Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family

Eric P Bennett1,2, Ulla Mandel1, Henrik Clausen3, Thomas A Gerken4,5,6, Timothy A Fritz7, and Lawrence A Tabak8

1Department of Odontology, Copenhagen Center for Glycomics, University of Copenhagen, Norre Alle 20, DK-2200 Copenhagen N, Denmark; 2Department of Cellular and Molecular Medicine, Copenhagen Center for Glycomics, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark; 3Department of Pediatrics, 4Biochemistry and 5Chemistry, Case Western Reserve University, School of Medicine, BRB 823, 2109 Adelbert Rd, Cleveland, OH 44106-4948, USA; 6Food and Drug Administration, Center for Biologics Evaluation and Research, Office of Vaccines Research and Review, Division of Vaccines and Related Products Applications, Rockville, MD 20852, USA; and 8Department of Health and Human Services, Section on Biological Chemistry, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD 20892, USA

Received on August 19, 2011; revised on December 14, 2011; accepted on December 14, 2011

Glycosylation of proteins is an essential process in all eukaryotes and a great diversity in types of protein glyco- sylation exists in animals, plants and microorganisms. Mucin-type O-glycosylation, consisting of glycans attached via O-linked N-acetylgalactosamine (GalNAc) to serine and threonine residues, is one of the most abundant forms of protein glycosylation in animals. Although most protein glycosylation is controlled by one or two genes encoding the enzymes responsible for the initiation of glycosylation, i.e. the step where the first glycan is attached to the relevant amino acid residue in the protein, mucin-type O-glycosylation is controlled by a large family of up to 20 homologous genes encoding UDP-GalNAc:polypeptide GalNAc-transferases (GalNAc-Ts) (EC 2.4.1.41). Therefore, mucin-type O-glycosylation has the greatest potential for differential regulation in cells and tissues. The GalNAc-T family is the largest glycosyltransferase enzyme family covering a single known glycosidic linkage and it is highly conserved throughout animal evolution, although absent in bacteria, yeast and plants. Emerging studies have shown that the large number of genes (GALNTs) in the GalNAc-T family do not provide full functional redundancy and single GalNAc-T genes have been shown to be important in both animals and human. Here, we present an overview of the GalNAc-T gene family in animals and propose a classification of the genes into subfamilies, which appear to be conserved in evolution structurally as well as functionally.

Keywords: GalNAc-T / GalNAc-transferase / GALNT / monoclonal antibodies / O-glycoproteins / O-glycosylation

Mammalian protein O-glycosylation

Mucin-type O-glycosylation is initiated by a large homologous polypeptide N-acetylgalactosamine (GalNAc)-transferase (GalNAc-T; GalNAc-Ts have also appeared in the literature as polypeptide ppGalNAcT, ppGalNAc-T, ppGalNAcT and GalNAcT) family that catalyzes the first step in the biosynthesis forming the GalNAcα1-3-O-serine (Ser)/threonine (Thr) linkage in O-glycoproteins. Thus, the large number of enzymes controlling the initiation step makes mucin-type O-glycosylation unique among other types of protein glycosylation. All other types of protein glycosylation are controlled by one or two isoenzymes or in the case of N-glycosylation a complex of proteins. The O-GalNAc residues are further processed by the addition of different monosaccharides catalyzed by 30 or more distinct glycosyltransferases (Figure 1). GalNAc O-glycosylation is initiated in the Golgi apparatus after most protein folding events have taken place (Figure 1B). N-Glycosylation and other types of O-glycosylation [including O-mannose, O-fucose, O-Glc (glucose) and O-Gal (galactose added to hydroxylysine, Hyl)] of proteins in the secretory pathway are initiated in the endo-plasmic reticulum (ER). Only proteoglycan (O-Xyl) biosynthesis is initiated in the Golgi (Gotting et al. 2007). The abundant O-GlcNAc (N-acetylglucosamine) glycosylation occurs in the cytosol and nucleus and is in animals catalyzed by a single cytosolic enzyme without known homologs (Hu et al. 2010). This type of protein glycosylation is therefore not found on proteins processed in the secretory pathway, although a recent finding of the O-GlcNAc-type glycosylation on Notch is puzzling (Matsuura et al. 2008; Sakaidani et al. 2010).

These unique features of GalNAc O-glycosylation pose interesting possibilities with respect to regulation and functions of O-glycans as well as effects of genetic deficiencies in the large GalNAc-T gene family. Thus, the most pertinent and perhaps not mutually exclusive questions are whether the large number of GalNAc-T genes are used to dynamically regulate O-glycosylation to achieve differential modification...
of proteins and thereby serve specific functions or alternatively whether the large number of genes is needed to cover a wide spectrum of acceptor sequences and provide “back-up” in the face of deleterious mutations.

Deficiencies in genes controlling the initiation of other types of protein glycosylation generally cause severe phenotypes. Thus, deficiencies in the oligosaccharyltransferase complex are lethal in eukaryotes (Heesen et al. 1993; Kelleher and Gilmore 2006); deficiency in the O-mannosyltransferases (protein-O-mannosyltransferase T1 and T2 function in a heteromeric complex) leads to severe muscular dystrophies (Reeuwijk et al. 2006) and targeted disruption leads to embryonic lethality in mice (Lommel et al. 2010); deficiency in either of the two fucosyltransferases initiating O-fucosylation [protein O-fucosyltransferases 1 and 2 function with different substrates] is incompatible with life (Shi and Stanley 2003; Du et al., 2010); deficiency in the enzyme initiating O-glucosylation (POGLUT) is lethal (Stanley 2008; Acar et al. 2008; Fernandez-Valdivia et al. 2011); deficiency in protein O-xylosyltransferase 2 (XYLT2; one of the two xylosyltransferases XYL1 and XYL2) initiating proteoglycan chains leads to polycystic liver and kidney disease (Condac et al. 2007); and inactivation of lysyl hydroxylase 3 (one of three isoenzymes that precedes core Hyl galactosylation) causes early lethality in mice (Rautavuoma et al. 2004). In striking contrast, as will be discussed in the “GalNAcTs and disease” section, deficiencies in some of the vertebrate GalNAc-Ts produce only subtle phenotypes, suggesting a considerable degree of redundancy as well as unique functions of individual isoforms. Underscoring this is the observation that only one of 14 GalNAc-T isoforms in Drosophila, the pgant35A (dGalNAc-T1) gene, is essential for normal development beyond the larvae stage (Schwientek et al. 2002; Ten Hagen and Tran 2002).

There are a number of obstacles to deciphering unique functions of the GalNAc-T genes. Our knowledge of the spatio/temporal expression patterns of individual GalNAc-Ts has been hampered, in part, by a lack of appropriate immune reagents for determining specific distributions in tissues at the individual cell level as well as intracellular topology. Our understanding of the GalNAc O-glycoproteome has been largely limited to abundant glycoproteins, and, until recently, proteome-wide strategies to map sites of O-glycosylation have been missing. We hope that this situation will improve with
the introduction of our recently described SimpleCell strategy, where O-glycosylation is simplified to GalNAc alone using zinc finger gene targeting in human cells (Steenoft et al. 2011). Another obstacle for the field has been a lack of general as well as isoform-specific substrate sequence motifs for predicting sites of O-glycosylation. Progress is also being made here with new strategies for in vitro studies of substrate specificities of individual GalNAc-Ts (Schwientek et al. 2007; Gerken et al. 2011). Thus, the field is advancing rapidly through the introduction of new technologies and hints of biological functions from genome-wide association studies (GWAS).

Here, we review the current state of the large GalNAc-T gene family, its evolution, the isoenzymes and their many functions.

The GalNAc-T enzymes

The GalNAc-Ts are classified as GT27 family members in the CAZY glycosyltransferase classification (http://www.cazy.org), which is based on sequence and structure similarities and today includes 94 distinct gene families. A total of 20 human GalNAc-T gene entries are available. Of these, 17 have already been reported in the literature, whereas the remaining three are reported in the accompanying manuscript by Raman et al. Most of these GalNAc-T genes have been found to encode active polypeptide GalNAc-Ts functioning in O-glycosylation (Table I). The GalNAc-T family is highly conserved throughout metazoan evolution and although the number of members varies, all completed genomes contain large families of highly homologous sequences indicating that the structure and function of GalNAc-Ts have been conserved.

GalNAc-T activity was first identified and characterized in the bovine submaxillary gland and rat extracts (McGuire and Roseman 1967; Hagopian and Eylar 1968; Hagopian 1969), and the enzyme activity was subsequently partially purified and characterized from ovine submandibular glands (Hill et al. 1977). The enzyme activity had a broad pH optimum (6.5–8.6) and required divalent metal ions as cofactor (Mn2+ optimal) (Sugiura et al. 1982; Elhammer and Kornfeld 1986). Different strategies and sources were used for purification of the GalNAc-T enzyme activity, which resulted in isolation of two different enzymes, bovine GalNAc-T1 and human GalNAc-T2 with >80% sequence similarity (Hagen et al. 1993; Homa et al. 1993; White et al. 1995). Analysis of human gene sequences confirmed the existence of an ortholog of GalNAc-T1 with >95% sequence identity to bovine GalNAc-T1 and this was shown to encode an active enzyme, demonstrating that the human genome contained multiple GalNAc-T genes (Sorensen et al. 1995; Wandall et al. 1997). With the use of degenerate polymerase chain reaction (PCR)-based cloning strategies and emerging bioinformatics and EST (expressed sequence tagged) information, a number of additional distinct GalNAc-Ts were identified (Bennett et al. 1996; Clausen and Bennett 1996; Hagen et al. 1997; Bennett, Hassan, et al. 1998; Ten Hagen et al. 1998, 1999, 2001; Bennett, Hassan, Hollingsworth, et al. 1999; Bennett, Hassan, Mandel, et al. 1999), eventually leading to the appreciation of the existence of the large number of GalNAc-T genes present in mammals (Table I). Several recent reviews have provided updates on the continuously expanding GalNAc-T gene family and their biology (Clausen and Bennett, 1996; Hassan, Bennett et al. 2000; Ten Hagen et al. 2002; Tian and Ten Hagen 2008; Tabak 2010).

Considerable efforts have been devoted to analysis of the characteristics and functions of these enzymes and some general properties have emerged: (i) individual GalNAc-T isoforms have distinct acceptor peptide substrate specificities, although considerable overlap exists; (ii) acceptor peptide substrate specificities include both unmodified peptide sequences and partially glycosylated GalNAc-peptide substrates; (iii) a C-terminal GalNAc-T lectin domain modulates the specificity toward partially glycosylated GalNAc-peptide substrates and (iv) GalNAc-T isoforms are differentially expressed in cells and tissues during development and differentiation and marked changes in expression are found in diseases including cancer. These features suggest that GalNAc-Ts function in a coordinated fashion driven by their acceptor substrate specificities and that the cellular repertoire of GalNAc-Ts is regulated to accommodate the need for O-glycosylation of diverse protein substrates.

The domain structure of GalNAc-transferases: catalytic and lectin domains

GalNAc-Ts share the common type II membrane structure of Golgi glycosyltransferases with a short N-terminal cytoplasmic tail, a hydrophobic non-cleaved signal sequence serving as a membrane-spanning domain, a stem region of variable length and a luminal catalytic domain (Paulson and Colley 1989). However, GalNAc-Ts are unique among eukaryote glycosyltransferases in having a C-terminal ricin-like lectin domain (~120 amino acids) in addition to a catalytic unit (Hazes 1996; Imberty et al. 1997). The subcellular topology of several GalNAc-Ts has been characterized by immunocytochemistry; GalNAc-T1, -T2 and -T3 are present throughout the Golgi with isoform-specific differences in relative amounts in the different Golgi stacks (Rotger et al. 1998). The short cytoplasmic tails of all GalNAc-Ts contain basic residues that may be involved in interaction with peripheral Golgi membrane protein tethering complexes (Smith and Lupashin 2008). The stem regions are variable in length and range from around 90 amino acids (GalNAc-T1, -T3 and -T6) to 170 amino acids (GalNAc-T3, -T8 and -T15) with the exception of an extended 470-amino acid stem region found in GalNAc-T5. The stem regions are thought to displace the catalytic domains into the lumen of Golgi stacks, but more specific functions directed by the great variability in GalNAc-T stem regions remain to be determined.

The GalNAc-T catalytic domains (~230 amino acids in length) contain a G-T-A structural motif (Bourne and Henrissat 2001) characterized by two tightly interacting β-α-β Rossmann-like folds. The G-T-A motif essentially corresponds to the previously described GT1 motif of the GalNAc-Ts (Hagen et al. 1999). The Rossmann-like folds contain residues which bind to the uracil moiety of the UDP-GalNAc donor substrate (Supplementary data, Figure S1; Fritz et al. 2006;
Kubota et al. (2006). A DхX Mн2⁺ ion binding motif, which is conserved in all human isoforms and interacts with the diphasphate moiety of the donor substrate UDP leaving the group via the coordination of the Mn²⁺ ion, is located near the phosphate moiety of the donor substrate UDP leaving the is conserved in all human isoforms and interacts with the di- change in response to UDP-GalNAc binding. Residues within the catalytic domain, the donor substrate-binding residues lie within or near a loop, which undergoes a large conformational along a pre-formed channel in the surface of the catalytic domain. A subset of these residues form a “proline pocket” (Fritz et al. 2006), which explains the preference of some acceptor Thr methyl group was revealed by the crystal structure of GalNAc-T residues responsible for catalytic activity which appears to explain why GalNAc-Ts often show prefer-ence for glycosylating Thr vs Ser residues. As with other retaining GT-A glycosyltransferases, de- however, most of the acceptor substrate-binding residues lie along a pre-formed channel in the surface of the catalytic domain. A subset of these residues form a “proline pocket” (Fritz et al. 2006), which explains the preference of some GalNAc-T isoforms for acceptors containing a proline 3 residue, but redesignated GALNT10 by Schwientek et al. (2002). GALNT17 has recently been published as GALNT20 by Peng et al. (2010), but we suggest this isoform to be assigned to improved prediction of isoform-spe- cific catalytic activity. The functional role of the DхX motif has been demonstrated and, interestingly, the GalNAc-T DхX cannot be substituted for the more common DхD motif found in other glycosyltransferases using the C-terminus of the second Rossmann fold. The functional role of the DxH motif has been demonstrated and, interestingly, the GalNAc-T DxH cannot be substituted for the more common DхD motif found in other glycosyltransferases, de-
The unique lectin domains of GalNAc-Ts belong to the Ricin-type lectin structural family (Dodd and Drickamer 2001). Hazes (1996) and Imberty et al. (1997) originally proposed the presence of lectin domains on GalNAc-Ts and these lectins have been classified as carbohydrate-binding module family 13 members in the CAZy database. This module is found in many kingdoms of life including prokaryotes (excluding archaea) and adopts a β-trefoil structure composed of three homologous repeats (α, β and γ) presumed to have evolved through a series of gene duplication events (Rutener et al. 1987). The β-trefoil fold is found in superfamilies such as cytokines (interleukins and fibroblast growth factors; Ponting and Russell 2000), ricin B-like toxins (like agglutinins and hemolytic lectins; Hazes 1996), Kunitz protease inhibitors and actin-binding proteins (such as β-crystallin and hisactosan-like actin-binding proteins; Mukhopadhyay 2000; Kureishy et al. 2002; Liu et al. 2002). General characteristics of β-trefoil folds are: (i) the presence of QxW and CLD motifs in each repeat of ~40 amino acids; (ii) limited sequence similarity (Ornitz and Itoh 2001) and (iii) binding specificities as diverse as heparan sulfate, sulfated GalNAc, Xyl, Gal, Glc, lactose and GalNAc. The GalNAc-T lectin domain sequences are poorly conserved except for QxW and CLD motifs and six invariant cysteine residues that form disulfide bridges within each of the three α, β and γ repeats (Fritz et al. 2004). The carbohydrate-binding specificities of the GalNAc-T lectin domains studied have so far demonstrated rather high specificity for GalNAc and GalNAc glycopeptides (Wandall et al. 2007), but not for glycopeptides carrying elongated O-glycans such as Galβ1-3GalNAcα-O-Ser/Thr, NeuAcα2-6GalNAcα-O-Ser/Thr or GlcNAcβ1,3GalNAcα-O-Ser/Thr (Pedersen et al. 2011). Blocking experiments and inactivating mutations in the lectin domains affect GalNAc-glycopeptide substrate specificities of the enzymes. Mutational analyses have demonstrated that the α-repeat is important for GalNAc-T2, -T3 and -T4 lectin binding (Hassan et al. 2000; Wandall et al. 2007; Pedersen et al. 2011), whereas in GalNAc-T1 both the α- and β-repeats are important (Tenno et al. 2002) and in GalNAc-T10 it is the β-repeat that is important (Kubota et al. 2006). The lectin domains clearly function to modulate and improve the catalytic efficiency of GalNAc-Ts with partially GalNAc-glycosylated substrates containing a high density of acceptor sites such as those found in mucin tandem repeat sequences. How this translates to the glycosylation of large mucins in cells is unknown but it seems likely that the lectin domains serve to improve binding of GalNAc-Ts to mucin substrates in order to complete initiation of O-glycosylation before O-glycans are elongated (the processing step), as this would interfere with further addition of GalNAc residues. Since the initiation process occurs simultaneously with the processing step in the Golgi, the combined binding affinities of the catalytic and lectin domains may improve competition as long as there are acceptor sites available. This hypothesis is supported by one study demonstrating that the biosynthesis of the MUC5Ac mucin appears to go through an intermediate GalNAc-glycosylated glycoform before elongation starts (Sheehan et al. 2004). Further support may be found in a recent study demonstrating that the selective relocation of GalNAc-Ts to ER results in increased density of O-glycosylation (Gill et al. 2010). It thus seems possible that O-glycan density is in part secured by two subsets of GalNAc-Ts, one subset “initiating” on naked or low-density acceptors before another subset executes “follow-up” glycosylation by adding GalNAc to adjacent GalNAc occupied sites.

The catalytic and lectin domains are connected by a linker sequence 10–25 amino acids in length, which based on structural data display variability in flexibility. The GalNAc-T1 catalytic and lectin domains are closely associated and interact through a relatively large interfacial surface area (Fritz et al. 2004). The GalNAc-T1 lectin domain also contains an N-linked glycan at N552 (Wragg et al. 1997). Hydrogen bonding between the chitobiose core of the N-glycan and the amide nitrogen of linker residue Y428 and between residues R426 and E416 is likely to contribute to the stable association of the GalNAc-T1 lectin and catalytic domains. In contrast, the two domains of GalNAc-T2 show little to no interaction (Fritz et al. 2006). GalNAc-T2 lacks an N-linked glycan corresponding to that of the GalNAc-T1 lectin domain, but GalNAc-T10 has an N-linked glycan at N593. The N-linked glycans of GalNAc-T10 were removed by endoglycosidase H treatment prior to crystallization; thus, it remains unclear whether the N593 glycan on T10 affects linker flexibility of this isoform. Sequons for N-linked glycans corresponding to that in the GalNAc-T1 lectin domain are also present in human GalNAc-T13 and -T17. The relative positioning of the two domains has been attributed, in part, to differences in linker sequence properties. Peripheral GalNAc-T1 linker residues were shown to interact with both the catalytic and lectin domains, whereas a more stretched linker in GalNAc-T10 results in weaker interaction with the catalytic domain (Kubota et al. 2006). Therefore, linker flexibility may function to control the relative orientation of the lectin and catalytic domains and serve to drive lectin-mediated substrate specificities of GalNAc-Ts. This has been shown in experiments involving the replacement of the linker in GalNAc-T10 with the linker of GalNAc-T1, which resulted in altered substrate specificity of the chimeric enzyme from glycopeptide to broad peptide specificity (Kubota et al. 2006). However, in contrast to this, a similar study demonstrated that deleting the GalNAc-T10 lectin domain or exchanging it with the corresponding GalNAc-T2 lectin domain did not alter the specificity of the mutant chimeric isoform compared with normal wild-type GalNAc-T10 (Raman et al. 2008). This discrepancy in results may be due to different experimental conditions and acceptor substrates used or due to differences in the GalNAc-T10 linker region start positioning (Phe449 in the former and Pro455 in the later study). Hence, further studies are needed to clarify this.

Although studies of the binding specificities of GalNAc-T lectins analyzed so far demonstrate binding to the GalNAc monosaccharide residue, recent novel experimental approaches based on glycopeptide beads and microarrays have
demonstrated differential binding specificities of GalNAc-T1, -T2, -T3 and -T4 lectins toward different subsets of small synthetic and recombinant GalNAc glycopeptides (Pedersen et al. 2011). This suggests that the lectins serve more complex roles in modulating functions of GalNAc-Ts.

Peptide and glycopeptide substrate specificities of GalNAc-transferases

The peptide and glycopeptide substrate specificities of the GalNAc-T family are still relatively poorly understood. Studies of the acceptor specificities of GalNAc-Ts have generally been limited to in vitro assays with a limited number of peptides. Nevertheless, substrate specificities identified with such peptides have been shown to correlate very well with functions on corresponding proteins in cells (Nehrke et al. 1998; DeFrees et al. 2006; Kato et al. 2006; Schjoldager et al. 2010). More recently, the application of random peptide and glycopeptide substrate libraries has allowed quantitative determination of neighboring residue preferences and emergence of GalNAc-T isoform-specific acceptor sequence motifs (Gerken et al. 2006, 2008, 2011; Perrine et al. 2009). This strategy is based on peptide substrates with the general motif: GAGA(\(X\))\(_n\)T(\(X\))\(_n\)GAGA (where \(X\) = subset of randomized AA residues including Ser-O-GalNAc and \(n = 3–5\)). These relatively short substrates are thought to reveal the catalytic domain peptide and glycopeptide specificity independent of the lectin domain. Presently, randomized peptide substrate preferences have been reported for GalNAc-T1, -T2, -T3, -T5, -T10 and -T12 and the fly orthologs of GalNAc-T1 and -T2 (pgant5 and pgant2, respectively; Gerken et al. 2008), whereas Ser/Thr-O-GalNAc preferences have been reported only for GalNAc-T1, -T2 and -T10 (Gerken et al. 2006, 2011; Perrine et al. 2009). Using these substrates, the mammalian and fly GalNAc-T1 and -T2 orthologs have been shown to have nearly identical peptide substrate specificities (Gerken et al. 2008). This approach suggests that the principal sequence motif determining the substrate specificity of GalNAc-Ts reside in the three neighboring N- and C-terminal residues flanking sites of glycosylation. Interestingly, all the transferases studied so far, except GalNAc-T10, display a roughly similar C-terminal sequence motif of Pro, Gly/Ala and Pro for the +1 to +3 positions relative to the site of glycosylation. This common proline motif correlates in these transferases with the presence of aromatic residues forming the previously mentioned “proline pocket” (Fritz et al. 2006). Several other uncharacterized GalNAc-Ts contain these residues, which are therefore expected to possess similar C-terminal proline motifs as well (Gerken et al. 2011). GalNAc-T10 lacks the key “proline pocket” residues and very strongly prefers Ser-O-GalNAc at the +1 position, clearly showing the catalytic domain of T10 directly binds glycopeptides (Perrine et al. 2009). This further suggests that the T10 lectin domain is dispensable for the “follow-up” glycosylation of certain glycopeptide substrates. Whether the remaining glycopeptide requiring GalNAc-Ts in subgroup IIb (see Figure 2 and the following section for a description of the subgroup definition), i.e. GalNAc-T7 and -T17, possess similar catalytic domain glycopeptide-binding properties has yet to be determined. In contrast to the C-terminal preferences, the N-terminal preferences show a wide range of isoform-specific enhancements with elevated preferences for Pro, Val and Tyr being the most common at the −3 to −1 positions (Gerken et al. 2011). Transferase specific enhancement values can be used to predict sites of glycosylation at the Isoform Specific O-glycosylation Prediction (ISOGlyP) web site, http://isoglyp. utep.edu. Thus, to date, it seems that for “proline pocket” containing transferases acceptor substrate specificity will be dominated by each isoform’s different sensitivity to residues N-terminal of the site of glycosylation. It was recently shown that the overall preferences for charged residues vary among isoforms and this can be correlated with the calculated surface charge of the transferase (Gerken et al. 2011).

The human GALNT gene family

The 20 human GALNT genes localize to different chromosomal loci except for two sets of clustered genes (Table 1). One cluster, GALNT7/T17, is separated by 100 kb and another, GALNT11/T20, is spaced 8.5 kb apart. In both cases, there are no predicted ORFs (open reading frames) between the clustered genes. GALNT5 and T13 are also located close to each other but the genes are interspaced by 1.5 Mb containing multiple ORFs encoding known proteins. The sizes of the individual gene loci for GALNTs vary from 5 kb for T4 to >1.2 Mb for T17. All genes contain multiple coding exons (8–16 exons) except for GALNT4. Analysis of GALNT intron/exon positioning reveals that three intron positions are conserved in all genes (conserved intron boundaries/positions C1–C3, Figure 2) and that others are conserved in subsets of genes (Bennett, Weghuis, et al. 1998). This suggests that all GALNT genes originate from one ancestral GALNT gene that contained at least these three introns and that additional introns were gained during the evolution of various paralogs. As described in the following section, the positioning of the C1–C3 introns has been conserved in lower species. The accumulation of introns observed during evolution thus suggests that a common ancestor of deuterostome GALNTs was intron poor and that introns were gained during the evolution of the mammalian GALNT gene family. For all GALNTs, the exons encoding the catalytic unit are flanked at the 5′ end by the conserved C1 intron and at the 3′ end by an intron that is semi-conserved in 16 of the GALNT genes. The majority of introns found within the lectin domains have been identically positioned in 16 of the GALNTs and in these cases the intron positions relative to the ORF (intron phase) are zero, i.e. the intron is positioned between two codons (Figure 2).

One gene, GALNT4, stands out among the genes by being encoded in a single exon (Figure 2). GALNT4 is unlikely to represent an ancestral gene since intronless GALNT genes are not found in lower organisms. We therefore previously proposed that GALNT4 arose through a retropositional event (Bennett, Hassan, et al. 1998). The most likely candidate gene giving rise to GALNT4 is GALNT12, which exhibits the highest sequence similarity and has nine introns (Figure 2). GALNT4 is located within the second intron of the WD40 repeat domain-containing cartwheel protein, Poc1, which is required for the structural maintenance of centrioles.
Sequence alignments of the 5′- or 3′-untranslated regions (UTRs) of \( \text{GALNT4} \) and \( \text{T12} \) and with other GalNAc-T UTRs did not reveal any similarities. Analysis of repetitive elements flanking the \( \text{GALNT4} \) and \( \text{T12} \) loci showed that AluY elements (Batzер et al. 1996) are found flanking the single \( \text{GALNT4} \) exon, whereas AluSx&j elements flank both \( \text{GALNT4} \) and \( \text{T12} \) transcription units. Since AluY elements have arisen late in evolution [25 million years ago (MYA); Batzer and Deininger 2002], \( \text{GALNT4} \) must have arisen relatively late and not earlier than 25 MYA. These latter findings provide further support for the hypothesis that \( \text{GALNT4} \) arose as a transposition event.

The \( \text{GALNT} \) genes have also given rise to a number of pseudogenes (\( \text{ps} \)genes). The first example described was a \( \text{ps} \)gene derived from \( \text{GALNT1} \) (Meurer et al. 1996) and we have identified additional four \( \text{ps} \)genes. One additional
GalNAc-T20 has an invariant glycine (G334R) residue in the transcriptional units does not occur (personal observation). That alternative splicing between the two closely spaced transcripts designed to demonstrate alternative splicing events suggest and the two genes are spaced 8.5 kb apart. Experiments to maintain unique intron positions. Thus, three unique introns in group into subfamilies in the phylogenetic analysis and characterization of the enzymes would benefit from grouping into subfamilies with related functions. Based on the example of the GALNT3 and T6 subfamilies (Bennett, Hassan, Mandel, et al. 1999), we therefore propose to classify all GALNT genes into two major groups, I and II, with predicted predominant peptide and GalNAc-peptide substrate specificities, respectively. Group I genes are further subdivided into seven distinct subfamilies (Ia, Ib, Ic, Id, Ie, If, and Ig) and group II genes into two subfamilies (Ii and IiB). Paralogs within the classified subfamilies share identical intron number and positioning except for minor variations in introns positioned in the most 5' region. These include subfamily Ib (GALNT2/ T14/T16) with 15 introns in GALNT2 and 14 introns in T14 and T16 and subfamily le (GALNT8/T9/T18/T19), where the first intron in GALNT8 is shifted 111 bp relative to the first intron position in T9/T18/T19.

Two genes, GALNT3 and T15, do not show a significant relationship with these subfamilies. These two genes did not group into subfamilies in the phylogenetic analysis and contained unique intron positions. Thus, three unique introns in GALNT5 (uT5a/b/c) were identified at positions not found in any other GALNT gene. One specific intron boundary in GALNT15 (sT15) was also found only in GALNT7. This particular intron in GALNT7 is not conserved in position in the two other members belonging to the IIb subfamily (GALNT10/T17). On the other hand, all three members (GALNT7/T10/T17) share identical positioning of an intron (intron 8 in GALNT7) not found in GALNT15 (Figure 2).

GALNT20 lacks the exons encoding the lectin domain. GALNT20 colocalizes on chromosome 7q36 with GALNT11 and the two genes are spaced 8.5 kb apart. Experiments designed to demonstrate alternative splicing events suggest that alternative splicing between the two closely spaced transcriptional units does not occur (personal observation). GalNAc-T20 has an invariant glycine (G334R) residue in the Gal/GalNAc-T motif WG(G/R)EN (Supplementary data, Figure S1). The corresponding residues in the GalNAc-T1 crystal structure have been shown to be involved in essential UDP-GalNAc binding and are also found in the seven members of the β4Gal-T family. β4Gal-T7 also possess the invariant WG(G/R)EN mutation and, interestingly, this isoform has a different acceptor substrate specificity from the other members of the β4Gal-T family in that it transfers Gal to Xyl, which is essential for synthesis of the link region of proteoglycans (Almeida et al. 1999) instead of to GlcNAc. We have been unable to express the secreted GalNAc-T20 protein in insect cells and in COS7 cells, and C-terminally tagged full coding constructs failed to demonstrate Golgi localization (unpublished and accompanying manuscript by Raman et al.). Thus, GalNAc-T20 may represent: (i) a functional gene with yet unknown function and subcellular localization; (ii) a conserved expressed psgene or (iii) a gene with the potentiality for becoming a functional gene, a so-called potogene (Balakirev and Ayala 2003).

The currently available data on the enzymatic functions of GalNAc-Ts support the proposed subfamily classification, although the GALNT3/T6 subfamily is the best characterized (Bennett, Hassan, Mandel, et al. 1999). Within group I, seven subfamilies are found (Ia–Ig). Subfamily Ia (T1/T13) enzymes have similar peptide specificities, although one study has claimed that there are differences in the ability to glycosylate Syndecan-3 and MUC7-derived peptides (Zhang et al. 2003). However, we have not been able to confirm this (unpublished). In subfamily Ib (T2/T14/T16), GalNAc-T2 is the only well-characterized isoform in the literature, but our preliminary studies of these three enzymes show related functions with many substrates (unpublished and accompanying manuscript by Raman et al.). Subfamily le (T8/T9/T18/T19) enzymes are the least characterized. GalNAc-T9 was shown to have low activity with a small panel of peptides (Zhang et al. 2003), and in the accompanying paper GalNAc-T18 is also shown to have low activity with peptide substrates. In subfamily If (T11/T20) only GalNAc-T11 has been characterized and shown to have unique specificity compared with other GalNAc-Ts (Schwientek et al. 2002; Ten Hagen and Tran 2002). GALNT20 is different from all other GALNT genes in that the ORF lacks the C-terminal sequence encoding the lectin domain, and it remains to be verified that it does not represent a psgene. GalNAc-T5, constituting subfamily Id, and GalNAc-T15, constituting subfamily Ig, both show efficient activity with several peptide substrates (Ten Hagen et al. 1998; Cheng et al. 2004; Gerken et al. 2011).

Within group II, two enzyme subfamilies are found (Ii and IiB). Subfamily IiA GalNAc-T4 and T12 (T4/T12) share common GalNAc-glycopeptide substrates (unpublished). Interestingly, GalNAc-T4 is the only enzyme that uses GalNAc glycopeptides from the MUC1 tandem repeat, and hence the only known isoform that can complete the glycosylation of the MUC1 tandem repeat (Bennett, Hassan, et al. 1998; Hassan et al. 2000). GalNAc-T12 was originally reported to have very low activity toward an MUC5AC peptide substrate (Guo et al. 2002), but our preliminary studies clearly demonstrate that this enzyme primarily
Evolution of the GalNAc-T gene family

Studies on the evolution of GalNAc-Ts have calculated the substitution rates (between human and mouse) to be as low as those observed for histone and actin genes, and it was estimated that the GalNAc-T genes appeared among the first glycosyltransferases in early eukaryote evolution around 1200 MYA (Kaneko et al. 2000). Conservation of the intron phase (0) between the catalytic and lectin domains of the majority of GalNAc-transferases (Figure 2) suggests that the unique lectin domains on GalNAc-Ts arose via fusion with lectin modules early in evolution. A phylogram based on 102 putative GalNAc-T sequences from completed or preliminary assembled genomes depicts the size of the gene family in diverse species including Pan troglodytes (chimpanzee), Mus musculus (mouse), Gallus gallus (chicken), Xenopus tropicalis (frog), Caenorhabditis elegans (worm), Drosophila melanogaster (fly), Danio rerio (fish) and Toxoplasma gondii (microbe) (Figure 3, Table I). This analysis clearly demonstrates that all these organisms have a large number of GalNAc-transferases available and that there is a functional requirement for multiple isoforms with members from most of the classified subfamilies.

The fish GalNAc-T gene family was found to be the largest, containing 21 members, followed by the human and chimpanzee gene families with 20 members. Nineteen members were identified in rodents, 17 in chicken, 16 in frog, 14 in fly, 9 in worm and 5 in the microbial parasitic protozoan T. gondii. Notable differences in the size of individual subfamilies in different species include: (i) subfamily Id is missing (lacks a GALNT5 ortholog) in frog; (ii) Ie lacks GALNT8 orthologs in frog, chicken and mouse and GALNT18 is lacking in frog; (iii) subfamily Ig is missing (lacks a GALNT15 ortholog) in fish; (iv) subfamily If lacks GALNT20 orthologs in frog, chicken and fish and (v) subfamily Ii lacks a GALNT4 ortholog in chicken. Perhaps more significantly, subfamilies Ic and Ia are not found in worm and fly, and the low bootstrap confidence at the in vertebrate Ie node in Figure 3 (1%) suggests that these genes evolved separate from the mammalian Ie subfamily. The microbial GALNT group in the Ie clade (Gondi-T1, -T2 and -T3) or Ie clade (Gondi-T4 and -T5). Due to the neglectable bootstrapping value (0) seen at the node branching the deuterostome and protostome Ie lineages (Figure 3), it thus seems that the Ic, Ia and most likely the Ie lineages only arose in deuterostomes (including mammals). The mouse genome contains 19 GALNTs all located at syngenic human loci and these genes have identical genomic organization to the human genes. This variation in the number of paralogs and subfamilies in different species is likely to be a consequence of two evolutionary mechanisms: (i) lineage-specific paralog loss, which in a given species results in the specific loss of certain paralogs within a gene family and (ii) whole genome duplication (WGD) events that can result in species-specific genome duplication (Catchen et al. 2009). The fact that seven human subfamilies contain at least two members is in line with the proposed WGD model, which occurred twice during human evolution (Abi-Rached et al. 2002; McLysaght et al. 2002; Lundin et al. 2003; Dehal and Boore 2005). Time estimates have suggested that a primary duplication event occurred prior to the fish-tetrapod split followed by a second distinct genome duplication event early in vertebrate evolution (Jaillon et al. 2004). These proposed events are supported by studies establishing that the protostome (including nematodes/worm and arthropod/fly) and deuterostome lineages diverged 974 MYA and that mammal and actinopterygian fish (ray-finned fish) diverged 450 MYA (Hedges et al. 2004; Roger and Hug 2006). Thus, the human/fish and fly/worm GALNT families most likely evolved differently, which has resulted in a significant expansion of the Ie subfamily in fish with four GALNT8 paralogs, whereas this subfamily has been excluded in mouse and most likely in fly and worm too (Figure 3). The fish GALNT8 paralogs form two closely related pairs that reside on D. rerio chromosome 25 (GALNT8a/b) and chromosome 4 (GALNT8c/d) and are spaced 30 and 10 kb apart, respectively. The fish GALNT8c/d catalytic domain DNA sequences show 96% similarity. In addition, a fifth D. rerio GALNT8 ortholog may exist (BX004976.1), although this gene lacks the lectin domain encoded sequences and appears not to be transcribed (http://www.ensembl.org/). The D. melanogaster subfamily Iib is another example of species-specific subfamily expansion. The human Iib subfamily contains three genes (GALNT7/T10/T17), whereas 7 of the 14 GalNAc-Ts in D. melanogaster belong to subfamily Iib (Figure 3). The fly Iib subfamily contains the pgant8 cluster (CG7579, CG7304 and CG7297) colocalized to a 10-kb chromosome 3L locus and the pgant4 cluster (CG31856 and CG31776) localized to a 5-kb chromosome 2L locus. It should be noted that the CG7304 and CG7579 isoforms lack the essential DxH motif (found as DAQ and NGH, respectively) and they may represent non-functional gene elements. The enzymatic functions of two genes CG6394 (pgant7) and CG2103 (pgant6) within the fly subfamily Iib have been characterized and shown to exhibit the predicted glycopeptide substrate specificities (Ten Hagen et al. 2003).

Analysis of intron/exon positioning in human, fish, fly and worm support the model that GALNTs evolved from a common ancestral gene (Clausen and Bennett 1996). Of the three conserved human GALNT intron/exon boundaries, C1–C3 (Figure 2), both C1 and C2 are identical in fish T1, T5, T11 and T14, as well as fly CG3254, CG31651, CG7480 and CG4445 genes (Supplementary data, Figure S2). C1 is identical in worm Gly5 and 8, as well as fly CG7279, CG2103, CG31776, CG7579 and CG31856. C2 is identical in fly CG30463 and worm Gly7 and Gly9 genes.
Fig. 3. Phylogenetic tree of 102 GalNAc-Ts from human (Homo sapiens), frog (X. tropicalis), chicken (G. gallus), fish (D. rerio), fly (D. melanogaster), worm (C. elegans) and Gondi (T. gondii). The following Genbank accession numbers were used in addition to those of Table I. GALNT1: X. tropicalis NP_001025547.1, D. rerio XP_687472.2, D. melanogaster NP_001036338.1 (CG31651), G. gallus NP_001006381.1, C. elegans NP_498722.1 (gly-3). GALNT2: X. tropicalis XP_002931524.1, D. melanogaster NP_001006381.1, C. elegans NP_498722.1 (gly-4). GALNT3: X.
The prediction of true GalNAc-T orthologs among the different species analyzed in Figure 3 involves obvious uncertainties. Several studies have provided evidence that predicted orthologs from human, mouse and fly represent true functional orthologs (Schwientek et al. 2002; Ten Hagen and Tran 2002; Gerken et al. 2008). Interestingly, however, we found that the predicted orthologous genes in subfamily I, human GALNT11 and fly CG7480 (pgant35A), which were shown to have similar enzymatic functions (Schwientek et al. 2002), failed to complement each other in the lethal l(2)35Aa fly phenotype (Bennett et al. 2010).

Several parasitic genomes contain multiple GALNTs. Some of these have been expressed and shown to possess functional enzymes including ones from the protozoan apicomplexan parasite T. gondii (Stwora-Wojczyk et al. 2004). O-Linked GalNAc has been detected in the parasitic helminth Schistosoma mansoni (Nyame et al. 1987) and the protozoan parasite Trypanosoma cruzi (Previato et al. 1998). Recent studies of T. cruzi genes have revealed the presence of two T. cruzi genes (TeOgNT1 and TeOgNT2) encoding enzymes distantly related to the mammalian GalNAc-T gene family. Both T. cruzi genes utilize UDP-GlcNAc and produce GlcNAc1-O-Ser/Thr linkages (Heise et al. 2009).

Although GALNT orthologs were not identified in bacteria, plants or yeast, the presence of core mucin-type O-glycans in the fungus Cryptococcus neoformans has been reported (Kawaguchi et al. 1986).

Collectively, genetic, phylogenetic and functional information on GalNAc-Ts suggest that certain orthologs have been functionally conserved during evolution and that the evolution of the GalNAc-T family has resulted in both loss and expansion in the size and the number of subfamilies in different species.

Expression of GalNAc-transferases

The GalNAc-Ts are differentially expressed in cells and tissues and marked changes in expression are found in cancer. Northern analyses of the first identified GalNAc-Ts showed that distinct isoforms have different expression patterns in mouse, rat (Hagen et al. 1997; Ten Hagen et al. 1998) and human organs (Bennett et al. 1996; Bennett, Hassan, et al. 1998). More comprehensive Northern studies indicate that some GalNAc-Ts are more ubiquitously expressed in organs, e.g., GALNT1 and T2 (Homa et al. 1993; White et al. 1995), whereas, for example, GALNT3, T4, T5, T7, T8 and T10 have more restricted expression patterns (Bennett et al. 1996; Bennett, Hassan, et al. 1998; Bennett, Hassan, Hollingsworth, et al. 1999; Ten Hagen et al. 1998, 2001; White et al. 2000; Table II). Northern analyses of human GALNT9 and rat GALNT19 indicate that expression of these genes is largely restricted to the brain (Toba et al. 2000; Nakamura et al. 2005), although human GALNT19 expression is also detected in a few other tissues (Table II; accompanying manuscript by Raman et al.). Northern and quantitative analyses of GALNT6 and T18 display broad expression profiles (unpublished and accompanying manuscript by Raman et al.).

Expression of GALNT20 by Northern analysis (unpublished) and quantitative reverse transcriptase (RT)-PCR analysis seems to be quite restricted to testis (accompanying manuscript by Raman et al.). Quantitative RT-PCR analysis of GALNT12 shows expression in several tissues (Guo et al. 2002), whereas GALNT13, T14 and T15 display broader expression patterns (Wang et al. 2003; Zhang et al. 2003; Cheng et al. 2004).

Real-time RT-PCR expression profiling in mouse organs has corroborated a rather ubiquitous expression pattern for galnt1 and t2 and detected differential profiles for several of the other mouse genes such as galnt4 (detected in the colon, lung and sublingual gland), t12 (detected in the colon, spleen, thymus and sublingual gland) and t13 (not detected in any of the organs studied) (Young et al. 2003). In situ hybridization was used to profile the spatiotemporal expression pattern of seven murine galnts (galnt1/t2/t3/t4/t5/t7/t9) during mouse development, and in particular, galnt3, t5 and t7 were found to display unique patterns (Kingsley et al. 2000). RT-PCR and in situ hybridization have also been used to profile expression in Drosophila, and comprehensive semi-quantitative PCR analysis revealed expression of most of the eight isoforms tested during various developmental stages (Ten Hagen et al. 2003). Later studies of 12 of the 14 isoforms in the fly demonstrated unique expression patterns for CG3254 (pgant2), CG4445 (pgant3), CG7480 (pgant35A) and CG30463 during development. In particular, the CG3254 and...
CG30463 isoforms were found to display specific expression patterns related to eye development (Tian and Ten Hagen 2006).

Gene expression profiles can also be predicted “in silico” from the vast amount of information available from EST gene expression surveys. These surveys represent up to 125 normal, disease and development stage tissues and tissue-specific abundance information of individual human EST genes are provided (NCBI web site).

Table II summarizes our current understanding of the expression patterns of GALNTs as evaluated by mRNA as well as more limited information on protein expression by immunohistochemistry. Overall, there is reasonable agreement between expression results obtained from conventional Northern and immunohistochemical analyses, but the latter provides further detailed information on specific cell-type expression as well as subcellular topology. Discrepancies are observed. For instance, according to immunohistochemical experiments, GaINAc-T4, -T6 and -T12 were not expressed in the kidney, whereas the corresponding RNAs appear to be expressed (Table II). Similarly, the GaINAc-T14 enzyme was not found in the lung (Stern et al. 2010), whereas RNA appears to be expressed. Several issues may give rise to the discrepancies observed, such as differences in detection sensitivity of the methods used, and biological issues related to RNA and protein transcriptional and translational control and stability, which underscore the need for further systematic studies in this area. Regardless, we have attempted to summarize GaINAc-T expression in Table II with a relative expression score for each of the GALNT genes.

Direct monitoring of enzyme protein in cells is clearly preferable in order to assess the actual enzyme expression level, topology and function. We and others have put efforts into developing antibodies suitable for immunohistochemistry. A general problem has been selecting antibodies that react with both the native and the denatured proteins, and only a few monoclonal antibodies (MAbs) have been identified with these characteristics. We have built a library of MAbs to human GaINAc-Ts (T1-T4, -T6, -T11, -T12 and -T14) and selected pairs of MAbs that react with either the native soluble enzymes or the denatured protein (Figure 4; Bennett, Hassan, et al. 1998; Bennett, Hassan, Mandel, et al. 1999; Mandel et al. 1999; Schwientek et al. 2002).

Immunohistochemical studies with these antibodies on frozen sections and cell lines have extended our understanding of differential expression of GaINAc-Ts and demonstrated that even the more ubiquitously expressed isoforms such as GaINAc-T1 and -T2 indeed have cell-specific expression patterns. An illustrative example is the multilayered stratified squamous epithelium of oral mucosa with its well-defined
differentiation pattern where GalNAc-Ts are differentially expressed. Thus, GalNAc-T2 is expressed in the lower immature proliferative cell layers, GalNAc-T1 in the differentiated superficial cell layers and GalNAc-T3 in all cell layers. The application of antibodies has, furthermore, clearly demonstrated that the repertoire of GalNAc-Ts in cells changes during malignant transformation (Mandel et al. 1999). In general, cells express multiple GalNAc-Ts. The transcriptome of the important Chinese hamster ovary cell line CHO-K1 was recently characterized and it was found that

---

**Fig. 4.** Illustration of expression patterns of eight GalNAc-T isoforms in normal (A) salivary glands, (B) kidney, (C) colon and in (D) colon adenocarcinomas evaluated by immunofluorescence histology using MAb's with well-characterized specificities. Each panel is labeled with the GalNAc-T isoform analyzed [GalNAc-T1 with MAb UH3 (4D8); GalNAc-T2 with MAb UH4 (4C4); GalNAc-T3 with MAb UH5 (2D10); GalNAc-T4 with MAb UH6 (4G2); GalNAc-T6 with MAb UH7 (2F3); GalNAc-T11 with MAb UH8 (1B2); GalNAc-T12 with MAb 1F9 (unpublished) and GalNAc-T14 with MAb 3D2 (unpublished)] (Bennett, Hassan, et al. 1998; Bennett, Hassan, Mandel, et al. 1999; Mandel et al. 1999; Schwientek et al. 2002), and the protocol used for staining fresh frozen sections was as described previously (Mandel et al. 1999). Positive FITC fluorescence is shown in green. (A) Neighboring sections were also stained with PAS, HE or MAb PMH1 to human GalNAc-glycosylated MUC2 as indicated (Reis et al. 1998). The PAS staining of salivary glands clearly marks mucous acini (indicated by asterisks), serous acini (indicated by arrows) and duct cells (indicated by crosses). (B) The HE staining of kidney marks glomeruli (indicated by asterisks) and tubules (indicated by arrows). (C and D) Staining for MUC2 in colon tissues marks goblet cells (indicated by an arrow). Colon tissues were counterstained with Dapi nuclear stain in blue. 20 μM scale bar is included in figures.
only four GalNAc-Ts (-T2, -T7, -T11 and -T19) are expressed (Xu et al. 2011). One exception to this may be sperm cells, where we have examined expression of six GalNAc-Ts (GalNAc-T1, -T2, -T3, -T4, -T6 and -T11) and found that only GalNAc-T3 is expressed in the acrosome of spermatozoa (Bennett, Hassan, Mandel, et al. 1999; Mandel et al. 1999). This may be one example where the existence of a subfamily of GalNAc-Ts (GalNAc-T3 and -T6) with similar properties cannot provide a complete functional back-up due to differential expression (Rajpert-De et al. 2007). Two mouse knockout models of galnt3 exhibit testicular calcifications and male infertility (Esapa et al. 2009; Ichikawa et al. 2009; Duncan et al. 2011), but it is unclear if this also applies to human.

Deiciency in GALNT3 is found in the related diseases FTC (familial tumoral calcinosis) and HHS (hyperostosis hyperphosphatemia syndrome) and one patient studied has been reported to exhibit testicular microlithiasis and oligozoosperma (Campagnoli et al. 2006). However, family studies appear to indicate that males with FTC have produced offspring (Ichikawa et al. 2005; Carmichael et al. 2009). Further studies are clearly needed to assess the function of GALNT3 in fertility.

Examples of the exquisite cell-specific expression of GalNAc-Ts found in salivary glands, kidney and colon are illustrated in Figure 4. Although a large number of GalNAc-Ts are expressed in salivary glands (GalNAc-T11
and -T14 are not expressed), several isoforms show specific expression in different cell types. Thus, GalNAc-T3 is observed in the duct, serous and mucous cells, whereas GalNAc-T4 (Bennett, Hassan, et al. 1998) and -T6 are only weakly expressed in serous and strongly in mucous cells (Figure 4A). Similarly, in the kidney, GalNAc-T1 and -T2 are found in both tubules and glomeruli, whereas GalNAc-T11 (Schwientek et al. 2002) and -T14 are found only in tubules.

The repertoire of GalNAc-Ts also changes markedly in cancer. A number of studies have documented altered expression of GalNAc-Ts in cancer (Kohsaki et al. 2000; Shibao et al. 2002; Gu et al. 2004; Ishikawa et al. 2004; Landers et al. 2004; Miyahara et al. 2004; Yamamoto et al. 2004; Inoue et al. 2007). For example, GalNAc-T6 is not expressed in normal colon, but highly expressed in colon adenocarcinomas (Figure 4C and D). GalNAc-T6 has been associated with gastric carcinogenesis (Gomes et al. 2009), and T6 has also been associated with tumor stage ductal breast carcinomas (Berois et al. 2006) and is thus suggested as a new marker for breast cancer cell detection (Freire et al. 2006). In relation to this, a recent study has presented data proposing that elevated expression of GalNAc-T6 in breast cancer cells correlates with increased glycosylation and surface expression of the mucin MUC1 (Park et al. 2010). In a subsequent study, the same authors found that GalNAc-T6 was also important for the function of fibronectin-mediated adhesion (Park et al. 2011). Although these findings are intriguing, RNA knockdown strategies are known to have off-target effects and further studies are needed to support the conclusions. It is well-established that cancer cells produce aberrant glycosylation of proteins and lipids (Hakomori 1989, 1996; Dabelsteen 1996). However, the current information on aberrant glycosylation is generally limited to the structures of glycans and not to sites of attachment to proteins. Thus, our understanding of the effects of altered GalNAc-T expression in cancer is highly limited. It could be predicted that altered repertoire or topology of GalNAc-Ts in cells will result in altered density and patterns of O-glycans on proteins, but the means to address this have been limited. The only documented example of this is the apparent increased density of O-glycans on MUC1 in breast cancer cells (Muller et al. 1999; Muller and Hansich 2002), but it is uncertain whether this relates to changes in the GalNAc-T repertoire or altered processing (Dalziel et al. 2001).

Evidence suggests that malignant transformation is associated with marked disorganization of the Golgi apparatus and that these changes may be responsible for the formation of cancer-associated glycosylation abnormalities (Kellokumpu et al. 2002). It has been put forward that a balanced signaling system monitors and controls Golgi vesicular trafficking rates. This signaling system involves protein chaperones, cytoskeletal proteins and a Golgi pool of Src kinases implicated in maintaining the dynamic equilibrium of the Golgi complex (Bard et al. 2003; Pulvirenti et al. 2008). It was recently shown that Src specifically regulates Golgi to ER retrograde traffic of GalNAc-Ts either directly (Gill et al. 2011) and/or through activation of the GTPase dynamin 2, a factor believed to be involved in Golgi vesiculation during secretion (Weller et al. 2010). Furthermore, a recent report suggests that certain GalNAc-T isoforms localize to the ER in cancer (Wang et al. 2011). Thus, the traditional view of an exclusive Golgi-localized subcellular topology of GalNAc-Ts is being challenged by these novel findings.

The regulatory events controlling GALNT expression still remain to be explored. So far, analysis of an upstream region of GALNT3 is the only report of the gene regulatory elements controlling GALNT gene expression (Nomoto et al. 1999). GALNT gene expression may also be modulated by microRNAs. One study proposed that the urothelium specific miR-129 targets pathways involved in cell death processes associated with poor bladder cancer outcome and one of the putative targets in this study was shown to be GALNT1 (Dyrskjot et al. 2009). Another study suggested that miR-378 involved in osteoblast differentiation targets GALNT7 (Kahai et al. 2009). More recently, miR-30d was shown to confer a pro-metastatic cancer effect through increased IL10 expression mediated by GALNT7 down regulation (Gaziel-Sovran et al. 2011). Interestingly, this study found that expression of multiple GALNTs was inversely correlated with miR-30d levels, suggesting that miR-30d regulates more isoforms.

Alternative splicing of GALNTs may affect functions. A large number of alternative splice variants are detected “in silico” (www.ensemble.com) for many GALNTs, but most of these are unlikely to encode functional proteins. In the accompanying paper, Raman et al. demonstrate that one GalNAc-T13V1 splice variant lacking part of its lectin domain had similar activity toward a panel of peptides and glycopeptides.

GalNAc-Ts and disease

A large number of congenital diseases of glycosylation (CDGs) have emerged in the last two decades. So far 45 CDGs have been described and most of these are found in the N-glycosylation pathway and in O-glycosylation pathways other than the mucin-type (Schachter and Freeze 2009; Jaeken 2010). Since the initiation steps of most types of protein glycosylation are catalyzed by only one or two glycosyltransferase genes, it is not surprising that defects in these genes have global effects on protein glycosylation and cause severe phenotypes. In contrast, the large number of GALNTs controlling the initiation step of mucin-type O-glycosylation may be expected to provide substantial biosynthetic back-up. Loss of a single GALNT gene may thus not produce discernable phenotypes or produce more discrete phenotypes in select organs. Early studies of GALNT-deficient mice confirmed the predicted redundancy, and several targeted GALNTs did not produce obvious phenotypes (Marth 1996; Lowe and Marth 2003). However, later investigations demonstrated that galnt1-deficient mice exhibit a bleeding disorder and have deficiency in B-cell maturation (Tenno et al. 2007). No overt phenotypes have been reported for mice deficient in galnt4 or t5 (Ten Hagen et al. 2002), deficient in both galnt4 and galnt5 (Mutant Mouse Regional Resource Centers stock number 029584-UCD) (Tabak 2010), deficient in galnt8 (Manzi et al. 2000), deficient in galnt10 (MMRC stock number 011647-UNC) or deficient in galnt14 (MMRC stock number 032320-UCD). Mice deficient in galnt13 exhibit decreased
expression of Tn carbohydrate in brain tissues, but no apparent phenotype (Zhang et al. 2003). A more dramatic phenotype was observed by chance in Drosophila in an early study demonstrating that the recessive lethal mutations in l(2)35Aa (CG7480/pgant35A) could be rescued with genomic DNA encoding l(2)35Aa (Flores and Engels 1999). l(2)35Aa was shown to encode a functional enzyme (CG7480) orthologous to human GalNAc-T11, and the catalytic function of the enzyme was shown to be required for development (Schwientek et al. 2002; Ten Hagen and Tran 2002; Ten Hagen et al. 2009; Bennett et al. 2010). These studies demonstrated for the first time that individual GalNAc-T isoforms serve unique and essential functions and that the existence of a large GALNT gene family with 14 members in fly does not provide complete genetic and functional redundancy. Subsequent detailed analysis of the molecular mechanism causing the l(2)35Aa phenotype revealed that CG7480 is necessary for processing of tracheal cell glycoproteins that affect epithelial morphogenesis, which is needed to form an intact diffusion barrier in the respiratory system (Tian and Ten Hagen 2007). More recently, deficiency in another Drosophila gene CG4445 (pgant3) was shown to cause a wing blistering phenotype (Zhang et al. 2010). Studies into the molecular mechanism behind this phenotype have determined that the extracellular matrix protein tiggrin is a specific substrate for the CG4445 enzyme and that the observed phenotype is caused by inappropriate secretion of extracellular matrix components, which alters cell adhesion events. Additionally, a recent study identified four Drosophila genes (CG31956, CG31651, CG6394 and CG30463) that are essential for viability, highlighting the essential developmental roles that are covered by O-glycosylation (Tran et al. 2011). A recent study of GalNAc-Ts in Xenopus has also proposed a specific function for the xgalnt11 isoform in neural and mesodermal tissue differentiation (GalNT16-predicted human ortholog. Herr et al. 2008). Although not clearly proven, this study suggests that xgalnt11 modulates the ActR-IIb receptor (TGF-β type II receptor family) at the molecular level, which implies that O-glycosylation may co-regulate TGF-β signaling in vertebrates. Xenopus left/right (LR) body patterning abnormalities were also observed in another Xenopus study demonstrating that xGalnt11 morpholino knockdown gave rise to abnormal heart and gut looping (Fakhro et al. 2011).

To date, only one human GALNT gene, GALNT3, has been shown to underlie disease in humans (Topaz et al. 2004). The molecular mechanism underlying the FTC and HHS phenotypes caused by deficiency in GALNT3 was shown to be a lack of O-glycosylation of a single Thr glycosylation site in a proprotein convertase (PC) processing site of the phosphaturic factor FGF23, which leads to excessive processing and inactivation of FGF23 (Kato et al. 2006). Importantly, later studies demonstrated that mutations in FGF23 preventing its secretion also lead to FTC. Thus, defects in either GALNT3 or FGF23 lead to insufficient circulating active FGF23 and the same overall clinical manifestations. GalNAc-T3 has broader substrate specificity than glycosylation of FGF23, but it appears that the main essential function of this isoform in humans is to co-regulate processing and blood levels of FGF23. A more recent study showed that GalNAc-T3 expression appears to be regulated by Ca²⁺ and vitamin D levels (Chefetz et al. 2009), indicating a direct regulatory role of GalNAc-T3 mediated O-glycosylation in phosphate homeostasis. The murine orthologous gene, galnt3, appears to serve similar functions and a knockout mouse exhibited reduced intact FGF23 blood levels (Ichikawa et al. 2009). The deficient mice did not develop the classical FTC features, but maleGalnt3−/− null mice showed growth retardation, infertility and increased bone mineral density. However, conflicting results were obtained in another preliminary study demonstrating that a galnt3 homozygous mutant mouse harboring a W589R lectin domain mutation phenotypically displays hyperphosphataemia and calcino-sis (Esapa et al. 2009; Duncan et al. 2011). Thus, the phenotypes of these mutant mice resemble the human disease caused by inactivating GALNT3 mutations. The example of GALNT3 serves to show that defects in this large gene family may result in subtle organ selective deficiencies in O-glycosylation and discrete phenotypes that may be difficult to identify. The example also shows that mucin-type O-glycosylation has highly specific regulatory functions in fundamental biological pathways such as PC processing, which exemplifies the cellular need for a large differentially regulated gene family regulating diverse-specific functions.

GWASs suggest that other GALNT genes may serve specific functions yet to be uncovered. Several studies have implicated GALNT2 as a candidate gene regulating plasma lipid levels and propose that its dysfunction links to cardiovascular disease (Kathiresan et al. 2008; Willer et al. 2008). We have recently identified a putative molecular mechanism for involvement of mouse GalNAc-T2 in lipid metabolism, which may be similar to the way GALNT3 regulates proprotein processing of FGF23. Thus, GalNAc-T2 appears to control O-glycosylation just adjacent to a PC processing site (RAPR224)TT in angiopoietin-like protein 3 (ANGPTL3) that activates this inhibitor of lipoprotein lipases (Schjoldager et al. 2010). A direct role for GALNT2 in lipid metabolism was shown in recent knockdown and overexpression studies of murine GalNAc-T2, where targeting in the liver affected plasma lipid levels (Teslovich et al. 2010).

Several other studies have linked GALNTs to disease or susceptibility to diseases including GALNT3 with bone mineral density and fracture risk (Duncan et al. 2011), GALNT4 with acute coronary disease (O’Halloran et al. 2009) and GALNT14 to resistance to death-receptor-mediated apoptosis (Wagner et al. 2007). Inactivating somatic and germline mutations have been found in GALNT5, T12, T15, T16 and T17 in cancer patients that potentially impairs function (Wood et al. 2007; Guda et al. 2009). A very recent report has linked rare GALNT11 exon deletions in congenital heart disease patients with heterotaxy caused by abnormalities in LR body pattern- ing (Fakhro et al. 2011). Besides GALNT11, this study identified four additional genes not previously implicated in LR patterning and all five genes are believed to act in the same pathway.

Perspectives

Research efforts during the last two decades have identified and characterized the GALNT gene family as the largest
glycosyltransferase gene family known catalyzing the formation of a single glycosidic linkage. Although progress has been made in understanding the functions of this large gene family, we believe that we have only begun to see the proverbial tip of the iceberg of the intricate biology of how site-specific protein O-glycosylation is regulated and the role of each of the GalNAc-T isoforms in biosynthesis of the O-glycoproteome. As discussed here, a number of new strategies are currently being applied to advance the field. We believe that these will eventually show that the GalNAc O-glycoproteome is vastly greater than currently understood and that individual GalNAc-Ts serve highly specific and dynamically regulated functions in producing the O-glycoproteome in health and disease. Site-specific O-glycosylation is likely to regulate many other biological processes similar to the abundant PC processing event. We therefore predict that the GALNT gene family underlies many diseases and susceptibilities to diseases, which seem to be supported by emerging GWAS and other linkage studies.

Supplementary data
Supplementary data for this article is available online at http://glycob.oxfordjournals.org/.

Funding
This work was supported by funds from the University of Copenhagen (Excellence program), Danish Research councils and the Carlsberg Foundation.

Acknowledgements
We would like to thank all present and prior members of our labs who have contributed to the work presented in this review. A special thank to Steven B. Levery and Maya Bonde Haaland for revision of the manuscript.

Note added in proof
After manuscript acceptance GalNAc-T18 was described by Li et al. (Glycobiology. 2011 Dec 14. Epub ahead of print). The nomenclature used by Li et al. for GalNAc-T18 follows the nomenclature suggested in this review.

Conflict of interest
None declared.

Abbreviations
CDG, congenital diseases of glycosylation; ER, endoplasmic reticulum; EST, expressed sequence tag; FTC, familial tumoral calcinosis; HHS, hyperostosis hyperphosphatemia syndrome; Gal, galactose; GalNAc, N-acetylgalactosamine; Glc, glucose; GlcNAc, N-acetylglucosamine; GalNAc-T, polypeptide GalNAc-transferase; GALNT, GalNAc-transferase gene; GWAS, genome-wide association study; Hyl, hydroxylsine; LR, left/right; MAb, monoclonal antibody; MYA, million years ago; ORF, open reading frame; PC, proprotein convertase; PCR, polymerase chain reaction; pgene, pseudogenes; RT, reverse transcriptase; Ser, serine; Thr, threonine; UTR, untranslated region; WGD, whole genome duplication; XYL T1/2, protein O-xylosyltransferase 1/2.

References


Muller S, Alving K, Peter-Katalinic J, Zachara N, Gooley AA, Hanisch FG. 1999. High density O-glycosylation on tandem repeat peptide from...