cDNA Cloning and Expression of a Family of UDP-N-acetyl-d-galactosamine:Polypeptide N-Acetylgalactosaminyltransferase Sequence Homologs from Caenorhabditis elegans*

(Received for publication, November 12, 1997, and in revised form, January 21, 1997)

Fred K. Hagen† and Keith Nehrke

From the Center for Oral Biology, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642

The initiation of mucin-type O-glycosylation is catalyzed by a family of UDP-GalNAC:polypeptide N-acetylgalactosaminyltransferases (ppGaNTase) (EC 2.4.1.41). By screening two mixed-stage Caenorhabditis elegans cDNA libraries, a total of 11 distinct sequence homologs of the ppGaNTase gene family were cloned, sequenced, and expressed as truncated recombinant proteins (gly-3, gly-4, gly-5a, gly-5b, gly-5c, gly-6a, gly-6b, gly-6c, gly-7, gly-8, and gly-9). All cloned encode type II membrane proteins that shared 60–80% amino acid sequence similarity with the catalytic domain of mammalian ppGaNTase enzymes. Two sets of cDNA clones (gly-5 and gly-6) contained variants that appeared to be produced by alternative message processing, gly-6c contained a reading frameshift and premature termination codon in the C-terminal lectin-like domain found in most other ppGaNTase proteins, and a second clone (gly-8) lacked the typical C-terminal region completely. Homogenates of nematodes and immunopurified preparations of the recombinant GLY proteins demonstrated that worms express functional ppGaNTase enzymes (GLY-3, GLY-4, GLY-5A, GLY-5B, and GLY-5C), which can O-glycosylate mammalian apomucin peptide sequences in vitro. In addition to demonstrating the existence of ppGaNTase enzymes in a nematode organism, the substantial diversity of these isoforms in C. elegans suggests that mucin O-glycosylation is catalyzed by a complex gene family, which is conserved among evolutionary-distinct organisms.

The diversity of O-linked oligosaccharides displayed on secreted and cell surface glycoproteins is determined by the repertoire of glycosyltransferases present in the Golgi apparatus. The biosynthesis of mucin-type oligosaccharides at specific O-glycosylation sites begins with the transfer of the monosaccharide N-acetylgalactosamine (GalNAc) to specific threonines and serines of an apo-protein. This initiation event is regulated in mammals by a family of at least seven enzymes, known as UDP-GalNAC:polypeptide N-acetylgalactosaminyltransferases (ppGaNTases) (EC 2.4.1.41). Four ppGaNTase isoforms have been cloned and functionally expressed from rodent and human cDNAs: ppGaNTase-T1 (1, 2), ppGaNTase-T2 (3), ppGaNTase-T3 (4, 5), and ppGaNTase-T4 (6). Evidence for three additional members of this gene family has been obtained, using a polymerase chain reaction approach. An eighth murine isoform having high amino acid sequence homology to ppGaNTase-T1 has also been suggested by nucleic acid cross-hybridization to a genomic library (7). Transcripts encoding specific ppGaNTase enzymes have distinct tissue patterns of expression, indicating that the acquisition of O-glycans can be regulated by differential expression of the ppGaNTase gene family. The substrate reactivity of the mammalian isoforms varies from those that are broad (ppGaNTase-T1) to those that recognize a narrow range of specific peptide sequences (ppGaNTases-T4); therefore, O-glycosylation of specific proteins in vivo requires the coordinate expression of polypeptide substrates and their cognate ppGaNTase enzymes. An analysis of the human and murine expressed sequence tag (EST) data base revealed numerous new sequence homologs, suggesting that an even larger and more complex family may exist in mammals (8).

The complexity and potential redundancy of the ppGaNTase gene family in mammalian systems is underscored by a ppGaNTase gene ablation study in mice, in which the deletion of an exon from a putative ppGaNTase gene produced mice that appeared normal and unaffected in their ability to O-glycosylate proteins (7, 9). The growing number of ppGaNTase isoforms isolated in mammals indicates that a genetic approach to ablating ppGaNTase activity in a murine model will be a lengthy undertaking.

In this present study, we have searched for a simple model organism that is suitable for using a genetic approach to study the roles of mucin-type O-glycosylation during development and differentiation. We selected Caenorhabditis elegans because nematodes express hyperabundant mucin-like glycoproteins and because C. elegans is amenable to classical and reverse genetic studies. In addition, the EST data base revealed numerous C. elegans clones, which encoded putative sequence homologs of the mammalian ppGaNTases, providing us with a system for identifying the size and properties of the complete ppGaNTase family in a whole organism. Biochemical studies performed here revealed that worms express ppGaNTase enzyme activity and that this activity is encoded by a family of enzymes. A total of 11 distinct C. elegans cDNAs (encoded by polycrylamide gel electrophoresis; Tricine, N-tris(hydroxymethyl)glycine; MES, 2-(N-morpholino)ethanesulfonic acid.

† To whom correspondence should be addressed: Center for Oral Biology, School of Medicine and Dentistry, University of Rochester, Rochester, New York 14642. Tel.: 716-275-0380; Fax: 716-473-2679; E-mail: fred_hagen@urmc.rochester.edu.


* This work was supported in part by National Institutes of Health Grant DE-08108. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF031833 (gly-3), AF031834 (gly-4), AF031835 (gly-5a), AF031836 (gly-5b), AF031837 (gly-5c), AF031838 (gly-6a), AF031839 (gly-6b), AF031840 (gly-6c), AF031841 (gly-7), AF031842 (gly-8), and AF031843 (gly-9).

The abbreviations used are: GalNAc, N-acetylgalactosamine; ppGaNTase, UDP-GalNac:polypeptide N-acetylgalactosaminyltransferase; EST, expressed sequence tag; PCR, polymerase chain reaction; UTR, untranslated region; nt, nucleotide(s); aa, amino acid(s); PAGE, polyacrylamide gel electrophoresis; Tricine, N-tris(hydroxymethyl)methylglycine; MES, 2-(N-morpholino)ethanesulfonic acid.
gly genes, containing complete open reading frames with sequence homology to mammalian ppGaNTase, were isolated, sequenced, and expressed. Two sets of these cDNAs appear to be splice variants. Functional analysis of recombinant worm enzymes demonstrated that five of the members of this family catalyzed the ppGaNTase reaction in vitro using mammalian peptides as acceptor substrates.

**EXPERIMENTAL PROCEDURES**

*C. elegans* Homogenate and Extract—*C. elegans* N2 worms were grown on 15-cm egg plates for 2–3 weeks, washed with M9 medium (22 mm KH2PO4, 22 mm Na2HPO4, 86 mm NaCl, 1 mm MgSO4), and harvested and resuspended from debris by sedimentation at 18,000 × g for 20 min. Floating worms were removed and resuspended in a large volume of 50 mM NaCl and centrifuged as above but only for 2 min. This final worm pellet was resuspended in 100 mM NaCl, incubated at 20 °C for 30 min, and then frozen at −70 °C.

To prepare a homogenate of mixed-stage nematodes, aliquots of frozen worms were thawed, and all subsequent steps were performed at 4 °C. Nematodes (0.5 grams) were placed in a 2-ml tube containing 1 ml of 1-mm glass beads and 1 ml of homogenization medium (500 mM sucrose, 25 mm Tris maleate, pH 6.4, 1 mm MgCl2, and 0.1% detergent (average molecular weight = 250,000), and 5 mM β-mercaptoethanol). Worms were disrupted three times with 1-min pulses, using a Mini-Beadbeater-8 cell disrupter (Biopsec Products). One volume of this homogenate was diluted with 10 volumes of H2O and centrifuged at 250,000 × g for 60 min. The pellet was resuspended in 230 μl of extraction solution (50 mM sodium cacodylate, 50 mM β-mercaptoethanol, and 2.5% Triton X-100) and homogenized with a 1.5-ml tube plunger. After 30 min of rocking, the tube was centrifuged for 10 min and the supernatant was removed and saved. The pellet fraction was re-extracted as above, and the supernatants were combined, mixed, and stored at −70 °C.

**ppGaNTase Enzyme Assays—**Enzyme activity was measured in vitro using the following assay conditions: a final volume of 25 μl containing a final concentration of 500 μM EA2 peptide, 50 μM UDP-[14C]GalaNaC (25,000 cpm), 10 mM MnCl2, 40 mM cacodylate pH 6.5, 40 mM β-mercaptoethanol, and 0.1% Triton X-100. Peptide substrates (amino acid sequences in parentheses) used in these assays are: EPO-T (PPDDAATA-APLR), EPO-S (PPDDASAAPLRR), and EA2 (PTTDSTTPAPPTTK) at a concentration of 500 μM and TPPP at 1 mM. Three μl of worm extract, purified bovine colostrum ppGaNTase (2), or immunopurified recombinant enzymes were used in this standard enzyme assay (Fig. 1). All enzyme assay points were performed in duplicate, and these were repeated with duplicate enzyme preparations from worm extracts or COS7 cell supernatants. Glycosylated [14C]-labeled peptides were separated from unincorporated UDP-[14C]GalaNaC by anion exchange chromatography on formate form AG 1x8 resin spin columns (Bio-Rad).

**Data Base Analysis—**The amino acid sequence of the mouse ppGaN-Tase-T1 was used as a query to perform a TBLASTN analysis of the data base of expressed sequence tags (dBEST) (10). All EST sequences derived from the *C. elegans* data base were conceptually translated and aligned to the four known mammalian ppGaNTase isoforms, ppGaNTase-T1, -T2, -T3, and -T4. Any clone that contained at least three homologous (eight amino acids in length) with a defined spacing between each segment was treated as a putative homolog of the ppGaNTase family and used to generate a cDNA hybridization probe.

Seven different expressed sequence tag (EST) clones (yk2211, yk3 g10, yk15e11, cm13e2, yk15i1a8, yk72f6, and cm16e9) were selected for probe design (probes B for gly-4, C for gly-7, D for gly-6, E for gly-5β and -5c, F for gly-6, and H for gly-9). EST clones yk2211, yk3 g10, and yk15e11 were obtained from Dr. Y. Kohara. EST clones cm13e2 and cm16e9 were obtained from Dr. L. Fulton and Dr. Robert Waterston.

**PCR Amplification of Sequence Tags and Preparation of cDNA Hybridization Probes—**Nucleic acid hybridization probes, specific for each EST, were prepared by isolating polymerase chain reaction (PCR) products of a 190–330-nucleotide region in each cDNA from the data base. EST clones described above were used as templates for amplifying probes for gly-4, gly-6, and gly-7, while first strand cDNA from mixed stage nematode total RNA was used as a template for amplifying fragments of the remaining EST clones. First Strand cDNA synthesis kit (CLONTECH), oligo(dT)16, and total RNA from frozen stage *C. elegans* N2 nematodes. The oligonucleotides used for the amplification of each isoform is described in Table 1. [32P]-Labeled probes were generated using purified PCR products, antisense oligonucleotides for cDNAs gly-4 through gly-9, and a PCR labeling protocol. Blockage of PCR product was added to a 12.5 reaction mixture containing 1 μM primer, 1× PCR buffer, 1.5 mM MgCl2, 50 μM dATP, dGTP, dTTP (each), 0.75 units of Taq DNA polymerase (Perkin-Elmer), and 5 μl of [α-32P]dCTP (3000 Ci/mmol, NEN Life Science Products). Reactions were amplified for 30 cycles at 94 °C for 40 s, 49 °C for 40 s, and 72 °C for 40 s.

**cDNA Library Screening and Clone Characterization—**Full-length cDNA sequences for 10 of the sequence homologs were obtained by screening two *C. elegans* cDNA libraries: an oligo(dT)-primed cDNA library and a random-primed cDNA library, λ-ACT-RB1 and λ-ACT-RB2, respectively (kindly provided by Dr. R. Barstead; Ref. 11). Seven hundred thousand phage of each library RB1 and RB2 were plated onto 24 × 24-cm NuPlate and a lawn of LE392 Escherichia coli cells. The phage plaques plated using Hybrid-N membranes (Millipore Pharmacia Biotech), and the membranes were hybridized overnight at 68 °C in 5× SSPE, 50% formamide, 5× Denhardt’s, 0.1% SDS, and 100 μg/ml salmon sperm DNA, containing 3 × 105 cpm/ml of each [32P]-labeled denatured probe. Filters were washed three times for 20 min each in 2× SSC and 0.1% SDS at the following three temperatures: 42, 64, and 42 °C. Initial screening was performed with a mixture of seven probes for isoforms gly-4, gly-5α, gly-5β, gly-6, gly-7, gly-8, and gly-9. Ninety-six positive plaques were cored, dot-blotted on multiple Hybrid-N membranes, and probed with individual isoform-specific probes, using the conditions above. Twenty-one clones, corresponding to 10 different cDNAs, were isolated to homogeneity. Cre-lox excision of the pACT plasmid from each cloake was accomplished by transduction into the *E. coli* strain RB4, which expresses the Cre recombinase. COS7 quality plasmid DNA was prepared in the RB4 host and used directly for infrared fluorescence DNA sequencing, using Bca DNA polymerase in a Laddermann Core sequencing protocol (PanVera) and a LICOR model 4000L DNA sequencer. IRD41 dye-labeled primers were designed for the pACT plasmid, using the following sequences: PACT-F primer d(TATCTAATTGATGATGAAAG) and PACT-R primer d(AACATGGTTGACATCGGTGTTG). Ninety-six positive plaques were cored, dot-blotted on multiple Hybrid-N membranes, and probed with individual isoform-specific probes, using the conditions above. Twenty-one clones, corresponding to 10 different cDNAs, were isolated to homogeneity. Cre-lox excision of the pACT plasmid from each clone was accomplished by transduction into the *E. coli* strain RB4, which expresses the Cre recombinase.

**Design of Expression Constructs—**Expression constructs were designed such that the cDNAs were expressed as secreted recombinant proteins, lacking their natural N-terminal membrane anchors (for amino acid sequence, see Fig. 2). To clone all of the isoforms, a new expression vector construct, pIMKFS3, was created. The SV40 promoter-driven expression plasmid pIMKFS is virtually identical to pIMKFS1 (6), except that the multiple cloning site was expanded to include four new unique restriction sites: ApaI, BglII, NotI, and SacII (Fig. 5). Seven cDNA clones (gly-3, gly-4, gly-5α, gly-5β, gly-6, gly-7, gly-8, and gly-9) were introduced into the 5′ MluI and 3′ multiple cloning site of pIMKFS3, creating the constructs pFS3-GLY3, pFS3-GLY4, pFS3-GLY5a, pFS3-GLY6a, pFS3-GLY7, pFS3-GLY8, and pFS3-GLY9, respectively. The MluI restriction site was introduced into all constructs by PCR amplification of the stem region with a PCR primer (Table I, column two) and a downstream antisense strand primer in the 3′ UTR or vector sequences (Table I, column four). Only the cDNA sequences encoding the full-length, catalytic region, C-terminal coding region, and part of the 3′-UTR was present in each expression construct (the first amino acids used in the secretion construct are indicated in Fig. 2). With the exception of pFS3-GLY3, the majority of the coding region was replaced by a restriction fragment of the cDNA clone ob-
tained from the original library screen, such that the PCR-derived fragment was minimal in size. All PCR-derived sequences and cloning sites were completely re-sequenced to verify that no random PCR-induced mutations or frameshift artifacts existed in the expression constructs. Expression constructs of the "b" and "c" splice variants of gly-5 and gly-6 were constructed by replacing the 3' end of the cDNA in the pF3-GLY5a and pF3-GLY6a clones with the original cDNAs that contain the variant regions. These splice variant constructs were sequenced to verify that the sequence and reading frames matched the original cDNAs. The pF3-GLY6b construct was unstable in E. coli.

**Transient Expression of Recombinant Proteins—** Recombinant enzymes were expressed by transient transfection of COS7 cells, using these pF3-GLY3 through pF3-GLY9 constructs and LipofectAMINE reagents. The recombinant proteins were partially purified by incubating the culture medium (1.5 ml) or the cell extract (375 µl) with 150 µl of Storage Buffer (20 mM MES, pH 6.5, 50 mM NaCl, 1% Triton X-100, 5% glycerol). The extract was clarified by centrifugation at 10,000 × g for 5 min.

**Immunopurification of Recombinant Proteins and Analysis of Expression Levels—** The recombinant proteins were partially purified by incubating the culture medium (1.5 ml) or the cell extract (375 µl) with 150 µl of anti-FLAG M2 antibody-agarose (Eastman Kodak Co.) for 3 h at 4 °C.

**Temperature Dependence of GLY-3 Protein—** Enzyme activities were undetectable. The pMKF3 expression construct containing this truncated GLY-3/ZK688.8 protein is labeled pF3-GLY3. Protein expression was obtained by transient transfection of COS7 cells, which secreted the recombinant GLY-3 protein into the culture medium. To assay the functional activity of GLY-3, the recombinant GLY-3 protein was immunopurified using an anti-FLAG M2 antibody to remove any potential COST cell endogenous enzyme contamination. In parallel, pF3-NT1, encoding the secreted form of the mouse ppGaNTase-T1 isozyme, was transfected into COS7 cells. The pF3-GLY3 clone expressed sufficient ZK688.8 protein to demonstrate ppGaNTase activity (Fig. 1B). Enzyme assays performed at a range of incubation temperatures revealed that the temperature optimum for the recombinant worm GLY-3 protein was approximately 23 °C, while the recombinant mouse ppGaNTase-T1 had a temperature optimum of 45 °C.

**cDNA Cloning of 10 Additional ppGaNTase-like Clones—** Putative sequence homologs of the ppGaNTase family were detected in a C. elegans expressed sequence tag (EST) database using the mouse ppGaNTase-T1 amino acid sequence (6).
as a query and the program tblastn (10). A total of 20 EST cDNA clones contained at least three blocks of sequences conserved with mouse ppGaNtase-T1; however, many of these were overlapping clones. Nucleic acid hybridization probes were designed from seven non-overlapping C. elegans EST clones: probes B, C, D, E, F, G, and H in Table II. Two C. elegans \( \lambda \) ACT cDNA libraries, RB1 and RB2, containing oligo(dT) and random-primed cDNA clones, respectively, were hybridized with a mixture of the seven \( ^{32} \)P-labeled EST probes. A total of 1.4 million phage were screened, resulting in 584 clones with a moderate to strong hybridization signal. Therefore, on average, one positive clone was detected for every 2400 clones plated. Dot-blot analysis of 96 clones (48 from each library) revealed the frequency of each EST in the cDNA library (Table II). Hybridization with individual probes revealed that some cDNA clones hybridized to two probes, E and G. The abundance of these RNA messages in the cDNA library suggested the following relative frequency of each clone: most abundant = gly-7 > gly-5 > gly-8 > gly-6 > gly-4 > gly-9 > least abundant. The frequency of these cDNAs in the RB1 cDNA library was not an accurate representation of the frequency of the message in the RNA population because the 3'-UTR of some of the cDNAs (gly-6 and gly-9) were unstable in high copy DNA vehicles. Surprisingly, gly-8 was not detected in the random-primed RB1 cDNA library, although 19 clones out of 48 encoded gly-8 in the RB2 library. DNA sequence analysis of the available 5'-UTR did not reveal any evidence of SL1 or SL2 splice leaders; however, not all clones recovered significant lengths of 5'-UTR sequences.

DNA sequence and restriction analysis of 26 cDNA clones revealed a total of 10 novel cDNAs (designated gly-4, gly-5a, gly-5b, gly-5c, gly-6a, gly-6b, gly-6c, gly-7, gly-8, and gly-9), which were distinct from gly-3, the ZK6888.8 C. elegans ppGaNtase gene. (The nucleotide sequence is available in the GenBank™/NCBI Data Bank, using the accession numbers listed in the title page footnote.) Comparisons of gly-5a, gly-5b, and gly-5c cDNA sequences revealed that these three clones were 100% identical, except for an internal segment (about 100 nt) within the 3' end of the coding region. Similarly, gly-6a, gly-6b, and gly-6 cDNAs were identical for the complete sequence, except for an approximately 100–150-nt span in the 3’end of the coding region. No sequence variants were found for gly-3, gly-4, gly-7, gly-8, and gly-9 cDNA.

Conceptual translation of all 11 worm cDNAs revealed a single large open reading frame in each transcription unit. Type II membrane proteins were encoded by each cDNA (amino acid sequence, shown in Fig. 2). The generalized protein sequence features of these putative ppGaNtase homologs are summarized in Fig. 3. The overall size of these proteins typically ranges from 578 to 634 amino acids in length, except for GLY-8, which is 421 amino acids. GLY-8 appears to lack about 100–150-nt span in the 3' end of the coding region. Similarly, gly-6a, gly-6b, and gly-6 cDNAs were identical for the complete sequence, except for an approximately 100–150-nt span in the 3’end of the coding region. No sequence variants were found for gly-3, gly-4, gly-7, gly-8, and gly-9 cDNA.

![Fig. 1. Temperature-dependent ppGaNtase activity of worm and mammalian enzymes.](http://www.jbc.org/)

**TABLE II**

<table>
<thead>
<tr>
<th>Probes</th>
<th>Genes</th>
<th>RB1(^\text{a}) clones</th>
<th>RB2(^\text{b}) clones</th>
<th>Total RB clones</th>
<th>EST clone names(^c)</th>
<th>Frequency of EST clones in dBEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>gly-3</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>cm17a3, yk139e11</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>gly-4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>yk2f11, yk50c10, yk119a10</td>
<td>16</td>
</tr>
<tr>
<td>C</td>
<td>gly-7</td>
<td>13</td>
<td>17</td>
<td>30</td>
<td>yk3g10, yk98e8, cm13h5</td>
<td>16</td>
</tr>
<tr>
<td>D(^d)</td>
<td>gly-6</td>
<td>1</td>
<td>8</td>
<td>9</td>
<td>yk15e11</td>
<td>1</td>
</tr>
<tr>
<td>E(^e)</td>
<td>gly-5b</td>
<td>7</td>
<td>11</td>
<td>18</td>
<td>cm13e2, yk31d3, yk726b, yk88c3, yk153g7</td>
<td>26</td>
</tr>
<tr>
<td>F(^f)</td>
<td>gly-8</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>yk138g7, yk151a8, yk167e8, cm15f2</td>
<td>21</td>
</tr>
<tr>
<td>G(^f)</td>
<td>gly-9</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>cm16e9</td>
<td>5</td>
</tr>
</tbody>
</table>

\(^a\) The RB1 clones represent the clones from the oligo(dT)\(_\text{\large A}\)-primed cDNA library.

\(^b\) The RB2 clones represent the clones from the random-primed library.

\(^c\) The EST clones listed are those from dBEST that have identical sequence to probes A–H.

\(^d\) The D probe detected two types of cDNAs, gly-6a and gly-6b.

\(^e\) Some cDNA clones cross-reacted with the E and G probes.

---

**Fig. 1. Temperature-dependent ppGaNtase activity of worm and mammalian enzymes.** Activity of the mammalian enzyme is indicated by a box, and that of the worm enzyme is indicated by a closed circle. Enzyme assays were conducted at the temperatures indicated on the abscissa, using a mammalian multisite peptide substrate EA2 with the amino acid sequence PTDTSSTPAPT. A, native enzymes were derived from whole worm homogenates or from a partially purified bovine colostrum ppGaNtase preparation (2). B, recombinant enzymes (worm GLY-3 or mouse ppGaNtase-mT1) were expressed in COS7 cells and isolated as truncated secreted proteins. Activity of recombinant enzymes was determined using 2.5 pmol of GLY-3 and 1 pmol of mouse ppGaNtase-mT1.
FIGURE 2
Amino acid sequence alignments of murine ppGaNTase-mT1 and 11 nematode sequence homologs. The gene names for each isoform appear in the title page footnote. **mT1** refers to the murine ppGaNTase-T1 isoform and the **GLY** gene name is used for the nematode clones. Nucleotide sequence is available in GenBank™, using the accession numbers indicated in the title page footnote. Alignments were performed using the Clustal method in MacVector. Amino acids below the alignment indicate consensus residues defined by a 67% threshold. The beginning and end of the 333-aa central conserved catalytic region are indicated by a left and right horizontal arrow on the consensus line. Boxes were drawn around the blocks of conserved sequences (AGGLF and CHG GGNQ) that were used to design hybridization probes and PCR primers in a previous study (6). Note that the "CHG GGNQ" sequence is not highly conserved in two clones and absent in one clone (GLY-8). The first amino acid bold circle codon that was used in the expression construct in Fig. 5 is enclosed in a box.}

**ppGaNTase Family in C. elegans**

8273

ppGaNTase-T1 protein (Figs. 2 and 3, see line labeled mT1 for mouse ppGaNTase-T1). The 333-aa conserved region begins with a "FN" amino acid sequence at position 84 in murine ppGaNTase-T1 and ends with a "KWYLNX" segment at aa position 417. The size of the evolutionarily conserved sequences in these C. elegans ppGaNTase proteins is substantially larger than the highly conserved 81-amino acid segment reported in mammals (14); however, it is smaller than the 420-amino acid region conserved in mammalian ppGaNTases-T1, -T2, -T3, and -T4 (6). Most isoforms contained putative N-glycosylation sequences in or close to the stem region (Fig. 3). Amino acid sequence alignments of the GLY-5A, -5B, and -5C proteins revealed that at position 491 these proteins varied only in 32-, 35-, and 33-aa segments, respectively, while the N-terminal and C-terminal translated sequences flanking these segments were identical (Figs. 2 and 3). Analysis of the cDNA sequence suggests that the variant gly-5 cDNAs are produced by alternative message processing, because on the nucleic acid sequence level the variant segment is introduced at the same cDNA position in each clone (Fig. 4A). Preliminary genome sequence data revealed that the variant regions of GLY-5B and GLY-5C were present in the same gene as alternatively spliced exons on a single 1.6-kilobase pair fragment (data not shown). The position of this sequence variation begins at amino acid 491 in GLY-5 and resides in a domain of the protein that has homology to sugar-binding domain of the lectin ricin (15). This ricin-like lectin domain in ppGaNTase clones is located on the C terminus and is composed of three 39–50-amino acid tandem repeats that contain a "CLD" amino acid sequence that is important in sugar binding in the ricin lectin. The functional significance of this region in the ppGaNTases is not known.

In a similar theme, the sequence variation for GLY-6A, -6B, and -6C is restricted to the C terminus in the first repeat in the ricin-like lectin domain. The nucleic acid sequences on either side of these variant regions are completely identical, indicating that three gly-6 cDNAs could also represent alternative splice variants. The splice variation in gly-6, however, appears to be more complex than for gly-5. gly-6A contains two segments of 51 and 58 nt in the variant region, while only a single 103-nt segment is present in gly-6B (Fig. 4B). The reading frames and flanking amino acid sequences are identical for these two clones (Figs. 2 and 4B). gly-6C contained two segments in the variant region; the first segment is identical to the first segment of gly-6A, and the second segment is identical to the gly-6B variant segment. However, the latter segment of gly-6C is missing one nucleotide, which produced a reading frameshift and utilization of a premature termination codon in the downstream sequence (Figs. 2, 3, and 4B). Therefore, the reading frameshift results in the translation of an 80-amino acid C-terminal segment in GLY-6C that strongly diverges from GLY-6A and GLY-6B and a premature termination codon results in a 56-amino acid truncation in the C-terminal end of GLY-6C, relative to GLY-6A. The positions marking the beginning and end of the variant segments in both gly-5 and gly-6 cDNAs correspond to comparable splice junction positions in the mammalian ppGaNTase-T1 gene and the C. elegans gly-3 gene (data not shown).

**Functional Expression**—Soluble recombinant protein expression was achieved by cloning the luminal domain of each coding region into the mammalian expression vector, pIMKF3. CDNA sequences were introduced downstream of an insulin secretion signal and a series of epitope tags (Fig. 5). The length of the 3′-UTR sequence incorporated into the expression vehi-
FIG. 3. Predicted protein domains of the ppGaNTases. Predicted amino acid sequences were used to define the size of the cytoplasmic, transmembrane, and luminal Golgi domains. A Kyte-Doolittle hydrophilicity analysis was used to size the transmembrane region (black box). The conserved catalytic domain was determined by amino acid alignments from Fig. 2. The size of each region is indicated below the box. Gaps are introduced between boxes to maximize alignments. A putative ricin-like lectin domain, predicted by Hazes (15), is indicated by a box with three segments, corresponding to three QXXW-like repeats (α, β, γ). The conserved cysteine, hydrophobic and polar residues (CLD) sequence at positions 13, 14, and 15 in each QXXW repeat are indicated below each repeat. The alternatively spliced regions in GLY-5 and GLY-6 are indicated by different hatched patterns. Predicted N-glycosylation sites are indicated by a solid lollipop. Position of N-glycosylation sites that have been experimentally mapped is indicated by an open lollipop (18).

Proteins GLY-5A, GLY-5B, and GLY-5C migrated with the smallest apparent molecular weight, while GLY-8 migrated with the smallest. The inferred molecular weight of GLY-8 supported the cDNA sequence data, indicating that the size of its coding region was significantly smaller than the other transferrase homologs, and confirmed our observations that a termination codon occurs immediately after a C-terminal HDEL sequence. This sequence confers endoplasmic reticulum retention in the yeast Saccharomyces cerevisiae (16). Immunoprecipitations of the cellular fraction of transfected COS7 cells revealed that GLY-8 and not ppGaNTase-T1 (Fig. 6B), nor any of the other nematode GLY proteins (data not shown), were efficiently retained in the cell, indicating that the HDEL-C-terminal sequence of GLY-8 could be functioning as a retrieval signal in COS7 cells. PhosphorImager quantitation indicated that approximately 85% of the GLY-8 truncated protein is retained in the cellular fraction, while for mouse ppGaNTase-T1, 85% is secreted (Fig. 6B).

To assess the functional identity of the nematode clones, ppGaNTase enzyme assays were initially performed with the recombinant GLY proteins and a multisite substrate (EA2) at two different temperatures, 23 °C and 37 °C (Table III). Because GLY-6B and GLY-6C clones were not stably expressed or did not express detectable levels of recombinant protein, only 9 of the 11 gly gene products were examined for ppGaNTase activity. Transferase assays showed that recombinant worm proteins (GLY-3, GLY-4, GLY-5A, GLY-5B, and GLY-5C) exhibited ppGaNTase activity in vitro at both temperatures; however, they catalyzed GalNAc transfer with a temperature optimum of 23 °C. GLY-7 and GLY-6A, on the other hand, showed a low rate of UDP-GalNAc hydrolytic activity, but Sep-Pak C18 reverse phase HPLC analysis revealed that this [14C]GalNAc was not transferred to the peptide acceptor substrate and the hydrolysis persisted in the absence of peptides. To assess if the GLY proteins could O-glycosylate different sequence motifs, ppGaNTase assays were performed with a panel of four diverse peptide substrates (Fig. 7). In enzyme assays at 23 °C, the worm enzymes GLY-3, -4, -5A, -5B, and -5C showed a similar preference for the multisite threonine-containing substrate EA2. GLY-4, -5A, -5B, and -5C were shown to glycosylate single-site threonine-containing peptides (EPO-T and TPPP) to a greater extent than GLY-3. The serine in EPO-S was glyco-
sylated by GLY-3, -4, -5A, -5B, and -5C, however, at an extremely low rate, relative to threonine. Therefore, none of these GLY clones appeared to have a preference of serine over threonine in the human erythropoietin-derived peptide sequence (PPDAAAPLR). The secreted truncated GLY-7 recombinant protein was abundantly expressed in the COS7 cell transfection system; however, it was not active with any peptide in this panel of substrates. Affinity-purified GLY-8 from either the secreted or intracellular fraction (Fig. 6, A and B) did not initiate O-glycosylation with the peptides tested (Fig. 7). GLY-6A and GLY-9 exhibited no detectable ppGaNTase activity; however, the level of recombinant proteins was significantly lower than the other affinity-purified enzyme preparations, and thus these enzymes may not have been as stable at the standard in vitro enzyme assay conditions.

To determine if C. elegans could be used as a model system for studying the role of glycosylation in development, we have
**Fig. 4.** Sequence alignments of the 3' coding region of the *gly-5* and *gly-6* variant clones. The complete nucleotide sequence flanking the variant regions is 100% identical and is aligned on the left and right sides of the figure. Dashes are introduced to indicate borders of the regions that vary in sequence. Borders were chosen to maximize sequence alignments and to agree with splice junction consensus sequence rules. A variant contains an amino acid sequence element (cysteine, hydrophobic residue, aspartic acid: CLD is underlined), which is present in the ricin-like motif of many ppGaNTase clones (15). The reading frame of the downstream sequences of *gly-5* and *gly-6a* are identical, while the second segment of *gly-6b* is unique, which shows that the second segment of *gly-6b* is identical to that in *gly-6c*, except that the *gly-6b* segment is one nucleotide shorter. This difference in one nucleotide produces a reading frameshift (at asterisk) in the downstream sequence of *gly-6c* (see right side of alignments).
ppGaNTase Family in C. elegans

FIG. 5. Design of vector for expression and secretion of recombinant proteins. Truncated cDNAs encoding the 10 worm GLY proteins were introduced into the expression vector plMKF3. A, plMKF1, the parent plasmid plMKF3, was previously constructed in pSVL, a SV40 promoter/enhancer-driven mammalian expression vector (6). B, plMKF3 contains an expanded multiple cloning site between MluI and BamHI. C, the nucleotide and N-terminal amino acid sequences are indicated for the secreted region of the recombinant proteins. Boxes indicate translated regions encoding for the following: I, the insulin signal peptide; M, the histidine-rich metal binding site; K, the heart muscle kinase site for 32P labeling of recombinant proteins; and F, the FLAG antibody recognition sequence, used for immunoprecipitations. The cDNA insert begins with the stem region of the GLY proteins, which is cloned into the MluI site of plMKF3 (circled residue in Fig. 2 indicates the first amino acid codon used in each expression clone); therefore, each construct directs secretion of a truncated GLY protein.

stead ends with a HDEL motif, which has been previously shown to act as a retrieval signal for lumenal endoplasmic reticulum proteins in S. cerevisiae (16). The C-terminal primary sequence has an additional source of sequence variation in two pairs of cDNAs, gly-5 and gly-6, which both contain three variant segments.

Preliminary analysis of shotgun genome sequence data indicates that at least the GLY-5B and GLY-5C proteins are encoded by a single gene that is alternatively spliced. If the variants of GLY-6 are similarly derived by alternative splicing, then the 11 homologous cDNAs will be encoded by seven different genetic loci in worms. The number of functionally active ppGaNTase isozymes detected in C. elegans (five isozymes catalyze in vitro ppGaNTase activity) is similar to the number that has been reported in mammals. However, the actual number of ppGaNTase sequence homologs in mammals is expected to be larger, because mammals have a more complex genome and because the human EST data base contains additional novel ppGaNTase-like cDNAs (8), which have not been functionally expressed at this time. The combined sequence data of the mammalian and worm ppGaNTase homologs is useful for identifying candidate amino acid residues in the active site. The amino acid sequence alignment of the ppGaNTases in mammals and C. elegans indicates numerous residues and positions that are invariant among evolutionarily diverse organisms. Site-directed mutagenesis of the invariant positions is currently being performed to identify those residues that are essential for enzyme function.

In this study, we observed that not all the recombinant GLY proteins were functionally active on the set of four mammalian peptide substrates tested. O-Glycosylation of a serine-containing human erythropoietin peptide and its threonine homolog were used to test if any of the ppGaNTases had a preference for serine. All worm transferases (as well as mammalian ppGaNTase-T1, -T3, and -T4) appeared to catalyze GaINAc transfer to threonine at a much higher rate than to serine in the erythropoietin-derived peptides, under in vitro conditions. Four members of this gene product family (GLY-6A, -7, -8, and -9) did not transfer GaINAc to the mammalian peptide substrates tested in this study, using an in vitro assay. This observation suggests that either the in vitro conditions do not reproduce the in vivo intracellular environment or that the correct substrates (pep-
Each recombinant protein was assayed at 23 °C against a panel of four peptide substrates. The peptide sequences of the acceptor substrates are: EA2 (PTTDSTTPAPTK), EPO-T (PPDAA8AAPLR), EPO-S (PPDA8AAPLR), and TPPP (TPPP). A homogeneous preparation of murine ppGaNTase-mT1 was used to quantitate the relative yield of the COS7 cell-expressed recombinant proteins, determined by PhosphorImager scanning of the Tricine-SDS-PAGE gel in Fig. 6. Activity is represented for 1 pmol of recombinant protein.

Following the analysis of the GLY-6A, -6B, -6C, -7, -8, and -9 proteins, we have observed with a polypeptide mannosyltransferase. Polypeptide mannosyltransferase 4 is functionally responsible for glycosylating the O-mannosyl protein Ggp1p in vivo in yeast, but would not glycosylate its cognate substrate in vitro, using a protocol that had functioned for other members of the polypeptide mannosyltransferase family (19). Therefore, the functional identity of the GLY-6A, -6B, -6C, -7, -8, and -9 proteins is not clear, despite their remarkable similarity to bona fide ppGaNTase enzymes. In the case of the C. elegans GLY-6A and GLY-7 recombinant proteins, we observed UDP-GalNAc hydrolytic activity in the absence of peptide substrates. This low rate of hydrolysis is a trait shared with many ppGaNTases from both worms and mammals identified to date. This suggests that the GLY-6A and GLY-7 enzymes are capable of recognizing UDP-GalNAc as a potential sugar donor. Future in vitro studies will be directed at identifying the substrates and reaction requirements of GLY-6A, -6B, -7, -8, and -9 isoforms.

Given the diversity of sequences that are O-glycosylated and the large number of ppGaNTase substrates expressed by a given cell or organism, the existence of a complex gene family of ppGaNTase isoforms is not surprising. However, it is not clear why both mammals and nematodes ppGaNTases display such a large overlap in their peptide substrate reactivity. Five members of the gene family from nematodes identified in this study (GLY-3, GLY-4, GLY-5A, GLY-5B, and GLY-5C) are each capable of glycosylating most of the peptide substrates tested, though the rates of transfer for each isoform differed. The other isoforms identified here lack observable ppGaNTase activity with those same peptide substrates. These may then represent members of the gene family with a more rigid or restricted specificity. Gene ablation studies of polypeptide mannosyltransferases in S. cerevisiae and Drosophila melanogaster have indicated that protein mannosylation is essential for viability in yeast and for the symmetry and alignment of the adult body plan and musculature in the fly (19, 20). More significantly, multiple mannosyltransferases from the yeast need to be ablated before phenotypic variance can be detected. This could prove to hold true for ppGaNTases in nematodes, as well, and may help to determine if ppGaNTase isoforms are functionally redundant in a biological model.

Acknowledgments—We thank Dr. Robert Barstead for the cDNA library; Dr. Yuji Kohara, Dr. Robert Waterston, and Dr. Lucinda Fulton for some of the EST clones; and Marlene Balys for technical assistance. We also gratefully thank Dr. Lawrence Tabak for supporting this effort and reviewing this work.

Note Added in Proof—Recent submissions to the C. elegans EST database revealed two additional ppGaNTase sequence homologs, placing the size of the family to a total of 13 isoforms, encoded by 9 genes.

REFERENCES
22. Acknowledgments—We thank Dr. Robert Barstead for the cDNA library; Dr. Yuji Kohara, Dr. Robert Waterston, and Dr. Lucinda Fulton for some of the EST clones; and Marlene Balys for technical assistance. We also gratefully thank Dr. Lawrence Tabak for supporting this effort and reviewing this work.

Note Added in Proof—Recent submissions to the C. elegans EST database revealed two additional ppGaNTase sequence homologs, placing the size of the family to a total of 13 isoforms, encoded by 9 genes.
cDNA Cloning and Expression of a Family of UDP-N-acetyl-d-galactosamine:Polypeptide-N-Acetylgalactosaminytransferase Sequence Homologs from Caenorhabditis elegans

Fred K. Hagen and Keith Nehrke

doi: 10.1074/jbc.273.14.8268

Access the most updated version of this article at http://www.jbc.org/content/273/14/8268

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 21 references, 10 of which can be accessed free at http://www.jbc.org/content/273/14/8268.full.html#ref-list-1