Int J Legal Med (2003) 117 : 302–305
DOI 10.1007/s00414-003-0372-9

SHORT COMMUNICATION

María T. Zarrabeitia · José A. Riancho · María V. Lareu
Francisco Leyva-Cobián · Angel Carracedo

Significance of micro-geographical population structure in forensic cases: a bayesian exploration

Received: 15 April 2002 / Accepted: 24 March 2003 / Published online: 15 July 2003

Abstract We studied the influence of population structure at the microgeographical level on the analysis of forensic cases. A total of nine autosomal STRs and seven Y-STRs were analyzed in the general mixed population and in two relatively isolated valleys of Cantabria, a region in Northern Spain. Statistically significant differences existed in the frequency distribution of four autosomal STRs, with an overall Fst value of 0.3%. A simulation of virtual trio cases revealed that it did not have a practical influence on the analysis of paternity disputes. Significant differences also existed in most Y-STRs, with an overall Fst value of 3%. Thus, using the general database instead of the specific valley database resulted in 5-fold or higher overestimation of the likelihood ratio of matching in up to 30% of cases. A bayesian analysis revealed that this had a significant impact on the estimation of the probability of identity in scenarios of low "a priori" odds of suspicion.

Keywords Microsatellites · STR · Y-chromosome · Evidence interpretation · Population structure

Introduction

Available PCR-based technologies allow the efficient identification of individuals through the analysis of STRs and other genetic markers. In order to interpret the results properly, they need to be compared with those obtained in pertinent reference populations. Fortunately, the frequency distributions of most autosomal STR do not show great variation within large ethnic groups (i.e., caucasian) [1]. Thus, conclusions can usually be drawn whether using a general reference database or a more local one. However, that may not be the case for Y-chromosome markers. The smaller effective population and the lack of recombination make them more prone to showing different frequency distributions related to population structure. Indeed, we have recently shown evidence of population structure even at the microgeographical level in small rural areas [2]. In the present study we used a bayesian approach to address the issue of the possible relevance of those differences for the interpretation of DNA profiles.

Materials and methods

Population

We studied unrelated subjects living in Cantabria, a region in northern Spain with a population of 530,000. This region of 5,000 Km^2 is situated between the sea and the Cantabrian mountains, and has a flat, well communicated, and densely populated coastal area. It has about 400,000 inhabitants, who constitute a mixed and relatively mobile population, living in urban or semi-urban habitats. Therefore, it can be regarded as the source of a general database for the regional population. On the other hand, the southern part of Cantabria is a mountainous area with several valleys that have traditionally had difficult communication. The inhabitants have had less opportunity for social and economic interaction with people from other areas [3, 4]. Among them are the Liébana and the Pas valleys, each with a population about 5,000. Male subjects from the coastal area and from these two valleys were studied (100 individuals from each area).

DNA typing

DNA was isolated from peripheral blood by the Qiagen method (Qiagen, Hilden, Germany). Autosomal STRs were amplified by a multiplex PCR with the Profiler plus kit (including systems D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, and D7S820), according to the manufacturer's instructions (Applied Biosystems). Seven Y-chromosome STRs were also typed using fluorochrome-labeled primers. Loci DYS390, DYS19, DYS 389-I, DYS389-II, and DYS393 were amplified by a

M. T. Zarrabeitia (✉)
Unit of Legal Medicine, Faculty of Medicine,
University of Cantabria, 39011 Santander, Spain
Tel.: +34-942-201984, Fax: +34-942-201903,
e-mail: zarrabet@unican.es

J. A. Riancho · F. Leyva-Cobián
Hospital Marqués de Valdecilla, University of Cantabria,
Santander, Spain

M. V. Lareu · A. Carracedo
Institute of Legal Medicine, University of Santiago de Compostela,
Santiago de Compostela, Spain

pentaplex PCR as described by Gusmao et al. [5]. DYS391 and DYS392 were amplified in a single reaction using the primers described by Gusmao et al. [6] and Kayser et al. [7]. The size of amplified fragments was determined in an ABI Prism 310 analyser, following the recommendations of the International Society of Forensic Genetics [8].

Data analysis

The differences in allelic frequencies of single loci and the corresponding haplotypes were estimated by an extension of Fisher's exact test based on a Markov chain method with 10,000 possible combinations, and carried out with SPSS software. Coancestry coefficients were computed with Arlequin (Schneider et al.: Arlequin ver. 2.000, a software for population genetic data analysis, Genetics and Biometry Laboratory, University of Geneva. http://anthro.unige.ch/arlequin) and FSTAT software (Goudet J: FSTAT. A program to estimate and test gene diversities and fixation indices, http://www.unil.ch/izea/softwares/fstat.html).

Virtual trios were generated with randomly selected individuals and their possible offspring. Allele frequencies were estimated from the general database (coastal area) and from the specific database (the valley of the alleged father). Paternity indices (likelihood ratios) derived from autosomal STRs were computed with PATCAN software [9].

Matching probabilities for each haplotype were calculated as the haplotype frequency in each database. For the purposes of this study, in the case of haplotypes not found in a given database, a minimum frequency of 0.01 was considered. Likelihood ratios were estimated as the inverse of the matching probability. A bayesian estimation of the posterior probability was done with the following formulae implemented in a spreadsheet:

$$\text{Posterior probability} = \text{Posterior odds}/(1 + \text{Posterior odds})$$

where

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio}$$
$$\text{Prior odds} = \text{Prior probability}/(1 - \text{Prior probability})$$

Estimations were computed for three scenarios with different assumptions of a priori probabilities: 0.1 (odds 1:9), 0.5 (odds 1:1), and 0.9 (odds 9:1).

Results

Autosomal STRs

As previously reported [10], allele frequencies in the coastal area were similar to those found in other Caucasian populations. However, statistically significant differences between the three populations studied were found in four out of the nine loci analysed (D3S1358, p=0.003; D18S51, p=0.001; D5S818, p=0.004; and D7S820, p=0.011). The overall Fst value was 0.4% (p=0.001). Pairwise comparisons resulted in the following Fst values: Coast-Liébana, 0.13% (p=0.03); Coast-Pas, 0.55% (p=0.02); Pas-Liébana, 0.52% (p=0.02).

Since the Pas valley population appeared to be the most differentiated, it was chosen to analyse the forensic impact of the differences. Thus, paternity indices (likelihood ratios) and probabilities were calculated from 80 virtual trios from the Pas valley population data. As shown in Fig. 1, using the general database resulted in a slight overestimation of paternity indices. However, it was of little practical importance. With an a priori probability of 10%, post-test probability of paternity was higher than 99.73% in 69 out of 80 trios, using the specific Pas valley database, and in 75 using the coast database frequencies. All trios associated with a paternity probability higher than 99.73% using the coast database were associated with a probability higher than 97.9% when the specific Pas database was used. With a priori probabilities of 0.5 or higher, post-test probabilities of paternity were higher than 99.73% in all cases using the Coast database, and in all but two cases using the Pas database (99.27% and 99.63%, respectively).

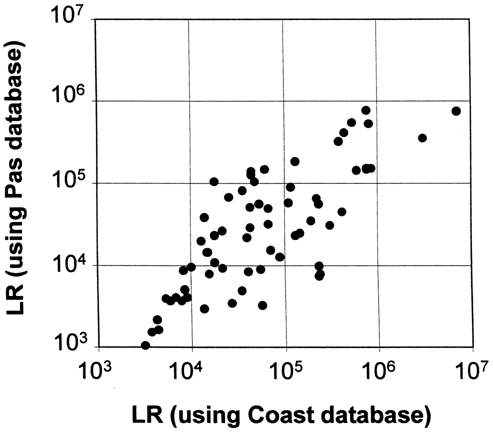

Fig. 1 Paternity indices (likelihood ratios) in 80 virtual trios with fathers from the Pas valley. The results obtained by using the specific valley database and the general database from the population in the coast area are compared

Y-chromosome STRs

There were marked differences among populations in allele frequencies at loci DYS19 (p=0.018), DYS389-II (p<0.001), DYS390 (p<0.001), DYS391 (p=0.006) and DYS392 (p=0.001). There were no significant differences in allele frequencies at loci DYS389-I (p=0.13) and DYS393

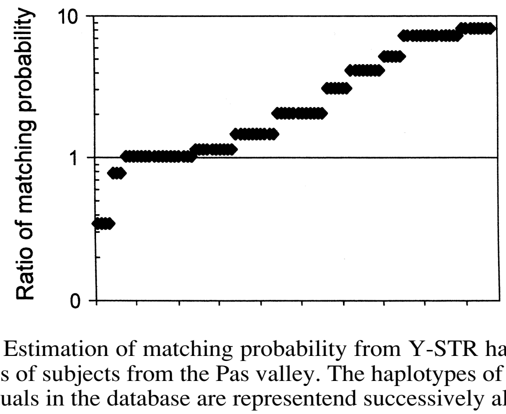

Fig. 2 Estimation of matching probability from Y-STR haplotype analysis of subjects from the Pas valley. The haplotypes of the 100 individuals in the database are representend successively along the horizontal axis. The Y axis show the ratio of the values estimated using the specific Pas valley database, to those obtained using the general coastal database (higher values of matching probability are associated with lower likelihood ratios and less evidence for identity)

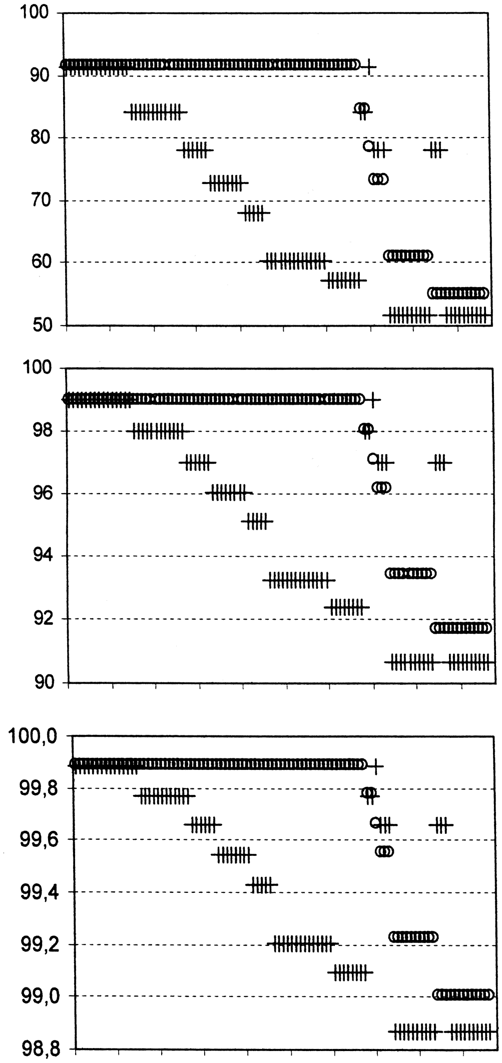

Fig. 3 Post-test identification probabilities (i.e., probability that the sample comes from the suspect) from Y-STR haplotype analysis if the general database is used as the reference (*circles*), in comparison with the results obtained when the specific Pas valley database is used (*crosses*). Results corresponding to the haplotypes included into Pas database are represented successively along the horizontal axis. Three different scenarios of "a priori" probabilities are assumed: 0.1 (*top panel*), 0.5 (*middle panel*) and 0.9 (*lower panel*). Note the different scales used

(p=0.16). Consequently, the populations also showed marked differences when the results were analysed at the haplotype level, with an overall Fst of 3% (p<0.001).

We estimated matching probabilities for each haplotype found in the Pas valley population using both the database of coastal haplotypes and the specific valley database. As shown in Fig. 2, using coastal database matching probabilities were usually underestimated (and subsequently resulted in higher likelihood ratios). In 30% of Pas haplotypes the differences were 5-fold or higher. Since sampling error might influence the results, we also analysed another 15 samples from unselected individuals belonging to the coastal population who had not been included in the database. Matching probabilities and hence likelihood ratios were estimated before and after including those haplotypes into the database. In 14 cases likelihood ratios did not appreciably change; in 1 case it was 1.3-fold higher before including the actual haplotype into the database.

To analyse the impact of those differences on the estimation of identity probabilities, we compared posterior probabilities estimated with the likelihood ratios derived from either the coastal or the specific valley databases, under different scenarios. As shown in Fig. 3, the bias caused by using the general database instead of the specific one was higher in the low pre-test probability scenario, than under those with medium or high a priori suspicion. With a priori probabilities of 50% or higher, little difference existed among the results obtained using different databases.

Discussion

In the present study STR analysis revealed a population structure at the microgeographical level, even in the absence of obvious language or ethnic differences. The differences were more marked in Y-chromosome than in autosomal loci, as indicated by Fst values 10-fold higher in the former. Thus, the results confirm the greater ability of Y-STRs to detect population structure. The smaller effective population size of Y-chromosomes (there are four autosomes and three X-chromosomes every Y-chromosome) and the lack of recombination make them more easily influenced by genetic drift, founder effects and other forces causing differences among populations.

There is general agreement about the value of the bayesian approach as a logical and coherent framework for the interpretation of genetic forensic evidence. Bayes' theorem has two components. On the one hand, the expert summarises DNA typing results, usually as a likelihood ratio, that represents the ratio of the probabilities of observing the data under the two competing hypotheses (i.e., the sample corresponds to the subject or to an unrelated person; the true father is the alleged father or an unrelated person). The other component, established not by the expert, but by the judge, represents the beliefs generated from all other evidence external to the test result. A recently published nomogram may help in relating both components [11].

It has been recommended to consider the influence of population structure and sampling error when analyzing the value of evidence resulting from non-recombining DNA regions, such as mitochondrial DNA and Y-STRs, and different approaches have been suggested and tested [12]. However, the real effect of ignoring population structure in the final decisions made by the judge is unknown.

In this study we used a simple bayesian approach and estimated the differences in post-test probabilities depending on the reference database considered, under different scenarios of pre-test probabilities. The results of our analysis suggest that at this level of differentiation, the bias caused by using a general database for autosomal STRs is not great in most cases. The data also suggest that evidence supplied by Y-STR analysis, added to other studies, may also help in resolving forensic cases, even in the absence of a specific subpopulation database. However, since Y-chromosome STRs are more likely to reveal population sub-structure, care should be taken in drawing conclusions merely on Y-STR haplotype data. Using a general database instead of the specific one may result in a relevant overestimation of the likelihood ratio. This may have a significant impact on post-test probability, particularly when a priori odds of suspicion are low. A further issue is the choice of how many loci should be analysed, as there is no a lineal direct relationship between the number of loci typed and haplotype diviersity [13].

In order to obtain reliable frequency estimates, international efforts are being made to build large databases of Y-STR haplotypes [14]. Nevertheless, as for mitochondrial DNA [12], corrections of estimates by sampling errors should be done when reporting the value of evidence in forensic cases. In addition, relatively isolated European populations should be studied in order to improve our understanding of population genetics of Y-STRs at the local level. The assumption of within-population haplotype frequency homogeneity may not hold for those isolated groups. Provided the data are available, this issue could be addressed by introducing Fst values into the calculations, as suggested by Balding and Nichols [15]. Our results suggest that if the issue is not taken into consideration, interpretation errors could occur, particularly with a low a priori odds of suspicion.

Acknowledgements Supported by a grant from the Fundación Pública Marqués de Valdecilla.

References

1. Budowle B, Shea B, Niezgoda S, Chakraborty R (2001) CODIS STR loci from 41 sample populations. J Forensic Sci 46:453–489
2. Zarrabeitia MT, Riancho JA, Leyva-Cobian F, Sanchez-Diz P, Carracedo A (2002) Differences in Y-chromosome haplotype frequencies at the microgeographical level. In: Brinkmann B, Carracedo A (eds) Progress in forensic genetics 9. Elsevier, Amsterdam, pp 409–412
3. Freeman S (1979) The Pasiegos. Chicago University Press, Chicago
4. Moure A, Suárez M (eds) (1995) De la Montaña a Cantabria. La construcción de una comunidad autónoma. Publicaciones de la Universidad de Cantabria, Santander
5. Gusmao L, Gonzalez-Neira A, Pestoni C, Brion M, Lareu MV, Carracedo A (1999) Robustness of the Y STRs DYS19, DYS389 I and II, DYS390 and DYS393: optimization of a PCR pentaplex. Forensic Sci Int 106:163–172
6. Gusmao L, Gonzalez-Neira A, Sanchez-Diz P, Lareu MV, Amorim A, Carracedo A (2000) Alternative primers for DYS391 typing: advantages of their application to forensic genetics. Forensic Sci Int 112:49–57
7. Kayser M, Caglia A, Corach D et al. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133
8. Gill P, Brenner C, Brinkmann B et al. (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. Int J Legal Med 114:305–309
9. Riancho JA, Zarrabeitia MT (2003) A windows-based software for common paternity and sibling analyses. Forensic Sci Int (in press)
10. Zarrabeitia MT, Riancho JA (2001) Population data on nine STRs from Cantabria, a mountainous region in northern Spain. Forensic Sci Int 122:175–177
11. Riancho JA, Zarrabeitia MT (2002) The prosecutor's and the defendant's bayesian nomograms. Int J Legal Med 116:312–313
12. Tully G, Bar W, Brinkmann B, Carracedo A, Gill P, Morling N, Parson W, Schneider P (2001) Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. Forensic Sci Int 124:83–91
13. Beleza S, Alves C, González-Neira A, Lareu M, Amorim A, Carracedo A, Gusmao L (2003) Extending STR markers in Y chromosome haplotypes. Int J Legal Med 117:27–33
14. Roewer L, Krawczak M, Willuweit S et al. (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118:106–113
15. Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64: 125–140