Int J Leg Med (1993) 105:251–256

International Journal of
Legal Medicine

Population genetics of three VNTR polymorphisms in two different Spanish populations

Emilio Valverde[1], **Carmen Cabrero**[2], **Ricardo Cao**[3], **María Sol Rodríguez-Calvo**[1], **Ana Díez**[2], **Francisco Barros**[1], **Jorge Alemany**[2], **and Angel Carracedo**[1]

[1] Institute of Legal Medicine, Faculty of Medicine, E-15705 Santiago de Compostela, Galicia, Spain
[2] Department of Molecular Biology, Pharmagen Corp, Calera 3, E-28760 Madrid, Spain
[3] Department of Mathematics, Faculty of Computing Sciences, E-15071 La Coruña, Galicia, Spain

Received February 11, 1992 / Received in revised form September 27, 1992

Summary. Two different Spanish populations, one from Galicia (NW Spain) and the other from the rest of Spain, have been analyzed at three different hypervariable loci (YNH24, MS43a and MS31) using the EDNAP electrophoretic protocol and *Hin*fI as restriction enzyme. Although the "rest of Spain" population is a clearly stratified population using classical blood groups, no evidence of stratification for these loci has been found and the differences to the Galician population were not significant, which suggests that a common Spanish population database could be possible. A semiparametric model is proposed for estimating frequencies, using the smoothed cross-validation of Hall et al. (1992) to calculate the size of the window utilized.

Key words: DNA polymorphisms – Single locus probes – Population study

Zusammenfassung. Zwei unterschiedliche spanische Populationen – aus Galizien (Nordwest-Spanien) und die andere aus dem restlichen Spanien – wurden auf drei hypervariable Loci (YNH24, MS43a und MS31) untersucht. Die Methode entsprach dem EDNAP-Protokoll, *Hin*fI wurde als Restriktionsenzym genommen. Obwohl die Population „restliches Spanien" in der klassischen Blutgruppenserologie eine deutlich geschichtete Population darstellt, ist kein Beweis für eine Schichtung dieser Loci gefunden worden, und die Unterschiede zur Bevölkerung von Galizien waren nicht significant. Dies legt nahe, daß eine allgemeine spanische Populationsdatenbank möglich sein könnte. Ein halb-parametrisches Modell wird vorgeschlagen, um die Frequenzen zu schätzen. Dieses benutzt die geglättete Kreuz-Validierung von Hall et al. (1992), um die Größe des benutzten Fensters zu berechnen.

Schlüsselwörter: DNA-Polymorphismen – Single-Locus-Sonden – Populationsstudie

Correspondence to: A. Carracedo

Introduction

Since the introduction of DNA polymorphisms in forensic biology, a great deal of effort has been placed on the standardization of criteria, both in laboratory protocols and in the statistical methods used for the evaluation of results. Standardization and the use of common protocols offer the possibility of sharing databases and the achievement of inter-lab comparisons. Although a common European electrophoretic protocol was finally achieved (Gill et al. 1992), different statistical criteria for the construction of databases and estimating allele frequencies have been proposed (Baird et al. 1986; Budowle et al. 1991; Gill et al. 1990; Pascali et al. 1991).

In our laboratory, we delayed the study of Spanish population samples until a common electrophoretic protocol had been adopted. This has now been achieved, and the results of a study of the Galician population (NW Spain), using *Hin*fI as restriction enzyme, the probes YNH24 (Nakamura et al. 1987), MS43a and MS31 (Wong et al .1987), and following the European electrophoretic EDNAP protocol are presented.

The second aim of this paper is to present the statistical model that we have adopted to estimate frequencies and to construct our databases for SLPs.

The substructure of populations and their influence in forensic cases has generated a great deal of controversy (Lander 1989; Lewontin and Hartl 1991; Chakraborty and Kidd 1991) and because of its peculiarities, the study of the Spanish population may be of interest. Spain is composed of populations with different historical backgrounds and even different languages. Therefore some of these show peculiar population-genetic characteristics (Lewontin and Hartl 1991). So it is necessary to take into account the possibility of stratification, if we are thinking of extending our database to the rest of Spain. Because of that, a mixed Spanish population from the rest of Spain (outside Galicia) was analysed and the results obtained were compared with those of the Galician population to test if a common Spanish database could be possible.

In short, the specific aims of this paper are:

1) Development and adoption of a solid method for estimation of frequencies.

2) Study of the Galician population ($n = 180$).

3) Comparison of results from this study and a stratified population from the rest of Spain, to decide whether or not a common database could be possible.

Materials and methods

Subjects. Samples of peripheral blood (10 ml) were taken from a total of 360 randomly selected subjects. Half of them (180) were from Galicia, a region located in the northwest of Spain. The other half come from the rest of Spain, and their geographical distribution is shown in Fig. 1. Blood was aliquoted (700 µl) and stored at −20°C.

Laboratory protocol. DNA was obtained from leucocytes by incubation in Tris.HCl 50 m*M*, NaCl 150 m*M* and EDTANa$_2$ 100 m*M*, with the addition of SDS (1.25%) and 0.3 mg/ml proteinase K, and precipitated with absolute ethanol after 2 extractions with phenol:chloroform:isoamyl alcohol (25:24:1) and chloroform:isoamyl alcohol (24:1), respectively.

DNA (3–7 µg) was digested with approximately 2 Units of *Hin*fI (Boehringer, Mannheim) per µg DNA, and run overnight in 20 × 25 cm 0.7% agarose/TBE gels, until the 2.0 Kb band of lambda/*Hin*dIII digested DNA saturated with ethidium bromide used as migration control had reached 15 cm from the application point.

Gels were depurinated, denatured, neutralized and transferred onto nylon membranes (Hybond N, Amersham). Membranes were sequentially hybridised with probes YNH24, MS43a and MS31, radioactively labelled using the Random Primed Kit of Boehringer Mannheim (Cat 1004760). The protocol is described elsewhere (Sambrook et al. 1989).

Autoradiography was carried out at −70°C for 2–5 days with 2 intensifying screens.

In order to determine the degree of measurement error, a series of experiments were performed, using the 1 kbp ladder from BRL (Cat 5615SA) and the Wide Range Ladder from Promega (Cat DG1931). The fragment lengths of the former were determined from the autoradiographs using the Promega Ladder as reference, and compared to their true lengths (as given by the manufacturer).

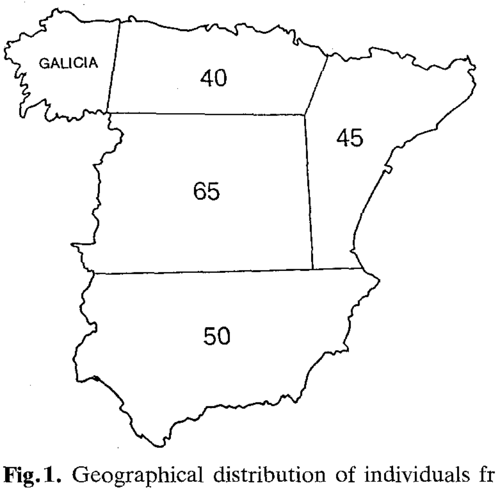

Fig. 1. Geographical distribution of individuals from the "rest of Spain" population

Analysis of bands. Migration distances of bands were measured with a densitometer (Elscript 400, Hirschmann) and converted into bp values using a computer program based on the local reciprocal method of Elder and Southern (1987), using as reference 3 lanes with lambda/ϕX174 ladder (Promega, Cat DG1931). Each gel contained at least one genomic control DNA.

Estimation of allele frequencies. The major statistical concern is the problem of estimating the probability density function of the fragments length (X) when the observed variable is only an approximation.

Let us denote Y to be the fragment length calculated based on the observed migration distance. Since the relation between the fragment length and the migration is, up to a certain level, logarithmic we may state the following model:

$$\ln Y = \ln g(X) + \varepsilon$$

or in other terms:

$$Y = g(X) \cdot \mathrm{e}^{\varepsilon} \tag{1}$$

where ε is a random error term independent of X (due to measurement errors) in the distance domain, assumed to be normal with zero mean and variance σ^2 and g is a real function which enables to take into account not only linear dependence between X and Y, but also more general relationships.

The semiparametric model given above reflects what occurs in practice with the estimated fragment lengths. Computing the conditional mean and variance, we get:

$$m(x) = E\{Y|_{X=x}\} = g(x) \cdot \mathrm{e}^{\sigma^2}/2 \tag{2}$$

$$\mathrm{Var}\{Y|_{X=x}\} = m(x)^2 \cdot \{\mathrm{e}^{\sigma^2} - 1\} \tag{3}$$

Using standard probabilistic tools, we find the formula for σ^2:

$$\sigma^2 = \ln[1 + \mathrm{Var}\{Y \cdot m(X)^{-1}\}] \tag{4}$$

Estimating the regression function, m, by nonparametric kernel methods, $\hat{m}$ (see Priestley and Chao 1972, Nadaraya 1964 or Watson 1964), substituting this estimator in the expression (4) and using a preliminary sample of the true values of X and the approximated lengths, Y, we end up with an estimator, $\hat{\sigma}^2$, of the error variance. Using the estimators $\hat{m}$ and $\hat{\sigma}^2$ in expression (2) we easily find an estimation for the function g.

A direct application of the kernel method (Parzen 1962) to the sample ($\ln Y_1, \ldots, \ln Y_n$) of the logarithms of the approximated fragment lengths leads to an estimator of the density function of $\ln Y$. Using the independence and characteristic functions, a simple estimator for the density function of $\ln g(X)$ is produced when the gaussian kernel is used:

$$\hat{f}^X(x) = \hat{f}^Y \widehat{\sqrt{h^2 \cdot \sigma^2}} (\ln \hat{g}(x)) \cdot \hat{g}'(x)/\hat{g}(x)$$

An important problem is how to select the band width parameter h. For practical purposes we used the smooth cross-validation method of Hall et al. (1992).

Results

Determination of measurement error

The results of the inter-ladder comparison can be seen in Fig. 2. There is a linear correlation between the magnitude of the error and the length of the fragment measured. This is supported by the data from the analysis of genomic control DNA (Table 1).

Population data

Density probability curves for each probe in Galician population are illustrated in Fig. 3.

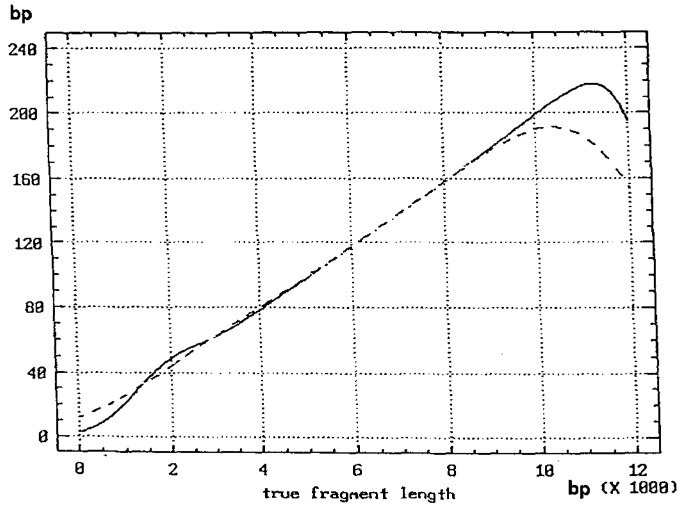

Fig. 2. Kernel regression estimator (—— $h_1 = 800$; ---- $h_2 = 1500$) of the conditional standard deviation observed in the measurement of 1 Kbp Ladder

Table 1. Standard deviations for different bands of genomic controls

Fragment length (bp)	Hybridizing probe	Standard deviation
3124	YNH24	0.33%
3890	YNH24	0.59%
4285	YNH24	0.68%
4327	YNH24	0.34%
5475	MS31	0.21%
6021	MS43a	0.37%
6797	MS43a	0.91%
7016	MS31	0.73%
7797	MS31	0.31%
8233	MS43a	0.82%
8569	MS43a	0.32%

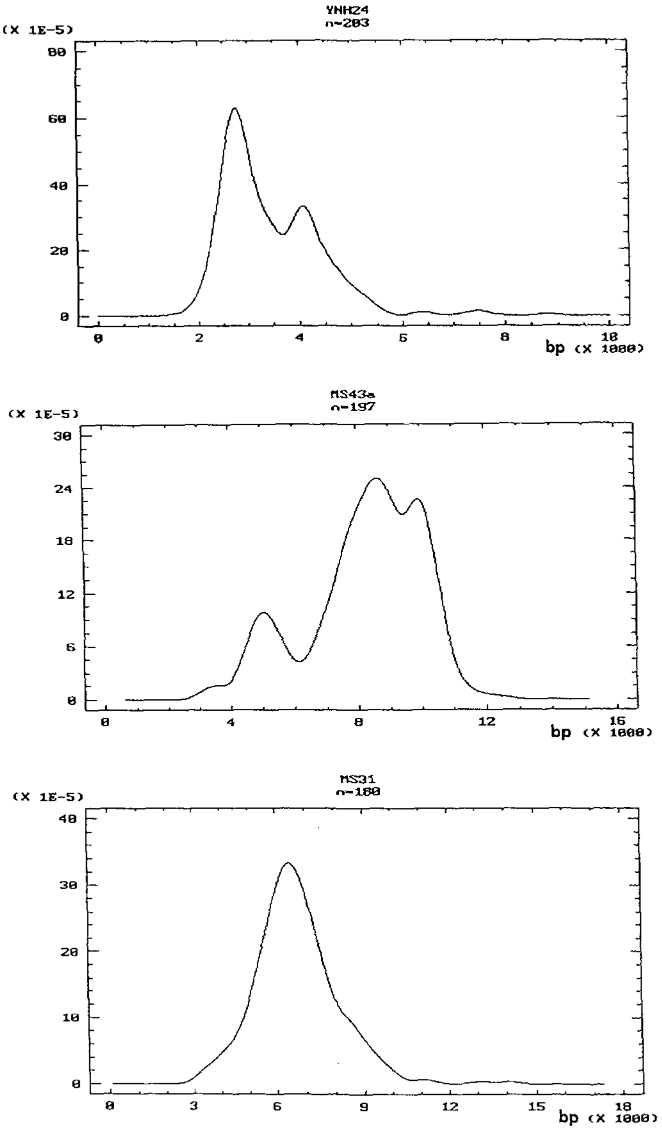

Fig. 3. Probability density curves for Galician population

The main characteristics of each diagram are:

– YNH24. Alleles ranged between 1500–8000 bp. The major peaks appear at 2720 bp and 4000 bp.

– MS43a. Alleles ranged between 3000–14000 bp. The database peaks at 10050 bp, and shows other peaks of less intensity distributed between 4000–9000 bp.

– MS31. Alleles ranged between 3000–14500 bp. The major peak is situated at 6290 bp.

Comparison of populations

To observe the possible differences between the Galician population and other Spanish populations, we have studied a random population ($n = 180$) representative of the rest of Spain.

The results obtained from this second group (Fig. 4) were compared with Galician results using 2 non-parametric tests: Kolmogorov-Smirnov test and Mann-Whitney U test (DeGroot 1986). The results can be seen in Fig. 5 and Table 2, and were not significant ($P > 0.408819$ for the Kolmogorov-Smirnov test and $P > 0.468707$ for the Mann-Whitney test).

Discussion

Estimation of allele frequencies

In estimating population frequencies, account must be taken of the experimental and measurement errors involved in determining fragment sizes. At least 4 methods have been developed for this and are referred to as floating bin (Baird et al. 1986), sliding window (Gill et al. 1990), fixed bin (Budowle et al. 1991) and point estimates (Pascali et al. 1991) approaches.

DNA profiling results can be analysed either by the above mentioned approaches or by the alternative Bayesian approach described by Evett and Werrett (1989) and Berry (1991). All these methods involve estimates of the population frequencies of the DNA fragments that

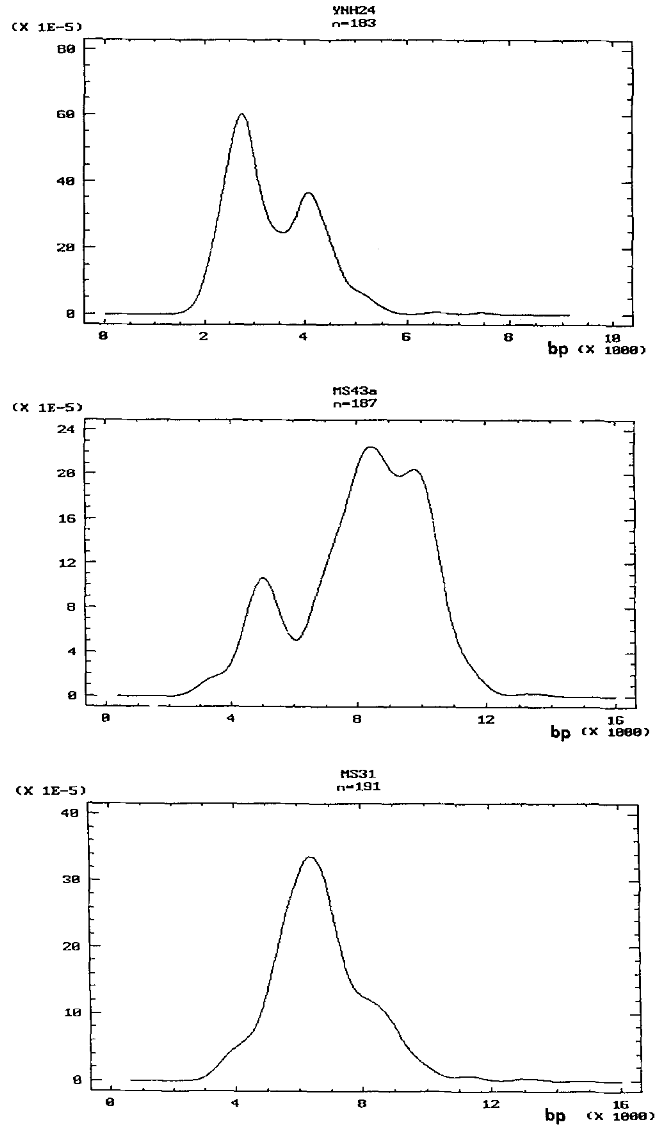

Fig. 4. Probability density curves for the "rest of Spain" population

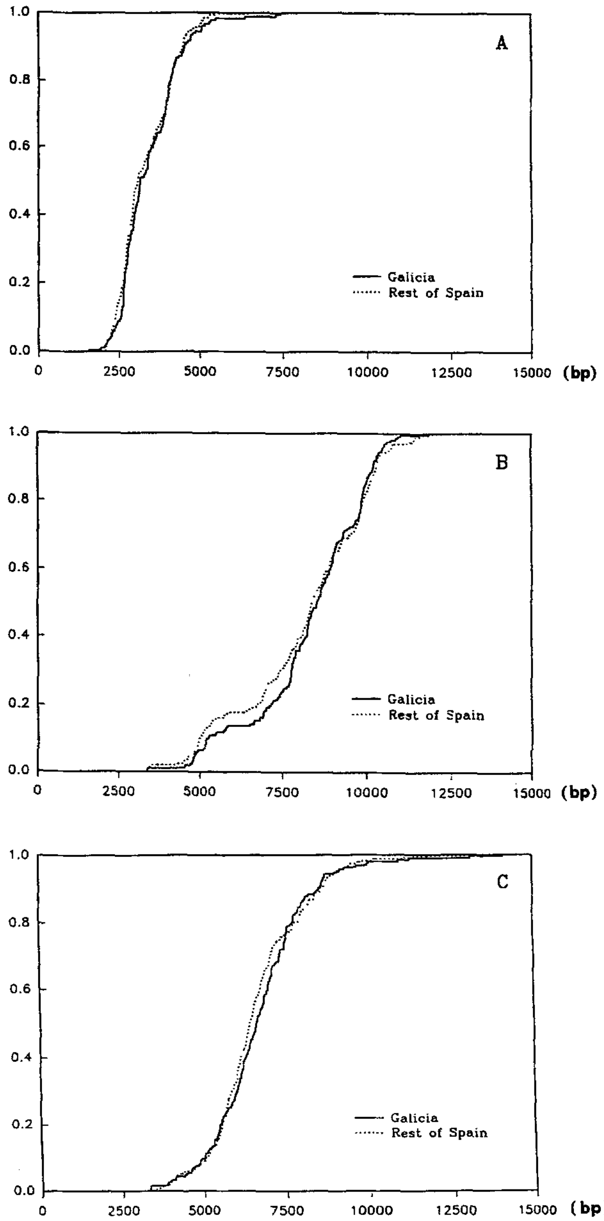

Fig. 5A–C. Accumulated frequency distributions of the two populations, used in the Kolmogorov-Smirnov test (**A**: YNH24; **B**: MS43a; **C**: MS31)

comprise the profiles. We have adopted a Bayesian approach to analyse the results and a semiparametric model to estimate the frequencies and thus the probability of occurrence of a band in our practical casework.

The semiparametric model given in (1) (see Material and methods) uses a parametric structure (normality) for the error ε, but is flexible enough for the marginal probability density function of the variable X. It also allows a great variety of regression functions of Y over X.

When the function g is linear, the model implies that the mean of the observed length, for a true fragment length, is this true value. On the other hand, when g is the identity function $g(x) = x$, the variability of the observed fragment length is proportional to the true fragment length, x. In general, the conditional standard deviation is proportional to $g(x)$. For our data, as Fig. 2 indicates, the assumption of linearity is very reasonable.

Table 2. P values for the Kolmogorov-Smirnov and Mann-Whitney tests, $P = 0.05$ being the maximum P value to refute the hypothesis of equality

	Kolmogorov-Smirnov test	Mann-Whitney test
YNH24	0.51521	0.649707
MS43a	0.408819	0.468707
MS31	0.411979	0.597911

The estimation procedure uses a preliminary sample of data consisting of the true fragment length (X) and the approximated value (Y) obtained using the migration distance in electrophoresis. This preliminary sample is used to estimate the regression function m, the error variance and the function g.

Up to this point, the only part of the model which has to be estimated is the distribution of X (or better the density, since X is a continuous variable).

This can be estimated using the kernel method which is a generalization of the classical histogram (which only counts frequencies in certain intervals) in which we not only count data in the neighbourhood of a particular point, but we also give different weights to the data depending on the distance from the point.

There are many ways of weighing data, this is done by means of the kernel function used. Nevertheless, the most important role in the estimation is played by the band width parameter h which is the scale used when weighing. The meaning of this parameter is closely related to the bin width in the histogram. Its choice is of great importance. Small values for h lead to very wavy estimates, with small bias and a lot of variance (undersmoothing). On the other hand large choices of band width produce flat estimations in the true modes, small variance but large bias (oversmoothing). The band width used in our study was the smoothed cross-validation of Hall et al. (1992). These authors have proved a good theoretical behaviour of this selector. Furthermore, several extensive simulation studies exist that show, in practice, good performance of this proposal.

Population substructure

The practical implications of population stratification in the evidential assesment of SLP results have been discussed by Lander (1989) and Evett and Gill (1991).

Here we have compared two different populations, one from Galicia (an unstratified population, panmitic and with immigration rates near to zero) and the other being a mixed population from the rest of Spain (this is assumed to be a stratified population as the blood group frequences in Southern Spain are similar to those in the Near East and North Africa, reflecting six centuries of Arab rule in the Southern Iberian Peninsula, and clearly different to those in the North of Spain).

The negative results obtained in the comparison of these two populations indicate that a correlation exists between the Galician population and the population from the rest of Spain. This can be supported in spite of the Hardy-Weinberg deviation observed which shows an excess of homozygotes (Table 3). Substructured populations exhibit an overall deficiency of heterozygosity whose proportional magnitude depends on the nature of substructuring as demonstrated by Chakraborty and Jin (1992). Since apparent heterozygote deficiency could be caused by some factors other than population substructuring (e.g., choice of boundaries of bin classes, inability to detect extreme sized alleles, or incomplete resolution of closely migrating alleles) and the homogeneous control (Galician population) shows a similar heterozygosity deficiency, we conclude that, at least in our case, the heterozygosity deficiency is not caused by population substructure, but by the other reasons mentioned above. This fact, and the correlation observed between the 2 populations studied, allow us to assume that a common Spanish database can be used in our practical casework.

Table 3. Comparison of the heterozygosity rates obtained from the two populations studied

System	Galicia	Rest of Spain
YNH24	90.15%	92.35%
MS43a	88.32%	94.09%
MS31	93.89%	91.10%

Acknowledgements. This study was supported by grants from the Ministerio de Educacĩon y Ciencia (CICYT PA86-0453) and Xunta de Galicia (XUGA8420689). We are grateful to Cellmark Diagnostics, Abingdon, Oxfordshire, England, and Dr. Paul Debenham for kindly supplying the probes MS43a and MS31.

References

Baird M, Balazs I, Giusti A, Miyasaki GL, Nicholas L, Wexler K, Kanter E, Glassberg J, Allen F, Rubinstein P, Sussman L (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am J Hum Genet 39:489–501

Berry DA (1991) Inferences using DNA profiling in forensic identification and paternity cases. Statist Sci 6:175–205

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, Deadman HA, Monson KL (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841–855

Chakraborty R, Jin L (1992) Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum Genet 88:267–272

Chakraborty R, Kidd K (1991) The utility of DNA typing in forensic work. Science 254:1735–1739

DeGroot MH (1986) Probability and statistics. Addison-Wesley, Reading Mass

Elder JK, Southern EM (1987) Computer-aided analysis of one-dimensional restriction fragment gels. In: Bishop MJ, Rawlings CJ (eds) Nucleic acid and protein sequence analysis. IRL Press, Oxford, pp 165–172

Evett IW, Werrett DJ (1989) Bayesian analysis of single locus profiles. In: Proceedings of the International Symposium on Human Identification 1989: Data adquisition and statistical analysis for DNA typing laboratories. Promega Corp, Madison, USA, pp 77–101

Evett IW, Gill P (1991) A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis 12:226–230

Gill P, Sullivan K, Werrett DJ (1990) The analysis of hypervariable DNA profiles: problems associated with the objective determination of a match. Hum Genet 85:75–79

Gill P, Woodroffe S, Bär W, Brinkmann B, Carracedo A, Eriksen B, Jones S, Kloosterman AD, Ludes B, Mevag B, Pascali VL, Schmitter H, Schneider PM, Thomson JA (1992) A report of an international collaborative experiment to demonstrate the uniformity obtainable using DNA profiling techniques. Forensic Sci Int 53:29–43

Hall P, Marron JS, Park B (1992) Smoothed cross-validation. Probab Theor Related Fields (in press)

Lander ES (1989) DNA fingerprinting on trial. Nature 339:501–505

Lewontin RC, Hartl DL (1991) Population genetics in forensic DNA typing. Science 254:1745–1750

Nadaraya EA (1964) On estimating regression. Theor Probab Applic 9:141–142

Nakamura Y, Leppert M, O'Connel P, Wolff P, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Statist 33:1065–1076

Pascali VL, d'Aloja E, Dobosz M, Pescarmona M (1991) Estimating allele frequencies of hypervariable DNA systems. Forensic Sci Int 51:273–280

Priestley MB, Chao MT (1972) Nonparametric function fitting. J R Statist Soc [B] 34:385–392

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor New York

Watson GS (1964) Smooth regression analysis. Sankhya [Ser A] 26:359–372

Wong Z, Wilson V, Patel I, Povey S, Jeffreys AJ (1987) Characterization of a panel of highly variable minisatellites cloned from human DNA. Ann Hum Genet 51:269–288