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
II. Introduction
III. Materials and Methods
   1. Cloning and sequencing
   2. Database search and data-submission
   3. Southern blot analysis and chromosome in situ hybridization
   4. GC% measurement
   5. Analysis of microsatellite alleles
IV. Results
   1. Chromosome walking from MHC class III to class II and newly found genes
      A) Gene for RAGE (AGER); B) Gene for PBX2 (HOX12)
   2. YAC clones carrying class II HLA-DRA and Southern blot analysis
   3. Chromosome in situ hybridization
   4. Human counterpart of mouse mammary tumor gene int-3;
      NOTCH3
      A) Structure of NOTCH3; B) Four subfamilies of mammalian Notch; C) Copy number of the human NOTCH3
   5. Boundary of long-range GC% mosaic domains assigned by GC% measurement
      A) Long-range GC% mosaic structures; B) Characteristic structures around the boundary of long-range GC% mosaic domains already published
   6. Other characteristic structures found around the boundary of long-range GC% mosaic domains
V. Discussion
   1. Newly found genes with respect to disease related genes
   2. Gross similarity of genes on 6p21.3 and those on 9q33-q34
VI. References
VII. Figures and Tables
VIII. Acknowledgments
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), II (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 cosmids, 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; Betteken 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, Cw63) (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 NsiI-generated insert fragment of an already obtained cosmid, as its probe, which were labeled with [α-32P] 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 EcoRI 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 32P-labeled probes (> 5×10^8 cpm / μg) of genomic fragments. Positive plaques were picked and purified by two additional rounds of plating and probing.

EcoRI, HindIII, BamHI or PstI fragments of cosmid or λ cloned insert, as well as smaller fragments produced by successive ScaI 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 *BcaBEST* 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°C for 15 min in 0.1 × SSPE (1 × SSPE is 0.18 M NaCl, 10 mM NaH₂PO₄, 1 mM EDTA, pH 7.7) 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 NotI 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 ZnCl2, 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 H3PO4 - 10 mM KH2PO4 (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 (Neuper 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, NFκB) (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 3H-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 AKIIBA 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 Schambocck 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 EcoRI-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 EcoRI 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 BamHI 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 ^3H-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 diffusable 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 EcoRI 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 (Notch 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 cde10/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 \( \lambda \) 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 HI 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 PstI 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 EcoRI fragment harboring the 5' region of NOTCH3 and (ATT)$_{11}$ in the 10-kb EcoRI 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)\textsubscript{10} repeat locating near the 5' region of NOTCH3. The sequences just surrounding the (CTG)\textsubscript{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 \textit{Sau}3AI or \textit{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 \textit{Sau}3AI digestion, four types of alleles, as well as RFLPs (restriction fragment length polymorphism) of \textit{Sau}3AI, were found (Fig. 15 B). To make sure of the \textit{Sau}3AI RFLPs and of knowing VNTRs (variable numbers of tandem repeat units) of the \textit{Sau}3AI-undigested fragments, analyses with \textit{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)\textsubscript{12} and the shortest to be (CTG)\textsubscript{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)\textsubscript{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 EcoRI 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 EcoRI 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 X1, 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


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.


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.


Fig. 1. Molecular map of contiguous cosmids and VAC clones that cover junction of MHC classes I and II. Ordered cosmids clones are represented by horizontal lines with vertical bars indicating EcoRI sites. Hatched boxes indicate locations of NOTCH3, HOX12 and RAGE genes and black box that of tenascin-X gene (TNX). Arrows show direction 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 EcoRI (E), PstI (P), HincII (H) and BamHI (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.
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 EcoRI, separated by electrophoresis on a 1% agarose gel, blotted onto a Hybond N nylon membrane and hybridized with 32P-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 32P-labeled RAGE, HOX12 or NOTCH3 probe: B, BamHI; Bg, BglII; D, Dral; E, EcoRI; H, HindII; P, PstI; Pvu, PvuII; K, KpnI. Positions of probes are listed in Fig. 2A or Fig. 6. For stringent washing, filters were washed at 65°C 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 PvuII 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.
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 below (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 1F (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 16 (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 17 (1.6 kb). CDC10-6 underlined is the last repeat in cdc10/ankyrin repeats. Inter-domain sequence is underlined by a broken line.
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 (Herm, 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 14 and in Notch/1in-12 repeats. + Molt 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, phage and YAC clones that cover junction of MHC classes II and III. Ordered clones are represented by horizontal lines with vertical bars indicating EcoRI sites; cosmid clones (M, KS, F series) are indicated by thicker horizontal lines than 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; 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 phage clones, were by Fukagawa (Fukagawa, Sugaya, et al., 1995). After completion of 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 P515P5T5, 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 EcoRI fragments with arrows. Hatched boxes indicate Alu cluster and LINE1 sequences.
Fig. 13 continued.
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.
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. D02086915, 10. MANIKA, 11. BM15, 12. AMA1, 13. PL117, 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 the (CTG)_{10} repeat, which sequence was determined by subcloning of cosm id clone KS74, are indicated. Vertical bars and numbers indicate Sau3AI and SacI sites. Oligonucleotide primers corresponding to chromosomal regions flanking 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 Sau3AI (B) or SacI (C). These products were analyzed by electrophoresis on a 8%(w/v) polyacylamide 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 CAG GGA CTC GGI AG-3'; PABL: 5'-CTT ACT CTT TGG CTT AC-3', 5'-ATG CAG ATG AAC TGA GA-3'.
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.
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

<table>
<thead>
<tr>
<th>No</th>
<th>position</th>
<th>(nt)</th>
<th>5' splice donor</th>
<th>(nt)</th>
<th>3' splice acceptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;6665-6715</td>
<td>&gt;52</td>
<td>CTG TGG Ggt gag cca</td>
<td>183</td>
<td>ctc cca gGG GCA GTA</td>
</tr>
<tr>
<td>2</td>
<td>6900-7006</td>
<td>107</td>
<td>TGG AAA CTG gta agc</td>
<td>130</td>
<td>ttc tag AAC ACA GGC</td>
</tr>
<tr>
<td>3</td>
<td>7137-7332</td>
<td>196</td>
<td>GTC TAC Cgt aag aat</td>
<td>166</td>
<td>cct tca gAG ATT CCT</td>
</tr>
<tr>
<td>4</td>
<td>7499-7563</td>
<td>65</td>
<td>CCC AAT AAG gta gtg</td>
<td>122</td>
<td>ccc cag GTG GGG ACA</td>
</tr>
<tr>
<td>5</td>
<td>7686-7773</td>
<td>88</td>
<td>GAG AAG Ggt gag tcc</td>
<td>90</td>
<td>atc ata gGA GTA TCT</td>
</tr>
<tr>
<td>6</td>
<td>7864-8046</td>
<td>183</td>
<td>GTC TGG Ggt gag cat</td>
<td>142</td>
<td>cca gAG CCT GTG CCT</td>
</tr>
<tr>
<td>7</td>
<td>8189-8319</td>
<td>131</td>
<td>ATG AAG GAT gtg agt</td>
<td>179</td>
<td>cac cag GTG GTG CCC</td>
</tr>
<tr>
<td>8</td>
<td>8499-8640</td>
<td>142</td>
<td>ATC ATC Ggt gag acc</td>
<td>615</td>
<td>ctt cca gAA CCA GGC</td>
</tr>
<tr>
<td>9</td>
<td>9256-9282</td>
<td>27</td>
<td>ACT GCA Ggt gag ggg</td>
<td>128</td>
<td>cgt aca gGC TCT GTG</td>
</tr>
<tr>
<td>10</td>
<td>9411-9537</td>
<td>127</td>
<td>GAG GAG AGg tga gtg</td>
<td>111</td>
<td>ctc ctc agG AAG GCC</td>
</tr>
<tr>
<td>11</td>
<td>9649-9920</td>
<td>272</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transcription start site has not been determined.
### TABLE 2

Exon-Intron Organization of the HOX12 Gene

<table>
<thead>
<tr>
<th>Exon No</th>
<th>Position (nt)</th>
<th>Exon Sequence</th>
<th>Intron Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>707-1198</td>
<td>CAG GCC AAg tga tgt</td>
<td>901 ctc cca agG AAACAC</td>
</tr>
<tr>
<td>2</td>
<td>2100-2173</td>
<td>AAA ACT Ggt atg tgg</td>
<td>218 atg cta gGC CTCAGC</td>
</tr>
<tr>
<td>3</td>
<td>2392-2639</td>
<td>TAT CAG CAG gta agg</td>
<td>100 cta cag GCA TGTAAT</td>
</tr>
<tr>
<td>4</td>
<td>2740-2930</td>
<td>GAT GCC AGg tgg gcc</td>
<td>183 ctc tgt agA CGAAAG</td>
</tr>
<tr>
<td>5</td>
<td>3114-3249</td>
<td>GTG TCT CAG gta tta</td>
<td>251 cca cag GTC TCCAAC</td>
</tr>
<tr>
<td>6</td>
<td>3501-3654</td>
<td>TCT GCA Ggt gga tcc</td>
<td>336 cca cca gGC TCTGGC</td>
</tr>
<tr>
<td>7</td>
<td>3991-4079</td>
<td>GCT TCC CAG gtc aga</td>
<td>122 ctc cag GTG GAATCA</td>
</tr>
<tr>
<td>8</td>
<td>4202-4288</td>
<td>GAA ATG AGG tgt agt</td>
<td>129 ttt cag GCA AATGGC</td>
</tr>
<tr>
<td>9</td>
<td>4418-6146</td>
<td>1729</td>
<td></td>
</tr>
</tbody>
</table>

Transcription start site has not been determined.
# TABLE 3

Properties of allele analyzed cell lines

<table>
<thead>
<tr>
<th>No</th>
<th>NAME</th>
<th>A</th>
<th>C</th>
<th>B</th>
<th>B6/6</th>
<th>B1</th>
<th>C2</th>
<th>C4A</th>
<th>C4B</th>
<th>DW</th>
<th>DR</th>
<th>DW2</th>
<th>DRW52/S3</th>
<th>DQ</th>
<th>DP</th>
<th>GLO</th>
<th>(CTG)n/Sau3A1</th>
<th>HLA DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MGAR</td>
<td>26</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>15</td>
<td>6</td>
<td>4</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
<td>F2</td>
</tr>
<tr>
<td>2</td>
<td>OMAR</td>
<td>2</td>
<td>45</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>13</td>
<td>24</td>
<td>52</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>3</td>
<td>COX</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td>S</td>
<td>C</td>
<td>Q0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>24</td>
<td>52</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MADURA</td>
<td>2</td>
<td>10</td>
<td>60</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>52</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>F1</td>
</tr>
<tr>
<td>5</td>
<td>OLGA</td>
<td>31</td>
<td>1,10</td>
<td>62</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>52</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>F2,F3</td>
</tr>
<tr>
<td>6</td>
<td>RSH</td>
<td>68,30</td>
<td>2</td>
<td>42</td>
<td>6</td>
<td>F</td>
<td>C</td>
<td>1</td>
<td>Q0</td>
<td>N</td>
<td>3</td>
<td>24</td>
<td>52</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td>F2</td>
</tr>
<tr>
<td>7</td>
<td>BSM</td>
<td>2</td>
<td>5</td>
<td>62</td>
<td>6</td>
<td></td>
<td></td>
<td>S</td>
<td>C</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>53</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SAVC</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>53</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>9</td>
<td>D0209915</td>
<td>25</td>
<td>-</td>
<td>18</td>
<td>6</td>
<td>S</td>
<td>Q0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>15</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>F2</td>
</tr>
<tr>
<td>10</td>
<td>MANIKA</td>
<td>3</td>
<td>-</td>
<td>50</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>53</td>
<td>233</td>
<td>NT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>11</td>
<td>BM15</td>
<td>1</td>
<td>7</td>
<td>49</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>11</td>
<td>25</td>
<td>52</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>12</td>
<td>AMA1</td>
<td>28</td>
<td>4</td>
<td>53</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>15</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F4</td>
</tr>
<tr>
<td>13</td>
<td>FE117</td>
<td>24</td>
<td>10</td>
<td>60</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>4</td>
<td>53</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>14</td>
<td>DEU</td>
<td>31</td>
<td>4</td>
<td>35</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>53</td>
<td>7</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>15</td>
<td>PF04015</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>24</td>
<td>52</td>
<td>2</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td>F4</td>
</tr>
<tr>
<td>16</td>
<td>SP0010</td>
<td>2</td>
<td>5</td>
<td>44</td>
<td>4</td>
<td>DB2</td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>25</td>
<td>52</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>F2</td>
</tr>
<tr>
<td>17</td>
<td>BM02</td>
<td>25</td>
<td>1</td>
<td>51</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>4</td>
<td>53</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>18</td>
<td>BTB</td>
<td>2</td>
<td>1</td>
<td>27</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>52</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>19</td>
<td>LUY</td>
<td>2</td>
<td>-</td>
<td>51</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>52</td>
<td>7</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>20</td>
<td>HO104</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>S</td>
<td>C</td>
<td>3</td>
<td>1</td>
<td></td>
<td>15</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>21</td>
<td>CBL</td>
<td>26</td>
<td>75</td>
<td>18</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>25</td>
<td>52</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>22</td>
<td>AKIBA</td>
<td>24</td>
<td>63</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
<tr>
<td>23</td>
<td>KS74</td>
<td>10</td>
<td>1</td>
<td>59</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
<td>52</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>F3</td>
</tr>
</tbody>
</table>

**Note.** The column "(CTG)n/Sau3A1" 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 Sau3A1. The allele of OLGA is heterozygous.
### TABLE 4

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

<table>
<thead>
<tr>
<th>6p21.3</th>
<th>chromosome 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene/Locus</td>
<td>Physical Location</td>
</tr>
<tr>
<td>VARS2</td>
<td>6p21.3</td>
</tr>
<tr>
<td>HSPA1, HSPA1L</td>
<td>6p21.3</td>
</tr>
<tr>
<td>C2, C4A, C4B</td>
<td>6p21.3</td>
</tr>
<tr>
<td>TNX</td>
<td>6p21.3</td>
</tr>
<tr>
<td>HOX12 (PBX2)</td>
<td>6p21.3</td>
</tr>
<tr>
<td>NOTCH3 (INT3)</td>
<td>6p21.3</td>
</tr>
<tr>
<td>COL11A2</td>
<td>6p21.3</td>
</tr>
<tr>
<td>RXRB</td>
<td>6p21.3</td>
</tr>
</tbody>
</table>
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 Eiichi 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.