How does one tell which of two (or more) marker values is the mutation?

The most straightforward and natural way is to just tally up the majority vote and label the minority marker value as the mutation, but until the patrilineage is analyzing reaches a decent size (say 7 or more) and is sufficiently diverse (with a maximum of ancestral lines that remain independent many generations back into the past), this simple procedure is perilous and liable to reversal. At a certain point in the growth of the patrilineage, though, as the majorities pile up, it becomes evident that one has made the right choice for the mutation.[1]

However, in cases where the mutation occurs near the top of the tree defined by the MRCA (Most Recent Common Ancestor) of the whole set of patrilineage haplotypes, determination of which is the mutated value again becomes problematic.

Lets consider an example. Suppose you have 7 haplotypes and they are identical except that one of them has a different value at one marker. One would unhesitatingly tend to call this a mutation, and the other 6 values the "normal", unmutated type. But suppose that 3 of the 7 haplotypes had the deviant value. Could you be so sure that the value of the 3 was the mutation, while the value of the 4 was the normal, unmutated, value?

In the nature of things, the descendants of any particular MRCA are very unevenly represented in the living population. If the MRCA of all goes back far enough, say 600 years, most of his patrilineal (all male) descendant lines will have died or daughtered out, but (on the average) one or two will have proliferated extensively. This isn't just vague theorizing; a study published in 2009[2] has discerned this pattern in the descendancies of some 40 British surname patrilineages, sampled at random. Now suppose that the MRCA began with three sons, A, B, and C, and that B experienced a mutation, while the other sons passed on the haplotypes they inherited from their father, unchanged. Further suppose that the descendancy of the A line eventually died or daughtered out, while that of C became greatly attenuated, leaving only a few male descendants. Under these circumstances, the vast majority of those who might show up and DNA test (assuming more or less random self-selection) would bear the mutated value, which to the modern DNA analyst would look like the normal, unmutated value.

Thus, we can never be sure that the most common values of any of the markers in a particular set of tested haplotypes are in fact the unmutated values, although more often than not the inherent probabilities, and "the law of large numbers" (where we have them) ensure that they are.

When we can't be sure which are the mutations, how do we determine the RPH?

This issue of determining which of two marker values is the mutation is intertwined with the issue of determining the overall haplotype of the MRCA of all. Of course if we knew what that haplotype looked like, it would be obvious which descendant marker values were the mutations. But short of exhuming the MRCA and testing his DNA directly, there is no way that we can know what his haplotype looked like, except through probabalistic inferences from the set of descendant haplotypes. For the most part we can infer which downstream marker values are the mutations to a high level of confidence , by considering both the overall mutational pattern, and the genealogy. But mutations that occur in the first couple of ancestral generations have no history to speak of, and thus would be subject to inference only if we could go back and test the man in the example above, and his three sons, A, B, and C.

---

[1] This is also the point at which one's choice of RPH (Root Prototype Haplotype) begins to become stable.

[2] Turi E. King and Mark A. Jobling, "Founders, Drift, and Infidelity: The Relationship between Y Chromosome Diversity and Patrilineal Surnames", in *Molecular Biology and Evolution* 26(May2009):1093-1102

Many project admins don't even try to solve this problem, which doesn't do the members of their project much good. Most of those who do, just pick the most common haplotype in the test set, the "modal" haplotype, and presume that this is the haplotype of the MRCA, and indeed, it often is. However, with just a few haplotypes (7 or less), or on the other hand, more than 15, the most common haplotype may not be the same as, or even the closest one to, the original haplotype of the MRCA, and with 111-marker haplotypes, all of them are likely to be unique, so there's no modal haplotype to select.

And without being able to analytically determine which variant marker values are mutations, the whole project of analyzing the mutational patterns that run through a set of patrilineage haplotypes to pick out the one emblematic of family sub-branches collapses. Some admins fall back on cladistic programs that purport to be able to infer the structure of inheritance trees from a set of descendant haplotypes, but these programs fall down in a number of ways, as I have argued, and illustrated in my paper "Fluxus Network Diagrams vs Hand-Constructed Mutation History Trees". To my knowledge, I am the only yDNA analyst who attempts to construct MHTs (Mutation History Trees) (my coinage) that take into consideration both the patterns of mutations and the collective genealogies behind all the patrilineage haplotypes, although I have picked up glimmerings of understanding of this issue on the ISOGG list for FTDNA project administrators by other thoughtful genetic genealogists. Although it is not the largest patrilineage, my best example of a mutation history tree to date is the DENNISON Patrilineage 1 Mutation History Tree: because of the relative genealogical maturity of this project, with many upper-level genealogical relationships having been worked out, it has been possible to interleave the names and dates of more intermediate patriarchs of this lineage than in any of my other MHTs to date. I only wish that there were more room for genealogical information on these trees, and that they weren't so time-consuming to construct by hand.

To return to the subject of identifying a standard haplotype that may be used to determine which variant marker values are mutations, I would also note that the modal haplotypes that most DNA project admins use who do anything whatsoever analytical with their data is subject to sampling bias in cases where many closely related genealogists, who might have gotten in touch before they ever thought of DNA testing, all test about the same time.

I've worked out an entirely different approach to determining the normal, or original unmutated, haplotype. Instead of selecting the most frequent haplotype in the project, and perhaps flipping a coin in case of ties, or on the other hand, synthesizing a haplotype from the set of the most common marker values, I select the standard haplotype by means of a procedure for determining which of the set of haplotypes is closest to all the others, collectively.

My term for the standard unmutated haplotype (used for determining which markers are mutated and which are normals is the RPH (Root Prototype Haplotype. This term acknowledges in its very name both its own meaning, and its own inherent imperfection. The RPH is an organic, actual haplotype, not some synthetic construct that may not even exist in the real world. And it's not necessarily the most common one in the set of haplotypes: it is the one haplotype of the test set that I believe to be the closest one to the actual Root Haplotype—the haplotype of the MRCA of all.

My selection of RPH[3] is based primarily on a tabulation of the GDs of all the haplotype pairs, with the haplotype having the lowest GD total (which means that it is the one closest to all the others, collectively) getting the nod—unless there is genealogical evidence pointing in a different direction. If the tabulation produces a tie, I break it by selecting (a) the person who has tested out to the greatest number of markers, or (b) (if there are more than one of those too) the one with the deepest and

[3] See my paper on the RPH for more on my method of determining the RPH, and for an example.

solidest genealogy. As it happens my methodology rarely produces a tie, and once a fair number of haplotypes has accrued to the project, the RPH seldom changes after that, although I reconsider it each time new members come in. My methodology would also be somewhat subject to distortion from sampling bias with smallish sets of haplotypes, except that before calculating the sums of the GDs to select the RPH, I first weed out all but one representative of each close cousin cluster.

My method of selecting an organic RPH—the best candidate of a set of haplotypes—to be the prototype of the MRCA of the set, means that I may end up with a haplotype that itself has a mutation, and thus deviates from the actual haplotype of the root ancestor, and it's important, due to the inherent indeterminacy of mutations near the top of the tree to keep this possibility always in mind.. However, for purposes of guiding genealogical research and theorizing, it doesn't actually matter much whether the RPH itself has a mutation. As long as it continues to be the one that is most closely related to all the other haplotypes collectively, it remains the best prototype.

The Genealogical Value of High Upstream, "Watershed", Mutations

What does matter, and it matters greatly, is that these top-of-the-tree mutations amount to genealogical watersheds, with all the bearers of one value falling on one side, and everyone else on the other side. Thus, where they exist, these are the most valuable of all mutations because they provide information that potentially affects every haplotype of the patrilineage, even where the details of the top of the tree aren't known. However, as with other mutations, just knowing that there is such a divide can help guide genealogical research to map it out.

There is just one problem with watershed mutations. Because they occur so early in the tree of descent from the MRCA of all, they afford the maximum chances for the same mutation to occur independently in one of the originally unmutated lines of descent, thus counterfeiting the watershed mutation. The chances of this, as usual, depend on the mutation rate of the particular marker involved. In fact, for this reason, the ultra-fast mutating CDYs cannot practically be used as watershed mutations, and the next fastest mutators, with DYS710, and DYS712 (in the 68-111 marker band), and also DYS575, 449, 534, 458, and 570, being suspect, in decreasing order. Over 14 generations (which would push the birth date of a typical MRCA back to about 1450), the chances that these fastest mutating markers would mutate spontaneously in any one of the unmutated lines, and in the right direction (up or down) to counterfeit a watershed mutation, range from about 5% for DYS710 to 2.5% for DYS570, which doesn't sound like much, except that if there were, say, 10 such unmutated haplotypes, the chance that any one of them would mutate spontaneously in the right way to counterfeit the watershed mutation would range from 37% down to 20%, perhaps taking other descendent haplotypes with them into the counterfeit camp, so it is good to keep these possibilities in mind.

For the most part, though, watershed mutations are bankable indicators of a major, far upstream, division of the patrilineage into two broad sub-lineages, and since they tend to divide fairly evenly, they can benefit those genealogists whose lines remain unconnected to an early patriarch by cutting the universe of connection possibilities in two. And for those attempting to determine the relationships between the misty, murky, unconnected patriarchs at the top of the tree, they can likewise be a godsend.