Choice of population database for forensic DNA profile analysis

Christopher D. Steele *, David J. Balding

UCL Genetics Institute, Darwin Building, Gower Street, London WC1E 6BT, UK

ABSTRACT

When evaluating the weight of evidence (WoE) for an individual to be a contributor to a DNA sample, an allele frequency database is required. The allele frequencies are needed to inform about genotype probabilities for unknown contributors of DNA to the sample. Typically databases are available from several populations, and a common practice is to evaluate the WoE using each available database for each unknown contributor. Often the most conservative WoE (most favourable to the defence) is the one reported to the court. However the number of human populations that could be considered is essentially unlimited and the number of contributors to a sample can be large, making it impractical to perform every possible WoE calculation, particularly for complex crime scene profiles. We propose instead the use of only the database that best matches the ancestry of the queried contributor, together with a substantial F_{ST} adjustment. To investigate the degree of conservativeness of this approach, we performed extensive simulations of one- and two-contributor crime scene profiles, in the latter case with, and without, the profile of the second contributor available for the analysis. The genotypes were simulated using five population databases, which were also available for the analysis, and evaluations of WoE using our heuristic rule were compared with several alternative calculations using different databases. Using $F_{ST} = 0.03$, we found that our heuristic gave WoE more favourable to the defence than alternative calculations in well over 99% of the comparisons we considered; on average the difference in WoE was just under 0.2 bans (orders of magnitude) per locus. The degree of conservativeness of the heuristic rule can be adjusted through the F_{ST} value. We propose the use of this heuristic for DNA profile WoE calculations, due to its ease of implementation, and efficient use of the evidence while allowing a flexible degree of conservativeness.

1. Introduction

In forensic DNA analysis, unknown contributors to a DNA profile are usually considered to come from one of several populations for which an allele frequency database is available. The choice of database can have an important impact on weight of evidence (WoE): the rarer an allele the stronger the evidence implicating a queried contributor (Q) if he has that allele and it is observed in the crime scene profile (CSP). The most appropriate population is the one that best matches the ancestry of X, the true source of the DNA. Under the prosecution case X is assumed to be Q, but under the defence case there is often little or no information about the ancestry of X. Many authors have noted that the database most appropriate for Q is not necessarily most appropriate for X [6,4]. Conversely, [3] argue for using the database of Q even if the ancestry of X is unknown, in part because the observation of the profile of Q introduces a size-bias effect: an observed profile tends to

be more common in the population in which it was observed than in a different population. Thus, having observed the profile of Q, on average the probability for X to have the same profile is higher if X is assumed to come from the same population.

In current forensic practice, when the ancestry of X is unknown, it is common to consider multiple population databases and choose the one that generates the lowest WoE. There should be no requirement to favour defendants in this way. Suppose for example that Q is Caucasian but it is discovered that the lowest WoE is obtained using a database of Vietnamese individuals. If the population local to the crime includes few Vietnamese and there is no evidence to suggest that a Vietnamese person was the source of the DNA, it may not be helpful to the court to report the WoE arising from the Vietnamese database. Similarly, the world's population can be categorised in a vast number of different ways, and it is not possible to investigate them all in order to report the smallest WoE. However, a forensic expert should make reasonable allowance for the different possible ancestries of X, given the available knowledge about the location and nature of the crime. It can be expedient to make approximations that favour the defence in order to permit simplified analyses while avoiding courtroom challenges. Here we propose a heuristic for WoE analysis that involves only one calculation, using the database most appropriate for

---

Q. We show that our heuristic tends to strongly favour defences compared with a range of alternative calculations.

For a one-contributor CSP when there are only, say, five population databases, it is usually easy to compute the WoE for each database and choose the one most favourable to the defence. However, for mixed profiles, the computational effort to consider multiple databases for each unprofiled contributor can be substantial. Thus our heuristic that computes the WoE only using the database of Q would be attractive, provided that it can be established to be conservative (favourable to the defence). If X is from the same population as Q then it becomes relevant to consider that they may also come from the same subpopulation, in which case an F_{ST} adjustment may be required [3]. We have recently published worldwide F_{ST} estimates appropriate for forensic use [7] and concluded that choosing $F_{ST} = 0.03$ is sufficiently large to be almost always conservative. The effect of the F_{ST} adjustment is to increase the probability assigned to the alleles of Q, and consequently decrease the probability for other alleles. Although the rationale for an F_{ST} adjustment is to allow for the possibility that X has ancestry similar to that of Q, we illustrate below that for $F_{ST} = 0.03$ our heuristic calculation is conservative even if X could have come from one of several different populations. It is for this reason that our heuristic uses the same value of F_{ST} whatever the population of Q, even though within-population F_{ST} values differ across populations.

A similar argument applies to other contributors to a mixed CSP. Consider a two-contributor profile, one of the contributors being X, who is alleged to be Q. If the reference profile of the other contributor is known, as is often the case for a victim or bystander, there are no probabilities to assess for the alleles of that individual and so the question of the appropriate population database is essentially the same as for the one-contributor case. When the reference profile of the other contributor, say U, is unavailable, then we show that it is conservative to use for both X and U the database best matching the ancestry of Q, again with $F_{ST} = 0.03$. The F_{ST} adjustment under our heuristic only increases the population allele fraction for the alleles of Q, which is helpful to defences because it increases the probability that X or U share alleles with Q, thus increasing the support for the defence explanation of the observed CSP.

It is not feasible or desirable to guarantee that a proposed WoE calculation is more favourable to the defence than any conceivable alternative calculation. We perform simulation experiments which show that for UK population databases our heuristic WoE calculation is, with probability ≫ 0.99, more favourable to defendants than a range of reasonable alternative calculations. We first simulate single-contributor CSPs matching the reference profile of the alleged contributor Q. Then the WoE for Q to be a contributor is calculated using the correct database (that used for the simulation) and is compared with the smallest WoE calculated using in turn four other databases. We repeated this exercise for one database using allele fractions that differ from the database values according to each of three values of F_{ST}, and show that our heuristic remains conservative compared to the WoE from the four alternative databases.

We then simulate two-contributor CSPs using all possible choices of two databases from the five available, and compare the WoE computed using the database of Q for both contributors (and $F_{ST} = 0.03$) with (a) the correct assignment of databases, (b) the minimum WoE using each of the four alternative databases for both X and U, and (c) the minimum WoE over the four databases for X, always using the correct database for U. In all our calculations, an adjustment using $F_{ST} = 0.03$ is applied to the alleles of Q when the database of Q is used for X.

When a calculation is performed using a database different from that of Q, perhaps because of evidence about the ethnic background of X, coancestry is not relevant and so it is appropriate to use $F_{ST} = 0$. It has been suggested [2] that even in this setting it would be cautious to use a low value of F_{ST} such as 0.01. This introduces some bias in favour of the defendant in order to allow for the ancestry of X to differ somewhat from the database population. Here we assume that there is no specific

suggestion of an alternative population for X, and since a bias in favour of defendants is introduced by taking the minimum WoE over four alternative database choices, we use $F_{ST} = 0$ in calculations using databases different from that of Q.

It is possible that the true ancestry of Q is unknown or misassigned, for example if he impersonates another individual, or an assessment of his physical appearance was incorrect. He may also be of mixed ancestry or some other ancestry not well represented in the available databases. In that case there is no size-bias effect tending to make the observed profile of Q more common in the population to which he is assigned than in other populations. However, although such an error may have an adverse impact on the calculated WoE, the generous value of F_{ST} is the main factor underlying the conservative nature of the WoE analysis that we propose, and so the impact of any population misassignment of Q will be relatively small.

2. Materials & methods

2.1. Databases

We have used frequency data at 16 STR loci for five UK populations: Caucasian (IC1), African and African Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6) (Table 1). For further details of the dataset, see [7]. We used these data to simulate 16-locus profiles assuming Hardy–Weinberg and linkage equilibria. Neither dropin nor dropout are included in the simulations, nor are they allowed for in the analyses.

The WoE is computed using the likelihood ratio framework [5], and reported in bans ($= log_{10}$(likelihood ratio)) comparing a hypothesis that includes Q as a contributor with an alternative in which Q is replaced by X, assumed to be unrelated to Q. We implement F_{ST} adjustment [2] to the population fractions of the alleles of Q whenever the database most appropriate for Q is used for X; the adjustment uses

Table 1
Number of allele observations at each locus for each population database: Caucasian (IC1), Afro-Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6).

Allele counts	IC1	IC3	IC4	IC5	IC6
D3S1358	6878	3941	520	599	1202
TH01	6816	3918	514	598	1202
D21S11	6870	3941	520	599	1199
D18S51	6808	3930	520	600	1195
D16S539	6818	3927	514	600	1199
VWA	6877	3936	520	600	1201
D8S1179	6871	3941	520	600	1202
FGA	6853	3938	516	600	1201
D19S433	6702	3868	507	595	1197
D2S1338	6443	3758	491	594	1176
D22S1045	1816	2482	421	498	954
D1S1656	1827	2508	426	504	959
D10S1248	1815	2499	416	500	912
D2S441	1800	2473	420	493	943
D12S391	1857	2543	437	499	945
SE33	368	872	237	394	268

Table 2
Mean weight of evidence (WoE) for the heuristic rule and the alternatives discussed in the text. The mean of the differences between the heuristic and alternative scenarios is also shown. The % Difference row shows the mean difference as a percentage of the average of the heuristic and alternative means.

Contributors under Hd	X	X + K	X + U		
			True both	True U	Same dbase
Heuristic (bans)	20.3	17.8	10.7	10.7	10.7
Alternative (bans)	24.5	20.7	12.8	14.1	14.0
Difference (bans)	4.2	3.0	2.1	3.4	3.2
Difference (%)	18.8	15.6	17.9	27.4	25.9