

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 patrilineage. 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 which 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, the modal is also ambiguous: should it be the most common haplotype of the set of tested subjects, or should it be constructed, scattershot, from the most common marker values of the set? Some project leaders, alternatively, abstract just the common marker values from the set, leaving the other “modal” marker values indeterminate, but this not only fails to provide any kind of a haplotype which might be considered a prototype for the that of the MRCA, it also provides no reference haplotype against which deviant marker values might be highlighted.

I believe that the more experienced project leaders have largely, and long-since, emancipated themselves from modal thinking, 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 which 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” which reliably determine the MRCAncstral 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 which 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 additional haplotypes accrue to a project, the prototype haplotype, the RPH, can be expected to change, until it eventually settles on one which 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 root haplotype, though we will never be able to tell for sure. As it turns out, though, this doesn’t matter too much. There is no way of reconstructing the original haplotype, for example, where a son or grandson of the MRCA who picked up a mutation became the sole surviving genetic representative of his line. Yet even in such a case, the fact that RPH deviates from the original is *genealogically* irrelevant, since we seek only to type the survivors.

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 sketch 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. I have developed a procedure, and a format for doing this in my paper [Mutation History Trees](#).

Determining the RPH

Dean McGee’s [Y-Utility](#) provides a way to compute RPH mechanically, by a process of simple addition. The input to Y-Utility is the actual set of haplotypes for the patrilineage. However, to minimize selection bias, and to prevent the modal from sneaking in by the back door, all but one of a set of close cousins should first be winnowed out, by choosing just one which is closest to the other haplotypes as representative. Right now I’m leaving out just 1st and 2nd cousins, but I may decide to widen that framework down the line. This step insures that the RPH is largely a function of the actual variance in the data set, with the focus kept on the original MCRA, not on anomalous mutational happenings downstream.

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	D	D	D	D
	0	0	0	0	0
	6	5	1	4	2
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.

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.

This procedure can also give us a feel for the outer dimensions of the tree. For example in the [Robb DNA Project, Lineage 2](#) 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. I have outlined this patrilineage as an example in my paper on Mutation History Trees.

The Desirability of Truncating the Base of the Tree: Generalizing the RPH Procedure

I have already noted that sets of close cousins should be collapsed into a single representative to minimize self-selection bias, with the one cousin whose haplotype is most like the others in the entire patrilineage set being chosen as representative of the others. Further reflection has convinced me that this sort of pruning of the base of the tree should be extended as far as possible, ideally to truncate the whole base of the tree as far up as possible, along with the downstream mutations which harbor there—*where and when we can identify such mutations as downstream*. What we are interested in determining is the RPH at the top 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 optimum result. This is not a step to be taken lightly, however. Determining that certain mutations are, in fact, downstream can only be done by a combination of extensive testing, and the construction of a mutation history tree which incorporates genealogical knowledge with the DNA data.

Further Thoughts on the Modal vs. the 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](#) which 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 aren't that large, 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.

Conclusion

In consideration of the above, I propose the abandonment of the inappropriate concepts of the "modal haplotype", and of "triangulating" to the theoretical MCRA haplotype, in favor of a mechanical procedure which can be readily followed each time a new result is posted, to more reliably identify the current "root prototype haplotype", or RPH.