Peer Reviewer List available upon request

March 2019

Y-DNA Phylogeny Reconstruction using likelihood-weighted phenetic and cladistic data – the SAPP Program

J. David Vance

Affiliations: ISOGG. Contact: davevance01@gmail.com

Abstract

Modern genetic genealogy conventional approaches to reconstructing the phylogeny of agnatic (male-line) ancestors for a group of Y-DNA-tested men have traditionally used either Y-STR or Y-SNP data only. This creates an occasional dilemma over which analysis - Y-STR or Y-SNP - more accurately reflects the phylogenetic tree of the group; an unnecessary dilemma since both sets of data are products of the same historical agnatic lines of descent and should therefore be complementary. Y-STRs and Y-SNPs also each have different strengths which can be used in concert to partially offset their separate weaknesses. An approach is presented that weighs phenetic and cladistic data characteristics from the available sources of data (Y-STR and Y-SNP) as well as from traditional genealogy information according to likelihood to reconstruct an agnatic phylogenetic tree which reaches 100% accuracy at maximum data availability while exploiting the strengths of each available data source. This approach has also been made publicly available as the free online software program *Still Another Phylogeny Program* (SAPP at http://www.jdvtools.com/SAPP).

1. Report

1.1. Introduction

The major value of commercial Y-DNA testing to the field of genealogy lies in the opportunity for the consumer, through aggregate data collected from one or more Y-DNA tests (collectively here called their "kit", although it may include test results from several companies), to match other tested men and gain more insight into their shared agnatic (male-line) ancestry. Discovering matches is therefore a key objective in the pursuit of genetic genealogy, and as affordable testing has improved and databases of matches have grown larger this objective has moved from a focus on "**Who** do I match?" to "**How** are we related?".

This second question has driven many approaches to reconstructing the phylogenetic tree of agnatic ancestors for a group of kits representing tested men, although until now most approaches have used only one type of available data from the Y chromosome – usually either Short Tandem Repeats (Y-STRs) or Single Nucleotide Polymorphisms (Y-SNPs), often paired with knowledge from traditional genealogy research. Conventionally, kits have been grouped into predicted or confirmed haplogroups by differing manual approaches and then further sorted within those haplogroups by genetic distance based on Y-STR marker allele differences. That sorting is often then further improved through more sophisticated analyses like Y-STR signature (motif) matching. At its most sophisticated, manual Y-STR mutation history trees can be created to map at least partial agnatic phylogenies for a group of men.

In parallel, Y-SNP haplotrees have also become a common structure for representing a group's phylogenetic tree especially as Y-SNP testing has gained in affordability and popularity.

At the current state of Y-DNA testing, any smaller haplogroup is typically formed of many kits at varying levels of Y-STR testing (often at Y12, Y37, Y67, or Y111 levels, though in some cases up to 561 Y-STRs), and Y-SNP testing - which even at its most extensive has typically uncovered a branching Y-SNP no more frequently than every 3-4 generations. In such cases one set of data may help determine branching in one subset of the phylogeny while another set of data carries more information about a different subset.

This also creates an occasional dilemma about whether Y-STR or Y-SNP analysis more closely reflects the actual phylogenetic tree of the group; a dilemma which is unnecessary since the approaches are complementary and can be combined along with further insights from the group's traditional genealogy research to recreate as much of the full likely

phylogenetic tree as possible based on all available sources of data.

Groups of men used in phylogenetic analysis may have been grouped for varying reasons; as examples, they may have self-selected themselves into one group (for instance by joining a surname or haplogroup project); they may have been indicated by a match list provided by a commercial company; or they may represent a subgroup created by a project administrator based on traditional genealogy information (e.g. common ancestor) or any number of predicted or actual Y-STR or Y-SNP criteria including Y-SNP test results or Y-STR genetic distance or allele marker similarities. This variety of origins means there can be no assumptions in a phylogenetic analysis about the amount of available data, variations or patterns in the available data, or even about the consistency of testing across the group. Some men may have tested anywhere from 12 to 561 Y-STR markers, or some men may only have a predicted Y-SNP haplogroup while others have done extensive Y-SNP discovery testing. There may also be an abundance or complete lack of traditional genealogy information linking the group. The only consistency that can typically be assumed (beyond the assumption that the Y-DNA testing results and traditional genealogy research themselves are accurate) is that more testing of any kind will provide further data to improve the phylogenetic analysis.

Traditional phylogenetic approaches to recreating the agnatic phylogenetic tree (e.g. maximum likelihood, parsimony, Bayesian, etc) would require that a consistent set of common data be available across the entire group for consistent analysis. Additionally, the accuracy of these methods is highly dependent on having statistically-significant volumes of data.

By contrast, the following properties are common in groups of men most frequently considered for Y-DNA phylogenetic analysis in genetic genealogy:

1. They are commonly known or believed to be related within at most a few thousand years, making them often smaller groups (generally fewer than 100 men up to perhaps 200, though sometimes larger) and therefore less suited to confidence through statistical analysis;

2. The differing levels of Y-DNA testing among group members significantly limits the availability of consistent data down to the lowest common denominator of the group's level of testing. This means either limiting phylogenetic analysis to the lowest amount of available data and ignoring other relevant data, or widening the analysis to more complex assessments than just the parameters which are available for all kits;

3. The differing levels of Y-DNA testing within the group also make certain sub-phylogenies more clearly-defined than others. It is not unusual for example for a family sub-group of the overall group to have a well-defined phylogeny for their own sub-group based on known relationships between the individuals or based on extensive Y-SNP testing within that family, while the rest of the group has fewer well-defined relationships.

4. At its most inclusive, however, the available data across all sources (Y-STR, Y-SNP, and traditional genealogy) is usually still insufficient to recreate a single unique and accurate phylogenetic tree for the entire group. There will generally be a finite set of equally-possible phylogenetic trees which meet likelihood criteria and which cannot be further distinguished by the available data.

5. The most accurate phylogenetic tree is neither the most parsimonious nor the most statistically-likely, it is instead the phylogenetic tree which most faithfully reflects the historical agnatic lines of descent for the group. None of the typical phylogenetic approaches are more suited than others to approach accuracy as defined this way especially in smaller groups and without considering all available sources of data.

6. Finally, the odds of real or apparent incorrect or inconsistent data is non-zero. Y-STR data is perhaps the most prone to this property given the common incidence of homoplasy (convergence); however, the positive or negative status of any Y-SNP may be misread or not reliably reported, and traditional genealogy research is regularly wrong as well. Any incorrect or inconsistent data can, of course, either misdirect the phylogenetic analysis or be internally inconsistent and therefore reduce the set of most-likely phylogenetic trees to the null set.

The goal of optimizing the set of most-likely phylogenetic trees to the minimum set possible therefore requires an approach which maximizes the available data from all sources, allows for differing criteria within sub-phylogenies, limits or eliminates incorrect or inconsistent data, and is not assumed to be restricted to the traditionally-optimal phylogenetic approaches.

1.2. Methods

A note: the genetic and biological underpinnings of data used in genetic genealogy are not generally discussed here unless specifically relevant to the purpose of modeling a group's agnatic male ancestry.

1.2.1. Characteristics of Available Data and Associated Likelihoods

1.2.1.1. Y-SNPs

The primary characteristic of Y-SNPs useful to modeling a phylogenetic tree is cladistic: Y-SNPs uniquely define a clade of agnatic descent. The group of men who share a Y-SNP mutation all share a more recent common ancestor with each other than they do with men who do not share that Y-SNP mutation.

There is much debate within genetic genealogy about how frequently the Y-SNP data reported by commercial testing meet this expected condition, since reported test data can typically include non-unique mutations, results from different levels of read technology and coverage over difficult-to-read and recombinant areas of the Y-chromosome. Therefore Y-SNP data must be filtered before analysis to identify the proper data subset for phylogenetic purposes. This filtering is currently inconsistently supported by commercial companies and generally requires manual intervention.

Equivalent Y-SNPs (two or more Y-SNPs within a phylogenetic block on the Y-SNP haplotree which have no discovered branching between them), *synonym* Y-SNPs (two Y-SNP labels for the same genetic mutation) and *recurrent* Y-SNPs (where the same physical mutation has occurred independently in two different clades of men) further complicate the cladistic information available from Y-SNPs. Recurrent Y-SNPs in particular invalidate this cladistic property and are not handled by our approach unless they are differentiated in the input data (for example, using the outmoded SNP1.1, SNP1.2 labeling convention). Equivalent and synonym Y-SNPs merely complicate but do not invalidate the Y-SNP cladistic property and are not currently included in our approach but could be added as a further enhancement.

If the purpose is to model the phylogenetic tree of the group under analysis, then a further limited set of Y-SNP data is important to that purpose. Y-SNPs older than the common ancestor of the group (and therefore positive for the entire group) and Y-SNPs that are private to a single individual in the group are generally not useful for modeling the group's phylogenetic tree. **Therefore, the available data need only include Y-SNPs for which at least two members of the group are known to be positive, and for which at least one other member of the group is known to be negative.** These Y-SNPs will define sub-clades within the branching of the group's phylogenetic tree. We note that this particular filtering is not a requirement for accuracy of our approach but only for efficiency as it serves just to limit the data considered to the set containing useful phylogenetic information.

Given the wide variations in Y-SNP coverage among men who have taken some form of Y-DNA test, however, the available useful data cannot be assumed to include the positive or negative status for every kit for every Y-SNP. For each kit, a Y-SNP's status may therefore be positive, negative, or unknown, and all three conditions are handled in our approach.

It is important for phylogenetic analysis to note that Y-SNPs provide both inclusionary and exclusionary cladistic information. For example, if one kit is positive for a Y-SNP, and another kit is positive for a different Y-SNP on a different branch of the Y-SNP haplotree, then those two kits cannot share a common ancestor any later than the Y-SNP(s) at the connection point of those two branches on the Y-SNP haplotree. This means the first kit will by definition share as a closer match any third kit which is positive for any Y-SNP further down on his own branch after that connection point.

At this time the likelihood of inaccurate or inconsistent Y-SNP information has not been extensively studied and so has not been factored into our approach. For analysis purposes therefore, the filtering described above must be done beforehand for Y-SNP data. Given that prior assumption, all provided Y-SNP data is treated as 100% accurate, with the acknowledgement that if inaccurate information is included it may limit the set of resulting phylogenetic trees down possibly to the null set if there is no solution which meets all apparent criteria.

1.2.1.2. Traditional Genealogy Information

Again, if our narrow purpose is to recreate the most likely agnatic phylogenetic tree among a group of men, then the necessary information from traditional genealogy research is again cladistic: which kits within the group are descended from more recent common ancestors?

This limited set of useful information from traditional genealogy has properties very similar to Y-SNPs. Identifying ancestors older than the common ancestor of the entire group is not very useful to phylogenetic analysis. Identifying ancestors unique to individual kits is also not very useful. **What IS useful is**

identifying common ancestors shared by at least two of the men in the group and NOT shared by at least one of the other men in the group. Like Y-SNPs, these ancestors will define sub-clades within the branching of the group's phylogenetic tree. This criterion again is an efficiency measure to limit the data to that which provides useful phylogenetic information and has no effect on the accuracy of our approach.

Given the family-oriented nature of traditional genealogy, it is perhaps more likely than with Y-SNPs that all of the descendants should be known for a particular common ancestor. However, we do need to allow for the same three status conditions: for each man in the group, their relationship to a specific common ancestor may be either positive (i.e. a descendant), negative (i.e. NOT a descendant), or unknown.

Also just as for Y-SNPs, there is no published data available on which to assess the likelihood of the accuracy of traditional genealogy information. Therefore we again have an assumption that the data provided is accurate and by extension that all positive and negative common ancestor information is 100% accurate. Just as for Y-SNPs, if this assumption is incorrect there may again be no solution which meets all apparent criteria.

1.2.1.3. Y-STRs

The established volume and affordability of Y-STR testing makes it currently the most consistently-available source of information for re-creating phylogenetic trees. Many groups may still have no phylogenetically-useful level of Y-SNP or traditional genealogy data collected, and in such cases the basis for agnatic ancestry reconstruction is limited to available Y-STR data alone.

Our approach therefore addresses Y-STR analysis in detail, and specifically makes use of four distinct characteristics of Y-STR data.

1.2.1.3.1. Y-STRs: Genetic Distance

Traditionally, *genetic distance*, an estimate of the number of mutational differences which separate two haplotypes (sets of Y-STR allele values), has been used by commercial companies to gauge the degree of relationship between any two tested individuals. While this may provide a rough assessment, the influence of both homoplasy (a.k.a. "convergence") and statistical variation in Y-STR mutation rates makes genetic distance inadequate as a stand-alone basis for phylogenetic analysis.

Genetic distance is considered in our approach as a general guide only and "last-ditch" prioritization for branching decisions that are still unclear after all other methods discussed here have been exhausted.

While a full discussion of the calculations of genetic distance are beyond the scope of this paper, the method used in our approach matches closely with commercial company calculations and takes into account the "step-wise" mutational model for most Y-STRs, the "infinite-alleles" mutational model for multi-copy Y-STRs and certain other special conditions (such as null values) as suggested by the STRBase reports of the NIST (National Institute of Standards and Technology). Microalleles are not handled in this approach.

To bridge across Y-STR testing levels, genetic distance between any two men is calculated as the ratio of their Y-STR mutational differences divided by the number of Y-STR allele values they share, or

$$GD_{ratio}(k1,k2) = \frac{\sum_{n=1}^{\min(N1,N2)} D^n}{\min(N1,N2)} \qquad (1)$$

where $GD_{ratio}(k1, k2)$ is the genetic distance between kits *k1* and *k2* expressed as a ratio, *N1* and *N2* are the number of Y-STRs tested for each kit, and D^n is the traditional mutational difference between each pair of Y-STR allele values between the two kits.

1.2.1.3.2. Y-STRs: Signature Matching

A more detailed assessment of relatedness requires comparison of the individual mutation differences between kits to find phenetic evidence suggesting common descent. This is often called *signature matching* or *motif matching*. As generations descend from a common ancestor and mutational variations in Y-STR allele values accumulate, these variations are passed on to descendant lines and, if not affected by homoplasy, create recognizable "signatures" which identify the sub-group who share a more recent common descent. In Figure 1 for example, the left-hand sub-group has formed a recognizable signature with STR A = x+1 and STR B = y-1.

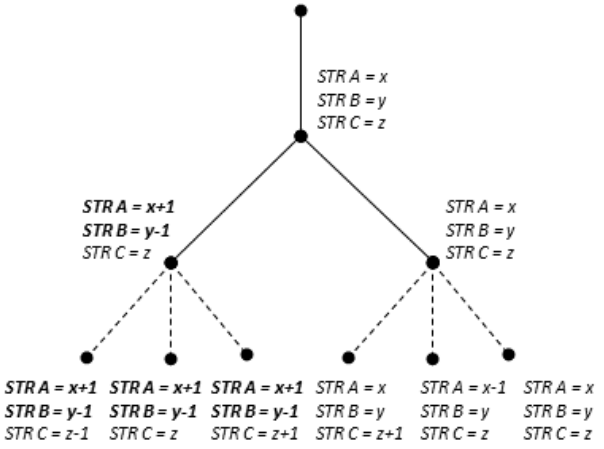


Figure 1. Y-STR Signature Matching

Signatures must be compared against a reference position in order to recognize that the signature mutations occurred *after* that reference point. For this reason, Y-STR allele values are compared against the starting Y-STR haplotype of the group's common ancestor as represented by either their known *ancestral haplotype* or, if the ancestral is not known, the calculated *modal haplotype* of the group. By this method, *off-modals* (Y-STR mutations which occurred within the branching of the phylogenetic tree) are identified and used to establish potential signatures. Note that exact adherence to the ancestral (or modal) haplotype is considered a special-case and treated as a signature also.

Signatures consisting of higher numbers of Y-STRs with lower mutation frequencies tend to be more recognizable. In fact where signatures exist, their relative importance for phylogenetic analysis is mainly dependent on two factors:

1. The "rarity" of the signature, meaning that patterns formed by higher numbers of lower-frequency Y-STR mutations are more likely to have been passed down by common descent instead of occurring independently, and

2. How far further independent Y-STR mutations may have obscured the signature since it was formed. In Figure 1, for example, if any of the left-hand descendants had a back-mutation of STR A from x+1 back to x, the signature would be harder to recognize.

Since both factors are directly related to the mutation rates of the individual Y-STRs which make up the signature, the relative importance of a signature can be expressed as the likelihood of the signature as given by

$$\text{Signature likelihood} = \prod_{k:1}^{n} \mu^k \qquad (2)$$

where the signature is made up of n Y-STR allele values, and μ is the mutation rate of each Y-STR. The relative importance of the signature is then inversely proportional to the likelihood, since smaller likelihoods are rarer and therefore more important as signatures.

Signatures passed down by common descent which consist of a single fast-mutating Y-STR are in practice rarely recognizable from artificial patterns caused by independent mutations. For this reason signatures of one single faster-mutating Y-STR are rarely considered unless indicated by special considerations (for instance, if the time back to common ancestor was so short as to warrant it). Signatures of single more slowly-mutating Y-STRs may be considered at lower genetic distances.

1.2.1.3.3. Y-STRs: Addressing Homoplasy (Convergence)

The effects of homoplasy in Y-STR analysis for genetic genealogy are generally known as *convergence*, which is defined by ISOGG as "the process whereby two different genetic signatures (usually Y-STR-based haplotypes) have mutated over time to become identical or near identical resulting in an accidental or coincidental match." For purposes of Y-STR phylogenetic analysis, this definition can be expanded to include any case where the mutations of Y-STRs over time obscure the identification of a group which descends from a common ancestor. This can occur either from Y-STR haplotypes of non-members outside the group converging to look sufficiently like members of the group, or by the Y-STR haplotypes of members of the group diverging sufficiently from other members.

While convergence cannot be completely eliminated, it can be significantly mitigated through two factors outside our approach, and three other factors which we address.

The factors outside our approach are:

1. Increasing levels of Y-STR testing among the group is probably the most effective method of reducing convergence. Most men who have tested with commercial companies will have hundreds or thousands of matches at the Y12

level but dozens or fewer at the Y67 or Y111 levels.

2. Most groups selected for phylogenetic analysis have already undergone some manual selection and sorting which identifies them as a group at least possibly descended from a common ancestor. This will eliminate the most unlikely of false matches.

Included in our approach are:

1. Combining Y-STR phylogenetic analysis with any level of accurate Y-SNP or traditional genealogy common ancestor information will provide phylogenetic context for clarifying the most likely sequences of Y-STR mutations over time.

2. The identification of Y-STR signatures will also provide the same phylogenetic context and further clarification of the most likely sequences of Y-STR mutations.

3. Further branching decisions are then prioritized to reduce the number of mutations in Y-STRs with lower mutation rates. This ensures that the resulting Y-STR mutation history in the absence of other relevant information is at least statistically most likely.

It is a recognized deficiency of our approach that, in the absence of phylogenetically-relevant Y-SNP or traditional genealogy information, and in the further absence of identifiable Y-STR signatures among the members of the group, convergence cannot be addressed except through statistically-likely branching prioritization based on Y-STR mutations. This indicates that our approach should be less effective given Y-STR-only data, at smaller numbers of kits and over shorter time spans (note this is borne out by the Test Data as will be shown).

1.2.1.3.4. Y-STRs: Addressing Long Branch Attraction

Initial testing of the other Y-STR analysis methods demonstrated the effects of *long branch attraction* at closer genetic distances; a form of systematic error whereby distantly related lineages are incorrectly inferred to be closely related because of the similarity of the amount of change they have undergone rather than the similarity of the changes themselves. For a more detailed review of long branch attraction see this Wikipedia entry (link).

For example, if two men in the group are more distantly related to the rest of the group and through convergence happen to have a lower genetic distance between each other than they do with the rest of the group, they may be sorted together under the same branch in the absence of Y-STR signatures or other relevant information even though an analysis of their individual Y-STR mutations would not indicate any common line of descent.

Addressing long branch attraction for two kits at close genetic distance to each other while more distantly related to the rest of the group requires assessing the degree of relatedness of the full set of Y-STR mutations among group members and how likely they are to have evolved along a common path.

To approximate this assessment on whether two kits have evolved along a common branch separate from a third kit, we consider their Y-STR haplotypes as individual *n*-dimensional vectors where *n* is the number of Y-STR allele values they have in common (i.e. the minimum of their levels of Y-STR testing). Then for each kit, the vector angle θ between their "haplotype vector" and the modal haplotype for the entire group is given as

$$cos\theta = \frac{\langle v,m \rangle}{||v|| \cdot ||m||} \quad (3)$$

where v is the kit's haplotype vector, m is the modal haplotype vector, and $||v||$ and $||m||$are the lengths of the respective vectors.

Our approach also weights the dimensions of the vectors using Y-STR mutation rates before calculating vector angles, to increase the effects of mutations in less frequently-mutating Y-STRs and cause Y-STR haplotypes which share slower-mutating Y-STR mutations to have closer vector angles.

If then two kits have evolved along a common line of descent compared to a third kit, especially if their line included slower-mutating Y-STR mutations, they will then have closer vector angles in n-space to the modal haplotype with each other than to the vector angles of the other members of the group against the modal haplotype. This provides an assessment of which kits may have evolved along more related paths than other kits, which allows for a further distinguishing decision basis than simply genetic distance alone.

Testing shows that in practice this comparison is necessary in fewer than 5% of branching decisions even in cases where no other relevant information exists to determine branching and genetic distance is

the last resort to distinguish the relative closeness of kits. However, in such cases this vector angle comparison distinguishes between long branch attraction and actual close descent of kits who are themselves more distantly related to the rest of the group.

1.2.2. Building the Agnatic Phylogenetic Tree

With the variations in levels of Y-DNA testing across all the members of a typical group needing phylogenetic analysis, no assumptions can be made about whether any data source is even available, or how much information is available from any individual data source. The analysis must work potentially standalone using any single data source (Y-SNP, Y-STR, or traditional genealogy), or potentially prioritize and integrate the characteristics of multiple sources.

Given the different data sources and multiple decision points that are necessary to incorporate all relevant data characteristics, our approach uses a classic weighted, multi-criteria neighbor-joining algorithm to select the highest priority available at any point in the new agnatic phylogenetic tree from amongst the various phenetic and cladistic characteristics already described.

The usual decision matrix is replaced here with a prioritized series of decision steps resulting in a final selection of the most-likely, most-closely related pair of kits still left to join. These two kits are then replaced by a new node representing their common ancestor, after which the two kits are taken out of consideration and replaced by the new node.

These decision steps are then applied repeatedly until either no further joining can be performed (in which case no solution is possible), or only two kits or branching points remain left to join, in which case they are joined to form the root of the phylogeny and the tree is complete.

The prioritized series of decisions consists of three steps:

1.2.2.1. Step 1: Handling Y-SNPs and Common Ancestors

For any Y-SNP or common ancestor specified, each kit has a status of positive, negative, or unknown.

If all branching points equated to a Y-SNP or common ancestor whose positive or negative status was known for all kits, there would be enough information to precisely rebuild the agnatic ancestral tree since the cladistic information would completely describe the phylogenetic tree. This is rarely the case, but the cladistic properties of Y-SNPs and common ancestors can still be exploited to limit the set of possible solutions.

Since Y-SNPs and common ancestors on different branches are mutually exclusive, as pairs of kits are joined on the tree under new branching points these branching points can themselves be assigned positive ("+") or negative ("-") status if the status of Y-SNPs or common ancestors below that branching point is all one or the other. In Figure 2, for instance, given the SNP1 status for all kits as shown, branching point BP1 is clearly SNP1+ and BP2 clearly SNP1-.

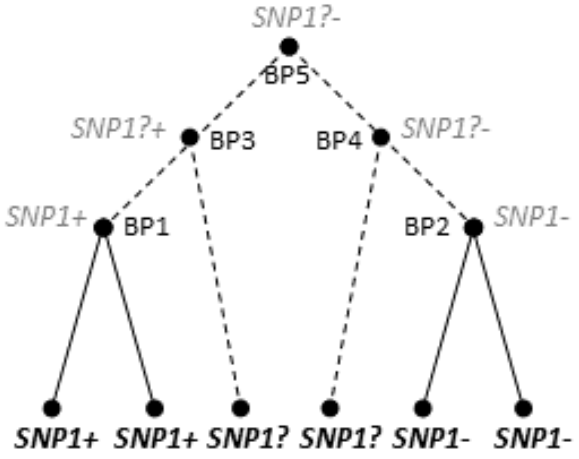


Figure 2. Assigning Y-SNP status to branching points

For unknown ("?") status, it becomes useful to track internally whether the branching point is made up of unknown and positive descendants ("?+") or unknown and negative descendants ("?-"), as illustrated by BP3 and BP4 in Figure 2. This "carries forward" the positive or negative status of subclades through areas of the phylogeny where the status of a Y-SNP or common ancestor may be unknown. In Figure 2, points BP4 and BP5 are clearly SNP1- although there are still unknowns below those branching points so the "?-" status is maintained.

It is important to note that "?+" and "?-" denote very different conditions. A branching point whose status is "?-" is presumed negative for that Y-SNP or common ancestor but is carrying the information that some kits below that branching point are unknown and some are negative. In Figure 2, points BP4 and BP5 are SNP1?- but clearly cannot be SNP1+. A branching point whose status is "?+", on the other hand, is marking an unknown range where the status of a Y-SNP mutation or common ancestor may still be

positive OR negative and therefore is marking the range that that Y-SNP mutation or common ancestor may have occurred on the tree. In Figure 2 for example, SNP1 clearly mutated at some point on the left-most branch below BP5 but may have occurred anywhere on that branch from BP5 to BP1. That range cannot be further reduced without additional information.

As new branching points are formed, their status for each Y-SNP or common ancestor is determined by the status of the two points being joined. The full set of choices is defined in Figure 3.

First Point/Second Point	New Branching Point
Same status for both	Same as both
+/- or -/+	-
Any combination of ?+ and + or ?	?+
+/? or ?/+	+ IF any other +'ve points remain to be joined, otherwise ?+
Any other combination	?-

Figure 3. Assigning New Branching Point Status

The other advantage of tracking these statuses for all kits and branching points is that it sets up two simple rules to implement the cladistic property of Y-SNPs and common ancestors.

Rule 1: Given the sets {S(+)}, {S(?)}, and {S(?+)} of kits and branching points whose status for a Y-SNP or common ancestor is +, ?, or ?+, respectively, then for any two kits or branching points A and B if the following is true:

$$\begin{gathered} A \in \{S(+)\} \\ B \in \{S(+)\} \\ and\ |S(+)| = 2, \{S(?)\} = \emptyset\ and\ \{S(?+)\} = \emptyset \end{gathered} \quad (4)$$

then A and B should be joined directly to each other as the next best pair.

In other words, clades are formed in the phylogeny as soon as their last two points are discovered.

The second rule is:

Rule 2: Given the sets {S(+)}, {S(-)}, {S(?+)}, and {S(?-)} of kits and branching points whose status for a Y-SNP or common ancestor is +, -, ?+, or ?-, respectively, then for any two kits or branching points A and B if the following is true:

$$\begin{gathered} A \in \{S(+)\}\ or\ A \in \{S(?+)\} \\ B \in \{S(-)\}\ or\ B \in \{S(?-)\} \\ and\ |S(+)| \geq 2 \end{gathered} \quad (5)$$

then A and B may not be joined directly to each other.

In other words, the branching point which would cause the clade to be fully formed for a Y-SNP or common ancestor cannot be created if there are still more kits or branching points left to join which are known to be part of that clade.

These rules are applied first to all possible pairs A and B without regard to their Y-STR status. We note again that if each branching point in the tree was fully described by a Y-SNP or common ancestor, there would be sufficient data provided through the Y-SNPs and common ancestors to completely recreate the one correct phylogenetic tree and the Y-STR data would not even be required.

If the data set provided of status of Y-SNPs and common ancestors for each kit is internally inconsistent, this approach may encounter conditions where there are no kits or branching points A and B left to join but the tree is not yet fully formed. For simple conflicts an individual status can be corrected, but in general a conflict in the input data set will prevent the approach from completing.

1.2.2.2. Step 2: Handling Y-STR Signatures

The next best data characteristic upon which to base the phylogeny is the presence of Y-STR signatures. Having already excluded from consideration the kits and branching points which would result in incompatible Y-SNP or common ancestor clades, we know that the consideration of any two points A or B will not violate the Y-SNP or common ancestor clade structure.

Having discovered Y-STR signatures in the input data and associated likelihoods as defined by Equation (2), the rule for joining any two points A and B becomes:

Rule 3: If {SS} is the set of all pairs of kits or branching points which share a Y-STR signature, then for any two kits or branching points A and B if the following is true

$$\begin{gathered} (A, B) \in \{SS\} \\ L(A, B) = Min(L(P_1, P_2))\ \textbf{\textit{within threshold}} \\ \forall (P_1, P_2) \in \{SS\} \end{gathered} \quad (6)$$

where L(A,B) is the likelihood of the signature shared by A and B as given in Equation (2), then A and B should be joined directly to each other as the next best pair. If there is more than one pair (A, B) with equal

minimum likelihood, then use the pair with lowest genetic distance.

"Within threshold" in Equation 6 means that to avoid confusion through convergence, the recognition of Y-STR signatures is dependent upon the degree of relationship between the pairs of points considered, where this degree of relatedness is approximated through their genetic distance. Two points with a higher genetic distance between them will be limited to more unique signatures due to the higher odds of convergence.

The current threshold used in our approach requires that:

$$GD_{ratio} <= 0.17 - (70.0 \times L(A, B)) \quad (7)$$

where GD_{ratio} is the genetic distance expressed as in Equation (1), and L(A, B) is the signature likelihood shared by points A and B as given in Equation (2).

Note that this means signatures are not usually recognized at all between kits or branching points whose GD_{ratio} is higher than 0.17 (or 6 for Y37, 11 for Y67, or 18 for Y111). This threshold can be adjusted if necessary.

The factor of 70.0 in Equation (7) is a mapping from average mutation rates of the Y111 set of Y-STRs to

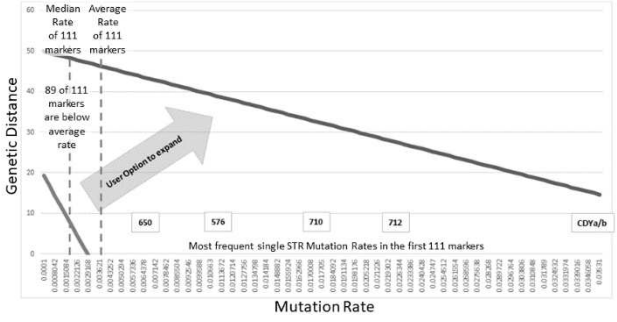


Figure 4. Y-STR recognition thresholds

genetic distances, set so that signatures based on only the most common single Y-STR markers are not recognized as signatures but less common individual Y-STR markers may be recognized at lower genetic distances. This threshold is illustrated by the lower line in Figure 4.

While Equation (7) provides an adequate threshold for signature recognition in most cases, it will not hold for all input data sets, particularly where the occurrence of Y-STR mutations has diverged significantly from statistical rates, or where convergence has occurred to a higher degree. The 70.0 factor in Equation (7) has therefore been made a configurable setting and can be adjusted for any given data set up to the higher line shown in Figure 4. This has the effect of increasing the number of Y-STR mutation patterns which are recognized as signatures.

Note that at lower genetic distances for kits closely related (which would especially apply within genealogical times), Equation (7) allows for Y-STR signatures consisting of one Y-STR mutation to be recognized except for the few most frequently-mutating Y-STRs. This will ensure that where two decisions must be made which each result in a necessary single Y-STR parallel mutation (i.e. a duplicate mutation to the same allele value on different branches), that the decision will be made which minimizes the parallel mutations for less frequently-mutating markers since their signature will be selected first.

1.2.2.3. Step 3: Handling Y-STR mutations without signatures

Once all Y-STR signatures have been recognized, the remaining kits and branching points which are not excluded under Step 1 are analyzed according to the following rules:

Rule 4: Given the set {S1} of remaining kits and branching points and any two kits or branching points A and B, if the following is true:

$$\begin{gathered} A \in \{S1\} \\ B \in \{S1\} \\ GD_{ratio}(A,B) = Min(GD_{ratio}(\{S1\})) \end{gathered} \quad (8)$$

where $GD_{ratio}(A,B)$ is calculated as in Equation (1), then A and B should be joined directly to each other as the next best pair.

If Rule 4 results in several possible pairs (A. B) with near-equal $GD_{ratio}(A,B)$ ("near-equal" is currently set at within 10%), then the next rule is applied

Rule 5: Given the set {S2} of pairs which closely satisfy Rule 4, if the following is true for any pair (A, B) within {S2}:

$$ABS\big(\theta(A) - \theta(B)\big) < ABS\big(\theta(P_1) - \theta(P_2)\big), \quad \forall P_1 \in \{S2\}, \forall P_2 \in \{S2\} \quad (9)$$

where $\theta(n)$ is the vector angle between n and the modal as calculated in Equation (3), then A and B should be joined directly to each other as the next best pair.

The application of both Rules 4 and 5 ensures that while genetic distance remains the general

prioritization for pairing in the absence of Y-STR signatures, where genetic distance is not sufficient to distinguish relatedness the approach also takes into account both the degree of related change and prioritizes less frequently-mutating Y-STRs, since both factors are included in the calculation given in Equation (3). Rule 5 also introduces the necessary correction to offset the systemic error of long branch attraction at close genetic distances.

1.3. Results

1.3.1. Field Testing

This approach was first released as the free SAPP program (http://www.jdvtools.com/SAPP) in March 2016, and has been run regularly by external users between 10 and 80 times every 24-hour period in 2018. Reported accuracy is high and in the author's experience in line with Test Data Runs reported below.

1.3.2. Test Data Production

To test the approach, four test data sets of Y-SNP and Y-STR data for 100 kits each were created assuming varying timeframes back to a single common ancestor as shown in Figure 5.

Test Data Set	Actual TMRCA
G10	10 generations
G35	35 generations
G70	70 generations
G180	180 generations

Figure 5. Time to Most Recent Common Ancestor (TMRCA) for each test data set

To generate these test data sets, the automated test generator started with two descendants along separate lines from a common ancestor a certain number of generations back (as given in Figure 5), and then randomly attached branching points for new descendants at random generations until 100 descendants was reached. That created a randomly-generated, known phylogeny for the 100 descendants. Then a Y-SNP mutation was assigned at each of the branching points and information reported for each descendant on their positive or negative status for each Y-SNP depending on which were found in their ancestral branches. The Y-SNP test data was then reported in three different ways to simulate the typical unknowns which would exist in actual group data:

1. Only the Y-SNP data for the upper branches of the phylogeny was reported first, to simulate groups with some minimal high-level Y-SNP testing;

2. Every other Y-SNP in the tree was reported next, to simulate where some deeper but still not complete Y-SNP testing had been performed;

3. And finally the status of all Y-SNPs was reported for all kits. Note that while complete Y-SNP data for all branching points is not typically found given current levels of Y-SNP testing, this test case was necessary to verify that the approach could achieve the predicted 100% accuracy.

Since common ancestors are handled identically to Y-SNPs in our approach, no traditional genealogy information was generated for the test data.

The automated test data generator then traversed the actual phylogenetic tree starting with an assigned Y-STR haplotype at the common ancestor and evolved the Y111 Y-STR set from Family Tree DNA over the generations, mutating backwards or forwards randomly according to their individual Y-STR mutation rates. This produced randomized but representative Y-STR data at the Y111 level for the entire group. For simplicity, the mutation approach used the step-wise mutation model (1 step at a time) for all Y-STRs, and did not address multi-copy infinite-allele mutations, RecLOHs, null values, microalleles, or the possible higher odds of back-mutations at higher allele values.

This Y-STR data was then reported at Y12, Y37, Y67, and Y111 levels for test purposes.

The full set of data produced is available as linked in the Supplementary Info section.

1.3.3. Test Cases

The following test case runs were then analyzed by the SAPP program for each of the four test data sets:

Y-STR Data	Y-SNP Data	Description
Y12	None	Only 12 Y-STRs provided for all 100 kits, no Y-SNPs
Y37	None	Only 37 Y-STRs provided for all 100 kits, no Y-SNPs
Y67	None	Only 67 Y-STRs provided for all 100 kits, no Y-SNPs
Y111	None	Only 111 Y-STRs provided for all 100 kits, no Y-SNPs
Y67	High-Level	67 Y-STRs and Y-SNP status for high-level Y-SNPs only provided for all 100 kits
Y111	High-Level	111 Y-STRs and Y-SNP status for high-level Y-SNPs only provided for all 100 kits
Y111	Every Other	111 Y-STRs and Y-SNP status for every other Y-SNP provided for all 100 kits
Y111	All	111 Y-STRs and Y-SNP status for all Y-SNPs provided for all 100 kits

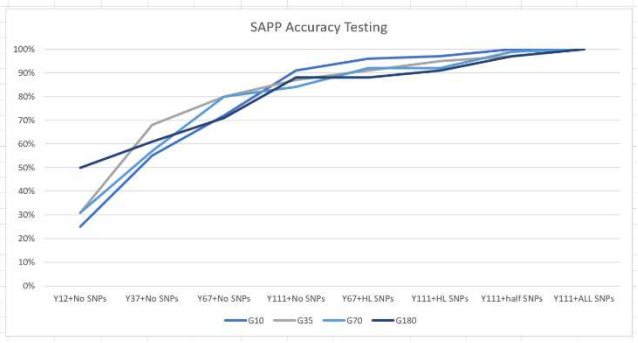
Figure 6. Description of Test Cases

Each run produced a phylogenetic tree, which was then compared to the original "actual" phylogenetic tree for accuracy. The measure of accuracy was how many of the 100 kits were correctly placed in the right position on the phylogenetic tree and therefore how closely the original "actual" tree was reproduced.

It should be noted that accuracy required only that the kits be grouped correctly, not that every single separate ancestral branching point be recreated. Kits which were correctly grouped into the right subclades but where several were placed under one branching point which represented several branching points of the original tree were still counted as accurately placed, since only in the last ("Y111 All") test case is there expected to be sufficient data to uniquely identify every single branching point of the original phylogeny.

1.3.4. Test Data Results

The test data runs yielded the following results for the test cases over the four data sets:



	SNPs	None				High-Level		Every Other	All
Dataset	STRs	Y12	Y37	Y67	Y111	Y67	Y111	Y111	Y111
G10		25%	55%	72%	91%	96%	97%	100%	100%
G35		31%	68%	80%	87%	91%	95%	97%	100%
G70		31%	57%	80%	84%	92%	92%	99%	100%
G180		50%	61%	71%	88%	88%	91%	97%	100%

Figure 7. Test Data Run Results (Graph and Numbers)

1.3.5. Testing Conclusions

Results of the test case runs conformed closely to expectations. Data aggregated from all sources improve accuracy, and more data improve accuracy further. Y12 levels of Y-STR data alone were insufficient to replicate the actual phylogeny with any meaningful accuracy, while Y37 Y-STR data alone achieved better but still highly variable accuracy. Y67 Y-STR data alone achieved 70-80% accuracy, and Y111 data alone achieved 85-90% accuracy.

The accuracy improvement of Y12 Y-STR data alone at longer timeframes back to the common ancestor is explained by the corresponding increase in Y-STR mutations over the smaller data set causing Y-STR signatures to appear and more definition between subclades. At larger numbers of Y-STRs present in the data, this improvement declines.

It was expected that increased convergence would cause a decline in accuracy in the data sets at higher timeframes back to the common ancestor; however, this effect was not observed although there is a modest decline.

The introduction of Y-SNP data yields an expected increase in accuracy, and the introduction of any level

of Y-SNP data appears to nearly close the gap between Y67 and Y111 accuracy without Y-SNPs.

The decline in accuracy in test cases with partial Y-SNP data at higher timeframes back to the common ancestor is explained by larger gaps between reported Y-SNPs and therefore more of the phylogenetic tree which must be recreated using Y-STRs alone.

The consistent 100% accuracy at the Y111-All level was predicted since every branching point is covered by the Y-SNP data and the phylogeny is completely described in the input data set.

It should be noted that accuracy from actual data may be lower than these results for at least three reasons:

1. Actual data typically consists of many levels of Y-STR and Y-SNP testing. With varying levels of test data, not only will some data not exist to make decisions, but the patterns of signatures and other decision criteria may not be obvious,

2. The kits in a group may not actually all be related to the same common ancestor but only seem to be related through convergence. While convergence can be significantly mitigated as already explained, it will be higher in actual data than in our test cases, if the full group has a longer than expected time to common ancestor or if a higher than usual amount of convergence has occurred,

3. The Y-STR mutation rates used to produce the Y-STR test data were the same as the rates used by the SAPP program to recreate the phylogeny. Variations in mutation rates would need to be significant to affect the approach, but if those variations are present in any actual group data, especially without relevant Y-SNP data, they will also reduce the accuracy achieved.

It is difficult to recommend whether this level of accuracy is sufficient for genetic genealogy. Genealogists of course will not and should not be satisfied with less than total accuracy, but we also believe that no data below the reporting of Y-SNPs at every branching point carries within it enough information for total accuracy. For this reason we recommend the SAPP approach as a modeling tool for a *likely* phylogenetic tree under varying conditions, and not as a predictor of the single best phylogeny.

We recognize that maximum-likelihood and Bayesian algorithms may in future improve the phylogenetic reconstruction. However, we would contend *a priori* that the aggregation of relevant phenetic and cladistic information from across multiple data sources currently yields higher accuracy than any existing consistent data source on which a pure phylogenetic approach would be applied. We look forward to a time when phylogenies can be reconstructed using simpler methods with equal or better accuracy.

It is also apparent that more immediate value in agnatic phylogenetic reconstruction can be gained by increasing the levels of Y-DNA testing across the genetic genealogy consumer base.

We also recognize that autosomal DNA testing of the men (or their close relatives) in the group under analysis may provide another potential source of phylogenetic information useful in recreating the agnatic ancestral tree. While autosomal DNA is limited usually to some 5-9 generations in how far back it can provide useful information, it could suggest or prioritize likely recent genealogical connections as a future enhancement to our approach. For now, this information would need to be manually provided into the current approach most likely as traditional genealogy information.

2. Supplementary Info

2.1. Test Data

Test data and results are available on Google Drive (link). Test data is provided in Excel format and SAPP-ready TXT files. Results include the Original ("Actual") phylogenetic tree as PNG images, and trees output by the SAPP runs with color-coded sub-groups, again as PNG images.

2.2. Y-STR-based TMRCA Calculations

While calculating TMRCAs (Time to Most Recent Common Ancestor) is not integral to the phylogenetic approach described in this paper, it is a popular additional requirement for agnatic phylogenies. Consumers want to know not only **how** they are related, but **when** their common ancestors branched off from each other.

Since most groups of kits used in this approach will have had some level of Y-STR testing and an unknown level of Y-SNP testing, a Y-STR-based TMRCA calculation is included in the SAPP program. The calculation uses an approach first described by Ken Nordtvedt which he called "Interclade Estimation" and implemented into his "Generations5" Excel program, and which was further extended and modified with error range estimations by Mark Jost.

The methodology was not published by Ken Nordtvedt for peer review but was reviewed in several online reports including Dienekes' Anthropology Blog (link). It first requires two separate haplogroups of Y-STR data that are known not to overlap but can be at any level of Y-STR testing. If there is no overlap between their phylogenies, then it can be assumed that all the Y-STR alleles of one haplogroup will at some point in their older phylogeny converge with all the Y-STR alleles of the second haplogroup, since they will all converge at the same time in the common ancestor of the two haplogroups.

This leads to the following formula for the number of generations back to that common ancestor:

$$G = \frac{\frac{1}{N_A N_B}\sum_{\forall x \in A}\sum_{\forall y \in B}(x-y)^2}{2\mu} \qquad (10)$$

where x and y are alleles for any given Y-STR sampled from the two haplogroups A and B, and N_A and N_B represent the number of different alleles in the two groups, μ is the mutation rate for the specific Y-STR, and G is the number of generations back to the common ancestor.

The advantage of using this TMRCA calculation with our phylogenetic analysis is that at each branching point in the calculated phylogeny, there are two distinct sub-groups which do not overlap (assuming tree accuracy). The calculation therefore can be performed and reported at each branching point.

The Dienekes blog linked above includes a review and test results for the methodology. We also compared the results of the TMRCA calculations back to the overall common ancestor of our four test data sets for the "Y111-All SNPs" test case with the following results:

Dataset	Actual TMRCA	Estimated TMRCA	Error Range	
G10	10	10	5-16	Within Error Range
G35	35	34	26-42	Within Error Range
G70	70	63	52-73	Within Error Range
G180	180	166	151-180	Within Error Range (Just)

Figure 8. TMRCA Calculation Results

Results were accurate overall with a slight compounding under-estimation observable at increasing timeframes back to the common ancestor. This may be due to the TMRCA calculation method, or it may be an artefact of our test data generation.

It should be noted that TMRCA accuracy was enhanced for the test data over actual data since the test data was generated using identical mutation rates to those used by the SAPP program in re-creating the phylogenetic tree and in calculating the TMRCAs. TMRCA accuracy using actual data will vary in part based on the differences in actual mutation frequency in the input data compared to rates used by SAPP.

We note also that although incorporating the individual mutation rates for the non-matching Y-STRs among the group has a higher likelihood of precision than approaches based on average mutation rates, in general no approach for estimating TMRCAs based on Y-STRs or Y-SNPs currently offers accuracy which most genetic genealogy consumers would consider acceptable, due to the lack of generational precision and wide error ranges. This approach is included in the SAPP program solely as a better "blunt instrument" among many.

2.3. Analyzing the additional 450 Y-STRs provided by Family Tree DNA's Big Y500 test

There is currently much debate about the value of the additional 450 Y-STRs for which at least a subset are reported by Family Tree DNA for each Big Y500 Next-Generation-Sequencing Y-DNA test.

One of the advantages of our approach is that it is independent of the number of Y-STRs used and in fact, can mix together any amount of tests of differing numbers of Y-STRs. So the SAPP program has been run many times with kits which include Big Y500 Y-STR data, and branching is reliably reported including those defined by mutations in the additional 450 Y-STRs.

Since the additional 450 Y-STR data includes many no-calls or unreported allele values, the program has been modified to triangulate values for the missing alleles using three other closest kits. While this may overlook occasional mutations that actually occurred, these were not included in the data in any case so the analysis is not degraded.

One current deficiency in the analyses that include these additional Y-STR values is that mutation rates for these Y-STRs are not publicly available. In the place of published rates, we are using rates calculated by citizen-scientists based on collections of Big Y500 data. This of course also affects the TMRCA calculations.

However, to date it appears that regardless of the ability to analyze these additional Y-STRs, their apparent mutation rate is so slow that they do not hold much value for phylogenetic purposes. The analyses conducted so far have only in about 5% of cases included a branching decision made based on the additional 450 Y-STRs which was not already apparent within the first 111.

2.4. Augmenting Y-SNP input using a Y-SNP Haplotree

Our approach for handling Y-SNPs intentionally exploits their cladistic properties without considering the mutation sequence of those Y-SNPs in relation to each other as they may have occurred among the common ancestors of the group of kits under analysis. This is because in many cases their relationship to each other is not known – they may have only occurred among a very small number of testers and have not yet been mapped into what is commonly known as a *Y-SNP haplotree,* or the ordered phylogenetic tree representing the known sequence and branching of Y-SNP mutation events which has been derived from previous group analysis. By not assuming any given pre-existing Y-SNP haplotree, our approach allows the individual kit Y-SNP results to dictate the logical ordering of Y-SNPs and so derives this Y-SNP haplotree as an identical overlay onto the calculated phylogenetic tree.

At maximum data availability where the actual phylogenetic tree connecting the group under analysis is fully described by positive and negative Y-SNP results for all kits at every branching point, the Y-SNP haplotree for the Y-SNPs within the phylogenetic tree is also fully described by the input data. However as we have noted this is rarely the case with groups considered for analysis; it is much more common for the Y-SNP test results among the kits to contain only partial information. As has also been noted, it is very common for the status of certain Y-SNPs to be known (positive or negative) for a subset of kits and unknown for the rest of the group outside that subset.

In such cases of partial information, knowledge of the Y-SNP haplotree can supplement the provided Y-SNP test results and add valuable cladistic information into the approach. This does not change our approach itself for building the phylogenetic tree, it merely maximizes the positive and negative status of Y-SNPs for the group under analysis and therefore optimizes the input data to further reduce the set of possible phylogenetic trees which satisfies the input criteria.

For example, if Kit1 is positive for SNP1 and Kit2 is positive for SNP2 and those were the only Y-SNP results given in the input data, then it would be assumed that Kit1 is SNP2? and Kit2 is SNP1? since those results had not been provided. However, by knowing from a pre-derived Y-SNP haplotree that SNP1 and SNP2 were on different (incompatible) branches, it can be derived that Kit1 must be SNP2- and Kit2 must be SNP1- even though those specific results had not been provided in the input data. Or, if SNP1 and SNP2 were known to be on the same branch of the Y-SNP haplotree and SNP2 was a child Y-SNP of SNP1 (i.e. SNP2 occurred as a later mutation in a man who was already SNP1+), it can be derived that Kit2 must be SNP1+ as well as SNP2+, although nothing additional could be derived in that situation for Kit1.

In order to supplement the provided Y-SNP input data into the approach SAPP has implemented an internal representation of the known Y-SNP haplotree and also allows users to provide their own knowledge of the Y-SNP haplotree as additional input.

It should be noted that the same rule mentioned earlier applies to Y-SNPs on this internal haplotree; that the only Y-SNPs useful to the phylogenetic analysis are those for which at least two members of the group are known to be positive, and for which at least one other member of the group is known to be negative. Also since the internal Y-SNP haplotree carries no information about the status of any Y-SNP result for any kits, the only Y-SNPs considered are those for which results have already been provided for at least one kit in the input data. Other Y-SNPs which may exist on the Y-SNP haplotree within the branching of the group's phylogenetic tree, but for which no positive or negative status has been provided for any kit, will not contribute any additional phylogenetically-relevant information simply by their position on the Y-SNP haplotree.

The SAPP program has therefore implemented a filter before our approach described in this paper which augments the input data using the internal Y-SNP haplotree (itself augmented by user input) according to two rules:

Rule 1: $\forall S_1 \in \{S\}, \forall S_2 \in \{S\}$*, where {S} is the set of Y-SNPs already specified in the input data, if the following is true for* S_1 *and* S_2*:*

$$S_1 \in \{H(S_2)\} \tag{11}$$

where $H(S_2)$ *is the set of all Y-SNPs in the Y-SNP haplotree at or under* S_2*, then for all kits k in the group, if*

$$k \in \{K(S_1+)\} \text{ and } k \in \{K(S_2?)\} \quad (12)$$

where $K(S_1+)$, $K(S_2?)$ are the sets of all kits in the group positive for S_1 and unknown for S_2, respectively, then set S_2 to positive for kit k.

and

Rule 2: $\forall S_1 \in \{S\}, \forall S_2 \in \{S\}$, where {S} is the set of Y-SNPs already specified in the input data, if the following is true for S_1 and S_2:

$$S_1 \notin \{H(S_2)\} \text{ and } S_2 \notin \{H(S_1)\} \quad (13)$$

where $\{H(S_2)\}$, $\{H(S_1)\}$ are the sets of all Y-SNPs in the Y-SNP haplotree at or under S_2 and S_1 respectively, then for all kits k in the group, if

$$k \in \{K(S_1+)\} \text{ and } k \in \{K(S_2?)\} \text{ or } k \in \{K(S_2+)\} \text{ and } k \in \{K(S_1?)\} \quad (14)$$

where $\{K(S_1+)\}$, $\{K(S_2?)\}$, $\{K(S_2+)\}$, $\{K(S_1?)\}$ are the sets of all kits in the group positive for S_1, unknown for S_2, positive for S_2, and unknown for S_1, respectively, then in the first instance set S_2 to negative for kit k or in the second instance set S_1 negative for kit k.

In other words, Rule 1 says that if a Y-SNP is positive for a given kit, then any Y-SNPs above the first one's position in the Y-SNP haplotree will be positive for that kit as well, and Rule 2 says that if two Y-SNPs are on incompatible branches of the Y-SNP haplotree and the status of one is positive for a given kit, then the status of the second will be negative.

3. Acknowledgements

The author wishes to acknowledge the many users who have reported feedback on the SAPP program over the nearly 3 years in which it has been publicly available. This has led to many usability improvements and features which were not contemplated in the original program.

Many leaders of the genetic genealogy community had either a known or unknown influence on the original development of this approach. We wish to thank in particular Mike Walsh, Maurice Gleeson, James Kane, James Irvine, Mark Jost, Dennis O'Brien, Robert Casey, and Ken Nordtvedt for both their previous work or direct commentary, as appropriate, which led to this approach.

While it is not relevant to the approach used to build the phylogenetic tree, the SAPP program uses a Reingold-Tilford algorithm to draw the resulting tree in graphical form (PNG). Credit goes to Stefan Loewe for his coding of this algorithm on GitHub.

4. Conflicts of Interest

The author declares no conflict of interest and no commercial interests. J. David (Dave) Vance is an IT services executive employed at a major services company. He is an officer of the Vance Family Association and runs their online blog and is the project administrator for the Vance Y-DNA surname project and co-administrator for the R1b-L513 haplogroup project. He is also an active member of the R1b-L21 Y-DNA haplogroup project and various genetic genealogy-oriented online forums and Facebook groups.

As of publication date this approach has been implemented as a PHP program called SAPP which is available online for free use as described in this paper.

5. References

STRBase, National Institute of Standards and Technology, *http://www.cstl.nist.gov/biotech/strbase/*

Dienekes' Anthropology Blog, *http://dienekes.blogspot.com/*

Family Tree DNA (Gene By Gene, Ltd), *https://www.familytreedna.com/*

International Society of Genetic Genealogy (ISOGG), *https://isogg.org/*