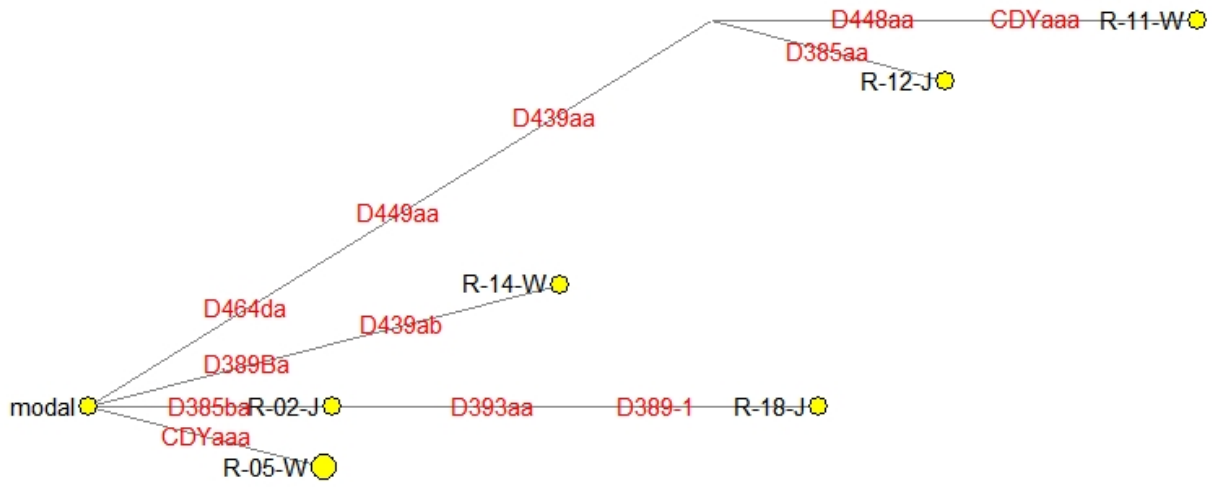
I ran this haplotype data set through Fluxus to produce "median joining" network or tree diagrams, with the "MP calculation" postprocessing option exercised. This quoted gibberish means nothing to me (and I can find no documentation on it either in the program help or in the Fluxus user guide), and whether the diagrams it produces are network diagrams or the descendancy tree charts that would be meaningful to genetic genealogists is also left undefined (the program has a radio button option to produce either a "Tree" or "Network" diagram, but the as often as not the two come out the same), and in no instance have I been able to get it to show all the nodes of a patrilineage (the tested descendant haplotypes) as end points of a descendancy tree. Presumably "tree" is meant in some other sense here known only to the programmers.

Provision is also made in Fluxus for optional weighting of the factors that underlie the chart—the mutation rates for the individual markers that vary across the haplotype set. At first this parameter specification facility appears to by dysfunctional. When I enter 353 as a weighting parameter for a CDYa mutation (since the mutation rate for CDYa is estimated to be about .0353) it doesn't appear to "take". However, it turns out that these weighting parameters are for some reason restricted to a range of 0-99, and *the user is not informed that he has entered an out-of-range parameter*—really quite unbelievable for a presumably professional program.

After a couple of days of playing around with this program (which is extremely awkward to use, as well as very poorly documented), I've finally gotten it to produce a diagram that comes close to resembling a true descendancy tree for my sample set:

Fluxus Tree Diagram for Robb Patrilineage 2

with inversely weighted mutation rates



[1] See my paper on RPH at http://www.johnbrobb.com/Content/TheRPH.pdf

It will be observed that this is still not quite a descendancy tree, even though I checked the "Tree" option in the program, since tested project member R-02 is shown as an ancestor of R-18, when in fact both are living contemporaries. Another problem is that this diagram fails to show that there were two mutations (or a single two-step mutation) at DYS439 for R-14—unless this is part of the secret coding of the letters "ab" appended to "D439". Also, we are not told whether the order in which multiple mutations are listed between nodes has any significance. Although there is clearly insufficient data to determine such an order, it may be that these markers are being listed in order of their mutation probabilities.

Furthermore, the top of the tree is labeled the "modal", even though if it were appropriate to care about the modal value of a haplotype set, the modal for this one could only be the common value of the R-05 and the R-24 haplotypes (the other 5 haplotypes are all different, and different from each other). Even more oddly, if this is to be considered a "Tree" chart is that R-24 was lost in the shuffle, at least by name—though the existence of two identical haplotypes does seem to have doubled the size of the R-05 node. Thus, both "modal" and "tree" are used here in senses that render the meaning of this chart obscure, and therefore inadequate as a representation of what I, in my own tree diagram, below, would call the hypothetical mutation history of this patrilineage.

In my attempts to get Fluxus to produce something useful, I ran this haplotype set through the program three times: first without changing the equal default weights for the variant markers, then again by weighting them proportional to the variant marker mutation rates (chopped down to fit the program's crude two-digit range; i.e. I multiplied by 1000 and rounded them to the nearest integer). Finally, I ran the program a third time after inverting the scaled mutation rate factors by dividing them by the fastest scaled mutation rate factor—35 for the CDYs.

Although the subordinate components of the diagram were identical for all three versions, they were attached in different ways. The unweighted version made R-18 subordinate to R-02, and both R-11 and R-12 subordinate to R-05, while the normally weighted version did the same except that it subordinated R-12 directly to the "modal" haplotype (the putative haplotype of the MRCA). In the third run, using the inverted weightings, the only subordination was of R-18 to R-02. This third chart (the one shown above) most closely resembles the one I have constructed, below.

In the end, the only components of the Fluxus diagram that make complete sense are those that associate sets of successive mutations with particular sublineages, but these associations can be readily inferred by a few moments inspection of a comparative haplotype chart like the one above. Since the Fluxus rationales for reconstructing the overall architecture of the tree remain undefined as well as indeterminate, I do not think that it's a particularly useful or reliable tool for genetic genealogical purposes.

Next, I will take up my proposed alternative to the cladogram: the mutation history tree.

Constructing a Mutation History Tree the Old-Fashioned Way—by Hand

All the mutation history trees I have constructed so far have been based on the FTDNA 37-marker panel, because none of the tests from other companies measure up in mutational sensitivity; however, there is no reason why the same procedure couldn't be used for the FTDNA 67-marker test.

The basic principle for constructing the tree is simple and straightforward: ***haplotypes with unique mutations or combinations of mutations constitute separate branches of the patrilineage.***

But mutations are nominally independent and the same mutation can occur independently in two different branches: how do we know when haplotypes with identical mutations occurred independently in two family branches, and when they represent a single mutation inherited from a common ancestor? And another, overlapping question is: how do we join up the various branches of the overall patrilineage? A general answer to the latter question is that in joining up sub-lineages we should be guided by our genealogical knowledge (where this is well-founded), but since we are using DNA partly as a check on that genealogy, let us first work out the principles of reconstruction as far as we can using the DNA results alone.

In deciding whether two overlapping mutations were probably independent, or whether they were inherited from a common ancestor, we must be guided by the mutation probabilities of the individual markers and combinations of markers, and these vary with the number of generations back to the Most Recent Common Ancestor (MRCA) of the patrilineage, or further, to the surname founder of the line. APPENDIX A, below, consists of a table showing, for each of the markers in the FTDNA 37-marker panel, its mutation probability, followed by the percentage chances of it mutating at least once over 20, 14, and 8 generations. These generation numbers were chosen because they represent, respectively:

20 generations back - the time when surname first began to come into general use in England
14 generations back - the time by which most Englishmen had surnames
8 generations back - the time by which most Scotsmen, Irishmen, and Welsh had surnames

I have worked out the mutation percentages for different numbers of generations back to provide guidance for patrilineages whose MRCA can be estimated from the degree of GD (Genetic Distance) in the haplotype set, or otherwise, to fall short of the earliest period of surname adoption. Values for estimates that fall in between these generational periods can be interpolated into the table, but when in doubt, a 20 generation estimate should be used.

Presumptions About ySTR Marker Mutations

The scientists know next to nothing about the causes of mutations to these markers, and little, even about the patterns of mutation. To simplify the analysis, and the probability calculations, I have adopted the following default hypotheses: that ySTR markers mutate independently of each other, according to individual mutation probability rates, that they mutate only one step at a time (the "stepwise mutation model"), and the mutation has an equal probability of gaining or losing a "repeat" (i.e. of begin "up" or "down"). The mutation rates I use are the same I've outlined in my paper ySTR Marker Panels Compared.

Problematic Markers (the CDYs and DYS464) and "Silent Mutations"

It will be seen from the table in APPENDIX A below that the CDY markers mutate at such a rapid rate (.03531 per generation) that it is extremely likely that they will appear independently in multiple family branches, and that some of these mutations may be quite recent, occurring just over

the last few generations. This lability renders them unsuitable for working out the architecture of the mutation history tree—unless, and for as long as, they are unique across all the haplotypes, and even then they must be treated with caution. Indeed, the CDYs are so likely to mutate than there is a reasonable chance that over 20 generations, one might mutate in a certain direction, and then mutate back in the opposite direction leaving not a trace, and thus disguising two mutations that we will never know about. These may be called "silent mutations".[2] Because there are so volatile, ***the CDYs are best ignored when working out the architecture of the mutation tree whenever the same CDY mutation appears singly in more than one haplotype***.

The other problematic marker (or set of markers) is DYS464a-d.[3] This is a four-component "multicopy" marker, so called because four different copies of the same marker are found at different points on the yChromosome. The problem is that all these copies are extracted when testing for 464, and jumbled together for evaluation, so that there is no way to link each measured values to its original location. By convention, the results of the four alleles are listed left to right in order of size, but with a DYS464 value of 14-14-14-15, the 15 allele value might occur at any of the four locations. Thus, we are faced with an anomaly: on the one hand, each component of 464 can mutate independently; nevertheless, we can only measure them as a set, which gives rise to frequent ambiguities: for a simple example, 13-14-14-15 might have originally been 14-14-14-14, with any one of the 14s gaining a repeat, and any one of the others losing one. Given the relatively high mutation rate for each individual component (.00566) it's quite likely that a 14-14-14-14 starting value might be measured as unchanged, even though any one of its components may have mutated up to 15, and then back down again to 14, creating a pair of silent mutations.

Clearly, DYS464 is a very troublesome marker, yet we can't afford to ignore it because it makes up 4 of the 37 markers and because it has a high mutation rate. It is usual in GD and TMRCA calculations to treat DYS464a-d as a single marker, with a quadrupled mutation rate (.02264), and to collapse multiple mutations into a single mutation (applying the so-called "infinite alleles mutation model"). Because over genealogical time 464 is unlikely to mutate more than once or twice, I hd previously elected to try to determine by comparative analysis, just how many times the composite marker had mutated, but experience has convinced me to just treat the whole marker as a binary value: either it has mutated from RPH value, or it hasn't. For purposes of constructing the mutation history tree, multiple patterns of deviation of 464 from the norm, are of course treated as so many separate, independent mutations.

DYS576, the fastest of the remaining markers (mutation rate .01022), has about an 18.6% chance of mutating over 20 generations, but only about a .5% chance of mutating back to its original state and thus disguising two mutations, so we may reasonably ignore the possibility of silent mutations for all markers except the CDYs.

Identifying a Family Branch Signature: Shared Single Mutations

Any mutation that is occurs more than once in a set of haplotypes can be considered a signature marker for a particular family branch. The problem is that, as the number of haplotypes increases, and additional instances of these singleton mutations appear, there is no foolproof way to tell, by considering all of the instances by themselves, which represent shared (inherited) mutations, and

---

[2] I calculate the probability of this to be 3.9% for each CDY. Although this probability is low for any given haplotype, if there are more than 15 in the set, one can expect at least one of them to have experienced one of these silent mutations.

[3] For simplicity, and to permit "apples-to-apples" comparison, I ignore here the possibility of additional copies, 464e, 464f, etc., which occur only rarely.

which may have occurred independently in more than one line. The best we can do towards resolving this dilemma is to consider: (1) the alternate probabilities; and (2) the underlying genealogies. Because we are trying at this point to construct a mutation history tree as far as possible independent of the genealogy, let's take a look first at the relative probabilities that the same mutation in two haplotypes either did, or didn't result from two independent mutations..

From APPENDIX A it can be seen that there is an exceptionally wide range of mutation probabilities across the 37 markers we are using here to analyze patrilineages. Indeed, the chances that the slower mutating markers would mutate even once, over a large number of haplotypes and generations, are quite minuscule. Thus, ***if we are lucky enough to pick up one of these slow markers (from DYS389I on down in the table in APPENDIX A below) in one of the project haplotypes, that by itself can be considered a reliable Family Branch Signature, because if we encounter it more than once, it's almost certain to represent a shared, because inherited, marker***.

On the other hand, the fastest-mutating markers can be expected to spring up like weeds across a large haplotype set. For example, the next fastest mutating marker after the CDYs, DYS576, has a 13.4% chance of mutating in any given haplotype over 14 generations. That means that if we have 20 project haplotypes, chances are that 20 x .134 = 2.68 of them will have a DYS576 mutation over those 14 generations. However, there are two ways any marker can mutate: up or down. So that we can really expect no more than 1-2 unique mutated values of 576 in this particular case. Still, two independent identical mutations to 576 is one too many, and we must thus be wary of the fastest mutating markers (DYS464, 576, 449, 458, 570, and 456) at least for large patrilineages with deep ancestral roots. The expected number of mutations for the least volatile of these (DYS456) across 20 haplotypes over 14 generations, and in the one (of two) directions that matches the mutation to another haplotype is 20 x 14 .00735 /2 = 1.03 so we are beginning to get onto safe ground with DYS456.

That leaves the dozen or so middle markers (from DYS439 on down in the table below) as the next best candidates to the slow markers for unequivocal singleton Family Branch Signatures.

Identifying a Family Branch Signature: Shared Multiple Mutations

If some of the single markers make problematic candidates for Family Branch Signatures, especially for large patrilineage sets, two or more markers shared by several haplotypes can be expected to do much better in delineating family sub-branches. And although these combinations only begin to appear when the patrilineage is fairly deep ancestrally, where it is shallow, single markers should probably suffice as Family Branch Indicators, since even the most mutable are much more likely to retain their uniqueness.

Combinations of shared markers that include the CDYs continue to be problematic, although one may wish to introduce them into the analysis at the point, with appropriate caution. As a gauge of the reliability of a pair of markers (as opposed to a single marker), consider that the probability that the second and third most mutable markers we are considering here (DYS576 and 449, passing over DYS464[4]) would mutate in exactly the same way over 20 generations in a second haplotype: it is only .7%. However, over a large set of 20 haplotypes, the probability that at least one of them will mutate independently to the same state, thus complicating their use as a family branch signature, is about 13.6%. But this is almost a worst case. Combinations involving less mutable markers or fewer

[4] I pass over DYS464 because it is not clear how one would calculate the probability of a specific mutational outcome, given the many different ways that it could be arrived at.

haplotypes, or shallower patrilineages,[5] should all fall in the probabilistic comfort zone. Therefore, ***in general, shared combinations of mutations, excluding the CDYs, should provide reliable family branch signatures***.

The Representational Conventions of the Mutation History Tree

Having determined what would constitute a reliable family branch signatures for a particular set of patrilineage haplotypes, the next step is to construct the tree itself. The tree runs from top to bottom, with the top occupied by the MRCA of the haplotype set (who may, or may not be the founder of the surname patrilineage). The haplotype of the MRCA, defined by the RPH procedure or otherwise, provides the initial reference values from which the mutations included in the tree deviate, and at least these initial values should be shown in proximity to the MRCA near the top of the tree. The actual mutations included in the tree may be written in an abbreviated form to allow large and complicated trees to be represented across a single line. For example, I use "576+" to represent an "up" mutation of DYS576 from the original state of the MRCA's haplotype ("576-" would represent a "down" mutation). Multistep mutations are written out as separate mutations, one after the other, in accordance with the stepwise mutation model.

Some method must be adopted of representing the branching of the tree. I use the Ascii characters "|" and "-" to draw pseudo lines, with "|" representing the line of descent, and "-" indicating a horizontal or collateral branching off of the main line of descent.

Each family branch is defined by the signature of one or more mutations, and the primary content of each branch (besides the structural components needed to show the branching activity itself) is a list of the mutations that have occurred in each branch, and each branch, or sub-branch, terminates with a list of the patrilineage haplotype IDs who share the cumulative set of mutations written into their branch.

These are the bare bones of the tree, but the intention is to exfoliate it with genealogical information, and a corresponding timeline, as this becomes possible.

Although we know the mutations that have occurred for each haplotype (leaving aside the possibility of silent mutations), if there were more than one we don't know the order in which they occurred, much less when in time. However, when there are multiple shared mutations, it may be presumed that the least likely of the markers to mutate is the one most characteristic of the branch, since there is an increasing probability that the others may have mutated independently. Consequently, multiple shared mutations should be listed in increasing order of mutability.

If the main line of the branch itself has offshoots, defined by additional mutations for certain haplotypes, those branches should be indicated by drawing appropriate lines, and the additional mutations should be written in for the sub-branch(es) following the same rules as for the main branch.

Finally, the identifiers for the haplotype(s) to which each set of mutations pertains should be written at the bottom of either the main branch line, or of one of the sub-branch lines.

---

[5] In most cases, the patrilineage MRCA will not go back as far as 20 generations, although the genealogical patrilineage itself might. A man born 20 generations before the typical yDNA testee, would have been born about 1270, when surnames were just beginning to become general in England, and they were adopted considerably later than other areas of Britain. And besides that, it's likely that over such a long stretch of time that all but one of the male lines of the founder have died out. Thus the probabilities I have calculated above are conservative, erring on the high side. I would estimate that in most cases, the MRCA goes back no more than about 16 generations, and in many cases no more than 6-12. It is well known that many surnames didn't finally gel in Wales or Scotland until the 18th century, only half a dozen generations back, and the K&J study found many lineages with a TMRCA estimate of just a couple of hundred years.

Adding in the Genealogy

The mutation history tree should initially be constructed as much as possible independent of the known genealogy, but hard choices in its construction should be guided by the genealogy where possible. Although incorporating genealogical knowledge may seem to make the implicit argument from the DNA evidence circular, the fact is that the DNA can provide no more than corrective guidance to the genealogy, and the purpose of constructing the tree is to facilitate that guidance.

Where a known ancestor of one or more of the people whose haplotypes are listed at the bottom of the tree can be located by inference either upstream or downstream of certain mutations, the name and (estimated) birth date of that individual should be interpolated at the appropriate point in the branching structure. Adding ancestors to the tree can be expected both to suggest certain research possibilities and also to assist with the further elaboration of the tree and the addition of other ancestors. As ancestors are added, the tree may begin to take on the characteristics of a timeline, and the branching lines of descent may even be calibrated accordingly to reflect this. However, if the tree is to be a reliable guide for any of these purposes, extreme conservatism should be used in adding ancestors to the tree. Please see my analysis of the completed mutation tree in the Robb example on the next page, for some remarks on the concrete benefits of being able to interpolate genealogical data.

Constructing the Mutation History Tree: An Example

I have constructed, below, a mutation history tree diagram for ROBB Patrilineage 2 independently of Fluxus, based only on an analysis of the above haplotype chart, and on my genealogically-derived knowledge of the patrilineage.

This example illustrate that the top of the tree is likely to be its most tricky part, and it may not be possible to resolve it to a high level of confidence, precisely because that is where the genealogical knowledge usually peters out. The Fluxus analysis of this Robb example has fortuitously brought to my attention one possible scenario involving the CDYs that should be kept in mind, and this in turn has suggested an additional principle for choosing between alternative organizational hypotheses for the top of the tree.

Because these CDY markers mutate so readily, there is a significant possibility that one of the CDYs has already mutated prior to the establishment of the RPH that I have chosen as representative of the MRCA or of the founder of the surname patrilineage before him. Ordinarily I would assume that the RPH *is* the ancestral haplotype, because that would be the simplest and most straightforward hypothesis to fit the data. However, Fluxus, which seems intent on in some way minimizing the length of the divergent paths (just how this is measured being left undefined), suggests one additional criterion for selecting from amongst several otherwise equivalent hypotheses regarding the interconnection of sublineages at the top of the tree: to adopt those postulates that would minimize the *imbalance* in the number of mutations down the descendant lines.

In this case, by postulating an initial value of CDYa=37 for the MRCAncestral haplotype, and consequently an initial mutation to 36 for the R-05 RPH, the number of mutations for R-11 and R-12 is reduced from 6 and 5, to 5 and 4, respectively, bringing them back within a more normal range for a surname patrilineage. Meanwhile, an additional mutation of CDYa for R-11 is indicated here in either case, although with the alternative assumption, that the MRCA haplotype was initially 36, the second mutation would have to have been a back mutation. The justification for preferring this more complex version (which, incidentally, violates my rule of listing the less mutable markers first) is simply that the smaller number of mutations within this small haplotype set is more probable, as even with these reductions, the number that remain suggest a very early founding for this surname patrilineage, perhaps as early as the 1100.

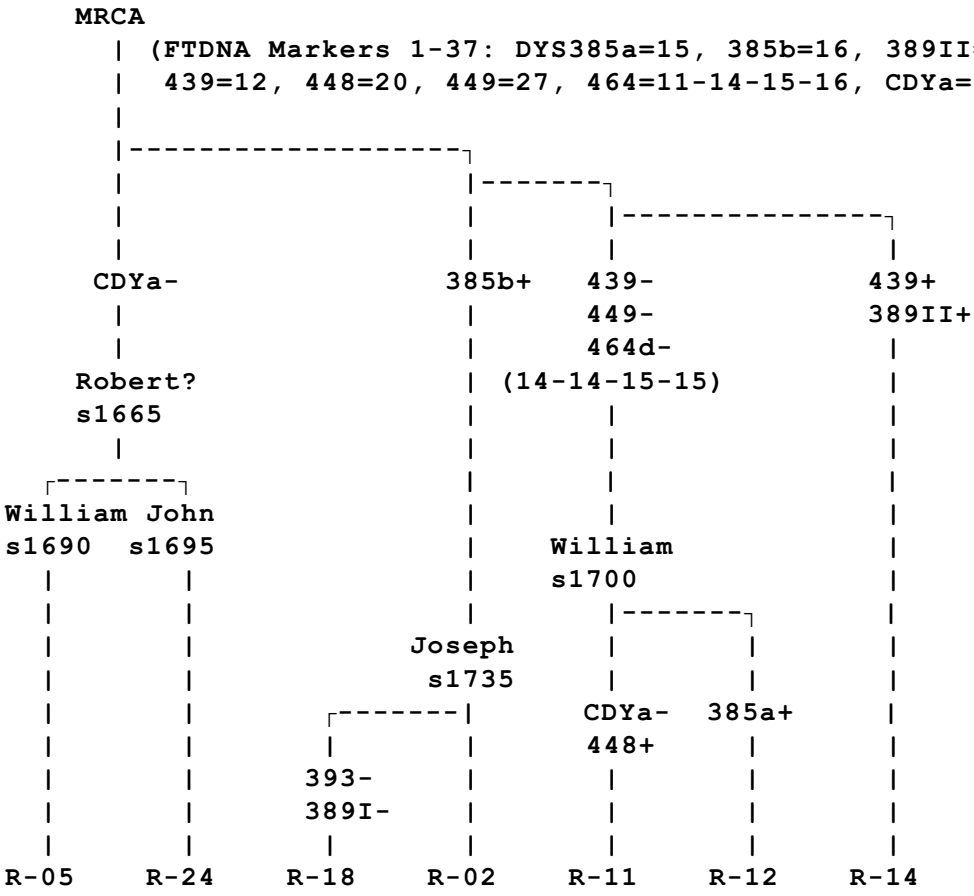
The positioning of the rare mutation to DYS393 so far downstream in the R-18 line (this is my own haplotype as a matter of fact) may seem highly dubious, and it certainly violates the rules suggested above for constructing a mutation history tree, but this is a case where strong genealogical evidence trumps any merely mechanical reconstruction, such as programs like Fluxus offer. On the other hand, the genealogical evidence linking R-11 and R-12 is weak to non-existent, yet the positioning of the downstream markers in the R-11 line, is virtually dictated by the collective heavy freight of the markers upstream of both R-11 and R-12. This is a case where the DNA results provide strong guidance to the research. As for the link between R-05 and R-06, there is quite strong circumstantial genealogical evidence for this relationship, and the identity of the DNA results all but make it conclusive.

Proposed Mutation History Tree of Descent For ROBB Patrilineage 2

constructed by inspection, assuming R-05 to be the RPH that was subject to a prior mutation of CDYa

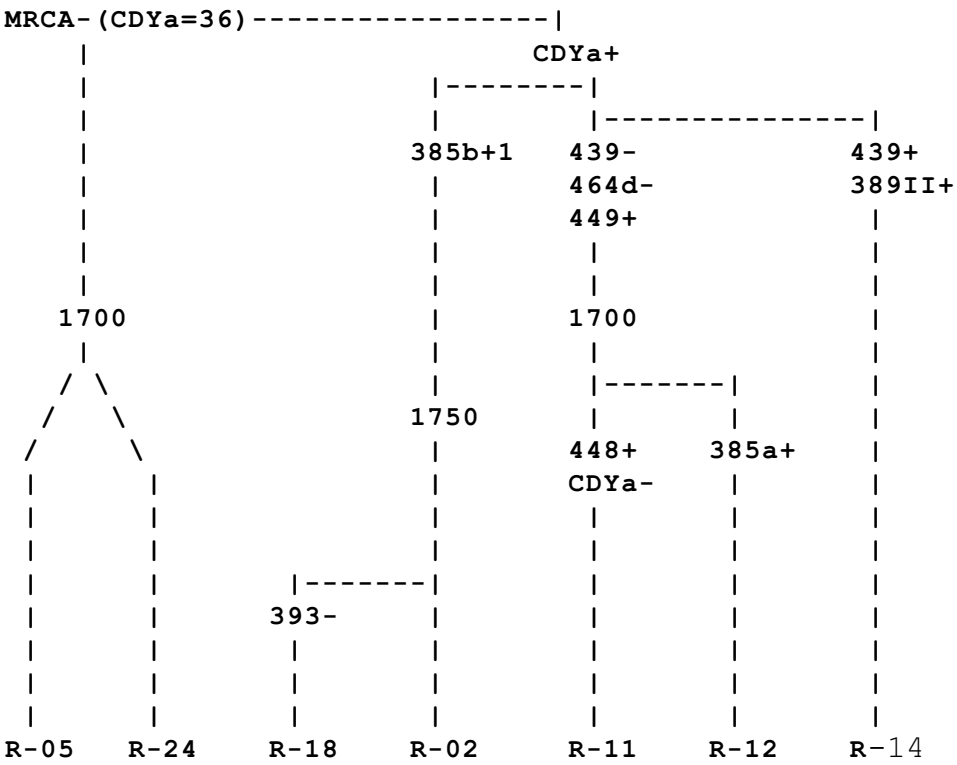
For each mutation in the following diagram, I have appended to the marker ID either "+" or "-" to indicate whether the marker gained or lost an allele value. Where several mutations are listed in succession, they are listed in increasing order of mutability of the underlying marker, although the mutations may actually have occurred in any order.

The year dates represent the approximate time when the family history of these lines begins (the prefix "c" means "circa", and "s" means "say"). Thus, branches shown after those times are grounded in research as well as consistent with this interpretation of the DNA evidence.



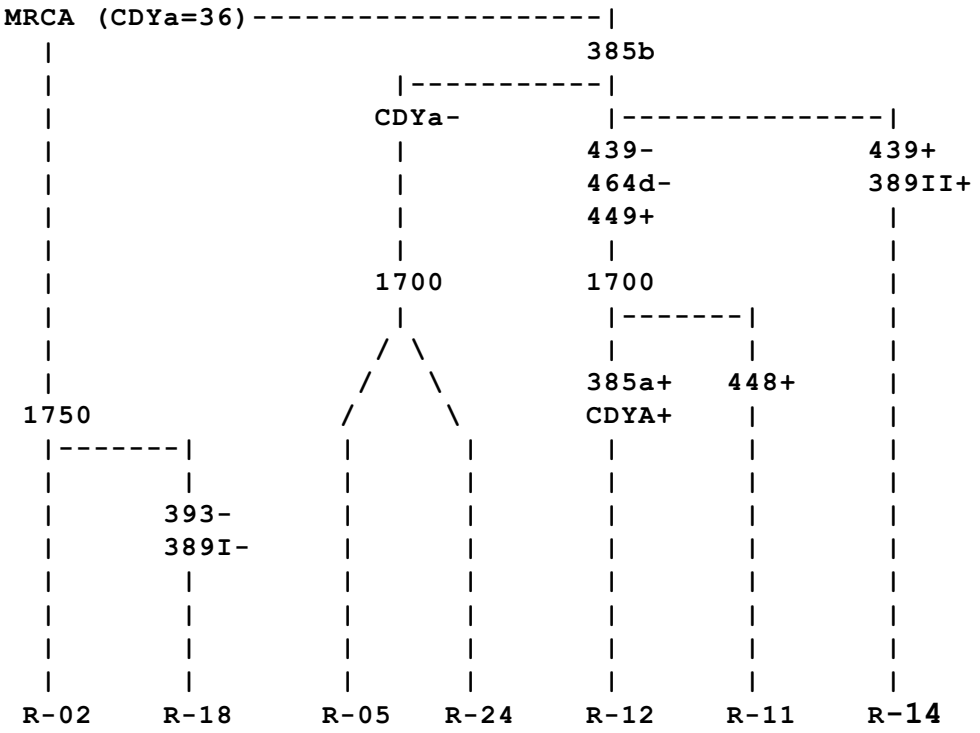
Alternate Mutation History Tree of Descent For ROBB Patrilineage 2

constructed by inspection, assuming **R-05** to be the RPH
(this version, with the CDYa mutation near the top rejected for imbalance)



Proposed Mutation History Tree of Descent For ROBB Patrilineage 2

constructed by inspection, assuming **R-02** to be the RPH
(some version of this tree might come into play if the RPH changes)



Conclusions: the Bottom Line

Cladogram software like Fluxus and Phylip may conceivably be useful for organizing a mass of haplotype data into a first approximation of a likely tree of mutational descent, but Fluxus, as applied above to a typical example of a genealogical patrilineage in an FTDNA Surname project, proved to be both unnecessary and unreliable.

Virtually all of the patrilineages in FTDNA surname projects large enough to be worth analyzing for the light they may shed on the mutation history of the patrilineage, fall into the range of 7-30 members, with the vast majority concentrated near the bottom end of this range. Thus I consider my Robb example above, which has been pruned down to 7 by pruning clusters of known cousins to just one representative, altogether typical of the practical situation facing us today.

And with such a small set, one can readily determine by inspection the main outlines of the mutation history tree just by sorting the haplotypes in a well-designed chart into mutation clusters, and drawing simple inferences from this about which mutations are likely to upstream, and which downstream. Fluxus may be able to do this instantaneously, and draw a nice diagram, but it takes so much time to input the data and set the necessary parameters that much of the speed advantage over a hand-constructed diagram is lost.

As to the finer details of the analysis, Fluxus produced three significantly different diagrams depending on the assumptions made, and the one which was closest to the actual tree (as determined by my exhaustive genealogical research on these Robbs, and on my by-hand analysis) happened to be the one with the most counterintuitive setting of the parameters.

Fluxus, and also Phylip, which I've taken a brief look at in online writeups, suffer greatly from poor documentation, with crucial terms and parameters inadequately defined, and the underlying algorithms unexplained. However, all such software, no matter how well documented and sophisticated it might be, necessarily suffers from a crucial defect: it is unable to take what is known about the underlying genealogies associated with the haplotypes into consideration. Although factoring in the genealogical knowledge when constructing a mutation history trees courts circular reasoning, it is the verifiability and overall plausibility of the results that count, and by this measure, these mechanical procedures fail to measure up. Constructing mutation history trees for genealogical patrilineages, like genealogy itself, is an art, not a science.[6]

---

[6] This is a slightly revised version of a paper written originally back in 2010, and since then my experience in constructing mutation history trees has produced somewhat more elegant examples that the one above. For example, the interested reader is referred to this mutation history tree for Patrilineage 1 of the DENNISON Surname project. The corresponding haplotype chart, and a detailed analysis of its mutations, will be found on the same page.

APPENDIX A: Mutation Probabilities & Chances of Mutation for FTDNA Markers 1-37

For each marker, with Mutation Probability of <MP>
the Probability that it has mutated one or more times over <Gens> generations

$$= 1 - (1 - \langle MP \rangle)^{\langle \#Gens \rangle}$$

		Mutation Percentage		
	Mutation Probability	20 Gens	14 Gens	8 Gens
CDYa-b*	.03531	51.3	39.5	25.0
DYS464a-d	.02264	36.7	27.4	16.7
DYS576	.01022	18.6	13.4	7.9
DYS449	.00838	15.4	11.1	6.5
DYS458	.00814	15.0	10.8	6.3
DYS570	.00790	14.6	10.5	6.1
DYS456	.00735	13.7	9.8	5.7
DYS439	.00477	9.1	5.0	3.8
DYS607	.00411	7.9	5.6	3.2
DYS460	.00402	7.7	5.5	3.2
DYS442	.00324	6.3	4.4	2.6
DYS390	.00311	6.0	4.3	2.5
DYS391	.00265	5.1	3.6	2.1
DYS447	.00264	5.1	3.6	2.1
DYS389II	.00242	4.7	3.3	1.9
DYS385a-b	.00226	4.4	3.1	1.8
YGATA-H4	.00208	4.1	2.9	1.7
DYS389I	.00186	3.7	2.6	1.5
DYS394(19)	.00151	3.0	2.1	1.2
DYS448	.00135	2.7	1.9	1.1
DYS459a-b	.00132	2.6	1.8	1.1
YCAIIa-b	.00123	2.4	1.7	1.0
DYS437	.00099	2.0	1.4	.8
DYS393	.00076	1.5	1.1	.6
DYS438	.00055	1.1	.8	.4
DYS392	.00052	1.0	.7	.4
DYS388	.00022	.4	.3	.2
DYS454	.00016	.3	.2	.1
DYS455	.00016	.3	.2	.1
DYS426	.00009	.2	.1	.1

* The mutation probabilities and percentages shown are for each multicopy marker component.