Studies on a Boundary of Long-Range G+C % Mosaic Domains in the Human Genome: Characterization of Pseudoautosomal Boundary-Like Sequence (PABL) Found near the Boundary

By

Tatsuo Fukagawa

Doctor of Philosophy

Department of Genetics

School of Life Science

The Graduate University for Advanced Studies

1995

CONTENTS

Chapter I

SUMMARY

The human genome, like those of warm-blooded vertebrates in general, is composed of long-range G+C% (GC%) mosaic structures related to chromosome bands. Several groups showed that the Giemsa-dark G bands are composed mainly of AT-rich sequences, and T bands (a subgroup of Giemsa-pale R bands) mainly of GC-rich sequences: ordinary R bands are heterogeneous and appear to be intermediate. Gene density, CpG island density, codon usage, chromosome condensation, DNA replication timing, repeat sequence density, and other chromosome behaviors such as recombination and mutation rate are related to chromosome bands and to long-range GC% mosaic domains. Gene-dense T bands with loose chromatin structures replicate early in S phase and are rich in *Alu* repeats, while G bands with condensed chromatin structures replicate late and are rich in LINE-1 repeats. Because chromosome bands are structures observed with microscopes, precise location of their boundaries may seem meaningless. However, considering various genome behaviors connected with chromosome bands, it appears possible to precisely locate their boundaries by putting informative landmarks on genome DNA. Boundaries may be structurally assigned as clear GC% transition points, and signals for punctuating and/or differentiating respective functions (e.g., a switching signal from early to late DNA replication) may be found in the boundaries.

The human major histocompatibility complex (MHC) has classes I (about 2 Mb), III (1 Mb), and II (1 Mb) from telomere to centromere. Ikemura and his colleagues had found the human MHC to be a typical example of long-range GC% mosaic structures by analyzing MHC sequences compiled by GenBank database and predicted a possible boundary of the Mb-level domains within an under-characterized 450 kb containing the junction of MHC classes II and III. To clone the mosaic boundary, bidirectional chromosome walking from the class III CYP21 to the class II and from the class II HLA-DRA to the class III was conducted in this study, and contiguous clones covering the 450-kb region bridging classes II and III were obtained. To analyze base-compositional distribution of the walked area, insert DNAs of the clones were

purified and digested by nuclease P1, and GC% was measured by a HPLC method. About 150 kb from the HLA-DRA was fairly homogeneously AT-rich (mostly less than 40% GC) showing extension of AT-rich sequences from the class II side. Then a sharp transition to about 50% GC occurred and this GC-rich level continued to the class III CYP21. To analyze the structures near and at the GC% transition, the cosmid and λ clones harboring the transitional region were sequenced. The following three types of characteristic structures were found; *Alu* repeats densely clustered in a 20-kb region, five LINE-1 repeats also clustered in a 30-kb region, and a sequence highly homologous with the pseudoautosomal boundary sequence of the short arms of the human sex chromosomes (PABX1 and PABY1); PABX1 and PABY1 are the interface sequences between sex-specific and pseudoautosomal regions. The author designated the sequence highly homologous with PABXY1 "PABL". There exists a possibility that the organization, a dense *Alu* cluster - a dense LINE-1 cluster - a PABL, is one characteristic of certain types of long-range GC% mosaic boundaries and of band boundaries. The author focused on characterization of the newly found sequence, PABL.

Human sex chromosomes are divided into two functionally distinct regions, sex-specific sequences and pseudoautosomal regions (PARs); within each male meiosis, X and Y chromosomes exchange DNA sequences by homologous recombination in PARs and thus PAR sequences are practically identical between X and Y chromosomes because of this obligatory recombination. The interface between PAR1 (about 2.6 Mb) and the sex-specific region is the pseudoautosomal boundary (PAB1) and therefore PAB1 is the proximal (centromeric) limit to X-Y homologous recombination in PAR1: PAB1 is very unique in the human genome as a strict physical site at which unusually high frequency recombination in the 2.6 Mb of PAR1 (known to be 20-fold greater than the genome average) terminates abruptly. Ellis and his colleagues reported sequences around the interface, i.e., PABX1 and PABY1 sequences. Interestingly, the sequence found near the boundary of the long-range GC% mosaic domains in the MHC is highly homologous (about 80% nucleotide identity) with the PABXY1 sequences which constitute the functional interface in the sex chromosomes. Using the sequence in the MHC as a probe, multiple copies of pseudoautosomal boundary-like sequences (PABLs) were detected

through Southern blot hybridization against genomic DNAs and cosmid cloning. The author defined a ca. 650-nt consensus sequence of the PABL core by determining and comparing eleven independent PABL sequences.

Although GenBank genomic sequences showing significant homology with the PABLs were confined to PABXY1, several human ESTs (expressed sequence tags) showed evident homology with separate portions of PABLs, indicating some, if not all, PABLs are transcribable. To clarify characteristics of the predicted PABL transcripts, six human cDNA libraries of different tissues and cells were screened using the PABL segment as a probe. Positive clones were obtained from all six libraries. Sequence analysis of the six cDNA clones showed there exists again a 650-nt core sequence which corresponds to that defined by the genomic PABLs. No ORFs with significant sizes could be found for the obtained cDNAs, not only for the PABL core sequences but also their flanks. When the cDNA sequences were searched with the BLASTX program against the protein sequence database, no significant homology with known proteins was detected. GRAIL, a computer program trained to identify protein-coding ORFs in human DNA, also could not detect reliable protein-coding capacity. These may suggest that their functional form, if present, is RNA molecules. To estimate intact sizes of PABL transcripts, northern blot analysis of human total or polyA$^+$ RNA fraction was conducted using the PABL probe. Broad bands, estimated to be 5-10 kb in length, were detected.

For study of evolutionary processes involved in forming the present PABXY1 and PABLs, their phylogenetic relationships were examined. In order to estimate evolutionary rates, the reported PABXY1 sequences of great apes and Old World monkeys were included; divergence between great apes and Old World monkeys is postulated here to be 25 million years ago. Phylogenetic trees were constructed using the neighbor-joining method. Using the evolutionary distance between human and Old World monkey PABX1, as well as that for PABY1, divergence time of PABLs including PABXY1 was estimated to be 60-120 million years: this is consistent with the result obtained through Southern hybridization that PABLs are present in the bovine genome. The evolutionary rates of individual PABLs and PABXY1 were then calculated

using the divergence time. The rates of some PABLs were far less than 1×10^{-9} substitutions per site per year, indicating evolutionary and functional constrains were executed on PABL sequences. Taking phylogenetic relationship between PABLs and PABXY1 into consideration, evolutionary process in the formation of the present pseudoautosomal boundary PAB1 is proposed by postulating an illegitimate recombination between two PABLs.

Chapter II

INTRODUCTION

1. Organization of the human genome

The word "genome" is over 70 years old. It was defined by Winkler (1920) to indicate the sum total of genes (of a haploid cell) of organism. Now, the non-coding sequences, whose existence was not known at that time, are included in the definition. It has become clear that the coding sequences represent only several percent of the human genome and therefore more than 90% of the genome corresponds to the non-coding sequences which contain large families of repeated sequences often being called "junk DNA". DNA sequence data including non-coding sequences have acceleratedly accumulated and non-coding sequences have begun to be recognized as functional components. The genome is now considered to be a comprehensive system in which nucleotide sequences including the non-coding regions obey rather precise rules that amount to a wide range of genomic codes. These views significantly arose from the base compositional analysis of large DNA fragments, coding and non-coding sequences in vertebrate genomes (Bernardi *et al.*, 1985). Analyzing human DNA by cesium salt density-gradient fractionation followed by Southern blot analysis, Bernardi *et al.* found that the genome of warm-blooded vertebrates is composed of mosaic structures of very long (> 300,000 bases) DNA sequences, each of which is fairly homogeneous in its base composition with a few different levels of G+C% (GC%), and that codon choice in each gene depends on GC% of the DNA segment harboring the gene (Bernardi *et al.*, 1985; Bernardi, 1989). They called the mosaic domains of the long-range regions homogeneous in GC% "isochores". Ikemura and his colleagues showed a positive correlation between the GC% at the codon third position and the GC% of both intron and wide flanking portions (Ikemura, 1985; Aota and Ikemura, 1986). These findings, together with those of later studies (Ikemura and Aota, 1988; Holmquist, 1989; Gardiner *et al.*, 1990; Ikemura *et al.*, 1990; Holmquist, 1992; Pilia *et al.*, 1993; Gardiner, 1995), indicate that the genomes of higher vertebrates including human have long-range mosaic structures of GC%, which are related to

chromosome bands and thought to constitute functional domains.

2. Chromosome band structures

The discovery of metaphase banding in 1970 marked a revolution in the understanding of human chromosome structures. The human karyotype is now defined by three structural sets of regions, the Giemsa (G) or Quinacrine (Q) bands, the Reverse (R) bands, and the Centromeric (C) bands. These bands are produced in metaphase chromosomes with fluorescent dyes, proteolytic digestion, or differential denaturing conditions (reviewed in Comings, 1978; Therman, 1986). Since their discovery, it has become increasingly clear that G/Q and R classes are associated with a broad range of inverse functional and biochemical attributes. Several researchers showed that G bands are composed mainly of AT-rich sequences, and T-type R bands (T bands; a heat-stable subgroup of R bands) mainly 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; Craig and Bickmore, 1993; Saccone *et al.*, 1993). Gene density, CpG island density, codon usage, chromosome condensation, DNA replication timing, repeat sequence density, and other chromosome behaviors such as recombination and mutation rates are related to chromosome bands and to long-range GC% mosaic domains (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). Gene-dense R (and especially T) bands with loose chromatin structures replicate early in S phase and are rich in *Alu* repeats, while G bands with condensed chromatin structures replicate late and are rich in LINE-1 repeats. These features are summarized in Table 1.

Because chromosome bands are structures observed with microscopes, precise location of their boundaries may seem meaningless. However, considering various genome behaviors connected with chromosome bands, we may be able to precisely locate band boundaries by using informative landmarks on the genome DNA. Boundaries may be structurally assigned as

clear GC% transition points, and functional signals may be found for punctuating and/or differentiating respective functions, e.g., a switching signal from early to late DNA replication. In this study the author focused on this problem and attempted to clarify characteristics of boundaries of chromosome bands and of long-range GC% mosaic structures.

3. The boundary of long-range G+C% mosaic domains in the human MHC

The human MHC has classes I (about 2 Mb), III (1 Mb), and II (1 Mb) from telomere to centromere (Campbell and Trowsdale, 1993). Ikemura *et al.* (1988, 1990) had analyzed the human MHC sequences and found a possible boundary of the GC% mosaic domains near the junction between MHC class II and class III in the following way. To examine the mosaic boundary, they ordered all non-redundant sequence (> 2 kb) of the MHC according to their genetic positions. All sequences from class I to III (spanning about 2.5 Mb) were evidently GC-rich. Sequences for class II (spanning about 1 Mb), however, had significantly lower GC% levels. Therefore, a possible boundary of the Mb-level GC% mosaic domains was assigned within an under-characterized 450 kb containing the junction of MHC classes II and III. At a standard 850-band level, the MHC is on a wide R band, 6p21.3, and by higher resolution banding, a narrow G subband 6p21.32 is located within the MHC. For these reasons, the human MHC is a good example for studying the long-range GC% mosaic structures. To precisely locate the domain boundary in the 450 kb, bidirectional chromosome walking from a centromeric class III gene CYP21 to class II (Matsumoto *et al.*, 1992; Sugaya *et al.*, 1994) and from a telomeric class II gene HLA-DRA to class III (Fukagawa *et al.*, 1995a) were conducted. In this study the author analyzed GC% of the contiguous clones covering the 450-kb region containing the domain boundary, and disclosed a sharp GC% transition.

4. Human pseudoautosomal boundary

Near the boundary of long-range GC% mosaic domains in the human MHC, Fukagawa *et al.* (1995a) found a sequence highly homologous with the pseudoautosomal boundary (PAB) sequence of the short arms of the human sex chromosomes. Human sex chromosomes are

divided into two functionally distinct regions, sex-specific sequences and pseudoautosomal regions (PARs). X and Y chromosomes exchange DNA sequences through homologous recombination in PARs within each male meiosis, and thus PAR sequences are practically identical between the two chromosomes because of the obligatory recombination (Cooke *et al.*, 1985; Simmler *et al.*, 1985; Ellis and Goodfellow, 1989; Freije *et al.*, 1992; Rappold, 1993; Kvalϕy *et al.*, 1994). There are two PARs for human sex chromosomes: PAR1 is at the distal ends of the short arms of the X and Y chromosomes and PAR2 of the long arms (Freije *et al.*, 1992; Rappold, 1993; Kvalϕy *et al.*, 1994). The interface between the PAR1 of about 2.6 Mb and the sex-specific region is the pseudoautosomal boundary PAB1, and therefore PAB1 is the proximal (centromeric) limit to recombination in PAR1. Ellis and Goodfellow (1989) and Ellis *et al.* (1989, 1990) reported sequences around the interface, PABX1 and PABY1 sequences (abbreviated PABXY1). As noted above, a sequence found in the boundary of long-range GC% mosaic domains in the MHC is highly homologous with the PABXY1 sequences which are considered to constitute the functional interface in the sex chromosomes. Fukagawa *et al.* (1995a) designated the sequence found in the MHC "pseudoautosomal boundary-like sequence 1 (abbreviated PABL1)".

Using the PABL1 segment as a probe, multiple copies of pseudoautosomal boundary-like sequences (PABLs) were detected through Southern blot hybridization against genomic DNAs (Fukagawa *et al.*, 1995a). Although multiple copies of PABLs were further confirmed by cosmid cloning, human genomic sequences in the databases showing evident homology with the PABL1 were confined to PABXY1. However, when Expressed Sequence Tag (EST) sequences were searched, two human ESTs showed evident homology with separate portions of PABL1, suggesting some, if not all, PABLs are transcribable and have some functions (Fukagawa *et al.*, 1995a). In this study, the author elucidated the ca. 650-nt core and consensus sequence of human PABLs, and characterized their transcripts by isolating cDNA clones. In addition, the author proposed a model of evolutionary formation of the present-day pseudoautosomal boundaries of the short arms of the human sex chromosomes considering

possible functions of PABLs.

Chapter III

MATERIALS AND METHODS

General procedures of molecular biology used in this study were performed according to "Molecular Cloning" (Sambrook *et al.*, 1989): extraction and purification of plasmid DNA, agarose or polyacrylamide gel electrophoresis, restriction enzyme digestion, polymerase chain reaction (PCR), modifying enzyme reaction, preparation and transformation of *E. coli* competent cells, and preparation of reagents. The reagents used in this study were analytical grade. The recovery and purification of DNA fractionated on an agarose gel was performed with QIAEX Gel Extraction kit (QIAGEN Gmbh, Germany), following the manufacture's protocol.

1. Chromosome walking

A cosmid library was constructed from the total human DNA of the HLA homogeneous B cell line AKIBA (HLA-A24, Bw52, DR2, Dw12, DQw1, Cp63) by using the cosmid vector pWE15 (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. A yeast artificial chromosome (YAC) library was constructed from the B-cell line CGM1 (A3, B8, Cw-, DR3, DQ2, DR52 and A29, B14, Cw-, DR7, DQ2, DR52) (Imai and Olson, 1990). The YAC library was screened using the HLA-DRA primer set (Ando *et al.*, 1994) and the positive clones characterized were YDR2 carrying about 450-kb insert and YDR3 carrying about 220-kb insert (Sugaya *et al.*, 1994; Fukagawa *et al.*, 1995a). The libraries from the yeast high-molecular-weight DNA containing these YACs were constructed by using λ DASH II (Stratagene, CA) and pWE15 vectors according to the manufacturer's protocol. Chromosome walking was conducted by using a restriction-enzyme digested fragment of an already obtained clone as its probe, which was labeled with [α-^{32}P] dCTP by the random-priming method. Hybridization with the labeled probe DNA was performed in hybridization buffer containing 5 × SSPE (1 ×

SSPE is 0.18 M NaCl, 1 mM EDTA, and 10 mM NaH_2PO_4 pH 7.7), 5 × Denhardt's solution (1 × Denhardt's solution contains 0.02% Ficoll, 0.02% polyvinylpyrrolidone, and 0.02% bovine serum albumin), 100 μg/ml freshly denatured salmon sperm DNA, 80-330 μg/ml freshly denatured human placental DNA, and 0.5% sodium dodecyl sulfate (SDS) at 60°C for 15 to 18 h after 1 h of prehybridization in the same buffer solution without the probe DNA as described by Sealey *et al.* (1985). The isolated cosmid or λ DNAs were subjected to *Eco*RI mapping.

2. Cloning of PABLs

A cosmid library constructed from the total human DNA of the HLA homogeneous B-cell line AKIBA on the pWE15 vector by Inoko *et al.* (1985) and a λ-EMBL3 library constructed from the human genomic DNA of peripheral blood cells by Tomatsu *et al.* (1989) were used. Six human cDNA libraries cloned on either the λgt10 or λgt11 vector were obtained from Clontech (Palo Alto, CA.); placenta 5'-stretch plus (λgt11), placenta (λgt11), monocyte (λgt11), B-cell (λgt10), spleen (λgt10) and skin fibroblast (λgt10) cDNA libraries. Cloning of PABLs from a human genomic or a cDNA library was done with a standard method (Sambrook *et al.*, 1989). Replica filters from libraries (1-5 × 10^5 pfu or cfu / 20 cm × 20 cm plate) were hybridized with the ^{32}P-labeled 360-nt portion of PABL1 corresponding to PABXY1 sequences. Positive clones were picked, purified, and subcloned into pUC118 or pT7Blue (Novagen, WI) vectors for nucleotide sequencing.

3. Nucleotide sequencing

*Eco*RI, *Hin*dIII, *Bam*HI, or *Pst*I fragments of cosmid or λ cloned inserts, as well as smaller fragments produced by the successive *Sau*3AI digestion, were subcloned into pUC118. The

deletion mutants of these clones were constructed by the ExoIII / Mung Bean nuclease system following the manufacture's protocol (Takara Shuzo Co., Ltd, Kyoto, Japan) and subcloned into pUC118. PCR fragments from the cDNA clones using λ vector primers were subcloned into pT7Blue. These plasmids were sequenced by *Taq* cycle sequencing kit, using fluorescence-labeled DyeDeoxy terminators for ABI 373A sequencer (Applied Biosystems, Foster City, CA). Sequences were aligned into contigs, and to bridge them, primers for sequencing were prepared. Fragmental sequences were connected and assembled into contigs by the ATSQ program of GENETYX (Software Development Co., Ltd., Japan). A database search of DDBJ, GenBank, EMBL, PIR, and SWISS-PROT was done using the FASTA and BLAST programs (Pearson and Lipman, 1988; Altschul *et al.*, 1990).

4. GC% measurement

Insert DNAs (30-40 kb) derived from cosmid clones were separated from RNA and the vector DNA by high-performance liquid chromatography (HPLC); cosmid DNAs were digested with *Not*I at the pWE15 linker, applied to 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 to 1 M) in 0.02 M Tris-HCl (pH 9.0). λ phage clones were digested with *Not*I, and their inserts were purified by electrophoresis on a 0.5% agarose. GC% of purified inserts was measured with a DNA-GC kit (Yamasa Shoyu Co., Chiba, Japan) as follows. According to the manufacturer's protocol, 20 μg of DNA (EDTA free) dissolved in 20 μl of distilled water was heated at 100°C for 5 min, rapidly cooled 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 h. P1 hydrolysates and a standard mononucleotides mixture supplied by the manufacturer were separately chromatographed on a YMC reversed-phase HPLC column (ODS-AQ-312, 0.6 × 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. Sequence alignment and construction of phylogenetic trees

Alignments and calculations of sequence identity were conducted using the MALIGN program available on the UNIX system of DNA Data Bank of Japan (DDBJ) in National Institute of Genetics. To determine the core sequence of PABLs, all pairs of sequences containing PABLs or PABXY1 were first aligned, and multiple alignments of the homologous portions thus found were calculated according to Hein (1990). For the construction of phylogenetic trees, the author divided individual PABLs into two regions, which were the downstream and upstream regions to the *Alu*-insertion site, according to Ellis *et al.* (1990). Phylogenetic trees were constructed by the neighbor-joining method (Saitou and Nei, 1987). A root was predicted by using the UPGMA method (Sneath and Sokal, 1973). Evolutionary distances (number of nucleotide substitutions) were estimated using the one-parameter (Jukes and Cantor, 1969) and two-parameter (Kimura, 1980) methods, and these distances were used to construct neighbor-joining trees. Bootstrap probabilities, based on 1000 resamplings, were calculated for each internal branch of neighbor-joining trees using the NJBOOT2 program (kindly provided by Dr. K. Tamura, Tokyo Metropolitan University). Estimation of evolutionary rates was carried out as described by Nei (1987).

6. Southern blot analysis

High-molecular-weight DNA was prepared from YAC-containing yeast strains grown to saturation in uracil- and tryptophan-deficient liquid media (Brownstein *et al.*, 1989), as described by Smith and Canter (1987). High-molecular-weight DNA from human placenta (Clontech, Palo Alto, CA) and bovine lung (Clontech, Palo Alto, CA) were purchased from the companies. Sample DNAs were digested by restriction enzymes to completion, size fractionated by electrophoresis on a 1% agarose gel, and blotted onto nylon membranes (Hybond-N$^+$, Amersham) using the Vacugene XL vacuum blotting system (Pharmacia, Co., Uppsala, Sweden) according to Pharmacia's instruction manual. Hybridization with the radiolabeled PABL probe was carried out in hybridization buffer containing 5 × SSPE, 5 × Denhardt's

solution, 100 μg/ml freshly denatured salmon sperm DNA, and 0.5% SDS at 60°C for 15 to 18 h after 1 h of prehybridization in the same buffer solution without the probe DNA, as described by Sambrook *et al.* (1989). Stringent washing was performed at 65℃ for 15 min in 0.1 × SSPE containing 0.1% SDS. The membranes were subjected to autoradiography using a Fuji Bio-Imaging Analyzer BAS 2000 (Fuji Photo Film Co., Japan) or X-ray films.

7. Northern blot analysis

Total RNA was extracted from GM01416D cells or from several tissues according to the AGPC method (Chomczynski and Sacchi 1987) as the following. The frozen cells (about 10^8 cells) were homogenized in 5 ml D solution; D solution contains 4 M guanidine thiocyanate, 25 mM sodium citrate (pH 7.0), 0.5% sarcosyl, and 0.1 M 2-mercaptoethanol. Then, 0.5 ml of 2 M sodium acetate (pH 4.0) was added and mixed by inversion. Five ml of phenol saturated with DEPC-treated water was added and mixed by inversion. One ml of mixture of chloroform and isoamyl alcohol (49 : 1, by volume) was added, mixed by inversion, shaken vigorously for 10 sec, and placed on ice for 15 min. Centrifugation was performed at 3,000 rpm for 40 min at 4 ℃. The upper, aqueous phase was carefully transferred into a fresh new tube. Two volumes of ice cold ethanol (EtOH) were added and stored at -20℃ for 1 h. The precipitate was collected by centrifugation at 3,000 rpm for 40 min at 4℃. The pellet was dissolved in 0.3 ml of D solution and transferred into a 1.5 ml tube. Three volumes of EtOH were added, mixed well and stored at -70℃ for 30 min. The RNA precipitate was collected by centrifugation at 12,000 rpm for 15 min. The RNA pellet was dissolved in DEPC-treated water. PolyA$^+$ RNA was purified using an oligo-dT column. Twenty five μg of total RNA from GM01416D cells, or 2 μg of the polyA$^+$ RNA fractions from various tissues were electrophoresed on a 1% agarose gel in a buffer containing 6% formaldehyde, 20 mM MOPS (pH 7.0), 1 mM EDTA, and 5 mM sodium acetate. After electrophoresis, the gel was transferred to Hybond-N$^+$ using the Vacugene XL vacuum blotting system. Hybridization was carried out in solution containing 5 × SSPE, 10

× Denhardt's solution, 100 μg/ml freshly denatured salmon sperm DNA, and 2% SDS at 60℃ for 18 to 24 h after 3 h of prehybridization in the same buffer solution without the probe DNA. Final washing was performed at 60℃ in 0.1 × SSC containing 0.1% SDS. The membranes of the northern blot were subjected to autoradiography with a Fuji Bio-Imaging Analyzer BAS 2000.

8. Chromosome in situ *hybridizations*

Fluorescence *in situ* hybridization (FISH) was used to assign the chromosome location of the cosmid or λ clones containing PABLs. Chromosome spreads were obtained from phytohemagglutinin-stimulated blood lymphocytes of a healthy male donor after thymidine synchronization and bromodeoxyuridine incorporation according to Takahashi *et al.* (1990, 1991). The cosmid or λ clones were labeled with biotin-16-dUTP (Boehringer Mannheim) by nick translation. *In situ* hybridization was performed according to Lichter *et al.* (1980) in the presence of human COT-1 DNA (GIBCO BRL, Gaithersburg, MD) as a competitor. The hybridized probe was detected with FITC-conjugated avidin (Boehringer Mannheim) without further signal amplification. Chromosomes were counterstained with 0.2 μg/ml propidium iodide for R-banding. In this method, R-banded chromosomes also show the counterpart G-banding pattern in their Hoechst 33258 staining. Fluorescence signals were imaged using a Zeiss Axioskop epifluorescence microscope equipped with a cooled Charge Coupled Device (CCD) camera (Photometrics, PXL 1400). Image acquisition was performed on a Macintosh Quadra 840 AV computer with the software program IPLab™ (Signal Analytics Co.). The images were then pseudocolored and merged using Adobe Photoshop™ 2.5J (Adobe Systems Inc.). Hoechst, FITC, and propidium iodide images were shown in blue, green, and red, respectively.

Chapter IV

RESULTS

1. Long-range GC% mosaic structures in the human MHC

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. As before, GC% of non-redundant genomic sequences longer than 3 kb were calculated and arranged by their genetic positions (Fig. 1); 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 are represented by rather thin vertical bars. An approximately 450-kb continuous black zone between class II HLA-DRA (abbreviated DRA) and class III CYP21 does not correspond to the GenBank sequences but to the region cloned by the bidirectional chromosome walk in this study, and will be focused on in the sections 2 and 3.

Analyzing human DNA by cesium salt buoyant density fractionation, Bernardi and his colleagues (Bernardi *et al.*, 1985; Bernardi, 1989; Bernardi, 1993) found five types of isochores with different GC% (H3, H2, H1, L2, and L1 in descending order of GC%). Figure 1 shows that most class I sequences are evidently GC-rich, belonging to the GC-richest isochore H3 (average 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 (average 40 and 45% GC, respectively). Class III sequences appear somewhat complicated, although they are richer in GC than class II sequences; the centromeric portion seems to belong to the GC-richest isochore H3 and the telomeric portion to the second GC-richest isochore H2 (average 49% GC). Confirming previous findings (Ikemura *et al.*, 1988; 1990), the boundary of long-range GC% mosaic domains, i.e., the transition between the AT-rich and GC-rich domains, is within about 450 kb, which is absent in the GenBank sequences and contains the junction between classes II and III.

The R band, 6p21.3, on which the human MHC lies, has been assigned to a T-type R band (T

band) (Holmquist, 1992). T bands are known to be composed mainly of the GC-rich H3 and H2 isochores (Bernardi, 1993; Saccone *et al.*, 1993); thus, the finding shown in Fig. 1 that three-fourths of the MHC (contiguous classes I and III) is composed mainly of these GC-rich isochores is consistent with the cytogenetic observation. It should be noted that high-resolution banding has indicated the presence of a thin Giemsa-positive subband within the human MHC (Spring *et al.*, 1985; Senger *et al.*, 1993).

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

To clone the mosaic boundary, cosmid walking from CYP21 to class II and YAC walking from DRA to class III were done (Matsumoto *et al.*, 1992; Sugaya *et al.*, 1994; Ando *et al.*, 1994; Fukagawa *et al.*, 1995a), and contiguous clones covering the 450 kb bridging classes II and III were obtained: M- and KS-series for cosmids and YDR2 and YDR3 for YACs (Fig. 2). Because YAC clones were too long for locating the GC% domain boundary precisely, in this work cosmid and λ phage libraries were constructed from the two YACs and walking from DRA to class III was performed using the libraries. About 30 kb telomeric from DRA, cosmid walking became difficult, although the reason was not clear, and λ phage walking was continued (P-series in Fig. 2), reaching the cosmids previously obtained by the walking from class III (M- and KS-series). Multicolor FISH analysis of interphase nuclei from lymphocytes of a normal male donor confirmed the contiguous array of the newly isolated λ clones with the previous cosmids.

To analyze the base-compositional distribution of the walked area, insert DNAs of 18 cosmid and 5 λ phage clones, marked by pound signs (#) in Fig. 2, were purified and digested with nuclease P1, and GC% was measured as described under Materials and Methods (Fig. 1). The area about 150 kb from DRA was fairly homogeneously AT-rich (mostly less than 40% GC, the AT-richest L1 isochore level), showing extension of AT-rich sequences from the class II side. Then a sharp transition to about 50% GC (the second GC-richest H2 isochore level) occurred;

this transition from L to H isochore was named the "L/H transition". After about 200 kb, the H2 level changed to the GC-richest H3 isochore level, previously reported as the "H2/H3 transition" (Ikemura *et al.*, 1992). The H3 level continued for about 50 kb, to class III CYP21 and beyond to at least to the centromeric half of class III. It should be noted that within individual isochores there are usually certain fluctuations of GC% between genes and their flanking regions (Ikemura and Aota, 1988; Bernardi, 1989).

3. Analysis of the L/H boundary

To analyze the structures near and at the L/H transition, the cosmid and λ clones with the transitional region spanning a total of about 80 kb have been sequenced. The following three types of characteristic structures were found by comparison of the obtained sequences with the entire GenBank data. At least 25 independent *Alu* repeats, 5 LINE-1 repeats, and one XY pseudoautosomal boundary-like sequence (designated "PABL1") were disclosed. Interestingly, most of the *Alu* repeats densely cluster in about a 20-kb region and the 5 LINE-1 repeats also cluster in a 30-kb region. The organization (Fig. 2) is as follows; **DRA (AT-rich) - 150 kb - L/H transition region (PABL1 - 30 kb of LINE-1 cluster - 20 kb of *Alu* cluster) - NOTCH 3 (moderately GC-rich) - 140 kb - H2/H3 transition (highly GC-rich) - 50 kb - CYP21**. Iris *et al.* (1993) found dense *Alu* clusters of several tens of kilobases in the telomeric portion of class III and detected all of the major *Alu* subfamilies classified by Jurka and Smith (1988). This is also the case for the *Alu* cluster near the L/H transition in this study (i.e. the most centromeric portion of class III), although the cluster has not yet been completely sequenced. The large-scale *Alu* clusters near both ends of class III, as well as GC% mosaic structures, may be related to particular genome characteristics of the MHC such as high level of polymorphism.

The density of LINEs is known to be high in AT-rich genome domains and that of *Alu* in GC-rich domains (see Table 1; Bernardi, 1989; Korenberg and Rykowski, 1988; Holmquist, 1992). In the L/H transition area, this general characteristic was accentuated by their dense clusterings. Due to the distinctive features of the pseudoautosomal boundary (PAB1), explained below, this

organization including PABL (PAB1-like sequence) may be a characteristic of certain types of long-range GC% mosaic boundaries and possibly of band boundaries. PABL1 is the first genomic sequence found to be highly homologous with the PAB1 sequence. In this study, the author will focus on characterization of this newly found sequence.

4. Characterization of PAB1-like sequences

Human sex chromosomes are divided into two functionally distinct regions, sex-specific regions and pseudoautosomal regions (Cooke *et al.*, 1985; Simmler *et al.*, 1985; Ellis and Goodfellow, 1989; Freije *et al.*, 1992; Rappold, 1993; Kvalϕy *et al.*, 1994). There are two pseudoautosomal regions (PARs) (Freije *et al.*, 1992; Rappold, 1993; Kvalϕy *et al.*, 1994): PAR1 is at distal ends of the short arms of X and Y chromosomes and PAR2, of their long arms. The existence of PAR1 (about 2.6 Mb) was deduced from observations of male meiosis, when sex chromosomes pair and form chiasmata between their short arms. X and Y sequences in PAR1 are practically identical because of the obligatory recombination event taking place at each male meiosis; sequences in PAR1 recombine between sex chromosomes while sex-specific sequences normally do not. An unusually high rate of homologous recombination in PAR1 (known to be 20-fold greater than the genome average) should be due to a special mechanism that promotes physical association and successive obligatory crossover between the sex chromosomes (Ellis *et al.*, 1990; Rappold, 1993). The interface between PAR1 and the sex-specific region is the pseudoautosomal boundary (PAB1) and therefore PAB1 is the proximal (centromeric) limit to recombination in PAR1. Ellis and Goodfellow (1989) and Ellis *et al.* (1989, 1990) reported sequences around the interface. An *Alu* element is known to be inserted in the PAB1 of Y chromosome (PABY1) but not in that of X chromosome (PABX1). Ellis *et al.* (1989) defined first the *Alu* element as the strict boundary of PAR1. A later study of PAB1 in Old World monkeys found no *Alu* repeats even on PABY1 (Ellis *et al.*, 1990), making the precise definition of PAB1 somewhat vaguer than originally thought. In Figure 3, entire sequences of approximately 450 nt reported by Ellis *et al.* (1989), except for the *Alu* element of

PABY1, are referred to as the PAB1 sequences. Figure 3 shows sequence alignment for PABX1, PABY1, and PABL1 found near the L/H transition in the MHC; the direction of PABX1 and PABY1 sequences (abbreviated PABXY1) is from telomere to centromere, i.e., from PAR1 to the sex-specific region. High homology between PABXY1 and PABL1 (close to or over 80% nucleotide identity, Table 2) spans almost the entire PABXY1 sequences, and the 3' end of the homology coincides with the reported PABXY1 homology terminus (Ellis *et al.*, 1989) where the X and Y sequences abruptly and completely diverge. In the case for the 5' terminus of PABXY1 sequences, there were no reasons to assign a specific terminal site, because the upstream 2.6-Mb sequence of PAR1 should be practically identical between X and Y chromosomes. In fact the 5' terminus of PABXY1 sequences reported by Ellis *et al.* (1989) was just an *Eco*RI site arbitrarily chosen in PAR1. However, if PABXY1 and PABL1 correspond to a certain functional signal, the 5' end of the respective signal may be defined. This will be answered in the section 6.

5. Southern hybridization analysis

If PABLs exist in the boundaries of long-range GC% mosaic domains, there should be multiple copies in the human genome. In addition, if PABLs are functionally important, they may exist in genomes of other organisms. Using the PABL1 segment as a probe, Southern blot hybridization was conducted against human and bovine genome DNAs. Southern blots of *Eco*RI, *Pst*I, or *Rsa*I digests of these DNAs were probed with a radio-labeled PABL1 segment and washed under rather stringent conditions (Fig. 4). Many bands with roughly equivalent intensities were found for human DNA but several strong bands with different intensities for bovine DNA, showing that many PABLs exist in both human and bovine genomes. Existence of homologous sequences in the bovine genome indicates their evolutionary stable maintenance and biological significance.

6. Core sequences of genomic PABLs

For analyzing characteristics of PABLs, the cosmid library of the human genomic DNA was

screened by using the PABL1 segment as a probe. Positive signals far exceeded the numbers having been observed for a single copy gene (e.g., NOTCH3 and TN-X in Fig. 2) that was obtained by screening the same cosmid library (Matsumoto *et al.*, 1992; Sugaya *et al.*, 1994); this finding is consistent with results of Southern blot hybridization against human genomic DNA. About 100 independent cosmid clones harboring PABL sequences were isolated and 10 clones were randomly selected for the following analyses. Their *Eco*RI, *Bam*HI, *Pst*I, or *Hin*dIII fragments that hybridized with PABL1 were subcloned and sequenced. All examined clones gave PABL sequences and were independent. Two of them (PABL2 and 3) were analyzed in detail. Figure 5 shows multiple alignment of the PABLs including PABXY1 sequences and PABLSp2 described in the section 7, which was calculated according to Hein (1990). Homology among PABLs, as well as with PABXY1, was high (about 80% nucleotide identity; Table 2) and spanned almost the entire region of PABXY1. Importantly, the 3' terminus of their homology corresponds approximately to the PABXY1 homology terminus reported by Ellis *et al.* (1989) where the X- and Y-chromosome sequences completely diverge.

In the case for the 5' terminus of PABXY1, as noted above, there were no reasons to assign a specific terminal site because of the PAR's nature, and the 5' terminus of the reported PABXY1 (Ellis *et al.*, 1989) was an *Eco*RI site arbitrarily chosen. Therefore, even if PABXY1 and PABLs have certain biological functions, the 5' end of the presumable functional signal may not be found by comparing only the X and Y sequences. In other words, comparison with and/or among PABLs can define the 5' terminus of the hypothesized signal. Homology among the three previously-analyzed PABLs (PABL1-3) actually extended upstream of the *Eco*RI site and ceased abruptly about 200 nt upstream. Thus, Fukagawa *et al.* (1995a) proposed that a region of about 650 nt including the extended 200 nt is presumably the complete form of PABLs. Recently Dr. N. A. Ellis kindly supplied me with his unpublished PAR1 sequence upstream of the *Eco*RI site. It should be noted that PABXY1 in Fig. 5 includes this sequence. Confirming the author's proposal, multiple alignment of six PABLs (Fig. 5) including the PABXY1 showed that the ca. 650 nt proposed is actually the core of PABLs, i.e., the sequence between the two arrows in Fig. 5. Conservation of precise termini is possibly due to their biological functions.

Ellis *et al.* (1990) reported that, although the *Alu* element itself is absent in PABY1 of Old World monkeys, sequences corresponding to the *Alu* insertion site are stably conserved among Old World monkeys and hominoids. Ten nucleotides at the respective site are also conserved for four PABLs (Fig. 5) in which the *Alu* element is absent. Homology levels of PABX1 and PABY1 are known to differ upstream and downstream to the *Alu* insertion site (Ellis *et al.*, 1989; Ellis *et al.*, 1990), that is also the case for PABLs (Table 3). This will be analyzed in detail in the section 9. Figure 6 summarizes their structural organization.

7. The transcripts of PABLs

Although multiple copies of PABLs were indicated by Southern blot hybridization against genomic DNA and by cosmid cloning, genomic sequences in the GenBank databases showing significant homology with the PABLs were confined to PABXY1. In the case of EST (Expressed Sequence Tag) sequences, however, two human ESTs, hbc671 (99 nt, Accession No. T11103) and HSAAAAWAH (266 nt, Accession No. Z19872), were previously found to have evident homology with separate portions of PABLs (Fukagawa *et al.*, 1995a). DNA sequences in the databases (DDBJ / EMBL / GenBank) have since accumulated significantly, and therefore human sequences in a recent GenBank (Release 87 including update data; 1995) were reexamined. Additional nine ESTs showing high homology with separate portions of PABLs (ca. 80% nucleotide identity) were found, supporting the previous supposition by Fukagawa *et al.* (1995a) that some, if not all, PABLs are transcribable. To clarify the characteristics of the predicted PABL transcripts, six human cDNA libraries constructed from different tissues and cells (placenta, monocyte, B-cell, spleen, and skin fibroblast; refer to Materials and Methods) were screened using the PABL1 segment as a probe. Positive clones were obtained from all six libraries, and insert sizes of the 20 cDNA clones analyzed were longer than the 650 nt, i.e., the size of the genomic PABL core. The author sequenced the following six cDNA clones: Mo1 and Mo2 from the monocyte library; Bc4 from the B-cell library; Sk13 from the skin library; Sp2 and Sp3 from the spleen library. Sequence data are presented in Appendix and have been deposited with the International Data Libraries DDBJ /

EMBL / GenBank under Accession Nos. D55638 (Bc4), D55639 (Mo1), D55640 (Mo2), D55641 (Sk13), D55643 (Sp2), and D55644 (Sp3).

It should be noted that a portion of the Mo2 cDNA sequence was practically identical to that of the above-mentioned EST, HSAAAAWAH. The identity spanned the entire region of 266-nt HSAAAAWAH (i.e., not only the PABL sequence but also its flanking region), showing the HSAAAAWAH to correspond to a portion of the Mo2 transcript itself or a very closely related transcript. Alignment around the 5' terminal portion of PABLs in the cDNA and two ESTs (T92306 and R12279) is listed in Fig. 7. Direction of the sequences is the same as that used for the genomic PABLs and, for comparison with the genomic PABLs, the PABL1 and PABX1 sequences are also presented. The homology terminus of the 5' portion of these transcribed PABLs corresponds closely to that defined for genomic PABLs. An arrow shows the 5' core terminus defined by genomic PABLs in Fig. 5.

Figure 8 shows an alignment around the 3' portion of transcribed PABLs and an EST (T47905). Again, the homology terminus of these transcribed PABLs corresponds closely to that defined for genomic PABLs. Therefore, results of Figs. 3, 5, 7, and 8 showed conservation of both termini of genomic and transcribed PABLs. Structural organizations of the transcribed PABLs and their flanks are summarized in Fig. 9.

Open reading frames (ORFs) with significant sizes could not be found in the obtained cDNAs, not only for the PABL core sequences but also for their flanking sequences. When these cDNAs were searched using the BLASTX program against the non-redundant protein sequence database compiled by the Human Genome Center of Japan, no significant homology with known proteins was detected. GRAIL, a computer program trained to identify protein-coding ORFs in human DNA (Uberbacher and Mural, 1991), also could not detect reliable protein-coding capacity. These may suggest that the functional form of the transcripts is the RNA molecule. Concerning the functional RNA molecules ever known, many of them have stable and characteristic secondary structures. Taking this into account, possible secondary structure for PABL cDNA sequences was calculated with an algorithm developed by Zuker (1989) using the GCGFOLD program managed by University of Wisconsin Genetics Computer Group

(UWGCG). The author first generated 100 randomized sequences with the same base-composition to each cDNA using the SHUFFLE program in the UWGCG, and calculated the most stable secondary structure both for the cDNA and for each of the randomized sequences. Table 4 lists the lowest energy level for each cDNA, as well as the average for the randomized sequences along with their standard deviation. Data obtained for the genomic PABLs are also listed. All cDNA and genomic sequences had significantly lower energy levels than the averages obtained for the randomized sequences, and differences of energy levels ranged 0.9-4.9 units of standard deviations for the respective randomized sequences (Table 4). IL4 and CYPDB1 mRNAs were similarly analyzed, and their energy levels were not significantly lower than the averages of the respective randomized sequences. On the other hand, the energy level found for 7SL RNA and 12S rRNA which are known to form stable secondary structures were significantly lower than the averages of the randomized sequences, and the differences were equivalent to those for PABLs (Table 4). These results suggest that PABL transcripts can form stable secondary structures.

When structures predicted for individual cDNA sequences were investigated, the PABL core portion was usually folded within itself separating from its flanking sequences. For example, the secondary structure of Sp2 cDNA (Fig. 10) shows that a major portion of the PABL core (nucleotide number 1-629) is folded separately from the flanking sequence (nucleotide number 630-1150). Figure 11 also shows a possible secondary structure of the PABL consensus sequence (see the section 9) calculated by the MFOLD program of UWGCG.

Sequences of the six cDNA clones were found to be distinct from those encoded by the previously characterized genomic PABLs (PABL1, PABL2, and PABL3; Fukagawa *et al.*, 1995). To know the genomic structures of the transcribed PABLs, a λ phage library of the human genomic DNA (Tomatsu *et al.*, 1989) was screened with a probe of the Sp2 cDNA fragment deprived of the PABL core. The genomic PABLSp2 sequence presented in Fig. 5 corresponds to the one thus isolated. The sequences of the 5' flank of about 50 nt and of the 3' of about 300 nt so far known for the Sp2 cDNA were identical to those of the genomic PABLSp2, showing there were no intron/exon structures near this PABL core. This sequence is

presented in Appendix and has been deposited with DDBJ / EMBL / GenBank under accession number D55642 (PABLSp2).

8. Northern hybridization analysis

PABL sequences of the six cDNA clones randomly selected were independent, showing many PABLs to be transcribed. Sizes of the cDNAs were larger than the PABL core. To estimate the intact sizes of PABL transcripts, northern blot analysis of human total or polyA$^+$ RNA fraction was conducted using the PABL1 probe. Figure 12A shows the result for the total RNA extracted from GM01416D cells. Broad bands mainly with mobilities slower than 28S rRNA, estimated to be 5-10 kb in length, were detected. The results for the polyA$^+$ RNA fractions were practically the same to the total RNA (Fig. 12B). All samples examined in Fig. 12 gave positive signals, though expression levels seemed to differ from each other. It should be noted that, in RT-PCR reactions for those RNA samples, all primer sets designed based on the six cDNA sequences gave clear positive bands with the expected sizes (data not shown). Observed broad bands of Figs. 12A and 12B should correspond to a large number of distinct transcripts hybridized with PABL sequences, and their long sizes were consistent with the finding that sizes of the PABL cDNAs were larger than the PABL core. To detect a single transcript corresponding to a unique cDNA, a Sp2 cDNA fragment deprived of its PABL core was used as a hybridization probe, and in this case a single sharp band of about 7.5 kb was detected (Fig. 12C).

9. Phylogenetic relationship among PABLs and PABXY1 and their consensus sequence

Obligatory pairing and crossover in the pseudoautosomal region (PAR) ensure accurate segregation of the sex chromosomes during male meiosis. An unusually high rate of homologous recombination in PAR1 (known to be 20-fold greater than the genome average) should be due to a special mechanism that promotes physical association and the successive obligatory crossover between the sex chromosomes. The strict limit to terminate the high-

frequency recombination in PAR1 is the pseudoautosomal boundary 1 (PAB1). An *Alu* element is known to be inserted in the PAB1 of the human Y chromosome (PABY1) but not in that of the X chromosome (PABX1) (Ellis *et al.*, 1989). Ellis *et al.* first defined the *Alu* element as the strict boundary of PAR1. A later study on the boundaries of PAR1 in Old World monkeys, however, found no *Alu* repeats on their PABY1 (Ellis *et al.*, 1990). In spite of lack of the *Alu* element, the *Alu*-insertion site itself was proposed as the strict boundary of PAR1 for the following reasons. In pairwise comparisons between the X and Y boundary sequences within each species of several hominoids and Old World monkeys, Ellis *et al.* (1990) found about 220-nt sequences downstream to the *Alu*-insertion site to be more divergent between the X and Y chromosomes than the sequences of the *Alu*-upstream region (~ 78% vs. ~97% nucleotide identity) whether or not the *Alu* element is present. By extensively analyzing the nucleotide substitution patterns, they concluded that the *Alu*-insertion site, which corresponds to the abrupt transition between the high- and reduced-homology regions, is the strict limit for the high-frequency recombination in PAR1 and thus the exact boundary of PAR1. The position of the boundary was practically the same in Old World monkeys and hominoids. This shows that the limit of the PAR recombination was situated at the *Alu*-insertion site before divergence of Old World monkey and great ape lineages and that an *Alu* element was inserted to the preexisting boundary of Y chromosome in the great ape lineage. According to their notion, this site is called the "*Alu*-insertion site" in this study whether or not the *Alu* element is present. The downstream of the reduced-homology region has no homology between X- and Y-sequences, i.e., these are sex-specific sequences (Fig. 6).

The *Alu*-insertion site is an exact interface which differentiates sex chromosomes into two functionally distinct regions. Sequences around the *Alu*-insertion site are known to be strictly conserved among species of hominoids and Old World monkeys. At the respective insertion site, a ten-nucleotide sequence is strictly conserved even for PABLs though the *Alu* element is again absent (see Fig. 5). Exact sequence conservation around this *Alu*-insertion site, as well as conservation of the 5' and 3' termini, is thought to be related to possible functions of PABXY1 and PABLs. For the studies of evolutionary processes involved in the formation of the present

PABXY1 and PABLs and of the functional constraints on the sequences, phylogenetic relationship and evolutionary rate were examined. In order to estimate the evolutionary rates, the reported PABXY1 sequences of great apes and Old World monkeys (Ellis *et al.*, 1990) were included for the analysis. Divergence between great apes and Old World monkeys is postulated here to be 25 million years ago. As noted above, PABX1 and PABY1 sequences upstream of the *Alu*-insertion site are highly homologous within a single species because of the PAR's nature, e.g. 99% nucleotide identity between human PABX1 and PABY1. To avoid inevitable complications derived from the PAR's characteristics, individual PABLs and PABXY1 were divided into two regions according to Ellis *et al.* (1990); 177-nt upstream and 155-nt downstream to the *Alu*-insertion site (for details, see the legend of Fig. 13). Unrooted phylogenetic trees were constructed by using the neighbor-joining method (Fig. 13). Trees obtained from evolutionary distances based on one- and two- parameter methods were identical in their branching patterns or topologies, and branch lengths were nearly the same. The root of the trees of Fig. 13 was tentatively predicted by the UPGMA method.

The downstream portion, and thus the non-PAR portion for the sex chromosomes, was first analyzed. PABX1 and PABY1 of all species are clearly separated into two distinct groups and the topology within each group well reflects the known phylogenetic relationship of the species (Fig. 13A). Interestingly, the distance between PABX1 and PABY1 is similar to the distance of PABX1 or PABY1 with PABLs. This is consistent with the notion of Ellis *et al.* (1990) that the genetic contact resulting in sequence homogenization between sex chromosomes did not occur in the *Alu*-downstream portion after the divergence of great ape and Old World monkey lineages and that the strict limit of PAR1 recombination, i.e., the *Alu*-insertion site, was situated at this site in the ancestral species to all extant higher primates. Using the evolutionary distance between human and Old World monkey PABX1 as well as that for PABY1, the divergence time of PABXY1 sequences and also of PABLs was estimated to be 60-120 million years. This is consistent with our previous finding that PABLs are present in the bovine genome (Fukagawa *et al.*, 1995a). Evolutionary rates of individual PABLs based on the divergence time thus obtained were estimated (Table 5).

The author then analyzed the upstream portion to the *Alu*-insertion site (Fig. 13B). Because of the PAR's nature, PABX1 and PABY1 sequences do not separate into two distinct groups. Evolutionary distances for several PABLs such as PABL2, PABL3, and Sp3 differ significantly between the upstream and downstream portions to the *Alu*-insertion site (Fig. 13A vs. 13B). This suggests that the *Alu*-insertion site corresponds to a functional and/or recombinational interface within PABLs as with PABXY1 although the *Alu* element itself is absent. Evolutionary rates for the upstream portion, which were estimated using the PABL's divergence time found for the downstream portion, are also presented in Table 5.

The evolutionary rates of some PABLs were estimated to be much smaller than 1×10^{-9} substitutions per site per year (e.g., 0.1×10^{-9} substitutions per site per year for the *Alu*-upstream portion of Sp3). This indicates that the sequences evolved at a significantly slower rate than typical non-coding regions, since the averages of evolutionary rates of mammalian pseudogenes and of introns of transcribed genes were estimated to be 4.9×10^{-9}, and 3.7×10^{-9} substitutions per site per year, respectively (Li *et al.*, 1985). These results suggest that PABLs and PABXY1 sequences would have been under selective constraints and have functions.

The author also analyzed a ca. 300-nt X-specific sequence immediately downstream of PABX1 (Fig. 6), which were reported for both hominoids and Old World monkeys (Ellis *et al.*, 1990). Pairwise-comparisons of the X-specific sequences between the species, as well as respective comparisons for PABX1 sequences (the upstream or downstream portion to the *Alu* insertion site), were conducted to estimate nucleotide divergence (%) in the regions. For 38 out of 42 pairs, nucleotide divergence (%) in the PABX1 sequences were lower than those of the X-specific sequence, supporting the exertion of evolutionary and functionary constraints on the PABX1 sequence (Table 6). In the case for the Y-specific region immediately downstream of PABY1, available sequences were rather short (ca. 100 nt), though *Alu* sequences inserted in individual PABY1 were known for hominoids. When nucleotide divergence for the composite sequences of these non-contiguous regions were analyzed, a tendency similar to for the X-specific sequences was found (data not shown), again supporting the supposition of

evolutionary constraints on the PAB1 sequence.

More than ten PABL sequences, i.e. genomic PABLs, PABXY1, and transcribed PABLs, have become available, and the homology among them was close to or over 80% nucleotide identity. This high homology allowed deduction of their consensus sequence of 646 nt (Fig. 14A); a unique base could be assigned to a 97% portion of the 646 nt. The tree topology in Fig. 13 shows that this consensus sequence is most likely related to their ancestor sequence. The consensus sequence thus obtained was aligned with Sp3 (Figs. 14B and C). Divergence found for the *Alu*-downstream portion was higher than that for the upstream portion. This is consistent with the results of Fig. 13 and Table 5, which show that the evolutionary rate of Sp3 for the *Alu*-upstream portion is significantly slower than that for the downstream portion. This indicates that the *Alu*-insertion site constitutes a functional interface within a PABL, which may relate with recombination and/or regulation of DNA replication as discussed later.

10. Chromosome in situ *hybridization*

To examine the chromosome locations of other PABLs, a standard fluorescence *in situ* hybridization (FISH) onto metaphase chromosomes was conducted (Fig. 15A-D). Table 7 lists map locations of ten PABLs and band characteristics. The results in Table 7 show that most of the known locations of PABLs including PABXY1 are on T-type and/or terminal R bands which have been shown to correspond to the evidently GC-rich genome portions (Ikemura and Wada, 1991; Holmquist, 1992; Bernardi, 1993) and no PABLs are on G bands at a standard banding level.

Chapter V

DISCUSSION

1. Boundary of long-range GC% mosaic domains

The human MHC containing PABL1 spans about 4 Mb, exceeding the average size of bands observed through high-resolution banding such as with a 2000-band level. At a standard 850-band level, the MHC is on a wide R band 6p21.3 which has been assigned to a T-type R band (Holmquist, 1992). The T-type R bands (T bands) are an evidently heat-stable subgroup of R bands and known to correspond to GC-rich regions (Ikemura and Wada, 1991; Bernardi *et al.*, 1993). These are called T bands since many of them are terminal R bands (the first R band from a telomere). By a higher resolution banding, a thin Giemsa-positive subband 6p21.32 was located within the MHC (Spring *et al.*, 1985; Senger *et al.*, 1993). Detailed base-composition analysis in Fig. 1 shows a telomeric portion of class II, including a junction with class III, to be most AT-rich in the MHC. The author predicts that the genome portion containing at least this evidently AT-rich 200-kb region corresponds to the thin Giemsa-positive subband.

Considering the wide range of functional behaviors of chromosome bands and GC% mosaic domains, their boundaries are most probably composed of multiplex signals and structures ensuring multiple functions. Characteristic structures such as PABL1 and large-scale clusters of two distinct types of repetitive sequences (*Alu* and LINE-1) were disclosed around a long-range GC% mosaic boundary. If these characteristics are found for other domain boundaries, it will add to our comprehensive understanding of DNA sequences and of chromosome structures. The present work is the first study to characterize such mosaic boundary at a nucleotide sequence level and, therefore, it is difficult to generalize the findings. In this situation, it is worthwhile to consider the band structures and GC% distribution around PAB1 of the short arms of the human sex chromosomes. PAR1 has been assigned to an R-band (Xp22.33 and Yp11.32), and judging from the high density of CpG islands, a major portion of 2.6 Mb of PAR1 would be GC-rich (Ellis and Goodfellow, 1989; Rappold, 1993; Schlessinger *et al.*, 1993). The boundaries with the neighboring Giemsa-positive subbands, Xp22.32 and

Yp11.31, appear rather close to PABXY1 (Rappold, 1993; Schlessinger, 1993). PABXY1 and their neighboring sex-specific sequences, including SRY (about 5 kb apart from PABY1), are known to be AT-rich (Ellis *et al.*, 1990; Sinclair *et al.*, 1990). The genome characteristics around PAB1 thus suggests that PABXY1 is near a boundary of the long-range GC% mosaic domains and of chromosome bands (Fig. 16). Ellis *et al.* (1990) noted that the region near PAB1 is densely populated with *Alu* repeats while chromosome Y as a whole is rather deficient in the repeats. Therefore, genome characteristics appear fairly similar in regions around PABXY1 and PABL1, indicating a certain generality of findings for the MHC. In order to directly characterize genome characteristics of regions having other PABLs, Nakamura (a member of the laboratory to which the author belongs) recently characterized a 3 Mb portion containing PABL2, by analyzing YAC and cosmid clones containing PABL2. He found the 3 Mb region to be composed of long-range GC% mosaic domains and the PABL2 to be situated near the domain boundary (should be published by Nakamura, Y., Fukagawa, T., and Ikemura, T.). The high density of *Alu* and LINE-1 repeats was also found near PABL2. Therefore, genome characteristics around PABL2 are analogous to those around PABL1 and PABXY1. The results obtained by analyzing GC% distribution around PABL1, PABL2, or PABXY1 suggest that PABLs generally exist near boundaries of long-range GC% mosaic domains as proposed by Fukagawa *et al.* (1995b).

It should be noted that not only the PABL core but also the cosmid-cloned fragments (30-40 kb) containing PABLs were found AT-rich, indicating PABLs belong to AT-rich domains. However, a major portion of PABLs mapped are situated on T-type and/or terminal R bands at a standard banding level (Table 7), which are known to be GC-rich. At a higher resolution level, the respective bands (e.g., 19q13.3, 11q25, 21q22.3) have internal narrow G bands as for 6p21.3 harboring the MHC (Yunis, 1981). Characteristics of the genome portions containing PABLs appear again to resemble each other. This may relate to PABL's functions. More than 2000 stained bands of human chromosomes have been observed through a high resolution banding (Yunis, 1981). Copy number of PABLs detected by the hybridization against the genomic DNA appears much smaller than the estimate number of all band boundaries, as well

as of GC% mosaic boundaries (Fukagawa *et al.*, 1995a). Certain types of boundaries (e.g., boundary of an AT-rich domain adjoining an evidently GC-rich genome portion such as that belonging to a T-band) may have PABLs or there may be many PABL-type sequences undetected by hybridization.

2. Possible functions of PABLs and PABXY1

Functionally important molecules evolve more slowly than less important molecules (Kimura, 1983). PABLs are thought to have biological functions since their evolutionary rates are significantly slower than non-functional regions (Tables 5 and 6). With regard to the functions of boundaries of GC%-mosaic domains and of chromosome bands, a switching of DNA replication timing is an important candidate (Holmquist, 1987; Holmquist, 1992; Craig and Bickmore, 1993). With the contiguous λ phage and cosmid clones of the MHC region (Sugaya *et al.*, 1994; Fukagawa *et al.*, 1995a) as fluorescent probes, FISH to interphase nuclei was conducted for analyzing DNA replication timing in this region. The principle of the analysis is explained in Fig. 17 (Selig *et al.*, 1992) and the results are presented in Fig. 18. It is shown that the GC-rich class III side replicates earlier than the AT-rich class II side and PABL1 is located in a switching region of DNA replication timing (to be published by Okumura, K., Hishida, K., Nogami, M., Taguchi, H., Fukagawa, T., Sugaya, K., Matsumoto, K., Ando, A., Inoko, H., and Ikemura, T.). A methylated and transcriptionally inactive X chromosome replicates very late, while PAR1 which escapes X inactivation replicates earlier (Goodfellow *et al.*, 1984; Schiebel *et al.*, 1993). A switch in DNA-replication timing, functioning at least in the inactivated X chromosome, presumably occurs near PAB1. PABLs including PABXY1 may be related to possible signals for DNA replication timing such as a pausing signal for the early replication.

Obligatory pairing and crossover along the PAR1 ensure accurate segregation of the sex chromosomes during male meiosis. Unusually high rate recombination in PAR1 between sex chromosomes (known to be 20-fold greater than the genome average) terminates abruptly at PAB1. There may exist a possibility that PABLs are involved in a process of recombination.

Postulating an illegitimate recombination between two PABLs, a model of evolutionary formation of the present pseudoautosomal boundaries will be proposed in the section 4.

3. Characteristics of PABL transcripts

PABX1 was found to be located within an intron of PBDX (pseudoautosomal-boundary divided on the X chromosome) gene, which encodes the Xg^a antigen and was recently renamed XG (Ellis *et al.*, 1994a; 1994b). When RT-PCR (RNA PCR) analyses were conducted on human total RNA of GM01416D or HeLa cells from which the contaminant genomic DNA was completely removed by using DNase I, the primers designed for the present cDNAs (one primer is within PABL core and the other is in its flanking sequence) reproductively produced clear PCR-products with the expected sizes under standard conditions recommended by Perkin-Elmer's protocol (data not shown). However, the primers designed from PABX1 and its flanking sequence only occasionally showed a very faint band with the expected size under the same or modified conditions. RNA molecules corresponding to the present cDNAs are presumably more abundant and/or have longer life-times than the XG intron RNA. In this connection, it should be noted that a portion of the Mo2 cDNA sequence was practically identical with the entire region of the 266-nt EST, HSAAAAWAH (i.e., not only the PABL sequence but also its flanking sequence), indicating that Mo2 transcripts are abundant. Sequences of three recently reported ESTs (R12279, R07404, and T99392 derived from at least two independent cDNA libraries; Hillier *et al.*, 1995) were found practically identical to each other, indicating that the respective transcripts are also abundant. Figure 12 shows that the abundance of Sp2 transcripts presumably differs among tissues.

Figure 9 shows that the cDNA clone having both edges of the PABL core was only Sp2 cDNA and an edge of other cloned inserts was within the PABL core. In the case of Mo2 cDNA which was predicted to be abundant, one edge was linked with polyA. Since a potential polyadenylation signal, AATAAA, is located 16-nt upstream of the polyA, this edge is thought to be formed by polyA-addition. This may be related to its abundance and/or stability. Another possibility for the lack of two intact core edges in the cDNAs may be due to the stable

secondary structures predicted for PABLs, which presumably inhibit efficient cDNA extension with reverse transcriptase. The latter possibility is consistent with the finding that sizes of PABL transcripts are much larger than that of the PABL core, e.g., approximately 7.5 kb for the Sp2 transcript (Fig. 12C). It should be noted that PABL1 and PABXY1 were both found in the studies searching for molecular signals differentiating global characteristics in the human genome. PABL transcripts with large sizes may be necessary for recognizing and/or differentiating global characteristics of the genome. In this connection, it is worthwhile to mention that XIST RNA with a large size of 17 kb is believed to play an important role in determining and differentiating global characteristics in the inactivated X chromosome (Brown *et al.*, 1992).

4. Evolutionary processes in forming sex-chromosome PABs

Recently, Ellis *et al.* (1994a) proposed a model for the evolutionary formation of the present-day pseudoautosomal boundary of the short arms, PABXY1. They hypothesized a pericentric inversion of the Y chromosome with one break point in earlier XG and the other in 5 kb distal to the SRY being supposed to be on the earlier long arm (Fig. 19 legend). The author here proposes a model in which the hypothesized inversion occurred utilizing an illegitimate recombination between the two PABL elements, one PABL in the earlier XG and the other near the earlier SRY; before the illegitimate recombination, the two elements on the earlier Y chromosome were ordinary PABLs (Fig. 19A). After the recombination, the characteristic of a pseudoautosomal boundary was acquired and the recombinant PABL became PABY1 (Fig. 19B and C). Ellis *et al.* (1990) presented an evidence that genetic contact resulting in homogenization between the sex chromosomes did not occur in the *Alu*-downstream region after divergence of Old World monkey and great ape lineages. The author's model proposed here and the results in Fig. 13 are consistent with their findings.

It should be stressed that the model proposed here is clearly distinct from the "attrition" model proposed for formation of the strict boundary of PAR (Ellis *et al.*, 1990). Ellis *et al.* mentioned

two types of models for this formation. One is based on genome rearrangements such as insertions, deletions, inversions, and translocations. The above-proposed process postulating the illegitimate recombination between two PABLs belongs to this category. The other model is based on the following "attrition" process. Recombination acts to maintain homology in PARs between the sex chromosomes, but this event is infrequent in the sequences very close to the pseudoautosomal boundary. The model predicts that when enough sex-specific differences accumulate in the region immediately distal to the boundary, the probability of recombination in the region with mismatches is reduced. Once recombination is limited, divergence between X and Y chromosomes accumulates more rapidly until recombination events no longer include the region of mismatches, resulting in formation of a new strict boundary. In this latter model, the reduced-homology regions of PABXY1 is considered to correspond to the attrited portion derived from a single PABL. When based on this model, the most probable explanation for the reason why the 3' homology terminus of PABXY1 is practically identical to those of PABLs is a strong functional constraint to preserve the 3' terminal position during course of evolution. Furthermore we should introduce an extra mechanism to put an evolutionally-stable interface between the high- and reduced-homology regions, i.e., the *Alu*-insertion site, during the attrition process. The explanation for the reduced-homology supposed in the model of Fig. 19 is distinct from the attrition model and is mainly based on the finding that the diversity between PABX1 and PABY1 is similar to the diversity of either PABX1 or PABY1 with PABLs (Fig. 13A). Table 3 shows that the homology between PABX1 and PABY1 (77.8% nucleotide identity) is lower than the homology between PABX1 and PABL2 (82.2%) or that between PABY1 and PABL1 (81.4%), indicating separate origins for the PABX1 and PABY1 downstream-sequences to the *Alu*-insertion site. In this connection, it is worthwhile to consider the sequence organization of the boundary of PAR2, the pseudoautosomal regions of the long arms of the human sex chromosomes. At the breakpoint of the X-Y homology in PAR2, Kvalϕy *et al.* (1994) found a portion of LINE-1 sequence (ca. 780 nt), and they proposed that an illegitimate recombination between two independent LINEs on the earlier X and Y chromosomes was involved in the formation of the present PAR2 boundary (PABXY2). The

reported LINE-1 sequences in PABX2 and PABY2 show about 92% nucleotide identity. One explanation for the reduced homology may be the attrition derived from a single LINE-1 sequence. To test this possibility, the author used the 780 nt of LINE-1 sequence of X or Y chromosome in searching GenBank sequences. Interestingly, the LINE-1 of the X chromosome showed 98% identity with LINE-1 of HSL1G (Dombroski *et.al.*, 1993) and that of the Y chromosome showed 96% identity with LINE-1 of HSRETBLAS (Toguchida *et al.*,1993), indicating separate origins for the LINE-1 sequences in PABX2 and PABY2. This finding is not consistent with the attrition model that postulates these two LINE-1 sequences with 92% identity to have been derived from a single LINE-1 but consistent with the model that the reduced homology was derived from an illegitimate recombination between two distinct LINEs. Boundaries of both PAR1 and PAR2 are therefore considered to have been produced by an analogous process, that is an illegitimate recombination between repetitive elements; PABLs for PABXY1 and LINEs for PABXY2. Analogous processes may also have been involved in forming the present-day human genome which is composed of GC% mosaic structures.

5. Conclusion

In this study, the author cloned and characterized the 450 kb containing the junction of MHC classes II and III in which a possible boundary of the Mb-level GC% mosaic domains had been located. The author first disclosed the boundary as a sharp GC% transition and found in the domain boundary, a sequence highly homologous with the pseudoautosomal boundary (PAB) sequence of human sex chromosomes. The author designated the sequence "PABL" and found many PABL-type sequences in the human genome. Analyzing a total of eleven PABLs (five genomic and six cDNA sequences), a ca. 650 nt of the PABL consensus sequence could be defined; a strict conservation of the 3' and 5' edges of the PABLs was found. Northern blot analysis showed the sizes of PABL transcripts to be 5-10 kb in length. The divergence time of PABLs was estimated to be 60-120 million years by analyzing five human PABLs and sex-chromosome PABs of seven primates. Deduced evolutionary rates showed PABLs to have been under selective constraints. A model for evolutionary formation of the present-day

pseudoautosomal boundary was proposed by postulating an illegitimate recombination between two PABLs (Fig. 19).

REFERENCES

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. *Nucleic Acids Res.*, **14**: 6345-6355, and 8702 (Erratum).

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 isochores organization of the human genome. *Annu. Rev. Genet.*, **23**: 637-661.

Bernardi, G. (1993) The isochores 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.

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

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. (1992) The human XIST gene: Analysis of 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**: 527-542.

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.

Chomczynski, P. and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-PhOH-chloroform extraction. *Anal. Biochem.*, **162**: 156-159.

Coming, D. E. (1978) Mechanisms of chromosome banding and implications for chromosome structure. *Annu. Rev. Genet.*, **12**: 25-46.

Cooke, H. J., Brown, W. R., and Rappold, G. A. (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature*, **317**: 687-692.

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.

Dombroski, B. A., Scott, A. F., and Kazazian H. H. (1993) Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl. Acad. Sci. U. S. A.*, **90**: 6513-6517.

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.

Ellis, N. A., Ye, T.-Z., Patton, S., German, J., Goodfellow, P. N., and Weller, P. (1994a) Cloning of *PBDX*, an *MIC2*-related gene that spans the pseudoautosomal boundary on chromosome Xp. *Nature Genet.*, **6**: 394-400.

Ellis, N. A., Tippett, P., Petty, A., Reid, M., Weller, P. A., Ye, T. Z., German, J.,

Goodfellow, P. N., Thomas, S., and Banting, G. (1994b) *PBDX* is the *X G* blood group gene. *Nature Genet.*, **8**: 285-290.

Freije, D., Helms, C., Watson, M. S., and Donis-Keller, H. (1992) Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science*, **258**: 1784-1787.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H., and Ikemura, T. (1995a) 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.

Fukagawa, T., Nakamura, Y., Okumura, K., Nogami, M., Ando, A., Inoko, H., Saitou, N., and Ikemura, T. (1995b) Human pseudoautosomal boundary-like sequences (PABLs): expression and a model of formation of present-day human pseudoautosomal boundary of sex chromosomes. *Hum. Mol. Genet.* (in press).

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

Gardiner, K. (1995) Human genome organization. *Curr. Opi. Genet. Dev.*, **5**: 315-322.

Goodfellow, P., Pym, B., Mohandas, T., and Shapiro, L. J. (1984) The cell surface antigen locus, *MIC2X*, escapes X-inactivation. *Am. J. Hum. Genet.*, **36**: 777-782.

Hein, J. (1990) Unified approach to alignment and phylogenies. *Meth. Enzymol.*, **183**: 626-645.

Hillier, L., Clark, N., Dubuque, T., Elliston, K., Hawkins, M., Holman, M., Hultman, M, Kucaba,T., Le, M., Lennon ,G., Marra, M., Parsons, J., Rifkin, L., Rohlfing, T., Soares, M., Tan, F, Trevaskis, E., Waterston, R., Williamson, A., Wohldmann, P., and Wilson, R. (1995) Unpublished; direct submission of EST sequences to GenBank database.

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. (1989) Evolution of chromosome bands: molecular ecology of noncoding DNA. *J. Mol. Evol.*, **28**: 469-486.

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

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 Univ. press, Oxford, U. K.

Imai, T. and Olson, M. V. (1990) Second-generation approach to the construction of yeast artificial-chromosome libraries. *Genomics*, **8**: 297-303.

Inoko, H., Ando, A., Kimura, M., and Tsuji, K. (1985) Isolation and characterization of 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., Rodigues-Tome, P., Claverie, M., Dausset, J., and Cohen, D. (1993) Dense *Alu* clustering and a potential new member of the *NFκB* family within a 90 kilobase HLA class III segment. *Nature Genet.*, **3**: 137-145.

Jurka, J. and Smith, T. (1988) A fundamental division in the *Alu* family of repeated sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **85**: 4775-4778.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro H.N. (ed) Mammalian Protein Metabolism, pp. 21-132. Academic Press, New York.

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.

Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**: 111-120

Kimura, M. (1983) The neutral theory of molecular evolution., Cambridge Univ. press, Cambridge, U. K.

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 .

Kvalφy, K., Galvagni, F., and Brown, W. R. A. (1994) The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. *Hum. Mol. Genet.*, **3**: 771-778.

Li, W.-H., Luo, C.-C., and Wu, C.-I. (1985) Evolution of DNA sequences. In MacIntyre, R. J. ed. Molecular evolutionary genetics. pp. 1-94. Plenum press, New York.

Lichter, P., Cremer, T., Borden, J., Manuelidis, L., and Ward, D. C. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by *in situ* suppression hybridization using recombinant DNA libraries. *Hum. Genet.*, **80**: 224-234.

Matsumoto, K., Arai, M., Ishihara, N., Ando, A., Inoko, H., and Ikemura, T. (1992) 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., 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.

Nei, M. (1987) Molecular evolutionary genetics, Columbia Univ. press, New York.

Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.*, **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.

Rappold, G. A. (1993) The pseudoautosomal regions of the human sex chromosomes. *Hum. Genet.*, **92**: 315-324.

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

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

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

Schiebel, K., Weiss, B., Wohrle, D., and Rappold, G. (1993) A human pseudoautosomal gene, ADP/ATP translocase, escapes X-inactivation whereas a homologue on Xq is subject to X-inactivation. *Nature Genet.*, **3**: 82-87.

Schlessinger, D., Mandel, J.-L., Monaco, A. P., Nelson, D. L., and Willard, H. F. (1993) Report of the fourth international workshop on human X chromosome mapping 1993. *Cytogenet. Cell Genet.*, **64**: 148-163.

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

Selig, S., Okumura, K., Ward, D. C., and Cedar, H. (1992) Delineation of DNA replication time zones by fluorescence *in situ* hybridization. *EMBO J.*, **11**: 1217-1225.

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

Simmler, M.-C., Rouyer, F., Vergnaud, G., Nystrom-Lahti, M., Ngo, K. Y., de la Chapelle,

A., and Weissenbach, J. (1985) Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. *Nature*, **317**: 692-697.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffith, B. L., Smith, M. J., Foster, J. W., Frischauf, A.-M., Lovell-Badge, R., and Goodfellow, P. N. (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, **346**: 240-244.

Smith, C. L. and Canter, C. (1987) Purification, specific fragmentation, and separation of large DNA molecules. *Meth. Enzymol.*, **155**: 449-467.

Sneath, P. H. A. and Sokal, R. R. (1973) Numerical taxonomy. Freeman, San Francisco.

Spring, B., Fonatsch, C., Muller, C., Pawelec, G., Kompf, J., Wernet, P., and Ziegler, A. (1985) Refinement of HLA gene mapping with induced B-cell line mutants. *Immunogenet.*, **21**: 277-291.

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.

Takeda, J., Yano, H., Eng, S., Zeng, Y., and Bell, G. I. (1993) A molecular inventory of human pancreatic islets : sequence analysis of 1000 cDNA clones. *Hum. Mol. Genet.*, **2**: 1793-1798.

Therman, E. (1986) Human chromosomes: structure, behavior, effects. Springer-Verlag,

New York.

Toguchida, J., McGee, T. L., Paterson, J. C., Eagle, J. R., Tucker, S., Yandell, D. W., and Dryja, T. P. (1993) Complete genomic sequence of the human retinoblastoma susceptibility gene. *Genomics*, **17**: 535-543.

Tomatsu, S., Kobayashi, Y., Fukumaki, T., Yubisui, T., Orii, T., and Sakaki, Y. (1989) The organization and the complete nucleotide sequence of human NADH-cytochrome b5 reductase gene. *Gene*, **80**: 353-361.

Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U. S. A.*, **88**: 11262-11265.

Yunis, J. J. (1981) Mid-prophase human chromosome. The attainment of 2000 bands. *Hum. Genet.*, **56**: 293-298.

Winkler, H. (1920) Vererbung und Ursache der Parthenogenese im Pflanzen und Tierreich, Fischer, Jena.

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

Zuker, M. (1989) Computer prediction of RNA structure. *Meth. Enzymol.*, **180**: 262-288.

Figure 1

Base-compositional map of the human MHC. 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 under Materials and Methods. Seven thick vertical bars in class II marked by asterisks (*) 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 isochore 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).

Figure 1

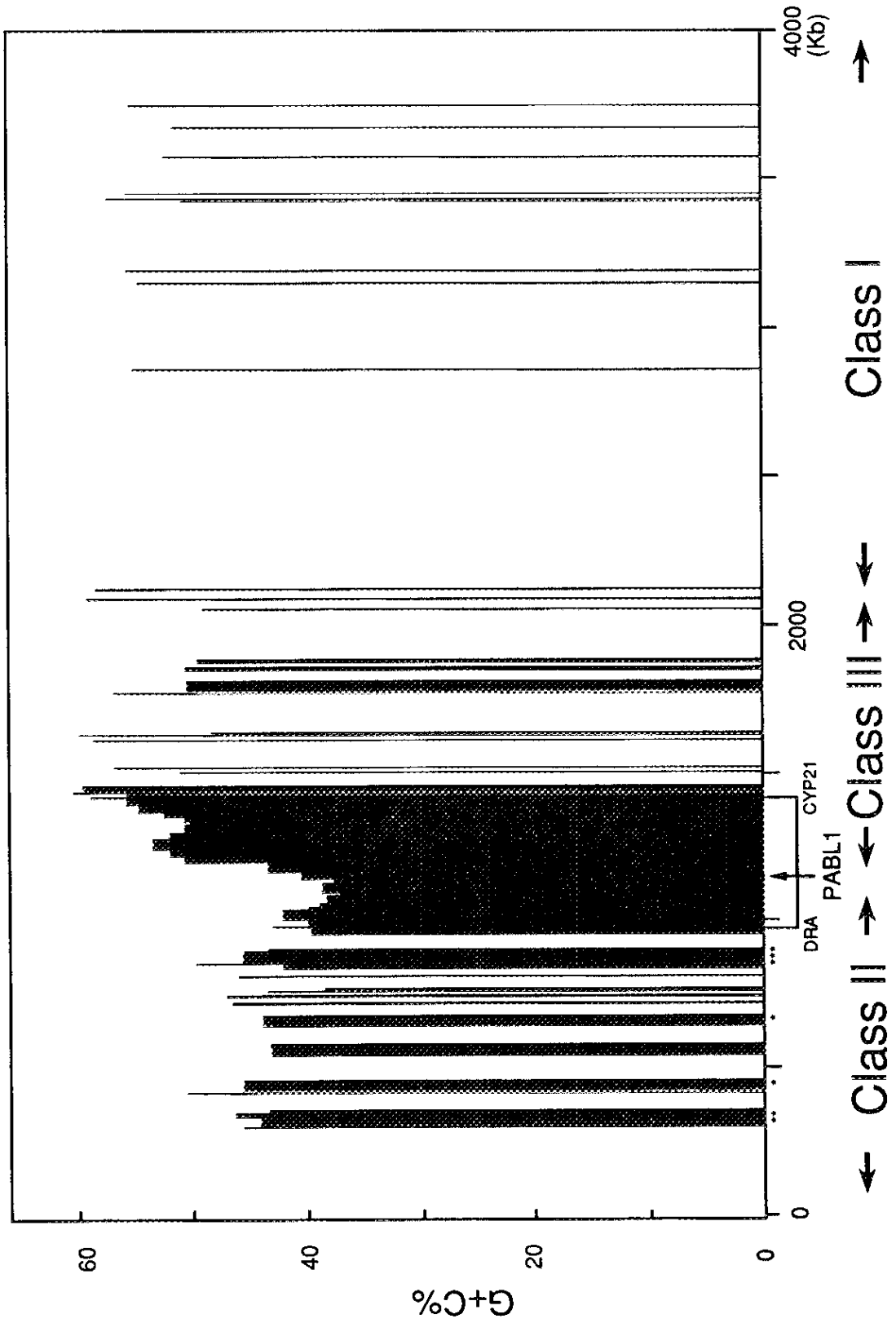

G+C%
0
20
40
60
0
2000
4000
(Kb)
Class II
Class III
Class I
DRA
PABL1
CYP21

Figure 2

Molecular map of contiguous cosmid, λ phage, and YAC clones that cover the 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 DNAs were used for the λ-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. After completion of λ phage walking, cosmid clones (F-series) were isolated using the phage clones as probes; the region where the author first encountered difficulty in cosmid walking could not be cloned even by this procedure. PABL1 was found in P515PT5. Notch 3 is the human counterpart of mouse *int-3* reported by our group (Sugaya *et al.*, 1994). LINE-1 cluster was partially sequenced and at least five independent LINE-1 repeats were found. Clones used for GC% measurements are marked by pound sign (#).

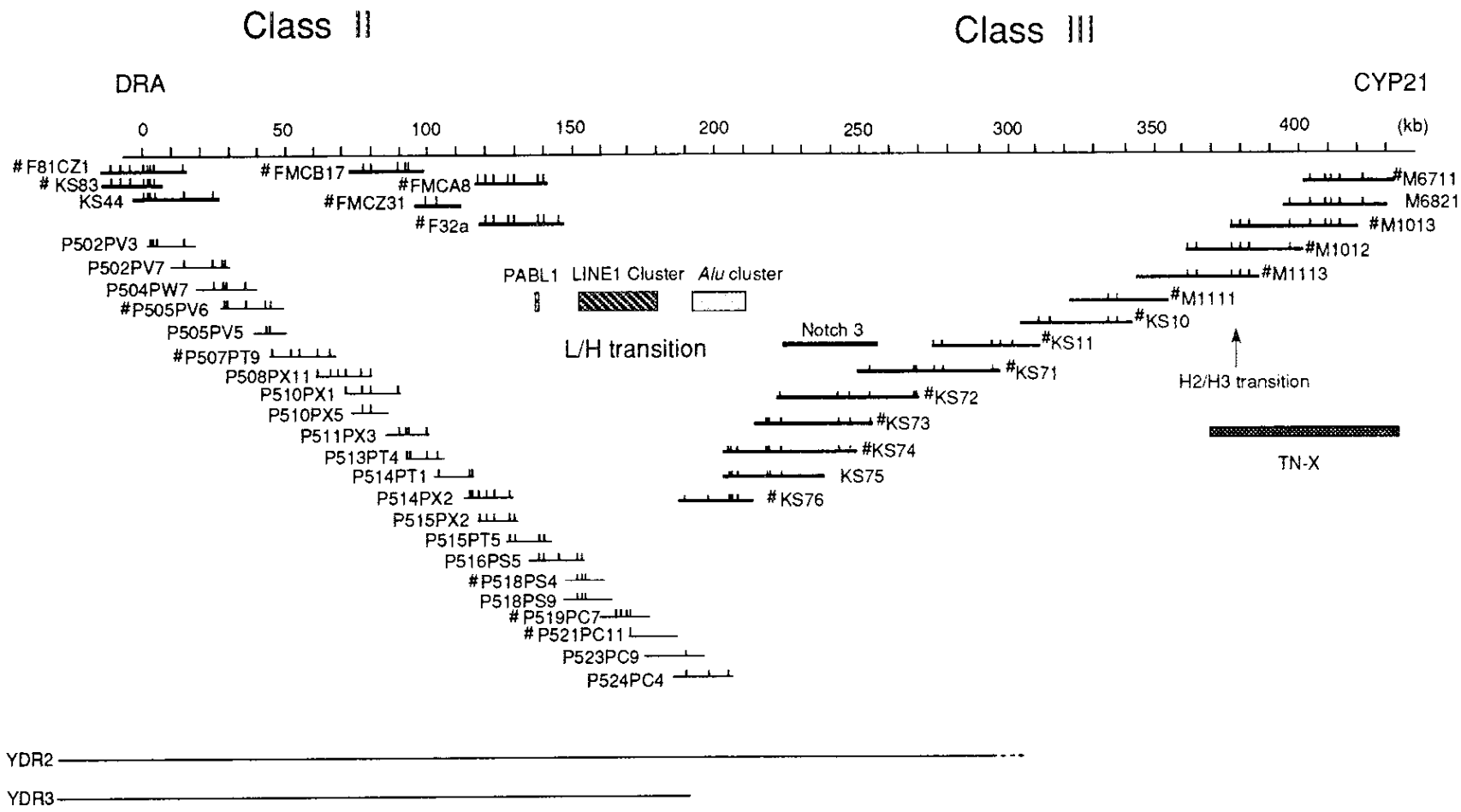

Class II
Class III
DRA
CYP21
0
50
100
150
200
250
300
350
400
(kb)
#F81CZ1
KS83
KS44
#FMCB17
#FMCA8
#FMCZ31
#F32a
P502PV3
P502PV7
P504PW7
#P505PV6
P505PV5
#P507PT9
P508PX11
P510PX1
P510PX5
P511PX3
P513PT4
P514PT1
P514PX2
P515PX2
P515PT5
P516PS5
#P518PS4
P518PS9
#P519PC7
#P521PC11
P523PC9
P524PC4
PABL1
LINE1 Cluster
Alu cluster
L/H transition
Notch 3
#KS11
#KS71
#KS72
#KS73
#KS74
KS75
#KS76
#M6711
M6821
#M1013
#M1012
#M1113
#M1111
#KS10
H2/H3 transition
TN-X
YDR2
YDR3

Figure 2

Figure 3

Comparison of PABX1, PABY1, and PABL1 sequences. Multiple alignment of nucleotide sequences was performed using the MALIGN program of DDBJ. An *Alu* sequence present in PABY1 is omitted in this alignment. Positions of identical nucleotides are marked by asterisks (*). PABXY1 sequences upstream to the *Eco*RI site (GAATTC) were not reported. An arrow shows the 3' terminus of the homology. The 3' terminus corresponds approximately to the PABXY1 homology terminus reported by Ellis *et al.* (1989).

Figure 3

```
PABX1    GAATTCTTAACAGGACCCATTTAGGATT-AAACAAGTTTTACTGGGGGTCTGCAGAAACT
PABY1    GAATTCTTAACAGGACCCATTTAGGATT-AAACAAGTTTTACTGGGGGTCTGCAGAAACT
PABL1    GAATTCTTAACAGGACCCGTTTAGGATTAAAACAAGTTTATTGGGGGGTCTGAAGAAACT
         ****************** ********* **********    ********* *******

PABX1    CCCCAGGCCTCCACAAACAAGTTTATTGGGGCTTTGAAGGAACTCTGCAAACCTCCTGGA
PABY1    CCCCAGGCCTCCACAAACAAGTTTATTGGGGCTTTGAAGGAACTCTGCAAACCTCCTGGA
PABL1    CCCCAGGCCTTCACAAACAAGTTTATTGGGGGTCT-TCTGAA----G-GAA-CTCC-ATA
         ********** ********************* * *    ***    *  ** ****   *

PABX1    TTTAGCAGGAGACAACATGAGGGTAATCACCCCGGCACCTGGACCCA-TTAGATTAAGTC
PABY1    TTTAGCAGGAGACAACATGAGGGTAATCACCCCGGCACCTGGACCCA-TTAGATTAAGTC
PABL1    TTTAGCAGGAGACAAGATAAGGGTAATCACTCCAGCACCTGGACCCATTTAGATTAAGTA
         *************** ** *********** ** ************* ***********

PABX1    AATTTACTGAGGCTCCTGAGGAAGATCCTCAGGACTCAGACCTTAGTTATAGATTAAAAG
PABY1    AATTTACTGAGGCTCCTGAGGATGATGCTCAGGACTCAGACCTTAGTTATAGATTAAAAG
PABL1    AATTTACTGAAGCTCTAGAGGAAAGCCTTCAGGACTCACATCTTAGTCACAGATTAGAAG
         ********** ****  *****      ********** * ****** * ****** ***

PABX1    AAGTTAATCACTTATGTCTTTAGATAAAT--GCACACACATATCTCCACATAGCTTGGAA
PABY1    AAGTTAATCATTTATGTCGTTACATAAATGGGCACTTACACATAGACGTATAGCTCAGAA
PABL1    AAGTTAATGACTTATGTCTTTAGATGAATGCACACTTACACGTAGACATATAGCTTAGAA
         ******** * ******* *** ** ***   ***  ***  *   *  ******  ***

PABX1    GGTATATAAGCTCTGGAAAAC--TATAATTTTGAGTTAGTCTGGTGATA--ATTTCCAGG
PABY1    GGTATATAAGCTCTATAAAACTTTATCATTTTGAGTGGGTCTGGTAATATTATCTACATG
PABL1    GGTATATTGGCTCTGGAAAACTTTGTAATTTTCAGTTGGTCTGG-CA-AAAATTTCCAGG
         *******  *****  *****  * * ***** ***  ******  * *  ** * ** *

PABX1    CCTTCTCCCTGTAACAGGTTGCAGAAATAAAAACTCTCTTCCTCCCCAGTTCATCGGTGT
PABY1    CCTTCTCCCTGTAACCACTTGTAGAAATAAAAACTCTCTTCCTTCTCAGTTCATCTGCAT
PABL1    CCTTCTCTCTGTACCTACTTATATAAATAAAAACTGTCTTCTTTCTCAGTTCATCTGCAT
         ******* ***** *   **  * *********** ***** * * ********* *  *

PABX1    TTCATTATTGGG-CTGTGAGAAATAGCAACC-CAGTTGGGTCCGGGAACAGAAGGTTTG
PABY1    CTCATTATTGGGCCACAAAAAAATAGCAGCCTGACCCTCAATTTGCTCAGGAAAGACAA
PABL1    CTCGTTATTGGGCCATGAAGAAA-AGCAGCCCGATTCTC--CTACCTC-AGCCTCCCAA
          ** ******** *    * *** **** **  *  *
```

Figure 4

Southern blot analysis of human and bovine DNAs with PABL1 probe. For each lane, 10 μg of high-molecular-weight DNA from human placenta or bovine lung (Clontech, CA) was digested with *Eco*RI (E), *Pst*I (P), or *Rsa*I (R), separated by electrophoresis on a 1% agarose gel, blotted onto a Hybond-N^+ nylon membrane, and hybridized with a ^{32}P-labeled 360-nt portion of PABL1 corresponding to PABXY1 sequences (see Fig. 3); the same hybridization probe was used for cloning of other PABLs. For stringent washing, filters were washed at 65℃ in 0.1 × SSPE containing 0.1% SDS for 15 min.

Figure 4

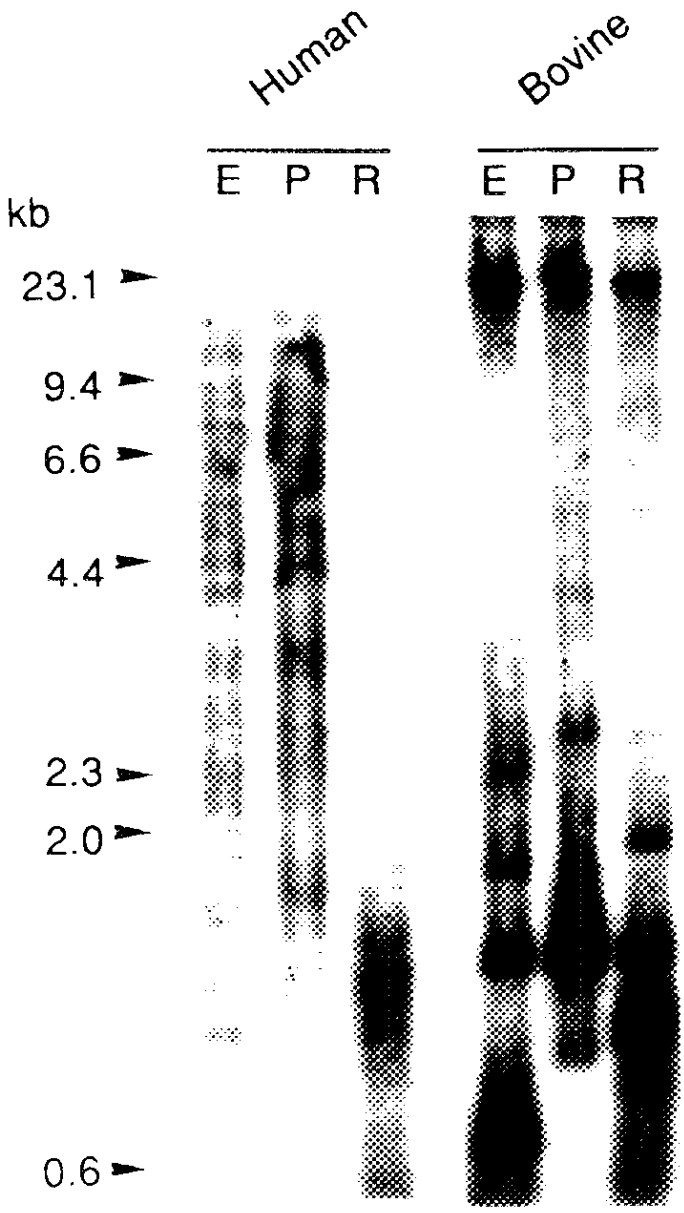


Figure 5

Comparison of genomic sequences of PABXY1 and PABLs. Multiple alignment of nucleotide sequences was performed using the MALIGN program. An *Alu* sequence present in PABY1 is omitted in this alignment. Positions of identical nucleotides are marked by asterisks(*), and those at which five of six nucleotides are identical, by dots (•). PABXY1 sequences upstream to *Eco*RI site (GAATTC) were kindly supplied by Dr. N. A. Ellis. Arrows show termini of the core of PABLs proposed. Nucleotide sequences for PABL1-3 and PABLSp2 have been deposited with DDBJ / EMBL / GenBank under Accession Nos. D30042, D30043, D30044, and D55642, respectively.

Figure 5

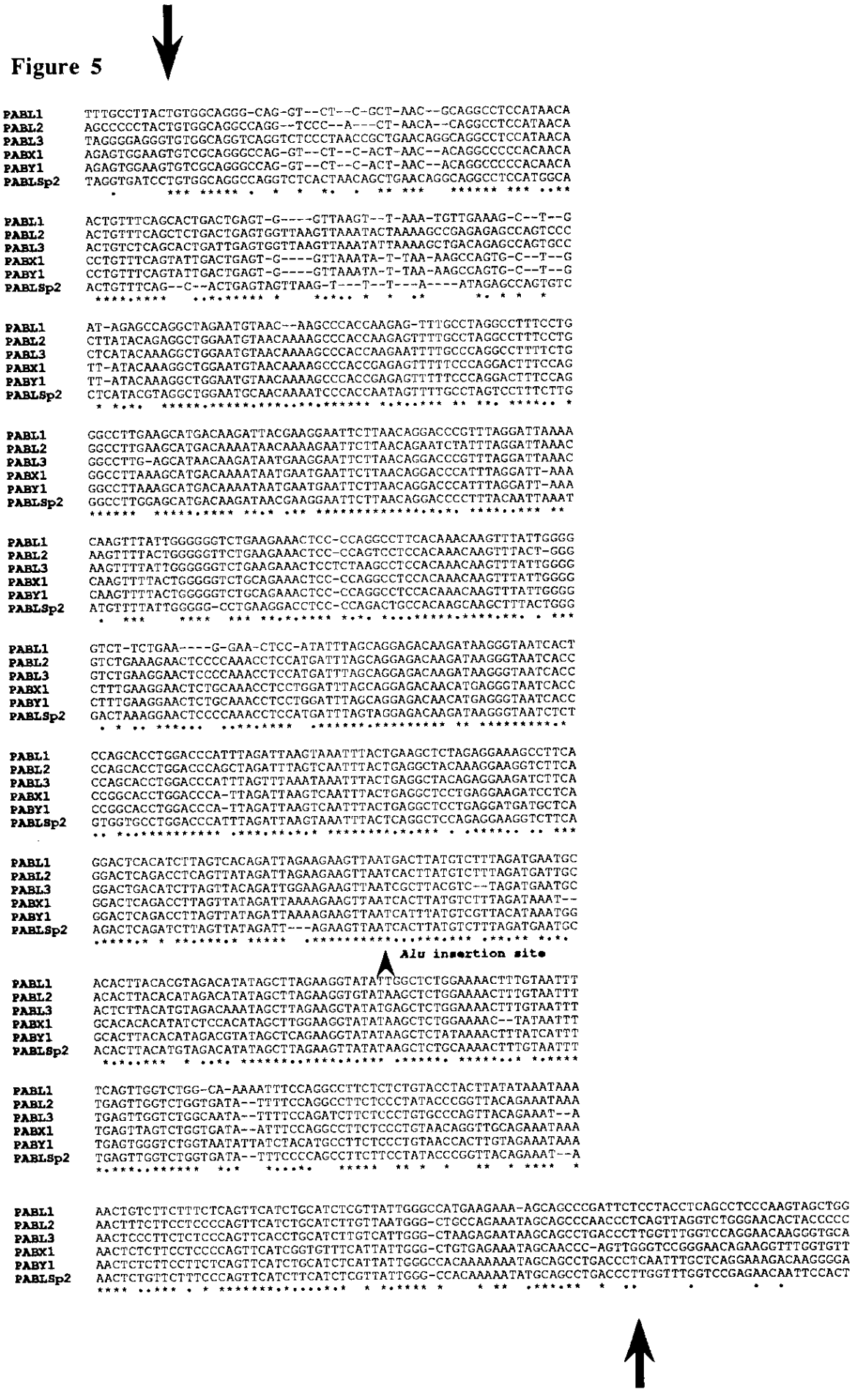


Figure 6

Genomic structures of PABXY1 and PABLs. In both PABXY1 and PABLs, regions upstream and downstream to the *Alu*-insertion site are separately marked (see the Text). There is a core unit of PABL (ca. 650 nt) in each sequence, which is flanked by the regions divergent between the clones.

Figure 6

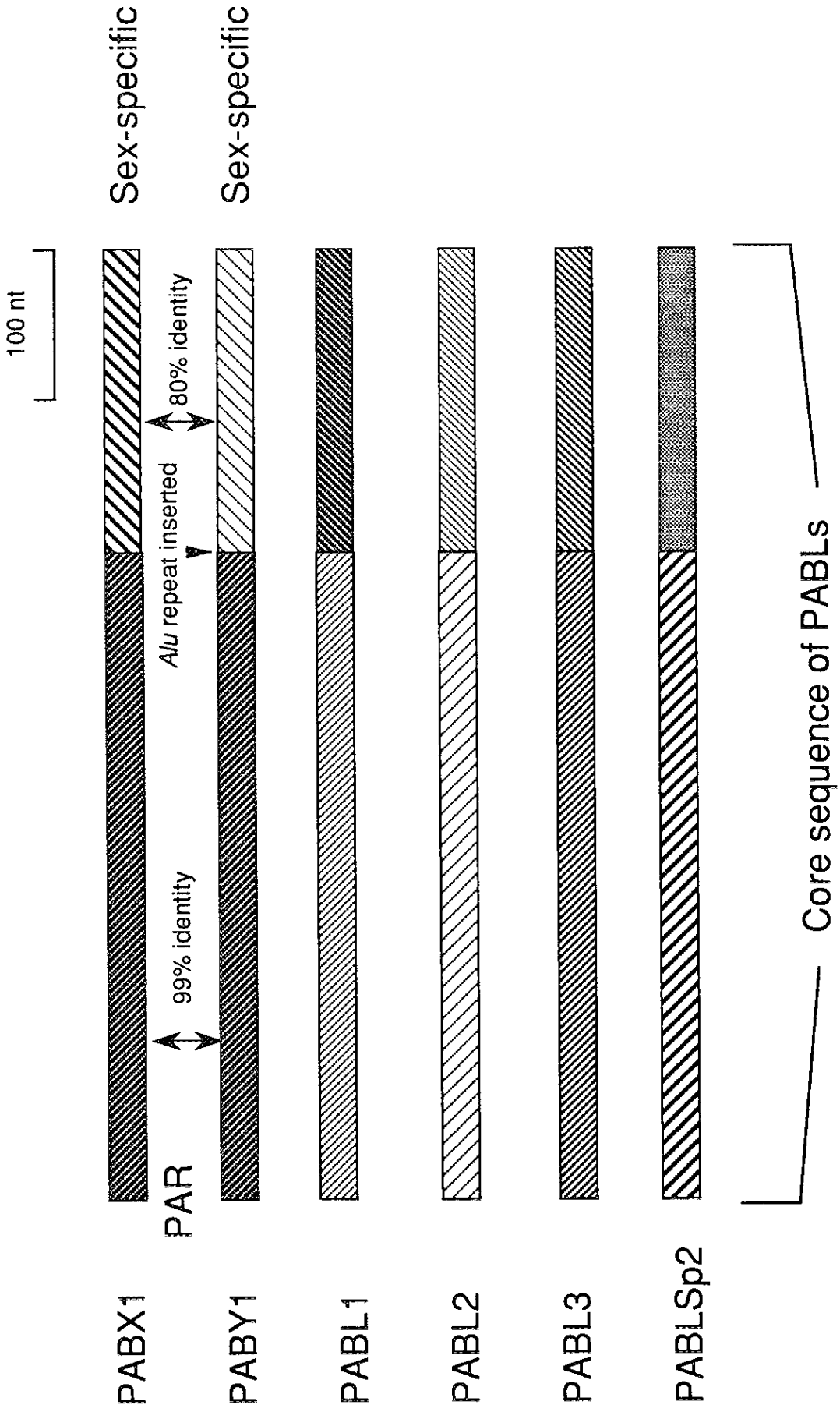

Sex-specific
Sex-specific
100 nt
80% identity
Alu repeat inserted
99% identity
PAR
PABX1
PABY1
PABL1
PABL2
PABL3
PABLSp2
Core sequence of PABLs

Figure 7

Comparison of the 5' region of PABLs in cDNA clones with that of PABL1 and PABX1. Multiple alignment of nucleotide sequences was performed using the MALIGN program. Near the 5' edge, two EST sequences (T92306 and R12279) were added to the alignment. Positions of identical nucleotides are marked by asterisks, and those at which four of five nucleotides (or six out of seven nucleotides for the region with EST sequences) are identical, by dots. An arrow shows the 5' edge of the core unit of PABLs proposed. The multiple alignment only analyzing the cDNA and EST sequences showed the 5' homology terminus to be practically identical to that for the genomic PABLs.

Figure 7

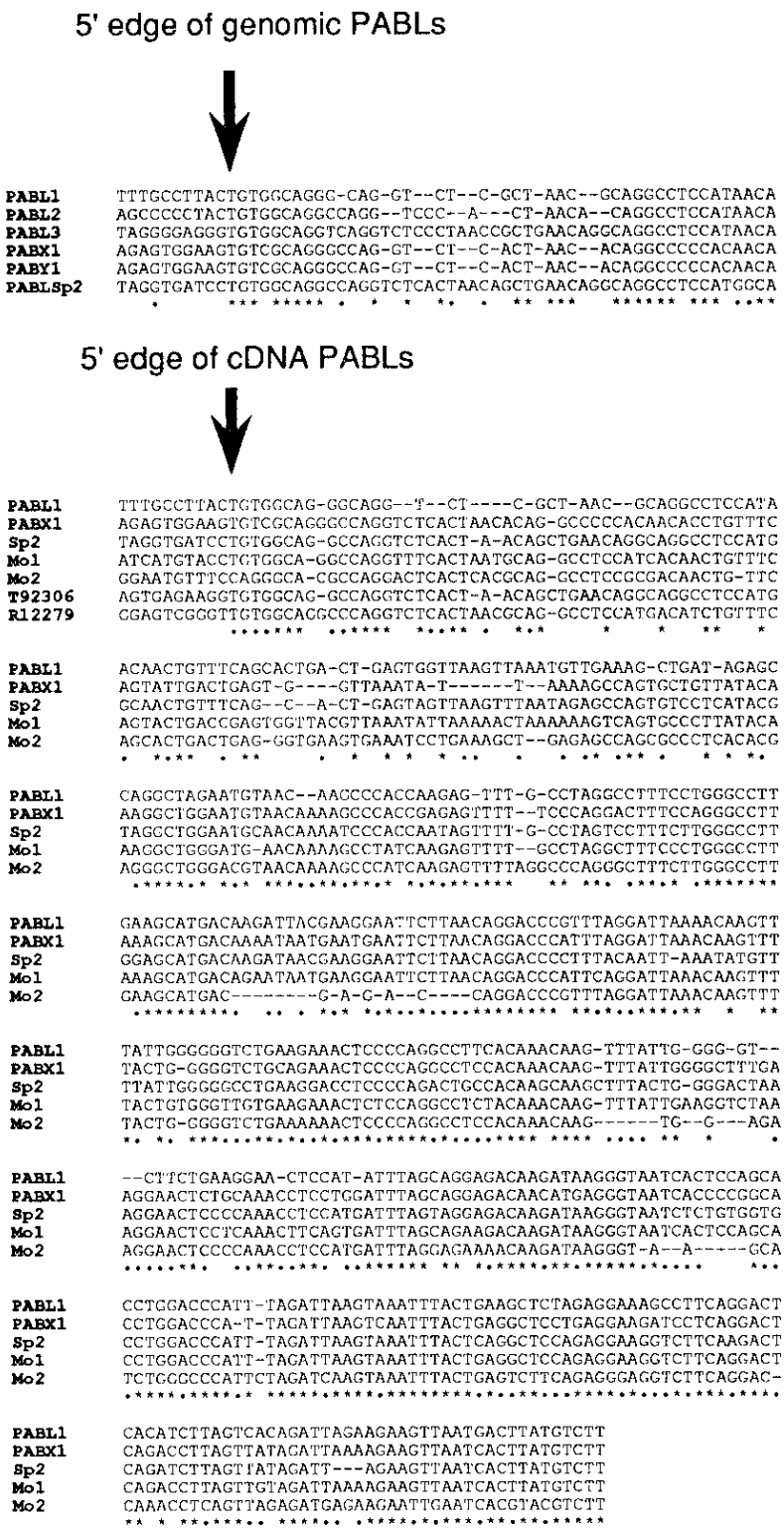

5' edge of genomic PABLs
PABL1
PABL2
PABL3
PABX1
PABY1
PABLSp2
5' edge of cDNA PABLs
PABL1
PABX1
Sp2
Mo1
Mo2
T92306
R12279

Figure 8

Comparison of the 3' region of PABLs in cDNA clones with that of PABL1 and PABX1. Multiple alignment and marking of nucleotide sequences are the same as those of Fig. 7. Near the 3' edge, one EST (T47905) was added to the alignment. The 3' homology terminus found by analyzing only cDNAs and EST approximately corresponds to that for the genomic PABLs. Based on individual cDNA and EST sequences obtained, directions of their transcription can often be predicted, and both directions appear to exist for the sequences analyzed in Figs. 7 and 8. It should be noted, however, that these assignments were solely dependent on the assumption that the construct of individual cDNAs exactly conforms to the design for the libraries (i.e., without complicated and unexpected events during formation of the respective cDNA clones). Undoubtedly the author need different types of experiments for conclusive assignment of direction for individual transcripts.

Figure 8

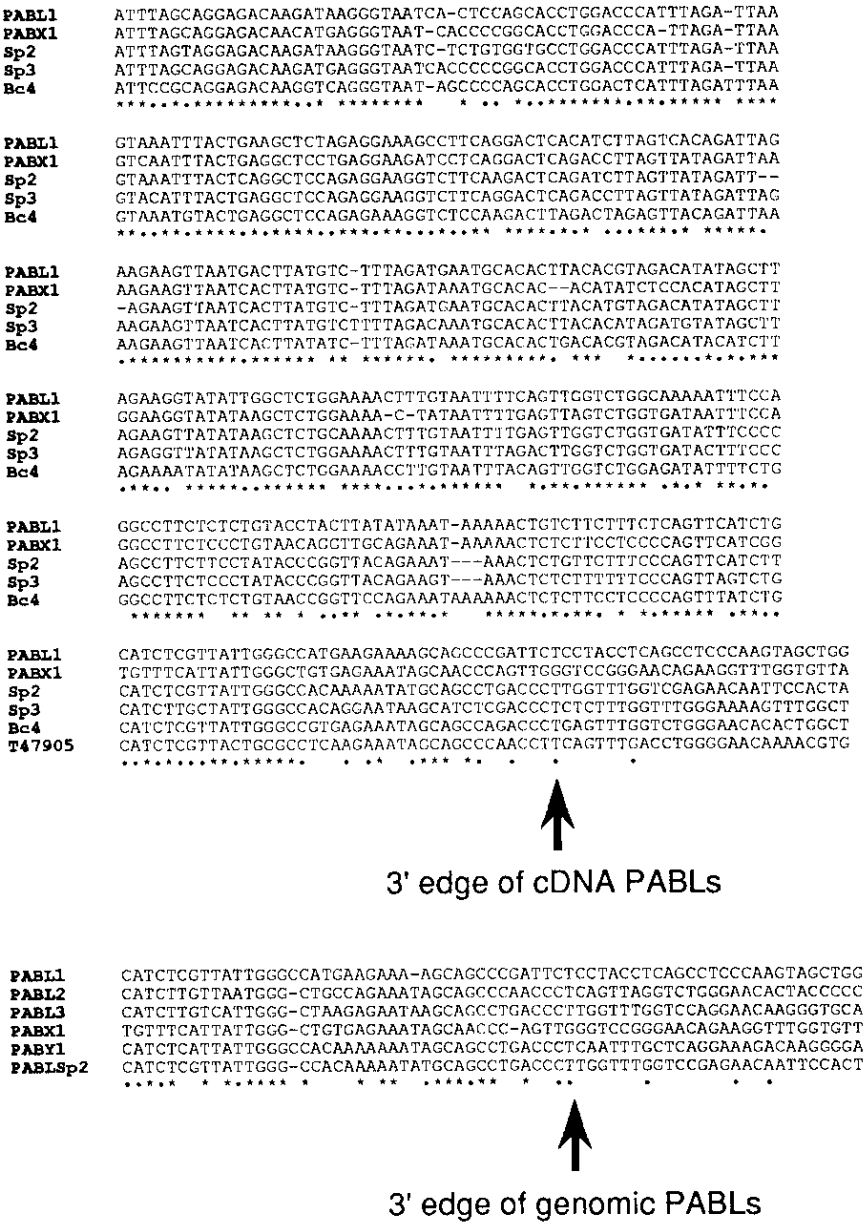

PABL1 ATTTAGCAGGAGACAAGATAAGGGTAATCA-CTCCAGCACCTGGACCCATTTAGA-TTAA
PABX1 ATTTAGCAGGAGACAACATGAGGGTAAT-CACCCCGGCACCTGGACCCA-TTAGA-TTAA
Sp2 ATTTAGTAGGAGACAAGATAAGGGTAATC-TCTGTGGTGCCTGGACCCATTTAGA-TTAA
Sp3 ATTTAGCAGGAGACAAGATGAGGGTAATCACCCCCGGCACCTGGACCCATTTAGA-TTAA
Bc4 ATTCCGCAGGAGACAAGGTCAGGGTAAT-AGCCCCAGCACCTGGACTCATTTAGATTTAA
PABL1 GTAAATTTACTGAAGCTCTAGAGGAAAGCCTTCAGGACTCACATCTTAGTCACAGATTAG
PABX1 GTCAATTTACTGAGGCTCCTGAGGAAGATCCTCAGGACTCAGACCTTAGTTATAGATTAA
Sp2 GTAAATTTACTCAGGCTCCAGAGGAAGGTCTTCAAGACTCAGATCTTAGTTATAGATT--
Sp3 GTACATTTACTGAGGCTCCAGAGGAAGGTCTTCAGGACTCAGACCTTAGTTATAGATTAG
Bc4 GTAAATGTACTGAGGCTCCAGAGAAAGGTCTCCAAGACTTAGACTAGAGTTACAGATTAA
PABL1 AAGAAGTTAATGACTTATGTC-TTTAGATGAATGCACACTTACACGTAGACATATAGCTT
PABX1 AAGAAGTTAATCACTTATGTC-TTTAGATAAATGCACAC--ACATATCTCCACATAGCTT
Sp2 -AGAAGTTAATCACTTATGTC-TTTAGATGAATGCACACTTACATGTAGACATATAGCTT
Sp3 AAGAAGTTAATCACTTATGTCTTTTAGACAAATGCACACTTACACATAGATGTATAGCTT
Bc4 AAGAAGTTAATCACTTATATC-TTTAGATAAATGCACACTGACACGTAGACATACATCTT
PABL1 AGAAGGTATATTGGCTCTGGAAAACTTTGTAATTTTCAGTTGGTCTGGCAAAAATTTCCA
PABX1 GGAAGGTATATAAGCTCTGGAAAA-C-TATAATTTTGAGTTAGTCTGGTGATAATTTCCA
Sp2 AGAAGTTATATAAGCTCTGCAAAACTTTGTAATTTTGAGTTGGTCTGGTGATATTTCCCC
Sp3 AGAGGTTATATAAGCTCTGGAAAACTTTGTAATTTAGACTTGGTCTGGTGATACTTTCCC
Bc4 AGAAAATATATAAGCTCTGGAAAACCTTGTAATTTACAGTTGGTCTGGAGATATTTTCTG
PABL1 GGCCTTCTCTCTGTACCTACTTATATAAAT-AAAAACTGTCTTCTTTCTCAGTTCATCTG
PABX1 GGCCTTCTCCCTGTAACAGGTTGCAGAAAT-AAAAACTCTCTTCCTCCCCAGTTCATCGG
Sp2 AGCCTTCTTCCTATACCCGGTTACAGAAAT---AAACTCTGTTCTTTCCCAGTTCATCTT
Sp3 AGCCTTCTCCCTATACCCGGTTACAGAAGT---AAACTCTCTTTTTTCCCAGTTAGTCTG
Bc4 GGCCTTCTCTCTGTAACCGGTTCCAGAAATAAAAAACTCTCTTCCTCCCCAGTTTATCTG
PABL1 CATCTCGTTATTGGGCCATGAAGAAAAGCAGCCCGATTCTCCTACCTCAGCCTCCCAAGTAGCTGG
PABX1 TGTTTCATTATTGGGCTGTGAGAAATAGCAACCCAGTTGGGTCCGGGAACAGAAGGTTTGGTGTTA
Sp2 CATCTCGTTATTGGGCCACAAAAATATGCAGCCTGACCCTTGGTTTGGTCGAGAACAATTCCACTA
Sp3 CATCTTGCTATTGGGCCACAGGAATAAGCATCTCGACCCTCTCTTTGGTTTGGGAAAAGTTTGGCT
Bc4 CATCTCGTTATTGGGCCGTGAGAAATAGCAGCCAGACCCTGAGTTTGGTCTGGGAACACACTGGCT
T47905 CATCTCGTTACTGCGCCTCAAGAAATAGCAGCCCAACCTTCAGTTTGACCTGGGGAACAAAACGTG
3' edge of cDNA PABLs
PABL1 CATCTCGTTATTGGGCCATGAAGAAA-AGCAGCCCGATTCTCCTACCTCAGCCTCCCAAGTAGCTGG
PABL2 CATCTTGTTAATGGG-CTGCCAGAAATAGCAGCCCAACCCTCAGTTAGGTCTGGGAACACTACCCCC
PABL3 CATCTTGTCATTGGG-CTAAGAGAATAAGCAGCCTGACCCTTGGTTTGGTCCAGGAACAAGGGTGCA
PABX1 TGTTTCATTATTGGG-CTGTGAGAAATAGCAACCC-AGTTGGGTCCGGGAACAGAAGGTTTGGTGTT
PABY1 CATCTCATTATTGGGCCACAAAAAAATAGCAGCCTGACCCTCAATTTGCTCAGGAAAGACAAGGGGA
PABLSp2 CATCTCGTTATTGGG-CCACAAAAATATGCAGCCTGACCCTTGGTTTGGTCCGAGAACAATTCCACT
3' edge of genomic PABLs

Figure 9

Structures of cDNA clones having PABLs. Results of multiple alignments described in Figs. 7 and 8 are summarized. There is the core unit of PABL (ca. 650 nt) flanked by the regions divergent between cDNA clones, as for the genomic clones. Significant ORFs could not be found for these cDNA, not only for the core PABLs but also for their flanks. EST sequences showing a high homology with PABLs are shown by horizontal lines: GenBank Loci HSAAAAWAH (Accession No. Z19872), hbc671 (T11103), HSGS04158 (D25790), and GenBank Accession Nos. R07404, R12279, T99392, T47905, R38816, and R43643. Sequence of the HSAAAAWAH, not only the PABL core but also its flanking region, was practically identical to a portion of Mo2 cDNA.

Figure 9

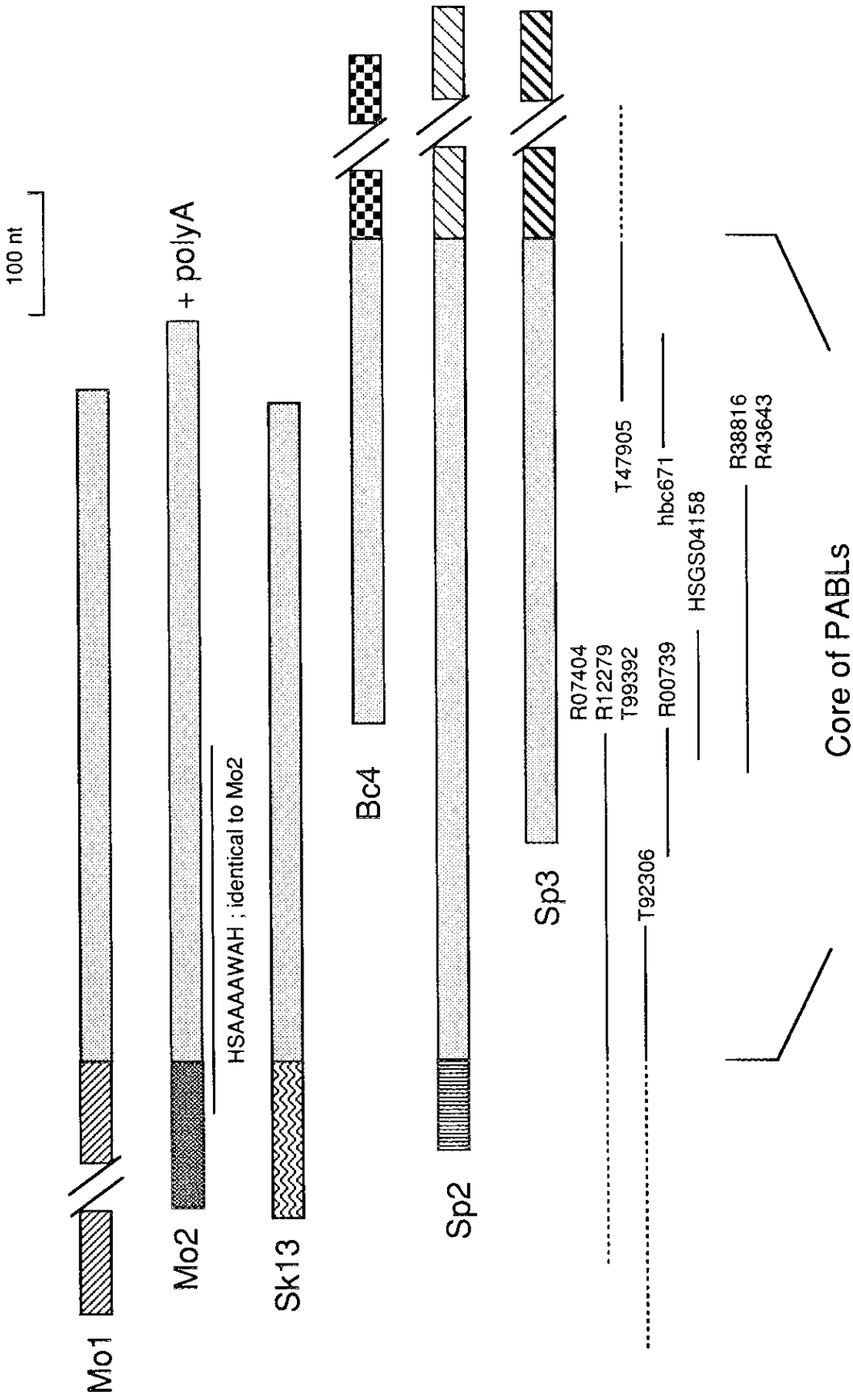

100 nt
+ polyA
HSAAAAWAH ; identical to Mo2
Mo1
Mo2
Sk13
Bc4
Sp2
Sp3
R07404
R12279
T99392
R00739
HSGS04158
hbc671
T47905
T92306
R38816
R43643
Core of PABLs

Figure 10

RNA secondary structure of Sp2 cDNA predicted by the FOLDRNA program in UWGCG. PABL core (nucleotide No. 1-629) was folded mostly within itself separating from its flanking sequence (nucleotide No. 630-1150).

Figure 10

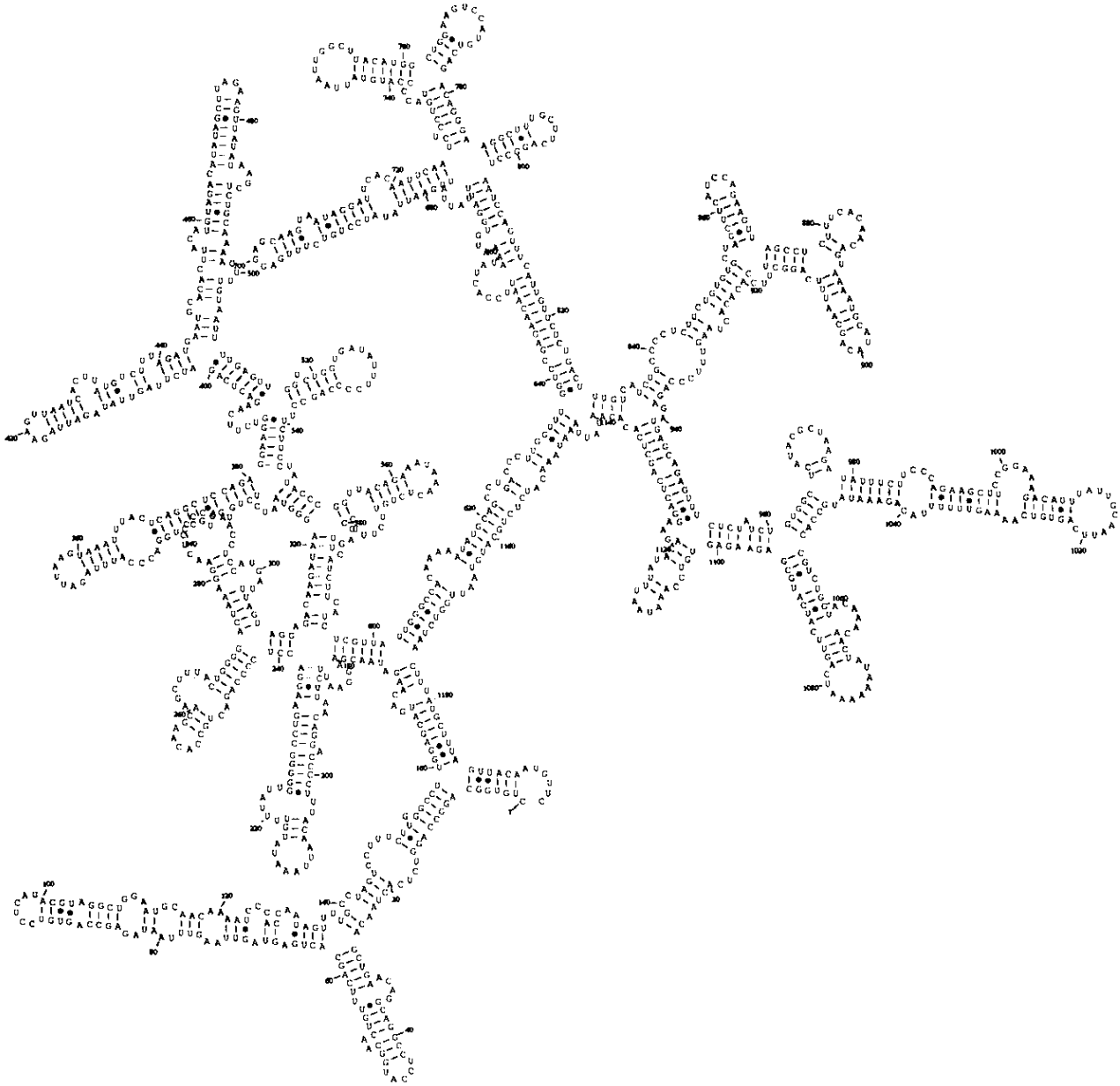


Figure 11

A possible RNA secondary structure of the PABL consensus sequence predicted by the MFOLD program in UWGCG. The consensus sequence of PABL is listed in Fig. 14A.

Figure 11

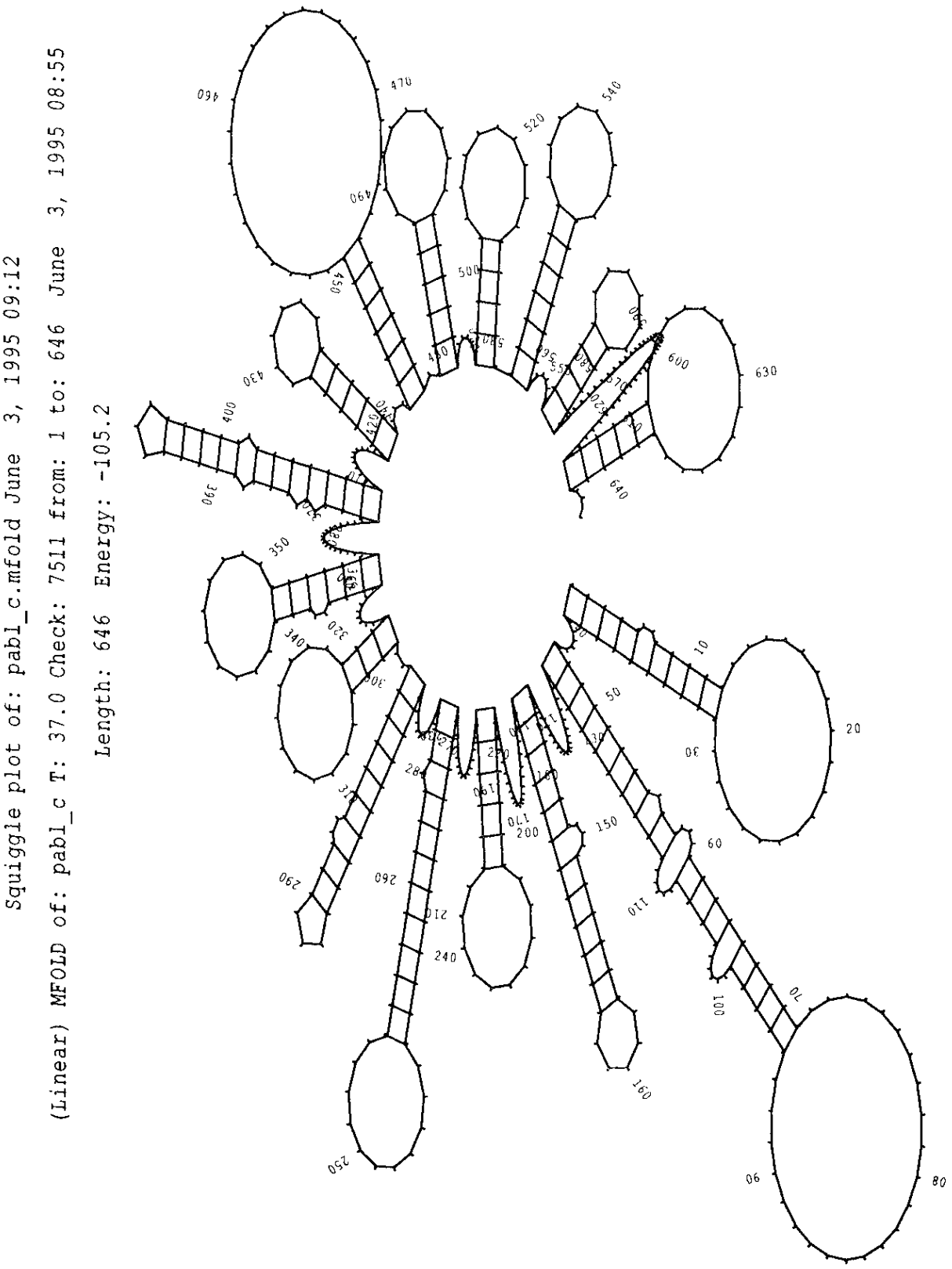

Squiggle plot of: pabl_c.mfold June 3, 1995 09:12
(Linear) MFOLD of: pabl_c T: 37.0 Check: 7511 from: 1 to: 646 June 3, 1995 08:55
Length: 646 Energy: -105.2

Figure 12

Northern blot analysis of human total or polyA$^+$ RNA. (A) Total RNA prepared from GM01416D cells was employed. The probe used for the hybridization was a 598-bp PCR fragment of PABL1. Broad bands mainly with mobilities slower than that of 28S rRNA were detected. (B) PolyA$^+$ RNAs of different tissues (lane1, spleen; lane2, thymus; lane3, prostate gland; lane4, testis; lane5, ovary; lane6, small intestine; lane7, colon; lane8, peripheral blood leukocyte) obtained from Clontech (Palo Alto, CA.) were tested. The probe was the PABL1 segment described above. (C) The polyA$^+$ RNA filter same as that of (B) was used after removal of the PABL1 probe. The probe used for this hybridization was a unique ca. 800-nt fragment of Sp2 cDNA deprived of the PABL sequence.

Figure 12

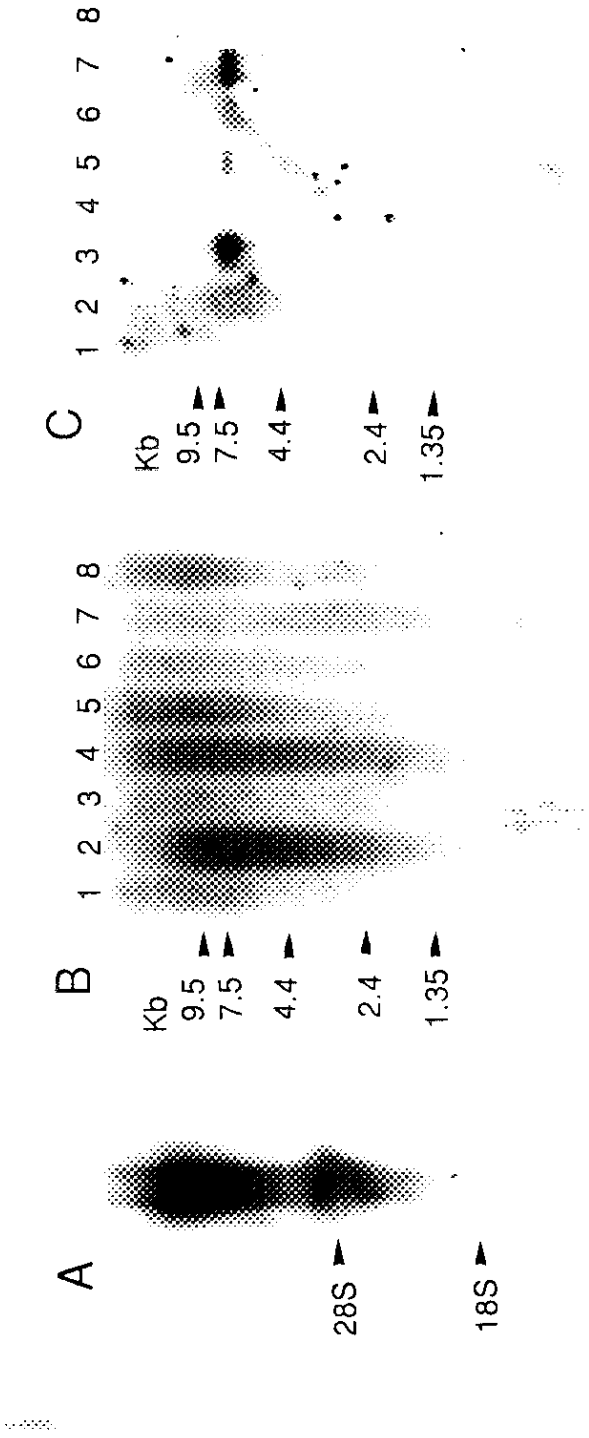

A
28S
18S
B
1 2 3 4 5 6 7 8
Kb
9.5
7.5
4.4
2.4
1.35
C
1 2 3 4 5 6 7 8
Kb
9.5
7.5
4.4
2.4
1.35

Figure 13

Phylogenetic trees of PABLs and PABXY1 sequences, based on the downstream (A) or the upstream (B) portion to the *Alu*-insertion site. Trees were constructed for PABXY1 of four hominoids and three Old World monkeys (Ellis *et al.*, 1990) and five human PABLs, using the neighbor-joining method; mainly because of limit of available PABXY1 sequences of great apes and Old World monkeys and partly of PABL cDNA sequences, the analyzed portions (151 nt for the downstream and 177 nt for the upstream) were shorter than those of the PABL core. Abbreviations for the species are as follows: *Homo sapience* (Human; Hum), *Pan troglodytes* (Chimpanzee; Chi), *Gorilla gorilla* (Gorilla; Gor), *Pongo pygmaeus* (Orangutan; Ora), *Theropithecus gelada* (Gelada baboon; Gel), *Macaca sylvanus* (Barbary macaque; Bar), and *Cercopithecus cephus* (Moustached guenon; Gue). The roots were predicted by using the UPGMA method. Branch lengths were proportional to the estimated number of nucleotide substitutions using the Kimura's two-parameter method. To estimate the divergence time of PABLs, the nucleotide substitutions for the downstream portion were used.

Figure 13A

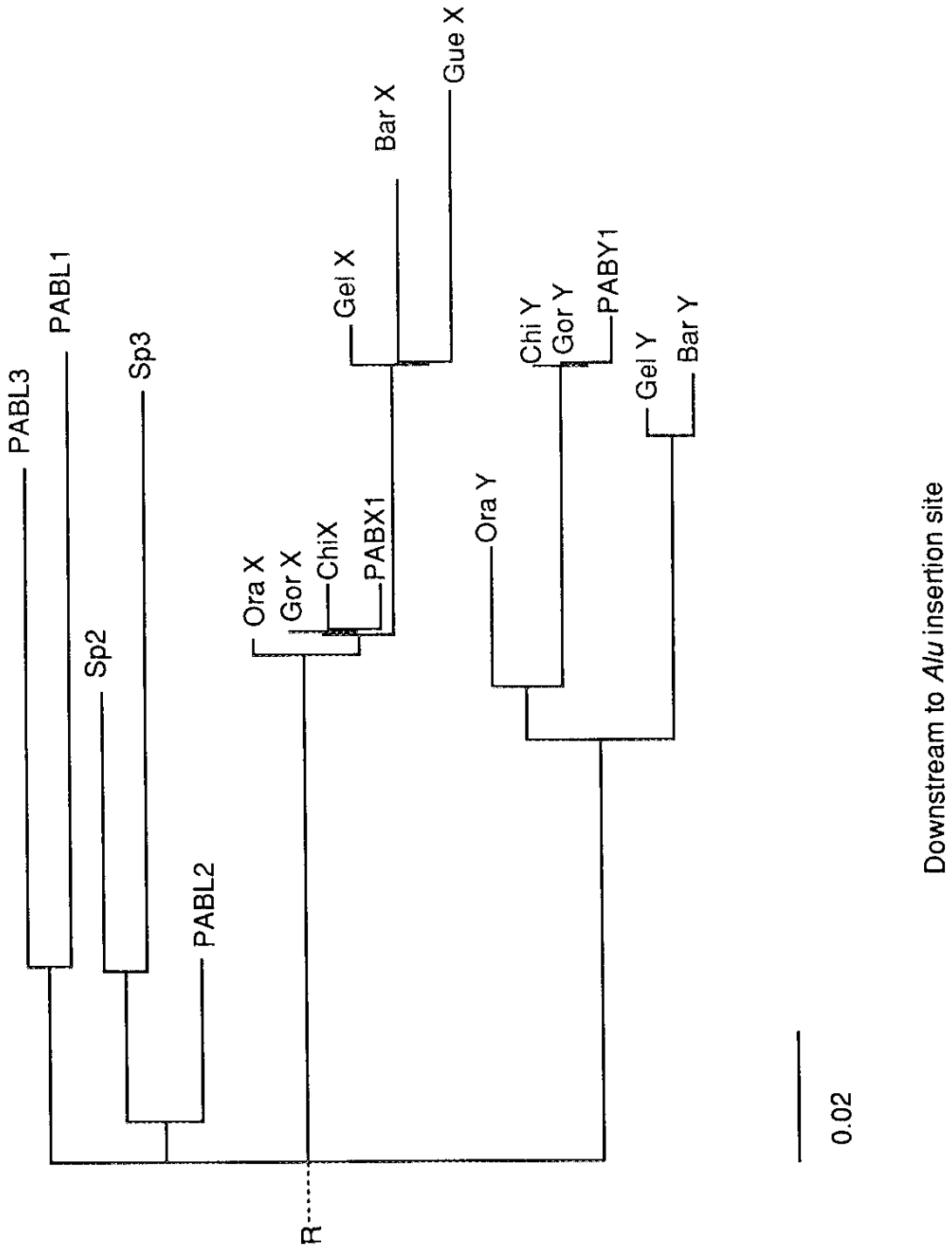

PABL3
PABL1
Sp3
Sp2
PABL2
Ora X
Gor X
ChiX
PABX1
Gel X
Bar X
Gue X
Ora Y
Chi Y
Gor Y
PABY1
Gel Y
Bar Y
R
0.02
Downstream to Alu insertion site

Figure 13B

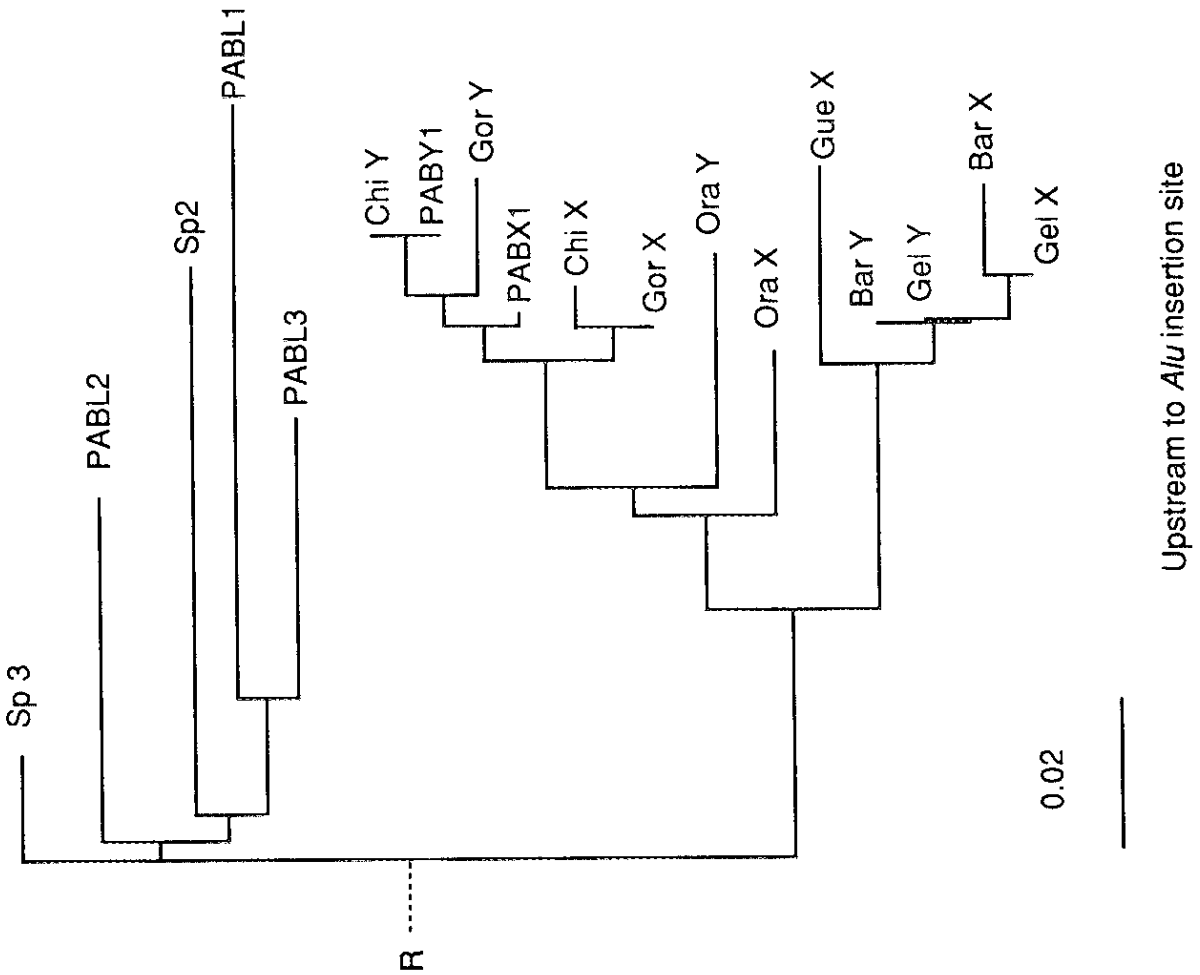

Sp 3
PABL2
Sp2
PABL1
PABL3
R
Chi Y
PABY1
Gor Y
PABX1
Chi X
Gor X
Ora Y
Ora X
Gue X
Bar Y
Gel Y
Bar X
Gel X
0.02
Upstream to Alu insertion site

Figure 14

Consensus sequence of PABLs and its alignment with Sp3. (A) Consensus sequence of PABLs was deduced by utilizing both genomic and cDNA sequences. (B) Comparison of the consensus sequence with the Sp3 upstream portion to the *Alu*-insertion site showed the divergence to be a 5.7% nucleotide difference. (C) Comparison with the Sp3 *Alu*-downstream portion showed a 11.9% divergence.

Figure 14

A) Consensus sequence of PABLs

```
TGTGGCAGGGCCAGGTCTCACTAACAGRCAGGCCTCCATAACAACTGTTTCAGCACTGACTG
AGTGGTTAAGTTAAATAYTAAAAGCYGASAGAGCCAGTSCCCTCATACAAAGGCTGGAATGT
AACAAAAGCCCACCAAGAGTTTTGCCTAGGCCTTTCCTGGGCCTTGAAGCATGACAAGATAA
CGAANGAATTCTTAACAGGACCCNTTTAGGATTAAAANAAKTTTATTGGGGGGTCTGAAGAA
ACTCCCCAGGCCTCCACAAACAAGTTTATTGGGGGTCTGAAAGAACTCCCCAAACCTCCATG
ATTTAGCAGGAGACAAGATAAGGGTAATCACCCCRGCACCTGGACCCATTTAGATTAAGTMA
ATTTACTGAAGCTCCAGAGGAAGRTCTTCAGGACTCAGAYCTTAGTTAYAGATTAGAAGAAG
TTAATCACTTATGTCTTTAGATGAATGCACACTTACAYRTAGACATATAGCTTAGAAGGTAT
ATAAGCTCTGGAAAACTTTGTAATTTTGAGTTGGTCTGGTGATAWTTTCCAGGCCTTCTCCC
TGTACCCRGTTACAGAAATAAAAACTCTCTTCYTTCCCAGTTCATCTGCATCTCGTTATTGG
GCTNNNAGAAATAGCAGCCTGACCCT
```

B) Alignment of the consensus sequence with the Sp3 upstream portion to the *Alu*- insertion site (a 5.7% divergence).

```
Sp3         TGAAGAAACTTCCCAGACCTCCACAAACAAGTTTTATTGGAGGATCTAAAGGAACTTCCC
            ********** ***** ************** ******** ** *** ** ***** ***
Consensus   TGAAGAAACTCCCCAGGCCTCCACAAACAAG-TTTATTGG-GGGTCTGAAAGAACTCCCC

Sp3         AAACCTCCATGATTTAGCAGGAGACAAGATGAGGGTAATCACCCCCGGCACCTGGACCCA
            ****************************** ********** *****************
Consensus   AAACCTCCATGATTTAGCAGGAGACAAGATAAGGGTAATCA-CCCCRGCACCTGGACCCA

Sp3         TTTAGATTAAGTACATTTACTGAGGCTCCAGAGGAAGGTCTTCAGGACTCAGACCTTAGT
            ************* ********* ***************************************
Consensus   TTTAGATTAAGTMAATTTACTGAAGCTCCAGAGGAAGRTCTTCAGGACTCAGAYCTTAGT

Sp3         TATAGATTAGAAGAAGTTAA
            ********************
Consensus   TAYAGATTAGAAGAAGTTAA
```

C) Alignment of the consensus sequence with the Sp3 downstream portion to the *Alu*- insertion site(a 11.9% divergence).

```
Sp3         TCACTTATGTCTTTTAGACAAATGCACACTTACACATAGATGTATAGCTTAGAGGTTATA
            *********** ******  ******************** *********** * ****
Consensus   TCACTTATGTC-TTTAGATGAATGCACACTTACAYRTAGACATATAGCTTAGAAGGTATA

Sp3         TAAGCTCTGGAAAACTTTGTAATTTAGACTTGGTCTGGTGATACTTTCCCAGCCTTCTCC
            ************************* ** ************** *****  *********
Consensus   TAAGCTCTGGAAAACTTTGTAATTTTGAGTTGGTCTGGTGATAWTTTCCAGGCCTTCTCC

Sp3         CTATACCCGGTTACAGAAGT--AAACTCTCTTTTTTCCCAGTTAG
            ** *****.********** *  ********** .*********
Consensus   CTGTACCCRGTTACAGAAATAAAAACTCTCTTCYTTCCCAGTTCA
```

Figure15

Fluorescence *in situ* hybridization of genomic clones having PABL1 (A), PABL2 (B), PABL3 (C), and PABLSp2 (D) onto metaphase chromosomes. Chromosome bands shown in (A)-(C) were the patterns counterstained with propidium iodide for R-banding, and the bands in (D) were those counterstained with Hoechst 33258 for G-banding pattern. PABL1, PABL2, PABL3, and PABLSp2 were located on chromosome bands 6p21.3, 20p11.21, 19q13.3, and 17q23, respectively; arrows indicate signals. These results and those in Table 7 were based on observations made on more than ten (pro)metaphase chromosomes, and fluorescence signals were not consistently observed on other chromosomes.

Figure 15

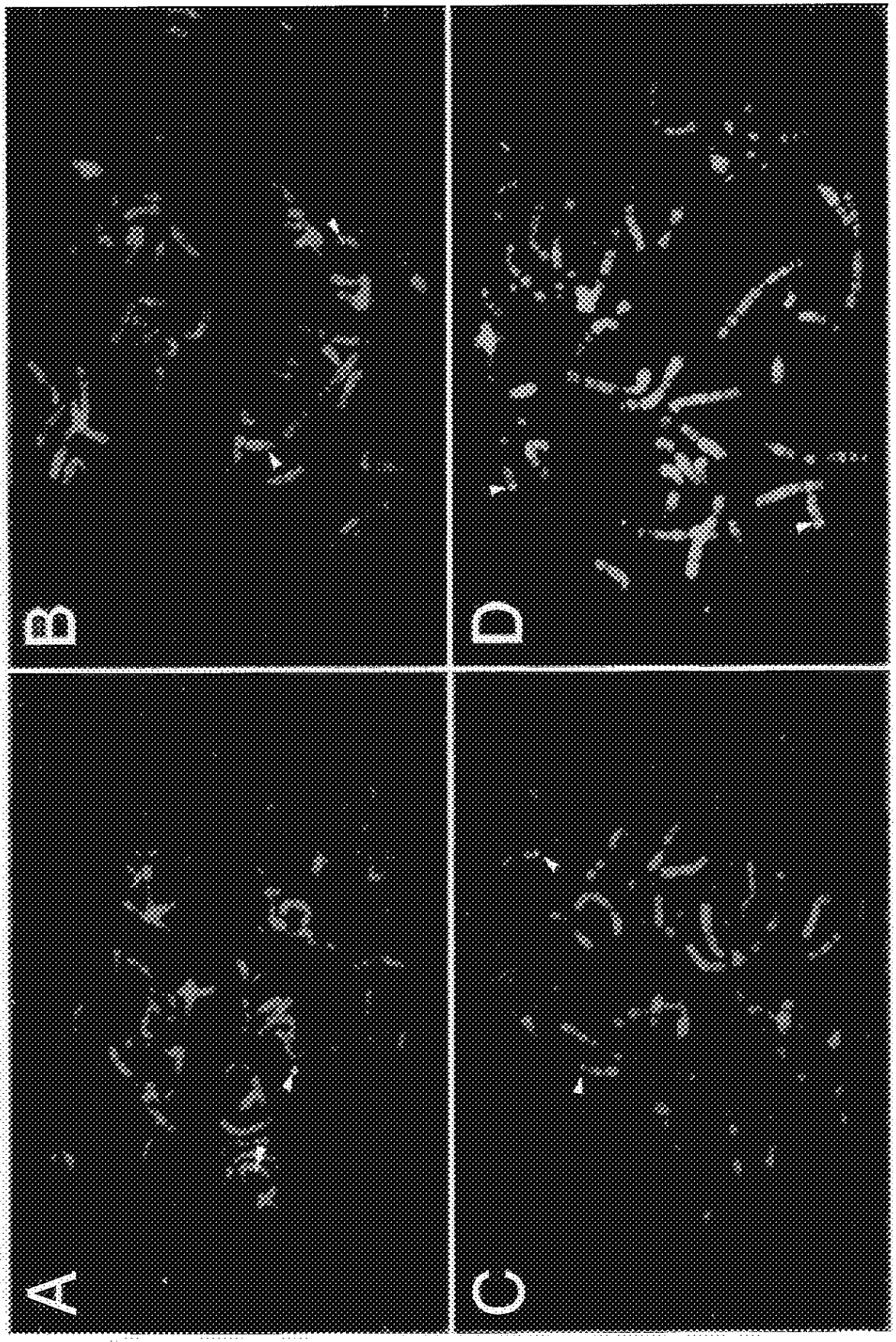


Figure 16

Chromosome bands and long-range GC% mosaic structures around the pseudoautosomal boundary (PAB1) of the short arms of the human sex chromosomes. Pseudoautosomal regions (PAR1) have been assigned to R-bands (Xp22.33 and Yp11.32) and, judging from the high density of CpG islands, a major portion of 2.6 Mb of PAR1 is thought to be GC-rich (Ellis and Goodfellow, 1989; Rappold, 1993; Schlessinger *et al.*, 1993). The boundaries with the neighboring Giemsa-positive subbands, Xp22.32 and Yp11.31, appear rather close to PABXY1 (Rappold, 1993; Schlessinger, 1993). PABXY1 and their neighboring sex-specific sequences, including SRY (about 5 kb apart from PABY1), are known to be AT-rich (Ellis *et al.*, 1990; Sinclair *et al.*, 1990). These observations suggest the PABXY1 to be near a boundary of the long-range GC% mosaic domains.

Figure 16

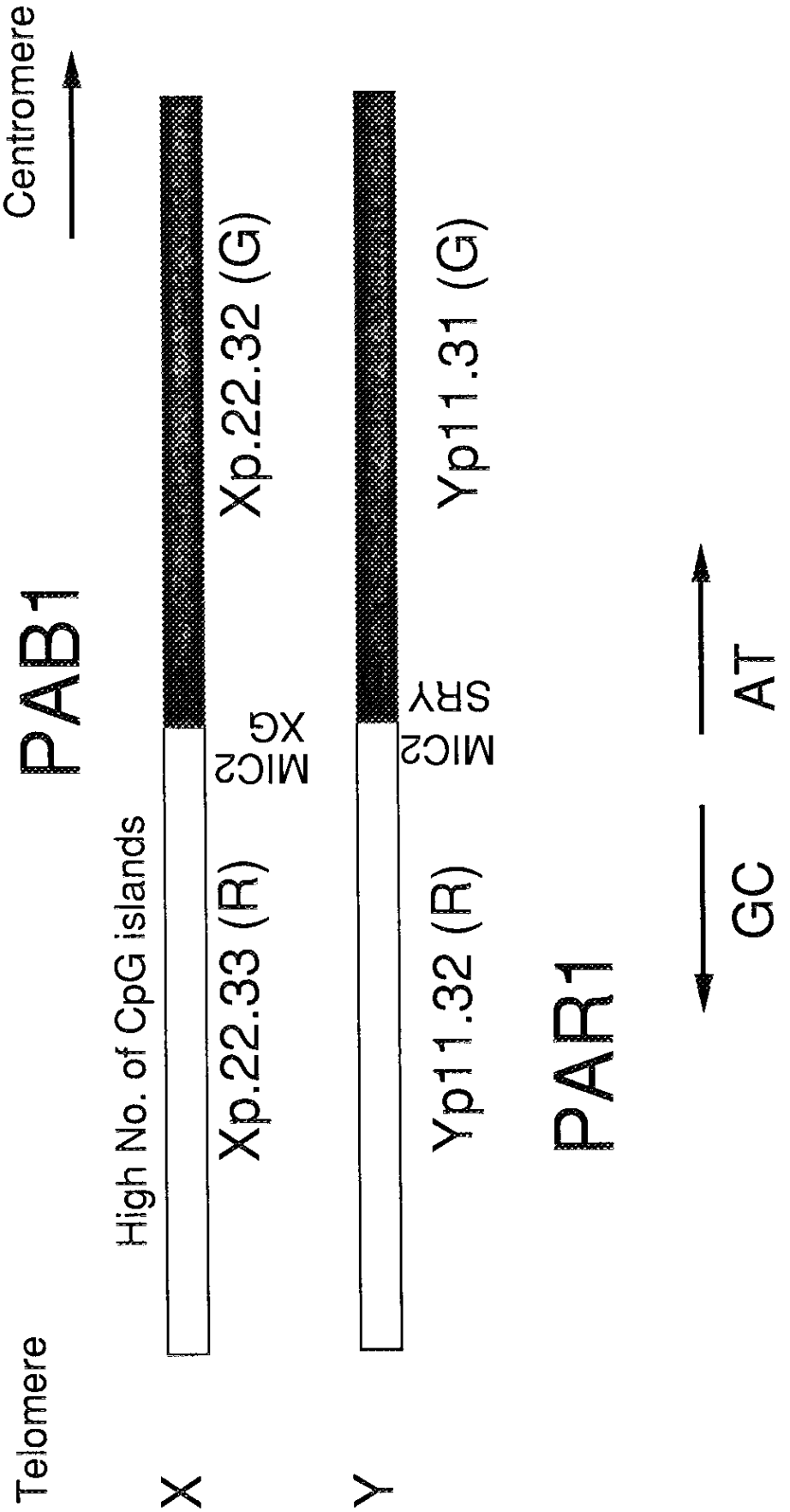
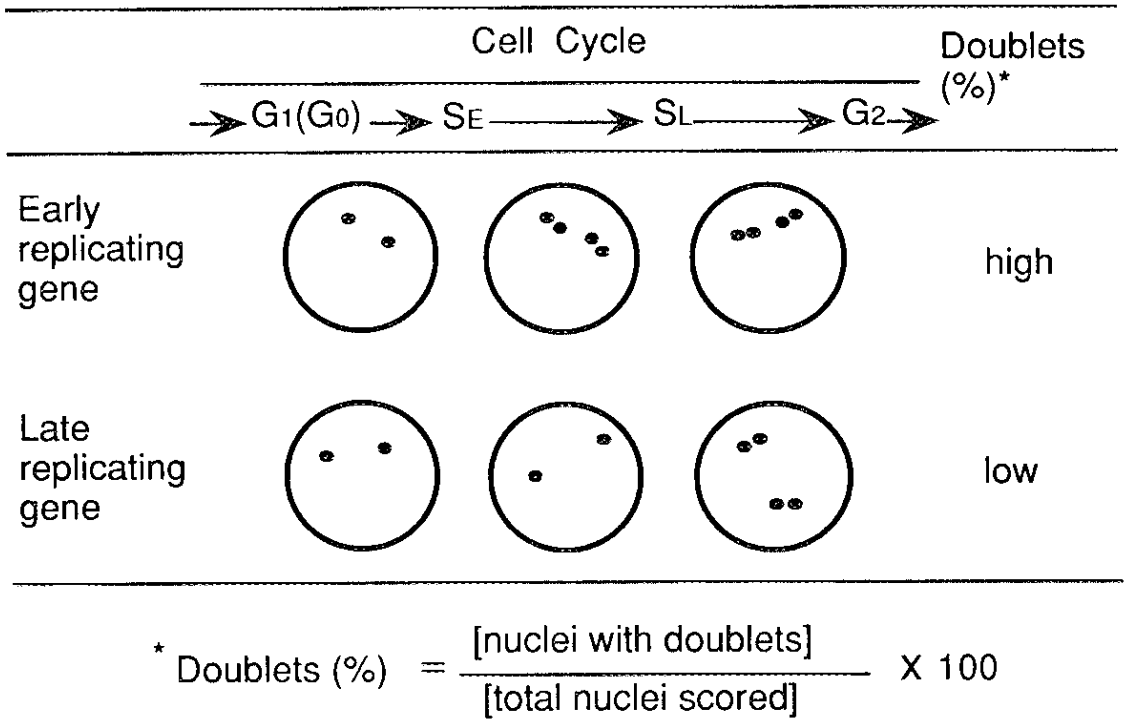

Centromere
PAB1
Xp.22.32 (G)
Yp11.31 (G)
MIC2
XG
SRY
MIC2
High No. of CpG islands
Xp.22.33 (R)
Yp11.32 (R)
PAR1
GC
AT
Telomere
X
Y

Figure 17

The principle of the analysis of DNA replication timing by fluorescence *in situ* hybridization (FISH) onto interphase nuclei. *In situ* hybridization to interphase nuclei was conducted using cloned cosmid or λ DNAs as probes. Singlet signals show the unreplicated genome portions, while doublet adjacent signals show the replicated portions. Three hundreds nuclei were examined and the doublet percentage was scored. A probe clone with the higher doublet percentage corresponds to the earlier replicating zone. When relative difference of replication timing for a certain pair of DNA segments was small, the two-color FISH method was used; the specimen with interphase nuclei was hybridized simultaneously with two different probes, each of which was labeled with biotin or digoxigenin. Utilizing different fluorescent colors of the probes, the replication pattern of both DNA segments in a single nucleus could be precisely compared by only changing filers of the fluorescence microscope. This multi-color FISH clearly showed their temporal order of DNA replication. By this procedure the relative replication time of each pair of DNA clones randomly selected from MHC region could be determined, and the result is shown in Fig. 18.

Figure 17

Determination of DNA replication timing by FISH



$$^{*}\text{Doublets (\%)} = \frac{\text{[nuclei with doublets]}}{\text{[total nuclei scored]}} \times 100$$

Figure 18

Differential DNA-replication timing in the MHC region found for HL60 cells. The GC-rich class III replicates earlier than the AT-rich class II, and PABL1 is located in the boundary region of the two distinct DNA-replication time-zone. The long-range GC% distribution is listed for the comparison. There is a close correlation between the temporal order of DNA replication, and GC% distribution in the human MHC. SE and SM show early and middle S phase, respectively.

Figure 18

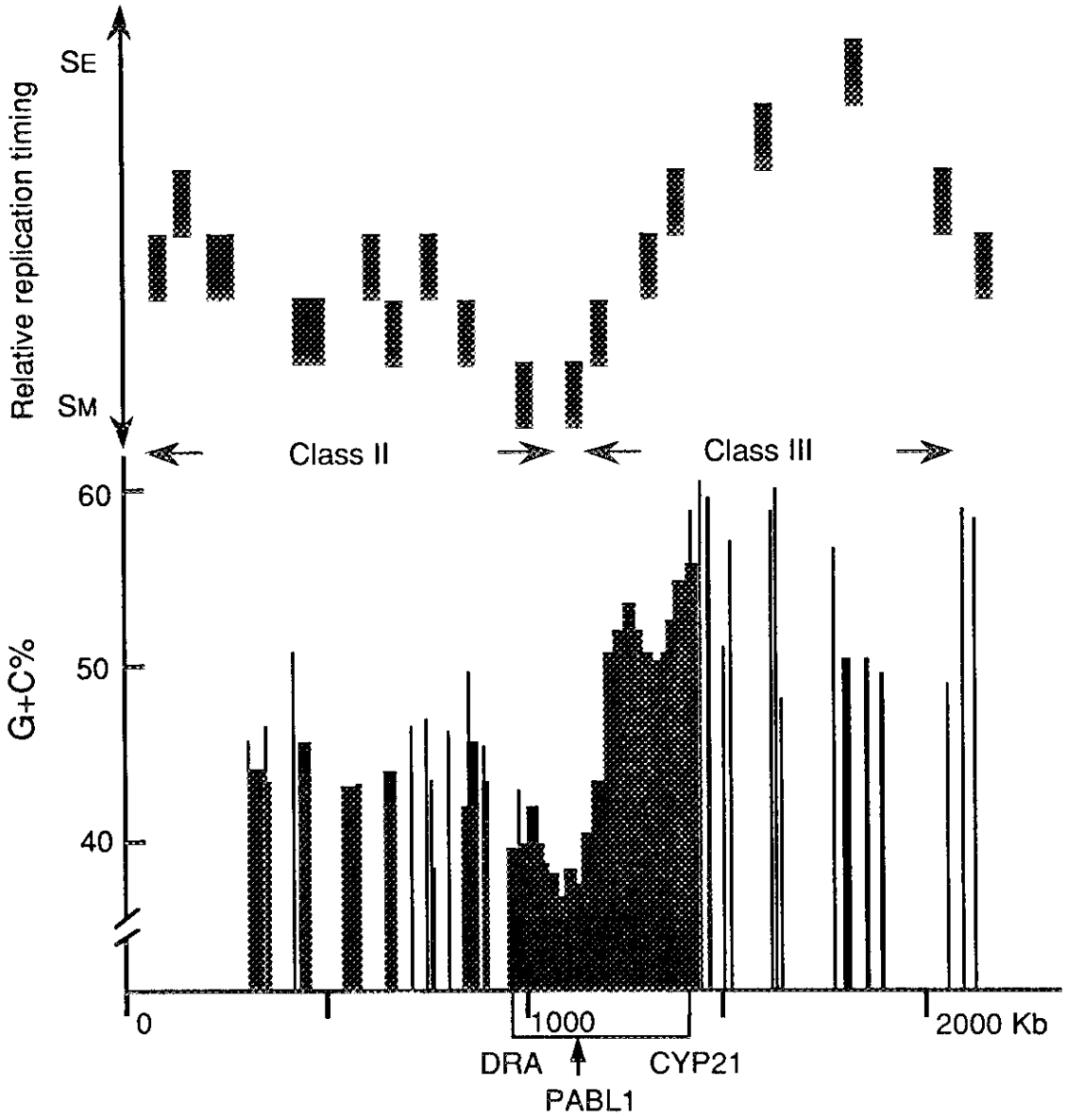
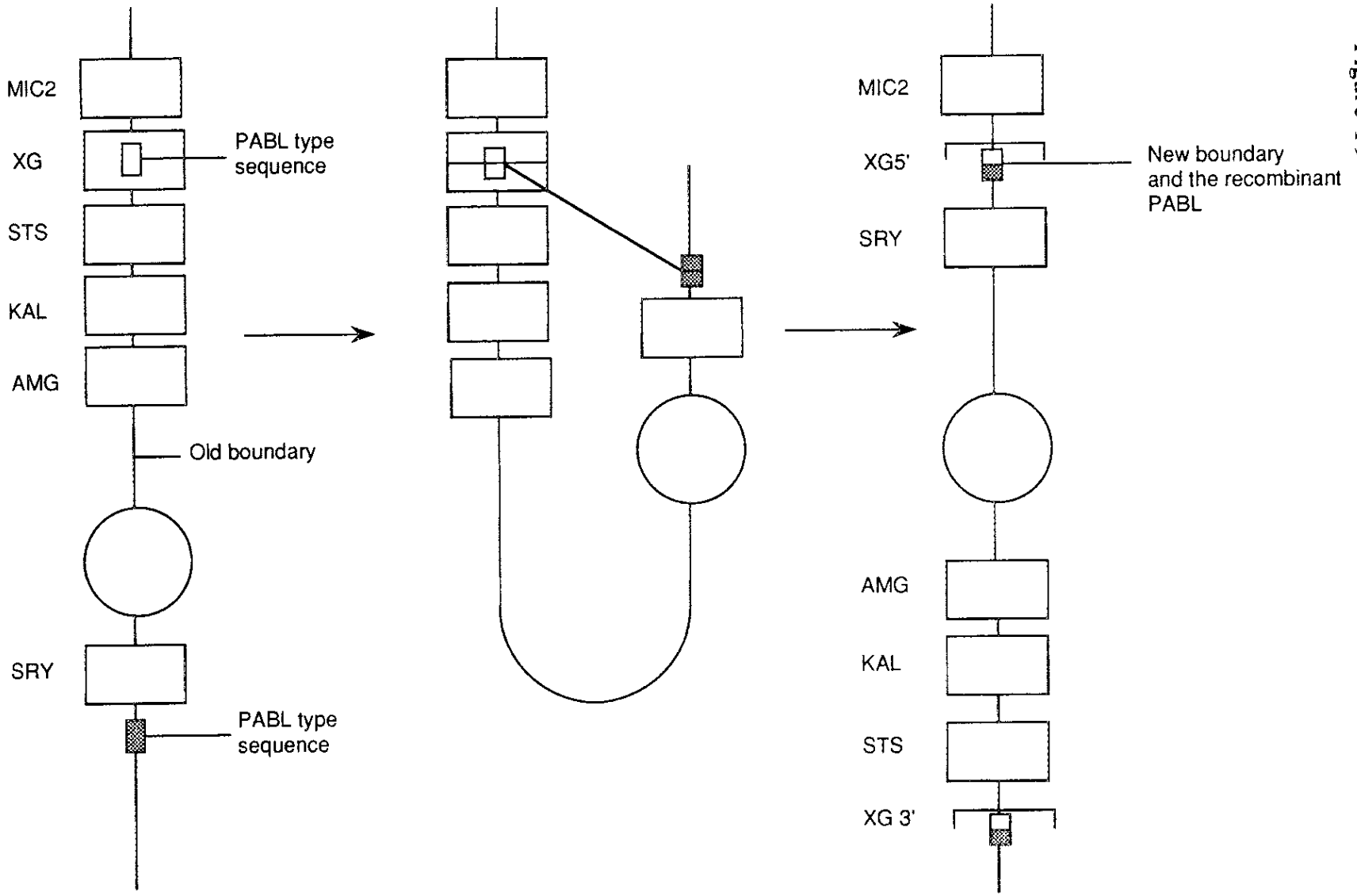

Relative replication timing
SE
SM
Class II
Class III
G+C%
60
50
40
0
1000
2000 Kb
DRA
CYP21
PABL1

Figure 19

Model of the evolutionary formation of the pseudoautosomal boundary PABY1. This is a modification of the model proposed by Ellis *et al.* (1994a). As a molecular process in their hypothesized pericentric inversion, an illegitimate recombination between two PABLs is postulated. (A) According to Ellis *et al.* (1994a), in early primates, PAB1 was hypothesized to be situated somewhere proximal to the present-day amelogenin gene on the X-chromosome (AMGX). The author supposes that, in the old Y chromosome, one PABL was located in the earlier XG and another was 5 kb distal to the earlier SRY. (B) After the divergence of higher primate and prosimian lineages a new boundary was supposed to be formed by a pericentric inversion on the Y chromosome initiated by breakpoints inside XG and 5 kb distal to SRY (Ellis *et al.*, 1994a). As a molecular process of this inversion, the author proposes an illegitimate recombination between the two PABLs above-noted. (C) An inverted Y chromosome with the present PABY1 was thus formed. Further evolutionary events such as another pericentric inversion transferring AMGY to the present location were described by Ellis *et al.* (1994a).



(A) Ancestral Y chromosome

(B) Crossing-over between two PABLs

(C) Pericentric inversion in Y chromosome and production of new boundary

Figure 19

Table 1. Properties of human chromosome G/Q, R, and T bands.

G/Q band	R band	T band (a subgroup of R band)
Late replicating	Early replicating	Very early replicating
AT-rich	Intermediate GC%	GC-rich
Low gene density	High gene density	Highest gene density
Low no. of CpG islands	Higher no. of CpG islands	Highest no. of CpG islands
L1 (LINE type)-rich	*Alu* (SINE type)-rich	Very *Alu* (SINE type)-rich
Compact chromatin structure	Loose chromain structure	Loose chromatin structure

Table 2

Table 2. Pairwise comparison of PABLs and PABXY1. Nucleotide identity (%) is listed.

	PABL1	PABL2	PABL3	PABLSp2	PABX1
PABL1					
PABL2	82.9				
PABL3	82.6	85.9			
PABLSp2	78.8	82.2	82.3		
PABX1	80.0	85.1	80.6	78.4	
PABY1	82.6	83.9	80.4	79.4	92.5

Table 3

Table 3. Pairwise comparison of PABLs and PABXY1. Nucleotide identity (%) for the upstream to the *Alu* insertion site is below, and that for the downstream is above the diagonal.

	PABX1	PABY1	PABL1	PABL2	PABL3
PABX1		77.8	77.9	82.2	74.7
PABY1	99.2		81.4	80.5	77.0
PABL1	83.3	82.4		84.1	82.4
PABL2	89.1	88.3	82.1		82.0
PABL3	88.4	87.6	86.5	89.2	

Upstream to *Alu* insertion site

Downstream to *Alu* insertion site

Table 4. Lowest free energy levels for the PABL core sequences and averages for those of the randomized sequences along with their standard deviation.

a) PABL sequences

	Length	GC%	Lowest free energy for PABLs	Lowest free energy for shuffled 100 seqences		Δenergy/SD
				Average	SD	
PABL1	638	42.5	-151.60	-146.84	5.64	0.89
PABL2	662	43.1	-171.30	-145.11	5.30	4.94
PABL3	668	43.4	-161.40	-153.94	6.63	1.13
Bc4 core	337	40.4	-75.70	-67.79	3.98	1.99
Mo1 core	538	41.1	-138.30	-117.19	5.37	3.93
Mo2 core	555	49.2	-145.90	-137.43	5.58	1.52
Sp2 core	629	42.6	-151.90	-143.57	6.00	1.39
Sp3 core	453	41.3	-105.70	-97.23	4.92	1.72
PABX1	620	42.7	-159.70	-137.93	5.78	3.77
PABY1	628	41.7	-154.70	-136.65	6.12	2.95

b) Control sequences

7SL RNA	303	63.0	-113.30	-105.80	4.34	1.73
12S rRNA	954	45.4	-200.70	-192.38	8.57	0.97
IL4 mRNA	462	52.4	-122.50	-121.94	5.10	0.11
CYPDB1 mRNA	1494	63.9	-566.60	-569.64	9.09	-0.33

Table 5

Table 5. Evolutionary rates of PABLs and PABXY1. All rates are in units of 10^{-9} substitutions per site per year. Divergence time of PABLs is assumed to be 60-120 million years.

	Upstream	Downstream
PABL1	1.2±0.4	1.5±0.5
PABL2	0.6±0.2	0.4±0.2
PABL3	0.7±0.3	1.3±0.4
Sp2	0.9±0.3	0.8±0.3
Sp3	0.1±0.05	1.5±0.5
PABX1	0.9±0.3	1.0±0.4
PABY1	0.9±0.3	1.5±0.5

Table 6

Table 6. The percent (%) divergence in pairwise comparison of sequences around pseudoautosomal boundary 1 of X chromosome.

	PABL sequence		Sex-specific
	Upstream to *Alu* insertion	Downstream to *Alu* insertion	
HumX/ChiX	1.5	1.2	1.0
HumX/GorX	1.0	0.6	2.1
ChiX/GorX	0.5	0.6	1.7
HumX/OraX	4.1	1.2	5.5
ChiX/OraX	4.6	1.2	5.1
GorX/OraX	4.1	0.6	6.2
HumX/BarX	10.7	8.1	11.6
HumX/GelX	9.6	6.9	10.8
HumX/GueX	11.2	9.4	14.2
ChiX/BarX	10.2	8.1	11.2
ChiX/GelX	9.1	6.9	10.5
ChiX/GueX	9.6	9.4	13.9
GorX/BarX	10.7	7.5	12.2
GorX/GelX	9.6	6.2	11.5
GorX/GueX	10.2	8.8	14.5
OraX/BarX	9.6	8.1	10.8
OraX/GelX	8.6	6.9	10.6
OraX/GueX	10.2	9.4	12.8
BarX/GelX	2.0	4.4	3.8
BarX/GueX	4.6	6.2	5.1
GelX/GueX	3.6	5.6	5.8

Table 7. Map positions of genomic PABL clones and band characteristics.

	Map position	Band type
PABL1	6p21.3	T
PABL2	20p11.21	ordinary R
PABL3	19q13.3	T
PABL4	3p21.3	T
PABL5	6p21.3	T
PABL6	6p21.3	T
PABL7	20p11.2	ordinary R
PABL8	14q24.3	ordinary R
PABL9	11q25	T and terminal R
PABL10	21q22.3	T and terminal R
PABLSp2	17q23	T
PABX1	Xp22.33	terminal R
PABY1	Yp11.3	terminal R

APPENDIX

Nucleotide sequences of PABLs shown in this section have been deposited with the DDBJ / EMBL / GenBank Data Libraries under Accession Nos. D30042 (PABL1), D30043 (PABL2), D30044 (PABL3), D55639 (Mo1), D55640 (Mo2), D55643 (Sp2), D55644 (Sp3), D55641 (Sk13), D55638 (Bc4), and D55642 (PABLSp2).

PABL1 (D30042)

```
CTTACTCTTT GCCTTACTGT GGCAGGGCAG GTCTCGCTAA CGCAGGCCTC 50
CATAACAACT GTTTCAGCAC TGACTGAGTG GTTAAGTTAA ATGTTGAAAG 100
CTGATAGAGC CAGGCTAGAA TGTAACAAGC CCACCAAGAG TTTGCCTAGG 150
CCTTTCCTGG GCCTTGAAGC ATGACAAGAT TACGAAGGAA TTCTTAACAG 200
GACCCGTTTA GGATTAAAAC AAGTTTATTG GGGGGTCTGA AGAAACTCCC 250
CAGGCCTTCA CAAACAAGTT TATTGGGGGT CTTCTGAAGG AACTCCATAT 300
TTAGCAGGAG ACAAGATAAG GGTAATCACT CCAGCACCTG GACCCATTTA 350
GATTAAGTAA ATTTACTGAA GCTCTAGAGG AAAGCCTTCA GGACTCACAT 400
CTTAGTCACA GATTAGAAGA AGTTAATGAC TTATGTCTTT AGATGAATGC 450
ACACTTACAC GTAGACATAT AGCTTAGAAG GTATATTGGC TCTGGAAAAC 500
TTTGTAATTT TCAGTTGGTC TGGCAAAAAT TTCCAGGCCT TCTCTCTGTA 550
CCTACTTATA TAAATAAAAA CTGTCTTCTT TCTCAGTTCA TCTGCATCTC 600
GTTATTGGGC CATGAAGAAA AGCAGCCCGA TTCTCCTACC TCAGCCTCCC 650
AAGTAGCTGG GATTACAGGT GTGTGCC                          677
```

PABL2 (D30043)

```
CCTTACCAGC CCCCTACTGT GGCAGGCCAG GTCCCACTAA CACAGGCCTC 50
CATAACAACT GTTTCAGCTC TGACTGAGTG GTTAAGTTAA ATACTAAAAG 100
CCGAGAGAGC CAGTCCCCTT ATACAGAGGC TGGAATGTAA CAAAAGCCCA 150
CCAAGAGTTT TGCCTAGGCC TTTCCTGGGC CTTGAAGCAT GACAAAATAA 200
```

CAAAAGAATT CTTAACAGAA TCTATTTAGG ATTAAACAAG TTTTACTGGG 250

GGTTCTGAAG AAACTCCCCA GTCCTCCACA AACAAGTTTA CTGGGGTCTG 300

AAAGAACTCC CCAAACCTCC ATGATTTAGC AGGAGACAAG ATAAGGGTAA 350

TCACCCCAGC ACCTGGACCC AGCTAGATTT AGTCAATTTA CTGAGGCTAC 400

AAAGGAAGGT CTTCAGGACT CAGACCTCAG TTATAGATTA GAAGAAGTTA 450

ATCACTTATG TCTTTAGATG ATTGCACACT TACACATAGA CATATAGCTT 500

AGAAGGTGTA TAAGCTCTGG AAAACTTTGT AATTTTGAGT TGGTCTGGTG 550

ATATTTTCCA GGCCTTCTCC CTATACCCGG TTACAGAAAT AAAAACTTTC 600

TTCCTCCCCA GTTCATCTGC ATCTTGTTAA TGGGCTGCCA GAAATAGCAG 650

CCCAACCCTC AGTTAGGTCT GGGAACACTA CCCCCCAACA CACACACACA 700

CAC 703

PABL3 (D30044)

TGGTGGTTAG GGGAGGGTGT GGCAGGTCAG GTCTCCCTAA CCGCTGAACA 50

GGCAGGCCTC CATAACAACT GTCTCAGCAC TGATTGAGTG GTTAAGTTAA 100

ATATTAAAAG CTGACAGAGC CAGTGCCCTC ATACAAAGGC TGGAATGTAA 150

CAAAAGCCCA CCAAGAATTT TGCCCAGGCC TTTTCTGGGC CTTGAGCATA 200

ACAAGATAAT GAAGGAATTC TTAACAGGAC CCGTTTAGGA TTAAACAAGT 250

TTTATTGGGG GGTCTGAAGA AACTCCTCTA AGCCTCCACA AACAAGTTTA 300

TTGGGGGTCT GAAGGAACTC CCCAAACCTC CATGATTTAG CAGGAGACAA 350

GATAAGGGTA ATCACCCCAG CACCTGGACC CATTTAGTTT AAATAAATTT 400

ACTGAGGCTA CAGAGGAAGA TCTTCAGGAC TGACATCTTA GTTACAGATT 450

GGAAGAAGTT AATCGCTTAC GTCTAGATGA ATGCACTCTT ACATGTAGAC 500

AAATAGCTTA GAAGGTATAT GAGCTCTGGA AAACTTTGTA ATTTTGAGTT 550

GGTCTGGCAA TATTTTCCAG ATCTTCTCCC TGTGCCCAGT TACAGAAATA 600

AACTCCCTTC TCTCCCAGTT CACCTGCATC TTGTCATTGG GCTAAGAGAA 650

TAAGCAGCCT GACCCTTGGT TTGGTCCAGG AACAAGGGTG CATGGTCAGA 700

TGTTGAG 707

Mo1 (D55639)

GAATTCCGTT TTTTTTTTTT TTCTGTAAGG TATATTATTT ATTGTTAGTT 50

ACATGTGGTC AATGGTGACA TACTTTCAAT AATTAAAAAT CGAATAATAC 100

TGAAATAACC ACAGCAGCTT TCAGTATAAT TTGCTTAAGT TGTTCTAGAA 150

AACACTGCTA ATTTTTTGTT TCTGCAGAGT CAAGTATAGA AGTGAGCAGA 200

TTGGTAAATT TATAAACATC ATGGAAAGAA TTATAGAGAG GAACTGGAGC 250

CACTCGAATG CCATTTGGAT TCCGCTTGTC ACAAACCACT CCTCTTTTTT 300

CTAGTTCTTG GAAAACATCT TCGTTTGGAA CAGAAAATGT TATTGTTAGC 350

TGGCACCCCC GCTCCTCTAC ATGAGACGGA GTAATTATGT TCACAACTGG 400

TTTCTTGGTT GCTGCTTTAT CTTTGCCATA GTTATGCTTG ATCAGGTATT 450

CCAGATAGCC AGTTAGCAAA ACAGATTTTT TCCGCAATGC CTTCATTGTC 500

GCTTGCTTAA AGATCTGTGG AATGAGAAAC AATCAGATTA AAGTGTCTTA 550

TTGATTTACA AAGAAGCAAA AATTAAGAGT AAATCATGCA TAAAAAGGAA 600

ATGCCCATAT TTGATCTTTC CCAATAAATA TCTCCCTTTA CATAATCTTT 650

TTGTGTAAAA TCATGCCCTC CTTTAAATGC TTTGCTATTA AATACTGAAG 700

GTAAAATTTC CATGACAGAA AAATTGTACA AATGTCACTT AAATAAAACT 750

TGCAAAGGAA GAAAATGTAA GTTAATATCA TGTACCTGTG GCAGGCCAGG 800

TTTCACTAAT GCAGGCCTCC ATCACAACTG TTTCAGTACT GACCGAGTGG 850

TTACGTTAAA TATTAAAAAC TAAAAAAGTC AGTGCCCTTA TACAAAGGCT 900

GGGATGAACA AAAGCCTATC AAGAGTTTTG CCTAGGCTTT CCCTGGGCCT 950

TAAAGCATGA CAGAATAATG AAGGAATTCT TAACAGGACC CATTCAGGAT 1000

TAAACAAGTT TTACTGTGGG TTGTGAAGAA ACTCTCCAGG CCTCTACAAA 1050

CAAGTTTATT GAAGGTCTAA AGGAACTCCT CAAACTTCAG TGATTTAGCA 1100

GAAGACAAGA TAAGGGTAAT CACTCCAGCA CCTGGACCCA TTTAGATTAA 1150

GTAAATTTAC TGAGGCTCCA GAGGAAGGTC TTCAGGACTC AGACCTTAGT 1200

TGTAGATTAA AAGAAGTTAA TCACTTATGT CTTCAGATGA ATGCACACTT 1250

ACACCTGTAA AACTTTGTAA TTTTGAGTTG GGTCTGGTGA TAATTTTCCA 1300

GGCCTTCCTC CCTGCGGGAA TTC 1323

Mo2 (D55640)

GAATTCCGTC ACTCGGGCCC ACATGTGCCA AGGGGGGCTG GTTTTGGGGC 50

CGGCGGGTTC TGCCTGATGC TCAGAGGGTA ACTGGATGCT GAAAACTGTG 100

AGTCTTTCTT CAACTCAGGG GAATGTTTCC AGGGCACGCC AGGACTCACT 150

CACGCAGGCC TCCGCGACAA CTGTTCAGCA CTGACTGAGG GTGAAGTGAA 200

ATCCTGAAAG CTGAGAGCCA GCGCCCTCAC ACGAGGGCTG GGACGTAACA 250

AAAGCCCATC AAGAGTTTTA GGCCCAGGGC TTTCTTGGGC CTTGAAGCAT 300

GACGAGACCA GGACCCGTTT AGGATTAAAC AAGTTTTACT GGGGGTCTGA 350

AAAAACTCCC CAGGCCTCCA CAAACAAGTG GAGAAGGAAC TCCCCAAACC 400

TCCATGATTT AGGAGAAAAC AAGATAAGGG TAAGCATCTG GGCCCATTCT 450

AGATCAAGTA AATTTACTGA GTCTTCAGAG GGAGGTCTTC AGGACCAAAC 500

CTCAGTTAGA GATGAGAAGA ATTGAATCAC GTACGTCTTT AGACGAATGC 550

ACACTGACAC GGAGGACACA GAGGTTAGAA GGTGTGTAAG CTCTGGAAAA 600

ACTGTAATCT GGAGTTGGTC TGGTGATCAT ACCAGGGCTT CTCCCTGTTA 650

CCAGTTACAG AAATAAAATC CCTCTTCTTC CCCAAAAAAA AAAAAAAAAA 700

AACCCCCCCC CCCCCCC 717

Sp2 (D55643)

TCCTCCATGA TACTATAAAT GTATCAGGTC CACTTTTTTC TTCTTACTTT 50

AGGTGATCCT GTGGCAGGCC AGGTCTCACT AACAGCTGAA CAGGCAGGCC 100

TCCATGGCAA CTGTTTCAGC ACTGAGTAGT TAAGTTTAAT AGAGCCAGTG 150

TCCTCATACG TAGGCTGGAA TGCAACAAAA TCCCACCAAT AGTTTTGCCT 200

AGTCCTTTCT TGGGCCTTGG AGCATGACAA GATAACGAAG GAATTCTTAA 250

CAGGACCCCT TTACAATTAA ATATGTTTTA TTGGGGGCCT GAAGGACCTC 300

CCCAGACTGC CACAAGCAAG CTTTACTGGG GACTAAAGGA ACTCCCCAAA 350

CCTCCATGAT TTAGTAGGAG ACAAGATAAG GGTAATCTCT GTGGTGCCTG 400

GACCCATTTA GATTAAGTAA ATTTACTCAG GCTCCAGAGG AAGGTCTTCA 450

AGACTCAGAT CTTAGTTATA GATTAGAAGT TAATCACTTA TGTCTTTAGA 500

TGAATGCACA CTTACATGTA GACATATAGC TTAGAAGTTA TATAAGCTCT 550

GCAAAACTTT GTAATTTTGA GTTGGTCTGG TGATATTTCC CCAGCCTTCT 600

TCCTATACCC GGTTACAGAA ATAAACTCTG TTCTTTCCCA GTTCATCTTC 650

ATCTCGTTAT TGGGCCACAA AAATATGCAG CCTGACCCTT GGTTTGGTCC 700

GAGAACAATT CCACTATAAA ATGTGGATTT ATTATTGAAT TATATCCTGT 750

CTTTGAGGAG CAAGTAATAG GATTCACAAT TCAATCTCCT GTACCCCATG 800

TATTAATTGG CTTACATGGC TGGAAGTCCA TGTCAGACAG GGAAGGCTTT 850

GCTTCAGGCC TAATCCAGTT TCATTGTTCT CTGACTTTGT CATCTGCCCC 900

TCTTCTGTGT GCTAGCTTCA TCCAGAAGTT AGCCTCTTTC ACAACAGTAA 950

AATGCATACA GCAATTTCAG GCTTTCCACA CACTAAGTTT CCCAGAGATG 1000

AGCAGAGTTT CTCTATCTGT GGCTCATACG CTAAGATATT TCTTCCCAGA 1050

AGCTT 1055

Sp3 (D55644)

GAATTCTTAA CAGGACACGT TTAGGATTAA ACAGGTTTTA CTGGGGGTCT 50

GAAGAAACTT CCCAGACCTC CACAAACAAG TTTTATTGGA GGATCTAAAG 100

GAACTTCCCA AACCTCCATG ATTTAGCAGG AGACAAGATG AGGGTAATCA 150

CCCCCGGCAC CTGGACCCAT TTAGATTAAG TACATTTACT GAGGCTCCAG 200

AGGAAGGTCT TCAGGACTCA GACCTTAGTT ATAGATTAGA AGAAGTTAAT 250

CACTTATGTC TTTTAGACAA ATGCACACTT ACACATAGAT GTATAGCTTA 300

GAGGTTATAT AAGCTCTGGA AAACTTTGTA ATTTAGACTT GGTCTGGTGA 350

TACTTTCCCA GCCTTCTCCC TATACCCGGT TACAGAAGTA AACTCTCTTT 400

TTTCCCAGTT AGTCTGCATC TTGCTATTGG GCCACAGGAA TAAGCATCTC 450

GACCCTCTCT TTGGTTTGGG AAAAGTTTGG CTGTTGTCTG TTATGGGAAG 500

GTCTGGGAAA AGATCTCGAC TGCTTAATTG GAAGAGTCAG AACTACAGTA 550

ACCATTCCAC AGACTAGAAC ACAGAAAGAA ACAGGTCT 588

Sk13 (D55641)

GAATTCCGTT TTTTTTTTTT TTTTTTTTTT TTTTTTTTTT TTTTTTTTTT 50

TTTTTTTTTT TTTTTTTTTT TCGGAATTCC GGCCACAACC AAGATAGAGG 100

AAATTCCCTA TGCCCAAAAG TTTCTTGTAC CCCTTTGCAG TTCATCAGCC 150

TTTCAACCCT CAGCCCCAAG GAGCCACTGT CACTTCAGGT CCGTTTGCAT 200

TTTTAACCAT TTTCTATAAA TGAAATGATA CCTGTGTTCT TTGGTGTCTG 250

CCTTCTTTCA TGCATCAATG TAATTTTAAA ATCCATCTGT AATGTCTTGT 300

AAATACTGAG TAGTATTCCT TCATGTGGCT ATACCATGTA TGTTGGGACT 350

TTCACTTATT ATTGGGACAT TTCTATCATT CCACGTTTGG GCTATTATGA 400

AGAAACTATC ATGAGCATCC ATACGTAGCA GGCCAGGTCT CACTAACACA 450

GGCCTCCCTA AAAACTGTTT CAGTAGCGAC TGAGTGGTTC AATTAAATAT 500

TAAGAGGGAA AAAGGAAAAA AAAAAAAGGA GGCCAGCGCC CTTTTCAAAA 550

GGTTGGATTG TAAAAAAAGC CCACCAAGAG TTTGGGCTAG GTTTTTCCTG 600

GGCCTTAAAG CATGACGAAA TAACGAAGGC ATTCTTAACA GGAGCCATTT 650

AGTATTAACC GAGTTTCACT GGGGGTCTGA AGAAACTCCC CAGGCCTCCA 700

CAGACAAGTT TATTGGAGAT CTGAAGGGAC TCTCCAAACC TCTGTGATTT 750

AGCAGGAGAC AAGATAAGGG CCCCCAGCAC CTAGACCCAT TTAGATTAAG 800

TGAATTTAAC TGAGGTTCCA GAGGAAGGTC TTTAGGACTC AGACTTAGTT 850

ATAGATTAAA AGAAGTTAAT CACTTATGTA TTTAGATGAA TCGGAATTC 899

Bc4 (D55638)

GAATTCCGCA GGAGACAAGG TCAGGGTAAT AGCCCCAGCA CCTGGACTCA 50

TTTAGATTTA AGTAAATGTA CTGAGGCTCC AGAGAAAGGT CTCCAAGACT 100

TAGACTAGAG TTACAGATTA AAAGAAGTTA ATCACTTATA TCTTTAGATA 150

AATGCACACT GACACGTAGA CATACATCTT AGAAAATATA TAAGCTCTGG 200

AAAACCTTGT AATTTACAGT TGGTCTGGAG ATATTTTCTG GGCCTTCTCT 250

CTGTAACCGG TTCCAGAAAT AAAAAACTCT CTTCCTCCCC AGTTTATCTG 300

CATCTCGTTA TTGGGCCGTG AGAAATAGCA GCCAGACCCT GAGTTTGGTC 350

TGGGAACACA CTGGCTGGGC TTCCTGCCTG GCGGTAGAGC TTCACTGTCC 400

TCCTTCCCAG CCAAAATGTG GCCCTTGAAG GCTTCCTCCC CCAACTGACC 450

GGTTTTGGGT TCTAGGTACT CAGGGAACAC CGTCTTCAAG GAGGCCATTT 500

TCATACTGTT CCTGGTTTGA TTTTTCTGAA ATTAGGAAAT GCCCCTTGAG 550

GCATCCAGGA TTGTCTTTAT TTCCCCTGAA GCTTCCCCTC CTGAAGAAAA 600

AAGTTCTCAA ACACCCTCAG GACTCAAACC CCAAGTTCAC CCTGTGTAAA 650

CATTCCCATT TCCTAAACCC AGGCAGGCTC ATGGCCCTCA GCCTCTCAGA 700

CCCCATGGAA AGGGGGGGGG GGGGGGGGGG GG 732

PABLSp2 (D55642)

TGGAAAGCAG AGGTTGCAGT GAGCCGAGAT TGTGCCATGG CACTCCAGCC 50

TGGGCAACAG AGCAAGACTC CATCTCCAAA AAAACAAACC AAAAAAATTG 100

ATACTCGTTC ATAAGTCCTT TCCTCCATGA TACTATAAAT GTATCAGGTC 150

CACTTTTTTC TTCTTACTTT AGGTGATCCT GTGGCAGGCC AGGTCTCACT 200

AACAGCTGAA CAGGCAGGCC TCCATGGCAA CTGTTTCAGC ACTGAGTAGT 250

TAAGTTTAAT AGAGCCAGTG TCCTCATACG TAGGCTGGAA TGCAACAAAA 300

TCCCACCAAT AGTTTTGCCT AGTCCTTTCT TGGGCCTTGG AGCATGACAA 350

GATAACGAAG GAATTCTTAA CAGGACCCCT TTACAATTAA ATATGTTTTA 400

TTGGGGGCCT GAAGGACCTC CCCAGACTGC CACAAGCAAG CTTTACTGGG 450

GACTAAAGGA ACTCCCCAAA CCTCCATGAT TTAGTAGGAG ACAAGATAAG 500

GGTAATCTCT GTGGTGCCTG GACCCATTTA GATTAAGTAA ATTTACTCAG 550

GCTCCAGAGG AAGGTCTTCA AGACTCAGAT CTTAGTTATA GATTAGAAGT 600

TAATCACTTA TGTCTTTAGA TGAATGCACA CTTACATGTA GACATATAGC 650

TTAGAAGTTA TATAAGCTCT GCAAAACTTT GTAATTTTGA GTTGGTCTGG 700

TGATATTTCC CCAGCCTTCT TCCTATACCC GGTTACAGAA ATAAACTCTG 750

TTCTTTCCCA GTTCATCTTC ATCTCGTTAT TGGGCCACAA AAATATGCAG 800

CCTGACCCTT GGTTTGGTCC GAGAACAATT CCACTATAAA ATGTGGATTT 850

ATTATTGAAT TATATCCTGT CTTTGAGGAG CAAGTAATAG GATTCACAAT 900

TCAATCTCCT GTACCCCATG TATTAATTGG CTTACATGGC TGGAAGTCCA 950

TGTCAGACAG GGAAGGCTTT GCTTCAGGCC TAATCCAGTT TCATTGTTCT 1000

CTGACTTTGT CATCTGCCCC TCTTCTGTGT GCTAGCTTCA TCCAGAAGTT 1050

AGCCTCTTTC ACAACAGTAA AATGCATACA GCAATTTCAG GCTTTCCACA 1100

CACTAAGTTT CCCAGAGATG AGCAGAGTTT CTCTATCTGT GGCTCATACG 1150

CTAAGATATT TCTTCCCAGA AGCTTCCAGC AAAACTCCCC TCGAGTCATA 1200

TGTGTGTGAA TAGAGTCATA AGCCAACCTC TGAATCAATC TTTGTATCAA 1250

AGTGATCAAA GTGGGTGGAA TTAACTCTGG ATCAAATGAT GCCTCATCCT 1300

TGGGAACTAG TAATGAAGTC AAGTTTCTGA TCAAGCTTCC CTTAACAGAA 1350

GTCAAACAAA ATGATCAAAG TGGGTGGAAT TAACTCTGGA TCAAATGATG 1400

CCTCATCCTT GGGAACTAGT AATGAAGTCA AGTTTCTGAT CAAGCTTCCC 1450

TTAACAGAAG TCAAACAAAA 1470

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor, Prof. Toshimichi Ikemura, for his continuous guidance and encouragement during all the stages of this work.

I greatly appreciate Prof. Naruya Saitou for guidance of evolutionary computation and consideration. I also thank Dr. Kimihiko Sugaya, Mr. Yasukazu Nakamura, Dr. Toyoaki Tenzen, Dr. Ken-ichi Matsumoto and other members in the Department of Evolutionary Genetics for discussions and comments. I am very grateful to Dr. Nathan A. Ellis for kindly providing me with his unpublished sequence and valuable comments. I also thank Prof. Katsuzumi Okumura, Dr. Asako Ando, and Prof. Hidetoshi Inoko for their collaboration and discussions. I also thank Mrs. Yoko Miyauchi for her technical assistance.

I also thank Director Jun-ichi Tomizawa, Prof. Kensuke Horiuchi, Prof. Takashi Gojobori, Prof. Susumu Hirose, Prof. Takeshi Seno, Prof. Toshihiko Shiroishi, and Prof. Asao Fujiyama, for serving on my supervisory committee.

In this work, computers at the DDBJ and the Human Genome Center of Japan were used. This work was supported by a JSPS Fellowship for Japanese Junior Scientists (No. 1675) and by a Grant-in-Aid of Scientific Research from the Ministry of Education, Science and Culture of Japan.

Finally, I would like to express my sincere appreciation to my parents Takeo and Toshiko Fukagawa. They gave me this opportunity and their warmest support for this work.

September 1995

Tatsuo Fukagawa