Family 6 Glycosyltransferases in Vertebrates and Bacteria: Inactivation and Horizontal Gene Transfer May Enhance Mutualism between Vertebrates and Bacteria

Glycosyltransferases (GTs) control the synthesis and structures of glycans. Inactivation and intense allelic variation in members of the GT6 family generate species-specific and individual variations in carbohydrate structures, including histo-blood group oligosaccharides, resulting in anti-glycan antibodies that target glycan-decorated pathogens. GT6 genes are ubiquitous in vertebrates but are otherwise rare, existing in a few bacteria, one protozoan, and cyanophages, suggesting lateral gene transfer. Prokaryotic GT6 genes correspond to one exon of vertebrate genes, yet their translated protein sequences are strikingly similar. Bacterial and phage GT6 genes influence the surface chemistry of bacteria, affecting their interactions, including those with vertebrate hosts.

The surfaces of cells are covered with carbohydrates, so glycans have key roles in the interactions between cells and of cells with the extracellular matrix, regulatory molecules, toxins, viruses, and antibodies (1). In the intestines of mammals, glycans of the outer membranes of Gram-negative bacteria contribute to the ability of the host to distinguish between commensal bacteria and pathogens, a key factor in intestinal homeostasis (2). Glycans can differ in structure between species and between individuals within a species; the histo-blood groups (HBGs) A, B, AB, and O (see Fig. 1A) are a well studied example of phenotypic variation between individuals that results in the display of complex glycans with different structures on cell membranes and extracellular glycoproteins (3, 4). Because glycans are antigenic, individuals produce antibodies directed against non-self-HBGs, with A or O individuals producing antibodies against the B antigen and B or O individuals against the A antigen. The presence of such antibodies indicates exposure to exogenous HBG-like antigens (5). HBG polymorphism appears to reflect multiple selective effects, including resistance to bacterial pathogens that bind to HBGs on host cells and interactions of the immune system with enveloped viruses that carry HBGs from a previous host (5).

Complex glycans are ubiquitous and have a greater potential for structural variation than polypeptides and nucleic acids because several different bonds can link pairs of adjacent monosaccharides. Nevertheless, glycans have received less attention than other macromolecules partly because they are heterogeneous secondary gene products with structures determined by the specificities of glycosyltransferases (GTs). Besides being antigenic, glycans have vital roles in biological systems encompassing protein folding and stability, cell adhesion, molecular trafficking and clearance, and signal transduction (6). Their biological and biomedical significance is reflected in the 1–2% of the open reading frames in all domains that encode enzymes involved in their synthesis and in the many diseases linked to aberrations in glycosylation (6, 7). The Carbohydrate-Active enZymes (CAZy) Database classifies GTs into 89 different families (92 listed but families 36, 46, and 86 are deleted) based on sequence interrelationships (8, 9). In this minireview, we discuss the properties and evolution of the GT6 family, which includes those responsible for the synthesis of HBGs. Genome databases indicate that the GT6 family has an irregular distribution, diagnostic of a nonlinear evolutionary history; GT6 enzymes are more remarkable for their non-functionality than activity in humans and some other primates. The products of the GT6 family are of particular biomedical interest because of their role in interactions of the immune systems of vertebrates with beneficial and pathogenic prokaryotes.

Mammalian GT6 Enzymes

The HBG locus on chromosome 9 has three alleles, A, B, and O. The A and B alleles encode active GTs (HBGTA and HBGTB) that catalyze the transfer of N-acetylgalactosamine and galactose, respectively, from their UDP derivatives into an α-linkage with the 3-OH group of a galactosyl residue on a glycan containing an H antigen (Fig. 1A). In the O allele, a frameshift mutation generates an inactive GT (10). HBGTA and HBGTB are closely similar in structure, differing in only 4 of 354 residues (10). Although the A, B, and O alleles predominate, giving rise to A, B, AB, and O phenotypes, >100 other variants have been described with mutations in the coding and noncoding regions, generating additional phenotypes (www.ncbi.nlm.nih.gov/projects/gv/mhc/xslcgi.cgi). The expression of HBGs is also modulated by fucosyl-α1,2-transfereases (FTs) that catalyze the synthesis of the H antigen. Although the FT in erythrocytes (H-type FT) is present in most individuals, secretor FT, which catalyzes H antigen production in saliva, gastrointestinal secretions, milk, and epithelial cells, is inactive in 20% or more of the population (11). Non-secretors (secretor FT-
negative individuals) have HBGs only on erythrocytes and their precursors. The Bombay phenotype, found in a minority of individuals, results from inactivity in both H-type and secretor FTs. Although all HBGs are lacking, in these individuals, they are generally healthy (12). The protective or other role of the ABO system is therefore subtle, but correlations have been found between various diseases, including pancreatic cancer, and blood groups (e.g. Refs. 3, 5, and 13).

Mammals have genes for three additional GT6 family members that catalyze the transfer of \(\alpha\)-galactose or GalNAc to the 3-OH group of a \(\alpha\)-linked galactose or GalNAc in an acceptor substrate (14). These genes encode a \(\beta\)-galactosyl-\(\alpha\),3-galactosyltransferase (\(\alpha\)3GT) that catalyzes the synthesis of the xenoantigen or \(\alpha\)-Gal epitope, a \(\beta\)-Gal-\(\alpha\),3-galactosyl entity (15), Forssman glycolipid synthase (FS) (16), and isogloboside 3 synthase (iGb3S) (17) (Fig. 1A). \(\alpha\)3GT and iGb3S both catalyze the formation of a similar structure but in glycoproteins and glycolipids, respectively. In humans, the enzymes expressed from the \(\alpha\)3GT, FS, and iGb3S genes are inactive as the result of frameshift and missense mutations (14–17). The lack of active \(\alpha\)3GT and iGb3S results in the absence of \(\alpha\)-Gal epitopes in human tissues and the presence of antibodies (~1–3% of the IgG) against the \(\alpha\)-Gal epitope in the circulations of all humans.

Our closest relatives among the primates, including chimps, gorillas, and orangutans, also produce inactive forms of these three GTs, but other mammals have active enzymes (14, 18). The inactivation of \(\alpha\)3GT and iGb3S is thought to have been an evolutionary adaptation that confers resistance to enveloped viruses derived from mammalian hosts that carry \(\alpha\)-Gal epitopes on their cell membranes. FS is also inactivated in humans and our closer relatives among the primates (19); this may be linked to the effects of Forssman antigen on vulnerability to toxins (16, 20). The inactivation of FS preceded the divergence of the chimps, orangutans, macaques, and rhesus monkeys, and it has been suggested that the catalytically inactive protein may have acquired a new function (19).

**Structural Basis of Specificity and Catalysis in Mammalian GT6 Enzymes**

Like most GTs from vertebrates, GT6 family members are type 2 membrane proteins with N-terminal cytosolic domains, a transmembrane helix, a stem, and a C-terminal catalytic domain (14). Their catalytic domains have a GT-A fold, one of the two predominant fold types (GT-A and GT-B) of the catalytic domains of structurally characterized GTs. Both folds encompass two Rossmann-like \(\alpha\)/\(\beta\)-domains; in the GT-A fold, these are closely associated through a large interaction interface, but in the GT-B fold, they are more distinct and are flexibly linked. Lairson et al. (21) have recently published a comprehensive review of the structures and mechanisms of GTs.

**FIGURE 1. Structure of the catalytic domain of mammalian GT6 and acceptor substrates and products.** A, chemical structures of the type 1 H antigen (i), type 1 HBG A (ii), type 1 HBG B (iii), \(\alpha\)-Gal epitope (xenoantigen) (iv), and Forssman glycolipid (v). Monosaccharides are color-coded as follows: Gal, red; GalNAc, purple; Fuc, green; GlcNAc, blue; and Glc, black. B, schematic representation of superposition of the \(\alpha\)3GT apo-form in cyan (Protein Data Bank code 2JCO) and \(\alpha\)3GT in complex with UDP-Gal in brick red (code 2VS5) showing the C-terminal (color-coded) flexible loop. The two invariant Arg and Glu residues are shown in the inset.
Crystallographic studies have provided extensive information about the structure and interactions of HBGTA and HBGTB (22–25) and θH92513GT (26–32) (Fig. 1B), revealing strong conservation of structure and structure-function relationships within the GT6 family. Like most GT-A fold GTs, both require a divalent metal ion (Mn\(^{2+}\)) for catalytic activity (33). The interactions of HBGTA/B and θH92513GT with donor substrates and metal cofactor are closely similar. Their N-terminal domains interact with the uracil, ribose, and diphosphate moieties of the UDP-sugar (donor) substrate. An essential Asp-Val-Asp (DVD) sequence motif (present as a D\(\times\)D motif in most GT-A fold GTs (21)) at the junction of the two domains (Fig. 2A, region C) interacts directly with the ribose moiety and indirectly, via the Mn\(^{2+}\), with the UDP phosphates. The binding of UDP or donor substrate increases order in two sections of the polypeptide chain, a region close to the active site and the C-terminal 10 residues of both enzymes (Fig. 1B) (25, 27, 28, 31). These changes organize the binding site for acceptor substrates; conformational changes in disordered loops adjacent to the active site are a shared feature of many GTs (see Ref. 25). Most of the interactions with the Gal or GalNAc of the donor substrate involve residues of the C-terminal domain, except for an invariant Arg of the N-terminal domain (Figs. 1B and 2A, region B) that interacts with the 3-OH group of the donor monosaccharide and may be important for galacto versus gluco specificity. Residues that function in acceptor substrate binding and catalysis are located in the C-terminal domain, including an invariant Glu within a highly conserved region of sequence (Figs. 1B and 2A, region F) and two basic residues close to the C terminus (Fig. 2B, region G).

GT6 enzymes catalyze a sequential bi-bi ordered mechanism in which the UDP-sugar binds, prior to acceptor, to an enzyme-Mn\(^{2+}\) complex (28). Inverting GTs are thought to utilize a single displacement \(S_2\)-like mechanism, but the mechanisms of retaining GTs are controversial (see Ref. 21). The expected mechanism is a double displacement reaction involving two inversions in which the monosaccharide is initially transferred to a nucleophile, such as a protein carboxyl group, and subsequently to the acceptor (21). Much evidence indicates that GT6 family members do not use this mechanism, but one that is analogous to that of the well characterized glyco-
cogen phosphorylases (34, 35), involving cleavage of the UDP–galactose bond, with all substrates present in the active site, generating a planar cationic oxycarbenium ion that is stabilized by anionic groups. A conformational change in the active site (21) may shift the location of this reactive species in the active site, possibly via Mn$^{2+}$ or basic residues in the C-terminal region that interact with UDP. The 3-OH group of the terminal galactose of the acceptor is deprotonated by UDP, and a bond is formed between this oxygen and the oxycarbenium moiety to generate the product.

A key region in donor substrate specificity is region E in Fig. 2A (HAAI in α3GT); the corresponding sequences in HBGTA and HBGTB, LGGF and MGAf, respectively, determine their respective specificities for UDP-GalNAc and UDP-Gal (22). The dominant role of this region in donor substrate specificity in α3GT was demonstrated by the identification of a mutant that is specific for UDP-GalNAc rather than UDP-Gal by screening a library of mutants with alternative sequences for residues 280–288. This mutant had the sequence AGGL replacing HAAI; structural studies suggested that the smaller side chains generate a pocket on the protein surface that can accommodate the 2-acetamido group of GalNAc (36).

Turcot-Dubois et al. (14) have investigated genes encoding different GT6 family members in non-mammalian vertebrates. These species have genes that may be orthologs of blood group enzymes and others that encode FS-like or iGb3S/α3GT-like GTs. Atypical members of the GT6 family in mammals were identified that have substitutions for highly conserved active-site residues, including the DXD motif (14). Two of these were expressed but did not display GT activity, so their functions are presently unknown (14). It is possible that they are pseudogenes or encode proteins that have lost catalytic activity, becoming binding proteins (lectins) like some members of glycoside hydrolase family 18 in vertebrates and invertebrates (37).

GT6 Family Members from Other Eukaryotes and Prokaryotes

With new information from genome sequences, genes were identified that encode members of the GT6 family in two non-vertebrate eukaryotes, Branchiostoma floridana (an amphioxus) and Capaspora owczarzaki (a unicellular opisthokont that is a symbiont of a snail) (38), plus 16 species of bacteria and the cyanophage P-SSM2 (supplemental Table 1); many other homologs are also in the human intestinal and marine metagenomes (39). The distribution pattern in eukaryotes is puzzling because the amphioxus is a chordate like the vertebrates, but C. owczarzaki represents an early branch from the line leading to metazoa and fungi. If the genes from eukaryotes, bacteria, and phages were related to vertebrate GT6 genes by the normal evolutionary process (vertical gene transfer), the GT6 family would have a broad distribution in eukaryotes and prokaryotes. However, GT6 genes are absent in currently sequenced genomes of a wide array of other invertebrates, plants, fungi, protozoa, archea, and most bacteria. Thus, outside of the chordates, the GT6 family has a sporadic distribution, suggesting the horizontal transfer of GT6 genes between eukaryotes and prokaryotes as well as between bacteria. It is interesting to compare the distribution of GT6 family members with that of another retaining GT family, GT8. These enzymes catalyze the formation of Gal-α1,4, Gal-α1,3 and Glc-α1,2 linkages and include GTs from plants, fungi, vertebrates, invertebrates, protozoa, bacteria, and viruses. The GT8 family encompasses glycogen and bacterial α1,4-galactosyltransferases related to LgtC (40). The GT8 family has a continuous distribution, consistent with evolution through vertical gene transfer.

Comparisons of the catalytic domain sequences of eukaryotic GT6 family members with the complete sequences of their prokaryotic counterparts show a high level of identity, up to 35% in some pairwise comparisons. Many, but not all, highly conserved sites are components of the active sites in the structurally characterized mammalian enzymes (Fig. 2A). The bacterial and phage GT6 enzymes lack the non-catalytic domains of the vertebrate proteins and are also truncated relative to the minimal functional mammalian catalytic domains by ~47 residues. In fact, with a single exception, the coding sequences of the prokaryotic GT6 enzymes correspond closely to the final exon in the genes for the vertebrate enzymes, consistent with horizontal transfer of this exon of the vertebrate gene (38). Phylogenetic trees constructed using GT6 protein sequences indicate that they form two clades, one containing the phage, eukaryotic, and Parachlamydia acanthamoebae GT6 enzymes and the second containing the other bacterial GT6 enzymes (Fig. 3).

The significance of the 47 extra residues in the vertebrate catalytic domains is unclear; an extension of similar length is present in GT6 from C. owczarzaki, but it is not significantly similar in sequence to the extension in the vertebrate catalytic domains. Therefore, although this part of the mammalian GT6 catalytic domains is well structured and tightly associated with the rest of the domain, it does not appear to be functionally important.

Properties and Structure-Function Relationships of Bacterial GT6 Enzymes

At present, little information is available on the properties of prokaryotic GT6 family members. Wang and co-workers (41) first expressed GT6 from Escherichia coli strain O86 and found that it is a galactosyltransferase that catalyzes the synthesis of HBG B antigens. This enzyme was shown to catalyze a step in the synthesis of the bacterial O antigen because gene disruption modified its structure (42). The same group cloned GT6 from Helicobacter mustelae and showed that it is a GalNac-transferase that can produce HBG A antigen structures (43). These enzymes were found to be useful for efficient enzymatic glycan synthesis and were applied for this purpose in the presence of Mn$^{2+}$ ions, as used with mammalian enzymes.

Comparison of the sequences of the bacterial and mammalian GT6 enzymes shows some striking differences (Fig. 2A). First, the DXD motif of the eukaryotic GTs is replaced in the bacterial proteins by NXX, with the single exception of GT6 from P. acanthamoebae; also cyanophage P-SSM2 GT6 retains the DXD motif. In the mammalian enzymes, the aspartates of the DXD motif are essential for activity (24, 33). The genome of Bacteroides ovatus, a Gram-negative commensal bacterium of the distal mammalian gut, encodes two GT6 family members; Tumable and Brew (39) expressed one of these (desig-
Based on the replacement of the Daccharized with regard to structure-function relationships. there are strong similarities in structure-function relationships suggesting that the bacterial enzymes may differ in catalytic mechanisms of basic and nonpolar amino acids; truncation of this region in BoA shows that it is not required for activity (39). These features are reminiscent of previous observations with the unrelated bacterial GT8 enzymes. The C-terminal regions of bacterial GT8 enzymes show little similarity to each other apart from having a high content of basic and aromatic amino acids. In two of these enzymes, LgtC and WaaJ, the C-terminal regions were found to have a role in interactions with the bacterial cell membrane, and they appear to form amphipathic α-helices that bind to the cell membrane (40, 44). Although the bacterial GT6 enzymes do not have a high content of aromatic residues, protein-membrane interactions can also be mediated by unstructured regions of polypeptide that typically have basic-hydrophobic-basic (BHB) amino acid sequence motifs (45). Brzeska et al. (46) have identified BHB lipid-binding regions in various proteins, including myosin I from Acanthamoebae and Dictyostelium; their web-based program (helixweb.nih.gov/bhsearch) uses the Wimley and White hydrophobicity scale (47) to scan protein sequences for BHB membrane-binding regions. Analysis of bacterial GT6 sequences indicates that their C-terminal regions contain such motifs, suggesting that bacterial GT6 enzymes may interact with the cytoplasmic face of the cell membrane through these regions. This is consistent with the role of GT6 enzymes from Gram-negative bacteria in catalyzing steps in the synthesis of the O antigen component of the lipopolysaccharide component of the outer membrane; this is synthesized on the cytosolic side of the cell membrane prior to assembly with the lipid A-core oligosaccharide intermediate (48).

### Functions of GT6 Family Members across the Evolutionary Spectrum

Bacterial glycans display greater diversity and have little similarity in chemical structure compared with eukaryotic glycans. Nevertheless, some bacteria have surface structures that resemble those from mammals and other vertebrates and are thought to facilitate their survival in eukaryotic hosts. The O antigens of Gram-negative bacteria are highly variable in structure, generating variations in the chemistry of the bacterial surface between different species and strains and within a single strain under different conditions (48). Changes in the complement of GTs and other enzymes are the ultimate source of these differences in structure and can arise from evolutionary changes in GT specificity, gene loss or inactivation, and the acquisition of new GT genes by horizontal gene transfer (HGT). The GT6 genes in bacteria appear to be an example of a GT gene of probable vertebrate origin participating in O antigen synthesis and enhancing molecular mimicry of host glycans. Based on current information on three bacterial GT6 enzymes, it appears
that they catalyze the synthesis of either HBG A or B-like glycans; current knowledge of structure-function relationships in the GT6 family suggests that the other bacterial GT6 enzymes have similar specificities. Consequently, their mimicry of host glycans and avoidance of immune responses will be confined to hosts that carry the corresponding HBG phenotype. Therefore, eluding immune responses from the host may not be the primary role of GT6 genes in commensal bacteria (5, 7). As discussed previously, the polymorphism of ABO HBGs in humans has been suggested to arise from selective pressure to protect the species against pathogens that utilize a particular HBG antigen as a receptor combined with protection, via antibodies against other HBGs, against enveloped viral pathogens that carry HBGs from another host (5). HBGs have other more indirect effects; for example, A or B antigens can modulate sialic acid-mediated interactions between pathogens, including malarial parasites and eukaryotic cell membranes (49). The selective advantage to a bacterium of carrying A or B antigens is less clear, but it may be linked to mutualism. For example, they may represent a major source of non-self-HBG antigens that promote the production of anti-HBG antibodies that protect their hosts from enveloped viruses and other pathogens. The recognition of commensal bacteria by their host is mutually beneficial, and perturbation of their cell surface chemistry can lead to inflammatory bowel disease (50); glycans such as HBG antigens appear to contribute to this recognition process (2). Also, some enteric pathogens, including noroviruses (51) and E. coli heat-labile and cholera-derived toxins (52–54), initially bind to HBGs on mammalian intestinal cells, and it may be advantageous for HBG-decorated intestinal bacteria to protect their hosts by acting as decoy receptors.

Cyanophage P-SSM2 GT6 is part of an array of P-SSM2 genes for enzymes involved in LPS synthesis that are absent in the genomes of two other cyanophages, P-SSM4 and P-SSM7. It has been suggested that in P-SSM2, they affect infection and establishment of the prophage stage by altering the surface of its host cyanobacterium to prevent other phages from binding (55). Many GT6 genes similar to that of P-SSM2 are found in the Environmental Samples Database, mostly in the marine metagenome. Their sequences group into two clades: one that is 60–84% identical in protein sequence to P-SSM2 GT6, whose members are probably from currently unidentified cyanophages, and one that consists of GT6 family members that are <40% identical to those from P-SSM2. Based on sequences in key substrate interaction regions, the members of each group appear to be similar in substrate specificity to each other, whereas the two groups may differ in both donor and acceptor substrate specificity; whether the second group is from phage and/or bacterial sources is unclear. Other homologs in the human intestinal metagenome have sequence characteristics of bacterial GT6 enzymes (39).

Members of the GT6 family fall into several evolutionary subgroups. One consists of the vertebrate enzymes with distinct multidomain structures, and a second consists of the bacterial enzymes, apart from P. acanthamoebae, which have BHB membrane-binding and NXN sequence motifs. Phylogenetic trees constructed using sequences from all available sources, including environmental databases, show a clear division into two principal clades with DXD and NNX sequence motifs, with the deepest branches in the DXD group leading to cyanophage P-SSM2, P. acanthamoebae, and C. owczarzaki. These analyses do not provide a robust answer regarding whether the ancestor of the GT6 family was from the metal-dependent or metal-independent GT6 groups. GT6 from P. acanthamoebae strain Hall’s coccus is unique among the bacterial enzymes in having a DXD motif. This bacterium normally infects free-living amoebae but can also infect human macrophages, pneumocytes, and lung fibroblasts (56) and has been linked to pneumonia and other diseases (57). It has a GT6 that is most similar to cyanophage P-SSM2 GT6, and its products may mediate interactions of this Parachlamydia strain with mammalian and other hosts.

GT6 from Sulfurospirillum deleyianum is, at present, an anomaly because it is from a species that does not inhabit the gastrointestinal tract or skin of vertebrates but is from a sulfur-reducing bacterium that is found in mud, lake material, and subsurface ground water (58).

The features of the GT6 family reviewed here indicate that HGT involving eukaryotes, bacteria, and phages occurred during their evolutionary development; although it seems likely that the direction is eukaryote to prokaryote, the evidence is not conclusive. Examples of HGT from eukaryotes to prokaryotes have been reported previously, including a fructose-bisphosphate aldolase gene between red algae and marine cyanophages (59). It is interesting that the deeper branches in sequence-based phylogenetic trees (Fig. 3) lead to endosymbions with potential for mediating HGT: P. acanthamoebae, associated with acanthamoebae and humans, and C. owczarzaki, a symbiont of the snail Biomphalaria glabrata, which is also an intermediate host of Schistosoma mansoni, the causative agent of schistosomiasis. Capsaspora is a predator of S. mansoni (38). The presence of many GT6 homologs from uncharacterized species in the marine metagenome indicates that there is more to be discovered about the distribution and evolution of the GT6 family.

Acknowledgments—We thank Professor Louise N. Johnson for constructive comments and Dr. Nethaji Thiyagarajan for generating Fig. 1B.

REFERENCES
1. Lowe, J. B., and Marth, J. D. (2003) Annu. Rev. Biochem. 72, 643–691
2. Munford, R. S. (2008) Infect. Immun. 76, 454–465
3. Hakomori, S. (1999) Biochim. Biophys. Acta 1473, 247–266
4. Marionneau, S., Cailleau-Thomas, A., Rocher, J., Le Moullac-Vaidye, B., Ruvinin, N., Clément, M., and Le Penu, J. (2001) Biochimie 83, 565–573
5. Seymour, R. M., Allan, M. J., Pomiankowski, A., and Gustafsson, K. (2004) Proc. R. Soc. Lond. B Biol. Sci. 271, 1065–1072
6. Ohtsubo, K., and Marth, J. D. (2006) Cell 126, 855–867
7. Varki, A. (2006) Cell 126, 841–845
8. Campbell, J. A., Davies, G. J., Bulone, V., and Henrissat, B. (1997) Biochem. J. 326, 929–939
9. Coutinho, P. M., Deleury, E., Davies, G. J., and Henrissat, B. (2003) J. Mol. Biol. 328, 307–317
10. Yamamoto, F., McNeill, P. D., and Hakomori, S. (1995) Glycoconjugate J. 12, 51–58
11. Henry, S., Oriol, R., and Samuelsson, B. (1995) Vox Sang. 69, 166–182
12. Kelly, R. J., Ernst, L. K., Larsen, R. D., Bryant, J. G., Robinson, J. S., and Lowe, J. B. (1994) Proc. Natl. Acad. Sci. U.S.A. 91, 5843–5847
