Brief Definitive Report

A Variable Number of Tandem Repeats Locus within the Human Complement C2 Gene Is Associated with a Retroposon Derived from a Human Endogenous Retrovirus

By Zeng Bian Zhu,* Shie-Liang Hsieh,‡ David R. Bentley,§ R. Duncan Campbell,‡ and John E. Volanakis*

*From the *Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama 35294; the ‡Medical Research Council Immunochemistry Unit, Department of Biochemistry, University of Oxford, Oxford OX1 3AU, United Kingdom; and the §Division of Medical and Molecular Genetics, United Medical and Dental Schools of Guy's and St. Thomas's Hospitals, Guy's Hospital, London SE1 9RT, United Kingdom*

Summary

We have previously described multiallelic restriction fragment length polymorphisms of the C2 gene, suggesting the presence of a variable number of tandem repeats (VNTR) locus. We report here the cloning and sequencing of the polymorphic fragments from the two most common alleles of the gene, a and b. The results confirm the presence of a VNTR locus consisting of a nucleotide sequence, 41 bp in average length, repeated tandemly 23 and 17 times in alleles a and b, respectively. The difference in the number of repeats between the two alleles is due to the deletion/insertion of two noncontiguous segments, 143 and 118 bp long, of allele a, and of a 40-bp segment of allele b. The VNTR region is associated with a SINE (short interspersed sequence)-type retroposon, SINE-R.C2, located within the third intron of the C2 gene. SINE-R.C2 is a member of a previously described large retroposon family of the human genome, apparently derived from the human endogenous retrovirus, (HERV) K10, which is homologous to the mouse mammary tumor virus.

The gene encoding the human second component of complement, C2, spans ~19 kb of DNA and maps within the class III region of the MHC on the short arm of chromosome 6 (1). The C2 gene is polymorphic, displaying four RFLPs (2–5) which give rise to at least nine C2 gene haplotypes (5). Two of these RFLPs detected by SstI and BamHI, are multiallelic and map at the 5′ region of the gene (3). It has been suggested that these multiallelic RFLPs detect a single variable number of tandem repeats (VNTR) locus within the C2 gene (3, 5). Many VNTR loci have been detected throughout the human genome (6, 7), and most of them are highly polymorphic and thus, useful in linkage studies, and particularly in forensic medicine (8). VNTRs are thought to result from deletion/insertions of the repeated nucleotide sequences or minisatellites via a mechanism of slipped-strand mispairing (9), or through frequent unequal crossing over primarily during meiosis (6, 10). To account for the high levels of allelic variability in repeat copy number, Jeffreys et al. (6) suggested that VNTRs encode hotspots of recombinational activity.

Five alleles of the putative VNTR locus of the C2 gene have been detected, characterized by SstI fragments of 2.75, 2.70, 2.65, 2.60, or 2.50 kb corresponding to BamHI fragments of 3.45, 3.40, 3.35, 3.30, or 3.20 kb (3, 5). The distribution of these alleles among 143 unrelated individuals was found to be skewed with 250 (87.4%) chromosomes having 2.70/3.40-kb SstI/BamHI fragments (allele a) and an additional 19 (6.6%) chromosomes having 2.5/3.2-kb SstI/BamHI fragments (allele b). The present study was undertaken to ascertain the presence of a VNTR locus in the C2 gene and to investigate the molecular basis for the low rate of polymorphic variation of the locus. Subcloning and nucleotide sequencing of the polymorphic fragments of the two most frequently observed alleles, a and b, demonstrated that a VNTR locus is indeed present within intron 3 of the C2 gene. The data also indicated that the repeated nucleotide sequences constitute part of a SINE (short interspersed sequence)-type retroposon, apparently derived from a human endogenous retrovirus, HERV-K10 (11), which is homologous to the mouse mammary tumor virus.

 J. Exp. Med. © The Rockefeller University Press • 0022-1007/92/06/1783/05 $2.00
Volume 175 June 1992 1783–1787

Materials and Methods

Three human genomic cosmid clones, S22A, provided by Dr. Thomas Spies, Harvard University (12), cos1a (1), and cosK101 (13) were used to isolate DNA fragments containing the hypervariable region of the C2 gene. Southern blotting analysis indicated that S22A contained the 2.5-kb SstI and the 3.2-kb BamHI fragments (allele b), while cos1a and cosK101 contained the 2.7- and 3.4-kb SstI and BamHI fragments, respectively (allele a). Additional RFLP analysis indicated that S22A contained haplotype b, and cos1a and cosK101 haplotype a of the C2 gene (5).

The cosmids S22A, cos1a, and cosK101 were digested with SstI, and the resulting restriction fragments were subcloned into pUC18. Subclone pS2.5, containing the polymorphic 2.5-kb SstI fragment, and subclone pS2.7, containing the 2.7-kb SstI fragment, were isolated by using a 300-bp BamHI/KpnI fragment as a probe (Fig. 1). This fragment was isolated from BamHI/KpnI double digests of plasmid pG850 (3) by electrophoresis on low-temperature melting agarose. The probe was labeled with ^{32}P-dCTP by using the random hexanucleotide priming method (14). The 2.5-kb insert of subclone pS2.5 and a 1.5-kb BamHI/SstI fragment of the insert of subclone pS2.7 were subcloned into M13mp18 and M13mp19 for nucleotide sequencing.

Nucleotide sequencing was carried out by the dideoxy chain termination method using modified bacteriophage T7 DNA polymerase (15). Buffer gradient gels were used to resolve the sequence reactions. The universal primer of M13 and a series of oligonucleotides, which were synthesized as sequence data became available, were used as primers for obtaining the complete sequence of both strands. Oligonucleotides were synthesized at the Oligonucleotide Synthesis Core Facility of the Comprehensive Cancer Center of this university. To resolve compression bands, dITP was substituted for dGTP in the sequence reactions (16). Nucleotide sequencing data were analyzed by using the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group, (Madison, WI), and the MacVector Sequence Analysis Software (International Biotechnologies, Inc., New Haven, CT).

Results and Discussion

The nucleotide sequences of the 2.5-kb SstI polymorphic fragment of allele b, consisting of 2,468 bp, and of the 1.5-kb BamHI/SstI fragment of allele a, consisting of 1,531 bp, were elucidated. Data on the structure and exon/intron organization of the C2 gene (17), allowed us to map the 5′ SstI and the BamHI restriction sites ~1,565 and 2,719 bp, respectively, downstream of exon 3 (Fig. 1). Comparison of the nucleotide sequences of the BamHI/SstI fragments of the two alleles demonstrated that two segments, 143 and 118 bp long, of allele a are absent from allele b, and a 40-bp segment of allele b is absent from allele a. These and eight additional single nucleotide deletions/insertions account for a net difference of 217 bp between the two alleles, which is in good agreement with the 0.2-kb difference determined by Southern blotting (5). The nondeleted nucleotide sequences of the two alleles differ from each other in 11 of 1,262 positions.

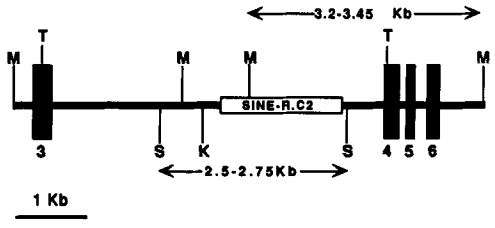

Figure 1. Partial restriction map of the C2 gene region that contains the SINE-R.C2 retroposon. (*Open box*) SINE-R.C2. (*Black boxes*) exons. (*M*) BamHI; (*T*) TaqI; (*S*) SstI; and (*K*) KpnI. (*Arrows*) SstI and BamHI multiallelic polymorphic fragments.

A computer-assisted search of the EMBL and the GenBank data bases indicated that a region of the SstI multiallelic polymorphic fragment of the C2 gene is highly homologous to R11, a member of a previously described SINE-type retroposon family (18). The C2 gene and the retroposon are in opposite orientations. The retroposon homology region spans 808 bp of allele b, starting at nucleotide 798, and the corresponding 593-bp segment of the available sequence of allele a, starting at the BamHI restriction site (nucleotide 1155). SINE-R11 is one of three homologous human nonviral retroposons, isolated from a human fetal liver genomic library by screening with DNA probes derived from a human endogenous retrovirus, HERV-K10 (11). An estimated 4–5,000 additional copies of this retroposon are present per haploid human genome. To our knowledge, this is the first member of this large retroposon family mapped to a chromosomal site. HERV-K10 is a 9.2-kb genome present in ~50 copies per haploid human genome, and is homologous to both type A retroviruses and to the type B mouse mammary tumor virus (11). Both SINE-R11 and the retroposon of the C2 gene (SINE-R.C2) are homologous to a region near the 3′ end of HERV-K10. An alignment of the homologous nucleotide sequences of SINE-R.C2 of the two C2 alleles, SINE-R11, and HERV-K10 is shown in Fig. 2. As shown, the homology region of the HERV-K10 consists of two noncontiguous subregions. The first is 424 bp in length, starting at nucleotide 8407, and includes 141 bp of the 3′ end of the *env* gene and the adjacent and partially overlapping 330-bp long segment of the 3′ LTR of HERV-K10. The following 367 bp of HERV-K10 (8831–9197), which include a putative TATA box, are absent from both C2 and SINE-R11. This segment is followed by the second homology subregion of HERV-K10, which spans 102 bp of the 3′ LTR (9198–9299) and includes a putative polyadenylation signal, AATAAA.

The homology between SINE-R.C2 and SINE-R11 extends upstream and downstream of the HERV-K10 homology region, spanning the entire length of SINE-R11 except for 19 bp at its 3′ end. The overall nucleotide identity between SINE-R.C2 and SINE-R11 is 91%. At the 5′ end of the C2 gene homology region which constitutes the 3′ end of SINE-R11 there is a poly(T) tract possibly representing the poly(A) tail of a partial transcript of HERV-K10. The 3′ end of the C2 homology region and the corresponding 5′ end of SINE-R11 are composed of tandemly repeated, G- and C-rich nucleotide sequences. SINE-R11 contains six tandem repeats as compared to 23 and 17 for the C2 alleles a and b, respectively.

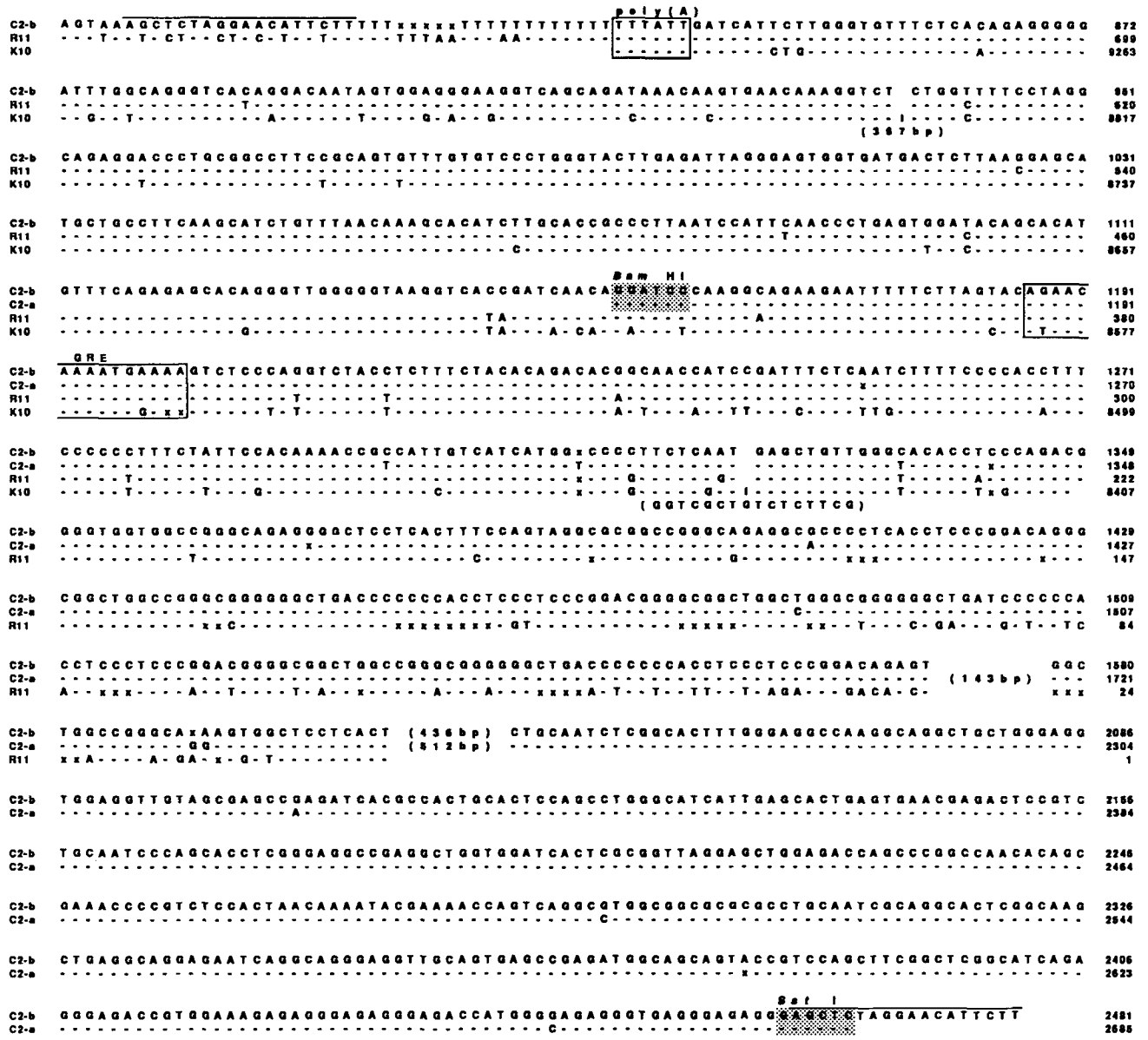

Figure 2. Comparison of the nucleotide sequences of the two C2 alleles, a and b, the retroposon SINE-R11 (18), and the human endogenous retrovirus, HERV-K10 (11). Numbering of the C2 sequence is from the upstream SstI site. (*Dashes*) nucleotide identity. (*x*) deletions. Segments unique to any of the nucleotide sequences are indicated by the number of bp in parentheses. (*White boxes*) Polyadenylation signal of HERV-K10, and glucocorticoid response element (GRE) motif. (*Shaded boxes*) BamHI and SstI restriction sites of the C2 gene. (*Overbar*) The 18-bp direct repeat possibly corresponding to the target site duplication of the SINE-R.C2. The nucleotide sequence of C2b downstream of the SstI site was obtained from a subclone, pM3.2, containing the BamHI polymorphic fragment (Fig. 1). These sequence data are available from EMBL under accession numbers Z11739 and Z11740.

The tandem repeats of both C2 alleles are followed by a 427-bp-long segment (2042–2468, allele b) which appears to consist of remnants of nucleotide repeats. The presence of a 18-bp-long nucleotide sequence, AGCTCTAGGAACATTCTT, at the 3′ end of this segment which is a direct repeat of a 5′ end sequence (Fig. 2), suggests that this region is part of the retroposon. Assuming that this 18-bp long sequence represents a target site duplication, the length of SINE-R.C2, 1,884 and 1,667 bp for allele a and b, respectively, is larger than that of the average SINE-type retroposon (19).

An alignment of the tandem repeats of the two C2 alleles is shown in Fig. 3. Inspection of the repeated sequences indicates two types of repeats that are homologous to each other but differ in length. The most common, type I repeat, consists of 40 bp, while the type II repeat consists of 49 bp. There are 16 type I and 6 type II repeats in allele a, and 12 and 5, respectively, in allele b. In addition, a half repeat is present only in allele a. Type I repeats of both alleles are 79% homologous to their consensus sequence, which is 85% homologous to that of the tandem repeats of the three previously

sequenced members of the SINE-R family of retroposons (18). The substantial divergence of type I repeats may explain the relatively low allelic variation of the C2 VNTR locus. However, the longer type II repeats are more homogeneous, displaying 91% homology to their consensus sequence (Fig. 3). This raises the possibility that the locus may be more polymorphic than indicated by presently available data. Two conserved core sequences are present in all repeats. The first, C$\overset{T}{C}$CC$\overset{T}{C}$CAC is near the 5′ end, and the second, GGCCGGGC, is near the 3′ end of each repeat. These core sequences could have been involved in the observed deletions/insertions. The 143-bp segment present only in allele a includes 3.5 repeats and could be explained by an unequal crossing-over involving the 3′ conserved core. The 40- and 118-bp segments present only in allele b and a, respectively, include one and three intact type I repeats, respectively, and could be explained by separate unequal crossing-over events involving the 5′ conserved core (Fig. 3). The three observed deletions/insertions probably represent three distinct genetic events. Each of these events should have resulted in two reciprocal alleles leading to a larger total number of alleles than the reported five (3, 5). Thus, it seems possible that additional allelic variation of the C2 gene exists not represented in the rather limited population samples studied or alternatively, not detected by the methodology used. Given the value of a highly polymorphic marker within the MHC in disease-linkage analyses, additional population studies seem warranted.

Certain additional structural features of interest are present in the VNTR region of the C2 gene. Five imperfect palindromes, GGGGGGCTGA(T)CCCCCC, are present in both alleles (Fig. 3). They extend from the 3′ end of a repeat to the 5′ end of the following repeat. These palindromes have

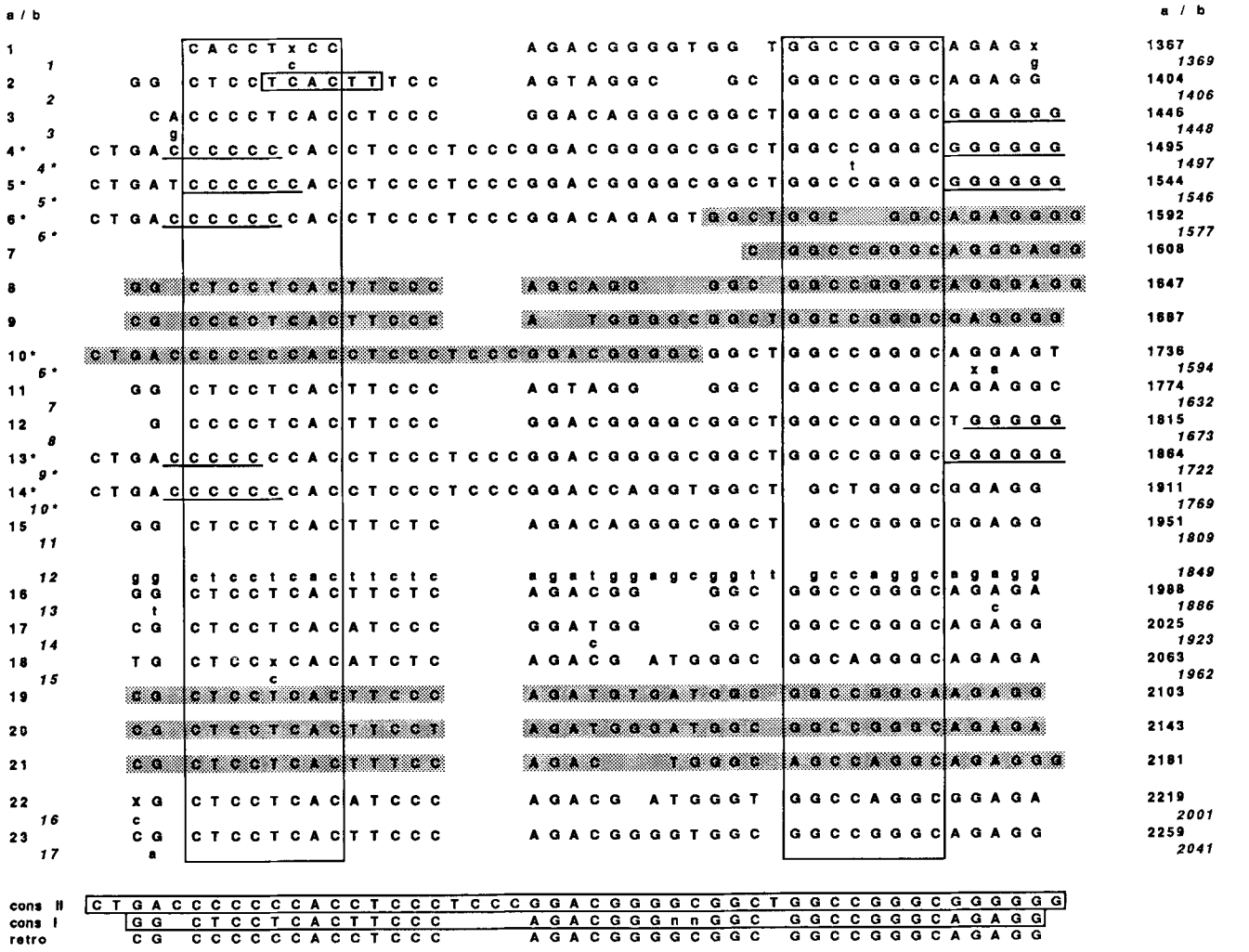

Figure 3. Alignment of the nucleotide repeats of alleles a and b of the C2 gene. For allele b, only nucleotide substitutions and insertions are shown in lower case letters. (*x*) Deletions in either allele. Repeats are numbered on the left. Allele b repeats are in italics. Stars indicate type II repeats. Nucleotide numbering (allele b in italics) is from the upstream SstI site. The two segments present only in allele a are shaded. The sequence of the repeat present only in allele b is given in lower case letters. (*Vertical boxes*) The two conserved cores. The sequence motif complementary to the 6-bp inducible enhancer of the interferon-β gene is boxed. The five imperfect palindromes are underlined. (*Bottom*) The consensus sequences of type I and II repeats and of the three previously characterized (18) members of the SINE-R family (retro).

the potential to form cruciform structures (20, that could contribute to the restricted allelic variation of the VNTR locus. A 15-bp nucleotide motif, AGAACAAAATGAAAA, starting at nucleotide 1187 of the C2 sequence (Fig. 2) is complementary to the consensus 15 mer of binding sites for glucocorticoid receptors of steroid-regulated genes (21). A sequence motif, TCACTT, complementary to the AAGTGA inducible enhancer element of the interferon-β gene (22), is present in 11 and 7 copies in allele a and b, respectively. A number of potential Sp1 binding sites (GGGCGG) are also present in both alleles. The presence of these elements raises the possibility that this region of the SINE-R.C2 has regulatory activity for the C2 gene. This possibility is currently under investigation.

We thank Mrs. Paula Kiley for expert secretarial assistance.

This work was supported in part by US Public Health Service grants AI-21067 and AR-03555.

Address correspondence to Dr. John E. Volanakis, Division of Clinical Immunology & Rheumatology, University of Alabama at Birmingham, UAB Station, THT 437, Birmingham, AL 35294.

Received for publication 10 February 1992.

References

1. Carroll, M.C., R.D. Campbell, D.R. Bentley, and R.R. Porter. 1984. A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2, and factor B. *Nature (Lond.).* 307:237.
2. Woods, D.E., M.D. Edge, and H.R. Colten. 1984. Isolation of a complementary DNA clone for the human complement protein C2 and its use in the identification of a restriction fragment length polymorphism. *J. Clin. Invest.* 74:634.
3. Bentley, D.R., R.D. Campbell, and S.J. Cross. 1985. DNA polymorphism of the C2 locus. *Immunogenetics.* 22:377.
4. Cross, S.J., J.H. Edwards, D.R. Bentley, and R.D. Campbell. 1985. DNA polymorphism of the C2 and factor B genes: detection of a restriction fragment length polymorphism which subdivides haplotypes carrying the C2C and factor BF alleles. *Immunogenetics.* 21:39.
5. Zhu, Z.B., and J.E. Volanakis. 1990. Allelic associations of multiple RFLPs of the gene encoding complement protein C2. *Am. J. Hum. Genet.* 46:956.
6. Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985. Hypervariable "minisatellite" regions in human DNA. *Nature (Lond.).* 314:67.
7. Nakamura, Y., M. Lepert, P. O'Connell, R. Wolf, T. Holms, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R. White. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science (Wash. DC).* 235:1616.
8. Jeffreys, A.J., A. MacLeod, K. Tamaki, D.L. Neil, and D.G. Monckton. 1991. Minisatellite repeat coding as a digital approach to DNA typing. *Nature (Lond.).* 354:204.
9. Levinson, G., and G.A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4:203.
10. Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature (Lond.).* 332:278.
11. Ono, M., T. Yasumaga, T. Miyata, and H. Ushikubo. 1986. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *J. Virol.* 60:589.
12. Spies, T., M. Bresnahan, and J.L. Strominger. 1989. Human major histocompatibility complex contains a minimum of 19 genes between the complement cluster and HLA-B. *Proc. Natl. Acad. Sci. USA.* 86:8955.
13. Sargent, C.A., I. Dunham, and R.D. Campbell. 1989. Identification of multiple HTF-island associated genes in the major histocompatibility complex class III region. *EMBO (Eur. Mol. Biol. Organ.) J.* 8:2305–2312.
14. Feinberg, A.P., and B. Vogelstein. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* 132:6.
15. Tabor, S., and C.C. Richardson. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. USA.* 84:4767.
16. Mills, D.R., and F.R. Kramer. 1979. Structure-independent nucleotide sequence analysis. *Proc. Natl. Acad. Sci. USA.* 76:2232.
17. Ishii, Y., Z.B. Zhu, K.J. Macon, and J.E. Volanakis. 1991. Structure of the gene for human complement component C2. *Complement Inflammation.* 8:167. (Abstr.).
18. Ono, M., M. Kawakami, and T. Takezama. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res.* 15:8725.
19. Weiner, A.M., P.L. Deininger, A. Efstratiadis. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* 55:631.
20. Lilley, D.M. 1980. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc. Natl. Acad. Sci. USA.* 77:6468.
21. Beato, M., G. Chalepakis, M. Schaver, and E.P. Slater. 1989. DNA regulatory elements for steroid hormones. *J. Steroid Biochem.* 32:737.
22. Fujita, T. H. Shibuya, H. Hotta, K. Yamanishi, and T. Taniguchi. 1987. Interferon-β gene regulation: tandemly repeated sequences of a synthetic 6 bp oligomer function as a virus-inducible enhancer. *Cell.* 49:357.