Research paper

Population-specific F_{ST} values for forensic STR markers: A worldwide survey

John Buckleton [a,b], James Curran [c], Jérôme Goudet [d], Duncan Taylor [e,f], Alexandre Thiery [g], B.S. Weir [h],*

[a] Environmental and Scientific Research Ltd., Auckland, New Zealand
[b] National Institute of Standards and Technology, Gaithersburg, MD, USA
[c] Department of Statistics, University of Auckland, Auckland, New Zealand
[d] Department of Ecology and Evolution, UNIL-Sorge, Lausanne, Switzerland
[e] Forensic Science South Australia, Adelaide, SA, Australia
[f] School of Biological Sciences, Flinders University, Adelaide, SA, Australia
[g] Bioinformatics Core Facility, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[h] Department of Biostatistics, University of Washington, Seattle, WA, USA

ARTICLE INFO

ABSTRACT

The interpretation of matching between DNA profiles of a person of interest and an item of evidence is undertaken using population genetic models to predict the probability of matching by chance. Calculation of matching probabilities is straightforward if allelic probabilities are known, or can be estimated, in the relevant population. It is more often the case, however, that the relevant population has not been sampled and allele frequencies are available only from a broader collection of populations as might be represented in a national or regional database. Variation of allele probabilities among the relevant populations is quantified by the population structure quantity F_{ST} and this quantity affects matching proportions. Matching within a population can be interpreted only with respect to matching between populations and we show here that F_{ST}, can be estimated from sample allelic matching proportions within and between populations. We report such estimates from data we extracted from 250 papers in the forensic literature, representing STR profiles at up to 24 loci from nearly 500,000 people in 446 different populations. The results suggest that theta values in current forensic use do not have the buffer of conservatism often thought.

1. Introduction

The use of microsatellite, or short tandem repeat (STR), loci for forensic identification is now well established and analyses of large STR frequency databases have been published. Silva et al. [1] made use of STR allele frequencies accessed from a then-online database strdna-db (Pamplona et al. [2]), while Steele et al. [3] used data from people living in the United Kingdom (UK), or wishing to migrate to the UK on the basis of relatedness to a UK resident, as well as "reference" data collected by the UK Forensic Science Service. Both these papers provide useful reviews of the forensic uses of STR profiles and both describe the population genetic structure revealed by STR data. Silva et al. [1] commented on the overall similarity between conclusions that can be drawn from

forensic and other population databases about the structure of human populations and its relationship to the movement of people from East Africa, the most likely place of origin of anatomically modern humans [4]. This is in spite of forensic markers being chosen to maximize the diversity among individuals. Steele at al. [3] gave detailed analyses of the population structure quantity F_{ST} that can be regarded as measuring the evolutionary relatedness between two individuals in the same population.

Here we continue the discussion of F_{ST} by describing a new approach to estimating this quantity and then by applying this approach to allele frequencies at 24 STR loci in forensic use published for 446 populations. Central to our approach is the recognition that F_{ST} values are statements about pairs of alleles within a population *relative to* some other population or collection of populations, and this reference frame needs to be specified. Our results suggest that a somewhat higher value of F_{ST} than previously thought should be used in forensic match probability calculations,

especially when the continental ancestry of the relevant population is unknown.

1.1. Allelic matching theory

Population genetic theory (e.g. [5]) distinguishes between three types of probability of A_u, the uth allele at locus **A**. First, there is the sample frequency $\tilde{p}_{iu}$ in a sample of individuals from population i. Second, there is the actual frequency $\ddot{p}_{iu}$ of A_u in population i from which this sample has been taken. Note that $\ddot{p}_{iu}$ is the expected value of $\tilde{p}_{iu}$ over samples from that population, irrespective of sample size, so $\tilde{p}_{iu}$ is an unbiased and consistent estimate of $\ddot{p}_{iu}$. With random sampling of individuals, and if the population in question is in Hardy–Weinberg equilibrium, then the number of A_u alleles in a sample of n_i genotypes has a binomial distribution with parameters $2n_i$ and $\ddot{p}_{iu}$ so the within-population variance of $\tilde{p}_{iu}$ is $\text{Var}_W(\tilde{p}_{iu}) = \ddot{p}_{iu}(1-\ddot{p}_{iu})/(2n_i)$.

Any population genetic theory aimed at calculating match probabilities must address another level of sampling, namely that inherent to the underlying evolutionary process [6]. An actual allele frequency $\ddot{p}_{iu}$ is just one of many values that are possible for a given evolutionary history, and the expected value of $\ddot{p}_{iu}$ taken over many different realizations of this same history, written as p_u, is usually referred to as the allele frequency of A_u. It may be better to say "allele probability" here but we will defer to common usage and use "frequency" both for sample and population proportions and the corresponding expected values or probabilities. The same value p_u applies here to all populations and it is the probability that an allele drawn at random from any population is of type A_u. Note that the total expectation of the sample frequency $\tilde{p}_{iu}$, taken over samples and over populations is also equal to p_u. There is an implicit assumption that natural selection has not been acting differentially among populations. For a wide class of evolutionary models, the variance of the actual allele frequency $\ddot{p}_{iu}$ among replicates of that population has the form $\text{Var}_A(\ddot{p}_{iu}) = p_u(1-p_u)\theta_i$, where θ_i is the probability that two alleles drawn randomly from population i are identical by descent (ibd). Alleles are ibd if they have a common origin – this is necessarily a statement about the past and there is variation among all the possible populations that might descend from an ancestral population. The total variance of the sample frequency $\tilde{p}_{iu}$, over samples from a population and over realizations of a population, is $\text{Var}_T(\tilde{p}_{iu}) = p_u(1-p_u)[\theta_i + (1-\theta_i)/(2n_i)]$.

The evolutionary perspective given by this introduction of θ_i allows account to be taken of some alleles in the population having a shared ancestry and being ibd. Shared ancestry of some alleles increases the chance of allele matching, or identity in state (ibs), for alleles drawn randomly from the population.

If there is Hardy–Weinberg equilibrium within a (large) population i for which the frequency of allele A_u is $\ddot{p}_{iu}$, then the probability that two alleles in that population are of type A_u is $Pr(A_uA_u)_i = \ddot{p}_{iu}^2$, and this has an expected value over replicates of the population of $p_u^2 + p_u(1-p_u)\theta_i$. If, instead of a population frequency $\ddot{p}_{iu}$ for population i, or its expected value p_u, we have only a sample value $\tilde{p}_u$ from a larger population, or collection of populations, then we must consider the sampling variance of $\tilde{p}_u$. Some of the alleles contributing to that sample value are from the target population i for which θ_i is appropriate and some are from different populations j, $j \neq i$ for which an analogous quantity θ_{ij} is appropriate. This new measure is the probability that an allele from population i is ibd to an allele from population j. The estimated two-allele probability for population i is $\tilde{p}_u^2 + \tilde{p}_u(1-\tilde{p}_u)\beta_i$, where $\beta_i = (\theta_i - \theta_B)/(1 - \theta_B)$ and θ_B is the average over all pairs of populations within the sampled collection of population-pair θ_{ij}'s. We are assuming a large (unknown) number of populations within the sampled collection. The corresponding allelic match probability is $Pr(A_u|A_u) = p_u + \theta_i(1 - p_u)$ and we can estimate this as $\tilde{p}_u + \beta_i(1-\tilde{p}_u)$.

Balding and Nichols [7] extended the discussion of matching to genotypes and gave the match probabilities for a population as:

$$Pr(A_uA_u|A_uA_u)_i = \frac{[2\theta_i + (1-\theta_i)p_u][3\theta_i + (1-\theta_i)p_u]}{(1+\theta_i)(1+2\theta_i)}$$
$$Pr(A_uA_v|A_uA_v)_i = \frac{2[\theta_i + (1-\theta_i)p_u][\theta_i + (1-\theta_i)p_v]}{(1+\theta_i)(1+2\theta_i)}, \quad u \neq v \tag{1}$$

If the particular population within a collection of populations is not specified, and we wish to express a match probability that would apply to any of a large number of populations within the collection, then we replace θ_i by the average value θ_W over populations. To estimate the matching probabilities with the total sample allele frequencies $\tilde{p}_u$ in place of the true values p_u, we use $\beta_W = (\theta_W - \theta_B)/(1 - \theta_B)$, the average of the β_i's, in place of θ_W.

We note that most discussions in the literature, e.g. [1,2], do not make explicit mention of between-population values θ_{ij} or θ_B, and we also note that β_W is the quantity usually referred to as F_{ST}. Our formulation stresses that the concept of identity by descent of alleles within a population requires a point of reference, which we take as identity by descent between pairs of populations. We see below that there is an immediate translation to allelic matching within and between populations.

There is widespread use in forensic science of the profile probabilities $Pr(A_uA_u)$ or $Pr(A_uA_v)$, $u \neq v$, and their product rule estimates $\tilde{p}_u^2$ or $2\tilde{p}_u\tilde{p}_v$, rather than the match probabilities shown in (1). To allow for departures from Hardy–Weinberg equilibrium in the population collection that results from variation in allele frequencies among populations, the profile probabilities in a subpopulation can be estimated as

$$Pr(A_uA_u)_i \triangleq \tilde{p}_u[F_i + (1-F_i)\tilde{p}_u], \quad Pr(A_uA_v)_i \triangleq 2\tilde{p}_u\tilde{p}_v(1-F_i) \tag{2}$$

although, following the National Research Council [8], the $(1 - F_i)$ is often omitted in the expression for heterozygotes. In these last equations, the symbol $\triangleq$ means "is estimated by" and the equations provide estimates for population genotype frequencies in terms of population-collection sample allele frequencies. The quantity F_i is the total inbreeding coefficient for population i. It is the average of Eq. (2) that would be appropriate if the particular population was not identified and here we write the average over populations of F_i as F_W, although it is generally written as F_{IT} [6]. If there is Hardy–Weinberg equilibrium within populations, then $F_i = \theta_i$, $F_W = \theta_W$.

Regardless of whether the average $(1 - F_W)$ term is included for heterozygotes, it is not clear that estimated profile probabilities are of great forensic interest since the relevant quantities are match probabilities: the chance that an untyped person will have a certain genotype given that a typed person has that genotype. When account is taken of the evolutionary history of a population, F_i, $\theta_i > 0$, the match probabilities exceed the profile probabilities. Match probability estimates require the quantities β_i, or their average β_W.

Since their introduction by Balding and Nichols [7], the match probabilities in Eq. (1) have been extended to allow for inbreeding [9], mixtures [10], and relatedness [11]. They have been of substantial benefit for the interpretation of matching autosomal profiles and they were endorsed by the US National Research Council [8].

1.2. Estimation of F_{ST}

There is a logical difficulty in estimating $F_{ST} = \beta_W$ or the population-specific values β_i. We have shown how it arises in match probabilities to account for variation in allele frequencies over evolutionary replicate populations. We could estimate β_i from standard methods, such as those of Weir and Cockerham [12] (who

assumed $\theta_B = 0$ and $\theta_i = \theta$ for all values of i) or Weir and Hill [13] (who relaxed those assumptions), *if* we had data from more than one actual population to provide an indication of evolutionary variation, but if we had those frequencies we could use them directly in a within-population analysis and not need the θ formulation. We adopt a pragmatic work-around by exploiting the many published sets of allele frequencies, grouping them into broad ancestry or geographic sets: an estimate from a set of European-ancestry samples, for example, can be used in situations where a European-ancestry θ is required.

We take this opportunity to simplify the expressions for our earlier estimation procedures [12,13] while allowing for different values θ_i for different populations i and for analogous quantities θ_{ij} for pairs i, j of populations. Bhatia et al. [14] found it convenient to work with sample heterozygosities within and among populations, and we use the complementary matching proportions (i.e. sample homozygosities). The within-population sample matching proportions can be found from the counts n_{ilu} of allele A_{lu} at locus l sampled from population i as $\tilde{M}_{il} = \sum_u n_{ilu}(n_{ilu}-1)/[n_{il}(n_{il}-1)]$, where $n_{il} = \sum_u n_{ilu}$ is the total number of alleles sampled at that locus. The sample allele frequencies are $\tilde{p}_{ilu} = n_{ilu}/n_{il}$, and we can write $\tilde{M}_{il}$ as $n_{il}[\sum_u \tilde{p}_{ilu}^2 - 1/n_{il}]/(n_{il}-1)$ or, approximately, as $\tilde{M}_{il} = \sum_u \tilde{p}_{ilu}^2$. For populations i and j, the between population-pair sample matching proportions are $\tilde{M}_{ijl} = \sum_u n_{ilu}n_{jlu}/(n_{il}n_{jl})$ or $\tilde{M}_{ijl} = \sum_u \tilde{p}_{ilu}\tilde{p}_{jlu}$. For sets of r populations, we take averages over populations and over pairs of populations of the matching proportions for locus l: $\tilde{H}_{Wl} = \sum_{i=1}^{r} \tilde{H}_{il}/r$ and $\tilde{H}_{Bl} = \sum_{i \neq j} \tilde{H}_{ijl}/[r(r-1)]$.

We estimate the β's by comparing within-population matching proportions to between population-pair matching proportions with the ratios $\hat{\beta}_{il} = (\tilde{M}_{il}-\tilde{M}_{Bl})/(1-\tilde{M}_{Bl})$ and $\hat{\beta}_{Wl} = (\tilde{M}_{Wl}-\tilde{M}_{Bl})/(1-\tilde{M}_{Bl})$. Approximating the expectation of these ratios as the ratio of expectations, the Weir and Hill model [13] leads to $\hat{\beta}_{il}, \hat{\beta}_{Wl}$ being unbiased estimates of β_{il}, β_{Wl}.

We will see below that there is variation among loci of the locus-specific values $\hat{\beta}_{il}$ and these differences may reflect both different locus-specific mutation rates as well as large sampling variances. We provide estimates of locus-average values, evaluated as explained in the Supplementary Material, that will be unbiased if each locus has the same value. The average matching proportions over loci are $\tilde{M}_i = \sum_{l=1}^{L} \tilde{M}_{il}/L$, $\tilde{M}_W = \sum_{i=1}^{r} \tilde{M}_i/r$, and $\tilde{M}_B = \sum_{l=1}^{L} \tilde{M}_{Bl}/L$ for a set of L loci. The β estimates are $\hat{\beta}_i = (\tilde{M}_i-\tilde{M}_B)/(1-\tilde{M}_B)$ and $\hat{\beta}_W = (\tilde{M}_W-\tilde{M}_B)/(1-\tilde{M}_B)$: these "ratio of averages" estimates have smaller variances than the averages of the $\hat{\beta}_{il}$ or $\hat{\beta}_{Wl}$ ratios.

It is not possible to estimate the within-population θ's other than as being "relative to" the between-population-pair values. The compound quantities β measure how much more matching there is within populations than there is between pairs of populations. Weir and Hill [13] showed that it is these compound parameters that, for a pure drift model of evolution, are functions of the time since the set of populations diverged from an ancestral population. For two populations i and j in the pure drift case, the parameter $(\beta_i + \beta_j)/2$ is proportional to the time since those two populations diverged from an ancestral population and so it can be used as a distance measure in constructing evolutionary dendrograms (see Supplementary Material).

To illustrate how populations i differ in match probabilities it would be preferable to examine values of θ_i. Although only estimates of θ_i relative to θ_B are available, comparisons are still possible. If $\beta_i > \beta_j$, for example, then it can be inferred that $\theta_i > \theta_j$ and it will be assumed reasonable that $\hat{\beta}_i > \hat{\beta}_j$ also implies $\theta_i > \theta_j$ when the same between-population-pair matching proportion $\tilde{M}_B$ is used for each. In the empirical values shown below, the estimates $\hat{\beta}_i$ for population i, in a group of populations such as those with common continental ancestry, use data from all the populations in that group in order to calculate $\tilde{M}_B$. Within-population match probabilities are

appropriately estimated with population-collection sample allele frequencies $\tilde{p}_{lu}$ in place of population frequencies p_{lu} and estimates $\hat{\beta}_W$ values in place of θ_W. The formulation given here of β_W or F_{ST} shows this quantity describes allelic matching within populations relative to matching between pairs of populations within a collection of populations.

2. Data and data handling

Data were obtained from 250 published population reports in the following journals: Forensic Science International, Forensic Science International: Genetics, International Journal of Legal Medicine, Journal of Forensic Sciences, and Legal Medicine. The data were gathered from electronic sources, either from the PDF text or from published electronic files, or from hard copy using OCR and post-processing. We are aware that we have not been comprehensive and that even more data are available. The populations used and the associated references are given in the Supplementary Material. This process yielded 446 distinct populations. The populations and the papers from which we extracted data are also shown in the Supplementary Material. We do not have access to the individual genotypes, so many of the data-cleaning steps described by Pemberton et al. [15] are not available.

The data required a significant amount of post-data handling from cleaning to storage. In particular, not all of the data collected in this study were originally collected for forensic purposes. A proportion was collected more probably for anthropological reasons than validation of forensic multiplexes for casework. The net effect of this fact is that we had some issues regarding non-standard loci and non-standard allele designations, as well as small samples.

Errors in the data were also apparent in a significant number of cases. Most of these errors were typographical and in many cases the correct data could be deduced. In some cases we determined that an error was present because the allele frequencies did not add sufficiently close to one or because allele frequencies multiplied by twice the number of individuals in the sample were not sufficiently close to integers. In one case the allele frequencies had not been correctly determined from the genotype frequencies and in another case D3S1358 and D13S317 had been swapped. In approximately half of the cases of perceived error we received a helpful answer from the corresponding author to our enquiries. In the other cases we have either omitted the data or used our own deductions as to the correct values. We have also struggled with different nomenclatures for rare alleles, however this should have only a minor effect on the analysis. We assigned the populations to geographic groups using our best judgment, and we show below that these groups of populations cluster together with principal component analysis. We therefore have some confidence that the grouping has both genetic and geographic coherence although we concede that further adjustments could be made. The motivation for grouping at all, apart from simplifying the presentation of results, is that we wish to provide θ values (now interpreted as β_W) that would result in appropriate match probability estimates for any population within a broadly-defined group of populations. The group names, numbers of populations within groups, and sample sizes within groups, are shown in Table 1. The data we discuss represent observations on 494,473 individuals. The loci for which we present results are displayed in Table 2.

3. Results

3.1. Principal coordinates analysis

As a first examination of the data, and as a check on our assignment of populations to geographic groups, we performed a

Table 1
Sample characteristics for geographic groups.

Group	Name	Number of		Sample size[†]	
		Populations	Loci[*]	Minimum	Maximum
Africa	African	37	24	342	7998
Andaman Islands	Andam	3	9	61	97
Asian	Asian	85	24	1153	23,557
Australian Aborigine	AusAb	17	15	1686	18,441
Caucasian	Caucn	173	24	861	233,753
Hispanic	Hisp	41	24	236	167,872
India Pakistan	IndPk	30	23	488	4525
Inuit	Inuit	3	21	194	403
Native American	NatAm	34	17	618	4234
Polynesian	Polyn	4	15	2614	33,593
Unknown	Unknn	15	24	205	9072

[*] Number of loci scored in at least one population in the group.
[†] Number of individuals scored at a single locus over all populations in the group.

Table 2
Numbers of alleles and numbers of populations for each locus.

Locus	No. of alleles	Number of populations per geographic group									
		African	Andam	Asian	AusAb	Caucn	Hisp	IndPk	Inuit	NatAm	Polyn
CSF1PO	27	33	–	76	2	109	39	25	2	33	2
D1S1656	26	3	–	7	–	34	2	2	1	–	–
D2S441	33	3	–	9	–	36	2	2	1	–	–
D2S1338	33	20	–	63	2	97	21	14	1	25	4
D3S1358	30	36	3	82	17	161	41	27	3	31	4
D5S818	25	34	2	81	17	131	40	27	2	34	2
D6S1043	32	1	–	4	–	2	1	–	–	–	–
D7S820	37	34	3	81	17	131	40	27	2	34	2
D8S1179	30	37	3	77	17	161	41	27	3	34	4
D10S1248	16	3	–	9	–	36	2	2	1	–	–
D12S391	38	3	–	7	–	31	3	2	1	–	–
D13S317	28	34	3	81	17	131	40	27	2	34	2
D16S539	26	35	–	74	2	135	39	28	3	30	4
D18S51	56	36	2	80	17	161	41	28	3	34	4
D19S433	36	20	–	65	2	99	21	14	1	25	4
D21S11	70	37	3	80	17	161	41	28	3	34	4
D22S1045	13	3	–	8	–	36	2	2	1	–	–
FGA	86	37	3	81	17	162	41	27	3	34	4
PENTAD	40	6	–	22	–	37	10	7	–	4	–
PENTAE	46	6	–	23	–	37	10	7	–	4	–
SE33	85	2	–	3	–	11	2	2	1	–	–
TH01	30	36	–	77	2	142	41	28	3	34	4
TPOX	22	33	–	75	2	109	39	25	2	34	2
vWA	31	36	3	81	17	162	41	30	2	34	4

principal component analysis of the allele frequencies in the whole data set. We used only those populations for which at least 25 individuals were scored. In Fig. 1 we show a parallel coordinates plot of the first 10 principal coordinates. There is clear separation of populations by geographic group: principal coordinate (PC) 1 separates out Asian, Caucasian and Indo-Pakistan groups, PC2 separates out the Hispanic and Native American groups, PC3 separates out the African group, PC4 separates out the Australian Aborigine group, PC5 separates out the Inuits and Polynesians. There does not appear to be any population assigned to the wrong group and the overall impression is that the forensic STR markers can distinguish geographic groups, as has been noted previously [1]. The results depicted in Fig. 1 suggest that further F_{ST}–based analyses are appropriate.

3.2. Population-specific F_{ST}

A complete list of the population- and locus-specific estimates $\hat{\beta}_{il}$ is given in the Supplementary Material. In Fig. 2 we display estimates of the population-specific parameters β_i for populations i. The estimates use all the loci scored in each population, as

explained in Appendix A, and we bootstrap over loci to provide the 95% confidence intervals indicated by vertical lines in the figure. The between-population-pair matching values $\tilde{M}_B$ used to estimate the β's were for all pairs of populations in the survey. The estimates have been ordered by size and colored by geographic group. The smallest values are for the African populations, reflecting the greater diversity within those older populations, whereas the largest values are for the smaller and less diverse Native American and Inuit populations. The Asian values are generally higher than the Caucasian values, with both lying between the African and Native American values. This pattern is not unexpected but is based on larger sets of data than have been used previously. Even when up to 24 loci are used to estimate β_i, there are large sampling variances. We now turn to a closer look among loci and among geographic groups.

3.3. Locus-specific F_{ST}

In Fig. 3 we display box-plots of the estimates of β_{il} for loci l and populations i. The between-population pair matching values $\tilde{M}_B$ used to estimate these β's were for all pairs of populations in the
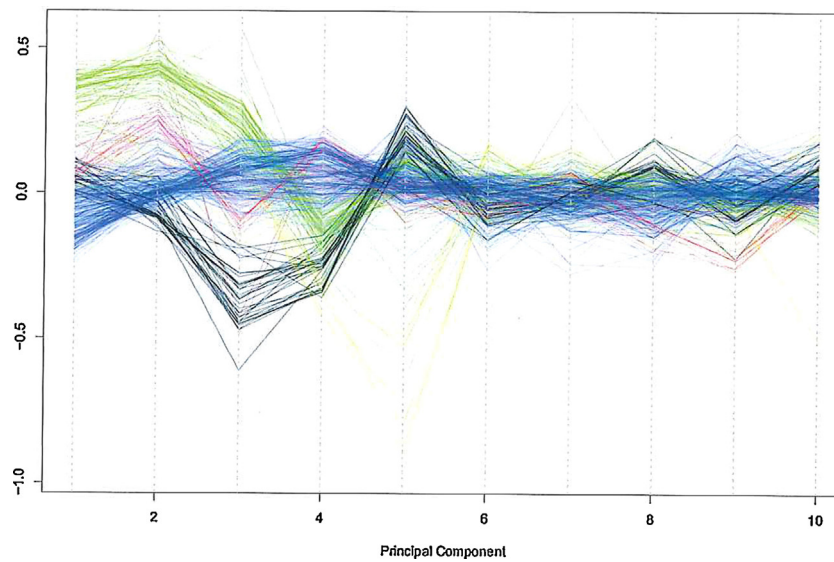
Fig. 1. Parallel coordinate plot for first 10 principal coordinates for all populations with sample sizes at least 50. Each line in the plot represents one population. Color code: Black = African, Gray = AusAb, Yellow = Asian, Blue = Caucn, Purple = Hisp, Brown = IndPk, Red = NatAm, Orange = Inuit, Brown = Andam, Green = Polyn.
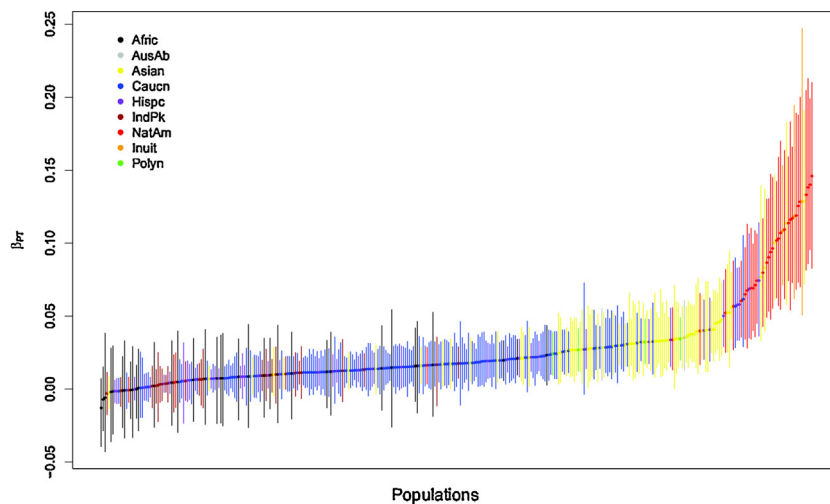

Fig. 2. Estimated values of the population-specific β_i, ordered by size. Each vertical line in the plot represents one population, and the length of the line is the 95% confidence interval obtained by bootstrapping over loci.

survey. The colored dots show the mean values over populations for each of the nine geographic groups.

There is relatively little variation over loci in the averages over populations of locus-specific values, but there is substantial variation in the range of values over populations at each locus. The variation appears to reflect the numbers of geographic groups and/or numbers of populations typed at each locus. Locus D3S1358, for example, was typed in over 400 populations from all geographic groups, whereas locus D6S1043 was typed in only eight populations in four geographic groups. This sampling range seems to mask any effect of mutation rate in the figure. The loci have been ordered on the X axis in decreasing order of published mutation rates (Table 14.5 in [16]): SE33 on the left has the highest reported rate, of 0.0016, whereas the D1S1656, D2S441, D10S1248, D22S1045 loci on the right have no reported mutations.

3.4. Geographic-group-specific F_{ST}

In Fig. 4 we display population-specific estimates of β_i for each continental group of populations. Two sets of estimates are shown: those on the right, with a shaded box plot, use only the populations

in the group to calculate the between-population-pair matching proportion $\tilde{M}_B$. The resulting β_W estimate is therefore the average over populations of θ relative to that group or region, and we could write the estimates as $\hat{\beta}_{PR}$. These are the estimates to be used if the relevant group of populations is known. The estimates on the left for each group, with an open box, use all pairs of populations across all groups to calculate $\tilde{M}_B$. Now the estimates are relative to the total collection of populations and we could write them as β_{PT}. These are appropriate if there was no information about the group to which the relevant population belongs.

As the population versus total estimates are to be used when there is no information about the group to which a population belongs, it is reasonable that those estimates β_{PT} are greater than the population versus region estimates $\hat{\beta}_{PR}$: this increases the matching probability and reduces the evidentiary strength of a match. A substantial difference in the two estimates $\hat{\beta}_{PT}$ and $\hat{\beta}_{PR}$, is seen for the Inuit group of populations: the average within-population matching proportion, averaged over loci, is $\tilde{M}_W = 0.4379$ whereas the average between-population-pair matching proportions are 0.1726 for pairs within the group and 0.0090 for all pairs in the study. The Inuit populations are more similar to each
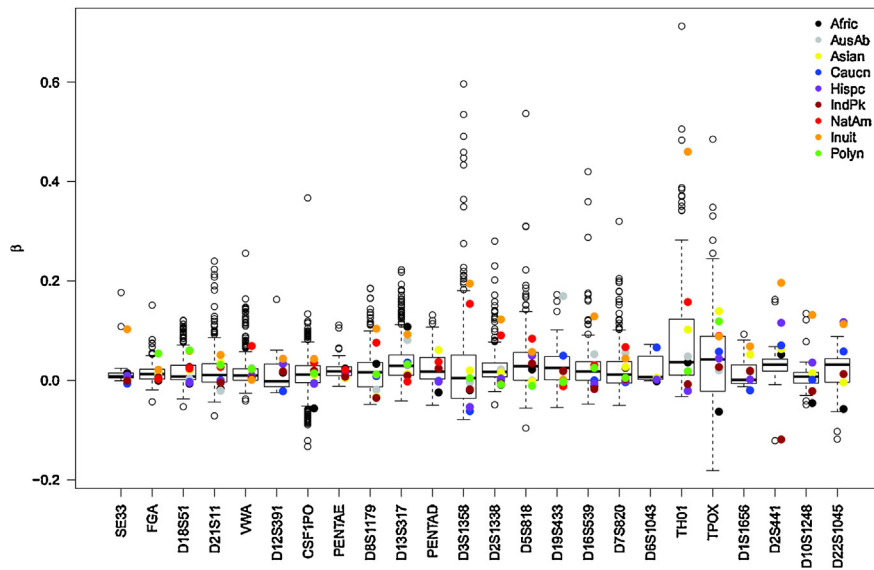
Fig. 3. Estimated values of the locus-specific β_{il}, ordered by locus. The black dots, and box-plots with whiskers, are for populations. The colored dots are for geographic groups. The box plots show the medians and interquartile ranges: the whiskers extend out from the box plots to 1.5 times the interquartile range.
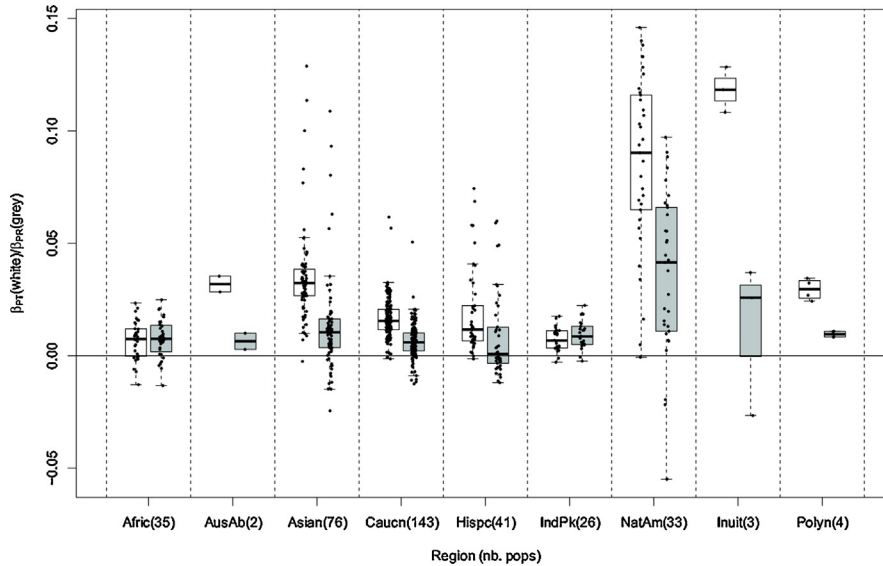

Fig. 4. Estimated values of the population-specific β's, for each geographic group, using all loci. Each box plot indicates the inter-quartile variation among populations within the group, with whiskers extending out from the box plots by 1.5 times the interquartile range. For each geographic group, the left hand plot (white box plot) compares each within-population matching proportion to the average matching proportion among all pairs of populations. The right hand plot (gray box plot) compares each within-population matching proportion to the average matching proportion among all pairs of populations in that region.

other than are any pair of populations in the study. We find that $\hat{\beta}_{PR} = 0.0205$ and $\hat{\beta}_{PT} = 0.1057$. The theta correction has much less effect if attention can be confined to the Inuit group.

The values for the African group show the opposite relationship from the Inuit group. Now the average within-population matching proportion is $\tilde{M}_W = 0.1884$ and the average between-population-pair averages are 0.1691 within the African region and 0.1726 for all pairs of populations. There is more divergence between pairs of African populations than there is among all pairs of populations and $\hat{\beta}_{PR} = 0.0082$ and $\hat{\beta}_{PT} = 0.0020$. The theta corrections for African populations have little effect regardless of the reference group.

The locus- and group-specific values of $\hat{\beta}_{Wl}$ are shown in Tables 3 and 4, where the group or the world serve as reference collections of populations. We advocate the use of the locus-average values. Except for Africa and India-Pakistan, the values are

greater when the world rather than the group serves as a reference, and we note the different rankings of the groups in the two tables: Africa clearly has the lowest average value in Table 4 whereas it is comparable to the Australian Aboriginal and Hispanic values in Table 3. The Inuit group has the highest value in Table 4, but is second behind Native Americans in Table 3.

From the perspective of a forensic scientist wanting to assign an F_{ST} value for a broad geographic group, if an upper limit (the upper end of each whisker in Fig. 4) was considered to be the most appropriate [3], then we would suggest values around 0.05 except for Native Americans and Inuits when the value is above 0.10. Rather than the upper limit, however, we have some preference for the median values (the center of each box in Fig. 4), and these suggest values around 0.01–0.03 except for Native Americans and Inuits when the value are higher and more susceptible to the group or the world being the reference set. The

Table 3
Estimated values β_{wl} of locus-specific β's for each geographic region, and the value for all loci within a region. Each estimate is for θ_W relative to pairs of populations within that region. There are no estimates if a locus is scored in less than two populations for a region, or if less than 25 individuals were typed in a region. Values in the body of the table are for each population, averaged over populations in a region. The bottom row is the average over loci. The method of averaging is described in the Supplementary Material.

Locus	Africa	AusAb	Asian	Caucn	Hispn	InPak	NatAm	Inuit	Polyn
CSF1PO	0.0042	0.0020	0.0073	0.0031	0.0018	0.0064	0.0256	0.0066	0.0256
D1S1656	0.0047	–	0.0027	0.0039	0.0032	−0.0018	–	–	–
D2S441	0.0129	–	0.0082	0.0032	0.0212	0.1933	–	–	–
D2S1338	0.0101	0.0039	0.0129	0.0075	0.0097	0.0158	0.0661	–	0.0051
D3S1358	0.0042	0.0002	0.0104	0.0044	0.0111	0.0178	0.0548	0.0058	0.0014
D5S818	0.0044	0.0018	0.0090	0.0055	0.0125	0.0109	0.0553	0.0127	0.0073
D6S1043	–	–	0.0051	0.0005	–	–	–	–	–
D7S820	0.0055	0.0008	0.0117	0.0064	0.0070	0.0067	0.0282	0.0092	0.0047
D8S1179	0.0050	0.0005	0.0123	0.0059	0.0021	0.0122	0.0197	0.0273	0.0077
D10S124	0.0018	–	0.0137	0.0026	0.0036	−0.0018	–	–	–
D12S391	0.0090	–	0.0039	0.0016	0.0204	−0.0018	–	–	–
D13S317	0.0068	0.0008	0.0157	0.0071	0.0144	0.0062	0.0335	0.0220	0.0136
D16S539	0.0085	0.0015	0.0205	0.0047	0.0094	0.0049	0.0343	0.0402	0.0042
D18S51	0.0047	0.0003	0.0145	0.0050	0.0025	0.0053	0.0306	0.0076	0.0026
D19S433	0.0103	0.0011	0.0103	0.0091	0.0136	0.0002	0.0310	–	0.0143
D21S11	0.0151	0.0001	0.0140	0.0055	0.0064	0.0128	0.0530	0.0002	0.0184
D22S104	0.0270	–	0.0498	0.0017	−0.0022	−0.0018	–	–	–
FGA	0.0033	0.0002	0.0116	0.0038	0.0058	0.0076	0.0235	0.0147	0.0063
PENTAD	0.0075	–	0.0145	0.0107	0.0022	0.0171	0.0259	–	–
PENTAE	0.0015	–	0.0152	0.0056	0.0029	0.0115	0.0217	–	–
SE33	0.0062	–	0.0084	0.0264	−0.0009	−0.0018	–	–	–
TH01	0.0209	0.0006	0.0237	0.0176	0.0176	0.0128	0.0640	0.0366	0.0168
TPOX	0.0104	0.0871	0.0164	0.0076	0.0074	0.0174	0.0511	0.0359	0.0307
VWA	0.0042	0.0002	0.0114	0.0046	0.0076	0.0059	0.0198	0.0479	0.0045
All	0.0081	0.0064	0.0133	0.0066	0.0077	0.0171	0.0368	0.0199	0.0106

Table 4
Estimated values β_{wl} of locus-specific β's for each geographic region, and the value for all loci within a region. Each estimate is for θ_W relative to all pairs of populations in the survey. There are no estimates if a locus is scored in less than two populations for a region, or if less than 25 individuals were typed in a region. Values in the body of the table are for each population, averaged over populations in a region. The right-most column is the average over regions, and the bottom row is the average over loci. The method of averaging is described in the Supplementary Material.

Locus	Africa	AusAb	Asian	Caucn	Hisp	IndPk	NatAm	Inuit	Polyn	World
CSF1PO	−0.0668	0.0130	0.0154	0.0127	0.0165	0.0197	0.0616	0.0406	0.0291	0.0117
D1S1656	0.0339	–	0.0658	−0.0018	0.0189	0.0316	–	0.0812	–	0.0157
D2S441	0.0153	–	0.0265	0.0316	0.1005	−0.0285	–	0.1625	–	0.0332
D2S1338	0.0029	0.0313	0.0319	0.0129	0.0234	0.0134	0.1210	0.1255	0.0035	0.0292
D3S1358	0.0145	0.0279	0.0578	−0.0345	0.0239	0.0227	0.2200	0.2196	0.0426	0.0254
D5S818	0.0102	−0.0229	−0.0132	0.0465	0.0474	0.0197	0.1192	0.0461	−0.0243	0.0337
D6S1043	−0.0006	–	0.0126	0.0669	0.0030	–	–	–	–	0.0233
D7S820	0.0244	0.0557	0.0345	0.0001	0.0165	0.0039	0.0842	0.0443	−0.0078	0.0222
D8S1179	0.0405	−0.0153	−0.0187	0.0169	0.0273	−0.0207	0.0885	0.1264	0.0227	0.0179
D10S1248	−0.0397	–	0.0383	0.0047	0.0473	−0.0195	–	0.1345	–	0.0102
D12S391	0.0317	–	0.0448	−0.0097	0.0745	0.0258	–	0.0522	–	0.0120
D13S317	0.1221	0.0806	0.0235	0.0445	0.0051	0.0093	0.0252	0.0990	0.0375	0.0384
D16S539	−0.0018	0.0597	0.0237	0.0288	0.0093	−0.0025	0.0720	0.1635	0.0227	0.0250
D18S51	−0.0012	0.0064	0.0345	0.0064	0.0026	0.0323	0.0503	0.0733	0.0538	0.0181
D19S433	−0.0095	0.1661	0.0226	0.0410	0.0053	0.0166	0.0132	−0.0013	0.0015	0.0254
D21S11	−0.0076	−0.0225	0.0422	0.0084	0.0126	0.0013	0.0702	0.0492	0.0393	0.0200
D22S1045	−0.0626	–	−0.0078	0.0300	0.0872	−0.0211	–	0.0836	–	0.0204
FGA	0.0027	0.0038	0.0183	0.0164	0.0011	0.0072	0.0226	0.0296	0.0655	0.0142
PENTAD	−0.0402	–	0.0567	0.0180	0.0015	0.0160	0.0380	–	–	0.0227
PENTAE	0.0185	–	0.0163	0.0235	0.0136	0.0137	0.0409	–	–	0.0202
SE33	0.0234	–	0.0138	0.0205	0.0152	0.0041	–	0.1081	–	0.0219
TH01	0.0731	0.0679	0.1465	0.0189	0.0369	0.0199	0.2084	0.5200	0.0464	0.0755
TPOX	−0.1336	−0.0031	0.0911	0.0578	0.0064	−0.0369	0.0736	0.0395	0.0412	0.0339
VWA	−0.0021	0.0246	0.0195	0.0087	0.0373	0.0055	0.0808	0.0231	0.0213	0.0198
All loci	0.0038	0.0328	0.0328	0.0193	0.0258	0.0065	0.0804	0.1050	0.0265	0.0244

table in the Supplementary Material shows much higher variation for locus- and population-specific values. As we discuss below, our recommendation is to use the multi-locus values in the last lines of Tables 3 and 4.

4. Discussion

We have presented an analysis of an extensive set of autosomal forensic STR allele frequencies with the aim of assisting forensic scientists assign F_{ST} or theta values for use in match probability calculations.

If allele frequencies $\ddot{p}_{iu}$ are known for the population thought to be relevant, meaning that it is the population from which an unknown contributor to an evidence profile is supposed to be drawn randomly, then those frequencies may be used directly. Assuming independence of alleles within and between loci, and assuming the population is large, the probability of the two profiles (for the person of interest and the unknown person with a

matching profile) is just the product of these frequencies and the match probability is the same as the profile probability.

Allele frequencies in a population are not known, however, and sampling variation must be considered. For reasonably large sample sizes, the evolutionary variation is larger than the sampling of individuals from the population and the "theta corrections" of Eq. (1) are appropriate. Values of θ_i need to be assigned: they cannot be estimated with data from a single population. Taking evolutionary variation into account, in effect, allows the evolutionary relatedness of the person of interest and the unknown person to be accommodated. If the relevant population is not sampled, but allele frequencies are available from some collection of populations, maybe of some specific geographic region or ancestry, then θ is replaced by β, with an average value β_W seeming most appropriate.

Silva et al. [1] showed that measures of population structure based on STR data provide good separation of geographic groups whether the loci were selected for evolutionary or forensic studies, with attenuated separation for forensic markers because they have been chosen to have the high diversity needed for individual identification. These authors made most use of the statistic R_{ST}, in which alleles are compared by the squared difference in the number of repeat units, rather than simply matching or not. As did these authors we also examined the survey of non-forensic STR data reported by Pemberton et al. [15] and, like them, found similar but attenuated separation for forensic markers with β instead of R_{ST} (results not shown). It is interesting that "global" F_{ST} value in [1] of 0.027 from allele frequency data is comparable to our 0.024 value shown in Table 4. Silva et al. did not focus on estimation of match probabilities.

Steele at al. [3] reported F_{ST} values for a collection of populations self-identified by UK residents or potential immigrants. Their "direct" method of estimation made use of data from a reference population with similar continental ancestry for each population under consideration, so their approach is closest to our $\hat{\beta}_{PR}$ values shown in Table 3. Their "indirect" method uses allele frequencies from the whole set of populations and so corresponds to the approach we used to generate the $\hat{\beta}_{PT}$ values shown in Table 4. Steele et al. use a Bayesian method and we would expect the medians of their posterior distributions to be comparable to the moment estimates we display. They concluded with a general recommendation of an F_{ST} of no more than 0.03 being generally applicable. For populations within a group, we are in general agreement with Steele et al, provided data from the appropriate group are used. If a world-wide dataset is to be used then higher values are necessary for Native American and Inuit populations. For individual Asian populations, however, we often find higher values than those shown in [3].

Steele and Balding [17] repeated the recommendation in [3] that an F_{ST} of 0.03 is sufficiently large to be almost always conservative. They advocate the use of a database "most appropriate" for the unknown donor of an evidentiary sample.

As the forensic task is to predict profile matching probabilities from an available set of data, we find intuitive appeal in phrasing our approach in terms of allelic matching proportions, especially by emphasizing the need to consider matching within and between populations. A matching proportion within a population is high in the present context only if it is substantially higher than matching between pairs of populations. The importance of the appropriate terms of reference is illustrated by the plots in Fig. 4.

The situation most likely to confront a forensic scientist is having to decide on a value of F_{ST} for match probability predictions for a population when the allele frequencies are going to be based on a sample from a collection of populations of similar ancestry or geographic region. We show the locus-specific values for each region in Tables 3 and 4. These $\hat{\beta}_{Wl}$ values are highly variable across loci and values are ranked differently within each region. We note a few negative locus-specific values, indicating less allelic matching at those loci within populations than between pairs of populations in a region. Because of the large variances in locus-specific values, we place little confidence in any one of them. We suggest, instead, the values we show that are based on all the loci typed for a population. In Fig. 2 we showed the 95% confidence intervals based on bootstrapping over loci and we note that Steele et al. [3] advocated using quantities analogous to the upper ends of these intervals. Our recommendation, however, is to use the all-loci values in the last rows of Tables 3 and 4.

We have approached allelic match probabilities from the expression $Pr(A_uA_u) = \breve{p}_{iu}^2$ for pairs of alleles in a population, i, for which each allele has probability $\breve{p}_{iu}$ of being of type A_u. The actual population frequencies $\breve{p}_{iu}$ are generally unknown, so we took expectations over replicates of the evolutionary process to obtain $Pr(A_uA_u) = p_u^2 + p_u(1-p_u)\theta_i$. If only an estimate $\tilde{p}_u$ is available for p_u, we had the estimate $\tilde{p}_u^2 + \tilde{p}_u(1-\tilde{p}_u)\beta_W$. Steele et al. [2] also considered the case of matching probabilities for individuals in two populations i, j. Our development leads to $Pr(A_uA_u) = p_u^2 + p_u(1-p_u)\theta_{ij}$ in that case, with sample values requiring $\beta_{ij} = (\theta_{ij} - \theta_B)/(1 - \theta_B)$. If populations i, j are in the same region R then we have an estimation problem: we cannot estimate β_{ij} other than by comparison to the average θ_B over all pairs of populations in that region and the only possibility we have is to compare each $\hat{\beta}_{ij} = (\tilde{M}_{ij} - \tilde{M}_B)/(1 - \tilde{M}_B)$ to their average. This average is zero by construction. The reference point for a single population is (the average over) pairs of populations: for a single region there is no meaningful reference point for pairs of populations. We can make progress in estimating β_{ij} for two populations in the same region by taking θ_B to refer to pairs of populations, one from each of two regions. If we have a hierarchical structure limited to populations and regions, however, we cannot address estimation of θ_{ij} for populations in different regions. Steele et al. [2] do not address the decomposition of F_{ST} into components within and between populations and so do not fully address the issue of allelic matching between populations.

The concept we stress in this paper is that forensic match probabilities, designed to recognize the shared ancestry between a person of interest and an unknown donor of an evidentiary sample, can be expressed in terms of identity by descent of pairs of alleles drawn from the relevant population. If sample allele frequencies, drawn from a larger set of populations, are to be used in calculating match probabilities then it is necessary to consider the identity between alleles drawn from pairs of these populations. It is the compound parameter F_{ST}, comparing allelic identity within populations to that between populations, that is of forensic importance. The value of this parameter clearly depends on the set of populations furnishing allele frequencies. The same point was made by Steele et al. [1] and we endorse their suggestion for the sampled set of populations to be generally of similar ancestry to the population of interest. A possible exception is for populations of African ancestry that tend to have retained greater diversity than those in the rest of the world.

Acknowledgements

Appendix A

A.1. Combining estimates over loci and populations

The within- and between-population-pair matching proportions for populations i and population-pairs i, j at locus l are $\tilde{M}_{il}$ and $\tilde{M}_{ijl}$. These have expectations $[1 - H_l(1 - \theta_{il})]$ and $[1 - H_l(1 - \theta_{ijl})]$ respectively and they provide the population/locus specific estimates

$$\hat{\beta}_{il} = \frac{\tilde{M}_{il} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}}$$

where $\tilde{M}_{Bl} = \sum_i \sum_{j \neq i} M_{ijl} / [r_l(1 - r_l)]$ if locus l is scored in r_l populations. These estimates have expectation $\beta_{il} = (\theta_{il} - \theta_{Bl}) / (1 - \theta_{Bl})$ if $\theta_{Bl} = \sum_i \sum_{j \neq i} \theta_{ijl} / [r_l(1 - r_l)]$.

Averaging over populations. Averaging matching proportions leads to average estimates over populations for locus l. If locus l is scored in r_l populations:

$$\tilde{M}_{Wl} = \frac{1}{r_l} \sum_i \tilde{M}_{il}$$

$$\hat{\beta}_{Wl} = \frac{1}{r_l} \sum_i \hat{\beta}_{il} = \frac{\tilde{M}_{Wl} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}}$$

$$\mathcal{E}(\hat{\beta}_{Wl}) = \frac{\theta_{Wl} - \theta_{Bl}}{1 - \theta_{Bl}}$$

Here the sums are over populations for which locus l is scored (e.g. $\theta_{Wl} = \sum_i \theta_{il} / r_l$), and the estimate is for the average over populations of the θ_{il} values relative to a reference value θ_{Bl}. These estimates allow a comparison among loci *if* the same populations are scored for each locus. In reality, this may not be the case and comparisons need to be interpreted carefully, although there is less of an issue as the number of populations increases.

Averaging over loci. Combining estimates over loci for population i is not as straightforward if the θ's differ among loci. A unweighted average over loci is

$$\hat{\beta}_i^u = \frac{1}{L_i} \sum_l \hat{\beta}_{il}$$

where L_i is the number of loci scored in population i. This average has expectation

$$\mathcal{E}(\hat{\beta}_i^u) = \frac{1}{L_i} \sum_l \beta_{il} = \frac{1}{L_i} \sum_l \frac{\theta_{il} - \theta_{Bl}}{1 - \theta_{Bl}}$$

This is not the average over loci of the θ_{il} values relative to a reference value.

Variance is reduced by weighting estimates by their denominators:

$$\hat{\beta}_i^w = \frac{\sum_l (\tilde{M}_{il} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})}, \quad \mathcal{E}(\hat{\beta}_i^w) = \frac{\sum_l H_l(\theta_{il} - \theta_{Bl})}{\sum_l H_l(1 - \theta_{Bl})}$$

If there is no variation among loci, $\theta_{il} = \theta_i$, and $\theta_{Bl} = \theta_B$, as might be expected for neutral loci with similar mutation rates, each of $\hat{\theta}_i^u, \hat{\theta}_i^w$ have the same expected value of $(\theta_i - \theta_B)/(1 - \theta_B)$.

A single estimate $\hat{\beta}_W$, for a set of populations, using all the scored loci, can be calculated as

$$\hat{\beta}_W^w = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B} = \frac{\sum_l (\tilde{M}_{Wl} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})}$$

$$\mathcal{E}(\hat{\beta}_W^w) = \frac{\sum_l H_l(\theta_{Wl} - \theta_{Bl})}{\sum_l H_l(1 - \theta_{Bl})}$$

An unweighted single estimate that does not have an expectation dependent on H_l is

$$\hat{\beta}_W^u = \frac{1}{L} \sum_l \hat{\beta}_{Wl} = \frac{1}{L} \sum_l \frac{\tilde{M}_{Wl} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}}$$

$$\mathcal{E}(\hat{\beta}_W^u) = \frac{1}{L} \sum_l \frac{\theta_{Wl} - \theta_{Bl}}{1 - \theta_{Bl}}$$

This is not the average value over loci, θ_W, relative to a reference value. If $\theta_{il} = \theta_i$ for all l, as might be expected for neural loci with equal mutation rates, each of $\hat{\beta}_W^w$ and $\hat{\beta}_W^u$ has the same expected value $(\theta_W - \theta_B)/(1 - \theta_B)$.

A.2. Which estimate should be used?

The results shown in this paper show variation in $\hat{\beta}_{il}$, suggesting variation in θ_{il}, over both loci and populations. The estimates are ratios of quadratic forms and large variances are expected. The figures suggest more variation over populations than over loci and this is not unexpected. One goal of this paper is to suggest values of "theta" for the calculations of match probabilities with the "theta correction." We have shown that the appropriate quantity is β_W, an average over populations in some relevant group of populations.

Our interest is primarily in the effect of population, rather than locus, on match probabilities. We would prefer to use a single theta, or β, value than separate values for each locus. [There is a need for separate values for each locus if we wish, for example, to detect signatures of natural selection at individual loci.] Which value should be used? The mean or median, or some other percentile, of the collection of single-locus values for a population or group of populations are valid measures, and have been advocated by Balding and his colleagues. Collections of values are shown in Figs. 2 and 3. Extensive sets of numerical values are shown in Supplementary Material section 4. We argue instead for the best single value calculated from all the loci, meaning the $\hat{\beta}_W^w$ values with smallest variance, and these are the values shown in Tables 3 and 4.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.03.004.

References

[1] N.M. Silva, L. Pereira, E.S. Poloni, M. Currat, Human neutral genetic variation and forensic STR data, PLoS ONE 7 (2012) e49666.

[2] J.P. Pamplona, F. Freitas, L. Pereira, A worldwide database of autosomal markers used by the forensic community, Forensic Sci. Int.: Genet. Suppl. Ser. 1 (2008) 656–657.

[3] C.D. Steele, D. Syndercombe Court, D.J. Balding, Worldwide F_{ST} estimates relative to five continental-scale populations, Ann. Hum. Genet. 78 (2014) 468–477.

[4] C.B. Stringer, P. Andrews, Genetic and fossil evidence for the origin of modern humans, Science 239 (1988) 1263–1268.

[5] C.C. Cockerham, Variance of gene frequencies, Evolution 23 (1969) 72–84.

[6] S. Wright, Evolution in Mendelian populations, Genetics 16 (1931) 97–159.

[7] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.

[8] National Research Council, National Research Council Committee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.

[9] K.L. Ayres, A.D.J. Overall, Allowing for within-subpopulation inbreeding in forensic match probabilities, Forensic Sci. Int. 103 (1999) 207–216.

[10] J.M. Curran, C.M. Triggs, J.S. Buckleton, B.S. Weir, Interpreting DNA mixtures in structured populations, J. Forensic Sci. 44 (1999) 987–995.

[11] B.S. Weir, The rarity of DNA profiles, Ann. Appl. Stat. 1 (2007) 358–370.

[12] B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution 38 (1984) 1358–1370.

[13] B.S. Weir, W.G. Hill, Estimating F-statistics, Ann. Rev. Genet. 36 (2002) 721–750.

[14] G. Bhatia, N. Patterson, S. Sankararaman, A.L. Price, Estimating and interpreting F_{ST}: the impact of rare variants, Genome Res. 23 (2013) 1514–1521.

[15] T.J. Pemberton, M. DeGiorgio, N.A. Rosenberg, Population structure in a comprehensive genomic data set on human microsatellite variation, G3 Genes Genomes Genet. 3 (2013) 891–907.

[16] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Academic Press, New York, 2014.

[17] D.C. Steele, D.J. balding, Choice of population database for forensic DNA profile analysis, Sci. Justice 544 (2014) 487–493.