
Journal: www.joqq.info

Originally Published: Volume 5, Number 2 (Fall 2009)

Reference Number: 52.009

CLUSTER ANALYSIS AND THE TMRCA PROBLEM: Y-STR MOUNTAINS IN HAPLOSPACE, PART I: METHODS

Author(s): Peter Gwozdz

Y-STR Mountains in HaploSPACE, Part I: Methods

Peter Gwozdz

Abstract

This article introduces a "mountains in haploSPACE" analysis method for evidence that a Y-STR cluster corresponds to a clade. The word "type" is used here for those clusters that can be shown to be relatively isolated "mountains" on the basis of correlated STR values. Excel file analysis tools are available as Supplementary Data, including tools for calculation of size and age of a type. This article presents a new quantitative measure: the statistical background percent (SBP), a rough high estimate of the percent of samples in the data cluster that do not belong to the hypothetical clade for statistical reasons. SBP is presented as an additional method, not to replace other cluster assessment methods. Types and subtypes introduce complication in the estimation of the age of a clade, so this article also presents a discussion of caveats for age estimations based on Y-STR data.

Introduction and Definitions

Short Tandem Repeat (STR) data for Y chromosomes are available on the web, for example at Ysearch (www.ysearch.org), Yhrd (www.yhrd.org) - Willuweit (2007), and Family Tree DNA (FTDNA) (www.familytreedna.com). The methods introduced by this article are generally applicable to Y-STR data; the tools in the **Supplementary Data** follow the FTDNA format of 12, 25, 37, and 67 standard marker sets, plus a few additional rare compound markers, and minus rare missing markers.

The **Summary** on page 156 provides a concise two-page statement of the method of the article, where terms in boldface are defined specifically for this method. Where these terms appear in the text, they are usually shown in *italics*.

"Haplotype" has many meanings. In this article a haplotype is a set of Y-STR values for a specified set of markers. Another meaning of "haplotype" is a "sample"—the set of STR values from a particular man (plural—samples or data or database). For clarity, I avoid that latter meaning, using only the former meaning of "haplotype." A haplotype may have no corresponding sample in a particular database, for example

when it has been artificially constructed, as in a modal haplotype. Markers are also called loci—singular locus.

"Clusters" are sets of Y-DNA samples classified by STR marker value. This definition is imprecise because the word "clusters" is used loosely in the literature and on the web. Some uses of "clusters" do not fit even this imprecise definition. Y-STR clusters are valuable for genetic genealogy because the Y chromosome does not recombine. Clusters are generally provided as hypothetical subdivisions of accepted clades, where accepted clades are based on Single Nucleotide Polymorphism (SNP) markers or other Unique Event Polymorphism (UEP) markers. STR clusters provide clues on where to look in the search for SNP subdivision markers. Clusters are published mostly via web sites; there are scores of web pages that provide cluster classifications. Ysearch has some "modal" STR haplotypes for clusters. FTDNA has a large number of links to "projects," many of which offer proposed cluster classifications.

There does not seem to be a single preferred method for the initial selection of clusters as candidate clades, and indeed I have no method to offer other than the obvious methods: sorting STR data in a search for correlated STR marker values; searching for very common haplotypes (many samples in a database) that are relatively isolated (few neighbors - fewer samples at haplotypes with one mutation step genetic distance); searching for unusual values of a slowly mutating STR marker; using a network - joining computer program in a search for long, isolated branches. Cluster discovery is as much an art as an objective method; many cluster classifications

Address for correspondence: Peter Gwozdz, pete2g2@comcast.net

Received: June 5, 2009; accepted: September 13, 2009

on the web seem to spring from the experienced intuition of the author.

We presume that proposed Y-STR clusters will someday be confirmed or refuted as clades, with the discovery of SNP markers, taken to be the gold standard for identification of clades, called haplogroups. Haplogroups can be treated as clusters insofar as an STR haplotype can be assigned to a haplogroup with high probability albeit not with certainty, for example Athey (2005, 2006). The haplogroup of a cluster is called the stem haplogroup, or parent haplogroup. STR data is rapidly accumulating on the web, so hypothetical cluster subdivisions will no doubt "stay ahead" of SNP haplogroup division for the immediate future.

Cruciani (2004) is a classic example of cluster analysis, with subsequent SNP validation of three of four clusters by Cruciani (2006). The latter article is titled as an evaluation of a network approach because a network-joining analysis of the former article is largely verified. However, it is not clear that the former clusters (defined by STR values) were initially chosen on the basis of the network, as opposed to being chosen by STR values and validated by the network analysis and subsequent SNPs. It is not completely clear how one may select objectively some cluster candidates based on network branches, but reject others. The three 2004 clusters that were validated in 2006 were concentrated in three different geographic areas. The fourth 2004 cluster was not geographically concentrated and was shown in 2006 to be composed of a mix of SNP clades and therefore not a valid clade. Also, the four 2004 clusters are the four largest clusters in the networks; smaller clusters were not proposed as hypothetical clades, so statistical sampling confidence is implicit. Cruciani (2006) is really a validation of the consideration of multiple lines of evidence for cluster analysis, in this case network-joining analysis, geographic concentration, STR correlation, and cluster size.

The "modal haplotype" for a cluster is the set of most common STR values for a cluster, at any specified set of markers.

I use the word "*type*" to mean: (1) a hypothetical Y-DNA clade, proposed as a subdivision of an accepted SNP-defined haplogroup; and (2) a proposed modal haplotype for that clade at any specified set of STR markers; and (3) a set of haplotypes including the modal haplotype and all those STR haplotypes that differ slightly from the modal haplotype, as further defined below (*mountain* in HaploSpace); and (4) a cluster of Y-DNA samples, from a specified database, matching any of the set of haplotypes (at the same specified markers).

All *types* correspond to clusters, but not all clusters (as the word is used in the literature and on the web) are

types because of the restriction (3) to *mountains*, as explained below.

Methods for identification of STR cluster candidates are mentioned only briefly here. This article concentrates on methods to validate the quality of particular "*types*" of cluster candidates with objective formal evidence. Validation in this article means evidence by statistical assessment. Statistical isolation of a *type* is considered evidence (not proof) that the *type* is a clade. Comparing *types*, those *types* with relatively stronger isolation evidence are considered relatively more likely to represent clades.

There are other methods of cluster assessment, but there does not seem to be an accepted method of assessing clusters found by various methods. For example, a cluster based on a single rare STR mutation may correspond to a valid young clade, because for a very young clade a mutation in a slowly mutating marker is almost as good as an SNP, although such a cluster may score poorly in my SBP method and may be missed by a network - joining program. Network joining programs offer assessment as branch length, indicating genetic distance, but such assessment cannot be applied to judge the merit of a cluster identified by a different program or by another means. Also, most network - joining programs connect all samples into branches (clusters and subclusters) so adding just one new sample may rearrange quite a few branches. This is not unique to network - joining programs; any cluster analysis is sensitive to the statistical uncertainty associated with sampling. Statistical analysis is required to judge how robust a cluster is to the vagaries of sample collection. This article presents the SBP method for statistical assessment, not to replace but to add to existing methods of cluster assessment, with the caveat that the SBP method applies to most but not all clusters.

In the published literature I could find no formal article mentioning isolation assessment by consideration of a *gap* ("*gap*" defined below) in the distribution of genetic distance. Perhaps this idea is briefly mentioned in web discussions; for example Mayka (2007) has been brought to my attention.

It is understood that STR clusters (and *types*) are statistical, so even with confirmation by an SNP marker, at least a low percentage of men who closely match an STR cluster will turn out to not belong to the corresponding clade; these are called *background* in this article, and a method is proposed to estimate the *background*. The *statistical background percent* (SBP) is proposed as a high estimate of that *background*.

A relatively small SBP represents a relatively isolated *type*, which is considered to be relatively strong evidence that the *type* corresponds to a clade. This article discuss-

es reasons other than simple statistics why it is not possible to calculate the exact probability that a *mountain type* corresponds to a clade. So for example a 5% SBP does not mean 95% probability that a hypothetical *type* corresponds to a clade.

However, a relatively high SBP can eliminate a cluster from serious consideration: If a proposed STR cluster has SBP greater than 50% (not an isolated *mountain*, or not enough data for statistical significance) that means the *statistical background* (foreigners that do not belong to the hypothetical clade) is significant. The true *background* is probably less than the SBP, because SBP is defined here as an objective high estimate of the *background*.

Also, at least a low percentage of men who seem too distant to belong to the STR cluster will in fact belong to the clade; these are called *type outliers* in this article, and a method is proposed to estimate how many there are.

The size of a *type* is the number of samples in the data cluster for the *type*, with estimated corrections for *background* and for *type outliers*. Size can also be expressed as a percentage of the database from which the *type* is extracted, or as percentage of the stem haplogroup in that database.

It is also understood that most young Y-DNA clades cannot be represented as valid STR clusters. The distribution of STR values tends to be continuous. Particularly with population growth, unique combinations of STR values are statistically unlikely. Very old clades may become STR clusters due to the statistics of long time without population growth, but the oldest clades are already identified as haplogroups. Young STR clusters can appear due to founder effects, such as population bottlenecks or migration. A list of hypothetical clades based on STR *types*, as subdivisions of a current haplogroup, provides only those clades with particularly strong founder effects.

Averaged Squared Distance (ASD) in STR data is used to estimate the age of haplogroups or clusters. This is the same method used by others (e.g., Nordtvedt, 2008). This article describes the methods in detail, and provides references, application notes, and caveats.

Methods: Mountains in HaploSPACE

I introduce in this section a method for objective validation of the quality of hypothetical STR *types*. A **Summary** with boldface terms defined is presented page 156 as a simple method for use by genetic genealogists without expertise in statistics. The different sub-sections of the Methods section below have detailed explanations of the

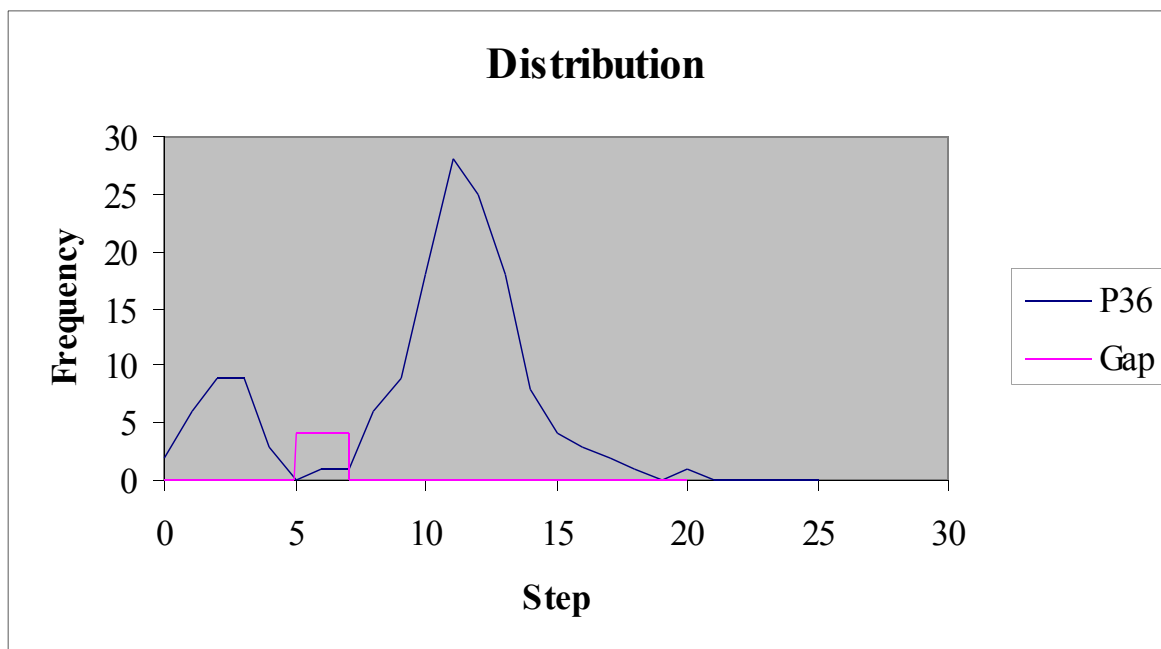


Figure 1. Example of a Mountain Type. The P type modal haplotype uses 36 of the 67 available markers. There are no samples at the pass (mutation step = 5). There are two samples in the gap (steps 5 to 7). There are 29 samples in the mountain (steps = 0 - 4). The P type mountain on the left is well isolated from the rest of the haplogroup on the right.

use of haploSPACE, an outline of more rigorous mathematical models, and comments on statistical issues.

By "*mountain* in haploSPACE" I mean an STR *type* from a particular database, defined by all samples that differ from the *type* modal haplotype by less than a *cutoff* value in genetic distance (total number of step difference in STR values). Modal haplotype selection is explained below. In other words, the graph of step frequency (number of samples per step) vs step from the modal haplotype goes down to a relatively low frequency at the *cutoff* distance. The graph looks like a *mountain*. Figure 1 is an example of a *mountain* from the companion article, "Y-STR Mountains in HaploSPACE, Part II: Application to Common Polish Clades" (Gwozdz, 2009). Steps are counted following the method used by Ysearch.

A "*type*" in this article (see Introduction and Definitions) is restricted to such *mountains*. "*Type*" refers to the modal haplotype, and to the full set of haplotypes that satisfy the criterion of being less than the *cutoff* from the modal haplotype, and to the set of STR samples that match the set of haplotypes, and to the hypothetical clade.

A *mountain* in haploSPACE is isolated from other samples in the stem haplogroup. A *mountain* is considered evidence (not proof) that the *type* corresponds to a clade. The evidence is stronger for a larger number of samples in the database, and for a larger number of samples in the *type*. The evidence is stronger for a lower *cutoff* step value. The evidence is stronger for a low, wide "*gap*" of step values extending beyond the *cutoff*, meaning the *type* is well isolated within the stem haplogroup. The evidence is stronger insofar as essentially the same sample set can be produced by various marker combination choices in the modal haplotype that is used for the definition of the *type*.

Cutoff and *gap* have precise definitions in the Summary.

Rapidly mutating markers provide larger step values (a magnified and blurred image of a *mountain*), so the *mountain* effect is less visible if rapidly mutating markers are included in the modal haplotype used to define the *type*. As an extreme and rare but simple example, suppose a relatively young *type* just happens to differ in value at four slowly mutating markers from the parent haplogroup, and suppose several samples in a large database match these four values perfectly and a few more samples miss by one marker, but no samples miss by two or three, and all other samples in the haplogroup database miss by all four. In this example, a haplotype of only those four markers provides a *mountain* in haploSPACE that seems likely to represent a clade. However, including many rapidly mutating markers in the modal haplotype would provide distributions of values at each marker, with subsequently large total genetic distance, so most samples in the *mountain* would have a

large genetic distance from any such modal haplotype, and some samples from outside the *mountain* would have smaller genetic distance, just due to statistical distribution. Such a *type* cannot be distinguished from the parent haplogroup using many rapidly mutating markers. As data accumulates over time in a database, the *type* modal haplotype values may statistically change for rapidly mutating markers. It is generally necessary to remove the rapidly mutating markers from the modal haplotype definition of a *type* in order to provide an objective and practical *cutoff* value. An objective method for marker ranking is provided 5 paragraphs below.

The very slowly mutating markers may not vary significantly within a haplogroup, in which cases it makes little difference if they are individually included or not in the modal haplotype definition of a *type*. About half the 67 standard markers seem to be like this, in my experience. A large number of such neutral markers in the modal haplotype improves the image of the *mountain* for young *types*, because some of these markers may be mutated for other clades in the stem haplogroup.

Obviously, the best markers for definition of a *type* are slowly mutating markers with modal haplotype values for the *type* that differ from the modal haplotype values for the stem haplogroup. In fact such a correlation of a set of markers would be a nice definition for a "cluster" (Not all users of the word "cluster" follow this idea). I call this the "*signature*" modal haplotype.

We expect relatively younger *types* to be well defined by relatively faster mutating markers, and we expect relatively older *types* to be well defined by relatively slower mutating markers.

There is an important exception: Sometimes a marker that is slowly mutating in most haplogroups has a bimodal distribution in the *type* of interest. A bimodal distribution, particularly a preponderance of two values, is evidence that the *type* is composed of at least two subtypes (population structure). Such a marker can be useful in an attempt to subdivide the *type*, as discussed further below. Such a marker should not be used as a "*mountain*" definition for the *type* itself. With random mutations (no population structure) the values for a given marker should be in a "tent" distribution for low fraction mutated, and in a "bell" distribution for high fraction mutated, as shown for example by Campbell (2007) and Watkins (2007). A marker with a distribution of values that is obviously very non-random may be a poor choice as a marker to define a *type*.

Rank of markers. The ideas of the previous 5 paragraphs are captured in a column of equations for objectively and automatically ranking all 67 markers for a *type*, provided in the tools in the **Supplementary Data**. The ranking is by "concentration." The equations calculate, for each marker: the fraction A of samples in

the proposed *type* that have the modal haplotype value; the fraction B of samples in the full stem haplogroup database that have that value; the "concentration" = $A \cdot A \cdot A / B$; the rank (1 to 67), largest to smallest concentration; and a row with only the best N markers (1 to N; N = 12 by default but is changed by the user during evaluation). I have never seen an objective marker ranking equation published. I have no theoretical justification for that "concentration" equation beyond my comments in the previous paragraphs. I find in practice that this "concentration" captures nicely the ideas of the previous five paragraphs. Other ranking methods may be used, because the *mountain* method does not restrict how markers are ranked or selected.

Recommended method: Try a few markers that seem to be correlated in a cluster of data from a database being analyzed. Use the tools to extract the data with perfect correlation and then try about 9 of the best markers for extracting data for the corresponding *mountain type* cluster from the database. Then try more and fewer markers, adding and subtracting markers by rank. Exclude markers that have problems. Select the modal haplotype that produces the lowest SBP, as defined below. Usually there is a *breadth* of marker sets that produce the same *mountain* data. The lowest SBP is usually produced by a modal haplotype with a number of markers on the high end of the *breadth*, but not necessarily the highest number of markers.

The tools provide a method to quickly calculate step distance for all samples in a database, using various marker sets, and to copy the results to columns for comparison and sorting. The result is a "*mountain graph*" frequency column for each set of markers. A "mask" method is provided for the "calculator" that calculates step count from the modal haplotype; a mask allows rapid removal / addition of markers to the calculator.

When too few markers are used, subtracting one more marker produces a relatively large change, blurring the "*mountain*" and moving some samples into and out of the *type*. When too many markers are used, adding one more marker produces a relatively large change. With a good set of markers adding or subtracting one more marker should make a relatively small change, if any, in the *mountain*.

For example, the *mountain* in Figure 1 has exactly the same 29 samples extracted into the *mountain* using from 13 to 41 markers. Using a 42nd marker captures 28 samples, missing one and adding none.

A self-consistent modal haplotype can be determined iteratively. The modal haplotype is the set of most common STR values for the markers used to define a *type*. However, the cluster data established by a tentative modal haplotype and *cutoff* may produce a new

type data set, with a new modal haplotype that differs slightly from the tentative modal haplotype. Theoretically this may lead to an endless loop, but I have found that inconsistencies are due to non-random markers as explained above. So far, in my experience, cluster candidates either fail to produce a self consistent *mountain* or produce a self-consistent *mountain type* after a few iterations, or produce a high SBP and are slightly inconsistent.

The definition of a *type* (a *type* is a *mountain*) is the modal haplotype, *cutoff*, and *gap*, using the set of markers that produces the best *mountain*. "Best" means smallest SBP, defined below. Usually there is a *breadth* of number of markers for which modal haplotypes produce the same *mountain*. Often within the *breadth* more than one choice produces the same SBP, in which case the definition is the one with the largest number of markers. The rank of markers depends on the method of calculation of mutation steps (see below). The ranking method is intended as only an aide for rapid evaluation; some human judgment is generally helpful in selecting the best markers for a wide *breadth*. The definition may change slightly as more data accumulates in the database over time. If the definition changes substantially with more data that may mean the original *mountain* definition was a statistical fluke, as discussed further below.

The *signature* of a *type* is the modal haplotype using only the best markers. A *signature* is a compact way to publish a *type*. For publication, the *signature* can be restricted to the three to five best-ranked markers. The companion article has a table of *signatures*, for example. A *signature* need not be good enough to extract the *mountain* data exactly. The companion article gives a rare example of a *type* where a *signature* of one marker extracts the *type* data exactly from the parent haplogroup, and any two other markers from the *signature* extract the same data, and the full 67 markers also extract the same samples into the *type*.

Examples and Tools

The companion article provides examples of how the *mountain* method works, with discussion. The **Supplementary Data** includes an Excel file for each *type* from that article, showing the details of the analysis.

Those Excel files for each *type* in the **Supplementary Data** are all copies of the tool "Type.xls," which serves as the master. "Doc" sheets provide documentation and detailed instructions for use of the "Type.xls" file. Documentation is deleted from the copies.

"Type.xls" has a "TypeRank" sheet for the cluster data. *Type* data is sorted by mutation step in a "Calculator" sheet that has the full database. Data for a *type* is copied from the Calculator to the TypeRank sheet, which auto-

matically calculates the full 67-marker modal haplotype, rank of markers, and other useful parameters. An "SBP" sheet calculates the *statistical background Percent* using a column of genetic distance values copied from the "Calculator" sheet. An "ASD" sheet estimates the age various ways, as explained below.

Methods: Background

A *type* is statistical at best. Even for a true clade (verified in the future with an SNP) with a strong *mountain* effect (wide, low *gap*) there are bound to be statistical *type outlier* STR samples from that (hypothetical) clade that fall into or beyond the *gap* if the database is large enough. Conversely, *statistical outliers* from distantly related clades (just beyond the *gap*) are expected sometimes to be included within the *mountain type* in question. In addition to *statistical outliers*, there may also be true foreigners, samples within the *type* data that belong to rare small clades that just happen to have similar marker values, but are distantly related to the main clade (as distantly as clades beyond the *gap*). A *type* admits at least a low probability of including samples in the *type* cluster that do not belong to the corresponding (hypothetical) clade. I call this the *background* from the stem haplogroup, defined more precisely below. The simplest assumption is a uniform *background*, so I assume the *gap background* just beyond the *mountain* measures the *background* present in the *mountain* itself. However, some samples in the *gap* may belong to the *mountain type*, outliers that have randomly more mutations. I use the simplest very rough approximation: that the *gap background* is half the samples in the *gap*, and the other half in the *gap* belongs to the *type mountain*. In order to estimate how many *background* samples are in the *mountain*, I need to digress into a discussion of *haplospace frequency*.

Haplospace Frequency

Haplospace frequency is distinct from step frequency.

Step frequency is the number of samples in the data for a *type* at a particular mutation step value from the modal haplotype. Figure 1 is an example of step frequency vs step.

Haplospace frequency is the number of samples per haplotype in a particular database. Using only the *signature* markers the modal haplotype is almost always the most common haplotype in a *type* cluster for a *mountain*, with the highest haplospace frequency. Haplotypes one step away have lower frequency. Table 1 has total possible haplotype counts. Haplotype count increases with Arabian value, and increases dramatically with the number of markers used. When many markers are used, the vast majority of possible haplotypes beyond two steps have no corresponding samples in a database.

Each cell in Table 1 is the sum of the three nearest neighbor cells above and to the left (except the first two cells). A larger version of this table is available in the "Haplotype Counts" sheet in the file "HaplotypeGenerator.xls" in the Supplementary Data, where the recursive cell formula can be copied to any size table. The "Documentation" sheet has the detailed proof of this simple recursive formula. That file also has a "Type List" sheet that uses a macro to generate the complete list of haplotypes up to a reasonable maximum step for any reasonable number of markers. "Reasonable" depends on the speed and memory of the computer.

Table 1 provides a dramatic demonstration of why most clusters do not form *mountains*. As step size increases, the number of possible haplotypes becomes very large, so even though most haplotypes are not present in the data the net sum of samples at high step count tends to be significant. In that respect, even a high *mountain pass* (a dip in the step frequency curve) is worth consideration.

In multi-dimensional haplospace there is a dimension for each marker being used. Step number is a 1-dimensional projection of haplospace. The graph of step frequency vs one-dimensional step may have a maximum at a non-zero step value. On the other hand, in the haplospace using the definition markers, a *mountain* is very tall; the haplospace frequency near the modal haplotype is much higher than elsewhere for a *type*. Using only the *signature* markers the *mountain* for a *type* is expected to be tent shaped with a very high maximum at the modal haplotype, and with a few smaller peaks near by. The other peaks around the modal haplotype may be due to subtypes or may just be due to sampling statistics.

With a limited database and all 67 markers, most possible haplotypes do not have a corresponding data sample. Even the modal haplotype may not appear in the data. Most of the data samples are singletons, with only one sample at that haplotype. Very few haplotypes will occur more than once, so we cannot determine haplospace frequency in practice for all 67 markers. Sampling statistics will not be much better when using the definition markers except at the modal haplotype.

With a very large database, haplospace frequency (data) is the same as haplospace probability (theory - expectation for more data). But that is only true at the modal haplotype and at a few steps away from the modal haplotype, and only true if few markers are used. Table 1 gives an idea of how large a database must be to accurately measure the haplospace probability. For example, the last row of Table 1 provides the possible number of haplotypes using a modal haplotype with 30 markers. At a step of 1, there are 60 possible haplotypes; for good statistics at least 10 samples per haplotype are needed, so about 600 samples at step 1 are

needed in a *type* to determine the haplotype frequency at step 1 with 30 markers. At step 3 Table 1 shows 36,020 possible haplotypes, so a sample at step 3 for a reasonable size *type* is expected to be a singleton, and most of those 36,020 possible haplotypes will not have a corresponding sample in a reasonable size database, because a *mountain type* with that many samples at step 3 would probably have been identified as a haplogroup by now. The lower right number in Table 1, possible haplotypes at step 10 (still with a 30-marker modal haplotype) is 1.8×10^{11} , which is larger than the human population on Earth.

Step frequency concentrates the data, so sampling statistics can be used to estimate step probability from step frequency. But with the *mountain* method, we study *types* with a low *gap*, so sampling statistics are weak for good *types* at the *gap*.

The uniform *background* assumption does not mean uniform *background* per step. It is unfair to take the *background* from a large number of possible haplotypes per step in the *gap* and apply it to the one (modal) haplotype at step zero. It is even unfair to apply the average *gap* step frequency to any one step in the *mountain*. Much of the *background* is probably at the last step of the *mountain*, just before the *cutoff*. Much of the remainder is probably at the previous step, much of the remainder after that at the previous step, etc. Similarly, the *background* at the *cutoff* is expected to be much larger than the *background* at the last step of the *mountain*. The *background* should increase with step count in the *gap* if the *gap* is more than one step.

The total *background* (in the *mountain*) is therefore expected to be less than (at most not much larger than) the *gap background* (with enough data for statistical significance), because the *gap* has so many more haplotypes. The *background* is certainly expected to be much less than average step frequency in a wide *gap*.

Nevertheless, the *mountain* method defines the *background* as the average step frequency in the *gap*. The *background* is overstated in order to compensate for the unknown small foreign clades (discussed above) that might be hiding in the *mountain*. This overstatement of the *background* punishes poor clusters but has little effect on good clusters. This overstatement of the *background* automatically becomes smaller for *types* with a wide, low *gap*, consistent with the assumption that such isolated *types* are more likely to represent pure clades.

The uniform *background* approximation does not mean uniform for all haplotypes. Uniform means weighted for mutation rate. Most *mountain types* are expected to be much younger than the stem haplogroup. The definition most likely has many slowly mutating markers that are almost always at the value that is modal for the stem haplogroup. The haploSPACE for these markers is like a

plain between *mountain* ranges. There are no hills out there. That's why they work well in a *type* definition, even with the same value in the *type* modal haplotype as in the haplogroup modal haplotype.

Table 1 provides a hint for better approximation methods. Perhaps average haplotype frequency (number of samples in the *gap* divided by the Table 1 number of haplotypes in the *gap*) could be assumed uniform and proportioned in the steps of the *mountain*. I tried that. I could not get it to work in a manner simple enough to propose as a standard.

See **Mathematical Models** for a brief hint at an even better validation theory.

Although not rigorous, the uniform *background* assumption based on average step frequency in the *gap* is offered here as what seems to be a valuable quantitative measure to estimate the percentage of *background* foreign samples in a *type*.

Background is sensitive to the number of markers; it comes out differently when different marker sets are used, so the *mountain* method specifies use of a definition that yields the lowest SBP. Yes, this means selection bias, so the *background* is overstated to compensate for selection bias, and that compensation is smaller for more isolated *types*, where selection bias is not expected to be as much of a problem.

The number of definition markers and the *background* may vary from month to month as more data accumulates in the database. Different people may find different definitions for the same *type* from the same database. The *mountain* method is not perfect. I think it works quite well if the intention is to reliably distinguish *backgrounds* 0.1% vs 1% vs 10% vs >50%.

The **Supplementary Data** has detailed instructions for use of the *mountain* method in the "Documentation" sheet for "Type.xls."

Mountain Discussion

So far, there is no published evidence for fitness or natural selection by Y-STR marker in humans. Vences (2009) found that gene expression can depend upon STR length when an STR falls in the promoter region of a gene, but this finding is unlikely to be particularly relevant to human Y-DNA because the Y chromosome has relatively few genes. There is no published evidence that any STR haplotype may cause a bias toward male descendants in humans. Following most published articles, this article does not consider a *mountain* as evidence of genetic natural selection. This article follows the usual assumption that a cluster is due to founder effects.

There is no a priori reason to expect a Y-STR clade to be a well defined *mountain*. On the contrary, surely most young clades are not *mountains*. A *mountain* is evidence of strong founder effects, where closely related clades happen to be rare in the population. Other clades may be rare by comparison to past rapid population growth that produced the *mountain*. Alternately, if a population becomes extinct, a *mountain* may represent an emigration of a tribe that moved before the extinction. In regional data, an immigrant population may produce a regional *mountain* clade even with closely related clades in the home region.

By definition, every male with male descendants founds a new Y-DNA clade, but we do not expect every such new clade to be a distinct *mountain*. A tent (or bell) shaped *mountain* in values for a selected set of marker dimensions is evidence of rapid initial population growth or isolation or migration or other founder effects.

A "highland in haplospace" is a set of clades that cannot be isolated as *mountains* in a particular database. It may well be that most clades in a particular database belong to highlands. With no strong founder effects, an entire haplogroup may contain no identifiable *mountains* for a limited set of markers and a limited database.

An "island in haplospace" is a *mountain* where the *gap* has no samples. An island is impressive evidence of a clade, but zero data is misleading, because zero data does not really mean zero frequency. With time, as more samples from the same population are added to the database, zero should increase to a small number. The Summary has better definition of an island as less than 5% SBP. P *type* in Figure 1 does not quite meet the definition. The P *type* SPB, now 6.7%, has been declining for the past two years as data accumulates, so it may continue to decline below 5%.

More markers and more data are better. Theoretically, if an infinite number of Y-STR markers were available with data for all men, every male with male descendants would have founded an island *type* clade defined by the unique infinite subset of markers that mutated between his father and him. In practice, more markers and more data should provide more *mountain types* for most haplogroups.

I consider all 67 FTDNA markers when I search for markers to define a *type*. I find that some *types* can also be defined using only the standard 25 markers. Although fewer well ranking markers are available at 25 compared to 67 markers, there is more data at 25 markers so statistics are better. The Excel functions ignore blanks so modal haplotypes can be determined, and ranking can be done, with 25-marker data combined with 67-marker data. The *gap* and SBP obviously need to be done independently, with all data at the same number of markers. Determining the *type* at 25 markers

using only the data with 67 markers provides an estimate of the probability that a sample that falls into the *type* at 25 will also fall into the *type* at 67. Estimating age with combined data is tricky; see the discussion below.

The 37-marker data is not much better than the 25-marker data because most of the additional 12 markers have high mutation rates.

The standard 12 markers offer the most data, so that data is a good place to look for evidence of *mountains*. There is plenty of evidence for *mountain types*, although I have never found a *mountain* at 12 markers with a low SBP. The *gap* at 12 markers is not impressive at first glance. However, there are 24 possible haplotypes at step 1. As shown in the companion article, it is possible for a common haplotype at 12 markers to have about the same number of samples at the modal haplotype as the total at step 1, so the average haplospace frequency (per haplotype, not per step) figured at step 1 is 24 times lower than the frequency at the modal haplotype. All the *types* I have determined at 67 markers look good even at only 12. However, A and K *types* in the companion article share identical 12-marker modal haplotypes. A 12-marker modal haplotype may contain multiple *types* that are not closely related.

Evidence of *mountains* can be found using only the standard nine markers (European) in Yhrd, as shown in the companion article.

A bimodal distribution for a marker (preponderance of two values) is evidence (not proof) of two subtype *mountain* peaks at that marker (mentioned above and discussed further below).

It is possible but unlikely that a *mountain* in haplospace represents two equally large clades that are very distantly related and just happen to have common marker values by chance. It is more probable that a *mountain* includes, in addition to one large clade, one or more distantly related smaller clades, bringing us back to the *background* discussion above, except here I point out that a *background* clade may be large enough to look like a false subclade, and concentrated enough (young) to not overlap into the *gap*, and not contribute much to the *background* calculation.

The *mountain* effect depends on the database. The companion article provides as examples STR *mountains* in the Polish Project that may represent Polish clades, common in Poland but relatively rare elsewhere. A regional *mountain type* may be weaker in the Ysearch database if there are overlapping *mountains* from other regions. Concentration of an STR *mountain type* in one region (or in one cohesive ethnic population) is good objective evidence (not proof) that the *type* is in fact a clade that experienced rapid population expansion. A

Table 1

Haplotype Count for a Given Number of STR Markers and Step Count. The table provides the total number of possible haplotypes at a particular mutation step count from the modal haplotype.

Mar- kers	Number of Haplotypes (by Step Column)										
	0	1	2	3	4	5	6	7	8	9	10
1	1	2	2	2	2	2	2	2	2	2	2
2	1	4	8	12	16	20	24	28	32	36	40
3	1	6	18	38	66	102	146	198	258	326	402
4	1	8	32	88	192	360	608	952	1408	1992	2720
5	1	10	50	170	450	1002	1970	3530	5890	9290	14002
6	1	12	72	292	912	2364	5336	10836	20256	35436	58728
7	1	14	98	462	1666	4942	12642	28814	59906	115598	209762
8	1	16	128	688	2816	9424	27008	68464	157184	332688	658048
9	1	18	162	978	4482	16722	53154	148626	374274	864146	1854882
10	1	20	200	1340	6800	28004	97880	299660	822560	2060980	4780008
11	1	22	242	1782	9922	44726	170610	568150	1690370	4573910	1.10E+07
12	1	24	288	2312	14016	68664	284000	1022760	3281280	9545560	2.60E+07
13	1	26	338	2938	19266	101946	454610	1761370	6065410	1.90E+07	5.40E+07
14	1	28	392	3668	25872	147084	703640	2919620	1.10E+07	3.60E+07	1.10E+08
15	1	30	450	4510	34050	207006	1057730	4680990	1.80E+07	6.50E+07	2.10E+08
16	1	32	512	5472	44032	285088	1549824	7288544	3.00E+07	1.10E+08	3.90E+08
17	1	34	578	6562	56066	385186	2220098	1.10E+07	4.90E+07	1.90E+08	6.90E+08
18	1	36	648	7788	70416	511668	3116952	1.60E+07	7.60E+07	3.20E+08	1.20E+09
19	1	38	722	9158	87362	669446	4298066	2.40E+07	1.20E+08	5.10E+08	2.00E+09
20	1	40	800	10680	107200	864008	5831520	3.40E+07	1.70E+08	8.00E+08	3.30E+09
21	1	42	882	12362	130242	1101450	7796978	4.80E+07	2.60E+08	1.20E+09	5.40E+09
22	1	44	968	14212	156816	1388508	1.00E+07	6.60E+07	3.70E+08	1.90E+09	8.50E+09
23	1	46	1058	16238	187266	1732590	1.30E+07	8.90E+07	5.20E+08	2.70E+09	1.30E+10
24	1	48	1152	18448	221952	2141808	1.70E+07	1.20E+08	7.30E+08	4.00E+09	2.00E+10
25	1	50	1250	20850	261250	2625010	2.20E+07	1.60E+08	1.00E+09	5.70E+09	3.00E+10
26	1	52	1352	23452	305552	3191812	2.80E+07	2.10E+08	1.40E+09	8.10E+09	4.30E+10
27	1	54	1458	26262	355266	3852630	3.50E+07	2.70E+08	1.90E+09	1.10E+10	6.30E+10
28	1	56	1568	29288	410816	4618712	4.30E+07	3.50E+08	2.50E+09	1.60E+10	9.00E+10
29	1	58	1682	32538	472642	5502170	5.40E+07	4.50E+08	3.30E+09	2.20E+10	1.30E+11
30	1	60	1800	36020	541200	6516012	6.60E+07	5.70E+08	4.30E+09	2.90E+10	1.80E+11

regional *mountain type* may also be evidence (not proof) of one or more immigrations from a distant region.

The **Supplementary Data** "Type" files each have a copy of all the database from only one haplogroup. Using the full database is tedious. When first analyzing a *type*, I use the full Polish Project database. It is important to include the data from sister haplogroups, and from the upstream stem haplogroup. The latter is the samples that have not been assigned to any of the downstream sister haplogroups. Once it is established that all the samples from other haplogroups fall far beyond the gap, the detailed analysis can be done using only the sub-database from the one parent haplogroup for that *type*. Obviously, this needs to be checked from time to time. If there is overlap in the *gap* with a sister haplogroup or with the upstream unassigned stem haplogroup, "cousin" haplogroups need to be checked, and data from all relevant haplogroups should always be included in analysis for that *type*. Publication of a *type* should include a comment about the step distance to the nearest outliers from other haplogroups, using the definition.

It would be interesting (beyond the scope of this article) to treat SNP haplogroups as *types* and to figure the SBP, as a quantitative measure of confidence that a sample can be assigned to a haplogroup on the basis of STR markers.

The rank of markers generally depends on the database, changing even when other haplogroups are added from the same parent database. For example, a marker that ranks 9th best in the haplogroup database may jump to 3rd place when another haplogroup is added to the database, if that marker ranks well for distinguishing the two haplogroups. If multiple *types* are identified, rank of markers for one *type* also changes if other *types* are removed from the haplogroup database. The rank of markers for a subtype come out very differently if the *type* is used as the database instead of the full haplogroup; in other words, the best markers for distinguishing a subtype from the *type* are not the same as the best markers for distinguishing the subtype from the haplogroup.

The rank is a tool to find a *breadth* of marker sets for a *type*; human intervention can eliminate troublesome markers and add good ones that for some reason rank poorly.

When two *types* are identified with overlap, the isolation of one generally improves when viewed with the other removed from the database. The companion article gives an example, where samples from one *type* tend to fall in or near the *gap* of the other *type*. The two *types* conditionally reinforce each other; if one is valid then the other is more likely valid. I do not propose a quantitative complication of SBP that takes this effect into consideration. It seems better to keep SBP simple. Consideration

of the overlap of *types* (or lack of overlap), however, provides qualitative evidence for consideration.

Cluster search is as much an art as a formal method. There are a number of ways to come up with clusters that are candidates for *types*. This article concentrates on validation. Here are four brief comments concerning a cluster search:

Obvious candidates are the most common haplotypes, using only the standard 12 markers, or the Yhrd (European) 9 markers, particularly if all the 1-step haplotypes (nearest neighbors) are relatively rare.

Sorting by various markers in an Excel sheet is efficient. Most sorts of STR data by marker value do not produce *mountains*. Lack of a *mountain* is missing evidence (not disproof) that a marker combination represents a clade.

"Evaluator.xls" in the **Supplementary Data** has a macro that figures, for each sample in a database, how many other samples have "infinite alleles" mutation step distance 0, 1, 2, 3 ... up to a user specified set of maximums. Clusters stand out in the sorted result.

Software is available on the web to arrange the data from a haplogroup into a median joining network. A good *type* should provide a bush on such a network, with the bush sticking out on a long branch, if the software displays genetic step distance as line length. The long branch may have a few sparse side shoots. Such software results are highly sensitive to whether or not rapidly mutating markers are included.

The search for STR clusters is prone to false positive bias, also called type I statistical errors. The false positive probability increases with the effort spent searching for clusters. As a reader, you need to consider how many people are searching for clusters, reporting only the ones that seem statistically significant. After all, with enough effort it should be possible to find *mountain types* even in small databases randomly generated by computer simulations of populations without any actual structure. False positive bias is difficult (close to impossible) to quantify because it depends on search effort and also depends on the population structure of the haplogroup. The *mountain* method allows choosing the best SBP from any number of trial modal haplotypes, which introduces selection bias, another type I error. The *mountain* method compensates for these biases by applying worst case statistics, but the same compensation is objectively applied to all *types*, so situations with more type I bias are not distinguished.

Web Databases

The Ysearch database (www.ysearch.org) includes modal haplotypes, fictitious entries for research purposes. These are a convenient way to publish results, but they

confuse inexperienced users with false matches, because by construction they match well with a lot of samples. Also, these need to be removed from the data for valid analysis. We don't want modal haplotypes included in our *mountain* graphs. I hope this article does not inspire a lot of modal haplotype clutter on Ysearch. Please enter only your most significant results, with all definition markers. Be sure to name it as a modal haplotype, and include comments. You can enter an experimental modal haplotype, work awhile with it, and easily delete it when finished.

Signatures should not be left in Ysearch, because they would produce too much clutter. Anyway, it is easier to use Ysearch "freeentry" URLs for *signatures*. Example URLs are available in the **Supplementary Data**. These examples can be easily edited to create new URLs for research.

FTDNA projects (www.familytreedna.com) do not include modal haplotypes. FTDNA projects include family sets, apparently solicited by one person. Such family sets may falsely bias data toward a larger cluster, because the data were not randomly sampled. The large cluster in the database implies a corresponding large cluster in the population. Other clusters are thereby falsely biased (very slightly) to a smaller percent of the database, and therefore seem to be smaller in the population. Family sets can be identified by sorting the data and noting identical family names for sets of very similar haplotypes. "Evaluator.xls" in the **Supplementary Data** can find these. All but one of each family should be removed from the downloaded database, leaving a representative with the most markers. An exception would be a family name that is very common, where the data may be independent. Pairs of similar haplotypes without identical names may be independent entries that perhaps should not be removed. On the other hand, a person may contact men who match him at 12 markers and encourage them to expand their data to 67 markers, thereby creating an "extended family" cluster in a 67-marker database. I see no way of identifying these false 67-marker clusters other than contacting the men and asking them about this. Obviously, any editing should be explained for each STR database that is studied. If all clusters have equal probability of containing small family sets editing would not be significant, but I have seen unusually large family sets that clearly should be edited. Insofar as this editing may leave some family sets, data is biased slightly, toward larger clusters. Insofar as this editing may remove men who submitted data independently, data is biased slightly, toward smaller clusters.

Ysearch also includes family sets. Some men enter their data more than once into Ysearch. In Ysearch it is possible to individually check the contact person name for each data sample. It is tedious. I have not developed an automatic method to find Ysearch family sets by

contact person name. Without editing, Ysearch data is biased toward larger clusters.

Yhrd (www.yhrd.org) has neither modal haplotypes nor family sets. Yhrd in Europe is mostly forensic data. The documentation indicates that only one sample from matching forensic pairs is included. Any matching pairs would bias the data toward larger clusters.

Methods Compound Markers; DYS464

Compound markers have potential for confusion. A "Calculator" tool was developed to calculate mutation steps four ways: treating a compound pair as individuals with step difference, or using the infinite alleles method, or using an assigned equivalent mutation step for a recLOH, or using a maximum count for a marker. "Calculator.xls" in the **Supplementary Data** has a "Documentation" sheet including a detailed explanation.

An example of recombination loss of heterozygosity (recLOH): DYS385(a,b) = (10,14) can turn into (10, 10) with a single mutation that copies the shorter chain to the location of the longer copy, yielding a step count of 4. Using infinite alleles, the mutation count is only one for the pair.

The "Calculator" by default follows the methods used by Ysearch to be compatible with the step counts at that web site. In determining matches in searches in Ysearch, compound markers are treated as individuals except for YCAII and DYS464. For each compound marker pair treated as individuals, it is necessary to check to make sure there is no problem in the data for each *type*, as discussed for example in the companion article. The most common problem is a sample that fits a *type* well except a compound marker recLOH mutation that causes a spurious high step count. A valuable ranking marker could be missed; a solution is to use one of the alternate step counting formulae. ASD age can be anomalously increased by a low fraction of samples with recLOH; a solution is to disregard that marker for the ASD age, or to modify the ASD equation at that marker, or to manually adjust the data.

Ysearch treats YCAII using infinite alleles, which is equivalent to counting matching pairs and subtracting from 2 for genetic distance. YCAII seems to have a high proportion of recLOH mutations, so using step differences for each of the two markers would provide larger genetic distance.

DYS464 is special. I found that using one or two of the DYS464 markers as individuals turns out to be very useful in ranking markers for some *types*. These are not really individual markers. The FTDNA format sorts DYS464 copies in order of STR length, so for example DYS464c is just the 3rd largest in the set of copies, and

could be any one of the physical copies. DYS464 is the only compound marker in the standard 67 with more than two copies; four copies are standard; up to seven copies occur. I have seen comments on the web that this set of markers is highly prone to evaluation errors during the scoring process following the tests.

Searches in Ysearch treat DYS464 using infinite alleles. I have not found documentation for this Ysearch method; it was pointed out to me by Mayka (2008) and I verified it. Here, infinite alleles genetic distance is equivalent to: [counting matching pairs (2 equal values, one from each of the two haplotypes being compared) using all markers for DYS464, with the matching pair count subtracted from the number of copies in the haplotype that has more copies]. However, I notice that Ysearch reports the number of markers as four even if more are present and used for the count. DYS464 mutates rapidly. When more than one of the DYS464 marker copies rank well, care should be exercised, because a recLOH mutation often causes a misleading step count.

It is a bit surprising that treating a single DYS464 value (in the FTDNA format) as an individual marker has value for *types*. I figure the *type signatures* and definitions either with or without the full DYS464 set, in case the reader is skeptical of using an individual DYS464 value to define a *type*. For *types* that are better defined with lower SBP using only one or two values from the set, I point this out as an additional comment.

Infinite alleles is quite useful for evaluation of compound *types*, perhaps better than step count in many cases, but I generally use the conventional step count in order to be compatible with Ysearch (where infinite alleles is used for YCAII and DYS464).

DYS389-2 is unique, because the corresponding STR is not resolved by commercial DNA tests, it is scored as a sum with DYS389-1 included. The true STR is a difference: DYS389-2 minus DYS389-1. Example: DYS389(1,2) = (13,30) vs (14,31) is 1 step in DYS389-1 but no step in DYS389-2, difference 17 for both. No problem after coding into Excel. For some *types*, that difference STR may be a *signature* marker, while DYS389-1 may not be useful. In my Excel files it is easy to use DYS389-2 in a modal haplotype definition, masking out the modal DYS389-1 value, because the data for DYS389-1 is always there. However, I see no way to use DYS389-2 in a definition on Ysearch without DYS389-1. For compatibility with Ysearch, I do not use DYS389-2 without DYS389-1 in definitions. For *types* that are better defined using only the difference STR, I point this out as an additional comment.

The mutation rates of Chandler (2006) are for an individual marker in compound markers. The rate can be multiplied by two as a rough estimate for the rate for a

compound marker pair, but not exactly because a single mutation may change the pair sequence and thereby change both. For DYS464 multiply by four for rough estimate of net mutation rate.

Some haplotypes have nulls, missing markers. Calculator.xls provides methods to use an equivalent step count for nulls.

Methods: Statistics

For a given database, the *mountain number* (number of data samples in a *mountain*), or the *gap number* (number of samples at the *gap*), or other data counts, may not be representative of the population due to sampling statistics. For large data sample counts, the one sigma confidence, 70%, is the square root of the count. Two sigma is two times the square root of the count. The 95% confidence is close to two sigma. For small samples, standard statistical tables can be used to determine the confidence interval, for example at <http://health.utah.gov/opha/IBIShelp/ConfIntns.pdf>.

In the **Summary**, I recommend Poisson statistics for each data count, because most data counts will be small numbers from a large database. Since samples in the *gap* are weakly correlated to the samples in the *mountain*, Poisson statistics are not rigorously valid, but close. Also, the full database may not be large enough for Poisson statistics to be exactly correct. However, these are nit-picking considerations compared to other much larger statistical considerations discussed a few paragraphs below. As a simple standard, I recommend the Poisson 70% confidence interval, applied to each sample count, with more reasons below.

Table 2 has the 70% Poisson confidence intervals for sample counts from 0 to 32. For counts larger than 32 an approximation is the count plus and minus the square root of the count. For example, the square root of 32 is 5.657, so the approximate confidence interval for 32 is (26.3 - 37.7), close to (26.2 - 38.9) in Table 2. The **Supplementary Data** sheet "SBP" within the file "Type.xls" uses the Poisson distribution for rapid exact calculation of SBP from user data input.

Background percent is calculated from both the *mountain number* and the *gap number*, so technically a root of the squares factor should be used for the confidence interval of the *background*. This isn't exactly true because the numbers are weakly correlated. However, the *mountain* method uses the worst case - 70% minimum *mountain number* and 70% maximum *gap number*. There are two reasons. The first reason is to keep it simple for people using Table 2 instead of an Excel file. The second, more important reason, a simple objective method to account for other statistics, takes a few paragraphs to explain:

The "SBP" sheet calculates SBP for several combinations of *cutoff* and *gap* values. The user is allowed to reconsider *cutoff* and *gap* values in order to obtain a better SBP. Also, the user is expected to try the sheet "SBP" for a number of different marker sets that look good for the definition of the *type*. The best SBP may motivate the user to reconsider the samples considered part of the *mountain*, used in the definition of the *type*, so the search for minimum SBP is iterative.

That's selection bias, a type I statistical error. False positive bias, another type I statistical error, was discussed above.

Reminder: SBP is an assessment method, not a method to initially find clusters. The tools in the **Supplementary Data** may be useful to find clusters based on correlation of STR values. SBP may be applied to a cluster defined by any other means. SBP may be used to "fine tune" the definition of a cluster identified by another means.

A 3rd type I error: Maybe men with a common 12-marker haplotype are more likely to join a DNA project and purchase all 67 markers, compared to men with rare 12-marker haplotypes. There might be a distortion in the data, toward larger more isolated *types*, particularly at 67 markers. We don't know if this is true. It might be opposite - toward smaller *types*.

Type I errors are very difficult, close to impossible, to quantify. It is assumed in this article that a relatively more isolated *type* is relatively less affected by type I errors.

One more big statistical issue; self consistency: The sampling confidence interval for the number of data samples in a *mountain* is available in Table 2, but that is only the confidence for the *mountain* size as defined. The number is also uncertain due to the self consistency issues discussed above, associated with parameter choices - modal markers and marker values and *cutoff* value and *gap*. This is tedious and subjective, so I do not recommend it as a standard method. A "self consistency" confidence interval can optionally be estimated by noting the minimum and maximum *mountain number* and *gap number* as parameters are varied within and outside the *breadth*. A subjective judgment needs to be made for 70% confidence. This self consistency confidence interval is independent of the sampling confidence interval from Table 2, so they should be combined as root of the squares. This is complicated. If done, it makes sense to also report all this in a publication along with my recommended simple standard SBP.

We need to estimate the net confidence interval for SBP, including the issues of the previous paragraphs. A simple way to do that is to downgrade the statistical confidence interval in a consistent and objective way. It

makes sense to select a method that downgrades large SBP's more than small SBP's, because it is assumed in this article that a large, well isolated *type* with small SBP is less likely to suffer from type I error and self consistency error than a relatively smaller, less isolated *type*.

One way would be to use a root of the squares method for the various statistical issues, with some kind of objective standard for type I errors, where that standard is wider for larger SBP. In practice this turns out to be too complicated. The worst case method is much simpler.

The worst- case calculation at 70% confidence for each number works fine in my experience, yielding SBP values that make sense to human intuition. Statistically, the worst- case calculation is equivalent to 80% to 95% sampling confidence, depending on the details of the data. However, considering the type I and self-consistency issues, the actual net confidence is unknown, surely closer to 70% than to 90%. So SBP as prescribed by the Summary equations is presented as a rough estimate of the statistical uncertainty, all things considered. Calling the SBP a 70% confidence is not rigorous, but it meets the intention of a reasonable, simple standard that includes a downgrade to account for type I errors. I also judge the Summary method to be understandable and reasonable to users without statistical expertise, who do not fully understand the discussion in this section.

I hope this SBP number can serve as a standard comparison of published *types*. *Types* with lower numbers can be considered better than *types* with higher numbers, more likely to be confirmed in the future with discovery of an SNP. More data should decrease the SBP for valid *types*. Authors with statistical expertise can add more complex confidence intervals to publications.

Reminder: there are other methods for validation evidence for clusters, independent of SBP, discussed above, so SBP is not the full story.

SBP subtracted from 100% should not be misconstrued as the probability that a man whose sample falls in the *type mountain* belongs to a corresponding clade. The man's probability of belonging to the clade is the probability that the *type* is in fact a clade multiplied by the statistical probability that his sample belongs to the *type*. The first probability is difficult if not impossible to calculate (discussed above and below). SBP subtracted from 100% is intended as a very rough low statistical estimate of that second probability. A better estimate would be to use the actual counts for the *background* percent instead of the 70% worst case. An even better estimate would be to use haploSPACE frequency instead of step frequency. With haploSPACE frequencies, *background* percent can be calculated as a function of step, which is 100% minus the

percent probability that a man with a sample at that step belongs to the hypothetical clade. An even better estimate would use a mathematical model (next topic) for *background* percent vs haplotype. These better estimates are too tedious to propose as a standard; I mention them here only to clarify the meaning of SBP.

I tried using 90% or 95% confidence intervals, but that did not work well, because SBP comes out badly too often using high confidence intervals. The computer insists that a wider *gap* is better, because a higher count at 95% confidence comes out better when averaged over a wider *gap*. However, human intuition with common sense cannot accept a wide *gap* that obviously includes the high counts beyond the *gap* on either or both sides. High confidence intervals require subjective judgment.

Size is number of samples in the *type* adjusted for *background*. Confidence interval in the size should inherit the root of the squares of the uncertainty factors from both the *mountain number* and the *gap number*. But again, the Summary simplifies the statistics, using the *mountain number* confidence interval as a size confidence interval.

Mathematical Models for *Types*

One way to quantify the self consistency of a *mountain type* is to fit the data to a mathematical model. I started to develop an Excel model using equations from the literature, for example Campbell (2007) and Watkins (2007). That unfinished project is too large to be included here. My Excel-based mathematical model is available in the **Supplementary Data**, but it is tedious to use, even for a person very familiar with Excel.

I do have an important observation to report. When fitting the distribution of haplotypes from data of a *type* to a mathematical model, I invariably noticed that specific markers did not fit at all. This lack of fit by specific markers is evidence of population structure, which is to say subtypes with strong founder effects, discussed above and discussed again below. Multiple peaks within a *mountain*. It makes no sense to fit the data to a simple *type* model when there is obvious evidence of subtypes. It was necessary to subdivide the model into a compound model. Invariably, the evidence for a subtype is based on minimal data. As data accumulates on the web after a few months, adding or subtracting evidence of the first subtypes, more minimal data subtypes are noticed, requiring further subdivision. It may be I just happened to be working on complicated clades. On the other hand, it may be that human Y-STR data usually has complex population structure (subtypes within *types* - hills within *mountains*). When we are lucky enough to find a well isolated *mountain*, it may be asking too much for that *mountain* to not have multiple peaks. There is a bright side to this observation: If a compound model needs to be developed in order to quantitatively model a

typical isolated *mountain type*, then that compound model should be applicable to any haplogroup, modeling the haplogroup as a set of *mountains* that overlap as a highland *mountain* range even when there are no low *gaps*. My motivation for trying to develop such a model is that it automatically provides age (next topic).

Methods: Infinite Alleles Age

Age of a *type* can be estimated from the number of mutations. An older *type* has more mutations than a younger *type*. There are a number of caveats to such age estimations. Caveats are discussed in the next three sections, after an outline of two methods: infinite alleles and ASD.

For an STR marker with a low fraction of mutations in a clade, the infinite alleles model can be used, where the value of the mutated marker is ignored. Any mutation gets a step count of one. A compound marker gets zero only if all copies are identical to the modal haplotype, and a single one otherwise. The age of the clade is fraction mutated divided by rate (mutations per generation) times a standard number for years (for example 25 or 30) per generation.

There are two ways to average over a set of markers: (a) average of age by marker, or (b) average fraction divided by average rate. The resulting age is not exactly the same for the two methods. It is easy to modify a spreadsheet to change from one method to the other. Arguments can be made for (a). The most common approach in the literature is (b), so I use (b).

The ASD age, next topic, should be used if there is a significant fraction of mutated samples for a particular marker, because of the probability of back mutation to the modal value, and because of the probability of multiple mutations at the same site.

Infinite Alleles and ASD both have "population structure" and other considerations, discussed in the next sections. Without these considerations, there should be an equal number of +1 and -1 mutations within statistical expectation, and very few multiple step mutations, for markers with low fraction mutated.

The **Supplementary Data** "ASD." sheet within the file "Type.xls" has automatic calculation of infinite alleles age, using only those markers that pass a test. The default test is less than 10% mutated, but that 10% cell can be changed by the user. The user can specify which markers to exclude for statistical considerations, as discussed in the following sections for ASD. Mutation rates of Chandler (2006) are used. Generation time can be changed from the default 25 years.

The infinite alleles age is useful as a sanity check comparison to the ASD age. A large difference (for each

marker with low fraction of mutations) may be due to multiple step mutations, such as recLOH. Infinite alleles age is useful for Compound Markers, particularly DYS464. Averaging markers gives a bias toward young infinite alleles age in the "ASD" sheet because of selection by low fraction mutated.

Methods: ASD Age

Averaged Square Distance (ASD) in STR data is traditionally used to estimate the age of a clade, explained for example by Goldstein (1995). Age of the clade in years is ASD divided by rate (mutations per generation) times a standard number for years (for example 25 or 30) per generation. This method is not exactly correct for technical reasons. The age of a clade is assumed to be the time of the most recent common ancestor (TMRCA), although this may be misleading for a number of reasons, some of which are discussed here in this article. Athey (2007a) is an example of a recent application of ASD to determine the age of a set of Y-DNA haplotypes. A classic application was by Thomas (1998).

Variance (standard deviation squared) is available on modern spread sheets, for example as the function VAR in Excel. Population variance, available in Excel as VARP, is identically equal to ASD. Although arguments can be made that VAR should be used, I use VARP for ASD to be consistent with the practice in the literature and on the web. The difference depends upon the number of samples N, where $VAR * (N-1) = VARP * N$. With 10 samples the ASD difference using VAR vs VARP is 10%; with 33 samples the difference is only 3%. A user can easily change VARP to VAR in the **Supplementary Data**. It does not really matter because sampling confidence interval is larger for small data sets and because systematic age uncertainties (next section) are large even for large samples.

If the data for a marker has an equal number of +1 and -1 mutation steps and no multiple steps ASD is identically equal to the fraction of samples mutated, so the ASD age is identical to the "infinite alleles" age. If the ASD age is significantly different than the infinite alleles age for a marker with low fraction of mutations, that is evidence of population structure or sampling variation, discussed below.

ASD is the average of squared distances from the mean for all samples. A different definition of ASD (not used here) is average squared distances of all pairs of samples from each other. $ASD (all\ pairs) = 2 \times VAR$ is a curious mathematical identity that causes a factor of 2 confusion in some web discussions.

ASD can also be defined between two clades, where ASD is the average of squared distances from all pairs, one sample from each clade, and where again TMRCA in generations can be estimated as ASD divided by muta-

Table 2
70% Confidence Intervals, Small Samples, Poisson Statistics

Number	Low	High
0	0	1.9
1	0.2	3.4
2	0.7	4.7
3	1.3	6
4	2	7.3
5	2.8	8.5
6	3.6	9.7
7	4.3	10.9
8	5.2	12.1
9	6	13.2
10	6.8	14.4
11	7.6	15.6
12	8.5	16.7
13	9.3	17.9
14	10.2	19
15	11.1	20.1
16	11.9	21.3
17	12.8	22.4
18	13.7	23.5
19	14.5	24.6
20	15.4	25.7
21	16.3	26.8
22	17.2	28
23	18.1	29.1
24	19	30.2
25	19.9	31.3
26	20.8	32.4
27	21.7	33.5
28	22.6	34.6
29	23.5	35.7
30	24.4	36.8
31	25.3	37.9
32	26.2	38.9

tion rate. In this respect, ASD for a single clade can be conceptualized as the distance of the clade from the mean, where the mean is the "other clade," and where the mean is a fractional expectation value for the modal value (the founder - MRCA) at that marker.

The Thomas (1998) ASD is 0.2226, which I use as a paradigm for a young clade. Thomas averaged ASD for 5 STRs (6 in the data, 5 for ASD), a subset of the FTDNA standard 12. In other words, a cluster proposed as a hypothetical clade looks relatively young if the STR ASD averaging the same 5 markers is less than 1/4. The Thomas age is 2,650 years based on an average effective mutation rate of 0.0021 per 25 year generation. This Thomas method is available in the "ASD" sheet in the **Supplementary Data**.

The (a) versus (b) averaging method comment for infinite alleles above also applies to ASD. Some other statistical objections to the averaging of ASD from multiple markers are discussed in this article. Age by marker may vary greatly, as discussed in the next section.

There are rare reasons, such as a point mutation within an STR, whereby a particular marker may have a unique mutation rate in a particular haplogroup, not discussed further in this article.

ASD age can be calculated using a copy of the master file "Type.xls." The "ASD" sheet automatically takes data from the "TypeASD" sheet. The user copies the STR data cluster into the latter. The "ASD" sheet calculates: ASD by marker, average ASD per the Thomas 5 markers, and average ASD using all markers or using a set of markers specified by the user. Age is calculated. Generation time can be changed from the default 25 years.

Obviously, if a cluster is defined by particular STR values, those markers cannot be used for estimating the age. However, with the *mountain* method introduced here, all markers for the cluster of a *type* can be used, because mutations in the defining markers are included in the cluster data.

For *mountains* with high *background*, age calculation using either infinite alleles or ASD is only a very rough age approximation, because the *mountain* cluster data probably contains outliers (*background*) from other clades (neighboring *mountains*). The selection of only those samples below the *cutoff* provides an age approximation that is biased too young.

Age estimation of a clade is very sensitive to the outliers—those samples that happen to have the most mutations. If age is calculated including the *gap* in order to capture outliers, the data likely includes foreigners that do not belong to the hypothetical clade, providing an age approximation that is biased too old.

Age both with and without the *gap* samples can be reported as an estimate of this bias. If the *gap* is small, age can be reported with the highest steps of the *mountain* excluded for a young estimate, and with a few steps beyond the *gap* for an old estimate.

For an island *type*, with negligible *background*, the age calculation comes out almost the same with and without the *gap*. Age calculation is still approximate because of the possibility that a foreign clade is hiding in the island with similar definition markers, as discussed above and again below.

Even age calculation for a known haplogroup can have bias to younger age if data is not restricted to only those samples with SNP results. Databases such as Ysearch and FTDNA include "predicted" haplogroup assignments based on proprietary STR methods. STR outliers without SNP data cannot be "predicted" with high probability, so these are "predicted" into the stem haplogroup, artificially decreasing outliers in the downstream haplogroup, thereby reducing ASD age in that haplogroup.

Methods: Mutation Rates

Mutation rates for each of the standard 37 markers are provided by Chandler (2006), calibrated to father-son pairs. Chandler's methods have been applied to 30 additional markers (from the FTDNA 67-marker set). Both the Thomas rates and the Chandler rates are incorporated in the "ASD" sheet mentioned above. These rates can be easily modified in my "Type.xls" master file as better rates become available in the future. Age is inversely proportional to mutation rate.

Calculated ASD age has a statistical uncertainty due to sample size. Even for very large sample sizes with small confidence interval, ASD age systematically comes out too young due to effects called "population dynamics" or "population structure." Population structure is traditionally treated by a "factor." The factor is applied to mutation rates producing "effective mutation rates" smaller than father-son rates, explained for example by Goldstein (1995), based on Moran (1975). There is a recent brief review of the subject of effective mutation rates by Athey (2007b). Literature recommendations for the effective population factor range to smaller than 1/4. Zhivotovsky (2006) nicely demonstrates the stochastic reduction of effective mutation rate for small populations with computer simulations. Zhivotovsky includes references to prior literature. Zhivotovsky found a population factor of 1/3.6 for simulated small haplogroups within a larger population without net population growth. He demonstrates that the factor is larger for populations with growth, and larger for larger populations, the stochastic limit being no reduction (factor = 1) for large growing populations, as theoretically required.

Thomas does not use this factor. The Thomas population may be a young clade that grew rapidly so does not need a factor. Actually, the Thomas 1998 average rate is 1.23 times the average of the Chandler rates for the five Thomas markers, but 1.23 is really an insignificant factor.

Nordtvedt (2008b) provides another method of statistical correction, with weighting factors for ASD age calculations for the TMRCA between two clusters. Briefly, the Nordtvedt correction accounts for the a priori probability that the founding haplotypes were in fact closer to the haplogroup modal haplotype than to the modal haplotypes calculated from the data. The Nordtvedt factors have relatively smaller effect for clusters with modal haplotypes relatively far from the haplogroup modal haplotype, and produce relatively older ages for clusters relatively closer to the haplogroup modal haplotype. In other words, a cluster with an unusual modal haplotype is evidence of population expansion or migration, and therefore probably younger, so less correction is required to the raw ASD. The Nordtvedt correction can be applied to a single cluster. The Nordtvedt correction should not be applied in addition to an effective mutation rate factor, because the two corrections are not independent.

For a Nordtvedt correction to regional data, the regional haplogroup modal haplotype, if known, should be used for a regional population expansion. For a population that migrated from a known foreign region, the foreign region haplogroup modal haplotype should be used. For a population that migrated from an unknown foreign region, with an unusual modal haplotype, the Nordtvedt correction from comparison to the known haplogroup modal haplotype provides little correction to the age calculated from the raw ASD. Raw ASD age is younger than an age adjusted with a traditional mutation rate factor.

Nordtvedt (2008a) also estimates multiple "down-weighting" factors by marker to account for population structure.

Technical comments: ASD divided by rate provides a good estimate of clade age, but this method is not exactly correct because the underlying stepwise mutation model (SMM) is not exactly correct for a number of reasons. The Chandler rates are for total mutations, equivalent to the infinite alleles model, but rare single mutations of more than one step are possible and are not rare for compound markers with recLOH. Mutation rate is generally asymmetric and varies with STR value, as demonstrated, for example, by Whittaker (2003, Figure 3). ASD age can be incorrect for small ASD (ASD about <3), as demonstrated for example by Campbell (2007) and Watkins (2007). However, according to Campbell's Figure 2, the error is less than 20% for ASD > 0.18. These technical issues tend to cause ASD age to be older, compensating in part for the stronger tendency toward younger ASD age due to population structure.

Methods: Subtypes for ASD Age

By "raw ASD age" I mean the age calculated from ASD without a correction (factor = 1), either for a single marker or averaged ASD for multiple markers. Age = ASD divided by father-son mutation rate, for example the Chandler (2006) rates. This analysis also applies to age based on infinite alleles - equivalent to ASD for markers with low mutation fraction.

In the following three paragraphs I present hypothetical examples as an analysis, to justify the consideration of subtypes for age estimation based on ASD. This discussion applies to large, young clades with strong founder effects, such as rapid population expansion, with or without migration.

1. If a cluster corresponds to a clade that was produced by a single recent vigorous population expansion from one founding individual, the raw ASD age should be used.

2. Suppose a cluster-clade was produced by a single recent vigorous population expansion from two individuals in the same haplogroup that differed at only a few markers. These may be the only two from a previous subgroup population who participated in the population expansion. Alternatively, in regional data, these may be the only two who migrated into the region from far away before the population expansion. A population correction factor as explained in the previous section can be used. The factor is unknown, and highly sensitive to how closely those two founding members are related. The corrected older age is the TMRCA. In this example, those markers that differed in the founders have a bimodal distribution, with older raw ASD age than the markers that were common in the founders. With enough data, each of the common founder markers provide the same raw ASD age, which is the time of the population expansion. With limited data, all the markers that do not have bimodal distributions can be averaged for the best estimate of the time of the population expansion. The markers with older ASD age are correlated, so the data can be split into two *mountains* in haploSPACE, although there may be significant overlap (a high "*mountain pass*" gap) if there are only a few rapidly mutating markers that differ in the founders. This may seem like a silly example, because we cannot know a priori that there were exactly two founders, and because with limited data ASD age will vary widely by marker just due to sampling statistics. The important point to be made with this simple example: TMRCA is likely older than average raw ASD age, but time of the single population expansion is likely less than the average raw ASD age because the ASD includes more variation at those markers that differ in the two founders.

3. More likely, a population expansion progresses in two or more stages separated in time, or less vigorously

throughout a range of time. More likely, there were more than two founders in a population expansion of a clade now being analyzed as a hypothetical *type*. Founders may have lived or immigrated at different times. The analysis is more complex. The analysis, however, is similar to the previous paragraph. TMRCA may well be quite a bit older than the raw ASD age. The approximate time of the population expansion, however, may well be younger than the raw ASD age. Subtypes may have different ASD ages.

Breaking a *type* into subtypes is conceptually the same as figuring out the population structure. In this light, it seems careless to only calculate TMRCA using a correction factor for all mutation rates and averaging ASD for all markers. A simple examination of the raw ASD age by marker may provide two valuable additional pieces of information. First, outlier older markers, particularly those with obvious bimodal distributions, provide justification for the population structure correction factor. Second, ASD age with old outlier markers removed provides an estimate for the time of a significant population expansion. If ASD age of subtypes are similar, that is important evidence that the population expansion involved a tribe of relatives; if ASD age of subtypes are not the same, that is evidence of multiple expansions or multiple immigrations.

As mentioned in the **Mathematical Models** section, it would be nice to have an automatic objective statistical measure that shows which markers exhibit population structure and which do not. Lacking that, a simple glance at the columns of data can be helpful. Without population structure a marker with few mutations should have a graph of values that looks like a tent, and a marker with many mutations should have a graph of values that looks like a bell. Some markers may have a bias toward step up or toward step down mutations, distorting the graph, but not look like a subtype.

In addition to systematic biases due to population structure, raw ASD age has statistical sampling uncertainty, which is very tedious to calculate. For example, if a set of 50 samples has three samples with one particular marker mutated by one step, the 70% confidence interval on that mutation count of 3 is 1.3 to 6.0, while the 95% confidence interval is 0.6 to 8.8! Averaging many markers helps to reduce the confidence interval. Calculation of confidence interval for average ASD age is beyond the scope of this article. A simple, incomplete rough estimate is: the statistical confidence interval for the total number of mutated values involved in the averaged ASD. The "ASD" sheet automatically provides the count of the number of mutated values from the modal value for each marker, and total for all or for selected marker sets.

Although 67 markers may be required in a search for markers to provide a good definition of a *mountain*

type, some *types* are well defined with only 25 markers available. In such cases it makes sense to include the data with the fewer markers, providing better statistics for ASD. It would make sense to weight those markers that have more data, in figuring the average, but the "ASD" sheet does not support such weighting. If the 25-marker data have significantly more samples in the *gap* than the 67-marker data, it is not wise to combine the data, because the range of ages (*mountain* only vs *mountain* plus *gap*) will be wider in the 25-marker data in such case.

Another kind of age uncertainty concerns self-consistency of *types*. It is helpful to estimate the sensitivity of ASD age to selection of modal haplotype and *gap* for the *type*, as discussed above. This uncertainty can be quantified by varying the parameters, as discussed above, and noting the variation produced in the raw ASD.

A *subtype* cannot be identified with high confidence just because one particular marker has a much older age. Even with a statistical analysis showing high confidence that the one marker has an older age beyond expectation, it is difficult to rule out a false positive bias, as discussed above. As is the case for *types*, correlation of multiple markers, with a low *mountain gap*, is required for confident identification of *subtypes*. But there is an important distinction: *types* should be assumed to be invalid unless there is reasonable validation. For the purpose of quantifying the time of population expansion of a *type*, however, all subtypes should be considered even with only weak evidence. That is to say, any marker with unusually high ASD age, or any marker with a bimodal distribution, should be suspected of harboring two or more subtypes.

Hypothetical population expansion time can be reported two ways. The raw average ASD age is an old upper limit. The average ASD age with all suspicious markers excluded is a young lower limit estimate for the time of population expansion.

TMRCA can also be reported two ways. The raw average ASD age is a young lower limit. ASD age using a correction factor from 1 to 1/4 for the father-son mutation rate is the traditional way, providing an older but more reasonable TRMCS. Although the subject of correction factor is beyond the scope of this article, I point out here that examination of the distribution of STR values for markers with older raw ASD age provides a justification for the use of correction factors. If a few markers stand out with roughly the same age, older than all other markers, that may mean those markers are the ones that differed in the tribe of founders before a population expansion or migration, and an older age with only those markers and without a correction factor may be a good estimate of TMRCA.

Rapidly mutating markers are generally valuable for determining age of young *types* because more mutations mean better sampling statistics.

As an extreme example, consider “A type” in the companion article. A type is very young, certainly less than 2,000 years old and perhaps much younger than that. However, DYS459b comes out 15,300 years old. DYS459b is bimodal, with near 50-50 split between the values of 10 and 11, and no other values. DYS459 is slowly mutating, yielding an old raw ASD age (TMRCA) when 50% mutated. The data is insufficient to consider this marker as proof of two subtypes. It could be a statistical fluke. It could be that DYS459b has an unusual rate only in this population. The simplest interpretation: the values 10 vs 11 represent two *subtypes* that predate the population expansion, or split due to a mutation early in the expansion. Whatever the reason, it certainly demonstrates a caveat when averaging ASD from many markers.

The previous paragraph seems to recommend removal of old markers when calculating ASD age. In general, this is not fair. The ASD method takes care of the statistics. Rapidly growing young clades theoretically have no population structure (factor = 1). Zhivotovsky (2006) demonstrates this nicely with simulations. Averaged over many clades (or over many *types*, or over many simulations) the clades that come out too old are balanced by the clades that come out too young. Athey (2007b, also private communications, also this issue, page 131) points out that mutations early in a family tree can produce excess mutation counts. The previous paragraph may be an extreme example of this effect, if the binomial distribution at that one marker is not due to a founder but due to a mutation that occurred early in the population expansion. Briefly, in the statistics of rapidly growing young clades (male line families): (1) many clades have fewer than typical mutations early in the population expansion so have slightly lower raw ASD age, (2) many clades have a bit more than typical mutations early in the population expansion so have slightly higher raw ASD age, and (3) a few clades have a mutation in a slowly mutating marker very early in the population expansion, so end up with a bimodal distribution at that one marker (rarely at a few markers), so have a significantly higher raw ASD age. This statistics is not easy to grasp, but full understanding is not necessary: If you are analyzing a hypothetical young clade (or calculating the ASD for your family project), calculating the TMRCA of the *type* (or unknown time of the patriarch of your family data set), take the time to look at the columns of data. If one column for one marker is obviously bimodal, that marker may well represent a mutation that occurred in the earliest generations of that clade. If the marker has a low mutation rate, it will significantly increase the TMRCA. It provides you no satisfaction that many other clades have no mutations in

the early generations, and thereby slightly younger TMRCA, averaging out over all clades. With reasonable care, removal of a marker with an obvious bimodal distribution seems justified.

Conclusions

I offer this *mountain* in haploSPACE method as a quantitative measure of the validity of Y-STR *types*. If others publish the *Arabian* for their best STR candidates, we will soon have a relative target for how good a *type* should be. This method is applicable to haplogroups in retrospect.

TMRCA of a *mountain type* is probably older than the raw ASD age due to population structure. As usual for ASD calculation of TMRCA, population structure factors need to be used, perhaps a factor of 1/4, compared to father-son mutation rates. *types* and *subtypes* provide justification for such factors.

The time of population expansion for a *type* may well be younger than the raw ASD age. A rough estimate for hypothetical time of population expansion can be had by removing from the ASD calculation the STR markers that are bimodal or otherwise obviously not in a tent or bell distribution.

Supplementary Data

The [Supplementary Data](#) file has an index with links to other on-line files. Those at the JoGG web site do not change, but support the article at the time of its publication. The directory includes tools, data, analysis, and detailed results for this article, and for the companion article. For similar information that is updated, see the author's web site:

<http://www.gwozdz.org/PolishCladesUpdate>.

Web Resources

Whit Athey's Haplogroup Predictor
<http://www.hprg.com/hapest5>

Polish DNA Project
<http://www.familytreedna.com/public/polish/>

YHRD Y-STR Haplotype Reference Database
<http://www.yhrd.org>

Ysearch Y-STR Database
<http://www.ysearch.org>

References

Athey W (2005) Haplogroup prediction from Y-STR values using an allele frequency approach. *J Genet Geneal*, 1:1-7.

- Athey W (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *J Genet Geneal*, 2:32-39.
- Athey W (2007a) A major subclade of Haplogroup G2. *J Genet Geneal*, 3:14-18.
- Athey W (2007b) Mutation rates - who's got the right values? *J Genet Geneal*, 3(2):i-iii.
- Campbell J (2007) The SMM model as a boundary value problem using the discrete diffusion equation. *Theor Popul Biol*, 72:539-546.
- Chandler JF (2006) Estimating per-locus mutation rates. *J Genet Geneal*, 2:27-33. See also [on-line extension to 67 markers](#).
- Cruciani F, et al, 19 authors (2004) Phylogeographic analysis of Haplogroup E3b (E-215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet*, 74:1014-1022.
- Cruciani F, et al, 5 authors (2006) Molecular dissection of the Y chromosome Haplogroup E-78 (E3b1a): a *posteriori* evaluation of a microsatellite - network - based approach through six new biallelic markers. *Human Mutation in Brief Online* #916:1-10.
- Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139:463-471.
- Gwozdz P (2009) Y-STR Mountains in HaploSpace, Part II: application to common Polish clades. *J Genet Geneal*, 5:159-185.
- Mayka L (2007) Web discussion on the objective definition of a cluster.
- Mayka L (2008) Private email communication. Mayka is administrator of the Polish Project.
- Moran P (1975) Wandering distributions and the electrophoretic profile. *Theor Popul Biol*, 8:318-330.
- Nordtvedt K, Cullen J (2008a) Extended haplotype estimation of clade TMRCA's confirmation by computer simulation.
- Nordtvedt K (2008b) More Realistic TMRCA Calculations. *J Genet Geneal*, 4:96-103. see also <http://knordtvedt.home.bresnan.net/>
- Thomas MG, Skorecki K, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. *Nature*, 394:138-140.
- Vincent MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 324:1213-1216.
- Watkins JC (2007) Microsatellite evolution: Markov transition functions for a suite of models. *Theor Popul Biol*, 71:147-159.
- Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibby RM (2003) Likelihood-Based Estimation of Microsatellite Mutation Rates. *Genetics*, 164:781-787.
- Willuweit S, Roewer L, on behalf of the International Forensic Y Chromosome User Group (2007) Y chromosome haplotype reference database (Yhrd): Update. *Foren Sci Int: Genetics* 1:83-87 (<http://dx.doi.org/10.1016/j.fsigen.2007.01.017>)
- Zhivotovskiy L, Underhill P, Feldman W (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol Biol Evol*, 23:2268-2270.

Summary: Mountain (Type) Method

Step Frequency is the number of samples in a database that differ from a proposed modal haplotype by a step number of mutations. Mutation steps are counted following the method used by Ysearch. A *type* is a hypothetical clade that forms a cluster with a step frequency curve that looks like mountain. Figure 1 is an example.

A *mountain pass* is the step number at the minimum in the step frequency curve just beyond the *mountain*. A *gap* is the number of steps in a span of continuous step values including the pass, where the step frequency is low, as defined below. A *gap* may be a single step for a steep *mountain* pass, or it may be two or more steps. Figure 1 has a *gap* of three. The *cutoff* is the lowest step number of the *gap*, which is also the number of steps in the *mountain*. The *cutoff* is not part of the *mountain*. The cluster and the modal haplotype for the *type* are defined by the *mountain*, those samples less than the *cutoff*. The *mountain* number is the total number of samples in the *mountain*. The *gap number* is the number of samples in the *gap* (including the pass, and only the pass if the *gap* equals one).

The *background* refers to foreign samples that do not belong to the *type* (hypothetical clade) but fall within the *mountain*. The *gap background* refers to samples that do not belong to the *type* and fall within the *gap*. The *type outliers* refers to samples that belong to the *type* but fall beyond the *mountain*.

As a rough approximation, explained above, *type outliers* are estimated as half the *gap* number, and *background* is estimated as the average step frequency in the *gap*. So the *background* is the *gap number* divided by the *gap*. This method very likely overstates the background, in order to compensate for the unknown systematic statistical errors. The background divided by the *mountain* number is the *background fraction*, which is expressed as background percent. Background percent cannot be applied to a particular step, and cannot be applied as a probability to individual samples, because much of the background is expected in the last step of the *mountain*, much of the remainder is expected in the previous step, and after that, much of each corresponding remainder is expected in the corresponding previous step, with very little background expected at step zero.

The *statistical mountain number* is the minimum of the *mountain number* confidence interval. The *statistical gap number* is the maximum of the *gap number* confidence interval. From the background assumptions above, the *statistical background* is the *statistical gap number* divided by the *gap*. I propose 70% Poisson statistics applied individually to each number as a standard for publication and comparison.

The *statistical background percent* (SBP) is the *statistical background* divided by the *statistical mountain number*.

SBP is proposed as a simple measure of validity, on the premise that more isolated *types*, with low SBP, are more likely to represent clades. SBP combines the additional premise that larger data samples are better; databases with more data, and *types* with a higher percent in a database, provide narrower confidence intervals and hence lower SBP, but even a *type* with a few samples should be considered if it is very well isolated as evidenced by a low SBP.

The two confidence intervals are combined as worst case in order to compensate for selection bias and other systematic statistical errors. SBP should not be called a 70% sampling confidence worst case, even though it was designed to give the impression of overall 70% worst case confidence including systematic statistical errors that cannot be measured and including a consideration for small foreign clades.

SBP often comes out greater than 100%, which is a psychological discouragement to publication. A *type* with SBP less than 50% is worthy of monitoring as data accumulates, in my experience, and a *type* below 25% is rare enough to be worthy of publication. An island is a *type* with 5% or lower SBP.

Statisticians may, of course, also provide more rigorous statistical calculations. Confidence intervals other than 70% could be specified.

The definition of a *type* is the modal haplotype, *cutoff*, and *gap* that provides the smallest SBP. Exception is when a *cutoff*/*gap* pair that provide a minimum SBP is not the best pair to human judgment, for example involving the last step of the *mountain*. Markers chosen by automatic ranking for minimum SBP provide an objective and credible modal haplotype, but it is not necessary to restrict marker choice to automatic ranking. Specific markers with obvious problems, for example recLOH, can be selectively omitted from the definition. Obviously, it is not fair to cherry pick those markers that do not rank well but just happen to have no mutations

at the *gap*, or to remove markers that rank well but just happen to have mutations at the *gap*. There is selection bias, and such bias is compensated by the way SBP is defined as worst case. The person publishing a *type* with SBP evaluation is free to choose the definition, as long as SBP is objectively calculated, and as long as a comment is included for excluded markers and for non-minimum SBP. Large *types* comparable to the size of the haplogroup are misleading because the "down slope" on the far side of the haplogroup *mountain* provides a false low SBP.

SBP is not the probability that the *type* is not a clade. SBP is a high estimate of the contingent probability that samples in the *mountain* do not belong to the clade for statistical reasons, even if the *type* really does represent a clade.

The size of a *type* is the *mountain number*, minus the background, plus the *type outliers*. The size confidence interval is the *mountain number* confidence interval.

A *signature* of a *type* is the modal haplotype using only a few markers from the definition that best separate the corresponding data cluster from a database. Usually but not always, a signature value differs from the modal haplotype value of the stem haplogroup at that marker. A signature may or may not actually produce exactly the same *mountain* as the definition. It is a good idea to say so if it does, since that is also a quality of a good *type*.

The *breadth* of a *type* is the set of continuous number of markers that can be used for a modal haplotype to produce the same *mountain* cluster as the definition modal haplotype. The breadth is expressed as the smallest and largest number of markers. The breadth includes the definition. For example the *mountain* cluster for Figure 1 has a breadth of 3 to 51. This means not just the same *mountain number*, but the same samples less than a *cutoff*, better than 90%. Allow <10% outliers, which are additional samples in the *mountain* plus samples missing from the *mountain*. (Allow one outlier for 11-20 *mountain number*, two outliers for 21-30, three for 31-40, etc.) The *cutoff* value is not necessarily the same for each modal haplotype in the breadth. The step frequency vs step below the *cutoff* need not be identical for each modal haplotype in the breadth. File "Type.xls" has a cell where the user can quickly overtype the number of markers, and immediately observe the *mountain* data for various number of markers. The breadth is a measure of quality of a *type*; the wider the better. Usually, but not always, a *type* with a wide breadth has a low SBP.

Mathematical Summary

$$B_{SP} = \frac{N_{SG}}{g N_{SM}}$$

$n(s)$ = step frequency vs step

$N_M = \sum n(s) [s = 0 \text{ to } (c - 1)]$

$N_G = \sum n(s) [s = c \text{ to } (c + g - 1)]$

c and g selected for minimum B_{SP}

B_{SP} = **SBP** = Statistical Background Percent = proposed measure of quality of a type
 = percent of the cluster that does not belong to the hypothetical clade
 adjusted (increased) for statistics, sampling, selection bias, etc.
 a type with a smaller SBP is more likely to represent a clade

N_{SG} = statistically adjusted (increased) value of N_G ; 70% Poisson statistics
 = the higher value corresponding to N_G from Table 2
 N_G = Number of samples in the gap = gap number

N_{SM} = statistically adjusted (decreased) value of N_M ; 70% Poisson statistics
 = the lower value corresponding to N_M from Table 2
 N_M = Number of samples in the mountain = mountain number

c = cutoff

g = gap

$$S = N_M - N_G / g + N_G / 2$$

$$S_{Ci} = N_{Mci} - N_G / g + N_G / 2$$

S = Size of the type (hypothetical clade)

S_{Ci} = Size confidence interval

N_{Mci} = Mountain number confidence interval; 70% Poisson statistics
 = the interval of values corresponding to N_M from Table 2