Variable Number of Tandem Repeat (VNTR) Markers for Human Gene Mapping

YUSUKE NAKAMURA, MARK LEPPERT, PETER O'CONNELL, ROGER WOLFF, TOM HOLM, MELANIE CULVER, CINDY MARTIN, ESTHER FUJIMOTO, MARK HOFF, ERIKA KUMLIN, RAY WHITE

A large collection of good genetic markers is needed to map the genes that cause human genetic diseases. Although nearly 400 polymorphic DNA markers for human chromosomes have been described, the majority have only two alleles and are thus uninformative for analysis of genetic linkage in many families. A few known marker systems, however, detect loci that respond to restriction enzyme cleavage by producing a fragment that can have many different lengths. This polymorphism is due to variation in the number of tandem repeats of a short DNA sequence. Because most individuals will be heterozygous at such loci, these markers will provide linkage information in almost all families. Ten oligomeric sequences derived from the tandem repeat regions of the myoglobin gene, the zeta-globin pseudogene, the insulin gene, and the X-gene region of hepatitis B virus, were used to develop a series of single-copy probes. These probes revealed new, highly polymorphic genetic loci whose allele sizes reflected variation in the number of tandem repeats.

A COMPLETE DESCRIPTION OF THE ANATOMY OF THE HUMAN genome is under way. A number of genes causing inherited disease have been localized on human gene maps, enabling clinicians to identify many individuals at risk. DNA cloning and sequencing technologies are revealing our common genetic heritage and are detecting genetic variation among individuals at its most fundamental level, that of variation in the sequence of DNA bases.

Many important human genes, however, including the great majority of mutant genes that cause human genetic disease, are not at present readily amenable to cloning and sequencing. For most human genetic diseases no gene product is available and biochemical knowledge is limited, making cloning of these genes through established procedures difficult to envision. Loci for such genes can often be mapped, however, by correlating the inheritance (segregation) of a disease trait with the inheritance of a specific chromosomal region, that is, through studies of genetic linkage in families. When a mutant gene is localized to a specific region of a chromosome with sufficient precision, application of powerful DNA technologies that are becoming available for working with very large DNA segments may lead to identification of the gene responsible for the disease.

The authors are at the Howard Hughes Medical Institute and Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84132.

Gene Mapping Through Linkage

Within the last few years, linkage analysis has localized the genes responsible for several major genetic diseases, including Huntington's disease (*1*), Duchenne muscular dystrophy (*2*), polycystic kidney disease (*3*), and cystic fibrosis (*4–6*). To map a disease gene by the linkage approach, we ask a similar question for each of a number of chromosomal regions: within each family in which a genetic disease is segregating, is there a positive correlation in inheritance of a specific chromosomal region with the inheritance of the disease gene? If so, the gene causing the disease may be located within that chromosomal region. In order to determine that the co-segregation seen is not due to chance, however, it is necessary to examine many inheritances and, often, many affected families. Furthermore, it is essential to be able to determine which of the two copies of a chromosomal region has been inherited by an individual from each parent; in other words, the inherited segment must carry a distinguishable version (allele) of a known marker locus.

For example, Fig. 1, A and B, represents a family segregating a dominant disease gene. If the two copies of the chromosomal region in question (marked here by a locus with two alleles) cannot be distinguished, as in Fig. 1A, it will not be possible to determine which copy has been inherited and no linkage information can be obtained. However, in a situation illustrated by Fig. 1B, the affected parent carries two distinguishable alleles (is heterozygous) for the marker locus; allele 1 appears to be co-segregating with the disease trait. The importance of heterozygosity is also illustrated in Fig. 1C. If the family represented by the pedigree were segregating an autosomal recessive disorder, probe B (both parents are heterozygous for the marker) would be more useful than probe A (one parent is homozygous for that marker) in linking the marker to the disease.

The frequency of heterozygosity for a marker system among individuals in the population is, therefore, extremely important in defining that marker's usefulness in providing linkage information from family studies. As alleles at genetic loci are distributed essentially by chance among humans, the expected frequency of heterozygosity at a marker locus (its informativeness) is directly related to the number of common alleles present in the population. The more alleles a locus exhibits, the more likely it is that an individual will be heterozygous.

Furthermore, because most genetic diseases are rare, the informativeness of a marker system for a disease locus can determine whether a linkage study to map a genetic disease is even feasible, given limited family resources. It follows that development of a large number of highly polymorphic marker systems for the human

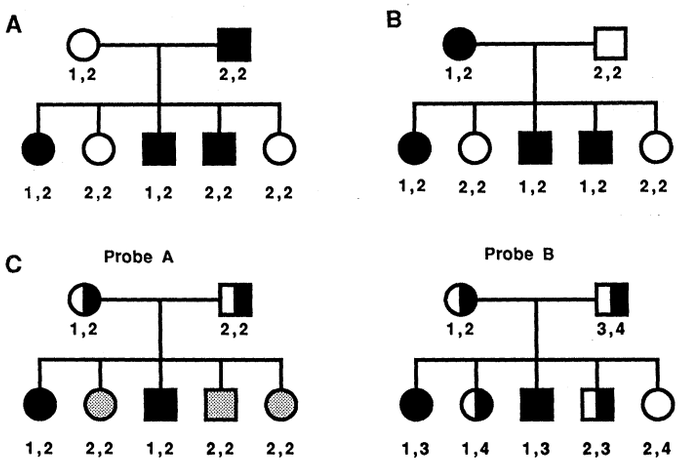

Fig. 1. Informativeness of the probes. Pedigree charts illustrate the models of autosomal dominant (**A** and **B**) and recessive (**C**) genetic disease in a family. Probe A is a site-polymorphic DNA marker that has two alleles and probe B is a highly polymorphic DNA marker that has many alleles. ●, ■ : Affected individual; ○, □ : normal individual; ◑, ◨ : carrier; ◎, ▣ : carrier status uncertain. Round symbols designate females; squares, males.

genetic map would accelerate the localization of disease genes, because the likelihood would increase that the genes of interest could be localized by genetic linkage to known markers.

DNA Markers

Variations in DNA sequence along the genome are common (*7*). Most of the known DNA sequence variations are apparently due to base-pair (bp) changes that create or destroy a cleavage site for a specific restriction enzyme, causing a change in the length of a DNA fragment. A marker locus based on a single such variant will have only two alleles (a single cleavage site is either present or absent) and the chance that a parent will have two different alleles at the locus is always less than 50%. It is sometimes possible to identify several such restriction site polymorphisms very close to one another. This approach can be very useful when it is important that a specified locus, such as β-hemoglobin (*8*), be developed into a highly informative genetic marker. However, genotyping with marker systems that reflect multiple site polymorphisms can be cumbersome; it often requires several probes and/or the digestion of DNA samples with several restriction enzymes.

Hypervariable Loci

In contrast, a few known DNA marker systems reveal a DNA restriction fragment whose length is highly variable within the population. The first report of a polymorphic DNA locus discovered with an arbitrary DNA probe revealed such a system (*9*), with fragments of more than 15 different lengths observed in a small sample of unrelated individuals. Subsequently, similar systems were observed at several other loci, including the insulin gene (*10*), the Harvey-*ras* oncogene (*11*), the zeta-globin pseudogene (*12*), and the myoglobin gene (*13*) loci.

The DNA base sequences of these hypervariable loci indicate that the restriction fragments contain a set of tandem repeats of a short (11 to 60 bp) oligonucleotide sequence. The length of the restriction fragment is a function of the number of copies of the tandem repeats present within the fragment. A marker system based on this type of polymorphism, in contrast to change in a single base pair, can be highly informative in linkage analysis when it exhibits multiple alleles (*3*, *14*). To identify a genetic sequence that contains tandem repeats but represents *only a single locus*, we designate it a variable number of tandem repeats (VNTR) locus.

Jeffreys *et al.* (*13*) recently demonstrated that a DNA probe based on a set of tandem repeats associated with the myoglobin gene locus can detect, by hybridization to human genomic DNA, a number of loci containing tandem repeats of similar sequence. The restriction fragment pattern revealed by the sum of the VNTR loci containing such related sequences, scattered throughout the genome, can constitute a genetic "fingerprint" unique to an individual. This "fingerprint" may be useful for forensic and other applications. However, the large number of restriction fragments that are jointly revealed at a number of loci by partial homology to the tandem repeat sequence probe make interpretation of the resultant allelic series a formidable task. Fortunately, the same sequence from the myoglobin hypervariable region that reveals genetic "fingerprints" in total human DNA can also serve to screen DNA libraries and to identify clones representing unique loci (*15*).

We have implemented and extended this general approach by using synthetic oligonucleotide sequences from several of the known VNTR loci, as well as other candidate sequences, as probes of human genomic libraries. Many new, individual VNTR loci have been identified and developed as genetic marker systems.

Detection and Characterization of VNTR Loci

To determine whether any of the previously known VNTR sequences would identify new families of VNTR loci, we constructed a series of oligonucleotides corresponding to the consensus sequences of the tandem repeats present at the loci of the insulin gene, the pseudogene of zeta-globin, the myoglobin gene, and part of the core sequence of pYNZ22, a plasmid identified by the zeta-globin oligonucleotide. We also synthesized 16-base and 20-base oligonucleotides that correspond to a portion of the X-gene region of hepatitis B virus (HBV), on the basis of an apparent similarity of these sequences to the consensus sequence of the myoglobin family. The X-gene region of HBV is of special interest because it is thought to be the region within the viral genome used for integration of the

Table 1. DNA sequences and hybridization conditions for oligonucleotide probes. Hybridization was carried out in a solution of 5× SSC (1× SSC = 0.15*M* NaCl and 0.015*M* sodium citrate); 50 m*M* tris-HCl (*p*H 7.4); 1× Denhardt's solution (0.02% bovine serum albumin, 0.02% polyvinylpyrrolidone, 0.02% Ficoll); yeast tRNA (10 μg/ml); and 1×10^5 cpm/ml; ^{32}P 5′-end labeled probe (specific activity 2×10^6 cpm/pmol) for 16 hours. Washing was done three times in 5× SSC–0.1% SDS for 5 minutes.

Probe	Sequences	Temperature (°C) of:	
		Hybridization	Washing
Zeta-globin (18mer)	TGGGGCACAGG$^{T}_{C}$TGTGAG	42	48
Insulin (14mer)	ACAGGGGTGTGGGG	30	37
Myoglobin-1 (16mer)	GGAGGTGGGCAGGAAG	37	44
Myoglobin-2 (14mer)	GGAGGCTGGAGGAG	37	42
HBV-1 (16mer)	GGAGTTGGGGGAGGAG	37	44
HBV-2 (20mer)	GGACTGGGAGGAGTTGGGGG	50	60
HBV-3 (15mer)	GGTGAAGCA$^{G}_{C}$AGGTG	37	42
HBV-4 (15mer)	GAGAGGGGTGTAGAG	37	42
HBV-5 (15mer)	GGTGTAGAGAGGGGT	37	42
YNZ22 (15mer)	CTCTGGGTGTCGTGC	37	42

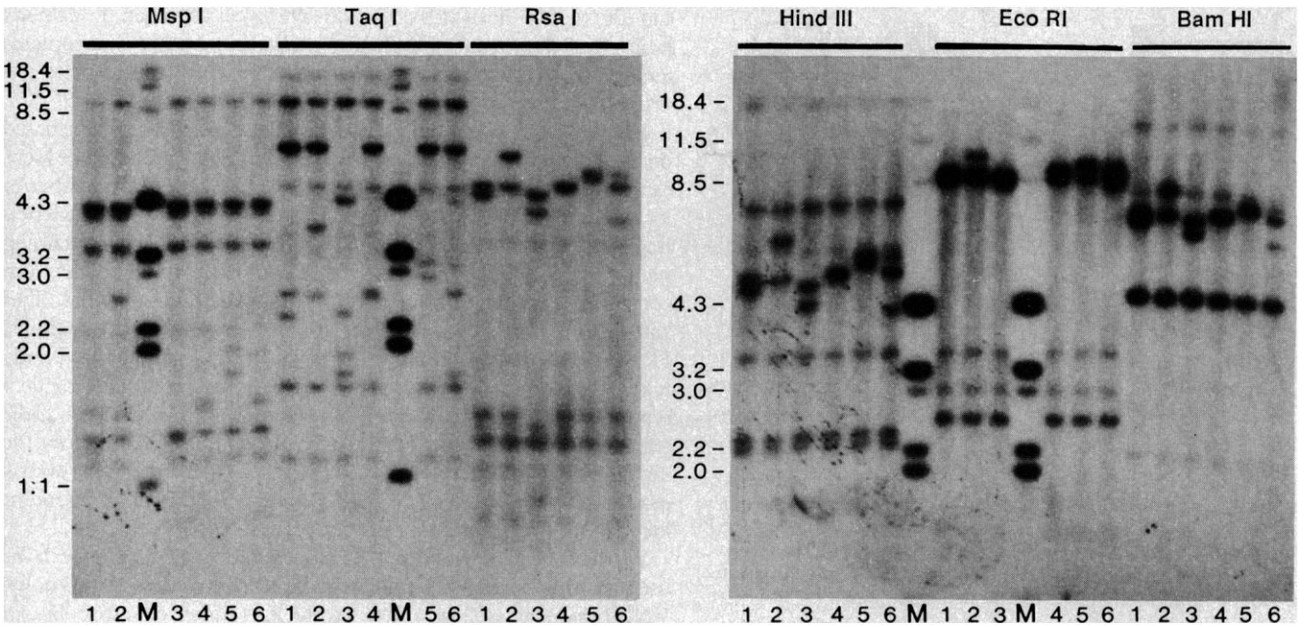

Fig. 2. Autoradiograms of Southern transfers from six unrelated individuals, with the whole cosmid DNA cTHH33 as a probe. The lymphocyte DNAs were digested with restriction enzymes Msp I, Taq I, Rsa I, Eco RI, Bam HI or Hind III. Transfer and hybridization were done as described (*22*), except that we used 500 μg of human placental DNA in the prehybridization solution instead of salmon sperm DNA. M, size markers (in kilobases, left columns) of mixed DNA from digested λ phage and pBR322.

virus into the host genome (*16–18*). The specific sequences of the synthesized oligonucleotides are shown in Table 1.

We initially screened a human genomic phage library (*19*) with the myoglobin-1 oligonucleotide. Low-stringency conditions were chosen for hybridization in order to detect relationships even when homology between the oligonucleotides and genomic sequences was relatively low. About 50 positive clones per genome equivalent (almost 250,000 clones) were obtained, but none revealed VNTR polymorphisms when used to probe restriction digests of DNA from unrelated individuals. Because phages having repeated DNA sequences do not grow well in *rec*A$^+$ bacteria (*20*), it appeared possible that the library constructed in *rec*A$^+$ bacteria had lost many of the phages that contained the repeat sequence.

Screening of a human cosmid library (*21*) growing in *rec*A$^-$ bacteria was more productive (Table 2). Although we used the same conditions as for phage screening, we obtained three times as many positive clones per genome equivalent from the cosmid library as from the phage library. Screening of the human genomic cosmid libraries with the six indicated oligonucleotides has yielded more than 1000 candidate clones per genome. As there are some similarities in sequence among the several probe oligonucleotides (see Table 1), it is important to show that most of the cosmid clones identified with one oligonucleotide probe are different from those identified with another oligonucleotide probe. Fewer than 4% of cosmid clones from a set of 100 hybridized with more than one of the oligonucleotide probes under our conditions. A total of 372 candidate clones have now been used as probes of restriction enzyme–digested DNA from several individuals to determine whether the clones reflect VNTR loci.

Table 2. Summary of screening for VNTR polymorphism.

Probe	Positive clones per genome*	Clones tested for polymorphism	VNTR polymorphism	Percent†	Site polymorphism‡
Zeta-globin	180	57	12	21	20 (6)
Insulin	220	20	4	20	8 (1)
Myoglobin-1	150	65	8	12	19 (8)
Myoglobin-2	18	18	8	44	8 (2)
HBV-1	200	75	13	17	38 (13)
HBV-2	40	15	2	13	10 (3)
HBV DR 1	50	26	6	23	6 (2)
HBV DR 2 and 3	150	58	15	26	40 (19)
YNZ22	38	38	9	24	25 (11)
Totals		372	77	21	174 (65)

*We screened one genomic equivalent (almost 75,000 colonies) to two genomic equivalents for each probe. The number of positive clones per genome means the number per 75,000 colonies. †The proportion of VNTR DNA markers among the tested cosmids. ‡The number of cosmids that showed site polymorphisms with two or more restriction enzymes is shown in parentheses.

The cosmids containing possible VNTRs were screened for whether they would reveal DNA polymorphism by hybridization of radiolabeled cosmid DNA with Southern transfers of human genomic DNA. Screening with the whole cosmid as probe can be achieved by prehybridization of the filters with human placental DNA (500 μg/ml) in order to eliminate background hybridization caused by repetitive human sequences (*22*). Restriction site polymorphisms of the conventional kind (single base pair variation) as well as VNTR loci were identified. VNTR loci, in addition to exhibiting more than two allele sizes, can be revealed by several restriction enzymes, while a single base change is likely to be revealed by only one. Figure 2 illustrates a pattern typical for a VNTR locus, as revealed by hybridization with a whole cosmid probe.

Once a cosmid was found to identify a VNTR locus, we removed the common highly repetitive DNA sequences by subcloning the fragment responsible for the polymorphism. Although it was not obvious at the outset that the DNA fragment containing the VNTR region from the cosmid would, even at high stringency, identify only a single locus when used as probe, in practice it usually did so. Apparently, even though the DNA sequences of a VNTR region are members of a large set of similar sequences located at many different loci, there is enough variation in sequence among the loci that hybridization at high stringency often permits the unique identification of a single locus.

Figure 3A shows the result of a Southern transfer with the probe pYNH24, which had been identified by the HBV-2 oligonucleotide. This locus shows 19 resolvable alleles among 16 unrelated individuals, and all individuals are heterozygotes. Probe pMLJ14 (Fig. 3B), which was identified by the myoglobin-1 oligonucleotide, also shows heterozygosity in all 16 individuals. Further tests indicated that each of the two loci was demonstrably heterozygous in 113 (94%) of 120 unrelated individuals. When we use VNTR DNA markers as single-locus probes, a higher molecular weight allele gives a stronger intensity of signal on the blot than does a smaller allele, reflecting a larger copy number of repeating sequences (see Fig. 3). The marker loci defined by these probes are expected to be highly informative in linkage studies.

Figure 3, C and D, demonstrates the Mendelian inheritance of the DNA loci detected by pYNH24 and pYNZ22, respectively; the latter probe had been identified by the zeta-globin oligonucleotide. In the three-generation pedigrees pictured, pYNH24 showed eight alleles and pYNZ22 revealed six. In each case, the alleles, and therefore the chromosome regions within which they are imbedded, can be followed unambiguously through the family. For example, the maternal grandmother is heterozygous at the pYNH24 locus (Fig. 3C), having alleles 1 and 5. Allele 1 is inherited by her daughter, who passes it on to four of her children: individuals III-2, -4, -5, and -7. If there were a dominant disease gene tightly linked to the chromosomal region defined in this family by allele 1, we would see the disease expressed in each of the above individuals and we could begin to document localization of the mutant gene to the region defined by probe pYNH24.

Table 2 provides a summary of the results we obtained from screening cosmids identified by the oligonucleotide probes. The frequency of VNTR loci detected within each group of cosmid probes ranged from 12% in the case of the myoglobin-1 oligonucleotide to as high as 44% with the cosmids identified by the myoglobin-2 oligonucleotide. Of the 372 cosmid DNAs tested thus far, 77 (21%) have revealed VNTR polymorphism, with nearly 90% of these showing three or more alleles. These are minimum estimates, both of frequency of clones revealing VNTR loci and of frequency of heterozygosity at these loci. Of the cosmids not identifying a VNTR locus, more than half do reveal restriction enzyme site polymorphisms with heterozygosities greater than 40%, and 65

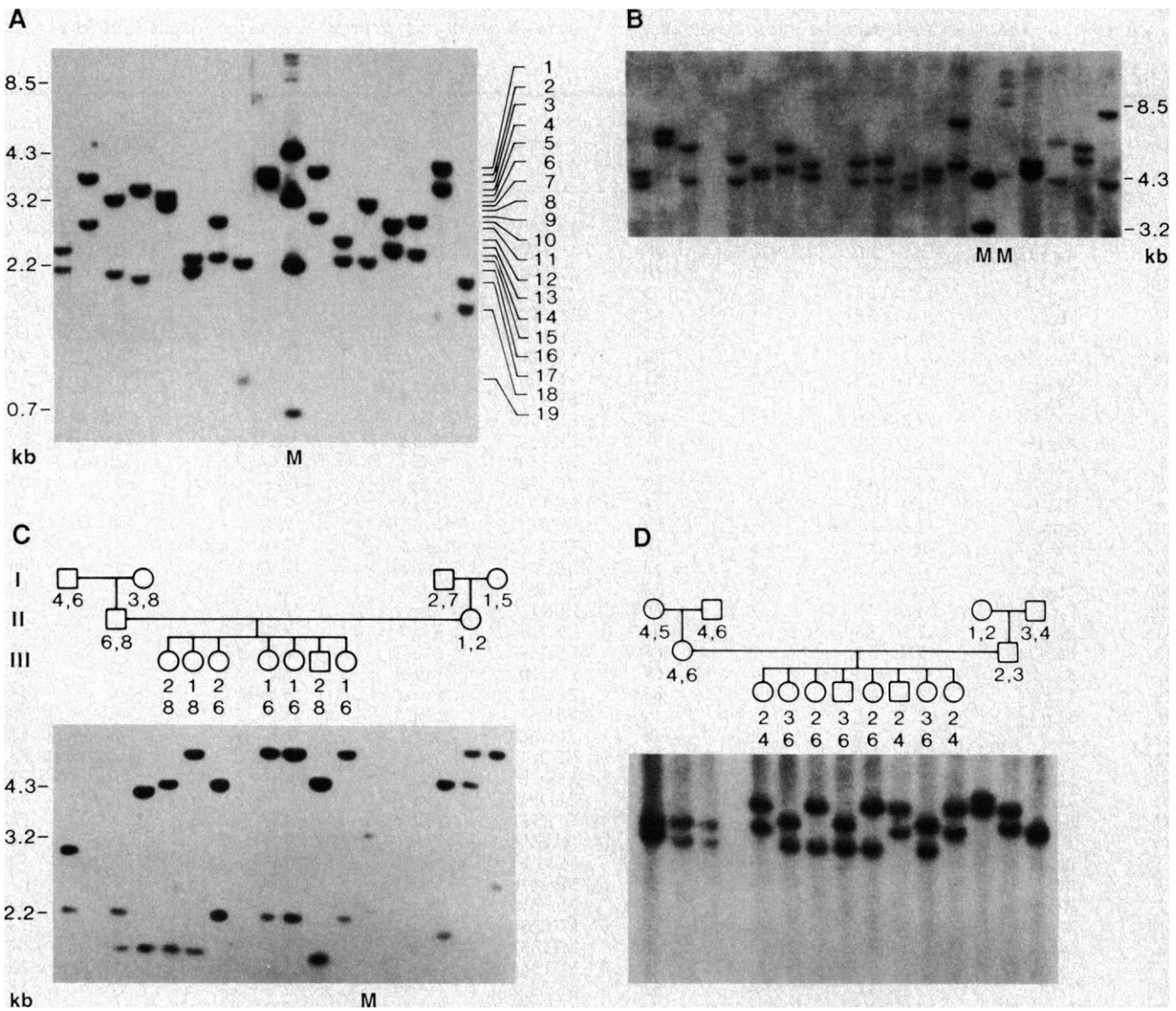

Fig. 3. Autoradiograms of Southern transfers from 16 unrelated individuals (**A** and **B**) and from three-generation families (**C** and **D**). The DNA was digested with Msp I (A and C), Rsa I (B), or Bam HI (D). Filters were hybridized with pYNH24 (A and C), pYNZ22 (D), or pMLJ14 (B). The genotypes of individuals in each three-generation family are shown directly below their symbols in the pedigree. M, marker lane(s); kb, kilobases.

cosmid clones show polymorphism with more than two enzymes.

Figure 4 and Table 3 provide an initial description of the allelic characteristics of the 77 new VNTR loci. A minimum estimate of the average heterozygosity shown by the 67 DNA markers in Table 3 that have three or more alleles is over 70%. The number of alleles detected thus far in the tested population varies from two at some loci to more than 20 at others.

Seventeen of the VNTR loci, indicated in Table 3 by the dagger (†), have now been characterized in 822 individuals within a set of 58 complete, three-generation families with large sibships (*14*, *23*). Linkage analysis indicates that the 17 loci are well distributed over at least ten chromosomes; the loci are not clustered in only one or a few regions of the genome.

Table 4 shows representative sequences of the tandem repeats from six of the VNTR loci, three ascertained with the zeta-globin oligonucleotide probe, two with the HBV probe, and one with the insulin probe. The repeat units range in length from 17 bp to 40 bp. Several repeats show striking homology; the zeta-globin repeat sequence matches the YNZ2 sequence at 11 of the possible 17 base pairs, and the YNZ22 and YNZ132 sequences match at 14 of 25 possible base pairs. Because each of these sequences was ascertained by hybridization with an oligonucleotide representing the zeta-globin sequence, thus introducing a bias, the statistical significance of the similarities is difficult to evaluate. Nonetheless, a strand bias in the distribution of G and C is apparent, with 71% of the G's on the same strand. Furthermore, the 62% G + C is also unusual, deviating markedly from the average composition of 40% G + C that is characteristic of mammalian DNA.

Discussion

These results fulfill the expectation that many VNTR loci exist within the human genome; they can be efficiently detected by screening cosmid libraries with oligonucleotides representing the sequences of known VNTRs. Oligonucleotide probes based on previously reported sequences from the myoglobin VNTR (*13*), and from the zeta-globin and insulin VNTRs, as well as sequences found within the X-gene region of HBV, have effectively identified cosmids that will detect new VNTR loci.

Although similarities exist among the individual VNTR sequences found in the sets, cross-screening indicates that overlap between clones is no more than a few percent at the hybridization stringencies used. Only one instance of multiple ascertainment of the same locus with cosmids derived from different screenings has been observed; this occurred at an exceptionally large locus, YNI10

Table 3. VNTR DNA markers.

Marker	Enzyme*	Allele		Hetero-zygosity (%)	Marker	Enzyme*	Allele		Hetero-zygosity (%)
		Size (kb)	Num-ber				Size (kb)	Num-ber	
YNZ2	Rsa I	1.0– 3.0	5	64†					
YNZ21	Msp I	1.0– 4.0	>10	89	THH51	Msp I	5.0– 7.0	3	67
YNZ22	Bam HI	1.1– 2.0	>10	86†	THH59	Pvu II	0.8– 1.8	6	71†
YNZ23	Pst I	2.0/ 2.5	2	40†	THH104	Rsa I	0.8– 1.8	6	67
YNZ32	Taq I	2.3– 2.8	5	75†	THH116	Rsa I	2.5– 2.8	3	67
YNZ86	Msp I	0.6– 0.8	3	52†	THH123	Msp I	0.8– 1.2	4	56
YNZ132	Taq I	1.8– 2.3	6	69†	THH129	Taq I	3.2– 5.0	3	64†
YNZ186	Bam HI	1.2– 2.0	6	83	YNH24	Msp I	1.0– 7.0	>20	94†
YNZ195	Taq I	1.0– 2.5	>10	83	YNH37	Taq I	2.0– 4.0	5	78†
JCZ3	Taq I	1.5– 3.0	7	78	EKZ101	Rsa I	2.0/ 2.2	2	44
JCZ16	Rsa I	1.3– 2.0	5	61	EKZ103	Msp I	2.0– 2.4	3	50
JCZ19	Bgl II	1.8– 3.1	6	83	EKZ107	Taq I	2.5– 4.4	4	67
YNM3	Rsa I	2.5– 2.8	4	34†	EKZ109	Taq I	2.0/ 2.5	2	50
YNM4	Taq I	2.3/3.0/5.0	3	60	EKZ127	Msp I	3.0– 4.5	4	61
MLJ14	Rsa I	4.0–15.0	>20	94†	EKZ130	Rsa I	1.0– 2.0	5	78
MLJ101	Msp I	2.2– 3.5	6	89	EFD4	Pvu II	2.0– 2.5	3	61
MLJ102	Bgl II	6.0– 8.0	6	78	EFD6	Rsa I	2.0– 3.4	3	67
MLJ103	Taq I	0.6– 0.8	4	67	EFD7	Pvu II	1.0– 1.8	3	56
MLJ205	Msp I	1.5– 4.5	>10	83	EFD9	Taq I	1.5– 6.0	>10	89
MLJ208	Pst I	4.5– 7.0	7	78	EFD11	Msp I	2.0– 2.7	4	56
CMM1	Bam HI	2.1/ 2.3	2	33	EFD13	Taq I	2.0– 3.0	3	67
CMM3	Bam HI	1.9– 3.3	7	83	EFD19	Msp I	3.0– 4.5	6	83
CMM5	Rsa I	2.5– 3.4	4	67	EFD20	Msp I	3.2– 3.7	3	67
CMM6	Taq I	2.5– 4.3	5	83	EFD52	Pst I	4.0–10.0	>10	94
CMM8	Msp I	2.3/ 2.8	2	39	EFD61	Msp I	1.0– 2.3	6	78
CMM12	Bam HI	3.5– 6.0	4	78	EFD63	Rsa I	2.0– 4.0	4	72
CMM19	Rsa I	1.2– 2.0	4	61	EFD64	Msp I	1.0– 5.0	>10	83
CMM22	Msp I	2.0/ 2.8	2	44	EFD75	Pvu II	1.9– 2.4	3	50
YNI10	Taq I	10.0–15.0	>10	85†	EFD91	Pvu II	2.5– 2.9	3	61
CMI37	Rsa I	2.3– 3.0	6	60†	EFD95	Msp I	4.0– 7.0	4	78
CMI40	Taq I	2.5– 4.5	4	67	MHZ10	Pst I	3.2– 4.0	4	72
CMI214	Bgl II	4.0– 5.0	3	61	MHZ13	Pst I	2.9– 4.5	5	78
THH5	Pvu II	1.1– 2.0	4	55†	MHZ15	Msp I	3.0– 3.6	3	67
THH7	Rsa I	3.0– 4.3	4	66	MHZ16	Msp I	1.2/ 1.3	2	38
THH18	Msp I	4.5/ 4.8	2	50	MHZ30	Taq I	2.7/ 3.9	2	33
THH27	Msp I	3.5– 4.2	5	78†	MHZ32	Msp I	3.2– 4.3	4	67
THH33	Rsa I	3.5– 5.0	>10	78	MHZ44	Msp I	2.4– 3.3	4	78
THH39	Pst I	2.1– 3.0	4	61†	MHZ47	Msp I	1.5– 3.2	6	83
THH50	Taq I	4.0/ 4.4	2	50	MHZ48	Pst I	3.0– 3.3	3	61

*Only the enzymes that gave the best resolution are shown. Probes YNZ and JCZ were isolated by the zeta-globin oligonucleotide; YNM and MLJ, by myoglobin-1; CMM, by myoglobin-2; THH, by HVB-1; YNH, by HBV-2; YNI and CMI, by insulin; EKZ, by HBV-3; EFD, by HBV-4 and -5; MHZ, by YNZ22 oligonucleotide. †The result in 120 unrelated individuals; entries without the † represent the result in 60 unrelated individuals.

(*24*). Furthermore, multiple ascertainment of the same VNTR locus by cosmids detected with the same oligonucleotide is also rare, again with the exception of locus YNI10, indicating that there are many different VNTRs within each set.

Relative efficiency of prescreening. The results reported here indicate that approximately 21% of cosmids ascertained by prescreening with appropriate oligomers will detect VNTR loci under our initial screening conditions. We have also tested 70 unselected cosmid clones for their ability to detect VNTRs; only one unselected cosmid showed a VNTR polymorphism.

These data suggest that the detection rate of VNTR loci with unselected cosmids, under our primary screening conditions, is approximately 2 to 4%. Prescreening the cosmid libraries with oligonucleotides increases by five- to tenfold the frequency of cosmids that will detect VNTR loci.

The absolute values obtained from this study, both for yield of VNTR loci and for heterozygosities, are minimum estimates. In a subsequent test of several VNTR loci, individuals that had scored as homozygotes under our initial test conditions could be identified as heterozygotes with high resolution gel systems.

The only other reported large-scale screening for human DNA sequence polymorphisms has been that of Braman *et al.* (*25*), in which some 1500 arbitrary loci were examined with large insert, unselected phage clones. Although the results of that study have not yet been fully reported, 14 of 1500 loci defined by the genomic phage clones revealed VNTR loci with frequencies of heterozygosity of 0.75 or greater, supporting the conclusion that many VNTR loci are present in the human genome.

DNA sequences of the VNTRs. It is tempting to speculate that the tandem repeats increase or decrease in number through a high frequency of unequal crossing over. This proposal has the added merit that it could account for the maintenance of sequence identity throughout the set of repeats at a specific locus, as well as account for the generation of multiple alleles. Jeffreys *et al.* (*13*) have further suggested that the VNTR sequences might encode hotspots for recombinational activity and thus account for their genesis and high mutability through unequal exchange. Any mechanism of diminution and amplification of the tandem repeats would have the same effect, however. An almost invariant core sequence, GGGCAGGAXG, seems to be common to nearly all of the VNTR sequences ascertained in the Jeffreys study. Furthermore, that group of authors also observed that the DNA sequences of several VNTRs ascertained by hybridization with the myoglobin sequence show an apparent relationship to the *chi* sequence of phage lambda, GCTGGTGG. This octamer sequence has been implicated as a hotspot for *rec*A-mediated recombination in *Escherichia coli*.

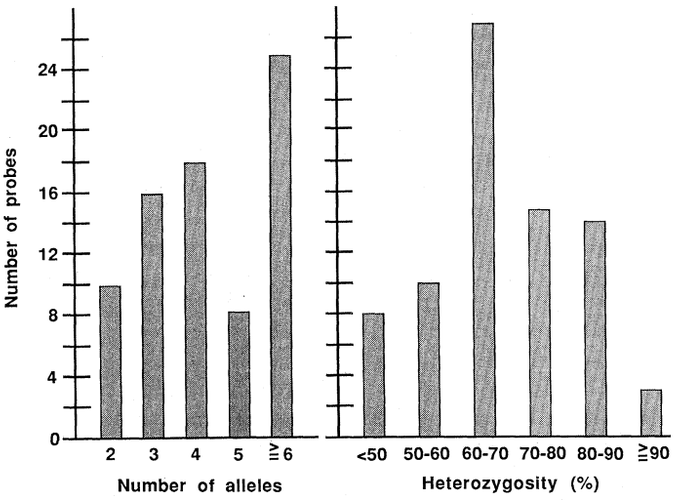

Fig. 4. Characteristics of 77 new VNTR marker loci.

In our study, however, a similar but somewhat different pair of core sequences was found: a somewhat variable common core sequence GGG_GTGGGG. Table 4 contains the almost invariant sequence G____TGGG. These sequences bear some, although not striking, similarity to the lambda *chi* sequence.

It is perhaps not surprising that the DNA sequences of these VNTRs are similar, as the majority thus far examined have been ascertained through hybridization with specific probe sequences. However, we can also see the *G____GTGGG* motif among the core sequences of three more VNTR loci (p3.4BHI, α-globin, and D14S1) ascertained by chance (Table 4) (*26–28*). Although most VNTRs reported so far belong to the G-rich family, two of them [apolipoprotein B (*29*) and collagen type II (*30*)] have a completely different, AT-rich, motif. The motif *G____TGGG* is almost invariant among the G-rich sequences, providing additional support for its importance. Although the statistical significance of finding such a sequence in any particular pairwise comparison test is not high, the likelihood that a specific sequence motif would be present by chance in eight independently ascertained VNTR sequences seems remote.

Finally, the fact that sequences from the X-gene region of HBV serve to identify VNTR loci is of some interest. This region of the HBV genome is thought to be the viral site of integration into the

Table 4. Comparison of the core DNA sequences of VNTR loci.

Locus	Sequence
Zeta-globin (*12*)	GGTT*GTGAGTGGGG*CACA (9G; 2C/18 nucleotides)
YNZ2	GA*GGCTCATGGGG*CACA (7G; 4C/17)
YNZ22	TGGA*GTCTCTGGGT*GTCGTGCGTCAGAGT (12G; 5C/29)
YNZ132	TGCA*GGCTGTGGGT*GTGATGGGTGA (13G; 2C/25)
Insulin (*10*)	ACA*GGGGTGTGGGG* (9G; 1C/14)
YNI10	CTG*GGGGTGTGGGT*GCTGCTCCAGGCTGTCAGATGCTCAC (16G; 10C/40)
THH59	CTGGG*GAGCCTGGGG*ACTTTCCACACC (9G; 9C/27)
YNH24	CAG*GAGCAGTGGG*AAGTACAGTGGGGTTGTT (14G; 3C/31)
Myoglobin (*13*)	CTAAAGCT*GGAGGTGGG*CAGGAACGACCGAGGT (14G; 6C/33)
Common sequence	GGGNNGTGGGG
chi	GCTGGTGG
Harvey-*ras* (*11*)	GGG*GGAGTGTGGCGT*CCCCTGGAGAGAA
p3.4BHI (*26*)	AACA*GTGCGTGGGC*CACGTGAGCGGAGCAGGCTC
α-globin (*27*)	AACAGC*GACACGGGGGG*
D14S1	$(^{T}_{C}GG)n=GG^{T}_{C}GG^{T}_{C}GG^{T}_{C}GG^{T}_{C}G$ .....
Apolipoprotein B (*29*)	TTTTATAATTAATATTTTATAAT .....
Collagen type II (*30*)	CAATATAGATAATATATACCTATATATTATTATA

human genome and thus could be considered recombinogenic in a manner similar to the long terminal repeat regions of retroviruses (*31*). Although these similarities could be due to functional constraints, viral or extrachromosomal origins for the VNTRs should be considered and explored.

Usefulness of VNTR loci for mapping disease loci. We expect the VNTR loci to be valuable as anchor points on the human genetic linkage map, and to be informative within the small sets of families that are available for mapping genetic disease loci, because the high frequency of heterozygosity both found and estimated in our study indicates that genotypic information will be obtained from most meioses in any sample set. Such genetic markers, made available to the research community, should contribute substantially to the construction of the human genetic linkage map and to the localization of genes that cause human genetic disease.

Note added in proof: Cosmids ascertained with the zeta-globin-, insulin-, and myoglobin-related oligonucleotides were rescreened with the oligonucleotides prior to testing for polymorphism. However, 212 cosmids ascertained with the HBV- and YNZ22-related oligonucleotides were tested for polymorphism without rescreening. Recently, we rescreened at random 106 of these cosmids and found that only 57 of the 106 in fact hybridized with the respective oligonucleotides. Eighteen of the 57 (32%) showed VNTR polymorphism, while only 2 (4%) of the 49 cosmids that failed to show oligonucleotide homology demonstrated VNTR polymorphism.

REFERENCES AND NOTES

1. J. F. Gusella *et al.*, *Nature (London)* **306**, 234 (1983).
2. J. M. Murray *et al.*, *ibid.* **300**, 69 (1982).
3. S. Reeders *et al.*, *ibid.* **317**, 542 (1985).
4. L.-C. Tsui *et al.*, *Science* **230**, 1054 (1985).
5. R. White *et al.*, *Nature (London)* **318**, 382 (1985).
6. R. Williamson *et al.*, *ibid.*, p. 384.
7. A. Jeffreys *et al.*, *Cell* **18**, 1 (1979).
8. S. E. Antonarakis, H. H. Kazazian, Jr., S. H. Orkin, *Hum. Genet.* **69**, 1 (1985).
9. A. Wyman and R. White, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6754 (1980).
10. G. I. Bell, M. J. Selby, W. J. Rutter, *Nature (London)* **295**, 31 (1982).
11. D. J. Capon, E. Y. Chen, A. D. Levinson, P. H. Seeburg, D. V. Goeddel, *ibid.* **302**, 33 (1983).
12. N. J. Proudfoot *et al.*, *Cell* **31**, 553 (1982).
13. A. Jeffreys, V. Wilson, S. Thein, *Nature (London)* **314**, 67 (1985).
14. R. White *et al.*, *ibid.* **313**, 101 (1985).
15. Z. Wong *et al.*, *Nucleic Acids Res.* **14**, 4605 (1986).
16. R. Koshy *et al.*, *Cell* **34**, 215 (1983).
17. H. Mizusawa *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 208 (1985).
18. K. Yaginuma *et al.*, *ibid.*, p. 4458.
19. R. Lawn *et al.*, *Cell* **15**, 1157 (1978).
20. A. Wyman L. Wolfe, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2880 (1985).
21. Y.-F. Lau and Y. W. Kan, *ibid.* **80**, 5225 (1983).
22. M. Litt and R. White, *ibid.* **82**, 6206 (1985).
23. Family panels are in the archives of the Centre d'Etude du Polymorphisme Humain (CEPH), Paris.
24. Y. Nakamura *et al.*, in preparation.
25. J. Braman, D. Barker, J. Schumm, R. Knowlton, H. Donis-Keller, *Cytogenet. Cell Genet.* **40**, 589 (1985).
26. A. Silva, personal communication.
27. A. P. Jarman *et al.*, *EMBO J.* **5**, 1857 (1986).
28. A. Wyman, J. Mulholland, D. Botstein, *Am. J. Hum. Genet.* **39** (Suppl.), A-226 (1986).
29. T. J. Knott *et al.*, *Nucleic Acids Res.* **14**, 9215 (1986).
30. N. G. Stoker *et al.*, *ibid.* **13**, 4613 (1985).
31. T. Tamura and T. Takano, *ibid.* **10**, 5333 (1982); C. Shoemaker *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3932 (1980).

32. We are grateful to J.-M. Lalouel for helpful discussion and for linkage analyses, to Y. W. Kan and Y. C. Lau for providing cosmid libraries, and to T. Maniatis for the phage libraries. We also wish to thank M. Belman and B. Loeffler for construction of oligonucleotides and R. Foltz for editorial assistance in the preparation of this manuscript. R.W. is an investigator of the Howard Hughes Medical Institute. As appropriate subclones free of repetitive sequences are developed, they will be added to the American Type Culture Collection (ATCC) for unrestricted access by investigators.

Research Articles

Oncogenesis of the Lens in Transgenic Mice

KATHLEEN A. MAHON, ANA B. CHEPELINSKY, JASPAL S. KHILLAN, PAUL A. OVERBEEK, JORAM PIATIGORSKY, HEINER WESTPHAL

Neoplastic tumors of the ocular lens of vertebrates do not naturally occur. Transgenic mice carrying a hybrid gene comprising the murine αA-crystallin promoter (−366 to +46) fused to the coding sequence of the SV40 T antigens developed lens tumors, which obliterated the eye cavity and even invaded neighboring tissue, thus establishing that the lens is not refractive to oncogenesis. Large-T antigen was detected early in lens development; it elicited morphological changes and specifically interfered with differentiation of lens fiber cells. Both α- and β-crystallins persisted in many of the lens tumor cells, while γ-crystallin was selectively reduced. Accessibility, characteristic morphology, and defined protein markers make this transparent epithelial eye tissue a potentially useful system for testing tumorigenicity of oncogenes and for studying malignant transformation from its inception until death of the animal.

THE TRANSGENIC MOUSE PROVIDES AN EXPERIMENTAL SYSTEM to study the molecular genetics of malignant growth in the intact organism, from the earliest to the terminal stages of the disease. Both viral and cellular oncogenes have been introduced into the germline of mice, resulting in malignancies of various types (*1–7*). Although the introduced oncogenes were present throughout the organism, tumors were generally found only in cell types specified by the regulatory regions controlling oncogene expression. This indicates that oncogenesis can be targeted by the judicious use of DNA sequences regulating the tissue-specific expression of genes.

A naturally occurring malignant tumor has never been reported in the ocular lens. However, abnormal proliferation of lens epithelium

K. A. Mahon, J. S. Khillan, and H. Westphal are in the Laboratory of Molecular Genetics, National Institute of Child Health and Development, National Institutes of Health, Bethesda, MD 20892. A. B. Chepelinsky and J. Piatigorsky are in the Laboratory of Molecular and Developmental Biology, National Eye Institute, National Institutes of Health, Bethesda, MD 20892. P. A. Overbeek is in the Howard Hughes Medical Institute, Houston, TX 77030.