Fundamental problem of forensic mathematics –
The Evidential Value of a Rare Haplotype

Charles H. Brenner*

* School of Public Health, Forensic Science Group, U.C. Berkeley and DNA-VIEW
6801 Thornhill Drive, Oakland, California 94611-1336
☎+1 510 339 1911, fax -1181
cbrenner@berkeley.edu

Abstract

Y-chromosomal and mitochondrial haplotyping offer special advantages for criminal (and other) identification. For different reasons, each of them is sometimes detectable in a crime stain for which autosomal typing fails. But they also present special problems, including a fundamental mathematical one: When a rare haplotype is shared between suspect and crime scene, how strong is the evidence linking the two? Assume a reference population sample is available which contains \( n-1 \) haplotypes. The most interesting situation as well as the most common one is that the crime scene haplotype was never observed in the population sample. The traditional tools of product rule and sample frequency are not useful when there are no components to multiply and the sample frequency is zero. A useful statistic is the fraction \( \kappa \) of the population sample that consists of “singletons” – of once-observed types. A simple argument shows that the probability for a random innocent suspect to match a previously unobserved crime scene type is \( (1-\kappa)/n \) – distinctly less than \( 1/n \), likely ten times less. The robust validity of this model is confirmed by testing it against a range of population models.

Keywords: haplotype; stain matching; mtDNA; Y-haplotype; forensic mathematics; likelihood ratio; matching probability

1 Notation ................................................................. 3
2 Introduction .............................................................. 4
  2.1 mtDNA & Y-chromosomes ........................................ 4
  2.2 The problem ..................................................... 4
3 Materials and Methods ..................................................... 5
  3.1 Data ............................................................ 5
  3.2 Population simulations ........................................ 6
  3.3 Neutrality test ................................................... 8
1 Notation

$S_i$  a once-observed haplotype. $S_0$ is the haplotype from the crime scene.

$T$  haplotype of an innocent suspect

$y$  probability that $T$ matches $S_0$, given that $S_0$ has been observed once

$y_p$  probability that $T$ matches a $p$-times observed type. ($y_1 = y$)

$N$  population size.

$\varnothing$  denotes a population of haplotypes. $\varnothing_1, ..., \varnothing_m, ...$ are various simulated populations.

$\varnothing_t$  population of $t$ equally rare traits

$\varnothing_{\gamma t}$  population of a common trait of frequency $\gamma$ and rare traits of frequency $1/t$

$n$  sample size

$D$  a sample drawn from some $\varnothing$. $D^*$ is a sample before including the crime stain.

$\alpha$  number of once-observed haplotypes (singletons) in a sample

$p$  popularity of an allele – number of times it occurs in a sample

$\alpha_1, \alpha_2, ..., \alpha_p, ...$ the number of singletons, doubletons, ... haplotypes of popularity $p$, in a sample. ($\alpha_1 = \alpha$)

$\kappa$  sample proportion $\alpha/n$ of singletons in a sample

$f_S$  population frequency of type $S$

$f_{s1}, f_{s2}, f_{s\rho}$  total population proportion of the haplotypes that are singletons, non-singletons, of popularity $p$ in some sample.

$N_e$  effective population size, a population parameter

$q$  probability two randomly selected types match (absent knowing $p$)

$m_e$  effective number of types – reciprocal of the probability two randomly selected types match

$\mu$  mutation rate per generation for a haplotype system
2 Introduction

2.1 mtDNA & Y-chromosomes

In recent years there has been increasing interest in using Y-chromosomal haplotypes [1-5] or mtDNA for forensic identification. These haplotype systems are also much used for body identification, especially for old graves [6, 7]. The advantages for some kinds of problems are considerable. Both methods are desirable for attacking kinship problems involving remote relatives because sex-linked traits are not diluted 50% each generation. Both methods are useful for some scant crime stains — mtDNA because of the high copy number, the Y-chromosome because it can be detected and amplified unambiguously even when it is a minor component compared to the female victim in a rape sample.

However, an mtDNA or a Y-chromosomal haplotype must be treated mathematically as a single indivisible (“atomic”) trait; so unlike those traditional DNA methods which examine several traits that are approximately independent of one another, no multiplication of probabilities is possible. Therefore it is vital to have a sound fundamental understanding of atomic trait matching probabilities in order to make a reasonable assessment of the strength of identification evidence.

2.2 The problem

Evidence — such as a Y-haplotype — is identical between a crime scene and a suspected donor. How strong is the evidence that the suspect is the donor? In particular, this paper discusses the critical matching probability question, "What is the probability that a random non-donor would by chance match a newly observed type?"

Such interesting questions and complications as possible dependence among traits and suitability of the reference population sample are outside the domain of this paper, essential though they may be to the comprehensive evaluation of DNA evidence. Moreover, to keep the discussion focused and simple, I ignore complications such as mitochondrial heteroplasmy, or adjustments appropriate because of multiple samples or preliminary phenotypic analysis. I take the DNA sequence of a haplotype as being no more than an arbitrary name; in this respect my model differs from that of Krawczak [8].

The matching probability will be derived by analysis of a population sample. For the purpose of this paper, assume that the population sample is appropriate, that it is randomly representative of possible innocent suspects. It may well be true that geographical clustering or other population genetic phenomena make it difficult to obtain a perfectly representative reference sample (subset, or “database”) but for the purpose of this paper put such difficulties aside.

The problem attacked here is therefore a modest one. But it is also a fundamental one, for without understanding the proper analysis of an individual indivisible trait under the simplest assumptions it is
impossible to give a proper analysis of DNA evidence in any situation. I therefore consider this the fundamental question of forensic mathematics.

I shall focus on the case – typical for mtDNA and Y-haplotype – that the observed crime scene type is not found in the population sample.

Evidentiary strength of a matching type is given by a likelihood ratio (LR), called here the matching LR:

$$LR = \frac{x}{y},$$

where $x$ and $y$ are likelihoods

$$x = \Pr(\text{suspect matches crime scene type} \mid \text{suspect is the donor})$$

$$y = \Pr(\text{suspect matches crime scene type} \mid \text{suspect unconnected to the crime}).$$

The simple view is to assume $x=1$. That isn’t strictly true – one could model typing fallibility (laboratory error), heteroplasm (in the case of mitochondrial evidence), and mutation (when identification is via relatives). But to keep to the point, settle for $x=1$. That being the case, we have simply $LR=1/y$, where $y$ is the matching probability.

The formulation of $y$ here states “suspect unconnected to the crime.” I prefer this to saying “suspect unrelated to the donor” since after all everyone is genetically related, and when dealing with a Y-chromosomal or mtDNA match to pretend otherwise is particularly artificial. As an assumption that is both realistic enough to be useful and ideal enough to permit analysis, think of the suspect, if innocent, as being randomly selected, hence having the same random chance as anyone else in the population to have any particular relationship to the donor. (Similarly, while a reference population sample should be gathered without a bias toward related people it also should not be purged of relatives. It should contain people who are related randomly.)

3 Materials and Methods

3.1 Data

The application and applicability of the methods developed here are evaluated partly by consideration and analysis of real data. Available population studies include Y haplotype and mtDNA datasets as listed in Table 2. The Y-haplotype databases are ABI data obtained per [9], Charité from [8], ancient Mongolian [7], Reliagene provided by S Sinha, Portuguese from H Gaeda, Macedonian from Z Jakovski, Krakow from M Sanak and from T Kupiec, Malaysian from P Krishnan, and the Ashkenazy data is M Hammer’s passed to me via AFDIL. The information about mtDNA databases is from M Coble (AFDIL) and T Melton (Mitotyping Technologies).
For the purpose of this paper a haplotype database is sufficiently described by it’s spectrum – the numbers $\alpha_1, \alpha_2, ...$ of traits seen once, twice, etc. The spectrum was determined for each Y haplotype database after eliminating the handful of irregular observations (profiles with a missing, off-ladder, or extra allele). See Table 1.

### Table 1: Spectra and Singleton Proportion ($\kappa$) of Some Y Haplotype Population Samples

<table>
<thead>
<tr>
<th>Population</th>
<th>$n$</th>
<th>$\kappa = \alpha / n$</th>
<th>#loci</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
<th>$\alpha_6$</th>
<th>$\alpha_7$</th>
<th>$\alpha_8$</th>
<th>$\alpha_{10}$</th>
<th>$\alpha_{11}$</th>
<th>$\alpha_{12}$</th>
<th>$p &gt; 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black (US) Reliagene</td>
<td>2239</td>
<td>0.61</td>
<td>11</td>
<td>1373</td>
<td>187</td>
<td>44</td>
<td>28</td>
<td>19</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Caucasian Reliagene</td>
<td>1826</td>
<td>0.55</td>
<td>11</td>
<td>1005</td>
<td>113</td>
<td>44</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hispanic Reliagene</td>
<td>454</td>
<td>0.77</td>
<td>11</td>
<td>348</td>
<td>17</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Native Am Reliagene</td>
<td>104</td>
<td>0.74</td>
<td>11</td>
<td>77</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Portuguese</td>
<td>315</td>
<td>0.94</td>
<td>17</td>
<td>297</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Malaysian (Malaysia)</td>
<td>332</td>
<td>0.96</td>
<td>17</td>
<td>319</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chinese (Malaysia)</td>
<td>331</td>
<td>0.89</td>
<td>17</td>
<td>295</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Indian (Malaysia)</td>
<td>305</td>
<td>0.94</td>
<td>17</td>
<td>287</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Krakow (IES)</td>
<td>523</td>
<td>0.93</td>
<td>17</td>
<td>485</td>
<td>19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ashkenazi Jewish</td>
<td>496</td>
<td>0.35</td>
<td>10</td>
<td>176</td>
<td>28</td>
<td>19</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Macedonian</td>
<td>100</td>
<td>0.83</td>
<td>17</td>
<td>83</td>
<td>7</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Krakow (Jagellonian)</td>
<td>189</td>
<td>0.95</td>
<td>17</td>
<td>179</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ancient Mongolian</td>
<td>27</td>
<td>0.67</td>
<td>10</td>
<td>18</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Black (US) ABI</td>
<td>985</td>
<td>0.94</td>
<td>17</td>
<td>925</td>
<td>28</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>African ABI</td>
<td>59</td>
<td>0.66</td>
<td>17</td>
<td>39</td>
<td>6</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Asian ABI</td>
<td>330</td>
<td>0.95</td>
<td>17</td>
<td>312</td>
<td>7</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Caucasian ABI</td>
<td>1276</td>
<td>0.90</td>
<td>17</td>
<td>1152</td>
<td>49</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Filipino ABI</td>
<td>105</td>
<td>0.96</td>
<td>17</td>
<td>101</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SE Hispanic ABI</td>
<td>597</td>
<td>0.91</td>
<td>17</td>
<td>543</td>
<td>22</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Native Am ABI</td>
<td>106</td>
<td>1.00</td>
<td>17</td>
<td>106</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Vietnamese ABI</td>
<td>103</td>
<td>0.94</td>
<td>17</td>
<td>97</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.2 Population simulations

The conclusions of this paper are partly checked and justified by computer simulation experiments. Each experiment consists of modeling the evolution of a population of haplotypes. Thirty-nine different simulated populations $\Theta_1, ..., \Theta_{39}$ were thus generated under a variety of modeling conditions of mutation rate, mutation model, population size, and rate of population growth.

All populations were generated following a Wright-Fisher approach, simulating a generation at a time. Each new generation is obtained as a sample with replacement from the previous generation with some of the haplotypes then modified by a mutation rule. Some populations are grown, starting with a single
haplotype and growing at a chosen rate per generation until the target size is reached. Other populations are obtained from a grown population by stabilizing for several hundred generations without growth. Mutation rules include: infinite many alleles (every mutation produces a brand new haplotype [10]), and modified stepwise with a specified number $l$ of loci ($95\%$ of mutations add or subtract one step at one locus, $5\%$ two steps). Analysis of the first dozen (preliminary) models suggested creating 27 systematically designed population models by choosing all combinations of three values of each of three parameters as follows:

- population growth of $0$, $3\%$ [11], or $10\%$/generation,
- population size $N$ of $30K$, $100K$ or $300K$ men,
- mutation rate ($\mu$) slow, medium, or fast. “Medium” is $\mu=17/300$ per generation, imitating the roughly average $1/300$ per locus per generation mutation rate typical of current human identification STR loci including those of the 17-locus Y-filer system. “Slow” is $1/3$ of the medium rate, simulated by 17 slower-mutating loci. “Fast” is triple the medium rate, modeled by triple the number of loci. Analysis of the early models showed that the stepwise model isn’t significantly different from infinite alleles.

Within each population $\varOmega_m$ we can then assess the matching probability formula that I suggest. Assessment is done by inspecting samples from the simulated populations. From each $\varOmega_m$, 100 samples were drawn and analyzed for each of the sample sizes $n=100, 300, 1000, 3000,$ and $10000$. Each sample $D=D(m,n,j), j=1,...,100,$ has some collection $S_0,...,S_n,...,S_{n,1},$ of $a(m,n,j) \ (0 \leq \alpha \leq n)$ distinct haplotypes each of which occur once in $D$. For any particular type $S \in D$ the population frequency $f_{S} = (1/N) \sum_{T \in \varOmega_m} I(T=S)$ – here $I(\cdot)$ is the indicator function – can be determined by peeking into $\varOmega_m$ (as would not be possible in real life). The formula for matching probability ($\alpha$) – based on observable data – suggested in this paper is assessed by comparison with $f_{S}$.

A matching formula is valid to the extent that it has the same expected value (or in some contexts a smaller, “conservative” value is valid) as the expected value of $f_{S}$. The appropriate interpretation of the term “expected” is

(a) since likelihood ratios are multiplied, evaluated in the logarithmic domain, in this case $\log(1-\kappa)$;
(b) averaged over indistinguishable cases, i.e. for fixed $m, n$, all singletons $i$, and arguably all samples $j$, are indistinguishable.

Hence the formula is valid if, for a realistic population model $\varOmega$ and sample size $n$, the probability it
The point of view here is prospective, contrary to the classical preference of some mathematical writers who evaluate the likelihood ratio in retrospect after $S_0$ and (matching) suspect type $T$ are both observed. I prefer my approach because I think it is slightly less formal, but the mathematics end up the same either way.

3.3 Neutrality test

mtDNA haplotypes or a Y-haplotype system consisting of a large number of STR loci are logically equivalent to a locus with a very high mutation rate, a large number of possible mutational changes and a very large number of potential alleles. As such, it is plausible that these haplotype systems might approximate the infinitely many selectively neutral alleles model [12]. Slatkin’s test [13] is a Monte Carlo procedure which evaluates a population sample by comparing the distribution of very rare and not-so-rare traits with model expectations. The model predicts a proliferation of rare types (since many are created and many go extinct at each generation), fewer slightly more common types, etc. If the population sample spectrum is relatively unexpected, that is evidence against the model. For the simulated populations which are stabilized, the Slatkin test is used to decide when equilibrium has been reached.

4 Analysis

4.1 Extended database

Denote by $S_0$ the crime scene haplotype, and let $D^-$ (“database”) be a population sample of size $n-1$. Since $D^-$ is stipulated to be an appropriate and representative population sample, there’s no harm in thinking of $D^-$ as a collection of crime scene haplotypes from other similar but unrelated crimes. Assume that the type $S_0$ does not appear in $D^-$. There is some collection of types that occur exactly once in $D^-$. Call them singletons (Figure 1), and let $\alpha_1-1$ be the number of singletons. To avoid double subscripts I shall generally omit the subscript 1 and write $\alpha$ for $\alpha_1$. Label the singleton types $S_1, S_2, \ldots, S_{\alpha-1}$. These and $S_0$ are all by definition distinct. We shall eventually\(^1\) determine the haplotype $T$ of some innocent suspect, a person unrelated to the crime scene donor, or more precisely at most randomly related like

\[
\begin{array}{ccc}
DYS390 & DYS391 & DYS392 \\
22 & 10 & 11 & * \\
23 & 10 & 11 & \\
23 & 10 & 11 & \\
23 & 10 & 11 & \\
23 & 10 & 11 & \\
23 & 11 & 12 & * \\
24 & 9 & 11 & * \\
24 & 11 & 11 & * \\
25 & 10 & 11 & \\
25 & 10 & 11 & \\
\end{array}
\]

Figure 1 Example population sample $D$ with $n=10$ and $\alpha=4$ singletons (starred). Hence $\kappa=0.4$.

\(^1\) The point of view here is prospective, contrary to the classical preference of some mathematical writers who evaluate the likelihood ratio in retrospect after $S_0$ and (matching) suspect type $T$ are both observed. I prefer my approach because I think it is slightly less formal, but the mathematics end up the same either way.
any two people in \( \mathbf{D}^* \). The situation of concern is when \( T=S_0 \), when our innocent suspect is unlucky inculpated. Therefore the matching probability \( y=\Pr(T=S_0) \) is of vital interest.

The first key observation is that the probabilities \( \Pr(T=S_i) \) are all the same. Probability is a summary of information and the information about the all the \( S_i \) is exactly the same: all been observed exactly once and all lack any connection to the innocent suspect. Consequently it is convenient and appropriate to think of the extended database \( \mathbf{D}=\mathbf{D}^-\cup\{S_0\} \) obtained by tossing \( S_0 \) into \( \mathbf{D}^- \). \( \mathbf{D} \) is a population sample of size \( n \), with \( \alpha \) singletons: \( S_0, S_1, S_2, ..., S_{\alpha-1} \). The matching probability to a crime sample of a new type and with database \( \mathbf{D}^- \) is equivalent to the matching probability to a singleton in database \( \mathbf{D} \) when there has been no crime.

4.2 A simplified approach – counting

Imagine comparing \( T \) with the \( n \) types in \( \mathbf{D} \). Since \( T \) by stipulation has no connection to the crime or to \( \mathbf{D}^* \), even if we assume that \( T \) matches some type among the \( n \), the probability \( y \) is at most \( 1/n \) that \( T \) match any particular unique type – call it a singleton – such as \( S_0 \). Therefore if LR is the likelihood ratio that a matching suspect is the source of \( S_0 \) we can say confidently that

\[
LR > LR_c = n
\]

where the subscript \( c \) indicates reference to what we may refer to as the “counting model.” Note though that the count includes \( S_0 \).

It is surely possible to do even better. A major task of this paper is deriving (§ 4.4) and justifying an inflation factor (§ 5.1) by which (1) can be improved if we additionally take into account the considerable chance \( T \) doesn’t match any of the \( n \) previously observed types.

It could be that even a further improvement is possible by arguing that, even when \( T \) matches some type among the \( n \), singletons such as \( S_0 \) have less than their \( 1/n \) fair share of probability to be matched. This possibility is examined in § 5.2.2 However, while there may be such an effect if some haplotypes are associated with a fitness advantage, it is not very large and is hard to model in a convincing way.
Figure 2 Change in $\kappa(n)$, the expected fraction of once-observed types, with increasing sample size $n$ for Y-filer haplotypes (ABI Caucasian). Each $\kappa(n)$ is estimated by averaging the $\kappa$ values for several hundred random $n$-subsets of the original 1272 haplotypes. The best fit to a curve of the form $\kappa(n)=\theta/(\theta+n)$ is with $\theta=10930$.

4.3 Analysis of singletons

Let $\kappa=\alpha/n$ be the proportion of singletons among a sample of $n$ haplotypes. Obviously $\kappa$ depends on the testing system. All other things being equal, testing 17 loci on the Y-chromosome produces a larger proportion of singletons than does testing 7 loci. Also, $\kappa$ depends on the sample size $n$. If $n$ is small it may be that $\kappa=1$, but we concentrate on $\kappa<1$. Conversely if $n$ is a substantial fraction of the world population most sampled men will have a son, father, or brother also sampled, so there will be relatively few singleton Y-haplotypes. In the limit $\kappa$ is about twice the per-generation mutation rate. Certainly $\kappa$ changes only slowly with $n$ (see Figure 2). Appendix §6.2 indicates why it is natural to expect that $\kappa(n)$ behaves approximately as $2/(n+2)$ for some constant $\theta$, and interprets $\theta$.

What is the probability that the next haplotype observed will be new? Answer: Since $\kappa$ grows only slowly, it should be about the same as the probability that the last one was, which is to say $\kappa$. This is close to a theorem of H. Robbins [14]. It follows as an immediate corollary that about $\kappa$ of the types in the population are not represented in $D$, have not been seen. That’s how typical it is that a crime scene type is previously unseen.

4.4 Fundamental question

Now we can evaluate $y=Pr(S_0=T)$. The analysis proceeds in three simple steps. If $S_0=T$, then each of the following must be true:

A. (Observed) $T$ matches something in $D$.

Since $T$ is a new type with probability $\kappa$, $Pr(Observed)=1-\kappa$. (2)

B. (Singleton) $T$ matches some singleton in $D$.

An exact answer may be difficult, but an obvious claim is
Pr(Singleton | Observed) \approx \kappa \quad (3)

which is considered in more detail in §5.2.

C. (Match) \quad T=S_0.

Given that T matches some singleton S_i in D, the subscript i is equally likely to have one value as another, hence

Pr(Match | Observed & Singleton) = 1/\alpha. \quad (4)

Ploddingly putting the above together –

\[ y = \Pr(\text{Match}) \]
\[ = \Pr(\text{Match} \& \text{Singleton} \& \text{Observed}) \]
\[ = \Pr(\text{Match} | \text{Singleton} \& \text{Observed})\Pr(\text{Singleton} | \text{Observed})\Pr(\text{Observed}) \]
\[ = (1 - \kappa)\kappa / \alpha \]
\[ = (1 - \kappa) / n. \quad (5) \]

Consequently the evidential strength for matching a previously unseen haplotype

\[ LR = LR_\kappa = n / (1 - \kappa) \quad (6) \]

where the subscript \kappa corresponds to the above analysis according the “kappa model.”

5 Discussion

5.1 Inflation factor

It has occasionally been said that the matching LR can be no larger than \( n - 1 \), the size of the database. Formula \( (6) \) gives matching LR that are larger than \( n \) by a factor of \( 1 / (1 - \kappa) \), which I therefore call the inflation factor in Table 2.

<table>
<thead>
<tr>
<th>sample size</th>
<th>singletons</th>
<th>singleton proportion</th>
<th>inflation factor</th>
<th>matching LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n - 1 )</td>
<td>( \alpha - 1 )</td>
<td>( \kappa = \alpha / n )</td>
<td>( 1 / (1 - \kappa) )</td>
<td>( LR_\kappa = n / (1 - \kappa) )</td>
</tr>
<tr>
<td>( \kappa = 1/q )</td>
<td>( \kappa = 1/q )</td>
<td>( \kappa = 1/q )</td>
<td>( \kappa = 1/q )</td>
<td>( \kappa = 1/q )</td>
</tr>
</tbody>
</table>

\[ t \text{ model} \]
\[ \text{empirical – per pairwise match rate} \]
<table>
<thead>
<tr>
<th>Mitotyping Technologies</th>
<th>mtDNA</th>
<th>2000</th>
<th>~1200</th>
<th>~60%</th>
<th>2.5</th>
<th>~5000</th>
<th>3917</th>
</tr>
</thead>
<tbody>
<tr>
<td>mtDNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFDIL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all races</td>
<td>7867</td>
<td>4941</td>
<td>0.63</td>
<td>2.7</td>
<td>21157</td>
<td>16919</td>
<td></td>
</tr>
<tr>
<td>Caucasians</td>
<td>1219</td>
<td>750</td>
<td>0.62</td>
<td>2.6</td>
<td>3174</td>
<td>2514</td>
<td></td>
</tr>
<tr>
<td>7 Y loci Charité</td>
<td>2438</td>
<td>1101</td>
<td>0.45</td>
<td>1.8</td>
<td>4449</td>
<td>3070</td>
<td></td>
</tr>
<tr>
<td>Caucasians</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-Plex (11 loci) Reliagene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>1605</td>
<td>1205</td>
<td>0.75</td>
<td>4.0</td>
<td>6448</td>
<td>5607</td>
<td>1410</td>
</tr>
<tr>
<td>Caucasian</td>
<td>1242</td>
<td>691</td>
<td>0.56</td>
<td>2.3</td>
<td>2804</td>
<td>2122</td>
<td>285</td>
</tr>
<tr>
<td>Hispanic</td>
<td>452</td>
<td>348</td>
<td>0.77</td>
<td>4.4</td>
<td>1973</td>
<td>1737</td>
<td>465</td>
</tr>
<tr>
<td>Native American</td>
<td>104</td>
<td>77</td>
<td>0.74</td>
<td>3.9</td>
<td>408</td>
<td>353</td>
<td>153</td>
</tr>
<tr>
<td>Yfiler 17-plex; ABI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>977</td>
<td>917</td>
<td>0.94</td>
<td>16.3</td>
<td>15941</td>
<td>15447</td>
<td>14020</td>
</tr>
<tr>
<td>American</td>
<td>59</td>
<td>39</td>
<td>0.67</td>
<td>3.0</td>
<td>180</td>
<td>148</td>
<td>90</td>
</tr>
<tr>
<td>African</td>
<td>327</td>
<td>309</td>
<td>0.95</td>
<td>18.2</td>
<td>5977</td>
<td>5811</td>
<td>4100</td>
</tr>
<tr>
<td>Caucasian</td>
<td>1272</td>
<td>1148</td>
<td>0.90</td>
<td>10.3</td>
<td>13069</td>
<td>12421</td>
<td>8883</td>
</tr>
<tr>
<td>Filipino</td>
<td>105</td>
<td>101</td>
<td>0.96</td>
<td>26.5</td>
<td>2809</td>
<td>2756</td>
<td>2730</td>
</tr>
<tr>
<td>SE Hispanic</td>
<td>597</td>
<td>543</td>
<td>0.91</td>
<td>11.1</td>
<td>6622</td>
<td>6319</td>
<td>5233</td>
</tr>
<tr>
<td>Native American</td>
<td>106</td>
<td>106</td>
<td>~0.99*</td>
<td>107*</td>
<td>11500*</td>
<td>11400*</td>
<td>~</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>103</td>
<td>97</td>
<td>0.94</td>
<td>17.3</td>
<td>1803</td>
<td>1750</td>
<td>1751</td>
</tr>
<tr>
<td>17 Y loci Gaeda</td>
<td>313</td>
<td>295</td>
<td>0.94</td>
<td>17.4</td>
<td>5478</td>
<td>5319</td>
<td>2713</td>
</tr>
<tr>
<td>Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashkenazi Jews</td>
<td>496</td>
<td>176</td>
<td>0.36</td>
<td>1.6</td>
<td>772</td>
<td>481</td>
<td>97</td>
</tr>
</tbody>
</table>

* To avoid $\kappa=1$, instead of $\alpha=n$ analyzed on the basis that $\alpha>n-1$.

Table 2 some haplotype populations sample statistics.

5.1.1 Pathological inflation

For small enough samples or exceedingly polymorphic traits it may happen that $\kappa=1$ in which case formula (6) gives an infinite LR and infinite inflation factor, which obviously cannot be accurate. Possible remedies include a careful and refined statistical treatment, or simply avoiding $\kappa$ too close to
1. For the present study I prefer the latter approach.

5.2 Accuracy – sample overrepresentation

Let’s consider more carefully question B, \( \Pr(\text{Singleton} \mid \text{Observed}) \). If there are any types in the population that don’t appear in the sample, they are \textit{a fortiori} under-represented in the sample. Therefore to compensate, types that do occur in the sample tend to be over-represented [15]. This would be most obviously true for represented rare types. The intuition behind (3) is that singleton types in \( D \) are on average at least as over-represented, compared to their population frequency, as are non-singleton types. That is, let \( f_{\geq 1} \) denote the (normally unknown) combined population frequency of the types \( \{S_i\} \) which are singletons in \( D \), and \( f_{>1} \) denote the combined population frequency of the types that occur multiply in \( D \). The corresponding sample frequencies are \( \kappa \) and \( 1 - \kappa \). Then \( \kappa / f_{\geq 1} \) is the over-representation rate for singletons and \( (1 - \kappa) / f_{>1} \) is the over-representation rate for non-singletons. In claiming (3) we are assuming that

\[
\frac{\kappa}{f_{\geq 1}} \approx \frac{1 - \kappa}{f_{>1}},
\]

though for the courtroom context we will normally also be quite happy if the inequality

\[
\frac{\kappa}{f_{\geq 1}} \geq \frac{1 - \kappa}{f_{>1}}
\]

holds as it is “conservative” in the sense that it implies that the LR given by (6) at worst understates the evidence connecting suspect to crime.

Is (8) justified? That depends on the frequency spectrum – the distribution of frequencies of types – that the evolutionary mechanisms such as mutation, drift, and selection tend to produce. First consider two unrealistic extreme frequency spectra, then reality.

5.2.1 A diverse artificial population

At one extreme, the worst case for (8), imagine the sample \( D \) of size \( n \) to be drawn from an artificial population \( \Omega' \) consisting of some large unknown number \( t \), approximately in the range \( n < t < n^2 / 2 \), of equally rare types. Then \( D \), in addition to \( a_1 \) singletons we may have some \( a_2 > 0 \) doubletons and even some tripletons, i.e. duplicated types, which are over-represented two and three times more than the singletons are over-represented. In short, the singletons are over-represented the least and (8) fails; the recommendation (6) would be anti-conservative. Formula (3) is therefore not as obvious as it looks. However, even if we erroneously apply the \( \kappa \) model formula (6) to the \( \Omega' \) model, the expected error in using \( LR_{\kappa} \) instead of the correct \( LR_{\Omega'} \) would be small (Appendix, §6.1.1).
5.2.2 Very popular haplotypes

At the opposite extreme consider a population $\Theta^N$ with a common type $Z$ of substantial frequency and a large number of very rare types of frequency $1/t$. As an example, $\Theta^{0.2,10000}$ has a 20% type and 8000 rare types. A sample $D$ from $\Theta^{0.2,10000}$ of size $n=100$ rates to include nearly 80 singletons (thus $k=0.8$ and $f_1=80/10000$), about 20 copies of $Z$, and perhaps a few doubletons of rare types ($f_1=0.2+\epsilon$). Hence $k/f_1=100$ while $(1-k)/f_1=1$, so (8) holds with a factor of 100 to spare.

5.2.3 Probable populations

The frequency spectrum of a real haplotype systems lies in between. Constant $t$ is unrealistic. Different lineages will have differing numbers of descendants so by random drift $\Theta^N$ is not stable. Significantly common types $Z$ as in the $\Theta^N$ model tend to crumble out of existence because of mutation, but this depends on the rate of mutation, which in turn is proportional to the number of variable sites in the haplotype. For example if we momentarily limit our attention to the 9-locus “minimal haplotype” [8] projection of 1272 17-locus ABI Caucasians (Table 1), one type occurs 66 times (5%) and another 36 times (3%) – common types like $Z$. Considering all 17 loci breaks these two large groups into 86 groups, mostly singletons of course, and the largest clump is 4 men (the second most popular 17-locus type; the most popular occurs 7 times). Consequently, the two sides of (7) are not far from equal and therefore there is not a lot of scope for improving (6) when the mutation rate is high.

5.3 Observations from the simulation models

To assess the behavior of the formula for realistic frequency spectra, populations were simulated by computer under various evolutionary models and numerous samples (“databases”) examined from each one. See Materials and Methods. In this section the population simulations are used as a proving ground to evaluate the $\kappa$ model.

5.3.1 Preliminary observations

Following are a few preliminary observations based on examination of the simulations.

- There are no common types. This is ensured by the high rate of mutation. Even if selection – which was not modeled – favored some mitochondria or Y chromosomes, common types per §5.2.2 are impossible (also see Table 1) so $LR_\kappa$ cannot be highly conservative.

---

2 providing we except extremes such as sequencing the entire Y chromosome which might realize the model $\Theta^N$ – a different type for everyone.
All else being equal,

- A smaller population has commoner haplotypes. A larger population supports a proportionally larger number of types, hence the commonness of haplotypes (whether measured by the commonest ones or by the rarer ones) tends to be inversely proportional to population size.

- Slow mutation means commoner haplotypes. For fixed population size and growth rate haplotype frequencies vary inversely with mutation rate – about proportionally if no growth, more than proportionally if the population is growing.

- Smaller samples means larger \( \kappa \). This is obvious from Figure 2.

- The number of loci across which the mutation is distributed doesn’t matter as long as it is large. 17 loci is about the same as infinite alleles.

### 5.3.2 Evaluation of the \( \kappa \) model

For a given population \( \Theta_m \) and fixed sample size \( n \), there are samples \( D(m, n, j)\) each with some fraction \( \kappa(j) \) of singletons. For each such singleton \( S_i \in D_j \), there is a population frequency \( f_{S_i} \). If (6) is a good formula then on average \( n/(1-\kappa(j))=1/f_{S_i} \), i.e. the inflation factor \( 1/(1-\kappa(j)) \approx 1/nf_{S_i} \). It turns out, a little surprisingly, that the average value of the population frequencies is independent of \( \kappa(j) \); if \( D_n \), by accident of sampling, has more singletons than \( D_\alpha \), nonetheless the singletons in those two sets are on average equally rare. Therefore we can estimate the expected population frequency for a singleton by the average \( f \) of all the \( f_{S_i} \) taken over \( i \) and \( j \), and correspondingly define the ideal or effective kappa \( \kappa_e(m, n) \) as that value of \( \kappa_e \) which satisfies \( 1/(1-\kappa)=1/nf \).

The test of the \( \kappa \) model (6) then consists in comparing the average value of \( 1/(1-\kappa(j)) \) with the ideal value \( 1/(1-\kappa_e) \). Compute \( \kappa \) such that \( 1/(1-\kappa)=\text{Average}(1/(1-\kappa(j))) \). To compare this with the effective \( \kappa \) consider, for each \( m, n \),

\[
\text{relative } \kappa \text{ error } = (1-\kappa_e)/(1-\kappa) - 1.
\]

\( \text{Error}>0 \) means the \( \kappa \) model exaggerates the strength of the evidence, \( \text{error}<0 \) means it understates.

Figure 4 shows the suitability of the \( \kappa \) model for the 27 simulated populations which systematically investigate the three population parameters mutation rate, growth rate, and population size, as well as sample size. The effects of these parameters are

---

\(^3\) “average” in the sense of geometric mean
Increasing mutation increases error. At the extreme “fast” mutation rate (17%/generation, more than triple the mutation rate of Y-filer or other haplotype systems in current forensic use) the $\kappa$ model is moderately anti-conservative – i.e. LR 30% too high – under the population simulation conditions that were used.

Increasing growth rate decreases error. At zero growth the $\kappa$ model is close to accurate. Appendix §6.2 gives some indication why this is theoretically expected. Rapid population growth implies negative error, i.e. “conservative” LR’s.

Larger population size slightly decreases error. It is fortunate that population size isn’t a significant factor, because it is by no means clear what population size is appropriate for real populations.

Sample size – no clear or significant pattern.

There are various qualities that are likely in a real population that were not modeled. The story of the Genghis Khan Y-haplotype [16] supports the plausible thesis that some Y chromosomes – or mitochondria – may be associated with superior fitness. If some
haplotypes are preferentially selected to reproduce, as is likely through hitchhiking, the consequent excess of multiply-occurring types would supply a safety margin, i.e. tend to make the $\kappa$ model more conservative.

Statistics for sample size $n=100$ are poorly defined as discussed in §5.1.1 because $\kappa=1$ so often and are omitted from the graphs. A similar explanation applies to sporadic very negative bars for $n=300$ such as for the models with $N=300,000$ and fast (17%/generation) mutation.

5.4 Matching probabilities for non-singletons

5.4.1 Matching LR for multiply observed type

The sample popularity $p$ of an allele is the number of times it is observed in a particular population sample. We are sometimes interested in the matching probability

$$y_p = \Pr(\text{innocent suspect matches crime scene type } S \mid S \text{ observed } p \text{ times in } D)$$

for $p>1$. Remember $D$ has been augmented by the crime scene type; $y_p$ generalizes $y=y_1$. From preliminary analysis of the simulated populations $\mathcal{P}$, the simple rule

$$y_p \approx py = p(1-\kappa)/n.$$  \hspace{1cm} (9)

is, empirically, about right for small $p$. It is intuitively implausible for large $p$ (for common traits sample frequency must be a good estimate for population frequency meaning that $y_p \rightarrow p/n$; i.e. the inflation factor collapses to 1 for large enough $p$), but absent selection and rapid population growth all types are rare. That’s a fact which contradicts some people’s intuition (myself for a long time for example, and [17]) and impression from older data (‘minimal’ Y-haplotypes of 7 loci). It is predicted by the neutral theory that there should be no common alleles [18]. Of course, the Y-chromosome as a whole is not neutral. No doubt it includes genes with considerable fitness variation to which the loci of the identification haplotype are hitchhikers par excellence [16]. The association is never broken by recombination, but nonetheless because of the high mutation rates for the haplotype systems here under consideration common types are not to be expected and are not found.

5.4.2 Average matching LR

We have (5):

$$y = \Pr(\text{innocent suspect matches crime scene type } S \mid S \text{ not observed in } D^c) = (1-\kappa)/n$$

and would like to estimate
\[ q = \Pr(\text{innocent suspect matches allelic type } S) \]

in general, without regard to the specific type \( S \). Therefore \( 1/q = m_e \), the \textit{effective number of types} (§6.1.3).

The most straightforward way to estimate \( m_e \) is empirical: count the pairwise matches in a population sample – formula (13). Alternatively there are several theoretical approaches. Assuming the infinitely many neutral alleles model (§18), Appendix, §6.2.3 \( m_e = \kappa / (1 - \kappa) \). Therefore \( q = \gamma / \kappa \). Note that the empirical formula uses all the \( \alpha_p \) but the theoretical formula seems to use only \( \alpha \). If they tend to agree it must be because the infinite alleles model implies a relationship among the \( \alpha_p \) and the infinite alleles model approximately holds. Another way is to assume a relationship among the \( \alpha_p \); suppose (9). Then \( q = 1/m_e \) can be written as a weighted average of the \( \gamma_p \). Appendix §6.1.3 carries out this calculation and the consequence, (15), is \( 1/\gamma \approx m_e + n - 1 \). This formula has a simple intuitive meaning. Suppose that somehow – by magic or from memory – we know \( m_e \) but lack any reference sample \( D^- \). Then the evidentiary significance of a suspect type \( T \) matching the crime scene type \( S \) is the matching likelihood ratio \( 1/\gamma = m_e \) which indeed corresponds to formula (15) with an empty database \( D^- \), whose size \( n-1 = 0 \). If we then sequentially observe reference types each of which does not match \( S \), the effect of each observation is to increment the likelihood ratio \( 1/\gamma \) by one.

5.5 Comparison with other methods

Other methods for presenting rare haplotype evidence have been suggested.

5.5.1 Haplotype survey

“Haplotype surveying” is an idea proposed in [8] to estimate evidentiary significance for Y-haplotypes. In contrast to my approach, the name (sequence) of a haplotype is assumed to contain information. The model assumption is that haplotypes which are near neighbors in mutations steps are likely to have similar population frequencies. A common allele begets numerous copies of its neighbors at every generation, so the neighbors are also common. However for haplotype systems nearly all types are rare and certainly the types of interest are the rare ones. The frequencies of rare haplotypes are relatively far more affected by drift than by mutation because the effect of drift – essentially sampling variation – is to change the number of instances \( \epsilon \) of a trait by something like \( \sqrt{\epsilon} \), which is a large percentage change when \( \epsilon \) is small. Therefore a haplotype surveying approach does not seem founded on a plausible model. It may be roughly equivalent to guessing \( \gamma = 1/m_e \) ± random variation.
5.5.2 The “FBI method”

If the population frequency of a trait is more than $3/n$, then the probability that it will be unobserved in a sample of size $n$ is about 5%. Sometimes this inference is illogically inverted to claim that when the sample frequency is $0/n$ the population frequency is 95% to be in the range $[0/n, 3/n]$. Treating $3/n$ as the matching probability amounts to offering $\text{LR}_F = n/3$ as the matching likelihood ratio. Compared to the model (6), this FBI method understates the evidence by a factor of $\text{LR}_k/\text{LR}_F = 3/(1-\kappa)$ or roughly from 7.5 to 30 or more for current systems and data. It has no logical underpinning— it doesn’t correspond to any model. If a very conservative approach is deemed desirable, then formula (1) has the advantage of being logical and simple, and incidentally squanders three times less evidence than the FBI method.

5.6 Comment – Probability vs. Frequency

“Probability” means the long-run success rate of some conceptually repeatable experiment (“trial”). The starting conditions are the same for every trial, and consist of fixing whatever information is known to the experimenter. Probability is therefore a summary of the data that is available to the experimenter. Population frequency on the other hand is unknown so obviously cannot be the probability of anything. Traditionally the forensic community answers the matching probability question by consulting a population sample, typically of several hundred or thousand people. The sample frequency is then taken as an estimate for population frequency which in turn is used as a surrogate for probability. As a result it has become habit in the forensic community to conflate frequency and probability. There is an institutionalized misconception that population frequency is matching probability. For common traits such as an allele at an STR locus the frequency approximation is simple and is reasonable enough. But frequency isn’t the goal, probability is, and if matching probability can be decided without the distraction and detour of considering population frequency, so much the better— especially for rare traits for which population frequency can’t be accurately estimated.

Reviewing the logic for step C (4) of the $\kappa$ model derivation is the clearest way to see that population frequency is irrelevant to the matching probability question. We assume, at point C in the logic, that the innocent suspect has type $T$ and that $T$ coincidentally matches one of the $\alpha$ singletons $S_0, ..., S_i, ..., S_{\alpha-1}$. Which one? The available data is identical for the various $S_i$. They all come from crimes or circumstances with which the suspect is equally unconnected, and they all have been observed only once. Given our stipulation that names don’t matter the subscripts could as well be permuted at random. Can it be more obvious that the probability for $T$ to match each of the $S_i$ is the same – therefore $1/\alpha$? Of course the singletons have different population frequencies but that, and confidence intervals, would
only be relevant if we planned to bet on what the population frequency of \( S_0 \) is, which we do not. We plan to bet on \( T \) matching \( S_0 \) (given that \( T \) matches some singleton), and the chance to win that bet is exactly \( \frac{1}{\alpha} \), neither more nor less.

5.7 Conclusion

The fundamental question to decide the evidentiary significance of a trait linking suspect to crime is not one of frequency but of probability: What is the probability for such a match to happen by coincidence when the suspect is innocent? Of JS Mill’s description that probability of an event “means the degree of expectation of its occurrence, which we are warranted in entertaining by our present evidence” [19] the last two words are worth particular emphasis. It is not relevant what our probability estimate would be if we had different population data than we have (which is the motive for confidence intervals), any more than makes sense to speculate about alleles in loci that have not been tested. The evidentiary strength of the match is a summary of present evidence.

A second key insight is that from the perspective of an innocent suspect (the key perspective), the crime scene trait as a datum stands on equal footing with the reference types in the “database” or population sample. Hence it should be considered as part of the sample. In the language of statistics, you must condition on the crime scene observation.

Third is the fact that, contrary to ordinary experience, sample frequency may not be a good indicator of matching probability (or of population frequency for that matter). In particular, it’s necessary to consider the entire sample, not just the one matching trait, and if the sample has predominately unique types then their sample frequency overstates their matching probability – the “inflation factor”.

Four versions of matching likelihood ratio for matching to a previously unseen trait have been developed in this paper, three of which may be useful depending on circumstance and practicalities. The best estimate is the \( \kappa \) model \( (6) \), that \( LR_\kappa = \frac{n}{1-\kappa} \). The counting formula \( (1) \) – \( LR_c = n \) – is very robust and even easier to explain because it doesn’t take advantage of the inflation factor, and can be used if the highly conservative value it supplies is sufficient for a

**Figure 5** comparison of matching LR from four different formulas by showing the inflation factors
particular forensic case. An intermediate possibility, robust and slightly conservative would be to use \[ LR = m \varepsilon \] for new haplotypes. It is certainly the right number to use absent knowing \( p \), which can happen. For large \( \kappa \) it is only a little smaller than \( LR \), and may be easier to explain. Finally although I can’t justify the \( t \) model (Appendix §6.1.1) mathematically, it is included in the graph for interest and comparison.

For the less frequent situation \( p > 1 \), (9) is an idea but not a recommendation. All we can say for now is that \( n/p < LR < n/(1-\kappa)p \).

Finally, it is worth noting that the \( \kappa \) model analysis presented here applies perfectly well to traditional STR systems. If a handful of common alleles comprise 99% of the allele population, then \( \kappa \) is near zero so take the inflation factor as 1 and the \( \kappa \) model degenerates to the counting model. For an allele observed \( p-1 \) times in a reference database \( D^- \) of size \( n-1 \), the matching probability is therefore \( p/n \).

6 Appendix

6.1 Statistical estimates

6.1.1 Maximum likelihood estimate of single trait probability

Suppose each individual in a certain population \( \Theta \) possesses one or another of \( t \) distinguishable and equally rare traits. Select a random sample \( D \) of \( n \) individuals. Let \( \kappa \) be the proportion of singletons, that is, of traits that occur exactly once in \( D \). The expected value of \( \kappa \), \( E(\kappa) \), is given by

\[
E(\kappa) = (1-1/t)^{n-1}, \text{ or } E(\kappa) \approx e^{-(n-1)/t} \text{ if } t \text{ is not small.}
\]

Proof: Let \( P(\dot{i}) \) be the probability that the \( i \)-th element in the sample is a singleton:

\[
P(\dot{i}) = (1-1/t)^{n-1}.
\]

Over a large number of \( n \)-samples, \( P(\dot{i}) \) is also the expected number of singletons at the \( i \)-th elements, per sample. These expectancies are additive (regardless that the events being summed are not independent), so

\[
E(\kappa) = \sum_{i=1,2,\ldots,n} P(\dot{i}) = nP(\dot{i}), \text{ and}
\]

\[
E(\kappa) = E(\alpha/n) = E(\alpha)/n = P(\dot{i}) = (1-1/t)^{n-1}\quad(10)
\]

as claimed. Using the fact that \( (1-1/t)^t \to e^{-1} \) (e=2.71828...) as \( t \to \infty \), \( (1-1/t)^{n-1} = (1-1/t)^{(n-1)/t} \approx e^{-(n-1)/t} \), so from (10),
E(κ) ≈ e^{(n-1)/t} \tag{11}

Q.E.D.

The value of κ can be observed in a sample. If we assume that the observed value is close to the expected value, then t can be estimated from formula (11).

κ ≈ E(κ) ≈ e^{(n-1)/t} ≈ e^{n/t}, whence

t ≈ -n/\ln κ. \tag{12}

6.1.2 Comparing the t model with the κ model

Suppose that true state of nature is \( \mathcal{D} \). Then the correct likelihood ratio for a match would be LRₜ = t.

Perhaps from ignorance of the true state of nature we instead estimate the likelihood ratio using the κ model (6). What error would we commit?

\[
\frac{LR_\kappa}{LR_\tau} = \frac{n}{(1-\kappa)t}
\]

= - (ln \kappa)/(1-\kappa) \quad \text{(by (12))}

= [(1-\kappa)+(1-\kappa)^2/2 + ...]/(1-\kappa) \quad \text{(expanding ln \kappa in Taylor series around \kappa=1),}

i.e.

\[
LR_\kappa \approx LR_\tau \{1 + (1-\kappa)/2\}
\]

Although the above is well short of rigorous it does seem in the domain of interest – κ close to 1 – that LRₜ is not far different from LR_κ. See Figure 5. Hence even in the pathological \( t \)-model, where the assumption (7) underlying the κ model is violated to the maximum extent possible, the formula (6) is nonetheless not far off.

6.1.3 Effective number of types

The effective number of types, \( m_e \), is the likelihood ratio supporting identity when two randomly selected types match – i.e. it is the number of types which, if they were equally frequent, would provide diversity equal to the actual diversity.

The effective number of types can easily be estimated directly from the population sample \( \mathbf{D} \). Simply count the number of ways that a matching pair can be drawn from the sample and compare it to the number \( \binom{\theta}{2} \) of ways to draw any pair. A matching pair is obtained only when both members have the
same popularity $p$ and belong to the same one of the $\alpha_p$ identity cohort groups of that popularity. Hence (summation is always over $p$)

$$m_e = \binom{n}{2}/\sum\alpha_p\binom{p}{2}.$$  \hfill (13)

Thus $1/m_e$ is the probability for the type T of an innocent suspect to match a database type S absent knowing the sample popularity of S.

6.1.4 Matching LR for new type from $m_e$

When the sample popularity of S is given as $p$, then the matching probability is some different amount $\gamma_p$, 1$\leq p$, where $1/m_e$ is a weighted average of the $\gamma_p$, bigger than $\gamma_1=\gamma$ and smaller than $\gamma_p$ for large $p$. Specifically there are $p\alpha_p$ objects S in D of popularity $p$ each of which has probability $\gamma_p$ to match T, so

$$1/m_e = \sum\gamma_p p\alpha_p/n.$$  \hfill (14)

Assume the condition $\gamma_p \approx \gamma$ of equal over-representation (9). Then

$$\sum\gamma_p p\alpha_p/n = \gamma \sum p^2\alpha_p/n$$

$$= \gamma \sum (p\alpha_p + 2\binom{p}{2}\alpha_p)/n$$

$$= \gamma [n + 2\binom{n}{2}/m_e]/n$$

$$= \gamma [1 + (n-1)/m_e]$$

Substituting this into (14) gives the very elegant result

$$1/\gamma \approx m_e + n - 1.$$ \hfill (15)

Alternatively, (15) follows from the neutral theory (§6.2). On that assumption follows $\beta(0,\theta)$ is the prior probability distribution for haplotype frequencies ([20], page 116), from which (15) follows ([21] “Use of databases”) by standard properties of the beta function.

Apparently the neutral model and equal over-representation are related assumptions.

6.2 Population genetic estimates

Assume the infinitely many neutral alleles model [12] or “neutral theory” for short.
6.2.1 Effective population size and effective number of alleles

A classical formula from population genetic theory is \( m_e = 1 + 2N_e \mu \) (for haploid populations), \( N_e \) being the effective population size – the size of an ideal equilibrium population with equivalent \( m_e \), so roughly \( N_e = m_e / 2\mu \). This formula is the motivation for the example population sizes chosen for simulations. From for example the ABI U.S. Caucasian Yfiler database, \( m_e = 8883 \) \textbf{[Table 2]} and \( \mu = 17 / 300 \), hence \( N_e = 80,000 \).

6.2.2 Expected value of \( \kappa \)

Under the neutral theory \[20\] (p 94), gives

\[
\Pr(\text{ith sampled allele is previously unobserved}) = \frac{\theta}{(\theta+n)},
\]

where we can define \( \theta = m_e - 1 = 2N_e\mu \). Equating with Robbins’ result, we have

\[
\kappa = \frac{\theta}{(\theta+n)} = 1 / [1 + n / 2N_e\mu].
\]

(16)

Hence \( \kappa \) will be smaller for larger samples, and larger for larger populations or larger mutation rates. Note that the mutation rate \( \mu \) for a Y-haplotype is the combined mutation rate for all of the included loci, hence \( \mu = 17 / 300 \) per generation for the 17-locus Y-filer haplotype.

6.2.3 Value of \( \theta \)

We can solve (16) for \( \theta \):

\[
\theta = n\kappa / (1 - \kappa),
\]

(17)

or, perhaps more robustly, we could use the same relation to infer \( \theta \) from \( \kappa \) values for a range of \( n \) using the bootstrap and curve-fitting method of \textbf{Figure 2}.

Since for rare haplotypes \( m_e >> 1 \), we can say \( m_e \approx \theta \), so for situations of interest for this paper

\[
m_e = n\kappa / (1 - \kappa).
\]

(18)

7 Acknowledgments

I am indebted to Terry Speed whose broad hint in pointing me to several key references \[14, 18, 21\] years ago suggest that he anticipated the main idea of this paper. Thanks to Tom Parsons, Monty Slatkin, Keiji Tamaki, Bruce Weir and Steve Lee for encouragement and discussions. This work was supported in part by the DNA-VIEW User’s Group.
References


