CHAPTER 11

Statistical Interpretation: Evaluating the Strength of Forensic DNA Evidence

It would not be scientifically justifiable to speak of a match as proof of identity in the absence of underlying data that permit some reasonable estimate of how rare the matching characteristics actually are.

—NRC II, p. 192

In Chapter 10 on short tandem repeat (STR) data interpretation, we concluded with a section entitled 'To match or not to match: That is the question.' If a DNA profile from a suspect does not match the evidence from a crime scene (and the testing has been performed properly), then we can reliably conclude that the individual in question did not contribute the biological sample recovered from the crime scene.

However, the more interesting outcome of a DNA profile comparison is what to conclude when the profiles between suspect and evidence match. Are they from the same individual or is there someone else out there who might just happen to match the evidence in question? Since we do not have the luxury of access to DNA profiles of everyone living on planet Earth, we must use smaller population data sets to extrapolate the possibility of a random match.

To estimate this match probability, allele frequencies are collected from various ethnic/racial sample sets. Based on their allele frequencies from validated databases, population genetic principles are applied to infer how reasonable it is that a random, unrelated individual could have contributed the DNA profile in question (Figure 11.1).

It is important to distinguish between unrelated and related individuals in assumptions being made for the calculations that follow. Obviously related individuals have DNA profiles that are more similar than unrelated individuals who are compared. In most equations that will be used in this chapter,

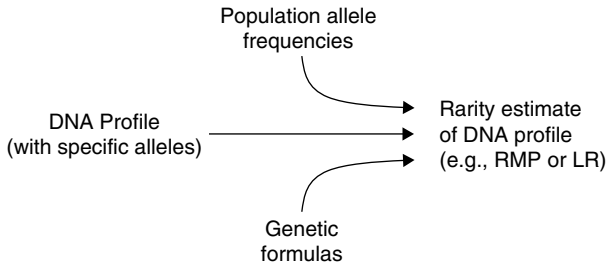

FIGURE 11.1

The rarity estimate for a specific DNA profile (in the form of a random match probability, RMP, or likelihood ratio, LR) is determined based on the alleles present in the profile, the population allele frequencies used, and genetic formulas that account for population substructure or degree of relatedness.

we will be assuming that unrelated individuals are involved. In Chapter 17, applications are considered where assumptions are made that individuals are related, such as paternity testing and disaster victim identification using biological relatives.

Of the three possible outcomes of a DNA test—'exclusion,' 'inconclusive,' or 'inclusion' between samples examined—only the third requires statistics. Statistics attempt to provide meaning to the match. These match statistics are usually provided in the form of an estimate of the random match probability or, in other words, the frequency for the particular genotype (DNA profile) in a population. However, different approaches may be taken in stating rarity of a match between a questioned (Q) and known (K) sample.

This chapter discusses basic principles that are important when considering statistical interpretation of forensic DNA profiles and estimating the rarity of a particular set of STR alleles. Example equations are examined and discussed using the population allele frequency information contained in Table 11.1. The goal is not to perform an in-depth examination of each genetic and statistical principle but rather to keep things in an understandable format for beginners in the field. Those readers desiring more extensive information on the topics discussed within this book may refer to additional references listed at the end of this chapter or to the accompanying volume, *Advanced Topics in Forensic DNA Typing, 3rd Edition*. Appendix 3 at the back of this book also covers basic principles of probability and statistics.

POPULATION DATA

To assess how common or rare a particular allele and allele combination are, data is gathered from representative groups of individuals. It is possible to run a small subset of the population and reliably predict allele and genotype

Table 11.1 STR allele frequencies from Caucasian and African American U.S. population samples.

Allele	Caucasian (*N* = 302)	African American (*N* = 258)
CSF1PO		
7	—	0.05253
8	0.00497*	0.06031
9	0.01159	0.03696
10	0.21689	0.25681
11	0.30132	0.24903
12	0.36093	0.29767
13	0.09603	0.03696
14	0.00828	0.00973
FGA		
18	0.02649	0.00194*
18.2	—	0.01163
19	0.05298	0.06202
20	0.12748	0.05620
21	0.18543	0.11628
22	0.21854	0.19574
22.2	0.01159	0.00388*
23	0.13411	0.17054
23.2	0.00331*	0.00194*
24	0.13576	0.12209
25	0.07119	0.12403
26	0.02318	0.08140
27	0.00331*	0.02326
28	—	0.01163
TH01		
6	0.23179	0.12403
7	0.19040	0.42054
8	0.08444	0.19380
9	0.11424	0.15116

(Continued)

Table 11.1 Continued

Allele	Caucasian ($N = 302$)	African American ($N = 258$)
9.3	0.36755	0.10465
10	0.00828	0.00194*
TPOX		
6	0.00166*	0.10078
7	—	0.01744
8	0.53477	0.37209
9	0.11921	0.17829
10	0.05629	0.08915
11	0.24338	0.21899
12	0.04139	0.02132
VWA		
14	0.09437	0.07752
15	0.11093	0.18605
16	0.20033	0.24806
17	0.28146	0.24225
18	0.20033	0.15504
19	0.10430	0.06202
20	0.00497*	0.01550
D3S1358		
14	0.10265	0.08915
15	0.26159	0.30233
16	0.25331	0.33527
17	0.21523	0.20543
18	0.15232	0.06008
19	0.01159	0.00388*
D5S818		
8	0.00331*	0.04845
9	0.04967	0.03876
10	0.05132	0.06977
11	0.36093	0.23256

Table 11.1 Continued

Allele	Caucasian (*N* = 302)	African American (*N* = 258)
12	0.38411	0.35271
13	0.14073	0.23837
14	0.00662*	0.01550
D7S820		
7	0.01821	0.01550
8	0.15066	0.23643
9	0.17715	0.10853
10	0.24338	0.33140
11	0.20695	0.20349
12	0.16556	0.08721
13	0.03477	0.01357
D8S1179		
8	0.01159	0.00194*
9	0.00331*	0.00581*
10	0.10099	0.02907
11	0.08278	0.04457
12	0.18543	0.14147
13	0.30464	0.21705
14	0.16556	0.30039
15	0.11424	0.18411
16	0.03146	0.06977
D13S317		
8	0.11258	0.03295
9	0.07450	0.03295
10	0.05132	0.02326
11	0.33940	0.30620
12	0.24834	0.42442
13	0.12417	0.14535
14	0.04801	0.03488
15	0.00166*	—

(Continued)

Table 11.1 Continued

Allele	Caucasian (*N* = 302)	African American (*N* = 258)
D16S539		
8	0.01821	0.03876
9	0.11258	0.19574
10	0.05629	0.11628
11	0.32119	0.31783
12	0.32616	0.19574
13	0.14570	0.11822
14	0.01987	0.01744
D18S51		
10	0.00828	0.00584*
11	0.01656	0.00195*
12	0.12748	0.07782
13	0.13245	0.05253
14	0.13742	0.07198
15	0.15894	0.16148
16	0.13907	0.15759
17	0.12583	0.15175
18	0.07616	0.12257
19	0.03808	0.09922
20	0.02152	0.06420
21	0.00828	0.00973
22	0.00828	0.00584*
D21S11		
27	0.02649	0.07752
28	0.15894	0.25775
29	0.19536	0.19767
30	0.27815	0.17442
30.2	0.02815	0.00969
31	0.08278	0.08140
31.2	0.09934	0.04651
32	0.00662*	0.00775*

Table 11.1 Continued

Allele	Caucasian (N = 302)	African American (N = 258)
32.2	0.08444	0.05814
33	0.00166*	0.00581*
33.2	0.02649	0.03488
34	—	0.00581*
34.2	0.00497*	—
35	0.00166*	0.02326
36	—	0.00969

Reported in Butler, J.M., et al. (2003). Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. Journal of Forensic Sciences, 48, *908–911; Genotypes for the individuals used to generate these allele frequencies are available on STRBase at http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm*
Note: Allele frequencies denoted with an asterisk (*) are below the 5/2N minimum allele threshold recommended by the NRC II report. Some additional rare alleles were removed to save space.

frequencies in the entire population—much like a telephone survey of several hundred individuals is used to try to predict the outcome of a political election. The key is collecting information from enough individuals to reliably estimate the frequency of the major alleles for a genetic locus.

The primary goal of generating a population database is to find all 'common' alleles and sample these alleles multiple times in order to reliably estimate the frequency of alleles present in the population under consideration. It is worth noting that some alleles, particularly variant alleles, have only been observed a few times and are therefore rather rare. Table 11.1 lists allele frequencies for U.S. Caucasian and African American populations at the 13 core STR loci used in the United States. These allele frequencies are used throughout the book in calculations made for various purposes.

Generating a population database

The primary steps in generating and testing a population database are illustrated in Figure 11.2. A laboratory must first decide on the number of samples that will be tested and what particular ethnic/racial groups are relevant to estimating DNA profile frequencies that might be encountered by the lab. Population databases are often generated by gathering a set of biological samples in the form of liquid blood from a local hospital or blood bank. Usually the individuals selected are healthy and, it is hoped, unrelated to one another so that they reliably represent the population of interest. These 'convenience' samples are deemed reliable since they are similar to other data sets from

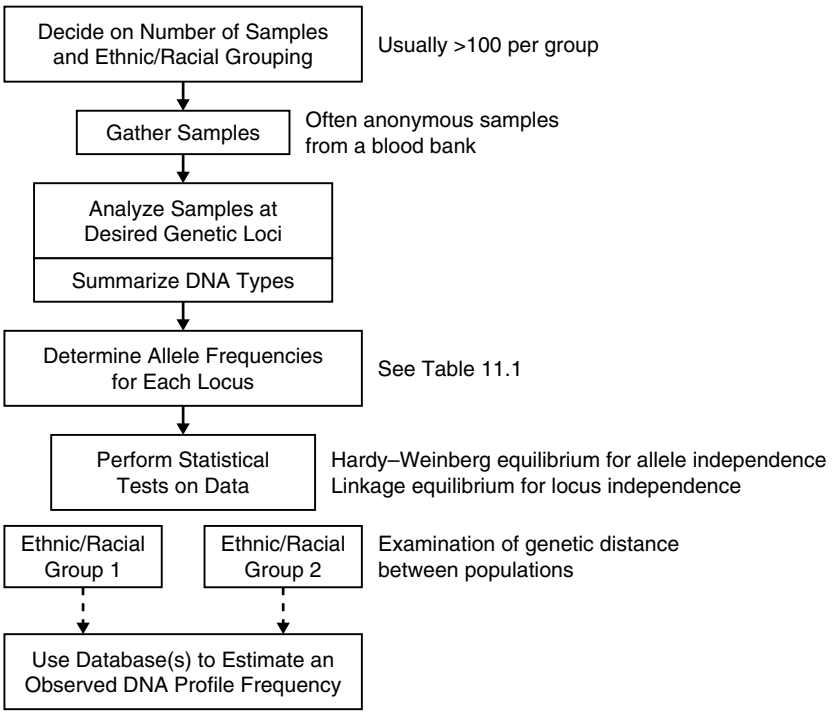

FIGURE 11.2

Steps in generating and validating a population database that can then be used to estimate the frequency of an observed DNA profile in the population.

similar population groups. Usually the individual samples are devoid of identifiers that could be used to link the DNA typing results back to the donor.

After the samples have been gathered, they are extracted, PCR-amplified, and genotyped at the STR loci of interest, such as the 13 core loci used in the FBI's Combined DNA Index System (CODIS). These single-source samples are typically processed using commercial STR kits and standard interpretation guidelines to designate alleles.

Following the gathering of the genotype data, the information is converted into allele frequencies by counting the number of times each allele is observed. D.N.A. Box 11.1 shows an example of allele counting with the STR locus D13S317 used to determine the Caucasian data in Table 11.1.

Allele frequency information allows for more compact data storage and enables Hardy–Weinberg equilibrium testing. Typically the sample genotypes and allele frequencies associated with a particular ethnic/racial group are segregated to enable both intragroup and intergroup comparisons.

D.N.A. Box 11.1 Converting Collected STR Genotypes into Observed Allele Frequencies

Following the gathering of the genotype data at each STR locus, the information is converted into allele frequencies by counting the number of times each allele is observed. The table below shows an example of allele counting for the locus D13S317 used to determine the Caucasian data in Table 11.1.

The observed alleles, ranging from 8 to 15 repeats, are listed across the top and down the left side. At the intersection of the rows and columns, the numbers of observed genotypes are listed. For example, starting in the top left-hand corner, the genotype 8,8 is seen 9 times in the set of 302 individuals examined, while the genotype 11,14 is seen 12 times. On the right side, the numbers of observed alleles are counted by summing the row and column containing the allele of interest. Thus, the number of chromosomes containing allele 8 is equal to 68 from 9 + 9 + 1 + 17 + 13 + 10 + 0 + 0 for the row containing allele 8 plus 9 for the column with the 8,8 genotype. The number of 8,8 genotypes is counted twice since both chromosomes contain an allele 8 at the D13S317 marker. The frequency for allele 8 is determined by dividing 68 into the total number of chromosomes, which are 604 since there are two chromosomes for each of the 302 individuals typed. Note that there is only one allele 15 observed in this study, which comes from a 10,15 genotype. The frequency for allele 15 is marked with an asterisk since it is observed only once and falls below the minimum allele frequency of $5/2N$ or 0.00828, where $N = 302$ individuals tested.

Source:
Butler, J. M., et al. (2003). Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations: All allele frequencies and genotypes. *Journal of Forensic Sciences, 48,* 908–911. Available at http://www.cstl.nist.gov/biotech/strbase/NISTpopdata/JFS2003IDresults.xls

Genotype Array	8	9	10	11	12	13	14	15		Allele Count	Observed Frequency
	8,8	8,9	8,10	8,11	8,12	8,13	8,14	8,15			
8	9	9	1	17	13	10	0	0	**8**	68	**0.11258**
		9,9	9,10	9,11	9,12	9,13	9,14	9,15			
9		1	2	15	10	4	3	0	**9**	45	**0.07450**
			10,10	10,11	10,12	10,13	10,14	10,15			
10			2	12	6	3	2	1	**10**	31	**0.05132**
				11,11	11,12	11,13	11,14	11,15			
11				37	54	21	12	0	**11**	205	**0.33940**
					12,12	12,13	12,14	12,15			
12					21	18	7	0	**12**	150	**0.24834**
						13,13	13,14	13,15			
13						7	5	0	**13**	75	**0.12417**
							14,14	14,15			
14							0	0	**14**	29	**0.04801**
								15,15			
15								0	**15**	1	**0.00166***
									Total = 604		

Sample sizes used for allele frequency estimation

Most published population data include on the order of 100 to 200 STR types per locus per population examined. In a key paper in 1992 entitled 'Sample size requirements for addressing the population genetic issues of forensic use of DNA typing,' Ranajit Chakraborty concluded that 100 to 150 individuals per population could provide an adequate sampling for a genetic locus provided that allele frequencies below 1% were not used in forensic calculations. Others have arrived at similar conclusions; namely, that 100 to 120 individuals per locus per population are sufficient for robust likelihood calculations. Collecting information from more samples usually only improves the accuracy of frequency estimates for rare alleles. Comparisons of data collected with typical population sizes versus thousands of individuals show similar allele frequency results (D.N.A. Box 11.2).

Population comparison DNA databases are often generated by individual forensic laboratories to assess variations in common local populations. This is particularly important to locales that may have an isolated population within its jurisdiction. For example, in Arizona it would be helpful to have a population database involving Native Americans such as Apaches and Navajos since they live in fairly close-knit communities within Arizona and would be expected to have different genotype frequencies compared to Caucasians or African Americans living in Arizona.

Minimum allele frequency

To obtain a reliable estimate of an allele frequency, it is important to collect more than one data point for that allele. A conservative minimum allele frequency is used to ensure that an allele has been sampled sufficiently to be used reliably in statistical tests. The 1996 National Research Council report (NRC II) states that an estimate of an allele frequency can be very inaccurate if the allele is so rare that it is represented only once or a few times in a database, and some rare alleles might not be represented at all. Thus, it is recommended that each allele be observed at least five times to be included in reliable statistical calculations. The minimum allele frequency is therefore $5/(2N)$, where N is the number of individuals sampled from a population and $2N$ is the number of chromosomes counted because autosomes are in pairs due to inheritance of one allele from one's mother and one from one's father.

When an observed allele frequency falls below the minimum allele frequency of $5/2N$, such as the D13S317 allele 15 in D.N.A. Box 11.1, then the minimum allele frequency is used instead. Thus, with the D13S317 allele 15 example, a value of 0.00828 ($5/[2 \times 302]$) would be used in allele frequency

D.N.A. Box 11.2 Comparison of STR Allele Frequencies from a 'Normal' versus a Larger Population Study

Most population data sets used in estimating STR allele frequencies come from roughly a hundred to a few hundred individuals. To demonstrate that collection of data from a few hundred individuals can provide reliable STR allele frequency estimates, comparisons of individual allele frequencies can be made to much larger data sets. In the table below, a comparison of a typical population study with a few hundred individuals (Butler et al., 2003) is made to a much larger study involving thousands of samples (Einum & Scarpetta, 2004). In both African American and Caucasian data sets, most of the allele frequencies are very similar. For example, D13S317 allele 12 in African Americans was seen in 42.9% with the 7833 sample study and 42.4% with the 258 sample study. Note that with the larger sample set, more rare alleles were observed (e.g., alleles 7 and 16).

D13S317	African American		Caucasian	
Alleles	*N* = 7833	*N* = 258	*N* = 7814	*N* = 302
7	0.0001	—	0.0003	—
8	0.0260	0.0330	0.1200	0.1126
9	0.0218	0.0330	0.0754	0.0745
10	0.0273	0.0233	0.0618	0.0513
11	0.2940	0.3062	0.3110	0.3394
12	0.4290	0.4244	0.2830	0.2483
13	0.1520	0.1454	0.1040	0.1242
14	0.0486	0.0349	0.0443	0.0480
15	0.0010	—	0.0014	0.0017
16	0.0002	—	—	—
Minimum allele frequency (5/2N)	0.0003	0.0096	0.0003	0.0083

Sources:

Butler, J. M., et al. (2003). Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *Journal of Forensic Sciences, 48,* 908–911.

Einum, D. D., & Scarpetta, M. A. (2004). Genetic analysis of large data sets of North American Black, Caucasian, and Hispanic populations at 13 CODIS STR loci. *Journal of Forensic Sciences, 49,* 1381–1385.

calculations rather than the 0.00166 actually reported in the study described in D.N.A. Box 11.1.

Sources of samples for population databases

Individuals whose DNA profiles will be used to construct a population database for allele frequency estimation purposes should be selected without prior

knowledge of genotypes at the loci under examination to ensure randomness of the samples. A frequent practice is to collect samples from blood donors or hospital volunteers. For example, the samples used to generate the STR typing data used in Table 11.1 were purchased from two different blood banks and represent anonymous blood donors with self-identified ethnicities. Well-characterized population samples with anthropological descriptions would be desirable in many cases to carefully define population groups but are not necessary to obtain valid information in forensic DNA population databases. Self-declaration of ethnicity can be a suitable method of categorizing samples on the basis of ethnicity.

Broad racial/ethnic categories are usually adequate for most forensic population databases, unless an isolated population is of interest, such as an Amish community. It is also desirable to use unrelated individuals in creating a population database in order to improve the precision of allele frequency estimates by increasing the number of independent alleles sampled.

Statistical tests on population data

Once STR genotypes have been generated from population samples, the data are typically evaluated with statistical tests to ensure that the allele frequencies are reasonable based on genetic inheritance principles (D.N.A. Box 11.3). Computer programs are used to conduct statistical tests for Hardy–Weinberg equilibrium (HWE) and linkage equilibrium in order to assess independence of alleles and loci (Figure 11.3).

With the assumption of independence, it then becomes possible to equate the overall match probability with the product of the locus-specific match probabilities. This combination of locus-specific match probabilities is referred to as the *product rule*. In other words, the match probability for the STR locus D13S317 can be combined with additional STR loci such as TH01 and D18S51 to decrease the odds of a random match to an unrelated individual.

GENETIC FORMULAS

Most of the examples that are worked in this chapter utilize the simple HWE model for the genetic formulas applied (i.e., 2pq for heterozygotes and p^2 for homozygotes). However, real-world populations often need adjustments to correct for what is known as *population substructure*.

Genetic mixing of alleles is not completely random because parents often share some common ancestry. The consequence of this nonrandom mating is that there is usually a decrease in heterozygotes and an increase in homozygotes. This population substructure can be adjusted for with the use of a correction factor referred to as theta (θ). The National Research Council (NRC II) report

D.N.A. Box 11.3 Hardy–Weinberg Equilibrium Testing

STR alleles are inherited by an individual from his or her mother and father in a Mendelian fashion and frequencies of occurrence follow a predictable pattern of probability. If two alleles A and a occur with frequencies p and q in the population, then the genotype AA (a homozygote) should occur p^2 and the genotype Aa (heterozygote) should occur with frequency 2pq. Allele frequencies are used to generate expected genotype frequencies that are then compared to the observed genotype frequencies. If observed and expected values are similar, then it is assumed that alleles within the genetic locus are stable or in other words 'in equilibrium.'

Hardy–Weinberg equilibrium (HWE) predicts the stability of allele and genotype frequencies from one generation to the next. The primary purpose in testing for HWE is to determine if alleles within a locus are independent of each other. Frequencies should not change over the course of many generations if the locus is genetically stable. However, natural populations usually violate HWE to some degree and thereby cause allele frequencies to change over time.

HWE assumes a random mating population of infinite size with no migration or mutation to introduce new alleles, which of course does not exist in real human populations. The reasons for each of these HWE assumptions are listed below:

If minor departures are seen from HWE, there is generally no major cause for concern with using a particular database. Some authors will do little more than note that there is a statistically significant departure from HWE for a particular locus in their population data set. It is important to keep in mind that there are three principal reasons for observations of major differences (departures) from Hardy–Weinberg equilibrium: (1) Parents might be related, leading to inbreeding and a higher than expected number of homozygotes; (2) population substructure; and (3) selection because persons with different genotypes might survive and reproduce at different rates.

Another purpose of performing an HWE test is to look for any indications of excess homozygosity. The primary explanation for excess homozygosity is allelic dropout or 'null alleles' (see D.N.A. Box 10.3) where only one allele is observed from a truly heterozygous individual.

Source:
National Research Council Committee on DNA Forensic Science (1996). *The evaluation of forensic DNA evidence*. Washington, DC: National Academy Press.

The Assumption	The Reason
Large population	Lots of possible allele combinations
No natural selection	No restriction on mating so all alleles have equal chance of becoming part of next generation
No mutation	No new alleles being introduced
No immigration/emigration	No new alleles being introduced or leaving
Random mating	Any allele combination is possible

entitled *The Evaluation of Forensic DNA Evidence* discusses issues that surround population structure. The NRC II report makes several recommendations for taking population substructure into account.

The NRC II Recommendation 4.1 substructure adjustments replace p^2 for homozygote calculations with $p^2 + p(1 - p)\theta$, where θ is an empirically

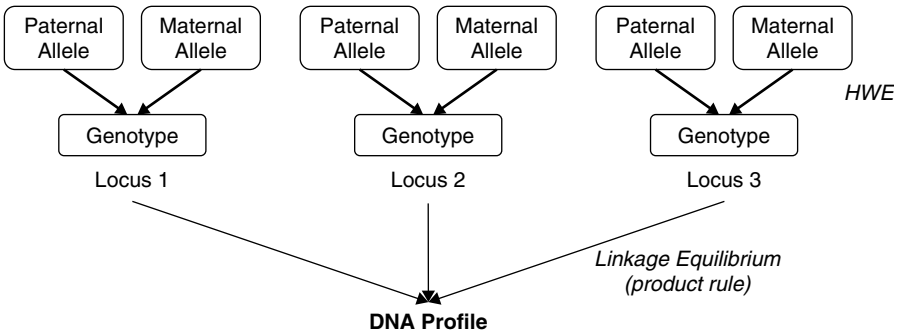

FIGURE 11.3

A DNA profile is made up of genotypes from individual genetic loci. The genotype at each locus results from inheritance of paternal and maternal alleles. Hardy–Weinberg equilibrium (HWE) tests evaluate the independence of alleles within a genetic locus, while linkage equilibrium tests ascertain the independence of alleles between loci. Independent loci and alleles enable use of the product rule.

determined measure of population subdivision. A conservative value for θ is 0.01 for typical at-large populations and 0.03 with smaller, isolated, and more inbred groups of people. A number of studies have demonstrated that $\theta = 0.01$ is a reliable and conservative estimate of population substructure with extensive population data.

The impact of these recommendations on homozygote and heterozygote frequency calculations is illustrated in Table 11.2. Examples are given with a TH01 homozygote 6,6 and a D13S317 heterozygote 11,14.

Impact of relatives on STR profile frequency estimates

If the suspect and the true perpetrator of a crime are related, then their genotype frequencies are not independent and a different calculation is required. Since STR profiles from relatives are expected to be more similar to the individual in question than a random, unrelated individual, NRC II Recommendation 4.4 covers probability calculations from various scenarios of individuals related to the suspect. These calculations are discussed in the companion volume *Advanced Topics in Forensic DNA Typing, 3rd Edition*.

If the possibility exists that a close relative of the accused had access to the crime scene and may have been a contributor of the evidence, then the best action is usually to obtain a reference sample from the relative. For example, a scenario involving a brother as a potential evidence contributor should be sufficient probable cause for obtaining a reference sample from the brother and typing it with the same STR markers as used for the evidence. This information could then be used to resolve the question of whether or not the relative carries the same DNA profile as the accused.

Table 11.2 Comparison of statistical treatment for homozygotes and heterozygotes under different assumptions.

	Under HWE	Unconditional (NRC II Recommendation 4.1)	Conditional with Substructure Adjustment
			(NRC II Recommendation 4.10a)
Homozygote	p_i^2	$p_i^2 + p_i(1-p_i)\theta$	$\frac{[p_i(1-\theta)+2\theta][p_i(1-\theta)+3\theta]}{(1+\theta)(1+2\theta)}$
TH01 **6,6** p_i = **0.23** θ = **0.01**	$(0.23)^2$ = **0.053**	$(0.23)^2$ + (0.23)(1 − 0.23)(0.01) = 0.053 + 0.0018 = **0.055**	[(0.23)(1 − 0.01) + 2(0.01)][(0.23)(1 − 0.01) + 3(0.01)] (1 + 0.01)(1 + 2(0.01)) = (0.2477)(0.2577)/(1.01)(1.02) = **0.062**
			(NRC II Recommendation 4.10b)
Heterozygote	$2p_i\,p_j$	$2p_i\,p_j$	$\frac{2[p_i(1-\theta)+\theta][p_i(1-\theta)+\theta]}{(1+\theta)(1+2\theta)}$
D13S317 **11,14** p_i = **0.34** p_j = **0.05** θ = **0.01**	2(0.34)(0.05) = **0.0340**	2(0.34)(0.05) = **0.0340**	2[(0.34)(1 − 0.01) + 0.01][(0.05)(1 − 0.01) + 0.01] (1 + 0.01)(1 + 2(0.01)) = 2(0.3466)(0.0595)/(1.01)(1.02) = **0.0400**

Note: Allele frequency values (p_i, p_j) for the TH01 and D13S317 example data are from Table 11.1 (U.S. Caucasians). Note that if θ is zero, then unconditional and conditional formulas collapse to their Hardy–Weinberg equilibrium (HWE) functions.

METHODS FOR STATING THE RARITY OF A DNA PROFILE

When a match is observed between an evidence sample (the 'unknown' or questioned sample—Q) and a reference sample (the 'known'—K), then statistical methods are typically invoked to provide information regarding the relevance of this match (see Figure 11.1). The prosecution argues that the Q and K samples have a common source, while the defense typically argues that the samples happen to match by chance. The possibility of another *unrelated* individual pulled at random from the population possessing an identical genotype can be determined by calculating the frequency with which the observed genotype occurs in a representative population database. When a DNA profile is fairly common, then it is easier to imagine that the suspect might not be connected to the crime scene. If, on the other hand, the genotype is found to be extremely rare, then the evidence is stronger that the suspect contributed the crime scene sample in question.

A number of population databases have been generated in recent years to which a DNA profile may be compared. Some basic U.S. population allele

frequencies, which are listed in Table 11.1, will be used to illustrate how profile frequencies are determined. For calculations performed in one's own laboratory, a relevant population database, usually specific to possible populations in one's local area, would be used instead.

It is important to keep in mind that methods for reporting DNA evidence vary between laboratories. Some laboratories present random match probabilities that are based on genotype frequency estimates. Another approach is to report likelihood ratios to convey relative support for the weight of DNA evidence under the hypothesis that the defendant is the source of the DNA profile versus an unrelated individual from the population at large. The FBI Laboratory has opted for a source attribution approach when random match probabilities are sufficiently rare. In the following sections, we will discuss the issues surrounding each approach and go through the statistical calculations performed with each method.

FREQUENCY ESTIMATE CALCULATIONS

DNA profile frequency estimates are calculated by first considering the genotype frequency for each locus and then multiplying the frequencies across all loci. The most effective method to understand how the probability of a random match is calculated is to work through an example.

The frequency for any DNA profile can be calculated with knowledge of the alleles from the DNA profile and allele frequencies seen in a population database. Of course, a different size database or one with different allele frequencies can result in a different expected genotype frequency for each tested locus and hence a different DNA profile frequency. It is therefore important that the database used is large enough and representative of the population of the suspect(s). Frequencies from multiple populations are typically reported to provide a range of possibilities since the population of the true perpetrator is unknown.

In Table 11.3 and Table 11.4 the DNA profile frequencies for 13 STR loci are determined using allele frequencies from the two different population groups listed in Table 11.1. One database contains allele frequencies from DNA profiles generated from 302 U.S. Caucasians or 604 measured alleles. The other set of allele frequencies in Table 11.1 comes from 258 African Americans or 516 measured alleles. The DNA profile in question contains the following alleles: 11 and 14 at D13S317 (heterozygous), 6 at TH01 (homozygous), and 14 and 16 at D18S51 (heterozygous).

In the population sample of 604 alleles (302 U.S. Caucasian individuals), allele 11 for D13S317 was observed 205 times, which is 0.33940 or approximately 34% of the time (see D.N.A. Box 11.1 and Table 11.1). In other words, we can assume that there is a 34% chance that any particular D13S317 allele

Table 11.3 Random match probability for a 13-locus STR profile using the U.S. Caucasian allele frequencies found in Table 11.1.

	Allele 1	Allele 2	Allele 1 Frequency (*p*)	Allele 2 Frequency (*q*)	Formula	Expected Genotype Frequency
D13S317	11	14	0.33940	0.04801	2pq	0.0326
TH01	6	6	0.23179		p^2	0.0537
D18S51	14	16	0.13742	0.13907	2pq	0.0382
D21S11	28	30	0.15894	0.27815	2pq	0.0884
D3S1358	16	17	0.25331	0.21523	2pq	0.1090
D5S818	12	13	0.38411	0.14073	2pq	0.1081
D7S820	9	9	0.17715		p^2	0.0314
D8S1179	12	14	0.18543	0.16556	2pq	0.0614
CSF1PO	10	10	0.21689		p^2	0.0470
FGA	21	22	0.18543	0.21854	2pq	0.0810
D16S539	9	11	0.11258	0.32119	2pq	0.0723
TPOX	8	8	0.53477		p^2	0.2860
VWA	17	18	0.28146	0.20033	2pq	0.1128
AMEL	X	Y				
Product rule						1.20×10^{-15}
Combined frequency						1 in 8.37×10^{14}

selected at random from an unrelated individual will be an 11. In the same manner, the chance for observing an allele 14 in a Caucasian population is q = 0.04801 since this allele was seen 29 times in 604 allele measurements (see D.N.A. Box 11.1 and Table 11.1).

If the individual with the 11,14 D13S317 genotype received these alleles at random from each of his parents, then the chance of receiving an 11 from his mother and a 14 from his father is pq and of receiving the 14 from his mother and the 11 from his father is another pq. With either combination possible, the probability to be 11,14 by chance is pq + pq or 2pq.

Plugging the frequency values of p = 0.33940 and q = 0.04801 into the formula 2pq ($2 \times 0.33940 \times 0.04801$) results in an estimated genotype frequency of 0.0326 or, in other words, approximately 3% of people from a Caucasian population are expected to have an 11,14 genotype at the D13S317 locus. Conducting the same analysis with an African American population database will result in a similar genotype frequency of 0.0214 or 2% (Table 11.4).

Table 11.4 Random match probability calculations for the same 13-locus STR profile shown in Table 11.3 but using the African American allele frequencies found in Table 11.1.

	Allele 1	Allele 2	Allele 1 Frequency (p)	Allele 2 Frequency (q)	Formula	Expected Genotype Frequency
D13S317	11	14	0.30620	0.03488	2pq	0.0214
TH01	6	6	0.12403		p^2	0.0154
D18S51	14	16	0.07198	0.15759	2pq	0.0227
D21S11	28	30	0.25775	0.17442	2pq	0.0899
D3S1358	16	17	0.33527	0.20543	2pq	0.1377
D5S818	12	13	0.35271	0.23837	2pq	0.1682
D7S820	9	9	0.10853		p^2	0.0118
D8S1179	12	14	0.14147	0.30039	2pq	0.0850
CSF1PO	10	10	0.25681		p^2	0.0660
FGA	21	22	0.11628	0.31783	2pq	0.1244
D16S539	9	11	0.19574	0.32119	2pq	0.0723
TPOX	8	8	0.37209		p^2	0.1385
VWA	17	18	0.24225	0.15504	2pq	0.0751
AMEL	X	Y				
Product rule						6.04×10^{-17}
Combined frequency						1 in 1.66×10^{16}

With the TH01 locus, a homozygous allele 6 was observed (Table 11.3). The same comparison of the profile's observed allele to a measured allele frequency in a population database is performed with TH01 but in this case the combined probability of inheriting allele 6 from both parents is pp or p^2 (see Figure 2.8). Since allele 6 was observed 140 times out of 604 allele measurements in U.S. Caucasians, $p = 0.23179$ and $p^2 = 0.0537$ (Table 11.3). Using the African American allele frequency for TH01 allele 6 (Table 11.1), $p = 0.12403$ and $p^2 = 0.0154$ (Table 11.4). Thus, allele is more rare in African Americans.

Since these two STR loci are on separate chromosomes (e.g., chromosome 13 for D13S317 and chromosome 11 for TH01), they will segregate independently during meiosis, allowing the genotype frequencies to be multiplied. In the case of a U.S. Caucasian population, the chance of a person having the combined genotype of 11,14 at D13S317 and 6,6 at TH01 is 5% of 3% (i.e., 0.0537×0.0326) or 0.175%.

D.N.A. Box 11.4 Names of Big Numbers with Their Corresponding Scientific Notation

10^{6}	Million	10^{39}	Duodecillion
10^{9}	Billion	10^{42}	Tredecillion
10^{12}	Trillion	10^{45}	Quattuordecillion
10^{15}	Quadrillion	10^{48}	Quindecillion
10^{18}	Quintillion	10^{51}	Sexdecillion
10^{21}	Sextillion	10^{54}	Septendecillion
10^{24}	Septillion	10^{57}	Octodecillion
10^{27}	Octillion	10^{60}	Novemdecillion
10^{30}	Nonillion	10^{63}	Vigintillion
10^{33}	Decillion	10^{100}	Google
10^{36}	Undecillion		

Source:
http://www.sizes.com/numbers/big_numName.htm

Similar calculations for the D18S51 locus with alleles 14 and 16 result in an estimated genotype frequency of 3.82% (see Table 11.3). The combined profile frequency with these three loci thus becomes 0.000067—the product of the three individual genotype frequencies ($0.0326 \times 0.0537 \times 0.0382$) or about 1 in 15,000. Note that when using the African American allele frequencies, the DNA profile frequency in this example drops to 0.0000076 ($0.0214 \times 0.0154 \times 0.0227$) or about 1 in 131,000 individuals (see Table 11.4). Because the alleles observed in the specific STR profile under consideration in Table 11.4 are rarer in an African American population, the profile has a rarer frequency estimate.

Working through the rest of the STR loci in Table 11.3 and Table 11.4, the final combined frequency estimate for the DNA profile in question is 1 in 837 trillion (8.37×10^{14}) using Caucasian allele frequencies (Table 11.3) and 1 in 16.6 quadrillion (1.66×10^{16}) using African American allele frequencies (Table 11.4).

Often the rarity of a calculated DNA profile goes beyond one in billions (10^{9}) or trillions (10^{12}) to numbers that are not frequently used because they are so large. A list of some big number names is contained in D.N.A. Box 11.4 to aid in verbal descriptions of rare DNA profiles. For example, the inverted value of 1.20×10^{-15} is 1 in 8.37×10^{14} or 0.84×10^{15} (one in 0.84 quadrillion).

D.N.A. Box 11.5 OmniPop: Calculating STR Profile Frequencies against Multiple Population Databases

The ability to determine simultaneously the frequency for a particular STR profile in multiple population databases was recently made easier with the development of a Microsoft Excel macro called OmniPop. Below the cumulative profile frequency range is calculated for the particular STR profile listed against 202 published population studies involving Profiler Plus kit loci and 120 published reports containing all 13 CODIS core loci. The cumulative profile frequencies obtained with U.S. Caucasian allele frequencies presented in the Table 11.1 data set are listed as well.

These profile frequencies were all calculated with a theta value of 0.01. When using a theta value of 0.03 as recommended by NRC II for more inbred populations, the range for the computed profile with all 13 STR loci across the 120 published population data sets is 1.19×10^{14} to 1.27×10^{21}.

It is worth noting that the computed profile is part of the U.S. Caucasian data set used to generate the allele frequencies described in Table 11.1 and thus this database would be expected to compute fairly conservative values for this particular 13-locus STR profile as demonstrated below.

STR Locus	Profile Computed	Number of Populations Used	Cumulative Profile Frequency Range (1 in …)	Cumulative Profile Frequency against U.S. Caucasians (Table 11.1)
D3S1358	16,17	202	4.53 to 62.6	9.19
VWA	17,18	202	37.6 to 1,080	81.8
FGA	21,22	202	737 to 119,000	1,010
D8S1179	12,14	202	8,980 to 5,430,000	16,400
D21S11	28,30	202	165,000 to 248,000,000	186,000
D18S51	14,16	202	3.85×10^{6} to 2.68×10^{10}	4.88×10^{6}
D5S818	12,13	202	2.28×10^{7} to 4.22×10^{11}	4.51×10^{7}
D13S317	11,14	202	4.32×10^{8} to 1.69×10^{13}	1.38×10^{9}
D7S820	9,9	202	1.17×10^{10} to 2.98×10^{16}	4.22×10^{10}
D16S539	9,11	120	3.14×10^{11} to 1.11×10^{18}	5.82×10^{11}
TH01	6,6	120	3.53×10^{12} to 1.45×10^{19}	1.05×10^{13}
TPOX	8,8	120	9.13×10^{12} to 1.54×10^{20}	3.63×10^{13}
CSF1PO	10,10	120	1.42×10^{14} to 2.65×10^{21}	7.43×10^{14}

Source:
OmniPop 200.1 was used for these calculations. Created by Brian Burritt of the San Diego Police Department and freely available at http://www.cstl.nist.gov/biotech/strbase/populationdata.htm

As more and more loci match during a Q and K sample comparison, it becomes less and less likely that an *unrelated*, random person in the population contributed the crime scene sample. Thus, either the suspect contributed the evidence or a very unlikely coincidence occurred.

Impact of various population databases

From the combined STR profile frequencies calculated in Tables 11.3 and 11.4, it is apparent that different populations can yield different frequency estimates due to variations in allele frequencies in these populations. A calculation of the same STR profile as used in the previous examples against 202 different published population databases (including the two present in Table 11.1) found that the cumulative profile frequency ranged from 1 in 3.43×10^{14} to 1 in 2.65×10^{21} (D.N.A. Box 11.5).

It is probably worth noting that the final calculated value in the far right column of D.N.A. Box 11.5 (1 in 7.43×10^{14}) differs slightly from that determined in Table 11.3 (1 in 8.37×10^{14}) due to the number of significant figures carried throughout the calculations. Thus, to obtain consistent frequency estimates with the same allele frequency information, it is essential to maintain the same significant figures between calculations.

Another source of population databases that enables an online search is the European Network of Forensic Science Institutes (ENFSI) DNA Working Group STR Population Database located at http://www.str-base.org/index.php. An estimated random match probability for a DNA profile of interest can be calculated using allele frequencies produced from 5700 profiles covering 24 European populations that have been generated with the SGM Plus loci.

General match probability

DNA profile probabilities can be calculated for a variety of scenarios. There are five different sets of people and possible relationships to a suspect: (1) the suspect's siblings, (2) his other relatives, (3) other members of his subpopulation, (4) other members of his racial group, and (5) everyone else.

Instead of having to calculate all of these case-specific match probabilities, some authors have proposed using general match probabilities that have been calculated from the theoretically most conservative method involving the most two common alleles for each locus (D.N.A. Box 11.6). The primary advantage of this approach is that repeated calculations are not required for each profile observed. Rather the general match probability is provided in court as being a very conservative estimate on the rarity of the observed DNA profile. Another reason that this approach is advocated is that some statisticians feel that it is difficult to provide any sound statistical support for probabilities of

D.N.A. Box 11.6 General Match Probability Values

In a paper performing statistical analyses to support forensic interpretation of the 10 loci present in the SGM Plus kit, Foreman and Evett (2001) advocate the use of general probability values when reporting full-matching STR profiles. With the 10 STR loci present in the SGM Plus kit used in the United Kingdom and Europe, the probabilities are shown in the table below (see Foreman & Evett, 2001, Table 4).

They argue that adoption of such figures would eliminate the need to perform case-specific match probabilities making it much easier to present information to the court. The match probabilities for specific STR profiles are typically several orders of magnitude smaller than those given above, which were calculated from the theoretically most common SGM Plus profile. Thus, these probabilities should provide a fair and reasonable assessment of the weight of DNA evidence for each category and in the end would probably be favorable to the suspect (defendant).

A similar calculation for a full match with the 13 CODIS loci using the most common alleles observed in U.S. population databases, such as Table 11.1, would result in even higher general match probability values since more STR loci are being examined.

Sources:

Balding, D. J. (1999). When can a DNA profile be regarded as unique. *Science & Justice, 39,* 257–260.

Foreman, L. A., & Evett, I. W. (2001). Statistical analyses to support forensic interpretation for a new 10-locus STR profiling system. *International Journal of Legal Medical, 114,* 147–155.

Relationship with suspect	Match probability
Sibling	1 in 10 000
Parent/child	1 in 1 million
Half-sibling or uncle/nephew	1 in 10 million
First cousin	1 in 100 million
Unrelated	1 in 1 billion

such a small magnitude (e.g., 10^{-21}) given the limited sampling that has been performed.

What a random match probability is not

It is important to realize what a random match probability is not. It is not the chance that someone else is guilty or that someone else left the biological material at the crime scene. Likewise, it is not the chance of the defendant not being guilty or the chance that someone else in reality would have that same genotype. Rather a *random match probability* is simply the estimated frequency at which a particular STR profile would be expected to occur in a population. This random match probability may also be thought of as the theoretical chance that if you sample one person at random from the population he or she will have the particular DNA profile in question.

Switching the language and meaning of a random match probability is something referred to as the *prosecutor's fallacy* or the fallacy of the transposed conditional. Statements such as 'there is only a 1 in 15,000 chance that the DNA profile came from someone else' or 'there is only a 1 in 15,000 chance that the defendant is not guilty' are examples of the prosecutor's fallacy. A correct statement would be 'the probability of selecting the observed profile from a population of random unrelated individuals is expected to be 1 in 15,000 based on the alleles present in this sample.…' Note that with a 13 STR locus-match instead of just the three used in the above example the random match probabilities are in the range of trillions, quadrillions, and beyond.

The *defense attorney's fallacy* is equally problematic where the assumption is made that everyone else with the same genotype has an equal chance of being guilty or that every possible genotype in a mixture has an equal chance of having committed the crime. Access to the crime scene, motive, and legitimate alibis all play a role in an investigation, suggesting that it is unwise to consider DNA evidence and corresponding frequency estimates in a vacuum devoid of other information. A suspect is usually under suspicion and investigation prior to his DNA profile being known, and thus the DNA results are most often used to corroborate and connect a criminal perpetrator to his crime scene rather than as the sole evidentiary material.

LIKELIHOOD RATIO

When matching STR profiles are obtained between a suspect (known sample, K) and the crime scene evidence (question sample, Q), it is necessary to quantify the evidentiary value of this match. Another approach in assessing the weight of the Q-K comparison besides the match probability profile frequency estimate just described is the use of a likelihood ratio (LR). LRs involve a comparison of the probabilities of the evidence under two alternative propositions. These mutually exclusive hypotheses represent the position of the prosecution—namely, that the DNA from the crime scene originated from the suspect—and the position of the defense—that the DNA just happens to coincidently match the defendant and is instead from an unknown person out in the population at large.

A likelihood ratio is a ratio of two probabilities of the same evidence under different hypotheses (see Appendix 3). For example, if a DNA profile generated from a crime scene evidence sample matches a suspect's DNA profile, then there are generally two possible hypotheses for why the profiles match each other: (1) the suspect matches because he left his biological sample at the crime scene or (2) the true perpetrator is still at large and just happens to match the suspect at the DNA markers examined.

Typically, the first hypothesis (and that championed by the prosecution) is placed in the numerator of the likelihood ratio while the second hypothesis—that someone else other than the defendant committed the crime (which is of course the defense's position)—is placed in the denominator. Thus, in mathematical terms:

$$\text{LR} = \frac{H_p}{H_d}$$

or verbally the likelihood ratio equals the hypothesis of the prosecution divided by the hypothesis of the defense. Since the hypothesis of the prosecution is that the defendant committed the crime, then $H_p = 1$ (assumes 100% probability). On the other hand, the hypothesis of the defense that the profile originated from someone else can be calculated from the genotype frequency of the particular STR profile. If the STR typing result is heterozygous, then this probability would be $2pq$, where p is the frequency of allele 1 and q is the frequency of allele 2 in the relevant population for the locus in question. Alternatively, for a homozygous STR type the H_d would be p^2. Therefore,

$$\text{LR} = \frac{H_p}{H_d} = \frac{1}{2pq}$$

If the STR type in question was D13S317 alleles 11 and 14, then p is 0.3394 and q is 0.04801 for the Caucasian population (Table 11.1). The likelihood ratio for the D13S317 genotype match then becomes

$$\text{LR} = \frac{H_p}{H_d} = \frac{1}{2pq} = \frac{1}{2(0.33940)(0.04801)} = \frac{1}{0.03259} = 30.7$$

If the value for a likelihood ratio is greater than one, then it provides support to the prosecution's case. If on the other hand, the LR is less than one, then the defense's case is supported. In the example shown here, if there is a match between a crime stain possessing D13S317 alleles 11 and 14 and the suspect who also possesses a D13S317 genotype of 11,14, then it is 30.7 times more likely that the suspect left the evidence than that it came from some unknown person out of the general Caucasian population.

Note that the rarer the particular STR genotype is, the higher the likelihood ratio will be since there is a reciprocal relationship. In its simplest form, an LR is the inverse of the estimated genotype frequency for each locus, and if discrete alleles and independent marker systems are utilized, then the LR is simply the inverse of the relative frequency of the observed genotype in the relevant

population. Of course, LRs can become much more complicated if mixtures or alternative scenarios for the evidence are possible. The product of all locus-specific LRs results in the full profile LR, which in the example of the Caucasian data shown in Table 11.3 comes to 8.37×10^{14} (the inverse of 1.20×10^{-15}).

When considering the strength of a likelihood ratio in terms of supporting the prosecution's position, the following guidelines have been suggested:

If likelihood ratio is…	Then the evidence provides…
1 to 10	limited support…
10 to 100	moderate support…
100 to 1000	moderately strong support…
1000 to 10,000	strong support…
10,000 or greater	very strong support…

With a 13-locus STR match likelihood ratio of 8.37×10^{14} based on a full profile with unambiguous results (e.g., no mixture present), the evidence has extremely strong support from the proposition that the suspect supplied the evidentiary sample.

SOURCE ATTRIBUTION

With average random match probabilities for unrelated individuals of less than one in a trillion using the 13 core STR loci, there comes within the context of a particular case and Q-K comparison a high degree of confidence that an individual is the source of an evidentiary DNA sample with reasonable degree of scientific certainty. When the rarity of a specific DNA profile (based on frequency estimates) exceeds a predefined threshold (D.N.A. Box 11.7), a laboratory may set a policy to declare that the Q sample can be attributed to the K reference sample. Such a declaration is known as *source attribution*.

Even with very small probabilities, it is important to keep in mind that absolute certainty is outside the realm of scientific inquiry. Yet a high degree of confidence in individualization of a DNA profile can be obtained when the rarity of a profile exceeds the world population many fold. For this reason, the FBI Laboratory adopted a policy in 2000 that when a specific profile's probability was less than a thousand times the U.S. population size, the sample's questioned source (Q sample) was attributed to the reference known (K sample).

A statement provided with a report involving a source attribution might include the following words: 'In the absence of identical twins or close

D.N.A. Box 11.7 Source Attribution

Many laboratories in the United States use a source attribution statement when a DNA profile frequency estimate exceeds a predefined threshold. The logic for this approach, as described by Budowle et al. (2000), is discussed next.

If p_x is the random match probability for a given evidentiary profile X, then $(1 - p_x)^N$ is the probability of not observing the particular profile in a sample of N *unrelated* individuals.

When this probability is greater than or equal to a $1 - \alpha$ confidence level (with α being 0.01 for 99%), then $(1 - p_x)^N \geq 1 - \alpha$ or $p_x \leq 1 - (1 - \alpha)^{1/N}$, which enables the calculation that if N is approximately the size of the U.S. population ($N = 300{,}000{,}000$), then a random match probably of less than 3.35×10^{-11} will confer at least 99% confidence that the evidentiary profile is unique in the population. The table below lists the random match probability thresholds for various population sizes and confidence levels:

Random match probability thresholds for source attribution at various population sizes and confidence levels. With a random match probability of 1.20×10^{-15} in U.S. Caucasians, the example STR profile would be considered 'unique.'

		Confidence Levels ($1 - \alpha$)			
	Sample Size (*N*)	**0.90**	**0.95**	**0.99**	**0.999**
	2	5.13×10^{-2}	$2.53 \infty 10^{-2}$	$5.01 \infty 10^{-3}$	5.00×10^{-4}
	3	3.45×10^{-2}	1.70×10^{-2}	3.34×10^{-3}	3.33×10^{-4}
	4	2.60×10^{-2}	1.27×10^{-2}	2.51×10^{-3}	2.50×10^{-4}
	5	2.09×10^{-2}	1.02×10^{-2}	2.01×10^{-3}	2.00×10^{-4}
	10	1.05×10^{-2}	5.12×10^{-3}	1.00×10^{-3}	1.00×10^{-4}
	25	4.21×10^{-3}	2.05×10^{-3}	4.02×10^{-4}	4.00×10^{-5}
	50	2.10×10^{-3}	1.03×10^{-3}	2.01×10^{-4}	2.00×10^{-5}
	100	1.05×10^{-3}	5.13×10^{-4}	1.00×0^{-4}	1.00×10^{-5}
	1000	1.05×10^{-4}	5.13×10^{-5}	1.01×10^{-5}	1.00×10^{-6}
	100,000	1.05×10^{-6}	5.13×10^{-7}	1.01×10^{-7}	1.00×10^{-8}
	1,000,000	1.05×10^{-7}	5.13×10^{-8}	1.01×10^{-8}	1.00×10^{-9}
	10,000,000	1.05×10^{-8}	5.13×10^{-9}	1.01×10^{-9}	1.00×10^{-10}
	50,000,000	2.11×10^{-9}	1.03×10^{-9}	2.01×10^{-10}	2.00×10^{-11}
U.S. (1999)	260,000,000	4.05×10^{-10}	1.97×10^{-10}	3.87×10^{-11}	3.85×10^{-12}
U.S. (2005)	300,000,000	3.51×10^{-10}	1.71×10^{-10}	3.35×10^{-11}	3.33×10^{-12}
	1,000,000,000	1.05×10^{-10}	5.13×10^{-11}	1.01×10^{-11}	1.00×10^{-12}
World pop	6,000,000,000	1.76×10^{-11}	8.55×10^{-12}	1.68×10^{-12}	1.67×10^{-13}

Source:

Budowle, B., et al. (2000). Source attribution of a forensic DNA profile. *Forensic Science Communication, 2*(3). Available at http://www.fbi.gov/hq/lab/fsc/backissu/july2000/source.htm

relatives, it can be concluded to a reasonable scientific certainty that the DNA from (Q) and from (K) came from the same individual' or 'Reasonable scientific certainty means that you are (x%) certain that you would not see this profile in a sample of (y) unrelated individuals.'

ADDITIONAL STATISTICAL CALCULATIONS

The previous sections have all focused on different approaches to providing statistical interpretation on single-source autosomal STR data. A number of other scenarios may exist, however, for which a forensic scientist might need to provide statistical support for a DNA result. Several of these other scenarios are briefly discussed below. Greater detail on these topics may be found by consulting the articles or texts cited in the reference list for this chapter. These topics are also discussed in greater detail in the *Advanced Topics of Forensic DNA Typing, 3rd Edition*.

Database match probability

As noted in Chapter 12, the development of national DNA databases filled with profiles from both convicted offenders and unsolved casework samples permits searches for matches between evidentiary and database profiles. To calculate what might be termed a *database match probability*, NRC II Recommendation 5.1 advocates that the random match probability be multiplied by N, the number of persons in the database. The FBI's DNA Advisory Board in their February 2000 recommendations on statistical approaches endorsed this NRC II report recommendation.

Mixture statistics

When mixed DNA profiles are observed, several approaches are taken for statistical evaluation of the strength of the evidence. These include likelihood ratios and combined probabilities of inclusion/exclusion. A worked example using the combined probability of exclusion can be found in D.N.A. Box 14.2. The combined probability of exclusion is sometimes referred to as the 'random man not excluded' (RMNE) approach.

The DNA Advisory Board in their 2000 recommendations on statistical approaches noted that both the probability of exclusion and likelihood ratio calculations are acceptable and recommended that 'one or both calculations be carried out whenever feasible and a mixture is indicated.' The DNA Commission of the International Society of Forensic Genetics (ISFG) in their 2006 recommendations on mixture interpretation emphasized the value of likelihood ratios.

Lineage markers and the counting method

As noted in Chapter 16, lineage markers include mitochondrial DNA and Y-chromosome haplotypes that are transferred directly from generation to generation either from mother to child in the case of mitochondrial DNA or from father to son in the case of the Y chromosome. The counting method in conjunction with an upper bound confidence limit is typically used when estimating the rarity of a mtDNA or Y-chromosome haplotype. The counting method relies on the size of the database and involves counting the number of times the profile (haplotype) has been observed within the database. A frequency estimate with a confidence interval is then made based on this count. Worked examples for Y-STRs and mtDNA are found in D.N.A. Boxes 16.2 and 16.3.

Points for Discussion

- What might be some of the advantages and disadvantages of combining population databases (i.e., having a universal or mixed ethnicity database)?
- Why is the counting method used for estimating Y-STR and mtDNA haplotype frequencies rather than a random match probability calculation using the product rule?

READING LIST AND INTERNET RESOURCES

General Information

Balding, D. J. (2005). Weight-of-evidence for forensic DNA profiles. Hoboken, NJ: John Wiley & Sons.

Buckleton, J., Triggs, C. M., & Walsh, S. J. (Eds.), (2005). *Forensic DNA evidence interpretation*. Boca Raton, FL: CRC Press.

Brenner, C. H. (2003). Forensic genetics: Mathematics. In D. N. Cooper (Ed.), *Nature encyclopedia of the human genome: Vol. 2* (pp. 513–519). New York: Macmillan Publishers Ltd., Nature Publishing Group.

DNA Advisory Board. (2000). Statistical and population genetic issues affecting the evaluation of the frequency of occurrence of DNA profiles calculated from pertinent population database(s). *Forensic Science Communications, 2*(3). Available at http://www.fbi.gov/programs/lab/fsc/backissu/july2000/dnastat.htm

Evett, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.

Fung, W. K., & Hu, Y.-Q. (2008). *Statistical DNA forensics: Theory, methods and computation*. Hoboken, NJ: John Wiley & Sons.

National Research Council Committee on DNA Forensic Science (1996). *The evaluation of forensic DNA evidence*. Washington, DC: National Academy Press.

Population Genetics

Gonick, L., & Wheelis, M. (1983). *The cartoon guide to genetics* (updated ed.). New York: HarperCollins Publishers.

Hartl, D. L., & Clark, A. G. (1997). *Principles of population genetics* (3rd ed.). Sunderland, MA: Sinauer Associates.

Weir, B. S. (1996). *Genetic data analysis II: Methods for discrete population genetic data*. Sunderland, MA: Sinauer Associates.

Statistics

Lucy, D. (2005). *Introduction to statistics for forensic scientists*. Hoboken, NJ: John Wiley & Sons.

Tracey, M. (2001). Short tandem repeat-based identification of individuals and parents. *Croatian Medical Journal, 42*, 233–238.

Weir, B. S. (2003). DNA evidence: Inferring identity. In D. N. Cooper (Ed.), *Nature Encyclopedia of the human genome: Vol. 2* (pp. 85–88). New York: Macmillan Publishers Ltd., Nature Publishing Group.

Weir, B. S., et al. (2006). Genetic relatedness analysis: Modern data and new challenges. *Nature Reviews Genetics, 7*, 771–780.

Weir, B. S. (2007). Forensics. In D. J. Balding, M. Bishop, & C. Cannings (Vol. Eds.), *Handbook of statistical genetics* (3rd ed.) (pp. 1368–1392). Hoboken, NJ: John Wiley & Sons.

Population Databases

Butler, J. M., et al. (2003). Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *Journal of Forensic Sciences, 48*, 908–911.

Chakraborty, R. (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Human Biology, 64*, 141–159.

Devlin, B. (1993). Forensic inference from genetic markers. *Statistical Methods in Medical Research, 2*, 241–262.

Einum, D. D., & Scarpetta, M. A. (2004). Genetic analysis of large data sets of North American Black, Caucasian, and Hispanic populations at 13 CODIS STR loci. *Journal of Forensic Sciences, 49*, 1381–1385.

Fung, W. K. (1996). Are convenience DNA samples significantly different? *Forensic Science International, 82*, 233–241.

Profile frequency estimates

Balding, D. J., & Nichols, R. A. (1994). DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International, 64*, 125–140.

Balding, D. J. (1999). When can a DNA profile be regarded as unique? *Science & Justice, 39*, 257–260.

Curran, J. M., et al. (2007). Empirical testing of estimated DNA frequencies. *Forensic Science International: Genetics, 1*, 267–272.

Likelihood ratios

Evett, I. W., & Weir, B. S. (1998). *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.

Evett, I. W. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, *40*, 233–239.

Source attribution

Budowle, B., et al. (2000). Source attribution of a forensic DNA profile. *Forensic Science Communications*, *2*(3). Available at http://www.fbi.gov/hq/lab/fsc/backissu/july2000/source.htm

Subpopulation issues

Buckleton, J. S., et al. (2006). How reliable is the sub-population model in DNA testimony? *Forensic Science International*, *157*, 144–148.

Curran, J. M., et al. (2003). What is the magnitude of the subpopulation effect? *Forensic Science International*, *135*, 1–8.

Prosecutor's fallacy

Balding, D. J., & Donnelly, P. (1994). The prosecutor's fallacy and DNA evidence. *Criminal Law Reviews*, *1994*, 711–721.

Leung, W. C. (2002). The prosecutor's fallacy—A pitfall in interpreting probabilities in forensic evidence. *Medicine, Science, and the Law*, *42*, 44–50.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law & Human Behavior*, *11*, 167–187.