(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 March 2002 (21.03.2002)

PCT

(10) International Publication Number
WO 02/22830 A2

(51) International Patent Classification[7]: C12N 15/54, 9/10, C12Q 1/48, C07K 16/18, G01N 33/53

(21) International Application Number: PCT/GB01/04120

(22) International Filing Date:
14 September 2001 (14.09.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

0022768.6	15 September 2000 (15.09.2000)	GB
0111995.7	16 May 2001 (16.05.2001)	GB

(71) Applicant *(for all designated States except US)*: **UNIVERSITY COLLEGE CARDIFF CONSULTANTS LTD.** [GB/GB]; 55 Park Place, Cardiff CF10 3AT (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **AESCHLIMANN, Daniel, Peter** [CH/GB]; Connective Tissue Biology Laboratories, School of Biosciences, Museums Avenue, Cardiff CF10 3US (GB). **GRENARD, Pascale, Marie** [FR/GB]; Connective Tissue Biology Laboratories, School of Biosciences, Museums Avenue, Cardiff CF10 3US (GB).

(74) Agents: **TOMBLING, Adrian, George** et al.; Withers & Rogers, Goldings House, 2 Hay's Lane, London SE1 2HW (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/22830 A2

(54) Title: TRANSGLUTAMINASE GENE PRODUCTS

(57) Abstract: The invention provides a nucleotide sequence comprising at least a portion of the nucleotide sequence of Fig. 10A, Fig. 6B or Fig. 10A or Fig. 10B; nucleotides which hybridise to the nucleotide sequences of Fig. 6A, Fig. 6B or Fig. 10A or Fig. 10B; nucleotides which are degenerate to the nucleotide sequences of Fig. 6A, Fig. 6B or Fig. 10A or Fig. 10B; all of which nucleotides encode a polypeptide having transglutaminase activity.

Transglutaminase Gene Products

The present invention relates to the identification of novel transglutaminase enzymes TG_Z and TG_Y.

Transglutaminases are a family of structurally and functionally related enzymes that catalyze the post-translational modification of proteins via a Ca^{2+} dependant transferase reaction between the γ-carboxamide group of a peptide-bound glutamine residue and various primary amines. Most commonly, γ-glutamyl-ε-lysine cross links are formed in or between proteins by reaction with the ε-amino group of lysine residues. Analysis of the three-dimensional structure of the a-subunit of factor XIII showed that transglutaminases contain a central core domain containing enzymatic activity, and a N-terminal β-sandwich domain and two C-terminal β-barrel domains, which are thought to be involved in the regulation of enzyme activity and specificity.

Seven different transglutaminase genes have been characterised in higher vertebrates on the basis of their primary structure (Aeschlimann, D, and Paulsson, M (1994) Thromb. Haemostasis **71**: 402-415 Aeschlimann *et al*: (1998) *J. Biol Chem* **273,** 3542). Transglutaminases can be found throughout the body, but each transglutaminase is characterised by its own typical tissue distribution, although each may be present in a number of different tissue types often in combination with other transglutaminases. Transglutaminase gene products have specific functions in the cross linking of particular proteins or tissue structures. For review see Aeschlimann and Paulsson (1994) (*supra*) and Aeschlimann and Tholmazy (2000) *Connective Tissue Res.* **41**, 1-27. For example, factor XIIIa stabilises the fibrin clot in haemostasis, whereas prostate transglutaminase (TG_P)[1] is involved in semen coagulation. Other transglutaminases have adopted additional functions such as the tissue transglutaminase (TG_C), which is involved in GTP-binding in receptor signalling, and band 4.2 protein which functions as a structural component of the cytoskeleton. Four transglutaminases have been shown to be expressed during the different stages of epidermal growth and differentiation. Three of these, keratinocyte transglutaminase (TG_K), epidermal transglutaminase (TG_E) and TG_X, are associated with keratinocyte terminal differentiation and the cross-linking of structural proteins to form the

cornified envelope. The fourth enzyme TG_C, is expressed in skin primarily in the basal cell layer, and plays a role in the stabilisation of the dermo-epidermal junction. The importance of proper cross-linking of the cornified envelope is exemplified by the pathology seen in patients suffering from a severe form of the skin disease referred to as congenital ichthyosis, which has been linked to mutations in the gene encoding TG_K.

All transglutaminase enzymes appear to be encoded by a family of closely related genes. Alignment of these genes demonstrates that all members of the transglutaminase family exhibit a similar gene organisation, with remarkable conservation of intron distribution. Furthermore, phylogenetic analysis indicates that an early gene duplication event subsequently gave rise to two different transglutaminase lineages; one comprising TG_C, TG_E, and band 4.2 protein; the other, factor XIIIa, TG_K and possibly also TG_P (Aeschlimann and Paulsson (1994) (*supra*)). The genes encoding TG_K and factor XIIIa have been mapped to human chromosome 14q11.2 and chromosome 6p24-25 respectively, whereas TG_C and TG_E have been mapped to chromosome 20q11, and TG_P has been mapped to chromosome 3p21-22.

Comparison of the structure of the individual transglutaminase genes shows that they may be divided into two subclasses, wherein the genes encoding TG_C, TG_E, TG_P and band 4.2 protein comprise 13 exons and 12 introns, and the genes encoding factor XIIIa and TG_K contain two extra exons. Exon IX of the former group is separated into two exons (X and XI) in TG_K and factor XIIIa, and the amino-terminal extensions of TG_K and factor XIIIa comprise an additional exon. However, except for the acquisition of an additional intron and the recruitment of an exon by the genes encoding factor XIIIa and TG_K, the gene structure is remarkably conserved among all members of the transglutaminase gene family. Not only is the position of intron splice points highly conserved, but also the intron splice types. This similarity in gene structure and homology of the primary structure of the transglutaminases provides further support for the proposition that the different transglutaminase genes are derived from a common ancestral gene.

The inventors have previously isolated a cDNA encoding a novel member of the transglutaminase gene family TG_X, from human foreskin keratinocytes (Aeschilmann *et al* (1998) J. Biol. Chem., **273,** 3452-3460). Two related transcripts with an apparent size of 2.2 and 2.8 kb were obtained. The deduced amino acid sequence for the full-length gene product encodes a protein with 720 amino acids and a molecular mass of 81kDa. A sequence comparison of TG_X to the other members of the transglutaminase gene family revealed that the domain structure and the residues required for enzymatic activity and Ca^{2+} binding are conserved and show an overall sequence identity of about 35%, with the highest similarity being found within the enzyme's catalytic domain.

The inventors subsequently determined that TG_X is the product of a ~35kb gene located on chromosome 15, comprising 13 exons and 12 introns. The intron splice sites were found to conform to the consensus for splice junctions in eukaryotes. The transcription initiation site is localised to a point 159 nucleotides upstream of the initiator methionine and the likely polyadenylation site is localised ~600 nucleotides downstream of the stop codon. The two mRNA isoforms are the result of alternative splicing of exon III and give rise to 2 protein variants of TG_X which comprise catalytic activity. TG_X is expressed predominately in epithelial cells, and most prominently during foetal development, in epidermis and in the female reproductive system.

The inventors have now localised the TGM5 gene to chromosome 15q15 by fluorescent *in situ* hybridisation. Band 4.2 protein has previously been mapped to this chromosomal region (Sung L. A. *et al* (1992) Blood **79**: 2763-2770; Najfeld V. *et al* (1992) Am. J. Hum. Genet **50**: 71-75) and has subsequently been assigned to position 15q15.2 by expression mapping of the LGMD2A locus on chromosome 15 (Chiannikulchai N. *et al* (1995) Hum. Mol. Genet **4**: 717-725). A short sequence encompassing the left arm of one of the YAC clones ($926G10^2$) used for expression mapping matched with the sequence of intron 12 of the TGM5 gene placing the genes encoding TG_X and band 4.2 protein in close proximity on chromosome 15 (Fig. 5C). PCR with specific primers for 5' (exon I) or 3' (exon XIII) sequences of band 4.2 protein as well as southern blot analysis revealed that the BAC clones containing the TGM5 gene also contained the EPB42 gene and that the 2 genes are arranged in tandem.

Further analysis by the inventors has recently led to the identification of two novel transglutaminase genes TGM7 and TGM6 which encode the proteins TG_Z and TG_Y respectively. Alternative mRNA sequences of the TGM7 gene are given in Fig. 6A and Fig. 6B. The TGM7 derived mRNA (Fig. 6A and Fig. 6B) comprises an open reading frame of 2130 nucleotides and a polyadenylation signal (AATAAA) 158 nucleotides downstream of the termination codon (TGA). The deduced protein for TG_Z consists of 710 amino acids. The deduced protein for TG_z from Fig. 6A has a molecular mass of 79, 908 Da and an isoelectric point of 6.7. The deduced protein for TG_z from Fig. 6B has a molecular weight of 80,065 and an isoelectric point of 6.6.

The TGM6 full length transcript (Fig. 10A) comprises an open reading frame of 2109 nucleotides. The deduced protein for the long form of TG_Y consists of 708 amino acids and has a calculated molecular mass of 79, 466 Da and an isoelectric point of 6.9. The transcript for the short form of TG_Y (Fig. 10B) comprises an open reading frame of 1878 nucleotides and the deduced protein consists of 626 amino acids with a molecular mass of 70, 617 Da and an isoelectric point of 7.6.

To analyse the relationship between the different transglutaminase genes, the inventors calculated their amino acid similarity based upon sequence alignments, and calculated their evolutionary distances using different algorithms. All the algorithms used predicted a close relationship between TG_X, TG_Z, TG_Y, TG_E, band 4.2 protein and TG_C, and factor XIIIa and TG_K, respectively. The grouping of TG_X, TG_Z, TG_Y, TG_E, TG_C, and band 4.2 protein in one subclass and factor XIIIa and TG_K in another is supported by the results of this analysis and by the gene structure and genomic organisation of the different transglutaminase genes.

The inventors have therefore determined the structure of the human TGM5 gene, and its flanking sequences, and have mapped the gene to the 15q15 region of chromosome 15. Further, the inventors have determined that the human TGM5 gene comprises 13 exons separated by 12 introns spanning roughly 35kb, and that the structure of the TGM5 gene is identical to that of EPB42 (band 4.2 protein), TGM2 (TG_C) and TGM3 (TG_E) genes. Southern blot analysis has also shown that TGM5 is a single copy gene in the haploid genome. The inventors developed a method for detection and identification of

transglutaminase gene products based on RT-PCR with degenerate primers and using this method have discovered the gene product of the TGM5 gene in keratinocytes (Aeschlimann *et al* (1998) J. Biol. Chem. **273**, 3452-3460). Using this method, the inventors have identified another new transglutaminase gene product in human foreskin keratinocytes and in prostate cacrinoma tissue which has been designated TG_z or transglustaminase type VII. A full-length cDNA for this gene product was obtained by anchored PCR. Long range genomic PCR was used comprising different combinations of primers designed from the flanking sequences of the TGM5 - EPB42 gene sequence and the TG_z cDNA sequence to explore whether the gene encoding TG_z (TGM7) was present in close proximity to the other two transglutaminase genes. This placed the TGM7 gene approximately 9kb upstream of the TGM5 gene and demonstrated that the genes are arranged in tandem fashion (Fig. 5C). The inventors have therefore determined that the transglutaminase genes, TGM5 (TG_X), TGM7 (TG_Z) and EPB42 (band 4.2 protein) are positioned side by side within approximately 100 kb on chromosome 15. It has also been found that the mouse homologues of these genes are similarly arranged on mouse chromosome 2. Finally, the inventors have identified and determined the nucleotide and amino acid sequences as well as tissue distribution for the novel transglutaminase gene products TG_Z and TG_Y.

According to a first aspect of the invention there is provided a nucleotide sequence comprising at least a portion of the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B; a nucleotide sequence which hybridise to the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B; a nucleotide sequence which is degenerate to the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B; all of which nucleotide sequences encode a polypeptide having transglutaminase activity.

Preferably the nucleotide sequence consists of the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B.

The first aspect of the present invention also provides a nucleotide sequence which hybridises under stringent conditions to the nucleotide sequences of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B and which encodes a polypeptide having transglutaminase activity. Preferably the nucleotide sequence has at least 80%, more preferably 90% sequence homology to the nucleotide sequence shown in Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B. Homology is preferably measured using the BLAST program.

The invention further provides a method of expressing a polypeptide comprising inserting a nucleotide sequence according to the first aspect of the present invention into a suitable host and expressing that nucleotide sequence in order to express a polypeptide having transglutaminase activity.

The invention also provides a vector comprising a nucleotide sequence according to the first aspect of the present invention.

According to another aspect of the invention there is provided a polypeptide having an amino acid sequence comprising at least a portion of the amino acid sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B, wherein the polypeptide has transglutaminase activity.

The invention also provides a polypeptide sequence which is at least 90% identical to the amino acid sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B and which has transglutaminase activity. The amino acid sequence of the polypeptide having transglutaminase activity may differ from the amino acid sequence given in Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B by having the addition, deletion or substitution of some of the amino acid residues. Preferably the polypeptide of the present invention only differs by about 1 to 20, more preferably 1 to 10 amino acid residues from the amino acid sequence given in Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B.

The invention also provides a composition comprising the polypeptide of the present invention for use in transamidation reactions on peptides and polypeptides.

The invention also provides a polypeptide comprising exons VII through to exon X of the sequence shown in Fig. 6A or Fig. 6B. The position of the exons on the sequence shown in Fig. 6A or Fig. 6B can be determined from Fig. 8 where intron splice sites are marked with arrow heads.

According to a further aspect of the invention, there is provided a polypeptide comprising exons II through to exon IV or exons X through to exon XII of the sequence show in Fig. 10A or Fig. 10B. As indicated above, the positions of the exons on the sequence shown in Fig. 10A or Fig. 10B can be determined from Fig. 8.

According to another aspect of the invention there is provided a composition comprising the polypeptide according to the present invention for use in the cross-linking of proteins.

According to a further aspect of the invention there is provided a diagnostic method comprising detecting expression of the polypeptide according to the present invention in a subject or in cells derived from a subject.

The invention also provides an antibody directed against the polypeptide according to the present invention. The antibody may be any antibody molecule capable of specifically binding the polypeptide including polyclonal or monoclonal antibodies or antigen binding fragments such as Fv, Fab, $F(ab')_2$ fragments and single chain Fv fragments.

The invention further provides a method of gene therapy comprising correcting mutations in a non wild type nucleotide sequence corresponding to the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig 10B. Such gene therapy methods can be performed by homologous recombination techniques or by using ribozymes to correct small sequence mutation. Suitable techniques are well known to those skilled in the art.

In accordance with a further aspect of the invention there is provided a method of diagnosis of autoimmune disease comprising taking a sample from a subject and testing that sample for the presence of a transglutaminase encoded by the nucleotide sequence of Fig. 6A, Fig. 6B, Fig. 10A or Fig 10B or portions thereof. Preferably the transglutaminase is detected by using an antibody having affinity for the transglutaminase.

The invention also provides a competitive protein binding assay for the differential diagnosis of autoimmune diseases comprising the detection of antibodies against the transglutaminase encoded by the nucleotide sequence of Fig. 6A, Fig. 6B or Fig.10A or Fig 10B, or portions thereof.

Preferably the protein binding assay comprises using exogenous transglutaminase TGz or TGy, or both, as a competitive antigen.

The invention will now be described with reference to the accompanying Figures 1 to 11, in which:

Fig. 1 is a representation of the genomic organisation of the human TGM5 gene. The human TGM5 gene is represented with the exons numbered I to XIII indicated by solid boxes separated by the introns 1 to 12. The sizes of the introns and exons are given in bp (base pairs). The 5'- and 3'- untranslated regions in exon 1 and XIII, respectively, are represented by hatched boxes with functional elements defining the transcript indicated. Additional sequence elements found in the TGM5 gene are indicated as follows: *Alu*, Alu 7SL derived retroposon; STS, sequence tagged site. Below the genomic map, a representation of the sequences present in the individual BAC clones is depicted.

Fig. 2 is a representation of the structure of the 5' untranslated region of the human TGM5 gene and mapping of the transcriptional start site. **A.** Primer extension analysis of poly (A^+) RNA isolated from primary human keratinocyte prior to (lane 1) or after (lane 2) culture in suspension for 12h. Extension products were separated on denaturing polyacrylamide gel alongside a Sanger dideoxynucleotide sequencing reaction of the appropriate genomic DNA fragment primed with the same oligonucleotide. The transcriptional start site is

indicated by the arrow. **B.** Nucleotide sequence of the proximal 5' region of the TGM5 gene, 5' ends of mRNA from primary keratinocytes mapped by RACE are indicated by arrowheads. The major transcription start site identified by primer extension is highlighted with an *asterisk* (labelled +1). Consensus sequences for putative regulatory elements are underlined.

Fig. 3 is a representation of the structure of the 3' untranslated region of the human TGM5 gene. 3'-flanking sequence is shown with sequences homologous to known consensus sequences for 3' processing of transcripts (AATAAA, CAYTG and YGTGTTYY) underlined. The termination points of cDNA's isolated from human keratinocytes (Aeschlimann *et al* (1998) J. Biol Chem **273,** 3453-3460) by 3' RACE are indicated by arrowheads. A pair of inverted long repeat sequences is highlighted in italics.

Fig. 4 is a southern blot analysis of human genomic DNA hybridised to genomic TG_X probes. Human genomic DNA was digested with *Bam*HI, *Eco*RI, and *Hind*III restriction enzymes and hybridised with short ^{32}P-labelled DNA fragments corresponding to intron 2 and flanking sequences of exon II and III (left panel) and exon X (right panel), respectively. The migration positions of the *Hind*III DNA size markers is indicated on the left.

Fig. 5 shows the chromosomal localisation of the human TGM5 gene by fluorescence in situ hybridisation. **A.** representative picture of fluorescine-labelled genomic DNA of BAC-228(P20) (fluroescence, arrows) hybridised to metaphase spreads of human chromosome stained with propidium iodide. **B.** An ideogram of banded chromosome 15 showing the localisation of the fluorescent signal on 13 chromosomes. **C.** Is a schematic map of the respective locus showing the organisation of the genes encoding TG_x (TGM5), band 4.2 protein (EPB4.2) and TG_z (TGM7) as well as other genes [mitochondrial ATPase subunit D pseudogene; D-type cyclin-interacting protein 1 (DIP1); EST Genbank AA457639, AA457640)], L1 repetitive element and genetic markers.

Fig. 6 shows A. the nucleotide sequence and deduced amino acid sequence for human TG_Z. B. an alternative nucleotide sequence and deduced amino acid sequence for human TG_Z.

The initiation and termination codons as well as polyadenylation signal (AATAAA) are underlined.

Fig.7 is a representation of the different tissue expression patterns for TG_X, band 4.2 protein TG_Y and TG_Z in different fetal and mature human tissues. Human tissue Northern dot blot normalised for average expression of 9 different housekeeping genes probed with a fragment corresponding to the C-terminal β-barrel domains of TG_X (**A**), TG_Z (**B**), (**C**) TG_Y and band 4.2 (**D**). A diagram showing the type of poly $(A)^+$ RNA dotted onto the membrane is shown in panel E.

Fig. 8 is a comparison of the structure of the different human transglutaminase genes. A. is an alignment of the nine characterised human gene products (TG_X, TG_Y, TG_Z (shown in Fig. 6A), TG_C, TG_E, band 4.2, factor XIII a-subunit, TG_K, TG_P,) is shown, with dashes indicating gaps inserted for optimal sequence alignment and underlined residues representing amino acids conserved in at least five gene products. The sequences are arranged to reflect the transglutaminase domain structure, based on the crystal structure of factor XIII a-subunit. N-terminal propeptide domain (**d1**), β-sandwich domain (**d2**), catalytic core domain (**d3**) and β-barrel domains 1 (**d4**) and 2 (**d5**) (from top to bottom). Known intron splice sites are marked by ˥. B. is an alignment of the nine characterised human gene products (TG_X, TG_Y, TG_Z (shown in Fig. 6B), TG_C, TG_E, band 4.2, factor XIII a-subunit, TG_K, TG_P,) is shown, with dashes indicating gaps inserted for optimal sequence alignment and underlined residues representing amino acids conserved in at least five gene products. The sequences are arranged to reflect the transglutaminase domain structure, based on the crystal structure of factor XIII a-subunit. N-terminal propeptide domain (**d1**), β-sandwich domain (**d2**), catalytic core domain (**d3**) and β-barrel domains 1 (**d4**) and 2 (**d5**) (from top to bottom). Known intron splice sites are marked by arrowheads.

Fig. 9 is a phylogenetic tree of the transglutaminase gene family and genomic organisation of the genes in man and in mouse. Sequences were aligned to maximise homology as shown in Fig. 8 except including sequences from different species as available: h, human; m, mouse; r, rat. Note, the mouse sequence for TG_x^3 is at present incomplete and no information is available for the N-terminal domain. In panel A5, a hypothetical pedigree for the gene family is given that is consistent with the data on the sequence relationship of the

individual gene products (**A**) as well as with the data on the gene structure and genomic organisation (**B**). Phylogenetic trees based on the amino acid sequence homology of the gene products have been constructed using the NJ method (Saitou and Nei, 1987) of the PHYLIP software package for (1) the N-terminal β-sandwich domain (2) the catalytic core domain, (3) the C-terminal β-barrel domains, and (4) the entire gene products, (**C**). Shows the similarity of TG_X to the other transglutaminase gene products. The domain structure is based on the X-ray crystallographic structure of the factor XIII a-subunit dimer and inferred on the other gene products based upon the sequence alignment shown in Fig. 8. The numbers reflect % sequence identity.

Fig. 10 shows the nucleotide sequence and deduced amino acid sequence of TG_Y. **A.** Shows the nucleotide and deduced amino acid sequence for the long form of TG_Y. **B.** Shows the nucleotide sequence and deduced amino acid sequence for the short form of TG_Y.

Fig. 11 is a schematic representation of the organisation of the identified transglutaminase gene clusters in the human genome.

Isolation and Determination of the Structure of the Human TGM5 Gene.

A unique insertion sequence of about 30 amino acids between the catalytic core domain and β-barrel domain 1 found in TG_x was used as a template to design specific primers for the screening of a human genomic library. The characterisation of several genes of the transglutaminase gene family showed that the positions of the introns has been highly conserved and a comparison of the TG_X sequence to the sequences of the other transglutaminases indicated that this unique sequence is present within an exon, exon X (see Fig. 8, aa 460-503) in TG_X. A PCR reaction from human genomic DNA using oligonucleotides P1 and P2 that match sequences at either end of this unique segment yielded a DNA fragment of expected size which was confirmed to be the correct product by sequencing (results not shown). Screening of a human genomic DNA BAC library by PCR using these oglionucleotides revealed two positive clones, BAC-33(P5) and BAC-228(P20) that were subsequently shown by Southern blotting with different cDNA probes to contain sequences spanning at least exon II to exon X of the TGM5 gene (results not shown).

Restriction analysis further indicated that each of the BAC clones contained substantially more than 50kb of human genomic DNA.

The similarity in the gene structure of the different transglutaminase genes prompted us to approach the characterisation of introns by PCR amplification using oligonucleotide primers corresponding to the flanking exon sequences at the presumptive exon/intron boundaries. All intron/exon boundaries were sequenced from the PCR products obtained in at least two independent PCR reactions, where applicable from both BAC clones, to exclude mutations introduced by *Taq* DNA polymerase, and the results compared. When sequences of PCR products comprising adjacent introns had no overlap, the intervening sequence (exon sequence) was determined by direct sequencing from isolated BAC plasmid DNA to confirm the absence of additional introns. Similarly, the 3'-untranslated region was obtained by step-wise extension of the known sequence using direct sequencing of BAC plasmid DNA. Both BAC clones terminated short of exon I and all attempts at isolating clones spanning exon I by screening of BAC and P1 libraries failed. Exon I and intron 1 sequences were finally derived by nested PCR from human genomic DNA using conditions optimised for long range genomic PCR.

We established that the TGM5 gene comprises approximately 35kb of genomic DNA and contains 13 exons and 12 introns (Fig. 1). All intron/exon splice sites conform to the known GT/AG donor/acceptor site rule and essentially to the consensus sequence proposed by Mount S.M (1982) Nucleic Acids Res. **10**, 459-472. (Table I). A sequence homologous to the branch point consensus CTGAC (Keller E. B and Noon. W. A (1984) *Proc. Natl. Acad. Sci. USA* **81**: 7417-7420) was found 24 to 44 nucleotides upstream of the 3' splice site in introns 1, 3-6, and 9-12. The size of the introns varied considerably, ranging from 106bp to more than 6kb (Fig. 1, Table II). The sequence obtained from the two different BAC clones matched with the exception of a deletion spanning the sequence from intron 6 to intron 8 in BAC-33(P5) (Fig. 1).

Further, we also resequenced the entire coding sequence of TG_X and found 3 point mutations as compared to the previously reported cDNA sequence (Aeschlimann. *et al* (1998) J. Biol Chem **273** 3452-3460) One of the nucleotide exchanges is silent, the other

two result in an amino acid exchange (Table III). The first two mutations were found in both BAC clones, the third was only present in BAC-228(P20) due to the deletion in the other BAC clone. These differences may result from sequence polymorphisms in the human gene pool as there was no ambiguity of the cDNA-derived sequence in this position determined from multiple independently amplified PCR products. However, the fact that a serine and alanine residue are changed into a proline and glycine residue that constitute the conserved amino acid in these positions in the transglutaminase protein family (see Fig. 8, aa 67 and aa 352 in TG_X) suggested that these may have been PCR-related mutations in the cDNA sequence. To clarify this issue, we have prepared cDNA from human foreskin keratinocytes from different individuals, amplified full-length cDNA with high fidelity DNA polymerase, and sequenced the respective portions of the cloned cDNAs. The data confirmed that allelic variants exist with differences in these positions (Table III).

The isolation and sequencing of cDNAs encoding TG_X and Northern blotting with TG_X cDNA probes revealed expression of at least two differentially spliced mRNA transcripts for TG_X in human keratinocytes. Solving the gene structure confirmed the short form of TG_X to be the result of alternative splicing of exon III as predicted. A third isolated cDNA that differed also at the exon III/exon IV splice junction turned out to be the result of incomplete or absent splicing out of intron 3 as the sequence upstream of exon IV in the cDNA matched with the 3' sequence of intron 3. Exon 3 encodes part of the N-terminal β-barrel domain of TG_X and the absence of the sequence encoded by exon 3 is expected to result in major structural changes in at least this domain of the protein. Nevertheless, expression of TG_x in 293 cells using the full-length cDNA resulted in synthesis of two polypeptides with a molecular weight consistent with the predicted products from the alternatively spliced transcripts (results not shown).

Initially, 5' RACE was used to determine the 5' end of TG_X cDNAs. Transcripts starting 77,96 and 157 nucleotides upstream of the initiator ATG were isolated in addition to the previously described shorter transcript (Fig. 3B, arrowheads). All of these transcripts were recovered repeatedly in independent experiments. Finally, primer extension experiments located the major transcription initiation site used in keratinocytes 157 nucleotides upstream of the translation start codon (Fig. 3). The proximal promoter region was

analysed for potential binding sites of transcription factors using MatInspector (Genomatix, Munich Germany) and GCG (Genetics Computer Group, Inc., Madison, Wisconsin) software packages. No classical TATA-box sequence was found but a number of other potential transcription factor binding sites could be identified (Fig. 3b), suggesting that TG_X promoter is a TATA-less promoter. Interaction of C/BP (may bind to CAAT-box), nuclear factor I (NF1) and upstream stimulatory factor (USF) to form a core proximal promoter has been demonstrated in a number of TATA-less genes. c-Myb is found in TATA-less proximal promoters of genes involved in hematopoiesis and often interacts with Ets-factors, and these sites may be operative in the expression of TG_X in hematopoetic cells, e.g HEL cells AP1, Ets and SP1 elements are typically found in keratinocyte-specific genes and may be involved in transcriptional regulation in keratinocytes. Several AP1 sites are present within 2.5kb of the upstream sequence and could interact with the proximal AP1 factor for activation. SP1 sites are properly positioned upstream of the start points of the shorter transcripts raising the possibility that these could also be functional, though to a lesser degree.

The last exon, exon XIII, contained a consensus polyadenylation signal AATAAA-600bp downstream of the termination codon (Fig. 3). This is in good agreement with the size of the mRNA (2.8kb) encoding full-length TG_X expressed in human keratinocytes as detected by Northern blotting considering the length of the coding sequence (2160bp.). A CAYTG signal that binds to U4 snRNA which is identical for 4 out of 5 nucleotides is present in tandem in 3 copies 7 nucleotides downstream of the polyadenylation signal. A close match (YCTGTTYY) of another consensus sequence YGTGTTYY that is found in many eukaryotic transcripts and provides a signal for efficient 3' processing is present 46 nucleotides downstream of the polyadenylation signal. However, we have previously reported that all cDNAs isolated by RT-PCR with an oligo(dT) oligonucleotide from human keratinocytes ended within 9 to 34 nucleotides downstream of the pentanucleotide ATAAA at position 2169. It has been shown that this pentanucleoditde functions as a polyadenylation signal and these shorter transcripts are selectively enhanced by PCR amplification because of the smaller size of the PCR product.

Chromosomal Localisation of the TGM5 Gene.

To address the genomic organisation and identify the chromosomal localisation of the TGM5 gene or genes in the human genome, we performed Southern blot analysis of human genomic DNA cut with *Bam*HI, *Eco*RI and *Hind*III restriction enzymes using probes derived from intron 2 as well as from the sequence encoded by exon X that is unique to TG_X (Fig.4). Bands of 4.5, 6.0,10.5, 4.3, 9.3 and 2.6kb were revealed with the respective probes. The simple pattern of restriction fragments hybridising with the probes indicated that the haploid human genome contains only one TGM5 gene.

The TGM5 gene was subsequently localised to chromosome 15 by fluorescent *in situ* hybridisation of human metaphase chromosome spreads using genomic DNA derived from either BAC clone as a probe (Fig. 5A). A comparison of the probe signal to the DAPI banding pattern localised the TGM5 gene to the 15q15 region. The localisation was subsequently refined by determining the distance of the fluorescent signal to the centromere as well as to either end of the chromosome on 13 copies of chromosome 15 and expressing it as a fractional distance of the total length of the chromosome. These measurements placed the TGM5 gene close to the centre of the 15q15 region, *i.e.* to the 15q15.2 locus (Fig. 5B).

TGM5 is part of a cluster of transglutaminase genes.

The EPB42 gene has previously been assigned to locus 15q15.2 on chromosome 15. This raised the possibility that the EPB42 gene may be arranged in tandem with the TGM5 gene. Indeed, PCR with specific primers for sequences derived from the 5' and 3' of the EPB42 gene yielded products of appropriate size from both, BAC-33(P5) and BAC-228(P20) (results not shown) and sequencing confirmed the identity of the PCR products. Southern blotting of BAC plasmid DNA with cDNA probes comprising the 5' or 3' end of the EPB42 gene and subsequent comparison of the pattern of labelled restriction fragments with that of the TGM5 gene allowed us to map this locus in more detail. The EPB42 gene and TGM5 are arranged in the same orientation being spaced apart by ~11kb (Fig. 5C) approximately 30% of which was sequenced to characterise the 3' and 5' flanking UTR of the TGM5 and EPB42 gene, respectively.

To analyse the relationship between the most closely homologous genes of the transglutaminase gene family (TGM7, TGM5 and EPB42 on human chromosome 15q15.2 and TGM2 and TGM3 on chromosome 20q11/12), we mapped the respective mouse genes using radiation hybrid mapping. All genes mapped to the distal part of mouse chromosome 2. The genes for *tgm7*, *tgm5*, and epb42 showed a best fit location for the segment defined by D2Mit104 proximal and D2Mit305 (66.9cM) and an LOD of >20 to D2Ertd616e (69.0cM). This is in good agreement with the assigned locus 67.5cM distal from the centromere, White, R.A., *et al.*, (1992) *Nat. Genet.* **2**, 80-83. This *tgm3* gene showed a best fit location for the segment defined by D2Mit447 proximal and D2Mit258 distal, with a highest LOD of 14.8 and 12.2 to D2Mit258 (78.0cM) and D2Mit338 (73.9cM), respectively. The *tgm2* gene showed a best fit location for the segment defined by D2Mit139 proximal (86.0cM) and D2Mit225 distal (91.0cM), with a highest LOD of 17.0 to the anchor market D2Mit287, consistent with it's assigned locus 89.0cM from the centromere, Nanda, N., *et al.*, (1999). *Arch. Biochem. Biophys.* **366**, 151-156.

We developed a method for detection and identification of transglutaminase gene products based on RT-PCR with degenerate primers and using this method discovered the gene product of the TGM5 gene in keratinocytes. Using this same method, we have identified another new transglutaminase gene product in human foreskin keratinocytes and in prostate carcinoma tissue which we designated TG_z or transglutaminase type VII. A full-length cDNA for this gene product was obtained by anchored PCR (see below). We used long range genomic PCR with different combinations of primers designed from the flanking sequences of the TGM5 - EPB42 gene segment and the TG_z cDNA sequence to explore whether the gene encoding TG_z, TGM7, was present in close proximity to the other two transglutaminase genes. This placed the TGM7 gene approximately 9kb upstream of the TGM5 gene and demonstrated that the genes are arranged in tandem fashion (Fig. 5C). The 5' UTR of the TGM5 gene was sequenced (Fig. 2). The genes encoding TG_C, TG_E, which are more closely related to the TGM5, TGM7, and EPB42 genes than the other transglutaminase genes based on amino acid sequence comparison and similarity in gene structure (Fig. 9) have been mapped to human chromosome 20qll (Gentile V. *et al* (1994) Genomics **20**, 295-297; Kim I. G. *et al* (1994) J. Invest. Dermatol **103**, 137-142). The syntenic regions of the 15q15 and 20q11 locus in the mouse are present in a short segment,

2F1-G, of chromosome 2, which puts all five transglutaminase genes in proximity (Fig. 9B). The mouse homologue of band 4.2 protein has been mapped to this region of mouse chromosome 2 (White R. A. *et al* (1992) Nat. Genet **2**, 80-83). We have isolated a BAC clone containing the gene encoding the mouse homologue of TG_x and shown that the tgm5 gene is located next to the epb42 gene in tandem fashion, similar to the organisation in the human genome. Furthermore, this clone was shown to contain the gene encoding the mouse homologue of TG_z. Genomic sequences derived from this BAC clone and also cDNA sequences derived from cDNA prepared from mouse uterus showed that the mouse and human gene products are 85% identical on the nucleotide level. To analyse the relationship between the transglutaminase genes in more detail, we calculated the amino acid similarity (Fig. 9C) based on the sequence alignment shown in Fig. 8 and calculated evolutionary distances using different algorithms (Fig. 9A). All algorithms predicted a close relationship between TG_X and TG_E, and band 4.2 protein and TG_C, raising the possibility that a single transglutaminase gene initially locally duplicated to generate a cluster of 3 genes, followed by a duplication of a larger segment of the chromosomal region, gave rise to the organisation of the genes in mouse. In humans these chromosomal regions were apparently redistributed to two different chromosomes. This hypothesis led us to spectulate on the existence of an additional gene on human chromosome 20q11. Careful analysis of the chromosomal sequences of this locus derived by the human genome project revealed the presence of a candidate gene, TGM6, located approximately 45kb downstream of the TGM3 gene consistent with our hypothesis (Fig. 11). To confirm that this is in fact a functional gene and not a pseudogene, we screened a large number of cell lines for expression of a respective gene product by PCR. A corresponding gene product, TG_Y, or transglutaminase type VI could be identified in a small cell lung carcinoma cell line and a full-length cDNA was subsequently derived by anchored PCR.

Determination of cDNA and amino acid sequences of TGM6 (TG_Y) and TGM7 (TG_Z) gene products

A full-length cDNA sequence for TG_z was obtained by anchored PCR using oligo(dT)-*Not* I primed cDNA prepared from human foreskin keratinocytes, prostate carcinoma tissue and human carcinoma cell line PC3, essentially following the strategy previously described (Aeschlimann *et al* (1998) *J. Biol Chem* **273**: 3245-3460). The oligo(dT)-*Not* I primer was

used as the anchoring primer to obtain the 3' end of the cDNA. 5' RACE was used to determine the 5' end of the cDNA. The obtained sequence information (Fig. 6) contained an open reading frame 2130 nucleotides and a polyadenylation signal (AATAAA) 158 nucleotides downstream of the termination codon (TGA). The deduced protein consists of 710 amino acids. The cDNA and amino acid sequence in Fig. 6A was first determined and the deduced protein has a calculated molecular mass of 79,908 Da and an isoelectric point of 6.7. The cDNA and amino acid sequence in Fig. 6B was then determined. This sequence differs by a few nucleotides and amino acids from the sequence given in Fig. 6A. The protein deduced from the sequence given in Fig. 6B has a calculated molecular mass of 80,065 and an isoelectric point of 6.6. A number of aberrantly spliced gene products were isolated which lacked part of exon IX (5'end) or retained the whole or part of intron 11. These products are unlikely to be of physiological significance but may point out that splicing of certain introns in this gene is a difficult and inefficient process.

A full-length cDNA sequence for TG_Y was obtained by PCR using oligo(dT) primed cDNA prepared from the lung small cell carcinoma cell line H69, and using sequence specific primers based on the presumptive transcribed genomic sequence. 5' RACE was used to determine the 5' end of the cDNA. The obtained sequence information for the long form of TG_Y (Fig. 10A) contained an open reading frame of 2109 nucleotides. The deduced protein for the long form of TG_Y consists of 708 amino acids and has a calculated molecular mass of 79,466 Da and an isoelectric point of 6.9. A shorter transcript was also isolated which apparently resulted from alternative splicing of the sequence encoded by exon XII. The absence of exon XII results in a frame shift and thereby in premature termination within exon XIII. The obtained sequence information for the short form of TG_Y (Fig. 10B) contained an open reading frame of 1878 nucleotides. The deduced protein for the short form of TG_Y consists of 626 amino acids and has a calculated molecular mass of 70,671 Da and an isoelectric point of 7.6. The sequence alterations due to the splicing result in a short protein which terminates just after the first C-terminal β-barrel domain. The β-barrel domains have been implicated in the regulation of enzyme-substrate interaction, and the lack of the second C-terminal β-barrel domain (see Fig. 8, d5) is likely to be of biological significance.

The catalytic mechanism of transglutaminases has been solved based on biochemical data available for several transglutaminases and the X-ray crystallographic structure of the factor XIII a-subunit dimer. The reaction center is formed by the core domain and involves hydrogen-bonding of the active site Cys to a His and Asp residue to form a catalytic triad reminiscent of the Cys-His-Asn triad found in the papain family of cysteine proteases. The residues comprising the catalytic triad are conserved in TG_Y (Cys276, His335, Asp358) and TG_Z (Cys227, His336, Asp359) (Fig. 8) and the core domain shows a high level of conservation as indicated by a sequence identity of about 50% between these gene products and the other transglutaminases (Fig. 9). A Tyr residue in barrel 1 domain of the a subunit of factor XIII is hydrogen-bonded to the active site Cys residue and it has been suggested that the glutamine substrate attacks from the direction of this bond to initiate the reaction based on analogy to the cysteine proteases. In TG_Y, the Tyr residue is conserved (Tyr 540) while in TG_Z the Tyr residue has been replaced by His538 similar to TG_X (Fig. 8). This is expected to be a conservative change which is supported by our data demonstrating that recombinant TG_X from 293 cells has transglutaminase activity. Crystallization experiments with factor XIIIa further indicated that 4 residues are involved in binding of Ca^{2+}-ion, including the main chain carbonyl of Ala457 and the side chain carboxyl groups of Asp438, Glu485, and Glu490. All three acidic residues are conserved in TG_Y and in TG_Z (Fig. 8). None of the residues critical to enzyme function are affected by the alternative splicing of TG_Y. Based on the preservation of critical residues for enzyme function and domain folding and the extensive overall similarity of the TG_Y isoforms and TG_Z to the other members of the transglutaminase protein family, it can be predicted that the characterized cDNAs are encoding active transglutaminases.

Tissue Expression Patterns for TG_X, TG_Y and TG_Z.

We have previously shown that TG_X is expressed in a number of different cell types (Aeschlimann *et al* (1998) J. Biol. Chem. **273**, 3452-3460).. To obtain a more complete picture on the expression of TG_X and the novel gene products, we performed a dot blot Northern blot analysis of more than 50 adult and fetal human tissues. Band 4.2 protein was expressed at high level in bone marrow and fetal spleen and liver, consistent with its role in hematopoietic cells, and virtually undetectable in all other tissues. In contrast, TG_X, TG_Y and TG_Z showed widespread expression at low level, with highest levels of TG_X, TG_Y and

TG_Z mRNA present in the female reproductive system, in the central nervous system, and in testis, respectively (Fig. 7).

RT-PCR analysis on human cell lines and tissues shows that TG_Z is expressed in osteosarcoma cells (MG-63), dermal fibroblasts (TJ6F, HCA2), erythroleukemia cells (HEL), in primary keratinocytes, mammary epithelium carcinoma cells (MCF7), HELA cells, skin, brain, heart, kidney, lung, pancreas, placenta, skeletal muscle, fetal liver, prostate and in prostate carcinoma tissue. A similar analysis for TG_Y revealed expression only in a lung small cell carcinoma cell line (H69) and extremely low levels of expression in tissues.

In conclusion, TG_Z is expressed widely in cells and tissues and expression levels are not apparently affected by cellular differentiation, (i.e keratinocyte differentiation or fibroblast senescence). TG_Y expression, on the other hand, was very restricted and expression was only found in H69 cell line. This cell line has characteristics of neuronal cells such as the expression of neuron-specific enolase and brain isozyme of creatine kinase which together with widespread expression in tissues of the nervous system suggests that TG_Y expression may be specific to neuronal cells. Transglutaminase action has been implicated in the formation of aberrant protein complexes in the central nervous system leading to nerve cell degeneration, e.g in Alzheimers and Huntington's disease. Based on its expression pattern, TG_Y is a logical candidate to bring about the underlying transglutaminase-related pathological changes.

Reagents

Oligonucleotides were from Oligos. Etc. Inc. (Wilsonville, OR) or life technologies and restriction enzymes from Promega Corp. (Madison, WI).

Genomic Library Screening

A human BAC library established in a F-factor-based vector, pBeloBAC 11, and maintained in *E. coli* DH10B was screened by PCR (Genome Systems, Inc., St. Louis, MO). A 147bp DNA fragment unique to TG_X was amplified from 100ng of genomic DNA in 100µl of 10mM Tris/HC1, pH 8.3, 50mM KC1 containing 2mM $MgC1_2$, 0.2mM dNTPs

using 2.5 units of *Tag* DNA polymerase (Fisher Scientific Corp. Pittsburgh, PA) and 50pmol of upstream primer P1,5'- CCACATGTTGCAGAAGCTGAAGGCTAGAAGC and downstream primer P2, 5'-CCACATGTCCACATCACTGGGTCGAAGGGAAGG. PCR cycles were 45 sec at 94°C (denaturation), 2 min at 60°C (annealing), and 3 min at 72°C (elongation) for a total of 37 cycles, with the first cycle containing an extended denaturation period (6 min) during which the polymerase was added (hot start), and the last cycle contained an extended elongation period (10 min). Two positive clones were identified, BAC-33(P5) and BAC-228(P20) (Genome Systems), and their identity verified by Southern blotting. Plasmid DNA was prepared using a standard alkaline lysis protocol. 2µg plasmid DNA was restricted with *Bam*HI, *Eco*RI, and *Spe*I and probed with a ^{32}P-labelled-500bp *Nco*I/B*spH*I and ~600bp B*spHI*/N*de*I cDNA fragment of TG_X, respectively, as described below.

Amplification of TGM5 Intron Sequences

PCRs were carried out with 2.5 units of *Taq* DNA polymerase (Fisher Scientific) and 100-200ng of plasmid DNA from BAC clones in 100µl of 10mM Tris/HC1, pH 8.3 50mM KC1 containing 2mM $MgC1_2$, 0.2mM dNTPs and 50pmol of the desired oligonucleotide primers. The PCR cycles were 45 sec at 94°C (denaturation), 1 min at 60°C (annealing), and 5 min at 72°C (elongation). A total of 32 cycles were carried out, with the first cycle comprising an extended denaturation period (6 min) during which the polymerase was added (hot start) and the last cycle comprised an extended elongation period (10 min). The following oligonucleotides were used as upstream and downstream primers, respectively, in the individual reactions:

intron 2, 5'-GGACCACCTGCTTGTTCGCCGGGG,5'-AGGGGCTGGGGCTGTGATGGCGTG;

intron 3,5'-ACCTCTTGAAAATCCACATCGACTCCT, 5'-CAGTTCTTGCTGCCTTGGTAGATGAAGCC;

intron 4,5'-GACAGTGAACCCCAGAGGCAGGAG, 5'-TCTGTGGCTGGGTCAGTCTGGAAGTGCA (P3);

intron 5,5'-GCCTGCACTTCCAGACTGACCCAGCCACA, 5'-TCCAGTTTCCATTGAGCACCCCA;

intron 6,5'-TGCTGGGTCTTTGCTGCCGTCATGTGC, 5'-TCCTTCTTCTTATTCCCCAAAATCCTGCC;

intron 7, 5'-TAGATGAGTATTATGACAACACAGGCAGG, 5'-GCGTCCAGCACCTGCCAGCCTCC;

intron 8,5'-TGAGTGCTGGATGGCCCGGAAGG, 5'-CCCGCTCGTCACTCTGGATGCTC;

intron 9, 5'-TTCACCAGGACACGAGTTCTGTTGGCA, P2 (see above);

intron 10, P1 (see above), 5'-TCAGGACTGCTTTTCTCTTCACCC;

intron 11, 5'-ACCCCTGCAAAATCTCCTATTCCC, 5'-AATATCACCTGTATGGAGAGTGGCTGG;

intron 12, 5'-TTGAGGACTGTGTGCTGACTGTGGM 5'-AATGATGCTTGCTTGGTGTTGGGG.

PCR's were carried out with 1.25 units of *Pfu* Turbo DNA polymerase (Stratagene) and 260ng genomic DNA in a total of 100µl of supplied reaction buffer supplemented with 0.2mM dNTPs, 2µl DMSO and 50 pmol primers. The PCR cycles were 45 sec at 94°C (denaturation), 1 min at 68°C (annealing), and 2 min at 72°C (elongation). A total of 37 cycles were made, with the first cycle containing an extended denaturation period (6 min) during which the polymerase was added (hot start), and the last cycle containing an extended elongation period (10 min).

Rapid Amplification of 5'-mRNA End

A modified RACE protocol was used to determine the transcription start site and obtain additional sequence information of exon I. Double stranded cDNA was prepared from poly(A^+) RNA of cultured normal human keratinocytes (Aeschilmann *et al* (1998) J. Biol Chem. **273**, 3452-3460) with the Copy Kit (Invitrogen, San Diego, CA). The cDNA was purified from nucleotides using the GlassMax DNA Isolation Kit (Life Technologies, Inc.) and tailed in the presence of 200µM dCTP with 10 units of terminal deoxynucleotidyl transferase (Promega) for 30 min at 37°C to anchor the PCR at the 5'-end. The PCR reaction was anchored by performing a total of 5 cycles of one-sided PCR at a lower annealing temperature (37°C) with the abridged anchor primer (Life Technologies, Inc.) only. Following transfer of 25% of this reaction at 94°C to a new tube containing abridged anchor primer and TG_X-specific primer P3 (see above), the first round of amplification was carried out for a total of 37 cycles under the conditions described above except for annealing which was carried out at 55°C. Nested PCR was done with the universal amplification primer (Life Technologies, Inc.) and TG_X-specific primer P4, 5'-TGAAGTACAGGGTGAGGTTGAAGG, as described above (annealing at 60°C) using 1.0 µl from the first round PCR.

Primer Extension Analysis

Oligonucleotide P5 5'-CATGGTAGCTGCCTCCGGTTCCTG containing a 5'-infrared label (IRD 800) was purchased from MWG Biotech (Ebesberg, Germany). Primer P5 (5.3pmol) was hybridised to 1µg of poly (A^+)RNA from primary keratinocytes

(Aeschlimann *et al* (1998) J. Biol. Chem. **273,** 3452-3460) and reverse transcription performed with 200 units of Superscript II RNAse H reverse transcriptase (Life Technologies) in a total of 20µl for 90 min at 42°C according to the manufacturer's instructions. Enzyme was heat inactivated and primer extension products extracted with phenol chloroform, precipitated with ethanol, and then analysed on a 4.5% denaturing polyacrylamide gel adjacent to dideoxynucleotide chain termination sequencing reactions (Thermo Sequenase Cycle Sequencing Kit; Amersham) derived from a double-stranded genomic DNA fragment using the same primer.

DNA Preparation and Sequencing

Plasmid DNA from BAC clones was further purified for direct sequencing by digestion with 200µg/ml of RNase A (Sigma, St. Louis, MO) for 1h at 37°C and by subsequent micro-dialysis using Spectra/Por 2 membranes (Spectrum Medical Industries, Inc. Laguana Hills, CA). PCR produts were gel purified using the QIA quick Gel Extraction Kit (Qiagen, Inc. Chatsworth, CA) for sequencing. Cycle sequencing was performed by the dideoxy chain termination method using the Cyclist Exo-P*fu* DNA Sequencing Kit (Stratagene, LaJolla, CA) and pre-cast 6% polyacrylamide gels with the CastAway Sequencing System (Stratagene) or using the dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (PE Biosystems) and an ABI 310 automated sequencer.

Southern Blotting

18µg human genomic DNA was digested with *Bam*HI, *Eco*RI, and *Hind*III restriction enzymes, separated in a 0.8% agarose gel and transferred to a Zeta-probe membrane (Bio-Rad, Labs. Hercules, CA). The gel was calibrated using the Lambda DNA/*Hind*III markers (Promega). ^{32}P-labelled probes were prepared by random prime labelling using the Multiprime DNA Labelling System (Amersham, Int. Amersham, UK) and PCR products corresponding to intron 2, intron 12, and exon X (see above) as DNA templates. Probes were hybridised to the blot overnight at 65°C in 500 mM NaH_2PO_4, pH 7.5, containing 1mM EDTA and 7% SDS. The membrane was washed at 65°C to a final stringency of 40mM NaH_2PO_4, pH 7.5, 1mM EDTA, and 1% SDS, and the result developed by exposure of the membrane to BioMax MR film (Eastman Kodak, Rochester, NY).

Chromosomal Localisation

Human peripheral blood lymphocytes were used to prepare metaphase chromosome spreads (Bebbington C. R. and Hentschel, C. C. G. (1987) in DNA cloning (Volume III) 184-188, IRL Press, Oxford UK). Cells were cultured in PB-Max Karyotyping medium (Gibco, BRL, Gaithersburg, MD) for 72h, and synchronised by culture in the presence of 10^{-7}M amethophterin (Fluka) for another 24h. Cells were released from the mitotic block by extensive washing and subsequent culture in the above medium containing 10^{-5}M thymidine for 5h. Cells were subsequently arrested in metaphase by addition of colcemid to a final concentration of 0.1μg/ml (Gibco BRL). Harvested cells were incubated in 0.075M KCP for 25 min at 37°C, fixed in methanol/acetic acid (3:1) solution, and chromosome spreads prepared by dropping the cells onto the glass slides. After air drying, chromosomes were treated with 100μg/ml of RNase A in 2x SSC for 1h at 37°C, denatured in 70% (v/v) formamide in 2x SSC for 3 min at 75°C, and dehydrated in a graded ethanol series. DNA probes were prepared by random prime labelling of plasmid DNA of BAC-33(P5) and BAC-228(P20) with fluorescine-conjugated dUTP using the Prime-It Fluor Fluorescence Labelling Kit (Stratagene). Probes were denatured at 75°C for 10 min in hybridisation buffer consisting of 50% formamide (v/v) and 10% dextran sulphate (w/v) in 4x SSC and prehybridised at 42°C for 20 min to 0.2 μg/ml human competitor DNA (Stratagene) to block repetitive DNA sequences. Probes were subsequently hybridised to the chromosome spreads at 37°C overnight, followed by washing to a final stringency of 0.1x SSC at 60°C. Spreads were mounted in phosphate-buffered glycerol containing 200 ng/ml propidium iodide to counterstain chromosomes. Slides were examined by epifluorescence microscopy using a 100x objective and images captured with a DC-330 CCD camera (DAGE-MTI, Inc. Michigan City, IN) using a LG-3 frame grabber board (Scion Corp. Frederick, MD) in a McIntosh 8500 workstation and a modified version of the NIH image 1.6 software (Scion Corp.). Images representing fluorescine-labelling and propidium iodide staining of the same field were superimposed using Adobe Photoshop 3.0 (Adobe Systems, Inc. Mountain View, CA) to map the gene to a chromosomal region.

Cloning of Novel Transglutaminase Gene Products by Anchored PCR

For cloning of TG_Y, poly(A)+RNA was prepared from about 10_6 H69 cells (American Type Culture Collection, Rockville, MD) by oligo(dT)-cellulose column chromatography using

the Micro-Fast Track Kit (Invitrogen, San Diego, CA) and recovered in 20µl 10mM Tris/HCl, pH 7.5. The poly(A)+RNA (5.0µl) was reverse transcribed into DNA in a total volume of 20µl using the cDNA Cycle Kit (Invitrogen) with 1.0µl oligo(dT) primer (0.2µg/µl). Overlapping fragments of TG_Y were amplified by PCR using oligonucleotides 5'-ATCAGAGTCACCAAGGTGGAC, 5'-AGAAACACATCGTCCTCTGCACACC (P6), 5'-CAGGCTTTCCTCTCACCGCAAACAC, 5'-CGTACTTGACTGGCTTGTACCTGCC, 5'-TCTACGTCACCAGGGTCATCAGTGC, 5'-GCCTGTTCACCGCCTTGCTGT, 5'-CATCACTGACCTCTACAAGTATCC, 5'-ACGGCGTGGGATTCATGCAGG, 5'-CATCCTCTATACCCGCAAGCC, and 5'-AGGTTGAGGCAGGATTAACTGAGGCCTC. PCRs were carried out with 1.25 units of AmpliTaq Gold DNA polymerase (PE Biosystems) and 2.0µl cDNA in a total of 50µl of supplied reaction buffer supplemented with 2mM $MgCl_2$, 0.2mM dNTPs and 25 pmol of the appropriate gene-specific primers. A total of 40 PCR cycles were made, with an elevated annealing temperature of 65°C for the initial 5 cycles and an annealing temperature of 60°C for the remaining cycles. The 5'-end of the cDNA was isolated by 5'-RACE as described above with the exception of using the gene-specific oligonucleotides P6, 5'-GATGTCTGGAACACAGCTTTGG, and 5'-TCACAGTCCAGGGCTCTGCTCAG. The PCR-products were either directly sequenced or when desired, cloned by taking advantage of the 3' A-overhangs generated by *Taq* DNA polymerase using the Original TA-Cloning Kit (Invitrogen).

For cloning of TG_Z, we used a series of degenerate and gene-specific oligonucleotides to isolate overlapping DNA fragments, essentially following our previously described strategy. TG_Z-specific oligonucleotide primers were 5'-CAACCTTGCGGCTTGAGTCTGTCG, 5'-CAGCAGCTCTGACGGCTTGGGTC (P7), 5'-ATCACCTTTGTGGCTGAGACCG, 5'-CAAGGGTTAAAAAGTAGGATGAAAGTTC, 5'-CACAGTGTGACTTACCCGCTG, 5'-CATACACCACGTCGTTCCGCTG, 5'-CTTAAAGAACCCGGCCAAAGACTG, 5'-CGATGGTCAAGTTCCTATCCAXGTTG, 5'-TGTTGTTTCCAATTTCCGTTCCGC, 5'-TCTGGCACCCTCTGGATACGCAG, 5'-CTTAGGGATCAGCCAGCGCAGC,

5'-GCGGATGAACCTGGACTTTGG, 5'-GGGTGACATGGACTCTCAGCG, 5'-TGGGCAAGGCGCTGAGAGTCCATG, 5'-GCTGGAGGGCGGGTCTCAGGGAGC, and 5'-AGGACAGAGGTGGAGCCAAGACGACATAGCC. Preparation of cDNA from human foreskin keratinocytes and prostate carcinoma tissue has been described previously. The PCRs were performed under the conditions described above or for PCR with degenerate primers as described previously. Nested PCRs were done by replacing the cDNA with 1.0µl from the first PCR reaction. The 5'-end of the cDNA was isolated by 5'-RACE as described above with the exception of using the gene-specific oligonucleotides P7, 5'-TGAAGCTCAGCCGGAGGTAGAAG, and 5'-GACAGACTCAAGCCGCAAGGTTG.

Northern Hybridization

A human RNA Master Blot containing poly(A)$^+$ mRNA of 50 different tissues was obtained from Clonetech Laboratories, Inc (Palo Alto, CA). ^{32}P-labeled probes were prepared by random prime labelling of DNA fragments of the different transglutaminase gene products using the Multiprime DNA Labelling System (Amersham, Int., Amersham, UK). DNA fragments of 500-700bp compromising the 3'-end of TG_X, TG_Z, and band 4.2 protein, were generated by restriction with *Pst* I and *Acc* I, *Nco* I and *Not* I (exon XII and XIII), and *Xho* I, respectively. The cDNA encoding human band 4.2 protein (Korsgren *et al.* 1990) was kindly provided by Dr Carl M. Cohen, Boston, MA. A ~ 220bp ^{32}P-labeled fragment of TG_Y was generated by PCR using oligonucleotides 5'-CAGCCTCAGTCACCGCCATCCGC and 5'-GATACTTGTAGAGGTCAGTGATG. Hybridization was performed under the conditions recommended by the manufacturer. The labeled membrane was exposed to BioMax MR film (Eastman Kodak) and films developed after 15 to 24hr for first exposure and 3 to 5 days for second exposure.

Amplification of TG_Y and TG_Z from different tissues

cDNA from various cell lines and human tissue was prepared as previously described. A panel of cDNAs from human tissue (Multiple Tissue cDNA Panel I) were also obtained from Clonetech Laboratories. A 365 or 287bp fragment of TG_Z was amplified by PCR using oligonucleotides 5'-TGGGCAAGGCGCTGAGAGTCCATG and

5'-GCTGGAGGGCGGGTCTCAGGGAGC or 5'-AGGACAGAGGTGGAGCCAAGACGACATAGCC, respectively, with an annealing temperature of 60°C. A 218 or 170bp fragment of TG_Y was amplified by PCR using oligonucleotides 5'-CAGCCTCAGTCACCGCCATCCGC and 5'-GATACTTGTAGAGGTCAGTGATG or 5'-GTGAAGGACTGTGCGCTGATG and 5'-CGGGAAGTGAGGGCTTACAAG, respectively, and identical conditions as above.

Mapping of Transglutaminase Genes in Mouse Genome

The 100 radiation hybrid (RH) clones of the T31 mouse/hamster RH panel (McCarthy *et al.*, (1997), Genome Res., 7, 1153-1161) (Research Genetics, Huntsville, AL) were screened by PCR. A 139bp fragment of the *tgm5* gene was amplified with primers5'-TGAGGACTGTGTGCTGACCTTG (f) and 5'-TCCTGTGTCTGGCCTAGGG (r), a 149bp fragment of the *epb42* gene with primers 5'-CAGGAGGAGTAAGGGGAATTGG (f) and 5'-TGCAGGCTACTGGAATCCACG (r), a 400bp fragment of *tgm7* with primers 5'-GGGAGTGGCCTCATCAATGG (f) and 5'-CCTTGACCTCACTGCTGCTGA (r), a -600bp fragment with *tgm3* with primers 5'-TCGGTGGCAGCCTCAAGATTG (f) and 5'-AGACATCAATGGGCAGGCATGG (r), and 655bp and 232bp fragments of *tgm2* with primers 5'-TTGGGGAGCTGGAGAGCAAC (f) and 5'-ATCCAGGACTCCACCCAGCA (r) and primers 5'-(GCGGCCGCTAGT)CCACATTGCAGGGCTCCTGACT (f) and 5'-GCTAGCCTGTGCTCACCATGAGG (r), respectively. PCRs were carried out in a GeneAmp 9600 thermacycler with 0.035 units/μl AmpliTaq Gold polymerase in standard reaction buffer containing 2mM $MgCl_2$, 0.2mM dNTPs, 0.4μM of each primer and 2.5 ng/μl genomic DNA in a total reaction volume of 25μl. PCR conditions were: polymerase activation for 10min at 95°C, annealing at 60°C for 45sec, extension at 72°C for 1min and denaturation at 94°C for 30sec for 35 cycles with a final extension of 3.5min at 72°C. PCR reactions were analyzed by agarose gel electrophoresis using 1% or 1.5% gels. The hybrid cell panel was analyzed at least twice in each case to exclude PCT related errors. The data was submitted to the Jackson Laboratory Radiation Hybrid Database for analysis and mapped relative to known genomic markers (http://www.jax.org/resources/documents/cmdata.rhmap).

Table I. Splice donor and acceptor sequences in the human TGM5 gene. Residues consistent with the splice site consensus sequence (MAG/GTRAG and YAG/G) are underlined.

Intron number	Donor sequence	Acceptor sequence
	M A Q	G L E V A
1	GCTACCATGGCCCAAGgtagggaaagcccctgtggccactggagtt	ttttgtctaaccctggctgccccattgcagGGCTAGAAGTGGCC
	F V V E T	G P L P D
2	TTCGTGGTTGAAACTGgtaagaaccccagctggctcacaggggctg	tggagggcctcagctctacttccctcctagGACCGCTGCCAGAC
	N P W C P	E D A V Y
3	AATCCCTGGTGCCCAGgtaaggctgggtgcccaggcggtgcctcct	tgcttcgtgccctcccactctggttcctagAGGATGCTGTCTAC
	W N Y G Q	F E D K I
4	TGGAACTATGGACAGgtgagtctcagccctgcttatggcccatcc	tgccttccctctctgcctctccccccgaagTTTGAAGACAAAATC
	V V C A M	I N S N D
5	GTGGTGTGTGCCATGgtgaggtccctggcgtgcccggggaggagg	ctcacacttctctatatggcttctcttcagATCAACAGCAATGAT
	A V M C T	V M R C L
6	GCCGTCATGTGCACAGgtaggaggtagaaaggacctcacaaaaagg	acaggtgatttttttgtgccctttttgcagTGATGAGGTGTCTG
	K D T I W	N F H V W
7	AAGGATACTATCTGGtgagaaacaacctctcaacctatttctag	caacgtctcccttggctctgtttgatacagGAACTTCCATGTCTGG
	Q E M S N	G V Y C C
8	CAGGAGATGAGCAACGgtgaggctctccagaagaaaggcaggcccc	gcccaccgaggctcccctgttctccttcagGCGTCTACTGCTGT
	Y K Y E E	G S L Q E
9	TACAAGTATGAAGAAGgttagtaagcaagccagccctactcagagc	cagctggtgctgtgctctccccaacttcagGATCCCTCCAGGAG
	L S P K E	A K T Y P
10	CTCTCTCCTAAAGAAGgtacgcatgtgcacagtttgtgtacgcaga	tctcaccccatccttgtgttctttctttagCAAAGACCTACCCC
	S I T I N	V L G A A
11	AGCATCACGATTAATgtaggcaggagtcctgcaaatggcttgtgg	taattctccttcccctcctggtctgtttagGTTCTAGGAGCAGCC
	Q Q K V F	L G V L K
12	CAGCAGAAAGTCTTgtaagtgctgcaagtgctcagccttctcct	ttttctgacatgctccattctctgttgcagCCTTGGAGTCCTCAAA

Table II. Intron sizes and splice types in the human TGM5 gene. Sizes of introns are estimated to be within about a 100bp unless indicated to be sequenced entirely.

Intron number	Splice type	Size	Method
1	1	6,300bp	PCR
2	1	102bp	Sequencing
3	1	3,300bp	PCR
4	0	2,900bp	PCR
5	0	600bp	PCR
6	1	~11,800bp	PCR and Restriction Analysis
7	2	1,600bp	PCR
8	1	106bp	Sequencing
9	1	2,900bp	PCR
10	1	545bp	Sequencing
11	0	1100bp	PCR
12	2	209bp	Sequencing

Table III. Apparent polymorphisms in the cDNA and genomic DNA sequences for TG_X. The positions with nucleotide and amino acid variations are underlined.

Residue	cDNA (a)		Gene		cDNA (b)	
67	S	TCA	P	CCA	P	CCA
220	Y	TAC	Y	TAT	Y	TAC
352	A	GCA	G	GGA	A	GCA

a. Aeschlimann et al., 1998

b. additional sequence variant isolated in this work

Claims

1. A nucleotide sequence comprising at least a portion of the nucleotide sequence of Fig. 6A or Fig. 6B; a nucleotide sequence which hybridise to the nucleotide sequence of Fig. 6A or Fig. 6B; a nucleotide sequence which is degenerate to the nucleotide sequence of Fig. 6A or Fig. 6B; all of which nucleotide sequences encode a polypeptide having transglutaminase activity.

2. A nucleotide sequence according to claim 1 consisting of the nucleotide sequence of Fig. 6A or Fig. 6B.

3. A nucleotide sequence which hybridises under stringent conditions to the nucleotide sequence of Fig. 6A or Fig. 6B and which encodes a polypeptide having transglutaminase activity.

4. A method of expressing a polypeptide comprising inserting a nucleotide sequence according to any preceding claim into a suitable host and expressing that nucleotide sequence in order to express a polypeptide having transglutaminase activity.

5. A vector comprising a nucleotide sequence according to any one of claims 1 to 3.

6. A polypeptide having an amino acid sequence comprising at least a portion of the amino acid sequence of Fig. 6A or Fig. 6B and which has transglutaminase activity.

7. A polypeptide according to claim 6 which is at least 90% identical to the amino acid sequence of Fig. 6A or Fig. 6B and which encodes a polypeptide having transglutaminase activity.

8. A polypeptide according to claim 6 or 7 where the amino acid sequence differs from that given in Fig. 6A or Fig. 6B by about 1 to 20 amino acid additions, deletions or substitutions.

9. A polypeptide according to any one of claims 6 to 8 comprising exon VII through to exon X of the sequence shown in Fig. 6A or Fig. 6B.

10. A composition comprising a polypeptide according to any one of claims 6 to 9 suitable for use in cross-linking proteins.

11. A composition comprising a polypeptide according to any one of claims 6 to 9 suitable for use in a transamidation reaction on peptides and polypeptides.

12. A diagnostic method comprising detecting expression of a polypeptide according to any one of claims 6 to 9 in a subject or in cells derived from a subject.

13. An antibody directed against a polypeptide according to any one of claims 6 to 9.

14. A method of gene therapy comprising correcting mutations in a non wild type nucleotide sequence corresponding to a nucleotide sequence of Fig. 6A or Fig. 6B.

15. A nucleotide sequence comprising at least a portion of the nucleotide sequence of Fig. 10A or Fig. 10B; a nucleotide sequence which hybridise to the nucleotide sequence of Fig. 10A or Fig. 10B; a nucleotide sequence which is degenerate to the nucleotide sequence of Fig. 10A or Fig. 10B; all of which nucleotide sequences encode a polypeptide having transglutaminase activity.

16. A nucleotide sequence according to claim 15 consisting of the nucleotide sequence of Fig. 10A or Fig. 10B.

17. A nucleotide sequence which hybridises under stringent conditions to the nucleotide sequence of Fig. 10A or Fig. 10B and which encodes a polypeptide having transglutaminase activity.

18. A method of expressing a polypeptide comprising inserting a nucleotide sequence according to any one of claims 15 to 17 into a suitable host and expressing that nucleotide sequence in order to express a polypeptide having transglutaminase activity.

19. A vector comprising a nucleotide sequence according to any one of claims 15 to 17.

20. A polypeptide having an amino acid sequence comprising at least a portion of the amino acid sequences of Fig. 10A or Fig. 10B and which has transglutaminase activity.

21. A polypeptide according to claim 20 which is at least 90% identical to the amino acid sequences of Fig. 10A or Fig. 10B and which encodes a polypeptide having transglutaminase activity.

22. A polypeptide according to claim 20 or 21 wherein the amino acid sequence differs from that given in Fig. 10A or Fig. 10B by about 1 to 20 amino acid additions, deletions or substitutions.

23. A polypeptide according to any one of claims 20 to 22 comprising exons II through to exon IV of the sequence shown in Fig. 10A or Fig. 10B.

24. A polypeptide according to any one of claims 20 to 22 comprising exon X through to exon XII of the sequence shown in Fig. 10A or Fig. 10B.

25. A composition comprising a polypeptide according to any one of claims 20 to 24 suitable for use in cross-linking proteins.

26. A composition comprising a polypeptide according to any one of claims 20 to 24 suitable for use in a transamidation reaction on peptides and polypeptides.

27. A diagnostic method comprising detecting expression of a polypeptide according to any one of claims 20 to 24 in a subject or in cells derived from a subject.

28. An antibody directed against a polypeptide according to any one of claims 20 to 24.

29. A method of gene therapy comprising correcting mutations in a non-wild type nucleotide sequence corresponding to the nucleotide sequence of Fig. 10A or Fig. 10B.

30. A method of diagnosis of autoimmune disease comprising taking a sample from a subject and testing that sample for the presence of a transglutaminase encoded by the nucleotide sequences of Fig. 6A, Fig. 6B, Fig. 10A or Fig. 10B, or portions thereof.

31. A method according to claim 30 wherein the autoimmune disease to be diagnosed is selected from Addison's disease, AI haemolytic anaemia, AI thrombocytopenic purpura, AI thyroid diseases, atrophic gastritis - pernicious anaemia, Chron's disease, colitis ulcerosa, Goodpasture syndrome, IgA nephropathy or IgA glomerulonephritis, myasthenia gravis, partial lipodystrophy, polymyositis, primary biliary cirrhosis, primary sclerosing cholangitis, progressive systemic sclerosis, recurrent pericarditis, relapsing polychondritis, rheumatoid arthritis, rheumatism, sarcoidosis, Sjögren's syndrome, SLE, splenic atrophy, type I (insulin-dependent) diabetes mellitus, diabetis mellitus, Wegener granulomatosis, ulcerative colitis, vasculitis (both systemic and cutaneous), vitiligo.

32. A competitive protein binding assay for the differential diagnosis of autoimmune diseases comprising the detection of antibodies against the transglutaminase encoded by the nucleotide sequences of Fig.6A, Fig. 6B, Fig.10A or Fig. 10B, or portions thereof.

33. A competitive protein binding assay according to claim 32 comprising non-endogenous transglutaminase TG_Z or TG_Y, or both, as a competitive antigen.

34. Competitive protein binding assay according to claim 33, wherein the binding assay is a competitive immunoassay selected from RIA, EIA/ELISA, LiA and FiA.

Fig. 1

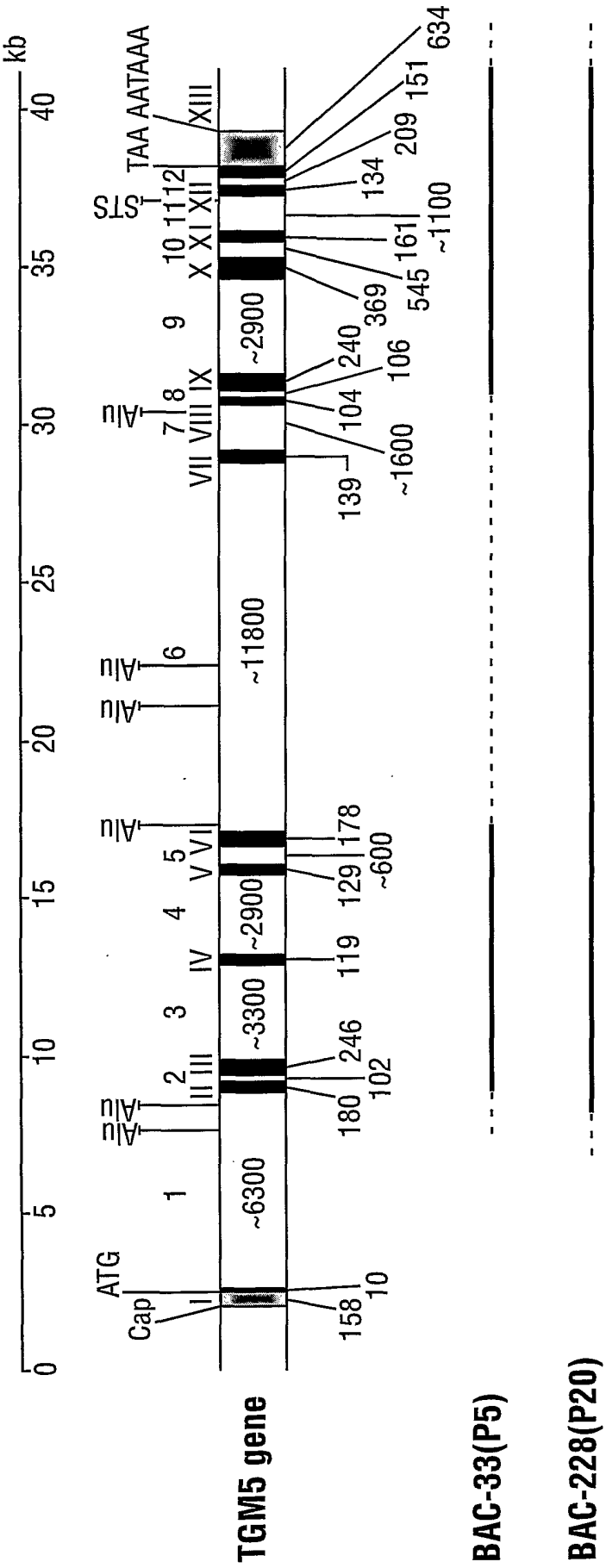

kb
0
5
10
15
20
25
30
35
40
ATG
Cap
TAA AATAAA
STS
Alu
~6300
~3300
~2900
~11800
~2900
~2900
~1600
~600
~1100
TGM5 gene
BAC-33(P5)
BAC-228(P20)

Fig. 2A

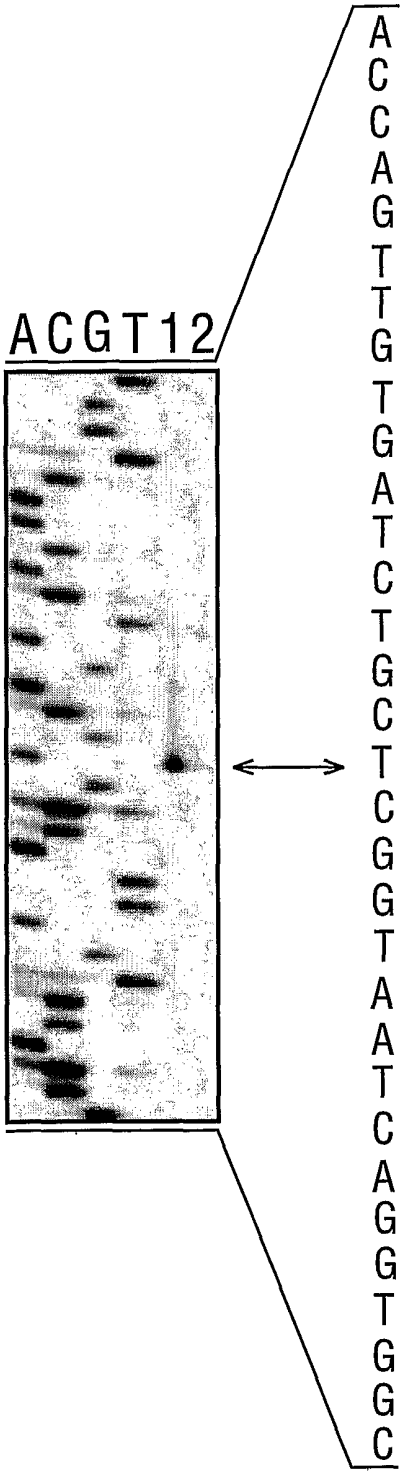


Fig. 2B

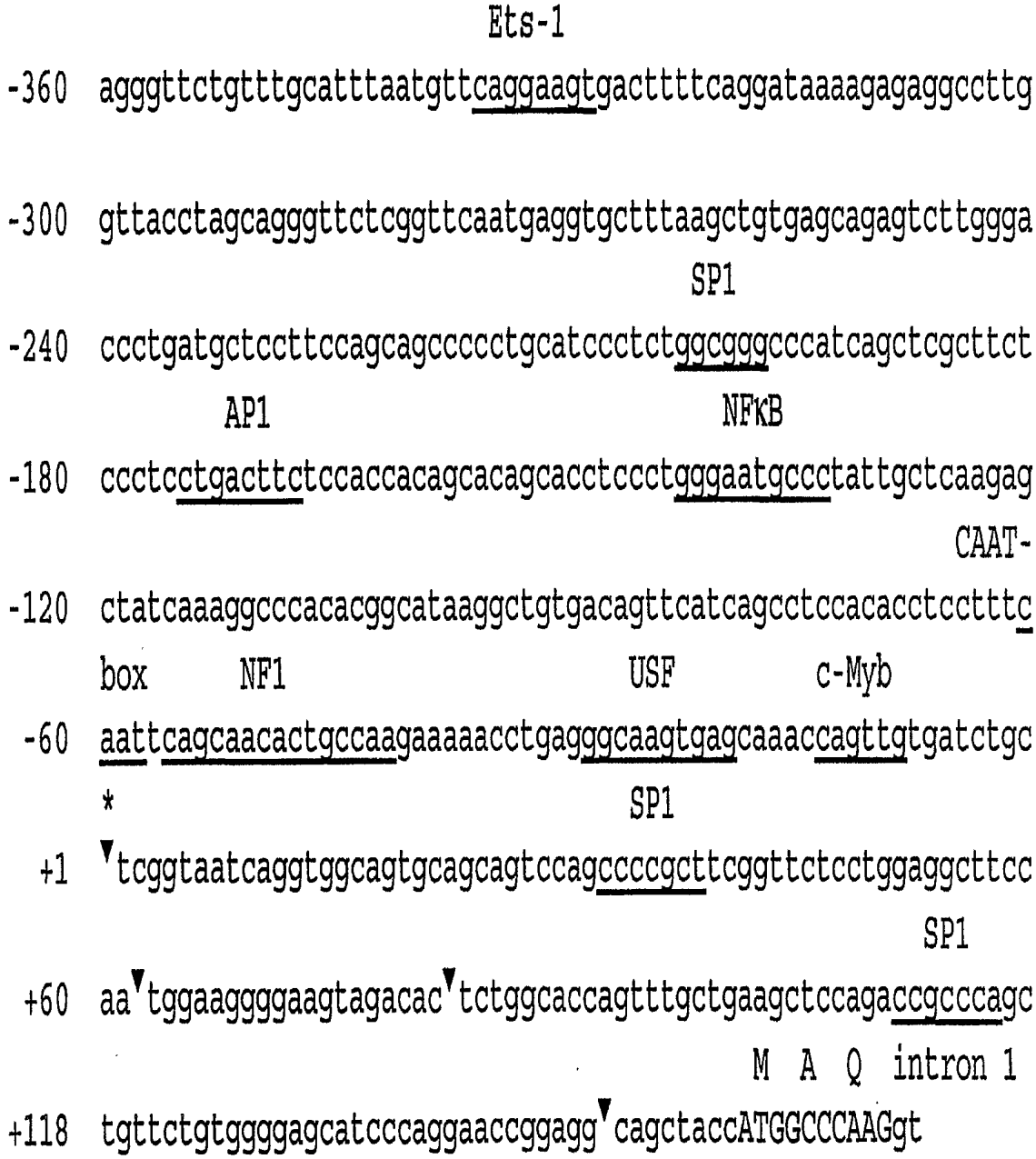

Ets-1
-360 agggttctgtttgcatttaatgttcaggaagtgacttttcaggataaaagagaggccttg
-300 gttacctagcagggttctcggttcaatgaggtgctttaagctgtgagcagagtcttggga
SP1
-240 ccctgatgctccttccagcagccccctgcatccctctggcgggcccatcagctcgcttct
AP1
NFκB
-180 ccctcctgacttctccaccacagcacagcacctccctgggaatgccctattgctcaagag
CAAT-
-120 ctatcaaaggcccacacggcataaggctgtgacagttcatcagcctccacacctcctttc
box
NF1
USF
c-Myb
-60 aattcagcaacactgccaagaaaaacctgagggcaagtgagcaaaccagttgtgatctgc
*
SP1
+1 tcggtaatcaggtggcagtgcagcagtccagccccgcttcggttctcctggaggcttcc
SP1
+60 aatggaaggggaagtagacactctggcaccagtttgctgaagctccagaccgcccagc
M A Q intron 1
+118 tgttctgtggggagcatcccaggaaccggaggcagctaccATGGCCCAAGgt

Fig. 3

V D F A L och

2303 GTAGACTTGCATTAtaaattctggaac▼aacgcgcc▼agacgtgtgaatttc▼aagcttcagg

2364 aaaaggagcaagttcaaatgcaagctgcgccaattcccccaccacacaagaggcttcacagggctc

2428 cagcaagagccacagaggggatgacgtgttcatttttctgtcttcctgactccactagaaattt

2492 aagctccatgagggcaaagactttgctttgtttactaccccctatactcagaaccatttctggca

2556 tatgctaggcactcaacaaaatatttttgaatgaataggagactccagcatccagagaac

2620 aggtaggaaatgtctatggatggatatttcccctggaccatttgcacagctccccctggactcttt

2684 tcagggcccagggattccactgtgtcccatccagagattccaggattccagtcacctatccaga

2748 agcgtgatttggcacagaggtcagaggatactggtaggacttggccatgaacttaactgccccct

2812 gccccagatatccaggaaagaaaaagacaggctgaacagctcactgttgtttgttgttgcgaa

2876 agctaaattcccctagtatgaataaacttcagaccttgctcttgcctcatgtagtcatcact

2940 ttcctatctgttccttggatagctgcagtctccattcattcaaaaaagtcatttattgagtgcc

3004 tagcatatgccagaagtggttctgagtgtaggggtacaaagtaaacaaagcaaagtccctgccc

3068 tcatggagcgcacatttctcagtggagg

Fig. 4

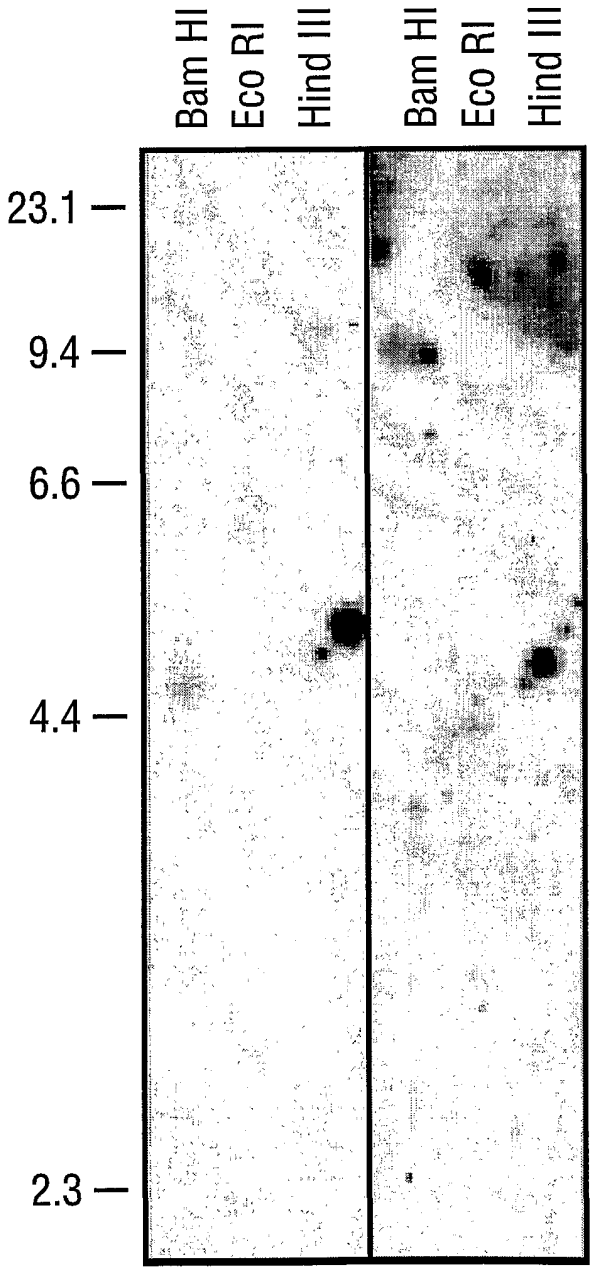
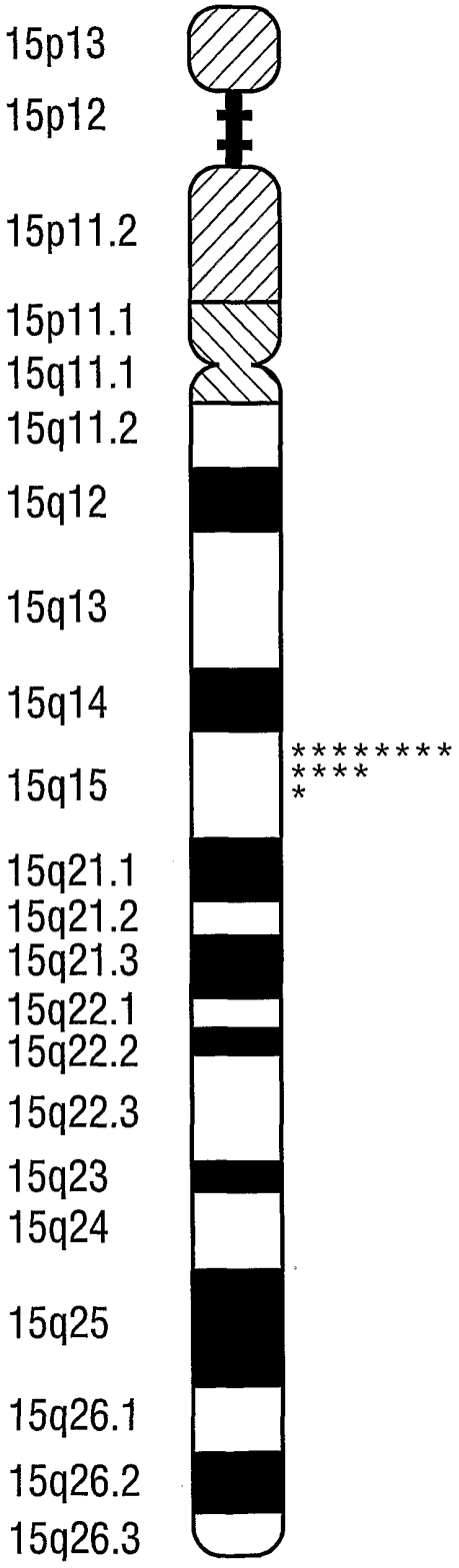


Fig. 5A

Fig. 5B


15p13
15p12
15p11.2
15p11.1
15q11.1
15q11.2
15q12
15q13
15q14
15q15
15q21.1
15q21.2
15q21.3
15q22.1
15q22.2
15q22.3
15q23
15q24
15q25
15q26.1
15q26.2
15q26.3
****
****
*

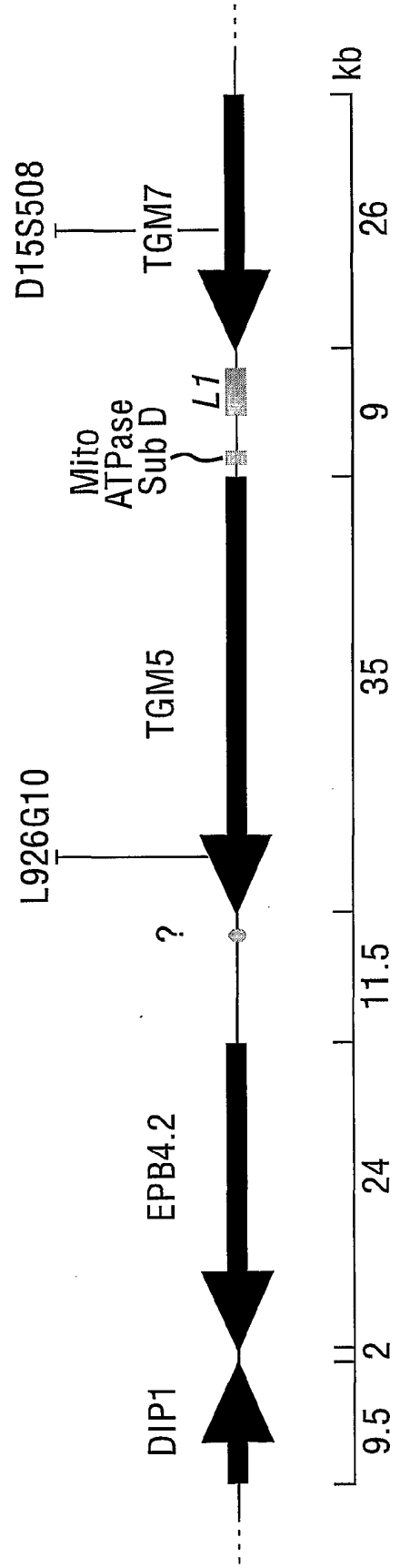

D15S508
TGM7
Mito ATPase Sub D
L1
L926G10
TGM5
?
EPB4.2
DIP1
kb
26
9
35
11.5
24
2
9.5

Fig 5C

Fig. 6A

```
                                                                      5
                                                                 AC CAC

          14          23          32          41          50          59
ATG GTG GGA ATG GCA ACC TTG CGG CTT GAG TCT GTC GAC CTG CAG AGC TCC AGG
 M   V   G   M   A   T   L   R   L   E   S   V   D   L   Q   S   S   R

          68          77          86          95         104         113
AAC AAC AAG GAG CAC CAC ACG CAG GAG ATG GGC GTC AAG CGG CTC ACT GTG CGC
 N   N   K   E   H   H   T   Q   E   M   G   V   K   R   L   T   V   R

         122         131         140         149         158         167
CGC GGC CAG CCC TTC TAC CTC CGG CTG AGC TTC AGC CGA CCC TTC CAG TCC CAG
 R   G   Q   P   F   Y   L   R   L   S   F   S   R   P   F   Q   S   Q

         176         185         194         203         212         221
AAC GAC CAC ATC ACC TTT GTG GCT GAG ACC GGA CCC AAG CCG TCA GAG CTG CTG
 N   D   H   I   T   F   V   A   E   T   G   P   K   P   S   E   L   L

         230         239         248         257         266         275
GGG ACC CGA GCC ACA TTC TTC CTC ACC CGG GTC CAG CCC GGG AAT GTC TGG AGC
 G   T   R   A   T   F   F   L   T   R   V   Q   P   G   N   V   W   S

         284         293         302         311         320         329
GCT TCT GAT TTC ACC ATT GAC TCC AAC TCT CTC CAA GTT TCC CTT TTC ACA CCA
 A   S   D   F   T   I   D   S   N   S   L   Q   V   S   L   F   T   P

         338         347         356         365         374         383
GCC AAT GCA GTT ATT GGC CAT TAC ACT CTG AAA ATA GAG ATC TCT CAG GGC CAA
 A   N   A   V   I   G   H   Y   T   L   K   I   E   I   S   Q   G   Q

         392         401         410         419         428         437
GGT CAC AGT GTG ACT TAC CCG CTG GGA ACT TTC ATC CTA CTT TTT AAC CCT TGG
 G   H   S   V   T   Y   P   L   G   T   F   I   L   L   F   N   P   W

         446         455         464         473         482         491
AGT CCA GAG GAC GAC GTC TAC CTG CCA AGT GAA ATA CTG CTG CAG GAG TAT ATC
 S   P   E   D   D   V   Y   L   P   S   E   I   L   L   Q   E   Y   I

         500         509         518         527         536         545
ATG CGA GAT TAT GGC TTT GTT TAC AAG GGT CAT GAA AGA TTC ATC ACC TCC TGG
 M   R   D   Y   G   F   V   Y   K   G   H   E   R   F   I   T   S   W

         554         563         572         581         590         599
CCC TGG AAC TAC GGG CAG TTT GAA GAG GAC ATC ATA GAC ATC TGC TTT GAG ATC
 P   W   N   Y   G   Q   F   E   E   D   I   I   D   I   C   F   E   I

         608         617         626         635         644         653
CTG AAC AAG AGC CTG TAT CAC TTA AAG AAC CCG GCC AAA GAC TGT TCC CAG CGG
 L   N   K   S   L   Y   H   L   K   N   P   A   K   D   C   S   Q   R

         662         671         680         689         698         707
AAC GAC GTG GTG TAT GTG TGC AGG GTG GTG AGT GCC ATG ATC AAC AGC AAC GAT
 N   D   V   V   Y   V   C   R   V   V   S   A   M   I   N   S   N   D

         716         725         734         743         752         761
GAC AAT GGC GTG CTG CAG GGG AAC TGG GGC GAG GAC TAC TCC AAA GGG GTC AGT
 D   N   G   V   L   Q   G   N   W   G   E   D   Y   S   K   G   V   S

         770         779         788         797         806         815
CCT CTG GAG TGG AAG GGC AGT GTG GCC ATC CTA CAG CAG TGG TCA GCC AGG GGC
 P   L   E   W   K   G   S   V   A   I   L   Q   Q   W   S   A   R   G

         824         833         842         851         860         869
GGG CAG CCT GTG AAG TAC GGA CAG TGC TGG GTC TTC GCC TCT GTT ATG TGC ACC
```

Fig. 6A(contd.)

```
 G   Q   P   V   K   Y   G   Q   C   W   V   F   A   S   V   M   C   T

         878         887         896         905         914         923
GTA ATG AGA TGC TTA GGT GTT CCA ACC CGT GTT GTT TCC AAT TTC CGT TCC GCG
 V   M   R   C   L   G   V   P   T   R   V   V   S   N   F   R   S   A

         932         941         950         959         968         977
CAC AAC GTG GAT AGG AAC TTG ACC ATC GAT ACG TAC TAT GAC CGA AAT GCC GAG
 H   N   V   D   R   N   L   T   I   D   T   Y   Y   D   R   N   A   E

         986         995        1004        1013        1022        1031
ATG CTG TCA ACT CAG AAA CGA GAC AAA ATA TGG AAC TTC CAC GTC TGG AAT GAG
 M   L   S   T   Q   K   R   D   K   I   W   N   F   H   V   W   N   E

        1040        1049        1058        1067        1076        1085
TGC TGG ATG ATC CGG AAA GAT CTC CCA CCA GGA TAC AAC GGG TGG CAG GTT CTG
 C   W   M   I   R   K   D   L   P   P   G   Y   N   G   W   Q   V   L

        1094        1103        1112        1121        1130        1139
GAC CCC ACT CCC CAG CAG ACC AGC AGT GGG CTG TTC TGC TGT GGC CCT GCC TCT
 D   P   T   P   Q   Q   T   S   S   G   L   F   C   C   G   P   A   S

        1148        1157        1166        1175        1184        1193
GTG AAG GCC ATC AGG GAA GGG GAT GTC CAC CTG GCC TAT GAC ACC CCT TTT GTG
 V   K   A   I   R   E   G   D   V   H   L   A   Y   D   T   P   F   V

        1202        1211        1220        1229        1238        1247
TAT GCC GAG GTG AAC GCC GAT GAA GTC ATT TGG CTC CTT GGG GAT GGC CAG GCC
 Y   A   E   V   N   A   D   E   V   I   W   L   L   G   D   G   Q   A

        1256        1265        1274        1283        1292        1301
CAG GAA ATC CTG GCC CAC AAC ACC AGT TCC ATC GGG AAG GAG ATC AGC ACT AAG
 Q   E   I   L   A   H   N   T   S   S   I   G   K   E   I   S   T   K

        1310        1319        1328        1337        1346        1355
ATG GTG GGG TCA GAC CAG CGC CAG AGC ATC ACC AGC TCC TAC AAG TAC CCA GAA
 M   V   G   S   D   Q   R   Q   S   I   T   S   S   Y   K   Y   P   E

        1364        1373        1382        1391        1400        1409
GGA TCC CCT GAG GAG AGA GCT GTC TTC ATG AAG GCT TCT CGG AAA ATG CTG GGC
 G   S   P   E   E   R   A   V   F   M   K   A   S   R   K   M   L   G

        1418        1427        1436        1445        1454        1463
CCC CAA AGA GCT TCT TTG CCC TTC CTG GAT CTC CTG GAG TCT GGG GGT CTT AGG
 P   Q   R   A   S   L   P   F   L   D   L   L   E   S   G   G   L   R

        1472        1481        1490        1499        1508        1517
GAT CAG CCA GCG CAG CTG CAG CTT CAC CTG GCC AGG ATA CCC GAG TGG GGC CAG
 D   Q   P   A   Q   L   Q   L   H   L   A   R   I   P   E   W   G   Q

        1526        1535        1544        1553        1562        1571
GAC CTG CAG CTG CTG CTG CGT ATC CAG AGG GTG CCA GAC AGC ACC CAC CCT CGG
 D   L   Q   L   L   L   R   I   Q   R   V   P   D   S   T   H   P   R

        1580        1589        1598        1607        1616        1625
GGG CCC ATC GGA CTG GTG GTG CGC TTC TGT GCA CAG GCC CTG CTG CAT GGG GGT
 G   P   I   G   L   V   V   R   F   C   A   Q   A   L   L   H   G   G

        1634        1643        1652        1661        1670        1679
GGT ACC CAG AAG CCC TTC TGG AGG CAC ACA GTG CGG ATG AAC CTG GAC TTT GGG
 G   T   Q   K   P   F   W   R   H   T   V   R   M   N   L   D   F   G

        1688        1697        1706        1715        1724        1733
AAG GAG ACA CAG TGG CCG CTC CTC CTG CCC TAC AGC AAT TAC AGA AAC AAG CTA
 K   E   T   Q   W   P   L   L   L   P   Y   S   N   Y   R   N   K   L
```

```
      1742        1751        1760        1769        1778        1787
ACG GAC GAA AAG CTC ATC CGC GTG TCT GGC ATC GCG GAG GTT GAA GAG ACA GGG
 T   D   E   K   L   I   R   V   S   G   I   A   E   V   E   E   T   G

      1796        1805        1814        1823        1832        1841
AGG TCC ATG CTG GTC CTA AAA GAT ATC TGT CTG GAG CCT CCC CAC TTG TCT ATT
 R   S   M   L   V   L   K   D   I   C   L   E   P   P   H   L   S   I

      1850        1859        1868        1877        1886        1895
GAG GTG TCT GAG AGG GCT GAG GTG GGC AAG GCG CTG AGA GTC CAT GTC ACC CTC
 E   V   S   E   R   A   E   V   G   K   A   L   R   V   H   V   T   L

      1904        1913        1922        1931        1940        1949
ACC AAC ACC TTA ATG GTG GCT CTG AGC AGC TGC ACG ATG GTG CTG GAA GGA AGC
 T   N   T   L   M   V   A   L   S   S   C   T   M   V   L   E   G   S

      1958        1967        1976        1985        1994        2003
GGC CTC ATC AAT GGG CAG ATA GCA AAG GAC CTT GGG ACT CTG GTG GCC GGA CAC
 G   L   I   N   G   Q   I   A   K   D   L   G   T   L   V   A   G   H

      2012        2021        2030        2039        2048        2057
ACC CTC CAA ATT CAA CTG GAC CTC TAC CCG ACC AAA GCT GGA CCC CGC CAG CTC
 T   L   Q   I   Q   L   D   L   Y   P   T   K   A   G   P   R   Q   L

      2066        2075        2084        2093        2102        2111
CAG GTT CTC ATC AGC AGC AAC GAG GTC AAG GAG ATC AAA GGC TAC AAG GAC ATA
 Q   V   L   I   S   S   N   E   V   K   E   I   K   G   Y   K   D   I

      2120        2129        2138        2147        2156        2165
TTC GTC ACT GTG GCT GGG GCT CCC TGA GAC CCG CCC TCC AGC TGC CCT CCC TGG
 F   V   T   V   A   G   A   P   *

      2174        2183        2192        2201        2210        2219
CAC CCC TGC CCC ACC TGG CTC CTT TCT ACT CCT GGC TAT GTC GTC TTG GCT CCA

      2228        2237        2246        2255        2264        2273
CCT CTG TCC TCT CTC TAG CCT GCC TGG GAA TGA ATG AAG CTC TGT TAG AAA CAC

      2282        2291        2300        2309
CGT GTG CTT TGG GAA GAG ACA ATA AAG ATG TCT TTA TTT ATC AC
```

Fig. 6A(contd.)

Fig. 6B

```
                                                                          GG  GAG     5
ATG GAT CAG GTG GCA ACC TTG CGG CTT GAG TCT GTC GAC CTG CAG AGC TCC AGG AAC AAC AAG GAG CAC    74
 M   D   Q   V   A   T   L   R   L   E   S   V   D   L   Q   S   S   R   N   N   K   E   H     23
CAC ACG CAG GAG ATG GGC GTC AAG CGG CTC ACT GTG CGC CGC GGC CAG CCC TTC TAC CTC CGG CTG AGC   143
 H   T   Q   E   M   G   V   K   R   L   T   V   R   R   G   Q   P   F   Y   L   R   L   S     46
TTC AGC CGA CCC TTC CAG TCC CAG AAC GAC CAC ATC ACC TTT GTG GCT GAG ACC GGA CCC AAG CCG TCA   212
 F   S   R   P   F   Q   S   Q   N   D   H   I   T   F   V   A   E   T   G   P   K   P   S     69
GAG CTG CTG GGG ACC CGA GCC ACA TTC TTC CTC ACC CGG GTC CAG CCC GGG AAT GTC TGG AGC GCT TCT   281
 E   L   L   G   T   R   A   T   F   F   L   T   R   V   Q   P   G   N   V   W   S   A   S     92
GAT TTC ACC ATT GAC TCC AAC TCT CTC CAA GTT TCC CTT TTC ACA CCA GCC AAT GCA GTT ATT GGC CAT   350
 D   F   T   I   D   S   N   S   L   Q   V   S   L   F   T   P   A   N   A   V   I   G   H    115
TAC ACT CTG AAA ATA GAG ATC TCT CAG GGC CAA GGT CAC AGT GTG ACT TAC CCG CTG GGA ACT TTC ATC   419
 Y   T   L   K   I   E   I   S   Q   G   Q   G   H   S   V   T   Y   P   L   G   T   F   I    138
CTA CTT TTT AAC CCT TGG AGT CCA GAG GAC GAC GTC TAC CTG CCA AGT GAA ATA CTG CTG CAG GAG TAT   488
 L   L   F   N   P   W   S   P   E   D   D   V   Y   L   P   S   E   I   L   L   Q   E   Y    161
ATC ATG CGA GAT TAT GGC TTT GTT TAC AAG GGT CAT GAA AGA TTC ATC ACC TCC TGG CCC TGG AAC TAC   557
 I   M   R   D   Y   G   F   V   Y   K   G   H   E   R   F   I   T   S   W   P   W   N   Y    184
GGG CAG TTT GAA GAG GAC ATC ATA GAC ATC TGC TTT GAG ATC CTG AAC AAG AGC CTG TAT CAC TTA AAG   626
 G   Q   F   E   E   D   I   I   D   I   C   F   E   I   L   N   K   S   L   Y   H   L   K    207
AAC CCG GCC AAA GAC TGT TCC CAG CGG AAC GAC GTG GTG TAT GTG TGC AGG GTG GTG AGT GCC ATG ATC   695
 N   P   A   K   D   C   S   Q   R   N   D   V   V   Y   V   C   R   V   V   S   A   M   I    230
AAC AGC AAC GAT GAC AAT GGC GTG CTG CAG GGG AAC TGG GGC GAG GAC TAC TCC AAA GGG GTC AGT CCT   764
 N   S   N   D   D   N   G   V   L   Q   G   N   W   G   E   D   Y   S   K   G   V   S   P    253
CTG GAG TGG AAG GGC AGT GTG GCC ATC CTA CAG CAG TGG TCA GCC AGG GGC GGG CAG CCT GTG AAG TAC   833
 L   E   W   K   G   S   V   A   I   L   Q   Q   W   S   A   R   G   G   Q   P   V   K   Y    276
GGA CAG TGC TGG GTC TTC GCC TCT GTT ATG TGC ACC GTA ATG AGA TGC TTA GGT GTT CCA ACC CGT GTT   902
 G   Q   C   W   V   F   A   S   V   M   C   T   V   M   R   C   L   G   V   P   T   R   V    299
GTT TCC AAT TTC CGT TCC GCG CAC AAC GTG GAT AGG AAC TTG ACC ATC GAT ACG TAC TAT GAC CGA AAT   971
 V   S   N   F   R   S   A   H   N   V   D   R   N   L   T   I   D   T   Y   Y   D   R   N    322
GCC GAG ATG CTG TCA ACT CAG AAA CGA GAC AAA ATA TGG AAC TTC CAC GTC TGG AAT GAG TGC TGG ATG  1040
 A   E   M   L   S   T   Q   K   R   D   K   I   W   N   F   H   V   W   N   E   C   W   M    345
ATC CGG AAA GAT CTC CCA CCA GGA TAC AAC GGG TGG CAG GTT CTG GAC CCC ACT CCC CAG CAG ACC AGC  1109
 I   R   K   D   L   P   P   G   Y   N   G   W   Q   V   L   D   P   T   P   Q   Q   T   S    368
```

```
AGT GGG CTG TTC TGC TGT GGC CCT GCC TCT GTG AAG GCC ATC AGG GAA GGG GAT GTC CAC CTG GCC TAT 1178
 S  G  L  F  C  C  G  P  A  S  V  K  A  I  R  E  G  D  V  H  L  A  Y  391
GAC ACC CCT TTT GTG TAT GCC GAG GTG AAC GCC GAT GAA GTC ATT TGG CTC CTT GGG GAT GGC CAG GCC 1247
 D  T  P  F  V  Y  A  E  V  N  A  D  E  V  I  W  L  L  G  D  G  Q  A  414
CAG GAA ATC CTG GCC CAC AAC ACC AGT TCC ATC GGG AAG GAG ATC AGC ACT AAG ATG GTG GGG TCA GAC 1316
 Q  E  I  L  A  H  N  T  S  S  I  G  K  E  I  S  T  K  M  V  G  S  D  437
CAG CGC CAG AGC ATC ACC AGC TCC TAC AAG TAC CCA GAA GGA TCC CCT GAG GAG AGA GCT GTC TTC ATG 1385
 Q  R  Q  S  I  T  S  S  Y  K  Y  P  E  G  S  P  E  E  R  A  V  F  M  460
AAG GCT TCT CGG AAA ATG CTG GGC CCC CAA AGA GCT TCT TTG CCC TTC CTG GAT CTC GAG TCT GGG 1454
 K  A  S  R  K  M  L  G  P  Q  R  A  S  L  P  F  L  D  L  L  E  S  G  483
GGT CTT AGG GAT CAG CCA GCG CAG CTG CAG CTT CAC CTG GCC AGG ATA CCC GAG TGG GGC CAG GAC CTG 1523
 G  L  R  D  Q  P  A  Q  L  Q  L  H  L  A  R  I  P  E  W  G  Q  D  L  506
CAG CTG CTG CGT ATC CAG AGG GTG CCA GAC AGC ACC CAC CCT CGG GGG CCC ATC GGA CTG GTG GTG 1592
 Q  L  L  R  I  Q  R  V  P  D  S  T  H  P  R  G  P  I  G  L  V  V  529
CGC TTC TGT GCA CAG GCC CTG CTG CAT GGG GGT GGT ACC CAG AAG CCC TTC TGG AGG CAC ACA GTG CGG 1661
 R  F  C  A  Q  A  L  L  H  G  G  G  T  Q  K  P  F  W  R  H  T  V  R  552
ATG AAC CTG GAC TTT GGG AAG GAG ACA CAG TGG CCG CTC CTC CTG CCC TAC AGC AAT TAC AGA AAC AAG 1730
 M  N  L  D  F  G  K  E  T  Q  W  P  L  L  L  P  Y  S  N  Y  R  N  K  575
CTA ACG GAC GAA AAG CTC ATC CGC GTG TCT GGC ATC GCG GAG GTT GAA GAG GTG TCT GAG AGG GCT GGC TCC ATG CTG 1799
 L  T  D  E  K  L  I  R  V  S  G  I  A  E  V  E  E  T  G  R  S  M  L  598
GTC CTA AAA GAT ATC TGT CTG GAG CCT CCC CAC TTG TCT ATT GAG GTG TCT GAG AGG GCT GAG GTG GGC 1868
 V  L  K  D  I  C  L  E  P  P  H  L  S  I  E  V  S  E  R  A  E  V  G  621
AAG GCG CTG AGA GTC CAT GTC ACC TCC ACC AAC ACC TTA ATG GTG GCT CTG AGC AGC TGC ACG ATG GTG 1937
 K  A  L  R  V  H  V  T  L  T  T  N  T  L  M  V  A  L  S  S  C  T  M  V  644
CTG GAA GGA AGC GGC CTC ATC AAT GGG CAG ATA GCA AAG GAC CTT GGG ACT CTG GTG GCC GGA CAC ACC 2006
 L  E  G  S  G  L  I  N  G  Q  I  A  K  D  L  G  T  L  V  A  G  H  T  667
CTC CAA ATT CAA CTG GAC CTC TAC CCG ACC AAA GCT GGA CCC CGC CAG CTC CAG GTT CTC ATC AGC AGC 2075
 L  Q  I  Q  L  D  L  Y  P  T  K  A  G  P  R  Q  L  Q  V  L  I  S  S  690
AAC GAG GTC AAG GAG ATC AAA GGC TAC AAG GAC ATA TTC GTC ACT GTG GCT GGG GCT CCC TGA GAC CCG 2144
 N  E  V  K  E  I  K  G  Y  K  D  I  F  V  T  V  A  G  A  P  *        710
CCC TCC AGC TGC CCT CCC TGG CAC CCC TGC CCC ACC TGG CTC CTT TCT ACT CCT GGC TAT GTC TTG 2213
GCT CCA CCT CTG TCC TCT CTC TAG CCT GCC TGG GAA TGA ATG AAG CTC TGT TAG AAA CAC CGT GTG CTT 2282
TGG GAA GAG ACA ATA AAG ATG TCT TTA TTT                                                  2312
```

Fig. 6B (contd.)

Fig. 7A

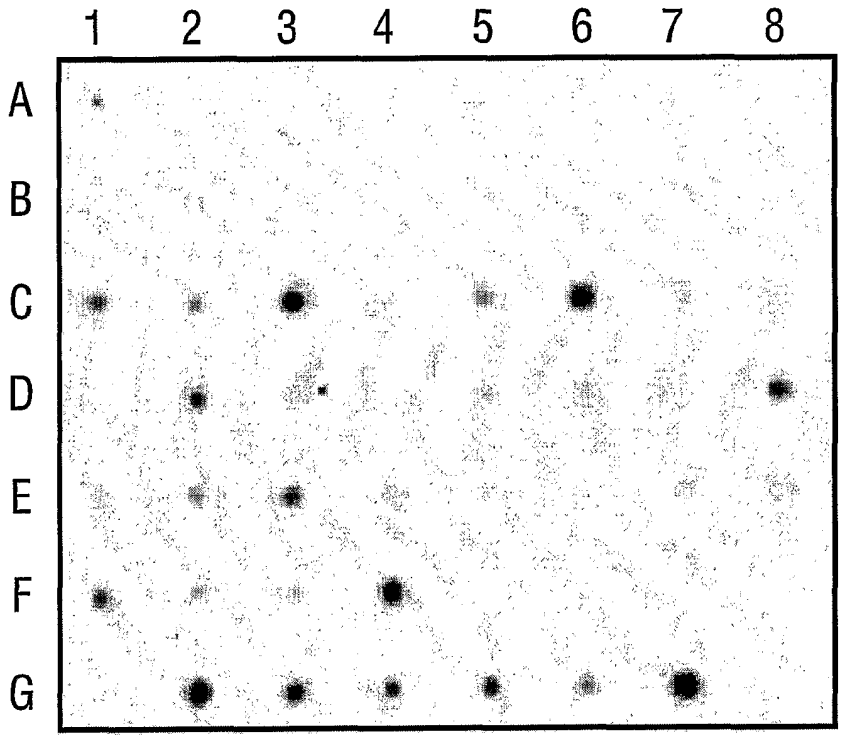


Fig. 7B

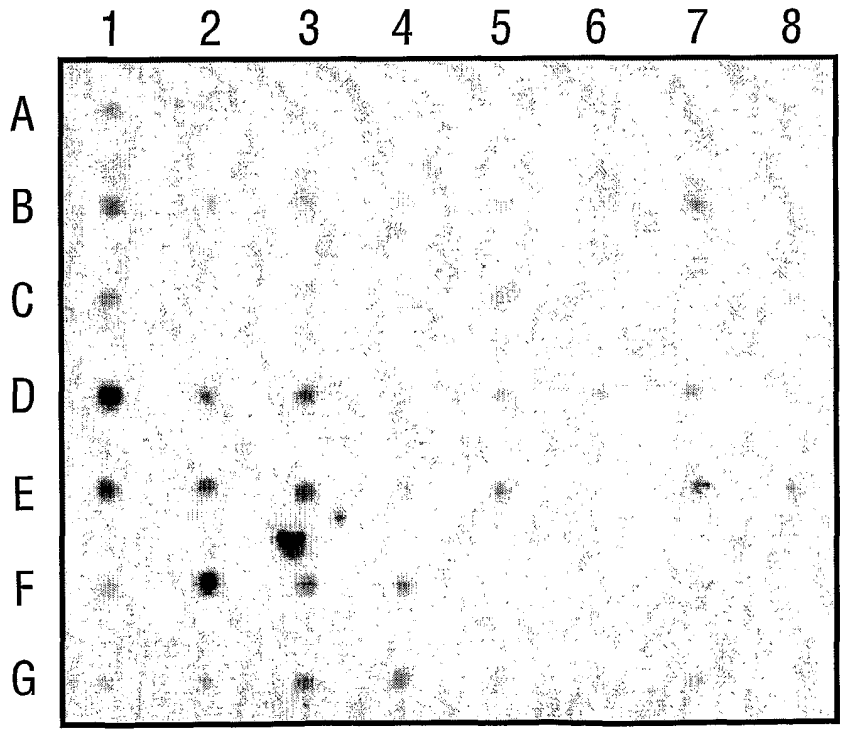


Fig. 7C

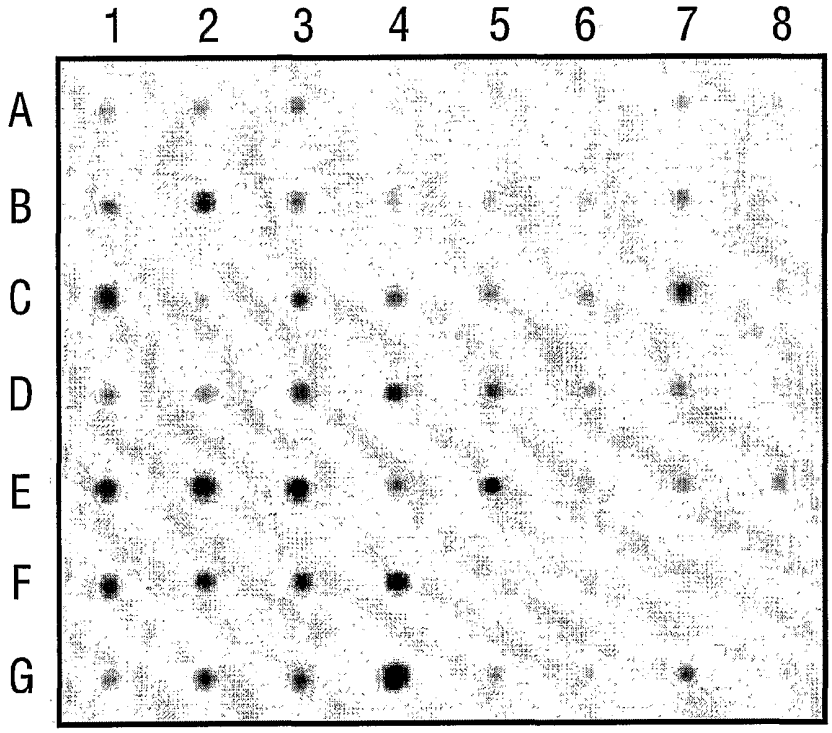


Fig. 7D

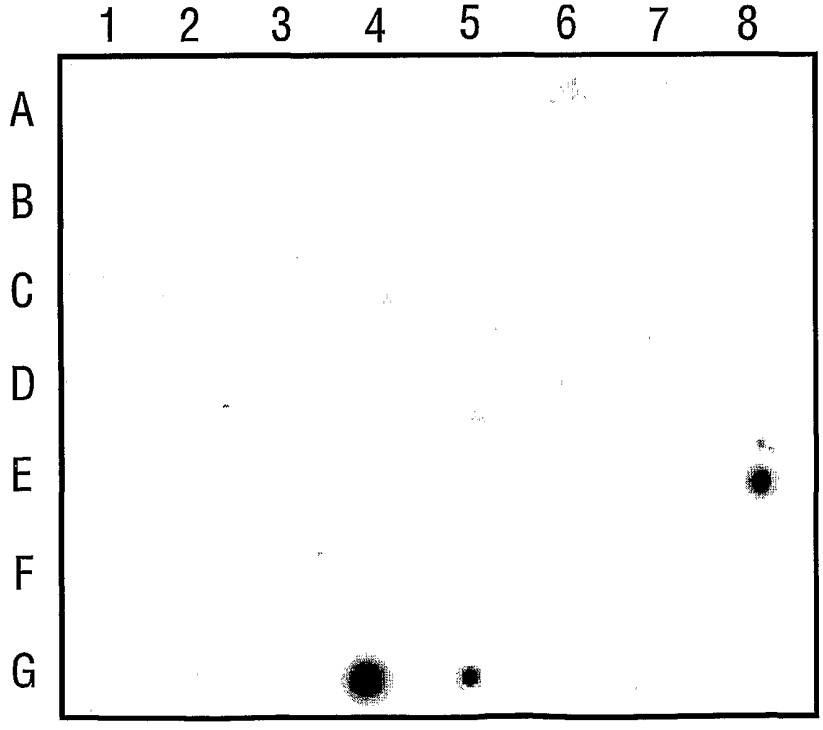


Fig. 7E

brain	amygdala	caudate nucleus	cere-bellum	cerebral cortex	frontal lobe	hippo-campus	medulla oblongata
occipital lobe	putamen	substantia nigra	temporal lobe	thalamus	Sub-thalamic nucleus	spinal cord	
heart	aorta	skeletal muscle	colon	bladder	uterus	prostate	stomach
testis	ovary	pancreas	pituitary gland	adrenal gland	thyroid gland	salivary gland	mammary gland
kidney	liver	small intestine	spleen	thymus	peripheral leukocyte	lymph node	bone marrow
appendix	lung	trachea	placenta				
fetal brain	fetal heart	fetal kidney	fetal liver	fetal spleen	fetal thymus	fetal lung	

Fig. 8A

```
TGX  SDERDDITENYKYPE GSLQERQVFLKALQKLKARSFHGSQRGAELQPSRPTSLSQDSPRSLHTPSLRPSDV  503 d3
TGZ  SDQRQSITSSYKYPE GSPEKRAVFMKASRKMLGPQRASLPFLDLLESGGLRD----------------------  487
B4.2 SDRCEDITQNYKYPE GSLQEKEVLENVEKEKMERKDNGIRPPSLETA-----------------------------  471
TGY  SDSRVDITDLYKYPE GSRKERQVYSKAVNRLFGVEASGRRIWIRRAGGRCLWRDDLLEPATK-------------  494
TGE  SNARMDVTDKYKYPE GSDQERQVFQKALGKLKPNTFPAATSSMGLETEEQE-----------------------  479
TGC  RDEREDITHTYKYPE GSSEEREAFTRANHLNKLAEKEE-------------------------------------  469
FXIII GDGMMDITDTYKFQE GQEEERLALETALMVGAKKPLNTEGVMKSR-------------------------------  515
TGK  SNMREDITYLYKNPE GSDAERKAVETAAAHGSKPNVYANRGSA-----------------------------------  573
TGP  QDRRRDITVEYKYPE GSSEERQVMDHAFLLLSSREHHRRPVKE-----------------------------------  469

TGX  VQVSLKFKLLDPPNMGQDICVLLALNMSSQFKD-----LKVNLSAQSLHDGSPLSPF-WQDTAFITLSPKE AKTYPCK--ISYSQVSQVLSTDKLIRISALGEEKSSPEKILVNKIITLS  617 d4
TGZ  QPAQLQLHLARIPEWGQDLQLLLRIQRVPDSTHPRGPIGLVVRFCAQALLHGGGTQKPF-WRHTVRMNLDFGK ETQWPLL--LPYSNYRNKLTDEKLIRVSGIAEVEETGRSMLVLKDICLE  606
B4.2 SPIKLLLKAPSSLPLRGDAQISVTLVNHSEQEKA-----VQLAIGVQAVHYNGVLAAKL-WRKKLHLTLSANL EKLITIG--LFFSNFERNPPENTFLRLTAMATHSESNLSCFAQEDIAIC  585
TGY  PSIAGKFKVLEPPMLGHDLRLAICAMLTSRAQR-----VRVNLSGATILYTRKPVAEI-LHESHAVRLGPQE EKRIPIT--ISYSKYKEDLTEDKKILLAAMCLVTK-GEKLLVEKDITLE  608
TGE  PSISGKLKVAGMLAVGKEVNLVLLLKNLGRDTKT-----VTVNMTAWTIIYNGTLVHEV-WKDSATMSLDPEE EAEHPIK--ISYAQVERVLKSDNMIRITAVCKVPD-ESEVVVERDIILD  592
TGC  TGMAMRIRVGQSMNMGSDFDVFAHITNNTAEEYV-----CRLLLCARTVSYNGILGPEGGTKYLLNLTLEPFS EKSVPLC--ILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDIYLE  584
FXIII SNVDMDFEV-ENAVLGKDFKLSITFRNNSHNRYT-----ITAYLSANITFTGVPKAEF-KKETFDVTLEPLS FKKEAVL--IQAGEYMGQLLEQASLHFTVTARINETRDVLAKQKSTVLT  628
TGK  EDVAMQVEA-QDAVMGQDLMVSVMLINHSSSRRT-----VKLHLYLSVTFVTGVSGTIF-KETKKSVELAPGA SDRVTMP--VAYKEYRPHLVDQGAMLINVSGHVKESGQVLAKQHTFRLR  686
TGP  NFLHMSVQS-DDVLLGRNSVNFTVILKRKTAALQN-----VNILGSFELQLYTGKKMAKL-CDLNKTSQIQGQ VSEVTLTLDSKTYINSLAILDDEPVIRGFIIAEIVESKEIMASEVFTSFQ  584

TGX  YPSITIN VLGAAVVNQPLSIQVIFSNPLSEQVEDCVLTVEGSGLFKKQDKVF LGV-LKPQHQASIILETVDPKSGQRQIQANMRSNKFKDIKGYRNVYVDPAL  719 d5
TGZ  PPHLSIE VSERAEVGKALRVHVTLTNTLMVALSSCTMVLEGSGLINGQIAKD LGT-LVAGHTLQIQLDLYPTKAGPRQLQVLISSNEVKEIKGYKDIPVTVAGAP  710
B4.2 RPHLAIK MPEKAEQYQPLTASVSLQNSLDAPMEDCVISILGRGLIHRERSYR FRS-VWPENTMCAKFQFTPTHVGLQRITVEVDCNMFQNLTNYKSVTVVAPELSA  690
TGY  -DFTTIK VLGPAMVGVAVTVEVTVVNPLIERVKDCALMVEGSGLLQEQLSID VPT-LEPQERASVQFDITPSKSGPRQLQVDIVSPHFPDIKGFVIVHVATAK  709
TGE  NPTLTIE VLNEARVRKPVNVQMLFSNPLDEPVRDCVLMVEGSGLLLGNLKID VPT-LGPKERSRVRFDILPSRSGTKQLLADFSCNKFPAIKAMLSIDVAE  692
TGC  NPEIKIR ILGEPKQKRKLVAEVSLQNPLPVALEGCTFTVEGAGLPEEQKTVEI PDPVEAGEEVKVRMDLVPLHMGLHKLVVNFESDKLKAVKGFRNVIIGPA  686
FXIII IPEILIK VRGTQVVGSDMTVTQETNPLKETLRNVWHLDGPGVT-RPMKKMFR E-IRPNSTQWEEVCRPWVSGHRKLIASMSDSLRHVYGELDVQIQRRPSM  731
TGK  TPDLSLT LLGAAVVGQECEVQVVFKNPLPVTLFNVVFRLEGSGLQ-RPKILNVG D-IGGNETVTLRQSFVPVRPGPRQLIASLDSPQLSQVHGVIQVDVAPAPGDGGFFSDAGGDSHLGRTIPMASRGGA  814
TGP  YPEFSIE LPNTGRIGQLLVCNCIFKNTLATPLTDVKFSLESLGIS-SLQTSDHG T-VQPGETIQSQIKCTPIKTGPKKFIVKLSSKQVKEINAQKIVLITK  683
```

Fig. 8A (contd.)

Fig. 8B


FXIII
TGK
TGX
TGZ
B4.2
TGY
TGE
TGC
FXIII
TGK
TGP
TGX
TGZ
B4.2
TGY
TGE
TGC
FXIII
TGK
TGP

```
TGX  SDERDDITEWVKYEE GSLQEEQVFLKALQKLKARSFHGSQRGAELQPSRPTSLSQDSPRSLHTPSLRPSDV                503  d3
TGZ  SDQRQSITSSYKYPE GSPEEEAVFMKASRKMLGPQRASLPFLDLLESGGLRD-------------------------   487
B4.2 SDRCEDITQWVKYPE GSLQEKEVLERVEKEKMERERKDVGIRPPSLRTA-----------------------------   471
TGY  SDSRVDITDLVKYPE GSRKEERQVYSKAVRHLFGVEASGRRIWIRRAGGRCLWRDDLLEPATK---------------   494
TGE  SNARMDVTDKVKYPE GSDQERQVFQKALGKLKPNTPFAATSSMGLETEEQR-------------------------   479
TGC  RDEREDITHTVKYPE GSSEEEEAFTRANHLNKLAEKEE-------------------------------------   469
FXIII GDGMMDITDTYKFQE GQEEERLALETALMYGAKKPLNTEGVMKSR-------------------------------   515
TGK  SNMREDITYLVKHPE GSDAERKAVETAAAHGSKPNVYANRGSA-----------------------------------   573
TGP  QDRRRDITYEVKYPE GSSEERQVMDHAFLLLSSEREHRRPVKR-----------------------------------   469

TGX  VQVSLKFKLLDPPNMGQDICFVLLALNMSSQFKD-----LKVNLSAQSLLHDGSPLSPF-WQDTAFTPLSPKE AKVYPCK--ISYSQVSQYLSTDKLIRISALGEEKSSPEKILVNKLITLS   617  d4
TGZ  QPAQLQLHLARIPEWGQDLQLLIRIQRVPDSTHPRGPIGLVVRPCAQALLHGGGTQKPF-WRHFTVRMNLDFGK ETQWPLL--LPYSNYRNKLTDEKLIRVSGIAEVEETGRSMLVLKDICLE   606
B4.2 SPLYLLLKAPSSLPLRGDAQISVTLVNHSEQEKA-----VQLAIGVQAVHYNGVLAAKL-WRKKLHLTLSANL EKLITIG--LPFSNFERNPPENTFLRLTAMATHSESNLSCFAQSDIAIC   585
TGY  PSIAGKFKVLEPPMLGHDLRLAICLANLTSRAQR-----VKVNLSGATILYTRKPVAEI-LHESHAVRLGPQE EKNIPIT--ISYSKYKEDLTEDKKILDAAMCLVTK-GEKLLVEKDITLE   608
TGE  PSISGKLKVAGMLAVGKEVNLVLLLKNLSRDTKT-----VTVNMTAWTIIYNGTLVHEV-WKDSATMSLDPEE EAEHPIK--ISYAQYERYIKSDNMIRITAVCKVPD-ESEVVVERDLILD   592
TGC  TGMAMRIRVGQSMNMGSDFDVFAHITNNTAEEYV-----CRLLLCARTVSYNGILGPECGTKYLLNLTLPPFS EKSVPLC--ILYEKYRDCLTESNLIKVRALLVEPVINSYILAERDLYLE   584
FXIII SNVDMDFEV-ENAVLGKDFKLSITFRNNSHNRYT-----ITAYLSANITFYTGVPKAEF-KKETFDVTLPLS FKKEAVL--IQAGEYMGQLLEQASLHFFVTARINETRDVLAKQKSTVLI   628
TGK  EDVAMQVEA-QDAVMGQDLMVSVMLINHSSSRRT-----VKLHLYLSVTFYTGVSGTIF-KETKKEVELAPGA SDRVTMP--VAYKEYRPHLVDQGAMLLNVSGHVKESGQVLAKQHTFRLR   686
TGP  NFLHMSVQS-DDVLLGNSVNFTVILKRKTAALQN-----VNILGSFELQLYTGKKMAKL-CDLNKTSQIQGQ VSEVTLTLDSKTYINSLAILDDEPVIRGFIAEIVESKEIMASEVFTSFQ   584

TGX  YPSITTIN VLGAAVVNQPLSIQVIFSNPLSEQVEDCVLTVEGSGLFKKQQKVF LGV-LKPQHQASIILETVPFKSGRQIQANMRSNKFKDIKGYRNVYVDFAL                   719  d5
TGZ  PPHLSIE VSERAEVGKAIRVHVTITNTIMVALSSCTMVLEGSGLINGQIAKD LGT-IVAGHTLQIQLDLYPTKAGPRQLQVLISSNEVKEIKGYKDIFVTVAGAP                   710
B4.2 RPHLAIK MPEKAEQYQPLTASVSLQNSLDAPMEDCVISLIGRGLIHRERSYR FRS-VWPENTMCAKFQFTPTHVGLQRLTVEVDCNMFQNLTNYKSVTVVAPELSA                   690
TGY  -DFTTIK VLGPAMVGVAVTVEVTVNPLIERVKDCALMVEGSGLLQEQLSID VPT-LEPQERASVQFDITPSKSGPRQLQVDLVSPHFPDIKGFVIVHVATAK                   709
TGE  NPTLTLE VLNEARVRKPVNVQMLFSNPLDEPVRDCVLMVEGSGLLLGNLKID VPT-LGPKERSRVRFDILPSRSGTKQLLADFSCNKFPAIKAMLSIDVAE                   692
TGC  NPEIKIR ILGEPKQRKRIVAEVSLQNPLPVALEGCTFTVEGAGLTEEQKTVEI PDPVEAGEEVKVRMDLVPLHMGLHKLVVNFEGDKLKAVKGFRNVIIGPA                   686
FXIII IPEIIIK VRGPQVVGSDMTVTQFTNPLKETLRNVWVHLDGPGVT-RPMKKMFR E-IRPNSTVQWEEVCRPWVSGHRKLIASMSSDSLRHVVGELDYQIQRRPSM                   731
TGK  TPDLSLT LLGAAVVGQECEVQIVFKNPLPVTLTNVVFRLEGSGLQ-RPKILNVG D-IGGNETVTLRQSFVPVRPGPRQLIASLDSPQLSQVHGVIQVDVAPAPGDGGFFSDAGGDSHLGETIPMASRGGA   814
TGP  YPEFSIE LPNTGRIGQLLVCNCIFKNTLAIPLTDVKFSLESLGIS-SLQTSDHG T-VQPGETIQSQIKCTPIKTGPKKFIVKLSSKQVKEINAQKIVLITK                   683
```

Fig. 8B (contd.)

Fig. 9A(1)

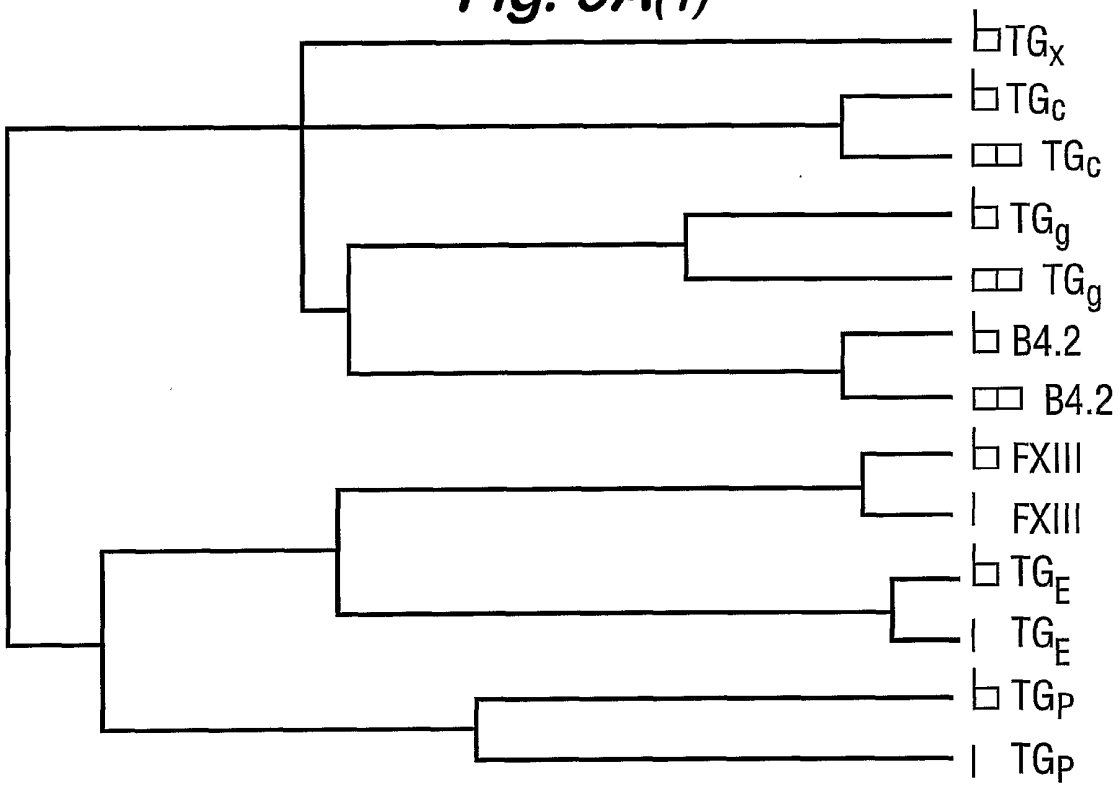


Fig. 9A(2)

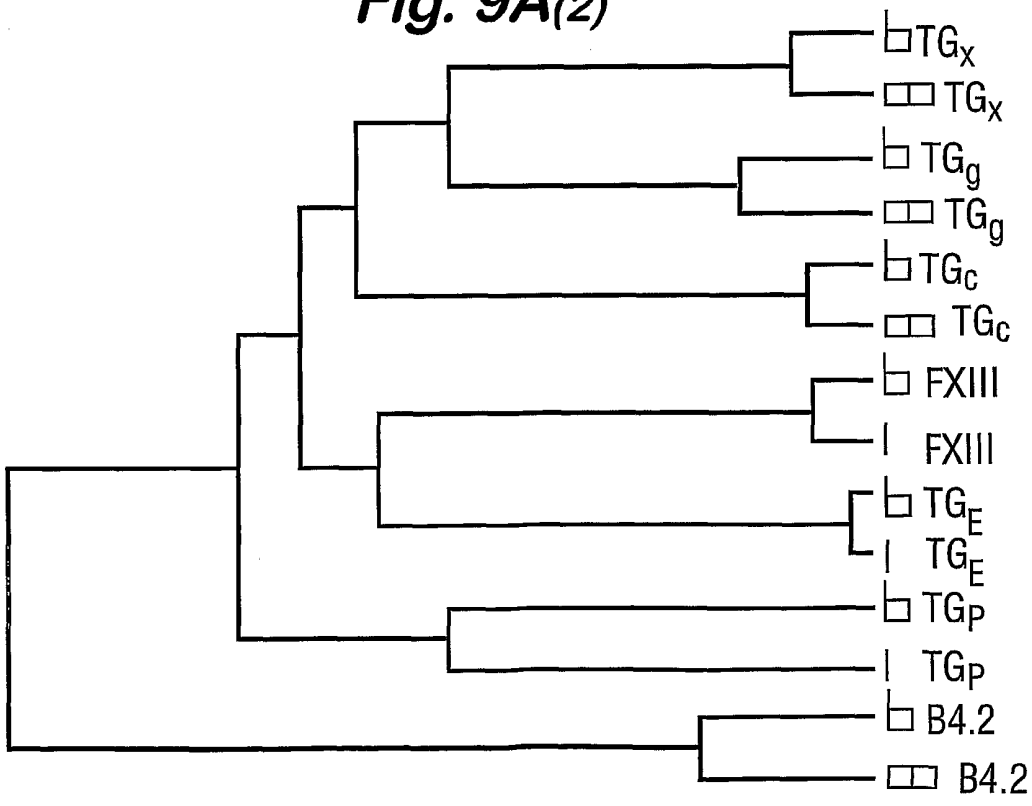


Fig. 9A(3)

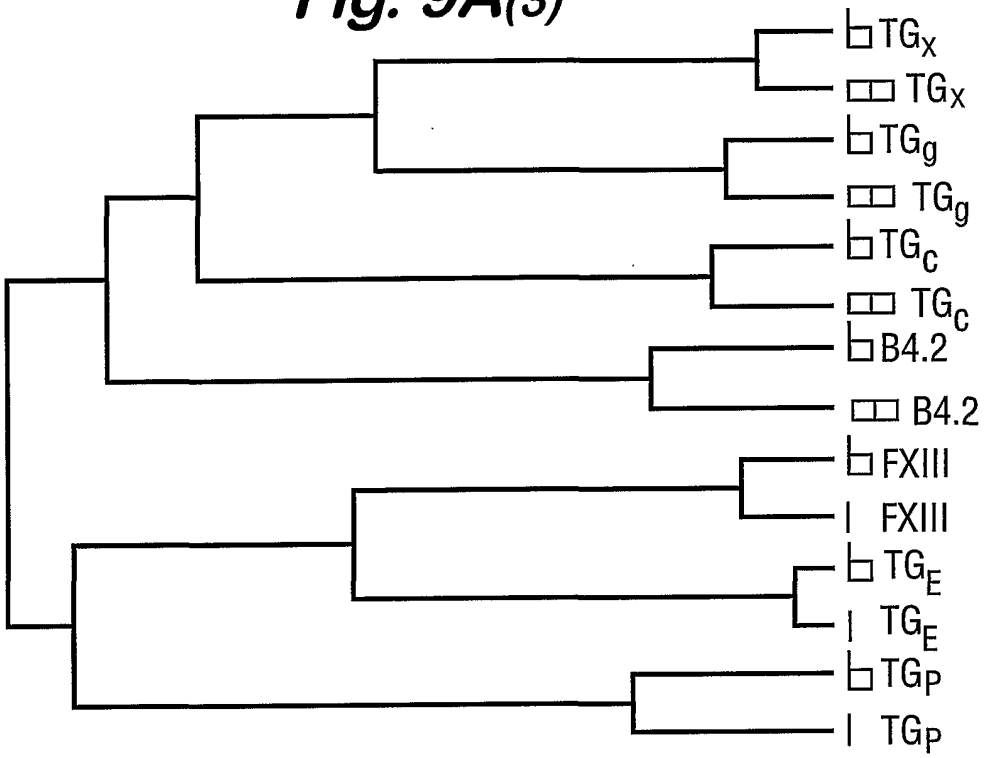
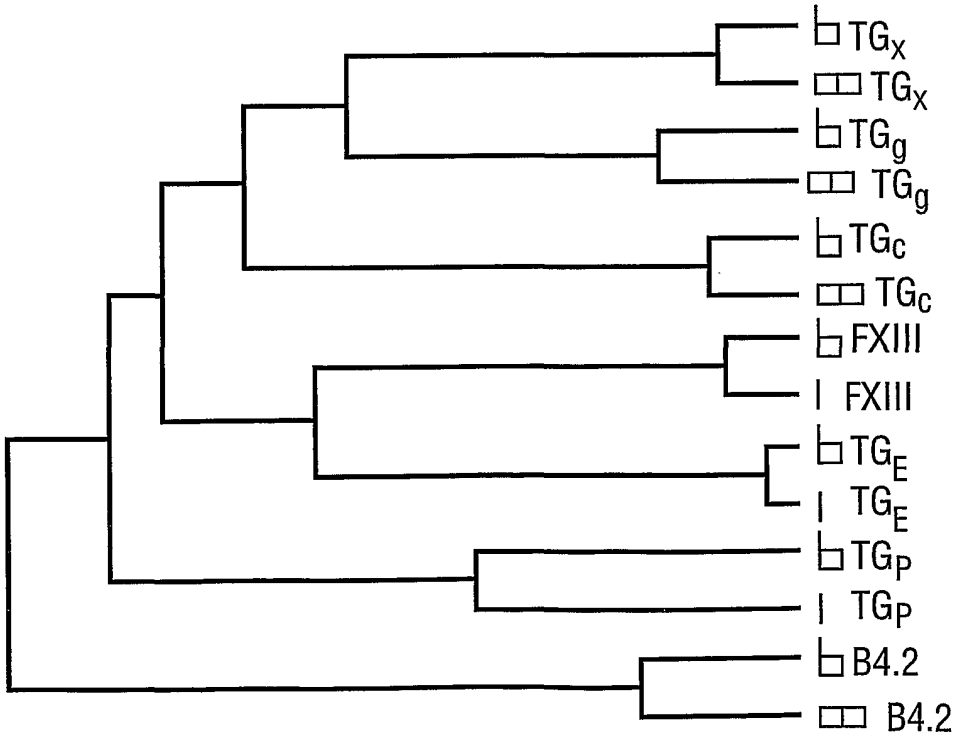


Fig. 9A(4)



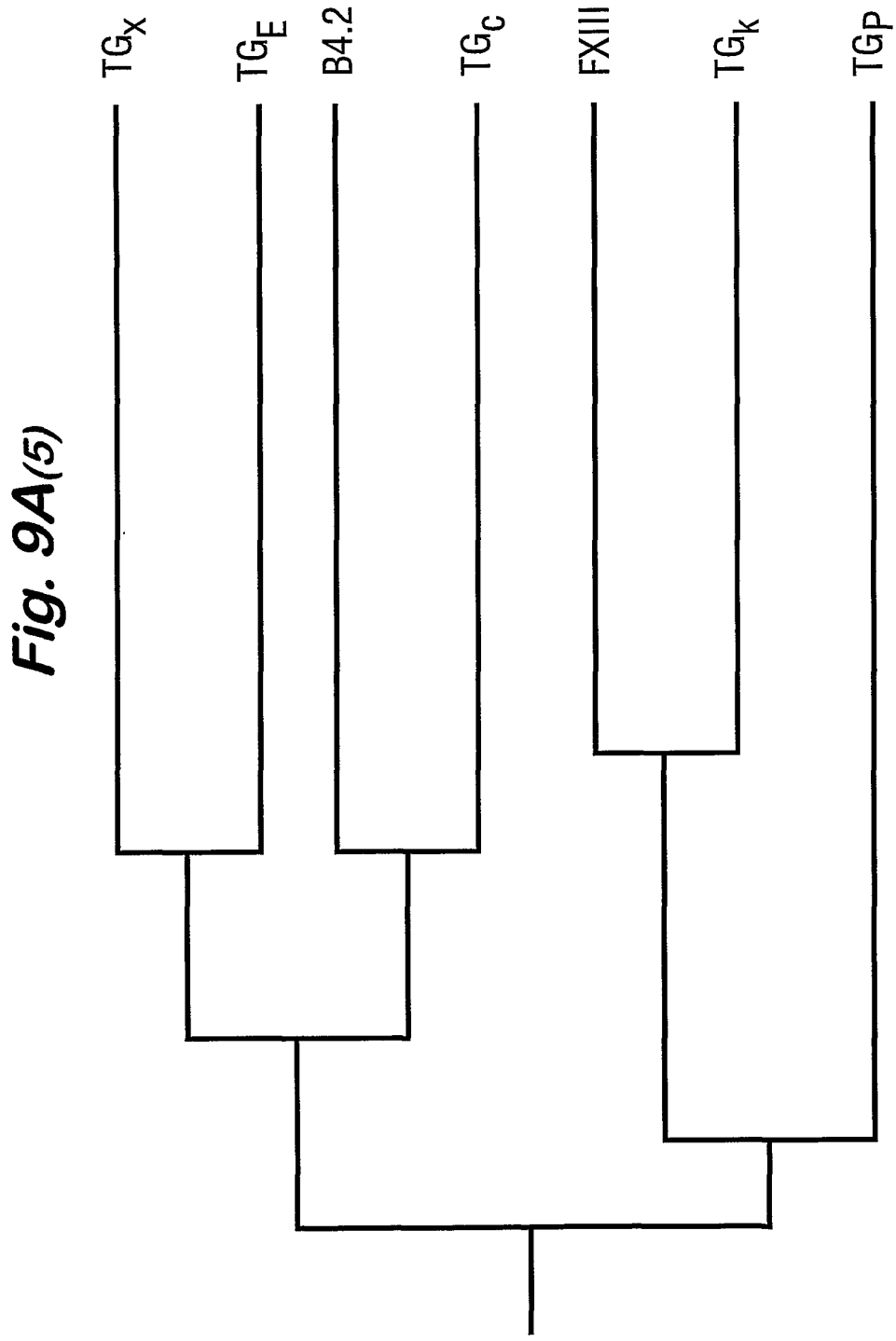


Fig. 9A(5)

Fig. 9B

	TGM5 (TG_X)	TGM7 (TG_Z)	EPB42 (band 4.2 protein)	TGM2 (TG_C)	TGM6 (TG_Y)	TGM3 (TG_E)	TGM4 (TG_P)	F13A1 (factor XIII a-subunit)	TGM1 (TG_K)
Chromosomal localization									
human	15q15 2, 67.5 cM	15q15	15q15 (a) 2, 67.5cM (g)	20q11-12(b) 2, 89cM	20q11(c)	20q11 (c)	3p21-22 (d)	6p24-25 (e)	14q11.2 (f)
mouse									
Gene size									
human	~35kb	~26kb	~20kb (h) ~22kb (m)	~37kb (b) ~34kb (o)	~45kb	~43kb (i)	~35kb (n)	~160kb (k)	~14kb (l)
mouse									
Number of exons									
human	13	13	13 13	13 13	13	13	13	15	15
mouse									

Fig. 9C

Gene Product	Overall	Protein Domains			
		β-sandwich	catalytic core	β-barrel 1	β-barrel 2
TG_X	100.0				
TG_Z	48.1	49.6	63.1	31.6	36.3
B4.2	31.6	33.1	41.1	16.7	23.5
TG_Y	45.2	41.0	56.6	34.2	44.1
TG_E	42.3	29.5	56.5	31.6	38.2
TG_C	40.1	35.3	55.6	20.2	31.4
FXIIIa	32.7	23.7	46.8	18.4	25.5
TG_K	34.9	25.9	49.6	18.4	29.4
TG_P	31.0	23.7	47.1	11.4	20.6

Fig. 10A

```
                                                                      5
                                                                   AA CCC

          14          23          32          41          50          59
ATG ACC CAG GGG ATC AGA GTC ACC AAG GTG GAC TGG CAG CGG TCG AGG AAT GGC
 M   T   Q   G   I   R   V   T   K   V   D   W   Q   R   S   R   N   G

          68          77          86          95         104         113
GCT GCC CAC CAC ACC CAG GAG TAC CCC TGC CCT GAG CTG GTG GTT CGC AGG GGC
 A   A   H   H   T   Q   E   Y   P   C   P   E   L   V   V   R   R   G

         122         131         140         149         158         167
CAG TCG TTC AGC CTC ACG CTG GAG CTG AGC AGA GCC CTG GAC TGT GAG GAG ATC
 Q   S   F   S   L   T   L   E   L   S   R   A   L   D   C   E   E   I

         176         185         194         203         212         221
CTC ATC TTC ACG GTG GAG ACA GGA CCC CGG GCT TCT GAG GCC CTC CAC ACC AAA
 L   I   F   T   V   E   T   G   P   R   A   S   E   A   L   H   T   K

         230         239         248         257         266         275
GCT GTG TTC CAG ACA TCG GAG CTG GAG CGG GGT GAG GGC TGG ACA GCA GCA AGG
 A   V   F   Q   T   S   E   L   E   R   G   E   G   W   T   A   A   R

         284         293         302         311         320         329
GAG GCT CAG ATG GAG AAA ACT CTG ACC GTC AGT CTC GCC AGC CCT CCC AGT GCT
 E   A   Q   M   E   K   T   L   T   V   S   L   A   S   P   P   S   A

         338         347         356         365         374         383
GTC ATT GGC CGC TAC CTG CTG AGC ATC AGG CTT TCC TCT CAC CGC AAA CAC AGC
 V   I   G   R   Y   L   L   S   I   R   L   S   S   H   R   K   H   S

         392         401         410         419         428         437
AAC CGG AGG CTG GGC GAG TTT GTT CTC CTT TTC AAC CCA TGG TGT GCA GAG GAC
 N   R   R   L   G   E   F   V   L   L   F   N   P   W   C   A   E   D

         446         455         464         473         482         491
GAT GTG TTT CTG GCC TCA GAG GAG GAG AGA CAG GAG TAC GTG CTC AGC GAC AGC
 D   V   F   L   A   S   E   E   E   R   Q   E   Y   V   L   S   D   S

         500         509         518         527         536         545
GGC ATC ATC TTC CGA GGC GTG GAG AAG CAC ATA CGA GCC CAG GGC TGG AAC TAC
 G   I   I   F   R   G   V   E   K   H   I   R   A   Q   G   W   N   Y

         554         563         572         581         590         599
GGG CAG TTT GAG GAG GAC ATC CTG AAC ATC TGC CTC TCC ATC CTG GAT CGA AGC
 G   Q   F   E   E   D   I   L   N   I   C   L   S   I   L   D   R   S

         608         617         626         635         644         653
CCC GGT CAC CAA AAC AAC CCA GCC ACC GAC GTG TCC TGC CGC CAC AAC CCC ATC
 P   G   H   Q   N   N   P   A   T   D   V   S   C   R   H   N   P   I

         662         671         680         689         698         707
TAC GTC ACC AGG GTC ATC AGT GCC ATG GTG AAC AGC AAC AAC GAC CGA GGT GTG
 Y   V   T   R   V   I   S   A   M   V   N   S   N   N   D   R   G   V

         716         725         734         743         752         761
GTG CAA GGA CAG TGG CAG GGC AAG TAC GGC GGC GGC ACC AGC CCG CTG CAC TGG
 V   Q   G   Q   W   Q   G   K   Y   G   G   G   T   S   P   L   H   W

         770         779         788         797         806         815
CGC GGC AGC GTG GCC ATT CTG CAG AAG TGG CTC AAG GGC AGG TAC AAG CCA GTC
 R   G   S   V   A   I   L   Q   K   W   L   K   G   R   Y   K   P   V
```

Fig. 10A(contd)

```
          824        833        842        851        860        869
AAG TAC GGC CAG TGC TGG GTC TTC GCC GGA GTC CTG TGC ACA GTC CTC AGG TGC
 K   Y   G   Q   C   W   V   F   A   G   V   L   C   T   V   L   R   C

          878        887        896        905        914        923
TTG GGG ATA GCC ACA CGG GTC GTG TCC AAC TTC AAC TCA GCC CAC GAC ACA GAC
 L   G   I   A   T   R   V   V   S   N   F   N   S   A   H   D   T   D

          932        941        950        959        968        977
CAG AAC CTG AGT GTG GAC AAA TAC GTG GAC TCC TTC GGG CGG ACC CTG GAG GAC
 Q   N   L   S   V   D   K   Y   V   D   S   F   G   R   T   L   E   D

          986        995       1004       1013       1022       1031
CTG ACA GAA GAC AGC ATG TGG AAT TTC CAT GTC TGG AAT GAG AGC TGG TTT GCC
 L   T   E   D   S   M   W   N   F   H   V   W   N   E   S   W   F   A

         1040       1049       1058       1067       1076       1085
CGG CAG GAC CTA GGC CCC TCT TAC AAT GGC TGG CAG GTT CTG GAT GCC ACC CCC
 R   Q   D   L   G   P   S   Y   N   G   W   Q   V   L   D   A   T   P

         1094       1103       1112       1121       1130       1139
CAG GAG GAG AGT GAA GGT GTG TTC CGG TGC GGC CCA GCC TCA GTC ACC GCC ATC
 Q   E   E   S   E   G   V   F   R   C   G   P   A   S   V   T   A   I

         1148       1157       1166       1175       1184       1193
CGC GAG GGT GAT GTG CAC CTG GCT CAC GAT GGC CCC TTC GTG TTT GCG GAG GTC
 R   E   G   D   V   H   L   A   H   D   G   P   F   V   F   A   E   V

         1202       1211       1220       1229       1238       1247
AAC GCC GAC TAC ATC ACC TGG CTG TGG CAC GAG GAT GAG AGC CGG GAG CGT GTA
 N   A   D   Y   I   T   W   L   W   H   E   D   E   S   R   E   R   V

         1256       1265       1274       1283       1292       1301
TAC TCA AAC ACG AAG AAG ATT GGG AGA TGC ATC AGC ACC AAG GCG GTG GGC AGT
 Y   S   N   T   K   K   I   G   R   C   I   S   T   K   A   V   G   S

         1310       1319       1328       1337       1346       1355
GAC TCC CGC GTG GAC ATC ACT GAC CTC TAC AAG TAT CCG GAA GGG TCC CGG AAA
 D   S   R   V   D   I   T   D   L   Y   K   Y   P   E   G   S   R   K

         1364       1373       1382       1391       1400       1409
GAG AGG CAG GTG TAC AGC AAG GCG GTG AAC AGG CTG TTC GGC GTG GAA GCC TCT
 E   R   Q   V   Y   S   K   A   V   N   R   L   F   G   V   E   A   S

         1418       1427       1436       1445       1454       1463
GGA AGG AGA ATC TGG ATC CGC AGG GCT GGG GGT CGC TGT CTC TGG CGT GAC GAC
 G   R   R   I   W   I   R   R   A   G   G   R   C   L   W   R   D   D

         1472       1481       1490       1499       1508       1517
CTC CTG GAG CCT GCC ACC AAG CCC AGC ATC GCT GGC AAG TTC AAG GTG CTA GAG
 L   L   E   P   A   T   K   P   S   I   A   G   K   F   K   V   L   E

         1526       1535       1544       1553       1562       1571
CCT CCC ATG CTG GGC CAC GAC CTG AGA CTG GCC CTG TGC TTG GCC AAC CTC ACC
 P   P   M   L   G   H   D   L   R   L   A   L   C   L   A   N   L   T
```

```
      1580        1589        1598        1607        1616        1625
TCC CGG GCC CAG CGG GTG AGG GTC AAC CTG AGC GGT GCC ACC ATC CTC TAT ACC
 S   R   A   Q   R   V   R   V   N   L   S   G   A   T   I   L   Y   T

      1634        1643        1652        1661        1670        1679
CGC AAG CCA GTG GCA GAG ATC CTG CAT GAA TCC CAC GCC GTG AGG CTG GGG CCG
 R   K   P   V   A   E   I   L   H   E   S   H   A   V   R   L   G   P

      1688        1697        1706        1715        1724        1733
CAA GAA GAG AAG AGA ATC CCA ATT ACA ATA TCT TAC TCT AAG TAT AAA GAA GAC
 Q   E   E   K   R   I   P   I   T   I   S   Y   S   K   Y   K   E   D

      1742        1751        1760        1769        1778        1787
CTG ACA GAG GAC AAG AAG ATC CTG TTG GCT GCC ATG TGC CTT GTC ACC AAA GGA
 L   T   E   D   K   K   I   L   L   A   A   M   C   L   V   T   K   G

      1796        1805        1814        1823        1832        1841
GAG AAG CTT CTG GTG GAG AAG GAC ATT ACT CTA GAG GAC TTC ATC ACC ATC AAG
 E   K   L   L   V   E   K   D   I   T   L   E   D   F   I   T   I   K

      1850        1859        1868        1877        1886        1895
GTT CTG GGC CCA GCC ATG GTG GGA GTG GCA GTT ACA GTG GAA GTG ACA GTA GTC
 V   L   G   P   A   M   V   G   V   A   V   T   V   E   V   T   V   V

      1904        1913        1922        1931        1940        1949
AAC CCC CTC ATA GAG AGA GTG AAG GAC TGT GCG CTG ATG GTG GAG GGC AGC GGC
 N   P   L   I   E   R   V   K   D   C   A   L   M   V   E   G   S   G

      1958        1967        1976        1985        1994        2003
CTT CTC CAG GAA CAG CTC AGC ATC GAC GTG CCT ACC CTG GAG CCT CAG GAG AGG
 L   L   Q   E   Q   L   S   I   D   V   P   T   L   E   P   Q   E   R

      2012        2021        2030        2039        2048        2057
GCC TCA GTC CAG TTT GAC ATC ACC CCC TCC AAA AGT GGC CCA AGG CAG CTG CAG
 A   S   V   Q   F   D   I   T   P   S   K   S   G   P   R   Q   L   Q

      2066        2075        2084        2093        2102        2111
GTG GAC CTT GTA AGC CCT CAC TTC CCG GAC ATC AAG GGC TTT GTG ATC GTC CAT
 V   D   L   V   S   P   H   F   P   D   I   K   G   F   V   I   V   H

      2120        2129        2138        2147        2156        2165
GTG GCC ACT GCC AAG TGA TGG ATC ATG AGG GAC TGA GAG GGG TGG ATT TGG CCC
 V   A   T   A   K   *

      2174        2183        2192        2201        2210        2219
CTG TCC TCC TCC TGC CCA TTC TTT GTC TCT TCC ACA TGG GAG CCA GGA GGC CTC

      2228        2237
AGT TAA TCC TGC CTC AAC CT
```

Fig. 10A(contd)

Fig. 10B

```
                                                                      5
                                                                   AA CCC
          14          23          32          41          50          59
ATG ACC CAG GGG ATC AGA GTC ACC AAG GTG GAC TGG CAG CGG TCG AGG AAT GGC
 M   T   Q   G   I   R   V   T   K   V   D   W   Q   R   S   R   N   G

          68          77          86          95         104         113
GCT GCC CAC CAC ACC CAG GAG TAC CCC TGC CCT GAG CTG GTG GTT CGC AGG GGC
 A   A   H   H   T   Q   E   Y   P   C   P   E   L   V   V   R   R   G

         122         131         140         149         158         167
CAG TCG TTC AGC CTC ACG CTG GAG CTG AGC AGA GCC CTG GAC TGT GAG GAG ATC
 Q   S   F   S   L   T   L   E   L   S   R   A   L   D   C   E   E   I

         176         185         194         203         212         221
CTC ATC TTC ACG GTG GAG ACA GGA CCC CGG GCT TCT GAG GCC CTC CAC ACC AAA
 L   I   F   T   V   E   T   G   P   R   A   S   E   A   L   H   T   K

         230         239         248         257         266         275
GCT GTG TTC CAG ACA TCG GAG CTG GAG CGG GGT GAG GGC TGG ACA GCA GCA AGG
 A   V   F   Q   T   S   E   L   E   R   G   E   G   W   T   A   A   R

         284         293         302         311         320         329
GAG GCT CAG ATG GAG AAA ACT CTG ACC GTC AGT CTC GCC AGC CCT CCC AGT GCT
 E   A   Q   M   E   K   T   L   T   V   S   L   A   S   P   P   S   A

         338         347         356         365         374         383
GTC ATT GGC CGC TAC CTG CTG AGC ATC AGG CTT TCC TCT CAC CGC AAA CAC AGC
 V   I   G   R   Y   L   L   S   I   R   L   S   S   H   R   K   H   S

         392         401         410         419         428         437
AAC CGG AGG CTG GGC GAG TTT GTT CTC CTT TTC AAC CCA TGG TGT GCA GAG GAC
 N   R   R   L   G   E   F   V   L   L   F   N   P   W   C   A   E   D

         446         455         464         473         482         491
GAT GTG TTT CTG GCC TCA GAG GAG GAG AGA CAG GAG TAC GTG CTC AGC GAC AGC
 D   V   F   L   A   S   E   E   E   R   Q   E   Y   V   L   S   D   S

         500         509         518         527         536         545
GGC ATC ATC TTC CGA GGC GTG GAG AAG CAC ATA CGA GCC CAG GGC TGG AAC TAC
 G   I   I   F   R   G   V   E   K   H   I   R   A   Q   G   W   N   Y

         554         563         572         581         590         599
GGG CAG TTT GAG GAG GAC ATC CTG AAC ATC TGC CTC TCC ATC CTG GAT CGA AGC
 G   Q   F   E   E   D   I   L   N   I   C   L   S   I   L   D   R   S

         608         617         626         635         644         653
CCC GGT CAC CAA AAC AAC CCA GCC ACC GAC GTG TCC TGC CGC CAC AAC CCC ATC
 P   G   H   Q   N   N   P   A   T   D   V   S   C   R   H   N   P   I

         662         671         680         689         698         707
TAC GTC ACC AGG GTC ATC AGT GCC ATG GTG AAC AGC AAC AAC GAC CGA GGT GTG
 Y   V   T   R   V   I   S   A   M   V   N   S   N   N   D   R   G   V

         716         725         734         743         752         761
GTG CAA GGA CAG TGG CAG GGC AAG TAC GGC GGC GGC ACC AGC CCG CTG CAC TGG
 V   Q   G   Q   W   Q   G   K   Y   G   G   G   T   S   P   L   H   W

         770         779         788         797         806         815
CGC GGC AGC GTG GCC ATT CTG CAG AAG TGG CTC AAG GGC AGG TAC AAG CCA GTC
 R   G   S   V   A   I   L   Q   K   W   L   K   G   R   Y   K   P   V
```

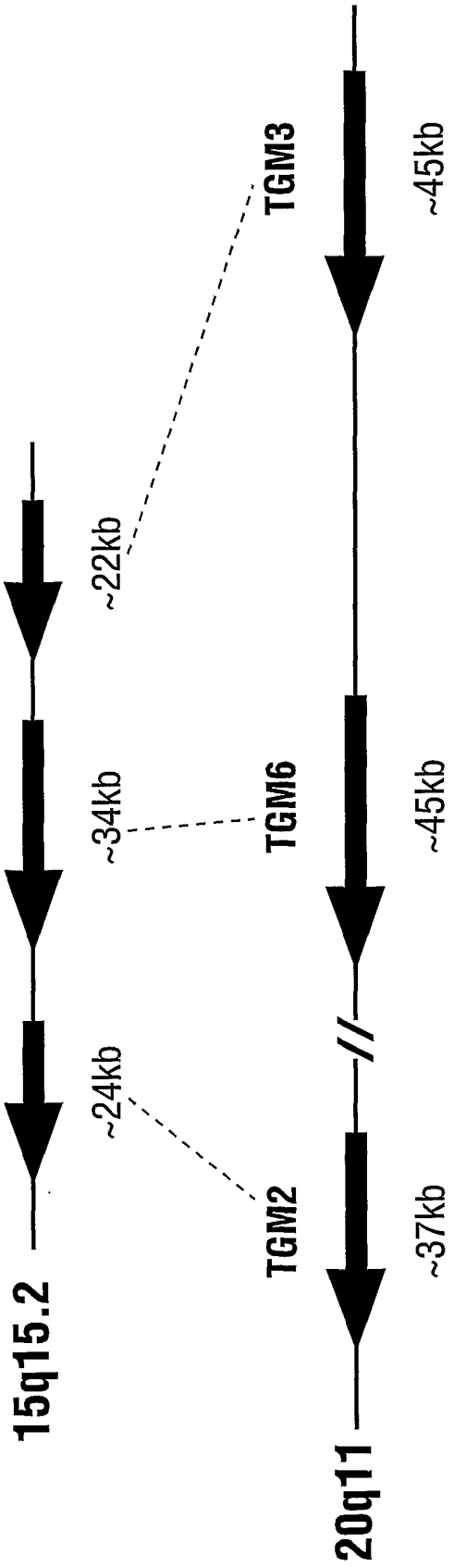
Fig. 10B(contd)

```
        824         833         842         851         860         869
AAG TAC GGC CAG TGC TGG GTC TTC GCC GGA GTC CTG TGC ACA GTC CTC AGG TGC
 K   Y   G   Q   C   W   V   F   A   G   V   L   C   T   V   L   R   C

        878         887         896         905         914         923
TTG GGG ATA GCC ACA CGG GTC GTG TCC AAC TTC AAC TCA GCC CAC GAC ACA GAC
 L   G   I   A   T   R   V   V   S   N   F   N   S   A   H   D   T   D

        932         941         950         959         968         977
CAG AAC CTG AGT GTG GAC AAA TAC GTG GAC TCC TTC GGG CGG ACC CTG GAG GAC
 Q   N   L   S   V   D   K   Y   V   D   S   F   G   R   T   L   E   D

        986         995        1004        1013        1022        1031
CTG ACA GAA GAC AGC ATG TGG AAT TTC CAT GTC TGG AAT GAG AGC TGG TTT GCC
 L   T   E   D   S   M   W   N   F   H   V   W   N   E   S   W   F   A

       1040        1049        1058        1067        1076        1085
CGG CAG GAC CTA GGC CCC TCT TAC AAT GGC TGG CAG GTT CTG GAT GCC ACC CCC
 R   Q   D   L   G   P   S   Y   N   G   W   Q   V   L   D   A   T   P

       1094        1103        1112        1121        1130        1139
CAG GAG GAG AGT GAA GGT GTG TTC CGG TGC GGC CCA GCC TCA GTC ACC GCC ATC
 Q   E   E   S   E   G   V   F   R   C   G   P   A   S   V   T   A   I

       1148        1157        1166        1175        1184        1193
CGC GAG GGT GAT GTG CAC CTG GCT CAC GAT GGC CCC TTC GTG TTT GCG GAG GTC
 R   E   G   D   V   H   L   A   H   D   G   P   F   V   F   A   E   V

       1202        1211        1220        1229        1238        1247
AAC GCC GAC TAC ATC ACC TGG CTG TGG CAC GAG GAT GAG AGC CGG GAG CGT GTA
 N   A   D   Y   I   T   W   L   W   H   E   D   E   S   R   E   R   V

       1256        1265        1274        1283        1292        1301
TAC TCA AAC ACG AAG AAG ATT GGG AGA TGC ATC AGC ACC AAG GCG GTG GGC AGT
 Y   S   N   T   K   K   I   G   R   C   I   S   T   K   A   V   G   S

       1310        1319        1328        1337        1346        1355
GAC TCC CGC GTG GAC ATC ACT GAC CTC TAC AAG TAT CCG GAA GGG TCC CGG AAA
 D   S   R   V   D   I   T   D   L   Y   K   Y   P   E   G   S   R   K

       1364        1373        1382        1391        1400        1409
GAG AGG CAG GTG TAC AGC AAG GCG GTG AAC AGG CTG TTC GGC GTG GAA GCC TCT
 E   R   Q   V   Y   S   K   A   V   N   R   L   F   G   V   E   A   S

       1418        1427        1436        1445        1454        1463
GGA AGG AGA ATC TGG ATC CGC AGG GCT GGG GGT CGC TGT CTC TGG CGT GAC GAC
 G   R   R   I   W   I   R   R   A   G   G   R   C   L   W   R   D   D

       1472        1481        1490        1499        1508        1517
CTC CTG GAG CCT GCC ACC AAG CCC AGC ATC GCT GGC AAG TTC AAG GTG CTA GAG
 L   L   E   P   A   T   K   P   S   I   A   G   K   F   K   V   L   E

       1526        1535        1544        1553        1562        1571
CCT CCC ATG CTG GGC CAC GAC CTG AGA CTG GCC CTG TGC TTG GCC AAC CTC ACC
 P   P   M   L   G   H   D   L   R   L   A   L   C   L   A   N   L   T
```

```
      1580        1589        1598        1607        1616        1625
TCC CGG GCC CAG CGG GTG AGG GTC AAC CTG AGC GGT GCC ACC ATC CTC TAT ACC
 S   R   A   Q   R   V   R   V   N   L   S   G   A   T   I   L   Y   T

      1634        1643        1652        1661        1670        1679
CGC AAG CCA GTG GCA GAG ATC CTG CAT GAA TCC CAC GCC GTG AGG CTG GGG CCG
 R   K   P   V   A   E   I   L   H   E   S   H   A   V   R   L   G   P

      1688        1697        1706        1715        1724        1733
CAA GAA GAG AAG AGA ATC CCA ATT ACA ATA TCT TAC TCT AAG TAT AAA GAA GAC
 Q   E   E   K   R   I   P   I   T   I   S   Y   S   K   Y   K   E   D

      1742        1751        1760        1769        1778        1787
CTG ACA GAG GAC AAG AAG ATC CTG TTG GCT GCC ATG TGC CTT GTC ACC AAA GGA
 L   T   E   D   K   K   I   L   L   A   A   M   C   L   V   T   K   G

      1796        1805        1814        1823        1832        1841
GAG AAG CTT CTG GTG GAG AAG GAC ATT ACT CTA GAG GAC TTC ATC ACC ATC AAG
 E   K   L   L   V   E   K   D   I   T   L   E   D   F   I   T   I   K

      1850        1859        1868        1877        1886        1895
CGT GCC TAC CCT GGA GCC TCA GGA GAG GGC CTC AGT CCA GTT TGA CAT CAC CCC
 R   A   Y   P   G   A   S   G   E   G   L   S   P   V   *

      1904        1913        1922        1931        1940        1949
CTC CAA AAG TGG CCC AAG GCA GCT GCA GGT GGA CCT TGT AAG CCC TCA CTT CCC

      1958        1967        1976        1985        1994        2003
GGA CAT CAA GGG CTT TGT GAT CGT CCA TGT GGC CAC TGC CAA GTG ATG GAT CAT

      2012        2021        2030        2039        2048        2057
GAG GGA CTG AGA GGG GTG GAT TTG GCC CCT GTC CTC CTC CTG CCC ATT CTT TGT

      2066        2075        2084        2093        2102        2105
CTC TTC CAC ATG GGA GCC AGG AGG CCT CAG TTA ATC CTG CCT CAA CCT
```

Fig. 10B(contd)

Fig. 11


TGM7
TGM5
EPB4.2
15q15.2
~22kb
~34kb
~24kb
TGM3
TGM6
TGM2
20q11
~45kb
~45kb
~37kb