Sequences of human immunoglobulin switch regions: implications for recombination and transcription

Frederick C.Mills, Jeffrey S.Brooker and R.Daniel Camerini-Otero*

Genetics and Biochemistry Branch, NIDDK, Building 10, Room 9D15, NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA

Received September 19, 1990; Revised and Accepted November 12, 1990

EMBL accession nos X56795 – X56797

ABSTRACT

We have sequenced the entire human S_μ and $S_{\gamma 4}$ immunoglobulin heavy chain class switch regions, and have also completed the sequence of human S_ϵ . S_μ is composed predominantly of GAGCT and GGGCT pentameric repeats, with these units also being found in S_ϵ at a much lower density. S_μ - $S_{\gamma 4}$ matches are infrequent, but $S_{\gamma 4}$ contains a cluster of repeated sequences similar to units in mouse γ switch sites and unrelated to the S_μ repeats, suggesting that S_μ - S_γ homology is not important in μ - γ switching. We examined our ϵ and $\gamma 4$ sequences for features that could regulate production of 'sterile' transcripts preceding switch recombination. There is an Evolutionarily Conserved Sequence (ECS) upstream from the human and mouse S_ϵ regions that overlaps and extends 5' to the start sites for human and mouse ϵ sterile transcripts. Similarly, an ECS upstream from $S_{\gamma 4}$ is homologous to a mouse sequence that overlaps and extends 5' to the start sites for mouse $\gamma 2b$ sterile transcripts. The ϵ and $\gamma 4$ conserved segments contain potential Interferon Stimulable Response Elements (ISRE's) that are identical between human ϵ and $\gamma 4$.

INTRODUCTION

Immunoglobulin heavy chain gene expression utilizes two distinct kinds of DNA recombination (for reviews see 1, 2). The first type of recombination that occurs during B lymphocyte ontogeny is the site specific assembly of V (variable), D (diversity), and J (joining) DNA segments into VDJ units coding for the variable regions of heavy chain proteins. Early in B cell development, RNA transcripts containing VDJ sequences and C_μ heavy chain constant region sequences are processed to form mRNA species coding for IgM heavy chain molecules. Subsequent to IgM expression, there is a second type of recombination known as heavy chain class switching, in which the C_μ constant region is replaced by a C_δ , C_γ , C_α , or C_ϵ constant region. A class switch thus results in production of an antibody in which a heavy chain variable region that was originally part of an IgM heavy chain is attached to an IgD, IgG, IgA or IgE heavy chain. Since variable regions of heavy chain proteins fit together with variable regions

of immunoglobulin light chain proteins to form the antigen binding sites of antibody molecules, switching allows a given antigen specificity to be joined to heavy chains which have different functions in the immune system. Switching from membrane-bound IgM to IgD results from differential splicing of a primary transcript containing VDJ, C_μ , and C_δ coding sequences, while secretion of IgD frequently involves C_μ deletion resulting from recombination between sequences upstream from C_μ and sequences upstream from C_δ (3, 4). Switching to γ , α , and ϵ isotypes is directed by lymphokines such as γ -interferon and interleukin 4 (1). These lymphokines induce transcription of RNA species containing sequences of unrearranged C_γ , C_α , or C_ϵ constant region genes (5–13). The mechanism for production of γ , α , or ϵ transcripts from unrearranged constant region genes has not been defined, but it is clear that following transcriptional activation, DNA rearrangement deletes C_μ and C_δ , bringing C_γ , C_α , or C_ϵ coding sequences into proximity with the VDJ segments (1, 2). After this switch recombination, B cells typically secrete high levels of a single IgG, IgA, or IgE isotype, and the switching process is complete.

Switch recombination occurs between a large switch sequence, or S region, upstream from the C_μ constant region coding region sequence, and large S regions similarly positioned upstream from the C_γ , C_α , and C_ϵ constant region sequences (2). Unlike the relatively precise molecular events that occur in immunoglobulin VDJ joining, switch recombination typically involves an imprecise joining at different positions throughout several thousand bases of S_μ sequence and similarly variable positions within the non- μ S regions. In addition, switching can also occur with some frequency at positions outside the S regions. The sequenced S regions in the literature are composed of many tandem repeats that show varying degrees of homology between the different switch regions. Homologous recombination in mammalian cells, as well as lower eukaryotes and bacteria, seems to require only small patches of homology (14–17) so one might on this basis suspect class switching to be a physiological gene rearrangement which involves homologous recombination. Furthermore, data from our laboratory indicates that in a cell free system, eukaryotic and prokaryotic recombinases are in fact able to pair DNA sequences that share only small segments of homology (18; Mills, F. unpublished results).

* To whom correspondence should be addressed

A first step in assessing the possible role of homologous recombination in class switching is a comparison of the DNA sequences for potentially recombining switch regions. In order to assess the homologies between switch regions, we have sequenced both the entire human μ switch region and $S\gamma 4$, the human switch region least homologous to $S\mu$. We have also completed the sequence of human $S\epsilon$, a switch region highly homologous to $S\mu$. These data represent the first extensive sequencing of human switch regions, and provides the only complete sequence for the 'donor' $S\mu$ region in any species. These sequences have allowed us to define $S\mu$ - $S\epsilon$ and $S\mu$ - $S\gamma 4$ sequence homologies, and evaluate whether or not these homologies fall within the range of matches which could support general recombination.

In order to fully define structural aspects of the switch sites, it is also necessary to recognize that class switching, in common with other recombination events (19, 20), appears tied to transcriptional activation of recombinogenic sequences. We have therefore attempted to identify features in the vicinity of the human ϵ and $\gamma 4$ switch regions that might function in regulating their transcription. The initiation sites for sterile transcripts from human and mouse ϵ loci, as well as the human $\gamma 1$, $\gamma 3$ and mouse $\gamma 2b$ loci have been defined (10-12,45) making it possible to examine our human ϵ and $\gamma 4$ sequences for evolutionarily conserved regions that could serve as promoters for sterile transcripts. We have searched our sequences for well-characterized general regulatory signals, such as TATA, CAAT, and GGCGG motifs associated with specifying and modulating transcription (21). Furthermore, recent studies have shown that interferon-induced transcription of a number of genes, including those for metallothionein, 2', 5'-oligo A synthetase, and an MHC protein (22-24), are controlled by small, defined sequences, called Interferon Stimulable Response Elements (ISRE's; ref.25).

Since γ -interferon is known to regulate switching (26, 27), ISRE's could in fact function as specific regulatory signals for switch site transcription. With this in mind, we have also searched our human ϵ and $\gamma 4$ locus sequences for potential ISRE's.

MATERIALS AND METHODS

Enzymes and Reagents

Restriction endonucleases, T4 DNA polymerase, Klenow fragment of *E. coli* DNA polymerase I, and polynucleotide kinase were purchased from New England Biolabs (Beverly, MA). DNA ligase was from Bethesda Research Laboratories (Gaithersburg, MD), calf intestinal phosphatase was from Boehringer Mannheim (Indianapolis, IN), and low gelling temperature agarose was from FMC Bioproducts (Rockland, ME). 35 S-labelled dATP was from New England Nuclear (Dupont, Wilmington, DE). Sequenase and TaqI DNA sequencing kits were from United States Biochemicals (Cleveland, OH).

General Methods

Restriction enzyme digests, and DNA subcloning operations such as purification of DNA fragments from agarose gels, vector preparation, ligation and bacterial transformations were carried out by standard methods (28). Oligonucleotides were synthesized by Mr. George Poy (National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH, Bethesda, MD) on a 380 B Synthesizer from Applied Biosystems (Foster City, CA).

Genomic Phage Clones

The lambda phage clone $\lambda J_H-C\mu$ is a germline 12.2 kilobase fragment from the human μ locus cloned in the λ CH4A vector. This clone was isolated by one of the authors (F.C.M.) from the Maniatis human Hae III-AluI library by screening with a

Fig.1. Genomic clones from the human heavy chain locus and location of sequenced regions for $S\mu$, $S\gamma 4$, and $S\epsilon$ (30, 36, 37). The λ phage clone $\lambda J_H-C\mu$ contains a 12.2 kilobase fragment from the human μ locus cloned in the λ CH4A lambda phage vector. The λ phage clone $\lambda 81A$ contains a 17.5 kilobase fragment from the $\gamma 4$ locus, and $\lambda 38A$ contains a 13.5 kilobase fragment from the ϵ locus. Both $\lambda 81A$ and $\lambda 38A$ were cloned in the $\lambda 28$ lambda phage vector.

XbaI-EcoRI fragment containing $S\mu$ (Fig. 1A), a 4757 base pair Bam HI-Hind III segment containing human $S\gamma 4$ (Fig. 1B), and a 2739 base pair BamHI-Pvu II segment containing most of $S\epsilon$ (Figure 1C) were subcloned into pUC12 and pUC13 plasmids, as well as M13mp8, mp18, and mp19 for DNA sequencing from M13 universal primers. Sequencing of a central $S\mu$ 2.0 kb Sac I fragment cloned into M13mp18 was performed on ordered deletants (31). Sequence data were also obtained by using oligonucleotides complementary to sequences within these segments. The μ and ϵ segments were completely sequenced on both DNA strands. The $\gamma 4$ segment was completely sequenced on both strands, except for a 352 base pair region extending from positions 2740–3091 on the $\gamma 4$ sequence (Fig. 5). The difficulty in obtaining sequence data from both strands is most likely due to the presence of multiple runs of G residues in this region.

DNA sequencing

DNA sequencing was performed using ³⁵S-dATP according to the protocols in United States Biochemicals Sequenase and TaqI sequencing kits. Restriction fragments spanning a 4502 base pair

Pentameric Repeats

Fig. 2. Human S_{μ} sequence and Repeating Units. (A) A 4502 base pair XbaI-EcoRI fragment contains S_{μ} . Typical S_{μ} pentameric repeats extend from position 331 to 3716. From position 3471 to position 3595 there are five 25 base pair repeats of the consensus TATACTGGGCTTAGCTGGGCTGGGC, with the most 3' repeat lying 130 bases upstream from a known DNaseI hypersensitive/nuclear factor I core binding site (39). (B) Human S_{μ} Dot Matrix. The S_{μ} sequence was analyzed for repetitive structural features by self-comparison using the dot matrix method of Pustell (33). The dots on this S_{μ} dot matrix indicate matches between 30 basepair segments with homologies of 90% or greater. This high limit for homology scores is necessary to distinguish larger structural features above a background of matches due to a high density of segments containing 6 or more contiguous pentameric repeats. The large box of matches extending from position 331 to position 3716 indicates the borders of S_{μ} as defined by pentameric repeats. The dot matrix reveals one small, denser box of pentameric repeats at the 5' border of S_{μ} , and a second dense box of repeats at positions 2250–2500. The 25 base pair repeats aligned in the S_{μ} sequence are visible as a third, small dense box at the 3' border of S_{μ} . There is also an off-diagonal line extending from positions 1500–1800: this indicates the presence of a large direct repeat at this location. For reference XbaI, Sac I, and EcoRI restriction sites have been placed along the x and y axes of the matrix.

DNA sequence analysis

Sequence similarity between DNA sequences was analyzed using the University of Wisconsin Genetics Computer Group software package implemented on a VAX minicomputer at the National Cancer Institute, Frederick, Maryland (32) and a dot matrix program (33) supplied by International Biotechnologies Incorporated (New Haven, CT) and implemented on a Macintosh IIX computer. Dot matrices were run so as to search for sequence similarities on both strands of a given sequence. Sequences were scanned for repeats using the RELATE program (34), and candidates for directly repeated sequences in S_{μ} were aligned using the GENALIGN multiple sequence alignment program (35). Both RELATE and GENALIGN are implemented on the National Cancer Institute VAX computer at Frederick, Maryland.

RESULTS

S_{μ} sequence

The 4502 base pair fragment containing human germline S_{μ} was completely sequenced (Fig. 2A). A 3385 base pair segment beginning with nucleotide 331 and ending with nucleotide 3716 contains densely clustered pentameric repeats, with the motifs GGGCT and GAGCT being most common. When the S_{μ} sequence is run against itself using dot matrix analysis (33), the borders of the pentameric repeat region are seen graphically as a large rectangle on the dot matrix (Fig. 2B). Since mouse physiological switching occurs predominantly within the pentameric repeat region (2), we take the pentameric repeats to define the borders of S_{μ} . The (C/T)AGGTTG consensus

heptamer sequences clustered 5' to mouse S_{μ} and postulated to serve as signals in switch recombination (38) are not clustered 5' to human S_{μ} . At the 5' boundary of S_{μ} as well as at an internal region covering approximate positions 2250–2500 there are particularly dense boxes of direct pentameric repeats. There is also a large internal segment which is directly repeated, as seen by the presence of a line extending parallel to the main diagonal from position 1500 to position 1800. Finally, the 3' border of S_{μ} has an array of repeats that are seen to be five 25 base pair units of the form

with the units being exactly repeated except for a single nucleotide at the 3rd position in the second unit (Fig. 2A). These 25 base pair repeats lie 255–130 base pairs 5' from a previously described strong Nuclear Factor I (NF1) binding site / DNaseI hypersensitive site (39).

A search for repeated sequences indicates that there are repeats in S_{μ} whose second pentameric unit contains the AA dinucleotide. By organizing the entire S_{μ} sequence into segments of variable length with AA in the second pentameric unit, we have been able to discern a weak higher-order structure in the human S_{μ} pentameric repeat region (Fig. 3A). A representative alignment of S_{μ} repeated structures is shown in Fig. 3B. We note that published (although incomplete) sequence data for mouse S_{μ} also shows a higher-order structure in which pentameric repeats are organized into segments of variable length, with these segments having sequences that are similar to each other at their 5' ends (40).

Fig. 3. Organization of S_{μ} Repeats. (A) Higher-Order Structure of the S_{μ} Pentameric Region. It is possible to organize the S_{μ} pentameric sequences from position 350 to position 3470 into 53 segments whose second pentameric unit contains (with three exceptions) the AA dinucleotide. This organization excludes the nearly exact 25 base pair repeats at the 3' boundary of S_{μ} . (B) Sample Alignment of S_{μ} Higher-Order Repeats. Sets of the higher-order repeat segments were simultaneously aligned using the GENALIGN program (35), in the pairwise alignment mode. All 53 segments could not be aligned simultaneously, and there was considerable variation in the quality of alignment between different sets. The alignment shown for S_{μ} segments 10–30 covering the sequence 881–1914 was one of the better alignments for a large set of segments. The number of each aligned repeat (as designated in Fig. 3A) is found on the left-hand side of the sequence, while the number on the right hand side of each repeat indicates the length of that sequence.

Fig. 4. S_{ϵ} sequence and analysis. (A) Sequence of a 2739 base pair S_{ϵ} BamHI-PvuII segment. Positions 405–538 on our human ϵ sequence comprise an Evolutionarily Conserved Sequences (ECS) that shows 84% homology to a 130 base pair sequence found approximately 2000 base pairs 5' to mouse S_{ϵ} (10, 11). The human and mouse ECS's overlap sequences containing start sites for ϵ sterile transcripts, and extend 5' to these start sites in both human and mouse (10–12). The numbering for the mouse sequence is the numbering used in reference 10. We have used dotted overlining to indicate the positions of human ϵ transcription start sites, and have used solid overlining to indicate potential regulatory features in the putative ϵ transcriptional promoter. These features include an ISRE-like element also found upstream from $S_{\gamma 4}$ (Fig. 5A). A vertical line along the right hand side of the ϵ sequence indicates that S_{ϵ} (as defined by pentameric repeats homologous to those in S_{μ}) extends from position 1201 onward. (B) Dot matrix comparison of S_{μ} and S_{ϵ} . Shown is a dot matrix of S_{μ} versus S_{ϵ} with matches of 65% or greater homology between 30 base pair segments. For this analysis we joined the S_{ϵ} sequence reported in this paper with the previously published 3' portion of S_{ϵ} (43). A region of matches with S_{μ} extends for approximately 2200 base pairs in the ϵ locus, defining the borders of S_{ϵ} by the criterion of S_{μ} -like pentameric repeats. There is a long, faint diagonal extending from positions 1000 and 2700 on the S_{ϵ} sequence, and between positions 1000 and 2700 on the S_{μ} sequence. This diagonal is an apparent remnant of the evolutionary relationship between human S_{ϵ} and S_{μ} .

There appear to be no significant inverted repeats in *Sμ*: when the sequence of the XbaI-EcoRI fragment is run only against its reverse complement on a dot matrix, one obtains an essentially blank dot matrix (analysis not shown).

S_ε Sequence

Next, we sequenced a 2739 base pair BamHI-Pvu II segment containing most of Se (Fig. 4A). Non-contiguous 533 and 645

base pair sequences within this segment have been published previously (12, 41). Clustering of *Sμ*-like pentameric repeats is discernible beginning with nucleotide 1201, although these are both considerably less dense and less regularly spaced than in *Sμ*. Although we were able to identify several duplicated and triplicated sequences within the pentameric region, we were unable to discern higher order multiple repeats such as those found in the mouse *Se* sequence (42).

Fig. 5. S γ 4 sequence. A 4757 base pair γ 4 BamH I-Hind III segment contains S γ 4 as defined by sequence similarity with S μ . In agreement with Akohiri et al. (44) at positions 89–246 we find twelve 21 base pair repeats of the consensus. CTGGA (T/C) (G/A) CTTCTGTTCCA (C/T) (G/T). From positions 1354–1562 we have aligned the human γ 4 sequence with a sequence approximately 2500 base pairs upstream from mouse S γ 4 that contains the start sites for mouse γ 2b sterile transcripts. We also show the alignment of the γ 4 sequence with sequences approximately 800 base pairs 5' to human S γ 3 (45; Mills, F. unpublished results). The numbering for the mouse γ 2b sequence is from ref. 9, and the numbering for the human γ 3 sequence is from ref. 45. The mouse γ 2b segment is 63% homologous to human γ 4, and the human γ 3 segment is 93% homologous to human γ 4. We have overlined a segment in the region of γ 3- γ 4- γ 2b Evolutionarily Conserved Sequence which is a potential ISRE (Interferon Stimulable Response Element). A vertical line extending along the right hand side of the sequence from approximate positions 2300–4600 indicates a region containing sequences similar to the repetitive sequences found within S μ . We take this 2300 basepair region to comprise S γ 4. A vertical line along the left hand side of the sequence extending from positions 2768–3466 indicates γ 4 repeated sequences. These repeats are similar to the repeating units in mouse S γ switch regions (Fig. 6C).

We combined our sequence with published ϵ locus sequence spanning the approximate 1100 base pairs between the 3' end of our sequence and C ϵ (43) in order to define the borders of S ϵ by means of a dot matrix comparison with S μ (Fig. 4B). A sharp decrease in the density of matches with S μ at positions 1201 and 3426 of the combined sequence indicates that human S ϵ as defined by homology with S μ is approximately 2200 basepairs in length. We obtain similar boundaries for S ϵ by running a dot matrix comparison of the S ϵ sequence versus itself (analysis not shown). The lower density of repeating units in S ϵ as opposed to S μ may be appreciated from the fact that in order to define the borders of S ϵ the dot matrix analysis in Fig. 4B was done at 65% matching stringency. When a similar analysis was carried out at 90% stringency (which was sufficient to define the borders of S μ in Fig. 2B), the number of S μ -S ϵ matches was reduced so drastically that it was impossible to define the borders of S ϵ .

An important control point for regulation of class switching seems to be activation of unrearranged non- μ genes and production of sterile transcripts that contain sequences from α , γ or ϵ constant regions and their associated switch regions (5–13). Transcription of the unrearranged human and mouse ϵ genes has recently been characterized, with the ϵ transcription start sites shown to lie within an 81 base pair segment -610 to -691 base pairs upstream from the 5' border of human S_{ϵ} and within a 250 base pair segment approximately -2000 to -2250 base pairs upstream from the 5' border of mouse S_{ϵ} (10–12). We have found a 133 base pair Evolutionarily Conserved Sequence (ECS) extending from positions 405 to 538 on our human ϵ sequence (-669 to -803 base pairs upstream from the 5' border of human

Se) that shows 84% homology to a 130 base pair segment approximately 2000 base pairs upstream from mouse Se. Both the human and mouse ECS regions overlap the sequences containing the start sites for ϵ sterile transcripts and also extend 5' to these sequences. The significantly different locations of the human and mouse ECS regions relative to Se indicates that the distance between Se and the start sites for sterile transcripts is not critical. The human-mouse alignment is shown in Fig. 4A. We also note that a 131 base pair segment upstream from the human yel pseudogene is 95% identical to the ϵ Evolutionarily Conserved Sequence (see Fig. 1 and ref. 41). A putative human ϵ Interferon Stimulable Response Element (ISRE) is discussed below in the context of an identical element found 5' to human S γ 4.

S_γ4 Sequence

The sequence of a 4757 base pair BamHI-Hind III segment from the $\gamma 4$ locus is shown in Fig. 5. In agreement with Akohiri et al., (44) we find twelve 21 base pair tandem repeats of the consensus

CTGGA (T/C) (G/A) CTTCTGTCCA (C/T) (G/T)

at positions 89–346 in the BamHI fragment. These repeated units have no significant homology to units in $S\mu$ or $S\epsilon$.

Sterile transcripts of one unrearranged mouse γ gene ($\gamma 2b$, ref. 9) and two human γ genes ($\gamma 1$ and $\gamma 3$, ref. 45) have been shown to initiate in sequenced regions 5' to their respective switch sites. Nucleotides 1354–1562 of the sequenced human $\gamma 4$ BamHI-Hind III segment (–724 to –932 base pairs upstream

Table 1.[illegible]

Fig. 6. Dot matrix analysis of $S_{\gamma 4}$. (A) Comparison of S_{μ} and $S_{\gamma 4}$. Dot matrix comparison between S_{μ} and the sequence of the 4757 base pair BamHI-Hind III fragment reported in this paper detects only a limited number of 30 base segments with 65% or greater homology. The best matches are found in the vicinity of position 2300 on the $S_{\gamma 4}$ sequence, but matches also extend approximately 2300 base pairs downstream from this location. (B) $\gamma 4$ versus $\gamma 4$ comparison. Dot matrix comparison of the $\gamma 4$ sequence versus itself was performed in order to detect direct and inverted repeated segments within the sequence. The cluster of 21 base pair repeats near the 5' BamHI site form a small, dense box extending from positions 89–246, while the repeats typical of S_{γ} sites form a larger box extending from positions 2768–3466. A long, off-diagonal element running opposite to the main diagonal in this self-self comparison indicates that a segment extending from approximate positions 400–1100 is repeated in an inverted orientation at approximate positions 3500–4200. (C) Mouse $S_{\gamma 1}$ versus human $S_{\gamma 4}$ comparison. Dot matrix comparison of human $S_{\gamma 4}$ versus mouse $S_{\gamma 1}$ (46) indicates that the tandemly repeated sequences extending from positions 2768–3466 in the human $\gamma 4$ sequence are similar to repetitive units extending from approximate positions 4000–8200 in the mouse $\gamma 1$ sequence, as judged by the presence of long, vertical columns of small diagonals on this matrix. The human $S_{\gamma 4}$ repeated elements are also similar to repeated elements found in mouse $S_{\gamma 2a}$, $S_{\gamma 2b}$, and $S_{\gamma 3}$ (analysis not shown), as well as elements found in human $S_{\gamma 3}$ (Mills, F. unpublished results).

from the 5' border of $S_{\gamma 4}$, see below) show 63% homology to a 223 base pair mouse $\gamma 2b$ segment that contains most of the start sites for $\gamma 2b$ sterile transcripts (9), and are also 93% homologous to sequences upstream from the start site for human $\gamma 3$ sterile transcripts (45). A 173 base pair segment of the $\gamma 4$ ECS has been published previously (45), but our extensive sequencing of the $\gamma 4$ locus allows us to locate this ECS precisely with reference to the $\gamma 4$ switch region. We also find that the conserved $\gamma 4$ sequence shows extensive homology to sequences

upstream from mouse $S_{\gamma 2a}$ and $S_{\gamma 1}$ (42, 46, analysis not shown).

Extending from positions 2200–4600 on the $\gamma 4$ sequence, there is a modest density of sequences similar to S_{μ} repeats (Fig.'s 5, 6A), suggesting these 2400 nucleotides roughly represent $S_{\gamma 4}$. The best S_{μ} - $S_{\gamma 4}$ match is at positions 2286–2411 in the $\gamma 4$ sequence, with 66% homology to a segment from S_{μ} (Table I). Since the best matches upstream from this position are less than half the length of this region of homology (88% over 16 base pairs, 74% over 27 base pairs, and 70% over 48 base pairs),

we take this 126 base pair segment to represent the 5' border of $S\gamma 4$. Moreover, the fact that this sequence lies downstream from the $\gamma 4$ segment homologous to the start sites for sterile transcripts in other γ genes (9, 45) also argues that this segment of $S\mu$ - $S\gamma 4$ homology is in the correct position to comprise the 5' border of $S\gamma 4$.

There is a cluster of G-rich tandemly repeated sequences extending from positions 2768–3466 in the $\gamma 4$ sequence (Fig. 5 and Fig. 6B). These repeats, varying from 68–97 base pairs in length, are located within $S\gamma 4$, and are homologous to the repeats found in mouse γ switch sites (ref. 46, Fig. 6C). Inspection of the 9 $S\gamma 4$ repeats (Fig. 5) shows that one repeat contains no GAGCT or GGGCT pentamers typical of $S\mu$, 8 repeats contain a single GAGCT pentamer, and no repeat contains a GGGCT pentamer, indicating that there is very little homology between the $S\mu$ and $S\gamma 4$ repeats.

Potential ISRE's Near Human $S\epsilon$ and $S\gamma 4$

It has been shown in both mouse and human systems that γ -interferon regulates heavy chain class switching with effects reciprocal to those of interleukin 4 (26, 27). Therefore it is possible that γ -interferon, like interleukin 4, controls non- μ heavy chain gene expression at a transcriptional level. Since transcriptional regulation of interferon-inducible genes is mediated by small Interferon Stimulable Response Elements (25), we asked whether these sequences are found upstream from the switch regions, where they could function as regulatory elements controlling transcription of unrearranged immunoglobulin heavy chain genes.

We find that there are segments resembling Interferon Stimulable Response Elements (ISRE's) within the Evolutionarily Conserved Sequences in the mouse ϵ locus (10, 11), as well as the human ϵ and $\psi\epsilon 1$ loci (this study and ref. 41). An experimentally demonstrated ISRE which regulates the expression of a human interferon-inducible gene designated 9–27 (25) matches the mouse ϵ ISRE-like sequence at 15 out of 23 positions, while the human ϵ ISRE-like sequence matches at 14 out of 23 positions (see Fig. 7A). Surprisingly, we also find that a 15 base pair segment in our $\gamma 4$ sequence is identical to the putative ISRE found 5' to human $S\epsilon$ (Fig. 4A). This ISRE-like sequence is located in the Evolutionarily Conserved Sequence found 5' to $S\gamma 4$ (Fig. 5). On the other hand, there is no flanking sequence homology between the ϵ and $\gamma 4$ ISRE's, suggesting that this identity is functionally important. We also find that a sequence upstream from the start sites for mouse α immunoglobulin sterile transcripts (13) is identical to the human ϵ ISRE in 19 out of 24 nucleotides. An alignment of the human ϵ and $\gamma 4$ ISRE-like sequences with the other putative immunoglobulin ISREs, as well as the ISRE 5' to the interferon responsive 9–27 gene, is shown in Fig. 7A.

DISCUSSION

The sequences in this paper complete approximately 27 kilobases of continuous sequence extending from the J_H - C_μ heavy chain enhancer through Cd (47, 48), 9 kilobases of continuous sequence in the $\gamma 4$ locus (44, 49), and 6 kilobases of continuous sequence in the ϵ locus (30, 43). These extensive tracts of sequence not only contain the entire $S\mu$, $S\gamma 4$, and $S\epsilon$ switch regions, but also place the switch regions in the context of considerable flanking sequence. Putative control functions for DNA segments in these switch regions and their flanking sequences are discussed below.

A

ACAGAGA AAAA cAgCTTCCAG	mouse α .	153-175.	ref. 13
ACAGAGAGAAAAGGGAACCTCCAG	human ϵ .	440-463.	this paper, ref. 41
AACAGAAcAAAAGGGAACCTTC	mouse ϵ .	641-662.	refs.10,11
CACCAGAA GAAAAGGGAACCTTC	human $\gamma 3$.	81-102.	ref. 45, Mills, F. unpublished
CACCAGAA GAAAAGGGAACCTTC	human $\gamma 4$.	1430-1451.	this paper, ref. 45
CAACAGAA GAAAGAGGAACCTTC	mouse $\gamma 2b$.	418-439.	ref. 9
aAACAGcAgGAAATAGAACTTa	9-27.	-180-157.	ref. 25
cacACAGaaagAAAagggAaCTTccag	consensus of above sequences		
GGAAAN(N)GAAACT	functional consensus		ref. 25

B

Fig. 7. Features of the human $\gamma 4$ and ϵ loci. (A) Alignment of ISRE elements. Immunoglobulin ISRE-like segments found approximately 2500 base pairs upstream from mouse $S\gamma 2b$, 800 base pairs upstream from human $S\gamma 3$, 800 base pairs upstream from human $S\gamma 4$, 750 base pairs upstream from human $S\epsilon$, 2000 base pairs upstream from mouse $S\epsilon$, and 2000 basepairs upstream from mouse $S\alpha$ were aligned with an ISRE element shown to confer interferon inducibility on the human gene 9–27 (25) using the GENALIGN program in the region alignment mode. A consensus of demonstrated ISRE elements is also shown beneath the alignment. (B) Summary of human $\gamma 4$ and ϵ transcriptional features. ECS designates Evolutionarily Conserved Sequences found 5' to $S\gamma 4$ and $S\epsilon$. $I\epsilon$ designates a small, non-coding exon in human ϵ sterile transcripts (12). Germline ϵ locus transcripts initiate at sites within a segment overlapping the ECS and are presumed to continue through $C\epsilon$. These primary transcripts are processed to yield approximate 1.7 kilobase RNA species which contain $I\epsilon$ sequences spliced onto $C\epsilon$ sequences. We postulate that a similar $\gamma 4$ noncoding exon, designated 'putative $I\gamma 4$ ', is likely to result from initiation of germline transcripts near the $\gamma 4$ ECS.

We also anticipate that as further data on class switching becomes available, additional control elements will be found in the sequences for these loci.

Structure of Human $S\mu$

The sequence of human $S\mu$ is characterized by a dense clustering of pentameric repeats. Except for five rigorous 25 base pair repeats at the 3' border of $S\mu$, there is only a loose organization of the $S\mu$ pentameric units into a higher-order structure. This loose higher-order structure may be functionally significant in switch recombination. Alternatively, the lack of rigorous higher-order structure may indicate that the $S\mu$ pentamers by themselves are sufficient for recognition by switch-recombination machinery.

Homologies between the Human Switch Sites

Although the $S\mu$ and $S\gamma$ switch regions are least homologous (2, 36, and this study), the μ - γ class switch is most frequent (50). Since the frequency of switching is not related to the degree of inter-S region homology, there has been considerable doubt as to whether homologous recombination is involved (35, 51). However, general eukaryotic recombinases can join DNA sequences with as little as 13 base pairs of continuous identical sequence, or sets of three smaller patches separated by deletions and/or loop-outs (18; Mills, F., unpublished results). The use of small homologous patches has also been documented for general recombination in mammalian cells in culture (14).

In Table I we show our best $S\mu$ - $S\epsilon$ and $S\mu$ - $S\gamma$ alignments. These matches were obtained by stepping through our $S\epsilon$ or $S\gamma$ sequences in 500 base pair intervals to obtain the optimal alignment of each 500 base pair segment with the complete $S\mu$ sequence (32). We found empirically that this procedure was more sensitive than searching for an alignment with the entire $S\epsilon$ or $S\gamma$ versus $S\mu$. Using this procedure we obtained an alignment for $S\epsilon$ versus $S\mu$ that is 80% identical over 186 base pairs, while the optimal alignment for $S\gamma$ versus $S\mu$ is 66% identical over 126 base pairs. Therefore we cannot exclude the possibility that the homologous recombination machinery utilizes these matches as a basis for switch site recombination.

However, we note that human $S\mu$ and mouse $S\mu$ share substantial homology (analysis not shown), as do human $S\gamma$ and mouse $S\gamma$, while intra species homology between human $S\mu$ and $S\gamma$, or between mouse $S\mu$ and mouse $S\gamma$, is relatively weak. Moreover, switch regions are found in the amphibian *Xenopus laevis* upstream from both $C\mu$ and $C\alpha$, a non- μ constant region gene, suggesting that switch region duplication occurred at least 350 million years ago (50, 52). Taken together, these observations argue that: (A) $S\mu$ and $S\gamma$ were duplicated from an ancestral sequence much earlier than the human / mouse divergence; and, (B) structural features peculiar to $S\mu$ and $S\gamma$ have been evolutionarily conserved, while $S\mu$ - $S\gamma$ homologies have not been conserved. It follows that $S\mu$ - $S\gamma$ homology *per se* is probably not functionally important in switch recombination.

An homology-independent model would involve the formation of a ternary complex between a switch-region specific recombination protein, $S\mu$ motifs, and $S\gamma$ -specific motifs. Homology-independent recombination would occur within the complex as a result of nuclease cutting and DNA ligation. We extend this model to $S\mu$ - $S\epsilon$, or $S\mu$ - $S\alpha$ recombination, but allow for the possibility that the homologous recombination machinery could act within the ternary complex (the fact that human $S\epsilon$ has little internal structure other than clusters of pentameric units homologous to those in $S\mu$ would be consistent with this hypothesis). In summary, we place the likelihood that class switching events involve homologous recombination on a gradient, with recombination between nearly exact repeats during the human μ - δ switch using homologous recombination machinery (3), μ - γ switching using homology-independent machinery, and μ - ϵ and μ - α switching not clearly attributable to either mechanism.

Putative initiation sites for sterile transcripts upstream from $S\epsilon$ and $S\gamma$

A DNA segment extending -669 to -803 base pairs upstream from human $S\epsilon$ overlaps the start sites for human ϵ sterile transcripts (12) and shows 84% homology to a segment overlapping the start sites for mouse ϵ sterile transcripts (10, 11).

This 134 base pair evolutionarily conserved sequence contains CAAT, GGCGG, and ATTTTA transcription signals, as well as a CTGTTGC 'X box' protein binding domain conserved upstream from Class II MHC genes (53), and a potential ISRE. We also find that our sequence -724 to -900 base pairs upstream from $S\gamma$ shows 63% homology to a DNA segment containing start sites for mouse γ 2b sterile transcripts. This segment contains an exact copy of the putative ISRE found in the human ϵ sequence. We believe these Evolutionarily Conserved Sequences represent control regions for regulating the production of sterile transcripts from unrearranged human ϵ and γ 4 genes. We also suggest that the presence of a segment upstream from the $\psi\epsilon$ 1 pseudogene which is very similar to the ϵ ECS may indicate that transcription of this pseudogene and the true ϵ gene are regulated in the same way.

Significance of Potential Immunoglobulin ISRE's

We observe that ISRE-like sequences are conserved upstream from the start sites for sterile transcripts initiating in the mouse α , ϵ , γ 2b, human ϵ , and human γ 3 loci, and are also found within a putative transcriptional control sequence located 5' to the human γ 4 switch region. Since the lymphokine γ -interferon is known to negatively regulate switching to mouse ϵ , mouse γ 2b, and human ϵ (26, 27), we suggest that ISRE's mediate a γ -interferon dependent negative control over the germline transcription of these immunoglobulin genes. Although most published studies on transcriptional control mediated by ISRE's only describe positive regulation (54, 55), there is a precedent for both positive and negative control mediated by the same ISRE (56). Moreover, the GGCGG transcriptional control element and immunoglobulin octamer sequence have both been found to function either as activators or repressors of transcription, depending on the DNA-binding proteins expressed by the cells in which these elements are assayed (57, 58). Taken together, these results suggest that regulation of switching by γ -interferon could be mediated by ISRE's as a tissue-specific or stage-specific phenomenon. We also postulate that the reciprocal effects of interleukin 4 and γ -interferon on switching could result from binding of different proteins to the same DNA sequence. Supporting this model is data showing that interleukin-4 treatment of mouse B lymphocytes induces a stage-specific protein which binds to a sequence 5' from a Class II MHC gene (59): this sequence contains a homopurine octamer AAGAAAAA reminiscent of the homopurine 'core' found within our putative immunoglobulin ISRE's (Figure 7A).

Our hypothesis that ISRE motifs and other sequences found upstream from the human ϵ and γ 4 switch sites regulate transcription of switch sites prior to recombination points to the importance of attaching these sequences to reporter genes and transfecting these constructs into lymphokine-induced cells. Complementary studies would involve investigating the interaction of these sequences with lymphokine-induced DNA binding proteins.

Summary

We conclude from analysis of our human μ , γ 4, and ϵ switch region sequences that homology is not important in μ - γ switch recombination, although it may mediate μ - ϵ recombination. However, regardless of the role that homology plays in switching, the extensive body of sequences presented in this paper make it possible to precisely map naturally occurring switch recombination junctions, switch-oncogene junctions, and junctions resulting from cell-transfection or cell-free switch

recombination systems, as well as providing a basis for analyzing the regulation of transcription during human class switching.

ACKNOWLEDGEMENTS

We would like to thank Edward Max and Ilyan Kirsch for providing λ phage clones containing human Se and $\text{S}\gamma 4$, respectively. We also thank George Poy for synthesizing oligonucleotides, Mark Boguski for providing advice on the use of the computer program RELATE, Nancy Valente for help with sequencing $\text{S}\mu$, and Linda Taylor for secretarial assistance. We gratefully acknowledge critical reviews of the manuscript by Peggy Hsieh, Moon Kim, Richard Proia, Philip Tucker, and Edward Max. One of the authors (F.C.M.) was supported by a Burroughs-Wellcome Senior Fellowship during part of the work reported in this paper.

REFERENCES

- Blackwell, T.K. and Alt, F.W. (1989) *Ann. Rev. Genet.*, **23**, 605–636.
- Gritzmacher, C. A. (1989) *Critical Reviews in Immunology*, **9**, 173–200.
- White, M. B., Word, C.J., Humphries, C.G., Blattner, F.R., and Tucker, P.W. (1990) *Mol. Cell. Biol.*, **10**, 3690–3699.
- Mountz, J.D., Mushinski, J.F., Owens, J.D., and Finkelman, F.D. (1990) *J. Immunol.* (in press).
- Stavnezer-Nordgren, J., and Sirlin, S. (1986) *EMBO J.*, **5**, 95–102.
- Lutzker, S., Rothman, P., Pollock, R., Coffman, R., and Alt, F.W. (1988) *Cell*, **53**, 177–184.
- Stavnezer, J., Radcliffe, G., Lin, Y.-C., Nietupski, J., Berggren, L., Sitia, R., and Severinson, E. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7704–7708.
- Esser, C. and Radbuch, A. (1989) *EMBO J.*, **8**, 483–488.
- Lutzker, S. and Alt, F.W. (1988) *Mol. Cell. Biol.*, **8**, 1849–1852.
- Gerondakis, S. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1581–1585.
- Rothman, P., Chen, Y.-Y., Lutzker, S., Li, S.C., Stewart, V., Coffman, R. and Alt, F.W. (1990) *Mol. Cell. Biol.*, **10**, 1672–1679.
- Gauchat, J.-F., Leberman, D.A., Coffman, R.L., Gascan, H., and de Vries, J.E. (1990) *J. Exp. Med.*, **172**, 463–473.
- Radcliffe, G., Lin, Y.-C., Julius, M., Marcu, K.B. and Stavnezer, J. (1990) *Mol. Cell. Biol.*, **10**, 382–386.
- Anderson, R.A., Kato, S., and Camerini-Otero, R.D. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 206–210.
- Pont, G., Degroote, F., and Picard, G. (1988) *Nucl. Acids Res.*, **16**, 8817–8833.
- Ahn, B.-Y., Dornfeld, K.J., Fagrelus, T.J., and Livingston, D.M. (1988) *Mol. Cell. Biol.*, **8**, 2442–2448.
- Watt, V.M., Ingles, C.J., Urdea, M.S., and Rutter, W.J. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4768–4772.
- Hsieh, P., Camerini-Otero, C.S., and Camerini-Otero, R.D. (1990) *Genes and Development* (in press).
- Voelkel-Meiman, K., Keil, R.L., and Roeder, G.S., (1987) *Cell*, **48**, 1071–1079.
- Klar, A.J.S., Strathern, J.N., and Abraham, J.A., (1984) *Cold Spring Harbor Symp. Quant. Biol.*, **49**, 77–88.
- Mitchell, P.J. and Tjian, R. (1989) *Science*, **245**, 371–378.
- Friedman, R.L. and Stark, G.R. (1985) *Nature*, **314**, 637–639.
- Benech, P., Vigneron, M., Peretz, D., Revel, M., and Chebath, (1987) *J. Mol. Cell. Biol.*, **7**, 4498–4504.
- Blonar, M.A., Baldwin, A.S., Jr., Flavell, R.A. and Sharp, P.A. (1989) *EMBO J.*, **8**, 1139–1144.
- Reid, L.E., Brasnett, A.H., Gilbert, C.C., Porter, A.C.G., Gewert, D.R., Stark, G.R., and Kerr, I.M. (1989) *Proc. Nat. Acad. Sci. USA*, **86**, 840–844.
- Snapper, C., and Paul, W. (1987) *Science*, **236**, 944–947.
- Thyphronitis, G., Tsokos, G. C., June, C.H., Levine, A.D., and Finkelman, F.D. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5580–5584.
- Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, published by Cold Spring Harbor Laboratory.
- Ford, A.M., Molgaard, H.V., Greaves, M.F., and Gould, H.J. (1983) *EMBO J.*, **2**, 997–1001.
- Max, E. E., Battey, J., Ney, R., Kirsch, I.A., and Leder, P. (1982) *Cell*, **29**, 691–699.
- Dale, R.M.K., Mclure, B.A., and Houchins, J.P. (1985) *Plasmid*, **13**, 31–40.
- Devereux, J., Haverli, P. and Smithies, O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
- Pustel, J., and Kafatos, F. C. (1984) *Nucleic Acids Res.*, **12**, 643–655.
- Dayhoff, M.O., (1979) *Atlas of Protein Sequence and Structure*, **5**, 353–358.
- Sobel, E. and Martinez, H.M. (1985) *Nucleic Acids Res.*, **14**, 363–374.
- Flanagan, J.G., and Rabbitts, T. H. (1982) *Nature*, **300**, 709–713.
- Hofker, M.H., Walter, M.A., and Cox, D. (1989) *Proc. Natl. Acad. Sci. USA*, **12**, 5567–5571.
- Marcu, K.B., Lang, R.B., Stanton, L. W., and Harris, L.J. (1982) *Nature*, **298**, 87–89.
- Henninghausen, L. Siebenlist, U., Danner, D., Leder, P. Rawlins, D., Rosenfeld, P. and Kelly, K.Jr. (1984) *Nature*, **314**, 289–292.
- Nikaido, T., Nakai, S., and Honjo, T. (1981) *Nature*, **292**, 845–848.
- Hisajima, H., Nishida, Y., Nakai, S., Takahashi, N., Ueda, S., and Honjo, T. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2995–2999.
- Nikaido, T., Yamakami-Kataoka, Y., and Honjo, T. (1982) *J. Biol. Chem.*, **257**, 7322–7329.
- Ueda, S., Nakai, S., Nishida, Y., Hisajima, H., and Honjo, T. (1982) *EMBO J.*, **1**, 1539–1544.
- Akahori, Y., Handa, H., Imai, Kenji, I., Abe, M., Kohzoh, K., Makemoto, H., Yasui, H., Okamura, K., Naito, M., Matsuoka, H., and Kurosawa, Y. (1988) *Nucleic Acids Res.*, **16**, 9499–9511.
- Sideras, P., Mizuta, T.-R., Kanamori, H., Suzuki, N., Okamoto, M., Kuze, K., Ohno, H., Doi, S., Fukuhara, S., Hassan, M. S., Hammarstrom, L., Smith, E., Shimizu, A., and Honjo, T. (1989) *Int. Immunology*, **1**, 631–642.
- Mowat, M. and Dunnick, W. (1986) *J. Immunol.*, **136**, 2674–2683.
- Mills, F.C., Fisher, L.M., Kuroda, R., Ford, A.M. and Gould, H.J. (1983) *Nature*, **306**, 809–812.
- Word, C. J., White, M.B., Kuziel, W.A., Shen, A.L., Blattner, F.R., and Tucker, P.W. (1989) *Int. Immunol.*, **1**, 296–309.
- Ellison, J., Buxbaum, J. and Hood, L. (1981) *DNA*, **1**, 11–18.
- Nisonoff, A. (1984) *Molecular Immunology*, Second Edition, pub. by Sinauer Associates, Inc., Sunderland, MA.
- Ravetch, J.V., Kirsch, I.R., and Leder, P. (1980) *Proc. Nat. Acad. Sci. USA*, **77**, 6734–6738.
- Schwager, J. (1987) *Basel Institute of Immunology Annual Report*, 40.
- Celada, A., Shiga, M., Imagawa, M., Kop, J. and Maki, R.A. (1988) *J. Immunol.*, **140**, 3995–4002.
- Dale, T.C., Rosen, J.M., Guille, M.J., Lewin, A.R., Porter, A.G.C., Kerr, I.M., and Stark, G.R. (1989) *EMBO J.*, **8**, 831–839.
- Decker, T., Lew, D.J., Cheng, Y.-S., E., Levy, D.E., and Darnell, J.E., Jr. (1989) *EMBO J.*, **8**, 2009–2014.
- Levy, D.E., Kessler, D.S., Pine, R., Reich, N., and Darnell, J.E. Jr. (1988) *Genes and Devel.*, **2**, 383–393.
- Lenardo, M.J., Staudt, L., Robbins, P., Kuang, A., Mulligan, R.C., and Baltimore, D. (1989) *Science*, **243**, 544–546.
- Kageyama, R. and Pastan, I. (1989) *Cell*, **59**, 815–825.
- Boothby, M., Gravalles, E., Liou, H.-C. and Glimcher, L.H. (1988) *Science*, **242**, 1559–1562.