Likelihood ratio and posterior odds in forensic genetics: Two sides of the same coin

Amke Caliebe[a]*, Susan Walsh[b,c], Fan Liu[d,e,c], Manfred Kayser[c], Michael Krawczak[a]

[a]*Institute of Medical Informatics and Statistics, Kiel University, Brunswiker Strasse 10, 24105 Kiel, Germany*
[b]*Department of Biology, Indiana-University-Purdue-University-Indianapolis (IUPUI), 723 W. Michigan St. Indianapolis, Indiana, USA*
[c]*Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands*
[d]*Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, No.1 Beichen West Road, Chaoyang District, Beijing 100101, China*
[e]*University of Chinese Academy of Sciences, 19 A Yuquan Rd, Shijingshan District, Beijing 100049, China*

* Corresponding author

E-mail addresses: caliebe@medinfo.uni-kiel.de (AC), walshsus@iupui.edu (SW), liufan@big.ac.cn (FL), m.kayser@erasmusmc.nl (MKa), krawczak@medinfo.uni-kiel.de (MKr)

Highlights

- Likelihood ratio (LR) and posterior odds (PO) are discussed in a forensic genetics context.
- Probabilistic models serve to illustrate the population-dependency of PO and LR.
- Valid PO can be derived empirically in the context of Forensic DNA Phenotyping.
- Owing to a lack of sensible priors, only LR, but not PO, is considered in DNA profiling.

This is the author's manuscript of the article published in final edited form as:

Caliebe, A., Walsh, S., Liu, F., Kayser, M., & Krawczak, M. (2017). Likelihood ratio and posterior odds in forensic genetics: Two sides of the same coin. Forensic Science International: Genetics, 28, 203–210. https://doi.org/10.1016/j.fsigen.2017.03.004

Abstract

It has become widely accepted in forensics that, owing to a lack of sensible priors, the evidential value of matching DNA profiles in trace donor identification or kinship analysis is most sensibly communicated in the form of a likelihood ratio (LR). This restraint does not abate the fact that the posterior odds (PO) would be the preferred basis for returning a verdict. A completely different situation holds for Forensic DNA Phenotyping (FDP), which is aimed at predicting externally visible characteristics (EVCs) of a trace donor from DNA left behind at the crime scene. FDP is intended to provide leads to the police investigation helping them to find unknown trace donors that are unidentifiable by DNA profiling. The statistical models underlying FDP typically yield posterior odds (PO) for an individual possessing a certain EVC. This apparent discrepancy has led to confusion as to when LR or PO is the appropriate outcome of forensic DNA analysis to be communicated to the investigating authorities. We thus set out to clarify the distinction between LR and PO in the context of forensic DNA profiling and FDP from a statistical point of view. In so doing, we also addressed the influence of population affiliation on LR and PO. In contrast to the well-known population dependency of the LR in DNA profiling, the PO as obtained in FDP may be widely population-independent. The actual degree of independence, however, is a matter of (i) how much of the causality of the respective EVC is captured by the genetic markers used for FDP and (ii) by the extent to which non-genetic such as environmental causal factors of the same EVC are distributed equally throughout populations. The fact that an LR should be communicated in cases of DNA profiling whereas the PO are suitable for FDP does not conflict with theory, but rather reflects the immanent differences between these two forensic applications of DNA information.

Keywords: likelihood ratio, posterior odds, causality, Forensic DNA Phenotyping, forensic DNA profiling, genetic evidence

1. Introduction

For decades, DNA profiling has served in forensic practice to facilitate the identification of trace donors. A trace DNA profile, typically comprising a selected number of highly-polymorphic short tandem repeats (STRs), is either compared to the DNA profiles of one or more suspects, or is gauged against one or more databases of DNA profiles of previously convicted persons. When a perfect match is found, i.e. when trace and target individual are of the same genotype at every STR considered, the forensic expert reports the evidential value of their result in the form of a likelihood ratio (LR).

In a forensic context, likelihoods allow weighing of the prosecution and defense hypotheses (Hp and Hd) against each other, which is not feasible by way of probabilities owing to a lack of sensible priors [1, 2]. With G denoting the genetic evidence (i.e. the match between trace and target DNA profile), each likelihood is defined by the conditional probability of G given the respective hypothesis, i.e. $L(Hp|G)=P(G|Hp)$ and $L(Hd|G)=P(G|Hd)$. The likelihood ratio $LR=L(Hp|G)/L(Hd|G)$ then quantifies the relative evidential value of G with a view to decide between Hp and Hd. It must be emphasized, however, that likelihoods are not probabilities because they are not additive (i.e. the joint likelihood of some mutually exclusive hypotheses usually does not equal the sum of the individual likelihoods) and therefore fail a critical formal requirement of probability theory. Instead, the likelihoods of Hp and Hd as well as the resulting LR should be viewed as a measure of rationale belief in either of the two hypotheses.

The primary interest of the court, of course, should be in posterior odds (PO) $P(Hp|G)/P(Hd|G)$. However, this quantity is difficult to grasp in the context of DNA profiling because of its dependence upon the prior odds $P(Hp)/P(Hd)$ which are usually difficult to specify. Prior odds, moreover, are solely in the domain of the judge or jury. Therefore, a consensus has been reached among forensic experts that likelihoods and the LR should constitute the only case-relevant outcome of their experimental work.

In the simplest situation, a case of interest involves a single suspect and a single trace donor. Under Hd, the former would be presumed to have been drawn at random from a certain suspect population. Under Hp, the suspect is the trace donor. If the trace and suspect DNA profiles match, the LR simplifies to the inverse of the so-called '(random) match probability'. The forensic expert would then report this probability and would leave it to the court to evaluate whether the suspect left the trace or not. In principle, the same reasoning can be applied to any courtroom evidence that does not exclude a suspect or a group of suspects from trace donorship.

Forensic DNA Phenotyping (FDP) is a relatively recent development in forensic genetics. It aims at predicting selected externally visible characteristics (EVCs) of a trace donor from their DNA as left behind at the crime scene. We will continue to use the expression 'predict' in this context despite the fact that some scholars have argued that 'prediction' should only be used for future events [3]. This is because any resulting (true or perceived) logical problems can be resolved by referring to the future disclosure of the EVC of the trace donor once they have been identified. The FDP approach bears great potential in cases where DNA profiling failed, for example, because the police have no suspect at all or neither suspect DNA profile matches the trace DNA profile [4-7].

There are some major conceptual differences between DNA profiling and FDP. First, whereas identification by DNA profiling involves at least two DNA samples, namely from trace and suspect, FDP usually works with just the trace DNA. Second, FDP is not meant to yield courtroom evidence but rather to guide the police investigations in cases where DNA profiling failed or was not feasible in the first place. Most notably, from a statistical perspective, FDP yields posterior probabilities of trait phenotypes (i.e. EVCs) from genotypes by way of statistical techniques such as, for example, logistic regression analysis [8-15]. Since this seems to contradict the forensic genetics paradigm of only reporting likelihood ratios, however, some forensic experts have felt uncomfortable about reporting the PO that somebody has a certain EVC level (such as blue eye color).

Currently, practical FDP is feasible only for eye and hair color. In fact, two dedicated DNA systems have been developed and forensically validated for these EVCs, namely IrisPlex for eye color alone [12, 16, 17] and HIrisPlex for simultaneous eye and hair color prediction [11, 18]. For skin color, efficient predictive DNA markers have been proposed as well [19], but these have not been forensically validated yet. For all other EVCs, studies to understand their genetic basis are not advanced enough to allow practical implementation of FDP [7]. This is not to say that no genetic associations have been documented yet for non-pigmentation traits such as male pattern baldness [20, 21], hair structure [22] and extreme body height [23]. Rather, their prediction accuracies are not yet high enough for FDP, which would require more predictive DNA markers to be identified and added to the respective models. Moreover, for any other EVC, genetic studies have only identified the first few genes, providing very limited prediction information, or genetic studies are lacking completely (see recent overview in [7]).

In the following, we will discuss the apparent discrepancy between DNA profiling and FDP from a statistical perspective. In so doing, we will focus upon the dependency, or not, of LR and PO on population affiliation. For DNA profiling, the role of the source population of the trace donor has been well worked out before, not least including a considerable debate about the appropriateness of the so-called ‘random man’ assumption. Thus, it has been argued repeatedly that the population of a potential donor is rarely if ever identical to the database population used for obtaining the match probability [24, 25], and a mathematical procedure known as the ‘theta correction’ was proposed to address this problem analytically [26-29]. So, whilst presenting a general framework of the relationship between LR and PO, we will nevertheless put our considerations into the specific context of FDP. We also propose some recommendations as to how forensic experts should report the results of their experimental work in the courtroom (DNA profiling) or to the investigating authorities (FDP).

2. Material and Methods

For the present study, the performance of IrisPlex-based FDP for eye color [17] was assessed empirically in eight previously published European population samples from Norway (n=547), the United Kingdom (n=498), Estonia (n=579), France (n=616), Italy (n=542), Greece (n=547), Spain (n=511) – all from the EUREYE Study [12] – and the Netherlands – from the Rotterdam Study (n=2364) [17]. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under the receiver operating curve (AUC) were estimated for a prediction model obtained before from a Dutch training set [9] that did not overlap with the Dutch sample used here. All analyses were performed with the R statistics software [30], particularly packages ROCR [31] and `caret` [32].

3. Probabilistic Models

3.1 General set-up

Couched in probability theoretical terms, every forensic application of DNA typing draws upon the causal relationship between at least two random variables, say X and Y. We will henceforth assume that X is causal for Y in the sense that the factual entity represented by X has a biological effect on that represented by Y, and that X is not merely statistically associated with Y. Various methods have been proposed in the scientific literature to arrive at a reasonable degree of belief in causality [33] and we will stipulate that, in the following, sufficient evidence for a causal relationship between X and Y exists.

In DNA profiling, X is an indicator of the identity, or not, of suspect and trace donor whereas Y is an indicator of matching DNA profiles. In FDP, X is a composite genotype whereas Y denotes the EVC of interest expressed by the unknown trace donor. Usually, composite genotype X will comprise both causal and merely associated variants so that its causality for Y is only partly warranted.

3.2 Likelihood ratio (LR) and posterior odds (PO)

Statistical inference is tantamount to the deduction from sample data of the characteristics of a so-called 'random variable', a concept borrowed from probability theory to relate chance experiments (e.g. the roll of a die or a blood pressure measurement) to their possible outcome (e.g. a label or a number). When considering two random variables X and Y at a time, it may be relevant to distinguish whether the variable of interest (henceforth termed the 'inference variable') is causal (i.e. X is the inference variable) or consequential (i.e. Y is the inference variable). The other variable, henceforth referred to as the 'lead variable', corresponds to the data used for inference making. Note that, in the context of forensic genetics, the different levels of the inference variable are often termed 'hypotheses' and the lead variable itself is called 'evidence'.

DNA profiling clearly aims at characterizing a causal inference variable X, namely trace donor-suspect identity, whereas FDP is concerned with a consequential inference variable Y, namely the EVC of interest. It should be pointed out here that, in both cases, the lead variable is a (composite) genotype, which confines the validity of the inference making process to a so-called 'random man' scenario. This means that the predicted phenotype in FDP is the phenotype of an individual randomly drawn from the

population of interest. If additional lead variables, such as eye witness reports or CCTV footage, were to be taken into account for FDP, then the underlying mathematical model would have to be changed accordingly (which, admittedly, can be very complicated).

If X is the inference variable and Y is the lead variable then P(X=x|Y=y) is the posterior probability, and the conditional probability P(Y=y|X=x) can be interpreted as the likelihood L(X=x|Y=y) in generalization of the likelihood concept alluded to in the introduction. Similarly, P(Y=y|X=x) is the posterior probability and P(X=x|Y=y) equals the likelihood L(Y=y|X=x) if Y is the inference variable and X is the lead variable. Note that, in both cases, the posterior probability is the sought-after quantity and the likelihood is only of indirect interest, if any.

Bayes' theorem connects PO and LR *via* the prior odds. For dichotomous random variables X and Y, of which Y is the inference variable, the theorem takes the form

$$\frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{L(Y=1|X=1)}{L(Y=0|X=1)} \frac{P(Y=1)}{P(Y=0)}.$$

Whilst PO and LR can take widely different values *per se*, Bayes' theorem allows one to be calculated from the other if the prior odds P(Y=1)/P(Y=0) are fixed. In this sense, PO and LR are two sides of the same coin. As was pointed out above, this relationship between LR and PO is particularly important in the context of trace-donor identification by DNA profiling, where the forensic expert only reports the LR and lets the court specify the prior odds necessary to turn LR into PO.

In principle, both P(X=x|Y=y) and P(Y=y|X=x) can be estimated empirically by measuring X and Y in a population-representative sample. In many cases, however, it may be either inefficient, or impracticable, or both, to involve a population-representative sample. In this situation, the sample may be stratified by to the lead variable and still allow valid characterization of the inference variable, i.e. estimation of the sought-after posterior probabilities. In some settings, however, even the measurement of the random variables itself may be intricate. For example, empirical estimation of P(X=0|Y=1) in the case of DNA profiling would require a set of persons, representative for the case at hand, with DNA profiles that match each other (i.e. Y=1) but originated from different persons (X=0). Without doubt, such a sample would be very hard to come by.

Fortunately, the possibility of appropriate mathematical modeling obviates the necessity of empirical data in the case of DNA profiling. The conditional probability of a match given a lack of trace donor-suspect identity, i.e. P(Y=1|X=0), can be readily calculated under the random man assumption as the population frequency of the shared genetic profile. Since X is the inference variable, however, P(Y=1|X=0) would still have to be transformed into a posterior probability using Bayes' theorem, which in turn would require knowledge of P(X=0). As has been mentioned above, priors such as this are outside the domain of the forensic experts so that they have to make do, in DNA profiling, with 1/P(Y=1|X=0) as the inverse conditional probability-turned-LR.

3.3 Cofactor Z

The dependency of LR and PO upon population affiliation will be addressed more generically in the following by way of considering a cofactor Z that is associated with Y, and possibly also with X. We will specify two mathematical models relevant to forensic genetics in which Z is explicitly lacking a role in inference making about either X (model A) or Y (model B).

(A) Causal variable X is the inference variable (Fig. 1A) and the conditional distribution of Y given X is independent of Z, i.e. $P(Y=y|X=x,Z=z)=P(Y=y|X=x)=L(X=x|Y=y)$. This means that the association between Y and X is not modified by Z, and that all information included in Z that is relevant to Y is already included in X. In this case, the make-do likelihoods $L(X=x|Y=y)$ are independent of the actual level z of cofactor Z.

(B) Consequential variable Y is the inference variable (Fig. 1B) and the conditional distribution of Y given X is again independent of Z, i.e. $P(Y=y|X=x,Z=z)=P(Y=y|X=x)$. Then, the sought-after posterior probabilities $P(Y=y|X=x)$ are independent of the actual level z of cofactor Z.

3.4 Additional Cofactors

In reality, more than one cofactor will be associated with consequential variable Y. For the sake of simplicity, we will consider only one such cofactor Z' and assume that $P(Y=y|X=x,Z'=z')\neq P(Y=y|X=x)$ for at least one level triplet (x,y,z'), i.e. Z' modifies the conditional distribution of Y given X. We will confine our considerations to model B, but the conclusions drawn from this can be generalized to multiple cofactors or to model A as well.

3.4.1 Z' is not associated with Z

If Z' is not associated with Z, then Z does not itself modify the conditional distribution of Y given X (Fig. 2A), i.e. model B applies and the sought-after posterior probabilities $P(Y=y|X=x)$ are independent of cofactor Z. For example, if Z denotes population affiliation and Z' is a population-independent cofactor, such as sex, this means that the above posterior probabilities can be estimated in any population. This notwithstanding, since Z' modifies the conditional distribution of Y given X, including Z' in a statistical model used to estimate the above posterior probabilities would improve the precision of the latter, and hence their predictive utility. The expanded model would also be applicable in all populations. Generalizing this statement to multiple cofactors, we may conclude that model B applies, i.e. $P(Y=y|X=x,Z=z)=P(Y=y|X=x)$ for any triplet (x,y,z), if all causes Z' of Y, other than Z, are independent of Z.

3.4.2 Z' is associated with Z

If Z' is associated with Z, then the conditional distribution of Y given X is modified by Z under virtually all realistic assumptions about their stochastic relationship (Fig. 2B). In other words, model B no longer applies. For a discrete cofactor Z such as, for example, population affiliation, this would imply that the sought-after conditional probabilities $P(Y=y|X=x)$ must be estimated separately in each Z-defined stratum. Again, the results would be more precise if Z' was included in the underlying model and, under certain conditions, consideration of Z' may even render the posterior probabilities independent of Z.

3.4 Examples

3.4.1 Example 1: Diagnosis of a disease using a classical biomarker [model A]

Although the prediction of disease phenotypes from forensic DNA samples is ethically problematic, legally banned in most countries, and therefore typically not considered in FDP, we nevertheless chose a medical example to illustrate model A. We assume that the presence of a certain disease X is causal for some biomarker Y, and that this causal relationship is independent of the population Z in which the biomarker is measured, i.e. $P(Y=y|X=x,Z=z)=P(Y=y|X=x)$. For example, the immunoassay-based HIV screening test is sensitive to detect HIV infection because the presence of viral antigen (i.e. X=1) is causal for the presence of antibody (i.e. Y=1), which is the biomarker in question.

For medical diagnostic purposes, infection status X is the inference variable and biomarker status Y is the lead variable. The quantities of interest are the posterior probabilities $P(X=x|Y=y)$. Since model A applies, the likelihoods $L(X=x|Y=y)=P(Y=y|X=x)$ are population-independent which implies that the LR is population-independent as well and can be estimated from case-control or population-representative data (prospective, cross-sectional). The LR also does not depend upon the prior probability (i.e. the prevalence) of HIV infection. Estimation of the PO, on the other hand, would require prevalence information which can only be obtained from population-representative data. Moreover, if the prevalence of HIV infection in population z is known, then the sought-after PO can also be calculated from the LR using Bayes' theorem.

In a medical diagnostic context, such as HIV detection, likelihoods $P(Y=1|X=1)$ and $P(Y=0|X=0)$ would be referred to as 'sensitivity' and 'specificity' of the test. Both parameters are thus population-independent if model A applies. The positive and negative predictive values $P(X=1|Y=1)$ and $P(X=0|Y=0)$, on the other hand, depend upon the prevalence of HIV infection and, therefore, upon the population in which the test is applied.

3.4.2 Example 2: DNA profiling for trace-donor identification [neither model A nor B]

As was mentioned above, in the context of trace-donor identification via DNA profiling, lead variable Y is an indicator of a perfect DNA profile match and inference variable X indicates trace donor-suspect identity. Clearly, X is causal for Y, and not *vice versa*. The likelihoods $L(X=x|Y=y,Z=z)=P(Y=y|X=x,Z=z)$ can be calculated provided that the necessary parameters are known for the population of origin z of the trace donor. Since no priors for X can sensibly be specified, calculation of PO from the LR using Bayes' theorem is not feasible, which is why the interpretation of DNA results in crime cases is usually confined to the reporting by the forensic expert of the LR. Unfortunately, information on population affiliation z is also rarely available in forensic practice, and several ways have been suggested to deal with this uncertainty. The most frequently followed approach is to assume that trace donor and suspect belong to the same population [26]. Another, more pragmatic but formally problematic suggestion has been to maximize the likelihood over different populations [28].

4. Forensic DNA Phenotyping (FDP)

4.1 Example 3: Prediction of eye color as an idealized, fictional example of FDP [model B]

The FDP paradigm implies that the EVC of interest is determined by a (single-locus or multi-locus) genotype and that all other determinants of the EVC are population-independent (model B, Fig. 1B). In the following, prediction of blue eye color (Y) will be used as a generic example of FDP. For the sake of simplicity, we will assume that blue eye color is caused by a single, haploid, biallelic marker X with alleles 1 and 0 knowing that, in reality, human eye color variation is due to diploid sequence variation at several genes [9, 34]. Thus, the posterior probabilities P(Y=y|X=x) would be the same in all populations and one study in one population would suffice to estimate these probabilities by a valid method. The results would allow practical use of the genetic marker in any other population, provided that each possible level of Y (i.e. blue and non-blue eye color) was sufficiently frequent in the original population to warrant accurate parameter estimation.

In contrast to the PO, the LR would be population-dependent even in the present simplistic example because it is tied to the prior odds by Bayes' theorem. This matter of fact shall be illustrated by two fictional populations with different frequencies of allele 1 of 0.1 and 0.4 (Table 1). The PO equal 9.00 in both populations despite the different allele frequencies and the different prevalence of blue eye color (0.54 in population 1, 0.66 in population 2). Due to the prevalence difference, however, the LR is considerably higher in population 1 than in population 2.

4.2 General Forensic DNA Phenotyping [possibly model B]

The goal of FDP is to predict an EVC (Y) from a genetic profile (X). If X is causal of Y except for some population-independent factors, then model B applies and P(Y=y|X=x,Z=z)=P(Y=y|X=x). The same holds true, of course, when the genetic profile is the exclusive cause of the EVC level of interest. As illustrated by fictional example 3 (section 4.1), the posterior probabilities can be estimated based on the premise of model B from any sufficiently large sample, irrespective of whether this sample is population-representative for the EVC or the genetic profile, or both. Moreover, no priors are required. The PO are independent of the population whereas the LR is not.

For a dichotomous trait such as presence of a specific eye color level (e.g. blue or brown), FDP can be viewed as a diagnostic test where the presence of a certain genotype is equivalent to a 'positive' test result. In diagnostics terminology, likelihoods L(Y=1|X=1)=P(X=1|Y=1) and L(Y=0|X=0)=P(X=0|Y=0) correspond to sensitivity and specificity, respectively, of the FDP test whilst the posterior probabilities P(Y=1|X=1) and P(Y=0|X=0) are the positive and negative predictive value, respectively. However, there is a crucial difference between FDP and medical diagnostics as exemplified by the HIV example of section 4.1: Whereas sensitivity and specificity were population-independent and the predictive values were dependent upon the population in which the test was to be applied (*via* the prevalence), it is just the other way round with FDP. Here, as long as model B applies, the sought-after predictive values are population-independent whereas sensitivity and specificity are not.

There are two main scenarios in which model B would be violated in the case of FDP (Fig. 3).

- The first scenario is concerned with the amount of causal genetic information captured by the genetic markers used for EVC prediction. Even if the EVC is entirely genetic, i.e. without any contribution by the environment at all, its genetic basis will rarely be fully covered by the DNA markers used for FDP. Then, if the unknown part (Z') of the genetic basis is distributed differently in different populations, $P(Y=y|X=x,Z'=z',Z=z)=P(Y=y|X=x)$ may hold for all populations z, but still $P(Y=y|X=x,Z=z) \neq P(Y=y|X=x)$ for at least one level triplet (x,y,z) (Fig. 2B, Fig.3).

- The other scenario relates to the possibility of a non-genetic (i.e. environmental) component Z'' that determines the EVC of interest in addition to its genetic component (Fig. 3) but that is not incorporated in the EVC prediction model. Again, if such a non-genetic component exists, it must be distinguished whether Z'' is distributed differently in different populations or not (Fig. 2A and 2B). For example, there may be an influence of sex on the EVC but because the sex ratio is virtually identical all over the world, this would not affect the validity of 'exporting' FDP from the original to any other populations. However, if Z'' has a different distribution in different populations (e.g. temperature), this would render model B invalid and the predictive model must be derived separately in each population.

Adding the unknown genetic component (scenario 1) or environmental component (scenario 2) to the EVC prediction model would result in model B being correct again. In both scenarios, the possible degree of violation of model B is critically dependent upon both the amount of causal information about the EVC that is captured by the genetic markers used for FDP and the degree of variation of the remaining (genetic or non-genetic) information between populations. The ensuing degree of population-dependence of the PO therefore needs to be evaluated separately for each EVC and each genetic marker system used for EVC prediction. If model B does not apply for the EVC in question, one possibility would be to report a table of posterior probabilities calculated for different possible populations of origin of the unknown sample donor.

4.3 Population dependence of eye color prediction using the IrisPlex System

Of all EVCs, eye color is currently the best to predict by way of FDP. The IrisPlex DNA test system, which includes six SNPs in six different genes, was found to achieve high levels of prediction accuracy for blue and brown eye color in eight European populations (Table 2). The respective performance measures were obtained with the initial IrisPlex model [9] resulting from multinomial logistic regression analysis.

Note that, if model B were to apply, the positive (PPVs) and negative predictive values (NPVs) should not vary between populations. For blue eye color, the predictive capability of IrisPlex is generally good and there are only moderate population differences in terms of both the predictive values and sensitivity and specificity (Table 2). The rather low specificity of 0.59 in Estonia may be regarded as an outlier. For brown eye color, the predictive capability of IrisPlex is more varied. The PPV, in particular, ranges from 0.65 in Norway to 0.96 in Greece whilst the NPV was found to lie between 0.87 (France) and 0.98 (Spain). Notably, the predictive capability was no better in the Dutch than in the other population samples despite the fact that the model was developed in a Dutch sample (not overlapping the sample used here).

In summary, model B does not fully apply for the IrisPlex DNA system but the violations seem to be limited and the ensuing posterior probabilities (i.e. PPV and NPV) vary only moderately between populations.

5. Discussion

5.1 Population-dependency of Forensic DNA Phenotyping (FDP)

We argued that the parameters of a DNA-based predictive model of an externally visible characteristic (EVC) are likely population-dependent unless all causes of the EVC other than the genetic markers included in the model are equally distributed in different populations. In the case of the IrisPlex DNA system to predict blue and brown eye color, we noticed a modest level of variability between the eight European populations tested, but this variation was not so strong as to impede the predictive capability of the system. Nevertheless, since the predictive values were found to vary by up to 30% for brown eye color, empirical adaption of the respective model to individual populations should represent a worthwhile refinement for this particular EVC. With blue eye color, in contrast, the IrisPlex system seems to be well suited for EVC prediction as is. One explanation for the exceptional predictability of blue eye color is its strong heritability and the minor influence of environmental cofactors. In fact, a causal genetic variant is known for this EVC, namely rs12913832 in the *HERC2* gene, that has high predictive accuracy on its own [8, 9, 17, 35-37]. This variant, which is also included in the IrisPlex and the HIrisPlex system, is located in a long-distance enhancer regulating the expression of the nearby *OCA2* pigmentation gene, thereby rendering the variant itself a molecular switch between blue and brown/intermediate eye color [36]. Another possible reason for the difference in predictive performance between blue and brown eye color may be the higher outward complexity of the latter phenotype. If brown eye color is more difficult to determine unambiguously in the field than blue eye color, and if this difficulty varies between populations, then the ensuing uncertainty about an individual phenotype may well translate into a population difference of the corresponding predictive values.

The other five SNPs included in the IrisPlex system are a mix of (likely) causal and non-causal albeit statistically associated genetic variants of varying effect size. Depending upon the strength of the individual genotype-phenotype association, some of the alleles predisposing to a certain eye color are also present in populations where this eye color is absent (see Figure 4 in [17]). Thus, alleles associated with non-brown eye color in Europeans were also found in populations where non-brown eye color is rarely if ever observed, a phenomenon also documented at the world-wide level using the CEPH-HGDP samples analyzed by Walsh et al. [17]. That very same study also revealed, however, that non-brown eye color is almost never predicted by IrisPlex in populations lacking non-brown eye color, as was attested for East Asians, Oceanians, Sub-Saharan Africans and Native Americans [17]. On the other hand, both brown and non-brown eye color was predicted in Europeans and in neighboring populations where the latter EVC level is indeed prevalent (see Figure 5 in [17]). These results led the authors to tentatively conclude that IrisPlex-based eye color prediction is not strongly population-dependent [16, 17], a notion essentially concurring with our own data from eight European populations as presented here (Table 2).

At present, the second best predictable EVC after eye color is hair color. Thus, the HIrisPlex DNA test system [11] was devised and forensically validated [18] to predict blond, brown, red and black hair color on the basis of 22 causal and non-causal DNA variants of up to medium effect. This includes four of the six IrisPlex SNPs, particularly the previously mentioned SNP rs12913832 in *HERC2*, an eye color predictor that is also associated with light and dark hair color. In addition, the *MC1R* gene was found to contain several DNA variants that are causal for red hair color, and 11 of these were included in HIrisPlex. Similar to eye color, some alleles associated with non-black hair color in Europeans were found to occur in non-European populations, where this phenotype level is absent [11]. On the other hand, the combination of all 22 HIrisPlex SNPs again predicted nothing but black hair in East Asians, Oceanians, Sub-Saharan Africans and Native Americans, where black is indeed the only hair color. Both non-black and black hair was predicted in European and surrounding populations [11]. This notwithstanding, the conclusion that HIrisPlex-based hair color prediction was mostly population-independent [11] requires further validation in independent samples.

In contrast to eye color, hair color is determined to some degree by non-genetic component(s). Some people are blond as children but grow dark blond or brown as adults. This age-dependent hair color change is not addressed by HIrisPlex or any other currently known molecular markers so that it would escape predictability for now. Moreover, the loss of hair color with advanced adult age in some (but not all) individuals is not currently predictable from DNA sequence data. This is because the genetic basis of hair greying or whitening is still mostly unclear despite the recent finding of a gene with significant but small effect [38].

Whether EVCs other than eye and hair color can also be predicted from DNA variation with sufficient accuracy and reliability still needs to be determined. With the possible exception of skin color, the genetic basis of other candidate EVCs has not yet been characterized well enough to yield practically useful levels of predictive accuracy and further research to find more and more strongly associated genes is required [7].

5.2 Forensic DNA profiling versus Forensic DNA Phenotyping

The mathematical models underlying DNA profiling and FDP are fundamentally different in terms of the causal relationship between the available genetic data and the target of inference making. In the case of DNA profiling, the inference variable is the identity, or not, of trace donor and suspect, which in turn determines the nature of their DNA profiles. With FDP, in contrast, the inference variable is the EVC, which is (causally) influenced by the genotype(s) used to infer the EVC. As a consequence, only LR should be calculated for DNA profiles whereas PO should be calculated in FDP. The two approaches also differ regarding the relevance of population affiliation. The fact that match probabilities and, hence, the LR depend upon the reference population considered has always been recognized by the forensic community. With FDP, in contrast, the sought-after PO may be widely independent of population affiliation, although the actual degree of independence is likely to vary between EVCs and needs to be investigated empirically. At least for eye and (possibly) hair color, the PO do not appear to be strongly dependent upon the population in which FDP is carried out.

5.3 Use of either LR or PO in forensic genetics

For genetic matches in DNA profiling, it has become standard to report LR rather than PO. Since this has been advocated by scientists for decades, a certain kind of mistrust to use PO in the first place has arisen in the forensic genetics community. Part of this view on PO may be due to the fact that forensic geneticists often think of the LR as the only quantity that can be calculated without making assumptions outside their field of expertise. We hope that we have been able to clarify that, whilst LR is the outcome of choice in the case of DNA profiling, there are other applications of DNA typing in forensics, such as FDP, where the sought-after PO can be calculated and therefore should be reported. Use of different statistical measures for different practical purposes is not in conflict with theory. Instead, it reflects that there may be immanent differences between different forensic uses of DNA information.

The same argument, of course, also applies to other, non-genetic information. Thus, it may well be sensible to complement DNA-based inference measures by LR or PO values derived from other sources of information, such as eye witnesses or CCTV. Whether and how these numerical figures can be combined with the genetics-based results depends upon the complexity of the underlying mathematical relationships. In any case, when communicating statistical results, the forensic expert should aim at clarifying this issue to the investigating authorities.

6. Acknowledgments

This study was supported in part by Erasmus MC University Medical Center Rotterdam (to MKa), Indiana-University-Purdue-University Indianapolis (to SW), and The National Institute of Justice, USA (to SW). FL is supported by the Chinese "1000-Talent Plan" for distinguished young scholars.

7. References

[1] Evett IW. Towards a uniform framework for reporting opinions in forensic science casework. Sci Justice. 1998;38:198-202.
[2] Caliebe A, Krawczak M. Probability and likelihood. In: Amorim A, Budowle B, editors. Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation. London: Imperial College Press; 2016.
[3] Biedermann A, Bozza S, Taroni F. Prediction in forensic science: a critical examination of common understandings. Front Psychol. 2015;6.
[4] Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2011;12:179-92.
[5] Kayser M, Schneider PM. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet. 2009;3:154-61.
[6] Murphy E. Legal and ethical issues in forensic DNA phenotyping. NYU School of Law, Public Law Research Paper No 13-46 Available at SSRN: http://ssrncom/abstract=2288204 or http://dxdoiorg/102139/ssrn2288204. 2013.
[7] Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015;18:33-48.
[8] Caliebe A, Harder M, Schuett R, Krawczak M, Nebel A, von Wurmb-Schwark N. The more the merrier? How a few SNPs predict pigmentation phenotypes in the Northern German population. Eur J Hum Genet. 2016;24:739-47.
[9] Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, et al. Eye color and the prediction of complex phenotypes from genotypes. Curr Biol. 2009;19:R192-3.
[10] Branicki W, Liu F, van Duijn K, Draus-Barini J, Pospiech E, Walsh S, et al. Model-based prediction of human hair color using DNA variants. Hum Genet. 2011;129:443-54.
[11] Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, et al. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet. 2013;7:98-115.
[12] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, et al. DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet. 2012;6:330-40.
[13] Kastelic V, Drobnic K. A single-nucleotide polymorphism (SNP) multiplex system: the association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model. Croat Med J. 2012;53:401-8.
[14] Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, et al. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Sci Int Genet. 2011;5:472-8.
[15] Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, et al. Further development of forensic eye color predictive tests. Forensic Sci Int Genet. 2013;7:28-40.
[16] Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, et al. Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. 2011;5:464-71.
[17] Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet. 2011;5:170-80.
[18] Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, et al. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet. 2014;9:150-61.
[19] Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, et al. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014;13:34-44.

[20] Liu F, Hamer MA, Heilmann S, Herold C, Moebus S, Hofman A, et al. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet. 2016;24:895-902.
[21] Marcińska M, Pośpiech E, Abidi S, Andersen JD, van den Berge M, Carracedo Á, et al. Evaluation of DNA Variants Associated with Androgenetic Alopecia and Their Potential to Predict Male Pattern Baldness. PLoS One. 2015;10:e0127852.
[22] Pośpiech E, Karłowska-Pik J, Marcińska M, Abidi S, Andersen JD, Berge Mvd, et al. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. Forensic Sci Int Genet. 2015;19:280-8.
[23] Liu F, Hendriks AEJ, Ralf A, Boot AM, Benyi E, Sävendahl L, et al. Common DNA variants predict tall stature in Europeans. Hum Genet. 2014;133:587-97.
[24] Balding DJ. Comment on: Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. Law, Probability and Risk. 2004;3:63-4.
[25] Balding DJ. Weight-of-Evidence for Forensic DNA Profiles: Wiley; 2005.
[26] Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int. 1994;64:125-40.
[27] Steele CD, Balding DJ. Choice of population database for forensic DNA profile analysis. Sci Justice. 2014;54:487-93.
[28] National Research Council. The Evaluation of Forensic DNA Evidence. Washington DC: National Academy Press; 1996.
[29] Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS. Population-specific FST values for forensic STR markers: A worldwide survey. Forensic Sci Int Genet. 2016;23:91-100.
[30] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
[31] Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940-1.
[32] Kuhn M. caret: Classification and Regression Training. R package version 6.0-64, https://CRAN.R-project.org/package=caret. 2016.
[33] Pearl J. An introduction to causal inference. Int J Biostat. 2010;6:Article 7.
[34] Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007;39:1443-52.
[35] Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. Hum Genet. 2008;123:177-87.
[36] Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 2012;22:446-55.
[37] Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am J Hum Genet. 2008;82:424-31.
[38] Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J-C, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun. 2016;7:10815.

8. Figure Legends

Figure 1: Schematic representation of mathematical models A and B of the relationship between three forensically relevant variables (for details, see text). X: causal variable, Y: consequential variable, Z: cofactor, red arrow: causal relationship, orange arrow: statistical association, blue arrow: direction of statistical inference, green arrow: modification of the statistical association between X and Y.

Figure 2: Schematic representation of the influence of an additional cofactor Z' on the conditional distribution of Y given X (for details, see legend to Figure 1)

Figure 3: Scenario of Forensic DNA Phenotyping (FDP) with cofactors (for details, see legend to Figure 1).

Figure 1

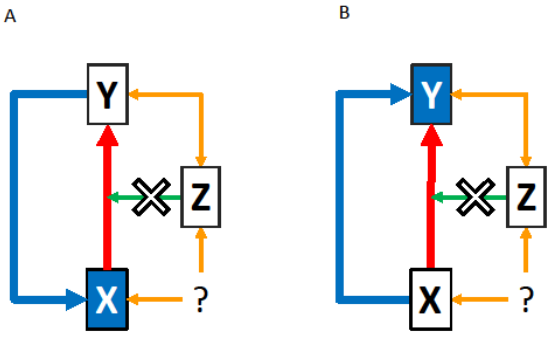


Figure 2

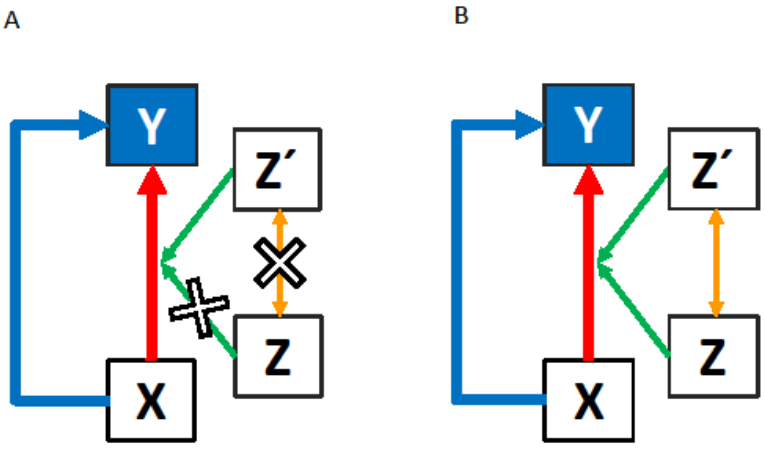
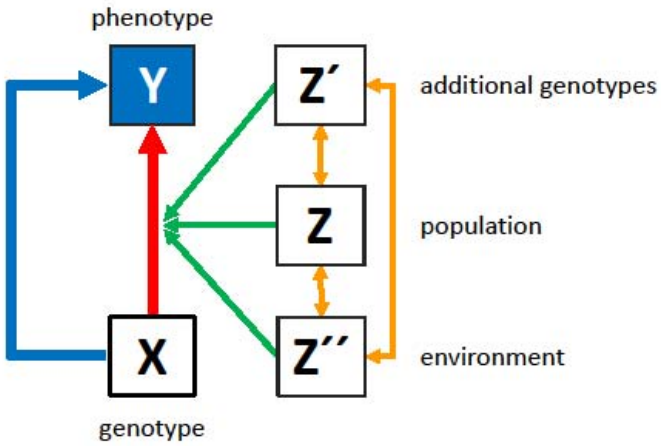


Figure 3



9. Tables

Table 1: Fictional example of FDP for blue eye color in two different populations

X (allele)	Y (eye color)					
	frequency (population 1)			frequency (population 2)		
	1 (blue)	0 (non-blue)	total	1 (blue)	0 (non-blue)	total
1	0.09	0.01	0.10	0.36	0.04	0.40
0	0.45	0.45	0.90	0.30	0.30	0.60
total	0.54	0.46	1.00	0.66	0.34	1.00
P(Y=1|X=1)	0.900			0.900		
P(Y=0|X=1)	0.100			0.100		
P(X=1|Y=1)	0.167			0.545		
P(X=1|Y=0)	0.022			0.118		
PO	9.00			9.00		
prior odds	1.17			1.94		
LR	7.59			4.62		

PO: posterior odds P(Y=1|X=1)/P(Y=0|X=1)

LR: likelihood ratio L(Y=1|X=1)/L(Y=0|X=1) = P(X=1|Y=1)/P(X=1|Y=0)

Table 2: Capability of the IrisPlex system to predict blue and brown eye color in eight European populations

sample	eye color									
	blue					brown				
	AUC	Se	Sp	PPV	NPV	AUC	Se	Sp	PPV	NPV
Norway (n=547)	0.93	0.95	0.81	0.94	0.83	0.93	0.88	0.90	0.65	0.97
Estonia (n=579)	0.89	0.95	0.59	0.90	0.78	0.89	0.71	0.94	0.69	0.95
UK (n=498)	0.94	0.94	0.82	0.93	0.84	0.94	0.85	0.89	0.67	0.96
France (n=616)	0.96	0.90	0.93	0.84	0.96	0.92	0.94	0.77	0.88	0.87
Italy (n=542)	0.95	0.86	0.95	0.87	0.94	0.92	0.98	0.68	0.82	0.95
Greece (n=547)	0.98	0.88	0.98	0.87	0.98	0.96	0.99	0.76	0.96	0.91
Spain (n=511)	0.97	0.87	0.97	0.87	0.97	0.95	1.00	0.70	0.92	0.98
NL (n=2364)	0.91	0.94	0.77	0.90	0.85	0.92	0.87	0.88	0.68	0.96

UK: United Kingdom, NL: Netherlands, AUC: area under the receiver operating characteristic (ROC) curve, Se: sensitivity, Sp: specificity, PPV: positive predictive value, NPV: negative predictive value