CHARACTERIZING THE (GLYCO)PEPTIDE SUBSTRATE SPECIFICITIES OF THE ppGalNAc T FAMILY OF GLYCOSYLTRANSFERASES

By

LESLIE REVOREDO

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thesis Advisor: Dr. Thomas A. Gerken

Department of Chemistry

CASE WESTERN RESERVE UNIVERSITY

August, 2016
We hereby approve the thesis/dissertation of

Leslie Revoredo

candidate for the degree of Ph.D.

Committee Chair
Irene Lee
Committee Member
Thomas A. Gerken
Committee Member
Michael Zagorski
Committee Member
Paul Carey
Committee Member
Michael E. Harris

Date of Defense
May 17, 2016

*We also certify that written approval has been obtained for any proprietary material contained therein.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>COMMITTEE SIGN OF SHEET</td>
<td>insert</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xvi</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>xxii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xxiii</td>
</tr>
<tr>
<td>CHAPTER 1: PROTEIN GLYCOSYLATION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 N-linked Protein Glycosylation</td>
<td>4</td>
</tr>
<tr>
<td>1.1.2 O-linked Protein Glycosylation</td>
<td>7</td>
</tr>
<tr>
<td>1.1.3 Non-Mucin Protein O-linked Glycosylation</td>
<td>9</td>
</tr>
<tr>
<td>1.2 MUCINS</td>
<td>11</td>
</tr>
<tr>
<td>1.2.1 Mucin-Type Protein O-Glycosylation</td>
<td>11</td>
</tr>
<tr>
<td>1.2.2 Mucin-Type O-Glycan Biosynthesis</td>
<td>14</td>
</tr>
<tr>
<td>1.2.3 Structure</td>
<td>17</td>
</tr>
<tr>
<td>1.2.4 Mucins in Health and Diseases</td>
<td>24</td>
</tr>
<tr>
<td>1.3 ppGalNAc T TRANSFERASES</td>
<td>26</td>
</tr>
<tr>
<td>1.3.1 History</td>
<td>26</td>
</tr>
<tr>
<td>1.3.2 Expression and Distribution</td>
<td>27</td>
</tr>
</tbody>
</table>
1.3.3 Role of ppGalNAc T’s in Cancer and Diseases ........................................30
1.3.4 ppGalNAc T Structure .............................................................................32
1.3.5 ppGalNAc T Peptide Substrate Specificity .............................................37
1.4 O-GLYCAN PREDICTION SOFTWARES ..................................................43
1.5 SUMMARY .................................................................................................45
1.6 POTENTIAL LIMITATIONS OF OUR WORK ..........................................46
1.7 REFERENCES .............................................................................................47

CHAPTER 2: CHARACTERIZATION OF THE PEPTIDE SUBSTRATE
SPECIFICITIES OF ppGalNAc T4, T7, T6, T11, T13, T14, T16, dT1 (PGANT35A),
PGANT9A AND PGANT9B USING RANDOM PEPTIDE SUBSTRATES .............67

2.1 BACKGROUND AND SIGNIFICANCE .......................................................68
2.2 MATERIALS AND METHODS .................................................................76
  2.2.1 Transferases .........................................................................................76
  2.2.2 Reagents and Peptide Substrates .........................................................78
  2.2.3 Instrumentation ...................................................................................79
  2.2.4 Glycosylation Assays With Recombinant ppGalNAc T’s and Random Peptide
      Substrates (PVI, PVII, PVIII) ..................................................................79
  2.2.5 Isolation of the Glycosylated Random Peptide Product .......................81
  2.2.6 Edman Amino Acid Sequencing .........................................................81
2.3 RESULTS AND DISCUSSION .................................................................82
  2.3.1 Full Characterization of the Catalytic Domain Peptide Substrate Preferences
       of ppGalNAc T11, T13, T14, T16, dT1, PGANT 9A and PGANT9B ..........82
     2.3.1.1 ppGalNAc T13 in Subfamily Ia .......................................................83
     2.3.1.2 ppGalNAc T14 and T16 in Subfamily Ib .........................................88
     2.3.1.3 ppGalNAc T11 and dT1 (PGANT35A) ............................................92
     2.3.1.4 PGANT9A and PGANT9B ............................................................97
  2.3.2 Partially Characterized Transferases ppGalNAc T4, T6 and T7 .............101
     2.3.2.1 Comparison of the Catalytic Domain Preferences of ppGalNAc T4 with
            Subfamily Iia Member T12 ..............................................................102
2.3.2.2 Comparison of the Catalytic Domain Preferences of ppGalNAc T6 with Subfamily Ic Member T3 ................................................................. 107
2.3.2.3 Comparison of the Catalytic Domain Preferences of ppGalNAc T7 with Subfamily IIb Member T10 ......................................................... 108

2.4 SUMMARY AND CONCLUSIONS .......................................................... 109
2.5 REFERENCES ..................................................................................... 119


3.1 BACKGROUND AND SIGNIFICANCE ................................................. 124
3.2 MATERIALS AND METHODS ............................................................. 127
  3.2.1 Transferases ................................................................................. 127
  3.2.2 Reagents and (Glyco)peptide Substrates ......................................... 128
  3.2.3 Instrumentation ........................................................................... 129
  3.2.4 Glycosylation Assays with Recombinant ppGalNAc T’s and Glycopeptide Substrate GP(T*10)C-Ser, Non-glycosylated Control GP(A10)C-Ser and GP(T*10)C-Thr ................................................................. 129
  3.2.5 Determination of Transferase Specific (Glyco)peptide Glycosylation Sites by Edman Sequencing ......................................................... 132

3.3 RESULTS AND DISCUSSION .............................................................. 133
  3.3.1 Design and Use of (Glyco)peptide Substrates ............................... 133
  3.3.2 Glycosylation of the GP(T*10)C-Ser and GP(T*10)C-Thr and GP(A10)C-Ser with ppGalNAc T2, T3, T4, T5, T7, T10, T11, T12, T13, T16, dT1 and PGANT7 ............................................................................................................. 135
    3.2.2.1 ppGalNAc T13 in Subfamily Ia .................................................. 136
    3.2.2.2 ppGalNAc T2 and T16 in Subfamily Ib ...................................... 140
    3.2.2.3 ppGalNAc T3 in Subfamily Ic ................................................... 142
    3.2.2.4 ppGalNAc T5 in Subfamily Id ................................................... 145
    3.2.2.5 ppGalNAc T11 and dT1 in Subfamily If ...................................... 147
3.2.2.6 ppGalNAc T4 and T12 in Subfamily IIa.................................152
3.2.2.7 ppGalNAc T7, T10 and PGANT7 in Subfamily IIb......................156

3.4 SUMMARY AND CONCLUSIONS..................................................162
3.5 REFERENCES.............................................................................168

CHAPTER 4: DETERMINATION OF THE GLYCOPEPTIDE PREFERENCES
MEDIATED BY THE LECTIN DOMAINS OF THE “PEPTIDE-PREFERRING” AND
“GLYCOPEPTIDE-PREFERRING” ppGalNAc T SUBFAMILY MEMBERS……171

4.1 BACKGROUND AND SIGNIFICANCE...........................................172

4.2 MATERIALS AND METHODS.....................................................176

4.2.1 Transferases........................................................................176
4.2.2 Reagents and (Glyco)peptide Substrates.....................................178
4.2.3 Instrumentation......................................................................178
4.2.4 Glycosylation Assays with Recombinant ppGalNAc T’s and (Glyco)peptide
Substrates GP(T*22)R, GP(T*10)L and Non-glycosylated Controls GP(A22)R and
GP(A10)L.................................................................................178
4.2.5 Determination of Optimal Reaction Conditions and Initial Observations.....181
4.2.6 Determination of the GalNAc Acceptor Sites of Incorporation by Edman
Sequencing.................................................................................183

4.3 RESULTS AND DISCUSSION.....................................................184

4.3.1 Design of the Lectin Domain Probing Random (Glyco)peptide Substrates...184
4.3.2 Glycosylation of the GP-L and GP-R Series of (Glyco)peptide with ppGalNAc
T2, T3, T4, T5, T6, T7, T10, T11, T12, T13, T13 Ex10B, T14, T16, dT1
(PGANT35A), PGANT7, PGANT9A and PGANT9B.................................188
4.3.2.1 ppGalNAc T1 and T13 in Subfamily Ia......................................188
4.3.2.2 ppGalNAc T2, T14 and T16 in Subfamily Ib..............................192
4.3.2.3 ppGalNAc T3 and T6 in Subfamily Ic........................................195
4.3.2.4 ppGalNAc T5 in Subfamily Id..................................................199
4.3.2.5 ppGalNAc T11 and dT1 in Subfamily If.....................................201
4.3.2.6 ppGalNAc T4 and T12 in Subfamily IIa…………………………………….204
4.3.2.7 ppGalNAc T7, T10 and PGANT7 in Subfamily IIb………………………….207
4.3.2.8 Non-glycosylated Controls GP(A22)R and GP(A10)L…………………….209
4.3.3 Lectin Domain Splice Variant Studies on ppGalNAc T13 & T13 Ex10B and
PGANT9A & PGANTB…………………………………………………………...212
  4.3.3.1 ppGalNAc T13 vs. T13 Ex10B………………………………………………212
  4.3.3.2 PGANT9A vs. PGANT9B…………………………………...217
4.5 SUMMARY AND CONCLUSIONS………………………………………………219
4.6 REFERENCES………………………………………………………………….230

CHAPTER 5: CHARACTERIZATION OF SOMATIC AND GERM-LINE
MUTATIONS ON GLYCOSYLTRANSFERASE GENES B3GNT2, B4GALT2,
ST6GALNACII AND GALNT12……………………………………………………...235
5.1 BACKGROUND AND SIGNIFICANCE………………………………………….236
5.2 MATERIALS AND METHODS………………………………………………...242
  5.2.1 Transferases………………………………………………………………….242
  5.2.2 Transferase Reaction Reagents and Substrates…………………………….244
  5.2.3 Instrumentation……………………………………………………………….245
  5.2.4 Glycosylation Assays with Recombinant Proteins………………………...245
    5.2.4.1 β3GnT2 (β-1,3-N-Acetylglucosaminyltransferase 2) Assays…………..247
    5.2.4.2 β4GalT2 (β-1,4-Galactosyltransferase 2) Assays………………………247
    5.2.4.3 ST6GalNAcII (α-N-Acetylgalactosaminidyl-α-2,6-Sialyltransferase 2)
                  Assays………………………………………………………………………….248
    5.2.4.4 ppGalNAc T12 Assays…………………………………………………249
5.3 RESULTS AND DISCUSSION………………………………………………250
  5.3.1 β3GnT2 (β-1,3-N-Acetylglucosaminyltransferase 2)…………………………250
5.3.2 β4GalT2 (β-1,4-Galactosyltransferase 2) ........................................255
5.3.3 ST6GalNAcII (α-N-Acetylgalactosaminidyl-α-2,6-Sialyltransferase 2) ....259
5.3.4 Phenotypic Characterization of WT and Mutant Glycosyltransferases β3GnT2, 
β4GalT2 and ST6GalNAcII .......................................................................261
5.3.5 ppGalNAc T12 (polypeptide N-acetylgalactosaminyl transferase 
12) ..................................................................................................................263
5.4 SUMMARY AND CONCLUSIONS .........................................................266
5.5 REFERENCES .................................................................................................270

CHAPTER 6: SUMMARY AND FUTURE DIRECTIONS ............................274
6.1 SUMMARY .................................................................................................275
6.2 FUTURE DIRECTIONS .................................................................................282
6.3 REFERENCES .................................................................................................288

APPENDIX A: NUMERICAL EVP VALUES FOR ppGalNAc T13, T14, T16, T11,  
dT1, T4, T6, T7, PGANT9A AND PGANT9B FOR EACH AMINO ACID AT THE 
FLANKING POSITIONS ..............................................................................290

APPENDIX B: FIRST AND SECOND AUTHOR PUBLICATIONS ..................295

  Gerken et al. JBC 2013 ..................................................................................296
  Revoredo et al. Glycobiology 2016 .................................................................312
  Venkitachalam et al. Scientific Reports 2016 ..............................................334
BIBLIOGRAPHY
LIST OF TABLES

CHAPTER 1

Table 1.1  Table of selected peptide and glycopeptide substrates used in in-vitro studies with recombinant ppGalNAc T’s .................................................................39

CHAPTER 2

Table 2.1  Random peptide substrates used in this work for the characterization of the catalytic domain specificity of the ppGalNAc T’s and fly orthologs ......................70

Table 2.2  Table of ppGalNAc T’s used in most of this work describing the source and expression system ........................................................................................................77

Table 2.3  Transferase specific catalytic domain preferences determined for subfamily members Ic, Ila and IIb derived from random peptide substrates PVI, PVII and PVIII ............................................................................................................106

Table 2.4  Correlation of Catalytic Domain Residues with the “Classic” T-P(G/A)P Substrate Motif for Selective ppGalNAc T’s and PGANT’s characterized by the Gerken Lab ..............................................................111

Table 2.5  Random peptide enhancement value products predict experimental transferase-specific sites of glycosylation ..............................................................117

CHAPTER 3

Table 3.1  Random peptide and glycopeptide substrates used to characterize catalytic domain glycopeptide specificity .............................................................130

CHAPTER 4

Table 4.1  ppGalNAc T lectin domain probing (glyco)peptide substrates .............179

Table 4.2  ppGalNAc T linker and lectin subdomain alignments ..........................225

CHAPTER 5

Table 5.1  Somatic mutations found in CRC patients ......................................241

Table 5.2  Table of glycosyltransferases used in this work describing the construct and expression ....................................................................................................243

Table 5.3  Table of substrates used and expected glycosylated products ...........246
## LIST OF FIGURES

### CHAPTER 1

| Figure 1.1 | Mammalian Protein Glycan Linkages | ................................................... | 3 |
| Figure 1.2 | N-linked Protein Glycosylation in the ER and Processing in the Golgi | ........................................ | 5 |
| Figure 1.3 | Examples of different types of protein Ser/Thr O-linked glycosylation | ........................................ | 8 |
| Figure 1.4 | Mucin-type O-glycan elongation pathway | .................................................. | 13 |
| Figure 1.5 | Schematic representation of the biosynthesis of mucin-type O-glycans | .................................. | 15 |
| Figure 1.6 | Examples of mucin domain structures of secreted MUC2 mucin | ................................. | 19 |
| Figure 1.7 | Examples of mucin domain structures of membrane-bound mucins | .......... | 20 |
| Figure 1.8 | Structural roles of mucin-type O-glycans. Secreted mucins (A-C) have stiffened and extended protease resistant structures and membrane-bound mucins (D-E) serve as scaffold and tethers for the interactions of ligands | ................................. | 22 |
| Figure 1.9 | Phylogenetic organization of the ppGalNAc T gene family | ................................. | 28 |
| Figure 1.10 | Crystal structure of ppGalNAc T2 (PDP 2FFU) | ........................................... | 33 |

### CHAPTER 2

| Figure 2.1 | Hydrophobic catalytic domain amino acid residue preferences for subfamily Ia member ppGalNAc T1 (A) subfamily Ib member T2 (B) subfamily Ic member T3 (C) subfamily Id member T5 (D) subfamily Ilb member T10 (E) and subfamily member Ila member T12 (F) obtained from random peptide substrates PVI, PVII and PVIII | .................................................. | 72 |
| Figure 2.2 | Hydrophilic catalytic domain amino acid residue preferences for subfamily Ia member ppGalNAc T1 (A) subfamily Ib member T2 (B) subfamily Ic member T3 (C) subfamily Id member T5 (D) subfamily Ilb member T10 (E) and subfamily member Ila member T12 (F) obtained from random peptide substrates PVI, PVII and PVIII | .................................................. | 73 |
| Figure 2.3 | Chromatography plots for ppGalNAc T13 using random peptides PVI, PVII and PVIII | .................................................. | 84 |
| Figure 2.4 | Catalytic domain amino acid residue preferences for subfamily Ia members ppGalNAc T13 (A-B) and T1 (C-D) obtained from random peptide substrates PVI, PVII and PVIII | .................................................. | 86 |
Figure 2.5 Catalytic domain hydrophobic amino acid residue preferences for subfamily Ib members ppGalNAc T14 (A), T16 (B) and T2 (C) obtained from random peptide substrates PVI, PVII and PVIII.  

Figure 2.6 Catalytic domain hydrophilic amino acid residue preferences for subfamily Ib members ppGalNAc T14 (A), T16 (B) and T2 (C) obtained from random peptide substrates PVI, PVII and PVIII.

Figure 2.7 Catalytic domain hydrophobic amino acid residue preferences for subfamily If members ppGalNAc T11 (A) and dT1 (C) and hydrophilic amino acid residue preferences for ppGalNAc T11 (B) and dT1 (D) obtained from random peptide substrates PVI, PVII and PVIII.  E) Comparison plot of the enhancement factors for all amino acids for ppGalNAc T11 plotted against the enhancement factors for all amino acids for dT1 (PGANT35A).

Figure 2.8 Multi-species ppGalNAc T phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Red arrows in subfamily If are pointing to the human T11 and dT1 (PGANT35A).  

Figure 2.9 Catalytic domain hydrophobic amino acid residue preferences for PGANT9A (A) and PGANT9B (C) and hydrophilic amino acid residue preferences for PGANT9A (B) and PGANT9B (D) obtained from random peptide substrates PVI, PVII and PVIII.  

Figure 2.10 Multi-species ppGalNAc T phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Red arrows in subfamily If are pointing to PGANT9 and human T15.

Figure 2.11 Partial catalytic domain hydrophobic amino acid residue preferences for (A) ppGalNAc T4 (subfamily IIa), (B) ppGalNAc T6 (subfamily Ic) and (C) ppGalNAc T7 (subfamily IIb) obtained from random peptide substrates PVI, PVII and PVIII.

Figure 2.12 Partial catalytic domain hydrophilic amino acid residue preferences for (A) ppGalNAc T4 (subfamily IIa), (B) ppGalNAc T6 (subfamily Ic) and (C) ppGalNAc T7 (subfamily IIb) obtained from random peptide substrates PVI, PVII and PVIII.

Figure 2.13 ISOGlyP homepage containing all of the parameters necessary for transferase-sequence specific EVP value determinations.

Figure 2.14 ISOGlyP output for ppGalNAc T1, T2, T3, T5, T10, T11, T12, T13, T14 and T15, along with their determined optimal sequences.
CHAPTER 3

Figure 3.1 Subfamily Ia ppGalNAcT13 catalytic domain glycopeptide preferences..........................................................................................................................................................137

Figure 3.2 Subfamily Ib, ppGalNAc T2 and T16 catalytic domain glycopeptide preferences........................................................................................................................................................................140

Figure 3.3 Subfamily Ic, ppGalNAc T13 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................144

Figure 3.4 Subfamily Id ppGalNAc T5 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................146

Figure 3.5 Subfamily Id ppGalNAc T11 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................148

Figure 3.6 Subfamily Id dT1 (PGANT35A) catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................149

Figure 3.7 Subfamily IIa ppGalNAc T4 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................153

Figure 3.8 Subfamily IIa ppGalNAc T12 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................154

Figure 3.9 Multi-species ppGalNAc T subfamily IIb phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Red arrows are pointing to the human T7 and fly PGANT7..........................................................................................................................................................................................................................158

Figure 3.10 Subfamily IIb ppGalNAc T7 and PGANT7 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................159

Figure 3.11 Subfamily IIb ppGalNAc T10 catalytic domain glycopeptide preferences..........................................................................................................................................................................................................................160

Figure 3.12 Efficiency of GalNAc transfer to acceptor glycopeptide varies with ppGalNAc T families..........................................................................................................................................................................................................................163

Figure 3.13 Correlation of glycopeptide specificities with the ppGalNAc T phylogenetic family tree..........................................................................................................................................................................................................................164

Figure 3.14 Schematic representations of catalytic domain-directed glycopeptide activities and peptide activities of the ppGalNAc T’s..........................................................................................................................................................................................................................165
CHAPTER 4

Figure 4.1  Glycosylation sites observed with versions of ppGalNAc T2. (A) Library of MUC5AC (glyco)peptides used…………………………………………………………………………174

Figure 4.2  Substrate concentration dependence for ppGalNAc T1 and T2. (Glyco)peptides GP(T*22)R (red diamonds), GP(A22)R (purple triangles), GP(T*10)L (blue squares) and GP(A10)L (light blue downward triangles)………………………………182

Figure 4.3  Lectin domain substrate design based in the X-ray crystal structure of ppGalNAc T2 with EA2 peptide bound in the catalytic domain……………………………………185

Figure 4.4  Effects on adding free GalNAc on ppGalNAc T2 (glyco)peptide utilization. (A) GP(T*22)R (red diamonds) and (B) GP(A22)R (purple triangles), in the presence of 100 mM GalNAc……………………………………………………………………………………………………..187

Figure 4.5  Subfamily Ia ppGalNAc T1 and T13 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T1 and T13 showing the net [3H]-GalNAc utilization……………………………………190

Figure 4.6  Subfamily Ib ppGalNAc T2, T14 and T16 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T2, T14 and T16 showing the net [3H]-GalNAc utilization………………………………………………………………………………193

Figure 4.7  Subfamily Ic ppGalNAc T3 and T6 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T3 and T6 showing the net [3H]-GalNAc utilization………………………………………………………………………………196

Figure 4.8  Subfamily Id ppGalNAc T5 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T5 showing the net [3H]-GalNAc utilization…………………………………………………………………………………200

Figure 4.9  Subfamily If ppGalNAc T11 and dT1 (PGANT35A) lectin domain preferences…………………………………………………………………………………………………………………………203

Figure 4.10  Subfamily IIa ppGalNAc T4 and T12 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L with ppGalNAc T4 and T12 showing the net [3H]-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating [3H]-GalNAc transfer to glycopeptide (fraction 25-31) and high hydrolysis (35-41) for ppGalNAc T4 and T12. (C-D) [3H]-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with
ppGalNAc T4 and T12 showing the optimal distance from a T* to the X acceptor residues is approximately 8-9 residues…………………………………………………………205

Figure 4.11  Subfamily IIb ppGalNAc T7, T10 and PGANT7 lectin domain preferences………………………………………………………………………………………………………208

Figure 4.12  Plots of [3H]-GalNAc incorporation into the Xaa residues of random glycopeptides GP(T*22)R and GP(T*10)L (open bars) and after normalization by subtraction of control peptides GP(A22)R and GP(A10)L that were glycosylated by ppGalNAc T13 (filled bars)………………………………………………………………………………211

Figure 4.13  Splice variants ppGalNAc T13 and T13 Ex10B lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L with ppGalNAc T13 and T13 Ex10B showing the net [3H]-GalNAc utilization……………………………………………………………………………………………………215

Figure 4.14  Splice variants PGANT9A and 9B lectin domain preferences………218

Figure 4.15  Efficiency of GalNAc transfer to acceptor glycopeptides varies with ppGalNAc T family……………………………………………………………………………220

Figure 4.16  Correlation of glycopeptide specificities with the ppGalNAc T phylogenetic family tree…………………………………………………………………………221

Figure 4.17  Cartoon and X-ray crystal modeling for the ppGalNAc T lectin domain-assisted property…………………………………………………………………………222

CHAPTER 5

Figure 5.1  Protein glycosylation pathways involving ppGalNAc T12, β3GnT2, β4GalT4 and ST6GalNAcII…………………………………………………………………………238

Figure 5.2  Identification of B3GNT2 mutants and localization of R6X mutant……………………………………………………………………………………………………252

Figure 5.3  In vitro glycosylation assays with P186T and D47H mutants. (A) Enzymatic activity of WT β3GnT2 against LacNAc-PNP and Lactose-PNP at ON incubations……………………………………………………………………………………………………254

Figure 5.4  Glycosylation assays with β4GalT2…………………………………………………………256

Figure 5.5  Western blot of glycosidase treatments on WT and mutant β4GalT2……………………………………………………………………………………………………258
Figure 5.6  Glycosylation assays with ST6GalNAcII. (A) Reaction catalyzed by ST6GalNAcII…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………...
ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Thomas A. Gerken for his support and guidance throughout the years. I would also like to thank him for believing in me and letting me find my own ways to approach and solve problems. This made me work harder and ultimately helped shape me into the professional I am today.

I would also like to thank my family and close friends for the endless support and encouragement, especially in difficult times not only in my graduate career but also in life. Through their love and patience, I was able to overcome many setbacks and keep myself focused.
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFGP</td>
<td>Antarctic fish Anti-Freeze Glycoprotein</td>
</tr>
<tr>
<td>AFM</td>
<td>Atomic Force Microscopy</td>
</tr>
<tr>
<td>AMOP Domain</td>
<td>Adhesion Associated Domain</td>
</tr>
<tr>
<td>ANGPT3</td>
<td>Angiopoietin-like Protein 3</td>
</tr>
<tr>
<td>ApoC-III</td>
<td>Apolipoprotein C-III</td>
</tr>
<tr>
<td>ASF</td>
<td>Asialofetuin</td>
</tr>
<tr>
<td>β3GNT2</td>
<td>β-1,3-N-Acetlyglucosaminyltransferase 2 Gene</td>
</tr>
<tr>
<td>β3GnT2</td>
<td>β-1,3-N-Acetylglucosaminyltransferase 2</td>
</tr>
<tr>
<td>β3GnT6</td>
<td>β-1,3-N-acetylglucosaminyltransferase 6 (Core 3 synthase)</td>
</tr>
<tr>
<td>β4GALT2</td>
<td>β-1,4-Galactosyltransferase 2 Gene</td>
</tr>
<tr>
<td>β4GalT2</td>
<td>β-1,4-Galactosyltransferase 2</td>
</tr>
<tr>
<td>BIS-TRIS</td>
<td>2,2-Bis(hydroxymethyl)-2,2′,2″-nitrilotriethanol</td>
</tr>
<tr>
<td>BW5147</td>
<td>Mouse Lymphoma T-cell Line</td>
</tr>
<tr>
<td>CA125</td>
<td>Carcinoma Antigen 125 (i.e. MUC16)</td>
</tr>
<tr>
<td>CaCl₂</td>
<td>Calcium Chloride</td>
</tr>
<tr>
<td>C1GalT</td>
<td>β-1,3- galactosyltransferase (Core 1 synthase)</td>
</tr>
<tr>
<td>C2GnT</td>
<td>β-1,6-N-acetylglucosaminyltransferase</td>
</tr>
<tr>
<td>CDG</td>
<td>Congenital Disorders of Glycosylation</td>
</tr>
<tr>
<td>CG30463</td>
<td>PGANT9 (<em>Drosophila melanogaster</em> ppGalNAc T)</td>
</tr>
<tr>
<td>CHO</td>
<td>Chinese Hamster Ovary Cells</td>
</tr>
<tr>
<td>CMD</td>
<td>Congenital Muscular Dystrophy</td>
</tr>
<tr>
<td>CMP</td>
<td>Cytidine Monophosphate</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic Obstructive Pulmonary Disease</td>
</tr>
<tr>
<td>COS7</td>
<td>Carcinoma Cell Line</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal Cancer</td>
</tr>
<tr>
<td>CSM</td>
<td>Canine Submaxillary Mucin</td>
</tr>
<tr>
<td>Dol-P-Man</td>
<td>Dolichol Monophosphate Mannose</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl Sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DPM</td>
<td>Disintegrations Per Minute</td>
</tr>
<tr>
<td>DR5</td>
<td>Death Receptor 5 (TRAIL Receptor 2)</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic Acid</td>
</tr>
<tr>
<td>EGF Domain</td>
<td>Epidermal growth factor-like</td>
</tr>
<tr>
<td>EF</td>
<td>Enhancement Factor</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic Reticulum</td>
</tr>
<tr>
<td>ERBB2</td>
<td>Erb-B2 Receptor Tyrosine Kinase 2</td>
</tr>
<tr>
<td>EV</td>
<td>Enhancement Value</td>
</tr>
<tr>
<td>EVP</td>
<td>Enhancement Value Product</td>
</tr>
<tr>
<td>FGF23</td>
<td>Fibroblast Growth Factor Receptor 23</td>
</tr>
<tr>
<td>FGFR3</td>
<td>Fibroblast Growth Factor Receptor 3</td>
</tr>
<tr>
<td>Fuc</td>
<td>Fucose</td>
</tr>
<tr>
<td>GDP</td>
<td>Guanosine Diphosphate</td>
</tr>
<tr>
<td>Gal</td>
<td>Galactose</td>
</tr>
<tr>
<td>GalNAc</td>
<td>N-acetylgalactosamine</td>
</tr>
<tr>
<td>GalNAz</td>
<td>N-azidoacetylgalactosamine Tetraacylated</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>GALNT</td>
<td>ppGalNAc T gene</td>
</tr>
<tr>
<td>Glc</td>
<td>Glucose</td>
</tr>
<tr>
<td>GlcNAc</td>
<td>N-acetylglucosamine</td>
</tr>
<tr>
<td>GlcNAc T1</td>
<td>Mannosyl (α-1,3-)Glycoprotein β-1,2-N-Acetylglucosaminyltransferase</td>
</tr>
<tr>
<td>GLT25D1</td>
<td>Collagen β(1-O)Galactosyltransferase 1</td>
</tr>
<tr>
<td>GLT25D2</td>
<td>Collagen β(1-O)Galactosyltransferase 1</td>
</tr>
<tr>
<td>GPI</td>
<td>Glycosylphosphatidylinositol</td>
</tr>
<tr>
<td>GT-A</td>
<td>Glycosyltransferase Fold Type A</td>
</tr>
<tr>
<td>HCl</td>
<td>Hydrochloric Acid</td>
</tr>
<tr>
<td>HDL</td>
<td>High-Density Lipoprotein</td>
</tr>
<tr>
<td>HDPE</td>
<td>High-Density Polyethylene</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>HEK293f</td>
<td>Human Embryonic Kidney 293 Cells</td>
</tr>
<tr>
<td>hEPO</td>
<td>Human Erythroprotein</td>
</tr>
<tr>
<td>HIV gp120</td>
<td>Human Immunodeficiency Virus Envelope Glycoprotein 120</td>
</tr>
<tr>
<td>HPA</td>
<td>Helix Pomatia (Lectin Beads)</td>
</tr>
<tr>
<td>ISOGlyP</td>
<td>Isoform Specific O-Glycosylation Prediction Software</td>
</tr>
<tr>
<td>LAcNAc</td>
<td>N-acetyl-D-lactosamine</td>
</tr>
<tr>
<td>LDL Receptor</td>
<td>Low-Density Lipoprotein Receptor</td>
</tr>
<tr>
<td>Man</td>
<td>Mannose</td>
</tr>
<tr>
<td>MCF-7</td>
<td>Breast Cancer Cell Line</td>
</tr>
<tr>
<td>MES</td>
<td>2-(N-morpholino)ethanesulfonic Acid</td>
</tr>
<tr>
<td>MonoQ 5/50</td>
<td>Monodisperse Porous Polystyrene/Divinyl Benzene Beads</td>
</tr>
<tr>
<td>Symbol</td>
<td>Name</td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>MnCl₂</td>
<td>Manganese Chloride</td>
</tr>
<tr>
<td>MUC</td>
<td>Mucin Protein</td>
</tr>
<tr>
<td>NaCl</td>
<td>Sodium Chloride</td>
</tr>
<tr>
<td>NaN₃</td>
<td>Sodium Azide</td>
</tr>
<tr>
<td>NeuG1</td>
<td>N-glycolylneuraminic Acid</td>
</tr>
<tr>
<td>Neu5Ac</td>
<td>N-acetyleneuraminic Acid</td>
</tr>
<tr>
<td>NIDO Domain</td>
<td>Nidogen-like Domain</td>
</tr>
<tr>
<td>Ni-NTA</td>
<td>Nickel Nitrilotriacetic Acid</td>
</tr>
<tr>
<td>OGA</td>
<td>O-β-N-acetylglicosaminidase</td>
</tr>
<tr>
<td>OGT</td>
<td>β-N-acetylglicosaminyl Transferase</td>
</tr>
<tr>
<td>OPT12</td>
<td>Transferase-Specific Optimal Substrate 12</td>
</tr>
<tr>
<td>OST</td>
<td>Oligosaccharyltransferase Enzyme Complex</td>
</tr>
<tr>
<td>PGANT</td>
<td>Drosophila GalNAc Transferase</td>
</tr>
<tr>
<td>PNGaseF</td>
<td>Peptide -N-Glycosidase F</td>
</tr>
<tr>
<td>PNP</td>
<td>p-Nitrophenol</td>
</tr>
<tr>
<td>POFUT1</td>
<td>Protein O-Fucosyltransferase 1</td>
</tr>
<tr>
<td>POFUT2</td>
<td>Protein O-Fucosyltransferase 2</td>
</tr>
<tr>
<td>POGLUT1</td>
<td>Protein O-Glucosyltransferase 1</td>
</tr>
<tr>
<td>polyHis</td>
<td>Polyhistidine Tag</td>
</tr>
<tr>
<td>POMT1</td>
<td>Protein O-mannosyltransferase 1</td>
</tr>
<tr>
<td>POMT2</td>
<td>Protein O-mannosyltransferase 2</td>
</tr>
<tr>
<td>POMGnT1</td>
<td>Protein O-mannose β-1,2-N-acetylglicosaminyltransferase</td>
</tr>
<tr>
<td>ppGalNAc T</td>
<td>Polypeptide N-α-acetylgalactosaminyl Transferases</td>
</tr>
</tbody>
</table>
proBNP   Pro-brain Natriuretic Peptide
PSM     Porcine Submaxillary Mucin
PTH     Phenythiohydantoin
PVDF    Polyvinylidene Fluoride
R^2     Coefficient of Determination
S*      GalNAc-O-Ser
SBA     Glycine Max (Lectin Beads)
SDS     Sodium Dodecyl Sulfate
SDS-PAGE Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SEA Domain Sea Urchin Sperm Protein Enterokinase Domain
Sf9     Baculovirus Cell System
SJA     Sophora Japonica (Lectin Beads)
SP-Sepharose Sulphopropyl Spharose
Src     Proto-oncogene Tyrosine-Protein Kinase
ST3GalI β-galactoside α-2,3-sialyltransferase 1
ST6GalI β-Galactosamidase-α-2,6 sialyltransferase 1
ST6GalNAcI α-N-Acetylgalactosaminidyl-α-2,6-sialyltransferase 1
ST6GALNACII α-N-Acetylgalactosaminidyl-α-2,6-sialyltransferase 2 Gene
ST6GalNAcII α-N-Acetylgalactosaminidyl-α-2,6-sialyltransferase 2
ST-antigen Neu5Acα2-3Galβ1-3GalNAc-Ser/Thr
STn-antigen Neu5Acα2-6GalNAcα-Ser/Thr
SV40 promoter Simian Vacuolating Virus 40
SW480    Human Colon Cancer Cell Line
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>T*</td>
<td>GalNAc-O-Thr</td>
</tr>
<tr>
<td>T-antigen</td>
<td>Galβ1-3GalNAc-O-Ser/Thr</td>
</tr>
<tr>
<td>T-synthase</td>
<td>Core 1 synthase, see C1GalT</td>
</tr>
<tr>
<td>TGF-β R II</td>
<td>Transforming Growth Factor β Receptor II</td>
</tr>
<tr>
<td>Tn-antigen</td>
<td>GalNAc-O-Ser/Thr</td>
</tr>
<tr>
<td>TPR</td>
<td>Tetratricopeptide Repeat Domain</td>
</tr>
<tr>
<td>UC</td>
<td>Ulcerative Colitis</td>
</tr>
<tr>
<td>UDP</td>
<td>Uridine Diphosphate</td>
</tr>
<tr>
<td>UV-Vis</td>
<td>Ultraviolet Visible Spectroscopy</td>
</tr>
<tr>
<td>VNTR</td>
<td>Variable Number of Tandem Repeat</td>
</tr>
<tr>
<td>vWF</td>
<td>von Willebrand Factor</td>
</tr>
<tr>
<td>VVA</td>
<td><em>Vicia Villosa</em> (Lectin Beads)</td>
</tr>
<tr>
<td>WT</td>
<td>Wild-Type Enzyme</td>
</tr>
<tr>
<td>WWS</td>
<td>Walker-Warburg Syndrome</td>
</tr>
<tr>
<td>XT-I</td>
<td>Xylosyltransferase I</td>
</tr>
<tr>
<td>XT-II</td>
<td>Xylosyltransferase II</td>
</tr>
</tbody>
</table>
Characterization of the (Glyco)peptide Substrate Specificities of the ppGalNAc T Family of Glycosyltransferases

Abstract

by

LESLIE REVOREDO

Many proteins of eukaryotic cells are known to be O-glycosylated. Glycoproteins with heavily O-glycosylated mucin domains provide important biological functions in a cell: i.e., protection from pathogens, cell-to-cell adhesion and intracellular protein trafficking. Mucin-type O-glycosylation occurs in the Golgi complex and begins with the transfer of GalNAc from UDP-GalNAc onto Ser/Thr residues of polypeptides. This step is catalyzed by a large family (20) called N-α-acetylgalactosaminyl transferases (ppGalNAc-T’s) and forms the GalNAc-α-O-Ser/Thr product. Subsequent elongation is performed by specific glycosyltransferases, producing a variety of glycans. Family members have been classified into peptide- and glycopeptide-prefering subfamilies, although both subfamilies possess variable activities against glycopeptide substrates. Structurally, 19 isoforms contain a C-terminal catalytic domain linked via a flexible linker to an N-terminal ricin-like lectin domain. The (glyco)peptide substrate specificities of the ppGalNAc-T transferases and the roles of the catalytic and lectin domains in glycopeptide glycosylation still remain largely unknown. Based on the systematic
random peptide approach created by the Gerken Lab, I have determined the glycopeptide substrate specificities of several ppGalNAc-T isoforms. A series of (glyco)peptides were created in order to specifically probe the functions of the catalytic and lectin domains in terms of neighboring (1-5 residues) and remote prior glycosylation (6-17 residues) from an acceptor site, respectfully. Using several glycopeptide-preferring isoforms, glycosylation was observed from -4, -3, -1 and +1 relative to a neighboring GalNAc-O-Thr, which I attributed to specific GalNAc-O-Thr binding at the catalytic domain. The other series of glycopeptides contained a GalNAc-O-Thr near the C- or N- terminus of the substrate to address the directionality preferences of the lectin domain. Results with several peptide- and glycopeptide-preferring isoforms revealed preferences that varied among transferase isoform, where some preferred a C-terminally placed GalNAc-O-Thr, or a N-terminally placed GalNAc-O-Thr and others equally preferred the C-/N- terminally placed GalNAc-O-Thr. These directionality preferences are due to the GalNAc-O-Thr interactions at the lectin domain. Results of these studies revealed for the first time the site-specific glycopeptide glycosylation preferences by some ppGalNAc-T’s and has demonstrated that both domains of the ppGalNAc-T’s have specialized and unique functions that work in concert to control and order mucin-type O-glycosylation.
CHAPTER 1

PROTEIN GLYCOSYLATION
1.1 Introduction

Most proteins that are expressed by eukaryotic cells are glycosylated in one way or another. Protein glycosylation occurs in the ER and Golgi producing glycoproteins with glycans containing monosaccharide(s), disaccharides, oligosaccharide chains, etc. The sugar moieties are attached to polypeptides and other glycans via specific linkages by enzymes called glycosyltransferases. Glycoproteins are essential for life as they are involved in important biological functions and have important physiological roles in cells and tissues (1). Protein glycosylation is a complex and multifaceted post-translational modification, where several variables can be modified including the type of sugar-protein linkage, the sugars that are transferred and the resulting glycan structure and length. There are five different classes of glycans that are produced, see Figure 1.1. These classifications are based on their peptide-sugar linkage and are dependent on the amino acid functional group to which the sugar is linked, the most common being and \( N \)- (nitrogen atom) and \( O \)- (oxygen atom) linkages. The latter being further defined by the sugar moiety being attached (see Figure 1.1). A less common glycosidic linkage is through a \( C \)- (carbon atom) where a sugar moiety is added to a carbon on the tryptophan side-chain (1,2). The other two classes of glycan linkages are glypiation and phosphoglycosylation, which are special kinds of sugar to protein bonds that arise from the linkage of a phospholipid (usually GPI anchors) to a protein via a sugar core and the attachment of a sugar to a protein via a phosphodiester bond, respectively (3). Since protein glycosylation and elongation is non-templated (unlike DNA) these variables are highly dependent on the expression levels of the glycosyltransferase(s) in the ER and Golgi, and because the presence of a
Figure 1.1- Mammalian Protein Glycan Linkages. The first three are glycosidic bonds with N-, O- and C- linkages. The last two are glycan linkages that are linked to proteins via a sugar or phosphodiester bonds.

* Note that O-linked glycosylation can also occur with other amino acids with the –OH functional group such as tyrosine, hydroxyproline (Hyp) and hydroxylserine (Hyl). These are less common and not shown here.
particular glycosyltransferase can vary from cell-to-cell and cell type (4), this can create a wide variety of heterogeneous glycans and glycoproteins. This chapter will focus on N- and O-linked glycosylation.

1.1.1 N-linked Protein Glycosylation

In N-linked glycosylation, a pre-assembled common structure, Glc$_3$Man$_9$GlcNAc$_2$ (containing two glucose, nine mannose, and two N-acetylglucosamine sugar moieties) is transferred to the amide group of asparagine residues on protein substrates via the action of the oligosaccharyltransferase enzyme complex (OST) (5,6) (Figure 1.2). The site of attachment on a protein occurs at the consensus sequence, Asn-X-(Ser/Thr) where the X can be any amino acid except for Pro. This process occurs co-translationally; while the protein itself is being synthesized and passing through the lumen of the ER. Once the transfer occurs, the resulting glycan is further processed where the sugar moieties are either trimmed or added, in different compartments within the ER and Golgi (7). This process is highly dependent on the concentration and location of the enzymes within the different compartments (8), highlighting the complexity of N-linked glycosylation. While all N-linked glycans have the same core structure, there are three main types of N-linked glycans that can result from the different core terminal elaborations. These include: i) oligomannose, where only mannose moieties are attached to the core, ii) complex, where the core consists of three mannose sugar moieties and two N-acetylglucosamine moieties, or iii) hybrids, where one arm attached to the core contains only mannose and other the arm attached to the core contains other sugar attachments (often times a fucose is attached to N-acetylglucosamine that is directly attached to the Asn of the protein) (9).
Figure 1.2 - N-linked Protein Glycosylation in the ER and Processing in the Golgi. The pre-assembled structure, Glc\textsubscript{3}Man\textsubscript{9}GlcNAc\textsubscript{2}, is initially attached to a dolichylphosphate (a hydrophobic lipid carrier molecule) and is transferred to an asparagine residue of the nascent polypeptide by the oligosaccharyltransferase (OST) complex. Further processing trims or adds sugars in the ER and Golgi. Examples of oligomannose and hybrid N-linked glycans are marked as 1 and 2 respectively.
N-linked glycosylation occurs in archa and eukaryotes with variations in the synthesis pathway occurring in bacteria and yeast (6,10). Defects in N-linked glycans, usually caused by mutation(s) in the enzymes responsible for the formation of either the core precursor and the core terminal structure in the later processing phase, have multiple effects both in the mouse and in man. Defects in GlcNAc-T1 (Mgat-1), an essential glycosyltransferase in forming the complex and hybrid N-glycans, causes embryonic lethality in the mouse (11,12). While in man, N-glycans play a critical role in monitoring correct protein folding of newly synthesized proteins in the ER, determining whether a protein will be degraded or allowed to enter the Golgi and ultimately allowing the protein to reach a functional state (13). Studies have shown that a complete loss of N-glycosylation is lethal in yeast and mammals (14), while substantial deficiency in the pathway steps (primarily from enzyme defects) cause many human congenital disorders of glycosylation (known as CDG) (15,16). CDG is the largest known disorder that affects N-linked glycosylation, with 19 types already identified. Clinical problems include mental retardation, liver dysfunction and intestinal disorders (17). Other disorders that arise from defects in the glycan core and terminal structures are mucolipidosis II (I-cell disease), pseudo-Hurler polydystrophy, type II congenital dyserythropoietic anemia, and galactosemia (18-21).

The identification, characterization and analysis of N-linked glycans is relatively simple compared to mucin-type O-linked glycans on proteins. The main advantage is the strict Asn-X-(Ser/Thr) motif, which makes it easier to identify and mutate N-glycosylation sites. Furthermore, identifying N-linked glycans in cellular and tissue samples containing glycoproteins is relatively easy as N-glycan release can be performed
under mild conditions using an enzyme called PNGaseF that practically cleaves nearly all N-linked glycans (22). The released N-linked glycans can then be isolated via a combination of several approaches including permethylation analysis, exoglycosidase digestion and mass spectrometry. Mass spectrometry is the most widely used method to characterize the N-linked glycan structures (22-24).

1.1.2 O-linked Protein Glycosylation

Protein O-linked glycosylation is defined as the attachment of a sugar moiety onto the hydroxyl group of a polypeptide Ser or Thr residue. Currently, there are seven different types of Ser/Thr O-linked glycans, due to the different sugars that can be attached to the protein core. The most common is a mucin-type O-glycan where an N-acetylgalactosamine is transferred to the protein, forming an O-GalNAc linkage. All of the others are known as non-mucin-type and these include O-GlcNAc (N-acetylgalactosamine), O-Mannose, O-Glucose, O-Galactose, O-Fucose and O-Xylose (see Figure 1.3). Proteoglycans are another class of heavily O-linked glycans but will not be discussed in my thesis. Most of these modifications mainly occur in the ER and Golgi, while others have been shown to occur in the cytoplasm of the cell and some of these modifications occur in both prokaryotes and eukaryotes (25). The most common mammalian modifications include the mucin-type O-glycans and non-mucin O-GlcNAc (N-acetylgalactosamine) and O-Mannose modifications. Each of these modifications encompasses ranges of proteins that have distinct biological roles. These will be briefly discussed in the sections below.

1 O-linked glycosylation also occurs in other amino acids that contain the –OH functional group such as tyrosine, hydroxyproline (Hyp) and hydroxylserine (Hyl). These will not be discussed in my thesis.
**Figure 1.3** - Examples of different types of protein Ser/Thr O-linked glycosylation. UDP or GDP sugar donors are shown as well as the major enzymes responsible for the initial sugar additions to Ser or Thr residues. Note that the –OH is from a hydroxylserine (Hyl) in O-Galactose glycosylation (F) and the –OH is from a Ser in O-Xylose glycosylation (G).
1.1.3 Non-Mucin Protein O-linked Glycosylation

O-GlcNAc

In comparison to the other types of O-linkages, O-GlcNAcylation is distinct from all other forms of protein glycosylation in many ways. First, there are only two enzymes responsible for the modulation of O-GlcNAcylation; the first is β-N-acetylglucosaminyl transferase (O-GlcNAc transferase, known as OGT), which transfers the GlcNAc sugar moiety onto the Ser/Thr acceptor on the protein substrate, and O-β-N-acetylglucosaminidase (O-GlcNAcase, known as OGA), which removes the peptide bound O-GlcNAc (see Figure 1.3 B). Secondly, this post-translational modification is the most abundant modification on nuclear, cytoplasmic and mitochondrial proteins of all plants and animals (26). In the normal life cycle of cytosolic and nuclear proteins, the GlcNAc is attached and removed multiple times at different rates and at different sites (27). Lastly, the GlcNAc is generally not elongated or modified to form complex structures (28). It is however, commonly involved a “cross-talk” with O-phosphorylation. Studies have shown that O-phosphorylation can compete with O-GlcNAcylation for the same site or they can occur at adjacent sites to one another on proteins.

The protein targets for O-GlcNAcylation (i.e. OGT) seem to have no consensus sequence; studies have shown that approximately half of the known O-GlcNAc sites contain a Pro-Val-Ser type of motif while the other half has nothing in common other than the high presence of Ser and Thr residues (29). Moreover, much work in this area has discovered that both enzymes have a TPR (a tetratricopeptide repeat domain), a protein interaction domain that is specifically targeted to substrates via many transient
protein/protein interactions with accessory (binding) proteins that produce holoenzymes, each with high degree of substrate specificity (30,31). The substrate selectivity of OGT has also been shown to be controlled by the cellular concentration of the UDP-GlcNAc (32).

Biologically O-GlcNAcylation is required for life as it regulates nearly every cellular process in response to nutrients and stress in plants and animals (32). It has several important roles in blocking O-phosphorylation, modulating signaling, transcription and translation, protein trafficking and turnover (28). Because of this, numerous diseases and disorders have been linked to O-GlcNAcylation including: diabetes, neurological diseases such as Parkinson’s dystonia and Alzheimer’s disease and cancer (28,33,34).

**O-Mannose**

O-mannosylation is initiated in the ER and it begins with the conversion of GDP-Mannose to the dolicholphosphate- mannose (Dol-P-Mannose) donor on the cystolic face of the ER, followed by the transfer of mannose from Dol-P-Mannose by the actions of two enzymes POMT1 and POMT2 to Ser and Thr residues on acceptor substrates (35,36) (Figure 1.3 C). This forms the core m1 structure, which is then elongated by the addition of other monosaccharide and functional groups (such as phosphates) forming a wide variety of O-mannosylated glycans. This is an important modification that occurs on proteins that are secreted, anchored to the plasma membrane wall (and cell wall) and other proteins in the secretory pathway in bacteria, fungi and mammals (37). Much work has been done in identifying substrates that are O-mannosylated and some studies on the
substrate specificity on POMT1, POMT2 and POMGnT1 (an elongating glycosyltransferase) have proposed sequence motifs but have yet to be confirmed (35,38,39).

In mammals, approximately 30% of all glycans are O-mannosylated in the brain (40). This has important implications in the development of the nervous system and disorders such as cobblestone lissencephaly, mental retardation and multiple sclerosis (35,40). Protein O-mannosylation also plays an important role in muscle structure and function, thus, defects in the glycosyltransferase enzymes responsible for proper O-mannosylation of α-dystroglycan (the best known O-mannosylated protein) have been shown to cause congenital muscular dystrophy (known as CMD), a disease that causes progressive weakness and wasteness of skeletal muscle (40). Similarly, mutations in the initiating enzymes POMT1 and POMT2 were identified in patients with Walker-Warburg syndrome (known as WWS); a recessive but severe form of CMD that is characterized by ocular abnormalities and severe brain malformation in patients (41). Other less severe mutations in these enzymes result in milder forms of CMD such as muscle-eye-brain disease (42).

Although O-GlcNAcylation and O-Mannosylation are biologically important, these were not the focus of my thesis.

1.2 Mucins

1.2.1 Mucin-Type Protein O-Glycosylation

Protein O-glycosylation of the mucin-type is one of the most common post-translational modifications that is initiated in the Golgi apparatus (and subsequently
elongated as it moves through the Golgi) with the stepwise addition of sugar residues by a series of specific glycosyltransferases. This results in the creation of a wide variety of heterogeneous O-glycan structures. The first step in this cascade is initiated by a large family of glycosyltransferases, called UDP-GalNAc:polypeptide N-α-acetylgalactosaminyl transferases (ppGalNAc T’s) (or GALNT1-20) (Figure 1.3 A). These enzymes define the sites of O-glycosylation by the transfer of N-acetylgalactosamine (GalNAc) from the sugar donor UDP-GalNAc onto the hydroxyl groups of Ser or Thr residues on acceptor protein substrates. This reaction creates the simplest form of mucin-type O-glycan, commonly referred to as the Tn-antigen due to the fact that most of these α-GalNAc structures are found in cancer cells. This Tn-antigen in cancer cells is commonly modified by sialic acid (N-acetylneuraminic acid and abbreviated as Neu5Ac) at the GalNAc C6 position forming the sialyl Tn-antigen (STn-antigen), which does not elongate. However, in normal cells, specific glycosyltransferases may glycosylate the GalNAc (Tn) structure creating a variety of core structures (Figure 1.4), which are further elongated. One of the most common mucin type O-glycans is the Core 1 structure, where a β-Galactose is added to the 3rd position of the GalNAc (i.e. Galβ1-3GalNAc). This is called the T-antigen, and it can also be sialylated (which stops elongation) or further elongated. These different core structures result from the different sugar attachments giving 8 possible core structures. Cores 1-4 are the most common which are found on the secreted mucins of mucin secreting tissues such as the bronchi, colon and salivary glands (43,44). These core structures are then further elongated in a stepwise yet competitive manner to create large complex O-linked glycoproteins. Because
Figure 1.4- Mucin-type O-glycan elongation pathway. Elongation of the Tn-antigen occurs through the stepwise addition of other specific glycosyltransferases to create 8 core structures. The enzymes shown in purple are being studied in the Gerken Lab.
the expression levels of the individual ppGalNAc T’s and glycosyltransferases vary from cell types and tissues, many different types of sugar attachments can occur in an O-linked glycoprotein, a phenomenon known as microheterogeneity.

The focus of my dissertation will be on determining the specificity of the enzymes involved in the initiation of mucin-type O-glycans, therefore, when O-glycans are mentioned in the rest of the chapters, they will refer to mucin-type O-glycan unless specified.

1.2.2 Mucin Type-O-Glycan Biosynthesis

Mucin type-O-glycosylation elongation was first thought of to mainly occur in the luminal regions of the cis-Golgi, but more recent studies have indicated that it occurs throughout the cis-, medial- and trans- Golgi regions, as studies have shown that the localization and compartmentalization of the ppGalNAc T’s and glycosyltransferases varies among these regions of the Golgi (45-47). Early acting enzymes are responsible for the formation of cores 1 and 2 O-glycans (such as the ppGalNAc T’s, C1GalT and C2GnT transferases) and are localized in the cis- and medial- Golgi regions while the later acting glycosyltransferases that elongate and cap the core O-glycan structures with other sugar units, are found in the trans-Golgi regions (45,48-52) (Figure 1.5). Over expression of the ST6GalNAc I sialyltransferase enzyme causes it to localize to the cis-Golgi, resulting in the sialylation of the early O-GalNAc glycans and subsequently terminating glycosylation in cancerous cells (53).

There are many factors that regulate the biosynthetic pathway of mucin O-glycosylation in the Golgi making this a very complex and widely not understood
**Figure 1.5**- Schematic representation of the biosynthesis of mucin-type O-glycans.
process. It is well established that the main factors that contribute to the regulation of O-glycosylation in the Golgi is the competing activities of the glycosyltransferases and their localization. The most common glycan chain terminators are sialic acid (N-acetyl neuraminic acid; abbreviated as Neu5Ac), a derivative of Neu5Ac called N-glycolylneuraminic acid (abbreviated as NeuG1) and Fucose (Fuc) (3,46). Other factors include the concentrations of the donor substrate and acceptor substrate, the substrate transport rate through the Golgi, metal ion Mn\(^{2+}\) concentrations, the luminal environment of the Golgi (including pH) and the overall membrane structural organization (3,45,47,54,55). Changes in pH can really alter glycosylation; several studies have identified an anion channel that regulates Golgi pH (56,57). Cells that are deficient in these ion channels lack the correct acidic environment and transport glycoproteins very slowly through the secretory pathway, resulting in truncated O-glycans. More recently, it has been discovered using cancer cell lines that the ppGalNAc T’s can relocate to the ER, through the regulation of Src (a proto-oncogene) (58,59). This in turn increases the levels of the Tn-antigen and enhances tumor cell migration and invasiveness. These findings indicate that the localization of the ppGalNAc T’s relies heavily on the physiological state of the cell (55).

Unlike N-linked glycosylation, this process occurs post-translationally and most importantly, there is no highly specific consensus motif determined to date. This makes sense, since there are twenty isoforms that can attach GalNAc as opposed to one transferase complex for N-linked glycosylation. Furthermore, our lab and others have shown that the specificity of one ppGalNAc T isoform may be different from another, making the analysis of target glycosylation sites difficult. However, there are a number of
semi-predictive database approaches (60-63) that have been developed which have revealed some predictive trends. The main observations are that Thr is a much better acceptor than Ser and that these glycosylation sites are associated with high contents of Ser, Thr, Pro and Gly. The other predictive trend that is observed among most ppGalNAc T’s is a common motif where a Pro residue is preferred at +1 and +3 C-terminal positions from the site of glycosylation (i.e. XXX-T-P(+1)XP(+3)) (64).

1.2.3 Structure

Mucins are very large heavily mucin-type O-glycosylated glycoproteins that are found in all animals. Mucins may be secreted or membrane-bound and serve to protect cell surfaces and epithelium. Mucin-type O-glycosylation occurs in mainly in the Golgi, but recently studies have shown that it can also occur in the ER and blood plasma (58). The mucin protein core typically compromises 20-10% of the molecular mass and contains regions with high contents of Pro, Gly, Ser and Thr, regions known as the “Variable number of tandem repeat” or VNTR regions. These regions have high numbers of O-glycans that compromise most of their mass approximately 80-90% (65,66). The clustering of these O-glycans forms the typical “bottle brush” conformation that provides stiffened and extended structures in solution that are highly resistant to proteases. The secreted mucins commonly have N- and C-terminal globular domains that oligomerize forming large networks. Structural examples of a secreted mucin (dimer, trimer and multimerization structures) and membrane bound mucins are shown in (Figure 1.6 and Figure 1.7 respectfully). Figure 1.8 A-E summarizes the structural motifs that mucins and mucin domains may play.
The secreted protein core may contain globular domains that are poorly 
glycosylated (including both O- and N-linked) known as the von Willebrand factor 
(vWF) C- and D-domains (Figure 1.6 A). In addition, cysteine rich knot domains are 
sometimes found in the VNTR region. The oligomerization (trimer formation) and 
multimerization (dimer formation) of the N- and C-terminal globular domains (Figure 1.6 
B-E) respectfully, via the vWF domains is instrumental for the formation of their large 
structures, leading to their viscous properties (67-70). Both secreted and membrane-
bound mucin domains (Figure 1.8) provide structural components that protect and 
hydrate cell surfaces, provides a scaffold for the modulation of cell-cell interactions 
including signaling/trafficking, protein sorting and targeting, regulates the inflammatory 
and immune responses while the lack of mucins (abnormal glycosylation) is involved in 
tumorogenesis and metastasis in cancer (46,71-76).
Figure 1.6- Examples of mucin domain structures of secreted MUC2 mucin. (A) Secreted Monomer of MUC2. (B) Dimerized monomer forms through the vWF C-domains. (C) Trimerized monomers form through the vWF N-terminal domains. (D) Both the dimers and trimers come together to form hexamer rings. (E) The hexamer rings come together through disulfide bonds to form a polymerized network (Johansson et al. (68) and Ambort et al. (69)).
Figure 1.7- Examples of mucin domain structures of membrane-bound mucin. (A) MUC1 and (B) MUC4 (note that MUC1 does not contain an EGF domain). (Jonckhree et al. (92) and Bafna et al. (93))
Secreted mucins are usually found on the epithelial cell surfaces of ocular, respiratory, digestive, reproductive and urogenital systems of all vertebrates (66,77). Specialized cells that secrete mucins in the intestinal and respiratory tracks are known as goblet cells. Mucins exist as secretions of gels (mucin network-like structures) (Figure 1.6 E and Figure 1.8 B) that primarily behave as barriers and play fundamental roles in cell lubrication, hydration and protection from pathogens (78,79). In humans, the most widely known and studied gel-forming secreted mucins are MUC2, MUC5AC, MUC5B, MUC6 and MUC19 (80) and interestingly, only the N- and C- globular domains are evolutionarily conserved within mammals (67). MUC2 is the principal gel-forming mucin in the epithelium of intestinal cells and has been shown to be an important barrier to protect against pathogenic invasion in the highly bacterial nature of the colon (81). Glycosyltransferase knock out studies using mice has shown that the deletion of C1GalT1 (the enzyme responsible for the T-antigen formation Galβ1-3GalNAc) within intestinal cells causes defective mucus barriers that leads to spontaneous colitis that resembles human ulcerative colitis (UC) (82). Similarly, MUC5AC and MUC5B, the two major mucins found in the lungs, have been shown to have protective functions against bacterial infections (83,84). However, the over expression of either MUC5AC or MUC5B can be detrimental as well, as the increased secretion of mucus can cause airway obstruction and are linked to diseases such as Cystic Fibrosis and Chronic Obstructive Pulmonary Disease (COPD) (85-88).

MUC6 is a major gastric mucin (89) and is also highly expressed in breast cancer tumors (90). MUC19 has been shown to be expressed in the human salivary glands and
Figure 1.8- Structural roles of mucin-type O-glycans. Secreted mucins (A-C) have stiffened and extended protease resistant structures and membrane-bound mucins (D-E) serve as scaffold and tethers for the interactions of ligands.
tracheal submucosal glands (91). In keeping with their functions, these mucins typically contain the N- and C- terminal vWF, VNTR and cysteine rich domains.

Membrane associated or membrane bound mucins are all type I membrane proteins containing one short cytoplasmic tail, a transmembrane domain and a large glycosylated domain that reaches far out of the cell surface (Figure 1.7) (67). They too function as barriers to protect the cell against other cells and larger molecules surrounding the cell surface. The most common membrane bound mucins in humans are MUC1, MUC3, MUC3A, MUC3B, MUC4, MUC12, MUC13, MUC15, MUC16, MUC17, MUC20, MUC21 and MUC22 (67,92). The membrane bound mucins contain additional extracellular regions that may contain a SEA domain (sea urchin sperm protein Enterokinase domain) and/or a epidermal growth factor-like (EGF) domains (Figure 1.7 A and B) (93). Some mucins have additional domains and in the case of MUC4 two additional domains where identified; the NIDO (nidogen-like domain) and AMOP domains (67,92). Although the overall functions of these domains remain to be fully elucidated, so far it has been shown that these domains can serve as protein/protein interacting modulators or interacting domains and involved in cleavage events and association of subunits (92). The surface of the membrane bound mucins can also provide a scaffold surface where enzymes, receptors or specific carbohydrate epitopes or ligands can bind to (Figure 1.8 E) (72).

The membrane mucins may be involved in detecting changes in the cellular environment. Changes in pH, ionic concentration, hydration, or physical interactions (with biological active molecules) can cause the release of the extra cellular domains such as the SEA domain (72,92). Both of these surface receptor and sensory functions of
mucins may also be involved in signaling pathways, either through signal transduction or through ligand and receptor binding to communicate to the epithelial cells about the condition of the external environment. These signals are transmitted to the intracellular domain through via post-translational modifications of the cytoplasmic tail or by other means. Two mucins that have been extensively studied and have been shown to have different signaling mechanisms are MUC1 and MUC4. MUC1 has been shown to be a docking protein for signaling molecules such as β-catenin (a regulator of transcription), FGFR3 (Fibroblast growth factor receptor 3) and other proteins involved in cancer regulation via the phosphorylation of its cytoplasmic tail (93-95). MUC4 on the other hand, has been shown to be a receptor ligand for various signaling proteins to control cell proliferation, differentiation and apoptosis. Important well studied examples are the interactions of the MUC4 EGF domain with the tyrosine kinase ERBB2 growth factor receptors (93,94).

1.2.4 Mucins in Health and Diseases

Aberrant glycosylation is a hallmark of several disease states, especially cancer since several tumor antigens are also mucins (96). This is largely due to the deregulation of the expression of the mucin core and the glycosyltransferases that modify them. For mucin-type O-glycans, incomplete glycosylation occurs in proteins that produces shortened or truncated glycans, known as the T-antigen (core 1 Galβ1–3GalNAcα-Ser/Thr) and Tn-antigen (GalNAcα-Ser/Thr) and their sialylated forms known as ST-antigen (Neu5Acα2-3Galβ1–3GalNAc-Ser/Thr) and STn-antigen (Neu5Acα2-6GalNAcα-Ser/Thr). This is due to the overexpression of ST6GalNAcI which is responsible for forming the STn-antigen and ST3Gal1 which forms the ST-antigen (97).
These activities cause the aberrant expression of tumor-associated epitopes on mucins (such as MUC1) in breast (98) and gastric cancers (99). The production of the STn-antigen is correlated with a decrease in cancer cell adhesion, increased tumor growth, increased tumor cell migration, invasion and poor prognosis (76). Because the STn-antigen is commonly detected in most patients with cancers such as pancreas, stomach, colon, breast, bladder and ovary (100-109), it has been proposed to be used a prognostic marker and a target for the design of anticancer vaccines (110,111). Indeed, CA125 (carcinoma antigen 125, i.e. MUC16) is a well established biomarker for ovarian cancer (112) and CA15-3 (MUC1) is a prognostic marker for breast cancer (113).

Many secreted and membrane bound proteins passing through the secretory pathway (63,114) have been shown to contain mucin type O-glycans outside of the classified mucin domains where they serve largely unknown functions. Several studies have suggested that such single sites of O-glycosylation may be involved in distinct biological processes and may play modulating and signaling roles. One of these mechanisms has been shown to occur in hormone like secreted proteins where glycosylation modulates the cleavage of proprotein convertase cleavage sites, resulting in hormone up or down regulation. For example, phosphaturic factor FGF23 (115), angiopoietin-like protein 3 (ANGPTL3) (116) and pro-brain natriuretic peptide proBNP (117). Studies on Drosophila Melanogaster have demonstrated that O-glycosylation protects an essential component of the secretory apparatus from furin-mediated proteolysis, thereby enabling proper secretory vesicle formation and polarized secretion of extracellular matrix components (118). Therefore, disruptions in O-glycosylation can affect the secretion and composition of the cellular environment. Despite the biological
relevance, we know very little about the actual site(s) of O-glycosylation, how they are regulated and even less is known of the actual molecular mechanism(s) of O-glycan involvement in development.

1.3 ppGalNAc T Transferases

1.3.1 History

Over 45 years ago, McQuire and Roseman published the first reported ppGalNAc T activity (119), where the enzyme was extracted from sheep maxillary glands. For their assays, they prepared de-glycosylated substrates also from sheep submaxillary mucin. They noticed that the enzyme had a high specificity towards UDP-[3H]-GalNAc and that radiolabel incorporation of GalNAc was proportional to the enzyme concentration and incubation time. A second attempt for the purification would follow in 1982 by Sugiura et al. (120), where a ppGalNAc T was purified to homogeneity from ascites hepatoma with a reported molecular weight of ~ 55 kDa. Upon their analysis of this enzyme with different protein acceptors, they noticed that peptide conformation around the acceptor site affected the rate of transfer. In 1986, Elhammer and Kornfield (121) purified a ppGalNAc T from two sources; a soluble form purified from bovine colostrum and a membrane-bound form purified from BW5147 mouse lymphoma cells. Both had a corresponding molecular weight of 70 kDa but there were reported differences in their enzymatic properties. This early research would not be fully appreciated until a decade later, when the Clausen laboratory in 1995 examined ovine and porcine submaxillary glands and discovered two ppGalNAc T transferases that could be differentiated from one another based on their activities and levels of tissue expression (122). These two
transferases would turn out to be the most ubiquitously expressed transferases in tissues and are known as ppGalNAc T1 and T2. Shortly after, a number of distinct ppGalNAc T’s were identified by the Clausen and Tabak laboratories including ppGalNAc T3 (123), T4 (124), T5 (125), T6 (126), T7 (127) and T9 (128), revealing a large number of mammalian transferases. Currently, there are 20 identified members of the ppGalNAc T family of transferases, with 15 of them being cloned, expressed and at least partially characterized (129-138) (Figure 1.9). Moreover, several ppGalNAc T orthologs (with very similar structures and conserved functions) have been identified in higher eukaryotes such as Drosophila melanogaster (i.e. PGANT) (139,140), Caenorhabditis elegans (125), Echinococcus granulosus (141) (i.e. Eg-ppGalNAc T), Toxoplasma gondii (i.e. T.gondii ppGalNAc T) (142-144) and Cryptosporidium species (i.e. Cp-ppGalNAc T’s) (145). O-linked GalNAc linkages have been identified in parasitic species such as Schistosoma mansoni (146) and Trypanosoma cruzi (147), where the enzymes have been shown to be distantly related to the mammalian ppGalNAc T’s.

1.3.2 Expression and Distribution

The mouse and Drosophila melanogaster are the biological models that have been extensively used to study the mammalian ppGalNAc T’s or fly PGANT’s specifically their expression and distribution and will be discussed separately below.

The mammalian ppGalNAc T’s are distributed among all organs with isoforms having different expression levels in each tissue. Earlier studies with Northern blots show that ppGalNAc T1 and T2 mRNAs are ubiquitously (>70%) expressed among all organs analyzed
**Figure 1.9**- Phylogenetic organization of the ppGalNAc T gene family (modified from 152). Peptide-preferring transferases in subfamilies Ia-Id and If-Ig are boxed in red while glycopeptide-preferring transferases in subfamilies IIa-IIb are boxed in blue. Subfamily Ie (or “Y” series) is boxed in gray due to their presently unknown activities (131).
This is also true in the case of the mouse, where the organs identified with the most ppGalNAc T1 and T2 include the brain, heart, liver, lungs, ovary, prostate, spleen, thymus, thyroid and colon (150). Interestingly, ppGalNAc T7 has been shown to be ubiquitously expressed, although in low levels, in the human (151,152) and also holds true for the mouse (150). The expression of other transferase isoforms, such as ppGalNAc T5, T8, T9, T10, T13, T15, T17, T19 and T20 however, are less ubiquitous and more regulated to a specific organs (153). ppGalNAc T8, T9, T13 and T19 have been shown to be specific to the brain (130,134,151,154), while T20 and T3 have been shown to be highly expressed in the testis (151). The expression levels of ppGalNAc T5, T10, T11, T15 and T17 seem to be high specifically in the kidney, brain, colon and heart. In the mouse, tissues with the highest transferase expression levels are the testis with ppGalNAc T3 (having the highest expression of all transferases seen in this study), the lung with high levels of T8 and the ovary with T10. (150). Lastly, the remaining transferases, ppGalNAc T3, T4, T6, T11, T12, T16 and T18 have more broad expression levels (<30%) spread across tissues (153). Interestingly, the colon has a high expression of ppGalNAc T4 in both man and mouse. How the unique substrate specificity of the ppGalNAc T’s relate to their tissue expression distributions will be discussed in later chapters.

*Drosophila melanogaster* has been used as a model to study the biological roles/functions of *O*-linked glycosylation since there are fewer PGANT’s, with 12 family members (155) and because they can be readily mutated or knocked out. Expression patterns for the different developmental stages, especially in the digestive and salivary tracks reveal that nearly all PGANTs are expressed. However, there are unique
expression patterns that vary for certain members that are observed within the embryonic and imaginal disc (portion that will become part of the adult insect) developmental stages. For example, PGANT3 is uniquely expressed in the pharynx and esophagus at embryonic stages 14-17 while PGANT2 expression is first detected in the developing tracheal branches during stages 12-13 (156). PGANT35A is also detected in the longitudinal dorsal trachea stages 16-17. In situ hybridization data revealed that PGANT 1-3 and CG30463 (a predicted family member, PGANT9) are specifically expressed in the regions of the wing imaginal disc that is destined to become the wing blade, wing hinge or body wall (156). Furthermore, PGANT2 and CG30463 have expression patterns in the eye-antennal disc. Early developmental studies have shown that inactive mutations in the PGANT35A (ppGalNAc T11 orthologue in man) is essential for viability due to disruption of the respiratory system (157) and mutations in PGANT3 (close orthologue of ppGalNAc T1 in man) causes decreased secretion of extracellular matrix components and loss of proper epithelial cell adhesion causing wing blistering (158,159). Four other PGANT transferases, PGANT4, PGANT5, PGANT7 and CG30463 have been identified to be individually essential for viability. PGANT5 (a ppGalNAc T1 orthologue) was identified to be essential for cell regulation and acidification of the digestive system, while more than one PGANT’s are necessary for the development of the mesoderm, digestive system and respiratory system (43,155).

1.3.3 Role of ppGalNAc T's in Cancer and Diseases

Several ppGalNAc T isoforms have been shown to be involved in distinct biological processes including signaling functions, development and normal cellular processes. For example, O-glycosylation of cell surface receptors may be regulated by
specific ppGalNAc T isoforms that appear to modulate receptor signaling by unknown mechanisms, these include: TGF-β R II (160) by ppGalNAc T16, death receptor DR5 (161) by ppGalNAc T14 and LDL-receptor (162) by ppGalNAc T2. A number of distinct ppGalNAc T isoforms have been associated with human diseases. The first being the loss of ppGalNAc T3 causing familial tumoral calcinosis due to an unregulated cleavage of FGF23 (163). The others include ppGalNAc T2, T5 and T19 which have been linked to levels of HDL cholesterol and coronary artery disease (164-166), hereditary multiple exostoses (167) and Williams-Beuren Syndrome (138) respectively. ppGalNAc T20 (GALNTL5) and ppGalNAc T3 have been shown to be required for mouse and human sperm motility and fertility (137,168). Furthermore, several ppGalNAc T’s such as T1-T3, T5-T7, T9, T11-T16 and T19 have been shown to be associated with various cancers (74,169-172). In an ongoing study, ppGalNAc T12 and other glycosyltransferases, β4GALT2, β3GNT2 and ST6GalNAcII are being investigated for their linkage to colon cancer (see chapter 5). The mechanisms of disease causation by the loss or over expression of the transferases in these latter examples are presently unknown.

O-glycosylation is critical for development, where as discussed above in the fly, several ppGalNAc T (PGANT) isoforms are essential for viability (43,140) while in the mouse the loss of the a single elongating Core 1 transferase (T-synthase, which adds the β-Gal to the 3rd position of the peptide GalNAc) is embryonic lethal (173). Interestingly, other studies in the mouse have shown that ppGalNAc T1 modulates salivary gland organogenesis (174) and is required for proper heart and valve development (175) while the overexpression of ppGalNAc T3 has been linked to dwarfism (176).
Although there are several examples of how single site(s) of glycosylation are involved in several disease states, we still know very little about the actual site(s) of \( O \)-glycosylation, how they are regulated and even less is known of the actual molecular mechanism(s) of \( O \)-glycan involvement in development. Further confounding the issue are the reports that ppGalNAc T18, T19 and T20 (and other members of the ‘Y’ (Ie) subfamily, see Figure 1.9 (132)) lack readily detectable GalNAc transferase activity against typical substrates (132,137,138) suggesting that these transferases may have unusual substrates and/or play completely different roles than the other family members.

1.3.4 ppGalNAc T Structure

All ppGalNAc T’s are type II Golgi membrane proteins, with a domain structure that includes a short N-terminal cytoplasmic tail, a hydrophobic transmembrane domain, a stem domain with variable lengths and a luminal N-terminal catalytic domain (8) that is connected by short linker to a C-terminal ricin-like lectin domain (Figure 1.10).

All transferases except for ppGalNAc T20 contain this unique lectin domain structure (151). The catalytic domain is much larger than the lectin domain, consisting of ~230 amino acids while the lectin domain is ~120 amino acid residues in length. The catalytic domain is responsible for all catalytic activity; the binding of the acceptor substrate and the UDP-GalNAc and also performs the subsequent transfer of GalNAc onto the acceptor substrate. Both domains are linked via a short flexible segment that varies in length among isoforms creating flexibility between domains and functions to position the lectin domain relative to the catalytic domain which could play important roles in driving the selection of subsequent GalNAc-modification sites in already
Figure 1.10- Crystal structure of ppGalNAc T2 (PDP 2FFU) by (178). The EA2 peptide (STTPAPTTK) is bound to the catalytic domain and are space-filled: N-terminal Ser is blue, C-terminal Lys is red, the Pro residues are purple and the Thr are in brown. The Asp (in the α-subdomain) residue involved in glycopeptide binding is space-filled pink. Approximate locations of the β and γ subdomains are shown.
glycosylated substrates (55). The lectin domain is responsible for the recognition and interaction with a GalNAc residue on a prior GalNAc glycosylated substrate. Much of my work has focused on understanding the roles of the catalytic and lectin domains on glycopeptide substrates, see chapters 3 and 4.

The X-ray crystal structures of ppGalNAc T1, T2 and T10 (177-180) have revealed that the catalytic domain has a GT-A structural motif that is characterized by having two interacting α-β-α Rossman-like folds (181) which contain residues that will bind to the uracil of the UDP-GalNAc. Another important motif is the metal coordinating (Mn\(^{2+}\)) DxH motif, which is responsible for binding the phosphate moiety of the UDP donor substrate group via the coordination of the Mn\(^{2+}\) ion. This motif is located near the C-terminus of the second Rossman fold. Studies have shown that the metal ions required for catalysis are ones that can form the octahedral geometry and the Mn\(^{2+}\) ion is the most preferred divalent metal ion (181). The histidine residue in the DxH motif is absolutely required; changing it to an Asp residue (to produce the more common DxD motif found in other glycosyltransferases) will result in significantly reduced or no catalytic activity within the family members studied (181,182). The next important motif is the Gal/GalNAc binding motif, this is the region that interacts with the GalNAc moiety of the UDP-GalNAc donor (153,178,179). Transferases that are classified as glycopeptide-preferring transferases (this is defined in the next section) may contain an additional binding pocket or surface that interacts with the glycopeptide GalNAc moiety (179,183-185). Another unique structural feature of this family of enzymes is a flexible loop which extends out of the UDP-GalNAc and substrate binding site on the catalytic domain which undergoes conformational changes between a so called open (inactive) and closed
(active) conformation upon binding of the UDP-GalNAc and substrate (178,186). This loop is located near some of the acceptor substrate binding residues, but mostly on the residues that bind to the UDP-GalNAc, as the UDP-GalNAc has been shown to be the key factor in keeping the loop closed in the active conformation (186). Recent work by Lira-Navarrete et al. has shown that the flexible loop of ppGalNAc T2 exists in multiple conformations along the catalytic cycle; either semi-open, open and closed and that this is tied to a key catalytic residue identified as Trp331, which essentially drives the closed conformation as the Trp331 undergoes conformational changes as well. When the enzyme is in its active (closed) form, the Trp331 adopts an “in” conformation where it is brought closer to the UDP-GalNAc and acceptor substrate. When the enzyme is in its inactive (open) conformation, the Trp331 adapts the “out” conformation (see Figure 1D of (180)) and is away from the UDP-GalNAc and acceptor substrate. Most of the residues that are directly involved with acceptor substrate binding are located in a pre-formed channel on the surface of the catalytic domain; these are Lys363, Phe361, Phe280 and Trp282. These residues are important for forming a “proline binding pocket” that dictates acceptor substrate preferences (178), see chapter 2.

The C-terminal lectin domain of the ppGalNAc T’s structure belongs to the Ricin-type lectin family which consists of a folded β-trefoil structure that is built from three homologous repeat units known as the α, β and γ repeats (153). The ricin-type family lectins are the most diverse family of lectins and are characteristically multi-domain proteins that have sugar recognition modules (187). In fact, these sugar binding modules are even commonly observed in many prokaryotic systems such as bacterial glycoside hydrolases, where it has been shown that these modules serve to increase local enzyme-
substrate concentrations and the targeting the enzyme to specific carbohydrate substrates (188-191). In the case of the ppGalNAc T’s, there have been numerous studies reporting on the lectin domain having similar functions by modulating glycopeptide recognition and specificity and also serving to increase transferase efficiency (184,192-196) (this will be further discussed in chapter 4). The β-trefoil repeats all share a common binding motif, known as the CLD and QXW motifs (where X can be any amino acid) that are ~40 amino acids apart (197-200) and the Asp residue of the CLD motif is critical for lectin binding activity. Mutagenesis studies on the ppGalNAc T’s have shown that only specific subdomains actively bind GalNAc. For ppGalNAc T1, the α and β subdomains are important for GalNAc binding, for ppGalNAc T2, T3 and T4 its the α-subdomain and interestingly for ppGalNAc T10 the β-subdomain was determined to be important for GalNAc binding (55,177,178,192,194,201,202). However, the reasons as to why some domains are more active than others is still unknown. This is difficult to determine since the lectin domain should not bind tightly to the peptide GalNAc (194) as this would be inhibiting and non-productive as I will discuss in chapter 4.

The short linker domain is a flexible segment that connects the catalytic and lectin domains and varies from ~10 to 25 amino acids in length depending on isoform. The flexibility and length of the linker is believed to control the relative orientation of the catalytic and domains by essentially bringing both domains together or apart while the extent of this motion is what controls the glycosylation pattern and preference of the enzyme against glycopeptide substrates (180,203). The recent three dimensional topography AFM images of ppGalNAc T2 with ligand or no ligand present have revealed how these domains may mechanistically work together (180). With these images, three
conformational states were observed; a highly compact structure where the two domains closely interact, a less compact structure that displays a clear slot between both domains and an extended structure with both domains clearly separated (180). This supports the extended state visualized in the X-ray crystal structure of ppGalNAc T2 by Fritz et al. (178) and also for the first time, could explain how these enzymes may behave in solution. This does not however, provide information of the distribution between the different conformational states.

1.3.5 ppGalNAc T Peptide Substrate Specificity

The large ppGalNAc T family consists of 20 different isoforms in man that are classified into two major families, designated as I for peptide-preferring transferases and II for glycopeptide (GalNAc-O-Ser/Thr containing) -preferring transferases based on their primary substrate preferences. These families are further divided into subfamilies (Ia-g and IIa-b) based on their amino acid sequence similarities shown in Figure 1.9 (153). Although these terms have been historically used, more distinction is needed as we have recently shown that there is significant overlap in peptide/glycopeptide specificity within transferase isoforms (185,203). Despite the abundance of O-glycosylation and ppGalNAc T isoforms, we still do not fully understand the peptide/glycopeptide acceptor specificity. Multiple studies have been performed both in vitro and in vivo to characterize the substrate preferences of the ppGalNAc T’s.

Most of the early in vitro studies began by using tissue extracts as the enzyme source and peptides from secreted mucins as substrates, specifically the tandem repeats of MUC1, MUC2, MUC5AC as substrates for substrate preference determination. These
studies all found that the primary amino acid sequence played a significant role in the activity of the ppGalNAc T’s (204-210). The first attempts at *in vivo* studies were conducted by Nehrke et al. (211,212). They used COS7 and MCF-7 cell lines and a human von Willebrand factor (vWF) reporter substrate with the sequence; PHMAQVTVGPG containing one single glycosylation site to determine the effects of the flanking amino acid sequences. They determined that the placement of charged residues at positions -1 and +3 reduced glycosylation. Further studies with more substrates and cell lines determined that regardless of cell type and substrate, a glutamic acid at positions -1 and +3 (from the site of glycosylation) would reduce glycosylation. Determining a transferase isoform-specific substrate *in vivo* turned out to be difficult, as the *in vivo* studies reflected work represented the action of multiple transferases with difference expression levels acting on only one substrate. Because of this, workers moved into performing *in vitro* assays using purified native recombinant ppGalNAc T’s with a variety short peptide substrates derived from mucin glycoproteins and other *O*-linked glycoproteins. Most of these *in vitro* studies utilized recombinant ppGalNAc T1-T7, T10 and T11 and short peptide (and glycopeptide) substrates such as the EA2, MUC7, PSGL-1-1b, HIV gp120, MUC1a, MUC2, MUC5AC, human von Willebrand factor and erythroprotein (126,128,195,213-222) these are given in Table 1.1. These studies were useful in determining which peptide and/or glycopeptide substrate(s) was most active with a ppGalNAc T isoform. However, most of these substrates contained multiple Ser and Thr acceptor sites and in most cases, these sites were not determined leading to an incomplete characterization with respect to their preferred acceptor glycosylation sites.
**Table 1.1** – Table of selected peptide and glycopeptide substrates used in *in vitro* studies with recombinant ppGalNAc T’s. Table taken and modified from Ten Hagen et al. (214).

<table>
<thead>
<tr>
<th>Substrate</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T10</th>
<th>T11</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA2-1 (PTTDSTTPAPT)k</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EA2-2 (DTTPAPT)k</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC1 (TAPAHGVTASADRTAPGASTAPP)</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC2 (PTTTPSTTTMVTPTPTPTC)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC5AC (GTTSPVPTTSSTTSAP)</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC5AC-3 (GTT*PSPVPTTSSTTSAP)</td>
<td>+</td>
<td>+++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC5AC-13 (GTTSPVPTTSSTTSAP)</td>
<td>+</td>
<td>+</td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC5AC-3,13 (GTT*PSPVPTTSSTTSAP)</td>
<td>+</td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUC7 (CPPTPSATTAPPSSSAPPETTAA)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSGL-1-1B (Ac-QATEYEYLDYDFLPETEPPEM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV gp120 (Ac-CIRIQRGPGRGVTIGKGNMR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1 represents multiple studies, therefore the comparison of GalNAc incorporation can only be made within each study. a) Bennett et al. (123), b) Ten Hagen et al. (125) & (127), c) Ten Hagen et al. (129), d) Bennett et al. (194), e) Ten Hagen et al. (212) and f) Bennett et al. (215). A (+) indicates GalNAc incorporation into substrate. A (-) indicated no activity and a blank indicates the transferase was not studied using the particular substrate. A (*) indicates the best substrate observed for a given transferase and substrate.
A more systematic study was first performed by Yoshida et al. (219), where ppGalNAc T1 was tested against over 100 synthetic peptides derived from the simple sequence (PDAASAAP) of the human erythroprotein (hEPO). By performing systematic substitutions on the amino acids surrounding the acceptor residues on the parent sequences AAATAAA and AAASAAA and by the eliminating multiple acceptor sites on the sequence, they were able to evaluate the best sequence for GalNAc transfer to ppGalNAc T1. Their results were useful in determining that Thr is a much better acceptor than Ser by ~40 fold and when Pro was in the flanking +1 and +3 positions from the site of glycosylation, the GalNAc incorporation was ~90 fold higher. This T-PXP motif was tested in vivo (in CHO cells) and was found to be an efficient O-glycosylation sequence (223). Interestingly, when a Pro was at either the +1 position or the +3 position, the GalNAc incorporation was increased by 7 fold or 33 fold respectively, from the parent peptide. This study was influential in showing the importance of Pro residues surrounding the acceptor site.

Additional studies performed in the Gerken Lab with the 81 residue mucin tandem repeat of PSM (porcine submaxillary mucin, containing over 31 sites of glycosylation) and CSM (canine submaxillary mucin, containing over 20 sites of glycosylation) showed differences in the glycosylation sites (or glycosylation patterns) between ppGalNAc T1 and ppGalNAc T2 (224,225), which were attributed to both neighboring sequence and prior sites of GalNAc glycosylation. Analysis of these sites revealed that Pro enhancements were also observed at the +1 position (from the site of glycosylation) with ppGalNAc T1 and Pro enhancements were also observed at the -1, -3
and +3 positions (from the site of glycosylation) for ppGalNAc T2. These results were consistent with the previous studies of Yoshida et al.

Studies performed by Pratt et al. (226) with several ppGalNAc T isoforms (ppGalNAc T1, T2, T3, T4, T5, T7 and T11) systematically determined the effects of neighboring prior GalNAc glycosylated residues on transferase activity. A glycopeptide library of 56 members with 0 to 4 prior GalNAc glycosylated sites (known as T*) on the EA2 substrate (PTTDSTTPAPPTK) was used against these seven ppGalNAc T isoforms. The results showed that the ppGalNAc T transferases had different glycopeptide substrate preferences to which they can be categorized into four types/categories. The first category belongs to the “early” transferases, ppGalNAc T1, T2 and T5 which preferred to glycosylate substrates that contained 0 to 1 previously glycosylated sites. The next category belongs to the “intermediate” transferases, ppGalNAc T3 and T4, which preferred to glycosylate glycopeptide substrates that contained 2 prior sites of glycosylation (a di-glycopeptide). The third category is for the “late” transferase, ppGalNAc T10, which preferred to glycosylate substrates that contained 3 to 4 sites of prior glycosylation (tri- and tetra- glycopeptides). The last category belonged to the outlier transferases ppGalNAc T7 and T11. ppGalNAc T7 was only active against glycopeptide substrates that contained prior sites of glycosylation at the -4 and +1 positions from the acceptor site. ppGalNAc T11 on the other hand, was the only transferase that did not show any specific substrate requirement and its activity was very limited in the number of glycopeptide substrates. This study was useful in showing that working together, the ppGalNAc T transferases can produce the densely glycosylated mucins.
Altogether these studies were useful in determining which substrates were most active with which transferase and that a Pro at the +1 and +3 sites enhances glycosylation for some transferases, yet no transferase specific consensus sequence was revealed. The work of the Gerken Lab using random peptide substrates had led to the ability to quantify these transferase specificities (which will be discussed below).

The Gerken lab has been working for several years on determining the substrate preferences of the ppGalNAc T’s using a systematic method of \textit{in vitro} experiments with a library of oriented random peptide substrates that is quantitative and highly reproducible. Currently, there are 10 transferases fully characterized; ppGalNAc T1, T2, T3, T5, T10, T11, T12, T13, T14, T16, (64,183,227), other transferases that are partially characterized; ppGalNAc T4, T6, T7 (unpublished data) and the fly orthologs of ppGalNAc T1, T2, T11 and T15 (PGANT5, PGANT2, PGANT35A, PGANT9A and PGANT9B) which are fully characterized (228)(unpublished). The oriented random substrates used to characterize the substrate preferences follow the general sequence: GAGA(\textit{X}_n)T(\textit{X}_n)AGAGK, where \textit{n} = 3 or 5 and the \textit{X}’s are randomized amino acid positions (with 8-10 different residue compositions of each library). This combinatorial chemistry yields a very large library of unique peptide sequences, \(~10^{10}\) different sequences. These studies proceed by partially glycosylating the random peptide substrate, isolating the glycopeptide product via lectin affinity chromatography and using Edman degradation amino acid sequencing to determine the change in compositions of the \textit{X} residues. Specific preferences or enhancement values are obtained by comparing the glycopeptide \textit{X} residue composition to the control pre-lectin (non-glycosylated) random peptide \textit{X} residue compositions at each Edman degradation cycle. From these
studies, we have observed transferase specificity that is due to acceptor characteristics such as peptide sequence, charge and composition and even neighboring glycosylation.

1.4 O-glycan Prediction Softwares

In collaboration with the Ming Ying lab in University of Texas at El Paso, the preferences for ppGalNAc T1, T2, T3, T5, T10, T11, T12, T13, T14 and T16 have been put into a web-based prediction software called ISOGlyp (Isoform Specific O-Glycosylation Prediction) found at http://isoglyp.utep.edu. Details of this approach will be discussed in chapter 2.

Recent work by the Clausen laboratory has introduced the SimpleCell technology which has helped to characterize the individual substrate targets of ppGalNAc T’s *in vivo*. This technology allows for the analysis of all O-glycosylated proteins (known as the O-glycoproteome) in a cell line of interest. Basically, the method relies on using human cell lines that can be made deficient in a specific gene (i.e. a specific ppGalNAc T transferase) and by subsequently terminating the elongation steps via a targeting knock out of Cosmc (a chaperone that is required for the Core 1 transferase which elongates O-glycans) using zinc finger nucleases. This produces truncated GalNAc-O-Ser/Thr (Tn antigen) and Sialyl-GalNAc-O-Ser/Thr (STn-antigen) glycans on proteins that can then be readily analyzed (229,230). The isolation of the truncated glycosylated proteins from either cell lysates or secretions is achieved through lectin affinity chromatography and the identification of glycan specific sites is achieved through tandem mass spectrometry. Using these methods they have been able to identify hundreds of unique O-glycoproteins and O-glycan sites using 12 human cell lines with knock out ppGalNAc T1-T3, T5, T11,
T12, T14 and T16 (63). Additionally, SimpleCell was also applied to CHO cells providing one of the first detailed insights into the O-glycosylation capacity of these cells (231).

This method has helped to validate several O-glycoproteins that were known to specifically be glycosylated by ppGalNAc T2, including ApoC-III and ANGPTL3; which have significant roles in lipid metabolism (116,232,233). Another major finding using the SimpleCell technology was the discovery of the O-glycosylation sites in the linker regions of LDLR receptor (63) and upon further investigation, it was revealed that ppGalNAc T11 was the enzyme responsible for the glycosylation (234) which is highly important for the function of LDLR. This demonstrates the importance of precise gene editing and its ability to uncover biological functions of specific ppGalNAc T members. This O-glycoproteome data work is being placed in a vector machine based predictor called NetOGlyc4.0 (http://www.cbs.dtu.dk/services/NetOGlyc-4.0/).

In an effort to better predict glycosylation sites and improve both prediction software programs, both the Gerken and the Clausen lab have begun to compare the ppGalNAc T transferase specific in vitro glycosylation preferences with the in vivo O-glycoproteomic transferase-specific proteins determined (235). A large panel of transferase-specific substrates determined by SimpleCell were put into ISOGlyP and the predictions were compared. In some cases, ISOGlyP was in agreement with the SimpleCell approach, where it was predicted that a transferase will glycosylate that specific protein that was determined for that transferase in vivo. In other cases, both were not in agreement, where ISOGlyP predicted that a protein will be glycosylated by a
specific transferase that was not determined for that transferase in vivo. Clearly, more work is needed on both ends in order to improve these methods.

1.5 Summary

Since most studies used a limited set of (glyco)peptide substrates (that most often differed among research groups) our understanding of each isoforms specificity is still unclear. A systematic study of the catalytic and lectin domain specificities of these transferases, using a common series of (glyco)peptide substrates, is needed to fully understand their properties. This data will be instrumental in understanding how each transferase glycosylates its target. It is important to know what factors govern O-glycan site selection by the ppGalNAc T’s and what factors modulate specific O-glycan elongation. Without a clear systematic understanding of the ppGalNAc T peptide and glycopeptide specificity and hence enzymology, it is difficult to deduce the molecular mechanisms underlying their biological functions. Results of these studies could lead to novel strategies for the treatment of a wide range of diseases of aberrant O-glycosylation.

My thesis will focus on the functional characterization of the catalytic and lectin domains of the different ppGalNAc T isoforms using peptide and glycopeptide substrates that will independently probe the functions of each domain. For the glycopeptide substrates, the main focus will be to study the effects of a prior GalNAc glycosylated residue (T*= GalNAc-O-Thr) that is placed nearby (1-5 residues) an acceptor site or remote (6-17 residues) of an acceptor site. Chapter 2 will discuss the catalytic domain peptide preferences determined for 7 ppGalNAc T’s using the random peptide approach and show that they display specific substrate preferences that vary among isoforms and in
some cases, are identical within subfamily members. In chapter 3, I have studied the catalytic domain glycopeptide preferences of 9 ppGalNAc T isoforms and two PGANT orthologs which belong to either the glycopeptide-preferring and peptide-preferring subfamilies and have shown that there are unique catalytic domain glycopeptide preferences between transferase isoform and subfamily (185). Chapter 4 will discuss the glycopeptide directionality preferences mediated by the lectin domain of 11 ppGalNAc T isoforms and four PGANT orthologues and show that the presence of a remote N- or C-terminal placed GalNAc-O-Thr can be an important determinant of overall catalytic activity and specificity which differ between transferase isoforms (203). Chapter 5 will discuss a collaboration project with the Guda and Markowitz Labs (CWRU) and the Woods Lab (Memorial University of Newfoundland) where several mutations on glycosyltransferase genes were identified in individuals with colon cancer. The effects of the mutations on the encoded proteins will be examined in order to determine their roles in colon cancer progression.

1.6 Potential Limitations of Our Work

The random (glyco)peptide substrates utilized in my thesis are relatively short and contain a relatively high content of Gly and Pro residues compared to most proteins and are therefore expected to possess extended random structures that cannot fully represent the complexity of a folded protein. We recognize that this may be a limitation to our methods, especially since previous work has shown that ppGalNAc T11 recognizes the tertiary structure of the linker regions of the LDLR receptor (234). Nevertheless, these (glyco)peptide substrates have provided highly useful information and previously not understood data that has helped us to more fully understand the unique roles of the
catalytic and lectin domains of these transferases. These data will be substantially useful in future studies of the ppGalNAc T against both folded and extended proteins and peptides.

1.7 References

insights into mechanism and specificity of O-GlcNAc transferase. *The EMBO Journal* 27, 2780-2788


that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40, 161-169


recognition that varies among members of the polypeptide GalNAc transferase family. *Glycobiology* **26**, 360-376


adjacent to serine and threonine residues. *Journal of Biological Chemistry* **268**, 22979-22983


CHAPTER 2

CHARACTERIZATION OF THE PEPTIDE SUBSTRATE SPECIFICITIES OF ppGalNAc T4, T7, T6, T11, T13, T14, T16, dT1 (PGANT35A), PGANT9A AND PGANT9B USING RANDOM PEPTIDE SUBSTRATES
2.1 Background and Significance

The peptide substrate specificity of the ppGalNAc T’s has been the subject of many research groups since the early 1990’s. Much of the early *in vitro* work on ppGalNAc T’s consisted of using short peptide sequences from mucins and other known O-linked glycoproteins that gave useful information (mainly trends) about the ppGalNAc T transferases and their substrates. These trends were that a glutamic acid at flanking positions -1 and +3 from the site of glycosylation (a Thr) significantly reduced glycosylation (1,2), that a Pro at the flanking +1 and +3 position from the site of glycosylation (a Thr) enhanced glycosylation and that a Thr residue was a much better acceptor and Ser (3-15). These studies also gave information as to which peptide (see Table 1.1 in chapter 1) and GalNAc glycopeptide (16) substrates were more active for each ppGalNAc T. Despite the large multiple efforts by these labs, these studies were not useful in determining transferase specific sequences. This was in part, due to the fact that each research group used their own series of substrates, leading to an incomplete overlap in the characterization of these transferases. Furthermore, this was a slow progressing field, as it took approximately 15 years to fully identify all 20 ppGalNAc T isoforms.

The Gerken lab became interested in the ppGalNAc T’s when their analysis of the porcine submaxilliary gland mucin tandem repeat revealed unique patterns of O-glycosylation (17). This was reported about the same time the first isoforms, ppGalNAc T1 and T2 were being reported in literature. Because peptide sequence was clearly modulating site specific glycosylation in the porcine mucin, the Gerken Lab began focusing on approaches to characterize the ppGalNAc specificities, with the goals to understand the following: 1) Why are there so many ppGalNAc T transferase isoforms?
2) What do they glycosylate, what are their substrate targets? 3) Can we identify any non-redundant functions? 4) And are their functions conserved across species? The Gerken lab has been working on these goals for several years beginning with a random peptide approach that has successfully provided peptide substrate preferences of the catalytic domain of several mammalian ppGalNAc T’s (T1, T2, T3, T5, T10 and T12) and fly orthologs (PGANT5 and PGANT2) (18-20). This approach consists of using random peptide substrates that have the general form: GAGA(Xₙ)T(Xₙ)AGAGK, where the central Thr acceptor is flanked by 3 or 5 (n) randomized amino acid positions (X) from the C- and N- terminal positions (i.e. if n=5, +[1-5] and –[1-5] positions, respectfully) of the central Thr. These random peptide substrate sequences currently in use are shown in Table 2.1. The random peptide substrates are partially glycosylated by a ppGalNAc T transferase using UDP-[³H]-GalNAc and the products are passed through several types of column chromatography techniques such as ion exchange chromatography (Dowex), size-exclusion chromatography (Sephadex G10) and lectin-affinity chromatography (with lectins SJA, SBA, HPA and VVA, see materials section for full names of lectins and procedures) to isolate the glycosylated glycopeptide. The glycopeptide product and starting non-glycosylated peptide control are Edman sequenced and the chromatograms integrated to calculate mole fractions of each amino acid at a given cycle (i.e. the randomized X positions). So called enhancement factors (EF) are obtained by dividing the obtained mole fractions of each amino acid at each cycle for the control by the glycosylated product.
Table 2.1- Random peptide substrates used in this work for the characterization of the catalytic domain specificity of the ppGalNAc T’s and fly orthologs.

<table>
<thead>
<tr>
<th>Peptide (name)</th>
<th>Sequence</th>
<th>Number of Unique Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVI</td>
<td>X = G, A, P, V, L, Y, E, Q, R, H</td>
<td>1.0 X 10^{10}</td>
</tr>
<tr>
<td>PVII</td>
<td>X = G, A, P, I, M, F, D, N, R, K</td>
<td>1.0 X 10^{10}</td>
</tr>
<tr>
<td>PVIII</td>
<td>X = G, A, P, V, Y, E, N, S, R, K</td>
<td>1.0 X 10^{10}</td>
</tr>
</tbody>
</table>

Note that the Thr acceptor position is underlined in red and the flanking amino acid residues are shaded in blue.
With these randomized substrates, unique preference data, essentially “fingerprints” of transferase-specific preferences, for all amino acids (except for Thr, Trp and Cys) can obtained, see Figures 2.1 and 2.2 for previously determined ppGalNAc T preferences (i.e. ppGalNAc T1, T2, T3, T5, T12 and T10). For simplification purposes, Thr, Trp and Cys were left out of the amino acid pool primarily because Cys residues will form disulfide bonds, forming higher-level complex peptide structures. A randomized Thr residue would create another acceptor site, further complicating the analysis. However, a Ser residue is allowed due to the fact that a Thr residue will be glycosylated much faster than Ser. Lastly, Trp residues create sequencing issues since the Trp is oxidized in the workup and is lost upon sequencing (earlier attempts in the Gerken Lab to reduce this oxidation failed to improve the detection of Trp).

Overall, we have found from these studies that transferase hydrophobic residue preferences vary from the -3 to +3 positions (from the site of glycosylation), depending on the isoform (Figure 2.1 A-F). A common trend that is observed with most ppGalNAc T transferases (except for ppGalNAc T10) is a C-terminal Pro-Gly/Ala-Pro motif, with a much higher preference for Pro at the +3 position than the +1 position. This correlates well with the glycosylation trends observed by Yoshida et al., as mentioned above. Transferase-specific preferences are observed at the N-terminal positions, especially at the -1 positions, where ppGalNAc T1, T2 and T5 (Figures 2.1 A, B and D) have preferences for Pro and where ppGalNAc T3 and T12 (Figures 2.1 C and F) have preferences for Val (note however that for some ppGalNAc T’s Pro and Val have similar preferences). Furthermore, at the -2 position, ppGalNAc T2 has a modest enhancement (~1.5 or less) for Gly, while ppGalNAc T3 has modest enhancements for Ala and Tyr.
Figure 2.1 - Hydrophobic catalytic domain amino acid residue preferences for subfamily Ia member ppGalNAc T1 (A) subfamily Ib member T2 (B) subfamily Ic member T3 (C) subfamily Id member T5 (D) subfamily Iib member T10 (E) and subfamily Iia member T12 (F) obtained from random peptide substrates PVI, PVII and PVIII. These are hydrophobic enhancement factors for amino acid residues Gly, Ala, Pro, Val, Ile, Leu, Met, Phe and Tyr. Note the similar C-terminal (+1 to +3) pattern of preferences for all transferases except for ppGalNAc T10. Data taken from (18, 20, 21).
Figure 2.2- Hydrophilic catalytic domain amino acid residue preferences for subfamily Ia member ppGalNAc T1 (A) subfamily Ib member T2 (B) subfamily Ic member T3 (C) subfamily Id member T5 (D) subfamily IIb member T10 (E) and subfamily member IIa member T12 (F) obtained from random peptide substrates PVI, PVII and PVIII. These are hydrophilic residues Glu, Asp, Gln, Asn, Arg, Lys, His and Ser. Data taken from (18, 20, 21).
ppGalNAc T12, on the other hand, has a large enhancement for Tyr at the -2 position. At the -3 position, enhancements are observed for Pro, Ile and Val for ppGalNAc T2, while ppGalNAc T3 and T12 have modest enhancements for Tyr. ppGalNAc T3 has an additional modest enhancement at the -3 position for Ala. For ppGalNAc T10, however, none of these trends were observed, in fact, the enhancement factors were significantly lower than the other transferases determined. There are only modest Pro enhancements at the -1, +1 and +2 positions (Figure 2.1 E), in complete contrast to the 2-5 fold Pro enhancements determined for the other transferase isoforms at the +1 and +3 positions. Other modest enhancements are observed for Gly and Ala at random C- and N- terminal positions. These results suggest that ppGalNAc T10 does not have highly specific peptide motifs. Perhaps this is due to its strict glycopeptide preferences (see chapter 3), where its catalytic domain was shown to accommodate a GalNAc-O-Ser/Thr (21). The hydrophilic residue preferences shown in Figure 2.2 A-F, do not display and obvious trends, only modest preferences are observed at random N- and C- terminal positions. Nevertheless, it was discovered that the ratio of basic to acidic residue enhancements varied between isoforms, suggesting overall charge also plays a role. Basically, the Gerken Lab discovered an inverse correlation between the charged residue enhancement values with the ppGalNAc T transferase isoelectric point and electrostatic surface charge, where the transferase electrostatic charge plays a role in modulating substrate specificity. ppGalNAc T1 and T2 mostly preferred acidic residues (Glu and Asp, Figure 2.2 A and B) and this corresponded with an electrostatic surface charge that was mostly basic for ppGalNAc T1 and T2. ppGalNAc T3 and T5 mostly preferred basic residues (Arg, Lys
and His Figure 2.2 C and D ) and this corresponded with an overall acidic surface charge on the transferases (20).

From these initial studies, it became clear that peptide sequence, neighboring glycosylation and overall charge serve to modulate each ppGalNAc T catalytic domain substrate specificity (20,21). The fact that we observe different peptide substrate specificities for each ppGalNAc T isoform, has given rise to predictive approaches. It was reasoned that by simply multiplying the enhancement values of a peptide sequence for a given transferase would provide a value that would represent the propensity for the sequence to be glycosylated. This value is called an enhancement value product, or EVP. From these enhancement values, optimal substrates as well as selective peptide substrates were identified for ppGalNAc T1 and T2, which were confirmed by experiments (18). Recently, we used these values to derive additional optimal substrates for ppGalNAc T3, T5 and T12. This approach has also successfully predicted/correlated experimental data on a number of additional peptides (20,22). Most importantly, the transferase specific enhancement values have been implemented in a web-based software, called ISOGlyP (Isoform Specific O-Glycosylation Prediction) that can be used to predict glycosylation sites for individual ppGalNAc T transferases for an entered peptide sequence. This is found in the web at http://ISOGlyP.utep.edu and available for the community, especially for the researchers in the glycobiology field.

In my studies, this random peptide approach was used to fully characterize the catalytic domain substrate preferences of ppGalNAc T11, T13, T14, T16 and fly orthologs PGANT35A (also called dT1 and is the mammalian ppGalNAc T11 ortholog), PGANT9A and PGANT9B (ortholog to the mammalian ppGalNAc T15) and partially
characterize the catalytic substrate preferences of ppGalNAc T4, T6 and T7. The results for ppGalNAc T11, T13, T14 and T16 have been added to the ISOGlyP database. In this chapter, the catalytic domain substrate preferences determined for these ppGalNAc T’s will be compared to their subfamily members that have been previously determined. Also, comparisons of the catalytic domain substrate preferences determined for the fly PGANT transferases will be compared to their mammalian ppGalNAc T orthologs.

2.2 Materials and Methods

2.2.1 Transferases

As in most of our work, the transferases were obtained from multiple sources and expression systems that were provided by our collaborators, see Table 2.2 for the list of transferases and sources. Transferases are either N-terminal truncated and affinity tagged constructs that were bound to affinity beads or affinity purified. Although the transferases were from different sources, the experiments completed using different sources of transferases yielded comparable results. Soluble N-terminal polyHis-tagged recombinant transferases ppGalNAc T6, T7, T11, T14, T16 and dT1 were expressed from High five insect cells as secreted proteins and purified from Ni-NTA aragose (Invitrogen) or SP-Sepharose (Sigma-Aldrich) and MonoQ 5/50 ion exchange chromatography (GE Healthcare) as previously described (5,23-25). These were stored at 4°C for varying lengths of time. Poly-His tagged ppGalNAc T4, T6, T7 and T13 were expressed from baculovirus-infected Sf9 cell system as secreted proteins. The transferases were N-terminally His-tagged catalytic domains bound to Ni-NTA-affinity beads (Thermo-
Table 2.2- Table of ppGalNAc T’s used in most of this work describing the source and expression system.

<table>
<thead>
<tr>
<th>ppGalNAc T Transferase</th>
<th>Supplier</th>
<th>Expression System</th>
<th>Secretion/Purification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bovine ppGalNAc T1</td>
<td>Elhammer Lab, Michigan</td>
<td>Baculovirus Sf9</td>
<td>Soluble 2</td>
</tr>
<tr>
<td>T2</td>
<td>Tabak Lab, Maryland Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 2</td>
</tr>
<tr>
<td>T3</td>
<td>Clausen Lab, Denmark Tabak Lab, Maryland</td>
<td>High Five Insect Cells Pic5ia pastoris</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>T4</td>
<td>Clausen Lab, Denmark Tabak Lab, Maryland Jarvis Lab, Wyoming Moremen Lab, Georgia</td>
<td>High Five Insect Cells Baculovirus Sf9 HEK293f (human)</td>
<td>Soluble 1 Beads 4 Soluble 2</td>
</tr>
<tr>
<td>T5</td>
<td>Moremen Lab, Georgia</td>
<td>HEK293f (human)</td>
<td>Soluble 2</td>
</tr>
<tr>
<td>T6</td>
<td>Jarvis Lab, Wyoming</td>
<td>Baculovirus Sf9</td>
<td>Beads 4</td>
</tr>
<tr>
<td>T7</td>
<td>Clausen Lab, Denmark Tabak Lab, Maryland Jarvis Lab, Wyoming</td>
<td>High Five Insect Cells Baculovirus Sf9 COS7 (mammalian)</td>
<td>Soluble 1 Beads 4 Soluble 2</td>
</tr>
<tr>
<td>T10</td>
<td>Moremen Lab, Georgia</td>
<td>HEK293f (human)</td>
<td>Soluble 2</td>
</tr>
<tr>
<td>T11</td>
<td>Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>T12</td>
<td>Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>T13</td>
<td>Clausen Lab, Denmark Tabak Lab, Maryland Jarvis Lab, Wyoming</td>
<td>High Five Insect Cells Baculovirus Sf9</td>
<td>Soluble 1 Beads 4</td>
</tr>
<tr>
<td>T14</td>
<td>Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>T16</td>
<td>Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>dT1 (PGANT35A)</td>
<td>Clausen Lab, Denmark</td>
<td>High Five Insect Cells</td>
<td>Soluble 1</td>
</tr>
<tr>
<td>PGANT7</td>
<td>Ten Hagen Lab, Maryland</td>
<td>COS7 (mammalian)</td>
<td>Media 3</td>
</tr>
<tr>
<td>PGANT9A</td>
<td>Ten Hagen Lab, Maryland</td>
<td>COS7 (mammalian)</td>
<td>Media 3</td>
</tr>
<tr>
<td>PGANT9B</td>
<td>Ten Hagen Lab, Maryland</td>
<td>COS7 (mammalian)</td>
<td>Media 3</td>
</tr>
</tbody>
</table>

Storage Conditions:
1. Stored at 4°C
2. Mixed with 40% glycerol and stored at -20°C
3. Stored frozen at -20°C
4. Used immediately
Fisher)\(^1\) (26). After extensive washing, the transferases were used for enzyme assays. These transferases were typically utilized immediately. ppGalNAc T4, T5 and T10 was expressed using HEK293f cells and purified Ni-NTA superflow (Qiagen) nickel affinity chromatography\(^1\), methods were similar to the expression of rat ST6GalI (27). These transferases were mixed with 40% glycerol and stored at -20°C. PGANT9A and 9B were expressed in COS7 cells and the recombinant secreted protein of the cell culture media was frozen for experiments (28). Reactions with the control media gave no significant UDP-[\(^3\)H]-GalNAc utilization compared to the expression media.

2.2.2 Reagents and Peptide Substrates

Random peptide substrates PVI, PVII and PVIII listed in Table 2.1, were custom synthesized by Sussex Research, Ottawa, ON (Canada) and by New England Peptides (NEP), Garner, MA. Stock solutions of 50 mg/mL or 100 mg/mL were prepared by lyophilizing from water several times and adjusting to pH 7-7.5 with either dilute NaOH and/or HCl. Fully N-acetylated and radiolabeled UDP-[\(^3\)H]-GalNAc (C\(_6\)-[\(^3\)H]-CH\(_2\)-labeled) was purchased from American Radiolabeled Chemicals Inc. (St. Louis, MO). Protease inhibitor cocktails P8340 and P8849 and UDP-N-acetyl-D-galactosamine (GalNAc) was purchased from Sigma-Aldrich (St. Louis, MO). Dowex anion exchange (1x8, 200 mesh) resin was purchased from Arcos Organics. Sephadex G10 was purchased from Amersham Biosciences (Little Chalfont, UK) and GE Healthcare. Sephadex G10 column running buffer was made with 50 mM acetic acid (Fisher) and pH adjusted to 4.5

\(^1\) For detailed methods and cDNA constructs, visit the University of Georgia’s Resource for Integrated Glycotechnology, “Glyco-Enzyme Repository” (http://glycoenzymes.ccrc.uga.edu/).
with ammonium hydroxide (Fisher). Immobilized lectins SJA (*Sophora japonica*), SBA (*Glycine max*) and HPA (*Helix pomatia*) were purchased from EY Laboratories (San Mateo, CA). Immobilized VVA (*Vicia villosa*) lectin was purchased from Vector Laboratories (Burlingame, CA). The lectin column running buffer was 250 mM Tris Base, 1M NaCl, 1 mM CaCl\(_2\) and 0.002% NaN\(_3\). Scintillation vials used are 7- mL HPDE and purchased from Fisher.

2.2.3 Instrumentation

Edman sequencing was performed on a Procise 494 protein sequencer from Applied Biosystems (Foster City, CA). The sequencer is equipped with a reverse-phase C18 PTH column for PTH-amino acid analysis. Samples were lyophilized on a Virtis Freezemobile 12EL (Gardinger, NY). Liquid Scintillation counting was performed on a Beckman Model LS6500. UV-Vis measurements were taken using a Beckman Coulter Du 530 spectrophotometer (Fulleron, CA). Reaction vials were placed in a thermostatted TAITEC shaking Microincubator M-36 (San Jose, CA). Fraction collectors that are hooked up to the chromatography columns are TELEDYNE ISCO Foxy 200 fraction collector (Lincoln, NE).

2.2.4 Glycosylation Assays with Recombinant ppGalNAc T’s and Random Peptide Substrates (*PVI, PVII, PVIII*)

Reaction mixtures for the random peptide glycosylation consisted of 10 mM MnCl\(_2\), 50 mM Sodium Cacodylate, pH 6.5, 1.3 mM 2-mercaptoethanol, 2 mM UDP-GalNAc (containing a total of 100 µCi UDP-[^3]H]-GalNAc), 1/100 dilution of Protease
Inhibitor Cocktails P8340 and P8849, 0.3% Sodium Azide, 5 mg/mL (~15 mM) of either random peptide substrates PVI, PVII and PVIII and either 50 µL- 325 µL of soluble purified transferase (or transferase bound to affinity beads). Total reaction volumes ranged from 200 µL to 450 µL and were carried out in 1 mL capped Eppendorf tubes. Experiments performed with all three substrates were performed independently of one another with the same transferase concentrations and the same UDP-[³H]-GalNAc stock. The transferases used were T4, T7, T6, T11, T13, T14, T16, dT1 (Drosophila PGANT35A, the ortholog to the human T11), PGANT9A and PGANT9B. Detailed information for each specific transferase will be discussed at the appropriate sections, below.

All reactions were incubated at 37°C in the shaking incubator and quenched after ~20-24 hrs (overnight) reaction with one volume of 250 mM EDTA. After diluting to 5 mL, UDP and non-hydrolyzed UDP-GalNAc were removed from the sample by passing through a ~3 mL column of Dowex anion exchange (1x8) resin, Cl⁻ form. [³H]-GalNAc incorporation was determined by scintillation counting 1/50 of the sample before and after the Dowex column. Total UDP-GalNAc utilization (transfer to substrate) typically ranged from 1-3%. The eluent was lyophilized and chromatographed on a Sephadex G10 column (113 X 0.7 cm) collecting ~2 mL fractions. A portion of each fraction (typically 200 µL) was counted via scintillation counting and also monitored for absorbance at 220 nm and 280 nm. Fractions that contained both [³H]-GalNAc transfer to substrate activity and OD 220/280 absorbance were pooled and lyophilized.
2.2.5 Isolation of the Glycosylated Random Peptide Product

The lyophilized pooled fractions from the Sephadex G10 were passed over on an immobilized mixed bed lectin column (10 mL, 22 X 0.8 cm) containing a mixture of 2 mL or more or each lectin: SJA, SBA, HPA and VVA equilibrated at 4°C. The lyophilized sample was diluted to 1 mL with dH₂O and as a control for amino acid sequencing, a 100 µL aliquot was taken from the sample before loading onto the column and labeled as “Pre-Lectin”. The remaining sample was loaded onto the column. After several hours, when the fraction collector was around tube #17, the glycopeptide product (GalNAc-O-Ser/Thr) was released from the column by washing with a solution 40-50 mg (~45 mM) of GalNAc in 4 mL of column running buffer. The fractions were monitored and pooled based on [³H]-GalNAc radioactivity and 220 nm and 280 nm absorbance. The [³H]-GalNAc glycopeptide product was lyophilized and separated from the free GalNAc and running buffers on a second Sephadex G10 run and pooled as described above. This pooled fraction would then be labeled as “Post-Lectin”. Samples were lyophilized an additional three to four times from water to remove the acetic acid and ammonium hydroxide from the Sephadex G10 running buffer and to further concentrate the sample for Edman sequencing.

2.2.6 Edman Amino Acid Sequencing

Pulsed liquid phase Edman amino acid sequencing was performed on an Applied Biosystems Procise 494 protein sequencer using a C18 PTH column temperature of 55°C on both the “Pre-Lectin” and “Post-Lectin” samples. The chromatographs were integrated for the peak areas using the ABI Procise software. When necessary, base-line errors were
corrected by estimation or cutting and weighing the individual peaks. The mole fraction of each residue type for the “Pre-Lectin” and “Post-Lectin” was calculated from the obtained composition of each amino acid from the Edman sequencing analysis at each randomized position (i.e. cycles 5-9 and 11-15). The central Thr residue (at cycle 10) was assigned a 0.

Transferase-specific enhancement factors (EF) were obtained by dividing the mole fractions of each individual amino acid residue at each random cycle position of the “Post-Lectin” by the “Pre-Lectin” (Post/Pre). Multiple reactions against all three random peptides were performed for each transferase. For a full characterization, at least 2 determinations were made for each individual substrate (PVI, PVII and PVIII) for a given transferase and the final data was averaged. All of the data analysis was performed using custom spreadsheets using Microsoft Excel and Lotus 123 software that was developed in the Gerken Lab.

The resulting plots of transferase preferences are shown in Figures 2.1, 2.2, 2.4, 2.5, 2.6, 2.7, 2.9, 2.11 and 2.12. In these plots, 0 represents the site of glycosylation and the positions from the +1 to +5 are the C-terminal sites while positions -1 to -5 are the N-terminal sites (or the flanking X positions). Enhancement factors greater than 1 indicate an elevated preference for the residue given by the transferase while a value less than 1 indicates a decreased preference. All numerical EVP values for the amino acids determined at each randomized position for all transferases are shown in Appendix A.

2.3 Results and Discussion

2.3.1 Full Characterization of the Catalytic Domain Peptide Substrate Preferences of ppGalNAc T11, T13, T14, T16, dT1, PGANT9A and PGANT9B
2.3.1.1 ppGalNAc T13 in Subfamily Ia

These reactions were performed using the standard reaction conditions described in the methods section using 50-150 µL of soluble affinity purified ppGalNAc T13 (Clausen Lab) or transferase bound to affinity beads (Jarvis Lab). After overnight incubations, the reactions were quenched with 1X EDTA. Samples were passed thorough the Dowex column to remove the unreacted UDP-[^3]H-GalNAc and UDP, lyophilized and passed through a Sephadex G10 column to separate the glycosylated products from the buffers and salts. Typical G10 plots are shown in Figure 2.3 A. A typical G10 plot will show two peaks of[^3]H-GalNAc radioactivity, where the first eluting peak corresponds to transfer to substrate while the second eluting peak is non-productive UDP-GalNAc hydrolysis (i.e. transfer to water). This is commonly observed with the ppGalNAc T transferases and is dependent on the ppGalNAc T transferase and substrate that was used. This will be later discussed in chapter 3. The glycopeptide product was isolated from the non-glycosylated product via lectin affinity chromatography. Note a 100 µL aliquot “Pre-Lectin” control was removed prior to loading. The glycopeptide product binds tightly to the lectin beads and is eluted by the addition a solution of ~45 mM (40-50 mg) GalNAc. This peak appears later in the chromatogram see Figure 2.3 B (after the arrow). The peak is pooled and lyophilized based on[^3]H-GalNAc content and 220 nm and 280 nm absorbance. This “Post-Lectin” sample was run on Sephadex G10 column, where the glycosylated product is separated from the free GalNAc and also further cleaned up from lectin buffers. The sample was pooled based on[^3]H-GalNAc and 220 nm and 280 nm absorbance (Figure 2.3 C) and subject to four lyophilizations from water to remove the Sephadex G10 buffers and to further concentrate the sample.
Figure 2.3- Chromatography plots for ppGalNAc T13 using random peptides PVI, PVII and PVIII. (A) First sephadex G10 run on random peptides separating the glycopeptide product (first eluting peak, fractions ~25-34) from the buffers, reagents and un-glycosylated substrate (shown in the purple downward triangle, fractions ~40-50). Hydrolysis peaks are marked with the orange diamonds (fractions ~36-40). (B) Lectin affinity chromatography on the pooled glycopeptide fraction from the first G10. The eluted glycopeptide/GalNAc peak is in fractions ~20-25. A small un-glycosylated peptide peak is observed at fractions ~5-10 (marked with a purple line). Red arrow indicates the approximate time the GalNAc is added to the column. (C) Last Sephadex G10 run on the pooled glycopeptide/GalNAc peak from the prior lectin run. The glycopeptide product in fractions ~25-32 (black line) is separated from the free GalNAc in fractions ~34-45 (purple circles).
Several reactions were performed with the random peptides against ppGalNAc T13 but only the 10 best experiments (based on $[^3]$H content and gel filtration peak shape) were selected to be analyzed on the Edman sequencer (3 reactions with PVI, 2 reactions with PVII and 5 reactions with PVIII). Both the control “Pre-Lectin” sample and the “Post-Lectin” sample of each reaction were sequenced on the Procise 494 protein sequencer. The chromatograms and integrated data for the “Pre-Lectin” and “Post-Lectin” samples were treated as described above to obtain the amino acid preference ratios (or enhancement factors). These are all averaged and plotted in Figure 2.4 A and B.

In these plots 0 represents the site of glycosylation, flanking the site are the ±5 randomized amino acid (X) positions. Enhancement factors greater than 1 indicate an elevated preference for a given residue by the transferase while values less than 1 indicate a decreased preference. In Figure 2.4 A, hydrophobic amino acid preferences are plotted for T13 which show that the most sequence sensitivity between the -3 to +3 positions and in particular there is observed elevated preferences for Pro at the -1, +1 and +3 positions. There are also elevated preferences for Gly at the +2 position and a slight preference for Val at the -1 position. This common C-terminal Pro-Gly/Ala-Pro motif is observed with ppGalNAc T1, T2, T3, T5 and T12 which have been previously characterized by our lab (20). In keeping with common trends among transferases, there are very low enhancement factors for hydrophilic residues Glu, Asp, Gln, Asn, Arg, Lys, His and Ser, where no common patterns are observed (Figure 2.4 B). However, upon closer analysis there are seemingly higher preferences for an acidic Glu and Asp residues at positions -5 to +1, which is consistent with the trend observed for ppGalNAc T1 which had higher preferences for acidic residues for Glu and Asp.
Figure 2.4- Catalytic domain amino acid residue preferences for subfamily Ia members ppGalNAc T13 (A-B) and T1 (C-D) obtained from random peptide substrates PVI, PVII and PVIII. (A) The hydrophobic enhancement factors for amino acid residues Gly, Ala, Pro, Val, Ile, Leu, Met, Phe and Tyr. The determined preference motif for ppGalNAc T13 is (P/V)-T-PGP. (B) Enhancement factors for the hydrophilic residues Glu, Asp, Gln, Asn, Arg, Lys, His and Ser. (C) The hydrophobic enhancement factors for ppGalNAc T1, the determined preference motif is (P/V)-T-PGP. (D) The hydrophilic enhancement factors of amino acids Glu, Asp, Gln, Asn, Arg, Lys, His and Ser. Note that the substrate preferences of ppGalNAc T1 was determined by other lab personnel and has been published (18). (E) Enhancement factor value of each amino acid determined for ppGalNAc T13 and T1 are plotted against each other in this comparison plot.
Comparison of the Catalytic Domain Preferences of ppGalNAc T13 with Subfamily Ia

Member ppGalNAc T1

The GALNT genes that express both ppGalNAc T1 (GALNT1) and T13 (GALNT13) are classified into peptide-preferring subfamily Ia because of their 88% sequence identity (29,30). ppGalNAc T1 is one of the most characterized transferase and our lab has determined its preferences previously (Figure 2.4 C-D). Using our library of random peptide substrates, I have shown that the catalytic domains of ppGalNAc T1 and T13 have very similar if not nearly identical substrate preferences compare Figures 2.4 A and D. Even though the enhancement factors are not exactly the same for ppGalNAc T1 and T13, the overall trend is remarkably similar. This is confirmed in the comparison plot in Figure 2.4 E, where the enhancement factor values of each amino acid determined for ppGalNAc T13 and T1 are plotted against one another. Although the $R^2$ value is 0.583 (and not 1), it is still a good indication of similarity between the data points. The C-terminal preferences have elevated preferences for Pro at the +1 and +3 positions and at the +2 position, an elevated preference for Gly. A slight difference is observed N-terminal to the site of glycosylation, specifically at the -1 position. While both ppGalNAc T1 and T13 have enhancement factors that are close (or above 2) for Pro and Val, ppGalNAc T1 has a slightly higher preference for Pro while ppGalNAc T13 has a slightly higher preference for Val.

This is in contrast to studies performed by Zhang et al. (31) where they used several derived mucin and human syndecan-3 (glyco)peptides with more than one acceptor site (and glycosylated sites) with ppGalNAc T1 and T13 and observed differences in their ability to glycosylate different sites. Although these differences could
not be explained (or proven) at the time, I have determined that these differences are largely due to their lectin domains binding initially to the glycosylated substrate. This will be further discussed in chapter 4, in my studies of the lectin domain preferences.

2.3.1.2 ppGalNAc T14 and T16 in Subfamily Ib

These reactions were performed using the standard reaction conditions described in the methods section using 150-250 µL of soluble affinity purified ppGalNAc T14 (Clausen Lab) and 100-150 µL of soluble affinity purified T16 (Clausen Lab). After overnight incubations, the reactions were quenched with 1X EDTA. All chromatography runs, pools and lyophilizations were conducted the same way as described in the above sections. Due to redundancy, the chromatographs will not be shown, but are similar to those shown in Figure 2.3 for ppGalNAc T13.

For ppGalNAc T14, a total of 6 determinations were made: 2 reactions for PVI, 2 reactions for PVII and 2 reactions for PVIII. For ppGalNAc T16, a total of 11 determinations were made: 4 for PVI, 4 for PVII and 3 for PVIII. Plots containing the preferences for the hydrophobic and hydrophilic residues for ppGalNAc T14 and T16 are given in Figures 2.5 and Figure 2.6 respectfully.

In Figure 2.5 A-B, the common C-terminal hydrophobic amino acid preferences are observed for ppGalNAc T14 and T16; enhancement factors are close to 2 for Pro at the +1 position, higher enhancement factors (closer to 3) for Pro at the +3 position and preferences for Gly at the +2 position. The N-terminal preferences are also similar for the -1 and -2 positions. At the -1 position, there is a large (2-fold) enhancement for Pro for both ppGalNAc T14 and T16 while at the -2 position; there is an elevated preference for
Gly. At -3 positions however, there are slight differences. ppGalNAc T14 has a slightly preference for Val while ppGalNAc T16 has a preference for both Val and Ile (although Ile is slightly higher).

The hydrophilic residue preferences for ppGalNAc T14 and T16 shown in Figure 2.6 A-B, which are very low and do not have any obvious common patterns. There is however, a trend for a higher preference for acidic residues Glu and Asp at the randomized +5 to −5 flanking positions, which is similar to what was observed for ppGalNAc T2.

**Comparison of the Catalytic Domain Preferences of Subfamily Ib Members ppGalNAc T2, T14 and T16**

These transferases are classified into the peptide-preferring subfamily Ib, due to their very similar gene sequences (GALNT2, GALNT14, GALNT16 [or GALNTL1]). The catalytic domains of ppGalNAc T2 and T16 share a 56% sequence identity (30) while ppGalNAc T2 and T14 share a 67% sequence identity (29) and ppGalNAc T14 and T16 share a sequence identity of 53%. ppGalNAc T2 is the only member that is well characterized in this family, however with this work, I have been able to show that the catalytic domain preferences of ppGalNAc T2, T14 and T16 are very similar. The comparison plots of the enhancement factors of all amino acids determined for ppGalNAc T14 and T16 are plotted against T2 in Figure 2.5 D and E respectfully, while the enhancement factors of all amino acids for ppGalNAc T16 are plotted against T14 in Figure 2.5 F. The similarity between ppGalNAc T14 and T2 gives
Figure 2.5 - Catalytic domain hydrophobic amino acid residue preferences for subfamily Ib members ppGalNAc T14 (A), T16 (B) and T2 (C) obtained from random peptide substrates PVI, PVII and PVIII. (A) The determined preference motif for ppGalNAc T14 is GP-T-PGP. (B) The determined preference motif for ppGalNAc T16 is (V/I)GP-T-PGP. (C) The determined preference motif for ppGalNAc T2 is (P/V/I)GP-T-PGP. Note that the substrate preferences of ppGalNAc T2 was determined by other lab personnel and has been published (18). (D) Comparison plot of the enhancement factors determined for all amino acids with ppGalNAc T14 are plotted against T2. (E) Comparison plot of the enhancement factors determined for all amino acids with ppGalNAc T16 are plotted against T2. (F) Comparison plot of the enhancement factors determined for all amino acids for ppGalNAc T16 plotted against T14.
**Figure 2.6**- Catalytic domain hydrophilic amino acid residue preferences for subfamily Ib members ppGalNAc T14 (A), T16 (B) and T2 (C) obtained from random peptide substrates PVI, PVII and PVIII. Note that the substrate preferences of ppGalNAc T2 was determined by other lab personnel and has been published (18).
a coefficient of determination ($R^2$) of 0.657, meaning that the values are not far from one another. For ppGalNAc T16, the similarity with ppGalNAc T2 is much higher, with an $R^2$ of 0.800, even though their sequences share a lower percentage of sequence identity (56%) than ppGalNAc T14 and T2 (67% sequence identity). For ppGalNAc T16 and T14, the $R^2$ is 0.796, again a relatively high number considering that their sequence identity is 53%, the lowest percentage from all.

All three transferases display the common C-terminal Pro-Gly-Pro motif, including the common trend where the enhancement factors for the Pro at the +3 position is higher than the Pro enhancement factors at the +1 position. What is unique about this family is the very large (2 fold) N-terminal enhancement for Pro at the -1 position and the preferences for the -3 position, which vary among a Pro/Val/Ile preference. These subtle differences however may also be due to experimental error.

Much prior work of other research groups has focused on looking at the glycopeptide activities of ppGalNAc T2 and T16 using several mucin derived substrates and other substrates and have shown that ppGalNAc T16 may have a distinct glycopeptide specificity from that of ppGalNAc T2 (30). It has been speculated that this is due to the differences in their lectin domains. These differences in the lectin domains of ppGalNAc T2, T14 and T16 will be examined and discussed in chapter 4.

2.3.1.3 ppGalNAc T11 and dT1 (PGANT35A)

These reactions were performed using the standard reaction conditions described in the methods section using 100-150 µL of soluble affinity purified ppGalNAc T11 (Clausen Lab) and 100-200 µL of soluble affinity purified dT1 (Clausen Lab). After
overnight incubations, the reactions were quenched with 1X EDTA. All chromatography runs, pools and lyophilizations were conducted the same way as described in the above sections. Again, due to redundancy, the chromatographs will not be shown, but see Figure 2.3 for examples.

For ppGalNAc T11, a total of 6 determinations were made: 2 reactions for PVI, 2 reactions for PVII and 2 reactions for PVIII. For ppGalNAc dT1, a total of 11 determinations were made; 4 for PVI, 4 for PVII and 3 for PVIII. The plots containing the preferences for the hydrophobic and hydrophilic residues for ppGalNAc T11 and dT1 are in Figure 2.7. In Figure 2.7 A and C, the hydrophobic amino acid residue preferences are shown for both ppGalNAc T11 and dT1 respectfully. Again, both transferases have the common C-terminal Pro-Gly-Pro motif, however, these transferases are unique because they both have enhancement factors that are higher than 2 for Pro at the +4 position. This high of a preference for Pro at this position has not been observed with any of the other transferases, thus it is unlikely that it is due to sequencing lag. The N-terminal preferences show a range of preferences specifically at the -1 position. For ppGalNAc T11, there is a high Pro enhancement, followed by Ala and Gly, on the other hand, for dT1 there is about an equal preference for Gly, Ala, Pro and Val.

The hydrophilic residue preferences are shown in Figure 2.7 B and D. Again, there were no strong patterns or preferences observed for both ppGalNAc T11 and dT1. Interestingly, ppGalNAc T11 seems to have a higher preference for acidic residues Glu and Asp while dT1 (PGANT35A) has a higher preference for the basic residue Lys.
Figure 2.7- Catalytic domain hydrophobic amino acid residue preferences for subfamily If members ppGalNAc T11 (A) and dT1 (C) and hydrophilic amino acid residue preferences for ppGalNAc T11 (B) and dT1 (D) obtained from random peptide substrates PVI, PVII and PVIII. E) Comparison plot of the enhancement factors for all amino acids for ppGalNAc T11 plotted against the enhancement factors for all amino acids for dT1 (PGANT35A).
Comparison of the Catalytic Domain Preferences of ppGalNAc T11 with Subfamily If Member dT1 (PGANT35A)

As mentioned above, *Drosophila dT1* (PGANT35A) is the ortholog of the human ppGalNAc T11 and is in the larger multi-species phylogenetic tree in subfamily If along the ppGalNAc T’s from frog, chicken, fish, fly, *C. elegans* and *Toxoplasma Gondii* (29). Figure 2.8. The sequence similarities between ppGalNAc T11 and dT1 are ~71% within the conserved GT1 fold regions and the Gal/GalNAc-T (catalytic region and UDP-Gal binding) motif (28). Previous work has determined that they both have similar specificity (28,32,33) against a handful of common substrates, such that most substrate preferences are retained.

The fact that these basic catalytic functions are conserved across the species suggests that these functions may be biologically important. Interestingly, their *in vivo* functions seem to differ as previous work has shown that dT1 deficient files cannot be rescued with human ppGalNAc T11, resulting in lethality (34). With my work, I have been able to show that the substrate preferences of the catalytic domain of dT1 and T11 are indeed very similar with a small subtle difference at the N-terminal -1 position. This is confirmed in the comparison plots in Figure 2.7 E, where the R² value is 0.86, meaning that their preferences are nearly identical. It is unlikely that this difference is what is responsible for apparent different *in vivo* functions; perhaps these differences are in their glycopeptide (GalNAc-containing) preferences modulated by their catalytic or lectin domains. This will be explored in chapters 3 and 4.
Figure 2.8- Multi-species ppGalNAc T phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Taken and modified from Bennett et al. (29). Red arrows in subfamily If are pointing to the human T11 and dT1 (PGANT35A).
2.3.1.4 PGANT9A and PGANT9B

These two fly transferases PGANT9A and 9B were provided to us by the Ten Hagen Lab as a collaboration project, where our goals were to use our library of random (glyco)peptide substrates to characterize the catalytic and lectin domain preferences. PGANT9 is the essential main gene (CG30463) in the fly and the gene is alternatively spliced into two variants, PGANT9A and PGANT9B. The location of the splice is in the lectin domain. In the fly, PGANT9 has been shown to be required in specific tissues for development (35), while PAGNT9A has been shown to be the dominant transferase in larval tissues and in salivary larval tissues, however, both PGANT9A and PGANT9B are highly expressed (unpublished). This work would accompany the first manuscript of the characterization of PGANT9 function in the fly.

These reactions were performed using the standard reaction conditions described in the methods section using 164 µL of PGANT9A in media (Ten Hagen Lab) and 164-220 µL of PGANT9B in media (Ten Hagen Lab). After overnight incubations, the reactions were quenched with 1X EDTA. All chromatography runs, pools and lyophilizations were conducted the same way as described in the above sections. Again, due to redundancy, the chromatographs will not be shown, but see Figure 2.3 for examples.

For PGANT9A, a total of 10 determinations were made: 4 reactions for PVI, 2 reactions for PVII and 4 reactions for PVIII. For PGANT9B, a total of 10 determinations were made: 5 for PVI, 2 for PVII and 3 for PVIII. The plots containing the preferences for the hydrophobic and hydrophilic residues for PGANT9A and PGANT9B are in Figure 2.9. In Figure 2.9 A and C, the hydrophobic amino acid residue preferences are
**Figure 2.9**- Catalytic domain hydrophobic amino acid residue preferences for PGANT9A (A) and PGANT9B (C) and hydrophilic amino acid residue preferences for PGANT9A (B) and PGANT9B (D) obtained from random peptide substrates PVI, PVII and PVIII. (E) Comparison plot of the enhancement factors for all amino acids for PGANT9B plotted against the enhancement factors for all amino acids for PGANT9A. Note that hydrophobic residues Ile and Phe were removed from the plots due to their very low enhancement factors for PGANT9A and PGANT9B (see Appendix A for values).
shown for both PGANT9A and PGANT9B respectfully. Note that both Ile and Phe were removed from the plots, due to the visibly lower enhancement factors of PGANT9A and PGANT9B with PVII, presumably due to the lack of a Val in the sequence. Again the common C-terminal Pro-Gly-Pro motif is observed with both fly transferases and similarly to T11 and dT1, there is a high preference for Pro (>1.5) at the +4 position. The N-terminal -1 position for both PGANT9A and PGANT9B has a high preference (>2) for Val and just like ppGalNAc T1 and ppGalNAc T12, a smaller enhancement (>1.2) for Tyr.

The hydrophilic residue preferences shown in Figure 2.9 B and D does not show any strong patterns but does show some differences between PGANT9A and PGANT9B. For PGANT9A, there are high preferences that are close to 2 for the acidic Asp residue at the -1 and -2 positions, see Figure 2.9 B. For PGANT9B, there are modest preferences that are greater than 1.5 for the acidic Asp residue at the -1, +1 and +2 positions. Interestingly, there are also modest preferences for a Ser (that are greater than 1.5) at the -1, +1 and +2 positions for PGANT9A.

**Comparison of the Catalytic Domain Preferences of PGANT9A and PGANT9B**

In the larger multi-species phylogenetic tree containing the ppGalNAc T’s from frog, chicken, fish, fly, *C. elegans* and *Toxoplasma Gondii* (29), PGANT9 (CG30462 gene) is closely related to the human ppGalNAc T15 in subfamily Ig shown in Figure 2.10. Neither PGANT9 or ppGalNAc T15 have been characterized and will not be discussed. PGANT9A and PGANT9B are both lectin domain splice variants of the
Figure 2.10- Multi-species ppGalNAc T phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Taken and modified from Bennett et al. (29). Red arrows in subfamily If are pointing to PGANT9 and human T15.
CG30462 gene and results with a BLAST of their protein sequences show that they both have similarities to ppGalNAc T1 and T13 (Ten Hagen Lab, unpublished). Based on our results, PGANT9A and PGANT9B have very similar preferences with a correlation coefficient $R^2$ value of 0.699. Since these are lectin domain splice variants, their lectin domain preferences will be tested and compared in chapter 4.

2.3.2 Partially Characterized Transferases ppGalNAc T4, T6 and T7

This section will focus on discussing the partial characterizations of three transferases, ppGalNAc T4, T6 and T7 each in a different ppGalNAc T subfamily. This data is incomplete, since only about half of the number of reactions have been obtained to plot the final data due to the limited amount of transferase or lack of active transferase. Additional reactions are needed to fully characterize the substrate preferences. Therefore, what is shown are preliminary substrate preferences; as they could potentially change with additional reactions with the random peptide substrates. The catalytic domain substrate preferences will be discussed but not fully compared with their subfamily members in this chapter; however, full comparisons of their glycopeptide activities of their catalytic and lectin domains with subfamily members will be discussed in chapters 3 and 4.

*ppGalNAc T4, T6 and T7*

Reactions were performed as described above with 100 µL of soluble ppGalNAc T4 (Clausen Lab) or 125-325 µL T4 in beads (Jarvis Lab), 125-230 µL of ppGalNAc T6 in beads (Jarvis Lab) and 100-125 µL of soluble ppGalNAc T7 (Clausen Lab) and 100
µL of ppGalNAc T7 in beads from the Jarvis Lab or Tabak Lab. After 20-24 hrs of incubation, the reactions were quenched with 1X EDTA. All chromatography runs, pools and lyophilizations were conducted the same way as described in the above sections.

For ppGalNAc T4, a total of 4 determinations were made: 1 reaction for PVI, 2 reactions for PVII and 1 reaction for PVIII. For ppGalNAc T6, a total of three determinations were made: 2 reactions for PVI and 1 reaction for PVIII. For ppGalNAc T7 a total of 4 determinations were made: 1 reaction for PVI, 1 reaction for PVII and 2 reactions for PVIII. The plots containing the preferences for the hydrophobic and hydrophilic residues for ppGalNAc T4, T6 and T7 are given in Figures 2.11 and Figure 2.12 respectively.

2.3.2.1 Comparison of the Catalytic Domain Preferences of ppGalNAc T4 with Subfamily IIa Member T12

These transferases are part of subfamily IIa, one of the subfamily groups that are considered to be mostly “glycopeptide-preferring” transferases. Their catalytic domain sequences are 72% identical (29). The catalytic domain of ppGalNAc T12 has been previously characterized by our lab using our library of random peptide substrates which shows that it does indeed have significant preferences towards peptide substrates (20) although it occurs with high hydrolysis. Apart from containing the common C-terminal Pro-Ala-Pro motif, ppGalNAc T12 displayed a unique preference for Arg at the +2 position and unique N-terminal preferences, which included elevated preferences for Tyr at the -3 and -2 positions and about an equal preference for Val, Ile and Met at the -1
Figure 2.11 - Partial catalytic domain hydrophobic amino acid residue preferences for (A) ppGalNAc T4 (subfamily IIa), (B) ppGalNAc T6 (subfamily Ic) and (C) ppGalNAc T7 (subfamily IIb) obtained from random peptide substrates PVI, PVII and PVIII.
Figure 2.12- Partial catalytic domain hydrophilic amino acid residue preferences for (A) ppGalNAc T4 (subfamily Ila), (B) ppGalNAc T6 (subfamily Ic) and (C) ppGalNAc T7 (subfamily IIb) obtained from random peptide substrates PVI, PVII and PVIII.
position see Table 2.3. Further studies that I have performed with glycopeptide substrates on the catalytic domain and lectin domains of ppGalNAc T12 also shows that there are unique glycopeptide specificities in both domains. These will be discussed in chapters 3 and 4.

Our random peptide results for ppGalNAc T4 confirms that this transferase will also glycosylate naked peptides. In Figure 2.11 A, the hydrophobic amino acid residue preferences show the common C-terminal Pro-Gly-Pro motif, while the N-terminal preferences only show an elevated preference for Pro and Val at the -1 position. More work is needed to fully characterize these preferences before making comparisons to the other ppGalNAc T’s. The hydrophobic residue preferences shown in Figure 2.11 A show a higher than usual preference for Gln especially at the +1 position. Again more work needs to be done in order to confirm whether this is real or not. In chapters 3 and 4, the unique glycopeptide specificity of the catalytic and lectin domains of ppGalNAc T4 will be discussed.

The fact that we are observing peptide activities with transferases that are classified as mostly “glycopeptide-preferring” transferases brings in a significant overlap in the classification of these transferases. In order to indicate their peptide specificities, the classifications should be changed/renamed. My discussion on this issue is found in chapters 3 and 4.
Preferences for ppGalNAc T3, T10 and T12 were previously determined by our lab (18, 20, 21). The X’s represent the randomized amino acids. Amino acids preferences that were not observed in the motifs were left with an X.

**Table 2.3** - Transferase specific catalytic domain preferences determined for subfamily members Ic, IIa and IIb derived from random peptide substrates PVI, PVII and PVIII.

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Transferase</th>
<th>Preference Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ic</td>
<td>ppGalNAc T3</td>
<td>X(A/Y)(V/I)TPGP</td>
</tr>
<tr>
<td></td>
<td>ppGalNAc T6</td>
<td>XXVTP(G/A)P</td>
</tr>
<tr>
<td>IIa</td>
<td>ppGalNAc T4</td>
<td>XX(P/V)TP(G/A)P</td>
</tr>
<tr>
<td></td>
<td>ppGalNAc T12</td>
<td>YY(V/I/M)TP(A/R)P</td>
</tr>
<tr>
<td>IIb</td>
<td>ppGalNAc T7</td>
<td>XXXTXPY</td>
</tr>
<tr>
<td></td>
<td>ppGalNAc T10</td>
<td>XXXTXPX</td>
</tr>
</tbody>
</table>
2.3.2.2 Comparison of the Catalytic Domain Preferences of ppGalNAc T6 with Subfamily Ic Member T3

PPGalNAc T6 is in subfamily Ic which is classified to be a “peptide-preferring” transferase family. The sequence conservation of the catalytic domains of ppGalNAc T3 and ppGalNAc T6 is 77% (29). Previous work with in vitro substrates showed that they both had similar substrate specificities (that were not commonly seen with other transferases) and that their expressions patterns are different among cell/tissue type (5). The determined catalytic domain preferences of ppGalNAc T3 by our lab has shown that the hydrophobic amino acid preference patterns are unique. The C-terminal preferences for ppGalNAc T3 contained the commonly observed Pro-Gly-Pro motif, however, the N-terminal preferences showed a unique elevated (2 fold) enhancement factor for Val and a smaller enhancement factor for Ile at the -1 position. At the -2 position, there were equal preferences for Ala and Tyr (see Table 2.3). In Figure 2.11 B, the hydrophobic amino acid preferences for ppGalNAc T6 are shown. Although not all enhancement factors for all amino acids were determined, the beginnings of a similar trend could be observed, where the C-terminal has Pro enhancements at the +1 and +3 and a Val enhancement is observed at the N-terminal -1 position. Regardless, more determinations need to be made in order to confirm these preliminary results. The hydrophilic residue preferences of ppGalNAc T6 (in Figure 2.12 B) show an elevated preference for a Ser at the +1, again, this was only determined from one reaction and could not be a real trend. One reason for the lack of strong enhancements observed is the poor quality of the sequencing of many of these runs due to low signal to noise and presence of contaminating overlapping peaks in the chromatograms.
Although ppGalNAc T3 and T6 are considered to be predominantly “peptide-preferring” transferases, their respective glycopeptide activities with their catalytic and lectin domains will be discussed in chapters 3 and 4.

2.3.2.3 Comparison of the catalytic domain preferences of ppGalNAc T7 with subfamily IIb member T10

This subfamily is the second subfamily of transferases that are considered “glycopeptide-preferring” (29). They both share a 48% sequence identity and interestingly seem to have similar substrate characteristics. The catalytic domain preferences of ppGalNAc T10 has been previously reported by our lab using our library of random peptide substrates and has determined that there are no large enhancements observed for the hydrophobic and hydrophilic residues except for a slight Pro enhancement at the +2 position (21) (see Table 2.3). This makes sense, if ppGalNAc T10 is truly a “glycopeptide-preferring” transferase then perhaps there shouldn’t be any peptide preferences observed. Note however, we have shown that ppGalNAc T10 has an 8 fold preference for GalNAc-O-Ser/Thr at the -1 position (21). Thus, ppGalNAc T10 binds glycopeptides in its catalytic domain. In Figure 2.11 C, the hydrophobic amino acid residue preferences for ppGalNAc T7 show no specific patterns, except for small enhancements for Tyr at the -1 and +3 positions and Pro at the +2. However, since these are preliminary determinations and not all hydrophobic amino acids were surveyed it is difficult to say whether these enhancements are real. The hydrophilic amino acid preferences for ppGalNAc T7 are shown in Figure 2.12 C and again no large patterns are observed.
Further studies on the glycopeptide preferences of the catalytic and lectin domains of ppGalNAc T7 will be further discussed in chapters 3 and 4 in comparison with ppGalNAc T10. Indeed we do observe the strong GalNAc-O-Ser/Thr preference at the -1 position for ppGalNAc T7.

**2.4 Summary and Conclusions**

By utilizing the library of random peptide substrates, I have been able to fully characterize an additional four ppGalNAc T transferases (T11, T13, T14 and T16), three fly transferase orthologs dT1 (PGANT35A), PGANT9A and PGANT9B and partially characterize three ppGalNAc T transferases T4, T6 and T7. I have determined preferences for all amino acids except for Thr, Trp and Cys. The common C-terminal Pro-Gly/Ala-Pro motif observed on almost all ppGalNAc T transferases fully characterized (except for ppGalNAc T10 and possibly T7) is governed by several conserved catalytic domain residues that are involved in the peptide substrate binding. We and others have determined that these key catalytic residues are Phe280, Trp282 and Phe361 see Table 2.4 to see these residues modeled onto the catalytic domain of ppGalNAc T2 (20,36,37). The proline residues 8 and 10 in the EA2 substrate essentially “clasp” on onto the conserved Trp residue (Trp282 in ppGalNAc T2). This Trp residue is found in the peptide binding cleft of most ppGalNAc T transferases except for ppGalNAc T7 and T10. The other two Phe residues (Phe280 and Phe361 in ppGalNAc T2) were also thought of to provide a structural motif that is predictive for the Pro-Gly/Ala-Pro motif because they are conserved among all transferases characterized except for ppGalNAc T7 and T10. Table 2.4 shows that ppGalNAc T10 and T7 has a Tyr in replacement of the
Phe361, an Arg in replacement of the Trp282 and a Trp in replacement for Phe280 (for ppGalNAc T7) and a Tyr in replacement for Phe280 (for ppGalNAc T10), which results in the loss of the observed Pro-Gly/Ala-Pro motif. These differences in preferences can be seen in Figure 2.1 for ppGalNAc T10. Other reports show that there are two other additional residues (besides the aromatic Trp282 and Phe361) that are important for forming the “proline-pocket” and they are Val255, Leu270, based on the X-ray crystal structure of ppGalNAc T2 with EA2 peptide bound (36,38). The Pro10 of EA2 is inserted into a cavity surrounded by hydrophobic residues Leu270, Val255 and Phe361 and is stacked against the side chain of Trp282, giving the observed enhancement for Pro at the +3. It has also been mentioned that the role of the Pro residues in the T-P(G/A)P motif expose the Ser/Thr residues in a β-turn conformation leading to more efficient glycosylation (39).

There are also similarities in specificity within subfamily Ia members (ppGalNAc T1 and T13), subfamily Ib members (ppGalNAc T2, T14 and T16) as well as specificities conserved among the ppGalNAc T’s and the fly orthologs, such as T11 and dT1 (PGANT35A) (and prior studies for ppGalNAc T1 & PGANT5 and ppGalNAc T2 & PGANT2). Although specificities are similar among family members, there are subtle differences in the preferences between other ppGalNAc T subfamilies. Perhaps more determinations are required with higher incorporation of GalNAc and better sequencing in order to really make these differences less subtle and more obvious. These differences are key for predicting sites of glycosylation.
Table 2.4- Correlation of Catalytic Domain Residues with the “Classic” T-P(G/A)P Substrate Motif for Selective ppGalNAc T’s and PGANT’s characterized by the Gerken Lab.

The top figure displays the X-ray crystal structure of the ppGalNAc T2 catalytic domain bound to EA2 peptide substrate (36). The prominent residues thought to be involved in the binding of EA2 peptide substrate are space filled as are the EA2 substrate residues. ppGalNAc T2 substrate PAP interacting residues are: Phe 361 (steel blue), Phe 280 (olive green) and Trp 282 (green). The EA2, PAP residues are: Pro8 (purple), Ala9 (orange) and Pro10 (purple). The N-terminal and C-terminal EA2 residues, Ser5 (blue), Lys10 (red) are labeled white. The remaining EA2 residues are in orange. The table below the figure lists the conservation of each isoform’s P(G/A)P substrate motif as Yes (Y) or No (N) along with the identity of the conserved P(G/A)P interacting aromatic residues. Note the high correlation for the residues across subfamilies and ortholog pairs. Preferences were obtained from (18, 20, 21) or from unpublished preliminary data, indicated with an *. 

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>ppGalNAc T</th>
<th>P(G/A)P motif</th>
<th>Phe361</th>
<th>Phe280</th>
<th>Trp282</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ib</td>
<td>T2</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T14</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T16</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>Ia</td>
<td>T1</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T13</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>Ic</td>
<td>T3</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T6</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>Id</td>
<td>T5</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>If</td>
<td>T11</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>dT1 (PGANT35A)</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>Ig</td>
<td>PGANT9</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IIa</td>
<td>T4</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T12</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IIb</td>
<td>T7</td>
<td>N*</td>
<td>Y</td>
<td>W</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>T10</td>
<td>N*</td>
<td>Y</td>
<td>Y</td>
<td>R</td>
</tr>
</tbody>
</table>

EA2 Peptide: $^{S_5}TT_PAP_{13}$
For the transferases that were partially characterized, more work needs to be performed in order to fully characterize and compare the substrate preferences within family members. How these data are being used to calculate probabilities and predict sites of O-glycosylation is described below.

Enhancement Value Products

For prediction purposes with the web-based program ISOGlyP, the predictive transferase-specific enhancement value product (EVP) is obtained for a given sequence by multiplying together the transferase specific enhancement factors (EF) of the amino acids flanking the site of glycosylation. (These EF values determined for each transferase are taken from our random peptide determinations in Figures 2.1, 2.2, 2.4, 2.5, 2.6, 2.7, 2.9, 2.11, 2.12 & Appendix A for each transferase). EVP’s are believed to reflect the rate of glycosylation (or the likelihood) that a particular sequence will be glycosylated by a particular transferase. For amino acids that do not have any determined EF values such and Thr, Trp and Cys they are assigned a value of 1. Basically, the higher the EVP value the more likely (and faster) that the given sequence will be glycosylated by the particular transferase isoform. A number close to 1 is regarded as a neutral number, it’s not an inhibitory site nor is it enhanced. A number less than 1 indicates that the transferase does not prefer that particular sequence and will most likely not glycosylate that site. An EVP value of 2 would indicate a 2-fold higher rate of glycosylation while very high EVP values (~10 or higher) indicates that the particular sequence is an exceptional one for the specific transferase.
Examples of how this works are shown in Figures 2.13 and 2.14. Figure 2.13 displays the home webpage for ISOGlyP. In this page, you can insert the sequence to be analyzed in two ways: Option 1 you can upload the file with the input sequence in FASTA format, or Option 2 you can simply type in the sequence. In this area, I have typed in the optimal sequences that were determined for ppGalNAc T1, T2, T3, T5, T10, T11, T12, T13, T14 and T16. After inputting the sequence, the next thing to do is to pick the prediction constraints, exactly how many N- and C-terminal positions from the site of glycosylation would you like to analyze. Next to this are the Cys and Thr values, they are automatically chosen as 1, but you have the option to change it to a Ser residue (since Cys and Ser differ only by the functional group i.e. –SH to –OH respectfully and because Ser is similar to Thr). Lastly, you can choose any (or all) of the transferases that you would like to analyze and hit submit. You will be directed to a page that has the prediction output similar to what is shown in Figure 2.14, where the first column will have name of the inputted sequence, followed by the Ser/Thr position, then a column that shows the sequence pattern used to calculate predictions and lastly all of the transferases selected with the EVP values underneath. If you look only within a ppGalNAc T transferase column, the EVP values can predict (and correlate with) the optimal sequences for the ppGalNAc T transferases (indicated by the high EVPs in the red boxes). However, if you compare each individual sequence with all of the transferases, there are a lot of EVP values that are higher than the expected ppGalNAc T transferase. For example, with ppGalNAc T1, the optimal sequence is DAP-T-PGP. If you compare the EVP’s for the ppGalNAc T1 column only with all the other 9
Figure 2.13- ISOGlyP homepage containing all of the parameters necessary for transferase-sequence specific EVP value determinations.
optimal sequences, you still see that the highest EVP value (36.13) will only be for the T1 optimal sequence, boxed in red. However, when you consider only looking at T1 OptimalSequence and look across the other ppGalNAc T transferases, there are four other transferases (ppGalNAc T2, T11, T14 and T16) with higher EVP values than ppGalNAc T1, these are underlined in red.

The only optimal sequence that is successfully shown to be transferase specific is the ppGalNAc T12 optimal sequence as seen in Figure 2.14, where no other transferase has a high EVP value for this sequence and T12 does not have any high EVP values with any of the other sequences. We are presently attempting to confirm these predictions experimentally.

There are several cases where our ISOGlyP predictions match experimental in vivo results. One is FGF23, which is a protein that regulates phosphate concentration in plasma and is specifically glycosylated by ppGalNAc T3. Loss of FGF23 activity by furin cleavage is thought to lead to increased phosphate levels and the clinical syndrome called familial tumoral calcinosis. The glycosylation of the Thr-178 by ppGalNAc T3 prevents the cleavage of FGF23 at a neighboring site and increases its secretion. Glycosylation studies with ppGalNAc T1, T2 and T3 on FGF23 by Kato et al. (40) has shown that the glycosylation of Thr-178 in FGF23 is performed by only ppGalNAc T3 and not by T1 or T2. However, all three transferases will glycosylate a nearby site, Thr-171. This result can be predicted using ISOGlyP and is shown in Table 2.5. For the Thr-171 site, all three transferases have product preferences that are greater than 1 indicating that all three transferases should glycosylate this site. For the Thr-178 site, the ISOGlyP EVP predictions for ppGalNAc T1 and T2 are less than 1 (0.29 and 0.16,
**Figure 2.14**- ISOGlyP output for ppGalNAc T1, T2, T3, T5, T10, T11, T12, T13, T14 and T15, along with their determined optimal sequences.
**Table 2.5** - Random peptide enhancement value products predict experimental transferase-specific sites of glycosylation. Table taken from (20).

| Transferase Sites Glycosylated | L  | I  | H  | F  | N  | T  | P  | I  | P  | R  | R  | H  | T  | R  | S  | A  | E  | D  | D  | S  | E  | R  | P  | D  | L  |
|-------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| T1 (Thr-171)                  | 1.66 | 1.67 | 1.68 | 1.69 | 1.70 | **1.71** | 1.72 | 1.73 | 1.74 | 1.75 | 1.76 | 1.77 | 1.78 | 1.79 | 1.80 | 1.81 | 1.82 | 1.83 | 1.84 | 1.85 | 1.86 | 1.87 | 1.88 | 1.89 | 1.90 |
| T1                            | 2.41 | 0.29 | 0.5 | 0.68 |
| T2 (Thr-171)                  | 1.1 | 0.16 | 0.38 | 0.28 |
| T3 (Thr-171, Thr-178)         | 2.69 | 1.77 | 0.21 | 0.01 |

*a* Data from Kato et.al.

*b* Product transferase-specific enhancement values are +/- 3 positions from the site of Ser or Thr glycosylation; values greater than 1 suggest enhanced glycosylation.
respectfully), suggesting that both of these transferases will not glycosylate this site. For ppGalNAc T3, the EVP value is 1.77, clearly indicating that T3 should be capable of glycosylating this site.

This demonstrates the uniqueness of the peptide preferences of the individual ppGalNAc T transferases detected from our work and the prediction potential of ISOGlyP. Still, there are several variables that need to be adjusted or added in order to give optimal predictions. These EVP calculations are contingent on a couple of things; first, all amino acids and positions are weighted the same way (all multiplied out) to obtain an EVP value. Whether one site is weighted more than the other still remains to be determined and whether a residue in one position will affect the preference of another in a different position still remains to be studied. Secondly, predictions do not take into account end effects of a peptide. Therefore predictions that are 3-5 residues from the C- or N- terminal of the substrate might be off. Moreover, the EVP calculations weigh the Ser residues equally to the Thr residue, although we know that Thr is generally glycosylated at a higher rate than Ser. Thus, when comparing Ser and Thr sites, the EVP’s for a Ser site should be decreased by at least a factor of 10 (the Gerken Lab is presently in the process of determining these factors for each ppGalNAc T). Lastly, these calculations only apply to peptide sequences (i.e. non-glycosylated substrates) we have not included the effects of neighboring glycosylated sites. These are issues that the Gerken Lab will try to work on and incorporate in the near future.
2.5 References


CHAPTER 3

DETERMINATION OF THE CATALYTIC DOMAIN GLYCOPEPTIDE
PREFERENCES OF SELECTED “PEPTIDE-PREFERRING” AND
“GLYCOPEPTIDE-PREFERRING” ppGalNAc T SUBFAMILY MEMBERS
3.1 Background and Significance

As discussed in the introduction, the ppGalNAc T’s (except for T20) are structurally composed of an N-terminal catalytic domain tethered by a short (flexible) linker to a C-terminal lectin domain. Most studies performed on ppGalNAc T2 and T4 showed that the glycopeptide activities of these transferases was solely responsible by the lectin domain (1,2); as the lectin domain contains the three potential carbohydrate binding sites known as the α, β, γ subdomains. This was believed to be the case for all other transferases, including ppGalNAc T7 and T10 (3-5), as crystal structure of ppGalNAc T10 with a GalNAc-O-Ser (free, non-peptide bound) shows the GalNAc-O-Ser specifically bound to the β subdomain of the lectin domain (4). However, when a ppGalNAc T10 β subdomain lectin domain mutant, Y536A, was used against a glycopeptide substrate (VPS₃T*PPTSPSTPPTSPS) (where T*= GalNAc-O-Thr), glycopeptide activities were observed, where glycosylation occurred at the Ser3 (-1 or one N-terminal position from the T*). It was therefore concluded from the observed activities with the glycopeptides and X-ray crystal structures that both the catalytic and lectin domains of ppGalNAc T10 must each contain GalNAc recognition sites. Further studies on ppGalNAc T10 (with the participation of the Gerken Lab) (6) demonstrated that the removal of the lectin domain did not alter the ability of T10 to glycosylate glycopeptides that contained a acceptor residue that is directly N-terminal (-1 position) from the prior site of glycosylation (T*). This result was influential in determining that the catalytic domain of ppGalNAc T10 may have specific requirements for a glycopeptide substrate and that T10 might have a specific preference site. This was later confirmed by the Gerken Lab (7) where the series of random peptides (PVI-PVIII) and
two new glycopeptide substrates (GPI-GPII) were used to determine the peptide and glycopeptide specificities of ppGalNAc T10. In this work, they observed that for the glycopeptide substrates, ppGalNAc T10 exhibited a single large preference for a GalNAc-O-Ser/Thr at the -1 position (N-terminal) relative to the Ser/Thr acceptor site.

Because the Gerken Lab has previously established a random peptide approach for the determination of the catalytic domain substrate preferences, this transition into determining the glycopeptide specificity would not be as difficult. By using the same random peptide design, a new series of complementary glycopeptide substrates GPI and GPII were created. GPII has the sequence GAGA(X_n)T*(X_n)AGAG, where now a T* = GalNAc-O-Thr will be used in the center, n= 4-5 and the X acceptor residues contain a randomized Ser acceptor. The methods would be very similar as well to the random peptide substrates; where GPII will be partially glycosylated with [3H]-UDP-GalNAc and the glycopeptide product will be isolated via ion-exchange chromatography (Dowex) and size-exclusion chromatography (Sephadex G10). No lectin affinity chromatography is performed since the GPII substrate is already glycosylated thus preventing product separation. The glycosylation site preferences would be determined by sequencing the full reaction mixture where the PTH derivatives were diverted at each cycle to a fraction collector and counted for [3H]-GalNAc incorporation. The other glycopeptide substrate, GPI has a very similar design to the random glycopeptide substrate: GAGA(X_n)T(X_n)AGAG, where the center Thr will be the acceptor residue, n=4-5 and the X residues will contain a randomized GalNAc-O-Ser (or S*). This substrate would be the complementary substrate to GPII. Basically in GPI, the glycosylation position is fixed while the glycosylated S* is randomized and in GPII the glycosylated residue T* is fixed.
while the acceptor positions are randomized. These substrates are expected to give the same result, for a transferase-specific glycosylation site preference. The methods for the product glycopeptide isolation, however, will be different than GPII, because the lectin affinity chromatography is not suitable for product separation. This new approach consisted to using a modified analog of GalNAc, known as GalNAz, where the N-acetyl group in the GalNAc is modified with an azide. Because the ppGalNAc T’s can catalyze UDP-GalNAz, this method would be feasible (8). The glycosylation by UDP-GalNAz would not be radiolabeled (as only trace amounts were used for tracking of product), therefore, the detection method would require the attachment of an alkyne-biotin probe, through the copper(I)–catalyzed cycloaddition approach. This “click-chemistry” approach between the azide and alkyne group has been well established (7-9). The GalNAz-alkyne biotin product can then be isolated via an avidin column and the glycosylation preferences would be determined by regular Edman sequencing (no PTH collection).

Using these substrates, the effects of neighboring prior glycosylation on the catalytic domain the ppGalNAc T’s could be studied in a systematic manner. While this approach was useful in determining that the specific catalytic domain binding site preference for ppGalNAc T10 was strictly at the -1 site (from the site of glycosylation), and showed that none existed for ppGalNAc T1 and T2, the question still remained as to whether there are other ppGalNAc T transferases that contain additional catalytic domain binding sites that are different or similar to ppGalNAc T10.

In this chapter, I will address the role of the catalytic domain of the “glycopeptide-preferring” transferases in subfamily IIa (ppGalNAc T4 and T12) and
subfamily IIb (ppGalNAc T17 and T10) and a number of “peptide-preferring”
transferases (T2, T3, T5, T11, T13, T16 and dT1 (PGANT35A)) in subfamilies Ia, Ib, Ic,
Id and If. Although both glycopeptide-preferring subfamilies readily glycosylate
GalNAc-O-Thr/Ser glycopeptides, they differ in their ability to glycosylate non-
glycosylated peptide substrates (see chapter 2). I have determined using a series of newer
and improved glycopeptide substrates in Table 3.1, that members of these glycopeptide-
preferring subfamilies prefer to add GalNAc in close proximity to a previously
glycosylated residue (i.e. neighboring glycosylation), ranging from the -3, -1 and +1
residues. This suggests that these ppGalNAc T transferases possess a mechanism
whereby the catalytic domain preferentially recognizes an existing GalNAc in
glycopeptide substrates. Comparative studies were also performed on the fly and human
ppGalNAc T ortholog pairs, PGANT7 and ppGalNAc T7 and dT1 (PGANT35A) and
T11, further demonstrating conservation of these mechanistic specificities across species
(although slight differences were observed with between dT1 and T11).

3.2 Materials and Methods

3.2.1 Transferases

As in the previous chapter, ppGalNAc T transferases were obtained from
multiple sources and expression systems (see Table 2.2 in chapter 2 for the list of
transferases and sources). They were expressed as N-terminal truncated and affinity
tagged constructs and provided as either transferase bound to affinity beads or affinity
purified transferase. Again, the same transferases that were supplied from different
sources gave comparable results. ppGalNAc T2 and T3 were expressed from Pichia
*pastoris* and supplied in cell culture media and expressed in High Five insect cells and supplied as a soluble affinity purified transferase (6,10,11). ppGalNAc T4, T7, T12 and T13 were expressed in the baculovirus Sf9 cell system and secreted as N-terminally HIS-tagged catalytic domains bound to Ni-NTA-affinity beads

1 (12). N-terminal polyHis tagged ppGalNAc T4, T5, T10 and T12 were expressed in HEK293 cells and purified by Ni-NTA superflow (Qiagen) nickel affinity chromatography and supplied as soluble transferases 1. These methods were similar to the expression of rat ST6GalI (13). N-terminal polyHis-tagged ppGalNAc T4, T7, T11, T12, T13, T16 and dT1 were expressed in High Five insect cells and purified by Ni-NTA aragose (Invitrogen) and MonoQ 5/50 ion exchange chromatography and supplied as soluble transferase (3,14-16). *Drosophila* PGANT7 was expressed in COS7 cells and supplied as secreted protein in cell culture media (17). Control media reactions with no recombinant protein did not show significant UDP-[3H]-GalNAc utilization.

3.2.2 Reagents and (Glyco)peptide Substrates

The glycopeptide substrates and non-glycosylated peptide control used in this work are listed in Table 3.1. These substrates were custom synthesized from Sussex Research, Ottawa, ON (Canada) and New England Peptides (Gardner, MA). Stock solutions of 50 mg/mL (2.3-2.6 mM) of random (glyco)peptide substrates were prepared by adjusting to pH 7-7.5 with dilute NaOH and/or HCl and lyophilizing from water several times and stored frozen. The materials required for this work such as radiolabeled

---

1 For detailed methods and cDNA constructs, visit the University of Georgia’s Resource for Integrated Glycotechnology, “Glyco-Enzyme Repository” (http://glycoenzymes.crc.uga.edu/).
UDP-[\(^{3}\)H]-GalNAc (C-\(^{3}\)H)-CH\(_2\)-labeled), Dowex 1x8 anion exchange resin, Sephadex G10 beads and column buffers are all described in chapter 2.

3.2.3 Instrumentation

All instrumentation including the Applied Biosystems Procise 494 sequencer, scintillation counter, lyophilizer and spectrophotometer used in this work are described in chapter 2.

3.2.4 Glycosylation Assays with Recombinant ppGalNAc T's and Glycopeptide Substrate

\(GP(T^{*10})C\)-Ser, Non-glycosylated Control \(GP(A10)C\)-Ser and \(GP(T^{*10})C\)-Thr

The general reaction conditions used to glycosylate random (glyco)peptide substrates, \(GP(T^{*10})C\)-Ser, \(GP(A10)C\)-Ser and \(GP(T^{*10})C\)-Thr were: 68 mM sodium cacodylate, pH 6.5, 1.8 mM 2-mercaptoethanol, 10 mM MnCl\(_2\), 50 \(\mu\)M \([\(^{3}\)H]-UDP-GalNAc (\(\sim 6 \times 10^8\) DPM/\(\mu\)mol), 2.3-2.6 mM (5 mg/mL) of substrates \(GP(T^{*10})C\)-Ser, \(GP(A10)C\)-Ser and \(GP(T^{*10})C\)-Thr and up to 400\(\mu\)L of transferase (either soluble or transferase bound to affinity beads).
**Table 3.1** - Random peptide and glycopeptide substrates used to characterize catalytic domain glycopeptide specificity.

<table>
<thead>
<tr>
<th>Peptide/ Glycopeptide (Name)</th>
<th>Designated Color</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP(A10)C-Ser</td>
<td>Yellow</td>
<td>GAGA\text{XXXXX}A_{10}\text{XXXXX}AGAG</td>
</tr>
<tr>
<td>GP(T*10)C-Ser</td>
<td>Green</td>
<td>GAGA\text{XXXXXT}^{*}_{10}\text{XXXXX}AGAG</td>
</tr>
<tr>
<td></td>
<td>(T^{*} = \text{Thr-O-GalNAc})</td>
<td></td>
</tr>
<tr>
<td>GP(T*10)C-Thr</td>
<td>Dark Green</td>
<td>GAGA\text{XXXXXT}^{*}_{10}\text{XXXXX}AGAG</td>
</tr>
<tr>
<td></td>
<td>(X = G, A, R, P, N, E, Y, V &amp; T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(T^{*} = \text{Thr-O-GalNAc})</td>
<td></td>
</tr>
</tbody>
</table>

Note that the acceptor positions are underlined in red and indicate the presence of a Ser or Thr acceptor residue.
Reaction volumes ranged from 75-600 µL and were carried out in 2 mL capped Eppendorf tubes. The reaction mixtures were incubated at 37°C and were left shaken in a TAITEC Microincubator M-36. Time point aliquots of either 15-120 µL were removed at 15, 45, 120 (2 hr), 240 (4 hr) and ~1200 (overnight) minutes and quenched with an equal volume of 250 mM EDTA. Reaction controls were carried out with no peptide/glycopeptide present as well (i.e. Blank), as an internal control for enzymatic activity. In other experiments where low incorporation was expected, only overnight reactions were carried out. Experiments performed with the glycopeptide substrates were performed one substrate at a time or in some cases GP(T*10)C-Ser and GP(A10)C-Ser were performed at the same time with the same transferase concentrations and the same UDP-[3H]-GalNAc stock. The transferases used were ppGalNAc T2, T3, T4, T5, T7, T10, T11 T12, T13, T16, dT1 (PGANT35A) and PGANT7. Detailed information about each transferase reactions will be discussed in the appropriate sections below.

The final transferase concentrations to be used in the reaction assays was determined by several trial and error runs that would typically transfer 10-50% of the total [3H]-GalNAc to the substrate after an overnight incubation (giving a range of ~0.003-0.016 mol of GalNAc transferred per mole of glycopeptide). After the determination of the optimal transferase concentrations, typically 2-4 independent experiments were performed for the GP(T*10)C-Ser substrate. Only 1-2 determinations were performed with the non-glycosylated control, GP(A10)C-Ser and the other substrate GP(T*10)C-Thr. The results are averaged (when applicable) and represented in the final plots. In some reactions, a higher specific activity of UDP-[3H]-GalNAc was used for

---

2 For the explanation on how these optimal conditions were determined, see section 4.2.5 “Determination of Optimal Reaction Conditions and Initial Observations” in chapter 4.
reactions that were to be analyzed by Edman sequencing. The samples were passed through a ~3 mL column of Dowex 1x8 anion exchange resin to remove the unreacted UDP-GalNAc and free UDP. The total UDP-[³H]-GalNAc utilization (including transfer to substrate and hydrolysis) was determined by difference after scintillation counting 1/50 of the sample before and after the dowex column. These data are typically reported as plots of mole fraction of total UDP-[³H]-GalNAc utilized vs. time. To distinguish between [³H]-GalNAc transfer to (glyco)peptide and [³H]-UDP-GalNAc hydrolysis, samples were passed through a Sephadex G10 column for selected time points (typically overnight). The fractions for transfer to (glyco)peptide were monitored and pooled based on [³H]-GalNAc radioactivity and 220 nm and 280 nm absorbance. The pooled fractions were lyophilized from water several times before Edman sequencing as described below.

3.2.5 Determination of Transferase Specific (Glyco)peptide Glycosylation Sites by Edman Sequencing

Samples from the Sephadex G10 isolated glycosylated product were analyzed by Edman sequencing where the PTH derivatives were diverted at each cycle to a ISCO Foxy 200 fraction collector and counted for incorporated radioactivity. It was observed that collecting directly into 7-mL HDPE scintillation vials (Fisher) significantly reduced nonspecific losses of the radiolabeled PTH-derivatives compared with collecting in glass or plastic test tubes. Even though there is commonly sample-to-sample variability in the [³H] content (due to sample losses and different initial UDP-[³H]-GalNAc specific activities) loaded on the sequencer, the observed sites of incorporation were found to be identical between experiments with the same ppGalNAc T transferases. The same holds
true for transferases that were supplied from different sources and batches (such as for ppGalNAc T3, T4, T10 and T12) that were performed a year or more apart. To improve the signal/noise ratio, counting was performed for 5 min/vial and typically repeated 2-3 times, with the resulting DPM (disintegrations per minute) values averaged before plotting. A common noticeable feature of the plots is a sequencing lag in [3H] content following a peak of maximum incorporation. This we believe is due to the lower solubility of the PTH-Ser-O-GalNAc derivative compared with the standard amino acid PTH derivatives, in the organic solvents used to extract the sample filter during the Edman sequencing.

3.3 Results and Discussion

3.3.1 Design and Use of (Glyco)peptide Substrates

The design of GP(T*10)C-Ser and GP(T*10)C-Thr was based on a substrate analog GPII (see introduction) that has been previously used to characterize T* binding at the catalytic domain of ppGalNAc T10 (7), where T10 was shown to glycosylate directly N-terminal of the prior site of GalNAc-O-Thr/Ser glycosylation. This was attributed to the unique GalNAc-O-Ser/Thr recognition site of the catalytic domain of ppGalNAc T10. Prior work of Raman and co-workers (6) has also shown the same glycopeptide specificity observed for ppGalNAc T10.

GP(T*10)C-Ser and GP(T*10)C-Thr, where the C stands for the center placed GalNAc-O-Thr (T*) at position 10, has the general form GAGAXXXXXXT*XXXXXAGAG, where the random X residues flaking the central GalNAc-O-Thr contains a Ser acceptor residue and a Thr acceptor residue, respectively.
Using this substrate, the role of a single T* on the glycosylation of its neighboring residues, X, can be determined via the monitoring of UDP-[\(^3\)H]-GalNAc utilization and Edman amino acid sequencing to determine the Ser-[\(^3\)H]-GalNAc or Thr-[\(^3\)H]-GalNAc content at each X position. Since the acceptor sites X are directly flanking the central T*, the glycosylation of the neighboring sites of GP(T*10)C-Ser and GP(T*10)C-Thr are thus thought to be modulated by T* binding at the catalytic domain and not the T* binding at the lectin domain. Support for this conclusion is found in the recent studies by Lira-Navarrette et al. (18), where their modeling of a glycopeptide bound to the human ppGalNAc T2 lectin domain showed that no sites closer than 5 residues from the T* would be predicted to be glycosylated by the assistance of the lectin domain of the transferase. In their plot of predicted enzymatic activity vs. residue separation, a broad peak of glycosylation was predicted between 7 to 10 residues from the T* (site of prior glycosylation). This aspect will be discussed in more detail in chapter 4. For clarity, results with GP(T*10)C-Ser substrate will be plotted in bright green (circles), while results with GP(T*10)C-Thr will be plotted in dark green (diamonds). A non-glycosylated control peptide for GP(T*10)C-Ser was also used where the center T* is replaced by an Ala residue. This peptide will be called GP(A10)C-Ser and the results will be plotted in yellow (stars).

Since Thr is a much better acceptor than Ser, the substrate GP(T*10)C-Thr was created to increase the radiolabel incorporation and to use against transferases that showed low incorporation with the GP(T*10)C-Ser. Note that GP(T*10)C-Thr has less randomized X amino acids which could also cause higher activities with the transferases.
3.3.2 Glycosylation of the GP(T*10)C-Ser and GP(T*10)C-Thr and GP(A10)C-Ser with ppGalNAc T2, T3, T4, T5, T7, T10, T11, T12, T13, T16, dT1 and PGANT7

The results of the glycosylation with the GP(T*10)C-Ser and GP(T*10)C-Thr and GP(A10)C-Ser substrates by selected ppGalNAc T transferases in family I and II will be presented in a series of plots. The first plots are the representative time course plots where the total $[^3]$H-UDP-GalNAc utilization vs. time are shown, where the total utilization (i.e. counts passing through the Dowex column) represents both the amount of transfer of $[^3]$H-GalNAc to (glyco)peptide and $[^3]$H-GalNAc transfer to water (i.e. UDP-GalNAc hydrolysis product). A commonly observed feature of the transferase reaction time course plots shown in the results section is the plateau in the transfer of GalNAc that is often observed with high transferase activities, even for the least active substrates. This can be attributed to several factors. For the most active substrates (approaching 50% utilization), the plateau in incorporation is most likely due to the depletion of the UDP-$[^3]$H-GalNAc donor. For the less active substrates, the slowing of the rate is due to the sequential loss of the optimal or best acceptor sites as glycosylation of the random (glyco)peptide progresses. Furthermore, because of the half-lives of ppGalNAc T1 and T2 are about 5 and 2.5 days (19), it is unlikely that the transferase inactivation is the source of the plateau with these transferases. However, transferase inactivation cannot be ruled out for the remaining transferases. The second series of plots are the representative Sephadex G10 gel filtration chromatograms showing the separation of the $[^3]$H-GalNAc (glyco)peptide product (in the first eluting peak) from free hydrolyzed $[^3]$H-GalNAc (second eluting peak). The last plot will show the Edman amino acid sequencing, where the $[^3]$H-GalNAc incorporation into the X acceptor residues of the substrate will be
determined, revealing specific sites of glycosylation. All of the Sephadex G10 runs and Edman sequencing was performed on all overnight incubations and time points, unless otherwise specified. All the data presented will be broken down by common substrate classification (i.e. peptide-preferring or glycopeptide-preferring) and by subfamily (i.e. Ia, Ib, Id, IIa and IIb).

3.2.2.1 ppGalNAc T13 in Subfamily Ia

ppGalNAc T1 and T13 are both classified and grouped into subfamily Ia, the first subfamily of the peptide-preferring transferases. They share an 88% sequence identity and in chapter 2, I showed that their catalytic domain peptide specificities are very similar.

Two determinations with ppGalNAc T13 and GP(T*10)C-Ser were performed under the following conditions as described above with either 150 µL of transferase bound to affinity beads or 100 µL of soluble transferase. ppGalNAc T13 was not as active with this substrate as the other transferases, therefore only overnight incubations were performed in order to maximize the [³H] incorporation signal. Figure 3.1 A displays the Sephadex G10 chromatograms with GP(T*10)C-Ser and the Blank (with no substrate present). For the GP(T*10)C-Ser chromatogram, the first eluting weak peak corresponds to [³H]-GalNAc transfer to GP(T*10)C-Ser while the second eluting peak corresponds to [³H]-GalNAc transfer to water (see right panel of Figure 3.1 A). The [³H]-GalNAc transfer to GP(T*10)C-Ser is very low compared to the [³H]-GalNAc transfer to water. This is an indication that this substrate is not very good with ppGalNAc T13.
Figure 3.1- Subfamily Ia ppGalNAcT13 catalytic domain glycopeptide preferences. (A) Left and right pane (with different DPM scales) show the Sephadex gel filtration chromatograms of the overnight incubations of ppGalNAc T13 with GP(T*10)C-Ser (in green) and blank (in purple) demonstrating the elevated GalNAc hydrolysis (fractions 35-45) and low \(^{3}\)H-GalNAc transfer to substrate (fractions 26-33). (B) \(^{3}\)H-GalNAc incorporation plots into the X acceptor residues of GP(T*10)C-Ser showing a small potential enhancement at the -4 position. The right panel shows a representation of the observed specific site of glycosylation marked with the red arrow.
The Edman amino acid sequencing analysis performed on the $[^3\text{H}]-\text{GalNAc}$ incorporation into the acceptor X residues of GP(T*10)C-Ser are shown in Figure 3.1 B. The right panel shows the representation of a potential site of glycosylation observed for GP(T*10)C-Ser (marked with a red arrow) with ppGalNAc T13. These plots reveal the specific sites of glycosylation on the X residues performed by the transferase. In this plot, there are unexpectedly high counts coming off the initial GAGA residues. This is perhaps due to the sample not binding onto the filter and eluting out in the early cycles. There is however, a small enhancement observed at the -4 position, suggesting that ppGalNAc T13 may prefer to glycosylate four (-4) N-terminal positions from a prior site of glycosylation. Although this is not large enhancement, this could be indicative of a glycosylation site preference and needs to be fully confirmed with additional studies with GP(T*10)C-Ser, GP(A10)C-Ser and GP(T*10)C-Thr. Because ppGalNAc T13 is a peptide-prefering transferase, this would make sense as to why there is essentially no glycopeptide activity with this substrate. Perhaps the catalytic domain cannot accommodate a nearby GalNAc-O-Thr (T*) residue, causing the substrate to bind weakly thus enhancing transfer of $[^3\text{H}]-\text{GalNAc}$ to water. Our interpretation of UDP-$[^3\text{H}]-\text{GalNAc}$ hydrolysis will be discussed in the summary section. Due to the lack of ppGalNAc T1, studies using these substrates with T1 could not be performed. However, earlier work with ppGalNAc T1 with GPII showed an enhancement at the -4 position (7). In chapter 4, I will discuss that ppGalNAc T1 and T13 do indeed glycosylate glycopeptide substrates but only when the site of previous glycosylation is 6 (or more) residues away from the acceptor site.
3.2.2.2. ppGalNAc T2 and T16 in Subfamily Ib

In chapter 2, ppGalNAc T2 and T16 were shown to have very similar peptide catalytic domain specificities. Here, their catalytic domain glycopeptide specificities were examined with GP(T*10)C-Ser, GP(A10)C-Ser and GP(T*10)C-Thr.

For ppGalNAc T2, a total of 3 determinations were made with GP(T*10)C-Ser, 2 with GP(A10)C-Ser and one determination with GP(T*10)C-Thr with either 40 or 50 µL of soluble ppGalNAc T2 using the reaction conditions described above. For ppGalNAc T16 a total of 2 determinations for GP(T*10)C-Ser were made with either 150 or 200 µL of soluble T16. Only one reaction was performed with no substrate present. With ppGalNAc T2, time point aliquots were removed at 15, 45, 120 (2 hr), 240 (4 hr) and ~1200 (overnight) minutes and quenched with an equal volume of 250 mM EDTA. These are represented in a time course plot that shows the total UDP-[³H]-GalNAc utilization (including GalNAc transfer to substrate and hydrolysis) and shown in Figure 3.2 A. For ppGalNAc T16, only overnight reactions were performed due to its lower activities with GP(T*10)C-Ser and no reactions with the control was performed. The corresponding Sephadex gel filtration chromatograms of the overnight incubations are shown in Figure 3.2 B for ppGalNAc T2 and Figure 3.2 C for ppGalNAc T16. Again, there are low peaks of incorporation accompanied by very large hydrolysis peaks, indicating that GP(T*10)C-Ser is not a good substrate for both of these transferases. However, when GP(T*10)C-Thr is glycosylated by ppGalNAc T2, the [³H]-GalNAc transfer is enhanced by 5 fold, compared to the GP(T*10)C-Ser, indicating that GP(T*10)C-Thr is not only a much better substrate but confirms that a Thr residue is a better acceptor.
Figure 3.2- Subfamily Ib, ppGalNAc T2 and T16 catalytic domain glycopeptide preferences. (A) Representative time course plot for overnight incubations with ppGalNAc T2 and GP(T*10)C-Ser (green) and GP(A10)C-Ser (yellow) showing the net $[^3]$H-GalNAc utilization. Sephadex gel filtration chromatograms with GP(T*10)C-Ser (green), GP(T*10)CThr (dark green), GP(A10)C-Ser (yellow) and blank (purple) demonstrating the elevated GalNAc hydrolysis (fractions 34-45) and lower $[^3]$H-GalNAc transfer to substrate (fractions 26-32) for ppGalNAc T2 (B) and ppGalNAc T16 (C). For ppGalNAc T2 with GP(T*10)C-Thr (middle), a larger peak of $[^3]$H-GalNAc incorporation is observed at fractions 27-34 and a small peak of hydrolysis at fractions 35-44. (D) $[^3]$H-GalNAc incorporation plots into the X acceptor residues for ppGalNAc T2 with GP(T*10)C-Ser (green) (left) showing a large enhancement at the -5 position and with GP(A10)C-Ser showing a range of incorporation into the X residues. Below GP(T*10)C-Ser (yellow) (middle) is a representation of the specific site of glycosylation marked with a red arrow. (E) $[^3]$H-GalNAc incorporation plots into the X acceptor residues for ppGalNAc T16 with GP(T*10)C-Ser (green) showing no specific peaks of incorporation. (F) $[^3]$H-GalNAc incorporation plots into the X acceptor residues for ppGalNAc T2 with GP(T*10)C-Thr (dark green) showing a large enhancement at the -5 position.
For ppGalNAc T2, the gel filtration plots shows that T2 transfers slightly more $[^3\text{H}]$-GalNAc to the control substrate GP(A10)C-Ser rather than the glycosylated analog GP(T*10)C-Ser, keeping with the peptide-preferring functions. However, a large peak of $[^3\text{H}]$-GalNAc transfer is observed with GP(T*10)C-Thr.

Parts D, E and F of Figure 3.2 shows the incorporation in the acceptor X residues on GP(T*10)C-Ser, GP(A10)C-Ser and GP(T*10)C-Thr for ppGalNAc T2 and on GP(T*10)C-Ser for T16 as determined by Edman amino acid sequencing. For ppGalNAc T2, a large peak of $[^3\text{H}]$-GalNAc incorporation is observed at the -5 position from the T* with both GP(T*10)C-Ser and GP(T*10)C-Thr. The glycosylation of this remote position is thought to be due to the lectin domain interacting with the center T* and not the catalytic domain (the corresponding lectin domain interactions will be further discussed in chapter 4). Interestingly, in both plots, there are low $[^3\text{H}]$-GalNAc peaks of incorporation from the -4 to -1 positions and peaking at the +4 position. The incorporation at the +4 is nevertheless is quite low. The incorporation between the -3 and -1 positions is not thought of to be entirely due to sequencing lag. Thus, ppGalNAc T2 may have limited abilities to glycosylate directly near a prior site of glycosylation similar to what our lab observed with glycopeptide GPII (7). Edman sequencing of the non-glycosylated control, GP(A10)C-Ser with ppGalNAc T2 which does not contain a neighboring T*, gives a broad peak of incorporation with a maximum observed at the -1 and -2 positions from the center Ala. This proves that a neighboring T* (or $\text{S}^*$) is nearly inhibitory and that the catalytic domain of ppGalNAc T2 can only readily glycosylate non-glycosylated “naked” substrates.
A similar pattern is observed for ppGalNAc T16 with GP(T*10)C-Ser in Figure 3.2 E where $[^3]$H-GalNAc incorporation is observed at the -5 position, however, it is hard to tell whether this is a real result or due to sequencing lag since the initial GAGA residues also contain high counts of $[^3]$H-GalNAc perhaps due to the sample not fully sticking onto the filter. There is a very small enhancement at the -1 position, however, since this is only based on one sequence run, no specific sites of glycosylation can be fully determined. More assays and sequencing runs need to be performed.

Both ppGalNAc T2 and T16 belong to the peptide-preferring subfamily Ib of transferases. These results with GP(T*10)C-Ser (which contains a neighboring prior site of glycosylation) and non-glycosylated control GP(A10)C-Ser, shows that ppGalNAc T2 has limited abilities of glycosylate a glycopeptide substrate where the prior site of glycosylation (T*) is no closer than 5 C-terminal positions away. Whether this remote glycopeptide preference is primarily due to the catalytic domain alone or by the assistance of the lectin domain will be discussed in chapter 4.

3.2.2.3 ppGalNAc T3 in Subfamily Ic

ppGalNAc T3 along with T6 belong to subfamily Ic of peptide-preferring transferases. Due to the lack of ppGalNAc T6, no reactions were performed with these substrates and only results for ppGalNAc T3 will be presented here. Chapter 4 will show the comparisons of the lectin domains of both transferases.

A total of three determinations were made with ppGalNAc T3 and GP(T*10)C-Ser, while only one was made with GP(A10)C-Ser with 100 µL or 150 µL of either soluble purified ppGalNAc T3 or transferase bound to affinity beads with the same
reaction conditions as mentioned above. For the first two of the reactions with GP(T*10)C-Ser, overnight incubations were performed while another in another assay both GP(T*10)C-Ser and GP(A10)C-Ser were conducted simultaneously where the five standard reaction time points were removed and quenched with an equal volume of 250 mM EDTA. Figure 3.3 A shows the resulting reaction plot, with the total $[^3H]$-GalNAc utilization being much higher for GP(T*10)C-Ser than the non-glycosylated control GP(A10)C-Ser. The corresponding Sephadex G10 chromatograms of the overnight incubations in Figure 3.3 B show low $[^3H]$-GalNAc transfer to both GP(T*10)C-Ser and GP(A10)C-Ser. Interestingly, we observe large UDP-$[^3H]$-GalNAc hydrolysis for the GP(T*10)C-Ser substrate but none for the control GP(A10)C-Ser. Why this difference occurred is not fully understood but may relate to lectin interactions occurring on GP(T*10)C-Ser.

In Figure 3.3 C, a peak of $[^3H]$-GalNAc incorporation is observed at the +5 position relative to the T*, which is unique for any of the ppGalNAc T transferases characterized thus far. However, because this +5 position is remote from the prior site of glycosylation (T*) this glycosylation could potentially reflect the lectin domain interactions with the T* of GP(T*10)C-Ser. Chapter 4 will discuss how this +5 position correlates with the lectin domain interactions of ppGalNAc T3. The right panel shows the $[^3H]$-GalNAc incorporation with the non-glycosylated control GP(A10)C-Ser showing weak to no glycosylation preferences.

These results suggest that ppGalNAc T3 which is considered a peptide-preferring transferase, also has capability to glycosylate glycopeptide substrates where the prior site of glycosylation (T*) is no closer than 5 N-terminal positions away. Whether these
**Figure 3.3**- Subfamily Ic, ppGalNAc T13 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T3 and GP(T*10)C-Ser (green) and GP(A10)C-Ser (yellow) showing the net [\(^3\)H]-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green) and GP(A10)C-Ser (yellow) demonstrating the elevated GalNAc hydrolysis (fractions 36-45) and low [\(^3\)H]-GalNAc transfer to substrate (fractions 27-34) for ppGalNAc T3. (C) [\(^3\)H]-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T3 with GP(T*10)C-Ser (green) (left) showing a large enhancement at the +5 position and with GP(A10)C-Ser (yellow) (right) showing low incorporation into the X residues. The bottom left panel shows a representation of the peak site of glycosylation marked with the red arrow.
glycopeptide preferences are determined by the assistance of the lectin domain will be discussed in chapter 4. It is interesting though, that the Edman amino acid sequencing with the GP(A10)C-Ser substrate, gives no significant peaks of incorporation.

3.2.2.4 *ppGalNAc T5 in Subfamily Id*

*ppGalNAc T5* belongs to the subfamily Id of peptide-preferring transferases and is the only member in its group. To study its catalytic glycopeptide preferences, a total of 4 determinations were made with GP(T*10)C-Ser, 2 determinations performed with GP(A10)C-Ser and one determination with GP(T*10)C-Thr with 100 µL or 150 µL of soluble affinity purified *ppGalNAc T5* with the same reaction conditions as mentioned above. For these reactions, the five standard reaction time points were removed and quenched with an equal volume of 250 mM EDTA. Representative time course plots are shown in Figure 3.4 A, displaying the total $[^{3}H]$-GalNAc utilization. Overall, the highest apparent activity is observed with both glycopeptide substrates, GP(T*10)C-Ser and GP(T*10)C-Thr and lower activity is observed with GP(A10)C-Ser. The corresponding Sephadex G10 chromatographs of the overnight incubations are shown in Figure 3.4 B. The chromatograms show the same trend of low $[^{3}H]$-GalNAc transfer to substrate and high hydrolysis for GP(T*10)C-Ser and GP(A10)C-Ser. However, for GP(T*10)C-Thr, the chromatogram shows a higher amount of $[^{3}H]$-GalNAc transfer to substrate (nearly 5 fold higher) than hydrolysis. This indicates that GP(T*10)C-Thr is a much better substrate for *ppGalNAc T5*, presumably because a Thr acceptor is much more preferred than Ser.
Figure 3.4- Subfamily Id ppGalNAc T5 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T5 and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) showing the net $[^{3}\text{H}]$-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and blank (purple) demonstrating the elevated GalNAc hydrolysis (fractions 36-45) and low $[^{3}\text{H}]$-GalNAc transfer to substrate (fractions 27-34) for ppGalNAc T5. For GP(T*10)C-Thr (right), a larger peak of $[^{3}\text{H}]$-GalNAc incorporation is observed at fractions 27-34 and a small peak of hydrolysis at fractions 35-44. (C) $[^{3}\text{H}]$-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T5 with GP(T*10)C-Ser (green) (left) and GP(T*10)C-Thr (dark green) (right) showing a large enhancement at the -5 position and with GP(A10)C-Ser (yellow) (middle) showing low incorporation into the X residues. The bottom left and right panels show a representation of the peak site of glycosylation marked with the red arrow.
The Edman amino acid sequencing plots in Figure 3.4 C show that there is a -5 glycosylation site preference relative to the T* for ppGalNAc T5. In fact, this is seen for both GP(T*10)C-Ser and GP(T*10)C-Thr with the latter having a 10 fold enhancement in signal. For the GP(A10)C-Ser control, an enhancement at the -1 and -2 N-terminal positions from the Ala are observed although the signal is much lower than the other plots.

These results suggest that ppGalNAc T5, a peptide-preferring transferase, also has capabilities of glycosylating glycopeptide substrates where the prior site of glycosylation (T*) is no closer than 5 C-terminal positions. Whether these glycopeptide preferences are directed by the catalytic domain alone or by the assistance of the lectin domain will be discussed in chapter 4.

3.2.2.5 ppGalNAc T11 and dT1 in Subfamily If

ppGalNAc T11 is part of peptide-preferring subfamily If and dT1 (PGANT35A), is the fly ortholog to T11 also belongs to the same subfamily in the ppGalNAc T interspecies phylogenetic tree (see chapter 2 for the phylogenetic tree). In chapter 2, I showed that both of the transferases had nearly identical catalytic domain peptide specificities. Here, I will study their glycopeptide catalytic domain specificity, specifically when a prior site of glycosylation is nearby an acceptor site and show that there may be differences between the human ppGalNAc T and fly PGANT.

For ppGalNAc T11, a total of two determinations were made with GP(T*10)C-Ser, and one with GP(A10)C-Ser and GP(T*10)C-Thr using either 200 µL - 400 µL of soluble ppGalNAc T11 with the same reaction conditions as mentioned above. Only one
Figure 3.5- Subfamily If ppGalNAc T11 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T11 and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) showing the net \[^3\text{H}\]-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) demonstrating the GalNAc hydrolysis (fractions 35-40) and \[^3\text{H}\]-GalNAc transfer to substrate (fractions 25-30) for ppGalNAc T11. (C) \[^3\text{H}\]-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T11 with GP(T*10)C-Ser (green) and GP(T*10)C-Thr showing a large enhancement at the -5 and -4 positions. The bottom panels show a representation of the observed specific site of glycosylation marked with the red arrow. ND stands for not determined.
Figure 3.6- Subfamily If dT1 (PGANT35A) catalytic domain glycopeptide preferences. (A) Representative time course plot with dT1 (PGANT35A) and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) showing the net $[^3]$H-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) demonstrating hydrolysis (fractions 35-40) and $[^3]$H-GalNAc transfer to substrate (fractions 26-32) for dT1 (PGANT35A). (C) $[^3]$H-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of dT1 (PGANT35A) with GP(T*10)C-Ser (green) and GP(T*10)C-Thr showing a large enhancement at the -5, -4 and -3 positions. The bottom panels show a representation of the observed specific site of glycosylation marked with the red arrow. ND stands for not determined.
reaction with GP(T*10)C-Ser was performed as an overnight incubation while the others has the five standard time points removed and quenched with an equal volume of 250 mM EDTA. This was performed identically with dT1, with 200 µL - 400 µL of soluble dT1 (PGANT35A). Representative time course plots are shown in Figure 3.5 A and Figure 3.6 A, displaying the total [³H]-GalNAc utilization. Overall for both transferases, the highest activity is observed with GP(T*10)C-Thr while lower activities are observed with GP(T*10)C-Ser and GP(A10)C-Ser. The corresponding Sephadex G10 chromatographs of the overnight incubations are shown in Figure 3.5 B and 3.6 B. The chromatograms show the same trend of low [³H]-GalNAc transfer to substrate and high hydrolysis for GP(T*10)C-Ser and GP(A10)C-Ser. In contrast, for GP(T*10)C-Thr, there are larger amounts of [³H]-GalNAc transfer to substrate while [³H]-GalNAc hydrolysis is decreased, in the case of ppGalNAc T11, both of these peaks are equal (Figure 3.5 B).

The Edman amino acid sequencing plots for GP(T*10)C-Ser and GP(T*10)C-Thr are shown in Figures 3.5 C for T11 and Figure 3.6 C for dT1. Due to low counts on the GP(A10)C-Ser, no sequencing was performed on these samples. For ppGalNAc T11, there is a peak enhancement observed at the -5 position with GP(T*10)C-Ser and a peak enhancement observed at the -5 (although modest) and -4 positions with GP(T*10)C-Thr. There is slight variability between both substrates, perhaps the activity with the GP(T*10)C-Ser is too low that we cannot detect the enhancement at the -4 position. This makes sense because the Edman sequencing plots with GP(T*10)C-Thr does show a modest -5 position enhancement although the enhancement at the -4 is larger. Based on the results from these two substrates, this suggests that ppGalNAc T11 prefers to glycosylate four N-terminal positions (-4) to five N-terminal positions (-5) from a prior
site of glycosylation. For $dT_1$ (PGANT35A), the peak glycosylation site preference observed for both substrates, GP(T*10)C-Ser and GP(T*10)C-Thr is at the -5 and -4 positions, suggesting that $dT_1$ prefers to equally glycosylate five (-5) and four (-4) N-terminal positions from a prior site of glycosylation. Whether these -5 N-terminal preferences correspond to lectin domain interactions will be explored in chapter 4. Interestingly, for the GP(T*10)C-Thr plot with $dT_1$, there is an equal incorporation observed at the -4 and -3 positions as well, it could very well be due to sequencing lag, but in most cases, it tends to gradually decrease which is not what is seen here. It is more believable that the -4 enhancement is real since the amount of radiolabel is equal to the -5 enhancement. Whether the -5 enhancement observed for T11 and $dT_1$ is due to lectin domain interactions will be discussed in chapter 4. Therefore, the catalytic domain of both transferases glycosylates the -4 position, but the differences in glycosylation patterns indicate that perhaps there are some slight differences potentially in the lectin domain glycopeptide glycosylation activities of ppGalNAc T11 and $dT_1$ (PGANT35A). ppGalNAc T11 modestly glycosylates the -5 position while $dT_1$ equally glycosylates -5 position (compared to the -4 position) from the site of glycosylation. If this is proven to be true, this work represents the first difference observed between T11 and $dT_1$, where their catalytic domain glycopeptide preferences are similar but their lectin domains slightly differ in their glycosylation abilities. These results could potentially give some insight as to why ppGalNAc T11 cannot replace $dT_1$ in in vivo experiments.
3.2.2.6 ppGalNAc T4 and T12 in Subfamily IIa

Subfamily IIa is one of the two subfamilies of glycopeptide-preferring transferases and includes ppGalNAc T4 and T12. I have demonstrated in chapter 2 that these transferases are also capable of glycosylating “naked” peptide substrates and have very similar peptide catalytic domain specificities. Here, I will test their ability to glycosylate residues neighboring a prior site of glycosylation (T*).

For ppGalNAc T4, a total of 4 determinations were made with GP(T*10)C-Ser, 2 determinations performed with GP(A10)C-Ser and one determination with GP(T*10)C-Thr with 100 µL or 150 µL of soluble ppGalNAc T4 with the same reaction conditions as mentioned above. For ppGalNAc T12, a total of 2 determinations were performed with GP(T*10)C-Ser and one determination was made for GP(A10)C-Ser with 60 µL or 150 µL of soluble ppGalNAc T12 with the same reaction conditions as mentioned above. For these reactions, the five standard reaction time points were removed and quenched with an equal volume of 250 mM EDTA. Representative time course plots are shown in Figures 3.7 A and 3.8 A, displaying the total $[^3]H$-GalNAc utilization for ppGalNAc T4 and T12 respectively. The reaction plots for both ppGalNAc T4 and T12 show that the highest activities are with glycopeptide substrates, GP(T*10)C-Ser and GP(T*10)C-Thr, while activities with the control peptide remain low. The corresponding Sephadex G10 plots of the overnight incubations, in Figures 3.7 B and 3.8 B, show a higher amount of $[^3]H$-GalNAc transfer to substrate than observed for the peptide-preferring transferases, and even though $[^3]H$-GalNAc hydrolysis is still high. $[^3]H$-GalNAc hydrolysis is visibly reduced when GP(T*10)C-Thr was used with ppGalNAc.
**Figure 3.7** - Subfamily IIa ppGalNAc T4 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T4 and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) showing the net $[^3]$H-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) demonstrating the decreased GalNAc hydrolysis (fractions 36-42) and increased $[^3]$H-GalNAc transfer to substrate (fractions 25-32) for ppGalNAc T4. (C) $[^3]$H-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T4 with GP(T*10)C-Ser (green) and GP(T*10)C-Thr (dark green) showing a large enhancement at the +1 and +5 positions and with GP(A10)C-Ser (yellow) showing low incorporation into the X residues. The bottom panels show a representation of the observed specific site of glycosylation marked with the red arrow.
Figure 3.8- Subfamily IIa ppGalNAc T12 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T12 and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and blank (purple) showing the net $[^{3}H]$-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and blank (purple) demonstrating the decreased GalNAc hydrolysis (fractions 35-43) and elevated $[^{3}H]$-GalNAc transfer to substrate (fractions 25-33) for ppGalNAc T12. (C) $[^{3}H]$-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T12 with GP(T*10)C-Ser (green) showing a large enhancement at the -3 position and with GP(A10)C-Ser (yellow) showing low incorporation into the X residues. The bottom left panel shows a representation of the observed specific site of glycosyilation, marked with the red arrow.
T4 again suggesting that a Thr acceptor is much better than Ser. Another interesting observation is that there is almost no hydrolysis in the Sephadex G10 chromatogram with ppGalNAc T12 and the non-glycosylated control substrate GP(A10)C-Ser. This was also observed with ppGalNAc T3 and the reason as to why this is observed is still unknown.

Striking differences between both transferases are observed in the[^3H]-GalNAc incorporation plots, in Figures 3.7 C left and right panels and 3.8 C left panel which show that both transferases posses different preferred sites of glycosylation. For ppGalNAc T4, there is nearly equal[^3H]-GalNAc incorporation at the +1 and +5 positions (with both substrates GP(T*10)C-Ser and GP(T*10)C-Thr) suggesting that T4 prefers to glycosylate a Ser and Thr residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a prior site of glycosylation (T*). It is likely, however, that the +5 preference may be due to the lectin domain interactions, which is consistent with findings that will be discussed in chapter 4. These sites of glycosylation are consistent with ppGalNAc T4’s previously reported (2) high activity against glycopeptides A2 (PT*2TDST7TPAPTTKK) and A4 (PTTDST*6T7PAPTTKK), where both of these substrates possess a +1 and +5 acceptor site from a T*. In A2 the Thr-7 glycosylation site is five C-terminal positions (+5) from the T*2 and in A4 the Thr-7 glycosylation site is one C-terminal (+1) position from the T*6. ppGalNAc T12, on the other hand, shows a maximum[^3H]-GalNAc incorporation at the -3 position relative to the T* indicating that T12 prefers to glycosylate a Ser (or Thr) residue three N-terminal residues from a prior site of glycosylation (T*). This difference is surprising, since both ppGalNAc T4 and T12 belong to the same subfamily of ppGalNAc T transferases and share a 72% sequence identity. Earlier studies in our lab against ppGalNAc T12 with GPIII show the same
glycosylation site preference (unpublished). In contrast to ppGalNAc T4, ppGalNAc T12 does not glycosylate the +5 position of GP(T*10)C-Ser, which is consistent with what I have observed with the lectin domain probing substrates in chapter 4.

Edman sequencing of GP(A10)C-Ser (Figure 3.7 C middle panel and 3.8 C right panel) for ppGalNAc T4 and T12 reactions reveal no peaks of [3H]-GalNAc incorporation, meaning that they have limited abilities to glycosylate the GP(A10)C-Ser substrate. (Note however, that ppGalNAc T4 and T12 have activities with the random peptide substrates). This explains the previously observed glycopeptide activities of these transferases, as high incorporation is only observed with the glycopeptide substrates. This means that these transferases prefer glycopeptides that have a T* within +/- 1-4 residues from the acceptor site. Therefore, ppGalNAc T4 and T12 must serve as “filling in” transferases, glycosylating +1, +5 and -3 sites from a prior site of glycosylation, respectively. These specificities have been unknown until this study.

3.2.2.7 ppGalNAc T7, T10 and PGANT7 in Subfamily IIb

ppGalNAc T7 and T10 both belong to subfamily IIb of glycopeptide preferring transferases and PGANT7 the fly orthologue to the human ppGalNAc T7, is also classified under the same subfamily with other mammalian GALNT genes (20) see Figure 3.9. In chapter 2, I showed that ppGalNAc T7 does not have any significant peptide catalytic domain specificities, as previously determined for ppGalNAc T10 by our lab (11). Using these glycopeptide substrates in Table 3.1, I have determined all of these three transferases are active against these glycopeptide substrate and unlike ppGalNAc T4 and T12, have identical catalytic domain glycopeptide specificities.
To study the catalytic domain glycopeptide preferences of this family, a total of 3
determinations were made with GP(T*10)C-Ser and one determination with GP(A10)C-
Ser with 100 µL or 250 µL of either soluble ppGalNAc T7 or transferase bound to
affinity beads. For ppGalNAc T10, a total of 2 determinations were made with
GP(T*10)C-Ser and one determination with GP(A10)C-Ser and GP(T*10)C-Thr with
100 µL or 150 µL of soluble T10. For PGANT7, only one determination was made with
GP(T*10)C-Ser with 500 µL of transferase in cell culture media bound to affinity beads.
All reactions were performed under the same reaction conditions as described above and
all standard time points were removed and quenched with an equal volume of EDTA.
The resulting time course plots and Sephadex G10 gel filtration chromatograms are
shown in Figures 3.10 A-D and 3.11 A-B. For all three transferases, the time course plots
show that the highest activity is with the glycopeptide substrates and not the control. The
 corresponding Sephadex G10 chromatograms of the overnight incubations show that
there is an appreciable higher amount of [\(^3\)H]-GalNAc transfer to substrate similar to the
subfamily IIa members, and varied amounts of [\(^3\)H]-GalNAc hydrolysis within
transferase and glycopeptide substrates.
Figure 3.9- Multi-species ppGalNAc T subfamily IIb phylogenetic tree from humans, chicken, fish, frog, worm (C. elegans) and T. Gondii. Taken and modified from Bennett et al. (20). Red arrows are pointing to the human T7 and fly PGANT7.
Figure 3.10- Subfamily IIb ppGalNAc T7 and PGANT7 catalytic domain glycopeptide preferences. Representative time course plot with ppGalNAc T7 (A) and PGANT7 (B) with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and blank (purple) showing the net $[^3]$H-GalNAc utilization. Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and blank (purple) demonstrating the decreased GalNAc hydrolysis (fractions 35-43) and elevated $[^3]$H-GalNAc transfer to substrate (fractions 25-33) for ppGalNAc T7 (C) and PGANT7 (D). $[^3]$H-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T7 (E) and PGANT7 (F) with GP(T*10)C-Ser (green) showing a large enhancement at the -1 position and with GP(A10)C-Ser (yellow) showing low incorporation into the X residues. The bottom panels show a representation of the observed specific site of glycosylation marked with the red arrow.
Figure 3.11- Subfamily IIb ppGalNAc T10 catalytic domain glycopeptide preferences. (A) Representative time course plot with ppGalNAc T10 and GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) showing the net $[^{3}H]$-GalNAc utilization. (B) Corresponding Sephadex gel filtration chromatograms of the overnight incubations with GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow), GP(T*10)C-Thr (dark green) and blank (purple) demonstrating the decreased GalNAc hydrolysis (fractions 35-43) and elevated $[^{3}H]$-GalNAc transfer to substrate (fractions 25-33) for ppGalNAc T10. (C) $[^{3}H]$-GalNAc incorporation plots into the X acceptor residues for chromatographed overnight incubations of ppGalNAc T10 with GP(T*10)C-Ser (green) and GP(T*10)C-Thr (dark green) showing a large enhancement at the -1 position and with GP(A10)C-Ser (yellow) showing low incorporation into the X residues. The bottom left panel shows a representation of the observed specific site of glycosylation marked with the red arrow.
The plots of $[^3]$H-GalNAc incorporation in Figures 3.10 E and F and 3.11 C left and right panels show for all transferases, the same conserved -1 glycosylation site preferences for GP(T*10)C-Ser (with all three transferases) and GP(T*10)C-Thr (for ppGalNAc T10). This suggests that ppGalNAc T7, T10 and PGANT7 prefer to glycosylate Ser and Thr residues that are exactly one N-terminal position from prior site of glycosylation (T*). It is noteworthy that these transferases do not exhibit the -5 and +5 preferences suggesting that there is little or no lectin domain involvement in glycosylating GP(T*10)C-Ser and GP(T*10)C-Thr. These results are interesting, given that ppGalNAc T7 and T10 share a 48% sequence identity and ppGalNAc T7 and PGANT7 share a 41% sequence identity which is in complete contrast to what was observed for subfamily IIa members, where ppGalNAc T4 and T12 share a 72% sequence identity. Previous studies in our lab have shown that ppGalNAc T10 will glycosylate a Ser (or Thr) residue directly N-terminal (-1) from the T* (7) and the same was also shown in studies by Pratt et al. (2) where their large glycopeptide library showed that the best substrates for ppGalNAc T7 and T10 were the ones that had a Thr acceptor one N-terminal position from the site of glycosylation. Other studies on ppGalNAc T10 have shown that even with the removal of the lectin domain, there is no effect on the glycosylation at the -1 position from the prior site of glycosylation (6).

Comparisons with the Edman sequencing non-glycosylated control, GP(A10)C-Ser (Figure 3.10 E middle panel and 3.11 C middle panel) products with ppGalNAc T7 and T10 confirm the glycopeptide requiring activities of these transferases as there is virtually no $[^3]$H-GalNAc incorporation observed. This means that these transferases
have identical glycopeptide specificities and function as strictly “filling in” transferases, glycosylating directly N-terminal of an existing T* (or S*).

3.4 Summary and Conclusions

With the two glycopeptide substrates GP(T*10)C-Ser and GP(T*10)C-Thr in Table 3.1, I have been able to systematically investigate the role of the catalytic domain of several peptide-preferring and glycopeptide-preferring transferases in glycosylating glycopeptide substrates in terms of neighboring (1-5 residues) prior glycosylation. I have found that all five of the glycopeptide-preferring family II transferases studied have elevated activities towards GP(T*10)C-Ser and GP(T*10)C-Thr (see summary of glycopeptide activities in Figure 3.12) and that each isoform glycosylates only one or two unique sites on these substrates (see glycosylation preferences summary in Figure 3.13). This suggests that the binding of the T* (GalNAc-O-Thr) residue of GP(T*10)C-Ser and GP(T*10)C-Thr at specific sites in the catalytic domain directs subsequent neighboring glycosylation by these transferases 1-5 residues away. In contrast, the peptide-preferring family I transferases (in subfamilies Ia, Ib, Ic, Id and If) all have lower activities towards the GP(T*10)C-Ser glycopeptide substrate and much higher activities with GP(T*10)C-Thr glycopeptide substrates (Figure 3.12). They furthermore do not specifically glycosylate directly neighboring sites (Figure 3.13), except for ppGalNAc T11 and dT1. This suggests that these ppGalNAc T transferases have no GalNAc-O-Thr/Ser binding sites in their catalytic domains.

Apart from the simple peptide binding in the catalytic domain (in Figure 3.14 B) that must occur for all transferases, these results demonstrate that for the glycopeptide-preferring transferases, there is an additional mode for glycopeptide substrate recognition
Figure 3.12- Efficiency of GalNAc transfer to acceptor glycopeptide varies with ppGalNAc T families. Bars represent the percent (i.e. efficiency) of GalNAc transfer to the indicated (glyco)peptide substrates: GP(T*10)C-Ser (green), GP(A10)C-Ser (yellow) and GP(T*10)C-Thr (dark green). Left grouping of bars represents the “peptide-preferring” transferases (ppGalNAc T13, T2, T16, T3, T5, T11 and dT1) and the right grouping represents the “glycopeptide-preferring” transferases (T4, T12, T7, T10 and PGANT7). Values shown are obtained from 1 to 3 G10 filtration runs of standard overnight incubations.
**Figure 3.13** - Correlation of glycopeptide specificities with the ppGalNAc T phylogenetic family tree. To the right of the ppGalNAc T subfamily phylogenetic tree (20) are the glycopeptide preferences obtained from this work. Glycopeptide substrate cartoons shown in green represent GP(T*10)C-Ser and GP(T*10)C-Thr where the T* represents the position of the initial GalNAc-O-Ser/Thr, while the arrows indicate the position(s) glycosylated by the indicated transferase isoform. The grey arrows represent the modest glycosylation site preferences. The arrows in parenthesis represent the proposed +/- 5 lectin-mediated glycosylation site preferences, which will be discussed in chapter 4. Transferases whose preferences have not been determined against these substrates are left as a blank space.
A. Catalytic Domain Glycopeptide Recognition
"Catalytic Domain-Directed"

B. Catalytic Domain Peptide Recognition

**Figure 3.14**- Schematic representations of catalytic domain-directed glycopeptide activities and peptide activities of the ppGalNAc T’s. (A) Catalytic domain glycopeptide recognition by the mixed (glyco)peptide-preferring subfamily IIa transferases (ppGalNAc T4 and T12) and the strict glycopeptide-preferring subfamily IIb transferases (ppGalNAc T7 and T10). (B) Peptide substrate binding to the catalytic domain directing glycosylation by family I (ppGalNAc T1-T6, T11-T14 and T16) and subfamily IIa members (ppGalNAc T4 and T12). The large grey oval represents the catalytic domain while the smaller right oval represents the lectin domain (note that the lectin domain could also be on the left). The (glyco)peptide acceptor is represented with a thick black line. Taken from (16).
by ppGalNAc T4, T7, T10 and T12 that would be the direct binding of the glycopeptide GalNAc within the catalytic domain, targeting specific neighboring glycosylation, see Figure 3.14 A. This glycopeptide activity would not be driven by the lectin domain as is discussed in chapter 4. We would like to refer to these activities as catalytic domain-directed. A so-called lectin-domain assisted property will be discussed in chapter 4.

Another issue that needs to be addressed is the observed overlapping peptide/glycopeptide substrate preferences for the different transferase isoforms and we have suggested to rename or reclassify family II transferases. Since ppGalNAc T4 and T12 have been shown to active against a number of non-glycosylated peptide substrates and therefore not strictly glycopeptide-preferring transferases, we suggest that family IIa members be termed mixed (glyco)peptide-preferring isoforms (abbreviated as mixed (G)P-preferring). Since subfamily IIb members, ppGalNAc T7 and T10 are poorly active towards naked peptide substrates, we suggest calling them strict glycopeptide-preferring isoforms (abbreviated as strict GP-preferring).

**UDP-GalNAc-Hydrolysis**

During our studies of the ppGalNAc T’s, I observed variable non-productive hydrolysis of the UDP-[³H]-GalNAc donor (i.e. transfer to water) vs. transfer to peptide/glycopeptide substrate. This varies among transferase isoform and is also dependent on peptide/glycopeptide substrate concentrations and the concentrations of UDP-[³H]-GalNAc donor. Presently, we do not know what factors are involved but there is clearly competition with transfer of GalNAc to water and transfer to substrate.
Interestingly, the peptide-preferring and glycopeptide-preferring transferases behave very differently against our peptide/glycopeptide substrates, displaying different degrees of hydrolysis. For the peptide-preferring transferases, hydrolysis is usually increased against the glycopeptides (i.e. GP(T*10)C-Ser/Thr whose glycosylation sites are neighboring of the acceptor), compared to peptide substrates see Figures 3.1-3.6. In addition, transfer to glycopeptide is significantly decreased. This suggests that the catalytic domain of these family I transferases cannot accommodate a GalNAc-O-Ser/Thr (except for ppGalNAc T11 and dT1 which have a -4 glycosylation site preference). For the family II glycopeptide preferring transferases T4, T7, T10, T12 and PGANT7, significant rates of hydrolysis are observed against most peptide and glycopeptide substrates although greater transfer to glycopeptide is observed for GP(T*10)C-Ser and GP(T*10)C-Thr (Figures 3.7-3.8 and 3.10-3.11). As no structural studies on the ppGalNAc T’s have revealed a specific GalNAc-O-Thr/Ser binding site on the catalytic domain active site cleft, it is likely that these glycopeptides bind presumably much weaker because of a more open peptide binding cleft. This may explain the weak peptide binding/activity of the transferases. Thus, we believe these results suggest that the rate of hydrolysis depends on how well the substrate binds to the catalytic domain binding cleft. This would be modulated by both T* binding at the catalytic domain and at the lectin domain (chapter 4) which would serve to help align the acceptor peptide into the catalytic domain and increase transferase efficiency.

For the first time, I have identified multiple catalytic domain GalNAc binding

---

3 In the process of writing my thesis, an X-ray crystal structure of ppGalNAc T4 with a di-glycopeptide bound was given to us in confidentiality by the Hurtado-Guerrero Lab. The structure has not been fully analyzed and has not been published.
sites for the “glycopeptide-preferring” transferases in subfamilies IIa and IIb that likely account for their glycopeptide activities. These transferases are commonly called filling-in transferases and in this work I have identified site preferences for their respective activities. I have also demonstrated the partial conservation of specificity between the fly and human orthologs PGANT7 and ppGalNAc T7, which display identical catalytic domain-directed glycopeptide specificities. For the other pair of transferase orthologs, dT1 (PGANT35A) and ppGalNAc T11, small differences were observed in their site preferences, revealing for the first time a potential difference between both transferase orthologs. With this work, we now have a better understanding of the roles of each ppGalNAc T in this large family.

3.5 References


4. Kubota, T., Shiba, T., Sugioka, S., Furukawa, S., Sawaki, H., Kato, R.,
Wakatsuki, S., and Narimatsu, H. (2006) Structural Basis of Carbohydrate
Transfer Activity by Human UDP-GalNAc: Polypeptide a-N-
Acetylgalactosaminytransferase (pp-GalNAc-T10). *Journal of Molecular Biology*
359, 708-727

5. Cheng, L., Tachibana, K., Zhang, Y., Guo, J., Kahori, T. K., Kameyama, A.,
Wang, H., Hiruma, T., Iwasaki, H., Togayachi, A., Kudo, T., and Narimatsu, H.
(2002) Characterization of a novel human UDP-GalNAc transferase, pp-GalNAc-
T10. *FEBS Letters* 531, 115-121

alpha -N-Acetylgalactosaminytransferase Function in Concert to Direct
Glycosylation Site Selection. *J.Biol.Chem.* 283, 22942-22951

GalNAc Transferase 10 (ppGalNAc T10), Involved in Mucin-type O-
Glycosylation, Has a Unique GalNAc-O-Ser/Thr-binding Site in Its Catalytic
Domain Not Found in ppGalNAc T1 or T2. *J.Biol.Chem.* 284, 20387-20397

approach toward proteomic analysis of mucin-type O-linked glycosylation.
*Proceedings of the National Academy of Sciences of the United States of America*
100, 14846-14851

Stepwise Huisgen Cycloaddition Process: Copper-2-Catalyzed Regioselective

catalytic and lectin domains of the human UDP-GalNAc:polypepide a-N-
acetylgalactosaminytransferase-2. *Journal of Biological Chemistry* 281, 8613-
8619

11. Gerken, T. A., Jamison, O., Perrine, C. L., Collete, J., Moinova, H., Ravi, L.,
Paradigms for the Initiation of Mucin Type Protein O-Glycosylation by the
Polypeptide GalNAc Trasnfereases (ppGalNAc T) Family of Glycosyltransferases.
*Journal of Biological Chemistry* 286, 14493-14507

line glycoengineering approach provides baculovirus-inducible glycogene
expression and increases human type glycosylation effiency. *Journal of
Biotechnology* 182-183, 19-29

13. Barb, A. W., Meng, L., Gao, Z., Johnson, R. W., Moremen, K. W., and
Motion on Enzymatic Siaylation. *Biochemistry* 51, 4618-4626

14. Hassan, H., Reis, C. A., Bennett, E. P., Mirgorodskaya, E., Roepstorff, P.,
Hollingsworth, M. A., Burchell, J., Taylor-Papadimitriou, J., and Clausen, H.
(2000) The lectin domain of UDP-N-acetyl-D-galactosamine: polypeptide N-
acetylgalactosaminytransferase-T4 directs its glycopeptide specificities. *Journal
of Biological Chemistry* 275, 38197-38205


DETERMINATION OF THE GLYCOPEPTIDE PREFERENCES MEDIATED BY
THE LECTIN DOMAINS OF THE “PEPTIDE-PREFERRING” AND
“GLYCOPEPTIDE-PREFERRING” ppGalNAc T SUBFAMILY MEMBERS
4.1 Background and Significance

The carbohydrate binding lectin domain of the ppGalNAc T transferases is a unique feature that most enzymes do not possess. As mentioned throughout my thesis, the lectin domain is composed of three homologous repeat sequences known as the α, β and γ subdomains, which contain the sugar binding motifs. These simple sugar binding motifs are known as the CLD and QXW motifs (where X can be any amino acid). Much of the earlier work on the lectin domains of ppGalNAc T1, T2 and T4 showed that mutations on the CLD and QXW motifs of the lectin α, β and γ subdomains compromised the glycopeptide activity of the transferases and that the lectin domain activity was inhibited by the addition of free GalNAc (2-4,8,9). From these studies, it was concluded that function of the lectin domain was to exclusively interact with a GalNAc on glycosylated substrates and thus the lectin domain would help align the glycopeptide substrate into the catalytic domain, increasing transferase efficiency (3,11). Other studies that followed focused on looking at the binding properties, specifically whether the lectin domains have any glycopeptide sequence specificity and if there was any directionality preferences toward glycosylated substrates. Studies by Pedersen et al. (14) suggested that the lectin domains of ppGalNAc T2 and T3 may recognize glycopeptide sequence context. A more detailed study on ppGalNAc T2 by Raman et al. (11) showed that the lectin domain of ppGalNAc T2 was involved in directing glycopeptide site selection in a direction-dependent manner, where the placement of a N- or C- terminal prior GalNAc glycosylated residue (T*) influenced the subsequent site of glycosylation.

The latter study with ppGalNAc T2 was a collaborative effort between the Tabak and Gerken Labs. A series of MUC5AC (glyco)peptides, in Figure 4.1 A, were tested
against a WT ppGalNAc T2 and a lectin domain deleted ppGalNAc T2 (catalytic domain only) and differences were found between the glycosylation site preferences, see Figure 4.1 B and C, respectively. In the presence of the lectin domain of ppGalNAc T2 (WT), the sites of glycosylation shifted depending on where the prior site of glycosylation was located on the substrate, see Figure 4.1 B. The biggest differences were seen between MUC5AC-3 and MUCAC-13, where glycosylation sites were observed at the C- and N-terminal positions of the substrate, specifically 10 residues away from the prior site of glycosylation. This suggested that the lectin domain assists in directing glycosylation in a N- and C-terminal direction. Without the lectin domain of ppGalNAc T2, however, the sites of glycosylation by the catalytic domain was the same in all MUC5AC (glyco)peptide substrates, at the Thr9, Figure 4.1 C. This clearly suggested that the lectin domain is involved in glycopeptide recognition and involved in mediating directionality preferences. However, when the dissociation constants were measured between the ppGalNAc T2 and the lectin domain deleted ppGalNAc T2, differences were observed with the MUC5AC-3 and MUC5AC-13 glycopeptides. Basically, the binding (K\textsubscript{d}) of the MUC5AC-13 was reduced by 4.9 fold in the absence of the lectin domain while the binding of MUC5AC-3 was not. This shows that the binding of MUC5AC-13 is lectin domain-mediated while the binding of MUC5AC-3 is not. The authors explained that the glycosylation of MUC5AC-3 (at Thr13) was driven by the increased turnover rate of the enzyme due to the lectin domain
Figure 4.1- Glycosylation sites observed with versions of ppGalNAc T2. (A) Library of MUC5AC (glyco)peptides used in the work of (11). The purple T* indicates and prior site of glycosylation while the red Thr indicates the observed sites of glycosylation. Glycosylation that is catalytic domain-directed, Thr9 are bold (red), while the lectin domain-assisted sites of glycosylation are underlined in red. (B) Edman sequencing of the glycosylation products of the MUC5AC (glyco)peptide library with ppGalNAc T2. Observed sites of glycosylation are marked in red. Note the shift in glycosylation sites between glycopeptides. (C) Edman sequencing of the glycosylation products of the MUC5AC (glyco)peptide library with only the catalytic domain of ppGalNAc T2. Observed sites of glycosylation are marked in red. Note that the Thr9 is the dominant site of glycosylation for all substrates. Data modified from (11).
assisting in other means such as product release (i.e. the lectin domain pulling off the product). Because of this study, the Gerken Lab became interested in expanding their studies to the lectin domain and because no truly systematic study of the ppGalNAc T glycopeptide substrate utilization had existed, this seemed like the perfect project for the lab.

In the previous chapter, a series of randomized (glyco)peptides was designed to interrogate the role of the ppGalNAc T catalytic domains in determining glycopeptide specificity for several family I and family II ppGalNAc T subfamily members. This was primarily based on characterizing the effects of neighboring prior glycosylation and I showed that for the “glycopeptide-preferring” transferases (ppGalNAc T4, T7, T10, T12 and PGANT7) and “peptide-preferring” transferases ppGalNAc T11 and dT1 (PGANT35A), the catalytic domain was responsible for the binding of the glycopeptide T* (GalNAc-O-Thr) and subsequently glycosylating unique sites that were +/-4 residues away. For the peptide-preferring transferases (ppGalNAc T2, T3 and T5), the neighboring GalNAc on the acceptor site results in inhibition of glycosylation, where lower rates of glycosylation were observed. Nonetheless, sites of glycosylation were observed at the maximum distances, at -5/+5. This chapter will focus on characterizing the effects of remote prior glycosylation, where a site of prior glycosylation T* will be 6-17 residues from an acceptor site of glycosylation. Here, I will show that while Family I members (ppGalNAc T1, T2, T3, T5, T6, T11, T13, T14, T16, dT1 (PGANT35A), PGANT9A and PGANT9B) readily glycosylate peptide substrates and have been termed “peptide-preferring” isoforms; several members of this family (Ia-Id) are influenced by remote, 6-17 residues of a prior GalNAc-O-Ser/Thr glycosylation in an N- or C- terminal
direction that is mediated by their lectin domains. Subfamily IIa members (ppGalNAc T4 and T12), on the other hand, which commonly display sole activities against a neighboring GalNAc-O-Ser/Thr containing glycopeptides, will also be shown to be influenced by remote prior glycosylation, specifically in a C-terminal direction, while subfamily IIb (ppGalNAc T7, T10 and PGANT7) members have no activity or weak lectin domain preferences. Other studies comparing the lectin domain preferences were performed on three human ppGalNAc T’s and/or fly ortholog pairs; ppGalNAc T11 and dT1(PGANT35A), a lectin domain splice variant of ppGalNAc T13 (known as ppGalNAc T13 Ex10B) and lectin domain splice variants PGANT9A & PGANT9B (fly transferases).

Overall, these studies show that the presence of the N- or C- terminally placed T* in these glycopeptides are important determinants of the catalytic activity and specificity of these transferases, which can significantly differ between transferase isoforms. Because there were large differences in glycosylation rates observed for some transferases between the glycopeptide substrates, we believe we have now uncovered another level of control of mucin-type of glycosylation that will further advance our understanding of the regulation of this important modification.

4.2 Materials and Methods

4.2.1 Transferases

As in most of our work, the transferases were obtained from multiple sources and expression systems that were provided by our collaborators see Table 2.2 in chapter 2 for list of transferases and sources. Transferases are either N-terminal truncated and affinity tagged constructs that were bound to affinity beads or affinity purified. Although the
transferases were from different sources, the experiments completed using the 
transferases yielded comparable results. Bovine ppGalNAc T1, which possesses a 98.9%
sequence identity to human T1, was a gift from Ake Elhammer (15). This transferase was 
expressed in Sf9 insect cells and supplied as a soluble purified transferase. ppGalNAc T2 
and T3 were supplied from two sources and expressed either from Pichia pastoris or 
High Five insect cells (11,16,17). They were supplied either in cell culture media or 
soluble affinity purified. Soluble N-terminal polyHis-tagged recombinant transferases 
ppGalNAc T6, T7, T11, T14, T16 and dT1 were expressed from High five insect cells as 
secreted proteins and purified from Ni-NTA aragose (Invitrogen) or SP-Sepharose 
(Sigma-Aldrich) and MonoQ 5/50 ion exchange chromatography (GE Healthcare) as 
previously described (2,17-19). Poly-His tagged ppGalNAc T4, T6, T7 and T13 were 
expressed from baculovirus-infected Sf9 cell system as secreted proteins. The 
transferases were N-terminally His-tagged catalytic domains bound to Ni-NTA-affinity 
beads (Thermo-Fisher)¹ (20). After extensive washing, the transferases were used for 
enzyme assays. ppGalNAc T4, T5 and T10 was expressed using HEK293f cells and 
purified Ni-NTA superflow (Qiagen) nickel affinity chromatography¹, methods were 
similar to the expression of rat ST6GalI (21). PGANT9A, PGANT9B and PGANT7 was 
expressed in COS7 cells and the recombinant secreted protein of the cell culture media 
was used (22). Reactions with the control media gave no significant [³H]-UDP-GalNAc 
utilization compared to the expression media.

¹ For detailed methods and cDNA constructs, visit the University of Georgia’s Resource 
for Integrated Glycotechnology, “Glyco-Enzyme Repository” 
(http://glycoenzymes.ccrc.uga.edu/).
4.2.2 Reagents and (Glyco)peptide Substrates

Random (glyco)peptide substrates GP(T*22)R, GP(A10)R, GP(T*10)L and GP(A10)L (found in Table 4.1) were custom synthesized by Sussex Research, Ottawa, ON (Canada). R and L refer to the right and left locations of the T* (GalNAc-O-Thr) in the glycopeptides. These (glyco)peptide substrates were formerly named GPIV, GPIV Cont, GPV and GPV Cont, respectively. Stock solutions of (glyco)peptide substrates were prepared by lyophilizing from water several times and adjusting to pH 7-7.5 with either dilute NaOH and/or HCl and stored frozen at a concentration of 50 mg/mL (1.5-1.7 mM). The other materials such as [3H]-UDP-GalNAc (C6-[3H]-CH2-labeled), protease inhibitor cocktails, dowex anion exchange (1x8, 200 mesh) resin, Sephadex G10 and running buffer is the same as described in chapter 2.

4.2.3 Instrumentation

All instrumentation including the Applied Biosystems Procise 494 sequencer, scintillation counter, lyophilizer and spectrophotometer used in this work are the same as described in chapter 2.

4.2.4 Glycosylation Assays with Recombinant ppGalNAc T’s and (Glyco)peptide Substrates GP(T*22)R, GP(T*10)L and Non-glycosylated Controls GP(A22)R and GP(A10)L

Standard reaction conditions using the four lectin (glyco)peptides are as follows: 100 mM HEPES, pH 7.5 (ppGalNAc T1, T2 and T3) or 68 mM sodium cacodylate
Table 4.1- ppGalNAc T lectin domain probing (glyco)peptide substrates.

<table>
<thead>
<tr>
<th>Peptide/ Glycopeptide (Name)</th>
<th>Designated Color</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP(T*22)R</td>
<td>Red</td>
<td>GAGA\ldots\ldots T^{*}_{22} \ldots AG</td>
</tr>
<tr>
<td>GP(A22)R</td>
<td>Purple</td>
<td>GAGA\ldots\ldots A_{22} \ldots AG</td>
</tr>
<tr>
<td>GP(T*10)L</td>
<td>Blue</td>
<td>GAGAZ\ldots \ldots T^{*}_{10} \ldots AG</td>
</tr>
<tr>
<td>GP(A10)L</td>
<td>Light Blue</td>
<td>GAGAZ\ldots \ldots A_{10} \ldots AG</td>
</tr>
</tbody>
</table>

\(X = G, A, R, P, N, E, Y, V \& T\)
\(Z = G, A, R, P, N, E, V, Y\)
\(T^* = \text{Thr-O-GalNAc}\)

Note that the acceptor positions are underlined in red and indicate the presence of a randomized Thr residue.
(ppGalNAc T3, T4, T5, T6, T7, T10, T11, T12, T13, T14, T16, dT1, PGANT7, PGANT9A and PGANT9B), pH 6.5, 1.8 mM 2-mercaptoethanol, 10 mM MnCl$_2$, 50 µM [$^3$H]-UDP-GalNAc (~ 6 x 10$^8$ DPM/µmol), 1.5 mM (5 mg/mL) of substrates GP(T*22)R, GP(A22)R, GP(T*10)L and GP(A10)L and up to 450 µL of transferase (either soluble or transferase bound to affinity beads). Reaction volumes ranged from 75-600 µL and were carried out in 2 mL capped Eppendorf tubes. The reaction mixtures were incubated at 37°C and were left shaken in a TAITEC Microincubator M-36. Time point aliquots of either 15-120 µL were removed at 15, 45, 120 (2 hr), 240 (4 hr) and ~1200 (overnight) minutes and quenched with an equal volume of 250 mM EDTA. Reactions with all four substrates were conducted at the same time and with the same transferase concentrations and the same UDP-$[^3$H]-GalNAc stock. Detailed information about each specific transferase reactions will be discussed in the appropriate sections below.

Incubations were performed using a range of transferase concentrations (determined by trial and error) that would transfer between 10-50% of the total UDP-$[^3$H]-GalNAc to the optimal substrate after an overnight incubation (giving a range of ~0.003-0.016 mol of GalNAc/mol of glycopeptide). In total, 2-10 independent experiments were performed at 2-3 different transferase concentrations (except for ppGalNAc T14 and PGANT7). UDP-GalNAc concentrations were kept low (0.5-50 µM) to achieve high specific activity for subsequent Edman sequencing analysis (described below). After quenching, the total [$^3$H]-GalNAc utilization and separation of the product glycosylated glycopeptides were determined and treated the same way as described in chapter 3. The results are averaged and represented in the final plots.
4.2.5 Determination of Optimal Reaction Conditions and Initial Observations

Before beginning this work, several initial reactions were performed by other lab members in order to find the best conditions. The initial use of the GP(T*22)R and GP(T*10)L series of glycopeptide substrates in Table 4.1 with ppGalNAc T1 and T2 gave results that seemed to vary with glycopeptide concentration. When using the lowest substrate concentrations of ~0.16 or 0.016 mM of (glyco)peptide substrate, all four substrates were inactive. At the highest glycopeptide concentrations of 1.5-1.7 mM, differential activities were observed where the glycopeptide substrates displayed significant higher activities than the non-glycosylated peptide controls. These results are summarized in Figure 4.2. These differential effects are attributed to substrate concentrations being initially below the $K_m$ value of the substrate and we have therefore chosen to perform these studies with substrate concentrations of 1.5-1.7 mM (5 mg/mL). Because of the random nature and multiple acceptor sites of these substrates, a continuum of $K_m$ and $V_{max}$ values is expected; we have therefore not pursued any detailed kinetic analysis of these substrates. It was also noted that the relative rates of glycosylation of the four substrates varied somewhat with the amount of transferase used in the assay; nevertheless, overall the same transferase preferences for the different substrates were observed (data not shown).

As mentioned previously in chapter 3, the transferase reaction time course plots shown in the results section generally show a characteristic plateau in the transfer of GalNAc that is often observed with high transferase activities and is even observed the least active substrates. This is again due to to same factors. For the most active substrates (approaching 50%
Figure 4.2- Substrate concentration dependence for ppGalNAc T1 and T2. (Glyco)peptides GP(T*22)R (red diamonds), GP(A22)R (purple triangles), GP(T*10)L (blue squares) and GP(A10)L (light blue downward triangles). (A) ppGalNAc T1 with substrate concentration of 5 mg/mL (1.5-1.7 mM) and (B) 0.5 mg/mL (0.15-0.17 mg/mL). ppGalNAc T2 with substrate concentrations of 5 mg/mL (1.5-1.7 mM) (C), 0.5 mg/mL (0.15-0.17 mg/mL) (D) and 0.05 mg/mL (0.015-0.017 mM) (E). The overall trend shows the specificities at higher substrate concentrations. Note that the UDP-[³H]-GalNAc concentration for these experiments was 0.5 µM (12). This experiment was previously performed by another lab member in the Gerken Lab.
utilization), the plateau in incorporation is most likely due to the depletion of the UDP-\[^3\text{H}^\]-GalNAc donor. For the less active substrates, the slowing of the rate is thought to be due to the sequential loss of the optimal or best acceptor sites as glycosylation of the random (glyco)peptide progresses. Furthermore, because of the half-lives of ppGalNAc T1 and T2 are about 5 and 2.5 days (23), it is unlikely that the transferase inactivation is the source of the plateau with these transferases. However, transferase inactivation cannot be ruled out for the remaining transferases.

4.2.6 Determination of the GalNAc Acceptor Sites of Incorporation by Edman Sequencing

The lyophilized post-Dowex \[^3\text{H}^\]-GalNAc glycosylated glycopeptide were processed and analyzed as described in chapter 3. To reduce nonspecific binding of the \[^3\text{H}^\]-GalNAc-O-Thr-PTH derivative to the glass loading disk (determined by counting the disk after sequencing), ~0.25 mg of non-reacted “cold” (glyco)peptide substrate was commonly added to the sample to be analyzed. Nevertheless, significant variability in recovery of radiolabel was observed. A preview in radioactivity observed in the N-terminal residues, particularly noticeable with the GP(T*10)L and GP(A10)L (glyco)peptides, is thought to be due to non-specific release of the \[^3\text{H}^\]-GalNAc (glyco)peptide from the glass filter during the Edman sequencing. Edman sequencing on the non-glycosylated controls GP(A22)R and GP(A10)L were performed on selected samples and discussed in a separate section below.
4.3 Results and Discussion

4.3.1 Design of the Lectin Domain Probing Random (Glyco)peptide Substrates

The goal of these studies was to develop a series of “universal” substrates that could be used with any ppGalNAc T isoform to access the role of each isoform’s lectin domain. In a nutshell, these glycopeptide substrates will address the role of a remote GalNAc-O-Thr (abbreviated as T*) on transferase activity, based on the assumption that the T* would bind at the lectin domain, therefore directing the catalytic domain to glycosylate the X residues. These substrates were designed to be sufficiently long, specifically 29 residues, to span the catalytic and lectin domains based on the X-ray crystal structures of ppGalNAc T1 and T2 (1,9), in Figure 4.3. As shown in Table 4.1, each acceptor substrate contained a T* placed near the C- or N-terminal (GP(T*22)R and GP(T*10)L, where R is for right and L is for left placed T*, respectively) which in turn, were flanked by five randomized Z residues lacking an acceptor Ser or Thr acceptor. An additional 12 randomized X residues, also including the acceptor Thr, were extended from the Z’s in a N- or C- terminal direction, respectively. These X residues thus serve to probe glycosylation 6-17 residues from an existing T* in an N- or C- direction, respectively. To confirm that the T* in these glycopeptides were indeed interacting with the lectin domain, non-interacting control peptides were also used where the T* was replaced by an Ala residue, GP(A22)R and GP(A10)L. The rationale behind utilizing the randomized residues, X and Z rather than a specific peptide sequence was based on our goal to eliminate (or at least reduce) transferase-specific bias due to a particular peptide sequence. With these (glyco)peptides, we can systematically determine the extent that
**Figure 4.3**- Lectin domain substrate design based in the X-ray crystal structure of ppGalNAc T2 with EA2 peptide bound in the catalytic domain (10). Drawn in the catalytic domain are lines that represent glycopeptides GP(T*22)R (red line), and GP(T*10)L (blue line), showing the expected glycopeptide binding, where the remote prior sites of glycosylation are either extended to the right of the catalytic domain with GP(T*22)R (prior remote glycosylated C-terminus ~XZZZZZT*ZZZZZAG) or to the left of the catalytic domain with GP(T*10)L (prior remote glycosylated N-terminus GAGAZZZZZT*ZZZZZ~). Approximate locations of the lectin subdomains are marked.
remote GalNAc-O-Thr interactions, at the lectin domain can enhance catalytic domain glycosylation of the X residues and whether there is a preferred N- or C- directionality. In addition, the optimal number of residues between the lectin bound T* and the residues glycosylated by the catalytic domain can be determined from these lectin domain substrates. For clarity, results for the GP-R series of substrates will be presented in shades of red and those for GP-L will be in shades of blue.

Confirmation that these lectin domain substrates are indeed reporting on lectin domain binding/interactions of a GalNAc residue is shown in Figure 4.4. In these plots, the effects of adding a free GalNAc competitor were examined using ppGalNAc T2, where the addition of free GalNAc significantly reduced the activity of GP(T*22)R while not affecting the activity of the non-glycosylated control peptides. It was also revealed that under the higher (glyco)peptide and transferase concentrations and lower UDP-[3H]-GalNAc concentrations utilized in this and in many of the initial experiments performed by other lab members, that GP(T*10)L also displayed an elevated activity relative to the control that is completely inhibited by added free GalNAc (Figure 4.4 C and D). These results suggest both GP(T*22)R and GP(T*10)L glycopeptide substrates can interact with the lectin domain of ppGalNAc T2 under certain conditions, also leading to rate enhancements. This is consistent with the kinetic observations in Raman et. al (11), where they reported a lectin domain assistance in an C-terminal T* glycosylated MUC5AC-13 and a lectin domain involvement in product release for an prior N-terminal T* glycosylated MUC5AC-3 (GTT*PSPVPTTSTSAP) glycopeptide substrate. However, the molecular mechanisms of how the lectin domain-mediated product release
**Figure 4.4** - Effects on adding free GalNAc on ppGalNAc T2 (glyco)peptide utilization. (A) GP(T*22)R (red diamonds) and (B) GP(A22)R (purple triangles), in the presence of 100 mM GalNAc. (C) GP(T*10)L (blue squares) and (D) GP(A10)L (light blue downward triangles) in the presence of 100 mM GalNAc. Note the decrease in $[^3]$H-GalNAc transferred in the presence of 100 mM GalNAc. Substrate concentrations was 5 mg/mL (1.5-1.7 mM) and the UDP-$[^3]$H-GalNAc concentration for these experiments was 0.5 µM (12). This experiment was previously performed by another member in the Gerken Lab.
occurs and contributes to ppGalNAc T2’s C-terminal long range specificity is presently unknown and awaits further study.

4.3.2 Glycosylation of the GP-L and GP-R series of (Glyco)peptides with ppGalNAc T1, T2, T3, T4, T5, T6, T7, T10, T11, T12, T13, T13 Ex10B, T14, T16, dT1 (PGANT35A), PGANT7, PGANT9A and PGANT9B

The results of the glycosylation with the lectin domain probing glycopeptides, GP(T*22)R, GP(A10)R, GP(T*10)L and GP(A10)L with several ppGalNAc T transferases from families I and II and fly orthologs will be presented in a series of plots, in the same format as chapter 3, where the first data displayed will be the reaction time course plots, followed by the Sephadex G10 chromatograms and the Edman amino acid sequencing. The Sephadex G10 runs and Edman sequencing were performed on the overnight incubations (and occasionally for the 4hr incubations for some transferases) for selected (glyco)peptide substrates. All of the data presented will be broken down subfamily classifications (i.e. Ia, Ib, Ic, Id, If, IIa and IIb).

4.3.2.1 ppGalNAc T1 and T13 in Subfamily Ia

As discussed in the previous chapters, both ppGalNAc T1 and T13 are considered to be peptide-preferring transferases and share an 88% sequence identity. In chapter 2, I showed that both ppGalNAc T1 and T13 have nearly identical catalytic domain peptide preferences, while in chapter 3, I showed that catalytic domain ppGalNAc T13 might not have a preference for a glycopeptide with a neighboring prior site of glycosylation. In this chapter, I will show that both transferases have a preference for a glycopeptide(s) with a
remote prior site of glycosylation and demonstrate that these transferases have different directionality preferences.

A total of 6 determinations were made (with all four (glyco)peptide substrates at the same time) with 60-70 µL (of a stock 1:100 fold dilution) of bovine ppGalNAc T1 and 80-125 µL of either ppGalNAc T13 in beads or soluble affinity purified, using the standard reaction conditions as described above. The resulting reaction time course plots and accompanying overnight Sephadex G10 chromatograms are shown in Figure 4.5 (left and right panels). In Figure 4.5 A (left and right panels), the reaction time course plots show that ppGalNAc T1 has elevated activities towards GP(T*22)R while T13 has about equal activities towards both GP(T*22)R and GP(T*10)L. This is confirmed in the Sephadex G10 plots in Figure 4.5 B (left and right panels), which shows that the net [³H]-GalNAc utilization observed in the reaction time course plots is due to transfer to glycopeptide substrate and not hydrolysis. For ppGalNAc T1, the peak for glycopeptide product is the largest for GP(T*22)R while the others are essentially lower. For ppGalNAc T13, both the GP(T*22)R and GP(T*10)L glycopeptide product peaks are about equal. This suggests that the lectin domain of ppGalNAc T1 interacts with a glycopeptide that has a prior T* that is in a C-terminal direction meaning that the lectin domain is positioned to the right of the catalytic domain while the lectin domain of T13 interacts with a glycopeptide that has a prior T* in both an N- and C-terminal direction suggesting that the lectin domain may have extensive lectin domain conformational dynamics where it can be equally populated to the right or to the left of the catalytic domain.
Figure 4.5- Subfamily Ia ppGalNAc T1 and T13 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T1 and T13 showing the net $[^3]H$-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating mostly $[^3]H$-GalNAc transfer to glycopeptide and low hydrolysis for ppGalNAc T1 and T13. (C-D) $[^3]H$-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T1 and T13 showing the optimal distance from a T* to the X acceptor residues is approximately 9-10 residues. (E) $[^3]H$-GalNAc incorporation into the X residues of GP(A22)R and GP(A10)L with and T13. Note the higher DPM scale in (C).
The Edman amino acid sequencing analysis of GP(T*22)R and GP(T*10)L for both ppGalNAc T1 and T13 is shown in Figure 4.5 C (left and right panels). Rather than observing a sharp peak of maximal $[^3]$H-GalNAc incorporation (as mostly observed in chapter 3), a relatively broad peak of distribution of incorporation is observed for all substrates. The distance from the T* to the peak of incorporation in the X residues of GP(T*22)R and GP(T*10)L glycosylated by ppGalNAc T1 and T13 is 9-10 residues. Nevertheless, there are different glycosylation patterns observed with both transferases. Note the much higher DPM values for ppGalNAc T1 with GP(T*22)R (Figure 4.5 C left) compared to GP(T*10)L. The Edman amino acid sequencing analysis of GP(A22)R and GP(A10)L with ppGalNAc T13 are shown in the left and right panels in Figure 4.5 E, displaying low peaks of incorporation. How this control data was used to reveal the extent of lectin domain involvement will be discussed in a separate section below.

Prior evidence showing that there may be lectin domain targeting of N- or C-terminal prior T* glycosylation for ppGalNAc T1 and T13 was indicated in the results of Zhang et al. (24) for the multiple glycosylation reactions of the MUC7 peptide (PTPSATT$^7$PAPPSS$_{13}$S$_{14}$APPET$_{19}$T$_{20}$AAK). Their studies of the MUC7 peptide showed that both ppGalNAc T1 and T13 were capable of glycosylating Thr7. When this glycosylated substrate was used again against both transferases, only ppGalNAc T13 was capable of further glycosylating the C-terminal Ser13-14 and Thr19-20 residues (shown in bold) of MUC7; ppGalNAc T1 did not glycosylate these residues. These findings are consistent with the transferase glycopeptide preferences observed in my studies; where ppGalNAc T1 prefers to glycosylate sites N-terminal of a prior site of glycosylation,
whereas ppGalNAc T13 is capable of glycosylating sites both N- and C-terminal of a prior site of glycosylation.

4.3.2.2 *ppGalNAc T2, T14 and T16* in *Subfamily Ib*

In the previous chapters, I have shown that all of these three subfamily Ib members, ppGalNAc T2, T14 and T16, have very similar catalytic domain peptide preferences and that the catalytic domain of ppGalNAc T2 (and perhaps ppGalNAc T16) may have limited abilities to glycosylate a glycopeptide substrate that has a T* near the acceptor site. These similarities can be attributed to the fact that their sequence identities are well above 53%. In this chapter, I will demonstrate that the remote directionality preferences for glycopeptide substrates with ppGalNAc T16 is different from ppGalNAc T2 and T14.

Using all four glycopeptide substrates in one assay, a total of 5 determinations were made with ppGalNAc T2 using 60-70 µL (of a stock 1:10 fold dilution), a total of 2 determinations for ppGalNAc T14 with 205 µL of soluble affinity purified transferase and 3 determinations were made with ppGalNAc T16 using 60-70 µL (of a stock 1:10 fold dilution) of soluble affinity purified transferase. All assays were performed using the standard reaction conditions as described above. The resulting series of plots are shown in Figure 4.6. The reaction time course plots and Sephadex G10 chromatograms for ppGalNAc T2 and T14 (Figures 4.6 A-B left and middle panels, respectively) both demonstrate that the highest activity is with the GP(T*22)R substrate while the activities for other (glyco)peptide substrates are much lower. In contrast for ppGalNAc T16, the plots show equal activities for both GP(T*22)R and GP(T*10)L (see Figure 4.6 A-B right.
Figure 4.6- Subfamily Ib ppGalNAc T2, T14 and T16 lectin domain preferences.
(A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T2, T14 and T16 showing the net [3H]-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating mostly [3H]-GalNAc transfer to glycopeptide and low hydrolysis for ppGalNAc T2, T14 and T16. (C-D) [3H]-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T2, T14 and T16 showing the optimal distance from a T* to the X acceptor residues is approximately 9-13 residues. ND stands for not determined. (Note the DPM scale differences between panels C and D). Red arrow in ppGalNAc T2 GP(T*22)R represents the closest distance the lectin domain is interacting.
panels). This suggests that the lectin domains of ppGalNAc T2 and T14 both interact with a glycopeptide substrate with a prior C-terminal T* while the lectin domain of T16 interacts with glycopeptides that have a prior T* in either N- or C-terminal directions. This represents the first difference observed with these subfamily members.

The Edman amino acid sequencing plots show that the optimal glycosylation distances from the N- or C-terminal T* are ~10 residues for ppGalNAc T2 (Figure 4.6 C-D left panels), 13 residues for ppGalNAc T14 with GP(T*22)R (Figure 4.6 C middle panel), and 9 residues for ppGalNAc T16 (Figure 4.6 C-D right panel). Note that the sequencing for GP(T*10)L glycosylated with ppGalNAc T14 was not performed due to very low amounts of [3H]-GalNAc transfer. The biggest differences in glycosylation patterns are observed for GP(T*22)R between the three transferases, especially with ppGalNAc T14. Back in chapter 3, I showed that for ppGalNAc T2 glycosylated with GP(T*10)C-Ser (see Figure 3.2), there was an observed -5 glycosylation site preference and I mentioned that this might be lectin domain mediated. With these glycosylation results with GP(T*22)R, I can confirm that this -5 glycosylation site preference for ppGalNAc T2 observed earlier is indeed lectin domain mediated since it overlaps with the glycosylation patterns observed for ppGalNAc T2 glycosylated with GP(T*22)R, where glycosylation can occur as early as -6 residues from the prior site of glycosylation. Thus, the -5 glycosylation site preference is the closest distance at which the lectin domain can interact with a nearby glycosylated residue, see red arrow for ppGalNAc T2 in Figure 4.6 C.

The catalytic and lectin domain of ppGalNAc T2 was extensively studied with a series of MUC5AC glycopeptides in Raman et al. (11). Based on comparative studies, it
was concluded that ppGalNAc T2 bound glycopeptide MUC5AC-13 (GTT₃PSPVPTTSTT*₁₃SAP) in a lectin domain-assisted manner, where the acceptor site was 10 residues N-terminal of an existing site of prior GalNAc glycosylation (glycosylation by WT ppGalNAc T2 was observed at the Thr3) but not a glycopeptide with the reverse orientation and this was due to the lectin domain being involved in product release. This is consistent with our findings with ppGalNAc T2 with GP(T*₂²)R and GP(T*₁⁰)L. Recently, an X-ray crystal structure of ppGalNAc T2 with MUC5AC-13 bound was recently reported (25) in the expected orientation that we would expect.

4.3.2.3 ppGalNAc T3 and T₆ in Subfamily Ic

ppGalNAc T3 and T₆ belong to the family Ic of peptide-preferring transferases and share a 77% sequence identity. So far the catalytic domain peptide preferences and glycopeptide preferences have only been determined for ppGalNAc T3. Until now, no comparisons could be made between both subfamily members due to the limited amount of ppGalNAc T₆. In this chapter, the lectin domain preferences of ppGalNAc T3 and T₆ will be determined and compared with each other.

Using the same standard reaction conditions, a total of 3 determinations were made with ppGalNAc T₃ using 125-230 µL of soluble affinity purified transferase and 125-460 µL of ppGalNAc T₆ on beads. The resulting series of plots are shown in Figure 4.7, where the reaction time course plots and Sephadex G10 chromatograms (Figure 4.7 A-B left and right panels) both demonstrate that ppGalNAc T₃ and T₆ both have preferences for the GP(T*₁⁰)L while the activities for the other glycopeptide substrates are low. This is the only subfamily of peptide-preferring transferases that possesses this lectin domain preference, where the lectin domain interacts with a prior glycosylated
Figure 4.7- Subfamily Ic ppGalNAc T3 and T6 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T3 and T6 showing the net [³H]-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating mostly [³H]-GalNAc transfer to glycopeptide and low hydrolysis for ppGalNAc T3 and T6. (C-D) [³H]-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T3 and T6 showing the optimal distance from a T* to the X acceptor residues is approximately 7-16 residues. ND stands for not determined. Red arrow in ppGalNAc T3 GP(T*10)L represents the closest distance the lectin domain is interacting.
residue (T*) that is N-terminal to the catalytic domain, suggesting that the lectin domain might be positioned to the left of the catalytic domain.

The Edman amino acid sequencing plots in Figure 4.7 C-D left and right panels show that the glycosylation patterns are different for GP(T*10)L between ppGalNAc T3 and T6, where the optimal glycosylation distances is ~8 residues and ~7 residues respectively. Interestingly, glycopeptide GP(T*22)R glycosylated with ppGalNAc T3 shows a broad distribution and a sharp peak of [³H]-GalNAc incorporation that is ~16 residues N-terminal of a T*. This could suggest that the lectin domain of ppGalNAc T3 could also weakly interact with GP(T*22)R and glycosylate remote sites.

In chapter 3, I showed that for ppGalNAc T3 glycosylating GP(T*10)C-Ser, there was a +5 glycosylation site preference relative to the T* (see Figure 3.3) and mentioned that this might be lectin domain mediated. From the glycosylation results with GP(T*10)L, I can confirm that indeed this is lectin domain mediated since the [³H]-GalNAc incorporation plots of GP(T*10)L overlaps with GP(T*10)C-Ser (i.e. the glycosylation is observed in GP(T*10)L begins sharply at the +6 residue from the prior site of glycosylation). Thus, the +5 preference observed with GP(T*10)C-Ser may represent the closest distance at which the lectin domain of ppGalNAc T3 may interact relative to a prior T*, see red arrow in Figure 4.7 D.

Previous studies on the glycan binding requirements of the lectin domain of ppGalNAc T3 also suggested that this transferase utilizes its lectin domain to glycosylate residues that are C-terminal of the prior site of glycosylation (T*) (8), which is in agreement with our findings with GP(T*10)L. Other studies with ppGalNAc T3 and Fibroblast Growth Factor 23 (FGF23) showed that T3 was the only transferase capable of
glycosylating the proprotein processing region (i.e. Thr178) of FGF23 (IHFNT171PIPRRHT178R_SAEDD) [where the cleavage site in underlined] (26-28) and unpublished work by Kato et al. (see chapter 2 for a longer discussion on this). They found that the glycosylation must occur in an ordered manner, where ppGalNAc T3 first glycosylates the peptide Thr171 in a lectin independent manner and then Thr-178 site in a lectin-dependent manner for both in vitro enzyme reactions and ex vivo cell line experiments. The Thr171 site can be glycosylated by other transferases (26) and in previous work by our lab (16), we have shown using our transferase specific preferences obtained from the random peptide substrates, that Thr171 would be a moderate to good substrate for ppGalNAc T1, T2 and T3 where as Thr178 would not be a good substrate for ppGalNAc T1 and T2 but a modest substrate for ppGalNAc T3 (see chapter 2, Table 2.5 for ISOGlyP EVP’s). These results suggest that the prior glycosylation of Thr171 may serve to target ppGalNAc T3 glycosylation of Thr178, 7 residues C-terminal of Thr171. The glycosylation products of ppGalNAc T3 with GP(T*10)L shown in Figure 4.7 D left panel, is entirely consistent with these experiments, where a peak of optimal glycosylation is shown to be ~8-9 residues C-terminal of the original site of glycosylation. In contrast, the prior glycosylation of Thr171 would fail to enhance the glycosylation of Thr178 by ppGalNAc T1 and T2 because these transferases possess the reverse glycopeptide preferences. Thus, the glycosylation of Thr171 enhances the glycosylation of Thr178 by ppGalNAc T3 while effectively reducing the activities of those transferases with the reverse glycopeptide preferences. These conclusions were published in the paper that I co-authored with Dr. Gerken (12).
4.3.2.4 *ppGalNAc T5 in Subfamily Id*

For *ppGalNAc T5*, a total of 10 determinations were made using 60-80 µL of a diluted (1:10) stock solution of soluble transferase using the standard reaction conditions. The series of plots are shown in Figure 4.8, where both the reaction time course plots and accompanying Sephadex G10 chromatograms display that *ppGalNAc T5* has a nearly equal preferences for lectin glycopeptides GP(T*22)R and GP(T*10)L (Figure 4.8 A-B). This suggests that the lectin domain of *ppGalNAc T5* must be equally populated to the right and left of the catalytic domain.

The Edman amino acid sequencing plots demonstrate that the optimal glycosylation distances are 9-10 residues from the prior site of glycosylation (T*). In chapter 3, I showed that for *ppGalNAc T5* glycosylated with GP(T*10)C-Ser and GP(T*10)C-Thr, that there was an observed -5 glycosylation site preference (see Figure 3.4) and mentioned that this might be lectin domain mediated. With the glycosylation results with GP(T*22)R, I can confirm that indeed this is lectin domain mediated since the [³H]-GalNAc incorporation plots of GP(T*22)R overlap with GP(T*10)C-Ser and GP(T*10)C-Thr, suggesting that the -5 preference observed for these glycopeptides represents the closest distance at which the lectin domain of *ppGalNAc T5* begins to interact, see red arrow in Figure 4.8 C. It is interesting that we were only able to detect the -5 preference site in GP(T*10)C-Ser and GP(T*10)C-Thr and not the +5 site, given the fact that *ppGalNAc T5* is equally active against both lectin domain probing substrates.
Figure 4.8- Subfamily Id ppGalNAc T5 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T5 showing the net $[^3]$H-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating mostly $[^3]$H-GalNAc transfer to glycopeptide and low hydrolysis for ppGalNAc T5. (C-D) $[^3]$H-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T5 showing the optimal distance from a T* to the X acceptor residues is approximately 9-10 residues. Red arrow in ppGalNAc T5 GP(T*22)R represents the closest distance the lectin domain is interacting.
Why this occurred is not known, however, there might be more of a preference for GP(T*22)R than GP(T*10)L (both the reaction time course plots and Sephadex G10 chromatograms show the activities with GP(T*22)R to be slightly higher), which could explain why we are seeing this directionality preference with GP(T*10)C-Ser and GP(T*10)C-Thr. Another possibility could be that the preferred distance for GP(T*10)L is just longer than that of GP(T*22)R.

4.3.2.5 ppGalNAc T11 and dT1 in Subfamily I

ppGalNAc T11 and dT1 have nearly identical catalytic domain peptide preferences, as I showed in chapter 2. It is possible that the lectin domain glycopeptide activities could be different between both transferases as I showed in chapter 3 and represents the first difference observed between both transferases. In this chapter, I will further focus on comparing the lectin domain preferences.

For ppGalNAc T11 and dT1, a total of 6 determinations were made using 300-450 µL of soluble affinity purified T11 and dT1 using the standard reaction conditions. The resulting series of reaction time course plots and Sephadex G10 chromatograms are shown in Figure 4.9 A-B left and right panels. Both of these plots show that both ppGalNAc T11 and dT1 mostly prefer the GP(T*22)R glycopeptide substrate over GP(T*10)L. This suggests that the lectin domains of ppGalNAc T11 and dT1 interact with a glycopeptide substrate with a prior C-terminal T*. In order to show any further differences, the glycosylation products of GP(T*22)R and GP(T*10)L glycosylated with ppGalNAc T11 and dT1 that I obtained were counted for [3H]-GalNAc and monitored for OD220 and OD280. These values were then divided; [3H]-DPM/OD220 and [3H]-
DPM/OD2280 to obtain ratios to better quantify the actual $[^{3}H]$-GalNAc content of each glycopeptide. These values were obtained for each glycopeptide, GP(T*22)R and GP(T*10)L and then ratioed. Interestingly, these ratios revealed that $dT1$ prefers to glycosylate the GP(T*22)R glycopeptide substrate almost 2 fold greater than that of ppGalNAc T11 (see Figure 4.9 C), where the ratio of GP(T*22)R/GP(T*10)L for T11 is ~1.4 and $dT1$ is ~2.6. Whether this has a significant impact on the in vivo activities of these transferases is unknown. It could be that these differences are in the recognition of tertiary structures as recently work has shown that ppGalNAc T11 can recognize the linker regions of the LDLR receptor (29). Note that there were large standard deviation errors in Figure 4.9 C are due to the variability of the OD’s for the glycopeptide substrates, which were often times difficult to obtain due to their low absorbances at 220 and 280nm. Still, this data was useful in determining differences of both transferases.

The Edman sequencing plots demonstrate very similar optimal glycosylation site preferences for ppGalNAc T11 and $dT1$. While the distribution is broad, there are sharp peaks of optimal distances of glycosylation for ppGalNAc T11 is approximately 16 residues for GP(T*22)R and 7 residues for GP(T*10)L while for $dT1$ distances are approximately 17 residues GP(T*22)R and 7 residues for GP(T*10)L, see Figure 4.9 D-E left and right panels. In chapter 3, I showed that there was a -5 enhancement for both ppGalNAc T11 and $dT1$ in Figures 3.5 C and Figure 3.6 C and mentioned that this may be lectin domain mediated. With glycosylation results with GP(T*22)R with both ppGalNAc T11 and $dT1$, I can confirm that this -5 preference is lectin domain mediated since the $[^{3}H]$-GalNAc incorporation plots of GP(T*22)R overlap with GP(T*10)C-Ser and GP(T*10)C-Thr, suggesting that the -5 preference observed for these glycopeptides
Figure 4.9 - Subfamily If ppGalNAc T11 and dT1(PGANT35A) lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T1 and dT1 (PGANT35A) showing the net [3H]-GalNAc utilization. (B) Sephadex G10 chromatograms with (glyco)peptide substrates demonstrating high [3H]-GalNAc transfer to GP(T*22)R and high hydrolysis for selected substrates for ppGalNAc T11 and dT1. (C) Ratios of transfer GP(T*22)R vs. GP(T*10)L with ppGalNAc T11 and dT1 obtained from DPM/OD values, where N=5 for ppGalNAc T11 and dT1. (D-E) [3H]-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T11 and dT1 showing the optimal distance from a T* to the X acceptor residues is approximately 7-17 residues. Red arrow in ppGalNAc T11 and dT1 GP(T*22)R represents the closest distance the lectin domain is interacting.
represents the closest distance at which the lectin domain of ppGalNAc T11 and dT1 begins to interact, see red arrow in Figure 4.9 D-E. The fact that we can pick up this -5 preference site in GP(T*10)C-Ser and GP(T*10)C-Thr and not the +5 site, confirms that GP(T*22)R is the most preferred substrate. Because dT1 prefers to glycosylate the GP(T*22)R substrate at a higher rate than ppGalNAc T11, helps explain why there is a higher enhancement observed with the GP(T*10)C-Thr substrate at the -5 position, that is not seen with T11, see Figures 3.5 and 3.6 in chapter 3. These results suggest that the lectin domains of ppGalNAc T11 and dT1 are mostly populated to the right of the catalytic domain.

4.3.2.6 ppGalNAc T4 and T12 in Subfamily IIA

In chapter 2, I showed that ppGalNAc T4 and T12 have very similar peptide catalytic domain preferences, while in chapter 3, I showed that they both have very different glycopeptide catalytic domain preferences. The differences between peptide and glycopeptide substrates are interesting, given that their catalytic domain sequence identities are 72% identical. In this chapter, I will show that they both have the same lectin domain preferences, but have different lectin domain glycosylation properties.

For ppGalNAc T4, a total of 5 determinations were made using 125-355 µL of soluble affinity purified transferase, while for ppGalNAc T12 a total of 6 determinations were made using 68-125 µL of soluble affinity purified or transferase in beads with the same standard reaction conditions. In Figure 4.10 A-B left and right panels, both the time course plots and the Sephadex G10 chromatograms show that the most active substrate is
Figure 4.10- Subfamily IIa ppGalNAc T4 and T12 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T4 and T12 showing the net $[^3]$H-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating $[^3]$H-GalNAc transfer to glycopeptide (fraction 25-31) and high hydrolysis (35-41) for ppGalNAc T4 and T12. (C-D) $[^3]$H-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T4 and T12 showing the optimal distance from a T* to the X acceptor residues is approximately 8-9 residues. Red arrow in ppGalNAc T4 GP(T*22)R represents the closest distance at which the lectin domain is interacting. Taken from (13).
GP(T*10)L, with ppGalNAc T4 having more $[^3]H$-GalNAc incorporation than T12. As mentioned in the previous chapter, with most of the glycopeptide preferring transferases there are two peaks observed in the Sephadex G10 chromatograms; the first eluting peak corresponds to $[^3]H$-GalNAc transfer to glycopeptide substrate (fractions ~24-31) while the second eluting peak corresponds to $[^3]H$-GalNAc transfer to water (fractions ~34-41) (i.e. UDP-$[^3]H$-GalNAc hydrolysis). Therefore, the total $[^3]H$-GalNAc utilization that is represented in Figure 4.10 A (left and right panels) represents an activity that includes both transfer to (glyco)peptide and transfer to water.

The Edman amino acid sequencing plots for GP(T*22)R with ppGalNAc T4 and T12 show no peaks of $[^3]H$-GalNAc incorporation. On the other hand, for GP(T*10)L there are significant peaks of $[^3]H$-GalNAc incorporation and the glycosylation patterns are different between both transferases see Figure 4.10 C-D left and right panels. The optimal distances of glycosylation from the initial T* is 8-9 residues. As discussed in the previous chapter, I observed a +5 preference for ppGalNAc T4 with GP(T*10)C-Ser and GP(T*10)C-Thr which is not observed with ppGalNAc T12 (see Figure 3.7 in Chapter 3). The results with GP(T*10)L in Figure 4.10 D (left and right panels) suggest that ppGalNAc T4 could indeed glycosylate at the +5 position while T12 might not (i.e. T4’s $[^3]H$-GalNAc incorporation plot could be easily extrapolated to the +5 position while T12’s plot could not). This suggests that the lectin domain can modulate glycosylation as early as the +5 residue from the T* (see red arrow in Figure 4.10 D left panel) while the weaker lectin domain mediated glycosylation of ppGalNAc T12 begins at (or after) the +6 residue. Therefore, using these glycopeptides, I have obtained consistent data showing that there are differences the ppGalNAc T4 and T12 lectin domain directing properties.
Altogether, with these results it is clear that ppGalNAc T4 and T12 have mixed roles, serving to glycosylate both peptide and glycopeptide substrates, where the glycopeptide substrates are glycosylated differently by both transferases.

4.3.2.7 ppGalNAc T7, T10 and PGANT7 in Subfamily IIb

The catalytic domains of ppGalNAc T7 and T10 were shown in previous chapters to have no strong peptide preference motifs, but they do have strict and identical glycopeptide requirements (where a prior glycosylated residue is required one N-terminal position from the acceptor site). This strict glycopeptide requirement was also shown for PGANT7, the fly ortholog of ppGalNAc T7. It is interesting that this subfamily has such high glycosylation similarities given that ppGalNAc T7 shares a 48% sequence identity with ppGalNAc T10 and shares a 41% sequence identity with PGANT7. The lectin domain N- or C- directionality preferences of these three subfamily members will be examined below.

For ppGalNAc T7, a total of 5 determinations were made using 75-125 µL of transferase in beads or soluble affinity purified, for ppGalNAc T10, a total of 3 determinations were made using 105 µL of soluble transferase and for PGANT7 a total of 2 determinations were made with 300 µL of transferase in media. All reactions were carried out using the standard reaction conditions as previously described. The time course plots and the Sephadex G10 plots in Figure 4.11 A-B (left and middle panels) show that both ppGalNAc T7 and T10 have low amounts of incorporation into both GP(T*22)R and GP(T*10)L, with very high hydrolysis activity. For PGANT7 (in the
Figure 4.11- Subfamily IIb ppGalNAc T7, T10 and PGANT7 lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A22)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T7, T10 and PGANT7 showing the net $[^3]$H-GalNAc utilization. (B) Sephadex G10 chromatograms with selected (glyco)peptide substrates demonstrating $[^3]$H-GalNAc transfer to glycopeptide (fractions 25-31) and high hydrolysis (35-41) for ppGalNAc T7, T10 and PGANT7. (C-D) $[^3]$H-GalNAc incorporation into the X residues of GP(T*22)R and GP(T*10)L with ppGalNAc T7 and PGANT7 showing the optimal distance from a T* to the X acceptor residues is approximately 8-11 residues. ND stands for not determined. Taken from (13).
right panel of Figure 4.11) shows that there is equal incorporation into both GP(T*22)R and GP(T*10)L with lower amounts of hydrolysis. These results suggest that for ppGalNAc T7 and T10 there is little to no lectin domain activity with these substrates, but for PGANT7, the lectin domain may be equally populated either to the right or to the left of the catalytic domain.

The Edman amino acid sequencing for ppGalNAc T7 gave no observed patterns of glycosylation (Figure 4.11 C-D left panel), therefore, since T10 had undetectable amounts of [3H]-GalNAc incorporation these were not analyzed. The shapes and distribution into GP(T*22)R and GP(T*10)L by PGANT7 (Figure 4.11 C-D right panel) suggest that maximum glycosylation may be more distant from the T* for both GP(T*22)R and GP(T*10)L which is consistent with PGANT7 not glycosylating the +/-5 positions of GP(T*10)C-Ser (see Figure 3.9 right panel in Chapter 3).

The fact that I was not able to detect any lectin domain activity with ppGalNAc T7 and T10 could suggest that these transferases function as strictly “filling-in” transferases that glycosylate directly N-terminal of an existing S* or T*. For the first time, I was able to show a big difference between a human and fly transferase ortholog pair, where ppGalNAc T7 and PGANT7 have identical catalytic domain filling in activates, but differ in their abilities to use their lectin domains.

4.3.2.8 Non-glycosylated Controls GP(A22)R and GP(A10)L

The nonglycosylated controls GP(A22)R and GP(A10)L were designed to be controls for GP(T*22)R and GP(T*10)L. These control peptide substrates are not intended to show any lectin domain activity since they do not contain a glycosylated Thr
(T*) in the sequence. They are meant to compared against the glycosylated analogs, 
GP(T*22)R and GP(T*10)L which are intended to interact with the lectin domain. 
Edman sequencing of GP(A22)R and GP(A10)L were not successfully obtained from 
most transferases due to very low incorporation; therefore, the successfully obtained data 
from both control peptides for ppGalNAc T13 will be used in this section. The only other 
successful obtained control data, from ppGalNAc T2, was nearly identical to T13. In 
order to reveal the extent that the lectin domain may alter the pattern of $[^3]H$-GalNAc 
incorporation, the distributions of the nonglycosylated control peptides GP(A22)R and 
GP(10)L, (glycosylated by ppGalNAc T13) were subtracted from the distributions 
obtained for the glycosylated glycopeptides GP(T*22)R and GP(T*10)L, respectively. 
These difference plots determined for only eight transferases are grouped by glycopeptide 
substrate utilization and are shown in Figure 4.12 (12). It was anticipated that these plots 
would reveal the effects of lectin domain binding on the pattern of glycosylation and 
perhaps further reveal differences between transferase subfamily classes and N- and C-
terminal preferences. Except for ppGalNAc T3 with GP(T*22)R, the difference plots 
were not appreciably different from the original distribution, although in few cases, the 
distributions may have shifted slightly (for ppGalNAc T2, T5 and T16 with GP(T*22)R). 
From these results, we conclude that the binding of the glycopeptide substrate for the 
most part does not greatly alter the pattern of glycosylation compared to the 
nonglycosylated substrate. Furthermore, no consistent differences were observed between 
isoform subfamily classes or between transferases within a class that showed different N-
or C- terminal preferences (ie. ppGalNAc T2 vs. T14 and ppGalNAc T1 vs. T13). These 
findings are consistent with a highly mobile lectin domain where the lectin-bound
standard specific activity of UDP
and B, substrates were glycosylated using 0.5 µM 100
for ppGalNAc T1, T2, T3, T5, T6, T13, T14 and T16, respectively. Note that for A
and B, substrates were glycosylated with standard 50 µM
standard specific activity of UDP-[^3]H]-GalNAc, where as in (C-H), substrates were glycosylated with 100-fold specific activity of UDP-[^3]H]-GalNAc.

**Figure 4.12**- Plots of[^3]H]-GalNAc incorporation into the Xaa residues of random glycopeptides GP(T*22)R and GP(T*10)L (open bars) and after normalization by subtraction of control peptides GP(A22)R and GP(A10)L that were glycosylated by ppGalNAc T13 (filled bars). (A-H) plots of GP(T*22)R (left) and GP(T*10)L (right) for ppGalNAc T1, T2, T3, T5, T6, T13, T14 and T16, respectively.

![Plots of ^[3]H]-GalNAc incorporation into the Xaa residues of random glycopeptides GP(T*22)R and GP(T*10)L (open bars) and after normalization by subtraction of control peptides GP(A22)R and GP(A10)L that were glycosylated by ppGalNAc T13 (filled bars). (A-H) plots of GP(T*22)R (left) and GP(T*10)L (right) for ppGalNAc T1, T2, T3, T5, T6, T13, T14 and T16, respectively. Note that for A and B, substrates were glycosylated with 50 µM standard specific activity of UDP[^3]H]-GalNAc, where as in (C-H), substrates were glycosylated with 100-fold specific activity of UDP[^3]H]-GalNAc.](image-url)
glycopeptide substrate would have a broad range of acceptor interacting distances with the catalytic domain. Alternatively, these results could be consistent with just a regiospecific increase in substrate concentration simply due to rapid binding and release from a conformationally fixed lectin domain.

4.3.3 Lectin Domain Splice Variant Studies on ppGalNAc T13 & T13 Ex10B and PGANT9A & PGANTB

Our lab is collaborating with several other labs (Osinaga, Clausen and Ten Hagen Labs) that are studying the human ppGalNAc T’s and the fly PGANT’s. Of particular interest, these groups are specifically focused on comparing lectin domain splice variants, which have been identified in both the human and fly. One particular splice variant is in human ppGalNAc T13 called ppGalNAc T13-V1 (or ppGalNAc T13 Ex10B). In the fly transferases, two lectin domain splice variants PGANT9A and PGANT9B have also been found from the main gene CG30463 which encodes for PGANT9. Since our work with the lectin domain probing glycopeptide substrates has proven to be successful in determining directionality preferences among transferase isoform, we can now expand our studies to examine these lectin domain splice variants and determine if they are different. Each splice variant pair will be discussed in the sections below.

4.3.3.1 ppGalNAc T13 vs. T13 Ex10B

Studies by Raman et al. (30) with ppGalNAc T13 showed that there are two splice variants, variant one called ppGalNAc T13-V1 and variant two called ppGalNAc T13-V2. In ppGalNAc T13-V1, one half of the lectin domain is altered, due to a 198bp intron
between exons 10 and 11 not being spliced out. This variant has also been identified in a library of human and mouse cDNA sequences pooled from various tissues (31). The subdomains that are conserved are the α and β subdomains, where the important binding/interacting cysteine residues in the CLD motif as well as the QXW are conserved. The other slice variant, ppGalNAc T13-V2 is truncated at exon 6 resulting in an incomplete catalytic domain, where no activity is expected. These splice variants were tested with a panel of substrates that included the EA2 peptide and MUC5AC (glyco)peptides (see Table 1 of (30)) and they reported that ppGalNAc T13 and T13-V1 were active on a wide range of peptides with little difference between the activities on the peptides and glycopeptides. Variant 2 on the other hand, showed no activity. While the activities were the same for ppGalNAc T13 and T13-VI, the observed $K_m$ values with ppGalNAc T13-V1 with glycopeptide substrates MUC5AC-3 and MUC5AC-13 were ~8 fold lower and ~3 fold lower, respectively compared to ppGalNAc T13. Their site preferences however, were identical.

ppGalNAc T13 has been shown to be highly and restrictively expressed in the neurons of the brain and expressed in all neuroblastoma cells (24) and has been shown to be involved in the metastatic process of some tumors. This involvement is thought to be due to the specific glycosylation of syndecan-1 by ppGalNAc T13. Studies by Berios and co-workers (32) showed that GALNT13 (gene for ppGalNAc T13) was upregulated in patients with neuroblastoma and that the GALNT13 expression in the bone marrow of patients is a strong predictor of poor clinical outcome. The Osinaga Lab from the Universidad de la Republica Montevideo in Uruguay, is studying the catalytic function of ppGalNAc T13 and its biological relevance to tumor progression. Recently, they have
identified nine GALNT13 splice variants that differ in their catalytic and lectin domains or both (unpublished). Since our glycopeptide substrates are designed for probing catalytic and lectin domain functionality, we were approached to collaborate on this project. We were asked to specifically compare the lectin domain activities of ppGalNAc T13 and one splice variant ppGalNAc T13 Ex10B (known as ppGalNAc T13-V1 from (30)). We were only supplied with the WT and ppGalNAc T13 Ex10B splice variant since it was demonstrated to have the highest activity out of the other splice variants. Using our (glyco)peptide substrates, I was able to determine that the activities of the WT and Ex10B variant were essentially the same with both of our lectin glycopeptide substrates.

For these studies, a total of 2 determinations were made with each transferase using 100/150 µL of ppGalNAc T13 and T13 Ex10B using the same standard reaction conditions. Only three time point aliquots of 120 min (2hr), 240 min (4hr) and 1200 min (overnight) were obtained. The resulting reaction time course plots and accompanying Sephadex G10 chromatograms (overnight incubations only) are shown in Figures 4.13 A and B left and right panels. Both plots are in agreement that both ppGalNAc T13 and ppGalNAc T13 Ex10B show equal transfer to both GP(T*22)R and GP(T*10)L, with T13 Ex10B displaying reduced activities. The glycosylation products of GP(T*22)R and GP(T*10)L glycosylated with ppGalNAc T13 and T13 Ex10B obtained were counted for [3H]-GalNAc and monitored for OD220 and OD280. These values were then divided; [3H]-DPM/OD220 and [3H]-DPM/OD2280 to obtain ratios to better quantify the actual [3H]-GalNAc content of each glycopeptide. The values that were obtained for each glycopeptide, GP(T*22)R and GP(T*10)L and then ratioed and shown in Figure 4.13 C.
Figure 4.13- Splice variants ppGalNAc T13 and T13 Ex10B lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with ppGalNAc T13 and T13 Ex10B showing the net $[^{3}H]$-GalNAc utilization. (B) Sephadex G10 chromatograms with (glyco)peptide substrates demonstrating high $[^{3}H]$-GalNAc transfer to GP(T*22)R and GP(T*10)L and low hydrolysis for selected substrates for ppGalNAc T13 and T13 Ex10B. (C) Ratios of transfer GP(T*22)R vs. GP(T*10)L with ppGalNAc T13 and T13 Ex10B (normalized to DPM/OD). For ppGalNAc T13, N=6 and for T13 Ex10B, N=8. (D) ppGalNAc T1, T13 and T13 Ex10B lectin subdomain sequences. Residues in red and green represent the conserved residues while the residues in bold represent the different residues (Festari et al. unpublished).
where the ratios of GP(T*22)R and GP(T*10)L of both transferases are nearly identical. Note that there were large standard deviation errors are due to the variability of the OD’s for the glycopeptide substrates, which were often times difficult to obtain due to their low absorbances at 220 and 280nm. Still, this data was useful in determining the similarities of the splice variants.

In chapter 2, I showed that both ppGalNAc T1 and T13 have nearly identical catalytic domain specificity and mentioned that they share an 88% sequence identity. Here, I showed that they have a different directionality preferences (see Figure 4.5) where ppGalNAc T1 has a preference for GP(T*22)R and T13 has a preference for both GP(T*22)R and GP(T*10)L. Previous mutational studies on the lectin domain of ppGalNAc T1 showed that the \( \alpha \) and \( \beta \) subdomains are the active domains in glycopeptide binding, while the \( \gamma \) subdomain is not involved (4,33). This could also be true for ppGalNAc T13 and its splice variant Ex10B, where the \( \gamma \) subdomain might not be involved in glycopeptide binding but the \( \alpha \) and \( \beta \) subdomains may be. Since the ppGalNAc T13 Ex10B splice variant exists in the lectin \( \gamma \) subdomain, one would expect to see different lectin domain properties but since none were observed, this suggests that the \( \gamma \) subdomain is not involved in glycopeptide binding this could explain why both ppGalNAc T13 and Ex10B still share the same identical preferences (i.e. either the \( \alpha \) and \( \beta \) subdomains are involved in glycopeptide binding). It is interesting though that the residues in the \( \alpha \), \( \beta \) and \( \gamma \) subdomains are conserved in ppGalNAc T1 and T13 see Figure 4.13 D.
4.3.3.2 PGANT9A vs. PGANT9B

The lectin domain studies will compliment the almost identical catalytic domain peptide preferences shown for PGANT9A and PGANT9B in chapter 2 in further examining the different functionalities of the splice variants.

For these studies, a total of 7 determinations were made using 100/180 µL of PGANT9A and PGANT9B respectively in cell culture media using the standard reaction conditions. Only three time point aliquots of 120 min (2hr), 240 min (4hr) and 1200 min (overnight) were removed. The resulting reaction time course plots and accompanying Sephadex G10 chromatograms (overnight incubations only) are shown in Figures 4.14 A and B. Both plots show that there are differences in the activities for GP(T*22)R and GP(T*10)L between the transferases. PGANT9A has a higher preference towards the GP(T*10)L while PGANT9B has a higher preference for GP(T*22)R. This is further shown in Figure 4.14 C where the ratios of GP(T*22)R and GP(T*10)L of both transferases show that PGANT9A prefers GP(T*10)L slightly more (with a ratio of ~0.6) while PGANT9B has a higher preference for GP(T*22)R (with a ratio of ~1.6).

Why the differences in preferences are observed could be explained by their lectin domain sequences. Figure 4.14 D shows the sequence alignment of the α subdomains of PGANT9A and PGANT9B. While the CLD and QXW motifs are conserved in both, the main differences are in the lengths and the conservation of amino acids. PGANT9A is 3 residues longer than PGANT9B and the sequences differ in small regions within the α subdomain where 17 residues are not conserved. Whether these differences are the reason that the change the directionality preferences of the lectin domains of PGANT9A and PGANT9B remains to be further studied. Therefore, no conclusions can be made from
Figure 4.14- Splice variants PGANT9A and 9B lectin domain preferences. (A) Representative time course plots for overnight incubations with GP(T*22)R (red), GP(A10)R (purple), GP(T*10)L (blue) and GP(A10)L (light blue) with PGANT9A and PGANT9B showing the net $[^3]$H-GalNAc utilization. (B) Sephadex G10 chromatograms with (glyco)peptide substrates demonstrating high $[^3]$H-GalNAc transfer to GP(T*22)R and GP(T*10)L and low hydrolysis for selected substrates for PGANT9A and PGANT9B (B). (C) Ratios of transfer GP(T*22)R vs. GP(T*10)L with PGANT9B and PGANT9B (normalized to DPM/OD). For PGANT9A, N=12 and for PGANT9B, N=14. (D) Cartoon depicting the location of the splice and $\alpha$ subdomain sequence alignment for PGANT9A and PGANT9B. CLD and QXW motifs are underlined in red. Conserved residues are marked with a *.
the sequence alignments, more work is needed in order to address this. This work represents the first difference ever observed with a transferase where alternative splicing in the lectin domain causes a change in the glycosylation preferences.

4.5 Summary and Conclusions

In this work, I have characterized ten of the peptide-preferring ppGalNAc T transferases, four of the glycopeptide-preferring ppGalNAc T transferases and four fly PGANT orthologs against this series of lectin domain probing glycopeptides listed in Table 4.1. My findings have un-ambiguously revealed that prior remote GalNAc-O-Thr substrate glycosylation can be recognized by these transferases in a specific N- or C-terminal direction that varies with ppGalNAc T isoform see summary Figures 4.15 and 4.16. For all of the transferases characterized (except for ppGalNAc T7 and T10), at least one of the glycopeptide substrates, GP(T*22)R and GP(T*10)L shows significant elevated activity over its nonglycosylated control. This specific N- or C-terminal elevated activity is attributed to the binding of the glycosylated Thr residue (T*) of the substrate to the lectin domain in such a manner that the acceptor region of the substrate is oriented at the catalytic domain for optimal glycosylation (Figure 4.17 A). This is clearly a mode of substrate recognition that must be mediated by the lectin domain (and not entirely by the catalytic domain as shown in chapter 3). Therefore, this mode of substrate recognition should be best be termed, a lectin domain-assisted activity, which operates together with the catalytic domain in an ordered fashion (Figure 4.17 B). Additionally, these findings strongly suggest that the tethered lectin domain of the ppGalNAc T’s may be mobile and its location relative to the catalytic domain varies among isoforms. Such
Figure 4.15- Efficiency of GalNAc transfer to acceptor glycopeptides varies with ppGalNAc T family. Bars represent (i.e. efficiency) of GalNAc transfer to the indicated glycopeptide substrates: GP(T*22)R (red) and GP(T*10)L (blue). Left grouping of the bars represents the “peptide-preferring” subfamily I transferases, ppGalNAc T1, T2, T3, T5, T6, T11, T13, T13 Ex10B, dT1, PGANT9A and PGANT9B, while the right grouping of bars represents the “glycopeptide-preferring” subfamily II transferases ppGalNAc T4, T7, T10, T12 and PGANT7. Values are obtained from 1 to 3 gel filtration runs of standard overnight reactions. Note that these values are displayed to reveal trends and may vary with the transferase reaction conditions. Error bars represent standard deviations from 2 to 4 experiments.
Figure 4.16 - Correlation of glycopeptide specificities with the ppGalNAc T phylogenetic family tree. To the right of the ppGalNAc T subfamily phylogenetic tree (44) are the glycopeptide preferences obtained from this work and work from chapter 3. Glycopeptide substrate cartoons representing GP(T*22)R are shown in red (left column), GP(T*10)L in blue (middle column) GP(T*10)C-Ser in green where the T* represents the position of the initial GalNAc-\(\text{O-Thr/Ser}\), while the arrows indicate the position(s) glycosylated by the indicated transferase isoform. Note that the arrows for GP(T*22)R and GP(T*10)L represent a broad distribution of glycosylation. The arrows in parenthesis represent the proposed +/- 5 lectin-mediated glycosylation site preferences. The grey arrows indicate weak lectin domain interactions. Transferases whose preferences have not been determined against these substrates are left as a blank space.
Figure 4.17- Cartoon and X-ray crystal modeling for the ppGalNAc T lectin domain-assisted property. (A) Model building of (glyco)peptide substrates GP(T*22)R and GP(T*10)L onto the EA2-bound ppGalNAc T2 crystal structure (9). The left panel shows GP(T*22)R modeled within its Xaa residues placed in the catalytic domain in the same N- to C- terminal orientation as the 9 residues of the bound EA2 peptide. The right panel shows GP(T*10)L modeled into the catalytic domain. For simplicity, the 9 EA2 residues were maintained as in the original structure, whereas 3 additional residues were added to complete the representation of the 12 Xaa residues, and a Ser residue (green) was used to represent the location of the GalNAc-O-Thr. Note that for simplicity, both regions were modeled as static extended structures but in reality would be flexible, comprising an ensemble of structures. (B) Schematic representation of the lectin domain-assisted property activities of the ppGalNAc T’s (upper panel: ppGalNAc T1, T2, T5, T11, T13, T14, T16, PGANT7, PGANT9A, PGANT9B and dT1, lower panel: ppGalNAc T3, T4, T5, T6, T11, T13, T16, PGANT7, PGANT9A and dT1). (C) Superimposition of the catalytic domains of ppGalNAc T2-bound-EA2 (yellow trace) (2FFU), ppGalNAc T2-bound-UDP (blue trace) (2FFV), ppGalNAc T10-bound-UDP and GalNAc (red trace) (2D71) and ppGalNAc T1 (green trace) (1XHB) showing the positional variability of the lectin domain. The critical lectin α-subdomain Asp residue is space-filled pink (12).
domain mobility is supported by the superimposition of the X-ray crystal structures of ppGalNAc T1, T2 and T10 (1,6,9), shown in Figure 4.17 C. For those transferases with similar N- and C-terminal glycopeptide enhancements, ppGalNAc T5, T13, T16, PGANT7 and PGANT9A, it is possible that the lectin domain may be sufficiently mobile that it can enhance glycosylation from both N- and C-terminal direction. Alternatively, the lectin domain may serve to equally increase the N- and C-terminal glycopeptide substrate concentrations by a simple bind and release mechanism, as shown for other glycosyltransferases with carbohydrate binding molecules (34-37). Further evidence of a highly flexible lectin domain may be found in the broad distributions of glycosylation that is observed in the Xaa region of both the glycosylated and non-glycosylated substrates. The only significant alteration in distribution is observed for ppGalNAc T3 glycosylating its non-preferred glycopeptide substrate GP(T*22)R (see Figure 4.7 C left panel), suggesting that the lectin domain is involved to some extent in directing the glycosylation of this substrate. Further studies are needed in order to fully understand the dynamics of the lectin domain of these transferases.

These studies have also revealed that the peptide-preferring transferases of subfamilies Ia-Id and If, all have abilities to glycosylate glycopeptide substrates with remote sites of glycosylation and are therefore, not strictly peptide preferring transferases. In order to address the overlapping peptide/glycopeptide substrate preferences, we suggest that for clarity, these subfamily members be reclassified as remote glycopeptide/peptide-preferring transferases, abbreviated as GP/P-preferring.

In an attempt to correlate the different N- and C-terminal glycopeptide substrate specificities to the transferase peptide sequence, we have compared the sequences of their
linker domain and the three lectin domain binding motifs, shown in Table 4.2. From the alignment of the linker domain, there are no obvious differences in length or sequence that can readily account for the different behavior observed within a class. For example, for family I members, ppGalNAc T2 and T14, which prefer glycopeptide GP(T*22)R are only one residue shorter than T16, which shows preferences for both glycopeptide substrates, whereas the linker domains of ppGalNAc T1 and T13 are identical except for two conservative substitutions, although they show different glycopeptide specificities. In contrast, the linker domains of ppGalNAc T3 and T6 are only 50% identical but have the same elevated preferences with GP(T*10)L. Family If members, ppGalNAc T11 and dT1, the residues in the linker regions are less conserved yet the preferences for GP(T*22)R remain the highest for both transferases. The same is true for subfamily II members, where there are no obvious differences in sequence or length. For example, the PGANT7 linker shows the greatest similarity to the ppGalNAc T10 linker, compared with T7, despite the fact that PGANT7 and T10 have significantly different lectin-domain assisted properties. The mixed (glyco)peptide-preferring subfamily members, ppGalNAc T4 and T12 (IIa) linkers appear to be more similar to the remote glycopeptide/peptide class I linkers than to the IIb members which are all ~5 residues longer.

Numerous studies on the binding lectin motifs on the α, β and γ repeat subdomains across multiple organisms have revealed binding motifs, including the CLD and QXW sequences (38-41). Mutagenesis studies on the lectin domains of ppGalNAc T1 (D444A), T2 (D458H) and T3 (D519H) have revealed that the Asp residue of the α-subdomain CLD motif is typically required for lectin binding activities (3,14,34,42,43). A similar mutation in the ppGalNAc T1 β-ricin subdomain (D484A) only modestly
A) Linker domain (L) residues shaded in dark grey are conserved (or similar) in the majority of the 13 transferases shown, while residues shaded in light grey are conserved only within a given transferase subfamily.
B) α, β and γ ricin lectin subdomain residues that are identical to the canonical CLD-QXW motif are shaded in green, while similar canonical motif residues are shaded in yellow. Non-canonical (and presumably inactivating) residues are shaded in red.
C) Summary column: green and yellow backgrounds represent the presence of the CLD-QXW canonical motif or a similar motif respectively that could be expected to bind sugar from prior studies. A red background indicates a likely inactive sugar binding motif. Subdomains labeled with + represent experimentally confirmed sugar binding motifs (1-7).
D) Proline residues in the linker regions are marked in red based on the assumption that the Pro residues are important for the linker dynamics (25).

### Table 4.2 - ppGalNAc T linker and lectin subdomain alignments.

<table>
<thead>
<tr>
<th>Class</th>
<th>Linker (L)</th>
<th>α-ricin b</th>
<th>β-ricin b</th>
<th>γ-ricin b</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>catalytic(C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>PDPOSITORY</td>
<td>PDPOSITORY</td>
<td>PDPOSITORY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T13</td>
<td>PDPOSITORY</td>
<td>PDPOSITORY</td>
<td>PDPOSITORY</td>
<td></td>
</tr>
<tr>
<td>Class Ib</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>PE_LRVBDQDFAFALQO---GTN_CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>T4</td>
<td>PE_BGRQKVMGRNLKGYRNGQﾄ---PLS_QCLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>Class Ic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>PE_VYVLNLFLVTVQIEKEVQPL---SCLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>T6</td>
<td>PE_MPVGDLTTFYAVKEKLGQTN---SCLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>Class Id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>PD_EIAIRVAS---QVI_AVALGK---</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>Class If</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T11</td>
<td>PE_MGSCGPVLQVGRQYVHLG---NK_CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>dT1</td>
<td>PE_EHNGVRYVDFTPQRLTG---EL_EAA</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>Class IIa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>PE_LNVYEDRTQWGDALAREG---SE_CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>T12</td>
<td>PE_LNVYEDRTQFRMAG---QXW_DY_CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>Class IIb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7</td>
<td>PD_ECSLTVPRMTTQ---CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>T10</td>
<td>PD_ECSLTVPRMTTQ---CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
</tr>
<tr>
<td>PGANT7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YD_WYDKFLGL_PANLHGEIRLRVASDG---CLD</td>
<td>PD</td>
<td>CLD</td>
<td>CLD</td>
<td>α⁺</td>
<td></td>
</tr>
</tbody>
</table>
decreased the lectin binding activity (4,43), whereas mutation of the γ-ricin subdomain of ppGalNAc T1 and T2 (D525A and D541A, respectively) shows no effects on their lectin activity (2). As shown in Table 4.2, the β-subdomains of ppGalNAc T2, T14, T16, T3, T4, T5, T6, T7, dT1, T12 and PGANT7 and the γ-subdomains of T3, T4, T6, T11, dT1, T12 and PGANT7 lack this Asp. Several transferases possess a Glu instead of Asp, such as ppGalNAc T14 and T16 in the α- subdomain and T5 and T16 in the γ subdomain, however, whether these domains can bind GalNAc is presently unknown. The β-subdomains of ppGalNAc T6, T5, T4, T12, T7 and PGANT7 lack the QXW motif and would not be expected to be active. Although I was not able to detect any lectin domain activity with ppGalNAc T10, there is a co-crystal structure of T10 that shows a GalNAc-O-Ser bound to its β-subdomain (having canonical motifs CFD and QLW) (6). The ricin subdomain motifs of ppGalNAc T5 would be expected to have the weakest lectin binding activities because all three subdomains lack the critical Asp residue (although in the γ-subdomain, the Asp is replaced by a Glu). Nevertheless, because ppGalNAc T5 displays clear glycopeptide specificities, one or more of its lectin subdomains (perhaps its γ-subdomain) must possess significant binding activity to provide the glycopeptide enhancements. This would be a difficult task to prove, since glycopeptide binding to any of the ppGalNAc T’s lectin domains has been difficult to detect directly, further suggesting its weak nature (3,8,14). This is sensible, since tight lectin domain binding of glycopeptide substrates would likely be non-productive, slowing glycosylation due to a delayed product release. With this analysis of the linker domains, we conclude that are no obvious correlation between the likely activities of specific lectin subdomain motif and a given transferase’s observed glycopeptide activity.
From the above discussion of the lectin domain motifs listed in Table 4.2, it is likely that for all transferases, except for ppGalNAc T5, that binding would occur at the α-subdomain. In Figure 4.17 C, the superimposed crystal structures of ppGalNAc T1, T2 and T10 show the critical Asp residue of the CLD motif in the α-subdomain (highlighted in the structures) is found in a wide range of positions relative to the catalytic domain. In solution, we would assume that the conformational flexibility of the lectin domain relative to the catalytic domain would be greater than what is observed in Figure 4.17 C. Altogether, both the structural and experimental results are consistent with a highly mobile lectin domain, whose dynamics and relative orientation with respect to the catalytic domain varies widely with transferase isoform. Whether additional domain-domain interactions may modulate the relative positions of the lectin domain to the catalytic domain is presently unknown. It is clear that further structural studies are needed in order to fully understand the roles of the lectin domains in these transferases.

The recently reported X-ray crystal structures, small angle X-ray scattering and solution modeling studies of ppGalNAc T2 interacting with the MUC5AC glycopeptides (25) is an important step in this direction, confirming the dynamic flexibility between the ppGalNAc T2 catalytic and lectin domains. In this work, it was confirmed that ppGalNAc T2 has a preference for a glycopeptide with a C-terminal prior site of glycosylation and that the overall glycosylation profile had a broad distribution with a maximum site of incorporation that was 10 residues from the prior site of glycosylation. They have revealed a detailed mechanism on how ppGalNAc T2 selectively glycosylates unused acceptor sites that were located on the N-terminal end of the substrate. It was shown that ppGalNAc T2 populated an ensemble of compact (monomeric) and extended
(dimeric) structures. In the presence of the acceptor substrate, the dimers disappear as the catalytic and lectin domain are compacted together and form a monomeric structure. The authors suggested that the formation of the monomeric structure is critical for enzymatic activity and that his flexibility was due to the flexible linker. The authors determined that the role of the flexible linker is to control the different conformations of the two domains, allowing the catalytic domain to glycosylate the different sites at different distances. Experiments with the linker swap-outs are currently in progress.

The control of site-specific mucin type O-glycosylation and the need for such a large family of initiating ppGalNAc T’s is not well understood. Previous work has shown that ppGalNAc T1 and T2 are ubiquitously expressed in nearly all mammalian tissues and cell lines, where as the remaining ppGalNAc T isoforms studies in this work are more selectively expressed (44-46). For example, ppGalNAc T3 is highly expressed in the testis and kidney, while ppGalNAc T16 is more expressed in the heart than T1 and ppGalNAc T13 is specifically expressed in the neurons (19,24,26,30,46). How the multiple transferases work in concert in glycosylating their target proteins is largely unknown, as is their regulation at both the protein and transcriptional level. With this work, we have identified a previously unappreciated level of control, where by remote prior O-glycosylation is used to target and enhance the glycosylation of specific N- or C-terminal sites in an isoform specific manner. Our studies and those of others strongly suggest that this N- or C- terminal selectivity is due to weak glycopeptide binding to the lectin domain, whose orientation relative to the catalytic domain is highly mobile and isoform-dependent. This glycopeptide selectivity can provide an additional level of control or fidelity for the glycosylation of biologically significant sites and suggests that
$O$-glycosylation in some instances may be exquisitely controlled. Furthermore, our observations that homologous ppGalNAc T isoforms within a given subfamily (that presumably have similar catalytic domain peptide substrate specificities, see chapter 2) may possess different N- or C- terminal glycopeptide preferences and even neighboring filling in glycopeptide activities (see chapter 3) which could explain the large number of ppGalNAc T family members, thereby maintaining peptide specificity while altering glycopeptide specificity. These studies clearly demonstrate that the biological control of mucin type O-glycosylation is highly complex and that further structural, biochemical, and biological studies are necessary to fully understand this important modification.
4.6 References


carcinoma antigen, by the lectin SNA-II from *Sambucus nigra*. Proteins: Structure, Function and Bioinformatics 75, 89-103


CHAPTER 5

CHARACTERIZATION OF SOMATIC AND GERM-LINE MUTATIONS ON GLYCOSYLTRANSFERASE GENES B3GNT2, B4GALT2, ST6GALNACII AND GALNT12
5.1 Background and Significance

Alternations or defects in glycosylation pathways has been consistently shown to be involved in many human cancers especially colon cancer (1-3). Normal colonic mucins have a wide range of O-glycan structures with high levels of glycosyltransferases; both initiating and elongating that produce the Tn-antigen and the common core structures 1-4 respectively (Figure 5.1). The roles of these core structures have been reported to be involved in the homeostasis of the intestines (4-6), with core 3 O-glycans being the most expressed in gastrointestinal mucosa and the major mucin-type O-glycan in colonic tissues (7-9). Previous work has shown that the expression of several ppGalNAc T’s (that synthesize the Tn-antigen) is increased not only in colon cancer but other carcinomas (11). In fact, many specific ppGalNAc T transferases have been shown to be involved in cancers, see chapter 1. The sialylated form of the Tn-antigen known as the STn-antigen (Figure 5.1), that is produced by the enzymes ST6GalNAc-I and II, have linked to the progression of many cancers (12). This sialic acid modification is O-glycan chain terminating, as it stops the elongation of the Tn-antigen. Both the Tn- and STn-antigen have been shown to be markers for poorly differentiated carcinomas and their high occurrence has been associated with the advanced progression of cancer, increased invasive and proliferative tumors, metastasis and poor clinical outcome (13). Moreover, the mRNA levels of ST6GalNAcII has been shown to be increased in cases of colorectal cancer (14).

The synthesis of the core structures 1-3 in colon cancer is altered; in normal colonic tissues the O-glycan mucins display mainly core 3 structures, while in cancerous colonic tissues, there is a high prevalence of core 1 and core 2 structures. The main
enzyme responsible for producing the core 3 structure, β3-N-acetylglucosamine (Core 3 GlcNAc T or Core 3 synthase) has been shown to be reduced in colon cancer, at dramatically low detection levels in cultured colon cancer cells (15). The absence of the core 3 enzyme not only increases the likelihood of Tn-antigens but also decreases the formation of the Core 4 structure see Figure 5.1 (1). Additionally, loss of Core 3 synthase has been shown to enhance the metastatic potential of colon carcinoma cells (4), while studies in the mouse show that Core 3 synthase deficient mice display reduced production of colonic MUC2 mucin and show increased susceptibility to colitis and colon adenocarcinoma (16,17). Cores 1-3 can be further elongated by other glycosyltransferases, producing polylactosamine chains. Polylactosamine (or Poly-N-acetyllactosamine) chains are unique glycans that arise from repeating N-acetyl-D-lactosamine (LacNAc) repeats comprised of galactose and N-acetylglucosamine sugar units (i.e. \([\beta\text{Gal}(1\rightarrow4)\beta\text{GlcNAc}(1\rightarrow3)]_n\text{-R}) where R can be glucose or N-acetylglucosamine from the Core structure) (18). These can be attached to N- and O- linked glycans as well as glycolipids. In the adenocarcinomas of colorectal cancers, two types of polylactosamine chains are produced, they are type 1: \([\beta\text{Gal}(1\rightarrow3)\beta\text{GlcNAc}(1\rightarrow3)]_n\) and type 2: \([\beta\text{Gal}(1\rightarrow4)\beta\text{GlcNAc}(1\rightarrow3)]_n\) (1). Examples of glycosyltransferases that synthesize these structures are the β4GalT (type 2 only), β3GalT (type 1 only) and β3GnT families, Figure 5.1. The repeated actions of these glycosyltransferases generate the polylactosamine sequence. It has been shown in colon cancer that the activities of β4GalT and β3GnT transferase families are up-regulated while β3GalT transferase family members are down-regulated (1,18,19).
Figure 5.1- Protein glycosylation pathways involving ppGalNAc T’s, β3GnT2, β4GalT4 and ST6GalNAcII. Shown is the initiation of mucin-type O-glycan by the ppGalNAc T’s, producing the Tn-antigen. Synthesis of the cores 1-3 are catalyzed by the C1GalT1, CnGnT1-3 and β3GnT6 glycosyltransferases (respectively). Synthesis of polylactosamine structures on tetraantennary N-linked glycans and on the core 1, core 2 and core 3 O-glycans, catalyzed by β3GnT2 and β4GalT2. Glycosylation by ST6GalNAcII of the Tn-antigen, core 1 and core 3 results in chain termination.
The Gerken, Guda and Markowitz labs have been collaborating for many years on several studies of glycosyltransferases and their roles in cancer. Of most interest is colon cancer, where several mutant glycosyltransferases (especially GALNT12) genes have been identified from tissue samples, or DNA samples from colon cancer patients (3,20). GALNT12 has been of interest to the Guda Lab for some time due to its high levels of expression in the normal colon (21) and because previous work by the Guda Lab identified eight somatic and germ-line mutations of GALNT12 out of 272 colon cancer patients (3). These mutants were: M1I, Y395X, R382H, T491M, R373H, R297W and D303N and the locations of these mutations were throughout the catalytic domain and even as early as an initiating codon (M1I). The Guda Lab expressed these mutated transferases and the Gerken Lab examined the mutant enzymatic activity. The impact of these mutations on the encoded ppGalNAc T12 enzyme were tested against the EA2 peptide (PTTDSTTPAPTTK) and MUC5AC peptide (GTTPSPVPTTSTTSA) and found that the mutations encoded inactive enzymes or enzymes with reduced activities (3). It was concluded that the reduced activities of these mutant ppGalNAc T12 enzymes would lead to improper glycosylation (i.e. under glycosylation) of proteins produced in colon cancers. Because of this study, the Woods Lab in Newfoundland Canada became interested in this type of work and screened for deleterious GALNT12 sequences in a cohort of 517 patient DNA samples from The Newfoundland Colorectal Cancer Registry (NFCCR). By Sanger sequencing, they identified an additional 8 germ-line missense GALNT12 mutations; H101Q, I142T, E239Q, T286M, V290F, R297W, D303N and Y396C, Table 5.1. All of these mutations were specific to one patient, while D303N was detected in 3 patients. The Woods Lab was also interested in testing the impact of the
encoded mutations on the protein activity (just like in the previous paper) and because this work requires the efforts from both the Guda and Gerken Labs, another collaboration was formed.

To further this work, the Guda Lab has also sequenced 430 glycosylation-associated genes in a series of 31 colon cancer (CRC) patient-derived cell lines and identified 12 new significantly mutated glycosylation associated genes. Of upmost interest were the enrichment of the somatic mutations in genes, B3GNT2 (3 mutants), B4GALT2 (one mutant) and ST6GALNAC2 (two mutants), which accounted for 16% of the CRC cell lines tested, see Table 5.1. These three genes encode for the glycosyltransferases; β3GnT2, β4GalT2 and ST6GalNAcII which involved in the elongation and termination of the N- and O-linked cores 1-3 glycans in the colon. This work was recently published (10).

The Gerken and Guda Labs systematically characterized the enzymatic and biological functional consequences of each of the identified CRC-associated mutant glycosyltransferases which showed that these mutations have effects on either (one or more): enzymatic activity, transferase localization, post-translational modification and/or the migratory potential of colon cancer cells. Altogether, these findings have uncovered new potential contributors to aberrant glycosylation and contribute to the pathogenesis of molecular subsets of colon and other gastrointestinal malignancies.

In this chapter, I will discuss my contribution to this work.
Table 5.1 – Somatic mutations found in CRC patients ((10) and unpublished).

<table>
<thead>
<tr>
<th>Glycosyltransferase (Gene)</th>
<th>Glycosyltransferase (Protein)</th>
<th>Protein Change (Mutation)</th>
<th>Variant Class</th>
<th>Location on Transferase</th>
<th>SIFT/Polyphen Prediction</th>
<th>Loss of WT allele on Cancer DNA/RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3GNT2</td>
<td>β3GnT2</td>
<td>R6X</td>
<td>Nonsense</td>
<td>TM Domain</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P180T</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D247H</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td>B4GALT2</td>
<td>β4GalT2</td>
<td>A146V</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Deleterious</td>
<td>Yes</td>
</tr>
<tr>
<td>ST6GALNAC2</td>
<td>ST6GalNAcII</td>
<td>D43H</td>
<td>Missense</td>
<td>Stalk</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td>GALNT12</td>
<td>ppGalNAc T12</td>
<td>H101Q</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y142T</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E239Q</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T236M</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V290F</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R297W</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D303N</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Neutral/Damaging</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y396C</td>
<td>Missense</td>
<td>Catalytic Domain</td>
<td>Deleterious</td>
<td>No</td>
</tr>
</tbody>
</table>
5.2 Materials and Methods

All procedures prior to the glycosylation assays, including the collection of patient samples, the targeted re-sequencing of genes, the mapping, filtering, detection, selection of candidate somatic mutations and the expression of the constructs, the western blots, de-glycosylation experiments and migratory experiments were all performed by the Guda Lab and only a few of their methods will be discussed here.

5.2.1 Transferases

All transferases were supplied by the Guda Lab and are described in Table 5.2. WT and mutant(s) constructs of B3GNT2, B4GALT2 and ST6GALNAC2 were expressed in COS7 cells and the cloning vectors were pcDN3.1/V5-HIS TOPO (for B3GNT2 and B4GALT2) and pIHV modified SV40 promoter-driven pZeoSV2 vector (for ST6GALNAC2) (10). The recombinant protein constructs were purified by immunoprecipitating from the lysates using anti-V5 agarose beads, subsequently washed with buffer and supplied as transferase bound to beads. WT and mutant(s) constructs of ppGalNAc T12 were expressed in SW480 cells (a human colon cancer cell line) and the cloning vector pIHV. The recombinant constructs were purified by immunoprecipitation from the cell culture medium using anti-V5 agarose beads, subsequently washed with buffer and supplied as transferase bound to beads. Typically for these assays the transferases were used immediately on the same day and were not stored. Control (CONT) samples were obtained where the COS7 cells were not transfected with the respective cDNA sequences.
**Table 5.2** – Table of glycosyltransferases used in this work describing the construct and expression.

<table>
<thead>
<tr>
<th>Glycosyltransferase</th>
<th>Construct</th>
<th>Expression System</th>
<th>Vector</th>
<th>Secretion/Purification</th>
</tr>
</thead>
<tbody>
<tr>
<td>β3GnT2</td>
<td>WT CONT P186T D247H</td>
<td>COS7</td>
<td>pcDNA3.1/V5-HIS TOPO</td>
<td>Beads</td>
</tr>
<tr>
<td>β4GalT2</td>
<td>WT CONT A146V</td>
<td>COS7</td>
<td>pcDNA3.1/V5-HIS TOPO</td>
<td>Beads</td>
</tr>
<tr>
<td>ST6GalNAcII</td>
<td>WT CONT D43H R115W</td>
<td>COS7</td>
<td>pIHV</td>
<td>Beads</td>
</tr>
<tr>
<td>ppGalNAc T12</td>
<td>WT CONT H101Q I142T E239Q T286M V290F R297W D303N Y396C</td>
<td>SW480</td>
<td>pIHV</td>
<td>Beads</td>
</tr>
</tbody>
</table>
**Western Blots of Expressed Recombinant Transferases**

The western blots for the expressed transferases were performed by the Guda Lab. Briefly, after immunoprecipitation, a fraction (1/10) of the recombinant protein was mixed with Laemmli sample buffer at 95°C and loaded onto a Bis-Tris SDS/4-12% polyacrylamide gel. After SDS/PAGE, proteins were transferred onto Immobilon-P PVDF membranes. Membranes were blocked for 1 hr with 5% nonfat milk and incubated with the appropriate dilution of mouse anti-V5 antibody conjugated to horseradish peroxidase to detect the V5-tagged proteins for both pcDNA3.1 and pIHV constructs. Enhanced Chemiluminescence Plus and Image J software were used to detect and quantitate respective protein bands.

5.2.2 Transferase Reaction Reagents and Substrates

All radiolabeled materials including UDP-[\(^{3}\)H]-GlcNAc, UDP-[\(^{3}\)H]-Gal, UDP-[\(^{3}\)H]-GalNAc and CMP-[\(^{3}\)H]-Neu5Ac were purchased from American Radiolabeled Chemicals (St. Louis, MO). Sugar donors UDP-GlcNAc, UDP-Gal, UDP-GalNAc and CMP-Neu5Ac and Protease Inhibitor cocktails P8340 and P8849 were purchased from Sigma-Aldrich (St. Louis, MO). LacNAc-PNP substrate was purchased from Toronto Research Chemicals (Toronto, ON Canada). Lactose-PNP and GlcNAc-PNP substrates were purchased from Carbosynth Limited (West Berkshire, UK). Antarctic Fish antifreeze glycoprotein was a gift from Arthur L. DeVries (University of Illinois at Urbana-Champaign). Asialofetuin from fetal calf was purchased from Sigma-Aldrich. OPT12 peptide was custom synthesized and purchased from RS Synthesis (Louisville, KY). All substrates were lyophilized from water several times and the stocks were
prepared by diluting with either water or DMSO to a concentration of 10 mM or 50 mg/mL and stored at -20°C, see Table 5.3 for substrate descriptions. Sep-pak C18 columns were purchased from Waters (Milford, MA). Dialysis tubing, MW3500 Snakeskin and G2 dialysis cassettes (Slide-A-Lyzer), MW cut off 3500, were purchased from ThermoFisher. Sephadex G10 gel-filtration columns were the same as described in chapter 2. PD MiniTrap G10 Sephadex columns (5 mL) were purchased from GE Healthcare. Immunoprecipitation buffer consisted of 50 mM Tris pH 7.5, 150 mM NaCl, 1 mM CaCl₂, 1 mM MnCl₂ and EDTA-free protease inhibitor pellets.

5.2.3 Instrumentation

All instrumentation used including the microplate incubator, lyophilizer, scintillation counter, fraction collector and spectrophotometer were the same as described in chapter 2.

5.2.4 Glycosylation Assays with Recombinant Proteins

Typically for these glycosylation assays, each WT and mutant glycosyltransferases were performed together on the same day of isolation using the same donor and substrate concentrations. A total of three repeats were performed for each WT and mutant glycosyltransferase assay.
### Table 5.3 – Table of substrates used and expected glycosylated products.

<table>
<thead>
<tr>
<th>Glycosyltransferase</th>
<th>Donor Substrate</th>
<th>Acceptor Substrate</th>
<th>Product Glycan Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>β3GnT2</td>
<td>UDP-GlcNAc</td>
<td>LacNAc-PNP: β-Gal(1-4)GlcNAc-β-PNP</td>
<td>β-GlcNAc(1-3) β-Gal(1-4)GlcNAc-β-PNP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lactose-PNP: β-Gal(1-4)Glc-β-PNP</td>
<td>β-GlcNAc(1-3) β-Gal(1-4)Glc-β-PNP</td>
</tr>
<tr>
<td>β4GalT2</td>
<td>UDP-Gal</td>
<td>GlcNAc-PNP: β-GlcNAc-PNP</td>
<td>β-Gal(1-4) β-GlcNAc-PNP</td>
</tr>
<tr>
<td>ST6GalNAc II</td>
<td>CMP-Neu5Ac</td>
<td>AFGP (antarctic fish antifreeze glycoprotein): β-Gal(1-3)α-GalNAc-O-Thr ASF (asialo fetuin): β-Gal(1-3)α-GalNAc-O-Ser/Thr</td>
<td>β-Gal(1-3)(α-Neu5Ac (2-6))α-GalNAc-O-Thr β-Gal(1-3)(α-Neu5Ac (2-6))α-GalNAc-O-Ser/Thr</td>
</tr>
<tr>
<td>ppGalNAc T12</td>
<td>UDP-GalNAc</td>
<td>GAGAYYTTPRAGAGK</td>
<td>GAGAYYT(GalNAc-α-T)PRAGAGK</td>
</tr>
</tbody>
</table>

**Nomenclature of Substrates:**

- **LacNAc-PNP:** p-nitrophenyl 2-Acetamido-2-deoxy-4-O-(β-D-galactopyranosyl)- β-D-glucopyranoside
- **Lactose-PNP:** 4-nitrophenyl- β-D-lactopyranoside
- **GlcNAc-PNP:** 4-nitrophenyl-2-acetamido-2-deoxy- β-D-glucopyranoside
5.2.4.1 β3GnT2 (β-1,3-N-Acetylglicosaminyltransferase 2) Assays (10)

Reaction conditions for the β3GnT2 activity assay consisted of 150 mM MES buffer pH 7.5, 10 mM MnCl₂, 1.8 mM UDP-GlcNAc (containing a total of 100 µCi of UDP-[³H]-GlcNAc per reaction), protease inhibitor cocktails P8340 and P8849 and either 0.5 mM LacNAc-PNP substrate or 0.5 mM Lactose-PNP substrate and ~100-150 µL of β3GnT2 (CONT, WT, P186T and D247H) bound to beads. Reaction volumes were ~250-300 µL and were carried out in 1.5 mL capped Eppendorf tubes. Reaction vials were placed in a thermostated microplate incubator and were shaken at 36°C. Time point aliquots of ~83-100 µL at 2hr, 5hr and 24 hr (ON) were removed and quenched with an equal volume of 250 mM EDTA. Samples were then diluted to 3 mL and a small aliquot removed for scintillation counting while the rest of the sample was subjected to reverse-phase chromatography on a Sep-pak C18 column. The glycosylated product was eluted using 100% methanol and was directly used for scintillation counting. The transferase activities of the WT and mutant β3GnT2 transferases were expressed as the ratio of total [³H]-GlcNAc DPM’s of Post- vs. Pre – Sep-pak columns and further normalized to protein levels with the corresponding Western Blot. The resulting plots demonstrating the wild-type and mutant protein activity are shown in Figures 5.2 and 5.3.

5.2.4.2 β4GalT2 (β-1,4-Galactosyltransferase 2) Assays (10)

The following reaction conditions were used for the B4GALT2 activity assay: 25 mM Tris-Base buffer pH 7.4, 0.2% Triton X-100, 10 mM MnCl₂, 2 mM UDP-Gal (containing a total of 100 µCi of UDP-[³H]-Gal), protease inhibitor cocktails P8340 and P8849, 3 mM of GlcNAc-PNP substrate and ~250 µL of β4GalT2 (CONT, WT and
A146V) bound to beads. Reaction volumes were ~250-300 µL and were carried out in 1.5 mL capped Eppendorf tubes. Reaction vials were placed in a thermostated microplate incubator and were shaken at 36°C. Time point aliquots of ~83-100 µL at 1hr, 4hr and 24 hr (ON) were removed and quenched with an equal volume of 250 mM EDTA. The glycosylated product isolation was carried out the same way as described above for β3GnT2. The resulting plots showing the wild-type activity compared to the control and mutants are shown in Figures 5.4 and 5.5.

5.2.4.3 *ST6GalNAcII* (*α-N-Acetylgalactosaminidyl-α-2,6-Sialyltransferase 2*) Assays (10)

Glycosylation assays consisted of a reaction mixture containing: 50 mM MES buffer pH 6.0, 10 mM MgCl₂, 2 mM CaCl₂, 2 mM CMP-Neu5Ac (sialic acid) (containing a total of 100 µCi of CMP-[³H]-Neu5Ac), 0.2% Sodium Azide, protease inhibitor cocktail P8340 and P8849, 5 mg/mL of Antarctic fish antifreeze glycoprotein or 5 mg/mL of asialofetuin and ~100-150 µL of ST6GalNAcII (CONT, WT, D43H and R115W) bound to beads. Reaction volumes were ~200-250 µL and were carried out in 1.5 mL capped Eppendorf tubes. Reaction vials were placed in a thermostated microplate incubator and were shaken at 35°C. Time point aliquots of ~66-83 µL at 2hr, 4hr and 24hr (ON) were removed and quenched with an equal volume of 250 mM EDTA. Samples were diluted by 10-fold and a small aliquot removed for scintillation counting. The remainder of the sample was either dialyzed in MW3500 dialysis tubing or in 3 mL G2 dialysis cassettes against 2L of distilled water for 73 hours at 4°C with continuous stirring. The water was changed ~6 times over the course of dialysis. After dialysis, the samples were lyophilized and diluted to 1 mL and a small aliquot was removed for
scintillation counting while the rest was monitored for absorbance at 220 and 280nm. WT and mutant transferase specific activities were determined as the ratio of post-dialysis counts vs. the product glycoprotein absorbance at 220 and 280 nm and further normalized to the transferase protein levels from the Western blotting. The resulting plots showing the wild-type activity compared to the control and mutants is shown in Figure 5.6.

5.2.4.4 ppGalNAc T12 Assays (unpublished)

The glycosylation reaction for ppGalNAc T12 consisted of 76 mM Sodium cacodylate buffer pH 6.8, 0.61 mM 2-mercaptoethanol, 0.06% Triton X-100, 12 mM MnCl₂, 2 mM UDP-GalNAc (containing 300 μCi of UDP-[³H]-GalNAc), protease inhibitor cocktails P8340 and P8849, 5 mg/mL of OPT12 substrate and 200-300 μL of ppGalNAc T12 (CONT, WT, H101Q, I142T, E239Q, T286M, V290F, R297W, D303N and Y396C) bound to beads. OPT12 is an optimal peptide substrate for ppGalNAc T12 whose sequence was obtained using the random peptide substrate preferences reported for ppGalNAc T12 (Gerken et al. (22)) and available at the ISOGlyP web based server for predicting ppGalNAc isoform specific glycosylation. OPT12 was shown to be an excellent substrate against an authentic purified recombinant h-ppGalNAc T12 expressed in insect cells (kindly supplied to us by Henrik Clausen) (Revoredo et al. unpublished). Furthermore, this substrate shows very little hydrolysis activity compared to earlier substrates (Gerken and Revoredo unpublished).

The pH was adjusted to ~6.8 after the addition of all of the reagents. Reaction volumes were typically ~460-560 μL and carried out in 1.5 mL Eppendorf tubes. The reaction vials were incubated at 37°C and kept shaking in the microincubator. Following
an overnight incubation, reaction mixtures were quenched with an equal volume of 250 mM EDTA and frozen for later processing. UDP and non-hydrolyzed UDP-GalNAc were removed by passing the sample through a column of ~3mL of Dowex 1x8 anion exchange resin. \[^{3}H\]-GalNAc incorporation was determined by scintillation counting 1/50 of the sample before and after passing over the Dowex column. The eluent was lyophilized and passed over either a Sephadex G10 gel filtration column or passed over a PD MidiTrap Sephadex G-10 mini-column (1.0 x 5.5 cm) where the eluent was collected manually in one fraction (~12 mL). The fractions were monitored for \[^{3}H\]-GalNAc incorporation onto the glycosylated OPT12 substrate by scintillation counting and monitored for absorbance at 220 nm and 280 nm. Samples were pooled and lyophilized multiple times to remove acetic acid buffer and concentrate the sample. The lyophilized fractions were then diluted to 2 mL and a 1 mL aliquot was removed and counted for \[^{3}H\]-GalNAc incorporation and the initial solution monitored for absorbance at 220 nm and 280 nm. Net transferase activity was determined as the ratio of total sample \[^{3}H\]-GalNAc content in \[^{3}H\]-DPM/OD220 and \[^{3}H\]-DPM /OD280. With this approach any losses in peptide substrate during sample processing are readily corrected by normalizing to the peptide OD values. Mutant transferase-specific activities were then normalized to the WT transferase to yield an approximate relative specific activity of each mutant, see Figure 5.8.

5.3 Results and Discussion

5.3.1 β3GnT2 (β-1,3-N-Acetylglucosaminyltransferase 2)
This enzyme belongs to the gene family B3GNT, where family members share a “β3GT” motif that includes a type II membrane topology and one potential N-glycosylation site that is conserved (23). Structurally the human β3GnT2 consists of 369 amino acids (23). This enzyme is an elongating glycosyltransferase and is primarily responsible for catalyzing the addition of a β3-GlcNAc onto a terminal β4-linked galactose residue of polylactosamine chains that are found in N- and O- glycans and glycolipids (see Figure 5.2 A for reaction scheme). For O-glycans, the polylactosamine structures can be attached to core 1, core 2 and core 3 shown in Figure 5.1 A. Three mutations in the β3GNT2 gene were identified in colon cancer patients by the Guda Lab, these were: R6X, P186T and D247H. The R6X mutation exists in the early transmembrane region of the protein (Figure 5.2 B), because of this it was speculated to produce a truncated version of the protein and this was confirmed in the western blots shown in Figure 5.2 C, clearly showing that the R6X is truncated and expressed at a lower level compared to the WT β3GnT2. Furthermore, the Golgi-targeting signal sequence of β3GnT2 resides in the N-terminal transmembrane motif, were the Guda Lab hypothesized that the truncated R6X protein product that lacks this signal sequence and would not localize to the Golgi. This is shown in Figure 5.3 D, where the immunofluorescence analysis shows that indeed the mutant R6X in COS7 cells has diffused sub-cellular localizations while the β3GnT2 WT only localized in the Golgi. From these two studies, the Guda Lab concluded that the R6X β3GnT2 most likely does not have any catalytic function and may lack access to endogenous substrates within the Golgi. Therefore, in vitro glycosylation assays were not performed with the R6X mutant.
A. $\text{UDP-GlcNAc} + \text{Galβ4-R} \rightarrow \beta\text{GlcNAc(1-3)Galβ4-R} + \text{UDP}$

B. 

C. 

D. 

**Figure 5.2**- Identification of B3GNT2 mutants and localization of R6X mutant. (A) Reaction catalyzed by β3GnT2. (B) Cartoon of the somatic mutations mapped to the β3GnT2 protein coding regions (black line). Colored boxes represent the transferase domains, the TM is the transmembrane motif in green while the glycosyltransferase catalytic domain is in orange. (C) Western blot against the V5-tagged WT β3GnT2 and R6X mutant protein expression. Note the smaller size band for the R6X mutant compared to WT β3GnT2. (D) Immunofluorescence analysis of V5-tagged WT and mutant R6X show that the β3GnT2 WT (in green) is co-localized with the Golgi marker in red. Note the aberrant sub-cellular localization of the R6X mutant in green. Both glycosyltransferases were expressed in COS7 cells transfected with the respective cDNA constructs. Data taken from (10).
The other two mutants identified, P186T and D247H are missense mutations that are located in the catalytic domain of the β3GnT2 protein shown in Figure 5.2 B. The impact of the mutations on the protein activity was tested using two substrates, LacNAc-PNP and Lactose-PNP. These two substrates were selected based on the high activity of WT β3GnT2 against these substrates as reported previously (24,25) shown in Figure 5.3 A. The expected glycosylated product structures as well as the un-glycosylated substrate structures are shown in Table 5.3. Figure 5.3 B shows the expression levels of both mutants being comparable to the WT β3GnT2. The D47H mutant has no detectable enzymatic activity against both substrates; see left and right panels of Figure 5.3 C for results with LacNAc-PNP and Lactose-PNP respectively. For the P186T mutant, the enzymatic activities varied with substrate. For the LacNAc-PNP substrate, activities of the P186T mutant and WT β3GnT2 were comparable with the WT β3GnT2 displaying slightly higher activities. For the Lactose-PNP substrate, the P186T mutant displayed higher activities than the WT β3GnT2, see left and right panels of Figure 5.3 C.

Altogether, these results suggest that two mutants, the R6X and D47H could impair the downstream function of β3GnT2 in the cell. Interestingly, earlier work showed that a family member, β3GnT8, was identified to be up-regulated in colon cancer patients (18), where the authors concluded that its up-regulation is involved in malignancy due to the synthesis of β1-6 branched N-glycans. Therefore, the elevated activity found in the P186T mutant may play a different role in colon cancer compared to the other mutants.
Figure 5.3- *In vitro* glycosylation assays with P186T and D47H mutants. (A) Enzymatic activity of WT β3GnT2 against LacNAc-PNP and Lactose-PNP at ON incubations. Data was normalized to vector controls. (B) Western blot of the V5-tagged WT β3GnT2, P186T and D47H mutant protein expression. Both glycosyltransferases were expressed in COS7 cells transfected with the respective cDNA constructs. (C) Mean activities of WT, P186T and D47H mutants with LacNAc-PNP (left panel) and Lactose-PNP substrates (right panel). Error bars represent standard error derived from three independent experiments. Data taken from (10).
5.3.2 β4GalT2 (β-1,4-Galactosyltransferase 2)

This enzyme belongs to the β4GALT gene family consisting of seven members, which based on their amino acid sequences, are clustered into subgroups: β4GalT1 and β4GalT2, β4GalT3 and β4GalT4, β4GalT5 and β4GalT6 and β4GalT7 all of which are differentially expressed (26). Structurally, these are type II membrane proteins that reside in the Golgi (27) (except for β4GalT1 which can also be secreted in lactating mammary tissues) (28), while most members have conserved N-linked glycosylation sites (26). β4GalT2 catalyzes the transfer of β4-Gal onto a terminal β-linked GlcNAc (N-acetylglucosamine) residue of complex oligosaccharides and polylactosamine chains that are found in N- and O- glycans and glycolipids (see Figure 5.4 A for reaction scheme). Only one missense mutation, A146V was identified in the B4GALT2 gene in colon cancer patients by the Guda Lab. This mutation is located in the catalytic domain of the β4GalT2 glycosyltransferase, see Figure 5.4 B for cartoon. Out of all the mutants in this study, this mutation was accompanied by a genomic loss of the wild-type allele in the mutant CRC cell line (but will not be discussed here, see Figure 3B of (10)). The impact of the A146V mutation on the β4GalT2 enzyme was tested against one glucopyranoside substrate (see Table 5.3 for structure), GlcNAc-PNP, based on the previously reported high activity of the WT β4GalT2 with this substrate (29) shown in Figure 5.4 C. The glycosylation assays, in Figure 5.5 E, revealed no detectable activity with the A146V β4GalT2 mutant.
Figure 5.4 - Glycosylation assays with β4GalT2. (A) Reaction catalyzed by β4GalT2. (B) Cartoon of the somatic mutations mapped to the β4GalT2 protein coding regions (black line). Colored boxes represent the transferase domains, the TM is the transmembrane motif in green while the glycosyltransferase catalytic domain is in orange. (C) Enzymatic activity of WT β4GalT2 against GlcNAc-PNP at ON incubations. Data was normalized to vector controls. (D) Western blot of the V5-tagged WT β4GalT2 and A146V mutant protein expression. (E) In vitro glycosylation assays with WT β4GalT2 and A146V mutant. Mean activities of WT and A146V mutants with GlcNAc-PNP. Error bars represent standard error derived from three independent experiments. Data taken from (10).
Upon closer analysis of the Western blots in Figure 5.4 D, the Guda Lab noticed a slightly different migratory pattern for the WT β4GalT2 than the mutant, perhaps suggesting that there is an additional post-translational modification on the WT β4GalT2. The mass spectrometry analysis of the WT and mutant β4GalT2 bands in the Western blot in Figure 5.4 D confirmed the identities as β4GalT2 protein, but was not able to resolve the specific post-translational modification (data not shown). Upon the sequence analysis of β4GalT2, three potential N-linked glycosylation sites (N-X-S/T) were identified at amino acids, 66, 71 and 357 (29). The Guda Lab decided to assess the N-linked glycosylation sites by expressing the β4GalT2 WT and A146V mutant in the corresponding V957 CRC cell line and performing Western Blots after treating the expressed transferases with pan N- and O- glycosidase or a specific N-linked glycosidase (PNGaseF). The resulting Western blots are shown in Figure 5.5. The Western shows that both the WT and mutant A146V β4GalT2 are N-glycosylated due to the similarities of the migratory patterns of the bands with both pan-glycosidase (lane 5 for WT and lane 8 for A146V) and PNGaseF treatments (lane 6 for WT and lane 9 for A146V). Despite this, the WT β4GalT2 still showed a higher sized shift than the mutant suggesting an additional, undetermined post-translational modification for WT β4GalT2.

Altogether, these findings suggest that the A146V β4GalT2 mutant not only significantly decreases enzymatic activity but also potentially disrupts an additional undetermined post-translational modification. On the other hand, prior work has shown that in colon cancer adenocarcinomas, the activity of the β4GalT’s are up-regulated in the synthesis of polylactosamines (1) such as β4GalT1 (19) and β4GalTIV (30). It is still
Figure 5.5- Western blot of glycosidase treatments on WT and mutant β4GalT2. Note the different band sizes between WT and mutant β4GalT2 in untreated cells: lanes 4 and 7. Pan glycosidase (lanes 5 and 8) and PNGaseF (lanes 6 and 9) treatments on both WT and mutant β4GalT2 respectively, show that both of these proteins are N-glycosylated. Overall, WT β4GalT2 bands show a higher size-shift than mutant indicating the presence of another additional post-translational modification. Data taken from (10).
unclear as to why some β4GalT’s are up-regulated and down-regulated in colon cancer and more work needs to be done in order to understand this.

5.3.3 ST6GalNAcII (α-N-Acetylgalactosaminidyl-α-2,6-Sialyltransferase 2)

ST6GalNAcII belongs to the ST6GalNAc family of transferases in humans that transfer sialic acids onto the T- and Tn- antigen structures and oligosaccharide chains shown in Figure 5.6 A, with the other family members being ST6GalNAc-I, III-VI (31). ST6GalNAcII catalyzes the transfer of sialic acid into the peptide linked α-GalNAc-O-Ser/Thr forming the STn-antigen and onto the Core 1 and Core 3 O-glycan structures (Figure 5.1) (i.e. β-Gal(1,3)[α-Neu5Ac(2,6)]α-GalNAc-O-Ser/Thr and β-GlcNAc(1,3)[α-Neu5Ac(2,6)]α-GalNAc-O-Ser/Thr, respectively). The other family member, ST6GalNAcI, has been shown to have similar but not identical specificities (32). Like the other glycosyltransferases, ST6GalNAcII has a type II membrane topology, containing the characteristic motifs of other sialyltransferases (known as sialyl motifs L, S, and VS) and a large C-terminal catalytic domain of approximately ~280 amino acids (31,33,34). The ST6GalNAc T family and the STn antigen has been studied intensively since both the enzyme and product antigen is highly expressed in carcinoma of patients with breast, gastric, colorectal, ovarian and pancreatic cancers (13,34-39). Two missense mutations in the ST6GALNAC2 gene were identified in colon cancer patients by the Guda Lab, D43H and R115W. The D43H mutant is located in the stalk of the transferase between the transmembrane domain and the catalytic domain while the R115W is located in the catalytic domain of the transferase as shown in Figure 5.6 B.
A. CMP-Neu5Ac + $\text{R(1-3)}\alpha\text{GalNAc-O-Ser/Thr}$ $\rightarrow$ $\text{R(1-3)[}\alpha\text{-Neu5Ac(2,6)]}\alpha\text{GalNAc-O-Ser/Thr}$ + CMP

$R=$ Gal, GlcNAc

B. ST6GALNAC2

C. ST6GALNAC2

D. Western blot of the V5-tagged WT ST6GALNACII, D47H and A115W mutant protein expression.

E. Enzymatic activity of WT ST6GALNACII against AFGP and ASF at ON incubations. Data was normalized to vector controls. (D) Western blot of the V5-tagged WT ST6GALNACII, D47H and A115W mutant protein expression. (E) In vitro glycosylation assays with WT ST6GALNACII, D47H and A115W mutants. Error bars represent standard error derived from three independent experiments. Data taken from (10).

Figure 5.6- Glycosylation assays with ST6GALNACII. (A) Reaction catalyzed by ST6GALNACII. (B) Cartoon of the somatic mutations mapped to the ST6GALNACII protein coding regions (black line). Colored boxes represent the transferase domains, the TM is the transmembrane motif in green while the glycosyltransferase catalytic domain is in orange. (C) Enzymatic activity of WT ST6GALNACII against AFGP and ASF at ON incubations. Data was normalized to vector controls. (D) Western blot of the V5-tagged WT ST6GALNACII, D47H and A115W mutant protein expression. (E) In vitro glycosylation assays with WT ST6GALNACII, D47H and A115W mutants. Error bars represent standard error derived from three independent experiments. Data taken from (10).
The impact of these mutations on the enzymatic activity of the encoded ST6GalNAcII was assed using two substrates, antifreeze glycoprotein from Antarctic fish (AFGP) and asialofetuin (ASF), see Table 5.3 for substrate structures and Figure 5.6 C (32). From the results, no significant differences in activities between the WT and ST6GalNAcII D47H and R115W mutants could be established due to variable results obtained against both substrates, see Figure 5.6 E. Hence, no conclusions could be made regarding the mutants with these glycosylation assays.

5.3.4 Phenotypic Characterization of WT and Mutant Glycosyltransferases β3GnT2, β4GalT2 and ST6GalNAcII

Since aberrations in cell surface glycans have been shown to primarily affect the migratory and metastatic potential of cancer cells, the effects of the WT and mutant genes on cancer cell migration were tested (by the Guda Lab) using the widely used colon cancer cell line SW480. Briefly, SW480 cells were transfected with the respective WT or mutant versions of β3GnT2, β4GalT2 and ST6GalNAcII (and an empty vector control) and the cell migration was assessed in a scratch wound assay over the course of a 48 hour period using the IncuCyte live cell kinetic imaging system (see (10) for more detailed information). The results are shown in Figure 5.7 A-C. The β3GnT2 mutants, P186T and D247H significantly enhanced the migratory potential of the SW480 cell while the WT did not (Figure 5.7 A). This suggests that both the β3GnT2 mutants potentially have a role in the oncogenic functions. In contrast, the WT ST6GalNAcII suppressed the
Figure 5.7- Effects of WT vs. mutant enzymes on colon cancer cell migration. SW480 colon cancer cell lines transfected with (A) WT β3GnT2 and mutants P186T and D247H (B) WT β4GalT2 and mutant A146V (C) WT ST6GalNACII and mutants D43H and R115W. Proteins were quantified using the IncuCyte scratch wound assay over a 48 hour period. Each of the β3GnT2 expressing cells showed a significant increase in migratory potential when compared to the WT and vector control, while the WT ST6GalNACII reduced cell migration and not the mutants. Note the reduced migratory potential of WT and mutant β4GalT2. Error bars represent the standard error of the means derived from five replicated per experimental group. Data taken from (10).
migration of the SW480 cells, while the mutants D43H and R115W failed to inhibit
cancer cell migration thus indicating a loss in phenotype function (Figure 5.7 C). The
results with the WT ST6GalNAcII is consistent with its proposed function as a tumor
suppressor in breast cancer (40). Results with the WT β4GalT2 and mutant A146V did
not show any difference (Figure 5.7 B), suggesting that perhaps β4GalT2 does not play
any role in cell migration.

5.3.5 ppGalNAc T12 (polypeptide N-acetylgalactosaminyltransferase 12)

ppGalNAc T12 as discussed earlier in my thesis, is a member of the ppGalNAc T
family of transferases that transfers GalNAc onto Ser or Thr residues on protein
substrates. This enzyme has been extensively studied in the Gerken Lab and much work
has been put into characterizing the catalytic domain peptide and glycopeptide
preferences (chapters 2 (22) and 3 (41)) as well as the lectin domain preferences (chapter
4). My work has revealed T12 to have a neighboring C-terminal T* catalytic domain
glycopeptide preferences as well as a remote N-terminal T* directionality preferences
due to its lectin domain. Earlier work by the Gerken lab, using the random peptide
approach (22) has suggested that the optimal ppGalNAc T12 preference motif of
:YYITPRP, see Figure 2.1 of chapter 2. This optimal motif has incorporated into the
peptide sequence: GAGAYYITPRPAGAG, called OPT12 and has been tested against
WT ppGalNAc T12 (from the Clausen and Guda Labs). Our results show OPT12 to be an
excellent substrate for WT ppGalNAc T12; where the sephadex G10 chromatograms of
the products show complete transfer to substrate with no hydrolysis of UDP-GalNAc
suggesting that this would be an excellent substrate to use with this work for monitoring mutant ppGalNAc T12 activity (see Figure 5.8 A).

Eight additional somatic and germ-line mutations were identified in colon cancer patients by the Woods Lab. These are H101Q, I142T, E239Q, T286M, V290F, R297W, D303N and Y396C, see Table 5.1. All of the mutations are located in the catalytic domain of ppGalNAc T12, see Figure 5.8 B. The impact of the mutations on the encoded protein were tested with the OPT12 as this is a better substrate than the originally tested substrates. The results are given in Figure 5.8 C. Although all mutants showed lower activity (<60%) compared to the WT, the I142T and Y396C mutants showed the lowest level of enzymatic activity (<45%). The results with the D303N mutant is consistent with the previous studies performed with the EA2 and MUC5AC substrates, where a 37% activity was reported (3). However, with the R297W mutant, we observed an enzymatic activity of ~40%, while the prior study showed an enzymatic activity of 7% for the R297W mutant (3). This variability is probably due to the different substrates. We may be observing more activity in our present assays because we are using the optimal substrate for ppGalNAc T12. The mutants with the greatest differences in enzymatic activity as compared to the WT, based on the student t-Test (P ≤0.05) are H101Q, I142T, V290F, R297W and Y396C. These results confirm that the lowered activities of the mutant ppGalNAc T12 are associated with colon cancer in some manner, perhaps due to the inability to glycosylate important protein targets.
Figure 5.8- Biochemical characterization of WT and mutant versions of ppGalNAc T12 associated with colon cancer. (A) Sephadex G10 Chromatograms of selected ppGalNAc T12 mutants glycosylated with OPT4 showing efficient transfer to substrate and no hydrolysis. (B) Cartoon of the somatic mutations mapped ppGalNAc T12 protein coding regions (black line). Colored boxes represent the transferase domains; the TM is the transmembrane motif in green while the glycosyltransferase catalytic domain is in orange. (C) Y-axis depicts the enzyme activities of mutant versions of ppGalNAc T12, relative to the wild-type (WT) protein (X-axis). Error bars represent standard error of the means derived from three independent replicate experiments (performed by Revoredo, L and Venkitachalam, S). A (*) indicates significant difference (P≤0.05) in enzyme activity of respective mutant ppGalNAc T12 as compared to the WT protein, estimated by one-tailed Student’s t-test. (D) Shown below the graph is a representative Western blot image demonstrating protein expression of WT and mutant versions of ppGalNAc T12 in SW480 cells transfected with respective cDNA constructs. Unpublished data.
5.4 Summary and Conclusions

It is well known that aberrant glycosylation contributes to the progression of colon cancers (1-3); however, the molecular mechanisms underlying aberrant glycosylation by the glycosyltransferases and their role(s) in tumor progression have yet to be understood. In this study, 31 patient-derived colorectal cancer (CRC) cell lines were re-sequenced and out of the 430 genes tested, 12 genes were found to be significantly mutated in colorectal cancers (performed by the Guda Lab). These 12 genes contained a significant number of mutations in the N- and O-glycosylation pathways that were especially related to the polylactosamine synthesis. The three genes detected, B3GNT2, B4GALT2 and ST6GALNAC2 were in 5 out of the 31 CRC cases, with 3 mutations in B3GNT2 (R6X, P186T & D247H), 1 mutation in the B4GALT2 (A146V) and 2 mutations in ST6GALNAC2 (D43H & R115W) (Table 5.1). Only the B4GALT2 A146V mutant was accompanied by a loss of the WT allele (data not shown). The additional analysis of the independent large-scale cancer data sets also showed recurrent somatic mutations in B3GNT2, B4GALT2 and ST6GALNAC2, where all three account for ~3% of CRC cases (10,42,43). For GALNT12, 8 additional germline mutations were identified in a screening of 517 patient DNA samples from The Newfoundland Colorectal Cancer Registry (NFCCR) (performed by the Woods Lab). These were H101Q, I142T, E239Q, T286M, V290F, R297W, D303N and Y396C each of which accounted for one patient except for D303N, which was found in 3 patients.

The effects of these mutations on the encoded proteins; β3GnT2, β4GalT2, ST6GalNAcII and ppGalNAc T12 were tested (in the Gudgen Lab by: Revoredo, L and the Guda Lab by: Venkitachalam, S) *in vitro* against their respective substrates. The
results suggested that the mutational effects on the glycosyltransferases indeed have an impact on enzymatic activity and/or migratory potential of the colon carcinoma cells (10).

β3GnT2
For the β3GnT2 mutants, the R6X mutant encoded for a N-terminal truncated protein with no catalytic domain, which showed diffused subcellular localizations. The other two mutants, P186T and D247H, were located in the catalytic domain. Against the two substrates tested, the D247H consistently showed a loss of enzymatic activity, while the activity was retained in the P186T mutant (Figure 5.3). However, the cell migration phenotypic analyses of all three mutants showed significant increase in the migratory potential of colon adenocarcinoma cells compared to the WT (Figure 5.7). This suggests that the observed oncogenic function in the original colon cancer cell line is indeed likely caused by the decreased activities of the mutant β3GnT2 (10).

β4GalT2
The only mutant detected for β4GalT2 was the A146V and was the only mutant in this study that was accompanied by a genomic loss of the WT allele in the corresponding colorectal cancer sample (data not shown). Enzymatic activity of the A146V mutant was lost against its substrate, see Figure 5.4. Additionally, glycosidase studies, in Figure 5.5, suggest that the A146V mutation may be altered by an unknown post-translational modification (Figure 5.5). The cell migration phenotypic analyses, interestingly see Figure 5.7, did not show a difference in mutants the migratory potential compared to the WT, indicating that the endogenous targets of β4GalT2 may not be involved in regulating
cell motility or that they require a differentiated tissue microenvironment for functioning (10).

**ST6GalNAcII**

The two ST6GalNAc II mutants tested, D43H and R115W did not show any loss of enzymatic activity against both substrates (Figure 5.6) in our glycosylation activity assays, although the D43H mutant consistently showed slightly higher activities than the WT. The results with the D43H mutant is keeping with prior findings that show increased protein sialylation in colorectal cancers (14,44). The phenotypic analyses of these mutants are quite unexpected as the mutants showed higher migratory potentials than the WT ST6GalNAcII (10) (Figure 5.7). Perhaps the WT and mutant ST6GalNAcII may have different affinities/specificities towards endogenous targets, perhaps due to altered trafficking or altered accessory protein interactions.

**ppGalNAc T12**

All of the 8 mutants tested showed lowered activities compared to the WT ppGalNAc T12, with 5 of them showing significant differences: H101Q, I142T, V290F, R297W and Y396C see Figure 5.8 C. The dramatic effects on the activities of these mutations on the encoded mutant proteins further supports the involvement of ppGalNAc T12 (GALNT12) in colon cancer.

While the studies may not fully capture or explain the functional complexities of N- and O-linked glycosylation, these studies still provide some insight into the genes that are involved in the synthesis of the Tn-antigen, cores 1 and 3 for O-glycosylation pathways and how they may contribute to colorectal cancers. Clearly, further
characterization of the glycosyltransferases with their endogenous substrates and the evaluation of these mutant phenotypes in animal models would provide additional insights into the biology of these genes and their roles in colon cancer initiation and/or progression.
5.5 References


CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS
6.1 Summary

The work in this thesis explored the peptide and glycopeptide specificities of the catalytic and lectin domains of the ppGalNAc T transferases. This large family (20 different isoforms in man) of transferases is responsible for initiating the first step in mucin-type O-glycosylation in the Golgi, where a GalNAc is transferred from UDP-GalNAc onto Ser and Thr residues on protein substrates. Members of the ppGalNAc T family are classified into two major families (I and II) based on their peptide/glycopeptide preferences respectively (1). Although this distinction has been historically used, some overlap in substrate preferences exists among transferase isoforms. The families have been further divided into subfamilies (Ia-g and IIa-b) based on their amino acid sequence similarities, which tend to correlate with their peptide/glycopeptide substrate preferences. Family I members readily glycosylate peptide substrates and have been termed "peptide-preferring" isoforms, however, several members of this family (Ia-d) have been shown to be influenced by remote, 6-17 residues of prior GalNAc-O-Ser/Thr glycosylation via utilization of their lectin domains. Family IIa-b members on the other hand, commonly display sole activities against GalNAc-O-Ser/Thr containing glycopeptides and hence have been termed "glycopeptide-preferring" isoforms, although some isoforms (subfamily IIa) can also readily glycosylate non-glycosylated peptide substrates.

Prior work has shown that the ppGalNAc T’s are influenced by neighboring GalNAc-O-Ser/Thr glycosylation (S* and T* respectively) near a acceptor site which causes a series of transferase specific effects: an inhibition of glycosylation (i.e. ppGalNAc T1 and T2) (2), increase in glycosylation rates (i.e. ppGalNAc T4) (3) or a
complete requirement for a neighboring T* (i.e. ppGalNAc T10). Prior work on the
lectin domain of the ppGalNAc T transferases (ppGalNAc T1, T2, T3 and T4) (4-7)
showed that some transferases preferred certain substrates to others, and that the lectin
domain is involved in glycopeptide recognition and specificity, but no other detailed
information was known. In order to fully systematically study the functions of the
catalytic and lectin domains and address the overlapping specificities; a systematic
approach with glycopeptide substrates was developed. Each substrate was specifically
designed to independently probe the functions of the catalytic and lectin domains, in
terms of neighboring (1-5 residues) and remote (6-17 residues) of prior glycosylation.

The glycopeptide substrates used to study the catalytic domain glycopeptide
preferences was designed similarly to GPI, a glycopeptide substrate that was previously
used in the Gerken Lab containing the same design as the random peptide substrates.
These glycopeptide substrates, GP(T*10)C-Ser and GP(T*10)C-Thr, have a central
GalNAc glycosylated Thr residue (T*) that is flanked by five (+/- 5 residues) randomized
amino acid positions, X, where one of the randomized residues will contain a
randomized Ser acceptor (for glycopeptide GP(T*10)C-Ser)) or a randomized Thr
acceptor (for glycopeptide GP(T*10)C-Thr)). Edman sequencing of the substrates
determined the $[^3]$H-GalNAc incorporation in the X positions, where unique preference
data were obtained for 9 ppGalNAc T’s and two PGANT orthologs that were classified as
either “peptide-preferring” or “glycopeptide-preferring” transferases. For the first time,
isoform-specific glycosylation site preferences were determined from the -4, -3, -1 and
+1 sites of glycosylation (from the prior T*) confirming the roles the “filling-in
transferases” ppGalNAc T4, T12, T7, T10 and PGANT7 (8) have and roles of the
“peptide-preferring” transferases ppGalNAc T11 and dT1(PGANT35A). These glycopeptide substrates also confirmed the closest distance at which the lectin domain begins to assist the catalytic domain to glycosylate the -5 and +5 positions for ppGalNAc T2, T3, T4, T5, T11 and dT1 (PGANT35A).

For the lectin domain studies, a universal series of substrates were created in order to address the role of the lectin domain in glycopeptide glycosylation. These substrates address the modulating role of a remote GalNAc-O-Thr (T*) on transferase activity, based on the assumption that the T* would bind at the lectin domain. These glycopeptides were designed to span the catalytic and lectin domains, 29 residues in length. Each (glyco)peptide substrate contained a T* placed near the C- (GP(T*22)R) or N- (GP(T*10)L) terminal, where R is for right and L for left placed T*, respectively. These sites were in turn, flanked by 5 randomized Z residues lacking an acceptor Ser or Thr. An additional 12 randomized X residues, also including the acceptor Thr, were extended from the Z’s in an N- or C- terminal direction respectively. These X residues thus serve to probe glycosylation 6 to 17 residues from an existing T* in an N- or C-direction, respectively. With these (glyco)peptides the extent that remote GalNAc-O-Thr interactions at the lectin domain can enhance catalytic domain glycosylation of the X residues was determined. Specifically, differences in N- or C- directionality preferences between transferase isoforms could be observed. Notably, significant activities with these (glyco)peptides were observed with the “peptide-preferring” transferases. For ppGalNAc T1, T2, T11, T14, dT1 (PGANT35A) and PGANT9B, preferences were observed for a C-terminally placed GalNAc-O-Thr, while for ppGalNAc T3, T4, T6, T12 and PGANT9A, preferences were observed for a N-terminally placed GalNAc-O-Thr. Several other
transferase isoforms, ppGalNAc T5, T13, T16 and PGANT7 displayed equally enhanced N- or C-terminal activities relative to the non-glycosylated controls (9). Lastly, two other ppGalNAc T isoforms, ppGalNAc T7 and T10, displayed no significant activity with either of the (glyco)peptide substrates, suggesting no lectin domain activity. This N- or C-terminal selectivity is presumably due to the weak glycopeptide biding to the lectin domain, whose orientation relative to the catalytic domain is dynamic and isoform dependent. Such N- or C-terminal selectivity provides an additional level of control for O-glycosylation.

With these studies, a clearer understanding the functions of the lectin and catalytic domains in modulating glycopeptide substrate specificity has been revealed. For example we can distinguish between the effects of neighboring (+/- 5 residues) prior glycosylation, due to glycopeptide binding at the catalytic domain to the longer range effects of remote glycosylation (+/- 6 to 17 residues), where glycopeptide binding is mediated by the lectin domain. Of particular interest we have observed clear and unexpected differences in activity and directionality among the transferases studied to date. Furthermore, the results of this work demonstrates that apart from the peptide binding at the catalytic domain, there are at least two modes of glycopeptide substrate recognition by the ppGalNAc T’s. These would be the direct binding of the glycopeptide GalNAc within the catalytic domain, targeting specific neighboring glycosylation and the recognition of a glycopeptide at the lectin domain, directing remote glycosylation to the catalytic domain. We would like to refer to these activities as catalytic domain-directed and lectin domain-assisted. Another important aspect of this work is the evidence of the overlapping peptide/glycopeptide substrate preferences observed for the different
transferase isoforms. Because of this, we would like to suggest the renaming of the classifications of family I and family II transferases. Since subfamily Ia-Id members have elevated activities towards glycopeptide substrates (with a remote site of prior glycosylation) and thus not strictly “peptide-preferring” transferases, we suggest that for clarity, the family Ia-Id members be called remote glycopeptide/peptide-preferring isoforms, abbreviated as remote GP/P-preferring. Similarly, the subfamily IIa members have been shown to be relatively active towards a number of nonglycosylated peptide substrates and therefore not strictly “glycopeptide-preferring” transferases, we suggest that the family IIa members be called mixed (glyco)peptide-preferring isoforms, abbreviated as mixed (G)P-preferring. Lastly, since subfamily IIb members are poorly active against naked peptide substrates, we suggest calling them strict glycopeptide-preferring transferases, abbreviated as strict GP-preferring.

Other aspects that this thesis explored was the 1) Catalytic domain preferences determined for new ppGalNAc T isoforms (ppGalNAc T11, dT1 (PGANT35A), T13, T14, T16, PGANT9A and PGANT9B), 2) The comparison studies between the human ppGalNAc T’s and fly ortholog pairs, ppGalNAc T7 and PGANT7 and T11 and dT1(PGANT35A), 3) The comparison studies of the lectin domain splice variants of ppGalNAc T13 and T13 10XB and the fly PGANT9A and PGANT9B and 4) Studies on mutant glycosyltransferases (identified from colon cancer patients) and their roles in colon cancer was examined. Each will be discussed briefly below.

The catalytic domain preferences were determined for several undetermined ppGalNAc T isoforms (and fly PGANT’s) such ppGalNAc T11, dT1 (PGANT35A), T13, T14, T16, PGANT9A and PGANT9B using the random peptide approach determined by
the Gerken Lab. Overall, preferences were very similar if not identical between individual transferase subfamily members and varied between the different subfamilies. The determinations of the human ppGalNAc T transferases, T11, T13, T14 and T16 were added to the ISOGlyP database, for predictions of O-glycosylation.

This similarity of catalytic domain preferences was also true for the human and fly transferase ortholog pairs ppGalNAc T11 and dT1, where the preferences were nearly identical, demonstrating the conservation among the transferases across species. However, the glycopeptide catalytic domain preferences showed slight differences between ppGalNAc T11 and dT1. Specific sites of glycosylation were observed with GP(T*10)C-Thr at the -5 and -4 positions from the site of glycosylation for both transferases, however, the glycosylation patterns were slightly different. ppGalNAc T11 preferred to glycosylate the -4 position more than the -5 position while dT1 equally preferred the -5 and -4 positions. The catalytic domain peptide preferences of the other human and fly ortholog pair, ppGalNAc T7 and PGANT7 was examined with GP(T*10)C-Ser and it was shown that they both have a strict conserved -1 glycosylation site preference. However, their lectin domain preferences differed, where ppGalNAc T7 showed no activities with the substrates, while PGANT7 showed an equal preference for both the N- and C- terminally placed GalNAc-O-Thr. This work represents the first differences observed between both transferase and fly ortholog pairs.

Studies with the lectin domain splice variants showed very interesting results. For the first splice variant pair studied, ppGalNAc T13 and T13 Ex10B, no differences were observed in their lectin domain preferences, both showing an equal preference for both the C- (GP(T*22)R) or N- (GP(T*10)L) terminally placed T*.
conservation of their α and β subdomains in the lectin domain of both transferases thus conserving the lectin domain activities between both transferases (since the ppGalNAc T13 Ex10B splice is in the γ subdomain). The lectin domain splice variants of the fly PGANT9A and PGANT9B gave nearly identical catalytic domain peptide preferences, however, the lectin domain preferences showed that PGANT9A has a slightly higher preference for the N- (GP(T*10)L) terminally placed T* while the PGANT9B had a significantly higher preference (~1.5 fold) for the C- (GP(T*22)R) terminally placed T*.

This work represents the first difference observed with a lectin domain splice variant.

With these studies, we now have a better understanding perhaps why the large family of transferases exists. Each isoform contains a unique combination of catalytic domain-directed (glyco)peptide specificity and lectin domain remote glycopeptide specificity. Thus, for example, the isoforms in the Ia (ppGalNAc-T1 and T13) and Ib (ppGalNAc-T2, T14, T16) subfamilies of remote GP/P-preferring transferases all have nearly indistinguishable random peptide sequence preferences but possess different lectin domain preferences. Furthermore, the fact that nearly all ppGalNAc-T isoforms recognize either prior neighboring glycosylation or prior remote glycosylation, or a combination of both, suggests that these transferases must operate in an ordered process to perform their glycosylation functions, where one transferase produces an optimal glycopeptide substrate for the next. This suggests that mucin type O-glycosylation is potentially highly orchestrated in a cell and helps to further explain why the patterns of ppGalNAc-T isoform expression varies from cell type to cell type. Thus, the over expression or suppression of a single transferase in this cascade could lead to multiple downstream alterations of subsequent glycosylation.
Lastly, several somatic and germ-line mutations on glycosyltransferase genes were identified from actual colon cancer patients and the effects on the encoded proteins was studied in order to assess their roles in colon cancer. The transferases identified included one ppGalNAc T member, ppGalNAc T12 and three other glycosyltransferases which are likely involved in mucin-type O-glycan elongation (i.e. polylactosamine structures) or glycan termination, they were β3GnT2, β4GalT4 and ST6GalNAcII. Glycosylation assays were performed on each mutant and compared to their WT versions. Results showed that the mutations in these transferases either altered enzymatic activity (i.e. decreased/increased) or migratory potentials of the cells expressing them (8) (and unpublished data). Overall, these studies have uncovered new potential contributors to aberrant glycosylation that may contribute to the molecular pathogenesis of subsets of colon and other gastrointestinal malignancies.

6.2 Future Directions

The ultimate goals of our studies are to fully understand the properties of the transferases that initiate mucin-type O-glycosylation so that fully predictive models of mucin type O-glycosylation can be developed. The Gerken lab will continue to work on this aspect by continuing to add on to their systematic approach of random peptide substrates and continuing to work on several projects. These studies will include, 1) The characterization of the catalytic and lectin domains of the transferases that have not been characterized, 2) The improvement of ISOGlyP with the determination of the Ser vs. Thr utilization with the use of further specialized substrates, 3) Development of novel peptide
and glycopeptide substrates to further characterize all ppGalNAc T’s and 4) New collaboration projects with the Hurtado-Guerrero Lab to validate our work.

The Gerken Lab would like to continue to characterize all of the available ppGalNAc T transferases that have not been characterized before using our (glyco)peptide library in order to fully understand why there are so many transferase isoforms. Of upmost interest are 1) The low activity ‘Y’ subfamily Ie of ppGalNAc T transferases T8, T9, T18 and T19, 2) Subfamily If member T20 and subfamily Ig member ppGalNAc T15 and 3) Subfamily IIb member ppGalNAc T17 who have not been sufficiently characterized. Most of these transferases have been shown to have unique neuronal and testis expressions (10-13). Part of the reason why the “Y” series of transferases have lower activities against the common Ser and Thr peptides substrates is because of their replacement of the conserved Trp (W) residue by a Tyr (Y) in the UDP-GalNAc binding site (1,10,14). Because our random peptide approach has been successful in characterizing many other transferases, we think we are well positioned to use these substrates to characterize these transferases since we can essentially “push” the reactions by increasing the UDP-GalNAc and substrate concentrations. This has been previously performed for ppGalNAc T10 (2). Previous work on ppGalNAc T18 showed chaperone like activities against co-expressed ppGalNAc T2 and T10 (10), cooperatively increasing their activities. We would like to characterize ppGalNAc T18 in the presence of these and other transferases against our library of (glyco)peptides to confirm the rate increasing effects and determine whether the interaction alters ppGalNAc T2 and T10’s catalytic and lectin domain specificity. Additionally, we would like to explore the mannose specificities of these transferases since all of these transferases have increased
expressions in neuronal/brain tissues, which have elevated O-mannosylated proteins (15). Additional unique glycopeptides will be designed as substrates with a Mannose-O-Thr as needed. The results of these studies could be of high interest to the field by perhaps revealing novel specificities and biological roles for these transferases.

Studies have shown that for ppGalNAc T1, Ser residues are typically glycosylated at lower rates than Thr (16,17), but an actual experimental quantification of the differences for each isoform has not been performed. Our lab would like to address this using new series of peptide substrates, where one will contain a center Thr and another with a center Ser, that will be flanked by randomized X amino acid residues from the -3 to +3 from the acceptor site (i.e. XXXTXXX and XXXSXXX). With this approach, both substrates will be glycosylated under identical reaction conditions eliminating many confounding variables. After the isolation of the glycopeptides on the Sephadex G10, the ratio of Thr to Ser incorporation will be determined. This rate data will be incorporated into the ISOGlyP program to further enhance its predictions, since ISOGlyP does not presently distinguish between the rate differences between Ser or Thr acceptor sites. Preliminary work on this by the new graduate student in the lab, Earnest James Paul Daniel, has shown that the rates of glycosylation of Ser vs. Thr are different among the ppGalNAc T transferases.

As presented in Chapter 2, the determined transferase specific peptide substrate motifs to date do not always give the best predictions nor are these motifs always the best transferase specific substrates in vitro (unpublished). In order to further determine the transferase specific optimal motif (and further reduce variability), the Gerken Lab has obtained a series of four peptide substrates. With the first original series of substrates,
PVI-PVIII, we have determined that an optimal C-terminal P-(G/A)-P motif exists for almost all transferases and that the N-terminal positions are the most important for determining transferase specific motifs. With these two pieces of information, we have designed specialized random peptide substrates with the general format: $XAGTPGP$, $GXGTPGP$, $GAXTPGP$ and $GAGTPXP$, where the central Thr acceptor is in bold and the randomized X amino acids will only be fixed at one position in the substrate (as opposed to five). This new series of substrates will not require any new materials since they can be treated exactly like the other series of peptide substrates. We believe that this series of substrates will be very useful in further determining transferase specific motifs as well as improving ISOGlyP predictions and will be used in the lab very soon.

Several new ideas have been thought of to continue to take the random peptide approach to the next level. One of the new ideas explored by the Gerken Lab was to use di-glycopeptide substrates, where two prior sites of glycosylation T* will be placed on both the N- and C- terminal ends of the substrate that are 10 residues apart. A total of 10 randomized X’s (i.e. $XXXXXT^*XXXXX$) and Z’s (i.e. $ZZZZT^*ZZZZZ$) residues are going to be placed on either N- or C- terminal ends from the prior site of glycosylation, (i.e. $ZZZZT^*ZZZZZXXXXXT^*XXXXX$ and $XXXXXT^*XXXXXZZZZT^*ZZZZZ$) where the underlined X’s will contain a Thr acceptor and the Z’s will not. The idea is that these di-glycopeptide substrates will systematically and simultaneously address the whether the catalytic and lectin domain preferences observed previously change/not change with a substrate that has two prior sites of glycosylation. As a kind gift from the Tabak Lab at the NIH, these di-glycopeptide substrates have been synthesized. Preliminary work with these di-glycopeptide substrates with ppGalNAc T4 has confirmed
the preferred sites of glycosylation observed with the other mono-glycopeptide substrates. We are currently in the process of figuring out how to best analyze this data.

Another idea for the random (glyco)peptide substrates is to design substrates that have three dimensional structures since it was shown that ppGalNAc T11 can recognize the tertiary structures of the linker regions of the LDLR receptor (18). Although the lab needs to do more brainstorming on this idea, we have not yet come up with a way to design these substrates but will in the future. We think that these substrates will provide additional structural information about the substrate specificity of the ppGalNAc T transferases.

An essential collaboration has emerged from my thesis work is with the Hurtado-Guerro Lab in the University of Zaragoza (in Zaragoza, Spain). The lab is interested in studying the ppGalNAc T transferases as well, except that they are more focused on obtaining crystal structures of the ppGalNAc T transferases with optimal (glyco)peptides, guided from our work, to generate models to explain the roles of both the catalytic and lectin domains in (glyco)peptide binding and catalytic activity. In their most recent publication, a structure of ppGalNAc T2 was co-crystallized with a glycopeptide MUC5AC-13 (GTTPSPVPTTSTT*TSA) where the T* was bound to the lectin domain of ppGalNAc T2 (19). This structure proved our work that showed ppGalNAc T2 to have a preference for a substrate that contained a prior site of glycosylation that was on the C-terminal end of the substrate. The crystal structure also revealed that the α subdomain of the lectin domain was bound to the C-terminal T*. Furthermore, with molecular modeling, they were also able to predict the glycosylation patterns of ppGalNAc T2
showing a broad distribution of potential incorporation just as we observed with our lectin domain substrate. This type of distribution was attributed to the flexibility of the linker joining the catalytic and lectin domains.

While writing this thesis, we were asked by the Hurtado-Guerrero lab to provide an optimal sequence for ppGalNAc T4, which they have subsequently obtained a co-crystalized structure of ppGalNAc T4 with bound glycopeptide exactly as we have predicted. They are in the process of analyzing the structure and plan on publishing this in the future. This collaboration is essential since we basically determine the best optimal substrate for a given transferase while they used this information to obtain a substrate bound crystal structure of the transferase of interest. The fact that they can do this means that our work is valuable and truly capable of determining transferase specific motifs. This will help in advancing the field and understanding the mechanisms on how the ppGalNAc T catalytic and lectin domains function.

The results of these proposed studies will have a significant impact on the field and will assist our collaborators and other investigators who are attempting to understand the molecular mechanisms underlying the biology of O-glycosylation in a wide range of biological contexts. Knowing the (glyco)peptide substrate specificities and using this data to predict individual ppGalNAc T isoform glycosylation sites, using ISOGlyP will be an important tool for identifying and confirming specific glycosylation sites.
6.3 References


NUMERICAL EVP VALUES FOR ppGalNAc T13, T14, T16, T11, dT1, T4, T6, T7, PGANT9A AND PGANT9B FOR EACH AMINO ACID AT THE FLANKING POSITIONS (CHAPTER 2)
Appendix A- Numerical EVP values for ppGalNAc T13, T14, T16, T11, dT1, T4, T6 and T7 for each amino acid at the flanking positions.

**A. ppGalNAc T13**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>T13</th>
<th>T14</th>
<th>T15</th>
<th>T16</th>
<th>T17</th>
<th>T18</th>
<th>T19</th>
<th>T20</th>
<th>T21</th>
<th>T22</th>
<th>T23</th>
<th>T24</th>
<th>T25</th>
<th>T26</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.85</td>
<td>0.95</td>
<td>1.75</td>
<td>1.25</td>
<td>1.25</td>
<td>1.45</td>
<td>1.65</td>
<td>2.15</td>
<td>1.95</td>
<td>1.95</td>
<td>1.75</td>
<td>1.95</td>
<td>1.75</td>
<td>1.95</td>
</tr>
<tr>
<td>A</td>
<td>1.50</td>
<td>1.20</td>
<td>1.50</td>
<td>1.60</td>
<td>1.20</td>
<td>1.30</td>
<td>1.40</td>
<td>1.70</td>
<td>1.60</td>
<td>1.60</td>
<td>1.50</td>
<td>1.60</td>
<td>1.50</td>
<td>1.60</td>
</tr>
<tr>
<td>P</td>
<td>1.95</td>
<td>1.20</td>
<td>1.60</td>
<td>1.60</td>
<td>1.60</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
</tr>
<tr>
<td>V</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>L</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>M</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>S</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

**B. ppGalNAc T14**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>T14</th>
<th>T15</th>
<th>T16</th>
<th>T17</th>
<th>T18</th>
<th>T19</th>
<th>T20</th>
<th>T21</th>
<th>T22</th>
<th>T23</th>
<th>T24</th>
<th>T25</th>
<th>T26</th>
<th>T27</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td>A</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
</tr>
<tr>
<td>P</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>V</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>L</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>M</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>S</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

**C. ppGalNAc T16**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>T16</th>
<th>T17</th>
<th>T18</th>
<th>T19</th>
<th>T20</th>
<th>T21</th>
<th>T22</th>
<th>T23</th>
<th>T24</th>
<th>T25</th>
<th>T26</th>
<th>T27</th>
<th>T28</th>
<th>T29</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td>A</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
</tr>
<tr>
<td>P</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>V</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>L</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>M</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>F</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>S</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

291
### D. ppGalNAc T11

<table>
<thead>
<tr>
<th>ppGalNAc T11</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
<td>T15</td>
</tr>
<tr>
<td>6</td>
<td>1.20</td>
<td>1.18</td>
<td>1.08</td>
<td>1.24</td>
<td>1.59</td>
<td>0.94</td>
<td>1.97</td>
<td>1.16</td>
<td>1.30</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.28</td>
<td>1.26</td>
<td>1.41</td>
<td>1.78</td>
<td>1.72</td>
<td>1.79</td>
<td>1.36</td>
<td>1.46</td>
<td>1.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
<td>0.99</td>
<td>0.95</td>
<td>0.87</td>
<td>1.35</td>
<td>1.25</td>
<td>0.90</td>
<td>0.67</td>
<td>0.96</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.69</td>
<td>0.52</td>
<td>1.04</td>
<td>0.45</td>
<td>0.32</td>
<td>0.60</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.74</td>
<td>0.80</td>
<td>0.61</td>
<td>0.47</td>
<td>0.52</td>
<td>0.47</td>
<td>0.50</td>
<td>0.61</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.63</td>
<td>0.67</td>
<td>0.70</td>
<td>0.73</td>
<td>0.71</td>
<td>0.70</td>
<td>0.59</td>
<td>0.38</td>
<td>0.65</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
<td>0.84</td>
<td>0.70</td>
<td>0.60</td>
<td>0.54</td>
<td>0.32</td>
<td>0.34</td>
<td>0.36</td>
<td>0.57</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.97</td>
<td>1.09</td>
<td>1.21</td>
<td>1.10</td>
<td>0.95</td>
<td>0.44</td>
<td>0.70</td>
<td>0.71</td>
<td>1.01</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.93</td>
<td>0.85</td>
<td>0.94</td>
<td>0.86</td>
<td>0.61</td>
<td>0.59</td>
<td>0.63</td>
<td>0.47</td>
<td>0.67</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.82</td>
<td>0.59</td>
<td>0.62</td>
<td>0.61</td>
<td>0.49</td>
<td>0.64</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.92</td>
<td>0.77</td>
<td>0.78</td>
<td>0.81</td>
<td>0.51</td>
<td>0.41</td>
<td>0.57</td>
<td>0.41</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### E. dt1 (PGANT35A)

<table>
<thead>
<tr>
<th>dt1 (PGANT35A)</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
<td>dT1</td>
</tr>
<tr>
<td>6</td>
<td>1.20</td>
<td>1.16</td>
<td>1.66</td>
<td>1.50</td>
<td>0.88</td>
<td>1.80</td>
<td>1.03</td>
<td>1.10</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.24</td>
<td>1.16</td>
<td>1.06</td>
<td>1.54</td>
<td>1.17</td>
<td>1.33</td>
<td>0.94</td>
<td>1.02</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
<td>1.00</td>
<td>0.85</td>
<td>1.48</td>
<td>1.37</td>
<td>0.85</td>
<td>0.63</td>
<td>1.00</td>
<td>1.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.92</td>
<td>0.88</td>
<td>0.93</td>
<td>0.75</td>
<td>1.26</td>
<td>0.62</td>
<td>0.50</td>
<td>0.88</td>
<td>1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.60</td>
<td>0.70</td>
<td>0.70</td>
<td>0.60</td>
<td>0.70</td>
<td>0.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.72</td>
<td>0.75</td>
<td>0.76</td>
<td>0.51</td>
<td>0.69</td>
<td>0.66</td>
<td>0.48</td>
<td>0.77</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.74</td>
<td>0.82</td>
<td>0.80</td>
<td>0.69</td>
<td>0.43</td>
<td>0.50</td>
<td>0.44</td>
<td>0.67</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.70</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td>0.91</td>
<td>0.78</td>
<td>0.67</td>
<td>0.91</td>
<td>1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.90</td>
<td>0.81</td>
<td>0.85</td>
<td>0.75</td>
<td>0.88</td>
<td>1.00</td>
<td>0.56</td>
<td>0.89</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.56</td>
<td>0.71</td>
<td>0.70</td>
<td>0.57</td>
<td>0.56</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.16</td>
<td>1.27</td>
<td>1.20</td>
<td>1.19</td>
<td>1.00</td>
<td>1.01</td>
<td>0.87</td>
<td>1.13</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.59</td>
<td>0.64</td>
<td>0.68</td>
<td>0.60</td>
<td>0.41</td>
<td>0.46</td>
<td>0.34</td>
<td>0.43</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### F. ppGalNAc T4

<table>
<thead>
<tr>
<th>ppGalNAc T4</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
<th>PVIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
<td>T4</td>
</tr>
<tr>
<td>4</td>
<td>1.63</td>
<td>1.50</td>
<td>1.38</td>
<td>1.35</td>
<td>1.26</td>
<td>1.17</td>
<td>1.55</td>
<td>1.39</td>
<td>1.45</td>
<td>1.43</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.94</td>
<td>0.81</td>
<td>0.95</td>
<td>0.96</td>
<td>0.91</td>
<td>1.18</td>
<td>1.41</td>
<td>1.35</td>
<td>1.15</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.28</td>
<td>1.39</td>
<td>1.25</td>
<td>1.65</td>
<td>1.78</td>
<td>2.45</td>
<td>1.58</td>
<td>2.68</td>
<td>2.03</td>
<td>1.72</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.65</td>
<td>0.72</td>
<td>0.70</td>
<td>0.72</td>
<td>1.41</td>
<td>0.84</td>
<td>0.50</td>
<td>0.90</td>
<td>0.68</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.71</td>
<td>0.72</td>
<td>0.75</td>
<td>0.87</td>
<td>0.78</td>
<td>0.59</td>
<td>0.40</td>
<td>0.37</td>
<td>0.49</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.45</td>
<td>0.47</td>
<td>0.56</td>
<td>0.57</td>
<td>0.82</td>
<td>0.54</td>
<td>0.49</td>
<td>0.29</td>
<td>0.46</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.39</td>
<td>0.42</td>
<td>0.52</td>
<td>0.55</td>
<td>0.54</td>
<td>0.36</td>
<td>0.31</td>
<td>0.28</td>
<td>0.45</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.91</td>
<td>0.99</td>
<td>1.23</td>
<td>1.23</td>
<td>0.98</td>
<td>0.44</td>
<td>0.45</td>
<td>0.43</td>
<td>0.70</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.75</td>
<td>0.73</td>
<td>0.86</td>
<td>0.55</td>
<td>0.83</td>
<td>0.72</td>
<td>0.64</td>
<td>0.47</td>
<td>0.58</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.52</td>
<td>0.65</td>
<td>0.66</td>
<td>0.73</td>
<td>0.61</td>
<td>0.73</td>
<td>0.69</td>
<td>0.79</td>
<td>0.98</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.75</td>
<td>1.60</td>
<td>0.96</td>
<td>0.79</td>
<td>1.07</td>
<td>1.75</td>
<td>1.23</td>
<td>0.98</td>
<td>1.06</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.65</td>
<td>0.81</td>
<td>0.78</td>
<td>0.81</td>
<td>0.83</td>
<td>0.64</td>
<td>0.61</td>
<td>0.44</td>
<td>0.61</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.08</td>
<td>0.93</td>
<td>0.96</td>
<td>0.86</td>
<td>0.71</td>
<td>0.71</td>
<td>0.87</td>
<td>0.70</td>
<td>0.70</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.97</td>
<td>1.20</td>
<td>1.14</td>
<td>1.05</td>
<td>0.98</td>
<td>0.84</td>
<td>0.70</td>
<td>0.76</td>
<td>0.84</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.75</td>
<td>0.91</td>
<td>0.76</td>
<td>0.83</td>
<td>0.67</td>
<td>0.55</td>
<td>0.40</td>
<td>0.43</td>
<td>0.46</td>
<td>0.47</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.12</td>
<td>1.08</td>
<td>1.05</td>
<td>0.86</td>
<td>0.74</td>
<td>1.21</td>
<td>1.44</td>
<td>1.01</td>
<td>1.09</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>
G. ppGalNAc T6

<table>
<thead>
<tr>
<th>N=3</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
<th>T6</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.33</td>
<td>1.26</td>
<td>1.28</td>
<td>1.12</td>
<td>1.15</td>
<td>1.42</td>
<td>1.36</td>
<td>1.41</td>
<td>1.41</td>
<td>1.41</td>
<td>1.41</td>
</tr>
<tr>
<td>P</td>
<td>1.18</td>
<td>1.13</td>
<td>1.23</td>
<td>1.02</td>
<td>1.14</td>
<td>1.39</td>
<td>1.29</td>
<td>1.25</td>
<td>1.25</td>
<td>1.25</td>
<td>1.25</td>
</tr>
<tr>
<td>V</td>
<td>1.44</td>
<td>1.41</td>
<td>1.19</td>
<td>1.29</td>
<td>1.53</td>
<td>1.76</td>
<td>1.30</td>
<td>1.72</td>
<td>1.58</td>
<td>1.44</td>
<td>1.44</td>
</tr>
<tr>
<td>L</td>
<td>0.38</td>
<td>0.33</td>
<td>0.47</td>
<td>0.47</td>
<td>0.41</td>
<td>0.43</td>
<td>0.39</td>
<td>0.31</td>
<td>0.32</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>M</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
</tr>
<tr>
<td>F</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
<td>ERR</td>
</tr>
<tr>
<td>Y</td>
<td>0.79</td>
<td>0.85</td>
<td>0.86</td>
<td>0.85</td>
<td>0.78</td>
<td>0.61</td>
<td>0.68</td>
<td>0.78</td>
<td>0.88</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>E</td>
<td>1.18</td>
<td>0.87</td>
<td>0.84</td>
<td>0.74</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.79</td>
<td>0.84</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>S</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>S</td>
<td>1.95</td>
<td>1.44</td>
<td>1.47</td>
<td>1.41</td>
<td>1.37</td>
<td>2.09</td>
<td>1.55</td>
<td>1.69</td>
<td>1.17</td>
<td>1.23</td>
<td>1.23</td>
</tr>
</tbody>
</table>

H. ppGalNAc T7

<table>
<thead>
<tr>
<th>N=4</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
<th>T7</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.62</td>
<td>1.42</td>
<td>1.48</td>
<td>1.53</td>
<td>1.28</td>
<td>1.22</td>
<td>1.28</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
<td>1.37</td>
</tr>
<tr>
<td>A</td>
<td>1.22</td>
<td>1.20</td>
<td>1.17</td>
<td>1.12</td>
<td>1.08</td>
<td>1.42</td>
<td>1.19</td>
<td>1.11</td>
<td>1.11</td>
<td>1.11</td>
<td>1.11</td>
</tr>
<tr>
<td>P</td>
<td>1.15</td>
<td>1.15</td>
<td>1.14</td>
<td>1.17</td>
<td>1.30</td>
<td>1.67</td>
<td>1.23</td>
<td>1.34</td>
<td>1.34</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>T</td>
<td>1.01</td>
<td>1.05</td>
<td>1.08</td>
<td>1.17</td>
<td>0.87</td>
<td>0.97</td>
<td>1.19</td>
<td>1.34</td>
<td>1.34</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>I</td>
<td>0.73</td>
<td>0.71</td>
<td>0.68</td>
<td>0.64</td>
<td>0.50</td>
<td>0.36</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>L</td>
<td>0.94</td>
<td>1.02</td>
<td>1.05</td>
<td>0.78</td>
<td>0.81</td>
<td>0.60</td>
<td>0.47</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>M</td>
<td>0.51</td>
<td>0.50</td>
<td>0.52</td>
<td>0.42</td>
<td>0.43</td>
<td>0.38</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>F</td>
<td>0.53</td>
<td>0.54</td>
<td>0.46</td>
<td>0.60</td>
<td>0.43</td>
<td>0.36</td>
<td>0.59</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>Y</td>
<td>1.01</td>
<td>0.99</td>
<td>1.26</td>
<td>1.74</td>
<td>1.06</td>
<td>1.03</td>
<td>1.93</td>
<td>1.59</td>
<td>1.59</td>
<td>1.59</td>
<td>1.59</td>
</tr>
<tr>
<td>C</td>
<td>0.80</td>
<td>0.71</td>
<td>0.73</td>
<td>0.96</td>
<td>0.84</td>
<td>0.82</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>D</td>
<td>0.58</td>
<td>0.59</td>
<td>0.63</td>
<td>0.64</td>
<td>0.84</td>
<td>0.59</td>
<td>0.95</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>Q</td>
<td>0.96</td>
<td>0.86</td>
<td>0.86</td>
<td>0.10</td>
<td>1.25</td>
<td>0.86</td>
<td>0.84</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>S</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>1.53</td>
<td>1.41</td>
<td>1.28</td>
<td>1.28</td>
<td>1.28</td>
<td>1.28</td>
<td>1.28</td>
</tr>
</tbody>
</table>

I. PGANT9A

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.38</td>
<td>1.40</td>
<td>1.22</td>
<td>1.33</td>
<td>1.17</td>
<td>1.01</td>
<td>1.55</td>
<td>1.19</td>
<td>1.31</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>P</td>
<td>1.21</td>
<td>1.09</td>
<td>1.10</td>
<td>1.11</td>
<td>1.11</td>
<td>0.98</td>
<td>1.20</td>
<td>1.00</td>
<td>1.02</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>V</td>
<td>1.04</td>
<td>1.04</td>
<td>1.02</td>
<td>1.04</td>
<td>1.38</td>
<td>2.61</td>
<td>1.19</td>
<td>3.51</td>
<td>2.12</td>
<td>1.43</td>
<td>1.43</td>
</tr>
<tr>
<td>L</td>
<td>0.93</td>
<td>1.01</td>
<td>1.20</td>
<td>0.87</td>
<td>2.14</td>
<td>1.55</td>
<td>0.87</td>
<td>0.85</td>
<td>1.00</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>M</td>
<td>0.27</td>
<td>0.31</td>
<td>0.79</td>
<td>0.63</td>
<td>0.73</td>
<td>0.54</td>
<td>0.57</td>
<td>0.51</td>
<td>0.52</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>F</td>
<td>0.36</td>
<td>0.70</td>
<td>0.50</td>
<td>0.73</td>
<td>0.42</td>
<td>0.57</td>
<td>0.47</td>
<td>0.46</td>
<td>0.70</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Y</td>
<td>1.07</td>
<td>0.96</td>
<td>1.09</td>
<td>1.12</td>
<td>1.24</td>
<td>0.55</td>
<td>0.57</td>
<td>0.86</td>
<td>1.09</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>D</td>
<td>1.24</td>
<td>1.22</td>
<td>1.12</td>
<td>0.99</td>
<td>0.66</td>
<td>1.05</td>
<td>0.94</td>
<td>0.72</td>
<td>1.06</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>Q</td>
<td>0.98</td>
<td>1.06</td>
<td>1.10</td>
<td>0.94</td>
<td>0.60</td>
<td>1.04</td>
<td>0.81</td>
<td>0.56</td>
<td>0.82</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>K</td>
<td>1.13</td>
<td>1.04</td>
<td>1.99</td>
<td>1.19</td>
<td>0.88</td>
<td>0.83</td>
<td>0.74</td>
<td>0.63</td>
<td>0.87</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>E</td>
<td>0.82</td>
<td>0.86</td>
<td>0.94</td>
<td>0.96</td>
<td>0.77</td>
<td>0.84</td>
<td>1.18</td>
<td>0.71</td>
<td>0.80</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>M</td>
<td>0.56</td>
<td>0.62</td>
<td>0.57</td>
<td>0.50</td>
<td>0.42</td>
<td>0.44</td>
<td>0.45</td>
<td>0.40</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>H</td>
<td>0.68</td>
<td>0.68</td>
<td>0.72</td>
<td>0.69</td>
<td>0.46</td>
<td>0.40</td>
<td>0.37</td>
<td>0.30</td>
<td>0.46</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>S*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>S</td>
<td>1.05</td>
<td>0.96</td>
<td>1.07</td>
<td>1.07</td>
<td>1.24</td>
<td>0.91</td>
<td>1.36</td>
<td>1.05</td>
<td>1.06</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>
J. PGANT9B

<table>
<thead>
<tr>
<th>N=10</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
<th>PGANT9B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td>+2</td>
<td>+3</td>
</tr>
<tr>
<td>G</td>
<td>1.97</td>
<td>1.60</td>
<td>1.34</td>
<td>1.56</td>
<td>1.28</td>
<td>1.06</td>
<td>1.80</td>
<td>1.26</td>
<td>1.26</td>
<td>1.37</td>
</tr>
<tr>
<td>A</td>
<td>0.96</td>
<td>0.92</td>
<td>0.94</td>
<td>1.04</td>
<td>0.96</td>
<td>0.93</td>
<td>1.26</td>
<td>0.91</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>P</td>
<td>0.96</td>
<td>1.52</td>
<td>0.98</td>
<td>0.98</td>
<td>1.34</td>
<td>2.06</td>
<td>1.13</td>
<td>3.11</td>
<td>1.82</td>
<td>1.46</td>
</tr>
<tr>
<td>V</td>
<td>0.94</td>
<td>0.99</td>
<td>1.17</td>
<td>0.78</td>
<td>2.11</td>
<td>1.52</td>
<td>0.90</td>
<td>0.80</td>
<td>0.99</td>
<td>1.18</td>
</tr>
<tr>
<td>I</td>
<td>0.10</td>
<td>0.04</td>
<td>0.05</td>
<td>0.08</td>
<td>0.10</td>
<td>0.21</td>
<td>0.21</td>
<td>0.20</td>
<td>0.22</td>
<td>0.19</td>
</tr>
<tr>
<td>L</td>
<td>0.61</td>
<td>0.70</td>
<td>0.75</td>
<td>0.76</td>
<td>0.38</td>
<td>0.57</td>
<td>0.49</td>
<td>0.48</td>
<td>0.72</td>
<td>1.00</td>
</tr>
<tr>
<td>M</td>
<td>0.39</td>
<td>0.47</td>
<td>0.59</td>
<td>0.59</td>
<td>0.64</td>
<td>0.64</td>
<td>0.69</td>
<td>0.65</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td>F</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Y</td>
<td>1.02</td>
<td>1.03</td>
<td>1.02</td>
<td>1.08</td>
<td>1.17</td>
<td>0.49</td>
<td>0.57</td>
<td>0.87</td>
<td>1.17</td>
<td>1.21</td>
</tr>
<tr>
<td>E</td>
<td>1.09</td>
<td>1.04</td>
<td>0.89</td>
<td>0.76</td>
<td>0.45</td>
<td>0.76</td>
<td>0.78</td>
<td>0.66</td>
<td>0.82</td>
<td>0.95</td>
</tr>
<tr>
<td>D</td>
<td>1.34</td>
<td>1.10</td>
<td>1.18</td>
<td>1.34</td>
<td>1.41</td>
<td>1.74</td>
<td>1.64</td>
<td>1.42</td>
<td>1.29</td>
<td>1.25</td>
</tr>
<tr>
<td>Q</td>
<td>1.09</td>
<td>1.10</td>
<td>1.08</td>
<td>0.96</td>
<td>0.61</td>
<td>1.08</td>
<td>0.87</td>
<td>0.63</td>
<td>0.77</td>
<td>1.00</td>
</tr>
<tr>
<td>N</td>
<td>0.73</td>
<td>0.76</td>
<td>0.82</td>
<td>0.94</td>
<td>0.75</td>
<td>0.91</td>
<td>0.95</td>
<td>0.97</td>
<td>1.13</td>
<td>1.10</td>
</tr>
<tr>
<td>K</td>
<td>1.96</td>
<td>1.90</td>
<td>1.11</td>
<td>1.03</td>
<td>0.74</td>
<td>1.01</td>
<td>0.83</td>
<td>0.72</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td>H</td>
<td>0.60</td>
<td>0.65</td>
<td>0.78</td>
<td>0.67</td>
<td>0.46</td>
<td>0.64</td>
<td>0.61</td>
<td>0.52</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>D</td>
<td>1.05</td>
<td>0.75</td>
<td>0.77</td>
<td>0.74</td>
<td>0.48</td>
<td>0.43</td>
<td>0.37</td>
<td>0.35</td>
<td>0.47</td>
<td>0.54</td>
</tr>
<tr>
<td>S*</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>S</td>
<td>0.96</td>
<td>1.01</td>
<td>1.08</td>
<td>1.13</td>
<td>1.47</td>
<td>1.63</td>
<td>1.52</td>
<td>1.14</td>
<td>1.17</td>
<td>0.88</td>
</tr>
</tbody>
</table>
APPENDIX B

FIRST AND SECOND AUTHOR PUBLICATIONS
Glycobiology and Extracellular Matrices: The Lectin Domain of the Polypeptide GalNAc Transferase Family of Glycosyltransferases (ppGalNAc Ts) Acts as a Switch Directing Glycopeptide Substrate Glycosylation in an N- or C-terminal Direction, Further Controlling Mucin Type O-Glycosylation

Thomas A. Gerken, Leslie Revoredo, Joseph J. C. Thome, Lawrence A. Tabak, Malene Bech Vester-Christensen, Henrik Clausen, Gagandeep K. Gahlay, Donald L. Jarvis, Roy W. Johnson, Heather A. Moniz and Kelley Moremen

J. Biol. Chem. 2013, 288:19900-19914. doi: 10.1074/jbc.M113.477877 originally published online May 20, 2013

Access the most updated version of this article at doi: 10.1074/jbc.M113.477877

Find articles, minireviews, Reflections and Classics on similar topics on the JBC Affinity Sites.

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

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

This article cites 74 references, 46 of which can be accessed free at http://www.jbc.org/content/288/27/19900.full.html#ref-list-1
The Lectin Domain of the Polypeptide GalNAc Transferase Family of Glycosyltransferases (ppGalNAc Ts) Acts as a Switch Directing Glycopeptide Substrate Glycosylation in an N- or C-terminal Direction, Further Controlling Mucin Type O-Glycosylation

Received for publication, April 16, 2013, and in revised form, May 17, 2013 Published, JBC Papers in Press, May 20, 2013, DOI 10.1074/jbc.M113.477877


From the Departments of Pediatrics (W. A. Bernbaum Center for Cystic Fibrosis Research), Biochemistry, and Chemistry, Case Western Reserve University, Cleveland, Ohio 44106, the Section on Biological Chemistry, NIDCR, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20892, the Copenhagen Center for Glycomics (CCG), Departments of Cellular and Molecular Medicine and Dentistry, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen N, Denmark, the Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, and the Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602

Mucin type O-glycosylation is initiated by a large family of polypeptide GalNAc transferases (ppGalNAc Ts) that add α-GalNAc to the Ser and Thr residues of peptides. Of the 20 human isoforms, all but one are composed of two globular domains linked by a short flexible linker: a catalytic domain and a ricin-like lectin carbohydrate binding domain. Presently, the roles of the catalytic and lectin domains in peptide and glycopeptide recognition and specificity remain unclear. To systematically study the role of the lectin domain in ppGalNAc T glycopeptide substrate utilization, we have developed a series of novel random glycopeptide substrates containing a single GalNAc-O-Thr residue placed near either the N or C terminus of the glycopeptide substrate. Our results reveal that the presence and N- or C-terminal placement of the GalNAc-O-Thr can be important determinants of overall catalytic activity and specificity that differ between transferase isoforms. For example, ppGalNAc T1, T2, and T14 prefer C-terminally placed GalNAc-O-Thr, whereas ppGalNAc T3 and T6 prefer N-terminally placed GalNAc-O-Thr. Several transferase isoforms, ppGalNAc T5, T13, and T16, display equally enhanced N- or C-terminal activities relative to the nonglycosylated control peptides. This N- and/or C-terminal selectivity is presumably due to weak glycopeptide binding to the lectin domain, whose orientation relative to the catalytic domain is dynamic and isoform-dependent. Such N- or C-terminal glycopeptide selectivity provides an additional level of control or fidelity for the O-glycosylation of biologically significant sites and suggests that O-glycosylation may in some instances be exquisitely controlled.

Mucin type protein O-glycosylation, as defined by the α-GalNAc-O-Ser/Thr linkage, is one of the most common types of protein glycosylation found in higher organisms. This modification is initiated in the Golgi by a large family (~20 members in mammals and about half the number in the fly and Caenorhabditis elegans (see Ref. 1 for a review) of polypeptide GalNAc transferases (ppGalNAc Ts) that transfer GalNAc from UDP-GalNAc to Ser and Thr. Subsequent chain elongation of the GalNAc residue by the stepwise action of additional specific transferases results in a vast array of heterogeneous O-linked glycan structures. In contrast to the initiating ppGalNAc Ts, nearly all of these O-glycan elongating transferases occur in small families of 1–4 members. Despite its prevalence, the biological roles and control of mucin type O-glycosylation remain incompletely understood.

§§This work was supported, in whole or in part, by National Institutes of Health Grants R01-CA78834 (to T. A. G.) and P41-GM103390 and P41-RR015351 (to James Prestegard (Complex Carbohydrate Center, University of Georgia) and K. M. and D. L. J.) and a National Institutes of Health, NIDCR, intramural program grant (to L. A. T.). This work was also supported by Danish National Research Foundation Grant DNRF107 (to H. C.).

†To whom correspondence should be addressed: Dept. of Pediatrics, Case Western Reserve University, School of Medicine, BRB 823, 2109 Adelbert Rd. Cleveland, OH 44104-4948. Tel.: 216-368-4556; Fax: 216-368-4223; E-mail: txg2@cwru.edu.

‡Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

**To whom correspondence should be addressed: Dept. of Pediatrics, Case Western Reserve University, School of Medicine, BRB 823, 2109 Adelbert Rd. Cleveland OH 44104-4948. Tel.: 216-368-4556; Fax: 216-368-4223; E-mail: txg2@cwru.edu.

††Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

‡‡Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

¶¶Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

 §§Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

##Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

###Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

####Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

**The abbreviations used are:** ppGalNAc T, UDP-α-GalNAc-polypeptide N-acetylgalactosaminyltransferase; PTH, phenylthiohydantoin amino acid derivative of Edman sequencing; Ni-NTA, nickel-nitrilotriacetic acid; GPIV, GPIV-C, GPV, and GPV-C, random (glyco)peptides as defined in Table 1; HDPE, high density polyethylene.

*This work was supported, in whole or in part, by National Institutes of Health Grants R01-CA78834 (to T. A. G.) and P41-GM103390 and P41-RR015351 (to James Prestegard (Complex Carbohydrate Center, University of Georgia) and K. M. and D. L. J.) and a National Institutes of Health, NIDCR, intramural program grant (to L. A. T.). This work was also supported by Danish National Research Foundation Grant DNRF107 (to H. C.).

To whom correspondence should be addressed: Dept. of Pediatrics, Case Western Reserve University, School of Medicine, BRB 823, 2109 Adelbert Rd. Cleveland, OH 44104-4948. Tel.: 216-368-4556; Fax: 216-368-4223; E-mail: txg2@cwru.edu.

Present address: Dept. of Microbiology and Immunology, Columbia University Graduate School of Arts and Sciences at the College of Physicians and Surgeons, Columbia University, New York, NY 10027.

**The abbreviations used are:** ppGalNAc T, UDP-α-GalNAc-polypeptide N-acetylgalactosaminyltransferase; PTH, phenylthiohydantoin amino acid derivative of Edman sequencing; Ni-NTA, nickel-nitrilotriacetic acid; GPIV, GPIV-C, GPV, and GPV-C, random (glyco)peptides as defined in Table 1; HDPE, high density polyethylene.
Mucin type O-glycans have typically been associated with so-called mucin domains, where high numbers of Ser and Thr residues are glycosylated, producing a chemical- and protease-resistant extended peptide structural motif. Glycoproteins containing O-glycosylated mucin domains commonly function in the protection of the cell surface and the modulation of cell-cell interactions and hence play important roles, for example in inflammation, the immune response, metastasis, and tumorigenesis (2–6). Because the role of the mucin domain in these glycoproteins has been thought to produce a stable extended hydrophilic motif, the actual glycosylation pattern has usually been thought to be of relatively low importance. This is supported by the very low sequence conservation across mammalian and insect species of the glycosylated domains of the secreted mucins (7, 8). Nevertheless, it cannot be ruled out that in some cases, O-glycosylated domains may present a molecular code for the specific recognition of additional binding partners, thus playing potential roles in protein sorting, targeting, and turnover (9–15).

Recent proteome-wide analysis demonstrates that mucin type O-glycosylation is widely distributed on most proteins passing through the secretory pathway and that most have only one or few isolated O-linked glycans (16–18). Thus, it is likely that O-linked glycans at specific sites may play other important biological roles. One emerging role is the co-regulation of the proprotein processing process (19, 20), where O-glycans modulate cleavage of nearby proprotein convertase cleavage sites (i.e., phosphaturic factor, FGF23, and angiopoietin-like protein 3 (ANGPTL3) (21, 22)). Other studies have shown that specific O-glycosylation can modulate receptor activities (i.e. TGF-βR II and the death receptor DR5) by presently unknown mechanisms (23, 24). Several of the above O-glycosylation events are targeted by specific ppGalNAc T isoforms (i.e. ppGalNAc T2, T3, T14, and T16); therefore, some occurrences of O-glycosylation may be actively regulated to serve as modulators of complex biological processes and perhaps even signaling. Unsurprisingly, a number of distinct ppGalNAc T isoforms have been shown to cause or be associated with human disease. The loss of ppGalNAc T3 causes familial tumoral calcinosis due to abnormal processing of FGF23 (21, 25), ppGalNAc T2, T5, and T19 (also known as WBSCR17) are linked to Williams-Beuren syndrome (15), hereditary multiple exostoses (26), and levels of HDL cholesterol and coronary artery disease (22, 27, 28), respectively, whereas ppGalNAc T3, T6, T7, T9, and T12 are associated with various cancers (29–34). Except for the known role of ppGalNAc T3 in modulating the cleavage and inactivation of FGF23, the mechanisms for causing disease by the other transferase isoforms are still uncertain.

Recent studies have further demonstrated the critical role of O-glycosylation in vertebrate and invertebrate development (35–39). In the fly, several ppGalNAc T isoforms are required to complete development (38), whereas in the mouse, ppGalNAc T1 modulates salivary gland organogenesis (40). Interestingly, the loss of the elongating transferase, T-synthase (which adds a β-Gal to the 3-position of the peptide GalNAc) leads to embryonic lethality in the mouse (35). Although a few target proteins have been implicated in the above studies, little is known about the actual site(s) of glycosylation and their specificity, and even less is known of the actual molecular mechanism(s) leading to their biological function.

Structurally, members of the ppGalNAc T family (except h-ppGalNAc T20 (41)) possess a unique two-domain architecture consisting of a transmembrane tethered N-terminal catalytic domain linked via a short flexible segment to a C-terminal ricin-like lectin domain containing three potential carbohydrate binding sites (see Fig. 1 for the crystal structure of ppGalNAc T2 with bound EA2 peptide substrate (42)). Presently, the roles of the catalytic and lectin domains in peptide and glycopeptide recognition and specificity remain unclear. Our recent studies utilizing short random (glyco)peptide substrates have shown that the ppGalNAc Ts possess specific binding preferences that vary among isoforms for peptide and even GalNAc-O-Ser/Thr-containing glycopeptide substrates (43–45). Interestingly, we and others have shown that the placement of a neighboring GalNAc-O-Ser/Thr residue near a glycosylation site produces a range of transferase-specific effects (i.e. a relative inhibition of glycosylation (i.e. ppGalNAc T1 and T2 (46, 47)), an alteration or shift in glycosylation site (ppGalNAc T2 and T4 (48–51)), or even a large apparent rate enhancement (i.e. requirement for glycosylated substrate) (ppGalNAc T7 and T10 (45, 51–53)). Interestingly, the alteration in site preference observed for ppGalNAc T2 and T4 resides in their lectin domain (48, 50, 51), whereas the nearly absolute glycopeptide requirement of ppGalNAc T10 resides completely in its catalytic domain (45, 51). Recent studies characterizing the properties of the lectin domains of ppGalNAc T2 and T3 further suggest that they may recognize glycopeptide sequence context (54) and only glycopeptides having O-linked α-GalNAc residues (55), respectively. Because, to date, there have been no truly systematic studies of
Lectin Domain Control of Mucin Type O-Glycosylation

ppGalNAc T glycopeptide substrate utilization, we have now extended our studies to the function of the lectin domain, utilizing a series of random glycopeptides containing N- or C- terminally placed GalNAc-O-Thr residues. Our results show that the presence and the N- or C-terminal location of a GalNAc-O-Thr site in these glycopeptides are indeed important determinants of the overall catalytic activity and specificity of these enzymes, which can significantly differ between transferase isoforms. Because of the large differences in glycosylation rates observed for some isoforms between an N- or C-terminally placed GalNAc-O-Thr, we believe we have now uncovered another level of control of mucin type O-glycosylation that will further advance our understanding of the regulation of this important modification.

EXPERIMENTAL PROCEDURES

Reagents and Random Peptide Substrates—Fully N-acetylated UDP-[3H]GalNAc (C6-[3H]CH2-labeled) and UDP-[14C]GalNAc (NAc-[14C]CH3-labeled) were purchased from American Radiolabeled Chemicals Inc. Dowex 1X8 anion exchange resin was purchased from Acros Organics. Sephacryl G-10 was obtained from Amersham Biosciences and GE Healthcare. Random (glyco)peptide substrates GPIV, GPIV-C, GPV, and GPV-C (see “Results” and Table 1 for sequences) were custom synthesized by Sussex Research (Ottawa, CA). Random (glyco)peptide substrates were lyophilized from water several times, adjusted to pH 7.5 with dilute NaOH and/or HCl, and stored frozen at a concentration of 50 mg/ml from water several times, adjusted to pH 7.5 with dilute NaOH.

GPIV, GPIV-C, GPV, and GPV-C (see “Results” and Table 1 for sequences) were custom synthesized by Sussex Research (Ottawa, CA). Random (glyco)peptide substrates were lyophilized from water several times, adjusted to pH 7.5 with dilute NaOH and/or HCl, and stored frozen at a concentration of 50 mg/ml (15–17 mM). Similar random (glyco)peptide X and Z residue compositions were confirmed by Edman amino acid sequencing (data not shown). Edman amino acid sequencing was performed on an Applied Biosystems Procise 494 peptide sequencer as described previously (43–45, 56). Liquid scintillation counting was performed in a Beckman model LS6500 scintillation counter.

Transferases—N-terminally truncated and affinity-tagged ppGalNAc Ts used in this work were obtained from multiple sources and multiple expression systems and used as one or more of the following forms: medium supernatant, affinity-purified soluble transferase, or transferase bound to affinity beads. Experiments completed using transferases from different sources yielded consistent results. Soluble affinity-purified bovine ppGalNAc T1 (from Sf9 insect cells) and human ppGalNAc T2 and T3 (from Pichia pastoris) have been described previously (43–45, 56). ppGalNAc T3 was also used as medium supernatant from P. pastoris as described (43). Poly-His-tagged ppGalNAc T3, T6, and T13, expressed from Xpress SF+ insect cells (Protein Sciences Corp., Meriden, CT), were used directly bound to Ni-NTA affinity beads (Thermo Fisher) after extensive washing. Soluble N-terminal poly-His-tagged ppGalNAc T5 and T16 were expressed from HEK293F cells and purified by Ni-NTA superflow (Qiagen) nickel affinity chromatography by procedures analogous to those for the production of rat ST6Gal1 (57). Soluble N-terminal poly-His-tagged ppGalNAc T14 and T16 were expressed from High Five insect cells and purified on Ni-NTA-agarose (Invitrogen) and MonoQ 5/50 GL ion exchange chromatography (GE Healthcare) essentially as described previously (58).

Glycosylation of Random Lectin Glycopeptide Substrates—Typical reaction conditions utilizing random lectin glycopeptide substrates GPIV, GPIV-C, GPV, and GPV-C were as follows: 100 mM HEPES, pH 7.5 (ppGalNAc T1, T2, and T3) or 68 mM sodium cacodylate, pH 6.5 (ppGalNAc T3, T5, T6, T13, T14, and T16), 1.8 mM 2-mercaptoethanol, 10 mM MnCl2, 50 μM [3H]radiolabeled UDP-GalNAc (~6 × 108 dpm/μmol), and 1.5–1.7 mM (5 mg/ml) of glycopeptide substrates and up to 50 μl of soluble transferase or transferase-bound affinity beads. UDP-GalNAc concentrations were kept low (0.5–50 μM) to achieve high specific activity for subsequent Edman sequence analysis (described below). After the addition of transferase, reaction mixtures (75 μl in 500 μl of capped Eppendorf tubes) were incubated at 37 °C in a TAITEC shaking Microincubator M-36. Aliquots of 15 μl were taken out at time points of 15, 45, 120, 240, and ~1200 min (overnight) after initiating the reaction and quenched by the addition of 1 volume of 250 mM EDTA. UDP and non-hydrolyzed UDP-GalNAc were removed by passing the sample through a column of ~3 ml of Dowex 1X 8 anion exchange resin. [3H]GalNAc incorporation was determined by scintillation counting 1/10 or 1/20 of the sample before and after passing through the Dowex column. Data are typically reported as plots of the mole fraction of total UDP-GalNAc transferred (including hydrolysis) versus time. Confirmation of [3H]GalNAc transfer to peptide and the extent of UDP-[3H]GalNAc hydrolysis was determined by Sephadex G10 gel filtration analysis of selected reaction time points as described previously (44). For all transferases reported in this study, the extent of hydrolysis was negligible (see “Results”). Incubations were performed using a range of transferase concentrations (determined by trial and error) that would transfer between 10 and 50% of the total [3H]GalNAc to the optimal substrate after an overnight incubation (giving a range of ~0.003–0.016 mol of GalNAc/mol of glycopeptide). In total, 2–4 independent experiments were performed at 2–3 different transferase concentrations (except for ppGalNAc T14).

Determination of Optimal Reaction Conditions and Initial Observations—Initial use of the GPIV and GPV series of substrates (see Table 1) with ppGalNAc T1 and T2 gave results that seemed to vary with (glyco)peptide concentration. For example, at the lowest substrate concentrations (~0.16 or ~0.016 mM (glyco)peptide with T1 and T2, respectively), all four substrates were essentially inactive, whereas at the highest substrate concentration (1.5–1.7 mM), differential activities were observed, with the glycopeptide substrates displaying significantly greater activity compared with the control peptide substrates (see Fig. 2). We attribute these differential effects to peptide concentrations initially below the minimum Kapp value of the substrate and therefore have chosen to perform our studies utilizing substrate concentrations of 1.5–1.7 mM (5 mg/ml). Because of the random and multiple acceptor site nature of these substrates, a continuum of Kapp and Vmax values is expected to be observed; therefore, we have not pursued a detailed kinetic analysis of these substrates. It was further observed that the relative rates of glyco-
ylation of the four substrates varied somewhat with the amount of transferase used in the assay; nevertheless, overall transferase preferences were preserved (data not shown).

A commonly observed feature of the transferase activity plots shown below is the plateau in transfer of GalNAc that is often observed with high transferase activities, even for the least active substrates. We attribute this plateau to a combination of several likely factors. For the most active substrates (approaching 50% GalNAc utilization), the plateau in incorporation is most likely dominated by the depletion of UDP-GalNAc donor. For the less active substrates, we attribute the slowing of the rate to the sequential loss of the optimal or best acceptor sites as glycosylation of the random (glyco)peptide progresses. Because the half-lives of ppGalNAc T1 and T2 are about 5 and 2.5 days (59), it is unlikely that transferase inactivation is the source of the plateau with these transferases. However, transferase inactivation cannot be ruled out for the remaining transferases.

**Determination of GalNAc Acceptor Sites of Incorporation by Edman Sequencing**—Lyophilized post-Dowex \[^3H\]GalNAc-glycosylated peptide products were isolated by Sephadex G10 chromatography, and the glycopeptide fraction was pooled according to its \[^3H\]radioactivity. After multiple lyophilizations from water, glycopeptides were subject to Edman degradation using a custom PTH analyzer program that diverted the PTH-derivatives from each cycle directly into 7-ml HDPE scintillation vials (Fisher) loaded in a 100-tube rack on an ISCO Foxy 200 fraction collector. It was observed that collecting directly into scintillation vials significantly reduced nonspecific losses.
of the radiolabeled PTH-derivatives compared with collecting in glass or plastic test tubes. To improve the signal/noise ratio, counting was performed for 5 min/vial and typically repeated 2–3 times, with the resulting dpm values averaged before plotting. To reduce nonspecific binding of the \(^{3}H\)GalNAc-O-Thr-PTH-derivative to the glass fiber sample loading disk (determined by counting the disk after sequencing), ~0.25 mg of nonreacted “cold” (glyco)peptide substrate was commonly added to the sample to be analyzed. Nevertheless, significant variability in recovery of radiolabel was observed. A preview in radioactivity observed in the N-terminal residues, particularly noticeable in the GPV and GPV-C (glyco)peptides, is thought to be due to nonspecific release of (glyco)peptide from the glass filter during the Edman sequencing.

**RESULTS**

**Design of Lectin Glycopeptide Substrates**—The goal of these studies was to develop a series of “universal” substrates that could be used with any ppGalNAc T isoform to access the role of each isoform’s lectin domain. Peptides were designed with a single C- or N-terminal GalNAc-O-Thr residue (T*), designed to interact with the lectin domain, flanked by five positions of randomized residues (Z) containing no acceptor (Ser or Thr) residues (Table 1). The Z regions flanking the T* contained no acceptor Thr sites specifically to eliminate glycopeptide activities arising from the direct binding of GalNAc-O-Thr in the catalytic domain as is observed for ppGalNAc T10, which optimally glycosylates a Ser/Thr directly preceding a GalNAc-O-Ser/Thr (45, 51). Beyond these residues, in an N- or C-terminal direction, respectively, are 12 positions of randomized X residues also containing Thr that would serve as the acceptor region by its interaction with the catalytic domain. As a control for lectin domain specificity, peptides (GPIV-C and GPV-C) were also synthesized with the T* replaced with an Ala residue. The rationale behind utilizing randomized residues, X and Z, rather than a specific peptide sequence was based on our goal to eliminate (or at least reduce) transferase-specific bias due to a particular peptide sequence. With this series of (glyco)peptides the role of the lectin domain of the ppGalNAc Ts against glycopeptide substrates can now be systematically accessed, specifically addressing 1) the utilization of the lectin domain on overall transferase activity; 2) the preferred N- or C-terminal orientation, if any, of the T* relative to the site of glycosylation; and 3) the optimal number of residues between the lectin-bound T* and those residues glycosylated by the catalytic domain.

**Glycosylation of the GPV-GPIV Series of (Glyco)peptides by ppGalNAc Ts**—Representative time course plots and gel filtration profiles demonstrating the differential transfer of \(^{3}H\)GalNAc into peptides GPIV, GPIV-C, GPV, and GPV-C by the eight ppGalNAc Ts under study are given in Fig. 3. As
shown by the right-hand panels of Fig. 3, the G10 gel filtration analysis reveals relatively little UDP-GalNAc hydrolysis activity for all of the transferases characterized (i.e. high $^3$H in peptide fractions ~27–33 and little $^3$H where free GalNAc would appear, fractions ~37–43). Therefore, the time plots of the ion exchange column pass-through (Fig. 3, left-hand panels) can be taken as a direct measure of transfer of GalNAc to (glyco)peptide. To aid in the discussion, a phylogenetic tree and a subfamily classification of the human transferases (1) are shown in Fig. 4, where a graphical summary of glycopeptide utilization deduced from the data in Fig. 3 is also given. We will discuss transferase isoforms with common behaviors below.

GPIV-preferring Transferases; ppGalNAc T1, T2, and T14—As shown in Fig. 3, A–C (and data not shown), these subfamily Ia and Ib transferases are significantly more active against the C-terminal glycosylated glycopeptide, GPIV, compared with its nonglycosylated control and the N-terminal glycosylated analog (glyco)peptide substrates, GPV and GPV-C. This observation suggests that the binding of the C-terminal T* of GPIV to the transferase lectin domain enhances its activity compared with the other substrates. This may be due to both an effective increase in substrate concentration due to repeated lectin binding and release or to an optimal alignment of the acceptor in the catalytic domain due to specific lectin binding.

To confirm that lectin binding is involved with the enhanced GPIV activity, the effects of adding free GalNAc competitor were examined on ppGalNAc T2. As shown in Fig. 5, A and B, the addition of free GalNAc significantly reduced the activity of GPIV while not affecting the activity of the nonglycosylated control peptide. It should be further noted that under the higher (glyco)peptide and transferase and lower UDP-GalNAc concentrations utilized in this and many of our initial experi-
ments, GPV also shows an elevated activity relative to GPV-C (see Fig. 2C) that is completely inhibited by added free GalNAc (Fig. 5, C and D). This suggests that the GPV glycopeptide substrate can interact with the lectin domain of ppGalNAc T2 under certain conditions, also leading to a rate enhancement.

The x-ray crystal structure of ppGalNAc T2 (see Fig. 1) shows nine residues of the EA2 peptide substrate bound to the catalytic domain in a specific N to C terminus orientation (42). By model building, we have extended the EA2 peptide to mimic the bound GPIV and GPV glycopeptides, as shown in Fig. 6A (left- and right-hand panels, respectively). In the models, we observe that the glycosylated C terminus (~XZZZZZZTZ*ZZZZZZXAG) of GPIV is extended to the right of the catalytic domain, whereas the glycosylated N-terminal (GAGAZZZZZTZ*ZZZZZX/H11011) of GPV is extended to the left of the catalytic domain. Note that for simplicity, the extended regions were modeled as Ala residues in an extended conformation; however, in reality, an ensemble of dynamic structures would be expected. The observed elevated activity of GPIV over GPV for ppGalNAc T1 and T14, therefore, would suggest that the lectin domain of these transferases would be positioned toward the right of the catalytic domain such that its Xaa residues would be aligned optimally in the catalytic domain. Such domain flexibility seems plausible in solution, because the relative positions of the catalytic and lectin domains are highly variable among the several crystal structures of ppGalNAc T1, T2, and T10 shown in Fig. 6B (42, 60, 61). This orientational flexibility is undoubtedly the result of the ~20-residue extended linker that connects the domains (see Table 2).

Glycopeptide GPV-preferring Transferases; ppGalNAc T3 and T6—These subfamily Ic transferases (Fig. 3, D and E) (and data not shown) display the reversed preferences of ppGalNAc T1, T2, and T14, showing GPV as the most active substrate with the remaining substrates being relatively inactive. This suggests that the lectin domains of ppGalNAc T3 and T6 preferentially interact with T* residues N-terminal of the site of subsequent glycosylation, suggesting that the lectin domain of these transferases must be positioned toward the left of the catalytic domain, as shown in the right panel of Fig. 6A.

Transferases with Preferences for both Glycopeptide Substrates GPIV and GPV; ppGalNAc T5, T13, and T16—As shown by the plots in Fig. 3, F–H (and data not shown), ppGalNAc T5, T13, and T16 possess nearly equally enhanced preferences for both glycopeptide substrates GPIV and GPV compared with their control nonglycosylated peptides. ppGalNAc T5 is the only member of subfamily Id, whereas ppGalNAc T13 and T16 are members of subfamilies Ia and Ib, respectively, whose other members, ppGalNAc T1 and ppGalNAc T2 and T14, were shown above to prefer the GPIV glycopeptide substrate (see Fig. 3, A–C). This indicates that even within a subfamily, the N- and C-glycopeptide preferences can vary. The lack of directional glycopeptide specificity for ppGalNAc T5, T13, and T16 may be due to extensive lectin domain conformational dynamics, suggesting that the domain may be equally populated at the right and left of the catalytic domain and/or due to an elevated local concentration of glycopeptide substrate due to lectin binding and release.
Edman Sequencing of GPIV and GPV Series of (Glyco-)peptides—In an attempt to determine whether there is an optimal distance between the lectin-bound T* residue and the Xaa residues glycosylated by the catalytic domain, selected (glyco-)peptide products were Edman sequenced, and each residue's glycosylation pattern was determined. For simplicity, the 9 EA2 residues were maintained as in the original structure (see Fig. 1), whereas 3 additional Ala residues were added to complete the representation of the 12-Xaa acceptor region. Likewise, Ala residues were used to represent the Zaa residues, whereas a Ser residue was used to represent the location of the GalNAc-O-Thr. Note that for simplicity, both regions were modeled as static extended structures but in reality would be relatively flexible, comprising an ensemble of structures.

In an attempt to reveal the extent that the lectin domain may alter the pattern of \[^3H\]GalNAc incorporation, the distributions of the nonglycosylated control peptides, GPIV-C and GPV-C (glycosylated by ppGalNAc T13, Fig. 7E), were subtracted from the distributions obtained for the glycosylated peptides GPIV and GPV, respectively. These difference plots for all eight transferase isoforms, grouped by glycopeptide utilization, are shown in Fig. 8. It was anticipated that these plots would reveal the effects of lectin domain binding on the patterns of glycosylation and perhaps further reveal differences between transferase class and N- or C-terminal preference. Except for ppGalNAc T3 on GPIV (Figs. 7C and 8D), the difference plots were not appreciably different from the original distribution, although in a few cases, the distributions may have shifted slightly (i.e. for ppGalNAc T2, T5, and T16 on GPIV). From these results, we conclude that the binding of glycopeptide substrate for the most part (i.e. except for ppGalNAc T3 on GPIV) does not greatly alter the pattern of glycosylation compared with the nonglycopeptide substrate. Furthermore, no consistent differences were observed between isoform classes or between transferases within a class that had different N- or C-terminal preferences.

![Diagram](http://www.jbc.org/)

**FIGURE 6.** A, model building of (glyco)peptide substrates GPIV and GPV onto the EA2-bound ppGalNAc T2 crystal structure (42). Left, GPIV modeled with its Xaa residues placed in the catalytic domain in the same N- to C-terminal orientation as the 9 residues of the bound EA2 peptide. Right, GPIV modeled in the catalytic domain. For simplicity, the 9 EA2 residues were maintained as in the original structure (see Fig. 1), whereas 3 additional Ala residues were added to complete the representation of the 12-Xaa acceptor region. Likewise, Ala residues were used to represent the Zaa residues, whereas a Ser residue was used to represent the location of the GalNAc-O-Thr. Note that for simplicity, both regions were modeled as static extended structures but in reality would be relatively flexible, comprising an ensemble of structures.

**TABLE 2**

<table>
<thead>
<tr>
<th>ppGalNAc T linker domain and lectin domain motif alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>catalytic(C) linker(L) α-ricin β-ricin γ-ricin</td>
</tr>
<tr>
<td>-CCLLLLLLLLLLLLLLLLLLLLL_CLD--QXW-- -CLD--QXW-- -CLD--QXW--</td>
</tr>
<tr>
<td><strong>Class 1a</strong></td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td><strong>Class 1b</strong></td>
</tr>
<tr>
<td>T2</td>
</tr>
<tr>
<td>T4</td>
</tr>
<tr>
<td>T6</td>
</tr>
<tr>
<td><strong>Class 1c</strong></td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td>T6</td>
</tr>
<tr>
<td><strong>Class 1d</strong></td>
</tr>
<tr>
<td>T5</td>
</tr>
</tbody>
</table>

*Note that the last residues of the catalytic domain (C) and the first residues of the lectin domain (L), are shaded in red, whereas residues conserved only within a transferase class are shaded light grey. α-, β-, and γ-ricin lectin subdomain canonical motif residues, CLD and QXW, are shaded in grey, those residues that are similar to the canonical motif are shaded in yellow, and dissimilar, non-canonical residues are shaded in green (66–69).*
C-terminal glycopeptide preferences (i.e. ppGalNAc T1, T2, and T14 versus ppGalNAc T13, T16, and T5). These findings are consistent with a highly mobile lectin domain where the lectin-bound glycopeptide substrate would have a broad range of acceptor interacting distances with the catalytic domain. Alternatively, these results could be consistent with just a regiospecific increase in substrate concentration simply due to rapid binding and release from a conformationally fixed lectin domain.

DISCUSSION

Among eukaryote glycosyltransferases, the ppGalNAc T family is unique by possessing a separate carbohydrate binding lectin domain. Interestingly, such carbohydrate binding mod-
ules are common to bacterial glycoside hydrolases, serving to both increase local enzyme-substrate concentrations and to impart substrate specificity to nominally nonspecific catalytic domains by targeting the enzyme to specific substrates or sub-

FIGURE 8. Plots of [3H]GalNAc incorporation into the Xaa residues of random glycopeptides GIV and GPV (open bars) and after normalization by subtraction of control peptides GIV-C or GPV-C that were glycosylated by ppGalNAc T13 (filled bars). A–H, plots of GIV (left) and GPV (right) for ppGalNAc T1, T2, T14, T3, T6, T5, T13, and T16, respectively. Note that for A and B, substrates were glycosylated using 0.5 μM, 100-fold specific activity UDP-[3H]GalNAc, whereas in C–H, substrates were glycosylated using 50 μM standard specific activity UDP-[3H]GalNAc.
Lectin Domain Control of Mucin Type O-Glycosylation

ilar functions to modulate glycopeptide substrate recognition and specificity (48–53), but to date there has been no systematic study of the family against a common set of (glyco)peptide substrates. In this work, we have characterized eight nonglyco-
peptide-requiring ppGalNAc T isoforms against the series of random (glyco)peptide substrates listed in Table 1. Our findings have unambiguously revealed that prior GalNAc-O-Thr(Ser) substrate glycosylation can be recognized by these transferses in a specific N- or C-terminal direction that varies with ppGalNAc T isoform (see Figs. 3 and 4). Thus, for all eight transferses, at least one of the glycopeptide substrates, GPIV and/or GPV, shows significantly elevated activity over its non-glycosylated control. We attribute this N- or C-terminal specificity (i.e. elevated activity) to the binding of the glycosylated Thr residue of the substrate to the lectin domain in such a manner that the acceptor region of the substrate is oriented at the catalytic domain for optimal glycosylation. These findings strongly suggest that the tethered lectin domain on the ppGalNAc Ts may be mobile and that its location relative to the catalytic domain varies among isoforms. Such domain mobility is supported by the superimposition of the x-ray crystal structures of ppGalNAc T1, T2, and T10, as shown in Fig. 6B (42, 60, 61). For those transferses with similar N- and C-terminal glycopeptide enhancements (Fig. 4), it is possible that the lectin domain may be sufficiently mobile that it can enhance glycosylation from both an N- and C-terminal direction. Alternatively, the lectin domain may serve to equally increase the N- and C-terminal glycopeptide substrate concentrations by a simple bind and release mechanism, as shown for the glycoside hydro-
lases (62–65). Further evidence for a highly flexible lectin domain may be found in the broad distribution of glycosylation that is observed in the Xaa region of both the glycosylated and nonglycosylated substrates (see Figs. 7 and 8). The only signif-
icant alteration in distribution is observed for ppGalNAc T3 glycosylating its nonpreferred glycopeptide substrate GPIV (Fig. 7). This suggests that the lectin domain of this transferser is involved to some extent in directing the glycosylation of this substrate. Obviously, further studies are necessary to fully understand the dynamics of the lectin domain of these transferses.

In an attempt to correlate the different N- and C-terminal glycopeptide specificities to transferase peptide sequence we have compared the sequences of their linker domain and selected lectin domain motifs, as shown in Table 2. From the alignment of the linker domains, no obvious differences in length or sequence can readily account for the different behavior observed within a class (i.e. ppGalNAc T2 and T14, which prefer glycopeptide GPIV, are only 1 residue shorter than T16, which shows preferences for both glycopeptide substrates, whereas the linker domains of ppGalNAc T1 and T13 are identical except for two conservative substitutions, although they show different glycopeptide specificities). By contrast, the linker domains of ppGalNAc T3 and T6 are only 50% identical but have the same elevated GPV preferences.

Numerous studies of the ricin lectin α-, β-, and γ-subdomain repeats across multiple organisms have revealed several binding motifs, including the CLD and QXW sequences (66–69). Mutagenesis studies of the ppGalNAc T lectin domains have shown that the Asp residue of the α-subdomain CLD motif is typically required for lectin binding activities (i.e. ppGalNAc T1-D444A (54, 70, 71), ppGalNAc T2-D458H (50, 54), and ppGalNAc T3-D519H (54, 55) are lectin-inactivating mutants). A similar mutation in the ppGalNAc T1 β-rinic subdomain (D484A) only modestly decreases lectin-modulated activity (71, 72), whereas mutation of the γ-rinic subdomains of ppGalNAc T1 and T2 (D525A and D541A, respectively) shows no effects on their lectin domain activities (50, 71, 72). Mutation of the α-subdomain of ppGalNAc T4 (D459H) also eliminates its lectin activity (48). As shown in Table 2, the β-subdomains of ppGalNAc T2, T14, T16, T3, T6, and T5 and the γ-subdomains of T3 and T6 lack this Asp. Several transferses possess Glu instead of Asp (i.e. ppGalNAc T14 and T16 in the α-subdomain and T16 and T5 in the γ-subdomain); whether these domains can bind GalNAc is presently unknown. The β-subdomains of ppGalNAc T6 and T5 lack the QXW motif having EEW and LKW and would be expected to be inactive. Interestingly, the co-crystal structure of ppGalNAc T10 shows GalNAc-O-Ser bound to its β-subdomain (having canonical motifs CFD and QLW) (61). The ricin subdomain motifs of ppGalNAc T5 would be expected to have the weakest lectin binding activities because all three subdomains lack the critical Asp residue, although, in its γ-subdomain, the Asp is replaced by a Glu. Nevertheless, because ppGalNAc T5 displays clear glycopeptide specificities, one or more of its lectin subdomains (perhaps its γ-subdomain) must possess significant binding activity to provide glycopeptide enhancements. Generally, glycopeptide binding to any of the ppGalNAc Ts lectin domains has been difficult to directly detect, further suggesting its weak nature (50, 54, 55). As with the analysis of the linker domains, we con-
clude that we cannot discern any obvious correlation between the likely activities of specific lectin subdomain motif and a given transferase’s observed glycopeptide specificity.

From the above discussion of the lectin domain motifs listed in Table 2, it is likely, for all but ppGalNAc T5, that glycopeptide binding would occur predominantly at the α-subdomain. As shown in the superimposed crystal structures of the ppGalNAc Ts in Fig. 6B, the critical Asp residue of the CLD motif in the α-subdomain (highlighted in the structures) is found in a wide range of positions relative to the catalytic domain. One could easily assume that the conformational flex-
ibility of the lectin domain relative to the catalytic domain in solution would be greater than that observed in Fig. 6B. Thus, both the structural and experimental results are consistent with a highly mobile lectin domain, whose dynamics and relative position with respect to the catalytic domain varies widely with transferase isofrom. Whether additional domain-domain inter-
actions may modulate the relative positions of the lectin domain to the catalytic domain is presently unknown, but it seems highly plausible that such interactions could assist in the positioning of the lectin domain. Future studies will address these questions.

Prior Evidence for the Lectin Domain Targeting N- or C-terminal Glycopeptide Glycosylation—The glycosylation of the Muc 7 peptide (PPTSSAT2PAPPSS13S14APPET19T20AAK) by the subclass Ia transferses ppGalNAc T1 and T13 has been examined by Zhang et al. (73). Interestingly, they observed that
Lectin Domain Control of Mucin Type O-Glycosylation

The control of site-specific mucin type O-glycosylation and the need for such a large family of initiating ppGalNAc T is not well understood. In this work, we have identified a previously unappreciated level of control, whereby prior O-glycosylation is used to target and enhance the glycosylation of specific N- or C-terminal sites in an isoform-specific manner. Our studies and those of others strongly suggest that this N- or C-terminal selectivity is due to weak glycopeptide binding to the lectin domain, whose orientation relative to the catalytic domain is highly mobile and isoform-dependent. This glycopeptide selectivity can provide an additional level of control or fidelity for the glycosylation of biologically significant sites and suggests that O-glycosylation in some instances may be exquisitely controlled. Furthermore, our observations that homologous ppGalNAc T isoforms within a given subfamily (that presumably have similar peptide substrate specificities; see Ref. 56) may possess different N- or C-terminal glycopeptide preferences may explain the large number of ppGalNAc T family members thereby maintaining peptide specificity while altering glycopeptide specificity. These studies clearly demonstrate that the biological control mucin type O-glycosylation is highly complex and that further structural, biochemical, and biological studies are necessary to fully understand this important modification.

Acknowledgments—We thank Oliver Jamison for initial assistance with designing the lectin peptide substrates and Dr. Vivian Yee for assistance with the molecular modeling.

REFERENCES


Note that Raman et al. (51) have also proposed in their kinetic analysis that the ppGalNAc T3 lectin domain may also direct glycosylation to sites C-terminal of a GalNAc-O-Thr residue by assisting in product release rather than assisting in substrate binding.


ppGalNAc T1 and T2 are ubiquitously expressed in nearly all mammalian tissues and cell lines, whereas the remaining ppGalNAc T isoforms studied in this work are selectively expressed (see Refs. 1, 16, and 75). For example, ppGalNAc T3 is the dominant transferase expressed in the testis and is also highly expressed in the kidney (21, 75, 76), and ppGalNAc T16 is more highly expressed than ppGalNAc T1 in the heart and brain (41), whereas ppGalNAc T13 is specifically expressed in neurons (73). How the multiple transferases work in concert in glycosylating their target proteins is largely unknown, as is their regulation at both the protein and transcriptional levels, although, judging from observations in our laboratory and other laboratories, large differences in protein stabilities exist.

although Thr7 was initially glycosylated by both transferases, only ppGalNAc T13 was capable of further glycosylating the boldface C-terminal Ser and Thr residues of the peptide. These findings are entirely consistent with the transference glycopeptide preferences summarized in Fig. 4, where ppGalNAc T1 prefers to glycosylate sites N-terminal of a prior site of glycosylation, whereas ppGalNAc T13 is capable of glycosylating sites both N- and C-terminal of a prior site of glycosylation. Raman et al. (51) have extensively characterized the glycosylation of a series of MUC5AC glycopeptides by ppGalNAc T2 and its catalytic domain. Based on comparative binding studies, it was concluded that ppGalNAc T2 bound glycopeptide when the acceptor site was 10 residues N-terminal of an existing site of GalNAc glycosylation but not a glycopeptide with the reverse orientation and that this was due to the presence of the lectin domain, an observation consistent with our studies of ppGalNAc T2 (see Fig. 4). Likewise, studies on the glycan binding requirements of the lectin domain of ppGalNAc T3 suggest that this transferase utilizes its lectin domain to glycosylate residues C-terminal of the site of prior glycosylation (55), again in keeping with our findings for ppGalNAc T3. ppGalNAc T3 was previously shown to be the only isoform capable of glycosylating the proprotein processing region of FGF23 (IHFNT171PIPR RHT178R ↓ SAEDD) (21, 25, 74), and in unpublished work, Kato et al. have found that ppGalNAc T3 glycosylates the Thr178 site in a lectin-dependent manner by first glycosylating Thr171 in a lectin-independent manner, both in vitro enzyme assays and ex vivo in CHO cells stably transfected with ppGalNAc T3 variants. The Thr171 site is also glycosylated by several other ppGalNAc-T isoforms (21). Our previous work (43) utilizing transferase specific preferences, obtained from random peptide studies, also suggests that Thr171 would be a moderate to good substrate for ppGalNAc T1, T2, and T3, whereas Thr178 would not be a substrate for ppGalNAc T1 or T2 and would be only a modest substrate for ppGalNAc T3 (see Table 2 of Ref 43). This suggests that the prior glycosylation of Thr171 may serve to target ppGalNAc T3 glycosylation of Thr178, 7 residues C-terminal of Thr171. The glycosylation distribution plots for ppGalNAc T3 for glycopeptide GPV (Figs. 7C and 8D) are entirely consistent with this, showing a peak of glycosylation 8–9 residues C-terminal of the original site of glycosylation. By contrast, the prior glycosylation of Thr171 would fail to enhance the glycosylation of Thr178 by ppGalNAc T1 and T2 because these transferases possess the reverse glycopeptide specificities. Thus, glycosylation of Thr171 enhances the glycosylation of Thr178 by ppGalNAc T3 while effectively reducing the activities of those transferases with the reverse glycopeptide preferences. Thus, prior glycosylation of Thr171 serves as a targeting switch or enhancer for the glycosylation of Thr178, whose glycosylation is required to inhibit the proprotein convertase cleavage of FGF23 in vivo and in vitro (20, 21, 74).


Lectin Domain Control of Mucin Type O-Glycosylation

31.

32.

33.

35.

36.

37.
38.

39.
40.

41.

42.

43.

44.

45.

46.

47.

JULY 5, 2013 • VOLUME 288 • NUMBER 27

48.

49.

50.

51.

52.

53.

54.

55.

56.

57.

58.

59.

60.

61.

62.

310

neighboring residue glycosylation on the substrate specificity of the uridine 5⬘-diphosphate-␣-N-acetylgalactosamine:polypeptide N-acetylgalactosaminyl transferases T1 and T2. Kinetic modeling of the porcine
and canine submaxillary gland mucin tandem repeats. Biochemistry 43,
9888 –9900
Hassan, H., Reis, C. A., Bennett, E. P., Mirgorodskaya, E., Roepstorff, P.,
Hollingsworth, M. A., Burchell, J., Taylor-Papadimitriou, J., and Clausen,
glycosylation-dependent activities of polypeptide N-acetylgalactosaminyltransferases rGalNAc-T2 and -T4 on mucin glycopeptides, Glycobiology 11, 731–740
Wandall, H. H., Irazoqui, F., Tarp, M. A., Bennett, E. P., Mandel, U.,
Takeuchi, H., Kato, K., Irimura, T., Suryanarayanan, G., Hollingsworth,
binding to GalNAc-glycopeptide substrates is required for high density
GalNAc-O-glycosylation. Glycobiology 17, 374 –387
Raman, J., Fritz, T. A., Gerken, T. A., Jamison, O., Live, D., Liu, M., and
polypeptide ␣-N-acetylgalactosaminyltransferase function in concert to
direct glycosylation site selection. J. Biol. Chem. 283, 22942–22951
Pedersen, J. W., Bennett, E. P., Schjoldager, K. T., Meldal, M., Holmér,
Lectin domains of polypeptide GalNAc-transferases exhibit glycopeptide
binding specificity. J. Biol. Chem. 286, 32684 –32696
Yoshimura, Y., Nudelman, A. S., Levery, S. B., Wandall, H. H., Bennett,
E. P., Hindsgaul, O., Clausen, H., and Nishimura, S. (2012) Elucidation of
the sugar recognition ability of the lectin domain of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase 3 by using unnatural glycopeptide substrates. Glycobiology 22, 429 – 438
peptide acceptor preferences between Drosophila and mammalian polypeptide-GalNAc transferase orthologue pairs. Glycobiology 18, 861– 870
Barb, A. W., Meng, L., Gao, Z., Johnson, R. W., Moremen, K. W., and
Prestegard, J. H. (2012) NMR characterization of immunoglobulin G Fc
glycan motion on enzymatic sialylation. Biochemistry 51, 4618 – 4626
Vester-Christensen, M. B., Bennett, E. P., Clausen, H., and Mandel, U.
(2013) in Glycosyltransferases: Methods and Protocols (Brockhausen, I.,
ed), Methods in Molecular Biology, Humana Press, in press
O-glycosylation is modulated by neighboring residue glycosylation status.
Kinetic modeling of the site-specific glycosylation of the apo-porcine
sumbaxillary mucin tandem repeat by UDP-GalNAc:polypeptide Nacetylgalactosaminyltransferases T1 and T2, J. Biol. Chem. 277,
49850 – 49862
Kubota, T., Shiba, T., Sugioka, S., Furukawa, S., Sawaki, H., Kato, R.,
Wakinaka, T., Kiyohara, M., Kurihara, S., Hirata, A., Chaiwangsri, T.,
Ohnuma, T., Fukamizo, T., Katayama, T., Ashida, H., and Yamamoto, K.
(2013) Bifidobacterial ␣-galactosidase with unique carbohydrate-binding
module specifically acts on blood group B antigen. Glycobiology 23,

JOURNAL OF BIOLOGICAL CHEMISTRY

19913

Downloaded from http://www.jbc.org/ at Cleveland Health Sciences Library on December 11, 2014

34.

R. S., Bhardwaj, N., Shao, Y., Darvishian, F., Zavadil, J., Erlebacher, A.,
Mahal, L. K., Osman, I., and Hernando, E. (2011) miR-30b/30d Regulation
of GalNAc Transferases Enhances Invasion and Immunosuppression during Metastasis, Cancer Cell 20, 104 –118
Polypeptide N-acetylgalactosaminyltransferase 6 disrupts mammary acinar morphogenesis through O-glycosylation of fibronectin. Neoplasia 13,
320 –326
Taniuchi, K., Cerny, R. L., Tanouchi, A., Kohno, K., Kotani, N., Honke, K.,
Berois, N., Mazal, D., Ubillos, L., Trajtenberg, F., Nicolas, A., Sastre-Garau,
317–328
Berois, N., Gattolliat, C. H., Barrios, E., Capandeguy, L., Douc-Rasy, S.,
Valteau-Couanet, D., Bénard, J., and Osinaga, E. (2013) GALNT9 gene
expression is a prognostic marker in neuroblastoma patients. Clin. Chem.
59, 225–233
fatal embryonic hemorrhage in mice lacking core 1-derived O-glycans.
the fly. Developmental and mechanistic insights from Drosophila. Glycobiology 19, 102–111
J. Biol. Chem. 287, 5243–5252
during development. J. Biol. Chem. 288, 6921– 6929
modulates integrin and FGF signalling by influencing the secretion of
UDP-N-acetyl-␣-D-galactosamine:polypeptide N-acetylgalactosaminyltransferases. Completion of the family tree. Glycobiology 22, 768 –777
between the catalytic and lectin domains of human UDP-GalNAc:
281, 8613– 8619
Gerken, T. A., Jamison, O., Perrine, C. L., Collette, J. C., Moinova, H., Ravi,
paradigms for the initiation of mucin type protein O-glycosylation by the
polypeptide GalNAc transferase (ppGalNAc T) family of glycosyltransferases. J. Biol. Chem. 286, 14493–14507
of common and unique peptide substrate preferences for the UDPGalNAc. Polypeptide ␣-N-acetylgalactosaminyltransferases T1 and T2
(ppGalNAc T1 and T2) derived from oriented random peptide substrates.
J. Biol. Chem. 281, 32403–32416
Tabak, L. A., and Gerken, T. A. (2009) The glycopeptide preferring polypeptide-GalNAc transferase-10 (ppGalNAc T10), involved in mucin typeO-glycosylation, has a unique GalNAc-O-Ser/Thr binding site in its catalytic domain not found in ppGalNAc T1 or T2. J. Biol. Chem. 284,
20387–20397
site-specific oligosaccharide distribution of the O-glycans attached to
the porcine submaxillary mucin tandem repeat. Further evidence for the
modulation of O-glycan side chain structures by peptide sequence. J. Biol.
Chem. 277, 7736 –7751
Gerken, T. A., Tep, C., and Rarick, J. (2004) Role of peptide sequence and




Glycan Synthesis

Mucin-type O-glycosylation is controlled by short- and long-range glycopeptide substrate recognition that varies among members of the polypeptide GalNAc transferase family

Leslie Revoredo², Shengjun Wang⁴, Eric Paul Bennett⁴, Henrik Clausen⁴, Kelley W Moremen⁵, Donald L Jarvis⁶, Kelly G Ten Hagen⁷, Lawrence A Tabak⁸, and Thomas A Gerken¹,²,³

²Department of Chemistry, ³Department of Pediatrics and Biochemistry, Case Western Reserve University, Cleveland, OH 44106, USA, ⁴Copenhagen Center for Glycomics (CCG), Departments of Cellular and Molecular Medicine and Dentistry, Faculty of Health Sciences, University of Copenhagen, Copenhagen N DK-2200, Denmark, ⁵Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA, ⁶Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA, ⁷Developmental Glycobiology Section, and ⁸Section on Biological Chemistry, Department of Health and Human Services, NIDCR, National Institutes of Health, Bethesda, MD 20892, USA

¹To whom correspondence should be addressed: Tel: +1-216-368-4556; Fax: +1-216-368-4223; e-mail: txg2@cwru.edu

Received 9 September 2015; Revised 3 November 2015; Accepted 16 November 2015

Abstract

A large family of UDP-GalNAc:polypeptide GalNAc transferases (ppGalNAc-Ts) initiates and defines sites of mucin-type Ser/Thr-O-GalNAc glycosylation. Family members have been classified into peptide- and glycopeptide-preferring subfamilies, although both families possess variable activities against glycopeptide substrates. All but one isoform contains a C-terminal carbohydrate-binding lectin domain whose roles in modulating glycopeptide specificity is just being understood. We have previously shown for several peptide-preferring isoforms that the presence of a remote Thr-O-GalNAc, 6–17 residues from a Ser/Thr acceptor site, may enhance overall catalytic activity in an N- or C-terminal direction. This enhancement varies with isoform and is attributed to Thr-O-GalNAc interactions at the lectin domain. We now report on the glycopeptide substrate utilization of a series of glycopeptide (human-ppGalNAc-T4, T7, T10, T12 and fly PGANT7) and peptide-preferring transferases (T2, T3 and T5) by exploiting a series of random glycopeptide substrates designed to probe the functions of their catalytic and lectin domains. Glycosylation was observed at the −3, −1 and +1 residues relative to a neighboring Thr-O-GalNAc, depending on isoform, which we attribute to specific Thr-O-GalNAc binding at the catalytic domain. Additionally, these glycopeptide-preferring isoforms show remote lectin domain-assisted Thr-O-GalNAc enhancements that vary from modest to none. We conclude that the glycopeptide specificity of the glycopeptide-preferring isoforms predominantly resides in their catalytic domain but may be further modulated by remote lectin domain interactions. These studies further demonstrate that both domains of the ppGalNAc-Ts have specialized and unique functions that work in concert to control and order mucin-type O-glycosylation.

Key words: glycoprotein biosynthesis, glycosyltransferase, lectin, mucin
Introduction

Mucin-type protein O-glycosylation is a common and essential post-translational modification in higher organisms that begins with the transfer of the sugar N-acetylgalactosamine (GalNAc) onto the hydroxyl groups of Ser or Thr residues of protein substrates. This type of protein glycosylation is initiated in the Golgi complex by a large family of enzymes known as the UDP-GalNAc:polypeptide N-acetylgalactosaminyl-transferases (ppGalNAc-Ts). These enzymes therefore define the sites of O-glycosylation.

Mucin-type O-glycosylation is critical for development, where in the fly, several ppGalNAc-T (PGANT) isoforms are essential for viability (Ten Hagen and Tran 2002; Tran and Ten Hagen 2013; Zhang et al. 2014), cell adhesion, and proper secretory apparatus function (Zhang et al. 2014). In addition, O-glycosylation of cell surface receptors may be regulated by specific ppGalNAc-T isoforms that appear to modulate receptor signaling (Kato et al. 2006; Semenov et al. 2009; Schjoldager et al. 2010). In the mouse, ppGalNAc-T1 modulates salivary gland organogenesis (Tian et al. 2012) and is required for proper heart and valve development (Tian et al. 2015). Inactivating mutations in ppGalNAc-T3 are responsible for the human disease familial luminal calcinosis, which is characterized by defects in phosphate regulation. This is the first human disease linked to a specific ppGalNAc-T and is mechanistically caused by the loss of the specific glycosylation of the phosphaturic factor, FGFR3, which leads to its proteolytic cleavage (Kato et al. 2006). Such modulation of protease cleavage sites may be a common function for O-glycosylation in biology (Semenov et al. 2009; Schjoldager et al. 2010). Finally, genome wide association studies have identified a number of distinct ppGalNAc-T isoforms that are associated with certain human diseases and conditions (Simmons et al. 1999; Topaz et al. 2004; Kathiresan et al. 2008; Willer et al. 2008; Holleboom et al. 2011; Nakayama et al. 2012; Miyazaki et al. 2013; Takasaki et al. 2014), including many cancers (Kohsaki et al. 2000; Berois et al. 2006; Brockhausen 2006; Wood et al. 2007; Libisch et al. 2014). Altogether, these studies indicate essential (but not well understood) roles for this type of protein modification across diverse tissues and species.

Members of the ppGalNAc-T family (20 different isoforms in man) have been classified into two major families (I and II) based on their peptide/glycopeptide preferences, respectively. Although this distinction has been historically used, some overlap in substrate preferences exists among transferase isoforms. The families have been further divided into subfamilies (Iα–g and IIA–b, Figure 1A) based on their amino acid sequence similarities which tend to correlate with their overall peptide/glycopeptide substrate preferences (Bennett et al. 2012). Family I members readily glycosylate peptide substrates and have been termed “peptide-prefering” isoforms; however, several members of this family (Iα–d) have been shown to be influenced by remote, 6–17 residue, prior Ser/Thr-O-GalNAc glycosylation via utilization of their lectin domains (see Figures 1B and below) (Gerken et al. 2013). Family IIA–b members on the other hand commonly display nearlysole activities against Ser/Thr-O-GalNAc containing glycopeptides and hence have been termed “glycopeptide-prefering” isoforms, although some isoforms (particularly subfamily IIA) can also readily glycosylate unglycosylated peptide substrates (Hassan et al. 2000; Gerken et al. 2011). Because of the increasingly complex peptide and glycopeptide specificities now being observed for these transferases, and taking into account the results of the present work, we will suggest below, a more precise reclassification of the ppGalNAc-T family. This takes in account our observations, that depending on the isoform, glycopeptides can interact with either the lectin or catalytic domains (or both) thereby resulting in significantly different glycopeptide specificities (see Figure 9 and Discussion).

Structurally, ppGalNAc-T family members (except h-ppGalNAc-T20) (Raman et al. 2012), consist of an N-terminal catalytic domain linked to a C-terminal ricin like lectin domain containing three potential carbohydrate Ser/Thr-O-GalNAc-binding sites (Figure 1B; Fritz et al. 2006; Kubota et al. 2006; Yoshimura et al. 2012). Both domains are linked via a short flexible segment that varies in length among isoforms creating flexibility between domains (Fritz et al. 2004,2006; Kubota et al. 2006; Lira-Navarrete et al. 2013). How these domains modulate the peptide and glycopeptide activities of the different subfamilies is just being understood (Raman et al. 2008; Gerken et al. 2013; Lira-Navarrete et al. 2015).

We and others have shown that the ppGalNAc-T’s display specific peptide substrate preferences that vary among isoforms (Perrine et al. 2009; Gerken et al. 2011). This modulation of substrate specificity is based on peptide sequence, charge and importantly neighboring glycosylation. The placement of a Ser/Thr-O-GalNAc residue near a glycosylation site results in multiple effects. These effects vary from a relative inhibition of glycosylation to a requirement for a neighboring GalNAc, and may also include an alteration or shift in glycosylation site. Thus the family I “peptide-prefering” ppGalNAc-T2 utilizes its lectin domain for modulating directionality and site preference against Ser/Thr-O-GalNAc glycopeptides, while the family II “glycopeptide-prefering” ppGalNAc-T10 utilizes its catalytic domain for its near absolute requirement for a neighboring Ser/Thr-O-GalNAc (Kubota et al. 2006; Raman et al. 2008; Perrine et al. 2009).

Recently, we reported that the presence of a remote N- or C-terminal placed Thr-O-GalNAc can be an important determinant of overall catalytic activity and specificity in eight of the characterized transferases in “peptide-prefering” subfamilies Iα–Id which differed among isoforms (Gerken et al. 2013). This N- or C-terminal selectivity is thought to be due to glycopeptide Ser/Thr-O-GalNAc binding to the lectin domain whose orientation may be highly dynamic and isoform dependent. We have previously proposed that this N- or C-directional specificity may serve to control the specificity of mucin-type O-glycosylation suggesting that O-glycosylation may actually be highly ordered (Gerken et al. 2013).

Here, we address the roles of the catalytic and lectin domains of the “glycopeptide-prefering” transferases in subfamilies IIA (ppGalNAc-T4 and T12) and IIB (ppGalNAc-T7 and T10). Although both subfamilies will readily glycosylate Ser/Thr-O-GalNAc glycopeptides, they differ in their ability to readily glycosylate nonglycosylated peptide substrates (Bennett et al. 1998,1999; Raman et al. 2008; Perrine et al. 2009; Gerken et al. 2011; Raman et al. 2008; Perrine et al. 2009). Using our series of lectin and catalytic domain probing glycopeptide substrates, we are able to define the role of the lectin and catalytic domains in the recognition of extant GalNAc on glycopeptide substrates. Specifically, we have determined that members of these subfamilies prefer to add GalNAc in close proximity to a previously glycosylated residue (i.e. neighboring glycosylation), suggesting a mechanism whereby the catalytic domain preferentially recognizes an existing GalNAc in glycopeptide substrates. Comparative studies on the fly orthologue of ppGalNAc-T7, PGANT7, have also been performed further demonstrating conservation of these mechanistic specificities across species.

1 We will be using the transferase numbering system of Bennett et al. (2012) for consistency.
This work in conjunction with our prior studies will be highly useful for refining predictive approaches for identifying sites of mucin-type O-glycosylation that thus far have been limited by a lack of isoform-specific peptide and glycopeptide glycosylation data (Elhammer et al. 1993; Hansen et al. 1998; Chen et al. 2008; Steentoft et al. 2013).

Results

Design and use of (Glyco)peptide substrates

In previous studies (Perrine et al. 2009; Gerken et al. 2013), we designed a series of random (glyco)peptide substrates to interrogate the roles of the lectin and catalytic domains in modulating glycopeptide specificity for several family I and one family II ppGalNAc-Ts (Table I). As described below, these were based on characterizing the effects of remote and neighboring prior Ser/Thr-O-GalNAc glycosylation, respectively, on transerase activity.

Lectin domain probing random glycopeptide substrates GP(T*22)R and GP(T*10)L (formerly named GPIV and GPV (Gerken et al. 2013))

These substrates address the modulating role of a remote Thr-O-GalNAc (abbreviated as T*) on transerase activity, based on the assumption that the T* would bind at the lectin domain. These substrates were designed to be sufficiently long (29 residues) to span the catalytic and lectin domains based on the crystal structures of ppGalNAc-T1 and T2 (Fritz et al. 2004, 2006; Figure 1B). As shown in Table I, each acceptor substrate contained a T* placed near the C- or N-terminal (GP(T*22)R and GP(T*10)L, where R is for right and L for left placed T*, respectively) which in turn were flanked by five randomized Z residues lacking an acceptor Ser or Thr. An additional 12 randomized X residues, also including the acceptor Thr, were extended from the Z’s in an N- or C-terminal direction, respectively. These X residues thus serve to probe glycosylation 6–17 residues from an existing T* in an N- or C-direction, respectively. To confirm that the T* in these glycopeptides are indeed interacting with the lectin domain, non-interacting control peptides were also used where the T* was replaced by an Ala residue (GP(A22)R and GP(A10)L, respectively). With these (glyco)peptides, we can determine the extent that remote Thr-O-GalNAc interactions, at the lectin domain can enhance catalytic domain glycosylation of the X residues and whether there is a preferred N- or C-directionality. (For clarity, results for the GP-R series of substrates will be presented in shades of red and those for GP-L will be in shades of blue.)

Confirmation that these lectin domain substrates indeed are reporting on lectin domain binding/interactions of a GalNAc residue was shown in our prior work (Gerken et al. 2013). Furthermore, these substrates correctly report the observed remote T* lectin binding directionality for ppGalNAc-T2 and T3. Thus, ppGalNAc-T2, which prefers to bind a C-terminal T* based on its preference for GP (T*22)R (Gerken et al. 2013), is shown bound to a C-terminal glycosylated MUC5AC-13 glycopeptide in a recently reported X-ray crystal structure (Lira-Navarrete et al. 2015). In addition, the kinetic studies
of Raman et al. (2008) of ppGalNAc-T2 show a similar lectin domain binding preference for the MUC5AC-13 glycopeptide. ppGalNAc-T3 prefers an N-terminal T* based on its preference for GP(T*10)L and is known to sequentially glycosylate an FGF23 peptide at two sites that are seven residues apart in an N-to-C order (Gerken et al. 2013). The glycosylation of the latter C-terminal site was shown to require a ppGalNAc-T3 possessing an active lectin domain, in agreement with its preference for GP(T*10)L (Gerken et al. 2013).

Catalytic domain probing random glycopeptide substrate GP(T*10)C

The design of GP(T*10)C is based on a substrate analog GPII that has been previously employed to characterize T* binding at the catalytic domain of ppGalNAc-T10 (Perrine et al. 2009). GP(T*10)C (where C is for a center placed T*) has the general form GAGAXXXXXT*10XXXXXXXXAG where the X residues flank the central Thr-O-GalNAc contain unglycosylated Ser residues that can serve as an acceptor. Using this substrate, we can address the role of a single T* on the glycosylation of its directly neighboring residues, X, using UDP-[3H]-GalNAc and Edman amino acid sequencing to determine the Ser-[3H]-GalNAc content on the X positions. Since the acceptor (X) regions of gpGalNAc-T3 are directly flanking T*, the glycosylation of the closely neighboring sites by ppGalNAc-T2 lectin domain reported by Lira-Navarrete et al. (2015). In their plot of predicted enzymatic activity vs. residue separation, no sites closer than 5 residues from the T* were predicted to be glycosylated by T2, while a broad peak of glycosylation was predicted between 7 and 14 residues from the T* (see Figure 5C of Lira-Navarrete et al. 2015).

2 Note however that the kinetic studies of Raman et al. (2008) also report lectin domain involvement in product release for an N-terminal T* glycosylated MUC5AC-3 glycopeptide substrate. This explains the observation that the lectin domain of T2 can also direct remote glycosylation C-terminal of a T* with certain substrates. The molecular mechanisms of how the lectin domain-mediated product release occurs and contributes to the T2’s C-terminal long-range glycopeptide specificity is presently unknown and awaits further study.

3 Note, however, that the random X acceptor sites on GP(T*10)C contain Ser residues which are typically much less reactive than the Thr residues in the acceptor sites of the lectin probing glycopeptides GP(T*22)R and GP(T*10)L. In addition, the X positions of GP(T*10)C contain additional randomized residue types compared with GP(T*22)R and GP(T*10)L which could further reduce GP(T*10)C's relative activity. Therefore, when comparing the activities of the lectin domain glycopeptides to GP(T*10)C, the actual activity of GP(T*10)C could be expected to be much higher than observed.
L is the most favorable substrate, showing the highest GalNAc incorporation (see Figure 3B right panel, blue series). For the remaining "glycopeptide-preferring" family II transferases, variable incorporation into the lectin probing substrates GP(T*22)R and GP(T*10)L, is observed (Figure 3C–E), with the exceptions of the fly PGANT7 which shows substantial and nearly equal incorporation into GP(T*22)R and GP(T*10)L (Figure 3F).

Also shown in Figures 2 and 3 are the plots for the control nonglycosylated peptide substrates GP(A22)R and GP(A10)L (pink and turquoise series, respectively) which show reduced activities and little transfer to peptide compared with their glycosylated analogues. This suggests that for the "glycopeptide-preferring" transferases, the lectin domain plays variable roles in remote glycopeptide site selection, ranging from very high for ppGalNAc-T4 and PGANT7 and essentially none for ppGalNAc-T10.

To further quantify and compare each transferase substrate preference, we determined the percent (or efficiency) of GalNAc transferred to substrate (relative to total UDP-GalNAc utilization) from the gel filtration plots shown in Figure 3 for GP(T*22)R, GP(T*10)L, and GP(T*10)C. These data are plotted in Figure 4 which clearly shows that the efficiency of transfer varies depending on the glycopeptide substrate and ppGalNAc-T isoform family and subfamily. For the "peptide-preferring" subfamilies Ia–d, transfer is highest with glycopeptide substrates GP(T*22)R and GP(T*10)L (60–90%), while a much lower percentage of transfer (3–20%) is observed for GP(T*10)L. By comparison, the "glycopeptide-preferring" transferases...
Fig. 3. Sephadex G10 gel filtration chromatograms demonstrating elevated GalNAc hydrolysis (transfer to water) for the "glycopeptide-preferring" ppGalNAc-Ts. Chromatograms are shown for the “peptide-preferring” ppGalNAc-T5 (A) and “glycopeptide-preferring” ppGalNAc-T4 (B), T12 (C), T10 (D), T7 (E) and the fly PGANT7 (F). Left panels are chromatograms for random (glyco)peptide substrates GP(T*22)R (red diamond), GP(A22)R (pink upward triangle) and GP(T*10)C (green circle) while the right panels show chromatograms for GP(T*10)L (blue square), GP(A10)L (turquoise downward triangle) and GP(T*10)C (green circle). These plots demonstrate the incorporation of [3H]-GalNAc into (glyco)peptide substrate (fractions 24–31) and hydrolysis (i.e. transfer to water, fractions 36–43) for the overnight incubations (1200 min) shown in Figure 2. These plots are representative plots from a single experiment, although two or more experiments displaying nearly identical results were performed.
in subfamilies IIa–b, generally show less transfer to GP(T*22)R and GP(T*10)L and greater transfer to GP(T*10)C compared with the family I transferases. Nonetheless, the subfamily IIa transferases, ppGalNAc-T4 and T12 show the highest percent transfer to GP(T*10)L (40–75%) while having substantial transfer to GP(T*10)C (30–35%). The more strict “glycopeptide-prefering” transferases in subfamily IIb, ppGalNAc-T7 and T10 have high transfers to GP(T*10)C (20–40%) and much lower (or almost no transfer) to GP(T*22)R and GP(T*10)L (3–30%). Interestingly, the fly PGANT7, which is in the same subfamily as ppGalNAc-T7 and T10, possesses very high transfers to all three substrates GP(T*10)C, GP(T*22)R and GP(T*10)L (50–70%).

Glycosylation site determination on GP(T*10)C

Edman sequencing of the GP(T*10)C products of ppGalNAc-T2, T3, T4, T5, T7, T10, T12 and PGANT7 were performed to locate transferase specific glycosylation sites relative to the central Thr-O-GalNAc (T*). PTH derivatives at each cycle were counted for [3H]-GalNAc incorporation at the +1 and +5 positions. This suggests that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (right panel of Figure 5D). These sites of glycosylation are consistent with ppGalNAc-T4’s high activity against glycopeptides A2 and A4 of the glycopeptide library of Pratt and coworkers which possess a +1 and +5 acceptor site from a T* (Pratt et al. 2004).

ppGalNAc-T12, which belongs to the same subfamily as ppGalNAc-T4 (72% identity), shows maximum incorporation at the −3 position relative to T* (Figure 5E), indicating that T12 prefers to glycosylate a Ser (or Thr) residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a T* residue. It is likely that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (data not shown). In contrast to ppGalNAc-T4, ppGalNAc-T12 does not glycosylate the +5 position of GP(T*10)L which is consistent with its weaker activity against lectin probing GP(T*10)L.

Our previous studies with ppGalNAc-T10 on GPII have shown that it will glycosylate a Ser (or Thr) residue directly N-terminal (−1) of a Ser or Thr-O-GalNAc residue (Perrine et al. 2009). This is sterically hindered by the presence of the T* (Gerken et al. 2002; Gerken 2004; Perrine et al. 2009). Interestingly, for ppGalNAc-T2 we observe very low [3H]-GalNAc incorporation spread across the −4 to −1 positions with a comparatively small peak at the +4 position. The [3H]-GalNAc incorporation at these positions is not thought to be entirely due to sequencing lag, since similar positions were shown to be weakly glycosylated in our previous studies with ppGalNAc-T2 on a similar glycopeptide, GPII (Perrine et al. 2009), thus T2 may have a very limited ability to glycosylate directly near a prior site of glycosylation.

ppGalNAc-T4 displays a completely different behavior with the GP(T*10)C substrate (Figure 5D), showing nearly equal incorporation at the +1 and +5 positions. This suggests that ppGalNAc-T4 prefers to glycosylate a Ser (or Thr) residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a T* residue. It is likely that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (right panel of Figure 5D). These sites of glycosylation are consistent with ppGalNAc-T4’s high activity against glycopeptides A2 and A4 of the glycopeptide library of Pratt and coworkers which possess a +1 and +5 acceptor site from a T* (Pratt et al. 2004).

ppGalNAc-T12, which belongs to the same subfamily as ppGalNAc-T4 (72% identity), shows maximum incorporation at the −3 position relative to T* (Figure 5E), indicating that T12 prefers to glycosylate a Ser (or Thr) residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a T* residue. It is likely that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (data not shown). In contrast to ppGalNAc-T4, ppGalNAc-T12 does not glycosylate the +5 position of GP(T*10)L which is consistent with its weaker activity against lectin probing GP(T*10)L.

Our previous studies with ppGalNAc-T10 on GPII have shown that it will glycosylate a Ser (or Thr) residue directly N-terminal (−1) of a Ser or Thr-O-GalNAc residue (Perrine et al. 2009). This is sterically hindered by the presence of the T* (Gerken et al. 2002; Gerken 2004; Perrine et al. 2009). Interestingly, for ppGalNAc-T2 we observe very low [3H]-GalNAc incorporation spread across the −4 to −1 positions with a comparatively small peak at the +4 position. The [3H]-GalNAc incorporation at these positions is not thought to be entirely due to sequencing lag, since similar positions were shown to be weakly glycosylated in our previous studies with ppGalNAc-T2 on a similar glycopeptide, GPII (Perrine et al. 2009), thus T2 may have a very limited ability to glycosylate directly near a prior site of glycosylation.

ppGalNAc-T4 displays a completely different behavior with the GP(T*10)C substrate (Figure 5D), showing nearly equal incorporation at the +1 and +5 positions. This suggests that ppGalNAc-T4 prefers to glycosylate a Ser (or Thr) residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a T* residue. It is likely that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (right panel of Figure 5D). These sites of glycosylation are consistent with ppGalNAc-T4’s high activity against glycopeptides A2 and A4 of the glycopeptide library of Pratt and coworkers which possess a +1 and +5 acceptor site from a T* (Pratt et al. 2004).

ppGalNAc-T12, which belongs to the same subfamily as ppGalNAc-T4 (72% identity), shows maximum incorporation at the −3 position relative to T* (Figure 5E), indicating that T12 prefers to glycosylate a Ser (or Thr) residue directly C-terminal to (+1) and five (+5) C-terminal positions away from a T* residue. It is likely that ppGalNAc-T4’s +5 preference may be due to lectin domain interactions, consistent with its high preference for GP(T*10)L (data not shown). In contrast to ppGalNAc-T4, ppGalNAc-T12 does not glycosylate the +5 position of GP(T*10)L which is consistent with its weaker activity against lectin probing GP(T*10)L.

Our previous studies with ppGalNAc-T10 on GPII have shown that it will glycosylate a Ser (or Thr) residue directly N-terminal (−1) of a Ser or Thr-O-GalNAc residue (Perrine et al. 2009). This is sterically hindered by the presence of the T* (Gerken et al. 2002; Gerken 2004; Perrine et al. 2009). Interestingly, for ppGalNAc-T2 we observe very low [3H]-GalNAc incorporation spread across the −4 to −1 positions with a comparatively small peak at the +4 position. The [3H]-GalNAc incorporation at these positions is not thought to be entirely due to sequencing lag, since similar positions were shown to be weakly glycosylated in our previous studies with ppGalNAc-T2 on a similar glycopeptide, GPII (Perrine et al. 2009), thus T2 may have a very limited ability to glycosylate directly near a prior site of glycosylation.

Glycosylation site determination on GP(T*10)C

Edman sequencing of the GP(T*10)C products of ppGalNAc-T2, T3, T4, T5, T7, T10, T12 and PGANT7 were performed to locate transferase specific glycosylation sites relative to the central Thr-O-GalNAc (T*). PTH derivatives at each cycle were counted for [3H]-GalNAc incorporation and plotted in Figure 5A–H. For clarity, diagrams indicate the sites of glycosylation on GP(T*10)C are shown in the right panels of Figure 5A–H, while for panels A–D diagrams representing the glycosylation of the lectin domain probing GP(T*22)R and GP(T*10)L are also shown below the GP(T*10)C notation. For the family I transferases, a peak of GalNAc incorporation is observed at the +5 position from the T* for ppGalNAc-T2 and T5, and at the +5 position for ppGalNAc-T3 (see Figure 5A–C). As discussed below, these remote +5 positions correlate with the lectin domain substrate preferences for these isoforms and may therefore reflect remote lectin domain interactions with the T* of GP(T*10)C. Note that these family I transferases show no peaks of [3H]-GalNAc incorporation at X positions 1–4 residues adjacent to the T* suggesting that the catalytic domains of these transferases do not recognize T* and may indeed be

4 A noticeable feature for the plots is the sequencing lag in [3H] content following a peak of maximum incorporation. This is due to the lower solubility of the PTH-Ser-O-GalNAc derivative, compared with the standard amino acid PTH derivatives, in the organic solvents used to extract the sample filter during the Edman sequencing.

5 Glycosylation of A2 (PT*2TDSTTPAPTTKK) and A4 (PTTDST* TPAPTTKK) by ppGalNAc-T4 is observed at Thr-7 in both substrates. In A2, Thr-7 is five C-terminal positions (+5) from the T*2 and in A4 Thr-7 is one C-terminal (+1) position from the T*B.
Fig. 5. [3H]-GalNAc incorporation into the X acceptor residues of GP(T*10)C revealing specific sites of glycosylation that vary with isoform. Edman amino acid sequence analysis of [3H]-GalNAc incorporation into GP(T*10)C by ppGalNAc-T2 (A), T3 (B), T5 (C), T4 (D), T12 (E), T10 (F), T7 (G) and the fly PGANT7 (H). At the right of each panel is representations depicting major sites of glycosylation (arrows) for GP(T*10)C. Below these representations, for panels A–D, representations depicting the sites of glycosylation (X’s and arrows) of the lectin probing glycopeptides GP(T*22)R and GP(T*10)L are given (determined previously (Gerken et al. 2013) or from this work). All plots are for overnight incubations of GP(T*10)C with the indicated transferase and are representative of at least two sequence determinations. This figure is available in black and white in print and in color at Glycobiology online.
confirmed in the present study with GP(T*10)C (see Figure 5F). ppGalNAc-T7, in the same subfamily as ppGalNAc-T10 shows identical behavior with the GP(T*10)C substrate, (Figure 5G) as does PGANT7 the fly orthologue to human T7 (Ten Hagen et al. 2003; Bennett et al. 2012) (Figure 5H). It is noteworthy that these transferases do not exhibit the −5 or +5 preferences suggesting that little lectin domain involvement in glycosylating GP(T*10)C. Thus, with GP(T*10)C we have shown that both subfamilies of "glycopeptide-prefering" transferases studied to date glycosylate unique and specific neighboring sites on glycopeptide substrates. Because these sites are only 1–3 residues from the T* we conclude that this specificity must be due to T* binding in the substrate binding cleft of the catalytic domain of these transferases.

**Comparison with nonglycosylated control GP(A10)C**

To further confirm that the glycosylation site specificity observed with GP(T*10)C is due to the T* binding to the transferase's catalytic (or lectin) domain and not due to end effects, or any other artifact, we preformed glycosylation reactions using a subset of transferases against both GP(T*10)C and GP(A10)C, where in the latter substrate the central T* was replaced by an Ala residue. Due to limited transferase availability, reactions were performed using ppGalNAc-T2, T4 and T10 representing subfamily I, IIa and IIb, respectively. Their gel filtration and Edman sequencing plots are shown in Figure 6A–C (right and left panels, respectively). The gel filtration plots show that ppGalNAc-T2 transfers slightly more [3H]-GalNAc to GP(A10)C than GP(T*10)C, while with ppGalNAc-T4 and T10, transfer to GP(T*10)C is significantly greater than to the control GP(A10)C which is very low. As expected, the Edman sequencing of the GP(A10)C products of ppGalNAc-T4 and T10 (Figure 6B and C) displays very low incorporation plots and shows no specific peaks of glycosylation which is in contrast to what is observed with GP(T*10)C (see Figure 5D and F). These results clearly confirm the glycopeptide requiring activity of these transferases, where a prior T* (or S*) must be within 1 to 4 residues of the acceptor site. On the other hand, with ppGalNAc-T2, a neighboring T* (or S*) is nearly inhibitory, thus only the remote flanking −5 site is glycosylated in GP(T*10)C (see Figure 5A), which is likely lectin domain mediated. Therefore, the ppGalNAc-T2 glycosylation of GP(A10)C, which does not contain the inhibiting T*, gives a broad peak of [3H]-GalNAc incorporation with a maximum observed directly N-terminal to the center Ala (Figure 6A). These results clearly indicate that ppGalNAc-T2's
catalytic domain can only readily glycosylate nonglycosylated “naked” peptide substrates.

**Lectin domain-mediated glycosylation site determination**

As performed previously for the family I ppGalNAc-T's (Gerken et al. 2013), Edman amino acid sequencing of the GP(T*22)R and GP(T*10)L products glycosylated by ppGalNAc-T4, T7, T12 and PGANT7 were performed to determine the optimal distance between the lectin bound T* and the [3H]-GalNAc glycosylated X residues. These plots are shown in Figure 7. Unlike the plots for GP(T*10)C, these plots show a broad distribution of maximum incorporation for the preferred substrate for ppGalNAc-T4 and T12. The plots of the GP(T*10)L [3H]-GalNAc incorporation for the two transferases, Figure 7A and B, suggest that ppGalNAc-T4 could indeed glycosylate at the +5 position while T12 may not (i.e. T4’s [3H]-GalNAc incorporation plot could be easily extrapolated to the +5 position while T12’s plot would not). This suggests that lectin domain-mediated glycosylation for ppGalNAc-T4 begins at the +5 residue while the weaker lectin domain-mediated glycosylation of ppGalNAc-T12 begins at (or after) the +6 residue. Therefore, using the two different glycopeptides, we have obtained consistent data showing that there are differences between the lectin domains of these two ppGalNAc-T’s. We have summarized these results in cartoon form against the ppGalNAc-T family phylogenetic tree in Figure 8.
In this work, we have extended our prior studies of the “peptide-preferring” family I, transferases (ppGalNAc-T1, T2, T3, T5, T6, T13, T14 and T16) to members of the “glycopeptide-preferring” family II, transferases (ppGalNAc-T4, T7, T10, T12 and PGANT7) using our series of lectin domain probing glycopeptide substrates (Table I and Figures 2 and 3). In addition, we have introduced a new glycopeptide substrate, GP(T*10)C (Table I), to address the role of directly neighboring glycosylation and to reveal specific sites of glycosylation catalyzed by the catalytic domain (Figure 5). Using these glycopeptide substrates, the roles of the catalytic and lectin domains in glycosylating glycopeptide substrates can be systematically characterized in terms of neighboring (1–5 residues) and remote (6–17 residues) prior glycosylation.

We have found that 4 of the 5 “glycopeptide-preferring” family II transferases have elevated activities towards GP(T*10)C and that each glycosylates unique sites on this substrate. This suggests that the...
binding of the Thr-O-GalNAc residue of GP(T*10)C at specific sites in the catalytic domain directs subsequent neighboring site selection in these transferases. We have also observed that the lectin domain probing glycopeptides, GP(T*22)R and GP(T*10)L, show variable activities with the “glycopeptide-prefering” transferases; hence, the involvement of the lectin domain in directing remote glycosylation for these transferases ranges from nearly none to relatively high depending on isoform.

In contrast, the “peptide-prefering” family I transferases characterized to date show the greatest activity for one or both of the lectin domain probing glycopeptides, indicating that Thr-O-GalNAc binding at the lectin domain directs and enhances remote glycosylation. The “peptide-prefering” transferases all have very low activities against the catalytic domain probing glycopeptide, GP(T*10)C, and do not glycosylate directly neighboring (1–4) sites, strongly suggesting that there are no Ser/Thr-O-GalNAc-binding sites in the catalytic domain of these transferases.

These results demonstrate that there are at least two different modes of glycopeptide substrate recognition by the ppGalNAc-Ts. These would be the recognition of a glycopeptide GalNAc at the lectin domain, directing remote glycosylation to the catalytic domain (Figure 9A) and the direct binding of glycopeptide GalNAc within the catalytic domain, targeting specific neighboring glycosylation (Figure 9B). We shall refer to these activities as lectin domain assisted and catalytic domain directed which can operate singly or together depending on the isoform.

We would also like to address the overlapping peptide/glycopeptide substrate preferences for the different transferase isoforms and suggest renaming of the classifications of family I and family II transferases. Since subfamily Ia–d members have elevated activities against glycopeptide substrates with a remote prior glycosylation site and thus not strictly peptide-prefering transferases, we suggest that for clarity, these family Ia–d members be reclassified as remote glycopeptide/peptide-prefering isoforms (abbreviated as remote GP/P-prefering). Similarly, the subfamily Ila members have been shown to be relatively active towards a number of nonglycosylated peptide substrates and therefore not strictly glycopeptide-prefering transferases, we suggest that the subfamily Ila members be termed mixed (glyco)peptide-prefering isoforms (abbreviated as mixed (G)P-prefering). Since subfamily Ilb members are poorly active towards naked peptide substrates, we suggest calling them strict glycopeptide-prefering isoforms (abbreviated as strict GP-prefering).

Mixed (Glyco)peptide-prefering transferases

ppGalNAc-T4 and T12 (subfamily Ila)

Previous work on ppGalNAc-T4 has shown it to be highly active against Ser/Thr-O-GalNAc glycosylated substrates and that its lectin domain is required for this activity (Hassan et al. 2000; Wandall et al. 2007). With our catalytic and lectin domain probing substrates, we have presented data that strongly suggests that ppGalNAc-T4 glycopeptide glycosylation is directed by both its catalytic and lectin domains depending on the distance between the acceptor sites and the existing Ser/Thr-O-GalNAc. Thus the catalytic domain of ppGalNAc-T4 directs glycosylation to one residue (+1) C-terminal of an existing T* while its lectin domain assists in directing glycosylation from +5 to over +13 residues C-terminal from the T*. Interestingly, ppGalNAc-T4 displays the highest lectin domain activity (against GP(T*10)L) of all the “glycopeptide-prefering” family II members studied, perhaps explaining prior conclusions that its lectin domain was solely responsible for the glycopeptide activity.

ppGalNAc-T12 is 72% identical to ppGalNAc-T4 and also utilizes its catalytic and lectin domains to recognize and glycosylate...
Ser/Thr-O-GalNAc containing glycopeptides. In spite of its high-
sequence conservation to ppGalNAc-T4, its catalytic domain glyco-
ylates three residues (−3) N-terminal from an existing T* while its lectin
domain assists in directing glycosylation from +6 to over +13 residues
C-terminal from a T*, ppGalNAc-T4 and T12 therefore serve the so-
called glycopeptide filling-in roles whose actual specificities had been
unknown until this study. In keeping with this activity, both trans-
ferases are highly expressed in tissues that produce densely glyco-
sylated mucin glycoproteins (Young et al. 2003).

Our previous studies on ppGalNAc-T12 revealed that it also has
good activities towards nonglycosylated random peptide substrates
showing the “classical” ppGalNAc-T preferences for Pro at the +1 and
+3 positions relative the site of glycosylation (Gerken et al. 2011)
and possessing the conserved aromatic residue motifs that have been
linked to this specificity as shown in Supplementary material,
Figure S1 (Gerken et al. 2011; Lira-Navarrete et al. 2015). This
aromatic residue motif is also conserved in ppGalNAc-T4 but not in
ppGalNAc-T7, T10 or PGANT7 further confirming that ppGalNAc-T4 and T12 have mixed roles serving to glycosylate both peptide and glycopeptide substrates.

Strict glycopeptide-preferring transferases

ppGalNAc-T7 and T10 and PGANT7 (subfamily IIb)

ppGalNAc-T7 and T10 possess 48% sequence identity and will both
glycosylate one residue N-terminal (−1) of a prior site of glycosylation
which is highly suggestive of a catalytic domain-directed activity.
The lectin domain-assisted glycosylation for these transferases is very low
in contrast to the subfamily IIa mixed (glyco)peptide-prefering trans-
ferases. Therefore, the removal of the lectin domain of ppGalNAc-T10
is found to only slightly alter its long-range glycosylation specificity
against glycosylated substrates but it does not affect its ability to gly-
cosylate at the −1 position relative to a T* (Raman et al. 2008). We
have previously shown (utilizing additional random (glyco)peptides
not described here) that ppGalNAc-T10 is highly selective for glyco-
sylating directly N-terminal (−1) of an S* or T*, but nevertheless
can only very poorly glycosylate nonglycosylated random peptides
(Perrine et al. 2009) and lacks the “classical,” (T/S)PXP, Pro-binding
motifs as does ppGalNAc-T7.7 Therefore, ppGalNAc-T10 and T7
likely possess nearly identical specificities and similarly function as
a strict filling-in transferases, glycosylating directly N-terminal of an
existing S* or T*. This is again in keeping with their high expression
in tissues that produce densely glycosylated mucins (Young et al. 2003).

The Drosophila PGANT7 is classified in the same IIb subfamily
with other mammalian GALNT genes (Bennett et al. 2012) and shares
a 41% sequence identity with ppGalNAc-T7. Unsurprisingly,
PGANT7 displays the identical catalytic domain-directed strict
requirement for an N-terminal (−1) glycosylation site relative to T*
that is observed for ppGalNAc-T7 and T10. In contrast to
ppGalNAc-T7 and T10, PGANT7 will nearly equally glycosylate
remote sites from 6 to ~13 residues N- or C-terminal from a T*, indicat-
ing that PGANT7 possesses the ability for long-range glycosylation
that must be mediated by its lectin domain. Thus, PGANT7 possesses
both strong catalytic domain-directed and lectin domain-assisted

6 Preliminary studies on random peptide substrates show that
ppGalNAc-T4 possesses the same TPXP substrate preference motif
(data not shown).

7 Preliminary random peptide studies also suggest similar behavior for
ppGalNAc-T7 which also lacks the Pro residue sequence recognition
motif (data not shown).

activities, with the latter activity lacking in its mammalian ortholo-
gues. Whether PGANT7 (which also lacks the “classical” Pro-binding
motifs) can glycosylate nonglycosylated peptides is presently
unknown.

Correlation of lectin domain binding preference
with transferase sequence

In our previous work, we attempted to determine whether any linker
or lectin domain motifs might relate to a transferase’s remote glyco-
peptide N- or C-directional preference by aligning their linker and lec-
tin domains (Gerken et al. 2013). As we have previously observed for
the family I transferases, the linker domains within the subfamily IIa or
IIb transferases do not reveal any obvious differences in sequence or
length that can account for the differences in their remote lectin
domain-assisted glycosylation behavior (see Supplementary material,
Table S1). Indeed, the PGANT7 linker shows the greatest similarity to
the ppGalNAc-T10 linker, compared with ppGalNAc-T7, despite the
fact that PGANT7 and T10 have significantly different lectin
domain-assisted activities. It is also worth noting that the mixed
(glyco)peptide-substrate ppGalNAc-T4 and T12 (subfamily IIa) linkers
appear to be more similar to the remote glycopeptide/peptide
class I linkers than to the more strict glycopeptide-prefering subfamily
IIb linkers which are all ~5 residues longer.

Several studies of the ricin lectin α, β and γ-subdomains in multiple
species, from bacteria to mammals, have revealed a number of sugar
binding sequence motifs, the most common being the CLD and QXW
sequences (Hazes 1996; Imberty et al. 1997; Fujimoto et al. 2000;
Maveyraud et al. 2009). Mutagenesis studies have shown that the
Asp residue of the CLD motif is typically required for lectin-binding
activity (see Figure 1B). These motifs are variously found in the lectin
domains of the ppGalNAc-Ts where mutual and structural studies have
revealed that only specific lectin subdomains may actively bind
GalNAc (Hassan et al. 2000; Tenno et al. 2002; Fritz et al. 2004;
Pratt et al. 2004; Kubota et al. 2006; Wandall et al. 2007; Pederson
et al. 2011).

In our previous work on the family I transferases, we found no obvi-
ous correlation in the ricin domain motifs and the transferase’s glyco-
peptide directional preferences (Gerken et al. 2013) and come to the
same conclusion with the family II transferases characterized in this
work (see Supplementary material, Table S1). In fact, we may again ob-
serve an inverse correlation between the mammalian ppGalNAc-T7,
T10 and the fly PGANT7. For ppGalNAc-T10 all three lectin subdo-
 mains possess likely binding motifs and its X-ray structure shows
Ser-O-GalNAc bound to its β-subdomain (Kubota et al. 2006). How-
ever, no significant activity is observed with our lectin domain probing
substrates. Likewise, ppGalNAc-T7 has only one canonical lectin
motif, the γ-subdomain, and it shows little activity with either of
our lectin domain substrates. In contrast, PGANT7 contains no
truly canonical motifs (its α-subdomain having the closest: CLD-QLV)
but displays highly significant activities against both lectin domain
probing glycopeptides. Since PGANT7 displays clear long-range glyco-
peptide specificities, at least one of its lectin subdomains must pos-
sess significant, although weak, binding activity to account for its
long-range glycopeptide specificity.

Such weak lectin binding is consistent with prior studies on the
ppGalNAc-Ts, as direct glycopeptide binding to any of the
ppGalNAc-Ts lectin domains, regardless of motif, cannot be readily
detected (Wandall et al. 2007; Pederson et al. 2011; Yoshimura
et al. 2012), confirming the weak binding nature of the ppGalNAc-T
lectin domains. This is sensible, since tight lectin binding of
glycopeptide substrates would likely be nonproductive, slowing glyco-
sylation due to delayed product release. As we discussed earlier (Ger-
ken et al. 2013), lectin domain binding may simply increase the local
concentrations of substrate while orientating it in the catalytic domain
cleft. Clearly, further detailed structural studies are needed to begin to
understand the roles of the lectin domains in these transferases. The
recently reported X-ray crystal structures, small angle X-ray scattering
and solution modeling studies of ppGalNAc-T2 interacting with
MUC5AC glycopeptides is an important step in this direction (Lira-
Navarrete et al. 2013) confirming the dynamic flexibility between the
ppGalNAc-T2 catalytic and lectin domains (Gerken et al. 2013).

**UDP-GalNAc hydrolysis**

In our studies of the ppGalNAc-T’s, we observed variable non-
productive hydrolysis of the UDP-[3H]-GalNAc donor (i.e. transfer
to water) vs. transfer to peptide/glycopeptide substrate. Presently, we
do not know what factors are involved but there is clearly competition
with transfer of GalNAc to water and transfer to substrate. Interest-
ingly, the family I “peptide-prefering” and family II “glycopeptide-
depending” transferases behave very differently, where the lowest
hydrolysis is found with their preferred substrates. Thus, the rate of
nonproductive hydrolysis may be a measure of how well an acceptor
substrate productively binds to the active site of the catalytic domain.

**Potential limitations**

The random glycopeptide substrates utilized in these studies are rela-
tively short and contain a relatively high content of Gly and Pro com-
pared to most proteins, hence these are expected to possess extended
random structures that cannot fully represent the complexity of a
folded protein. We recognize that this may be a limitation to our meth-
ods, particularly since previous work has shown that ppGalNAc-T11
recognizes the tertiary structure of the linker regions of the LDLR re-
ceptor (Pedersen et al. 2014). Nevertheless, these (glyco)peptide sub-
strates have provided highly useful and previously not understood
data that has helped us more fully understand the unique roles of the
catalytic and lectin domains of these transferases. These data will be
substantially useful in future studies of the ppGalNAc-T against both folded and extended proteins and peptides.

**Conclusion**

With these studies, a clearer understanding the functions of the lectin
and catalytic domains in modulating glycopeptide substrate specificity
has been revealed. For example, we can distinguish between the effects of
neighboring (+/−5 residues) prior glycosylation, due to glycopep-
tide binding at the catalytic domain, to the longer range effects of re-
move glycosylation (+/−6 to 17 residues), where glycopeptide binding
is mediated by the lectin domain (Figure 8). Of particular interest, we
have observed clear and unexpected differences in activity and direc-
tionality among the transferases studied to date.

For the first time, we have identified multiple catalytic domain
GalNAc-binding sites for the “glycopeptide-prefering” transferases
in subfamilies IIa and IIb that likely account for their glycopeptide-
specific activities. These transferases are commonly called filling-in
transferases and in this work we now have identified site preferences
for their respective activities. We have also observed that for the
“peptide-prefering” (i.e. remote GP/P-prefering) family I, trans-
ferases that the lectin domain is nearly always involved in remote gly-
cosylation of glycopeptides and that this may be modulated in an N-or
C-terminal direction. This lectin domain involvement is only variably
observed for the “glycopeptide”-prefering family II transferases.

We have further demonstrated the partial conservation of speci
cificity between the fly and human orthologues PGANT7 and ppGalNAc-T7,
which display identical catalytic domain-directed glycopeptide specificities, but quite different remote lectin domain glycopeptide specificities.

With these studies, we now have a better understanding perhaps
why this large family of transferases exists. Each isoform contains a
unique combination of catalytic domain-directed (glyco)peptide speci-
ficity and lectin domain remote glycopeptide specificity. Thus, for ex-
ample, the isoforms in the Ia (ppGalNAc-T1 and T13) and Ib
(ppGalNAc-T2, T14, T16) subfamilies of remote GP/P-prefering transferases all have nearly indistinguishable random peptide se-
quence preferences (Gerken et al. 2011) but possess different lectin
domain preferences (Gerken et al. 2013). Furthermore, the fact that
nearly all ppGalNAc-T isoforms recognize either prior neighboring
glycosylation or prior remote glycosylation, or a combination of
both, suggests that these transferases must operate in an ordered pro-
cess to perform their glycosylation functions, where one transferase
produces an optimal glycopeptide substrate for its self or another
transferase. Importantly, experimental evidence indeed exists suggest-
ing an ordered process for ppGalNAc-T glycosylation (Iida et al.
2000; Kato et al. 2001; Norden et al. 2015). This suggests that mucin-
type O-glycosylation is potentially highly orchestrated in a cell and
helps to further explain why the patterns of ppGalNAc-T isoform ex-
pression varies from cell type to cell type (Bennett et al. 2012). Thus,
the overall expression or suppression of a single transferase in this cas-
cade could lead to multiple downstream alterations of subsequent gly-
cosylation. The ultimate goals of our work are, therefore, to fully
understand the properties of these transferases so that fully predictive
models of mucin-type O-glycosylation can be developed.

**Materials and methods**

**Reagents and random peptide substrates**

All the random (glyco)peptide substrates listed in Table I (Gerken et al.
2013) were custom synthesized by Sussex Research, Ottawa, ON
(Canada). Note that GP(T*22)R, GP(A22)R, GP(T*10)L and GP(A10)L were previously named GPV, GPV-Cont, GPV and
GPV-Cont (Gerken et al. 2013). Stock solutions of 50 mg/mL (15–
17 mM) of random (glyco)peptide substrates were prepared by lyoph-
ilizing from water several times and adjusting to pH 7−7.5 with dilute
NaOH/HCl. Fully N-acetylated UDP-[3H]-GalNAc was purchased
from American Radiolabeled Chemicals Inc. Dowex 1 × 8 anion ex-
change resin (100–200 mesh) was purchased from Acros Organics
and Sephadex G10 was obtained from Amersham Biosciences and
GE Healthcare. Edman amino acid sequencing was performed on an
Applied Biosystems Procise 494 peptide sequencer as previously de-
scribed (Gerken et al. 2006, 2011; Perrine et al. 2009). Liquid scintil-
lation counting was performed using a Beckman Model LS650.

**Transferases**

As in our previous work (Gerken et al. 2013), ppGalNAc-T’s were ob-
tained from multiple sources and multiple expression systems as
N-terminal truncated and affinity tagged constructs. They were used

8 T. Gerken and L. Revoredo unpublished data, available on the World
Wide Web “Isoform Specific O-Glycosylation Prediction, ISOGlyP”
(http://isoglyp.utep.edu) M-Y Leung, J. Mohl, G. Cardenas, and
I. Almeida, Boarder Biomedical Research Center, University of Texas
at El Paso.
as transferase bound to affinity beads or affinity purified soluble transferase. Experiments completed using transferases from different sources gave comparable results (data not shown). Soluble affinity purified human ppGalNAc-T2 and T3 were expressed from *Pichia pastoris* (Fritz et al. 2006; Raman et al. 2008; Gerken et al. 2011). PolyHis-tagged ppGalNAc-T4, T7, T12 and T13 were expressed in the baculovirus-infected Sf9 cell system9 (Toth et al. 2014). These transferases were expressed as secreted, N-terminally His-tagged catalytic domains, bound to Ni-NTA-affinity beads (Thermo-Fisher) and used directly for enzyme assays after extensive washing. Control media exposed affinity beads revealed no significant activities against UDP-[3H]-GalNAc (data not shown). Soluble N-terminal polyHis tagged ppGalNAc-T4, T5, T10 and T12 were expressed using methods similar to the expression of rat ST6GalI (Barb et al. 2012) using HEK293f cells and purified9 by Ni-NTA superflow (Qiagen) nickel affinity chromatography. Soluble N-terminal polyHis-tagged ppGalNAc-T4, T7, T12 and T16 were expressed from High Five insect cells and purified on Ni-NTA agarose (Invitrogen) and MonoQ 5/50 GL ion exchange chromatography (GE Healthcare) as described previously (Bennett et al. 1999; Hassan et al. 2000; Vester-Christensen et al. 2013). Detailed prior unpublished methods of the cloning and expression of ppGalNAc-T12 are found in the Supplementary Materials section. Bovine ppGalNAc-T1, which possesses 98.9% sequence identity with human T1, was a kind gift of Ake Elhammer (Kalamazoo, MI) (Homa et al. 1993).

*Drosophila* PGANT7 was expressed in COS7 cells and the recombinant secreted protein of the cell culture media was used (Ten Hagen et al. 2003). Control media reactions gave no significant UDP-[3H]-GalNAc utilization in overnight incubations compared with expression media (data not shown).

**Glycosylation of lectin and catalytic domain probing random glycopeptide substrates**

The following conditions were used to glycosylate random glycopeptide substrates GP(T+22)R, GP(T+10)L (and their nonglycosylated controls) and GP(T+10)C: 68 mM sodium cacodylate, pH 6.5, 1.8 mM 2-mercaptoethanol, 10 mM MnCl2, 50 µM [3H]-radio labeled UDP-GalNAc (~6 × 10^6 DPM/µmol), 1.5–1.7 mM (5 mg/mL) of (glyco)peptide substrates and up to 68 µL of soluble transferase or 100 µL transferase bound affinity beads. Total reaction volumes were 75–200 µL and carried out in 500 µL capped Eppendorf tubes. Reaction mixtures were incubated at 37°C in a TAITEC shaking Mi croincubator M-36. Time-point aliquots of 15–40 µL were removed at 15, 45, 120, 240 and ~1200 (overnight) min after initiating the reaction and quenched by the addition of 1 volume of 250 mM EDTA. Reaction controls were carried out without glycosylpaset present. Typically, experiments were performed with all (glyco)peptide substrates (except GP(A10)C) at the same time with the same transferase concentrations and the same UDP-[3H]-GalNAc stock. Separate selected transferase reactions comparing GP(T+10)C and GP(A10)C were also performed.

Experiments were performed using transferase concentrations (determined by trial and error) that would transfer between 10 and 50% of the total UDP-GalNAc to the optimal substrate after an overnight incubation (giving a range of ~0.003–0.016 mol GalNAc transferred per mole of glycopeptide). In some cases, higher specific activity of UDP-[3H]-GalNAc was used for reactions that were to be analyzed by Edman sequencing. UDP and non-hydrolyzed UDP-GalNAc were removed by passing the sample through a column of ~3 mL of Dowex 1 × 8 anion exchange resin. Total UDP-[3H]-GalNAc utilization (i.e. transfer to substrate and hydrolysis) was determined by difference, after scintillation counting 150 of the sample before and after passing over the Dowex column. These data are typically reported as plots of mole fraction of total UDP-GalNAc utilized vs. time. Determination of the extent of [3H]-GalNAc transfer to peptide and the extent of UDP-[3H]-GalNAc hydrolysis was determined by Sephadex G10 gel filtration analysis of selected reaction time points as previously described (Gerken et al. 2013). Typically, 2–4 independent experiments were performed at the optimal transferase concentration, after several trial and error determinations of each transferase’s activity against each substrate. Representative plots of these final experiments are given in the figures. The substrate ranking trends in the trial and error determinations followed the same trends that were observed in the final presented data, although their relative ratios varied with transferase concentration as reported in prior work (Gerken et al. 2013).

Determination of substrate glycosylation sites was performed by Edman amino acid sequencing of the G10 isolated [3H]-GalNAc glycosylated substrate as previously described (Gerken et al. 2013). Although there is commonly sample-to-sample variability in the [3H] content loaded on the sequencer (due sample losses and different initial UDP-[3H]-GalNAc-specific activities), the observed sites of incorporation were found to be identical between samples. This is also true when comparing data obtained from entirely different batches of transferases (i.e. ppGalNAc-T3, T10 and T12) performed a year or more apart. Sequence lag in [3H] content is also commonly observed (A noticeable feature for the plots is the sequencing lag in [3H] content following a peak of maximum incorporation. This is due to the lower solubility of the PTH-Ser-O-GalNAc derivative, compared with the standard amino acid PTH derivatives, in the organic solvents used to extract the sample filter during the Edman sequencing.).

**Supplementary data**

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

**Funding**

This work was supported by grants from the National Institutes of Health: R01-CA78834 and U01-GM113534 (to T.A.G.), P41-GM103390 and P41-RR005351 (to K.W.M. and D.L.J.), P01-GM107012 (to G.J. Boons, Complex Carbohydrate Research Center, Athens, GA) and NIDCR intramural program grant (to L.A. T. and K.T.H.). Funding by the Danish National Research Foundation (DNRF107) is also acknowledged (to H.C.).

**Acknowledgements**

We thank the following for their involvement in producing the transferases utilized in this work: Melissa Stuart (Jarvis laboratory), Duy Tran and Liping Zhang (Ten Hagen laboratory), Malene Bech Vester-Christensen and Yun Kong (Claussen Laboratory), Roy W. Johnson, Shuo Wang and Heather A. Moniz (Moremen Laboratory).
Conflict of interest statement

None declared.

Abbreviations

GalNAc, N-acetylgalactosamine; GP(T*22)R, GP(T*22)R-C; GP(T*10)L, GP(T*10)C, random glycopeptides as defined in Table 1 [note that these glycopeptide substrates are color-coded throughout Figures 2–7 and Table 1 as: GP(T*22)R(red), GP(A22)R (pink), GP(T*10)L (blue), GP(A10)L (turquoise), GP(T*10)C (green) and GP(A10)C (yellow)]; Ni-NTA, nickel-nitrilotriacetic acid; ppGalNAc-T, UDP-GalNAc-polypeptide N-acetylgalactosaminyl-transferrases; PTH, phenylthiohydantoin amino acid; T*, Thr-O-GalNAc; S*, Ser-O-GalNAc.

References


**Supplemental Methods (Revoredo et. al.)**

**Expression and Purification of Human ppGalNAc-T12 Expressed in Insect Cells**

A secreted His-tagged expression construct encoding amino acid residues 33-581 of ppGalNAc-T12 (GenBank AJ132365) was generated from MKN45 cDNA using primers EBHC1051 (5'-GCGGAATTC GGCTCGGTGCTGCGGGCGCAGC-3')/EBHC1031 (5'-GCGGCGGCCGC CGGTTCTGGTCACTGCTTAGC-3'), artificial flanking restriction sites are underlined and using EBHC1031 for RT mediated cDNA synthesis. A His-tag, thrombin, T7-tag linker (5'-ggatcc cgggtacc ttctagagacgacgacagacgacgacgacgacgacgacgacgacgagctgtgatctctggcgctgtgactgtgactgtgactgtgactgtgactgtgaatc-3' (restriction sites underlined)) encoding -DPGYLLESSHHHHHSSGLVPRGSHMASMT-GGQQMGEF- was inserted into the BamHI-EcoRI site of the Baculoviral expression vector pAcGP67A (Pharmingen/BD Biosciences, USA), generating pAcGP67A-HIS. Subsequently the secreted GalNAc-T12 EcoRI/NotI cDNA insert was inserted into the EcoRI/NotI site of pAcGP67A-HIS, generating pAcGP67A-HIS-GalNAc-T12. pAcGP67A-HIS-GalNAc-T12 was cotransfected with Baculo-Gold™ DNA (Pharmingen/BD Biosciences, USA) and recombinant Baculovirus obtained after two successive amplifications in Sf9 cells, as described previously (Schwientek, T. et al. 2002).

The isolation of ppGalNAc T12 was performed as previously described for ppGalNAc T1-T3 and T11 (Vester-Christensen, M.B. et. al. 2013) specifically modified for ppGalNAc T12 as described below. Supernatant from infected High Five cells (4 x 250 mL) were harvested 96 h post-infection by centrifugation (2,000g, 4 ºC, 10 min). The supernatant (3 x 330 mL) were dialyzed (Spectra/Por dialysis membrane, 10 kDa cut-off; Spectrum Laboratories) against 3 x 5 L 50 mM Tris pH 8, 150 mM NaCl, and applied to an Ni-NTA column (2.5 ml) equilibrated in 25 mM Tris pH 8.0, 300 mM NaCl, 10 mM imidazole using gravity flow. The column was washed with 25 column volumes of 25 mM Tris pH 8.0, 300 mM NaCl, 10 mM imidazole. Protein was eluted with 25 mM Tris pH 8.0, 300 mM NaCl, 250 mM imidazole. Fractions (0.5 mL) were collected in tubes containing 4.5 mL 25 mM Tris pH 8.0, 300 mM NaCl. ppGalNAc-T12 containing fractions were pooled and concentrated to 13 mL by Centricon filtration (10 kDa cut-off, EMD Millipore Corp.) before diluted 30-fold in 25 mM Tris pH 7.5 (360 mL total volume). The sample was applied (1 mL/min) to a MonoQ column (5/50 GL, GE Healthcare) equilibrated in buffer containing 25 mM Tris pH 7.5, 10 mM NaCl and the column was washed in the same buffer for 15 column volumes at the same flow rate. Protein was eluted with linear gradient of 25 mM Tris pH 7.5, 10 mM NaCl to 500 mM NaCl over 30 column volumes. An ÄktaFPLC (GE Healthcare) interfaced by UNICORN 4.0 control software (GE Healthcare) was used for the chromatographic procedure. ppGalNAc-T12 containing fractions were pooled and stored at 4C. Quantification of enzyme protein concentration was performed by comparative SDS-PAGE Coomasie stained gels using BSA as standard.

References


**Supplemental Figure S1 (Revoredo et. al )**

**Figure S1. Correlation of Catalytic Domain Residues with the "Classic" T-P(A/G)P Substrate Motif.** The top figure displays the X-ray crystal structure of the ppGalNAc T2 catalytic domain bound to EA2 peptide substrate (Fritz, T.A., et al. 2006). The prominent residues thought to be involved in the binding of the EA2 peptide substrate are *space filled* as are the EA2 substrate residues. ppGalNAc T2 substrate PAP interacting residues are: Phe 361 (steel blue), Phe280 (olive green), and Trp 282 (green). The EA2, PAP residues Pro 8 (purple), Ala 9 (orange) and Pro 10 (purple) are also indicated. The N- terminal and C-terminal EA2 residues, Ser 5 (blue), Lys 13 (red) are labeled while the remaining EA2 residues are colored orange. The table below the figure lists the conservation of each isoform's P(A/G)P substrate motif as Y or N (for Yes or No, respectively) along with the identity of the conserved P(A/G)P interacting aromatic residues. Note the high correlation for the residues across multiple subfamilies. P(A/G)P motif, preferences were obtained from (Gerken, T.A., et al. 2011) or from unpublished preliminary data the latter being indicated as Y* and N*. 


EA2 Peptide: $\text{PTTD-S}_5\text{TT}_7\text{PAPTTK}_{13}$

<table>
<thead>
<tr>
<th>Sub-Family</th>
<th>Transferase</th>
<th>P(A/G)P Motif</th>
<th>Phe361</th>
<th>Phe280</th>
<th>Trp282</th>
</tr>
</thead>
<tbody>
<tr>
<td>IB</td>
<td>T2</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T14</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T16</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IA</td>
<td>T1</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T13</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IC</td>
<td>T3</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>ID</td>
<td>T5</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IF</td>
<td>T11</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IIA</td>
<td>T4</td>
<td>Y*</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>T12</td>
<td>Y</td>
<td>F</td>
<td>F</td>
<td>W</td>
</tr>
<tr>
<td>IIB</td>
<td>T7</td>
<td>N*</td>
<td>Y</td>
<td>W</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>T10</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td>PGANT 7</td>
<td>-</td>
<td>Y</td>
<td>Y</td>
<td>E</td>
</tr>
</tbody>
</table>
### Supplemental Table S1 (Revoredo et al.)

**Table S1  ppGalNAc T Linker Domain and Lectin Domain Motif Alignments**

<table>
<thead>
<tr>
<th></th>
<th>Catalytic Domain</th>
<th>Linker Domain</th>
<th>α-rinic</th>
<th>β-rinic</th>
<th>γ-rinic</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>PD_SQIPRYFSL</td>
<td>GEIRNVETNQ</td>
<td>CLD444</td>
<td>QVF</td>
<td>CLD484</td>
<td>QLW</td>
</tr>
<tr>
<td>T13</td>
<td>PD_SQIPRYFSL</td>
<td>GEIRNVETNQ</td>
<td>CLD443</td>
<td>QVF</td>
<td>CLD483</td>
<td>QLW</td>
</tr>
<tr>
<td>Ib</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>PE_LRVPDHQDIAFGALQQ--GTN</td>
<td>CLD458</td>
<td>QBE</td>
<td>CLT498</td>
<td>QKW</td>
<td>CLD541</td>
</tr>
<tr>
<td>T14</td>
<td>PE_LSIPKESSIQKGNIRQ--RQK</td>
<td>CLE432</td>
<td>QVW</td>
<td>CLS478</td>
<td>QKW</td>
<td>CLD519</td>
</tr>
<tr>
<td>T16</td>
<td>PE_LTVPVKEALP-SIIQK--GVN</td>
<td>CLE443</td>
<td>QAM</td>
<td>CLA488</td>
<td>QKW</td>
<td>CLE532</td>
</tr>
<tr>
<td>Ic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>PE_VYVPDLNPVISGYIKSGQPL</td>
<td>CLD519</td>
<td>QYF</td>
<td>CLH563</td>
<td>QIW</td>
<td>CLS607</td>
</tr>
<tr>
<td>T6</td>
<td>PE_MFVPDLTPTFYGAIKNLGTNQ</td>
<td>CLD511</td>
<td>QYF</td>
<td>CLH555</td>
<td>EEW</td>
<td>CLT599</td>
</tr>
<tr>
<td>Id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>PD_LRAPIVRAS</td>
<td>--GVLINVALGK</td>
<td>CIS824</td>
<td>QFQ</td>
<td>CIA860</td>
<td>LKW</td>
</tr>
<tr>
<td>IIa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>PN_LHVPEDRPGWHG AIRS-RG--SE</td>
<td>CLD459</td>
<td>QFF</td>
<td>CAE505</td>
<td>IIW</td>
<td>CLS549</td>
</tr>
<tr>
<td>T12</td>
<td>PE_LHVPEDRPGFFGMLQNGK--DY</td>
<td>CFD460</td>
<td>QFF</td>
<td>CIA508</td>
<td>QKF</td>
<td>CVT549</td>
</tr>
<tr>
<td>IIb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7</td>
<td>YD_ITSHYP-LPPKN DGEIRGFETAY</td>
<td>CID547</td>
<td>QLF</td>
<td>CLT587</td>
<td>KEW</td>
<td>CLD627</td>
</tr>
<tr>
<td>T10</td>
<td>WD_LPKFYPVPEPPAA WGEIRNVGTGL</td>
<td>CAD473</td>
<td>AAW</td>
<td>CFD525</td>
<td>QLW</td>
<td>CM565</td>
</tr>
<tr>
<td>Pgant7</td>
<td>YD_VYDKFPGG-LPANLHWGESLR SVASDG</td>
<td>CLD481</td>
<td>QLV</td>
<td>CVE521</td>
<td>SPQ</td>
<td>CM560</td>
</tr>
</tbody>
</table>

**Notes:**
- α, β, γ refer to different lectin domains.
- CLD, QXW, QQW indicate motif alignments.
- α+βγ indicates the presence of both α and β motifs with γ.

332
Table S1 Notes

a) Linker domain (L) residues shaded in dark grey are conserved (or similar) in the majority of the 13 transferases shown, while residues shaded in light grey are conserved only within a given transferase subfamily.

b) α, β and γ ricin lectin subdomain residues that are identical to the canonical CLD-QXW motif are shaded in green, while similar canonical motif residues are shaded in yellow. Non-canonical (and presumably inactivating) residues are shaded in red.

Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer

Srividya Venkitachalam1, Leslie Revoredo2, Vinay Varadan1,3, Ryan E. Fecteau1, Lakshmeswari Ravi4, James Lutterbaugh4, Sanford D. Markowitz1,4, Joseph E. Willis1,5, Thomas A. Gerken2,6,7 & Kishore Guda1,3

The molecular basis of aberrant protein glycosylation, a pathological alteration widespread in colorectal cancers (CRC), and the mechanisms by which it contributes to tumor progression remain largely unknown. We performed targeted re-sequencing of 430 glycosylation-associated genes in a series of patient-derived CRC cell lines (N = 31) and matched primary tumor tissues, identifying 12 new significantly mutated glycosylation-associated genes in colon cancer. In particular, we observed an enrichment of mutations in genes (B3GNT2, B4GALT2, ST6GALNAC2) involved in the biosynthesis of N- and Cores 1–3 O-linked glycans in the colon, accounting for ~16% of the CRCs tested. Analysis of independent large-scale tumor tissue datasets confirmed recurrent mutations within these genes in colon and other gastrointestinal cancers. Systematic biochemical and phenotypic characterization of the candidate wild-type and mutant glycosyltransferases demonstrated these mutations as either markedly altering protein localization, post-translational modification, encoded enzymatic activities and/or the migratory potential of colon carcinoma cells. These findings suggest that functionally deleterious mutations in glycosyltransferase genes in part underlie aberrant glycosylation, and contribute to the pathogenesis of molecular subsets of colon and other gastrointestinal malignancies.

Protein glycosylation is a key post-translational modification that plays a fundamental role in regulating multiple cellular processes including cell adhesion, migration, cell-cell recognition and immune surveillance1. Glycosylation of newly synthesized peptides may be initiated in both the Endoplasmic reticulum (ER) and Golgi apparatus and is catalyzed by a series of specific glycosyltransferases that may display overlapping specificities depending on the transferase1–4. These enzymes typically transfer single sugar residues from nucleotide-sugar donors to protein and sugar acceptors, the latter resulting in glycan elongation forming a vast array of glycan structures5. The resulting glycans are typically characterized as N-linked or O-linked based on the amino acid residues (Asn or Ser/Thr) the glycans are attached to, which also corresponds to their origin of initiation in the ER or Golgi respectively1. Aberrant protein glycosylation is a hallmark of many human cancers including colorectal cancers (CRC)6–8. However, the molecular basis of aberrant glycosylation and the mechanisms by which it contributes to tumor progression remain largely unknown. We previously reported the first finding of somatic and germline inactivating mutations in the gene encoding for GALNT12, a key enzyme involved in the initiating step of mucin type O-glycosylation, in a subset of colon cancer cases9. Our initial findings strongly suggest that mutations in O-glycosylation pathway genes may in part underlie aberrant protein glycosylation commonly seen in colon and other cancers, and potentially contribute to the development of a subset of these malignancies. Since both

1Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland OH 44106, U.S.A. 2Department of Chemistry, Case Western Reserve University, Cleveland OH 44106, U.S.A. 3Division of General-Medical Sciences-Oncology, Case Western Reserve University, Cleveland OH 44106, U.S.A. 4Division of Hematology and Oncology, Case Western Reserve University, Cleveland OH 44106, U.S.A. 5Department of Pathology, Case Western Reserve University, Cleveland OH 44106, U.S.A. 6Department of Pediatrics, Case Western Reserve University, Cleveland OH 44106, U.S.A. 7Department of Biochemistry, Case Western Reserve University, Cleveland OH 44106, U.S.A. Correspondence and requests for materials should be addressed to K.G. (email: gkishore@yahoo.com)
we identified 12 of the 36 genes to be significantly mutated in CRCs (P ≤ 0.05) (Table 1, Supplementary Table S5). Interestingly, we observed three of these genes B3GNT2, ST6GALNAC2, and B4GALT2 mapping to protein glycosylation pathways that are involved in the formation of polylactosamine chain extensions on N- and O-linked glycans (B3GNT2 and B4GALT2) and in the termination of O-glycan Core 1 and 3 structures (ST6GALNAC2) (see Fig. 1)14–16. Together, mutations in these three genes accounted for five CRC cases, with two missense and one nonsense mutation in ST6GALNAC2, two missense mutations in B3GNT2, and B4GALT2, and in the termination of O-glycan Core 1 and 3 structures (ST6GALNAC2) (see Fig. 1)14–16. Together, mutations in these three genes accounted for five CRC cases, with two missense and one nonsense mutation in ST6GALNAC2, and one missense mutation with a loss of the wild-type allele in B4GALT2 (Table 2, Supplementary Table S5). Furthermore, in silico prediction by SIFT and/or PolyPhen revealed four of the five missense mutations within these genes to significantly alter protein function (Table 2). Mutual exclusivity analysis showed that mutations affecting B3GNT2, ST6GALNAC2, B4GALT2 individually, or any of the 36 candidate glycosylation genes as a group, as not being independent of known driver oncogenic mutations in KRAS or BRAF in colon cancer (Supplementary Table 4, Supplementary Fig. 1), indicating that the glycosylation defects may play a complementary role to other mitogenic signaling pathways in the multi-step colon cancer progression model. Evaluation of independent large-scale cancer datasets17,18 revealed recurrent somatic mutations in B3GNT2, B4GALT2 and ST6GALNAC2, accounting for ~3% of CRC cases (Supplementary Table S6). These findings suggest that genetic defects in glycosyltransferases involved in the biosynthesis of Core 1–3 O-glycans potentially contribute to aberrant glycosylation and colon tumor progression.

**Results**

**Catalog of somatic mutations in glycosylation pathway genes in colon cancer.** We performed targeted re-sequencing of 430 glycosylation genes in a set of patient-derived microsatellite stable (MSS) CRC cell lines (N = 31) (Supplementary Tables S1–S3) to determine the type and extent of glycosylation pathway defects in colon cancer, and to assess for the prevalence of bi-allelic defects in these glycosylation pathway genes. Following sequential filtering and confirmation in antecedent primary colon tumors (see Methods), we identified a total of 41 non-silent mutations mapping to 36 unique genes, with the majority of mutations being missense alterations (Supplementary Table S4). Eighteen of the missense mutations were predicted to be deleterious in nature by SIFT and/or Polyphen algorithms10,11. Five mutations were highly likely deleterious in nature including, 3 nonsense mutations (ALG13, B3GNT2 and MAN2B2), a splice site mutation (ALG6), and a frame shift deletion mutation (ST8SIA3) (Supplementary Table S4). Colon cancers with mutant B4GALT2, MGAT2, or ST8SIA3 showed genomic loss of respective wild-type alleles, while colon cancers with mutant B3GALT1, GAL3ST1, GLT2SD2, or PIGO showed loss of transcript expression of respective wild-type alleles, providing evidence for bi-allelic defects in these genes in colon cancer (Supplementary Table S4).

We next determined which among the 36 candidate genes are mutated at a significantly higher rate than the expected background rate in the CRGs under study. Using the statistical framework as previously described by our group12,13, we identified 12 of the 36 genes to be significantly mutated in CRCs (P ≤ 0.01, FDR < 0.05) (Table 1, Supplementary Table S5). Interestingly, we observed three of these genes B3GNT2, ST6GALNAC2, and B4GALT2 mapping to protein glycosylation pathways that are involved in the formation of polylactosamine chain extensions on N- and O-linked glycans (B3GNT2 and B4GALT2) and in the termination of O-glycan Core 1 and 3 structures (ST6GALNAC2) (see Fig. 1)14–16. Together, mutations in these three genes accounted for five CRC cases, with two missense and one nonsense mutation in B3GNT2, two missense mutations in ST6GALNAC2, and one missense mutation with a loss of the wild-type allele in B4GALT2 (Table 2, Supplementary Table S5). Furthermore, in silico prediction by SIFT and/or PolyPhen revealed four of the five missense mutations within these genes to significantly alter protein function (Table 2). Mutual exclusivity analysis showed that mutations affecting B3GNT2, ST6GALNAC2, B4GALT2 individually, or any of the 36 candidate glycosylation genes as a group, as not being independent of known driver oncogenic mutations in KRAS or BRAF in colon cancer (Supplementary Table 4, Supplementary Fig. 1), indicating that the glycosylation defects may play a complementary role to other mitogenic signaling pathways in the multi-step colon cancer progression model. Evaluation of independent large-scale cancer datasets17,18 revealed recurrent somatic mutations in B3GNT2, B4GALT2 and ST6GALNAC2, accounting for ~3% of CRC cases (Supplementary Table S6). These findings suggest that genetic defects in glycosyltransferases involved in the biosynthesis of Core 1–3 O-glycans potentially contribute to the pathogenesis of molecular subsets of gastrointestinal cancers.

**Biochemical characterization of wild-type and mutant glycosyltransferases.** As suggested by our previous findings in GALNT129, mutations in glycosylation-associated genes could alter enzymatic activity of the encoded glycosyltransferase leading to aberrant glycosylation of protein substrates. We therefore proceeded to assess for differences in enzymatic activities of respective wild-type versus each of the mutant versions of

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Glycosylation Pathway sub-network</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3GALT1</td>
<td>Glycosphingolipid synthesis, N-glycan trimming and branching</td>
</tr>
<tr>
<td>B3GNT2</td>
<td>N-/Core 1,2,3, O-linked glycan biosynthesis, polylactosamine</td>
</tr>
<tr>
<td>B4GALT2</td>
<td>N-/Core 1,2,3, O-linked glycan biosynthesis, polylactosamine</td>
</tr>
<tr>
<td>CHST12</td>
<td>Sulfation of Chondroitin and Dermatan</td>
</tr>
<tr>
<td>CMAS</td>
<td>Nucleotide sugar biosynthesis</td>
</tr>
<tr>
<td>GAL3ST1</td>
<td>Glycosphingolipid synthesis, Isoglucoside synthesis</td>
</tr>
<tr>
<td>GAL3ST3</td>
<td>Isoglucoside synthesis</td>
</tr>
<tr>
<td>GLB1</td>
<td>Hydrolysis of Beta-galactose</td>
</tr>
<tr>
<td>HS3ST5</td>
<td>Sulfation of Heparan</td>
</tr>
<tr>
<td>LYZ</td>
<td>Lysozyme activity</td>
</tr>
<tr>
<td>ST6GALNAC2</td>
<td>Core 1,3, O-linked glycan biosynthesis</td>
</tr>
</tbody>
</table>

Table 1. Functional annotation of glycosylation pathway genes found significantly mutated in colon cancer.

protein N- and O-glycosylation are complex processes involving a multitude of enzymes, we initiated a study to characterize the extent and significance of genetic defects in the colon cancer glycome. By employing comprehensive genomic, biochemical, and functional approaches in a series of patient-derived colon cancer cell lines and matched primary tumors, we identified significant molecular and functional defects in 3 genes that likely control the biosynthesis (termination and elongation) of N- and Core 1–3 O-linked glycans expressed in the colon, thus uncovering mechanisms potentially contributing to aberrant glycosylation and colon tumor progression.
B3GNT2, B4GALT2, and ST6GALNAC2 transferases identified in this study (Table 2), using in vitro derived substrates (Supplementary Fig. S2).

B3GNT2 (β-1,3-N-Acetylglucosaminyltransferase 2) catalyzes the addition of β-3 N-acetylglucosamine onto a terminal β-4 linked galactose residue forming extended polylactosamine (polyLacNAc) chains composed of repeats of N-acetyllactosamine (β-Gal (1–4) β-GlcNAc(1–3))ₙ. PolyLacNAc chains may be found on N- and O-linked glycans, the latter potentially attached to Core 1, Core 2 and Core 3 base structures as shown in Fig. 1. As mentioned above, we identified 3 mutations in B3GNT2: R6X, P186T, and D247H (Table 2).

Given the putative deleterious nature of the R6X stop-gain mutation (Fig. 2a), we first tested if this mutation leads to nonsense-mediated decay (NMD) of the transcript. RNA expression analysis of B3GNT2 in the corresponding mutant CRC cell line however showed retention of the mutant allele (Supplementary Fig. S3), suggesting this mutation may not activate NMD but may rather encode a truncated version of the protein via utilization of an alternative downstream translation start site. Western blot analysis of ectopically expressed R6X B3GNT2 into COS7 cells indeed showed a truncated protein product, albeit expressed at a significantly lower level than wild-type B3GNT2 protein (Fig. 2b). Mass spectrometry analysis further confirmed protein translation of R6X B3GNT2, but attempts to identify the start codon in the R6X mutant product were unsuccessful (data not shown). Nevertheless, given that the Golgi-targeting signal sequence is contained within the N-terminal B3GNT2 transmembrane motif (Fig. 2a), we hypothesized that the truncated R6X protein product may be devoid of the signal sequence and therefore may not localize to Golgi. Immunofluorescence analyses of ectopically expressed wild-type and R6X B3GNT2 in COS7 cells in fact showed wild-type B3GNT2 as being exclusively localized to the Golgi, in contrast to the R6X mutant which showed aberrant and diffuse sub-cellular localization (Fig. 2b).

**Table 2. Colon cancer associated somatic mutations in B3GNT2, ST6GALNAC2 and B4GALT2 genes.**

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>CRC Sample I.D</th>
<th>Variant class</th>
<th>Protein change</th>
<th>UniProt Domain</th>
<th>SIFT/Polyphen Prediction</th>
<th>Loss of wild-type allele in cancer DNA/RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3GNT2</td>
<td>V964</td>
<td>Nonsense</td>
<td>R6X</td>
<td>Topological</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>B3GNT2</td>
<td>V915</td>
<td>Missense</td>
<td>P186T</td>
<td>Transferase</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td>B3GNT2</td>
<td>V920</td>
<td>Missense</td>
<td>D247H</td>
<td>Transferase</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td>ST6GALNAC2</td>
<td>V915</td>
<td>Missense</td>
<td>D43H</td>
<td>Stalk</td>
<td>Deleterious</td>
<td>No</td>
</tr>
<tr>
<td>ST6GALNAC2</td>
<td>V451</td>
<td>Missense</td>
<td>R115W</td>
<td>Transferase</td>
<td>Neutral</td>
<td>No</td>
</tr>
<tr>
<td>B4GALT2</td>
<td>V957</td>
<td>Missense</td>
<td>A146V</td>
<td>Transferase</td>
<td>Deleterious</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 1. Protein glycosylation pathways involving B3GNT2, B4GALT2 and ST6GALNAC2. Synthesis of polylactosamine structures (polyLacNAc) on tetraantennary N-linked glycans and on the Core 1, 2 and 3 O-glycans, catalyzed by B3GNT2 and B4GALT2 enzymes. The addition of sialic acid (NeuNAc) to the Core 1 and 3 O-glycan chains by ST6GALNAC2 results in chain termination. 

**Table 2. Colon cancer associated somatic mutations in B3GNT2, ST6GALNAC2 and B4GALT2 genes.**
Figure 2. Biochemical characterization of wild-type and mutant versions of B3GNT2. (a) Somatic mutations mapped to B3GNT2 protein coding regions (black line). Colored boxes indicate annotated protein structural domains. (TM) transmembrane motif. (b) (top) Western blot analysis showing V5-tagged wild-type and R6X mutant protein expression in COS7 cells transfected with respective cDNA constructs. Note the relatively smaller size of mutant R6X protein as compared to wild-type B3GNT2. (b) (bottom) Immunofluorescence analysis of V5-tagged wild-type and R6X mutant protein in COS7 cells transfected with respective cDNA constructs. Note the co-localization of wild-type B3GNT2 (green) with the Golgi marker Giantin (red), and the aberrant sub-cellular localization of R6X mutant (green). (c) (top) Western blot analysis showing protein expression of V5-tagged wild-type and missense mutant versions of B3GNT2 in COS7 cells transfected with respective cDNA constructs. (c) (bottom) Mean enzyme activity of wild-type, P186T, and D247H mutants assessed using Lactose-PNP and LacNAc-PNP substrates as a function of reaction time normalized to vector control. Error bars represent standard error of the means derived from three independent replicate experiments. Note the marked loss of D247H mutant enzyme activity against both the substrates, when compared to wild-type B3GNT2 (P ≤ 0.05).
Taken together, these findings suggest that the miss-localized R6X B3GNT2 mutant may potentially lack access to endogenous substrates within the Golgi, or may aberrantly glycosylate unintended substrates within the cell.

The two missense B3GNT2 mutations (P186T and D247H) mapped to the catalytic domain of the B3GNT2 protein (Table 2, Fig. 2a, Supplementary Fig. S3). We assessed the impact of these mutations on encoded enzymatic activities using two different substrates, Lactose-PNP (Lactose para-nitrophenol) and LacNAc-PNP (LacNAc para-nitrophenol), selected based on the positive activity of wild-type B3GNT2 against each of these substrates (Supplementary Fig. S2). As shown in Fig. 2c, biochemical analysis revealed that while activities of the wild-type and P186T B3GNT2 proteins were comparable, the D247H mutant exhibited no detectable enzymatic activity against either of these substrates (P < 0.05). Taken in total, these findings suggest that R6X and D247H mutations may markedly impair B3GNT2 downstream function in the cell.

B4GALTL2 (β-1,4-Galactosyltransferase 2) catalyzes the transfer of galactose to N-acetylgalactosamine residues forming the β-(Gal(1–4))3-(GlcNac)-R moiety on N- and O-linked glycans likely including the polyLacNac structure (Fig. 1). The missense mutation, A146V, maps to the transferase catalytic domain of the B4GALTL2 gene (Fig. 3a). Interestingly, this mutation was accompanied by a genomic loss of the wild-type allele in the mutant CRC cell line (Fig. 3b). We next examined the impact of A146V mutation on the encoded B4GALTL2 enzyme activity using a glucopyranoside substrate, selected based on positive activity of wild-type B4GALTL2 against this substrate (Supplementary Fig. S2). As shown in Fig. 3c, biochemical analysis revealed robust enzyme activity of the wild-type protein, with the A146V exhibiting no detectable enzymatic activity (P < 0.05).

Interestingly, we also consistently noted that the wild-type B4GALTL2 but not the A146V mutant as exhibiting a differential migratory pattern on SDS-PAGE, suggesting a potential post-translational modification of the wild-type protein (Fig. 3c). To test this, we ectopically expressed wild-type or A146V into the corresponding B4GALTL2-mutant V957 CRC cell line, and performed Western blot analyses in COS7 cells (Fig. 3c), wild-type B4GALTL2 protein exhibited a differential migratory pattern than the A146V mutant in V957 (Fig. 3d, lane 4 vs. 7 from left). Mass spectrometry analysis of respective protein bands in the wild-type and A146V mutant transfections confirmed their identity as B4GALTL2 protein (data not shown), but was however unable to resolve the specific post-translational modification of wild-type B4GALTL2. Nonetheless, given that B4GALTL2 contains three potential N-linked glycosylation sites (NXS/T) at amino acids 66, 71, and 357–364, we treated V957 cells ectopically expressing wild-type or A146V mutant proteins with either a pan N- and O-glycosidase or a specific N-linked glycosidase (PNGase F) to assess for N-linked glycosylation of wild-type versus mutant protein. While Western blot analysis showed both wild-type and mutant proteins as being predominantly N-glycosylated (i.e., similar shifts with the pan-glycosidase and PNGase F), the wild-type protein still showed a higher size-shift than the mutant suggesting additional, as yet undetermined, post-translational modification of wild-type B4GALTL2 (Fig. 3d, lanes 5 vs. 8, 9 from left). These findings, besides revealing bi-allelic defects in B4GALTL2, also suggest that mutational changes in B4GALTL2 may potentially disrupt post-translational modification of the encoded protein, resulting in impaired enzymatic activity.

ST6GALNAC2 (α-N-Acetylgalactosaminidyl α-2,6-Sialyltransferase 2) catalyzes the addition of sialic acid residues to the 6 position of the peptide linked GalNAc in the Core 1 and Core 2 O-glycan structures: (β-1–3)α-GalNAc-O-Thr/Ser and (β-1–3)α-GalNAc-O-Thr/Ser respectively (see Supplementary Fig. S2). We identified two missense mutations in the ST6GALNAC2 gene, D43H located in the stalk between the transmembrane and transferase domain, and R115W located within the transferase domain (Fig. 4a, Supplementary Fig. S4). We assessed the impact of these mutations on encoded ST6GALNAC2 enzyme activity using antifreeze glycoprotein from Antarctic fish (AFGP) and asialofetuin (ASF) substrates. AFGP consists of the (β-1–3)α-GalNAc-(GlcNAc)-R m repeat, while ASF contains multiple O-glycan structures including (β-1–3)α-GalNAc-O-Thr-Ala-Ala)n repeat, both of which display incorporation of radiolabeled NeuNAc when wild-type ST6GALNAC2 is expressed (Supplementary Fig. S2). As shown in Fig. 4b, no significant differences in enzyme activities between wild-type and mutant ST6GALNAC2 was observed against these substrates, although we consistently observed an increased enzyme activity of the D43H mutant over the wild-type sialyltransferase against the AFGP substrate in our assays, which we did not observe with the ASF substrate. The significance of this apparent gain of activity in the mutant against the homogeneous AFGP is yet to be determined.

**Phenotypic characterization of wild-type and mutant glycosyltransferases.** We next proceeded to examine the phenotypic consequences of the mutant glycosyltransferases identified in our CRC dataset (Table 2). Since aberrations in cell surface glycans have been shown to primarily affect the migratory and metastatic potential of cancer cells, we compared the effects of wild-type versus mutant genes on cancer cell migration using the widely employed SW480 CRC cell line model. Of note, SW480 parental CRC cells show retention of endogenous B3GNT2 and B4GALT2 mRNA expression with marked loss of expression of ST6GALNAC2 when compared to normal colon epithelia (Supplementary Fig. S5). SW480 cells were transiently transfected with respective wild-type or mutant versions of B3GNT2, ST6GALNAC2, and B4GALTL2, or with an empty vector control, and cell migration was assessed in a scratch wound assay over a course of 48 hours using the highly quantitative IncuCyte live cell kinetic imaging system. While wild-type B3GNT2 showed no effect on cell migration, the mutant versions of B3GNT2 however significantly enhanced the migratory potential of SW480 cells (Fig. 5, P < 0.05), suggesting a potential gain of oncogenic function of the respective B3GNT2 mutant proteins. In contrast, wild-type ST6GALNAC2 markedly suppressed CRC cell migration (Fig. 5, P < 0.05), consistent with its proposed function as a tumor suppressor in breast cancer, while the two ST6GALNAC2 mutants failed to inhibit cancer cell migration (Fig. 5) indicating potential loss of phenotypic function of the ST6GALNAC2 mutant proteins. No change in cell migration was observed in CRC cells carrying either wild-type or mutant B4GALTL2 (Fig. 5). Taken together, these findings suggest that the endogenous protein targets of B3GNT2 and ST6GALNAC2 may likely be involved in regulating cell migration.
Figure 3. Biochemical characterization of wild-type and mutant B4GALT2.  
(a) A146V somatic mutation mapped to B4GALT2 protein coding regions (black line). Colored boxes indicate annotated protein structural domains. (TM) transmembrane motif.  
(b) Sequencing chromatograms depicting A146V somatic mutation in the matched primary tumor and cell line DNA, and cell line RNA. Note the loss of wild-type B4GALT2 allele in both the DNA and RNA from the cell line.  
(c) (top) Western blot analysis showing V5-tagged wild-type and mutant B4GALT2 protein expression in COS7 cells transfected with respective cDNA constructs.  
(c) (bottom) Mean enzyme activity of wild-type and mutant B4GALT2 proteins assessed using GlcNAc-PNP substrate as a function of incubation time, normalized to vector control. Error bars represent standard error of the means derived from three independent replicate experiments. Note the significant loss of A146V mutant enzyme activity when compared to wild-type protein (P ≤ 0.05).  
(d) Protein lysates from V957 cells transiently transfected with V5-tagged empty vector, wild-type or A146V B4GALT2 were immunoprecipitated with anti-V5 agarose and treated with either a pan-glycosidase, N-linked glycosidase PNGase F or left untreated. Western blot analysis was performed using anti-V5 antibody (see Methods). Note the significant difference in protein sizes between wild-type versus mutant B4GALT2 in untreated cells (lane 4 vs. 7 from left). Although both wild-type and mutant B4GALT2 proteins appear to be N-glycosylated (lanes 5, 6 vs. 8, 9 from left), a substantial fraction of glycosidase-treated wild-type protein still showed a higher size-shift than the mutant, suggesting wild-type B4GALT2 as selectively undergoing additional post-translational modifications.
Aberrant protein glycosylation is a frequent pathological alteration associated with the onset and progression of colon cancers. Yet, the molecular mechanisms underlying aberrant glycosylation and their potential role in tumor progression remain poorly understood. Here, we performed targeted re-sequencing of 430

Figure 4. Biochemical characterization of wild-type and mutant versions of ST6GALNAC2. (a) Somatic mutations mapped to ST6GALNAC2 protein coding regions (black line). Colored boxes indicate annotated protein structural domains. (TM) transmembrane motif. (b) (top) Western blot analysis showing protein expression of V5-tagged wild-type and missense mutant versions of ST6GALNAC2 in COS7 cells transfected with respective cDNA constructs. (b) (bottom) Mean enzyme activity of wild-type and mutant ST6GALNAC2 protein assessed using antifreeze glycoprotein (AFGP) and asialofetuin (ASF) substrates at the indicated time points normalized to vector control. Error bars represent standard error of the means derived from three independent replicate experiments.

Figure 5. Effects of wild-type versus mutant enzymes on colon cancer cell migration. Migratory kinetics of SW480 CRC cells transiently expressing respective wild-type or mutant B3GNT2, ST6GALNAC2, B4GALT2 proteins were quantified using the IncuCyte scratch wound migration assay over a 48 hour time-course. Each of the mutant B3GNT2 expressing cells showed a significant increase in migratory potential when compared to the vector control (R6X, P = 0.001; P186T, P = 0.006, D247H, P = 0.047), whereas wild-type but not mutant ST6GALNAC2 significantly reduced cell migration as compared to vector control (P < 0.05) at the 48hr time-point. Error bars represent standard error of the means derived from five replicates per experimental group.

Discussion
Aberrant protein glycosylation is a frequent pathological alteration associated with the onset and progression of colon cancers. Yet, the molecular mechanisms underlying aberrant glycosylation and their potential role in tumor progression remain poorly understood. Here, we performed targeted re-sequencing of 430
glycosylation-associated genes in 31 patient-derived CRC cell lines and matched primary colon tumors to characterize the type and extent of glycosylation pathway defects in colon cancer. Of the 430 genes tested, 12 genes were significantly mutated in CRCs (Table 1). In particular, we noticed an enrichment of mutations in the poly-lactosamine and N- and O-glycosylation pathway genes, including B3GNT2, ST6GALNAC2, and B4GALT2 in CRCs (Table 1, Fig. 1). Together, mutations in these genes were detected in 5 of the 31 CRC cases tested, with 3 mutations in B3GNT2 (R6X, P186T, D247H), 2 mutations in ST6GALNAC2 (D43H, R115W), and 1 mutation in B4GALT2 (A146V) accompanied by a loss of the wild-type allele (Table 2, Fig. 3b). Additional evaluation of independent large-scale cancer datasets17,18 revealed recurrent somatic mutations in B3GNT2, B4GALT2 and ST6GALNAC2, accounting for ~3% of CRC cases (Supplementary Table S6).

Functionally, a B3GNT2 and to a lesser extent B4GALT2 are involved in the synthesis of polyLacNAc chains on N-linked tetrantenary structures and on Core 1, 2 and 3 O-glycan core structures (Fig. 1)15,16,20. ST6GALNAC2 on the other hand adds a NeuNAc to the 6 position of the peptide GalNAc of O-glycan Core 1 or 3 structures thus terminating chain elongation26,33. Although polyLacNAc biosynthesis and O-glycan core termination are independent processes they nevertheless may be linked, as the Core 1, 2 and 3 O-glycans may be further elongated with polyLacNAc chains (Fig. 1)15,16.

The O-glycosylation pathway is fundamental to several critical processes in the cell and aberrations in the O-glycosylation pathway are known to be associated with both early as well as later stages of cancer progression15,34. In particular, Core 3 O-glycans have been implicated in the maintenance of intestinal homeostasis28,29,35. In fact, Core 3 O-glycans are primarily expressed in gastrointestinal mucosa, and are the major core structures of mucin-type glycoproteins in colon mucosal tissue36–39. Existing evidence also suggests that aberrations in Core 3 O-glycans likely play a key role in CRC development. For example, deregulated expression of Core 3 structures is frequently observed in colon cancers40,41. In particular, reduced expression of Core 3 synthase, an enzyme involved in the initial step of Core 3 biosynthesis, has been observed frequently in colon, gastric, and pancreatic ductal adenocarcinomas, with loss of Core 3 synthase expression highly correlating with the grade of colon neoplasia in familial adenomatous polyposis patients28,29. Furthermore, loss of activity of Core 3 synthase has been shown to enhance the metastatic potential of colon carcinoma cells29, and mice deficient in Core 3 synthase display reduced production of colonic MUC2 protein and show increased susceptibility to colitis and colon adenocarcinoma42,43. These findings, together with our observation of a significant enrichment of CRC-associated mutations in genes likely involved in Core 3 termination or Core 3 polyLacNAc elongation (Table 1, Fig. 1), strongly suggest that aberrations in these glycosyltransferases play an important role in CRC progression. We therefore proceeded to systematically characterize the functional consequences of each of the CRC-associated mutant glycosyltransferases identified in this study (Table 2).

As mentioned above, we detected 3 somatic mutations in B3GNT2 (R6X, P186T, D247H). The R6X mutation, despite being a nonsense variant, encoded an N-terminal truncated protein (Fig. 2, Supplementary Fig. S3). Importantly, as opposed to the Golgi-specific localization of wild-type B3GNT2, the R6X mutant exhibited aberrant and diffuse sub-cellular localization (Fig. 2), suggesting that the mis-localized R6X mutant may be unable to access its endogenous substrates in the Golgi besides also potentially altering the glycosylation patterns of unintended substrates within the cell. Next, biochemical analyses of the missense mutants (P186T, D247H) using two in vitro derived B3GNT2 substrates (LacNAc-PNP and Lactose-PNP) showed a loss of enzymatic activity of the D247H mutant against these selected substrates (Fig. 2). Intriguingly, phenotypic analyses showed all three B3GNT2 mutants as significantly enhancing the migratory potential of colon adenocarcinoma cells (Fig. 5), indicating a gain of oncogenic function likely resulting from dominant negative activities of the mutant enzymes against wild-type B3GNT2 and/or other glycosyltransferases. Moreover, since B3GNT2 is involved in the synthesis of polyLacNAc chains, genetic defects in B3GNT2 could lead to aberrations in cell surface polyacatosamines, critical signaling molecules that are often implicated in tumor cell migration and possibly metastasis44–46. Further studies to identify the actual endogenous substrates of B3GNT2 would help delineate the role of this transferase in the pathogenesis of CRCs.

The missense mutation (A146V) in B4GALT2 was accompanied by a genomic loss of the wild-type allele in the corresponding CRC sample (Fig. 3). Biochemical analyses using an in vitro derived B4GALT2 substrate, GlcNAc-PNP, showed loss of enzymatic activity of the A146V mutant (Fig. 3). In addition, de-glycosylation studies revealed A146V mutation as markedly affecting post-translational modification of the B4GALT2 protein (Fig. 3), which in turn could have a negative impact on its enzymatic activity. Phenotypic analysis however showed neither wild-type nor mutant B4GALT2 as affecting the migratory potential of CRC cells (Fig. 5), suggesting that the endogenous targets of B4GALT2 may not likely be involved in regulating cell motility or that they require tissue microenvironment for functioning.

The two missense mutations detected in ST6GALNAC2 (D43H and R115W) showed no apparent loss of enzymatic activities when tested against the AFGP and ASF substrates, although ST6GALNAC2 D43H mutant consistently showed an increase in enzyme activity over the wild-type transverse against the AFGP substrate in our assays (Fig. 4). This finding of enhanced D43H enzyme activity appears to be in keeping with prior studies where sialylation is increased while Core 3 structures decreased in CRC37. Interestingly, phenotypic analyses demonstrated wild-type ST6GALNAC2, but not the mutants, as markedly impeding the migratory potential of colon carcinoma cells (Fig. 5). It is likely that the wild-type and mutant ST6GALNAC2 proteins may exhibit differential specificities/affinities towards actual endogenous protein targets involved in regulating cell migration in vivo. Nonetheless, our phenotypic findings are consistent with the reported metastasis suppressor role of ST6GALNAC2 in breast cancer44 and further indicate a loss of phenotypic function of ST6GALNAC2 mutants identified in CRCs. In summary, we have comprehensively characterized the mutational landscapes of glycosylation-associated genes in colon cancer, identifying three glycosyltransferases as significant mutational targets in CRCs. Functional studies demonstrate these mutant glycosyltransferases as having a significant impact on the encoded enzymatic activity and/or the migratory potential of colon carcinoma cells. Although our study may not fully capture the
functional complexities and kinetics of N- or O-linked glycosylation, the finding of functionally deleterious CRC mutations in genes that are likely fundamental to maintaining intestinal homeostasis, suggests that genetic defects in polylactosamine and Core 1 and 3 O-glycosylation pathway potentially contribute to CRC pathogenesis. Of note, given our prior studies identifying mutations in GALNT2 gene as being associated with susceptibility to familial colon neoplasia, future studies can be designed to explore whether genetic defects in Core 1/3 glycosylation pathway also play a role in susceptibility to unexplained inherited forms of colon cancer. Further characterization of the actual endogenous substrates of these glycosyltransferases and evaluation of phenotypic consequences of these mutant glycosyltransferases in pre-clinical animal models should provide additional insights into the biologic role of these genes in colon cancer progression.

Materials and Methods

Detailed methods are provided in Supplementary Methods section.

Patient samples and nucleic acid extraction. Patient-derived VACO series of colon cancer cell lines were propagated as previously described. Colon tumor and normal tissue specimens matched to respective VACO cell lines were obtained from a formalin-fixed paraffin embedded (FFPE) archive that were collected under an Institutional Review Board (IRB) approved protocol at the Case Medical Center. All participants provided written informed consent prior to participating in the study and all methods were carried out in accordance with the approved guidelines. Genomic DNA from the cell lines and FFPE tissues was extracted as previously described. Demographics of DNA samples used for the study are provided in Supplementary Table S1.

Targeted re-sequencing of glycome pathway genes. A custom Agilent SureSelect XT array (Agilent Technologies, Inc. Santa Clara, CA) was designed to capture and sequence the coding and splice site regions of 430 candidate glycosylation pathway genes (Supplementary Table S2) in a series of 31 patient-derived VACO CRC cell lines (Supplementary Table S1).

Somatic mutation detection. Burrows-Wheeler Aligner was used to align the raw FASTQ files to the human reference genome (build hg19). Sample coverage metrics are provided in Supplementary Table S3. Nucleotide variations were detected using SOAPsnp, Genome Analysis Toolkit and mPILEUP. Somatic mutations were identified using a series of variant-filtering steps, and were confirmed by Sanger sequencing in both cell lines and matched primary colon tumor tissues. All together, 41 somatic protein-altering mutations in 36 genes were identified amongst the 31 CRC cases tested (Supplementary Table S4).

Significantly mutated genes and selection of gene candidates for functional studies. Significantly mutated genes were identified using the statistical framework previously described by our group. Twelve candidate genes showed a significantly higher mutation rate than the background (Supplementary Table S5). Three of these were identified as genes involved in the polylactosamine chain extension on N- and Core 1–3 O-linked glycans or in chain termination of Core 1/3 O-glycans, and were selected for further functional characterization (Supplementary Table S5).

Sanger sequencing. Custom PCR primers flanking respective mutant loci in candidate glycosyltransferase genes were designed for Sanger sequencing (Supplementary Table S7).

Pyrosequencing to test for KRAS/BRAF hotspot mutations. Pyrosequencing assays were designed using the PSQ Assay Design software (QIAGEN, Chatsworth, CA) to test for hotspot mutations in KRAS (codons 12, 13, 61, and 146) and BRAF (codon 600). For each assay, one of the PCR primers was biotinylated at the 5′ end and purified using high performance liquid chromatography. All PCR reactions were performed using FastStart Taq (Roche). Following PCR, amplification products were sequenced on a PyroMark MD pyrosequencing instrument (QIAGEN) and mutation analysis was conducted as previously described. Sanger sequencing was used to confirm all mutations detected by pyrosequencing analysis.

Mutual Exclusivity Evaluation. To test if mutations affecting the glycosylation genes occur in a mutually exclusive fashion with respect to other known oncogenic driver mutations in CRCs (KRAS and BRAF), we applied CoMEt, which employs an exact statistical test for mutual exclusivity that has been shown to be more sensitive in detecting mutually exclusive events within combinations containing rare alterations.

Generation of expression constructs and recombinant protein purification. Full length cDNA fragments, encoding wild-type (WT) or mutant B3GNT2, ST6GALNAC2 and B4GALT2 transcripts were PCR amplified from total RNA derived from a reference normal colon sample or from corresponding mutant CRC cell lines, respectively, and cloned into pcDNA3.1 or pHHV vectors. Transfection was performed in COS7 cells using Lipofectamine 2000 (Life Technologies, Carlsbad, CA); recombinant proteins were isolated using immunoprecipitation with anti-V5 antibody.

Western blot analysis. 1/10th of the immunoprecipitated recombinant protein was subjected to SDS/PAGE analysis and immunoblotted with mouse anti-V5 antibody.

Enzyme assays. The donor and acceptor substrates used for assaying wild-type B3GNT2, B4GALT2, and ST6GALNAC2 enzyme activities are given in Supplementary Fig. S2A. Briefly, reaction mix containing 100–150 μl of immunoprecipitated proteins were subjected to either reverse-phase chromatography (B3GNT2, B4GALT2)
or dialysis (ST6GALNAC2) to measure the incorporation of radio-labelled sugars (UDP-[3H]-GlcNAc, UDP-[3H]-Gal, CMP-[3H]-NeuAc) by respective wild-type and mutant enzymes over a period of 24 hrs.

Mass spectrophotometry. COS7 cells transfected with empty vector, wild-type or mutant expression constructs of B3GNT2 or B4GALT2 were immunoprecipitated with anti-V5 antibody and subjected to SDS-PAGE. Relevant Coomasie G250 stained protein bands were excised for subsequent Mass spectrometry analysis using liquid chromatography-tandem mass spectrometry (LC-MS/MS).

**In vitro glycosidase assay.** V957 CRC cells were transfected with pcDNA3.1/V5- His/empty vector, or B4GALT2 wild-type or B4GALT2 A146V mutant constructs. Immunoprecipitated wild-type and mutant B4GALT2 protein were treated with either Peptide N-glycosidase F or a pan glycosidase protein deglycosylation mix or left untreated at 37°C for 4 hours followed by Western blot analysis using anti-V5 antibody.

**Confocal Imaging.** COS7 cells transfected with V5-tagged pcDNA3.1 empty vector or B3GNT2 wild-type or B3GNT2 R6X mutant were immunostained with anti-V5 antibody, anti-Giantin antibody and DRAQ5 (nuclear counterstain). Immunostained cells were visualized using the Zeiss LSM 510 confocal microscope.

**Scratch wound cell migration assay.** Scratch wound assay was performed in SW480 cells, transfected with either pcDNA3.1/ pIHV empty vector, respective wild-type or mutant constructs of B3GNT2, B4GALT2, and ST6GALNAC2, using the automated IncuCyte ZOOM live cell kinetic imaging system (Essen BioScience, Ann Arbor, MI) as per the manufacturer’s instructions over a period of 48hrs.

**Statistical analyses.** Significant differences in enzyme activities and cell migration between wild-type and mutant proteins were estimated using a Student’s t-test; a P value < 0.05 was considered statistically significant.

**References**

Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer.

How to cite this article

Biochemistry and functional characterization of glycosylation-associated mutational landscapes in colon cancer.


Inherited deleterious variants in GALNT12 are associated with CRC susceptibility.

Inherited deleterious variants in GALNT12 are associated with CRC susceptibility.

Inherited deleterious variants in GALNT12 are associated with CRC susceptibility.

Inherited deleterious variants in GALNT12 are associated with CRC susceptibility.

The authors declare no competing financial interests.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Supplementary Methods and Figures

Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer

Srividya Venkitachalam, Leslie Revoredo, Vinay Varadan, Ryan Fecteau, Lakshmeswari Ravi, James Lutterbaugh, Sanford D. Markowitz, Joseph E. Willis, Thomas A. Gerken and Kishore Guda

SUPPLEMENTARY METHODS

CRC cell lines, patient samples, and nucleic acid extraction

The patient-derived VACO series of colon cancer cell lines were propagated as previously described \(^1,2\). Colon tumor and normal tissue specimens matched to respective VACO cell lines were obtained from a formalin-fixed paraffin embedded (FFPE) archive that were collected under an Institutional Review Board (IRB) approved protocol at the Case Medical Center. Genomic DNA from the cell lines and FFPE tissues was extracted as previously described \(^3,4\). DNA from the CRC cell lines was tested for microsatellite instability at microsatellite markers: BAT26, BAT40, D2S123, D5S346, and D17S250 \(^5\). Only microsatellite stable (MSS) cell lines were included in this study. Total RNA from the cell lines was extracted using the mirVana Kit (Ambion, Austin, TX). The tumor stage, gender, race, MSI status, and tissue source of DNA samples used for the study are provided in Supplementary Table S1.

Targeted re-sequencing of glycome pathway genes

A total of 430 candidate genes, belonging to known protein glycosylation pathways, were identified (Supplementary Table S2) by reviewing published literature \(^6-8\) and annotated databases, including the Consortium for Functional Glycomics (http://www.functionalglycomics.org/fg/). A custom Agilent SureSelect XT array was designed
to capture and sequence the coding and splice site regions of the 430 genes (Agilent Technologies, Inc. Santa Clara, CA) in a series of 31 MSS VACO CRC cell lines (Supplementary Table S1). Sample library generation was carried out by the Gene Expression and Genotyping Core Facility at CWRU cancer center as per manufacturer’s instructions. Briefly, sample DNA’s were quantified using a picogreen fluorometric assay and 3µg of genomic DNA were randomly sheared to obtain a base pair peak of 150 bp using a Covaris S2 sonicator (Covaris Inc, Woburn, MA). Sonicated DNA was then end-repaired, A-tailed, and ligated with indexing-specific adapters. Each sample was captured using the custom capture array kit followed by library amplification and quality assessment using Agilent 2100 Bioanalyzer High Sensitivity DNA Chip (Agilent). The captured libraries were precisely quantified using a qPCR-based Kapa Biosystems library quantification kit (Kapa Biosystems Inc, Woburn, MA) on a Roche Lightcycler 480 (Roche Applied Science, Indianapolis, IN). Amplified libraries were pooled in batches of 10 or 11 samples per pool and sequenced using the Illumina HiScan instrument (Illumina Inc, San Diego, CA) to generate 100bp single-end reads.

**Read mapping and annotation**

Burrows-Wheeler Aligner (BWA) ⁹ was used to align individual 100bp reads from the raw FASTQ files to the human reference genome (build hg19) with default parameters. Following the conversion of aligned reads in to binary Sequence Alignment/Map (BAM) format, coverage metrics of target bases were calculated using the Picard algorithm (http://samtools.sourceforge.net). Sample coverage metrics provided in Supplementary Table S3 show 94.31% of the target bases as being covered at 30X average read-depth, with 1.48% of target bases showing no coverage. Next, nucleotide variations (SNV, splice site, and
insertion/deletion) in respective samples were detected using three variant calling algorithms including, SOAPsnp\textsuperscript{10}, Genome Analysis Toolkit (GATK)\textsuperscript{11}, and mPILEUP\textsuperscript{12}, set to default parameters, and with the human reference sequence as the background. Genomic variants were mapped to the human transcriptome reference database (RefSeq, build hg19) using a variant annotation tool developed in house (SLATE), which identifies variants mapping to gene coding regions and splice-sites, including their corresponding positions and codon changes within respective transcripts.

**Filtering of variants and detection of somatic mutations**

Since targeted re-sequencing was performed only on the CRC cell lines, a series of variant filtering steps detailed below were performed to identify putative somatic mutations. First, all the variants identified in the cell lines were queried against the 1000 genome database\textsuperscript{13} as well as against an in-house variant database generated from platform-matched whole exome sequencing of more than 150 germline samples, to eliminate variants that are likely germline and not somatic. The use of platform-matched in-house database additionally aided in eliminating recurrent artifacts or false positives seen in the deep sequence data. Second, direct Sanger sequencing was performed on all the remaining putative somatic variants in cell lines, and any variants that were not confirmed by Sanger were eliminated. Third, each of the variants were Sanger sequenced in respective matched FFPE normal tissue to ascertain the somatic nature of the variants. Fourth, candidate somatic mutations were then Sanger sequenced in respective matched FFPE primary tumor tissue, and only mutations that were also present in the antecedent tumor tissue were included for further analysis. In the end, 41 somatic protein-altering mutations in 36 genes were identified amongst the 31 CRC cases (Supplementary Table S4).
Significantly mutated genes and selection of gene candidates for functional studies

In order to identify genes that were mutated at a significantly higher rate than the background, we followed a statistical framework similar to our published studies detailing mutational profiles of colon cancer. Briefly, to account for nucleotide composition influence on the likelihood of mutations, base coverage metrics for the six different nucleotide contexts including AT transitions, AT transversions, CG transitions, CG transversions, CpG transitions and CpG transversions were first obtained for each of the 36 genes (Supplementary Table S3) in the entire sample cohort using the Genome MuSiC suite. The background mutation rate for each of the above six mutation contexts and indels in colon cancers was derived from an in-house database of whole exome sequence data derived from 30 late-stage colon cancers (Supplementary Table S5). Next, for a given gene, we counted the number of mutations in each of the above seven categories (indels plus mutations in six different sequence contexts) and calculated the probability of the observed number of mutations in a particular category using an exact binomial distribution. The total probability (combined P value) of a gene exhibiting the observed number of mutations in all of the seven categories was then calculated to be the product of the seven context-specific probabilities. Finally, to correct these probabilities for multiple comparisons, we used the algorithm described by Benjamini and Hochberg. Overall, 12 of 36 candidate genes showed a significantly higher mutation rate than the background (P≤0.01, FDR <0.05) (Supplementary Table S5). 3 of these were identified as genes involved in the polylactosamine chain extension on N- and Core 1-3 O-linked glycans or in chain termination of Core 1/3 O-glycans, and were selected for further functional characterization (Supplementary Table S5).
Sanger sequencing

For DNA, gene-specific M13-tagged PCR primers flanking respective mutant loci were custom designed (Supplementary Table S7). PCR conditions included 95°C for 4 min, 35 cycles of 95°C for 45s, 62.3°C for 30s and 72°C for 45s. Each reaction was carried out in a 50µl reaction volume using 2U of Fast-TAQ DNA polymerase (Roche Applied Science, Indianapolis, IN) with 10-20ng of cell line, primary tumor or matched normal DNA. The PCR products were purified, and sequenced directly or sequenced following sub-cloning into pCR2.1-TOPO vector using universal M13 forward and reverse primers. For RNA, 2 µg DNase-treated RNA was reverse transcribed using SuperScript III First-Strand Synthesis (Life technologies, Carlsbad, CA) according to standard protocol to obtain cDNA. cDNA samples were PCR amplified using above conditions with custom M13-tagged primers (Supplementary Table S7), and resulting PCR products were purified and directly sequenced using universal M13 forward and reverse primers. Analysis of Sanger sequencing data was performed using Mutation Surveyor software package (SoftGenetics, State College, PA).

Pyrosequencing to test for KRAS/BRAF hotspot mutations

Pyrosequencing assays were designed using the PSQ Assay Design software (QIAGEN, Chatsworth, CA) to test for hotspot mutations in KRAS (codons 12, 13, 61, and 146) and BRAF (codon 600). For each assay, one of the PCR primers was biotinylated at the 5’ end and purified using high performance liquid chromatography. Primer sequences are as follows. KRAS codons 12 and 13: For 5’- TCGATGGAGGAGTTTGTAAATGA-3’, Rev 5’- biotin-TTCGTCCACAAAATGATTCTGA-3’, Seq 5’-CTTGTGGTAGTTGGAGC-3’; KRAS codon
61: For 5′- CAGACTGTGTTTCTCCCTTCTCA-3′, Rev 5′- biotin-
TCCTCATGTACTGGTCCCTCATTG-3′, Seq 5′- ATATTCTCGACACACGAG-3′; KRAS
codon 146: For 5′-AGGCTCAGGACTTAGCAAGAAGTT-3′, Rev 5′-biotin-
GCCCCTCTCAAGACAAAAACAT-3′, Seq 5′-AATTCCTTTTTATTGAAACAT-3′. BRAF
codon 600: For 5′- TTCATGAAGACCTACAGTAAAA-3′, Rev 5′- biotin-
CCACAAAATGGATCCAGACA-3′, Seq 5′- TGATTGTGCTAGCTACA-3′. All PCR
reactions were performed using FastStart Taq (Roche) and primer concentrations of 0.2 uM.
Cycling conditions included an initial denaturation step at 95 C for 4 min, and 49 cycles of 95 for
15 s, 54 C for 30 s, and 72 C for 20 s. Following PCR, amplification products were sequenced on
a PyroMark MD pyrosequencing instrument (QIAGEN) and mutation analysis was conducted as
previously described.

Sanger sequencing was used to confirm all mutations detected by pyrosequencing analysis. Isolated
genomic DNA from tumor samples was used for PCR amplification of regions
encompassing codons 12, 13, 61, and 146 of KRAS, and codon 600 of BRAF. Forward and
reverse primers used for PCR amplification were tagged with a 5′ M13 forward (5′-
GTAAAACGACGGCCAGT-3′) and 5′ M13 reverse (5′-CAGGAAACAGCTATGAC-3′)
universal primer sequence, respectively. Primer sequences were as follows: KRAS codons 12 and
13: For 5′-TGGTGGAGTATTTGATAGTGTA-3′, Rev 5′-
CATGAAAATGGTGAGACAGAA-3′; KRAS codon 61: For 5′- TCCAGACTGTGTTTCTCCCT-
3′, Rev 5′- AACCCACCTATAATGGTGAATATCT-3′; KRAS codon 146: For 5′-
AGAAGCAATGCCCCTCTCAAG-3′, Rev 5′-GGACTCTGAAGATGTACCTATGGTC-
3′ BRAF codon 600: For 5′- TCATAATGCTTGCTCGTAGAGA-3′, Rev 5′-
GGCCAAAATTTAATCAGTGGGA-3′. All reactions were carried out using 0.4 uM
concentration of each primer and FastStart Taq polymerase (Roche, Indianapolis, IN). Cycling conditions for all primer pairs consisted of an initial denaturation at 95 C for 4 min followed by 39 cycles of 95 C for 30 s, 58 C for 30 s, 72 C for 30 s, and a final elongation at 72 C for 3 min.

**Mutual Exclusivity Evaluation**

To test if mutations affecting the glycosylation genes occur in a mutually exclusive fashion with respect to other known oncogenic driver mutations in CRCs (KRAS, BRAF), we applied CoMET\textsuperscript{17}, which employs an exact statistical test for mutual exclusivity that has been shown to be more sensitive in detecting mutually exclusive events within combinations containing rare alterations.

**Generation of expression constructs encoding wild-type or mutant versions of candidate genes**

Full length cDNA fragments, without the C-terminal stop codon, encoding wild-type (WT) human B3GNT2, ST6GALNAC2 and B4GALT2 transcripts were PCR amplified from total RNA isolated from a reference normal colon sample. cDNA fragments encoding respective mutant alleles were generated either by direct PCR amplification of cDNA from respective colon cancer samples, or by site directed mutagenesis (QuikChange Lightning, Agilent Technologies, Santa Clara, CA) of the wild-type allele. For B3GNT2 and B4GALT2, full-length PCR products were cloned using the TA-cloning method into the \textit{pcDNA3.1/V5-His} TOPO cloning vector (Life technologies), in frame with C-terminal V5 and His6 epitope tags. For ST6GALNAC2, PCR products corresponding to aa 29-374 were cloned into a modified SV40 promoter-driven \textit{pZeoSV2} vector (\textit{pIHV}) (Life technologies) that contains an insulin secretion signal to direct the secretion of the recombinant protein into the cell culture medium, and an N-terminal His6 and
V5 epitope tags to facilitate purification and detection of the recombinant protein. All constructs were verified by Sanger sequencing.

**Plasmid DNA transfection**

Transfection of plasmid DNA was performed using Lipofectamine 2000 transfection reagent (Life Technologies) according to recommended protocol. Briefly, 10^6 COS7 cells (American Type Culture Collection) grown in DMEM (Life Technologies) were plated in 100mm dish for 24hr before transfection, and incubated in 5% CO2 at 37 °C overnight. 4µg of plasmid DNA in 10µl of lipofectamine 2000 reagent was used per 100-mm dish.

**Recombinant protein purification**

COS7 cells were transfected with respective wild-type or mutant expression constructs. For pcDNA3.1 constructs, the cell monolayers were washed twice with ice-cold PBS and incubated with lysis buffer (50mMTris, pH 7.5/150mMNaCl/1mM CaCl2/1mM MnCl2/EDTA-free protease inhibitor pellets/ 0.3% CHAPS) for 15 min on ice. After scraping, the lysates were clarified by centrifugation for 15 min at maximal speed. The recombinant protein was immunoprecipitated from the lysates using anti-V5 agarose beads (Sigma-Aldrich, St Louis, MO), and subsequently washed with wash buffer (50mMTris, pH 7.5/150mMNaCl/1mM CaCl2/1mM MnCl2/EDTA-free protease inhibitor pellets). For pIHV constructs, conditioned medium was collected 48hr post-transfection, and recombinant protein was immunoprecipitated from 9 mL of active culture medium using anti-V5 agarose beads (Sigma-Aldrich). The beads were subsequently processed with wash buffer as described above.
**Western blot analysis**

After immunoprecipitation, 1/10 fraction of the recombinant protein was mixed with equal volume of Laemmlsi sample buffer (Bio-Rad, Hercules, CA) at 95 °C for 5 min, and loaded onto a Bis-Tris SDS/4–12% polyacrylamide gel (Life technologies). After SDS/PAGE, proteins were transferred onto Immobilon-P PVDF membranes (EMD Millipore, Billerica, MA). Membranes were blocked for 1hr with 5% nonfat milk, and incubated with appropriate dilution of mouse anti-V5 antibody conjugated to horseradish peroxidase (Life technologies) to detect the V5-tagged proteins for both pcDNA3.1 and pIHV constructs. Enhanced Chemiluminescence Plus (GE Healthcare-BioSciences, Pittsburg, PA) and ImageJ software (National Institutes of Health, Bethesda, MA)\(^{19}\) were used to detect and quantitate respective protein bands.

**B3GNT2 enzyme assay**

For a summary of the activities, including donor and acceptor substrates and expected products for the transferases characterized in this work, please see Supplementary Figure S2A. B3GNT2-bound beads (100-150µl settled volume) were added to 250µl of B3GNT2 reaction buffer which contained 150mM MES buffer pH 7.5, 10mM MnCl\(_2\), 1.8mM UDP-GlcNAc (containing total of 100µCi of UDP-[\(^{3}H\)]-GlcNAc per reaction), protease inhibitors (P8340 and P8849, Sigma-Aldrich) and 0.5mM LacNAc-PNP substrate (p-nitrophenyl 2-Acetamido-2-deoxy-4-O-(\(\beta\)-D-galactopyranosyl)-\(\beta\)-D-glucopyranoside) (Toronto Research Chemicals, Toronto, ON) or 0.5mM Lactose-PNP substrate (4-nitrophenyl-\(\beta\)-D-lactopyranoside) (Carbosynth Limited, West Berkshire, UK)\(^{20,21}\). Reagent reaction mixtures were shaken at 37°C in a thermostated microplate shaker (Taitec Microincubator M-36) to maintain a suspension of beads. Aliquots of suspended beads were removed and quenched with an equal volume of 250mM EDTA and frozen for
processing. After dilution an aliquot was removed for measuring radioactivity on a Beckman LS5801 scintillation counter. The remainder of the sample (2ml) was subjected to reverse-phase chromatography on short C18 Sep-Pak columns (Waters, Milford, MA) \(^{22}\). Samples were eluted using 100\% methanol and measured for radioactivity. Wild-type and mutant B3GNT2 transferase-specific activities were expressed as the ratio of total post- to total pre- Sep-Pak column DPM, and further normalized to protein levels in the corresponding Western blot. Plots demonstrating wild-type protein activity in comparison to vector control for both substrates are given in Supplementary Figure S2. Wild-type and mutant transferase assays were performed together on the same day of transferase isolation using identical donor and acceptor concentrations.

**B4GALT2 enzyme assay**

B4GALT2-bound beads (100-150\(\mu\)l settled volume) were incubated with ~250\(\mu\)l of B4GALT2 reaction buffer which contained 25mM Tris Base buffer pH 7.4, 0.2\% Triton X-100, 10mM MnCl\(_2\), 2mM UDP-Gal (containing a total of 100\(\mu\)Ci of UDP-[\(^{3}H\)]-Gal), protease inhibitors (P8340 and P8849, Sigma-Aldrich) and 3mM GlcNAc-PNP (4-nitrophenyl-2-acetamido-2-deoxy-\(\beta\)-D-glucopyranoside) substrate (Carbosynth Limited) \(^{21,23}\). Transferase reactions and processing were performed as described for B3GNT2 above. Wild-type activity as compared to vector control is shown in Supplementary Figure S2. Wild-type and mutants transferase reactions were performed together on the same day of isolation using the same donor and acceptor concentrations.
ST6GALNAC2 enzyme assay

ST6GALNAC2-bound beads (100-150µl settled volume) were added to ~500µl reaction mixtures contained 50mM MES buffer pH 6.0, 10mM MgCl₂, 2mM CaCl₂, 2mM CMP-Sialic acid (CMP-NeuNAc) (containing a total of 100µCi of CMP-[³H]-NeuNAc per reaction), 0.2% Sodium azide, protease inhibitors (P8340 and P8849, Sigma-Aldrich) and 5 mg/mL antartic fish antifreeze glycoprotein (AFGP) (gift of Arthur L. DeVries, Uni. Illinois at Urbana-Champaign) or 5mg/ml asialofetuin (ASF) (Sigma-Aldrich). Reaction mixtures were shaken at 35°C as described above; aliquots of suspended beads were removed and quenched with an equal volume of 250mM EDTA. After a 10-fold dilution, aliquots were removed for measuring radioactivity and the remainder of the sample was dialyzed against MW3500 dialysis tubing (Thermo Fisher, Snakeskin) or 3 ml G2 dialysis cassettes (Thermo Fisher, Slide-A-Lyzer) against 2L of distilled water for 72 hours at 4°C with continuous stirring. After dialysis, samples were lyophilized and reconstituted in 1 ml of water. Aliquots were removed for determining incorporation of 3H-NeuNAc into glycopeptide substrates. Relative transferase-specific activities were expressed as a ratio of post-dialysis counts to glycoprotein absorbance at 280 or 220nm and further normalized to transferase protein levels from Western blotting. Supplementary Figure S2 demonstrates positive wild-type activity in comparison to vector control for both substrates. Wild-type and mutant transferase assays were performed together on the same day of transferase isolation.

Mass spectrophotometry

COS7 cells were transfected with empty vector, wild-type or mutant expression constructs of B3GNT2 or B4GALT2. After 48 hours, cells were harvested and lysed with RIPA lysis buffer.
Isolated protein was subjected to immunoprecipitation at 4°C overnight using anti-V5 agarose beads. Samples were subsequently washed with wash buffer (50mMTris, pH 7.5/150mMNaCl/1mMCaCl2/1mM MnCl2/EDTA-free protease inhibitor pellets), mixed with 30μl of Laemmli sample buffer (Bio-Rad, Hercules, CA) at 95 °C for 5 min, and loaded onto a Bis-Tris SDS/10% polyacrylamide gel (Life technologies). Gels were stained with Coomassie G250 dye for 1hr followed by destaining with destaining solution (40% acetic acid, 10% methanol), overnight at room temperature. Relevant protein bands were excised for proteomic analysis. Mass spectrometry was carried out using liquid chromatography-tandem mass spectrometry (LC-MS/MS) at the Center for Proteomics and Bioinformatics, Case Western Reserve University.

**In vitro glycosidase assay**

V957 colon cancer cells were plated on collagen in 6-well plates. Cells were transfected the following day with 2μg of pcDNA3.1/V5-His/ empty vector, wild-type or A146V B4GALT2 using Lipofectamine 2000. After 48 hours, cells were harvested and lysed with RIPA lysis buffer. Isolated protein was subjected to immunoprecipitation at 4°C overnight using anti-V5 agarose beads. Subsequently, beads were washed with a wash buffer (50mMTris, pH 7.5/150mMNaCl/1mMCaCl2/1mM MnCl2/EDTA-free protease inhibitor pellets) and equal volumes of the beads from each sample were treated with either Peptide N-glycosidase F (PNGaseF; removes high mannose, hybrid and complex N-glycans) or a pan glycosidase protein deglycosylation mix (contains O-glycosidase, PNGase F, Neuraminidase, β1-4 Galactosidase and β-N-Acetylglucosaminidase that removes short O-glycans and most N-glycans) (New England Biolabs, Ipswich, MA) or left untreated at 37°C for 4 hours. Finally, the reaction mix
was combined with 6X Laemmli sample buffer and Western blot analysis was performed using anti-V5 antibody to detect wild-type and mutant B4GALT2 proteins.

**Confocal Imaging**

COS7 cells were cultured on the surface of 12mm coverslips in 6-well plates to 90% confluency. Cells were transfected with 2μg of pcDNA3.1/V5-His/B3GNT2 using Lipofectamine 2000. After 48hr, cells were fixed with 4% paraformaldehyde (Electron microscopy sciences, Hatfield, PA) at 4°C for 15 min and permeabilized in 0.2% Triton X-100 for 5 min. Samples were incubated in Image-IT™ FX signal enhancer (Life Technologies) for 30min and blocked in 10% goat serum (Life Technologies) for 15 min. Immunostaining was performed with anti-V5 antibody (Sigma-Aldrich, diluted 1:500 in blocking buffer) and anti-Giantin antibody (Abcam, Cambridge, MA, diluted 1:200 in blocking buffer) at room temperature for 1hr. Samples were incubated at room temperature for 1hr with AlexaFluor 488-conjugated goat anti-mouse (Life Technologies) and AlexaFluor 594-conjugated goat anti-rabbit (Cell Signaling, Danvers, MA) secondary antibodies, diluted 1:800 in blocking buffer. Nuclei were counterstained with DRAQ5 (Cell Signaling, diluted 1:1000 in 1XPBS) and stained cells were mounted using DABCO anti-fade mounting medium (Sigma-Aldrich). Cells and immunostaining was visualized with the appropriate filters using the Zeiss LSM 510 confocal microscope and image browser.

**Scratch wound cell migration assay**

Briefly, 4x10^4 SW480 cells (American Type Culture Collection), grown in MEM media (Life technologies) with additional supplements, were plated in each well of an ImageLock 96-well plate (Essen Bioscience, Ann Arbor, MI). After 24hr, cells were transfected with either an
empty vector, respective wild-type or mutant constructs of B3GNT2, ST6GALNAC2 and B4GALT2 using Lipofectamine 2000 transfection reagent. After 12hr, the WoundMaker™ tool (Essen BioScience) was used to create an end-to-end scratch in each well. Culture media was replaced with serum-free MEM with supplements, and cell migration across the width of the scratch wound was quantified over a 48hr time-course using the automated IncuCyte ZOOM live cell kinetic imaging system (Essen BioScience).

**Statistical analyses**

Significant differences in enzyme activities and cell migration between respective wild-type and mutant versions of B3GNT2, ST6GALNAC2, and B4GALT2 were estimated using a Student’s t-test, and a P value < 0.05 was considered statistically significant.

**SUPPLEMENTARY REFERENCES**


Figure S1. Mutational status of 36 glycosylation genes and KRAS/BRAF in colon cancers. Mutual exclusivity analysis using an exact statistical test (CoMeT) shows that the mutations affecting any of the 36 glycosylation genes (see Supplementary Table 4) are not independent of KRAS/BRAF mutations ($P > 0.05$). (*) indicates V425 as the only colon cancer sample harboring mutation in the BRAF gene in this dataset.
**Figure S2. Optimization of respective enzyme substrates.** A) Donor and acceptor substrates selected for assessing enzymatic activities of respective glycosyltransferases. B) *In vitro* enzymatic activities of wild-type B3GNT2, ST6GALNAC2 and B4GALT2 proteins against respective substrates (X-axis) at 24hr following incubation (see Methods). Fold-change in enzyme-specific activities of respective wild-type proteins, normalized to vector controls, are shown on the Y-axis. PNP refers to para-nitrophenol.
Figure S3. Sanger sequencing of \(B3GNT2\) somatic mutations. Representative DNA and RNA sequencing chromatograms of \(B3GNT2\) R6X (A), P186T (B), and D247H (C) mutations identified in respective colon cancers.
Figure S4. Sanger sequencing of ST6GALNAC2 somatic mutations. Representative DNA and RNA sequencing chromatograms of ST6GALNAC2 D43H (A) and R115W (B) mutations identified in respective colon cancers.
Figure S5. Expression of B3GNT2, ST6GALNAC2, and B4GALT2 in SW480 CRC cell line. Shown are the Affymetrix-based normalized RNA expression (Y-axis) profiles of the candidate genes in a random set of 16 normal colon epithelial RNA and the SW480 CRC parental cell line. Note the marked loss of expression of ST6GALNAC2 in SW480. Error bars in normal colon indicate mean ± s.e.m.


Elhammer AP, Kezdy FJ, Kurosaka A. 1999. The acceptor specificity of UDP-
16:171-180.


Heise N, Singh D, van der Wel H, Sassi SO, Johnson JM, Feasley CL, Koeller CM, Previato JO, Mendonça-Previato L, West CM. 2009. Molecular analysis of a UDP-


Ichikawa T, Nakayama J, Sakura N, Hashimoto T, Fukuda M, Fukuda MN, Taki T. 1999. Expression of N-acetyllactosamine and β1,4-Galactosyltransferase (β4GalT-I) During


Johansson MEV, Larsson JMH, Hansson GC. 2011. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proceedings of the National Academy of Sciences*, 108:4659-4665.


