Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 5068–5072, August 1985
Genetics

Sequence and evolution of the human T-cell antigen receptor β-chain genes

(gene conversion/DNA rearrangement)

A. TUNNACLIFFE, R. KEFFORD*, C. MILSTEIN, A. FORSTER, AND T. H. RABBITTS

Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom

Contributed by C. Milstein, April 3, 1985

ABSTRACT We present the nucleotide sequences of the two genomic constant (C)-region gene segments, $C_{\beta}1$ and $C_{\beta}2$, encoding the β chain of the human T-cell antigen receptor. The two C_{β} genes are organized identically to each other and to the corresponding mouse genes, both having four exons, whose boundaries were confirmed from the sequence of a $C_{\beta}2$ cDNA clone from the T-cell line MOLT-4. The predicted amino acid sequences of human $C_{\beta}1$ and $C_{\beta}2$ differ at only five positions, which suggests that the proteins have very similar functions. This similarity is the result of strong nucleotide-sequence conservation in protein-coding regions, which extends to silent positions. A quantitative analysis of an alignment of the nucleotide sequences of the two human genes shows that whereas the 5′ ends (including the first exon) are extremely homologous, the 3′ ends are widely divergent, with other regions having intermediate levels of homology. Analysis of published data [Gascoigne, N. R. J., Chien, Y., Becker, D. M., Kavaler, J. & Davis, M. M. (1984) *Nature (London)* 310, 387–391] shows that the mouse $C_{\beta}1$ and $C_{\beta}2$ genes are also virtually identical in their first exons but more divergent in the remaining coding regions. Therefore, partial gene conversion events may have occurred during the evolution of both human and mouse C_{β} genes.

In the cellular immune response of vertebrates, antigen recognition by T cells involves a membrane-bound receptor consisting of two disulfide-linked protein chains, α and β. Recently, cDNA clones corresponding to both chains have been isolated from man and mouse (1–4). Analysis of their sequences has revealed that the T-cell receptor (TCR) proteins consist of variable (V) and constant (C) regions (1–5). For the β-chain clones, comparison with their corresponding genomic sequences has shown that the variable region is composed of three types of segment, namely V_{β}, D_{β}, and J_{β} (variable, diversity, and joining, respectively), which are separated in germ-line DNA, but which rearrange in T cells to become contiguous (6–8). This recombination system has a direct parallel in immunoglobulin (Ig) heavy chain *V* gene structure, where it is in part responsible for the generation of antibody diversity (9). It is likely, therefore, that TCR diversity is created by similar mechanisms.

Another parallel with Ig heavy chain (and λ light chain) genes is that the *C*-region genes for the TCR β chain are found as more than one copy. In both man and mouse, there are two tandemly arranged C_{β} genes, with each gene having its own cluster of J_{β} regions, either of which can support V_{β}-D_{β}-J_{β} rearrangement (10–12). The functional basis for the duplicated J_{β}-C_{β} gene structure is not well understood, however, and we therefore determined the sequences of the human genes in order to compare their protein products. The results also enabled us to examine the evolution of the human C_{β} genes with respect to each other and their mouse counterparts.

METHODS

DNA Filter Hybridizations. Genomic DNA was prepared from human T-cell leukemia lines MOLT-4 (13) and JM (14) and from human colon carcinoma line COLO 320 (15) as described (16), digested with *Bam*HI or *Eco*RI, fractionated in 0.8% agarose gels, and transferred to nitrocellulose filters (17). Filters were hybridized with a C_{β} cDNA probe (M131B10BB1) as described (10).

Isolation and Characterization of cDNA Clones. The cell line NH17b.5.5, a variant of MOLT-4 selected for high expression of the T-cell differentiation antigen HTA-1 (CD1) (18), was treated with interferon α (2000 units/ml) for 48 hr and cycloheximide (2 μg/ml) for 15 min before harvesting. Poly(A)$^+$ RNA was ioslated from a cell membrane pellet as described (19). Double-stranded cDNA was prepared, nuclease S1-treated, repaired with DNA polymerase Klenow fragment and avian myeloblastosis virus reverse transcriptase, size-selected, and ligated into *Sma* I-cleaved and phosphatase-treated pUC9 (20) essentially as described (21). Ampicillin-resistant colonies were selected after transformation of competent (22) TG1 (T. Gibson, personal communication) and screened at moderate colony density (4000 colonies per 82-mm filter) with a nick-translated full-length β-chain cDNA clone (10). From 30,000 colonies screened, 17 were positive, and one of these was selected for DNA sequence determination (pUCM4-4). Sequencing was performed with M13-derived vectors (23) by use of the dideoxy chain-termination method (24).

Sequence Analysis. The isolation of phage λ recombinants containing human C_{β} genes was described previously (10). Fragments hybridizing to a C_{β} probe were subcloned in M13mp18 or -mp19 (23) and further subcloned according to a partial restriction map (see below) for sequencing as above.

Pairs of sequences were aligned using the DIAGON program (25), with a setting of 14 matches in a 21-base-pair (bp) span (67%). The final fine adjustment was done manually, introducing gaps to increase homology. Silent-site divergence (S) in coding regions was calculated as in ref. 26. Divergence (N) of introns and 3′ noncoding regions was calculated directly, treating each gap as one site and as a single substitution, regardless of length. The splicing signals, -G-T- and -A-G-, at the extremities of introns were ignored, being conserved among unrelated genes (27).

RESULTS AND DISCUSSION

Isolation and Characterization of a Human $C_{\beta}2$ cDNA Clone. To determine the positions of exon/intron boundaries within the C_{β} genes, it is necessary to refer to corresponding

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

Abbreviations: TCR, T-cell antigen receptor; V, variable; C, constant; J, joining; D, diversity; bp, base pair(s); kb, kilobase(s).

*Present address: Ludwig Institute for Cancer Research, Blackburn Building, University of Sydney, NSW 2006, Australia.

cDNA or protein sequences. Although cDNA clones for human $C_\beta 1$ are available (1, 10), no $C_\beta 2$ sequences were known when this work was started. Therefore, we isolated a human $C_\beta 2$ cDNA clone from the T-cell line MOLT-4, which was shown by filter hybridization to have $C_\beta 2$ gene rearrangements on both chromosomes. Fig. 1 shows the hybridization of a $C_\beta 1$ cDNA clone (which hybridizes to both C_β genes) with *Eco*RI and *Bam*HI digests of various genomic DNAs. As previously described (10), after digestion with *Eco*RI, DNA from the T-cell line JM shows rearrangement of the $C_\beta 1$ gene [9.5-kilobase (kb) band] on one chromosome, while retaining the germ-line $C_\beta 1$ configuration (11.5-kb band) on the other chromosome (Fig. 1*A*). The 4-kb band represents the unrearranged $C_\beta 2$ genes from both chromosomes. In MOLT-4 DNA, there is no band corresponding to the $C_\beta 1$ gene, indicating deletion of this gene on both chromosomes, but the 4-kb *Eco*RI fragment containing the $C_\beta 2$ gene(s) is present. Rearrangement of the $C_\beta 2$ genes is not observed when the DNA is cleaved with *Eco*RI, since this enzyme cuts downstream of the $J_\beta 2$ cluster (unpublished results). However, MOLT-4 has two *Bam*HI fragments containing C_β genes that differ from the germ-line pattern (represented by the 24-kb band seen in COLO 320), suggesting rearrangement of both $C_\beta 2$ genes. RNA blot hybridization data (not shown) confirmed that at least one of the $C_\beta 2$ genes is expressed in MOLT-4.

A cDNA library prepared from MOLT-4 RNA was screened with a C_β probe, and several clones were isolated. The sequence of one clone, pUCM4-4, is given in Fig. 2. It covers the entire $C_\beta 2$ coding region and extends through the whole of the *V* region including the sequence encoding the hydrophobic leader segment and the probable initiator methionine. The cDNA *C*-region sequence is identical with the genomic $C_\beta 2$ coding sequence presented here (see below) and differs in only a small number of nonreplacement positions from two published human $C_\beta 2$ cDNA sequences (28). The *J* region of the MOLT-4 cDNA clone does not correspond to any previously defined sequence but has greatest similarity (13 out of 16 amino acids) with a $J_\beta 2.1$ of mouse (11, 12), being more related to this than to two human $J_\beta 2$ sequences (28). There must therefore be at least three $J_\beta 2$ sequences upstream of $C_\beta 2$. It is also noteworthy that the *V* region of pUCM4-4 shows little homology with published sequences, and therefore provides the basis for defining a new germ-line V_β gene.

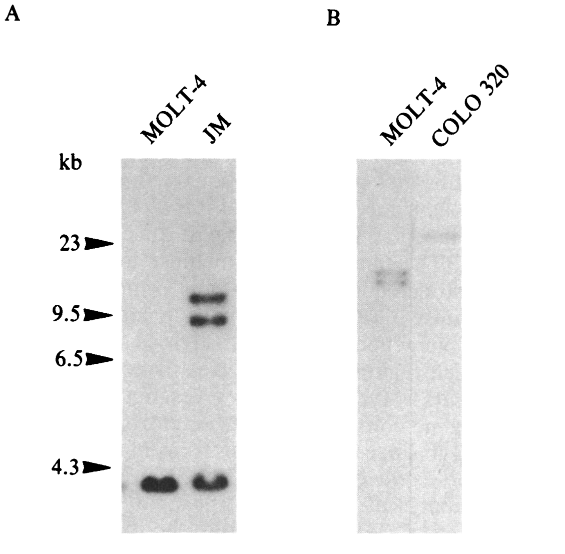


FIG. 1. Hybridization of TCR β-chain gene probes to DNA from T-cell lines MOLT-4 and JM and from colon carcinoma line COLO 320. DNA (10 μg) was digested with *Eco*RI (*A*) or *Bam*HI (*B*), fractionated in a 0.8% agarose gel, and transferred to nitrocellulose filters. The filters were hybridized with a C_β cDNA probe (M131B10BB1) (10). Sizes were calculated from coelectrophoresis of phage λ DNA digested with *Hin*dIII.

Sequence and Organization of Human C_β Genes. The $C_\beta 1$ and $C_\beta 2$ genes were cloned and sequenced as described in *Methods*. The sequencing strategies and complete sequences are shown in Fig. 3. The sequences have been aligned for direct comparison (see below), and gaps have been inserted to increase homology. The exon/intron boundary positions were confirmed by comparison with $C_\beta 1$ and $C_\beta 2$ cDNA sequences (this paper and refs. 1, 10, and 28).

Both genes have the same transcriptional orientation and consist of four exons, with introns located at identical positions in each gene: the first exon (387 bp) encodes both the immunoglobulin-like constant domain of the TCR and part of the so-called connecting peptide (11); the remainder of the connecting peptide is encoded by the second exon (only 18 bp) and a portion of the third exon (107 bp). Most of the transmembrane region (21 of 22 amino acids) of the protein is coded for by the rest of exon 3. The fourth exon encodes 1 amino acid of the transmembrane region and a cytoplasmic tail of 5 amino acids in $C_\beta 1$ and of 7 amino acids in $C_\beta 2$, followed by 3′ noncoding regions of about 190 bp up to the polyadenylylation signals. In the $C_\beta 1$ gene, the intron sizes are 442, 152, and 322 bp, respectively; in $C_\beta 2$, these are 516, 143, and 291 bp.

The nucleotide sequences of the coding regions of the two human C_β genes are remarkably similar, being about 97% identical. This results in just 5 amino acid replacements out of 177 shared positions. It seems likely, therefore, that the C_β proteins have very similar functions.

The overall organization of the human C_β genes is almost identical to that of the equivalent murine genes (11, 12), the numbers of exons and positioning of exon/intron boundaries being the same. However, two differences are that the first exon is 12 bp longer in the human genes and that mouse $C_\beta 2$ does not have the extra 2 amino acids found in the cytoplasmic tail of the human $C_\beta 2$ protein.

Evolution of C_β Genes. It is apparent, either by visual inspection (Fig. 3) or with the aid of a homology matrix computer program (25), that the degree of nucleotide sequence homology varies dramatically along the length of the C_β genes. To evaluate this phenomenon quantitatively, divergence was calculated in different regions of the aligned sequences as described in *Methods*. In noncoding regions (introns and 3′ untranslated regions), divergence (N) was calculated directly, but in coding regions, only silent-site divergence (S) was calculated, since the rate of substitution at these sites is not restricted by the need to conserve protein structure. Silent-site substitution rates are thus more directly comparable to the substitution rate at noncoding sites (30). S and N were calculated for the regions of the human C_β genes listed in Table 1. Due to the small sizes of the coding regions of exons 2 and 4, these were not included. Similar calculations were performed for human and mouse (11) $C_\beta 2$ genes, the exon sequences of human and mouse $C_\beta 1$ genes (alignments not shown), and the exon sequences of mouse $C_\beta 1$ and $C_\beta 2$ genes (based on the alignment of ref. 12).

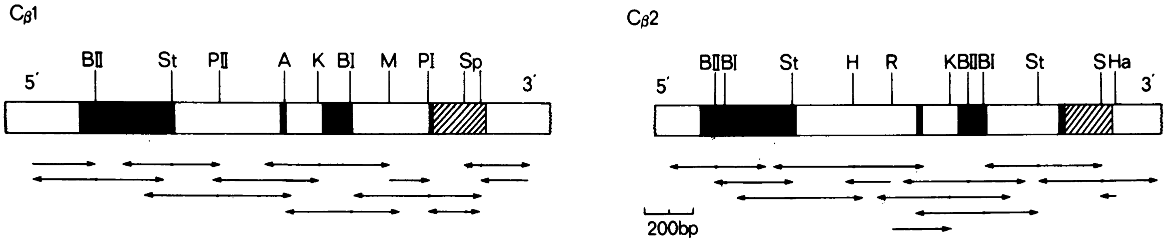
In the human/mouse $C_\beta 1$ and $C_\beta 2$ comparisons, S and N values are roughly constant throughout the gene lengths (Table 1). This shows that different parts of the genes have evolved at equivalent rates between the two species and serves as a control for the intraspecies comparisons. In the human $C_\beta 1/C_\beta 2$ alignment, however, S and N vary widely throughout the gene, a result which suggests unequal degrees of evolution in different regions. Since the two C_β genes must have arisen by gene duplication, we would expect all regions of the genes to be equally divergent in nonreplacement sites. This apparent paradox can be explained by invoking a

```
                     M  L  L  L  L  L  L  L  G  L  A  G  S  G  L  G  A  V  V  S  Q  H  P  S
GTGTGAGGCCATCACGGAAGATGCTGCTGCTTCTGCTGCTTCTGGGGCTAGCAGGCTCCGGGCTTGGTGCTGTCGTCTCTCAACATCCGA
        10        20        30        40  ┌──── 50        60        70        80        90
            o                             │
   W  V  I  C  K  S  G  T  S  V  K  I  E  Ⓒ  R  S  L  D  F  Q  A  T  T  M  F  W  Y  R  Q  F
GCTGGGTTATCTGTAAGAGTGGAACCTCTGTGAAGATCGAGTGCCGTTCCCTGGACTTTCAGGCCACAACTATGTTTTGGTATCGTCAGT
       100       110       120       130       140       150       160       170       180

   P  K  Q  S  L  M  L  M  A  T  S  N  E  G  S  K  A  T  Y  E  Q  G  V  E  K  D  K  F  L  I
TCCCGAAACAGAGTCTCATGCTGATGGCAACTTCCAATGAGGGCTCCAAGGCCACATACGAGCAAGGCGTCGAGAAGGACAAGTTTCTCA
       190       200       210       220       230       240       250 ────┐ 260       270
                                                                           │
   N  H  A  S  L  T  L  S  T  L  T  V  T  S  A  H  P  E  D  S  S  F  Y  I  Ⓒ  S  A  R  E  S
TCAACCATGCAAGCCTGACCTTGTCCACTCTGACAGTGACCAGTGCCCATCCTGAAGACAGCAGCTTCTACATCTGCAGTGCTAGAGAGT
       280 :     290       300       310       320       330       340       350       360
           : J                                             │ Cβ2
   T  S  D :P  K  N  E  Q  F  F  G  P  G  T  R  L  T  V  L │E  D  L  K  N  V  F  P  P  E  V
CGACTAGCGATCCAAAAAATGAGCAGTTCTTCGGGCCAGGGACACGGCTCACCGTGCTAGAGGACCTGAAAAACGTGTTCCCACCCGAGG
       370       380       390       400       410       420       430       440       450

   A  V  F  E  P  S  E  A  E  I  S  H  T  Q  K  A  T  L  V  C  L  A  T  G  F  Y  P  D  H  V
TCGCTGTGTTTGAGCCATCAGAAGCAGAGATCTCCCACACCCAAAAGGCCACACTGGTGTGCCTGGCCACAGGCTTCTACCCCGACCACG
       460       470       480       490       500       510       520       530       540

   E  L  S  W  W  V  N  G  K  E  V  H  S  G  V  S  T  D  P  Q  P  L  K  E  Q  P  A  L  N  D
TGGAGCTGAGCTGGTGGGTGAATGGGAAGGAGGTGCACAGTGGGGTCAGCACAGACCCGCAGCCCCTCAAGGAGCAGCCCGCCCTCAATG
       550       560       570       580       590       600       610       620       630

   S  R  Y  C  L  S  S  R  L  R  V  S  A  T  F  W  Q  N  P  R  N  H  F  R  C  Q  V  Q  F  Y
ACTCCAGATACTGCCTGAGCAGCCGCCTGAGGGTCTCGGCCACCTTCTGGCAGAACCCCCGCAACCACTTCCGCTGTCAAGTCCAGTTCT
       640       650       660       670       680       690       700       710       720

   G  L  S  E  N  D  E  W  T  Q  D  R  A  K  P  V  T  Q  I  V  S  A  E  A  W  G  R  A  D  C
ACGGGCTCTCGGAGAATGACGAGTGGACCCAGGATAGGGCCAAACCTGTCACCCAGATCGTCAGCGCCGAGGCCTGGGGTAGAGCAGACT
       730       740       750       760       770       780       790       800       810

   G  F  T  S  E  S  Y  Q  Q  G  V  L  S  A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V
GTGGCTTCACCTCCGAGTCTTACCAGCAAGGGGTCCTGTCTGCCACCATCCTCTATGAGATCTTGCTAGGGAAGGCCACCTTGTATGCCG
       820       830       840       850       860       870       880       890       900

   L  V  S  A  L  V  L  M  A  M  V  K  R  K  D  S  R  G  *
TGCTGGTCAGTGCCCTCGTGCTGATGGCCATGGTCAAGAGAAAGGATTCCAGAGGCTAGCTCCAAAACCATCCCAGGTCATTCTTCATCC
       910       920       930       940       950       960       970       980       990

TCACCCAGGATTCTCCTGTACCTGCTCCCAATCTGTGTTCCTAAAAGTGATTCTCACTCTGCTTCTCATCTCCTACTTACATGAATACTT
      1000      1010      1020      1030      1040      1050      1060      1070      1080

CTCTCTTTTTTCTGTTTCCCTGAAGATTGAGCTCCC
      1090      1100      1110
```

FIG. 2. Nucleotide and deduced amino acid sequences of the $C_{\beta}2$ cDNA clone (pUCM4-4) from MOLT-4. V (upstream of vertical dashed line), J, and C regions are indicated. The single-letter amino acid abbreviations are used. Cysteine residues presumed to be involved in V region intrachain disulfide bridges are circled; the circle above Cys-28 indicates a possible alternative residue for this disulfide bridge.

mechanism such as gene conversion, whereby the sequences of gene family members are corrected against one another during evolution (31). Such gene-conversion events involving the human C_{β} locus which affected only limited regions of each gene would account for the variable divergence along the gene lengths. The most dramatic example of an apparent gene conversion encompasses exon 1 and ≈100 bp of the first intron. Here, sequences are highly conserved between $C_{\beta}1$ and $C_{\beta}2$, which implies an evolutionarily recent gene conversion. The degree of homology in this region is in sharp contrast to that found at the 3' ends of the genes, which are widely divergent. It is reasonable to suppose that neither the sequences of the third introns nor those of the 3' noncoding regions have been corrected against one another since the ancient C_{β} gene duplication took place. Other regions of comparison show intermediate levels of homology, suggesting the occurrence of more than one gene-conversion event between the human C_{β} genes during evolution. One event may have covered intron 2 and exon 3 (and possibly extended upstream of exon 2), which have similar degrees of divergence (Table 1). It is difficult to be precise about the number and extent of such gene conversions with present knowledge of evolutionary mechanisms, but it is clear that gene conversion has played an important role in the evolution of human C_{β} genes. [The times of some of these evolutionary events can be estimated using published formulas (32) with the human/mouse $C_{\beta}2$ comparison as a standard, having an average divergence of 43.8% (weighted for the number of sites in each region), and a species divergence time of ≈75 million years (Myr) (33). This gives figures of 30–40 Myr since the gene conversion involving intron 2 and exon 3 and about 3 Myr since the conversion event in exon 1. The C_{β} duplication event occurred at least 100 Myr ago. Although these estimates of divergence times can be subject to considerable error (34), they illustrate the marked difference in evolutionary separation of different regions of the human C_{β} genes.]

Gene conversion also seems to have occurred between the two mouse C_{β} genes: exon 1 sequences are virtually identical (11, 12), as with the human genes, whereas exon 3 and the 3' noncoding region show about 40% divergence at nonreplacement sites (Table 1). This suggests that although the first



```
TGCATCCTAGGGACAGCATAGAAAGGAGGGGCAAAGTGGAGAGAGAGCAACAGACACTGGGATGGTGACCCCAAAACAATGAGGGCCTAGAATGACATAGTTGTGCTTCATTACGGCCCA 120
                                                            *  *  * **** *** ** **** *****  * ***** ************* ** **
----------------------------------------------------------ATGGCGTAGTCCCC-AAAGAACGAGGACCTAG--TAACATAATTGTGCTTCATTATGGTCCT

                                                                                              E  D  L  N  K  V  F  P  P
TTCCCAGGG--CTCTCTCTCACACACACAGAGCCCCTACCAGAACCAGACAGCTCTCAGAGCAACCCTGGCTCCAACCCCTCTTCCCTTTCCAGAGGACCTGAACAAGGTGTTCCCACCC 240
***** *    ******* **** ****************** ************************* ** ***   ************************** ** ************
TTCCCGGCCTTCTCTCTCACACATACACAGAGCCCCTACCAGGACCAGACAGCTCTCAGAGCAACCCTAGCCCCATTACCTCTTCCCTTTCCAGAGGACCTGAAAAACGTGTTCCCACCC
                                                                                              E  D  L  K  N  V  F  P  P

 E  V  A  V  F  E  P  S  E  A  E  I  S  H  T  Q  K  A  T  L  V  C  L  A  T  G  F  F  P  D  H  V  E  L  S  W  W  V  N  G
GAGGTCGCTGTGTTTGAGCCATCAGAAGCAGAGATCTCCCACACCCAAAAGGCCACACTGGTGTGCCTGGCCACAGGCTTCTTCCCCGACCACGTGGAGCTGAGCTGGTGGGTGAATGGG 360
********************************************************************************** *************************************
GAGGTCGCTGTGTTTGAGCCATCAGAAGCAGAGATCTCCCACACCCAAAAGGCCACACTGGTGTGCCTGGCCACAGGCTTCTACCCCGACCACGTGGAGCTGAGCTGGTGGGTGAATGGG
 E  V  A  V  F  E  P  S  E  A  E  I  S  H  T  Q  K  A  T  L  V  C  L  A  T  G  F  Y  P  D  H  V  E  L  S  W  W  V  N  G

 K  E  V  H  S  G  V  S  T  D  P  Q  P  L  K  E  Q  P  A  L  N  D  S  R  Y  C  L  S  S  R  L  R  V  S  A  T  F  W  Q  N
AAGGAGGTGCACAGTGGGGTCAGCACGGACCCGCAGCCCCTCAAGGAGCAGCCCGCCCTCAATGACTCCAGATACTGCCTGAGCAGCCGCCTGAGGGTCTCGGCCACCTTCTGGCAGAAC 480
************************** *********************************************************************************************
AAGGAGGTGCACAGTGGGGTCAGCACAGACCCGCAGCCCCTCAAGGAGCAGCCCGCCCTCAATGACTCCAGATACTGCCTGAGCAGCCGCCTGAGGGTCTCGGCCACCTTCTGGCAGAAC
 K  E  V  H  S  G  V  S  T  D  P  Q  P  L  K  E  Q  P  A  L  N  D  S  R  Y  C  L  S  S  R  L  R  V  S  A  T  F  W  Q  N

 P  R  N  H  F  R  C  Q  V  Q  F  Y  G  L  S  E  N  D  E  W  T  Q  D  R  A  K  P  V  T  Q  I  V  S  A  E  A  W  G  R  A
CCCCGCAACCACTTCCGCTGTCAAGTCCAGTTCTACGGGCTCTCGGAGAATGACGAGTGGACCCAGGATAGGGCCAAACCCGTCACCCAGATCGTCAGCGCCGAGGCCTGGGGTAGAGCA 600
******************************************************************************** ***************************************
CCCCGCAACCACTTCCGCTGTCAAGTCCAGTTCTACGGGCTCTCGGAGAATGACGAGTGGACCCAGGATAGGGCCAAACCTGTCACCCAGATCGTCAGCGCCGAGGCCTGGGGTAGAGCA
 P  R  N  H  F  R  C  Q  V  Q  F  Y  G  L  S  E  N  D  E  W  T  Q  D  R  A  K  P  V  T  Q  I  V  S  A  E  A  W  G  R  A

GGTGAGTGGGGCCTGGGGAGATGCCTGGAGGAGATTAGGTGAGACCAGCTACCAGGGAAAATGGAAAGATCCAGGTAGCAGACAAGACTAGATCCAAAAAGAAAGGAACCAGCGCACAC- 720
******************************************************************************* ****************  *******   **** *   **
GGTGAGTGGGGCCTGGGGAGATGCCTGGAGGAGATTAGGTGAGACCAGCTACCAGGGAAAATGGAAAGATCCAGGTAGCGGACAAGACTAGATCCAG-AAGAAAG---CCAGAGTGGACA

---------------------------------------------------------------CATGAAGG-AGAATTGGGCACCTGTGGTTCATTCTTCTCCCAGATTCTCAGCCCAAC 840
                                                               ***  ***  **** * ******   * *********   ***** ** *  **
AGGTGGGATGATCAAGGTTCACAGGGTCAGCAAAGCACGGTGTGCACTTCCCCCACCAAGAAGCATAGAGGCTGAATGGAGCACCTCAAGCTCATTCTTCCTTCAGATCCTGACACCTT-

AGAGCCAAGCAGCTGGGTCCCCTTTCTATGTGGCCTGTGTAACTCTCATCTGGGTGGTG--CCCCCCATCCCCCTCAGTGCTGCCACATGCCATGGATTGCAAGGACAATGTGGCTGACA 960
***** ****      *** ** *        ***           ****** ******  *  ***** * * ** * * * *    * **   ***   *  *** *  * **** **
AGAGCTAAGCTTTCAAGTCTCCCTGAGGACCAGCCATACAGCTCAGCATCTGAGTGGTGTGCATCCCATTCTCTTCTGGGGTCCTGGTTTCCTAAGATCATAGTGACCACTTCGCTGGCA

TCTGCATGGCAGAAGAAAGGAGGTGCT-GGGCTGTCAGAGGAAGCTGGTCT--------------------------GGGCCTGGGAGTCTGTGCCAACTGCAAATCTGACTTTACTTTT 1080
   *    ***  **  **   *  *  ***** *** **** ****   *                           ****** *********      *      ** * ****
CTGGAGCAGCATGAGGGAGACAGAACCAGGGCTATCAAAGGAGGCTGACTTTGTACTATCTGATATGCATGTGTTTGTGGCCTGTGAGTCTGTGATGTAAGGCTCAATGTCCTTAC----

                                                        D  C  G  F  T  S
AATTGCCTATGAAAATAAGGTCTCTCATTTATTTTCCTCTCCCTGCTTTCTTTCAGACTGTGGCTTTACCTCGGGTAAGTAAGCCCTTCCTTTTCCTCTCCCTCTCTCATGGTTCTTGAC 1200
           **       ********  ***** **  ***** ******************** ***** ******* ** *  ******* *****  *** **   **   **
----------AAAGCAGCATTCTCTCATCCATTTTTCTTCCCCTGTTTTCTTTCAGACTGTGGCTTCACCTCCGGTAAGTGAGTCTCTCCTTTTTCTCTCTATCTTTCGCCGTCTCTGCT
                                                        D  C  G  F  T  S

                                                                                                          V  S  Y  Q  Q
CTAGAACCAAGGCATGAAGAACTCACAGACACTGGAGGGTGGAGGGTGGGAGAGACCAGAGCTACCTGTGCACAGGTACCCACCTGTCCTTCCTCCGTGCCAACAGTGTCCTACCAGCAA 1320
** ****** ****** ****  *** ***** ** * ***  **  *  **** *       ***************** ** **  **   * * ** ****** *** *********
CTCGAACCAGGGCATGGAGAATCCACGGACACAGGGGCGTGAGGGAGGCCAGAGCC-------ACCTGTGCACAGGTACCTACATGCTCT--GTTCTTGTCAACAGAGTCTTACCAGCAA
                                                                                                          E  S  Y  Q  Q

 G  V  L  S  A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  A  L  V  L  M  A  M
GGGGTCCTGTCTGCCACCATCCTCTATGAGATCCTGCTAGGGAAGGCCACCCTGTATGCTGTGCTGGTCAGCGCCCTTGTGTTGATGGCCATGGTAAGCAGGAGGGCAGGATGGGGCCAG 1440
********************************* ***************** ******* *********** ***** *** **************** *******  **** **** **
GGGGTCCTGTCTGCCACCATCCTCTATGAGATCTTGCTAGGGAAGGCCACCTTGTATGCCGTGCTGGTCAGTGCCCTCGTGCTGATGGCCATGGTAAGGAGGAGGGTGGGATAGGGC-AG
 G  V  L  S  A  T  I  L  Y  E  I  L  L  G  K  A  T  L  Y  A  V  L  V  S  A  L  V  L  M  A  M

CAGGCTGGAGGTGACACACTGACACCAAGCA-CCCAGAAGTATAGAGTCCCTGCCAGGATTGGAGCTGGGCAGTAGGGAGGGAAGAGATTTCATTCAGGTGCCTCAGAAGATAACTTGCA 1560
  *   ** *  *        *** **  **       *    *   **   *****        **      *           *          *       ****           *
ATGATGGGGGCAGGGGATGGAACATCACACATGGGCATAAAGGAATCTCAGAGCCAGAGCACAGCCTAATATATCCTATCACCTCAATGAAACCATAATGAAGCCAGACTGGGGAGAAAA

CCTCTGTAGGATCACAGTGGAAGGGTCATGCTGGGAAGGAGA--AGCTGGAGTCACCAGAAAACCCAATGGATGTTGTGATGAGCCTTACTATTTGTGTGGTCAATGGGCCCTACTACTT 1680
     * *  *******   * *  *****   *** *****  * * *    * * *   **   *         *       **       *  *   *             *
TGCAGGGAATATCACAGA--ATGCATCATGGGAGGATGGAGACAACCAGCGAGCCCTACTCAAATTAGGCCTCAGAGCCCGCCTCCCCTGCCCTACTCCTGCTGTGCCATAGCCCC----

                                                           V  K  R  K  D  F  ter
TCTCTCAATCCTCACAACTCCTGGCTCTTAATAACCCCCAAAACTTTCTCTTCTGCAGGTCAAGAGAAAGGATTTCTGAAGGCAGCCCTGGAAGTGGAGTTAGGAGCTTCTAACCCGTCA 1800
                           * **   *    **   * *******  ******************* * ** *  *** *   **         *    *  * * **
---------------------------TGAAACCCTGAAAATGTTCTCTCTTCCACAGGTCAAGAGAAAGGATTCCAGAGGCTAGCTCCAAAACCATCCCAGGTCATTCTTCATCCTCAC
                                                           V  K  R  K  D  S  R  G ter

TGGTTTCAATACACATTCTTCTTTTGCCAGCGCTTCTGAAGAGCTGCTCTCACCTCTCTGCATCCCAATAGATATCCCCCTATGTGCATGCACACCTGCACACTCACGGCTGAAATCTCC 1920
         * *     ** **           *** ** **    ******    ** *    *      *  *        * *   *     *       *  *           *
CCAGGATTCTCCTGTACCTGCTCCCAATCTGTGTTCCTAAAAGTGATTCTCACTCTGCTTCTCATCTCCTACTTACATGAATACTTCTCTCTTTTTTCTGTTTCCCTGAAGATTGAGCTC

CTAACCCAGGGGGACCTTAGCATGCCTAAGTGACTAAACCAATAAAAATGTTCTGGTCTGGCCTGACTCTGACTTGTGAATGTCTGGATAGCTCCTTGGCTGTCTCTGAACTCCCTGTGA 2040
* *****               * *  **  * ***************  *    **  ******  *     *   ** ******   ** **      *  **      **
CCAACCC----------CCAAGTACGAAATAGGCTAAACCAATAAAAAATTGTGTGTTGGGCCTGGTTGCATTTCAGGAGTGTCTGTGGAGTTCTGCTCATCACTGACCTATCTTCTGAT

CTCTCCCCATTCAGTCAGGATAGAAACAAGAGGTATTCAAGGAAAATGCAGACTCTTCACGTAAGAGGGATGAGGGGCCCACCTTGAGATCAATAGCAGAA
 *      *           * * *            **       * *   *    **     *     *   *  *   *     **    *      *
TTAGGGAAAGCAGCATTCCCTTGGACATCTGAAGTGACAGCCCTCTTTCTCTCCACCCAATGCTGCTTTCTCCTGTTCATCCTGATGGAAGTCTCAACACA
```

FIG. 3. Sequencing of human $C_{\beta}1$ and $C_{\beta}2$ genes. (*Upper*) Sequencing strategies for $C_{\beta}1$ (*Left*) and $C_{\beta}2$ (*Right*). Partial restriction maps are shown, giving restriction sites used in sequencing: A, *Ava* I; BI, *Bal* I; BII, *Bgl* II; H, *Hin*dIII; Ha, *Hae* III; K, *Kpn* I; M, *Mbo* I; PI, *Pst* I; PII, *Pvu* II; R, *Rsa* I; S, *Sac* I; Sp, *Sph* I (there are two adjacent *Sph* I sites in $C_{\beta}1$); St, *Stu* I. Arrows indicate direction of sequencing. Shaded boxes represent exons; hatched boxes denote the 3′ untranslated regions. (*Lower*) Sequences and alignment of human $C_{\beta}1$ (upper line) and $C_{\beta}2$ (lower line) genes. Nucleotide sequences, with predicted amino acid sequences, are aligned to maximize homology (shown as asterisks), with insertion of gaps (signified by dashes) where necessary. Splicing signals (GT and AG) (27) and the polyadenylylation signal AATAAA (29) are underlined.

Table 1. Sequence divergence between human and mouse C_{β} genes

	S or N, % sequence divergence			
	Human $C_{\beta 1}$ vs. human $C_{\beta 2}$	Human $C_{\beta 2}$ vs. mouse $C_{\beta 2}$	Human $C_{\beta 1}$ vs. mouse $C_{\beta 1}$	Mouse $C_{\beta 1}$ vs. mouse $C_{\beta 2}$
5′ flanking region	19.1	53.6		
Exon 1	2.3	43.6	42.5	1.2
Intron 1				
5′ region	2.9	44.7		
Remainder	42.9	38.8		
Intron 2	29.1	40.7		
Exon 3	24.4	44.6	39.3	39.4
Intron 3	63.3	47.7		
3′ noncoding region	67.7	42.6	39.1	44.0

S refers to silent-site divergence in coding regions, and N to noncoding region divergence. Regions for comparison were defined by exon/intron boundaries, except for the 5′ region of intron 1, which ends at position 705 of Fig. 3. Coding regions of exons 2 and 4 were not included due to their small size. The 3′ noncoding region was delineated by the end of the termination codon (of human $C_{\beta 2}$ in the first two comparisons) and the beginning of the polyadenylylation signal AATAAA. Complete intron sequences of mouse $C_{\beta 1}$ were not available.

exons have undergone very recent gene conversion, this did not include the 3′ ends of the genes, which may have been evolving independently since about the time of the human/murine divergence.

It has been suggested that the ancestral J_{β}-C_{β} region duplication was exploited as a means of increasing the number of J_{β} segments, resulting in increased diversity after genetic drift (11, 12). However, our data show that the duplicated C_{β} genes have remained extremely similar in sequence (and, presumably, in function). In this context it may be important that the TCR β chain is part of a complex multiprotein structure on T-cell surfaces. This structure comprises the TCR α and β chains and the T3 molecule, itself consisting of three or four subunits (35). It may be desirable that both $C_{\beta}1$ and $C_{\beta}2$ interact equally well with T3, in order to maintain the advantage of a duplicated J_{β} locus. Identity of $C_{\beta}1$ and $C_{\beta}2$ over critical regions would ensure this, and could be maintained by coevolution of the β-chain genes. In support of this argument is the observation that both human and mouse C_{β} genes have undergone very recent conversion events in the same regions.

We wish to thank A. Stinson for excellent technical help, F. Calabi for advice and help in cDNA cloning, M. Bishop for discussions on the calculation of sequence divergence, and J. Firth and E. Wickham for preparing the typescript. A.T. is the recipient of a Medical Research Council Training Fellowship. R.K. was supported by a Florey Fellowship from the Royal Society and a Buck Studentship from St. John's College, Cambridge.

1. Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S. P., Aleksander, I. & Mak, T. W. (1984) *Nature (London)* **308**, 145–149.
2. Hedrick, S. M., Cohen, D. I., Nielsen, E. A. & Davis, M. M. (1984) *Nature (London)* **308**, 149–153.
3. Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N. & Tonegawa, S. (1984) *Nature (London)* **312**, 36–40.
4. Sim, G. K., Yague, J., Nelson, J., Marrack, P., Palmer, E., Augustin, A. & Kappler, J. (1984) *Nature (London)* **312**, 771–775.
5. Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. (1984) *Nature (London)* **308**, 153–158.
6. Siu, G., Clark, S. P., Yoshikai, Y., Malissen, M., Yanagi, Y., Strauss, E., Mak, T. W. & Hood, L. (1984) *Cell* **37**, 393–401.
7. Chien, Y., Gascoigne, N. R. J., Kavaler, J., Lee, N. E. & Davis, M. M. (1984) *Nature (London)* **309**, 322–326.
8. Kavaler, J., Davis, M. M. & Chien, Y. (1984) *Nature (London)* **310**, 421–423.
9. Tonegawa, S. (1983) *Nature (London)* **302**, 575–581.
10. Sims, J. E., Tunnacliffe, A., Smith, W. J. & Rabbitts, T. H. (1984) *Nature (London)* **312**, 541–545.
11. Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M. B., Yoshikai, Y., Fitch, F., Mak, T. W. & Hood, L. (1984) *Cell* **37**, 1101–1110.
12. Gascoigne, N. R. J., Chien, Y., Becker, D. M., Kavaler, J. & Davis, M. M. (1984) *Nature (London)* **310**, 387–391.
13. Minowada, J., Ohnuma, T. & Moore, G. E. (1972) *J. Natl. Cancer Inst.* **49**, 891–895.
14. Schneider, U., Schwenk, H. U. & Bornkamm, G. (1977) *Int. J. Cancer* **19**, 621–626.
15. Alitalo, K., Schwab, M., Lin, C. C., Varmus, H. E. & Bishop, J. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1708–1711.
16. Bentley, D. & Rabbitts, T. H. (1981) *Cell* **24**, 613–623.
17. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
18. Burrone, O. R., Calabi, F., Kefford, R. F. & Milstein, C. (1983) *EMBO J.* **2**, 1591–1595.
19. Berger, S. L. & Birkenmeier, C. S. (1979) *Biochemistry* **18**, 5143–5149.
20. Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268.
21. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
22. Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557–580.
23. Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26**, 101–106.
24. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
25. Staden, R. (1982) *Nucleic Acids Res.* **10**, 2951–2961.
26. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
27. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
28. Yoshikai, Y., Anatoniou, D., Clark, S. P., Yanagi, Y., Sangster, R., van den Elsen, P., Terhorst, C. & Mak, T. W. (1984) *Nature (London)* **312**, 521–524.
29. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214.
30. Miyata, T., Yasunaga, T. & Nishida, T. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7328–7332.
31. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. (1983) *Cell* **33**, 25–35.
32. Kimura, M. & Ohta, T. (1972) *J. Mol. Evol.* **2**, 87–90.
33. Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 1–8.
34. Wilson, A. C., Carlson, S. S. & White, T. J. (1977) *Annu. Rev. Biochem.* **46**, 573–639.
35. Borst, J., Alexander, S., Elder, J. & Terhorst, C. (1983) *J. Biol. Chem.* **258**, 5135–5141.