Int J Legal Med (2002) 116 : 153–160

TECHNICAL NOTE

Jae Won Lee · Hye-Seung Lee · Juck-Joon Hwang

Statistical analysis for estimating heterogeneity of the Korean population in DNA typing using STR loci

Received: 27 February 2001 / Accepted: 24 July 2001

Abstract The coancestry coefficients for the Korean population are estimated by using 3 statistical methods for 17 loci (D3S1358, D21S11, D13S317, D18S51, D7S820, D8S1179, D5S818, FGA, VWA, F13A1, FES/FPS, THO1, TPOX, CSF1PO, D12S391, GABA and ACTBP2). The subpopulations were considered by last name and home origin, respectively. Our results show that the values for the coancestry coefficient for the Korean population are too large to ignore although they do not show substantial heterogeneity. These estimated values are also applied to simulated forensic cases.

Keywords Coancestry coefficient · Allele frequency · Korean population

Introduction

In forensic case work, DNA typing has become the most important tool for identification. Although it is similar to the analysis by conventional serological techniques, there are two differences: the scale of implementation of the new methods and the enormous power of the evidence. In particular, STR (short tandem repeat) loci in conjunction with PCR (polymorphism chain reaction) have made identification using DNA more useful to type very small amounts of DNA.

J. W. Lee (✉)
Department of Statistics, Korea University,
5-1, Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea
e-mail: jael@mail.korea.ac.kr,
Tel.: +82-2-32902237,
Fax: +82-2-9249895

H.-S. Lee
Division of Biostatistics, Columbia University,
600 West 168th Street, New York, NY 10032, USA

J.-J. Hwang
Department of Legal Medicine,
Korea University College of Medicine,
126-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea

In order to use DNA profiles appropriately, the forensic scientist should consider the relationships between two compared persons since there is a chance that the two persons might have DNA profile patterns (i.e. genetic types) that match in the small number of loci examined. Typically, the relationship can be discriminated into three levels; unrelated, unrelated in the same subpopulation and related in the same family. Among these, the relationship that two compared persons are unrelated in the same subpopulation comes from the correlation of alleles induced by heterogeneity among the subpopulations (Lewontin & Hartl 1991; Nichols & Balding 1991; Hartl & Lewontin 1993). Disregarding this correlation of alleles in the forensic calculation for DNA evidence results in exaggerating the strength of the evidence against the compared person (e.g. suspect or alleged father), even though it is not as important as the relatedness in the same population. Thus, it is important to evaluate the coancestry effects which incorporate the population structure for a specific population and apply it to the evaluation of DNA evidence. Evaluation of the coancestry effects allows the use of established population genetics and also makes the conventional assumption of unrelatedness of two compared persons unnecessary.

In this paper, we estimated the coancestry coefficient for a South Korean population. Since Korea is known to be a single nation sharing an approximately 5000-year history, the argument that the Korean population is homogeneous seems to be persuasive and thus current forensic work in Korea does not allow for the coancestry coefficient. However, an extensive shared history tends to reveal a greater coancestry coefficient (Evett & Weir 1998) and a substructure may be present in Korea because the Korean people might have been less mobile for a long period of time. Thus it is desirable to routinely implement the coancestry coefficient in the forensic calculations. When we use the notation of Wright (1951, 1965), the relationship between a pair of alleles is denoted as F_{IS} for alleles within individuals within subpopulations, F_{IT} for alleles within individuals relative to the total and F_{ST} for alleles between individuals within subpopulations relative

to the total. Since F_{IS} is affected primarily by the mating system within subpopulations, when we consider people tend to avoid marrying relatives, it might be expected to be slightly negative for human populations (Evett & Weir 1998). Therefore, we assume that F_{IS} is very small if not zero and thus F_{IT} is roughly the same as F_{ST}, which is the coancestry coefficient θ we intend to estimate. Most methods for estimating θ require the subpopulation assignment, but the subpopulations are hard to define and the choice of subpopulations and the estimation method is up to the researcher's discretion. It is therefore useful to investigate correlations for a variety of ways of specifying a subpopulation and for various estimation methods and the results can then be used to obtain a reasonable value of the coancestry coefficient to substitute into more appropriate forensic calculations. Thus, we considered two ways of defining a subpopulation and used the methods of Weir & Cockerham (1984), Balding et al. (1996) and Roeder et al. (1998) to estimate the coancestry coefficient. In addition, we compared these methods in the simulated forensic scenarios of criminal cases and paternity trio cases.

Materials and methods

Data

The STR genotypes at 17 loci (D3S1358, D21S11, D13S317, D18S51, D7S820, D8S1179, D5S818, FGA, VWA, F13A1, FES/FPS, THO1, TPOX, CSF1PO, D12S391, GABA, ACTBP2) were obtained from 1,164 persons in the South Korean population. Among them, we selected 492 apparently unrelated persons to investigate the subpopulation effect.

Most subpopulation effects in a population result from the mating system, and thus we considered the home origin and the last name as the factor causing heterogeneity in South Korean population, respectively. Geographical location is the most reasonable factor in this study and the choice of the last name comes from the fact that the possibility that two persons with the same family name share the same allele could be higher than that from two persons with different family names. Actually, although there are hundreds of last names in South Korea, we considered only three groups ("Kim", "Lee", "Park") to which the majority of the Korean people belong. For the home origin, we divided South Korea into five areas (except for Jeju Island): Kangwon and northern area (north-east area), Seoul and Kyungee (north-west area), Kyungsang (south-east area), Junla (south-west area) and Chungchung (central area). In South Korea, more than half of the whole population have the above three dominating family names and the others have hundreds of different family names, and sharing the same family name does not imply that they are related. Instead, each family name is separated into hundreds of groups according to their locality. For example, if the first ancestor was born in Hansan 700 years ago and his family name is Lee, then all his descendants belong to the Hansan-Lee family. In South Korea, all the people in this Hansan-Lee family are considered as relatives, and they cannot marry each other by law. For this reason, the Korean people believe that the population may be genetically differentiated by family name. In our data, all the people who share the same family name have different localities and thus they are unrelated. In this sample we have 105 Kims, 64 Lees and 40 Parks, with 37 individuals from Kangwon and North area, 146 from Seoul and Kyungee, 67 from Kyungsang, 69 from Junla and 54 from Chungchung. Although these five areas cover most of the areas in South Korea, we have only 373 (out of 492) samples since the information on the home origin are missing in many samples. In addition, in order to simulate forensic works, 492 criminal cases and 490 paternity trio cases were generated from the original 1,164 persons.

Estimation methods

We used the three methods of Weir & Cockerham (1984), Balding et al. (1996) and Roeder et al. (1998). Strictly speaking, while the methods of Weir & Cockerham (1984) and Balding et al. (1996) estimate F_{ST} using the classified subpopulation information, the method of Roeder et al. (1998) estimates F_{IT}. However, as mentioned above, since we assume that F_{IS} is very small if not zero, we regard that F_{IT} is almost the same as F_{ST} in the South Korean population. In addition, while the method of Weir and Cockerham (1984) is a classical procedure which rests on the method of moments, the other two methods are based on the Bayesian approach. Balding et al. (1996) estimated subpopulation-specific and locus-specific values from the subpopulation information. They used lognormal distribution and Roeder et al. (1998) used beta distribution as the prior information. We have used various prior distributions to examine how this affects the estimation of the coancestry coefficient and found that it has no practical impact on the forensic calculation. Thus, we followed their choices of prior distributions in our calculations. We also used the sequential plot of each generated output to check convergence to the target distribution and confirmed that there is no trend.

For a locus (L), let n and r be the total sample size and the number of subpopulations, respectively and p_K and n_K denote the frequency and counts of allele K in the population, p_{Ks} is the frequency for allele K in the s-th subpopulation and m_S is the number of all the alleles in s-th subpopulation.

The method of Weir and Cockerham (1984)

This uses directly the theory that the coancestry coefficient refers to pairs of alleles in different individuals in the same subpopulation, relative to pairs of alleles in the whole population. Thus, it requires data from more than one subpopulation and it can be estimated by comparing allelic variation within and between populations. To do this, two mean squares are calculated: *MSA* among subpopulations and *MSW* within subpopulations.

$$MSA = \frac{2n}{r-1}\sum_{S=1}^{r}(\hat{p}_{K_S} - \overline{p_K})^2, \text{ where } \overline{p_K} = \frac{\sum_{S=1}^{r}\hat{p}_{K_S}}{r} \quad (1)$$

$$MSW = \frac{2n}{r(2n-1)}\sum_{S=1}^{r}\hat{p}_{K_s}(1-\hat{p}_{K_s}) \quad (2)$$

From the above two mean squares, the coancestry coefficient can be estimated for each locus as follows:

$$\hat{\theta} = \frac{\sum_{K}(MSA - MSW)}{\sum_{K}(MSA + (2n-1)MSW)} \quad (3)$$

The coancestry coefficient for all loci is as follows:

$$\hat{\theta} = \frac{\sum_{L}\sum_{K}(MSA - MSW)}{\sum_{L}\sum_{K}(MSA + (2n-1)MSW)} \quad (4)$$

In addition, Li (1996) derived sampling properties of this estimator and suggested the confidence interval for the estimator θ_K (Weir 2001).

$$\text{Lower confidence limit} = \frac{(r-1)\hat{\theta}_K}{U(1-\hat{\theta}_K) + (r-1)\hat{\theta}_K} \quad (5)$$

$$\text{Upper confidence limit} = \frac{(r-1)\hat{\theta}_K}{L(1-\hat{\theta}_K) + (r-1)\hat{\theta}_K} \quad (6)$$

Here, the quantities L, U are the $\alpha/2$ and $1-\alpha/2$ percentiles $\chi^2_{(r-1)}$ of distribution.

The method of Balding et al. (1996)

This is a likelihood-based method combining information over loci and subpopulations. The likelihood they implement is multinomial-Dirichlet for each locus i and subpopulation j, for the probability of observing a sample of alleles. That is, allele frequencies in each subpopulation are assumed to be unknown and are modelled by a Dirichlet distribution with parameters $(1/\theta-1)p_K$. The likelihood form for locus i and subpopulation j is proportional to:

$$\frac{\Gamma(1/\theta-1)}{\Gamma(1/\theta-1+m_S)}\prod_K \frac{\Gamma((1/\theta-1)p_K+n_{K_S})}{\Gamma((1/\theta-1)p_K)} \quad (7)$$

(from personal discussions with Dr. Karen Ayres).

To estimate θ, they used a Metropolis-Hastings algorithm to carry out a Bayesian analysis. They used a model for θ which includes "hyperparameters" for each locus and subpopulation, $\theta_{ij} = 1/(1 + a_i + b_j)$. In our calculations, the priors are on a_i and b_j and are assumed to be independent lognormal (3.5, 1.5) for θ between 0 and 1, although a_i and b_j should be between 0 and infinity. The autocorrelation plot between a_i and b_j showed that there is not much correlation.

The method of Roeder et al. (1998)

This is an indirect method based on the excess of homozygosity using Bayesian techniques. Another similar method using this Bayesian technique was suggested by Foreman et al. (1997). However, since they generated the subpopulation data from the assumed number of subpopulations, this is still based on the estimation of F_{ST}. Roeder et al. (1998) suggested this method without generating the subpopulation data and it uses a mixture of two functions. When a genotype of one person is (x_1, x_2), the mixture is defined as:

$$f(x_1,x_2|G) = \theta f_1(x_1,x_2) + (1-\theta) f_2(x_1,x_2), \text{ where}$$
$$f(x_1,x_2) = \begin{pmatrix} p_K & if \;\; x_1 = x_2 = K \\ 0 & if \;\; x_1 \neq x_2 \end{pmatrix} \quad (8)$$
$$f(x_1,x_2) = \begin{pmatrix} p_K^2 & if \;\; x_1 = x_2 = K \\ 2p_K p_{K'} & if \;\; (x_1,x_2) = (K,K') \end{pmatrix}$$

Then, for a given θ, the probability that an observation (x_1,x_2) is $f_1(x_1,x_2)$ is:

$$\delta = \frac{\theta f_1(x_1,x_2)}{\theta f_1(x_1,x_2) + (1-\theta) f_2(x_1,x_2)} \quad (9)$$

Gibbs sampler is used to obtain the posterior distribution of θ, and the Markov Chain Monte Carlo (MCMC) procedure is also involved. As in Roeder et al. (1998), we have used Beta (1, 49) as our prior for θ. Although not tabulated here, we have also examined the effect of prior distributions on the estimation of coancestry coefficient through simulations and found the choice of prior distribution has no practical impact on the estimation. The effect of prior was also discussed in Roeder et al. (1998). Note also that the lognormal (3.5,1.5) prior used with the Balding method and Beta (1,49) prior used with the Roeder method give almost the same prior means. We ran the Gibbs sampler for an initial 300 cycles as a burn-in to achieve a stationary distribution; these results were discarded. Then another 29,700 cycles were run to obtain an estimate of the posterior distribution. As discussed in Geyer (1992), discarding the initial 1 or 2% of runs would usually suffice.

Results

Comparison of allele frequencies among the subpopulations

When the allele proportions are different among the subpopulations, there might appear to be Hardy-Weinberg disequilibrium in the population as a whole even if there is an equilibrium within each subpopulation. This phenomenon is known as the Wahlund effect caused by an excess homozygosity. Hence, we first examined Hardy-Weinberg equilibrium for each locus by using the method suggested by Guo and Thompson (1992). Results from this independence test for alleles within a locus are shown in Table 1 and the ACTBP2 locus shows a significant disequilibrium (p-value = 0.009). In addition, we examined the significance of the difference in allele frequencies among the subpopulations classified by the last name and the home origin, respectively. We applied Fisher's exact test (Freeman & Halton 1951) using StatXact software 4.01. The results are shown in Table 1 and there is no significant difference in allele frequencies among the subpopulations.

Estimation of coancestry coefficient in the Korean population

The coancestry coefficient values (θ) for each locus are shown in Tables 2, 3, 4 and 5 when the methods of Weir & Cockerham (1984), Balding et al. (1996) and Roeder et al. (1998) are applied, respectively. For the method of Weir & Cockerham (1984) and Balding et al. (1996),

Table 1 Comparison of allele frequency among subpopulations: Fisher's exact test

Locus	Number of alleles	H-W equilibrium test	Fisher's exact test using Monte Carlo method (p-value)	
			Last name	Home
D3S1358	7	0.287	0.1441	0.6301
D21S11	16	0.832	0.5652	0.4969
D13S317	9	0.765	0.1595	0.4231
D18S51	17	0.481	0.9872	0.8308
D7S820	9	0.853	0.3195	0.7896
D8S1179	9	0.113	0.8900	0.4133
D5S818	10	0.172	0.8096	0.2454
FGA	17	0.828	0.6627	0.7499
VWA	8	0.533	0.3377	0.1400
F13A1	5	0.409	0.4848	0.7124
FES/FPS	7	0.492	0.6868	0.2348
THO1	7	0.125	0.3740	0.3607
TPOX	7	0.093	0.7716	0.1072
CSF1PO	9	0.100	0.8314	0.3110
D12S391	14	0.676	0.7708	0.8124
GABA	7	0.485	0.6002	0.2953
ACTBP2	34	0.009	0.4280	0.7257

Table 2 The estimation of coancestry coefficient (θ) by the Weir and Cockerham (1984) method

Locus	Number of alleles	Last name	Confidence interval		Home origin	Confidence interval	
			Lower	Upper		Lower	Upper
D3S1358	7	0.00882	0.00241	0.26252	0.00301	0.00082	0.10787
D21S11	16	0.00570	0.00155	0.18664	0.00500	0.00136	0.16731
D13S317	9	0.01544	0.00423	0.38546	0.00509	0.00138	0.16976
D18S51	17	0.00380	0.00103	0.13230	0.00375	0.00102	0.13094
D7S820	9	0.00856	0.00233	0.25667	0.00569	0.00155	0.18631
D8S1179	9	0.00456	0.00124	0.15490	0.01261	0.00345	0.33811
D5S818	10	0.00457	0.00124	0.15507	0.00498	0.00135	0.16678
FGA	17	0.00784	0.00214	0.24007	0.00542	0.00148	0.17910
VWA	8	0.01064	0.00291	0.30087	0.01211	0.00331	0.32901
F13A1	5	0.00182	0.00049	0.06807	0.00234	0.00063	0.08560
FES/FPS	7	0.00146	0.00040	0.05519	0.00330	0.00090	0.11701
THO1	7	0.01038	0.00283	0.29545	0.00227	0.00062	0.08343
TPOX	7	0.00609	0.00166	0.19676	0.00733	0.00200	0.22801
CSF1PO	9	0.00437	0.00119	0.14947	0.00728	0.00198	0.22676
D12S391	14	0.00274	0.00075	0.09908	0.00900	0.00245	0.26641
GABA	7	0.00729	0.00199	0.22701	0.00781	0.00213	0.23939
ACTBP2	34	0.00613	0.00167	0.19797	0.00620	0.00169	0.19960

Table 3 Estimates of percentiles of the posterior distribution of coancestry coefficient (θ) by the Balding et al. (1996) method for last name

Percentile	D3S1358	D21S11	D13S317	D18S51	D7S820
95%	1.861E-03	1.857E-03	1.869E-03	1.800E-03	1.882E-03
90%	1.537E-03	1.530E-03	1.522E-03	1.510E-03	1.555E-03
75%	1.090E-03	1.074E-03	1.078E-03	1.064E-03	1.092E-03
50%	7.027E-04	6.977E-04	7.003E-04	6.887E-04	7.063E-04
25%	4.167E-04	4.110E-04	4.133E-04	4.077E-04	4.187E-04
10%	2.473E-04	2.467E-04	2.480E-04	2.460E-04	2.490E-04
5%	1.783E-04	1.773E-04	1.750E-04	1.767E-04	1.767E-04
Mean	0.000817	0.000811	0.000814	0.000800	0.000825

Percentile	D8S1179	D5S818	FGA	VWA	F13A1
95%	1.837E-03	1.819E-03	1.830E-03	1.854E-03	1.866E-03
90%	1.520E-03	1.506E-03	1.525E-03	1.530E-03	1.537E-03
75%	1.077E-03	1.069E-03	1.072E-03	1.065E-03	1.086E-03
50%	6.970E-04	6.897E-04	7.080E-04	7.013E-04	7.050E-04
25%	4.167E-04	4.080E-04	4.163E-04	4.133E-04	4.197E-04
10%	2.453E-04	2.480E-04	2.473E-04	2.433E-04	2.457E-04
5%	1.7504E-04	1.760E-04	1.750E-04	1.757E-04	1.743E-04
Mean	0.000812	0.000803	0.000809	0.000812	0.000819

Percentile	FES/FPS	THO1	TPOX	CSF1PO	D12S391
95%	1.841E-03	1.842E-03	1.834E-03	1.837E-03	1.840E-03
90%	1.536E-03	1.503E-03	1.514E-03	1.515E-03	1.514E-03
75%	1.067E-03	1.047E-03	1.071E-03	1.068E-03	1.068E-03
50%	6.947E-04	7.013E-04	6.940E-04	6.957E-04	6.930E-03
25%	4.147E-04	4.123E-04	4.120E-04	4.133E-04	4.113E-04
10%	2.447E-04	2.447E-04	2.467E-04	2.467E-04	2.460E-04
5%	1.760E-04	1.740E-04	1.743E-04	1.753E-04	1.753E-04
Mean	0.000813	0.000812	0.000806	0.000808	0.000806

Percentile	GABA	ACTBP2
95%	1.867E-03	1.790E-03
90%	1.538E-03	1.498E-03
75%	1.084E-03	1.060E-03
50%	6.997E-04	6.960E-04
25%	4.160E-04	4.107E-04
10%	2.490E-04	2.457E-04
5%	1.773E-04	1.757E-04
Mean	0.000819	0.000797

Table 4 Estimates of percentiles of the posterior distribution of coancestry coefficient (θ) by the Balding et al. (1996) method for home origin

Percentile	D3S1358	D21S11	D13S317	D18S51	D7S820
95%	7.282E-03	2.305E-03	5.024E-03	4.656E-02	3.855E-03
90%	5.677E-03	1.861E-03	3.831E-03	4.000E-02	3.028E-03
75%	3.795E-03	1.310E-03	2.484E-03	2.996E-02	2.036E-03
50%	2.341E-03	8.136E-04	1.477E-03	2.164E-02	1.241E-03
25%	1.366E-03	4.756E-04	8.618E-04	1.555E-02	7.220E-04
10%	8.324E-04	2.744E-04	4.942E-04	1.120E-02	3.946E-04
5%	5.646E-04	2.064E-04	3.410E-04	9.127E-03	2.794E-04
Mean	0.002926	0.000982	0.001931	0.023980	0.001549
Percentile	D8S1179	D5S818	FGA	VWA	F13A1
95%	4.969E-03	3.971E-03	2.730E-03	8.082E-03	8.752E-03
90%	3.960E-03	3.176E-03	2.286E-03	6.522E-03	6.649E-03
75%	2.649E-03	2.139E-03	1.603E-03	4.522E-03	4.218E-03
50%	1.646E-03	1.325E-03	1.019E-03	2.900E-03	2.532E-03
25%	9.518E-04	7.890E-04	6.032E-04	1.721E-03	1.437E-03
10%	5.460E-04	4.398E-04	3.512E-04	1.043E-03	8.508E-04
5%	4.074E-04	3.140E-04	2.464E-04	7.412E-04	5.710E-04
Mean	0.002016	0.001631	0.001196	0.003465	0.003287
Percentile	FES/FPS	THO1	TPOX	CSF1PO	D12S391
95%	5.164E-03	6.847E-03	6.806E-03	5.484E-03	4.023E-03
90%	4.035E-03	5.402E-03	5.244E-03	4.269E-03	3.314E-03
75%	2.622E-03	3.556E-03	3.369E-03	2.718E-03	2.279E-03
50%	1.562E-03	2.154E-03	1.989E-03	1.666E-03	1.473E-03
25%	8.732E-04	1.219E-03	1.094E-03	9.578E-04	9.106E-04
10%	4.938E-04	6.794E-04	6.216E-04	5.358E-04	5.690E-04
5%	3.518E-04	4.780E-04	4.542E-04	3.904E-04	3.896E-04
Mean	0.002000	0.002723	0.002591	0.002131	0.001753
Percentile	GABA	ACTBP2			
95%	6.231E-03	1.816E-03			
90%	4.913E-03	1.544E-03			
75%	3.225E-03	1.117E-03			
50%	1.999E-03	7.250E-04			
25%	1.159E-03	4.402E-04			
10%	6.798E-04	2.684E-04			
5%	4.694E-04	1.918E-04			
Mean	0.002502	0.000832			

these values were obtained for the subpopulations classified by the last name and the home origin, respectively. First, we estimated the θ values for each locus independently. As shown here, the estimates using the Weir & Cockerham method are in the range of 0.00146–0.01544 for the last name classification and 0.00227–0.01261 for the home origin classification. For the Balding method with the home origin classification, the average median θ for the 17 loci is 0.00285, the average mean value is 0.00338 and the mean values are in the range of 0.000913–0.00764. For the last name classification, the average median θ for the 17 loci is 0.00070, the average mean value is 0.00081 and the mean values are in the range of 0.00018–0.00184. For the Roeder method, the average median θ for the 17 loci is 0.00167, the average mean value is 0.00219 and the mean values are in the range of 0.00178–0.00314. For the Weir-Cockerham method, it is interesting to see that D3S1358, D21S11, D13S317, D18S51, D7S820, FGA and TH01 have larger values when the subpopulations are classified by three last names, and the other 10 loci have larger values when the subpopulations are classified by 5 home origins. However, these two values are not very different. On the other hand, for the Balding method, the difference between these two classifications of subpopulation is quite large. As shown in Tables 3 and 4, θ values with the five home origins are about 3 times as much as those with the three last names, and it shows the Balding method is more sensitive to the number of subpopulations than the Weir-Cockerham method. Thus, for the Balding method, while the performance with the home origin classification is close to that with using the Roeder method, the performance with the last name classification is close to that with the method assuming unrelatedness. In addition, Tables 3, 4 and 5 all show the longer tail distribution that the median is smaller than the mean for each locus.

The confidence intervals from the Weir-Cockerham method are wide and greatly overlap and it shows no evidence against the constancy assumption (Table 2). Also, the distribution of the θ values from using the Balding

Table 5 Estimates of percentiles of the posterior distribution of coancestry coefficient (θ) by the Roeder et al. (1998) method

Percentile	D3S1358	D21S11	D13S317	D18S51	D7S820
95%	5.43653E-03	7.78466E-03	5.37153E-03	7.43504E-03	5.76448E-03
90%	4.28062E-03	6.44648E-03	4.22233E-03	6.12696E-03	4.41077E-03
75%	2.61505E-03	4.44988E-03	2.65697E-03	4.23635E-03	2.65535E-03
50%	1.36667E-03	2.66869E-03	1.35404E-03	2.50849E-03	1.39798E-03
25%	5.65555E-04	1.31352E-03	5.57003E-04	1.12954E-03	6.02410E-04
10%	2.04650E-04	2.99826E-04	2.03555E-04	2.76773E-04	2.10181E-04
5%	9.94349E-05	1.16340E-04	9.91589E-05	1.22939E-04	1.00462E-04
Mean	0.001888	0.003141	0.001880	0.002956	0.001952
Percentile	D8S1179	D5S818	FGA	VWA	F13A1
95%	6.41794E-03	7.72999E-03	5.26264E-03	6.89665E-03	5.26608E-03
90%	5.16829E-03	6.41002E-03	4.14654E-03	5.40568E-03	4.14776E-03
75%	3.50467E-03	4.43585E-03	2.60671E-03	3.38522E-03	2.60723E-03
50%	1.93420E-03	2.65442E-03	1.33211E-03	1.71069E-03	1.33266E-03
25%	6.68663E-04	1.29933E-03	5.52259E-04	7.23363E-04	5.51995E-04
10%	2.11494E-04	2.81019E-04	2.02688E-04	2.21156E-04	2.02540E-04
5%	1.01154E-04	1.13775E-04	9.89528E-05	1.00466E-04	9.88129E-05
Mean	0.002379	0.003119	0.001849	0.002365	0.001848
Percentile	FES/FPS	THO1	TPOX	CSF1PO	D12S391
95%	5.26378E-03	5.18669E-03	6.96252E-03	5.14765E-03	5.04743E-03
90%	4.08380E-03	4.06955E-03	5.54506E-03	4.00936E-03	3.97548E-03
75%	2.53792E-03	2.54641E-03	3.44316E-03	2.49509E-03	2.49502E-03
50%	1.31150E-03	1.31111E-03	1.81922E-03	1.29570E-03	1.29072E-03
25%	5.53138E-04	5.51315E-04	8.03698E-04	5.51385E-04	5.45340E-04
10%	2.02540E-04	2.02684E-04	2.38545E-04	2.02758E-04	2.02248E-04
5%	9.86807E-05	9.90228E-05	1.07833E-04	9.91589E-05	9.89528E-05
Mean	0.001822	0.001819	0.002446	0.001798	0.001779
Percentile	GABA	ACTBP2			
95%	6.96328E-03	5.04641E-03			
90%	5.54528E-03	3.96000E-03			
75%	3.44888E-03	2.47680E-03			
50%	1.82094E-03	1.28562E-03			
25%	8.03794E-04	5.44470E-04			
10%	2.38989E-04	2.02248E-04			
5%	1.07833E-04	9.88129E-05			
Mean	0.002448	0.001775			

Table 6 The estimation of coancestry coefficient (θ) combined for all loci for the Korean population

Method		Estimated	Lower	Upper
Weir & Cockerham (1984)	Last name	0.00653	0.00178	0.20818
	Home	0.00619	0.00169	0.19945
Balding et al. (1996)	Last name	0.000811	0.0001758	0.001843
	Home	0.00338	0.000913	0.007641
Roeder et al. (1998)		0.002192	0.000104	0.006057

Table 7 Forensic casework example in the same subpopulation – a criminal case

G_E	G_S	LR
A_iA_i	A_iA_i	$\frac{(1+\theta)(1+2\theta)}{[2\theta+(1-\theta)p_i][3\theta+(1-\theta)p_i]}$
A_iA_j	A_iA_j	$\frac{(1+\theta)(1+2\theta)}{2[\theta+(1-\theta)p_i][\theta+(1-\theta)p_j]}$

method (Tables 3 and 4) and the Roeder method (Table 5) does not vary substantially over loci. Thus, it looks reasonable to assume the constant coancestry coefficient and combine information across loci to get more precise estimates than the individual estimates. We used the mean value to combine them by assuming that the θ values are the same for all loci (Table 6). As shown in Table 6, the Weir-Cockerham method gives the largest θ value and the Balding method with last name classification gives the smallest value, and the Balding method with home origin classification and the Roeder method give similar values. Note that from Tables 4 and 5, the Balding method with home origin classification gives smaller values than the Roeder method for D21S11, D18S51, D7S820, D8S1179, D5S818, FGA, D12S391 and ACTBP2 and gives larger values for the other loci.

From the coancestry coefficients estimated by using the above three methods, we can see that these values for Korean populations are too large to be ignored although they do not exhibit substantial heterogeneity. In forensic

Table 8 Forensic casework example in the same subpopulation – a paternity case (trio)

G_C	G_M	G_{AF}	PI
A_iA_i	A_iA_i	A_iA_i	$\frac{1+3\theta}{4\theta+(1-\theta)p_i}$
		A_iA_j $i \neq j$	$\frac{1+3\theta}{2(3\theta+(1-\theta)p_i)}$
	A_iA_j $i \neq j$	A_iA_i	$\frac{1+3\theta}{3\theta+(1-\theta)p_i}$
		A_iA_j $i \neq j$	$\frac{1+3\theta}{2(2\theta+(1-\theta)p_i)}$
A_iA_j $i \neq j$	A_iA_i	A_jA_j	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$
		A_iA_j	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$
		A_jA_k $k \neq i, j$	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$
	A_iA_j $i \neq j$	A_iA_i	$\frac{1+3\theta}{4\theta+(1-\theta)(p_i+p_j)}$
		A_iA_j	$\frac{1+3\theta}{4\theta+(1-\theta)(p_i+p_j)}$
		A_jA_k $k \neq i, j$	$\frac{1+3\theta}{2(3\theta+(1-\theta)(p_i+p_j))}$
	A_iA_k $k \neq i, j$	A_jA_j	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$
		A_jA_l $l \neq j$	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$

casework, the estimation of θ values for the population can make the assumption of unrelatedness of two compared persons unnecessary when there is no evidence that the two persons belong to the same family. If so, "what is the best method for the estimation of θ value?" might be naturally asked. As shown above, in order to estimate the exact F_{ST}, the information about the subpopulation labels is required, and the method of Foreman et al. (1997) also requires the number of subpopulations. However, especially in a country with only one ethnic group such as Korea, the researcher's choice for subpopulation labels is difficult to pass by a unanimous vote. Thus, even though it needs the assumption that F_{IT} is the same with F_{ST}, the method of Roeder et al. (1998) seems to be also reasonable because it does not require any assumption about the subpopulation. In using any method discussed above, researchers can choose an appropriate value in a range of estimated values.

Forensic casework and discussion

Forensic casework is generally classified into criminal cases and paternity cases. While, for the criminal case, the genotype of the suspect is compared with the evidence obtained at the crime scene, for the paternity case the genotype of the alleged father is compared with that of the child when the allele that the child must have received from the mother is excluded. All the conventional methods have assumed that the compared persons are unrelated. However, even though the level of coancestry is small, disregarding this effect can exaggerate the strength of the evidence against the compared person (e.g. suspect or alleged father). Moreover, if there is evidence that the two persons belong to the same family, the exaggeration could be more serious. Thus, the various situations for forensic cases where the relationship is considered have been studied (Balding and Nichols 1994, 1995; Balding and Donnelly 1995; Belin et al. 1997; Brookfield 1994; Lee HS et al. 2000; Lee JW et al. 1999, 2001a, 2001b). The simple forensic calculations (*LR*) when the coancestry coefficient is considered are shown in Tables 7 and 8 (Balding and Nichols 1994, 1995).

In order to examine the performance of *LR* depending on the θ values, we calculated the *LR*s by using the mean value of θ for 17 loci in Table 6, respectively. Here, to

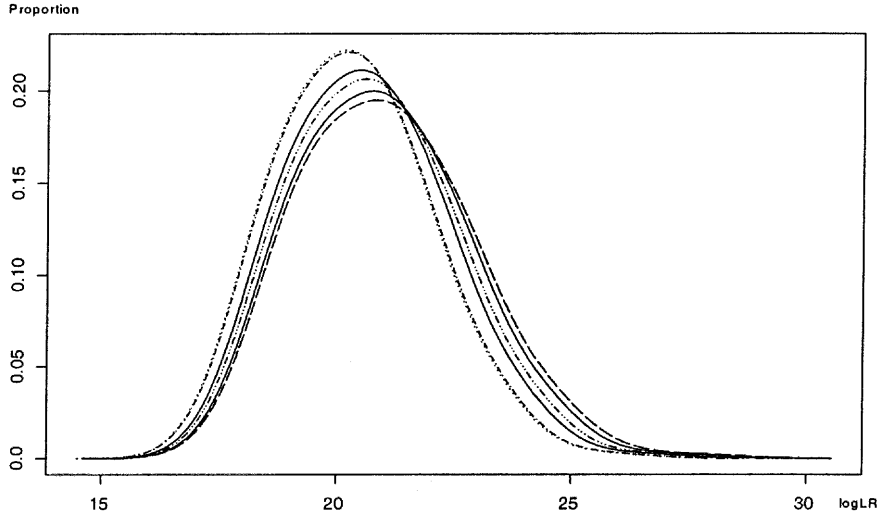

Fig. 1 Probability density for Log10 (LR) in criminal case using θ from each estimation method. Balding et al. (1996) (———), Unrelated ($\theta = 0$) (– – – –), Roeder et al. (1998) (–··–··–), Weir & Cockerham (1984) (-------)

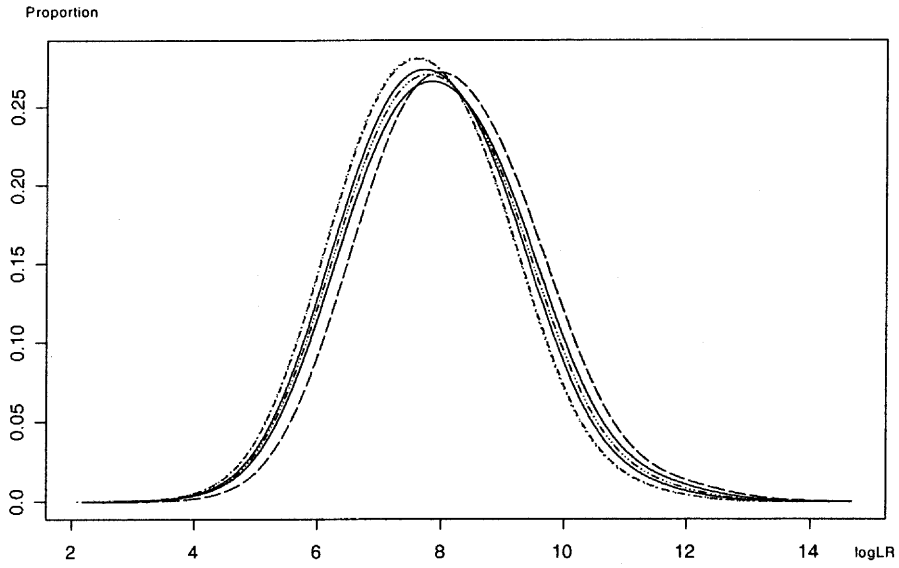

Fig. 2 Probability density of Log10 (LR) in paternity case using θ from each estimation method. Balding et al. (1996) (———), Unrelated ($\theta = 0$) (– – – –), Roeder et al. (1998) (–·–·–·–), Weir & Cockerham (1984) (-------)

simulate forensic case work, 492 criminal cases and 490 paternity trio cases were generated from the original 1,164 persons. The results are shown in Figs. 1 and 2. Figure 1 shows the distribution of the index values *LR* when 492 criminal cases are simulated and Fig. 2 shows the distribution when 490 paternity trio cases are simulated.

As shown here, the larger θ value makes the distribution of the *LR* values shift to the left. The shift is bigger when the criminal cases are considered (Fig. 1). Note that the lower *LR* values give a more favourable decision to the suspect or the alleged father in the court. Also, while the Weir-Cockerham method does not show any difference in the distributions of *LR* between the last name and the home origin subpopulation, the Balding method shows a big difference.

Acknowledgements This work was supported by Korea Research Foundation grant (KRF-99-015-DP0059). We express our sincere thanks to Dr. Karen Ayres for her help to implement the Balding method. We also thank the referees who provided helpful suggestions for revising the original manuscript.

References

Balding DJ, Donnelly P (1995) Inference in forensic identification. J R Stat Soc A 158:21–53

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

Balding DJ, Nichols RA (1995) The method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12

Balding DJ, Greenhalgh M, Nichols RA (1996) Population genetics of STR loci in Caucasians. Int J Legal Med 108:300–305

Belin TR, Gjertson DW, Ming-yi Hu (1997) Summarizing DNA evidence when relatives are possible suspects. J Am Stat Assoc 92:706–716

Brookfield JFY (1994) The effect of relatives on the likelihood ratio associated with DNA profile evidence. Sci Justice 34:193–197

Evett IW, Weir BS (1998) Interpreting DNA evidence. Sinauer Associates, Sunderland Mass

Foreman LA, Smith AFM, Evett IW (1997) A Bayesian approach to validating STR multiplex databases for use in forensic casework. Int J Legal Med 19:244–250

Freeman GH, Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38:141–149

Geyer CJ (1992) Practical Markov Chain Monte Carlo. Stat Sci 7:473–511

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hartl DL, Lewontin RC (1993) NRC report on DNA typing. Science 260:473–474

Lee HS, Lee JW, Han GR, Hwang JJ (2000) Motherless case in paternity testing. Forensic Sci Int 114:57–65

Lee JW, Lee HS, Park M, Hwang JJ (1999) Paternity probability when a relative of father is an alleged father. Sci Justice 39: 223–230

Lee JW, Lee HS, Park M, Hwang JJ (2001a) Evaluation of DNA match probability. Forensic Sci Int 116:139–148

Lee JW, Lee HS, Park M, Hwang JJ (2001b) Paternity determination when the alleged father's genotypes are unavailable. Forensic Sci Int 123:202–210

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

Li Y (1996) Characterizing the structure of genetic population. PhD thesis, North Carolina State University, Raleigh, NC

Nichols RA, Balding DJ (1991) Effects of population structure on DNA fingerprint analysis in forensic science. Heredity 66: 297–302

Roeder K, Escobar M, Kadane JB (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 85:269–297

Weir BS (2001) Forensics. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics. John Wiley, Chichester, pp 721–739

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1951) The genetical structure of populations. Ann Eugen 15:32–354

Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19:395–420