(Reprinted from Nature, Vol. 277, No. 5698 pp. 627-633 February 22 1979)

Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments

Hitoshi Sakano, John H. Rogers*, Konrad Hüppi, Christine Brack, André Traunecker, Richard Maki, Randolph Wall* & Susumu Tonegawa

Basel Institute for Immunology, 487 Grenzacherstrasse, Postfach, CH-4005 Basel, Switzerland and *Molecular Biology Institute and Department of Microbiology and Immunology, School of Medicine, University of California, Los Angeles, California 90024

A 6.8-kilobase DNA fragment containing the sequence coding for the constant region of the mouse immunoglobulin γ_1 heavy chain was cloned from total cellular DNA. Electron microscopic and nucleotide sequencing studies showed that the three protein domains and the hinge region are encoded in separate DNA segments.

SOME higher cell genes consist of informational DNA interspersed with silent sequences. Eukaryotic cistrons are often transciption units containing alternate regions to be excised from the primary transcript, introns, and to be left in the mature messenger RNA for expression, exons[1]. The first indication of such a mosaic gene structure was obtained in the 28S ribosomal RNA gene of *Drosophila melanogaster*[2,3]. This observation was followed by discoveries of an intron within the RNA leader regions of the human adenovirus late genes[4,5]. Introns were also found in the protein-encoding regions of mouse immunoglobulin λ-chain genes[1,6], rabbit and mouse β-globin genes[7,8], hen ovalbumin genes[9,10], and in yeast tRNA genes[11,12]. More examples are accumulating. These studies show that introns are widespread in higher cells and their viruses, but the functions and origin of introns remain unclear.

Both light and heavy chains of immunoglobulin molecules consist of two functionally differentiated regions: the amino terminal variable (V) region and the carboxyl terminal constant (C) region. The V region recognises and binds antigens while the C region exerts various effector functions such as interaction with cell surface membrane and complement. Amino acid sequence studies have shown that both light and heavy chains consist of homology units[13,14]. For instance, the C region of a γ-class heavy chain contains three homology regions, CH1, CH2 and CH3, each of which is similar in size (about 110 residues) and homologous to the C region of a light chain[15]. Furthermore, the V regions of both light and heavy chains may also be related to the C region based on the arrangement of intrachain disulphide bonds and their size, and the presence of a weak homology between these regions[13,15]. These studies suggested two important points on the structure and evolution of immunoglobulin molecules. First, homology regions may have similar three-dimensional structures, each consisting of a compact globular domain. This hypothesis was directly confirmed by X-ray crystallography[16,17]. Second, the presence of the internal homology suggested that both light and heavy chain genes evolved by duplication of a primordial gene that codes for a polypeptide chain with a size of a single homology unit[13]. Our previous studies on λ and κ light chain genes demonstrated that the V and C domains are encoded in separate exons[6,18,19]. We here report that the three C region domains of mouse γ_1 heavy chain as well as the chain-linking hinge region between the CH1 and CH2 domains are encoded in separate DNA segments.

Table 1 Length measurements of R-loop molecules

Segment	Length (kilobases)	*N*
A	3.29 ± 0.13	48
B	1.92 ± 0.12	68
R1	0.28 ± 0.05	64
R2	0.34 ± 0.05	53
R3	0.42 ± 0.08	60
I1	0.51 ± 0.09	45
I2	0.18 ± 0.03	14
I1a	0.37 ± 0.06	29

The R-loops R1–R3 were measured in double-stranded molecules, in which the intervening sequence I_1 was stretched out or seen as an intron-loop, and I_2 was seen only in stretched out molecules (Fig. 2*b*). In the single-strended DNA–mRNA hybrids the intron I_{1a} is shortened because of the annealing of the hinge coding region. I_{1b} is not measurable, but probably is about 100 base pairs long. The hinge region must be within 140 base pairs from the 5′ end of the R2 coding region. Values are given ±s.d. *N*, numbers of measured molecules. Measurements of R-loop molecules were made with a Numonics digitiser and the data analysed with a Hewlett Packard HP9825 calculator.

Identification and cloning of an *Eco*R1 fragment carrying a $C\gamma_1$ gene

High molecular weight total cellular DNA extracted from MOPC 21 myeloma was digested to completion with *Eco*R1 and the resulting DNA fragments were analysed by the Southern gel blotting technique[20] using as the hybridisation probe a nick-translated cDNA clone (pH21-1) prepared from MOPC 21 γ_1 chain mRNA. As shown in Fig. 1*A*, we detected a single band of 6.8 kilobases that hybridised with the cDNA probe. Rogers *et al.* sequenced the γ_1 cDNA and found that it contains sequences coding for part of the $C\gamma_1$ region that extends from amino acid residue 287 to 439[21]. We conclude that the 6.8-kilobase *Eco*R1 fragment contains at least part of the $C\gamma_1$ DNA sequence.

DNA fragments in the 6.8-kilobase fraction were recovered from the preparative agarose gel and were ligated with the isolated DNA arms of an EK-2 vector, phage λgt WES[22]. The recombinant DNAs were packaged *in vitro* into phage coats[23] and plaques were screened by *in situ* hybridisation using the nick-translated $C\gamma_1$ cDNA clone as probe[24]. Screening of about 2×10^5 independently arising plaques lead to the identification of eight positive clones. When digested with *Eco*R1, DNA from the eight phage clones gave, in addition to the left and the right DNA arms of the vector, a 6.8-kilobase DNA fragment that hybridised with the $C\gamma_1$ cDNA probe. The results of one of the eight phage clones, $IgC\gamma_1$-1, are shown in Fig. 1*B*.

The $C\gamma_1$ gene is split

The position of the $C\gamma_1$-coding region in the 6.8-kilobase DNA fragment of the $IgC\gamma_1$-1 clone was determined by electron microscopy. R-loop molecules were formed between the cloned DNA and a γ_1 chain mRNA purified from MOPC 21 myeloma cells[25]. Figure 2 shows the various R-loop structures observed. Most of the molecules have three small R-loops, R1, R2 and R3, that are separated by two double-stranded intervening DNA segments (I1 and I2) (Fig. 2*a*). In all molecules a short RNA tail could be observed at the largest R-loop (R3), and in some

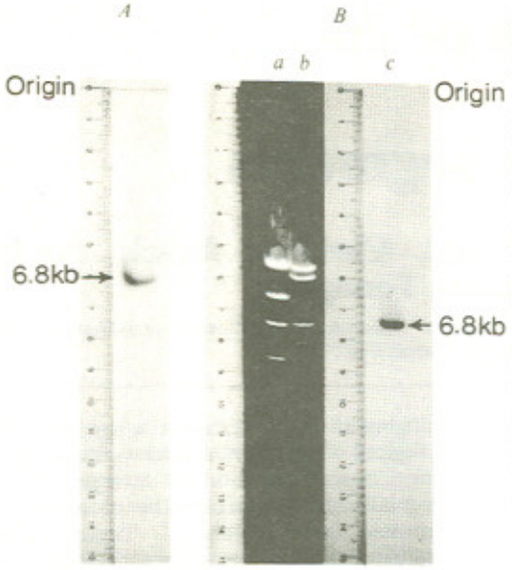

Fig. 1 Southern hybridisation of pH21-1 cDNA to myeloma MOPC 21 or cloned IgCγ_1-1 DNA. *A*, 2 mg of high molecular weight DNA extracted from myeloma MOPC 21 was digested to completion with *Eco*R1 and separated on 0.8% horizontal agarose gel slab (20×40×1 cm) in TA buffer (40 mM Tris-HCl, 5 mM Na-acetate, 1 mM EDTA, *p*H 7.9) at a constant current of 50 mA for 4 days. To detect the γ_1 gene-positive DNA fragments, a 0.5-cm wide section of the gel was cut out from the centre in parallel with the direction of electrophoresis, and was blotted onto a nitrocellulose filter (Schleicher-Schuell, BA85) for 3 h as described by Southern[20]. The filter was incubated at 65 °C for 12 h in the presence of 2×10^7 c.p.m. of denatured, nick-translated pH21-1 cDNA (specific activity 2×10^8 c.p.m. per μg) as described elsewhere[19]. The figure shows an autoradiogram of the filter after 6 h of exposure. The molecular size of the γ_1 gene-positive band was estimated from its relative electrophoretic mobility to *Hin*dIII fragments of phage λ DNA. For the cloning, DNA was recovered from the preparative slab gel as follows. Hydroxyapatite (Bio Rad, DNA grade) washed with TA buffer was mixed with an equal volume of 0.8% melted agarose (in TA buffer) and the mixture was poured into the 0.5-cm wide vertical slot in the centre of the gel slab and into a parallel slot of 0.5-cm width located near one edge of the gel. The gel slab was rotated 90° and electrophoresis was continued at 150 mA for 4 days to let DNA fragments migrate into and bind to the embedded hydroxyapatite. Hydroxyapatite containing γ_1 gene-positive DNA fragments was identified from the autoradiogram of the Southern gel blot, and a 0.5-cm thick slice of the hydroxyapatite cake in this region was cut out. The slice was suspended in 1 ml of saturated KI solution and heated at 65°C for 15 min to melt the agarose. The hydroxyapatite was collected by centrifugation and was washed with 3 ml of 10 mM KPO_4 buffer (*p*H 6.9). The DNA was eluted from the hydroxyapatite with 0.5 ml of 0.6 M KPO_4 (*p*H 6.9) buffer and the hydroxyapatite was removed by centrifugation. The DNA was dialysed against 10 mM Tris-HCl *p*H 7.9 containing 1 mM EDTA and was concentrated by ethanol precipitation. *B*, 0.5 μg of cloned Ig Cγ_1-1 DNA was digested with *Eco*RI, separated in a 1% agarose gel and stained with 0.5% ethidium bromide (lane *b*). DNAs were transferred onto nitrocellulose filter and hybridised with 2×10^6 c.p.m. of nick-translated pH 21-1 cDNA (2×10^8 c.p.m. per μg) (lane *c*). *Hin*dIII-digested phage λ DNA (0.5 μg) was separated in the same gel as a size marker (lane *a*). The third largest *Hin*dIII-λ fragment is 6.6 kilobases long.

molecules a longer RNA tail extends also from the shorter R-loop (R1).

The sites of R-loops and intervening DNA segments are summarised in Table 1. The intervening DNA segment between R2 and R3 (I2) is very short and cannot be visualised as a clear double-stranded loop, rather it appears as a little knob separating the two R-loops. The length of the two intervening DNA segments can best be measured in those hybrids where several mRNA fragments have annealed to the same DNA fragment (Fig. 2*b*, *c*). The lengths of R1 (280 base pairs) and R2 (340 base pairs) roughly correspond to the size of a DNA segment coding for a single domain (100–120 residues) of an immunoglobulin chain, while the length of R3 (420 base pairs) is slightly larger. We interpret the various parts of the observed R-loop structure in the following way. R1 and R2 correspond to the DNA segments coding for the CH1 and CH2 domains, respectively, while R3 probably encodes the CH3 domain and the presumed 3′-untranslated region of γ_1 chain mRNA. This interpretation is supported by the length and the location of the two RNA tails observed in the intact R-loop structure (Fig. 2*a*). The short (~200 bases) and the long (~400 bases) RNA tails can be interpreted to represent the poly A and the V-coding sequence at the 3′- and the 5′-end of the mRNA.

Correlation between exons and the domain- and hinge-coding DNA segments

In order to confirm the interpretation of the R-loop structure described above, we constructed a restriction enzyme cleavage map of the IgCγ_1-1 clone and determined the nucleotide sequences of the relevant regions (Fig. 3*b*, *c*).

In determining the nucleotide sequences we first concentrated on the five regions (indicated in Fig. 3*c* as I, II, IV, V, and VI) that cover the ends of the three major exons. In so doing we discovered that coding at the right end (see Fig. 3*c* for orientation) of exon 1 and the left end of exon 2 leaves about 13 amino acid residues around the hinge region (Fig. 3*d*) unaccounted for. DNA sequencing revealed an additional exon (exon H) that codes for this cysteine-rich region within the 0.5-kilobase DNA segment separating R1 from R2. Electron micrographs of single-stranded DNA-mRNA hybrids (Fig. 2*d*, *e*) are consistent with the presence of exon H. When γ_1 mRNA was hybridised to denatured IgCγ_1-1 DNA, it annealed to the hinge region, thus forming another short intron I1b that is similar in size and appearance to I2. We thus call the four protein-encoding DNA segments exon 1, H, 2 and 3, and the three non-coding intervening DNA segments intron B, C and D as indicated in Fig. 3*c*. We also call the non-coding DNA segment to the left of exon 1 intron A because we anticipate the presence of another short exon further upstream (that is, towards the left in Fig. 3) that codes for the V–C junction region (J region)[18] in analogy to the structure of a λ light chain gene.

The DNA sequences in the six regions (I–VI) indicated in Fig. 3*c* are shown in Fig. 4. Also shown in Fig. 4 are part of the amino acid sequences of MOPC 21γ_1 chain as determined by Adetugbo[26], and the nucleotide sequences predicted from the amino acid sequence. Coding in exon 1 begins with Ala at residue 121 and ends in Val at residue 215. Coding in exon H starts with the Val and ends in Val at residue 228. Coding in exon 2 starts with the Val codon and ends in Gly at residue 335. Finally, coding in exon 3 starts with the Gly residue. The entire amino acid sequence of the MOPC 21 γ_1 chain was determined by Adetugbo[26]. While the presence of the sequence homology among the three C region domains is obvious, the exact domain boundaries are ambiguous. Nevertheless, examination of the internal sequence homology in the γ_1 chain and comparison of its sequence with that of a human γ_1 chain[15], for which the three dimensional structure of the Fab fragment is available[16,17] makes it possible to draw every domain boundary within short regions that contain no more than a few amino acid residues. Thus CH1–hinge, hinge–CH2 and CH2–CH3 boundaries can be placed around residues 214–218, residues 227–228, and residues 332–335, respectively. Furthermore, the amino terminal end of the CH1 domain is around residues 115–122. The nucleotide sequences show that in all cases coding ends or begins with a residue that lies within one of these inter-domain regions deduced from structural analysis of the protein. Since it is extremely unlikely that such a correlation is a coincidence, we conclude that each of the three Cγ_1 domains and the hinge region are encoded by separate exons. Indeed one may look at it in the opposite sense and say that the exact ends of domains and the hinge region are defined by the intron–exon boundaries on the DNA.

In the coding regions the fit between the determined and the predicted nucleotide sequences is good. However, note the discrepancies in 10 codons in the regions shown in Fig. 4. Of the 10 differences, 9 are due to exchange or permutation of residues at nearby positions (for instance the Pro–Ser exchange in exon 1) and the tenth is between Ser and Thr. Thus these discrepancies are probably due to errors in determination of the amino acid sequence. Arg–Lys exchange in exon 3 was found previously[21] in the sequence of cDNA clone, pH21-1.

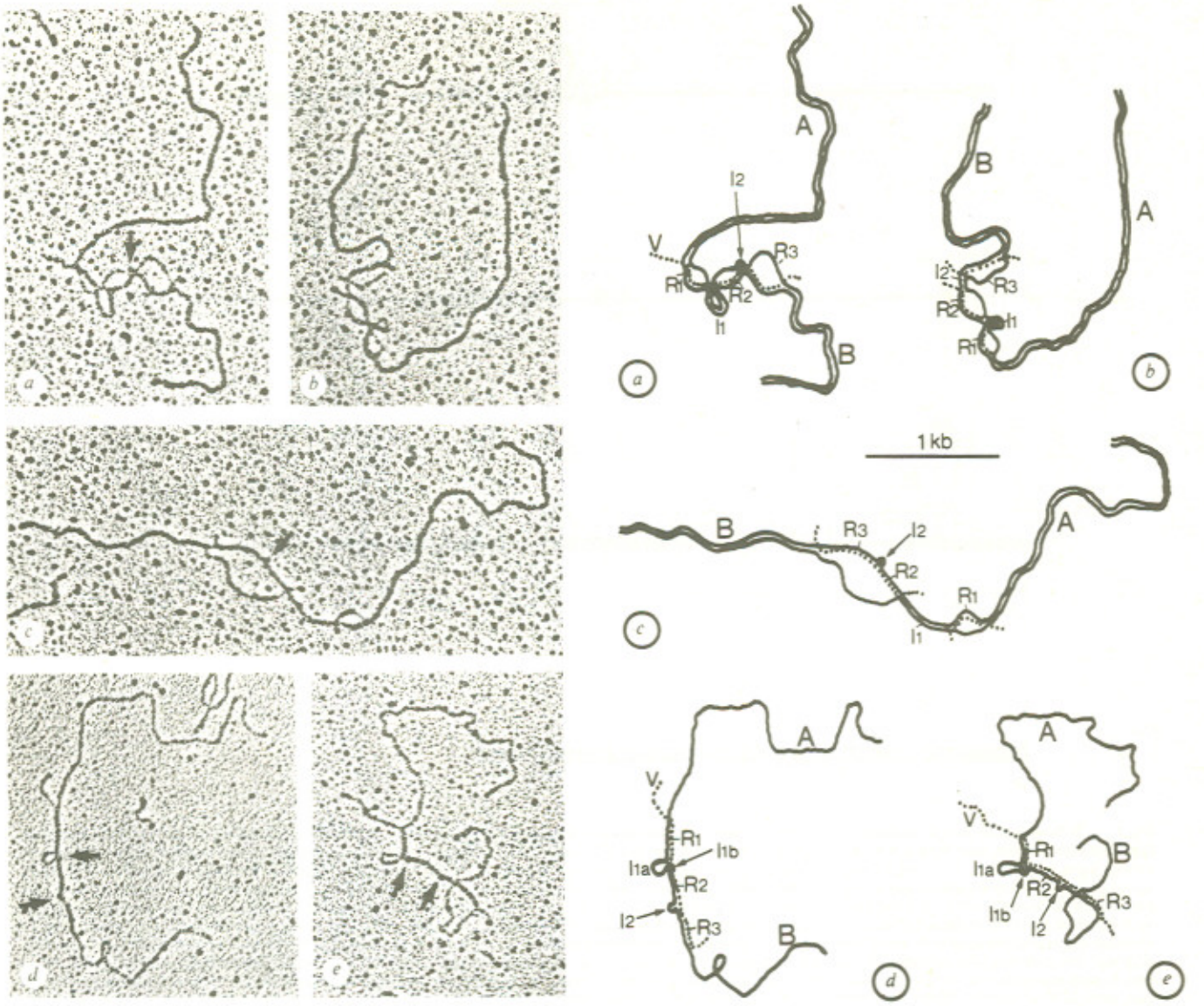

Fig. 2 Left: electron micrographs of R-loop molecules (*a–c*) and RNA–DNA hybrids (*d, e*). The *Eco*RI-digested IgCγ_1-1 DNA was incubated with MOPC21 γ_1 mRNA under R-loop conditions at 57 °C[18]. Single-stranded DNA–mRNA hybrids were formed after heat denaturation and incubation of the R-loop mixture at 60 °C. The molecules were mounted for electron microscopy by the formamide-cytochrome *c* spreading method[40,41]. Three R-loops, R_1, R_2, R_3 are separated by two intervening sequences I_1 and I_2. When two separate mRNA molecules hybridise to different R-loops, the intervening sequences are stretched out and can be measured (*b, c*). In single-stranded hybrids (*d, e*) it is possible to resolve also the small intron I_{1b} which separates the hinge coding region from the intron I_{1a}, and which appears as a small knob beside I_{1a}. Right: schematic interpretation of the hybrid molecules.

Are there any additional introns within the DNA segments coding for the three domains? The entire sequences of exon 1 and exon 2 segments have been determined (K. H. and H. S., unpublished), and no coding gaps were found in these regions. In exon 3, sequencing has not been completed, but the 220-base pair *Hae*III–*Hae*III fragment arising from exon 3 co-migrates with the equivalent fragment from the pH21-1 cDNA clone in a 6% acrylamide gel. It seems that there is no intron within domain-encoding segments.

Nucleotide sequences at the splice sites

The coding gaps in split genes are considered to be eliminated during processing of the primary RNA transcript by splicing of exon sequences[27–29]. The nucleotide sequences of the putative splice sites in the Cγ_1 gene are listed in Fig. 5 together with the splice site sequences of the λ_I and λ_{II} light chain genes which we previously determined[1,30]. As in all other known split genes, the exact points of coding interruption are ambiguous because of the redundancy near the opposite ends of the introns. Thus a single nucleotide G is repeated near the two ends of intron B, a dinucleotide AG near the ends of introns C, and a trinucleotide AGG near the two ends of intron D. By comparing sequences near the ends of six ovalbumin gene introns Breathnach *et al.*[31] found that it is always possible to choose cutting points so that the base sequence of an intron begins with GT and ends with AG. This rule applies to the two small introns of λ chain genes[1,30], the six introns of SV40[32] and the small introns of β-globin genes[33]. While intron C and intron D of the Cγ_1 chain gene also obey this rule, intron B does not. Because of the G repetition intron B either begins with GA and ends with CA or begins with AG and ends with AG. Another generalisation that can be made for the sequence near the ends of introns is an apparent presence of a basic sequence of several bases from which all known marginal sequences can be derived with no more than one or two base changes[31,34]. In particular we noted that all three introns of immunoglobulin λ chain genes have the sequence AGGTNAG near the 5′ end (relative to the direction of transcription)[30]. While introns C and D of the Cγ_1 gene have the same sequence at the analogous positions, intron B seems to have a quite different sequence. Sequence similarity near the

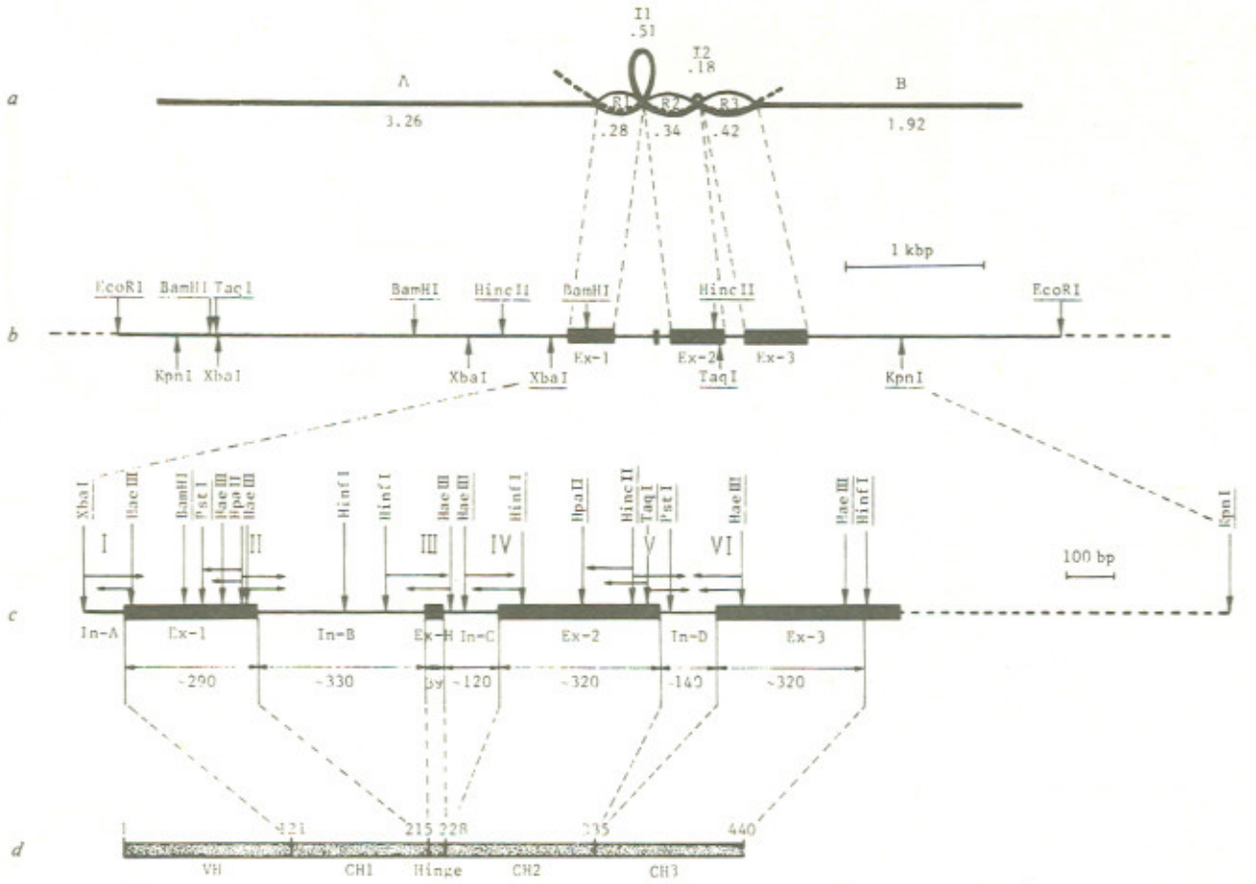

Fig. 3 *a*, R-loop map of the 6.8-kilobase insert. Heavy broken lines and heavy solid lines represent MOPC 21 γ_1 mRNA and double-stranded DNA, respectively. Duplexes between fine solid lines and broken lines are DNA–RNA hybrids. Fine solid lines are replaced single-stranded DNA. The numbers indicate lengths of various parts in kilobases. *b*, A map of the 6.8-kilobase insert obtained by the combination of restriction enzyme digestion and R-loop formation. The whole IgCγ_1-phage DNA was digested either with one of the indicated enzymes or with that enzyme plus *Eco*RI and separated in a 1% agarose gel. Comparison of the single and the double digestion patterns allowed us to localise the outmost cleavage sites of each enzyme on the 6.8-kilobase insert. Additional cleavage sites were identified by comparison of the electrophoresis patterns obtained after digestion of the isolated 6.8-kilobase insert with one enzyme or various combinations of two enzymes. To correlate the restriction map with the R-loop map, we analysed R-loops formed between the γ_1 mRNA and the insert which had been digested with respective enzymes. The cleavage sites identified this way are underlined. R1, R2 and R3 in (*a*) correspond to Ex1, Ex2 and Ex3 in (*b*) respectively. *c*, Detailed restriction enzyme map around the exons. Sites for *Hae*III and *Hin*fI were mapped in the five isolated pieces generated by *Taq* and *Bam* cleavage of the 6.8-kilobase insert. For each piece in turn, parallel digests were done with *Hae*III and *Hinf*, and the products were electrophoresed on 6% polyacrylamide. They were compared with similar digests of the whole insert. By noting the size of the fragments which were present only in the digests of the whole insert, and of the fragments generated from them in the *Bam*- and *Taq*-cut material, most of the sites around the *Bam* and *Taq* sites could be located. Within and between exon 1 and exon 2 all the remaining sites could also be identified, since there was only one way in which the fragments observed on single and double digests could be self-consistently ordered. Two small (30–40 base pairs) *Hae*III fragments had to be postulated to make up the total length of this region. One of these was confirmed, and some remaining ambiguities were resolved by labelling the 5′ ends of *Bam*-cut pieces of the 6.8-kilobase insert with [γ-^{32}P]ATP, cutting with *Hinf*, isolating the single end-labelled fragments by preparative polyacrylamide gel electrophoresis, and analysing polyacrylamide gel electrophoresis, and analysing partial digests of them by *Hae*III. Within exon 3, the 217-base pair *Hae*III fragment was identified by comparison with a *Hae*III digest of pH21-1, the cDNA clone, in which the same fragment is present. The underlined cleavage sites were confirmed by Southern gel blotting analysis using the pH21-1 cDNA probe. The lengths (in base pairs) of introns and exons are based on the DNA sequencing data as well as on the size estimation of restriction fragments. Hinge coding region (exon H) was located between exon 1 and exon 2 by DNA sequencing (see text). The roman letters indicate six regions that were sequenced. Horizontal arrows indicate actual fragments used for sequencing. In, intron; Ex, exon. *d*, Correlation of domains and hinge of MOPC 21 γ_1 heavy chain to four exons of Cγ_1-1 DNA. Numbers represent the amino acid positions at the domain-hinge or domain–domain boundaries. Numbering starts with the first NH_2-terminal asparagine of the γ_1 chain.

3′-end of introns is somewhat less, but YAG is shared by all introns of the λ and the γ_1 chain genes. In addition, as in most other introns, sequences rich in pyrimidines precede the basic sequence near the 3′ ends of Cγ_1 introns.

Deletion mutants

The correlation between exons and domain- or hinge-encoding DNA segments immediately drew our attention to the fact that mutant or variant immunoglobulin chains are often found in which an entire domain or the hinge region is deleted. Franklin and Frangioni reported several examples of so-called human heavy chain disease proteins with such characteristics[35]. In the present context, one mouse γ_1 chain mutant isolated from cultured MOPC 21 myeloma cells is of particular interest[36]. This mutant, called IF2, lacks the CH1 domain extending from residue 121 to residue 214. The deleted amino acids are exactly those that are encoded by exon 1 (Fig. 4). The so-called C_κ fragment found in myeloma MPC 11 may be another example[37]. The sequence of the amino terminal region of the C_κ fragment that had been synthesised in a cell-free translation system under the direction of purified C_κ fragment mRNA has been determined [38,39]. The sequence suggests that the C_κ fragment is synthesised in the cells as a precursor carrying a hydrophobic leader peptide at the amino terminus. In the precursor a 17-residue long leader is directly attached to the C_κ region peptide. Our previous results showed that both the leader and the V regions are encoded in separate exons in mouse λ chain genes[1,30]. If the same applies to κ chain genes, the C_κ fragment precursor may be considered as a deletion variant in which the entire peptide encoded in the V-coding exon is deleted.

Assuming that the splicing enzyme can potentially bridge two non-adjacent exons, then all deletions which begin in one intron

Fig. 4 Partial nucleotide sequences of clone IgCγ_1-1. Nucleotide sequences of the six regions, I, II, III, IV, V, and VI indicated in Fig. 3*c* are shown. For DNA sequencing, 2 mg of phage IgCγ_1-1 DNA was cleaved with *Kpn*I and *Xba*I and separated on a 5% polyacrylamide gel (20×40×0.5 cm). The 2.5-kilobase *Xba*I–*Kpn*I fragment (see Fig. 3*c*) containing the Cγ_1-coding sequences was eluted electrophoretically from the gel and washed by ethanol precipitation. The fragment was then cleaved with *Hin*fI, *Hae*III, *Hpa*II, *Taq*I or *Hin*cII, treated with bacterial alkaline phosphatase (Worthington), end-labelled with T_4-polynucleotide kinase (P-L Biochemicals) and [γ-^{32}P]ATP (Amersham), and sequenced by the method of Maxam and Gilbert[42] with a few modifications. Cleavage reaction at A and G was performed as follows (A. Maxam, personal communication). A mixture of end-labelled DNA and 1 μg of carrier calf thymus DNA was incubated in 23 μl of 87 mM pyridine formate *p*H 2 at 20 °C for 90 min. After lyophilisation, 20 μl of 0.5 M piperidine was added and the mixture was heated at 90 °C for 30 min in a sealed capillary. The sample was then lyophilised three times, dissolved in 10 μl of 0.1 M NaOH and 1 mM EDTA, and combined with 10 μl of urea–dye mixture. For G, C and T and C reactions, the conventional method was used except that Mg-acetate was omitted from dimethyl sulphate and hydrazine stop solutions. Reacted samples were separated in thin polyacrylamide gels (30×40×0.035 cm) at 1.5 kV as described by Sanger and Coulson[43]. The figure also shows partial amino acid sequences of the γ_1 chain from myeloma MOPC 21[26] as well as the nucleotide sequences predicted from the amino acid sequence. Amino acids which do not match those predicted from our nucleotide sequence are indicated by asterisks, and the predicted amino acids are presented in italics. Vertical lines indicate possible intron–exon boundaries. Cleavage sites for some restriction endonucleases are underlined. Amino acid numbering begins with the asparagine at the NH_2-terminal end of the mature γ_1 chain from MOPC 21. N: A, G, C, or T. R: A or G. Y: C or T.

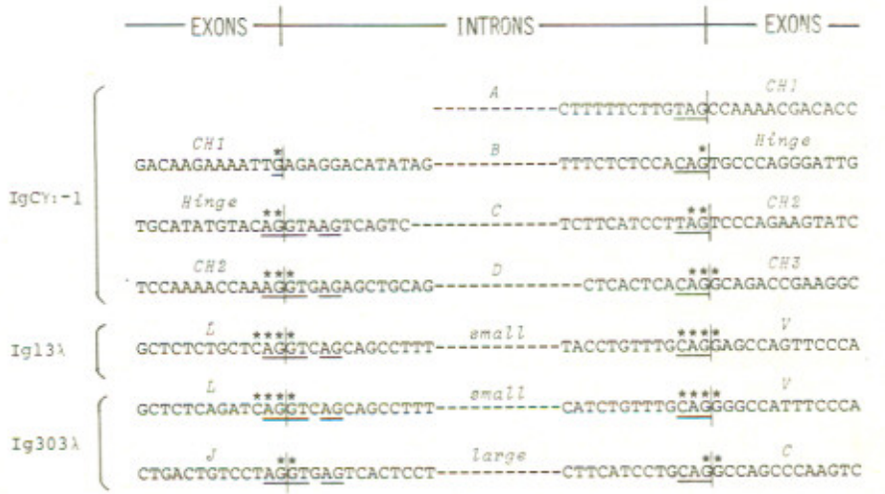

Fig. 5 Sequences around the splice sites of three immunoglobulin genes. Nucleotide sequences around the splice sites of the IgCγ_1-1 clone as well as those of Igl3λ^1 (contains an embryonic $V_{\lambda II}$ gene) and Ig303λ^{30} (contains a myeloma (V+C) λ_I gene) are summarised. Intron–exon boundaries defined by the GT–AG rule[31] are indicated by vertical lines. The short 'common' sequences near the 5'- or 3'-end of introns are underlined. In case of intron B which does not follow the GT–AG rule, the boundaries indicated are based on the assumption that the intron ends with AG (see text). Nucleotides repeated near the opposite ends of an intron are marked with asterisks. CH1, CH2 and CH3 designate the three constant region domains of the γ_1 heavy chain. L, V, J, and C designate the hydrophobic leader, variable, joint, and constant region of the λ_I or λ_{II} light chain.

and end in another could generate a single mutant protein in which the entire polypeptides encoded in exons between the two introns are deleted. Alternatively, a mutation at or near a splice site, whether it is a deletion or a base change, could perturb the splicing pattern such that the entire sequence in one exon is skipped in the mRNA. For instance, certain mutations near the intron B–exon H boundary may block splicing of exon 1 with exon H and in turn may promote its splicing with exon 2. Cloning and analysis of the relevant DNA fragments from mutant cells will test these models directly.

Evolution of immunoglobulin genes and origin of introns

How splicing arose in evolution is unknown. Crick has suggested that splicing was originally a protection mechanism, evolved by the host organism to reduce the damage being inflicted by the insertion of nonsense sequences in the middle of the gene (F. Crick, personal communication). In this hypothesis, the introns are viewed as DNA elements that were once inserted more or less randomly within a pre-existing gene. The placement of introns in ovalbumin and β-globin genes indeed seems to be random. However, the presence of introns at the regularly spaced boundaries of domain-coding DNA segments and at the boundaries of the domain- and hinge-encoding DNA segment is in apparent contrast to such a 'random' insertion of DNA elements. To explain the regularity observed in the location of the introns we propose the following scheme as a possible and likely process of the evolution of an immunoglobulin heavy chain gene. It is likely that when the putative primordial gene duplicated, the unit of duplication was substantially larger than the unit of translation. As DNA duplication is likely to involve some form of illegitimate recombination which makes use of accidental short sequence homology, there is no *a priori* reason why a domain coding DNA segment would be duplicated immediately adjacent to itself. Most probably, the duplicated DNA would contain a spacer between the two domain-encoding DNA segments. One obvious way to fuse the DNA segments is to delete the spacer. But, for exactly the same reason as given above for the duplication step, such a deletion is probably a rare event. Instead, the organism might very well have resorted to a pre-existing splicing mechanism that might have developed under an evolutionary pressure such as that suggested by Crick.

How did the necessary splice sites arise at or near the ends of the domain-encoding DNA segments? One possibility is the primordial domain gene was itself an intact exon bound by splice sites. Alternatively, splice sites might have been created *de novo* by drift at the proper regions after duplication. In most cases, there seems to be a basic sequence of four to six bases to which sequences near the margins of introns are related in varying degrees. While it may not be difficult to imagine that a proper sequence of this length is generated rather frequently by drift, the specificity provided by these sequences cannot be sufficient. The same sequence is also found within exons[34]. The obvious possibility is that combination of some higher order structure and the sequence information determine the specificity.

To generate the four-domain heavy chain gene the process outlined above is imagined to have been repeated. An additional aspect of the evolution of heavy chain genes is the generation of the hinge-coding exon. While it is possible to imagine that this exon arose from part of the ancestral intron between exon 1 and exon 2, there are two more interesting possibilities. One is that the hinge evolved by reduction of an extra C region domain corresponding to the CH2 domain which is found in place of hinge in μ and ε chains[44]. The other is the DNA segment containing the hinge-coding sequence with or without ready-made splice sites at its margins was inserted into the ancestral intron between the exon 1 and exon 2.

Whichever variation of the model described above is more likely to be correct, it suggests that the splicing mechanism, with or without recombination at the DNA level, enables an organism to create a new gene by assembling DNA pieces each coding for a polypeptide chain of some functional use[1]. If such a gene creation process indeed took place in evolution, some of the present day introns should be regarded as spacers between assembled exons and not as insertion sequences *per se*. As suggested by the hinge-coding exon and the hydrophobic leader-coding exon[1,30], the correlation between a clearly defined protein domain and an exon is not a necessary element of the proposed gene creation process. A DNA segment that codes for a protein of some function and flanked by splice sites, whether they arose from the margins of insertion elements or by mutational drift, can combine with another such segment coding for a protein of a similar or an entirely different function. The function could be for binding with a specific ligand, for providing disulphide bridges or even for simply providing a peptide that can stabilise the overall structure of the newly created protein.

We thank G. Dastoornikoo and Lorna Gibson for technical assistance and M. Komaromy and J. Strommer for advice and materials. R.W. and J.H.R. were supported by NIH grants AI-13410 and CA-12800.

Received 27 December 1978; accepted 17 January 1979.

1. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74,** 3518–3522 (1978).
2. Glover, D. M. & Hogness, D. S. *Cell* **10,** 167–176 (1977).
3. Wellauer, P. K. & Dawid, I. B. *Cell* **10,** 193–212 (1977).
4. Berget, S. M., Moore, C. & Sharp, P. A. *Proc. natn. Acad. Sci. U.S.A.* **74,** 3171–3175 (1977).
5. Chow, L. T., Gelinas, R. E., Brocker, T. R. & Roberts, R. J. *Cell* **12,** 1–8 (1977).
6. Brack, C. & Tonegawa, S. *Proc. natn. Acad. Sci. U.S.A.* **74,** 5652–5656 (1977).
7. Jeffreys, A. J. & Flavell, R. A. *Cell* **12,** 1097–1108 (1977).
8. Tilghman, S. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **75,** 725–729 (1978).
9. Breathnach, R., Mandel, J. L. & Chambon, P. *Nature* **270,** 314–319 (1977).
10. Doel, M. T., Houghton, M., Cook, E. A. & Garey, N. H. *Nucleic Acids Res.* **4,** 3701–3713 (1977).
11. Goodman, H. M., Olson, M. V. & Hall, B. D. *Proc. natn. Acad. Sci. U.S.A.* **74,** 5453–5457 (1977).
12. Valenzuella, P., Venegas, A., Weinberg, F., Bishop, R. & Rutter, W. J. *Proc. natn. Acad. Sci. U.S.A.* **75,** 190–194 (1978).
13. Hill, R. L., Delaney, R., Fellows, R. E. & Lebovitz, H. E. *Proc. natn. Acad. Sci. U.S.A.* **56,** 1762–1768 (1966).
14. Singer, S. J. & Doolittle, R. F. *Science* **153,** 13–17 (1966).
15. Edelman, G. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **63,** 78–85 (1969).
16. Poljak, R. *et al. Proc. natn. Acad. Sci. U.S.A.* **70,** 3305–3310 (1973).
17. Padlan, E. A. *et al. Nature new Biol.* **245,** 165–167 (1973).
18. Brack, C., Hirama, M., Lenhard-Schuller, R. & Tonegawa, S. *Cell* **15,** 1–14 (1978).
19. Lenhard-Schuller, R., Hohn, B., Brack, C., Hirama, M. & Tonegawa, S. *Proc. natn. Acad. Sci. U.S.A.* **75,** 4709–4713 (1978).
20. Southern, E. M. *J. molec. Biol.* **98,** 503–517 (1975).
21. Rogers, J. H., Clark, P. & Salser, W. (submitted).
22. Tiemeier, D. C., Enquist, L. & Leder, P. *Nature* **263,** 526–528 (1976).
23. Hohn, B. & Murray, K. *Proc. natn. Acad. Sci. U.S.A.* **74,** 3259–3263 (1977).
24. Benton, W. D. & Davis, R. W. *Science* **196,** 180–192 (1977).
25. White, R. L. & Hogness, D. S. *Cell* **10,** 177–182 (1977).
26. Adetugbo, K. *J. biol. Chem.* **253,** 6068–6075 (1978).
27. Tilghman, S. M., Curtis, P. J., Tiemeier, D. C., Leder, P. & Weissmann, C. *Proc. natn. Acad. Sci. U.S.A.* **75,** 1309–1313 (1978).
28. Knapp, G., Beckmann, J., Johnson, P. F., Fuhrman, S. A. & Abelson, J. *Cell* **14,** 221–236 (1978).
29. O'Farrell, P. Z., Cordell, B., Valenzuela, P., Rutter, W. J. & Goodman, H. M. *Nature* **274,** 438–445 (1978).
30. Bernard, O., Hozumi, N. & Tonegawa, S. *Cell* **15,** 1133–1144 (1978).
31. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. *Proc. natn. Acad. Sci. U.S.A.* **75,** 4853–4857 (1978).
32. Reddy, V. B. *et al. Science* **200,** 494–502 (1978).
33. Van den Berg, J. *et al. Nature* **276,** 37–44 (1978).
34. Catterall, J. F. *et al. Nature* **275,** 510–514 (1978).
35. Franklin, E. C. & Frangione *Contemp. Topics Molec. Immun.* **4,** 89–126 (1975).
36. Adetugbo, K., Milstein, C. & Becker, D. S. *Nature* **265,** 299–304 (1977).
37. Kuehl, W. M. & Sharff, M. D. *J. molec. Biol.* **89,** 409–421 (1974).
38. Burstein, Y. & Schechter, I. *Biochemistry* **17,** 2392–2400 (1978).
39. Rose, S. M., Kuehl, M. & Smith, G. P. *Cell* **12,** 453–462 (1977).
40. Davis, R. W., Simon, M. & Davidson, N. *Meth. Enzym.* **21D,** 413–428 (1971).
41. Westmorland, B. C., Szybalski, W. & Ris, W. *Science* **163,** 1343–1346 (1969).
42. Maxam, A. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74,** 560–564 (1977).
43. Sanger, F. & Coulson, A. R. *FEBS Lett.* **87,** 107–116 (1978).
44. Bennich, H. & von Bahr Lindsm, H. in *Progress in Immunology II,* Vol. 1 (eds Brent, L. & Holborow, J.) 49–58 (North-Holland, Amsterdam, 1974).

Printed in Great Britain by Macmillan Production Ltd., Basingstoke, Hampshire