Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 4840–4844, June 1991
Biochemistry

Organization of the gene for human erythrocyte membrane protein 4.2: Structural similarities with the gene for the a subunit of factor XIII

(transglutaminase gene family/gene mapping/intron–exon junctions)

CATHERINE KORSGREN* AND CARL M. COHEN*†‡

*Department of Biomedical Research, St. Elizabeth's Hospital, Boston, MA 02135; and †Tufts University School of Medicine, Boston, MA 02111

Communicated by Daniel Branton, February 19, 1991 (received for review October 5, 1990)

ABSTRACT **Human erythrocyte band 4.2 is a major membrane-associated protein with an important, but still undefined, role in erythrocyte survival. We previously sequenced the complete cDNA for band 4.2 and showed that the protein has a strong sequence identity with the transglutaminase family of proteins but lacks transglutaminase activity. Here we have analyzed the genomic organization of band 4.2. The band 4.2 gene is ≈20 kilobases, consisting of 13 exons and 12 introns. Reticulocytes contain two different sized messages for band 4.2, and our results show that the major, smaller, message is produced by alternative splicing within band 4.2 exon I. The upstream region of the gene has several prospective promoter elements arranged in a pattern similar to that of two other erythroid genes, β-globin and porphobilinogen deaminase. Alignment of the band 4.2 amino acid sequence with that of the a subunit of human coagulation factor XIII and division of the sequences into exons reveal a remarkable correspondence, and in most cases identity, in the sizes of the paired exons. Moreover, each corresponding intron of the two genes is of an identical splice junction class. These and other similarities suggest that the gene for band 4.2 is closely related to and possibly derived from that for the a subunit of factor XIII and that the proteins may share common structural and functional properties.**

Human erythrocyte band 4.2 is a major membrane-associated protein of 72 kDa and is present at ≈200,000 copies per cell (1, 2). Band 4.2 is clearly important for normal erythrocyte function since patients whose erythrocytes are deficient in or lack band 4.2 are anemic due to accelerated erythrocyte destruction and have abnormally shaped, possibly fragile, erythrocytes (3–5). In spite of this, however, the exact function of band 4.2 in the erythrocyte remains unclear. We have recently cloned and sequenced the full-length cDNA for band 4.2 (ref. 6; see also ref. 7). We found that the amino acid sequence of human erythrocyte band 4.2 has homology with two transglutaminases, guinea pig liver transglutaminase, and the a subunit of human coagulation factor XIII. The region of most striking identity includes a 49-amino acid domain of band 4.2, which is 69% and 51% identical to the corresponding domains of guinea pig liver transglutaminase and of the a subunit of factor XIII, respectively, and which spans the regions containing the active sites (8, 9) of both enzymes. However, band 4.2 has no demonstrable crosslinking activity when tested *in vitro* (6).

In principle, the sequence similarities of band 4.2 with the transglutaminases could have arisen as the result of convergent evolution, in which each class of proteins evolved common, but as yet undetermined, functions independently, or, more likely, by divergent evolution, in which case the proteins would be expected to have arisen from a common ancestral gene. This question can be resolved by comparing the organization of the gene for band 4.2 with those of the transglutaminases, since a similar gene organization and sequence would imply a common gene ancestor. In this paper, we present the complete gene organization of human erythrocyte band 4.2§ and compare it with that of the a subunit of human factor XIII, the only transglutaminase whose gene structure has been determined (10).

METHODS

Library Screening. Four hundred thousand plaque-forming units of a library constructed from human placental genomic DNA, partially digested with *Sau*3A and cloned in EMBL-3 (Clontech), were screened by hybridization with a ^{32}P-labeled probe made from the full-length human erythrocyte band 4.2 cDNA (see below). Six positive reacting clones with an average insert size of 15 kilobases (kb) were isolated and subcloned in p-Gem4Z (Promega) with the restriction enzyme *Sac* I (New England Biolabs) for restriction enzyme mapping and sequencing. The full-length band 4.2 cDNA was labeled with [^{32}P]dCTP by primer extension using the multiprime DNA labeling system (Amersham). All other partial cDNA fragments used for Southern mapping were labeled in the same fashion.

Analysis of Genomic Subclones. Cesium chloride-banded plasmid DNA was prepared from three of the positive genomic clones—11A, 3B, and 5C (see below). These were digested with restriction endonucleases, fractionated on an agarose gel, transferred to nitrocellulose, and hybridized (11) with each of four different [^{32}P]cDNA probes (see *Results*). Those DNA fragments that hybridized were sequenced by the dideoxynucleotide chain-termination method (12).

Sequence Analysis. Nucleotide sequence was analyzed by the DNASTAR (IntelliGenetics) program. Amino acid sequences were aligned and gapped by using the AALIGN program in DNASTAR.

Amplification of the 5′ Region of Band 4.2 cDNA by Primer Extension. Reticulocyte RNA (13) was primed with an antisense oligonucleotide complementary to band 4.2 cDNA nucleotides +511 to +529,¶ reverse transcribed, and subsequently 3′ tailed with poly(G) using terminal deoxynucleotidyltransferase according to the method of Loh *et al.* (14). Second-strand synthesis and succeeding amplifications by the polymerase chain reaction were primed with a sense

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed.

§The sequence reported in this paper has been deposited in the GenBank data base (accession no. M60298).

¶We have numbered the nucleotides such that nucleotide 1 is the first A of the translation start site. This numbering differs from that used in our previous publication (6). In addition, we have included in our nucleotide and amino acid numbering a 30-amino acid splice in exon I, which was not included in our previous work; see text.

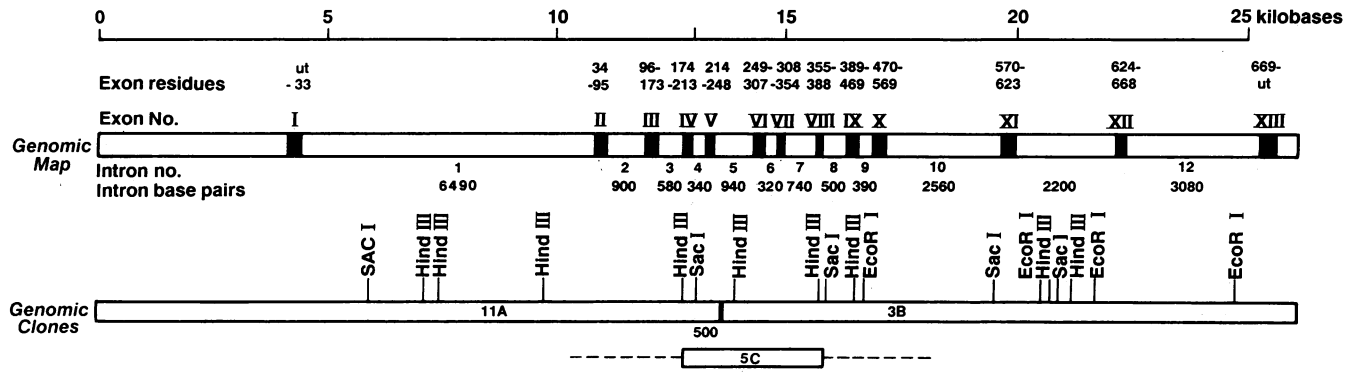

FIG. 1. Genomic organization of the gene to human erythrocyte band 4.2. The 13 exons are represented by solid rectangles and are numbered I–XIII with the included residues enumerated above each (ut, untranslated sequence). The residues have been numbered to include the 30-amino acid splice (see text). The 12 introns are labeled 1–12 and beneath each is designated the approximate size in nucleotides. Below the genomic map is a restriction map of the two nonoverlapping genomic clones, 11A and 3B, in which is included the entire band 4.2 cDNA sequence. Also shown is the overlapping clone 5C. Clone 5C was ≈15 kb but only that portion shown by the open box was mapped.

oligonucleotide complementary to the poly(G) tail and an antisense primer 5′ to that initially used for the reverse transcriptase (nucleotides +268 to +285). Oligonucleotides were synthesized by an Applied Biosystems 391 DNA synthesizer.

RESULTS

Characterization of the Human Band 4.2 Gene. Two nonoverlapping clones isolated from the human placental genomic library contained the complete coding sequence to human erythrocyte band 4.2. The genomic clones were digested with restriction endonucleases and the fragments were analyzed by the method of Southern (11) with four different [^{32}P]cDNA probes made from the full-length band 4.2 cDNA. The four probes had the following composition (nucleotide 1 is the first A of the translation start): nucleotides −187 to +456, +504 to +1308, +1309 to +2197, and +1761 to +2197. Fragments of the genomic clones that hybridized with any one of the four probes were sequenced. In most cases, nonhybridizing fragments containing introns were not fully sequenced, and the approximate sizes of the introns were determined by electrophoresis following restriction endonuclease digestion. Fig. 1 shows that the human erythrocyte band 4.2 gene is ≈20 kb and contains 13 exons representing the total human reticulocyte cDNA. The coding sequence from the genomic DNA is identical to the cDNA sequence (6, 7). Exons I–V were contained within the genomic clone 11A and exons VI–XIII were included in genomic clone 3B. A third overlapping genomic clone, 5C, was mapped with probes made from exons V and VI in order to establish the approximate size of intron 5. The exons ranged in size from 104 to 314 base pairs with an average size of 170 base pairs, while the introns varied from 6.4 to 0.3 kb.‖ Fig. 2*A* shows that all the exon–intron boundaries follow the consensus 5′ donor–3′ acceptor splice junction sequence for eukaryotic genes of gt-ag (15). Fig. 2*B* shows that all three classes of intron, 0, I, and II (16), are represented.

It was previously shown by Sung *et al.* (7) that reticulocytes contain two forms of band 4.2 mRNA, a small form encoding a protein of 691 amino acids, and a larger form, which contains an additional 90 nucleotides following nucleotide +9, encoding a protein of 721 amino acids. We have confirmed the presence of the larger band 4.2 transcript by primer extension of total reticulocyte RNA using a band 4.2-specific primer (data not shown). Further confirmation of the authenticity of the larger transcript was obtained from our analysis of the band 4.2 genomic sequence. Our genomic data show that band 4.2 exon I contains 5′ noncoding sequence, the translation start site, and 99 nucleotides encoding 33 amino acids. These 33 amino acids are identical to the first 33 amino acids of the larger band 4.2 transcript (7). The 3′-most 90 nucleotides of exon I, coding for 30 amino acids, must be removed by splicing to generate the smaller transcript, coding for the 691-amino acid protein (see Fig. 4). Our results therefore show that the smaller message is generated by splicing within exon I, and not by alternative exon usage. With the exception of the 30-amino acid insertion after amino acid 3, the sequence of the larger band 4.2 protein is identical to that of the smaller. Since purified band 4.2 migrates as a single band in SDS/polyacrylamide gels (2) one of the transcripts is expected to predominate over the other in the reticulocyte and, presumably, in the mature erythrocyte. Northern analysis of total human reticulocyte RNA using a full-length cDNA encoding the 691-amino acid band 4.2 form revealed only one band of ≈2.4 kb (data not shown). However, it was not possible to determine whether this band represented the smaller or larger form. To determine which form of the two band 4.2 message sizes predominated, reticulocyte RNA was reverse transcribed with an antisense primer from the 5′ coding sequence (see *Methods*). Electrophoresis of the PCR reaction products in an agarose gel followed by staining with ethidium bromide revealed the presence of two bands (data not shown). By visual inspection, it was clear that the product corresponding to the smaller transcript was present in much greater abundance than was that corresponding to the larger. Both products were subcloned and sequenced to verify that they represented the two transcripts in question. We conclude from this that the smaller gene product of 691 amino acids is the predominant band 4.2 protein seen on SDS/polyacrylamide gels of human erythrocyte ghosts.

The primer-extension experiment described above was also used to derive the nucleotide sequence close to the transcription start site of the band 4.2 gene (Fig. 3*A*). The most 5′ sequences of the two transcripts obtained were identical (verifying that they are derived from a single gene) and extended to nucleotide −214, numbering backwards from the translation start site. However, this position was 12 nucleotides downstream of the most 5′ sequence generated by Sung *et al.* (7), which extended to nucleotide −226. This apparent difference in the 5′ origin of the band 4.2 transcript likely reflects a slight degradation of our reticulocyte RNA and in the following discussion we assume that the transcription start site is at nucleotide −226 (marked by an arrow in Fig. 3*A*). Here, nucleotide 1 is the first A of the translation

‖Limited amounts of intron sequence beyond that shown in Fig. 2 are available from the authors.

A

EXON (SIZE, B.P.)	5' DONOR SEQUENCE	INTRON (SIZE, B.P.)	3' ACCEPTOR SEQUENCE	EXON
	100		101	
I	GGGATGGATGgtaggcgggagatggtggt	1	ttctctctttcccacaacagCCCTGGGTATC	II
*	G M D A	(6490)	L G I	
	286		287	
II	GCACAAACTGgttagtgagtcaggcctaga	2	taatcaccgtgtgactacagGAGAGCAGCCT	III
(186)	A Q T G	(900)	E Q P	
	520		521	
III	TGGAATAGAGgtaagtttggacctgggcca	3	ccctacccacctctcaccagAGGATGCTGTG	IV
(234)	W N R E	(580)	D A V	
	639		640	
IV	TTTGGCCAGgtaccagggccctctagggg	4	gcctgtgtgatccttaacagTTCGAGGGG	V
(119)	F G Q	(340)	F E G	
	744		745	
V	GGTGCCTTGgtaagcgagaggctggggac	5	ccacacatttccccaaacagCTGCATTTT	VI
(104)	G A L	(940)	L H F	
	922		923	
VI	GCTTGCACAGgttggtccccttgtctcctg	6	ctccgcccctcttctcacagTGCTGCGATGC	VII
(178)	A C T V	(320)	L R C	
	1061		1062	
VII	GGCAGAATCTGgtcgagaagcccaacccaga	7	tttctctttgctgcattcagGATCTTCCAG	VIII
(139)	G R I W	(740)	I F Q	
	1165		1166	
VIII	GGAGGTGGAGgtaaagcctaggggtgggtc	8	ccaacctttctctccctcagTCCTGGGGTCC	IX
(104)	G G G V	(500)	L G S	
	1408		1409	
IX	TATCCTGAAGgtacgggccagggaagggcc	9	ttgtttcttcccttctccagGGTCTCTTCAG	X
(243)	Y P E G	(390)	S L Q	
	1708		1709	
X	GCCAACCTGGgtaatgctctatggccccac	10	ccacctgtcatgtgtttcagAAAAGATAATA	XI
(300)	A N L E	(2560)	K I I	
	1869		1870	
XI	GCCATCAAGgtaggcaccagtcctcccct	11	ggtattgcctcttcccacagATGCCAGAG	XII
(161)	A I K	(2200)	M P E	
	2003		2004	
XII	AGCTACAGgtaagtgatggcaggcttct	12	tgccttttccttttatttcagATTCCGTTCA	XIII
(134)	S Y R	(3080)	F R S	

B

	FACTOR XIII		BAND 4.2	
	INTRON	CLASS	INTRON	CLASS
	B	I	1	I
	C	I	2	I
	D	I	3	I
	E	O	4	O
	F	O	5	O
	G	I	6	I
	H	II	7	II
	I+J	I+0	8	I
	K	I	9	I
	L	I	10	I
	M	O	11	O
	N	II	12	II

FIG. 2. (*A*) Nucleotide sequence of the exon donor/acceptor splice junctions. The nucleotide sequences of the splice junctions are shown, and the number of each 3′ and 5′ boundary nucleotide of each exon is indicated above the nucleotide. The exons are numbered I–XIII and the intervening introns are numbered 1–12. The approximate size in base pairs of introns and exons is indicated below each and in parentheses. *, The exact size of the exon has not been determined. (*B*) Intron splice junction classes. The class designation of each of the introns of the band 4.2 gene is shown. Also shown is the class of the corresponding intron of the gene for the a subunit of human factor XIII (see text for discussion of intron–exon alignment).

start; nucleotides and amino acids are numbered to include the unspliced sequence of exon I (see Fig. 4 and ref. 7).

In Fig. 3*B*, we compare the upstream sequence of the band 4.2 gene with the genomic sequences for two other erythrocyte proteins, β-globin and porphobilinogen deaminase (21). For comparative purposes only, in Fig. 3*B* we have renumbered the band 4.2 upstream sequence such that nucleotide −226, the transcription start site, is numbered +1. The figure shows that the upstream region of the band 4.2 gene contains five elements that are similar in sequence to the upstream elements of the genes for β-globin and porphobilinogen deaminase (21). The elements are spaced a similar distance from the transcription start site and have a similar relative spacing and order. The positions of these sequences in the upstream region of the band 4.2 gene are also shown in Fig. 3*A* by underlining. This comparison has enabled us to identify five possible regulatory elements in the band 4.2 gene: (*i*) a possible TATA element (22); (*ii*) a short G+C-rich domain, which could be an Sp1 binding site (20); (*iii*) a likely (but nonstandard) CAAT box (22); (*iv*) a CAAC box (18, 21); (*v*) two GF-1 binding domains (17), one at −23 to −28 (shown only in Fig. 3*A* at positions −249 to −254; note numbering difference between Fig. 3 *A* and *B*, as discussed above) and one at −173 to −178 (Fig. 3*B*). At this time, the identification of these elements as having a regulatory function in band 4.2 gene expression is speculative and future studies will be directed to assessing the function of these elements experimentally.

Comparison of the Organization of the Gene for Human Erythrocyte Band 4.2 with That of the a Subunit of Human Coagulation Factor XIII. The gene for the a subunit of human factor XIII is 160 kb and has 15 exons and 14 introns (10), while the gene for band 4.2 is only 20 kb and contains 13 exons and 12 introns. To compare the exon sizes of the two proteins, we aligned the amino acid sequences of the proteins using the AALIGN program in DNASTAR and then divided the aligned sequences at their appropriate exon boundaries (Fig. 4). It should be noted that exon I of the a subunit of factor XIII contains only untranslated sequence, so the exon pairing starts with exon II of factor XIII. Exon I of band 4.2 contains 5′ untranslated sequence (data not shown) and sequence coding for 33 amino acid residues. Of these, residues 3–33 are the ones removed by splicing to produce the major 691-amino acid transcript, as discussed above, and are indicated by a superscribed line. Examination of Fig. 4 reveals that the degree of sequence identity between the paired exons varies considerably. However, examination of the sizes of the paired exons reveals that, with only one exception, each exon of band 4.2 is very similar and in many cases identical in size to the exon of the a subunit of factor XIII with which it is paired. The one exception is that the a subunit of factor XIII has an intron between exons X and XI (intron position marked by arrow in Fig. 4), whose combined sequence corresponds to band 4.2 exon IX. It should also be noted that while the last exon of band 4.2 (XIII) and the last exon of the a subunit of factor XIII (XV) encode a comparable number of amino acid residues, they are dissimilar in nucleotide number since each contains different numbers of noncoding residues.

We have also compared the splice junction classes (16) of the corresponding introns of the genes for band 4.2 and the

A

```
-461  tggatatgggactcctgtttgtggcctgcctctcttccagagtcttccagtgcagtccag  -402
-401  ataacgatctctgagagtctcttgagaccactgaagcaagaggcttcaaggtttgttcta  -342
-341  ggcttctctagttggcaggtcttccccgatcccaccccctcccccagggctgccaggcgt  -282
-281  attcctgcctgggcctcctggcacaagagataaaatgaacagggttactcccagtaactg  -222
-221  gtccaggagataccagcagagagggagtaggagagaagaaacatgtcagggtgctcacag  -162
```

B

```
Band 4.2 gene          -178 -173  -88      -79    -66     -58    -47   -41    -27  -21
                         AGATAA-85--ATCCCACCCC--13--TGCCAGGCG--10--TGGGCCT--15--ATAAAA

Porphobilinogen        -186 -181  -106     -97    -90     -82    -79   -74    -30  -25
deaminase gene           AGATAA--75-GCCCCACCCT--7---GGCCAGGCT-3----TGGGCCT-44---ACAACT

Human ß-globin                 -94      -85    -78     -70                    -31  -26
gene                             GCCACACCCT--7---GGCCAATCT----------39-------ATAAAA

Promoter element:        GF-1        CAAC Box        CAAT Box        Sp1 Binding     TATA Box

Consensus sequence       AGATAA[1]   GCCTCACCC[2]    GGTCAATCT[3]    CCGCCC[4]       TATAAT[2]
                         T   G         A               C
```

FIG. 3. Upstream sequences of the gene for human erythrocyte band 4.2. (*A*) Upstream sequence of the gene for band 4.2 is shown with the prospective promoter elements (described in text and in Fig. 3*B*) underlined. The numbering is based on the A of the translation start site being designated as nucleotide +1 (see text). The arrow indicates nucleotide −226, the most 5′ cDNA sequence obtained (7), and the putative transcription start site (see text for discussion). (*B*) The upstream sequences for three human erythrocyte genes are compared. Similar sequences of each are aligned above the labeled respective promoter elements. The 5′ and 3′ nucleotides of each putative element are numbered and the distance between each element is given in nucleotides. The band 4.2 gene upstream sequence has been numbered from the putative transcription start site (nucleotide −226 of Fig. 3*A*) to facilitate the comparison of the three erythroid gene upstream sequences. The consensus sequence superscripts indicate the following references: 1, ref. 17; 2, ref. 18; 3, ref. 19; 4, ref. 20.

a subunit of human factor XIII (Fig. 2*B*). In every case, the corresponding intervening introns are of the same splice junction class, the only difference being in intron J of the a subunit of factor XIII, which has no corresponding intron in the band 4.2 gene. The similarity in sequence, exon size, and intron class suggests that the genes for these two proteins are closely related and may have evolved one from the other or from a common ancestral gene.

DISCUSSION

We have shown that human erythrocyte band 4.2 is the product of a 20-kb gene containing 13 exons and 12 introns. The exon–intron boundaries all have the eukaryotic consensus splice junction sequence (15), and all three classes of intron are represented by the 12 band 4.2 introns (16). Exon I of band 4.2 is subject to alternative splicing within its coding sequence, removing the 90 most 3′ nucleotides, resulting in the presence in reticulocytes of two distinct message sizes and, presumably, protein products. We have found that the lower molecular weight spliced transcript is the predominant human reticulocyte species by comparing the relative abundance of the two products generated by primer extension of the 5′ region of band 4.2 RNA from reticulocytes. Both transcripts have a common transcription start site, confirming that they are derived from a single gene. Our analysis of the band 4.2 gene also reveals the presence of several cis-element consensus sequences starting ≈20 nucleotides upstream from the transcription start site. These elements have a sequence and arrangement similar to the upstream elements for two other erythroid genes, β-globin and porphobilinogen deaminase, suggesting the use of common cis elements in these three erythroid genes.

In principle, the gene for band 4.2 could be related to the gene for the a subunit of factor XIII by either divergent or convergent evolution (the latter could also include exon

```
                  No. bp   Exon
Band 4.2            *       I    1 MGQGEPSQRSTGLAGLVAAPAASPVFIKGSGMD 33
Factor XIII       148       II   1 SETSRTAFGGRRAVPPNNSNAAEDDLPTVELQGVVPRGVNLQ 42
Band 4.2          186       II  34     ALGIKSCDFQAARNNEEHHTKALSSRRLFVRRGQPFTIILYFRAPVRAFLPALKKVALTAQT 95
Factor XIII       189       III 96 EFLNVTSVHLFKERWDTNKVDHHTDKYENNKLIVRRGQSFYVQIDL---SRPYDPRRDLFRVEYVI 106
Band 4.2          234       III 96 GEQPSKINRTQATFPISSLGDRKWWSAVVEERDAQSWTISVTTPADAVIGHYSLLLQVSGRKQLL------LGQFTLLFNPWNR 173
Factor XIII       252       IV 107 GRYPQENKGTYIPVPIVSELQSGKWGAKIVMREDRSVRLSIQSSPKCIVGKFRMYVAVWTPYGVLRTSRNPETDTYLLFNPWCE 190
Band 4.2          119       IV 174 EDAVFLKNEAQRMEYLLNQNGLIYLGTADCIQAESWDFGQ 213
Factor XIII       119        V 191 DDAVYLDNEKEREEYVLNDIGVIFYGEVNDIKTRSWSYGQ 232
Band 4.2          105        V 214 FEGDVIDLSLRLLSK-DKQVEKWSQPVHVARVLGAL 249
Factor XIII       108       VI 233 FEDGILDTCLYVMDRAQMDLSGRGNPIKVSRVGSAM 266
Band 4.2          178       VI 250 LHFLKEQRVLPTPQTQATQEGALLNKRRGSVPILRQWLTGRGRPVYGQAWVLAAVACT 307
Factor XIII       175       VII 267 VNAKDDEGVLVGSWDNIYAYGVPPSAWTGSVDILLEYRSSENPVRYGQCWVFAGVFNT 324
Band 4.2          139       VII 308 VLRCLGIPARVVTTFASAQGTGGRLLIDEYYNEEGLQNGEQQRGRIW 354
Factor XIII       139       VIII 325 FLRCLGIPARIVTNYFSAHDNDANLQMDIFLEEDGNVNSKLTKDSVW 371
Band 4.2          104       VIII 355 IFQTSTECWMTRPALPQGYDGWQILDPSAPNGGG 388
Factor XIII       104        IX 372 NYHCWNEAWMTRPDLPVGFGGWQAVDSTPQENSD 405
Band 4.2          242        IX 389 VLGSCDLVPVRAVKEGTVGLTPAVSDLFAAINASCVVWKCCEDGTLELTDSNTKYVGNNISTKGVGSDRCEDITQNYKYPE 469
Factor XIII     89+154  X+XI 406 GMYRCGPASVQAIKHGHVCFQFDAPFVFAEVNSDLIYITAKKDGTHVVENVDATHIGKLIVTKQIGGDGMMDITDTYKFQE 486
Band 4.2          300         X 470 GSLQEKEVLERVEKEKMEREKDNGIRPPSLETASPLYLLLKAPSSLPLRGDAQISVTLVNHSEQEKAVQLAIGVQAVHYNGVLAAKLWRKKLHLTLSANL 569
Factor XIII       288       XII 487 G--QEEERLA-LETALMYGAKKPLNTEGVMKSRSNVDMDFEVENAV-LGKDFKLSITFRNNSHNRYTITAYLSANITFYTGVPKAEFKKETFDVTLEPLS 582
Band 4.2          161        XI 570 EKIITIGLFFSNFERNPPENTFLRLTAMATHSESNLSCFAQEDIAICRPHLAIK 623
Factor XIII       161       XIII 583 FKKEAVLIQAGEYMGQLLEQASLHFFVTARINETRDVLAKQKSTVLTIPEIIIK 636
Band 4.2          134       XII 624 MPEKAEQYQPLTASVSLQNSLDAPMEDCVISILGRGLIHRERSYR 668
Factor XIII       137       XIV 637 VRGTQVVGSDMTVTVQFTNPLKETLRNVWVHLDGPGVTRPMKK-MFR 682
Band 4.2            *       XIII 669 FRSVWPENTMCAKFQFTPTHVGLQRLTVEVDCNMFQNLTNYKSVTVVAPELSA 721
Factor XIII       1688       XV 683   EIRPNSTVQWEEVCRPWVSGHRKLIASMSSDSLRHVYGELDVQIQRRPSM 732
```

FIG. 4. Comparison of corresponding exons of the genes for band 4.2 and the a subunit of factor XIII. The complete amino acid sequences of band 4.2 and the a subunit of factor XIII were aligned by using the AALIGN program in DNASTAR. The aligned sequences were divided into exons using our data for band 4.2 and the data of Ichinose and Davie (10) for factor XIII. The number of base pairs (bp) in each exon is shown; *, the exact size of the exon is not known. Amino acid numbers are indicated at the beginning and end of each exon. The line above band 4.2 residues 4–33 denotes the residues absent from the principal band 4.2 translation product. The arrow in factor XIII subunit a exon X+XI indicates the position of the additional intron (see text). The arrow in factor XIII subunit a exon VII indicates the active site cysteine, for which band 4.2 has an alanine substituted. •, Amino acid identity between exons.

shuffling). The fact that the sequence identities between the two proteins are limited to half or fewer of the paired exons shown in Fig. 4 could be an argument for exon shuffling between these two genes or between these genes and another, as yet unidentified, gene. In view of the similarity in gene structure (see *Results*), a more likely explanation is that one of the two genes was derived from the other or that the two were derived from a common ancestor. The lineage of these proteins can be further explored by comparing the overall sequence identity of band 4.2 with two transglutaminases of human origin, the a subunit of factor XIII (23, 24) and endothelial cell transglutaminase (25). Analysis of these sequences by the ALP3 program (Bionet) revealed that the a subunit of factor XIII is 36% identical to endothelial cell transglutaminase but only 25% identical to band 4.2, while endothelial cell transglutaminase is 32% identical to band 4.2. This suggests that endothelial cell transglutaminase is more similar to the a subunit of factor XIII than is band 4.2, while band 4.2 is more similar to endothelial cell transglutaminase than to the a subunit of factor XIII. Ordering the proteins in this way suggests that band 4.2 may have evolved from endothelial cell transglutaminase, which in turn evolved from the a subunit of factor XIII. In this context, it may be of some significance that the gene for the a subunit of human factor XIII is considerably larger than that for band 4.2 (160 kb versus 20 kb for band 4.2) due mostly to the larger sizes of its introns (10) and has a somewhat larger coding sequence (732 amino acids versus 721 amino acids for band 4.2), suggesting greater antiquity (26, 27).

The exon pairing that results from the optimal alignment of the amino acid sequences of the two proteins pairs exon II of the a subunit of factor XIII with exon I of band 4.2 (see Fig. 4). While these exons have no sequence identity, they do have several features in common. In both proteins, most of the amino acids encoded by these exons are subject to elimination, albeit by two distinct mechanisms. As noted above, the predominant form of band 4.2 is generated by alternative splicing of the nucleotides within exon I encoding 30 of 33 amino acids. The a subunit of factor XIII on the other hand is subject to posttranslational proteolysis of its first 37 amino acids encoded by exon II (28). While the genes differ in several respects, one prominent difference relates to the pairing of band 4.2 exon IX with the a subunit of factor XIII exons X and XI, between which the a subunit of factor XIII has an intron. If the genes for band 4.2 and the a subunit of factor XIII are closely related, then at some time in evolutionary history this intron was either deleted from the band 4.2 gene or was inserted into the gene for the a subunit of factor XIII. Examples of both types of splicing events have been reported (29–32). In the overall comparison of the gene for human erythrocyte band 4.2 with the gene for the a subunit of human factor XIII, it is evident that the coding sequences are less similar than is the organization of the genes (exon size and intron splice junction class). This suggests the possibility that there has been greater selective pressure to retain the genomic organization than to preserve sequence identity. The significance of the similarities in genomic organization and/or amino acid sequence may be clarified when further data regarding the function of erythrocyte band 4.2 and intracellular transglutaminases become available.

The authors would like to thank Ms. Lucille Paul for preparation of the manuscript and Ms. Joan Joos for preparing the figures. The authors also thank Dr. V. Gentile and P. Davies for sharing their cDNA sequence of human endothelial cell transglutaminase with us prior to publication. This work was supported by National Institutes of Health Grant HL 37462.

1. Korsgren, C. & Cohen, C. M. (1988) *J. Biol. Chem.* **263,** 10212–10218.
2. Korsgren, C. & Cohen, C. M. (1986) *J. Biol. Chem.* **261,** 5536–5543.
3. Rybicki, A. C., Heath, R., Wolf, J. L., Lubin, B. & Schwartz, R. S. (1988) *J. Clin. Invest.* **81,** 893–901.
4. Ideguchi, H., Nishimura, J., Nawata, H. & Hamasaki, N. (1990) *Br. J. Haematol.* **74,** 347–353.
5. Ghanem, A., Pothier, B., Marechal, J., Ducluzeau, M. T., Morle, L., Alloisio, N., Feo, C., Ben Abdeladhim, A., Fattoum, S. & Delaunay, J. (1990) *Br. J. Haematol.* **75,** 414–420.
6. Korsgren, C., Lawler, J., Lambert, S., Speicher, D. & Cohen, C. M. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 613–617.
7. Sung, I. A., Chien, S., Chang, L. S., Lambert, K., Bliss, S. A., Bouhassira, E. E., Nagel, R. L., Schwartz, R. S. & Rybicki, A. C. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 955–959.
8. Lorand, L. & Conrad, S. M. (1984) *Mol. Cell. Biochem.* **58,** 9–35.
9. Folk, J. E. (1980) *Annu. Rev. Biochem.* **49,** 517–531.
10. Ichinose, A. & Davie, E. W. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 5829–5833.
11. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
12. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
13. Goosens, M. & Kan, Y. W. (1981) *Methods Enzymol.* **76,** 805–817.
14. Loh, E. Y., Elliott, J. F., Cwirla, S., Laniee, L. L. & Davis, M. M. (1989) *Science* **243,** 217–220.
15. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4853–4857.
16. Sharp, P. A. (1981) *Cell* **23,** 643–646.
17. Evans, T., Reitman, M. & Felsenfeld, G. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 5976–5980.
18. Dierks, P., van Ooyen, A., Cochran, M. D., Dobkin, C., Rever, J. & Weissman, C. (1983) *Cell* **32,** 695–706.
19. Bonoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8,** 127–142.
20. Dynan, W. S. & Tjian, R. (1983) *Cell* **35,** 79–87.
21. Chretien, S., Dubart, A., Beaupain, D., Raich, N., Grandchamp, B., Rosa, J., Goossons, M. & Romeo, P. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6–10.
22. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383.
23. Grundmann, V., Amunn, E., Zehlmeissl, G. & Kupper, H. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8024–8028.
24. Ichinose, A., Hendrickson, L. E., Fujikawa, K. & Davie, E. W. (1986) *Biochemistry* **25,** 6900–6906.
25. Gentile, V., Saydak, M., Chiocca, E. A., Akande, O., Birckbichler, P., Lee, K. N., Stein, J. P. & Davies, P. J. A. (1991) *J. Biol. Chem.* **266,** 478–483.
26. Craik, C. S., Choo, Q., Swift, G. H., Quinto, C., MacDonald, R. J. & Ruttor, W. J. (1984) *J. Biol. Chem.* **259,** 14255–14264.
27. Jaworski, C. J. & Piatigorsky, J. (1989) *Nature (London)* **337,** 752–754.
28. Takagi, T. & Doolittle, R. (1974) *Biochemistry* **13,** 750–756.
29. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20,** 555–566.
30. Lonberg, N. & Gilbert, W. (1985) *Cell* **40,** 81–90.
31. Rogers, J. (1985) *Nature (London)* **315,** 458–459.
32. Blake, C. C. F. (1979) *Nature (London)* **277,** 598.