

# Clinically Significant Missense Variants in Human *GALNT3*, *GALNT8*, *GALNT12*, and *GALNT13* Genes: Intriguing In Silico Findings

Muhammad Ramzan M. Hussain,<sup>1\*</sup> Jamal Nasir,<sup>2</sup> and Jumana Yousuf Al-Aama<sup>1</sup>

<sup>1</sup>Princess Al-Jawhara Al Brahim Center of Excellence in Research of Hereditary Disorders, Faculty of Medicine, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

<sup>2</sup>Division of Biomedical Sciences (BMS), Human Genetics Research Centre, St. George's University of London (SGUL), London, UK

## ABSTRACT

Aberrant glycosylation by *N*-acetylgalactosaminyl transferases (GALNTs) is a well-described pathological alteration that is widespread in hereditary diseases, prominently including human cancers, familial tumoral calcinosis and hyperostosis-hyperphosphatemia. In this study, we integrated different computational tools to perform the in silico analysis of clinically significant mutations (nsSNPs/single amino acid change) at both functional and structural levels, found in human *GALNT3*, *GALNT8*, *GALNT12*, and *GALNT13* genes. From function and structure based insights, mutations encoding R162Q, T359K, C574G, G359D, R297W, D303N, Y396C, and D313N substitutions were concordantly predicted highly deleterious for relevant GALNTs proteins. From intriguing findings, T359K-*GALNT3* was simulated with high contribution for disease susceptibility (tumor calcinosis) as compared to its partner variant T272K [Ichikawa et al. [2006] J. Clin. Endocrinol. Metab. 91:4472–4475]. Similarly, the prediction of high damaging behavior, evolutionary conservation and structural destabilization for C574G were proposed as major contributing factors to regulate metabolic disorder underlying tumor calcinosis and hyperostosis-hyperphosphatemia syndrome. In case of R297W-*GALNT12*, prediction of highly deleterious effect and disruption in ionic interactions were anticipated with reduction in enzymatic activity, associated with bilateral breast cancer and primary colorectal cancers. The second *GALNT12* mutation (D303N)—known splice variant—was predicted with disease severity as a result of decrease in charge density and buried behavior neighboring the catalytic B domain. In the lack of adequate in silico data about systematic characterization of clinically significant mutations in GALNTs genes, current study can be used as a significant tool to interpret the role of GALNTs reaction chemistry in disease-association risks in body. J. Cell. Biochem. 115: 313–327, 2014. © 2013 Wiley Periodicals, Inc.

**KEY WORDS:** CLINICALLY SIGNIFICANT MUTATIONS; *GALNT3*; *GALNT12*; *GALNT13*

Aberrant glycosylation by various glycosyltransferases is a well-described glycan alteration widespread in hereditary diseases, prominently including human cancers and familial tumoral calcinosis [Wood et al., 2007; Guda et al., 2009; Ichikawa et al., 2010]. The molecular phenomena underlying such altered glycosylation patterns have been credited to modify the processes like cell growth, differentiation, transformation, adhesion, and metastasis [Guo et al., 2002; Argueso et al., 2003; Zhang et al., 2003; Potapenko et al., 2010; Ichikawa et al., 2010; Hussain, 2012; Raman et al., 2012; Hussain et al., 2013].

Mucin-type *O*-glycosylation is a well-documented hallmark of protein glycosylation initiated by the large family of Golgi-associated

GALNTs (UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases) enzymes, which regulate the specific-transfer of GalNAc residue from UDP-GalNAc to hydroxyl group of Ser/Thr in polypeptide acceptors [Fritz et al., 2004; Yoshimura et al., 2010]. To date, different members of this family have been recognized with discrete-conserved regions in transmembrane, stem, and luminal domains. The Golgi-luminal region is further described as catalytic (GT1 and GalNAc-T) and ricin B-type lectin domains, making strong contributions in enzyme activities [Hagen et al., 1999; Tenno et al., 2002; Fritz et al., 2004]. Mutational variants in catalytic and ricin B-type lectin domains have been described with the loss of enzyme functions, regulated by the changes in charge density and kinetics of amino acid

\* Correspondence to: Muhammad Ramzan M. Hussain, Princess Al-Jawhara Al Brahim Center of Excellence in Research of Hereditary Disorders, Faculty of Medicine, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.

E-mail: geniouschemist26@gmail.com, mrmhussain@kau.edu.sa

Manuscript Received: 10 July 2013; Manuscript Accepted: 6 September 2013

Accepted manuscript online in Wiley Online Library (wileyonlinelibrary.com): 12 September 2013

DOI 10.1002/jcb.24666 • © 2013 Wiley Periodicals, Inc.

residues [White et al., 2000]. Aberrant glycosylation as a result of these mutations in GALNTs genes are found responsible for different types of inherited diseases including ovarian, breast and colorectal cancers the most spreading cancer types world-wide [Wood et al., 2007; Guda et al., 2009; Phelan et al., 2010]. For example, the single nucleotide polymorphism (OMIM# 602273/rs17647532) in *GALNT1* gene has been genotyped in 6965 cases, and reported in Ovarian Cancer Association Consortium [Phelan et al., 2010]. Hyperphosphatemic familial tumor calcinosis (HFTC; OMIM# 211900)—an autosomal recessive metabolic disorder—has been characterized by nonsense and missense mutations in *GALNT3* gene [Chefetz et al., 2005; Ichikawa et al., 2005]. A recent clinical investigation for breast cancer has revealed the loss in enzymatic activity of *GALNT5* as a result of missense mutation identified in *GALNT5* gene [Guda et al., 2009]. Similar to *GALNT3* and *GALNT5*, different missense and nonsense mutations (encoding stem and Golgi-luminal domains) have also been reported in *GALNT8* gene (OMIM# 606250), and nucleotide variants in *GALNT8* gene are found to be associated to the response to IFN therapy against chronic hepatitis C patients [White et al., 2000; Nakano et al., 2013]. Moreover, deleterious mutations in *GALNT12* gene (OMIM# 610290) have been identified in germline of 6 out of 272 unrelated patients with colon cancer [Guda et al., 2009].

Occurrence of GALNTs in large number of body tissues [Potapenko et al., 2010] and their substrate specificity for wide range of protein targets reflect high level diseases-risks that could be mediated by this class of enzymes as a result of aberrant glycosylation [Chefetz et al., 2005; Guda et al., 2009; Phelan et al., 2010; Potapenko et al., 2010]. Hence, to understand the molecular phenomena underlying such altered glycosylation patterns that have been credited to changes in cell growth, differentiation, transformation, adhesion, and metastasis, is an important task to carry out.

To gain function and structure based insights about the clinically significant variants, found in *GALNTs* genes, we performed a

comprehensive in silico study by involving SIFT, PolyPhen-2, PANTHER, I-Mutant, and various other tools. Captivatingly, we extended our study in the form of protein–protein and protein–ligand docking to visualize the surface cavities for lock and key fitting of GalNAc residue in native and mutant models of *GALNT3*, *GALNT12*, and *GALNT13*.

## METHODOLOGY

### DATA MINING AND ANALYSIS

To collect the clinically significant (disease-related) missense variants, Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim>) Ensemble, NCBI-dbSNP and literature data were considered as potential web sources. Detail of the SNPs (rs IDs, allele change and origin, and amino acid variation) and their concerned protein regions were retrieved from Human Genome Variation database (HGVBbase; <http://hgvbbase.cgb.ki.se>), NCBI-dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), and UniProt (<http://www.UniProt.org/>). For functional analysis, SIFT, PolyPhen, PANTHER, and I-Mutant tools were combined to perform the in silico characterization of clinically significant point mutations. Clustal Omega, ConSurf, and NetSurfP were used to study the conservation and solvent accessibility behavior. STRING and Kegg were employed to study protein interaction network. Project Hope, I-TASSER, Patchdock, FireDock, Autodock Vina, Chimera, Swiss-Pdb viewer were used to predict the structural impacts of mutant analogues on native proteins and their ligand complexes (Fig. 1).

**Analysis of functional consequences of missense variants by SIFT.** Based on degree of conservation and sequence homology, Sorting intolerant from tolerant (SIFT) simulates whether an amino acid substitution affects protein function. Our input carried the Ensembl ENSP IDs with amino acid change to get SIFT predictions for non-synonymous amino acid substitutions of *GALNTs* genes. SIFT algorithm generates score-based binary classification (tolerant and

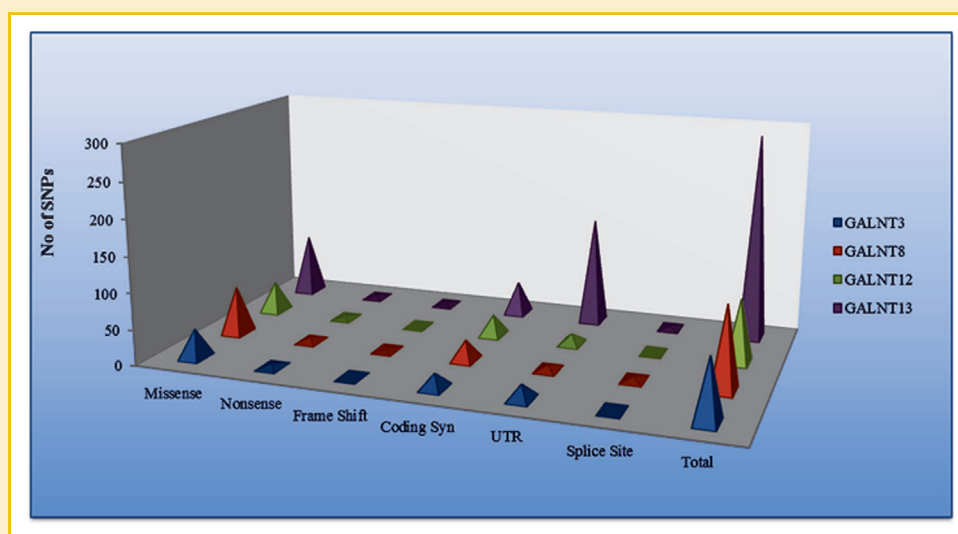


Fig. 1. A three-dimensional map of SNPs in 5 GALNTs genes: *GALNT13* is recorded with (highest total of) 288 SNPs, and *GALNT3* with (lowest total of) 85 SNPs. Eighty-six missense SNPs are found in *GALNT13* gene, 67 SNPs in *GALNT8*, 43 SNPs in *GALNT12*, and 40 SNPs in *GALNT3*. UTR-region SNPs are observed in high number (150) in case of *GALNT3* gene.

deleterious) for amino acid substitutions. The SIFT value  $\leq 0.05$  indicates the deleterious effect of non-synonymous variants on protein function [Ng and Henikoff, 2003, 2006].

**Estimation for functional impacts of non-synonymous amino acid substitutions by PolyPhen-2.** On the basis of Naïve Bayes machine-learning, Polymorphism Phenotyping-2 or PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>) predicts the consequence of an amino acid change on protein function by calculating false positive rate (FPR) and true positive rate (TPR) thresholds. Input options for PolyPhen2 constitute the FASTA sequence or UniProt accession number and detail of amino acids substitution [Alanazi et al., 2011; Hussain et al., 2012; Adzhubei et al., 2013].

**Classification of deleterious variants by PANTHER.** PANTHER or Protein ANalysis THrough Evolutionary Relationships (<http://www.pantherdb.org/>) is a unique database tool to characterize genes by using their functions. Amino acid substitution and protein sequence are the possible inputs to run the cSNP tools of PANTHER. PANTHER subPSEC score  $\leq -3$  classifies the amino acid substitution as deleterious or intolerant, whereas the score  $\geq -3$  is predicted to be less deleterious (0 score indicates neutral impact) [Thomas et al., 2003; Mi et al., 2005].

**Simulation for neural-network-based stability analysis of disease-causing mutations by I-Mutant.** I-Mutant (<http://folding.uib.es/i-mutant/i-mutant2.0.html>) is a widely used support vector machine based (SVM) tool which calculates alteration in stability and free energy of mutated proteins by excising the single-site mutations. The UniProt-FASTA sequence and amino acid change were used as inputs to assess the data from I-Mutant. I-Mutant output contains free energy change (DDG), stability alteration and reliability index at particular temperature and pH [Bava et al., 2004; Hussain et al., 2012].

#### **MULTIPLE SEQUENCE ALIGNMENT AND EVOLUTIONARY CONSERVATION OF PROTEINS BY CLUSTAL OMEGA**

Clustal Omega is a useful tool to produce biologically meaningful multiple sequence alignments and evolutionary relationships for divergent proteins. FASTA sequences of GALNT3, GALNT8, GALNT12, and GALNT13 proteins with UniProtKB/Swiss-Prot format were given as inputs to sequence input window to retrieve Cladograms or Phylograms. The conservation output carries "\*" (asterisk) to indicate positions which have a single, fully conserved residue; ":" (colon) to show conservation between groups of strongly similar properties; and "." (period) to specify conservation between groups of weakly similar properties [Goujon et al., 2010; Sievers et al., 2011].

**Identification of functional regions by ConSurf-DB.** Based on the phylogenetic relations between homologous sequences, ConSurf-DB tool estimates the evolutionary conservation of amino acids/nucleotides in protein/DNA/RNA. The database uses input options like UniProt-FASTA/PDB file/Clustal-MSA to collect the continuous conservation scores decorated with discrete color grading system, starting from grade 1 (with turquoise color) to the grade 9 (with maroon color). To anticipate the conserved behavior of selected non-synonymous variants in particular GALNTs proteins, we applied the ConSurf database and used the FASTA sequence of UniProt as an input [Ashkenazy et al., 2010].

**Solvent accessibility prediction by NetSurfP.** NetSurf is an artificial neural networks (ANNs) tool that assigns reliability scores applied to solvent accessibility predictions as an inherent part of the training process. FASTA sequences of selected GALNTs proteins were given to input window of NetSurfP to predict solvent accessibility of key amino acid substitutions. The NetSurfP output defines two subclasses for solvent accessible residues, the first class includes buried (low accessible), and second class comprises exposed (high accessible) residues [Petersen et al., 2009].

#### **STRUCTURAL SIMULATION FOR WILD-TYPE AND MUTANT RESIDUES BY PROJECT HOPE**

Project Hope (<http://www.cmbi.ru.nl/hope/home>) collects the structural information from series of sources to give the effect of a certain mutation on the protein structure. After curating the whole protein sequence from UniProt, HOPE server builds the protein models by collecting information from Yasara and WHAT IF Twinset web sources, UniProt database and DAS servers [Venselaar et al., 2010].

#### **INTERACTOME (PROTEIN-PROTEIN INTERACTION)**

**Protein interaction network by STRING.** STRING database "<http://string-db.org/>" determines the protein (physical and functional) interactions by integrating the information from Genomic Context, High-throughput Experiments, (Conserved) Co-expression, and Previous literature Knowledge (reported in the PubMed and relevant web-sources). Detail of protein name and species were given to STRING database to obtain the binary and multiple interactions for candidate proteins [Franceschini et al., 2013].

**Mechanistic investigation of GALNTs reactions and KEGG.** To simulate the functional networking of GALNTs enzymes, we used Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY and LIGAND, and curated the data for molecular reaction and interaction networks, including metabolic pathways, regulatory pathways, and molecular complexes for biological systems [Kanehisa et al., 2006].

#### **BUILDING OF PDBS FOR GALNTs (NATIVE AND MUTANT) STRAINS AND IN SILICO ANALYSIS OF MUTANT AND WILD-TYPE RESIDUES**

UniProt database and I-TASSER tool were used to map out the quality based 3D models for GALNTs proteins that were further analyzed for solvent accessibility, ligand interactions and force field energy calculations in wild-type and mutant strains.

**Simulation of full-length secondary structures by I-TASSER.** The iterative threading assembly refinement (I-TASSER) combines various approaches such as threading, ab initio modeling and atomic-level structure refinement for computerized prediction of protein structure and function. FASTA sequences of relevant GALNTs proteins were given to I-TASSER tool to generate full-length secondary and tertiary structures, and functional annotations. The higher value of C-score depicts the good confidence level of relevant-predicted model. In order to approximate the correct topology of generated models, C-score has correlated with the TM-score and RMSD [Zhang, 2008; Roy et al., 2010].

**Molecular docking for protein-protein and protein-ligand interactions by PatchDock, FireDock and AutoDock Vina.** PatchDock algorithm entails object recognition and image segmentation techniques to carry out rigid docking having three-stage filtering

process: (a) Molecular Shape Representation, (b) Surface Patch Matching, and (c) Filtering/scoring.

After retrieving the top 1000 complexes, the PatchDock output was redirected to FireDock (as an input) to generate the top 10 refined complexes of associated GALNT–ligand [Schneidman-Duhovny et al., 2005; Trott and Olson, 2010]. Moreover, AutoDock Vina [Trott and Olson, 2010], a new program with improved accuracy of docking and new scoring function was used to predict the unbound receptor–ligand structures for selected GALNTs.

**Structural visualization for native and mutant analogues by Swiss-Pdb Viewer and Chimera.** Swiss-Pdb Viewer (4.0.4) is a unique computational application that offers a user-friendly interface to investigate several proteins at the same time. In Swiss-Pdb Viewer, proteins can be superimposed and mutated in order to annotate structural alignments, and locate their active sites or any other relevant parts [Guex et al., 2009].

UCSF Chimera (<http://www.cgl.ucsf.edu/chimera>) is a quality-oriented program for interactive analysis of macromolecular features, including density plots, supramolecular structures, sequence alignments, docking outcomes, trajectories and conformational ensembles [Pettersen et al., 2004]. After generating the PDB files for protein–ligand complexes, UCSF chimera program was employed to see the including density plots and conformational ensembles.

## RESULTS

Twelve clinically significant missense variants in four *GALNTs* genes (*GALNT3*, *GALNT8*, *GALNT12*, and *GALNT13*) were found with their pathogenic relevance in dbSNP–NCBI, Ensemble and experimentally proved data (Table I). From four candidate genes, *GALNT13* is retrieved with the highest number (86) of missense SNPs, and *GALNT3* with lowest number of SNPs (40) in dbSNP–NCBI database (Fig. 1). Among 12 missense variants, only 3 (c.1076G>A, c.937G>A, and c.1427G>A) were curated with their clinical significance in dbSNP. The remaining 09 non-synonymous variants (c.485G>A, c.815C>A, c.1076C>A, c.1720T>G, c.3G>A, c.889C>T, c.907G>A, c.1187A>G, and c.1472C>T) in *GALNT3* and *GALNT12* genes were

selected for their disease-relevance in documented data (Table I). From total (12) variants, only two experimentally proved disease-variants (c.485G>A and c.1187A>G) did not determine with rs IDs detail in both dbSNP–NCBI and Ensemble databases (Table I).

## FUNCTIONAL IMPACTS OF CLINICALLY SIGNIFICANT VARIANTS ON PROTEIN FUNCTION

A non-synonymous (missense variant) is an important single amino acid change in a coding region that can affect protein function and generate phenotype consequences. Analysis of such amino acid substitutions (AAS) by structure and sequence based methods contributes to identify the damaging nsSNPs involved in human disease. In fact the integration of results obtained from amino acid substitution methods (AASMs) and their concordance analysis can help us to validate the outcomes curated by AASMs. In our study, we applied four in silico sequence and structure based AASMs (SIFT, PolyPhen, PANTHER–cSNP, and I-Mutant) to annotate the functional impacts of key non-synonymous variants, by evaluating the significance of the amino acids they carry.

**Analysis of functional consequences by SIFT.** To assess the effect of AAS on protein function, SIFT entails the key positions in a protein sequence based on sequence homology of amino acids. After submitting 12 missense variants to SIFT, a total of 11 (≈92%) nsSNPs (c.485G>A, c.815C>A, c.1076C>A, c.1720T>G, c.1076G>A, c.3G>A, c.889C>T, c.1187A>G, c.1472C>T, c.937G>A, and c.1427G>A) were predicted as intolerant or damaging with SIFT score ranging from 0.00 to 0.04 (Table II). According to SIFT algorithm these variants could potentially modify the protein function, hence, enabled them biologically significant to further look at.

**Analysis of deleterious variants by PolyPhen–2.** Based on Naïve Bayes posterior probability calculation, PolyPhen–2 estimates that a given missense variant is damaging (probably/possibly damaging) or benign with the calculation of false positive and true positive rates (the chance that the mutation is classified as damaging when it is indeed damaging). Unlike SIFT, PolyPhen–2 (HumDiv/HumVar) score ranges from 0 to 1, and 0 value indicates the tolerated/neutral effect of nucleotide variation on protein function. According to PolyPhen–

TABLE I. Twelve Clinically Significant Variants With Detail of rsIDs, Allele Change and Phenotype, and Amino Acid Variations

Gene	rs ID	Allele change	Amino acid change	Region	Phenotype	Refs.
<i>GALNT3</i>	ND	c.485G>A	R162Q	Stem region	TC	Ichikawa et al. [2010]
	rs137853090	c.815C>A	T272K	Catalytic A (GT1)	TC	Ichikawa et al. [2006]
	rs137853091	c.1076C>A	T359K	Catalytic B		
	rs267606841	c.1720T>G	C574G	Ricin B-type lectin	TC/HHS	Ichikawa et al. [2010]
<i>GALNT8</i>	rs142567239	c.1076G>A	G359D	Catalytic B	PA	NCBI-dbSNP
	rs267606839	c.3G>A	M1I	Stem region	PCC and DC	Guda et al. [2009]
<i>GALNT12</i>	rs149726976	c.889C>T	R297W	Linker A region (region between Catalytic A and B domains)	PCCC, ICRB and ICLB	Guda et al. [2009]
	rs145236923	c.907G>A	D303N	Linker A region	CRC	Clarke et al. [2012]
<i>GALNT13</i>	ND	c.1187A>G	Y396C	Linker B region (region between Catalytic B and lectin domains)	CRC	Clarke et al. [2012]
	rs267606840	c.1472C>T	T491M	Lectin domain	PCCC and CC	Guda et al. [2009]
	rs140738540	c.937G>A	D313N	Catalytic B	PA	NCBI-dbSNP
	rs148805770	c.1427G>A	R476Q	Ricin B-type lectin	PA	NCBI-dbSNP

ND, not determined; TC, tumoral calcinosis; HSS, hyperostosis–hyperphosphatemia syndrome; PCC, proximal (transverse) colon carcinoma; DC, ductal carcinoma right breast; CC, colon cancer at the splenic flexure; PCCC, proximal (cecum) colon cancer; ICRB, invasive carcinoma right breast; ICLB, intraductal carcinoma left breast; CRC, colorectal cancer; PA, pathogenic.

TABLE II. SIFT, PolyPhen-2 and PANTHER-cSNP Analysis of Curated Variants

Allele change	SIFT score	SIFT prediction	PolyPhen score (HumDiv)	PolyPhen prediction (HumDiv)	subPSEC	PANTHER prediction
R162Q	0	Damaging	1	Probably damaging	-7.16859	Highly deleterious
T272K	0	Damaging	0.993	Probably damaging	-5.61087	Highly deleterious
T359K	0	Damaging	0.993	Probably damaging	-7.58197	Highly deleterious
C574G	0	Damaging	1	Probably damaging	-7.96936	Highly deleterious
G359D	0.04	Damaging	0.985	Probably damaging	-6.81052	Highly deleterious
M1I	0	Damaging	0.118	Benign	-	Weak prediction
R297W	0	Damaging	0.999	Probably damaging	-6.6485	Highly deleterious
D303N	0.47	Tolerated	1	Probably damaging	-4.71957	Highly deleterious
Y396C	0	Damaging	0.999	Probably damaging	-8.75743	Highly deleterious
T491M	0	Damaging	0.999	Probably damaging	-6.04206	Highly deleterious
D313N	0.01	Damaging	0.955	Possibly damaging	-4.86872	Highly deleterious
R476Q	0.01	Damaging	0.955	Possibly damaging	-4.41714	Highly deleterious

Damaging variants classified as highly functional on the basis of coherence in results of SIFT, PolyPhen-2, PANTHER-cSNP and I-Mutant (Table III) tools.

2, 9 (75%) variants (c.485G>A, c.815C>A, c.1076C>A, c.1720T>G, c.1076G>A, c.889C>T, c.1187A>G, c.907G>A, and c.1472C>T) were categorized as probably damaging; 2 (17%) nsSNPs (c.1427G>A, c.937G>A, and c.1427G>A) were exhibited possibly damaging, and remaining 1 (8%) variants (c.3G>A) were found with benign effect to the protein function (Table II). Totally, 11 (~92%) variants were predicted damaging by PolyPhen-2. The scoring of SIFT and PolyPhen2 displays the quantitative calculation of severity-effect of AAS on protein function.

**Identification of functional variants by PANTHER-cSNP.** PANTHER calculates the Substitution Position-Specific Evolutionary Conservation (subPSEC) scores from alignments to hidden Markov models (HMMs) and identifies the functional impact of SNPs. From PANTHER results only 1 (8%) variant "c.3G>A" was determined with weak prediction. Based on the subPSEC score ranging from -4.41714 to -8.75743, 11 (92%) SNPs (c.485G>A, c.815C>A, c.1076C>A, c.1720T>G, c.1076G>A, c.889C>T, c.907G>A, c.1187A>G, c.1472C>T, c.937G>A, and c.1427G>A) were predicted with deleterious effect on the protein function (Table II).

**Simulation for neural-network-based stability analysis by I-Mutant.** By using the I-Mutant SVM-based tool, we have added another layer of refinement to assess the functional consequences of key mutations. After analyzing the functional consequences of candidate variants by SIFT, PolyPhen and PANTHER, the amino acid substitutions were further submitted to I-Mutant 2.0 to validate their deleterious effect. From I-Mutant results, 10 (83%) out of 12 mutations (c.485G>A, c.1076C>A, c.1720T>G, c.1076G>A, c.3G>A, c.889C>T, c.907G>A, c.1187A>G, c.1472C>T, and c.937G>A) were evaluated with decrease in stability, and considered as highly deleterious variants to GALNTs proteins (Table III). This decrease in stability was predicted with DDG and RI values ranging from -1.96 to 0.71 and 1 to 9, respectively.

**PREDICTION FOR EVOLUTIONARY CONSERVED REGIONS AND SOLVENT ACCESSIBILITY BY CLUSTAL OMEGA, CONSURF-DB AND NETSURFP**

Clustal Omega creates high-quality accurate alignments by aligning diverse range of protein sequences. From output file of multiple sequence alignment (MSA), 2 residues of GALNT3 (R162 and C574) with asterisk notation "\*" were found in highly conserved region (Fig. 2A). While one residue (T359) was observed with the

conservation between groups of strongly similar properties, indicated with ":". However, the prediction of G359 with asterisk sign indicated the high conservation of this residue in GALNT8 protein. Similar to G359-GALNT8, presence of asterisk mark for Y396-GALNT12 indicated the conservation of this residue in linker B domain; and colon mark for T491 of GALNT12 highlighted conservation between groups of strongly similar properties. Except Met1 of GALNT12, all remaining amino acids of selected GALNTs proteins were located in the vicinity of conserved residues.

By submitting the MSA file of Clustal Omega to Clustal Phylogeny tool, rooted phylogenetic tree was obtained to understand the conservation behavior among curated GALNTs proteins. Phylogenetic tree in Figure 2B indicated evolutionary implications for GALNTs proteins: the first lineage gave rise to GALNT13, that splitted into two tips: first lineage was given rise into single GALNT8 node, and second lineage is further subdivided into two nodes, separated into GALNT12 and GALNT3. Both GALNT12 and GALNT3 were found with almost same taxonomic length, and share the common ancestor with each other. All three nodes of GALNT3, GALNT12, and GALNT8 shared common ancestor lineage with GALNT13.

ConSurf algorithm projects the coloring grades for highly conserved residues of protein assemblies by taking into account the phylogenetic homology among the sequences and similarity of

TABLE III. I-Mutant Prediction to Anticipate the Effect of GALNTs Variants on DDG and Protein Stability

Allele change	Amino acid change	pH	Temp (°C)	Stability	RI	DDG (Kcal/mol)
c.485G>A	R162Q	7.0	25	Decrease	9	-1.74
c.815C>A	T272K	7.0	25	Increase	4	0.06
c.1076C>A	T359K	7.0	25	Decrease	6	-1.49
c.1720T>G	C574G	7.0	25	Decrease	7	-0.67
c.1076G>A	G359D	7.0	25	Decrease	7	-1.17
c.3G>A	M1I	7.0	25	Decrease	6	-0.26
c.889C>T	R297W	7.0	25	Decrease	6	-0.47
c.907G>A	D303N	7.0	25	Decrease	3	-1.96
c.1187A>G	Y396C	7.0	25	Decrease	1	0.31
c.1472C>T	T491M	7.0	25	Decrease	3	0.71
c.937G>A	D313N	7.0	25	Decrease	8	-1.31
c.1427G>A	R476Q	7.0	25	Increase	0	-0.34

Deleterious variants causing high decrease in protein structure stability.





core of catalytic domain B, and charge of the buried wild-type residue is lost by this mutation.

The third mutation (Tyr396Cys) of *GALNT12* matches a previously described entry: “Colorectal cancer type 1 (CRCS1) [MIM:608812].” This buried residue is found in the linker region between catalytic B and ricin B-type lectin. The hydrophobicity of the wild-type and mutant residue differs and will cause loss of hydrogen bonds and associated mis-folding in the linker region. In the fourth mutation, wild-type (Thr491) residue is annotated as “Ricin B-type lectin” in UniProt with Q81XK2. The introduction of methionine as a mutant analogue can disturb the hydrogen bonding with the Glutamic acid at 495 position. This mutation is also described previously with “Colorectal cancer type 1 (CRCS1) [MIM:608812]” in OMIM and ExPASy (VAR\_064363) databases.

In the first disease-reported mutation of *GALNT13*, the wild-type residue is observed with 2 H-bonding interactions with Lys363 and Arg190. The wild-type residue also forms a salt bridge with Arginine at 190 and 367 positions. This amino acid substitution is occurring in the catalytic B domain of *GALNT13*, hence, the loss of charge and ionic interactions can abolish the catalytic tendency of protein. *GALNT13* second substitution (Arg476Glu) lies in the ricin B-type lectin. The wild-type residue forms a hydrogen bond with the Glutamic acid on position 474. The mutation introduces loss in charge between the wild-type and mutant amino acid, which can disturb this domain and abolish its function.

#### PROTEIN–PROTEIN INTERACTION NETWORK BY STRING AND KEGG REACTION PATHWAY

STRING database determines the protein–protein interaction network by retrieving the evidences from Genomic Context, High-throughput Experiments, Co-expression, and Previous literature Knowledge. According to STRING results, *GALNT3*, *GALNT8*, *GALNT11*, *GALNT12*, and *GALNT13* proteins indicated the similar but strong interactions for B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1 (Fig. 4A). However, the difference was seen in case of interactions for MUC1, MUC2, MUC7, MUC5B, CHPF, B4GALNT1, MYL4, KLRC3, FGF7, FGF23, SDC3, NUP188, ROCK2, ABO, and MYL4 proteins. Except *GALNT3* protein, all other *GALNTs* members showed the weak interaction with CHPF. *GALNT3* was observed with weak interactions for FGF7, FGF23 and ABO. *GALNT8* was specified with MYL4 and KLRC3. *GALNT11* was observed with weak interactions for ROCK2, NUP188, ZNF804B, and CEP152. MUC1 and MUC5B revealed the weak interaction behavior with *GALNT12*. *GALNT13* indicated the weak interactions with SDC3, SGCB and XIRP2 (Fig. 4A).

KEGG is a unique database source to simulate the functional networking of *GALNTs* enzymes in the form of molecular reaction and interaction networks, including metabolic pathways. Like STRING, KEGG showed the enzymatic pathway for curated *GALNTs* enzymes in mucin biosynthesis (KEGG ID: hsa00512) by involving the same set of enzymes like B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1 (Fig. 4B).

#### PROTEIN–LIGAND FLEXIBLE DOCKING

In the light of STRING and KEGG results for protein–protein interactions, we switched our analysis to protein–ligand docking to

further examine the particular transfer of GalNAc residue mediated by native and mutant structures of *GALNTs*. For that purpose we used different docking flavours like Patchdock, Firedock, and Autodock Vina to analyze the ligand–receptor interaction behavior for *GALNT3*, *GALNT12*, and *GALNT13* models.

After efficient rigid-docking of *GALNT3*-wt (wild type) with FGF23 (Fig. 5), FireDock was employed to select the top 10 models from the basket of 1,000 *GALNT3*–FGF23 complexes to further refine the rigid-body docking candidates on the basis of scoring matrix. The highest scoring (stable) complex of *GALNT3*–FGF23 was selected to examine the possible surface fitting of GalNAc for FGF23 (Thr178). From our results, unbound surface fitting of GalNAc molecular was found within the pentad (five residues) cavity, comprising the Thr178-FGF23, Arg175-FGF23, His106-FGF23, Asp344-*GALNT3*, and Gln348-*GALNT3*, for wild-type *GALNT3* (Fig. 6). In mutant models, GalNAc residue was observed at different sites other than Thr178 of *FGF23*. In R162Q mutant model, the FGF23 binding pocket for GalNAc residue was comprised of tetrad cavity carrying Met89, Arg91, Ile173, and Pro174 of *FGF23*. For the second mutant model (T272K), GalNAc residue was found in contact with three residues (Met1, Arg91 and Tyr107 of *FGF23*) making a triad cavity for surface fitting. In the third mutation “T359K,” binding cavity was formed by His106, Tyr107 and Arg176 to make surface fitting for GalNAc residue. For the last mutant model, four residues (His106, Tyr107, Arg176, and His177) were found in development of tetrad binding cavity for GalNAc fitting. In all mutant complexes, FGF23 showed different binding pattern for *GALNT3* protein (Fig. 5) and therefore affected the behavior of GalNAc ligand for its flexible fitting with FGF23.

After MM2 energy minimization of MUC7 (TTAAPPTPSATTPA) to minimum RMS Gradient of 0.100, *GALNT12*-wt (wild type) protein was docked with the *MUC7*. The filtering of energetically most favourable conformation of *GALNT12*–*MUC7* complex among the variety of conformations was performed on the basis of Patchdock and Firedock. The wild-type model of *GALNT12* was found in contact with *MUC7* tandem repeat domain through Arg373, Asn379, Tyr395, Asn399, Gln473, Ile517, His519, and Leu520 residues (Fig. 7), located in catalytic B, linker B, and ricin B-type lectin regions. For first mutant model, no significant change was observed for hand-in-glove GalNAc fitting to *MUC7*. In the second mutant model (R297W-*GALNT12*), the binding pockets for *MUC7* was constituted with Arg295, Gln275, Ser271, Thr286, Arg497, Glu504, Glu520, Cys521, and Glu522 to make unbound surface fitting. Upon complexation of near-to-bound (unbound) conformation of GalNAc residue with *MUC7* in *GALNT12*–*MUC7* complex, both Thr-11th and Thr-12th were seen in contact with GalNAc. Like *GALNT12*(wt)–*MUC7* complex, *GALNT12*(D303N)–*MUC7* intermediate showed the same behavior for GalNAc attachment, situated on Thr-12th of *MUC7*. The remaining two mutant models–Y396C and T491M–indicated the aberrant fitting of GalNAc on *MUC7* in relevant *GALNT12*–*MUC7* complexes (Fig. 7).

For *GALNT13*-wt, geometric fitting of *MUC7* was observed within the main stretch of Leu24, Glu184, Leu193, Asn300, and Glu303, while the unbound geometric fitting of GalNAc residue was visualized at Thr-12th of *MUC7* (Fig. 8), as reported in the literature [Zhang et al., 2003]. The first mutant model D313N was



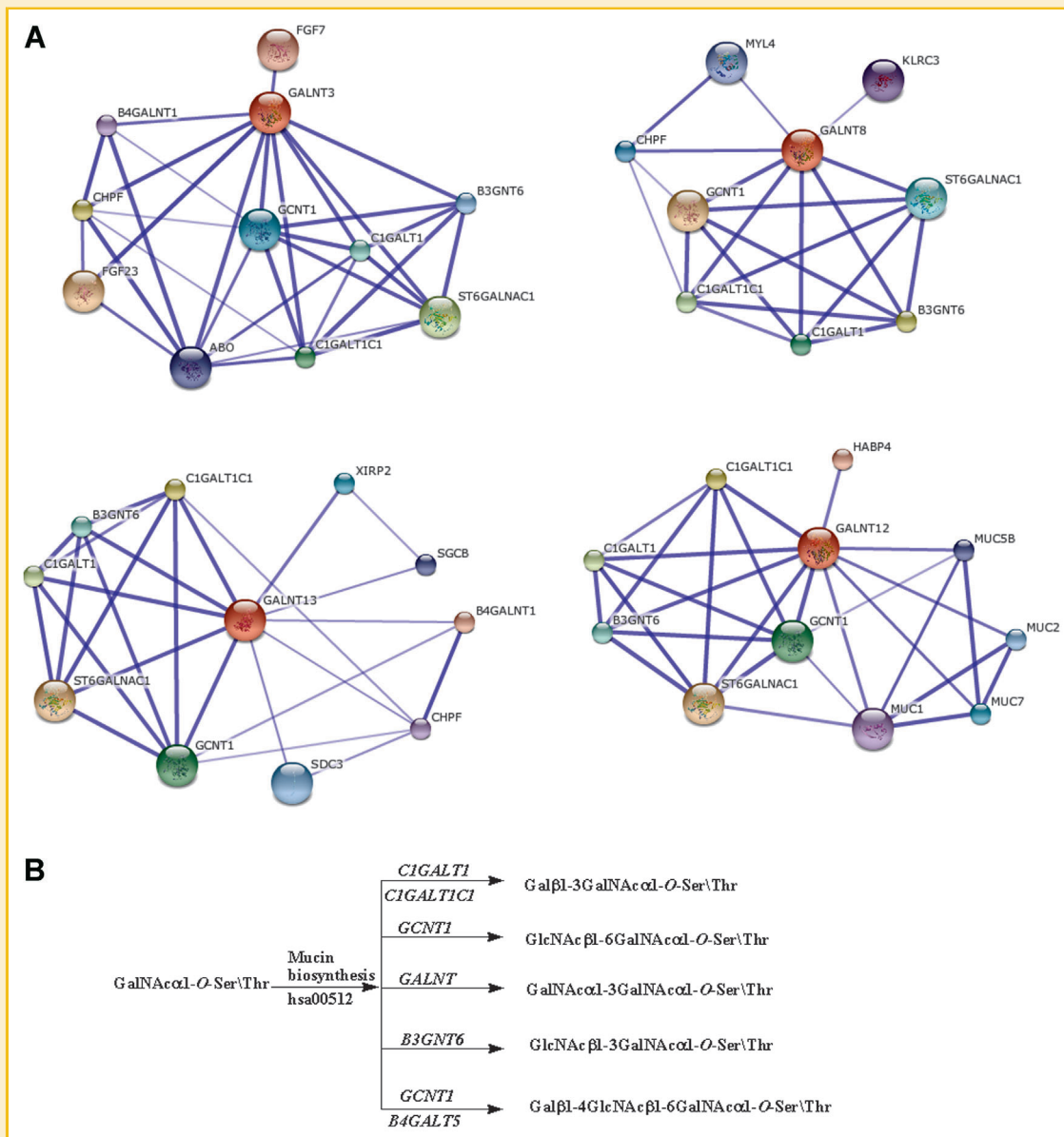


Fig. 4. Protein-protein interaction (interactome) for GALNT3, GALNT8, GALNT12, and GALNT13 proteins. A: Similar and strong interactions are observed for B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1 proteins. The difference in interaction behavior is seen for the MUC1, MUC2, MUC7, MUC5B, CHPF, B4GALNT1, MYL4, KLRC3, FGF7, FGF23, SDC3, NUP188, ROCK2, ABO, and MYL4 proteins. Except GALNT3 protein all other members are predicted with weak interaction for CHPF. GALNT3 is detected with weak interactions for FGF7 and ABO. B: Involvement of *C1GALT1*, *C1GALT1C1*, *GCNT1*, *B3GNT6*, *GALNTs*, and *B4GALT5* to mediate the mucin biosynthetic pathway (KEGG ID: hsa00512).

seen interacting with MUC7 through GLys261, Arg265, Tyr267, Pro280, Pro419, Asp420, and Glu422. Due to change in binding sites between *MUC7* and *GALNT13*, GalNAc residue was docked in surface contact with Ala-10th of *MUC7* instead of Thr-12th. In the second mutant model (R476Q), *MUC7* was fitted in the cavity of eight residues—Val5, Val9, Lys8, Leu26, Glu184, Asn300, Glu303, and Asn319. Instead of surface fitting with *MUC7*, the unbound fitting of GalNAc residue was visualized in triad cavity carrying Asn216-*GALNT13*, Asn319-*GALNT13*, and Asn365-*GALNT13*.

## DISCUSSION

With the advent of high throughput technology such as whole exome and genome sequencing, the list of genetic variations is increasing day by day in an active manner. After delineating the disease linked missense variants through sequencing techniques, the next step is to figure out structural and functional consequences at protein level to further understand the biological mechanisms underlying the disease association that may help to discover vital drug elements through in silico approaches [Alanazi et al., 2011; Hussain et al., 2012].

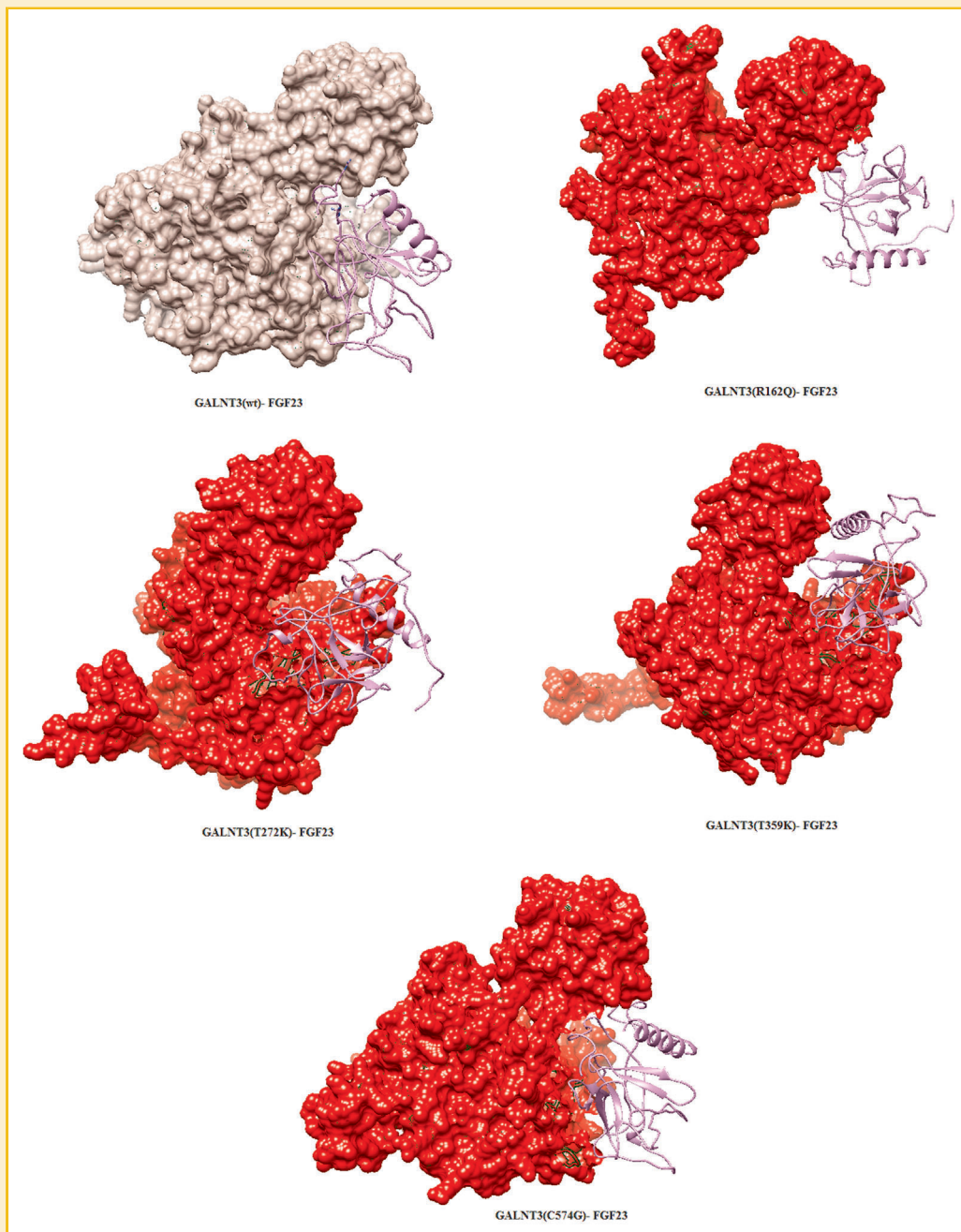


Fig. 5. Different binding behavior of FGF23 to create unbound hand-in-glove *GALNT3*-FGF23 complexes.

In the present study, we initialized the computational screening for clinically reported missense variants by using sequence and structure based tools (SIFT, PolyPhen-2, PANTHER-cSNP, and I-Mutant). The second level screening is performed by using Clustal Omega, ConSurf, NetsurfP tools to determine the conservation and solvent accessibility of selected SNPs. After second level screening, we prolonged our study for protein-protein and protein-ligand docking to decipher the structural consequences of mutations. For *GALNT3*, except T272K, all three missense variants are concordantly predicted deleterious by

SIFT, PolyPhen-2, PANTHER, I-Mutant (Table I). Likewise, R162-*GALNT3*, T359-*GALNT3*, and C574-*GALNT3* residues are found in highly conserved domain by Clustal Omega and ConSurf tools (Figs. 2 and 3). Although, T272K and T359K have been previously reported as compound mutation [Ichikawa et al., 2006], our concordance in silico predictions for these two mutations reveals the high deleterious effects of T359K in tumor calcinosis susceptibility (increased renal tubular phosphate reabsorption) as compared to its partner variant (T272K). High deleterious function (Table I), conservation behavior

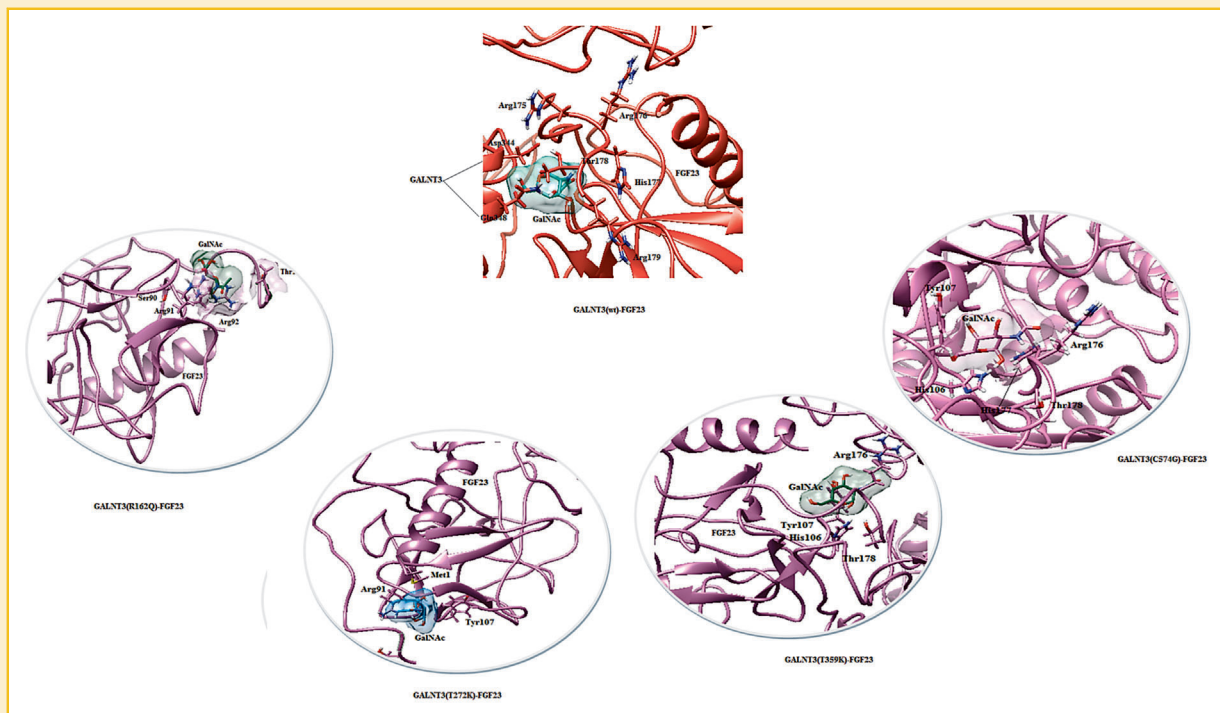


Fig. 6. Hand-in-glove surface-fitting of GalNAc within the cavities generated by FGF23 and/or GALNT3 residues in GALNT3(wt)-FGF23, GALNT3(R162Q)-FGF23, GALNT3(T272K)-FGF23, GALNT3(T359K)-FGF23, and GALNT3(C574G)-FGF23 complexes.

(Figs. 2 and 3) and loss of H-bonding interactions by buried T359 residue anticipate the more contribution of T359K instead of T272K, screened as compound mutation in tumor calcinosis patient [Ichikawa et al., 2006]. Unlike first three nucleotide variants of *GALNT3*, C574G has been characterized in both tumor calcinosis and hyperostosis-hyperphosphatemia syndrome [Ichikawa et al., 2010] which may be due to the high damaging behavior by SIFT, PolyPhen-2, PANTHER, and I-Mutant, high evolutionary conservation by Clustal Omega and Consurf, structure destabilization (loss of cysteine rigidity) by Hope server, and presence of C574 residue in close proximity of *FGF23* in *GALNT3*-*FGF23* complex—to affect *O*-glycosylation mechanism of *FGF23* at Thr178, by affecting the phosphate level circulation associated with both tumoral calcinosis and hyperphosphatemia syndrome [Topaz et al., 2004]. Finally, the localization of R162 in stem region, T272 in catalytic A domain, T359K in catalytic B domain, and C574G in lectin region, indicate the involvement of all regions of *GALNT3* protein in disease regulation (Table I and Fig. 2).

The G359D variant of *GALNT8* is simulated with high damaging (Table II) effect, and observed in catalytic B region with highly conserved behavior. So, G359D-*GALNT8* is predicted to alter the considerable change in energy, electrostatic intra-molecular interactions, and functional interaction behavior of *GALNT8* enzyme. Only R297W and Y396C of *GALNT12* are delineated as functionally deleterious variants by concordant results of SIFT, PolyPhen-2, PANTHER, and I-Mutant. The recognition of R297W in an African patient with three cancer phenotypes—proximal (cecum) colon cancer,

invasive carcinoma right breast and intraductal carcinoma left breast—may indicate high level disease risks linked with this mutation [Guda et al., 2009]. From our results, increased damaging effects predicted by SIFT, PolyPhen, PANTHER, and I-Mutant for R297W may support the underlying pathogenicity of this variant. Correspondingly, the distortion in H-bonding interactions due to R297W at the border region of linker A and catalytic B domain, and improper transfer of GalNAc to protein targets (like MUC1 and MUC7) in mucin biosynthesis pathway (KEGG ID: hsa00512), is often reported in tumors of colon, breast, ovarian, lung and pancreatic origin [Gendler, 2001; Ju et al., 2011; Walsh et al., 2013], hence, support the disease severity associated with R297W in regulation of different cancer diseases including bilateral breast cancers and primary colorectal cancer [Guda et al., 2009]. Within experimental data [Guda et al., 2009], only 7% reduced enzymatic activity has been seen for R297W while 37% is observed for D303N of *GALNT12*. As both residues—R297 and D303—are found in linker A region (bordering the catalytic B regions), the enzymatic activity of *GALNT12* is highly associated with this domain. For D303N mutation reported in colorectal cancer and predicted as non-tolerant (NT) variant in one of the recent report [Clarke et al., 2012], our SIFT result is indicating the tolerant effect. According to our SIFT analysis (cross validated with Ensemble SIFT results), D303N carries tolerant effect on protein function instead of NT. However, except SIFT, all three tools (PolyPhen-2, PANTHER, and I-Mutant) corroborate the damaging or intolerant behavior of D303N. Similar to *GALNT3*, five variants of *GALNT12* are found in stem (M1I), linker A (R297W; D303N) and B (Y396C), and lectin (T491M) regions,



Fig. 7. Unbound surface fitting of MUC7 for *GALNT12*-wt (wild type) and its mutant (R297W, D303N, Y396C, and T491M) protein models. Except *GALNT12*(R297W)-MUC7, all three complexes showed the surface fitting of GalNAc residue with Thr-12th of MUC7 (TTAAPPTPSATTPA).

therefore, specifying the involvement of these regions in enzymatic catalysis of protein targets (Fig. 4). For the pathogenic variants (D313N and R476Q) of *GALNT13*, SIFT, PolyPhen-2, PANTHER, and docking tools are used to see the damaging effects of D313N and R476Q on *GALNT13* protein function that may regulate the aberrant glycosylation of respective target proteins like MUC1, MUC2, MUC5B, and MUC7.

From overall functional analysis, both SIFT and PolyPhen predicted 10 (83%) variants to be highly deleterious. Although there is significant difference in SIFT and PolyPhen-2 algorithms [de Alencar and Lopes, 2010], the concordance analysis of SIFT and PolyPhen results showed 10 (83%) variants (c.485G>A, c.815C>A, c.1076C>A, c.1720T>G, c.1076G>A, c.889C>T, c.1187A>G, c.1472C>T, c.937G>A, and c.1427G>A) with damaging effects on

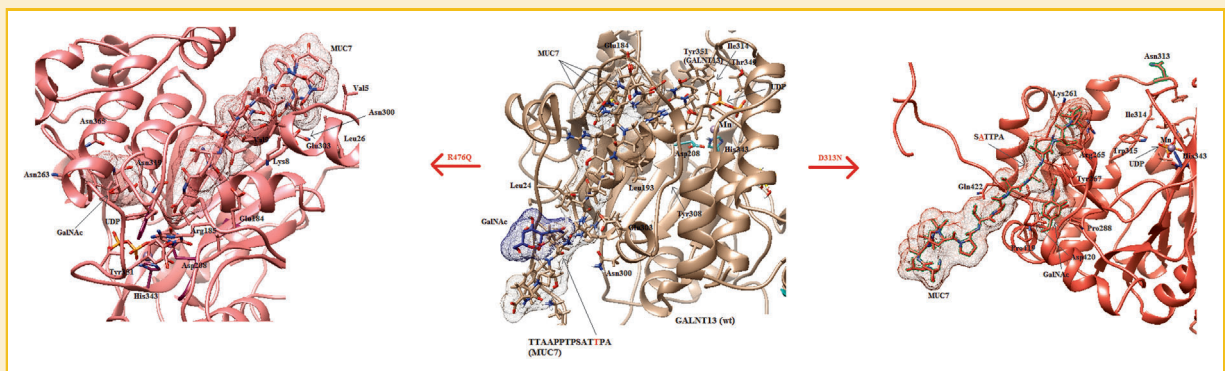


Fig. 8. Surface fitting of MUC7 with *GALNT13* models (wild type and mutants). Unlike *GALNT13*-wt (wild type) structure, GalNAc is found within the tetrad cavity comprising Ala-10th-MUC7, Tyr267-GALNT13, Pro288-GALNT13, and Asp420-GALNT13 for *GALNT13* (D313N) mutant model. In second model (R476Q), the binding conformation of GalNAc within the triad cavity of Asn263-GALNT13, Asn319-GALNT13, and Asn365-GALNT13 residues, reflects about the level of aberration that may cause by this mutation.

relevant proteins. Moreover, good coherence is found for predictions taken from PANTHER and PolyPhen-2, simulated 11 (92%) nsSNPs (except c.3G>A) to be deleterious. By comparing outputs from all four tools (SIFT, PolyPhen-2, PANTHER-cSNP and I-Mutant), nucleotide variations that encode R162Q, T359K, C574G, G359D, R297W, Y396C, T491M, and D313N variants are found highly damaging. The differences in prediction capabilities can be attributed to the fact that every method uses different set of sequences and alignments. In silico investigation tools that integrate both sequence and structure based approaches will of added advantage in providing a reliable prediction results with wider coverage of different aspects of nucleotide variation analysis. However, variability in prediction outputs of these algorithms reflects both advantages and disadvantages.

Second stage concordance analysis is performed by involving three different tools—Clustal Omega, ConSurf and NetSurfP—to validate the quality of results predicted for evolutionary conservation and solvent accessibility parameters. According to Clustal Omega and ConSurf findings, five residues (R162-GALNT3 T359-GALNT3, C574-GALNT3, G359-GALNT8, and Y396-GALNT12) are observed in highly conserved region and predicted to have potential impact on relevant GALNTs proteins. Except R162-GALNT3 and D313-GALNT13, all residues are simulated with the same solvent accessibility behavior (six buried and five exposed residues) in ConSurf and NetSurfP results. In rooted phylogenetic tree, GALNT13 is inferred as most common ancestor for all decedents (GALNT3, GALNT8, and GALNT12). Different branch lengths for taxonomic units of phylogenetic tree may be interpreted as time estimates for evolutionary conservation of relevant GALNTs.

Form structural analysis based on Project Hope results, R162 residue is annotated with hydrogen bonding and salt bridge for Glu281 and Glu315. Whereas T272 residue is observed with hydrogen bonding for Thr186 and Thr187, and T359 is predicted with 2 hydrogen bonding for Gly329 and Thr361. As a result of substitutions for T272K and T359K, mutant residue can distort the intra-molecular hydrogen bonding in the catalytic subdomain A and B. In case of last variant (C574G) of GALNT3, introduction of Gly and loss of Cys can induce destabilization for native structure by minimizing the rigidity. For G359-GALNT8, 2 hydrogen bonds are observed for Leu331 and Leu360. Due to difference in flexibility, charge density and density, D359 (mutant) can influence the H-bonding and catalytic activity maintained by wild-type residue (G359). Substitution for R297, Y396, and T491 residues in GALNT12 protein are observed with alteration in H-bonding interactions underlying the protein mis-folding in linker and catalytic domains. In GALNT13, Lys363, and Arg190 are found in interaction with wild-type residue (D313) through H-bonding interactions.

From STRING protein-protein interaction analysis, GALNT3, GALNT8, GALNT11, GALNT12, and GALNT13 proteins are predicted with the strong identical interactions for protein targets like B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1 (Fig. 4A). However, the difference lies for interactions with MUC1, MUC2, MUC7, MUC5B, CHPF, B4GALNT1, MYL4, KLRC3, FGF7, FGF23, SDC3, NUP188, ROCK2, ABO, and MYL4 proteins. Except GALNT3, all other members of GALNTs proteins are found with weak association with CHPF. GALNT3 is observed with weak contact lines for FGF7 and ABO proteins. GALNT8 is seen in weak interaction

pattern for MYL4 and KLRC3. GALNT11 is observed with weak association links for ROCK2, NUP188, ZNF804B, and CEP152. MUC1 and MUC5B are annotated with weak interaction behavior for GALNT12. GALNT13 is indicated the weak interactions with SDC3, SGCB and XIRP2 proteins. The STRING interaction result is further validated by using KEGG pathways for GALNTs. From KEGG results, GALNTs and other enzymes including B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1 are involved in mucin biosynthesis pathway (Fig. 4B). Intriguingly, same set of proteins (B3GNT6, ST6GALNAC1, C1GALT1C1, C1GALT1, and GCNT1) are predicted in strong interaction network for GALNT3, GALNT8, GALNT11, GALNT12, and GALNT13, by STRING database.

Basing on retrospective data about GalNAc detection with FGF23 (in GALNT3-FGF23 complex) and MUC7 (for GALNT12 and GALNT13) substrates, we extended our study to see the flexible fitting of GalNAc residue. From experimental data [Kato et al., 2006], glycosylation site for FGF23 is Thr178, and our finding about hand-in-glove surface contact of GalNAc through flexible docking corroborate the specific involvement of pentad cavity formed by Thr178-FGF23, Arg175-FGF23, His106-FGF23, Asp344-GALNT3, and Gln348-GALNT3 residues. In case of mutant models, this cavity is varied between triad and tetrad binding pockets with the variations in underlying residues (Fig. 6).

In GALNT12-MUC7 docking, MUC7 is visualized in surface contact with Arg373, Asn379, Tyr395, Asn399, Gln473, Ile517, His519, and Leu520 residues, located in catalytic B, linker B, and ricin B-type lectin regions of GALNT12. In literature [Zhang et al., 2003], the GalNAc is specifically transferred to Thr-12th of MUC7 (TTAAPPTPSATTPA). In our results, GALNT12 (wt)-MUC7 complex is shown the identical behavior for GalNAc residue but the surface cavity is equipped with the Thr11-MUC7, Pro13-MUC7, Ile517-GALNT12, and His519-GALNT12 (Fig. 7). In mutant models, the binding pattern of amino acids of GALNT12 is changed to regulate the different fitting behavior of GalNAc residue on MUC7. For R297W, Y396C, and T491M containing mutant models, improper unbound conformation of GalNAc is observed with MUC7, while the D303N is visualized with almost same behavior as shown by wild-type GALNT12.

For GALNT13, MUC7 tandem repeat region (TTAAPPTPSATTPA) is observed as an acceptor peptide for GalNAc residue, and a striking difference in substrate specificity is seen for both wild-type and mutant models (D313N and R476Q) of GALNT13, possessing MUC7 with different binding pockets (Fig. 8). In case of D313N, GalNAc residue is found in surface fitting with 10th Ala of tandem repeat region instead of 12th Thr. whereas, for R476Q, GalNAc residue is fitted within the tetrad (Asn263, Asn319, Asn365, and Tyr300) cavity of GALNT13 instead of interacting the 12th Thr of MUC7 (Fig. 8).

## CONCLUSION

This study shows the aberrant effects of missense variants affecting the glycosyltransferase property of GALNT3, GALNT8, GALNT12, and GALNT13 enzymes. The mutant GALNTs structures may regulate the improper or incomplete glycosylation of protein targets (Fig. 4) in different tissues, and may be the cause of variable manifestations

including hyperphosphatemia, dental anomalies, calcification of bones and eyelids, and different cancer diseases. In our study, intriguing findings propose the main contribution of T359K variant in a compound mutation (T272K and T359K) characterized in tumoral calcinosis, and the role of C574G in regulation of both tumoral calcinosis and hyperostosis–hyperphosphatemia syndrome. Similar to the T359K and C574G of *GALNT3*, the prediction of R297W-*GALNT12* variant with highly pathogenic effects suggests the possible reduction in enzymatic activity of *GALNT12*, hence, the improper glycosylation of target proteins (including MUC1, MUC5B, and MUC7) may regulate the different cancer diseases including bilateral breast cancers as well as primary colon cancer. Additionally, pathogenic variants (D313N and R476Q) of *GALNT13* are predicted with high damaging effects that can modify expression and function of *GALNT13* enzyme to modulate the aberrant glycosylation of MUC1 and MUC7. Finally, annotation of binding cavities for GalNAc residue on *GALNT3*-FGF23 and *GALNT12*-MUC7 complexes as a result of unbound surface fitting may be helpful to further generate the stereochemical reaction chemistry associated with different diseases and disorders.

## REFERENCES

- Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet Chapter 7: Unit7.20*.
- Alanazi M, Abduljaleel Z, Khan W, Warsy AS, Elrobb M, Khan Z, Al Amri A, Bazzi MD. 2011. In silico analysis of single nucleotide polymorphism (SNPs) in human beta-globin gene. *PLoS ONE* 6:20.
- Argueso P, Tisdale A, Mandel U, Letko E, Foster CS, Gipson IK. 2003. The cell-layer- and cell-type-specific distribution of GalNAc-transferases in the ocular surface epithelia is altered during keratinization. *Invest Ophthalmol Vis Sci* 44:86–92.
- Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38:W529–W533.
- Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. 2004. ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res* 32:D120–D121.
- Cheftetz I, Heller R, Galli-Tsinopoulou A, Richard G, Wollnik B, Indelman M, Koerber F, Topaz O, Bergman R, Sprecher E, Schoenau E. 2005. A novel homozygous missense mutation in FGF23 causes Familial Tumoral Calcinosis associated with disseminated visceral calcification. *Hum Genet* 118:261–266.
- Clarke E, Green RC, Green JS, Mahoney K, Parfrey PS, Younghusband HB, Woods MO. 2012. Inherited deleterious variants in *GALNT12* are associated with CRC susceptibility. *Hum Mutat* 33:1056–1058.
- de Alencar SA, Lopes JC. 2010. A comprehensive in silico analysis of the functional and structural impact of SNPs in the IGF1R gene. *J Biomed Biotechnol* 2010:1–8.
- Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. 2013. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41:D808–D815.
- Fritz TA, Hurley JH, Trinh LB, Shiloach J, Tabak LA. 2004. The beginnings of mucin biosynthesis: The crystal structure of UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-T1. *Proc Natl Acad Sci USA* 101:15307–15312.
- Gendler SJ. 2001. MUC1, the renaissance molecule. *J Mammary Gland Biol Neoplasia* 6:339–353.
- Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* 38:W695–W699.
- Guda K, Moinova H, He J, Jamison O, Ravi L, Natale L, Lutterbaugh J, Lawrence E, Lewis S, Willson JK, Lowe JB, Wiesner GL, Parmigiani G, Barnholtz-Sloan J, Dawson DW, Velculescu VE, Kinzler KW, Papadopoulos N, Vogelstein B, Willis J, Gerken TA, Markowitz SD. 2009. Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci USA* 106:12921–12925.
- Guex N, Peitsch MC, Schwede T. 2009. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis* 30:S162–S173.
- Guo JM, Zhang Y, Cheng L, Iwasaki H, Wang H, Kubota T, Tachibana K, Narimatsu H. 2002. Molecular cloning and characterization of a novel member of the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase family, pp-GalNAc-T12. *FEBS Lett* 524:211–218.
- Hagen FK, Hazes B, Raffo R, deSa D, Tabak LA. 1999. Structure–function analysis of the UDP-N-acetyl-d-galactosamine:polypeptide N-acetylgalactosaminyltransferase. Essential residues lie in a predicted active site cleft resembling a lactose repressor fold. *J Biol Chem* 274:6797–6803.
- Hussain MRM. 2012. The role of galactose in human health and disease. *Cent Eur J Med* 7:11.
- Hussain MR, Shaik NA, Al-Aama JY, Asfour HZ, Khan FS, Masoodi TA, Khan MA, Shaik NS. 2012. In silico analysis of single nucleotide polymorphisms (SNPs) in human BRAF gene. *Gene* 508:188–196.
- Ichikawa S, Lyles KW, Econs MJ. 2005. A novel *GALNT3* mutation in a pseudoautosomal dominant form of tumoral calcinosis: Evidence that the disorder is autosomal recessive. *J Clin Endocrinol Metab* 90:2420–2423.
- Ichikawa S, Imel EA, Sorenson AH, Severe R, Knudson P, Harris GJ, Shaker JL, Econs MJ. 2006. Tumoral calcinosis presenting with eyelid calcifications due to novel missense mutations in the glycosyl transferase domain of the *GALNT3* gene. *J Clin Endocrinol Metab* 91:4472–4475.
- Ichikawa S, Baujat G, Seyahi A, Garoufali AG, Imel EA, Padgett LR, Austin AM, Sorenson AH, Pejini Z, Topouchian V, Quartier P, Cormier-Daire V, Dechaux M, Malandrinou F, Singhellakis PN, Le Merrer M, Econs MJ. 2010. Clinical variability of familial tumoral calcinosis caused by novel *GALNT3* mutations. *Am J Med Genet A* 152A:896–903.
- Ju T, Otto VI, Cummings RD. 2011. The Tn antigen-structural simplicity and biological complexity. *Angew Chem Int Ed Engl* 50:1770–1791.
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res* 34:D354–D357.
- Kato K, Jeanneau C, Tarp MA, Benet-Pages A, Lorenz-Depiereux B, Bennett EP, Mandel U, Strom TM, Clausen H. 2006. Polypeptide GalNAc-transferase T3 and familial tumoral calcinosis. Secretion of fibroblast growth factor 23 requires O-glycosylation. *J Biol Chem* 281:18370–18377.
- Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremiex O, Campbell MJ, Kitano H, Thomas PD. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 33:D284–D288.
- Hussain MRM, Asfou H, Yasir M, Khan A, Mohamoud HSA, Al-Aamaa JY. 2013. The microbial pathology of Neu5Ac and Gal epitopes. *J Carbohydr Chem* 32:15.
- Nakano R, Maekawa T, Abe H, Hayashida Y, Ochi H, Tsunoda T, Kumada H, Kamatani N, Nakamura Y, Chayama K. 2013. Single-nucleotide polymorphisms in *GALNT8* are associated with the response to interferon therapy for chronic hepatitis C. *J Gen Virol* 94:81–89.
- Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814.
- Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 7:61–80.

- Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 9:51.
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612.
- Phelan CM, Tsai YY, Goode EL, Vierkant RA, Fridley BL, Beesley J, Chen XQ, Webb PM, Chanock S, Cramer DW, Moysich K, Edwards RP, Chang-Claude J, Garcia-Closas M, Yang H, Wang-Gohrke S, Hein R, Green AC, Lissowska J, Carney ME, Lurie G, Wilkens LR, Ness RB, Pearce CL, Wu AH, Van Den Berg DJ, Stram DO, Terry KL, Whitman DC, Whittemore AS, DiCioccio RA, McGuire V, Doherty JA, Rossing MA, Anton-Culver H, Ziogas A, Hogdall C, Hogdall E, Kruger Kjaer S, Blaakaer J, Quaye L, Ramus SJ, Jacobs I, Song H, Pharoah PD, Iversen ES, Marks JR, Pike MC, Gayther SA, Cunningham JM, Goodman MT, Schildkraut JM, Chenevix-Trench G, Berchuck A, Sellers TA. 2010. Polymorphism in the GALNT1 gene and epithelial ovarian cancer in non-Hispanic white women: The Ovarian Cancer Association Consortium. *Cancer Epidemiol Biomarkers Prev* 19:600–604.
- Potapenko IO, Haakensen VD, Luders T, Helland A, Bukholm I, Sorlie T, Kristensen VN, Lingjaerde OC, Borresen-Dale AL. 2010. Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression. *Mol Oncol* 4:98–118.
- Raman J, Guan Y, Perrine CL, Gerken TA, Tabak LA. 2012. UDP-*N*-acetyl-alpha-d-galactosamine:polypeptide *N*-acetylgalactosaminyltransferases: Completion of the family tree. *Glycobiology* 22:768–777.
- Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat Protoc* 5:725–738.
- Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363–W367.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
- Tenno M, Saeki A, Kezdy FJ, Elhammer AP, Kurosaka A. 2002. The lectin domain of UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase 1 is involved in *O*-glycosylation of a polypeptide with multiple acceptor sites. *J Biol Chem* 277:47088–47096.
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13:2129–2141.
- Topaz O, Shurman DL, Bergman R, Indelman M, Ratajczak P, Mizrahi M, Khamaysi Z, Behar D, Petronius D, Friedman V, Zelikovic I, Raimer S, Metzker A, Richard G, Sprecher E. 2004. Mutations in GALNT3, encoding a protein involved in *O*-linked glycosylation, cause familial tumoral calcinosis. *Nat Genet* 36:579–581.
- Trott O, Olson AJ. 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multi-threading. *J Comput Chem* 31:455–461.
- Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. 2010. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 11:548.
- Walsh MD, Clendenning M, Williamson E, Pearson SA, Walters RJ, Nagler B, Packenas D, Win AK, Hopper JL, Jenkins MA, Haydon AM, Rosty C, English DR, Giles GG, McGuckin MA, Young JP, Buchanan DD. 2013. Expression of MUC2, MUC5AC, MUC5B, and MUC6 mucins in colorectal cancers and their association with the CpG island methylator phenotype. *Mod Pathol*. 2013:1–15.
- White KE, Lorenz B, Evans WE, Meitinger T, Strom TM, Econs MJ. 2000. Molecular cloning of a novel human UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferase, GalNAc-T8, and analysis as a candidate autosomal dominant hypophosphatemic rickets (ADHR) gene. *Gene* 246:347–356.
- Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezzo Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
- Yoshimura Y, Matsushita T, Fujitani N, Takegawa Y, Fujihira H, Naruchi K, Gao XD, Manri N, Sakamoto T, Kato K, Hinou H, Nishimura S. 2010. Unexpected tolerance of glycosylation by UDP-GalNAc:polypeptide alpha-*N*-acetylgalactosaminyltransferase revealed by electron capture dissociation mass spectrometry: Carbohydrate as potential protective groups. *Biochemistry* 49:5929–5941.
- Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:1471–2105.
- Zhang Y, Iwasaki H, Wang H, Kudo T, Kalka TB, Hennes T, Kubota T, Cheng L, Inaba N, Gotoh M, Togayachi A, Guo J, Hisatomi H, Nakajima K, Nishihara S, Nakamura M, Marth JD, Narimatsu H. 2003. Cloning and characterization of a new human UDP-*N*-acetyl-alpha-d-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase, designated pp-GalNAc-T13, that is specifically expressed in neurons and synthesizes GalNAc alpha-serine/threonine antigen. *J Biol Chem* 278:573–584.