TECHNICAL NOTE

Ziaur Rahman,[1] *Ph.D.; Talat Afroze,*[1,2] *Ph.D.; and B.S. Weir,*[3] *Ph.D.*

DNA Typing Results from Two Urban Subpopulations of Pakistan*

REFERENCE: Rahman Z, Afroze T, Weir BS. DNA typing results from two urban subpopulations of Pakistan. J Forensic Sci 2001;46(1):111–115.

ABSTRACT: A population genetic characterization of the Araeen and Raajpoot ethnic subpopulations of Lahore City, Pakistan was undertaken in order to assess the utility of DNA typing for forensic purposes in Pakistani populations. One hundred unrelated individuals from each group were genotyped for four independently assorting loci: HLA DQA1, CSF1PO, TPOX, and TH01. Allele frequencies were calculated, one- and two-locus tests for association were conducted, and the samples were compared by contingency table tests and *F*-statistic estimation. Although there is expected to be some genetic divergence between the two groups, forensic needs may be satisfied with a single Pakistani database of DNA profiles. The present data suggest that nine independently assorting loci will be sufficient to provide estimated profile probabilities of the order of 10^{-9} but a set of 13 loci, as employed in the U.S., would better compensate for the dependencies introduced by family membership and evolutionary history.

KEYWORDS: forensic science, DNA typing, Pakistan, population structure, Hardy-Weinberg

The use of DNA profiles for Pakistani criminal trials requires a study of the population genetics of these human identification markers in the Pakistani population. Data on the prevalence within the Pakistani population of the various alleles of genes used in human identification have not been available, although there have been studies on samples classed simply as "Pakistani" (1,2). The degree of evolutionary relatedness among the 30 or 40 subpopulations existing within the Pakistani population is not known. Although we consider that there is now ample published data pointing to the robustness of DNA profile probabilities for use with any human population, we undertook the present study to provide information on probabilities for Pakistani resident and expatriate populations.

The Pakistani population is a rich mixture of different ethnic subpopulations that have coexisted in this region of the world for many centuries, having survived numerous demographic disturbances caused by periodic conquests of invading armies from Central Asia, the Middle East, and Europe. The earliest documented conquest of this region was by Alexander the Great, followed much later by the Moghuls in the 16th century and the most recent demographic upheaval occurred about 50 years ago with the Partition of 1947 whereby the Indian sub-continent was divided into the two independent states of India and Pakistan at the end of British rule. The major demographic trend since the 1970's in the Pakistani urban populations has been a constant influx of migrants from rural areas of Pakistan and, from the mid-1980's, from Afghanistan.

The Pakistani population is highly structured, with subdivision into various ethnic subpopulations. In rural areas, people invariably marry within their own subpopulation. This trend is replicated to some extent in urban areas, but there has been a lot of ethnic mixing as well, especially after 1947. There are no population genetic data on the extent of ethnic mixing in Pakistani urban areas.

Pakistani subpopulations have a relatively high level of inbreeding: first-cousin marriages form a significant percentage of all marriages, especially in rural areas. One study (3) found 411 of 681 married women in hospital wards in Lahore to be married to various classes of cousins, with an average coancestry between spouses of 0.038. This coancestry coefficient is also the inbreeding coefficient of the next generation (4). A more recent study (5) reported 60% of marriages for 1011 ever-married women living in four multi-ethnic and multi-lingual squatter settlements of Karachi to be consanguineous, and over 80% of these to be between first cousins, leading to an inbreeding coefficient of 0.032. The question of forensic interest is whether this level of inbreeding results in detectable departures from the Hardy-Weinberg law.

It is in this context that we selected the Araeen and Raajpoot ethnic subpopulations of Lahore, the second biggest city in Pakistan with an estimated population of six million. These two subpopulations together may constitute 20 to 40% of the total population of Lahore City. Blood samples were collected from 100 unrelated Araeens and 100 unrelated Raajpoots and these samples were genotyped for four independently assorting loci: HLA DQA1 (henceforth written as DQA1), CSF1PO, TPOX, and TH01.

Materials and Methods

Blood Samples

Individuals belonging to either Araeen or Raajpoot subpopulations of Lahore and either born in Lahore City or residing in Lahore for the past ten years were screened by interview at blood banks of a few major hospitals or by a 45-min seminar delivered to four

[1] Centre of Excellence in Molecular Biology, Canal Bank Road, Thokar Niaz Baig, Lahore-53700, Pakistan.

[2] Present Address: MBRC 3-900, 101 College Street, Toronto General Hospital, Toronto, M5G 2C4, Canada.

[3] Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC.

* This work was supported in part by NIH Grant GM 45344.

Received 23 April 1999; and in revised form 13 Jan. 2000; accepted 31 Feb. 2000.

large colleges in the city. A total of 100 unrelated Araeen persons and 100 unrelated Raajpoot persons were sampled. Three to 5 mL of whole blood was collected with informed written consent into sterile tubes containing 70 μL of 0.5 M ethylene diamine tetraacetic acid (EDTA), pH 8.0. Part of the blood was frozen down within 24 h of collection in 700 μL aliquots at −70°C and the rest was spotted (30 μL per spot) onto autoclaved cotton strips which were air dried and also stored frozen at −70°C.

Genomic DNA Extraction

High molecular weight human genomic DNA was isolated from whole blood by an FBI protocol (6) involving selective hypo-osmotic lysis of red blood cells, a 2 h Proteinase K digestion on a turning wheel at 56°C followed by phenol-chloroform extraction and ethanol precipitation. The concentration of the resuspended genomic DNA was quantitated by resolution on a 0.4% agarose gel along with human genomic DNA mass standards (K562 DNA Quantitation Standards; Gibco/Life Technologies, Gaithersburg, MD).

DQA1 Genotyping

All samples were genotyped for the DQA1 locus (six alleles, chromosome 6). Two ng of human genomic DNA was used in the Polymerase Chain Reaction (PCR) to amplify a portion of the human DQA1 gene by using a DQA1 genotyping kit (AmpliType HLA DQα Typing Kit; PE Applied Biosystems/Roche Molecular Systems, Branchburg, NJ) according to the manufacturer's instructions. The biotinylated PCR product was hybridized to nylon strips containing immobilized DNA corresponding to various alleles of this gene viz. alleles 1.1, 1.2, 1.3, 2, 3, and 4. Allele scoring was done by color development of the nylon strips. All developed strips were photographed. This kit did not discriminate among DQA1 alleles 4.1, 4.2 and 4.3 and grouped all three as allele 4. Alleles 4.1, 4.2, and 4.3 are quite common in the Lahore subpopulations we studied, so combining all three alleles as "4" may obscure subpopulation differences. Allele frequencies are given in Table 1.

Short Tandem Repeat Loci Genotyping

All samples were genotyped for three independently assorting human short tandem repeat (STR) loci: CSF1PO (10 alleles; chromosome 5), TPOX (8 alleles; chromosome 2), and TH01 (8 alleles; chromosome 11) by using a commercial PCR-based kit (GenePrint STR Multiplex System CTT Kit; Promega, Madison, WI) according to the manufacturer's instructions. Human genomic DNA was used to simultaneously amplify CSF1PO, TPOX, and TH01 STR loci in a single tube. The concentration of the PCR products was determined by agarose gel estimation using appropriate DNA mass standards. Exactly 20 ng of the PCR products from each Araeen or Raajpoot sample were resolved along with allelic ladders for all three loci on a 5% polyacrylamide gel followed by silver staining and documentation of the gel image with a video imaging system linked to a computer. Allele scoring was done by matching the mobility of the stained amplified bands with the allelic ladders. Allele frequencies are given in Table 1.

Statistical Analysis

Tests for Association

A population with a substantial degree of inbreeding is expected to show departures from Hardy-Weinberg, with more homozygotes than expected. Exact tests for Hardy-Weinberg (4) were conducted at each locus, and the *p*-values are shown in Table 2. These are the probabilities of the observed set of genotype counts (or a set with greater departures from Hardy-Weinberg) conditional on the allele counts and on the assumption of Hardy-Weinberg. Small values would suggest departures from Hardy-Weinberg, but were not found in either of the two samples. The values for DQA1 are greater than those reported for Pakistanis living in Abu Dhabi (1), but those for TH01 are comparable to the value reported in (2).

Although the loci typed in this study are unlinked, it is possible that allelic frequencies at different loci are dependent. Exact tests for association among all four alleles at pairs of loci were conducted (4). These tests compare two-locus genotype counts with those expected under complete allelic independence, so include an examination of Hardy-Weinberg relationship. The *p*-values for these tests are shown in Table 3. In each of the two samples, one of the six tests has a low value.

TABLE 1—*Allele frequencies.*

		Sample		
Locus	Allele	Raajpoot	Araeen	Combined
CSF1PO	8	.000	.010	.005
	9	.040	.055	.048
	10	.215	.270	.242
	11	.345	.335	.340
	12	.350	.270	.310
	13	.040	.055	.048
	14	.010	.005	.007
THPOX	6	.000	.010	.005
	7	.005	.015	.010
	8	.390	.370	.380
	9	.125	.100	.113
	10	.055	.100	.077
	11	.355	.380	.367
	12	.070	.025	.048
TH01	6	.325	.240	.282
	7	.165	.255	.210
	8	.160	.105	.133
	9	.215	.250	.232
	9.3	.130	.135	.133
	10	.005	.015	.010
DQA1	1.1	.185	.160	.172
	1.2	.095	.045	.070
	1.3	.210	.165	.187
	2	.150	.100	.125
	3	.085	.050	.068
	4	.275	.480	.378

TABLE 2—*One-locus test* p-*values.*

	Sample		
Locus	Raajpoot	Araeen	Combined
CSF1PO	0.10	0.10	0.05
TPOX	0.39	0.14	0.10
TH01	0.55	0.56	0.11
DQA1	0.79	0.44	0.51
Over 8 tests	0.24		

Care needs to be taken when multiple tests are performed. In this study, there are eight single-locus tests, and 12 two-locus tests. Although each is of interest in its own right, there is also interest in the overall hypothesis that all of the eight, or 12, hypotheses are true. Fisher (7) gave a procedure for combining *p*-values, and these overall *p*-values are shown in Tables 2 and 3. These calculations confirm the lack of detectable departures from Hardy-Weinberg, and the presence of two detectable departures from two-locus independence.

Tests for Population Structure

Substructure within either the Raajpoot or Araeen populations could lead to evidence of allelic association (4), but single-locus associations were not detected. Differences in allelic frequencies between the two samples, however, can lead to evidence of allelic association when the two samples are combined into a single sample. Indeed, the *p*-values for the Hardy-Weinberg tests are all smaller in the combined sample than in either sample separately, with the CSF1PO result suggesting significant association at the 5% level (Table 2). At pairs of loci, four of the six *p*-values are 0.05 or less.

TABLE 3—*Two-locus test* p-*values.*

	Sample		
Locus	Raajpoot	Araeen	Combined
CSF1PO, TPOX	0.28	0.02	0.01
CSF1PO, TH01	0.11	0.24	0.01
CSF1PO, DQA1	0.29	0.40	0.03
TPOX, TH01	0.00	0.22	0.05
TPOX, DQA1	0.40	0.44	0.33
TH01, DQA1	0.34	0.45	0.45
Over 12 tests	0.00		

A purely statistical means of comparing two samples is to compare their allelic counts with a goodness-of-fit test and a chi-square test statistic (4). Calculations are set out in Table 4. Only the DQA1 locus gave a significant value ($\chi^2 = 19.96$ with 5 df). A genetic approach is to estimate the coancestry coefficient θ (or F_{ST}) between the two populations (4). These values are 0.002, 0.000, 0.007 and 0.028 for loci CSF1PO, TPOX, TH01, and DQA1, and are consistent with the goodness-of-fit test results. It is interesting to note that the relatively large DQA1 value did not lead to a detectable departure from Hardy-Weinberg when the two samples were combined. The four-locus average value of θ is 0.009, which is very close to an often-recommended value of 0.01 (8). However, this ignores the very long tail to the right of the distribution of θ (9) and 0.03 may be a more appropriate value to use in conditional probability equations.

How Many Loci are Needed?

Although this study has made use of only four markers, more are available. A natural question to ask is: how many are sufficient for forensic purposes? In other words, how many markers will be sufficient to provide a high probability of being able to distinguish between people, or to provide a high probability of there not being another person in the population with the same profile? For a single marker with allele frequencies p_i, the average profile has a probability of

$$P = \sum_i p_i^2 \,(p_i^2) + \sum_{i<j} 2p_ip_j \,(2p_ip_j)$$

$$= \left(\sum_i p_i^2\right)^2 - \sum_i p_i^4$$

This is also the probability that two people, chosen randomly, will have the same profile. For the four markers in this study, the com-

TABLE 4—*Comparison of sample allele counts.*

	CSF1PO			TPOX			TH01			DQA1		
	Allele	Raajpoot	Araeen	Allele	Raajpoot	Araeen	Allele	Raajpoot	Araeen	Allele	Raajpoot	Araeen
O*	8	.00	2.00	6	.00	2.00	6	65.00	48.00	1.1	37.00	32.00
E		1.00	1.00		1.00	1.00		56.50	56.50		34.50	34.50
X		1.00	1.00		1.00	1.00		1.28	1.28		.18	.18
O	9	8.00	11.00	7	1.00	3.00	7	33.00	51.00	1.2	19.00	9.00
E		9.50	9.50		2.00	2.00		42.00	42.00		14.00	14.00
X		.24	.24		.50	.50		1.93	1.93		1.79	1.79
O	10	43.00	54.00	8	78.00	74.00	8	32.00	21.00	1.3	42.00	33.00
E		48.50	48.50		76.00	76.00		26.50	26.50		37.50	37.50
X		.62	.62		.05	.05		1.14	1.14		.54	.54
O	11	69.00	67.00	9	25.00	20.00	9	43.00	50.00	2	30.00	20.00
E		68.00	68.00		22.50	22.50		46.50	46.50		25.00	25.00
X		.01	.01		.28	.28		.26	.26		1.00	1.00
O	12	70.00	54.00	10	11.00	20.00	9.3	26.00	27.00	3	17.00	10.00
E		62.00	62.00		15.50	15.50		26.50	26.50		13.50	13.50
X		1.03	1.03		1.31	1.31		.01	.01		.91	.91
O	13	8.00	11.00	11	71.00	76.00	10	1.00	3.00	4	55.00	96.00
E		9.50	9.50		73.50	73.50		2.00	2.00		75.50	75.50
X		.24	.24		.09	.09		.50	.50		5.57	5.57
O	14	2.00	1.00	12	14.00	5.00						
E		1.50	1.50		9.50	9.50						
X		.17	.17		2.13	2.13						
		$\chi^2 = 6.62$(6df) Not Significant			$\chi^2 = 10.71$(6df) Not Significant			$\chi^2 = 10.24$(5df) Not Significant			$\chi^2 = 19.96$(5df) Significant at 0.005	

* O:observed; E:expected; X:$(O-E)^2/E$.

bined sample values of P are 0.125, 0.141, 0.079, 0.085 with a geometric average very close to 0.10.

Two individuals can be distinguished if they differ for at least one locus. The probability of not being able to distinguish two individuals, therefore, is the probability that they have the same genotype at all loci. This is $\Pi_l P_l$, where P_l is the probability of the same genotype at the lth locus. For markers like those studied here, the product will be close to $(0.1)^m$ for m loci. A value of 10^{-9} would require 9 loci.

A report of the U.S. National Research Council (NRC) (8) suggested that a DNA profile could be considered unique in a population if there was only a small probability of there being a second occurrence of the profile. For a profile with probability P and a population of size N, this led the NRC to suggest that P should be, approximately, α/N in order for the probability of a second occurrence to be α. Typical values of α might be 0.001, and N of the order of 10^6, leading to P being of the order of 10^{-9} − again suggesting the use of $m = 9$. The problem with this argument is that all profiles in a population are assumed to be independent. This cannot be true, especially in Pakistani populations where cousin marriages are common. It is suggested that more than 9 loci be used to allow for the dependencies introduced by both family and evolutionary relatedness. The use of a core set of 13 loci by the FBI in the U.S. appears to be a reasonable precaution. Further discussion of this issue is contained in (4,10,11).

Discussion

We embarked on the present population genetic characterization of two cohabiting ethnic subpopulations of Lahore, Pakistan in order to investigate the establishment of population genetic databases for the urban Pakistani population in general, and to investigate the use of "Pakistan" databases for expatriate populations. Establishment of such databases would be a crucial step in the successful application of DNA typing for violent crimes in Pakistan and countries with Pakistani immigrants. There is a question of whether one general database for the population at large would suffice or would it be necessary to set up 30 to 40 databases, one for each ethnic subpopulation. In other words, are the subpopulations so divergent genetically that one would have to know which ethnic subpopulation was relevant for a particular crime and then use allele frequencies from that particular subpopulation?

We decided to focus on the Muslim (Moslem) subpopulations of Lahore City where the most recent demographic events have been mass migration of Muslim refugees from the eastern part of the subcontinent during 1947 and urbanization from rural areas of Punjab province from the 1970's onwards. Analysis of data from two ethnic subpopulations which have cohabited in the city for at least the past 400 years was aimed at three questions:

- Are the subpopulations in Hardy-Weinberg equilibrium or is inbreeding so high that equilibrium would not be a reasonable assumption?
- How many (STR) loci would one need to analyze for crime scene and suspect samples so that the probability of a person other than the suspect having a particular genotype would be low?
- Would one have to use allele frequencies empirically determined for each specific subpopulation in order to analyze crime casework or would the Araeen and Raajpoot subpopulations turn out to be so similar that allele frequencies from either subpopulation could be used interchangeably?

The data collected in this study, although not extensive, confirm the general robustness of forensic DNA databases. There is evidence of significantly different allelic frequencies only at the DQA1 locus, and even then, the two sets of frequencies found in this study are very close to those reported for 100 Pakistanis living in the United Arab Emirates (1) (the frequency for allele 1.2 in Table 2 of that paper should be 0.130 instead of 0.297). In 43 of 48 published studies (12), the "4" allele was the most frequent as it is for the two samples described here. More importantly, the estimated θ value from this worldwide survey was 0.042.

As has been stressed in Ref 4, it is appropriate to focus on conditional profile probabilities when interpreting DNA profiles: what is the probability that an unknown person has a profile when it is known that one person (often the defendent in a trial) has that profile? In equations endorsed by the U.S. NRC (8), these conditional probabilities involve *population* allele frequencies and θ. The results apply on average to any subpopulation within that population, and they eliminate the need to have data from all subpopulations. The NRC suggested a θ value of 0.01 or 0.03 and either value appears to be sufficiently high for a major racial group. The even more conservative value of 0.05 should allow worldwide frequencies to be used for almost any subpopulation. The only question remaining would be whether to use local Pakistani or worldwide frequencies. The local frequencies may prove easier to defend in local courts, but it is doubtful if the confidence intervals around the results from either set of frequencies would exclude the result from the other.

It has already been noted that other studies (3,5) have estimated within-population inbreeding levels in Pakistan to be 0.038 and 0.032 on the basis of reported frequencies of cousin marriages. The value based on all four loci and both samples in this study is 0.038 (using the methods described in (4)), and this could be used to estimate profile probabilities according to Recommendation 4.1 in the NRC report (8). It is acknowledged that tests for Hardy-Weinberg would not be expected to detect inbreeding levels of the order of 0.03 with samples of size 100, so that the non-significant results in Table 2 are not surprising. However, the tests are powerful enough to detect departures from independence sufficiently large to have a meaningful forensic effect on probabilities estimated on the assumption of independence (5). Nonetheless, the issue is somewhat moot. The use of the "θ" equations for conditional probabilities (4,8) carries an implicit assumption of allelic dependence in a population due to population structure, and obviates the need for testing allelic independence within populations.

The present study suggests that Araeen and Raajpoot data from Lahore City can be used to estimate probabilities for either group, and reinforces the belief that any published frequencies for human identification loci can be used for forensic purposes in Pakistan populations, inside or outside that country, provided a sufficiently high value of the population structure parameter θ is used.

References

1. Tahir M, Caruso J, Budowle B, Novick GE. Distribution of HLA-DQA1 and polymarker (LDLR, GYPA, HNBB, D7S8, and GC) alleles in Arab and Pakistani populations living in Abu Dhabi, United Arab Emirates. J Forensic Sci 1997;42:914–8.
2. Evett IW, Gill PD, Scranage JK, Weir BS. Establishing the robustness of short-tandem-repeat statistics for forensic applications. Am J Hum Genet 1986;58:398–407.
3. Shami SA, Schmitt LH, Bittles AH. Consanguinity, spousal age at marriage and fertility in seven Pakistani Punjab cities. Ann Hum Biol 1990;17:97–105.
4. Evett IW, Weir BS. Interpreting DNA evidence. Sunderland, MA: Sinauer, 1998.

5. Hussain R, Bittles AH. The prevalence and demographic characteristics of consanguineous marriages in Pakistan. J Biosoc Sci 1998;30:261–75.
6. Kirby LT. DNA fingerprinting: an introduction. London: Oxford University Press, 1997.
7. Fisher RA. Statistical methods for research workers. London: Oliver and Boyd, 1932.
8. National Research Council. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press, 1996.
9. Balding DJ, Nichols RA. Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 1997;78:583–9.
10. Weir BS. Are DNA profiles unique? Proceedings of Ninth International Symposium on Human Identification. Madison, WI: Promega, 1999;114–7.
11. Balding DJ. When can a DNA profile be regarded as unique? Science Justice 1999;39:257–60.
12. Weir BS et al. A survey of forensic databases: AmpliType™ and DQA1. In press.

Additional information and reprint requests:
B.S. Weir, Ph.D.
Box 8203
Raleigh, NC 27695-8203