Chapter Four

DISCUSSION AND CONCLUSION

CHAPTER FOUR: DISCUSSION

4.1 Introduction

The basis of most of the current DNA profiling or fingerprinting systems mainly relies on the polymorphism provided by repetitive sequences within the human genomic DNA, namely the minisatellite and microsatellite loci. The high variability within these polymorphic loci is primarily due to variation in repeat copy numbers (VNTR), which can be detected by conventional methods of Southern blotting or assisted by PCR-based techniques (Jeffreys *et al.*, 1991). Whilst these techniques have been reasonably established and accepted by the general public, there exist several limitations in terms of interpreting the data generated by conventional electrophoresis and DNA band analysis systems (see Section 1.4.4).

4.2 Minisatellite variant repeat mapping

Recently, Prof. Alec Jeffreys had overcome the limitations of the current profiling system and described a derivative of DNA fingerprinting method. This approach is called minisatellite variant repeat (MVR) mapping and is based on the analysis of a second level of variability within the internal sequence content of minisatellite repeat units (Jeffreys *et al.*, 1991; Neil and Jeffreys, 1993). MVR-PCR mapping of two minisatellite loci, D1S8 (MS32) and D7S21 (MS31), revealed that two major types of repeat-unit sequence variants can be recognised, and they intersperse within the minisatellite loci (Jeffreys *et al.*, 1991; Neil and Jeffreys, 1993).

PCR amplification with variant-specific primers and a fixed primer located upstream of the minisatellite sequence generated a mixture of amplified products differing in lengths, depending on the variant-specific primer annealing sites along the minisatellite, therefore revealing the minisatellite repeat variant array (see Section 1.4.4 and Figure 3). It has been suggested that this minisatellite variant array might provide an unlimited source of individual specific polymorphisms, which would be the ideal identification system. More importantly, the data interpretation process no longer relies on visual assessment of electrophoretic bands. Instead, a digital approach could be utilised to represent individual-specific array of repeat-unit variants, which would be more discriminatory for identification purposes.

The minisatellite MS31 has several advantages for MVR-PCR studies compared to MS32 and MS205. MS31 had been shown to contain a more even balance of the two repeat unit types assayed (a type and t type) and have a better interspersion pattern of variant repeat units, with fewer clusters of a particular repeat type. Studies have also shown that MS31A possesses a heterozygosity of at least 99.9% (Neil and Jeffreys, 1993), thus the diploid codes are generally more informative than those of MS32. Flanking base substitutional polymorphisms have enabled the 5' structure of a large number of MS31A alleles to be derived from genomic DNA by allele-specific MVR-PCR. In addition, the MS31 locus has properties that satisfy the criteria for MVR-PCR, and they include:

i) highly polymorphic, with an allele length heterozygosity greater than 95% (this ensures that most or all alleles are rare),

ii) repeat unit heterogeneity is not too extensive,

iii) the sites of variation are suitably positioned to allow the design of repeat unit specific primers, and

iv) all primers used for MVR mapping have been shown to work at the discriminatory annealing temperature of the MVR-specific primers.

As mentioned in Section 1.4.4, to obtain definitive MVR-PCR data, one has to be able to characterise the variation within a single allele (allele-specific MVR-PCR). This requires the use of allele-specific primers, which would produce unambiguous information on the array of repeat variants within a minisatellite locus on a particular chromosome. For this purpose, much effort has been channelled towards characterising the sequence content upstream of the minisatellite region (5' flanking region) in search of a suitable site for allele-specific primer design, which has also been the main aim of conducting the present study.

4.3 Flanking substitutional polymorphic site analyses

Three flanking substitutional polymorphic sites (*Alu*I, *Hga*I, and *Psp*1406I) of MS31 (D7S21) were analysed in this study. *Alu*I and *Hga*I polymorphisms have been suggested to be common in the Malaysian, Caucasian and Japanese populations in previous studies conducted by Neil and Jeffreys (1993) and Koh *et al.* (1993, 1994).

For *Alu*I and *Hga*I assays, results obtained from the Chi-square analysis were comparable to those reported by Neil and Jeffreys (1993) who studied the Caucasian and Japanese populations, and by Koh *et al.* (1993, 1994) who studied the Malaysian population. As mentioned earlier, *Psp*1406I is a new site studied for MS31A. Chi-

square results for *Psp*1406I showed similar results as in *Alu*I and *Hga*I assays. These three polymorphic sites exhibited Hardy-Weinberg equilibrium.

The frequencies of heterozygous individuals obtained from the *Alu*I assay were the lowest (ranging from 28 to 34%) compared to the *Hga*I assay (ranging from 44 to 51%) and the *Psp*1406I assay (ranging from 43 to 51%). Normally, the expected maximum frequency of heterozygous individuals obtainable in a population in the absence of selection is 50% (Jeffreys, 1987).

The combined heterozygosity over each site (*Alu*I, *Hga*I, and *Psp*1406I) indicated that up to 70% (212/310 x 100%) of the Malaysian individuals were heterozygous at one or more of the flanking polymorphism sites, compared to 51% in Caucasian and 59% in Japanese individuals (Neil and Jeffreys, 1993). This suggests more that half of Malaysian population can be mapped by allele-specific MVR-PCR (Monckton *et al.*, 1993).

Table 10 shows the distributions of alleles (*Alu*I + and -, and *Hga*I + and -) in the Malaysian, Japanese, and Caucasian population samples. The results obtained in this study for the Malaysian population samples showed little deviation from those reported by Koh *et al.* (1993, 1994).

Pairwise comparisons by the heterogeneity G-test for the *Alu*I polymorphism site did not reveal any significant deviation between the races (Table 11) except between Chinese/Caucasian.

However, pairwise comparisons for the *Hga*I polymorphism site indicated that significant differences existed between Malay/Chinese, Chinese/Indian, Indian/Japanese, and Caucasian/Japanese (Table 11).

For the *Psp*1406I polymorphism assay, data from the Caucasian and Japanese populations are not available for comparison. The distributions of the *Psp*1406I + and - alleles in the Malays, Chinese and Indians are given in Table 12.

Pairwise comparisons for the *Psp*1406I polymorphism site (Table 13) indicate that only the Malays and Chinese shared similar distribution, and that significant differences were noted when both these races were compared to the Indians.

Table 10: Comparison of the allele frequencies at *Alu*I and *Hga*I sites in the 5' flanking DNA of MS31A among the Malaysian, Caucasian, and Japanese population samples.

Locus	Allele	Malaysian						Caucasian *		Japanese *	
		Malay		Chinese		Indian					
		Freq.	No.	Freq.	No.	Freq.	No.	Freq.	No.	Freq	No.
*Alu*I	+	0.22 (0.23)	46	0.25 (0.23)	54	0.20 (0.18)	40	0.13	22	0.25	48
	-	0.78 (0.77)	160	0.75 (0.77)	158	0.80 (0.82)	162	0.87	142	0.75	148
*Hga*I	+	0.61 (0.72)	125	0.70 (0.64)	149	0.54 (0.51)	110	0.58	95	0.7	137
	-	0.39 (0.28)	81	0.30 (0.36)	63	0.46 (0.49)	92	0.42	69	0.3	59

Note: -*Based on data from Neil and Jeffreys (1993), who investigated 82 unrelated Caucasians and 98 unrelated Japanese.

Freq. = Frequency and

No.= Number.

- Numbers in brackets denote the frequencies reported by Koh *et al.* (1993, 1994).

Table 11: Pairwise comparisons by the heterogeneity G-test of the distributions of *Alu*I + and - alleles and *Hga*I + and - alleles between different population samples.

Pairwise comparison	*Hga*I (+ and -)			*Alu*I (+ and -)		
	G_H	df	p	G_H	df	p
Malay vs Chinese	4.274*	1	0.05>p>0.025	0.566	1	0.50>p>0.30
Malay vs Indian	1.618	1	0.25>p>0.20	0.390	1	0.70>p>0.50
Malay vs Caucasian	0.228	1	0.70>p>0.50	1.904	1	0.20>p>0.10
Malay vs Japanese	3.774	1	0.10>p>0.05	3.124	1	0.10>p>0.05
Chinese vs Indian	11.112*	1	p<0.001	0.174	1	0.70>p>0.50
Chinese vs Caucasian	3.09	1	0.10>p>0.05	5.378*	1	0.025>p>0.02
Chinese vs Japanese	0.006	1	0.95>p>0.90	0.034	1	0.90>p>0.80
Indian vs Caucasian	0.444	1	0.70>p>0.50	1.690	1	0.20>p>0.10
Indian vs Japanese	10.136*	1	0.005>p>0.001	0.850	1	0.50>p>0.30
Caucasian vs Japanese	5.578*	1	0.02>p>0.01	3.582	1	0.10>p>0.05

$G_H = \chi^2$ of the heterogeneity G-test

df = Degree of freedom

p = Probability

* = Significant in the distribution of alleles between two population samples.

Note: Appendix A shows the method of calculations (Sokal and Rohlf, 1981).

Table 12: Comparison of the allele frequencies at the *Psp*1406I sites in the 5' flanking DNA of MS31A among the Malaysian population samples.

Locus	Allele	Malaysian					
		Malay		Chinese		Indian	
		Freq.	No.	Freq.	No.	Freq.	No.
*Psp*1406I	+	0.379	78	0.434	92	0.287	58
	-	0.621	128	0.566	120	0.713	144

Table 13: Pairwise comparisons by the heterogeneity G-test of the distribution of *Psp*1406I + and – alleles between different population samples.

Pairwise comparison	*Psp*1406I (+ and -)		
	G_H	df	p
Malay vs Chinese	1.326	1	0.25>p>0.20
Malay vs Indian	3.854*	1	0.05>p>0.025
Chinese vs Indian	9.712*	1	0.005>p>0.001

$G_H = \chi^2$ of the heterogeneity G-test

df = Degree of freedom

P = Probability

* = Significant in the distribution of alleles between two population samples

Note: 1. No *Psp*1406I data for the Caucasian and Japanese populations are available for comparison.

2. Appendix A shows the method of calculations.

4.4 Haplotype assays

Haplotype assays were performed to determine the alleles at the three polymorphic sites. From the results, the association between two or among three polymorphic sites was determined.

Tables 8 and 9 show that significant associations occurred between alleles at the *Alu*I and *Hga*I sites and at the *Hga*I and *Psp*1406I sites. These observations indicated that the three restriction endonuclease sites were not totally independent of each other.

The haplotype distributions at the *Hga*I and *Alu*I sites of five ethnic groups were calculated (Table 14) and used for pairwise comparisons between the different groups (Table 15). The heterogeneity G-test results revealed that only Malays/Chinese and Malays/Indians showed a similar distribution of haplotype frequency.

For *Hga*I-*Psp*1406I assay, pairwise comparisons could only be carried out between the different races in the Malaysian population since data on *Psp*1406I from the Caucasian and Japanese populations are not available. The results indicated that the haplotype distributions were different except between the Malays and Indians (Tables 16 and 17).

Table 14: Distribution of the four haplotypes at the MS31A 5' flanking *Hga*I and *Alu*I sites in the Malay, Chinese, Indian, Caucasian, and Japanese population samples.

Haplotype	Malaysian						Caucasian *		Japanese *	
-220 -4	Malay		Chinese		Indian					
	Obsd	f	Obsd	f	Obsd	f	Obsd	f	Obsd	f
- +	10	0.049	10	0.047	11	0.054	20	0.122	39	0.199
+ +	36	0.175	43	0.203	24	0.119	1	0.006	9	0.046
- -	66	0.320	54	0.255	80	0.396	75	0.457	98	0.500
+ -	94	0.456	105	0.495	87	0.431	68	0.415	50	0.255
Total	206	1.000	212	1.000	202	1.000	164	1.000	196	1.000

Obsd = observed number

f = frequency

* Based on data from Neil and Jeffreys (1993).

Table 15: Pairwise comparisons by the heterogeneity G-test of the distributions of four haplotypes at the *Hga*I and *Alu*I sites between different population samples.

Pairwise comparison	G_H	df	p
Malay vs Chinese	2.342	3	0.70>p>0.50
Malay vs Indian	4.038	3	0.30>p>0.25
Malay vs Caucasian	45.482*	3	p<0.001
Malay vs Japanese	55.384*	3	p<0.001
Chinese vs Indian	12.036*	3	0.01>p>0.005
Chinese vs Caucasian	60.114*	3	p<0.001
Chinese vs Japanese	74.750*	3	p<0.001
Indian vs Caucasian	289.306*	3	p<0.001
Indian vs Japanese	35.548*	3	p<0.001
Caucasian vs Japanese	16.564*	3	p<0.001

$G_H = \chi^2$ of the heterogeneity G-test

df = Degree of freedom

p = Probability

* = Significant in the distributions of alleles between different population samples

Note: Appendix B shows the method of calculations (Sokal and Rohlf, 1981).

Table 16: Distribution of the four haplotypes at the MS31A 5' flanking *Hga*I and *Psp*1406I sites in Malays, Chinese, and Indians from Malaysia.

Haplotype	Malay		Chinese		Indian	
-220 -108	Obsd	f	Obsd	f	Obsd	f
+ -	58	0.282	79	0.373	64	0.317
- +	12	0.058	21	0.099	10	0.050
+ +	70	0.340	71	0.335	47	0.233
- -	66	0.320	41	0.193	81	0.401
Total	206	1.000	212	1.000	202	1.000

Table 17: Pairwise comparisons by the heterogeneity G-test of the distributions of four haplotypes at the *Hga*I and *Psp*1406I sites between different population samples.

Pairwise comparison	G_H	df	p
Malay vs Chinese	11.532*	3	0.01>p>0.005
Malay vs Indian	6.522	3	0.10>p>0.05
Chinese vs Indian	23.602*	3	p<0.001

$G_H = \chi^2$ of the heterogeneity G-test

df = Degree of freedom

p = Probability

* = Significant in the distributions of alleles between different population samples

Note: No data for the *Alu*I and *Psp*1406I studies within the Caucasian and Japanese population are available for comparison.

Furthermore, the distributions of the eight possible haplotypes at the three polymorphic sites within the races were also determined. However, haplotyping of alleles at the *Alu*I and *Psp*1406I sites cannot be performed since allele specific primers for either of these sites are not available for this study. Therefore, the haplotypes spanning these three polymorphic sites were inferred from haplotyping results obtained from *Hga*I-*Alu*I and *Hga*I-*Psp*1406I studies.

However, this has reduced the number of individuals available for study from 310 to 282. The haplotypes for the remaining 28 individuals could not be ascertained.

Table 18 shows that the Chi-square value from the data was significantly higher than that at df=7, $p<0.001$. This indicated that significant linkage equilibrium occurred among the three polymorphic sites, even though sometimes incomplete linkage disequilibrium did occur among them.

Data from Table 19 were used to carry out pairwise comparisons between the three ethnic groups from Malaysia, as shown in Table 20. No significant differences were noted except for Chinese/Indian.

Table 18: χ^2 analysis of eight haplotypes at the MS31A flanking polymorphic sites in the Malaysian population.

Haplotype *Hga*I-*Alu*I-*Psp*1406I	Observed	Freq.Observed	Expected	$(O-E)^2/E$
+ + +	59	0.105	24.156	50.253
+ + -	23	0.041	42.026	8.613
+ - +	110	0.195	101.006	0.801
+ - -	151	0.268	175.724	3.479
- + +	10	0.018	15.575	1.996
- + -	17	0.030	27.095	3.761
- - +	27	0.048	65.123	22.317
- - -	167	0.296	113.295	7.788
Total	564	1.00	564	99.008

No. of alleles at *Hga*I: + = 343 (0.608) - = 221 (0.392)

No. of alleles at *Alu*I: + = 109 (0.193) - = 455 (0.807)

No. of alleles at *Psp*1406I: + = 206 (0.365) - = 358 (0.635)

The expected numbers were calculated by assuming random association between *Hga*I, *Alu*I, and *Psp*1406I. The example of calculation:

$$E(+ + -) = (0.608) \times (0.193) \times (0.635) \times 564$$

$$= 42.026$$

The Chi-square value is 99.008, greater than the χ^2 values at both 5 and 1% levels of significance. At df=7, $p<0.001$, $\chi^2 = 24.32$.

Table 19: Distribution of eight haplotypes at the MS31A 5' flanking polymorphic sites in Malays, Chinese, and Indians from Malaysia.

Haplotype *Hga*I-*Alu*I-*Psp*1406I	Malay		Chinese		Indian	
	Obsd.	f	Obsd.	f	Obsd.	f
+ + +	21	0.113	22	0.116	16	0.085
+ + -	5	0.027	11	0.058	7	0.037
+ - +	41	0.220	39	0.206	28	0.148
+ - -	48	0.258	54	0.286	47	0.264
- + +	2	0.011	7	0.037	2	0.011
- + -	8	0.043	3	0.016	7	0.037
- - +	8	0.043	14	0.074	8	0.042
- - -	53	0.285	39	0.206	74	0.392
Total	186	1.000	189	1.000	189	1.000

Table 20: Pairwise comparisons by the heterogeneity G-test of the distributions of eight haplotypes at the *Hga*I, *Alu*I, and *Psp*1406I sites between different Malaysian populations samples.

Pairwise comparison	G_H	df	p
Malay vs Chinese	11.804	7	0.20>p>0.10
Malay vs Indian	7.016	7	0.50>p>0.30
Chinese vs Indian	21.414*	7	0.005>p>0.001

$G_H = \chi^2$ of the heterogeneity G-test

df = Degree of freedom

p = Probability

* = Significant in the distributions of alleles between different population samples

Note: Appendix C shows the method of calculations.

4.5 Sequencing

Samples that showed irregular banding patterns in the *Hga*I assay were selected for sequencing analysis to determine the source of variation. A total of 11 samples were sequenced and they included 7 samples representing irregular heterozygous and homozygous banding patterns and 4 normal samples that served as the controls.

From the sequencing results, a 12 bp deletion starting from positions -230C to -241A was detected in all samples showing irregular banding patterns. However, the deletion was only observed in one allele in each of the samples showing the irregular homozygous banding patterns. For irregular heterozygous samples, the deletion was noted in the same alleles bearing the *Hga*I site.

The sequencing results also ruled out the presence of a second *Rsa*I or *Hga*I site within the amplified fragment as suggested earlier (refer Section 3.1.2 *Hga*I +/- assay).

It is noteworthy that this 12 bp deletion was only noted in Malays and Indians. The number of samples included in this study, however, is not sufficiently large to conclude that this deletion is absent in the Chinese population.

This finding suggested that the Malays and Indians are more closely related than to the Chinese. It would be interesting to investigate this deletion in a larger population sample to determine its distribution within the three major races in the Malaysian population.

Since the variant only occurs in Malays and Indians, it might be a useful marker for studying the population structure and origin of these two racial groups.

Figures 18 and 19 illustrate the effects of the 12 bp deletion on the PCR products after *Rsa*I digestion in regular and irregular heterozygous and homozygous samples.

Figure 18: Regular and irregular heterozygous samples

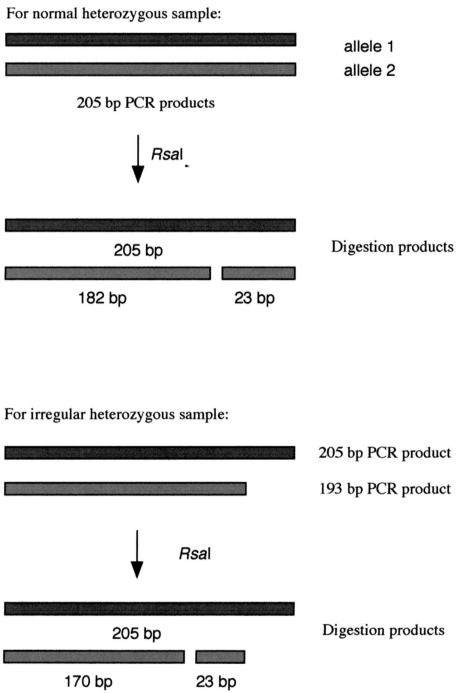
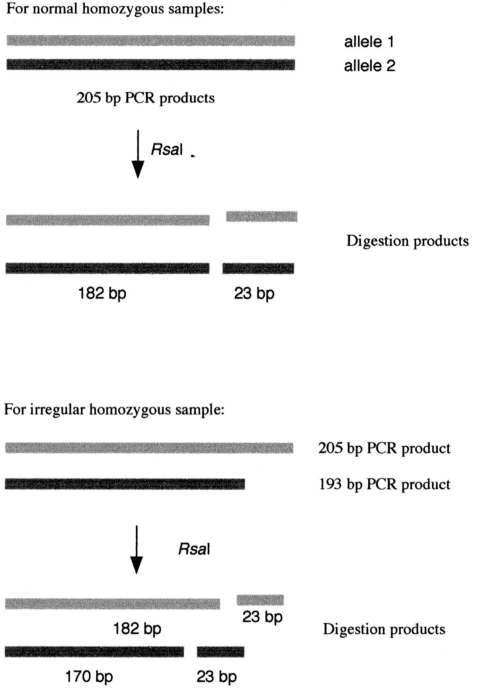


Figure 19: Regular and irregular homozygous samples



4.5 Conclusion

This study was conducted to determine the degree of variation within the 5' MS31A flanking DNA region in the Malaysian population, and for its possible use in MVR-PCR.

Results obtained thus far suggested that there is little variation among the races and the distribution of allele frequencies is in concordant to that expected of Hardy Weinberg equilibrium. The haplotype analysis however indicates, to a certain extent, that there exists some association between the polymorphic sites, although the linkage disequilibrium is incomplete.

Sequencing analysis revealed a 12 bp deletion, which was observed in individuals of Malay and Indian origin. Further studies are needed in order to determine the distribution of this deletion in the Malaysian population.