Why the RPH should replace the Modal as a reference haplotype for yDNA projects

I am proposing here a construct that I call the "root prototype haplotype" (RPH) to replace the "modal" as a reference haplotype for a collection of tested yDNA haplotypes with a common genealogical patrilineage.[1] That we need a reference haplotype is undeniable, unless the raw data are to remain just a set of meaningless strings of numbers signifying that the tested individuals are patrilineal cousins.

We need a reference haplotype in the first place in order to be able to highlight (usually through color coding) marker value deviations across the set of tested haplotypes. And what we really want to do is to be able to think of the highlighted deviations as *mutations* from the haplotype of the common patriarch of the lineage—the MRCA.

Adopting the modal value(s) of the set as the reference haplotype, as is the current convention, is both misleading and inappropriate, because the most common values of the set are to a large extent a function of the self-selection bias that induces first one, and then others, of a group of allied researcher cousins to volunteer for testing. And even where test volunteers accrue more "randomly", the modal concept doesn't begin to grapple with the possibility that the majority of the surviving descendants of an ancestor who lived, say, ten generations ago, may bear a highly skewed load of mutations compared to the broad range of descendants.

Not surprisingly for such an ill-conceived application of an irrelevant statistical concept, references to "a modal haplotype" is also ambiguous: is that supposed to mean the most common haplotype in a set of haplotypes, or is it supposed to refer to a synthetic haplotype constructed from the mot common (modal) values of each individual constituent marker across the set? Many FTDNA surname projects duck the issue of a reference haplotype by abstracting from the set of markers those which are identical across the set, while merely noting the variant values of the others, without bothering to try to determine which values are mutations and which the original (unmutated) values. But this slack approach frustrates any attempt to determine the historical pattern of mutations down the known geneealogical generations, and so provide a purely DNA basis for sorting yDNA-tested descendants into particular family sub-branches.

I would hope that the more experienced project leaders have largely, and long-since, emancipated themselves from modal thinking (though they mostly seem to cling to this inappropriate terminology), and have adopted the "triangulation" metaphor in its place. But while that is a vast improvement, triangulation isn't quite the thing either. In the first place, the metaphor doesn't quite fit. Triangulation is a process of geometric construction by which an unknown point is determined by projecting two known points, each in a given direction, so as to form convergent lines that intersect at the point to be determined. But which, exactly, are the known points amongst a collection of variant haplotypes, and how, exactly, do we determine the specific "directions" that reliably determine the MRCAncestral haplotype?

My alternative, the "root prototype haplotype" at least puts the right name on the construct we are truly aiming at. Whether my method of deriving it be the best one remains to be seen, but because I am not relying on an inappropriate metaphor, at least I am able to specify such a method, and so far I have found it to be far preferable analytically to the confused attempts I have seen to twist the modal or triangulation concepts into what they ought to be.

As an approach to determining the RPH, I propose that, of the current collection of haplotypes, the one that is most closely related to all the others collectively be adopted as the MRCA prototype. I will present a specific procedure below, for calculating the RPH. As an alternative, though not

[1] Ideally, surname projects should be broken down into a set of individual patrilineage projects, and I shall use the term "project" hereinafter to refer to a genealogical patrilineage project, rather than to an entire surname project.

necessarily mutually exclusive, approach to positing an RPH, I will also discuss creating a synthetic haplotype either from scratch—from the most likely set of individual marker values—or by tweaking the selected haplotype to reset any apparent mutation(s) it has picked up to the most common value of the set.

As additional haplotypes accrue to a project, the RPH, can be expected to change, until it eventually settles on one that is at least the closest approximation to the original root MRCA haplotype that we are likely to get. Hopefully, in many, if not most cases, the eventual RPH will be identical to the actual root haplotype, though we may never be able to tell for sure.

If the sole purpose of having a reference haplotype were to be able to fix on some standard against which we might highlight deviant marker values, then any reference haplotype would do—the modal, the first one listed, whatever. With the RPH, though, we can highlight certain deviations as ***mutations***, and provide in our haplotype chart the means for sketching the outlines of the actual genetic tree by considering the number and patterns of specific mutational deviations from the RPH across the member haplotype set. Characteristic patterns in particular branches should appear and the length of the branches, measured in # of generations or years, should correspond, at least very roughly, to the number of mutations, or better, the genetic distance. I have developed a procedure and a format for doing this in my paper on mutation history trees.

For purposes of calculating the RPH it is important to prune the set of descendant haplotypes of those known to be close cousins, thus eliminating any mutations that may have occurred during the most recent, genealogically worked out, generations. There figures to be a tendency for genealogists working on the same sub-branches of the tree, some of them whose interrelationships are known, to discover the benefits of yDNA testing in the same early stages of a patrilineage project, and by thus flooding the project with their own similar haplotypes, creating a self-selection bias within the full set of project haplotypes. Since the purpose of testing is to extend the genealogical tree back in time, into the unknown, or problematic, generations, basing the RPH determination on mutations which can be inferred to have occurred downstream, during the known, more recent generations, will in many cases cause the wrong haplotype to be chosen as RPH. I will have more to say below on pruning, or actually truncating, the base of the tree.

Selecting the RPH from a set of haplotypes

Dean McGee's Y-Utility can be used to generate genetic distance charts that facilitate the computation of RPH by a process of simple addition and comparison. The input to Y-Utility is the actual set of haplotypes for the patrilineage.

Y-Utility also has the great merit of allowing data for a whole large set of haplotypes to be cut from a Y-results display, and pasted into its input box, after a few tweaks to ensure that each of the row headings is exactly one continuous string, and the marker values are each separated by exactly one space. I first paste my Y-result gleanings into a text file, wherein I join all the separated elements of the row/column headings with dashes or dots and make them equal length, and I then proceed to align all the marker values, leaving exactly one space between each (Y-Utility is picky and will generate an out-of-range error for your marker values if you fail to do this). For example:

```
D05-Alan-Daniel,b.s1688... 13 24 14 11 ...
D04-JohnA................. 13 24 14 11 ...
```

In its default mode, Y-Utility generates a variety of charts. The easiest one to calculate from is the Genetic Distance report, which comes out in the penultimate position. A set of parameters allows one to specify various mutation rates and other variables, but only one of these has an effect on the computation of RPH—the choice of the "infinite alleles" or the "hybrid mutation model" (I choose

the latter). One non-default option I always take is to uncheck "Create Modal haplotype" to eliminate this meaningless and distracting addition to the data, and I also set "Highlight Reference" to "None". Here is an example in which I have reduced the headers to simple project numbers:

Genetic Distance					
ID	D-06	D-05	D-01	D-04	D-02
D-06	37	1	4	5	7
D-05	1	37	3	4	8
D-01	4	3	37	3	7
D-04	5	4	3	37	6
D-02	7	8	7	6	37
Related	Probably Related	Possibly Related			

The diagonal line represents the number of markers used in each cross-comparison. The color coding helps guide the eye to the most promising columns (or rows—the calculation can be made for either), with the closest relationships coded green, then yellow, red, and beige. D-05 here wins by a nose with a low total of 16—the best single candidate to represent the actual haplotype of the common ancestor of these tested descendants. Where the totals are equal for two or more haplotypes, I choose the one which has been extended to the greatest number of markers as the RPH.

With such a small data set, the RPH is likely to change a number of times as additional members accrue. However, each change can be expected to improve the estimate, and eventually the RPH will become stable, and hopefully optimally representative of the haplotype of the original MRCA of all.

This procedure for choosing the RPH can also give us a feel for the outer dimensions of the tree. For example in the subset of the ROBB DNA Patrilineage 2 group that I've used in my paper on Mutation History Trees, there are seven members, with six different haplotypes. The RPH is R-05, and the two most divergent are R-11 and R-18, which have a genetic distance of 7. In fact, without R-05, from which they each diverge by three, we might be inclined to consider them as constituting different lineages. Considered in relation to the RPH, though, R-11 and R-18 just define the present outer limits of a single tree.

Pruning, and Truncating, the Tree, in Preparation for RPH Determination

I have already noted that sets of close cousins should be collapsed into a single representative of their line to minimize self-selection bias, and to weed out from the RPH determination procedure any mutations that can be inferred to have occurred downstream, during the recent generations when the genealogy is known. I would just like to emphasize here that this pruning should be pushed as far upstream as possible—consistent with a critical and conservative assessment of what is genealogically *known*. By pruning, I mean discarding all but one of the haplotypes of this subset of known close cousins. Where these haplotypes diverge, those which vary from the subset, and/or from the norm of the complete project set, should be the first to go, since one may infer with great confidence that any variations represent mutations that have occurred in the most recent generations since the known MRCA of the subset members.[2]

In fact, for purposes of determining the MRCA of all the current (and even possible) patrilineage cousins, we would ideally want to work with a subset of haplotypes, each representing the more recent MRCA of subsets of genealogically known cousins. And it is with that in mind that we take our pruning shears to the base of the tree. The results of such an operation are that we have in effect truncated both the genealogical and the genetic base of the tree, thus attaining a tighter focus on its upper, unknown, reaches. What we are interested in, after all, is in determining is the RPH at the top

[2] I realize that there is a bit of formal circularity here in my use of the term "project norm" since it is that (the RPH) which we are trying here to determine. In practice, though, there is no difficulty in identifying a unique allele value found in one or more haplotypes of a close cousin cluster as a mutation that has occurred just in the generations since their mutual MRCA.

of the tree, and for that purpose, the closer to the top of the tree we can begin our analysis, the more likely the RPH procedure is to yield the closest possible approximation to the actual haplotype of the MRCA of all.

It may be worth stepping back at this point and recollecting what we are attempting to use yDNA testing for, beyond the straightforward classification of a haplotype into a patrilineage (or, with enough testing and luck, into a particular familial sub-branch): it is to help elucidate the relationship probabilities at the amorphous top of the tree, where the genealogical research becomes difficult and problematic, and the evidence scant. At one extreme, there are some very mature surname and patrilineage projects where the genealogy is almost perfectly known back to earliest generations. Typically, these are large patrilineages, with dozens of haplotypes, yet applying this pruning and truncating technique the best determination of the RPH may be based on just a handful of haplotypes far up the tree where the genealogy becomes murky.

At this point, the tree whose base I am talking about truncating is the mutation history tree, and determining the RPH, and with it, which haplotype marker values are mutations, is only the first step in working out the mutation history tree. As genealogical and genetic knowledge hopefully accumulates, the tree is gradually leafed out and filled in, and truncation farther up, with a re-determination of RPH enabled, in a series of iterations.

Determining the RPH by this procedure is the easy part. The real art lies in the construction of the mutation history tree.

Further Thoughts on the Modal Haplotype vs. a selected RPH

In analyzing large patrilineages of 15 or more, I have found that the modal value tends to converge with the RPH. This suggests that over the period of genealogical time that we are interested in, for many if not most lines, either the original root haplotype of the surname founder, or a close and early variant, has tended to survive in large enough numbers to predominate over later variants. And to the degree that this is so, the initial modal value may be a reasonable choice for the RPH. However, most patrilineages are much smaller than that, and there are altogether too many "if"s here for me to be comfortable relying on the modal value, especially if no attempt has been made to prune close cousins and thus avoid self-selection bias.

Carrying the methodology described in this paper farther back, before genealogical time, becomes problematic as the increasing probability for the more mutable markers to back-mutate to their original state (leaving no trace that they have ever mutated) renders the choice of any particular descendant haplotype as representative of the haplotype of an ancient, pre-surname patriarchal MRCA increasingly wide of the mark. Under these circumstances, single, central modals are likely to be replaced by multiple nodal haplotypes, and while the RPH procedure will still choose the haplotype closest to the average GD of all, the RPH haplotype itself will predictably develop additional mutational deviations from that of the original.

The nightmare scenario for any attempt to determine a reference haplotype for a set of surviving descendants occurs when one or more mutations occur in the first few generations after the MRCA, and certain other conditions are met that produce an asymmetrical descent tree.

In the worst case, suppose the MRCA has four sons, one of whom picks up a mutation, and that this son and his descendants happen to be the most fertile, leaving the greatest number of descendants. Then suppose, to make the situation worse, that the lines of two of his brothers die out completely. Under these circumstances, most of the descendant haplotypes will carry the mutation, which will appear, by any method of determining the reference haplotype, to be that of the original MRCA. Furthermore, since even a well-worked-out genealogy is likely to stop short of this ultimate MRCA generation, there will never be any way of proving otherwise. Even in cases where the first mutation

to this patrilineage occurs a generation or three downstream, there is still a good chance that asymmetrical fertility, or other factors that bias the test sample, will produce similar distortions that render the choice of reference haplotype misleading.

However, if instead of choosing the modal haplotype to represent the MRCA, one were to apply a progressive iterative (recursive) procedure of determining an RPH, working out a mutation history tree, then as the tree became filled in, truncating its base, and recalibrating the RPH, one might hope to push the genealogy back far enough to get a fix on such early mutations. Although the subset of high-level haplotypes got smaller with each truncation of the tree, the modal value would become increasingly unreliable and ambiguous, my RPH-selecting procedure would continue to select that one of the remaining haplotypes that was most closely related to the others.

An Alternative Approach: Constructing a Synthetic RPH

Many ySTR DNA analysts have, I believe, constructed patrilineage reference haplotypes from the most common (or modal) values of each individual marker across the current set of haplotypes, if only to provide resolution in situations where all haplotypes are unique, or where there are more than one equal size cluster of identical haplotypes. There is still, in this approach, danger of self-selection bias, but also still, as the number and more important the diversity of haplotypes increase, a probable convergence of the resulting synthetic RPH to the actual root haplotype. This approach can be improved upon by, instead of slavishly adopting the modal value of each marker, considering the overall patterning of resulting mutations against the backdrop of known genealogy. Another factor that ought to be considered in creating a synthetic haplotype is that where a particular marker has three different values, that by the laws of probability, it is more likely that the middle one is the root value, with mutations occurring in both directions from it, up and down.

Tweaking the Selected RPH in Consideration of Modal Marker Values

As often as not, an RPH determined by selecting the one that is closest collectively to all of the other haplotypes in the set will nonetheless have one, or perhaps even two, marker values that are unique or nearly so. In such cases, it is reasonable to create a synthetic variant of the selected haplotype that eliminates these individual marker anomalies by resetting them to the most common (modal) or the most likely value for the marker across the set. This might be considered a hybrid approach, though technically a selected marker with even one marker value changed, would have to be considered a synthetic haplotype.

A Synthetic RPH versus a (Tweaked) Natural RPH

One might even consider such a technically synthetic RPH superior, because it abstracts the RPH from any particular descendant haplotype, but it is important whenever a synthetic RPH is used in preference to a selected, natural, one, that some account be made of the principals followed in constructing the synthetic, or in deviating from the selected haplotype with respect to one or two markers.

Conclusions

In consideration of the above, I propose the replacement of the term “modal haplotype” as the reference haplotype for a set of tested patrilineage haplotypes, by the term “root prototype haplotype (RPH)”. I further offer my procedure for selecting the best natural haplotype to represent the test set’s RPH, based on its being the haplotype most closely related to all the others collectively. I also suggest tweaking this selected haplotype if it appears to include an anomalous (probably mutated)

marker value or two, by resetting that (or those) marker values to their most likely original value(s) thus creating from the selected haplotype a synthetic haplotype that closely resembles it.

Key to making the best selection of the RPH from the tested patrilineage set, is to prune out all but one representative of each genealogically known sub-branch of the family, and with it, discarding all marker variations that can be reliably inferred to fall within the known family branch.

Finally, to make the most of this method, the RPH should be re-determined whenever significant additions are made either to the genealogical or the genetic knowledge of the patrilineage.[3] Where the significant additions are genealogical breakthroughs, connecting up previously unconnected sub-branches to the main tree, or otherwise completing the base (most recent) part of the tree at a higher level, the entire new known tree should be truncated, with each of it's branches represented by a single haplotype and the RPH re-computed from those.

[3] One can usually determine when this needs to be done without actually carrying through the calculating procedure.