Volume 13 Number 20 1985 Nucleic Acids Research

The genomic organisation and nucleotide sequence of the HLA-SB(DP) alpha gene

Simon K.Lawrance[1], Hriday K.Das[2*], Julian Pan[2] and Sherman M.Weissman[2]

Departments of [1]Biology and [2]Human Genetics, Yale University, New Haven, CT 06520, USA

Received 9 July 1985; Revised 9 September 1985; Accepted 11 September 1985

ABSTRACT

We have isolated a unique fragment of the HLA-DR alpha gene and probed human genomic DNA at low stringency to search for homologous sequences. A minimum of six non-polymorphic cross-hybridizing high molecular weight fragments were found in all DNAs examined. In order to obtain molecular clones of these cross-hybridizing fragments, we constructed lambda and cosmid libraries of human DNA and screened them at low stringency with the HLA-DR alpha gene specific subclone. We have isolated clones corresponding to each of the six fragments and, in this paper, describe those which contain the gene encoding HLA-SB(DP) alpha.

INTRODUCTION

The major histocompatibility complex immune response (class II) region encodes a group of heterodimeric cell-surface glycoproteins which are essential elements of mammalian immunity and which are implicated in many of its failures (reviewed in 1). The polymorphic character of these gene products has allowed the functional, structural, and genetic dissection of this region. The HLA-SB(DP) immune response subregion, originally described by Shaw and coworkers (2), can be detected by the stimulatory effect of these antigens in primed lymphocyte cultures. Its products have been demonstrated to function as restriction elements in the presentation of foreign antigen to T cells (3). Five alleles of SB have been demonstrated (4) and the complex has been mapped approximately 3.3 cM centromeric to HLA-DR (5). Coincident with these functional and genetic studies, the SB subregion has been approached by molecular biologists exploiting the homology between the sequences of the genes encoding these proteins and

those of the DR subregion. First cDNA (6,7), and more recently, genomic clones (8,9,10,11) sharing homology with DR alpha have been isolated. Those encoding SB gene products have been identified by comparison with the amino acid sequence of Hurley et. al. (12). This paper reports the isolation of cosmid and lambda clones encompassing over 70kb of DNA from the SB subregion including individual clones which contain pairs of complete SB alpha and SB beta genes. We have determined the nucleotide sequence of over 14kb of DNA encoding the SB alpha gene. This sequence, together with that of Kelly and Trowsdale (13), completes the sequence of the functional SB alpha and beta genes. Furthermore, we provide evidence for additional SB related sequences and evidence suggestive of novel evolutionary mechanisms operative in this gene family.

MATERIALS AND METHODS

Construction of Libraries of Human DNA

High molecular weight human DNA was isolated by the method of Blin and Stafford (14) from the HLA-hemizygous lymphoblastoid cell line 3.1.0 (HLA-A2, B27, Cw1, DR1, SB4) (15) kindly provided by Dr. D. Pious. Following partial Sau3A digestion, fragments of 15-25kb and 30-50kb were fractionated by sucrose density centrifugation. Appropriately sized fragments were ligated into the cosmid vectors pJB8 (16) or pCOS2 (17), or the lambda vector charon 28 (18) with T4 DNA ligase (New England Biolabs). Recombinant DNAs were packaged into lambda phage particles utilising freeze-thaw and sonic extracts prepared by the method of Hohn (19) from E.coli strains BHB2688 and BHB2690. Host cells for lambda libraries were LE392 and for cosmid libraries, HB101. Recombinant clones were obtained at efficiencies of 10^5 cosmids and 10^7 phage/microgram insert DNA. Lambda clones were also obtained from the HaeIII-AluI library of human placental DNA constructed by Dr. T. Maniatis (20).

Sources of Probes

DNA fragments for Southern blotting, library screening, and sequencing were isolated by subcloning into plasmid vectors pBR322 or pBR328 and subsequent extraction from low melting temperature agarose. In order to completely eliminate contaminating vector sequences from probes used to screen libraries at low stringencies, fragments were subjected to an additional subcloning into lambda gt WES (21) followed by isolation of the fragment from low melting temperature agarose. DR alpha like sequences were detected using the 3.1kb EcoR1-EcoR1 fragment of a DR alpha genomic clone which we have descibed previously (22,23). DR beta like sequences in the lambda and cosmid clones were detected using the DR beta cDNA probe (24) kindly provided by Dr. H. McDevitt. DNA fragments were nick translated to a specific activity of greater than 10^8 cpm/microgram.

Restriction Mapping of Clones

Restriction maps of lambda clones were generated by probing Southern blots of partial digests of lambda DNA. Restriction maps of cosmid clones were determined by probing blots of single, double, and triple digests of cosmid DNAs prepared by standard procedures.

DNA Sequencing

Libraries of DNA fragments for sequence analysis were constructed by sonication of intact plasmid subclones, followed by ligation of S1 treated, Klenow filled, size selected fractions into SmaI-cut phosphatase treated M13mp8 (Amersham). Recombinant DNAs were transfected into JM101 and clones for template preparation were selected by hybridization of appropriate probes to plaque lifts. Templates were prepared by published methods and sequenced by the method of Sanger and Coulson (25). Electrophoresis in Hong's buffer gradient gels (26) with slight

modifications enabled over 500 nucleotides to be determined from a single template with no detectable loss of accuracy. Overlapping sequences were aligned using the DBCOMP and DBUTIL programs of Staden (27). The resulting database consisted of 247 templates and 59507 determinations (average = 241 determinations/template). Each of the 14646 assigned nucleotides was determined from at least three templates (average = 4.1 determinations/nucleotide). The juxtaposition of the three sequenced subclones was confirmed by subcloning and sequencing the overlapping HindIII-BamHI and HindIII-HindIII fragments as indicated in figure 1. The resulting sequence was analysed using the Wisconsin group of programs (28).

RESULTS AND DISCUSSION

We have previously described the isolation of a lambda phage clone of HLA-DR alpha utilising the primer extension method (29,22,23). Hybridization of subclones of this clone in whole genome Southern blots indicate that while both the 4.4kb EcoRI-EcoRI and the 1.9kb SstI-EcoRI fragments containing the first exon and intron of DR alpha include sequences highly repeated in the human genome, the 3.1kb EcoRI-EcoRI fragment containing the second through fifth exons hybridizes to a single band under conditions of high stringency (0.1xSSC, 68°C). Under lowered stringency conditions (3xSSC, 50°C), the 3.1kb fragment, when labelled to high specific activity, hybridizes weakly, but distinctly, to a minimum of six nonpolymorphic Eco RI bands of greater than 5kb in a panel of DNA's from five HLA-nonidentical cell lines.

Lambda and cosmid libraries of human DNA were screened at low stringency with the 3.1kb EcoRI-EcoRI fragment in order to obtain molecular clones of these fragments. Two strongly hybridizing and six weakly hybridizing clones were isolated from a cosmid library of HLA-hemizygous 3.1.0. DNA containing

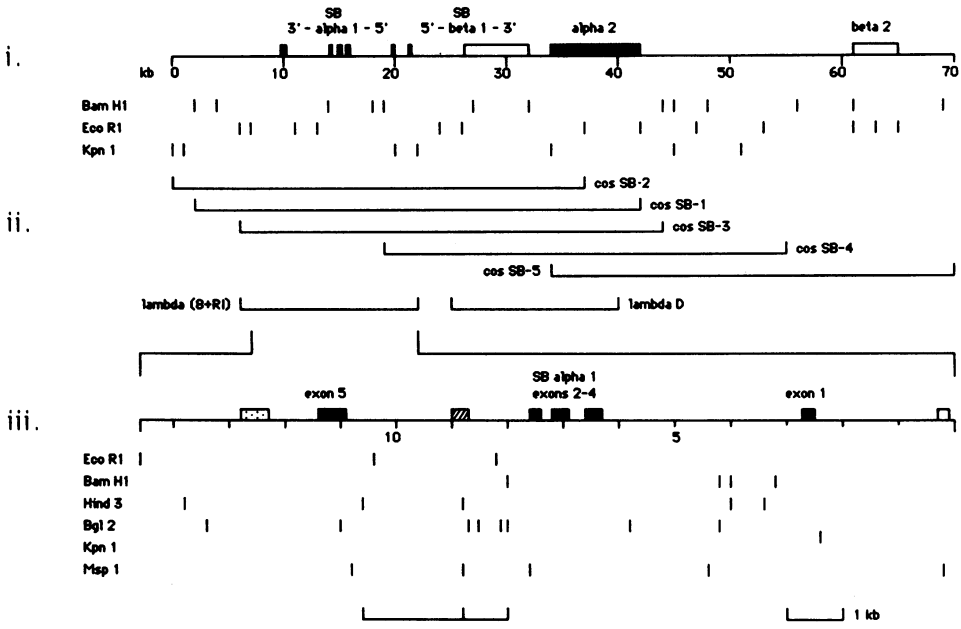

Figure 1. Molecular map of the HLA-SB(DP) subregion. (i) Restriction map of 70kb from the SB subregion. Closed boxes indicate regions of hybridization with the alpha probe and open boxes indicate regions hybridizing with the beta probe. Individual exons and introns are indicated where determined by sequence analysis. (ii) indicates the extent of the individual overlapping cosmid and lambda clones. Clones cos SB1-SB5 and lambda D were isolated from libraries of 3.1.0 DNA, lambda (8+RI) was isolated from the library of human placental DNA (20). (iii) Restriction map of the sequenced 11kb, 4kb, and 2kb EcoRI-EcoRI fragments containing SB alpha. The open box indicates the first exon of SB beta, the closed boxes indicate the exons of SB alpha, the striped box indicates the position of the region of sequence sharing homology with the IgC epsilon genes (see text and figure 3), and the stippled box indicates the location of a Kpn element. The solid bar at the bottom shows the position of the HindIII-HindIII and HindIII-BamHI fragments sequenced to confirm the alignment of the three EcoRI-EcoRI fragments.

approximately 500,000 clones. An additional five weakly hybridizing clones were isolated from the lambda libraries. Five of the weakly hybridizing cosmid clones and two of the lambda clones were found to overlap in a single group, as determined by restriction fragment analysis (figure 1). As indicated in figure 1, both the 3.1kb EcoRI-EcoRI fragment of DR alpha and the DR beta cDNA probe hybridize to two regions within the cluster.

Linkage of the alpha-1 and beta-1 genes was confirmed by DNA sequencing (see below). The relative direction of transcription of these two genes was determined as head to head by comparing the hybridization patterns of the 4.2kb MspI-MspI fragment isolated from lambda (8+RI), which contains the first exons of both genes, to the patterns obtained with the DR alpha 3.1kb EcoRI-EcoRI fragment and the DR-beta cDNA probe. The completeness of the SB alpha-1 and SB beta-1 genes within individual cosmid clones (cos SB1-3) was determined by sequence localization of the 3' untranslated exon of SB alpha and by the presence of additional alpha hybridizing sequences downstream from the SB beta-1 gene. The alpha-2 and beta-2 regions presumably correspond to the SB alpha pseudogene described by Servenius et. al. (10) and the SB beta pseudogene described by Kappes et. al. (30). The overall arrangement of the cluster is in close agreement with the recent data of several groups (8,9,10,31).

The DR alpha hybridizing 11kb EcoRI-EcoRI fragment of lambda (8+RI) was initially selected for sequence analysis. Templates constructed from this fragment and selected with the DR alpha probe corresponded predominantly to the highly conserved third exon (encoding the alpha-2 domain). Comparison of this sequence with the SB alpha cDNA sequence (6) allowed the preliminary identification of the clone as SB alpha. As shown in figure 2, the complete sequence confirms this identification. The translated genomic sequence is identical to the SB alpha cDNA sequence of Auffray et. al. (7) and differs by one nucleotide in the fourth exon from that of Erlich (6). The untranslated sequences are also identical with the exception of a dinucleotide insertion, TC, in the 3' untranslated regions of both our and Auffray's sequence relative to Erlich's. These identities, together with the absence of restriction fragment length


-ProThrAlaArgGlyGlnVal
ValSerThrLeuLeuValMetLeuLeuAlaThrLeuAlaValThrArgProAlaAlaSerValGlnLeuValMetMet
MetArgProGluAspArgMetPheHisIleArgAlaValIleLeuArgAla
LeuSerLeuAlaPheLeuLeuSerLeuArgGlyAlaGlyAlaIleLysA
laAspHisValSerThrTyrAlaAlaPheValGlnThrHisArgProThrGlyGluPheMetPheGl
uPheAspGluAspGluMetPheTyrValAspLeuAspLysLysGluThrValTrpHisLeuGluGluPheGlyGlnAlaPheSerPheGluAlaGlnGly
GlyLeuAlaAsnIleAlaIleLeuAsnAsnAsnLeuAsnThrLeuIleGlnArgSerAsnHisThrGlnAlaThrAsnA
(a)
(b)
(c)
(d)
(e)
(f)

```
                    spProProGluValThrValPheProLysGluProValGluLeuGlyGlnProAsnThrLeuIleCysHisIleAspLysP
6901 TGGTGCTGTCTCCTTCCAGATCCCCCTGAGGTGACCGTGTTTCCCAAGGAGCCTGTGGAGCTGGGCCAGCCCAACACCCTCATCTGCCACATTGACAAGT
     hePheProProValLeuAsnValThrTrpLeuCysAsnGlyGluLeuValThrGluGlyValAlaGluSerLeuPheLeuProArgThrAspTyrSerPh
7001 TCTTCCCACCAGTGCTCAACGTCACGTGGCTGTGCAACGGGGAGCTGGTCACTGAGGGTGTCGCTGAGAGCCTCTTCCTGCCCAGAACAGATTACAGCTT (g)
     eHisLysPheHisTyrLeuThrPheValProSerAlaGluAspPheTyrAspCysArgValGluHisTrpGlyLeuAspGlnProLeuLeuLysHisTrp
7101 CCACAAGTTCCATTACCTGACCTTTGTGCCCTCAGCAGAGGACTTCTATGACTGCAGGGTGGAGCACTGGGGCTTGGACCAGCCGCTCCTCAAGCACTGG
     G
7201 GGTATGCAACTGCTTTTCTCTCCATAATCTCCTGGCATCCTCTATTCCAAAGACCTGGTGTCCTCTGCACCAGCTTTCCGCACTGGCTGGGTCTCAGTCC
7301 TCTCCTCGTCCTAACATCCAATTAACTGGTCCATAACCTTCAATTCCCACAACCATCCCAGGCCATCACCACCCTCACTGCACCCTCCTGACCCTATCTCT
                  luAlaGlnGluProIleGlnMetProGluThrThrGluThrValLeuCysAlaLeuGlyLeuValLeuGlyLeuValLeuGlyIleIl
7401 TCATTCTTCCCCCAGAGGCCCAAGAGCCAATCCAGATGCCTGAGACAACGGAGACTGTGCTCTGTGCCCTGGGCCTGGTGCTGGGCCTAGTCGGCATCAT (h)
     eValGlyThrValLeuIleIleLysSerLeuArgSerGlyHisAspProArgAlaGlnGlyThrLeu ✲
7501 CGTGGGCACCGTCCTCATCATAAAGTCTCTGCGTTCTGGCCATGACCCCCGGGCCCAGGGGACCCTGTGAAATACTGTAAAGGTGGGAATGTAAAGAGGA
7601 GGCCCTAGGATTTGTAGAATGTAAGGAAGGGAGGAAAAATTCAATCTGATAAGTGTTCATTGATCTTCTAATGGGTTAAAAGCATTCAGCCACATAACAA
7701 CAACAACACCGATAACTAACTGAGTAGTTAATATGGTCAGGCGCTATTCTGAGGATTTACATTTATTAACTCACTTTATTCTCACACATAGTCTTTGAGG
7801 TAGGTACTATTATTTTCACTATTTCACATGAGAGATACTTACATCTTTTTACATACACAGAGACTTTAAGCACTTTGATCAAGTTCCCACAGCTATGAAG
7901 TAGTAGGGCTAGCTTCCAATCCAGAAAGTCTGGATCCAAGACTGTTTATCCACTGTCCTATTCACCCTATTTTGTGAAGGAAAAGACCAAGTTCAAATTC
8001 TCCAGAGTCCATTGCCAAATAATGGAGTCAGATCTATATTTCTATACATAATTACAACACAGTGTGGTGGGTGCCTGTAACTACTTACTGTCTCTACTTG
8101 GACTCATTCCATGGCAATGTTCACACAAAAAATGCCCCTCCAGAGATCTTACAGGTTTCTATTTATCATAACACTCACCATGCTTTATATTTTTATATGT
8201 TTTGGGAATTCTCTTAGCATTAGACAGTGAACTTCCATGCAGATGACCACATCTAATTCATTATTATTATTGTTATTCATGCTGGACCTCAGGTACAAAA
8301 GGTTAAGAACTTCTCAGTTCATTATATGATCATCATTGGTGCCTCCGAGCTCTCTCTCTCCCTTGATTTATTTGGTCCCTTTATCTCCAGTCCTTAC
8401 TCCCATATCTAACCTCTTACCCCTACCTCATAGGTAAACATTTTAATGAATTTGATGTTTCCTTTTATTTGCATAGATCCTCTGTAATATGTAGTAGTGT
8501 CCAGTGTACATGTATTTTTAATTAACCAAAATGGCATTAAATTATAGATCTAATTTTGTACATCCAGTTTGTTTCTTCCAAATCTTCCATAGTATTTTAC
8601 |TTTATATGTCCATGCATTAGTCCATTTTGCATTGCTATAAAGGAATATCTGAAGTTACCTAATTTACGAAGAAAAGAGCTTTAAATGGCTCACAGATCTG|
8701 |CAGGCTGTACGCGAAACATGGCACTAGCATCTGCTTCTGTTGGGGGATTCTGGAAGCTTTTACTCATGGTGGAAGGCAAGTGGAGCCAGTGCATCACATG|
8801 |GTCATAGAGGGAGAAAGAGACATAGAAAGAGGTGCCAGCCTCTTTTTAACAACCAGGTTTCATGTGCACTAATAAGATGAGAACTCAACTCATTACCCGGA| (i)
8901 |GAGGGGACAAAGCCATTCATGAGGGTCTCCTCCATGATTCAAATACCTCCCACCAGGCCCCACCTCCAACACTGGGGATCAATTTTCAACATGAGACTTG|
9001 |GAAGTGACAAATATCCAAATCATATTAATCCACATATCTACATTGCTCCTGGGATACCTGGATCATTCCTGGTTCTCTACTATTGCAAGCAATGCTTGTA|
9101 TCTCACATGGAACTGCATATACATGTGGGCCTGACCTGCATCCCTGGAATGTATGTATCCTAGAAAGGGGTTGCAGGGTTGCTGGAGATGCAGCTCCTTA
9201 ATTTGACTAAACACTGCTCATCTTCTCATCAGAATGGCTGTACTCATCTGAACTTCCTTTGTCAGTACTCTAATTGTCCTGCAACTCCTAAATGGACTTC
9301 AACACTGGACATTATCCAGTTTTCTAACTTTTGCCAATTTCATGTGCATAAAGAAATATGCTGTTTTATTTTGCATTTCTTTAATTACTAATAATTGGGG
9401 CTATAATTAGGACTGATTAGCCACTTGGGGGTTCCTTTTCTATAAATTGCCTGTTCACATTCATTGTCCATTTTTGTACTATGTGCTTCCATCATTTTCT
9501 TATTGATTTGCAGGTGATCCTTATATAGTCCTGCTAGTAGTCCCTTGTCAGTTTAGGCATTGCAAATGTTTTCCTCTAATCTGACTTCTGGCAACTGTC
9601 TCCTTGGTTTCCTTTATTGAAGAGAAATCCTTAATATTTTGTAATGAAGTCCATCAACTGTATTTTTGTTGTGTCTTTTTTAAAAGAAGTCTTCCCT
9701 ATACTGAGATATCAAAGATACTCTTAAAACATCTCCTACAGTTTTAAATTTCACATTTACTACTTTAATTCATCTGGGATTCATCTTTGTGTTTGATGGG
9801 GATCATGTTTTATTTTTCTTTATATAATGGGCCAGTGTGTTCCCACAACTACTAAATAGTTCACCTTTTCCCCATAGGTTAGTAGTGTCTCCTTTGCTAT
9901 ACTGAAAGCTCCCATTATAGGTGGGCCTGTGTCTGAGTTCCATCTTGTTCCACTGTTCTGTTTGTCTCTTCTTGTGCCAGTGTCCTAGTATTTTGATTAC
10001 TATGACATTGTAGTGTGTGTTAGTATCCAGTAGGACAAATTCTTGTTTATTTTTCTTAGTTCACACACATTTATAATTATATCTATAATGATTTGTAACA
10101 GAGTGAAGTGAATGTAGAATGTCAGATGTTAAGAGGAAGAATGGAAAAGAGGGCTGGGACTAGGGTGATGTAGGGGATGCACCTGGCTTAGGTGCAAAAT
10201 TTGGGGGATACCAAAAGAACTCAGTAATAAXTCATATTTTAATGAAATATCTTGAAAAGGCAAAATTAATGCAAAGATACATGATTAACAAAACATCCAA
10301 AGAGGAGTATTTAACAAAAATGGAGAAGCAGAGAAGCAGAAGAATTAGGAGAATATGCTGTCACATGAGCCAAGGAATTAAAGAATTCAGGAAGGAGGAA
10401 GTACTGCTGTCAGATGTTCAACAGAGGTCATTTTAGAAAATTTACCTTGGTTTTTGAAATCCTTTCAAAGAGCAGTATACACAATGTGAGCAAGTATCCT
10501 TCGTTCATTGCCGTCATTGATATGGTTTGGATATTTGTCCCTTCCAATTTCTCATTCCAGGGTTAAGCTTCTTCTCTGCCCTCAGTAATGTGGCCCTTCCC (i)
10601 CTTGTCTGTATATTTTGGAGACATGAAGCATGTGGGATGGCCTCACAGTCAGCTGGGGTTTGAGGGTGAAATTCAATGACTTTCGTGAACTCCTTGGCTC
10701 CTATGTGCTCTTCACTGGAGGACCAGGGCATGTGCAGGGATGACCACCTTCTCCCTGGGACCTGAACAGGGCAGAGAAATGGGAAGCTCGGGTGCAAAGG
10801 GAGTGGGGAAGATGGGTCCGGGCTTACAGTACTGAACCCAGGAATGACAATAACTGTGTGTGTTGCTGCAGGTGACAAAATATCTGAACAGAAGAGGACT
10901 TAGGAGAGATCTGAACTCCAGCTGCCCTACAAACTCCATCTCAGCTTTTCTTCTCACTTCATGTGAAAACTACTCCAGTGGCTGACTGAATTGCTGACCC (j)
11001 TTCAAGCTCTGTCCTTATCCATTACCTCAAAGCAGTCATTCCTTAGTAAAGTTTCCAACAAATAGAAATTAATGACACTTTGGTAGCACTAATATGGAGA
11101 TTATCCTTTCATTGAGCCTTTTATCCTCTGTTCTCCTTTGAAGAACCCCTCACTGTCACCTTCCCGAGAATACCCTAAGACCAATAAATACTTCAGTATT
11201 TCAGAGCGGGGAGACTCTGAGTCATTCTTACTGGAAGTCTAGGACCAGGTCACATGTGAATACTATTTCTTGAAGGTGTGGTTTCAACCTCTGTTGCCGA
11301 TGTGGTTACTAAAGGTTCTGATCCCACTTGAACGGAAAGGTCTGAGGATATTGATTCAGTCCTGGGTTTTTCCCTAACTACAGGATAGGGTGGGGTAGAG
11401 AAAGGATATTTGGGGGAAATTTTACTTGGATGAAGATTTTCTTGGATGTAGTTTGAAGACTGCAGTGTTTGAAGTCTCTGAGGGAAGAGATTTGGTCTGT
11501 CTGGATCAAGATTTCAGGCAGATTAGGATTCCATTCACAGCCCCTGAGCTTCCTTCCCAAGGCTGTATTGTAATTATAGCAATATTTCATGGAGGATTTT
11601 TCTACATGATAAACTAAGAGCCAAGAAATAAAATTTTTAAAATGCCCTAATTCATTGCAATTTTTACCAGCCATAGTCACTCCATGTGGGAGAACTTAAA
11701 TCATGATTACCAGAGCTTTCAAAGGTTTGAGAATAGTGATGATTATGAAGAAAAATATCTTATTTGAGCAAGGATTTTGTTTCTTTATGAGTGTTCATTA
11801 GATATTACGATGAAAAAGCATGAAATGGTAAAAATTCAGATAAATATAAAAACATGTTCTCTAGTTTTTTTAAGTTAAAAAAGGAATTGTTTAAAGTA
11901 AAAATTATTTGGGGGTTTATAACATACCCAGAAGTAAAATATGATGACAATGGCACAAAGAATAGAAGGGAGAAATGGAAGTATAATGTTGTAAGTTTCT
12001 TATACATGTTAAGTGGTGTGTTATTATTTGAAGGTAGAATGTATTAAGATGAATATTTTAAGCTCCTGATAACTATTGAAAAAAAAGAGGTATAGCCAA
12101 GAGGCCAATGGAGAAGATAAAATAGAACACTAAGCATAATTAATTCAAAATAAAGAAATAAAAAAGGGAAAGTCTGGTAAGACAAAAAGAAAACAAACTG
12201 TAAGATGGTAGAGTTTAAAACAACCATACTAATAATTGAATTAAATGCACATGGCCTAAATATTCTAATGAAAAGGGAAAGATTGTCAGAATGCACAAAA
12301 |AAATCTACAGGCCAACTTCATGCTCTCTACATAATGCCCTCTTTAAATATGAAGGCAAAGACAGGTAAAAAGTAAAAGAATGGGAAAATCACATGTATACC|
12401 |GTGGAATGCTATGCAGCCATAAAAAAATGAGTTCATGTTGTTTGTGGGGACATGGATGAAGCTGGAAGCCATCCTTCACAGCAAACTAACACAGGAACAG|
12501 |AAAACCAAACACCACACGTTCTCACTCGTAAGTGGGAGTTCAACAATTAGAACACATGGACACAGGGAGGGGAACACCTCACACCAGGGTCTGTCAGGGC| (k)
12601 |ATGGGGAGCAAGGGGAGGGAGAGCATTAGGACACATACCGAATGTATTGCATGGCTTAAAACCTAGATGATGGGTTGATAGATGCAGCAAACCACATGGC|
12701 |ACATGTATAACTATGTAACAAACCTGCACATTCTGCACATGTATCCCAGAACTTAAAGTAAAAAAAAAAAAAACGAAAATAATGCCAACCATGGAAGTAT|
12801 TGGTGGCTGTGTTAATATCAGAAATATAAGACTCAGAAATATTACCAAGGAGAAAGAAGGATAGTTCATAATGATAAAAGATCATTTTATTATCAACATA
12901 TAACAATCCTAAATGTGTTTGTTCTAGAAAATATGTCTTAAATCACATTATACCAAAAATGATAAAAATAAATCAGAAATAGACAAATTCACAATTATAT
13001 TTTAGTATTCTAGCACTCAGTAAACAATAAAATATTTAGGAAAAAACTTCATGAGGACATGATAGATTTAAATAACATTATCAATGAACCAACGTGATCT
13101 AATCAAGATCTGTAGAATATTCCACCCAATAGTGGCAGAATACACATTATTTTCAAATGCTCAAGAATATTCCACAGGACAGACTATACACTGGGTCATA
13201 AACACATATAAATAAATGTCTCTATTGAAATCATACAGAATACATTCTCGGACCACAATAGCATTAAATTAGAAACCAATAACA
13301 GAAAAATACCTTGAAAGTCCCAAATACCTAGAAATTAAAAAGTATACTGCTAAACAGCACCTGGATTTAAAAAGAGTCAGAAGGAAAATTAGAAAATATT
13401 TTGAACTGAGTGAATATGAAAGCACACTATCAAAATTAGTATGATACACTAATTACACAATAATTTATAACTTTACATAATTGGAAAGGAGGGAAACTCT
13501 AAAATCAACCATCTATGTTCCCATCTTAAGAAGCTAGAAAAAAAAGTCAAATGAATCCCAAGATTAATAAGATCAGAAATAAATGCAATAAAATGGACA
13601 AACAATAAAGAAAATAAACAAAGTCAATTGCTGGTTTTCAATAAGGCTCAATACATTCATGAATCTCTAGGTAGATGGATCAAGAAAAAGAGATAAGACT
13701 CAAATCCCCAATATCAGAAATGGAAGTGGGTACGTCACAACAAATCATACAGACATTAAAAGTATTATGACAGAATGCTATGAAATTGCCAATAAACAAA
13801 AATGACAATAAATTTGACAATTTACATTGTTAAATTAAATTAAATTTGGTGTAAAGCTTTCTCCATATCTTAAATTCCTACATAGCAAACTAACCCAACT
13901 TAACATAACTGCGTTATGCAAACAAACTACAGCCTAACTTAAGAGTGTTGTAATAAATAGCTGAGTCTCAGCCAATCACAGGCTGCCAAGTGATCATATT
14001 ATGTCCCCCATAAGCAAATCCCTCATCACGCCATGCCCATATAAGGCAAACACTGAGCTGTAATAAATTGGCTGGTTTTGAGTATCACTTCCTGTTTTTA
14101 TCTATAAACACTGCCTTCACATGTTGCTGGACAGAGCTTTCTGAATCTTTCTGGGTTCTGAGGGCTCCCCAATTCATGAATTGTTCTTTGCTAAATAAAC
14201 TCTGTTAAATTCAACTTCTCTAAAATTTTATTTTAACAACATGAAATAGAAACATTTCTTGAACGATTCAAATTACCAAAACTAAATCAAGAAGAAATC
14301 TCTGTCAATGGAAGAAATTAAGTTTGTAATTTTAAAAATCCTTCTCACAGAGAAGGCCAAATGGTTTCCAGGTCAGATGGCTTCATTGATTAATCCATAG
14401 TATTCTGTCATAATACTGTCATAATACTTTTAATGTCTGTATGATTAATTCTATTATACATTTAAGGAAGAAACAATACAAACTCAATACAAACTCTTTC
14501 AACAAAAGAGCAGAAGAAAACACATCCCAATAATTTACAAGTCCAATATACTAACATCAAACAATGACATTAAAAGAAAAGGAAACTAAAGACACACAAA
14601 CATAGCCATGAAAATGTTTAACAAAATATTTTTTTAAAATTGAATTC
```

polymorphisms between the alpha regions of the 3.1.0 cosmid clones and those of other haplotypes (8,9,10,31) further attest to the extremely limited polymorphism of the SB alpha immune response genes.

The complete nucleotide sequence of the SB alpha gene is presented in figure 2. The overall organisation of the gene is analagous to other alpha genes with a large (3584bp) intron separating the leader peptide and first two codons of the mature protein from the tightly linked second, third, and fourth exons. A comparison of intron and exon sizes of available alpha genomic sequences is presented in table 1. The conservation of intron sizes, with the notable exception of SB alpha intron 4 (see below), is impressive in these otherwise considerably diverged genes. A similar conservation of intron and exon sizes has been observed in other gene families such as the globin genes (reviewed in 32).

Our results confirm the tight linkage of the SB alpha and beta genes, as well as the presence of several putative

Figure 2. The complete nucleotide sequence of the HLA-SB(DP) alpha gene from the lambda clone (8+RI) including the alpha-beta intergenic region and the first exon of the SB beta gene. Letters indicate: (a) the first exon of SB beta, as indicated by Kelly and Trowsdale (13); (b) the locations of the beta consensus sequence, the single nucleotide substitution and the beta proximal 13bp inverted repeat; (c) the analagous alpha consensus sequence, single nucleotide substitution and inverted repeat; (d) the first exon of SB alpha as determined by comparison with the cDNA sequence of Erlich (6). A leader peptide of 31 amino acids is indicated although a shorter, 25 amino acid leader may be utilised. As is the case for other related alpha genes, the first two amino acids of the mature protein are also encoded in this exon. (e) indicates the position of sequences of unknown origin or function involved in a complex 27bp inverted and 19bp direct repeat; (f) the location of the second exon encoding 82 amino acids of the alpha-1 domain of the mature protein; (g) the third exon encoding 94 amino acids of the highly conserved alpha-2 domain; (h) the fourth exon encoding 51 amino acids of the transmembrane and cytoplasmic domains. The box at (i) indicates the location of the sequences sharing homology with the IgC epsilon genes (see figure 3). The 39bp inverted repeats are indicated by the arrows. (j) indicates the position of the fifth exon encoding the 3'-untranslated region as determined by comparison with the cDNA sequence (6). The box at (k) indicates the position of a repetitive element of the Kpn family.

	E1(a)	I1	E2	I2	E2	I3	E4	I4	E5(b)
SB	82	3584	246	340	282	214	167	3289	312
DR	82/50(c)	2200*	246/68	491	282/75	288	166/61	738	263(366)/46
DX	82/65	3500*	249/61	250*	282/74	200*	175/60	-	-
DC	-	-	-	-	282/72	-	175/59	-	-
I-E	82/57	2242	246/61	487	282/64	537	169/59	626	312/38

Table 1. Intron/exon sizes and % homologies of genes related to HLA-SB(DP) alpha. (a) values omit 5' untranslated sequences. (b) values from exon 5 start to first nucleotide of polyadenylation signal. (c) nucleotides/%homology with SB alpha. Data compiled from the NIH-GenBank database (28).

regulatory elements, including the alpha consensus sequence (c in figure 2) noted by Kelly and Trowsdale (13). This consensus sequence of 54 nucleotides is 69% homologous to a sequence located at an equivalent position immediately upstream from the DR alpha gene. Our sequence of the 2.3kb alpha-beta intergenic region is identical to that of Kelly and Trowsdale (13) with exception of substitutions of C>G at position 457 and G>A at position 2263. These substitutions are notable in that they occur in analagous positions between the perfect 13bp inverted repeat common to the two SB genes and the alpha and beta consensus sequences immediately upstream of each gene (b and c in figure 2). Bestfit analysis (28) of the beta-proximal half of the intergenic region with the alpha-proximal half selects an imperfect 23bp inverted repeat that includes these two nucleotide positions, further emphasizing the possible biological significance of these intergenic polymorphisms. We have not detected any additional sequence elements shared between this region or other noncoding regions in the SB alpha sequence and other sequenced immune response genes.

As noted above, the sequences five prime to the DR alpha gene are highly repeated in the human genome. In contrast, the sequences five prime to the SB alpha gene are not. In fact, hybridization of the 4.2kb MspI-MspI fragment containing the first exons of SB alpha and SB beta and the intergenic region in

whole genome Southern blots reveals the presence of three strongly hybridizing high molecular weight (7-12kb) Eco RI bands in DNA from the 3.1.0 cell line, at least one of which is polymorphic. One of these bands is accounted for by the SB intergenic region and another by the second region of hybridization of the 4.2kb MspI-MspI fragment in the cosmid cluster near the beta-2 gene (see figure 1). The third band remains unaccounted for and, consequently, suggests the presence of another alpha, beta, or intergenic region closely related to the SB complex in the human genome.

Hybridization of nick translated human DNA to restriction digests of the cosmid and lambda clones indicates that highly repetitive sequence elements are interspersed throughout the 70kb SB subregion. Analysis of the available genomic sequence indicates the presence of several Alu and Kpn family repetitive elements (13) as well as intermediately repetitive elements (9) within the introns of SB beta. In contrast, the introns of SB alpha appear to be free of highly repetitive sequences as determined by hybridization and sequence analysis, suggesting the possible restriction of such elements to the polymorphic members of this gene family. A Kpn family repetitive element is located approximately one kilobase downstream from the fifth SB alpha exon (figure 2).

In contrast to the 3.1kb EcoRI-EcoRI fragment of DR alpha which contains exons two through five and is essentially unique in the genome, the roughly analagous 4.8kb HindIII-HindIII fragment of SB alpha, which contains exons two through four, appears to hybridize as an intermediately repeating element. Between ten and twenty Eco RI bands are apparent over a diffuse background in whole genome Southern blots. As noted above, a major structural distinction between the two genes in this region is the presence of an unusually large fourth intron in SB alpha

Figure 3. Alignment of the homologous IgC epsilon and HLA-SB(DP) alpha sequences. Ce1 indicates the sequence approximately 8kb upstream from the functional IgC epsilon gene. Ce2 indicates the homologous sequence directly upstream from the deleted pseudogene. Both sequences are from Hisajima et. al. (33). SBa indicates the sequence from the fourth intron of SB alpha (pos. 8641-9078 in figure 2). Arrows indicate the positions of the 39bp inverted repeats. The TGGGG pentanucleotide at the IgC-epsilon pseudogene deletion endpoint is underlined (see text for further details).

(737bp in DR alpha versus 3289bp in SB alpha). Comparison of the sequence of this intron with the NIH-GenBank nucleotide sequence data library (28) revealed a striking homology with sequences studied by Hisajima et. al. (33) upstream from the IgC epsilon pseudo and IgC epsilon functional genes. As shown in figure 3, there is a 73% homology between the SB alpha intron sequence and the IgC epsilon gene over 300bp. In the pseudogene, this homology ends abruptly at the pentanucleotide TGGGG characteristic of Ig switch regions. A deletion at this site is thought to be in part responsible for the inactivation of the second IgC epsion gene (33). The homology of the SB alpha intron sequence with the IgC epsilon functional gene extends beyond the TGGGG deletion endpoint through a 39bp sequence which is repeated nearly perfectly (90%) in an inverted fashion 1.5kb downstream

near the beginning of SB alpha exon 5 (figures 2 and 3). This repeat does not appear to be present in the available IgC epsilon sequence nor are these sequences found elsewhere in the SB region sequence. As Hisajima et. al. (33) have noted, it is conceivable that the Ig switch mechanism may be involved in germline rearrangements such as the deletion of sequences near IgC epsilon, and, in this case, the apparent insertion of DNA in the SB alpha gene relative to DR alpha. Although it seems more likely that these sequences are involved primarily in such rearrangements, we cannot rule out the possibility that they are involved in a functional relationship between the SB and IgC epsilon genes or gene products.

In conclusion, we have described the structure and genomic organisation of the gene encoding HLA-SB(DP) alpha. The clones described, containing pairs of complete SB alpha and SB beta genes, together with the ability to obtain integration and expression of exogenous DNA in transgenic animals, will provide a favourable model system in which to study the function and regulation of the SB genes and gene products.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the excellent technical assistance of Ms. Julia Wang and Ms. Sherry Rucker and to thank Ms. Emily Lawrance for help in preparing the manuscript. We also thank Dr. Henry Erlich for providing the SB alpha cDNA sequence. This work was supported by National Institutes of Health Grants AM 33871-02.

*Present address: Rockefeller University, New York, NY 10021, USA

REFERENCES

1. Kaufman, J.F., Auffray, C., Korman, A.J., Shackelford, D.A. and Strominger, J. (1984) Cell 36, 1-13.
2. Shaw, S., Johnson, A.H. and Shearer, G.M. (1980) J. Exp. Med. 152, 565-580.
3. Eckels, D.D., Lake, P., Lamb, J.R., Johnson, A.H., Shaw, S., Woody, J.N. and Hartzman, R.J. (1983) Nature 301, 716-718.
4. Shaw, S., Pollack, M.S., Payne, S.M. and Johnson. A.H. (1980) Human Immunology 1, 177-185.
5. Termijtelen, A., Khan, P.M., Shaw, S. and van Rood, J.J. (1983) Immunogenetics 18, 503-512.

6. Erlich, H., personal communication.
7. Auffray, C., Lillie, J.W., Arnot, D., Grossberger, D., Kappes, D. and Strominger, J.L. (1984) Nature 308, 327-333.
8. Gorski, J., Rollini, P., Long, E. and Mach, B. (1984) Proc. Natl. Acad. Sci. USA 81, 3934-3938.
9. Trowsdale, J., Kelly, A., Lee, J., Carson, S., Austin, P. and Travers, P. (1984) Cell 38, 241-249.
10. Servenius, B., Gustafsson, K., Widmark, E., Emmoth, E., Andersson, G., Larhammar, D., Rask, L. and Peterson, P.A. (1984) Eur. Mol. Biol. Org. J. 3, 3209-3214.
11. Okada, K., Boss, J.M., Prentice, H., Spies, T., Mengler, R., Auffray, C., Lillie, J., Grossberger, D. and J.L. Strominger (1985) Proc. Natl. Acad. Sci. USA 82, 3410-3414.
12. Hurley, C.K., Shaw, S., Nadler, L., Schlossman, S. and Capra, J.D. (1982) J. Exp. Med. 156, 1557-1562.
13. Kelly, A. and Trowsdale, J. (1985) Nucleic Acids Res. 13, 1607-1621.
14. Blin, N. and Stafford, D.W. (1976) Nucleic Acids Res. 3, 2303-2308.
15. Gladstone, P., Fueresz, L. and Pious, D. (1982) Proc. Natl. Acad. Sci. USA 79, 1235-1239.
16. Ish-Horowicz, D. and Burke, J.F. (1981) Nucleic Acids Res. 9, 2989-2998.
17. Poustka, A., Rackwitz, H-R., Frischauf, A-M., Hohn, B. and Lehrach, H. (1984) Proc. Natl. Acad. Sci. USA 81, 4129-4133.
18. Rimm, D.L., Horness, D., Kucera, J. and Blattner, F.R. (1980) Gene 12, 301-309.
19. Hohn, B. (1979) Methods Enzymol. 68, 229-309.
20. Maniatis, T., Hardison, R.C., Lacy, E. Lauer, J., O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978) Cell 15, 687-701.
21. Leder, P., Tiemeier, D. and Enquist, L. (1977) Science 196, 175-177.
22. Das, H.K., Lawrance, S.K. and Weissman, S.M. (1983) Proc. Natl. Acad. Sci. USA 80, 3543-3547.
23. Das, H.K., Lawrance, S.K. and Weissman, S.M. (1983) Proc. Natl. Acad. Sci. USA 80, 7024.
24. Bell, J.I., Estess, P., St. John, T., Saiki, R., Watling, D.L., Erlich, H.A. and McDevitt, H.O. (1985) Proc. Natl. Acad. Sci. USA 82, 3405-3409.
25. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5465-5467.
26. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
27. Staden, R. (1982) Nucleic Acids Res. 10, 4731-4751.
28. Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12(1), 387-395.
29. Das, H.K., Biro, P.A., Cohen, S.N., Erlich, H.A., VonGabain, A., Lawrance, S.K., Lemaux, P.G., McDevitt, H.O., Peterlin, B.M., Schulz, M.F., Sood, A.K. and Weissman, S.M. (1983) Proc. Natl. Acad. Sci. USA 80, 1531-1535.
30. Kappes, D.J., Arnot, D., Okada, K. and Stominger, J.L. (1984) Eur. Mol. Biol. Org. J. 3, 2985-2993.
31. Okada, K., Prentice, H.L., Boss, J.M., Levy, D.J., Kappes, D., Spies, T., Raghupathy R., Mengler, R.A., Auffray, C. and Strominger, J.L. (1985) Eur. Mol. Biol. Org. J. 4, 739-748.
32. Collins, F.S. and Weissman, S.M. (1984) Prog. Nucleic Acids Res. and Mol. Biol. 31, 315-462.
33. Hisajima, H., Nishida, Y., Nakai, S., Takahashi, N., Ueda, S. and Honjo, T., (1983) Proc. Natl. Acad. Sci. USA 80, 2995-2999.