An Immunoglobulin Heavy Chain Variable Region Gene Is Generated from Three Segments of DNA: V\textsubscript{H}, D and J\textsubscript{H}

P. Early, H. Huang, M. Davis, K. Calame and L. Hood
Division of Biology
California Institute of Technology
Pasadena, California 91125

Summary

We have determined the sequences of separate germline genetic elements which encode two parts of a mouse immunoglobulin heavy chain variable region. These elements, termed gene segments, are heavy chain counterparts of the variable (V) and joining (J) gene segments of immunoglobulin light chains. The V\textsubscript{H} gene segment encodes amino acids 1-101 and the J\textsubscript{H} gene segment encodes amino acids 107-123 of the S107 phosphorolcholine-binding V\textsubscript{H} region. This J\textsubscript{H} gene segment and two other J\textsubscript{H} gene segments are located 5' to the \( \mu \) constant region gene (C\textsubscript{\mu}) in germline DNA. We have also determined the sequence of a rearranged V\textsubscript{H} gene encoding a complete V\textsubscript{H} region, M603, which is closely related to S107. In addition, we have partially determined the V\textsubscript{H} coding sequences of the S107 and M167 heavy chain mRNAs. By comparing these sequences to the germline gene segments, we conclude that the germline V\textsubscript{H} and J\textsubscript{H} gene segments do not contain at least 13 nucleotides which are present in the rearranged V\textsubscript{H} genes. In S107, these nucleotides encode amino acids 102-106, which form part of the third hypervariable region and consequently influence the antigen-binding specificity of the immunoglobulin molecule. This portion of the variable region may be encoded by a separate germline gene segment which can be joined to the V\textsubscript{H} and J\textsubscript{H} gene segments. We term this postulated genetic element the D gene segment, referring to its role in the generation of heavy chain diversity. Essentially the same noncoding sequences are found 3' to the V\textsubscript{H} gene segment and as inverse complements 5' to two J\textsubscript{H} gene segments. These are the same conserved nucleotides previously found adjacent to light chain V and J gene segments. Each conserved sequence consists of blocks of seven and ten conserved nucleotides which are separated by a spacer of either 11 or 22 nonconserved nucleotides. The highly conserved spacing, corresponding to one or two turns of the DNA helix, maintains precise spatial orientations between blocks of conserved nucleotides. Gene segments which can join to one another (V and J, for example) always have spacers of different lengths. Based on these observations, we propose a model for variable region gene rearrangement mediated by proteins which recognize the same conserved sequences adjacent to both light and heavy chain immunoglobulin gene segments.

Introduction

The immunoglobulin molecule is a complex entity with two major functions—recognition of foreign substances (antigen binding) and the elimination or destruction of these foreign substances (effector functions). This bipartite nature is reflected in the structure of immunoglobulins, which are composed of polypeptide chains with many alternative sequences of amino acids near their N terminal ends (the variable or V region), but only a few possible sequences for the remainder of the chain (the constant or C region). Typical immunoglobulin molecules contain two identical heavy chains and two identical light chains. Pairs of light and heavy chains fold into discrete structural domains—the V region domain which binds antigens, and the C region domains responsible for the effector functions (Gally, 1973). V region domains consist of approximately 107 amino acids from a light chain and approximately 125 amino acids from a heavy chain. Within the V region domain, three short polypeptide loops from both the heavy and light chains form the antigen-binding pocket (Amzel et al., 1974; Segal et al., 1974). These short portions of the V region are termed the hypervariable regions, since extensive protein sequence studies of immunoglobulins have shown them to be the most frequent sites for amino acid substitutions between V regions (Wu and Kabat, 1970; Copra and Kohoo, 1974).

The heavy chains and the two types of light chains (\( \kappa \) and \( \lambda \)) which occur in vertebrate immunoglobulin molecules are each encoded by a separate multigene family (Hood, Campbell and Elgin, 1975). The functionally distinct V and C regions of immunoglobulin polypeptides are encoded by independent genetic elements within each gene family (Dreyer and Bennett, 1965). In the \( \kappa \) light chain gene family of mouse, which has been the most extensively studied, one germline gene encodes the C region. V regions are encoded by genes which are created during differentiation by the fusion of two independent elements, V and J, which we will call gene segments. A large number of germline V\textsubscript{\kappa} gene segments encoding amino acids 1-95 can join to one of four germline J\textsubscript{\kappa} gene segments encoding amino acids 96-107 of the V region (Max, Seidman and Leder, 1979; Sakano et al., 1979). This process presumably occurs by deletion of the RNA originally separating V\textsubscript{\kappa} and J\textsubscript{\kappa} gene segments (Sakano et al., 1979), and results in the formation of a \( \kappa \) gene which encodes all 107 amino acids of a \( \kappa \) region as an uninterrupted nucleotide sequence. The J\textsubscript{\kappa} gene segments are closely linked to the C\textsubscript{\kappa} gene, so V-J joining results in a single transcription unit containing both the C\textsubscript{\kappa} gene and the newly formed V\textsubscript{\kappa} gene. It has been postulated that conserved blocks of noncoding nucleotides found 3' to undifferentiated V\textsubscript{\kappa} gene segments and as inverse complements 5' to undifferentiated J\textsubscript{\kappa} gene segments play a role in the mechanism
of V-J joining (Max et al., 1979; Sakano et al., 1979).

Our studies have focused on the heavy chain gene family of mouse. In contrast to the \( \kappa \) gene family, the heavy chain family contains at least eight C region genes—\( \mu, \delta, \gamma_{m}, \gamma_{ho}, \gamma_{s}, \alpha, \) and \( \epsilon \). These \( \epsilon \) genes form a tightly linked cluster (Mage et al., 1973) which is linked to various \( \gamma_{h} \) region markers (Riblet, 1977). The heavy chain genes also undergo DNA rearrangements which bring a \( \gamma_{h} \) and a \( \epsilon \) gene close together to form a single transcription unit (Early et al., 1979). During B cell differentiation, the first constant region expressed is \( \gamma_{h} \), leading to the synthesis of an IgM molecule. Subsequently, other DNA rearrangements can produce combinations of different \( \gamma_{h} \) genes with the same \( \gamma_{h} \) gene (Davis et al., 1980) and thereby lead to the expression of other classes of immunoglobulins.

In this paper we examine the nature of the DNA rearrangements which result in the formation of a complete \( \gamma_{h} \) gene. The \( \gamma_{h} \) genes that we have chosen to study encode \( \gamma_{h} \) regions from immunoglobulins which bind phosphorylcholine. The amino acid sequences of nine \( \gamma_{h} \) regions derived from myeloma proteins binding phosphorylcholine have been determined (Figure 1). Four of the \( \gamma_{h} \) regions are identical to the S107 sequence. The remaining variants differ from the S107 sequence by 1–13 residues and in certain cases by insertions or deletions. Some of the variations correlate with differences in binding constants for phosphorylcholine and related antigens (Padlan et al., 1976; Goetzle and Richards, 1977). Many of the amino acid substitutions and all of the deletions and insertions occur in the third hypervariable region, but amino acid substitutions also occur in the second hypervariable region and outside hypervariable regions. Since the S107 \( \gamma_{h} \) region has been sequenced from four independent myeloma proteins (S107, T15, S63 and Y5236), it appears to be encoded by a germline gene.

The mouse genome contains 8–9 \( \gamma_{h} \) gene segments homologous to a cloned cDNA probe for the phosphorylcholine-binding \( \gamma_{h} \) regions (Davis et al., 1979). We have determined the nucleotide sequence of one of these germline \( \gamma_{h} \) gene segments, which encodes the S107 \( \gamma_{h} \) region. We also have determined the sequence of a germline \( \gamma_{h} \) gene segment which is expressed in most phosphorylcholine-binding \( \gamma_{h} \) regions. These gene segments are compared with a rearranged \( \gamma_{h} \) gene encoding the complete M603 phosphorylcholine-binding heavy chain, and with sequences from mRNAs encoding two other phosphorylcholine-binding heavy chains. We conclude that \( \gamma_{h} \) genes are created during B cell differentiation by joining three DNA segments: \( \gamma_{h}, D \) and \( \gamma_{h} \). The mechanism for DNA joining is probably the same in heavy and light chains, and involves protein recognition of blocks of conserved nucleotides adjacent to germline gene segments.

Results

Clones Containing Heavy Chain Gene Segments

Genomic clones were isolated from Charon 4A phage libraries (Maniatis et al., 1978) by the Benton and Davis (1977) filter screening procedure using \(^{32}P\)-labeled cloned cDNAs (Early et al., 1979; Calame et al., 1980; Davis et al., 1980). Phage in these libraries contain large (12–20 kb) inserts of either BALB/c mouse sperm DNA (germline) or DNA from M603 myeloma tumors (for details see Early et al., 1979 and Davis et al., 1980). Figure 2 depicts three genomic clones containing genes or gene segments which encode \( \gamma_{h} \) regions from myeloma proteins binding phosphorylcholine (these will be termed PC genes or gene segments). Homologies between these clones were partly determined by electron microscopy of heteroduplexed DNAs and by restriction mapping (Davis et al., 1980).

The ChSpVpc3 clone, isolated from a sperm DNA library, contains a germline PC \( \gamma_{h} \) gene segment and will be referred to as the germline PC \( \gamma_{h} \) clone. The ChSpv27 clone, also derived from a sperm library, contains the \( \gamma_{h} \) gene. As will be shown later, ChSpv27 includes a PC \( \gamma_{h} \) gene segment located 6 kb 5' to the \( \gamma_{h} \) gene. ChSpv27 will be called the germline \( \gamma_{h} \) clone. Southern blot comparisons between sperm DNA and ChSpv27 show that the clone has deleted approximately 2 kb of mouse DNA (Davis et al., 1980). The deletion is located about 1 kb 5' to the \( \gamma_{h} \) gene (M. Davis, unpublished observations). In the mouse genome, therefore, the PC \( \gamma_{h} \) gene segment is approximately 0 kb 5' to the \( \gamma_{h} \) gene. The Ch003a0 clone was isolated from an M603 myeloma DNA library and encodes both the M603 \( \gamma_{h} \) region and the \( \gamma_{h} \) region.

![Figure 1. The Phosphorylcholine-Binding (PC) Group of VH Regions](image)
Variable Region Gene Rearrangement

983

Germline VH
ChSpVpc3

Germline JH
ChSpμ27

M603 VHCg
Ch603a6/
Ch603a125

Figure 2. Genomic Clones and Sequencing Strategies

Homology between the myeloma and germline clones (indicated by shading) was determined by electron microscopy of heteroduplexes and by restriction site comparisons (Davis et al., 1980) as well as by nucleotide sequences. The raised boxes are coding sequences, with the 5' ends to the left. ChSpVpc3 (germline V<sub>H</sub>) and ChSpμ27 (germline J<sub>H</sub>) were isolated from a germline (sperm) library. The map of Ch603a6/Ch603a125 is a composite of two overlapping clones from an M603 myeloma library. Ch603a6 contains the rightward three Eco RI fragments, and Ch603a125 the leftward three fragments. In the text, this composite is referred to as “Ch603c<sub>Y6</sub>” or the M603 V<sub>H</sub> clone. The expanded maps under each clone show the locations of coding sequences more exactly. The direction and extent of sequence determinations are shown by arrows. The expanded map around the J<sub>107</sub> gene segment uses crosshatches to show the region of overlap between restriction fragments which hybridize to a PC J<sub>H</sub> probe (Figure 4). The J<sub>107</sub> gene segment is the only J<sub>H</sub> gene segment within these bounds. The numbers above the line indicate distances of the restriction sites from the Hha I site in J<sub>107</sub>. 

3.6kb

6.2kb

220 30 60 90 120 150 180 210

11.3kb

5kb

446.2 kb
This clone contains the rearranged \( V_{H} \) gene presumably expressed in the M603 myeloma tumor (Early et al., 1979). Ch603a6 will be referred to as the M603 \( V_{H} \) clone.

We used the method of Maxam and Gilbert (1977) to determine the sequences of \( V_{H} \) and \( J_{H} \) gene segments in these three genomic clones. The sequencing strategy employed for each clone is outlined in Figure 2.

The Germline PC \( V_{H} \) Gene Segment Encodes 101 Amino Acids of the \( V_{H} \) Region

Figure 3 depicts the sequence of the \( V_{H} \) gene segment in the germline PC \( V_{H} \) clone. This gene segment, denoted \( V_{H07} \), encodes the first 101 amino acids of the PC heavy chains from S107, T15, S63, Y5236 and H8 myelomas (Figure 1).

The coding sequence of \( V_{H07} \) ends with the first amino acid of the third hypervariable region. The first 75 nucleotides to the 3' side of this codon do not encode any part of a phosphorylcholine-binding \( V_{H} \) region, and include a termination codon near the end of the \( V_{H} \) gene segment. The remainder of the \( V_{H} \) region must be encoded elsewhere in the germline genome. The noncoding sequence which follows \( V_{H07} \) includes the heptanucleotide CACAGTG and the decanucleotide GACACAAACC (Figure 3), both of which resemble conserved nucleotide sequences 3' to light chain \( V \) gene segments (Max et al., 1979; Sakano et al., 1979; see Table 1). These nucleotides may play a role in DNA rearrangements of immunoglobulin gene segments (see Discussion).

A Hydrophobic Signal Peptide Is Encoded with the \( V_{H} \) Region

Immunoglobulin polypeptides are translated with a hydrophobic amino terminal "signal peptide" which is subsequently removed to generate the mature immunoglobulin chain (Milstein et al., 1972). Such peptides appear to play a role in the mechanism of protein secretion (Blobel and Dobberstein, 1975). The signal peptide encoded by S107 \( \alpha \) mRNA is shown in Figure 3. This figure depicts the nucleotide sequence for part of p107aR5, a cDNA plasmid derived from S107.
Variable Region Gene Rearrangement

mRNA (Early et al., 1979). Comparison with the germ-line VH107 sequence in Figure 3 shows that intervening DNA interrupts the signal peptide codons at an AGGT sequence 12 nucleotides from the beginning of the mature VH region. An RNA splice site at this position is also found in light chain genes (Bernard, Hozumi and Tonegawa, 1978; Seidman, Max and Leder, 1979). We have not determined the germ-line location of the mRNA sequence 5' to this splice point.

The PC JH Gene Segment Is Linked to the C, Gene in Germline DNA

In the light chain gene families, J gene segments are closely linked to C genes (Bernard et al., 1978; Max et al., 1979; Sakano et al., 1979). Since C, is the first heavy chain constant region to be expressed by lymphocytes (Raff, 1976), we reasoned that JH gene segments would be adjacent to the C gene. This gene organization would allow DNA rearrangements to form a complete VH gene initially associated with the C, gene. Subsequently, other DNA rearrangements could replace the C, gene with different CH genes, leading to the expression of other immunoglobulin classes with the same VH region (Davis et al., 1980).

To locate any PC JH gene segments in the “germline JH” clone, which we knew contained a C, gene (Cameron et al., 1980), restriction fragments were hybridized to a 32P-labeled cDNA plasmid encoding the entire S107 heavy chain. Figure 4 shows a blot (Southern, 1975) of DNA from the germline JH clone digested with three different restriction enzymes and separated on a polyacrylamide gel. Paired lanes contained digests either of the whole cloned DNA or of a 3.6 kb Hha I fragment (see Figure 2). Each lane shows only a single band of hybridization to the S107 cDNA plasmid. For each restriction enzyme, the band produced from the 3.6 kb Hha I fragment is smaller than the band from whole DNA. This result indicates that the region of the germline JH clone hybridizing to the S107 cDNA plasmid includes one of the Hha I ends of the 3.6 kb restriction fragment. In other experiments (P. Early, unpublished results), the S107 cDNA plasmid only hybridized to a 6.2 kb Eco RI fragment of this clone. This Eco RI fragment includes the 5', but not the 3', end of the 3.6 kb Hha I fragment (see Figure 2).

These data allow the region of the germline JH clone containing the PC JH gene segment to be localized. Since only one band from each restriction digest of whole DNA hybridized to the S107 cDNA plasmid, the region of hybridization must be shared by the three different restriction enzymes. Accordingly, any possible PC JH gene segments must be limited to a sequence of 220 nucleotides between the Hinf I and Hae III sites shown in Figure 2. This sequence is at the 5' end of a region of homology between the germline JH and the M603 VH clone (Figure 2).

Germline JH Gene Segments Encode Part of VH Regions

Figure 5 shows the sequence of 530 nucleotides in the germline JH clone, including the 220 nucleotide Hinf I/Hae III fragment which hybridizes with the S107 cDNA plasmid. There is a single PC JH gene segment in this fragment. This germline gene segment, denoted JH107, encodes amino acids 107–123 of the S107 VH region, and terminates at the 3' end of the VH region with an AGGT sequence associated with RNA splicing (Breathnach et al., 1978; Catterall et al., 1978). Neither the VH107 nor the JH107 gene segments encode amino acids 102–106 for any PC heavy chain. These amino acids comprise a major portion of the third hypervariable region and thus partly determine antigen-binding specificity (Padlan et al., 1976; Goetze and Richards, 1977). We will discuss the origin of the nucleotides encoding this portion of the VH region below.

A second JH gene segment, denoted JH151, is located 270 nucleotides 3' from JH107 (Figure 5). This JH gene segment does not encode part of any PC VH region and is two codons shorter than JH107. These two JH gene segments differ at 11 of 45 nucleotides.
These nucleotides apparently originate from a third germline gene segment (P. Early, unpublished results) (Figure 5). As in light chains (Bernard et al., 1978; Sakano et al., 1979), these noncoding nucleotide sequences are approximately inverse complements of the nucleotides 3' to VHIO, (see Table 1). The amino acid sequences encoded by JH15 has been found in other heavy chain VH regions, including M315, which binds dinitrophenol (Francis et al., 1974), and some VH regions from myeloma proteins which bind α,β dextran (Schilling et al., 1980). Diversity in heavy chains from myeloma proteins binding phosphorylcholine are compared in Figure 7. The data indicate that the germline JH15 gene segment is at the 5' end of a 5 kb region of DNA 500-1400 bp 3' to the JH15 gene segment (P. Early, unpublished results).

Both JH15 and JH315 are preceded by noncoding nucleotide sequences similar or identical to nucleotide sequences 5' to the light chain J gene segments (Figure 5). The amino acids encoded by the JH15 gene are numbered from the beginning of the mature S107 heavy chain. The Hinfl and Hae III sites mark the boundaries of the PC J1 gene segment (see Discussion). We have shown by heteroduplex and restriction analyses (Davis et al., 1980) that the germline JH15 gene segment is at the 5' end of a 5 kb region of homology between the germline J1 clone and the M603 VH region (Figure 2). The rearranged M603 VH gene does not include the first tyrosine codon in JH15, but the last 16 codons of the M603 gene are identical to the remainder of the JH15 sequence. The absence of the tyrosine codon in M603 may be an example of the sort of junctional variation also seen in light chain V-J joining (Weigert et al., 1978, 1980; Max et al., 1979; Sakano et al., 1979).

Sequences of Three PC mRNAs Show Evidence for Two Sites of DNA Joining in VH Genes

The partial sequences of three mRNAs encoding heavy chains from myeloma proteins binding phosphorylcholine are compared in Figure 7. The data include the codons for amino acids 95-123 of the S107 heavy chain and the homologous portions of the M603 and M167 heavy chains. These sequences include the third hypervariable regions, which differ in size for each case (Figure 1). The mRNA sequences were determined by extension of a cloned cDNA primer in the presence of dideoxynucleotide triphosphates (Sanger, Nicklen and Coulson, 1977).

The deletions and insertions in these mRNAs are relatively homologous to the S107 sequence. The scattered single nucleotide differences between these three mRNAs may reflect their distinct origins from different germline gene segments. The M603 VH region almost certainly is derived from an-
other germine \( V_h \) gene segment, since the remainder of this \( V_h \) region differs extensively from S107 (Figure 1). Alternatively, some of the differences between these mRNAs might have arisen from somatic mutation during B cell differentiation prior to neoplastic transformations, or during growth and propagation of the myeloma tumors. The \( J_h \) sequences of these three mRNAs are identical except in the third hypervariable region. All could be derived from the \( J_{H107} \) gene segment if DNA joining occurred at slightly different positions in each case (Figure 7). The \( J_{H107} \) gene segment may be the only \( J_h \) gene segment used in the PC heavy chains analyzed to date (compare the protein sequences in Figure 1). There are no other closely related \( J_h \) gene segments in the germline.

Discussion

Germline Origins of Rearranged \( V_h \) Genes

We have determined the DNA sequence of two germine gene segments, \( V_{H107} \) and \( J_{H107} \), which together encode most of the variable region of the S107 heavy chain. The \( V_{H107} \) gene segment contains the codons for the first 101 amino acids of the mature heavy chain, including the first and second hypervariable regions. The \( J_{H107} \) gene segment is approximately 8 kb to the 5' side of the \( C_h \) gene in germline DNA, while the \( V_{H107} \) gene segment is not closely associated with the \( C_h \) or \( C_{H3} \) genes in the germline.

We also have determined the sequence of one rearranged PC \( V_h \) gene from the M603 myeloma tumor. As shown in Figures 6 and 7, the M603 V gene includes the last 16 codons of the \( J_{H107} \) gene segment. The first 300 nucleotides to the 3' side of the \( J_{H107} \) sequence show only three substitutions between the germline \( J_h \) clone and the M603 \( V_h \) clone (Figure 6). We also have determined the sequence of one rearranged PC \( V_h \) gene from the M603 myeloma tumor. As shown in Figures 6 and 7, the M603 V gene includes the last 16 codons of the \( J_{H107} \) gene segment. The first 300 nucleotides to the 3' side of the \( J_{H107} \) sequence show only three substitutions between the germline \( J_h \) clone and the M603 \( V_h \) clone (Figure 6). These differences could be due to polymorphism among BALB/c mice or to mutation in the myeloma tumors. The region of homology between the myeloma M603 and the germline \( J_{H107} \) clone extends 5 kb from the 3' side of the \( J_{H107} \) gene segment (Figure 8). There are no other \( J_h \) gene segments closely related to \( J_{H107} \) in this region of homology or elsewhere in the germline. The boundaries between the gene segments probably differ somewhat for each mRNA, but cannot be precisely located from these data. The sequence of the M603 gene also was determined from the M603 \( V_h \) clone. For comparison, portions of the germline \( V_{H107} \) and \( J_{H107} \) gene segments are shown beneath, with the conserved noncoding sequences in boxes.
on the 3' side of the JH sequence hybridizes strongly only to the Cg band when germline DNA is digested with either of two restriction endonucleases (Davis et al., 1980). These results indicate that the JH107 gene segment is the unique germline JH precursor for the rearranged M603 Vh gene. The JH107 gene segment is also probably the precursor for the other known PC variable regions, as suggested by mRNA (Figure 7) and protein (Figure 1) sequences.

The germline Vh107 gene segment is not completely identical to the corresponding portion of the M603 Vh gene. There are seven differences in 303 coding nucleotides, creating three amino acid substitutions (Figure 6). The germline Vh107 clone and the rearranged M603 Vh clone, however, share 11.3 kb of homology 5' to the Vh sequences (Figure 2). If another germline Vh gene segment is the precursor for the M603 gene, it must be very closely related to Vh107. If the Vh107 gene segment is the germline precursor of the rearranged M603 Vh gene, it must have accumulated seven nucleotide differences (three of which are silent), either during B cell differentiation in the mouse from which the M603 tumor was initially isolated, or during growth and propagation of the tumor as it was passaged more than 100 generations. The Vh107 gene segment does encode the first 101 amino acids of the most frequently expressed phosphorylcholine-binding Vh region, exemplified by SI07, without somatic mutation. Thus the first and second hypervariable regions are not encoded by “minigene” inserted into germline Vh genes during lymphocyte differentiation, as postulated by Kabat and co-workers (Wu and Kabat, 1970; Kabat, 1975; Rao et al., 1979; Schilling et al., 1980). Where do the “extra” nucleotides in the PC Vh genes come from? Other explanations are possible, but they probably originate from a third germline gene segment (D, or diversity) which joins between Vh and Jh gene segments. Note that in the Vh and Jh gene segments (Figures 3 and 5), coding sequences are separated by at most three nucleotides from the conserved heptanucleotide sequence which is associated with the point of DNA joining in light chains (Max et al., 1979; Sakano et al., 1979). The position of this conserved sequence, which probably plays the same role in both heavy and light chain DNA rearrangements, leaves no room for the D nucleotides to arise from DNA directly contiguous to either the Vh or Jh gene segments. Our model of Vh-D-Jh joining is illustrated in Figure 8. We propose that germline D gene segments are located between the Vh and Jh gene segments so that the Vh, D and Jh gene segments can be joined in the same order as they are found in the rearranged Vh gene. The D gene segments might be associated with the Cg gene, as are Jh gene segments, or, alternatively, they might be closely linked to Vh gene segments. If the latter possibility is correct, each group of Vh gene segments might have its own set of D gene segments. The Vh regions from myeloma proteins binding levan may lack D segments (Schilling et al., 1980). Thus some Vh gene segments might join directly to Jh gene segments in a manner similar to light chain V-J joining.

DNA Rearrangement As a Generator of Antibody Diversity

DNA joining can produce variable region diversity by at least two means: combinatorial joining of germline gene segments and codon alterations at the site of joining. Combinatorial joining suggests that any gene segment may be joined to any other. Thus 200 Vh, and 4 Jh gene segments could generate 800 V-Jh genes. If Vh regions are encoded by three sets of germline gene segments, Vh, D and Jh, the extent of combinatorial diversity will be greater than for light chains which are encoded only by V and J gene segments.
The Mechanism of DNA Rearrangement in V Genes

Light and heavy chain V and J gene segments all share one important feature. Blocks of relatively conserved noncoding nucleotides occur 3' to V gene segments and 5' to J gene segments (Bernard et al., 1978; Max et al., 1979; Sakano et al., 1979). The most common nucleotides adjacent to either light or heavy chain V gene segments are CACAGTG...CACACGTG, where the conserved hepta- and decanucleotides are separated by either 11 or 22 essentially random nucleotides (Table 1).

Similarly, light and heavy chain J gene segments are adjacent to GGTGTTTTGA....CAGCTG, where, again, the blocks of conserved nucleotides are separated by either 11 or 22 random nucleotides (Table 1). The conserved nucleotides next to V and J gene segments are nearly inverse complements of one another, as seen in Figure 8. By contrast, the 11 or 22 nucleotides which separate blocks of conserved nucleotides show no particular homologies, except between some related V gene segments (see V,41, V,42, and V,43 in Table 1).

What is remarkable, though, is the highly conserved spacing between the hepta- and decanucleotide sequences. In every case, the spacing is either 11 or 22 nucleotides, deviating from these values by at most one nucleotide (Table 1). The only exception is a Jq gene segment which appears to be nonfunctional (Max et al., 1979; Sakano et al., 1979). V gene segments have an 11 nucleotide spacer, and Jq gene segments have a 22 nucleotide spacer. The Vm gene segment has a 22 nucleotide spacer, and the Jq gene segment an 11 nucleotide spacer. Heavy chain V and J gene segments have 22 nucleotide spacers. No information is yet available for the putative heavy chain D gene segments.

One turn of the DNA helix requires about 10.4 nucleotide pairs (Wang, 1979). Thus the conserved spacing of 11 or 22 nucleotides corresponds to either one or two turns of the helix, plus an extra nucleotide. This spacing means that the near ends of the conserved hepta- and decanucleotides maintain a precise relative orientation on the same side of the DNA helix. While separated by one or two turns of DNA, the two blocks of conserved nucleotides are in much the same relative orientation on the helix as if they formed a continuous stretch of 17 conserved nucleotides.

The foregoing points may be summarized as follows. The same noncoding nucleotides are conserved in all three immunoglobulin gene families, which diverged in evolution sometime prior to the origin of vertebrates (Marchalonis and Cone, 1973). Accordingly, these sequences have been conserved for more than 500 million years. The conserved nucleotides 3' to the V gene segments and 5' to J gene segments are approximately inverse complements of one another. The conserved nucleotides always occur as a heptanucleotide and a decanucleotide, separated by either 11 or 22 essentially random nucleotides. This spacing corresponds closely to one or two turns of the DNA helix. The data currently available indicate that gene segments of one type within a gene family all have spacers of the same size. At least for the light chains, gene segments which can join to one another have different sized spacers: V, spacers are "one turn;" J, spacers are "two turns;" the Vm spacer is "two turns;" the Jq "one turn."

Those observations suggest a model for DNA rearrangement in V genes. Specific joining proteins exist in the precursors of B lymphocytes which can recognize and bind to the blocks of conserved nucleotides adjacent to V and J gene segments (and probably D...
segments as well). The same joining proteins could bind in opposite orientations to both V and J gene segments, since the recognition nucleotides are nearly inverse complements of one another (Figure 8). The variations from the "prototype" recognition sequence \( \text{TACAAAAGC} \) are most pronounced for V gene segments (Table 1). This may reduce joining protein binding to individual V gene segments to compensate for their larger numbers, if optimal DNA joining requires roughly equal numbers of protein molecules bound to each type of gene segment. One form of joining protein may bind to blocks of nucleotides with a "one turn" spacer, and another form to blocks with a "two turn" spacer. These forms may be different aggregations of the same subunits. The strong conservation of spacer lengths in the sequences shown in Table 1 probably reflects a fairly rigid orientation of two binding sites in the joining proteins, one for each block of conserved nucleotides. Greater deviations of spacer length would not allow both sites to interact with the DNA simultaneously. A joining protein bound across a "two turn" spacer can interact with a joining protein bound across a "one turn" spacer to form a complex (Figure 8) in which the two gene segments are subsequently cut and ligated together. Slight differences in the points of cutting and joining can produce junctional codon variations (Weigert et al., 1978, 1980; Max et al., 1979; Sakano et al., 1979).

If there is a functional difference between 11 and 22 nucleotide spacers as we suggest, gene segments of the same type would be prevented from joining to one another. Otherwise, if, for example, two V\textsubscript{γ} gene segments were to join, breakage and inversion of the chromosome could result. Our model predicts that the heavychain D gene segments will have 11 nucleotide spacers on both sides. This would allow both V\textsubscript{v}D and D-J\textsubscript{v} joining to occur by the "one turn plus two turns" mechanism. It would also prevent V\textsubscript{γ} and J\textsubscript{γ} gene segments from joining without a D gene segment, unless some V\textsubscript{γ} or J\textsubscript{γ} gene segments had 11 nucleotide spacers.

One important implication of this model is that similar or identical enzyme systems would mediate the DNA rearrangements of the V-J (and D) gene segments in all three immunoglobulin families. This supposition would account both for the conserved recognition sequences and the conserved spacing relationships.

There is one additional interesting observation. The D nucleotides in S107 mRNA form a palindrome around a central G (Figure 9). The same sort of symmetry is also seen in the D nucleotides of M603 and M167 mRNAs (Figure 7). Whether this symmetry might be important for rearrangement of the postulated D gene segments is unknown.

There are many cases in proaryotes in which proteins bind specifically to DNA sequences with a 2 fold axis of symmetry (see Lewin, 1977). The structure shown in Figure 8 differs from these procaryotic binding sites mainly by including a large loop of DNA between the two nearly symmetric halves. A mechanism for DNA rearrangement which involves protein recognition of specific sites in duplex DNA appears more plausible than the cruciform structure postulated by other investigators (Max et al., 1979; Sakano et al., 1979). Such a structure is not likely to form spontaneously between widely separated regions of DNA containing immunoglobulin gene segments. Furthermore, the cruciform model would not account for the conservation of "one turn" and "two turn" spacers or of the same blocks of noncoding nucleotides in widely divergent gene families. An analog may also be drawn with the operator sequences in \( \lambda \) phage, where repressor binding occurs to the symmetric duplex DNA rather than to an alternative cruciform structure (Maniatis and Ptashne, 1973).

The Evolution of Systems for Specific Protein-DNA Interactions

In the immunoglobulin gene segments, a system of protein-DNA interactions has apparently evolved which recognizes two specific blocks of conserved nucleotides and the spacing between them. Probably two separate protein domains or subunits exist, one binding to \( \text{TACAAAAGC} \) and one to \( \text{TACAAAAGG} \). These separate binding sites may have been present on independent proteins, but have since been combined to produce two alternative kinds of joining proteins. We postulate that one of these kinds of joining proteins binds to blocks of specific nucleotides separated by a "one turn" spacer, and the other kind to blocks separated by a "two turn" spacer. This may represent a general strategy in eucaryotes for generating highly specific DNA-binding proteins by combining smaller sequence-specific proteins. The specificity of the combined protein is determined not only by the subunit specificities, but also by the overall protein con-

\[
\begin{array}{cccccccc}
C & - & - & - & - & - & - & - \\
T & - & - & - & - & - & - & - \\
C & - & - & - & - & - & - & - \\
T & - & - & - & - & - & - & - \\
\end{array}
\]

Figure 9. Symmetry of the S107 D Sequence

The partial sequence shown was determined from S107 mRNA. Portions of this sequence which could derive from the V\textsubscript{H107} and J\textsubscript{H107} germline gene segments are listed horizontally. The vertical stem (D) is coint complementary. An alternative representation of the D nucleotides would be as a palindrome in duplex DNA.
formation, which determines the length of any spacer between blocks of recognition nucleotides.

Mouse light and heavy chain gene families are known to be on separate chromosomes (Mage et al., 1973; Hengartner, Meo and Müller, 1978; Swan et al., 1979). This chromosomal separation may be a necessary consequence of the use of the same or similar recognition signals for DNA rearrangement in all these gene families. Otherwise, joining might occur between V and J (or D) gene segments from different families, leading to nonfunctional immunoglobulins.

Experimental Procedures

Cloned cDNA Probes

p107aR5 contains 1550 nucleotides of the S107a mRNA sequence, including the entire translated portion (Early et al., 1979). The insert is bordered by Eco RI synthetic linkers in the Eco RI site of pMB9. p107aR5 was subcloned to produce a plasmid which hybridizes only to PC V, gene segments. 4.45 bp V, fragment of p107aR5 can be prepared by digestion with Eco RI + Hha I. This fragment is bounded by an Eco RI linker and a Hha I cleavage site in amino acid codons 114-115, the middle of the J, gene sequence in the S107 V, gene. Terminal transferase (P/L Biochemicals) was used to add oligo(dG)10-20 to Eco RI + Hha I-digested p107aR5 (Rochebiochemistry, Jav and Wu, 1976). The tailed DNA's were separated on an 8% polyacrylamide gel, and the 445 bp V, fragment was eluted by the method of Maxam and Gilbert (1977), omitting SDS. Oligo(dG)10-20 was added to pBR322 which had been linearized with Pst I (W. Rowekamp and R. Firtel, personal communication). Equimolar amounts of these two tailed DNAs were annealed and used directly to infect E. coli strain 1776 (Villa-Komaroff et al., 1978). One of the resulting plasmid clones, p107V1, contains the 445 bp V, fragment of p107aR5, flanked by oligo(dG)10-20 tracts and the regenerated pBR322 Pst I sites. This plasmid was labeled with 32P by nick translation (Maniatis, Jeffrey and Kleid, 1978) as a probe for PC V, gene segments. It hybridizes only to the J, gene segment under the conditions used (Engel and Dodgson, 1978). 32P-labeled p107aR5 was used to detect PC JH gene segments (it will also hybridize to PC VH gene segments and the C, gene).

Germline Clones

The germline Charon 4A phage library from which ChSpPC3 and ChSpo7 were isolated was derived from BALB/c mouse sperm DNA partially digested with Eco RI (Davis et al., 1980). The cells used to prepare DNA were assayed by light microscopy to be greater than 90% sperm (Cowan et al., 1960). We have also isolated genomic clones with gene segments homologous to V,JH and Jg from a germine library containing more than 99% sperm DNA (P. Early, M. Davis and K. Calame, unpublished observations). Southern blots of Eco RI-digested sperm and embryo DNA's show no difference in the bands observed by hybridization to S107 PC DNA probes (Davis et al., 1979, 1980).

mRNA Sequencing

A modification of the dideoxynucleotide procedure (Sanger et al., 1977) was used to determine the sequences of cDNAs made from a mRNA template. An 80 bp Hind I fragment of p107aR5 encoding C, amino acids 126-150 was isolated from 10% polyacrylamide gel by the procedure of Maxam and Gilbert (1977), omitting SDS. Approximately 0.3 pmole each of an mRNA and the 80 bp Hind I fragment were mixed in 2 μl of 33 mM NaCl, 33 mM Tris-HCl (pH 7.5) and sealed in a siliconized glass capillary. This was boiled 2 min and annealed at 85°C for 30 min. The capillary was broken and the contents blown into a tube containing 1 μl 66 mM MgCl₂, 66 mM dithiothreitol and 6 μCi 32P-dCTP (Amersham, 350 Ci/mmole). The volume was adjusted to 10 μl with H₂O and 2 μl were distributed to each reaction tube. Deoxynucleotide triphosphates, dithiothreitol, dCTP and AMV RNA-dependent DNA polymerase (J. Beard) were added and incubated essentially as described by Sanger et al. (1977). Reactions were hypotized and suspended in 3.5 μl of 80% formamide, 10 mM EDTA plus dyes prior to electrophoresis on a 0.4 mm 8% Trit-silicate urea gel (Sanger and Coulson, 1978).

Acknowledgments

This work was supported by an NSF grant. P. E. and M. D. are supported by an NIH Training Grant. H. C. is supported by an NIH Fellowship. H. H. is supported by an American Cancer Society Fellowship. We thank Tom Maniatis, Norman Davidson, Max Delbrück, Jim Posakony, Jim Schilling and Steve Crews for helpful discussions. During the preparation of this manuscript, we were advised by N. Newell, J. Richards and F. Blattner of their discovery of additional J(V) gene segments.

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

Received December 14, 1979

References
