Tenascin-X: A Novel Extracellular Matrix Protein Encoded by the Human XB Gene Overlapping P450c21B

James Bristow,* Meng Kian Tee,* Stephen E. Gitelman,* Synthia H. Mellon,‡§ and Walter L. Miller*§

*Departments of Pediatrics, and of ‡Obstetrics, Gynecology and Reproductive Sciences, and the §Metabolic Research Unit University of California, San Francisco, California

Abstract. A human gene termed XB overlaps the P450c21B gene encoding steroid 21-hydroxylase and encodes a protein that closely resembles extracellular matrix proteins. Sequencing of phage and cosmid clones and of cDNA fragments shows that the XB gene spans 65 kb of DNA, consisting of 39 exons that encode a 12-kb mRNA. The predicted protein of over 400 kD consists of five distinct domains: a signal peptide, a hydrophobic domain containing three heptad repeats, a series of 18.5 EGF-like repeats, 29 fibronectin type III repeats, and a carboxy-terminal fibrinogen-like domain. Because the structure of the protein encoded by the XB gene closely resembles tenascin, we term this protein tenascin-X (TN-X), and propose a simplified nomenclature system for the family of tenascins. RNase protection experiments show that the TN-X transcript is expressed ubiquitously in human fetal tissues, with the greatest expression in the fetal testis and in fetal skeletal, cardiac, and smooth muscle. Two adrenal-specific transcripts, P450c21B (steroid 21-hydroxylase) and Y (an untranslated transcript) overlap the XB gene on the complementary strand of DNA, yielding a unique array of overlapping transcripts: a "polygene." In situ hybridization histochemistry experiments show that the TN-X transcript and the P450c21 and Y transcripts encoded on the complementary DNA strand are all expressed in the same cells of the human adrenal cortex. Genetic data suggest that TN-X may be essential for life.

TENASCIN, a large glycoprotein of the extracellular matrix, has been studied extensively for about 10 years (for review see reference 11). Each of the six monomers that constitute tenascin exists in multiple size variants of 220-320 kD and contains four domains. The amino terminus comprises the hydrophobic "head group," which facilitates polymerization into the tenascin hexabrachion; this is followed in turn by a series of EGF-like repeats, a series of fibronectin type III (Fn III)[1] repeats, and a carboxy-terminal fibrinogen-like domain (11, 16, 22, 41, 52, 58). Tenascin appears to mediate interactions between cells and the extracellular matrix (and possibly between cells) (7, 47), through an RGD-dependent (3, 29) or -independent (46, 52) receptor. Initial studies suggested a crucial role in embryonic development (6, 53, 58), although recent tenascin gene knockout experiments in transgenic mice suggest that tenascin serves no mandatory, irreplaceable role in development or in the adult (48). Tenascin appears to be one member of a family of related proteins, encoded by related genes. Each family member has the same general structure as tenascin: NH2-head group-EGF domain-Fn III domain-fibrinogen domain-COOH. In addition to tenascin, a closely related protein termed restrictin (42) or J1-160/180 (13) has a similar structure. The structure of undulin, which contains a von Willebrand factor domain at the NH2 terminus followed by Fn III repeats (23), indicates that there are variations on this theme.

We now describe the gene structure and tissue distribution of another member of this family termed tenascin X (TN-X), which is encoded by an unusual gene termed XB in the human major histocompatibility locus. This gene was discovered because it overlaps the gene encoding steroid 21-hydroxylase (39), and was later shown to encode a tenascin-like sequence (14, 60; Bristow, J., S. E. Gitelman, Y. Shi, and W. L. Miller, 1990. *Pediatr. Res.* 27:76a). Recent preliminary reports from another group have confirmed this (30) and extended the findings to show that this gene has the expected EGF-like domain (31). In addition to overlapping the P450c21B gene (formally designated CYP21B) (40) encoding steroid 21-hydroxylase, the human XB gene also extensively overlaps the YB gene (see Fig. 1), which encodes an untranslated adrenal-specific RNA (4). In this paper, we delineate the organization, sequence, and tissue distribution of expression of the XB gene. Furthermore, we propose a unified nomenclature system for these various tenascins and assign the name TN-X to the product of this new gene.

---

1. *Abbreviations used in this paper*: Fn III, fibronectin III; TN-X, tenascin X.

Materials and Methods

Genomic and cDNA Library Screening

Genomic libraries prepared from human leukocyte DNA in λ-DASH, and λ-Charon 4A were plated at a density of 3 × 10⁵ pfu/plate. Duplicate lifts onto nitrocellulose filters were made and cross-linked by baking. A cosmid library in pWE15 (Stratagene, San Diego, CA) and a cDNA library in λZAP (4) were plated at 5 × 10⁵ colonies/plate, and duplicate replica filters were made and processed by standard procedures. All filters were probed in 50 mM Tris (pH 7.5)/50% formamide/5× SSC/1× Denhardt's solution/0.1% NaDodSO₄, at 42°C, with probes labeled to a specific activity of >5 × 10⁸ cpm/μg using random primers (Pharmacia Fine Chemicals, Piscataway, NJ). Washing was carried out in 0.1% SSC/0.1% NaDodSO₄ for 10 min at room temperature and for 30 min at 65°C. Positive plaques were picked and purified further by two additional rounds of plating and probing. Bacteriophage DNA was prepared by the method of Helms et al. (17). Cosmid DNA was prepared by alkaline lysis.

DNA Sequencing

Genomic and cDNA fragments for sequencing were subcloned into pBluescript KS or SK vectors (Stratagene) and purified over a CsCl/ethidium bromide gradient. Sequencing was done on double-stranded templates (5) with Sequenase (United States Biochemical, Cleveland, OH) and ³⁵S-labeled dATP as recommended by the manufacturer, except that termination reactions were carried out at 42°C. Reaction products were analyzed by electrophoresis on 5% polyacrylamide gels containing 7 M urea, and subsequent autoradiography. Coding regions and portions of all introns of the XB gene, and the entirety of all cDNAs were sequenced on both strands. Sequences were analyzed with the DNA Inspector (TEXTO, West Lebanon, NH) and the "Eugene" sequence-analysis software (Molecular Biology Information Resource, Baylor College of Medicine, Department of Cell Biology, Waco, TX). Phylogenetic analysis of Fn III repeats was performed using the PAUP program (Swofford, D. L., Illinois Natural History Survey, Champaign, IL, 1991).

Reverse Transcription/Polymerase Chain Reaction (RT/PCR)

Human fetal tissues were quickly frozen in liquid N₂, and total RNA was prepared by homogenization in 4 M guanidinium thiocyanate, followed by ultracentrifugation over a 5.7 M CsCl cushion, as described (33). Pellets were resuspended in 10 mM Tris-HCl (pH 7.6)/1 mM EDTA/0.1% NaDodSO₄ extracted twice with phenol/CHCl₃, precipitated with ethanol, quantitated by adsorption spectrometry at A_{260}/A_{280}, and stored in ethanol before use. PolyA⁺ RNA was obtained by oligo-dT chromatography. Random hexamers were used to prime cDNA synthesis from 100 ng of polyadenylated RNA or 2 μg of total RNA using Superscript reverse transcriptase (Bethesda Research Laboratories, Gaithersburg, MD) for 1 h at 42°C. Aliquots of cDNA were amplified by PCR using Taq polymerase (IBI, New Haven, CT) and specific oligonucleotide primers based on the sequences of various cosmid subclones. PCR products were subcloned into pBluescript (Stratagene) or pCRII (Invitrogen, San Diego, CA) for restriction mapping and DNA sequencing.

Rapid Amplification of cDNA Ends

Rapid amplification of cDNA ends (RACE) was performed as described (12) using a reagent kit (Bethesda Research Laboratories); cDNA synthesis from 2 μg of total muscle RNA was primed with a gene-specific antisense primer and reverse transcribed at 42°C for 1 h. After removal of the initial primer and RNA, the cDNA was tailed with dCTP and terminal transferase. Subsequent PCR used nested antisense primers and a 5′ adaptor primer. PCR products were subcloned into pBluescript for sequence analysis.

RNase Protection

RNase protection experiments were done essentially as described (14). For RNase protection of XB gene transcripts, relevant genomic or cDNA fragments were subcloned into pBluescript and linearized. An antisense probe was synthesized with ³²P-UTP using bacteriophage T7 RNA polymerase. After purification, the probe was hybridized to 15 μg of total RNA in 80% formamide/50 mM PIPES (pH 6.4)/0.4M NaCl/1 mM EDTA overnight at 42°C. Single-stranded RNA was digested with 5–8 μg of DNase-free RNase

A (Boehringer-Manheim, Indianapolis, IN) and 50 U RNase T1 (Pharmacia, Piscataway, NJ) at 37°C for 1 h, and the protected fragments were separated by electrophoresis on a 7% polyacrylamide gel containing 7 M urea.

In Situ Hybridization Histochemistry

Probes for in situ hybridizations were synthesized from linearized pBluescript plasmids containing cDNAs for human P450scc (8), P450c21 (32), TN-X, and Gene Y (4). Antisense riboprobes labeled with ³⁵S-labeled UTP were synthesized with T3 or T7 RNA polymerase as described for the RNase protection probes.

Human tissues (20–23-wk gestation) were obtained through approved protocols of the UCSF Department of Obstetrics, Gynecology and Reproductive Sciences at the time of elective cervical dilatation and evacuation. Tissues were transported in medium 199 on ice, dried to remove excess moisture, and frozen in Tissue Tek O.C.T. Compound (Miles, Inc., Elkhurst, IN) in a methanol/dry-ice bath. 10-μm sections were cut with a cryostat, collected onto microscope slides, and coated with a gelatin chrom alum mixture (0.4%/0.04% wt/vol). Frozen sections were fixed for 5 min in 4% paraformaldehyde in PBS, pH 7.4, and dehydrated in 70 and 100% ethanol. Dried sections were stored with desiccant at −70°C, then treated with proteinase K (2.5 μg/ml in 2× SSC) at 37°C for 15 min, rinsed in 2× SSC, and acetylated for 30 min as described (34). Probes were boiled for 5 min and applied to tissue sections at 100 pg/ml in prehybridization solution 50% formamide, 4× SSC, 1× Denhardt's solution, 140 μg/ml heparin, 25 μg/ml denatured *Escherichia coli* DNA, 2.5 μg/ml polyadenylic acid, and 100 mM DTT, and then hybridized overnight in a humidified environment at 37°C. Slides were washed twice for 15 min at 37°C in 2× SSC containing 100 mM β-mercaptoethanol, twice for 15 min at 37°C in 0.5× SSC/100 mM β-mercaptoethanol, and once for 15 min in 2× SSC at 37°C. Slides were then treated with 20 μg/ml RNase A in 2× SSC at 37°C for 30 min, washed one final time in 2× SSC for 15 min at room temperature, and coated with Ilford K-5 nuclear emulsion diluted 1:1 with 2× SSC. Slides were stored at 4°C, developed after 2 wk of autoradiography, and counterstained with cresyl violet to define tissue architecture.

Results

Arrangement of the Gene Cluster

Tenascin X is encoded by the large XB gene in the class III region of the major histocompatibility locus. This gene is part of a complex array of genes encoding the fourth component of serum complement (C4), the form of cytochrome P450 having steroid 21-hydroxylase activity (P450c21), and a transcript of unknown function operationally termed gene Y (4, 14, 36, 39). This locus is duplicated in the human genome, resulting in the complex array of genes shown in Fig. 1. We initially discovered the X genes overlapping the 3′ ends of the P450c21 genes by cloning a 2.7-kb cDNA fragment identified with a P450c21B cDNA probe (39). The 2.7-kb cDNA is encoded by the large XB gene; the XA gene in turn was truncated by the nonhomologous recombination event that led to the duplication of the human gene cluster, so that it does not appear to encode protein (14). By contrast, the XB gene appeared to encode a 12-kb RNA (14), which in turn encoded a protein resembling the extracellular matrix protein tenascin (14, 60).

Structure of the XB Gene

To clone the XB gene, we first screened a human genomic DNA library in λ-Dash using a 1.8-kb BamHI/EcoRI fragment of the 2.7-kb XB cDNA (39). This probe, which was chosen because it excluded the regions overlapping P450c21, identified λXB-1 as an XB genomic clone. This clone hybridized to an 850-bp PstI/EcoRI fragment from the 5′ end of the 2.7-kb X cDNA and did not hybridize to a P450c21-specific cDNA fragment or a 5′ C4A cDNA (Fig. 1). Probing of a

A

10   20   30   40   50   60   70   80   90   100   110   120

YB   YA

TN-X →   21B   C4B   XA   21A   C4A
                ←   ←   →   ←

B

5kb   λXB-2   λXB-1

cTNX:6-2   cTNX:2-1

cTNX:1A

cTNX:5

5'   0.3   0.6   2.7   3'

0.6

Bam HI | 5.4 | 3.2 |1.8|| 4.0 | 3.0 |1| 3.0 |1| 5.1 |3.2 |1|1| 3.1 | 3.4 | 6.9 | 2.2 |

Eco RI | 12.0 | 3.2 | 3.5 | 13.0 | 7.0 | 5.2 |2|2| 7.0 | 8.5 |

0.6

C

500 Amino Acids

HEPTAD   EGF-LIKE   FN-TYPE III   FG
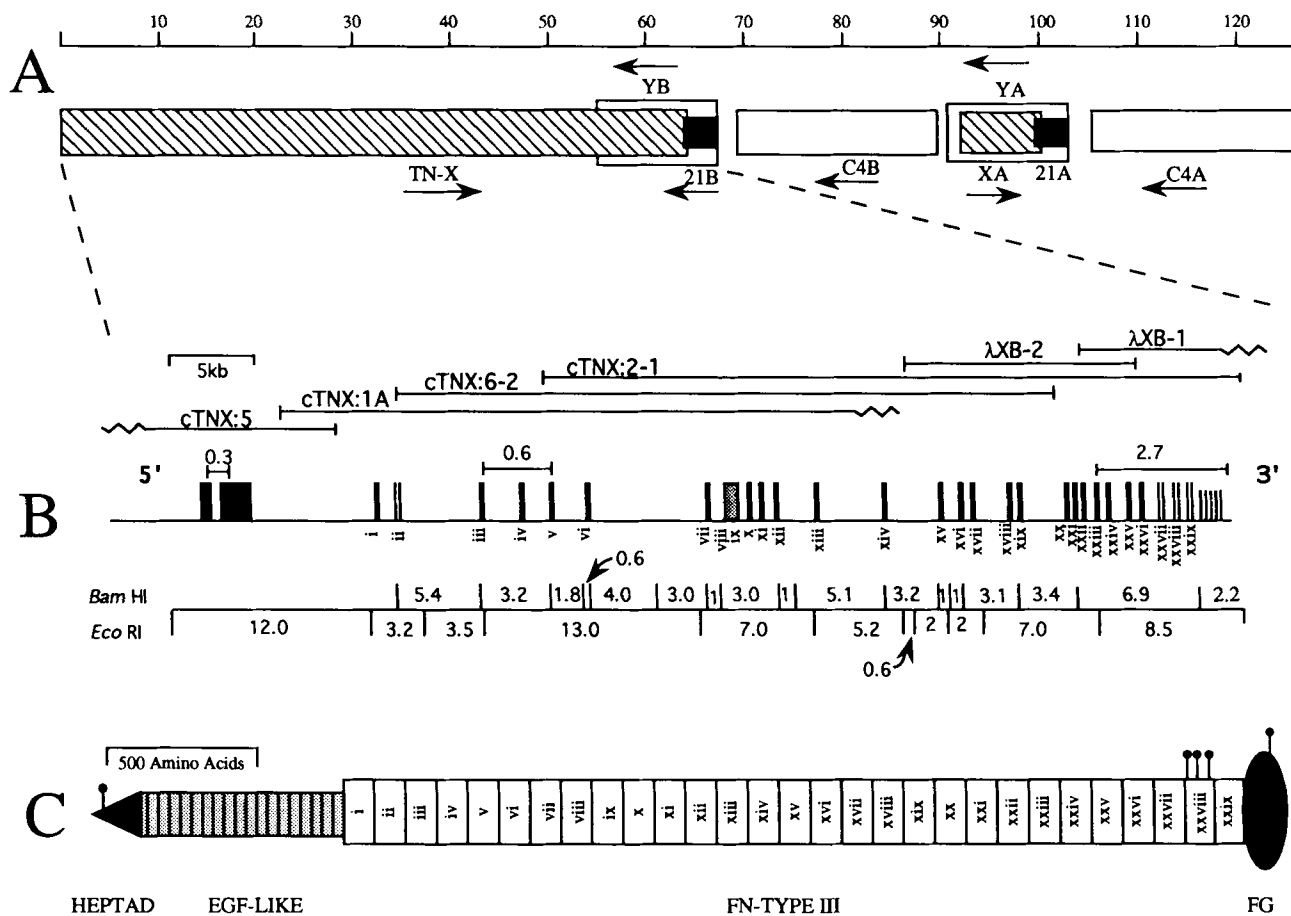
Figure 1. Map of the XB gene-encoding TN-X. (*A*) Organization of the duplicated C4/P450c21/X/Y locus on the short arm of human chromosome 6; the centromere is to the left and the telomere to the right. The upper line shows the scale in kilobases. The diagram shows the TN-X and XA genes as hatched boxes, the P450c21 genes (21A and 21B) as black boxes, and the Y genes as open boxes. The arrows indicate transcriptional orientation. In *B*, the TN-X gene is enlarged to show its exon/intron structure (note the scale marker at the left). The upper lines show the extent of the genomic DNA in the various lambda phage (λ) and cosmid (c) clones discussed in Results. The exons of the TN-X gene are shown as boxes, and the Roman numerals below the line identify the Fn III repeats. cDNAs are indicated as horizontal bars over the relevant exons. Exon 1 encodes the leader peptide and heptad domain; exon 2 encodes the EFG-like repeats, exons 3–34 encode the 29 Fn III repeats, and exons 35–39 encode the fibrinogen-like domain. Note that Fn III repeats ii, xxvii, xxviii, and xxix are each encoded by a pair of "split" exons, while repeats viii and ix are encoded by a single "fused" exon. A map of the BamHI and EcoRI restriction sites is shown below. (*C*) Diagram of the structure of a TN-X monomer, using the symbols popularized by Spring et al. (52). The five "balloons" on the heptad, on Fn III repeat xxviii, and on the fibrinogen-like domain (FG) designate potential N-linked glycosylation sites.

Southern blot of multiple restriction enzyme digests of λXB-1 with the 1.8-kb cDNA probe revealed a 7-kb BamHI fragment and other RFLP markers predicted to be in the DNA 3' of the P450c21B gene (45). DNA sequencing could be initiated by a 20-mer corresponding to the 5' end of the 2.7-kb cDNA, further confirming that phage λXB-1 corresponded to the XB region, and that the 2.7-kb cDNA was encoded by XB.

Mapping of λXB-1 showed it contained 11 kb of human DNA. Sequencing of this DNA showed the 2.7-kb cDNA was encoded by 15 exons spanning 6.6 kb (sequence included in Fig. 2). We found only two bona fide nucleotide differences between the gene and the cDNA. One of these differences was in the third position of a codon and does not result in an amino acid change, while the other results in a substitution of isoleucine for valine (amino acid number 3260 in Fig.

2). A sequencing error (bases 25,817–25,819 in Fig. 2) was also detected. The intron/exon junctions matched the GT/AG rule perfectly, and showed no evidence that the 2.7-kb cDNA corresponded to an alternately spliced mRNA.

λXB-1 also contained the 5' portion of the exon encoding the 5' end of the 2.7-kb cDNA, thus slightly extending the open reading frame. However, Northern blots suggested that the XB transcript might be substantially longer. To elucidate the sequence of XB, we then screened two additional human adrenal cDNA libraries, a human testis cDNA library and a human placenta cDNA library with 5' fragments of the 2.7-kb cDNA; all of these yielded 3' fragments of XB cDNA, but none was longer than the original 2.7-kb clone. Therefore, we extended our knowledge of XB by sequencing the XB gene.

Sequencing of the remainder of the 5' DNA in λXB-1 re-

```
   1: CAGAGCAGGTGCAGAGGCACTGCAGCTGCTCGGTTGCCCAGCCTCTGAATGATGCCAGCCCAGTATGCTCTAACCTCCAGCCTGGTTCTCCTGGTGCTGCTGAGCACAGCCAGAGCAGGCCCCTTCTCTTCA
   1:                                             M  M  P  A  Q  Y  A  L  T  S  S  L  V  L  L  V  L  L  S  T  A  R  A/ G  P  F  S  S
 133: CGGTCCAATGTGACACTGCCAGCCCCCCGGCCCCCTCCCCAGCCAGGGGGCCACACAGTGGGGGCTGGAGTGGGAAGCCCCTCTTCTCAGCTTTACGAGCACACAGTGGAAGGAGGGGAGAAGCAGGTGGTA
  29: R  S (N) V  T  L  P  A  P  R  P  P  P  Q  P  G  G  H  T  V  G  A  G  V  G  S  P  S  S  Q  L  Y  E  H  T  V  E  G  G  E  K  Q  V  V
 265: TTCACCCACCGCATTAACCTGCCCCCTTCCACTGGCTGTGGCTGTCCCCCAGGCACCGAGCCCCCAGTCCTTGCTTCAGAGGTACAGGCCCTGAGGGTCCGTCTAGAGATCCTGGAGGAGTTGGTGAAGGGG
  73: F  T  H  R  I  N  L  P  P  S  T  G  C  G  C  P  P  G  T  E  P  P  V  L  A  S  E  V  Q  A  L  R  V  R  L  E  I  L  E  E  L  V  K  G
 397: CTCAAGGAACAGTGCACTGGGGGATGTTGTCCTGCCCTCTGCCCAAGCTGGCACAG GTGAGCAGGTGATCACACGAAGAGGGTGGAGAGGTGGGGTGGGGTGGGCATTGCTAGTCCATAAAGGTCCTTGGTATG
 117: L  K  E  Q  C  T  G  G  C  C  P  A  S  A  Q  A  G  T
 529: AATTAGAAGAAGGCACTCCTTCCTCACCGTGAGGGGTGTGGATGCAGCCCAGAACACAACTTGGAGAGCAGAGCTGGGCTACATGTCAACCAAAGCATATGCGAGGGCCCTGAGAGGCTGCATACATTACAC
 661: ATACTAGCAACTGGGAGCAGACATGGCCTTATGAGATGAGGCTAGCCTGGCTAGGGGCCTGACACTAGGAGCACTACATAGGCTCAGCCTCTCTCCCAGGAGAGAGACTGAGACTTGCCTCTCCCCTCCTAC
 793: TCCAG GTCAGACAGATGTCCGGACCCTCTGCAGTCTCCATGGTGTGTTTGATCTGAGCCGCTGCACCTGTTCCTGTGAGCCAGGCTGGGGTGGGCCCACCTGCTCAGACCCCACAGATGCTGAGATCCCTCC
 135:       G  Q  T  D  V  R  T  L  C  S  L  H  G  V  F  D  L  S  R  C  T  S  C  C  E  P  G  W  G  G  P  T  C  S  D  T  D  A  E  I  P  P
 925: CTCTTCCCCACCCTCAGCCTCGGGGTCCTGCCCAGATGACTGCAATGATCAGGGTCGCTGTGTCCGTGGTCGTTGCGTGTGCTTTCCCGGCTACACTGGCCCCAGCTGTGGCTGGCCATCCTGTCCCGGGGA
 178: S  S  P  P  S  A  S  G  S  C  P  D  D  C  N  D  Q  G  R  C  V  R  G  R  C  V  C  F  P  G  Y  T  G  P  S  C  G  W  P  S  C  P  G  D
1057: CTGCCAAGGCCGTGGGCGCTGCGTGCAGGGCGTGTGTGTGCCGGGCAGGCTTCTCAGGCCCCGACTGCAGCCAGCGCTCCTGCCCTCGAGGTTGCAGCCAGAGGGGACGCTGTGAGGGTGGGCGCTGCGT
 222: C  Q  G  R  G  R  C  V  Q  G  V  C  V  C  R  A  G  F  S  G  P  D  C  S  Q  R  S  C  P  R  G  C  S  Q  R  G  R  C  E  G  G  R  C  V
1189: GTGTGACCCAGGCTACACTGGTGACGACTGTGGCATGAGGAGCTGCCCTCGCGGTTGCAGTCAGAGGGGCGCTGTGAGAATCGGCGCTGCGTGTGTAACCCCGGCTACACTGGCGCAGGACTGTGGGGTGAG
 266: C  D  P  G  Y  T  G  D  D  C  G  M  R  S  C  P  R  G  C  S  Q  R  G  R  C  E  N  G  R  C  V  C  N  P  G  Y  T  G  E  D  C  G  V  R
1321: GAGCTGCCCTCGGGGCTGCAGCCAGCGGGGACGCTGCAAGGACGGGCGCTGCGTGTGTGACCCCGGCTACACTGGCGAGGACTGTGGTACGCGGAGCTGCCCCTGGGACTGTGGCGAGGGCGGGCGCTGCGT
 310: S  C  P  R  G  C  S  Q  R  G  R  C  K  D  G  R  C  V  C  D  P  G  Y  T  G  E  D  C  G  T  R  S  C  P  W  D  C  G  E  G  G  R  C  V
1453: GGACGGCCGCTGCGTGTGCTGGCCCGGGTACACAGGCGAGGACTGCAGCACGCGGGACATGTCCGAGGGACTGCCGGGGCCGCGGGCGCTGCGAGGACGGCGAATGCATTTGCGACACGGGCTACAGCGGGGA
 354: D  G  R  C  V  C  W  P  G  Y  T  G  E  D  C  S  T  R  T  C  P  R  D  C  R  G  R  G  R  C  E  D  G  E  C  I  C  D  T  G  Y  S  G  D
1585: CGACTGCGGGCGTGCGCAGCTGCCCTGGCGACTGCAACCAAAGGGGCCGCTGCGAGGACGGCCGCTGCGTGTGCTGCTGGCCGGGGTACACTGGAACCGATTGCGGGCTCGCCGCCCTGCCCACGCGACTGTAGAGG
 398: D  C  G  V  R  S  C  P  G  D  C  N  Q  R  G  R  C  E  D  G  R  C  V  C  W  P  G  Y  T  G  T  D  C  G  S  R  A  C  P  R  D  C  R  G
1717: TCGCGGGCGCTGCGAGAACGGCGTGTGTGTTGCAATGCGGGCTACAGCGGCGAGGACTGCGGTGTGCGCAGCTGTCCTGGGGACTGTCGTGGCCGGGCGCTGTGAGAGTGGCCGCTGCATGTGTTGGCC
 442: R  G  R  C  E  N  G  V  C  V  C  N  A  G  Y  S  G  E  D  C  G  V  R  S  C  P  G  D  C  R  G  R  G  R  C  E  S  G  R  C  M  C  W  P
1849: GGGGTACACAGGCCGGGACTGCGGCACGCGCGCCTGTCCTGGCGACTGTCGCGGGCGCGGGCGCTGCGTGGATGGCCGCTGCGTGTGCAACCCGGGCTTCACCGGTGAGGACTGTGGGAGCCGTCGCTGTCC
 486: G  Y  T  G  R  D  C  G  T  R  A  C  P  G  D  C  R  G  R  G  R  C  V  D  G  R  C  V  C  N  P  G  F  T  G  E  D  C  G  S  R  R  C  P
1981: CCGGGACTGCCGTGGGCACGGGCCTTTGCGAGGATGGCGTGTGCGTGTGTGACGGCAGGCTACTCAGGGGAAGACTGCAGCACGCGGCAGCTGCCCCGGGGGCTGCCGAGGCCGCGGCCAGTGCCTAGATGGGCG
 530: G  D  C  R  G  H  G  L  C  E  D  G  V  C  V  C  D  A  G  Y  S  G  E  D  C  S  T  R  S  C  P  G  G  C  R  G  R  G  G  Q  C  L  D  G  R
2113: GTGTGTGTGCGAGGACGGCTACTCTGGCGAGGATTGCGGTGTGAGGCAGTGCCCGAATGACTGCAGCCAGCACGGCGTGTGCCAGGACGGTGTGTGCATCTGTTGGGAAGGCTACGTGAGTGAGGACTGCAG
 574: C  V  C  E  D  G  Y  S  G  E  D  C  G  V  R  Q  C  P  N  D  C  S  Q  H  G  V  C  Q  D  G  V  C  I  C  W  E  G  Y  V  S  E  D  C  S
2245: CATCCGCACCTGCCCCTCCAACTGCCACGGGAGGGGCCGCTGTGAGGAAGGGCGCTGCCTGTGCGACCCAGGCTACACCGGCCCTACCTGTGCCACCCGCATGTGCCCGGCTGACTGCCGGGGACGTGGGCG
 618: I  R  T  C  P  S  N  C  H  G  R  G  R  C  E  E  G  R  C  L  C  D  P  G  Y  T  G  P  T  C  A  T  R  M  C  P  A  D  C  R  G  R  G  R
2377: GTGTGTGCAAGGAGTGTGCCCTGTGCCACGTGGGCTATGGCGGTGAGGACTGCGGGCAGGAAGAGCCTCCAGCCAGCGCCTGCCCTGGAGGCTGCGGGCCCCGGGAACTGTGCCGGGCAGGCCAGTGTGTGTG
 662: C  V  Q  G  V  C  L  C  H  V  G  Y  G  G  E  D  C  G  Q  E  E  P  P  A  S  A  C  P  G  G  C  G  P  R  E  L  C  R  A  G  Q  C  V  C
2509: TGTAGAGGGCTTCCGAGGCCCTGACTGTGCCATCCAGACATGCCCAGGGGACTGCCGTGGCCGAGGAGAGTGTCACGATGGCAGCTGTGTCTGCAAAGATGGGTATGCTGGCGAAGACTGCCGGAGAAG GTGA
 706: V  E  G  F  R  G  P  D  C  A  I  Q  T  C  P  G  D  C  R  G  R  G  E  C  H  D  G  S  C  V  C  K  D  G  Y  A  G  E  D  C  G  E
2641: GCAGGCAGCCTTCCCCAGTGTACTCTGG    about 6kb    GAATTCTGGCAGATGAACACGGCGTGCCCATGAGAGGGGACTCTGACAGATAAATAGGGGGTGCCCATGAGAGGGGACTCTGGGATC
2773: ACACTGCTAGGGCCTAAATTCTTGCCATTTTCTTGGTGTTGGTGACGTTGGCAAGATACTTGTCCTTCTGTGTCTCAGTTTCCTTATCTGTAAAATGGGACTAATAGGACCTGTTTCGTAGGGTTGTTGTGA
2905: GAATTAAGTGAGTTCATGTGTGTACCTCGCCTAGAACAGTGGCTCACAAGGAGTTAGCGCTCTCAGCATGTTTGCTCCCCAGCTTGTGCAGGGTCAAGTTCCACACAGTACCATGCTGACACTGTTTAGAAT
3037: TTGGAGAAAGAAGAGATAAGGGGGATTGGAGCAGCAGAGGCAAGAGTGCCAGCCCCTGAGCCACCTGGTGCTCTCTCTCACAAGACAGAGGGGGGGAGCCCCCATTCACAAGAAGGGTTCCCAG  about
3169:  100 bp    ACTAGAGAGGGAACCTCCCCTCCGGGGAGGGCAAGAAGGAAGGGCCTGGATGGGGCCTAGGCTCTTGGAAGGAAAGATGATGATTGAAGAACCAGACACCCACCTGAGTCCTCCTCTTTAA
3301: CCTGGCTAG TGATCGATGGCCCCCAGGACCTCCGAGTGGTGTCTGTGACACCGACAACACTGGAGCTTGGCTGGCTGCGTCCCCAGGCTGAGGTGGACCGAGTTTGTGGTGTCCTACGTCAGTGCCGACAACC
 748:          V  I  D  G  P  Q  D  L  R  V  V  S  V  T  P  T  T  L  E  L  G  W  L  R  P  Q  A  E  V  D  R  F  V  V  S  Y  V  S  A  D  N
3433: AGAGGGTGAGGCTGGAAGTCCCCCCTGAAACAGACGGGACGCTGCTGACTGACCTGATGCCAGGCGTAGAATATGTGGTGACTGTCACAGCGGAGCGGGGCCGGGCAGTCAGCTACCCAGCTTCTGTCAGGG
 789: Q  R  V  R  L  E  V  P  P  E  T  D  G  T  L  L  T  D  L  M  P  G  V  E  Y  V  V  T  V  T  A  E  R  G  R  A  V  S  Y  P  A  S  V  R
3565: CCAACACAG GTATGGCTGGCCAGAGGTTAGGGAAGGGCCCTGGTTTCCCAGCCTTGGATCC    about 5kb
 833: A  N  T                                                                   g  p  p  r  l  g  e  l  t  v  t  d  r  t  s  d  s  l  l  l  r  w  t  v  p  e  q  e  f
3626:        This amino acid sequence previously reported (30).    GGATCCTGGCCGCAAGTACAAATTTGTCCTGTATGGGTTTGTTGGCAAGAAGAGGCATG
 865: d  s  f  v  i  q  y  k  d  r  d  g  q  p  q  v  v  p  v  e  g  p  q  r  s  a  v  i  t  s  l  D  P  G  R  K  Y  K  F  V  L  Y  G  F  V  G  K  K  R  H
3697: GTCCGCTGGTGGCTGAAGCCAAGATCT GTGAGTGACAGCAGTAACACCCTGCCCTCTGTACTGCCCTGAAGAGGTTTTCTCAGTGCTTGGGAACCTGCTTTGGGAGCTCCGGGAAGGGTTCCTGGAGGCAGT
 916: G  P  L  V  A  E  A  K  I
3829: GGTGCATGAGCTGAATTCTGAAGGGAAAGTAGTGGTGGGGCAGGCAAAGAGTTTGAGAAGCTCAGTGCAGCCTGAGGG    about 1kb    TGCCTCAGAGTGACCCAAGTCCAGGGACTCCACCCCGC
 925:                                                                                        L  P  Q  S  D  P  S  P  G  T  P  P  R
3961: CTGGGAAACCTGTGGGTGACAGACCCTACCCCAGATTCACTGCACCTCTCCTGGACTGTCCCTGAGGGCCAGTTTGACACCTTCATGGTCCAGTACAGGGACAGGGATGGACGGCCCCAGGTGGTACCTGTG
 938: L  G  N  L  W  V  T  D  P  T  P  D  S  L  H  L  S  W  T  V  P  E  G  Q  F  D  T  F  M  V  Q  Y  R  D  R  D  G  R  P  Q  V  V  P  V
4093: GAAGGGCCCGAGCGTTCATTTGTTGTCTCCTCACTGGACCCTGACCACAAGTACAGATTCACTCTGTTTGGAATTGCGAACAAGAAGCGGTATGGCCCCCTCACGGCCGATGGCACCACTG  about 1kb
 982: E  G  P  E  R  S  F  V  V  S  S  L  D  P  D  H  K  Y  R  F  T  L  F  G  I  A  N  K  K  R  Y  G  P  L  T  A  D  G  T  T
4225:        CTCCAGAGAGGGAAAGAGGAGCCCCCCGCCCTGAGTTCCTGGAGCAGCCCCTCCTGGGGGAACTGACAGTGACCCGGCGTGACCCCAGACTCCTTGCGTCTCTCATGACAGTGGCCCAGGGCCC
1022:         A  P  E  R  K  E  E  P  P  R  P  E  F  L  E  Q  P  L  L  G  E  L  T  V  T  G  V  T  P  D  S  L  R  L  S  W  T  V  A  Q  G  P
4357: CTTCGACTCATTACTGGTCCAGTACAAGGATGCACAGGGGCAGCCCCAGGCAGTGCCCTGTTGCGGGGGATGAGAATGAGGTTACTGTCCCCGGCCTGGATCCCGACCGGAAGTATAAGATGAACCTCTACGG
1064: F  D  S  F  M  V  Q  Y  K  D  A  Q  G  Q  P  Q  A  V  P  V  A  G  D  E  N  E  V  T  V  P  G  L  D  P  D  R  K  Y  K  M  N  L  Y  G
4489: GCTTCGAGGCAGGCAGCGTGTGGGGCCCGAGTCTGTGGTGGCCAAGACTG GTGAGTCATGCTGCCAGGCCTCCCTCCCCTGCTAGCCCCATCCTGTGAGCGGGACTTGGCTGGGGCTCCTTCCTGCTGGGT
1108: L  R  G  R  Q  R  V  G  P  E  S  V  V  A  K  T
4621: CCATCTT    about 1.8kb    GGGCCGTCATTCTTGCCAATTCTTTCCTCAATCTGCTATCCCAAAGTAGTCCCCCATGTAGCAGGTCACAGAAGGGCTGGTTTCCCTTGTGCCACAGTCTCCCC
4753: ACTCCTAGGAGCCTCAGCAGAGACCTAACTGTAGGGGGCGGGTAGGAGCCCACAGGGCTGTGATTCCAGAGCCAGCCACCCAAGGACATGTCTCTCCTTGGATCCTTATGCCTCTATCTCTTCTCTCAGACT
4885: CCTCAGGAGGATGTGGACGAG ACCCCCAGCCCCACAGAACTGGCACGGAGGCCCCGGAGTCCCCCGAGGAGCCGCTCCTGGGGGAGCTGACAGTGACAGGATCCTCCCCTGATTCGCTGAGCCTCTTCTGGA
1124:                    D  P  Q  P  H  R  T  G  T  E  A  P  S  P  E  E  P  L  L  G  E  L  T  V  T  G  S  S  P  D  S  L  S  L  F  W
5017: CCGTCCCCCAGGGCAGCTTCGACTCTTTCACCGTGCAGTACAAGGACAGGGATGGGCGGCCCCCGGGCGGTGCGTGTTGGGGGCAAGGAGAGTGAGGTCACCGTGGGAGGCCTAGAGCCCGGGCACAAGTACA
1161: T  V  P  Q  G  S  F  D  S  F  T  V  Q  Y  K  D  R  D  G  R  P  R  A  V  R  V  G  G  K  E  S  E  V  T  V  G  G  L  E  P  G  H  K  Y
5149: AGATGCACCTGTACGGCCTCCACGAGGGGCAGCGCGTGGGCCCGGTGTCCGCCGTGGGGCTGACAG GTGAGTGAGGGG    about 4kb    GGCAGGCAATAGGGAATGATCTAAGAGATGAATAT
1205: K  M  H  L  Y  G  L  H  E  G  Q  R  V  G  P  V  S  A  V  G  L  T
5281: GCCAGTGCAGGACTTGGATTTCACCCTGAGTGAGGTAGAAGCCACTGCGGGAACCAGCTCTGGCTAATGTGTTAACAGGATCC    about 2.5kb    CCCTGGAGCCACGCCTAGGGGAGCTGACAGTAAC
1227:                                                                              A  L  E  P  G  L  E  L  T  V  T
5413: GGATATAACCCCTGACTCTGTGGGCCTCTCATGGACAGTCCCTGAGGGTGAATTCGACTCCTTTGTGGTTCAGTACAAGGACAGGGACGGGCAGCCCCAGGTCGTGCCCGTGGCTGCAGATCAGCGGGAGGT
1238: D  I  T  P  D  S  V  G  L  S  W  T  V  P  E  G  E  F  D  S  F  V  Q  Y  K  D  R  D  G  Q  P  Q  V  V  P  V  A  A  D  Q  R  E  V
5545: CACTATCCCTGACCTGGAACCCTCCACCAAGTACAAGTTCCTGCTCTTTGGGATCCAGGATGGGAAACGACGCAGCCCAGTCTCTGTGGAGGCAAAGACGG GTGAGATGGGCCCTACACAGCTGAGCCTGAG
1282:  T  I  P  D  L  E  P  S  T  K  Y  K  F  L  L  F  G  I  Q  D  G  K  R  R  S  P  V  S  V  E  A  K  T
5677: GCCACAGCCTTCCCACCCTTCTCTTCACTCCCTCTGCCTGAGTCTTCCCTTGGTCCACCTGCTCTGTCTCATTTGAAAAGCCACATGGGGCTAGCACAGTGACTCATGCCTGTAATCCCAGCACTTTGGGAG
5809: TCCAAGGCAGGTCAGGAGTTTGAGACCAGCTCTGGCCAACATGGCAAAACCCTGTCTCTACTAAAAAGACAAAATTAGCTGGGTGTGTGATTGGGTGCTAGTAATTCCAGCTACT    about 700 bp
5941: AGATCCTTCCCTCCGTGAGCTCCCGGCATCATCCCCTCAACATTCTCACAGGCAGCCCACGCTCCTTCACCAGGCCCAGTGGAGCCTTAGTTCTCCCGGGCCAGCCGGTGCAGAATAAGAAAGGGAACGGAG
6073: AAGGGTGAGAACTACCTGTGTCGTAACATCCTGCCTCCCTGACCTGCCGAGTGCCCCCATCAGGGATCCTGCGCTCCTCGGGGCCGGCCAAATCACCCCTTTCCTCAGGGGCCGGCCAAATCACGGCCAAAT
6205: CGTGTGCTCCTGGGTAAGAGGAGATAGAGACCCAGTCTGTAAGAAGCTGCACCCTGCTGGGGAAGCAGTCTGTGAGAGGCCGGAAAGAGGCTGGACAGAAAA    about 500 bp    GCGATCCT
1315:                                                                                                                           ... A  I  L
```

Figure 2. Sequence of the XB gene. The approximate sizes of gaps in the nucleotide sequence are indicated. Because of the gaps, nucleotide numbers indicate only relative distances. BamHI and EcoRI sites used in mapping the gene are underlined. Nucleotide differences in the 2.7-kb cDNA (39), are shown as small letters above the gene sequence. Encoded amino acids are indicated below the second base of the corresponding codon, using the IUPAC single letter code. In exon 1 the signal peptide is separated from sequence of the mature protein by "/," cysteine residues that may participate in inter-chain polymerization are in bold, and hydrophobic amino acids of the heptad repeats are underlined. Amino acid sequence of repeat iii from Matsumoto et al. (30) is in small letters. Incomplete nucleotide sequences of repeats viii, xii, and xiii are indicated by (. . .) preceding or following known amino acid sequences. Potential sites for N-linked glycosylation

6337: CTCCTGGACGGTTCCTGAGGCCAGTTCGACTCAAAAAGTGGTCCAGTTCAAGGACAAAGACGGCCCCCAGGTGGTGCCCGTGGAGGGCCCATGAGCGCTCTGTCACTGTCACCCCTCTGGATGCCGGCCGCAA
1318: S W T V P E A S S T Q K V V Q F K D K D G P Q V V P V E G H E R S V T V T P L D A G R K
6469: GTACAGATTCCTCCTCTATGGCCTCCTGGGCAAGAAGCGCCATGGCCCTCTCACTGCCGACGGCACCTCGGGTGAGGGGCATTCCCTGCAGGTGACCACCGTGACCCAGAACTCCGTGGGCCTCTCCTGGAC
1362: Y R F L L Y G L L G K K R H G P L T A D G T S G E G H S L Q V T T V T Q N S V G L S W T
6601: AGTCCCTGAGGGCCAGTTTGACTCCTTTGTGGTCCAGTACAAAGACAGGGACGGGCAGCCCCAGGTGGTGCCCGTGGAGGGCAGCCTCAGGGAGGTCAGCGTGCCGGGCCTGGACCCTGCCCACAGGTACAA
1406: V P E G Q F D S F V V Q Y K D R D G Q P Q V V P V E G S L R E V S V P G L D P A H R Y K
6733: GCTGCTGCTCTACGGGCTGCACCACGGCAAGCGTGTGGGCCCCATCTCGGCCGTCGCCATTACTG GTGAGTGTGCGGCAGCTGGAACACCTGTGCCTCCTTCCCGCCTGGCTCTCCTGGTCTGACTGAGCCA
1450: L L L Y G L H H G K R V G P I S A V A I T
6865: TAAGATCTCTGAGCTTCCCATTTTATCATATCGAGTCCTTGTGCCCCTTTAGAGTCAGTTGCCCCAGTCCCAGCCTCCAGCAACCACTGTCTGCTTTCGATCATCCTGGAGACAGCATTTTGACTCTCCATG
6997: TAATGGAATCACAATACGCAGTCTCTTGTGTTTCATTTCTTGCAATTAGCACAATGCTTTTGAAATTCACCCATGTGCAGCAAAAGCTGGTCTTTCGGATTGCTGACTGCTTGCCCTCACTCCCTCTCCTCCC
7129: TCACTCCTCCTGACAGTCTGGGTGCAGCGACACACCAGCCATCTGTCTCTACCTGTCTCTGTGAACCAGCCGGCTGGGAAGAAACGGAAACTGAGACCACGGCCCCGACCCCTCCAG CGCCTGAGCCCCACC
1471:   A P E P H
7261: TCGGGGAGTTGACAGTGGAGGAGGCCACGTCACACACCCTGCATCTCTCCTGGATGGTGACTGAGGGAGAATTTGACTCCTTCGAAATCCAGTACACAGATAGAGACGGGCAACTCCAAATGGTCCGGCATAG
1476: L G E L T V E E A T S H T L H L S W M V T E G E F D S F E I Q Y T D R D G Q L Q M V R I
7393: GAGGTGACCGGAATGACATCACCCTCTCTGGCCTGGAATCCGACCACAGATACCTGGTGACCCTGTATGGTTTCAGTGATGGGAAGCATGTAGGTCCTGTCCATGTCGAGGCCCTGACAG GTGAGAACTCTG
1520: G G D R N D I T L S G L E S D H R Y L V T L Y G F S D G K H V G P V H V E A L T
7525: CCCACTATGCCTCCTTTCAGATGGCTGGGAGAGTCCAGAGGACAGCAGAGTCCGCGGATACCTGCCCACCTCAGTCCTCTCTTTCCATGTCTCTGTCCAGTCCCGGAGGAGGAGAAGCCTTCAG AACCTCCC
1560:   E P P
7657: ACCGCAACCCCCGAGCCCCCCATCAAGCCTCGCCTCGGGGGAGCTGACCGTGACAGATGCCACCCCTGACTCCCTCAGCCTGTCCTGGACAGTTCCCGAGGGACAGTTTGACCACTTCCTGGTCCAGTACAGG
1563: T A T P E P P I K P R L G E L T V T D A T P D S L S L S W T V P E G Q F D H F L V Q Y R
7789: AATGGAGATGGGCAGCCCAAGGCAGTGAGGGGTCCAGGGCACGAGGAAGGGGTCACCATCTCGGGCCTGGAGCCAGACCATAAATACAAGATGAACCTGTACGGCTTCCACGGTGGCCAGCGCATGGGCCCT
1607: N G D G Q P K A V R G P G H E E G V T I S G L E P D H K Y K M N L Y G F H G G Q R M G P
7921: GTGTCTGTCGTCGGGGTGACAG GTGAGTGGATGATGGGAGCCCCAGGGTGGGAGCCATGGGAGGGTCACCCTCTTGCTCACCCTCTTGCTCTTTGGTGATGACTGGTGGGGAATGGGACAAGGGTCTGGTCA
1651: V S V V G V T
8053: GCACCACAGACCTGCTTGTGGCTGGGGCTGGGGCTCCCCTTGGGCCTTCCTGTGAGGTTGACCACTGGCTCCTCCTGAACAGAGAGGGGGCCATCGGGAATTTTGCTGTGCTGGTGGCTGTCCCAGGTCCCCC
8185: ACAGCTGACCCTGGAACTTGTCATGTGTGTTAGCTGTCAGCTGAGCAGGACCACCCAGCCCCAAGAGTAGGCCTCTCTGAACTGACCTCGGTCCCCCAGTCATAGCCTTGGCTTCTCCCTCCTTTTCCCCAG
8317: ATCCCAAGGACATCCCCCAGTCATAGCCTTGGCTTCTCCCTCCTTTTCCCAGATCCCAAGGACATCCCCCTCACTCTCTCTTCCTCCTTCCTCAG CTGCAGAGGAAGAGACCCCCAGCCCCCACAGAACCCAGC
1658:   A A E E E T P S P T E P S
8449: ATGGAGGCCCCGGAGCCCGCTGAGGAGCCGCTCCTGGGGGAGCTAACAGTGACAGGATCCTCCCCTGACTCGCTGAGCCTCTCCTGGACCGTCCCCCAGGGCCGCTTCGACTCCTTCACCGTGCAGTACAAG
1671: M E A P E P A E E P L L G E L T V T G S S P D S L S L S W T V P Q G R F D S F T V Q Y K
8581: GACAGGGACGGGCGGCCCCAGGTGGTGCGTGTTGGGGGC about 0.5kb CTTCCTCCTTTCTTTCTTTTCTCTTTCTTTCTTTCTTTCACTTCTTTTCTTTCTTCGTGTTTAGACAAGGTCTCACTCT
1715: D R D G R P Q V V R V G G...
8713: ATCGCCTGGGCTAGAGATCTGGCACAGTCACAATTCACTGCACTGCAGCCTCAACCTCCTGGGCTTAAGGGGATCCTCCCACTTCAGCCTCCCGAATAGCTGGGACTACAAGTGCACTCCACCATGCCCAGC
8845: TAATTTTTTGTATTTTTTGTAGAGACAGGGTTAGCCATGTTGCCCAGGCCAGTCTCGAACTCT about 1 kb GATGTCTCATAATTGTGTTCCCATTTGTTGCTGAATCAGGTCTATAGAAGGAG
8977: GATACATTCGAATTGAGAAAACTGAAATTTTAAATAGAGTAATCAGGGAAGGCCCACAGAAGAAGGTGGCTTTTGAACAAAGACTTAAAGAAGTGTGGGGATGAGCCATGAAGAATATCTGGGGAAAGAACAT
9109: TTAGGTAGAGGGAACAGTCATTGTAAAAGTACTGAGGTAGGAGGCTGCCTGGCATGGTTGAGGGACCATAAGGAAGCCACAAAATCTGATTATGATTATAATCAGAAATCGATCTTAAGTTAGGGCCAGACT
9241: GGGCCCACCTTCTGACCTGAATTCAAGACAGCTGGGGCCTCAACACCTCCTTGCAGCAAGAGAAAAGATTCTGTTTCTATCTCTTCCTAG GGCCCGAAGAGGAGTCCCCTGATGCTCCTCTTGCAAAGCTGC
1728:   P E E E S P D A P L A K L
9373: GCCTAGGGCAGATGACAGTGAGAGACATCACCTCCGACTCCCTCAGCCTCTCCTGGACAGTCCCCGAGGGCCAGTTTGACCATTTCTTGGTCCAGTTTAAGAATGGGGACGGGCAGCCCAAGGCGTTGCGGG
1741: R L G Q M T V R D I T S D S L S L S W T V P E G Q F D H F L V Q F K N G D G Q P K A L R
9505: TGCCGGGACACGAGGATGGG about 3.2 kb GGACCCTGCCATCCTCTCTGTGTCCCTTTTTCTCCAG CTGCAGAGGAAGAGACCCCCAGCCCCACAGAACCCAGCATGGAGGCCCCGGAGCCCCC
1785: V P G H E D G...   A E E E T P S P T E P S M E A P E P P
9637: TGAGGAGCCGCTCCTGGGGGAGCTAACAGTGACAGGATCCTCCCCTGACTCGCTGAGCCTCTCCTGGACCGTCCCCCAGGGCCGCTTCGACTCCTTCACCGTGCAGTACAAGGACAGGGACGGGCGGCCCCA
1811: E E P L L G E L T V T G S S P D S L S L S W T V P Q G R F D S F T V Q Y K D R D G R P Q
9769: GGTGGTGCGTGTTGGGGGCGAGGAGAGCGAGGTCACCGTGGGGGGCCTGGAGCCTGGGCGCAAATACTTGATGCACCTGTATGGCCTCCACGAGGGGCGGCGCGTGGGCCCGGTGTCAGCCGTGGGCGTGAC
1855: V V R V G G E E S E V T V G G L E P G R K Y L M H L Y G L H E G R R V G P V S A V G V T
9901: TG GTAGTAGTGCTT about 1 kb ATTTTAGTAGAGAGGAAGTTTCTCCATGTTGGTCAAGGCTGCTCAAACTCCTGACCTCAGGTGATCTGCCCGCCTCGGCCTCCCAAAGTGCTGGGATTACA
10033: GGCATGAGCCACCGCGCCCAGCCAGACTTATTTCATTTCTAATTATGAGAAGTTCAAAATTTCTAGTCCATAAGAAATTGGAGGGGTGGAGATCATGAGTTAATTGAATTCCTACCGATGGGGAAG
10165: AATCAGCCCTGCTGAGAGGCGGTCTCTTGCAGGGCAGGTCGCCACCTCTGTGGCACCAAGTGAGGTGCGAGCTCATCTCCCAAAGGCCATCCAGACAGAGGAAGCACTCTCAATCGGCATCCCCGATTCCTG
10297: GGGGAGAGGCCTCATTTCCATATGCATGTTCAGAGTTGTGTTCCTGTTCCTCCACTGTGGGGAACCACAGAACTGACAGAGGACAGGCTGAGGACCCCACCCCTGCCCCTCCTGTTCTATGTGTATTGCTGCCCC
10429: ACCCCCAACCCCACACTCTATTATTAATTGTTGTAGCCCAGAATTTCTTTCTTTCTTTTTCTTTTTTTTTTTTTAGACTGAGTCTCACTCTGTTGCCCAGGCTGGAGTGCAGTGGTGTGATCTCAGCTAACT
10561: GCAACCTCTGCCTCCCAGGTTCAAGTGATTGTCCTGCCTCAGCCTCCTGAGTAGCTGGGATTACAGGTGCATGCCACCACGCCTGGCTAATTTTTGTATTTTCAGTAGAGACCGGGTTTCATGATATTGGCC
10693: AAGCTAGACTCGAATTCCACCTCAGGTGATCCACCCGCCCGCCTCGGGACATGCCTGGCCTGTGCTCGGGATTACAGGCGTGAACCACTGCTCCCAGCCCAGAACTTCTTTTTTAGCCCACGTTTTTCTAGTGAAAATAATA
10825: CAGCAACATTATGTGGAAAGCTTGAAAAACAGAAACAAAGAATCTCATAGTCCTACTTCGGGCCCAGCTGCCGGCATTAATAACAATGTGTGTAGTGCACATTCCCTTCCTGTATTTTGCATGCATAGCGTGT
10957: TTTAAAGTTGCAATCAGTGTTCATAAAGTTCTTGGCACTTGCTTGTTTGTACACAAGTACATTGTAATCATTCACCTCACGGCTACACAACCAGCATTCATCATCGTTCAATGGTTATTTATCGCGTTGCGG
11089: TGAAAGCACTATAACAAAATTAATCATCTTCTACGGTCATTTGTGTTCCTGACACATCTGCTGTCATGAATAGCCCATCGTGAGTTCTTCCATGCTATAAACTTTTCCTGCTTTAATAATTTTCTTAGGACAG
11221: ATGCCAGAACTGGCAGGTCAAGGAATTGAGAATTACTTTGGCTCCTGACACTATGTCTCAGTTGCCTTCTGGAGGTTTCTAACAGTGCAGCCTCTGCCCAGAGGCCGGGGGGCTCGGGGATACCTCACCGGCTCT
11353: TATTCCAAGAGTCACTGAACGGCGAACACAAAGCTGCCCCAGCCCTCAGCCTGCTCTGGAGGGGCGCATTTGATGTATGACCTCTGTTGACAGCACCAGCAAAGCAAGTTGCCCTTAAACCCTTAAACTCTG
11485: TATCCCCCTATATTACCTTTCAGCCCCACAAGAGGATGTGGACGAGACCCCCAGCCCTACAG AACCAGGCACAGAGGCCCCAGAGCCCCCCGAGGAGCCTCTCCTGGGGGAGCTGACAGTGACAGGATCCTC
1899:   E P G T E A P E P P E E P L L G E L T V T G S S
11617: CCCTGACTCGCTGAGCCTTTCCTGGACCGTCCCCCAGGGCCGCTTTGACTCCTTCACCGTGCAGTACAAGGACAGGGACGGGCGGCCCCAGGCGGTGCGTGTTGGGGGCCAGGAGAGCAAGGTCACTGTGAG
1923: P D S L S L S W T V P Q G R F D S F T V Q Y K D R D G R P Q A V R V G G Q E S K V T V R
11749: GGGCCTGGAGCCTGGGCGCAAGTACAAAGATGCACCTGTACGGCCTCCACGAGGGGCGGCGCCTGGGCCCGGTGTCTGCCGTGGGCGTCACAG GTGAGTGAGTGTGGGTGGGGCAGGGTTGGAAGACAGCCCT
1967: G L E P G R K Y K M H L Y G L H E G R R L G P V S A V G V T
11881: AGAAAATGTGCCCTTCTCTACCATTTTCCTATACTATTTCTGTCTTGATGGGGCTCACAGTGAAAAGACATGTCATGGAGAGACAGGCTGCAATCCAGCAAATGAAGCAA
12013: AGGCGGGTGAGCATGTGATAGGGAGGCCCAGGGCTCAGGTCAGGACCAGACAGGGACGCCTAAAGTCACCCTGCCCATGGGTACCCAGGGGACAGCCAGGACCTGAGGCCCAGGCATGCCTTAGCTTGGTGAC
12145: AGCTTTAGAGAGAAGGTGAAGTGTGTCAGATAATCACAGCTGGTGCAAAGGCCAGGAGGCTAGAAAGAGCATGGCCACTGAGGATTGAGTGGGGTGTCTTTACGGTGAGTATCTCATCCTGAAGTGTGGGAGC
12277: AGAGGAGGGGACCACTCACCAGGCCTGGGCCTCCCCAGGGATGAGGATGTGGTTGCCCCAGGTCTGTGACATGGCCCCAGCAGCTGTGCCCTGTGTGAAGCTCATCCCGTGGGGCTGAAGATGGG
12409: GATGGGGTGGCAGGGAGCCTGGAGGCAGCGAGGCCCAGTAGGCAGTTGGTGGCCCTGGTGAGAGGTGACAGTGGCTCAAACTAGGATCGGGGACTGGAGGTGGGGTAGGAAGGTATCCAGGGGAATTCAGGGT
12541: AAAGATGCTCTGAGGCTGCTGGCAGCTGGTGAGGAGCTGGATCCAGGAACACACCCCAGGCTCTGGCCTCGGGAGGAGTGTGCTGAGCTTGTTGCGGACGAAAGACAGAAGCCCAGTGAACAAAAGATGGCG
12673: AAGAGACCCCAGTGCTGGGAGGCCAGGGGTGCAGAGGCCCGAGTGGGGCTGTGCTCAAAAGAGAGGCGGTGCTGGAGGGACAGGAGAGGTGGCCTGGGTGTTGGGAGGTGGGGGTGAGGTGGGGGCTGAGGG
12805: CAGGAGGGTCAGGGTGAGGGATAGGAAAGGCCACAGGAGGAGGAGGGGATGAAGACTGTGCTGGAGGGCCTGTGGGCAGCATGCTCCTGGCTCTTGGGCACTTTGTGTTTTGTGACACATCCTTCATTCGAT
12937: GGGGCCAGGACCTCACTGTCCCCACACGTGTCTGTCCAACTCCAG AGGATGAAGCCGAGACCACCCAAGCAGTGCCTACCATGACCCCTGAGCCCCCCATCAAGCCTCGCCTGGGGGAGCTGACCATGACAG
1997:   E D E A E T T Q A V P T M T P E P P I K P R L G E L T M T
13069: ATGCCACCCCTGACTCCCTCAGCCTGTCCTGGACGGTTCCCGAGGGCCAGTTTGACCACTTCCTGGTCCAGTACAGGAATGGGGATGGGCAGCCCAAGGCGGCGCGGGTGCCGGGGCACGAGGACGGGGTCA
2026: D A T P D S L S L S W T V P E G Q F D H F L V Q Y R N G D G Q P K A A R V P G H E D G V
13201: CCATCTCAGGCCTGGAGCCAGACCATAAATACAAGATGAACCTGTACGGCTTCCACGGTGGCCAGCGCGTGGGCCCCATCTCTGTCATTGGGGTGACGG GTGAGTGGATGATGGCAGCCCCAGGGTGGGAGC
2070: T I S G L E P D H K Y K M N L Y G F H G G Q R V G P I S V I G V T
13333: CGTGGGAGGGTCACCCTCTTGCTCTTTGGTGATGACTGGTGGGGAATGGGCCAGGGGTCCGGTCAGCACCACAGACCTGCTTGTGGGCTGGGGCTCCCCTTGGGCCTTCCTCTGAGGCTGACCCCTGGCTCC
13465: TCCTGAGCAGGGAGGGGCCATCAGGAGTTCTGCTGTGCTGGTGACTGTCCCAGGTCCCCCACAGCTGACCCTGGAACTTGTCATGTGTGTTAGCTGTCAGTTGAGCAGGACCACCCAGCCCCAAGAAGTGGGC
13597: TTTTCTGAAATGACCTCACATACCCAGTAGTGGCCATGGTTTCTCCCTCCTTCCCTTGAAGACCTGAGCACATCCCCCAGGCACCTGGCATCCTCTCTATATCTCCTTTTCTCAGCTGCAG AGGAAGAGACC
2093:   E E E T
13729: CCCAGCCCCACGGAACTCAGCACTGAGGCCCCGGAGCCCCCTGAGGAGCCCGCTCCTGGGGGAGCTGGAGCTGACAGTGACAGGATCCTCCCCTGACTCGCTGAGCCTCTCCTGGACCATCCCCCAGGGCCACTTCGAC
2107: P S P T E L S T E A P E P P E E P L L G E L T V T G S S P D S L S L S W T I P Q G H F D
13861: TCCTTCACCGTGCAGTACAAGGACAGGGACGGGCGGCCCCAGGTGATGCGTGTCAGGGGCGAGGAGAGGCAGGTCACCGTGGGGGGCCTGGAGCCCGGGCGCAAATACAAGATGCACCTGTACGGCCTCCAC
2151: S F T V Q Y K D R D G R P Q V M R V R G E E R Q V T V G G L E P G R K Y K M H L Y G L H

Figure 2.

are circled. The polyadenylation signal of TN-X is boxed, and that of P450c21B is indicated by a dashed overline. EMBL accession numbers are as follows: X71923 = exons 1 and 2; X71924 = intron 2; X71926 = Fn repeat i; X71925 = Fn repeat iii; X71927 = Fn repeat iv; X71930 = Fn repeat v; X71938 = Fn repeat vi; X71928 = intron 9; X71929 = Fn repeat vii; X71931 = intron 10; X71936 = Fn repeats viii–xii; X71936 = Fn repeats viii–xii; X71932 = intron 14; X71934 = Fn repeat xiii; X71935 = Fn repeat xiv; X71933 = Fn repeats xv–xvii; X71937 = Fn repeats xviii–xxix; fibrinogen domain, 3' untranslated region.

```
13993: GAGGGGCGGCGTGTGGGCCCGGTGTCCACCGTGGGTGTGACAG GTGAGTGTTTGTGAGTGAGGAAGATGGCCCTAGAAGATGTTGCTTTCTCTGCAACTTCATGAAAACAAAAATATTCTCACCAGCCGGGC
 2195:  E  G  R  R  V  G  P  V  S  T  V  G  V  T
14125: TTCTTTTGCACGTTTCCATGCTTGGGGATCTTGCTAAGAGGCAGATTGGGGGTTCAACAGGTTTGGGTCTAGGCCTTATTGAGATTCTGCATTTCCAACAAGAAATGTTGCGGATTGCTGCTGGCCCATGATC
14257: ACGCCCCTTGAGTAGCAAAGTTCTTCACGACAAAGGAATTGGACCCCTCTTTTGAAATCTGTTGAAGAGAATTTTCTGCGTCCCTGATTCCTGGTAGTGTGCTTTCTCTGTGGAGTTGACTAGGGGCCGTGA
14389: AGGAAGACAGAAGGCAGTGAGGGGCAGCGGTCGACTGCACACTCTGGAAGCCCACTATTGGAATAGGAATAGGAACAGACCTTTCTCACCAATGGGCCAATTTGTTCATTCAGCAAAGAATTCCTTGCCCT
14521: GATCCGGACACTGTAATTTTAGGGTATCAATGTCCTCTGGCCAACGACCTCCAGAAACTCACCTAAAATTGAATGGATGGAATTATGCCTCCCATGCCAGTGCAGGAGGACTGTGGGGTGACTTAGGAAGAG
14653: GCTTTAGAAAGAGAATTGTAAGGAACTGGCTTGTGTTTGGTGTTTAGGAAGACTTAGAGAGCTTCTCTTATAGACC about 800 bp AAGGCCGCAGGAGAGGAGAGGATGAAGAGCTGTGCTGGAG
14785: GCGTGTGGGCAGCATCGTCCTGCTCTTGGGCACTTTGTGTTTTGTGACACATCCTTTCTATGCTGAACTGAGAACCCAGGGACCTCACTCTCCCTATATGTGTCGTCCAGCTCCAG AGGATGAAGCAGAGA
 2209:                                                                                                                    E  D  E  A  E
14917: CCACCCAAGCAGTGCCCACCACAACCCCTGAGCCCCCCAACAAGCCTAGCCTCGGGGAGCTGACCGTGACAGATGCCACCCCTGACTCCCTCAGCCTGTCCTGGATGGTCCGAGGCCAGTTTGACCACTTCC
 2214:  T  T  Q  A  V  P  T  T  T  P  E  P  P  N  K  P  S  L  G  E  L  T  V  T  D  A  T  P  D  S  L  S  L  S  W  M  V  R  G  Q  F  D  H  F
15049: TGGTCCAGTACAGGAATGGGGATGGGCAGCCCAAGGTGGTGCGGTGGCCGGGGCACGAGGACGGGGTCACCATCTCAGGCCTGGAGCCAGACCACAAGTACAAGATGAACCTGTACGGCTTCCACGGTGGCC
 2258:  L  V  Q  Y  R  N  G  D  G  Q  P  K  V  V  R  W  P  G  H  E  D  G  V  T  I  S  G  L  E  P  D  H  K  Y  K  M  N  L  Y  G  F  H  G  G
15181: AGCGCGTGGGCCCCATCTCTGTCATTGGGGTGACAG GTGAGTGTACGATGGGACCCGCAGAGTGGGGCCCTGTGGGAGGGTCTCCCTTTCTCTGGTGATGGGTGAACTGGCCCAGGAAGCCCCTCTGCTCTTGG
 2302:  Q  R  V  G  P  I  S  V  I  G  V  T
15313: CTGAGCCATGGTACTTTTTTGTCTTTCCCCACTTCCCTGAGGACTGACAGATCTTCCTGGGTGGACAAGGGCCCTGTGAGCTCTGTTGGTGGCTGTCCCAAGTTCCCCAGCACTGACCTCAGAGCTTGTCAT
15445: GTGTGTTGACTGTAAACTGAGCAAGACCACCCAGCTCCAAAGATGGGCCTCTCCAAGCTGACCCCAGACCCCACTCATGGCCACCAGTTCGGCTCTCCTTCCTCACAAGACCCAAGGACATCCCCCAGGGAA
15577: GCTGCCTCACCTTCTCTGTCCCCTCTTCTCAG CTGCAGAGGAAGAAACTCCCGCCCCCACAGAACCCAGCACGGAGGCCCCGGAGCCCCTGAGGAGCCGCTCCTGGGGGAGCTGACAGTGACAGGATCCTC
 2314:                                 A  A  E  E  E  T  P  A  P  T  E  P  S  T  E  A  P  E  P  P  E  E  P  L  L  G  E  L  T  V  T  G  S  S
15709: CCCTGACTCGCTGAGCCTCTCCTGGACCATCCCCCAGGGCCGCTTCGACTCCTTCACTGTGCAGTACAAGGACAGGGACGGGCGGCCCCAGGTGGTGCGTGTCAGGGGCGAGGAGAGCGAGGTCACCGTGGG
 2348:  P  D  S  L  S  L  S  W  T  I  P  Q  G  R  F  D  S  F  T  V  Q  Y  K  D  R  D  G  R  P  Q  V  V  R  V  R  G  E  E  S  E  V  T  V  G
15841: GGGCCTGGAGCCCGGGCGCAAATACAAGATGCACCTGTACGGCCTCCACGAGGGGCAGCGCGTGGGCCCAGTGTCCGCTGTGGGTGTGACAG GTGAGTAAGTGTGAGTGAGGCGAGGTGGGGAAGATGCCC
 2392:  G  L  E  P  G  R  K  Y  K  M  H  L  Y  G  L  H  E  G  Q  R  V  G  P  V  S  A  V  G  V  T
15973: TGGAAGACACTGCTGTCTCTCCAGTCTTCGTGAAAACATACTCTGAAGAGTATTCCTTCCTTTGTGTATCCAAATGCCTGGAGATTTTTGTTAAAAAGCAGATTTGGATTCAGCAGGCTTGGCTAGGGCAAG
16105: AGACCATGCATTTCTTTCTTTTTCTTTCTTTTTTTTTTTGAGATAGAGTCTCGCTTTGTCACCCAGGCTGGAGTACAGTGGCTCAATCTCGGCTCACTGGAACCTCCGCCTCCTGGGTTCAAGTG
16237: ATTCTCCTACCTCAGCCTCCTGAGTAACTGGGACTACAGGGCGTGCGCACCCGCGCCCAGCTAATTTTTTCATATTTTTAGTAGAGATGGGTTTCACCCGTGTTAGCCAGGATGGTCTCAATCTCCTGACCT
16369: CGTGATCTGCCCTCCTCGGCCTCCCAAAGTGCTGGGGATCATAGGCGTGAGCCACGGCGCCCAGCCGACCATCCATTTCTAACAAGATTCTCATCTATCTTGATGCCACTGGTCCAAGGAACACACTTTTTT
16501: TGTTGTTTTTTATTGTAGTAAAGTACACTTGACATAATTCGCCATTCTAATCGCTTTCTAGTTTGCAGTCCAGCATGAAGTCCACTCACATTGCTGTGGACTGTCACCCTCCACTTCCAGAACTCCTCTTCC
16633: CAGACTGAAACTTCCTACCCACTAGACACGAACTCCCATTCTCCCCTTGCCCCAGCCCTGGCAACACCATTCTACTCTCTGTCTCTATGAATTTGACTCTAGGTACCTCATATAAGTGCAATCATACAATA
16765: TTTGTCTTTTTTGAATGGTTCATTTCATTAAATACAATGTCTTTAAGGTTCATCTATGTCGTAGCATCAGTCAGAATTTCCTTTTTATTTTATTTTTCTGTGGCAATGGGGGTCTCGGTATGTTGCCCAGGCT
16897: GGTCTCAAACTCCTGGCCTCAAGCGAGTCTCCCACCTTGACCTCCCAAAGTGCTGGGATTATAGGCAAGAGCCACTGCACCTGGCCAGAACTTCCTTCCTTTTCAGGCTGAATAATGTTTCGTTTTACACAT
17029: ACAGACCACACTTTGCTCATCCTTTCATCCATTGATGGACATCTGGGTTCGTTCCACCTCTTAGCTATCAAGAATAATGCTGCTATGAATATTTGTGTCGAAATCACAGAACACCTTTGTAGCAAAGCTCCT
17161: CCCAGTAACAGATTTGAAACTCTCTTAGAATTTGTTGAAGCGAATTTTTCTTGCATTCATGAATCCTCACAGAGGTTAGGCTCAGAGTTAGGGTTCCTGTCCTAATGGGAAAAGTTCCTGTCCTAATGGGGC
17293: TAACGTCATGGGGGACAGGCTGTGGACCAGTAAGAAAGCAAAGGCGGGTGAGCATGTGACAAGAAGCCCAGAGCCAGGCAGGAATACCTAAACCACCCTACCTGTGGATACTCAGGGGACAGTCAGGATCTG
17425: AGGCCAGGCAGGCCTTAGCTTGGTGACAGCTTTAGAGAGAAGGTGAAGCGTGTCAGATCAGCACAGGAGGACAAAGGCCAGGAGGCTAGAAAGAGCATGGTGTGTGAGAAGTACTGAGGATTGAGTGGGGT
17557: GTCCTTTGCAGTGAATAGATAGTCCTGAAGTGTGGGAGCAGAGGAGGCCTGGGGTCTCCCAGGGATGAGGATGTGATTGCCCAGGTCTGTGCTGAGATGGCCCCAGCAGCTGTGCCTGTGTGAAGTTCATCC
17689: GTGGGAGCTAAAGATGGGGATGGGGTGGCAGGGAGCCTGGAGGCAGGGAGGCCAGTAGGCAGTTGGTGGCCCTGGTGAGAGGTGACAGTGGCTCAAACTAGGATTGGGGACTGGAGGTGGGGGAGGAAGGTA
17821: TCCAGGGAAACTCGGGGGAGAGGTACTCTGAGGCTGCTGGCAGCTGGTGAGGGGCTGGGTCCAGGAACACACCCCAGGCGTGGCCTCGGGAGGAGTGTGTGAGCTTGTTGCAGAGCAAAGACAGAAGCCCAG
17953: TGAACAAAAGATGGCTGAGGAGACCCCAGTGCAGGGAGGGTAAGGGTGCAGAGGCCAAGTGGAGCTGTGCTCAAAAGAGGCGGTGCTGGAGGGACAGGGAGGAGGTGGCCTGGGTGTTGGGAGGTGGGGGT
18085: GAGGGTGGGGCCGAGGGCAGGGGGGTCAGGGTGAGGGATAGGAAAGGCCCGCAGGAGAGGAGAGGATGAAGAGCTGTGCTGGAGGCGCTGTGGGCAGCCATCGTCCTGCTCTTGGGCACTTTGTGTTTTGTGA
18217: CACATCCTTTCTATGCTGAACTGAGAACCCAGGACCTCACTCTCCCACACGTGTCTGTCCAG CTCCAGAGGATGAAGCAGAGACCACCCAAGCAGTGCCCACCACAACCCCTGAGCCCCCCAACAAGCCTCG
 2422:                                                                A  P  E  D  E  A  E  T  T  Q  A  V  P  T  T  T  P  E  P  P  N  K  P  R
18349: CCTCGGGGAGCTGACCGTGACAGATGCCACCCCTGACTCCCTCAGCCTGTCCTGGATGGTCCCCGAGGGCCAGTTTGACCACTTCCTGGTCCAGTACAGGAATGGGGATGGGCAGCCCAAGGTGGTGCGGTG
 2446:  L  G  E  L  T  V  T  D  A  T  P  D  S  L  S  L  S  W  M  V  P  E  G  Q  F  D  H  F  L  V  Q  Y  R  N  G  D  G  Q  P  K  V  V  R  W
18481: GCCGGGGCACGAGGACGGGGTCACCATCTCAGGCCTGGAGCCAGACCACAAGTACAAGATGAACCTGTACGGCTTCCACGGTGGCCAGCGCGTAGGCCCTGTGTCTGCCATTGGGGTGTCGG GTGAGTGAAT
 2490:  P  G  H  E  D  G  V  T  I  S  G  L  E  P  D  H  K  Y  K  M  N  L  Y  G  F  H  G  G  Q  R  V  G  P  V  S  A  I  G  V  S
18613: GATGGGAGCCCCAGGGTGGGAGCTGTGGGAGGGCCACCTCTTGCTCTTTGGTGATGACTGGTGGGGAATGGGACAGGGGTCTGGTCAGCACCACAGAACTGCTTGTGGCTGGGGCTGGGACTCCCCTTGGGC
18745: CTTCCTATGTGGTTGACCCCTGGCTCCCCCTGAGCAGGGAGGGGCCATCAGGAGTTTTGCTGTGCTGGTGCTGTGCCAGTCCCCCCACAGCTGACCCTGGAACTTGTTACGTGTGTTAGCTGTCAGCTGAGC
18877: AGGACCACCCAGCCCCAAGAATGGACTTCTCTGAAATGACCTCAGGTCCCCCAGTCATAGCCTTGGCTTCTCCCTCCTTTTCCCCAGGACCCAAGGACATCCCCCTCACTCTCTCTCCCTCCTTCTCCACTG
19009: CAG AGGAAGAGACCCCCAGCCCCCACAGAACCCAGCACTGAGGCCCCGGAGCCCCTGAGGAGCCGGCTCCTGGGGGAGTTGACAGTGACAGGATCCTCCCCCTGACTCGCTGAGCCTCTCCTGGACCGTCCCCC
 2530:  E  E  E  T  P  S  P  T  E  P  S  T  E  A  P  E  A  P  E  E  P  L  L  G  E  L  T  V  T  G  S  S  P  D  S  L  S  L  S  W  T  V  P
19141: AGGGCCGCTTCGACTCCTTCACCGTGCAGTACAAGGACAGGGACGGGCAGCCCCAGGTGGTGCGTGTCAGGGGCGAGGAGAGCGAGGTCACCGTGGGGGGCCTGGAGCCCGGGCGCAAATACAAGATGCATC
 2573:  Q  G  R  F  D  S  F  T  V  Q  Y  K  D  R  D  G  Q  P  Q  V  V  R  V  R  G  E  E  S  E  V  T  V  G  G  L  E  P  G  R  K  Y  K  M  H
19273: TGTACGGCCTCCACGAGGGGCAGCGCGTGGGCCCAGTGTCCACCGTGGGCATCACGG GTGAGTGGGGGGACAGGCCCTCGTCCCCAGGTTTACCTCTGCAGCCCCCTTGTGTTTCTCCTTTGGATCTTGGCA
 2617:  L  Y  G  L  H  E  G  Q  R  V  G  P  V  S  T  V  G  I  T
19405: CCTCTTTTGACTGGGCCTCTAGGTTTCTGTCTTTTCTCCCATGTTGCTAATGATCCTGCCTCATCCTTGGATTCATGGAGACTGTGTGGAGTCAGATGGGCAGGTAGTCCATGCCCTGTTTGTTGTGAGCTTC
19537: CTTCTCCCATTGTGATCTGGAACCTCCATGTTGCTCGTGTGCTCATTTGTTAGTCTTCAGGCTCCCTTAAGGAGTATTTTAGTGGCTCAGAAAGTGCTTCAGATCCAGCTGCCCAGATCTGATGTGCCTTCG
19669: ATGTAGTGTTGAGCGACTTTGAGGCGTTTCCTAACCTTGAGCTTCAGATTCCTGCAAAGTTGGGGTGATGATAATAATGGCACCCACTTCATATGTTGTGCGAGGGTTAAATGCACCACTGTTTGTGAGCTG
19801: CTTACAGCAATGCAGGGCACAGATTCTAAAACAAGCGTTTTAGAGGAAGCGCTAAGAAAATGCTCACTCCAGTCCTGGCAGAGACTGCCTCCCTTGCGCCAGGGTCCTGCCTCCTGACCATGGGCCTGCTCTGC
19933: TCTTTTCAG CGCCCCTGCCCACACCACTGCCGGTGGAGCCCCGCCTGGGGGAGCTGGCGGTGGAGGCCGTGACCTCGGACTCAGTGGGCCTCTCATGGACGGTGGCCCAGGGCCCCTTTGACTCCTTCCTGG
 2636:           A  P  L  P  T  P  L  P  V  E  P  R  L  G  E  L  A  V  E  A  V  T  S  D  S  V  G  L  S  W  T  V  A  Q  G  P  F  D  S  F  L
20065: TACAGTACAGGGACGCGCAGGGGCAGCCCCAGGCAGTGCCTGTGAGCGGAGACCTCCGAGCCGTAGCCGTCTCGGGGCTGGTACCGGCCCGCAAGTACAAGTTCCTGCTCTTTGGACTCCAGAATGGGAAAC
 2677:  V  Q  Y  R  D  A  Q  G  Q  P  Q  A  V  P  V  S  G  D  L  R  A  V  A  V  S  G  L  V  P  A  R  K  Y  K  F  L  L  F  G  L  Q  N  G  K
20197: GCCACGGCCCAGTCCCTGTGGAGGCCAGGACCG GTGAGTGAGGGCTGGAGGCCTCCCGCGGCCCAGAGCCTTCGCCCCCTTGTGGCACCTGTTGGAATTTCACGTTCTGTGCCCCACACTCCAGTCCTCAGCA
 2721:  R  H  G  P  V  P  V  E  A  R  T
20329: CCCACTGATTTATTGGGTCAGGAAAACCCAGGCCCTGAGCTCTCCTCTCCCACTCACACTCATCCCTCTCCCTACTTTCCCTGCACCCCAGGGGACACTTGCTTTCTTGTCTGGCTCCTCTTTTATTCCTA
20461: CTCCTGGCCCAGTCACCCTCCCGTTCCTGGGACCGGCTCACAGAGCCATAGCAGCCCAGGAAGCTCCGTGGCCCCTTTGCCTCCATCCCACTCTCCATGCACCTCACTGTCTTTTCCAG CCCCAGACACCAA
 2732:                                                                                                                          A  P  D  T  K
20593: ACCGTCTCCCCGCCTGGGGGAGCTGACTGTGACAGATGCCACCCCTGACTCCGTGGGCCTCTCGTGGACGGTCCCTGAGGGCGAATTCGACTCCTTCGTGGTCCAGTACAAGGATAAGGATGGTCGGCTCCA
 2737:  P  S  P  R  L  G  E  L  T  V  T  D  A  T  P  D  S  V  G  L  S  W  T  V  P  E  G  E  F  D  S  F  V  V  Q  Y  K  D  K  D  G  R  L  Q
20725: GGTGGTGCCGGTGGCAGCCAACCAGCGGGAGGTCACAGTCCAGGCCCTGGAGCCCAGTAGGAAATACAAGGTTCCTGCTCTATGGTCTGTCAGGCAGGAAACGACTGGGCCCCATCTCTGCTGACAGCACCAC
 2781:  V  V  P  V  A  A  N  Q  R  E  V  T  V  Q  A  L  E  P  S  R  K  Y  K  F  L  L  Y  G  L  S  G  R  K  R  L  G  P  I  S  A  D  S  T  T
20857: AG GTGAGTCCCAGTCCAGCCTCCACCTTTTCCAGAGCTGCCTCTCATCCGAGCCCTCAGAGCTGGCCCTGCAGCCTTCCCGTGAGATTCCCTCTATCAGCCCTGACACAACCATCTTAACTCCGAAAGTGAG
20989: TCCCTCTGTGTGTCAGGCACAGTTCTGAAGCATATGCATTACTTCCTTTCATTCGCTGACTGCTGAGGAAATAAAAGGGACACGGGCTTTGGGGCCAGCAGATCTGTGTTCCAGTTCCAGTCCTAGCCTTTG
21121: CCAGGGTGACCATGGACAAGCCGCTTCATCCCCGGGACCTCAGTGTCCTCACCTGCATAATGCTGGACTCAGTAGTGGTTCCCATCCTCCTCCCTCTGACAGGGCTGAGGAGCCATGGGCCATGAGTCA
21253: GGAAGAAAAGGGAGGGCTTGGCTATGGAAAGACGATGGAAAAACCTGGTGTCACTATGCCTATGACTCAGCTCCCTGAGGAAGGGGTTGGTGGGTCACCGACCCTCACCCCACTCCCCAGGACCATGAGTCA
21385: CTATGCCCAGAGGAGGACAGAGCAGCCGGCCAGTCTGGGCCTGGGTCCCTGGGACTCCGTAATTCTCTTCCCAAAGAGTCACTGTGGGGACGGTACAGATCCCACCCATCGGGCCCCAGCCCCACCCGGGC
21517: AAGCCTCTGCTCTGCCCTCTGTTCTCCCAGCTAATCCCCTGCTCTTCTCCCTGCCCCACCCTGGCTGCCCCGGCCCAGTGGGGTCAGTGTGGGGCTGGGGAAGCAGGAGGCACGGATGTTCAGGCTGATGGC
21649: CACAAGGGGACAAGGGGGAGATCACAGCCTGGCAGTGATGGAGCCGTGCATTGGCCGGCTGCCCAGACCAGCTCTCGATGTTTGGGGTGTCTCCGTGACAACCCAGAGACCTCCGGGGAGTCTCTCTCAGC
21781: TCTGGCCTCATTCCTGCTCAGGGGCTGGGGGTCAGGGTAAACAAAGGCCCCTCTCTGCACCCCCAGCCACCCATCCCTCGGGAGATGATCTGTAATGAATTTGGCGGCATCCTCGATCACAGCAGGGAAGGGGC
21913: GGGGCAGGAAGGAGTTTGGGCAGCAGCCTCCAGAGGAGGAGGGGGCTGTTCTCCCTCATTTCCTGTGGGGCATGGCGGGGAGCAGGCCTGTGTGTCTCCTCAAGGAGCTGTCCCTGGGGCTACACTGGAGGGA
22045: CGATTTCCCAGAACCTCACACCTCCGGGAGGCTGCCAGGGCTTAGGCAAAGGCAGCATGTGACTAAGAGCTTTCCCTCCTCCCTCTGCACAG CTCCCCTGGAGAAGGAGCTACCTCCCCACCTGGGGGAACT
 2825:                                                                                                A  P  L  E  K  E  L  P  P  H  L  G  E  L
22177: GACCGTGGCTGAGGAGACCTCCAGCTCTCTGCGCCTGTCCTGGACGGTAGCCCAGGGCCCCTTTGACTCCTTCGTGGTCCAGTACAGGGACACGGACGGGCAGCCCAGGGCAGTGCCTGTGGCCCAGACCA
 2839:  T  V  A  E  E  T  S  S  S  L  R  L  S  W  T  V  A  Q  G  P  F  D  S  F  V  V  Q  Y  R  D  T  D  G  Q  P  R  A  V  P  V  A  A  D  Q
22309: GCGCACAGTCACCGTAGAGGACCTGGAGCCTGGCAAGAAATACAAGTTTCTGCTCTACGGGCTCCTTGGGGGAAAGCGCCTGGGCCCGGTCTCTGCCCTGGGAATGACAG GTGAGGCTGCTGTGCCTGGCTA
 2883:  R  T  V  T  V  E  D  L  E  P  G  K  K  Y  K  F  L  L  Y  G  L  L  G  G  K  R  L  G  P  V  S  A  L  G  M  T
```

Figure 2.

```
22441: TAGCAAGCCAGCTTGTGTGGGTTTCCTTGTGCATTTGGGCTGAAGACAAAGATGACTGCAGGAGTCGGCAGGCCGGAGTGGGGCGCNCTGNCCTGTCCCCAGGAAGGAGGAGGAGTCTGCAGCCCTGTGGGC
22573: TTCAACATCCATCAAGGAGTCCAGAGCAGGAGCCAGGCCAGGCCAGGAGGAAAGGCCCTGGGCAGGGGCTCTCTAATCTCCCAGCCCCGACTCTGCCCCGTCACTGCCCGCTGCTCCTCATTACTCGCTGGGGC
22705: TGCTGTCGCCTCCCCGAAGGGTGGCCTTGTCCAGATAGCGGCAAACCTCCCTGCCGTGGATGAGTCAGGAGCATTTTCTTAAGAGGAACATCACTGGAAAACAAAATGAGCGGGGACACAGAAACCAACAGC
22837: AGTGGCTGCATTTGTGGTACAGGCTCCTCTTCCAGAGCTCGCTGATGCCCACCTCAGACAGGCCTGACCACGGCACGGCTGGTGGGATTTGCCAGTCACCTCAACCAGCCAGTTCCACCCTCAGCTTCTCTC
22969: AGAAGGGAGCACCACACTCCTCAAGCTCAGTGAATGTATCCCGGCATGGCTGCGGCCCAGACCCTGTCGATATCTCGAGGTGGGCTCGGCAGCACACCCGGGTGTGGAAGGGGGAAGCCAGCACCTGACTCACA
23101: CAGCGCGGGAGCTCGCAGGAGTCACGAGGCCACAGCGACTTCATTGTCTGACTGGGACTGGGCCTGGACCTATAAACTTCCCACCTCAGCCTTGGGCCAAGCCTGGAAGATAAAAATGAGCACCCCATGGC
23233: GCCCCTCACTCAGATTCTCCCCTGGGCTTCTCCCACGCAG CCCCAGAAGAGGACACACCAGCCCCAGAGTTAGCCCCAGAGGCCCCTGAGCCTCCTGAAGAGCCCCGCCTAGGAGTGCTGACCGTGACCGAC
2919:                                            A  P  E  E  D  T  P  A  P  E  L  A  P  E  A  P  E  P  P  E  E  P  R  L  G  V  L  T  V  T  D
23365: ACAACCCCAGACTCCATGCGCCTCTCGTGGAGCGTGGCCCAGGGCCCCTTTGATTCCTTCGTGGTCCAGTATGAGGACACGAACGGGCAGCCCCAGGCCTTGCTCGTGGACGGCGACCAGAGCAAGATCCTC
2950:         T  T  P  D  S  M  R  L  S  W  S  V  A  Q  G  P  F  D  S  F  V  V  Q  Y  E  D  T  N  G  Q  P  Q  A  L  L  V  D  G  D  Q  S  K  I  L
23497: ATCTCAGGCCTGGAGCCCAGCACCCCCTACAGGTTCCTCCTCTATGGCCTCCATGAAGGGAAGCGCTCGGGGCCCCTCTCAGCTGAGGGCACCACAG GTACCACCAGGCGTCTCCGGCCTCTAGCCTAGGAC
2994:       I  S  G  L  E  P  S  T  P  Y  R  F  L  L  Y  G  L  H  E  G  K  R  L  G  P  L  S  A  E  G  T  T
23629: TCAGAAGGGAGAAACGGGGGCTCAGAAGGGGTGGTCGCAGGGAAAGAGCGTGAGGCGGGTACCAGGGAGAGAGGATGGATGGGCTGGATGCGAGTGGCCTTTAGCTCTGCCCCACAGGACCCCCCTGTGGCT
23761: GCAAGTCCCTGGTTACAGATAGAGAAACGGGGGCAGGGAGGGGGGTGGAAGGGACGTGCTCTGGGTCACCAAGCTGGTGTGCTTCTGTCTCCAATCCCTTCTCCCCCACCCACTCCGTGCAG GGCTGGCTCC
3026:                                                                                                                                   G  L  A  P
23893: TGCTGGTCAGACCTCAGAGGCAGTCAAGGCCCCGCCTGTCCCAGCTGTCTGTGACTGACGTGACCACCAGTTCACTGAGGCTCAACTGGGAGGCCCCACCGGGGGGCCTTCGACTCCTTCCTGCTCCGCTTTGG
3030:       A  G  Q  T  S  E  E  S  R  P  R  L  S  Q  L  S  V  T  D  V  T  T  S  S  L  R  L  N  W  E  A  P  P  G  A  F  D  S  F  L  L  R  F  G
24025: GGTTCCATCACCAAGCACACTGGAGCCGCATCCGCGTCCACTGCTGCAGCGCGAGCTGATGGTGCCGGGGACGCGGCACTCGGCCGTGCTCCGGGACCTGCGTTCCGGGACTCTGTACAGCCTGACACTGTA
3074:       V  P  S  P  S  T  L  E  P  H  P  R  P  L  L  Q  R  E  L  M  V  P  G  T  R  H  S  A  V  L  R  D  L  R  S  G  T  L  Y  S  L  T  L  Y
24157: TGGGCTGCGAGGACCCCACAAGGCCGACAGCATCCAGGGAACCGCCCGCACCCTCAGCCCAG GTAAGGACCCACACACACTCTGCCCCAAAGTGGGGGTCTTTGTACTTCACGGGGGGGACCTAGTGCCTCA
3118:     G  L  R  G  P  H  K  A  D  S  I  Q  G  T  A  R  T  L  S  P
24289: GCCAGCCGGTGGGGGTGGGCCGAGTTGGTGGTGGGCCTGGAGGAATCTGCACAGCCGACTTCCATTCCTGGGGACTAGAGGAAAAGGGGTGGTGAGCCTGTGCTGGAGCAGAGGCGAGGGGGGGACTCGCAGGGA
24421: GGAGCCTCCCTGCCCCTGCCTGCGTCATTGTTCCTTGACCCCTCTGCAG TTCTGGACAGCCCCCGTGACCTCCAATTCAGTGAAATCAGGGACGACTCAGCCAAGGTCAACTGGATGCCCCCACCATCCCGG
3138:                                                    V  L  E  S  P  R  D  L  Q  F  S  E  I  R  E  T  S  A  K  V  N  W  M  P  P  P  S  R
24553: GCCGGACAGCTTCAAAGTCTCCTACCAGCTGGCCGACGGAG GTGGTGCCTTTGCCATGTGCTCATCGCCTCGCATTTCCTCTCCCCCCTGCACTCTGCCCACCCTCCAGCTGCCCTGGGGTTCCCTGGGTAAC
3166:     A  D  S  F  K  V  S  Y  Q  L  A  D  G
24685: CCTCGATCCCCAATGTTTTCAG GGGAGCCTCAGAGTGTGCAGGTGGATGGCCAGGCCCGGACCCAGAAACTCCAGGGGCTGATCCCAGGCGCTCGCTATGAGGTGACCGTGGTCTCGGTCCGAGGCTTTGAG
3179:                         G  E  P  Q  S  V  Q  V  D  G  Q  A  R  T  Q  K  L  Q  G  L  I  P  G  A  R  Y  E  V  T  V  V  S  V  R  G  F  E
24817: GAGAGTGAGCCTCTCACAGGCTTCCTCACCACGG GTGACATGGACTGGCACCCGGGGCAAGAGGTGGGAGCCAAGAAAACGGCATGGGTGGGAGTTGAGAGAGAACGAGGAGGGTGAAAGGGAGGTGGTGGA
3216:  E  S  E  P  L  T  G  F  L  T  T
24949: GGCTCCGATTGCGGACGGGAGGCCAGTGGAGTCTGGGGAGGCACGGAGTAGAGAGAGCCGCGGGGACCCCTCTGAGCCCCTCCCCTTCCCCCAG TTCCTGACGGTCCCACACAGTTGCGTGCACTGAACTTG
3227:                                                                                                     V  P  D  G  P  T  Q  L  R  A  L  (N) L
25081: ACCGAGGGATTCGCCGTGCTGCACTGGAAGCCCCCCCAGAATCCTGTAGACACCTATGACATCCAGGTCACAGCCCCTGGGG GTGAGCAGGGCTGAGGCCTCTGGAGGGGACTTGTTCAGGGTGGGGATTGC
3240:  T  E  G  F  A  V  L  H  W  K  P  P  Q  N  P  V  D  T  Y  D  I  Q  V  T  A  P  G
25213: AGGGGGGGAGGCTGGACTCTGGCCGAGGATGGAGGGGGAGGCCTTGATGCCCCTCTCTACACTCCCAG CCCCGCCTCTGCAGGCGGAGACCCCAGGCAGCGCGGTGGACTACCCCCTGCATGACCTTGTCCT
3267:                                                                        A  P  P  L  Q  A  E  T  P  G  S  A  V  D  Y  P  L  H  D  L  V  L
25345: CCACACCAACTACACCGCCACAGTGCGTGGCCTGCGGGGCCCCAACCTCACTTCCCCAGCCAGCATCACCTTCACCACAG GTAGGGTCTGTGGGGTGTGTGGGACAGGGAGAGGAGGTAGAGGGAGCCAGGT
3289:    H  T  (N) Y  T  A  T  V  R  G  L  R  G  P  (N) L  T  S  P  A  S  I  T  F  T  T
25477: TGGGCCCTCATCCCCATCTCCTCTTCCTGCTTTCCCTCCTAG GGCTAGAGGCCCCTCGGGACTTGGAGGCCAAGGAAGTGACCCCCCGCACCGCCCTGCTCACTTGGACTGAGCCCCCAGTCCGGCCCGCAGG
3315:                                             G  L  E  A  P  R  D  L  E  A  K  E  V  T  P  R  T  A  L  L  T  W  T  E  P  P  V  R  P  A  G
25609: CTACCTGCTCAGCTTCCACACCCCTGGTGGACAGAACCAG GTGCCCCGGCCCCACTGACCCAACTCCCCTCCCTGGGTGATTCCAGGAGGTGCTGCTCTGGCCCTCCCGGAGGGTCTCCACCTCCCTCTCCC
3346:  Y  L  L  S  F  H  T  P  G  C Q  N  Q
25741: CTGACCCCCCCTTGTCTGTCCCACAGGAG ATCCTGCTCCCAGGAGGGATCACATCTCACCAGCTCCTTGGCCTCTTTCCCTCCACCTCCTACAATGCACGGCTCCAGGCCATGTGGGGCCAGAGCCTCCTGC
3359:                                  I  L  L  P  G  G  I  T  S  H  Q  L  L  G  L  F  P  S  T  S  Y  N  A  R  L  Q  A  M  W  G  Q  S  L  L
25873: CGCCCGTGTCCACCTCTTTCACCACGG GTACCTGACGCACGGGCCCGGGGCCGGGGGCTGGGTGGGCAGCCAGGGCCTAAGGCTTGGAAAAGGACTGGCCCCTGCTCTCCTCTCCCAG GTGGGCTGCGGATC
3393:  P  P  V  S  T  S  F  T  T                                                                                               G  G  L  R  I
26005: CCCTTCCCCAGGGACTGCGGGGAGGAGATGCAGAACGGAGCCGGTGCCTCCAGGACCAGCACCATCTTCCTCAACGGCAACCGCGAGCGGCCCCTGAACGTGTTTTGCCGACATGGAGACTGATGGGGGCCGGC
3407:  P  F  P  R  D  C  G  E  E  M  Q  N  G  A  G  A  S  R  T  S  T  I  F  L  N  G  N  R  E  R  P  L  N  V  F  C  D  M  E  T  D  G  G
26137: TGGCTG GTGGGTGGCATTCCCAGGGGTCTGTGCAGGGCAAGGGGCTGTTGGCCCCGGAGCCAGAGGCTGATGGTGCCCCCACTTGCTTCCCAG GTGTTCCAGCGCCGCATGGATGGACAGACAGACTTCTGG
3451:  W  L                                                                                            V  F  Q  R  R  M  D  G  Q  T  D  F  W
26269: AGGGACTGGGAGGACTATGCCCATGGTTTTGGGAACATCTCTGGAGAGTTCTGGCTGG GTCAGTGCCTCACAGGGACTGGGGAAGTACGGATGGGGATGGGGGCCCTGTGGACACCAGGACCCTGATGAGGC
3466:  R  D  W  E  D  Y  A  H  G  F  G (N) I  S  G  E  F  W  L
26401: CACGTATCCCACCCCCAG GCAATGAGGCCCTGCACAGCCTGACACAGGCAGGTGACTACTCCATCCGCGTGGACCTGCGGGCTGGGGACGAGGCTGTGTTCGCCCAGTACGACTCCTTCCACGTAGACTCGG
3485:                     G  N  E  A  L  H  S  L  T  Q  A  G  D  Y  S  I  R  V  D  L  R  A  G  D  E  A  V  F  A  Q  Y  D  S  F  H  V  D  S
26533: CTGCGGAGTACTACCGCCTCCACTTGGAGGGCTACCACGGCACCGCAG GTAAGCAGAGGCTGTGAGCCTGGGAGGGTGAGGCTGGCGAGGGGAGGCCCCTCATGGCTCCTTCCTCCACCCTGCCCAG GGGACT
3523:  A  A  E  Y  Y  R  L  H  L  E  G  Y  H  G  T  A                                                                                 G  D
26665: CCATGAGCTACCACAGCGGCAGTGTCTTCTCTGCCCGTGATCGGGACCCCAACAGCTTGCTCATCTCCTGCGCTGTCTCCTACCGAGGGGCCTGGTGGTACAGGAACTGCCACTACGCCAACCTCAACGGGC
3541:  S  M  S  Y  H  S  G  S  V  F  S  A  R  D  R  D  P  N  S  L  L  I  S  C  A  V  S  Y  R  G  A  W  W  Y  R  N  C  H  Y  A  N  L  N  G
26797: TCTACGGGAGCACAGTGGACCATCAG GTGAGGGGTGGGGAGGCGGCTCAGAGCTGGGGTGGCTGGGGGCTCGGCCTGCCTAGGTTTCAGCCCCACAGTGTAACAGGCAAGGGACTGAGTGGCTGGGTGAAATG
3585:  L  Y  G  S  T  V  D  H  Q
26929: GAACAATCATGCCAGCCTCGCAGAGGGAGCTGGAGTTGATTTATTGGCTGGAAAGGGCCAGCTCAGAATTAAGCCTCAATCCTCTGCAGCGGAGGGTCAGGAAGGGAGCTCTGCCGGGGAGGTTGGTTGAGTG
27061: CTGGGAGCTACCTCCTTAAGGGGAATGGGAAGAGCACGATGGGACATCCGGCTTTGACTCTCTCTTGACAACCCCTTTCCCAG GGAGTGAGCTGGTACCACTGGAAGGGCTTCGAGTTCTCGGTGCCCTTCAC
3594:                                                                                       G  V  S  W  Y  H  W  K  G  F  E  F  S  V  P  F  T
27193: GGAAATGAAGCTGAGACCAAGAAACTTTCGCTCCCCAGCGGGGGGAGGCTGAGCTGCTGCCCACCTCTCTCGCACCCCAGTATGACTGCCGAGCACTGAGGGGTCGCCCCGAGAGAAGAGCCAGGGTCCTTC
3611:  E  M  K  L  R  P  R  N  F  R  S  P  A  G  G  G  E
27325: ACCACCCAGCCGCTGGAGGAAGCCTTCTCTGCCAGCGATCTCGCAGCCACTGTGTTTACAGGGGGGAGGGGAGGGGTTCGTACAGGAGC AATAAA GGAGAAACTGAGGTACCCGA
```

Figure 2.

vealed two additional exons encoding 100 amino acid repeats (xxi and xxii in Fig. 1) similar to those found in the 5' end of the X cDNA. Genomic DNA lying 5' of that in λXB-1 was obtained by identifying an overlapping phage (λXB-2) from a library in λCharon 4A by hybridization to a 1.4-kb BglII fragment (bases 19,651–21,088 in Fig. 2) from the 5' end of λXB-1. This extended the available DNA another 10 kb. Complete sequencing of this DNA (Fig. 2) identified repeats xv–xx (Fig. 1). When this proved insufficient, three overlapping XB cosmids (cTNX: 2-1, 6-2, and 1A in Fig. 1) were identified extending 60 kb from the 3' end of P450c21B. Probing of restriction digests of these cosmids with a genomic probe spanning repeats xvi and xvii identified multiple bands, suggesting that additional repeats remained to be identified. Restriction endonuclease mapping of upstream DNA was followed by subcloning of overlapping BamHI and EcoRI fragments. The ends of each subclone were sequenced, and the fragments were ordered by further mapping and sequencing. Repeats i, iii, v–vii, and xii–xiv contained, or were very near BamHI or EcoRI sites, and each was identified by sequence homology to repeats xv–xxix. Another four repeats (viii–xi) were identified by additional sequencing of a 3-kb BamHI fragment, which showed particularly intense hybridization to the genomic probe containing repeats xvi and xvii. Repeat iv was identified by RT/PCR. Repeat ii was mapped but not sequenced (31), and has not been verified by us. Each repeat had canonical splice junction sites at its boundaries, and maintained the same open reading frame as the original 2.7-kb cDNA.

We searched for EGF-like repeats similar to those found near the amino terminus of tenascin by using a PCR-amplified probe encompassing 1.1 kb of the EGF domain of

human tenascin. However, low-stringency hybridization detected no such sequences in cosmids cTNX: 1A, 6-2, or 2-1. Based on the partial sequence of a recently reported genomic DNA fragment of TN-X (31), we synthesized PCR primers corresponding to the apparent EGF-like domain (bases 903–922) and the apparent upstream heptad domain of TN-X (bases 250–269). These primers amplified a 327-bp fragment from adrenal cDNA and a 673-bp fragment from genomic DNA. This 327-bp 5' cDNA was then used to identify our final cosmid cTNX-5 (Fig. 1).

To identify additional upstream coding regions, we used the RACE procedure (12) using oligonucleotide sequences within the 327-bp 5' cDNA fragment. Complementary DNA was synthesized from adrenal or muscle RNA using an oligonucleotide primer in the EGF domain (bases 831–850). This cDNA was 3' tailed with dC, PCR amplified with poly(dG) linked to the adaptor primer and a second oligonucleotide in the EGF repeat (bases 806–826) (the 3' primer). A second round of PCR was performed with the adaptor primer and an oligonucleotide in the heptad repeats (bases 263–282 in Fig. 2). The final product was cloned, and colonies were screened with the 327-bp 5' cDNA. The sequences of positively hybridizing clones extended 220 bp into and beyond a region reported as the 5'-most intron by Matsumoto et al. (31), indicating that this was actually an exonic sequence. To demonstrate that this sequence was actually part of an exon, we performed an additional RT/PCR experiment (Fig. 3). PCR amplification of fetal muscle cDNA using specific primers (bases 60–80 and 806–826 in Fig. 2) yielded a 419-bp fragment, but PCR amplification of fetal muscle RNA without reverse transcription yielded no product, demonstrating that the PCR product arose from newly transcribed cDNA and not from contaminating cDNA or genomic DNA. Amplification of genomic DNA or of cosmid cTNX-5 yielded a 765-bp product, differing from the 419-bp RT/PCR product by retention of the 346 bp in intron 1. Inclusion of this new DNA in exon 1 was also verified in fetal adrenal RNA by RNase protection assay (data not shown).

Determination of the correct intron/exon organization provides the sequence of the amino-terminus, including two ATG translational initiation codons, immediately preceded by a TGA stop codon. The algorithm of von Heijne (55) strongly predicted that the first 22 amino acids constitute a leader peptide for the secretory pathway, consistent with the hypothesis that TN-X is a secreted protein of the extracellular matrix. We mapped and sequenced a total of 37 exons encoding 3,626 amino acids of the XB gene (Fig. 2). In addition, Fig. 1 shows Fn III repeat ii, encoded by exons 4 and 5; this was mapped by Matsumoto et al. (30), but the sequence was not reported. Sequencing of three additional Fn III repeats is incomplete, although the lengths of missing sequences can be estimated from Fig. 4 A. Thus we estimate that the complete TN-X gene contains 39 exons encoding a protein of about 3,816 amino acids. Because fairly large regions of the gene remain unsequenced, additional exons encoding Fn III repeats may remain undiscovered.

Structural Features of the XB gene and Encoded TN-X Protein

The organization of the 39 exons of the XB gene is very similar to that of tenascin. Exon 1 contains a small amount of 5'-untranslated sequence and encodes the signal sequence
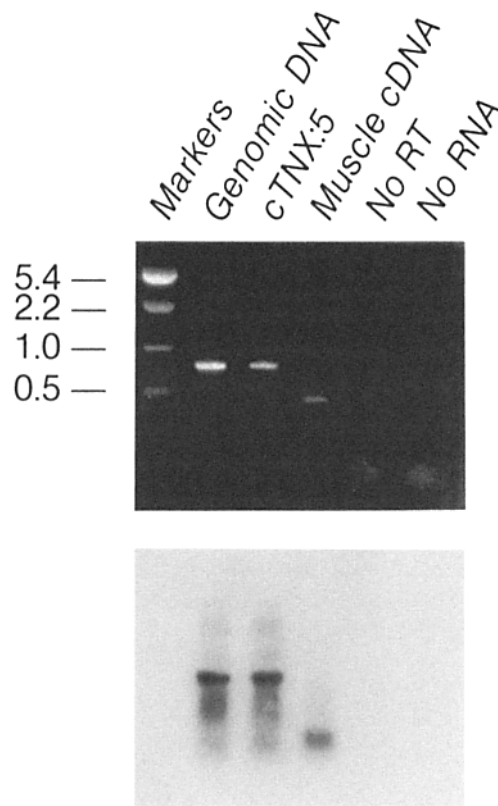
Figure 3. Exon/intron organization of the 5' end of the XB gene. (*Top*) PCR using a 5' primer in the signal peptide and a 3' primer in the EGF repeats. (*Markers*) Bacteriophage PM-2 cut with Hind-III. Other lanes are PCR products obtained with the following templates: human genomic DNA, cosmid cTNX:5 DNA, random-primed human fetal muscle cDNA, muscle RNA without reverse transcription (*RT*), and without added nucleic acid. (*Bottom*) Autoradiogram of the gel pictured above, probed with a 327-bp probe from the heptad domain (see text).

and the "head piece," which in tenascin and restrictin is responsible for polymerization of the protein monomers into multi-armed "brachion" structures. The carboxy-terminal half of the head group is similar in TN-X and tenascin, and contains three heptad repeats flanked by five conserved cysteine residues, suggesting that TN-X is also capable of forming multimers, as described for tenascin (52). Sequences upstream from these cysteine residues are unique in TN-X and lack the additional cysteine residues found in tenascin and restrictin which may permit assembly of higher-order brachion structures. Exon 2 encodes a series of 18.5 EGF-like repeats that are 55% similar to the 13.5 EGF-like repeats of tenascin.

Exons 3–34 encode a series of 29 repeats resembling Fn III repeats (aligned and compared in Fig. 4). The second repeat (exons 5 and 6) and the last three repeats (exons 29–34) are each encoded by a pair of "split exons." Splitting of Fn III repeats into two exons is also seen in the genes for chicken (52) and human tenascin (16). The maximum similarity between tenascin and TN-X Fn III repeats is in the 3' split exons, but is only 40%. This region encodes a cell-binding domain (tenascin repeat 14), identified by Spring et al. (52), but the most similar Fn III repeat in TN-X (xxviii) is only 34% similar. Unlike the tenascin (or restrictin) genes, re-

```
        1    10      20      30      40      50      60      70      80      90      100     110     120
        |    |       |       |       |       |       |       |       |       |       |       |       |
I          VIDGPQDLRVVSVTPTTLELGWLRPQAEVDRFVVSY------------VSADNQRVRLEVPPETDGTLLTDLMPGVEYVVTVTAERGRAVSYPASVRANT
III        gpprlgeltvtdrtsdslllrwtvpegefdsfviqy------------kdrdgqpqvvpvegpqrsavitslDPGRKYKFVLYGFVGKKRHGPLVAEAKI
IV         LPQSDPSPGTPPRLGNLWVTDPTPDSLHLSWTVPEGQFDTFMVQY------------RDRDGRPQVVPVEGPERSFVVSSLDPDHKYRFTLFGIANKKRYGPLTADGTT
V          APERKEEPPRPEFLEQPLLGELTVTGVTPDSLRLSWTVAQGPFDSFMVQY------------KDAQGQPQAVPVAGDENEVTVPGLDPDRKYKMNLYGLRGRQRVGPESVVAKT
VI         DPQPHRTGTEAPESPEEPLLGELTVTGSSPDSLSLFWTVPQGSFDSFTVQY------------KDRDGRPRAVRVGGKESEVTVGGLEPGHKYKMHLYGLHEGQRVGPVSAVGLT
VII        LEPRLGELTVTDITPDSVGLSWTVPEGEFDSFVVQY------------KDRDGQPQVVPVAADQREVTIPDLEPSTKYKFLLFGIQDGKRRSPVSVEAKT
VIII             ...AILSWTVPEGQFDSKVVQF------------KDKDGQPQVVPVEGHERSVTVTPLDAGRKYRFLLYGLLGKKRHGPLTADGTS
IX             GEGHSLQVTTVTQNSVGLSWTVPEGQFDSFVVQY------------KDRDGQPQVVPVEGSLREVSVPGLDPAHRYKLLLYGLHHGKRVGPISAVAIT
X          APEPHLGELTVEEATSHTLHLSWMVTEGEFDSFEIQY------------TDRDGQLQMVRIGGDRNDITLSGLESDHRYLVTLYGFSDGKHVGPVHVEALT
XI         EPPTATPEPPIKPRLGELTVTDATPDSLSLSWTVPEGQFDHFLVQY------------RNGDGQPKAVRGPGHEEGVTISGLEPDHKYKMNLYGFHGGQRMGPVSVVGVT
XII   AAEEETPSPTEPSMEAPEPAEEPLLGELTVTGSSPDSLSLSWTVPQGRFDSRVQY------------KDRDGRPQVVRVGG...
XIII       PEEESPDAPLAKLRLGQMTVRDITSDSLSLSWTVPEGQFDHFLVQF------------KNGDGQPKALRVPGHEDG...
XIV   AEEETPSPTEPSMEAPEPPEEPLLGELTVTGSSPDSLSLSWTVPQGRFDSFTVQY------------KDRDGRPQVVRVGGEESEVTVGGLEPGRKYLMHLYGLHEGRRVGPVSAVGVT
XV         EPGTEAPEPPEEPLLGELTVTGSSPDSLSLSWTVPQGRFDSFTVQY------------KDRDGRPQAVRVGGQESKVTVRGLEPGRKYKMHLYGLHEGRRLGPVSAVGVT
XVI   EDEAETTQAVPTMTPEPPIKPRLGELTMTDATPDSLSLSWTVPEGQFDHFLVQY------------RNGDGQPKAARVPGHEDGVTISGLEPDHKYKMNLYGFHGGQRVGPISVIGVT
XVII  EEETPSPTELSTEAPEPPEEPLLGELTVTGSSPDSLSLSWTIPQGHFDSFTVQY------------KDRDGRPQVMRVRGEEORQVTVGGLEPGRKYKMHLYGLHEGRRVGPVSTVGVT
XVIII EDEAETTQAVPTTTPEPPNKPSLGELTVTDATPDSLSLSWMVR-GQFDHFLVQY------------RNGDGQPKVVRWPGHEDGVTISGLEPDHKYKMNLYGFHGGQRVGPISVIGVT
XIX   AAEEETPAPTEPSTEAPEPPEEPLLGELTVTGSSPDSLSLSWTIPQGRHDSFTVQY------------KDRDGRPQVVRVRGEESEVTVGGLEPGRKYKMHLYGLHEGQRVGPVSAVGVT
XX    APEDEAETTQAVPTTTPEPPNKPRLGELTVTDATPDSLSLSWMVPEGQFDHFLVQY------------RNGDGQPKVVRWPGHEDGVTISGLEPDHKYKMNLYGFHGGQRVGPVSAIGVS
XXI   EEETPSPTEPSTEAPEAPEEPLLGELTVTGSSPDSLSLSWTVPQGRFDSFTVQY------------RNGDGQPKVVRVRGEEESEVTVGGLEPGRKYKMHLYGLHEGQRVGPVSTVGIT
XXII       APLPTPLPVEPRLGELAVEAVTSDSVGLSWTVAQGPFDSFLVQY------------RDAQGQPQAVPVSGDLRAVAVSGLVPARKYKFLLFGLQNGKRHGPVPVEART
XXIII      APDTKPSPRLGELTVTDATPDSVGLSWTVPEGEFDSFVVQY------------KDKDGRLQVVPVAANQREVTVQALEPSRKYRFLLYGLSGRKRLGPISADSTT
XXIV       APLEKELPPHLGELTVAEETSSSLRLSWTVAQGPFDSFVVQY------------RDTDGQPRAVPVAADQRTVTVEDLEPGKKYKFLLYGLLGGKRLGPVSALGMT
XXV   APEEDTPAPELAPEAPEPPEEPRLGVLTVTDTTPDSMRLSWSVAQGPFDSFVVQY------------EDTNGQPQALLVDGDQSKILISGLEPSTPYRFLLYGLHEGKRLGPLSAEGTT
XXVI   GLAPAGQTSEESRPRLSQLSVTDVTTSSLRLNWEAPPGAFDSFLLRFGVPSPSTLEPHPRPPRPLLQRELMVPGTRHSAVLRDLRSGTLYSLTLYGLRGPHKADSIQGTARTLSP
XXVII      VLESPRDLQFSEIRETSAKVNWMPPPSRADSFKVSY------------QLADGGEPQSVQVDGQARTQKLQGLIPGARYEVTVVSVRGFEESEPLTGFLTT
XXVIII     VPDGPTQLRALNLTEGFAVLHWKPPQNPVDTYDIQV------------TAPGAPPLQAETPGSAVDYPLHDLVLHTNYTATVRGLRGPNLTSPASITFTT
XXIX       GLEAPRDLEAKEVTPRTALLTWTEPPVRPAGYLLSF------------HTPGGQNQEILLPGGITSHQLLGLFPSTSYNARLQAMWGQSLLPPVSTSFTT
           !   * !     ! !*  !   !!! !  !!                          !             !! *  ! !         !        !
```

```
                                    ┌───i
                         ┌──100──┐   ├───iii
                         │       └100┤  ┌iv
                         │           └──┤viii
                         │              └v
                  ┌─100─┐│           ┌─────vi
                  │     ││    ┌──100──┤  ┌─xii
                  │     ││    │       └33┤  ┌─xiv
                  │     └100──┤         └51┤  ┌xvii
                  │          │         49┤54┤95┤xix
                  │          │            │  └─┤xxi
                  │          │            └────xv
            ┌─100─┤          └─100──┐     ┌x
            │     │                 │  ┌100┤xi
            │     │                 └50┤   └─xvi
            │     │                    │ ┌100┐xviii
      ┌─100─┤     │                    └100┤ └100┤xx
      │     │     ┌─100──┐                 └─────xiii
-100──┤     │     │      ├──────xxii
      │     └─100─┤      ├──────xxiv
      │          │      └─50─────xxv
      └─100──────┤              vii
                 │              xxiii
                 │              ix
                 │              xxvi
                 └──100────┐   xxvii
                          ┌100┤xxix
                          └───┤xxviii
```

Figure 4. (A) Sequence alignment of the Fn III repeats. Repeats were aligned manually using the highly conserved leucine, tryptophan, and tyrosine residues (*). Missing data are indicated by (. . .); gaps by (—); and identity at a position in 20 or more repeats by (!). The sequence of repeat iii previously reported but not verified by us is shown in small letters; repeat ii has only been identified by hybridization, and has not been sequenced (30). This alignment requires fewer gaps than that proposed by Matsumoto et al. (31). (B) Phylogenetic tree of Fn III repeats obtained with the PAUP program. An heuristic search was performed on 28 repeats based on the alignment in A and using the Dayhoff cost matrix. Initial trees were obtained by stepwise addition using closest neighbors, and the most parsimonious trees were found by tree bisection/reconnection. The consensus of the six trees identified by this process is depicted. Numbers are percentage of trees containing each bifurcation.

peats viii and ix of the XB gene are encoded by a single exon (Figs. 1 and 2). To our knowledge, such a "fused exon" has not been described in other genes encoding proteins with Fn III repeats. Each of the other 24 repeats is encoded by a single exon. The encoded amino acid sequences of the Fn III repeats suggests a series of evolutionary duplication events (Fig. 4 B). The dendrogram shows that Fn III repeats at the amino-terminal end (i–ix) and at the carboxy-terminal end (xxiv–xix) diverged earlier in evolution, while the repeats in the middle of TN-X arose more recently. Exons 35–39 encode the carboxy-terminal fibrinogen-like domain and the 3'-untranslated region. Exon 39 overlaps the P450c21B gene (39). The carboxy-terminal fibrinogen-like domain is widely conserved in evolution (2, 60), and the sequence and intron/exon structure of the 3' end of the XB gene is very similar to the structure of the genes for the β and γ chains of fibrinogen (43). This domain is 54% similar in human tenascin and TN-X.

Cellular Expression of TN-X

Expression of the XB gene encoding TN-X can be evaluated by highly specific RNase protection assays. We used a single-stranded probe from a region (Fn III repeat xxiii) that lies beyond the limit of duplication of this locus; hence, it will not hybridize to and be protected by XA mRNA species. This probe has <50% nucleotide sequence identity with the corresponding region of tenascin, so that tenascin mRNA is not detected. As shown in Fig. 5, the XB gene encoding TN-X is expressed at varying levels in all tissues examined, with greatest expression in fetal testis and in fetal smooth, striated, and cardiac muscle.

An unusual feature of this genetic system is that all of these overlapping RNA transcripts are expressed in a single tissue, the adrenal cortex. The P450c21 mRNA and the four Y transcripts (YA-short, YA-long, YB-short, YB-long) (4) all overlap the XA and XB transcripts. To determine whether these overlapping transcripts are all expressed in the same cells, we performed in situ hybridization histochemistry experiments using probes for P450c21, Y, XA, and XB. All of the probes homogeneously and diffusely stain adult human adrenal cortex (Fig. 6). The P450c21 probe is a 0.7-kb KpnI/EcoRI fragment of our human P450c21 cDNA (32) that excludes XA or XB sequences; because all Y transcripts contain the entire P450c21 sequence (4), it is not possible to construct a P450c21 probe that excludes Y. Fig. 6 A shows that P450c21/Y is ubiquitously expressed in all cells of the adrenal cortex, but not in the adrenal capsule. The XB probe
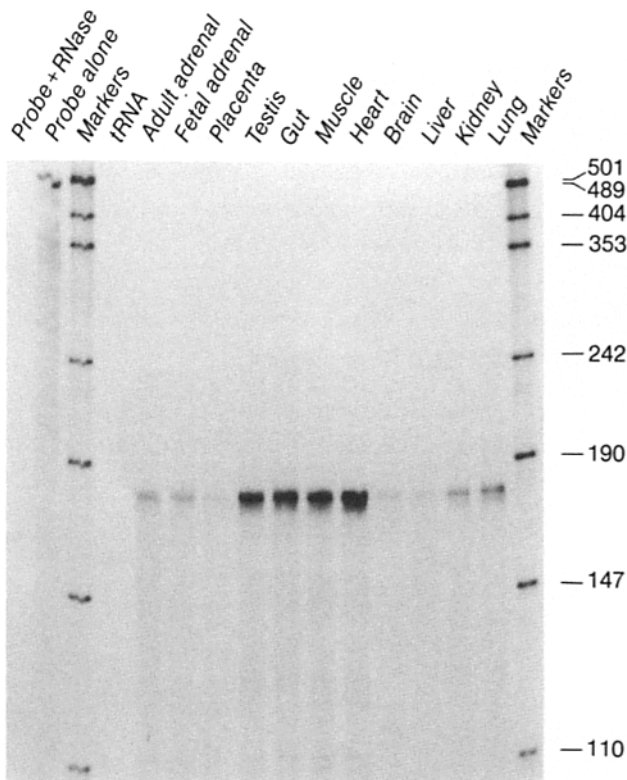
Figure 5. The tissue distribution of TN-X mRNA expression. Samples of 20 μg of RNA from various tissues from 20–22-wk-old human fetuses were assayed by RNase protection using a single-stranded 500-bp riboprobe transcribed from a 1.4-kb BglII genomic fragment encompassing Fn III repeat xxiii. Note that the protected 180-bp fragment is seen in all samples; however, the level of expression varies widely among tissues, and is greatest in fetal testis, gut (small intestine, mostly smooth muscle), muscle (striated), and heart (whole).

is a single-stranded copy of a 353-bp PstI fragment that encompasses exon 23 (Fn III repeat xxi, bases 19,006–19,368 in Fig. 2). This exon lies 3-kb upstream from the cap site of the XA transcript, and has only ~50% nucleotide sequence identity with any portion of XA; hence this probe is specific for XB. Fig. 6 B shows that >90% of cells in the adrenal cortex also contain XB (TN-X) mRNA, and that there is substantial TN-X expression in the connective tissue of the capsule. The XA and Y probes are the opposite strands of a 238-bp SacI fragment extending from bases 2,390 to 2,628 in the XA gene sequence (Fig. 3 A in reference 14); because this sequence lies in the region corresponding to intron 26 in XB (bases 22,876–23,114 in Fig. 2), it is not found in XB (TN-X) mRNA, so that one strand of the probe is specific for XA and the other is specific for Y. Fig. 6 C shows that all cells of the adrenal cortex also contain XA RNA, and Panel D shows that Y RNA is also expressed homogeneously throughout the adrenal cortex, but, as with P450c21, neither XA nor Y is expressed by the adrenal capsule. As each of these four probes hybridizes with RNA in over 90% of the adrenocortical cells, all of these transcripts appear to be expressed in most adrenocortical cells.

As indicated by the RNase protection experiment in Fig.

5, TN-X is widely expressed in human fetal tissues, including the adrenal. In Fig. 7 A, a 400-bp probe from repeats xxvi and xxvii that detects transcripts from both the XB and XA genes diffusely hybridizes to RNA throughout the human fetal adrenal cortex. Similar diffuse hybridization is seen with probes for Y (Fig. 7 B), the cholesterol side-chain cleavage enzyme, P450scc (Fig. 7 C), and P450c21 (Fig. 7 D). Thus the pattern of expression of the P450c21/X/Y genes is similar in both the adult and fetal adrenal cortex.

Discussion

Overlapping Genes and Polygenes

Overlapping genes occur commonly in prokaryots, bacteriophages, and viruses, but only a few examples have been described in higher eukaryotes. These include a cardiac-specific gene on the opposite strand of the gene for rat hypothalamic gonadotropin-releasing hormone (1) two Drosophila loci (18, 51), a mouse locus of unknown function (59), a member of the c-erb-A gene family on the opposite strand from the gene encoding rat (28) and human (37) thyroid hormone receptor, and transcripts on the opposite strand of rat TGF-β_3 (44) N-myc (27), and murine p53 (24). In prokaryotic systems, overlapping genes encoding "antisense" RNAs often are involved in gene regulation (for review see references 20 and 50); the roles of overlapping eukaryotic genes are now being studied.

All of these overlapping gene systems described to date consist of only two transcripts. By contrast, the TN-X/ P450c21/Y system appears to be the first eukaryotic example of three linked, overlapping genes: a "polygene." It is not wholly surprising that such a complex was first discovered in the HLA class III region. This region of the genome appears to be a hot spot for genetic recombination, as evidenced by the highly allelic nature of the HLA locus in general, and by the very high incidence of gene conversion events in the P450c21 and C4 loci (for reviews and references see references 35 and 38). This may have contributed to the very high density of genes in this region: as shown in Fig. 1, little intragenic DNA separates the various known transcriptional units in this locus.

Studies with antisense oligonucleotides, antisense RNAs generated in vitro, and transfected antisense genes (reviewed in reference 54) indicate that the abundance of an mRNA and its encoded protein can be decreased by introducing exogenous antisense sequences, presumably by a mechanism dependent on RNA/RNA hybridization (21, 25). However, endogenously produced antisense RNAs differ from these experimental systems in two important ways. First, heteroduplexes of mRNA and synthetic DNA expose the RNA component to rapid degradation by RNase H–like activities (9, 56); this would not occur with endogenously produced antisense RNA transcripts. Second, experimental inhibition of gene expression generally requires a large molar excess of the antisense sequence, probably because synthetic single-stranded oligonucleotides are unstable in vivo (25). However, antisense nucleotides are stabilized by integration into an open reading frame or by polyadenylation (10, 19). Hence endogenously produced P450c21B/TN-X duplexes or TN-X/Y duplexes are likely to be much more stable than are corresponding synthetic experimental systems. Our in situ hy-
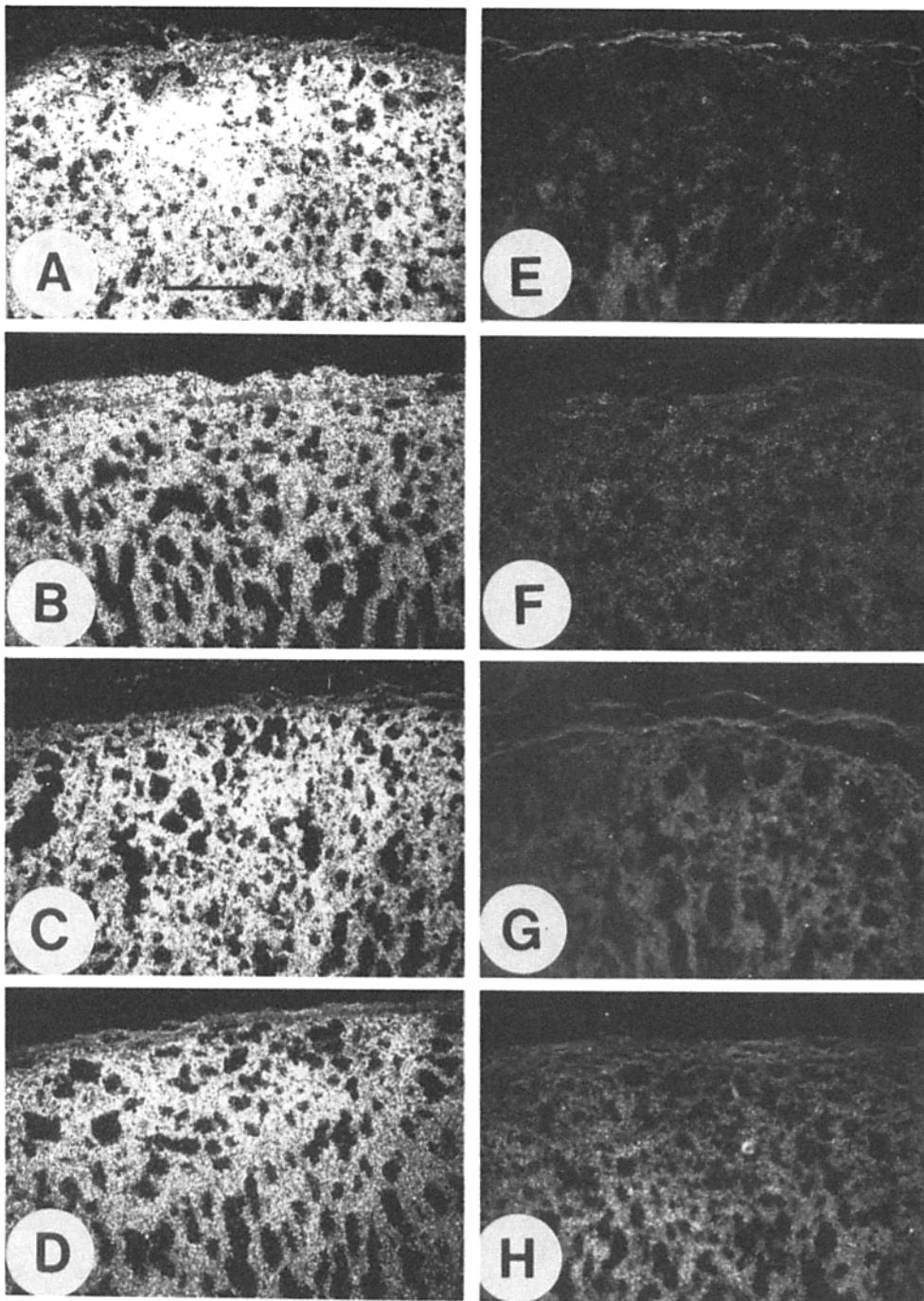
Figure 6. Darkfield photomicrographs of in situ hybridization histochemistry. A frozen surgical sample of human adult adrenal was sectioned and hybridized with the following probes: (*A*) P450c21 cDNA fragment that excludes XA and TN-X sequences; (*B*) a fragment of TN-X cDNA that does not overlap P450c21, XA, or Y; (*C*) a 5′ fragment of XA cDNA that excludes TN-X, P450c21 and Y; and (*D*) a fragment of YA cDNA that hybridizes to both YA and YB, but not to P450c21, XA, or TN-X sequences. The nonhybridizing "holes" in each photograph are due to holes in the desiccated frozen surgical sample. An opposite strand TN-X sense probe is used as a control in panel F. Because of the overlapping nature of the other transcripts, the controls in *E*, *G*, and *H* are sections treated with RNase A before hybridization with cDNAs for P450c21 (*E*), XA (*G*), and Y (*H*). All photomicrographs were taken at the same magnification. Bar, 200 μm.

bridization histochemistry experiments show that P450c21B, XB, and Y are all expressed in the same fetal adrenocortical cells. Thus, formation of such RNA duplexes in vivo is possible.

Some such endogenously produced antisense systems have been studied. N-myc and its antisense transcript form double-stranded hybrids in vivo, which influence N-myc mRNA processing (29). In contrast, the antisense transcript of p53 prevented normal translocation of p53 mRNA to the cytoplasm in murine erythroleukemia cells (24). Kimelman and Kirschner (26) described a small transcript arising from the opposite strand of the gene encoding *Xenopus* basic FGF.

This antisense transcript encoded a protein, but also directly interacted with the sense transcript converting half of the hybridized adenine residues to inosine, thus changing the informational content of the sense transcript. However, we have found no evidence of such RNA editing in the P450c21/ TN-X/Y system.

Tenascins

Tenascin was first described as an extracellular matrix protein in tendons and muscle and later in developing chick brain. It has been variously termed GMEM, tenascin,
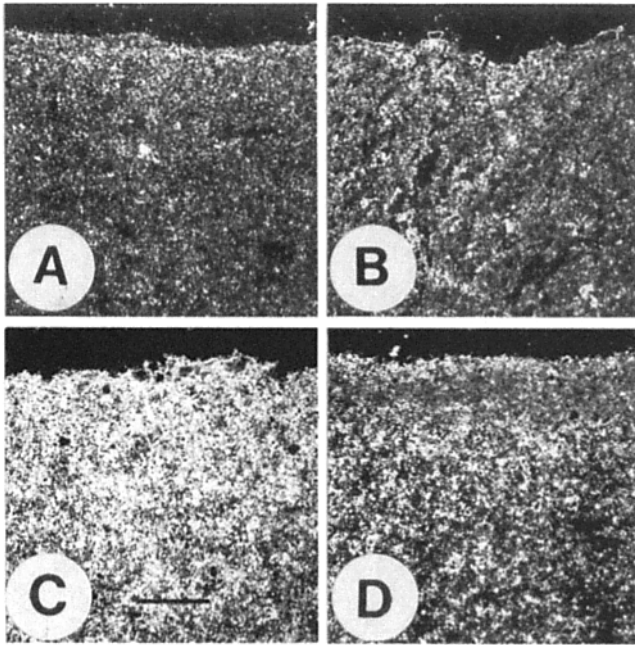
Figure 7. Darkfield photomicrographs of in situ hybridization histochemistry of human fetal adrenal sections. The probes used are: (*A*) a 350-bp HincII/PstI probe that detects transcripts from either the XB or XA genes; (*B*) the same Y-specific probe used in Fig. 6 *D*; (*C*) a 1.2-kb fragment of human P450scc cDNA (8); (*D*) the same P450c21 probe used in Fig. 6 *A*. The controls for the fetal tissues are the same as in Fig. 6, and are not shown. All photomicrographs were taken at the same magnification. Bar, 100 μm.

cytotactin, J1, hexabrachion, and myotendinous antigen (reviewed in reference 11). Elucidation of its gene, mRNA, and protein structure revealed that the protein is a polymer of six chains, each comprising a carboxy-terminal fibrinogen-like domain, a series of Fn III repeats, a series of EGF-like repeats, and an amino-terminal "head group" that permits the chains to aggregate into the characteristic "hexabrachion" seen in electron micrographs (52). The closely related protein restrictin (42), also termed J1-160 (13), is a trimer composed of similar but considerably smaller units: a "tribrachion." The XB gene described in this report encodes a similar but much larger protein that we term tenascin-X, or TN-X. Although there is no direct evidence to date, we predict that TN-X will also polymerize into a "brachion" comprising an unknown number of arms. We suggest that this category of proteins be termed "brachions" or "tenascins." We favor the latter. Tenascin, the first-described member, which is also widely termed "cytotactin," would be termed tenascin-C or TN-C to designate tenascin–cytotactin; restrictin would be termed tenascin-R or TN-R; and the product of the XB gene described in this paper would be tenascin-X or TN-X. This system would emphasize the relatedness among the monomeric units of these proteins and would, to the extent possible, incorporate terminologies and letterings favored by various groups. It seems unlikely that the number of tenascins will exceed the confines of the alphabet.

TN-X protein appears to have well-conserved heptad repeats capable of forming alpha helices. For TN-C and TN-X, the heptads appear to allow three polypeptides to form a tri-

ple coiled-coil which is stabilized by the flanking cysteine residues (46, 52). One interesting difference between TN-X and other members of this gene family is the sequence divergence at the extreme amino terminus of the mature protein. This region appears to mediate dimerization of or TN-C trimers joined in the heptad region to form the final hexabrachion. Cysteine residues are notably absent from this region of TN-X, and hence we predict that TN-X may form pairs or trimers, but not higher order polymers. While the available electron microscopic data show no heterogeneity in the lengths of the arms of TN-C and TN-R, it might be possible for monomers of TN-C, -R, and -X to form heteropolymers.

Several other features of TN-X merit comment. Unlike TN-C and TN-R, which are heavily glycosylated, TN-X has only five Asn-X-Ser (Thr) sites for N-linked glycosylation (Fig. 2). Whether this difference in the potential for glycosylation will confer unique biological properties to TN-X is not known. In addition, both TN-C and TN-R mRNAs undergo alternate splicing to yield different sized mRNAs and proteins. Because most of our structural data come from genomic DNA rather than from cDNA copies of mRNA molecules, we have no data to confirm or deny alternate splicing of TN-X. Finally, unlike human and chicken TN-C, TN-X does not contain an RGD sequence, and hence cannot bind to an RGD-dependent receptor. Recent evidence suggests that TN-C binding is not limited to the domain carrying the RGD sequence (46, 52). RGD-independent binding is consistent with the lack of an RGD sequence in mouse TN-C (57).

Overall, there is greater sequence similarity among the 29 Fn III repeats of TN-X (Table I) than among the 15 Fn III repeats of TN-C (15) or the 9 Fn III repeats of TN-R (42). Our phylogenetic analysis of Fn III repeats in TN-X and Gulcher's similar analysis of TN-C (15) suggest a recent reduplication of central Fn III repeats in these genes that has not occurred in TN-R. Table I shows that 15 of the 20 central Fn III repeats of TN-X (iv–xxiii) are most similar to repeat V of TN-R. This suggests that the central repeats of TN-X arose by duplication of the exon encoding the homolog of TN-R repeat V in the ancestral progenitor gene. These findings support a model in which TN-C, TN-R, and TN-X arose from a smaller primordial gene resembling TN-R, and that subsequent internal reduplications produced the central repeats of TN-C and TN-X after their divergence from one another. For TN-X, frequent duplication might be expected given the high incidence of genetic crossover and homologous recombination events in the MHC region of chromosome 6 (reviewed in reference 38). These events have led to homogenization of duplicated sequences in this region such as C4A and B, P450c21A and B, and the various Fn III repeats of TN-X.

No functional studies have been done to date on TN-X. However, previous studies of TN-C using polyclonal antibodies may have identified mixed populations of TN-C and TN-X. Our RNase protection experiments unambiguously establish that the TN-X mRNA is expressed in nearly all tissues, and is especially widely expressed in developing fetal tissues. It is conceivable that antibodies raised against TN-C might cross-react with TN-X (which might not be detected in Western blotting studies because of its size). Thus, previous studies of the functional role of tenascin based on immunologic reagents may need to be re-examined in the light

Table I. Relatedness of FnIII Repeats in the Tenascins

TN-X repeat	1	3	4	5	6	7	8	9	10	11
Most similar: TN-X	5	8	3	21	14, 19	23	3, 23	8	11, 18	16
Percent Identity	31	66	65	72	90	73	(66)	63	51	92
Most similar: TN-C	1	8	12	6	8	8	12	12	12	12
Percent Identity	31	35	35	35	35	34	(31)	33	33	34
Most similar: TN-R	4	6	5	5	1	5	5	6	6	5
Percent Identity	37	27	26	35	30	27	(22)	28	24	31

	12	13	14	15	16	17	18	19	20	21
Most similar: TN-X	14	12	15	14	11	19	20	21	18	19
Percent Identity	(99)	(88)	93	93	92	93	93	96	93	96
Most similar: TN-C	8	7	8	12	8, 12	8	8	8	12	8
Percent Identity	(46)	(21)	36	34	33	37	31	36	34	35
Most similar: TN-R	5	3	1	5	5	5	5	5	5	5
Percent Identity	(21)	(19)	30	33	30	33	31	31	27	31

	22	23	24	25	26	27	28	29
Most similar: TN-X	5, 7	7	7	24	11	1, 29	1	28
Percent Identity	65	73	65	61	53	27	31	29
Most similar: TN-C	9, 12	10	9	9	8	13	14	15
Percent Identity	33	33	30	39	30	43	34	42
Most similar: TN-R	5	5	9	9	4	6	7	9
Percent Identity	30	28	28	28	26	35	32	36

The alignment of FnIII repeats in Fig. 4 A was used to calculate percent identity of each repeat with the repeat it most closely resembles. Identity figures for incompletely sequenced repeats of TN-X are in parentheses; these figures may not be representative. Each TN-X repeat was also aligned with the most similar Fn III repeat of human tenascin and chicken restrictin, and percent identity calculated.

of the newly established existence of another, very similar, and widely expressed protein.

Crucial insights into function can come from deletion mutants—either created in the laboratory or spontaneously occurring in nature. The C4/P450c21/Gene X locus has been the subject of intensive genetic study because of its association with human disease, and especially with congenital adrenal hyperplasia (35, 36, 38). Among 811 informatively studied human chromosomes bearing mutated alleles causing congenital adrenal hyperplasia, 156 had large gene deletions and another 46 had other genetic rearrangements, yet none of these chromosomes has a deletion extending into the XB gene encoding TN-X (38), suggesting that TN-X may be necessary for survival.

Examination of the duplication junctions of the C4/P450c21/Gene X locus clearly shows that the duplicated human, bovine, and murine loci have different boundaries, suggesting gene duplications and genetic rearrangements that postdate mammalian speciation (14). Therefore, it is difficult to extrapolate from mouse genetics to human genetics in this locus. In the mouse, the P450c21A gene rather than the B gene encodes 21-hydroxylase, while in cattle, both genes are functional. In the mouse genome, the distance from P450c21A to C4B is about 60 kb, possibly enough room to accommodate a functional TN-X gene, whereas the corresponding human XA gene is truncated and does not encode a tenascin-like protein (14). Homozygous deletion of the TN-C gene in mice by gene knockout results in phenotypically normal mice, suggesting that TN-C has no irreplaceable function in the mouse (48). The ubiquitous expression of TN-X in the fetus makes TN-X a logical

candidate to assume the presumed biological function of TN-C. Mice lacking an 80-kb fragment extending from P450c21A to C4B, which may contain the mouse homolog of TN-X, can only be bred as heterozygotes; the homozygous condition is lethal (49). Thus it is possible that TN-X, and not TN-C, is the tenascin essential for development.

References

1. Adelman, J. P., C. J. Bond, J. Duglass, and E. Herbert. 1987. Two mammalian genes transcribed from opposite strands of the same DNA locus. Science (Wash. DC). 235:1514–1517.
2. Baker, N. E., M. Mlodzik, and G. M. Rubin. 1990. Spacing differentiation in the developing Drosophila eye: a fibrinogen-related lateral inhibitor encoded scabrous. Science (Wash. DC). 250:1370–1377.
3. Bourdon, M. A., and E. Ruoslahti. 1989. Tenascin mediates cell attachment through an RGD-dependent receptor. J. Cell Biol. 108:1149–1155.
4. Bristow, J., S. E. Gitelman, M. K. Tee, B. Staels, and W. L. Miller. 1993. Abundant adrenal-specific transcription of the human P450c21A "pseudogene." J. Biol. Chem. 268:12919–12924.
5. Chen, E. Y., and P. H. Seeburg. 1985. Supercoiled sequencing: a fast and simple method for sequencing plasmid DNA. DNA. 4:165–170.
6. Chiquet-Ehrismann, R., E. J. Mackie, C. A. Pearson, and T. Sakakura. 1986. Tenascin: an extracellular matrix protein involved in tissue integra-

tions during fetal development and oncogenesis. *Cell.* 47:131–139.

7. Chiquet-Ehrismann, R., P. Kalla, C. A. Pearson, K. Beck, and M. Chiquet. 1988. Tenascin interferes with fibronectin action. *Cell.* 53:383–390.

8. Chung, B., K. J. Matteson, R. Voutilainen, T. K. Mohandas, and W. L. Miller. 1986. Human cholesterol side-chain cleavage enzyme, P450scc: cDNA cloning, assignment of the gene to chromosome 15, and expression in the placenta. *Proc. Natl. Acad. Sci. USA.* 83:8962–8966.

9. Dash, P. I., I. Lotan, M. Knapp, E. R. Kandel, and P. Goellet. 1987. Selective elimination of mRNAs *in vivo*: complementary oligodeoxynucleotides promote RNA degradation by an RNase H-like activity. *Proc. Natl. Acad. Sci. USA.* 84:7896–7900.

10. Delauney, A. J., Z. Tabaeizadeh, and D. P. S. Verma. 1988. A stable bifunctional antisense transcript inhibiting gene expression in transgenic plants. *Proc. Natl. Acad. Sci. USA.* 85:4300–4304.

11. Erickson, H. P., and M. A. Bourdon. 1989. Tenascin: an extracellular matrix protein prominent in specialized embryonic tissues and tumors. *Annu. Rev. Cell Biol.* 5:71–92.

12. Frohman, M. A., M. K. Dush, and G. R. Martin. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA.* 85:8998–9002.

13. Fuss, B., E. Wintergerst, and M. Schachner. 1993. Molecular characterization of the neural recognition molecule J1-160/180: a modular structure similar to tenascin. *J. Cell Biol.* 120:1237–1249.

14. Gitelman, S. E., J. Bristow, and W. L. Miller. 1992. Mechanism and consequences of the duplication of the human C4/P450c21/gene X locus. *Mol. Cell. Biol.* 12:2124–2134.

15. Gulcher, J. R., M. J. Alexakos, M. M. Le Beau, R. S. Lemons, and K. Stefansson. 1990. Chromosomal localization of the human hexabrachion (tenascin) gene and evidence for recent reduplication within the gene. *Genomics.* 6:616–622.

16. Gulcher, J. R., D. E. Nies, M. J. Alexakos, N. A. Ravikant, M. E. Sturgill, L. S. Marton, and K. Stefansson. 1991. Structure of the human hexabrachion (tenascin) gene. *Proc. Natl. Acad. Sci. USA.* 88:9438–9442.

17. Helms, C., M. Y. Graham, J. E. Dutchik, and M. V. Olson. 1985. A new method for purifying lambda DNA from phage lysates. *DNA.* 4:39–49.

18. Henikoff, S., M. A. Keene, K. Fechtel, and J. W. Fristom. 1986. Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. *Cell.* 44:33–42.

19. Holt, J. T., T. V. Gopal, A. D. Moulton, and A. W. Nienhuis. 1986. Inducible production of c-fos antisense RNA inhibits 3T3 cell proliferation. *Proc. Natl. Acad. Sci. USA.* 83:4794–4798.

20. Inouye, M. 1988. Antisense RNA: its functions and applications in gene regulation—a review. *Gene.* 72:25–34.

21. Izant, J. G., and H. Weintraub. 1984. Inhibition of thymidine kinase gene expression by anti-sense RNA: a molecular approach to genetic analysis. *Cell.* 36:1007–1015.

22. Jones, F. S., M. P. Burgoon, S. Hoffman, K. L. Crossin, B. A. Cunningham, and G. M. Edelman. 1988. A cDNA clone for cytotactin contains sequences similar to epidermal growth factor-like repeats and segments of fibronectin and fibrinogen. *Proc. Natl. Acad. Sci. USA.* 85:2186–2190.

23. Just, M., H. Herbst, M. Hummel, H. Dürkop, D. Tripier, H. Stein, and D. Schuppan. 1991. Undulin is a novel member of the fibronectin-tenascin family of extracellular matrix glycoproteins. *J. Biol. Chem.* 266:17326–17332.

24. Khochbin, S., and J. J. Lawrence. 1989. An antisense RNA involved in p53 mRNA maturation in murine erythroleukemia cells induced to differentiate. *EMBO (Eur. Mol. Biol. Organ.) J.* 8:4107–4114.

25. Kim, S. K., and B. J. Wold. 1985. Stable reduction of thymidine kinase activity in cells expressing high levels of anti-sense RNA. *Cell.* 42:129–138.

26. Kimelman, D., and M. W. Kirschner. 1989. An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in Xenopus oocytes. *Cell.* 59:687–696.

27. Krystal, G. W., B. C. Armstrong, and J. F. Battey. 1990. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Mol. Cell. Biol.* 10:4180–4191.

28. Lazar, M. A., R. A. Hodin, D. S. Darling, and W. W. Chin. 1989. A novel member of the thyroid/steroid hormone receptor family is encoded by the opposite strand of the rat c-erbAα transcriptional unit. *Mol. Cell. Biol.* 9:1128–1136.

29. Leahy, D. J., W. A. Hendrickson, I. Aukhil, and H. P. Erickson. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science (Wash. DC).* 258:987–991.

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

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

32. Matteson, K. J., J. A. Phillips, III, W. L. Miller, B. Chung, P. J. Orlando, H. Frisch, A. Ferrandez, and I. M. Burr. 1987. P450XXI (steroid 21-hydroxylase) gene deletions are not found in family studies of congenital

adrenal hyperplasia. *Proc. Natl. Acad. Sci. USA.* 84:5858–5862.

33. Mellon, S. H., and W. L. Miller. 1989. Extraadrenal steroid 21-hydroxylation is not mediated by P450c21. *J. Clin. Invest.* 84:1497–1502.

34. Mellon, S. H., and C. Vaisse. 1989. cAMP regulates P450scc gene expression by a cycloheximide-insensitive mechanism in cultured mouse Leydig MA-10 cells. *Proc. Natl. Acad. Sci. USA.* 86:7775–7779.

35. Miller, W. L., and Y. Morel. 1989. The molecular genetics of 21-hydroxylase deficiency. *Annu. Rev. Genet.* 23:371–393.

36. Miller, W. L., S. E. Gitelman, J. Bristow, and Y. Morel. 1992. Analysis of the duplicated human C4/P450c21/X gene cluster. *J. Steroid Biochem. Mol. Biol.* 43:961–971.

37. Miyajima, N., R. Horiuchi, Y. Shibuya, S. Fukushige, K. Matsubara, K. Toyoshima, and T. Yamamoto. 1989. Two erbA homologs encoding proteins with different T3 binding capacities are transcribed from opposite DNA strands of the same genetic locus. *Cell.* 57:31–39.

38. Morel, Y., and W. L. Miller. 1991. Clinical and molecular genetics of congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Adv. Hum. Genet.* 20:1–68.

39. Morel, Y., J. Bristow, S. E. Gitelman, and W. L. Miller. 1989. Transcript encoded on the opposite strand of the human steroid 21-hydroxylase/complement component C4 gene locus. *Proc. Natl. Acad. Sci. USA.* 86:6582–6586.

40. Nebert, D. W., D. R. Nelson, M. J. Coon, R. W. Estabrook, R. Feyereisen, Y. Fujii-Kuriyama, F. J. Gonzalez, F. P. Guengerich, I. C. Gunsalus, and E. F. Johnson. 1991. The P450 superfamily: update on new sequences, gene mapping, and recommended nomenclature. *DNA Cell Biol.* 10:1–14.

41. Nies, D. E., T. J. Hemesath, J. H. Kim, J. R. Gulcher, and K. Stefansson. 1991. The complete cDNA sequence of human hexabrachion (Tenascin). A multidomain protein containing unique epidermal growth factor repeats. *J. Biol. Chem.* 266:2818–2823.

42. Nörenberg, U., H. Wille, J. M. Wolff, R. Frank, and F. G. Rathjen. 1992. The chicken neural extracellular matrix molecule restrictin: similarity with EGF-, fibronectin type III-, and fibrinogen-like motifs. *Neuron.* 8:849–863.

43. Pan, Y., and R. F. Doolittle. 1992. cDNA sequence of a second fibrinogen α chain in lamprey: an archetypal version alignable with full-length β and γ chains. *Proc. Natl. Acad. Sci. USA.* 89:2066–2070.

44. Potts, J. D., E. B. Vincent, R. B. Runyan, and D. L. Weeks. 1992. Sense and antisense TGF β₃ mRNA levels correlate with cardiac valve induction. *Dev. Dyn.* 193:340–345.

45. Prentice, H. L., P. M. Schneider, and J. L. Strominger. 1986. C4B gene polymorphism detected in a human cosmid clone. *Immunogenetics.* 23:274–276.

46. Prieto, A. L., C. Andersson-Fisone, and K. L. Crossin. 1992. Characterization of multiple adhesive and counteradhesive domains in the extracellular matrix protein cytotactin. *J. Cell Biol.* 119:663–678.

47. Rüegg, C. R., R. Chiquet-Ehrismann, and S. S. Alkan. 1989. Tenascin, an extracellular matrix protein, exerts immunomodulatory activities. *Proc. Natl. Acad. Sci. USA.* 86:7437–7441.

48. Saga, Y., T. Yagi, Y. Ikawa, T. Sakakura, and S. Aizawa. 1992. Mice develop normally without tenascin. *Genes Dev.* 6:1821–1831.

49. Shiroishi, T., T. Sagai, S. Natsuume-Sakai, and K. Moriwaki. 1987. Lethal deletion of the complement component C4 and steroid 21-hydroxylase genes in the mouse H-2 class III region, caused by meiotic recombination. *Proc. Natl. Acad. Sci. USA.* 84:2819–2823.

50. Simons, R. W., and N. Kleckner. 1988. Biological regulation by antisense RNA in prokaryotes. *Annu. Rev. Genet.* 22:567–600.

51. Spencer, C. A., R. D. Giriz, and R. B. Hodgetts. 1986. Overlapping transcription units in the dopa decarboxylase region of *Drosophila. Nature (Lond.).* 322:279–281.

52. Spring, J., K. Beck, and R. Chiquet-Ehrismann. 1989. Two contrary functions of tenascin: dissection of the active sites by recombinant tenascin fragments. *Cell.* 59:325–334.

53. Tan, S. S., A. L. Prieto, D. F. Newgreen, K. L. Crossin, and G. M. Edelman. 1991. Cytotactin expression in somites after dorsal neural tube and neural crest ablation in chicken embryos. *Proc. Natl. Acad. Sci. USA.* 88:6398–6402.

54. van der Krol, A. R., J. N. M. Mol, and A. R. Stuitje. 1988. Modulation of eukaryotic gene expression by complementary RNA or DNA sequences. *Biotechniques.* 6:958–976.

55. von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 14:4683–4690.

56. Walder, R. Y., and J. A. Walder. 1988. Role of RNase H in hybrid-arrested translation by antisense oligonucleotides. *Proc. Natl. Acad. Sci. USA.* 85:5011–5015.

57. Weller, A., S. Beck, and P. Ekblom. 1991. Amino acid sequence of mouse tenascin and differential expression of two tenascin isoforms during embryogenesis. *J. Cell Biol.* 112:355–362.

58. Weller, A., S. Beck, and P. Ekblom. 1991. Amino acid sequence of mouse tenascin and differential expression of two tenascin isoforms during embryogenesis. *J. Cell Biol.* 112:355–356.

59. Williams, T., and M. Fried. 1986. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. *Nature (Lond.).* 322:275–279.

60. Xu, X., and R. F. Doolittle. 1990. Presence of a vertebrate fibrinogen-like sequence in an echinoderm. *Proc. Natl. Acad. Sci. USA.* 87:2097–3101.