

Fitting the Balding-Nichols model to forensic databases

Rori V. Rohlf^a, Vitor R. C. Aguiar^c, Kirk E. Lohmueller^b, Amanda M. Castro^c, Alessandro C. S. Ferreira^c, Vanessa C. O. Almeida^c, 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

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 the specific implementation of the Balding-Nichols model must be considered when applied to 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.

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

1 **1. Introduction**

2 Forensic databases, rapidly increasing in size, invite powerful analyses of
3 rates of coincidental genotype matching [1, 2, 3]. Such analyses have vali-
4 dated some basic assumptions in forensic genetics, particularly the reasonable
5 over-estimation of genotype frequencies with existing methods. However,
6 these studies also illustrate how database population genetic diversity dif-
7 fers from what is expected under the basic model of forensic genetics: the
8 Balding-Nichols (BN) model.

9 The BN model simply and elegantly provides a framework for estimating
10 probabilities of observed genotypes, taking into account population structure
11 and variance in allele frequency estimates [4, 5]. The BN model can be inter-
12 preted as describing an ancestral population which has split into a number of
13 internally randomly mating sub-populations which evolve independently over
14 some time, resulting in a present-day total population made up of a num-
15 ber of cryptic sub-population groups. The sampling probabilities estimated
16 under the BN model then incorporate the deviations from Hardy-Weinberg
17 equilibrium expected due to population divergence.

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

25 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 individ-
31 uals in question are from the same sub-population [4]. Balding and Nichols
32 convincingly argue that this assumption is appropriate, saying “the ‘same
33 sub-population’ assumption is conservative, since the suspect’s profile will
34 tend to be more common in his/her sub-population than in other groups”
35 [4]. A number of studies have shown the importance and appropriateness of
36 this assumption and the corresponding θ correction in genetic identification
37 calculations [5, 6, 7, 8, 9, 10, 11, 12].

38 In database applications, typically all pairs of genotypes in a database
39 will be compared to each other and their degree of matching assessed. Previ-
40 ous applications of the standard BN model to forensic databases [1, 2] have
41 shown that the often-used θ correction of 0.01 usually adequately corrects
42 for coincidental allele-sharing, raising estimated probabilities of matching
43 genotypes above their observed levels, and therefore reducing false positive
44 rates below their expectation (in statistical terms, making the test ‘con-
45 servative’). Yet, these analyses show an excess of non-similarity between
46 observed pairs of individuals, as compared to the expectation [1, 2]. As we
47 will show, this is likely due to the fact that the standard formulation of the
48 BN model does not take decreased allele sharing between individuals from
49 different sub-populations into account. When applying the BN model to de-

50 scribe the amount of genotypic matching observed in a database, it is not
51 clear that the ‘same sub-population’ assumption is appropriate.

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

57 **2. Methods**

58 *2.1. Allele sharing matrix*

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

71 *2.2. Single locus allelic sharing probabilities*

72 *2.2.1. Individuals from the same sub-population group*

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

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

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

79 *2.2.2. Individuals from same or different sub-population groups*

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

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

86 We can allow individuals to be from different sub-populations by introduc-
 87 ing a parameter d , which describes the probability that a pair of individuals
 88 are from different sub-population groups. This way, we fully describe the
 89 BN model with some individuals from the same sub-population group and
 90 some from differing groups. Under a model with population differentiation,
 91 two-individual genotype probabilities are

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

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

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

97 *2.2.3. Chromosomes from same or different sub-populations*

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

108 *2.3. Likelihood framework*

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

118 Using the sampling probability of M as a likelihood function, we can es-
119 timate parameters of the model using maximum likelihood. Since the mod-
120 els described here are nested and fulfill standard regularity conditions, the
121 asymptotic distributions of likelihood ratio test statistics (LRTs) are known

122 to be chi square. Specifically, if we take the null hypothesis to be the typical
123 implementation of the BN model with a fixed value of θ (say $H_0 : \theta = 0.01$),
124 and the alternative to be the typical BN model implementation where θ
125 varies ($H_a : \theta \neq 0.01$), the LRT is distributed as a chi square with one degree
126 of freedom ($LRT \chi_1^2$). Similarly, to compare the typical implementation of
127 the BN model with our implementation with population differentiation, we
128 specify $H_0 : \theta \neq 0.01, d = 0$ and $H_a : \theta \neq 0.01, d \neq 0$, in which case the
129 $LRT \chi_0^2 + \chi_1^2$. The model allowing chromosomes within individuals from dif-
130 ferent sub-populations reduces to the model with population differentiation
131 under complete allelic correlation ($a = 1$). In this case, d is equivalent to
132 $(k - 1)/k$. This enables tests where $LRT \chi_0^2 + \chi_1^2$ between the chromosomal
133 model and the model with population differentiation

134 Additionally, we can obtain parameter estimates of $\hat{\theta}$, \hat{d} , \hat{a} , \hat{k} , and \hat{c} .
135 While we do not advocate interpreting these estimates too strongly, as the
136 underlying population models are very simple, we can compare them as a
137 reference.

138 2.4. Database

139 We consider genotype data from 99,275 Brazilian individuals undergo-
140 ing paternity testing during 2011-2013 in the Hermes Pardini Laboratory,
141 Vespasiano, MG, Brazil. The individuals genotyped reside in all 26 Brazil-
142 ian States and the Federal District (Brasilia). The genotypes were obtained
143 using a combination of two Life Technologies kits and ABI 3730 Genetic an-
144 alyzers (Life Technologies, CA, USA) for a total of 20 loci (the original 13
145 CODIS core loci and additionally D10S1248, D22S1045, D1S1656, D12S391,
146 D2S441, D2S1338, D19S433, PentaD, and PentaE) [14].

147 While there are no known relatives in this dataset, unknown relatives, or
148 multiple entries of the same individual are expected. As such, individuals
149 with 17-20 loci matching and the same birth dates (when available) were
150 removed as likely multiple entries or identical twins with some genotyping or
151 clerical errors. When dates were not available or inconsistent apparently due
152 to a typo, names were manually checked by the lab personnel and the most
153 complete profile was kept, resulting in a data set with 96,400 individuals [14].

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

160 **3. Results**

161 *3.1. Observed database matching*

162 We counted the number of zero, one, and two allele matches for each
163 locus for each pair of individuals in the dataset to create the observed matrix
164 M_{obs} , as shown in Supplemental Table 1. For example, in our dataset, out
165 of $\binom{90852}{2} = 4,126,997,526$ pairs of genotypes, 295,948 pairs have exactly
166 one loci matching at both alleles, two loci matching at one allele (partially
167 matching), and 17 markers matching at neither allele (Supplemental Table
168 1).

169 *3.2. Comparing data likelihood under different models*

170 Previous investigators have used the conventional implementation of the
171 BN model (without population differentiation) with θ fixed at 0.01 to describe
172 matching in databases [1, 2]. Under this model, setting $\theta = 0.01$, we calcu-
173 lated the log likelihood of the observed match matrix as $-47,246,554,388$
174 (Table 1). We can graphically compare our observed and expected results in
175 a dropping ball diagram [11, 15, 2] (Figure 1), or in a heatmap of the resid-
176 uals (Figure 2a). The heatmaps in this manuscript show a color gradient
177 along the log of the divergence of the observed and expected as $\log \frac{(obs-exp)^2}{exp}$.
178 Through these visualizations, we see that as in previous analyses [1, 2], under
179 the typical BN model implementation with θ set at 0.01, there is an excess
180 of observed pairs of individuals who share few alleles, as compared to the
181 expectation.

182 Using the maximum likelihood framework and optimizing over θ , we per-
183 formed a similar analysis (Table 1, Supplemental Table 2). This model where
184 θ may vary fits the observed data significantly better than with θ fixed at 0.01
185 ($LRT = 94,490,590,210$). However, we still observed an excess of individu-
186 als sharing few alleles (Figure 2b). Further, under the maximum likelihood
187 of this model, θ is estimated near zero as $\hat{\theta} = 4.6e - 10$, indicating that the
188 θ correction as implemented here is insufficient to describe the data. This
189 makes sense since the θ correction accounts for excess allelic sharing due to
190 common ancestry within a sub-population.

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

194 groups with probability d . Again, we find the maximum likelihood value un-
195 der this model (Table 1, Supplemental Table 3). The model fits the observed
196 data significantly better ($LRT = 2,444,098$) and corrects for the previous
197 excess of individuals sharing few alleles (Figure 2c). However, we still see
198 consistent differences between the observed and expected allelic matching.
199 Compared to the observed data, the population differentiation model pre-
200 dicta a more narrow range of allelic matching than what is observed.

201 In the population differentiation model, it is assumed that both alleles
202 within an individual derive from the same sub-population. This assumption
203 may not be valid in realistic cases of admixture, and can be relaxed using
204 the a model where chromosomes are considered separately with some corre-
205 lation. We fit such a model with k equally represented sub-populations and
206 intra-individual allelic correlation to the observed data. This model, allowing
207 chromosomes of different population origins within individuals, fits the data
208 significantly better than the model without admixture ($LRT = 148835$) (Fig-
209 ure 2d). Still, we observe a wider range of allelic matching than is expected
210 under these models.

211 4. Discussion and Conclusions

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

217 Using this procedure, we found that estimating the value of θ , unsurpris-

218 ingly, fits the data significantly better than a uniform value of 0.01. Further,
219 we found that estimate to be near zero. This initially surprising estimate is
220 explained by considering that the common implementation of the BN model
221 in forensic genetics accounts for excess allele sharing due to recent ancestry,
222 but not relatively less allele sharing for individuals with more distant an-
223 cestry. Under this implementation, every pair of individuals has increased
224 allelic sharing due to recent ancestry. Since many pairs of individuals do not
225 share recent ancestry, the maximum likelihood estimate of θ is driven to zero
226 to explain the lack of consistent excess allele sharing.

227 We show and implement several parameterizations of the full BN model
228 where individuals may or may not have excess allele sharing (equivalently,
229 may or may not derive from the same population group). This full BN model
230 fit the observed match matrix significantly better than the typical BN model
231 implementation. Under the full BN model, $\hat{\theta} = 0.008$, which is closer to the
232 typical value used in practice of $\theta = 0.01$.

233 Even under the full BN model implementation, we predict a more narrow
234 range of locus-matching than observed. In the BN model, all sub-populations
235 have equal excess allele sharing internally and are equally unrelated to each
236 other. While this model provides a simple and reasonable over-estimation
237 of coincidental genotype match rates, essential to forensic case work, it is
238 clearly a simplification of complex human population structures, where some
239 individuals are vastly more related than others. A more sophisticated model
240 allowing varying degrees of allele sharing between individuals would likely
241 better fit our observation of a broad range of allelic-matching. However,
242 such a model would begin to accumulate parameters, making use in forensic

243 case work impractical compared to the adequate typical BN model imple-
244 mentation.

245 Additionally, the full BN model does not explain a small observed excess
246 of people matching at many loci. For example, there are three pairs of
247 individuals who match both alleles at 13 loci and one allele at six loci, whereas
248 under the full BN model, $5.0e - 13$ are expected. There are several possible
249 explanations for these individuals. They may be genetic relatives who share a
250 large number of alleles IBD. They could share even more alleles than expected
251 if allele frequencies are mis-specified because they derive from a population
252 group divergent from the whole sample [16]. It is also possible that the same
253 individual was entered a number of times, with genotyping or clerical errors
254 resulting in differing alleles.

255 Other authors have considered the presence of genetic relatives within
256 a database when calculating genotype match probabilities with additional
257 parameters for the probabilities that a pair of individuals hold particular
258 genetic relationships [1, 17, 2]. This way, the total probability of genotype
259 matching takes into account the possibility of genetic relationships. However,
260 since the loci are still treated independently, the small probability of a genetic
261 relationship is factored in at each locus separately, rather than considering
262 how genetic relatives share alleles across loci. As a result, unless there are
263 extensive genetic relatives in a dataset, this does not dramatically affect the
264 expected allelic matching.

265 We have shown how the correct full implementation of the BN model
266 is crucial to understanding database-wide allelic matching. While this is
267 essential for database applications, it does not affect forensic case work where

268 the typical BN model implementation is adequate to reasonably overestimate
269 the probability of coincidental genotype matching.

270 **Acknowledgements**

271 We are immensely grateful to the individuals whose DNA samples were
272 used in this study, without which none of this work would be possible.
273 This work was supported in part by National Institutes of Health grant ???,
274 National Science Foundation award 1103767, and a CAPES-Brazil scholar-
275 ship (Programa Demanda Social and PhD Student Exchange Program BEX
276 8425/11-6) ??MORE??. The funders had no role in study design, data col-
277 lection and analysis, decision to publish, or preparation of the manuscript.
278 This study was approved by the Ethics Committee of the Federal University
279 of Espirito Santo (Brazil) Health Sciences Department (No. 448327).

- 280 [1] B. Weir, Matching and partially-matching DNA profiles, *Journal of*
281 *Forensic Science* 49 (2004) 1009–1014.
- 282 [2] T. Tvedebrink, P. S. Eriksen, J. M. Curran, H. S. Mogensen, N. Mor-
283 ling, Analysis of matches and partial-matches in a danish str data set,
284 *Forensic Science International: Genetics* 6 (2012) 387–392.
- 285 [3] J. Curran, T. Tvedebrink, DNAtools: Tools for empirical testing of
286 DNA match probabilities, R package.
- 287 [4] D. Balding, R. Nichols, DNA profile match probability calculation: How
288 to allow for population stratification, relatedness, database selection and
289 single bands, *Forsensic Science International* 64 (1994) 125–140.

- 290 [5] D. J. Balding, R. A. Nichols, A method for quantifying differentiation
291 between populations at multi-allelic loci and its implications for investi-
292 gating identity and paternity, in: B. Weir (Ed.), *Human Identification:
293 The Use of DNA Markers*, Vol. 4 of *Contemporary Issues in Genetics
294 and Evolution*, Springer Netherlands, 1995, pp. 3–12.
- 295 [6] P. Gill, I. Evett, Population genetics of short tandem repeat (str) loci, in:
296 B. Weir (Ed.), *Human Identification: The Use of DNA Markers*, Vol. 4 of
297 *Contemporary Issues in Genetics and Evolution*, Springer Netherlands,
298 1995, pp. 69–87.
- 299 [7] I. Evett, P. Gill, J. Scranage, B. Weir, Establishing the robustness of
300 short-tandem-repeat statistics for forensic applications, *The American
301 Journal of Human Genetics* 58 (1996) 398–407.
- 302 [8] I. Evett, P. Gill, J. Lambert, N. Oldroyd, R. Frazier, S. Watson, S. Pan-
303 chal, A. Connolly, C. Kimpton, Statistical analysis of data for three
304 british ethnic groups from a new str multiplex, *International Journal of
305 Legal Medicine* 110 (1) (1997) 5–9.
- 306 [9] D. Balding, R. Nichols, Significant genetic correlations among Cau-
307 casians at forensic DNA loci, *Human Heredity* 78 (1997) 583–589.
- 308 [10] J. M. Curran, J. S. Buckleton, C. M. Triggs, What is the magnitude of
309 the subpopulation effect?, *Forensic Science International* 135 (2003) 1 –
310 8.
- 311 [11] J. M. Curran, S. J. Walsh, J. Buckleton, Empirical testing of estimated

- 312 {DNA} frequencies, *Forensic Science International: Genetics* 1 (2007)
313 267 – 272.
- 314 [12] B. Weir, The rarity of DNA profiles, *The Annals of Applied Statistics* 1
315 (2007) 358–370.
- 316 [13] A. M. Adams, R. R. Hudson, Maximum-likelihood estimation of de-
317 mographic parameters using the frequency spectrum of unlinked single-
318 nucleotide polymorphisms, *Genetics* 168 (2004) 1699–1712.
- 319 [14] V. R. Aguiar, A. M. de Castro, V. C. Almeida, F. S. Malta, A. C.
320 Ferreira, I. D. Louro, New {CODIS} core loci allele frequencies for 96,400
321 brazilian individuals, *Forensic Science International: Genetics* in press.
- 322 [15] J. Curran, Are dna profiles as rare as we think? or can we trust dna
323 statistics?, *Significance* 7 (2010) 62–66.
- 324 [16] R. Rohlf, S. Fullerton, B. Weir, Familial identification: Population
325 structure and relationship distinguishability, *PLoS Genetics* 8 (2012)
326 e1002469.
- 327 [17] L. Mueller, Can simple population genetic models reconcile partial
328 match frequencies observed in large forensic databases?, *Journal of Ge-
329 netics* 87 (2008) 101–108.

330 **5. Tables**

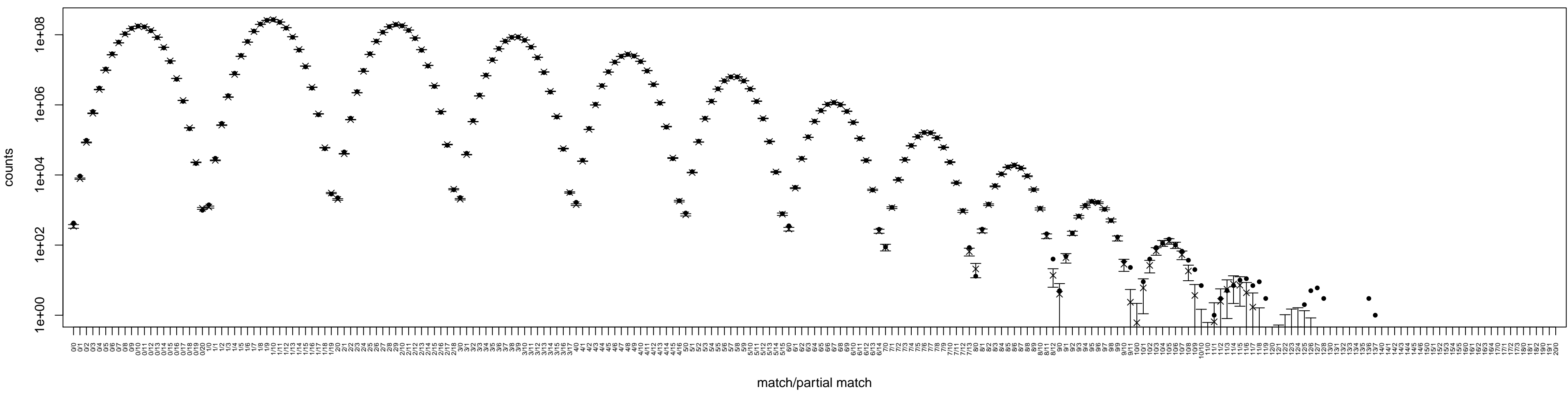
331 **6. Figure Legends**

model	log likelihood	$\hat{\theta}$	\hat{d}	\hat{a}	\hat{k}
typical implementation with $\theta = 0.01$	-47246554388	NA	NA	NA	NA
typical implementation with θ varying	-1259283	$4.6e - 10$	NA	NA	NA
population differentiation implementation	-37234	$7.6e - 03$	0.966	NA	NA
sub-population groups by chromosome	-37182	$7.9e - 03$	NA	0.939	3538.5

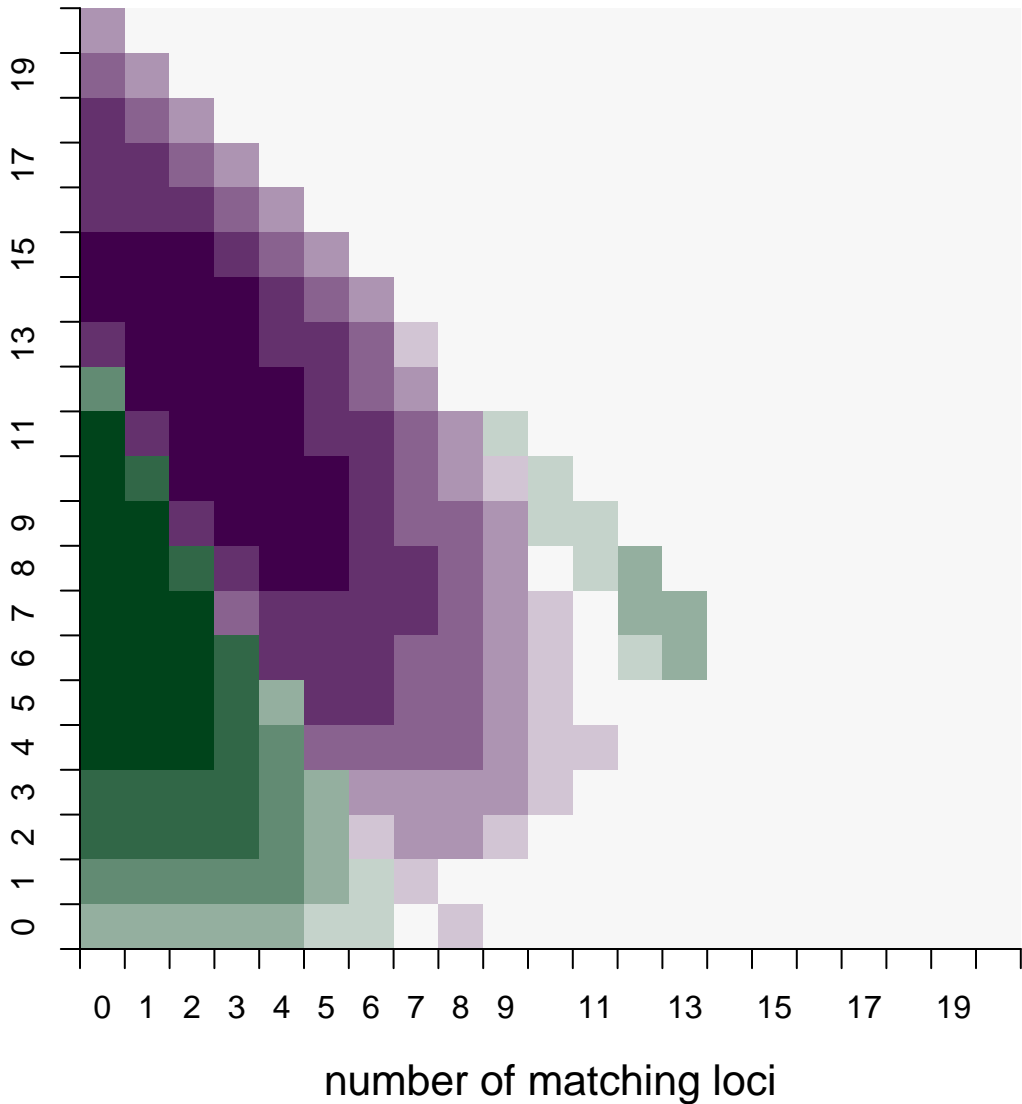
Table 1: The maximum log likelihoods and parameter estimates are listed for each model considered. For models which do not incorporate particular parameters, those estimates are listed as NA.

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.

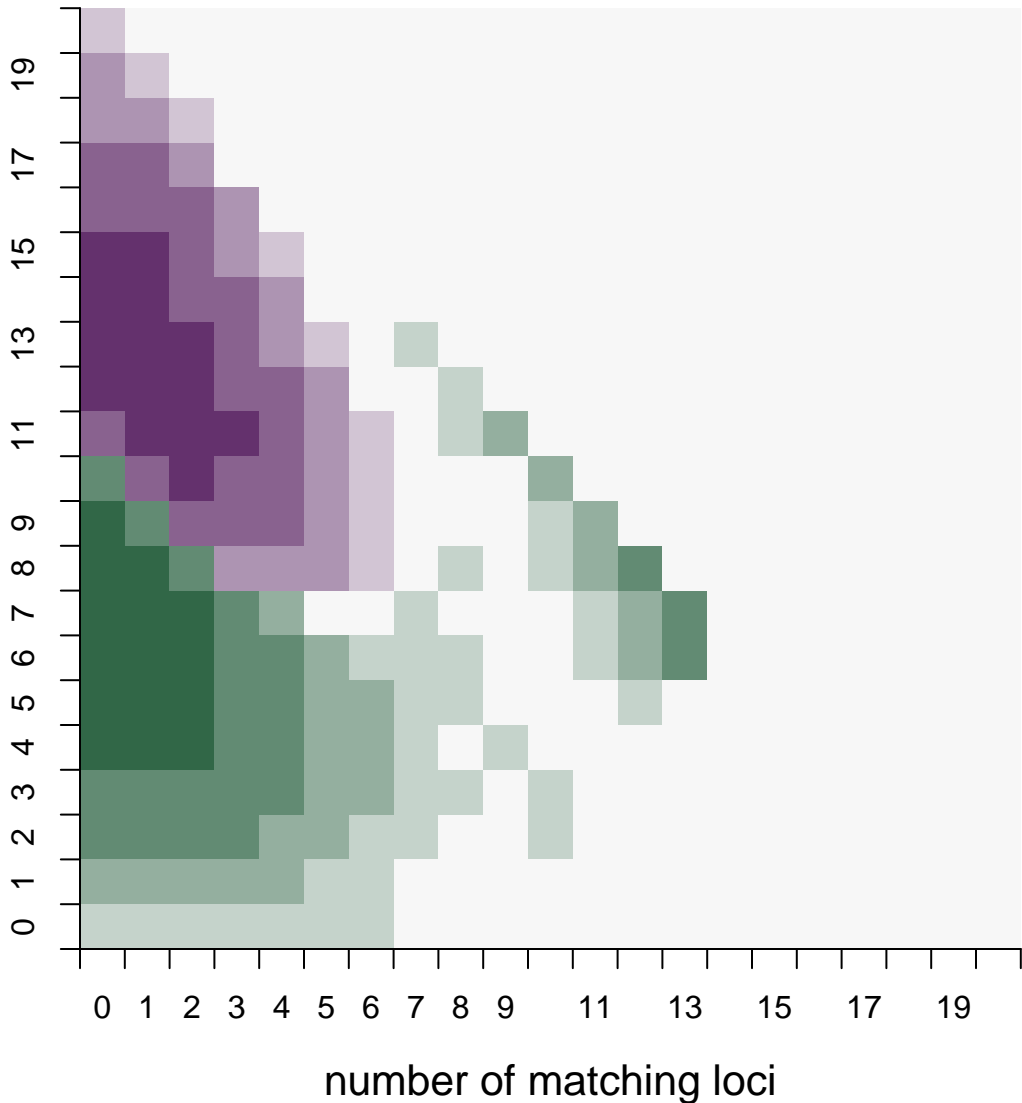
Figure 2: These heatmaps show the difference between the observed match matrix and that expected under, (a) the typical implementation of the BN model with $\theta = 0.01$, (b) the typical implementation of the BN model where θ varies, (c) the full implementation of the BN model, and (d) the full implementation 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.



number of partially matching loci



number of partially matching loci

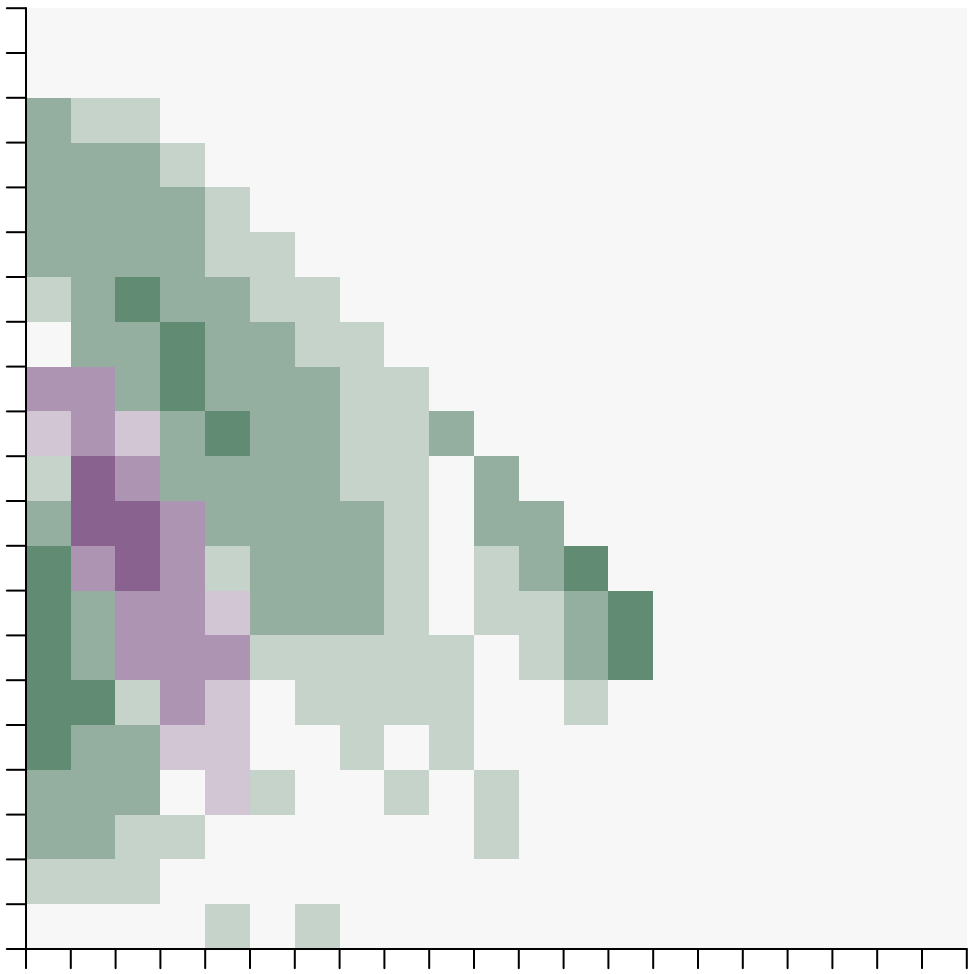


number of partially matching loci

19
17
15
13
11
9
8
7
6
5
4
3
2
1
0

0 1 2 3 4 5 6 7 8 9 11 13 15 17 19

number of matching loci



number of partially matching loci

19
17
15
13
11
9
8
7
6
5
4
3
2
1
0

0 1 2 3 4 5 6 7 8 9 11 13 15 17 19

number of matching loci

