The EMBO Journal vol.7 no.11 pp.3441–3449, 1988

Searching for coding sequences in the mammalian genome: the *H-2K* region of the mouse MHC is replete with genes expressed in embryos

Kuniya Abe, Jing-Fang Wei[1], Fu-Sheng Wei[1], Yu-Chih Hsu[1], Hiroshi Uehara, Karen Artzt and Dorothea Bennett

Department of Zoology, The University of Texas at Austin, Austin, TX 78712-1064, and [1]Department of Immunology and Infectious Diseases, The Johns Hopkins University, Baltimore, MD 21205, USA

Communicated by M.F.Lyon

We have searched for expressed genes in 170 kb of cosmid cloned DNA from the *H-2K* region of the mouse MHC. This region is known to contain two genes, *H-2K* and *K2*. We identified unique/low copy sequences evenly spaced along the cloned DNA, and used these as probes to search for conserved sequences in Southern blots from a variety of mammalian species. The majority of the unique sequences were found to have homologues and most of these were associated with CpG non-methylated islands. Northern blot analysis and isolation of clones from 5.5 and 10.5-day embryo cDNA libraries showed five additional genes encoded in the *H-2K* region. Four of these are abundant in embryos; the fifth is exclusively expressed in lymphoid cells. Our data indicate a minimum of seven genes in 170 kb, an unexpectedly high gene density. These results differ from two recent studies where similar lengths of cloned DNA were examined for expressed genes, and only one, or a part of one gene was found. The combined data suggest that the spatial organization of genes in the mammalian genome may not be random.

Key words: t-complex/*H-2* complex/mouse embryos

Introduction

Advances in molecular techniques have made it feasible to search any segment of cloned DNA for expressed genes even if the nature of the gene products is not defined. Thus, genes that are of interest and whose chromosomal location is known can be identified and characterized. This has been a particularly fruitful approach for finding human disease-causing genes, even with the limitations imposed by available human pedigrees (Friend *et al.*, 1986; Monaco *et al.*, 1986; reviewed by Orkin, 1986). Surprisingly, similar studies have not been reported until now in mice, where sensitive genetic breeding experiments can be designed and analysed.

An excellent primary target for a 'reverse genetics' approach in the mouse is the region of chromosome 17 occupied by the *H-2* complex and the mutant *t*-complex. The *H-2* complex has been subjected to closer scrutiny than any other genetic entity of the mouse, and was also the first substantial region of the mammalian genome to be cloned by 'chromosomal walking' (Hood *et al.*, 1983; Steinmetz *et al.*, 1986). Molecular studies using >1600 kb of cloned DNA have provided significant information for understanding the organization and evolution of the mammalian chromosome as well as genes essential for immune function. The *t*-complex has also been well-studied genetically, especially with respect to the locations of its various lethal mutations, and importantly in the present context, four of these lethal genes have been found to map very close to or within the *H-2* complex (Artzt, 1984; Shin *et al.*, 1984). In fact, the t^{w5} mutation, an early acting embryonic lethal, has proved to be recombinationally inseparable from *H-2K*. Statistical considerations thus locate t^{w5} within 0–250 kb of *H-2K* (Artzt *et al.*, 1988). We have cloned into cosmids ~170 kb of DNA containing the entire *H-2K* region from each of two *t*-haplotypes, t^{w5} and t^{12}, by taking advantage of available probes and information about the structure of the wild-type complex. The structure of the *H-2K* region of these two haplotypes was thoroughly analysed by high resolution restriction mapping to permit comparisons between the two mutant chromosomes, and between them and wild-type chromosomes (Uehara *et al.*, 1987).

In the study we report here, we first used this already cloned DNA to expand the chromosomal walk. We then tried to establish some basic strategies for searching for coding sequences in a relatively large stretch of DNA. We chose the *H-2K* region for two reasons. First, the mouse MHC is one of the most well-characterized regions in the mammalian genome, and information about novel genes linked to the MHC may provide clues to full understanding of this genetic complex. Second, we have been interested in genes that play important roles during embryogenesis, and one such 'developmental mutation', t^{w5}, has been localized very close to the *H-2K* gene. We searched as thoroughly as possible for expressed sequences in the *H-2K* region, and found five novel genes in ~150 kb. Four of these genes are expressed in early embryos, and can therefore be considered as candidates for the t^{w5} gene. The fifth is expressed predominantly in macrophages.

Results

'Chromosomal walking' to clone the H-2K region from the t^{w5} haplotype: identification of unique/low copy sequences

Since a long-term goal of ours is to identify and analyse the t^{w5} lethal gene, and since previous genetic analysis gave an estimate of <250 kb for the region around *H-2K* that contains t^{w5}, we extended a previously reported chromosome walk (Uehara *et al.*, 1987) to include the entire *H-2K* region and a part of the *I*-region. Overlapping cosmids that define ~240 kb of DNA have been aligned after restriction mapping. In the course of this experiment, we defined the centromeric limit of the t^{w5} gene to a point ~65 kb proximal to the *H-2K* gene (Artzt *et al.*, 1988) (see Figure 1, zig-zag line). We therefore chose to focus on the region leftward of this limit. Figure 1 shows the restriction map of the *H-2K* region of the t^{w5} haplotype. Comparison of the

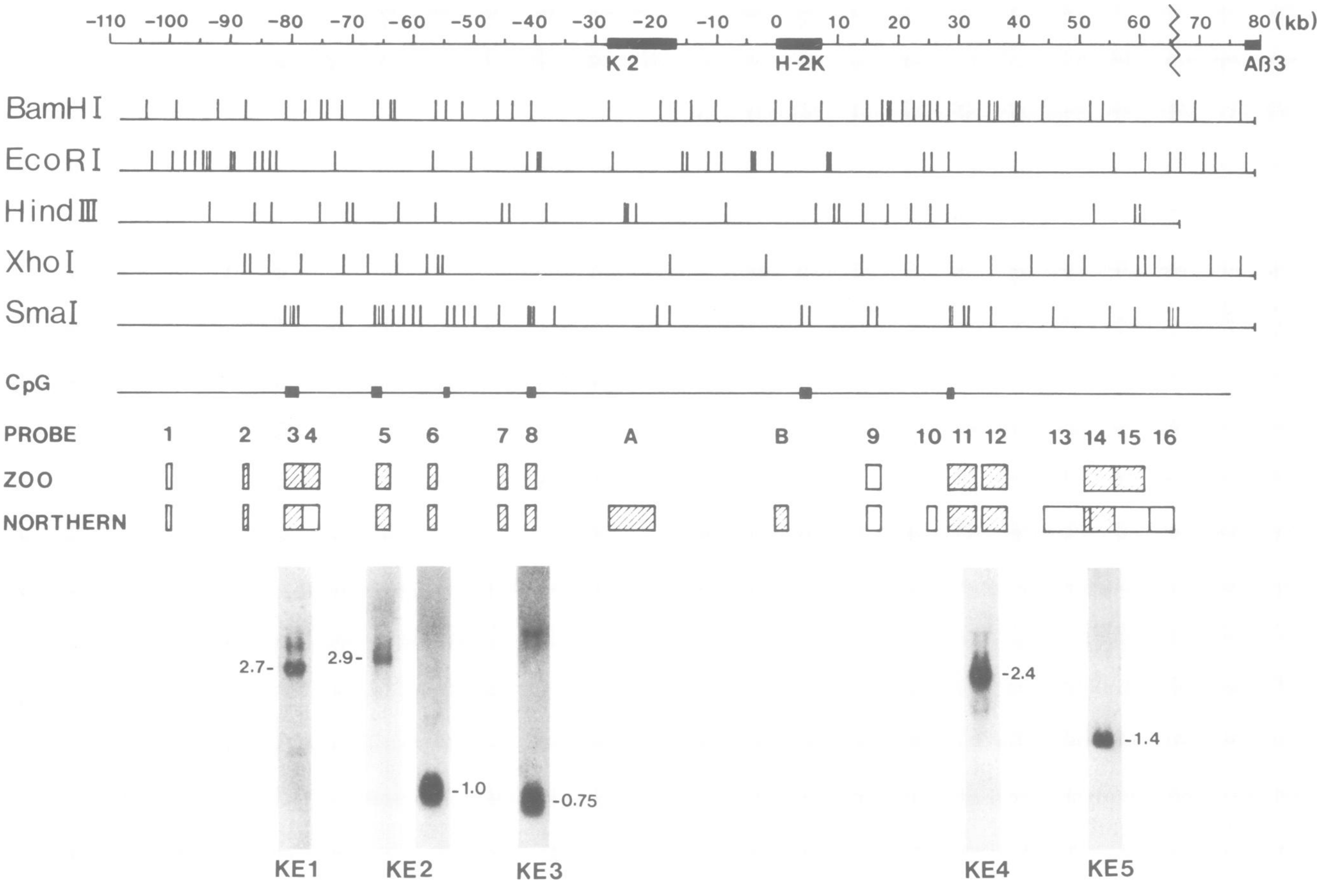
-110 -100 -90 -80 -70 -60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 (kb)
K 2
H-2K
Aβ3
BamHI
EcoRI
HindIII
XhoI
SmaI
CpG
PROBE
1 2 3 4 5 6 7 8 A B 9 10 11 12 13 14 15 16
ZOO
NORTHERN
2.7-
2.9-
-1.0
-0.75
-2.4
-1.4
KE1
KE2
KE3
KE4
KE5

Fig. 1. Molecular map of the *H-2K* region of the t^{w5} haplotype. The vertical lines indicate the points of restriction sites for *Bam*HI, *Eco*RI, *Hin*dIII, *Xho*I and *Sma*I. The coordinates show the distance in kb starting at 0 for the position of the 2.3 kb *Bam*HI fragment homologous with the 3′ portion of the *H-2K* class I gene. The locations of the *H-2K* and *K2* class I genes, and of the class II gene, *Aβ3*, are shown as solid boxes at the top of the figure. Since the MHC is inverted in *t*-haplotypes (Shin *et al.*, 1983), the centromere is on the right-hand side of the map, and the telomere is on the left. The vertical zig-zag line at about +65 kb indicates the centromeric limit of the region where the t^{w5} gene should reside (Artzt *et al.*, 1988). The locations of non-methylated CpG-rich islands are shown below the restriction map. The location of hybridization probes is indicated by numbers and their length corresponds to the size of the boxes below them. Positive hybridization of these probes to either zoo blots or Northern blots are shown by hatched boxes (▨), while open boxes (□) indicate no hybridization. Hybridization of probes 3, 5, 6, 8, 11 + 12 to total RNA from 10 day mouse embryos is shown at the bottom. Hybridization of probe 14 to spleen RNA is also shown. Probe A is a 8.9 kb *Bam*HI fragment from cosmid no. 12, which covers most of the *K2* class I sequence. Probe A was pre-competed with mouse DNA to eliminate the hybridization with repetitive sequences. Probe B (2.3 kb *Bam*HI fragment from cosmid 9-2-5) contains the 3′ portion of the *H-2K* gene.

t^{w5} restriction map to the reported map of inbred mice revealed that the structure of the *H-2K* region of the t^{w5} haplotype is extremely similar to that of inbred mice except that the whole region of the MHC is, of course, inverted (Artzt, 1984; Shin *et al.*, 1983). The locations of the two class I genes in this region, *K* and *K2*, and of the class II gene, *Aβ3*, are shown at the top of Figure 1. In order to identify additional genes in the ~170 kb of cloned DNA distal to the centromeric limit of the t^{w5} mutation, we first defined unique/low copy sequences by hybridizing radiolabeled total mouse genomic DNA to Southern blots of restriction enzyme digested cosmid DNA. Of 44 fragments which did not contain highly repetitive sequences, we chose 16 restriction fragments to use as probes. These were spaced evenly along the cloned DNA.

Most of the unique sequences studied in the H-2K region are evolutionarily conserved

If genes have important biological functions, their sequences should be expected to be conserved in evolution. Thus, any sequences that are conserved across species might represent coding exons of genes (Monaco *et al.*, 1986). To examine this possibility, we used the unique copy sequences obtained in the cosmid walk as hybridization probes on Southern blots of DNA from various mammals: rodents (rat and Chinese hamster), mink, cat and human. Hybridization and washing of such 'zoo blots' were performed under moderately high stringency. Of 13 probes examined in this way, all except one hybridized to rat DNA. This result was expected since mouse and rat share considerable homology even in noncoding sequences (Sheppard and Gutman, 1982). However, it was surprising that 11 out of 13 sequences also hybridized to counterparts in DNA from mammals phylogenetically distant from rodents (Figure 1). Figure 2 shows examples of zoo blots clearly demonstrating that mouse sequences represented by probe 5, probe 8 or probes 11 plus 12 were conserved in all mammalian genomes that we examined. In general hybridization signals tend to become weaker as evolutionary distance increases. The results of the zoo blot analyses are summarized in Table I.

Most of the conserved sequences studied are expressed

The highly conserved sequences thus detected were

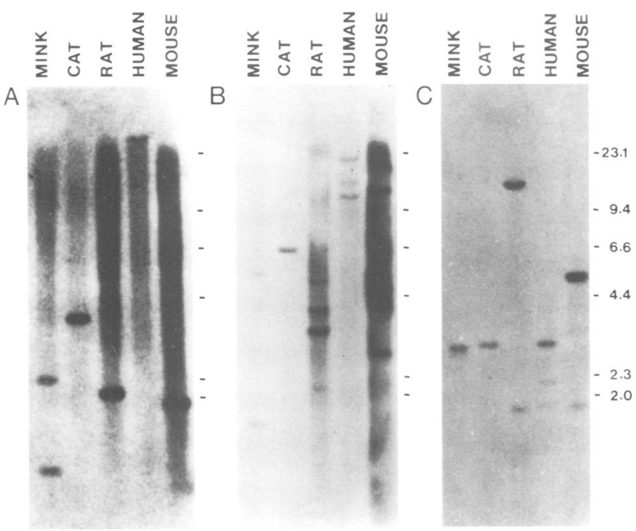

Fig. 2. Interspecies sequence conservation of the probes derived from the *H-2K* region. 10 μg of DNA from mouse, rat, cat, mink and human were digested with *Bam*HI, Southern blotted and hybridized with the probes 5(**A**), 8(**B**), and 11 + 12(**C**). Hybridizations were carried out under moderately high stringency (hybridized in 50% formamide. 5 × SSC at 42°C; wash in 0.1 × SSC, 0.1% SDS at 50°C or in 1 × SSC, 0.1% SDS at 65°C).

Table I. Hybridization of unique/low copy probes to DNA of various species

Probe	Mouse	Rat	Chinese hamster	Cat	Mink	Human	Detects transcripts in mouse RNA
1	+	–	–	–	–	–	–
2	+	+	ND	+	–	–	+
3	+	+	+	+	+	–	+
4	+	+	+	+	+	–	–
5	+	+	ND	+	+	+	+
6	+	+	+	+	+	+	+
7	+	+	ND	+	+	+	+
8	+	+	ND	+	+	+	+
9	+	+	ND	–	–	–	–
10	+	ND	ND	ND	ND	ND	–
11	+	+	+	+	+	+	+
12	+	+	+	+	+	+	+
13	+	+	ND	ND	ND	ND	–
14	+	+	+	–	–	+	+
15	+	+	ND	+	+	+	–
16	+	+	ND	–	–	–	–

(+) positive hybridization; (–) negative hybridization; ND, not determined.

The probes represent the following fragments: (**1**) 1.1 kb *Bam*HI–*Eco*RI fragment from cosmid clone, 9-2-3 (**2**) 0.9 kb *Xho*I fragment from clone 9-2-3; (**3**) 2.9 kb *Bam*HI fragment from 10-6-5; (**4**) 2.5 kb *Bam*HI fragment from 10-6-5; (**5**) 1.7 kb *Bam*HI fragment from 9-2-3; (**6**) 1.4 kb *Xho*I fragment from 3-6-3; (**7**) 1.8 kb *Pst*I fragment from 3-6-3; (**8**) 1.7 kb *Eco*RI fragment from 3-6-3; (**9**) 2.9 kb *Bam*HI fragment from 9-2-5; (**10**) 1.9 kb *Bam*HI fragment 9-2-5; (**11**) 4.9 kb *Bam*HI fragment from 3-1-2; (**12**) 3.9 kb *Pst*I fragment from clone 3-1-2; (**13**) 7.1 kb *Bam*HI fragment from 3-1-2; (**14**) 3.9 kb *Pst*I fragment from 3-1-2 (same size as probe 12, but derived from different location of the same cosmid clone); (**15**) 5.1 kb *Eco*RI fragment from 3-1-2; (**16**) 4.4 kb *Eco*RI fragment from 3-1-2. All the cosmid clones were described in the previous report (Uehara *et al.*, 1987).

Table II. Frequency of sites of CpG enzymes in the *K*-region DNA

Enzyme	Number of sites	Recognition sequence	Sites per genome[a]
*Not*I	3	GCGGCCGC	5 × 10^2
*Nru*I	2	TCGCGA	2.7 × 10^4
*Pvu*I	2	CGATCG	2.7 × 10^4
*Xma*III	5	CGGCCG	1.2 × 10^4
*Mlu*I	2	ACGCGT	2.7 × 10^4
*Sac*II	17	CCGCGG	1.2 × 10^4
*Bss*HII	10	GCGCGC	1.2 × 10^4
*Sma*I	38	CCCGGG	4.8 × 10^4

[a] Calculated distribution of CpG endonuclease sites in mammalian DNA (genome size = 3 × 10^9 bp), taken from Lindsay and Bird (1986).

subsequently used to probe Northern blots of RNA from 10 day mouse embryos or from embryonal carcinoma (EC) cells. The Northern blot data provide evidence that a number of transcripts are encoded by the evolutionarily conserved sequences. The locations of the transcriptionally active sequences and a summary of the Northern blot analyses are shown (Figure 1). Since several different transcript sizes were detected, a simplistic interpretation predicted the existence of several different genes in the region. Probes 2 and 3 both detected a transcript of 2.7 kb. Probe 5 hybridized to a 2.9 kb transcript, while probe 6 revealed a transcript of 1.0 kb. Probe 7 exhibited weak hybridization to a 1.2 kb transcript (data not shown). Probe 8 gave a signal at 0.75 kb and probes 11 and 12 both detected a transcript with a size of 2.4 kb. The size of each transcript was carefully measured. Especially when the size of one transcript was close to that of a transcript detected with a neighboring probe, the two different probes were used in turn in hybridizations to the same RNA blot to confirm the size difference. Although probe 14 was negative for embyro RNA, subsequent study showed that it detected a 1.4 kb transcript in RNA from spleen or thymus. Finally, probes 4 and 15 did not hybridize to any RNA that we examined. However, both of these are located very close to transcriptionally active sequences, which suggests that they might contain immediately flanking sequences that could account for their evolutionary conservation.

The H-2K region contains an unusually high density of sites for rare-cutting CpG enzymes

The vertebrate genome is low in G+C and the dinucleotide CpG is under-represented within the genome (Bird, 1986). Therefore, enzymes with CpG as part of their recognition motif (CpG enzymes) cleave the DNA infrequently. However, when the *K*-region restriction map was determined, we noticed a high density of sites for CpG enzyme *Sma*I (CCCGGG). Subsequent analysis using various other CpG enzymes clearly demonstrated that the *K*-region is uncommonly rich in CpG sequences (Table II). For instance, the frequencies of *Sac*II sites or *Bss*HII sites in the *K*-region are 20- to 30-times higher than the average calculated number (Lindsay and Bird, 1987). This was especially true within the segment located betweem map positions −35 to −80, which contained 21 of the total of 38 *Sma*I sites found in the entire 170 kb of the *K*-region examined. This implied a high G+C content and an absence of CpG suppression in the *K*-region DNA.

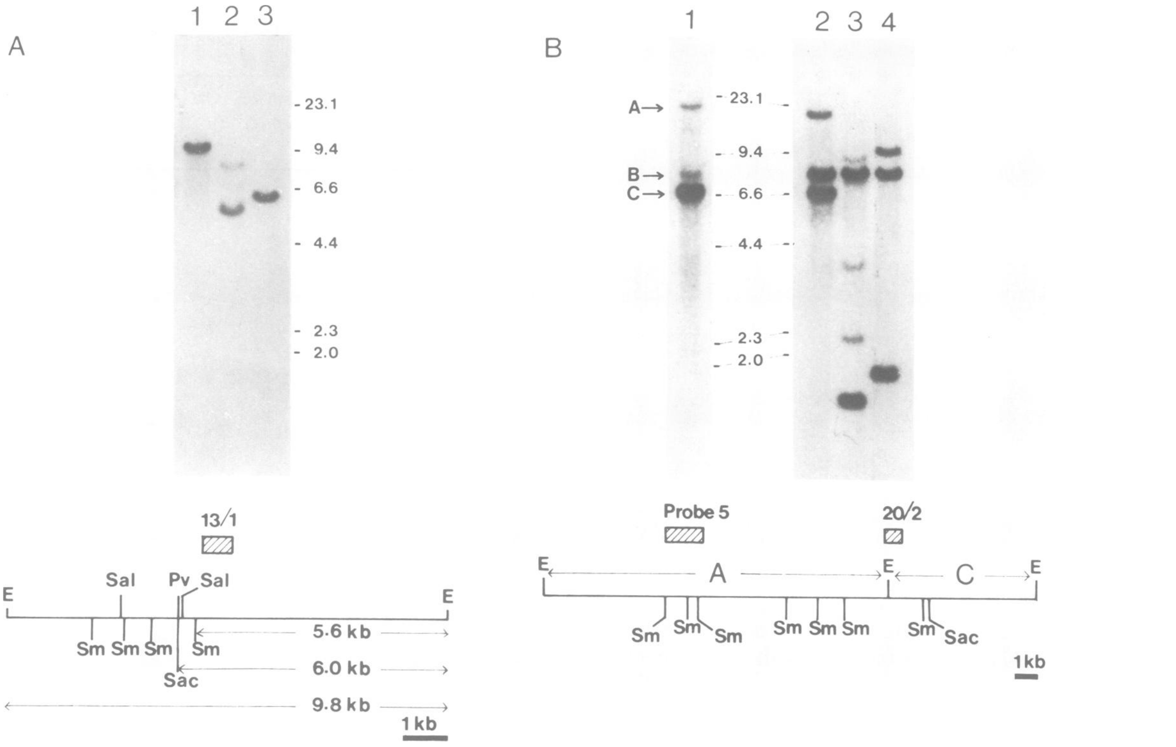

Fig. 3. Autoradiographs from hybridizations of the probes, 13/1(**A**), and 20/2 or probe 6(**B**), to EC cell DNA digested with *Eco*RI or doubly digested with *Eco*RI and C-G methylation sensitive enzymes. (A) lane 1, *Eco*RI; lane 2, *Eco*RI + *Sma*I; lane 3, *Eco*RI + *Sac*II. The restriction map of 9.8 kb *Eco*RI fragment with the enzymes, *Sma*I(Sm), *Sac*II(Sac), *Sal*I(Sal), *Pvu*I(Pv) and *Eco*RI(E), is shown below the autoradiograph. Hatched box indicates an approximate position of 13/1. (B) lane 1, *Eco*RI-digested DNA probed with the genomic fragment, probe 6; lanes 2, 3, and 4, EC cell DNA digested with *Eco*RI(2), *Eco*RI + *Sma*I(3), *Eco*RI + *Sac*II(4) was hybridized with cDNA probe, 20/2. A, 15.4 kb *Eco*RI fragment; C, 6.5 kb *Eco*RI fragment.

Non-methylated CpG-rich islands mark most of the transcribed sequences in the H-2K region

It has been shown that regions rich in non-methylated CpG (CpG-rich islands) are often associated with the 5′ portion of vertebrate genes (Bird, 1986). The fact that the *K*-region contains a number of transcribed sequences and is rich in CpG suggested the possibility of CpG-rich islands in this region.

To study this possibility further, we performed Southern blot analyses with methylation-sensitive restriction enzymes. DNA was prepared from EC cells in which most of the conserved sequences are expressed. It was digested with *Eco*RI or doubly digested with *Eco*RI and either *Sma*I or *Sac*II which are both sensitive to methylation at CpG. When blots of *Eco*RI digests were probed with cDNA clone 13/1 which was isolated by probe 3 (see below), the expected 9.8 kb *Eco*RI fragment was detected (Figure 3A). However, this fragment disappeared in doubly digested DNA, clearly showing the presence of sites cleavable by *Sma*I and *Sac*II within the fragment. A restriction map of the 9.8 kb *Eco*RI fragment (bottom, Figure 3A) shows a cluster of four *Sma*I sites. Since 13/1 hybridized strongly to the 5.6 kb band, the *Sma*I site 5.6 kb to the left of the right-end *Eco*RI site is non-methylated. In *Eco*RI−*Sac*II digests, 13/1 hybridized to a single intense 6.0 kb band. This places the cleavable *Sac*II site very close to the non-methylated *Sma*I site. The sites for *Sal*I and *Pvu*I were also found to be close to the cDNA clone (Figure 3A). Thus, this putative CpG-rich island was shown to be gene-associated. The same approach was used to identify four more CpG islands, and all of them were associated with the transcribed sequences found in the *K*-region (Figure 1 and 3B). In each island, sites for at least three different rare-cutting enzymes were clustered, and found to be non-methylated in EC cell DNA. In addition to these newly found CpG-rich islands, it has been reported that the *H-2K* class I gene is associated with CpG islands (Tykocinski and Max, 1984). Thus almost all the genes in the *K*-region are marked by nonmethylated CpG-rich islands; see Figure 1 for their location.

The important findings from the experiments described above are (i) almost all the *H-2K* region sequences that are conserved in mammals evolutionarily distant from rodents are either transcriptionally active or physically very close to active sequences, (ii) most of the transcribed sequences are expressed in the RNA of early embryos, (iii) most of the transcribed sequences are associated with CpG-rich islands, and (iv) the *K*-region DNA is unusually rich in CpG sequences.

Characterization of transcribed sequences: isolation of cDNA clones from 5.5 and 10.5 day mouse embryo cDNA libraries

An efficient method for screening. To characterize further the sequences transcribed from the *H-2K* region, we constructed λgt10 cDNA libraries from 5.5 and 10.5 day mouse embryos. The libraries were screened with eight probes already known to be expressed in embryo RNA, and with additional unique copy probes which fill the gap between the transcribed sequences. Since repetitive screening of the library with so many probes would have been a

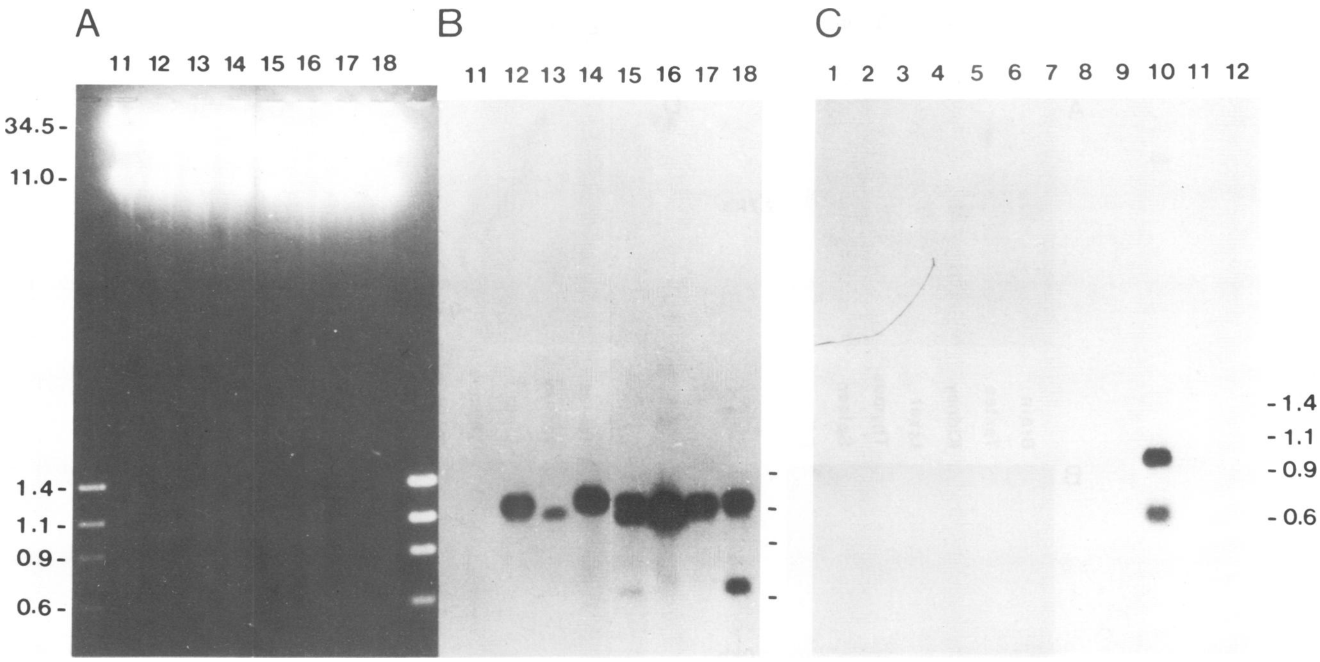

Fig. 4. Hybridization results from probe 11 to Southern blot of inserts DNA prepared from mouse embryo cDNA libraries. (**A**) Ethidium bromide picture of the gel used for making the Southern blot. 5 μg of the *Eco*RI-digested phage DNA derived from 10.5 day library was loaded per slot onto a 1% Tris−phosphate gel. Two heavy bands are vector arms, and a smear of inserts DNA is seen. Each lane corresponds to different screening plate (see text, and Materials and methods). (**B**) Autoradiograph of the Southern blot hybridized with probe 11. The blot was hybridized with probe 11, and washed in high stringency (final wash in 0.1 × SSC, 0.1% SDS at 65°C for 1 h). Exposure was for 15 h at −70°C with an intensifying screen. Positive hybridization to all the lanes except for lane 11 indicates that there are cDNA inserts homologous to probe 11 in most of the lanes. According to this information, a cDNA clone, 14/2, was isolated from the screening plate no. 14. (**C**) Hybridization of 14/2 cDNA to the blot of insert DNA from 5.5 day embryo library. Positive hybridization to lane 10 demonstrates that the gene represented by clone 14/2 is also expressed in 5.5 day mouse embryos.

laborious task, we employed a rapid method for library screening devised by K.Abe and Eric Lader of this laboratory. Essentially this entails dividing the whole library into 'books' of recombinant phages (each 'book' being one plate); a pool of cDNA inserts is prepared from each book by restriction enzyme digestion and then used for a single lane of a 'catalogue' Southern blot for screening with any given probe. This method not only greatly simplifies screening procedures, but also permits the immediate identification of the phage clone containing the longest cDNA insert.

Figure 4 shows an example of this procedure. In each lane of cloned cDNA in an ethidium-stained catalogue gel derived from the 10.5 day library, the two arms of λgt10 and a smear of DNA comprising about 25 000 different inserts are seen (Figure 4A). Probe 11 detected bands with similar intensity but usually different sizes in all lanes except for lane 11, indicating that all the screening plates except no. 11 contain positive clones (Figure 4B). Screening of replica filters from plate no. 14 yielded three positive clones with identical 1.0 kb inserts as predicted by the catalogue blot. (The identical clones presumably resulted from amplification of the library). One of these positive clones was subsequently used as a hybridization probe on another catalogue blot from a 5.5 day embryo cDNA library. The positive signals in lane 10 of Figure 4 (C) show that the gene detected by probe 11 is also expressed by 5.5 day embryos.

Based on the Northern blot data using genomic and cDNA probes, we assigned five novel genes along the *H-2K* region, and called them KE 1−5 (to denote *K*-region expressed). Details of these genes are separately described below.

KE 1 (transcript size, 2.7 kb). Probe 3 was used to isolate a partial cDNA clone with a 0.6 kb insert from the 10.5 day library. 13/1 was subsequently shown to hybridize on Northern blots to the same single 2.7 kb transcript detected by probe 3. As judged by Northern and Southern blot analyses, 13/1 represents a single copy gene that maps to the *H-2K* region. Since probe 2, about 12 kb distal to probe 4, detected a faint 2.7 kb transcript this gene seems to cover at least 12 kb of genomic DNA (map position from −75 to −87). We have called this gene KE 1. Clone 13/1 had its highest level of expression in testis, as shown in Figure 5A. Interestingly, its expression was also high in EC cells and in brain; it is not uncommon for genes with high levels of expression in testis also to have abundant transcripts in brain or EC cells (Stacey and Evans, 1984; Shackleford and Varmus, 1987). A low level of expression occurred in other adult somatic tissues such as kidney, liver, thymus and spleen. Quantitative slot blot analysis of RNA showed that the level of expression in testis was 3- to 4-fold greater than that in EC cells and brain, and at least eight times more than in other somatic tissues (data not shown).

KE 2 (transcript size, 1.0 kb). Probe 6 was used to isolate a 0.6 kb cDNA clone, 20/2, from the 10.5 day library, and this clone was in turn used to isolate cDNA clones from the 5.5 day library. 20/2 detected a prominent 1.0 kb transcript, and a fainter 2.9 kb band which was barely detectable when genomic probe 6 was used. The neighboring probe 5 hybridized more strongly to the 2.9 kb RNA, but not at all to the 1.0 kb transcript. Since the cDNA clone detected both 2.9 and 1.0 kb RNA, and did not cross-hybridize to probe

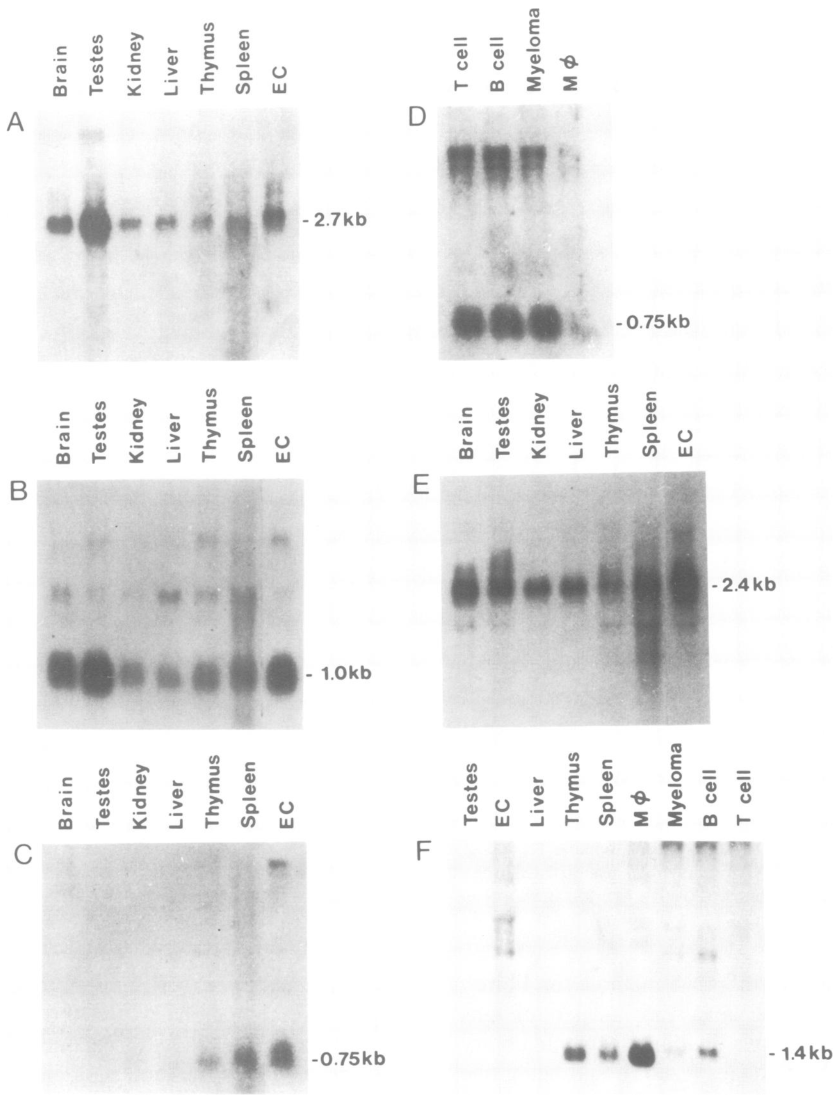

Fig. 5. Tissue distribution of KE gene transcripts. cDNA or genomic probes were used to screen RNA from various sources. Sources of RNA are shown at the top of each figure. (**A**) KE 1; cDNA, 13/1. (**B**) KE 2; cDNA, 20/2. (**C**) and (**D**) KE 3; probe 8. Signal at high mol. wt RNA may be a result of cross-hybridization of probe 8 to 18S rRNA. (**E**) KE 4; cDNA, 14/2. (**F**) KE 5; probe 14.

5, it is not likely that the two transcripts represent different genes, suggesting that the 2.9 kb RNA is either unprocessed or results from alternate splicing. The conservation of probe 5 sequences especially in the human genome lends support to the latter possibility. Clone 20/2 appears to define another gene, designated KE 2, which occupies at least 10 kb (position −54 to −64). Both clone 20/2 and probe 6 hybridize to 15.4 kb (A) and 6.5 kb (C) fragments on Southern blots of *Eco*RI digests of the cosmid from which probe 6 was obtained (data not shown). However, as shown in Figure 3B, on Southern blots of *Eco*RI digested genomic DNA, they both detect, in addition to the expected fragments A and C, an extra fragment of 7.0 kb (B). The same result was obtained using an independently isolated clone. This suggests the presence of a homologous sequence outside the *H-2K* region. Indirect evidence favors the notion that it is the *K*-region gene that is expressed. First, probe 6 hybridizes more strongly to the *K*-region fragment C than to the anonymous B fragment, and fragment C is always slightly more intense than the B fragment when probed with cDNA clones. This implies that sequence divergence exists between the two fragments, and also between the transcript and fragment B. Second, non-methylated CpG-rich islands, which often exist 5′ to expressed mammalian genes, were found within fragment C but not within fragment B (Figure 3B). The overall pattern of KE 2 expression was very similar to that of KE 1, although they have been proven to be different genes (Figure 5B).

KE 3 (transcript size, 0.75 kb). Probe 8 has been used to screen 3 × 10^5 plaques of the 10.5 day embryo library, but we have not yet succeeded in retrieving a corresponding cDNA. Nevertheless, it is likely that probe 8 represents parts of a transcriptional unit, for several reasons: it exhibits hybridization with DNA from all mammalian species examined (Figure 2B); it appears to represent a unique copy sequence when examined by hybridization to Southern blots of mouse genomic DNA; it detects a 0.75 kb transcript with a specific tissue distribution; and it also contains CpG-rich islands. Probe 8 has, therefore, been tentatively assigned to a third, but still putative gene, KE 3. KE 3 appears to be most abundant in EC cells, and to be expressed to a lesser extent in the lymphoid tissues of spleen and thymus (Figure 5C). Trace amounts of the same size transcript are constitutively expressed in all other tissues, but can be detected only after prolonged exposure of Northern blots (data not shown). To examine further the expression of KE 3 in lymphoid tissues, we isolated RNA from a T cell line (BW5147), a B cell line (MAO-A 1A9), a myeloma (NS1), and normal peritoneal macrophages. Northern blots of these RNAs hybridized with a probe to KE 3 showed that the gene is expressed in both T and B cell lineages, but is absent in macrophage RNA (Figure 5D).

KE 4 (transcript size, 2.4 kb). Probes 11 and 12 both identify a 1.0 kb cDNA clone, 14/2, that detects a 2.4 kb transcript which is the same as detected by the two probes on Northern blots of embryo RNA. As demonstrated on Southern blots of genomic digests, it apparently represents a single copy gene that we have designated KE 4. As shown in Figure 4C, KE 4 is expressed in 5.5 day mouse embryos. KE 4 was found to be constitutively expressed in all tissues examined, although expression was reproducibly two to three times higher in EC cells compared to adult tissues (Figure 5E).

KE 5 (transcript size, 1.4 kb). Probes 14 and 15 were both negative on Northern blots of embryonic RNA and, as expected, did not yield any positive signals when they were used to screen the 10.5 day embryo cDNA library. Since they represented evolutionarily conserved sequences, however, we were prompted to use them to screen Northern blots from various other tissues. These analyses showed that probe 14 detects a 1.4 kb transcript that is expressed only in lymphoid tissues and in a cultured embryonic fibroblast cell line, STO. This putative gene is tentatively designated KE 5. More detailed analysis by Northern blotting demonstrated that KE 5 expression is highest in macrophages although a low level of expression is seen in the B-cell line that we examined (Figure 5F).

Discussion

Distribution of genes in the mammalian genome: is the MHC a gene-rich region?

In the work presented here, we searched for unidentified genes in a 170 kb stretch of DNA, and showed for the first time five novel genes in the *K*-region of the mouse MHC. Similar kinds of studies have recently been done for the human Duchenne Muscular Dystrophy locus (Monaco *et al.*, 1986) and the region of the Y chromosome containing the testis determining factor (Page *et al.*, 1987). In both cases coding sequences were sought in about the same extent of cloned DNA as the *K*-region, but only one gene or a part of one gene was found. This is in sharp contrast to our findings. On statistical grounds, only 2% of the mammalian genome is thought to contain transcribed genes, and their average spacing is estimated to be 35 kb (Ohno, 1986). However, the studies described above and ours suggest that this figure is not at all representative of the entire genome, and that mammalian genes are very unevenly distributed. Bernardi *et al.* (1985) showed that mammalian genomes can be classified into different fractions based on G+C content, and that long (>200 kb) DNA segments with relatively homogeneous base composition can be identified. Their data suggested that genes are preferentially distributed in segments with high G+C content and relatively high CpG frequency, although such segments are relatively rare in the genome. Our present observations strongly support this view. The *K*-region is extraordinarily rich in CpG sequences, reflecting its high G+C content and the absence of CpG suppression. Also, the existence of seven genes in 150 kb represents a density comparable to that in other crowded areas of the *H-2* complex such as the Qa- and I-regions. Thus, it does appear that the mammalian genome is a mosaic of several classes of DNA segments, and that gene density varies in different segments according to their G+C content. In this context we should note that not only the *K*-region but a large part of the entire MHC may represent segments that have high G+C contents and are gene-rich. Not only is the frequency of CpG enzyme sites in the MHC higher than in other parts of the genome (Barlow and Lehrach, 1987; Muller *et al.*, 1987), but a number of clustered CpG enzyme sites (i.e. CpG-rich islands) have been observed within its detailed molecular map (Steinmetz *et al.*, 1986) that point to the existence of additional genes in an already heavily populated area.

Identification of coding sequences by Southern blot analysis

In a search for expressed genes in the *H-2K* region we examined the potential of two strategies: identifying unique/low copy sequences that were homologous across species, and defining the positions of methylation-free CpG islands. Both methods proved valuable in locating novel genes in that region. The finding of interspecific sequence homology was expected to be useful because in general the segments which comprise entire gene sequences evolve differently. Protein-coding and regulatory regions are conserved in evolution because of their functional constraints, whereas introns and non-regulatory regions diverge rapidly (Hayashida and Miyata, 1983). Thus, sequence divergence in introns and intergenic sequences between different species will increase in proportion to the time since the two species split apart from a common ancestor. For example, extensive sequence homology (~90%) is observed in both the coding and non-coding regions of immunoglobulin J_κ genes from rat and mouse, which diverged from each other ~17 million years ago (Sheppard and Gutman, 1982). On the other hand, a comparison of the dihydrofolate reductase genes of mouse and man, two species separated by 75 million years (Hayashida and Miyata, 1983), gave quite different figures; the protein coding and the immediate 5′ upstream non-coding region showed 89 and 65% homology, respectively, whereas the introns and the 3′-untranslated region had no significant homology (Yang *et al.*, 1984). Therefore, sequences with no functional significance will diverge to the point where residual homology cannot be detected by Southern blot hybridization. The results obtained in our 'zoo blot' analyses show that mouse and mammals other than rodents are, in practice, distant enough to permit distinguishing coding and non-coding sequences by Southern hybridization. The zoo blot analysis is not only a simple procedure, but also offers some technical advantages over RNA hybridization techniques in which the detection of message largely depends on its abundance, and, of course, on its tissue distribution. In fact, in the preliminary screens in which we used embryo cDNA to hybridize to restriction enzyme digested *K*-region cosmids, we obtained clear-cut results indicating the presence of a transcribed gene only for the relatively abundantly expressed KE 4 gene. However, the zoo blot approach pointed correctly to the presence of other transcribed genes. For example, the evolutionary conservation of probe 14 led us to the identification of the lymphoid-specific KE 5 sequence which was missed in the original screening for embryo-expressed genes.

The hypothesis that a high proportion of CpG islands are gene-associated has been proposed by Bird and colleagues (Bird, 1986; Lavia *et al.*, 1987), and recently supported by others (Rappold *et al.*, 1987). Our data showing that all of the KE genes examined are associated with apparent islands further sustains this hypothesis. CpG islands were originally thought to occur primarily in the 5′ end of 'house-keeping' genes, but our results as well as those of Rappold *et al.* (1987) demonstrate that genes with specific tissue-dependent expression can also be marked by such islands, thus enhancing the usefulness of CpG islands in searching for a broad range of genes.

The division of cDNA libraries into 'books' greatly facilitated screening procedures necessary for obtaining cDNA sequences. Their simultaneous use as an alternative to Northern blots was also helpful in examining gene expression in very young mouse embryos, since the small quantities of RNA that can be obtained from such embryos has made it very difficult to perform conventional Northern blot analysis of embryo-expressed genes. We also successfully applied similar procedures to screen the cosmid library during chromosomal walking.

KE 5, a novel gene in the MHC expressed specifically in lymphoid tissues

At least two loci coding for molecules expressed in lymphoid tissues have been mapped to the *I−K* boundary region (Monaco and McDevitt, 1982, 1986; Hayes and Bach, 1980). Of these, the Low Molecular Weight Proteins (LMP) show a very similar tissue distribution to that of the KE 5 gene transcript. Both are extremely abundant in macrophages, and are found to a lesser extent in fibroblasts and B cells. In addition, the message for LMPs is similar in size to the transcript detected by KE 5 (Monaco and McDevitt, 1986). However, the tentative map position of the LMP locus, between *Aβ3* and *Aβ*, is not consistent with the location of KE 5, which maps between *K* and *Aβ3*. Nevertheless, since the LMP antigens consist of a number of subunits, of which only two have been definitively mapped, it is possible that KE 5 corresponds to one of the remaining subunits.

A cluster of embryo-expressed genes in the proximity of the H-2K gene

The data reported here demonstrate the existence of closely linked, embryo-expressed genes in the *H-2K* region. The high levels of expression of some of their corresponding RNAs in embryos or testes are not typical of other known *H-2* linked genes. We have no information about their biological function at the moment. However, the restricted tissue expression of most of them, coupled with their evolutionary conservation, suggests that they do not simply represent house-keeping types of genes, but rather implies that they may play some important role in embryogenesis and/or testicular function.

Very few embryonic functions have been mapped to the *H-2K* region. Teratocarcinoma transplantation (Gt) antigens that are also expressed on adult tissues have been proposed to map to both the K and D ends of the *H-2* complex (Johnson *et al.*, 1983). However, recent studies using *H-2* class I gene mutants showed that Gt antigens are more like class I antigens themselves (Demant and Oudshoorn-Snoek, 1985; Moser *et al.*, 1986), ruling out a possible relationship between Gt antigens and the newly found KE genes. Extensive recombination analysis of the t^{w5} lethal mutation suggests that the genetic distance between that mutation and *H-2K* is <0.08 cM (Artzt *et al.*, 1988) and thus that the physical distance is 0−250 kb. Since nothing is known about the t^{w5} gene product, candidate genes can be assessed on their genomic position relative to *H-2K*, and on their time of expression. On these grounds KE 1−4 all qualify as candidates since all are expressed in EC cells. In fact, two of them have been shown to be expressed in 5.5 day embryos, a time when t^{w5} is thought to act. The *H-2K* or *K2* genes could be considered as candidates on the supposition that some mutation in regulatory sequences could lead to inappropriate expression in 5.5 day embryos with lethal consequences. We tested this possibility by collecting ~100 5.5 day embryos from litters where ~50% of the embryos were expected to be t^{w5}/t^{w5} or $t^{w5}/+$. About 2−3 μg of total RNA was obtained and used to prepare a single lane of an RNA blot. No hybridization signal was obtained after probing with two cDNA probes, pH2IIa and pH2III (Steinmetz *et al.*, 1981), which hybridize with all known class I sequences, thus suggesting that a regulatory mutation in *H-2K* does not represent the t^{w5} lethal mutation. We know the *H-2K* and *K2* genes are not expressed in normal embryos as 5.5 days, since the 5.5 day cDNA library was negative after screening with probes A and B (see Figure 1 for their location). We are currently undertaking further analysis of KE gene expression at different early stages of embryogenesis by a combination of Northern blotting and *in situ* hybridization. These experiments, coupled with DNA sequence analysis, should give information on the functional role of the highly conserved KE genes, and their relationship to one another and to the t^{w5} mutation.

Materials and methods

Cell lines

Cell lines used were as follows: pluripotent teratocarcinoma, PSA1, (Martin *et al.*, 1977); embryonic fibroblast cell, STO (Evans and Kaufmann, 1981); T-cell line, BW5147 (Altman *et al.*. 1982); Myeloma, NSI (Köhler and Milstein, 1976); B-cell line, MAO-A 1A9 (a gift of Dr P.Gottlieb, Department of Microbiology, University of Texas at Austin); embryonal carcinoma, JC44 (a gift of Dr G.Barry Pierce, Department of Pathology, University of Colorado). The conditions for cell culture were described in detail in the references cited above.

Probes

The probes used for chromosomal walking were probes 1 and 16 (see Figure 1) and *Aβ3*, an 0.8 kb *Eco*RI−*Bam*HI fragment (Widera and Flavell, 1985). All other probes described in this work (see Figure 1 and Table I) were derived from cosmids isolated from cosmid libraries of DNA from C3H·t^{w5}/t^{w12} compound mice or t^{w5}/t^{w5} EC cells (Uehara *et al.*, 1987).

Probe DNA was prepared by electro-elution from agarose gels or by melting agarose gel slices containing insert DNA. DNA was labeled either by nick translation (Rigby *et al.*, 1977) or by oligolabeling using random hexamer primer (Feinberg and Vogelstein, 1983).

cDNA libraries

Two libraries using λgt10 vector were constructed from mRNA prepared from 5.5 day egg cylinder stage embryos and 10.5 day embryos according to the method of Huynh *et al.* (1985). Embryos at egg cylinder stage (Stage 8 of Theiler) were obtained after a 4 day, *in vitro* culture of 1500 preimplantation embryos (Hsu, 1979). Embryos at 10.5 days of gestation were dissected free from decidual tissues. Details of the libraries will be described elsewhere.

cDNA library screening

Since we needed to screen the libraries with many different probes, we devised a simple and rapid method for library screening. Approximately 6×10^5 recombinant phages were plated out on 24 plates (25 000 plaques per 150 mm plate). After making duplicate plaque replicas, 10 ml of λ-dilution buffer was overlayed on each plate to prepare plate lysate, and the lysates were stored separately. The plates containing phages were kept at 4°C for picking positive clones as a considerable number of phages still remain in plaques after making the lysate. Subsequently, phage DNA was made from each plate lysate, a mixture of 25 000 different clones, according to the method of Davis *et al.* (1980). Phage DNA (5 μg) from each plate lysate was digested with *Eco*RI to release the inserts, separated on 1% agarose gels, and transferred. Each lane of the Southern blots thus made represents 25 000 different kinds of cDNA inserts derived from one plate. By hybridizing probes with Southern blots of cDNA inserts we could determine which filter contained a positive clone, and at the same time, the clone with the longest insert (see Figure 4). Knowing this information, we then screened several plaque replicas corresponding to the positive lanes, instead of screening all the filters. Use of several such Southern blots for different probes speeded up the whole screening procedure.

Southern blots

Restriction digested DNA separated on 0.7% Tris−phosphate agarose gels was transferred to Nitroplus 2000 membrane (MSI) in 20 × SSC. Hybridization was carried out in a mix containing 50% Formamide, 50 mM NaPO4, pH 6.5, 5 × SSC, 5 × Denhardts, 0.1% SDS, 1 mM EDTA, 10% Dextran sulfate and 100 μg/ml Salmon Sperm DNA at 42°C. Blots were washed twice in 2 × SSC, C.1% SDS for 30 min at room temperature with agitation, then in 0.1 × SSC, 0.1% SDS at 65° for 2 × 20 min. For 'Zoo blot' hybridization, the final wash was performed in either 0.1 × SSC, 0.1% SDS at 50°C or 1 × SSC, 0.1% SDS at 65°C.

Northern blots

Total RNA from embryos, cell lines and various adult organs was isolated by the guanidium thiocyanate method of Chirgwin *et al.* (1979). Poly(A)$^+$ RNA was selected by passage several times over an oligo(dT) column (Maniatis *et al.*, 1982).

15 μg of total RNA or 2−3 μg of poly(A)$^+$ RNA were fractionated on 1.1% Formaldehyde agarose gels and transferred to the MSI membrane. Hybridization and washing conditions were the same as for Southern blots. Addition of sonicated and denatured total mouse DNA into a prehybridization and hybridization mix at a concentration of 50 μg/ml helped to reduce the hybridization background when genomic fragments were used as probes.

Acknowledgements

We thank Drs P.Gottlieb and G.Barry Pierce for providing us with the cell lines. We also thank Taeko Abe and Sara Herlick for expert technical assistance throughout this study. The excellent assistance of Judy Camps in preparing the manuscript is greatly appreciated. This study was supported by NIH grants HD10668 and CA21651.

References

Altman,A., Sferruza,A., Weiner,R.G. and Katz,D.H. (1982) *J. Immunol.*, **128**, 1365−1371.

Artzt,K. (1984) *Cell*, **39**, 565−572.

Artzt,K., Abe,K., Uehara,H. and Bennett,D. (1988) *Immunogenetics*, **28**, 30−37.

Barlow,D.P. and Lehrach,H. (1986) *Trends Genet.*, **3**, 167−171.

Bernardi,G., Olofsson,B., Filipski,J., Zerial,M., Salinas,J., Cuny,G., Meunier-Rotival,M. and Rodier,F. (1985) *Science*, **228**, 953−958.

Bird,A.P. (1986) *Nature*, **321**, 209−213.

Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Ratler,W.J. (1979) *Biochemistry*, **18**, 5294−5299.

Davis,R.W., Botstein,D. and Roth,J.K. (1980) *Advanced Bacterial Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Demant,P. and Oudshoorn-Snoek,M. (1985) *Immunogenetics*, **22**, 543−552.

Evans,M.J. and Kaufmann,M.H. (1981) *Nature*, **292**, 154−156.

Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6−13.

Friend,S.H., Bernards,R., Rogolj,S., Weinberg,R.A., Rapaport,J.M., Albert,D.M. and Dryja,T.P. (1986) *Nature*, **323**, 643−646.

Hayashida,H. and Miyata,T. (1983) *Proc. Natl. Acad. Sci. USA.*, **80**, 2671−2675.

Hayes,C.E. and Bach,F.H. (1980) *J. Exp. Med.*, **151**, 481−485.

Hood,L., Steinmetz,M. and Malissen,B. (1983) *Annu. Rev. Immunol.*, **1**, 529−568.

Hsu,Y.C. (1979) *Dev. Biol.*, **68**, 453−461.

Huynh,T.V., Young,R.A. and Davis,R.W. (1985) In Glover,D.M. (ed.), *DNA Cloning−A Practical Approach.* IRL Press, Oxford, Vol. 1, pp. 49−78.

Johnson,L.L., Clipson,L.J., Dove,W.F., Feilbach,J., Maher,J. and Sheldolovsky,A. (1983) *Immunogenetics*, **18**, 137−145.

Köhler,G. and Milstein,C. (1976) *Eur. J. Immunol.*, **6**, 511−519.

Lavia,P., Macleod,D. and Bird,A. (1987) *EMBO J.*, **6**, 2773−2779.

Lindsay,S. and Bird,A.P. (1987) *Nature*, **327**, 336−338.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Martin,G.R., Wiley,L.M. and Damjanov,I. (1977) *Dev. Biol.*, **61**, 230−244.

Monaco,A.P., Neve,R.L., Colletti-Feener,C., Bertelson,C.J., Kurnit,D.M. and Kunkel,L.M. (1986) *Nature*, **316**, 842−845.

Monaco,J.J. and McDevitt,H.O. (1982) *Proc. Natl. Acad. Sci. USA.*, **79**, 3001−3005.

Monaco,J.J. and McDevitt,H.O. (1986) *Human Immunol.*, **15**, 416−426.

Moser,A.R., Shedlovsky,A. and Johnson,L.L. (1986) *Immunogenetics*, **23**, 271−273.

Muller,U., Stephan,D., Philippsen,P. and Steinmetz,M. (1987) *EMBO J.*, **6**, 369−373.

Ohno,S. (1986) *Trends Genet.*, **2**, 8.

Orkin,S.H. (1986) *Cell*, **47**, 845−850.

Page,D.C., Mosher,R., Simpson,E.M., Fisher,E.M.C., Mardon,G., Pollack,J., McGillivray,B., de la Chapelle,A. and Brown,L.G. (1987) *Cell*, **51**, 1091−1104.

Rappold,G.A., Stubbs,L., Labeit,S., Crkvenjakov,R.B. and Lehrach,H. (1987) *EMBO J.*, **6**, 1975−1980.

Rigby,P.W.I., Dieckmann,M., Rhodes,C. and Berg,P. (1977) *J. Mol. Biol.*, **113**, 237−251.

Shackleford,G.M. and Varmus,H.E. (1987) *Cell*, **50**, 89−95.

Sheppard,H.W. and Gutman,G.A. (1982) *Cell*, **29**, 121−127.

Shin,H.-S., Flaherty,L., Artzt,K., Bennett,D. and Ravetch,J. (1983) *Nature*, **306**, 380−383.

Shin,H.-S., Bennett,D. and Artzt,K. (1984) *Cell*, **39**, 573−578.

Stacey,A.J. and Evans,M.J. (1984) *EMBO J.*, **3**, 2279−2285.

Steinmetz,M., Frelinger,J.G., Fisher,D., Hunkapiller,T., Pereira,D., Weissman,S.M., Uehara,H., Nathenson,S. and Hood,L. (1981) *Cell*, **24**, 125−134.

Steinmetz,M., Stephan,D. and Fischer Lindahl,K. (1986) *Cell*, **44**, 895−904.

Tykocinski,M.L. and Max,E.E. (1984) *Nucleic Acid Res.*, **12**, 4385−4396.

Uehara,H., Abe,K., Park,C.-H., Shin,H.-S., Bennett,D. and Artzt,K. (1987) *EMBO J.*, **6**, 83−90.

Widera,G. and Flavell,R.A. (1985) *Proc. Natl. Acad. Sci. USA.*, **82**, 5500−5504.

Yang,J.K., Masters,J.N. and Attardi,G. (1984) *J. Mol. Biol.*, **176**, 169−187.

Received on June 10, 1988; revised on August 1, 1988