

# Understanding Y haplotype matching probability

Charles H. Brenner<sup>a,b</sup>

published *Forensic Sci. Int. Genet.* **8** 233–243, January 2014

<sup>a</sup>Human Rights Center, U.C. Berkeley, Berkeley, CA United States

<sup>b</sup>DNA-VIEW, 6801 Thornhill Drive, Oakland, CA 94611-1336, United States

---

## 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 modeling – 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 exchangeability 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.

**Keywords:** haplotype, Y-haplotype, likelihood ratio, weight of evidence calculation, probability, model

---

## 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  $\theta$  – the Cockerham/NRC II [4, 5] notation for allele sharing by common descent. The unexpected reason why it is not is explained in §2.4.

In the autosomal case, a simple model considers only allele probabilities (sometimes carelessly called frequencies – see §4) and taking  $\theta$  into account is a refinement whose 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  $\theta$  comes first and anything else is the minor refinement. Once we know  $\theta$  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)

---

Email address: cbrenner@berkeley.edu (Charles H. Brenner)

40 if they are identical and all the haplotypes along the (necessar- 87  
41 ily patrilineal) paths connecting them to their common ancestor 88  
42 are also identical. “Identical by state” (IBS) as used here is 89  
43 synonymous with “identical”, an umbrella meaning in that IBS 90  
44 thus includes IBD as a subset.<sup>1</sup> Adopting the umbrella defini- 91  
45 tion for IBS means some other term may be needed to mean 92  
46 IBS but not IBD and for this purpose I use the word “strict.” 93  
47 Identity strictly by state is also “convergence”.

48 The expression “unrelated man” is common in forensic prac- 94  
49 tice; a good enough approximation sometimes. In the Y haplo- 95  
50 type world, always remember that everyone is related for that is 96  
51 nearly the only reason men match. 97

52 The historical development of autosomal calculation meth- 98  
53 ods has been unsystematic, often intuitive, and to the extent 99  
54 that there are underlying models, which concepts have been in- 100  
55 corporated (such as random mating, subpopulations, mutation), 101  
56 has of course been motivated by relevance to the autosomal sit- 102  
57 uation. The appropriate and important concepts for a model, 103  
58 useful for haplotype analysis may be different ones. It makes, 104  
59 sense to begin by exploring the population genetics of haplo- 105  
60 types in order to gain an understanding then to consider models, 106  
61 and methods of calculation. 107

## 62 2. Exploring Y haplotype populations 109

63 Before discussing a few approaches, good and bad, that have, 111  
64 been suggested for Y haplotype matching probabilities, we will, 112  
65 delve into the nature of Y haplotype populations in order to, 113  
66 have some background that will be a helpful context for evalu- 114  
67 ating the approaches. 115

68 I see three ways to explore: data, theory, and simulations. 116

### 69 2.1. Exploring Y-haplotype data 118

70 Population samples (or “reference databases”) for the 17- 119  
71 locus Yfiler™ Y haplotypes for  $n > 1000$  men are conveniently 120  
72 available on the Internet [6] for several different populations. 121  
73 Examples, simulations, data and analysis mentioned in this pa- 122  
74 per assume Yfiler haplotypes unless otherwise specified. 123

75 Examining the data, a few points are quickly obvious: 124

76 (a) The vast majority of types that occur at all in a population 124  
77 sample occur exactly once.<sup>2</sup> I use the symbol  $\kappa$  for this 125  
78 singleton proportion.  $\kappa$  gradually becomes smaller as the 126  
79 number of collected reference haplotypes increases and is 127  
80 larger when haplotypes include more loci and hence are 128  
81 more polymorphic.  $\kappa > 0.8$  for reference databases con- 129  
82 sidered in this paper. 130

83 No type occurs many times. For example, among the 4102 131  
84 Caucasian full profiles,  $\kappa = 84\%$  are singletons (once oc- 132  
85 ccurring), and 98% of the men in the sample have a type 133  
86 shared with at most 5 men. Implications: 134

<sup>1</sup>Some writers use IBS differently – as an alternative or partially an alter- 135  
native to IBD – but the umbrella usage looks to me more common especially 136  
among careful writers.

<sup>2</sup>for presently available sample sizes. If – somewhat unrealistically – sample 137  
size  $n$  were increased without increasing the number of loci, then to be sure 138  
eventually  $\kappa < 1/2$  (for about  $n > 10930$  according to [1]). 139

(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 etc. model of infinitely many selectively neutral alleles generally shows very small  $p$ -values - i.e.  $p < 0.01$  for the Caucasian and other large

140 datasets.<sup>3</sup>

141 (d) Various published data is available for the mutation rates  
142 at the various Y STR loci, and not surprisingly they are  
143 similar to autosomal STR rates: STR mutation rates average  
144 1/350 per locus, so over the full 17 loci  $\mu = 17/350$ ,  
145 or 5%, per meiosis.

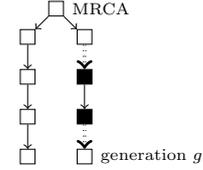


Figure 1: If the two IBS Y haplotypes of generation  $g$  are not IBD (are strictly IBS), then the lineage path connecting them through a common ancestor (MRCA) must include at least two mutational (dotted line) events.

## 146 2.2. Theoretical explorations

147 The main focus of this and the following section is exploring  
148 and understanding the claim in §1 that the story of rare haplo-  
149 type matching is essentially the story of common descent.

150 As a simplified but helpful model of a forensic Y-haplotype:  
151 Represent the data at each locus as an integer repeat amount;  
152 model a compound locus as if two loci; mutation occurs by single  
153 steps plus or minus at a locus and the mutation rate is the  
154 same 1/350 per meiosis at every locus. Under this model, a  
155 Yfiler haplotype is an ordered set of 17 integers, which is to say  
156 a point in a 17-dimensional integer lattice, and the evolutionary  
157 course of a haplotype as it is transmitted through generations  
158 with possible mutation is a 17-dimensional random walk. The  
159 mutation rate of  $\mu = 1/20$  per generation means that father and  
160 son carry the same haplotype with probability 0.95. If two con-  
161 temporary men trace back a combined  $2g$  generations to their  
162 most recent common patrilineal ancestor (MRCA), then with  
163 probability  $0.95^{2g}$  no mutation occurred at any of the  $2g$  meiotic  
164 events between them and their Yfiler haplotypes are IBD. For  
165 example, if the MRCA lived 100, 1000, or 10000 years ago then  
166  $g \approx 4, 40, \text{ or } 400$  and the IBD probabilities are 70%, 1/40 and  
167 1/10<sup>16</sup> (!) respectively. If they represent populations that have  
168 truly been separate for 10000 years, there is no chance for them  
169 to be identical by descent.

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

180 An example of a 3-dimensional integer lattice is a multi-  
181 story parking building with numbered rows and numbered slots  
182 within each row on each level. Now imagine 100 such build-  
183 ings, arrayed in a 10 by 10 grid. That's a 5-dimensional lattice  
184 of parking slots. Suppose you park your car and while you are  
185 gone someone randomly moves it a few times by step-wise muta-  
186 tion – “mutating” to the same slot in an adjacent row, then  
187 to the corresponding position one floor up or one structure to  
188 the East – like the meander through a haplotype's genealogi-  
189 cal mutational history. After even a few mutations your car is  
190 hopelessly lost. Equivalently, it is very unlikely that the car will

<sup>3</sup>However a few of the datasets show non-significant  $p$ -values such as three  
191 samples of size  $|D| \approx 300$  from Malaysia in Table 2 of [1], for which  $p = 0.08$ ,  
192 0.46, and 0.66.

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

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

Cancelling patterns involving more than two mutations are im-  
probable (and trickier to compute), but when considering long  
time spans – millenia – are, although a slight chance, virtu-  
ally the only chance of Y haplotype random matching and their  
probability is therefore at least of evolutionary interest. Figure 2  
shows the relationship. Matching across a patrilineal separation  
of several centuries is nearly always because there are no inter-  
vening 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 prob-  
abilities increase slightly to 70%, 1/30 and 1/10<sup>8</sup>.

Compare Y haplotype populations with the theoretically  
much discussed “infinitely many neutral alleles” model of  
Kimura, Crow, etc. (§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.

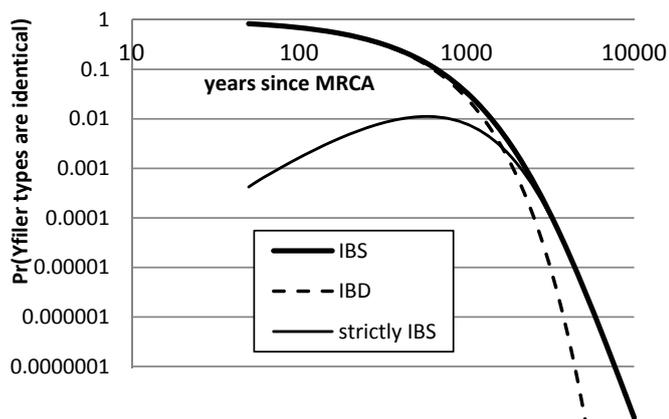


Figure 2: Comparing probability of any match (IBS) with IBD and strictly IBS. A match strictly by state is very unlikely although when the MRCA was more than 1500 years ago, IBD is even more unlikely.

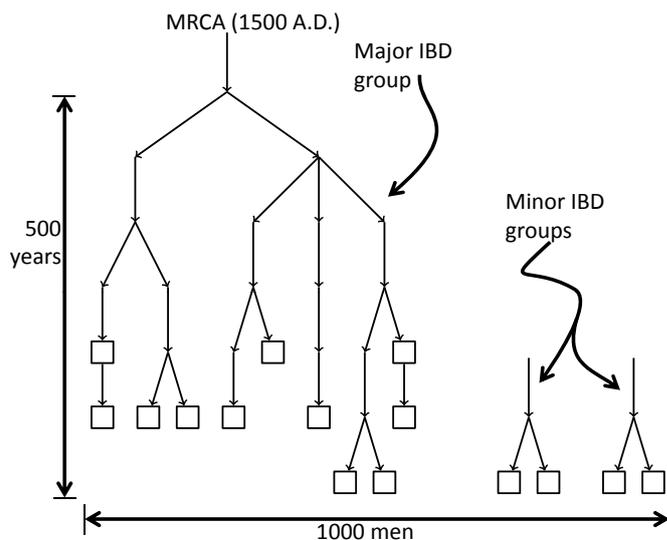


Figure 3: Schematic representation of the IBS cohort for a typical haplotype T. A single large IBD pedigree tree accounts for 98% of all T haplotypes.

### 2.3. Y-haplotype population simulations

Figure 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  $\Pr(\text{IBD}|\text{IBS}) \approx 33/34$ , and the probability would increase to the limit of 1 as the number of loci increases. This says that if two 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 (patrilinely 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.<sup>4</sup> Thus the almost invariable pattern is that almost all the men with a given type belong to one dominant IBD group. Figure 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

<sup>4</sup>as you can easily convince yourself with a little mental shuffling of IBD group size proportions  $d_1, d_2, \dots, \sum d_i = 1$ , realizing that  $\Pr(\text{IBD}|\text{IBS}) \approx \sum d_i^2$ .

(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 §1 that the story of many-locus haplotype matching is predominately a story of  $\theta$  and which thus tends to justify the simplifying modeling assumption under which the “ $\kappa$  method” (§3.6.2 and [1]) ignores haplotype structure. Consideration of haplotype structure (such as by looking at mutational neighbors; viz §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”.

### 2.4. Thinking about $\theta$

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

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

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  $\theta$  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  $\theta$  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  $\theta$  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  $\theta$  means the chance that two randomly selected men belong to the same patrilineal IBD group as in Figure 3. Hence in the case of Y-haplotype matching the first term of (1), the  $\theta$ , 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  $\theta$  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 §2.1(a)i) then it rates to be less than averagely common and  $\theta$  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 §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:

- 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.

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	(§3.1)	4102/0 – i.e. infinite
Brenner counting	(§3.1.1)	4102
C.I. from 0 (SWGDM)	(§3.2)	4102/3
frequency surveying	(§3.3)	~ 4100?
infinite alleles model	(§3.4)	13000
average matching chance	(§3.5)	8800
$t$ -model	(§3.6.1)	23000
$\kappa$ method	(§3.6.2)	25000
discrete Laplace model	(§3.7)	~ 25000

Table 2: Glossary of notation

notation	meaning
$LR_\Omega$	matching LR assuming model $\Omega$
$LR_\Omega(T, \mathcal{D})$	matching LR assuming $\Omega$ and using explicitly mentioned data
$T$	type to match (data)
$\mathcal{D}$	reference database augmented with the target haplotype (data)
$ \mathcal{D} $	size of $\mathcal{D}$
$\Omega$	population model
$\Omega^i, \Omega^k$	particular models
$\theta$	$\Pr(IBD)$ – ([4, 5] notation)
$\theta_{Ewens}$	$\theta$ as in [11]; essentially the effective number of alleles, minus 1
$\kappa$	proportion of singletons in $\mathcal{D}$
$p$	“popularity”; number of occurrences
$\propto$	proportional to; i.e. if $y = 2x$ then $y \propto x$

#### 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

339 appropriate question for evaluating a DNA match is the prob-388  
 340 ability that an innocent (meaning randomly selected with re-  
 341 spect to the crime scene donor) suspect would have the target389  
 342 type, conditional upon the crime scene occurrence. Condition-390  
 343 ing is equivalent to mentally extending the reference database  
 344 by adding the target type [1]. The consequence is that a new  
 345 type is reported as occurring one time in the extended database  
 346  $\mathcal{D}$ . What calculation to then make from  $\mathcal{D}$  remains to be de-392  
 347 cided and various possibilities will be discussed later, but of  
 348 course the most obvious idea is to consider the sample fre-393  
 349 quency,  $1/|\mathcal{D}|$ , and it is a conservative estimate (i.e. an over-394  
 350 estimate) of the probability that an innocent suspect will match  
 351 as noted in §2.1(b)i.

352 “Overestimate the probability” means that the expected value  
 353 of the sample frequency is greater than the probability in light  
 354 of the data and everything we know about nature. That’s not a  
 355 purely mathematical truth; the “nature” component is critical.  
 356 In fact maybe it’s impossible to have a non-trivial correct for-  
 357 mula without population genetic modelling assumptions – i.e.  
 358 impossible to prove validity with mere mathematics and with-  
 359 out appeal to population genetic reality (but §3.6.1 has the op-  
 360 posite speculation). As a thought experiment imagine a model  
 361 called *common-type* under which nature has a very strong ten-  
 362 dency to discourage types rarer than 25% population frequency.  
 363 Under that model, if we observe five instances of a type  $T$  in a  
 364 sample of size  $n = 50$ , it is much more likely that we have  
 365 under-observed a common trait than over-observed a rare one,  
 366 hence we would properly conclude that  $\Pr(T) > 5/50$ , i.e. the  
 367 counting method would be anti-conservative. It follows that  
 368 accepting even the counting method entails accepting at least  
 369 some weak modeling condition, e.g ruling out *common-type*. 415

In reality, because of drift and mutation, nature very much416  
 370 prefers rare Y STR haplotypes, like Figure 4(b), hence the417  
 371 counting method is quite conservative. The counting rule like-418  
 372 lihood ratio 419

$$LR_{\text{counting}} = |\mathcal{D}| \quad (2)420$$

370 is therefore an understatement of the evidence that the suspect  
 371 is the crime scene donor. 422

### 372 3.2. Confidence interval from zero 425

373 Something like the approach of SWGDAM recommenda-426  
 374 tions [12] has been recommended by several forensic statisti-427  
 375 cians but I do not like it as it has no mathematical basis. It does428  
 376 not correspond to any model and it confuses concepts. 429

377 As noted above, the case of interest is a crime stain not found430  
 378 in the  $n$ -haplotype reference database. The analysis begins with431  
 379 the correct idea that the value of the evidence against a matching432  
 380 suspect can be expressed as the probability that a random (i.e.433  
 381 non-donor) person would match, conditional on the observa-434  
 382 tion at the crime. Two errors in quick succession then abruptly435  
 383 derail the analysis: failing to condition, and replacing “proba-436  
 384 bility” by the different and inappropriate concept of “population437  
 385 frequency”. The right question 438

386 Based on the evidence that a type has been seen once,  
 387 what is the probability to see it again? 441

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 statis-  
 391 tician might estimate the frequency of a type that has been  
 392 observed several times, namely confidence intervals – an arti-  
 393 fice at best which becomes particularly dubious in the arena of  
 394 zero observations, the dreaded “confidence interval from zero.”  
 395 Statisticians know well that even at best a confidence interval  
 396 doesn’t really mean a range the frequency is likely to be in  
 397 (though that of course is how it will be understood), but rather  
 398 a converse of that: a range such that if the frequency is within it  
 399 the actual observation is likely. In some situations the approxi-  
 400 mation of one by the other may be plausible; zero observations  
 401 is not such a situation. Finally, the question arises as to what  
 402 confidence interval is appropriate. Historically 95% is an arbi-  
 403 trary choice. It’s not rooted in any principle. Possibly it has  
 404 proven its mettle in a practical way in arenas like manufactur-  
 405 ing process control, but there is no reasonable analogy, let alone  
 406 logic, for how such an experience would translate to DNA, evi-  
 407 dence, and justice. A comical debate lately arose about the  
 408 right size for the one-sided interval from zero in the SWGDAM  
 409 procedure. A reasonable analogy would be debating what kind  
 410 of tutu to put on a dog. In truth, adopting a confidence interval  
 411 from thin air is no more than a way to paper over and cover up  
 412 deeper seated illogic. See §4.

### 3.3. Frequency surveying

Haplotype “frequency surveying” [13, 14, 15] is a proposal  
 413 motivated by the notion that haplotypes which are near step-  
 414 wise neighbors will tend to have similar frequencies, presuma-  
 415 bly because of mutation among neighbors. Hence it seeks to  
 416 augment the paucity of reference observations for a particular  
 417 haplotype by considering as well the richness or sparseness of  
 418 nearby haplotypes. However this tempting idea doesn’t hold up  
 419 on close consideration. It suffers from a handful of possibly  
 420 reparable shortcomings in execution, and an insuperable funda-  
 421 mental error. 424

The shortcomings begin with no explicit model. There-  
 425 fore the curve-fitting approach isn’t derived mathematically, but  
 426 rather consists of the guesswork that it will be good enough to  
 427 assume that each one-step neighbor will contribute to a hap-  
 428 lotype’s frequency as much as two two-step neighbors, three  
 429 three-step neighbors, and so on. Presumably the image is that  
 430 there is some kind of traffic among the haplotypes in a densely  
 431 populated region of the 17-dimensional lattice space so that  
 432 they replenish (or otherwise influence) one another through mu-  
 433 tation. However, that image is tantamount to assuming a high  
 434 rate of convergence through mutation (even across several mu-  
 435 tational differences), contrary to the story described in §2.2. On  
 436 the other hand with rare traits – which haplotypes are – ge-  
 437 netic drift acts relatively quickly. Consequently I expect any  
 438 influence from a replenishment phenomenon to be dwarfed by  
 439 genetic drift. Haplotype frequencies are mostly just random.  
 440 Notwithstanding an unfortunate and false endorsement in [2]

442 that the simultaneously published paper [15] presents a valida-480  
 443 tion of frequency surveying ([15] itself makes no such claim),  
 444 the frequency surveying approach cannot work. [16, 17, 18, 3]  
 445 have further discussion. However, my impression from author  
 446 Krawczak is that he now agrees the method is invalid and there-  
 447 fore perhaps it is a dead issue.

### 448 3.4. The infinite alleles model approach

From Ewens' celebrated [11], the model of infinitely many 483  
 neutral alleles suggests a prior distribution of haplotype fre-484  
 quencies that amounts to  $\beta(0, \theta_{\text{Ewens}})$  (Figure 4(b) and formula485  
 7).  $\theta_{\text{Ewens}} \approx 8800$  for Caucasians per §2.1(b). Given such a486  
 prior distribution one can apply Bayes' theorem, incorporating487  
 sample data if available, to compute a posterior probability:

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

488  
 489  
 490  
 491  
 492  
 493  
 494  
 495  
 496  
 497  
 498  
 499  
 500  
 501  
 502  
 503  
 504  
 505

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

### 454 3.5. The average matching chance

455 The empirical pairwise matching experiment described in  
 456 §2.1(b) is the same as the experiment of comparing an innocent  
 457 man with a crime stain. Therefore it can fairly be used in re-  
 458 porting the evidence (either for a haplotype not in the database  
 459 as the estimate is then conservative, or if the reference database  
 460 is lost) and an appropriate statement is easy to formulate and  
 461 explain:

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

### 466 3.6. Models motivated by the data

467 Efron [19] makes the point that in the modern statistical  
 468 world of "big data", models are sometimes implied by the data  
 469 itself. Haplotype population sample data shows a prevalence of  
 470 singletons, and some of the implications of this have already  
 471 been mentioned §2.1(a). If every type in the database were  
 472 unique – as would be the case if we sampled complete genomes  
 473 for example – the hypothesis of maximum likelihood would be  
 474 that everyone in the world is unique (and a lower bound on the506  
 475 order of  $n^2$  effective types). The presence of some repeated ob-507  
 476 servations in the database implies an upper bound as well as a  
 477 lower on the amount of diversity in the population. I consid-  
 478 ered two approaches exploiting this idea, the "t-model" and the  
 479 "k method".

#### 3.6.1. t-model

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

$$t = \frac{n}{-\ln \kappa} \quad (3)$$

as is shown in [1]. The model of equal frequencies is artificial,  
 but it may be possible to prove that it is conservative by demon-  
 strating 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. There-  
 fore it is a possibility that (3) is conservative without any popu-  
 lation genetic assumptions.

#### 3.6.2. k method

The main result of [1] is the "k method", which refines the  
 counting method. As in the counting method augment the refer-  
 ence database to  $\mathcal{D}$  by including the crime scene type in it. Let  
 $\kappa$  be the proportion of singletons in  $\mathcal{D}$ . The following lemma  
 quantifies remark §2.1(a).

**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  $\kappa$ , one of the  
 singletons and hence was unrepresented before he was added.  
 The same will be true of the next man added – probability  $\kappa$  that  
 his type will be previously unrepresented. That's the same as  
 saying that  $\kappa$  of the population is unrepresented in  $\mathcal{D}$ . Therefore  
 the proportion of the population that is represented is  $1 - \kappa$ .  $\square^5$

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

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

as the matching likelihood for a singleton and in general

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

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:

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

If the inflation proportion  $I$  holds for each of the popularity  
 cohorts individually, then (5) is correct (unbiased). However,

<sup>5</sup>A rigorous version of this lemma would acknowledge that  $\kappa$  changes  
 slightly with the addition of a new man, and the proof appeals to Robbins'  
 theorem [20].

508 unless we can rule out that some of the more-observed types<sup>562</sup>  
509 may be legitimately common because of some special mech-<sup>563</sup>  
510 anism (Genghis Khan effect [21]? selection via hitchhiking?)<sup>564</sup>  
511 the full generality of (5) would be hard to show and therefore<sup>565</sup>  
512 is cautiously not recommended by [1]. For singletons though,<sup>566</sup>  
513 (4) is well-supported by simulations. Incidentally for a rapidly<sup>567</sup>  
514 growing population it may be notably conservative which im-<sup>568</sup>  
515 plies that for at least some  $p > 1$  the more general formula is<sup>569</sup>  
516 anti-conservative, another reason for caution in its use at least  
517 in court. For humanitarian body identification and small  $p$  (5)<sup>570</sup>  
518 seems reasonable though.

### 519 3.6.3. Doubts about $\kappa$ by Buckleton, Krawczak and Weir

520 While not written as formal mathematics, [1] lays out ar-<sup>574</sup>  
521 guments and results systematically – statement of problem,<sup>575</sup>  
522 premises, model, derivation of results, validation of the  $\kappa$ ,<sup>576</sup>  
523 and deductive organization are clear communication and facil-<sup>577</sup>  
524 itating a substantive and reasoned discussion or argument. But,<sup>578</sup>  
525 the assertion “we have shown, Brenner’s approach ... suffers,<sup>579</sup>  
526 from potential anti-conservativeness in the way it inherently es-<sup>580</sup>  
527 timates haplotype frequencies” as part of the ending discussion,<sup>581</sup>  
528 in [2] is not backed up by substantive or reasoned discussion.<sup>582</sup>  
529 Inquiry to the authors elicited that the remark referred to §5.2<sup>584</sup>  
530 of [2]. That section mentions three ideas which are respectively<sup>585</sup>  
531 pointless, confused, and wrong as follows:<sup>586</sup>

- 532 • “difficult to determine” – After a bit of mathematical ma-<sup>587</sup>  
533 nipulation aimed at gaining insight into the performance  
534 of the  $\kappa$  method, §5.2 of [2] comes to a rather complicated<sup>588</sup>  
535 but correct expression for the expected value of my for-  
536 mula in terms of the (unknown) population frequencies of<sup>589</sup>  
537 various haplotypes. If the complicated expression could<sup>590</sup>  
538 be shown by analysis to have too large or too small an ex-<sup>591</sup>  
539 pected value, that would show that my method is conser-<sup>592</sup>  
540 vative (ok) or anti-conservative (bad). The paper comes to<sup>593</sup>  
541 neither conclusion, but rather, frankly admits that “It is a<sup>594</sup>  
542 complex function ... therefore it is difficult to judge” based<sup>595</sup>  
543 on this, sadly, dead end approach. I sympathize – my note,<sup>596</sup>  
544 books are littered ideas that didn’t pan out, but I don’t see<sup>597</sup>  
545 any point in publishing them.<sup>598</sup>
- 546 • “Brenner demonstrated himself” an example where the<sup>599</sup>  
547 method is anti-conservative. For an artificial and unre-<sup>600</sup>  
548 alistic population consisting of 10000 equally rare haplo-<sup>601</sup>  
549 types ( $\Omega^{10000}$  in the notation of [1]), the  $\kappa$  formula would<sup>602</sup>  
550 not work. But [1] doesn’t claim it would; rather it ex-<sup>603</sup>  
551 plicitly points out that the modelling condition of “equal<sup>604</sup>  
552 over-representation” upon which the method rests can be<sup>605</sup>  
553 violated by artificially constructed populations but seems<sup>606</sup>  
554 valid for populations that might occur naturally. My own<sup>607</sup>  
555 pathological example is therefore not evidence against my<sup>608</sup>  
556 own method.<sup>609</sup>
- 557 • elementary algebraic blunder – [2] derives from the  $\kappa$  for-<sup>610</sup>  
558 mula the result that any singleton in a database rates to be<sup>611</sup>  
559 seen more than once in a second realization of the same  
560 database – a surprising and counterintuitive conclusion if  
561

correct. One could then presume that sampling a third  
time would include still more copies of the ever burgeon-  
ing 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 fre-  
quency 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.<sup>6</sup>

### 3.7. discrete Laplace model

I argue above (§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 collec-  
tion of subpopulations each of which is a collection of haplo-  
types clustered around a central (ancestral?) haplotype. The  
model incorporates neighborhood information in that it calcu-  
lates higher probabilities for haplotypes closer to a central hap-  
lotype. Analyzing the performance on many simulated popula-  
tions, the paper concludes that while both  $\kappa$  and Laplace hap-  
lotype probabilities are close to unbiased compared to the ac-  
tual frequency, the Laplace method is more accurate. Moreover  
and very usefully the Laplace method is applicable even for un-  
observed 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 “prob-  
ability is not frequency” but that slogan may incite accidental  
misunderstanding because of its freighted and potentially am-  
biguous words.

Probability and frequency are related concepts to be sure. In-  
deed, a common (the “frequentist”) definition of the probability  
of an event is the frequency with which that event would oc-  
cur in an unending series of repeated experimental trials. Even  
an anti-frequentist who refuses this as a definition probably ac-  
cepts the intuition it represents at least for a reasonably objec-  
tive instance of probability. Hájek [22]: “[F]inding out a rela-  
tive 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 se-  
lected person would match. That’s a fairly objective example  
of probability, the sort we should be able to handle without get-  
ting mired in the Bayesian/Frequentist ideological conflict of  
20th century statisticians.

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

612 Of course higher-probability events occur more frequently<sup>664</sup>  
613 whether in life or in the “repeated experiment” interpretation<sup>665</sup>  
614 of probability. But that abstract sense of “frequency” isn’t the<sup>666</sup>  
615 sense at issue, population frequency is. Population frequency<sup>667</sup>  
616 is not abstract; it’s an actual fact about nature, a number *but*<sup>668</sup>  
617 *an unknown one* that exists in the real world. And the insti-<sup>669</sup>  
618 tutionalized misconception in forensics is to confuse matching<sup>670</sup>  
619 probability with *population frequency*. For example consider  
620 the probability for a man to match a crime scene Y-haplotype<sup>671</sup>  
621  $T$  assuming that the man has been randomly selected from the<sup>672</sup>  
622 population. Population frequency means the number of men<sup>673</sup>  
623 with  $T$  in the population divided by the number of men in the<sup>674</sup>  
624 population.

625 “So what?” some will say. “Doesn’t the matching proba-<sup>675</sup>  
626 bility end up being the haplotype frequency in the population<sup>676</sup>  
627 anyway? If 500 men in a population of 500000 have type  $T$  –<sup>677</sup>  
628 population frequency is 1/1000 – then imagining a large number<sup>678</sup>  
629 of repetitions of the evidential situation won’t random suspects<sup>679</sup>  
630 match 1/1000 of the time?” That sounds quite reasonable. What<sup>680</sup>  
631 is wrong with it?<sup>681</sup>

632 It would be correct if the population frequency of  $T$  were<sup>682</sup>  
633 *known* to be 1/1000. Then and only then are haplotype fre-<sup>683</sup>  
634 quency and matching probability equivalent. For probability<sup>684</sup>  
635 necessarily is a summary of data – known (or stipulated) facts<sup>685</sup>  
636 – and cannot possibly depend on facts that are unknown. This<sup>686</sup>  
637 is an undisputable and familiar principle.<sup>687</sup>

638 For example we say team A has some probability to  
639 win a game coming up this weekend based on our  
640 knowledge of past encounters, information about the  
641 players’ conditions, and other information.

642 If the weekend passes and we receive no information  
643 whatever relevant to the game, then we can equally  
644 ask in retrospect “What is the probability that team A  
645 won?” and with the evidence the same, the answer  
646 will be the same. Sometimes people say there is no  
647 probability any more and either it is 0 or 1, depending  
648 on what happened, but talk like that is facetious or  
649 confused. The reality of winning changed over the  
650 weekend; the probability of winning did not.

651 The probability of the team to win isn’t a property of the team<sup>703</sup>  
652 or the game itself, but rather is a description of the particular<sup>704</sup>  
653 knowledge we have about them. As the 19th century thinker JS<sup>705</sup>  
654 Mill put it [23],<sup>706</sup>

655 Every event is in itself certain, not probable; if we  
656 knew all, we should either know positively that it will  
657 happen, or positively that it will not. But its probabili-<sup>707</sup>  
658 ty to us means the degree of expectation of its occur-<sup>708</sup>  
659 rence, which we are warranted in entertaining by our  
660 present evidence.<sup>709</sup>

661 Once we subscribe to the confused position that probability<sup>713</sup>  
662 can depend on unknown facts, we confront the impossible ques-<sup>714</sup>  
663 tion as to which unknown facts are privileged to be considered<sup>715</sup>

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 un-  
known. What can we do instead? Does it matter?

#### 4.1. What can we do instead?

The appropriate understanding of haplotype matching prob-  
ability isn’t how often a random match will occur to the partic-  
ular 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 §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 notwith-  
standing 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  $\kappa$  model includes the simplify-  
ing assumption that a Y haplotype “sequence” (the repeat num-  
bers) 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 sim-  
ilarity among Y haplotypes, then you might adopt a model un-  
der 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 evalu-  
ation means discarding a factor of 18 of evidentiary strength

<sup>7</sup>genotypes at additional untested loci?

718 from a terrestrial number. And it could be worse. I've han-765  
 719 dled cases with a single 1-step Y-haplotype discrepancy, i.e. 766  
 720 between alleged father and son; consequently  $LR = 25000 \times$  767  
 721  $Pr(\text{mutation}) \approx 70$ . That's still pretty strong evidence, but wan- 768  
 722 tonly reduce it a further 18-fold and you've got nothing.

723 Besides, it's a good principle to do things right. A clear and 770  
 724 correct understanding is a good foundation for progress. Ran- 771  
 725 dom half-wrong ideas and guesses are not, and they're also dan- 772  
 726 gerous to bring to court. A good defense attorney should be able  
 727 to shred an expert who talks nonsense, even if the nonsense is  
 728 numerically better for the defense than the right number would  
 729 be.

Schematically:

**unsound:** database  $\Rightarrow$  sample frequency  
 sample frequency  $\Rightarrow$  frequency  $\pm$  interval  
 frequency  $\pm$  interval  $\approx$  probability  $\pm$  interval  
**correct:** data+model  $\Rightarrow$  probability

730 The unsound paradigm for matching probability amounts to:  
 731 The type is very rare in the database. Therefore it is *probably*  
 732 at least *pretty* rare in the population. The feeble logical conclu-  
 733 sion is only that matching by an innocent suspect is *probably*  
 734 *unlikely*.

735 A correct footing takes the larger perspective in which the  
 736 database is viewed as evidence ("database as evidence" – [24])  
 737 drawn from a population that conforms to some modeling  
 738 premises. Then we can say: The type is very rare in the  
 739 database. That may be because the type is very rare in the popu-  
 740 lation or it may be because the database is a very abnormal sam-  
 741 ple. Either way, for an innocent suspect to match requires a very 773  
 742 unlikely circumstance, meaning the evidence is very strong that  
 743 a matching suspect is not innocent.

744 Two examples follow of situations in which the matching 775  
 745 probability to haplotype  $T$  has little relation to the population 776  
 746 frequency of  $T$ . The first is a little artificial, but simpler. 777

#### 747 4.3. Example – An island population 779

748 Thought experiment: In 14 A.D. a complete catalogue is 781  
 749 compiled of the Y haplotypes of every man in a closed com- 782  
 750 munity. One particular type  $T$  happens to have a population 783  
 751 frequency of exactly 1/100. Suppose by immigration controls 784  
 752 and magically stopping mutation, no new types will be intro- 785  
 753 duced. Fast forward 2000 years and the modern population, 786  
 754 because of genetic drift, has completely different haplotype fre- 787  
 755 quencies which are unknown – we have no new data. Just as a 788  
 756 puzzle, what is the probability that a randomly selected man is 789  
 757 type  $T$ ?

758 Answer: The random man can trace his lineage back to a 791  
 759 single man in 14 A.D., and that ancestor is 1/100 to be type 792  
 760  $T$ . Hence so is our man. Of course if we knew the modern 793  
 761 population frequency the answer would be different, but based 794  
 762 on the knowledge that we actually have the probability is 1/100. 795  
 763 Matching probability is not population frequency; rather it is 796  
 764 whatever inference is warranted from available evidence. 797

#### 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 cham-  
 bers occupied by lethal bullets (Figure 4(a)). Some agent ran-  
 domly 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

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

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",

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

Is it?

A simpler way to find  $K$  is to notice that the game is en-  
 tirely 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 con-  
 trols all fertilization or birth events (à la Maxwell's Daemon)  
 to ensure that the all haplotypes occur with one only two pos-  
 sible 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  $\text{freq}(T) = f$  (that part con-  
 stitutes an evolutionary replicate) then  $T$  turns up at the crime,  
 is the analog of choosing a gun in the roulette game. Select-  
 ing, testing, and disposing one way or the other of an innocent  
 (i.e. random) suspect, is the analog of firing the gun. Just as

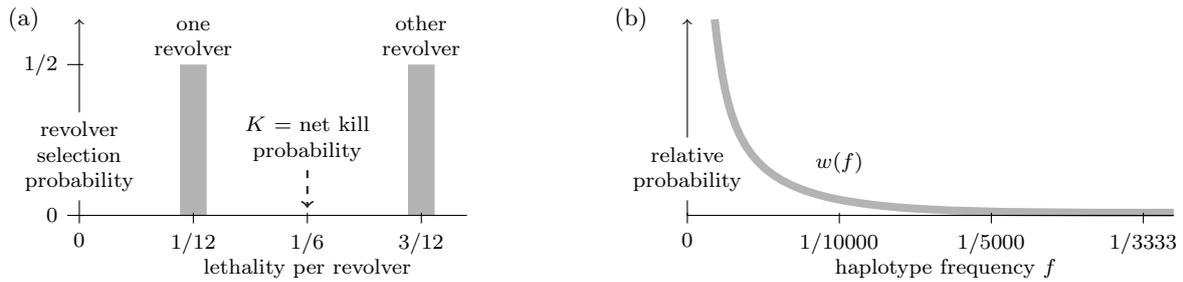


Figure 4: discrete and continuous frequency spectra: (a) 2-gun roulette (§4.4), (b) “infinite alleles” haplotype frequency distribution (§4.5)

798 in the roulette game, the haplotype matching probability isn't  
 799 the unknown number  $f$ ; it's an average of the values  $f$  might  
 800 have, known as the expected value of  $f$  (per equation (6)). And  
 801 as in the roulette game  $E(f) = 1/6$  exactly. Despite that we  
 802 have the extreme situation in which this is known to be signifi-  
 803 cantly different from  $\text{freq}(T)$  – a difference of  $\pm 1/12$  – there is  
 804 no more to be said, no confidence or credible interval. Such in-  
 805 tervals are reasonably used in statistics to describe uncertainty  
 806 in estimating a population parameter such as  $\text{freq}(T)$  by par-  
 807 tially describing the probability distribution of the parameter.  
 808 Matching probability however is not frequency, is not a param-  
 809 eter. Such intervals do not apply to probability itself because the  
 810 very idea of a “probability distribution of a probability” is to-  
 811 tally pointless since, as we have just seen in this example, such  
 812 a probability distribution would be exactly equivalent to its ex-  
 813 pected value. The mistaken practice of attaching an uncertainty  
 814 interval to a probability is not seen, so far as I can determine,  
 815 outside of forensic statistics.

#### 816 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 con-  
 tinuum of possible haplotype frequencies, so the two weights  
 $\{w_i\}$  above are replaced by a prior probability distribution (Fig-  
 ure 4(b)) – a continuous “frequency spectrum”  $w(f)$  defined for  
 all frequencies  $f \in [0, 1]$  – and the  $\Sigma$  becomes an  $\int$ . For ex-  
 ample if we assume the infinite alleles model discussed in §3.4  
 then the relative propensity for nature to create a haplotype with  
 frequency  $f$  would be something like

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

a distribution very strongly favoring very small values of  $f$  (Fig-  
 ure 4(b)). Next, in the real world of forensic genetics there is  
 usually a population sample – e.g.  $\mathcal{D}$  as in §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:

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

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

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

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 §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 proper-  
 ties of the beta distribution  $m(p)$  has the simple form given in  
 §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 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.

## 835 5. Concluding discussion

This paper has two related aims. First is to clarify the cor-  
 rectness of the  $\kappa$  method [1] in the face of published criticism  
 [2]. The criticism is unsound, resting on misunderstanding and  
 errors as §3.6.3 shows. The  $\kappa$  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.

848 The second aim grew out of the first. Understanding the basis<sup>899</sup>  
849 of the  $\kappa$  approach requires understanding the Y haplotype ter-<sup>900</sup>  
850 rain. And that in turn means casting aside preconceptions that<sup>901</sup>  
851 we hold from autosomal experience and resisting the impulse,<sup>902</sup>  
852 to assume obvious seeming parallels between the two. There<sup>903</sup>  
853 are a great many important differences.<sup>904</sup>

- 854 1. Obviously the product rule is out for Y haplotypes – noth-<sup>907</sup>  
855 ing new, everyone knew that.<sup>908</sup>
- 856 2. Sample frequency as an estimate for matching probabil-<sup>910</sup>  
857 ity (i.e. “blind counting”, §3.1) is not bad for autosomal<sup>911</sup>  
858 work although the augmented count (§3.1.1) is clearly bet-<sup>912</sup>  
859 ter. For Y haplotype work even the latter leaves a lot on<sup>913</sup>  
860 the table; sample frequency severely overestimates match-<sup>914</sup>  
861 ing probability.<sup>915</sup>
- 862 3. To clear the table better, it helps a lot in the Y case to look<sup>917</sup>  
863 at the whole reference database, not just observances of the<sup>918</sup>  
864 target haplotype. For autosomal work we never thought to<sup>919</sup>  
865 do that (for the good reason that it wouldn’t help much).<sup>920</sup>
- 866 4. Confidence or credible intervals represent careless think-<sup>921</sup>  
867 ing but are mostly harmless in the autosomal arena. For Y,<sup>922</sup>  
868 that careless thinking is an impediment to understanding<sup>923</sup>  
869 and to expressing the actual evidential strength.<sup>924</sup>
- 870 5. The “unrelated person” concept is an acceptable approxi-<sup>925</sup>  
871 mation for autosomal work; a fatal one for understanding<sup>926</sup>  
872 Y haplotypes.<sup>927</sup>
- 873 6. The autosomal idea of a  $\theta$  correction is almost backwards<sup>928</sup>  
874 from the reality of the Y haplotype matching world where<sup>929</sup>  
875  $\theta$  is the starting point, and where literal application of<sup>930</sup>  
876 “standard” autosomal theory gives negative probabilities.<sup>931</sup>
- 877 7. Thinking about models is vital for understanding and de-<sup>932</sup>  
878 veloping a Y haplotype matching probability approach.<sup>933</sup>  
879 Some autosomal practice acknowledges population sub-<sup>934</sup>  
880 structure – good, but a mere academic exercise by com-<sup>935</sup>  
881 parison with the importance of theory for Y.<sup>936</sup>
- 882 8. Finally there is another distinction outside the ambit of this<sup>937</sup>  
883 paper, the common concern about possible geographical<sup>938</sup>  
884 clustering of Y haplotypes. To what extent clustering com-<sup>939</sup>  
885 plicates evidential calculation and what to do about it are<sup>940</sup>  
886 topics to explore.<sup>941</sup>

## 887 Acknowledgements

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

- 890 [1] C. Brenner, Fundamental problem of forensic mathematics – the eviden-  
891 tial value of a rare haplotype, *Forensic Sci. Int. Genet.* 4 (2010) 281–291.
- 892 [2] J. Buckleton, M. Krawczak, B. Weir, The interpretation of lineage mark-  
893 ers in forensic DNA testing, *Forensic Sci. Int. Genet.* 5 (2011) 78–83.
- 894 [3] M. M. Andersen, P. S. Eriksen, N. Morling, The discrete Laplace expo-  
895 nential family and estimation of Y-STR haplotype frequencies, *Journal of*  
896 *Theoretical Biology* 329 (2013) 39–51.
- 897 [4] C. Cockerham, Variance of gene frequencies, *Evolution* 23 n1 (1969) 72–  
898 84.

- [5] J. F. Crow, et al, The evaluation of forensic DNA evidence, National Re-  
search Council, 1996.
- [6] Applied Biosystems, <http://www6.appliedbiosystems.com/yfilerdatabase/> (2009).
- [7] M. Slatkin, An exact test for neutrality based on the Ewens sampling dis-  
tribution, *Genetical Research* 64 (1994) 71–74.
- [8] Wolfram Inc., Polyas’s random walk constants, <http://mathworld.wolfram.com/PolyasRandomWalkConstants.html>.
- [9] Long Bing et al, Population genetics for 17 Y-STR loci (AmpFISTR-  
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, *Genetical Research* 68 (1996) 259–260.
- [11] W. Ewens, The sampling theory of selectively neutral alleles, *Theor Pop*  
*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/](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 com-  
ment, *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 haplo-  
types using coalescent theory, *Forensic Sci. Int. Genet.* 7(2) (2013) 264–  
271.
- [19] B. Efron, Modern science and the Bayesian-frequentist contro-  
versy, <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 fre-  
quentism, *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 Invest-*  
*igation*, 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.*  
*Statist. Soc. B* 58(2) (1966) 425–443.