Fitting the Balding-Nichols model to forensic databases

Rori V. Rohlf^{a,*}, Vitor R. C. Aguiar^{c,**}, 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

^b*University of California, Los Angeles*

Department of Ecology and Evolutionary Biology

^c*Universidade Federal do Espírito Santo*

Departamento de Ciências Biológicas

^d*Laboratório Hermes Pardini*

Departamento de Genética Molecular

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 ances-

*rrohlfs@berkeley.edu

**Current address: University of São Paulo
Department of Genetics and Evolutionary Biology

try, studies must account for relatively decreased allelic matching between individuals with more ancient shared ancestry.

Keywords: Balding-Nichols model, population genetics, partial match probability, DNA database

1. Introduction

Forensic databases, rapidly increasing in size, invite powerful analyses of rates of coincidental genotype matching [1, 2, 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].

26 In the BN model used in forensic applications, the probability of observing
 27 a particular genotype conditioning on having observed the same genotype is
 28 estimated using the θ correction to account for coincidental allelic sharing be-
 29 tween two individuals due to excess shared ancestry within a sub-population.
 30 In most forensic calculations, there is an implicit assumption that the in-
 31 dividuals in question are from the same sub-population [4]. Balding and
 32 Nichols convincingly argue that this assumption is appropriate, saying “the
 33 ‘same sub-population’ assumption is conservative, since the suspect’s pro-
 34 file will tend to be more common in his/her sub-population than in other
 35 groups” [4]. A number of studies have shown the importance and appro-
 36 priateness of this assumption and the corresponding θ correction in genetic
 37 identification calculations to provide an over-estimate of genotype frequen-
 38 cies [5, 6, 7, 8, 9, 10, 11, 12]. Curran *et al.* [10] in particular have argued that
 39 this implementation of the BN model over-corrects for population structure.

40 In database applications, on the other hand, typically all pairs of geno-
 41 types in a database will be compared to each other and their degree of match-
 42 ing assessed. Previous applications of the standard BN model to forensic
 43 databases [1, 2] have shown that the often-used θ correction of 0.01 usually
 44 adequately corrects for coincidental allele-sharing, raising estimated probab-
 45 ities of matching genotypes above their observed levels, and therefore reduc-
 46 ing false positive rates below their expectation (in statistical terms, making
 47 the test ‘conservative’). Yet, these analyses show an excess of non-similarity
 48 between observed pairs of individuals, as compared to the expectation [1, 2].
 49 As we will show, this is likely due to the fact that the standard formulation
 50 of the BN model does not take decreased allele sharing between individuals

51 from different sub-populations into account. When applying the BN model
 52 to describe the amount of genotypic matching observed in a database, it is
 53 not clear that the ‘same sub-population’ assumption is appropriate.

54 In this manuscript, we investigate how empirical genotype matching ob-
 55 servations can be explained by reconsidering the implementation of the BN
 56 model. We show that by accounting for the case of two individuals deriv-
 57 ing from different population groups, we significantly improve the ability to
 58 describe empirical matching rates in a database.

59 **2. Methods**

60 *2.1. Allele sharing matrix*

61 To quantify the degree of multi-locus genotype matching within a data set,
 62 consider the matrix M where each entry $M_{m,p}$ is the number of profile pairs
 63 with m markers matching at both alleles and p markers matching at one allele
 64 [1, 2]. Tvedebrink *et al.* [2] described a recursive algorithm to compute the
 65 probability $\pi_{m,p}$ that two multi-locus genotypes completely match at m loci
 66 and partially match at p loci, constructing a probability matrix π analogous
 67 to M . This method uses the single-locus probabilities of individuals matching
 68 two, one, and zero alleles as $P_{1,0}$, $P_{0,1}$, and $P_{0,0}$, respectively, following in the
 69 notation of Tvedebrink *et al.* [2]. Note the parallel notation to counts of
 70 matching and partially matching markers in $M_{m,p}$. Weir [1] described how
 71 to compute $P_{1,0}$, $P_{0,1}$, and $P_{0,0}$ at a locus by summing over the appropriate
 72 two-individual single locus genotype probabilities [1].

73 *2.2. Single locus allelic sharing probabilities*

74 *2.2.1. Individuals from the same sub-population group*

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

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

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

81 *2.2.2. Individuals from same or different sub-population groups*

82 Under the BN model, if two individuals are not in the same population
 83 group, the probability that their alleles coalesce more recently than a muta-
 84 tion or migration event is zero. In other words, there is no increased chance
 85 of allele-sharing due to shared ancestry for individuals in different population

86 groups. In that case, the probability of observing their genotypes is computed
 87 as a function of the observed allele frequencies without the θ correction.

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

$$\begin{aligned} P(A_{\cdot,\cdot}, A_{\cdot,\cdot}, A_{\cdot,\cdot}, A_{\cdot,\cdot}) &= (1 - d)P(A_{1,\cdot}, A_{1,\cdot}, A_{1,\cdot}, A_{1,\cdot}) \\ &\quad + dP(A_{1,\cdot}, A_{1,\cdot}, A_{2,\cdot}, A_{2,\cdot}) \end{aligned}$$

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

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

100 2.2.3. *Chromosomes from same or different sub-populations*

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

112

113 2.2.4. *Accounting for genetic relatives*

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

120 probability of a particular observed genotype G pair is

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

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

125 2.3. Likelihood framework

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

135 Using the sampling probability of M as a likelihood function, we can es-
136 timate parameters of the model using maximum likelihood. Since the mod-
137 els described here are nested and fulfill standard regularity conditions, the

138 asymptotic distributions of likelihood ratio test statistics (LRTs) are known
 139 to be chi square. Specifically, if we take the null hypothesis to be the typical
 140 implementation of the BN model with a fixed value of θ (say $H_0 : \theta = 0.01$),
 141 and the alternative to be the typical BN model implementation where θ varies
 142 ($H_a : \theta \neq 0.01$), the LRT is distributed as a chi square with one degree of
 143 freedom ($LRT \sim \chi_1^2$). Similarly, to compare the typical implementation of
 144 the BN model with our implementation with population differentiation, we
 145 specify $H_0 : \theta \neq 0.01, d = 0$ and $H_a : \theta \neq 0.01, d \neq 0$, in which case the
 146 $LRT \sim \chi_0^2 + \chi_1^2$. The model allowing chromosomes within individuals from
 147 different sub-populations reduces to the model with population differentia-
 148 tion under complete allelic correlation ($a = 1$). In this case, d is equivalent to
 149 $(k-1)/k$. This enables tests where $LRT \sim \chi_0^2 + \chi_1^2$ between the chromosomal
 150 model and the model with population differentiation.

151 Additionally, we can obtain [maximum likelihood](#) parameter estimates [and](#)
 152 [likelihood profile confidence intervals](#) of $\hat{\theta}$, $\hat{d}$, $\hat{a}$, and $\hat{k}$. While we do not advo-
 153 cate interpreting these estimates too strongly, as the underlying population
 154 models are very simple, we can compare them as a reference.

155 2.4. Database

156 We consider genotype data from 98,988 Brazilian individuals undergoing
 157 paternity testing during 2011-2013 in the Hermes Pardini Laboratory, Ves-
 158 pasiano, MG, Brazil. The individuals genotyped reside in all 26 Brazilian
 159 States and the Federal District (Brasilia). [Although the population genetic](#)
 160 [background of these particular individuals is unknown, generally Brazilian](#)
 161 [populations show ancestry from Indigenous South America, Africa, and Eu-](#)
 162 [rope \[14, 15, 16\]. It is unclear if these ancestries are represented proportion-](#)

ally 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.

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

187 M_{obs} , as shown in Supplemental Table 1. For example, in our dataset, out
 188 of $\binom{90852}{2} = 4,126,997,526$ pairs of genotypes, 295,948 pairs have exactly
 189 one locus matching at both alleles, two loci matching at one allele (partially
 190 matching), and 17 markers matching at neither allele (Supplemental Table
 191 1).

192 3.2. Comparing data likelihood under different models

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

205 Using the maximum likelihood framework and optimizing over θ , we per-
 206 formed a similar analysis (Table 1, Supplemental Table 2). This model where
 207 θ may vary fits the observed data significantly better than with θ fixed at
 208 0.01 ($LRT = 103,621,604$). However, we still observed an excess of individ-
 209 uals sharing few alleles (Figure 2b). Further, under the maximum likelihood
 210 of this model, θ is estimated near zero as $\hat{\theta} = 4.6 \times 10^{-10}$ (in 95% likelihood
 211 profile confidence interval $(0.0, 2.07 \times 10^{-9})$), indicating that the θ correction

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

215 We allow individuals in different sub-populations to share comparatively
216 fewer alleles through common ancestry using the population differentiation
217 model, where two random individuals derive from different sub-population
218 groups with probability d . Again, we find the maximum likelihood value un-
219 der this model (Table 1, Supplemental Table 5). The model fits the observed
220 data significantly better ($LRT = 2,444,098$) and corrects for the previous
221 excess of individuals sharing few alleles (Figure 2c). However, we still see
222 consistent differences between the observed and expected allelic matching.
223 Compared to the observed data, the population differentiation model pre-
224 dicts a more narrow range of allelic matching than what is observed.

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

235

236 3.3. Accounting for relatives

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

247 4. Discussion and Conclusions

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

253 Using this procedure, we found that estimating the value of θ , fits the
254 data significantly better than a uniform value of 0.01. This is intuitive since
255 the typically used $\theta = 0.01$ was not chosen to fit the observed number of co-
256 incidental matches between pairs of profiles in a forensic database. Further,
257 we found that estimate to be near zero. This initially surprising estimate is
258 explained by considering that the common implementation of the BN model
259 in forensic genetics accounts for excess allele sharing due to recent ancestry,

260 but not relatively less allele sharing for individuals with more distant an-
261 cestry. Under this implementation, every pair of individuals has increased
262 allelic sharing due to recent ancestry. Since many pairs of individuals do not
263 share recent ancestry, the maximum likelihood estimate of θ is driven to zero
264 to explain the lack of consistent excess allele sharing.

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

270 We further implemented a model allowing admixture by letting chromo-
271 somes within individuals have different population origins, which fits the data
272 significantly better than the model without admixture ($LRT = 148,835$).
273 However, we caution against over-interpreting the maximum likelihood pa-
274 rameter point estimates of this model, especially considering the sometimes
275 very wide likelihood profile confidence intervals (Supplemental Table 4). De-
276 spite similar log likelihoods, the parameter point estimates differ between
277 the models with and without genetic relatives. This indicates that the flex-
278 ible model accomadates a correlation structure in the data that may not
279 be explicitly described by the parameters, for example more complex popu-
280 lation structures including assymetric migrations or varying sub-population
281 group sizes. Again, this argues against a direct biological interpretation of
282 parameter estimates.

283 In the BN model, all sub-populations have equal excess allele sharing in-
284 ternally and are equally unrelated to each other. While this model provides a

285 simple and reasonable over-estimation of coincidental genotype match rates,
286 essential to forensic case work, it is clearly a simplification of complex hu-
287 man population structures, where some individuals are vastly more related
288 than others. A more sophisticated model allowing varying degrees of allele
289 sharing between individuals would likely better fit our observation of a broad
290 range of allelic-matching. However, such a model would begin to accumulate
291 parameters, making use in forensic case work impractical compared to the
292 adequate typical BN model implementation.

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

304 Like other authors [1, 11, 20, 2], we consider presence of genetic relatives
305 [within a database when calculating genotype match probabilities so that](#)
306 [the total probability of genotype matching takes into account the possibility](#)
307 [of genetic relationships. Our results fail to reject the null hypothesis of](#)
308 [no genetic relatives present.](#) Since the loci are still treated independently,
309 the small probability of a genetic relationship is factored in at each locus

310 separately, rather than considering how genetic relatives share alleles across
311 loci. As a result, unless there are extensive genetic relatives in a dataset, this
312 does not dramatically affect the expected allelic matching.

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

320 Acknowledgments

321 We are immensely grateful to the individuals whose DNA samples were
322 used in this study, without which none of this work would be possible.
323 This work was supported in part by National Institutes of Health grant
324 2R14003229-07, National Science Foundation award 1103767, and a CAPES-
325 Brazil scholarship (Programa Demanda Social and PhD Student Exchange
326 Program BEX 8425/11-6). The funders had no role in study design, data
327 collection and analysis, decision to publish, or preparation of the manuscript.
328 This study was approved by the Ethics Committee of the Federal University
329 of Espirito Santo (Brazil) Health Sciences Department (No. 448327).

330 [1] B. Weir, Matching and partially-matching DNA profiles, *Journal of*
331 *Forensic Science* 49 (2004) 1009–1014.

332 [2] T. Tvedebrink, P. S. Eriksen, J. M. Curran, H. S. Mogensen, N. Mor-

333 ling, Analysis of matches and partial-matches in a danish str data set,
334 Forensic Science International: Genetics 6 (2012) 387–392.

335 [3] J. Curran, T. Tvedebrink, DNAtools: Tools for empirical testing of
336 DNA match probabilities, R package.

337 [4] D. Balding, R. Nichols, DNA profile match probability calculation: How
338 to allow for population stratification, relatedness, database selection and
339 single bands, Forensic Science International 64 (1994) 125–140.

340 [5] D. J. Balding, R. A. Nichols, A method for quantifying differentiation
341 between populations at multi-allelic loci and its implications for investi-
342 gating identity and paternity, in: B. Weir (Ed.), Human Identification:
343 The Use of DNA Markers, Vol. 4 of Contemporary Issues in Genetics
344 and Evolution, Springer Netherlands, 1995, pp. 3–12.

345 [6] P. Gill, I. Evett, Population genetics of short tandem repeat (str) loci, in:
346 B. Weir (Ed.), Human Identification: The Use of DNA Markers, Vol. 4 of
347 Contemporary Issues in Genetics and Evolution, Springer Netherlands,
348 1995, pp. 69–87.

349 [7] I. Evett, P. Gill, J. Scrannage, B. Weir, Establishing the robustness of
350 short-tandem-repeat statistics for forensic applications, The American
351 Journal of Human Genetics 58 (1996) 398–407.

352 [8] I. Evett, P. Gill, J. Lambert, N. Oldroyd, R. Frazier, S. Watson, S. Pan-
353 chal, A. Connolly, C. Kimpton, Statistical analysis of data for three
354 british ethnic groups from a new str multiplex, International Journal of
355 Legal Medicine 110 (1) (1997) 5–9.

- 356 [9] D. Balding, R. Nichols, Significant genetic correlations among Cau-
357 casians at forensic DNA loci, *Human Heredity* 78 (1997) 583–589.
- 358 [10] J. M. Curran, J. S. Buckleton, C. M. Triggs, What is the magnitude of
359 the subpopulation effect?, *Forensic Science International* 135 (2003) 1 –
360 8.
- 361 [11] J. M. Curran, S. J. Walsh, J. Buckleton, Empirical testing of estimated
362 {DNA} frequencies, *Forensic Science International: Genetics* 1 (2007)
363 267 – 272.
- 364 [12] B. Weir, The rarity of DNA profiles, *The Annals of Applied Statistics* 1
365 (2007) 358–370.
- 366 [13] A. M. Adams, R. R. Hudson, Maximum-likelihood estimation of de-
367 mographic parameters using the frequency spectrum of unlinked single-
368 nucleotide polymorphisms, *Genetics* 168 (2004) 1699–1712.
- 369 [14] T. C. Lins, R. G. Vieira, B. S. Abreu, D. Grattapaglia, R. W. Pereira,
370 Genetic composition of Brazilian population samples based on a set of
371 twenty-eight ancestry informative SNPs, *American Journal of Human*
372 *Biology* 22 (2010) 187–192.
- 373 [15] S. D. J. Pena, G. Di Pietro, M. Fuchshuber-Moraes, J. P. Genro, M. H.
374 Hutz, F. d. S. G. Kehdy, F. Kohlrausch, L. A. V. Magno, R. C. Montene-
375 gro, M. O. Moraes, M. E. A. d. Moraes, M. R. d. Moraes, E. B. Ojopi,
376 J. A. Perini, C. Racciopi, A. K. C. Ribeiro-dos Santos, F. Rios-Santos,
377 M. A. Romano-Silva, V. A. Sortica, G. Suarez-Kurtz, The genomic an-

- 378 cestry of individuals from different geographical regions of Brazil is more
379 uniform than expected, PLoS ONE 6 (2011) e17063.
- 380 [16] S. R. Giolo, J. M. P. Soler, S. C. Greenway, M. A. A. Almeida, J. G. S.
381 Mariza de Andrade, C. E. Seidman, J. E. Krieger, A. C. Pereira, Brazil-
382 ian urban population genetic structure reveals a high degree of admix-
383 ture, European Journal of Human Genetics 20 (2012) 111–116.
- 384 [17] V. R. Aguiar, A. M. de Castro, V. C. Almeida, F. S. Malta, A. C.
385 Ferreira, I. D. Louro, New {CODIS} core loci allele frequencies for 96,400
386 brazilian individuals, Forensic Science International: Genetics in press.
- 387 [18] J. Curran, Are DNA profiles as rare as we think? Or can we trust DNA
388 statistics?, Significance 7 (2010) 62–66.
- 389 [19] R. Rohlfs, S. Fullerton, B. Weir, Familial identification: Population
390 structure and relationship distinguishability, PLoS Genetics 8 (2012)
391 e1002469.
- 392 [20] L. Mueller, Can simple population genetic models reconcile partial
393 match frequencies observed in large forensic databases?, Journal of Ge-
394 netics 87 (2008) 101–108.

395 **5. Tables**

396 **6. Figure Legends**

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

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

Figure 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.

Figure 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.

Figure 1

Figure 2a

number of partially matching loci

Figure 2b

number of partially matching loci

Figure 2c

Figure 2d