

# Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons

(*CD44* gene)

GAVIN R. SCREATON, MARTYN V. BELL, DAVID G. JACKSON, FRANCOIS B. CORNELIS, ULRICH GERTH, AND JOHN I. BELL

Molecular Immunology Group, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom

Communicated by David Weatherall, August 27, 1992 (received for review June 26, 1992)

**ABSTRACT** The CD44 molecule is known to display extensive size heterogeneity, which has been attributed both to alternative splicing and to differential glycosylation within the extracellular domain. Although the presence of several alternative exons has been partly inferred from cDNA sequencing, the precise intron–exon organization of the *CD44* gene has not been described to date to our knowledge. In the present study we describe the structure of the human *CD44* gene, which contains at least 19 exons spanning some 50 kilobases of DNA. We have identified 10 alternatively spliced exons within the extracellular domain, including 1 exon that has not been previously reported. In addition to the inclusion or exclusion of whole exons, more diversity is generated through the utilization of internal splice donor and acceptor sites within 2 of the individual exons. The variation previously reported for the cytoplasmic domain is shown to result from the alternative splicing of 2 exons. The genomic structure of *CD44* reveals a remarkable degree of complexity, and we confirm the role of alternative splicing as the basis of the structural and functional diversity seen in the CD44 molecule.

The human CD44 glycoprotein [Pgp-1 (1), HCAM (2), Hermes antigen (3), ECRM III (4)] has been proposed to function as a lymph node homing receptor on circulating lymphocytes. Expressed on a wide range of different tissues, the CD44 molecule also binds the extracellular matrix components hyaluronic acid (5), fibronectin (6), and collagen (4) as well as the cytoskeletal protein ankyrin (7). Several antibodies recognizing CD44 have been shown to induce lymphocyte activation (8, 9) and to inhibit lymphopoiesis (10). In addition, CD44 can mediate both homotypic and heterotypic cell adhesion (11, 12). The many functional roles of this molecule may relate to its considerable size heterogeneity, which cannot be accounted for simply by differences in glycosylation (5, 13, 14).

Recently, we and others have isolated a number of different isoforms of CD44 (15–18). These have been characterized by cDNA sequencing and appear to arise by alternative splicing in two different regions, the membrane proximal extracellular domain and the cytoplasmic tail. Some of this variation has been shown to produce functional changes in the molecule. For example, the presence of the 396-base-pair (bp) insert in the epithelial variant of CD44 reduces the affinity for hyaluronic acid (5). Study of these cDNA variants has given some insight into the genomic organization of *CD44* but has not allowed the precise characterization of the number and boundaries of exons that encode the variant region of the molecule.

In the present paper we have cloned the gene for CD44 from a human yeast artificial chromosome (YAC) and characterized its genomic structure.\* These studies reveal a

remarkable degree of complexity within the structure of the *CD44* gene and demonstrate conclusively the role of alternative splicing in generating the structural heterogeneity that is characteristic of the CD44 molecule.

## MATERIALS AND METHODS

**Isolation, Characterization, and Subcloning of a *CD44*-Containing YAC.** The primer pair F11–R11 was used to screen a YAC library provided by the U.K. Human Genome Mapping Resource Centre (19) by the PCR method (20). Total yeast DNA from positive YACs was prepared in agarose blocks (21) and analyzed by pulsed-field gel electrophoresis, Southern transfer, and hybridization to <sup>32</sup>P-labeled CD44 cDNA variant E (17). A partial restriction map was generated by the partial digestion and indirect end-labeling method (21).

To obtain subclones of the YAC, total yeast DNA in agarose blocks from one of the YAC clones was digested for 2 min with 1 unit of *Taq* I per  $\mu$ g of DNA. After phenol/chloroform extraction, the fragments were cloned into the unique *Cla* I site of pL53In (22) and used to transform *Escherichia coli* DH5 $\alpha$  cells. Clones containing *CD44* were identified by colony hybridization to <sup>32</sup>P-labeled CD44 variant E (17) and intron PCR products (this paper). Plasmid DNA was prepared by standard methods (23), and the *CD44* exon content was determined by dot-blot hybridization to <sup>32</sup>P-end-labeled oligonucleotides.

**PCR Amplification.** One-hundred nanograms of total yeast DNA or 500 ng of human placental DNA was used as template in a 50- $\mu$ l reaction mix that contained 10 mM Tris-HCl (pH 8.4), 50 mM KCl, 2.5 mM MgCl<sub>2</sub>, 250  $\mu$ M each dNTP, the appropriate primers at 200 nM, and 2 units of *Taq* DNA polymerase. The cycle parameters were 5 min at 95°C, 1 min at 63°C, and 3 min at 72°C (1 cycle) followed by 1 min at 95°C, 1 min at 63°C, and 3 min at 72°C for 40 cycles. PCR products were verified by direct sequencing across the splice junctions. Sequencing of double-stranded templates, either PCR products or plasmids, was performed by using <sup>32</sup>P-end-labeled primers and *Taq* DNA polymerase (24). PCR primers used to amplify introns are given in Table 1.

**Single Primer PCR.** In cases where it was not possible to amplify whole introns by PCR and no YAC subclones were available, we developed a single-primer PCR method to sequence across the splice junction. Buffer conditions were the same as for standard PCR (see above), and the primer concentrations were doubled. Three separate 60-cycle amplifications used annealing temperatures of 50°C, 60°C, and 70°C. The products of these three reactions were pooled and sequenced directly; an internal primer was used to provide specificity.

Abbreviations: YAC, yeast artificial chromosome; 3' UTR, 3' untranslated region.

\*The sequences reported in this paper have been deposited in the GenBank data base (accession nos. L05407–L05424 for *CD44* exons 2–19).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Primers for PCR amplification of introns

Primer	Sequence
F2	AACCTGCCGCTTTGCAGGTGTATT
R3	GAAGCAATATGTGTACTACTGGGAG
F4	CTCCACCTGAAGAAGATTGTACATC
F5	TTACACCTTTTCTACTGTACACCCC
R5	TAGCAGGGATTCTGTCTGTGCTGT
F6	CACTTTGATGAGACTAGTGTACAG
R6	CAGCCATTGTGTGTTGTGTGAAG
F7	GTACGTCTTCAAATACCATCTCAGC
R7	CTTCATCATCATCAATGCCTGATCC
F8	ACGGCCCTTTGACCCACACAAAACA
R8	TTGAATGGCTTGGGTTCCACTGG
F9	CCCCTCATTACCATGAGCATCAT
R9	GCTTGTAGAATGTGGGGTCTCTTC
F10	CCAGGCAACTCCTAGTAGTACAAC
R10	GAATGGGAGTCTTCTCTGGGTGTT
F11	CTCAGCTATACCAGCCATCCAAT
R11	CCATCCTTCTTCTGCTTGATGAC
F12	TGGACTCCAGTCATAGTACAACGC
R12	GTCATTGAAAGAGGTCTGTCTCTG
F13	GCAGAGTAATTCTCAGAGCTTCTCT
R13	TGTCAGAGTAGAAGTTGTTGGATGG
F14	ATGTCACAGGTGAAGAAGAGACC
R14	TGGAATCTCCAACAGTAACTGCAGT
F15	tcagtagtactgaccttctgattgc
R15	TCAGATCCATGAGTGTATGGGAC
F16	gggggaaatcctttgtgtaatgct
R16	TTTGGGGTGTCTTATAGGACCAG
F17	ccagaattacctgcctattggctg
R17	GACTGCAATGCAAAGTCAAGAATC
R18	AGGGAAAATGAGGAAGCTGGAAGG
F19	gatgatctgatgcctgagtcactg
R19	ccttaccatctcagctcttctg

Sequences begin with the 5' nucleotide. Primers in lowercase letters are based upon the intron sequence and will amplify across the nearest adjacent exon. The numbers correspond to the exon nearest to the primer (eg., the combination of F2 with R3 will amplify the intron between exons 2 and 3 and will have part of exons 2 and 3 at each end).

## RESULTS AND DISCUSSION

**Genomic Map of *CD44*.** A YAC clone containing the *CD44* gene was obtained by PCR screening of a library provided by the U.K. Medical Research Council Human Genome Mapping Resource Centre. Restriction and pulsed-field gel electrophoresis analyses (not shown) indicate that this clone is approximately 320 kilobases (kb) in size and that the *CD44*

gene is contained within a 105-kb *Sal* I fragment and occupies some 50–60 kb of genomic DNA upon adjacent *Cla* I fragments of approximately 29 and 31 kb.

A partial restriction map of the YAC is shown in Fig. 1A with the *CD44* portion enlarged in Fig. 1B. The 5' end of the gene is located by the sites for *Sma* I and *Sac* II, which are present in the published sequence approximately 200 bp 5' of the initiation codon (18). Portions of the *CD44* gene, either intron PCR products or plasmid subclones, were assigned to intervals in the map by restriction analysis (Fig. 1C).

Most of the introns, except those between exons 1 and 2, 3 and 4, 5 and 6, and 18 and 19, were readily amplified from genomic DNA; PCR products or subclones were sequenced to reveal the precise splice boundaries and the flanking splice sites. The coding sequence of the gene and intron sizes estimated from PCR are shown in Fig. 2 along with the flanking intron sequences; the intron at the 3' end of exon 3 was not amplified by either conventional or single-primer PCR.

**Membrane Proximal Extracellular Domain.** The region of the *CD44* gene encoding the membrane-proximal extracellular domain spans 25 kb of genomic DNA and contains at least 10 alternatively spliced exons (5–14) including one, exon 6, that has not been previously described (Fig. 3). Nine of these variable region exons (exons 6–14) can be skipped by alternative splicing, while exon 5 contains an alternative splice donor site; in addition, exon 7 contains two distinct splice acceptor sites. A variant has also been described in the rat, where the homologue of human exon 15 has been spliced out (25), giving even greater potential diversity to this region.

Exon 6 was identified in cDNA from the colon carcinoma cell line HT-29 by PCR, where we designed primers to search for new exons by amplification between variable region exons that were presumed to be adjacent. Amplification between exons 5 and 7 consistently produced a product 320 bp long in addition to the expected product of 191 bp (Fig. 4). The larger product was directly sequenced and shown to contain the 129-bp insert, exon 6 (Fig. 2). The sequence of exon 6 encodes a 43-amino acid peptide that contains a high proportion of serine and threonine residues, a feature that is shared with the other alternative exons, allowing for extensive O-linked glycosylation. This new exon was also detected in several other cDNAs including mammary and bladder tumors by PCR (data not shown). We have preliminary evidence for further alternatively spliced exons lying between exons 5 and 6 in the mouse (data not shown).

The cDNA sequences for this variable region can now be located on our new exon map (Fig. 3); exons 14, 13, 12, 7, and 11 correspond to exons 1–5 previously reported from our

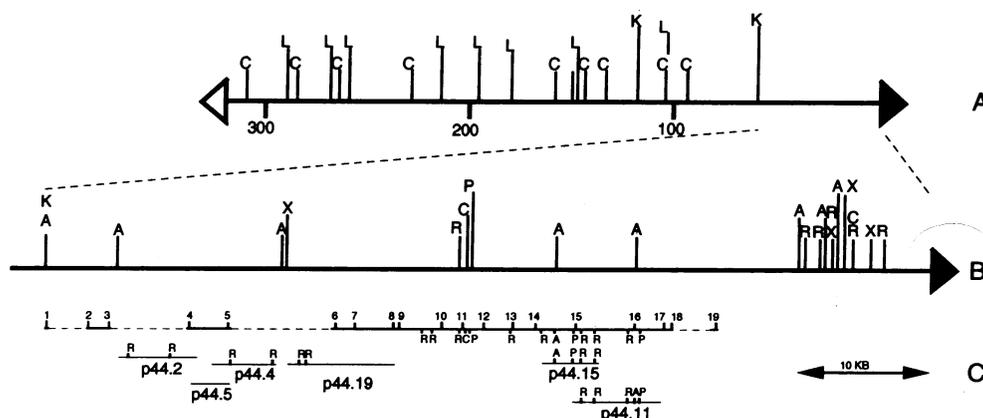


Fig. 1. Genomic organization of the *CD44* gene. (A) Partial restriction map of the whole YAC. (B) Terminal portion shown in more detail. (C) Exon distribution aligned with the YAC restriction map and plasmid subclones. Restriction site symbols are as follows: K, *Sac* II; A, *Sma* I; L, *Sal* I; C, *Cla* I; X, *Xho* I; R, *Eco*RI; P, *Kpn* I. The open and solid triangles represent the *URA3* and *TRP1* YAC vector arms, respectively. The dashed lines in C represent introns, the sizes of which can only be estimated from the restriction data and cannot be directly measured.



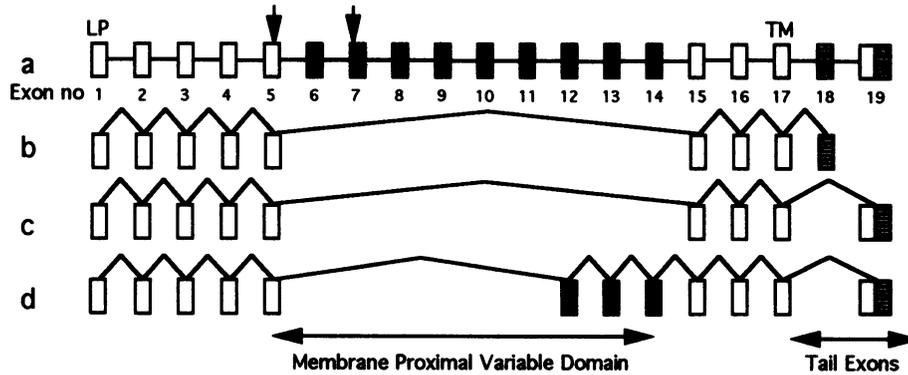


FIG. 3. (a) Schematic exon map showing all 19 exons of *CD44* (see Fig. 2). Open boxes represent constitutive exons, and solid boxes represent alternative exons that can be wholly spliced out. Alternative splice donor and acceptor sites on exons 5 and 7 are marked with arrows, and the crosshatching represents the 3' UTR sequence. Exons encoding the leader peptide LP, and transmembrane region TM are marked. (b and c) Hemopoietic variants of *CD44* encoding short (b) or long (c) cytoplasmic tails. (d) Exons encoding the epithelial form of the molecule.

(18). In the less abundant form, the DNA encoding the cytoplasmic tail is truncated to 9 bp and is followed by a distinct 3' UTR. As the junction between short and long tails contains the invariant part of the splice donor consensus GT in humans, baboons (28), and mice (29), it has been proposed that the short tail is generated by splicing from an internal splice donor sequence within the long tail to a downstream 3' untranslated exon (30). This is not the case; in fact characterization of the exons encoding the cytoplasmic tails reveals that the coding and 3' UTR of the long tail are both carried on exon 19, while exon 18 carries the 3' UTR associated with the short-tailed variant of *CD44* and exon 17 carries the transmembrane domain and first three amino acids of the cytoplasmic tail, which are shared by both long- and short-tailed variants (Fig. 2).

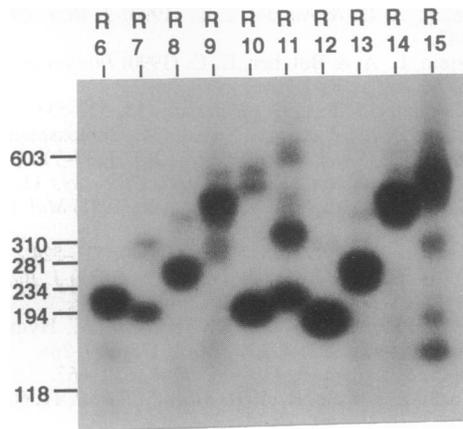


FIG. 4. Autoradiogram of PCR products generated by using cDNA from the colon cancer line HT-29 as template. A common primer F5 was used in all of the reactions along with exon-specific primers for the variable region exons, R6–R14. The product generated by the combination of F5 with R15 spans the whole of the insertion site in the membrane-proximal extracellular domain. Products were run on a 2% agarose gel and then transferred onto Hybond-N<sup>+</sup> (Amersham). After transfer, the membrane was probed with the oligonucleotide R5, which lies internal to F5 in the backbone sequence of the molecule and will detect all *CD44*-containing products. The figure shows the two products generated with the primer pair F5–R7, the upper band contains exon 6 and also shows the diversity of transcripts produced by this cell line with more than one product in most of the lanes. A major band of 534 bp in the products generated by using the primer pair of F5–R15, which span the whole insertion site, corresponds to the epithelial variant of *CD44*, containing variable exons 12–14, while the lowest band of 148 bp represents the hemopoietic variant lacking exons 6–14 (Fig. 3).

**Conclusions.** With 12 of 19 exons being capable of alternative splicing and the common use of alternative donor or acceptor splice sites within exons, *CD44* is potentially one of the most extensively alternatively spliced genes reported to date. Both the fibronectin gene (31) and *CD45* (32) contain three alternatively spliced exons, whereas the rat  $\alpha$  tropomyosin gene (33) and *NCAM* (34) can alternatively splice 10 of 15 exons and 12 of 27 exons, respectively. None of these examples contain a tandem array of at least 10 exons that appear to be selected at random as is the case with *CD44*, which gives enormous potential diversity. Tissue-specific patterns of alternative splicing have been shown to have important functional correlates in several systems including sex determination in *Drosophila* (35) and ligand specificity for fibronectin (31). In *CD44* the role of one alternatively spliced variant in enhancing the metastatic potential of rat tumors (25) indicates that the membrane proximal extracellular domain may have important effects in mediating cell adhesion and targeting. This is supported by the effect of the introduction of exons 12–14 in abrogating hyaluronic acid binding by the human epithelial variant (5). The functional roles of many other splice variants of *CD44* remain to be established.

The alternative splicing of the cytoplasmic tail may relate to the potential role of the *CD44* molecule in T-cell activation. Alteration in the length of the cytoplasmic tail may modulate intracellular signaling, as the cytoplasmic tail has potential sites for the action of both protein kinases A and C (14). Furthermore, it has been shown that *CD44* associates with cytoskeletal actin filaments and ankyrin, presumably via the cytoplasmic tail. In macrophages activation increases *CD44* phosphorylation and reduces contacts with the cytoskeleton (14). In lymphocytes *CD44* has been demonstrated to be associated with protein kinase C. Protein kinase C activation by anti-*CD44* monoclonal antibodies may be involved in both ankyrin binding (36) and the promotion of T-cell adhesion via LFA-1 (lymphocyte function-association antigen 1) (11). In addition, variation in the length of the cytoplasmic tail may also affect the ability of the molecule to bind specific ligands through the N-terminal domain, as the short tail has been shown to reduce hyaluronic acid binding by the hemopoietic variant (37). Interestingly the DNA sequence for the short-tail 3' UTR carries a long A+T-rich tract (Fig. 2). Similar A+U-rich elements have been found in the 3' UTRs of several mRNAs coding for growth factors and protooncogenes and are implicated in targeting the mRNA for rapid turnover (38).

The tissue-specific patterns of *CD44* alternative exon usage suggest that the regulation of splicing not only is responsible for the characteristic size heterogeneity but also may affect the function of the molecule. Although individual cell lines

appear to express a number of different CD44 transcripts (Fig. 4), it remains to be established whether single cells also express more than one isoform of CD44.

Comparison of splice donor and acceptor sites flanking the alternatively spliced exons has revealed no obvious variation from consensus sequences. However, analysis of the polypyrimidine tract from -14 to -4 bp before the splice acceptor site reveals a pyrimidine content of 87% for the constitutive exons (1-5 and 15-17) compared with 65% for the alternative exons (6-14, 18, and 19). Increasing the pyrimidine content of this tract has been demonstrated to reverse exon skipping in artificial constructs (39).

The genomic structure of *CD44* presented here should facilitate the characterization of isoforms expressed in a number of different cell types including tumors and lymphocytes. In lymphocytes the regulated expression of *CD44* isoforms may be involved in activation, targeting, and ontogeny, while in nonlymphoid tissues it may play an important role in the control of cell adhesion. Future studies on the nature and control of alternative splicing of *CD44* together with functional studies of the alternatively spliced exons, particularly their ligand specificities, may help to shed further light onto these events.

We thank Zoe Christolodoulou, Mike Francis, and Kay Davies and the Medical Research Council Human Genome Mapping Resource Centre for assistance in isolating the CD44 YAC; Michael Reth for the plasmid vector; David Simmons for the HT 29 cDNA; Ken Smyth for mammary and bladder tumor RNAs; Penny Catlin for producing numerous oligonucleotides; Charles Bangham for discussions on the single primer PCR sequencing method; and Sarah Wood for her help in preparing the manuscript. This work was supported by the Wellcome Trust. G.R.S. is a Medical Research Council Training Fellow.

- Zhou, D. F., Ding, J. F., Picker, L. J., Bargatze, R. F., Butcher, E. C. & Goeddel, D. V. (1989) *J. Immunol.* **143**, 3390-3395.
- Goldstein, L. A., Zhou, D. F., Picker, L. J., Minty, C. N., Bargatze, R. F., Ding, J. F. & Butcher, E. C. (1989) *Cell* **56**, 1063-1072.
- Jalkanen, S. T., Bargatze, R. F., Herron, L. R. & Butcher, E. C. (1986) *Eur. J. Immunol.* **16**, 1195-1202.
- Wayner, E. A., Carter, W. G., Piotrowicz, R. S. & Kunicki, T. J. (1988) *J. Cell Biol.* **107**, 1881-1891.
- Stamenkovic, I., Aruffo, A., Amiot, M. & Seed, B. (1991) *EMBO J.* **10**, 343-348.
- Jalkanen, S. & Jalkanen, M. (1992) *J. Cell Biol.* **116**, 817-825.
- Kalomiris, E. L. & Bourguignon, L. Y. (1988) *J. Cell Biol.* **106**, 319-327.
- Denning, S. M., Le, P. T., Singer, K. H. & Haynes, B. F. (1990) *J. Immunol.* **144**, 7-15.
- Huet, S., Groux, H., Caillou, B., Valentin, H., Priour, A. M. & Bernard, A. (1989) *J. Immunol.* **143**, 798-801.
- Miyake, K., Medina, K. L., Hayashi, S., Ono, S., Hamaoka, T. & Kincade, P. W. (1990) *J. Exp. Med.* **171**, 477-488.
- Koopman, G., van Kooyk, Y., de Graaff, M., Meyer, C. J., Figdor, C. G. & Pals, S. T. (1990) *J. Immunol.* **145**, 3589-3593.
- Shimizu, Y., Van Seventer, G., Siraganian, R., Wahl, L. & Shaw, S. (1989) *J. Immunol.* **143**, 2457-2463.
- Brown, T. A., Bouchard, T., St John T., Wayner, E. & Carter, W. G. (1991) *J. Cell Biol.* **113**, 207-221.
- Camp, R. L., Kraus, T. A. & Pure, E. (1991) *J. Cell Biol.* **115**, 1283-1292.
- Hofmann, M., Rudy, W., Zoller, M., Tolg, C., Ponta, H., Herrlich, P. & Gunthert, U. (1991) *Cancer Res.* **51**, 5292-5297.
- Dougherty, G. J., Landorp, P. M., Cooper, D. L. & Humphries, R. K. (1991) *J. Exp. Med.* **174**, 1-5.
- Jackson, D. G., Buckley, J. & Bell, J. I. (1992) *J. Biol. Chem.* **267**, 4732-4739.
- Shtivelman, E. & Bishop, J. M. (1991) *Mol. Cell. Biol.* **11**, 5446-5453.
- Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D. & Olson, M. V. (1989) *Science* **244**, 1348-1351.
- Green, E. D. & Olson, M. V. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1213-1217.
- Anand, R., Riley, J. H., Butler, R., Smith, J. C. & Markham, A. F. (1990) *Nucleic Acids Res.* **18**, 1951-1956.
- Auch, D. & Reth, M. (1990) *Nucleic Acids Res.* **18**, 6743-6744.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
- Murray, V. (1989) *Nucleic Acids Res.* **17**, 8889.
- Gunthert, U., Hofmann, M., Rudy, W., Reber, S., Zoller, M., Haussmann, I., Matzku, S., Wenzel, A., Ponta, H. & Herrlich, P. (1991) *Cell* **65**, 13-24.
- Cooper, D. L., Dougherty, G., Harn, H. J., Jackson, S., Baptist, E. W., Byers, J., Datta, A., Phillips, G. & Isola, N. R. (1992) *Biochem. Biophys. Res. Commun.* **182**, 569-578.
- Stamenkovic, I., Amiot, M., Pesando, J. M. & Seed, B. (1989) *Cell* **56**, 1057-1062.
- Isacke, C. M., Sauvage, C. A., Hyman, R., Lesley, J., Schulte, R. & Trowbridge, I. S. (1986) *Immunogenetics* **23**, 326-332.
- Wolffe, E. J., Gause, W. C., Pelfrey, C. M., Holland, S. M., Steinberg, A. D. & August, J. T. (1990) *J. Biol. Chem.* **265**, 341-347.
- Goldstein, L. A. & Butcher, E. C. (1990) *Immunogenetics* **32**, 389-397.
- Schwarzbauer, J. E. (1991) *Bioessays* **13**, 527-533.
- Rothstein, D. M., Saito, H., Streuli, M., Schlossman, S. F. & Morimoto, C. (1992) *J. Biol. Chem.* **267**, 7139-7147.
- Lees, M. J. & Helfman, D. M. (1991) *Bioessays* **13**, 429-437.
- Reyes, A. A., Small, S. J. & Akeson, R. (1991) *Mol. Cell. Biol.* **11**, 1654-1661.
- Baker, B. S. (1989) *Nature (London)* **340**, 521-524.
- Kalomiris, E. L. & Bourguignon, L. Y. (1989) *J. Biol. Chem.* **264**, 8113-8119.
- Lesley, J., He, Q., Miyake, K., Hamann, A., Hyman, R. & Kincade, P. W. (1992) *J. Exp. Med.* **175**, 257-266.
- Shaw, G. & Kamen, R. (1986) *Cell* **46**, 659-667.
- Dominski, Z. & Kole, R. (1991) *Mol. Cell. Biol.* **11**, 6075-6083.