for one hour in (0.1 x SSC, 0.5% SDS, 0.2 M Tris-HCL pH 7.5). All membranes were then rinsed in 2 x SSC and stored at 4°C in 2 x SSC.

2.7 Chromosomal localisation of single-copy probes by somatic cell hybrid DNA analysis

2.7.1 Preparation of somatic cell hybrid lines

A multiple chromosome mapping panel was constructed by the author's colleague (dos Santos, 1986). Hybrid cell lines were isolated from various somatic cell hybridization experiments: rodent parental cells used were from the mouse permanent lines RAG (HGPRT[-]), Cl.1D (TK[-]) and B82 (TK[-]) and the Chinese hamster permanent line Wg3-h (HGPRT[-]). The human parental cells were obtained from amniotic fluid samples, skin biopsies or peripheral blood specimens, presenting normal or abnormal karyotypes, from patients who had been referred for cytogenetic investigations. Cytogenetic characterisation of hybrid cell lines was carried out using a modification of the G-11 differential staining technique after preliminary G-banding. Panels were constructed with combinations of cell lines which provided unique bimodal signatures for all human chromosomes. The panels also included lines obtained as gifts from T. Mohandas, A. Retief and U. Francke.

2.7.2 Extraction of somatic cell hybrid DNA

The protocol of A. E. Retief (personal communication, 1987) was followed.

175 cm flasks of confluent monolayer cells were washed twice with 10 ml ice-cold PBS (0.14 M NaCl, 5 mM KCl, 20 mM Na_2PO_4, 1 mM KH_2PO_4). Proteinase K was added to 15 ml lysis buffer (10 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5% SDS), to a final concentration of 50 µg/ml, and this solution was added to the flask. The flask was kept flat for five minutes and then at an angle of 45° to collect the cells. The cells were transferred to a 50 ml tube and incubated for thirty minutes at 68°C, followed by an incubation at 45°C for ninety minutes.

The DNA from the cells was extracted using an equal volume of chloroform:isoamyl alcohol (24:1), gently mixing for five to ten minutes, and separating the phases by centrifugation at 2 000 *g* for ten minutes. The extraction was repeated four to five times. DNA in the final aqueous phase was concentrated by an ethanol precipitation, after addition of NaCl to a final concentration of 0.3 M. DNA was resuspended in TE and stored at 4°C.

2.7.3 Use of somatic cell hybrid DNA in localising single-copy DNA probes

Somatic cell hybrid DNA was digested with the restriction enzyme *Hind*III. The choice of enzyme was not significant. 15 - 20 µg of DNA was digested, which allowed for the detection of those chromosomes present in low percentages in some cell lines. The DNA in agarose was transferred to a nylon membrane, and hybridized to the DNA probe for which a chromosomal localisation was required. Examination of the resulting autoradiograph then showed the cell line combination which had been highlighted by the radiolabelled probe, and the chromosome from which the probe was derived could be deduced.

2.8 Linkage analysis

Linkage analysis was carried out on data collected on members of families in which the tyrosinase-positive albinism phenotype was segregating; and on data collected on DNA samples from the C.E.P.H. families.

2.8.1 Serogenetic marker studies

Serogenetic marker studies were carried out by colleagues in the Department on donated blood from members of families in which the tyrosinase-positive

albinism allele was segregating. The phenotypes of all available individuals were established on the following systems, which were known to identify polymorphic loci in Negroid populations:

Red cell antigen systems:

ABO

MNSs

Rh

Red cell enzyme systems:

Acid phosphatase (ACP_1)

Glutamate pyruvate transaminase (GPT_1)

Glutathione peroxidase 1 (GPX_1)

Glyoxalase I (GLO_1)

Peptidase A (Pep A)

Phosphoglucomutase 1 (PGM_1)

6-Phosphogluconate dehydrogenase (PGD)

Serum protein systems:

Properdin Factor B (Bf)

Group specific component (Gc)

Haptoglobin (Hp)

Transferrin (Tf)

Alpha 1 antitrypsin (α_1-AT)

Blood grouping was carried out using methods summarised by Race and Sanger (1968). For the red cell enzyme

systems protocols in Harris and Hopkinson (1976) were followed. Properdin factor B (Bf) and group specific component (Gc) phenotypes were determined after electrophoresis by the method of Teisberg (1970) and after immunofixation by the method of Alper and Johnson (1969). Haptoglobin (Hp) and transferrin (Tf) phenotypes were determined by the methods described in Giblett (1969). Alpha-1-antitrypsin (α1-AT) phenotypes were determined by isoelectric focusing according to the method in Keuppers (1976) and subtypes were identified using a modification of the method in Weidinger *et al.* (1985).

2.8.2 Lod score determination

Lod scores were calculated using the computer programme LIPED3, devised by Ott (1974) for the Elston-Stewart algorithm (Elston and Stewart, 1971), in order to evaluate the likelihood of linkage between the tyrosinase-positive albinism locus and the loci of those polymorphic markers typed by the author and colleagues which showed useful variation in any family. A recessive model and complete penetrance in both sexes was assumed. The LIPED programme statistically analyses data from families with a recessively inherited disease by taking into account the fact that carriers are indistinguishable from unaffected individuals. This results in lower lod scores than would have been obtained

if the mode of inheritance were dominant, and consequently a larger sample population is required.

The LINKAGE programme (Lathrop et al., 1984) was used for evaluating the likelihood of linkage between loci identified by single-copy probes (identified by the author) and loci available in the C.E.P.H. files.

CHAPTER 3

3. ISOLATION OF SINGLE-COPY PROBES AND DETECTION OF DNA SEQUENCE POLYMORPHISM

This chapter deals with the bacteriophage that were isolated from human genomic DNA libraries and used as RFLP-detecting probes. The results obtained are given and discussed, and estimates of genomic heterozygosity are calculated.

3.1 Results

3.1.1 Identification of lambda Charon 4A recombinant bacteriophage containing unique sequence human DNA inserts

Two human genomic libraries were screened: one was prepared by *Hae*III/*Alu*I and the other by *Eco*RI partial digestion (Lawn *et al.*, 1978).

The maximum number of plaques which were non-overlapping (approximately 300) were obtained on lawns of *E coli* LE392. Plaques were transferred *in situ* to nitrocellulose (Benton and Davis, 1977) and hybridized to radiolabelled total human DNA. Putative single-copy

bacteriophage were identified by their apparent lack of hybridization to total human genomic DNA, since detectable hybridization is to be expected only with moderately or highly repeated sequences in the human genome. The number of intensely hybridizing, weakly hybridizing and non-hybridizing regions in the autoradiographs were determined by re-aligning the petri dishes from which the plaques were transferred with the autoradiographs. Non-hybridizing regions were scored as corresponding to potentially single-copy phage.

Approximately 2000 bacteriophage were screened in each library. 23.8 per cent and 2.1 per cent putative single-copy phage were obtained in the *Eco*RI and *Hae*III/*Alu*I libraries respectively.

Putative single-copy phage were re-screened; only one phage was plated onto a bacterial lawn on any one petri dish. Figure 3.1 shows a typical autoradiograph obtained in such a screen. Only phage which failed to hybridize to total human genomic DNA a second time were investigated further. Twenty of these phage were isolated from each library and the DNA extracted.

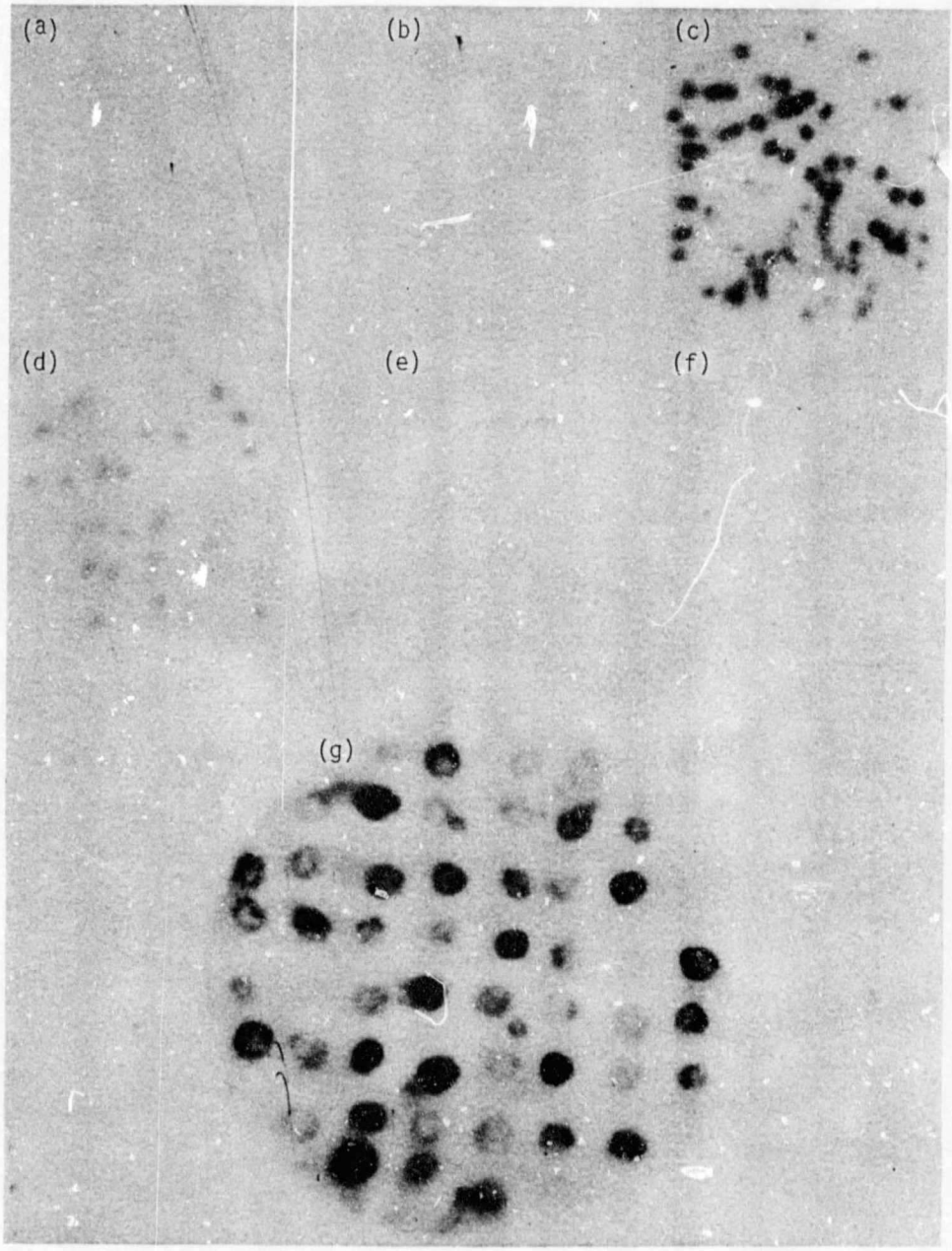

Figure 3.1 Autoradiograph showing nitrocellulose discs containing plaque DNA hybridised to total human genomic DNA. Discs a, b, e and f contain putative single-copy DNA. Discs c and d contain recombinant phage DNA with repetitive human DNA inserts. Disc g is a control with which to compare strongly-, weakly- and non-hybridising regions.

Putative single-copy phage DNA was digested with the restriction enzyme *Eco*RI and electrophoresed in 0.7% agarose. All digested phage had fragment sizes corresponding to the λCharon 4A arms (19.8 kb and 10.9 kb), and in some phage the arms were joined, giving a fragment of 30.7 kb. Those with additional fragments corresponding to the λCharon 4A 'stuffer region' (7.8 kb and 6.9 kb) were rejected as lacking human DNA inserts. Any phage with other *Eco*RI fragments were investigated further.

Figure 3.2 (a) shows the results of agarose gel electrophoresis of phage DNA after *Eco*RI digestion. All the phage in the photograph contain human DNA inserts. The DNA was transferred to nitrocellulose and hybridized to radiolabelled total human genomic DNA, in a final screen for repetitive sequence. The corresponding autoradiograph is shown in Figure 3.2 (b). Fragments containing repetitive DNA can be identified as bands to which the genomic DNA has hybridized.

Of the DNAs extracted from phage from the *Eco*RI library, only one (λRH52) contained a human insert. The other nineteen were non-recombinants. λRH52 was shown to constitute single-copy sequences by lack of hybridization to total genomic DNA. Although the

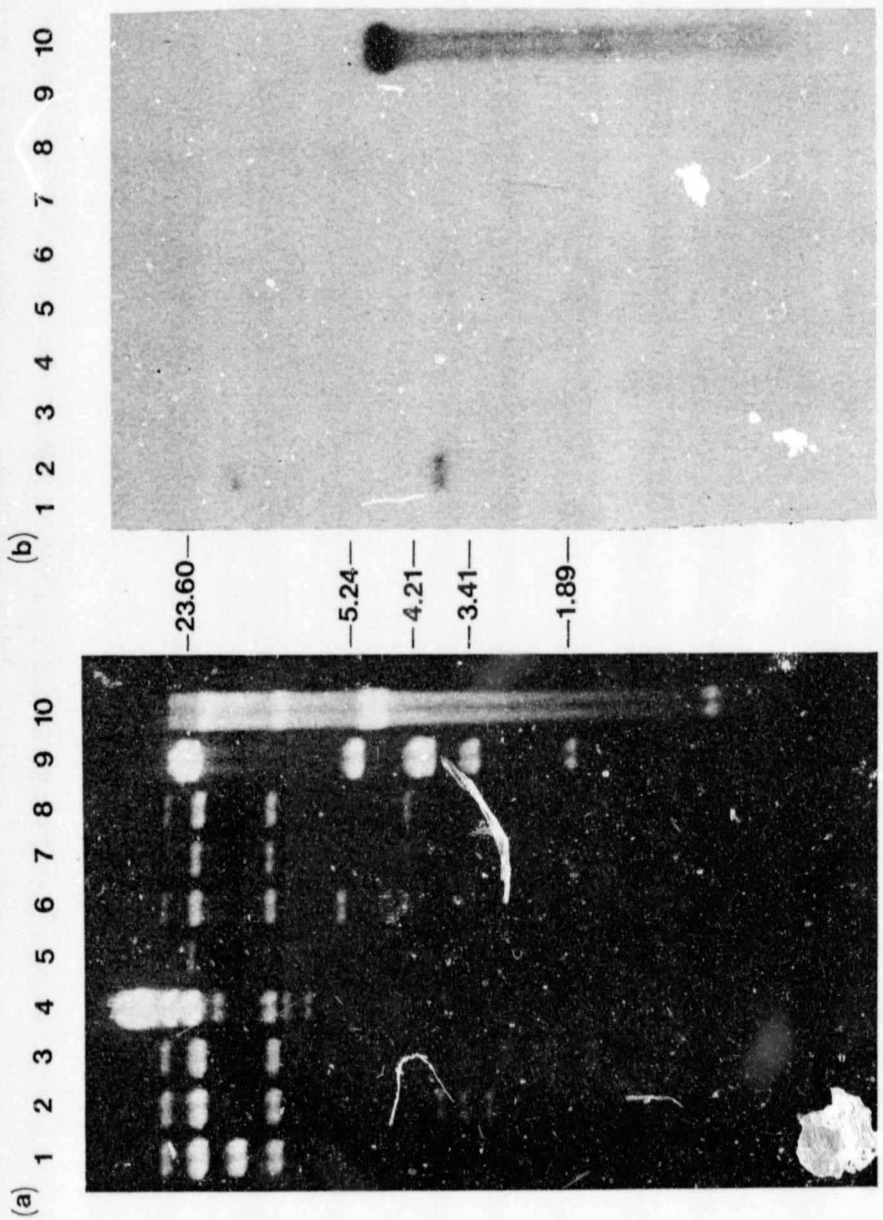

Figure 3.2 (a) Agarose gel electrophoresis of phage DNA after EcoRI digestion.

(b) Corresponding autoradiograph after Southern transfer and hybridization to radio-labelled total human genomic DNA. Fragments containing repetitive DNA are visible.

sample is too small for accurate extrapolation, it seems that a better figure for the percentage of single-copy phage in this library is less than two per cent ((1/20) x 23.8% = 1.2%).

Fourteen of twenty DNAs extracted from the *Hae*III/*Alu*I library contained human inserts. Of these, three were seen to contain repetitive sequences by hybridization to total genomic DNA. From these data, the percentage single-copy phage in the library can be estimated as approximately one per cent ((11/20) x 2.1 = 1.1%).

3.1.2 Screening for RFLPs in the South African Negroid population

High-molecular-weight DNA extracted from six or seven unrelated Negroid individuals was digested with each of the fourteen restriction enzymes *Taq*I, *Msp*I, *Bgl*II, *Hinc*II, *Sst*I, *Stu*I, *Pst*I, *Ava*II, *Bam*HI, *Eco*RI, *Hind*III, *Pvu*II, *Mbo*I, and *Rsa*I. A typical 'screening blot' constituted DNA from these 6 or 7 individuals digested with 3 different enzymes. Single-copy phage described in section 3.1.1 were used as RFLP-detecting probes: the phage were radio-labelled and hybridized to each screening blot. A typical autoradiograph is shown in Figure 3.3.

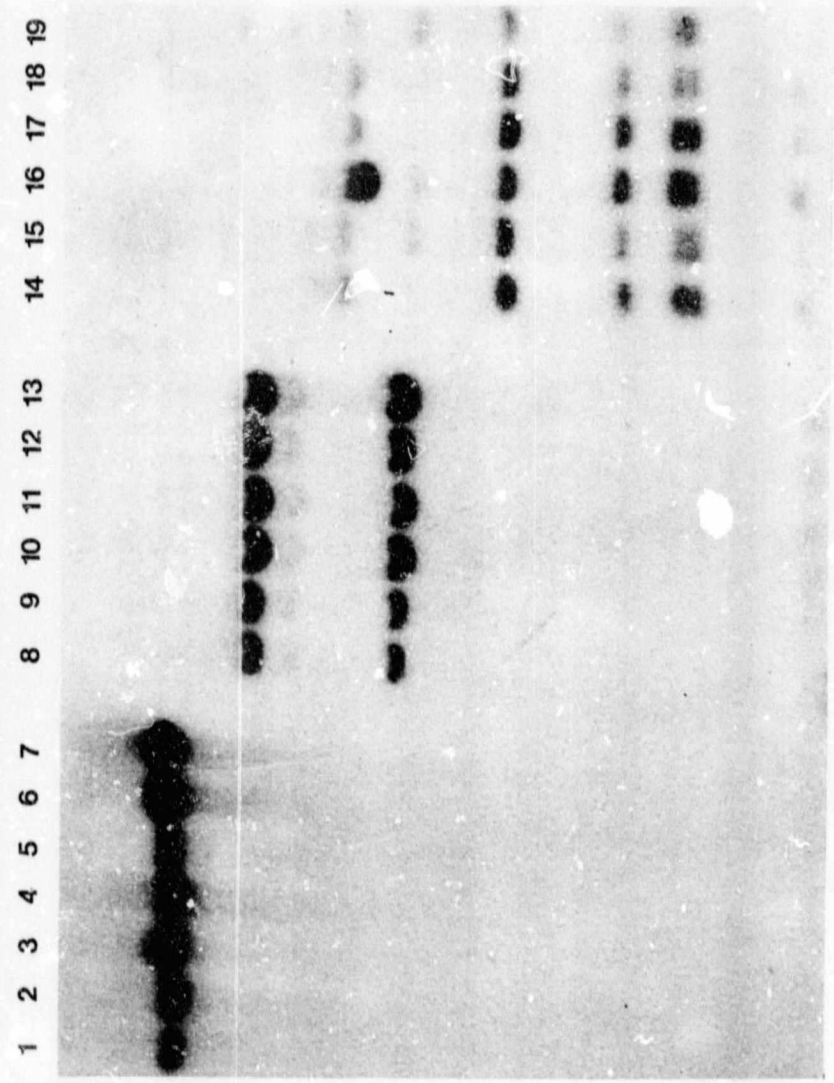

Figure 3.3 Autoradiograph showing an RFLP screening blot hybridized with the probe λMR3. Each lane contains unrelated Negroid DNA digested with BamHI (lanes 1 - 7), HincII (lanes 8 - 13), and PvuII (lanes 14 - 19).

Table 3.1 summarises the number of variants detected in the initial screening process. Six probes detected 14 variants with 10 enzymes. Sites that were not monomorphic in the DNA of the individuals screened are indicated by an asterisk. Only two variant alleles were seen for every probe/enzyme combination. The λMR3/*Stu*I, λMR3/*Mbo*I and λMR3/*Hin*dIII RFLPs were later shown to be in linkage disequilibrium (see section 4.1.2) and are therefore considered to represent one variant site. This site has been taken as a variant for a restriction enzyme with 6 bases in its recognition site, because the enzyme *Stu*I has been used most often in the RFLP's characterization. The number of variant sites, designated k, is given below the Table for enzyme recognition sites containing 4, 5 and 6 base pairs.

Table 3.2 shows the minimum number of sites screened (number of fragments seen plus one) with each probe/enzyme combination. Sites screened are designated m and the number screened with enzymes having recognition sites containing 4, 5 and 6 base pairs are given below the table.

Enzyme and probe variation has been calculated using data in Tables 3.1 and 3.2, and these estimates are given in Tables 3.3 and 3.4. H is an estimate of

Table 3.1 Variants detected by single-copy probes while screening for RFLP in the Negroid population.

Restriction Enzyme	Probe (N° of individuals screened)					
	λMR3	λMR7	λMR12	λMR17	λMR22	λRH52
TaqI	(6)	* (5/6)	(6)	(6)	(6)	* (2/6)
MspI	(6)	(6)	* (2/6)	* (3/6)	(6)	--
BglII	(7)	(6)	(6)	(6)	* (4/6)	(6)
HincII	(6)	* (4/6)	(7)	(6)	(6)	* (1/6)
SstI	(6)	(6)	(6)	(6)	(7)	(6)
StuI	* (5/7)	(7)	(6)	(6)	(6)	* (4/7)
PstI	(6)	(6)	(4)	(7)	* (1/6)	(6)
AvaII	(5)	(6)	(6)	(6)	(5)	(6)
BamHI	(6)	(7)	(6)	(7)	(6)	(6)
EcoRI	(6)	(6)	* (1/6)	(6)	(7)	(7)
HindIII	* (2/6)	(6)	(6)	(7)	(6)	(6)
PvuII	* (1/6)	(6)	(6)	(6)	(6)	(6)
MboI	* (5/6)	(6)	(6)	(6)	(7)	(7)
RsaI	(6)	(7)	(7)	(6)	(6)	(6)

Presence of a variant is indicated by an asterisk.

Breakdown of the number of variant sites :

$\Sigma k_4 = 4$

$\Sigma k_5 = 0$

$\Sigma k_6 = 8$

$\Sigma k_4 + \Sigma k_5 + \Sigma k_6 = 12$

Table 3.2 Minimum mumber of sites seen with each probe/enzyme combination (number of fragments + 1)

Restriction Enzyme	N° of bp recognised	Probe					
		λMR3	λMR7	λMR12	λMR17	λMR22	λRH52
TaqI	4	4	8	5	6	6	3
MspI	4	10	2	8	5	15	--
BglII	6	3	4	3	3	7	6
HincII	6	3	3	4	4	5	7
SstI	6	5	3	4	3	4	8
StuI	6	6	6	4	4	5	8
PstI	6	4	9	5	5	7	8
AvaII	5	7	3	9	6	11	3
BamHI	6	2	5	2	2	4	8
EcoRI	6	6	2	5	3	5	4
HindIII	6	9	2	4	3	4	6
PvuII	6	9	12	5	5	4	7
MboI	4	8	6	8	9	6	10
RsaI	4	10	6	11	11	12	12

Breakdown of the number of sites recognised :

$\Sigma m_4 = 181$

$\Sigma m_5 = 39$

$\Sigma m_6 = 268$

$8\Sigma m_4 + 10\Sigma m_5 + 12\Sigma m_6 = 5054$

Table 3.3 Enzyme variation in the detection of RFLPs

Enzyme	N° of probes used	N° of variant sites detected	N° of variant chromo-somes	Base pairs screened	Heterozygosity H	Heterozygosity $\hat{H}$
TaqI	6	2	7	1536	0,0078	0,0091
MspI	5	2	5	1920	0,0062	0,0052
BglII	6	1	4	1872	0,0032	0,0043
HincIII	6	2	5	1872	0,0064	0,0053
SstI	6	0	0	1944	0,0000	0,0000
StuI	6	2	9	2376	0,0051	0,0076
PstI	6	1	1	2736	0,0022	0,0007
AvaII	6	0	0	2340	0,0000	0,0000
BamHI	6	0	0	1656	0,0000	0,0000
EcoRI	6	1	1	1800	0,0033	0,0011
HindIII	6	1	2	2016	0,0030	0,0020
PvuII	6	1	1	2880	0,0020	0,0007
MboI	6	1	5	2256	0,0027	0,0044
RsaI	6	0	0	2976	0,0000	0,0000

$$H = \frac{(k_4 + k_5 + k_6)}{8m_4 + 10m_5 + 12m_6}$$

$$\hat{H} = 1 - \left[\left(\frac{a}{b}\right)^2 + \left(\frac{b - a}{b}\right)^2 \right]$$

Table 3.4 Probe variation in the detection of RFLPs

Probe	N° of enzymes used	N° of bp screened per individual ($8\Sigma m_4 + 10\Sigma m_5 + 12\Sigma m_6$)	N° of bp screened all individuals [b]	N° of variant chromosomes [a]	N° of variant sites ($\Sigma k_4 + \Sigma k_5 + \Sigma k_6$)	H	$\hat{H}$
3	14	890	5378	11	2	0.00224	0.00408
7	14	758	4728	11	2	0.00264	0.00464
12	14	778	4684	3	2	0.00251	0.00128
17	14	692	4272	3	1	0.00145	0.00140
22	14	962	5818	5	2	0.00208	0.00172
52	13	974	6068	7	3	0.00308	0.00213
		$\Sigma = 5054$	$\Sigma = 30948$	$\Sigma = 40$	$\Sigma = 12$		

$$H = \frac{(k_4 + k_5 + k_6)}{8m_4 + 10m_5 + 12m_6}$$

$$\hat{H} = 1 - \left[\left(\frac{a}{b}\right)^2 + \left(\frac{b-a}{b}\right)^2 \right]$$

heterozygosity calculated as $(k_4 + k_5 + k_6)/(8m_4 + 10m_5 + 12m_6)$, (Ewens *et al.*, 1981). This formula was used for consistency, so that other studies in which H has been calculated can be compared with data in the present study. H can be defined as the fraction of sites tested in the genome at which two or more nucleotide types appear. It takes no account of allele frequency and cannot be directly applied to a small group of individuals. Another estimate of heterozygosity was therefore used as well, the formula for $\hat{H}$ proposed by Nei (1975) and re-defined by Pearson (1985). $\hat{H}$ is becoming accepted as one of the best estimators of heterozygosity, and can be defined as the probability that when two homologous DNA sequences are compared they will show different nucleotide compositions at a given site. $\hat{H} = 1 - [(a/b)^2 + ((b-a)/b)^2]$, where a is the total number of variants on all chromosomes and b is the total number of base pairs examined on all chromosomes.

Estimates of overall genomic heterozygosity, calculated using both formulae, are given in Table 3.5.

Probe/enzyme combinations with which a DNA sequence polymorphism was detected were further investigated, and these results are presented in Chapter 4.

Table 3.5 Estimation of genomic heterozygosity

N° of probes	N° of bp screened per haploid genome [½(8Σm$_4$ + 10Σm$_5$ + 12Σm$_6$)]	N° of bp screened all individuals [b]	N° of variant sites (Σk$_4$ + Σk$_5$ + Σk$_6$)	N° of variant chromosomes [a]	H	$\hat{H}$
6	2527	30948	12	40	0.00237	0.00258

3.2 Discussion

3.2.1 Isolation of single-copy probes

Less than two per cent of the recombinant phage in the two genomic libraries obtained from T. Maniatis contained single-copy human DNA inserts. Initial estimates of putative single-copy phage were 2.1 and 23.8 per cent, figures which were revised as non-recombinant phage were identified. Such non-recombinants could have been produced during library construction and possibly selected for (by a growth advantage, for instance) during amplification of the library; alternatively, human inserts initially present may have been lost during amplification. The high frequency of phage lacking inserts was not expected, and may have been due, in addition to factors proposed above, to the age of the libraries used, which was approximately five years. Alterations in the constitutions of the libraries over time can only be speculated upon.

The percentages of single-copy phage obtained are consistent with reports in the literature. Botstein *et al.* (1980) obtained 1-3%, and Kao *et al.* (1982) obtained 1% single-copy phage from genomic libraries.

Slightly higher single-copy percentages were seen in the library prepared by partial digestion with *Eco*RI (1.2%) than with *Hae*III/*Alu*I (1.1%). It seems likely that recombinant phage containing an *Alu*I recognition site would also contain at least part of the ubiquitous *Alu* repetitive sequence. Such sequences are thought to be present in at least 95% of phage in the Maniatis human genomic libraries (Tashima *et al.*, 1981). The presence of part or all of an *Alu* family repeat in a high percentage of recombinants from the *Hae*III/*Alu*I library would explain a lower percentage of single-copy inserts obtained from this library. The percentage difference was not significant in this study, however.

The strategy used in the present study to obtain single-copy phage for the detection of RFLPs was that pioneered by R. White and co-workers (Wyman and White, 1980). RFLPs detected in this way are usually bi-allelic. Strategies for obtaining probes which identify highly polymorphic loci were developed subsequent to the work performed in the present study. These were described in section 2.2.2 and the implications for their use compared with the type obtained in this study will be discussed in Chapter 6.

The reason why the present strategy should identify predominantly two-allele polymorphisms has only become clear in recent years. Workers attempting to clone the region responsible for the variation detected by pAW101

(Wyman and White, 1980) found that hypervariable regions are unstable in the *recA*$^+$ hosts used in the construction of human DNA libraries (Wyman *et al.*, 1984, 1985). This would account for the relatively few highly polymorphic loci identified using probes from genomic, chromosomal or cDNA libraries constructed in *recA*$^+$ hosts. It is ironic that the first arbitrary RFLP locus (D14S1), identified by the probe pAW101 and identified in this way, was a notable exception to this finding.

3.2.2 Detection of RFLPs

Single-copy probes were used to screen DNA from a small number of individuals which had been digested with a relatively large number of enzymes. This was a strategy used by Aldridge *et al.* (1984). The rationale behind the strategy was that most polymorphisms identified in this way would have a rarer allele frequency of greater than 0.15 (Skolnick and White, 1982). In the present study, when the variants detected in the initial screening were investigated further, this was generally found to be true (see Chapter 4).

The enzymes used in the study had previously been shown to be useful in detecting DNA sequence polymorphism (Skolnick and White, 1982; Wijsman, 1984; Cooper and

Schmidtke, 1984). Their usefulness in the present study can be determined from the data in Table 3.3. It seems that the enzymes used in this study fall into two general groups: those with low or zero heterozygosity values, including *Pst*I, *Ava*II, *Bam*HI, *Eco*RI, *Sst*I, *Pvu*II and *Rsa*I; and those with relatively higher heterozygosity values, including *Taq*I, *Msp*I, *Bgl*II, *Hinc*II, *Hind*III, *Stu*I and *Mbo*I.

It is interesting that *Msp*I and *Taq*I, enzymes which contain CpG in their recognition sites, are thought to detect RFLPs at a higher frequency than other enzymes (Barker and White, 1982; Barker *et al.*, 1984). The rationale behind this is that the dinucleotide is heavily methylated (Van der Ploeg and Flavell, 1980; Cooper, 1983) and 5-methyl-cytosine is subject to frequent replacement by thymidine due to the de-amination of the methylated base (Coulondre *et al.*, 1978). This explains both the high frequency of the C to T transition (Vogel, 1972; Vogel and Kopun, 1977) and the resulting CpG deficiency in vertebrate genomes (Salser, 1977; Bird, 1980), and potentially explains a high frequency of variation in recognition sites containing CpG. Cooper *et al.* (1985) conclude that *Msp*I and *Taq*I are more informative than other enzymes; whereas Pearson (1985) concludes that *Msp*I and *Taq*I are not significantly more informative than *Bgl*II, *Eco*RI and *Pst*I. Results from the present study suggest that

*Msp*I and *Taq*I are not significantly more informative than *Hinc*II and *Stu*I. It seems reasonable to conclude that a bias is introduced by the use of some enzymes more frequently than others, and they are consequently thought to detect variation more often. Whether *Msp*I and *Taq*I are the most likely to detect variation has not been proven in any reported studies, but that they should be included in any screening panel is unquestionable. If the majority of information is found to be derived from a relatively small number of enzymes, then estimates of heterozygosity based on the assumption of a random distribution of sites r lative to nucleotide distribution will be biased and may have to be modified.

The findings from the present study should be useful together with other more extensive studies for selecting restriction enzymes for the detection of DNA polymorphism.

3.2.3 Estimation of genomic heterozygosity

Individual probe variation is shown in Table 3.4. The probes λMR3 and λMR7 have the highest heterozygosity values, and λMR12 has the lowest. Data from the six probes and fourteen enzymes used to screen 30948 bp were collated for the estimation of genomic heterozygosities of 0.0024 or 0.0026, shown in Table 3.5.

In perhaps the most comprehensive reviews of reported polymorphism with respect to human variation, Cooper and Schmidtke (1984) and Cooper *et al.* (1985) have given figures of 0.0026 and 0.0037 (the latter figure obtained from a larger sample) as estimates of heterozygosity in the human genome. The figure of approximately 0.0025 obtained in the present study is not dissimilar to Cooper and Schmidtke's first estimate.

To the writer's knowledge, the present study is the first in which genomic heterozygosity at the DNA level has been examined in a Negroid population. Other studies have been in Caucasoid populations. That similar amounts of variation appear to be present in the two population groups was to be expected, but these results are the first direct confirmation of such expectations.

In Jeffrey's (1979) systematic study of the β-globin loci, 60 individuals were screened with eight different restriction enzymes. He concluded from his results that at least one in 100 base pairs showed variation in this (largely coding) region. Murray *et al.* (1983) examined variation at the human serum albumin locus, and came to a similar conclusion. These studies did not however provide a direct estimate of heterozygosity, because the number of individuals screened

with each enzyme was not taken into account. Ewens *et al.* (1981) and Cooper and Schmidtke (1984) re-calculated Jeffreys' and Murray *et al.*'s results respectively as approximately 0.0040 and 0.0025.

Different estimates of heterozygosity are due to calculations derived from variation in coding and in single-copy regions: results referred to above and those in the study of Selentjaris and Gesteland (1983) seem to imply that there is at least twice as much variation outside coding regions as within them. If this is true then it would seem that selection is acting in such a way as to maintain sequence integrity within genes, but that there is relaxed selection pressure outside genes. Although much of the genome has been described as 'junk', there is increasing evidence for regions outside genes performing a large number of regulatory and structural functions. The question of whether variation in these regions is of neutral effect or whether it contributes to genomic evolution in a random or non-random fashion is not resolved. Study of increasingly large numbers of DNA sequence polymorphisms and a better understanding of the functional role of the DNA sequence itself will help to resolve some of these questions.

Comparisons of heterozygosity between chromosomes has led to the suggestion that the X chromosome has low

heterozygosity and is perhaps constrained with respect to sequence variability (Aldridge *et al.*, 1984; Pearson, 1985). A possible explanation for this is that the X chromosome contains a higher proportion of coding sequences than do the autosomes.

3.3 Summary

Two human genomic libraries obtained from T. Maniatis were screened for recombinant bacteriophage containing unique sequence DNA inserts. Less than two per cent of such phage were obtained from each library. Some of these phage were used as RFLP-detecting probes: DNA from 6 or 7 unrelated Negroid individuals was digested with 14 restriction enzymes, electrophoresed in agarose, transferred to a solid support, hybridized to the radiolabelled probe, and autoradiography was performed. Examination of autoradiographic bands enabled detection of restriction fragment variation. Fourteen variants were detected by 6 probes in combination with 10 enzymes.

The most informative enzymes with respect to detection of variation turned out to be *Taq*I, *Msp*I, *Bgl*II, *Hinc*II, *Hind*III, *Stu*I and *Mbo*I. Genomic heterozygosity was calculated to be 0.0025, so that 2 or 3 variants would be expected in every 1000 nucleotides. This is

consistent with the published estimations of Cooper and Schmidtke (1984) and Cooper *et al.* (1985).

Some of the probe/enzyme combinations which detected variation were investigated further, as described in the next chapter.

CHAPTER 4

4. STUDIES OF ANONYMOUS RFLPs DETECTED BY SINGLE-COPY PROBES

Six probes were used to detect DNA sequence variation as described in chapter 3. λRH52 was isolated from the *Eco*RI genomic library, and λMR3, λMR7, λMR12, λMR17, and λMR22 were isolated from the *Hae*III/*Alu*I genomic library. The fragment patterns revealed by *Eco*RI digestion of each of these probes are shown in Figure 4.1. Table 4.1 gives the *Eco*RI fragment sizes for each probe, thereby allowing estimation of the total size of each human insert.

λRH52, λMR3, λMR22 and λMR7 have been found to detect RFLPs that should be useful genetic markers. These RFLPs are described and discussed in section 4.1. The assignments of λRH52, λMR3, λMR7 and λMR22 are discussed in section 4.2, and population studies are described and discussed in section 4.3. Lastly, the inclusion of the probes λRH52 and λMR3 in the C.E.P.H. programme is described in section 4.4.

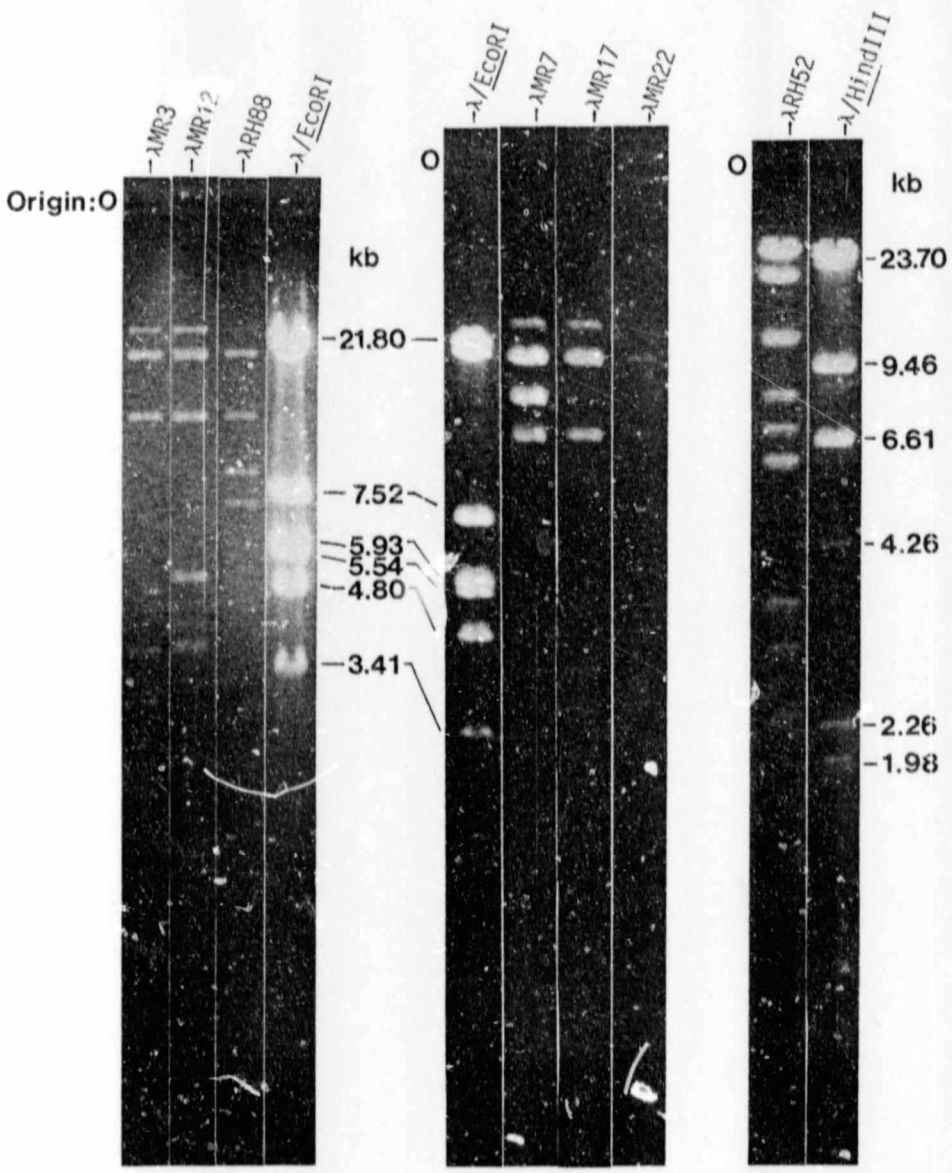

Figure 4.1 EcoRI restriction patterns of the recombinant single-copy phage used in the study. Lambda size markers and a non-recombinant, λRH88, are shown.

Table 4.1 EcoRI fragment sizes of λCharon4A recombinant bacteriophage used as single-copy probes.

Fragment sizes (kb)	Probe name					
	λMR3	λMR7	λMR12	λMR17	λMR22	λRH52
Phage DNA	30.70	30.70	30.70	30.70	30.70	30.70
	19.80	19.80	19.80	19.80	19.80	19.80
	10.90	10.90	10.90	10.90	10.90	10.90
Human DNA	3.60	14.75	5.00	4.30	8.70	8.60
	3.00		4.15	3.75	2.10	6.90
	2.90		3.70	3.30		5.85
	2.10		1.60	3.15		3.45
	1.05					3.10
	0.90					2.35
	0.60					
Minimum size of human DNA insert (kb)	14.15	14.75	14.45	14.50	10.80	30.25

4.1 Detection of DNA sequence polymorphism with single-copy probes

4.1.1 Random single-copy probe λRH52: variation detected with *StuI*, *TaqI* and *HincII*

λRH52 contains a large human DNA insert of approximately 30 kb (Table 4.1). It was isolated from a Maniatis library for which 15-20 kb human fragments were selected, but the cloning of a 30 kb fragment is not improbable. It would seem likely that a 30 kb DNA sequence should contain repetitive DNA. However, on the basis of three screening events in which the probe DNA did not hybridize to total human genomic DNA, even after very low stringency post-hybridization washes of 2 x SSC at room temperature, λRH52 was considered to be single-copy. In addition, the fact that the probe identifies only two fragments with both *TaqI* and *AvaII* would suggest that it contains unique sequence DNA.

λRH52 detects a two-allele RFLP with the restriction enzyme *StuI*. The polymorphic pattern is shown in Figure 4.2; the polymorphic fragment sizes are 5.8 kb and (4.0 + 1.8) kb. In Figure 4.3 (a) and (b) examples of family studies which can be interpreted as fulfilling Mendelian expectations are shown. The RFLP has been seen in the three populations studied, Negroid, San and Caucasoid.

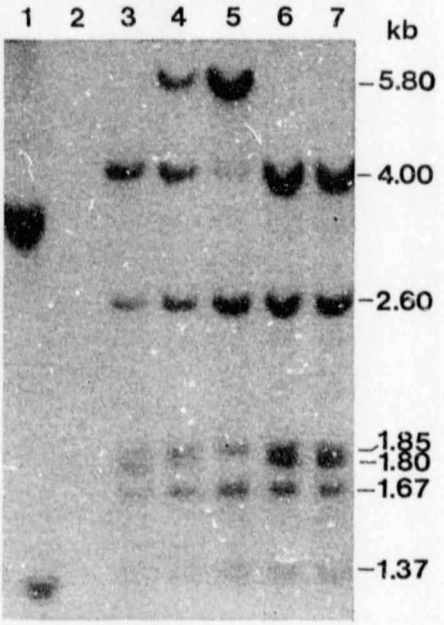

Figure 4.2 The polymorphic pattern revealed by λRH52 when hybridized to StuI-digested DNA from unrelated Negroid individuals. (lanes 3 - 7). Polymorphic 'alleles' of 5.8 kb and (4.0 + 1.8) kb can be seen. Lane 1 contains a lambda molecular weight marker. Lane 2 contains no sample.

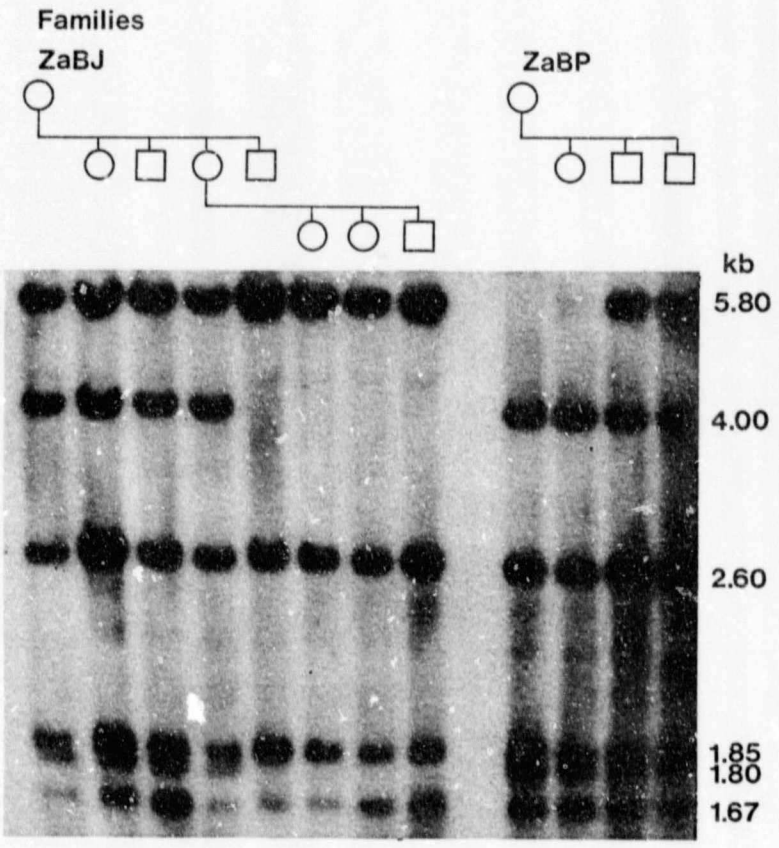

Figure 4.3(a) Study of two Negroid families showing the codominant Mendelian inheritance of the λRH52/StuI RFLP.

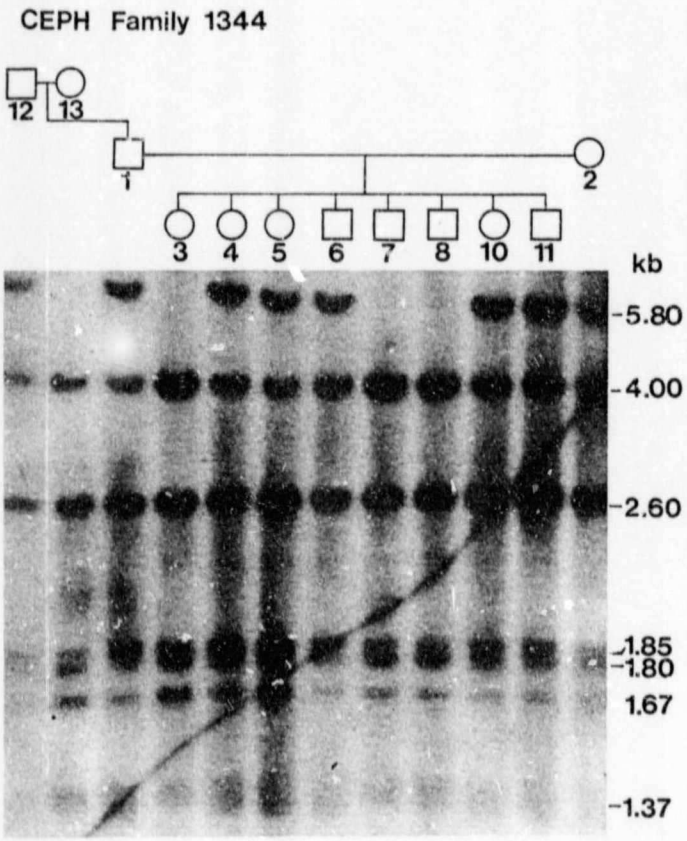

Figure 4.3(b) Study of a Caucasoid family showing the codominant Mendelian inheritance of the λRH52/StuI RFLP.

λRH52 also appears to detect a two-allele TaqI RFLP. This RFLP was detected in the Negroid population, but seems to occur at a low frequency in this population. The pattern in six unrelated Negroid individuals is shown in Figure 4.4 (a); the polymorphic fragments sizes are 20.05 kb and (13.10 + 6.95) kb. The rarer 'allele' in this RFLP occurs at a higher frequency in the South African Caucasoid population than in the Negroid population, and a third 'polymorphic allele' or variant band of 14.5 kb has been seen in one Caucasoid individual, whose family was, unfortunately, not available for further study. The pattern in unrelated S.A. Caucasoids including the individual with the 14.5 kb band is shown in Figure 4.4 (b). Eighty unrelated C.E.P.H. Caucasoids were also screened and found to be monomorphic for the (13.10 + 6.95) kb 'allele' (Figure 4.4 (c)). An apparently Mendelian inheritance pattern has been seen in two Negroid families, and one Caucasoid individual homozygous for the 20.05 kb fragment has been seen. The surprising difference between the frequency of allele 1 in S.A. Caucasoids and C.E.P.H. Caucasoids (see Table 4.7 for allele frequencies) is difficult to explain. It is possible that the South African Caucasoid is not as close genetically to the Western European as had been previously supposed. This is referred to again on page 181. Larger samples and more evidence will be needed, however, before such a conclusion can be drawn.

The haplotypes shown in the table below have been derived for 12 individuals in the Caucasoid population. The haplotypes have been numbered and their frequencies calculated. These are the 'observed' values in the table. The 'expected' values were calculated by multiplying together the known population frequencies

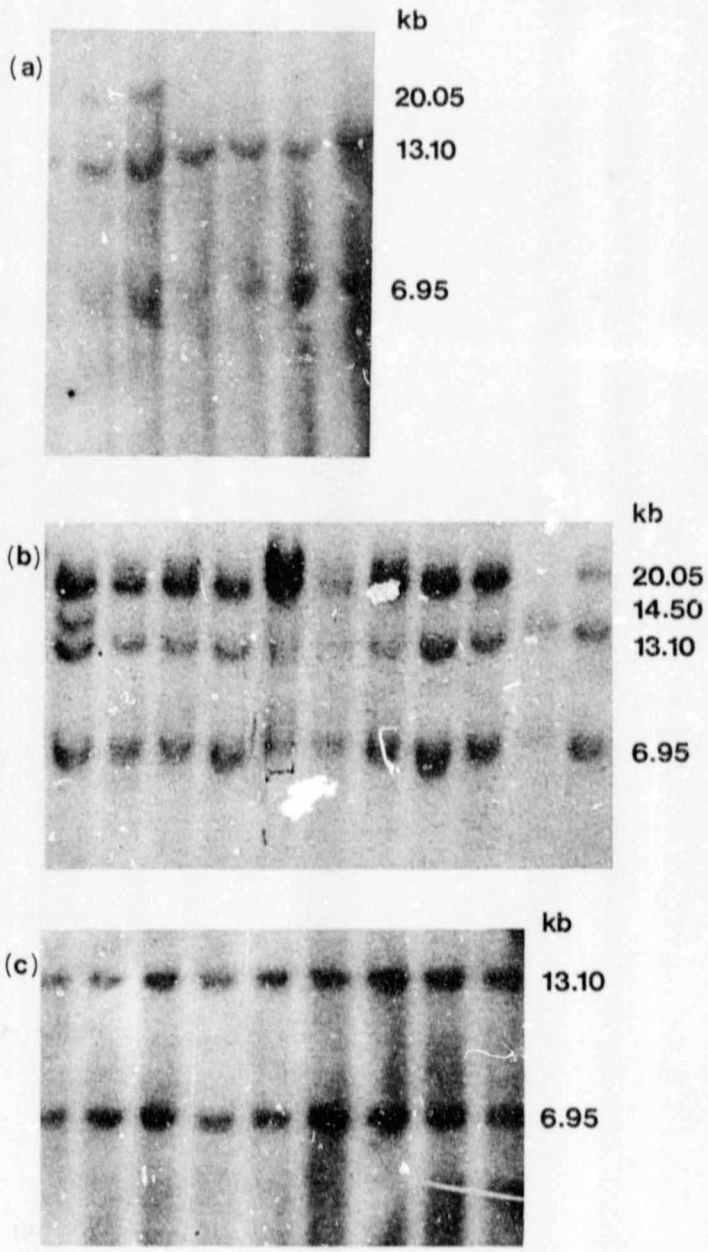

Figure 4.4 λRH52/TaqI RFLP
(a) Negroid DNA
(b) S.A. Caucasoid DNA
(c) C.E.P.H. Caucasoid DNA

(given in Table 4.7 on page 170) for each allele in the haplotype.

StuI allele	TaqI allele	Haplotype	Frequency	
(Fragment sizes in kb)			Obs.	Exp.
5.8	20.05	1	0.29	0.24
5.8	13.10; 6.95	2	0.17	0.29
4.0; 1.8	20.05	3	0.33	0.21
4.0; 1.8	13.10; 6.95	4	0.21	0.26

The chi square value (at 3 degrees of freedom) is 0.14, suggesting that there is no linkage disequilibrium at this locus and that the alleles are dissociating randomly. The sample size is very small, however, and this conclusion will only be justified if a larger sample is tested. Because each haplotype is represented it would indicate that haplotyping at this locus will markedly increase its informativeness. The PIC (Polymorphism Information Content) value calculated from the observed frequencies given above is 0.71, almost double the highest value possible for a two-allele polymorphism, which is 0.37.

A third variant revealed by λRH52 is detected with the restriction enzyme *Hin*cII. The variant is rare, and has only been seen in two individuals in the Negroid population (Figure 4.5). Neither individual's family is available for study of its inheritance.

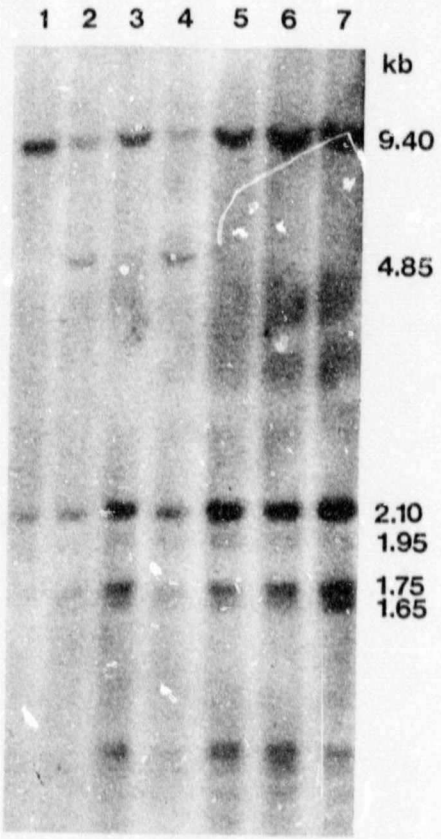

Figure 4.5 λRH52/HincII variant. DNA from unrelated Negroid individuals. The 4.85 kb variant band can be seen in lanes 2 and 4.