Am. J. Hum. Genet. 55:513–525, 1994

VNTR Alleles Associated with the α-Globin Locus Are Haplotype and Population Related

Jeremy J. Martinson,[1] Anthony J. Boyce,[2] and John B. Clegg[1]

[1]Institute of Molecular Medicine and [2]Institute of Biological Anthropology, University of Oxford, Oxford

Summary

The human α-globin complex contains several polymorphic restriction-enzyme sites (i.e., RFLPs) linked to form haplotypes and is flanked by two hypervariable VNTR loci, the 5′ hypervariable region (HVR) and the more highly polymorphic 3′HVR. Using a combination of RFLP analysis and PCR, we have characterized the 5′HVR and 3′HVR alleles associated with the α-globin haplotypes of 133 chromosomes, and we here show that specific α-globin haplotypes are each associated with discrete subsets of the alleles observed at these two VNTR loci. This statistically highly significant association is observed over a region spanning ~100 kb. With the exception of closely related haplotypes, different haplotypes do not share identically sized 3′HVR alleles. Earlier studies have shown that α-globin haplotype distributions differ between populations; our current findings also reveal extensive population substructure in the repertoire of α-globin VNTRs. If similar features are characteristic of other VNTR loci, this will have important implications for forensic and anthropological studies.

Introduction

Tandemly repetitive DNA loci, normally referred to as *VNTR* loci (Nakamura et al. 1987), are found widely dispersed throughout the genome. Their high polymorphism arises from variation in the number of the short sequence repeats from which each VNTR locus is composed. In addition, some VNTR loci are now known to be composed of more than one type of repeat sequence, so that an additional layer of polymorphism can be seen when the internal repeat structures of such loci are compared (Jeffreys et al. 1991; Neil and Jeffreys 1993).

Characterization of VNTR variation has usually been by measurement of allele sizes obtained by Southern blot hybridization (Balazs et al. 1989) or PCR (Budowle et al. 1991*a*), although more recent techniques have included direct sequencing or mapping of PCR-amplified alleles (Armour and Jeffreys 1992; Desmarais et al. 1993). Until recently, these analyses have all concentrated on the distribution of allele sizes or maps as an indication of the polymorphism present at a locus and have not included information on additional polymorphisms that may be present in the sequences flanking VNTR loci, despite early suggestions of such linkage between VNTR allele size and nearby RFLPs for the α-globin (Higgs et al. 1986), insulin (Cox et al. 1988), and, more recently, apolipoprotein B (Renges et al. 1992) loci. For the majority of VNTR loci, however, the extent of polymorphism near the locus is not known or is limited to a few hundred base pairs in the sequences flanking the VNTR array (Armour et al. 1993; Monckton et al. 1993), despite the fact that such flanking-marker information would yield many insights into VNTR mutation processes.

The human α-globin gene complex (fig. 1) is a good system with which to study these phenomena. Detailed mapping of this complex has revealed nine polymorphic sites throughout a 30-kb region; extensive linkage disequilibrium between the sites has resulted in only a limited number of haplotypes (Higgs et al. 1986). In addition to these site polymorphisms, the complex is flanked by two VNTR loci, the 5′HVR (Jarman and Higgs 1988) and the 3′HVR (Jarman et al. 1986). The 5′HVR array is composed of copies of a 57-bp repeat unit and has a heterozygosity of ~70%. The 3′HVR is composed of copies of a 17–21-bp repeat and is more polymorphic, with heterozygosities approaching 100% in some populations (Jarman et al. 1986) (table 1).

α-Globin haplotypes provide a powerful system with which to study population amalgamation and admixture (O'Shaughnessy et al. 1990). Our previous studies of α-globin haplotypes and genotypes, as well as other genetic markers, in the populations of Oceania have demonstrated that several aspects of the colonization history of the region are evident from the gene pools of the different populations (O'Shaughnessy et al. 1990). The α-globin gene markers in Polynesia are predominantly those seen in Southeast Asia (e.g., haplotypes Ia and IIa) but include a substantial proportion of markers found only in Melanesia (haplotypes IIIa, IVa, and Vc; fig. 1) (Hill et al. 1989). This is indicative of admixture between the ancestors of the present-day Polynesians, who originated in island South-

Received October 11, 1993; accepted for publication May 16, 1994.

Address for correspondence and reprints: Dr. Jeremy J. Martinson, MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, England.

0002-9297/94/5503-0013$02.00

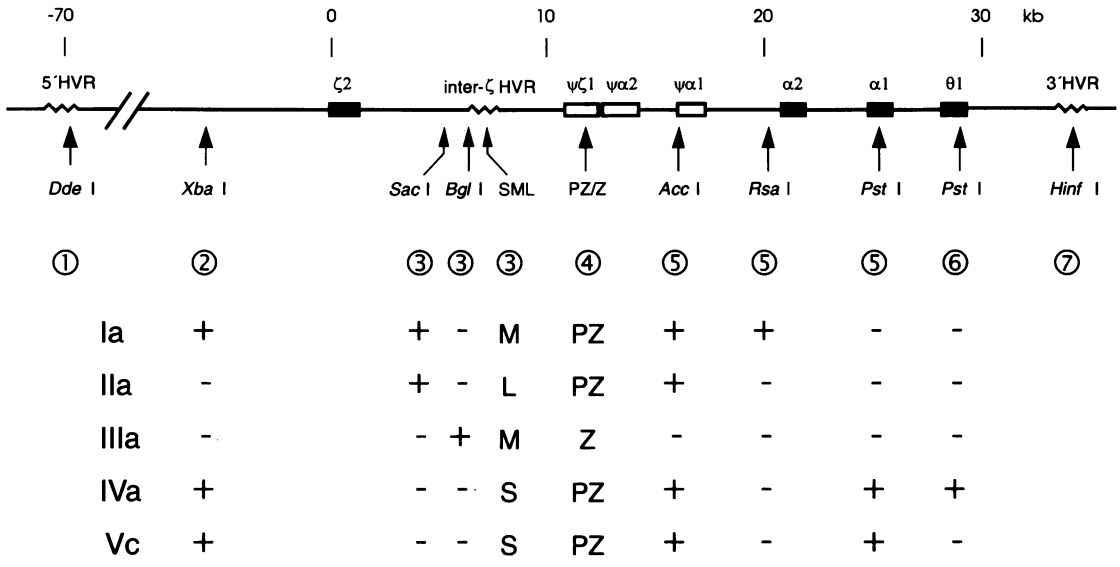

Figure 1 α-Globin cluster, showing the location of the haplotype sites and the flanking HVRs. Functional genes are denoted by blackened boxes; and pseudogenes are denoted by unblackened boxes. Zigzag lines denote the location of VNTR loci. The locations of the polymorphic loci that make up the haplotype, together with the restriction enzymes with which they are detected, are shown below the map. The numbers 1–7 below the restriction sites refer to the probes used to detect allelic variation; details for each probe can be found in Jarman and Higgs (1988) (1), Higgs et al. (1986) (2 and 6), Goodbourn et al. (1983) (3), Proudfoot et al. (1982) (4), Lauer et al. (1980) (5), and Jarman et al. (1986) (7). S, M, and L = small, medium, and large inter-ζ HVR allele sizes, respectively, discernible at the resolution used. Z and PZ = length polymorphism produced by a small VNTR locus within the intron of the ζ-gene, whose size variation is associated with the presence of either a ζ-gene (Z) or a ψζ pseudogene (PZ) at the location shown (Hill et al. 1985*b*). Numbers above the map represent length in kilobases. Below the map are shown some of the haplotypes common in Polynesia and discussed in the text. α-Globin haplotype nomenclature classifies the major haplotypes according to the alleles present at the five sites at the 3′ end of the haplotype; thus the group I haplotypes all end in the motif PZ++−−, group II in PZ+−−−, group III in Z−−−−, etc. (Higgs et al. 1986). Subdivision within major groups is on the basis of the four 5′ RFLP sites.

east Asia, and the inhabitants of the islands of Melanesia (Bellwood 1989; Gibbons 1994). Our studies have also shown that the populations of eastern Polynesia have a lower diversity at several minisatellite VNTR loci, consistent with a reduction in population size during the colonization of the region (Flint et al. 1989). Together these data show that the inhabitants of present-day Polynesia contain a variety of identifiable markers from their different ancestral populations but that their overall genetic diversity has been reduced as a consequence of the genetic bottleneck through which they passed during the colonization process.

This detailed anthropological information available as a result of our previous surveys makes the populations of Oceania a particularly useful model system with which to study in detail the genetic events that have given rise to the observed diversity at the α-globin complex. In particular, Polynesians have one marker chromosome, thought to have originated in Melanesia (Hill et al. 1985*a*) but carried into Polynesia and present there at high levels, probably as a consequence of genetic drift (Hill et al. 1987). The high frequencies of this unique marker have enabled us to develop an approach for assigning the sizes of both the 5′HVR and 3′HVR alleles to a variety of haplotypes in the populations of Oceania, and we show here that different haplotypes have their own characteristic range of VNTR alleles. Because of the lower polymorphism of the 5′HVR, some alleles at this locus are associated with more than one haplotype. At the 3′HVR, however, few occurrences of two different haplotypes sharing an identically sized allele are seen.

Material and Methods

Haplotype Determination

The samples studied here were collected from the populations of French Polynesia, in collaboration with the Institut Territorial de Recherches Médicales Louis Malardé, Papeete, Tahiti, French Polynesia. Samples were collected from healthy adult donors resident in the archipelagoes of French Polynesia and also from umbilical cords obtained from the maternity hospital in Papeete. Genealogical and ethnic-status information was also collected; this allowed for non-Polynesian individuals to be excluded from the survey. The blood samples were collected and transported on dry ice and were stored at −70°C prior to extraction. Leukocyte nuclear DNA was extracted by standard proteinase K digestion and phenol/chloroform extraction (Old and Higgs 1983) using an Applied Biosystems model 340A nucleic acid extractor.

Table 1

Properties of the 5'HVR and 3'HVR Loci

A. Repeat Size and Variability

Locus	Human Genome Mapping Symbol	Repeat Length (bp)	Size Range (repeats)	Heterozygosity (%)
5'HVR	D16S85	57	5–60	70
3'HVR	D16S85	17–21	15–400	90

B. Internal Repeat Structure

Locus and Type	Sequence[a]	Length (bp)
3'HVR:		
a	AACAGCGACACGGGGGG	17
b1	AACAGCGACACGGGAGG	17
b2	AACAGCGACACGGGGAGG	18
b3	AACAGCGACACGGGGGAGG	19
b4	AACACGCACACGGGGGGAGG	20
c	AACACGCACACGGGAGGGAGG	21
5'HVR:		
consensus	GGGGAGCATTCAGGAGGCCTTCCCGGAGGTAGGGTGGTGGGAAGAAGGGGGTCAGCGT	57

[a] Determined by Jarman et al. (1986) from a cloned 3'HVR allele and by Jarman and Higgs (1988) from a cloned 5'HVR allele. The recognition sequence for *Mnl*I in the 3'HVR repeats is shown underlined. The 5'HVR is composed of many other repeats that differ from the consensus sequence by up to four nucleotides; these are all substitutions and do not alter the length of the repeat.

α-Globin genotypes and haplotypes were determined by restriction-enzyme digestion, agarose gel electrophoresis, Southern blotting, and radiolabeled-probe hybridization using standard techniques (Sambrook et al. 1989). The probes used are described in the legend to figure 1. The 3'HVR and 5'HVR allele sizes were initially determined by agarose gel blotting to obtain estimates of allele size: more accurate sizes of the smaller (<1-kb) alleles were later obtained by electrophoresing restriction enzyme–digested DNA through polyacrylamide gels prior to alkali blotting (Martinson and Clegg 1990). The increased resolution enables alleles differing by a single repeat unit to be distinguished. Further accurate sizing of small 3'HVR alleles was obtained by electrophoresis of PCR-amplified alleles.

PCR Sizing of the 3'HVR

3'HVR alleles are composed of varying numbers of several different repeat units, shown in table 1, that vary both in sequence and in length. It is therefore possible that two alleles, while containing the same number of repeats, may vary slightly in size, because of differences in the type of repeats present; for example, the replacement of a 17-bp *a*-type repeat with an 18-bp *b*2-type repeat will result in two alleles that differ in length by only 1 bp. Such small differences in size were detected for smaller 3'HVR alleles by electrophoresis of radioactively labeled PCR-amplified alleles through sequencing-type polyacrylamide gels.

Genomic DNA (100–250 ng) was amplified in a 50-μl reaction volume containing 45 mM Tris-Cl pH 8.8, 11 mM $(NH_4)_2SO_4$, 4.5 mM $MgCl_2$, 4.5 μM Na_2EDTA, 6.7 mM β-mercaptoethanol, 110 μg of BSA/ml, 1 mM each dNTP, 5 U of thermostable DNA polymerase, and 1 μM each primer (Jeffreys et al. 1990), together with 1 μg of the single-strand DNA binding protein gp32 (Schwarz et al. 1990). One of the primers had been radiolabeled using ^{32}P γ-ATP and T4 polynucleotide kinase (Sambrook et al. 1989) prior to amplification. The sequence of the 5' primer was tggaacaagtaccctgagtcccacctg, and that of the 3' primer was gccgccctgtacaggagtcacttg. Initially, *Taq* polymerase (Cetus) was used, but higher yields were later obtained with VENT polymerase (New England BioLabs) under the same reaction conditions.

The reaction volume was overlaid with mineral oil and was amplified for 25 cycles in a Hybaid OmniGene thermal reactor. The initial cycle had a denaturing temperature of 95°C for 3 min, annealing temperature of 60°C for 1 min, and extension temperature of 72°C for 2 min. The subsequent 24 cycles had identical conditions, except that the denaturing time was reduced to 1 min and the extension time was incremented by 1 s each cycle. Allele sizes were determined by electrophoresis of 2.5-μl aliquots through 3.5% nondenaturing polyacrylamide gels, followed by direct autoradiography of the dried gel. More accurate sizing of smaller fragments was obtained by electrophoresis through 5% denaturing polyacrylamide sequencing gels, using M13 mp8 DNA sequence as a reference standard.

Table 2

5'HVR and 3'HVR Allele Sizes, from 51 Polynesian $-\alpha^{3.7}$ III Homozygotes

Allele Size (bp)	No. Observed	Frequency (%)
5'HVR:		
660	1	.98
1,290	63	61.76
1,460	38	37.25
3'HVR:		
540	33	32.35
580	1	.98
585	1	.98
620	64	62.75
630	2	1.96
650	1	.98

Allele sizes were corrected to take into account differences in size between *Hinf*I-digested and amplified alleles, thus facilitating comparison between amplified 3'HVR alleles and those determined by RFLP analysis.

Statistical Analysis of the 3'HVR and 5'HVR Allele Distributions

The association between VNTR alleles and the intervening haplotype was analyzed quantitatively for each separate VNTR locus by using nonparametric tests. Pairwise comparisons, using the Mann-Whitney test, for each VNTR locus were made for each haplotype observed at more than three chromosomes. The overall difference in VNTR distribution between all such haplotypes was analyzed using the Kruskal-Wallis test.

The association between *both* VNTRs and the intervening haplotype was analyzed using a multivariate approach. Wilk's λ (Mardia et al. 1979) was calculated for the overall distribution of haplotypes. This statistic describes the partition of variation, in both VNTR alleles, between the different haplotypes. The empirical distribution of expected λ values was obtained using computer simulation to reassign haplotypes randomly to each observed VNTR allele pair. The observed value of λ could then be compared with the distribution of expected values, and the probability of generating the observed value by chance was calculated.

Results

Restricted Distribution of 3'HVR Alleles in $-\alpha^{3.7}$III Homozygotes

The populations of Oceania contain high frequencies of many α-globin gene rearrangements that remove one of the two α-genes and cause the phenotype of mild α^+-thalassemia (Flint et al. 1986). Of the many different types of single-α-gene chromosomes present in Oceania, only one, designated "$-\alpha^{3.7}$III," is found at high levels in the populations of eastern Polynesia (Hill et al. 1989). No instances of this deletion chromosome have been found outside Melanesia and Polynesia, and all examples of it—which are thought to be the descendants of a single mutant—are associated with the same α-globin haplotype, the IIIa (Hill et al. 1989).

In the course of this study, we determined the α-globin genotype of 2,163 individuals and obtained 51 individuals homozygous for the $-\alpha^{3.7}$III deletion. All of these homozygotes were found to be homozygous for the type IIIa haplotype, confirming the conclusions drawn from previous surveys (Hill et al. 1989; O'Shaughnessy et al. 1990). These deletion chromosomes have an extremely restricted distribution of alleles at both the 3'HVR and 5'HVR loci (table 2): two alleles at each locus account for >98% of the number seen, with the variants differing slightly, in size, from the two common alleles. For example, the 3'HVR alleles associated with several $-\alpha^{3.7}$III homozygotes and $\alpha\alpha/-\alpha^{3.7}$III heterozygotes are shown in figure 2; the two deletion-associated alleles can be easily identified in heterozygous individuals.

Linkage Relationships between α Haplotypes and Flanking VNTR Alleles

The exclusive restriction of the $-\alpha^{3.7}$III deletion to the IIIa haplotype, as well as the limited distribution of 3'HVR and 5'HVR alleles associated with the deletion, allows both the haplotype and associated HVR alleles of the normal (nonthalassemic) $\alpha\alpha$-gene chromosome in individuals heterozygous for the deletion to be determined simulta-

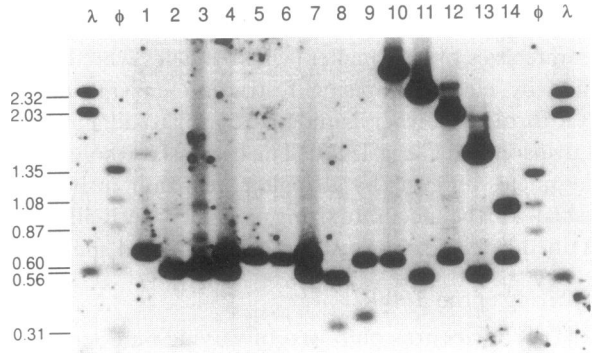

Figure 2 3'HVR alleles seen in homozygous and heterozygous $-\alpha^{3.7}$III deletion samples. Samples 1–7 are $-\alpha^{3.7}$III deletion homozygotes, and samples 8–14 are $\alpha\alpha/-\alpha^{3.7}$III heterozygotes. Samples 1, 5, and 6 are homozygous for the larger of the two deletion-associated alleles that are common in Polynesians; samples 2 and 3 are homozygous for the smaller allele, and samples 4 and 7 are heterozygous. The haplotypes of the $\alpha\alpha$ chromosomes of the heterozygous samples are as follows: samples 8 and 9, Ia; sample 10, IIa; sample 11, IIe; sample 12, IIIa; and samples 13 and 14, IVa. DNA molecular-weight markers shown are λ-*Hin*dIII and ϕX174-*Hae*III digests. Sizes, in kilobases, of the marker bands are given in the left margin. The autoradiograph has been deliberately overexposed in order to reveal the small bands in lanes 8 and 9.

neously. The normal αα haplotype is determined by subtracting the component RFLPs of the IIIa haplotype associated with the $-\alpha^{3.7}$III chromosome from the genotype obtained in the heterozygote, and the 5′HVR and 3′HVR allele sizes can be measured directly from autoradiographs such as that shown in figure 2. We determined the haplotype and the 3′HVR and 5′HVR allele sizes for the Tahitian umbilical cord DNAs in this manner: we concentrated on these samples as they were collected from unrelated individuals, and the ethnic status information obtained allowed us to exclude from the study non-Polynesian genotypes. Of the 1,023 umbilical cord DNAs analyzed, 18 were homozygous for the $-\alpha^{3.7}$III deletion, and 174 were heterozygous, but 41 of these were excluded as having non-Polynesian ancestry, leaving 133 Polynesian chromosomes that were analyzed in detail.

The haplotypes seen were consistent with those determined in previous surveys (Hill et al. 1989; O'Shaughnessy et al. 1990; Hertzberg et al. 1992). A total of 17 different haplotypes, 21 5′HVR alleles, and 46 3′HVR alleles were detected (described in detail in table 3). For each chromosome, these could be combined to give a "superhaplotype" comprising the 5′HVR and 3′HVR allele sizes together with the intervening RFLP haplotype (fig. 3). For each VNTR locus, there is a degree of haplotype-specific association (fig. 3*A* and *B*). However, each chromosome is actually described by three data points: 3′HVR allele size, 5′HVR allele size, and haplotype (fig. 3*C*). The distribution of haplotypes is clearly not random; "clustering" of alleles associated with particular haplotypes occurs. This can be seen most clearly with the IVa haplotype, characterized by a small range of 5′HVR alleles (660–995 bp), and an associated range of 3′HVR alleles (970–2,350 bp). In contrast, the Ia haplotype has an L-shaped distribution in which one arm consists of a small number of 5′HVR alleles associated with many 3′HVR alleles, and the other arm has the converse relationship. The majority of chromosomes with this haplotype have small alleles for both VNTR loci and lie in the bottom left-hand corner of the distribution seen in figure 3*C*.

Statistical Significance of Haplotype-VNTR Allele Association

Pairwise haplotype comparisons, made using the Mann-Whitney test, for each VNTR locus (summarized in table 4) show that the majority of haplotypes differ significantly from each of the others at one or other of the two VNTRs, even when these VNTRs are analyzed independently. Kruskal-Wallis analysis of the whole set of haplotypes (observed at more than three chromosomes each) indicates that the observed extent of haplotype restriction is very highly significant for each separate VNTR (5′HVR $H_{[7]} = 44.34$, $P < .001$; 3′HVR $H_{[7]} = 78.32$, $P \ll .001$), although the lower polymorphism seen at the 5′HVR is reflected in a lower, but still significant, value for H.

The degree of haplotype association becomes more striking when the distribution of *both* VNTRs is considered simultaneously. Simulation of the distribution of Wilk's λ showed that the empirical probability of the observed value being produced by chance was extremely low ($\lambda = .2767$; $P \ll .001$). Thus the association between the two VNTRs and the intervening haplotype as suggested by figure 3*C* is indeed very highly significant.

High-Resolution Sizing of VNTR Alleles Associated with Different Haplotypes

In figure 3*C* there is apparent overlap between some of the haplotype groups; that is, chromosomes bearing different haplotypes appear to share identically sized VNTR alleles. Because of the lower polymorphism of the 5′HVR, there are several instances of different haplotypes sharing the same-sized allele at this locus. This is not the case with the 3′HVR. When 3′HVR allele sizes are determined to a high level of resolution by PAGE, the apparent overlap between different, unrelated haplotypes is due to the different alleles being *similar* in size, rather than *identical;* that is, the distributions interdigitate rather than overlap. This can be seen for the smaller 3′HVR alleles in figure 4, where high-resolution sizing shows that alleles associated with the Ia and IId haplotypes, which are so similar in size as to be indistinguishable by agarose gel electrophoresis, are in fact different. This difference is not a simple multiple of a repeat length but, instead, reflects variation in the internal composition of these alleles (table 1). It is therefore possible for two alleles to be very similar in size (even to the extent that they appear identical unless viewed at high resolution) but actually to belong to different subpopulations of chromosomes when linked RFLPs are considered. The high-resolution sizing approach shown in figure 4 has also revealed slight variation in HVR allele size within a haplotype, indicating subtle changes in repeat-unit composition.

New Haplotype Relationships Revealed by Flanking VNTR Loci

The fine-scale size differences between 3′HVR loci revealed at high resolution explain many, but not all, of the apparent overlaps seen in figure 3. In figure 4*A*, haplotypes Ia and IIc have 3′HVR alleles that are identical in size at the base-pair level. Although the nomenclature used for α-globin haplotypes implies that these haplotypes are very different (see legend to fig. 1), their structures are similar and may reflect common ancestry. In fact, haplotypes Ia (++−MPZ++−−) and IIc (++−MPZ+−−−) could differ from each other by as little as one single base pair affecting the *Rsa* I site; the IIc could in fact have more sequence similarity with the Ia haplotype than with the other type II haplotypes, where both site changes and length changes to the inter-ζ HVR would be required to produce, say, the IIa (−+−LPZ+−−−). The HVR distri-

Table 3

Allele Distributions of 5′HVR and 3′HVR on Polynesian αα Haplotypes

Haplotype	RFLP	5′HVR							3′HVR					
		No. of Chromosomes	No. of Alleles	Mean	Median	SD	Minimum	Maximum	No. of Alleles	Mean	Median	SD	Minimum	Maximum
Ia	++−MPZ++−−	27	7	889.3	775	350.9	715	2,500	9	457.7	310	284.7	308	1,450
Inovel	+−−SPZ++−−	1	1	775.0	775	...	775	775	1	1,580.0	1,580	...	1,580	1,580
IIa	−+−LPZ+−−−	9	8	1,279.0	940	704.0	715	2,700	8	1,809.0	1,480	800.0	650	2,850
IIc	++−MPZ+−−−	4	1	775.0	775	...	775	775	2	398.3	310	153.0	310	575
IId	−−−SPZ+−−−	9	2	781.1	775	18.3	775	830	2	353.3	320	100.0	320	620
IIe	+−−SPZ+−−−	11	4	727.3	715	71.2	660	830	9	1,282.0	1,070	449.0	750	2,300
ζζζ	++−MPZ/−−SPZ+−−−	21	3	814.5	775	73.0	775	1,110	9	2,185.0	1,500	905.0	1,400	3,500
ζζζ+	++−MPZ/−−SPZ+−−−	3	3	1,017.0	1,110	211.0	775	1,165	3	3,267.0	3,200	208.0	3,100	3,500
IIg	+−−MPZ+−−−	1	1	715.0	715	...	715	715	1	1,800.0	1,800	...	1,800	1,800
IIIa	−−+M Z−−−−	10	9	1,436.0	1,225	789.0	660	2,550	9	1,807.0	2,050	712.0	620	2,600
IIIb	+−+M Z−−−−	1	1	660.0	660	...	660	660	1	1,050.0	1,050	...	1,050	1,050
IIIc	−+−L Z−−−−	2	2	857.5	857	38.9	830	885	2	2,250.0	2,250	707.2	2,200	2,300
IIIf	−−−M Z−−−−	2	2	1,463.0	1,463	972.0	775	2,150	2	1,550.0	1,550	849.0	950	2,150
IIIi	+++M Z−−−−	1	1	830.0	830	...	830	830	1	2,050.0	2,050	...	2,050	2,050
IVa	+−−SPZ+−++	28	6	719.1	660	80.6	660	995	14	1,482.2	1,150	499.1	970	2,350
IVb	−−−SPZ+−++	1	1	885.0	885	...	885	885	1	2,200.0	2,200	...	2,200	2,200
Vc	+−−SPZ+−−+	1	1	775.0	775	...	775	775	1	1,650	1,650	...	1,650	1,650

NOTE.—The triplicated-zeta chromosome (denoted as "ζζζ" above) is a recombinant chromosome found at high (10%–20%) frequencies in the populations of Southeast Asia and Polynesia. Previous studies have shown that this occurs on a recombinant Ia/IIe haplotype (O'Shaughnessy et al. 1990). In Polynesia, ~10% of the ζζζ chromosomes carry an additional *Bgl*II polymorphism (denoted here by ζζζ+).

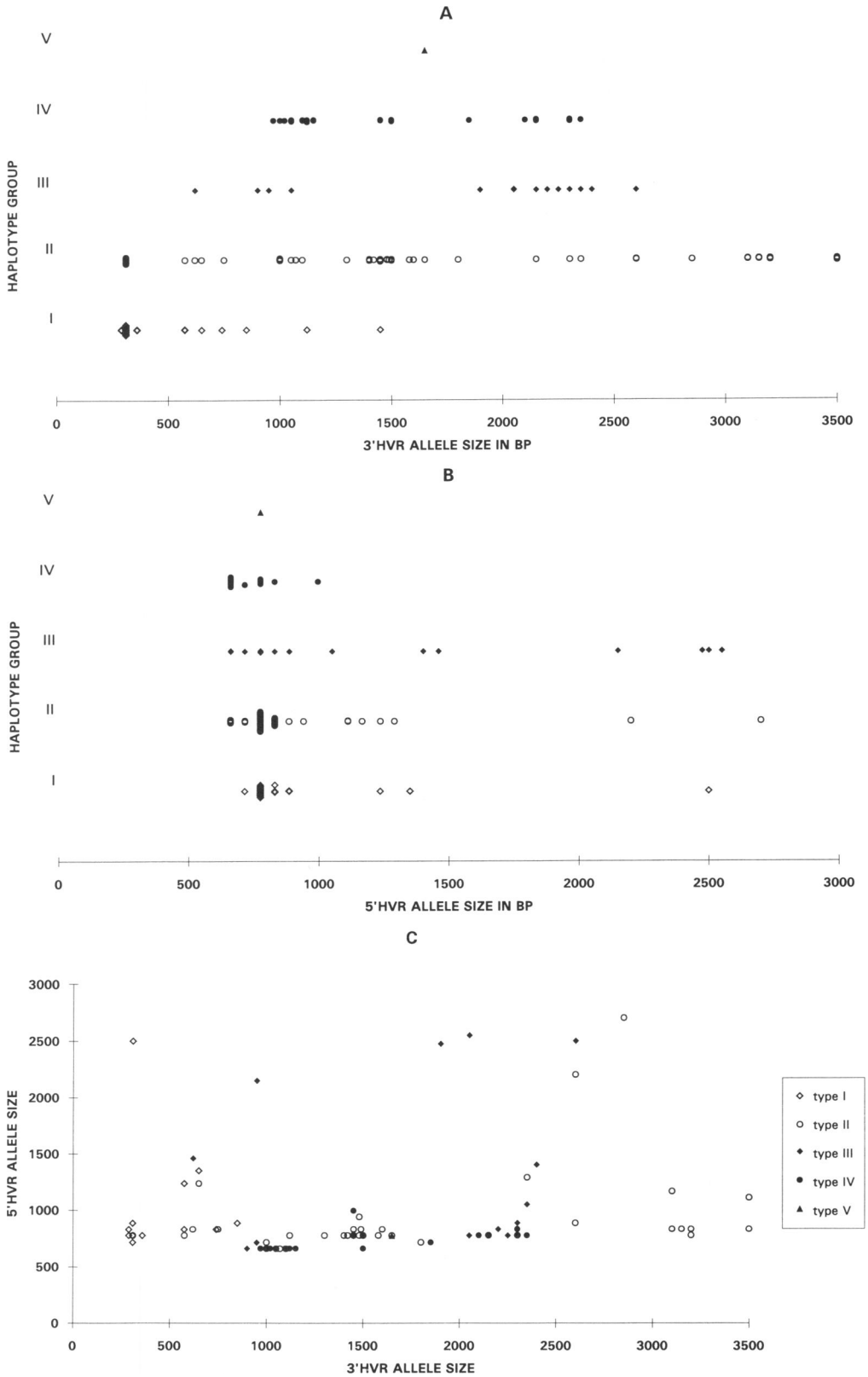

Figure 3 5′HVR and 3′HVR alleles in Polynesia, showing haplotype associations. The allele frequency distributions of the 3′HVR (*A*) and 5′HVR (*B*) for the normal, αα chromosomes of 133 Polynesian αα/$-\alpha^{3.7}$III–deletion heterozygotes are shown separately for each haplotype group. In panel C, the data for the two VNTRs are combined in a scatterplot where 5′HVR allele size in base pairs is plotted against 3′HVR allele size in base pairs. The major haplotype groups are identified.

Table 4

Pairwise Mann-Whitney Comparisons of 3′HVR and 5′HVR Allele distributions on Polynesian αα Haplotypes

	Ia	IIa	IIc	IId	IIe	ζζζ	IIIa	IVa
Ia		<.05	. . .[a]	NS	<.05	NS	NS	<.001
IIa	<.001		. . .[a]	<.05	<.01	NS	NS	<.001
IIc	NS	<.01		. . .[a]	. . .[a]	. . .[a]	. . .[a]	. . .[a]
IId	NS	<.001	NS		<.05	NS	NS	<.01
IIe	<.001	NS	<.01	<.001		<.01	<.05	NS
ζζζ	<.001	NS	<.01	<.001	<.05		NS	<.001
IIIa	<.001	NS	<.01	<.001	NS	NS		<.001
IVa	<.001	NS	<.01	<.001	NS	<.001	NS	

NOTE.—Data are *P* values at which *W* values are significant. Comparisons between 5′HVR allele distributions are shown above the diagonal, and comparisons for the 3′HVR are shown below the diagonal. NS = not significant.

[a] 5′HVR alleles associated with the IIc haplotype are all identical; hence the Mann-Whitney test cannot be performed, as the variance of that allele distribution is zero.

butions of the Ia, IIc, and IIa haplotypes in figure 5*A* show that the distribution associated with the IIc is a subset of that seen with the Ia and has no overlap with the distribution seen with the IIa. The HVR allele distributions associated with the IIc haplotype are not significantly different from those associated with the Ia; however, they differ significantly from the other group II haplotype allele distributions (table 4). Conversely, the VNTR allele distributions of the IId (−−−SPZ+−−−) and IIe (+−−SPZ+−−−) haplotypes, which could also theoretically differ from each other by a single base change, are significantly different, suggesting that these two haplotypes are similar by convergence, rather than as a result of a recent common ancestry.

The extensive distribution of group II haplotypes, shown in figure 3*C*, arises partly from the heterogeneous composition of this group in particular. As haplotype nomenclature is arbitrarily based on the 3′ end RFLP sites, a mutation that affects these RFLPs will produce a member of a different haplotype group, whereas a mutation affecting the 5′-end RFLPs will merely result in a novel member of the same group; this may explain why the Ia and IIc haplotypes described above are so similar. Group II haplotypes have a more extensive range of 5′ RFLPs than does any other haplotype group, and most group II haplotypes are more similar to members of other haplotype groups than they are to other group II haplotypes. This heterogeneity can clearly be seen in figure 5*B*, where the different group II haplotypes are identified separately. There is very little overlap between the different haplotypes, with the exception of the IIa, which has the broadest distribution of VNTR alleles seen in this survey. This may reflect its ubiquitous distribution in many of the world's peoples. These effects can also be seen in the quantitative analysis shown in table 4: the distribution of 3′HVR alleles associated with the IIa haplotype is not significantly different from the distributions seen with the other group II haplotypes, with the notable exceptions of the IIc haplotype (which, as explained above, resembles the Ia haplotype more than it does other group II haplotypes) and the IId haplotype, which has a particularly restricted distribution of 3′HVR alleles that lie outside the range of the IIa.

Discussion

The α-globin 3′HVR is an extremely polymorphic VNTR locus with heterozygosities approaching 100% in several populations (Jarman et al. 1986). As a consequence of this high mutability, distributions of alleles have been generated that we have shown are associated with different nearby RFLP haplotypes. In the absence of this haplotype information, the distribution of 3′HVR alleles and, to a lesser extent, 5′HVR alleles appears random or unstructured. Because of the apparent overlap between allele distributions associated with different haplotypes, it is difficult to discriminate between 3′HVR alleles on the basis of allele size alone, particularly if the size determination for these loci is on the basis of agarose-gel Southern blot hybridization. The extensive substructuring that exists within the overall distribution of alleles becomes apparent only when haplotype information is considered. This is relevant when one is considering the utility of VNTR loci in forensic work, as 3′HVR alleles originating in different populations can be similar or indistinguishable in size: for example, the 3′HVR-allele size distributions associated with the group III haplotypes (confined in this region to Melanesian and Polynesian populations) overlap considerably with those associated with the group I and II haplotypes (found in this region only in the populations of Southeast Asia and Polynesia). The current forensic practice of "binning" allele sizes conservatively (Budowle et al. 1991*b*) will, however, tend to reduce the impact of this

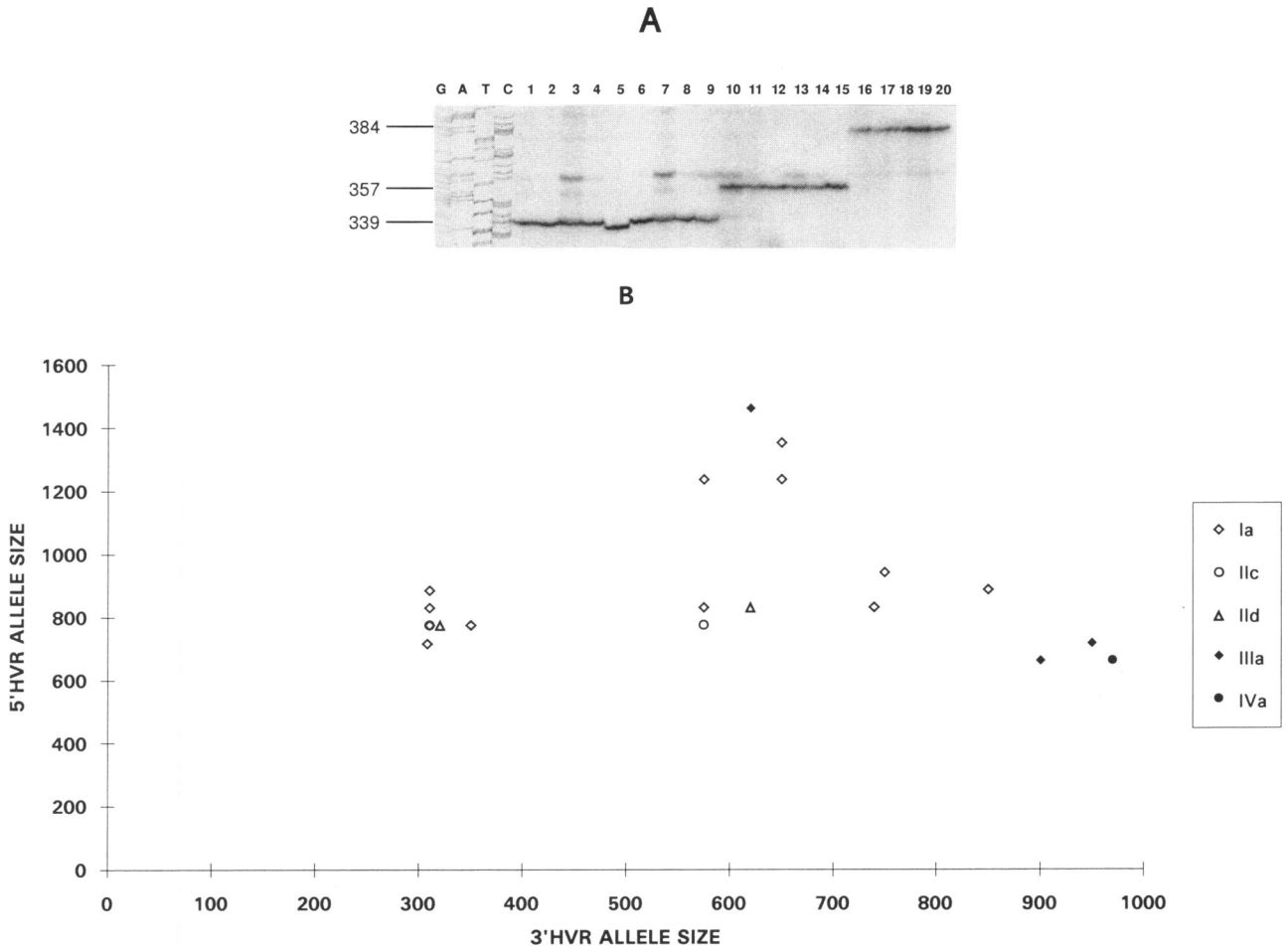

Figure 4 HVR allele sizes at high resolution. *A*, PCR-amplified 3′HVR alleles electrophoresed through a 5% polyacrylamide denaturing sequencing gel, with DNA sequence from bacteriophage M13mp18 as a molecular size marker. Each lane shows the 3′HVR allele associated with the normal, αα chromosome in individuals heterozygous for the $-\alpha^{3.7}$III deletion. Samples 1–7 and 16–20 have the Ia haplotype on their αα chromosome, samples 8 and 9 have the IIc haplotype, and samples 10–15 have the IId haplotype. The sizes of the three common bands are indicated in the margin. On the basis of the original restriction-digest and agarose-gel screening of these samples, the sizes of the IId haplotype 3′HVR alleles were erroneously assigned to one or other of the Ia haplotype groups; further, in 3′HVR allele the 2-bp size difference between sample 5 and the other Ia haplotypes was only detectable by the PCR-denaturing gel approach shown here. *B*, 3′HVR allele sizes determined by the aforementioned approach, plotted here at a higher resolution than that shown in the previous figures. With the exception of the Ia and IIc haplotypes discussed in the text, no other haplotypes have identical VNTR allele pairs.

phenomenon when the VNTR alleles are very similar in size.

The variety of repeat units that make up the 3′HVR array provides additional structural complexity, as two similarly or identically sized arrays may have radically different internal structures. As the different repeats vary in size as well as sequence, it is also possible for alleles to differ from one another by less than the length of a repeat unit; for example, the replacement of an *a*-type repeat with a *c*-type repeat would alter the size of the allele by 4 bp, a size difference that would be undetectable by most approaches. Such small internal structure differences can already be seen in figure 5, where one rare 3′HVR allele associated with the Ia haplotype differs from the more common allele by 2 bp. The full extent of such differences becomes apparent only when the allele internal structures are determined by internal MVR mapping (Jeffreys et al. 1991; Desmarais et al. 1993) or sequencing.

We are currently developing such an approach for the study of the α-globin VNTRs, using a modification of the techniques described by Smith and Birnstiel (1976) for use with PCR-amplified DNA fragments (authors' unpublished data): preliminary results (table 5) show that the size heterogeneity seen at the 3′HVR is indeed reflected in substantial differences in internal repeat composition. These differences reflect the population of origin of the different haplotypes, at least on the basis of the data so far obtained: the Melanesian $-\alpha^{3.7}$III deletion/IIIa haplotype chromosome 3′HVR is composed entirely of *a*- and *b*-type repeats,

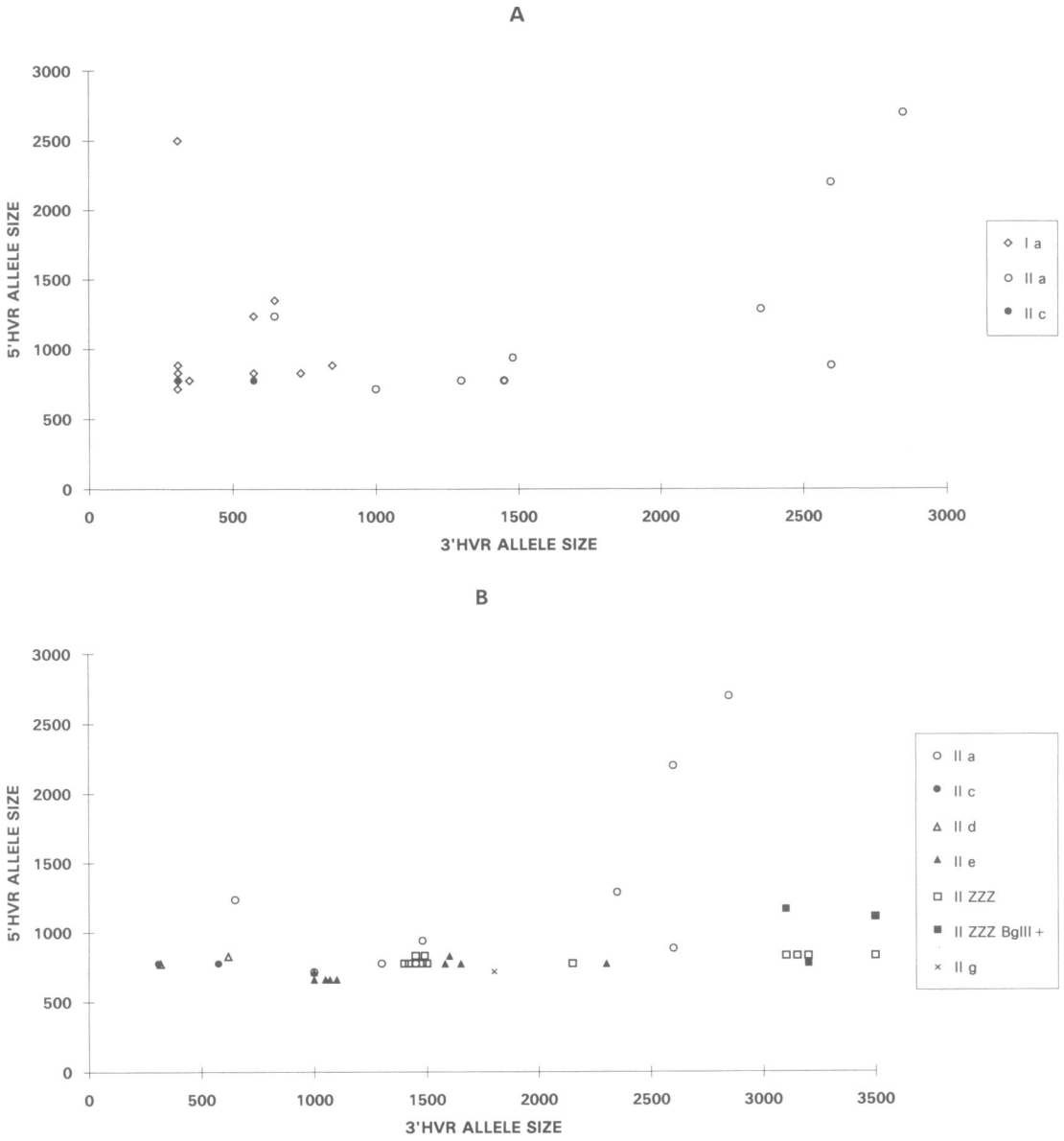

Figure 5 HVR alleles associated with group I and II haplotypes. *A*, 5′HVR and 3′HVR allele sizes, plotted here as in fig. 3*C*, for the Ia, IIc, and IIa haplotypes. The IIc allele distribution is wholly contained within that of the Ia and has no overlap with the distribution of the IIa. There is slight overlap, however, between the distribution seen with the Ia haplotype and that seen with the IIa haplotype. *B*, 5′HVR and 3′HVR allele sizes, plotted for the different group II haplotypes. The distribution of alleles associated with the ζζζ chromosome is bimodal, with clusters centered on 3′HVR allele sizes of 1,500 and 3,250 bp. The ζζζ/*Bgl*II+ chromosomes, however, all cluster around the larger 3′HVR distribution, indicating that the recent mutation that gave rise to this chromosome occurred on a ζζζ chromosome with a larger 3′HVR allele.

whereas the Southeast Asian Ia haplotype 3′HVR contains *a*-, *b*-, and *c*-type repeats. The map data so far obtained also reveal and confirm relationships between 3′HVR alleles: the 2-bp size difference between the Ia-associated alleles shown in figure 4*A* is due solely to the replacement of one *b*-type repeat with another (indicated as "B" in table 5; the resolution of the electrophoretic system used to determine the maps did not allow the identification of the precise repeat types involved). Also, the internal structures of the identically sized 3′HVR alleles associated with the Ia and IIc haplotypes are identical, confirming a strong relationship between these haplotypes in this population.

Further analysis of the variation in internal structure of the 3′HVR will doubtless reveal in more detail the population structure and genetic mutation processes of this locus. Already, the nonrandom association of haplotypes and HVRs indicates that interchromosomal recombination between heterologous haplotypes—which would exchange HVR alleles and thus homogenize the distributions associated with each haplotype—are not occurring at a de-

Table 5

Internal Map Structure of Polynesian 3′HVR Alleles

Haplotype	Allele Size[a] (bp)	No. Studied	Map
$-\alpha^{3.7}$III[b]	620	12	baaabbaabaaaaaabbababbabaabbba
$-\alpha^{3.7}$III[b]	540	10	baaabbaabaaaaaabbaba-----abbba
Ia	310	6	bbaacccabaa
IIc	310	2	bbaacccabaa
Ia	308	1	Bbaacccabaa
Ia	355	4	bbaaaaabccaba

[a] Based on *Hin*fI digests of genomic DNA.
[b] On IIIa haplotype.

tectable rate in this population. Also, the discrete distributions of 3′HVR alleles associated with different haplotypes show that reciprocal recombination between different 3′HVR alleles does not occur at a high rate. Additional evidence for this is the observation of characteristic repeat patterns in 3′HVR alleles associated with different haplotypes (table 5). High levels of interchromosomal recombination would exchange repeats and remove this association. From the data presented here, it is apparent that interchromosomal recombination is not responsible for the hypervariability at these loci but that intrachromosomal processes such as unequal sister-chromatid exchange (USCE) or replication slippage are much more prevalent (Wolff et al. 1988, 1989). These processes are presumably responsible for the clustering of rare alleles around common ones of similar size, as is seen with the type I haplotypes.

Similar conclusions have been drawn from analyses of VNTR allele polymorphism and association with nearby sequence polymorphism for other VNTR loci. Studies on the insulin locus have revealed several polymorphisms that are in strong linkage disequilibrium with each other; this linkage extends across the insulin VNTR and encompasses a region of ~40 kb (Cox et al. 1988; Lucassen et al. 1993). At the HRAS1 VNTR, nearby linked markers have been used to determine the progenitor allele involved in the production of new alleles (Kasperczyk et al. 1990), and mutant alleles have been implicated in susceptibility to cancer (Krontiris et al. 1993), either by linkage to nearby deleterious mutations or possibly by direct effect on gene expression (Green and Krontiris 1993). Renges at al. (1992) have studied extensively the relationship between the apolipoprotein B 3′VNTR and four polymorphic sites within the gene region and have shown that strong linkage disequilibrium exists between each diallelic site and the VNTR. Further, they show that certain VNTR alleles occur predominantly on specific haplotypes, defining a region of close association extending over ⩾40 kb. Our work on the α-globin complex here shows that such a relationship exists over a region approximately two and a half times as long and that it encompasses two highly polymorphic VNTRs, as well as the less polymorphic ones used in the construction of the haplotype (fig. 1).

Small, nonintegral variations of allele size are not characteristic of the 5′HVR, which has an unvarying 57-bp repeat motif (Jarman and Higgs 1988). Within this size constraint, however, there are several sequence variants, many of which contain the recognition sequence for the restriction enzyme *Stu*I. This locus is therefore as amenable as the 3′HVR to internal repeat mapping and sequencing approaches that will facilitate the characterization of apparently identically sized alleles. Although the 5′HVR has a lower heterozygosity (and presumably a lower mutation rate) than the 3′HVR, it is clear from our data that there is haplotype specificity within its allele size distributions; this can be seen, for example, for the group III and IV haplotypes in figure 3C. This is likely to become more pronounced when the additional layer of sequence complexity is incorporated into the 5′HVR descriptions.

Thus both the 5′HVR allele and the 3′HVR allele exhibit considerable haplotype dependency in their respective allele size characteristics. Moreover, the α-globin haplotypes, like their β-globin counterparts (Wainscoat et al. 1986), can be used to create a phylogeny that well describes human population relationships. It follows, then, that these two VNTR loci have very significant in-built population-specific information in their allele distributions.

The phenomena observed here in the Polynesian population are particularly apparent because of the overall reduced diversity of 3′HVR and 5′HVR alleles in this population (Hertzberg et al. 1992). This consequence of the specific demographic history of the population studied here merely makes more noticeable a process that is likely to be occurring in other populations but that may be obscured by higher allelic diversity. It seems likely that a detailed survey of different loci in different populations would reveal that VNTR alleles are generally associated with restricted subsets of flanking markers. If this is indeed the case, then current assumptions of random allele distributions are likely to mislead.

The results shown here for two VNTR loci on chromosome 16 indicate that the distributions of VNTR allele sizes alone do not describe the total amount of variation present at that locus. Previous studies have demonstrated that the internal structures of some VNTR loci are themselves polymorphic (Jeffreys et al. 1990; Desmarais et al. 1993), and this is true of the α-globin VNTRs (table 5). Here we present the first indication that these internal structure and size polymorphisms are linked to discernible polymorphisms some considerable distance from the VNTR loci. This linkage association is not merely seen with polymorphisms in the sequences flanking the VNTR arrays, but extends over a 100-kb region encompassing several RFLP sites and two separate VNTR loci. In the absence of such flanking-marker information, false assignments may be made if allele size alone is the criterion of identity: we have shown here that two VNTR alleles may be so similar in size as to be indistinguishable by conventional blotting approaches but may also be totally unrelated to one another in terms of their internal structures, linked polymorphisms and, most important, the populations in which the different alleles originally arose.

Acknowledgments

We are grateful to Drs. J. Roux and G. Philippon and the staff of the Institute Territorial de Recherche Médicales Louis Malardé, Papeete, French Polynesia, for their generous assistance with the collection and transport of the samples studied here. We are also grateful to Mr. D. L. Neil for technical assistance with the early stages of this project; to Dr. A. Grafen for advice on the statistical analysis of the HVR data; and to Drs. J. Flint, R. M. Harding, and T. E. A. Peto for their comments on an earlier draft of the manuscript. This work was supported by the Wellcome Trust and the Medical Research Council, including a Research Studentship to J.J.M.

References

Armour JAL, Harris PC, Jeffreys AJ (1993) Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized diversity. Hum Mol Genet 2:1137–1145

Armour JAL, Jeffreys AJ (1992) Biology and applications of human minisatellite loci. Curr Opin Genet Dev 2:850–856

Balazs I, Baird M, Clyne M, Meade E (1989) Human population genetic studies of five hypervariable DNA loci. Am J Hum Genet 44:182–190

Bellwood PS (1989) The colonization of the Pacific: some current hypotheses. In: Hill AVS, Serjeantson SW (eds) The colonization of the Pacific: a genetic trail. Oxford University Press, Oxford, pp 1–59

Budowle B, Chakraborty R, Giusti AM, Eisenberg AJ, Allen RC (1991*a*) Analysis of the VNTR locus D1S80 by PCR followed by high-resolution PAGE. Am J Hum Genet 48:137–144

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

Cox NJ, Bell GI, Xiang K-S (1988) Linkage disequilibrium in the human insulin/insulin-like growth factor II region of human chromosome 11. Am J Hum Genet 43:495–501

Desmarais E, Vigneron S, Buresi C, Cambian F, Cambou JP, Roizes G (1993) Variant mapping of the Apo(B) AT rich minisatellite: dependence on nucleotide sequence of the copy number variations: instability of the non-canonical alleles. Nucleic Acids Res 21:2179–2184

Flint J, Boyce AJ, Martinson JJ, Clegg JB (1989) Population bottlenecks in Polynesia revealed by minisatellites. Hum Genet 83:257–263

Flint J, Hill AVS, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, et al (1986) High frequencies of α-thalassemia are the result of natural selection by malaria. Nature 321:744–749

Gibbons A (1994) Genes point to a new identity for Pacific pioneers. Science 263:32–33

Goodbourn SEY, Higgs DR, Clegg JB, Weatherall DJ (1983) Molecular basis of length polymorphism in the human ζ-globin gene complex. Proc Natl Acad Sci USA 80:5022–5026

Green M, Krontiris TG (1993) Allelic variation of reporter gene activation by the HRAS1 minisatellite. Genomics 17:429–434

Hertzberg MS, Mickleson KNP, Trent RJ (1992) Limited genetic diversity in Polynesians reflected in the highly polymorphic 3'HVR α-globin marker. Hum Hered 42:157–161

Higgs DR, Wainscoat JS, Flint J, Hill AVS, Thein SL, Nicholls RD, Teal H, et al (1986) Analysis of the human α-globin gene cluster reveals a highly informative genetic locus. Proc Natl Acad Sci USA 83:5165–5169

Hill AVS, Bowden DK, Trent RJ, Higgs DR, Oppenheimer SJ, Thein SL, Mickleson KNP, et al (1985*a*) Melanesians and Polynesians share a unique α-thalassemia mutation. Am J Hum Genet 37:571–580

Hill AVS, Gentile B, Bonnardot JM, Roux J, Weatherall DJ, Clegg JB (1987) Polynesian origins and affinities: globin gene variants in eastern Polynesia. Am J Hum Genet 40:453–463

Hill AVS, Nicholls RD, Thein SL, Higgs DR (1985*b*) Recombination within the human embryonic ζ-globin locus: a common ζ-ζ chromosome produced by gene conversion of the ψζ gene. Cell 42:809–819

Hill AVS, O'Shaughnessy DF, Clegg JB (1989) Haemoglobin and globin gene variants in the Pacific. In: Hill AVS, Serjeantson SW (eds) The colonization of the Pacific: a genetic trail. Oxford University Press, Oxford, pp 246–285

Jarman AP, Higgs DR (1988) A new hypervariable marker for the human α-globin gene cluster. Am J Hum Genet 43:249–256

Jarman AP, Nicholls RD, Weatherall DJ, Clegg JB, Higgs DR (1986) Molecular characterization of a hypervariable region downstream of the human α-globin gene cluster. EMBO J 5: 1857–1863

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60:478–485

Kasperczyk A, DiMartino NA, Krontiris TG (1990) Minisatellite

allele diversification: the origin of rare alleles at the HRAS1 locus. Am J Hum Genet 47:854–859

Krontiris TG, Devlin B, Karp DD, Robert NJ, Risch N (1993) An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. N Engl J Med 329:517–523

Lauer J, Shen C-KJ, Maniatis T (1980) The chromosomal arrangement of human α-like globin genes: sequence homology and α-globin gene deletions. Cell 20:119–130

Lucassen AM, Julier C, Beressi J-P, Boitard C, Froguel P, Lathrop M, Bell JI (1993) Susceptibility to insulin dependent diabetes mellitus maps to a 4.1kb segment of DNA spanning the insulin gene and associated VNTR. Nature Genet 4:305–310

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London

Martinson JJ, Clegg JB (1990) Alkaline transfer of small restriction fragments from polyacrylamide gels. Nucleic Acids Res 18:1307

Monckton DG, Tamaki K, Macleod A, Neil DL, Jeffreys AJ (1993) Allele-specific MVR-PCR analysis at minisatellite D1S8. Hum Mol Genet 2:513–519

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

Neil DL, Jeffreys AJ (1993) Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. Hum Mol Genet 2:1129–1135

Old JM, Higgs DR (1983) Gene analysis. In: Weatherall DJ (ed) The thalassaemias. Churchill Livingstone, Edinburgh, pp 74–102

O'Shaughnessy DF, Hill AVS, Bowden DK, Weatherall DJ, Clegg JB (1990) Globin genes in Micronesia: origins and affinities of Pacific island peoples. Am J Hum Genet 46:144–155

Proudfoot NJ, Gil A, Maniatis T (1982) The structure of the human ζ-globin gene and a closely linked, nearly identical pseudogene. Cell 31:553–563

Renges H-H, Peacock R, Dunning AM, Talmud P, Humphries SE (1992) Genetic relationship between the 3′-VNTR and diallelic apolipoprotein B gene polymorphisms: haplotype analysis in individuals of European and south Asian origin. Ann Hum Genet 56:11–33

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Schwarz K, Hansen-Hagge T, Bartram C (1990) Improved yields of long PCR products using gene 32 protein. Nucleic Acids Res 18:1079

Smith HO, Birnstiel ML (1976) A simple method for DNA restriction site mapping. Nucleic Acids Res 3:2387–2398

Wainscoat JS, Hill AVS, Boyce AJ, Flint J, Hernandez M, Thein SL, Old JM, et al (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319:491–493

Wolff RK, Nakamura Y, White R (1988) Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking DNA sequence. Genomics 3:347–351

Wolff RK, Plaetke R, Jeffreys AJ, White R (1989) Unequal crossing-over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. Genomics 5:382–384