The Genomic Organization of Platelet Glycoprotein IIIa*

(Received for publication, January 2, 1990)

Ann B. Zimrin‡, Susan Gidwitz§, Susan Lord¶, Elias Schwartz∥, Joel S. Bennett‡, Gilber C. White II§, and Mortimer Poncz∥**

From the §Hematology-Oncology Section, the Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, the Departments of ¶Medicine and ∥Pathology, the University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, and the ||Division of Hematology, the Children’s Hospital of Philadelphia and the University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104

The platelet membrane glycoprotein (GP) IIb/IIIa complex, a member of the integrin family of adhesive receptors involved in cell-cell and cell-matrix interactions, contains binding sites for fibrinogen, von Willebrand factor, fibronectin, and vitronectin. Absence or defects of this receptor result in the platelet bleeding disorder Glanzmann’s thrombasthenia. In this report, we describe the isolation of genomic DNA coding for the entire mature GPIIIa protein. Mature GPIIIa is encoded by 14 exons which range in length from 90 to 3618 base pairs, which are contained within an approximately 46-kilobase (kb) stretch of genomic DNA on chromosome 17. All of the exon/intron junctions were found to conform to the consensus splice donor and acceptor sequences. The coding region of the GPIIIa gene is identical with the previously described cDNA sequence except for three silent substitutions. One substitution creates a TaqI site which may be the site of a known GPIIIa polymorphism. A second substitution eliminates a Smal site. Aside from the start of the first exon described, which begins at the second base of the first codon of the mature protein, there is no correlation between the organization of the exons in this gene and proposed functional domains of the protein based on analysis of the primary amino acid sequence. The less frequently used polyadenylation signal AAATTAAA was present at the 3’-end of the major RNA transcript. Recently, an alternatively processed GPIIIa transcript has been described. We demonstrate that this transcript results from nonsplicing of the final intron. The description of the GPIIIa gene organization should be of importance in understanding the evolution of the integrin family of receptors and should be useful in the molecular biology analysis of thrombasthenic patients who have a defect in the GPIIIa gene.

* This work was supported in part by Grants HL27419, AM16691, HL40837, HL28157, HD07107, and DE02815 from the National Institutes of Health, the March of Dimes (Basic Research Grant to M. P.), and the Council for Tobacco Research-U.S.A., Inc. (1570 to J. S. B.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBankTM/EMBL Data Bank with accession number(s) J05427.

** To whom correspondence and reprint requests should be addressed: Division of Hematology, the Children’s Hospital of Philadelphia, 34th St. and Civic Center Blvd., Philadelphia, PA 19104.

1 The abbreviations used are: GP, glycoprotein; kb, kilobase(s); bp, base pair(s).
the complete intron-exon organization of the GPIIIa gene encoding the mature protein.

METHODS

Isolation of Genomic Clones for GPIIIa—Human genomic libraries that were screened include a library constructed from human liver in Charon 4A (kindly provided by Tom Maniatis (13)) and a library constructed from peripheral white cells of a patient with β-thalassemia in EMBLISA (14). Screening was done using cDNA clones for GPIIIa (7) that were labeled with [α-32P]dCTP using DNase-generated calf thymus DNA random primers (15). Positive clones were plaque-purified.

Characterization of the Genomic Clones—Restriction maps of GPIIIa-positive clones were obtained by digestion with one or more restriction enzymes. The DNA fragments were size-fractionated on agarose gels, blotted onto GeneScreen-Plus (Du Pont-New England Nuclear) or Hybond-N (Amersham), and hybridized with radiolabeled GPIIIa cDNA or oligonucleotides to determine the location of the exon sequences.

Subcloning of Restriction Fragments and Sequence Determination—Restriction fragments of the genomic clones were subcloned into M13mp18 and -mp19 bacteriophage or Bluescript plasmid (Stratagene, La Jolla, CA) and sequenced using the dideoxy chain termination techniques (16). Several inserts were completely sequenced using Bal-31-generated mutations (17) or by subcloning restriction fragments as previously described. The intron/exon boundaries and partial intronic sequence around exons 4, 5, 11, and 12 were determined using appropriate oligonucleotides based on GPIIIa cDNA sequence and M13 bacteriophages known to contain that particular exon. Restriction enzyme sites defined by the DNA sequence analysis were also used to localize exons 4, 5, and 12 in their M13 clones.

Determination of the 3'-End of Transcription—To obtain the sequence for the 3'-end of the GPIIIa gene, the HEL cDNA library constructed in the Agt11 phage, from which the original GPIIIa cDNA clones were isolated, was screened with a 1.6-kb 3'-EcoRI fragment of the GPIIIa cDNA (7). Positive clones were then rehybridized to a 1.1-kb HindIII fragment of genomic DNA from the 3'-end of the gene (Fig. 2). Clones containing sequence downstream from the original cDNA clone were isolated, subcloned into M13 bacteriophage, and sequenced.

Data Analysis—Storage and analysis of the DNA sequence data was done using a Macintosh IIX and the DNA inspector Ile program (Textco, West Lebanon, NH) and an IBM PC using the Pastell Sequence Analysis Program (International Biotechnologies Inc., New Haven, CT). The total determined DNA sequence has been submitted to GenBank® data bank.

RESULTS AND DISCUSSION

Twenty-one overlapping GPIIIa-positive clones were identified and isolated after screening 1.8 × 10⁶ plaques. These clones encompassed over 50 kb of DNA. Greater than 30 kilobases of this region was sequenced (Figs. 1 and 2) using the various techniques described of which 3.9 kilobases represents previously published cDNA sequence. More than 80% of the new sequence was determined on both strands and/or in overlapping clones in one orientation. A restriction endonuclease map derived from an analysis of these clones and the regions sequenced are indicated in Fig. 2. Analysis of these clones indicates that the coding region that represents mature GPIIIa protein is contained in 14 exons, which are distributed over approximately a 46 kb region. Exons are numbered using small Roman numerals to indicate that because the exon(s) involving the 5'-untranslated region and the signal peptide have not been determined these exons cannot yet be assigned a definitive number. The first 13 determined GPIIIa exons range in length from 90 to 425 base pairs with the 5' most exon corresponding to the second nucleotide of the first codon of the mature protein. The sequence in the coding regions is identical with that of the corrected cDNA sequence (7, 18) except for three silent substitutions. At 1159 of the cDNA sequence (7), there is an A → C (italicized in exon viii, Fig. 1), which is approximately at base 20,500 of the determined sequence; at position 1549, there is an A → G (italicized in exon ix, Fig. 1, approximately at base 21,500); and at position 1661, there is a G → C (italicized in exon ix, Fig. 1). The first substitution involves a restriction site change creating a TaqI site. This site may account for a previously described TaqI GPIIIa polymorphism (19). The last substitution also involves a restriction site change, eliminating a Smal site.

In the determined sequence, there are 20 complete or incomplete Alu repeats (arrow underline in Fig. 1). Two complete inverted Alu repeats were found near the polyadenylation signal in the 3'-untranslated region (dot underline in Fig. 1 at approximate position 43,500 and circled asterisk in Fig. 2). Unlike GPIIb which contains two GC splicing donors (21), all the splice junctions in GPIIIa contained the expected GT splice donor and AG splice acceptor (underlined in Fig. 1).

The 3'-end of the full length GPIIIa transcript and its polyadenylation site have not been previously determined. To identify additional cDNA clones that may contain additional 3'-sequence, we rescreened our Agt11 HEL cell cDNA library with the previously published 1.6-kb 3'-EcoRI GPIIIa cDNA fragment (7). A clone was isolated that contained an additional 2 kb of the 3'-untranslated region and had a stretch of 7 adenine residues at its 3'-end. The corresponding genomic sequence contained only 5 adenine residues. Although this difference in the number of adenine residues could be a polymorphism or a cloning artifact, we believe that it may represent the polyadenylation site for the following reasons. First, while the AATTTAAA sequence was not found, an AAATTTAAA sequence is present 15 nucleotides upstream of this proposed polyadenylation site. AATTTAAA has been found as the polyadenylation signal for a number of other genes (25). Second, the length of the GPIIIa cDNA based upon the proposed polyadenylation site would be 5.9 kb and is concordant with the size band seen for GPIIIa on Northern blot analysis of HEL cell or platelet RNA. Third, an additional 2.4 kilobases downstream to the proposed 3'-end of the gene has been sequenced, and no AATTTAAA sequence has been found. Thus, it appears that the 3' most exon of GPIIIa is 3618 nucleotides in length and contains 63 nucleotides encoding a portion of the cytoplasmic domain and a 3555-bp 3'-untranslated region, which is 1944 bp longer than the previously published partial 3'-untranslated sequence.

An alternate GPIIIa cDNA sequence was recently published by von Kuppevelt et al. (22). This sequence is identical with the previously published cDNA sequence (6, 7, 18) through comparison of the alternate cDNA sequence and the last intron of the GPIIIa is consistent with the alternate cDNA 3'-end representing nonsplicing at this junction as these two sequences are virtually identical. The one difference between them is in the copy number of a simple repeat that is present just prior to the polyadenylation AATAAA signal (TAA5) in the published cDNA sequence versus (TAA5) in the determined genomic sequence). This difference may represent a polymorphism in that the copy number of simple repeats often varies within a population (23). Alternatively, this difference in the number of repeats may affect the ability to use the alternative polyadenylation site, which by Northern blot analysis should result in a shorter 3.0-kb GPIIIa band. This
FIG. 1—continued

The determined DNA sequence of the GPIIIa gene. The determined DNA sequence is shown in capital letters. Short areas of undetermined sequence are represented as N, while the size of larger unsequenced regions is indicated by an approximation based on restriction fragment sizing. An approximate numbering system is shown in bold at the left hand side of the page. Exons are noted and numbered in small Roman numerals from I to xiv, and the encoded amino acid sequences are indicated below the corresponding sequence. The three silent substitutions noted in the coding region are italicized, as is the alternative C terminus coding sequence 13 amino acid residues (see "Results and Discussion"). The GT splice donor and AG splice acceptor of each exon is underlined. Alu repeats are underlined with an arrow and whether they are complete or incomplete is indicated. Those Alu repeats which occur at the end of sequences are indicated as partially sequenced. The two potential polyadenylation signals are dot underlined.

FIG. 2. Genomic organization of the GPIIIa gene. The restriction sites shown were determined from endonuclease analysis of genomic clones and subfragments. E = EcoRI, B = BamHI, H = HindIII, S = StuI, K = KpnI, and P = PstI. The sequenced areas are shown as solid lines, and the exons are shown as the solid boxed areas. The strategy used in determining the sequence is indicated below as arrows. The polyadenylation sites from the full length GPIIIa transcript and for the alternatively processed product are indicated by a circled asterisk and a plain asterisk, respectively. The scale shown at the bottom is in kilobases.

smaller size mRNA has been noted in some, but not all, previously studied patients (7, 24). In addition, the alternative polyadenylation signal site (dot underline in intron xiii of Fig. 1 at the approximate position 37,700 and asterisk in Fig. 2) is at the 3'-end of a complete direct Alu repeat. Therefore, both GPIIIa mRNAs contain an Alu repeat in their 3'-untranslated regions; the full length transcript having two inverted Alu repeats, while the alternatively processed RNA has a single direct repeat.

GPIIb and GPIIIa proteins appear to be expressed in thrombasthenic patients at approximately the same level (26), and the genes are within 250 kb of each other on the long arm of chromosome 17, with the GPIIIa gene 5' to the GPIIb gene (27). In addition, the 5'-flanking region of GPIIIa has been cloned and appears to lack the typical TATA box promoter of tissue-specific genes (21, 28). Therefore, because it is possible that GPIIb and GPIIIa are transcribed as a single functional unit, we have been interested in identifying the 5'-flanking region for the GPIIIa gene. Multiple genomic librar-
ies each of which had been constructed using different restriction endonucleases and λ phage variants were screened in an attempt to identify the 5’-end of the gene. Most of these libraries were general human genomic libraries that had been previously amplified; however, one unamplified partial Sau3A library in EMBL3A and grown on a recA negative host and one amplified chromosome 17 library (American Type Culture Collection, NIH Repository of Human DNA Probes and Libraries, Rockville, MD) were also screened. A 94-base pair Apol 5’-GPIIIa cDNA fragment (7), a 129-base pair ScaI 5’-GPIIIa cDNA fragment (7), and three GPIIIa cDNA oligonucleotides of 21–24 nucleotides representing the 5’-end of the known cDNA were used as probes without success. Perhaps, the 5’ most genomic region, like 5% of the human genome, is uncloneable, containing unusual sequences that prevent its cloning or that are toxic to λ phage growth (29–31).

In many other reported genes, a correlation has been found between the organization of exons and functional domains in the protein (32). Therefore, we examined the exon distribution and compared it to potential functional domains of GPIIIa deduced from its original cDNA sequence (Fig. 3). Aside from the beginning of the first exon shown, which corresponds to the second base of the first codon of the mature protein, we found no obvious correlation between exon organization and the boundaries of presumed biological domains. The first two cysteine-rich repeats and a portion of the third repeat are located within a single exon (exon ix). The remainder of the third repeat and the entire fourth repeat are found in exon x. The recently described RGD-binding domain based on cross-linking peptides (33) is found in exon iii. The transmembrane domain is found entirely within exon xiii, but sequence for the neighboring extracellular and cytoplasmic regions are also contained within that exon. Integrin receptors are present in invertebrates (34) as well vertebrates, and this long time evolution of these genes may have obfuscated any function-exon organization relationship.

Lanza et al. (35) have reported in abstract form the isolation of genomic DNA coding for GPIIIa. They found that the mature protein was coded for by 12 exons, ranging in length from 87 to 375 nucleotides, and calculated a gene length of 38 kb. We are unable to explain these differences between their preliminary report and our data in terms of overall gene size and in the number and length of individual exons.

Although the complete intron-exon structure of the genes of other members of the integrin superfamily have not yet been published, a recent report (36) described a family with leukocyte adhesion deficiency in whom an abnormal splice junction resulted in the aberrant splicing out of a 90-base pair exon of the β subunit of LFA 1. This exon corresponds to exon vii in the GPIIIa gene, which is also 90 base pairs in length. Therefore, the borders of at least this exon occurs in homologous locations in these two different integrin β genes. Correlation of the organization of GPIIIa with other members of the integrin β family may provide additional insight into their evolution, biological function, and the molecular basis of inherited integrin-related disorders.

Acknowledgments — We would like to thank Robin L. Eisman, Monica Jain, and Catherine Gordon for assistance in these studies.

REFERENCES


FIG. 3. Relationship between coding region and exon distribution. The organization of the cDNA is shown schematized in the upper portion of the figure with the coding region shown as a boxed area. The various proposed functional domains in the coding region are designated in the stick figure. The exons are shown below and are numbered from i through vii. The i refers to the exon(s) containing the 5’-untranslated and the signal peptide regions which have not been cloned. Consequently, the full number of exons is unknown.

8594 Genomic Organization of Platelet Glycoprotein IIIa
Genomic Organization of Platelet Glycoprotein IIIa