Int J Legal Med (2010) 124:63–69
DOI 10.1007/s00414-009-0359-2

ORIGINAL ARTICLE

Population stratification in Argentina strongly influences likelihood ratio estimates in paternity testing as revealed by a simulation-based approach

Ulises Toscanini · Antonio Salas · Manuel García-Magariños · Leonor Gusmão · Eduardo Raimondi

Received: 9 December 2008 / Accepted: 26 May 2009 / Published online: 20 June 2009

Abstract A simulation-based analysis was carried out to investigate the potential effects of population substructure in paternity testing in Argentina. The study was performed by evaluating paternity indexes (*PI*) calculated from different simulated pedigree scenarios and using 15 autosomal short tandem repeats (STRs) from eight Argentinean databases. The results show important statistically significant differences between *PI* values depending on the dataset employed. These differences are more dramatic when considering Native American versus urban populations. This study also indicates that the use of *Fst* to correct for the effect of population stratification on *PI* might be inappropriate because it cannot account for the particularities of single paternity cases.

Keywords Argentina · Population stratification · Autosomal STRs · Paternity index

Ulises Toscanini and Antonio Salas contributed equally to this work

U. Toscanini · E. Raimondi (✉)
PRICAI-FUNDACIÓN FAVALORO,
Av. Belgrano 1782-1er Subsuelo, (1093),
Capital Federal-Buenos Aires, Argentina
e-mail: eraimondi@ffavaloro.org

L. Gusmão
IPATIMUP, Instituto de Patologia e Imunologia Molecular
da Universidade do Porto,
Porto, Portugal

A. Salas · M. García-Magariños
Unidade de Xenética, Instituto de Medicina Legal
and Departamento de Anatomía Patolóxica y Ciencias Forenses,
Facultade de Medicina, Universidade de Santiago de Compostela,
Galicia, Spain

Introduction

Historically, non-exclusion in paternity testing was statistically evaluated by means of probability of paternity according to the Essen–Möller formula [1, 2]. Later, the use of the ratio between the probability of the hypothesis of paternity (*X*) and non-paternity (*Y*), with the form *X/Y*, was proposed [3] and this ratio, called the paternity index (*PI*), was considered to be sufficiently appropriate to report a result [4]. Recently, the Paternity Testing Commission of the International Society for Forensic Genetics (ISFG; http://www.isfg.org) has issued a series of recommendations on biostatistics [5, 6] suggesting that the biological evidence should be based on likelihood ratio principles.

Calculation of *PI* requires knowing the allele frequency distributions in the reference population. Caution must be taken when population substructure exists, so that appropriate corrections on *PI* values can be applied [7]. The use of *Fst* to measure (and correct for) the effect of substructure in reference populations is commonly used in forensic genetics [7]. *Fst* measures population differentiation based on allele frequencies. However, in routine casework, one case is generally evaluated at a time and global patterns of variability in the population do not necessarily represent the idiosyncrasies of particular cases and genetic profiles, in the same way as for haploid data [8–10]. Therefore, the use of *Fst* to account for population stratification does not always correctly adjust the *PI* values in every single case.

It is well documented that in Argentina differences exist between allele frequency distributions in populations, for common genetic markers used in forensics, that can have important consequences in routine forensic

casework [11, 12]. This view, however, is controversial since other authors claim that population differences within the country are irrelevant in this context [13]. Recently, we used a simulation-based approach to show that these differences actually have implications in the computation of likelihood ratios in forensic casework [12]. The goal of the present study was to determine the impact of the population substructure on paternity testing, using a different simulation-based approach that compares the results obtained when using different datasets for the computation of *PI* values in several pedigree scenarios. Some analytical expressions can be obtained in order to address these problems [14] in a general population context; these other approaches aim generally to investigate the expected average effect of using different levels of population stratification and mutation rates in hypothesized situations (e.g. artificially created populations). The study by Karlsson et al. [15] described a very interesting approach related to the evaluation of the risk of erroneous conclusions on DNA testing for immigration cases. The aim of the present study was, however, to exactly measure the real impact of using different datasets from Argentina on particular *PI* values by simulating paternity cases that could be real in this country, and given the fact that it is a particular *PI* value that is generally communicated to the courtroom. Therefore, the most theoretical general approaches, although necessary in science, do not help by definition to evaluate singular forensic cases where particular individuals are being judged. On the other hand, the present approach has the advantage that cases where parents come from different populations can easily be handled by sampling from different databases.

Materials and methods

Population samples and genotyping data

The study was based on 1,906 genotypes belonging to individuals of six urban populations from Buenos Aires (N=879), Neuquén (N=355), Tucumán (N=75), San Luis (N=61), Santa Cruz (N=132), and La Pampa (N=232) and two Native American populations from Colla (N=43) and Toba (N=129) in Argentina.

The genotype data consisted of a set of 15 autosomal STRs from the Powerplex® 16 System kit (Promega, Madison, WI, USA): D3S1358, HUMTH01, D21S11, D18S51, PENTA E, D5S818, D13S317, D7S820, D16S539, CSF1PO, PENTA D, HUMvWA, D8S1179, HUMTPOX, and FGA. No deviations from Hardy–Weinberg equilibrium were detected in any of these population samples.

Data simulation

Data simulation involved the following steps:

1. *Generation of artificial profiles.* For each of the 1,906 real profiles in the database, a set of new profiles was created by a computer-assisted procedure. First, allele frequencies were obtained for all the original datasets. Second, compatible profiles for both parents of each individual were built as follows: each of the two alleles was randomly assigned to each parent then the other allele of each parent was randomly taken from a vector of allele population frequencies of each STR locus.

Parents' sets were tested for Hardy–Weinberg equilibrium and no departures were observed.

2. *Definition of pedigrees to calculate the PI.* With the individuals generated as described in step 1, we constructed two different types of pedigrees: alleged father–mother–child (trio) and alleged father–child (duo).
3. *Frequency databases.* A total of 50 different allelic frequency matrices were built from each population sample constructed by selecting at random 80% of the individuals of the original datasets. This bootstrap-based approach aim to control for the variability involved in the estimation of allele frequencies due, for instance, to differences in samples size.
4. *PI calculation. PI* values were calculated by contrasting two mutually exclusive hypotheses in trios and duos: (1) the alleged father is the true father of the child and (2) the father is an unrelated individual.

*PI*s for all the pedigrees in one population were calculated with the corresponding reference database, and also using the databases from the seven other populations. Since 50 different frequency matrices were available for each population, each pedigree yielded 50 *PI*s for each population database. For each population database, the mean *PI* value was also calculated for every single pedigree.

Statistical analyses

As explained, for each individual (N=1,906) a set of 50 *PI* values were obtained using each of the eight datasets. Three goodness-of-fit tests were employed in order to examine if each set of 50 *PI* values fit with normality, namely Kolmogorov–Smirnov, Shapiro–Wilks and Pearson's χ^2 (see e.g. [16]). The normality assumption was rejected in most of the situations, even for the most conservative test of Kolmogorov–Smirnov. Therefore, all the *PI* values were converted into logarithms and the normality was checked again using the same goodness-of-fit tests. The normality

assumption (required to properly carry out the statistical tests below) could then be accepted for the logarithm of the *PI* values ($\log_{PI}$).

Next, for each individual an ANOVA analysis was carried out between the eight sets of 50 $\log_{PI}$. ANOVA allowed testing significant differences between the $\log_{PI}$ values obtained when using the different datasets. Due to the fact that the null hypothesis of equality between sets $\log_{PI}$ was always rejected, we next used four different statistical tests (namely Tukey, LSD Fisher, Duncan Ranks, and Newman; see e.g. [16]), in order to explore statistical differences between all pairwise comparisons involving the 1,906 profiles.

The decision to use several tests for testing normality and several post hoc tests was based on two facts: (a) the need for testing inconsistencies when using different statistical approaches that could reveal, for instance, some technical or conceptual problem in the design of the simulations and (b) select the test providing the most conservative results. Bonferroni's adjustment was used in order to account for multiple test corrections and setting the nominal significant value α to 0.01.

Additionally, for each profile we computed an ad hoc index, the weighted mean difference (WMD) between pairs of populations that quantifies the magnitude of the differences between pairs of *PI* values. This index is defined here as follows:for each pair of populations *i*,*j*,

$$\mathrm{WMD} = \frac{\bar{PI}_i - \bar{PI}_j}{\max\left(\bar{PI}_i, \bar{PI}_j\right)}$$

where $\bar{PI}$ indicates the mean value for the set of 50 *PI*s obtained of each individual in each dataset.

Double checking the results

All the simulations and statistical analysis were carried out using Visual Basic programming in Microsoft Excel and the freely available statistically package R (http://www.r-project.org/). A random subset of the pedigrees was selected from the original pedigree simulations, and the accuracy of the results was double checked by using the shareware software Familias v. 1.81, http://www.math.chalmers.se/~mostad/familias/ [17].

Rationale

The aim of the statistical analysis was to evaluate the impact on *PI*s using a single regional database for every forensic case in the country compared to using a regional database. In fact it is common for example that a laboratory in Buenos Aires receives paternity cases from all over the country. If all the populations in Argentina were homogeneous no significant differences would be observed on *PI*s; on the contrary, if population substructure exists, we would expect to find important differences depending on the database employed. The latter would involve the need to develop local frequency tables representing the main regions from the country instead of using a global one.

One could envisage another simpler potential solution to the problem, i.e. to build a global database of the country and use it as reference population for any paternity test carried out in the territory. However, as demonstrated below, the differences in *PI* values when using different datasets can be dramatic, and so, the use of a single database would just aggravate the problem; e.g. if one has a case from Buenos Aires, it will be more appropriate to use the Buenos Aires database than a global one. Similar problems were addressed from a theoretical point of view by Ayres [18].

The whole simulation algorithm employed in the present study is summarized in the scheme of Fig. 1.

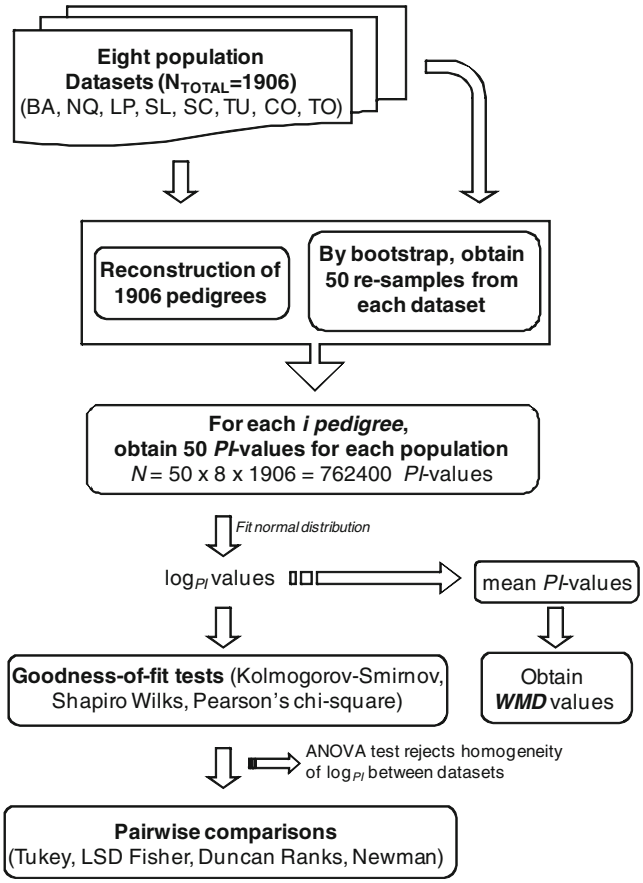


Fig. 1 Scheme showing the main steps considered in the simulation and statistical analysis carried out in the present study

Results and discussion

PI values vary significantly depending on the reference population

Several statistical tests were used to measure the percentage of pedigrees from which the *PI* values statistically differ when using different population datasets. For instance, the Tukey test (Table 1) indicates that most of the times the $\log_{PI}$ values differ significantly between populations (e.g. using a nominal value of α=0.01 coupled with a Bonferroni's correction assuming 1,906 comparisons). As expected, the largest percentages of statistically significant *PI* differences almost always involved comparisons between the two Native American populations versus the other datasets. The largest differences occurred between these two Native American populations.

The other statistical tests employed yielded less conservative results than Tukey (data not shown), since the Tukey test internally controls for global error type I (given the 28 comparisons carried out each time). The different statistical tests are, however, consistent in showing the number of percentages of *PI* values statistically significant as showed by a Mantel test; for instance, in trios, $r^2>0.997$ and $p<0.001$ (Pearson's correlation, 10,000 permutation tests) for all the comparisons (Tukey versus LSD of Fisher, Tukey versus Duncan Ranks, Tukey versus Newman).

The distribution of $-\log_{10}(P$ values) obtained using the Tukey test are shown in Table 2 (below the diagonal), for trios and duos. The most outstanding feature of these figures is that the slopes of the distributions are more pronounced in those comparisons involving more distant populations (see also [19]); for instance, those involving Native Americans. Also remarkable is the large number of $-\log_{10}(P$ values) that falls below the most conservative Bonferroni's correction.

Measuring inter-population differences in *PI* values

The main aim of the present analytical approach is to evaluate the magnitude of the differences in *PI* values and to what extent statistical significances between populations have an impact in substantial *PI* differences that could be relevant for decisions in court.

WMD values were computed for each individual profile. These values measure the magnitude of the difference between every single pair of mean *PI* values among populations. For instance, a WMD value of 0.7 indicates that the difference between the two mean values considered is 70% of the absolute value of the largest mean. Therefore, high WMD values indicate large differences between populations and vice versa.

Tables 2 and 3 (above the diagonal) for trios and duos respectively, show the distributions of WMD values between pairs of datasets. Note again that the two Native American populations show the most skewed distributions towards high WMD values. In particular Toba is more distinct than Colla with respect to the other populations. In general, the histograms of Tables 2 and 3 indicate large differences between *PI* values independently of the population dataset used. Table 1 (data below the diagonals for trios and duos) indicates the percentage of WMD values

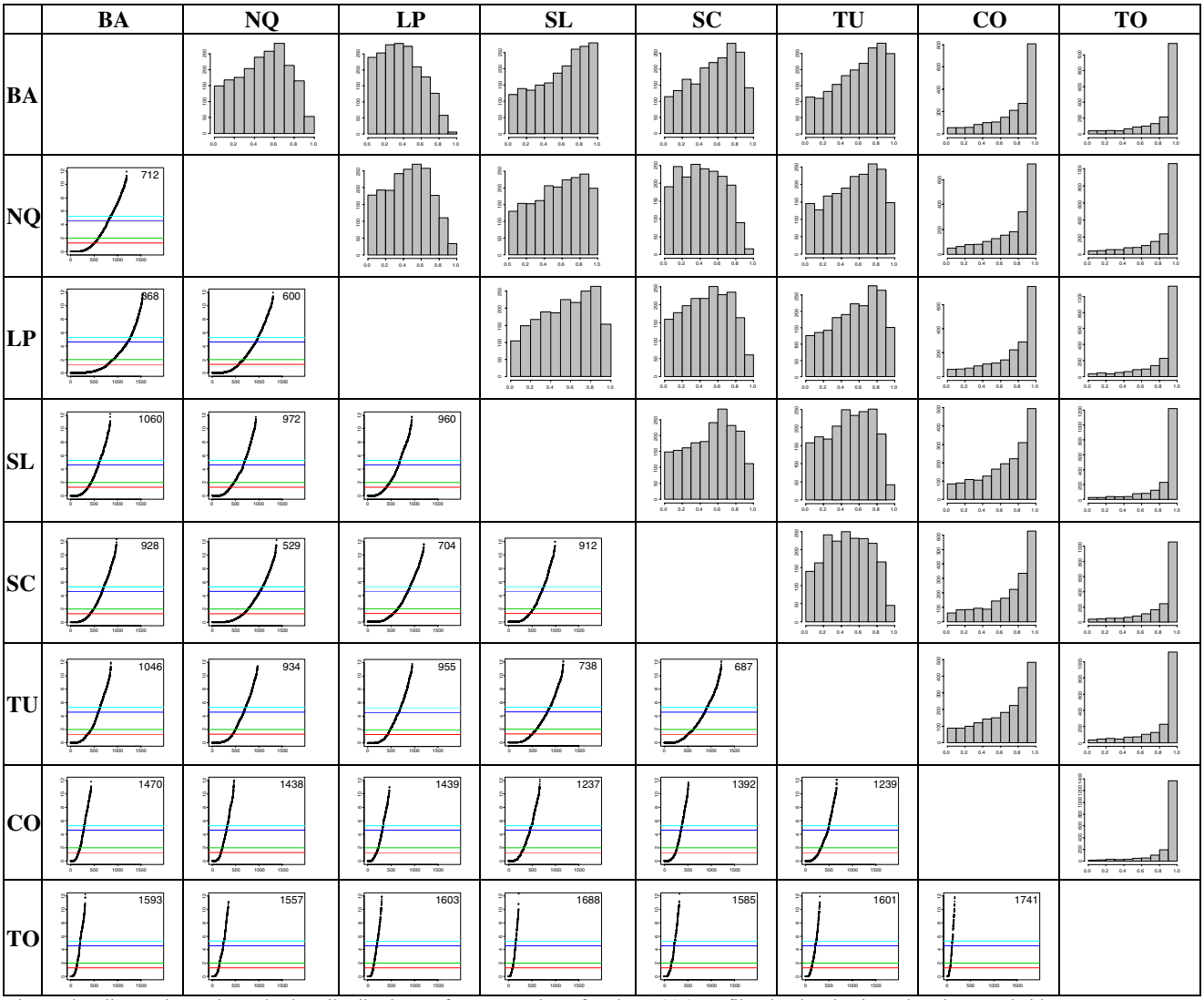
Table 1 Significant difference between populations

	BA	NQ	LP	SL	SC	TU	CO	TO
TRIOS								
BA	–	69.1/55	52.1/32.9	76.9/67.3	74.1/62.9	76.8/67.2	89.3/84.7	92.5/89.2
NQ	11.4	–	64.7/48.4	75.3/62.7	59.5/43.2	73.7/62.6	88.4/82.9	91.8/86.8
LP	3.4	7.5	–	74.9/63.7	65.9/52.3	74.8/63.2	88.0/82.6	92.8/89.0
SL	28.8	23.1	21.9	–	71.9/60.7	68.6/53.7	82.8/74.3	94.3/91.8
SC	20.7	5.6	11.8	17.1	–	67.4/52.1	85.7/80.4	92.0/88.1
TU	27.9	20.6	21.8	11.8	11.1	–	82.6/73.8	92.3/88.0
CO	56.7	56.0	54.8	42.2	50.6	42.7	–	95.6/94.0
TO	71.5	68.2	71.4	76.0	68.2	71.0	82.2	–
DUOS								
BA		67.9/54.6	51.7/34.3	76.8/66.3	74.2/63.1	76.9/67.5	86.7/81.2	93.3/90.2
NQ	11.5		65.2/48.2	73.2/61.2	60.6/43.3	73.3/61.5	86.7/79.1	93.1/88.8
LP	3.0	7.5		74.6/62.3	66.6/53.8	74.9/63.5	86.7/80.2	93.3/90.5
SL	28.6	21.7	20.4		71.5/59.4	67.5/53.1	82.2/72.8	94.8/92.3
SC	22.8	5.5	12.9	16.5		68.0/52.9	85.0/78.3	93.6/90.2
TU	29.7	20.4	23.2	12.7	10.5		80.2/72.9	94.2/91.2
CO	56.3	52.0	52.8	41.1	47.0	40.9		96.4/95.0
TO	74.3	72.7	75.5	78.6	73.6	75.1	84.3	

The upper diagonals values indicate the percentages of individuals that show significant differences in pairwise comparisons under the test of Tukey for trios (upper) and duos (bottom); the first term is for a α=0.01, while the second term is for the Bonferroni's correction assuming 1,906 comparisons. The below diagonals show the percentages of WMD values above 0.8

Population codes: *BA* Buenos Aires, *NQ* Neuquén, *LP* La Pampa, *SL* San Luis, *SC* Santa Cruz, *TU* Tucumán, *CO* Collas, *TO* Tobas

Table 2 WMD values for the 1,906 profiles in the database in trios



Above the diagonal are the pairwise distributions of WMD values for the 1,906 profiles in the database in trios. Each histogram represents therefore the impact on WMD for a given pair of frequency datasets over the 1,906. Below the diagonal are the distributions of $-\log_{10}(P$ values) for the test of Tukey; the horizontal lines represent from bottom to top the $\log_{10}$ values for $\alpha=0.05$, $\alpha=0.01$, and the respective values assuming Bonferroni corrections; the numbers in the top-right corner of these distributions pictures indicate the number of tests that fall out of the distribution and that in general correspond to values close to zero.

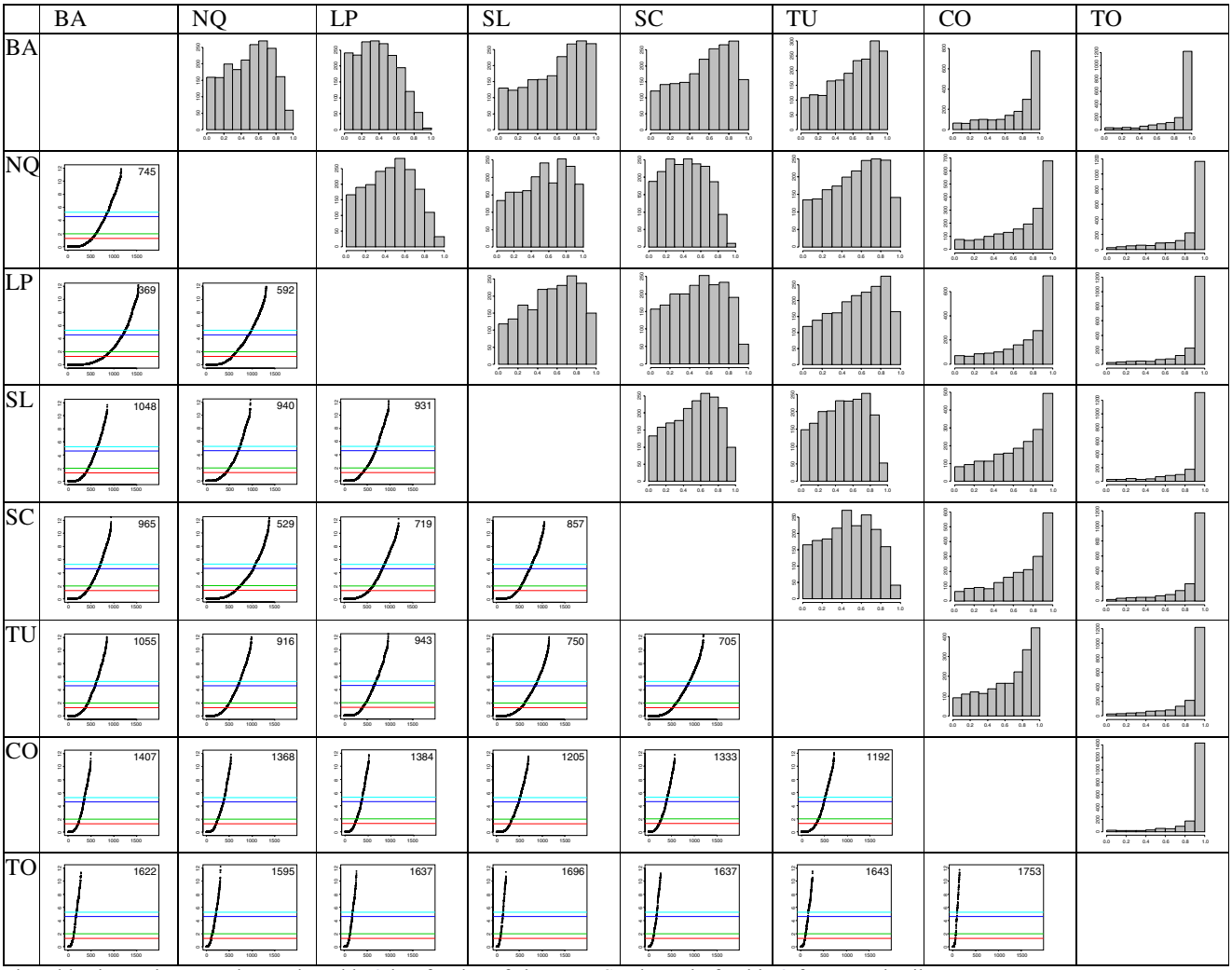
Population codes: *BA* Buenos Aires, *NQ* Neuquén, *LP* La Pampa, *SL* San Luis, *SC* Santa Cruz, *TU* Tucumán, *CO* Collas, *TO* Tobas

above 0.8. Note that these values correspond with the two last bars of the histograms presented in Tables 2 and 3 (data above the diagonal).

Reviewing previous finding concerning population substructure in Argentina

The importance of population substructure in Argentina has been minimized in previous studies [20–23]. More recently, Marino et al. [13] measured the impact of population substructure in Argentina, analyzing 15 autosomal STRs in ten population samples from the country, and concluded that no substructure could be detected supporting that a single database of the whole country could be suitable for the correct interpretation of paternity testing and forensic casework results. Nevertheless, they found a clear statistical differentiation between the Salta population sample and the rest of the population samples analyzed, which contradict their final conclusion about the possibility of using a unique database for the whole country. Moreover, our previous findings [11] revealed the existence of population substructure in Argentina at autosomal STR level. In addition, population stratification is also supported when looking at the population patterns of Y-STR [23, 24] and mitochondrial

Table 3 WMD values for the 1,906 profiles in the database in duos



The table shows the same data as in Table 2 but for duos father–son. See legend of Table 2 for more details.

DNA data, as can be inferred from the few studies carried out in populations from this country [25–28].

In the present study, we have employed exactly the same autosomal marker set used by Marino et al. [13] but our results and conclusions differ substantially. The main reason is that the statistical approaches employed in these studies are conceptually different. While Marino et al. [13] employed *Fst* genetic distances to detect and quantify genetic stratification, our approach aimed to measure the effect of population substructure directly on *PI* values. We here demonstrated that *Fst* corrections might not account for the singularities of the full universe of genetic profiles in a population. Thus, for instance, considering trios, ~22% of the *PI* values of the Toba's profiles differs more than three orders of magnitude if we use the database of Buenos Aires and some *PI* value can differ more than five magnitude orders. To cite one of the many outstanding examples of our results, we have observed a Toba profile with a *PI* value of 273 using the Toba dataset but 15,788,114 using Buenos Aires as the reference population in a case of alleged father–son.

It is worth stressing that in forensic routine work the results of the genetic test are directly communicated to the judge by way of a *PI* value, and these values are therefore those that are finally considered no matter what the values of *Fst* are in the populations. In other words, the use of *Fst* to correct for population stratification might not be appropriate in court.

Conclusions

The results of the present study clearly support the existence of population stratification in Argentina to a level that can be relevant in forensic routine work. On the other hand, the Argentinean populations show low *Fst* values, indicating that the use of this index to measure and correct

for population substructure might be inappropriate in forensics.

We have here simulated pedigree scenarios where a set of 15 different STRs are fully genotyped in all the individuals. These simulations emulate the most favorable scenario. However, in real paternity cases DNA profiles can be deficient (missing data) when using highly degraded DNA (e.g. exhumed remains). Moreover, the discrimination power of the 15-plex can be limited in pedigrees where e.g. a paternity relationship has to be inferred indirectly by genotyping family members related to the alleged father. In these cases, the consequences of using inappropriate databases can be even more dramatic because *PI* values are generally lower.

Using a single database for routine paternity testing in Argentine might not be justified and could lead to serious bias when estimating *PI* values. The approach used in the present study would be also appropriate to investigate the real effect of population stratification in the paternity testing routine work exercised in other countries.

Acknowledgements We would like to thank Thore Egeland for critically reading the article and providing us with useful suggestions. Two grants from the Fundación de Investigación Mútua Madrileña, and a grant from Xunta de Galicia (Grupos Emerxentes; 2008/XA122), given to AS supported this project. MGM is supported by an FPU grant from the Spanish Ministerio de Educación y Ciencia. IPATIMUP is partially supported by Fundação para a Ciência e a Tecnologia (POCI, Programa Operacional Ciência e Inovação 2010).

References

1. Essen-Möller E (1938) Die βeweiskraft der Âhnlichkeit im vaterschaftsnachweis- theoretische Grundlagen. Mitt Anthrop Ges (Wien) 68:9–53
2. Essen-Möller E, Quensel C (1939) Zur Theorie des Vaterschaftsnachweises aufgrund von Ähnlichkeitsbefunden. Z Ges Gerichtl Med 31:70–96
3. Gürtler H (1956) Principles of blood group statistical evaluation of paternity cases at the University Institute of Forensic Medicine Copenhagen. Acta Med Leg Soc (Liege) 9:83–94
4. Valentin J (1983) Positive evidence of paternity calculated according to Essen-Möller: the Bayesian approach. Arlington, Virginia
5. Morling N, Allen RW, Carracedo Á et al (2002) Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases. Forensic Sci Int 129:148–157
6. Gjertson DW, Brenner CH, Baur MP et al (2007) ISFG: recommendations on biostatistics in paternity testing. Forensic Sci Int Genet 1:223–231
7. Evett I, Weir B (1998) Interpreting DNA evidence. Statistical genetics for forensic scientists. Sunderland, Massachusets
8. Salas A, Bandelt H-J, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168:1–13
9. Egeland T, Salas A (2008) Statistical evaluation of haploid genetic evidence. TOForensicSJ 1:4–11
10. Egeland T, Salas A (2008) Estimating haplotype frequency and coverage of databases. PLoS ONE 3:e3988
11. Toscanini U, Gusmão L, Berardi G, Amorim A, Carracedo Á, Salas A, Raimondi E (2007) Testing for genetic structure in different urban Argentinian populations. Forensic Sci Int 165:35–40
12. Toscanini U, Salas A, Carracedo Á, Berardi G, Amorim A, Gusmão L, Raimondi E (2008) A simulation-based approach to evaluate population stratification in Argentina. Forensic Sci Int Genet SS 1:662–663
13. Marino M, Sala A, Bobillo C, Corach D (2008) Inferring genetic substructure in the population of Argentina using fifteen microsatellite loci. Forensic Sci Int Genet 1:350–352
14. Egeland T, Mostad PF (2002) Statistical genetics and genetical statistics: a forensic perspective. Scand J Stat 29:297–307
15. Karlsson AO, Holmlund G, Egeland T, Mostad P (2007) DNA testing for immigration cases: the risk of erroneous conclusions. Forensic Sci Int 172:144–149
16. Montgomery DC (2001) Design and analysis of experiments. Wiley, New York
17. Egeland T, Mostad PF, Mevåg B, Stenersen M (2000) Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int 110:47–59
18. Ayres KL (2000) Relatedness testing in subdivided populations. Forensic Sci Int 114:107–115
19. Toscanini U, Berardi G, Amorim A, Carracedo Á, Salas A, Gusmão L, Raimondi E (2006) Forensic considerations on STR databases in Argentina. Int Congress Series 1288. Elsevier, Amsterdam, pp 337–339
20. Marino M, Sala A, Corach D (2006) Population genetic analysis of 15 autosomal STRs loci in the central region of Argentina. Forensic Sci Int 161:72–77
21. Marino M, Sala A, Corach D (2006) Genetic analysis of the populations from Northern and Mesopotamian provinces of Argentina by means of 15 autosomal STRs. Forensic Sci Int 160:224–230
22. Marino M, Sala A, Corach D (2006) Genetic attributes of 15 autosomal STRs in the population of two patagonian provinces of Argentina. Forensic Sci Int 160:84–88
23. Marino M, Sala A, Corach D (2007) Genetic attributes of the YHRD minimal haplotype in 10 provinces of Argentine. Forensic Sci Int Genet 1:129–133
24. Toscanini U, Gusmão L, Berardi G, Amorim A, Carracedo Á, Salas A, Raimondi E (2008) Y chromosome microsatellite genetic variation in two Native American populations from Argentina: population stratification and mutation data. Forensic Sci Int Genet 2:274–280
25. Cabana GS, Merriwether DA, Hunley K, Demarchi DA (2006) Is the genetic structure of Gran Chaco populations unique? Interregional perspectives on native South American mitochondrial DNA variation. Am J Phys Anthropol 131:108–119
26. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1:44–55
27. Ginther C, Corach D, Penacino GA et al (1993) Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. EXS 67:211–219
28. Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo Á (2008) Gender bias in the multi-ethnic genetic composition of Central Argentina. J Hum Genet 53:662–674