Analysis of two different tandem repetitive elements within the human apolipoprotein B gene

Erwin H. Ludwig, Kurt Haubold, and Brian J. McCarthy[1]

Gladstone Foundation Laboratories for Cardiovascular Disease, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA 94140-0608

Summary A large number of copies of the sequence $(dTG\text{-}dAC)_n$, where n is between 10 and 60, exist in the human genome, and many are useful as polymorphic markers. One of these sequences occurs about 3 kilobases 5' of the human apolipoprotein (apo) B gene as seven distinguishable alleles containing from $(TG)_{12}$ to $(TG)_{18}$. This repeat is also present in the DNA of other primates. A second alternating purine-pyrimidine sequence with nine dinucleotide repeats and located in intron 4 is not polymorphic. Together with the apoB hypervariable repeat immediately 3' of the gene, the $(TG)_n$ sequence will provide a useful haplotype marker capable of distinguishing a large number of human apoB alleles, some of which may be associated with disease states.—**Ludwig, E. H., K. Haubold, and B. J. McCarthy.** Analysis of two different tandem repetitive elements within the human apolipoprotein B gene. *J. Lipid Res.* 1991. **32:** 374–379.

Supplementary key words minisatellites · polymerase chain reaction · DNA marker · hypervariable region · variable number of tandem repeats · purine-pyrimidine repeats · Z-DNA

Abbreviations: RFLP, restriction fragment length polymorphism; kb, kilobases; apo, apolipoprotein; HVR, hypervariable region; PCR, polymerase chain reaction; bp, base pair.

[1]To whom correspondence should be addressed at: Gladstone Foundation Laboratories, P.O. Box 40608, San Francisco, CA 94140-0608.

Progress in human genetic analysis has been greatly facilitated by the use of DNA polymorphisms. These have proved especially useful for mapping of linkage groups and localization of loci for genetic disease. However, such restriction fragment length polymorphisms (RFLPs) are usually dimorphic and display a heterozygosity index of less than 50%. A second class of markers consists of hypervariable regions of the human genome displaying a higher degree of polymorphism (1). These regions comprise tandem arrangements of short, repeating base sequences, and their hypervariability derives from allelic differences in the number of repeats. The resulting variation can readily be detected as a length polymorphism by cleaving the DNA with restriction enzymes outside the repeat region or by means of the polymerase chain reaction.

An extreme subclass of these tandemly repeated DNA sequences comprises very short simple repeats such as $d(TG)_n \cdot (AC)_n$. These simple sequence blocks of $(TG)_n$, where n is usually in the range of 15–30, occur roughly every 30–60 kilobases (kb) within the human genome (2, 3). Their occurrence within the β-globin gene cluster, and near the cardiac muscle α actin gene, the apolipoprotein (apo) A-II, the apoC-II, and the somatostatin genes, among others, makes them potentially useful markers, capable of resolving several alleles of these genes. In all the above cases, the heterozygosity index proved to be high (4, 5) and the alleles of different sizes were inherited according to Mendelian rules.

Mutations in the human apoB gene are of considerable interest because of the crucial role that this protein plays in the assembly, secretion, and catabolism of low density lipoproteins and in cholesterol homeostasis. For this reason, considerable effort has been made to distinguish apoB alleles through haplotype analysis using RFLP markers. In addition, a hypervariable region (HVR) immediately 3' of the apoB gene has also been used to resolve as many as 14 alleles (6–9). In this report, we describe a second hypervariable locus consisting of a $(TG)_n$ repeat located 3 kb 5' of the apoB gene translational start site. Several polymorphic human alleles were sequenced, as were five alleles of nonhuman primates. Together, these two 3'- and 5'-flanking HVRs provide an important addition to the array of haplotype markers available for analysis of the human apoB gene.

MATERIALS AND METHODS

Subjects

Total genomic DNA was isolated from white blood cells of 233 unrelated Austrian male Caucasians by standard methods.

DNA amplification

Portions of the subjects' DNA were amplified by the polymerase chain reaction (PCR). The conditions were as follows: 1-min denaturation at 94°C, 1-min annealing at 55°C, and 1-min extension at 72°C for 25 cycles. All amplifications contained 200–500 ng of genomic DNA, 200 μM of each dNTP, and the buffer and enzyme recommended by Saiki et al. (10). Three different reaction protocols were used. (a) For the 5' $(TG)_n$ amplification, 200 ng of PCR1, 50 ng of PCR2, and 1 ng of ^{32}P-labeled PCR2 (**Fig. 1**) were used. (b) For the intron 4 repeat $(CA)_n \cdot (GT)_n$, 200 ng of PCR3, 50 ng of PCR4, and 1 ng of ^{32}P-labeled PCR4 (Fig. 1) were used. (c) For sequencing of the 5' $(TG)_n$, 200 ng of PCR2 (Fig. 1) and 200 ng of PCR6 (GCTCGTCCCAGAGAAAGATTCAGT) 288 bp 5' of PCR1 were used.

Analysis of 5' $(TG)_n$ amplification product

One-tenth of the amplified product obtained through protocol (a) was dried, loading dye was added, and the sample was heated to 70°C, loaded on a 6% polyacrylamide gel in 1 × TBE (89 mM Tris base, 89 mM boric acid, 2mM EDTA (pH 8.0)) containing 7 M urea, and electrophoresed for 2 h at 45–50°C. A sequencing reaction was usually loaded on the gel as a marker. The gel was then dried and autoradiographed.

Sequence analysis of the 5' $(TG)_n$ amplification product

The product of amplification protocol (c) was extracted with phenol–chloroform, precipitated with ethanol, and resuspended in 20 μl of distilled H_2O. The sample was then digested with *Pst*I and *Sph*I and cloned into M13mp18 that had also been cut with *Pst*I and *Sph*I. Sequencing was performed on single-strand templates using United States Biochemical Corporation's (USB) protocol for DNA sequencing with Sequenase™.

The methods for analyzing intron 4 $(CA)_n \cdot (TG)_n$ amplification products were similar to those just described, except that the oligomers described in protocol (b) were used in the amplification.

Amplification, analysis, and sequencing of the 5' $(TG)_n$ in primates other than humans

The 5' $(TG)_n$ sequences were also obtained for other primates: baboon, chimpanzee, gorilla, orangutan, marmoset, and macaque. Amplifications were conducted with the oligomers described in protocol (a), except that 5 cycles were run with an annealing temperature of 37°C before the 25 cycles at 55°C. These amplifications also included as a primer another oligomer, PCR5 (TATCATCTTCCTCCTGAACCATTC), 37 bp 5' of PCR1, together with PCR2 under the same conditions as above. For sequencing, amplified products were extracted with phenol–chloroform, precipitated with ethanol, resuspended in 1 ml of distilled H_2O, washed with 3 ml of distilled H_2O in a Centricon 30 microconcentrator (Amicon), and resuspended in 20 μl of distilled H_2O. Be-

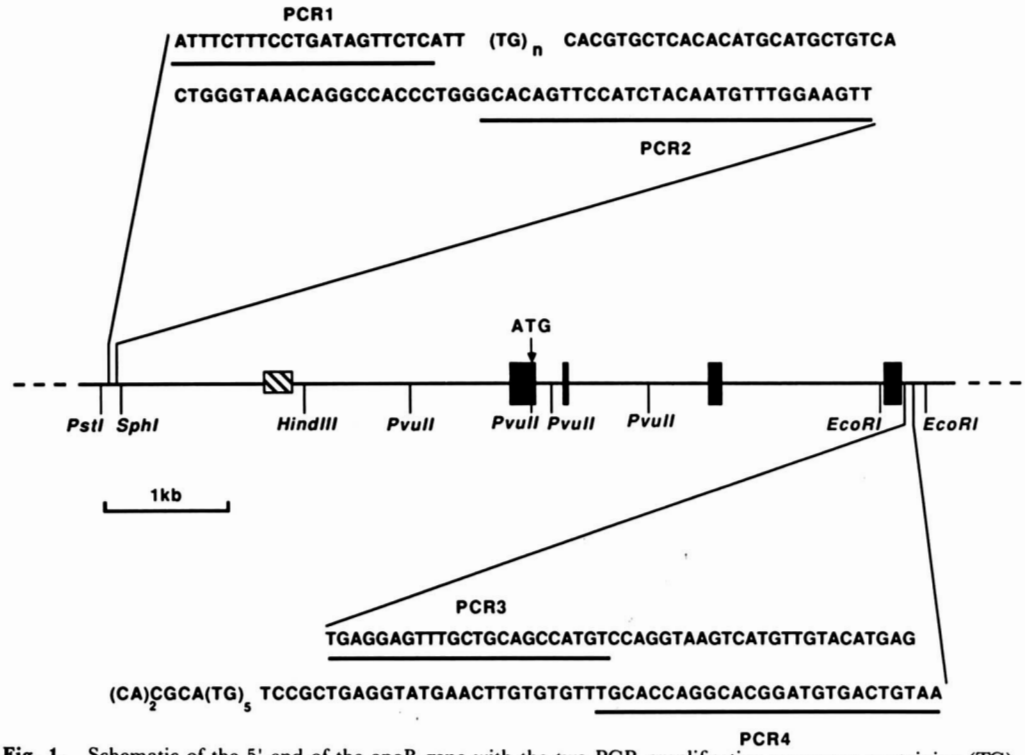

Fig. 1. Schematic of the 5' end of the apoB gene with the two PCR amplification sequences containing $(TG)_n$ repetitive elements. The center line represents the apoB gene up to intron 4 and 3500 bases flanking its 5' end. Exons are represented as black boxes; the ATG in exon 1 indicates the translational start site. The hatched box represents an Alu repeat; several restriction sites are also shown. The 5' $(TG)_n$ amplification sequence is shown above the gene, and the intron 4 $(TG)_n$ sequence below. Oligomers (PCR1–PCR4) used as amplification primers are underlined.

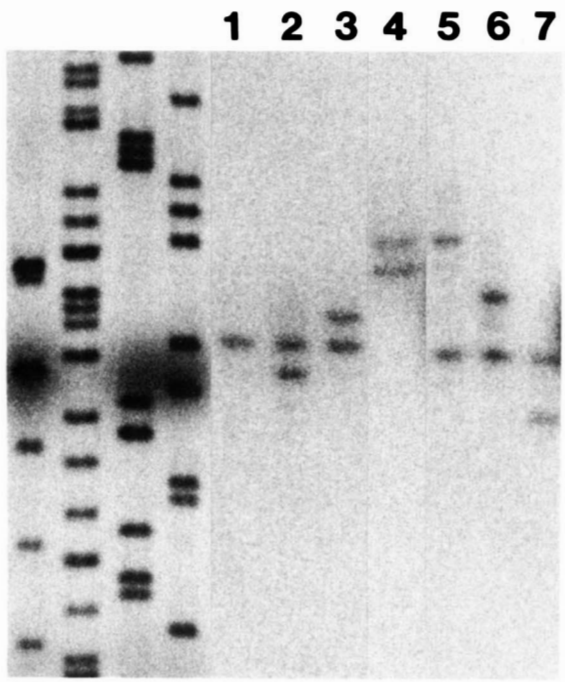

Fig. 2. The size distribution of the apoB 5' $(TG)_n$ HVR. Lanes 1 through 7 are amplification products of seven of the unrelated subjects and represent seven variants of the $(TG)_n$. A DNA sequencing reaction run in parallel acts as a marker. Lane 1 illustrates a $(TG)_{14}/(TG)_{14}$ polymorphism; lane 2, a $(TG)_{13}/(TG)_{14}$; lane 3, a $(TG)_{14}/(TG)_{15}$; lane 4, a $(TG)_{17}/(TG)_{18}$; lane 5, a $(TG)_{14}/(TG)_{18}$; lane 6, a $(TG)_{14}/(TG)_{16}$; and lane 7, a $(TG)_{12}/(TG)_{14}$.

tween a quarter and a half of the purified sample and sequencing primer was denatured at 95°C for 5 min and incubated at 4°C for 10 to 30 min. Sequencing was performed using the USB Sequence™ protocol.

The heterozygous gorilla DNA amplification product was subcloned using the oligomers described in protocol (c) and the same restriction enzymes and sequencing protocol described in the section on sequence analysis of the $(TG)_n$ amplification (above).

RESULTS

$(TG)_n$ repeat situated 5' of the human apolipoprotein B gene

Analysis of the 5'-flanking region of the human apoB gene revealed a $(TG)_{14}$ sequence (Fig. 1). By analogy with other such $(TG)_n$ repeats associated with several human genes (4, 5), the possibility arose that this sequence element might be polymorphic. This proposition was tested by amplifying the region surrounding the $(TG)_n$ element using one ^{32}P-labeled primer and resolving the fragments on a denaturing acrylamide gel (**Fig. 2**). From a total of 233 DNA samples, seven polymorphic forms were resolved; the size of the fragments revealed that the number of TG dinucleotides varied from 12 to 18. Each of the

size variants was subjected to DNA sequence analysis. Only one example each of $(TG)_{13}$ and $(TG)_{17}$ was available, but in all other cases several alleles were sequenced. In all cases the only observed difference in sequence was in the number of TG repeats (data not shown).

The distribution of repeat lengths within the Austrian population in this study is shown in **Table 1.** The $(TG)_{14}$ allele was the most frequent, followed by $(TG)_{15}$. Alleles of other sizes were relatively rare in this population, and the average heterozygosity index was 30.5%.

Presence of the 5' $(TG)_n$ hypervariable region in nonhuman primate DNA

To obtain some understanding of the origin and rate of divergence of the 5' HVR, attempts were made to amplify this same region of DNA of various primates. When PCR1 and PCR2 were used as primers, the DNA of chimpanzee, gorilla, orangutan, marmoset, and macaque produced amplified bands in the size range corresponding to that of the human DNA band (**Fig. 3**). Only the gorilla appeared to be heterozygous for this portion of the apoB gene, while baboon DNA failed to yield any amplified band. However, when PCR5 was used together with PCR2, all DNAs produced amplified bands except that of the marmoset (data not shown).

To verify that the difference in the size of the amplified bands among species was a reflection of different numbers of TG repeats, the bands amplified using primers PCR5 and PCR2 were sequenced. (For the gorilla, the two bands obtained from the DNA were subcloned and sequenced.) **Fig. 4** presents a comparison of the various DNA sequences. It is apparent that the differences in the number of TG repeats in the sequence are consistent with the apparent molecular weight of the amplified bands in Fig. 3. The numbers of repeats observed were 9 for chimpanzee and orangutan and 11 and 12 for the two gorilla alleles. The baboon sequence includes several changes within the alternating TG sequence characteristic of human and other primate species. The failure of baboon DNA to yield an amplified product when PCR1 was used as one of the primers is probably attributable to deletions in the sequence complementary to that primer (Fig. 4). As would be expected from evolutionary considerations, the baboon sequence is the most divergent among the five species compared in this manner.

TABLE 1. Frequency distribution of $(TG)_n$ hypervariable region alleles of the human apolipoprotein B gene

$(TG)_n$	No. of Alleles	Percent
12	3	0.6
13	1	0.2
14	364	78.1
15	88	18.9
16	7	1.5
17	1	0.2
18	2	0.4
	466	

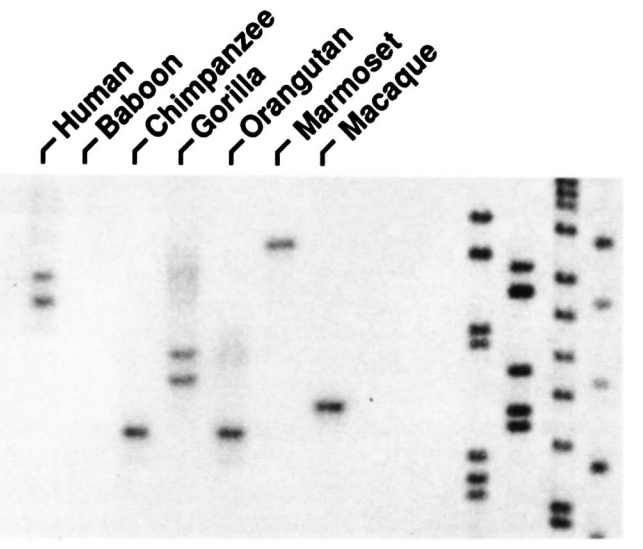

Fig. 3. Amplification products of the DNA of various primates using PCR1 and PCR2 as primers. A sequencing reaction was run in parallel as a size marker. The human control sample illustrates a $(TG)_{14}/(TG)_{15}$ polymorphism; no product was obtained for baboon. When compared with the human control repeats and the sequencing reaction, it appears that the chimpanzee and orangutan are homozygous for $(TG)_9$, the gorilla is heterozygous for $(TG)_{11}/(TG)_{12}$, the marmoset is homozygous for $(TG)_{16}$, and the macaque is homozygous for $(TG)_{10}$.

Potential dinucleotide repeat length polymorphism in intron 4 of the human apolipoprotein B gene

A second dinucleotide repeat detected in intron 4 of the human apoB gene was composed of the sequence $d(CA)_2CGCA(TG)_5$ (Fig. 1). Again, length polymorphism was investigated by amplifying the surrounding region using a ^{32}P-labeled oligonucleotide primer. However, in this case no size heterogeneity was observed among 102 DNA samples (data not shown).

DISCUSSION

A minisatellite length polymorphism immediately 3' of the apoB gene was first reported several years ago (6, 11). Because as many as 14 alleles containing different numbers of 15-bp tandem repeats can be resolved (8, 9), this polymorphism has proven to be useful for population studies (12, 13).

The present study concerned two other potential polymorphisms near and in the apoB gene that are based on purine-pyrimidine dinucleotide repeats. One of these, the 5' $(TG)_n$ repeat, proved to be polymorphic: seven alleles

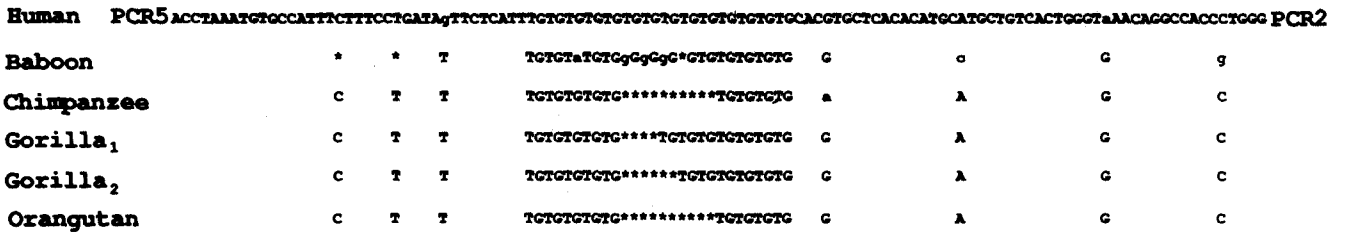

Fig. 4. Comparison of the base sequence surrounding the $(TG)_n$ HVR among various primates. The human sequence between primers PCR5 and PCR2 is displayed in upper-case letters. For flanking regions in nonhuman primate DNAs, only bases differing from those in human DNA are given (upper case). Any base that differs from that in the majority of the species surveyed is displayed in lower case. Asterisks designate deleted bases. Within the hypervariable TG repeat region itself, the complete sequence is represented as upper- and lower-case letters or asterisks for deletions.

were distinguishable within a group of 233 subjects (Table 1). Length polymorphism has been reported for several other $(TG)_n$ repeats in the human genome (4). On the other hand, we failed to detect any length heterogeneity in another repeating purine-pyrimidine sequence, located in intron 4. Although the reason for this lack of length heterogeneity is not obvious, it may be related to the limited number of alternating purines and pyrimidines in this sequence, i.e., $d(CA)_2CGCA(TG)_5$, with nine dinucleotide repeats. All of the purine-pyrimidine repeats demonstrated by Weber and May (4) to exhibit length polymorphism contain 14 or more repeats. Indeed, the apoB 5' $(TG)_n$ HVR is near the lower limit of this range. The polymorphism of these purine-pyrimidine repeats may be related to the fact that they can adopt a left-handed Z-DNA configuration (2, 14). Such repeats are highly recombinogenic in bacteria (15) and favor deletions. Furthermore, the frequency with which deletions occur increases dramatically in the range where the number of dinucleotide repeats increases from 8 to 13 (16). Although the details of the mechanism by which new length alleles of eukaryotic dinucleotide repeats are created by recombination or gene conversion are unclear, it does appear that topoisomerase II may be involved. The consensus sequence for vertebrate topoisomerase II recognition (17) approximates a purine-pyrimidine repeat comprising a total of 18 bp. In fact, sequences such as $(GT)_n$, $(AC)_n$, and $(AT)_n$, but not $(GC)_n$, are effective substrates for topoisomerase II (17).

To obtain data concerning the origin of the $(TG)_n$ polymorphic repeat, the DNA of several primates was examined for the presence of this element. The origin of this repeat clearly preceded the divergence of the human genome from that of other primates: the same repeat surrounded by essentially the same flanking sequence occurs in chimpanzee, gorilla, and orangutan DNA. Even in the one Old World monkey examined, the baboon, the sequence is highly conserved. Whether this conservation indicates that the TG element serves some essential function is unknown. Similar results have been reported for an AT-rich repeating sequence in the 3'-flanking region of the interleukin-6 gene (18). Again, the same basic repeated element is recognizable in both chimpanzee and baboon DNA. In contrast, although members of the most prevalent repeated sequences, such as the Alu family, are usually found at orthologous sites in several genes examined, exceptions have been reported. For example, an Alu and a Xba repetitive element are both present in introns of the human α-fetoprotein gene but absent from the corresponding sites in the gorilla α-fetoprotein gene (19).

The TG repeat will be valuable for high-resolution haplotype analysis of the human apoB gene. In fact, this TG repeat in conjunction with the HVR at the 3' site and with eight diallelic RFLP markers within the gene has already been used in haplotype analysis of a mutation in the human apoB gene that is associated with defective low density lipoproteins and hypercholesterolemia (13). Another possibly useful sequence is the recently identified polymorphic tetranucleotide repeat in intron 20 of the apoB gene (20). Finally, we note that the two flanking hypervariable markers occur very close to the boundaries of the apoB gene as defined by the DNaseI-sensitive domain and the location of chromosomal anchorage sites (21).

We would like to thank the Yerkes Regional Primate Research Center, Emory University, Atlanta, Georgia 30322, and Mary S. Anthony from the Bowman Gray School of Medicine for supplying primate tissues and plasma samples; also, Drs. Jerry L. Slighton, S. Embury, and S. Young for gifts of DNA samples. We also thank Tom Rolain and Charles Benedict for graphics, Will Doolittle for manuscript preparation, and Al Averbach and Sally Gullat Seehafer for editorial assistance. This research was supported by National Institutes of Health Grant HL38781.

Manuscript received 24 August 1990.

REFERENCES

1. Jeffreys, A. J., V. Wilson, and S. L. Thein. 1985. Hypervariable "minisatellite" regions in human DNA. *Nature.* **314:** 67–73.
2. Hamada, H., and T. Kakunaga. 1982. Potential Z-DNA forming sequences are highly dispersed in the human genome. *Nature.* **298:** 396–398.

3. Tautz, D., and M. Renz. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.* **12:** 4127–4138.
4. Weber, J. L., and P. E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388–396.
5. Litt, M., and J. A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44:** 397–401.
6. Huang, L. S., and J. L. Breslow. 1987. A unique AT-rich hypervariable minisatellite 3' to the apoB gene defines a high information restriction fragment length polymorphism. *J. Biol. Chem.* **262:** 8952–8955.
7. Jenner, K., A. Sidoli, M. Ball, J. R. Rodriguez, F. Pagani, G. Giudici, C. Vergani, J. Mann, F. E. Baralle, and C. C. Shoulders. 1988. Characterization of genetic markers in the 3' end of the apoB gene and their use in family and population studies. *Atherosclerosis.* **69:** 39–49.
8. Boerwinkle, E., W. Xiong, E. Fourest, and L. Chan. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3' hypervariable region. *Proc. Natl. Acad. Sci. USA.* **86:** 212–216.
9. Ludwig, E. H., W. Friedl, and B. J. McCarthy. 1989. High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* **45:** 458–464.
10. Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science.* **239:** 487–491.
11. Knott, T. J., S. C. Wallis, R. J. Pease, L. M. Powell, and J. Scott. 1986. A hypervariable region 3' to the human apolipoprotein B gene. *Nucleic Acids Res.* **14:** 9215–9216.
12. Friedl, W., E. H. Ludwig, B. Paulweber, F. Sandhofer, and B. J. McCarthy. 1990. Hypervariability in a minisatellite 3' of the apolipoprotein B gene in patients with coronary heart disease compared with normal controls. *J. Lipid Res.* **31:** 659–665.
13. Ludwig, E. H., and B. J. McCarthy. 1990. Haplotype analysis of the human apolipoprotein B mutation associated with familial defective apolipoprotein B100. *Am. J. Hum. Genet.* **47:** 712–720.
14. Wang, A. H. J., G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel, and A. Rich. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature.* **282:** 680–686.
15. McLean, M. J., J. A. Blaho, M. W. Kilpatrick, and R. D. Wells. 1986. Consecutive A • T pairs can adopt a left-handed DNA structure. *Proc. Natl. Acad. Sci. USA.* **83:** 5884–5888.
16. Freund, A. M., M. Bichara, and R. P. P. Fuchs. 1989. Z-DNA-forming sequences are spontaneous deletion hot spots. *Proc. Natl. Acad. Sci. USA.* **86:** 7465–7469.
17. Spitzner, J. R., I. K. Chung, and M. T. Muller. 1990. Eukaryotic topoisomerase II preferentially cleaves alternating purine-pyrimidine repeats. *Nucleic Acids Res.* **18:** 1–11.
18. Bowcock, A. M., A. Ray, H. Erlich, and P. B. Sehgal. 1989. Rapid detection and sequencing of alleles in the 3' flanking region of the interleukin-6 gene. *Nucleic Acids Res.* **17:** 6855–6864.
19. Ryan, S. C., and A. Dugaiczyk. 1989. Newly arisen DNA repeats in primate phylogeny. *Proc. Natl. Acad. Sci. USA.* **86:** 9360–9364.
20. Zuliani, G., and H. H. Hobbs. 1990. Tetranucleotide repeat polymorphism in the apolipoprotein B gene. *Nucleic Acids Res.* **18:** 4299.
21. Levy-Wilson, B., and C. Fortier. 1989. The limits of the DNase I-sensitive domain of the human apolipoprotein B gene coincide with the location of chromosomal anchorage loops and define the 5' and 3' boundaries of the gene. *J. Biol. Chem.* **264:** 21196–21204.