Forensic Science International: Genetics 19 (2015) 86–91

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Fitting the Balding–Nichols model to forensic databases

Rori V. Rohlfs [a,*], Vitor R.C. Aguiar [c,1], Kirk E. Lohmueller [b], Amanda M. Castro [d], Alessandro C.S. Ferreira [d], Vanessa C.O. Almeida [d], Iuri D. Louro [c], Rasmus Nielsen [a]

[a] University of California, Berkeley, Department of Integrative Biology, United States
[b] University of California, Los Angeles, Department of Ecology and Evolutionary Biology, United States
[c] Universidade Federal do Espírito Santo, Departamento de Ciências Biológicas, Brazil
[d] Laboratório Hermes Pardini, Departamento de Genética Molecular, Brazil

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form 15 April 2015
Accepted 5 May 2015
Available online 23 June 2015

Keywords:
Balding–Nichols model
Population genetics
Partial match probability
DNA database

ABSTRACT

Large forensic databases provide an opportunity to compare observed empirical rates of genotype matching with those expected under forensic genetic models. A number of researchers have taken advantage of this opportunity to validate some forensic genetic approaches, particularly to ensure that estimated rates of genotype matching between unrelated individuals are indeed slight overestimates of those observed. However, these studies have also revealed systematic error trends in genotype probability estimates. In this analysis, we investigate these error trends and show how they result from inappropriate implementation of the Balding–Nichols model in the context of database-wide matching. Specifically, we show that in addition to accounting for increased allelic matching between individuals with recent shared ancestry, studies must account for relatively decreased allelic matching between individuals with more ancient shared ancestry.

1. Introduction

Forensic databases, rapidly increasing in size, invite powerful analyses of rates of coincidental genotype matching [1–3]. Such analyses have validated some basic assumptions in forensic genetics, particularly the reasonable over-estimation of genotype frequencies with existing methods. However, these studies also illustrate how database population genetic diversity differs from what is expected under the basic model of forensic genetics: the Balding–Nichols (BN) model.

The BN model simply and elegantly provides a framework for estimating probabilities of observed genotypes, taking into account population structure and variance in allele frequency estimates [4,5]. The BN model can be interpreted as describing an ancestral population which has split into a number of internally randomly mating sub-populations which evolve independently over some time, resulting in a present-day total population made up of a number of cryptic sub-population groups. The sampling probabilities estimated under the BN model then incorporate the increase in allele sharing between individuals from the same sub-population due to their shared co-ancestry.

The amount of excess allele-sharing in a sub-population group beyond what is expected based on the total population allele frequencies can be quantified in the BN model by the parameter θ. θ can be thought of as the probability that two alleles in a sub-population are identical by descent (IBD) due to within sub-population shared ancestry. In a coalescent framework under simplifying assumptions, it represents the probability that two alleles sampled from within a sub-population coalesce before either mutates or migrates out of the sub-population [4].

In the BN model used in forensic applications, the probability of observing a particular genotype conditioning on having observed the same genotype is estimated using the θ correction to account for coincidental allelic sharing between two individuals due to excess shared ancestry within a sub-population. In most forensic calculations, there is an implicit assumption that the individuals in question are from the same sub-population [4]. Balding and Nichols convincingly argue that this assumption is appropriate, saying "the 'same sub-population' assumption is conservative, since the suspect's profile will tend to be more common in his/her sub-population than in other groups" [4]. A number of studies have shown the importance and appropriateness of this assumption and the corresponding θ correction in genetic identification calculations to provide an over-estimate of genotype frequencies [5–12]. Curran et al. [10] in particular have argued that this implementation of the BN model over-corrects for population structure.

* Corresponding author.
E-mail address: rrohlfs@berkeley.edu (R.V. Rohlfs).
[1] Current address: University of São Paulo, Department of Genetics and Evolutionary Biology, Brazil.

http://dx.doi.org/10.1016/j.fsigen.2015.05.005
1872-4973/

In database applications, on the other hand, typically all pairs of genotypes in a database will be compared to each other and their degree of matching assessed. Previous applications of the standard BN model to forensic databases [1,2] have shown that the often-used θ correction of 0.01 usually adequately corrects for coincidental allele-sharing, raising estimated probabilities of matching genotypes above their observed levels, and therefore reducing false positive rates below their expectation (in statistical terms, making the test 'conservative'). Yet, these analyses show an excess of non-similarity between observed pairs of individuals, as compared to the expectation [1,2]. As we will show, this is likely due to the fact that the standard formulation of the BN model does not take decreased allele sharing between individuals from different sub-populations into account. When applying the BN model to describe the amount of genotypic matching observed in a database, it is not clear that the 'same sub-population' assumption is appropriate.

In this manuscript, we investigate how empirical genotype matching observations can be explained by reconsidering the implementation of the BN model. We show that by accounting for the case of two individuals deriving from different population groups, we significantly improve the ability to describe empirical matching rates in a database.

2. Methods

2.1. Allele sharing matrix

To quantify the degree of multi-locus genotype matching within a data set, consider the matrix M where each entry $M_{m,p}$ is the number of profile pairs with m markers matching at both alleles and p markers matching at one allele [1,2]. Tvedebrink et al. [2] described a recursive algorithm to compute the probability $\pi_{m,p}$ that two multi-locus genotypes completely match at m loci and partially match at p loci, constructing a probability matrix π analogous to M. This method uses the single-locus probabilities of individuals matching two, one, and zero alleles as $P_{1,0}$, $P_{0,1}$, and $P_{0,0}$, respectively, following in the notation of Tvedebrink et al. [2]. Note the parallel notation to counts of matching and partially matching markers in $M_{m,p}$. Weir [1] described how to compute $P_{1,0}$, $P_{0,1}$, and $P_{0,0}$ at a locus by summing over the appropriate two-individual single locus genotype probabilities [1].

2.2. Single locus allelic sharing probabilities

2.2.1. Individuals from the same sub-population group

Under the typical implementation of the BN model, where all individuals are assumed to be in the same sub-population group, the two-individual genotype probabilities are

$$
\begin{aligned}
P(A_{1,1}A_{1,1}, A_{1,1}A_{1,1}) &= p_1(\theta + (1-\theta)p_1)\left(\frac{2\theta + (1-\theta)p_1}{1+\theta}\right)\left(\frac{3\theta + (1-\theta)p_1}{1+2\theta}\right)\\
P(A_{1,1}A_{1,1}, A_{1,1}A_{1,2}) &= \frac{4p_1p_2(1-\theta)(\theta + (1-\theta)p_1)(2\theta + (1-\theta)p_1)}{(\theta+1)(2\theta+1)}\\
P(A_{1,1}A_{1,2}, A_{1,1}A_{1,2}) &= \frac{4p_1p_2(1-\theta)(\theta + (1-\theta)p_1)(\theta + (1-\theta)p_2)}{(\theta+1)(2\theta+1)}\\
P(A_{1,1}A_{1,2}, A_{1,1}A_{1,3}) &= \frac{8p_1p_2p_3(1-\theta)^2(\theta + (1-\theta)p_1)}{(\theta+1)(2\theta+1)}\\
P(A_{1,1}A_{1,1}, A_{1,2}A_{1,2}) &= \frac{2p_1p_2(1-\theta)(\theta + (1-\theta)p_1)(\theta + (1-\theta)p_2)}{(\theta+1)(2\theta+1)}\\
P(A_{1,1}A_{1,1}, A_{1,2}A_{1,3}) &= \frac{4p_1p_2p_3(1-\theta)^2(\theta + (1-\theta)p_1)}{(\theta+1)(2\theta+1)}\\
P(A_{1,1}A_{1,2}, A_{1,3}A_{1,4}) &= \frac{24p_1p_2p_3p_4(1-\theta)^3}{(\theta+1)(2\theta+1)}
\end{aligned}
\tag{1}
$$

where $A_{1,i}$ is an allele i drawn from the single sub-population 1, so, for example $P(A_{1,i}A_{1,i}, A_{1,i}A_{1,j})$ is the probability observing a homozygote and heterozygote sharing one allele, and p_i is the frequency of allele i.

2.2.2. Individuals from same or different sub-population groups

Under the BN model, if two individuals are not in the same population group, the probability that their alleles coalesce more recently than a mutation or migration event is zero. In other words, there is no increased chance of allele-sharing due to shared ancestry for individuals in different population groups. In that case, the probability of observing their genotypes is computed as a function of the observed allele frequencies without the θ correction.

We can allow individuals to be from different sub-populations by introducing a parameter d, which describes the probability that a pair of individuals are from different sub-population groups. This way, we fully describe the BN model with some individuals from the same sub-population group and some from differing groups. This technique is analogous to that used by Curran et al. [11] to condition on different genetic relationships. Under a model with population differentiation, two-individual genotype probabilities are

$$P(A_{\cdot,\cdot}A_{\cdot,\cdot}, A_{\cdot,\cdot}A_{\cdot,\cdot}) = (1-d)P(A_{1,\cdot}A_{1,\cdot}, A_{1,\cdot}A_{1,\cdot}) + dP(A_{1,\cdot}A_{1,\cdot}, A_{2,\cdot}A_{2,\cdot})$$

where subscript dots indicate any option such that $A_{\cdot,\cdot}$ is any allele drawn from any sub-population and $A_{1,\cdot}$ is any allele drawn from sub-population 1. Genotype probabilities for individuals from the same population are the same as under the typical implementation of the BN model and for individuals from different sub-populations the genotype probabilities are

$$
\begin{aligned}
P(A_{1,1}A_{1,1}, A_{2,1}A_{2,1}) &= p_1^2(\theta + (1-\theta)p_1)^2\\
P(A_{1,1}A_{1,1}, A_{2,1}A_{2,2}) &= 4p_1^2p_2(1-\theta)(\theta + (1-\theta)p_1)\\
P(A_{1,1}A_{1,2}, A_{2,1}A_{2,2}) &= 4p_1^2p_2^2(1-\theta)(1-\theta)\\
P(A_{1,1}A_{1,2}, A_{2,1}A_{2,3}) &= 8p_1^2p_2p_3(1-\theta)(1-\theta)\\
P(A_{1,1}A_{1,1}, A_{2,2}A_{2,2}) &= 2p_1p_2(\theta + (1-\theta)p_1)(\theta + (1-\theta)p_2)\\
P(A_{1,1}A_{1,1}, A_{2,2}A_{2,3}) &= 4p_1p_2p_3(1-\theta)(\theta + (1-\theta)p_1)\\
P(A_{1,1}A_{1,2}, A_{2,3}A_{2,4}) &= 24p_1p_2p_3p_4(1-\theta)(1-\theta)
\end{aligned}
$$

2.2.3. Chromosomes from same or different sub-populations

In the previous formulation of joint genotype probabilities for two individuals, it is assumed that in each individual, both chromosomes derive from the same sub-population. This is not realistic in post-colonial societies, where few individuals can trace all their ancestry to their current location, creating strong admixture. We describe an alternative model allowing alleles within individuals to be drawn from different, but correlated, sub-population groups. This model essentially accounts for population structure and admixture between sub-population groups. In this model there are k sub-populations of equal size and relation to each other. The correlation of sub-population draws within individuals is described by the parameter a. We can use this model to compute joint genotype probabilities, as shown in Supplementary materials.

2.2.4. Accounting for genetic relatives

These implementations of the BN model can be expanded to account for genetic relatives using the method proposed by Curran et al. [11]. In brief, the proportion of pairs of individuals with a particular genetic relationship is parameterized, allowing the probability of a pair of genotypes to be computed conditioning on genetic relationship. In this manuscript we consider the relationships of parent–offspring, sibling, half-sibling, and cousin so that the probability of a particular observed genotype G pair is

$$P(G) = p_{unrel}(P(G|n_{IBD}=0)) + p_{p-o}(P(G|n_{IBD}=1)) + p_{sib}(0.25P(G|n_{IBD}=0) + 0.5P(G|n_{IBD}=1) + 0.25P(G|n_{IBD}=2)) + p_{hsib}(0.5P(G|n_{IBD}=0) + 0.5P(G|n_{IBD}=1)) + p_{\cos}(0.75P(G|n_{IBD}=0) + 0.25P(G|n_{IBD}=1))$$

where n_{IBD} is the number of alleles shared IBD and p_{p-o}, p_{sib}, p_{hsib}, and p_{cos} are the probability that a pair of individuals is genetic parent–offspring, siblings, half-siblings, and cousins, respectively, such that the probability of no genetic relationship is $p_{unrel} = 1 - (p_{p-o} + p_{sib} + p_{hsib} + p_{cos})$.

2.3. Likelihood framework

With match probabilities specified by the aforementioned models, we can calculate the expectation π of the match matrix M under varying assumptions regarding allele frequencies, and parameters of the models: θ for the typical implementation of the BN model without population differentiation, θ and d for the model with population differentiation, and θ, a, and k for the model allowing admixture between sub-populations (Table 1). By taking the entries π as categorical probabilities in a multinomial distribution, we can compute the sampling probability of an observed instance of M, an approach used effectively in other population genetic applications [13].

Using the sampling probability of M as a likelihood function, we can estimate parameters of the model using maximum likelihood. Since the models described here are nested and fulfill standard regularity conditions, the asymptotic distributions of likelihood ratio test statistics (LRTs) are known to be chi square. Specifically, if we take the null hypothesis to be the typical implementation of the BN model with a fixed value of θ (say $H_0 : \theta = 0.01$), and the alternative to be the typical BN model implementation where θ varies ($H_a : \theta \neq 0.01$), the LRT is distributed as a chi square with one degree of freedom ($LRT \sim \chi^2_1$). Similarly, to compare the typical implementation of the BN model with our implementation with population differentiation, we specify $H_0 : \theta \neq 0.01$, $d = 0$ and $H_a : \theta \neq 0.01$, $d \neq 0$, in which case the $LRT \sim \chi^2_0 + \chi^2_1$. The model allowing chromosomes within individuals from different sub-populations reduces to the model with population differentiation under complete allelic correlation ($a = 1$). In this case, d is equivalent to $(k-1)/k$. This enables tests where $LRT \sim \chi^2_0 + \chi^2_1$ between the chromosomal model and the model with population differentiation.

Additionally, we can obtain maximum likelihood parameter estimates and likelihood profile confidence intervals of $\hat{\theta}$, $\hat{d}$, $\hat{a}$, and $\hat{k}$. While we do not advocate interpreting these estimates too strongly, as the underlying population models are very simple, we can compare them as a reference.

2.4. Database

We consider genotype data from 98,988 Brazilian individuals undergoing paternity testing during 2011–2013 in the Hermes Pardini Laboratory, Vespasiano, MG, Brazil. The individuals genotyped reside in all 26 Brazilian States and the Federal District (Brasilia). Although the population genetic background of these particular individuals is unknown, generally Brazilian populations show ancestry from Indigenous South America, Africa, and Europe [14–16]. It is unclear if these ancestries are represented proportionally in this database. The genotypes were obtained using a combination of two Life Technologies kits and ABI 3730 Genetic analyzers (Life Technologies, CA, USA) for a total of 20 loci (the original 13 CODIS core loci and additionally D10S1248, D22S1045, D1S1656, D12S391, D2S441, D2S1338, D19S433) [17].

Multiple entries of the same individual are expected in this dataset. As such, pairs of individuals with 17–20 loci matching and the same birth dates (when available) were removed as likely multiple entries or identical twins with some genotyping or clerical errors. When birth dates were not available or inconsistent apparently due to a typo, names were manually checked by the lab personnel and the most complete profile was kept, resulting in a data set with 96,400 individuals [17].

While there are no known genetic relatives in this dataset, some unknown genetic relatives are likely present.

Since our analysis requires genotypes across the same number of loci for all individuals, we discarded all individuals with any missing data. In the remaining data set, extremely rare alleles observed exactly one time may be, in fact, genotyping errors. Profiles with these rare alleles were similarly eliminated. The final dataset considered in this analysis contained 90,852 individuals.

Table 1
The maximum log likelihoods are listed for each model considered, both accounting for genetic relatives and not.

Model	Log likelihood	
	Without relatives	With relatives
Typical implementation with $\theta = 0.01$	−53,070,085	−53,070,085
Typical implementation with θ varying	−1,259,283	−1,259,283
Population differentiation implementation	−37,234	−37,234
Sub-population groups by chromosome	−37,182	−37,182

3. Results

3.1. Observed database matching

We counted the number of zero, one, and two allele matches for each locus for each pair of individuals in the dataset to create the observed matrix M_{obs}, as shown in Supplementary Table 1. For example, in our dataset, out of $\binom{90852}{2} = 4,126,997,526$ pairs of genotypes, 295,948 pairs have exactly one locus matching at both alleles, two loci matching at one allele (partially matching), and 17 markers matching at neither allele (Supplementary Table 1).

3.2. Comparing data likelihood under different models

Previous investigators have used the conventional implementation of the BN model (without population differentiation) with θ fixed at 0.01 to describe matching in databases [1,2]. Under this model, setting $\theta = 0.01$, we calculated the log likelihood of the observed match matrix as −53,070,085 (Table 1). We can graphically compare our observed and expected results in a dropping ball diagram [11,18,2] (Fig. 1), or in a heat map of the residuals (Fig. 2a). The heat maps in this manuscript show a color gradient along the log of the divergence of the observed and expected as $((obs - exp)^2/exp)$. Through these visualizations, we see that as in previous analyses [1,11,2], under the typical BN model implementation with θ set at 0.01, there is an excess of observed pairs of individuals who share few alleles, as compared to the expectation.

Using the maximum likelihood framework and optimizing over θ, we performed a similar analysis (Table 1 and Supplementary Table 2). This model where θ may vary fits the observed data significantly better than with θ fixed at 0.01 ($LRT = 103,621,604$). However, we still observed an excess of individuals sharing few alleles (Fig. 2b). Further, under the maximum likelihood of this model, θ is estimated near zero as $\hat{\theta} = 4.6 \times 10^{-10}$ (in 95% likelihood profile confidence interval $(0.0, 2.07 \times 10^{-9})$), indicating

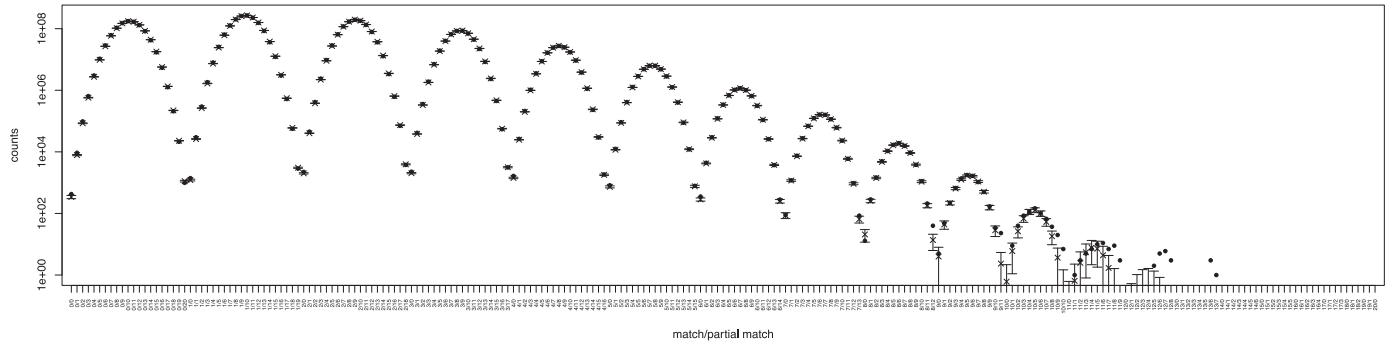

Fig. 1. This dropping ball plot shows the observed (dot) and expected (x) numbers of pairs of individuals sharing m matching loci and p partially matching loci where m/p is indicated on the x-axis.

that the θ correction as implemented here does not improve model fit (Supplementary Tables 3 and 4). This makes sense since the θ correction only accounts for excess allelic sharing due to common ancestry within a sub-population.

We allow individuals in different sub-populations to share comparatively fewer alleles through common ancestry using the population differentiation model, where two random individuals derive from different sub-population groups with probability

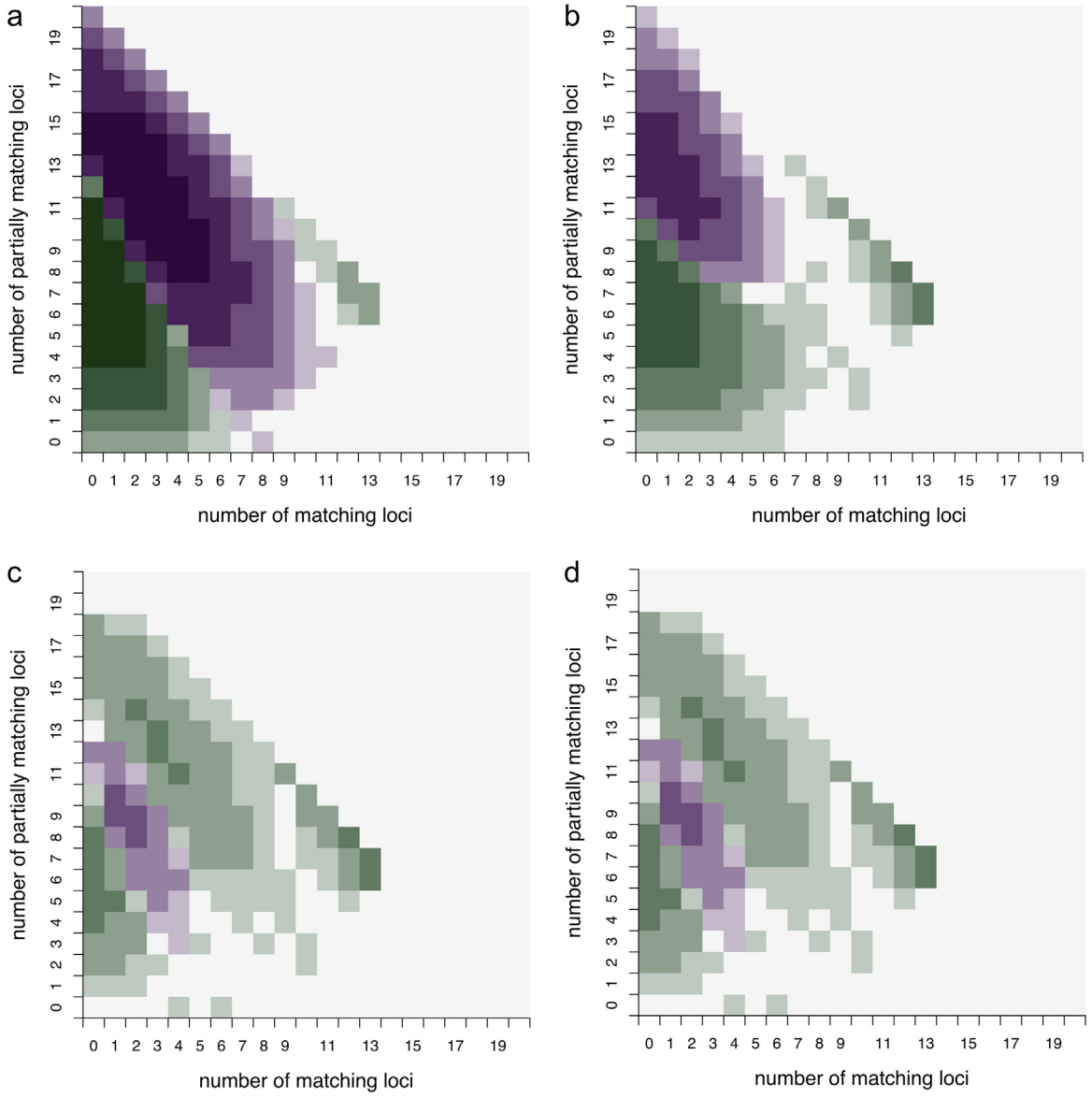

Fig. 2. These heat maps show the difference between the observed match matrix and that expected under, (a) the typical implementation of the BN model with $\theta = 0.01$, (b) the typical implementation of the BN model where θ varies, (c) the full implementation of the BN model, and (d) the full implementation of the BN model allowing for admixture. Purple indicates a lack of observed pairs of individuals and green indicates an excess of observed pairs of individuals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

d. Again, we find the maximum likelihood value under this model (Table 1 and Supplementary Table 5). The model fits the observed data significantly better ($LRT = 2{,}444{,}098$) and corrects for the previous excess of individuals sharing few alleles (Fig. 2c). However, we still see consistent differences between the observed and expected allelic matching. Compared to the observed data, the population differentiation model predicts a more narrow range of allelic matching than what is observed.

In the population differentiation model, it is assumed that both alleles within an individual derive from the same sub-population. This assumption may not be valid in realistic cases of admixture, and can be relaxed using the a model where chromosomes are considered separately with some correlation. We fit such a model with k equally represented sub-populations and intra-individual allelic correlation to the observed data. This model, allowing chromosomes of different population origins within individuals, fits the data significantly better than the model without admixture ($LRT = 148{,}835$) (Fig. 2d). Still, we observe a wider range of allelic matching than is expected under these models.

3.3. Accounting for relatives

Since it is likely that some genetic relatives are present in this large dataset, all of these analyses were repeated accounting for genetic relatives. The likelihood results are quite similar to those computed without accounting for relatives (Table 1). In most cases the maximum likelihood estimates for the frequency of genetic relatives were 0.0 with narrow likelihood profile confidence intervals, with an exception under the model allowing chromosomes within an individual different population origins (Supplementary Tables 3 and 4). In all cases, a likelihood ratio test comparing each model disallowing relatives to the analogous model accounting for relatives fails to reject the null hypothesis (Table 1).

4. Discussion and conclusions

We have shown how a multinomial distribution on the expected match matrix can be used to calculate the sampling probability of an observed match matrix. Further, we have shown how this probability can be maximized with respect to some parameters to provide maximum likelihood estimates of these parameters.

Using this procedure, we found that estimating the value of θ, fits the data significantly better than a uniform value of 0.01. This is intuitive since the typically used $\theta = 0.01$ was not chosen to fit the observed number of coincidental matches between pairs of profiles in a forensic database. Further, we found that estimate to be near zero. This initially surprising estimate is explained by considering that the common implementation of the BN model in forensic genetics accounts for excess allele sharing due to recent ancestry, but not relatively less allele sharing for individuals with more distant ancestry. Under this implementation, every pair of individuals has increased allelic sharing due to recent ancestry. Since many pairs of individuals do not share recent ancestry, the maximum likelihood estimate of θ is driven to zero to explain the lack of consistent excess allele sharing.

We show and implement several parameterizations of the full BN model where individuals may or may not have excess allele sharing (equivalently, may or may not derive from the same population group). This full BN model fits the observed match matrix significantly better than the typical BN model implementation.

We further implemented a model allowing admixture by letting chromosomes within individuals have different population origins, which fits the data significantly better than the model without admixture ($LRT = 148{,}835$). However, we caution against over-interpreting the maximum likelihood parameter point estimates of this model, especially considering the sometimes very wide likelihood profile confidence intervals (Supplementary Table 4). Despite similar log likelihoods, the parameter point estimates differ between the models with and without genetic relatives. This indicates that the flexible model accommodates a correlation structure in the data that may not be explicitly described by the parameters, for example more complex population structures including asymmetric migrations or varying sub-population group sizes. Again, this argues against a direct biological interpretation of parameter estimates.

In the BN model, all sub-populations have equal excess allele sharing internally and are equally unrelated to each other. While this model provides a simple and reasonable over-estimation of coincidental genotype match rates, essential to forensic case work, it is clearly a simplification of complex human population structures, where some individuals are vastly more related than others. A more sophisticated model allowing varying degrees of allele sharing between individuals would likely better fit our observation of a broad range of allelic-matching. However, such a model would begin to accumulate parameters, making use in forensic case work impractical compared to the adequate typical BN model implementation.

Additionally, the full BN model does not explain a small observed excess of people matching at many loci. For example, there are three pairs of individuals who match both alleles at 13 loci and one allele at six loci, whereas under the full BN model, $5.0e-13$ are expected. There are several possible explanations for these individuals. They may be genetic relatives who share a large number of alleles IBD. They could share even more alleles than expected if allele frequencies are mis-specified because they derive from a population group divergent from the whole sample [19]. It is also possible that the same individual was entered a number of times, with genotyping or clerical errors resulting in differing alleles. Of note, these individuals do share a common surname.

Like other authors [1,11,20,2], we consider presence of genetic relatives within a database when calculating genotype match probabilities so that the total probability of genotype matching takes into account the possibility of genetic relationships. Our results fail to reject the null hypothesis of no genetic relatives present. Since the loci are still treated independently, the small probability of a genetic relationship is factored in at each locus separately, rather than considering how genetic relatives share alleles across loci. As a result, unless there are extensive genetic relatives in a dataset, this does not dramatically affect the expected allelic matching.

We have shown how the correct full implementation of the BN model is crucial to understanding database-wide allelic matching. While this is essential for computing the number of expected matches in a large database, it does not affect forensic case work where the typical BN model implementation is adequate to reasonably overestimate the probability of coincidental genotype matching between a suspected contributor and the actual person who left the evidentiary profile.

Acknowledgments

We are immensely grateful to the individuals whose DNA samples were used in this study, without which none of this work would be possible. This work was supported in part by National Institutes of Health grant 2R14003229-07, National Science Foundation award 1103767, and a CAPES-Brazil Scholarship (Programa Demanda Social and PhD Student Exchange Program BEX 8425/11-6). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the

manuscript. This study was approved by the Ethics Committee of the Federal University of Espirito Santo (Brazil) Health Sciences Department (No. 448327).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2015.05.005.

References

[1] B. Weir, Matching and partially-matching DNA profiles, J. Forensic Sci. 49 (2004) 1009–1014.
[2] T. Tvedebrink, P.S. Eriksen, J.M. Curran, H.S. Mogensen, N. Morling, Analysis of matches and partial-matches in a Danish STR data set, Forensic Sci. Int. Genet. 6 (2012) 387–392.
[3] J. Curran, T. Tvedebrink, DNAtools: tools for empirical testing of DNA match probabilities, R package.
[4] D. Balding, R. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[5] D.J. Balding, R.A. Nichols, A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, in: B. Weir (Ed.), Human Identification: The Use of DNA Markers, Vol. 4 of Contemporary Issues in Genetics and Evolution, Springer, Netherlands, 1995, pp. 3–12.
[6] P. Gill, I. Evett, Population genetics of short tandem repeat (STR) loci, in: B. Weir (Ed.), Human Identification: The Use of DNA Markers, Vol. 4 of Contemporary Issues in Genetics and Evolution, Springer, Netherlands, 1995, pp. 69–87.
[7] I. Evett, P. Gill, J. Scranage, B. Weir, Establishing the robustness of short-tandem-repeat statistics for forensic applications, Am. J. Hum. Genet. 58 (1996) 398–407.
[8] I. Evett, P. Gill, J. Lambert, N. Oldroyd, R. Frazier, S. Watson, S. Panchal, A. Connolly, C. Kimpton, Statistical analysis of data for three British ethnic groups from a new STR multiplex, Int. J. Legal Med. 110 (1) (1997) 5–9.
[9] D. Balding, R. Nichols, Significant genetic correlations among Caucasians at forensic DNA loci, Hum. Hered. 78 (1997) 583–589.
[10] J.M. Curran, J.S. Buckleton, C.M. Triggs, What is the magnitude of the subpopulation effect? Forensic Sci. Int. 135 (2003) 1–8.
[11] J.M. Curran, S.J. Walsh, J. Buckleton, Empirical testing of estimated DNA frequencies, Forensic Sci. Int. Genet. 1 (2007) 267–272.
[12] B. Weir, The rarity of DNA profiles, Ann. Appl. Stat. 1 (2007) 358–370.
[13] A.M. Adams, R.R. Hudson, Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms, Genetics 168 (2004) 1699–1712.
[14] T.C. Lins, R.G. Vieira, B.S. Abreu, D. Grattapaglia, R.W. Pereira, Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs, Am. J. Hum. Biol. 22 (2010) 187–192.
[15] S.D.J. Pena, G. Di Pietro, M. Fuchshuber-Moraes, J.P. Genro, M.H. Hutz, F.d.S.G. Kehdy, F. Kohlrausch, L.A.V. Magno, R.C. Montenegro, M.O. Moraes, M.E.A.d. Moraes, M.R.d. Moraes, E.B. Ojopi, J.A. Perini, C. Racciopi, A.K.C. Ribeiro-dos Santos, F. Rios-Santos, M.A. Romano-Silva, V.A. Sortica, G. Suarez-Kurtz, The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected, PLoS ONE 6 (2011) e17063.
[16] S.R. Giolo, J.M.P. Soler, S.C. Greenway, M.A.A. Almeida, J.G.S. Mariza de Andrade, C.E. Seidman, J.E. Krieger, A.C. Pereira, Brazilian urban population genetic structure reveals a high degree of admixture, Eur. J. Hum. Genet. 20 (2012) 111–116.
[17] V.R. Aguiar, A.M. de Castro, V.C. Almeida, F.S. Malta, A.C. Ferreira, I.D. Louro, New CODIS core loci allele frequencies for 96,400 Brazilian individuals, Forensic Sci. Int. Genet. 13 (2014, November) e6–e12.
[18] J. Curran, Are DNA profiles as rare as we think? Or can we trust DNA statistics?, Significance 7 (2010) 62–66.
[19] R. Rohlfs, S. Fullerton, B. Weir, Familial identification: population structure and relationship distinguishability, PLoS Genet. 8 (2012) e1002469.
[20] L. Mueller, Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? J. Genet. 87 (2008) 101–108.