Forensic Science International: Genetics 8 (2014) 233–243

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensic Population Genetics - Original Research

Understanding Y haplotype matching probability

Charles H. Brenner [a,b,*]

[a] Human Rights Center, U.C. Berkeley, Berkeley, CA, United States
[b] DNA·VIEW, 6801 Thornhill Drive, Oakland, CA 94611-1336, United States

ARTICLE INFO

Article history:
Received 21 March 2013
Received in revised form 6 October 2013
Accepted 19 October 2013

Keywords:
Haplotype
Y-haplotype
Likelihood ratio
Weight of evidence calculation
Probability
Model

ABSTRACT

The Y haplotype population-genetic terrain is better explored from a fresh perspective rather than by analogy with the more familiar autosomal ideas. For haplotype matching probabilities, versus for autosomal matching probabilities, explicit attention to modelling – such as how evolution got us where we are – is much more important while consideration of population frequency is much less so. This paper explores, extends, and explains some of the concepts of "Fundamental problem of forensic mathematics – the evidential strength of a rare haplotype match" [1]. That earlier paper presented and validated a "kappa method" formula for the evidential strength when a suspect matches a previously unseen haplotype (such as a Y-haplotype) at the crime scene. Mathematical implications of the kappa method are intuitive and reasonable. Suspicions to the contrary raised in [2] rest on elementary errors.

Critical to deriving the kappa method or any sensible evidential calculation is understanding that thinking about haplotype population frequency is a red herring; the pivotal question is one of matching probability. But confusion between the two is unfortunately institutionalized in much of the forensic world. Examples make clear why (matching) probability is not (population) frequency and why uncertainty intervals on matching probabilities are merely confused thinking. Forensic matching calculations should be based on a model, on stipulated premises. The model inevitably only approximates reality, and any error in the results comes only from error in the model, the inexactness of the approximation. Sampling variation does not measure that inexactness and hence is not helpful in explaining evidence and is in fact an impediment.

Alternative haplotype matching probability approaches that various authors have considered are reviewed. Some are based on no model and cannot be taken seriously. For the others, some evaluation of the models is discussed. Recent evidence supports the adequacy of the simple exchangability model on which the kappa method rests. However, to make progress toward forensic calculation of Y haplotype mixture evidence a different tack is needed. The "Laplace distribution" model of Andersen et al. [3] which estimates haplotype frequencies by identifying haplotype clusters in population data looks useful.

1. Introduction – understanding Y haplotypes

For understanding the forensic use of Y-haplotype evidence, rather than adapt the methods and habits that have evolved for the analysis of autosomal DNA evidence it is more appropriate and productive to start over from the beginning. Evidence is quantified by a likelihood ratio built from the probability for a coincidental match by an innocent suspect; that fact remains. All else is up for grabs.

The genetic rules for the Y haplotype are different in several ways from the autosomal rules and these differences have population genetic consequences, which in turn affect matching probabilities. The most important genetic difference is of course the lack of recombination. One obvious consequence everyone knows: a matching probability can't be obtained by multiplication across loci. The haplotype must be treated as a unit. A Y-haplotype thus seems analogous to an autosomal allele, but we will see that copying the treatment of autosomal loci would be to fall into a trap, to err in several respects. For a start, the idea that sample frequency approximates population frequency approximates matching probability is too careless when most haplotypes in the population are completely unrepresented in the sample, as is the usual case with Y-haplotypes composed of multiple STR loci. Another habit from autosomal practice that doesn't apply sensibly to Y-haplotypes is the treatment of θ – the Cockerham/NRC II [4,5] notation for allele sharing by common descent. The unexpected reason why it is not is explained in Section 2.4.

In the autosomal case, a simple model considers only allele probabilities (sometimes carelessly called "frequencies" – see Section 4) and taking θ into account is a refinement whose

* Correspondence to: 6801 Thornhill Drive, Oakland, CA 94611, United States.
E-mail address: cbrenner@berkeley.edu

1872-4973/$ – see front matter © 2013 Published by Elsevier Ireland Ltd.
http://dx.doi.org/10.1016/j.fsigen.2013.10.007

introduction adds a bit of accuracy by acknowledging that identical alleles are occasionally identical by descent (IBD). The Y-haplotype situation is the opposite: Identical alleles are nearly always IBD. Hence θ comes first and anything else is the minor refinement. Once we know θ we are close to the matching probability.

Terminology: Two or more haplotypes are considered "identical" if they are identical as far as determined, i.e. concerning Yfiler haplotypes it means having the same repeat number(s) at each STR locus. Haplotypes are "identical by descent" (IBD) if they are identical and all the haplotypes along the (necessarily patrilineal) paths connecting them to their common ancestor are also identical. "Identical by state" (IBS) as used here is synonymous with "identical", an umbrella meaning in that IBS thus includes IBD as a subset.[1] Adopting the umbrella definition for IBS means some other term may be needed to mean IBS but not IBD and for this purpose I use the word "strict." Identity strictly by state is also "convergence".

The expression "unrelated man" is common in forensic practice; a good enough approximation sometimes. In the Y haplotype world, always remember that everyone is related for that is nearly the only reason men match.

The historical development of autosomal calculation methods has been unsystematic, often intuitive, and to the extent that there are underlying models, which concepts have been incorporated (such as random mating, subpopulations, mutation) has of course been motivated by relevance to the autosomal situation. The appropriate and important concepts for a model useful for haplotype analysis may be different ones. It makes sense to begin by exploring the population genetics of haplotypes in order to gain an understanding then to consider models and methods of calculation.

2. Exploring Y haplotype populations

Before discussing a few approaches, good and bad, that have been suggested for Y haplotype matching probabilities, we will delve into the nature of Y haplotype populations in order to have some background that will be a helpful context for evaluating the approaches.

I see three ways to explore: data, theory, and simulations.

2.1. Exploring Y-haplotype data

Population samples (or "reference databases") for the 17-locus Yfiler™ Y haplotypes for $n > 1000$ men are conveniently available on the Internet [6] for several different populations. Examples, simulations, data and analysis mentioned in this paper assume Yfiler haplotypes unless otherwise specified.

Examining the data, a few points are quickly obvious:

(a) The vast majority of types that occur at all in a population sample occur exactly once.[2] I use the symbol κ for this singleton proportion. κ gradually becomes smaller as the number of collected reference haplotypes increases and is larger when haplotypes include more loci and hence are more polymorphic. $\kappa > 0.8$ for reference databases considered in this paper.

No type occurs many times. For example, among the 4102 Caucasian full profiles, $\kappa = 84\%$ are singletons (once occurring), and 98% of the men in the sample have a type shared with at most 5 men. Implications:

(i) The crime scene type will usually not be found in the population sample.
(ii) Therefore this usual situation is the most important and most worthy of our attention.
(iii) Sample frequency is a poor estimate for population frequency. Among the haplotypes that are not represented, the sample frequency of 0 obviously underestimates the true frequency. Further, in order to compensate for the large fraction of the population that is under-represented in the sample, those sample frequencies that are not 0 must on average greatly overestimate the true frequency.
(iv) Point (a)iii may seem paradoxical, seemingly contradicting the intuition that sample frequency is an unbiased estimate for frequency. That intuition is correct in that averaging sample frequencies for a particular haplotype T over many repeated samplings is an unbiased estimate for population frequency. For a typical rare haplotype, repeated samplings give a sample frequency that is usually 0 but occasionally $1/n$ (where n = sample size), rarely more, with expected value equal to the population frequency – no bias. But the situation at hand is quite different; we fix our attention on a single sample and a single, necessarily non-zero, sample frequency such as $1/n$ (the common situation for casework), and consider the many haplotypes in the sample with that sample frequency. For example consider the set of population frequencies of the once-observed haplotypes in a database of size $n = 1000$. Not only are those frequencies well under 1/1000 on average, only a relative handful of individual frequencies exceed 1/1000.

(b) The probability that two randomly selected men have the same type is small – about 1/8800 among Caucasians, 1/3300 among Chinese, 1/13000 among US Blacks [6]. These numbers are calculated simply by comparing every pair of men in a database. Note that this calculation from pairwise comparisons is another way, different from sample frequency, for using the sample database to come up with matching probability. Implication:
(i) The average haplotype population frequency among observed Caucasians is 1/8800. For those haplotypes observed only once in the sample, the frequency must be even less. Therefore the singleton sample frequency of 1/4102 must be an overestimate by much more than two-fold. In fact, as is shown in [1], the actual amount of overestimate is by $1/(1-84\,\%)$ or about a factor of 6.

(c) Also interesting if not so obvious, using a test devised by Slatkin [7]: Comparing the population sample with expectation under the Kimura et al. model of infinitely many selectively neutral alleles generally shows very small p-values – i.e. $p < 0.01$ for the Caucasian and other large datasets.[3]

(d) Various published data is available for the mutation rates at the various Y STR loci, and not surprisingly they are similar to autosomal STR rates: STR mutation rates average 1/350 per locus, so over the full 17 loci $\mu = 17/350$, or 5%, per meiosis.

2.2. Theoretical explorations

The main focus of this and the following section is exploring and understanding the claim in Section 1 that the story of rare haplotype matching is essentially the story of common descent.

[1] Some writers use IBS differently – as an alternative or partially an alternative to IBD – but the umbrella usage looks to me more common especially among careful writers.

[2] for presently available sample sizes. If – somewhat unrealistically – sample size n were increased without increasing the number of loci, then to be sure eventually $\kappa < 1/2$ (for about $n > 10930$ according to [1]).

[3] However a few of the datasets show non-significant p-values such as three samples of size $|\mathcal{D}| \approx 300$ from Malaysia in Table 2 of [1], for which $p = 0.08$, 0.46, and 0.66.

As a simplified but helpful model of a forensic Y-haplotype: Represent the data at each locus as an integer repeat amount; model a compound locus as if two loci; mutation occurs by single steps plus or minus at a locus and the mutation rate is the same 1/350 per meiosis at every locus. Under this model, a Yfiler haplotype is an ordered set of 17 integers, which is to say a point in a 17-dimensional integer lattice, and the evolutionary course of a haplotype as it is transmitted through generations with possible mutation is a 17-dimensional random walk. The mutation rate of $\mu = 1/20$ per generation means that father and son carry the same haplotype with probability 0.95. If two contemporary men trace back a combined $2g$ generations to their most recent common patrilineal ancestor (MRCA), then with probability 0.95^{2g} no mutation occurred at any of the $2g$ meotic events between them and their Yfiler haplotypes are IBD. For example, if the MRCA lived 100, 1000, or 10000 years ago then $g \approx 4$, 40, or 400 and the IBD probabilities are 70%, 1/40 and $1/10^{16}$ (!), respectively. If they represent populations that have truly been separate for 10000 years, there is no chance for them to be identical by descent.

What is the additional probability that two such men's haplotypes are identical and not IBD? Identity strictly by state arises through convergent evolution, i.e. through multiple mutations whose net effect is to cancel one another out. The probability of convergence is easily seen to be very small. Since the mutation model is the same whether time goes forward or backward, the trail of haplotypes connecting two men is a random walk in the high-dimensional lattice, and convergence corresponds to a random walk returning to its starting point. That's an unlikely event when the number of dimensions is large [8].

An example of a 3-dimensional integer lattice is a multistory parking building with numbered rows and numbered slots within each row on each level. Now imagine 100 such buildings, arrayed in a 10 by 10 grid. That's a 5-dimensional lattice of parking slots. Suppose you park your car and while you are gone someone randomly moves it a few times by step-wise mutation – "mutating" to the same slot in an adjacent row, then to the corresponding position one floor up or one structure to the East – like the meander through a haplotype's genealogical mutational history. After even a few mutations your car is hopelessly lost. Equivalently, it is very unlikely that the car will mutate back to where you left it. Large-dimensional space is huge, cavernous and sparsely populated, and intuitions derived from one, two, or even three dimensions are a poor guide. In particular, in one or two dimensions random walks essentially always return to the start; in high dimensions rarely. In autosomal analysis each locus, such as vWA, can be considered as an independent unit and so considered identity between two alleles is only 10% or so to be IBD. (The same would be true for a single Y STR locus, but we are not considering Y loci singly.) The main chance for haplotype convergence is to have exactly two intervening mutations which cancel one another (Fig. 1). In the model there are 34 different possible mutations (gain or loss at each of 17 loci), hence even assuming that exactly two mutations occur it is only 1/34 that they cancel. Computing the

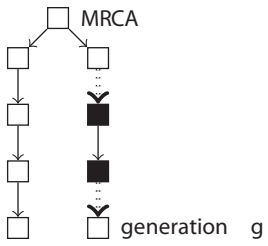

Fig. 1. If the two IBS Y haplotypes of generation g are not IBD (are strictly IBS), then the lineage path connecting them through a common ancestor (MRCA) must include at least two mutational (dotted line) events.

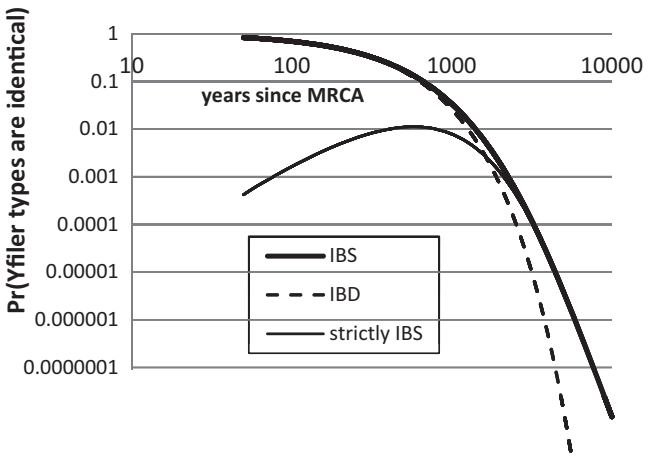

Fig. 2. Comparing probability of any match (IBS) with IBD and strictly IBS. A match strictly by state is very unlikely although when the MRCA was more than 1500 years ago, IBD is even more unlikely.

probability of 2-mutation convergence is a simple binomial exercise:

$$\begin{aligned} &Pr(\text{strictly IBS match}) \\ &\quad = Pr(\text{two haplotypes are convergent}) \\ &\quad \approx Pr(\text{2 cancelling mutations in } 2g \text{ generations}) \\ &\quad = \mu^2(1-\mu)^{2g-2}\binom{2g}{2}\tfrac{1}{34}. \end{aligned}$$

Cancelling patterns involving more than two mutations are improbable (and trickier to compute), but when considering long time spans – millenia – are, although a slight chance, virtually the only chance of Y haplotype random matching and their probability is therefore at least of evolutionary interest. Fig. 2 shows the relationship. Matching across a patrilineal separation of several centuries is nearly always because there are no intervening mutations at all, and matching across many millenia of separation is expected almost never to happen but will be after multiple mutations when it does. Indeed, comparing a recent study of 424 Chinese [9] with 10000 men in the various ABI populations showed no overlap of haplotypes. By comparison with the IBD probabilities above, the corresponding IBS probabilities increase slightly to 70%, 1/30 and $1/10^8$.

Compare Y haplotype populations with the theoretically much discussed "infinitely many neutral alleles" model of Kimura, Crow, etc. (Section 3.4). The infinite alleles model assumes that every mutation is to a new type, that there is no selection, and that the population size is fixed. From Slatkin's test [7,10] some of the assumptions don't hold very well for Y haplotypes. The first assumption is nearly true for Y haplotypes, the second is debatable, the third violated. It would be therefore be risky to assume that Y haplotype populations conform to the infinite alleles model.

2.3. Y-haplotype population simulations

Fig. 2 shows that IBD is the dominant category of IBS, especially in the situation of recent MRCA and correspondingly larger chance of IBS. But without knowing the distribution of MCRA times between pairs of haplotypes it can't be interpreted to say by how much IBD dominates strict IBS. To answer that question I generated populations of 17-locus Y haplotypes by simulation. By tagging each individual with his mutation history it is possible to count an observed rate of IBD among the pairs that are IBS. The experimental result is Pr(IBD|IBS) ≈ 33/34, and the probability would increase to the limit of 1 as the number of loci increases. This says that if two

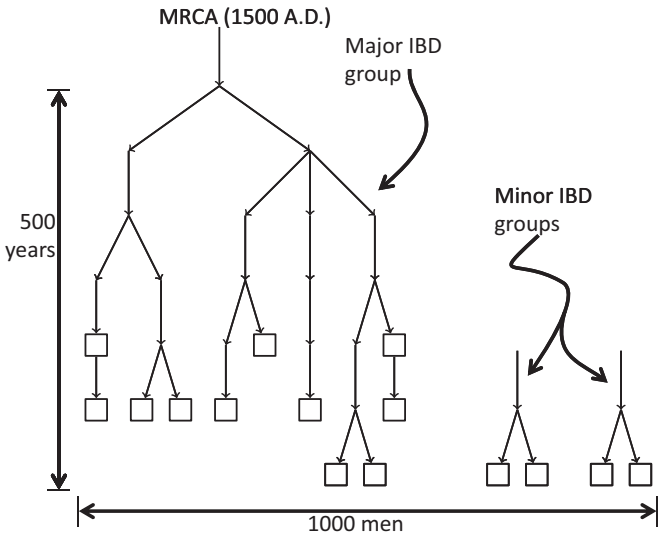

Fig. 3. Schematic representation of the IBS cohort for a typical haplotype T. A single large IBD pedigree tree accounts for 98% of all T haplotypes.

men share the same Y haplotype it is nearly always (e.g. 97%) because they are IBD – opposite to the autosomal situation in which only a few percent of matching is IBD. Let "IBD group" mean a group of (patrilineally related) men who share the same Y haplotype because it was transmitted among them without mutation. The entire collection of men with a particular haplotype T – an "IBS group" – thus consists of one or more IBD groups. The condition that IBD accounts for 97% of the TT pairs of men ensures that one of the IBD groups must dominate including over 98% of the IBS men.[4] Thus the almost invariable pattern is that almost all the men with a given type belong to one dominant IBD group. Fig. 3 illustrates the idea.

How closely are they related? As we have noted, two men 400 generations apart are never IBD, so an IBD group is much closer than 200th cousins. The chance for two random men to be IBS is about 1/8800, so their chance to be IBD is about 97% of that or 1/9000. Hence for even a medium sized population of ten million men, the dominant IBD group for a particular type is perhaps 1000 men – too many to be brothers or even close cousins. More distant cousins are far more numerous though less likely to be match. Computer simulations weighing these countervailing tendencies suggest that 10th to 30th cousin (measured via patrilineage) is roughly how related two matching men typically are. The story of a randomly selected man matching a crime scene haplotype is mostly the story of a randomly selected man being such a relative.

In summary Y-haplotype identity is overwhelmingly identity by descent, which confirms the assertion in Section 1 that the story of many-locus haplotype matching is predominately a story of θ and which thus tends to justify the simplifying modeling assumption under which the "κ method" (Section 3.6.2 and [1]) ignores haplotype structure. Consideration of haplotype structure (such as by looking at mutational neighbors; viz Section 3.3) is essentially an effort to evaluate the possibility of non-IBD matching, of convergence, but since that is so rare even a good job of estimating it would be only a minor refinement. Therefore I do not concur with the speculation in [2] that ignoring structure entails "a substantial loss of information".

[4] as you can easily convince yourself with a little mental shuffling of IBD group size proportions $d_1, d_2, \ldots, \sum d_i = 1$, realizing that $Pr(IBD|IBS) \approx \sum d_i^2$.

2.4. Thinking about θ

Classic allele matching probability for an autosomal locus [2]:

$$Pr(TT) = p_T Pr(T|T), \text{ where } Pr(T|T) = \theta + (1-\theta)\, p_T. \quad (1)$$

Here Pr(TT) means the probability that two alleles such as those of the genotype of a person are both T and p_T is supposed to be the probability that a randomly selected allele is T.

Formula (1) says that there are two ways that the second allele examined can also be T – with probability θ it is T because the two alleles are IBD; with probability $1-\theta$ they are not IBD in which case it is T with probability p_T. If TT is understood to mean the two alleles of a genotype then θ means the inbreeding coefficient, a measure of relatedness between the parents.

Can formula (1), by suitable analogy, apply to Y-haplotypes? More generally TT might be two alleles chosen by whatever rule; then θ is the probability that two alleles chosen by that rule are IBD. If the scenario is that the two "alleles" are a crime scene Y-haplotype and the Y-haplotype of a randomly selected suspect, then θ means the chance that two randomly selected men belong to the same patrilineal IBD group as in Fig. 3. Hence in the case of Y-haplotype matching the first term of (1), the θ, represents the chance of IBD matching, an interpretation proposed in [2]. But as we have discovered, the chance of IBD matching is already approximately the chance of matching. Hence the other term, $(1-\theta)p_T$, representing the chance of a non-IBD match, will be very small, dwarfed by comparison. Then what is p_T? It is something like the conditional probability for a second allele to match the first *given that they are not IBD*, but the exact meaning is even stranger and less intuitive than that. The parameter θ is defined as a population-wide average and it may be larger or smaller than a haplotype-specific matching chance. In particular if the crime scene haplotype is a type not observed in the reference database (viz Section 2.1(a)i) then it rates to be less than averagely common and θ is already larger than the probability that a randomly selected suspect matches. It follows that p_T would be negative

> Example: For the Caucasian example database of Section 2.1(a), $\theta = 1/9000$ and $\Pr(T|T) = 1/25000$ (Table 1) when the crime type T is new. So $p_T = -1/14000$ by formula (1).

and therefore is not even a probability. The analogy fails. The conclusion is that formula (1) was never an accurate formula even for autosomal alleles, but that fact has been overlooked because with forensic autosomal markers the second term dominates. This is another example of an approximation that is adequate in the autosomal arena but is not sensible when dealing with Y (or other) rare haplotypes.

3. Some haplotype matching calculation approaches

This section surveys various matching calculation methods that have been suggested for haplotype evidence. We assume here (as in [1]) that reference data exists that is suitably representative of some conceptual population of possible donors of the evidential haplotype, and that the hypotheses regarding a suspect are that he is either the donor or is in effect randomly selected from the population. This idealization is adopted not because other assumptions don't have practical importance – they do – but on the grounds of learning to walk before running.

Table 1 lists the methods discussed and gives a general sense of the range of likelihood ratios entailed, the largest of which represent the true strength of the evidence. Some of the smaller numbers may be acceptable as conservative. Through consideration of particular examples a few generalities become evident:

Table 1
Comparison of approaches assuming the Caucasian example database with $n = 4102$ and $\kappa = 0.84$ and a new haplotype.

Method		Likelihood ratio
Blind counting	(Section 3.1)	4102/0 – i.e. infinite
Brenner counting	(Section 3.1.1)	4102
C.I. from 0 (SWGDAM)	(Section 3.2)	4102/3
Frequency surveying	(Section 3.3)	~4100?
Infinite alleles model	(Section 3.4)	13000
Average matching chance	(Section 3.5)	8800
t-model	(Section 3.6.1)	23000
κ method	(Section 3.6.2)	25000
Discrete Laplace model	(Section 3.7)	~25000

Table 2
Glossary of notation.

Notation	Meaning
LR_{Ω}	Matching LR assuming model Ω
$LR_{\Omega}(T, \mathcal{D})$	Matching LR assuming Ω and using explicitly mentioned data
T	Type to match (data)
$\mathcal{D}$	Reference database augmented with the target haplotype (data)
$\lvert\mathcal{D}\rvert$	Size of $\mathcal{D}$
Ω	Population model
Ω^t, $\Omega\kappa$	Particular models
θ	Pr(*IBD*) – ([4,5] notation)
θ_{Ewens}	θ as in [11]; essentially the effective number of alleles, minus 1
κ	Proportion of singletons in $\mathcal{D}$
p	"Popularity"; number of occurrences
$\propto$	Proportional to; i.e. if $y = 2x$ then $y \propto x$

- It's important to have a model.
- It's sensible to consider the crime scene type as part of the sample database.
- A sample database can be treated as data in various ways.

I suggest that any sensible method in forensic mathematics must be grounded in a model – a presumed state of nature – and will also depend on data. Some of the approaches are based on a model. For notation used in the discussion refer to Table 2.

3.1. Blind counting method

The "counting method" traditionally means that the number of types in the population reference sample that match the crime scene target type, and the size n of the population sample, is reported as representing the matching evidence. In the interesting and usual case that $0/n$ is reported, there is at least a possible inference that the trait is infinitely rare and hence the evidence is infinitely strong against the suspect. Obviously that's not accurate or intended but it's fair to ask the reporting expert what the court is meant to conclude. If the expert doesn't have an answer then certainly the court won't either and the evidence presentation is deficient, arguably lacking probative value. If the expert does have an answer, the expert should give the answer rather than be coy.

3.1.1. Counting method (per Brenner)

[1] introduces a modified definition for "counting method" which we adopt henceforth. Begin with the premise that the appropriate question for evaluating a DNA match is the probability that an innocent (meaning randomly selected with respect to the crime scene donor) suspect would have the target type, conditional upon the crime scene occurrence. Conditioning is equivalent to mentally extending the reference database by adding the target type [1]. The consequence is that a new type is reported as occurring one time in the extended database $\mathcal{D}$. What calculation to then make from $\mathcal{D}$ remains to be decided and various possibilities will be discussed later, but of course the most obvious idea is to consider the sample frequency, $1/|\mathcal{D}|$, and it is a conservative estimate (i.e. an overestimate) of the probability that an innocent suspect will match as noted in Section 2.1(b)i.

"Overestimate the probability" means that the expected value of the sample frequency is greater than the probability in light of the data and everything we know about nature. That's not a purely mathematical truth; the "nature" component is critical. In fact maybe it's impossible to have a non-trivial correct formula without population genetic modelling assumptions – i.e. impossible to prove validity with mere mathematics and without appeal to population genetic reality (but Section 3.6.1 has the opposite speculation). As a thought experiment imagine a model called *common-type* under which nature has a very strong tendency to discourage types rarer than 25% population frequency. Under that model, if we observe five instances of a type T in a sample of size $n = 50$, it is much more likely that we have under-observed a common trait than over-observed a rare one, hence we would properly conclude that $\Pr(T) > 5/50$, i.e. the counting method would be anti-conservative. It follows that accepting even the counting method entails accepting at least some weak modeling condition, e.g. ruling out *common-type*.

In reality, because of drift and mutation, nature very much prefers rare Y STR haplotypes, like Fig. 4(b), hence the counting method is quite conservative. The counting rule likelihood ratio

$$LR_{\text{counting}} = |\mathcal{D}| \tag{2}$$

is therefore an understatement of the evidence that the suspect is the crime scene donor.

3.2. Confidence interval from zero

Something like the approach of SWGDAM recommendations [12] has been recommended by several forensic statisticians but I do not like it as it has no mathematical basis. It does not correspond to any model and it confuses concepts.

As noted above, the case of interest is a crime stain not found in the n-haplotype reference database. The analysis begins with the correct idea that the value of the evidence against a matching suspect can be expressed as the probability that a random (non-donor) person would match, conditional on the observation at the crime. Two errors in quick succession then abruptly derail the

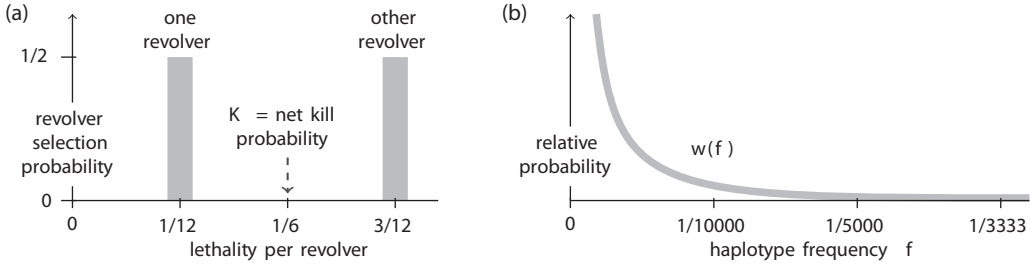

Fig. 4. Discrete and continuous frequency spectra: (a) 2-gun roulette (Section 4.4); (b) "infinite alleles" haplotype frequency distribution (Section 4.5).

analysis: failing to condition, and replacing "probability" by the different and inappropriate concept of "population frequency". The right question

> Based on the evidence that a type has been seen once, what is the probability to see it again?

thus has been twisted into the wrong question:

> What is the population frequency of a never-seen type?

An attempt is made to answer this by analogy with how a statistician might estimate the frequency of a type that has been observed several times, namely confidence intervals – an artifice at best which becomes particularly dubious in the arena of zero observations, the dreaded "confidence interval from zero." Statisticians know well that even at best a confidence interval doesn't really mean a range the frequency is likely to be in (though that of course is how it will be understood), but rather a converse of that: a range such that if the frequency is within it the actual observation is likely. In some situations the approximation of one by the other may be plausible; zero observations is not such a situation. Finally, the question arises as to what confidence interval is appropriate. Historically 95% is an arbitrary choice. It's not rooted in any principle. Possibly it has proven its mettle in a practical way in arenas like manufacturing process control, but there is no reasonable analogy, let alone logic, for how such an experience would translate to DNA, evidence, and justice. A comical debate lately arose about the right size for the one-sided interval from zero in the SWGDAM procedure. A reasonable analogy would be debating what kind of tutu to put on a dog. In truth, adopting a confidence interval from thin air is no more than a way to paper over and cover up deeper seated illogic. See Section 4.

3.3. Frequency surveying

Haplotype "frequency surveying" [13–15] is a proposal motivated by the notion that haplotypes which are near stepwise neighbors will tend to have similar frequencies, presumably because of mutation among neighbors. Hence it seeks to augment the paucity of reference observations for a particular haplotype by considering as well the richness or sparseness of nearby haplotypes. However this tempting idea doesn't hold up on close consideration. It suffers from a handful of possibly reparable shortcomings in execution, and an insuperable fundamental error.

The shortcomings begin with no explicit model. Therefore the curve-fitting approach isn't derived mathematically, but rather consists of the guesswork that it will be good enough to assume that each one-step neighbor will contribute to a haplotype's frequency as much as two two-step neighbors, three three-step neighbors, and so on. Presumably the image is that there is some kind of traffic among the haplotypes in a densely populated region of the 17-dimensional lattice space so that they replenish (or otherwise influence) one another through mutation. However, that image is tantamount to assuming a high rate of convergence through mutation (even across several mutational differences), contrary to the story described in Section 2.2. On the other hand with rare traits – which haplotypes are – genetic drift acts relatively quickly. Consequently I expect any influence from a replenishment phenomenon to be dwarfed by genetic drift. Haplotype frequencies are mostly just random. Notwithstanding an unfortunate and false endorsement in [2] that the simultaneously published paper [15] presents a validation of frequency surveying ([15] itself makes no such claim), the frequency surveying approach cannot work. [16–18,3] have further discussion. However, my impression from author Krawczak is that he now agrees the method is invalid and therefore perhaps it is a dead issue.

3.4. The infinite alleles model approach

From Ewens' celebrated [11], the model of infinitely many neutral alleles suggests a prior distribution of haplotype frequencies that amounts to $\beta(0,\ \theta_{\text{Ewens}})$ (Fig. 4(b) and formula 7). $\theta_{\text{Ewens}} \approx 8800$ for Caucasians per Section 2.1(b). Given such a prior distribution one can apply Bayes' theorem, incorporating sample data if available, to compute a posterior probability:

$$Pr(\text{random match}|\text{crime stain haplotype}) = 1/(n + \theta_{\text{Ewens}})$$

for a type previously unseen in a database of size n. This result, mentioned in [1], was the first method I discovered for the rare haplotype matching probability. I decided though that the result is too risky to recommend in forensic practice for the reasons given at the end of Section 2.2.

3.5. The average matching chance

The empirical pairwise matching experiment described in Section 2.1(b) is the same as the experiment of comparing an innocent man with a crime stain. Therefore it can fairly be used in reporting the evidence (either for a haplotype not in the database as the estimate is then conservative, or if the reference database is lost) and an appropriate statement is easy to formulate and explain:

> The suspect matches the crime stain. If the crime stain donor is not the suspect, such a match occurs only one time in 8800. Therefore either the suspect is the donor or a 1-in-8800 coincidence has occurred.

3.6. Models motivated by the data

Efron [19] makes the point that in the modern statistical world of "big data", models are sometimes implied by the data itself. Haplotype population sample data shows a prevalence of singletons, and some of the implications of this have already been mentioned Section 2.1(a). If every type in the database were unique – as would be the case if we sampled complete genomes for example – the hypothesis of maximum likelihood would be that everyone in the world is unique (and a lower bound on the order of n^2 effective types). The presence of some repeated observations in the database implies an upper bound as well as a lower on the amount of diversity in the population. I considered two approaches exploiting this idea, the "t-model" and the "κ method".

3.6.1. t-model

Suppose the population of haplotypes consisted of some large number t of equally frequent types, the model called Ω^t in [1]. $LR_{\Omega^t} = t$. An observation that κ is the proportion of singletons in a sample of size n, would be most consistent with

$$t = \frac{n}{-\ln \kappa} \tag{3}$$

as is shown in [1]. The model of equal frequencies is artificial, but it may be possible to prove that it is conservative by demonstrating that it represents a worst possible case for matching singletons. I have not proven that, but also I've not been able to construct even a contrived counterexample population. Therefore it is a possibility that (3) is conservative without any population genetic assumptions.

3.6.2. κ method

The main result of [1] is the "κ method", which refines the counting method. As in the counting method augment the

reference database to $\mathcal{D}$ by including the crime scene type in it. Let κ be the proportion of singletons in $\mathcal{D}$. The following lemma quantifies remark Section 2.1(a)i.

Lemma 1. *$\mathcal{D}$ includes the types of $1-\kappa$ of the population.*

Proof. Think of the database samples $\mathcal{D}$ accumulated one at a time. The last man added to $\mathcal{D}$ is, with probability κ, one of the singletons and hence was unrepresented before he was added. The same will be true of the next man added – probability κ that his type will be previously unrepresented. That's the same as saying that κ of the population is unrepresented in $\mathcal{D}$. Therefore the proportion of the population that is represented is $1-\kappa$. □[5]

Otherwise put, for types that do occur in $\mathcal{D}$ such as the crime scene type, the sample frequency on average over-represents the probability of being observed in a randomly selected man by an "inflation factor" $I = 1/(1-\kappa)$.

Modifying (2) by the factor I suggests the rule

$$LR_{\Omega\kappa}(1) = |\mathcal{D}|/(1-\kappa) \tag{4}$$

as the matching likelihood for a singleton and in general

$$LR_{\Omega\kappa}(p) = |\mathcal{D}|/p(1-\kappa) \tag{5}$$

for types of popularity p in the reference sample. To clarify, suppose $\mathcal{D}$ is the Caucasian population sample ($\kappa = 0.84$) so $I = 6$. Think of the partition of $\mathcal{D}$ into "popularity cohorts" – the singletons ($p = 1$) in aggregate, the doubletons ($p = 2$) in aggregate, etc.:

$$\mathcal{D} = \text{singletons} \cup \text{doubletons} \cup \cdots$$

If the inflation proportion I holds for each of the popularity cohorts individually, then (5) is correct (unbiased). However, unless we can rule out that some of the more-observed types may be legitimately common because of some special mechanism (Genghis Khan effect [21]? selection via hitchhiking?) the full generality of (5) would be hard to show and therefore is cautiously not recommended by [1]. For singletons though, (4) is well-supported by simulations. Incidentally for a rapidly growing population it may be notably conservative which implies that for at least some $p > 1$ the more general formula is anti-conservative, another reason for caution in its use at least in court. For humanitarian body identification and small p (5) seems reasonable though.

3.6.3. Doubts about κ by Buckleton, Krawczak and Weir

While not written as formal mathematics, [1] lays out arguments and results systematically – statement of problem, premises, model, derivation of results, validation of the κ method. Among various potential benefits of an explicit, linear, and deductive organization are clear communication and facilitating a substantive and reasoned discussion or argument. But the assertion "we have shown, Brenner's approach… suffers from potential anti-conservativeness in the way it inherently estimates haplotype frequencies" as part of the ending discussion in [2] is not backed up by substantitive or reasoned discussion. Inquiry to the authors elicited that the remark referred to Section 5.2 of [2]. That section mentions three ideas which are respectively pointless, confused, and wrong as follows:

- "*difficult to determine*" – After a bit of mathematical manipulation aimed at gaining insight into the performance of the κ method, Section 5.2 of [2] comes to a rather complicated but correct expression for the expected value of my formula in terms of the (unknown) population frequencies of various haplotypes. If the complicated expression could be shown by analysis to have too large or too small an expected value, that would show that my method is conservative (ok) or anti-conservative (bad). The paper comes to neither conclusion, but rather, frankly admits that "It is a complex function… therefore it is difficult to judge" based on this, sadly, dead end approach. I sympathize – my notebooks are littered ideas that didn't pan out, but I don't see any point in publishing them.
- "*Brenner demonstrated himself*" *an example where the method is anti-conservative.* For an artificial and unrealistic population consisting of 10000 equally rare haplotypes (Ω^{10000} in the notation of [1]), the κ formula would not work. But [1] doesn't claim it would; rather it explicitly points out that the modelling condition of "equal over-representation" upon which the method rests can be violated by artificially constructed populations but seems valid for populations that might occur naturally. My own pathological example is therefore not evidence against my own method.
- *elementary algebraic blunder* – [2] derives from the κ formula the result that any singleton in a database rates to be seen more than once in a second realization of the same database – a surprising and counterintuitive conclusion if correct. One could then presume that sampling a third time would include still more copies of the ever burgeoning haplotype. But how to reconcile that with my claim of validation? A possible course, when one comes up with a startling result, is to check your work. Given p for the frequency of some type, $1-(1-p)^n$ is the probability for an n-sample to include the type at all, not "more than once" as [2] alleges.[6]

3.7. Discrete Laplace model

I argue above (Section 3.3) that the structural neighborhood of a haplotype offers little information about the probability to match the haplotype. That isn't to say the structure contains no information at all.

Andersen *et al* [3] present a method called the "discrete Laplace model" based on modeling a population as a collection of subpopulations each of which is a collection of haplotypes clustered around a central (ancestral?) haplotype. The model incorporates neighborhood information in that it calculates higher probabilities for haplotypes closer to a central haplotype. Analyzing the performance on many simulated populations, the paper concludes that while both κ and Laplace haplotype probabilities are close to unbiased compared to the actual frequency, the Laplace method is more accurate. Moreover and very usefully the Laplace method is applicable even for unobserved haplotypes, a consideration that comes into play for analyzing Y mixtures.

4. Confusing probability with frequency

Matching probability is not population frequency, though confounding the two is a misconception that has long held sway in forensic genetic practice. I sometimes use the phrase "probability is not frequency" but that slogan may incite accidental misunderstanding because of its freighted and potentially ambiguous words.

Probability and frequency are related concepts to be sure. Indeed, a common (the "frequentist") definition of the probability

[5] A rigorous version of this lemma would acknowledge that κ changes slightly with the addition of a new man, and the proof appeals to Robbins' theorem [20].

[6] Through another slight confusion [2] instead writes $1-(1-p)^{n-1}$ but the difference is immaterial.

of an event is the frequency with which that event would occur in an unending series of repeated experimental trials. Even an anti-frequentist who refuses this as a definition probably accepts the intuition it represents at least for a reasonably objective instance of probability. Hájek [22]: "[F]inding out a relative frequency in a series of trials can often be the best... way of finding out the value of a probability." The "probability" at issue is matching probability. The simplest matching question concerns an atomic trait – an autosomal allele or a Y-haplotype. Such a DNA trait is found at a crime scene and we ask the probability that the corresponding DNA from a randomly selected person would match. That's a fairly objective example of probability, the sort we should be able to handle without getting mired in the Bayesian/Frequentist ideological conflict of 20th century statisticians.

Of course higher-probability events occur more frequently whether in life or in the "repeated experiment" interpretation of probability. But that abstract sense of "frequency" isn't the sense at issue, population frequency is. Population frequency is not abstract; it's an actual fact about nature, a number *but an unknown one* that exists in the real world. And the institutionalized misconception in forensics is to confuse matching probability with *population frequency*. For example consider the probability for a man to match a crime scene Y-haplotype T assuming that the man has been randomly selected from the population. Population frequency means the number of men with T in the population divided by the number of men in the population.

"So what?" some will say. "Doesn't the matching probability end up being the haplotype frequency in the population anyway? If 500 men in a population of 500000 have type T – population frequency is 1/1000 – then imagining a large number of repetitions of the evidential situation won't random suspects match 1/1000 of the time?" That sounds quite reasonable. What is wrong with it?

It would be correct if the population frequency of T were *known* to be 1/1000. Then and only then are haplotype frequency and matching probability equivalent. For probability necessarily is a summary of data – known (or stipulated) facts – and cannot possibly depend on facts that are unknown. This is an undisputable and familiar principle.

> For example we say team A has some probability to win a game coming up this weekend based on our knowledge of past encounters, information about the players' conditions, and other information.
>
> If the weekend passes and we receive no information whatever relevant to the game, then we can equally ask in retrospect "What is the probability that team A won?" and with the evidence the same, the answer will be the same. Sometimes people say there is no probability any more and either it is 0 or 1, depending on what happened, but talk like that is facetious or confused. The reality of winning changed over the weekend; the probability of winning did not.

The probability of the team to win isn't a property of the team or the game itself, but rather is a description of the particular knowledge we have about them. As the 19th century thinker JS Mill put it [23],

> Every event is in itself certain, not probable; if we knew all, we should either know positively that it will happen, or positively that it will not. But its probability to us means the degree of expectation of its occurrence, which we are warranted in entertaining by our present evidence.

Once we subscribe to the confused position that probability can depend on unknown facts, we confront the impossible question as to which unknown facts are privileged to be considered[7] and which are not. The only reasonable view when evaluating probability is therefore the same as the judiciary's principle for accepting evidence: only known facts are admissible.

Equating match probability to population frequency though nearly universal in forensic genetics, is thus a muddled view in that it says probability rests on a fact that is in practice unknown. What can we do instead? Does it matter?

4.1. What can we do instead?

The appropriate understanding of haplotype matching probability isn't how often a random match will occur to the particular haplotype T. Rather:

> The random matching probability to haplotype T is how often a random match will occur in general to a haplotype the data about which is the same as the data we have about T.

This is a more abstract formulation because it abstracts away consideration of T per se in favor of considering only *data* about T. That's progress because data is useful and valid fodder for probability. Thus the re-formulation is a question that has at least a chance of being answerable. And it has been answered to various extents by the various methods mentioned in Sections 3.1.1, 3.4, 3.5, 3.6.1, 3.6.2 and 3.7.

Note immediately one consequence: If the data we know about two different haplotypes is the same, it follows that the random match probability to each of them is the same notwithstanding that of course they may have very different population frequencies. Now, in practice we have some choice about what we regard as data since before applying mathematics to a real-world situation some simplification is necessary. That is, you have to choose a model. The κ model includes the simplifying assumption that a Y haplotype "sequence" (the repeat numbers) is merely a label, not data. Therefore the data specific to a haplotype is only the number of observations (including the reference database and the crime scene, i.e. in $\mathcal{D}$) so two singletons have the same matching probability. If you would prefer to account for the possible significance of sequence similarity among Y haplotypes, then you might adopt a model under which the data about T includes the number of observations not only of T itself, but also (for example) some function of the numbers of 1-step, 2-step, etc. neighbors of T. But the principle remains that if for another haplotype, U, the data is the same, then T and U have the same random matching probability. If there is no such U, if T is unique with respect to the data about it, the principle remains, that the matching probability for T is about the data, not about unknown population frequencies.

4.2. Does it matter?

Distinguishing probability from frequency doesn't matter so much for autosomal forensic work. The product rule covers up a lot of sins. The right astronomical number and the wrong astronomical number are not usually different in their practical evidential impact.

But for Y, as Table 1 shows, the SWGDAM idea of $4102/3 \approx 1400$ instead of $LR = 25000$ from a sound evidential evaluation means discarding a factor of 18 of evidentiary strength from a terrestrial number. And it could be worse. I've handled cases with a single 1-step Y-haplotype discrepancy, i.e. between alleged father and son; consequently $LR = 25000 \times Pr(\text{mutation}) \approx 70$. That's still pretty strong evidence, but wantonly reduce it a further 18-fold and you've got nothing.

[7] genotypes at additional untested loci?

Besides, it's a good principle to do things right. A clear and correct understanding is a good foundation for progress. Random half-wrong ideas and guesses are not, and they're also dangerous to bring to court. A good defense attorney should be able to shred an expert who talks nonsense, even if the nonsense is numerically better for the defense than the right number would be.

Schematically:

unsound : database ⇒ sample frequency
sample frequency ⇒ frequency ± interval
frequency ± interval ≈ probability ± interval
correct : data + model ⇒ probability

The unsound paradigm for matching probability amounts to: The type is very rare in the database. Therefore it is *probably* at least *pretty* rare in the population. The feeble logical conclusion is only that matching by an innocent suspect is *probably unlikely*.

A correct footing takes the larger perspective in which the database is viewed as evidence ("database as evidence" – [24]) drawn from a population that conforms to some modeling premises. Then we can say: The type is very rare in the database. That may be because the type is very rare in the population or it may be because the database is a very abnormal sample. Either way, for an innocent suspect to match requires a very unlikely circumstance, meaning the evidence is very strong that a matching suspect is not innocent.

Two examples follow of situations in which the matching probability to haplotype T has little relation to the population frequency of T. The first is a little artificial, but simpler.

4.3. Example – an island population

Thought experiment: In 14 A.D. a complete catalogue is compiled of the Y haplotypes of every man in a closed community. One particular type T happens to have a population frequency of exactly 1/100. Suppose by immigration controls and magically stopping mutation, no new types will be introduced. Fast forward 2000 years and the modern population, because of genetic drift, has completely different haplotype frequencies which are unknown – we have no new data. Just as a puzzle, what is the probability that a randomly selected man is type T?

Answer: The random man can trace his lineage back to a single man in 14 A.D., and that ancestor is 1/100 to be type T. Hence so is our man. Of course if we knew the modern population frequency the answer would be different, but based on the knowledge that we actually have the probability is 1/100. Matching probability is not population frequency; rather it is whatever inference is warranted from available evidence.

4.4. Example – two gun Russian roulette

Bored with the traditional form of the game, we invent a two-gun version of Russian roulette using two 12-chamber revolvers $\mathfrak{R}_1$, $\mathfrak{R}_2$ with respectively $f_1 = 1/12$ and $f_2 = 3/12$ of the chambers occupied by lethal bullets (Fig. 4(a)).

Some agent randomly selects one of the guns, and then we fire from a randomly selected chamber at a man "Innocent Suspect". What is the kill probability K?

- *expected frequency analysis* – We might call $\{f_i\}$ the "lethalities" of $\{\mathfrak{R}_i\}$. I trust we agree that using either gun alone to play the traditional one-gun Russian roulette, the kill probability K is the same as the lethality of the gun; $K = f_1$ or $K = f_2$ as the case may be. Back to the two-gun game, the probabilities to select each gun, $w_1 = w_2 = 1/2$, can be considered as weights. The kill probability is then the weighted average, the expected value

$$\begin{aligned} K &= E(f_i) \\ &= \text{average of } 1/12 \text{ and } 3/12 \\ &= \textstyle\sum w_i f_i \\ &= 1/6. \end{aligned} \tag{6}$$

At the risk of insulting my readers, I expect that some of them feel that since the constituent guns kill either more or less often than 1/6 there is more to be said, and that the kill probability is better expressed as something like a "credible interval",

$K = 1/6 \pm c$, for some small number c.

Is it?

A simpler way to find K is to notice that the game is entirely equivalent to a one-gun game with a 24-chamber gun, 4 of which are loaded. The kill probability is therefore exactly $K = 4/24 = 1/6$; no ifs, buts, or and-an-uncertainty-interval about it.

Now consider a thought experiment involving haplotypes. Suppose that we know that haplotype frequencies are controlled by a whimsical agent Daemon Nature who monitors and controls all fertilization or birth events (à la Maxwell's Daemon) to ensure that the all haplotypes occur with one only two possible population frequencies, either $f = 1/12$ or $f = 3/12$, in such a way as to ensure that the type T of a randomly selected individual is equally likely to occur in the population with one frequency or the other. A capital crime occurs and the crime scene haplotype is T. An innocent suspect is arrested, who will be executed if found to match T. What is the probability the innocent suspect will match, and therefore die?

That scenario is exactly analogous to the roulette game. The whole project in which Nature chooses one of the two values for f then generates a population with freq(T) = f (that part constitutes an evolutionary replicate) then T turns up at the crime, is the analog of choosing a gun in the roulette game. Selecting, testing, and disposing one way or the other of an innocent (i.e. random) suspect, is the analog of firing the gun. Just as in the roulette game, the haplotype matching probability isn't the unknown number f; it's an average of the values f might have, known as the expected value of f (per Eq. (6)). And as in the roulette game $E(f) = 1/6$ exactly. Despite that we have the extreme situation in which this is known to be significantly different from freq(T) – a difference of $\pm 1/12$ – there is no more to be said, no confidence or credible interval. Such intervals are reasonably used in statistics to describe uncertainty in estimating a population parameter such as freq(T) by partially describing the probability distribution of the parameter. Matching probability however is not frequency, is not a parameter. Such intervals do not apply to probability itself because the very idea of a "probability distribution of a probability" is totally pointless since, as we have just seen in this example, such a probability distribution would be exactly equivalent to its expected value. The mistaken practice of attaching an uncertainty interval to a probability is not seen, so far as I can determine, outside of forensic statistics.

4.5. Frequency spectrum as a prior probability distribution

The situation in real life is more intricate than the preceding example but not different in principle. Nature provides a continuum of possible haplotype frequencies, so the two weights $\{w_i\}$ above are replaced by a prior probability distribution (Fig. 4(b)) – a continuous "frequency spectrum" $w(f)$ defined for all frequencies $f \in [0, 1]$ – and the Σ becomes an $\int$. For example if we assume the infinite alleles model discussed in Section 3.4 then the relative propensity for nature to create a haplotype with frequency f would be something like

$$w(f) \propto (1 - f)^{8800} / f, \tag{7}$$

a distribution very strongly favoring very small values of f (Fig. 4(b)). Next, in the real world of forensic genetics there is usually a population sample – e.g. $\mathcal{D}$ as in Section 3.1.1 – that can be incorporated via Bayes' theorem. For example, let p, $p \geq 1$, denote the number of observations of the haplotype of interest in $\mathcal{D}$. That's an event whose probability of occurrence, $c(p, f)$, is easily written down:

$$\begin{aligned} c(p, f) &= Pr(p \text{ observations} | \text{population frequency} = f) \\ &= f^p (1 - f)^{|\mathcal{D}| - p} \binom{|\mathcal{D}|}{p} \end{aligned}$$

and by Bayes' theorem the matching probability $m(p)$ is

$$m(p) = \int_{f=0}^{1} w(f) c(p, f) df. \tag{8}$$

Note that this formula works even with no reference data. In that case $p = 1$ and $\mathcal{D} = \{T\}$ (because of "extending" the database with the crime scene observation per Section 3.1.1), $c(1, f) = f$ and formula (8) looks very like formula (6).

For the case of $w(f)$ as in formula (7), via standard properties of the beta distribution $m(p)$ has the simple form given in Section A.2.4 of [1]:

$$m(p) = \frac{p}{|\mathcal{D}| + 8799}.$$

And as above that's the matching probability, end. Any impulse to decorate a probability by credible or confidence intervals is only from momentum and confusion, not from mathematical reasoning.

Which is not to say the number is scientifically exact. It's not; it will be to some extent wrong. But the source of error is uncertainty about the model and the premises, not sampling variation. If $\mathcal{D}$ doesn't represent the relevant population, that can be a problem – the magnitude of which has nothing to do with sampling variation. Since the infinite alleles model isn't an accurate description of evolution, the result cannot be accurate. In this regard having a larger $\mathcal{D}$ may paper over deficiencies in the model. But even so the extent of that isn't measured by credible/confidence intervals or sampling variation.

5. Concluding discussion

This paper has two related aims. First is to clarify the correctness of the κ method [1] in the face of published criticism [2]. The criticism is unsound, resting on misunderstanding and errors as Section 3.6.3 shows. The κ method (4) holds up well for so simple a model. Recent work, especially concordant results in [3], confirm that the method is right within its assumptions, and moreover shows that the model simplification sacrifices some accuracy but usually not very much. An additional benefit of the Laplace method [3], and to me a more important one, is that it gives plausible probabilities for never-observed haplotypes (which must be considered as part of analyzing Y mixtures) and for multiply-observed haplotypes.

The second aim grew out of the first. Understanding the basis of the κ approach requires understanding the Y haplotype terrain. And that in turn means casting aside preconceptions that we hold from autosomal experience and resisting the impulse to assume obvious seeming parallels between the two. There are a great many important differences.

1. Obviously the product rule is out for Y haplotypes – nothing new, everyone knew that.
2. Sample frequency as an estimate for matching probability (i.e. "blind counting", Section 3.1) is not bad for autosomal work although the augmented count (Section 3.1.1) is clearly better. For Y haplotype work even the latter leaves a lot on the table; sample frequency severely overestimates matching probability.
3. To clear the table better, it helps a lot in the Y case to look at the whole reference database, not just observances of the target haplotype. For autosomal work we never thought to do that (for the good reason that it wouldn't help much).
4. Confidence or credible intervals represent careless thinking but are mostly harmless in the autosomal arena. For Y, that careless thinking is an impediment to understanding and to expressing the actual evidential strength.
5. The "unrelated person" concept is an acceptable approximation for autosomal work; a fatal one for understanding Y haplotypes.
6. The autosomal idea of a θ correction is almost backwards from the reality of the Y haplotype matching world where θ is the starting point, and where literal application of "standard" autosomal theory gives negative probabilities.
7. Thinking about models is vital for understanding and developing a Y haplotype matching probability approach. Some autosomal practice acknowledges population substructure – good, but a mere academic exercise by comparison with the importance of theory for Y.
8. Finally there is another distinction outside the ambit of this paper, the common concern about possible geographical clustering of Y haplotypes. To what extent clustering complicates evidential calculation and what to do about it are topics to explore.

Acknowledgements

I thank David DeGusta for invaluable advice with the manuscript and the referees for prodding me to greater clarity.

References

[1] C. Brenner, Fundamental problem of forensic mathematics – the evidential value of a rare haplotype, Forensic Sci. Int. Genet. 4 (2010) 281–291.
[2] J. Buckleton, M. Krawczak, B. Weir, The interpretation of lineage markers in forensic DNA testing, Forensic Sci. Int. Genet. 5 (2011) 78–83.
[3] M.M. Andersen, P.S. Eriksen, N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies, J. Theor. Biol. 329 (2013) 39–51.
[4] C. Cockerham, Variance of gene frequencies, Evolution 23 (1) (1969) 72–84.
[5] J.F. Crow, et al., The Evaluation of Forensic DNA Evidence, National Research Council, 1996.
[6] Applied Biosystems, http://www6.appliedbiosystems.com/yfilerdatabase/, 2009.
[7] M. Slatkin, An exact test for neutrality based on the Ewens sampling distribution, Genet. Res. 64 (1994) 71–74.
[8] Wolfram Inc., Polya's Random Walk Constants, http://mathworld.wolfram.com/PolyasRandomWalkConstants.html.
[9] Long Bing, et al., Population genetics for 17 Y-STR loci (AmpFISTRY-filer TM) in Luzhou Han ethnic group, Forensic Sci. Int. Genet. 7 (2) (2013) e23–e26.
[10] M. Slatkin, A correction to the exact test based on the Ewens sampling distribution, Genet. Res. 68 (1996) 259–260.
[11] W. Ewens, The sampling theory of selectively neutral alleles, Theor. Popul. Biol. 3 (1972) 87–112.
[12] SWGDAM, SWGDAM Haplotype Frequencies, http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/oct2009/standards/2009_01_standards01.htm/, 2009.
[13] L. Roewer, M. Kayser, P. de Knijff, K. Anslinger, A. Betz, A. Caglia, D. Corach, S. Fredi, L. Henke, M. Hidding, H. Krgel, R. Lessig, M. Nagy, V. Pascali, W. Parson, B. Rolf, C. Schmitt, R. Szibor, J. Teifel-Greding, M. Krawczak, A new method for the evaluation of matches in non-recombining genomes: to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (1) (2000) 31–43.
[14] M. Krawczak, Forensic evaluation of Y-STR haplotype matches: a comment, Forensic Sci. Int. 118 (2) (2001) 114–115.
[15] S. Willuweit, A. Caliebe, M.M. Andersen, L. Roewer, Y-STR frequency surveying method: a critical reappraisal, Forensic Sci. Int. Genet. 5 (2001) 84–90.
[16] C. Brenner, The "frequency surveying" approach cannot work, http://dna-view.com/downloads/documents/RareHaplotypes/Critique/%20haplotype/%20A4.pdf.

[17] A. Veldman, Evidential Strength of Y-STR Haplotype Matches in Forensic DNA Casework, Mathematisch Instituut, Universiteit Leiden, 2007.
[18] M.M. Andersen, A. Caliebe, A. Jochens, S. Willuweit, M. Krawczak, Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory, Forensic Sci. Int. Genet. 7 (2) (2013) 264–271.
[19] B. Efron, Modern Science and the Bayesian-Frequentist Controversy, http://www-stat.stanford.edu/ckirby/brad/papers/2005NEWModernScience.pdf.
[20] H. Robbins, Estimating the total probability of the unobserved outcomes of an experiment, Ann. Math. Stat. 39 (1968) 256–257.
[21] T. Zerjal, et al., The genetic legacy of the Mongols, Am. J. Hum. Genet. 72 (2003) 717–721.
[22] A. Hájek, "Mises redux" – redux: fifteen arguments against finite frequentism, Erkenntnis 45 (1996) 209–227.
[23] J. Mill, System of Logic: Ratiocinative and Inductive; Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, vol. 2, London, http://www.questia.com/PM.qst?a=o&d=5774540: Longmans, Green, Reader, and Dyer, 62, 1868.
[24] P. Dawid, J. Mortera, Coherent evaluation of forensic evidence, J. R. Stat. Soc. B 58 (2) (1966) 425–443.