Brief Definitive Report

CDR3 Length in Antigen-specific Immune Receptors

By Edwin P. Rock,* Peter R. Sibbald,‡ Mark M. Davis,* and Yueh-hsiu Chien

*From the Department of Microbiology and Immunology, and the *Howard Hughes Medical Institute, Stanford University, Stanford, California 94305; and the ‡European Molecular Biology Laboratory, 6900 Heidelberg, FRG*

Summary

In both immunoglobulins (Ig) and T cell receptors (TCR), the rearrangement of V, D, and J region sequence elements during lymphocyte maturation creates an enormous degree of diversity in an area referred to as the complementarity determining region 3 (CDR3) loop. Variations in the particular V, D, and J elements used, precise points of recombination, and random nucleotide addition all lead to extensive length and sequence heterogeneity. CDR3 loops are often critical for antigen binding in Igs and appear to provide the principal peptide binding residues in TCRs. To better understand the physical and selective constraints on these sequences, we have compiled information on CDR3 size variation for Ig H, L (κ and λ) and TCR α, β, γ, and δ. Ig H and TCR δ CDR3s are the most variable in size and are significantly longer than L and γ chains, respectively. In contrast, TCR α and β chain distributions are highly constrained, with nearly identical average CDR3 lengths, and their length distributions are not altered by thymic selection. Perhaps most significantly, these CDR3 length profiles suggest that γ/δ TCRs are more similar to Igs than to α/β TCRs in their putative ligand binding region, and thus γ/δ and α/β T cells may have fundamentally different recognition properties.

Specific vertebrate immune responses are initiated by B and T lymphocytes via Igs and TCRs, respectively. Although x-ray crystal structures are available only for Igs, TCRs are believed to share a similar tertiary and quaternary structure (1–3). Antigen-specific immune receptors confer specificity against a wide variety of potential pathogens by recombination of V, D, and J elements into a single Ig or TCR variable domain–encoding exon (4). Ig L (κ and λ), TCR α and TCR γ chains utilize only V and J gene elements, whereas Ig H, TCR β, and TCR δ also employ one or more D elements. X-ray structural analysis of antibody–antigen complexes shows that one or both of the CDR3 loops of Ig H and L chains are always involved in antigen contact (4). Similarly, the CDR3s of both α and β TCR chains seem critical for peptide recognition (5, 6). As a result of variation in numbers of D and J elements used, D element reading frames, junctional diversity, and N region nucleotide addition, the estimated number of possible CDR3 sequences is greatest for γ/δ TCRs and least for Igs (irrespective of somatic mutation), with α/β TCRs being intermediate (7).

Characterization of the length distribution of CDR3 in different immune receptor chains is of interest for two reasons. First, in computer modelling of antibodies, the length of a given CDR has a profound effect on its shape, with differences of even one amino acid able to produce significant changes in the overall structure (8, 9). Thus, the analysis of CDR3 length variation among different antigen receptors might shed light on structure-function relationships in different immune receptor classes. Second, Igs undergo affinity maturation via somatic mutation, which does not change CDR3 length. T cells, on the other hand, undergo thymic selection, which is also likely to affect receptor affinity. Since thymic selection may largely be driven by "peptide" (10–12) and the CDR3 regions of the TCR are involved in peptide contact, it is important to know whether or not thymic selection affects CDR3 length distributions.

In this report, we characterize CDR3 length distribution in immune receptor repertoires and examine their potential mechanistic basis. Our results indicate that thymic selection has no effect on CDR3 length of α/β TCRs. Interestingly, the analyses do indicate that γ/δ TCRs are much more similar in their CDR3 lengths to Ig than to α/β TCR. This latter result suggests that γ/δ and α/β T cells may recognize antigens differently.

Materials and Methods

CDR3 lengths were tabulated using a computer program written in Pascal running on a VAX 6000-420 under VMS 5.4. All chains were tabulated from either the online Sequences of Proteins of Immunological Interest as of January 1992 (13) or a more specific compilation (14, 15). After excluding sequence fragments without a complete CDR3 region, the program calculates CDR3 length as the distance from the J region–encoded GXG triplet (where G is glycine and X is any amino acid) to the nearest preceding V region

 J. Exp. Med. © The Rockefeller University Press • 0022-1007/94/01/0323/06 $2.00
Volume 179 January 1994 323–328

encoded cysteine (C). In the results shown, however, CDR3 length is defined as four amino acids less than the number of residues between the aligned C in the V element to the GXG triplet in the J region (3). Thus, the CDR3 region included in the sequence V_β-CASSLNWSQDTQYFGPG-J_β would count as nine amino acids. Results for each family were checked by hand and consecutive duplicates excluded. Additional sequences were included when CDR3 length could be identified unambiguously from the original reference or when alanine was identified as a substitute for one G in the GXG motif.

Statistics were calculated on a Macintosh computer using StatView II (Abacus Concepts, Berkeley, CA). The variance formula (a measure of dispersion) used by StatView II is for an unbiased sample estimate, rather than for the population mean (13). Skewing and kurtosis are measures to assess whether a data set conforms to a normal distribution. Skewing is the average of cubed standard scores (or z-values) of the distribution and reveals whether extreme values are evenly distributed above and below the mean (16). Kurtosis is three less than the average fourth power of the standard scores; it describes the peak and tails of a distribution (16). A normal distribution is symmetric (skewness = 0) and mesokurtic (modestly peaked with modest tails, kurtosis = 0). Results of Mann-Whitney and Kolmorgorov-Smirnov statistics are expressed as a p value, which represents the probability that the values tested could have occurred due to chance alone.

Results and Discussion

To assess the distribution of CDR3 lengths in immune receptor repertoires, we have developed a program that scans an online version of the Kabat database to tabulate CDR3 lengths measured from the conserved residues flanking CDR3. Results for mouse and human Ig, α/β, and γ/δ TCR are shown in Table 1. Histograms and cumulative distribution functions of CDR3 lengths are shown in Fig. 1.

To facilitate comparison between immune receptor chain families, we sought to characterize the relative order of CDR3 length distributions. Skewing and kurtosis were first calculated to determine whether the data sets conform to a normal distribution. Human κ and mouse L chain CDR3 length distributions are markedly platykurtic (sharply peaked); human L chains also show significant skewing (Table 1). Mouse β and γ chains show moderate platykurtosis and negative skewing. Thus the data sets obtained do not all conform to a normal distribution. We therefore compared chain-family lengths using the Mann-Whitney U test, a statistic for comparing data from two independent groups which assumes no particular, e.g., normal, distribution of values (16). At the 0.01 confidence level, CDR3 length in humans is in the order δ>H>α,β>γ,L. However, at the 0.05 confidence level, human γ chain CDR3 lengths are longer than those of L chains (p = 0.04). In the mouse, β chain CDR3 lengths are longer than those of α (p = 0.001), although the difference in means is small relative to that between L and H or γ and δ. Also, neither α nor β lengths were statistically different from those of H or γ chains (p = 0.07–0.62). Thus, mouse CDR3 length at the 0.01 confidence level is not simple but does conform to the rules: δ>H,α,β,γ,>L and β>α.

Table 1. *CDR3 Length in Antigen-specific Immune Receptor Chain Families from Human and Mouse*

	Family		Sequences tested	CDR3*	Range‡	Median§	Mean	Variance	Skewing	Kurtosis
Human CDR3 lengths	L	κ	319	100	2,7	6	5.9	0.6	−2.8	10
		λ	150	74	5,10	7	7.3	0.9	−0.1	−0.4
		Total	**469**	**174**	**2,10**	**6**	**6.5**	**1.2**	**−2.3**	**2.2**
	H		**325**	**123**	**3,25**	**12**	**12.7**	**20.7**	**0.6**	**0.003**
	α		**75**	**66**	**6,12**	**9**	**9.2**	**2.8**	**0.1**	**−0.9**
	β		**116**	**97**	**6,12**	**9**	**9.5**	**4**	**0.4**	**0.2**
	γ		**46**	**31**	**1,12**	**7**	**7.2**	**5.5**	**0.03**	**0.6**
	δ		**42**	**40**	**8,21**	**14**	**14.5**	**11**	**0.2**	**−0.8**
Mouse CDR3 lengths	L	κ	1,068	484	4,8	6	6	0.2	0.6	11.3
		λ	75	21	6	6	6	0	—	—
		Total	**1,143**	**505**	**4,8**	**6**	**6**	**0.1**	**0.6**	**12**
	H		**1,620**	**757**	**1,18**	**9**	**8.5**	**8.1**	**−0.01**	**−0.2**
	α		**123**	**95**	**6,12**	**8**	**8.5**	**1.6**	**−0.5**	**−0.3**
	β		**174**	**147**	**4,13**	**9**	**8.9**	**2**	**−0.5**	**1.7**
	γ		**46**	**46**	**4,11**	**9**	**8.8**	**1.8**	**−0.8**	**1.7**
	δ		**171**	**161**	**6,19**	**13**	**12.7**	**6.4**	**0.02**	**−0.05**

* CDR3, number of lengths analyzed.
‡ Range, lowest value, highest value (in amino acids).
§ Median CDR3 length (50th percentile).

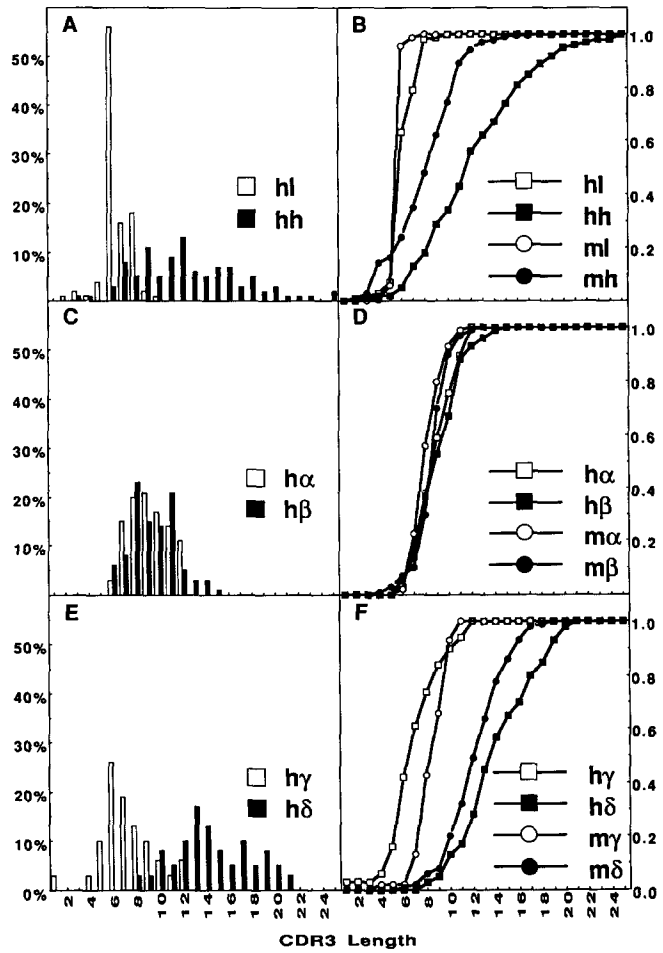

Figure 1. CDR3 lengths of antigen-specific immune receptor chains. (*A*, *C*, and *E*) Histograms showing percentages of CDR3 sequences at given lengths in human chain families. (*A*) Igs: (hl) human light; (hh) human heavy. (*C*) α/β TCR: (hα) human α; (hβ) human β. (*E*) γ/δ TCR: (hγ) human γ; (hδ) human δ. (*B*, *D*, and *F*) Cumulative distribution functions of CDR3 lengths. The vertical axis shows the proportion of sequences analyzed having a length less than or equal to the corresponding CDR3 length value on the horizontal axis. (*B*) Igs: (m*l*) Mouse light; (m*h*) mouse heavy. (*D*) α/β TCR: (mα) mouse α; (mβ) mouse β. (*F*) γ/δ TCR: (mγ) mouse γ; (mδ) mouse δ.

Among L chains as a group, $\lambda > \kappa$ CDR3 length in humans (p = 0.0001), and there is no difference in the mouse (p = 0.81). Human L, H, α, and δ chain CDR3 lengths are longer than those of the mouse (p = 0.0001–0.01) although this is not the case for β chains (p = 0.09). In contrast, mouse γ CDR3 lengths are longer than those of the human (p = 0.0003).

For chain families having similar lengths as determined above, we wished to establish whether the spread of distributions might nonetheless be different. The Kolmorgorov-Smirnov formula tests for whether two distributions are different in any way, again without requiring any particular grouping of data (16). There is a difference between CDR3 length distributions of mouse H and α chains (p = 0.04), as well as between H and β chains (p = 0.006). No other differences were found between CDR3 length distributions of chain families having similar lengths (p = 0.07–0.78).

Several points stand out. Ig H and TCR δ chains have the longest and most variable CDR3 lengths, whereas their respective paired L and γ chains are notably shorter and less variable. By contrast, α/β TCR chain CDR3 lengths have more similar ranges, medians (50th percentiles), and variance (dispersion). This dichotomy may be resolved by considering the nature of antigenic targets recognized by Igs versus α/β TCRs. Igs bind to a vast array of large and small unprocessed antigens. Thus the dispersed length distribution of Ig H CDR3 lengths is likely a function of the need to form crevices and protuberances that facilitate binding to a wide variety of antigenic surfaces. In contrast, α/β TCR are specific primarily for peptide–MHC complexes, in which the peptidic focus of antigenicity constitutes the central portion of a roughly planar surface (17, 18). We propose that α and β chain CDR3 lengths are constrained in size because of their evolutionary selection for binding to peptide–MHC complexes, since such lengths would be consistent with both α and β CDR3s contacting peptide directly (6).

Since both mouse and human α and β CDR3 lengths are remarkably uniform, we sought to ask whether thymic selection plays any role in this uniformity. We can address this by tabulating CDR3 lengths in subpopulations of T cells expressing the same V region gene element. Fig. 2 *A* shows the thymic $CD4^+8^+$, thymic $CD4^+8^-$, LN $CD4^+8^-$, and LN $CD4^-8^+$ CDR3 lengths of $V_\beta17^+$ T cells from SJL mice (14). Using the method of Kolmorgorov-Smirnov (16), we find no significant differences in length distribution between mouse β chains as a group and any of these four compartments (p = 0.41–0.78). Fig. 2 *B* shows CDR3 length distributions in $V_\beta8^+$ T cells from LN and spleen of wild-type C57BL/6 and C57BL/6 mice expressing the TCR α transgene (CDR3 length = 10) of the T cell hybridoma 2B4 (15). The β chain repertoire of the $CD4^+8^-$ and $CD4^-8^+$ cells in these mice was found to be limited, but that in the $CD4^-8^-$ population is as diverse as in nontransgenic mice. It is thought that the accumulation of $CD4^+8^-$ and $CD4^-8^+$ but not $CD4^-8^-$ cells is the result of positive selection (15). This comparison thus enables us to ask whether selective forces related to chain pairing could impinge on CDR3 length distribution in shaping the peripheral double-negative, $CD4^+8^-$, or $CD4^-8^+$ α/β TCR repertoires. Again, no differences in distribution are manifest between mouse β chains and any of the specific subsets examined (p = 0.08–0.18). Each of these two sets of data indicates that CDR3 length distributions of α and β chains are unaltered by thymic selection.

To assess the degree to which CDR3 length variation results from the relative lengths of the different D and J gene segments, we collated D (Table 2) and J region (Table 3) contributions to CDR3 size. There is a high correlation between median CDR3 length and the sum of median D and J contributions to CDR3 length (r = 0.96, excluding δ chains for which the number of D elements used is variable). A strong correlation also exists between median CDR3 length and me-

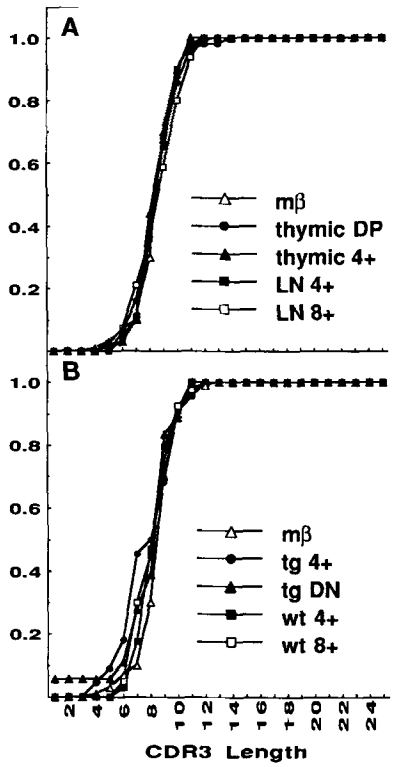

Figure 2. Effects of thymic selection on cumulative distribution functions of CDR3 lengths from subpopulations of T cells expressing the same V region gene element. The vertical axis shows the proportion of sequences analyzed having a length less than or equal to the corresponding CDR3 length value on the horizontal axis. (*A*) Vβ17-positive TCR β chains: mβ, all mouse β (as in Fig. 1); DP, $CD4^+CD8^+$ (double positive); 4^+, $CD4^+CD8^-$; 8^+, $CD4^-CD8^+$. mβ, n = 147; thymic DP, n = 49; thymic 4^+, n = 68; LN 4^+, n = 80; LN 8^+, n = 80. (*B*) Vβ-positive TCR β chains in 2B4 α transgenic mice. tg, 2B4 TCR α chain transgenic; DN, $CD4^-$ $CD8^-$ (double negative); (*wt*) wild type. tg 4^+ n = 22; tg DN n = 18; wt 4^+ n = 34; wt 8^+ n = 40.

dian J contribution for those chain families having only V-J junctions (r = 0.91 for L, α, γ).

These results indicate that gene segment usage is the greatest single determinant of CDR3 length distributions in all immune receptor chains. H chains, having long and widely dispersed CDR3 lengths, use the widest range of D elements. δ chains, which also exhibit a wide CDR3 length distribution, often use multiple D elements simultaneously (19, 20). Whereas human α and β CDR3 lengths show a similar distribution, D element usage by β chains is matched by greater J element length in α chains.

Interestingly, CDR3 length distributions do not correlate with genetic variability of the CDR3 region. The most constrained CDR3 length distributions are those of α and β chains. Yet α and β chain CDR3 regions are highly variable genetically with two possible D_β elements (12 nucleotides long) being translated in any of three reading frames, 12 J_β elements, and ~50 J_α elements (7, 21). In contrast, the dispersed CDR3 length distributions of Ig H and TCR δ chains are concomitant with genetic variability that is either less than or greater than that of α/β TCR. Ig H chain genetic variability is restricted by the fact that D elements typically cannot be translated in multiple reading frames. On the other hand, TCR δ variability is enhanced by its use of two or three D elements in a single chain, multiple reading frames, and three or four regions of N nucleotide insertion (19, 20).

These data also confirm that accurate modelling of TCR CDR loops will be even more difficult than those from Ig. First, "canonical" CDR loop conformations (8, 9) (i.e., conserved amino acid substructures) found in Igs are unknown in TCRs. This is particularly true for the CDR2 region of α/β TCRs, which shows high relative variability over a span of 13–15 amino acids (3, 13, 22). In addition, Fig. 1 shows that each TCR chain family has a CDR3 length distribution that is longer than that of Ig L CDR3. In the mouse, each chain family's distribution (except for Ig L) is at least as long as that of Ig H chain CDR3, which is yet to be modelled with any accuracy (9).

The antigen recognition requirements of γ/δ T cells are still poorly understood. Surprisingly we find here that γ and δ chain CDR3s are more like those of Ig than α/β TCR in both the disparity between heterologous chains of average CDR3 lengths and the pronounced variability of H and δ chain lengths. This suggests that γ/δ TCRs as a group may recognize antigens in a manner akin to Ig and different from that of α/β TCR with peptide–MHC complexes. Specifically,

Table 2. *D Subregion Lengths from Human and Mouse*

	Family	Number	Range	Median	Mean	Variance
Human D subregion lengths	H	31	3,14	10	8.5	6.5
	β	3	4,5	5	4.7	0.3
	δ	3	2,4	3	3	1
Mouse D subregion lengths	H	12	3,7	5	5.2	1.1
	β	2	4	4	4	0
	δ	2	3,5	4	4	2

Labels are as in Table 1.

Table 3. *J Subgene Contribution to CDR3 Length in Human and Mouse*

	Family	Number	Range	Median	Mean	Variance
Human J contribution	L	10	1	1	1	0
	H	6	3,8	4.5	4.7	3.5
	α	9	7,8	7	7.4	0.3
	β	13	4,6	5	4.9	0.5
	γ	4	5,9	5.5	6.3	3.6
	δ	3	4,7	5	5.3	2.3
Mouse J contribution	L	10	1	1	1	0
	H	4	3,5	4	4	1.3
	α	17	6,10	8	8	0.8
	β	12	4,6	5	4.8	0.3
	γ	2	7,9	8	8	2
	δ	2	4,7	5.5	5.5	4.5

Labels are as in Table 1.

if constraints on α and β CDR3 length reflect a functional requirement for contact between CDR3 residues of both chains and antigenic peptides bound to MHC molecules (6), then γ/δ T cells may not as a general rule recognize peptide in association with MHC. Consistent with this, Schild et al. (23) have shown that when γ/δ T cells recognize MHC molecules, bound peptides do not appear to play any role in conferring specificity. Thus γ/δ T cells may occupy a unique niche in the immune system combining Ig-like recognition properties with cellular effector functions such as cytotoxicity and cytokine release.

We thank Mark Segal of the University of California, San Francisco and Lincoln Moses of Stanford University for statistical advice; Sunil Maulik of Oxford Molecular, Inc. for Ig model building; and Fredrik Ivars for communication of data before publication.

E. P. Rock was supported by a Predoctoral Fellowship from the Howard Hughes Medical Institute. This work was supported by grants from the National Institutes of Health (Y.-H. Chien) and the Howard Hughes Medical Institute (M. M. Davis).

Address correspondence to Dr. Yueh-hsiu Chien, Dept. of Microbiology and Immunology, Stanford University Beckman Center, HHMI, Stanford, CA 94305-5428.

Received for publication 18 May 1993 and in revised form 7 October 1993.

References

1. Patten, P., T. Yokota, J. Rothbard, Y. Chien, K. Arai, and M.M. Davis. 1984. Structure, expression and divergence of T-cell receptor beta-chain variable regions. *Nature (Lond.).* 312:40.
2. Novotny, J., S. Tonegawa, H. Saito, D.M. Kranz, and H.N. Eisen. 1986. Secondary, tertiary, and quaternary structure of T-cell-specific immunoglobulin-like polypeptide chains. *Proc. Natl. Acad. Sci. USA.* 83:742.
3. Chothia, C., D.R. Boswell, and A.M. Lesk. 1988. The outline structure of the T-cell alpha beta receptor. *EMBO (Eur. Mol. Biol. Organ.) J.* 7:3745.
4. Davies, D.R., E.A. Padlan, and S. Sheriff. 1990. Antibody-antigen complexes. *Annu. Rev. Biochem.* 59:439.
5. Engel, I., and S.M. Hedrick. 1988. Site-directed mutations in the VDJ junctional region of a T cell receptor beta chain cause changes in antigenic peptide recognition. *Cell.* 54:473.
6. Jorgensen, J.L., U. Esser, B. Fazekas de St. Groth, P.A. Reay, and M.M. Davis. 1992. Mapping T-cell receptor–peptide contacts by variant peptide immunization of single-chain transgenics. *Nature (Lond.).* 355:224.
7. Davis, M.M., and P.J. Bjorkman. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature (Lond.).* 334:395.
8. Chothia, C., and A.M. Lesk. 1987. Canonical structures for

the hypervariable regions of immunoglobulins. *J. Mol. Biol.* 196:901.

9. Chothia, C., A.M. Lesk, A. Tramontano, M. Levitt, S.J. Smith-Gill, G. Air, S. Sheriff, E.A. Padlan, D. Davies, W.R. Tulip, et al. 1989. Conformations of immunoglobulin hypervariable regions. *Nature (Lond.).* 342:877.
10. Singer, A., T. Mizuochi, T.I. Munitz, and R.E. Gress. 1986. Role of self antigens in the selection of the developing T cell repertoire. *In* Progress in Immunology VI. Academic Press, New York. 60–66.
11. Nikolic-Zugic, J., and M.J. Bevan. 1990. Role of self-peptides in positively selecting the T-cell repertoire. *Nature (Lond.).* 344:65.
12. Berg, L.J., G.D. Frank, and M.M. Davis. 1990. The effects of MHC gene dosage and allelic variation on T cell receptor selection. *Cell.* 60:1043.
13. Kabat, E.A., T.T. Wu, H.M. Perry, K.S. Gottesman, and C. Foeller. 1991. Sequences of Proteins of Immunological Interest. 5th ed. National Institutes of Health Publication No. 91-3242, Bethesda, MD.
14. Candéias, S., C. Waltzinger, C. Benoist, and D. Mathis. 1991. The $V\beta17^+$ T cell repertoire: skewed $J\beta$ usage after thymic selection; dissimilar CDR3s in $CD4^+$ versus $CD8^+$ cells. *J. Exp. Med.* 174:989.
15. Ivars, F. 1992. T cell subset–specific expression of antigen receptor beta chains in alpha chain–transgenic mice. *Eur. J. Immunol.* 22:635.
16. Altman, D.G. 1991. Practical Statistics for Medical Research. Chapman and Hall, London.
17. Bjorkman, P.J., M.A. Saper, B. Samraoui, W.S. Bennett, J.L. Strominger, and D.C. Wiley. 1987. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature (Lond.).* 329:512.
18. Brown, J.H., T.S. Jardetzky, J.C. Gorga, L.J. Stern, R.G. Urban, J.L. Strominger, and D.C. Wiley. 1993. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature (Lond.).* 364:33.
19. Elliott, J.F., E.P. Rock, P.A. Patten, M.M. Davis, and Y.H. Chien. 1988. The adult T-cell receptor delta-chain is diverse and distinct from that of fetal thymocytes. *Nature (Lond.).* 331:627.
20. Loh, E.Y., J.F. Elliott, S. Cwirla, L.L. Lanier, and M.M. Davis. 1989. Polymerase chain reaction with single-sided specificity: analysis of T cell receptor δ chain. *Science (Wash. DC).* 243:217.
21. Koop, B.F., R.K. Wilson, K. Wang, B. Vernooij, D. Zallwer, C.L. Kuo, D. Seto, M. Toda, and L. Hood. 1992. Organization, structure, and function of 95 kb of DNA spanning the murine T-cell receptor C_α/C_δ region. *Genomics.* 13:1209.
22. Jores, R., P.M. Alzari, and T. Meo. 1990. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. *Proc. Natl. Acad. Sci. USA.* 87:9138.
23. Schild, H., N. Mavaddat, C. Litzenberger, E.W. Ehrich, M.M. Davis, J.A. Bluestone, L. Matis, R. Draper, and Y. Chien. The nature of MHC recognition by $\gamma\delta$ T cells. *Cell.* In press.