Studies on characteristic genome structures and on new genes found in the human MHC class III region near the junction with the class II

Kimihiko Sugaya

Doctor of Philosophy

Department of Genetics,
School of Life Science,
The Graduate University for Advanced Studies

1994

Contents

I. Summary

Genomes of higher vertebrates are composed of long-range mosaic structures of GC%, which are related to chromosome bands. Several groups, including the group to which I belong, showed chromosomal G bands to be mainly composed of AT-rich sequences and T bands (an evidently heat-stable subgroup of R bands), of GC-rich sequences: ordinary R bands are heterogeneous and appear to be intermediate. Gene density, DNA replication timing, repeat sequence density and other chromosome behaviors such as recombination are related to chromosomal bands and to the long-range GC% mosaic structures.

Human MHC spans about a 4 megabase (Mb) segment of the short arm of chromosome 6 (6p21.3), and the region is composed of classes I (about 2 Mb), III (1 Mb) and II (1 Mb) from telomere to centromere. Genes in classes I and II encode polymorphic antigens involved in genetic control of immune response: Class III, which is one of the regions most densely packed with genes in the human genome, encode proteins of diverse functions mostly unrelated to immune response. Susceptibility to a large number of diseases, including autoimmune disorders, is thought related to genes in the MHC. However, in many cases, it is not clear whether the susceptibility is due to known genes or to those not yet identified.

Previous studies of our group showed the human MHC to be a long-range mosaic of GC%. Contiguous classes I and III correspond to an evidently GC-rich domain and class II, to a domain with reduced GC%. Thus, borders of Mb-level GC% mosaic domains had been assigned within an under-characterized 450 kb

harboring the junction of classes II and III. To precisely locate the domain border and find new genes, chromosome walking to completely cover this 450-kb area were carried out by isolating cosmid, λ phage and YAC contigs. During characterization of the 450 kb, especially of the region near the border of the Mb-level GC% mosaic domains, three human MHC class III genes were found; the gene for receptor of advanced glycosylation end products of proteins (RAGE, a member of immunoglobulin superfamily molecules believed to be related to diabetes complication), PBX2 homeobox gene (designated HOX12 by me and suspected to be a proto-oncogene), and human counterpart of mouse mammary tumor gene *int-3*, designated as NOTCH3. Human RAGE and PBX2 sequences were previously determined sequencing their cDNA clones, but the gene structures and map locations had not been known. The contiguous RAGE and PBX2 (HOX12) genes were completely sequenced in this work, and a single copy number of these genes in the human genome was shown by Southern blot analysis.

Integration of the mouse mammary tumor virus (MMTV) into the *int-3* locus promotes the transcription of flanking mouse cellular *int-3* sequence that shares significant homology with the intracellular domain of *Drosophila* neurogenic Notch gene. The human sequences found in the present work contained not only the intracellular domain part present in the *int-3* sequence but also the extracellular part present in typical Notch-family genes, showing the sequence found in this study to correspond to the human counterpart of an uninterrupted form of the transmembrane protein gene predicted for the mouse *int-3* locus. The placental cDNA clones of the human NOTCH3 were isolated and sequenced. By constructing phylogenetic tree based on their sequences, four subfamilies for mammalian Notch genes were found.

Near a GC% transition of the long-range mosaic structures, being centromeric of NOTCH3, there were a 20 kb of dense *Alu* cluster and a 30 kb of dense LINE-1 cluster, as well as a pseudoautosomal boundary-like sequence (PABL) found by Fukagawa in our group. Summary of the organization of the walked 450 kb is as follows; [Class II; AT-rich side] HLA-DRA - 140 kb - PABL - 30 kb of LINE-1 cluster - 20 kb of *Alu* cluster - NOTCH3 - PBX2 - RAGE - 90 kb - TNX (tenascin X gene) - 70 kb - CYP21 [Class III; GC-rich side].

I have also found the gross similarity of genes on 6p21.3 and those on 9q33-q34. The human gene most closely related to NOTCH3 is TAN1 being precisely mapped on 9q34.3, that to PBX2 is PBX3 roughly mapped on 9q33-34, and that to TNX is HXB (tenascin C gene) on 9q32-q34. By searching human Genome Data Base (GDB), not only the three genes discovered by our group but also several others on 6p21.3 were found to have counterparts mostly mapped on 9q33-q34. This gross similarity should have been brought on by duplication of a wide range of the genome and thus gives a realistic knowledge concerning evolutionary processes to built up the present human genome. The similarity is also useful for finding undiscovered genes, especially candidate genes responsible for certain genetic diseases.

II. Introduction

Genomes of higher vertebrates are composed of long-range mosaic structures of G+C content, which are related to chromosome bands (Bernardi *et al.*, 1985; Ikemura, 1985; Aota and Ikemura, 1986; Ikemura and Aota, 1988; Bernardi, 1989). Several groups, including the group to which I belong, showed Giemsa-dark G bands to be mainly composed of AT-rich sequences and T bands (an evidently heat-stable subgroup of R bands), of GC-rich sequences: ordinary R bands are heterogeneous and appear to be intermediate (Ikemura and Aota, 1988; Bernardi, 1989; Ikemura *et al.*, 1990; Ikemura and Wada, 1991; Bernardi, 1993; Saccone *et al.*, 1993). Gene density, CpG island density, codon usage, chromosome condensation, DNA replication timing, repeat sequence density, and other chromosome behavior such as recombination and mutation rate are related to chromosomal bands and to long-range GC% mosaic structures (Bernardi *et al.*, 1985; Ikemura, 1985; Bird, 1987; Holmquist, 1987; Korenberg and Rykowski, 1988; Bernardi, 1989; Wolfe *et al.*, 1989; Gardiner *et al.*, 1990; Ikemura and Wada, 1991; Bettecken *et al.*, 1992; Pilia *et al.*, 1993; Craig and Bickmore, 1994). Concerning gene density, T-type R bands are known to be most densely packed with genes (Craig and Bickmore, 1993).

Previous studies of the group to which I belong (Ikemura *et al.*, 1988, 1990) showed the human major histocompatibility complex (MHC) to be a long-range mosaic of GC%. Human MHC spans about a 4 megabase (Mb) segment of the short arm of chromosome 6 (on a band 6p21.3), and the region is composed of classes I (about 2 Mb), III (1 Mb) and II (1 Mb) from telomere to centromere. Genes in classes I and II encode polymorphic antigens involved in genetic control of immune response, and those in class III encode proteins of diverse functions mostly unrelated to immune response. Contiguous class I and III (3 Mb) regions correspond to an evidently GC-rich domain and class II (1 Mb), to a domain with reduced GC%. Thus, borders of Mb-level GC% mosaic domains were assigned within under-characterized 450 kb harboring the junction of class II and III. To characterize the genome structures of the 450-kb region, cosmid walking from a centromeric class III gene CYP21 to class II and YAC walking from the most telomeric class II gene HLA-DRA to class III were carried out (Sugaya *et al.*, 1994; Matsumoto *et al.*, 1992), and λ phage walking from HLA-DRA to class III was done by Fukagawa *et al.* (1995). It should be noted that the band 6p21.3 on which the human MHC region lies has been assigned to a T-type R band above-noted to be densely packed with genes. At a high resolution level, a narrow G band (6p21.32) is thought to exist within the MHC (Senger *et al.*, 1993).

Susceptibility to a large number of diseases, including autoimmune disorders, is thought related to genes in the MHC (Klein, 1986). In many cases, it is not clear whether the susceptibility is due to known gene per se or to genes as yet unidentified. Genes in the MHC have been intensively studied, and many new genes have been found (see a review, Campbell and Trowsdale, 1993). The research group to which I belong (Matsumoto *et al.*, 1992a, 1992b) and another group (Bristow *et al.*, 1993) found an extracellular matrix protein tenascin-like gene, TNX, in the centromeric region to CYP21: Presence of the mouse counterpart in the respective position and its expression were recently confirmed (Matsumoto *et al.*, 1994). In the present work, the GC% boundary in the MHC was located about 180 kb centromeric to TNX, and the region about 90 - 180 kb centromeric to TNX was characterized disclosing three new genes and several characteristic genome structures (Sugaya, *et al.*, 1994; Fukagawa, Sugaya, *et al.*, 1995).

III. Materials and Methods

1. Cloning and sequencing

A cosmid library, which was used in this study, was constructed by Dr H. Inoko and his colleagues (Tokai Univ.), using the cosmid vector pWE15 (Stratagene, San Diego, CA) from the total human DNA of an HLA homozygous B cell line, AKIBA (HLA-A24, Bw52, DR2, Dw12, DQw1, Cp63) (Inoko *et al.*, 1985). This was done so as to reduce the chance of heterozygosity among different cosmids for shotgun sequencing as much as possible. Successive cosmid walking was done by standard procedures using a *Not*I-generated insert fragment of an already obtained cosmid, as its probe, which were labeled with [α-^{32}P] dCTP by the random-priming. Hybridization was performed after 1h of preannealing with human placental DNA (80-330 μg / ml of hybridization buffer), as described by Sealey *et al.* (1985). The isolated cosmid DNAs were subjected to *Eco*RI mapping.

Six human cDNA libraries cloned on λ gt10 or λ gt11 vector were obtained from Clontech, Palo Alto, CA.; those were placenta 5'-stretch plus (λ gt11), placenta (λ gt11), monocyte (λ gt11), B-cell (λ gt10), spleen (λ gt10) and skin fibroblast (λ gt10) cDNA libraries. Replica filters from the libraries (1-2 × 10^5 pfu / 20 cm × 20 cm plate) were hybridized with ^{32}P-labeled probes (> 5 × 10^8 cpm / μg) of genomic fragments. Positive plaques were picked and purified by two additional rounds of plating and probing.

*Eco*RI, *Hin*dIII, *Bam*HI or *Pst*I fragments of cosmid or λ cloned insert, as well as smaller fragments produced by successive *Sau*3AI digestion, were subcloned into pUC118 and sequenced by the *Taq* cycle sequencing, using the method of fluorescence-labeled DyeDeoxy terminators for ABI 373A automated sequencer (Applied Biosystems, Foster City, CA). When the sequence was not certain, it was determined by the standard dideoxy chain termination method using 7-DEAZA Sequencing Kit Ver. 2 or *Bca*BEST Dideoxy Sequencing Kit (TAKARA SHUZO CO., Kyoto, Japan). The obtained sequences were aligned into contigs, and to bridge them, primers for sequencing were prepared.

2. Database search and data-submission

Database search of GenBank, EMBL and PIR was made using FASTA and BLAST programs (Pearson and Lipman, 1988; Altschul *et al.*, 1990). Alignments and percentages of sequence identity were determined by ALIGN program of the DDBJ and Human Genome Center, Japan. The phylogenetic tree was constructed by using the program package CLUSTAL V (Higgins, 1994). Transcription factor binding sites were sought using the TFD database (Ghosh, 1991). The nucleotide sequence for the contiguous HOX12 and RAGE genes, 10108 nt, was deposited in DDBJ / GenBank / EMBL under accession number D28769. Map locations of RAGE, HOX12 and NOTCH3 were deposited in Genome Data Base (GDB) under GDB Id G00-306-354, G00-306-356 and G00-306-600, respectively, and their linkage was under GDB Id G00-306-676. The GDB committee has recently used the gene symbol "AGER" for the gene RAGE, according to the nomenclature rules.

3. Southern blot analysis and chromosome in situ *hybridization*

A YAC library was prepared by Drs T. Imai, T. Eki, K. Yokoyama and E.

Soeda of The Institute of Physical and Chemical Research (RIKEN, Tsukuba), and the clones harboring HLA-DRA were screened by Ando *et al.* (1994). High-molecular-weight DNA was isolated from YAC-containing yeast strains grown to saturation in uracil- and tryptophan-deficient liquid media (Brownstein *et al.*, 1989). YAC-containing yeast DNA and high-molecular-weight DNA from human placenta (Clontech, Palo Alto, CA) were digested by restriction enzymes to completion and size fractionated by electrophoresis on a 1% agarose gel. Southern blot and hybridization analysis were carried out according to standard methods (Maniatis *et al.*, 1989), and were conducted in the presence of 0.5 mg/ml human DNA for suppression procedure. Stringent washing was performed at 65℃ for 15min in 0.1×SSPE (1×SSPE is 0.18 M NaCl, 10mM NaH_2PO_4, 1mM EDTA, pH7.7) containing 0.1%SDS. The membranes were subjected to autoradiography with Fuji Bio-Imaging Analyzer BAS 2000 (Fuji Photo Film Co., Japan).

Chromosomal location of the cloned cosmid DNA was assigned by Drs K. Mita and E. Takahashi (National Institute of Radiological Sciences, Chiba) using direct R-banding fluorescence *in situ* hybridization (FISH) as described previously (Takahashi *et al.*, 1990). Direct R-banding FISH was based on FISH combined with replication R-bands. Suppression procedure with human Cot-1 DNA (about 30 times excess amounts) was according to Lichter *et al.* (1990), and procedures of labeling, hybridization, rinsing and detection were done in a routine manner (Takahashi *et al.*, 1991).

4. GC% measurement

Insert DNAs (30 - 40 kb) derived from pWE15 cosmid clones were separated from the vector DNA and RNA by high-performance liquid chromatography (HPLC): cosmid DNAs were digested by *Not*I at the pWE15 linker, put on a TSKgel DEAE-NPR HPLC column (0.46 × 3.5 cm; Tosoh Co., Tokyo), and eluted with a linear gradient of NaCl (from 0.5 M to 1 M) in 0.02 M Tris-HCl (pH 9.0). GC% of purified inserts was measured with DNA-GC kit (Yamasa Shoyu Co., Chiba, Japan) following the manufacturer's protocol. Briefly, 20 μg of DNA (EDTA free) being dissolved in 20 μl of distilled water was heated at 100° C for 5 min followed by rapid cooling in an ice bath, mixed with 20 μl of nuclease P1 solution (2 units/ml of 40 mM sodium acetate buffer containing 0.2 mM $ZnCl_2$, pH 5.3), and incubated at 50° C for 1 hour. P1 hydrolysate and a standard mononucleotides mixture supplied by the manufacturer were separately chromatographed on a YMC reversed-phase HPLC column (ODS-AQ-312, 0.6 x 15 cm; YMC Co., Kyoto) in 10 mM H_3PO_4 - 10 mM KH_2PO_4 (pH 3.5) at 26° C, and GC% was calculated according to the manufacturer's protocol.

5. Analysis of microsatellite alleles

Oligonucleotide primers corresponding to chromosomal regions flanking microsatellite repeats were synthesized. Genomic DNAs from 23 cell lines, listed in Table 3 and legend of Fig. 15 were used as a template in PCR amplification. PCR reactions were carried out on 200 ng of individual template DNAs with 20 pmol of a pair of primers according to the GeneAmp kit protocols (Perkin-Elmer Cetus). Amplification was performed in a Perkin-Elmer Cetus thermal cycler model 9600, using the following conditions; 30 cycles with 94° C (30 sec), 55° C (30 sec), 72° C (30 sec). The last polymerization step was extended to 7 min. After PCR, products were extracted by phenol/chloroform, digested by restriction enzymes if necessary, and analyzed by electrophoresis on a 8%(w/v) polyacrylamide gel.

IV. Results

1. Chromosome walking from MHC class III to class II and newly found genes

During the cosmid walking from the MHC class III CYP21 to class II, ten overlapping cosmids extending about 100 kb were previously obtained and TNX gene being immediately centromeric to CYP21 was found (Matsumoto *et al.*, 1992a, 1992b). In the successive walking to class II, sixteen cosmids extending about 180 kb from the TNX gene were isolated (Sugaya *et al.*, 1994). Figure 1 shows positions of the sixteen class III clones (KS-series cloned by me) along with those of previous ten clones (M-series cloned by Dr Matsumoto), as well as clones harboring class II HLA-DRA (KS-series cloned by me). To confirm proper walking and find new genes, the termini of individual class III cosmids were sequenced by using the T3 or T7 promoter primer designed for the cosmid pWE15 vector. Comparison of terminal sequences (about 300 nt) with GenBank data indicated evident similarities with various portions of the following genes. Those of cosmids KS11, KS72 and KS61 were essentially identical with separate portions of human RAGE cDNA reported by Neeper *et al.* (1992); RAGE is a receptor for advanced glycosylation end products of proteins. KS122 and KS120 were essentially identical with separate portions of the human PBX2 homeobox gene reported by Monica *et al.* (1991), while KS71, KS73, KS104, KS123, KS130 and KS132 were homologous with various portions of Notch-homologs of a wide range of species; *Drosophila* (Wharton *et al.*, 1985), *Xenopus* (Coffman *et al.*, 1990), zebrafish (Bierkamp *et al.*, 1993), human (Ellisen *et al.*, 1991; Stifani *et al.*, 1992; Larsson *et al.*, 1994), rat (Weinmaster *et al.*, 1991,1992) and mouse (Franco del Amo *et al.*, 1993; Lardelli *et al.*, 1993, 1994; Robbins *et al.*, 1992). These three genes were confirmed to be present in the respective regions. I will give a full detail of these genes in the following section.

A) Gene for RAGE (AGER)

Advanced glycosylation end products of proteins (AGEs) are nonenzymatically glycosylated proteins that accumulate in vascular tissue in aging and at an accelerated rate in diabetes. A 35-kDa receptor molecule for AGEs (RAGE) has been found on the surface of endothelial cells where it mediates binding (Schmidt *et al.*, 1992). Human and bovine RAGE gene sequences were identified as cDNAs (Neeper *et al.*, 1992; GenBank M91211 and M91212), but their genomic sequences and map positions have not been determined. RAGE is a member of the immunoglobulin superfamily of cell surface molecules and shares significant homology with MUC18, N-CAM and the cytoplasmic domain of CD20 (Neeper *et al.*, 1992). To determine whether sequences in the cosmids actually correspond to that of the RAGE gene, eight DNA primers of 20-mers for sequencing were prepared based on the reported cDNA sequence and sequencing was conducted using cosmid KS71 DNA as the template. All primers gave the exactly predicted sequences for RAGE exons as well as possible introns, indicating presence of the RAGE gene. To determine the total genomic sequence of the RAGE gene, fragments of cosmid KS71 after digestion by five restriction enzymes were subcloned into pUC118 and sequenced as described in Materials and Methods. The contiguous 10108-nt sequence obtained was deposited in DDBJ / GenBank / EMBL under the accession number D28769 (Sugaya *et al.*, 1994). Pairwise alignment of this genomic sequence with that of reported cDNA showed the entire cDNA sequence to be present in the genomic sequence, and the gene to be composed of 11 exons (Fig. 2A). All intron/exon junctions showed agreement with GT/AG rule. Table 1 shows 5' and 3' splice donor and acceptor sequences and nucleotide length of exons and introns. The reported human cDNA clone has a truncated form, and initiator methionine ATG is absent from the cDNA sequence. In this genomic sequence, ATG codon is present in position exactly corresponding to the initiator ATG of bovine RAGE cDNA (Neeper *et al.*, 1992). The following minor differences were observed between human cDNA and genomic sequences. The first two G bases of cDNA, absent in the genomic sequence, appeared to be generated during the cDNA cloning since the genomic sequence in the respective position was identical with bovine RAGE cDNA. In the protein coding region, a consecutive base difference producing one amino acid change (Arg <-> Gln at amino acid position 100) was present. In the 3' untranslated region (UTR) of 175 nt, there were no differences.

B) Gene for PBX2 (HOX12)

In the 10108-nt genomic sequence, the entire PBX2 cDNA sequence (Monica *et al.*, 1991; GenBank X59842) was found. Distance between initiator ATG of RAGE and 3' UTR of PBX2 was approximately 500 nt, thus indicating the inter-gene sequence to be significantly shorter than those of usual human genes. PBX2 and PBX3 genes were isolated as cDNAs based on extensive homology to PBX1, a human homeobox gene involved in t(l;19) translocation in acute pre-B-cell leukemias (Kamps *et al.*, 1990; Nourse *et al.*, 1990). Pairwise alignment of the present genomic sequence with that of the reported PBX2 cDNA showed the gene to be composed of 9 exons (Fig. 2A) and all intron/exon junctions to match GT/AG rule (Table 2). The first 25 nt of the reported 5' UTR corresponded to the vector sequence for cloning. After removing the vector sequence, there were two differences between the reported 5' UTR (271 nt) and the genomic sequence (99.3% identity). Genomic sequence extended 706 nt upstream the reported 5' UTR. In 1636 nt of the 3' UTR, there were only 7 base differences (99.6% identity). In the protein-coding region, there was one base difference resulting in an amino acid change (Ile <-> Met at amino acid position 393).

Figure 2B-D shows an evident CpG island in the 5' region of PBX2 sequence. Examination of TFD database (Ghosh, 1991) indicated a number of potential binding sites for transcriptional regulatory factors in the 5' region harboring the CpG island, ubiquitous factors (e.g., AP-1, AP-2, SP1) and specific factors (γ -IRE, GMCSF, BHLH, NF κ B) (Fig. 3A). Figure 3B shows potential binding sites detected in the inter-gene portion between PBX2 and RAGE sequences; because of very close locations of two genes, 3' UTR of PBX2 is also listed though the sequence is written in italics.

Chromosome *in situ* hybridization by Monica *et al.* (1991), using ^{3}H-labeled cDNA probes, showed PBX genes not to be clustered: PBX2, 3q22-23; PBX3, 9q33-34; PBX1, lq23 (Kamps *et al.*, 1990; Nourse *et al.*, 1990). The PBX2 sequence in my work was on 6p21.3 in contrast to 3q22-23 assigned by their hybridization using the radio-labeled probe. To clarify the reason for this discrepancy and examine whether my data is specific to AKIBA cells used for the present cosmid library, as well as to check for proper chromosome walking, two experiments were conducted: YAC walking from class II to class III using YAC library from a B-cell line CGM1 was done and *in situ* chromosome hybridization using fluorescence probe was carried out.

2. *YAC clones carrying class II HLA-DRA and Southern blot analysis*

Yeast Artificial Chromosomes (YACs) which harbor the most telomeric class II gene HLA-DRA, were screened by Ando *et al.* (1994), using pairs of PCR primers designed based on HLA-DRA sequence reported by Schamboeck *et al.* (1983). YAC clones carrying HLA-DRA but not HLA-DRB9 were further confirmed by using the respective radio-labeled DNA probes; HLA-DRB9 is a pseudogene of HLA-DRB and about 15 kb centromeric of HLA-DRA. The clones isolated should thus correspond to those extending from HLA-DRA to class III and possibly those bridging the gap not yet covered by the cosmid walking from class III. I analyzed two YAC clones, YDR2 (carrying about 450 kb insert) and YDR3 (about 220 kb). Based on the gene organization summarized by Campbell and Trowsdale (1993), PBX2 and RAGE sequences should be about 260 - 280 kb telomeric of HLA-DRA.

If AKIBA and CGM1 cells have the genome organization proposed by Campbell and Trowsdale (1993) in the respective area and the cosmid walking from the class III side is properly done, longer YDR2 but not shorter YDR3 should cover PBX2 and RAGE sequences. Southern blots of *Eco*RI-digested yeast genomic DNA of YDR2 or YDR3 were probed with radio-labeled cosmid KS 83 harboring class II HLA-DRA or class III cosmid KS75. Class II probe hybridized to YDR2 and YDR3, but class III probe only to YDR2 supporting the prediction. Hybridizing bands were accounted for restriction fragments of the cosmids, except for terminal fragments derived from the *Eco*RI site of cosmid linker (Fig. 4A). Thus, it was concluded that the presence of PBX2 and RAGE sequences in the MHC was neither specific to AKIBA genome nor due to cloning artifacts.

Copy numbers of PBX2 and RAGE segments in the human genome were examined by Southern blotting hybridization for human genome DNA digested with eight different restriction enzymes. Ten μg of high-molecular-weight DNA from human placenta, as well as 5 μg of YDR2-containing yeast DNA, were completely digested by each enzyme and size fractionated by electrophoresis on a 1% agarose gel. Southern blots of gels for the human or YAC DNA were hybridized with radio-labeled 0.8-kb RAGE probe (Fig. 2A). Hybridization patterns of human and YAC DNAs were essentially the same (Fig. 4B), and consistent with those predicted from the sequence determined; a single band was observed for four of eight enzyme digests. RAGE sequence is thus concluded to be present only once in the haploid human genome and to correspond to the real RAGE gene (Sugaya *et al.*, 1994).

Figure 4B shows also results obtained with PBX2 probe (1.1 kb). Hybridization pattern of YAC DNA was consistent with those from the genomic sequence, and five of eight digests had a single hybridization band. The pattern of human DNA was very similar to that of YAC while one or a few more bands with reduced intensity were seen after stringent washing. Essentially the same conclusion was drawn using another PBX2 probe of a 0.6-kb *Bam*HI fragment (data not shown). Weak bands reproducibly observed would not be related to CpG methylation of human DNA because enzymes used are not affected by methylation. Polymorphism may occur between the two chromosomes, but this is presumably not the case, since for no enzyme digests RAGE probe, as well as NOTCH3 probe mentioned later (Fig. 4B), did not give reproducible extra bands after stringent washing for all enzyme digests. A sequence closely related to, but distinct from, the PBX2 sequence may be present even in the haploid genome. If this is the case, the sequence may correspond to one previously mapped on 3q22-q23.

3. Chromosome **in situ** *hybridization*

As the second approach to determine locations of PBX2 and RAGE sequences (as well as *int-3* homolog described later), the direct R-banding fluorescence *in situ* hybridization (FISH), which is based on FISH combined with replication R-banding (Takahashi *et al.*, 1990) were carried out by Drs K. Mita and E. Takahashi (National Institute of Radiological Sciences, Chiba). They used the insert DNA of cosmid KS72 isolated by me as a probe. This clone contains the complete PBX2 sequence and portions of RAGE and *int-3* homolog sequences. They examined 100 typical R-banded (pro)metaphase plates. Of them, 54% showed complete double spots on both homologs, 41% were incomplete single or/and double spots on either or both homologs and in others (5%) no spots were detected. Fluorescent signals of KS72 were localized on 6p21.3 (Fig. 5), and no doublet signals were observed at other chromosome locations. Cultured cells for the *in situ* hybridization were from lymphocytes of a normal female donor (Takahashi *et al.*, 1990), and therefore the PBX2 sequence on 6p21.3 should be a general characteristics of the human genome. Monica *et al.* (1991) used the name PBX2 for cDNA selected by cross-hybridization to PBX1 cDNA. The present genomic sequence was essentially identical to PBX2 cDNA sequence, even in 5' and 3' UTRs. Thus, the present sequence should be the genomic PBX2 sequence. There is presently no direct information of sequence of the exact genome segment on 3q22-q23, which was hybridized with the ^{3}H-labeled PBX2 probe. To avoid possible confusion, I tentatively designated the gene in MHC "HOX12" rather than PBX2 in my paper (Sugaya *et al.*, 1994), after consulting with Genome Data Base (GDB) held by Johns Hopkins University. The GDB committee has recently noted, based on my mapping data, the HOX12 as the real PBX2 gene.

4. Human counterpart of mouse mammary tumor gene **int-3**; ***NOTCH3***

As noted above, nucleotide sequences of terminal portions of six cosmids being located centromeric to HOX12 were homologous with various portions of Notch-homologs of organisms such as *Drosophila*, *Xenopus*, zebrafish, rat, mouse and human. The highest nucleotide identity, about 80%, was noted for mouse mammary tumor gene *int-3*. Mouse mammary tumor development results from clonal outgrowth of tumor cells that frequently contain MMTV integrated at one or more specific genome regions called *int* loci. By mouse linkage analysis, one *int* locus, *int-3*, has been mapped between MHC class III *C4* and class II *H-2Aa* (Siracusa *et al.*, 1991), which corresponds to the area harboring the portion where I walked in the human genome. A consequence of MMTV integration at *int-3* is activation of expression of a 2.3-kb RNA species corresponding to 3' adjacent cellular sequence of the viral insertion. Its expression alters growth properties of HC11 mammary epithelial cells in culture. Nucleotide sequence of *int-3* RNA has been assigned to intracellular domain of a Notch-homologous gene (Robbins *et al.*, 1992). Notch was first found as a *Drosophila* neurogenic gene required for correct segregation of epidermal cells from neuronal cell precursors during embryogenesis, and subsequent studies demonstrated its roles in eye and sensillum development, in mesoderm differentiation and in oogenesis. *Drosophila* Notch is thus widely expressed during embryonic and adult development and mediates many different cell-cell interactions during normal fly development (Fortini and Artavanis-Tsakonas, 1993). Notch gene product is a transmembrane protein and composed of EGF-like repeats, Notch/lin-12 repeats, cdc10/ankyrin repeats and PEST regions (Fig. 6).

Notch homologs have been isolated from a variety of vertebrates and in human, two genes have been sequenced. The TAN1 on human chromosome 9 was shown to be involved in the translocation t(7;9)(q34;q34.3) of acute T cell lymphoblastic leukemia (T-ALL), and its cDNA sequence was reported (Ellisen *et al.*, 1991). The other human gene, hN, was isolated as a cDNA using PCR-primers designed from *Drosophila* and *Xenopus* Notch sequences, and its intracellular domain portion has been sequenced (Stifani *et al.*, 1992).

A) Structure of NOTCH3

To elucidate the organization of the Notch-homolog in MHC, various portions of cosmids KS71 and KS74 were subcloned and sequenced. Then, sequences were assembled into contigs. These sequences showed various degrees of identity with functional domains of Notch-homologs.

1) Intracellular domain; PEST sequences and cdc10/ankyrin repeats. A 0.7-kb sequence, f5 in Fig. 6, showed the highest homology with the nucleotide sequence from PEST domain to the first cdc10/ankyrin repeat (CDC10-1) of mouse *int-3* (77% nucleotide identity); alignment with *int-3* cDNA is shown in Fig. 7A. PEST domain was characterized by clusters of proline (P), glutamic acid (E), serine (S) and threonine (T) residues and was usually found in Notch-homologs. PEST sequences in *int-3* lie on either side of the cdc10/ankyrin repeats, and that of Fig. 7A showed higher homology to the proximal one. A 0.9-kb sequence, f6 in Fig. 6, showed homology with those for the cdc10/ankyrin repeats of several Notch-homologs and the highest homology was found with the first cdc10/ankyrin repeat (CDC10-1) of mouse *int-3* gene (93% identity, Fig. 7B); at the cDNA sequence

level, this is continuous from that of Fig. 7A, showing CDC10-1 interrupted by an intron. Intron-exon junctions are indicated by arrows. The cdc10/ankyrin repeats were first recognized in yeast cdc10/swi6 cell-cycle transcriptional regulatory proteins and in other proteins having roles in cell cycle control. These repeats are the most highly conserved domain among all of the known Notch-related proteins. A 1.6 kb sequence, f7, showed the highest homology to the last cdc10/ankyrin repeat (CDC10-6) of *int-3* (70% identity, Fig. 7C). Using the coding frame for the aligned *int-3* sequence human sequences were translated into amino acid sequences, which were sought in the PIR protein database; for five other frames, they were interrupted by termination codons (data not shown). Amino acids sequence found in PIR were those of PEST and cdc10/ankyrin domains of Notch proteins (Figs. 8C, D and E). The highest homology was noted for the mouse *int-3* protein sequence; 69.6% identity for the sum of continuous sequence of Figs. 8C and D, and 60.7% for that of Fig. 8E.

2) Extracellular domain; Notch/lin-12 cysteine repeats and cysteine-rich epidermal growth factor-like repeats (EGF cysteine repeats). A 1.2-kb sequence (f4 in Fig. 6) showed homology with nucleotide sequences for Notch/lin-12 cysteine repeats of human, mouse, rat, *Xenopus* and *Drosophila* Notch genes, and 1.0, 0.4 and 0.7-kb sequences (f1,f2 and f3), with those for Notch EGF-cysteine repeats of the species (data not shown). Depending on the coding frame for Notch genes thus aligned, the present human sequences were translated into amino acid sequences which were sought in PIR protein database. High levels of homology with EGF cysteine repeats

(Fig. 8A) and Notch/lin-12 repeats (Fig. 8B) of Notch-homologous proteins were found, and positions of cysteine were especially well conserved. Figures 8A and B indicate alignments for f3 and f4 sequences, respectively; EGF-like repeats occur more than thirty times in Notch genes, and that with the highest homology for each organism is presented (Fig. 8A). EGF-like repeats in Notch mediate extracellular interactions, as receptor for some diffusible ligands, binding sites in tissue matrices and recognition structures on certain cells.

The obtained sequences in cosmids KS71 and KS74 were thus homologous with Notch homologs virtually throughout all functional domains. Size of this human Notch homolog was estimated as more than 30 kb based on *Eco*RI map in Fig. 1. Owing to this large size, it is difficult to determine the full genomic sequence. Thus Notch cDNA clones were isolated from a human placenta cDNA library using several genomic fragments as probes, and the longest clone named PB5P4 was sequenced (3898 nt). Positions of cDNA sequences that have been determined at the present moment are indicated by horizontal lines at the bottom of Fig. 6.

B) Four subfamilies of mammalian Notch

A phylogenetic tree was constructed with the program package CLUSTAL V (Higgins 1994), analyzing amino acid sequences of the highly conserved cdc10/ankyrin repeats domain translated from the present cDNA sequences and the respective ones of known Notch-family proteins of a wide range of species (Fig. 9). This tree revealed four subfamilies of mammalian Notch genes. Human TAN1, rat Notch1, mouse Notch1, zebrafish Notch and *Xenopus* Notch are closely related and appear to constitute one group (designated "Notch subfamily 1"). Human Notch hN and rat Notch2 make another group ("subfamily 2"). Although the intracellular domain sequence of mouse Notch2 (Motch B) has not been determined, the mouse Notch2 likely belongs to subfamily 2 since it could be assigned to the mouse counterpart of rat Notch2 using EGF and Notch/lin-12 repeats sequences (Figs. 8A and 8B). The human gene found in MHC and mouse *int-3* comprise another group ("subfamily 3"), and thus the human gene was previously designated as "NOTCH3" (Sugaya *et al.*, 1994). A phylogenetic tree of extracellular Notch/lin-12 domain was basically the same to that presented by Fig. 9 (data not shown).

Main purpose for constructing the phylogenetic tree of Fig. 9 was to know evolutionary relationship of NOTCH3 with other Notch genes. Since NOTCH3 was rather distantly related with most of other Notch genes, region for which all amino acid sequences could be unambiguously aligned was confined to the highly conserved regions such as cdc10/ankyrin repeats domain. For this reason, sequences belonging to the same subfamily happened to be very close or identical (e.g. human Notch hN and rat Notch2). An example of the tree including less conserved regions was presented in Fig. 9 of Sugaya *et al.* (1994). To clarify the evolutionary origin and functions of the subfamily 3, it is important to know whether this subfamily exists in other species such as *Xenopus*. Because of sequence diversity, the subfamily 3 sequence may have not been detected by cross-hybridization with *Drosophila* Notch probes. Human NOTCH3 sequence should be useful for detecting subfamily 3 genes of other species.

Recently, a new mouse Notch homolog has been found and designated mouse Notch3 (Lardelli *et al.*, 1994). Their paper has been published after acceptance of my paper for printing (Sugaya *et al.*, 1994). In the present thesis, the phylogenetic tree of Fig. 9 was constructed including the newly reported mouse Notch3. The mouse Notch3 makes a subfamily clearly distinct from the subfamily 3 to which mouse *int-3* and human NOTCH3 in the MHC are grouped. Very recently, using mouse probes, the same group (Larsson *et al.*, 1994) cloned human counterparts of the mouse Notch2 and Notch3, and designated as human NOTCH2 and "NOTCH3", respectively. They mapped the human NOTCH2 on 1p13-p11 and the "NOTCH3" on 19p13.2-p13.1. Though the nucleotide sequence of their "NOTCH3" was not reported, the gene was clearly distinct from the "NOTCH3" found by me, judged both by its genetic position and by the partial amino acid sequence reported. Furthermore, the mouse Notch3 is known to be distinct from the mouse *int-3* gene. Therefore, existence of four subfamilies of mammalian Notch became clear, and the gene name "NOTCH3" happened to be used for the two different human genes. To avoid the confusion, the GDB committee has tentatively called my "NOTCH3" in the MHC region as INT3. The phylogenetic tree presented in Fig. 9 showed that the subfamily 3, to which my "NOTCH3" and mouse *int-3* are grouped, is most distantly related with other Notch genes. Based on this result, I now think the human Notch gene found in the MHC region to be named "NOTCH4" or "NOTCHR (Notch-related)" rather than "NOTCH3". It is now necessary to establish a nomenclature system of Notch-family genes for clarification of confusions having happened within and between various species.

C) Copy number of the human NOTCH3

The copy number of our "NOTCH3" was determined by hybridization to YAC-containing yeast DNA and human genomic DNA as done for RAGE and HOX12. Only one copy was found in the human genome (Fig. 4B). Combining this finding with the map locations and the phylogenetic tree of Fig.9, the NOTCH3 was assigned to the human counterpart of the uninterrupted form of the *int-3* gene. At present, the published sequence for extracellular domain of subfamily 3 is confined to that determined in this work though Robbins *et al.* (1992) noted, as their unpublished data, the uninterrupted form of *int-3* to presumably have Notch extracellular domain. It should be noted that Dr Shirayoshi in Prof. Nakatsuji Lab. of NIG cloned and characterized the extracellular domain of mouse *int-3* (personal communication).

5. Boundary of long-range GC% mosaic domains assigned by GC% measurement

A) Long-range GC% mosaic structures

The human MHC was found to be an example of long-range GC% mosaic structures by extensively analyzing sequences compiled by GenBank (Release 59, 1989) (Ikemura *et al.*, 1990). GenBank sequences have since accumulated significantly and, to confirm the mosaic structure, human MHC sequences in a recent GenBank (Release 80, 1994) were reexamined (Fukagawa, Sugaya, *et al.*, 1995). As before, GC% of non-redundant genomic sequences longer than 3 kb were calculated and arranged by their genetic positions (Fig. 10); the height of vertical bars corresponds to GC% and the width corresponds to sequence length. Most of the GenBank sequences were less than 10 kb and represented by rather thin vertical bars. About a 450-kb continuous black zone between class II HLA-DRA (abbreviated DRA in Fig. 10) and class III CYP21 does not correspond to GenBank sequences but to the region cloned by the present chromosome walk, and GC% distribution of the region will be focused on in this section.

Analyzing human DNA by cesium salt buoyant density fractionation, Bernardi and his colleagues (Bernardi *et al.*, 1985; Bernardi, 1989; Bernardi, 1993) defined five types of isochores with different GC% (H3, H2, H1, L2, and L1 in descending order of GC%). Figure 10 shows that most class I sequences are evidently GC-rich, with levels of the GC-richest isochores H3 (av. 53% GC; refer to Bernardi, 1993). In contrast, most sequences in class II are rather AT-rich and presumably correspond to L and H1 isochores (av. 40 and 45% GC, respectively). Class III sequences appear somewhat complicated though they are GC-richer than class II sequences; the centromeric portion seems to belong to the GC-richest isochores H3 and the telomeric portion to the second GC-richest isochores H2 (av. 49% GC). Therefore, confirming previous findings of the group to which I belong, the boundary of long-range GC% mosaic domain (i.e. the transition between the AT-rich and GC-rich domains) was assigned within about 450 kb harboring the junction between classes II and III. To clone the mosaic boundary, as extensively described in Introduction, cosmid walking from class III CYP21 to class II, and YAC and λ phage walking from class II HLA-DRA to class III were done, bridging classes II and III (Fig. 11). To analyze the base-compositional distribution of the walked area, insert DNAs of cosmid fragments indicated in Figure 11 were purified, digested by nuclease P1, and GC% was measured (Figs. 10 and 12) as described in Materials and Methods.

About 300 kb centromeric of CYP21, a fairly sharp GC% transition was found (Fig. 10), and this transition from H to L isochore was named the region "L/H transition" (Fig. 11; Fukagawa, Sugaya, *et al.*, 1995). So far concerning the Mb-level structures, this "L/H transition" was most apparent. It should be noted that, at a local level, this did not correspond to the direct transition from the GC-richest isochores H3 to the AT-rich isochore L, but appeared at least through the second GC-richest H2 level (Figs. 10 and 12): A ca. 60-kb region from CYP21, that harbors four fifth portion of TNX gene, was very GC-rich (mostly more than 55% GC, the GC-richest H3 isochores level) showing extension of the H3 level from the class III side, and a sharp transition to about 50% GC (the second GC-richest H2 isochores level) occurred. This transition from H3 to H2 isochore within the TNX gene was previously reported as the "H3/H2 transition" (Fig. 11; Ikemura *et al.*, 1992). Then, the H2 level continued about 160 kb (i.e., to the 5'-flanking region of NOTCH3) though there are certain local fluctuations of GC% between genes and their flanking regions (Figs. 11 and 12) as noted before (Ikemura and Aota, 1988; Bernardi, 1989). The successive 20 kb being centromeric of NOTCH3 was mainly composed of *Alu* repetitive elements (more than 25 repeats) as described below, and the GC% was about 45% (Fig. 12). This GC% level was equivalent to that of the H1 isochore but the domain size appeared too small to be assigned to an isochore.

Further chromosome walking from the dense *Alu* cluster of 20 kb to class II was difficult, possibly due to low density of unique sequences in the region. Thus, cosmid walking from class II HLA-DRA (i.e., from the AT-rich side) to class III, was done (KS43, KS44, KS83 and KS84 in Fig. 1) and, in a region about 30 kb telomeric of HLA-DRA, this walking became also difficult though the reason was not clear. Walking was continued using λ phage libraries constructed from the two YAC clones (YDR2 and YDR3) by Fukagawa, and he reached NOTCH3 (Fig. 11; Fukagawa *et al.*, 1995). His compositional analysis of the walked area showed the 180-kb region telomeric of HLA-DRA to be fairly homogeneously AT-rich and extension of AT-rich L isochore from class II. Just centromeric of the dense *Alu* cluster found by me, he found a dense LINE cluster of 30 kb. It has been thought that density of LINEs is high in AT-rich genome domains, while that of *Alu* is high in GC-rich domains (Bernardi, 1989; Korenberg and Rykowski, 1988; Holmquist, 1992). In the "L/H transition" area, this general characteristic was accentuated by their dense clusterings. There is a possibility that this characteristic is one of common features of boundaries of long-range GC% mosaics and therefore of chromosome bands.

B) Characteristic structures around the boundary of long-range GC% mosaic domains already published

To analyze the structures near and in the "L/H transition" area, the cosmid KS76 and the neighboring cosmid KS74, that span a total of about 80 kb, has been partially sequenced and the following three types of characteristic structures have already been published (Sugaya *et al.*, 1994; Fukagawa, Sugaya *et al.*, 1995). As noted above, at least 25 independent *Alu* repeats and 5 LINE-1 repeats were found. Interestingly, I found most of the *Alu* repeats densely cluster in about a 20-kb region. The 5 LINE-1 repeats also cluster in a 30-kb region. Furthermore, one pseudoautosomal boundary-like sequence was found and designated "PABL" (Fukagawa *et al.*, 1995). The interface between pseudoautosomal regions (PARs) of sex chromosomes and sex-specific regions is the pseudoautosomal boundary (PAB), and Goodfellow and his colleagues (Ellis and Goodfellow, 1989; Ellis *et al.*, 1989; Ellis *et al.*, 1990) reported sequences around the interface. Characteristic features of PAB of human sex chromosomes as a boundary of functional and structural domains, as well as a possible boundary of GC% mosaic domains and of chromosome bands, have been extensively explained by Fukagawa *et al.* (1995).

Summary of the organization of the walked 450 kb, including the GC% mosaic boundary (Fig. 11) is as follows;

[Class II; AT-rich side] HLA-DRA - 140 kb - PABL - 30 kb of LINE-1 cluster - 20 kb of *Alu* cluster - NOTCH3 - PBX2 - RAGE - 90 kb - TNX - 70 kb - CYP21 [Class III; GC-rich side]

6. Other characteristic structures found around the boundary of long-

range GC% mosaic domains

The structures mentioned above have already been published (Sugaya *et al.*, 1994; Fukagawa, Sugaya, *et al.*,1995). Other less characterized and unpublished structures around the "L/H transition" area are explained in this section. I have extensively characterized a segment extending about 140 - 180 kb centromeric to the TNX gene, that was cloned by both cosmids KS76 and KS74.

1) The 1.7 kb sequence of a *Pst*I fragment (PN112 in Figs. 12 and 13) in cosmid KS76, that located near the "L/H transition" area, showed evident homology with two human ESTs (expressed sequence tag); EST05686 (323 nt; EMBL HS7963; Adams *et al.*, 1993) and EST00737 (401 nt; EMBL HSXT00737; Adams *et al.*, 1992)(Fig. 13). Interestingly, the homology with EST00737 was split into two portions, and the split points on the genomic 1.7 kb sequence appeared to fit intron/exon junctions, so far judged by intron/exon consensus sequences. To examine whether the 1.7 kb region is transcribed, I screened cDNA libraries listed in Materials and Methods, using the 1.7 kb genomic fragment as a probe, and obtained a λ clone PN112ML2 (1042 nt) from monocyte cDNA library. Sequences of portions of PN112ML2, about 500 nt in total, were found to be homologous (about 70% nucleotide identity) but not identical with that of the genomic probe PN112, showing the cDNA to be transcribed from a genome segment distinct from the PN112. Sequences in the cDNA PN112ML2 were also homologous with those of the two ESTs mentioned. Figure 13 thus shows that major portions of the genomic 1.7-kb are homologous with the three separate ESTs (the two reported and the one

sequenced by me in this work), indicating the 1.7-kb region to be a portion of a transcribable or pseudo gene. When searched by BLASTX program against SWISS-PROT protein sequence database, a significant homology with a probable membrane antigen 3 encoded by saimiriine herpesvirus (VP03_HSVSA) was found (38% identities and 48% positives in 95 amino acids).

It is worthwhile to note that a 130-nt sequence of the cDNA PN112ML2, that is absent in the genomic 1.7-kb PN112, showed an evident homology with the 3' UTR of the prostaglandin-endoperoxide synthase 2 gene (GenBank HSU04636; 99% identity in 131 nt; Kosaka *et al.*, 1994). The homology was confined to this 130-nt sequence (Fig. 13C), and thus the newly found cDNA appeared not to correspond to the prostaglandin-endoperoxide synthase 2 gene itself. I tentatively designate the sequence of 130 nt as MERks1; MERs are the medium reiteracy frequency repeat sequences (Jurka *et al.*, 1993). A reduced level of homology with MERks1 (71% identity in 56 nt) was found for a portion of the 3' UTR of human BTEB (a GC box binding protein gene; GenBank HUMBTEB). Outside the MERks1, significant homologies were also found with a STS (sequence-tagged sites) of human chromosome 4 (GenBank HUM4STS735; 74% identity in 362 nt), with EST05686 (EMBL HS7963; 74% identity in 162 nt) and with a sequence near tandem GT repeat (GenBank HUMGTREPA; 67% identity in 168 nt) (Fig. 13C). So far searched by BLASTX program, no significant protein sequences homologous with those derived from the present cDNA were found.

2) During characterization of the 5' upstream region of NOTCH3 in cosmid KS74,

I found two kinds of trinucleotide microsatellite repeats; $(CTG)_{10}$ (CAG on the opposite strand) in the 3.6-kb *Eco*RI fragment harboring the 5' region of NOTCH3 and $(ATT)_{11}$ in the 10-kb *Eco*RI fragment within the *Alu* cluster (Figs. 12 and 14). Simple tandem repeats being termed microsatellites are abundant and polymorphic in eukaryotic genomes. Expansion of trinucleotide microsatellite repeats is an important mechanism of mutagenesis in humans and associated with human genetic diseases (interestingly, mainly with neurogenic diseases): e.g., fragile X syndrome (FMR-1; Kremer *et al.*, 1991), spinal and bulbar muscular dystrophy (SBMA; La Spada *et al.*, 1991), myotonic dystrophy (DM; Mahadevan *et al.*, 1992), Huntington disease (HD; Huntington's Disease Collaborative Research Group, 1993), spinal cerebellar ataxia type 1 (SCA-1; Orr *et al.*, 1993), and dentatorubral-pallidoluysian atrophy (DRPLA; Koide *et al.*, 1994). Four of these six syndromes (SBMA, HD, SCA-1, and DRPLA) result from expansion of $(CAG)_n$ repeats, while expansion of $(CCG)_n$ and $(CTG)_n$ (CAG on the opposite strand) repeats leads to FMR-1 and DM, respectively. Thus, five of the six diseases are associated with expansion of the $(CAG)_n$ repeat motif.

The two trinucleotide microsatellite repeats $(CTG)_{10}$ (CAG on the opposite strand) and $(ATT)_{11}$ found in the 5' upstream region of NOTCH3 (Figs. 12 and 14) may have certain roles of the gene expression. Notch was originally characterized as a neurogenic gene of *Drosophila* and, interestingly, genes responsible for neurogenic diseases such as narcolepsy have been suspected in the MHC class II or III region. The trinucleotide microsatellite repeats found in the 5' upstream of

NOTCH3 may become an informative marker useful for searching polymorphism of this gene and possibly for correlating it with certain genetic diseases. To ascertain this possibility, I have analyzed polymorphism of genomic DNAs from 23 different HLA haplotype B-cell lines, focusing on $(CTG)_{10}$ repeat locating near the 5' region of NOTCH3. The sequences just surrounding the $(CTG)_{10}$ repeat were evidently biased in the base-composition and appeared to be unsuitable for PCR primers. Thus a pair of PCR primers somewhat distant from the repeat unit was chosen (Fig. 14 B). To sensitively detect the expected polymorphisms, PCR products were digested by *Sau*3AI or *Sac*I followed by polyacrylamide gel electrophoresis. Only the fragment harboring the repeat unit showed difference of the gel mobility (Fig. 15 A). In the case of *Sau*3AI digestion, four types of alleles, as well as RFLPs (restriction fragment length polymorphism) of *Sau*3AI, were found (Fig. 15 B). To make sure of the *Sau*3AI RFLPs and of knowing VNTRs (variable numbers of tandem repeat units) of the *Sau*3AI-undigested fragments, analyses with *Sac*I digestion were separately carried out (Fig. 15 C). No RFLP existed in this case, and though the mobility difference of each allele became smaller, I could confirm its VNTRs. Sequence analysis on VNTRs showed the longest repeat unit to be $(CTG)_{12}$ and the shortest to be $(CTG)_6$. This difference between the longest and the shortest allele (i.e. 18-nt difference) was smaller than that expected from the mobility difference (Table 3). This may be due to a possible peculiarity of DNA structures and/or sequence difference other than the repeat unit. Polymorphism of this $(CTG)_n$ repeat was thus shown to be a useful polymorphic marker for NOTCH3. Its presumable association with narcolepsy or chromosome aberration in neoplasia (e.g., pleomorphic adenoma described later) is now under collaborative examination with Dr. Inoko's group of Tokai University. Other genetic diseases suspected in the walked area will be discussed later. It should be noted here that the PCR analysis for the exon 3 of RAGE gene and for the PABL region detected no polymorphism (Figs. 15D and E).

3) During characterization of 10-kb *Eco*RI fragment in cosmid KS74 within the *Alu* cluster, I found a dinucleotide repeat $(AT)_4(GT)_4(AT)_{22}$ (Figs. 12 and 14) near the telomeric side of trinucleotide microsatellite repeat $(ATT)_{11}$. In the porcine genome, Wilke *et al.* (1994) have shown that the majority of $(GT)_n$ repeats are either associated with SINEs or simple tandem repeats. A similar association with dispersed and tandem repeats was also known for human and bovine minisatellites (Armour *et al.*, 1989; Kaukinen and Varvio, 1992). Its dinucleotide repeat $(AT)_4(GT)_4(AT)_{22}$ found by me was also located near the *Alu* sequences, confirming the previous notion about the clustering of dispersed and tandem repeat types in mammalian DNA. Though the functional significance of these dinucleotide repeats is not clear, it is known that such dinucleotide repeats can form unusual DNA structures such as Z-DNA and 3- or 4- stranded DNA under physiological salt or pH conditions. Distinct repeat types have been proposed to be involved in a number of different cellular processes, including recombination, regulation of transcription and chromatin folding (for review on their structures and functions, see Vogt, 1990). However, their precise function (if they exist) and potential functional

interactions between associated repeat types remain to be elucidated.

4) Within the 20 kb of dense *Alu* cluster, I found also a sequence of ca. 270 nt homologous with the 5' flanking region of CD22 gene (EMBL S61408, 68.2% identity; tentatively designated MERks2; Figs. 12 and 16). The CD22 gene was mapped on chromosome 19, and around the gene a large number of *Alu* elements were found (Wilson *et al.*, 1993). Both positive and negative regulatory effects of *Alu* elements on transcription of genes such as CD22 and ε -globin genes have been proposed (Wu *et al.*, 1990; Saffer *et al.*, 1989; Oliviero *et al.*, 1988). The *Alu* elements, as well as trinucleotide repeats mentioned above, may have a certain regulatory role in the NOTCH3 transcription.

5) Within the 10.5-kb *Eco*RI fragment harboring LINE-1 repeats, I found a sequence of ca. 430 nt highly homologous with a segment of V region of TCRB (T-cell receptor β chain gene; GenBank HUMTCRB, 79% nucleotide identity in 429 nt) and of PZP (pregnancy zone protein gene; GenBank HSPZPA, 68% identity in 342 nt) (Figs. 12 and 17). Somewhat lower homology was also found with a segment ca. 90 kb upstream of TCRB (ca. 65% identity in 382 nt), a STS of human X chromosome (GenBank HUMSWX270, ca. 81% identity in 85 nt) and an intron of retinoblastoma susceptibility gene (GenBank HUMRETBLAS, ca. 61% identity in 401 nt). These should be assigned as a new class of MER and were tentatively designate it MERks3. Since a much reduced level of homology was found with the 5' end of LINE-1 repetitive sequence (GenBank HSKPNI03, ca. 58% identity in 279 nt), this MER may have been derived from a certain LINE.

V. Discussion

1. Newly found genes with respect to disease related genes

Using a ^{32}P-labeled cDNA library and hybridization techniques, Campbell and his co-workers (Kendall *et al.*, 1990) found MHC class III to be a very densely gene-packed region of the human genome, and predicted, in 140 - 170 kb centromeric of CYP21, several expressed genes (G16 - G18), though no sequence data were available. The new genes in the present work may correspond to genes predicted by the hybridization experiment, judging from their positions. In a region 80 - 130 kb centromeric to CYP21 (i.e., between TNX and RAGE genes), they predicted genes designated as G13 - G15 (Kendall *et al.*, 1990), and reported a partial sequence of G13 cDNA (Khanna *et al.*, 1992); I did no studies on this region. They found another gene, G9a, about 150 kb telomeric of CYP21 (the opposite side to my chromosome walk), and showed it to encode a protein containing cdc10/ankyrin repeats (Milner *et al.*, 1993). The cdc10/ankyrin repeats of this gene exhibited significant homology with those of Notch-homologs though other Notch domains such as Notch/lin-12 cysteine repeats and EGF cysteine repeats were absent. It should be noted here that very recently, Campbell and his co-workers (Aguado and Campbell, 1995) showed that the gene G17, which was assigned to 250-kb telomeric of the class II gene DRA, encodes PBX2. Also, after my publication (Sugaya *et al.*, 1994), Vissing *et al.* (1994) assigned RAGE cDNA to chromosome 6p21.3 by using FISH.

MHC genes including class III genes influence susceptibility to a wide range of diseases (Klein, 1986), and the newly found genes and closely related ones may influence certain diseases. Human NOTCH3 was assigned to the counterpart of mouse mammary tumor gene *int-3* being thought important to cell differentiation and tissue construction. Transgenic mice harboring *int-3* DNA fragment showed deregulation of normal development and hyperproliferation of glandular epithelia in salivary and mammary glands (Jhappan *et al.*, 1992). One chromosome aberration observed in neoplasia was located on human chromosome 6p21-p22. This was a t(6;8)(p21-p22;q12) translocation and resulted in pleomorphic adenoma in the salivary gland (Mitelman *et al.*, 1991). As found for Philadelphia-positive acute leukemias (Ph1$^+$ bcr$^-$ AL) caused by a chromosome translocation t(9;22)(q34;q11) (Chen *et al.*, 1989), *Alu* sequences may be involved for the translocation. In this connection, it is interesting to note that the YAC YDR2 and the cosmid KS76, having the 5' upstream region of NOTCH3 and the *Alu* cluster, were rather unstable, producing smaller versions of an initial clone during culture. Possible involvement of NOTCH3 gene in this neoplasia is now under collaborative examination with a group of Tokai University.

Recently, Larsson *et al.* (1994) reported the chromosome localization of the human NOTCH2 and their "NOTCH3" as 1p13-p11 and 19p13.2-p13.1, respectively. Because mice counterparts of these two genes are structurally very similar to the Notch1 that is the mice counterpart of human TAN1 (acute T-cell lymphoblastic leukemia), they suggest that these human genes may also be oncogenic if truncated in a fashion similar to TAN1 by translocation. They discussed the possible involvement of these genes in chromosomal rearrangements, mentioning human neoplasia-associated translocations related to 1p13-p11 or 19p13.2-p13.1, where the Notch genes were located. Then, all NOTCH genes are suspected to be proto-oncogenes and candidates for the sites of chromosome breakage in neoplasia-associated translocation. This view is consistent with my previous proposal that my "NOTCH3" may be involved for neoplasia-associated translocation (Sugaya *et al.*, 1994).

Linkage between MHC and diabetes has been intensively analyzed. In diabetes, formation of advanced glycosylation end products of protein (AGEs) is accelerated, thereby contributing to pathogenesis of diabetic angiopathy (Schmidt *et al.*, 1992). The receptor gene for AGEs in the MHC may influence diabetic complications, and therefore the linkage between MHC and diabetes should be studied also in this respect.

2. *Gross similarity of genes on 6p21.3 and those on 9q33-q34*

The human gene most closely related to NOTCH3 is TAN1 having been precisely mapped on 9q34.3, the gene related to HOX12 is PBX3 roughly mapped on 9q33-34, and that to TNX gene is HXB (the tenascin C gene) on 9q32-q34 (Table 4). This linkage similarity between the two sets of genes should reflect a common evolutionary origin, and appears to be part of paralogous chromosomal segments. Not only the four genes found by the group to which I belong but also the following genes on 6p21.3 were also found to have counterparts mostly mapped on 9q33-q34, by searching human Genome Data Base (GDB). As listed in Table 4, VARS2 (valyl-tRNA synthetase) is on 6p21.3 and VARS1 (valyl-tRNA synthetase) on 9; HSPA1 (heat shock 70kD protein-1) on 6p21.3 and GRP78 (glucose-regulated protein) on 9q33-34.1; C2 · C4A · C4B (complement component 2 · 4A · 4B, respectively) on 6p21.3 and C5 (complement component 5) on 9q33; COL11A2 (collagen XI, α -2 polypeptide) on 6p21.3 and COL5A1 (collagen V, α -1 polypeptide) on 9q34.2-34.3; RXRB (retinoide x receptor, β) on 6p21.3 and RXRA (retinoide x receptor, α) on 9q34.

The similarity between the two sets of genes mentioned should have been brought on by genome duplication and thus gives a realistic knowledge concerning evolutionary processes to built up the present human genome. The similarity also useful for finding undiscovered genes, especially candidate genes responsible for genetic disease, suspected in the walked area when possible hybridization probes of the respective counterpart were available.

The MHC region is evidently polymorphic, and various gene multiplications and deletions are noted. Large scale of *Alu* clustering and GC% mosaic structures may be related to the characteristic nature of this region. Iris *et al.* (1993) found 40 kb *Alu* dense clustering regions near the opposite end of class III, that borders on class I. Although the exact junctions of class I, II and III have not been characterized in detail, class III seems to be surrounded by large-scale dense *Alu* clusters at or near both ends. This may be related to evolutionary processes to built up the MHC. Transcriptional directions of NOTCH3, HOX12, RAGE, and TNX genes were the same (from centromere to telomere). This may be related to direction of DNA replication and/or regulation of gene expression.

VI. References

Adams, M. D., Dubnich, M., Kerlavage, A. R., Moreno, R. F., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. and Venter, J. (1992). Sequence identification of 2375 human brain genes. *Nature* **355**: 632-634.

Adams, M. D., Kerlavage, A. R., Fields, C. and Venter, J. C. (1993). 3400 expressed sequence tags identify diversity of transcripts from human brain. *Nature Genet.* **4**: 256-267.

Aguado, B. and Campbel, D. (1995). The novel gene G17, located in the human major histocompatibility complex, encodes PBX2, a homeodomain-containing protein. *GENOMICS* **25**: 650-659.

Ando, A., Kikuti, Y. Y., Kawata, H., Okamoto, N., Imai, T., Eki, T., Yokoyama, K., Soeda, E., Ikemura, T., Abe, K. and Inoko, H. (1994). Cloning of a new kinesin-related gene located at the centromeric end of the human MHC region. *Immunogenet.* **39**: 194-200.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Aota, S. and Ikemura, T. (1986). Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucl. Acids Res.* **14**: 6345-6355 & 8702 (for erratum).

Armour, J. A. L., Wong, Z., Wilson, V., Royle, N. J., and Jeffreys, A. J. (1989). Sequences flanking the repeat arrays of human minisatellites: Association with tandem and dispersed repeat elements. *Nucleic Acids Res.* **17**: 4925-4935.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953-958.

Bernardi, G. (1989). The isochore organization of the human genome. *Annu. Rev. Genet.* **23**: 637-661.

Bernardi, G. (1993). The isochore organization of the human genome and its evolutionary history - a review. *Gene* **135**: 57-66.

Bettecken, T., Aissani, B., Muller, C. R. and Bernardi, G. (1992). Compositional mapping of the human dystrophin- encoding gene. *Gene*, **122**: 329-325.

Bierkamp, C. and Campos-Ortega, J. A. (1993). A zebrafish homologue of the *Drosophila* neurogenic gene Notch and its pattern of transcription during early embryogenesis. *Mech. Dev.* **43**: 87-100.

Bird, A. P. (1987). CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3**: 342-347.

Bristow, J., Tee, M. K., Gitelman, S. E., Mellon, S. H. and Miller, W. L. (1993). Tenascin-X: a novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B. *J. Cell Biol.* **122**: 265-278.

Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D. and Olson, M. V. (1989). Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* **244**: 1348-1351.

Campbell, R. D. and Trowsdale, J. (1993). Map of the human MHC. *Immunol. Today* **14**: 349-352.

Chen, S. J., Chen, Z., d'Auriol, L., Coniat, M. L., Grausz, D. and Berger, R. (1989). Ph1$^+$ *bcr*$^-$ acute leukemias: implication of Alu sequences in a chromosomal translocation occurring in the new cluster region within the BCR gene. *Oncogene.* **4**: 195-202.

Coffman, C., Harris, w. and Kintner, C. (1990). Xotch, the *Xenopus* homolog of *Drosophila* Notch. *Science* **249**: 1438-1441.

Craig, J. M. and Bickmore, W. A.(1993). Chromosome bands - Flavours to Savour. *BioEssays*, **15**: 349-354.

Craig, J. M. and Bickmore, W. A.(1994). The distribution of CpG islands in mammalian chromosomes. *Nature Genet.*, **7**: 376-382.

Ellis, N. and Goodfellow, P. N. (1989). The mammalian pseudoautosomal region. *Trends Genet.*, **5**: 406-410.

Ellis, N. A., Goodfellow, P. J., Pym, B., Smith, M., Palmer, M., Frischauf, A.-M. and Goodfellow, P. N. (1989). The pseudoautosomal boundary in man is defined by an *Alu* repeat sequence inserted on the Y chromosome. *Nature*, **337**: 81-84 .

Ellis, N., Yen, P., Neiswanger, K., Shapiro, L. J. and Goodfellow, P. N. (1990). Evolution of the pseudoautosomal boundary in Old World monkeys and great apes. *Cell*, **63**: 977-986.

Ellisen, L. W., Bird, J., West, D. C., Soreng, A. L., Reynolds, T. C., Smith, S. D. and Sklar, J. (1991). TAN1, the human homolog of the *Drosophila* Notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell* **66**: 649-661.

Fortini, M. E. and Artavanis-Tsakonas, S. (1993). Notch: neurogenesis is only part of the picture. *Cell* **75**: 1245-1247.

Franco del Amo, F., Gendron-Maguire, M., Swiatek, P. J., Jenkins, N. A., Copeland, N. G. and Gridley, T. (1993). Cloning, analysis and chromosomal localization of Notch-1, a mouse homolog of *Drosophila* Notch. *Genomics* **15**: 259-264.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K, Ando, A., Inoko, H. and Ikemura, T. (1995). A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary.*Genomics* **25**: 184-191.

Gardiner, K., Aissani, B. and Bernardi, G. (1990). A compositional map of human chromosome 21. *EMBO J.* 9: 1853-1858.

Ghosh, D. (1991). New developments of a Transcription Factors Database.*TIBS* **16**: 445-447.

Hein, J. (1990). Unified approach to alignment and phylogenies. *In* "Methods in Enzymology" (R. F. Doolittle, Eds.) Vol. 183, pp.626-645, Academic press, New York.

Higgins, D. G. (1994) CLUSTAL V: multiple aligment of DNA and protein sequences. *Methods. Mol. Biol.* **25**: 307-318.

Holmquist, G. P. (1987). Role of replication time in the control of tissue specific gene expression. *Am. J. Hum. Genet.*, **40**: 151-173.

Holmquist, G. P. (1992). Chromosome bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.*, **51**: 17-37.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's Disease chromosomes. *Cell* **72**: 971-983.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13-34.

Ikemura, T. and Aota, S. (1988). Global variation in G+C content along vertebrate genome DNA: Possible correlation with chromosome band structures. *J. Mol. Biol.* **203**: 1-13.

Ikemura, T., Wada, K. and Aota, S. (1990). Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. *Genomics* **8**: 207-216.

Ikemura, T. and Wada, K. (1991). Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res.*, **19**: 4333-4339.

Ikemura, T., Matsumoto, K., Ishihara, N., Ando, A. and Inoko, H.(1992). In Tsuji, K., Aizawa, M. and Sasazuki, T. (eds) , HLA 1991. Vol.2, pp. 125-128 Oxford University Press, Oxford.

Inoko, H., Ando, A., Kimura, M. and Tsuji, K. (1985). Isolation and characterization of the cDNA clone and genomic clones of a new HLA class II antigen heavy chain, DO alpha. *J. Immunol.* **135**: 2156-2159.

Iris, F. J. M., Bougueleret, L., Prieur, S., Caterina, D., Primas, G., Perrot, V., Jurka, J., Rodriguez-Tome, P., Claverie, J. M., Dausset, J. and Cohen, D. (1993). Dense Alu clustering and a potential new member of the NFkB family within a 90 kilobase HLA class III segment. *Nature Genet.* **3**: 137-145.

Jhappan, C., Gallahan, D., Stahle, C., Chu, E., Smith, G. H., Merlino, G. and Callahan, R. (1992). Expression of an activated Notch-related *int-3* transgene interferes with cell differentiation and induces neoplastic transformation in mammary and salivary glands. *Genes Dev.* **6**: 345-355.

Jurka, J., Kaplan, D. J., Duncan, C. H., Walichiewicz, J., Milosavljevic, A., Gayathri, M. and Solus, J. F. (1993). Identification and characterization of new human medium reiteration frequency repeats. *Nucleic Acids Res.*, **21**: 1273-1279.

Kamps, M. P., Murre, C., Sun, X. and Baltimore, D. (1990). A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in Pre-B All. *Cell* **60**: 547-555.

Kaukinen, J., and Varvio, S.-L. (1992). Artiodactyl retoroposons: Association with microsatellites and use in SINEmorph detection by PCR. *Nucleic Acids Res.* **20**: 2955-2958.

Kawai, J., Ando, A., Sato, T., Nakatsuji, T., Tsuji, K.and Inoko, H. (1989). Analysis of gene structure and antigen determinants of DR2 antigen using DR gene transfer into mouse L cells. *J. Immunol.*, **142**: 312-317.

Kendall, E., Sargent, C. A. and Campbell, R. D. (1990). Human major histocompatibility complex contains a new cluster of genes between the HLA-D and complement C4 loci. *Nucl. Acids Res.* **18**: 7251-7257.

Khanna, A. and Campbell, D. (1992). Characterization of a novel gene G13 in the class III region of the human MHC. *In* "HLA 1991" (K. Tsuji, M. Aizawa and T. Sasazuki, Eds.), Vol.2, pp.198-201, Oxford Univ. Press, Oxford.

Klein, J. (1986). "Natural History of the Major Histocompatibility Complex" John Wiley, New York.

Koide, R., Ikeuch, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. and Tsuji, S. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.* **6**: 9-12.

Korenberg, J. R. and Rykowski, M. C. (1988). Human genome organization: *Alu*, Lines and the molecular structure of metaphase chromosome bands. *Cell*, **53**: 391-400 .

Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R. and Richards, R. I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* **252**: 1711-1714.

Kosaka, T., Miyata, A., Ihara, H., Hara, S., Sugimoto, T., Takeda, O., Takahashi, E. and Tanabe, T. (1994). Characterization of the human gene (PTGS2) encoding prostaglandin-endoperoxide synthase 2. *Eur. J. Biochem.* **221**: 889-897.

Lardelli, M. and Lendahl, U. (1993). Motch A and Motch B - two mouse Notch homologues coexpressed in a wide variety of tissues. *Exp. Cell Res.* **204**: 364-372.

Lardelli, M., Dahlstrand, J. and Lendahl, U. (1994). The novel *Notch* homologue mouse *Notch 3* lacks specific epidermal growth factor-repeats and is expressed in proliferating neuroepithelium.

Mech. Dev. **46**: 123-136.

Larsson, C., Lardelli, M., White, I. and Lendahl, U. (1994). The human NOTCH1, 2, and 3 genes are located at chromosome positions 9q34, 1p13-p11, and 19p13.2-p13.1 in regions of neoplasia-associated translocation. *Genomics* **24**: 253-258.

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. and Fischbeck, K. H. (1991). Androgen receptor gene mutation in X-linked spinal and bulbar muscular atrophy. *Nature* **352**: 77-79.

Lichter, P., Ledbetter, S. A., Ledbetter, D. H. and Ward, D. C. (1990). Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines. *Proc. Natl. Acad. Sci. USA* **87**: 6634-6638.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hey, K., Leblond, S., Earle-MacDonald, J., De Jong, P. J., Wieringa, B. and Korneluk, R. G. (1992). Myotonic dystrophy mutation: An unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**: 1253-1258.

Maniatis, T, Fritsch, E. F. and Sambrock, J. (1989). "Molecular Cloning: A Laboratory Manual", pp. 9.47-9.62, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Matsumoto, K., Arai, M., Ishihara, N., Ando, A., Inoko, H. and Ikemura, T. (1992a). Cluster of fibronectin type III repeats found in the human major histocompatibility complex class III region shows the highest homology with the repeats in an extracellular matrix protein, Tenascin. *Genomics* **12**: 485-491.

Matsumoto, K., Ishihara, N., Ando, A., Inoko, H. and Ikemura, T. (1992b). Extracellular matrix protein tenascin-like gene found in human MHC class III region. *Immunogenetics* **36**: 400-403.

Matsumoto, K., Saga, Y., Ikemura, T., Sakakura, T. and Chiquet-Ehrismann, R. (1994). The distribution of Tenascin-X is distinct and often reciprocal to that of Tenascin-C. *J. Cell. Biol.* **125**: 483-493.

Milner, C. M. and Campbell, R. D. (1993). The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats. *Biochem.J.* **290**: 811-818.

Mitelman, F., Kaneko, Y. and Trent, J. (1991). Report of the committee on chromosome changes in neoplasia. *Cytogenet. Cell Genet.* **58**: 1053-1079.

Monica, K., Galili, N., Nourse, J., Saltman, D. and Cleary, M. L. (1991). PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1. *Mol. Cell. Biol.* **11**: 6149-6157.

Neeper, M., Schmidt, A. M., Brett, J., Yan, S. D., Wang, F., Pan, Y. E., Elliston, K., Stern, D. and Shaw, A. (1992). Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. *J. Biol. Chem.* **267**: 14998-15004.

Nourse, J., Mellentin, J. D., Galili, N., Wilkinson, J., Stanbridge, E., Smith, S. D. and Cleary, M. L. (1990). Chromosomal translocation t(1;19) results in synthesis of a homeobox fusion mRNA that codes for a potential chimeric transcription factor. *Cell* **60**: 535-545.

Oliviero, S. and Monaci, P. (1988). RNA polymerase III promoter elements enhance transcription of RNA polymerase II genes. *Nucl. Acids. Res.* **16**: 1285-1293.

Orr, H. T., Chung, M., Banfi, S., Kwiatkowski, T. J., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4**: 221-226.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**: 2444-2448.

Pilia, G., Little, R. D., Aissani, B., Bernardi, G. and Schlessinger, D.(1993). Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics*, **17**: 456-462.

Robbins, J., Blondel, B. J., Gallahan, D. and Callahan, R. (1992). Mouse mammary tumor gene *int-3*: a member of the notch gene family transforms mammary epithelial cells. *J. Virol.* **66**: 2594-2599.

Saccone, S., De Sario, A., Wiegant, J., Raap, A.K., Della Valle, G. and Bernardi, G. (1993). Correlation between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, **90**: 11929-11933.

Saffer, J. D. and Thurston, S. J. (1989). A negative regulatory element with properties similar to those of enhancers is contained within an *Alu* sequence. *Mol. Cell. Biol.* **9**: 355-364.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.

Schamboeck, A., Korman, A. J., Kamb, A. and Strominger, J. L. (1983). Organization of the transcriptional unit of a human class II histocompatibility antigen: HLA-DR heavy chain. *Nucl. Acids Res.* **11**: 8663-8675.

Schmidt, A. M., Vianna, M., Gerlach, M., Brett, J., Ryan, J., Kao, J., Esposito, C., Hegarty, H., Hurley, W., Clauss, M., Wang, F., Pan, Y. E., Tsang, T. C. and Stern, D. (1992). Isolation and characterization of two binding proteins for advanced glycosylation end products from bovine lung which are present on the endothelial cell surface. *J. Biol. Chem.* **267**: 14987-14997.

Sealey, P. G., Whittaker, P. A. and Southern, E. M. (1985). Removal of repeated sequences from hybridization probes. *Nucl. Acids. Res.* **13**: 1905-1922.

Senger, G., Ragoussis, J., Trowsdale, J. and Sheer, D. (1993). Fine mapping of the human MHC class II region within chromosome band 6p21 and evaluation of probe ordering using interphase fluorescence *in situ* hybridization. *Cytogenet. Cell Genet.*, **64**: 49-53.

Siracusa, L. D., Rosner, M. H., Vigano, M. A., Gilbert, D. J., Staudt, L. M., Copeland, N. G. and Jenkins, N. A. (1991). Chromosomal location of the octamer transcription factors, Otf-1, Otf-2, and Otf-3, defines multiple Otf-3-related sequences dispersed in the mouse genome. *Genomics* **10**: 313-326.

Stifani, S., Blaumueller, C. M., Redhead, N. J., Hill, R. E. and Artavanis-Tsakonas, S. (1992). Human homologs of a *Drosophila* Enhancer of Split gene product define a novel family of nuclear proteins. *Nature Genet.* **2**: 119-127.

Sugaya, K., Fukagawa, T., Matsumoto, K., Mita, K., Takahashi, E., Ando, A., Inoko, H. and Ikemura, T. (1994). Three genes in the human MHC class III region near the junction with the class II: gene for receptor of advanced glycosylation end products, PBX2 homeobox gene and a Notch homolog, human counterpart of mouse mammary tumor gene *int-3*. *Genomics*. **23**: 408-419.

Takahashi, E., Hori, T., O'Connell, P., Leppert, M. and White, R. (1990). R-banding and nonisotopic in situ hybridization: precise localization of the human type II collagen gene (COL2A1). *Hum. Genet.* **86**: 14-16.

Takahashi, E., Yamauchi, M., Tsuji, H., Hitomi, A., Meuth, M. and Hori, T. (1991).

Chromosome mapping of the human cytidine-5'-triphosphate synthetase (CTPS) gene to band 1p34.1-p34.3 by fluorescence in situ hybridization. *Hum. Genet.* **88**: 119-121.

Vissing, H., Aagaard, L., Tommerup, N. and Boel, E. (1994). Localization of the human gene for advanced glycosylation end product-specific receptor (AGER) to chromosome 6p21.3. *GENOMICS* **24**: 606-608.

Vogt, P. (1990). Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on highly conserved "chromatin folding code". *Hum. Genet.* **84**: 301-336.

Weinmaster, G., Roberts, V. J. and Lemke, G. (1991). A homolog of *Drosophila* Notch expressed during mammalian development. *Development* **113**: 199-205.

Weinmaster, G., Roberts, V. J. and Lemke, G. (1992). Notch2: a second mammalian Notch gene. *Development* **116**: 931-941.

Wharton, K. A., Johansen, K. M., Xu, T. and Artavanis-Tsakonas, S. (1985). Nucleotide Sequence from the Neurogenic Locus Notch Implies a Gene Product That Shares Homology with Proteins Containing EGF-like Repeats. *Cell:* **43**: 567-581.

Wike, K., Jung, M., Chen, Y., and Geldermann, H. (1994). Porcine $(GT)_n$ sequences: structure and association with dispersed and tandem repeats. *Genomics* **21**: 63-70.

Wilson, G. L., Najfeld, V., Kozlow, E., Menniger, J., Ward, D. and Kehrl, J. H. (1993). Genomic structure and chromosome mapping of the human CD22 gene. *J. Immunol.* **150**: 5013-5024.

Wolfe, K.H., Sharp, P.M. and Li, W-H.(1989). Mutation rates among regions of the mammalian genome. *Nature*, **337**: 283-285.

Wu, J., Grindlay, G. J., Bushel, P., Mendelsohn, L. and Allan, M. (1990). Negative regulation of the human ϵ - globin gene by transcriptional interference: role of an *Alu* repetitive element. *Mol. Cell. Biol.* **10**: 1209-1216.

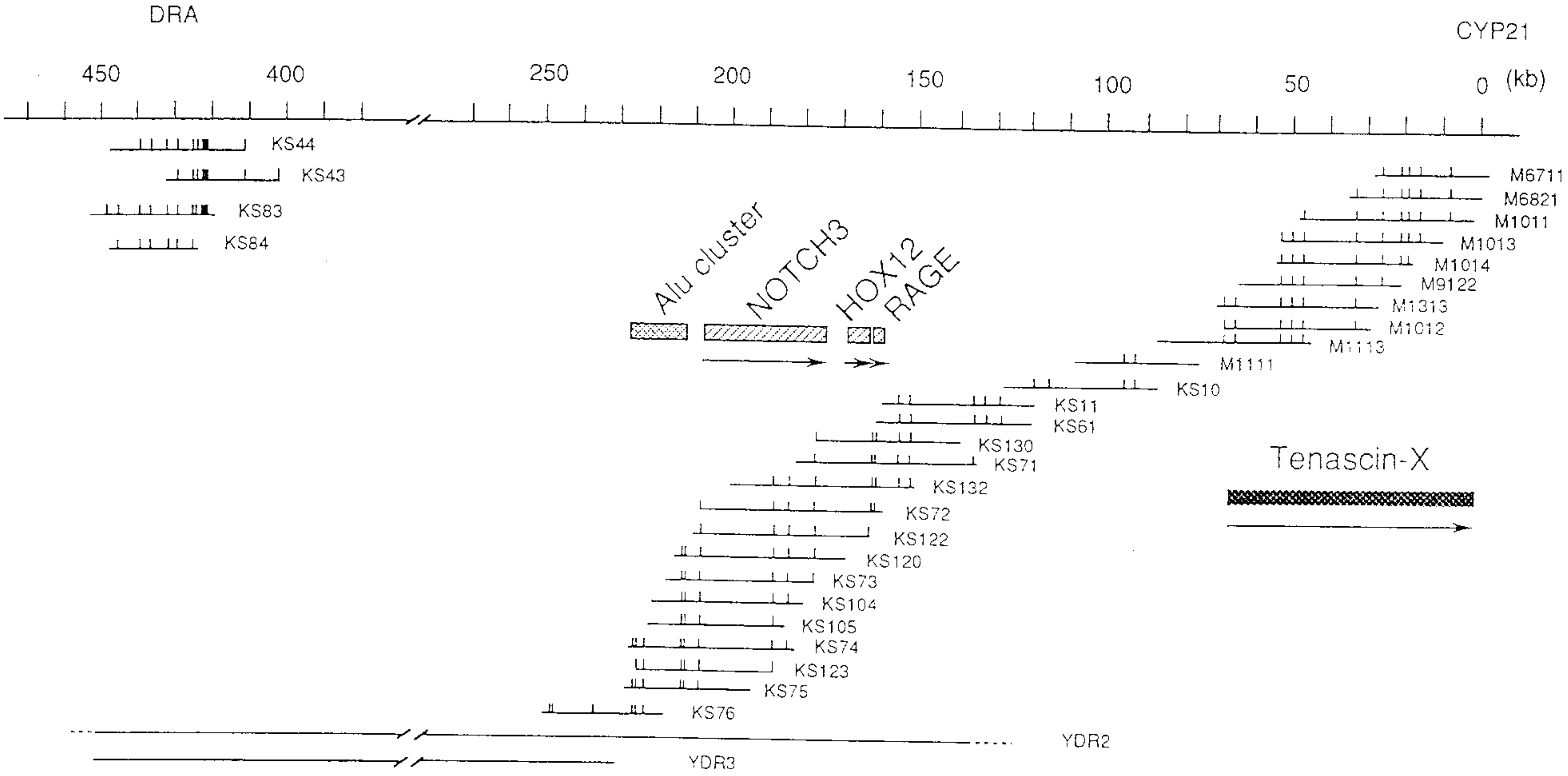


Fig. 1) Molecular map of contiguous cosmids and YAC clones that cover junction of MHC classes II and III. Ordered cosmid clones are represented by horizontal lines with vertical bars indicating *Eco*RI sites. Hatched boxes indicate locations of NOTCH3, HOX12 and RAGE genes and black box, that of tenascin-X gene (TNX). Arrows show directions of transcription.

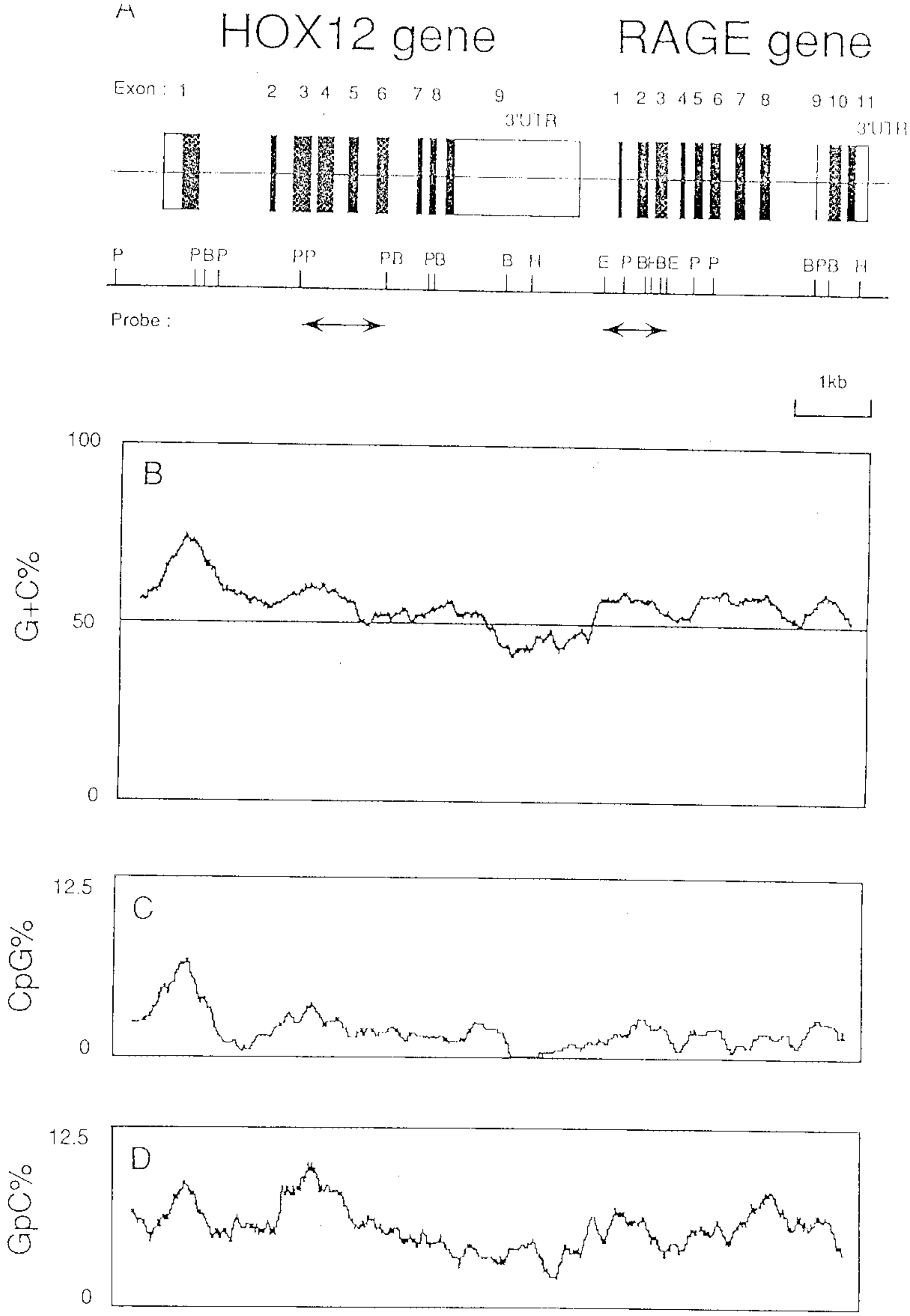


Fig. 2) Structures of HOX12 and RAGE genes and distributions of GC%, CpG% and GpC%. (A) Black boxes indicate protein-coding regions and open boxes, UTRs. Sites for *Eco*RI (E), *Pst* I (P), *Hid*III (H) and *Bam* HI (B) are shown by short vertical bars on top of the horizontal line, under which fragments for hybridization probes are marked. Distributions of GC%(G+C%) (B) and of dinucleotide CpG% (C) or GpC% (D) were calculated with window length of 500 nucleotides.

A

```
                                    GATCCCTGAGACTGAGGGGTTTACGGG -951
-950 CTGTGAATGGACCTTCAGCCCTGCCCACCCTCCCTCCCCACTGCTGCTGA -901
                             AP-1     TATA-BOX
-900 GTCTGTCTGATGTTTTGGTTGTGTGAATAAATATAATTCCCCTCTGGACT -851
                         AP-2              PEA3
-850 GCAGACTGGTATCTGGGGGGCCCAGGCGGGGTGAAAGGTAGGAAGGTGAG -801
                               NF1          AP-2
-800 GCCAGAGGCCTTTTCTCTCCCCAGTCTGGCCAGAGGCCAGCTCCCCTCCC -751
                         GMCSF r-IRE     r-IRE         NF-E1
-750 CGGCTGGTTAATTACTGGCTCATTAAGCAGCGGCTGGAGACCTCCCTAAT -701
       AP-2                                TCF-1
-700 TATCTCCCCCAGCCCCCCTCTTCGGTTTTAATTAAGTAGAACAGGGAGGG -651
       GMCSF                                 AP-2
-650 GAGTCATTAGAACAAGAAATATGAACTGAGCTGCCGGTGAACCCCAGGCA -601
                                       AP-1  r-IRE  BHLH
-600 TTCCAGCGGCCTGAGTCCACATCGCTTAGATCCCTGATTCAGGACCCAGG -551
               AP-2         TCF-1                 ETS-1
-550 TGACAGACGCCCCCAGCCGCCAACACAGCCCCACTCCTAGGCCGCGGAAG -501
             NFKB                BHLH  CF1  SIF
-500 TCCAGCCAGGGGGCTTTCCCATATCTTTCAGATGGCCCGTCTCCTCCCCT -451

-450 CATCCCCTCTTCCCTCTCCCCTCCTCCACTAGGTCTCAGTTCCTCTGTTT -401
                 GCF
-400 CTGTGTCTCTCTCTCCGCCCCCAGCTCCTCCCTGTTCCTCCTCTCTTCTC -351
                           AP-2     AP-2
-350 CCCTCCTCTTTCCTCTCCGGCTCCCCTCCCCCAGCCTCCCTCCCTCGTTC -301
     AP-2
-300 CCCCCCCTTCTCCCTCCTCCCTCCTCCCTCTCTCTCACACACACCCCCGC -251
                               GMCSF    GCF
-250 TTGGGCCTCCTCTCTCTCTCCGGCTCCATTTTCTCCGCCGCCGGGGGCCG -201
                      AP-2
-200 GGGTCTCCTGTGGGGGGCCCAGCCGGTATCCCAGGTCTCCCTTCAGTGCC -151
             AP-2                             SP1
-150 GGGGTGAACCCCCGGGGGAGCCGGGAGCCGGGGGCAGACGGGCGGGGGTT -101
     SP1                 GCF   AP-2
-100 GGGGCGGAGGGAGCAGCGGCCCCAGCGAGTTTGGGGGGAGAAGTAACCAG  -51
              GC-BOX   SP1                   AP-2
 -50 GCGGGGGGAGGGGCGGAGCAGAGAGGGGCCCTCAGGCCCCCCCCCCAGCT   -1

   1 atg
```

B

```
                                                  AP-2
-1800 GCAACTATTCTGTTTCTCTCCTGGGCTCCCCCACTTTCCCTTCCCCACCC -1751
      BHLH        LBF-1 r-IRE          AP-2
-1750 CACTTGTATGCTCTGGAATCTGTGGAGACGCCCAGCCCTGCCCAATCAGAG -1701
             CF1         r-IRE                 CF1
-1700 ATGCCAAAAATGGGGACATGACTTCTGGACAGAGGGACATGGGACACGCC -1651
              AP-2   SP1 AP-2 GCF AP-2
-1650 CCCATGCATCCCCACCCCCGCCCCTCCGGACGGCTTACTTACCTCATACG -1601
                                                  AP-1
-1600 CAGCTCATCTTAAACCAATAGAATCGCTCGGTGGACGAGAGTGTCTGACT -1551
      NF-GMB          NFKB           NF-IL6
-1550 CAGATATCTACCTCGGAGGGAGTTTCTGCTACTTTAGGGAATTATTGACT -1501
                                            TCF-1
-1500 GGGCTTTGGGGTTGAACTTTTTTTTTTTTAAAGAAAGAAAAAGAAACCCT -1451
            BHLH
-1450 GGGATCCATCTGTTTTTTTTTGTTGTTGTTGTTTTTTGTTGTTGGTGGT -1401
                                             NF-IL6
-1400 GGTGGTGGTGGTGGTTCTCAATTTTTAATTTAGTTTGGGGAAGTAGC -1351
                    TATA-BOX
-1350 TTGTTTTTTTTTTTATAAATATGTTGATTTCTTGTCTTTTTTTTTTTATTT -1301
                           GATA-1  TCF-1
-1300 CTTACTTTCCCATATTAGGGGTGATAGCCAAAGGGGTTCTGGTAAGAGAA -1251
                     LBP-1       AP-2
-1250 AGGGGGACAAACAGAACTGGTAAAGAGCCCCCCCTGGCTCCAGGCCTGTC -1201
       ETS-1 r-IRE PEA3
-1200 CATCAGGAAGTAAATTTTACAGGGCACCAAGCTTTGCCCCCTAAAATCCC -1151
                                           GMCSF BHLH
-1150 TTAGGTGTTCTTTGTTCATGCAGGCAGGTTTCTGCCGCATTTGATGTGGA -1101

-1100 GGCAGTGAAGGGCTTGCCCTGCTGGCCTCTCATCCCCCTTCTTCCCACAA -1051
                   r-IRE r-IRE         NF-IL6 PEA3
-1050 CCCTTGGGCAGGGCTGGACTCAGTAATTTTGAGGAAATTGAAGATGCCAT -1001
                            r-IRE
-1000 CTTCCCCTGTGAGTGACATGTCTTTAATTTTTTAAAAACTACTATTTGA  -951
                                                     CF1
 -950 AAATTGGAGGGGGAAGAATGGGAAGGGAGTTATTGCCAAATATGTTAAAT  -901

 -900 ATGGGTTGGGGTGCTTGTATATGTATCTTCCTCAATTTCCCCATAAATGA  -851

 -850 GGTATCTTTTTGTCACACCAAAATCAAGGGGTAGGGAGAGGGAGGAGGTT  -801
        TCF-1    BHLH        TCF-1
 -800 GCAAAAAGCCAGATGTGGGGAAAAGTAACATCAACACTGTCCCATCCTCA  -751
            LBP-1    BHLH                       r-IRE
 -750 GCCCTGAACTAGCTACCATCTGATCCCCTCAGACATTCTCAGGATCTTAC  -701
                          AP-2  GMCSF                 LBP-1
 -700 AAGACTGTCAGAGTGGGGAACCCCTCCCATTAAAGATCCGGCCAGGACTG  -651

 -650 GGACAGGTTGGAAGTGTGATGGGTGGGGGGTGGGAGGCATGGGCCGGGGG  -601
                   BHLH                 TCF-1
 -600 GCAGTTCTCTCCTCACTTGTAAACTTGTGTAGTTTCACAGAAAAAAAACA  -551

 -550 AAATGCAGTTTTAAATAAAGAAATTTCTTTTTTCCCTGGGTTTAGTTGAG  -501

 -500 AATTTTTTTCAAAAAACATGAGAAACCCCAGAAAAAAAATGATTTTCTTT  -451
                 TCF-1                AP-2              r-IRE
 -450 CACGAAGTTCCAAACAGGTTTCTCTCCTGTTCCCCAGCCTTGCCTTCATG  -401
               BHLH MRE            TCF-1 LBP-1 LF-A1
 -400 ATGCAGGCCCAATTGCACCCTTGCAGACAACAGTCTGGCCTGAACCCTAT  -351
                NF-IL6     CF1                       AP-2
 -350 TGATGCAACTTTGGGCAATCAAGATGGGGCTCCAGTGGGTCACCAGGCAG  -301
                 NF-IL6                      NF-IL6
 -300 CCCTGATGGACTGATGGAATAAATAGGATCGGGGGCTCTGAGGGAATGAG  -251
                      AP-2    AP-2 AP-2           r-IRE    SP1
 -250 ACCCTAGAGGGTACACTCCCCATCCCCCAGGGAAGTGACTGTACCCAGAG  -201
                AP-2         GATA-1         LBP-1
 -200 GCTGGTAGTACCCAGGGGTGGGGTGATAATTATTTCTCTAGTACCTGAAG  -151
                TCF-1  r-IRE
 -150 GACTCTTGTCCCAAAGGCATGAATTCCTAGCATTCCCTGTGACAAGACGA  -101
           CF1                                 SP1
 -100 CTGAAAGATGGGGGCTGGAGAGAGGGTGCAGGCCCCACCTAGGGCGGAGG   -51
       TCF-1                     r-IRE       PEA3
  -50 CCAGAGCAGGGAGAGGGGCAGACAGAGCCAGGACCCCTGGAAGGAAGCAGG    -1

    1 atg
```

Fig. 3) Factor binding sites in regulatory regions of HOX12 (A) and RAGE (B) genes. The nucleotide sequence is numbered, number 1 corresponding to a of initiator atg. Potential binding sites for transcription factors are underlined. The 3' UTR sequence of HOX12 is written in italics.

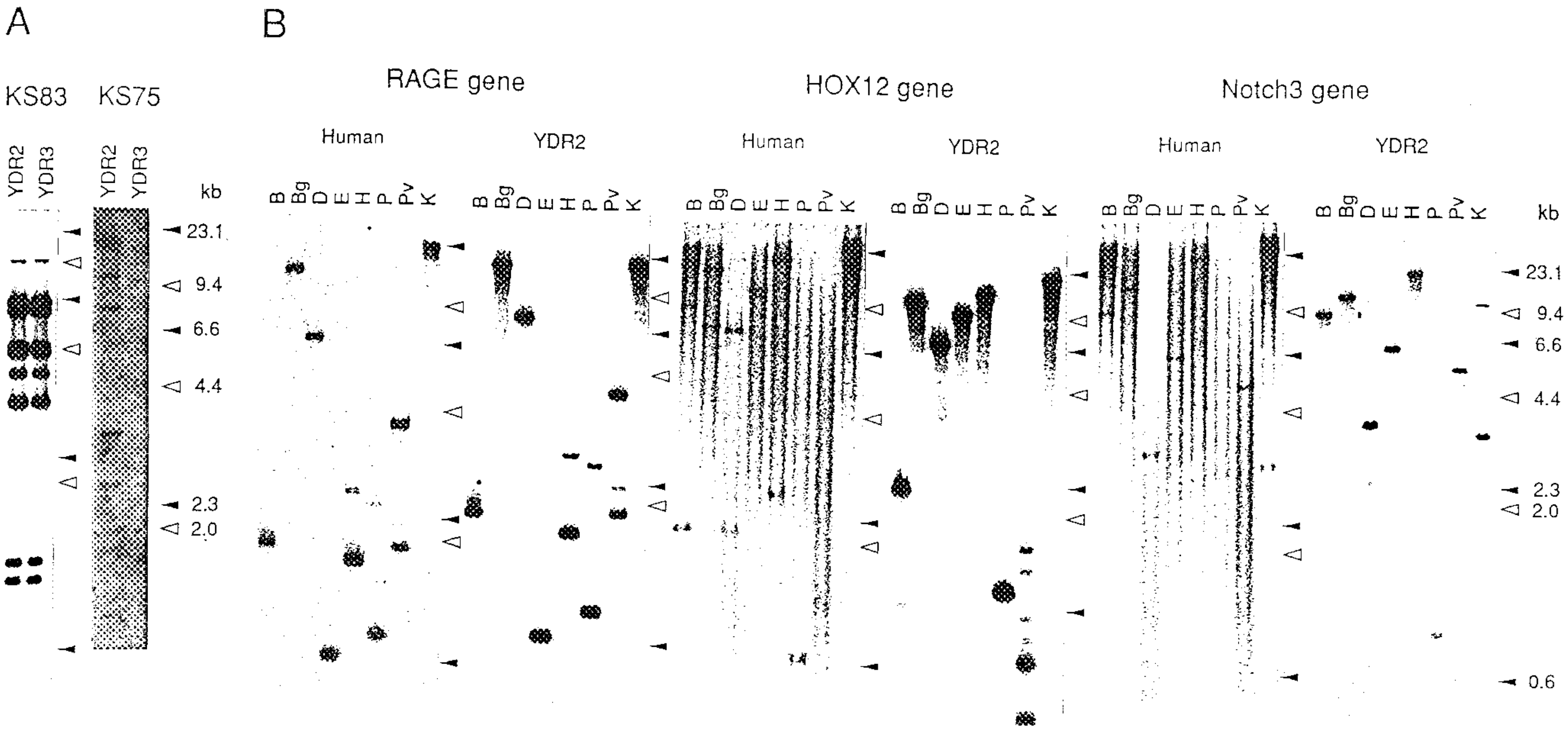


Fig. 4) Southern blot analysis of YAC and human DNAs. (A) For each lane, 5 µg of YDR2- or YDR3-carrying yeast DNA was digested with *Eco*RI, separated by electrophoresis on a 1% agarose gel, blotted onto a Hybond N nylon membrane and hybridized with ^{32}P-labeled KS83 or KS75 probe. (B) For each lane, 10 µg of human genomic DNA or 5 µg of YDR2-carrying yeast DNA was digested with the restriction enzyme indicated, electrophoresed, blotted and hybridized with ^{32}P-labeled RAGE, HOX12 or NOTCH3 probe: B, *Bam*HI; Bg, *Bgl*I; D, *Dra*I; E, *Eco*RI; H, *Hid*III; P, *Pst*I; Pv, *Pvu*II; K, *Kpn*I. Positions of probes are listed in Fig. 2A or Fig. 6. For stringent washing, filters were washed at 65℃ in 0.1 x SSPE for 15 min. Size in kb is indicated on the left. Owing to difference of electrophoretic conditions, the smallest two *Pvu*II bands observed for YDR2 by hybridization with HOX12 probe ran off the gel for human DNA. See FIG. 4 of Sugaya *et al.* (1994) for a figure with a better quality.

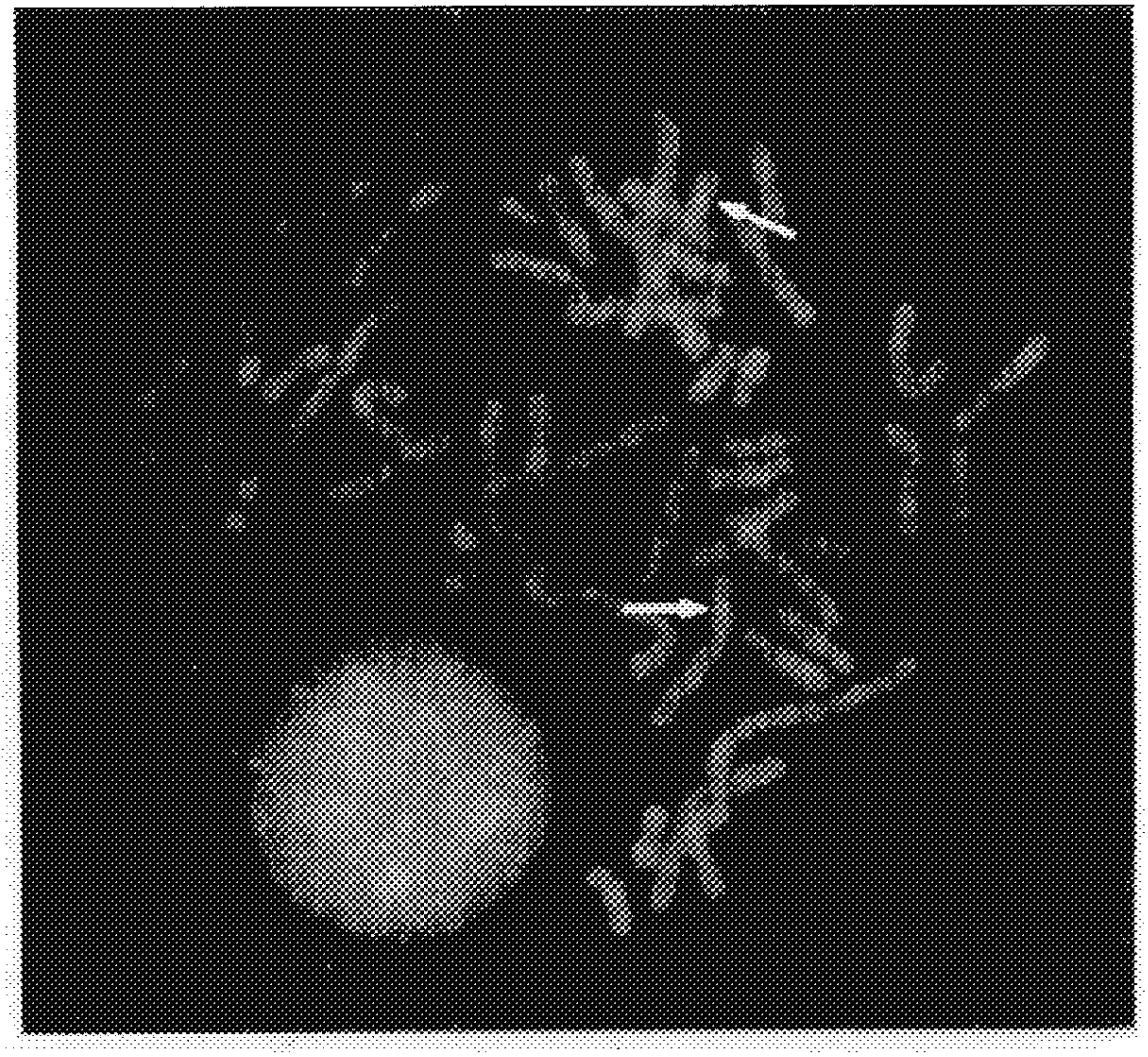


Fig. 5) Direct R-banding fluorescence *in situ* hybridization of cosmid KS72 onto (pro)metaphase chromosomes. Arrows indicate signals on 6p21.3. See FIG. 5 of Sugaya *et al.* (1994) for a photograph with a better quality.

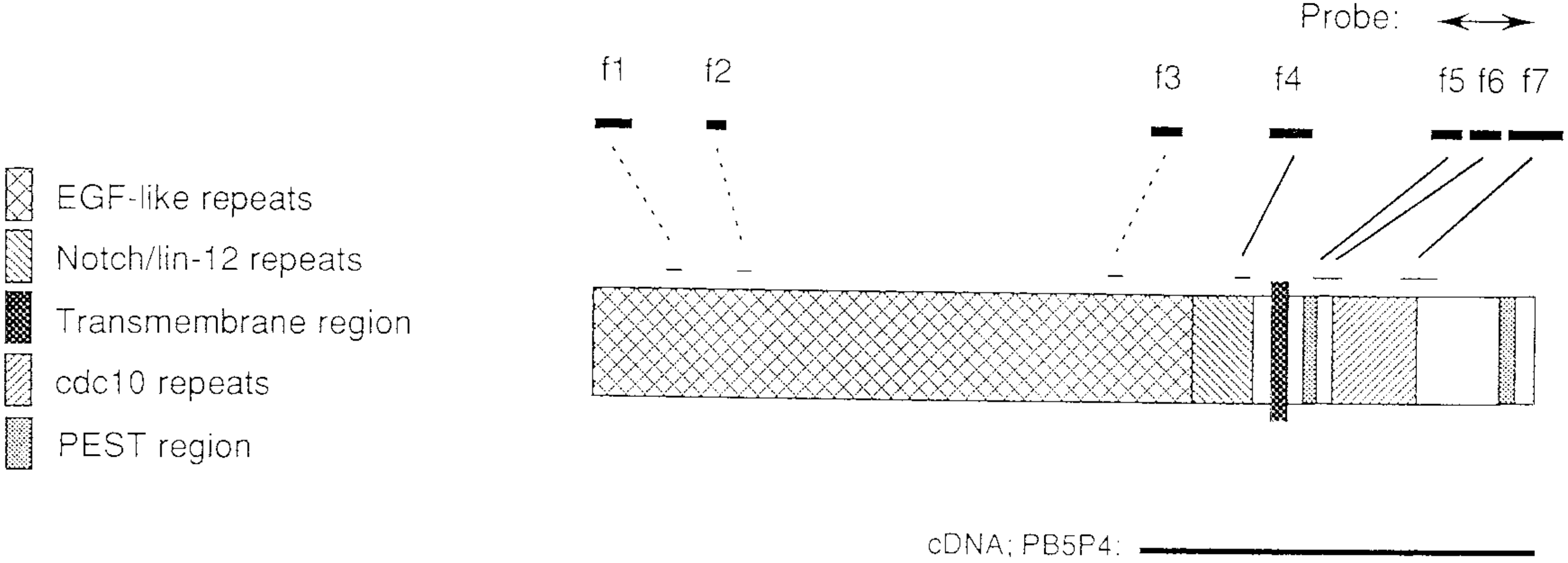


Fig. 6) General organization of Notch-family genes and locations of human NOTCH3 sequences determined. Genomic sequences (f1-f7) are indicated above the schematic representation and cDNA sequences are below the representation. Positions of functional domains showing the highest homology with individual genomic sequences are connected by slant lines. EGF-like repeats occur more than thirty times in Notch genes and positions showing the highest homology with NOTCH3 differ between species, and therefore positions of the highest homology with TAN1 are connected by dashed lines.

```
A
                    ↓
Human NOTCH3          A  E  E  T  G  P  P  S  T  C  Q  L  W
GATCCTTGCTGTTACCCAAGGGCTGAAGAAACAGGCCCACCCTCCACGTGCCAGCTCTGG
                    ************** * ** ****** ********* ***
                    GGCTGAAGAAACAGCCTCAGCCTCCAGGTGCCAGCTTTGG  521
Mouse int-3           A  E  E  T  A  S  A  S  R  C  Q  L  W    133
                    ______________________________----------
                      PEST

 S  L  S  G  G  C  G  A  L  P  Q  A  A  M  L  T  P  P  Q  E
TCTCTGAGTGGTGGCTGTGGGGCGCTCCCTCAGGCAGCCATGCTAACTCCTCCCCAGGAA
 * ** *   *  ******* * ****** ************** ** ***** *****
CCGCTCAACAGCAGCTGTGGAGAGCTCCCCCAGGCAGCCATGCTGACCCCTCCTCAGGAG  581
 P  L  N  S  S  C  G  E  L  P  Q  A  A  M  L  T  P  P  Q  E    153
------------------------------------------------------------

                                        ↓
 S  E  M  E  A  P  D  L  D  T  R  G  P  D
TCTGAGATGGAAGCCCCTGACCTGGACACCCGTGGACCTGGTATGTGAGTCAACCCAGA
* ***   *** *  *  **  ******** *********
TGTGAATCGGAGGTTCTGGATGTGGACACCTGTGGACCTG                      621
 C  E  S  E  V  L  D  V  D  T  C  G  P  D                     167
-------------------------________________
                         CDC10-1
B
                     ↓
Human NOTCH3            G  V  T  P  L  M  S  A  V
ATTACTCTGTCTTACCAACAGATGGGGTGACACCCCTGATGTCAGCAGTT
                     ************************* **
                     ATGGGGTGACACCCCTGATGTCAGCCGTC            650
Mouse int-3             G  V  T  P  L  M  S  A  V             176
                     _____________________________
                     CDC10-1


C
                  ↓
Human NOTCH3       E  Q  T  P  L  F  L  S     A  R  E  G  A
CCTGGCTCTTCTGTACAGGAGCAGACGCCGCTATTCCTGTCG---GCGCGGGAAGGAGCG
                  ** *********** ****** *    *     ********
                  GAACAGACGCCGCTTTTCCTGGCAGTCGTCGTCGAAGGAGCC  1211
Mouse int-3        E  Q  T  P  L  F  L  A  V  V  V  E  G  A     363
                  __________________________________________
                  CDC10-6

 V  E  V  A  Q  L  L  L  G  L  G  A  A  R  E  L  R  D  Q  A
GTGGAAGTAGCCCAGCTACTGCTGGGGCTGGGGGCAGCCCGAGAGCTGCGGGACCAGGCT
***** ** ** *****  ****** *** ***** ***** *  ***** ********
GTGGAGGTGGCGCAGCTGTTGCTGGAGCTCGGGGCGGCCCGGGGACTGCGAGACCAGGCC  1271
 V  E  V  A  Q  L  L  L  E  L  G  A  A  R  G  I  R  D  Q  A     383
________________________________________--------------------

 G  L  A  P  A  D  V  A  H  Q  R  N  H  W  D  L  L  T  L  L
GGGCTAGCGCCGGCGGACGTCGCTCACCAACGTAACCACTGGGATCTGCTGACGCTGCTG
***** ** ** *  ** ** ** * *** ** *  ********* ***** ********
GGGCTGGCCCCAGGAGATGTGGCCCGCCAGCGCAGTCACTGGGACCTGCTAACGCTGCTG  1331
 G  L  A  P  G  D  V  A  R  Q  R  S  H  W  D  L  L  T  L  L     403
------------------------------------------------------------

 E  G  A  G  P  P
GAAGGGGCTGGCCCACCA
*********** **  *
GAAGGGGCTGGACCGACT                                            1349
 E  G  A  G  P  T                                              409
------------------
```

Fig. 7) Alignment of nucleotide and amino acid sequences between human NOTCH3 and mouse *int-3*. Arrows indicate intron-exon junctions. Identical nucleotides are marked by * and amino acids are listed above (for NOTCH3) or bellow (for *int-3*) the nucleotide sequence. Amino acid and cDNA nucleotide positions of *int-3* are listed on the right according to Robbins *et al.* (1992). (A) Genomic sequence is from f5 (0.7 kb) of Fig. 6; for introns, about 20 nt surrounding the exon are presented in this and the following figures. PEST and cdc10/ankyrin sequences are underlined and the inter-domain sequence is by a broken line. CDC10-1 is the first repeat in cdc10/ankyrin repeats of mouse *int-3* (Robbins *et al.*, 1992). (B) Genomic sequence is from f6 (0.9 kb). This CDC10-1 sequence is 3' portion of the first cdc10/ankyrin repeat and continuous from that of (A) at cDNA level, showing CDC10-1 interrupted by an intron. (C) Genomic sequence is from f7 (1.6 kb). CDC10-6 underlined is the last repeat in cdc10/ankyrin repeats. Inter-domain sequence is underlined by a broken line.

A

```
Human NOTCH3         CQSQPCHNHGTCTPKPGGFHCACPPGFVGLRCEGDVDECLDQPCHPTGT
Rat Notch1           SRSPKCFNNGTCVDQVGGYTCTCPPGFVGERCEGDVNECLSNPCDPRGT   1281   61.2%
Mouse Notch1         SRSPKCFNNGTCVDQVGGYTCTCPPGFVGERCEGDVNECLSNPCDPRGT   1281   61.2%
Human TAN1           SRSPKCFNNGTCVDQVGGYSCTCPPGFVGERCEGDVNECLSNPCDARGT   1281   59.2%
Drosophila Notch     CKPGACHNNGSCIDRVGGFECVCQPGFVGARCEGDINECLSNPCSNAGT   1311   57.1%
Xenopus Xotch        TLEPKCFNNGKCIDRVGGYNCICPPGFVGERCEGDVNECLSNPCDSRGT   1281   55.1%
Zebrafish Notch      TGEPRCFNGGRCVDRVGGYGCVCPAGFVGERCEGDVNECLSDPCDPSGS   1279   53.1%
Rat Notch2           AGAPHCLNGGQCVDRIGGYSCRCLPGFAGERCEGDINECLSNPCSSEGS   1278   46.9%
Mouse MotchB         AGGPHCLNGGQCVDRIGGYTCRCLPGFAGERCEGDINECLSNPCSSEGS    961   46.9%
                          * * * *    **  * *  ** * *****  ***  **   *
```

B

```
Human NOTCH3         DQYCHDHFHNGHCEKGCNTAECGWD
Human TAN1           DQYCKDHFSDGHCDQGCNSAECEWD   1557   72.0%
Rat Notch1           DQYCKDHFSDGHCDQGCNSAECEWD   1557   72.0%
Mouse Notch1         DQYCKDHFSDGHCDQGCNSAECEWD   1557   72.0%
Xenopus Xotch        DQYCKDHFQDGHCDQGCNNAECEWD   1556   72.0%
Rat Notch2           DKYCADHFKDNHCDKGCNNEECGWD   1530   68.0%
Mouse MotchB+        DKYCADHFKDNHCDQ             1203   53.3%
Zebrafish Notch      DQYCRDHYADGHCDQGCNNAECEWD   1555   68.0%
Drosophila Notch     DAYCQKHYGDGFCDYGCNNAECSWD   1588   56.0%
                     * **  *     *  ***  ** **
```

C

```
Human NOTCH3         AEETGPPSTCQLWSLSGGCGALPQAAMLTPPQESEMEAPDLDTRGP
Mouse int-3          AEETASASRCQLWPLNSSCGELPQAAMLTPPQECESEVLDVDTCGP   166   67.4%
                     ****   * ****     ** ************ *    * **  **
```

D

```
Human NOTCH3         DGVTPLMSAV
Mouse int-3          DGVTPLMSAV    176  100.0%
Human TAN1           DGFTPLMIAS   1888   70.0%
Rat Notch1           DGFTPLMIAS   1879   70.0%
Mouse Notch1         DGFTPLMIAS   1879   70.0%
Zebrafish Notch      DGFTPLMIAS   1876   70.0%
Xenopus Xotch        DGFTPLMIAS   1885   70.0%
Human Notch hN       DGCTPLMLAS     29   70.0%
Rat Notch2           DGCTPLMLAS   1836   70.0%
Drosophila Notch     CGLTPLMIAA   1909   60.0%
                      * **** *
```

E

```
Human NOTCH3         EQTPLFLS-AREGAVEVAQLLLGLGAARELRDQAGLAPADVAHQRNHWDLLTLLEG----
Mouse int-3          EQTPLFLAVVVEGAVEVAQLLLELGAARGLRDQAGLAPGDVARQRSHWDLLTLLEG----    405
Drosophila Notch     DETPLFLA-AREGSYEACKALLDNFANREITDHMDRLPRDVASERLHHDIVRLLDE-HVP   2141
Zebrafish Notch      EETPLFLA-AREGSYETAKVLLDHLANRDIADHLDQLPRDIAHERMHHDIVRLLEEYNLV   2107
Human Notch hN       EETPLFLA-AREGSYEAAKILLDHFANRDITDHMDRLPRDVARDRMHHDIVRLLDEYNVT    261
Rat Notch2           EETPLFLA-AREGSYEAAKILLDHFANRDITDHMDRLPRDVARDRMHHDIVRLLDEYNVT   2069
Human TAN1           EETPLFLA-AREGSYETAKVLLDHFANRDITDHMDRLPRDIAQERMHHDIVRLLDEYNLV   2119
Rat Notch1           EETPLFLA-AREGSYETAKVLLDHFANRDITDHMDRLPRDIAQERMHHDIVRLLDEYNLV   2109
Xenopus Xotch        EFTSLFLA-AREGSYETAKVLLDHYANRDITDHMDRLPRDIAQERMHHDIVHLLDEYNLV   2116
Mouse Notch1         EETSLFLS-IRRESYETAKVLLDHFANRDITDHMDRLPRDIAQERMHHDIVRLLDEYNLV   2108
                       * ***        *    **   * *   *     * * *  * * *   **

Human NOTCH3         AGPP--EARHKATPGREVGPFPRART----VSVSV
Mouse int-3          AGPTTQEARAHARTTPGGGA---ARA----AGRCL    433   60.7%
Drosophila Notch     RSPQMLAMTPQAMIGSPPP---------GQQQPQL   2167   35.7%
Zebrafish Notch      RSPPLP---------LSPPLCCPN-TYLGIKPSPG   2132   32.1%
Human Notch hN       PSPP--GTVL--TSALSPVICGPNRSFLS------    286   32.1%
Rat Notch2           PSPP--GTVL--TSALSPVLCGPNRSFLS------   2094   32.1%
Human TAN1           RSPQLHGAPLGGTPTLSPPLCSPN-G---------   2144   29.8%
Rat Notch1           RSPQLHGTALGGTPTLSPTLCSPN-G---------   2134   29.8%
Xenopus Xotch        KSPTLHNGPLGAT-TLSPPICSPN-GY--------   2141   28.6%
Mouse Notch1         RSPQLHGTALGGTPTLSPTLCSPN-G---------   2133   26.2%
                       *
```

Fig. 8) Comparison of amino acid sequences with Notch-homologous proteins found by searching PIR. Identity levels (%) with human NOTCH3 and amino acid positions of reported sequences are indicated on the right. Multiple alignment of amino acid sequences was conducted by MALIGN program (Hein, 1990) of DDBJ. Identical amino acids are marked by *. (A) Human NOTCH3 sequence is from f3 region of Fig. 6 and thus in EGF-like repeats. (B) NOTCH3 sequence is from f4 and in Notch/lin-12 repeats. + Motch B has been partially sequenced. (C) Sequence is from f5 and that listed in Fig. 7A. (D) Sequence is from f6 and listed in Fig. 7B. (E) Sequence is from f7 and listed in Fig. 7C.

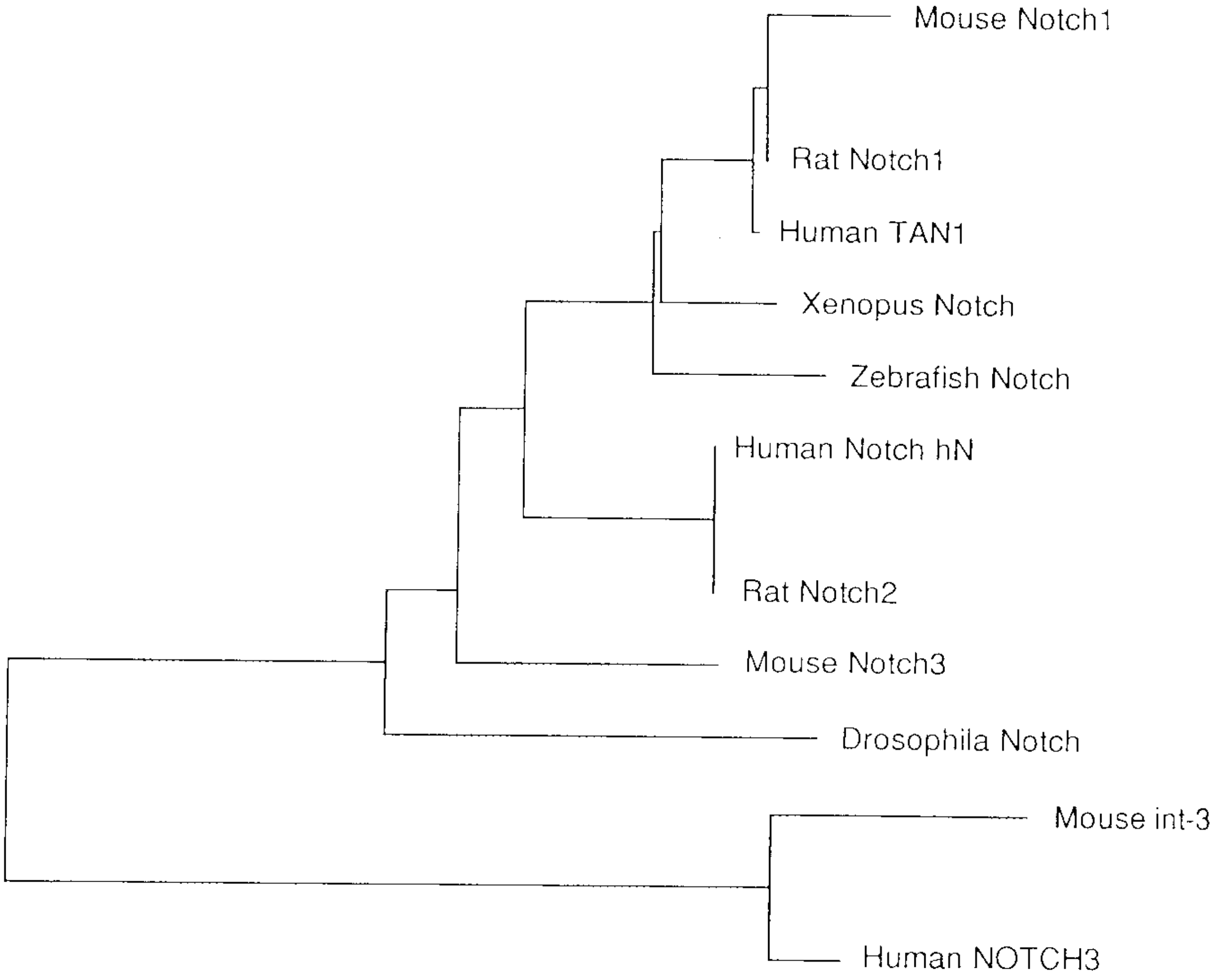


Fig. 9) Phylogenetic tree of intracellular domain of Notch genes. Amino acid sequences translated from cDNA sequences were analyzed using N-J method (Saitou and Nei, 1987). The branch lengths indicate the evolutionary distance between the different sequences.The Human NOTCH3 is not the human counterpart of the mouse Notch3 recently reported by Larsson *et al.* (1994).

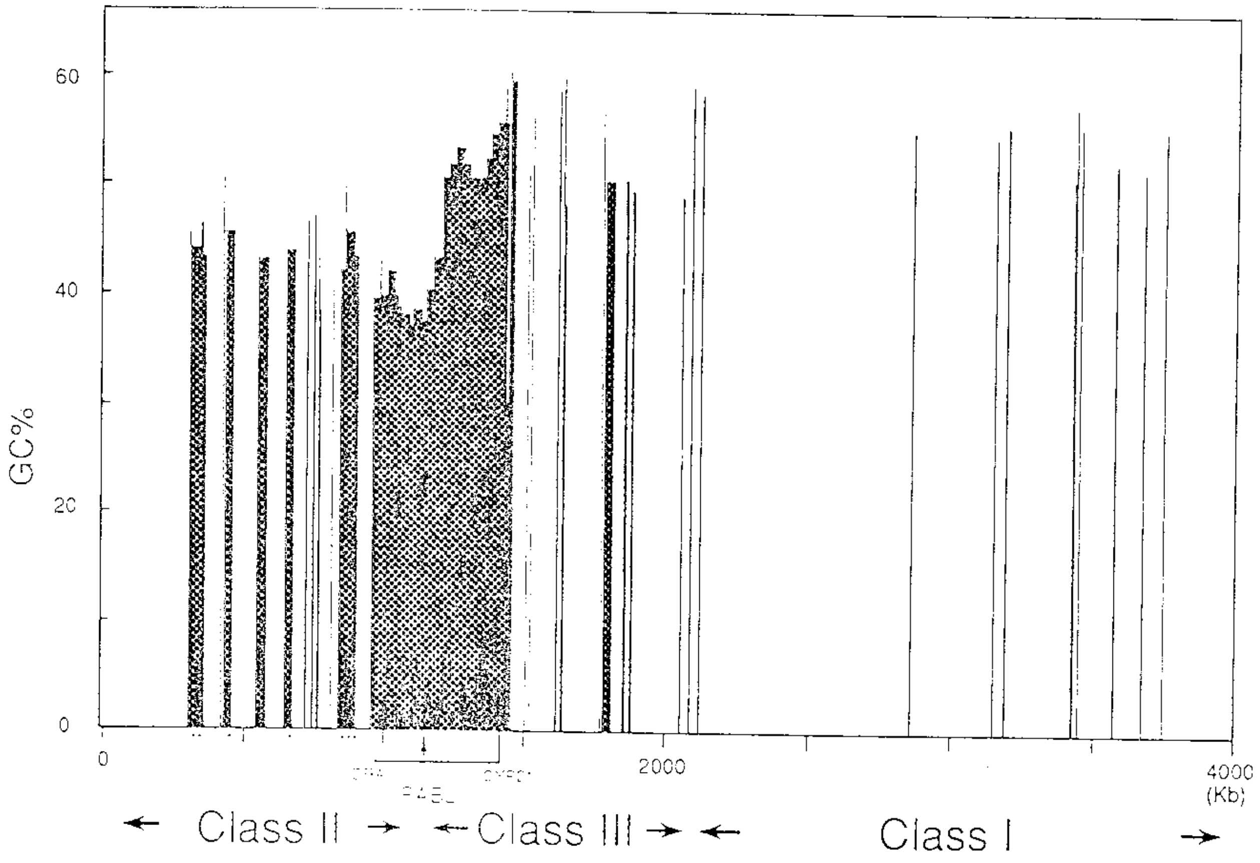


Fig. 10) Base-compositional map of the human MHC locus. Genomic MHC sequences in GenBank longer than 3 kb were selected, and their GC% was arranged by genetic position. GC% distribution between HLA-DRA and CYP21 (about 450 kb) was measured directly by GC contents of cloned fragments with the biochemical method described in Materials and Methods. Seven thick vertical bars in class II marked by * at the bottom correspond to GC% of previously isolated clones (Kawai *et al.* 1989), also measured biochemically. Gene-encoding regions are known to be often GC-richer than their flanks. This produces local GC% fluctuations within an isochores and a tendency for thinner bars (usually corresponding to gene sequences) to be GC-richer than thicker bars (corresponding to both genes and their long flanks). It should be noted that even focusing only on thin bars, GC% levels of classes I and III are higher than those of class II supporting long-range GC% mosaic structures (Ikemura *et al.*, 1990). This figure was from Fukagawa, Sugaya, et al. (1995).

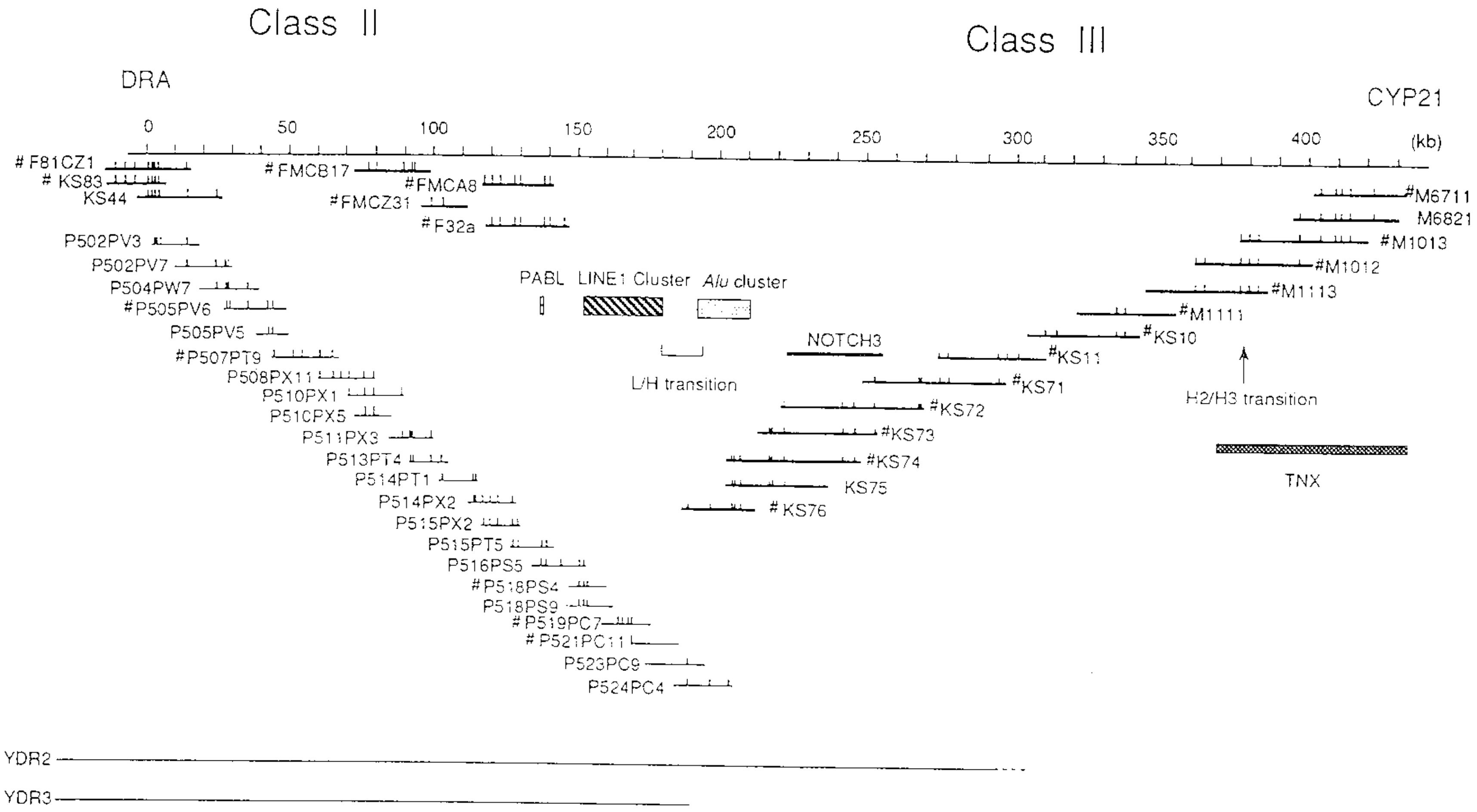


Fig. 11) Molecular map of contiguous cosmid, λ phage and YAC clones that cover junction of MHC classes II and III. Ordered clones are represented by horizontal lines with vertical bars indicating *Eco*RI sites; cosmid clones (M, KS, F series) are indicated by thicker horizontal lines than λ phage (P series) and YAC (YDR2 and YDR3) clones; the terminus of YDR2 has not been identified. The two YAC clones were used for the library construction to avoid artifacts caused by possible YAC chimerism; λ phage walking was done using these two independent libraries, examining mutual consistency by restriction maps. Cosmids of KS-series were cloned by me, those of M-series were by Dr. Matsumoto (Matsumoto *et al.*, 1992), and clone of F-series, as well as λ phage clones, were by Fukagawa (Fukagawa, Sugaya, *et al.*, 1995). After completion of λ phage walking, cosmid clones (F-series) were isolated using phage clones as probes; the region where we first encountered difficulty in cosmid walking could not be cloned even by this procedure. PABL1 was found in P515PT5. LINE-1 cluster was partially sequenced and at least five independent LINE-1 repeats were found. Clones used for GC% measurements are marked by #. This figure was taken from Fukagawa *et al.* (1995).

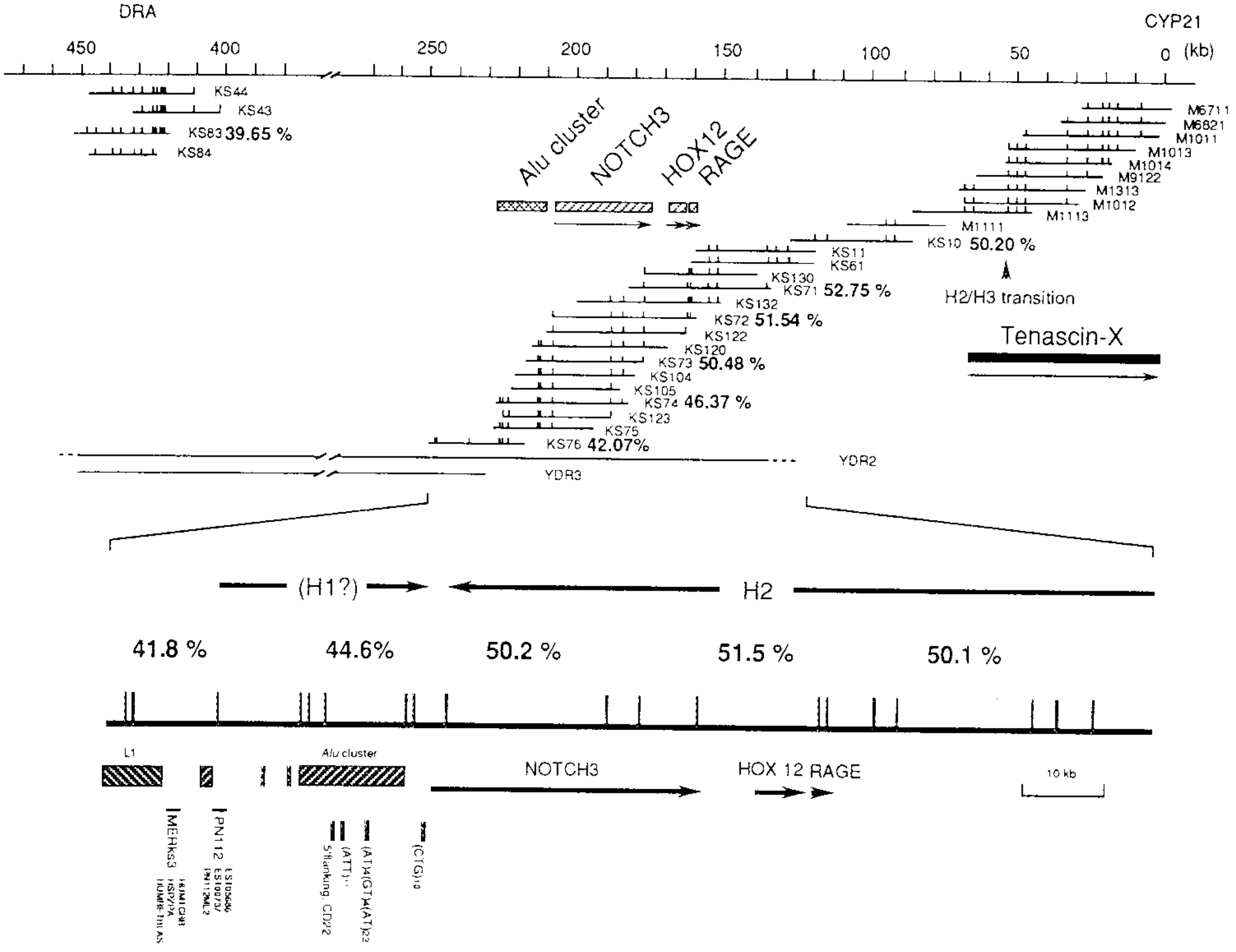
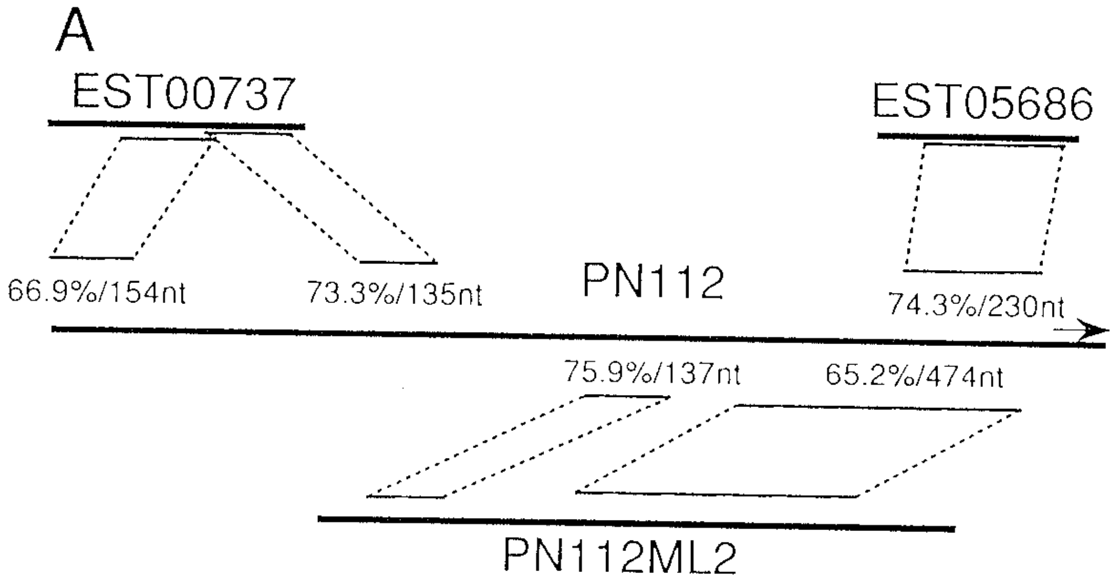


Fig. 12) GC% distribution and molecular map of the 5' upstream region of NOTCH3, that is near a boundary of the long-range GC% mosaic. Redrawn Fig. 1 with detail organization of the 5' upstream region of NOTCH3. GC% of cosmid inserts is presented right side, and that of EcoRI fragments is above the respective portions of the horizontal line. Characteristic genome structures are indicated by small black boxes and horizontal line. Isochore is indicated above the GC% of *Eco*RI fragments with arrows. Hatched boxes indicate *Alu* cluster and LINE1 sequences.



B

```
CTGCAGGTCC TTGGANCAMA AACTGGATGA GATTTCCCNG GTNTTGTTTT ATGTGGNTGA
GAGNNTGACW TTGTAACCAT GTNGGGGGAC TCTCTCTCTT GATCTCTGCC ATCTGGAGGG
ATGGAAATTC TTGGGNTCAA NTCAGGTGGC TGGTCTGAGA GGRNCTGGGA GTCTGAGACA
CATTAGCATA CTCTTCGTCC TGAATGTGTC GAGCCCTTAG GTGAGTTTTG TCTTAAAACG
TCCCATCTCT GCTGGTTGGA TTTATTAGGA CAAAAAAACA GTCCTATCTC TACAGGACTT
TTGTTGTATT TTGCTATCTT AAACCCATTT CCAAGAGGGA ATACTTGGGG GATGCCTCCT
CTAGGAATAC TTCTTGCTGC TTATATGCCA AAAACCTGGA AAATTACCAT CTGCAATTTA
AAAAAGGTGT TTGAGTCTCT ATTGGAACTA AGTACACCAT TGAAAGAAAA AGGATTTTAG
AGATCTCTTA TCTAAAACAA TTGAAGGAAG GTTAAACAGT AGTGTCGTGG GTAGCCTTAA
AAATTCTCTT GAGCAGTTAA AATCATTCGC AAGCTTGAAA ATGACTGCTC TAGATTCTTT
CTGGGAAGAG CACTGGCAAC CACCCTATGC TGTAGCACAG TAGCTAAATC TCTGCCCTTT
CACTATGGTG GCCTGGGATC ACTTCCCAGC TTAGGGGAAT GCGTCCTTTC TGGTTTTGTA
TTTGTGGGAC TTTTTGCCAT TCATTGATGG ACAGCTTCTG ATTTCCTGTC TTGAATTTTC
CTTGCTCTGA GATACCTTTG GGGTGATTCT AGATCTTGTA AAAAACTGCT TGGCATCTCT
TTGGAGATAC CTTGTGCATC TGTGGTTAAG TCATAACCCT AGTTAAGGCT CATTGGTTTC
AGGTGGGAGG TTATCCTTGG TAGAGAGTTC AAAAGCCAGA AATATCAGCT GTTTGTTCCA
GCTAAAAACT GGTAATAAGA GATCTGAAAG AATTTTCTTC AAGAGCTCTA TAGTTAAAAG
TCAACTTAAT TAAAACTGAT TTAGAATATA TGTGTACAGA TATTGTTTTA AAGCCTCTGC
TCTCTCTCTG TAAAAACATC TTAAGCAACT GAATTCTGTC TGCTTAAATT TTAATCTTGG
TATGTAAAAG CTAGGAAAGA AATATACTTT TAGAGATGGC TATTGACGTT GTTTACAGTG
AATAGTTATT ACTACAGGGT GGTACTGCTT TTTTTTTGCA CATTTAGATA AGAAAAGCAT
GCTTTCGGGC ACCTAGAAGG TATGGAATGA GGGTTAAGAC TCCCATGGAG CATTAAGTGA
TTACAGAATA GGCTGATTGC TATAGGGTTG CCCACCAGCC TCAGGGGAAT GTCCTTGCAG
TGAAGTGCAC CGTAAAAGCA TTGCACTGTC TTGTCCTGCG GTGTTCTCCT CTCTTGAGGA
CCCAGGATTC AGTGTAAAAG TTGGATCCTT AACTTTGGAG ATCTATTTTG CCTTCCAGCT
GTGCCTGCTT ATTAGGCCAT AGAAACTGCA TGCTTTCCTG GCCCTGTTCC TTAAAGGGCT
CCACCCTAAA GCCAGTAATT CAATTAAGAA ACTAACATCT TTAAAAAAAT TTCAGTGGGC
AAGTGTGTCT GTTTTCCTAA CCATCTTTTT TCTTTTTCTT CTTTTGAGAC AGGGTCTTGC
TCTGTCACCC AGGCTGGAAT GCAGTGGTGA GAACATAGCT TACCTGCAG
```

Fig. 13 continued.

C

```
GAATTCCGTG ATATATGTGT GACCTTCACC ACCTGTTAAT TCTCTTCTCA TCCATGAACC
ATCTTGAATT TTTCTTTCTC TGAGCACCTG GGTGGTTACT TTTGGTAAAA TTTAAAAGCC
           1
AGAAATATTG GACTTTTTGT CTGGCTGAAG TTAGGTAATA AGAAATTTGA AAGAATTTTT
TTTTAAGAGC ACTATGGTTA AAAGTCAGCT TAATTAAAAG CAGCTATTCA AGCTCTAACA
GCCTGGAACT CCTTGGAAAA AAACAGAGGA GGTGGCATAG ACCCTGTTTT GGGAAAAACA
TCTGCTTTCC TCATGAAACC TCAAGAATTG AAAGTGGATA GATCTCTCTC AAAAATCTAA
                                           2
GGCTCTGATC TTGTTTSGCA TGCATTATCT GATGTTTTTG ACTTTTGGGG ATATCAAGAA
ATTACTTTGC ATTATGAAAG AACTTTGGTG TGTAATAACT ACTACGTAGG AAATATACTT
TTGGGGATAG CTAGTGGCAA TTATGGGGAA ATACATGGCT TTTTGCACGT TTGGATCAGA
GAAACATGCT CTTGGCCAAC TTGGAAGGTA TGAAGATATT CCCACTCTCT CACTGAGAGA
TAAGACTGTC ATGGGGGGAT CAGCTAATCA CAGAATGGGC TTTGGGTTAT TTTGTAATGA
AATGCATGGT AAAATCATGT CACTGTCTTG TTCTTGTAGC ATTTCTCTTT TTGGGGATCT
                                       3
AGGATCTTGA TATAAAAATG GGACCCTTAA TTTTTGGGAT CTGTTTTGTC TTCCAGCTGT
GCCTGCTCAT TGGGCTGTGG AAACTGCATG CTTTCCTGGC CCTGTTCCTC TAACAGCTCC
ACCCTGAAGC CAATAATCCA ATTAAGAAAC TGGGAAATGA AAAATCTTAC AACTATTGGA
TCTTCTGCTG ACAAAACCTG GGAATTTGGG TTGTGTATGC GAATGTTTCA GTGCCTCAGA
   4
CAAATGTGTA TTTAACTTAT GTAAAAGATA AGTCTGGAAA TAAATGTCTG TTTATTTTTG
TACTATTAAA AAAACGGAAT TC
```

Fig. 13) A genomic 1.7 kb sequence near the "L/H" boundary showing evident homology with three ESTs. (A) Three ESTs are indicated above and under a line representing the genomic sequence PN112 (1729 nt). *Above* , EST00737 (401 nt; GenBank HSXT00737) and EST05686 (323 nt; HS7963): *Under*, PN112ML2 (1042 nt). Homologous regions are indicated by thin horizontal lines with nucleotide identity scores and connected by dashed lines. *Alu* element is indicated by an arrow. (B) The location of the 1.7 kb genomic segment (PN112) is shown in Fig. 12. Homologous sequences are underlined, and *Alu* element is in italics. (C) Nucleotide sequence of PN112ML2. Portions homologous with known sequences were underlined and numbered: No. 1 shows homology with a sequence near tandem GT repeat; No. 2, a STS of human chromosome 4; No.3, EST05686; No. 4, 3' UTR of prostaglandin-endoperoxide synthase 2 gene.

A

```
GATCTTGGCT CACTGCAACC TCTGCCTCCC AGGTTCAAGT GATTCTCCTG CCTCAGCCTC
CAAAGTAGCT GGAATTACAG GCATCTGCCA CCATGCCTGG CTAACTTTTT GTATTTTTAG
TAGAGACAGG GTTTCACCAT GTTGGCCAGG ATGGTCTTGA TCTCTGACCT CATGAGCCGG
CCACCTCCGC CTCTCCAAGT GCTGGGATTA CAGGTGTGAA CCACCTCACC TGGCCTGATT
CTTCTTTATA TCTTCCATTT CTTTGCCCAA ATTTCTGGTT TTCATTTGTT TCAAGAGAGT
TTGTAAATGC TTGTTAAATG TTGTTTTTTT TTTTTCCTTT TCTTTTTGAG ACAAGGTCTT
GCTCTGTTGC CCAGGCTGAA ATGCAATCAT GGCTCACTGC AGCCTTGACC TCCTAGGCTC
AAGTGATCCT CCCACCTCAG CCTTCAAGTA GCTGGTACCA CAAGTACACA CCACCATGTC
TGGCTAATTA AAAACATTTT TTTTTTCCAA GGGGCTGGGA CCACAAGTAC ACACTACCAT
TCCTGGGTAA TTATTATTAT TATTATTATT ATTATTATTA TTTTGGTGGT GTTTTTTGGA
GAGACAGATT TCCCTATGTT GGTGGGCAGA ACTTCCGGGT CAAG
```

B

```
GATCAACAAA CAGCAGGGCT GGGACTGCCC AGGGGGTTCC GAGATTCCTT CTCCCCTCCT
ATCACCTGCC CTCCAGGCAC ACCGTCCTAC TTCCCCCTAC TTCCCCAGGG GTTGTCAGGG
ACAGAAGGCC CCTCCTTCAT CCCCCCTAGT GTTCCTCCAC TCTTCCTCCG CCCCCCATTA
CTAGGGTGTC CAGGACATTG TGTGACTCAG GAAACAGCTC AGACGTGAGG CTTGCAGCAG
GCCGAGGAGG AAGAAGAGGG GCAGTGGGAG CAGAGGAGGT GGCTCCTGCC CCAGTGAGAG
CTCTGAGGGT CCCTGCCTGA AGAGGGACAG GGTACCGGGG CTTGGAGAAG GGGCTGTGGA
ATGCAGCCCC CTTCACTGCT GCTGCTGCTG CTGCTGCTGC TGCTGCTATG TGTCTCAGTG
GTCAGACCCA GAGGTGAGGC ATGGCGTGGG TGAGGTGAGG GGACCCAGCT CCCTTAGGAG
GATGATCAGT GGGGTGGGGG AAAGAGGGCC AAGCCCCAGG CCGTGTGAGG GATGCTGGAT
GGAGGAGATT CTCACTGCCC AAATAGAGAC GGCCTCCAGG GAAAGACGGC TCTGCCCATG
GAGCTGCTTT GGGCCTGGTG CCAGGGGTGG TGACTGCTGG GGGATGGGTG AGAGGGTGCC
CACCTCCAGG AAGAACCTCG TCAGCACTGG CACTGGAGGA CTCTTGCAGC CATAGGGAAG
AGGGGAAGAG GGAACACACT GACCACCTGC TTGGGGAGGA GATGAGAGGG AAGCAGGAGA
TGGGGACATG AAAGGTCAGG CCTACTAAGC CCTTTTCTTA GTCCAGCTTG TCCCCACCCC
CCCCGGATTG GCTCAATGCT TCGGCCTTTC CGGGAGGAAA TCTCTTCCGA AGTCTTCAGC
CATTCAACCT CCCCGGGAGG CAAC
```

C

```
GATCATGCCA CTGCACTCCA GCCTGGGTGA CAGAGCAAGA CTCCGTCTCA AAAAAAAAAA
ATCTGCTGAG CACCTACTTN GTGTGGCTAC TGTTCCAGGC CCTGGGGGAA ACACAAAGCA
AAAGAGATAA AGCAACTGCT CTCGTAGAGC TTTCATNCTA AAGAAAGACA GAAAATAAGT
AAGTTACAGA AAGAATATAT ATGTGTGTGT ATATATATAT ATATATATAT ATATATATAT
ATATATATAT ATATCTCCAA CTAGATATAT AGATGCATAT ATCTAG
```

Fig. 14) Trinucleotide and dinucleotide microsatellite repeats on the 5' upstream region of NOTCH3. (A) (ATT)$_{11}$, (B) (CTG)$_{10}$, and (C) (AT)$_{4}$(GT)$_{4}$(AT)$_{22}$ repeats are underlined. Primer sites are indicated as arrows.

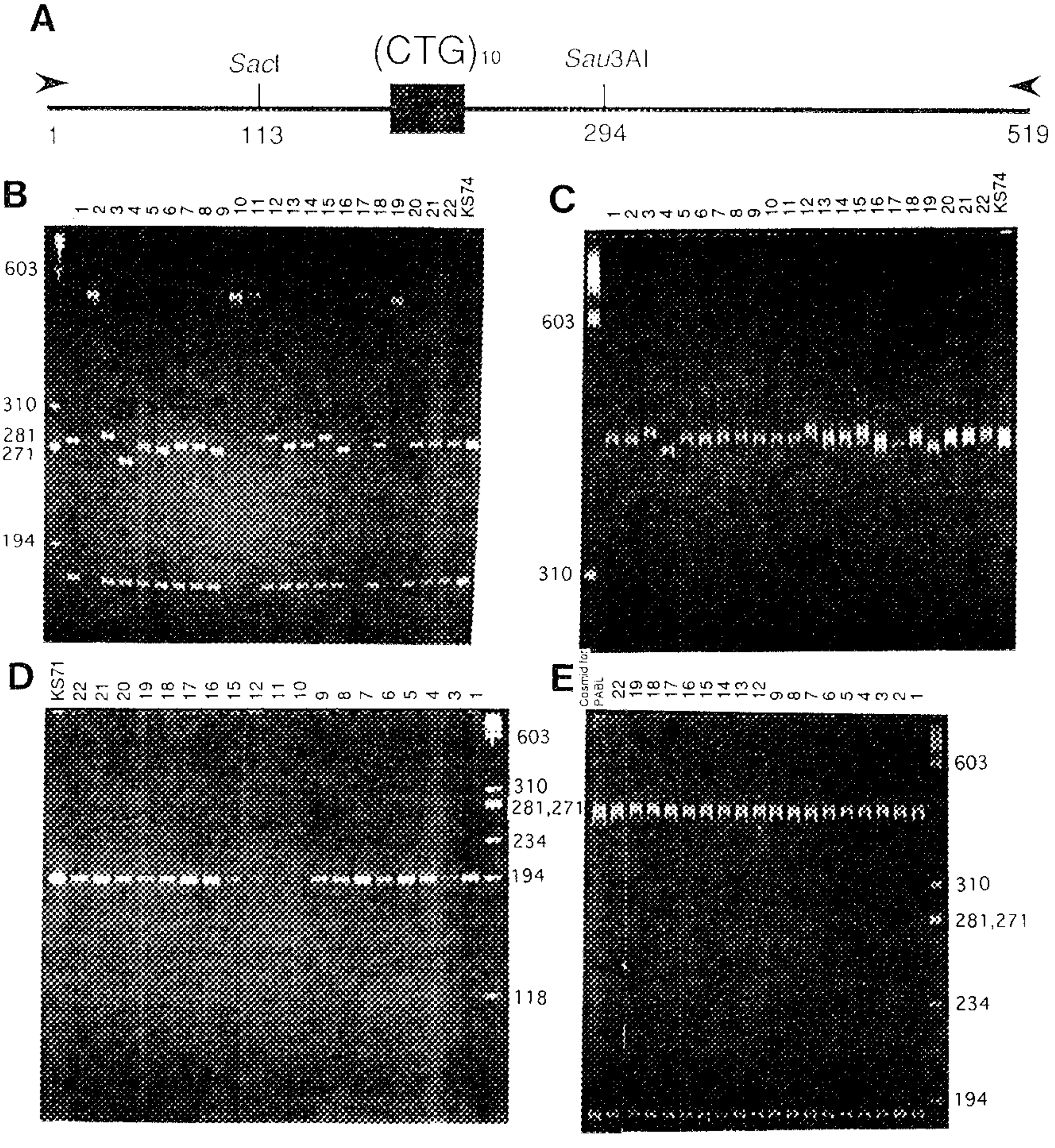
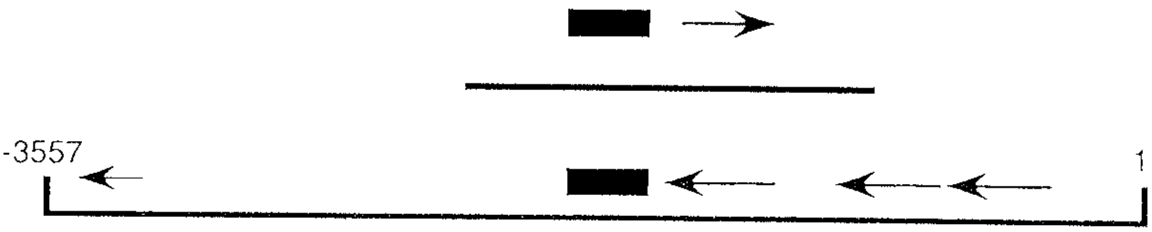


Fig. 15) Analysis of microsatellite alleles of (CTG)10 repeat region. PCR reactions were carried out on genomic DNAs from 22 cell lines; 1. MGAR, 2. OMW, 3. COX, 4. MADURA, 5. OLGA, 6. RSH, 7. BSM, 8. SAVC, 9. D0208915, 10. MANIKA, 11. BM15, 12. AMA1, 13. PE117, 14. DEU, 15. PF04015, 16. SP0010, 17. BM92, 18. BTB, 19. LUY, 20. HO104, 21. QBL, 22. AKIBA. (A) The restriction enzyme map of a chromosomal region flanking of the (CTG)10 repeat, which sequence was determined by subcloning of cosmid clone KS74, are indicated. Vertical bars and numbers indicate *Sau*3AI and *Sac*I sites. Oligonucleotide primers corresponding to chromosomal regions flanking of the (CTG)10 repeat region are indicated as arrows, and their sequences are indicated in Fig. 14. PCR reaction condition are described in Materials and Methods. After PCR reaction, products were extracted by phenol/chloroform, and were digested with *Sau*3AI (B) or *Sac*I (C). These products were analyzed by electrophoresis on a 8%(w/v) polyacrylamide gel. Allele analysis of RAGE exon 3 region (D), and of PABL (E). Oligonucleotide primers corresponding to these chromosomal regions were designed as follows; RAGE: 5'-AAC ACA GGC CGG ACA GAA GC-3', 5'-GGT AGA CAC GGA CTC GGT AG-3'; PABL: 5'-CTT ACT CTT TGC CTT AC-3', 5'-ATG CAG ATG AAC TGA GA-3'.

A

CTGACACACC CTGNCTGAGA GAGTAAGAAN CACCTCATTC CTGTTCCCCA CGGGGTCKCC
AGTGACACTG GGTTGTGATG GAGGGGGATT TCATACCACC CNVSRNGGGG GCTGAGAGTC
CCAGCTTCCT ACTTGCTGGG GNNGGGGTNN CCTCANCTGG GTGGGAGTTG ATGGCTAAAC
TCCCCACTCA TCCTTTATTG GCAGATATGG GGGTGAGAGT GTTTTTTTTT TTTTTTTTGC
CTGAAATAGA GTAGTCATTG TCTAAAAGTT TTGTCTTTCC AGGATGTTCC TTTGTTGGTC
CTTTGGCCAG AGACTTTCCG GGATTTTTTT CATCTTCCTG TTGGAGTTTC CTGGTTGCTG
GCTTTTCCAG CACCCAGTCT TGTATATATG AGGCAAAAGC CAAACCCAGG GAACTCACCA
CTATATTGTT TTTTCAGGTC TTGAGATTCC TAGCCAGTCT GCCTTCTCTG CATCTTTCAG
GATCTTCTTA TGTTTGTTTT ATATATACCA TCCAGGATTG TAGCTGTATT TAGCAGGAGG
AATCAGAAGA GCATCTACGT CATCTTGTCT TGGAGCTTGA AGGCAGCTGG TTAAGTCCTT
AAAACATTCA ACACAGATTT TCCATATGAC TCAGCAATTG GGTTCCTAGG TATCTACCTA
AGAAAAATGA AAGCAG*GCCG GGTGTGGTGG CTCATGCCTG TAATCCCAGG AATTTGGGAG*
GCCGAGGTGG GCGGATCACC TGAGGTCAGG AGTTTGAGAC CAGCCTGACC AACATGGAGA
AACCCCATCT CTACTAAAAA TACAAAAATT AGCTGGGCAT GGTGGTGCAT GCCTGTAATC
CCAGCTACTT GGGAGGCTGA GGCAGGAGAA TCACTTGAAC CCAGGAGGCG GAGGTTGCGG
TGAGCTGAGA TTGCGCTGTT GCACTCCAGC ATGGGCAACA AGAGCAAAAC TCTGTCTCAA
AAAAAAAAAA AAAAAAAARG AARGAAAAWT GAAAGCGTWT NGTCCACACA AATACTKGTN
TAAGAATTCA TAGCAGNGTT ATTCACAATA GGTNTGANGT AAAACAACCA ATTTTCCATG
GTCCCCGGGT ACCGAGCTCG ATTCCGGATC ATGGGCATGN TNTTCCTGNG TGAATTGTAT
CCNTCANATC CCAAAAAATT CGNGCCGGAG GATAAGTNTA AGCCCGGGNG CCATGNTGGN
TTATCNANTA ATGGGTGGGC CCNTGCCNTT CNGCNGAACC NTGGCNNTTA ATTATNNCNC
ACCNGGGGGG GTGGTTGGGT NCCTTCC

B



CD22 promoter region

Fig. 16) A genomic sequence highly homologous with the 5' flanking region of CD22 gene. (A) Genomic sequence homologous with the 5' flank of CD22 gene is underlined, and *Alu* elements is in italics. (B) The sequenced genomic fragment is indicated as a horizontal line above the line showing the CD22 promoter region; the number 1 corresponds to the translation start site of CD22 gene. The homologous regions are indicated by black boxes above the individual lines, and *Alu* elements are by arrows.

```
AACATTTACT  ACTGTAGATT  GCCAGGCCAA  GATGGCTGAT  TAGAAACAGC  TATGGTCCAC
AGCACTCACA  GAGAGGAACA  AAAGAGGCAA  GTGAATACAG  CATCTTCAAC  TGAAATATCC
AGGTACTTGC  ATTGGGACTC  ATCAGGAAAA  CAACTCGACC  CACAGAGAAC  AAAGAAAAGC
TGGATGGGGC  GACAGCCCCC  CTGGGAGCGA  CACAGAGCCA  AAGGAACCCC  CACTCCCAGC
CAAGGGAAGA  AGTGAGTGAT  GGTGCGACCT  CAGGAAACCA  TGCTTCTCCC  ATGGATCTTT
GCGACTGGTG  GATCAGGAGA  TTCCCTCATG  AGCCCATGCC  ACCAAGGCCT  TGGGTCCGAC
ACACACACAG  CTGTGTGGAG  TCTTGGCAGA  GCAGCTGCTC  AGACACACAC  AGAGACCCAC
GAGCTTTACA  TACTCTGGCC  CAGGGCGGAG  CGCAGGAAGG  GCCCCAAAAC  AGATC
```

Fig. 17) A sequence highly homologous with several reported sequences. Sequence homologous with a segment between TCRBV8S2 and TCRBV8S3 of V region of human T-cell receptor beta locus is underlined.

TABLE 1

Exon-Intron Organization of the RAGE Gene

Exon			Sequence at exon-intron junction and intron size (nt)		
No	position	(nt)	5' splice donor	(nt)	3' splice acceptor
1	<6665-6716	>52	CTG TGG Ggt gag cca	183	ctc cca gGG GCA GTA
2	6900-7006	107	TGG AAA CTG gta agc	130	ttc tag AAC ACA GGC
3	7137-7332	196	GTC TAC Cgt aag aat	166	cct tca gAG ATT CCT
4	7499-7563	65	CCC AAT AAG gta gtg	122	ccc cag GTG GGG ACA
5	7686-7773	88	GAG AAG Ggt gag tcc	90	atc ata gGA GTA TCT
6	7864-8046	183	GTC TGG Ggt gag cat	142	cca gAG CCT GTG CCT
7	8189-8319	131	ATG AAG GAT gtg agt	179	cac cag GGT GTG CCC
8	8499-8640	142	ATC ATC Ggt gag acc	615	ctt cca gAA CCA GGC
9	9256-9282	27	ACT GCA Ggt gag ggg	128	cgt aca gGC TCT GTG
10	9411-9537	127	GAG GAG AGg tga gtg	111	ctc ctc agG AAG GCC
11	9649-9920	272			

Transcription start site has not been determined.

TABLE 2

Exon-Intron Organization of the HOX12 Gene

Exon			Sequence at exon-intron junction and intron size (nt)		
No	position	(nt)	5' splice donor	(nt)	3' splice acceptor
1	<707-1198	>492	CAG GCC AAg tga gtg	901	ctc cca agG AAA CAC
2	2100-2173	74	AAA ACT Ggt atg tgg	218	atg cta gGC CTC AGC
3	2392-2639	248	TAT GAG CAG gta agg	100	cta cag GCA TGT AAT
4	2740-2930	191	GAT GCC AGg tgg gcc	183	ctc tgt agA CGA AAG
5	3114-3249	136	GTG TCT CAG gta tta	251	cca cag GTC TCC AAC
6	3501-3654	154	TCT GCA Ggt gga tcc	336	cca caa gGC TCT GGC
7	3991-4079	89	GCT TCC CAG gtc aga	122	ctg cag GTG GAA TCA
8	4202-4288	87	GAA ATG AGG gtg agt	129	ttt cag GCA AAT GGC
9	4418-6146	1729			

Transcription start site has not been determined.

TABLE 3

Properties of allele analyzed cell lines

No	NAME	A	C	B	B4/6	Bf	C2	C4A	C4B	DW	DR	DW2	DRW52/53	DQ	DP	GLO	(CTG)n/Sau3AI	HLA DISEASE
1	MGAR	26	7	8	6	S				2	15			6	4	1,2	F2	
2	OMW	2	-	45	6					18	13	24	52	6	1	2	S	
3	COX	1	7	8	6	S	C	Q0	1	3	3	24	52	2	3	2	F4	
4	MADURA	2	10	60	6					8	8		52	4	4		F1	21 HYDROXYLASE DEFICIENT
5	OLGA	31	1,10	62	6					8	8		52	4	3		F2,F3	
6	RSH	68,30	2	42	6	F	C	1	Q0	N	3	24	52	4	1	1	F2	
7	BSM	2	9	62	6	S	C	3	3	4	4		53	8	2		F3	
8	SAVC	3	7	7	6					4	4		53	8	-	1	F3	
9	D0208915	25	-	18	6	S	Q0	4	2	2	15			6	2,4		F2	C2 DEFICIENT
10	MANIKA	3	-	50	6						7		53	2?3	NT		S	
11	BM15	1	7	49	4					5	11	25	52	7	3		S	
12	AMA1	28	4	53	4					2	15			6	-		F4	
13	PE117	24	10	60	6					14	4		53	8	4		F3	
14	DEU	31	4	35	6					4	4		53	7	4		F3	
15	PF04015	1		8	6						3	24	52	2	1,4		F4	
16	SP0010	2	5	44	4					DB2	11	25	52	5	2		F2	
17	BM92	25	1	51	4					14	4		53	8	-		S	
18	BTB	2	1	27	4					8	8		52	4	4		F3	Mo BECHTERW
19	LUY	2	-	51	4					8	8		52	7	1,4		S	
20	HO104	3	7	7	6	S	C	3	1		15			6	4		F3	
21	QBL	26	75	18	6					3	3	25	52	2	2		F3	
22	AKIBA	24	63	52						12	2			1			F3	
23	KS74(cosmid)																F3	

Note. The column "$(CTG)_n$/*Sau3*AI" lists the polymorphism in the $(CTG)_n$ region shown in Fig. 15. F1 indicates the shortest allele; F2, the second; F3, the third; F4, the longest one; S, the RLFP of *Sau3*AI. The allele of OLGA is heterozygous.

TABLE 4

Gross similarity of genes on 6p21.3 and those on 9q33-q34

6p21.3		chromosome 9	
Gene/Locus	Physical Location	Gene/Locus	Physical Location
VARS2	6p21.3	VARS1	9
HSPA1, HSPA1L	6p21.3	GRP78	9q33-34.1
C2, C4A, C4B	6p21.3	C5	9q33
TNX	6p21.3	HXB	9q32-34
HOX12 (PBX2)	6p21.3	PBX3	9q33-34
NOTCH3 (INT3)	6p21.3	TAN1 (NOTCH1)	9q34.3
COL11A2	6p21.3	COL5A1	9q34.2-34.3
RXRB	6p21.3	RXRA	9q34

VIII. Acknowledgments

I wish to express my deep gratitude to Prof. Toshimichi Ikemura for guidance. I thank Dr Naruya Saitou for guidance of evolutionary computation and consideration, and also thank Dr Ken-ichi Matsumoto, Dr Toyoaki Tenzen, Tatsuo Fukagawa, Yasukazu Nakamura, and other members in the Division of Evolutionary Genetics for their encouragements and helpful suggestions. I also thank Drs Hidetoshi Inoko, Asako Ando, Katsuzumi Okumura, Kazuei Mita and Ei-ichi Takahashi for collaboration and discussions, and Mrs N. Ishihara and Y. Miyauchi for technical assistance. In this work, the computers at the DDBJ and the Human Genome center of Japan were used.