Statistical Evaluation of Haploid Genetic Evidence

T. Egeland[1] and A. Salas*[,2]

[1]*Department of Medical Genetics, Ullevaal University Hospital, 0407 Oslo, Norway*

[2]*Unidade de Xenética, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, 15782, and Grupo de Medicina Molecular, Hospital Clínico Universitario, 15706 Galicia, Spain*

Abstract: A variety of forensic cases need to be based on haploid DNA. The statistical evaluation of the genetic evidence in such cases requires particular attention and methods. There may be considerable differences between databases and this must be accounted for when there is uncertainty about the population origin of the perpetrator. We here assume access to the relevant databases sampled from populations that the perpetrator may conceivably come from. Intuitively, if there are strong reasons to claim that a specific database is particularly relevant, the likelihood ratio should be influenced accordingly. Moreover, the haploid evidence is typically weaker than for nuclear DNA and there is therefore a greater need to assess uncertainty; here we have chosen the bootstrap method. We also discuss the ability of the proposed method to account for population stratification when computing the *LR*. A pilot implementation of the haploeve software is freely available.

INTRODUCTION

Haploid genetic data, in particular mtDNA and the non-recombining part of the Y chromosome, is commonly used in many forensic applications (e.g. Carracedo *et al*. 2000; Gill *et al*. 2001) [1–2]. For instance, when the biological material is degraded and only mtDNA may be analysed, or in a rape case in which Y-chromosome data could be more appropriate than other markers (e.g. in order to facilitate the analysis of the DNA mixture coming from a female victim and a male suspect); such data may serve to strengthen the case against the suspect or the suspect may be cleared. Other forensic applications involve identification ranging from maternity/paternity cases to larger pedigrees, perhaps extending over several generations. Such applications may extend well beyond the forensic field. Thus, mtDNA and Y-chromosome data have played a central role in disentangling the ancient and recent past of human populations or their demographic movements and are commonly used in medical studies as well. Note that many of these applications are possible since these markers are transmitted from parents to their offspring in a matrilineal (mtDNA; see Bandelt *et al*. 2005 for a recent discussion) [3] or patrilineal way as pure haplotypic blocks. This mode of inheritance has important implications for the interpretation of evidence since siblings typically share mtDNA and brothers the non-recombining part of the Y chromosomes.

In a forensic context, it is generally important to assess uncertainty when evaluating the weight of the evidence. Although several suggestions (e.g. Roewer *et al*. 2000; Wilson *et al*. 2003; Balding 2005) [4–6] have been proposed in the last few years, we here aim to expand on this problem by considering more complex and common scenarios. In particular, our main aim is to provide a practical solution to the statistical evaluation of the mtDNA and Y-chromosome forensic tests when there is uncertainty regarding the appropriate reference database that should be used.

Before we present the approach of the paper, some words on alternative procedures and corrections for coancestry are in order. A simple method is to decide on one database for the evaluation of the evidence. According to Buckleton *et al*. (2005) [7], "…it is imperative that every effort should be made to use appropriate local databases and hence no correction or low value for…". While this may sometimes be adequate and lead to reasonable evaluation of evidence, it may be questionable to exclude alternative databases. At the other extreme, evidence for all conceivable databases may be evaluated and reported. However, this may be impractical. For instance, the number of databases may be prohibitively large; in our last example (Examples 1–4 are provided as supplementary material), there are 91 databases and it would not be practical to report 91 *LRs*. Besides, it may not be helpful for the decision makers. Moreover, it may appear reasonable to somehow weigh these databases as we propose. We offer a consistent and flexible approach that allows all relevant information to be included.

EVALUATING THE WEIGHT OF THE HAPLOID FORENSIC EVIDENCE

The context is the following: a crime has been committed and a DNA crime sample denoted E_C has been obtained. A suspect is apprehended and a suspect's DNA sample E_S is obtained. We do not consider the case when the suspect is found following the search of a database; see Storvik and Egeland (2007) [8] for a recent discussion of this problem. If the DNA profile of E_C and E_S differ, the apprehended suspect is exonerated (see, however, Salas *et al*. 2006b) [9] for some discussion of the potential misuses of mtDNA as an exclusion tool). We thus consider the case in which these samples coincide. We ignore the possibility of intergenerational mutation and heteroplasmic events as well as the possibility of lab errors (Bandelt *et al*. 2002; Bandelt *et al*. 2004a; Bandelt *et al*. 2004b; Salas *et al*. 2005a; Salas *et al*. 2005b; Salas *et al*. 2006b) [9–14]. The models and data required to include these rare effects are not available in our view.

*Address correspondence to this author at the Unidade de Xenética, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, 15782, and Grupo de Medicina Molecular, Hospital Clínico Universitario, 15706 Galicia, Spain; E-mail: apimlase@usc.es

1874-4028/08 2008 Bentham Science Publishers Ltd.

To evaluate the evidence, access to one or more databases is needed. The definition of the group or population will depend on the context and specific application. In Example 3 (see Supplementary Material), the following seven groups are given self-explanatory names: Andalucia (South Iberia), Basque Country, Catalonia (Northeast Iberia), Galicia (Northwest Iberia), Central Portugal, North Portugal and South Portugal (see below for references). Note that our choice here consists of samples of populations which are very closely (geographically and genetically) related, which would allow us to see the effect of different sample sizes, haplotype diversity and/or population substructure (if any) in a relatively small geographic area (Iberia). The groups are disjoint and their union encompasses all possibilities. The data used in Example 4 ((see p5) is quite different, consisting of more than 12,000 Y-STR haplotypes from 91 different populations. Our approach is relevant when databases based on samples from different populations or groups differ. In other words, the frequency of a sample may vary considerably for different databases. Consequently, what follows is particularly relevant for mtDNA and Y-chromosome samples. Consider the following hypotheses:

- H_P: the suspect is the source of the crime sample.
- H_D: some unknown person, unrelated to the suspect, is the source of the sample.

Observe that a large number of alternative suspects are not considered. For instance, Y-STR samples cannot be used to distinguish members of the same paternal lineage. The numerical evaluation of these hypotheses will be based on the following information and data:

1. Prior probabilities $p_j = P(E_C \text{ comes from population } j), j = 1,\ldots,G$. Prior here implies that the databases have not yet been considered (see Sheehan and Egeland 2007 [15] for a discussion of priors).
2. Databases $B = (B_1,\ldots,B_G)$.
3. E_C and E_S.

Observe that deciding in advance to use only one specific database, say the Central Portuguese, corresponds to assigning prior 0 to the others. A more thorough discussion of the role of the prior and other aspects of the model is deferred to the discussion section.

The likelihood ratio (LR) becomes:

$$LR = \frac{P(E_C, E_S \mid B, H_P)}{P(E_C, E_S \mid B, H_D)} =$$

$$\frac{P(E_C \mid E_S, B, H_P)P(E_S \mid B, H_P)}{P(E_C \mid E_S, B, H_D)P(E_S \mid B, H_D)} = \frac{1}{P(E_C \mid E_S, B, H_D)}$$

where the assumptions underlying the latter equation are the usual ones. In fact, the above equation essentially coincides with Eq. (2.3) of Evett and Weir (1998) [16]. The only differences are notational and the fact that databases are included explicitly. It may or may not be in order to remove the conditioning on E_S, as discussed in Evett and Weir (1998) [16]. Below, the conditioning is retained and the denominator may be written

$$P(E_C \mid E_S, B, H_D) =$$
$$\sum_{j=1}^{G} P(E_C \mid pop.\, j, E_S, B, H_D) P(pop.\, j \mid E_S, B, H_D) = \sum_{j=1}^{G} w_j p_j^* = L \qquad (1.1)$$

Here, w_j is the match probability assuming the sample to originate from population j. The estimation of w_j is discussed in the next section. Furthermore, p^*_{j} is the posterior probability for population *j*. Note that the approach based on (1.1) generalizes previous suggestions, specifically (1.1) degenerates to a match probability when only one database is used.

The evaluation of the posterior will differ depending on the application and the data. For the examples in this paper, principal component analysis (PCA) and discriminant analysis, as explained in Egeland *et al*. (2004) [17], is appropriate. A general disadvantage of PCA is the lack of scale invariance. A widely used approach to resolving this problem is to use variables as they are coded only if there are some natural or common units. If this is not the case, scaling to unit variance is advisable. The first three examples deal with dichotomized mtDNA data (the data is on the form explained in Egeland *et al*. 2004) [17] and the data should not be scaled prior to PCA. For the Y-chromosome data, the data matrix is scaled to have unit variance since the different markers are potentially on different scales (a transformation of the Y-STRs profiles to a binary form could also be carried out, as in the case of the mtDNA data). Again, it is possible to check different parameter settings. The specific implementation of the method will depend on the data type, and detailed instructions and examples are provided on the accompanying website (http://folk.uio.no/thoree/haploeve/).

To understand better the implications of (1.1), we discuss an example corresponding to two populations based on arti-

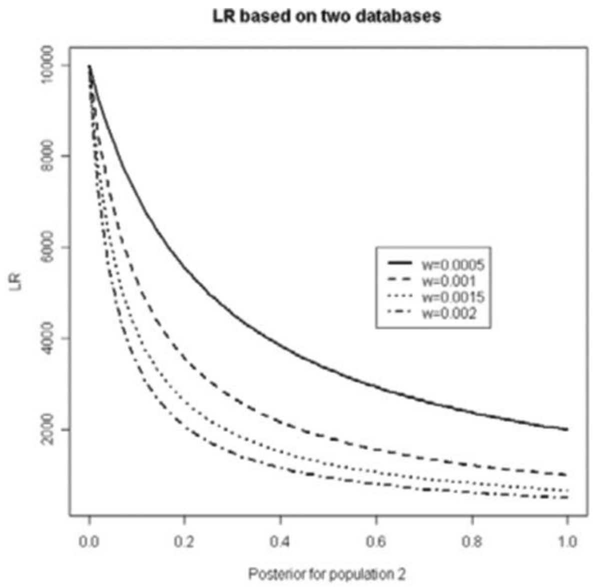

Fig. (1). There are two populations and the posterior for population 2, p^*_2, is on the x-axis. The four curves correspond to four different match probabilities in population 2. The match probability for population 1 is set to 0.0001. The match probability for the solid curve is 0.0005. When $p^*_2 = 1$

$$L = w_1 \times p_1^* + w_2 \times p_2^* = w_2 = 0.0005.$$

From this, LR = 1/0.0005 = 2000.

ficial data. The match probability for population 1, w_1, is 0.0001. The four curves of Fig. (**1**) correspond to match probabilities 0.0005, 0.0010, 0.0015 and 0.0020 in population 2. The curve giving the higher *LR* values corresponds to the lower match probability. The posterior probability for population 2, p^*_2, is on the x-axis. When $p^*_2 = 0$ then $p^*_1 = 1$ and, according to (1.1), $L = w_1 = 0.0001$ and $LR = 10000$, as can be seen from Fig. (**1**). For other values of the posteriors, the match probability of population 2 also matters.

UNSEEN SAMPLES

If the sample *E* is never seen, $L = 0$ and the *LR* is infinite. Different suggestions have been proposed in the literature to avoid this unfortunate *LR* estimate. A practical proposal is to let

$$\begin{aligned} w_i &= x_i / (n_j + k) \ if \ x_i > 0 \\ w_i &= k / (n_j + k) \ if \ x_i = 0 \end{aligned} \qquad (1.2)$$

where the haplotype is observed x_i times and n_j is the size of the database. The choice $k = 1$, which is our default, amounts to adding the evidentiary sample to the database; $k = 2$ corresponds to adding both the defendant and culprit profiles. This topic is discussed in several papers including Curran *et al*. (2002), Wilson *et al*. (2003) and Balding (2005) [5, 6, 18]. Our suggestion corresponds to Morton's proposal (Morton 1992) [19] for $k = 1$. While Morton acknowledges that (1.2) may be criticized, he also notes the resemblance with Laplace's Rule of Succession (see e.g. http://en.wikipedia.org/wiki/Rule_of_succession) and the justification this gives. The present context differs from the settings previously discussed since we are explicitly modelling several populations. However, our methods and software can also be used when there is only one database and results should coincide with previous recommendations for this case. It is also possible to compute match probabilities accounting for coancestry using formula (6.9) in Balding (2005) [6], as we will discuss in more detail in the last section.

INCLUDING UNCERTAINTY

There are various ways of including assessment of uncertainty, e.g. credibility intervals associated with (1.1) and the resulting *LR*. The method we have implemented is based on bootstrapping: a sample is drawn with replacement from the existing and the p-s and w-s of (1.1) are estimated. By repeating this, say 100 times, a distribution for the *LR* is obtained.

DISCUSSION

We have provided an intuitive method to compute *LRs* with uncertainty when there is also uncertainty about the appropriate database to use. There are alternative procedures and we will comment briefly on two of these before we discuss the model of the article in greater detail. Firstly, it is possible to report match probabilities rather than *LRs*. However, there is a simple relation between *LR* and match probabilities in the applications motivating the paper, so technically there is not much difference. Obviously, verbal versions of the numerical evidence will differ. Secondly, a Bayesian formulation entails calculating the posterior probability of the event 'NN is the source of the sample' (see Balding and Donnelly 1995) [20]. However, such an approach may not be admissible in crime cases as it may be seen to interfere with the role of the judge or jury.

As mentioned previously, haploid evidence cannot be used to prove individuality. Typically, a number of people share the same haplotype. Our formulation assumes that alternative (very closely related) suspects with the same haplotype have been cleared out of the case.

The use of prior probabilities in forensic settings has been criticized. We agree that the forensic expert should avoid introducing subjective prior information and at first sight it may therefore appear strange that priors appear in the model. However, this is due to a number of reasons: most importantly, the prior is not directly related to guilt. Furthermore, the prior need not be specified by the expert. Other approaches inevitably also involve the use of prior information, but in disguise: choosing one specific database corresponds to a very strict prior. Specifically, using, say, the Central Portuguese database in a case corresponds to assigning prior 0 to the others. The explicit use of priors can be avoided by presenting *LRs* (or match probabilities) for all databases and leaving it to the decision maker to integrate the evidence. However, as we have seen, this is impractical for some of the applications we have in mind. The use of a prior serves to make assumptions explicit, facilitating critical examination. As a technical aside, updating information or evidence is convenient using Bayes theorem. If one is ignorant of or hesitant to specify priors, as will often be the case, a flat prior can be used. As often happens, priors can be criticized: it may be considered more reasonable to weigh according to population size. However, what population size is relevant? In a particular vicinity of the crime scene? Again, priors are difficult, but we maintain that the problem cannot be avoided and, as always, several calculations can be made to examine the impact of the prior. Finally, priors and posteriors are, in a sense, relative terms. We have considered prior to precede the use of the database and the evidence of the case. However, if new evidence is made available, for example initially only mtDNA was available and at some later point Y-STRs were made available, then the posterior probability computed based on mtDNA may serve as a prior when the Y-STR data is introduced.

It can tentatively be said that this explicit modelling of population substructure could replace the correction for coancestry (assuming there is no further substructure within groups). In other words, theta corrections, as summarized by Eq. (6.9) in Balding (2005) [6], may not be needed. More explicitly, let us imagine we lump together different e.g. Iberian data sets (of different sample sizes, ranging from 50 to 196 individuals; see above and also Table **2**) into a single database. We might wish to correct our *LR* values using an *Fst* correction (*Fst* is ~0.007 for our Iberian data sets; data not shown). The range of *LR* estimates using *Fst* = 0.007 applied to (6.9) of Balding (2005) [6] approximates values computed using the approach of (1.1) applied to the seven Iberian data sets. Determining the appropriate correction for population stratification is not within the scope of this article. It is, however, worth mentioning that, to our knowledge, it has not been formally demonstrated that *Fst* appropriately corrects *LR* estimates in the presence of population stratification when using haploid data. For instance, distribution of population genetic variation contains information that is lost when characterized by a single measure such as *Fst* (Neigel 2002) [21]; and this is certainly the case with mtDNA and the Y-chromosome data, both containing variation with phylogeographic structure (Avise 2000) [22]. In other words, different lineages (haplogroups) can have different local geographical patterns as a result of different local demographic historical events (see, for instance, the Iberian distribution of haplogroup V (Torroni *et al*. 2001) [23]. Therefore, applying the same universal *Fst* correction to the whole, say Iberian, database could be inappropriate because different lineages could differ in their frequencies at a more regional or local geographical scale.

Note that there should be some a priori reason to include a particular population database in (1.1); for instance, it may not be reasonable to add e.g. a sub-Saharan database in the Iberian example above if the crime took place in a specific Iberian region. This is because if the evidentiary mtDNA consists by chance of a sub-Saharan haplotype (which is rare in Iberia), *L* in (1.1) could be overestimated due to an inflation of the term in (1.1) corresponding to the sub-Saharan data set added. A high posterior probability would then be assigned to this sub-Saharan database concomitant with a relatively high frequency for this profile within this database. Note in addition that this undesirable effect would not occur for a typical Iberian mtDNA profile (for these profiles, the posterior probability regarding the sub-Saharan subset and the frequency of these Iberian profiles within this African data set would be low, and then its effect on the global *L* would be insignificant). The use of phylogenetic/phylogeographic criteria could help to evaluate the implications of considering particular databases in (1.1) and to what extent undesirable effects could be affecting *LR* estimates.

It is difficult to provide a universal and practical recommendation as to which population databases to include when using (1.1). A practical choice would be to consider populations which are geographically related (such as the example given above for Iberia). In a broader geographical context (e.g. Europe), it may be tempting to lump together population samples (more or less genetically homogenous) and then apply (1.1). Obviously, to lump together for example the different SWGDAM databases and apply (1.1) is not feasible because of the uncertainty regarding the 'ethnicity' of the groups (if any; see Salas *et al*. 2006 [9]) that is supposed to be represented in these subsets. Moreover, the different sample sizes of these subsets, which need not be proportional to the 'ethnic' groups that they are supposed to represent, pose further problems. Population stratification within the US territory would be another drawback if we take into account the way the SGWDAM was conceived (Salas *et al*. 2006) [9].

An alternative to (1.1) would be to select the population data set with the highest posterior probability, and use only this data set to estimate the *LR*. This option tends to underestimate the *LR*; this choice could be overly conservative. Again, phylogeographic criteria could assist decision making.

Last but not least, there are problems related to forensic databases. Generally, these databases do not fit forensic aims and do not adequately represent the universe of possible perpetrators. This is, however, a matter of different nature beyond the scope of the present study.

Note that the use of e.g. phenotypic or cultural features (e.g. the US 'racial' standards) of the suspects as criteria for deciding which particular data set to apply makes no sense. Thus, the make-up of the SWGDAM database provides a good example. It is naïve to believe that the 'hispanic' SWGDAM data set consists of a homogeneous genetic unit (Salas *et al*. 2006) [9]; therefore, it makes no sense to classify an individual as 'Hispanic' with the purpose of using the 'Hispanic' data set to estimate a *LR*. The decision on which of the SWGDAM data sets to apply to a particular forensic case (occurring in a multi-ethnic city such as New York) could be completely arbitrary, and this will generally have important implications for the computation of the *LR*. Certainly, this problem adds to other important deficiencies of the different nature of the SWGDAM (Bandelt *et al*. 2001; Bandelt *et al*. 2004a; Salas *et al*. 2005a; Salas *et al*. 2006a) [9,11,13,24].

It is also worth mentioning that different sample sizes could obviously affect the uncertainty when using (1.1). For the example of Fig. (**1**), if one population is large, say population 1, then w_1 will be almost constant for all simulations, while if population 2 is small, then w_2 will vary in the simulation, so the greatest contribution to uncertainty will come from small databases with high posteriors. Therefore, it can be said that our approach implicitly accounts for variable sample sizes.

We realize that there may be additional problems related to the size and the representativity of databases. In particular, the databases may not be large enough to encompass all different haplotypes with acceptable probability. Then, adjustments to frequency estimates may be reasonable. Thus, the frequency estimate generally depends heavily on the sample size, and ignoring phylogeographic circumstances could lead to an overestimation of haplotype frequencies. For example, a typical sub-Saharan haplotype is probably rare in the Andalucian population in comparison with many other haplotypes that are probably relatively common in this region but that still remain unsampled due to the stochastic effects inherent in the sampling process.

ACKNOWLEDGEMENTS

The 'Ramón y Cajal' Spanish programme from the Ministerio de Educación y Ciencia (RYC2005-3), the grant from the Xunta de Galicia (PGIDIT06PXIB208079PR), and a grant from the Fundación de Investigación Médica Mutua Madrileña awarded to AS partially supported this project.

REFERENCES

[1] Carracedo, Á.; Bär, W.; Lincoln, P.; Mayr, W.; Morling, N.; Olaisen, B.; Schneider, P.; Budowle, B.; Brinkmann, B.; Gill, P.; Holland, M.; Tully, G.; Wilson, M. *Forensic Sci. Int.*, **2000**, *110*, 79-85.

[2] Gill, P.; Brenner, C.; Brinkmann, B.; Budowle, B.; Carracedo, Á.; Jobling, M.A.; de Knijff, P.; Kayser, M.; Krawczak, M.; Mayr. W.R.; Morling, N.; Olaisen, B.; Pascali, V.; Prinz, M.; Roewer, L.; Schneider, P.M.; Sajantila, A.; Tyler-Smith, C. *Forensic Sci. Int.*, **2001**, *124*, 5-10.

[3] Bandelt, H.-J.; Kong, Q.-P.; Parson. W.; Salas, A. *J. Med. Genet.*, **2005**, *42*, 957-60.

[4] Roewer, L.; Kayser, M.; de Knijff, P.; Anslinger, K.; Betz, A.; Caglia, A.; Corach, D.; Furedi, S.; Henke, L.; Hidding, M.; Kargel, H.J.; Lessig, R.; Nagy, M.; Pascali, V.L.; Parson, W.; Rolf. B.; Schmitt, C.; Szibor, R.; Teifel-Greding. J.; Krawczak. M. *Forensic Sci. Int.*, **2000**, *114*, 31-43.

[5] Wilson, I.; Weale, M.; Balding. D. *J. R. Statist. Soc. A.*, **2003**, *166*, 155-201.

[6] Balding, D.J. Weight-of-evidence for forensic DNA profiles; Wiley, London, **2005**.

[7] Buckleton, J.S.; Triggs, C.M.; Walsh, S.J. Forensic DNA evidence interpretation; CRC Press, **2005**.

[8] Storvik, G.; Egeland, T. *Biometrics*, **2007**, *63*, 922-5.

[9] Salas, A.; Bandelt, H.-J.; Macaulay, V.; Richards, M. *Forensic Sci. Int.*, **2006**, *168*, 1-13.

[10] Bandelt, H.-J.; Quintana-Murci, L.; Salas, A.; Macaulay, V. *Am. J. Hum. Genet.*, **2002**, *71*, 1150-60.

[11] Bandelt, H.-J.; Salas. A.; Bravi. C. *Science*, **2004**, *305*, 1402-4.

[12] Bandelt, H.-J.; Salas, A.; Lutz-Bonengel, S. *Int. J. Legal Med.*, **2004**, *118*, 267-73.

[13] Salas, A.; Carracedo, Á.; Macaulay, V.; Richards, M.; Bandelt, H.-J. *Biochem. Biophys. Res. Commun.*, **2005**, *335*, 891-9.

[14] Salas, A.; Prieto, L.; Montesino, M. ; Albarrán, C. ; Arroyo, E. ; Paredes-Herrera, M. R. ; Di Lonardo, A. M. ; Doutremepuich, C. ; Fernández-Fernández, I. ; de la Vega, A. G. ; Alves, C. ; López, C. M. ; López-Soto, M. ; Lorente, J. A. ; Picornell, A. ; Espinheira, R. M. ; Hernández, A. ; Palacio, A. M. ; Espinoza, M. ; Yunis, J. J. ; Pérez-Lezaun, A. ; Pestano, J. J. ; Carril, J. C. ; Corach, D. ; Vide, M. C. ; Álvarez-Iglesias, V. ; Pinheiro, M. F. ; Whittle, M. R. ; Brehm, A. ; Gómez, J. *Forensic Sci. Int.*, **2005**, *148*, 191-8.

[15] Sheehan, N.A.; Egeland, T. *Ann. Hum. Genet.*, **2007**, *71*, 501-18.

[16] Evett, I.W.; Weir, B.S. Interpreting DNA Evidence; Sinauer, **1998**.

[17] Egeland, T.; Bøvelstad, H.M.; Storvik, G.O.; Salas, A. *Ann. Hum. Genet.*, **2004**, *68*, 461-71.

[18] Curran, J.M.; Buckleton, J.S.; Triggs, C.M.; Weir, B.S. *Sci. Justice*, **2002**, *42*, 29-37.

[19] Morton, N.E. *Proc. Natl. Acad. Sci. USA*, **1992**, *89*, 2556-60.

[20] Balding, D.J.; Donnelly. P. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 11741-5.

[21] Neigel, J.E. *Conservation Genet.*, **2002**, *3*, 167-73.

[22] Avise, J.C. Phylogeography: the history and formation of species; Harvard University Press: Cambridge MA, **2000**.

[23] Torroni, A.; Bandelt, H.J.; Macaulay, V.; Richards, M.; Cruciani, F.; Rengo, C.; Martinez-Cabrera, V.; Villems, R.; Kivisild, T.; Metspalu, E.; Parik, J.; Tolk, H.V.; Tambets, K.; Forster, P.; Karger, B.; Francalacci, P.; Rudan, P.; Janicijevic, B.; Rickards, O.; Savontaus, M.L.; Huoponen, K.; Laitinen, V.; Koivumäki, S.; Sykes, B.; Hickey, E.; Novelletto, A.; Moral, P.; Sellitto, D.; Coppa, A.; Al-Zaheri, N.; Santachiara-Benerecetti, A.S.; Semino, O.; Scozzari, R. *Am. J. Hum. Genet.*, **2001**, *69*, 844-52.

[24] Bandelt, H.-J.; Lahermo, P.; Richards, M.; Macaulay, V. *Int. J. Legal Med.*, **2001**, *115*, 64-9.

Received: December 31, 2007 Revised: February 06, 2008 Accepted: February 06, 2008

SUPPLEMENTARY MATERIAL

In what follows, we provide some examples to illustrate the methods proposed in previous sections. Data sets, R (http://www.r-project.org/) code, and a tutorial are available from http://folk.uio.no/thoree/haploeve/.

Example 1

Excerpts from a simulated data set are shown in Table **1**. The data is only intended to illustrate the methods. There are three groups or populations, numbered 1, 2 and 3, all containing 100 individuals. We assume a priori that the three populations are equally likely. There are 10 sites considered and a 1 indicates a deviance from a reference sample. Individuals from population i have a 1 for site i with probability 0.8. For the remaining sites, the probability of a 1 is 0.2. Assume a person has a profile with a 1 for the first and last sites and 0 for the remaining sites. Such a profile is seen three times for population 1 and never for the two other populations. The following posteriors (based on PCA) are obtained for populations 1, 2 and 3, respectively: 0.941, 0.027 and 0.032. Using (1.1), the likelihood of the data under the alternative hypothesis becomes

$$L = w_1\,0.941 + w_2\,0.022 + w_3\,0.031.$$

If the match probabilities correspond directly to the observed frequencies, then $w_1 = 3/100$, $w_2 = w_3 = 0$, while $L = 0.941 \times 3/100 = 0.0282$ and LR = 1/0.0282 = 35.4. If, on the other hand, Eq. (1.2) is used with default option $k = 1$, then, $w_1 = 0.0297$, $w_2 = w_3 = 0.0099$ and $LR = 35.0$.

Table 1. Excerpts of the Data, Corresponding to Four Individuals (ID), Used in Example 1. The Column 'gr' Indicates Population Origin whereas X1–X10 Display the Haplotype

ID	gr	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	1	1	0	0	1	0	1	0	0	0	1
2	1	0	0	0	0	0	0	0	0	0	0
299	3	0	0	1	0	1	1	0	0	0	0
300	3	0	0	1	0	1	0	1	0	0	0

The bootstrap procedure with 100 samples and default options leads to a 95% interval ranging from 14.9 to 101.0. Using $Fst = \theta = 0.01$ reduces the LR (following Balding 2005) [1] further to 26.1 with a 95% bootstrap interval (13.1, 50.5).

Example 2

Consider next mtDNA from three populations: Germans (N = 1,314; Pfeiffer *et al.* 1999; 2001) [2–3], Icelandic (N = 396; Helgason *et al.* 2000) [4] and Mozambicans (N = 309; Salas *et al.* 2002) [5]. For our example below, we used the most frequent sequence of the Mozambicans as the evidentiary sample, one occurring in 46 or 15.0% of the individuals (C16148T T16172C C16187T T16188G C16189T C16223T A16230G T16311C C16320T; Salas *et al.* 2002) [5] and never seen in the Europeans (Salas *et al.* 2004) [6]. This sequence is classified as Mozambican with probability virtually equal to 1 and we expect and find an LR close to 1/0.15 = 6.666, namely 6.7. A 95% interval is 5.6–8.8 while a histogram is shown in Fig. (**2**). Using only the Mozambican database gives a virtually unchanged result. If, on the other hand, only the German and Icelandic databases are used, the resulting LR changes dramatically to 667.8.

This example illustrates the implications of considering particular data sets when computing the LR (see below). The scenario represented by these three data sets (or whatever combination of other data sets) may not be realistic and the expertise and choices of databases made could be critical.

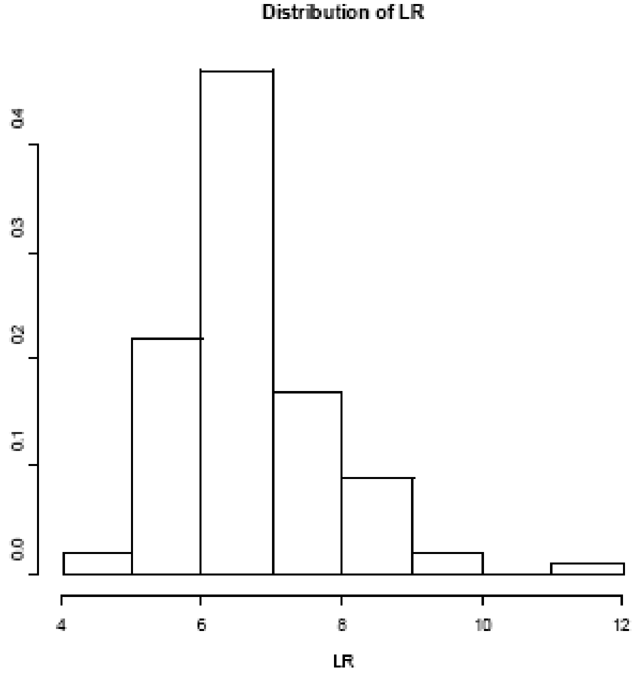

Fig. (2). Distribution of LR in Example 2. The evidentiary sample is the most frequent sequence of the Mozambicans. Three databases are combined to obtain the distribution of the LR. Using only the database of Mozambique would lead to almost identical distribution.

Example 3

Consider next the seven databases of Table **2** (the first hypervariable region of each specified data set): Portugal (N = 540; González *et al.* 2003, Pereira *et al.* 2004) [7–8], Basques (N = 171; Bertranpetit *et al.* 1995, Côrte-Real *et al.* 1996, Richards *et al.* 1996) [9–11], Catalonia (N = 118; Crespillo *et al.* 2000) [12], Galicia (N = 135; Salas *et al.* 1998, González *et al.* 2003) [7, 13]. These are more (genetically) homogeneous populations than those of the previous example. Assume a sample corresponding to the revised Cambridge Reference Sequence (rCRS; Andrews *et al.* 1999) [14] is derived from the defendant and the scene of the crime. The number of observations of this haplotype is given in Table **2**. Our software, with default options (e.g. flat priors for the probabilities to all the subsets), gives LR 4.9 and 95.0% interval 4.3 to 5.6.

It is also of interest to study the range of likelihood values that can be expected in a database. Thus, for example, there are 1,014 samples in the Iberian database. Their values range from 4.9, corresponding to the previously mentioned rCRS, to 1039.1. Fig. (**3**) shows the distribution corresponding to a random sample of 100 of the 1014 haplotypes.

Table 2. The Data Used for Example 3. The Frequency of the rCRS is Indicated in the Rightmost Column

Group	*N*	obs
Andalucia	50	7
Basques	171	38
Catalonia	118	20
Galicia	135	34
PortCent	160	32
PortNorth	184	40
PortSouth	196	41

Example 4

In this example, we consider a large database of Y-chromosome data (http://www.yhrd.org/; see also Roewer *et al.* 2000) [15]. Table **3** shows three haplotypes of the 12,727 haplotypes from 91 populations; the names of the populations are given in the rightmost column. The sample sizes in the various populations vary between 25 and 573. This data set differs from the previous ones and those of Egeland *et al.* (2004) [16] in two important senses: the number of markers per individual is much smaller, and more importantly, the markers are not dichotomous. It is therefore reasonable to scale to unit variance prior to PCA.

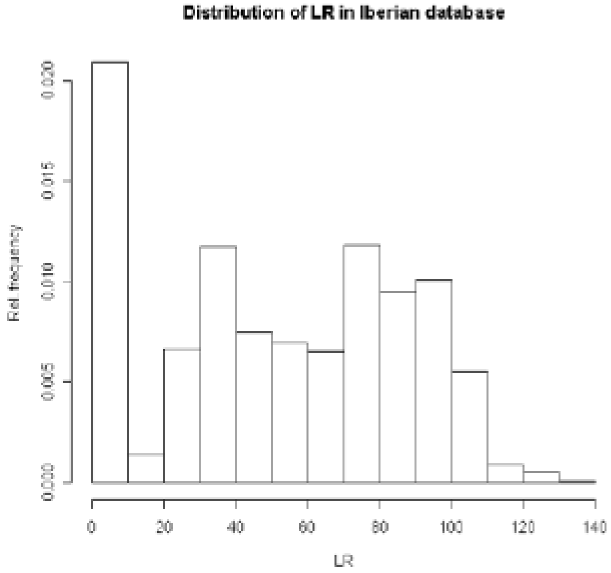

Fig. (3). The distribution of LRs for a random sample of 100 of the 1,014 Iberian samples (see Example 3).

Consider first a common haplotype. Assume first the evidentiary sample is the second in Table **3**, i.e. number 6,476. This is a common haplotype observed 661 times, corresponding to 5.1%, and seen in 80 of the 91 populations. The resulting *LR* 14.3 is lower than the naive *LR* obtained as 1/0.052 = 19.2, and this makes sense since the latter includes databases which may not be appropriate for this evidentiary sample. The 95% interval is 12.6 to 15.3.

Consider next a much rarer haplotype: the first of the below table. This is observed three times (Albania, Budapest [Hungary], Leipzig [Saxony]). Assume first there is no prior information on the origin of the sample. Then *LR* 94.8 (88.8, 98.0).

Prior information: if one has some prior information indicating that the sample is of, say, Albanian origin, the above analysis may not be appropriate. One could then choose to use only the Albanian population. This and more general prior information can be included, as explained on the website.

Table 3. Three of 12,727 Haplotypes used in Example 4.

Haplo. No.	Pop.	DYS 19	DYS 389I	DYS 389II	DYS 390	DYS 391	DYS 392	DYS 393	Pop.
1	1	12	13	30	24	10	11	13	Albania
6476	42	14	13	29	24	11	13	13	Madrid, Central East Spain
12727	91	17	13	30	25	10	11	13	Anatolia, Turkey

REFERENCES

[1] Balding, D.J. Weight-of-evidence for forensic DNA profiles; Wiley, **2005**.

[2] Pfeiffer, H.; Brinkmann, B.; Huhne, J.; Rolf, B.; Morris, A.A.; Steighner, R.; Holland, M.M.; Forster, P. Expanding the forensic German mitochondrial DNA control region database: genetic diversity as a function of sample size and microgeography. *Int. J. Legal Med.*, **1999**, *112*, 291-8.

[3] Pfeiffer, H.; Forster, P.; Ortmann, C.; Brinkmann, B. The results of an mtDNA study of 1200 inhabitants of a German village in comparison to other Caucasian databases and its relevance for forensic casework. *Int. J. Legal Med.*, **2001**, *114*, 169-72.

[4] Helgason, A.; Sigurethardottir, S.; Gulcher, J.R.; Ward, R.; Stefansson, K. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.*, **2000**, *66*, 999-1016.

[5] Salas, A.; Richards, M.; De la Fé, T.; Lareu, M.V.; Sobrino, B.; Sánchez-Diz, P.; Macaulay, V.; Carracedo. Á. The making of the African mtDNA landscape. *Am. J. Hum. Genet.*, **2002**, *71*, 1082-111.

[6] Salas, A.; Richards, M.; Lareu, M.V.; Scozzari, R.; Coppa, A.; Torroni, A.; Macaulay, V.; Carracedo, Á. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am. J. Hum. Genet.*, **2004**, *74*, 454-65.

[7] González, A.M.; Brehm, A.; Pérez, J.A.; Maca-Meyer, N.; Flores, C.; Cabrera, V.M. Mitochondrial DNA affinities at the Atlantic fringe of Europe. *Am. J. Phys. Anthropol.*, **2003**, *120*, 391-404.

[8] Pereira, L;. Cunha, C.; Amorim, A. Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *Int. J. Legal Med.*, **2004**, *118*, 132-6.

[9] Bertranpetit, J.; Sala, J.; Calafell, F.; Underhill, P.A.; Moral, P.; Comas, D. Human mitochondrial DNA variation and the origin of Basques. *Ann. Hum. Genet.*, **1995**, *59*, 63-81.

[10] Côrte-Real, H.B.S.M.; Macaulay, V.A.; Richards, M.B.; Hariti, G.; Issad, M.S.; Cambon-Thomsen, A.; Papiha, S.; Bertranpetit, J.; Sykes, B.C. Genetic diversity in the Iberian peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.*, **1996**, *60*, 331-50.

[11] Richards, M.; Côrte-Real, H.; Forster, P.; Macaulay, V.; Wilkinson-Herbots, H.; Demaine, A.; Papiha, S.; Hedges. R.; Bandelt, H.-J.; Sykes, B. Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.*, **1996**, *59*, 185-203.

[12] Crespillo, M.; Luque, J.A.; Paredes, M.; Fernández, R.; Ramirez, E.; Valverde, J.L. Mitochondrial DNA sequences for 118 individuals from northeastern Spain. *Int. J. Legal Med.*, **2000**, *114*, 130-2.

[13] Salas, A.; Comas, D.; Lareu, M.V.; Bertranpetit, J.; Carracedo, Á. mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur. J. Hum. Genet.*, **1998**, *6*, 365-75.

[14] Andrews, R.M.; Kubacka, I.; Chinnery, P.F.; Lightowlers, R.N.; Turnbull, D.M.; Howell, N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **1999**, *23*, 147.

[15] Roewer, L.; Kayser, M.; de Knijff, P.; Anslinger, K.; Betz, A.; Caglia. A.; Corach, D.; Furedi, S.; Henke, L.; Hidding, M.; Kargel, H.J.; Lessig, R.; Nagy, M.; Pascali, V.L.; Parson, W.; Rolf, B.; Schmitt, C.; Szibor, R.; Teifel-Greding, J.; Krawczak, M. A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci. Int.*, **2000**, *114*, 31-43.

[16] Egeland, T.; Bøvelstad, H.M.; Storvik, G.O.; Salas, A. Inferring the most likely geographical origin of mtDNA sequence profiles. *Ann. Hum. Genet.*, **2004**, *68*, 461-71.