Preface

I present below the fundamental principles I've derived for making the most of yDNA testing for genealogical purposes.

Although the more specific sections of my argument below, and all of my examples, pertain to Family Tree DNA's popular family of ySTR (short for yChromosome Satellite Repeat) tests (the 37-, 67-, and 111-marker tests), the underlying principles for deriving genealogical inferences apply equally well to the ySNP portions of FTDNA's BigY tests, or to competitor's ySNP tests. The definitional links are to the glossary of terms on my DNA testing information page, which provides an introduction to many different types of testing, as well as links to texts and papers like this one that are relevant to the interpretation of these tests. However, it is my goal here to bring together in one place everything that is germane to the interpretation of a set of yDNA results for a particular genealogical patrilineage.

Besides "patrilineage" there are a number of other terms related to DNA testing for genealogical purposes that it's essential to understand, and for all of these I have provided hyperlinks below at the point where they are first introduced.

The other prefatory point I wish to make is that much of my advice below runs counter to what you will read elsewhere, or even what FTDNA (the testing company I've always recommended for ySTR testing) lays out. All that I can say is that I worked out these principles many years ago and am unaware of anything that has happened since that would require me to change them, and that over the years I've seen signs of others, including FTDNA, coming around slowly to my point of view. Because one of my constant objects in all things is the pursuit of truth and genuine understanding, I will be grateful to anyone who can provide me with new information about these matters, or suggest better ways of thinking about them.

Finally, the fact that there is much in what follows that constitutes implied criticism of FTDNA and its works should not be construed as undermining my admiration and gratitude for what the founder of the company, Bennett Greenspan, and his many helpers have created for us. No one expects the pioneers in a new field to anticipate the best ways of making use of their discoveries and inventions. Rather it is their lot to stake their all on their vision and to hack through and blaze the trail to enlightenment and a brave new world, while it is for those who come after, the users (and particularly the many outstanding ISOGG list activists), to figure out how best to make use of those inventions.

Introduction: Key Concepts and Terms

The most essential term pertaining to ySTR testing and its uses for genealogical purposes is (genealogical) patrilineage. The term is my coinage, at least the way I have defined it, but the concept it designates is just as essential to others who use different terms or none. It is mostly one's patrilineage, not one's patrilineal surname per se, that genealogists care about, which is why I've broken down my FTDNA Surname Projects into sets of standalone patrilineage projects—although I also provide an umbrella page for the surname, and the surname project in general.

The DENNISON Surname DNA Project may serve as an example of the template I've worked out to use for all my projects. Because the individual DENNISON patrilineage projects are each standalone projects, I can use the templates I've created for those pages to create and manage patrilineage projects that are merely satellite to other established FTDNA Surname projects, like the two ALLEN Patrilineage projects I've set up, and I'll be using some of the data in these independent patrilineage projects as illustrations of my principles respecting the interpretation of ySTR test results.

Here are some more essential terms.

The commonly used term MRCA (Most Recent Common Patrilineal Ancestor), is the most recent patrilineal progenitor of all the members of a set (or subset) of tested male descendants in a patrilineage project. The results of an FTDNA test of a male's ySTR sites is a ySTR haplotype—a panel of 37, 67, or 111 ySTR marker values. It is because these marker values rarely mutate that the results of an FTDNA 37-marker test can be used to sort descendant males into their particular patrilineage, and it is because a few ySTR mutations can nonetheless be expected to occur over genealogical time that a patrilineage can sometimes be further divided into sub-branches just on the basis of these test results. Where this is possible, I have coined the term CCC (Closer Cousin Cluster) to represent a subset of descendant haplotypes within a patrilineage project that are more closely related to each other than to any of the other members of the patrilineage, because they all descend from a common MRCA who lived closer to them in time than the MRCA they share with all the other members.

The RPH (Root Prototype Haplotype) is my term for the MRCA's projected ySTR haplotype; the term more commonly used term is "modal" haplotype, or marker value, but this is a misnomer because there's no guarantee that the most common inherited marker or haplotype values in a set of haplotypes are the original ones of the MRCA, or even provide the best approximation to his haplotype.

Determining an RPH for a set of haplotypes is essential to differentiating mutated marker values from normal (or inherited) ones. There are various ways of doing this, but the most common way, which works well when there are enough haplotypes extended to a particular level (37, 67, or 111), and when it is known from the corresponding genealogies that the ancestries of all of those tested diverged many generations back, is to synthesize the RPH from the most prevalent (or, if you like, the modal) values. However, in cases where there is no overwhelming preponderance of a particular value, the most prevalent value may not be the best choice, and what does one do when there are two eequally prevalent values, or three pluralities of values? In such cases, discretion and consideration of both the overall mutational pattern and the associated genealogies is called for.

I will also have much to say below about TMRCA (Time back to the MRCA), whether expressed in generations or years, and GD (Genetic Distance), which in this context of ySTR testing is popularly taken to mean the number of different marker values in the test results of two haplotypes, for example, 35/37, or 105/111. These overt quantitative GD differences are typically equated with the number of mutual mutations that have occurred since the genetic divergence in the ancestral lines of two tested descendants, but as we shall see, counting mutations isn't quite as simple as that. If we take GD at its face value though, TMRCA is simply a function of GD. A probability calculation is applied to a GD difference based on the assumed rates of mutation of the markers involved, or of the marker panel as a whole, and from this the number of generations necessary to produce that many mutations is derived. The number of generations can then be turned into years back to the MRCA of the two descendant haplotypes by multiplying by the average number of years per generation. Various numbers for this have been used (25 years/generation is conventional in general parlance), but a great deal of empirical evidence points instead to about 34 years, as my paper "How Long is a Human Generation" argues.

Principle 1: The FTDNA 37-marker test is the one indispensable test for those who want to know for sure to which surname patrilineage they belong; this test is also sufficient to all but rule out the possibility that an NPE (Non-Paternity Event) has occurred in their patriline.

The FTDNA 37-Marker test is both necessary and sufficient for determining a male's patrilineage, and for the majority of testees this is the sole benefit of ySTR testing. How much of a benefit it is depends on the genealogical interactions prompted by the test results, whether these stem from contacts with FTDNA-reported matches, or from the surname or patrilineage projects organized by volunteers around the results. It should be noted though, that even when a person comes up as unmatched, implying that his is a unique one-person patrilineage, there can be great genealogical benefit to knowing that he does ***not*** belong to certain prominent patrilineages of his surname.

A secondary benefit of determining one's patrilineage is to ascertain whether an NPE (Non Paternity Event) has occurred in one's patriline. And as often as not, only an FTDNA 37-marker ySTR test is capable of doing this, because NPEs so frequently leave no paper trail. NPEs can occur in any generation past, and they typically do occur at rates that vary from about 2-7% or more (until modern times, at least, when rates have soared in many areas). What's more, the chances that an NPE has occurred somewhere up one's ancestral chain are cumulative and amount to between 20-50% for those who can trace their ancestries back 12 or more generations. It's also the case that if one is able to push one's ancestral line back all the way to the period of surname adoption (the 14th century in England (minus Wales), though centuries later in the rest of Great Britain) NPEs are likely to be the rule, not the exception, because the first man to adopt, say, the surname Smith, because he happened to be a blacksmith, is likely to have siblings or cousins who adopted other surnames.

Principle 2: GD comparisons (e.g. 35/37, 105/111) between two haplotypes (and TMRCA estimates based on them) do not provide genealogically useful estimates of closeness of relationship; they are therefore of little or no value in identifying the Closer Cousin Clusters (CCCs), which are genealogically meaningful.

This proposition runs counter to what most people believe, and to what FTDNA encourages people to believe. The company has gone to the length of creating a TIP calculator for calculating the TMRCA of pairs of haplotypes (yours and that of one of your matches), expressed as a range of probabilities that your common MRCA comes in at 1, 2, 3, ... N generations back. The company also makes a point of sorting your matches into closeness of match categories based on GD correspondences of (for the 37-marker test) 32/37 through 37/37. Yet neither of these approaches yield results accurate or reliable enough to be genealogically useful in identifying sub-branches within a patrilineage. And at the same time, the company fails to provide guidance on the analytical techniques that ***can*** be used to determine which tested members of a patrilineage are more closely related to the other members of what I call their CCC (Closer Cousin Cluster) than they are to all the other members of the patrilineage. I will outline those techniques under Principles 3 and 4 below.

There are many reasons why TMRCA and GD measures don't yield genealogically useful results.

In the first place, the number of marker value divergences between two haplotypes isn't necessarily the same as the number of mutations that have occurred since their ancestral lines diverged.

There are a number of different kinds of mutations, some of which (multistep mutations) can mimic multiple single mutations, and others (back mutations) that reverse previous mutations back to their original state, thus wiping out all traces that a mutation has occurred.

Then there are the mutations to the multicopy markers, especially the highly mutable DYS464, in which several marker values appear to be changed by a single mutation. There's also a type of

mutation called a reclOH event, that can change several multicopy marker values at once by deleting certain of their values, then effecting a repair of sorts by filling in the missing values from an alternate copy that runs backwards on the same palindromic loop; never mind the technicalities here: what's important is that these reclOH events, which are quite common (though we have only a vague notion of their frequency) constitute a kind of mutation that can change several marker values at once. Deletions can also occur at other places in the copying of the Y-chromosome, leaving holes (missing marker values).

Determining when these special kinds of mutations have occurred can be quite problematic, and deciding how to account for them—whether to add 0 or 1 or 2 or whatever to the total number of mutations for the haplotype—can fuzz those nice discrete GD comparison numbers and further bloat the already wide range of possibilities when it comes to estimating TMRCA.

In the second place, markers mutate at wildly different rates (some more than 50 time more frequently than others), and besides that, current estimates of these individual marker mutation rates are rather crude, and may be off by as much as 15% or more either way.[1]

However, these problems in counting mutations, and in applying appropriate mutation rates to the calculation of TMRCA, pale in comparison to this fatal flaw in making pairwise haplotype comparison: **mutations are so rare, and their occurrences so sporadic, that pairwise comparisons even of 111-marker haplotypes over many generations of genetic divergence between two ancestral lineages constitute too small a sample of the mutational process to provide meaningful estimates of closeness of relationship between the descendant haplotypes**.

For example, even a pair of 111-marker haplotypes that diverged from a MRCA who lived as many as 10 generations ago, provides only 111 x 9 x 2 = 1998 opportunities for mutation, and if one were to make, say, ten such pairwise comparisons, one could expect to find a range of 3-8 mutations separating any given pair, with lesser probabilities of even greater variation. Translating this 3-8 range (over 111 markers and 10 generations) back into TMRCA estimates expressed in years would yield a range of birth dates for one of these pairwise MRCAs between 1525 and 1800, which makes the estimate genealogically quite useless. And taking such TMRCA estimates seriously is no less misguided than supposing that a 108/111 match is necessarily "closer" than a 103/111 match; these and all the numbers in between are all within the range of normal and expected variability across a set of paired haplotypes.

Let's look at a few real-world examples of the meaninglessness of pairwise haplotype comparisons.

In my ALLEN Patrilineage 1 project, the haplotypes of about 20 members have been extended to 111 markers. On the genealogical evidence, which is very extensive and deep, all of these lineages converge backwards to a handful of closely related patriarchs who were born in or immigrated to Virginia between 1680-1710, and probably ultimately to a single immigrant ancestor born between 1620-1650. Although many members of this group remain unconnected to any of the deep ancestral lines, we do know that none of the members of this project shown in the GD matrix converge with any of the others within the last 7 generations, and most converge at about 9 generations on the

---

[1] It has been claimed that FTDNA's TIP calculator does take into consideration these varying mutation rates from marker to marker, but I also know that, at least in the past, the average mutation rate used by FTDNA for their 37-marker panel was considerably at odds with the rates worked out more accurately by DNA maven John Chandler. Although I myself use these so-called Chandler rates in my own calculations, and even cruder estimated rates for the markers beyond the first 37, it's likely that many of these individual marker mutation rates are wide of the mark, as the underlying mutational frequency data is still accumulating. I do have a fair amount of confidence, however, in the overall average rates of whole marker panels (37, 67, or 111), because there are enough markers in them that probably most of the mis-estimates of individual marker rates cancel out.

average. This large number of highly diverse 111-marker haplotypes has made it possible to calculate a TMRCA for the MRCA of the whole extended group that is probably accurate to within a generation, or at most two, either way. In fact the predicted birth date of the father of them all based purely on the DNA, which is in the range 1610-1630, dovetails almost perfectly with what the genealogical evidence suggests.

But now look at the range of pairwise GDs in the boxes of the 111-marker GD matrix for ALLEN Patrilineage 1, comprising 23 different highly divergent haplotypes all descended from the same immigrant ancestor, and where the actual differences in closeness of relationship vary only from 8-11 generations. If mutations were as regular as a metronome, we would expect no more than a single GD's worth of variance for these 111-marker haplotypes Yet their pairwise GDs range from 0 (identical haplotypes) through 12, where by FTDNA's criteria any GDs >9 aren't even "possibly related", and on there's only a moderate chance that a pair that is are say 110/111 (GD=1) are any more closely related than a pair that's 103/111 (GD=8).

Two of my other projects illustrate the extremely sporadic nature of yDNA mutations[2]. The haplotype chart for my ALLEN Patrilineage 2 Project shows that the haplotypes for two members (A-07 and A-24, surrounding the RPH in the center of the chart) have no mutations, yet their ancestries diverge by at least 7 generations, and probably for several more. On the other hand, one of the members of my RATHBONE Patrilineage 1 Project has tested himself, his father, and his brother through 67 markers: as expected, the haplotypes of father and brother are identical, yet the subject himself has experienced two mutations in just one generation, and the probability calculations show that there is a 2% chance of that happening.

In my opinion, FTDNA's reporting of pairwise GDs is extremely misleading, leading customers to suppose that they can measure closeness of relationship by the bare number of mutational differences between haplotypes taken two at a time. And the company used to lean even more heavily on these pairwise comparisons by providing a calculator that they labeled TIP that presumed to calculate the probability distribution that convergence of the ancestries of the pair to a common ancestor went back X generations. At some point, because in many cases a highly probable prediction of convergence after 2 or 3 generations (because GD=0) was belied by the fact that a pair were known from genealogical research to have no convergence for at least 7 or 8 generation, a parameter was added that allowed the customer to specify that genealogical constraint. But since most genealogists are able to project their ancestry back at least 5-6 generations this parameter swamped the calculated probabilities, and made nonsense of them, and as far as I know the TIP calculator was withdrawn (at least I can no longer find it): if so, I congratulate FTDNA on their evolved, if belated, understanding.

Although individual pairwise GD comparisons are of little or no genealogical value, constructing a matrix of ***all*** the pairwise comparisons across a large patrilineage, grouped into sub-clusters of claser (green) GDs, can be suggestive of the actual CCCs indicated by analysis of the mutational patterns (which I will explain next) and/or by genealogical knowledge. The 111-marker GD matrix of the ALLEN Patrilineage 2 project provides an example of this: the last two haplotypes listed appear, just from the chart, to constitute a small CCC independent of the remaining members of the project, and in fact, this is borne out both by their genealogies (which suggest that they descended from a different 17th century immigrant ancestor, as well as by those mutational patterns. Large GD matrices, like the one earlier mentioned for the ALLEN Patrilineage 1 project can also support quite accurate TMRCA estimates for the single MRCA of all of the project members.

---

[2] Although I'm focusing here on ySTR mutations, the other type, ySNP mutations that are inventoried by FTDNA's BigY test, are if anything even more sporadic.

Principle 3: Analysis of mutational patterns across a patrilineage can identify CCCs, but only by means of shared, because inherited, mutations across uncommonly large sets of tested descendants who are quite distantly related to each other.

The ALLEN Patrilineage 1 project cited above is the kind of project for which CCCs can be identified, as it has the chief desiderata: many haplotypes of diverse lineages that have been extended to 111 markers, backed up by solid and deep genealogical knowledge. In fact this project has 19 such haplotypes extended to 111 and is the largest that I've worked with or studied to date. How many is the "many" 111-marker haplotypes that are needed to identify CCCs? I'd say at least 5 or 6, and for 37 markers, 7-10—and this is presuming that all of the haplotypes in question are known to diverge from each other over at least 6 generations. With fewer than these numbers it becomes problematic to determine the composition of the RPH, and therefore to identify which values are mutations, and which the normals, and it is upon the mutational pattern that inferences to CCCs are predicated.

It's not enough that two or more haplotypes share the same mutation. The markers that are most likely to mutate, and which therefore account for most of the mutations, can, by the same token, mutated more than once within the same patrilineage, particularly when the lineage is deep and diverse, and where this happens two haplotypes with the same mutation may have come by their shared mutation independently, instead of inheriting it from a common ancestor. To assess the probability of this, the specific mutational probability of the marker in question needs to be taken into account, and also the overall pattern of mutations across the patrilineage. The reader will find a table that illustrates the mutational probabilities of each of the first 37 markers mutating more than once across varying number of generations, in the appendix to this paper, as well as the formula for calculating those probabilities for other markers whose mutational probabilities are given.

The additional requirement, the need to be sure that the descendant haplotypes have no common patrilineal ancestor within at least the last six or so generations implies a fourth principle of ySTR DNA interpretation:

Principle 4: The identification of and genealogical exploitation of Closer Cousin Clusters depends more on genealogical research than on ySTR DNA testing.

Beyond the sorting of haplotypes into patrilineage (which is conclusive by itself according to Principle 1), ySTR DNA testing can at most provide only ***guidance*** toward the most fruitful avenues of genealogical research: it cannot substitute for that research. When it comes to determining CCCs, DNA-based theories must generally yield to the genealogical evidence when there are conflicts, though in extreme cases, where the DNA probabilities point the other way with a 95-98% confidence level, the genealogical evidence and the conclusions drawn therefrom need to be scrutinized very critically.

Most of the time, however, the mutational patterns in the DNA that appear to demarcate CCCs confirm, or are at least consistent with what the genealogical evidence suggests, and the two modalities of evidence mutually strengthen each other.

To return to the methodology for analyzing the mutational patterns that characterize CCCs, in light of Principle 4 the set of (extended) haplotypes used as the basis for the analysis needs to be pruned of all known closer cousins (or patrilineal relatives) from recent generations, both to achieve the six or so generations of known separation between cousins, and also (as a preliminary) in deriving the RPH that serves as the standard for determining which marker values are the mutated ones, and which the normal values. In general, all but one of these recent clusters of known closer cousins need to be weeded out, leaving that one to represent the rest of his recent lineage. A corollary of this rule is that when it comes to ySTR testing, there is little or no benefit to testing known close patrilineal relatives—ones' father, brothers, or close cousins—unless one simply wants to rule out the possibility

that one of them may have experienced an NPE in his line, and for that purpose, the lowest resolution ySTR test that FTDNA offers should suffice, and that would seem to be the 37-marker test (too bad, since FTNDA used to offer a much cheaper 12- or 25-marker test).

Principle 5: Ultimately, the value of yDNA testing for genealogical purposes depends on the depth and quality of evidence that your patrilineage cousins are able and willing to share.

The first requirement here is that the tests that have been run to produce at least a 37-marker ySTR haplotype which turns up as a match to a particular patrilineage, have an active and knowledgeable genealogist behind it, whether that be the patrilineal descendant male who actually tested, or his sponsor or genealogical representative.

The second requirement is that this genealogist remains prepared to respond to inquiries at the email address he has supplied to FTDNA, or to the surname or patrilineage projects s/he may have joined that have been organized around FTDNA ySTR test results.

Unfortunately, all too many FTDNA customers order tests without understanding their limitations, or what they need to bring to the table to take advantage of them; many are looking for a genealogical shortcut or a magic talisman that will somehow get them past their brick wall. However, unless they are fortunate enough to be able to hook up with a robust and genealogically activist patrilineage project, the most that they can hope for is that one of their matches will bth respond to their email query; and be more genealogically knowledgeable about their particular line than they are. Alas, all too often, the few testees who have such knowledge have over time derived so little benefit themselves from their tests (precisely because they know more than their matches), that they don't bother to respond, or even keep their email address up-to-date.

Another reason FTDNA-reported matches often fail to respond is that in the common case where the interested genealogist has tested, not themselves, but a surrogate male who bears the surname that they're interested in, the one email address that FTDNA presents to matches is that of the surrogate, and not the genealogist—and many if not most surrogates, not having much interest in their own genealogy, fail either to respond or to pass on the contact to their genealogical representative.

This can come about because when FTDNA initially solicits these email addresses from the customer, it merely asks him/her to list one or more email addresses in any order, first in a field associated with the customer herself and with her physical address, and then in a second blank associated with the surrogate and his physical address, without explaining which of these addresses will be used as the "primary" email address for purposes of genealogical contact, or even making it quite clear which of the two physical addresses is to be used in shipping the test kit. I suppose that the primary email address would the first listed in the customer blank, but since this first blank is associated with a physical mailing address, which the customer may be forgiven for assuming that this first physical address is the one for the test kit, s/he may also assume that the first email address listed should correspond to that physical address, and also be that of the surrogate. All of this ambiguity could be avoid by clarifying the entry form, and dedicating a field to "Email address to be contacted at by your test result matches?" And while they are at it, it would be also be desirable to expand this contact information to include more than one email address, lest one become obsolete over time.

The most important factor regarding matches, though, is simply the quality and depth of their genealogical research. One advantage of independent projects organized by patrilineage rather than surname is that instead of just providing a few pairwise matches, the contact information for all of the principal genealogists interested in the line can be provided in a project directory that is also available to other researchers interested in the patrilineage who are as yet untested, but have found the web page through browsing.

Unfortunately, some years ago, through an overzealous interpretation of the European GDPR privacy guidelines, FTDNA banned the administrators of their surname projects from publishing on their project pages all the information needed to support fruitful yDNA-based genealogical projects. Thus, pretty much the only such projects that it is genealogically worthwhile to join, or organize, are the kinds of patrilineage projects that I've organized and administer entirely independent of FTDNA, and of course with the consent and desire of the members to share both their research and their yDNA classifications with other researchers interested in their particular surname patrilineage ancestries.

The CCC Analytical Method, applied to a particular patrilineage

Let's look at another example of a patrilineage for which CCC Analysis has proved fruitful—the DENNISON Patrilineage 1 Project. In this large patrilineage of 17 members, the haplotypes of 11 have been extended to 111 markers (here's the project 111-marker GD matrix). These lineages have been very thoroughly researched, and most members can trace back to their immigrant ancestor. Thus, we can also say of this set of 11 extended haplotypes that we know that their lines don't converge for the last six generations, and that the average distance back to the MRCA is probably about 9 generations.

The general idea in CCC analysis, however, is to study, not the GD matrix, but the mutational pattern shown by the project haplotype chart (or table) . In this table of marker value results for each haplotype, mutated marker values are those identified by the colored squares, and the different colors represent different types of mutations—as defined by the color-coding glossary in the text at the bottom of the table. It will be noted that I've grouped the haplotypes together into clusters in the light both of the mutational patterns suggested by the DNA data itself, and also from my knowledge and suppositions regarding the actual ancestral relationships among these subjects.

Definitive identification of CCCs depends on finding distinctive mutations that all the members of a particular cluster share, which they inherited from the MRCA of their cluster.

But this formulation begs several questions.

First, there is the problem discussed briefly above: how do we determine which marker values are the mutated ones, and which the normal — the values that comprise the RPH we wish to construct as our best guess for the composition of the haplotype of the MRCA ancestor of all members of the patrilineage?

As noted above, in most cases, one can simply use the preponderant value across all the haplotypes, but where the preponderance is slight (with one value having only a simple majority over the other, or where the subsets of values are of equal size), and/or where there are more than two marker values, other principles of analysis must be brought into play. Where there are, say, three values, the most likely original MRCA value is the middle one, since the mutational process is generally suppose to run either way, in roughly equal proportions, either adding or subtracting a number to the original marker value. However, certain directional biases seem to be emerging from the so far minimal mutational data favoring either gaining or losing a "repeat" (in the technical terminology), so we may need to keep an eye on this possibility going forward.

Where the balance is fine between two marker values, I take into account what I know of the genealogy, and of the likely genealogical clusterings and weightings within the particular set of haplotypes involved. Thus, if it appears that 6 of 111 haplotypes have marker value A, and 5 value B, but I know or believe that most of those who have value A look like the makings of a Closer Cousin Cluster either because their haplotypes are similar otherwise, or because their ancestries seem to converge short of the MRCA of all, while the smaller set of 5 haplotypes shows greater diversity

either of DNA or genealogy or both, I may chose the minority value B as the normal value, and use it in constructing the RPH. As noted above, drawing on the genealogical as well as the DNA evidence introduces an element of circularity if our aim was to reason out relationship strictly from the DNA evidence, but in view of Principle 4, our DNA-analytical goal here must be the more modest one of just looking to the DNA evidence for suggestive, or confirmatory patterns to support whatever we can learn from genealogical research.

It's not always going to be possible to pick the correct marker value as the mutated one, and it may be necessary to change one's initial picks as more, and particularly more diverse, haplotype accrue to the project. The reason for this inherent uncertainty about which values are the mutated ones, is analogous to uncertainty bedeviling pairwise TMRCA: in both cases, the sample size is too small.

The first problem one encounters in CCC analysis, correctly ascertaining the RPH, and thus identifying the mutated values, can be expected to evanesce as projects grow in size. However, there's a second type of problem may actually increase as more haplotypes come into play.

This biggest problem with using shared mutations to identify CCCs is that sometimes two haplotypes may share a mutation, not because they inherited it from a common ancestor, but because the mutation occurred independently in two or more sub-branches of the patrilineage that aren't closely related to each other.

Unfortunately, there is no definitive way to determine whether a shared mutation was inherited rather than generated independently a second or third time within the same set of haplotype. However, we can look to the widely varying mutabilities of the individual markers to assess the corresponding probability that a particular marker may have mutated more than once in the same patrilineage.

The Mutational Quirks ofCertain Markers

As it happens, the very most mutable markers in the FTDNA panels (DYS449, 464, the CDYs, and 710 and 712 in the 111-marker panel) are quite likely to have mutated independently more than once across a fairly large set of haplotypes (say 7-10), and there are many other fast-mutating markers that are suspect.

I roughly indicate the relative mutability of individual markers in the columnar marker headings of my haplotype charts— color coding the faster mutators red (the fastest are the deepest reds); the average mutators with the background text color (a dark reddish brown); while the notably slow mutators are color-coded blue. Naturally, I use the more exact (but still only approximately known) numeric probabilities for each marker in my probability calculations.

Thanks to the ambiguous possibilities here, **where shared mutations may or may not be inherited, reasonably reliable identification of a CCC from the DNA evidence alone typically depends on finding at least one a shared average or slow mutator, or more than one fast mutating marker**. Of course, where there is also strong genealogical evidence identifying a CCC, lesser DNA evidence may be used to reinforce that identification.

Finally, in mutational assessment, a special word is called for with respect to two sets of markers. The pair of CDY markers in the 1-37 panel mutate so frequently that they are best ignored in these analyses, except, perhaps as confirmatory grace notes. The probability that one of these will mutate in a single father-to-son generational transmission is about .03531, thus the probability of it mutating more than once in a single haplotype over, say, 10 generations is 7%. And if one is analyzing a set of

10 haplotypes, one can positively expect more than one such mutation to each CDY, and once in a while a second mutation in the same haplotype that's actual a back mutation to it's original value.[3]

When the DYS464 multicopy marker, which typically comprises four component values (and sometimes several more) mutates it often changes more than one of its component values at once, and "calling" the values of this marker even has an element of subjective interpretation on the test bench. Without getting into the technicalities of why this is so, DYS464 is best treated as a single highly mutable marker, with a per generation mutation probability of .02264—not far behind the CCCs (the next most mutable marker is DYS710, at .01570). Offsetting this, the way DYS464 is normally evaluated, back mutations rarely occur.

Precisely because DYS464 comprises 4+ marker values and is so mutable (two thirds as mutable as the CDYs) we can't afford to ignore it in our haplotype cluster analysis. And FTDNA offers an optional test called DYS464X that eliminates the ambiguities in the evaluation of DYS464 and more often than not reduces it to 2 marker values that are just as solid as any of the single-copy markers. I've therefore generally recommended, whenever a patrilineage has turned up several different DYS464 values according to the default test methodology, that at least certain of the project members order the additional DYS464X test, so that we can figure out what is really going on with this important marker. Unfortunately, FTDNA seems to have decided to bundle this often valuable test with re-evaluations of other multi-copy markers that aren't of such value, with a higher price on the bundle, so I expect to use more circumspection in recommending this test to my project members in the future.

With the above principles and caveats in mind, you may wish to read through my detailed CCC analysis of DENNISON Patrilineage 1 haplotypes as an extended example of what can be done with a large set of extended haplotypes. Since the analysis is specific to the mutational pattern across these 11 extended haplotypes, it would be well while reading this material to bring up the corresponding project haplotype table in a separate window for reference. The table scrolls to the right, allowing one to see which haplotypes have extended to 111, and showing the mutations that have occurred in that upper band.

[3] From the point of view of analysis, the instability of these CDYs isn't as bad as it seems: because a marker can mutate up or down the chances that the same CDY is a different haplotype will mutate in the same direction as a previous one is just half this ultrafast mutation rate; on the other hand, though, a mutated CDY marker has an equal, appreciable, chance of mutating back to its original value, leaving no trace that it ever mutated at all.