Forensic Science International: Genetics 4 (2010) 281–291

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Fundamental problem of forensic mathematics—The evidential value of a rare haplotype

Charles H. Brenner [a,b,*]

[a] School of Public Health, Forensic Science Group, U.C. Berkeley, Berkeley, CA United States
[b] DNA·VIEW, 6801 Thornhill Drive, Oakland, CA 94611-1336, United States

ARTICLE INFO

Article history:
Received 16 June 2009
Received in revised form 20 October 2009
Accepted 21 October 2009

Keywords:
Haplotype
Stain matching
mtDNA
Y-haplotype
Forensic mathematics
Likelihood ratio
Matching probability

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 κ 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.

This paper hinges above all on one key insight: probability is not frequency. The common but erroneous "frequency" approach adopts population frequency as a surrogate for matching probability and attempts the intractable problem of guessing how many instances exist of the specific haplotype at a certain crime. Probability, by contrast, depends by definition only on the available data. Hence if different haplotypes but with the same data occur in two different crimes, although the frequencies are different (and are hopelessly elusive), the matching probabilities are the same, and are not hard to find.

1. Introduction

1.1. Mitochondrial DNA and 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.

1.2. 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

* Correspondence address: 6801 Thornhill Drive, Oakland CA 94611-1336. USA. Tel.: +l 510 339 1911; fax: +l 510 339 1181.
E-mail address: cbrenner@berkeley.edu.

1872-4973/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2009.10.013

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-haplotypes – 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 = x/y, where x *and* y are likelihoods

x = *Pr*(suspect matches crime scene type|suspect is the donor),
y = *Pr*(suspect matches crime scene type|suspect unconnected to the crime).

The simple view is to assume $x = 1$. That is not strictly true—one could model typing fallibility (laboratory error), heteroplasmy (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.)

1.3. Guide to this paper

For the practical forensic scientist, the essential sections of this paper, amounting to about one third of the total bulk, are Section 1.2 The problem, Section 2 Analysis which derives the main result, the likelihood ratio for matching a previously unobserved trait, Section 4.4 Conclusions and Section 4.3 Probability vs. Frequency. It is also useful to be aware of Appendix A.1 Notation which gathers together the symbols encountered in the paper.

The remainder of the paper includes validation and consideration of non-new traits in Section 3, some theoretical context in Appendix A.3, and Section 4.2 comments on rival approaches. The mathematics in the paper does not go beyond straightforward algebra and probabilties, and for the most part is relegated to appendices.

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	

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

2. Analysis

2.1. Extended database

Denote by S_0 the crime scene haplotype, and let $\mathbf{D}^-$ ("database") be a population sample of size $n - 1$. Since $\mathbf{D}^-$ is stipulated to be an appropriate and representative population sample, there's no harm in thinking of $\mathbf{D}^-$ as a collection of crime scene haplotypes from other similar but unrelated crimes. Assume that the type S_0 does not appear in $\mathbf{D}^-$. There is some collection of types that occur exactly once in $\mathbf{D}^-$. Call them *singletons* (Fig. 1), and let $\alpha_1 - 1$ be the number of singletons. To avoid double subscripts I shall generally omit the subscript 1 and write α for α_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 any two people in $\mathbf{D}^-$. The situation of concern is when $T = S_0$, when our innocent suspect is unluckily 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 $\mathbf{D}^-$. **D** is a population sample of size n, with α singletons: $S_0, S_1, S_2, \ldots, S_{\alpha-1}$. 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 **D** when there has been no crime.

2.2. A simplified approach—counting

Imagine comparing T with the n types in **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

$$\mathrm{LR} > \mathrm{LR}_c = n \tag{1}$$

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 (Section 2.4) and justifying an *inflation factor* (Section 3.1) by which (1) can be improved if we additionally take into

[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.

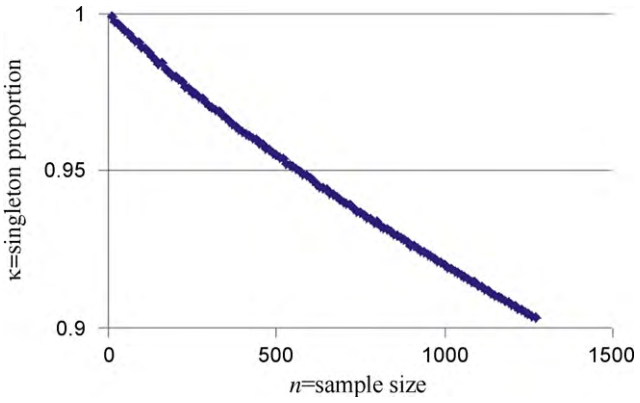

Fig. 2. Change in $\kappa(n)$, the expected fraction of once-observed types, with increasing sample size n for Yfiler® haplotypes (ABI Caucasian). Each $\kappa(n)$ is estimated by averaging the κ values for several hundred random n-subsets of the original 1272 haplotypes. The best fit to a curve of the form $\kappa(n) = \theta/(n + \theta)$ is with $\theta = 10{,}930$.

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 Section 3.2.2. 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.

2.3. Analysis of singletons

Let $\kappa = \alpha/n$ be the proportion of singletons among a sample of n haplotypes. Obviously κ 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, κ 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 κ is about twice the per-generation mutation rate.) Certainly κ changes only slowly with n (see Fig. 2). Appendix A.3 indicates why it is natural to expect that $\kappa(n)$ behaves approximately as $\theta/(n + \theta)$ for some constant θ, and interprets θ.

What is the probability that the next haplotype observed will be new? Answer: Since κ grows only slowly, it should be about the same as the probability that the last one was, which is to say κ. This is close to a theorem of Robbins [9]. It follows as an immediate corollary that about κ of the types in the population are not represented in **D**, have not been seen. That is how typical it is that a crime scene type is previously unseen.

2.4. 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 κ,

$$Pr(\text{Observed}) = 1 - \kappa. \tag{2}$$

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

An exact answer may be difficult, but an obvious claim is

$$Pr(\text{Singleton}|\text{Observed}) \approx \kappa \tag{3}$$

which is considered in more detail in Section 3.2.

C. (Match) $T = S_0$.

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

$$Pr(\text{Match}|\text{Observed \& Singleton}) = \frac{1}{\alpha}. \tag{4}$$

Ploddingly putting the above together:

$$\begin{aligned} y &= Pr(\text{Match}) \\ &= Pr(\text{Match \& Singleton \& Observed}) \\ &= Pr(\text{Match}|\text{Singleton \& Observed})Pr(\text{Singleton}|\text{Observed}) \\ &= Pr(\text{Observed}) \\ &\approx \frac{(1-\kappa)\kappa}{\alpha} \\ &= \frac{1-\kappa}{n}. \end{aligned} \tag{5}$$

Consequently the evidential strength for matching a previously unseen haplotype

$$\text{LR} \approx \text{LR}_\kappa = \frac{n}{1-\kappa} \tag{6}$$

where the subscript κ corresponds to the above analysis according the "kappa model."

3. Results and validation

3.1. 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 1.

3.1.1. 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 κ too close to 1. For the present study I prefer the latter approach.

3.2. Accuracy—sample overrepresentation

Let us consider more carefully question B, Pr(Singleton|Observed). If there are any types in the population that do not appear in the sample, they are *a fortiori* underrepresented in the sample. Therefore to compensate, types that do occur in the sample tend to be overrepresented [10]. 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 overrepresented, compared to their population frequency, as are non-singleton types. That is, let $f_{=1}$ 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 κ and $1 - \kappa$. Then $\kappa/f_{=1}$ is the overrepresentation rate for singletons and $(1 - \kappa)/f_{>1}$ is the overrepresentation rate for non-singletons. In claiming (3) we are assuming that (see Appendix A.4)

$$\frac{\kappa}{f_{=1}} \approx \frac{1-\kappa}{f_{>1}}, \tag{7}$$

Table 1
Some haplotype populations sample statistics.

	Sample size	Singletons	Singleton proportion	Inflation factor	Matching LR for new type ($1/y$)		Matching LR for any type ($m_e = 1/q$)
					κ model	t model	empirical—per pairwise match rate
	$n-1$	$\alpha-1$	$\kappa=\alpha/n$	$1/(1-\kappa)$	$LR_\kappa = n/(1-\kappa)$	$t=-n/\ln\kappa$	
Mitotyping technologies							
mtDNA	2000	~1200	~60%	2.5	~5,000	3,917	
mtDNA AFDIL							
All races	7867	4941	0.63	2.7	21,157	16,919	
Caucasians	1219	750	0.62	2.6	3,174	2,514	
7 Y loci Charité							
Caucasians	2438	1101	0.45	1.8	4,449	3,070	
Y-Plex (11 loci) Reliagene							
African American	1605	1205	0.75	4.0	6,448	5,607	1,410
Caucasian	1242	691	0.56	2.3	2,804	2,122	285
Hispanic	452	348	0.77	4.4	1,973	1,737	465
Native American	104	77	0.74	3.9	408	353	153
Yfiler® 17-plex; ABI							
African American	977	917	0.94	16.3	15,941	15,447	14,020
African	59	39	0.67	3.0	180	148	90
Asian	327	309	0.95	18.2	5,977	5,811	4,100
Caucasian	1272	1148	0.90	10.3	13,069	12,421	8,883
Filipino	105	101	0.96	26.5	2,809	2,756	2,730
SE Hispanic	597	543	0.91	11.1	6,622	6,319	5,233
Native American	106	106	>0.99[a]	107[a]	11,500[a]	11,400[a]	∞
Vietnamese	103	97	0.94	17.3	1,803	1,750	1,751
17 Y loci Gaeda							
Portuguese	313	295	0.94	17.4	5,478	5,319	2,713
10 Y loci Hammer Poland/Hungary							
Ashkenazi Jews	496	176	0.36	1.6	772	481	97
Y, 12plex Macedonia							
Caucasian	99	82	0.83	5.9	588	537	485
Yfiler® Krakow							
Caucasian	435	403	0.93	13.6	5,941	5,720	3,553

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

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

$$\frac{\kappa}{f_{=1}} \geq \frac{1-\kappa}{f_{>1}} \tag{8}$$

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.

3.2.1. 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 Ω^t 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 $\alpha = \alpha_1$ singletons may have some $\alpha_2 > 0$ doubletons and even some tripletons, i.e. duplicated types, which are overrepresented two and three times more than the singletons are overrepresented. In short, the singletons are overrepresented 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 κ model formula (6) to the Ω^t model, the expected error in using LR_κ instead of the correct $LR_t = t$ would be small (Appendix A.2.1).

3.2.2. Very popular haplotypes

At the opposite extreme consider a population $\Omega^{z,t}$ with a common type Z of substantial frequency and a large number of very rare types of frequency $1/t$. As an example, $\Omega^{0.2,\ 10{,}000}$ has a 20% type and 8000 rare types. A sample **D** from $\Omega^{0.2,\ 10{,}000}$ of size $n = 100$ rates to include nearly 80 singletons (thus $\kappa \approx 0.8$ and $f_{=1} \approx 80/10{,}000$), about 20 copies of Z and perhaps a few doubletons of rare types ($f_{>1} = 0.2 + \in$). Hence $\kappa/f_{=1} = 100$ while $(1-\kappa)/f_{>1} \approx 1$, so (8) holds with a factor of 100 to spare.

3.2.3. Probable populations

The frequency spectrum of a real haplotype systems lies in between. Constant t is unrealistic.[2]

Different lineages will have differing numbers of descendants so by random drift Ω^t is not stable. Significantly common types Z as in the $\Omega^{z,t}$ 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

[2] Providing we except extremes such as sequencing the entire Y chromosome which might realize the model Ω^N—a different type for everyone.

Table 2
spectra and singleton proportion (κ) of some Y-haplotype population samples.

population	n	$\kappa = \alpha/n$	#loci	α	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	$p > 12$
Black (US) Reliagene	2239	0.61	11	1373	187	44	28	19	5	6	3	1	1	1	1	$\alpha_{15}=1$
Caucasian Reliagene	1826	0.55	11	1005	113	44	19	11	4	1	6	2	2	1	1	α_{13}, α_{14}, α_{15}, α_{16},α_{18}, α_{19}, α_{26}, α_{29}, $\alpha_{42}=1$
Hispanic Reliagene	454	0.77	11	348	17	6	2	2	–	2	–	–	1	–	1	
Native Am Reliagene	104	0.74	11	77	8	–	1	–	–	1	–	–	–	–	–	
Portuguese	315	0.94	17	297	5	1	–	1	–	–	–	–	–	–	–	
Malaysian (Malaysia)	332	0.96	17	319	5	1	–	–	–	–	–	–	–	–	–	
Chinese (Malaysia)	331	0.89	17	295	18	–	–	–	–	–	–	–	–	–	–	
Indian (Malaysia)	305	0.94	17	287	9	–	–	–	–	–	–	–	–	–	–	
Krakow (IES)	523	0.93	17	485	19	–	–	–	–	–	–	–	–	–	–	
Ashkenazy Jewish	496	0.35	10	176	28	19	7	5	1	–	1	–	–	–	1	$\alpha_{13}=1$, $\alpha_{14}=2$, α_{15}, α_{16}, $\alpha_{18}=1$, $\alpha_{19}=2$
Macedonian	100	0.83	17	83	7	1	–	–	–	–	–	–	–	–	–	
Krakow (Jagellonian)	189	0.95	17	179	5	–	–	–	–	–	–	–	–	–	–	
ancient Mongolian	27	0.67	10	18	3	1	–	–	–	–	–	–	–	–	–	
Black (US) ABI	985	0.94	17	925	28	–	1	–	–	–	–	–	–	–	–	
African ABI	59	0.66	17	39	6	1	–	1	–	–	–	–	–	–	–	
Asian ABI	330	0.95	17	312	7	–	1	–	–	–	–	–	–	–	–	
Caucasian ABI	1276	0.90	17	1152	49	5	1	–	–	1	–	–	–	–	–	
Filipino ABI	105	0.96	17	101	2	–	–	–	–	–	–	–	–	–	–	
SE Hispanic ABI	597	0.91	17	543	22	2	1	–	–	–	–	–	–	–	–	
Native Am ABI	106	1.00	17	106	–	–	–	–	–	–	–	–	–	–	–	
Vietnamese ABI	103	0.94	17	97	3	–	–	–	–	–	–	–	–	–	–	

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.

3.3. Real and simulated validation data

3.3.1. 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 1. The Y-haplotype databases are ABI data obtained per [11], 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 its spectrum—the numbers α_1, α_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 2.

3.3.2. 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 $\Omega_1, \ldots, \Omega_{39}$ 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: infinitely many alleles (every mutation produces a brand new haplotype [12]), and modified stepwise with a specified number 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% [13], or 10%/generation,
- population size N of 30,000, 100,000 or 300,000 men,
- mutation rate (μ) 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 Yfiler® 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 is not significantly different from infinite alleles.

Within each population Ω_m we can then assess the matching probability formula that I suggest. Assessment is done by inspecting samples from the simulated populations. From each Ω_m, and for each sample size n = 100, 300, 1000, 3000, and 10,000, 100 n-samples were drawn and analyzed. Each sample $\mathbf{D} = \mathbf{D}(m,n,j)$, $j = 1, \ldots, 100$, has some collection $S_0, \ldots, S_i, \ldots, S_{\alpha-1}$, of $\alpha(m,n,j)$ $(0 \le \alpha \le n)$ distinct haplotypes each of which occur once in $\mathbf{D}$. For any particular type $S_i \in \mathbf{D}$ the *population* frequency $f_{S_i} = (1/N)\sum_{T \in \Omega_m} I(T = S_i)$ – here $I(\cdot)$ is the indicator function – can be determined by peeking into Ω_m (as would not be possible in real life). The formula for matching probability (6) – based on observable data – suggested in this paper is assessed by comparison with f_{S_i}.

A matching probability formula is valid to the extent that it has the same expected value (or in some contexts a larger, "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 Ω and sample size n, the probability it gives is a good estimate of the expected population frequency for singletons in the sample.

3.3.3. Neutrality test

Mitochondrial DNA 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 [14]. Slatkin's test [15] 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.

3.4. 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") drawn from each one (Section 3.3.2). In this section the population simulations are used as a proving ground to evaluate the κ model.

3.4.1. 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 Section 3.2.2 are impossible (also see Table 2) so LR_κ cannot be highly conservative.

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 κ. This is obvious from Fig. 2.
- The number of loci across which the mutation is distributed does not matter as long as it is large. Seventeen loci is about the same as infinite alleles.

3.4.2. Evaluation of the κ model

For a given population Ω_m and fixed sample size n, there are samples $\mathbf{D}(m,n,j) = \mathbf{D}_j$ each with some fraction $\kappa(j)$ of singletons. For each such singleton $S_i \in \mathbf{D}_j$ there is a population frequency f_{S_i}. If (6) is a good formula then on average the matching LR = $n/(1-\kappa(j)) \approx 1/f_{S_i}$, 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 $\mathbf{D}_1$, by accident of sampling, has more singletons than $\mathbf{D}_2$, 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 κ_e which satisfies $1/(1-\kappa_e) = 1/nf$. The test of the κ model (6) then consists in comparing the average of the calculated inflation factor values $1/(1-\kappa(j))$ with the ideal value $1/(1-\kappa_e)$. Compute κ such that $1/(1-\kappa) = \text{average}^3\ 1/(1-\kappa(j))$. To compare this with the effective κ consider, for each m, n,

$$\textit{relative}\ \ \kappa\ \ \textit{error} = \frac{1-\kappa_e}{1-\kappa} - 1.$$

Error > 0 means the κ model exaggerates the strength of the evidence, *error* < 0 means it understates.

Fig. 3 shows the suitability of the κ 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

- *Increasing mutation increases error.* At the extreme "fast" mutation rate (17%/generation, more than triple the mutation rate of Yfiler®) or other haplotype systems in current forensic use) the κ 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 κ model is close to accurate. Appendix A.3 gives some indication why this is theoretically expected. Rapid population growth implies negative error, i.e. "conservative" LRs.
- *Larger population size slightly decreases error.* It is fortunate that population size is not a significant factor, because it is by no means clear what population size is appropriate for real populations.
- *Sample size* has little effect on error, except for sporadic anomalous very negative errors when the sample size is small.

Those anomalous negative errors are artifactual. The artifact comes about when $\kappa = 1$ for a significant proportion of the 100 samples of size n and consequently, to avoid infinite inflation factors (see Section 3.1.1), the inflation factor is artificially limited to be at most n (which corresponds to $\alpha = n - 1$). This procedure happens to be conservative for the population models examined, as seen by the sporadic very negative bars for $n = 300$ such as for the models with $N = 300{,}000$ and fast (17%/generation) mutation. For the same reason most of the statistics for $n = 100$ are meaningless and are therefore omitted from the graphs.

Those anomalous data points excepted, the graphs visually exhibit a systematic and nearly regular pattern of change in relative error with change in one or another model parameter. Small irregularities therefore indicate the sampling variation that one would see under repeated simulation of the same population model.

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 κ model more conservative.

4. Discussion

4.1. Matching probabilities for non-singletons

4.1.1. 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

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

[3] "average" in the sense of geometric mean.

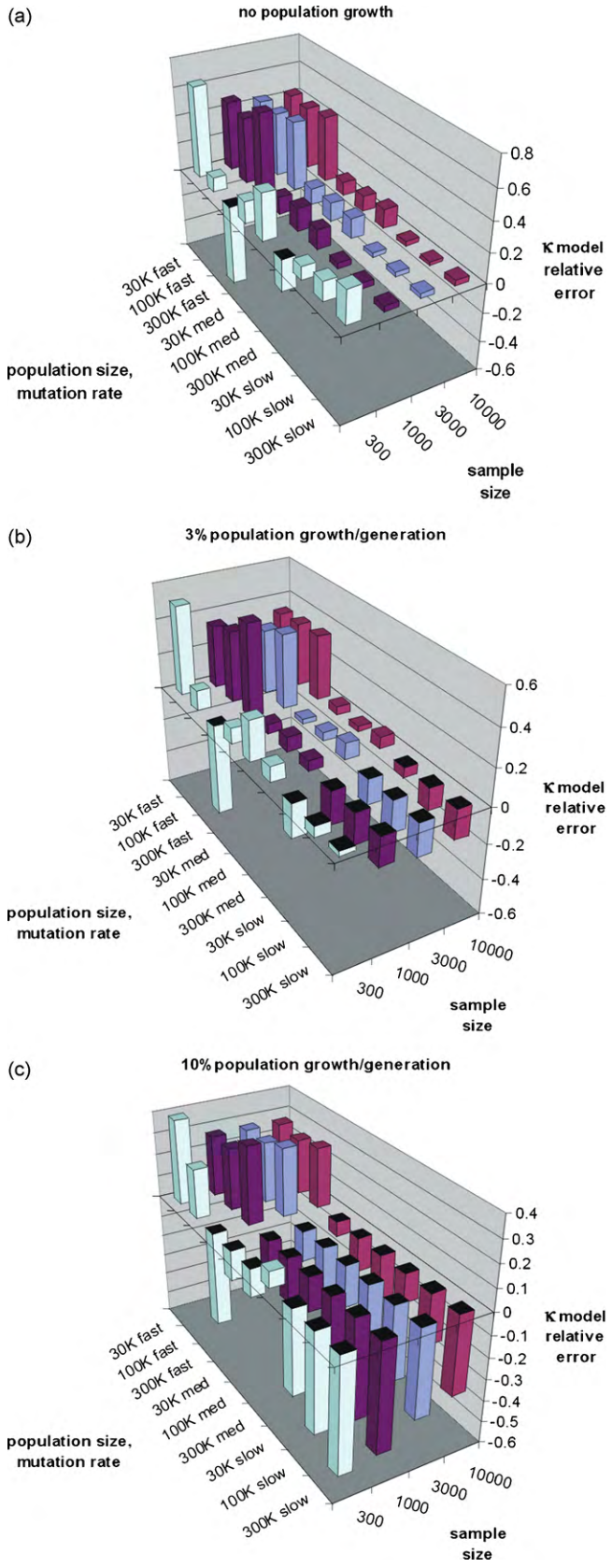

Fig. 3. Relative overstatement (if >0) of recommended LR formula (6) compared to "ideal" LR if the average population frequencies were known. Assessment of the performance of the formula under 27 (=3 × 3 × 3) models corresponding to various growth rates, N and μ per Section 3.3.2. Each bar represents an average computed over 100 samples drawn from one of the 27 simulated populations. Bars with a negative error indicate conservative performance. Omission of sample size $n = 100$ and sporadic very negative errors when $n = 300$ are discussed in the text.

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 Ω, the simple rule

$$y_p \approx py = \frac{p(1-\kappa)}{n}. \tag{9}$$

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 $y_p \to p/n$; 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.

4.1.2. Average matching LR

We have (5):

$$y = Pr(\text{innocent uspect matches crime scene type S} | \text{S not observed in } \mathbf{D}^-) = \frac{1-\kappa}{n}$$

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 *effective number of types* (Appendix A.2.3).

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, Appendix A.3.3 shows (18) $m_e \approx n\kappa/(1-\kappa)$. Therefore $q \approx y/\kappa$. Note that the empirical formula uses all the α_p but the theoretical formula seems to use only α. If they tend to agree it must be because the infinite alleles model implies a relationship among the α_p and the infinite alleles model approximately holds. Another way is to assume a relationship among the α_p; suppose (9). Then $q = 1/m_e$ can be written as a weighted average of the y_p. Appendix A.2.3 carries out this calculation and the consequence, (15), is $1/y \approx m_e + n - 1$. This formula has a simple intuitive meaning. Suppose that somehow – by magic or from memory – we know m_e but lack any reference sample $\mathbf{D}^-$. Then the evidentiary significance of a suspect type T matching the crime scene type S is the matching likelihood ratio $1/y = m_e$ which indeed corresponds to formula (15) with an empty database $\mathbf{D}^-$, whose sizen $n - 1 = 0$. 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 $1/y$ by one.

4.2. Comparison with other methods

Other methods for presenting rare haplotype evidence have been suggested.

4.2.1. 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 c of a trait by something like $\sqrt{c}$, which is a large pecrcentage change when c is small. Therefore a haplotype surveying approach does not seem founded on a plausible model. It may be roughly equivalent to guessing $y = 1/m_e \pm$ random variation.

4.2.2. The "FBI method"

If the population frequency of a trait is as much as $3/n$, 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 $LR_F = n/3$ as the matching likelihood ratio.

Compared to the κ model (6), this FBI method understates the evidence by a factor of $LR_\kappa/LR_F = 3/(1-\kappa)$ or roughly from 7.5 to 30 or more for current systems and data. It has no logical underpinning – it does not 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.

4.3. 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 certainly 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 cannot be accurately estimated.

Reviewing the logic for step C of the κ 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 α singletons $S_0, \ldots, S_i, \ldots, S_{\alpha-1}$. 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 $1/\alpha$, neither more nor less.

4.4. 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 it 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 sample frequency overstates 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 κ model (6), that $LR_\kappa = n/(1-\kappa)$. The counting formula (1) – $LR_c = n$ – is very robust and even easier to explain because it does not take advantage of the inflation factor, and can be used if the highly conservative value it supplies is sufficient for a particular forensic case. An intermediate possibility, robust and slightly conservative would be to use $LR = m_e$ (12) for new haplotypes. It is certainly the right number to use absent knowing p, which can happen. For large κ it is only a little smaller than LR_κ (Fig. 4) and may be easier to explain. Finally although I can't justify the t

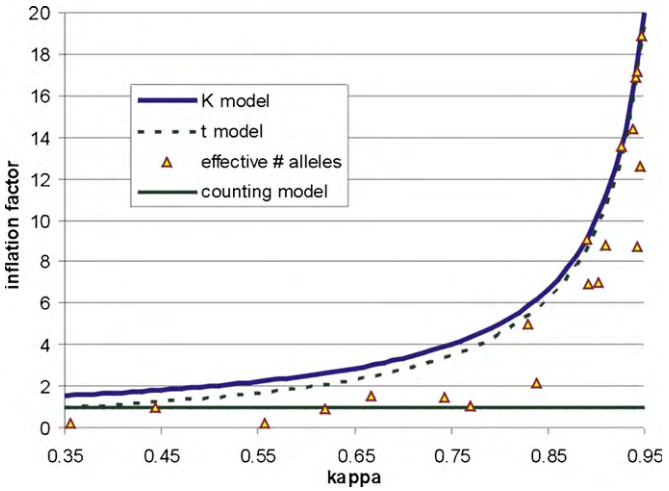

Fig. 4. comparison of matching LR from four different formulas by showing the inflation factors.

model (Appendix A.2.1) 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 $n/p < \mathrm{LR} < n/(1-\kappa)p$.

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

Acknowledgments

I am indebted to Terry Speed whose broad hint in pointing me to several key references [9,18,21] years ago suggest that he anticipated the main idea of this paper. Thanks to Tom Parsons, Monty Slatkin, Keiji Tamaki, Jim Crow, Bruce Weir and Steve Lee for encouragement and discussions. This work was supported in part by the DNA-VIEW User's Group.

Appendix A

A.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
Ω	denotes a population of haplotypes. $\Omega_1,\ldots$, $\Omega_m, \ldots$ are various simulated populations
Ω^t	population of t equally rare traits
$\Omega^{z,t}$	population of a common trait of frequency z and rare traits of frequency $1/t$
n	sample size
$\mathbf{D}$	a sample drawn from some Ω. $\mathbf{D}^-$ is a sample before including the crime stain
p	*popularity* of an allele—number of times it occurs in a sample
α	number of once-observed haplotypes (singletons) in a sample
$\alpha_1, \alpha_2, \ldots \alpha_p, \ldots$	the number of singletons, doubletons, ... haplotypes of popularity p, in a sample ($\alpha_1 = \alpha$)
κ	sample proportion α/n of singletons in a sample
f_S	population frequency of type S
$f_{=1}, f_{>1}$	total population proportion of the haplotypes that are singletons, non-singletons, 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
μ	mutation rate per generation for a haplotype system

A.2. Statistical estimates

A.2.1. Maximum likelihood estimate of single trait probability

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

$$E(\kappa) = \left(1 - \frac{1}{t}\right)^{n-1}, \quad \text{or} \quad E(\kappa) \approx e^{-(n-1)/t} \text{ if } t \text{ is not small.}$$

Proof Let $P(i)$ be the probability that the ith element in the sample is a singleton:

$$P(i) = \left(1 - \frac{1}{t}\right)^{n-1}.$$

Over a large number of n-samples, $P(i)$ is also the expected number of singletons at the ith elements, per sample. These expectations are additive (regardless that the events being summed are not independent), so

$$\begin{aligned} E(\alpha) &= \sum_{i=1,2,\ldots,n} P(i) = nP(i), \text{ and} \\ E(\kappa) &= E\left(\frac{\alpha}{n}\right) = \frac{E(\alpha)}{n} = P(i) = \left(1 - \frac{1}{t}\right)^{n-1} \end{aligned} \tag{10}$$

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

$$E(\kappa) \approx e^{-(n-1)/t} \tag{11}$$

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).

$\kappa \approx E(\kappa) \approx e^{-(n-1)/t} \approx e^{-n/t}$, whence

$$t \approx \frac{-n}{\ln \kappa}. \tag{12}$$

A.2.2. Comparing the t model with the κ model

Suppose that true state of nature is Ω^t. Then the correct likelihood ratio for a match would be $\mathrm{LR}_t = t$.

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?

$$\begin{aligned} \frac{\mathrm{LR}_\kappa}{\mathrm{LR}_t} &= \frac{n}{(1-\kappa)t} \\ &\approx -\frac{\ln \kappa}{1-\kappa} \text{ (by (12))} \\ &= \left[(1-\kappa) + \frac{(1-\kappa)^2}{2} + \cdots\right] / (1-\kappa) \\ &\qquad \text{(expanding } \ln \kappa \text{ in Taylor series around } \kappa = 1) \end{aligned}$$

i.e.

$$\mathrm{LR}_\kappa \approx \mathrm{LR}_t \left[1 + \frac{1-\kappa}{2}\right].$$

Although the above is well short of rigorous it does seem in the domain of interest – κ close to 1 – that LR_κ is not far different from LR_t. See Fig. 4. 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.

A.2.3. 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 **D**. Simply count the number of ways that a matching pair can be drawn from the sample and compare it to the number $\binom{n}{2}$ 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 α_p identity cohort groups of that popularity. Hence (summation is always over p)

$$m_e = \binom{n}{2} \Big/ \sum \alpha_p \binom{p}{2}. \tag{13}$$

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.

A.2.4. 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 y_p, $1 \le p$, where $1/m_e$ is a weighted average of the y_p, bigger than $y_1 = y$ and smaller than y_p for large p. Specifically there are $p\alpha_p$ objects S in **D** of popularity p each of which has probability y_p to match T, so

$$\frac{1}{m_e} = \sum \frac{y_p\, p\alpha_p}{n}. \tag{14}$$

Assume the condition $y_p \approx py$ of equal overrepresentation (9). Then

$$\begin{aligned}\sum \frac{y_p\, p\alpha_p}{n} &\approx y \sum \frac{p^2 \alpha_p}{n} \\ &= y \sum \left[p\alpha_p + 2\binom{p}{2}\alpha_p \right] / n \\ &\qquad (\text{now use } n = \textstyle\sum p\alpha_p \text{ and } (13)) \\ &= y \left[n + 2\binom{n}{2}/m_e \right] / n \\ &= y \left[1 + \frac{n-1}{m_e} \right].\end{aligned}$$

Substituting this into (14) gives the elegant result

$$\frac{1}{y} \approx m_e + n - 1. \tag{15}$$

Alternatively, (15) follows from the neutral theory (Appendix A.3). The neutral theory imples that $\beta(0,\theta)$ is the prior probability distribution for haplotype frequencies ([20], p. 116), from which (15) follows ([21] "Use of databases") by standard properties of the beta function.

Apparently the neutral model and equal overrepresentation are related assumptions.

A.3. Population genetic estimates

Assume the infinitely many neutral alleles model [14] or "neutral theory" for short.

A.3.1. Effective population size and effective number of alleles

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

A.3.2. Expected value of κ

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

$$Pr(n\text{th sampled allele is previously unobserved}) = \frac{\theta}{\theta + n},$$

where we can define $\theta = m_e - 1 \approx 2N_e\mu$. Equating with Robbins' result, we have

$$\kappa \approx \frac{\theta}{\theta + n} \approx \frac{1}{1 + n/2N_e\mu}. \tag{16}$$

Hence κ will be smaller for larger samples, and larger for larger populations or larger mutation rates.

Note that the mutation rate μ for a Y-haplotype is the combined mutation rate for all of the included loci, hence $\mu \approx 17/300$ per generation for the 17-locus Yfiler® haplotype.

A.3.3. Value of θ

We can solve (16) for θ:

$$\theta \approx \frac{n\kappa}{1 - \kappa}, \tag{17}$$

or, perhaps more robustly, we could use the same relation to infer θ from κ values for a range of n using the bootstrap and curve-fitting method of Fig. 2.

Since for rare haplotypes $m_e \gg 1$, we can say $m_e \approx \theta$, so for situations of interest for this paper

$$m_e \approx \frac{n\kappa}{1 - \kappa}. \tag{18}$$

A.4. Condition of equal overrepresentation

In Section 3.2 I claim that the condition (3), $Pr(\text{Singleton}|\text{Observed}) = \kappa$, is equivalent to the condition of equal overrepresentation (7) that $\kappa/f_{=1} = (1-\kappa)/f_{>1}$. These claims are relative to a fixed sample **D** and a haplotype T randomly selected from the general population. By way of proof, note that $f_{=1} = Pr(\text{Singleton}) = Pr(\text{Singleton \& Observed})$ while $f_{>1} = Pr(\sim\text{Singleton \& Observed})$. So (7) is

$$\frac{\kappa}{Pr(\text{Singleton \& Observed})} = \frac{1-\kappa}{Pr(\sim\text{Singleton \& Observed})} \quad \text{or}$$

$$\frac{\kappa}{Pr(\text{Observed})Pr(\text{Singleton}|\text{Observed})} = \frac{1-\kappa}{Pr(\text{Observed})Pr(\sim\text{Singleton}|\text{Observed})} \quad \text{or}$$

$$\frac{\kappa}{Pr(\text{Singleton}|\text{Observed})} = \frac{1-\kappa}{1 - Pr(\text{Singleton}|\text{Observed})},$$

which is true if and only if (3), Q.E.D.

References

[1] M. Prinz, Advantages and disadvantages of Y-short tandem repeat testing in forensic casework, Forensic Sci. Rev. 15 (2003) 189–196.
[2] K.A. Mayntz-Press, et al., Y-STR profiling in extended interval (> or =3 days) postcoital cervicovaginal samples, J. Forensic Sci. 53 (2) (2008) 342–348.
[3] S. Malsom, et al., The prevalence of mixed DNA profiles in fingernail samples taken from couples who co-habit using autosomal and Y-STRs, Forensic Sci. Int. Genet. 3 (2009) 57–62.
[4] M. Kayser, et al., Evaluation of Y-chromosomal STR's: a multi-center study, Int. J. Legal Med. 110 (1997) 125–133.
[5] M. Jobling, A. Pandya, C. Tyler-Smith, The Y chromosome in forensic analysis and paternity testing, Int. J. Legal Med. 110 (1997) 118–124.

[6] P. Ivanov, et al., Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II, Nat. Genet. 12 (1996) 417–420.
[7] C. Keyser-Tracqui, E. Crubézy, B. Ludes, Nuclear and mitochondrial DNA analysis of a 2,000-year-old Necropolis in the Egyin Gol Valley of Mongolia, Am. J. Hum. Genet. 73 (2003) 247–260.
[8] L. Roewer, et al., A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (2000) 31–43.
[9] H.E. Robbins, Estimating the total probability of the unobserved outcomes of an experiment, Ann. Math. Stat. 39 (1968) 256–257.
[10] N. Morton, Genetic structure of forensic populations, Proc. Natl. Acad. Sci. U.S.A. 89 (1992) 2556–2560.
[11] Applied Biosystems, Yfiler® haplotype database, http://www.appliedbiosystems.com/yfilerdatabase/.
[12] J. Crow, Twenty-five years ago in genetics: the infinite allele model, Genetics 121 (1989) 631–634.
[13] M. Slatkin, G. Bertorelle, The use of intraallelic variability for testing neutrality and estimating population growth rate, Genetics 158 (2) (2001) 865–874.
[14] M. Kimura, J. Crow, The number ofalleles that can be maintained in a finite population, Genetics 49 (1964) 725–738.
[15] M. Slatkin, An exact test for neutrality based on the Ewens sampling distribution, Genet. Res. 64 (1) (1994) 71–74.
[16] T. Zerjal, et al., The genetic legacy of the Mongols, Am. J. Hum. Genet. 72 (3) (2003) 717–721.
[17] A. Veldman, Evidential Strength of Y-STR Haplotype Matches in Forensic DNA Casework, Universiteit Leiden, 2007.
[18] W. Ewens, The sampling theory of selectively neutral alleles, Theor. Popul. Biol. 3 (1972) 87–112.
[19] J. Stuart Mill, A System of Logic: Ratiocinative and Inductive; Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, vol. 2, 1868, London, Longmans, Green, Reader, and Dyer, p. 62, http://www.questia.com/PM.qst?a=o&d=5774540 (Chapter 18).
[20] W. Ewens, Mathematical Population Genetics: I. Theoretical Introduction, 2nd ed., Springer, 2003.
[21] A. Dawid, J. Mortera, Coherent analysis of forensic identification evidence, J. R. Stat. Soc. B 58 (2) (1996) 425–443.