Volume 12 Number 10 1984 Nucleic Acids Research

CG dinucleotide clusters in MHC genes and in 5' demethylated genes

Mark L.Tykocinski+ and Edward E.Max*

Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20205, USA

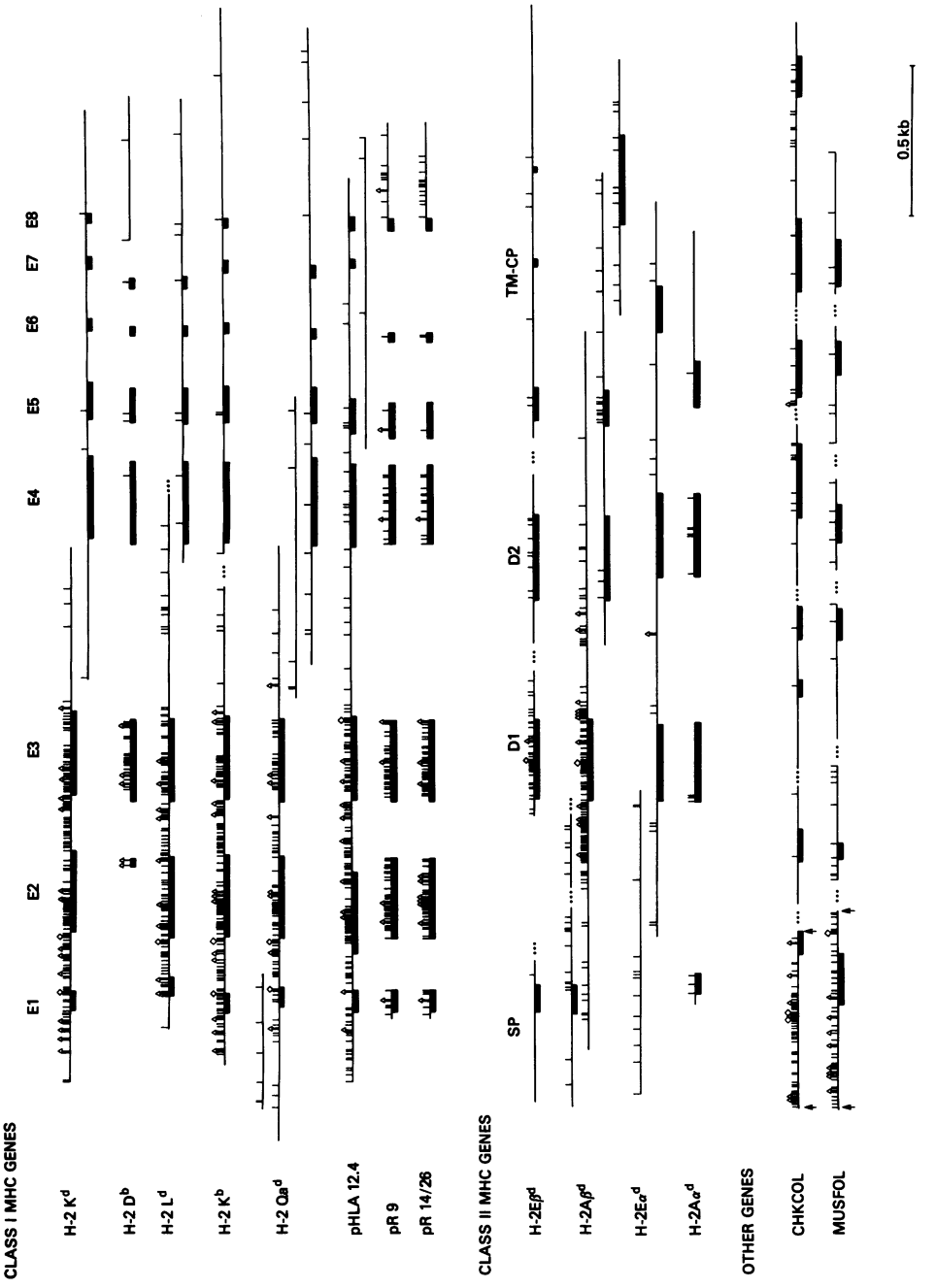
Received 23 January 1984; Revised and Accepted 24 April 1984

ABSTRACT

In the DNA of higher vertebrates the dinucleotide CG is unique in two respects: it occurs far less frequently than would be expected on the basis of the content of cytosine and guanine in a given DNA segment ("CG suppression") and it contains predominantly 5-methyl-cytosine, the only modified nucleotide common in vertebrate DNA. Here we point out the existence of CG clusters, i.e. localized lapses in the usual CG suppression, in two categories of DNA segments from vertebrates: around the polymorphic exons of major histocompatibility complex (MHC) genes and in the 5' regions of certain other genes. These observations contradict the recent suggestion that CG frequency is uniform over long contiguous segments of DNA containing several genes. A model for the origin of these CG clusters as a consequence of regional demethylation of germline DNA is supported by analysis of other sequence features of these regions as well as by previously published data on the methylation status in sperm DNA of two of these CG-rich regions.

INTRODUCTION

The dinucleotide CG is present in the genomic DNA of higher vertebrates as about 1% of all dinucleotides, compared with a frequency of about 4% expected from the content of C and G in the same DNA. Thus the ratio (CG)observed/(CG)expected is "suppressed" to about 0.25 (for review, see ref. 1). This CG suppression has been hypothesized to result from the fact that cytosine residues within the CG dinucleotide are largely methylated as 5-methylcytosine, with the dinucleotide ^{me}CG accounting for about 90% of the 5-methylcytosine in mammalian DNA. Over time, 5-methylcytosine tends to deaminate to thymidine, resulting in the depletion of CGs by conversion of ^{me}CG to either TG or CA (depending on which DNA strand underwent the mutation). This explanation has been supported (2) by the findings that: (i) the extent of CG suppression correlates with the content of 5-methylcytosine in various animal species (low in insects, higher in mammals) and (ii) CG suppression is generally accompanied by an excess of TG + CA dinucleotides.


CLASS I MHC GENES
E1
E2
E3
E4
E5
E6
E7
E8
H-2 K^d
H-2 D^b
H-2 L^d
H-2 K^b
H-2 Qa^d
pHLA 12.4
pR 9
pR 14/26
CLASS II MHC GENES
SP
D1
D2
TM-CP
H-2Eβ^d
H-2Aβ^d
H-2Eα^d
H-2Aα^d
OTHER GENES
CHKCOL
MUSFOL
0.5 kb

It has been suggested that CG frequency may be a property related to chromosomal constraints and thus may be uniform over long contiguous segments of DNA containing several genes (3). Here we point out several cases in which CG frequency varies significantly within genes. We have found clusters of CG dinucleotides in published sequences of the mouse dihydrofolate reductase (DHFR) gene, the chicken α2-collagen gene and in several genes of the major histocompatibility complex (MHC) from mouse and man. Such clusters apparently reflect very localized features within genes, features that may be important for understanding the molecular function of these genes.

CG Clusters

Class I MHC antigens are highly polymorphic integral membrane proteins that are present on virtually all mammalian cells and are major determinants of graft rejection (4,5). In our recent analysis of two rabbit class I MHC cDNA clones we observed a surprising clustering of the usually rare dinucleotide CG in the 5' half of the sequences, the region which includes the two domains exhibiting the greatest structural (5) and functional (6) polymorphism. Because of the potential significance of the CG dinucleotide in methylation-related gene regulation, Z-DNA formation and mutation rates, we examined the occurrence of this dinucleotide in the published sequences of other class I MHC genes and discovered that the asymmetry of CG distribution is a consistent finding in all these sequences.

The distribution of CG dinucleotides in reported class I MHC sequences is displayed in Fig. 1 in relation to the exon-intron structure of the

Figure 1. Display of CG occurrence in selected genes. The position of the dinucleotide CG is shown for each gene (identified on the left of the figure) by a small vertical tick mark. The open triangles and diamonds indicate CGCG and CGCGCG, respectively. Coding domains are indicated by bold rectangles below the line representing each gene. Exons for class I MHC genes are identified by E1, E2 ...; the exons for class II MHC genes are indicated by SP (signal peptide), D1 and D2 (first and second external domains) and TM-CP (transmembrane, cytoplasmic region. Long sequences have been broken into several lines; missing sequence data is indicated by "..." Arrows below the CHKCOL and MUSFOL genes indicate the limits of the DNA segments considered as "5' regions" in Table III. References for the sequences are as follows, where # indicates data accessed through the Los Alamos GENBANK database: H-2K^{d} -14; H-2D^{b} -15; H-2L^{d} -16; H-2K^{b} -13; H-2Qa^{d} -17; pHLA 12.4 -19; pR9 -18; pR14/26 -18; H-2Eβ^{d} -21; H-2Aβ^{d} -20; H-2Eα^{d} -26#; H-2Aα^{d} -25; CHKCOL (chicken α2-(I) collagen)- 27# 28#; MUSFOL (mouse dihydrofolate reductase) -29#,30#,31#.

Table I

CG frequencies: 5' and 3' regions of class I MHC genes

	5' region (5'UT-E3)				3' region (E4-3'UT)				Reference
	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	
H-2 (mouse)									
K^b (g)	.111	.114	.974	.98	.003	.062	.042	1.41	13
K^d (g)	.111	.116	.958	1.03	.004	.062	.064	1.45	14
D^b (c)[1]	.087	.095	.919	1.32	.007	.070	.106	1.38	15
L^d (g)	.099	.108	.920	1.06	.005	.053	.086	1.37	16
Qa^D (g)	.083	.109	.761	1.09	.005	.066	.076	1.45	17
RLA (rabbit)									
pR9 (c)	.103	.111	.932	1.11	.037	.093	.394	1.40	18
pR14/26(c)	.121	.116	1.043	1.06	.043	.094	.459	1.41	18
HLA (human)									
pHLA 12.4 (g)	.099	.114	.864	1.01	.010	.068	.152	1.27	19

[1] 5' sequence incomplete; only E3 used for 5' calculations.
g = genomic sequence - exons and introns included in calculations
c = cDNA sequence - calculations include only exons.

loci: E1 encodes a signal peptide; E2 and E3 encode the first two external domains including several polymorphic regions important in cytotoxic T cell recognition; E4 encodes the third external domain, which seems to be highly conserved (consistent with its proposed primary function of binding β_2-microglobulin); E5 encodes a transmembrane domain; and E 6, 7 and 8 encode short cytoplasmic domains and 3' UT. The clustering of CG dinucleotides in the 5' exons of all the genes is readily apparent from Figure 1.

The relatively high CG frequency observed in the 5' exons could reflect some functional significance of this dinucleotide or could simply result from the high content of C and G in these exons. To distinguish between these possibilities, we calculated the ratio between the observed CG frequency and that expected for a random sequence with the observed C and G content. (The expected CG frequency is simply the product [C frequency] X [G frequency] for the DNA segment being analyzed.) Table I demonstrates that the asymmetry of CG distribution in class I genes is not simply a result of regional differences in C and G content; although the expected CG frequency is somewhat higher on the 5' end of the genes (reflecting the higher C + G content), the ratio of (CG)observed/(CG)expected is consistently near 1 in the 5' regions but significantly and consistently below 1 in the 3' regions. The rabbit genes, in which CG clustering was first noted, in fact show the least dramatic asymmetry; the human HLA pseudogene pHLA 12.4 is intermediate, and the mouse sequences are most asymmetric in CG distribution. Fig. 2a graphically illustrates an analysis of mouse class I sequences with each exon and intron examined individually. This figure demonstrates that both exons and introns share the CG asymmetry.

Positional analysis of the CG dinucleotides in these class I genes reveals considerable variation in the precise localization of individual CGs. For instance, when sequences of the rabbit class I cDNA pR9 and the mouse H-2K^b gene are compared in the 5' exons E2 and E3, 47% of the CGs in the former are not found at the same site in the latter. This CG positional variability strongly argues for maintenance of the regional feature of CG-richness, rather than simply the conservation of multiple individual CG dinucleotides. Within coding domains, the clustered CG dinucleotides appear in all three phases relative to the coding frame, although they are most frequent in the third (codon-spanning) position. The clustered CGs appearing in the first two codon positions necessarily affect the statistics of codon usage for amino acids encoded by CG-containing triplets (Arg, Ser, Thr, Pro and Ala). An explanation of the CG clustering in terms of amino acid

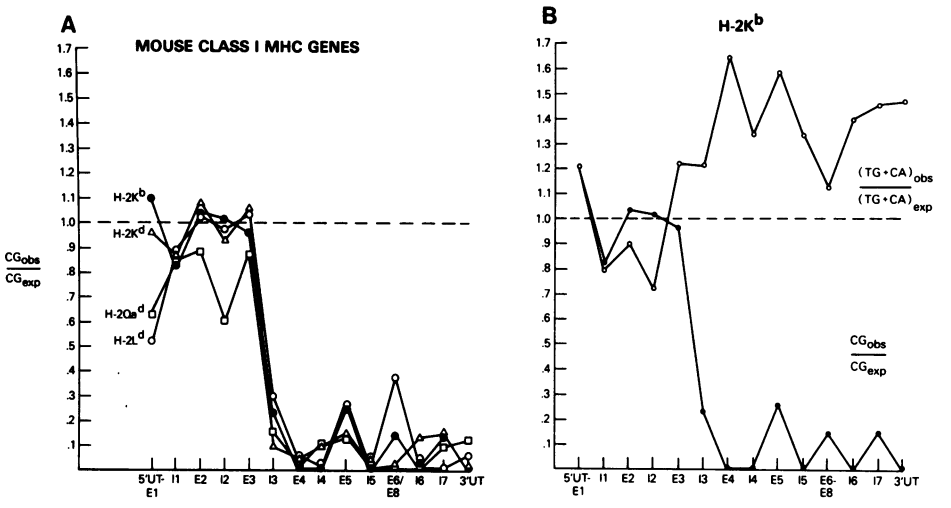


Figure 2. Dinucleotide frequencies corrected for mononucleotide content. The observed dinucleotide frequencies have been calculated individually for the indicated exon (E) and intron (I) regions and have been normalized to the expected frequencies as described in the text. Figure 2a demonstrates that the corrected CG freqencies are comparable for the four indicated murine MHC class I genes, and that the asymmetry is shared by both exons and introns. Figure 2b shows that CG suppression in the 3' region of the H-2K^b gene is accompanied by partially compensating elevation of the (observed/expected) ratio for TG + CA.

selection at the protein function level seems unlikely because (i) these 5 amino acids are all encoded by non-CG-containing triplets as well; (ii) exons E2 and E3 of the class I genes apparently tolerate many non-conservative amino acid replacements anyway; and (iii) as already pointed out, the elevated CG frequency is a property of introns as well as coding sequences.

A similar analysis was performed on the genes for class II MHC antigens. These proteins, found primarily on B lymphocytes and macrophages, mediate cellular immune interactions; they are composed of two chains (α and β) each of which has two external domains plus transmembrane and cytoplasmic domains (7). Class II MHC genes revealed similarly striking CG clustering (Fig. 1) with a somewhat different pattern from that seen in class I genes. The β chain genes showed CG clustering surrounding the exon encoding the first external domain, the main locus of polymorphism, but the clustering does not extend to the 5' end of the gene. The much less polymorphic α chains showed no comparable CG clustering. Table II compares the CG frequency (observed/expected) for the two external coding domains of some class II genes

Table II

CG Frequencies: first and second external domains of class II MHC genes

	First domain (5')				Second domain (3')				Reference
	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	
β chain genes									
$I\text{-}A_\beta^d$	.118	.107	1.10	1.14	.011	.079	.139	1.53	20
$I\text{-}E_\beta^d$	.097	.083	1.17	1.05	.032	.080	.400	1.43	21
$DR_\beta 79$	.067	.085	.788	1.18	.021	.079	.266	1.42	22
$DC_\beta 1$	.086	.089	.966	1.22	.028	.084	.333	1.50	23
$SB_\beta 1$	.084	.093	.903	1.11	.025	.081	.309	1.47	24
α chain genes									
I-Aα	.011	.050	.220	1.39	.018	.074	.243	1.60	25
I-Eα	.020	.050	.400	1.44	.028	.076	.368	1.33	26

Table III

CG frequencies: 5' and 3' regions of selected non-MHC genes

	5' region[1]				3' region[1]				Reference
	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	(CG)obs	(CG)exp	$\frac{(CG)obs}{(CG)exp}$	$\frac{(TG+CA)obs}{(TG+CA)exp}$	
α2-collagen (chicken)	.105	.103	1.02	.90	.012	.044	.273	1.24	27,28
DHFR (mouse)	.087	.100	.870	1.01	.013	.036	.360	1.08	29,30,31

[1] For this analysis the entire 5' sequenced segment of each gene (see arrows on Fig. 1) was arbitrarily designated "5' region" and the sequences remaining 3' were considered "3' regions".

of mouse and man; these data confirm that the CG elevation is confined to the first external domain, is not simply a consequence of the elevated C+G content, is found only in class II β genes (not α) and is conserved between mouse and man.

The observation of CG clustering in genes of the MHC prompted us to consider how general this feature might be in other DNA sequences. In a survey of 83 vertebrate gene sequences, Smith et al. (8) reported absence of CG suppression only in the human α-2-globin (but not β globin) genes, X. leavis rRNA genes and mouse κ J region coding sequences (although the latter represents a rather small sample of 195 bp including 5 highly conserved repeats of an even smaller segment). However, in a non-systematic survey we have noted two additional and particularly instructive examples of genes with CG clusters. As shown in Fig. 1, the 5' regions of the chicken α-2 collagen and mouse DHFR genes also demonstrate elevated CG frequency relative to more 3' regions of the same genes. As documented in Table III, the observed CG clustering in these genes is not just a consequence of elevated C+G frequency.

Asymmetries of CG distribution like those documented in the present paper--5' exons vs 3' exons--have not to our knowledge been previously noted for any other genes, although McClelland and Ivarie (9) demonstrated reduced CG suppression in 5' flanking and untranslated regions (vs 3' flanking and untranslated) in a composite calculation based on 15 mammalian gene sequences.

A model

We would like to suggest a model that could explain the CG clusters that we have observed. This model proposes that the 5' regions of certain genes may, in germline DNA, be maintained in a highly demethylated state relative to the 3' regions. As a result, the 5-methylcytosine deamination mutations would be rare on the 5' of the genes, so that CG dinucleotides would be preserved at their "natural" frequency, whereas the more methylated 3' regions would

undergo the usual deamination-related CG loss.

Support for the model

If the observed CG clusters resulted from regional reduction in the frequency of the mutation of ^{me}CG dinucleotides to TG or CA, then one would expect to find these regions lacking the usual elevation of TG+CA frequency that is found in CG-suppressed DNA. In fact, analysis of the nucleotide sequences of these genes (see Fig. 2b and Tables I, II, and III) demonstrates that the 5' regions showing CG clustering all have a lower [observed/expected] ratio for TG+CA occurrence than the 3' regions of the same gene. The highest TG+CA ratios in the 5' regions of the genes analyzed here occur in the class II α chains, which lack CG clustering.

If the attrition of CG dinucleotides in these regions is prevented by their demethylated state in germline DNA, then one might expect to be able to demonstrate this by examining methylation of these regions in sperm DNA. Although the multiplicity of sequences cross-hybridizing with MHC probes makes the methylation of these genes difficult to study, the mouse DHFR and chicken α-2 collagen genes have both been examined by testing their susceptibility to cleavage by MspI, HpaII or AvaI. The 5' regions of both genes in sperm DNA were found to be demethylated relative to the 3' regions, with identical results found in DNA of all other tissues examined (10,11). (It should be pointed out that the multiplicity of HpaII/MspI sites in these regions may result in an overestimate of demethylation as assessed by the HpaII/MspI method; the demethylation in these regions needs to be verified by other techniques, such as genomic sequencing methods [G. Church and W. Gilbert, personal communication]). Thus currently available data support the model attributing the observed CG clusters to germline demethylation of these regions.

Speculations about function

What is the significance of CG clusters? One hypothesis for considering these findings would be that the CG clustering is a functionally insignificant consequence of the demethylation of these regions in germline cells, which preserves CG dinucleotides from deamination-mediated mutation as discussed above. This hypothesis leads logically to the questions of (i) the function of the localized demethylation in these regions and (ii) the mechanism of their maintenance in the demethylated state. As to function, demethylation of genes, especially in 5' regions, has been correlated with gene activation; in

many genes the CG dinucleotides that can be assessed for methylation by restriction endonuclease digestion have been found to be demethylated in tissues where the genes are expressed but fully methylated in tissues where they are inactive. This notion led Stein et al. (10) to hypothesize that the genes for DHFR and adenine phosphoribosyl transferase (both of which were found to be 5' demethylated in sperm and other tissues) might remain demethylated as a result of their "housekeeping" status, i.e. because the encoded enzymes must be present everywhere for general metabolic functions necessary in all cells. While class I MHC genes are also widely expressed it is not clear that class II MHC genes or collagen genes should be considered "housekeeping" genes. Moreover, in at least one study (12), demethylation was found to be correlated with suppression rather than activation of gene expression, weakening the generality of the demethylation/gene activation correlation and the "housekeeping" hypothesis as an explanation applicable to CG clusters. The question of how these CG-rich regions might be maintained in the demethylated state in the germline remains open, since the mechanisms regulating methylation-demethylation are not yet understood.

In contrast to the view of CG clusters as an inconsequential result of regional demethylation, these clusters may be maintained directly by natural selection because they serve a specific function; the role of germline demethylation might then be to preserve the CG clusters from deamination-mediated mutation. A high density of CGs may confer significant structural features to DNA. An SV-40 nucleosome-free region that is associated with transcriptional enhancer activity is CG-rich (8).

Does the CG clustering have the same significance in MHC genes as in the other genes we have considered? Possibly, but several arguments suggest that the clusters in MHC genes may be somewhat different. Most striking is the correlation between CG clusters and the MHC domains that exhibit the greatest polymorphism: E2 and E3 of class I genes and the β1 domain of class II genes. It is tempting to speculate that the CG-rich environment in the MHC genes might facilitate a zonal diversifying mechanism that promotes the polymorphisms found in the exons of these regions, e.g. by establishing conditions that inhibit DNA repair. (Clearly such a mechanism would require recognition of other sequence features in addition to CG-richness to avoid detrimental diversification of CG-rich regions in genes like dihydrofolate reductase and collagen.) An alternative notion, that the CGs could (by their mutation to TG and CA) directly cause the extensive polymorphisms found in these regions, is contradicted by two points: the low ratio

[observed/expected] for TG+CA in these regions; and the fact that, of the differences between homologous MHC genes, only a small minority (e.g. about 11% for pR9 vs H-2K^b) represent CG-TG or CG-CA replacements.

Although the physiologic role of the CG clusters we have noted is not yet known, the evolutionary conservation of this feature in MHC genes argues for some functional significance, which we anticipate will be clarified by future investigation. At present, we conclude that peculiarities of CG distribution can apparently be quite localized, rather than always involving long contiguous multigene segments as has been suggested (3). Furthermore, we propose that regions of CG clusters may represent an "imprint"--produced over evolutionary time--of regions of DNA that in the germline are maintained in a demethylated state.

ACKNOWLEDGEMENTS

We gratefully acknowledge helpful discussions with Drs. Thomas Kindt, David Margulies, Eric Long, Ron Germain and Gary Felsenfeld, as well as the editorial skills and patience of Virginia Shaw.

+Present address: Institute of Pathology, Case Western Reserve University, Cleveland, OH 44106, USA

*To whom correspondence should be addressed

REFERENCES

1. Ehrlich, M. & Wang, R. Y.-H. Science 212, 1350-1357 (1981).
2. Bird, A.P. Nucl. Acids Res. 8, 1499-1504 (1980).
3. Lennon, G.G. & Fraser, N.W. J. Mol. Evol. 19, 286-288 (1983).
4. Klein, J. Science 203, 516-521 (1979).
5. Kimball, E.S. & Coligan, J.E. Contemp. Topics Immunol. 9, 1-63 (1983).
6. Evans, G.A., Margulies, D.H., Shykind, B., Seidman, J.G. and Ozato, K. Nature 300, 755-757 (1982).
7. Shackelford, D.A., Kaufman, J.F. & Strominger, J.L. Immunol. Rev. 66, 133-187 (1982).
8. Smith, T.F., Waterman, M.S. & Sadler, J.R. Nucl. Acids Res. 11, 2205-2220 (1983).
9. McCelland, M. & Ivarie, R. Nucl. Acids Res. 10, 7865-7877 (1982).
10. Stein, R., Sciaky-Gallili, N., Razin, A. and Cedar, H. Proc. Natl. Acad. Sci. USA 80, 2422-2426 (1983).
11. McKeon, C., Ohkubo, H., Pastan, I. & de Crombrugghe, B. Cell 29, 203-210 (1982).
12. Tanaka, K., Appella, E. & Jay, G. Cell 35, 457-465 (1983).
13. Weiss, E., Golden, L., Zakut, R., Mellor, A., Fahrner, K., Kvist, S. & Flavell, R.A. EMBO J. 2, 453-462 (1983).
14. Kvist, S., Roberts, L. & Dobberstein, B. EMBO J 2, 245-254 (1983).
15. Reyes, A.A., Schöld, M. & Wallace, R.B. Immunogenetics 16, 1-9, (1982).
16. Moore, K.W., Sher, B.T., Sun, Y.H., Eakle, K.A. & Hood, L. Science 215, 679-682 (1982).

17. Steinmetz, M., Moore, K.W., Frelinger, J.G., Sher, B. T., Shen, F-W, Boyse, E.A. & Hood, L. Cell 25, 683-692 (1981).
18. Tykocinski, M., Marche, P., Max, E. E. & Kindt, T. J. manuscript submitted (1983).
19. Malissen, M., Malissen, B. & Jordan, B.R. Proc. Natl. Acad. Sci. USA 79, 893-897 (1982).
20. Malissen, M., Hunkapiller, T. and Hood, L. Science 221, 750-754 (1983).
21. Saito, H., Maki, R.A., Clayton, L.K. & Tonegawa, S. Proc. Natl. Acad. Sci. USA 80, 5520-5524, (1983).
22. Long, E.O., Wake, C.T., Gorski, J. & Mach, B. EMBO J. 2:389-394 (1983).
23. Larhammar, D., Hyldig-Nielsen, J.J., Servenius, B., Andersson, G., Rask, L. & Peterson, P.A. Proc. Natl. Acad. Sci. USA 80, 7313-7317 (1983).
24. Gorski, J., Long, E.O. & Mach, B. manuscript in preparation.
25. Benoist, C.O., Mathis, D.J., Kanter, M.R., Williams V.E.II, & McDevitt, H.O. Cell 34, 169-177 (1983).
26. McNicholas, J., Steinmetz, M., Hunkapiller, T., Jones, P. & Hood, L. Science 218, 1229-1232 (1982).
27. Vogeli, G., Ohkubo, H., Sobel, M.E., Yamada, Y., Pastan, I. & de Crombrugghe, B. Proc. Natl. Acad. Sci. USA 78, 5334-5338 (1981).
28. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. & Doty, P. Nature 294, 129-135 (1981).
29. Crouse, G.F., Simonsen, C.C., McEwan, R.N. & Schimke, R.T. J. Biol. Chem. 257, 7887-7897 (1982).
30. Nunberg, H.J., Kaufman, R.J., Change, A.C.Y., Cohen, S.N. & Schimke, R.T. Cell 19, 355-364 (1980).
31. Simonsen, C.C. & Levinson, A.D. Proc. Natl. Acad. Sci. USA 80, 2495-2499 (1983).