Sequence comparison of alleles of the fourth component of complement (C4) and sex-limited protein (Slp)

C.Hemenway, M.Kalff, J.Stavenhagen, D.Walthall and D.Robins

Department of Biological Sciences, Columbia University, New York, NY 10027, USA

Received 17 December 1985; Revised 6 February 1986; Accepted 10 February 1986

ABSTRACT

cDNA clones specific for the fourth component of complement (C4) and its androgen-regulated isotype, sex-limited protein (Slp), have been isolated from two mouse haplotypes (H-2^d and H-2^{w7}) that show differential C4 activity and differential regulation of Slp. Clones were first isolated using a cDNA probe enriched by subtractive hybridization. Subsequent screening has resulted in cDNAs spanning the entire C4^d mRNA, as well as much of C4^{w7}, Slp^{w7} and a short region of Slp^d. The cDNAs for C4 and Slp show extensive sequence homology, but can be distinguished using oligonucleotide probes synthesized to regions of greatest sequence divergence. Sequence differences between C4 and Slp indicate structurally important features of C4 that have been altered in Slp such that Slp is unable to participate in the complement pathway. Of the few nucleotide differences between C4^d and C4^{w7}, a single base change resulting in one less glycosylation site in the C4^{w7} α chain could account for its 4-fold reduced hemolytic efficiency. Sequence comparison of multiple alleles of C4 and Slp indicates that possible gene conversion events occurred in the H-2^{w7} strain that has multiple Slp genes.

INTRODUCTION

The S region of the mouse major histocompatibility complex (H-2) includes the genes encoding two closely related proteins, C4 (the fourth component of complement) and Slp (sex-limited protein), that are distinct in their structure, function and regulation (1). The major site of production for both C4 and Slp is the liver, where they are synthesized as 200,000 MW single-chain precursors that are similarly processed to three-chain disulfide-linked glycoproteins that are secreted into the serum (2, 3). The α , β and γ subunits of C4 have MWs of 98,000, 74,000 and 34,000, respectively, whereas the corresponding peptide chains of Slp are 105,000, 72,000 and 32,000 MW (3). The order of subunits in both precursors is β - α - γ . Despite extensive structural similarities of C4 and Slp (4), Slp shows no complement activity (5) and is regulated by androgens; i.e., it is expressed in adult male but not female mice (6).

Allelic variants exist for both C4 and Slp that show regulatory as well as structural differences in the proteins. C4 serum concentrations vary about 20-fold between mice having either the C4-high (C4^h) or the C4-low (C4^l) alleles (1). Three major alleles have been described which affect regulation of Slp: (i) testosterone-regulated expression (Slp^a), (ii) con-

stitutive expression (Slp^{w7}), or (iii) no expression (Slp^o) (7, 8). Slp expression is found only in some C4^h strains (e.g. H-2^d but not H-2^b) and not in C4^l strains (H-2^k). The Slp^{w7} protein shows some structural variation from the Slp^a protein (as shown by tryptic peptide mapping), in addition to the difference in regulation (9). The C4 allele from the S^{w7} locus also shows structural variation in that while serum C4 levels are equivalent to those in an H-2^d haplotype, the hemolytic efficiency of the C4^{w7} protein is only 25% as great as that of the C4^d protein (10).

The multiplicity of structural and regulatory alleles of C4 and Slp existing in congenic mouse strains permits distinguishing DNA sequences involved in the structure, regulation and evolution of these proteins. Complete human and mouse C4 sequences have been recently derived from cDNA clones (11-13). Partial Slp cDNA sequences have been reported for two different mouse strains that express Slp constitutively (14,15). Here we compare C4 and Slp cDNAs isolated from H-2^d (C4^h, Slp^a) and H-2^{w7} (C4^h, Slp^{w7}) haplotypes. C4 and Slp cDNAs are greater than 95% sequence homologous, but oligonucleotide probes synthesized to regions of greatest sequence divergence can distinguish C4 and Slp, allowing straightforward examination of *in vivo* levels of expression. Regions of the protein essential for C4 hemolytic activity are underscored by sequence divergence between C4 and Slp. Differences between the C4 sequences of the two haplotypes highlight structural variations affecting C4 hemolytic efficiency. Structural differences between the sex-limited and constitutive forms of Slp may or may not relate to differences in expression. Ultimately, direct comparison of DNA sequences from genomic clones and analysis of their expression will allow us to define the cis-acting regions controlling tissue-specific, developmental and hormonal regulation of various C4 and Slp genes.

MATERIALS AND METHODS

Materials

Restriction and modification enzymes were purchased from New England Biolabs, Bethesda Research Laboratories, and Boehringer Mannheim Biochemicals. Reverse transcriptase was from Life Sciences. ³²P-nucleotides and ³⁵S-dATP were from Amersham and New England Nuclear.

Mice

B10.W7R mice (C4^h Slp^{w7}) were originally obtained from A. Ferreira and V. Nussenzweig (Dept. of Pathology, New York University School of Medicine) and have been maintained since in our animal facility. B10.D2 (C4^h Slp^a) and B10.BR (C4^l Slp^o) mice were purchased from Jackson Labs. B6/KH (C4^h Slp^o) mice were the gift of C. Levinthal.

Isolation of RNA

Total mouse liver RNA was prepared according to the method of Chirgwin et al. (16). Poly A+ RNA was purified by affinity chromatography on oligo(dT)-cellulose columns (17).

For preparation of the cDNA libraries, polyA⁺ RNA was fractionated by size through linear (5-20% wt/vol) sucrose-density gradients (18) and RNA >28S in size was collected.

cDNA library preparation and screening

First-strand cDNA was synthesized from size-fractionated liver polyA⁺ RNA of B10.W7R male and B10.D2 female mice, essentially as described by Buell et al. (18). Second-strand cDNA was synthesized by nick-translation repair of the cDNA/mRNA hybrid using *E. coli* RNaseH, *E. coli* PolI and *E. coli* DNA ligase, according to the method of Okayama and Berg (19). Double-stranded cDNAs were blunt-ended with T4 DNA polymerase, methylated with EcoRI methylase, ligated to EcoRI linkers, and inserted into the EcoRI site of the vector λ GT10 (20), which was then packaged *in vitro* and used to infect *E. coli* strain BNN 103 (Hfl). Approximately 7.5×10^5 recombinant phage were obtained for each library.

To screen the B10.W7R library, cDNA was reverse transcribed from B10.W7R size-fractionated polyA⁺ liver RNA to a specific activity of about 10^9 cpm/ μ g and was hybridized to 20 μ g of size-fractionated polyA⁺ B10.BR RNA in 30 μ l of 0.4 M PO₄ buffer for 16 hrs at 68°C. Unhybridized cDNA was collected by passage through hydroxylapatite, and represented 10% of the total cDNA counts. This cDNA was used to probe 50,000 phage of the B10.W7R library on duplicate filters from plates of 10,000 phage/plate (21). Positive plaques were plated at low density and rescreened on duplicate filters, one of which was hybridized to unenriched B10.W7R cDNA made from size-fractionated RNA and the other hybridized to B10.BR cDNA from size-fractionated RNA. Inserts of phage hybridizing to the first but not the second probe, and also hybridizing appropriately on Northern blots (see Fig.1), were subcloned into pUC8 (22) for further characterization. The cDNA libraries were subsequently rescreened using nick-translated probes of identified insert fragments, or oligonucleotide probes that were synthesized using an Applied Biosystems Model 380A DNA synthesizer.

Northern hybridizations

RNA was denatured in formamide, run on agarose formaldehyde gels (23) and transferred to nitrocellulose or nylon filters. Hybridization to nick-translated probes was at 42° in 50% formamide, according to standard conditions (24). Hybridization to oligonucleotides that were 5'-end-labeled with T4 polynucleotide kinase was at 37° in 5 x SSC without formamide; stringent washes decreased to 0.05 x SSC at 50°.

DNA sequence analysis

Restriction fragments were subcloned into M13 vectors (25). DNA sequence analysis was by the method of Sanger (26) using synthetic oligonucleotides as well as M13 primer.

Primer extension

A synthetic oligonucleotide corresponding to 25 nucleotides at the 5' end of pd131 (see Fig. 2) was purified on a 20% acrylamide/7M urea sequencing gel and ³²P-end-labeled to > 10^8 cpm/ μ g using T4 polynucleotide kinase. 5×10^5 cpm of this primer was hybridized to 10

µg of liver polyA⁺ RNA for 16 hrs at 30° in 10 µl of 80% formamide, 0.4 M NaCl, 40 mM PIPES (pH 6.4) and 1 mM EDTA. The hybrids were extended with reverse transcriptase according to the method of Treisman et al. (27) and analyzed on an 8% acrylamide/7 M urea sequencing gel.

RESULTS

Isolation of C4/Slp cDNAs

Two cDNA libraries were constructed in the high efficiency vector, λGT10 (20), from >28S liver mRNA of B10.W7R male (C4^h Slp^{w7}) and B10.D2 female (C4^h Slp^a) mice. The B10.W7R library was screened initially using a probe enriched by subtractive hybridization for C4/Slp sequences, as described in Methods. More than 200 phage showed positive hybridization on duplicate filters. Some of these phage were simultaneously plaque-purified and rescreened by plating phage at low density and by probing duplicate filters with either a positive C4/Slp probe (cDNA from the B10.W7R strain) or a negative C4/Slp probe (cDNA from the B10.BR strain). Recombinant inserts from phage hybridizing with the positive but not the negative cDNA probes were purified, nick-translated, and hybridized to Northern blots containing liver polyA⁺ RNA. C4/Slp clones were identified by their hybridization to a 5.7 kb mRNA (large enough to encode each protein) present in B10.D2 and B10.W7R males, less abundant in B10.D2 females, and only faintly detectable in the B10.BR female. These criteria were satisfied for at least 5 of the first 26 phage rescreened. The hybridization pattern of the first C4/Slp clone positively identified (λW39) is shown in Fig. 1. This hybridization pattern indicated that either all of the first clones contained C4 sequences and an Slp clone had not as yet been isolated, or that the two sequences showed sufficient homology to cross-hybridize under our conditions.

The first identified C4/Slp cDNA, λW39, contained an insert of approximately 600 bp. This insert was used to rescreen 10⁵ phage from each of the two cDNA libraries to obtain longer clones. As none of the clones obtained were long enough to encode full-length C4 or Slp molecules, further screening was carried out using terminal cDNA fragments as probes to obtain clones overlapping the length of the mRNA. Inserts were subcloned into pUC8 and characterized by restriction mapping. Restriction site polymorphisms allowed us to group the cDNAs from both libraries into two classes. Restriction maps of some of the cDNAs obtained from each library are shown in Fig. 2. We suspected one group to be Slp cDNAs since they were found in the H-2^{w7} male library and not in the H-2^d female library (with the exception of pd3, see below). Their designation as C4 or Slp in Fig. 2 is based on DNA sequence and hybridization data discussed below. The direction of transcription for these cDNAs was initially determined by an S1 nuclease analysis of end-labeled fragments hybridized to liver RNA (data not shown).

Figure 1. Identification of a C4/Slp cDNA by Northern blot hybridization. 5 μ g liver poly A+ RNA from several mouse strains were run per lane on an agarose formaldehyde gel, transferred to nitrocellulose and hybridized to the nick-translated insert of λ w39. This phage screened positively with 32 P-cDNA from B10.D2 (C4^hSlp^a) male liver, negatively with cDNA from B10.BR (C4^lSlp^o) female liver, and had a recombinant insert of 600 bp. Conditions of hybridization and washing were standard. H-2 haplotypes are given in parentheses.

Identification of C4 and Slp cDNAs by specific hybridization

The cDNAs of Fig. 2 were subcloned into M13 vectors (26) for sequencing by the dideoxy chain-termination method (25). Though little protein sequence of C4 or Slp was known when this work was initiated, certain structural features could be used as guidelines in the nucleic acid sequence analysis. For instance, as shown in Fig. 3A, four arginines are encoded approximately 1 kb upstream (bases 2560-2571 in clone pd60) of the polyA addition consensus sequence (AATAAA, at base 3525). This is a cleavage signal used in other proteolytically processed proteins, notably murine C3 (28), and indicates this as the putative α/γ junction site in pro-C4 or -Slp (vertical bar at nucleotide 2571 in Fig. 3A). The amino acids following these arginines are 85% in agreement with the partial protein sequence for the NH₂-terminus of the γ chain of human C4 (29), corroborating this sequence as the murine C4 or Slp γ chain.

Sequencing had shown extensive homology with only scattered base changes until a region where pd60 and pw29 diverge substantially from pw16 and pw32, approximately 150 bp upstream of the 4 arginines. In this region (bases 2452-2476), there are 7 mismatches within 25 bases between pd60, or pw29, and pw16. This cluster of base changes corresponds to an area of the α chain that undergoes proteolytic processing in C4 but not Slp (30). This region of H-2^d C4 and Slp (see below for Slp^d) is substantially similar to sequences previously published for C4^{w7} and Slp^{w7} (12,14). Oligonucleotide probes were synthesized for

Figure 2. Restriction maps of C4 and Slp cDNAs isolated from B10.D2 (pd clones) and B10.W7R (pw clones) libraries. cDNAs from each strain for each restriction site polymorphism class are shown aligned with a schematic diagram of the precursor protein and a composite restriction map derived for a full-length C4^d cDNA. In individual clones only sites which differ from the composite map are indicated. Slp cDNAs are distinguished by the lack of EcoRI, PstI and HindIII sites (denoted by black squares) that are present in C4 cDNAs. The circled BamHI site appears in C4^{w7}(pw29) and Slp^{w7}(pw32) but not C4^d(pd60); the circled PstI site appears only in C4^{w7}, 40bp upstream of the conserved PstI site. Identification of each clone as C4 or Slp is from sequence and hybridization data discussed in the text. pd125 is too 5'-ward to hybridize with the specific oligonucleotides but is identified as a C4 cDNA on the basis of sequence data from promoter regions of C4^d and Slp^d genomic clones (data not shown). Restriction sites : BamHI (B), BglII (L), EcoRI (E), HindIII (H), KpnI (K), PstI (P), PvuII (U), SmaI (S), XhoI (X).

bases 2452-2476 (Fig. 3) for the two different cDNAs, kinased to high specific activity, and hybridized to Northern blots of liver polyA⁺ RNAs (Fig. 4). These show specific hybridization patterns that correlate with serum levels of C4 and Slp. That is, Slp is only expressed in H-2^d males and both sexes of H-2^{w7} mice, whereas C4 is twice as high in male mice as in females, and about 1/20th the level in C4-low strains. Thus these probes identify pd60 and pw29 as C4 cDNA clones and pw16 and pw32 as Slp cDNA clones. We have since synthesized another set of C4- and Slp-specific hybridization probes (at positions 494-515, where pw23 differs from pd60)(see Fig. 3B) that correspond to a further 5' region in our clones downstream of the COOH-terminus of the C4a peptide. The C4a peptide is cleaved from C4 upon activation by the complement protein C1s, while such cleavage does not occur for Slp. Nonaka et al. (15) previously pointed out the striking divergence between C4 and Slp in this

region in partial cDNAs from the FM mouse strain (C4^h Slp^a), which exhibits constitutive expression of Slp due to at least 2 loci not linked to the H-2 (31). This set of oligonucleotide probes identifies pw23 as an Slp clone and pd131 as a C4 clone.

Comparison of C4 and Slp cDNA sequences

The entire sequence of pd60 (3555 bp) is shown in Fig. 3A, with functionally important features of the α chain diagrammed in Fig. 3B. This C4^d cDNA includes the coding sequence for the last 87 amino acids of the β chain, the entire α and γ chains, and extends as far as the polyadenylation site. The C4^{w7} cDNA pw29 overlaps entirely with this sequence; the few nucleotide differences are indicated by black dots above the pd60 sequence. Nucleotide differences for Slp^{w7} cDNAs sequenced thus far are shown below the C4 sequence; those base changes leading to an amino acid change are boxed as codons. These clones allow sequence comparison of about 2/3 of the mRNA of C4 and Slp that diverge by about 3%, with approximately half the base changes representing silent mutations. The α chain mismatch alone is actually greater than 4%, while the γ chains of C4 and Slp differ by just over 1%.

Many of the amino acid differences between C4 and Slp fall in regions known to be functionally significant in C4. The most striking cluster of base changes (bases 494-516 of Fig. 3A) occurs after the cleavage signal for the C4a peptide, noted above as the region from which we synthesized a second set of C4- and Slp-specific oligonucleotides. In addition to the mismatch between C4 and Slp at this position, the subsequent 9 nucleotides in the C4 sequence are deleted in Slp. The C4a peptide itself is highly conserved between C4 and Slp, with only 2 amino acid changes in the NH₂-terminus of the α chain. As the COOH-terminal cleavage signal does not differ between C4 and Slp, it is likely that Slp cannot be activated by C1s due to altered protein conformation around the cleavage signal rather than lack of the cleavage site itself.

The C4d peptide-encoding region shows scattered base changes between C4 and Slp, less than half of which are silent. Clone pw32 begins 2 amino acids after the NH₂-terminus of the C4d peptide; thus, we do not know yet whether the cleavage sequence, which is arg-thr in C4^d and C4^{w7}, is the same in Slp^{w7}. However, in the FM mouse strain, the Slp cleavage signal is altered (to gln-met) so that a C4d-like peptide cannot be cleaved from Slp^{FM} (15). As most of the other Slp^{FM} changes also occur in Slp^{w7}, the lack of cleavage of the C4d peptide may be generally typical of Slp. While the thiolester site that is involved in membrane binding is conserved in C4 and Slp (with one silent base change at position 1249), the two amino acids directly preceding it are different (gln-gly in C4 but arg-ser in Slp). However, Slp has been shown to have some ability to transesterify, indicating that these amino acid changes may not be significant with respect to lack of complement function.

Two cloning artifacts were observed during the analysis of these clones, which may provide some insight into the evolution and function of this region of C4. First, some C4

B-

2546

small black circle indicates this base change also occurs in Slp^{w7} (with the exception of the first change at position 42, where Slp has not yet been sequenced). A large black circle indicates that the base differs only in C4^{w7}, whereas the base in Slp^{w7} is the same as in C4^d. β/α and α/γ junctions are marked by vertical bars. Cleavage sites of the C4a and C4d peptides are marked by black triangles. The 4 potential sites of N-linked glycosylation are marked by horizontal brackets. The 13 bp inverted repeats discussed in the text are overlined with double-headed arrows. The stop codon is bracketed. B) Functional regions within the C4 α chain are diagrammed. Glycosylation sites are marked by stalked boxes: open boxes occur in all sequences; the "d" box is specific to C4^d; the "Slp" site occurs in Slp^{w7} but not C4 sequences. "TE" denotes the thiol-ester site. 1 and 2 indicate regions to which C4- or Slp-specific oligonucleotide probes were synthesized. "IR" marks the position of the 13 bp inverted repeats that flank a deletion discussed in the text. A vertical arrow indicates the position of an insertion sequence present in some of the clones.

clones, of both d and w7 haplotypes, had large deletions from about 100 bp 5' of the C4d NH₂-terminus and extending about 560 bp downstream, with the thiolester site centrally located. In comparing these deleted clones to intact C4 cDNAs, we noted several inverted repeats flanking this region. The longest is a 13 bp sequence occurring 12 bp downstream of the 5' deletion breakpoint (base 1010 in Fig. 3) and, in the opposite orientation, 6 bp upstream of the 3' deletion breakpoint (base 1576 in Fig. 3). Several pairs of perfect inverted repeats of 11 base pairs also occur within this cDNA sequence. Given the method by which

Figure 4. C4- and Slp-specific oligonucleotide probes hybridized to poly A⁺ RNA from different inbred mouse strains. Based on a region of greatest sequence divergence between pw29 and pw16, two oligomers 25 bases long were synthesized, differing in 7 bases (see Fig. 3). The oligomers were kinased to high specific activity and hybridized to two identical Northern blots on Gene Screen Plus (NEN). Each lane contains 4 μ g poly A⁺ liver RNA from adult mice, except the B10.W7R mice which were 4 wks of age. The filters were washed in decreasing salt and increasing temperature to a final wash of 0.05 x SSC at 50°C. A) Probed with oligonucleotide from pw29 sequence. B) Probed with oligonucleotide from pw16 sequence.

these libraries were constructed, we think it likely that this region was deleted during second-strand cDNA synthesis, as we have isolated this kind of clone fairly frequently and have not detected a C4 mRNA of 5.1 kb in the cytoplasm. A second type of cloning artifact observed indicates that some intron sequences may have been cloned, as these libraries were constructed from total polyA⁺ RNA. For example, a small insertion of 154 bp has been found in some clones at base 1273 (Fig. 3A), corresponding to 125 base pairs downstream of the thiolester site in both C4 and Slp cDNAs. Slp differs from C4 in 4 of the 9 bases immediately 3' to the site of insertion. That this region is surrounded by both extensive inverted repeats and introns may be indicative of secondary structure involved in mRNA processing, an enhanced rate of evolution for this region of the protein, or even an independent origin of this region early in the evolution of C4.

The difference in α chain molecular weight between C4 and Slp is known to be due primarily to increased glycosylation of Slp (32). Further, C4^{w7} is known to have fewer oligosaccharide complexes than C4^d (32). Though not all glycosylation consensus sequences (asn-x-thr or -ser) are sites of glycosylation *in vivo*, these sequences in our cDNAs correlate well with the known α chain molecular weight differences and with the differential activity of C4^d and C4^{w7}. Two glycosylatable sites are conserved in all our cDNA sequences (Fig. 3B): one occurs just before the C4a COOH-terminus and the other between the C4d COOH-terminus and the α/γ junction. A third potential glycosylation site within the α chain is found in Slp cDNAs but not C4 sequences, due to a single base change (at base 1968 of Fig. 3) of a T to an A (tyr to asn) about 250 bp before the C4d COOH-terminus. A fourth potential glycosylation site of C4^d is altered in C4^{w7} and Slp^{w7} by the conversion of asn to lys by a C to G base change at base 2202 (Fig. 3); this site occurs just before the COOH-terminus of the C4d peptide. Thus, in agreement with protein data (4,30), Slp has more glycosylation sites than C4 and C4^d has more sites than C4^{w7}. Further, this single additional glycosylation group on C4^d may be solely responsible for the four-fold greater hemolytic efficiency of C4^d than C4^{w7} (10). Our sequence comparison of C4^d and C4^{w7} cDNAs shows that they differ by only 8 nucleotides within 3.6 kb. Only 3 of these base changes occur in the α chain, where hemolytic activity resides, and 2 of these changes are silent mutations. The only nucleotide difference resulting in an alteration in the amino acid sequence within the α chain is the additional glycosylatable site in C4^d. In fact, C4^{w7} has no oligosaccharide groups on its C4d peptide. A sequence ambiguity noted by Sepich et al. (12) at base 4257 of their sequence (base 2431 in Fig. 3) is a G in both our C4^d and C4^{w7} cDNAs, and not the A that they had previously published (33).

The sequence of an Slp^d cDNA of almost 400 bp spanning the α/γ junction is presented in Fig. 5 in comparison with our other cDNA sequences and human C4 sequence (11). This clone (pd3) was isolated by using the Slp-specific oligonucleotide (probe 1 in Fig. 3B) to screen the B10.D2 female library. Three positive clones were isolated out of approximately

C4-d:	gly	thr	leu	lys	ile	leu	arg	thr	tyr	asn	val	leu	asp	met	lys	asn	thr	thr	cys	gln	asp	leu	gln	ile	glu	val	lys	val	thr	gly	ala	93	
C4-w7:	GGC	ACC	TTG	AAG	ATC	CTT	CGT	ACC	TAC	AAT	GTC	CTG	EAC	ATG	AAG	AAC	ACC	ACA	TGC	CAG	GAC	CTT	CAG	ATA	GAA	GTG	AAG	GTC	ACA	GGC	GCT		
Slp-d:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
Slp-w7:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
C4-human:	--A	---	C--	---	G--	---	---	---	---	---	---	---	---	---	---	---	--G	--C	---	---	---	---	---	---	---	---	---	---	---	---	---		
				val																													
C4-d:	val	glu	tyr	ala	trp	asp	ala	asn	glu	asp	tyr	glu	asp	tyr	tyr	asp	met	pro	ala	ala	asp	asp	pro	ser	val	pro	leu	gln	pro	val	thr	186	
C4-w7:	GTG	GAA	TAC	GCA	TGG	GAT	GCC	AAT	GAA	GAC	TAC	GAA	GAC	TAC	TAT	GAC	ATG	CCA	GCC	GCA	GAT	GAC	CCC	AGC	GTC	CCC	TTG	CAG	CCT	GTC	ACG		
Slp-d:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
Slp-w7:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
C4-human:	--C	--G	---	A--	AT--	A--	A--	T--C	--G	---	--T	--G	---	?	G--	--G	C--T	---	AAG	thr	lys	---	A--	GAT	--C	--T	C--	---	--C	--G	--A		
				thr	met	ser	glu	asn							asp	asp	asn	leu															
C4-d:	pro	leu	gln	leu	phe	glu	gly	arg	arg	ser	arg	arg	arg	arg	arg	glu	ala	pro	lys	val	val	glu	glu	gln	glu	ser	arg	val	gln	tyr	thr	val	279
C4-w7:	CCC	CTG	CAG	CTT	TTT	GAG	GGT	GCT	AGG	AGC	CGC	CGC	AGG	AGG	GAG	GCC	CCC	AAG	GTG	GTC	GAA	GAG	CAG	GAG	TCC	AGA	GTT	CAG	TAC	ACT	GTG		
Slp-d:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
Slp-w7:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
C4-human:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
										asn																							
C4-d:	cys	ile	trp	arg	asn	gln	lys	leu	gly	leu	ser	gly	met	ala	ile	ala	asp	ile	thr	leu	leu	ser	gly	phe	his	ala	leu	arg	ala	asp	leu	372	
C4-w7:	TGT	ATC	TGG	GGA	AAT	GGC	AAG	CTG	GGG	CTG	TCT	GGC	ATG	GCC	ATC	GCA	GAC	ATC	ACC	CTC	CTA	AGT	GGA	TTC	CAC	GCC	CTG	AGG	GCT	GAC	CTG		
Slp-d:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
Slp-w7:	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
C4-human:	--C	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---		
				his			val											val															

Figure 5. Sequence comparison of C4^d, C4^{w7}, Slp^d, Slp^{w7} and human C4 around the α/γ junction. The sequence printed is 360 nucleotides of C4^d, with the derived amino acid sequence above. Nucleotide differences in the other sequences are indicated, with amino acid changes printed below. The α/γ junction is marked by an arrow. A glycosylatable site is bracketed. A question mark in the human sequence indicates a 3 base gap introduced to allow best alignment with the mouse sequences.

10⁵ phage screened. It is not surprising that Slp is present, although in very low abundance, in this library because female mice of sex-limited strains show 0.2 to 4% the level of Slp expression of the male (34). Slp^d differs from Slp^{w7} in 4 bases within this region, leading to 2 amino acid changes within the γ chain (at positions 355 and 365 in Fig. 5). The 6th amino acid of the γ chain is val in C4^d and human C4, but ala in C4^{w7} and both Slps. It is interesting to note that within this region, the sequence of probe 1 (Fig. 3B) differs substantially not only between C4 and Slp but is also very different in human C4.

5'-end mapping of C4 and Slp mRNAs

To localize genomic promoter regions of C4 and Slp so that putative regulatory sequences may be compared, we have mapped the mRNA 5' end by primer extension. An oligonucleotide corresponding to the 5' terminal sequence of clone pd131 (see Fig. 2) was used as a primer for reverse transcription of liver polyA⁺ RNA from mice of several haplotypes (see Fig. 6). This oligonucleotide hybridizes to both C4 and Slp genomic clones (data not shown), and thus can extend both C4 and Slp cDNAs. The extended products are the same length as seen by the similar gel pattern obtained for samples containing both C4 and Slp mRNA (B10.D2 and B10.W7R males) as for samples with C4 only (B10.D2 and B10.BR females). Both C4 and Slp mRNAs extend approximately 240 bp further than the 5' terminus of pd131. The

Figure 6. 5' end mapping of C4 and Slp mRNAs. A 25b oligonucleotide, derived from the 5' end of pd131, was used as a primer for reverse transcription of the 5' end of C4 and Slp mRNA, as described in methods. 5×10^5 cpm of end-labeled primer was hybridized to 10 μ g of poly A+ RNA from livers of B10.D2 females and males (d), B10.W7R males (w), B10.BR females (k), and a mouse mammary cell line (G) as a negative control. Extended products were run on an 8% acrylamide/7M urea gel, with pBR322 DNA digested with *Hinf*I and 32 P-end-labeled as marker.

cDNA pd125, characterized subsequently to pd131, contains the start site of translation and 42 bases of 5' untranslated sequence (data not shown). We have used genomic clones from the d haplotype to map C4 and Slp transcription start sites by SI nuclease analysis and have sequenced these regions (manuscript in preparation). Both C4 and Slp have 5' untranslated regions of approximately 60 bases that differ only by the deletion of one nucleotide in the Slp sequence.

DISCUSSION

cDNA clones for mouse C4 and Slp will be extremely useful in studies of the structure, regulation and evolution of C4 and Slp genes. While in humans there are two functional

non-allelic C4 genes (11), in mice the gene duplication and subsequent divergence has led to loss of C4 function and acquisition of hormonal regulation in Slp. Analysis of regulatory differences in the androgen-insensitive Slp^{w7} allele is complicated by the occurrence of multiple C4/Slp genes in the B10.W7R strain (35,36). As a first step in understanding this complex gene family, we have isolated cDNA clones for both C4 and Slp from the d and w7 mouse haplotypes. This may allow us to correlate some sequence polymorphisms with functional and/or regulatory features and to examine molecular evolution at this locus.

C4 and Slp coding sequences are remarkably similar, displaying a few percent divergence within the 3 kb that we have compared (more than half the coding region). Several regions of greatest sequence mismatch correlate with regions essential for C4 hemolytic activity, such as cleavage sites for the C4a and C4d peptides. As Slp shows no complement activity, it is not surprising that structural differences may account for its inability to participate in the complement pathway. Although C1s may be unable to activate Slp due to significant amino acid changes, Slp does contain a thiol-ester site that enables it to bind to membranes. Identification of regions of sequence divergence between C4 and Slp enabled us to synthesize oligonucleotides that hybridize specifically and that allow us to distinguish the cDNAs, mRNAs and genomic sequences.

Although Slp has no known function, its sequence seems as highly conserved between some strains as that of C4. Nonaka et al. (15) have compared C4 and Slp from the FM mouse strain within two short regions of the α chain. For these total 825 bp, Slp^{w7} is identical to Slp^{FM} in all but one nucleotide. At that base, 1195 in our sequence (Fig. 3A), there is an A in Slp^{w7} that is a G in Slp^{FM} and the three C4 sequences; thus, this codon is glu in Slp^{w7} and gly in all the others. Further sequence conservation between Slp^{w7} and Slp^{FM} can be seen in restriction site differences noted by Nonaka et al (15). Slp^{FM}, like Slp^{w7}, lacks a PstI site 3' to the β/α junction and a centrally-located HindIII site, both of which are present in C4^d, C4^{w7} and C4^{FM}. The Slp^d cDNA, pd3, also lacks the internal HindIII site. Both Slp^{FM} and Slp^{w7} have an additional BamHI site that is approximately 270 bp 5' to a BamHI site common to all the C4/Slp sequences. It may be that in the duplication of C4 to Slp, certain Slp-specific alterations were introduced that have not yet had time to diverge in separate strains. Alternatively, these Slp-specific differences may be conserved due to a function of Slp yet to be elucidated.

Sequence information now available for three mouse C4 alleles (from the d, w7 and FM haplotypes) indicates a high degree of conservation in the C4 molecules. Of the 8 nucleotide differences we have noted between C4^d and C4^{w7}, the only amino acid change within the α chain is the loss of a glycosylatable site by a single base change in C4^{w7}. Therefore, this single change is likely to be responsible for the 4-fold reduction in hemolytic efficiency of C4 in the w7 haplotype. Within the 3.6 kb that we can compare, C4^{FM} (13) differs from C4^d by 9 nucleotides. C4^{FM} contains all 4 glycosylatable sites of the C4 α chain, including one site

found in C4^d but not C4^{w7} and a second site found in Slp^{w7} but not C4^d or C4^{w7}. It is interesting to note that human C4 also contains both of these glycosylatable sites (11). C4^{FM} is as efficient as C4^d in hemolytic assays (10), suggesting that the one additional glycosylation site in the C4^{FM} molecule does not raise its activity significantly above that of C4^d. Diversity in C4 molecules may offer other advantages for the immune system. For example, polymorphisms may relate to the specificity of the C4 protein in reacting with different invading microorganisms, as has been postulated for polymorphic forms of human C4 (37). Consequently, differences in glycosylation, as observed between C4^d and C4^{FM}, might be involved in recognition of foreign cells.

The availability of sequence data for numerous alleles of C4 and Slp enables us to examine this locus for potential gene conversion, which has been postulated for class I and class II genes of the major histocompatibility complex as a mechanism for generating diversity (38,39). Therefore, C4 allelic differences were compared to determine whether any of these changes were to an Slp-specific sequence, thus indicating recombination between the C4 and Slp genes within a strain. Two examples of possible gene conversion can be found within the sequence shown in Fig. 5, comparing two C4 and two Slp alleles. A convincing candidate for a gene conversion event occurs at bases 245-246 in this sequence (six amino acids downstream of the α/γ junction) where TC in C4^d and C4^{FM} (13) are CT in Slp^d, Slp^{w7} and also C4^{w7}. Because two bases are affected, it seems more likely that there was a single recombination between Slp^{w7} and C4^{w7} rather than the occurrence of two independent and identical mutations. The maximal limits of this putative recombination may be defined by the next Slp-specific differences from C4 that are not present in C4^{w7} (the affected gene). For this example, then, a conversion may have occurred over 42 or fewer nucleotides. A second example occurs at position 150 of the sequence in Fig. 5, where a T in C4^{w7} and Slp^{w7} is a C in Slp^d, C4^d, C4^{FM} (13) and human C4(11). In this case, it seems likely that a mutation in either Slp^{w7} or C4^{w7} was followed by conversion between the two. The subsequent base shows Slp-specific variation, indicating this conversion to be maximally 11 bases. A third example of a potential gene conversion event, again occurring in genes of the w7 haplotype, is at base 1110 of the sequence in Fig. 3A (immediately downstream of the C4d cleavage site). Here, an A in C4^d and C4^{FM} is a G in Slp^{w7}, Slp^{FM} (15) and C4^{w7}. Again, while this simply could be a coincident point mutation in C4^{w7}, it is striking that 4 of 8 nucleotide differences between C4^d and C4^{w7} occur at positions where Slp differs from C4. As relatively little Slp sequence data are available for alleles other than the w7, gene conversion events in the other haplotypes can be discussed only if one assumes that Slp-specific differences are relatively conserved in the other haplotypes. For example, C4^{FM} differs from C4^d in 9 bases, several of which also correspond to Slp changes. In one region (bases 1671-1675 of Fig. 3A), C4^{FM} differs from C4^d and C4^{w7} by 2 bases (3 nucleotides apart), which are both in Slp^{w7}. If these Slp-specific differences are also present in Slp^{FM}, they would

represent another potential gene conversion event over a short region of a C4 gene. C4^{w7} and C4^{FM} differ from C4^d by 8 and 9 nucleotides, respectively. However, only one of these differences is the same in both C4^{w7} and C4^{FM}, which indicates that this base might be a mutation in C4^d. Further, it suggests that the C4 differences that seem to be to an Slp-specific sequence do not simply represent mutational hot-spots that are coincident between C4 and Slp, as one might then expect further coincidence between C4^{FM} and C4^{w7}. Additional sequence data for the various alleles of Slp and C4 are necessary to strengthen the argument for short conversion events between these genes.

If the sequence differences described above do represent gene conversion events, it is possible that they are relatively frequent in the w7 haplotype because there are multiple Slp genes within the S^{w7} locus (35,36) and recombination may be more frequent. It is likely that multiple expressing Slp^{w7} genes are largely similar, as evidenced in part by the fact that restriction site polymorphisms among cosmid clones from this locus are scarce (36). While pw16 and pw32 are identical within the 600 bases that they overlap, they may or may not represent transcripts of separate genes. Further sequence comparison of C4^d and Slp^d will establish whether putative gene conversion events are enhanced by multiple gene copies. In addition, further comparisons of Slp^d and Slp^{w7} cDNAs and genomic sequences may provide information on the differential regulation of these genes.

ACKNOWLEDGEMENTS

We are grateful to Helen Doyle for technical assistance, Craig Squires for advice on use of the DNA Synthesizer, Janet Kurjan for informative discussions of gene conversion, and Mark and Eva Dworkin for comments on the manuscript. Mike Palazzolo and Scott Zeitlin were extremely helpful in the construction of the cDNA libraries. This work was supported by a grant from the NIH to D.R.

REFERENCES

1. Shreffler, D.C. (1976) *Transplant Rev.* 32, 140-167.
2. Hall, R.E. and Colten, H.R. (1977) *Proc. Natl. Acad. Sci.* 74, 1707-1710.
3. Roos, M.H., Atkinson, J.P. and Shreffler, D.C. (1978) *J. Immunol.* 121, 1106-1115.
4. Karp, D.R., Parker, K.L., Shreffler, D.C., Slaughter, C. and Capra, J.D. (1982) *Proc. Natl. Acad. Sci.* 79, 6347-6349.
5. Ferreira, A., Nussenzweig, V. and Gigli, I. (1978) *J. Exp. Med.* 148, 1186-1197.
6. Passmore, H.C. and Shreffler, D.C. (1970) *Biochem. Genet.* 4, 351-365.
7. Passmore, H.C. and Shreffler, D.C. (1971) *Biochem. Genet.* 5, 201-209.
8. Hansen, T.H. and Shreffler, D.C. (1976) *J. Immunology* 117, 1507-1513.
9. Ferreira, A., Michaelson, J. and Nussenzweig, V. (1980) *Immunogenetics* 11, 491-497.
10. Atkinson, J.P., McGinnis, K., Brown, L., Peterlin, J. and Shreffler, D. (1980) *J. Exp. Med.* 151, 492 - 497.
11. Belt, K.T., Carroll, M.C. and Porter, R.R. (1984) *Cell* 36, 907-914.
12. Sepich, D.S., Noonan, D.J. and Ogata, R.T. (1985) *Proc. Natl. Acad. Sci.* 82, 5895-5899.

13. Nonaka, M., Nakayama, K., Yeul, Y.D. and Takahashi, M. (1985) *J. Biol. Chem.* 260, 10936-10943.
14. Ogata, R.T. and Sepich, D.S. (1984) *Proc. Natl. Acad. Sci.* 81, 4908-4911.
15. Nonaka, M., Takahashi, M., Natsuume-Sakai, S., Nonaka, M., Tanaka, S., Shimizu, A. and Honjo, T. (1984) *Proc. Natl. Acad. Sci.* 81, 6822-6826.
16. Chirgwin, J.M., Przbyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) *Biochemistry* 18, 5294-5299.
17. Aviv, H. and Leder, P. (1972) *Proc. Natl. Acad. Sci.* 69, 1408-1412.
18. Buell, G.N., Wickens, M.P., Payvar, F. and Schimke, R.T. (1978) *J. Biol. Chem.* 253, 2471-2482.
19. Okayama, H. and Berg, P. (1982) *Mol. Cell. Biol.* 2, 161.
20. Huynh, T., Young, R. and Davis, R. (1984) in *Practical Approaches in Biochemistry*, ed. Grover, D. (IRL Press, Oxford)
21. Benton, W.D. and Davis, R.W. (1977) *Science* 196, 180-182.
22. Vierra, J. and Messing, J. (1982) *Gene* 19, 259-268.
23. Rave, N., Crkvenjakov, R. and Boedtker, H. (1979) *Nucleic Acids Res.* 6, 3559-3567.
24. Goldberg, D. (1980) *Proc. Natl. Acad. Sci.* 77, 5794-5798.
25. Messing, J. and Vieira, J. (1982) *Gene* 19, 269-276.
26. Sanger, F., Nickson, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci.* 74, 5463-5467.
27. Treisman, R., Proudfoot, N.J., Shander, M. and Maniatis, T. (1982) *Cell* 29, 903-911.
28. Domdey, H., Wiebauer, K., Kazmaier, M., Muller, V., Odink, K. and Fey, G. (1982) *Proc. Natl. Acad. Sci.* 79, 7619-7623.
29. Porter, R.R. and Reid, K. (1979) *Adv. Protein Chem.* 33, 1-71.
30. Chan, A.C., Mitchell, K.R., Munns, T.W., Karp, D.R. and Atkinson, J.P. (1983) *Proc. Natl. Acad. Sci.* 80, 268-272.
31. Brown, L.J. and Shreffler, D.C. (1980) *Immunogenetics* 10, 19-29.
32. Karp, D. (1983) *J. Immunol.* 131, 1405-1410.
33. Ogata, R.T., Shreffler, D.C., Sepich, D.S. and Lilly, S.P. (1983) *Proc. Natl. Acad. Sci.* 80, 5061-5065.
34. Ferreira, A., Eichinger, D. and Nussenzweig, V. (1982) *J. Immunol.* 129, 1506-1508.
35. Levi-Strauss, M., Tosi, M., Steinmetz, M., Klein, J. and Meo, T. (1985) *Proc. Natl. Acad. Sci.* 82, 1746-1750.
36. Robins, D., Malissen, M., Hood, L., Ferreira, A., Walthall, D. and Mitchell, M. (1986) *Mol. Cell. Biol.* 6, 134-141.
37. Porter, R.R. (1983) *Mol. Biol. Med.* 1, 161-168.
38. Pease, L.R., Schulze, D.H., Pfaffenbach, G.M. and Nathenson, S.G. (1983) *Proc. Natl. Acad. Sci.* 80, 242-246.
39. Weiss, E.H., Mellor, A., Golden, L., Fahrmer, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983) *Nature* 301, 671-674.

Note added in proof: A full-length SIp^{w7} cDNA sequence was presented by Ogata and Sepich in December, 1985 (*J. Immunol.* 135, 4239-4244). We believe this may represent a transcript from a different SIp^{w7} gene, as there are more than 30 base differences in comparison to the sequence presented here.