Cloning of a Human UDP-N-Acetyl-α-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase That Complements Other GalNAc-Transferases in Complete O-Glycosylation of the MUC1 Tandem Repeat* (Received for publication, June 30, 1998)

Eric Paul Bennett‡, Helle Hassan‡, Ulla Mandel‡, Ekatarina Mirgorodskaya§, Peter Roepstorff‡, Joy Burchell, Joyce Taylor-Papadimitriou, Michael A. Hollingsworth‡, Gerard Merkx‡‡, Ad Geurts van Kessel‡¶, Hans Eiberg‡‡, Rudi Steffensen§§, and Henrik Clausen‡¶¶

From the ‡Faculty of Health Sciences, School of Dentistry, Copenhagen, Denmark, the ‡Department of Molecular Biology, University of Odense, Denmark, the ‡Imperial Cancer Research Fund, London WC2A 3PX, United Kingdom, the ‡Eppeley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, Omaha, Nebraska 68198, the **University Hospital Nijmegen, Department of Human Genetics 6500HB Nijmegen, The Netherlands, the ‡‡Faculty of Health Sciences, Genetics Institute, University of Copenhagen, 2200 N Copenhagen, Denmark, and the §§Regional Center for Blood Transfusion, Aalborg Hospital, Aalborg, Denmark

A fourth human UDP-GalNAc-polypeptide N-acetylgalactosaminyltransferase, designated GalNAc-T4, was cloned and expressed. The genomic organization of GalNAc-T4 is distinct from GalNAc-T1, -T2, and -T3, which contain multiple coding exons, in that the coding region is contained in a single exon. GalNAc-T4 was placed at human chromosome 12q21.3-q22 by in situ hybridization and linkage analysis. GalNAc-T4 expressed in S9F cells or in a stably transfected Chinese hamster ovary cell line exhibited a unique acceptor substrate specificity. GalNAc-T4 transferred GalNAc to two sites in the MUC1 tandem repeat sequence (Ser in GVTSA and Thr in PDTR) using a 24-mer glycopeptide with GalNAc residues attached at sites utilized by GalNAc-T1, -T2, and -T3 (TAPPAHGVTSA-PDTR-PAGSTAPPA, GalNAc attachment sites underlined). Furthermore, GalNAc-T4 showed the best kinetic properties with an O-glycosylation site in the P-selectin glycoprotein ligand-1 molecule. Northern analysis of human organs revealed a wide expression pattern. Immunohistochemistry with a monoclonal antibody showed the expected Golgi-like localization in salivary glands. A single base polymorphism, G1516A (Val to Ile), was identified (allele frequency 34%). The function of GalNAc-T4 complements other GalNAc-transferases may represent a major determining factor for sites of O-glycan attachments.

O-Glycosylation of MUC1 has attracted attention because it is altered in cancer cells with smaller and fewer O-glycans (12–14). The change in O-glycosylation leads to the exposure of cancer-associated peptide epitopes within the tandem repeat region of MUC1, although the exact molecular nature of this phenomenon is still unknown (15, 16). Analysis of the in vitro O-glycosylation properties of various GalNAc-transferase preparations including purified recombinant GalNAc-T1, -T2, and -T3 suggests that only three of the potential five sites in the repeat were glycosylated (AHGVTSA-PDTR-PAGSTAPPA, in vitro glycosylation sites underlined). However, recently Muller et al. (17) elegantly established that all five sites can be occu-

* This work was supported by the Danish Cancer Society, the Mizutani Foundation, the Velux Foundation, the Danish Medical and Natural Science Research Councils, the Novo Nordisk Foundation, National Institutes of Health Grant 1 RO1 CA66234, funds from the EU Biotech 4th Framework, and the Dutch Cancer Society. 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 for Blood Transfusion, Aalborg Hospital, Aalborg, Denmark.

** The abbreviations used are: GalNAc-transferase, UDP-N-acetyl-α-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase; GalNAc-T1, -T2, and -T3 represents human GalNAc-transferases, GenBank accession numbers X85018, X85019, X92689, respectively (3, 4); EST, expressed sequence tags; RACE, rapid amplification of cDNA ends; PCR, polymerase chain reaction; bp, base pair(s); HPLC, high performance liquid chromatography; CHO, Chinese hamster ovary; MALDI, matrix-assisted laser desorption/ionization; PAGE, polyacrylamide gel electrophoresis.

†† To whom the correspondence should be addressed: School of Dentistry, Nørre Alle 20, DK-2200 Copenhagen N, Denmark. Tel.: 45-35326835; Fax: 45-35326505; E-mail: henrik.clausen@odont.ku.dk.

‡‡ This paper is available on line at http://www.jbc.org

30472
This paper is available on line at http://www.jbc.org
plied in purified milk MUC1. Two explanations for this discrepancy are: 1) that in vitro O-glycosylation assays fail to reflect the in vivo function of GalNAc-transferases, or 2) that additional GalNAc-transferases have substrate specificity for the two last sites in the MUC1 repeat. The present study presents evidence supporting the latter explanation.

Previously, a sequence motif shared between GalNAc-T1 and -T2 was used to identify and clone GalNAc-T3 (4). Two candidate cDNA sequences containing the GalNAc-transferase motif were identified: one was the GalNAc-T3 gene. Here, we report the cloning and expression of the second identified cDNA, which was designated GalNAc-T4 (1). A putative murine orthologue also designated GalNAc-T4 was recently reported by Hagen et al. (5). The recombinant human GalNAc-T4 had a highly restricted acceptor substrate specificity, which is distinct from that of previously reported GalNAc-transferases. Thus, GalNAc-T4 was found to complement GalNAc-T1, -T2, and -T3, in O-glycosylation of the MUC1 tandem repeat by transferring GalNAc to two sites in the MUC1 repeat not utilized by other known GalNAc-transferases. Importantly, GalNAc-T4 showed preference for the MUC1 glycopeptide, GalNAc\(\text{TAP24}\), which was previously glycosylated with GalNAc-T2.

**Experimental Procedures**

**Identification and Cloning of cDNA for GalNAc-T4**—A cDNA sequence (TE4) with extensive similarity to the GalNAc-transferase motif was previously identified following reverse transcriptase-PCR of MKN45 mRNA using degenerate primers (4). The PCR product was cleaved by \\textit{BstNI} to remove GalNAc-T1 and -T2 derived products. Uncleaved product was isolated using the prepare-A-gene kit (Bio-Rad) and cloned into the pT7E3U19 vector (Pharmacia). A rapid cDNA library screening strategy was performed as described previously (4). A human salivary gland \\textit{agt11} library (CLONTECH) was amplified and aliquoted into 40 sublibraries, which were screened to identify phage clones containing TE4. Fourteen sublibraries found to contain TE4 \\textit{agt11} clones were further assayed by PCR using EBHC201 (5'-GGCCGATGCCGCGGGACCATATGCTG) or EBHC203 (5'-GAGCGGTTGCTTGTAGAATGTTG) primers combined with primers based on flanking \\textit{agt11} vector sequence to estimate lengths of cDNA inserts for selection of sublibraries with the largest 5'-sequences. Amplifications were: 35 cycles of 95°C for 45 s; 53°C for 1 s; 72°C for 2 min. One sublibrary (number 22) generated a 3' PCR product (EBHC201/agt11 REV) of approximately 750 bp, and one sublibrary (number 34) generated a 5' PCR product (EBHC203/agt11 REV) of approximately 800 bp. Additional 5' sequence was obtained by 5' RACE using primers combined with primers based on flanking \\textit{agt11} vector sequence to estimate lengths of cDNA inserts for selection of sublibraries with the largest 5'-sequences. Amplifications were: 35 cycles of 95°C for 45 s; 53°C for 1 s; 72°C for 2 min. One sublibrary (number 22) generated a 3' PCR product (EBHC201/agt11 REV) of approximately 750 bp, and one sublibrary (number 34) generated a 5' PCR product (EBHC203/agt11 REV) of approximately 800 bp. Additional 5' sequence was obtained by 5' RACE using primers combined with primers based on flanking \\textit{agt11} vector sequence to estimate lengths of cDNA inserts for selection of sublibraries with the largest 5'-sequences. Amplifications were: 35 cycles of 95°C for 45 s; 53°C for 1 s; 72°C for 2 min. One sublibrary (number 22) generated a 3' PCR product (EBHC201/agt11 REV) of approximately 750 bp, and one sublibrary (number 34) generated a 5' PCR product (EBHC203/agt11 REV) of approximately 800 bp. Additional 5' sequence was obtained by 5' RACE using primers combined with primers based on flanking \\textit{agt11} vector sequence to estimate lengths of cDNA inserts for selection of sublibraries with the largest 5'-sequences. Amplifications were: 35 cycles of 95°C, 20 s; 55°C, 1 s; 72°C, 30 s) followed by Kpn1 restriction analysis was performed on DNA samples from 206 healthy blood donors, in order to perform the allele frequencies.

**In Situ Hybridization to Metaphase Chromosomes**—Fluorescence in situ hybridization was performed on normal human lymphocyte metaphase chromosomes, using procedures described previously (18). Briefly, P1 DNA was labeled with digoxigenin-14-dUTP (Boehringer Mannheim) using the BioNick labeling system (Life Technologies). The labeled DNA was precipitated with ethanol in the presence of herring sperm DNA. A total of 200 ng of P1 DNA was precipitated with 5μl of herring sperm DNA (Life Technologies). A solution was hybridized in 12 ml of hybridization solution (2× SSC, 10% dextran sulfate, 1% Tween 20, and 50% formamide, pH 7.0). Prior to hybridization the probe was heat-denatured at 80°C for 10 min, chilled on ice, and incubated at 37°C to allow re-annealing of highly repetitive sequences. After denaturation of the slides, probe incubations were carried out under a 18× 10 cm cover slip in a moist chamber for 45 h. Immunodetection of the probe was achieved using sheep anti-digoxigenin (fluorescein isothiocyanate) (Boehringer) and donkey anti-sheep (fluorescein isothiocyanate) (Jackson Laboratories) antibodies. For evaluation of the chromosomal slides, a Zeiss epifluorescence microscope equipped with appropriate filters for visualization of fluorescein isothiocyanate was used. Hybridization signals and 4,6-diamidino-2-phenylindole-counterstained chromosomes were transformed into pseudo-colored images using image analysis software. For precise localization and chromosome identification 4,6-diamidino-2-phenylindole-converted banding patterns were generated using the BDS-image software package (ONCOR).

**Linkage Analysis of GalNAc-T4 with the Base Pair G1516A Substitution**—From a normal family with an average of 10 children from the Copenhagen Family Bank (19) were analyzed for the G/A polymorphism at bp position 1516 using the PCR assay with EBHC300/EBHC307. To confirm in situ localization on 12q21.3-q22, a set of DNA markers with known physical localization were used. The order and distances for these were drawn from the Genethon Linkage map (20: D12S590 (18.8)-D12S580 (10.8)-D12S81 (7.5)-D12S101). The physical map of these markers are: D12S80, 12p13.2-q11.21; D12S81, 12q21; and D12S101, 12q22. Lod scores were calculated with the LIPEP software (21).

**Expression of GalNAc-T4 in S9 Cells**—Expression constructs designed to contain amino acid residues 32-578 of the coding sequence of the putative GalNAc-T4 gene were prepared by genomic PCR on the two different P1 clones using the primer pair EBHC318 (5'-AGCGGATCC-CTTTCATGCGCTCCGAGGAGGCG) and EBHC307 with BamHI restriction sites (Fig. 1). These PCR products were cloned into a BamHI site of the expression vector pAcGP67 (Pharmingen), and the expression construct was sequenced. The constructs, pAcGP67-GalNAc-T4\textsubscript{506V} and pAcGP67-GalNAc-T4\textsubscript{506V}, were designed to yield a putative soluble form of the GalNAc-T4 protein with an N-terminal end positioned immediately C-terminal to the potential transmembrane domain and including the catalytic domain. Control constructs pAcGP67-GalNAc-T1-sol, pAcGP67-GalNAc-T2-sol, pAcGP67-GalNAc-T3-sol, and pAcGP67-O\textsubscript{72} were prepared as described previously (3, 4, 22).

A full coding expression construct was prepared by PCR of P1 clone 2 (GalNAc-T4\textsubscript{506V}) using primer pair EBHC320 (5'-AGCGGATCCCGGAGGGCGAGGCGAGGCG) and EBHC307 with BamHI restriction sites (Fig. 1). These PCR products were cloned into a BamHI site of the expression vector pAcGP67 (Pharmingen). Co-transfection of S9 cells with pAcGP67-constrs or pVL-constrs and Baculo-Gold\textsuperscript{TM} DNA was performed according to the manufacturer's specifications. Briefly, 0.4 μg of construct was mixed with 0.1 μg of Baculo-Gold DNA and co-transfected in S9 cells in 24-well plates. Ninety-six hours post-transfection recombinant virus was amplified in 6-well plates at dilutions of 1:10 and 1:50. Titers of amplified virus was estimated by titration in 24-well plates with monitoring of GalNAc-transferase activities. Initial transfection assays were performed on supernatants of S9 cells in 6-well plates infected with first or second amplified virus. Titers representing end point dilutions giving optimal enzyme activities were used. Transfection assays of the full coding expression construct was performed by extracting washed cells in 1% Triton X-100 as described previously (7).

**Purification of GalNAc-T4 and Production of Monoclonal Antibodies**—pAcGP67-GalNAc-T4\textsubscript{506V} was expressed in High Five\textsuperscript{TM} cells grown in serum-free medium in upright roller bottles shaking at 140 rpm in 27°C waterbaths. GalNAc-T4\textsubscript{506V} was purified essentially as described previously (8), using successive chromatography on Amber- 12P (A95, Sigma), DEAE-Sephadex (Pharmacia), and S-Sepharose Fast Flow (Pharmacia) columns, followed by ion exchange on Mini-S\textsuperscript{TM} (PC 3.2/3, Pharmacia) using the Smart System (Pharmacia), and peak fractions used for immunizing BALB/c mice as described previously (5-10 μg/subcutaneous injection three times with one intravenous boost) (9). Monoclonal antibodies were selected and characterized by immunocytology on Sf9 cells infected with various GalNAc-transferase expression constructs and selective immunoprecipitation of active Gal-
N-ac-T4 activity. Immunofluorescence staining of frozen sections of submandibular glands was performed as described previously (9), and monoclonal antibody PABN2, reacting with MUC5B (23), was used as marker of mucus cells.

**Stable Expression of Secreted GalNAc-T4 in Chinese Hamster Ovary Cells**—The insert of pAcGP67-GalNAc-T4sol was excised and cloned into a modified pCDNA3 vector (Invitrogen), which included 19 amino acids of the interferon secretion signal sequence (24). CHO-K1 (ATCC) was transfected using 0.2 μg of DNA and 5 μl of Lipofectin (Invitrogen) in subconfluent 6-well plates according to the manufacturer’s protocol. After 48 h the medium was changed and 400 μg/ml G418 was added. At 72 h 10–20% of the wells were trypsinized and the percentage of cells expressing GalNAc-T4 was evaluated by immunocytochemistry as described previously (9), using a novel anti-GalNAc-T4 monoclonal antibody, UH6. Based on the frequency of positive cells the residual transfectant cells were trypsinized and plated in 96-well plates. Two rounds of screening and cloning by limiting dilution using immunoreactivity with UH6 were performed and clones reaching over 50% positive cells were selected, and tested for level of secreted enzyme in supernatant of confluent cultures.

**Polypeptide GalNAc-Transferase Assay**—Standard assays were performed in 50-μl total reaction mixtures containing 25 mM cacodylate (pH 7.5), 10 mM MnCl₂, 0.1–0.25% Triton X-100, 150–200 μg/ml UDP-[¹⁴C]GalNAc (4,000 cpm/nmol) (Amersham), and 0.06–1 μM acceptor peptides. Peptides were synthesized by ourselves, by Carlbiotech (Copen-hagen), or Neosystems (Strasbourg), and quality was ascertained by amino acid analysis and mass spectrometry. Peptides Muc1a, Muc1b, and TAP24, were derived from the tandem repeat of human MUC1 (25), Muc2, Muc4, and Muc7 from the tandem repeat of human MUC2, MUC5AC, and MUC7, respectively (26–28). GalNAc-T4 was produced by *in vitro* glycosylation using GalNAc-T4 as described previously (8). Products were quantified by scintillation counting after chromatography on Dowex-1, octadeyl silica cartridges (Bakerbond), or HPLC (PC3/23 or mRPC C2/C18 SC2.1/10 Pharmacia, Smart System). Products produced by *in vitro* glycosylation were in most cases also confirmed by mass spectrometry.

**Structure Determination**—Using matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF) the mass spectra were acquired on either Voyager-DE or Voyager-Elite MALDI time of flight mass spectrometers (Perseptive Biosystem Inc.), equipped with delay extraction. The MALDI matrix was a 9:1 mixture of 2,5-dihydroxybenzoic acid (2,5-DHB) 25 g/liter and 2-hydroxy-5-methoxybenzoic acid 25 g/liter (Aldrich) dissolved in 2:1 mixture of 0.1% trifluoroacetic acid in water and acetonitrile. Samples dissolved in 0.1% trifluoroacetic acid to a concentration of approximately 2 pmol/ml were prepared for analysis by placing 1 μl of sample solution on a probe tip followed by 1 μl of matrix.

**Northern Hybridization**—Multiple tissue Northern (MTN) blots were obtained from CLONTECH. The soluble expression construct contained nucleotides 92–1757 was used as the GalNAc-T4 probe. The probe was random prime labeled using [α-³²P]dCTP (Amersham) and Oligo labeling kit (Pharmacia). Blots were probed as described previously (4), and washed at 5 × at 42 °C with 2 × SSC, 0.1% SDS, once with 0.5 × SSC, 1% SDS, and once at 55 °C with 0.1 × SSC, 0.1% SDS, in a mini-hybridization oven (HYBAID).

**RESULTS**

**Identification and Cloning of cDNA for GalNAc-T4**—Previously, reverse transcriptase-PCR was performed on mRNA from a variety of human organs and cell lines, using a pair of degenerate primers (EBHC100/EBHC106) corresponding to sequences flanking a putative GalNAc-transferase motif. Two novel sequences were identified, TE3 and TE4, with approximately 80% similarity in sequence to GalNAc-T1 and -T2 (4). The reverse transcriptase-PCR product obtained from the human gastric carcinoma cell line MKN45 was cleaved with BstNI (known to cleave GalNAc-T1 and -T2 sequences at non-conserved restriction sites), and the remaining uncleaved product was subcloned and sequenced. Eight out of 40 independent clones contained sequences similar but not identical to GalNAc-T1 and -T2. Six of these clones were derived from GalNAc-T3 (TE3) (4), and two clones contained a novel sequence designated TE4.

Cloning and sequencing of the complete coding sequence of GalNAc-T4 was achieved by a combination of PCR screening of 40 sublibraries from a human salivary gland Agt11 library, 5’ RACE, and genomic P1 cloning. The combined sequences contained an open reading frame of 1734 bp (GenBank accession number Y08564) (Fig. 1). The entire coding sequence was confirmed by sequencing of P1 clones in both directions. The deduced sequence of GalNAc-T4 is predicted to be a type II transmembrane protein with a hydrophobic retention signal in...
residues 11–31. A BLAST search of the EST data base (GenBank/NCBI) with the coding region of GalNAc-T4 only detected one human GalNAc-T4 EST, which is in contrast to several other GalNAc-transferases that are highly represented by ESTs.

GalNAc-T4 is highly similar to the other GalNAc-transferases in the COOH-terminal region (amino acid residues 85–559, GalNAc-T1; 105–571, GalNAc-T2; 151–633, GalNAc-T3; and 103–578, GalNAc-T4), and there is conservation of 12 cysteine residues (Fig. 1) (1). The NH2-terminal regions vary considerably in length and have no significant similarity.

**Genomic Organization and Chromosomal Localization of GalNAc-T4**—Three P1 clones each covering the entire coding sequence of GalNAc-T4 were isolated. Sequencing of all three P1 clones showed that the entire coding region of GalNAc-T4 was contained in a single exon.

Fluorescence in situ hybridization revealed that the GalNAc-T4 gene resides at human chromosome 12q21.3-q22 (Fig. 2). No specific hybridization signals were observed at other chromosomal sites. A total of 20 cells in metaphase were analyzed. Further confirmation of the chromosomal location was achieved by linkage analysis using the PCR assay for the polymorphism at position 1516 combined with chromosome 12 microsatellite markers (Table I). Analysis of 10 families yielded significant Lod score (Z > 3) between D12S80, D12S81, D12S101, and the GalNAc-T4 polymorphism. One recombination to the marker D12S81 was detected in an intercross mating (Z = 10.50 at $\theta_{(M-F)} = 0.02$). A marker D12S7 (12q14-q24.1) showed no recombination with GalNAc-T4, but the marker is not on the Génétion map. The suggested order according to the Génétion map was as follows: D12S90-D12S80-D12S81-“GalNAc-T4,” D12S7-D12S101.

The identified polymorphism at bp position 1516 in GalNAc-T4 was common. Gene frequencies of the alleles were $p(G) = 0.663$ and $p(A) = 0.337$ (Table II). There was no significant deviation from the Hardy-Weinberg expectation ($X^2_{(1)} = 0.012; 0.05 < p < 0.90$).

Expression of GalNAc-T4506V—Expression of pAcGP67-GalNAc-T4506Vsol resulted in GalNAc-transferase activity in the culture medium of infected cells that was greater than background values obtained with uninfected controls or cells infected with the histo-blood group O2 gene (not shown). Activities measured with the mucin derived substrates, Muc7, EA2, and Muc2 (see Table III for structures), were only 2–7-fold over background values. Very low activities with a few other substrates were also observed, but for many of the “mucin-like” substrates a relative high background with SF9 cells infected with irrelevant constructs made assessment difficult. GalNAc-T4 was expressed in High Five™ cells and purified to near homogeneity, however, very low endogenous activity was detected with some substrates in the same fractions when medium from cells infected with irrelevant expression constructs were used in the same purification strategy. The activities presented in Table III were obtained with purified GalNAc-T4 and -T2, and are expressed as milliunits/mg (the concentration of enzyme protein estimated by SDS-PAGE using bovine serum albumin as standard). The highest activities with GalNAc-T4 were found (as expected) from assays of SF9 medium with the mucin peptides Muc7, EA2, and Muc2, however, the efficiencies with these substrates were considerably lower than GalNAc-T2.

**Table I**

| Marker | 0.01 | 0.05 | 0.10 | 0.20 | 0.30 | 0.40 |
|--------|------|------|------|------|------|------|
| D12S90 | -18.88 | -7.1 | -2.76 | 0.37 | 1.03 | 0.56 |
| D12S80 | -5.36 | 2.15 | 3.72 | 3.99 | 2.91 | 1.21 |
| D12S81 | 10.42 | 10.21 | 9.33 | 7.09 | 4.52 | 1.79 |
| D12S7  | 8.05 | 8.03 | 7.39 | 5.69 | 3.73 | 1.56 |
| D12S101| 1.46 | 4.28 | 4.9 | 4.44 | 3.09 | 1.27 |

**Table II**

| Genotype | Cases |
|----------|-------|
| G/G      | 91    |
| A/A      | 24    |
| A/G      | 91    |
| Total    | 206   |
T4, where the only efficient substrate identified was the EA2 sequence from rat submandibular glands (5). MUC1-derived substrates, Muc1a (19-mer) and Muc1b (15-mer) at 500 μM, yielded negligible activity with murine GalNAc-T4. The glycopeptide, GalNAc-T4, TAP24 (Table III), was tested because it has all the acceptor sites for GalNAc-T1, -T2, and -T3 occupied (8), and therefore is not a substrate for these enzymes. Surprisingly, a low but significant activity with this substrate was observed, indicating that although incorporation is observed with substrate even after prolonged incubation (24 h).

### Table III

| Peptide | Sequence | GalNAcT4 | GalNAcT2 |
|---------|----------|----------|----------|
| Muc1a'  | AHGVTSPADTPR | 6.7 ± 1  | NT       |
| Muc1b'  | RPAPGSTAPPA | 0.9 ± 0.8 | 0.7 ± 0.5 |
| TAP24   | TAPPAHGTSPAPDTPGASTAPP | 7.1 ± 0.8 | 4.4 ± 0.1 |
| GalNAcTAP24 | TAPPAHGTSPAPDTPGASTAPP | NT    | NT       |
| Muc2    | PPTTRIPPTITIIIMPTPTTPT | 29 ± 0.8 | 9 ± 0.5  |
| Muc7    | CFPPTPSATTPAPSSSPAPPTETTA | 53 ± 2  | 21 ± 1   |
| PSGL-1a | QATEYELDYDYLPEC | ND      | ND       |
| PSGL-1b | Ac-QATEYELDYLDFLPEETPEPM | 1.4 ± 0.1 | 1.2 ± 0.1 |
| HIV_int | Ac-CPTRGPRAPVTTGKIGNMR | ND     | ND       |
| VTHP | Ac-PFTVPFPGYD | ND      | ND       |
| hCG-β   | PRQFDRSSSAPPSPSRLP | ND      | ND       |
| EA2     | DSTTPAPPTK | 42 ± 4  | 15 ± 0.1 |

### Table IV

| Substrate | Sequence | GalNAcT1 | GalNAcT2 | GalNAcT3 | GalNAcT4 |
|-----------|----------|----------|----------|----------|----------|
| Acceptor substrate | | | | | |
| Muc7     | CPPTPSATTPAPPSSAPPETAA | 0.04 | 0.76 | 0.015 | 0.64 | 0.01 | 0.056 | 0.48 | 0.25 |
| EA2      | DSTTPAPPTK | 0.07 | 1.80 | 0.85 | 1.80 | 0.06 | 0.96 | 2.19 | 0.44 |
| GalNAcTAP24 | TAPPAHGTSPAPDTPGASTAPP | NA | NA | NA | NA | NA | NA | NA | 0.09 | 0.007 |
| PSGL-1b | Ac-QATEYELDYLDFLPEETPEPM | ND | ND | ND | ND | ND | ND | 0.02 | 0.003 |

| Donor substrate | UDP-GalNAc | 0.062 | 0.010 | 0.029 | 0.16 |

GalNAc-T4 O-Glycosylates Thr<sup>57</sup> in PSGL-1—Two peptide designs of the NH<sub>2</sub>-terminal sequence of mature PSGL-1 were tested (Table III). In a recent study, Thr<sup>57</sup> of PSGL-1 was identified as the carrier site of an O-glycan required for P-selectin binding (29). GalNAc-T1, -T2, and -T3, exhibited poor activities with both peptide designs, but HPLC analysis of prolong reaction indicated that both Thr<sup>44</sup> and Thr<sup>57</sup> served as substrates, although the reaction did not go to completion for either site (not shown, Tables III and IV). GalNAc-T4, in contrast, showed low activity with the PSGL-1b peptide containing both Thr<sup>44</sup> and Thr<sup>57</sup>, but not with PSGL-1a without the Thr<sup>57</sup> site. The reaction went to completion and the apparent <i>Kₘ</i> for this peptide was very low (Tables III and IV). Analysis of the exhaustively glycosylated PSGL-1b peptide by mass spectrometry confirmed that a single GalNAc was incorporated (not shown). A strong substrate inhibition with the negatively charged PSGL-1b peptide was observed at concentrations greater than 1 mM. The <i>Kₘ</i> for purified GalNAc-T4 for UDP-GalNAc was 160 μM using the Muc7 acceptor substrate. No incorporation with UDP-Gal or UDP-GlcNAc was found using the same peptide.

### Establishment of a Stably Transfected CHO Line Secreting High Levels of GalNAc-T4—GalNAc-T4 enzyme from insect cells was found to have poor kinetic properties compared with other GalNAc-transferases even with acceptor substrates that exhibited low <i>Kₘ</i> values. A stable CHO transfecant was produced to obtain a stable source of GalNAc-T4, and to determine if another cell type produced GalNAc-T4 with better kinetic properties.
expression of GalNAc-enzyme source is valuable for studies of substrates with low NAc-transferase activity in the medium of wild-type CHO, this H3–6. Since there is no detectable endogenously secreted Gal-T4/21A1 is comparable or better than that of CHO/GalNAc-T3/ with Muc7 (Table III), thus the secretion level of CHO/GalNAc-T4 appear to have a specific activity of 0.053 unit/mg NAc-T3 has a specific activity of 0.5 unit/mg with Muc1a, and an anti-GalNAc-T3 monoclonal antibody (not shown). Gal- established using the same immunoscreening procedure with CHO clones, CHO/GalNAc-T3/H3–6, secreting 5 milliunits/ml activity measured with Muc7 in the standard assay. A stable medium of confluent T-flasks contained up to 0.95 milliunits/ml other GalNAc-transferases (not shown). Further- not with cells infected with constructs for soluble or full-length GalNAc-T1, -T2, -T3, or irrelevant genes (not shown). Further- high-expressing variant, as found for the secreted construct (not shown), however, the assay is influenced by a very high endogenous GalNAc-transferase background. The expression level was evaluated by immunochemistry with UH6, and cells infected with the full coding construct stained stronger than those infected with the pAcGP67-GalNAc-T4 506Vsol construct. When a Triton X-100 homogenate of infected Sf9 cells was incubated with GalNAc,TAP24 for 24 h with one addition of UDP-GalNAc and extra enzyme, no incorporation was detected by MALDI-TOF of HPLC purified product.

Northern Blot Analysis of Human Organs—Northern blots with mRNA from 24 human adult and 5 human fetal organs were probed with GalNAc-T4 using similar blots previously used for probing GalNAc-T1, -T2, and -T3 (4). GalNAc-T4 hy-bridized to a single mRNA of approximately 6 kilobases with varying intensity in all organs (Fig. 4), suggesting a ubiquitous expression pattern similar to GalNAc-T1 and -T2. This pattern is different than the expression pattern found in Northern analysis of murine organs with the putative murine GalNAc-T4 orthologue, where a more restricted expression pattern was observed (5).

Immunolocalization of GalNAc-T4 with UH6—Hagen et al. (5) showed by Northern analysis that the putative murine GalNAc-T4 orthologue was strongly expressed in sublingual and not submandibular salivary glands. Human mRNA from these organs were not available for this study. Immunohistol- ogy on frozen sections of human submandibular glands show that GalNAc-T4 is strongly expressed in mucous cells, whereas staining of serous cells was very weak (Fig. 5). Thus, the failure of Northern analysis to detect expression in submandibular glands is likely to be result of the low number of mucous cells in this gland combined with low expression in serous cells (5).

GalNAc-T4 contains a hydrophobic signal sequence which is predicted to provide retention in Golgi (30). The observed subcellular staining pattern with supranuclear reactivity sugges- tive of Golgi localization was identical to the staining pattern found with antibodies to GalNAc-T1, -T2, and -T3 (9). Im-
DISCUSSION

In the present study, a fourth human GalNAc-transferase gene, designated GalNAc-T4, was cloned and expressed. GalNAc-T4 was shown to have polypeptide GalNAc-transferase activity with a unique acceptor substrate specificity distinct from other human GalNAc-transferases. We have previously identified unique substrates for the first three human GalNAc-transferases (4, 8). GalNAc-T4 was exceptional in that it had no or only poor activities with acceptor substrates derived from several mucin tandem repeats, which represent some of the best substrates for other GalNAc-transferases. In contrast, GalNAc-T4 showed unique activities for new acceptor sites in the MUC1 tandem repeat and in the P-selectin ligand, PSGL-1. Thus, in vitro studies of the activities of recombinant GalNAc-transferases provide increasing evidence that each GalNAc-transferase serve different roles in O-glycosylation (8, 33). The finding that each enzyme has unique acceptor sites may suggest that evolution of this large family of GalNAc-transferases was driven by the need to glycosylate a diverse set of protein sequences.

The molecular processes governing the specificity and kinetics of O-glycosylation remain obscure. It is clear that a single consensus peptide sequence does not exist for O-glycosylation (34–36). The extent to which the primary amino acid sequence determines O-glycosylation has long been debated, and the failure to identify consensus motifs has indicated that secondary structure and surface accessibility plays a larger role than primary sequence context in specifying acceptor sites for O-glycosylation. This hypothesis was initially supported by an in vivo study of O-glycosylation of a single acceptor site, where extensive amino acid substitutions in flanking sequences produced no or only little influence of the observed O-glycosylation (37). The in vivo result was in striking contrast to the effects of similar substitutions in an in vitro study with the same acceptor site using purified bovine GalNAc-T1, in which GalNAc-T1 was particularly inhibited by flanking residues with negative charge (38). It is important to note, however, that while in vitro studies can assess the specificity of a single GalNAc-transferase, in vivo studies mostly evaluate the accumulated activities of the entire repertoire of GalNAc-transferases expressed in the cell type chosen for study. An exception to this is the strategy developed by Rottger et al. (32), where a single GalNAc-transferase is relocated to endoplasmic reticulum. Nevertheless, the data indicate that in vitro assessed specificity of GalNAc-transferases failed to reflect their in vivo activity. In agreement with this, Muller et al. (17) recently showed that five potential sites in the MUC1 tandem repeat peptide sequence can be glycosylated in vivo, while previous in vitro studies indicated that only a total of three sites in the repeat were utilized by various GalNAc-transferase preparations (39–41) as well as by recombinant GalNAc-T1, -T2, and -T3 (8).

An important point is that most observed discrepancies between in vitro and in vivo O-glycosylation patterns have been examples where in vitro glycosylation is less effective than in vivo glycosylation. Since the complete repertoire of GalNAc-transferases and their accumulated substrate specificities is still unknown, it is possible that all in vivo identified sites can be accommodated by these enzymes. Furthermore, it may not be possible to evaluate the O-glycosylation capacity of a given cell or organ by merely analyzing the total activities in extracts. Studies addressing these issues in detail indicate that the primary amino acid sequence of acceptor sites may be the major determining factor of O-glycosylation. Previously, a unique acceptor site for human GalNAc-T3 was identified in the HIVsubV3-loop sequence (4). In a recent study, Nehrke et al. (33) used an in vivo model to show that in vivo glycosylation of the HIV peptide sequence was dependent on the co-expression of GalNAc-T3 in the host cell. Furthermore, it was demonstrated that the unique specificity of GalNAc-T3 for this substrate was directed by a single charged residue at position +3 of the acceptor site. As shown here, the previous failure of complete O-glycosylation of the MUC1 tandem repeat peptides in in vitro studies, may be explained by the unique acceptor specificity of GalNAc-T4 and its kinetic properties. Both human and murine GalNAc-T4 showed poor activity with various peptide designs from the MUC1 tandem repeat (Table III) (5), whereas the partially glycosylated repeat peptide, GalNAcTAP24, was one of the best substrates identified for the human homologue. The observed poor catalytic efficiency of GalNAc-T4 may explain why this enzyme activity cannot be measured in total extracts when the starting substrate is the naked peptide. This is further supported by the finding that the

![Figure 4](https://example.com/figure4.png)

FIG. 4. Northern blot analysis of human tissues. Multiple tissue Northern blots, MTNI, MTNII, immune MTN, endocrine MTN, and fetal MTN, from CLONTECH was hybridized with 32P-labeled GalNAc-T4 probe (TEB4) (Panel A). Panel B blots were hybridized with β-actin probe.
full coding construct of GalNAc-T4 expressed in Sf9 cells as monitored by immunocytology failed to yield significant activities above background values in extracts. Elucidating the mechanism for this require further studies. It is likely that the failure to identify a single consensus acceptor sequence for O-glycosylation may be explained by the existence of a large number of GalNAc-transferases that each have different acceptor specificities.

GalNAc-T4 exhibited unique acceptor substrate specificities different from other studied GalNAc-transferases. Thus, the acceptor site identified in PSGL-1 that exhibited the lowest $K_m$ is flanked by negatively charged residues (-ETE-), which is a sequence context rarely found in glycoproteins (36). Since P-selectin binding to PSGL-1 is dependent on the presence of an O-glycan at this site (29), identification of the GalNAc-transferase responsible for O-glycosylation is important. PSGL-1 is O-glycosylated at Thr$^{57}$ in HL-60 cells and in CHO (42), which may suggest that GalNAc-T4 will be expressed in these cell lines. The repertoire of GalNAc-transferases in HL-60 cells include GalNAc-T1, -T2, and -T3 (9), and immunocytology with UH6 also demonstrated weak expression of GalNAc-T4 (not shown). One of the unique acceptor sites identified in MUC1 (-PDTR-) is also flanked by charged residues. The finding that GalNAc-T4 showed strong preference for the partially glycosylated MUC1 repeat demonstrates that prior O-glycosylation may induce new acceptor sites or at least enhance the activity with these sites. Previously, a negative influence of prior glycosylation was shown with MUC2-derived glycopeptides using GalNAc-T1 (43). Presently, it is not clear which of the three sites in the MUC1 repeat glycosylated by GalNAc-T1, -T2, and -T3 are required to induce the GalNAc-T4 activity. The order in which GalNAc-T4 processes the two last sites is also not clear. Nevertheless, it is evident that complete O-glycosylation of the MUC1 tandem repeat requires several GalNAc-transferases perhaps acting in a specific order. The reduced number of O-glycans of MUC1 in cancer (14), may be explained by specific changes in GalNAc-transferase repertoire and/or topology in cancer cells.

A putative mouse orthologue of GalNAc-T4, with 91% amino acid sequence similarity to human GalNAc-T4, was previously identified (5). The murine GalNAc-T4 was found to exhibit unique specificity for the EA2 peptide, whereas no glycosylation was observed with several other mucin-related acceptor sequences including MUC1-derived peptides. Human GalNAc-T4 also utilized the EA2 peptide, but the tandem repeat sequence of human MUC7 was a better substrate. The expression pattern of murine GalNAc-T4 was restricted and high expression only found in salivary glands, colon, and gastric mucosa (5). The expression pattern of human GalNAc-T4 in the same organs as tested for the murine enzyme were correlated, although the same high level of expression in colon and stomach was not found (Fig. 4). A wider expression pattern was found in human since more organs were tested. The finding that organs of the immune and endocrine systems expressed
GalNAc-T4 is potentially highly significant. Since the human and mouse GalNAc-T4 are similar in sequence and their substrate specificities expression patterns appear to be similar, it is likely that they represent orthologous genes. However, it is important to note that several of the GalNAc-transferases exist in paired copies with high sequence similarity, and with identical genomic organization. A close homologue of GalNAc-T3, designated GalNAc-T5,4 has an identical genomic structure with 10 coding exons (11), and its kinetic properties are identical to GalNAc-T3. It may be difficult to identify orthologous GalNAc-transferase genes, and an example of this was the recent attempt to generate a knock-out of murine GalNAc-T1 based on the bovine sequence (44). Apparently, a close homologous gene, which is yet not fully cloned and characterized, was targeted and the resultant lack of phenotype is therefore still unexplained (45).

Studies of the genomic organization of the GalNAc-transferase family have revealed features that suggest a close evolutionary link between GalNAc-T1, -T2, and -T3, such as conserved intron/exon boundaries (11). Characterization of the GalNAc-T4 gene showed an organization entirely different from other GalNAc-transferase genes. This may shed light on the origin and function of GalNAc-T4. The organization of the GalNAc-T4 gene leads us to propose two possibilities regarding its origin. The first is that GalNAc-T4 arose earlier in evolution than the human GalNAc-T1, -T2, and -T3 genes, if introns are considered late introductions into these genes (46). We recently characterized the genomic organization of human GalNAc-T1, -T2, and -T3 (11). These three genes are encoded in multiple exons (10–16). Several intron/exon boundaries within the coding regions of these genes are conserved, suggesting that the genes arose through gene duplication from a common ancestral gene containing these introns. The genomic organization of several C. elegans homologues have been characterized and they all contain multiple coding exons (6), and at least one of these genes contain an intron/exon boundary conserved with human GalNAc-T1, -T2, and -T3 (11). This strongly suggest that the GalNAc-transferase family arose early in evolution. A second possibility regarding the origin of GalNAc-T4 is that it arose by retrotransposition from one of the GalNAc-transferases. A retrotransposed pseudogene with high similarity to GalNAc-T1 has been identified, however, it is not expressed and the gene contains inserted sequences, multiple frameshift mutations, and repeat sequences characteristic of transposons (10). Several findings are contrary to this hypothesis: it is unusual to find expressed and functional pseudogenes; common features of primate retrocopies and pseudogenes such as direct repeats and A rich boxes, were not found in the 5' and 3' sequences flanking the coding region (47); and GalNAc-T4 does not exhibit substantial identity to any of the identified and characterized GalNAc-transferases.

The GalNAc-transferase genes are not clustered at one locus. All genes reported thus far are located on different chromosomes (1, 11), as are two additional putative GalNAc-transferase genes that we are currently investigating, including a gene that shows a high degree of similarity to GalNAc-T3.4 The significance of this is presently unknown, but assignment of chromosomal localization and identification of marker microsatellites will allow correlation with diseases linked to particular chromosomal sites. Thus, GalNAc-T3 which is selectively strongly expressed in pancreas co-localizes to a recently identified diabetes susceptibility gene on 2q24–31 (11).

4 E. P. Bennett, H. Hassan, T. Hollingsworth, and H. Clausen, manuscript in preparation.
33. Nehrke, K., Hagen, F. K., and Tabak, L. A. (1998) *Glycobiology* 8, 367–371
34. Elhammer, A. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G., and Kezdy, F. J. (1993) *J. Biol. Chem.* 268, 10029–10038
35. Gooley, A. A., and Williams, K. L. (1993) *J. Biol. Chem.* 268, 10029–10038
36. Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L., and Brunak, S. (1996) *Glycoconj. J.* 13, 115–130
37. Nehrke, K., Hagen, F. K., and Tabak, L. A. (1996) *J. Biol. Chem.* 271, 7061–7065
38. O'Connell, B. C., Hagen, F. K., and Tabak, L. A. (1992) *J. Biol. Chem.* 267, 25010–25018
39. Nishimori, I., Johnson, N. R., Sanderson, S. D., Perini, F., Mountjoy, K., Cerny, R. L., Gross, M. L., and Hollingsworth, M. A. (1994) *J. Biol. Chem.* 269, 16123–16130
40. Nishimori, I., Perini, F., Mountjoy, K. P., Sanderson, S. D., Johnson, N., Cerny, R. L., Gross, M. L., Fontenot, J. D., and Hollingsworth, M. A. (1994) *Cancer Res.* 54, 3738–3744
41. Stadie, T. R., Chai, W., Lawson, A. M., Byfield, P. G., and Hanisch, F. G. (1995) *Eur. J. Biochem.* 229, 140–147
42. Li, P., Wilkens, P. P., Crawley, S., Weinstein, J., Cummings, R. D., and McEver, R. P. (1996) *J. Biol. Chem.* 271, 3255–3264
43. Broekhuisen, I., Toki, D., Broekhuisen, J., Peters, S., Bielfeldt, T., Kleen, A., Paulsen, H., Meldal, M., Hagen, F., and Tabak, L. A. (1996) *Glycoconj. J.* 13, 849–856
44. Hennet, T., Hagen, F. K., Tabak, L. A., and Marth, J. D. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 12070–12074
45. Marth, J. D. (1996) *Glycobiology* 6, 701–705
46. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W., and Gilbert, W. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 14632–14636
47. Nadir, E., Margalit, H., Gallily, T., and Ben-Sasson, S. A. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 6470–6475
48. Lowe, J. B. (1990) *Bailliere’s Clin. Hematol.* 4, 465–492
49. Clausen, H., Bennett, E. P., and Grunnet, N. (1994) *Transf. Clin. Biol.* 1, 79–89
50. Ott, J., Tan, J., Dunn, J., Jaeken, J., and Schachter, H. (1996) *Am. J. Hum. Gen.* 59, 610–617
51. Takeda, J., Miyata, T., Kawagoe, K., Iida, Y., Endo, Y., Fujita, T., Takahashi, M., Kitani, T., and Kinoshita, T. (1993) *Cell* 73, 703–711