entropy

MDPI

Article

A Nonparametric Bayesian Approach to the Rare Type Match Problem

Giulia Cereda * and Richard D. Gill

Mathematical Institute, Leiden University, Postbus 9512, 2300 RA Leiden, The Netherlands; gill1109@gmail.cin
* Correspondence: giulia.cereda7@gmail.com

Received: 27 January 2020; Accepted: 9 April 2020; Published: 13 April 2020

Abstract: The "rare type match problem" is the situation in which, in a criminal case, the suspect's DNA profile, matching the DNA profile of the crime stain, is not in the database of reference. Ideally, the evaluation of this observed match in the light of the two competing hypotheses (the crime stain has been left by the suspect or by another person) should be based on the calculation of the likelihood ratio and depends on the population proportions of the DNA profiles that are unknown. We propose a Bayesian nonparametric method that uses a two-parameter Poisson Dirichlet distribution as a prior over the ranked population proportions and discards the information about the names of the different DNA profiles. This model is validated using data coming from European Y-STR DNA profiles, and the calculation of the likelihood ratio becomes quite simple thanks to an Empirical Bayes approach for which we provided a motivation.

Keywords: forensic statistics; likelihood ratio; Bayesian nonparametric; rare type match problem; Y-STR

1. Introduction

The largely accepted method for evaluating how much some available data $\mathcal{D}$ (typically forensic evidence) helps discriminate between two hypotheses of interest (the prosecution hypothesis H_p and the defense hypothesis H_d) is the calculation of the *likelihood ratio* (LR), a statistic that expresses the relative plausibility of the data under these hypotheses, defined as

$$\mathrm{LR} = \frac{\Pr(\mathcal{D}|H_p)}{\Pr(\mathcal{D}|H_d)}. \tag{1}$$

Widely considered the most appropriate framework to report a measure of the 'probative value' of the evidence regarding the two hypotheses [1–4], it indicates the extent to which observed data support one hypothesis over the other. The likelihood ratio is supposed to be multiplied by the prior odds, in order to obtain the posterior odds. The latter is the quantity of interest for a judge, but the prior odds do not fall within the statistician's competence. Even if a judge does not explicitly do Bayesian updating, the likelihood ratio is still considered to be the correct way for the expert to communicate their evaluation of the weight of the evidence to the court. We refer the reader to Taroni et al. [5] for extensive arguments for the use of likelihood ratios in forensic statistics. Forensic literature presents many approaches to calculate the LR, mostly divided into Bayesian and frequentist methods (see Cereda [6,7] for a careful differentiation between these two approaches).

This paper proposes a first application of a Bayesian nonparametric method to assess the likelihood ratio in the rare type match case, the challenging situation in which there is a match between some characteristic of the recovered material and of the control material, but this characteristic has not been observed before in previously collected samples (i.e., in the database of reference). This constitutes a problem because the value of the likelihood ratio depends on the unknown proportion of the matching

characteristic in a reference population, and the uncertainty over this proportion, in standard practice for simpler situations, is dealt with using the relative frequency of the characteristic in the available database. In particular, we will focus on Y-STR data, for which the rare type match problem keeps turning up [7]. The problem is so substantial that it has been called "the fundamental problem of forensic mathematics" [8].

The use of our "Bayesian nonparametric" method involves the mathematical assumption that there are infinitely many different Y-STR profiles. Of course, we do not believe this literally to be true. We do suppose that there are so many profiles that we cannot say anything sensible about their exact number, except that it is very large. Hence, we pretend they are infinitely many, so that we can use the chosen Bayesian nonparametric method. The nomenclature is misleading since it gives the impression that there is no parameter. A nonparametric model is not a model without parameters, is a model with at least one infinite-dimensional parameter. Some might call this model "semiparametric".

The parameter of the model is the infinite-dimensional vector $\mathbf{p}$, containing the (unknown) sorted population proportions of all possible Y-STR profiles. As a prior over $\mathbf{p}$, we choose the two-parameter Poisson Dirichlet distribution, and we model the uncertainty over its hyperparameters α and θ through the use of a hyperprior. The information contained in the actual repeat numbers, a list of which form the name of each Y-STR profiles, is discarded thereby reducing the full data $\mathcal{D}$ to a smaller set D.

If compared to traditional Bayesian methods such as those discussed in Cereda [6], this method has the advantage of having a prior for the parameter $\mathbf{p}$ that is more realistic for the population we intend to model. Moreover, despite its technical theoretical background, we empirically derived an approximation that makes the method intuitive and simple to apply for practical use: indeed, simulation experiments show that a ("empirical Bayes") hybrid approach that plugs in maximum likelihood estimators for the hyperparamenter is justified, at least when using populations having features which we observe in combined European data. The last point in favor of the choice of the two-parameter Poisson Dirichlet prior over $\mathbf{p}$ is that it has the following sufficiency property: the probability of observing a new Y-STR profile only depends on the number of already observed Y-STR profiles and on the sample size, while the probability of observing a Y-STR profile that is already in the database only depends on its frequency in the database and on the sample size.

The paper is structured as follows: Section 2 discusses the state of the art regarding the rare type match problem and the evaluation of Y-STR matches. Section 3 presents our model, with the assumptions and the prior distribution chosen for the parameter $\mathbf{p}$ along with some theory on random partitions and the Chinese restaurant representation, useful to provide a prediction rule and a convenient and compact representation of the reduced data D. In addition, a lemma that facilitates computing the likelihood ratio in a very simple way is presented and proved. In Section 4, the likelihood ratio is derived. Section 5 illustrates the application of this model to a database sampled from an artificial population. We will discuss data-driven choices for the hyperparameters, and the derivation of the likelihood ratio values obtained both with and without reducing the data to partitions, in the ideal situation in which the vector $\mathbf{p}$ is known. In addition, the distribution of the likelihood ratios for different rare type match cases is studied, along with the analysis of two different errors.

2. State of the Art

Y-STR data have been our main motivation for studying the rare type match problem. Our model will reduce the relationship among various profiles to a binary "match" or "no match" equivalence relation. However, there is a big debate in the scientific community regarding whether it is acceptable to throw away the genetic structure of this kind of data. In this section we discuss the state of the art regarding the rare type match problem as a general issue and also the state of the art regarding methods for assessing evidential values of matching Y-STR profiles. Moreover, it should be realized that the rare type match problem is an interesting problem also outside the Y-STR profile setting and our model can perhaps also be applied to other kinds of data.

2.1. The Rare Type Match Problem

The evaluation of a match between the profile of a particular piece of evidence and a suspect's profile should depend on the relative frequencies of that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the suspect is in trouble. A big problem arises when the observed frequency of the profile in a sample (database) from the population of interest is 0 (there is already a problem when the number is small, since it means that we don't know the population frequency very well). This problem could have been named "the new type match problem", but we decided to use the name "rare type match problem", motivated by the fact that a profile that has zero occurrences is likely to be rare, even though it is challenging to quantify how rare it is. The rare type match problem is particularly important for new kinds of forensic evidence, such as results from DIP-STR markers (see, for instance, Cereda et al. [9]) for which the available database size is still limited. The problem also occurs when more established types of evidence, such as Y-chromosome (or mitochondrial) DNA profiles are used, as explained in Section 2.2: they have been our main motivation for the present study.

The rare type match problem has been addressed in well known non-forensic statistics domains, and many solutions have been proposed. The *empirical frequency estimator*, also called *naive estimator* that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace [10], and the *add-half* estimator of Krichevsky and Trofimov [11]), which add a constant to the count of each type, included the unseen ones. However, these methods require knowledge of the number of possible unseen types, and they perform badly when this number is probably large (an anyway unknown) compared to the sample size (see Gale and Church [12] for an additional discussion). Alternatively, Good [13], based on an intuition on A.M. Turing, proposed the *Good–Turing estimator* for the total unobserved probability mass, based on the proportion of singleton observations in the sample. An application of this estimator to the frequentist LR assessment in the rare type match case is proposed in Cereda [7].

More recently, Orlitsky et al. [14] have introduced the *high-profile estimator*, which extends the tail of the *naive estimator* to the region of unobserved types. Anevski et al. [15] improved this estimator and provided a consistency proof for their modified estimator (original authors only provided heuristic reasoning that turned out to be rather difficult to make rigorous).

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi [16] using Dirichlet processes, by Lijoi et al. [17], De Blasi et al. [18] using a general Gibbs prior, and by Favaro et al. [19] with specific focus on the two-parameter Poisson Dirichlet prior, for which Arbel et al. [20] provide large sample asymptotic and credible bands. In particular, Favaro et al. [21] show the link between the Bayesian nonparametric approach and the Good–Turing estimator. However, the LR assessment requires not only the probability of observing a new species, but also the probability of observing this same species twice (according to the defense, the crime stain profile and the suspect profile are two identical independent observations): to our knowledge, the present paper is the first to address the problem of assessment of the LR in the rare type match case using a Bayesian nonparametric model (i.e., with an infinite dimensional parameter). As a prior for $\mathbf{p}$, we will use the two-parameter Poisson Dirichlet distribution, which is proving useful in many discrete domains, in particular language modeling [22]. In addition, it predicts a power-law behavior that describes an incredible variety of phenomena [23] including the observed distribution of Y-STR haplotypes too.

2.2. Evaluation of Matching Probabilities of Y-STR Data

Y-STR profiles are typically used to detect male DNA in male–female DNA mixtures and are made of a number (usually varying from 7 to 23) of integers, that we treat as categorical observations, corresponding to STR polymorphisms belonging to the non-recombining part of the Y-chromosome.

There is no biological reason to assume independence among Y-STR loci, and even though the lack of recombination is in principle balanced by recurrent and backward mutations, the existence of such a dependency is studied and confirmed by Caliebe et al. [24]. As far as Y-STR population frequencies are concerned, the dependency between loci implies that no factorization (of the kind used for the autosomal markers) can be used to calculate these frequencies, and that the available databases are too small with respect to the large space of possible profiles (hence, a database will likely contain a high proportion of singletons). Indeed, the rare type match case is very frequent when using Y-STR data, and the use of simplistic methods such as the profile count is too conservative for practical use (it is bounded from below by the inverse of the database size) [24]. In Andersen et al. [25] and Andersen et al. [26], approximations of the joint distribution with second and third order dependencies between loci are explored. However, as admitted by the authors, there is a limitation due to the inadequacy of the sizes of available databases that makes it necessary to use simulations that in turns are oversimplification of real data.

Moreover, as highlighted already in 1994 by Balding and Nichols [27], match probabilities cannot be identified with population frequencies since a match can be due also to a certain degree of relatedness between the two donors of the stain. This is particularly true for Y-STR data, since Y-STR profiles are inherited almost identically from father to son. More recently, Andersen and Balding [28] investigate the influence of relatedness on matches and make a study concluding that 95% of matching profiles are separated by a relatively small number (50–100) of meiosis, hence the degree of relatedness is a very influential factor, according to their study. They thus propose a method to describe the distribution of the number of males with a matching Y-STR profile, extending the approach to mixtures in Andersen and Balding [29]. One limitation of this study is that it is based on extensive simulations which have to be performed anew in each new application, on assumptions about the genetic evolutionary model, and on parameters which are essentially unknown.

There are a huge number of methods developed to assess the evidential values for Y-STR data. Among those that are developed precisely for the rare type match case, there are Egeland and Salas [30], Brenner [8], Cereda [7], and Cereda [6]. These particular (just listed) methods do not take into account genetic information contained in the allelic numbers forming a Y-STR DNA profile. They do not use the fact that, due to relatedness, the observation of a particular Y-STR profile increases the probability of observing the same Y-STR profile again or Y-STR profiles that differ only for few alleles. We refer the reader to Roewer [31], Buckleton et al. [32], Willuweit et al. [33], Wilson et al. [34] for models that use population genetics for coancestry. These models are not designed to be used for the rare type match case, though the Discrete Laplace method presented in Andersen et al. [35] can be successfully applied to that purpose, as shown in Cereda [7].

After a careful study of the available methods for assessing likelihood ratios (or matching probabilities) for Y-STR matches, one can see that they are of different natures (some of them do their best to exploit the genetic structure, others don't) and based on different assumptions. In our opinion, none of them is fully satisfactory and at the same time useful for the rare type match and for general cases.

In this paper, we study what can be done if we reduce the data, taking into account only the equalities and inequalities among profiles rather than considering the specific Y-STR observed characteristics. We know part of the scientific community will not agree with our approach, preferring an approach such as the one of Andersen and Balding [28], but we believe that our method can also be useful. Indeed, we think it can be very useful for an analyst, in a particular case, to obtain results using several different methods relying on different assumptions: this is actually "sensitivity analysis". It provides further information which can be used by the court—even if it only shows that the evidential value of the match is almost unquantifiable. Moreover, even though Y-STR data have been the main motivation for this study, this model is actually applicable to different kinds of data (in principle for all forensic data that show power law behavior). When applied to data without genetic structures (such

as tire marks or glass fragments), these kinds of criticisms should fade away if, of course, one can also find empirical or theoretical support for power law behavior.

The Y-STR marker system will thus be employed here as an extreme but in practice common and important example in which the problem of assessing the evidential value of rare type match can arise. We believe that the analyst should perform several analyses using different models and different assumptions, and compare the performance of the different methods, in order to try to learn from the differences (or lack of differences) between the conclusions which would follow from each method individually.

The big issues of working with Y-STR data are the unavailability of reliable databases, which are representative of actual population. The YHRD database is in fact a collection of databases coming from police or laboratories. The individual databases are not actually random samples from well defined populations since different institutions and organizations use different selection criteria. For instance, is a prison population representative of the population outside? The sizes of the databases from different countries are not proportional to the sizes of the populations.e are well aware of this limitation.

3. The Model

3.1. Notation and Data

Throughout the paper, the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realization of the random vector $\mathbf{P}$. Probability is denoted with $\Pr(\cdot)$, while the density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subscript is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be interchangeably used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Notice that, in Formula (1), $\mathcal{D}$ was regarded as the event corresponding to the observation of the available data. However, later in the paper, $\mathcal{D}$ will be regarded as a random variable generically representing the data. The particular data at hand will correspond to the value d. In that case, the following notation will thus be preferred:

$$\mathrm{LR} = \frac{\Pr(\mathcal{D} = d|H = h_p)}{\Pr(\mathcal{D} = d|H = h_d)} \quad \text{or} \quad \frac{p(d|h_p)}{p(d|h_d)}.$$

Lastly, notice that "DNA types" are used throughout the paper as a general term to indicate Y-STR profiles.

The data used in the present study were obtained from the Y Chromosome Haplotype Reference Database (YHRD) [36,37]. Here, only seven of the markers included in the PowerPlex1Y23 system (PPY23, Promega Corporation, Madison, WI, USA) were investigated: DYS19, DYS389I, DYS389II.I, DYS390, DYS391, DYS392, DYS393. The dependence between these seven "core markers" is studied in Caliebe et al. [24] that concludes that "each of these seven markers contribute indispensable information about each other markers from the same set".

3.2. Model Assumptions

Our mathematical model is based on the two following mathematical assumptions:

Assumption 1. *There are infinitely many different DNA types in nature.*

This assumption, already used by e.g., Kimura [38] in the "infinite alleles model", allows the use of Bayesian nonparametric methods and is very useful for instance in "species sampling problems"

when the total number of possible different species in nature cannot be specified. This assumption is sensible also in case of Y-STR DNA profiles since the state space of possible different haplotypes is so large that it can be considered infinite.

Assumption 2. *The names of the different DNA types do not contain usable information.*

Actually, the specific sequence of numbers that forms a DNA profile carries information: if two profiles show few differences, this means that they are separated by few mutation drifts, hence the profiles share a relatively recent common ancestor. However, this information can be very difficult to use and it might be wiser not to try to use it in the LR assessment. This is the reason why we will treat DNA types as "colors", and only consider their partition into different categories. Stated otherwise, we do not use any topological structure on the space of the DNA types.

Notice that this assumption makes the model a priori suitable for any characteristic which has many different possible types showing power law behavior, as we will see, thus the approach described in this paper might be considered, in principle, after replacing "DNA types" with any other category.

3.3. Prior

In Bayesian statistics, parameters of interest are modeled through random variables. The (prior) distribution over a parameter should represent the prior uncertainty about its value.

The assessment of the LR for the rare type match involves two unknown parameters of interest: one is $h \in \{h_p, h_d\}$, representing the unknown true hypothesis, the other is $\mathbf{p}$, the vector of the unknown population frequencies of all DNA profiles in the population of potential perpetrators. The dichotomous random variable H is used to model parameter h, and the posterior distribution of this random variable, given the data, is the ultimate aim of the forensic inquiry. Similarly, a random variable $\mathbf{P}$ is used to model the uncertainty over $\mathbf{p}$. Because of Assumption 1, $\mathbf{p}$ is an infinite-dimensional parameter, hence the need for Bayesian nonparametric methods [39,40]. In particular because of Assumption 2, data can be reduced to partitions, as explained in Section 3.5, and it will turn out that the distribution of these partitions does not depend on the order of the p_i. Hence, we can define the parameter $\mathbf{p}$ as having values in $\nabla_\infty = \{(p_1, p_2, ...) \mid p_1 \geq p_2 \geq ..., \sum p_i = 1, p_i > 0\}$, the ordered infinite-dimensional simplex. The uncertainty about its value will be expressed by the two-parameter Poisson Dirichlet prior [41–44].

The two-parameter Poisson-Dirichlet distribution can be defined through the following stick-breaking representation [40]:

Definition 1 (two-parameter GEM distribution)**.** *Given α and θ satisfying the following conditions,*

$$0 \leq \alpha < 1, \text{ and } \theta > -\alpha. \tag{2}$$

The vector $\mathbf{W} = (W_1, W_2, ...)$ is said to be distributed according to the GEM(α, θ)*, if*

$$\forall i \quad W_i = V_i \prod_{j=1}^{i-1} (1 - V_j),$$

where V_1, V_2,... are independent random variables distributed according to

$$V_i \sim Beta(1 - \alpha, \theta + i\alpha).$$

It holds that $W_i > 0$, and $\sum_i W_i = 1$.

The GEM distribution (short for Griffin–Engen–McCloskey distribution) is well-known in the literature as the "stick-breaking prior" since it measures the random sizes in which a stick is broken iteratively.

Definition 2 (Two-parameter Poisson Dirichlet distribution)**.** *Given α and θ satisfying condition (2), and a vector $\mathbf{W} = (W_1, W_2, ...) \sim GEM(\alpha, \theta)$, the random vector $\mathbf{P} = (P_1, P_2, ...)$ obtained by ranking $\mathbf{W}$, such that $P_i \geq P_{i+1}$, is said to be* Poisson Dirichlet distributed *$PD(\alpha, \theta)$. Parameter α is called the* discount parameter*, while θ is the* concentration parameter.

For our model, we will not allow $\alpha = 0$, hence we will assume $0 < \alpha < 1$, in order to have a prior that shows a power-law behavior as the one observed in the YHRD database (see Section 5.1). We think that it could be interesting to look for about models from population genetics and evolution which seem similar in some respects and which predict power-law behavior.

The main reason that prompted us to choose the two-parameter Poisson Dirichlet distribution among the possible Bayesian nonparametric priors is given by model fitting (see Section 5.1). However, there is a very nice feature of this model. It is the only one that has the following very convenient sufficiency property [45]: the probability of observing a new species only depends on the number of already observed species and on the sample size, and the probability of observing an already seen species only depends on its frequency in the sample and on the sample size.

Lastly, we point out that, in practice, we cannot assume that we know the parameters α and θ: we will resolve this by using a hyperprior.

3.4. Bayesian Network Representation of the Model

The typical data to evaluate in case of a match is $\mathcal{D} = (E, B)$, where $E = (E_s, E_c)$, and

E_s = suspect's DNA type,
E_c = crime stain's DNA type (matching the suspect's type),
B = a reference database of size n, which is treated here as a random sample of DNA types from the population of possible perpetrators.

The hypotheses of interest for the case are:

h_p = The crime stain originated from the suspect,
h_d = The crime stain originated from someone else.

In agreement with Assumption 2, the model will ignore information about the names of the DNA types: data $\mathcal{D} = (E, B)$ will thus be reduced to D accordingly. The Bayesian network of Figure 1 encapsulates the conditional dependencies of the random variables $(H, A, \Theta, \mathbf{P}, X_1, ..., X_{n+2}, D)$, whose joint distribution is defined below in terms of a collection of conditional distributions (one for each node).

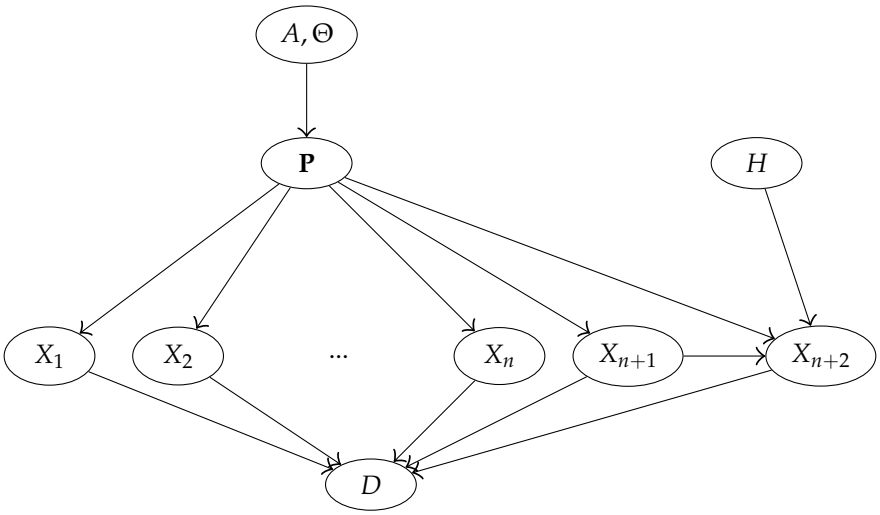


Figure 1. Bayesian network showing the conditional dependencies of the relevant random variables in our model.

H is a dichotomous random variable that represents the hypotheses of interest and can take values $h \in \{h_p, h_d\}$, according to the prosecution or the defense, respectively. A uniform prior on the hypotheses is chosen for mathematical convenience since it will not affect the likelihood ratio (the variable H being in the conditioning part):

$$\Pr(H = h) \propto 1 \quad \text{for } h \in \{h_p, h_d\}.$$

(A, Θ) is the random vector that represents the hyperparameters α and θ, satisfying condition (2). The joint prior density of these two parameters will be generically denoted as $p(\alpha, \theta)$:

$$(A, \Theta) \sim p(\alpha, \theta).$$

For obvious reasons, this will be called the 'hyperprior' throughout the text.

The random vector $\mathbf{P}$ with values in ∇_∞ represents the ranked population frequencies of Y-STR profiles. $\mathbf{P} = (p_1, p_2, ...)$ means that p_1 is the frequency of the most common DNA type in the population, p_2 is the frequency of the second most common DNA type, and so on. As a prior for $\mathbf{P}$, we use the two-parameter Poisson Dirichlet distribution:

$$\mathbf{P}|A = \alpha, \Theta = \theta \sim PD(\alpha, \theta).$$

The database is assumed to be a random sample from the population. Integer-valued random variables X_1, ..., X_n are here used to represent the (unknown) ranks in the population of the frequencies of the DNA types in the database. For instance, $X_3 = 5$ means that the third individual in the database has the fifth most common DNA type in the population. Given $\mathbf{p}$, they are an i.i.d. sample from $\mathbf{p}$:

$$X_1, X_2, ..., X_n|\mathbf{P} = \mathbf{p} \sim_{i.i.d.} \mathbf{p}. \tag{3}$$

X_{n+1} represents the rank in the population ordering of the suspect's DNA type. It is again an independent draw from $\mathbf{p}$:

$$X_{n+1}|\mathbf{P} = \mathbf{p} \sim \mathbf{p}.$$

X_{n+2} represents the rank in the population ordering, of the crime stain's DNA type. According to the prosecution, given $X_{n+1} = x_{n+1}$, this random variable is deterministic (it is equal to x_{n+1} with

probability 1). According to the defence, it is another sample from $\mathbf{p}$, independent of the previous ones:

$$X_{n+2}|\mathbf{P}=\mathbf{p}, X_{n+1}=x_{n+1}, H=h \sim \begin{cases} \delta_{x_{n+1}} & \text{if } h=h_p \\ \mathbf{p} & \text{if } h=h_d \end{cases}. \tag{4}$$

In order to observe X_1, ..., X_{n+2}, one would need, by definition, to know the rank, in terms of population proportions, of the frequency of each DNA type in the database. This is not known, hence $X_1, ..., X_n$ are not observed.

Section 3.5 recalls some notions about random partitions, useful before defining node D, the "reduced" data that we want to evaluate through the likelihood ratio.

3.5. Random Partitions and Database Partitions

A *partition of a set* S is an unordered collection of nonempty and disjoint subsets of S, the union of which forms S. Particularly interesting for our model are partitions of the set $S=[n]=\{1,...,n\}$, denoted as $\pi_{[n]}$. The set of all partitions of $[n]$ will be denoted as $\mathcal{P}_{[n]}$. Random partitions of $[n]$ will be denoted as $\Pi_{[n]}$, $n \in \mathbb{N}$. In addition, a *partition of* n is a finite nonincreasing sequence of positive integers that sum up to n. Partitions of n will be denoted as π_n, while random partitions as Π_n.

Given a sequence of integer valued random variables $X_1, ..., X_n$, let $\Pi_{[n]}(X_1, X_2, ..., X_n)$ be the random partition defined by the equivalence classes of their indices using the random equivalence relation $i \sim j$ if and only if $X_i = X_j$. This construction allows one to build a "reduction map" from the set of values of $X_1, ..., X_n$ to the set of the partitions of $[n]$ as in the following example ($n = 10$):

$$\mathbb{N}^{10} \to \mathcal{P}_{[10]} \tag{5}$$

$$X_1, ..., X_{10} \longmapsto \Pi_{[10]}(X_1, X_2, ..., X_{10}) \tag{6}$$

$$(2,4,2,4,3,3,10,13,5,4) \longmapsto \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\}\} \tag{7}$$

Similarly, and in agreement with Assumption 2, in our model, we can consider the reduction of data which ignores information about the names of the DNA types: this is achieved, for instance, by retaining from the database only the equivalence classes of the indices of the individuals, according to the equivalence relation "has the same DNA type". Stated otherwise, the database is reduced to the partition $\pi^{\text{Db}}_{[n]}$, obtained using these equivalence classes. However, the database only supplies part of the data. There are also two new DNA profiles that are equal to one another (and different from the already observed ones in the rare type match case). Considering the suspect's profile, we obtain the partition $\pi^{\text{Db+}}_{[n+1]}$, where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n+1$ constitutes a class by itself. Considering the crime stain profile too, we obtain the partition $\pi^{\text{Db++}}_{[n+2]}$ where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n+1$ and $n+2$ belong to the same (new) class. Random variables $\Pi^{\text{Db}}_{[n]}$, $\Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are used to model $\pi^{\text{Db}}_{[n]}$, $\pi^{\text{Db+}}_{[n+1]}$, and $\pi^{\text{Db++}}_{[n+2]}$, respectively.

Since prosecution and defense agree on the distribution of $X_1, ..., X_{n+1}$, but not on the distribution of $X_{n+2}|X_1, ..., X_{n+1}$, they also agree on the distribution of $\Pi^{\text{Db+}}_{[n+1]}$ but disagree on the distribution of $\Pi^{\text{Db++}}_{[n+2]}$ (see (4)).

The crucial points of the model are the following:

1. The random partitions defined through the random variables X_1, ..., X_{n+2} and through the database are the same:

$$\begin{aligned} \Pi^{\text{Db}}_{[n]} &= \Pi_{[n]}(X_1, ..., X_n), \\ \Pi^{\text{Db+}}_{[n+1]} &= \Pi_{[n+1]}(X_1, ..., X_{n+1}), \\ \Pi^{\text{Db++}}_{[n+2]} &= \Pi_{[n+2]}(X_1, ..., X_{n+2}). \end{aligned}$$

2. Although $X_1, ..., X_{n+2}$ were not observable, the random partitions $\Pi^{\text{Db}}_{[n]}, \Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are observable.

To clarify, consider the following example of a database with $k = 6$ different DNA types, from $n = 10$ individuals:

$$B = (b_1, b_2, b_1, b_2, b_3, b_3, b_4, b_5, b_6, b_2),$$

where b_i is the name of the ith DNA type according to the order chosen for the database. This database can be reduced to the partition of $[10]$:

$$\pi^{\text{Db}}_{[10]} = \{\{1,3\}, \{2,4,10\}, \{5,6\}, \{7\}, \{8\}, \{9\}\}.$$

Then, the part of the reduced data whose distribution is agreed on by prosecution and defense is

$$\pi^{\text{Db+}}_{[11]} = \{\{1,3\}, \{2,4,10\}, \{5,6\}, \{7\}, \{8\}, \{9\}, \{11\}\},$$

while the entire reduced data D can be represented as

$$\pi^{\text{Db++}}_{[12]} = \{\{1,3\}, \{2,4,10\}, \{5,6\}, \{7\}, \{8\}, \{9\}, \{11,12\}\}.$$

Now, suppose that we knew the rank in the population of each of the DNA types in the database: we knew that b_1 is, for instance, the second most frequent type, b_2 is the fourth most frequent type, and so on. Stated otherwise, we are now assuming that we observe the variables $X_1, ..., X_{n+2}$: for instance, $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, $X_{12} = 9$ (as in (5)). It is easy to check that $\Pi_{[10]}(X_1, ..., X_{10}) = \pi^{\text{Db}}_{[10]}$, $\Pi_{[11]}(X_1, ..., X_{11}) = \pi^{\text{Db+}}_{[11]}$, and $\Pi_{[12]}(X_1, ..., X_{12}) = \pi^{\text{Db++}}_{[12]}$.

Coming back to our model, data are defined as $D = \pi^{\text{Db++}}_{[n+2]}$, obtained partitioning the database enlarged with the two new observations (or partitioning $X_1, ..., X_{n+2}$). Node D of Figure 1 is defined accordingly.

Notice that, given $X_1, ..., X_{n+2}$, D is deterministic. An important result is that, according to Proposition 4 in Pitman [46], it is possible to derive directly the distribution of $D \mid \alpha, \theta, H$. In particular, it holds that, if

$$\mathbf{P} \mid \alpha, \theta \sim PD(\alpha, \theta),$$

and

$$X_1, X_2, ... \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p},$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\mathbb{P}^{\alpha,\theta}_n(\pi_{[n]}) := \Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{8}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to their least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a,b} := \begin{cases} \prod_{i=0}^{a-1}(x + ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 1 & \text{if } a = 0 \end{cases}$. This formula, also known as the *Pitman sampling formula*, is further studied in Pitman [47] and shows that $\mathbb{P}^{\alpha,\theta}_n(\pi_{[n]})$ does not depend on $X_1, ..., X_n$, but only on the sizes and the number of classes in the partitions. It follows that we can get rid of the intermediate layer of nodes $X_1, ..., X_{n+2}$. Moreover, it holds that $\Pr(D|\alpha, \theta, h_p) = \mathbb{P}^{\alpha,\theta}_{n+1}(\pi^{\text{Db+}}_{[n+1]})$, while $\Pr(D|\alpha, \theta, h_d) = \mathbb{P}^{\alpha,\theta}_{n+2}(\pi^{\text{Db++}}_{[n+2]})$.

The model of Figure 1 can thus be simplified to the one in Figure 2.

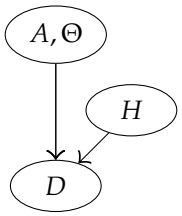


Figure 2. Simplified version of the Bayesian network in Figure 1.

3.6. Chinese Restaurant Representation

There is an alternative characterization of this model, called the "Chinese restaurant process", due to Aldous [48] for the one-parameter case, and studied in detail for the two-parameter version in Pitman [44]. It is defined as follows: consider a restaurant with infinitely many tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer-valued random variables that represent the seating plan: tables are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The process is described by the following transition matrix:

$$Y_1 = 1,$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } i = k + 1 \\ \\ \dfrac{n_i - \alpha}{n + \theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{9}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that at that time have been seated at table i. The process depends on two parameters α and θ with the same conditions (2). From (9), one can easily see the sufficientness property mentioned in Section 3.3.

$Y_1, ..., Y_n$ are not i.i.d., nor exchangeable, but it holds that $\Pi_{[n]}(Y_1, ..., Y_n)$ is distributed as $\Pi_{[n]}(X_1, ..., X_n)$, with $X_1, ..., X_n$ defined as in (3), and they are both distributed according to the Pitman sampling formula (8) [44].

Stated otherwise, we can obtain the same partition $\pi^{\text{Db}}_{[n]}$ through the seating plan of n customers or partitioning the $X_1, ..., X_n$ of the database. Similarly, $\pi^{\text{Db+}}_{[n+1]}$ is obtained when a new customer has chosen an unoccupied table (remember we are in the rare type match case), and $\pi^{\text{Db++}}_{[n+2]}$ is obtained when the $(n+2)$nd customer goes to the table already chosen by the $(n+1)$st customer (suspect and crime stain have the same DNA type). In particular, thanks to (9), we can write:

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_p, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = 1, \tag{10}$$

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_d, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = \frac{1 - \alpha}{n + 1 + \theta}, \tag{11}$$

since the $(n+2)$nd customer goes to the same table as the $(n+1)$st (who was sitting alone).

3.7. A Useful Lemma

The following lemma can be applied to four general random variables Z, X, Y, and H whose conditional dependencies are described by the Bayesian network of Figure 3. The importance of this result is due to the possibility of using it for assessing the likelihood ratio in a very common forensic situation: the prosecution and the defense disagree on the distribution of the entirety of data (Y) but agree on the distribution of a part it (X), and these distributions depend on parameters (Z).

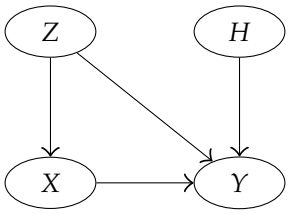


Figure 3. Conditional dependencies of the random variables of Lemma 1.

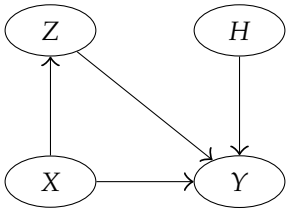
Lemma 1. *Given four random variables Z, H, X, and Y, whose conditional dependencies are represented by the Bayesian network of Figure 3, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\mathrm{lik}(h \mid x, y) \propto \mathbb{E}(p(y \mid x, Z, h) \mid X = x).$$

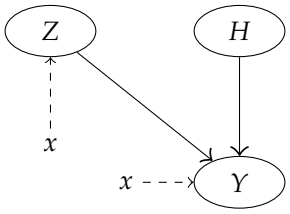
The Bayesian representation of the model, in Figure 3, allow for factoring the joint probability density of Z, H, X, and Y as

$$p(z, h, x, y) = p(z)\, p(x \mid z)\, p(h)\, p(y \mid x, z, h).$$

By Bayes' formula, $p(z)\, p(x \mid z) = p(x)\, p(z \mid x)$. This rewriting corresponds to reversing the direction of the arrow between Z and X:



The random variable X is now a root node. This means that, when we probabilistically condition on $X = x$, the graphical model changes in a simple way: we can delete the node X, and just insert the value x as a parameter in the conditional probability tables of the variables Z and Y which formerly had an arrow from node X. The next graph represents this model:



This tells us that, conditional on $X = x$, the joint density of Z, Y, and H is equal to

$$p(z, h, y \mid x) = p(z \mid x) p(h) p(y \mid x, z, h).$$

The joint density of H and Y given X is obtained by integrating out the variable Z. It can be expressed as a conditional expectation value since $p(z \mid x)$ is the density of Z given $X = x$. We find:

$$p(h, y \mid x) = p(h) \mathbb{E}(p(y \mid x, Z, h) \mid X = x).$$

Recall that this is the joint density of two of our variables, H and Y, after conditioning on the value $X = x$. Let us now also condition on $Y = y$. It follows that the density of H given $X = x$ and $Y = y$ is proportional (as function of H, for fixed x and y) to the same expression, $p(h) \mathbb{E}(p(y \mid x, Z, h) \mid X = x)$.

This is a product of the prior for h with some function of x and y. Since posterior odds equals prior odds times likelihood ratio, it follows that the likelihood function for h, given $X = x$ and $Y = y$ satisfies

$$\text{lik}(h \mid x, y) \propto \mathbb{E}(p(y \mid x, Z, h) \mid X = x).$$

Corollary 1. *Given four random variables Z, H, X, and Y, whose conditional dependencies are represented by the network of Figure 3, the likelihood ratio for $H = h_1$ against $H = h_2$ given $X = x$ and $Y = y$ satisfies*

$$\text{LR} = \frac{\mathbb{E}(p(y|x, Z, h_1)|X = x)}{\mathbb{E}(p(y|x, Z, h_2)|X = x)}. \tag{12}$$

4. The Likelihood Ratio

From now on, we will omit the superscripts Db, Db+, and Db++ for ease of notation.

Using the hypotheses and the reduction of data D defined in Section 3, the likelihood ratio will be defined as

$$\text{LR} = \frac{p(\pi_{[n+2]}|h_p)}{p(\pi_{[n+2]}|h_d)} = \frac{p(\pi_{[n+1]}, \pi_{[n+2]}|h_p)}{p(\pi_{[n+1]}, \pi_{[n+2]}|h_d)}.$$

The last equality holds due to the fact that $\Pi_{[n+1]}$ is a deterministic function of $\Pi_{[n+2]}$.

Corollary 1 can be applied to our model since defense and prosecution agree on the distribution of $\pi_{[n+1]}$, but not on the distribution of $\pi_{[n+2]}$, and data depends on parameters α and θ. Thus, if (A, Θ) play the role of Z, $X = \Pi_{[n+1]}$, and $Y = \Pi_{[n+2]}$, by using (10) and (11), we obtain:

$$\begin{aligned} \text{LR} &= \frac{\mathbb{E}(p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, h_p) \mid \Pi_{[n+1]} = \pi_{[n+1]})}{\mathbb{E}(p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, h_d) \mid \Pi_{[n+1]} = \pi_{[n+1]})} \\ &= \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}. \end{aligned}$$

The expected value is taken with respect to the posterior distribution of $A, \Theta \mid \Pi_{[n+1]} = \pi_{[n+1]}$. The solution we propose in this paper is to deal with the uncertainty about α and θ by using MLE estimators and plug those estimators into the formula. Notice that this is equivalent to a hybrid approach, in which the parameters are estimated in a frequentist way and their values are plugged into the Bayesian LR. In the future, we plan to use MCMC methods to calculate as exactly as possible the exact posterior distribution, given assumed priors on the hyperparameters.

By defining the random variable $\Phi = n\dfrac{1-A}{n+1+\Theta}$, we can write the LR as

$$\text{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

5. Analysis on the YHRD Database

In this section, we present the study we made on a database of 18,925 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe [37]. Our analyses are performed by considering only 7 Y-STR loci (DYS19, DYS389 I, DYS389 II, DYS3904, DYS3915, DY3926, DY3937), but similar results have been observed with the use of 10 loci.

5.1. Model Fitting

In Figure 4, the ranked frequencies of the 18'925 Y-STR profiles of the YHRD database are compared to the relative frequencies of samples of size n obtained from several realizations of

PD($\alpha_{MLE}, \theta_{MLE}$), where α_{MLE} and θ_{MLE} are the maximum likelihood estimates obtained using the entire database and the likelihood defined by (8).Their values are $\alpha_{\text{MLE}} = 0.51$ and $\theta_{\text{MLE}} = 216$.

To do so, we run several times the Chinese Restaurant seating plan (up to $n = 18,925$ customers): each run is used to approximate a new realization $\mathbf{p}$ from the PD($\alpha_{MLE}, \theta_{MLE}$). As explained in Section 3.6, the partition of the customers into tables is the same as the partition obtained from an i.i.d. sample of size n from $\mathbf{p}$. We can see that, for the most common haplotypes (left part of the plot), there is some discrepancy. However, we are interested in rare haplotypes, which typically have a frequency belonging to the right part of the plot. In that region, the two-parameter Poisson Dirichlet follows the distribution of the data quite well. The dotted line in Figure 4 shows the asymptotic behavior on the two-parameter Poisson Dirichlet distribution. Indeed, if $\mathbf{P} \sim \text{PD}(\alpha, \theta)$, then

$$\frac{P_i}{Zi^{-1/\alpha}} \to 1, \quad \text{a.s., when } i \to +\infty$$

for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$. This power-law behavior describes an incredible variety of phenomena [23].

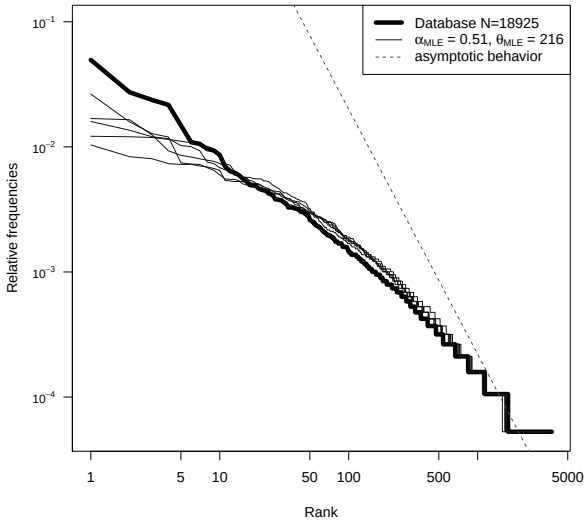


Figure 4. Log scale ranked frequencies from the database (thick line) are compared to the relative frequencies of samples of size $n = 18,925$ obtained from several realizations of PD($\alpha_{MLE}, \theta_{MLE}$) (thin lines). Asymptotic power-law behavior is also displayed (dotted line).

The thick line in Figure 4 also seems to have a power-law behavior, and, to be honest, we were hoping to get the same asymptotic slope of the prior. This is not what we observe, but, in Figure 5, it can be seen that, for such a big value of θ, we would need a bigger database (at least $n = 10^6$) to see the correct slope.

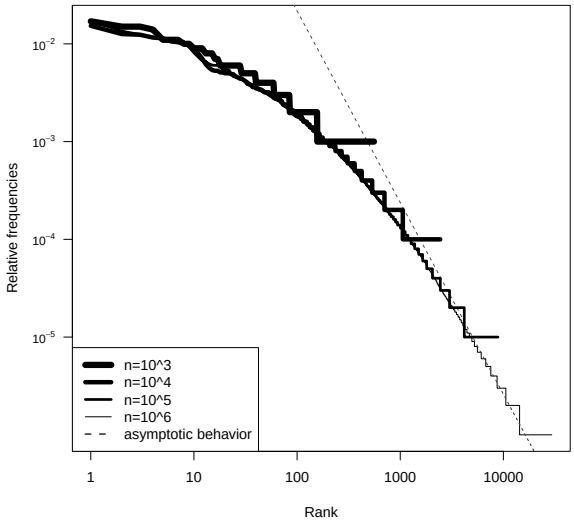


Figure 5. Log scale ranked frequencies from the two-parameter Poisson Dirichlet distribution with $\alpha = 0.51, \theta = 216$ approximated through a Chinese restaurant seating plan, each with its own number of costumers, corresponding to the different thickness of the lines.

5.2. Log-Likelihood

It is also interesting to investigate the shape of the log-likelihood function for α and θ given $\pi_{[n+1]}$. It is defined as

$$l_{n+1}(\alpha, \theta) := \log p(\pi_{[n+1]} | \alpha, \theta).$$

In Figure 6, the log-likelihood reparametrized using $\phi = n\dfrac{1-\alpha}{n+1+\theta}$ instead of α is displayed. A Gaussian distribution centered in the MLE parameters and with covariance matrix the inverse of the Fisher Information is also displayed (in dashed lines). This is not done to show an asymptotic property, but to show the symmetry of the log-likelihood, which validates approximation of $\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})$ with the marginal mode Φ_{MLE}, at least when we choose a hyperprior $p(\phi, \theta)$ that is flat around $(\phi_{MLE}, \theta_{MLE})$: indeed, it holds that $p(\phi, \theta \mid \pi_{[n+1]}) \propto l_{n+1}(\phi, \theta) \times p(\phi, \theta)$.

Hence, one could safely make this approximation if one believed that this symmetry would also be true in the real data situation at hand:

$$\mathrm{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}. \tag{13}$$

Notice that this is equivalent to a hybrid approach, commonly called "Empirical Bayes", in which the parameters are estimated through the MLE (frequentist) and their values are plugged into the Bayesian LR. We would like to reiterate that we are not using maximum likelihood estimates of the parameters because we consider the likelihood ratio from a frequentist point of view. Our aim is to calculate a Bayesian likelihood ratio, and we have observed empirically that, using the maximum likelihood estimates of the parameters, we can approximate this value.

Hence, in case of a rare type match problem, and using the YHRD database as the reference database, we have $\log_{10} \mathrm{LR} = 4.59$ that corresponds to say that it is approximately 40,000 times more likely to observe the reduced data under the prosecution hypothesis than under the defense hypothesis.

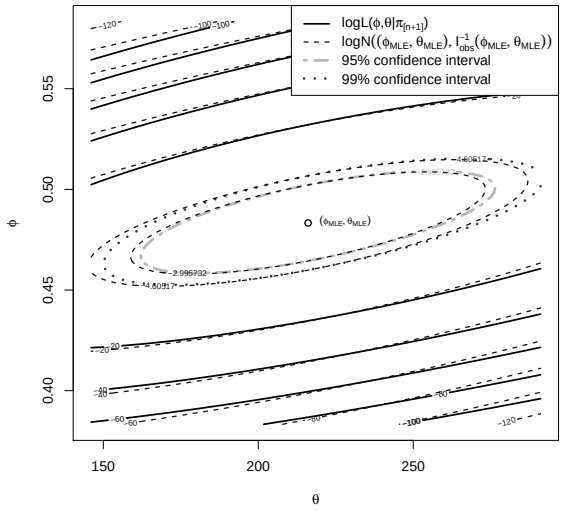


Figure 6. Relative log-likelihood for $\phi = n\frac{1-\alpha}{n+1+\theta}$ and θ compared to a Gaussian distribution displayed with 95% and 99% confidence intervals.

5.3. True LR

It is also interesting to study the likelihood ratio values obtained with out method according to formula (4), and to compare it with the 'true' ones, meaning the LR values obtained when the vector $\mathbf{p}$ is known, over simulated rare type match cases. This corresponds to the desirable (even though completely imaginary) situation of knowing the ranked list of the frequencies of all the DNA types in the population of interest. The model can be represented by the Bayesian network of Figure 7.

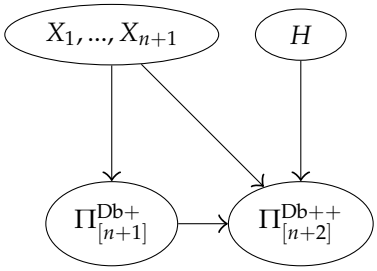


Figure 7. Bayesian network for the case in which $\mathbf{p}$ is known.

The likelihood ratio in this case can be obtained using again Corollary 1, where now X_1, ..., X_{n+1} play the role of Z. Indeed, now that $\mathbf{p}$ is known, the unobservable part of the model consists of the ranks of the types in the database.

$$\begin{aligned}
\text{LR}_{|\mathbf{p}} &= \frac{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_p, \mathbf{p})}{p(\pi_{[n+2]}, \pi_{[n+1]} \mid h_d, \mathbf{p})} \\
&= \frac{\mathbb{E}(p(\pi_{[n+2]} \mid \pi_{[n+1]}, X_1, ..., X_{n+1}, h_p, \mathbf{p}) \mid \Pi_{[n+1]} = \pi_{[n+1]}, \mathbf{p})}{\mathbb{E}(p(\pi_{[n+2]} \mid \pi_{[n+1]}, X_1, ..., X_{n+1}, h_d, \mathbf{p}) \mid \Pi_{[n+1]} = \pi_{[n+1]}, \mathbf{p})} \\
&= \frac{1}{\mathbb{E}(p_{X_{n+1}} | \Pi_{[n+1]} = \pi_{[n+1]}, \mathbf{p})}.
\end{aligned}$$

Notice that, in the rare type case, X_{n+1} is observed only once among the X_1, ..., X_{n+1}. Hence, we call it a singleton, and its distribution given $\mathbf{p}$, $\pi_{[n+1]}$ is the same as the distribution of each other

singleton. Let s_1 denote the number of singletons, and $\mathcal{S}$ the set of indices of singletons observations in the augmented database. It holds that

$$s_1 \mathbb{E}(p_{X_{n+1}} | \pi_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{i \in \mathcal{S}} p_{X_i} | \pi_{[n+1]}, \mathbf{p}).$$

Notice also that the knowledge of $\mathbf{p}$ and $\pi_{[n+1]}$ is not enough to observe $X_1, ..., X_{n+1}$.

Let us denote as X_1^*, ..., X_K^* the K different values taken by X_1, ..., X_{n+1}, ordered decreasingly according to the frequency of their values. Stated otherwise, if n_i is the frequency of x_i^* among $x_1, ..., x_{n+1}$, then $n_1 \geq n_2 \geq ... \geq n_K$. Moreover, in case X_i^* and X_j^* have the same frequency ($n_i = n_j$), then they are ordered increasingly according to their values. For instance, if $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, then $X_1^* = 4, X_2^* = 2, X_3^* = 3, X_4^* = 5, X_5^* = 9, X_6^* = 10, X_7^* = 13$.

By definition, it holds that

$$\mathbb{E}(\sum_{i \in \mathcal{S}} p_{X_i} | \pi_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{j:\, n_j = 1} p_{X_j^*} | \pi_{[n+1]}, \mathbf{p}).$$

Notice that $(n_1, n_2, ..., n_K)$ is a partition of $n+1$, which will be denoted as π_{n+1}. In the example, $\pi_{n+1} = (3, 2, 2, 1, 1, 1, 1)$. Since the distribution of $\sum_{j:\, n_j = 1} p_{X_j^*}$ only depends on π_{n+1}, the latter can replace $\pi_{[n+1]}$. Thus, it holds that

$$\mathrm{LR}_{|\mathbf{p}} = \frac{s_1}{\mathbb{E}(\sum_{j:\, n_j = 1} p_{X_j^*} | \pi_{n+1}, \mathbf{p})}. \tag{14}$$

A more compact representation for π_{n+1} can be obtained by using two vectors $\mathbf{a}$ and $\mathbf{r}$ where a_j are the distinct numbers occurring in the partition, increasingly ordered, and each r_j is the number of repetitions of a_j. J is the length of these two vectors, and it holds that $n+1 = \sum_{j=1}^{J} a_j r_j$. In the example above, we have that π_{n+1} can be represented by $(\mathbf{a}, \mathbf{r})$ with $\mathbf{a} = (1, 2, 3)$ and $\mathbf{r} = (4, 2, 1)$, $J = 3$.

There is an unknown map, χ, treated here as latent variable, which assigns the ranks of the DNA types, ordered according to their frequency in nature, to one of the number $\{1, 2, ..., J\}$ corresponding to the position in $\mathbf{a}$ of its frequency in the sample, or to 0 if the type if not observed. Stated otherwise,

$$\chi : \{1, 2, ...\} \longrightarrow \{0, 1, 2, ..., J\}$$

$$\chi(i) = \begin{cases} 0 & \text{if the } i\text{th most common species in nature is not observed in the sample,} \\ j & \text{if the } i\text{th most common species in nature is one of the } r_j \text{ observed } a_j \text{ times in the sample.} \end{cases}$$

Given $\pi_{n+1} = (\mathbf{a}, \mathbf{r})$, χ must satisfy the following set of J conditions:

$$\sum_{i=1}^{\infty} \mathbf{1}_{\chi(i) = j} = r_j, \qquad \forall j \in \{1, ..., J\}. \tag{15}$$

In addition, it should not be allowed that a profile observed k_N times in the population is observed $k_n > k_N$ times in the sample. Hence, we have to add a further condition:

$$N p_i > a_{\chi(i)}, \qquad \forall i \tag{16}$$

where N is the size of the entire population.

The map χ can be represented by a vector $\boldsymbol{\chi} = (\chi_1, \chi_2, ...)$ such that $\chi_i = \chi(i)$. In the example, we have that

$$\boldsymbol{\chi} = (0, 2, 2, 3, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, ...).$$

Notice that, given $\pi_{n+1} = (\mathbf{a}, \mathbf{r})$, the knowledge of $\boldsymbol{\chi}$ implies the knowledge of $X_1^*, ..., X_K^*$: indeed, it is enough to consider the position of the ranked positive values of $\boldsymbol{\chi}$, and to solve ties by considering the positions themselves (if $\chi_i = \chi_j$, than the order is given by i and j). For instance, in the example, if we sort the positive values of $\boldsymbol{\chi}$ and we collect their positions, we get $(4, 2, 3, 5, 9, 10, 13)$: the reader can notice that we got back to $X_1^*, ..., X_7^*$.

This means that, to obtain the distribution of $X_1^*, ..., X_K^* | \pi_{n+1}, \mathbf{p}$, which appears in (14), it is enough to obtain the distribution of $\boldsymbol{\chi} | \mathbf{a}, \mathbf{r}, \mathbf{p}$, and since we are only interested in the mean of the sum of singletons in samples of size $n+1$ from the distribution of $X_1^*, ..., X_K^* | \mathbf{a}, \mathbf{r}, \mathbf{p}$, we can just simulate samples from the distribution of $\boldsymbol{\chi} | \mathbf{a}, \mathbf{r}, \mathbf{p}$ and sum the p_i such that $\chi_i = 1$.

It holds that

$$p(\mathbf{a}, \mathbf{r} \mid \boldsymbol{\chi}, \mathbf{p}) \propto \prod_{1 \leq i \leq m} p_i^{a_{\chi_i}}, \tag{17}$$

where the proportionality factor is $\frac{(n+1)!}{\prod_{1 \leq j \leq J} (a_j!)^{r_j}}$.

5.3.1. Details of the Metropolis–Hastings Algorithm

Notice that for the model we assumed $\mathbf{p}$ to be infinitely long, but for simulations we will use a finite $\bar{\mathbf{p}}$, of length m. This is equivalent to assume that only m elements in the infinite $\mathbf{p}$ are positive, and the remaining infinite tail is made of zeros.

To simulate samples from the distribution of $\boldsymbol{\chi} | \mathbf{a}, \mathbf{r}, \mathbf{p}$, we use a Metropolis–Hastings algorithm on the space of the vectors $\boldsymbol{\chi}$ satisfying the $J + m$ conditions (15) and (16). Then, the state space of the Metropolis–Hastings Markov chain is made of all vectors of length m whose elements belong to $\{0, 1, ..., J\}$, and satisfy the conditions (15) and (16). If we start with an initial point $\boldsymbol{\chi}_0$ which satisfies (15) and, at each move t of the Metropolis–Hastings, we swap two different values χ_i and χ_j inside the vector, condition (15) remains satisfied while conditions (16) must be checked at every iterations. The Metropolis factor is the ratio of the two likelihoods $p(\mathbf{a}, \mathbf{r} \mid \boldsymbol{\chi}_t, \mathbf{p})$ and $p(\mathbf{a}, \mathbf{r} \mid \boldsymbol{\chi}_{t+1}, \mathbf{p})$, where $\boldsymbol{\chi}_t$ and $\boldsymbol{\chi}_{t+1}$ differ only because χ_i and χ_j are exchanged. Hence, using (17), the Metropolis factor for every move is

$$R = \frac{p_i^{a_{\chi_i}} p_j^{a_{\chi_j}}}{p_j^{a_{\chi_i}} p_i^{a_{\chi_j}}}.$$

Every exchange move is then accepted with probability R. The algorithm is iterated $N = 10^5$ times, with thinning steps of 10^3 and a burn-in period of 20000 iterations. Since it holds that

$$\mathbb{E}\Big(\sum_{j:\, n_j = 1} p_{X_j^*} | \pi_{n+1}^{\text{Db+}}, \mathbf{p} \Big) = \mathbb{E}\Big(\sum_{i:\chi_i = 1} p_i | \mathbf{a}, \mathbf{r}, \mathbf{p} \Big),$$

for every accepted $\boldsymbol{\chi}$, we calculate the sum of all p_is such that $\chi_i = 1$ and we use the average to approximate the denominator of (14). The algorithm is based on a similar one proposed in Anevski et al. [15].

This method allows us to approximate the 'true' LR when the vector $\mathbf{p}$ is known. This is almost never the case, but we can put ourselves in a fictitious world where we know $\mathbf{p}$ (such as the frequencies in the YHRD database, or as in the following section the frequencies from a smaller population) and compare the true values for the $\text{LR}_{|\mathbf{p}}$ with the one obtained by applying our Bayesian nonparametric model when $\mathbf{p}$ is unknown.

5.4. Frequentist-Bayesian Analysis of the Error

A real Bayesian statistician chooses the prior and hyperprior according to his beliefs. Depending on the choice of the hyperprior over α and θ, she may or may not believe in the approximation (13), but she does not really talk of "error". However, hardliner Bayesian statisticians are a rare species,

and most of the time the Bayesian procedure consists of choosing priors (and hyperpriors) which are a compromise between personal beliefs and mathematical convenience. It is thus interesting to investigate the performance of such priors. This can be done by comparing the Bayesian likelihood ratio with the likelihood ratio that one would obtain if the vector $\mathbf{p}$ was known, and for the same reduction of data. This is what we call "error": in other words, at the moment, we are considering the Bayesian nonparametric method proposed in this paper as a way to estimate (notice the frequentist terminology) the true $\mathrm{LR}_{|\mathbf{p}}$. If we denote by p_x the population proportion of the matching profile, another interesting comparison is the one between the Bayesian likelihood ratio and the frequentist likelihood ratio $1/p_x$ (here denoted as LR_f) that one would obtain knowing $\mathbf{p}$, but not reducing the data to partition. This is a sort of benchmark comparison and tells us how much we lose by using the Bayesian nonparametric methodology, and by reducing data.

In total, there are three quantities of interest ($\log_{10}\mathrm{LR}$, $\log_{10}\mathrm{LR}_{|\mathbf{p}}$, and $\log_{10}\mathrm{LR}_f$), and two differences of interest, which will be denoted as

- $\mathrm{Diff}_1 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_{|\mathbf{p}}$ (loss due to choice of the Poisson Dirichlet model and approximation (13)),
- $\mathrm{Diff}_2 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_f$ (overall loss).

In order to analyze these five quantities, we can study their distribution over different rare type match cases. However, there is an obstacle. The Metropolis–Hastings algorithm described in Section 5.3 is too slow to be used with the entire European database of Purps et al. [37] of size $n = 18,925$.

In order to make the computational effort feasible, we consider the haplotype frequencies for the sole Dutch population (of size $n = 2037$), and we pretend that they are the frequencies from the entire population of possible perpetrators. This population is summarized by the following $\mathbf{a}$, and $\mathbf{r}$:

$$\mathbf{a} = (1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,19,20,23,24,29,35,41,46,94,152,168,174)$$

$$\mathbf{r} = (356,80,31,20,13,11,5,6,3,5,4,3,2,3,1,1,1,1,1,1,1,1,1,1,2,1,1,1),$$

and the maximum likelihood estimators for α and θ are $\alpha_{\mathrm{MLE}} = 0.62$, $\theta_{\mathrm{MLE}} = 22$.

In this way, we can use the Metropolis–Hastings algorithm to simulate $\mathrm{LR}_{|\mathbf{p}}$. The model fitting is still good enough, as shown in Figure 8 (as a side note, notice that the asymptotic behavior is reached faster for this smaller value of $\theta_{\mathrm{MLE}} = 22$).

However, it is important to stress that the Gaussian shape and consequently the approximation (13) is not empirically supported for small databases of size $n = 100$.

In Table 1 and Figure 9a, we compare the distribution of $\log_{10}\mathrm{LR}_{|\mathbf{p}}$, $\log_{10}\mathrm{LR}$, and $\log_{10}\mathrm{LR}_f$ obtained by 96 samples of size 100 from the Dutch population . Each sample represents a different rare type match case with a specific database of reference of size $n = 100$.

The distribution of the benchmark likelihood ratio $\log_{10}\mathrm{LR}_f$ has more variation than the distribution of the Bayesian likelihood ratio, while $\log_{10}\mathrm{LR}_{|\mathbf{p}}$ appears to be the most concentrated around its mean.

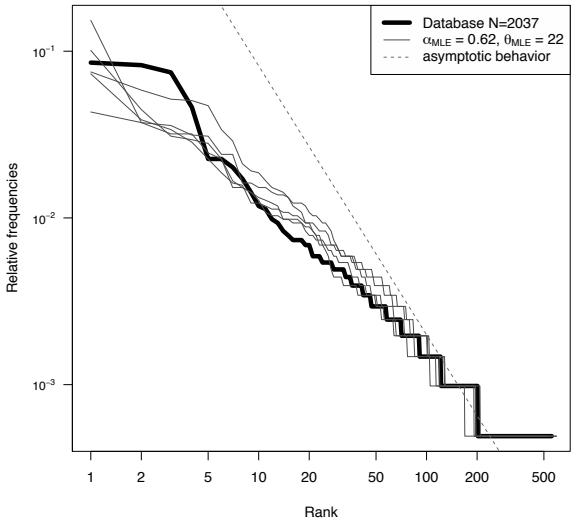


Figure 8. Log scale ranked frequencies from the Dutch database (thick line) compared to the relative frequencies of samples of size $n = 2037$ obtained from several realizations of PD($\alpha_{MLE}, \theta_{MLE}$) (thin lines). Asymptotic power-law behavior is also displayed (dotted line).

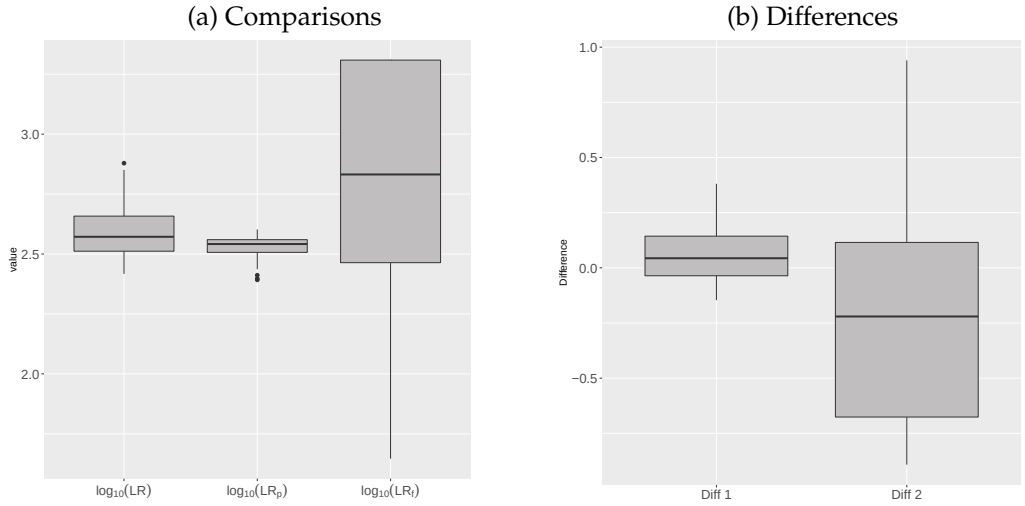


Figure 9. (**a**) comparison between the distribution of $\log_{10}$ LR, $\log_{10}$ LR$_{|\mathbf{p}}$, and $\log_{10}$ LR$_f$; (**b**) the error $\log_{10}$ LR $-$ $\log_{10}$ LR$_{|\mathbf{p}}$ and $\log_{10}$ LR $-$ $\log_{10}$ LR$_f$.

In Table 2 and Figure 9b, we consider the distribution of the two differences, Diff$_1$ and Diff$_2$. Diff$_1$ is the smallest and the most concentrated: it ranges between -0.146 and 0.381 and has a small standard deviation. It means that the nonparametric Bayesian likelihood ratio obtained as in (13) can be thought of as a good approximation of the frequentist likelihood ratio for the same reduction of data ($\log_{10}$ LR$_{|\mathbf{p}}$), even though we have not empirically validated the approximation for small databases of size 100. This difference is due to three things: the approximation (13), the MLE estimation of the hyperparameters, and the choice of a prior distribution (two-parameter Poisson Dirichlet) which is quite realistic, as shown in Figure 8, but not perfectly fitting the actual population.

Table 1. Summaries of the distribution of $\log_{10}\mathrm{LR}$, $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$, and $\log_{10}\mathrm{LR}_f$.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
$\log_{10}\mathrm{LR}$	2.417	2.512	2.572	2.59	2.658	2.879	0.102
$\log_{10}\mathrm{LR}_{\|\mathbf{p}}$	2.392	2.507	2.542	2.529	2.56	2.602	0.045
$\log_{10}\mathrm{LR}_f$	1.646	2.464	2.832	2.803	3.309	3.309	0.463

Table 2. Summaries of the distribution of Diff_1, Diff_2, and Diff_3.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
Diff_1	−0.146	−0.036	0.044	0.06	0.144	0.381	0.126
Diff_2	−0.891	−0.676	−0.221	−0.213	0.115	0.94	0.472

Notice that the difference increases if the Bayesian nonparametric likelihood is compared to the benchmark likelihood ratio (Diff_2). Here, the reduction of data comes into play too. However, the difference ranges within one order of magnitude, but most of the time lies between −0.676 and 0.115; thus, it is small.

6. Conclusions

This paper discusses the first application of a Bayesian nonparametric method to likelihood ratio assessment in forensic science, in particular to the challenging situation of the rare type match. If compared to traditional Bayesian methods such as those described in Cereda [6], it presents many advantages. First of all, the prior chosen for the parameter $\mathbf{p}$ seems to be quite realistic for the population whose frequencies we want to model. Moreover, though the theoretical background on which it rests may seem very technical and difficult, the method is extremely simple in practice, thanks to the use of an empirical Bayes approximation. More could be done in the future: in particular regarding approximation (13). The posterior expectation in the denominator could, for instance, be treated using MCMC algorithms or ABC algorithms. Then, we can try to improve the efficiency of the Metropolis–Hastings algorithm defined in Section 5.3 in order to be used with bigger and better populations. The big problem is how to use these methods when relevant populations are poorly defined and accessible databases are of doubtful relevance. We don't solve those problems.

It is not clear whether other methods are better. This is all very open and controversial. We suggest the analyst to perform several very different analyses and think carefully what the differences between the conclusions tells her. With this aim, we plan to compare this Bayesian nonparametric method to other existing methods for the rare type match problem, investigating calibration and validation through the use of ECE plots [49].

Author Contributions: The original version of this paper was written while the G. Cereda was PhD student at Leiden University under the supervision of R.D. Gill. The lead author, G. Cereda, contributed the original conception of this paper, performed most of the literature research, computation, and drafted the original manuscript. Intensive discussions and much rewriting and revision by the two authors together, during which both of them contributed further new ideas, led finally to this version.

Acknowledgments: We are indebted to Jim Pitman and Alexander Gnedin for their help in understanding their important theoretical results, to Mikkel Meyer Andersen for providing a cleaned version of the database of Purps et al. [37] and to Stephanie Van der Pas, Pierpaolo De Blasi, Giacomo Aletti, and Fabio Corradi for their useful opinions and comments. This research was supported by the Swiss National Science Foundation through Grant No. P2LAP2_178195.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Robertson, B.; Vignaux, G.A. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*; John Wiley & Sons: Chichester, UK, 1995.
2. Evett, I.; Weir, B. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*; Sinauer Associates: Sunderland, UK, 1998.
3. Aitken, C.; Taroni, F. *Statistics and the Evaluation of Evidence for Forensics Scientists*; John Wiley & Sons: Chichester, UK, 2004.
4. Balding, D. *Weight-of-Evidence for Forensic DNA Profiles*; John Wiley & Sons: Chichester, UK, 2005.
5. Taroni, F.; Aitken, C.; Garbolino, P.; Biedermann, A. *Bayesian Networks and Probabilistic Inference in Forensic Science*; John Wiley & Sons: Chichester, UK, 2006.
6. Cereda, G. Bayesian approach to LR in case of rare type match. *Stat. Neerl.* **2017**, *71*, 141–164.
7. Cereda, G. Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scand. J. Stat.* **2017**, *44*, 230–248.
8. Brenner, C.H. Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Sci. Int. Genet.* **2010**, *4*, 281–291.
9. Cereda, G.; Biedermann, A.; Hall, D.; Taroni, F. An investigation of the potential of DIP-STR markers for DNA mixture analyses. *Forensic Sci. Int. Genet.* **2014**, *11*, 229–240.
10. Laplace, P. *Essai Philosophique sur les Probabilites*; Mme. Ve Courcier: Paris, France, 1814.
11. Krichevsky, R.; Trofimov, V. The performance of universal coding. *IEEE Trans. Inf. Theory* **1981**, *27*, 199–207.
12. Gale, W.A.; Church, K.W. What's wrong with adding one? *Corpus-Based Research into Language*; Rodolpi: Amsterdam, The Netherlands, 1994.
13. Good, I. The population frequencies of species and the estimation of population parameters. *Biometrika* **1953**, *40*, 237–264.
14. Orlitsky, A.; Santhanam, N.P.; Viswanathan, K.; Zhang, J. On Modeling Profiles Instead of Values. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04) pp. 426–435.
15. Anevski, D.; Gill, R.D.; Zohren, S. Estimating a probability mass function with unknown labels. *Ann. Stat.* **2017**, *45*, 2708–2735.
16. Tiwari, R.C.; Tripathi, R.C. Nonparametric Bayes estimation of the probability of discovering a new species. *Commun. Stat. Theory Methods* **1989**, *A18*, 877–895.
17. Lijoi, A.; Mena, R.H.; Pruenster, I. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **2007**, *94*, 769–786.
18. De Blasi, P.; Favaro, S.; Lijoi, A.; Mena, R.H.; Pruenster, I.; Ruggiero, M. Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process. *IEEE Trans. Patterns Anal. Ans Mach. Intell.* **2015**, *37*, 212–229.
19. Favaro, S.; Lijoi, A.; Mena, R.H.; Pruenster, I. Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. R. Stat. Soc. Ser. (Methodol.)* **2009**, *71*, 993–1008.
20. Arbel, J.; Favaro, S.; Nipoti, B.; Teh, Y.W. Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Stat. Sin.* **2017**, *27*, 839–859.
21. Favaro, S.; Nipoti, B.; Teh, Y.W. Rediscovery of Good-Turing estimators via Bayesian nonparametrics. *Biometrics* **2016**, *72*, 136–145.
22. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* **2006**, *101*, 1566–1581.
23. Newman, M. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323–351.
24. Caliebe, A.; Jochens, A.; Willuweit, S.; Roewer, L.; Krawczak, M. No shortcut solutions to the problem of Y-STR match probability calculation. *Forensic Sci. Int. Genet.* **2015**, *15*, 69–75.
25. Andersen, M.M.; Curran, J.M.; de Zoete, J.; D., T.; Buckleton, J. Modelling the dependence structure of Y-STR haplotypes using graphical models. *Forensic Sci. Int. Genet.* **2018**, *37*, 29–36.
26. Andersen, M.M.; Caliebe, A.; Kirkeby, K.; Knudsen, M.; Vihrsa, N.; Curran, J.M. Estimation of Y haplotype frequencies with lower order dependencies. *Forensic Sci. Int. Genet.* **2019**, *46*, 102214.
27. Balding, D.J.; Nichols, R.A. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* **1994**, *64*, 125–140.
28. Andersen, M.M.; Balding, D.J. How convincing is a matching Y-chromosome profile? *Plos Genet.* **2017**, *13*, e1007028.

29. Andersen, M.M.; Balding, D.J. Y-profile evidence: Close paternal relatives and mixtures. *Forensic Sci. Int. Genet.* **2019**, *38*, 48–53.
30. Egeland, T.; Salas, A. Estimating Haplotype Frequency and Coverage of Databases. *PLoS ONE* **2008**, *3*, e3988.
31. Roewer, L. Y chromosome STR typing in crime casework. *Forensic Sci. Med. Pathol.* **2009**, *5*, 77–84.
32. Buckleton, J.; Krawczak, M.; Weir, B. The interpretation of lineage markers in forensic DNA testing. *Forensic Sci. Int. Genet.* **2011**, *5*, 78–83.
33. Willuweit, S.; Caliebe, A.; Andersen, M.M.; Roewer, L. Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Sci. Int. Genet.* **2011**, *5*, 84–90.
34. Wilson, I.J.; Weale, M.E.; Balding, D.J. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. Ser. (Stat. Soc.)* **2003**, *166*, 155–188.
35. Andersen, M.M.; Eriksen, P.S.; Morling, N. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *J. Theor. Biol.* **2013**, *329*, 39–51.
36. Willuweit, S.; Roewer, L. Y chromosome haplotype reference database (YHRD): Update. *Forensic Sci. Int. Genet.* **2007**, *1*, 83–87. doi:10.1016/j.fsigen.2007.01.017.
37. Purps, J.; Siegert, S.; Willuweit, S.; Nagy, M.; Alves, C.; Salazar, R.; Angustia, S.M.T.; Santos, L.H.; Anslinger, K.; Bayer, B.; et al. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci. Int. Genet.* **2014**, *12*, 12–23. doi:http://dx.doi.org/10.1016/j.fsigen.2014.04.008.
38. Kimura, M. The number pf alleles that can be maintained in a finite population. *Genetics* **1964**, *49*, 725–738.
39. Hjort, N.; Holmes, C.; Müller, P.; Walker, S. *Bayesian Nonparametrics*; Cambridge University Press: Cambridge, UK, 2010.
40. Ghosal, S.; Van der Vaart, A. *Fundamentals of Nonparametric Bayesian Inference*; Cambridge University Press, Cambridge, UK, 2017.
41. Pitman, J.; Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **1997**, *25*, 855–900.
42. Feng, S. *The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviors*; Springer: Berlin/Heidelberg, Germany, 2010.
43. Buntine, W.; Hutter, M. A Bayesian view of the Poisson-Dirichlet process. *arXiv* **2012**, arXiv:1007.0296.
44. Pitman, J. *Combinatorial Stochastic Processes*; Springer: Berlin/Heidelberg, Germany, 2006.
45. Zabell, S.L. *The Continuum of Inductive Methods Revisited; Cambridge Studies in Probability, Induction and Decision Theory*; Cambridge University Press: Cambridge, UK, 2005; pp. 243–274.
46. Pitman, J. The two-parameter generalization of Ewens' random partition structure. Technical report 345, Department of Statistics U.C. Berkeley CA, 1992.
47. Pitman, J. Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields* **1995**, *102*, 145–158.
48. Aldous, D.J. *Exchangeability and Related Topics*; Springer-Verlag: New York, NY, USA, 1985.
49. Ramos, D.; Gonzales-Rodriguez, J.; Zadora, G.; Aitken, C. Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods. *J. Forensic Sci.* **2013**, *58*, 1503–1517.