Glycoconjugate Journal (1995) **12**: 857–864

Isolation and characterization of a pseudogene related to human core 2 β-1,6-*N*-acetylglucosaminyltransferase

MARTI F.A. BIERHUIZEN*, KENTARO MAEMURA‡ and MINORU FUKUDA§

Glycobiology Program, La Jolla Cancer Research Foundation, La Jolla, California 92037, USA

Received 26 April 1995, revised 17 May 1995

In a previous study, we isolated genomic clones encoding core 2β-1,6-*N*-acetylglucosaminyltransferase (C2GnT) and blood group IGnT and proposed that these two genes were produced from a common ancestral gene by duplication, diversion and intron insertion. In the present study, we have isolated a pseudogene which is highly related to the gene of *C2GnT*. The sequence analysis of this pseudogene indicated that the pseudogene was produced by duplication of a common precursor gene for *C2GnT*. These results taken together strongly suggest that the ancestral gene was first duplicated and one of the duplicated genes directly evolved into the *IGnT* gene. The other duplicated gene was further duplicated to produce the *C2GnT* gene and the pseudogene.

Keywords: core 2β-1,6-*N*-acetylglucosaminyltransferase, pseudogene, gene evolution

Introduction

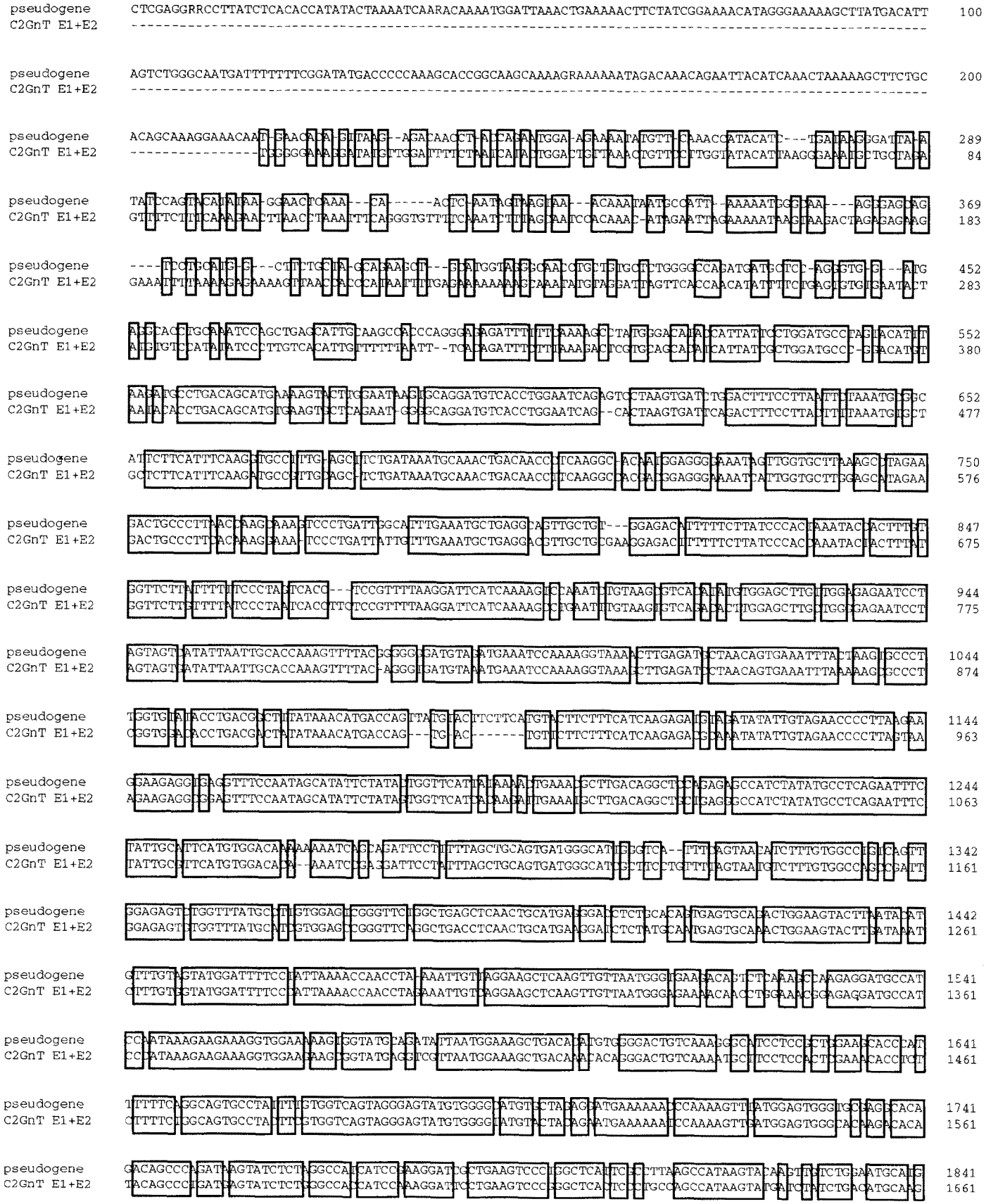
Cell surface carbohydrates are often characteristic of different cell lineage and different stages of differentiation [1–3]. Among these carbohydrates, those synthesized by β-1,6-*N*-acetylglucosaminyltransferases are particularly specific to cell-types. For example, a common *O*-glycan branch is formed by core 2 β1,6-*N*-acetylglucosaminyltransferase (C2GnT) and the addition of the β-1,6-*N*-acetylglucosaminyl linkage to the Galβ1 → 3GalNAc backbone results in the formation of the hexasaccharide, NeuNAcα2 → 3Galβ1 → 3 (NeuNAcα2 → 3Galβ1 → 4GlcNAcβ1 → 6)GalNAc [4, 5]. When β-1,3-*N*-acetylglucosaminyltransferase (extension enzyme) is also present, the β-1,6-*N*-acetylglucosaminyl branch is further extended to form poly-*N*-acetyllactosaminyl side chains in O-glycans [6, 7]. The poly-*N*-acetyllactosaminyl side chain is often modified to have a sialyl Le^x terminus, NeuNAcα2 → 3Galβ1 → 4(Fucα1 → 3)GlcNAc → R [4, 6]. A dramatic increase in core 2 branching was observed when human T-lymphocytes were activated from the resting state [5] and this increase leads into the expression of sialyl Le^x in activated T-lymphocytes [8]. The maturation of thymocytes from cortical to medullary thymus is associated with the turning off of C2GnT [9]. In pathological conditions such as leukaemia and immunodeficiency, leukocytes express an increased amount of C2GnT [10–13]. Moreover, highly metastatic tumour cells express much more branched oligosaccharides than low metastatic counterparts [14]. These results indicate that the increase of core 2 branches in hematomalignancy, and tumours in general, reflect the cell surface carbohydrates of immature cells. The results also suggest that core 2-based carbohydrates are involved in metastatic spreading of tumour cells.

During development of human erythrocytes, the blood group I-branching enzyme (IGnT), is substantially increased. While linear poly-*N*-acetyllactosamine (Galβ1 → 4GlcNAcβ1 → 3)$_n$ is expressed on fetal erythrocytes, it is replaced by branched, I-active poly-*N*-acetyllactosamine, Galβ1 → 4GlcNAcβ1 → 3(Galβ1 → 4GlcNAcβ1 → 6) on adult erythrocytes [15, 16]. Because of this importance, we have cloned cDNAs encoding the human *C2GnT* and *IGnT* genes [17, 18].

*Present address: Department of Hematology, Erasmus University, Rotterdam, The Netherlands.

‡Present address: Department of Internal Medicine, Osaka Medical College, Takatsuki, Japan.

§To whom correspondence should be addressed.


pseudogene
C2GnT E1+E2

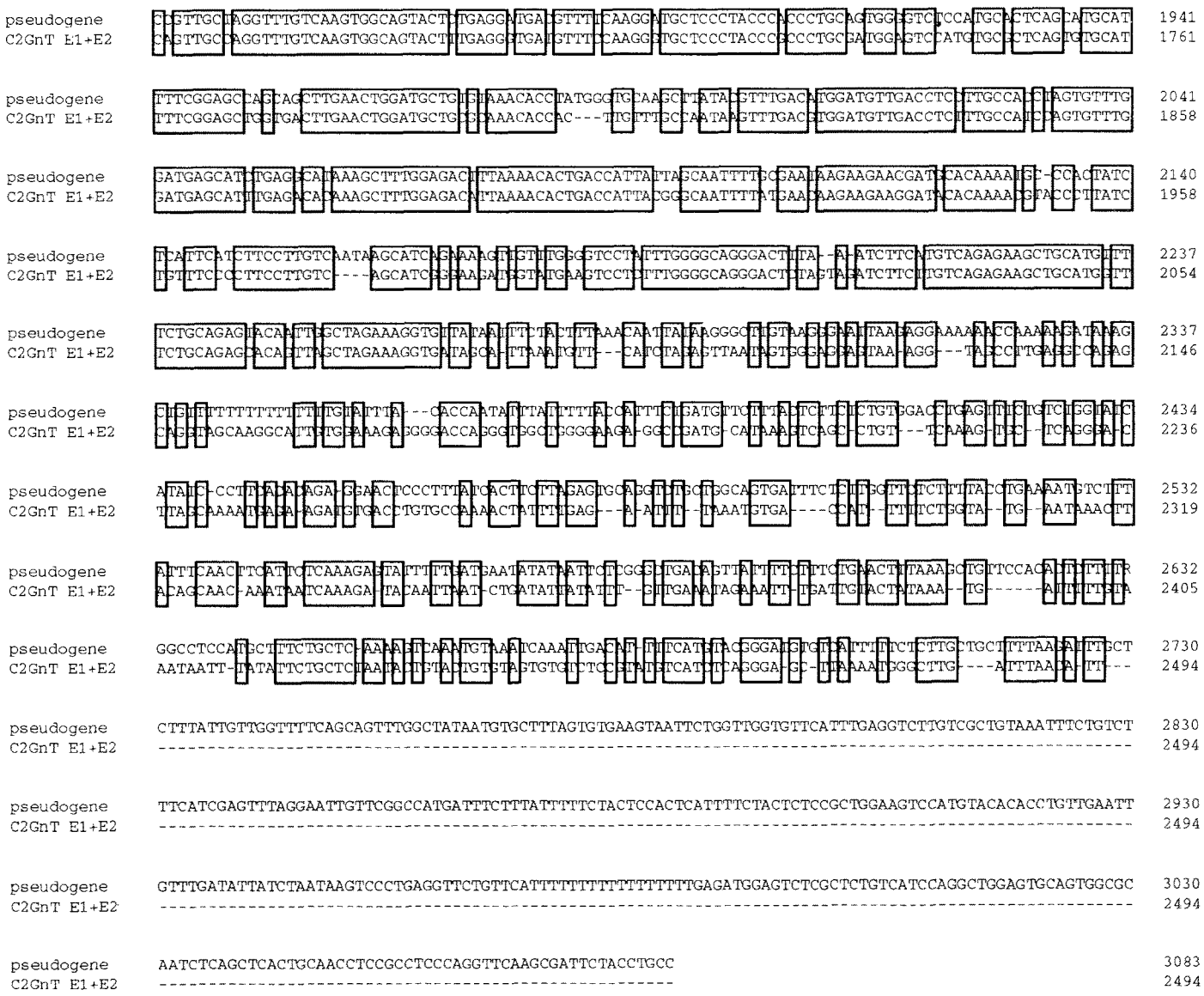


Figure 1. Comparison of nucleotide sequences of the pseudogene and C2GnT cDNA. *C2GnT* cDNA sequence was constructed by combination of exon 1 and exon 2 sequences obtained in the previous study [19]. The initiation methionine codon is nucleotides 792–794 in the pseudogene sequence. The pseudogene sequence was obtained in the *Xho* I–*Xba* I digested genomic DNA. Boxed sequences are identical.

More recently we have isolated genomic clones harbouring these two genes and found that the genomic organization of these two enzymes is diverse despite the fact that they share three regions of extensive homology in their catalytic domains; the highly homologous region B is split between exons 1 and 2 in the *IGnT* gene while the same region is encoded entirely by exon 2 in the *C2GnT* gene. Based on these results, we proposed that the common ancestral gene was first duplicated and then each duplicated gene evolved into the *C2GnT* and *IGnT* genes by intron insertion and divergence following the duplication [19].

During the above genomic cloning, we also isolated a genomic sequence that is highly related to the *C2GnT* gene. We describe here the characterization of this pseudogene. The sequence analysis of the pseudogene supports the conclusion that this pseudogene and *C2GnT* derived from one of the duplicated ancestral genes and *IGnT* evolved from the other duplicated gene.

Materials and methods

Isolation of genomic clones

A human placental genomic DNA library, constructed in λEMBL3, was purchased from Clontech, Inc. The library was screened with cDNA fragments specific for *C2GnT*

and *IGnT* after labelling by random oligonucleotide priming [20] using [α-^{32}P] dCTP and a kit from Boehringer Mannheim. The cDNA fragments were obtained by PCR amplification [21] of cDNA sequences which encode the beginning of the stem region of the protein to the 3′-untranslated region as described previously [19]. A probe for Southern hybridization was made by ^{32}P-labelling using random oligonucleotide priming [20].

Southern blot analysis

Phage DNAs were digested with various restriction enzymes and subjected to Southern blotting and hybridization as described previously [18]. Briefly, the blots were hybridized with ^{32}P-labelled cDNA inserts of *IGnT* or *C2GnT*. The hybridization was in 6 × SSPE, pH 7.4, 0.5% SDS, 50 μg ml^{-1} of denatured, sheared salmon sperm DNA containing 50% formamide at 42 °C for 16 h [22]. The blot was then washed several times in 2 × SSPE, pH 7.4, 0.5% SDS at room temperature for periods of 10 min and subsequently exposed to Kodak X-Omat AR film.

DNA sequencing

The DNA fragments of interest were subcloned into pcDNAI (Invitrogen) and nucleotide sequences were determined by the dideoxy chain termination method [23] utilizing T7 DNA polymerase (US Biochemical Corp.) and [α-^{35}S] dATP (DuPont – New England Nuclear). In order to judge if any portions of exon sequences were present, various primers used for cDNA sequencing [17, 18] were used as described [19]. Once the phage DNA yielded a sequence for a particular portion of *C2GnT* or *IGnT* exon sequence, sequencing was extended further using newly synthesized oligonucleotides based on the obtained sequence data. All of the oligonucleotides used were synthesized on an Applied Biosystems DNA synthesizer.

Results and discussion

Isolation and characterization of the human C2GnT *and* IGnT *genes*

As shown previously, eight genomic clones were initially isolated from about 1 × 10^6 plaques of the placental genomic DNA library. Among these, clone 20 was found to contain the *C2GnT* gene while clones 2, 3, 7, 9, 13 and 14 were found to contain various parts of the *IGnT* genes. The sequencing of those genomic clones and an additional clone containing *C2GnT* exon 1 showed the following results.

C2GnT is coded by two exons, of which the second exon encodes the whole translation product. In contrast, the complete coding sequence for *IGnT* is divided over three exons. As shown previously, *C2GnT* and *IGnT* share three regions of extensive homology in their catalytic domains [18]. However, the high homologous region B is split between exons 1 and 2 in the *IGnT* gene while the same region is encoded entirely by exon 2 in the *C2GnT* gene [19].

Isolation and characterization of a pseudogene related to C2GnT

During these studies, we also isolated clone 6 which hybridized with *C2GnT*. The nucleotide sequence of clone 6 differs from that of *C2GnT* or *IGnT*. However, when the nucleotide sequence of clone 6 was tested for homology with known sequences, it became evident that clone 6 contains a nucleotide sequence which is highly related to *C2GnT* (Fig. 1).

The sequence corresponding to the initiation methionine of *C2GnT* lies in nucleotides 792–794. The nucleotide sequence of clone 6 can be aligned by regarding this codon as the initiation methionine. The resultant translated sequence, however, becomes out of frame starting from nucleotides 990–992 (Fig. 2). Moreover, the nucleotide sequence of clone 6 is very similar to the *C2GnT* cDNA sequence but no sequence corresponding to *C2GnT* intron 1 can be found (see Fig. 1). These results strongly indicate that this newly isolated gene is a pseudogene which is highly related to *C2GnT*.

The striking feature of the nucleotide sequence of this gene is, however, that it lacks a polyadenylation signal after nucleotide 2539. The corresponding sequence in *C2GnT* contains a polyadenylation signal and polyadenylation takes place after nucleotide 2539 (see Fig. 1). These results strongly argue against the hypothesis that this pseudogene was produced from *C2GnT* mRNA.

The pseudogene contains a 5′-upstream sequence homologous to one of the *Kpn* I repetitive sequences (*Kpn* 13) and this sequence is present from nucleotide 1 to nucleotide 800 (Fig. 3). *Kpn* I repetitive sequence is the major human long interspersed repeated DNA sequence [24]. *Kpn* I repeats often have deletions and rearrangements at the 5′-end. The number of copies in the total genome is 5 × 10^4 to 1 × 10^4 for the 3′-end and 0.4 × 10^4 to 2.0 × 10^4 copies for the 5′-end. In this pseudogene, deletions took place apparently at the 3′-end of the *Kpn* I repeat. Interestingly, this *Kpn* I repeat is superimposed by an *Alu* repetitive sequence [25] from nucleotide 1 to 295 (Fig. 4). *Alu* repeat sequences represent the major human short interspersed DNA sequence and nearly 1 × 10^6 copies of the *Alu* repeat are present in the human genome [25]. We have analysed the 5′-sequence upstream from the exon 1 of *C2GnT*. However, there is no *Kpn* I or *Alu* sequence corresponding to the upstream sequence of the pseudogene. These results suggest that the *Alu* or *Kpn* I sequence was introduced after the ancestral gene was

Figure 2. Comparison of translated amino acid sequences of the pseudogene and *C2GnT*. Each set of three lines (from top to bottom) shows the nucleotide sequence of the pseudogene, its deduced amino acid sequence and *C2GnT* amino acid sequence. In order to maximize the homology between two predicted amino acid sequences, gaps are allowed. Stop codons and frame shifts are denoted by asterisks and slashes, respectively, in the pseudogene amino acid sequence.

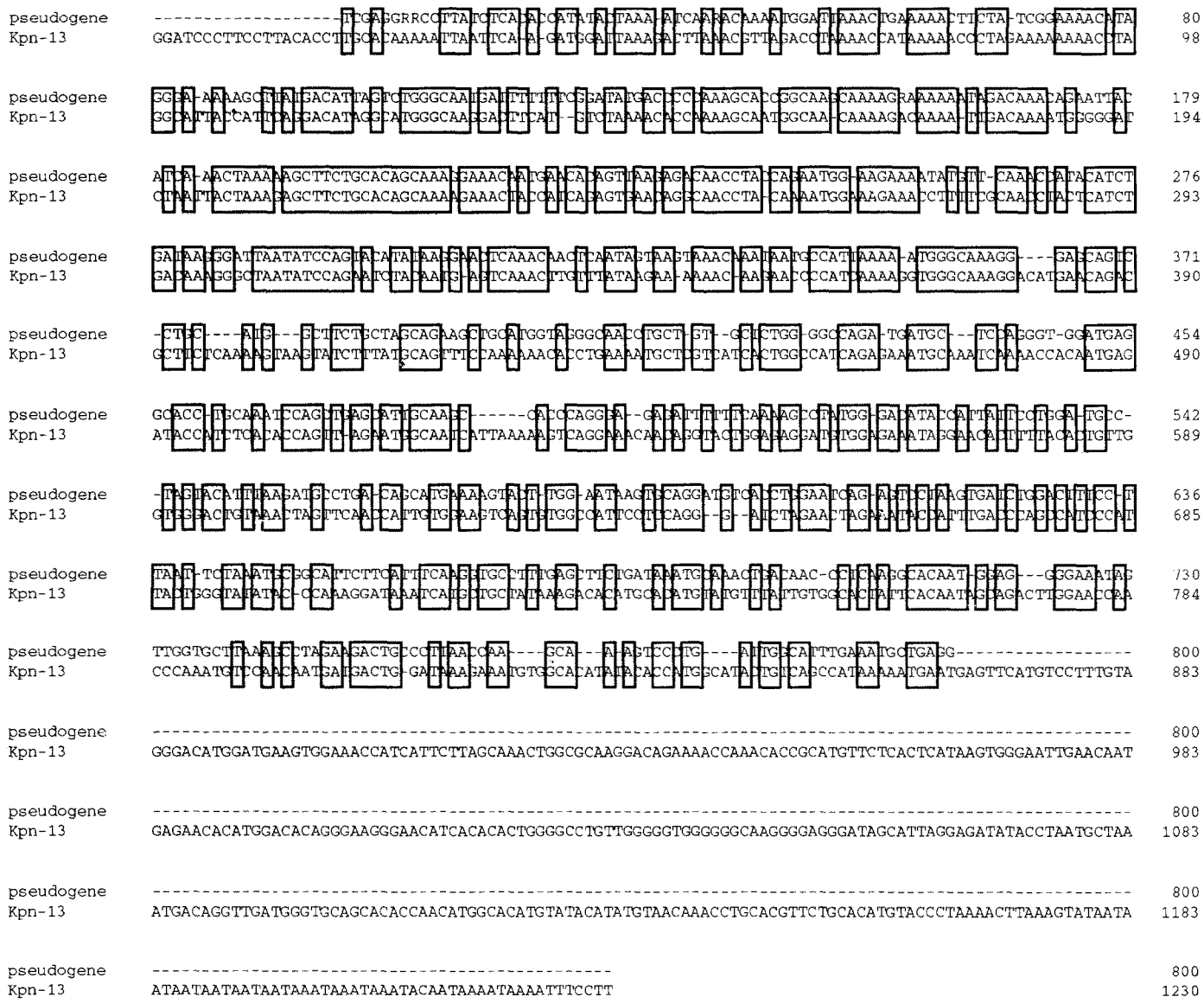


Figure 3. Comparison of the nucleotide sequence of the pseudogene and *Kpn*-13 repetitive sequence. The homology between these two sequences does not exist after nucleotide 800 in the pseudogene sequence. Thus 376 nucleotides in the 3′-region of the *Kpn*-13 sequence are missing. The homology is apparently absent right after initiation methionine (nucleotides 792–794) starts.

duplicated. Alternatively, the duplicated gene was rearranged by homologous recombination through *Alu* or *Kpn* I repeats as shown for the glycophorin genes [26].

In previous studies, we showed that *C2GnT* and *IGnT* are related to each other and they contain homologous sequences in regions in the catalytic domains, named A, B and C [18]. The sequence analysis of *C2GnT* and *IGnT* genes demonstrated that the highly homologous region B is split between exons 1 and 2 in the *IGnT* gene, while the same region is entirely coded by exon 2 in *C2GnT* gene. The other highly homologous regions, A and C, are also encoded by exon 2 in the *C2GnT* gene, while they are encoded by exon 1 and exon 3, respectively, in the *IGnT* gene (Fig. 5). Moreover, both *C2GnT* and *IGnT* genes reside at the band of q21 of chromosome 9. These results exclude the possibility that the *IGnT* and *C2GnT* were formed through exon shuffling [19]. These results taken together strongly suggest that the common ancestral gene was duplicated and one resultant gene directly evolved into *IGnT* after divergence and intron insertion. The other gene was further duplicated to produce *C2GnT* gene and the pseudogene (Fig. 5).

Similar results were obtained on α-2,6-sialyltransferases. Comparison of the deduced amino acid sequence demonstrated two regions of significant homology among

```
pseudogene  ----------------------------------------------------
Alu seq.    GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGA    50

pseudogene  ----------------------------------------------------
Alu seq.    GGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACAT   100

pseudogene  ------------------------------------CTCGAGGRRCCTTA    14
Alu seq.    GGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCG-GGCGTGGTG   149

pseudogene  TCTCACACCATATAATTAAAATCAARACAAAATGGATTAAACTGAAAAACT    64
Alu seq.    GCGCGCGCCTGRTAGTCCCAGCTACTCGGGA-GGCTGAGGCGGGAGAATC   198

pseudogene  TCTAATCGGAAAACATAGGGAAAAAGCT-TA-TGACATTAG-TCTGGCAA   111
Alu seq.    GCT-T--GAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCAC   245

pseudogene  TG-ATTTTTTTCGGATATGAC-CCCCAAAGCACCGGCAAGCAAAAGRAAA   159
Alu seq.    TGCACTCCAGCCTG-GGCGACAGAGCGAGACTCCGTC--TCAAAAAAAAA   292

pseudogene  AAATAGACAAACAGAATTACATCAAACTAAAAAGCTTCTGCACAGCAAAG   209
Alu seq.    AAA-------------------------AAAA-----------AA---   301

pseudogene  GAAACAATGAACACAGTTAAGAGACAACCTACCAGAATGGAAGAAAATAT   259
Alu seq.    ----------------------------------------------------   301
```

Figure 4. Comparison of the nucleotide sequences of the pseudogene and *Alu* repetitive sequence. *Alu* repetitive sequence has homology with the upstream sequence of the pseudogene.

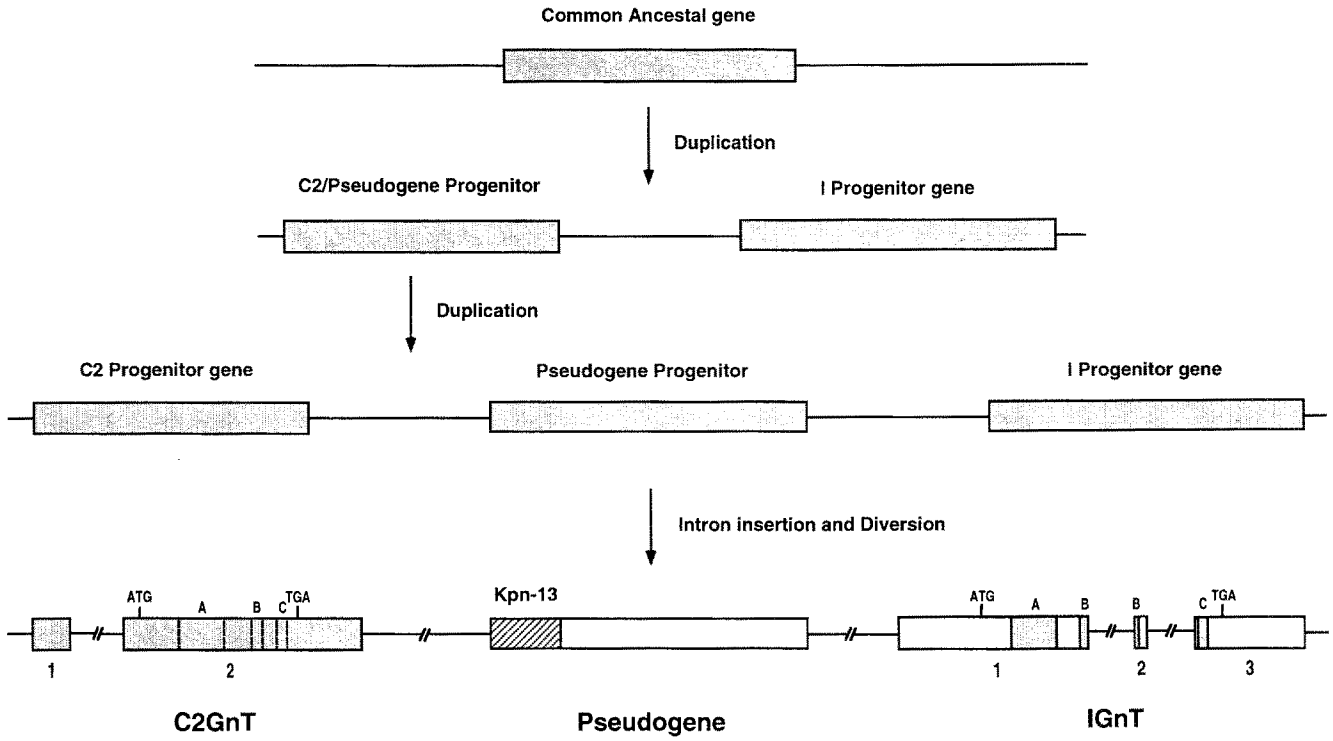


Figure 5. Evolutionary pathway of the *C2GnT* gene, the *IGnT* gene and the pseudogene. A common ancestral gene was duplicated and the resultant gene directly evolved into *IGnT* gene after intron insertion and diversion. The other gene was further duplicated to produce *C2GnT* progenitor gene and the pseudogene. The *C2GnT* progenitor gene evolved into *C2GnT* gene by intron insertion and diversion.

different sialyltransferases. One of them, sialyl motif L, corresponds to residues 178–225 in rat Galβ1 → 4GlcNAc α-2,6-sialyltransferase and was recently found to be the binding site for CMP-NeuNAc [27]. The residues 178 to 225 are, however, split between exon 2 and exon 3 [28, 29], supporting the conclusion that those homologous regions were not brought together by exon shuffling.

The studies on the genomic organization of glycosyltransferases revealed that there are two different types of genomic organization. One is represented by the *C2GnT* gene, in which the entire coding region is coded by one exon. These include *N*-acetylglucosaminyltransferase I [30], and fucosyltransferases III-VI [31]. The other is represented by the *IGnT* gene, in which the coding regions are split by several exons. These include β-1,4-galactosyltransferase [32], α-2,6-sialyltransferase [28, 29] and α-1,3-galactosyltransferase [33]. The present study strongly suggests that these seemingly different gene organizations among different glycosyltransferases were most likely produced from the common ancestral genes by gene duplication, diversion and intron insertion. Further studies are needed on the *C2GnT* and *IGnT* gene family to test this hypothesis by studying the chromosome localization of the pseudogene and the genomic organization of *C2GnT* and *IGnT* in the lower animal kingdom.

Acknowledgements

We thank Dr Kiyohiko Angata for useful discussion and Ms Bobbi Laubhan for secretarial assistance. This work was supported by Grants CA33000 and CA33895 from the National Cancer Institute.

References

1. Fukuda M (ed). (1992) *Cell Surface Carbohydrates and Cell Development* pp. 1–329. Boca Raton, FL: CRC Press.
2. Feizi T (1985) *Nature* **314**: 53–57.
3. Fukuda M (1985) *Biochim Biphys Acta* **780**: 119–50.
4. Fukuda M, Carlsson SR, Klock JC, Dell A (1986) *J Biol Chem* **261**: 12796–806.
5. Piller F, Piller V, Fox RI, Fukuda M (1988) *J Biol Chem* **263**: 15146–50.
6. Maemura K, Fukuda M (1992) *J Biol Chem* **267**: 24379–86.
7. Bierhuizen MFA, Maemura K, Fukuda M (1994) *J Biol Chem* **269**: 4471–79.
8. Ohmori K, Takada A, Yoneda T, Burna K, Harashima K, Tsuyuoka K, Hasegawa A, Kannagi R (1993) *Blood* **81**: 101–11.
9. Baum LG, Pang M, Perillo NL, Terry W, Uittenbogaart C, Fukuda M, Seilhammer JJ (1995) *J Exp Med* **181**: 877–87.
10. Piller F, Le Desit F, Weinberg KI, Parkman R, Fukuda M (1991) *J Exp Med* **173**: 1501–10.
11. Higgins EA, Siminovitch KA, Zhuang D, Brockhausen I, Dennis JW (1991) *J Biol Chem* **266**: 6280–90.
12. Saitoh O, Piller F, Fox RI, Fukuda M (1991) *Blood* **77**: 1491–99.
13. Brockhausen I, Kuhns W, Schacter H, Matta KL, Sutherland DR, Baker MA (1991) *Cancer Res* **51**: 1257–63.
14. Yousefi S, Higgins E, Daoling Z, Pollex-Kruger A, Hindsgaul O, Dennis JW (1991) *J Biol Chem* **266**: 1772–82.
15. Fukuda M, Fukuda MN, Hakomori S (1979) *J Biol Chem* **254**: 3700–3.
16. Fukuda M, Fukuda MN (1984) In *The Biology of Glycoproteins* (RJ Ivatt ed.) pp. 183–234. New York, Plenum Press.
17. Bierhuizen MFA, Fukuda M (1992) *Proc Natl Acad Sci USA* **89**: 9326–30.
18. Bierhuizen MFA, Mattei M-G, Fukuda M (1993) *Genes Dev* **7**: 468–78.
19. Bierhuizen MFA, Maemura K, Kudo S, Fukuda M (1995) *Glycobiology* **5**: 417–25.
20. Feinberg A, Vogelstein B (1983) *Anal Biochem* **132**: 6–13.
21. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) *Science* **239**: 487–91.
22. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
23. Sanger F, Nicklen S, Coulson AF (1977) *Proc Natl Acad Sci USA* **74**: 5463–67.
24. Sun L, Paulson KE, Schmid CW, Kadyk L, Leinwand L (1984) *Nucleic Acids Res* **12**: 2669–90.
25. Schmid CW, Jelinek WR (1982) *Science* **216**: 1065–70.
26. Kudo S, Fukuda M (1989) *Proc Natl Acad Sci USA* **86**: 4619–23.
27. Datta AK, Paulson JC (1995) *J Biol Chem* **270**: 1497–1500.
28. Svensson EC, Soreghan B, Paulson JC (1990) *J Biol Chem* **265**: 20863–68.
29. Wang X, O'Hanlon TP, Young RF, Lau JTY (1990) *Glycobiology* **1**: 25–31.
30. Hull E, Sarkar M, Spruijt MPN, Hoppener JWM, Dunn R, Schacter H (1991) *Biochem Biophys Res Commun* **176**: 608–15.
31. Weston BW, Smith PL, Kelly RJ, Lowe JB (1992) *J Biol Chem* **267**: 24575–84.
32. Hollis GF, Douglas JG, Shaper NL, Shaper JH, Stafford-Hollis JM, Evans RJ, Kirsch IR (1989) *Biochem Biophys Res Commun* **162**: 1069–75.
33. Yamamoto F, McNeill PD, Hakomori S (1991) *Biochem Biophys Res Commun* **175**: 986–94.