

**COMPUTATIONAL ANALYSIS OF GLYCOGENES FROM
MOUSE RNA-SEQ DATA**

Submitted in the partial fulfillment of the requirement
for the degree of

DOCTOR OF PHILOSOPHY

by

**Ahmad Firoz
(Regn.No. 901009008)**



**SCHOOL OF CHEMISTRY AND BIOCHEMISTRY
THAPAR UNIVERSITY
PATIALA, PUNJAB-147004, INDIA.
DECEMBER-2015**

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my guide and mentor Dr. Amjad Ali, Associate Professor, School of Chemistry and Biochemistry, Thapar University, Patiala, for giving me an opportunity to complete my thesis under his supervision. He has always been a pillar of support and constant source of inspiration. This thesis would never have been possible without his kind support, constructive criticism continuous advice and valuable guidance.

I am extremely thankful to the Director, Thapar University, Dean (Research & Sponsored Projects), Head, School of Chemistry and Biochemistry for extending the opportunity to undertake this doctoral research.

I am highly obliged to the doctoral committee members Dr. Rajesh Kumar, Dr. Kamaldeep Paul & Dr. Deepak Garg for monitoring my research work from time to time and giving their valuable suggestions. Grateful acknowledgement is made to all faculty members of school of Chemistry and Biochemistry, my colleague & friends.

Dr. Adeel Malik, Assistant Professor of Perdana University Centre for Bioinformatics, Selangor, Malaysia, deserves my sincere expression of thanks for providing valuable suggestion & help.

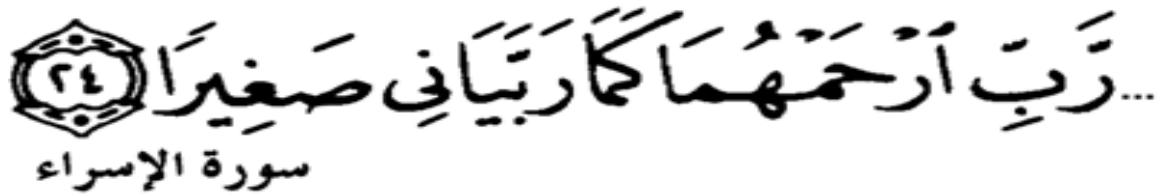
I am very much thankful to Dr. Vivekanand Jha. Professor, Department of Nephrology, PGIMER, Chandigarh, for providing necessary infrastructure and resources to accomplish my research work and valuable suggestion. I take this opportunity to sincerely acknowledge the Biomedical Informatics Centre of ICMR, PGIMER Chandigarh.

I thank Allah, the Almighty for giving me the grace and health to complete my studies. I am also extremely grateful to my parents, grandparents, my brothers, my wife, daughter, nephews, nieces, parents in law, brother & sister in law, relatives and friends for their infinite patience, love and unending support.



Ahmad Firoz

*I dedicate this thesis to my parents, for their
constant support and unconditional love.*



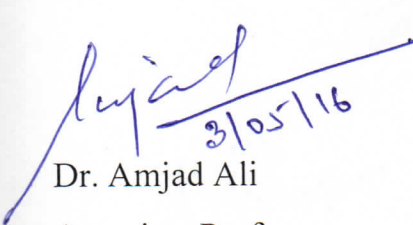
*"My Lord! Bestow on them Your Mercy as they did bring me up
when I was small." [Quran; Surah Al-Isra, Verse 24]*

Transliteration: "Rabbi irhamhuma kama rabbayanee sagheeran"

CERTIFICATE

This is to certify that thesis entitled "COMPUTATIONAL ANALYSIS OF GLYCOGENES FROM MOUSE RNA-SEQ DATA", being submitted by **Ahmad Firoz**, to the School of Chemistry and Biochemistry, Thapar University, Patiala for the award of degree of DOCTOR OF PHILOSOPHY, is a record of bonafide research work carried out by him. Mr. Ahmad Firoz has worked under my guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to my knowledge has reached the requisite standard.

The results embodied in the thesis have not been submitted in part or full to any other University or Institute for the award of any degree or diploma



Dr. Amjad Ali

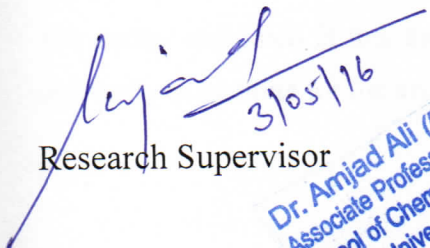
Associate Professor

School of Chemistry and Biochemistry

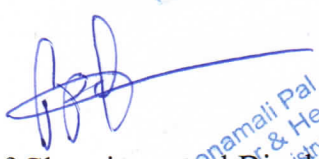
Thapar University, Patiala, India.

THESIS APPROVAL SHEET

Thesis entitled “**COMPUTATIONAL ANALYSIS OF GLYCOGENES FROM MOUSE RNA-SEQ DATA**” by **Ahmad Firoz** is approved for the degree of DOCTOR OF PHILOSOPHY.


Research Supervisor

Dr. Amjad Ali (Ph.D., IITB)
Associate Professor
School of Chemistry and Biochemistry
Thapar University, Patiala-147004


Head,
School of Chemistry and Biochemistry

Dr. Annamali Pal
Professor & Head
School of Chemistry & Biochemistry
Thapar University, Patiala-147004


Dean, Research and Sponsored Projects

Date.....3/05/16.....

Candidate's Declaration

I, hereby declare that the work presented in the thesis entitled "COMPUTATIONAL ANALYSIS OF GLYCOGENES FROM MOUSE RNA-SEQ DATA" in partial fulfillment of the requirement for the award of the degree of Doctor of Philosophy, School of Chemistry and Biochemistry, Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Amjad Ali, Associate Professor, School of Chemistry and Biochemistry, Thapar University, Patiala, India. The matter embodied in this thesis has not been submitted in part or full to any other university or institute for the award of any degree in India or Abroad.



Ahmad Firoz

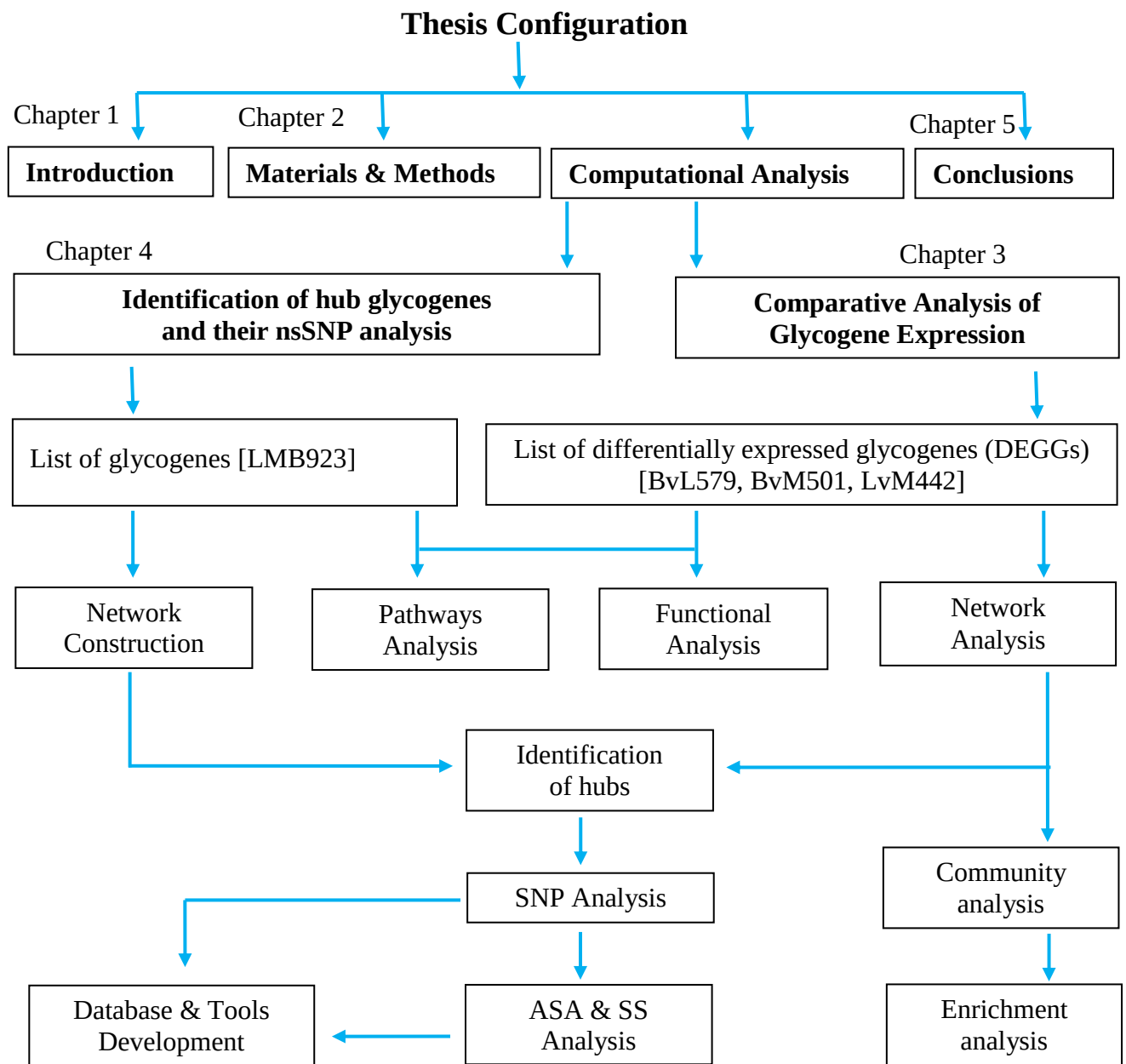
CONTENTS

Chapter	Title	Page number
	List of abbreviations	i
	Thesis configuration	ii
	Abstract	iii
1	Introduction and literature review	
1.1	Introduction	1
1.2	Glycobiology	2
1.3	Glycans	4
1.4	Glycosylation	6
1.4.1	N-glycosylation	7
1.4.2	O-glycosylation	7
1.4.3	C-glycosylation	7
1.4.4	Glypiation	8
1.4.5	Phosphoglycosylation	8
1.5	Glycation	8
1.6	Glycogene	8
1.6.1	Glycogene expression	9
1.7	Glycoinformatics	10
1.8	RNA-Seq technology and benefits	12
1.8.1	Computational challenges for RNA-Seq	14
1.9	Single Nucleotide Polymorphism	15
1.10	Conclusions	16
1.11	Objectives	17
	References	18
2	Materials and methods	
2.1	Datasets	25
2.2	Data Filtering	25
2.3	Fold change analysis	25
2.4	Annotation of differentially expressed glycogenes	26
2.5	Functional analysis	26
2.6	Pathway analysis	27
2.7	Network analysis and identification of hubs	27
2.8	Community and enrichment analysis	27
2.9	Computational analysis of nsSNPs of identified hub genes	28

2.10	Surface accessibility and secondary structure analysis of predicted nsSNPs	29
2.11	Database & Tools development	29
	References	30
3	Comparative analysis of glycogenes expression	
3.1	Introduction	32
3.2	Identification of differentially expressed glycogenes	32
3.3	Functional annotation and pathway analysis	33
3.4	Interaction network and identification of hub Genes.	36
3.5	Community analysis and functional annotation of the detected modules	42
3.6	Discussion	52
	References	56
4	Identification of hub glycogenes and their nsSNP analysis	
4.1	Introduction	63
4.2	Identification of glycogenes	63
4.3	Functional enrichment and pathway analysis	63
4.4	Analysis of interaction network and hub genes	72
4.5	SNP analysis	75
4.5.1	Deleterious or damaged nsSNPs by PROVEAN	76
4.5.2	Predicted nsSNPs that affect protein function by SIFT	77
4.5.3	Damaged nsSNPs by PolyPhen-2	77
4.6	Protein surface accessibility and secondary structure prediction by NetSurfP	79
4.7	Database and Tools development	82
4.8	Discussion	85
	References	86
5	Conclusions and futuristic aspect	
5.1	Introduction	89
5.2	Conclusions from the present work	89
5.3	Futuristic aspects	92
	List of publications	93

List of Abbreviations

Abbreviation	Description
RNA-Seq	RNA sequencing
DEGGs	Differentially expressed glycogenes
DAVID	Database for Annotation, Visualization and Integrated Discovery
GO	Gene ontology
KEGG	Kyoto Encyclopaedia of Genes and Genomics
KAAS	KEGG Automatic Annotation Server
KO	KEGG orthology
SNPs,	Single nucleotide polymorphisms
nsSNP	Non synonymous single nucleotide polymorphism
SBH	Single-directional best hit
BLAST	Basic Local Alignment Search Tool
SLC2A4	Solute carrier family 2
TNFRSF1B	Tumor necrosis factor receptor super family member 1B
TRFR2	Transferrin receptor protein 2
TNF	Tumor Necrosis Factor
UCHL1	Ubiquitin carboxylterminal hydrolase isozyme L1
HPXN	Hemopexin
NID2	Nidogen-2
FLT1	Vascular endothelial growth factor receptor 1
PROVEAN	Protein Variation Effect Analyzer
SIFT	Sorting Intolerant From Tolerant
PolyPhen-2	Polymorphism Phenotyping v2
ASA	Accessible surface area
FAC	Functional Annotation Clustering
MYOGkd	Myogenin knock-down



Abstract

Glycogenes regulate a wide array of biological processes in the development of organisms as well as different diseases such as cancer, primary open-angle glaucoma, and renal dysfunction. The objective of this study was to explore the role of differentially expressed glycogenes (DEGGs) in three major tissues such as brain, muscle, and liver using mouse RNA-seq data, and we identified 579, 501, and 442 DEGGs for brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups. DAVID functional analysis suggested inflammatory response, glycosaminoglycan metabolic process, and protein maturation as the enriched biological processes in BvL579, BvM501, and LvM442, respectively. These DEGGs were then used to construct three interaction networks by using GeneMANIA, from which we detected potential hub genes such as PEMT and HPXN (BvL579), IGF2 and NID2 (BvM501), and STAT6 and FLT1 (LvM442), having the highest degree. Community analysis of DEGGs suggests the significance of immune system related processes in liver, glycosphingolipid metabolic processes in the development of brain, and the processes such as cell proliferation, adhesion, and growth are important for muscle development. Additionally, an interaction network of non-redundant list of 923 glycogenes was also created by combining all identified glycogenes of brain, muscle and liver tissues, to detect potential hubs from different mouse tissues. SLC2A4, TNFRSF1B, TRFR2, and UCHL1 were identified as hub genes by calculating the node degree distribution using Network Analyzer plugin of Cytoscape. We also explored nsSNPs that may modify the expression and function of identified hubs using computational methods. We observe that the number of nsSNPs predicted by any two methods to affect protein function is 4, 7 and 2 for FLT1, NID2 and TNFRSF1B. Residues in the

native and mutant proteins were analyzed for solvent accessibility and secondary structure change. Analysis of hubs can help in determining their degree of conservation and understanding their functions in biological processes. Further studies are required to confirm the role of these predicted hub genes as well as the significance of biological processes. The nsSNPs proposed in this work may also be further targeted through experimental methods for understanding structural and functional relationships of hub mutants.

Keywords: glycogenes, gene expression, RNA-Seq, DAVID, hubs, interaction network, nsSNPs, solvent accessibility, secondary structure, mouse, liver, brain and muscle

Introduction and literature review

	Contents	Page number
1.1	Introduction	1
1.2	Glycobiology	2
1.3	Glycans	4
1.4	Glycosylation	6
1.4.1	N-glycosylation	7
1.4.2	O-glycosylation	7
1.4.3	C-glycosylation	7
1.4.4	Glypiation	8
1.4.5	Phosphoglycosylation	8
1.5	Glycation	8
1.6	Glycogene	8
1.6.1	Glycogene Expression	9
1.7	Glycoinformatics	10
1.8	RNA-Seq technology and benefits	12
1.8.1	Computational Challenges for RNA-Seq	14
1.9	Single Nucleotide Polymorphism	15
1.10	Conclusions	16
1.11	Objectives	17
	References	18

Abstract

A brief introduction about glycogenes and the pattern of its expression that provide clues about many biological functions, developmental changes and diseases in human, mouse and tissues of other organisms, high-throughput technology such as RNA-Seq experiment to investigate the functional roles of various glycogenes, and involvement of informatics to develop computational methods, tools as well as the construction of databases in order to specifically explore the role of glycogenes has been discussed in this chapter.

Keywords: Gene Expression, Glycogenes, RNA-Seq, Mouse, Glycans.

1.1 Introduction

Glycogenes are the genes that are responsible for the glycosylation of proteins, lipids, proteoglycans, and are associated with the synthesis of glycans such as glycosyltransferase, sugar-nucleotide transporters, sulfotransferases, etc (Taniguchi *et al.*, 2009). Cell function within an organism has huge variation due to gene expression pattern although all cells in an individual mammal have almost identical DNA (Salzman *et al.*, 2011). It is important to know that how cells and tissues differ in gene expression which is regulated during developmental changes in different tissues, consequently affecting their biological functions. The pattern of glycogene expression is one of important factors that provide clues about many biological functions, developmental changes and diseases in human, mouse and tissues of other organisms (Mantelli *et al.*, 2010; Ishii *et al.*, 2007; Nakayama *et al.*, 2013). A large variety of biological processes during the growth and development of organisms in addition to wide array of diseases such as cancer, primary open-angle glaucoma, and renal dysfunction, are regulated by glycogenes (Salzman *et al.*, 2011; Mantelli *et al.*, 2010). Alteration in the glycogene expression is also associated with drug resistance of various types of cancers including pancreatic cancer epithelial-mesenchymal transition (Maupin *et al.*, 2010), breast cancer (Ma *et al.*, 2013), and hepatocarcinoma (Guo *et al.*, 2013). The change in the glycogene expression is also associated with conjunctival epithelium (Mantelli *et al.*, 2010) and primary open-angle glaucoma (POAG) (Diskin *et al.*, 2006). Moreover, the defects in the normal expression of glycogenes in tonsillar B lymphocytes are also correlated with the proteinuria and renal dysfunction in IgA nephropathy (Inoue *et al.*, 2010; Inoue *et*

al., 2011). In general, expression of glycogenes is comparatively weak in comparison to other molecules; yet, some glycogenes are upregulated during particular circumstances (<http://www.stelic.com/cn27/pg122.html>). The sequencing of expressed RNAs with NGS approaches is abbreviated as RNA-Seq. Although several high-throughput microarray studies have been applied to investigate the functional roles of various glycogenes (Saravanan *et al.*, 2010; Kroes *et al.*, 2007; Tan *et al.*, 2014). However, to the best of our knowledge, previously no study has utilized RNA-Seq data in the existing databases to specifically explore the role of glycogenes. Therefore, in this work, we intend to investigate the role of glycogenes in various biological processes that are involved in the development of brain, muscle, and liver tissues using the precomputed expression values from mouse RNA-Seq data.

1.2 Glycobiology

The flow of genetic information, or gene expression, in normal cells is transcription of DNA into RNA and translation of RNA into amino acids to make proteins (Lewin, *et al.*, 2004). In fact, other two major classes of molecules such as lipids and carbohydrates are also required for creating a cell. These molecules can serve as intermediates in generating energy, as signaling molecules, or as structural components. Carbohydrates are often referred to as the third molecular chain of life, after DNA and proteins (Mitsuyama *et al.*, 2003). Living cells in all organisms are usually covered with one or another type of carbohydrate. Some viruses like influenza, HIV use sugars on the outside of human cells to gain entry. (<http://www.scripps.edu/news/press/2004/120704.html>) The functional roles of carbohydrates and their interactions with proteins are drawing more attention than before, since it has been recognized that carbohydrates are used as information carriers, rather than simple storage material (Mitsuyama *et al.*, 2003). and involved in broad range of biological processes include, infection by invading microorganisms and the subsequent immune response, leukocytic trafficking and infiltration, and tumor metastasis (Barondes *et al.*, 1994; Karlsson *et al.*, 1998). Altered carbohydrate and glycoprotein structures as compared to normal healthy cells are present in people with autoimmune conditions (www.deathtodisabetes.com/Glyconutrients_Glycobiology.html#.VmZ1i3YrLIo). Simply, highly dynamic protein-bound glycans are abundant in the nucleus and cytoplasm, where they appear to serve as regulatory switches. The development of a variety of new technologies for exploring the structures of these

glycan has opened up a new frontier of molecular biology which is called glycobiology. Thus an extended paradigm of molecular biology can be rendered as shown in Figure 1.1 as below:-

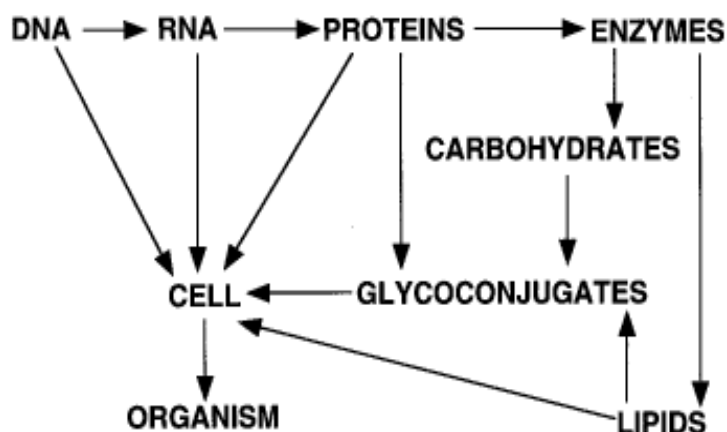


Figure 1.1: (Source: Varki *et al.*, 1999)

The four essential molecular building blocks of cells are nucleic acids, proteins, lipids, and carbohydrates, often also called glycans. Mostly nucleic acids and proteins are the heart of research; whereas glycan and lipid structures have been neglected. The glycobiology is study of understanding the structure, biosynthesis, biological function of glycans and their derivatives. The term glycobiology to recognize a new discipline coming together of the traditional disciplines of carbohydrate chemistry and biochemistry was coined by Prof. Raymond Dwek in 1988 (). Glycobiologists have several problems, making their research difficult and time consuming. The first major challenge is to develop highly sensitive analytical methods to analyze physiological Concentrations of glycans; there are currently no methods available to amplify glycans although DNA is amplified using polymerase chain reaction (PCR) techniques. Mass spectrometry, High Performance Liquid Chromatography (HPLC) and Nuclear magnetic resonance (NMR) are used in different areas of glycobiology to identify glycan structures. Another major challenge also lies in the complexity of glycan structures (www.deathtodibet.com/Glyconutrients_Glycobiology.html#.VmZ1i3YrLIo).

1.3 Glycans

Carbohydrates often also called glycans or saccharides are the most abundant and diverse of Nature's biopolymers. It fulfills numerous functions in nature. Protein glycosylation, the covalent attachment of oligosaccharides to proteins, is the most frequent co- and post-translational modification (PTM). Glycans play important role in many biological processes such as cell differentiation, signal transduction, cancer metastasis, immune responses, cell adhesion, pathogen-host recognition, toxin-receptor interactions and modulating the functions of many proteins and lipids through post-translational modifications (Hart GW *et al.*, 2007; Hashimoto *et al.*, 2010). All cells in nature are covered with a dense and complex coat of glycans.

Marked changes can occur in glycan structure during development, cellular activation, differentiation, cancer, and inflammation (Rademacher *et al.*, 1988). The great deal of information about expression of glycan-related genes, glycan structures and functions can be obtained through recent emerging analytical technique (Hashimoto *et al.*, 2010). Glycans basic structural units, Monosaccharides, that can be hydrolyzed into simpler unit are aldehydes or ketones that have two or more hydroxyl groups. The empirical formula of many is $(C-H_2O)_n$. Monosaccharides are also important fuel molecules and building blocks for nucleic acids.

Besides monosaccharides, two common disaccharides, lactose and sucrose, occur naturally (**Figure 1.2**). A disaccharide consists of two monosaccharides linked together by a C—O—C bridge called a glycosidic bond. The lactose is the major sugar in milk and sucrose is a common table sugar.

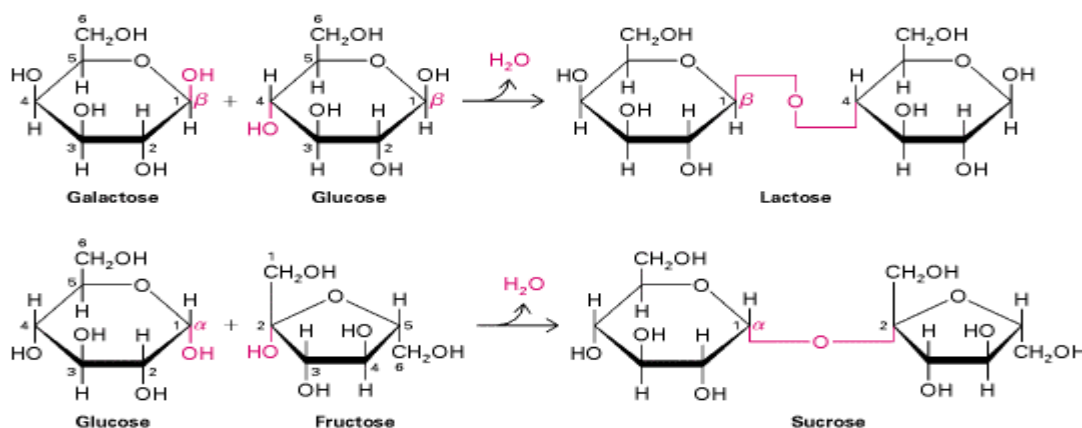


Figure 1.2

Figure 1.2: The formation of glycosidic linkages in disaccharides lactose and sucrose. The lactose linkage is $\beta(1 \rightarrow 4)$; the sucrose linkage is $\alpha(1 \rightarrow 2)$. In any glycosidic linkage, carbon 1 of one sugar molecule (in either α or β conformation) is linked to a hydroxyl oxygen on another sugar molecule. (Source: Lodish *et al.*, 1999)

In any glycosidic bond formation, the carbon 1 atom of one sugar molecule reacts with a hydroxyl group of another. The linkage is accompanied by the loss of water in the formation of most biopolymers. A large number of different glycosidic bonds can be formed between two sugar residues. Glucose could be bonded to fructose, for example, by any of the following linkages: $\alpha(1 \rightarrow 1)$, $\alpha(1 \rightarrow 2)$, $\alpha(1 \rightarrow 3)$, $\alpha(1 \rightarrow 4)$, $\alpha(1 \rightarrow 6)$, $\beta(1 \rightarrow 1)$, $\beta(1 \rightarrow 2)$, $\beta(1 \rightarrow 3)$, $\beta(1 \rightarrow 4)$, or $\beta(1 \rightarrow 6)$, where α or β specifies the conformation at carbon 1 in glucose and the number following the arrow indicates the fructose carbon to which the glucose is bound. Only the $\alpha(1 \rightarrow 2)$ linkage occurs in sucrose because of the specificity of the enzyme (the biological catalyst) for the linking reaction (Lodish *et al.*, 1999).

Similarly, one or more monosaccharide or oligosaccharide units (the glycone) are covalently linked to a non carbohydrate moiety (the aglycone) to form glycoconjugate. The portion of the molecule comprising the glycans can vary greatly in naturally occurring glycoconjugates. Indeed, the sugar chains make up a substantial portion of the mass of most glycoconjugates. Due to this reason, the surfaces of most types of cells (which are heavily decorated with different kinds of glycoconjugates) are effectively covered with a dense coating of sugars, giving rise to the so-called glycocalyx. (Varki *et al.*, 1999).

Same glycan can have different roles within an organism (Varki *et al.*, 2009). The expression of specific types of glycosylation on different glycoconjugates in different tissues at different times of development implies that these structures have diverse roles within the same organism. For example, Mannose-6-phosphatecontaining oligosaccharides were first found on lysosomal enzymes and are involved in lysosomal trafficking (Hoflack *et al.*, 1987). However, Mannose-6-phosphatecontaining glycans are found on a variety of apparently unrelated proteins, including proliferin, thyroglobulin, the Epidermal Growth Factor (EGF) receptor, and the Transforming Growth Factor beta (TGF- β) precursor. Likewise, the sialylated fucosylated lactosamines critical for selectin recognition (Fukuda *et al.*, 1992) are found in a variety of unrelated cell types, and the polysialic acid chains that play such an important

part in embryonic N-CAM, are found on egg jelly coat proteins and a sodium channel protein (Roth *et al.*, 1993).

Proteins and enzymes that bind glycan are present in large numbers in all living cells and are involved in important biological phenomena such as autoimmune disorders, host-parasite interactions, fertilization, and cellular differentiation (Dwek, *et al.*, 1996). In order to decode the specific molecular signal presented by a carbohydrate epitope, carbohydrate-binding proteins must present a non-catalytic binding site complimentary to the carbohydrate. The immunoglobulins employ their complimentary-determining regions (CDR) for carbohydrate recognition during the immune response. Other carbohydrate-recognizing proteins are known as the lectins, a diverse group of unrelated proteins that employ a binding site referred to as the carbohydrate recognition domain (CRD) (Audette *et al.*, 2003).

O and N-linked glycans are common in eukaryotes but may also be found less in prokaryotes. N-Linked glycans are attached in the endoplasmic reticulum to the nitrogen (N) in the side chain of asparagine in the sequon, and are derived from a core 14-sugar unit assembled in the cytoplasm and endoplasmic reticulum in eukaryotes. N-linked glycans are extremely important in proper protein folding in eukaryotic cells (Helenius *et al.*, 2004), and also in cell-cell interactions. For example, tumour cells make N-linked glycans that are abnormal. These are recognized by the CD337 receptor on Natural Killer cells as a sign that the cell in question is cancerous. The targeting of degradative lysosomal enzymes is also linked to N-linked glycans. O-linked glycans are attached to the hydroxyl oxygen of serine, threonine, tyrosine, hydroxylysine, or hydroxyproline side-chains, or to oxygens on lipids such as ceramide. O-linked glycans, mucin, have been found to be important in developing normal intestinal microflora. Phospho-glycans linked through the phosphate of a phospho-serine, C-linked glycans, a rare form of glycosylation where a sugar is added to a carbon on a tryptophan side-chain and glypiation, which is the addition of a GPI anchor that links proteins to lipids through glycan linkages. Glycosyltransferases are important a biological catalyst generates complex cell surface glycan involved in adhesion and signalling processes (<http://www.ncbi.nlm.nih.gov/books/NBK1897/>).

1.4 Glycosylation

Glycosylation involves enzyme controlled selective attachment of carbohydrate molecules to proteins and lipids. It is one of the most important and complex co and

post translational modifications occurring in cells, and is involved in various biological processes. It plays various critical roles in protein function (Taniguchi *et al.*, 2015). More than 50% of proteins are glycosylated. Types of glycosylation are classified according to the atom identity of the amino acid which binds the carbohydrate chain [<http://www.uniprot.org/help/carbohyd>]. There are two major types of glycosylation known as N-glycosylation and O-glycosylation. Another three unusual glycosylation forms are C-glycosylation, Glypiation and Glycation (Mazola *et al.*, 2011). Glycosylation of proteins occurs in all eukaryotes, many archaea and exceptionally in bacteria. Glycosylation site occupancy analysis is usually carried out by mass spectrometry (Herget *et al.*, 2009).

1.4.1 N-glycosylation

N-glycosylation occurs on secreted or membrane bound proteins and refers to attachment of a sugar moiety to the amide side chain of asparagine residue usually at the N4 atom within any of the consensus sequences Asparagine (Asn)-X- Serine/Threonine (Ser/Thr), where X is not Proline (Pro) although Threonine (Thr) is more common than Serine (Ser), more rarely above consensus sequences may be Asparagine (Asn)-X- Cysteine (Cys). These tripeptide sequences are known as sequon (Mazola *et al.*, 2011; <http://www.uniprot.org/help/carbohyd>). In eukaryotes, N- glycosylation begins in the endoplasmic reticulum by transferring the blocks of 14 sugars 2 N-acetylglucosamines, 9 mannoses and 3 glucoses (Glc3Man9GlcNAc2) to the protein. Next, after cleavage of the 3 glucose, one mannose residue the protein is transported to the Golgi apparatus for other mannose trimmings. Finally, the Man5GlcNAc2-Asn linked protein is the starting point to generate huge N-glycan types in the Golgi apparatus (<http://www.ncbi.nlm.nih.gov/books/NBK20709/>).

1.4.2 O-glycosylation

O-glycosylation is post translational event occur in Golgi and usually in mucin domains. The regions containing large numbers of sequential serine, threonine and proline residues known as mucin domains. These regions show little secondary structure. O-linked glycans are important in protein localization and trafficking, protein solubility, antigenicity and cell-cell interactions (<http://www.ncbi.nlm.nih.gov/books/NBK20709/>).

1.4.3 C-glycosylation

C-glycosylation refers to the covalent attachment of a mannose sugar to carbon on a tryptophan side-chain in the sequence W-X-X-W or W-S/T-X-C (W indicates

tryptophan; X is any amino acid). Where, first or both tryptophan residues become mannosylated. (<http://www.uniprot.org/help/carbohydr>)

C-glycosylation is unusual because the sugar is linked to a carbon rather than nitrogen or oxygen. Most common C-modified proteins are Thrombospondins (<http://www.ncbi.nlm.nih.gov/books/NBK20709/>).

1.4.4 Glypiation

This special type of glycosylation is the formation of glycosyl phosphatidyl inositol (GPI) anchor, where protein is attached to a lipid through glycan linkages. GPI anchors consist of Phosphoethanolamine linker that binds to the C-terminus of target proteins, Glycan core structure, and Phospholipid tail that anchors the structure in membrane. This glycosylation is widely detected on surface glycoproteins in eukaryotes and some archae (Kobayashi *et al.*, 1997).

1.4.5 Phosphoglycosylation

This type of glycosylation is limited to parasites and is characterized by the attachment of glycans to serine or threonine via phosphodiester bonds. phosphoglycosylation occurs by transfer of a prefabricated phosphoglycan from a membrane-bound molecule via a phosphoglycosyltransferase (PTase). The first reported example of phosphoglycosylation was an endopeptidase known as Proteinase I, isolated from the cellular slime mold *Dictyostelium discoideum* (Paul *et al.*, 1998).

1.5 Glycation

This is also called non-enzymatic glycosylation. Glycation refers to the non-enzymatic covalent bonding of protein or lipid molecule with sugars molecule, such as glucose or fructose. Glycation may occur either inside or outside the body. Over time, the sugar moieties attached to glycated proteins gradually modified to become Advanced Glycation Endproducts (AGEs), some of which are implicated in a variety of diseases, such as cancer, atherosclerosis, type II diabetes mellitus, Parkinson disease and Alzheimer disease (<http://www.uniprot.org/help/carbohydr>).

1.6 Glycogene

The central dogma of genetics indicated with the term “gene” a DNA portion whose corresponding mRNA encodes a protein. According to this view, RNA was considered a “bridge” in the transfer of biological information between DNA and proteins. Gene that is responsible for the glycosylation of proteins, lipids, proteoglycans and associated with the synthesis of glycans such as glycosyltransferase, sugar-nucleotide transporters,

sulfotransferases is called glycogene (Taniguchi *et al.*, 2009; <http://www.stelic.com/cn27/pg122.html>). The term glycogenome involves all genes that play a role in glycosylation, represents more than 600 glycogenes in the mouse genome (Janot *et al.*, 2009) and over 180 glycogenes in human which were identified, cloned and characterized (Narimatsu *et al.*, 2004).

1.6.1 Glycogene Expression

Cell function within an organism has huge variation due to gene expression pattern although all cells in an individual mammal contain nearly duplicate DNA (Salzman *et al.*, 2011) It is important to know how cells and tissues differ in gene expression which is regulated during developmental changes in different tissues, consequently affecting their biological functions. A large variety of biological processes during the growth and development of organisms in addition to wide array of diseases such as cancer, primary open-angle glaucoma, and renal dysfunction, are regulated by glycogenes. Alteration in the glycogene expression is also associated with drug resistance of various types of cancers (Maupin *et al.*, 2010) Expression pattern of these glycogenes is one of the essential factors that indicate about several biological functions, developmental changes, and diseases in human, mouse, and tissues of other organisms (Mantelli *et al.*, 2009; Ishii *et al.*, 2007; Nakayama *et al.*, 2013). One of the most abundant types of posttranslational modification in biological systems is glycosylation (Zhang *et al.*, 2013) and all genes that play essential roles in glycosylation represent the glycogenome (Janot *et al.*, 2009). The glycogenome comprises of genes involved in glycosylation of proteins, lipids, and proteoglycans and also consists of genes related to synthesis of glycans such as glycosyltransferases, sugar-nucleotide transporters, and sulfotransferases (Janot *et al.*, 2009; Taniguchi *et al.*, 2009; <http://www.stelic.com/cn27/pg122.html>). In general, expression of glycogenes is comparatively weak in comparison to other molecules; yet, some glycogenes are upregulated during particular circumstances [<http://www.stelic.com/cn27/pg122.html>]. Functional roles of glycogenes have been investigated by applying numerous high-throughput microarray studies (Saravanan *et al.*, 2010; Kroes *et al.*, 2007; Tan *et al.*, 2014). Recently, the role of glycogenes in skeletal muscle development in primary bovine MYOGkd muscle satellite cells (Lee *et al.*, 2014) by using RNA-Seq also have been explored. Additionally, we also used precomputed expression values from mouse RNA-Seq data to identify and understand the roles of differentially expressed

glycogenes (DEGGs) in brain, muscle, and liver tissues. These studies may help in providing insights that can aid in the detection of genes and their related biological functions and in turn, translate their impact on various biological processes.

1.7 Glycoinformatics

Glycoinformatics is involvement of informatics in glycobiology to develop computational methods and tools as well as the construction of databases in order to gain insight into the functions of glycoconjugates, function of glycogens, systems analysis of glycol-related pathways and complex carbohydrates (Kiyoko *et al.*, 2009). Glycoinformatics can help to elucidate the biological function of glycans, which are intertwined with the rest of the biological system such as interacting proteins and chemical compounds.

When digital documentation systems and search engines were introduced, the Complex Carbohydrate Structure Database (CCSD) also referred as CarbBank to access the data was established in the mid-1980s, the main purpose of which was to allow the user to find easily all publications in which specific carbohydrate structures were reported. The CCSD was developed and maintained by the Complex Carbohydrate Research Center of the University of Georgia (USA) and funded by the National Institutes of Health (NIH) (Claus-Wilhelm von der Lieth *et al.*, 2009). An important stimulation for glycoinformatics development was the establishment of the Consortium for Functional Glycomics (www.functionalglycomics.org) in 2001. This was the first large-scale project that clearly emphasized the need for informatics to manage and annotate automatically the vast amount of experimental data generated by glycomics research.

Kyoto Encyclopaedia of Genes and Genomes (KEGG) including a comprehensive glycan data resource called KEGG GLYCAN encompasses all aspects of the biological system, incorporating genomic information integrated with pathways, reactions and also chemical compounds. GlySeq provides analysis of experimentally determined occupied glycosylation sites using experimental data contained in the Protein Data Bank (PDB) and SWISS-Prot (Lutteke *et al.*, 2005).

There are many tools or resources for the prediction of glycosylation sites are available. Bioinformatics programmes such as NetNGlyc and NetOGlyc are available at www.cbs.dtu.dk/services/NetNGlyc/, www.cbs.dtu.dk/services/NetOGlyc/ respectively. GlycoEP tools are available at <http://www.imtech.res.in/raghava/glycoep/> which can

predict occupancy of glycosylation sites, and can be experimentally verified. The Glyco-Net programme available at www.glycoconjugate.jp/functions/. There are several research groups developing both public and commercial glycan databases. Some of the public databases are KEGG (Kanehisa *et al.*, 2000), SWEETD (Loss *et al.*, 2002) in the GLYCOSCIENCES.de (Lütteke *et al.*, 2006), the United States Consortium for Functional Glycomics (CFG) (functionalglycomics.org), and GlycoSuiteDB in The Expert Protein Analysis System (ExPASy) Proteomics Server (Cooper *et al.*, 2003). GlycoMinds (www.glycominds.com) is known as the commercial database. The Complex Carbohydrate Structure Database (CCSD) (Doubet *et al.*, 1989) is the first database of glycan structures. The data of the CCSD are currently included in the public glycan databases as mentioned above. The Carbohydrate Active Enzyme (CAZy) database is known as a database of enzymes relating to glycans, such as glycosyltransferases and lectins (Lombard *et al.*, 2013). All of these databases with the exception of CAZy are focused on glycan structures. The SWEET-DB mainly develops the tools with which to treat the glycan structures and geometry (Loss *et al.*, 2002). The CFG is constructing carbohydrate chips to investigate the interaction between carbohydrates and proteins for therapy, and databases for functional glycomics, such as an annotated database of mass spectrometry.

The EUROCarbDB, which also provides sophisticated, freely accessible, open-source informatics tools and databases to support glycobiology and glycomics research (Von der Lieth *et al.*, 2011). PROCARB, an open access database comprising three independently working components, PROCARB module consists of 3D structures of protein-carbohydrate complexes taken from PDB, Homology Models consists of manually developed 3D models of N-Linked and O-Linked glycoproteins of unknown three-dimensional structures, CBS-Pred It consists of a webserver to predict carbohydrate-binding sites using single-sequences or evolutionary information (PSSM) are available at www.procarb.org (Malik *et al.*, 2010). In this work we have also developed a database for the prediction of functional effect of nsSNPs in identified hub glycogens of mouse, available at www.procarb.org/glyco, In future this may be developed for many glycogens. Recently, emerging analytical techniques enabled us to obtain a great deal of information It is clear that there is an urgent need to develop databases and informatics for glycobiology and glycomics. The developments in glycomics will produce an enormous amount of data and there is a need to cut across multiple datasets to understand fully the structure–function relationships of complex carbohydrates about

the relationships, not only between the glycan structures and functions, but also among glycans, phenotypes of diseases and expression of glycan-related genes. In this situation, graphical networks describing the relationships among glycan related molecules, including genes, proteins, lipids and biological events are considered to become potential tools for accelerating the integrated study of Post-translational modification (PTMs). Although the KEGG PATHWAY and Biochemical Pathways in ExPASy have graphical network figures (Hashimoto *et al.*, 2010).

1.8 RNA-Seq technology and benefits

The field of DNA sequencing has a diverse history. In the early 1990s DNA sequencing was conducted dominantly through application of Sanger sequencing using capillary based semi-automated sequencers, Sanger sequencing was used for sequencing a large fraction of genomes currently used in modern databases, including the human genome. In the past few years a number of sequencing technologies have been developed and therefore able to create more sequence output compared to conventional Sanger sequencing. These are collectively called next-generation sequencing (NGS) approaches. These technologies have been released as commercial products, e.g., the Solexa Genome Analyzer (marketed by Illumina, San Diego), the SOLiD platform (marketed by Applied Biosystems; Foster City; CA, USA), 454 Genome Sequencers (Roche Applied Science; Basel), and the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA). These technologies create reads of length 25 - 250 bps and with up to 40 million reads per run (Barba *et al.*, 2014). The sequencing of expressed RNAs with NGS approaches is abbreviated as RNA-Seq, a number of research appeared that applied RNA-Seq to transcriptome of different organisms (Wang *et al.*, 2009; Mortazavi *et al.*, 2008; Cloonan *et al.*, 2008). These papers and reviews (Wang *et al.*, 2009; Costa *et al.*, 2011; Grabherr *et al.*, 2011). have established RNA-Seq experiments provide in-depth information on the transcriptional landscape with unprecedented sensitivity and throughput and outperform the previous techniques of Expressed Sequence Tag (EST) sequencing and splicing microarrays. RNA-Seq experiments are highly reproducible (Marioni *et al.*, 2008) and have increased sensitivity, therefore more statistical power for the detection of differentially expressed genes, alternative splicing events, and unannotated transcriptional units. RNA sequencing has revolutionized the exploration of gene expression, computational

methods for RNA-Seq analyses are read mapping, align reads directly to a reference transcriptome or genome, Transcriptome reconstruction, identify expressed genes and isoforms, Expression quantification, estimation of gene and isoform abundance, as well as methods for the analysis of differential expression across samples (Marioni *et al.*, 2008).

However, despite a number of researches appeared that applied RNA-Seq to transcriptome of different organisms and their obvious functional importance, the specific mechanism is still poorly understood. It is therefore extremely important to examine computationally the RNA-Seq data of mouse for glycogene expression.

What is RNA-Seq

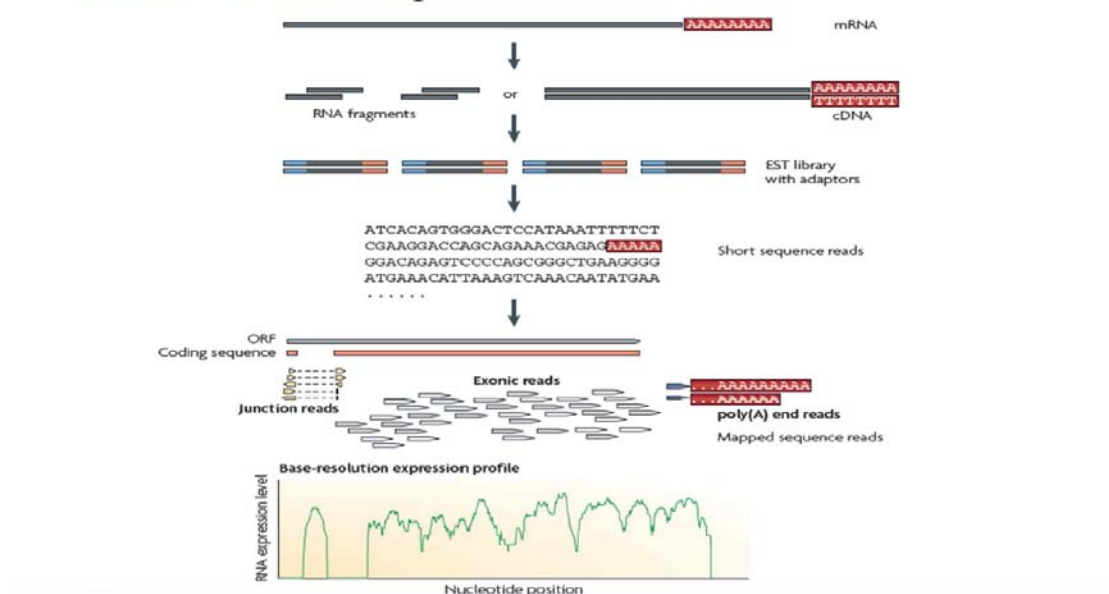


Figure-1.3 A typical RNA-seq experiment (Source: Wang Z et al., 2009)

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of transcriptome. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. RNA-Seq uses recently developed deep sequencing technologies. In general, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). The reads are typically 30–

400 bp, depending on the DNA sequencing technology used. In principle, any high-throughput sequencing technology can be used for RNA-Seq, and the Illumina IG, Applied Biosystems SOLiD and Roche 454 Life Science systems have already been applied for this purpose. The Helicos Biosciences tSMS system has not yet been used for published RNA-Seq studies, but is also appropriate and has the added advantage of avoiding amplification of target cDNA. Following sequencing, the resulting reads are either aligned to a reference genome or reference transcripts, or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene. Although RNA-Seq is still a technology under active development, it offers several key advantages over existing technologies.

1.8.1 Computational Challenges for RNA-Seq

Like other high-throughput sequencing technologies, RNA-Seq faces several informatics challenges, including the development of efficient methods to store, retrieve and process large amounts of data, which must be overcome to reduce errors in image analysis and base-calling and remove low-quality reads. Once high-quality reads have been obtained, the first task of data analysis is to map the short reads from RNA-Seq to the reference genome, or to assemble them into contigs before aligning them to the genomic sequence to reveal transcription structure. There are several programs for mapping reads to the genome, including ELAND, SOAP, MAQ and RMAP. However, short transcriptomic reads also contain reads that span exon junctions or that contain poly(A) ends, these cannot be analysed in the same way. For complex transcriptomes it is more difficult to map reads that span splice junctions, owing to the presence of extensive alternative splicing and *transsplicing*. One partial solution is to compile a junction library that contains all the known and predicted junction sequences and map reads to this library (Wilhelm *et al.*, 2008; Mortazavi *et al.*, 2008). A challenge for the future is to develop computationally simple methods to identify novel splicing events that take place between two distant sequences or between exons from two different genes. For large transcriptome, alignment is also complicated by the fact that a significant portion of sequence reads match multiple locations in the genome. One solution is to assign these multi-matched reads by proportionally assigning them based on the number of reads mapped to their neighbouring unique sequences. This method has been successful for low-copy repetitive

sequences (Mortazavi *et al.*, 2008). Short reads that have high copy numbers (>100) and long stretches of repetitive regions present a greater challenge. Obtaining longer sequence reads, for example using 454 technology, should help alleviate the multi-matching problem. Alternatively, a paired-end sequencing strategy, in which short sequences are determined from both ends of a DNA fragment, extends the mapped fragment length to 200–500 bp and is expected to be useful in the future. Sequencing errors and polymorphisms can present mapping problems for all genomes, not just for repetitive DNA. Generally, single base differences are not problematic, because most mapping algorithms accommodate one or two base differences. However, resolving larger differences will require better reference genome annotation for polymorphisms and deeper sequencing coverage.

One particularly powerful advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets. RNA-Seq has been used to accurately monitor gene expression during yeast vegetative growth, yeast meiosis (Wilhelm *et al.*, 2008) and mouse embryonic stem-cell differentiation, to track gene expression changes during development, and to provide a ‘digital measurement’ of gene expression difference between different tissues. Because of these advantages, RNA-Seq will undoubtedly be valuable for understanding transcriptomic dynamics during development and normal physiological changes, and in the analysis of biomedical samples, where it will allow robust comparison between diseased and normal tissues, as well as the sub classification of disease states.

1.9 Single Nucleotide Polymorphism

A single nucleotide polymorphism, or SNP, is a variation at a single position in a DNA sequence among individuals, are not just associated with genes but can also occur in non coding regions of DNA. Although a particular SNP may not cause a disorder, some SNPs are associated with certain diseases. A large amount of effort is being made in current epidemiology, medicine, and pharmacogenomics which focus on the discovery of genetic variations especially SNPs that are implicated in common and complex diseases (Lee *et al.*, 2008). Due to the large number of SNPs present in the genomes, prioritizing SNPs on the basis of their possible deleterious functional effects is essential (Bhatti *et al.*, 2006). Since there is lack of experimental data for the majority of SNPs, therefore, it becomes challenging to authenticate their deleterious effects. Many computational resources have

been developed and widely used that predict the assumed deleterious effects of these SNPs (Rebbeck *et al.*, 2004). In general, these computational tools scrutinize whether the SNP inhabits in functional genomics regions such as exons, splice sites, or transcription regulatory sites, and accordingly predict the possible related functional effects that the SNP may have using a range of machine learning methods (Lee *et al.*, 2008). However, to comprehensively evaluate the functional implication of SNPs, substantial amount of time and effort is required to apply multiple tools independently, and interpret the prediction results carefully (Lee *et al.*, 2008). It will be worth to obtain additional information regarding the selected SNPs before finalizing the decision. For example, SNPs in protein coding or regulatory regions are more prone to have deleterious effects, and, as such, the potential of triggering a disease is enhanced. Therefore, by using these computational methods and by acquiring additional information about the functionally significant SNPs that are expected to be linked with disease, a significant amount of genotyping and analysis overhead can be reduced (Lee *et al.*, 2008).

Single amino acid mutations within the proteins can considerably alter the stability of a protein structure and in turn may affect its function. Therefore, to understand the functional effects of variation, we also investigated the role of nsSNPs that may lead to an amino acid change in the glycoprotein.

1.10 Conclusions

Glycogens, that are responsible for the glycosylation of proteins, lipids, proteoglycans, and are associated with the synthesis of glycans A large variety of biological processes during the growth and development of organisms in addition to wide array of diseases such as cancer, primary open-angle glaucoma, and renal dysfunction, are regulated by glycogenes. Expression pattern of these glycogenes is one of the essential factors that indicate about several biological functions, developmental changes, and diseases in organisms, although expression of glycogenes is comparatively weak in comparison to other molecules yet, some glycogenes are upregulated during particular circumstances. Glycoinformatics is involvement of informatics in glycobiology to develop computational methods and tools as well as the construction of databases and can be helpful to gain insight into the functions of glycoconjugates, function of glycogens, many computational resources and methods are available for glycogens analysis. Single amino acid mutations within the proteins can considerably alter the stability of a protein structure and in turn

may affect its function. Therefore, the role of nsSNPs that may lead to an amino acid change in the glycoproteins. RNA-Seq experiments provide in-depth information on the transcriptional landscape with unprecedented sensitivity and outperform the previous techniques, so it is extremely important to examine glycogenes computationally using RNA-Seq data.

1.11 Objectives

In order to investigate the functional roles of glycogens, which are involved in various biological processes, computational analysis using RNA-Seq data of mouse with **following objectives** were defined for the present thesis.

- i. SNP identification & association studies
- ii. Quantifying the Gene expression
- iii. Novel Gene/isoform/exon prediction
- iv. Gene Ontology analysis of differentially expressed genes

References

- Audette, GF.; Delbaere, LT.; Xiang, J. Mapping protein: carbohydrate interactions. *Curr Protein Pept Sci.* 4, 11-20 (2003).
- Barba, M.; Czosnek, h.; Hadidi, A. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses.* 6. 106–136 (2014).
- Barondes, S.H.; Cooper, D.N.; Gitt, M.A.; Leffler, H. Galectins: Structure and function of a large family of animal lectins. *Journal of Biological Chemistry.* 269, 20807-20810 (1994).
- Bhatti, P.; Church, DM.; Rutter, JL.; et al, Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. *Am. J. Epidemiol.* 164. 794–804 (2006).
- Cloonan, N.; Forrest, ARR.; Kolle, G.; et al., Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods.* 5. 613 - 619 (2008).
- Cooper, CA.; Joshi, HJ.; Harrison, MJ.; et al., GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.* 31. 511–513 (2003).
- Costa, V.; Angelini, C.; DApice, L.; et al., Massive-Scale RNA-Seq Analysis of Non Ribosomal Transcriptome in Human Trisomy. *PLoS ONE.* 6. e18493 (2011).
- Diskin, S.; Kumar, J.; Cao, Z.; et al., Detection of differentially expressed glycogenes in trabecular meshwork of eyes with primary open-angle glaucoma. *Investigative Ophthalmology and Visual Science.* 47, 1491–1499 (2006).
- Doubet, S.; Bock, K.; Smith, D.; et al., The Complex Carbohydrate Structure Database. *Trends Biochem Sci.* 14. 475–477 (1989).

Dwek, MV.; Brooks, SA.; Streets, AJ.; et al. Oligosaccharide release from frozen and paraffin-wax-embedded archival tissues. *Anal Biochem.* 242, 8-14 (1996).

Fukuda, MN.; Gaetani, GF.; Izzo, P.; et al.; Incompletely processed N-glycans of serum glycoproteins in congenital dyserythropoietic anaemia type II (HEMPAS). *J Haematol.* 82, 745-52 (1992).

Grabherr, MG.; Haas, BJ.; Yassour, M.; et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.* 29. 644–652 (2011).

Guo, R.; Cheng, L.; Zhao, Y.; et al., Glycogenes mediate the invasive properties and chemosensitivity of human hepatocarcinoma cells. *The International Journal of Biochemistry and Cell Biology.* 45, 347–358 (2013).

Hart, GW.; Housley, MP.; Slawson, C. Cycling of O-linked beta-N-cetylglucosamine on nucleocytoplasmic proteins. *Nature.* 26, 1017-22 (2007).

Hashimoto, R.; Hirose, K.; Sato, T.; et al.,. Functional network of glycan-related molecules: glyco-net in glycoconjugate data bank. *BMC Syst Biol.* 29, 4:91 (2010).

Haynes, PA. Phosphoglycosylation: a new structural class of glycosylation?. *Glycobiology.* 8.1-5, (1998).

Helenius, A.; Aebi, M. Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev Biochem.* 73,1019-49 (2004).

Herget, S.; Ranzinger, R.; Thomson, R.; et al., Bioinformatics for Glycobiology and Glycomics: an Introduction. John Wiley & Sons, Ltd, (2009).

Hoflack, B.; Fujimoto, K.; Kornfeld, S. The interaction of phosphorylated oligosaccharides and lysosomal enzymes with bovine liver cation-dependent mannose 6-phosphate receptor. *Journal of Biological Chemistry.* 5, 123-9 (1987).

<http://www.ncbi.nlm.nih.gov/books/NBK1897/>

<http://www.scripps.edu/news/press/2004/120704.html>

<http://www.stelic.com/cn27/pg122.html>.

<http://www.uniprot.org/help/carbohydr>

Inoue, T.; Sugiyama, H.; Hiki, Y.; et al., Differential expression of glycogenes in tonsillar B lymphocytes in association with proteinuria and renal dysfunction in IgA nephropathy. *Clinical Immunology*. 136, 447–455 (2010).

Inoue, T.; Sugiyama, H.; Kitagawa, M.; et al., Abnormalities of glycogenes in tonsillar lymphocytes in IgA nephropathy. *Advances in Otorhinolaryngology*. 72, 71-74 (2011).

Ishii, A.; Hitoshi, S.; Fujimoto, I.; et al., Developmental changes in the expression of glycogenes and the content of N-glycans in the mouse cerebral cortex. *Glycobiology*. 17, 261-276 (2007).

Janot, M.; Audfray, A.; Lorient, C.; et al., Glycogenome expression dynamics during mouse C2C12 myoblast differentiation suggests a sequential reorganization of membrane glycoconjugates. *BMC Genomics*. 10. 483, (2009).

Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28. 27–30 (2000).

Karlsson, KA. Meaning and therapeutic potential of microbial recognition of host glycoconjugates. *Mol Microbiol*. 29, 1-11 (1998).

Kiyoko, F.; Kinoshita, A.; Kanehisa, M . Bioinformatics analysis of glycan structures from a genomic perspective. John Wiley & Sons, Ltd (2009).

Kobayashi, T.; Nishizaki, R.; Ikezawa, H. The presence of GPI-linked protein(s) in an archaeobacterium, *Sulfolobus acidocaldarius*, closely related to eukaryotes. *Biochim Biophys Acta*. 1334. 1-4, (1997).

Kroes, R. A.; Dawson, G.; Moskal, J. R. Focused microarray analysis of glyco-gene expression in human glioblastomas. *Journal of neurochemistry*. 103, 14–24 (2007).

Kroes, RA.; Dawson, G.; Moskal, JR. Focused microarray analysis of glyco-gene expression in human glioblastomas. *Journal of neurochemistry*. 103, 14–24 (2007).

Lee, EJ.; Malik, A.; Pokharel, S.; et al., Identification of genes differentially expressed in myogenin knock-down bovine muscle satellite cells during differentiation through RNA sequencing analysis. *PLoS ONE*. 9. Article ID e92447 (2014).

Lee, PH.; Shatkay, H. Ranking single nucleotide polymorphisms by potential deleterious effects. *AMIA Ann. Symp. Proc.* 6. 667–671 (2008).

Lewin, B. *GENES VIII*. International Edition, Pearson Prentice Hall. UK. (2004).

Lodish, H.; Berk, A.; Zipursky, S.; et al., *Molecular Cell Biology*. 4th ed. New York. W. H. Freeman (2000).

Lombard, V.; Ramulu, HG.; Drula, E.; et al., The carbohydrate-active enzymes database (CAZy). *Nucleic Acids Research*. 42. D490–D495 (2014).

Loss, A.; Bunsmann, P.; Böhne, A. et al., SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.* 30. 405-8 (2002).

Lutke, T., Böhne-Lang, A., Loss, A., et al., GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* . 16. 71R-81R (2006).

Lütke, T.; Frank, M.; Von der Lieth, CW. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.* 33 (Database issue). D242-6 (2005).

Ma, H.; Miao, X.; Ma, Q.; Zheng, W.; Zhou, H.; Jia, L. Functional roles of glycogene and N-glycan in multidrug resistance of human breast cancer cells. *IUBMB Life*. 65, 409–422 (2013).

Malik, A.; Firoz, A.; Jha, V.; et al., PROCARB: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools. *Advances in Bioinformatics*. Article ID 436036 (2010).

Mantelli, F.; Schaffer, L.; Dana, R.; Head, S. R.; Argueso, P. Glycogene expression in conjunctiva of patients with dry eye down regulation of Notch signaling. *PLoS ONE*. 50, 2666-2672 (2010).

Marioni, JC.; Mason, CE.; Mane, SM.; et al., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 18. 1509-17 (2008).

Maupin, K. A.; Sinha, A.; Eugster E.; et al., Glycogene expression alterations associated with pancreatic cancer Epithelialmesenchymal transition in complementary model systems. *PLoS ONE*. 5, Article ID e13002 (2010).

Mazola,Y.; Chinae, G.; Musacchio, A.; Integrating Bioinformatics Tools to Handle Glycosylation. *PLoS Computational Biology*. 7, e1002285 (2011).

Mortazavi , A.; Williams, BA.; McCue, K.; et al., Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, 5. 621-8 (2008).

Nakayama, Y.; Nakamura, N.; Tsuji, D.; Itoh, K. ; Kurosaka, A. Genetic diseases associated with protein glycosylation disorders in mammals. *InTechOpen*. 10 , (2013).

Narimatsu, H. Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J*. 21. 17-24, (2004).

Rademacher, T. W.; Parekh, R. B.; Dwek, R. A. Glycobiology. *Annual Review of Biochemistry*. 57, 785-838 (1988).

Rebbeck, TR.; Spitz, M.; Wu, X. Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet*. 5. 589–597 (2004).

Roth, GA.; Gonzalez, MD.; Monferran, CG.; et al., Myelin basic protein domains involved in the interaction with actin. *Neurochem Int.* 23, 459-65 (1993).

Salzman, J.; Jiang, H.; Wong, W. H. Statistical modeling of RNA-Seq data. *Statistical Science.* 26, 62-83 (2011).

Saravanan, C.; Cao, Z.; Head, S. R.; Panjwani, N. Analysis of differential expression of glycosyltransferases in healing corneas by glycogene microarrays. *Glycobiology.* 20, 13–23 (2010).

Saravanan, C.; Cao, Z.; Head, SR.; et al., Detection of differentially expressed wound-healing- related glycogens in galectin-3-deficient mice. *Investigative Ophthalmology and Visual Science.* 50. 5690–5696 (2009).

Shionyu-Mitsuyama, C.; Shirai, T.; Ishida, H.; Yamane, T. An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. *Protein Engineer.* 16, 467–478 (2003).

Tan, Z.; Lu, W.; Li, X.; et al., Altered N-glycan expression profile in epithelial-to-mesenchymal transition of NMuMG cells revealed by an integrated strategy using mass spectrometry and glycogene and lectin microarray analysis. *Journal of Proteome Research.* 13, 2783–2795 (2014).

Tan, Z.; Lu, W.; Li, X.; et al., Altered N-glycan expression profile in epithelial-to-mesenchymal transition of NMuMG cells revealed by an integrated strategy using mass spectrometry and glycogene and lectin microarray analysis. *Journal of Proteome Research.* 13. 2783–2795 (2014).

Taniguchi, N.; Endo, T.; Hart, GW.; et al., *Glycoscience: Biology and Medicine.* Springer (2015).

Taniguchi, N.; Suzuki, A.; Ito, Y.; et al., *Experimental Glycoscience.* Glycobiology. Springer, New York, NY, USA, (2009).

Varki, A.; Cummings, R.; Esko, J.; et al., Essentials of Glycobiology. Cold Spring Harbor Laboratory Press. ISBN-10:0-87969-559-5 (1999).

Varki, A.; Cummings, RD.; Esko, JD.; et al., Essentials of Glycobiology. 2nd edition. Cold Spring Harbor Laboratory Press. (2009).

Von der Lieth, CW. Glycobiology, Glycomics and (Bio)Informatics. John Wiley & Sons, Ltd (2009).

Von der Lieth, CW.; Ardá Freire, A.; Blank, D.; et al., EUROCarbDB: An open-access platform for glycoinformatics. Glycobiology. 21. 493–502 (2011).

Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 10. 57-63 (2009).

Wilhelm, BT.; Marguerat, S.; Watt, S.; et al., Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 26. 1239-43 (2008).

www.deathtodibabetes.com/Glyconutrients_Glycobiology.html#.VmZ1i3YrLlo

www.functionalglycomics.org

Zhang, Z.; Sun, J.; Hao, L.; et al., Modification of glycosylation mediates the invasive properties of murine hepatocarcinoma cell lines to lymph nodes. PLoS ONE. 20. Article ID e65218 (2013).

Materials and Methods

Contents		Page number
2.1	Datasets	25
2.2	Data filtering	25
2.3	Fold change analysis	25
2.4	Annotation of differentially expressed glycogenes	26
2.5	Functional analysis	26
2.6	Pathway analysis	27
2.7	Network analysis and identification of hubs	27
2.8	Community and enrichment analysis	27
2.9	Computational analysis of nsSNPs of identified hub genes	28
2.10	Surface accessibility and secondary structure analysis of predicted nsSNPs	29
2.11	Database & Tools development	29
	References	30

Abstract: This chapter includes Datasets, databases, Computational analysis tools, detailed description of methods and analytical techniques which are employed in present thesis.

Materials and Methods

2.1 Data sets

The expression value of RNAseq data sets with three different C57BL6 mouse tissues bearing accession numbers SRA001030 at NCBI Sequence Read Archive were used in this study, where RNAseq experiment were performed using Illumina Genome Analyzer and expression value computed using ERANGE package (Mortazavi *et al.*, 2008). The precomputed expression values with no spike for the brain, liver and muscles tissues were downloaded from WoldLab (<http://woldlab.caltech.edu/~alim/RNAseq/>) for our computational analysis.

2.2 Data filtering

To remove low expressed genes we retained only those genes where RPKM (Reads Per Kilobase of transcript per Million mapped reads) value for each gene were ≥ 1 in all the three tissues.

2.3 Fold change analysis

Fold Change is measure describing how much a quantity changes going from an initial to a final value. If initial value of x and a final value of $2x$ corresponds to a fold change of 2, or in common terms, a two-fold increase. It is often used in analysis of gene expression data of RNA-Seq experiments, for measuring change in the expression level of a gene and compare the expression of genes between two sets of arrays, e.g. case and control sets. Fold change analysis of genes commonly expressed between two different samples was done to compare the gene expression in glycogenes (Guo *et al.*, 2013; Lee *et al.*, 2014). Fold change was calculated as the ratio of brain versus liver, brain versus muscle, and liver versus muscle groups. The list of differentially expressed genes (DEGs) in each group was created for brain versus liver (BvL), brain versus muscle (BvM), and liver versus muscle (LvM) groups.

2.4 Annotation of all the differentially expressed glycogenes

To investigate the role of glycogenes in different tissues, the differentially expressed genes were manually verified in the UniProt database (<http://www.uniprot.org/>) to check whether they represent a glycogene or not. A gene was considered as a glycogene if it was annotated as a “glycoprotein” in the “Keywords” in PTM (Post translational Modification) section of the UniProt database, which is a comprehensive, high-quality and freely accessible database of protein sequence and information about the biological function of proteins, available at www.uniprot.org. The list of DEGs in each group was further filtered by retaining only those genes that were annotated as a “glycoprotein” in the UniProt database. Therefore, differentially expressed glycogenes (DEGGs) for brain versus liver (BvL), brain versus muscle (BvM), and liver versus muscle (LvM) groups were selected for the final analysis. We also created a non-redundant list of glycogenes by combining all identified glycogenes of brain, muscle and liver tissues.

2.5 Functional analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) is web-accessible programs, available at <http://david.abcc.ncifcrf.gov/home.jsp>, that provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind given list of genes. For any given gene list DAVID tools are able to Identify enriched biological themes, particularly GO term, discover enriched functional-related gene groups, cluster redundant annotation terms, visualize genes on BioCarta & KEGG pathway maps, search for other functionally related genes not in the list, list interacting proteins, link gene-disease associations, highlight protein functional domains and motifs, and other more utilities are also available (Huang *et al.*, 2009). DAVID Functional Annotation Clustering Tool can condense heterogeneous and redundant annotation terms into group for biological interpretation to explore the biological meaning of interesting gene lists from different angles. DAVID functional annotation cluster analysis was performed on the list of differentially expressed glycogenes (DEGGs) with a fold change of ≥ 2 . Only those clusters that show statistically significant, reported a *P* value of ≤ 0.05 gene ontology (GO) terms were selected, and count number ≥ 5 genes were selected for analysis. The gene ontology (GO) term

biological process (BP) in DAVID was used to categorize enriched biological processes in the different list of DEGs.

2.6 Pathway analysis

Pathway mapping was performed using the KEGG Automatic Annotation Server (KAAS) (<http://www.genome.jp/kegg/kaas/>) (Moriya *et al.*, 2007). The amino acid sequences of these DEGs were uploaded to the KAAS web server as an input using single-directional best hit (SBH) method to assign orthologs. KAAS offers functional annotation of genes in a genome via a Basic Local Alignment Search Tool (BLAST) similarity searches against a manually curated set of ortholog groups in the KEGG GENES database. KAAS assigned a KEGG orthology (KO) number to genes in the data sets, which were mapped to one of KEGG's reference pathways.

2.7 Network analysis and identification of hubs

The functional interactions between DEGs were analyzed by GeneMANIA webserver (Montejo *et al.*, 2010). The GO term biological process was used to create the interaction network between the DEGs and 50 additional genes by using mouse as a source species. The relationship between the genes in the network includes co-expression, physical and genetic interactions, pathways, co-localization, protein domain similarity, and predicted interactions. The network was filtered by removing all the interactions where weights < 0.1 .

Generally, the biological networks exhibit the scale-free property (Albert *et al.*, 2005) where only a few nodes in the network have many connections that represent hubs in the network. Hub genes were identified by calculating the node degree distribution (Wu *et al.*, 2005) by using the Network Analyzer plugin of Cytoscape (Smoot *et al.*, 2011). Top five genes with the highest degree distribution was considered as a hub in the current network.

2.8 Community and enrichment analysis

The functional modules within the network were further detected by using the greedy community-structure detection algorithm, implemented in the Cytoscape plugin GLay (<http://brainarray.mbni.med.umich.edu/sugang/glay/>) (Su *et al.*, 2010). GLay offers Cytoscape users a diverse group of community structure algorithms and graph layout

functions for network clustering and structured visualization (Su *et al.*, 2010). To identify the over represented biological functions within each cluster, the detected clusters were subjected to functional enrichment analyses by DAVID functional analysis tool. For enrichment analysis, only those communities were analyzed which have at least 10 nodes.

2.9 Computational analysis of coding non-synonymous SNPs (nsSNPs) of identified hub genes

Data mining the SNP information for hub genes identified for mouse was retrieved from National Centre for Biotechnology Information (NCBI) database dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP>). Three different programs Protein Variation Effect Analyzer (PROVEAN), Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping v2 (POLYPHEN-2) were used to predict the damaged or deleterious coding nsSNPs. For the prediction of deleterious or damaging coding nsSNPs using PROVEAN Protein and SIFT Sequence tools PROVEAN (Protein Variation Effect Analyzer) Protein [http://provean.jcvi.org/seq_submit.php] and SIFT (Sorting Intolerant From Tolerant) sequence [http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html] tools were used to provide PROVEAN and SIFT predictions for a list of protein variants. Both PROVEAN and SIFT are software tools that predict whether an amino acid substitution has an impact on the biological function of a protein if the score lies below a certain threshold value. In PROVEAN, the clustering of BLAST hits is carried out and the best 30 clusters of closely linked sequences form the supporting sequence set that is further employed to make the predictions. For every supporting sequence a delta alignment score is computed and then averaged within and across clusters generating final PROVEAN score (Choi *et al.*, 2012). A default score threshold of -2.5 or above is considered to be deleterious whereas anything less than this cut-off score has a neutral effect. On the other hand, SIFT is a multi-step algorithm, and uses sequence homology based method to classify amino acid substitutions. (Kumar *et al.*, 2009; Ng *et al.*, 2003). These tools accept a list of protein sequence variants as input for the predictions.

For the prediction of functional modification of coding nsSNPs by Polymorphism Phenotyping v2 (PolyPhen-2) The **standalone** PolyPhen-2 tool was installed for mouse and used to study the possible consequence of nsSNPs on protein structure and function.

PolyPhen-2 predicts the possible effect of amino acid replacement on the stability and function of mouse proteins by using structural as well as comparative evolutionary considerations (Adzhubei *et al.*, 2010). For every specific amino acid residue substitution in a protein, PolyPhen-2 mines a range of sequence and structural features of the replacement position and supplies them to a probabilistic classifier (Adzhubei *et al.*, 2010). For this work, we have selected the default option for generating the predictions.

2.10 Surface accessibility and secondary structure analysis of predicted nsSNPs

NetSurfP (ver.1.1) web server was used to predict protein surface accessibility or accessible surface area (ASA) and secondary structure predictions [<http://www.cbs.dtu.dk/services/NetSurfP>]. In addition to the prediction of ASA and secondary structure from sequence, NetSurfP also predicts the reliability for each prediction simultaneously in the form of a Z-score. The NetSurfP method consists of two sets of neural networks, the primary networks that are trained on sequence profiles and predicted secondary structure, and the secondary networks which exploits the outputs of primary networks as input collectively with sequence profiles to predict the relative surface exposure of the individual amino acid residues (Petersen *et al.*, 2009). The input sequences in fasta format of normal and predicted SNPs were submitted to the server for prediction.

2.11 Database & Tools development

The SNP data that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2), which may in turn affect protein function and also the prediction of protein surface accessibility or accessible surface area (ASA) and secondary structure by NetSurfP of hub genes are available online in the form of a simple database available at www.procarb.org/glyco. The web interface uses Hypertext Pre-processor (PHP) and JavaScript to allow users to search the database, stored in the of MySQL table, by applying a variety of text-based options like simple search and advanced search tools, where Data entries are displayed using dynamically generated pages which describe the relevant information.

References

Adzhubei, IA.; Schmidt, S.; Peshkin, L.; et al., A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**. 248–249, (2010).

Albert, R. Scale-free networks in cell biology. *Journal of Cell Science*. **118**. 4947-4957, (2005).

Choi, Y.; Sims, GE.; Murphy, S.; et al., Predicting the functional effect of amino acid substitutions and Indels. *PLoS ONE* .**7**. e46688, (2012).

Guo, Y.; Sheng, Q.; Li, J.; et al., Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS ONE*. **8**. Article ID e71462 (2013).

<http://www.uniprot.org/>.

Huang, DW.; Sherman, BT.; Lempicki, RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. **37**. 1-13, (2009).

Huang, DW.; Sherman, BT.; Lempicki, RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. **4**.44-57, (2009).

Kumar,P.; Henikoff, S.; Ng, PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc*. **4**. 1073–1081, (2009).

Lee, EJ.; Malik, A.; Pokharel, S.; et al., Identification of genes differentially expressed in myogenin knock-down bovine muscle satellite cells during differentiation through RNA sequencing analysis. *PLoS ONE*. **9**.Article ID e92447, (2014).

Montejo, J.; Zuberi, K.; Rodriguez, H.; et al., GeneMANIA cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*. **26**. Article ID btq562. 2927–2928, (2010).

Moriya, Y.; Itoh, M.; Okuda, S.; et al., KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*. 35.W182–W185, (2007).

Mortazavi , A.; Williams, BA.; McCue, K.; et al., Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, 5. 621-8 (2008).

Ng, PC.; Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 31. 3812–3814, (2003).

Petersen, B.; Petersen, TN.; Andersen, P.; et al., A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* 9. 51, (2009).

Smoot, ME.; Ono, K.; Ruscheinski, J. et al., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 27. 431–432, (2011).

Su, G.; Kuchinsky, A.; Morris, JH.; et al., GLay: community structure analysis of biological networks. *Bioinformatics*. 26. 3135–3137, (2010).

Wu, XR.; Zhu, Y.; Li, Y. Analyzing protein interaction networks via random graph model. *International Journal of Information Technology*. 11, 125–132, (2005).

Comparative Analysis of Glycogene Expression

	Contents	Page number
3.1	Introduction	32
3.2	Identification of differentially expressed glycogenes	32
3.3	Functional annotation and pathway analysis	33
3.4	Interaction network and identification of hub Genes	36
3.5	Community analysis and functional annotation of the detected modules	42
3.6	Discussion	52
	References	56

3.1 Introduction

The overall methodology adopted in the study discussed in this chapter is depicted as a flowchart in **Figure 3.1**.

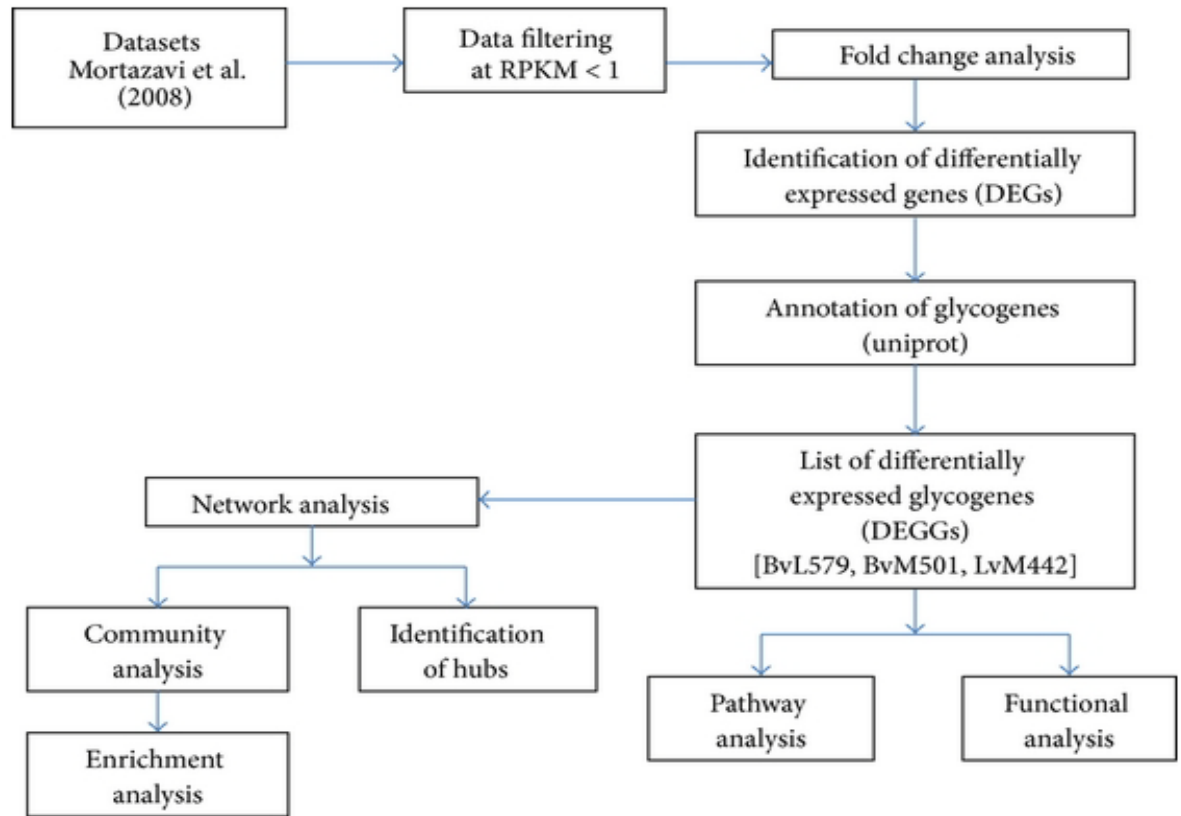


Figure 3.1 flowchart diagram

3.2 Identification of differentially expressed glycogenes

We downloaded the precomputed RPKM values for mouse transcriptome that was mapped and quantified previously by RNA-Seq analysis (Mortazavi *et al.*, 2009). This RNA-Seq data consists of RNA from adult mouse brain, liver, and skeletal muscle tissues and represents 33598 genes in each tissue sample. The data was further filtered by excluding the genes with RPKM values <1 from the analysis. As a result, 15,237, 11,920, and 12,202 genes were identified from brain, liver, and muscle samples, respectively (Table 3.1(a)). Then three groups (brain versus liver, brain versus muscle, and liver versus muscle) were created and the number of genes that were common in each group were identified and summarized in Table 3.1(b). From the table, it can be

observed that the number of common genes between brain versus liver is 10653, brain versus muscle is 10820, and liver versus muscle is 9833, respectively. These shared genes in each group were then used to calculate the fold change, which was defined as the ratio of RPKM values of brain versus liver, brain versus muscle, and liver versus muscle groups. In this study, the total fold change of ≥ 2 was considered to classify the differentially expressed genes (DEGs). Based on this definition, there are 5228 DEGs between brain versus liver, 5283 between brain versus muscle, and 3840 between liver versus muscle. The list of DEGs in each group was further filtered by retaining only those genes that were annotated as a “glycoprotein” in the UniProt database. Therefore, 579, 501, and 442 differentially expressed glycogenes (DEGGs) for brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups were selected for the final analysis (Table 3.1(b)).

Table 3.1: Gene expression summary. a) Total number of genes in each sample before and after filtering. Each tissue sample consists of 33598 genes that were filtered by removing all those genes that showed RPKM value < 1 . **b)** Total number of differentially expressed genes between different tissues.

(a) Data filtering			
Tissue	Total no. of genes		Total no. of genes after filtering
Brain	33598		15,237
Liver	33598		11,920
Muscle	33598		12,202
(b) Genes commonly expressed between different tissues			
Tissue	No. of expressed genes	No. of Differentially expressed genes (DEGs) (≥ 2 Fold Change)	No. of differentially expressed glycogenes (DEGGs) (≥2 Fold Change)
Brain Vs Liver	10,653	5228	579
Brain Vs Muscle	10,820	5283	501
Liver Vs Muscle	9,833	3840	442

3.3 Functional annotation and pathway analysis

To classify biological processes that are enriched in the brain vs liver, brain vs muscle, and liver vs muscle tissue groups, we classified all known DEGGs (fold change ≥ 2) using the Functional Annotation Cluster (FAC) tool available in the Database for Annotation, Visualization and Integrated Discovery (DAVID)

[<http://david.abcc.ncifcrf.gov/home.jsp>]. The GO term “Biological Process” was used for annotations, and top 10 GO terms having statistically significant *p*-values from the resulting functional analysis for each group is listed in **Table 3.2**. From this table, it can be seen that the GO terms that are enriched in brain vs liver glycogenes represent functions necessary for response to wounding, inflammatory response, and blood coagulation (**Table 3.2a**). The enriched GO terms in brain vs muscle include functions related to various metabolic (glycoprotein, glycosaminoglycan, sphingolipid, etc) processes (**Table 3.2b**). Similarly, functions related protein maturation, inflammatory response, complement activation are enriched in liver vs muscle samples (**Table 3.2c**).

Table 3.2. Top 10 significantly enriched gene ontology (GO) terms detected by FAC in differentially expressed glycogenes between **a)** Brain Vs Liver, **b)** Brain Vs Muscle, and **c)** Liver Vs Muscle. Only those terms which reported a *p*-value of ≤ 0.05 and count number ≥ 5 genes were selected for the analysis.

a) Brain Vs Liver		
GO Term	Total no. of genes	P-value
GO:0009611~response to wounding	53	2.98E-23
GO:0002526~acute inflammatory response	23	4.83E-16
GO:0006954~inflammatory response	35	1.43E-15
GO:0051604~protein maturation	23	2.51E-14
GO:0007596~blood coagulation	20	5.40E-14
GO:0050817~coagulation	20	5.40E-14
GO:0007599~hemostasis	20	7.17E-14
GO:0006952~defense response	45	7.57E-13
GO:0050878~regulation of body fluid levels	20	5.84E-12
GO:0016485~protein processing	20	7.21E-12
b) Brain Vs Muscle		
GO:0009100~glycoprotein metabolic process	30	1.76E-18
GO:0009101~glycoprotein biosynthetic process	23	8.1E-14
GO:0070085~glycosylation	17	6.09E-10
GO:0006486~protein amino acid glycosylation	17	6.09E-10
GO:0043413~biopolymer glycosylation	17	6.09E-10
GO:0030203~glycosaminoglycan metabolic process	11	3.1E-08
GO:0006665~sphingolipid metabolic process	12	2.79E-07
GO:0006643~membrane lipid metabolic process	12	3.9E-07
GO:0006022~aminoglycan metabolic process	11	4.45E-07
GO:0051604~protein maturation	14	6.1E-07
c) Liver Vs Muscle		
GO:0051604~protein maturation	20	3.17E-13
GO:0016485~protein processing	19	1.12E-12

GO:0009100~glycoprotein metabolic process	23	1.63E-12
GO:0006954~inflammatory response	27	6.69E-12
GO:0009611~response to wounding	33	1.07E-11
GO:0002526~acute inflammatory response	17	2.53E-11
GO:0009101~glycoprotein biosynthetic process	19	1.14E-10
GO:0051605~protein maturation by peptide bond cleavage	15	1.33E-10
GO:0006956~complement activation	11	4.25E-09
GO:0002541~activation of plasma proteins involved in acute inflammatory response	11	4.25E-09

in addition to DAVID functional analysis, we also identified the biological pathways of brain vs liver, brain vs muscle, and liver vs muscle DEGs annotated in the present study. FASTA formatted amino acid sequences of DEGs in these three sets were fed into the KAAS for prediction of distinct pathways. A total of 210 pathways were predicted for brain vs liver, whereas 193 and 198 pathways were predicted for brain vs muscle and liver vs muscle, respectively. The top 10 KEGG pathways for each of the three categories are shown in **Table 3.3** and a complete list of all pathways is provided in **Supplementary Table 3S1**. From these tables, it can be observed that the majority of the DEGs were found to be associated with important biological processes, many being classified in signaling (e.g. PI3K-Akt and Rap1 signaling) pathways, Cytokine-cytokine receptor interaction, Extra Cellular Matrix (ECM) receptor interaction or being involved in adhesion related functions.

Table 3.3 Top 10 KEGG pathways for DEGs in each category. a) Brain Vs Liver, b) Brain Vs Muscle, and c) Liver Vs Muscle.

(a) Brain Vs Liver	
Pathway Name	No. of mapped Genes
04151 PI3K-Akt signaling pathway	34
04610 Complement and coagulation cascades	34
04142 Lysosome	31
04060 Cytokine-cytokine receptor interaction	23
05200 Pathways in cancer	21
04015 Rap1 signaling pathway	18
04510 Focal adhesion	18
05205 Proteoglycans in cancer	18
04514 Cell adhesion molecules CAMs	16
04976 Bile secretion	16

(b) Brain Vs Muscle

04142 Lysosome	29
04151 PI3K-Akt signaling pathway	24
04060 Cytokine-cytokine receptor interaction	19
04514 Cell adhesion molecules CAMs	19
05200 Pathways in cancer	17
04141 Protein processing in endoplasmic reticulum	15
04510 Focal adhesion	15
05166 HTLV-I infection	13
04512 ECM-receptor interaction	12
04610 Complement and coagulation cascades	12

(c) Liver Vs Muscle

04142 Lysosome	33
04151 PI3K-Akt signaling pathway	26
04060 Cytokine-cytokine receptor interaction	23
04141 Protein processing in endoplasmic reticulum	17
05205 Proteoglycans in cancer	17
04610 Complement and coagulation cascades	15
04514 Cell adhesion molecules CAMs	14
04510 Focal adhesion	14
04512 ECM-receptor interaction	13
04145 Phagosome	13

3.4 Interaction network construction and identification of hub genes

DEGGs were mapped to the GeneMANIA to investigate how these genes interact with each other and additional genes that are related to a set of query genes by using a very large set of functional interaction data (Warde-Farley *et al.*, 2010). The genes showing significant interactions with weights higher than 0.1 were selected only for the network analysis. By integrating these relationships, a network between DEGGs and additional related genes was constructed for all the three groups viz., brain vs liver, brain vs muscle, and liver vs muscle (**Figure 3.2**). A GeneMANIA network analysis for interactions among the glycogene products suggested enrichment of GO terms related to positive regulation of locomotion, cell motility and migration for brain vs liver (**Table 3.4a**), as well as liver vs muscle samples (**Table 3.4c**). On the contrary, the processes related to extracellular matrix were over-represented in brain vs muscle samples (**Table 3.4b**).

Table 3.4. Top 10 enriched gene ontology (GO) terms for DEGs and additional related genes as reported by GeneMANIA for each category. a) Brain Vs Liver, b) Brain Vs Muscle, and c) Liver Vs Muscle.

(a) Brain Vs Liver		
GO id	Description	q-value
GO:0040017	positive regulation of locomotion	1.83E-23
GO:2000147	positive regulation of cell motility	1.83E-23
GO:0030335	positive regulation of cell migration	3.40E-23
GO:0051272	positive regulation of cellular component movement	3.40E-23
GO:0050817	coagulation	1.94E-20
GO:0010876	lipid localization	3.80E-20
GO:0007596	blood coagulation	4.54E-20
GO:0007599	hemostasis	7.56E-20
GO:0009897	external side of plasma membrane	2.15E-18
GO:0000323	lytic vacuole	1.05E-17
(b) Brain Vs Muscle		
GO:0031012	extracellular matrix	2.43E-23
GO:0005578	proteinaceous extracellular matrix	1.37E-17
GO:0009100	glycoprotein metabolic process	1.99E-16
GO:0000323	lytic vacuole	2.87E-16
GO:0005764	lysosome	2.87E-16
GO:0044420	extracellular matrix part	3.18E-15
GO:0005604	basement membrane	1.20E-14
GO:0005773	vacuole	4.56E-14
GO:0030335	positive regulation of cell migration	5.17E-12
GO:0005178	integrin binding	5.17E-12
(c) Liver Vs Muscle		
GO:0000323	lytic vacuole	1.09E-20
GO:0005764	lysosome	1.09E-20
GO:0051272	positive regulation of cellular component movement	8.78E-20
GO:0030335	positive regulation of cell migration	1.04E-19
GO:2000147	positive regulation of cell motility	1.79E-19
GO:0040017	positive regulation of locomotion	6.43E-19
GO:0005773	vacuole	1.70E-18
GO:0019838	growth factor binding	2.80E-18
GO:0009100	glycoprotein metabolic process	2.11E-15
GO:0031012	extracellular matrix	1.48E-14

The three interaction networks created from GeneMANIA were then exported to Cytoscape 2.8.2, a bioinformatics package for biological network visualization and data integration (Smoot *et al.*, 2011). The initial network for brain vs liver consists of 395 nodes and 1149 edges which was further filtered to 395 nodes and 1109 edges by

removing duplicate edges (**Figure 3.2a**). In case of brain vs muscle, the initial network consists of 352 nodes and 794 edges, whereas the filtered network has 352 nodes and 748 edges (**Figure 3.2b**). Similarly, the initial network for liver vs muscle network has 307 nodes and 695 edges, while the final network is represented by 307 nodes and 653 edges (**Figure 3.2c**). All the genes in the network are represented by circles and the interactions between them are represented as edges. Additionally, each query gene is shown in red where as the additional related genes predicted by GeneMANIA are shown in cyan (**Figure 3.2**).

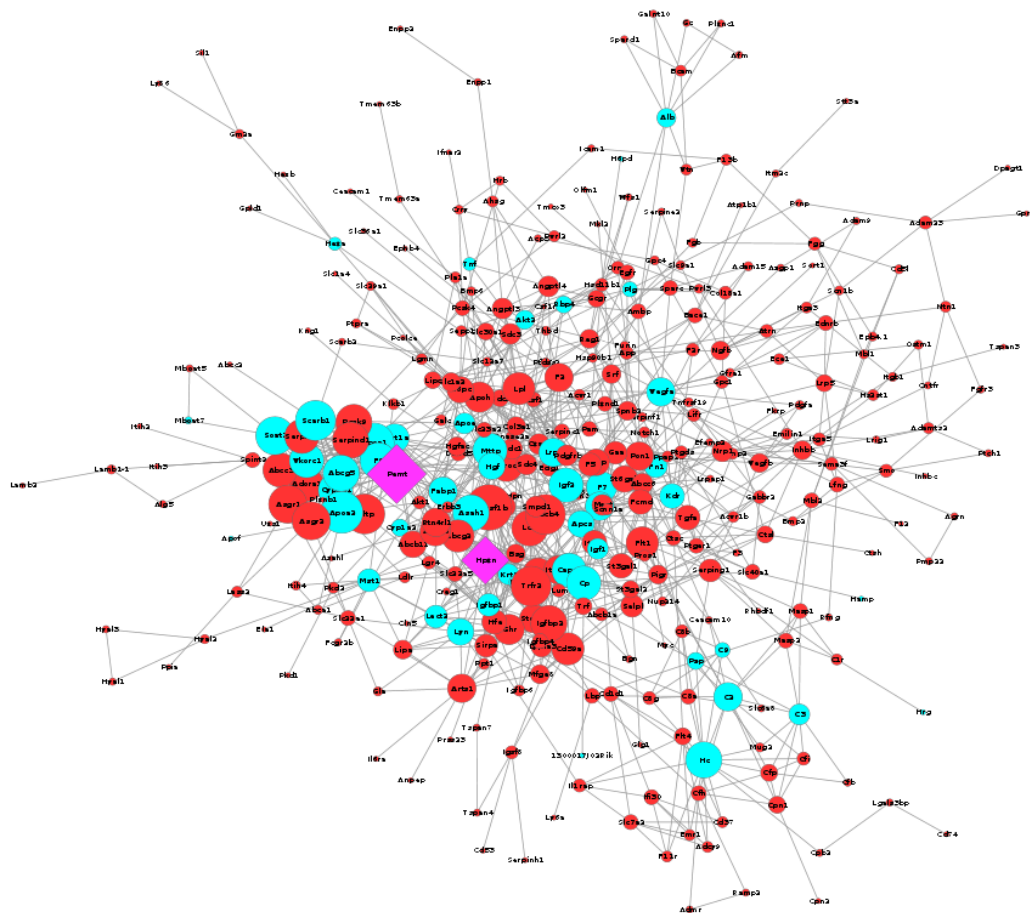


Figure 3.2(a)

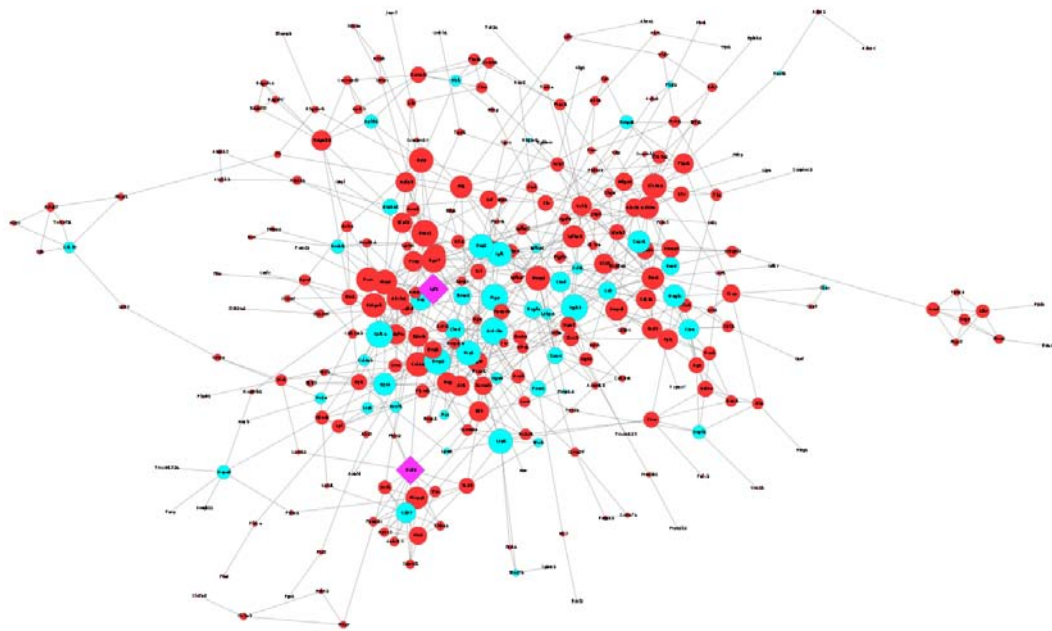


Figure 3.2(b)

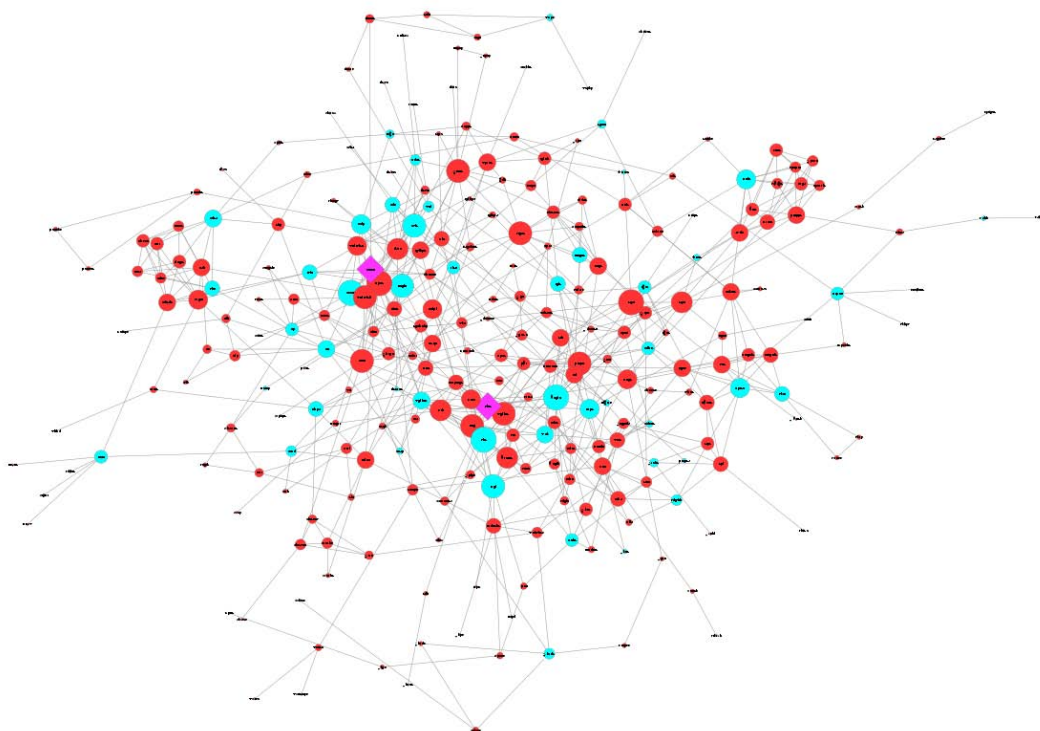


Figure 3.2(c)

Figure 3.2: Interaction networks between DEGs and additional related genes: (a) BvL579, (b) BvM501, and (c) LvM442. In each of these networks, red indicates a

DEGG whereas the GeneMANIA predicted genes are shown in cyan. The top two hubs are shown as purple diamonds. The size of the nodes represents the node degree.

The node degree was then calculated for all the nodes in each network by using Network Analyzer plugin of Cytoscape. The higher the node degree, the more important the gene was, and the gene was denoted as a hub gene. The genes with highest node degree include *PEMT* (Brain vs liver), and *IGF2* (brain vs muscle), having node degree of 29, and 14, respectively. Both *PEMT* and *IGF2* were predicted by GeneMANIA as related genes in their respective networks and do not represent the glycogenes. However, the genes with second highest node degree of 23 and 13 are represented by glycogenes *HPXN*, and *NID2* in brain vs liver and brain vs muscle networks (**Figure 3.3a and 3.3b**). Additionally, *NID2* was also predicted by GeneMANIA as a related gene in brain vs muscle network because it was removed from the query list as it showed a fold change of less than 2 in our analysis. In liver vs muscle network, the top node degree genes are *STAT6*, and the glycogene *FLT1* (**Figure 3.3c**), both having the node degree of 13.

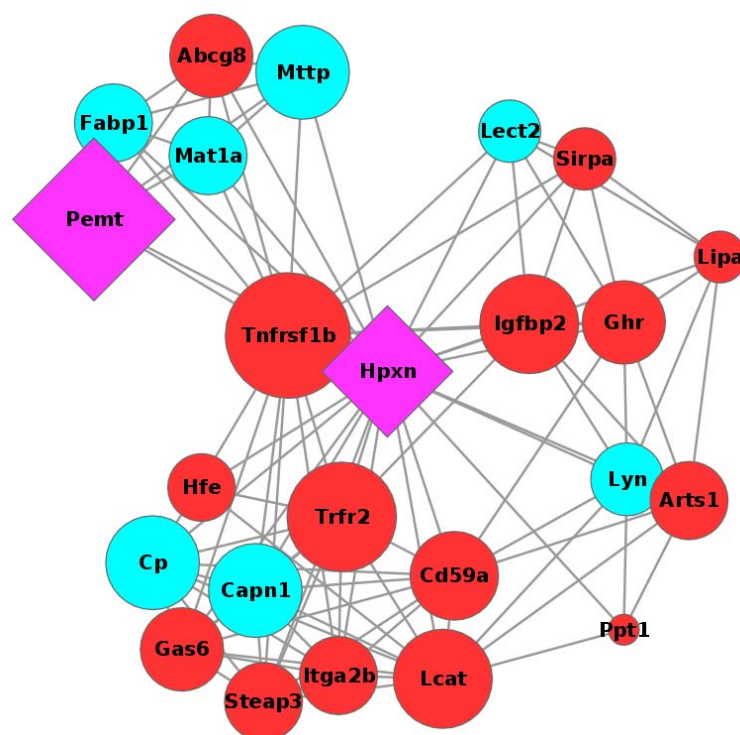


Figure 3.3(a)

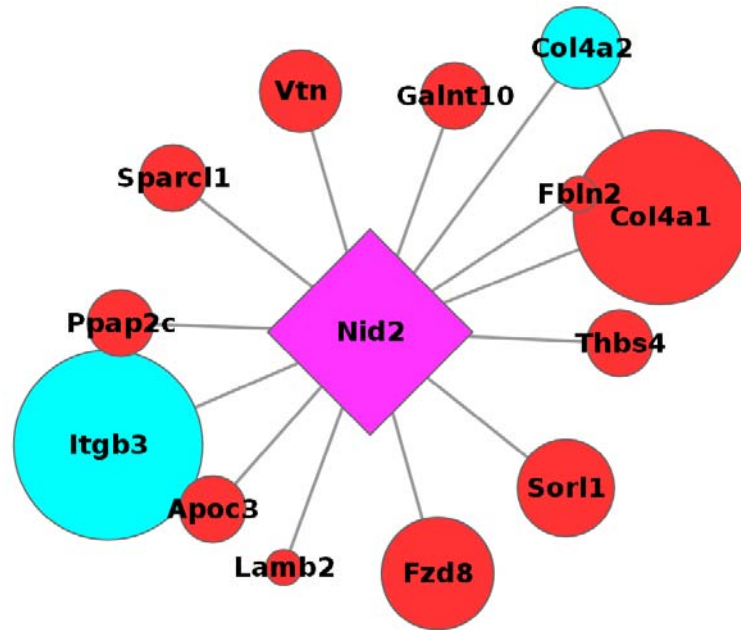


Figure 3.3(b)

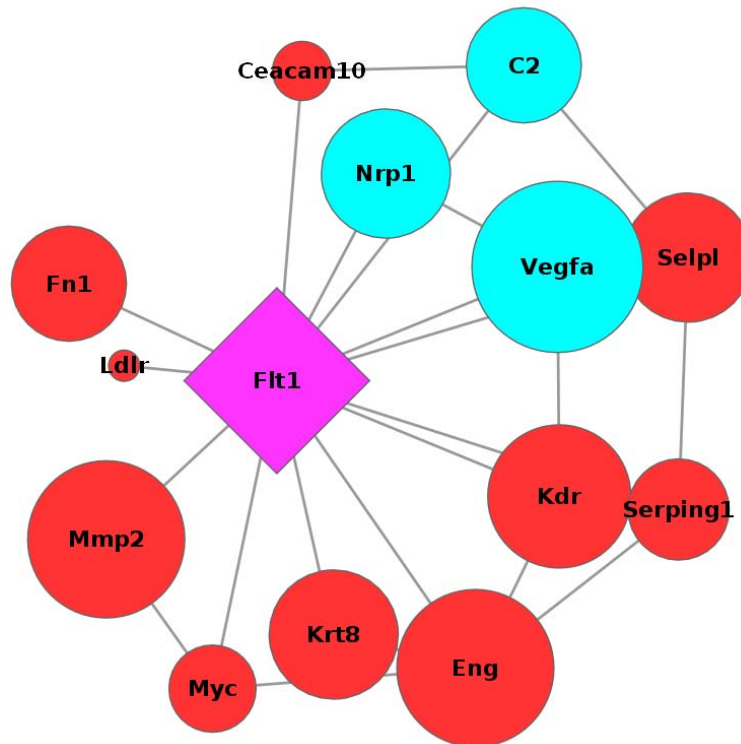


Figure 3.3(c)

Figure 3.3: Top glycogenes as hubs and their first neighbors. (a) *HPXN* gene as hubs in BvL579 interaction network. (b) *NID2* as hub gene in BvM501 interaction network. (c) *FLT1* gene as hub in LvM442 interaction network.

3.5 Community analysis and functional annotation of the detected modules

In order to identify the biologically related genes in the three networks, we employed Fast Greedy community-structure identification algorithm, implemented in the Cytoscape plug-in GLayer (<http://brainarray.mbni.med.umich.edu/sugang/glayer>) (Su *et al.*, 2010) to identify functional modules. 31 (**Figure 3.4a**), 40 (**Figure 3.4b**), and 31 (**Figure 3.4c**) clusters were detected for brain vs liver, brain vs muscle, and liver vs muscle networks, respectively. However, only those communities were selected for enrichment analyses which have at least 10 nodes. Therefore, based on this criterion, 8, 9, and 8 communities for brain vs liver, brain vs muscle, and liver vs muscle networks were finally analyzed for over-representation of GO terms.

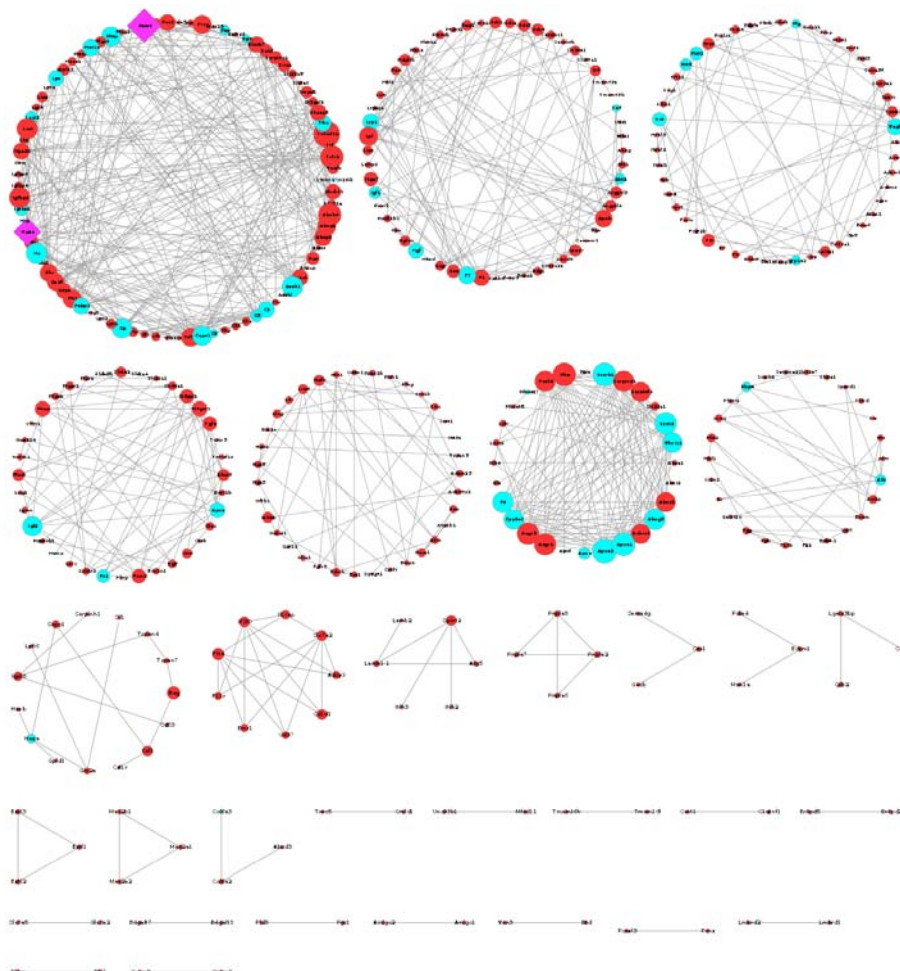


Figure 3.4(a)

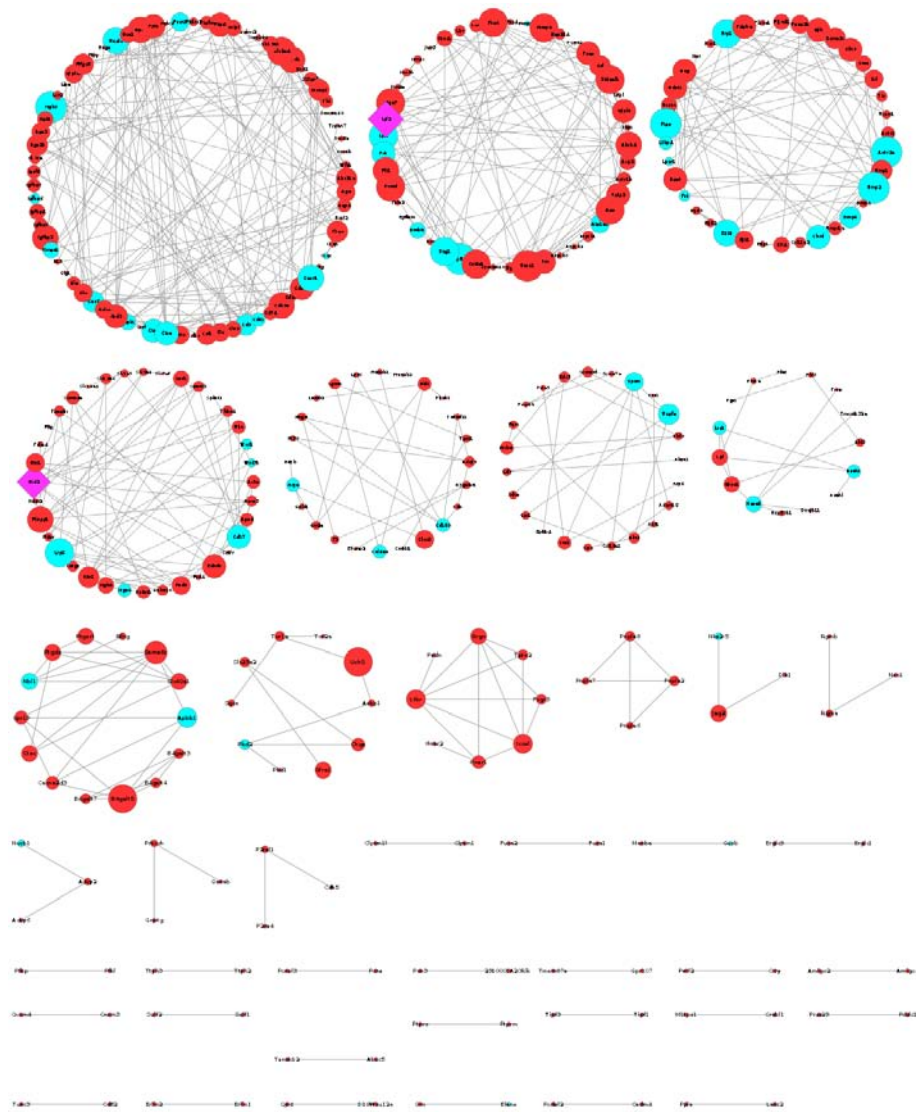
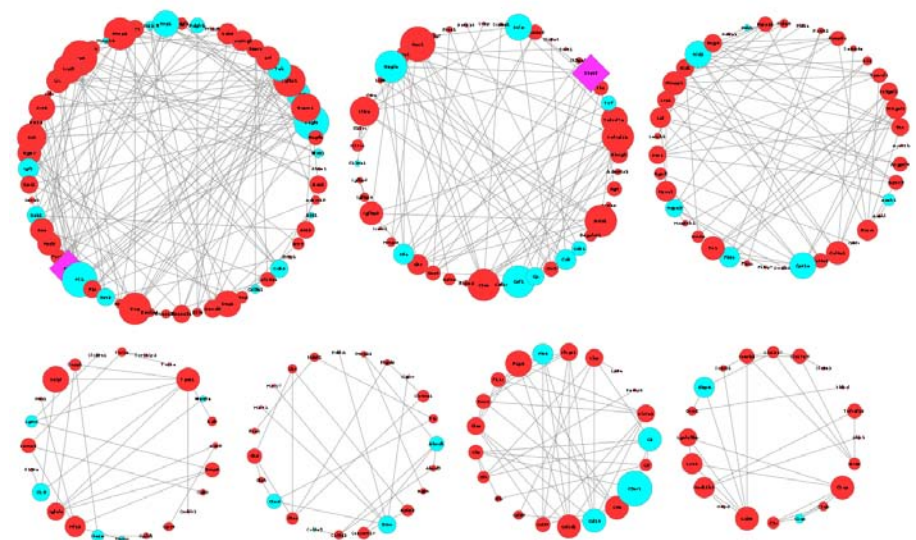


Figure 3.4(b)



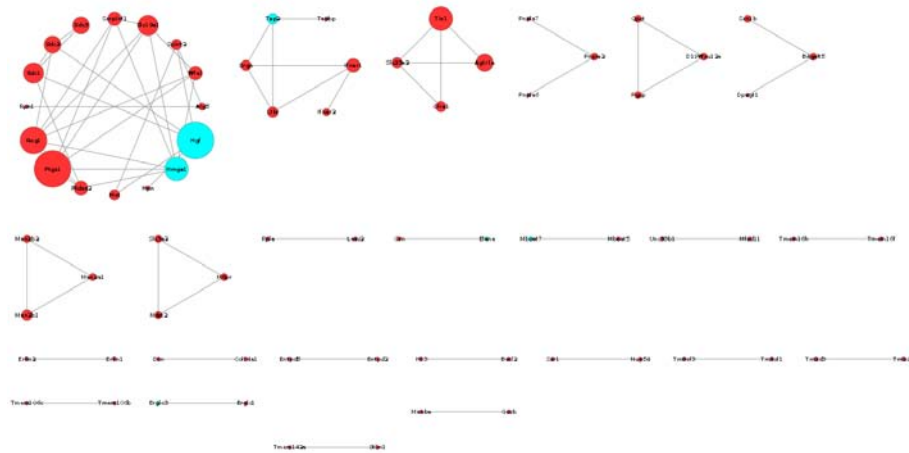


Figure 3.4(c)

Figure 3.4: Communities generated by fast greedy (G_{Lay}) clustering algorithm are shown. Clusters for (a) BvL579 network, (b) BvM501 network, and (c) LvM442 network are shown. In each community, DEGG nodes are represented by red circles whereas the cyan nodes represent GeneMANIA predicted genes. Hub genes are shown as purple diamonds.

To biologically categorize these clusters, DAVID functional analysis tool was used to classify the genes in each module, and observed the enrichment of GO term “Biological Process” in all the selected modules. The enrichment analysis for the modules for three networks is described as follows:

Brain vs liver (BvL579): The top 10 statistically significant enriched GO terms for DEGs in all 8 clusters for brain vs liver community analysis are summarized in (**Table 3.5a**). The most statistically significant GO terms that were enriched in cluster 1 were related to immune response, complement activation, and protein maturation. Cluster 2, 3 and 7 show enrichment for GO terms response to wounding, blood coagulation, and hemostasis. Cluster 2 and 3 also show over-representation of terms such as blood vessel development and vasculature development. The GO terms of cluster 4 were mostly related to hair cycle, development of epidermis and ectoderm. The GO functions of cluster 5 were most related to localization, motility and migration of cells, and regulation of cell proliferation. Cluster 6 was significantly enriched with functions related to homeostasis, transport, and metabolism of cholesterol, lipid as well as sterol. Cluster 8 showed over-representation of functional terms related to the metabolism of glycosphingolipid, glycolipid, sphingolipid, and ganglioside.

Brain vs muscle (BvM501): The top 10 statistically significant enriched GO terms for DEGs in all 9 clusters for brain vs muscle community analysis are summarized in (Table 3.5b). Cluster 1 of this network shows over-representation of functions related to regulation of cell growth, protein maturation, and immune response. Cluster 2 is enriched in GO terms related to proteoglycan metabolic process as well as the development of blood vessel and vasculature. Cluster 3 is enriched in the functions related to the development of bone, cartilage and skeletal system. The GO terms that are over-represented in cluster 4 show enrichment for functions related to adhesion, such as biological adhesion, cell adhesion, cell-substrate adhesion, and cell-matrix adhesion. The processes related to localization, motility and migration of cells are also over-represented in cluster 4. Cluster 5 showed over-representation of functional terms related to the metabolism of glycosphingolipid, glycolipid, sphingolipid, and ganglioside. Additionally, the functions related to the metabolism of lipid and oligosaccharides are also enriched in this group. In cluster 6, functions related to homeostasis, remodelling and regulation of transport are over-represented. Response to hypoxia and oxygen levels, apoptosis, and metabolic processes such as phospholipid, and organophosphate metabolic process are enriched in cluster 7. similar to cluster 4, the GO terms that are over-represented in Cluster 9 show enrichment for functions related to biological adhesion. Protein folding and metal ion transport are the other important enriched functions present in this cluster. Cluster 8 did not yield the enrichment of any statistically significant GO category.

Liver vs muscle (LvM442): Table 3.5c summarizes the top 10 statistically significant enriched GO terms for DEGs in all 8 clusters for liver vs muscle community analysis. The GO terms related to development and morphogenesis of blood vessels, vasculature and tube are enriched in cluster 1. Cluster 2 shows enrichment of functions related to various types of responses, such as response to wounding, external stimulus, and defence as well as inflammatory response. Other enriched functional groups in this cluster are regulation of cell growth, proliferation and cell activation. One of the over-represented functional groups in cluster 3 show enrichment for functions related to biological adhesion, lipid catabolic process, and 4-hydroxyproline metabolic process. Cluster 4 represents the enrichment of functions related to protein folding and bone

remodeling and resorption. Cluster 5 shows over-representation of GO terms such as regulation of T cell mediated cytotoxicity, or immunity, regulation of leukocyte mediated cytotoxicity, and regulation of cell killing. Cluster 6 and 7 are enriched in functions related to immune or inflammatory response, and complement activation. Finally, cluster 8 exhibited over-representation of signaling pathways such as hepatocyte growth factor receptor signaling, and mesenchymal-epithelial cell signaling, in addition to processes concerning MAP kinase activity.

Table 3.5: Enrichment of GO terms for Brain Vs Liver (BvL579), Brain Vs Muscle (BvM501) and Liver Vs Muscle (LvM442) cluster analysis.

(a) Brain Vs Liver

Cluster No.	Term	P-Value
1	GO:0002541~activation of plasma proteins involved in acute inflammatory response	5.02E-22
	GO:0006956~complement activation	5.02E-22
	GO:0002526~acute inflammatory response	2.26E-20
	GO:0006959~humoral immune response	2.25E-19
	GO:0051604~protein maturation	3.43E-19
	GO:0051605~protein maturation by peptide bond cleavage	3.19E-18
	GO:0002253~activation of immune response	3.20E-18
	GO:0016485~protein processing	6.27E-18
	GO:0006957~complement activation, alternative pathway	8.91E-18
	GO:0006954~inflammatory response	1.44E-17
2	GO:0051094~positive regulation of developmental process	5.33E-05
	GO:0042592~homeostatic process	4.25E-04
	GO:0009611~response to wounding	7.29E-04
	GO:0001568~blood vessel development	9.88E-04
	GO:0001944~vasculature development	0.001101764
	GO:0050817~coagulation	0.001394123
	GO:0007596~blood coagulation	0.001394123
	GO:0007599~hemostasis	0.001452434
	GO:0050878~regulation of body fluid levels	0.002774109
	GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	0.003100357
3	GO:0050817~coagulation	2.74E-11
	GO:0007596~blood coagulation	2.74E-11
	GO:0007599~hemostasis	3.08E-11
	GO:0050878~regulation of body fluid levels	1.97E-10
	GO:0042060~wound healing	1.26E-09
	GO:0009611~response to wounding	5.75E-08
	GO:0007167~enzyme linked receptor protein signaling pathway	9.33E-08
	GO:0001568~blood vessel development	5.52E-07
	GO:0001944~vasculature development	6.64E-07
	GO:0001525~angiogenesis	4.71E-05

4	GO:0022404~molting cycle process	2.16E-04
	GO:0022405~hair cycle process	2.16E-04
	GO:0001942~hair follicle development	2.16E-04
	GO:0042633~hair cycle	2.42E-04
	GO:0042303~molting cycle	2.42E-04
	GO:0008544~epidermis development	0.003102493
	GO:0007398~ectoderm development	0.00369522
	GO:0050678~regulation of epithelial cell proliferation	0.009595347
	GO:0048732~gland development	0.01092817
	GO:0007173~epidermal growth factor receptor signaling pathway	0.025611634
5	GO:0007166~cell surface receptor linked signal transduction	1.34E-05
	GO:0016477~cell migration	2.69E-05
	GO:0042127~regulation of cell proliferation	3.44E-05
	GO:0051674~localization of cell	4.88E-05
	GO:0048870~cell motility	4.88E-05
	GO:0006928~cell motion	6.35E-05
	GO:0008284~positive regulation of cell proliferation	2.52E-04
	GO:0007229~integrin-mediated signaling pathway	4.87E-04
	GO:0040007~growth	6.74E-04
	GO:0007507~heart development	0.001229375
6	GO:0055088~lipid homeostasis	2.16E-16
	GO:0042632~cholesterol homeostasis	2.64E-15
	GO:0055092~sterol homeostasis	2.64E-15
	GO:0030301~cholesterol transport	3.97E-13
	GO:0015918~sterol transport	3.97E-13
	GO:0033344~cholesterol efflux	7.23E-13
	GO:0006869~lipid transport	3.89E-12
	GO:0008203~cholesterol metabolic process	4.74E-12
	GO:0010876~lipid localization	8.01E-12
	GO:0016125~sterol metabolic process	1.02E-11
7	GO:0009611~response to wounding	3.88E-07
	GO:0007596~blood coagulation	2.94E-06
	GO:0050817~coagulation	2.94E-06
	GO:0007599~hemostasis	3.11E-06
	GO:0050878~regulation of body fluid levels	7.69E-06
	GO:0042060~wound healing	1.91E-05
	GO:0051818~disruption of cells of other organism during symbiotic interaction	3.08E-05
	GO:0051851~modification by host of symbiont morphology or physiology	3.08E-05
	GO:0051883~killing of cells in other organism during symbiotic interaction	3.08E-05
	GO:0031640~killing of cells of another organism	4.30E-05

8	GO:0006689~ganglioside catabolic process	5.11E-06
	GO:0046479~glycosphingolipid catabolic process	1.79E-05
	GO:0019377~glycolipid catabolic process	1.79E-05
	GO:0001573~ganglioside metabolic process	2.38E-05
	GO:0046466~membrane lipid catabolic process	1.15E-04
	GO:0030149~sphingolipid catabolic process	1.15E-04
	GO:0019915~lipid storage	1.29E-04
	GO:0006687~glycosphingolipid metabolic process	2.73E-04
	GO:0006664~glycolipid metabolic process	3.90E-04
	GO:0050885~neuromuscular process controlling balance	4.98E-04

(b) Brain Vs Muscle

1	GO:0040008~regulation of growth	3.73E-08
	GO:0051604~protein maturation	7.27E-08
	GO:0001558~regulation of cell growth	1.29E-06
	GO:0016485~protein processing	2.32E-05
	GO:0016044~membrane organization	7.38E-05
	GO:0016064~immunoglobulin mediated immune response	9.61E-05
	GO:0051605~protein maturation by peptide bond cleavage	1.09E-04
	GO:0019724~B cell mediated immunity	1.09E-04
	GO:0002684~positive regulation of immune system process	1.29E-04
	GO:0006958~complement activation, classical pathway	2.00E-04
2	GO:0009100~glycoprotein metabolic process	2.16E-06
	GO:0050654~chondroitin sulfate proteoglycan metabolic process	9.34E-06
	GO:0006029~proteoglycan metabolic process	8.82E-05
	GO:0048754~branching morphogenesis of a tube	1.02E-04
	GO:0019800~peptide cross-linking via chondroitin 4-sulfate glycosaminoglycan	1.89E-04
	GO:0001763~morphogenesis of a branching structure	3.20E-04
	GO:0001568~blood vessel development	4.29E-04
	GO:0001944~vasculature development	4.79E-04
	GO:0030204~chondroitin sulfate metabolic process	6.09E-04
	GO:0006813~potassium ion transport	8.13E-04
3	GO:0001501~skeletal system development	1.07E-16
	GO:0030509~BMP signaling pathway	6.34E-14
	GO:0001503~ossification	4.02E-12
	GO:0060348~bone development	7.43E-12
	GO:0030500~regulation of bone mineralization	6.08E-11
	GO:0007178~transmembrane receptor protein serine/threonine kinase signaling pathway	7.75E-11
	GO:0070167~regulation of biomineral formation	9.25E-11
	GO:0007167~enzyme linked receptor protein signaling pathway	1.84E-10
	GO:0007369~gastrulation	2.22E-10
	GO:0051216~cartilage development	3.31E-10

4	GO:0006928~cell motion	9.88E-05
	GO:0016055~Wnt receptor signaling pathway	2.59E-04
	GO:0016477~cell migration	4.37E-04
	GO:0048870~cell motility	7.09E-04
	GO:0051674~localization of cell	7.09E-04
	GO:0007160~cell-matrix adhesion	0.001189285
	GO:0009100~glycoprotein metabolic process	0.001250564
	GO:0031589~cell-substrate adhesion	0.001570006
	GO:0007155~cell adhesion	0.005368157
	GO:0022610~biological adhesion	0.005405426
5	GO:0006689~ganglioside catabolic process	1.66E-05
	GO:0019377~glycolipid catabolic process	5.78E-05
	GO:0046479~glycosphingolipid catabolic process	5.78E-05
	GO:0001573~ganglioside metabolic process	7.69E-05
	GO:0009611~response to wounding	2.01E-04
	GO:0046466~membrane lipid catabolic process	3.70E-04
	GO:0030149~sphingolipid catabolic process	3.70E-04
	GO:0019915~lipid storage	4.16E-04
	GO:0006687~glycosphingolipid metabolic process	8.77E-04
	GO:0009311~oligosaccharide metabolic process	0.001247784
6	GO:0042592~homeostatic process	6.03E-06
	GO:0043270~positive regulation of ion transport	0.001725638
	GO:0051270~regulation of cell motion	0.002725255
	GO:0048771~tissue remodeling	0.003048492
	GO:0001894~tissue homeostasis	0.003842174
	GO:0051050~positive regulation of transport	0.004095371
7	GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	0.004147057
	GO:0032869~cellular response to insulin stimulus	0.004461779
	GO:0042127~regulation of cell proliferation	0.004903945
	GO:0010765~positive regulation of sodium ion transport	0.005901212
	GO:0001666~response to hypoxia	0.006023043
	GO:0070482~response to oxygen levels	0.006648281
	GO:0016042~lipid catabolic process	0.009864913
	GO:0006644~phospholipid metabolic process	0.011805904
8	GO:0019637~organophosphate metabolic process	0.013020542
	GO:0006916~anti-apoptosis	0.013774773
	GO:0043066~negative regulation of apoptosis	0.037899056
	GO:0043069~negative regulation of programmed cell death	0.038883095
	GO:0060548~negative regulation of cell death	0.03908111
8	NA	NA

9	GO:0007160~cell-matrix adhesion	0.001495186
	GO:0031589~cell-substrate adhesion	0.001809131
	GO:0051085~chaperone mediated protein folding requiring cofactor	0.006635187
	GO:0051084~'de novo' posttranslational protein folding	0.008618097
	GO:0007155~cell adhesion	0.00916406
	GO:0022610~biological adhesion	0.009200307
	GO:0050982~detection of mechanical stimulus	0.010597488
	GO:0006458~'de novo' protein folding	0.010597488
	GO:0007259~JAK-STAT cascade	0.025656487
	GO:0030001~metal ion transport	0.036163834
	GO:0030001~metal ion transport	0.036163834
(c) Liver Vs Muscle		
1	GO:0001568~blood vessel development	1.06E-18
	GO:0001944~vasculature development	1.64E-18
	GO:0048514~blood vessel morphogenesis	8.45E-16
	GO:0001525~angiogenesis	3.48E-12
	GO:0048754~branching morphogenesis of a tube	9.17E-11
	GO:0007167~enzyme linked receptor protein signaling pathway	5.50E-10
	GO:0035239~tube morphogenesis	1.14E-09
	GO:0001763~morphogenesis of a branching structure	1.32E-09
	GO:0035295~tube development	5.53E-09
	GO:0001569~patterning of blood vessels	3.29E-08
2	GO:0040008~regulation of growth	4.27E-07
	GO:0009611~response to wounding	4.18E-06
	GO:0006954~inflammatory response	3.42E-05
	GO:0032101~regulation of response to external stimulus	1.89E-04
	GO:0048585~negative regulation of response to stimulus	7.82E-04
	GO:0006952~defense response	0.001399518
	GO:0001558~regulation of cell growth	0.002134607
	GO:0050867~positive regulation of cell activation	0.002784166
	GO:0032102~negative regulation of response to external stimulus	0.004518678
	GO:0008283~cell proliferation	0.004842417
3	GO:0007160~cell-matrix adhesion	7.84E-05
	GO:0031589~cell-substrate adhesion	1.14E-04
	GO:0044242~cellular lipid catabolic process	9.80E-04
	GO:0016042~lipid catabolic process	9.93E-04
	GO:0051346~negative regulation of hydrolase activity	0.006807864
	GO:0008218~bioluminescence	0.007742186
	GO:0018401~peptidyl-proline hydroxylation to 4-hydroxy-L-proline	0.007742186
	GO:0019471~4-hydroxyproline metabolic process	0.007742186

	GO:0009101~glycoprotein biosynthetic process	0.007780584
	GO:0007155~cell adhesion	0.00843311
	GO:0001501~skeletal system development	7.15E-04
	GO:0006955~immune response	0.011579
	GO:0051085~chaperone mediated protein folding requiring cofactor	0.013230724
4	GO:0051084~'de novo' posttranslational protein folding	0.01716759
	GO:0045453~bone resorption	0.019783921
	GO:0006458~'de novo' protein folding	0.021089619
	GO:0046849~bone remodeling	0.027593503
	GO:0006952~defense response	0.045862312
	GO:0001916~positive regulation of T cell mediated cytotoxicity	4.08E-05
	GO:0001914~regulation of T cell mediated cytotoxicity	6.22E-05
	GO:0002711~positive regulation of T cell mediated immunity	8.81E-05
	GO:0002474~antigen processing and presentation of peptide antigen via MHC class I	1.53E-04
5	GO:0002709~regulation of T cell mediated immunity	1.72E-04
	GO:0031343~positive regulation of cell killing	2.84E-04
	GO:0001912~positive regulation of leukocyte mediated cytotoxicity	2.84E-04
	GO:0001910~regulation of leukocyte mediated cytotoxicity	4.86E-04
	GO:0031341~regulation of cell killing	4.86E-04
	GO:0002821~positive regulation of adaptive immune response	6.25E-04
	GO:0002526~acute inflammatory response	4.67E-14
	GO:0045087~innate immune response	4.65E-13
	GO:0002684~positive regulation of immune system process	1.53E-12
6	GO:0050778~positive regulation of immune response	3.29E-12
	GO:0006954~inflammatory response	3.41E-12
	GO:0002541~activation of plasma proteins involved in acute inflammatory response	4.05E-12
	GO:0006956~complement activation	4.05E-12
	GO:0048584~positive regulation of response to stimulus	4.13E-11
	GO:0006952~defense response	4.75E-11
	GO:0006959~humoral immune response	5.29E-11
	GO:0009611~response to wounding	2.97E-06
	GO:0006958~complement activation, classical pathway	7.07E-06
	GO:0002526~acute inflammatory response	7.33E-06
7	GO:0002455~humoral immune response mediated by circulating immunoglobulin	8.69E-06
	GO:0006956~complement activation	2.20E-05
	GO:0002541~activation of plasma proteins involved in acute inflammatory response	2.36E-05

	GO:0016064~immunoglobulin mediated immune response	4.70E-05
	GO:0019724~B cell mediated immunity	5.25E-05
	GO:0002449~lymphocyte mediated immunity	1.02E-04
	GO:0002250~adaptive immune response	1.36E-04
8	GO:0051450~myoblast proliferation	0.002206368
	GO:0048012~hepatocyte growth factor receptor signaling pathway	0.002206368
	GO:0060665~regulation of branching involved in salivary gland morphogenesis by mesenchymal-epithelial signaling	0.003674845
	GO:0060638~mesenchymal-epithelial cell signaling	0.004408354
	GO:0060693~regulation of branching involved in salivary gland morphogenesis	0.006605966
	GO:0060688~regulation of morphogenesis of a branching structure	0.018252999
	GO:0001889~liver development	0.031208893
	GO:0000187~activation of MAPK activity	0.036205594
	GO:0043406~positive regulation of MAP kinase activity	0.042595839

3.6 Discussion

The current study offers the first thorough insight into the glycogene analysis of brain, muscle and liver tissues from mouse RNA-Seq data. Understanding the structure and function of these glycogenes is essential for studying the development of various tissues as well as their functional roles. Most of the serum glycoproteins originate from liver suggesting that liver diseases associated with aberrant glycosylation can be reflected by the changes in serum glycoproteins (Adamczyk *et al.*, 2012). Recently, the significance of glycomics of central nervous system (CNS) to identify potential glyco-biomarkers in neurological diseases (Barone *et al.*, 2012), and alterations in brain glycoproteins resulting from the aging process was shown (Sato *et al.*, 2010). Recent studies in the glycomics field also offered insights into the biological significance of the glycome in the pathogenesis of diseases in humans (Adamczyk *et al.*, 2012). Additionally, many studies have highlighted the significance of glycoconjugates during skeletal muscle development (Cornelison *et al.*, 2004; Janot *et al.*, 2009; Lee *et al.*, 2014).

With the evolution of bioinformatics during the past decades, molecular target discovery and targeted therapeutics have become a critical remedial treatment for diseases (Dinh *et al.*, 2007). In this work, we screened DEGs between brain, muscle

and liver tissues using RNA-Seq data and performed functional and pathway enrichment analysis using DAVID and KEGG tools. The most significant enriched GO terms for brain vs liver are processes related to response to wounding or inflammatory response. These processes are represented by genes that play a prominent roles in innate immunity such as Mannan-binding lectin serine protease 1 (MASP1), Mannan-binding lectin serine protease 2 (MASP2), Mannose-binding protein A (MBL1), and Mannose-binding protein C (MBL2) (Nauta *et al.*, 2003; Møller-Kristensen *et al.*, 2007) Brain vs muscle is enriched in functions related to glycoprotein metabolic processes and includes genes that code for various transferases such as N-acetylglucosamine-1-phosphotransferase subunit gamma (GNPTG) (Qian *et al.*, 2010), Beta-galactoside alpha-2,6-sialyltransferase 1 (ST6GAL1) (Kuhn *et al.*, 2013), and Beta-1,3-galactosyltransferase 6 (B3GALT6) (Bai *et al.*, 2001) Similarly, the top statistically significant enriched processes in liver vs muscle are related to protein maturation, or processing and are represented by genes such as *Battenin* (CLN3), and *Methionine aminopeptidase 2* (METAP2). CLN3 codes for batten disease protein which is involved in microtubule-dependent, anterograde transport of late endosomes and lysosomes (Uusi-Rauva *et al.*, 2012), where as METAP2 protein catalyzes the hydrolytic cleavage of the N-terminal methionine from newly synthesized polypeptides (Xiao *et al.*, 2010).

Additionally, three interaction networks of DEGs were constructed for brain vs liver, brain vs muscle, and liver vs muscle samples, and node degrees of genes were calculated. Hemopexin (HPXN), Nidogen-2 (NID2), and Vascular endothelial growth factor receptor 1 (FLT1) were the glycoproteins with the highest degree in brain vs liver, brain vs muscle, and liver vs muscle networks, respectively. However, these genes represent the second highest node degree in their respective networks as genes (PEMT, IGF2, and STAT6) with top node degree do not code for glycoproteins.

HPXN (~60kDa glycoprotein) is mainly synthesized by liver cells (Chen-Roetling *et al.*, 2012), secreted to the plasma where it binds free heme or hemin and inhibits its role in free radical reactions (Tolosano *et al.*, 2010) HPXN is also reported to be expressed by the cells of immune systems, ganglionic and photoreceptor cells of the retina, cells of the peripheral nervous system, and by the mesangial cells of kidney (Camborieu *et al.*, 1997; Hunt *et al.*, 1996; Kapojos *et al.*, 2003). Recently, it was shown that HPX protein

is also present in various regions of mouse brain (Li *et al.*, 2009). NID2 is homologous to another member of the nidogen family, NID1, and both are found in all basement membranes (BMs) (Bechtel *et al.*, 2012). Both nidogens have a similar distribution in various organs during development, however, in adult tissues nidogen-2 distribution becomes more confined (Kohfeldt *et al.*, 1998; Murshed *et al.*, 2000; Salmivirta *et al.*, 2002). Additionally, these nidogens show a broad range of interacting partners including other BM proteins such as laminin, collagen IV, and perlecan (Kohfeldt *et al.*, 1998; Salmivirta *et al.*, 2002; Fox *et al.*, 1991; Ho *et al.*, 2008). They are involved in various functions including the regulation of cell attachment (Chakravarti *et al.*, 1990), neutrophil chemotaxis (Senior *et al.*, 1992), trophoblast outgrowth (Yelian *et al.*, 1993), angiogenesis (Nicosia *et al.*, 1994), osteoblast and myogenic differentiation (Kimura *et al.*, 1998; Neu *et al.*, 2006). Vascular endothelial growth factors (VEGFs) constitute a family of six polypeptides (VEGF-A, -B, -C, -D, -E and PlGF), that regulate blood and lymphatic vessel development (Grünewald *et al.*, 2010). VEGF signaling occurs by binding to various cellular receptors such as VEGFR1 (FLT1), VEGFR2 (FLK1) and VEGFR3 (FLT4) (Terman *et al.*, 1991; Shibuya *et al.*, 1990), and neuropilin-1 (Pajusola *et al.*, 1992), and -2 (NRP1 and NRP2) (Neufeld *et al.*, 2002), and heparan sulfate proteoglycans (HSPG) (Tessler *et al.*, 1994). FLT1 and FLK1 are closely related receptor tyrosine kinases and both share common and specific ligands (Nishi *et al.*, 2008). FLT1 has weaker kinase activity than FLK1 (Nishi *et al.*, 2008), however, FLT1 is essential for normal development and angiogenesis as reported in previous FLT1 null mutant mice studies (Fong *et al.*, 1995; Fong *et al.*, 1999; Kearney *et al.*, 2002).

Recognizing the structure and function of biological networks is indispensable for the investigation of biological processes. In this work, we identified 8, 9, and 8 functional modules or communities for brain vs liver, brain vs muscle, and liver vs muscle networks using fast greedy algorithm implemented as GLAY (Su *et al.*, 2010), plugin for cytoscape. Furthermore, the function of each module in all the three networks was explored by using functional annotation tool DAVID. Our analysis shows that the modules that are common in brain vs liver and liver vs muscle networks show enrichment for processes related to immune and inflammatory response, and response to wounding. Liver is known to play an important role in innate immunity, an important primary line of defense against infection (Gao *et al.*, 2008). Additionally, Kupffer cells in the liver are one of the earliest to be affected by bacterial or sterile insults and add to

the inflammatory response (Bieghs *et al.*, 2013). This sterile inflammation can be responsible for liver injury as well as may also play a role in liver repair (Gao *et al.*, 2008). In case of the modules of brain vs muscle and liver vs muscle networks, the enrichment of processes related to cell adhesion, proliferation, and regulation of cell growth is observed. The expression of myogenic regulatory factors (MRFs) describe the different stages of skeletal muscle development that includes myoblast proliferation, cell-cycle exit, cell fusion and the maturation of myotubes to form myofibers (Lee *et al.*, 2014). Cell adhesion molecules such as cadherins are glycoproteins that mediate homotypic cell–cell adhesion through their extracellular domain (Charrasse *et al.*, 2003), and this cadherin-dependent adhesion is necessary for a diverse types of cellular functions (Knudsen *et al.*, 1990). Finally, glycosphingolipid or glycolipid metabolic processes show over-representation in modules of brain vs liver and brain vs muscle networks. Glycosphingolipids are universally expressed in all vertebrate cells as well as body fluids, but they are more rich in the nervous system (Yu *et al.*, 2009). Previous studies have established the role of glycosphingolipids in the development of the brain among various species (Svennerholm *et al.*, 1989; Yu *et al.*, 1988).

References

Adamczyk, B.; Tharmalingam, T.; Rudd, PM. Glycans as cancer biomarkers. *Biochimica et Biophysica Acta* **820**. 1347–1353, (2012).

Bai, X.; Zhou, D.; Brown, JR.; et al., Biosynthesis of the linkage region of glycosaminoglycans: cloning and activity of galactosyltransferase II, the sixth member of the β 1,3-galactosyltransferase family (β 3GalT6). *Journal of Biological Chemistry*. **276**. 48189-48195, (2001).

Barone, R.; Sturiale, L.; Palmigiano, A.; et al., Glycomics of pediatric and adulthood diseases of the central nervous system. *Journal of Proteomics*, **75**. 5123–5139, (2012).

Bechtel, M.; Keller, M. V.; Bloch W.; et al. Different domains in nidogen-1 and nidogen-2 drive basement membrane formation in skin organotypic cocultures. *The FASEB Journal*. **26**.3637-3648, (2012).

Bieghs, V.; Trautwein, C. The innate immune response during liver inflammation and metabolic disease. *Trends in Immunology*. **34**. 446-452, (2013).

Camborieu, L.; Bertrand, N.; Swerts, J. P. Changes in expression and localization of hemopexin and its transcripts in injured nervous system: a comparison of central and peripheral tissues. *Neuroscience*. **82**.1039-1052, (1997).

Chakravarti, S.; Tam, M. F.; Chung, A. E. The basement membrane glycoprotein entactin promotes cell attachment and binds calcium ions. *The Journal of Biological Chemistry*. **265**. 10597-10603, (1990).

Charrasse, S.; Causeret, M.; Comunale, F.; et al., Rho GTPases and cadherin-based cell adhesion in skeletal muscle development. *Journal of Muscle Research and Cell Motility*. **24**. 309-313, (2003).

Chen-Roetling, J.; Liu, W.; Regan, R. F. Hemopexin decreases hemin accumulation and catabolism by neural cells. *Neurochemistry International*. 60. 488-494, (2012).

Cornelison, DDW.; Wilcox-Adelman, SA.; Goetinck, PF. Et al., Essential and separable roles for Syndecan-3 and Syndecan-4 in skeletal muscle development and regeneration. *Genes and Development*. **18**. 2231–2236, (2004).

Dinh, PC.; Sotiriou.; Piccart, MJ. The evolution of treatment strategies: aiming at the target. *Breast*. **16**.10–16, (2007).

Fong, G.; Zhang, L.; Bryce, D.; et al., Increased hemangioblast commitment, not vascular disorganization, is the primary defect in flt-1 knock-out mice. *Development*. **126**. 3015-3025, (1999).

Fong, G.H.; Rossant, J.; Gertsenstein, M.; et al., Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium. *Nature*. **376**. 66–70, (1995).

Fox, J. W.; Mayer, U.; Nischt, R.; et al., Recombinant nidogen consists of three globular domains and mediates binding of laminin to collagen type IV. *EMBO Journal*. **10**. 3137-3146, (1991).

Gao, B.; Jeong, W.; Tian, Z. Liver: an organ with predominant innate immunity. *Hepatology*. **47**. 729-736, (2008).

Grunewald, F. S.; Prota, A. E.; Giese, A.; et al., Structure-function analysis of VEGF receptor activation and the role of coreceptors in angiogenic signaling. *Biochimica et Biophysica Acta: Proteins and Proteomics*. **1804**. 567–580, (2010).

Ho, M. S. P.; Bose, K.; Mokkapati, S.; et al., Nidogens-extracellular matrix linker molecules. *Microscopy Research and Technique*. **71**. 387-395, (2008).

Hunt, R. C.; Hunt, D. M.; Gaur, N.; et al., Hemopexin in the human retina: protection of the retina against hememediated toxicity. *Journal of Cellular Physiology*. **168**. 71–80, (1996).

Janot, M.; Audfray, A.; Lorient, C.; et al., Glycogenome expression dynamics during mouse C2C12 myoblast differentiation suggests a sequential reorganization of membrane glycoconjugates. *BMC Genomics*. **10**. 483, (2009).

Kapojos, J. J.; van Den Berg, A.; van Goor, H.; et al., Production of hemopexin by TNF- α stimulated human mesangial cells. *Kidney International*. **63**. 1681–1686, (2003).

Kearney, J. B.; Ambler, C. A.; Monaco, K.-A.; et al., Vascular endothelial growth factor receptor Flt-1 negatively regulates developmental blood vessel formation by modulating endothelial cell division. *Blood*. **99**, 2397–2407, (2002).

Kimura, N.; Toyoshima, T.; Kojima, T.; et al., Entactin-2: a new member of basement membrane protein with high homology to entactin/nidogen. *Experimental Cell Research*. **241**. 36–45, (1998).

Knudsen, K. A.; Myers, L.; McElwee, S. A. A role for the Ca²⁺-dependent adhesion molecule, N-Cadherin, in myoblast interaction during myogenesis. *Experimental Cell Research*. **188**. 175–184, (1990).

Kohfeldt, E.; Sasaki, T.; Gohring, W.; et al., Nidogen-2: a new basement membrane protein with diverse binding properties. *Journal of Molecular Biology*. **282**. 99–109, (1998).

Kuhn, B.; Benz, J.; Greif, M.; et al., The structure of human α -2, 6-sialyltransferase reveals the binding mode of complex glycans. *Acta Crystallographica D*. **69**. 1826–1838, (2013).

Lee, E.J.; Malik, A.; Pokharel, S.; et al., Identification of genes differentially expressed in myogenin knock-down bovine muscle satellite cells during differentiation through RNA sequencing analysis. *PLoS ONE*. **9**. Article ID e92447, (2014).

Li, R.-C.; Saleem, S.; Zhen G.; et al., Heme-hemopexin complex attenuates neuronal cell death and stroke damage. *Journal of Cerebral Blood Flow & Metabolism*. **29**. 953–964, (2009).

Møller-Kristensen, M.; Thiel, S.; Sjöholm, A. et al., Cooperation between MASP-1 and MASP-2 in the generation of C3 convertase through the MBL pathway. *International Immunology*. **19**. 141–149, (2007).

Mortazavi, A.; Williams, B.A.; McCue, K.; et al., Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, **5**. 621-8 (2008).

Murshed, M.; Smyth, N.; Miosge, N.; et al., The absence of nidogen 1 does not affect murine basement membrane formation. *Molecular and Cellular Biology*, **20**. 7007-7012, (2000).

Nauta, A.J.; Raashou-Jensen, N.; Roos, A. et al., Mannose-binding lectin engagement with late apoptotic and necrotic cells. *European Journal of Immunology*. **33**, 2853-2863, (2003).

Neu, R.; Adams, S.; Munz, B. Differential expression of entactin-1/nidogen-1 and entactin-2/nidogen-2 in myogenic differentiation. *Differentiation*, **74**. 573–582, (2006).

Neufeld, G.; Kessler, O.; Herzog, Y. The interaction of neuropilin-1 and neuropilin-2 with tyrosine-kinase receptors for VEGF. *Advances in Experimental Medicine and Biology*. **5**. 81–90, (2002).

Nicosia, R. F.; Nicosia, S. V.; Smith, M.; Vascular endothelial growth factor, platelet-derived growth factor, and insulin-like growth factor-1 promote rat aortic angiogenesis in vitro. *The American Journal of Pathology*. **45**. 1023–1029, (1994).

Nishi, J.; Minamino, T.; Miyauchi, H.; et al., Vascular endothelial growth factor receptor-1 regulates postnatal angiogenesis through inhibition of the excessive activation of Akt. *Circulation Research*. **103**, 261–268, (2008).

Pajusola, K.; Aprelikova, O.; Korhonen J.; et al., FLT4 receptor tyrosine kinase contains seven immunoglobulin-like loops and is expressed in multiple human tissues and cell lines. *Cancer Research*. **52**. 5738–5743, (1992).

Qian, Y.; Lee, I.; Lee W.; et al., Functions of the α , β , and γ subunits of UDP-GlcNAc:Lysosomal enzyme N-acetylglucosamine-1-phosphotransferase. *The Journal of Biological Chemistry*. **285**. 3360–3370. (2010).

Salmivirta, K.; Talts, J. F.; Olsson, M.; et al., Binding of mouse nidogen-2 to basement membrane components and cells and its expression in embryonic and adult tissues suggest complementary functions of the two nidogens. *Experimental Cell Research*. **279**. 188–201, (2002).

Sato, Y.; Endo, T. Alteration of brain glycoproteins during aging. *Geriatrics and Gerontology International*. **10**. S32–S40, (2010).

Senior, R. M.; Gresham, H. D.; Griffin, G. L.; et al., Entactin stimulates neutrophil adhesion and chemotaxis through interactions between its Arg-Gly-Asp (RGD) domain and the leukocyte response integrin. *Journal of Clinical Investigation*. **90**. 2251–2257, (1992).

Shibuya, M.; Yamaguchi, S.; Yamane, A.; et al., Nucleotide sequence and expression of a novel human receptor-type tyrosine kinase gene (flt) closely related to the fms family. *Oncogene*. **5**. 519–524, (1990).

Smoot, ME.; Ono, K.; Ruscheinski, J. et al., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. **27**. 431–432, (2011).

Su, G.; Kuchinsky, A.; Morris, J. H.; et al., GLay: community structure analysis of biological networks. *Bioinformatics*. **26**. 3135–3137, (2010).

Su, G.; Kuchinsky, A.; Morris, JH.; et al., GLay: community structure analysis of biological networks. *Bioinformatics*. **26**. 3135–3137, (2010).

Svennerholm, L.; Bostrom, K.; Fredman, P.; et al., Human brain gangliosides: developmental changes from early fetal stage to advanced age. *Biochimica et Biophysica Acta*. **1005**. 109–117, (1989).

Terman, B. I.; Carrion, M. E.; Kovacs, E.; et al., Identification of a new endothelial cell growth factor receptor tyrosine kinase. *Oncogene*. **6**. 1677–1683, (1991).

Tessler, S.; Rockwell, P.; Hicklin, D.; et al., Heparin modulates the interaction of VEGF165 with soluble and cell associated flk-1 receptors. *Journal of Biological Chemistry*. **269**, 12456–12461, (1994).

Tolosano, E.; Fagoonee, S. Morello, N.; et al. Heme scavenging and the other facets of hemopexin. *Antioxidants & Redox Signaling*. **12**. 305-320, (2010).

Uusi-Rauva, K.; Kytala, A.; Van Der Kant, V.; et al., Neuronal ceroid lipofuscinosis protein CLN3 interacts with motor proteins and modifies location of late endosomal compartments. *Cellular and Molecular Life Sciences*. **69**. 2075-2089, (2012).

Warde-Farley, D., Donaldson, S.L., Comes, O.; et al., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. **38**. W214–W220 (2010).

Xiao, Q.; Zhang, F.; Nacev, BA.; et al., Protein N terminal processing: substrate specificity of Escherichia coli and human methionine amino peptidases. *Biochemistry*. **49**. 5588-5599, (2010).

Yelian, F. D.; Edgeworth, N. A.; Dong, L.J.; et al., Recombinant entactin promotes mouse primary trophoblast cell adhesion and migration through the Arg-Gly- Asp (RGD) recognition sequence. *Journal of Cell Biology*. **121**. 923–929, (1993).

Yu, R. K.; Macala, L. J.; Taki, T.; et al., Developmental changes in ganglioside composition and synthesis in embryonic rat brain. *Journal of Neurochemistry*. **50**. 1825-1829, (1988).

Yu, R. K.; Nakatani, Y.; Yanagisawa, M. The role of glycosphingolipid metabolism in the developing brain. *Journal of Lipid Research*. **50**. S440–S445, (2009).

Identification of Hub Glycogenes and their nsSNP Analysis

Contents		Page number
4.1	Introduction	63
4.2	Identification of glycogenes	63
4.3	Functional enrichment and pathway analysis of glycogenes	63
4.4	Analysis of interaction network and hub genes	72
4.5	SNP analysis	75
4.5.1	Deleterious or damaged nsSNPs by PROVEAN	76
4.5.2	Predicted nsSNPs that affect protein function by SIFT	77
4.5.3	Damaged nsSNPs by PolyPhen-2	77
4.6	Protein surface accessibility and secondary structure prediction by NetSurfP	79
4.7	Database and Tools development	82
4.8	Discussion	85
	References	86

4.1 Introduction

The overall methodology adopted in the study discussed in this chapter is depicted as a flowchart in Figure 4.1.

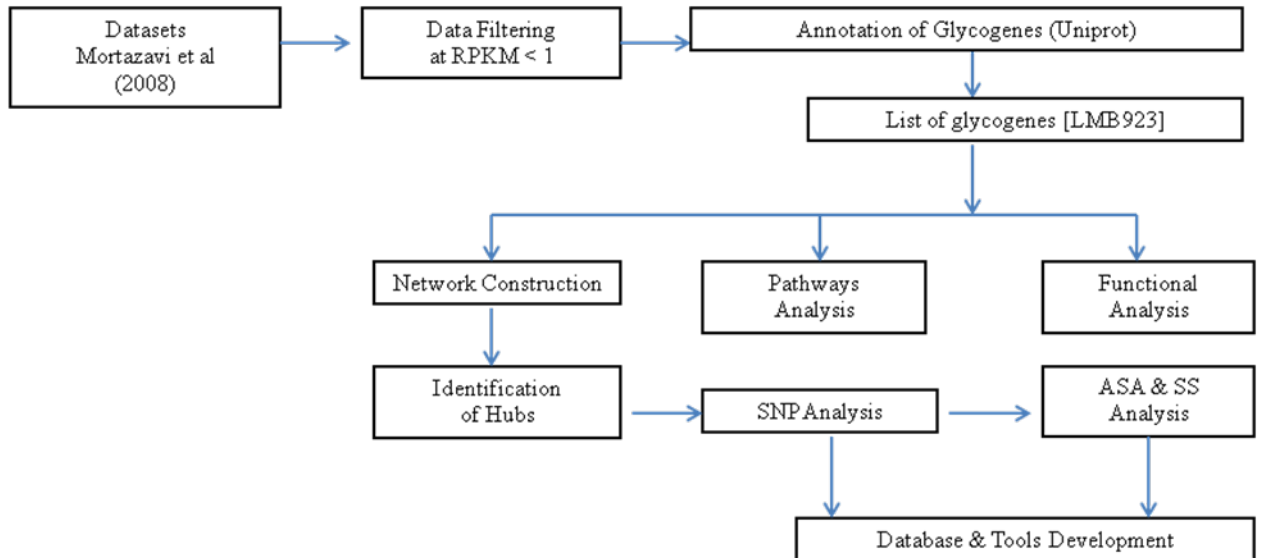


Figure 4.1 flowchart diagram

4.2 Identification of glycogenes

In the previous chapter 3, we explored the role of differentially expressed glycogenes (DEGGs) and identified hub genes using the precomputed expression values from mouse RNA-Seq data (Mortazavi *et al.*, 2008). In this work, we created a non-redundant list of 923 glycogenes (**Supplementary Table 4S1**) from this RNA-Seq data by combining all identified glycogenes of brain, muscle and liver tissues. The data was then filtered by removing any duplicate gene.

All these 923 glycogenes are annotated as “glycoproteins” in the UniProt (Magrane *et al.*, 2011) database and were further employed for functional, pathway and network analysis.

4.3 Functional enrichment and pathway analysis of glycogenes

In order classify enriched biological processes that may be regulated by glycogenes, a non-redundant list was created by pooling all glycogenes from mouse brain, liver and muscle tissues and then removing the duplicates. Functional Annotation Clustering (FAC) tool available in the Database for Annotation, Visualization, and Integrated

Discovery (DAVID) was used to classify over-represented biological processes using GO term “Biological Process” for the annotations. The functional annotation cluster analysis of top ten clusters (enrichment score ≥ 4.60) reveals a diverse range of biological processes in which these glycogenes are involved (**Table 4.1**). Cluster 1 (enrichment score = 12.72) and 7 (enrichment score = 5.11) mainly consists of genes that either involved in glycoprotein metabolic process or protein glycosylation. Cluster 2 with an enrichment score of 12.18 is probably the largest cluster among the 10 clusters as it consists of maximum number of statistically significant GO terms specifically related to immune response and includes terms such as acute inflammatory response, complement activation, immunoglobulin mediated immune response, etc. The third functional cluster (enrichment score = 10.03) consists of genes related to glycan metabolic processes, whereas cluster 4 (8.88) and cluster 5 (6.75) include genes with known functions in blood coagulation, hemostasis blood vessel development and blood vessel morphogenesis. Cluster 6 (enrichment score = 5.92) and 8 (enrichment score = 4.98) are either involved in organic acid and amino acid transport, or endocytosis, membrane organization, and vesicle-mediated transport. Cluster 9 (enrichment score = 4.88) has over-represented GO terms related to metabolic processes of cholesterol, sterol or steroid, whereas cluster 10 (enrichment score = 4.62) shows enrichment of processes that are concerned with antigen processing and presentation of peptide antigen.

In addition to DAVID FAC analysis, we attempted to map glycogenes on various biological pathways by submitting FASTA formatted amino acid sequences to KAAS web server that predicted 222 pathways in which these glycogenes may play a functional role. Top 20 KEGG pathways are summarized in **Table 4.2** and a complete list of all pathways is provided in **Supplementary Table 4S2**. From these tables, it can be seen that glycogenes are involved in diverse types of biological pathways such as signaling pathways (PI3K-Akt and Rap1 signaling pathway), complement and coagulation cascades, cytokine-cytokine receptor interaction, pathways in cancer or adhesion related functions (cell adhesion molecules CAMs and focal adhesion) (**Figure 4.2a-c**).

Table 4.1 Top 10 functional annotation clusters by DAVID

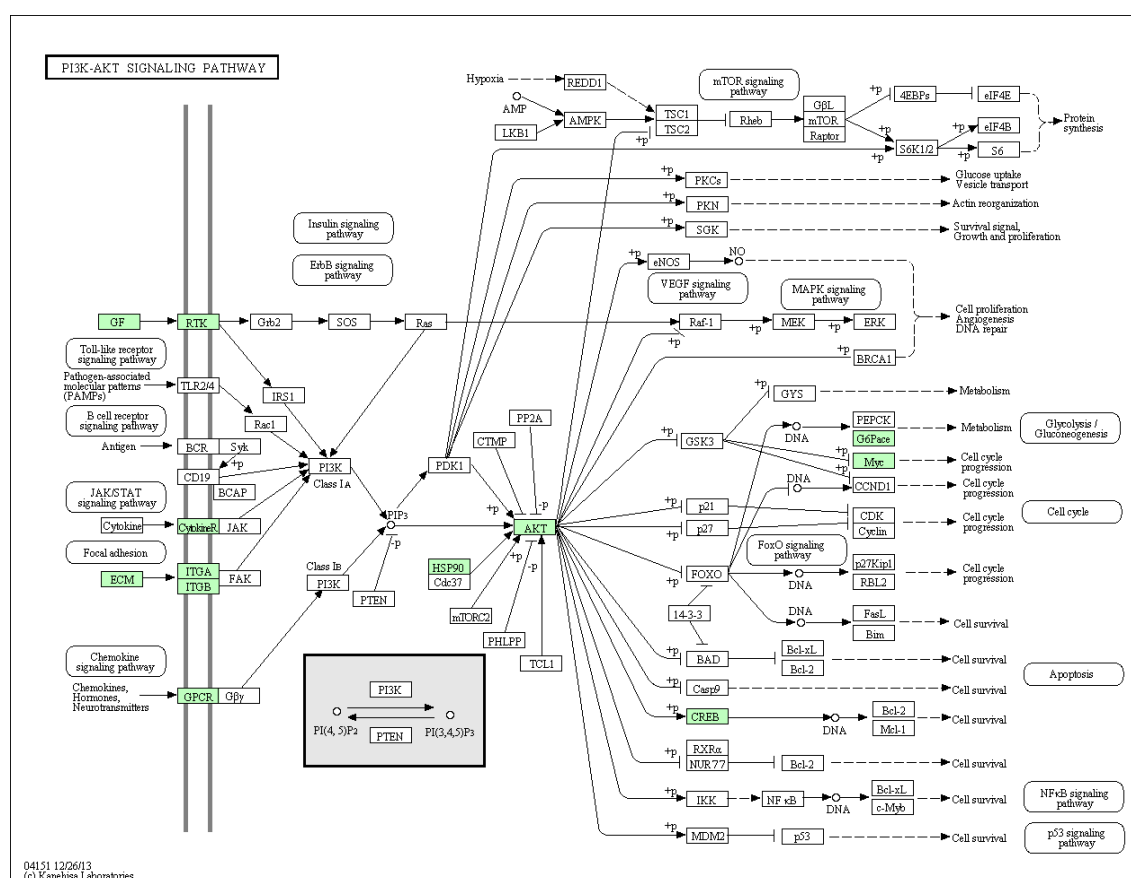
Annotation Cluster (Enrichment Score)	GO Term	P-value
1 (12.72)	GO:0009100~glycoprotein metabolic process	2.74E-21
	GO:0009101~glycoprotein biosynthetic process	1.27E-16
	GO:0043413~biopolymer glycosylation	2.24E-12
	GO:0006486~protein amino acid glycosylation	2.24E-12
	GO:0070085~glycosylation	2.24E-12
	GO:0006487~protein amino acid N-linked glycosylation	1.13E-05
2 (12.18)	GO:0009611~response to wounding	2.62E-23
	GO:0051604~protein maturation	1.66E-20
	GO:0002526~acute inflammatory response	7.90E-20
	GO:0016485~protein processing	2.50E-18
	GO:0006954~inflammatory response	4.29E-17
	GO:0051605~protein maturation by peptide bond cleavage	4.32E-16
	GO:0002541~activation of plasma proteins involved in acute inflammatory response	3.44E-15
	GO:0006956~complement activation	3.44E-15
	GO:0006958~complement activation, classical pathway	7.26E-13
	GO:0016064~immunoglobulin mediated immune response	3.54E-12
	GO:0006952~defense response	3.73E-12
	GO:0019724~B cell mediated immunity	6.81E-12
	GO:0002455~humoral immune response mediated by circulating immunoglobulin	1.22E-11
	GO:0006959~humoral immune response	1.69E-11
	GO:0002252~immune effector process	1.92E-11
	GO:0002449~lymphocyte mediated immunity	2.03E-11
	GO:0002460~adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.16E-11
	GO:0002250~adaptive immune response	2.16E-11
	GO:0002443~leukocyte mediated immunity	7.47E-11
	GO:0050778~positive regulation of immune response	1.23E-10
	GO:0002253~activation of immune response	1.75E-09
	GO:0045087~innate immune response	3.26E-09
	GO:0002684~positive regulation of immune system process	6.94E-09
	GO:0048584~positive regulation of response to stimulus	9.42E-09
3 (10.03)	GO:0006955~immune response	4.11E-08
	GO:0006957~complement activation, alternative pathway	1.79E-06
	GO:0030203~glycosaminoglycan metabolic process	9.03E-12
4 (8.88)	GO:0006022~aminoglycan metabolic process	8.84E-11
	GO:0005976~polysaccharide metabolic process	9.90E-10
	GO:0050817~coagulation	2.60E-10
	GO:0007596~blood coagulation	2.60E-10
	GO:0007599~hemostasis	3.39E-10

	GO:0042060~wound healing	7.98E-09
	GO:0050878~regulation of body fluid levels	2.02E-08
5 (6.75)	GO:0001568~blood vessel development	9.69E-09
	GO:0001944~vasculature development	1.81E-08
	GO:0048514~blood vessel morphogenesis	1.70E-06
	GO:0001525~angiogenesis	3.35E-06
6 (5.92)	GO:0046942~carboxylic acid transport	4.10E-09
	GO:0015849~organic acid transport	4.97E-09
	GO:0006865~amino acid transport	4.03E-06
	GO:0015837~amine transport	2.09E-05
	GO:0015807~L-amino acid transport	0.0014557
7 (5.11)	GO:0009311~oligosaccharide metabolic process	1.03E-07
	GO:0006487~protein amino acid N-linked glycosylation	1.13E-05
	GO:0006491~N-glycan processing	3.89E-04
8 (4.98)	GO:0006897~endocytosis	1.75E-07
	GO:0010324~membrane invagination	1.75E-07
	GO:0016044~membrane organization	1.05E-05
	GO:0006909~phagocytosis	1.19E-05
	GO:0016192~vesicle-mediated transport	0.02974658
9 (4.88)	GO:0008203~cholesterol metabolic process	3.38E-06
	GO:0016125~sterol metabolic process	1.09E-05
	GO:0008202~steroid metabolic process	5.92E-05
10 (4.622)	GO:0048002~antigen processing and presentation of peptide antigen	1.84E-10
	GO:0002474~antigen processing and presentation of peptide antigen via MHC class I	1.70E-08
	GO:0019882~antigen processing and presentation	7.94E-08
	GO:0019884~antigen processing and presentation of exogenous antigen	2.88E-05
	GO:0002478~antigen processing and presentation of exogenous peptide antigen	5.87E-05
	GO:0019886~antigen processing and presentation of exogenous peptide antigen via MHC class II	0.00542203
	GO:0002495~antigen processing and presentation of peptide antigen via MHC class II	0.00542203
	GO:0002504~antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	0.01033599
	GO:0042590~antigen processing and presentation of exogenous peptide antigen via MHC class I	0.01972544

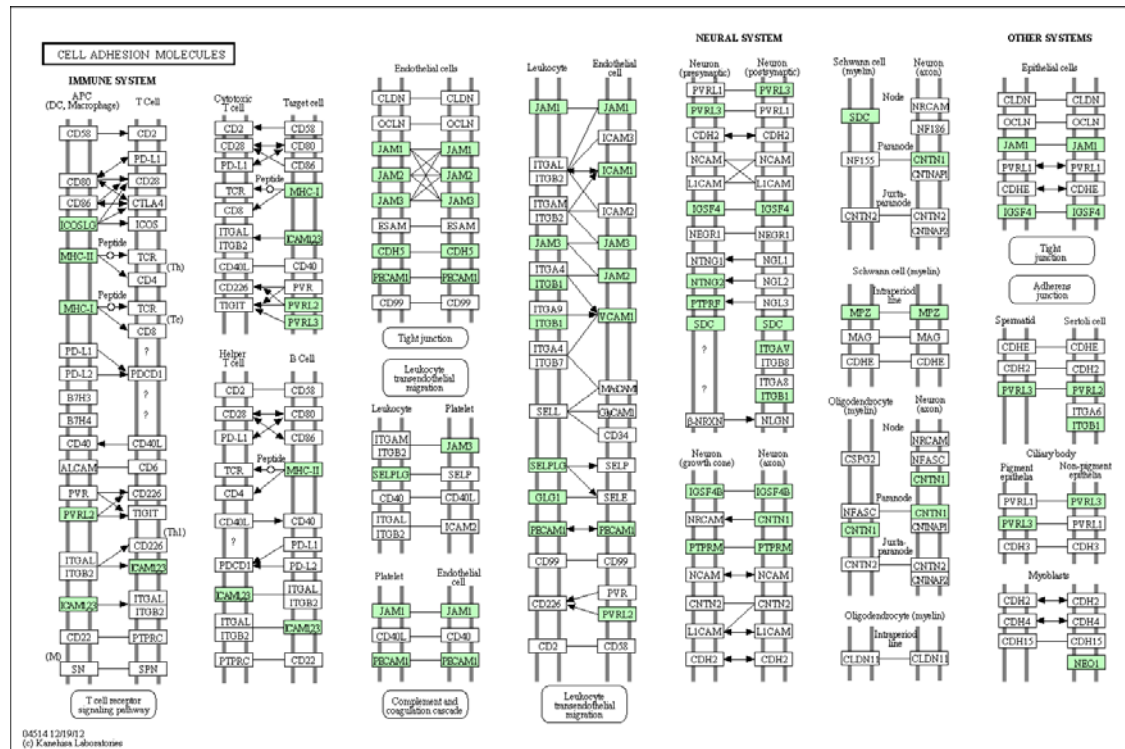
Table 4.2: Top 20 KEGG pathways for glycogens

Pathway Name	No. of Mapped Genes
04142 Lysosome	49
04151 PI3K-Akt signaling pathway	45
04610 Complement and coagulation cascades	39
04060 Cytokine-cytokine receptor interaction	38
05200 Pathways in cancer	30
04514 Cell adhesion molecules CAMs	28
05205 Proteoglycans in cancer	27
04510 Focal adhesion	26
04141 Protein processing in endoplasmic reticulum	24
04015 Rap1 signaling pathway	23
04512 ECM-receptor interaction	21
04976 Bile secretion	21
05166 HTLV-I infection	21
04145 Phagosome	20
04080 Neuroactive ligand-receptor interaction	18
05150 Staphylococcus aureus infection	18
04014 Ras signaling pathway	17
04360 Axon guidance	17
04810 Regulation of actin cytoskeleton	16
04640 Hematopoietic cell lineage	16

a)



b)



c)

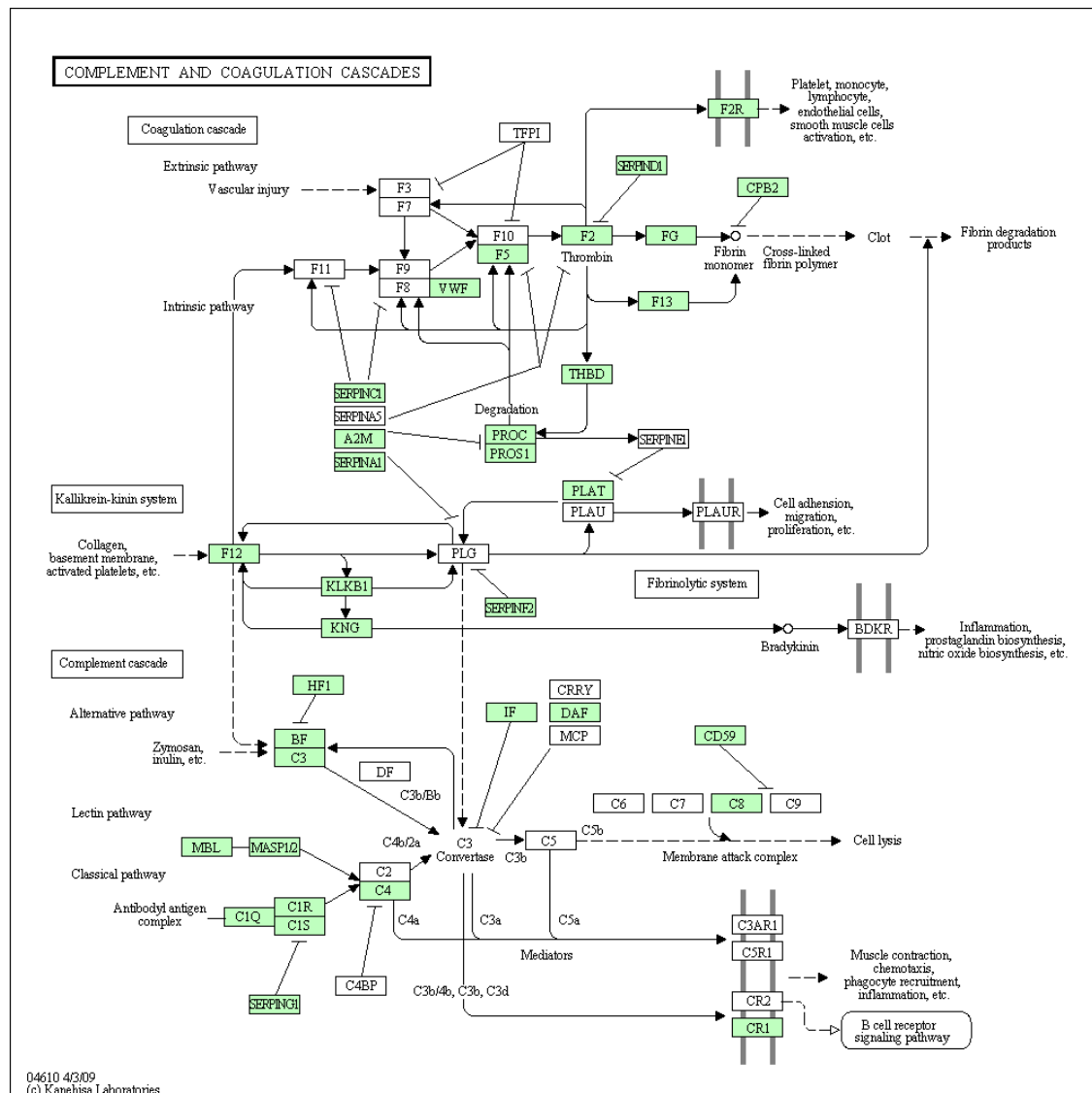


Figure. 4.2. Biochemical pathway analysis of glycogenes. Representative KEGG pathway for a) PI3K–AKT signaling, b) Cell adhesion molecules, and c) Complement and coagulation cascades.

4.4 Analysis of interaction network and hub genes

GeneMANIA plug-in for Cytoscape (Montejo *et al.*, 2010) was used to map glycogenes to probe interactions between these as well as additional genes that may be associated to a list of query genes by using a very large set of functional interaction data (Wardle-Farley *et al.*, 2010). Only genes that exhibited noteworthy interactions having weights ≥ 0.1 were selected for the network analysis. Furthermore, these relationships were integrated to construct an interaction network between glycogenes and additional correlated genes (**Figure 4.3**). Network analysis by GeneMANIA for interactions among the glycogene products suggested enrichment of diverse GO terms that may be involved in lysosome, locomotion, extracellular matrix, glycoprotein metabolic process and adhesion related processes (**Table 4.3**).

The interaction network was further analyzed in Cytoscape 2.8.2, a bioinformatics package for biological network visualization and data integration (Smoot *et al.*, 2011). All the island nodes, duplicate edges and self-loops were removed from the initial network. The final network consists of 658 nodes (circles) that represent genes and 2114 edges indicating the interactions between them. Moreover, the query glycogenes are shown as red and the genes predicted by GeneMANIA are shown in cyan (**Figure 4.3**).

Table 4.3: Top 20 functional categories of glycogenes as additional related genes as predicted by GeneMANIA.

GO id	Description	q-value
GO:0000323	lytic vacuole	4.75E-29
GO:0005764	lysosome	4.75E-29
GO:0009897	external side of plasma membrane	3.57E-28
GO:0051272	positive regulation of cellular component movement	7.39E-26
GO:0040017	positive regulation of locomotion	8.39E-26
GO:0031012	extracellular matrix	8.39E-26
GO:2000147	positive regulation of cell motility	8.39E-26
GO:0030335	positive regulation of cell migration	1.40E-25
GO:0005773	Vacuole	1.19E-24
GO:0010876	lipid localization	6.67E-21
GO:0009100	glycoprotein metabolic process	1.30E-20
GO:0051604	protein maturation	6.11E-20
GO:0016485	protein processing	3.97E-18
GO:0001525	angiogenesis	6.31E-18
GO:0031589	cell-substrate adhesion	9.23E-18
GO:0019838	growth factor binding	9.23E-18

GO:0050817	coagulation	1.75E-17
GO:0007596	blood coagulation	3.48E-17
GO:0007599	hemostasis	6.45E-17
GO:0005578	proteinaceous extracellular matrix	8.65E-17

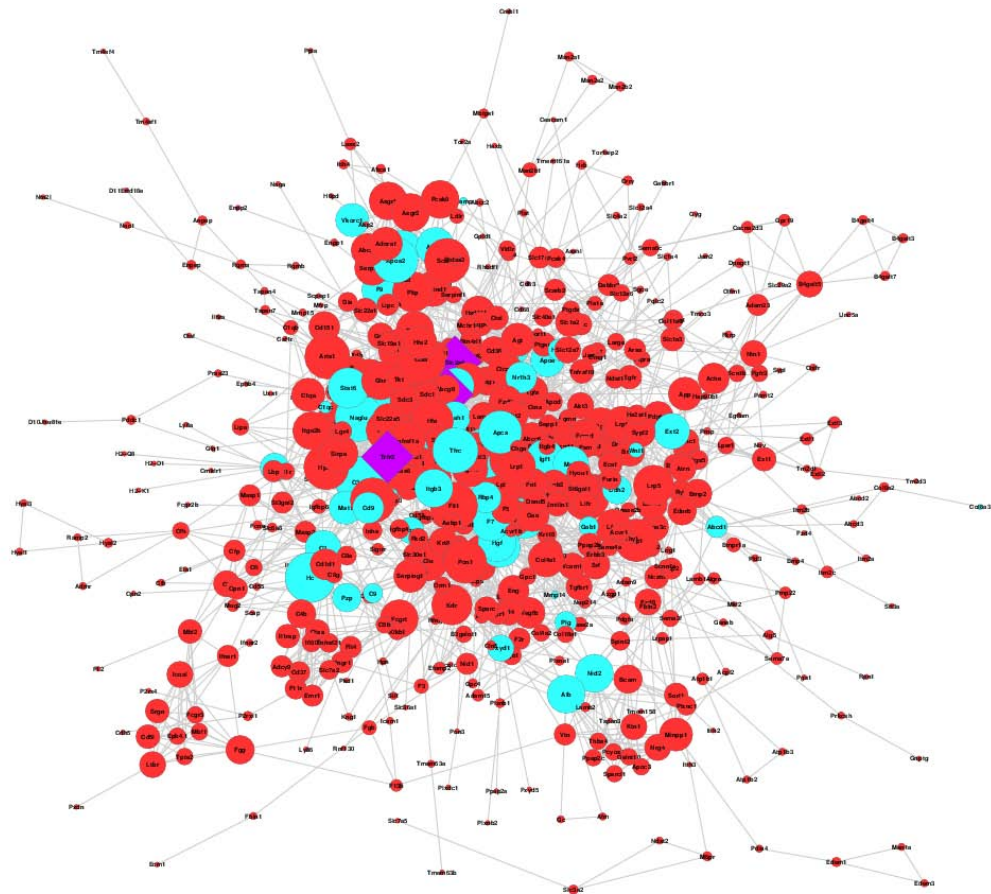


Figure. 4.3. Interaction networks between glycogenes and additional related genes. A red node indicates a query glycogenes and the genes predicted by GeneMANIA are shown in cyan. The top two hubs are shown as purple diamonds.

Network Analyzer plugin of Cytoscape was then used to calculate node degree for all nodes in the network. Nodes with higher node degree were considered as more important genes and were denoted as hub genes. The glycogenes *SLC2A4*, *TNFRSF1B*, *TRFR2* and *UCHL1* showed a node degree of 21 and therefore represent hubs in the current network. **Figure 4.4** shows the interaction of these hubs (purple nodes) and their first neighbors. These four hub glycogenes were also a part of initial list of input genes submitted to GeneMANIA for predicting interactions.

d)

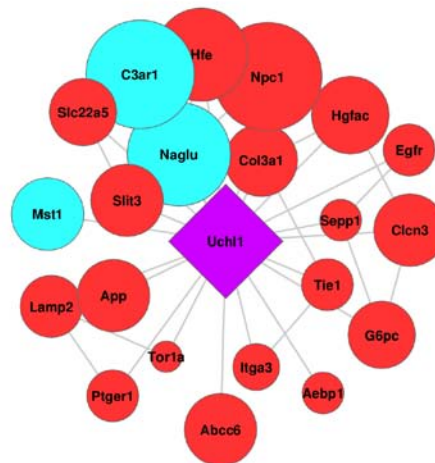


Figure. 4.4. Top glycocones as hubs identified in the current work and their first neighbors. A) SLC2A4, B) TNFRSF1B, C) TRFR2 and D) UCHL1. The size of the nodes represents the node degree.

4.5 SNP Analysis

In this work, we have selected coding non synonymous SNPs (nsSNPs) for our computational analysis. **Figure 4.5** shows the distribution of total number of SNPs and nsSNPs identified for mouse hub genes. It can be observed that HPXN has 278 SNPs and 7 nsSNPs, whereas FLT1 has 4929 SNPs and 23 nsSNPs which are highest among all identified hub genes in both SNPs and nsSNPs. NID2 has 1624 SNPs but has only 21 nsSNPs, whereas TNFRSF1B has 830 SNPs and only 11 nsSNPs. However, no nsSNPs were detected in the remaining three hubs (SLC2A4, UCHL1 and TRFR2) for mouse.

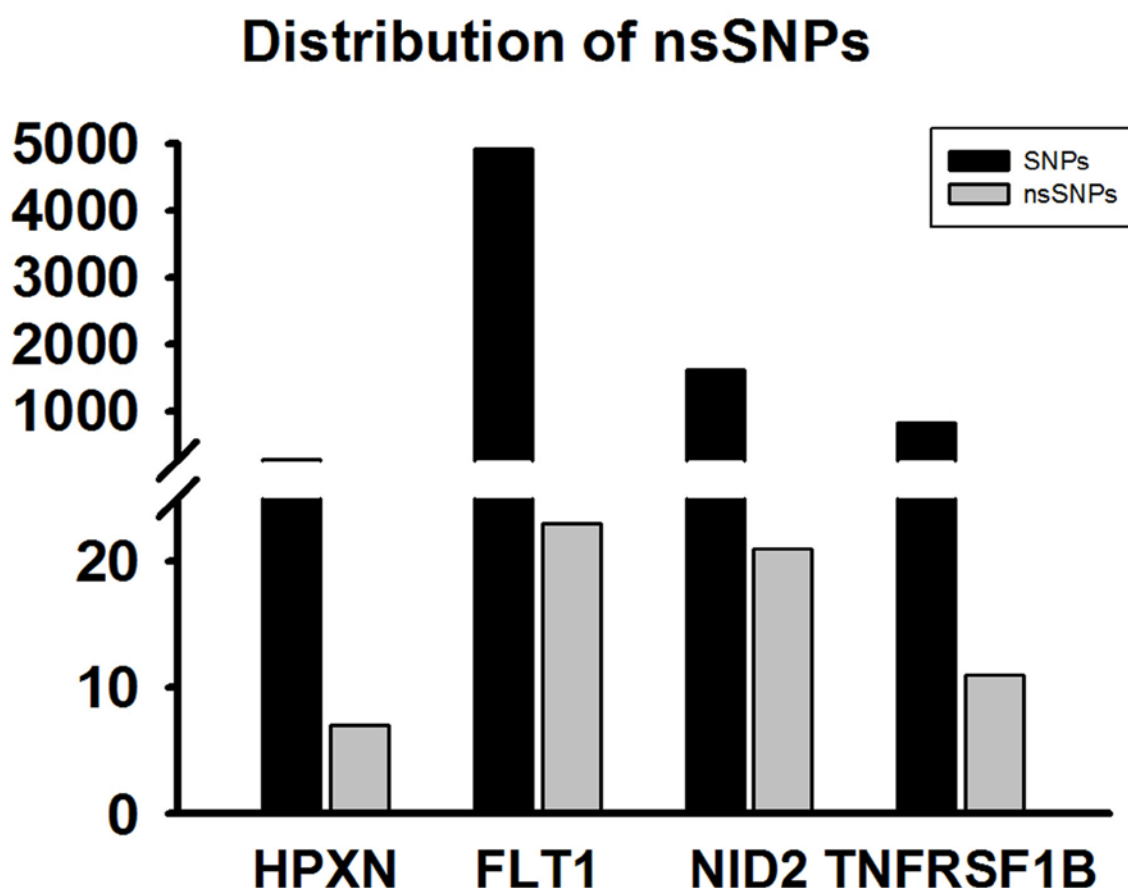


Figure. 4.5. Distribution of nsSNPs in hubs. Among the hubs identified in this chapter, SNPs were found for TNFRSF1B only. No SNP was detected for hubs SLC2A4, TRFR2 and UCHL1. HPXN, FLT1 and NID2 were identified as hub glycogenes in previous chapter 3.

4.5.1 Deleterious or damaged nsSNPs by PROVEAN

The PROVEAN protein tool was used for the prediction of damaged or deleterious nsSNPs in the hub genes. The program provides prediction for a protein sequence from any organism. The input to PROVEAN consists of detected coding nsSNPs and protein sequence for hubs, that is, 7 nsSNPs for HPXN (Uniprot ID: Q91X72), 23 for FLT1 (Uniprot ID: P35969), 21 for NID2 (Uniprot ID: O88322) and 11 for TNFRSF1B (Uniprot ID: P25119) in mouse. The input for each hub gene was submitted independently to PROVEAN with default parameters. No Deleterious SNP was identified for HPXN by PROVEAN (**Supplementary Table 4S3**). Among, 23 nsSNPs of FLT1 gene in mouse 3 were identified by PROVEAN to be Deleterious (**Table 4.4**).

For NID2, 3 out of 21 nsSNPs in mouse were identified to be Deleterious (**Table 4.5**) whereas, 1 out of 11 nsSNPs were identified to be Deleterious for TNFRSF1B (**Table 4.6**). All predictions were made by PROVEAN program by using -2.5 as a prediction cut-off.

4.5.2 Predicted nsSNPs that affect protein function by SIFT

The SIFT Sequence tool provides SIFT predictions for a given protein FASTA sequence. The input protein sequences and variants in required format for the same number of nsSNPs and hub genes as mentioned above were submitted with default parameters at http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html. The amino acids with probabilities < 0.05 are predicted to be deleterious. No SNP was identified for HPXN that would affect protein function by SIFT (**Supplementary Table 4S3**). Among 23 nsSNPs of FLT1 gene in mouse, 3 were identified to affect protein function (**Table 4.4**). For NID2, 8 out of 21 nsSNPs in mouse were identified to affect protein function (**Table 4.5**). For TNFRSF1B, 2 out of 11 nsSNPs in mouse were identified to affect protein function (**Table 4.6**) by SIFT program.

4.5.3 Damaged nsSNPs by PolyPhen-2

The protein sequences of each hub gene and their SNP substitution were submitted to standalone polyphen-2 installed in required format for mouse. No damaging SNP was identified for HPXN by PolyPhen-2 (**Supplementary Table 4S3**). 7 out of 23 nsSNPs for FLT1 gene were identified to be damaging (**Table 4.4**). Similarly, 7 out of 21 nsSNPs in mouse were identified to damaging (**Table 4.5**) for NID2 and 2 out of 11 nsSNPs in mouse were identified to be damaging for TNFRSF1B (**Table 4.6**) by Polyphen-2.

Table 4.4: nsSNPs of FLT1 (UniProt ID: P35969) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 3, 3 and 7 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 8 nsSNPs in mouse FLT1 gene out of total 23 were found to be deleterious, damaging or affect protein function and 15 nsSNPs were predicted to be tolerated by these three tools

(**Supplementary Table 4S4**). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA Change	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (Cutoff= < .05)	PPH2_Prob	PPH2
rs214988215	S80L	-2.053	Neutral	0.11	Tolerated	0.81	possibly damaging
rs233917867	F89L	-3.699	Deleterious	0.1	Tolerated	0.942	possibly damaging
rs213622875	T150M	-1.885	Neutral	0.02	Affect protein function	0.638	possibly damaging
rs233071760	A277V	-1.296	Neutral	0.08	Tolerated	0.994	probably damaging
rs32394398	D284E	-0.069	Neutral	1	Tolerated	0.821	possibly damaging
rs232914047	L453F	-3.129	Deleterious	0	Affect protein function	1	probably damaging
rs222631065	K575N	-0.425	Neutral	0.5	Tolerated	0.792	possibly damaging
rs32389134	Q1164K	-2.926	Deleterious	0.02	Affect protein function	0.232	benign

Table 4.5: nsSNPs of NID2 (UniProt ID: O88322) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 3, 8 and 7 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 11 nsSNPs in mouse NID2 gene out of total 21 were found to be deleterious, damaging or affect protein function and 10 nsSNPs were predicted to be tolerated by these three tools (**Supplementary Table 4S5**). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA Change	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (Cutoff= < .05)	PPH2_Prob	PPH2
rs262873422	P13L	-0.254	Neutral	0.01	Affect protein function	0.93	possibly damaging
rs30692168	S14P	-0.536	Neutral	0	Affect protein function	0.734	possibly damaging
rs30689780	V266A	-3.217	Deleterious	0	Affect protein function	0	benign
rs223532410	S290T	-0.531	Neutral	0	Affect protein function	0.807	possibly damaging
rs254798552	Y317D	-2.252	Neutral	0	Affect protein function	0.987	probably damaging
rs241645935	D351G	-1.699	Neutral	0	Affect protein function	0.003	benign
rs30689779	V357I	-0.137	Neutral	0	Affect protein function	0	benign
rs231633635	D406N	-0.805	Neutral	0	Affect protein function	0	benign
rs261439016	T525A	-2.642	Deleterious	0.68	Tolerated	0.983	probably damaging
rs30667034	E810K	-2.586	Deleterious	0.17	Tolerated	0.95	possibly damaging
rs255251970	A869T	0.255	Neutral	1	Tolerated	0.931	possibly damaging

Table 4.6: nsSNPs of TNFRSF1B (UniProt ID: P25119) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 1, 2 and 2 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 2 nsSNPs in mouse TNFRSF1B gene out of total 11 were found to be deleterious, damaging or affect protein function and 9 nsSNPs were predicted to be tolerated by these three tools (**Supplementary Table 4S6**). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA Change	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (Cutoff=< .05)	PPH2_Prob	PPH2
rs250384859	S396N	-0.651	Neutral	0	Affect protein function	0.858	possibly damaging
rs50901502	P431L	-2.631	Deleterious	0.01	Affect protein function	0.996	probably damaging

4.6 Protein Surface Accessibility and Secondary Structure prediction by NetSurfP

The protein sequences of each gene in normal and their predicted SNP substitutions were individually uploaded in Fasta format to NetSurfP web server for ASA and secondary structure predictions. Only those SNPs that were predicted by at least two software tools (4 SNPs of FLT1, 7 SNPs of NID2 and 2 SNPs of TNFRSF1B gene) to affect protein function were used for Accessible surface area (ASA) and secondary structure analysis. The predicted highest probabilities of secondary structure from NetSurfP were converted by perl script to one letter code to represent helix (H), beta sheet (E) and coil (C). Additionally, the data was further filtered by selecting only those residues that showed an ASA change of $\geq 10 \text{ \AA}^2$, change from buried to exposed state and vice versa, and change in their secondary structure (**Table 4.7-4.9**).

Table 4.7: Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of $\geq 10 \text{ \AA}^2$, and also show secondary structure change in FLT1 gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP = residues in normal and SNP sequence; ASA_N and ASA_SNP = ASA of residues in normal and SNP sequence; Class_N and Class_SNP = Class of residues in normal and SNP sequence [Class=buried (B) or exposed (E)]; SS_N and

SS_SNP=secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change as well as change from E to H or H to E are highlighted in bold.

SNP ID	Residue_N	Position	ASA_N (Å ²)	Class_N	SS_N	Residue_SNP	SNP	ASA_SNP (Å ²)	Class_SNP	SS_SNP	ASA-Change (Å ²)
rs233917867	G	532	20.092	B	E	G	F89L	33.629	E	C	-13.537
	T	771	30.819	B	E	T	F89L	45.507	E	C	-14.688
	L	772	42.278	B	E	L	F89L	60.405	E	C	-18.127
	K	832	57.452	B	E	K	F89L	95.445	E	H	-37.993
	V	859	31.585	B	E	V	F89L	60.589	E	C	-29.004
	V	861	34.921	B	E	V	F89L	67.705	E	C	-32.784
	V	1162	45.096	B	H	V	F89L	55.501	E	C	-10.405
rs213622875	L	772	42.278	B	E	L	T150M	58.134	E	C	-15.856
	K	832	57.452	B	E	K	T150M	95.239	E	H	-37.787
	S	846	40.317	E	C	S	T150M	29.898	B	E	10.419
	V	859	31.585	B	E	V	T150M	58.16	E	C	-26.575
	M	863	62.331	E	E	M	T150M	52.206	B	C	10.125
	L	876	47.057	B	H	L	T150M	72.489	E	C	-25.432
	K	881	104.331	E	H	K	T150M	48.052	B	E	56.279
rs232914047	I	882	63.085	E	C	I	T150M	22.478	B	E	40.607
	V	749	52.535	E	C	V	L453F	29.049	B	E	23.486
	L	828	32.775	B	E	L	L453F	59.947	E	C	-27.172
	K	832	57.452	B	E	K	L453F	71.069	E	C	-13.617
rs32389134	N	891	29.968	B	E	N	L453F	41.314	E	C	-11.346
	L	772	42.278	B	E	L	Q1164K	67.198	E	C	-24.92
	W	774	63.588	B	E	W	Q1164K	86.147	E	C	-22.559
	I	882	63.085	E	C	I	Q1164K	40.904	B	H	22.181
	K	914	69.3	E	C	K	Q1164K	53.585	B	E	15.715
	Y	921	66.76	E	C	Y	Q1164K	54.878	B	H	11.882
	L	922	53.502	E	C	L	Q1164K	40.813	B	H	12.689

Table 4.8: Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of ≥ 10 Å², and also secondary structure in NID2 gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP=residues in normal and SNP sequence; ASA_N and ASA_SNP=ASA of residues in normal and SNP sequence; Class_N and Class_SNP=Class of residues in normal and SNP sequence [Class=buried (B) or exposed (E)]; SS_N and SS_SNP=secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change are highlighted in bold.

SNP ID	Residue_ N	Positio n	ASA_N (Å ²)	Class _N	SS_ N	Residu e_SNP	SNP	ASA_SNP (Å ²)	Class_SNP	SS_SNP	ASA- Change (Å ²)
rs262873422	T	11	68.643	E	C	T	P13L	13.496	B	H	55.147
	P	12	50.644	E	C	P	P13L	4.583	B	H	46.061
	P	13	47.778	E	C	L	P13L	10.107	B	H	37.671
	R	27	56.907	E	C	R	P13L	43.144	B	H	13.763
rs30689780	M	1183	47.404	B	E	M	V266A	68.254	E	C	-20.85
rs254798552	E	308	51.624	B	E	E	Y317D	94.949	E	C	-43.325
	M	1183	47.404	B	E	M	Y317D	66.633	E	C	-19.229
	T	1193	35.59	E	E	T	Y317D	24.772	B	C	10.818
rs261439016	E	308	51.624	B	E	E	T525A	70.963	E	C	-19.339

Table 4.9: Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of ≥ 10 Å², and also secondary structure in TNFRSF1B gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP = residues in normal and SNP sequence; ASA_N and ASA_SNP = ASA of residues in normal and SNP sequence; Class_N and Class_SNP=Class of residues in normal and SNP sequence [Class=buried (B) or exposed (E)]; SS_N and SS_SNP=secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change are highlighted in bold.

SNP ID	Residue_ N	Positio n	ASA_N (Å ²)	Class _N	SS_ N	Residu e_SNP	SNP	ASA_SNP P (Å ²)	Class_SNP	SS_SNP	ASA- Change (Å ²)
rs50901502	Q	318	38.149	B	C	Q	P431L	60.224	E	H	-22.075
	R	368	63.273	B	C	R	P431L	93.844	E	E	-30.571
	T	379	49.585	E	C	T	P431L	26.575	B	E	23.01

All residues (G532, T771, L772, K832, V859, V861 and V1162) in FLT1 having F89L mutation show a change in ASA from a buried state to exposed state and a change in secondary structure from β -sheet (E) to coils (C) in majority of the cases. The only exception is K832 in which E in the native protein is changed to helix (H) conformation in the mutant protein and also shows the maximum change in ASA as a result of this mutation (**Table 4.4**). Similarly, in case of T150M mutation, K832 again shows change from buried to exposed state and change in E to H conformation. Residue K881 on the other hand shows opposite change from exposed to buried state and H to E conformation. However, in L453F and Q1164K, all of the residues that exhibit a

significant ASA change show a structural change from/to coiled conformation (**Table 4.4**).

It is interesting to note that all residues (**Table 4.8**) in NID2 exhibited a change in ASA from exposed to buried state and change from C to H conformation in the mutant protein P13L. The maximum ASA change of $55.14 \text{ \AA}^2$ is shown by T11 for this mutation. The mutations V266A and T525A both show ASA change from buried to exposed state ($\sim 20 \text{ \AA}^2$ change) with E to C conformational shift for residues M1183 and E308, respectively. Additionally, E308 also represents the residue with highest ASA change from buried to exposed state in mutant protein Y317D with change in E to C conformation (**Table 4.8**). No significant change was observed for NID2 protein with S14P, S290T and E810K mutations.

In case of TNFRSF1B, residues Q318 and R368 showed a ASA change from buried to exposed state in the mutant protein P431L (**Table 4.9**). Also, R368 represents the residue with maximum change in its ASA as a consequence of this mutation with change in its secondary structure from C to E conformation. No significant data was observed for S396N mutant protein.

4.7 Database and Tools development

Additionally, the SNP data that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2) and which may in turn affect protein function for hub genes (FLT1, TNFRSF1B and NID2) is available online in the form of a simple database available at www.procarb.org/glyco. **Figure 4.6** shows the home page of developed database. The web interface uses PHP and JavaScript to allow users to search the database by applying a variety of text-based options like simple search and advanced search. Simple search can be performed by using any of the three fields such as SNP ID, UniProt ID or Gene name. In advance search there are two fields, “Gene name” and “Residue”. The user can search a specific residue in any gene or all genes (currently the hub genes) available in this database. Both these search methods use a two step approach to display the SNP data. In first step, SNP ID, UniProt ID, Gene name, AA change, PPH2 Prob, PPH2, PROVEAN Score, PROVEAN Cutoff, SIFT Score and SIFT Cutoff will be displayed (**Figure 4.7**) and in second step by clicking AA change, secondary structure information of particular amino acid change

namely SNP ID, Residue N, Position, ASA N, Class N, SS N, Residue SNP, SNP, ASA SNP, Class SNP and SS SNP will be displayed (**Figure 4.8**). User can also filter search results by selecting any field such as Residue N, ASA change, Class change or SS change or any combination for a specific gene. Data entries are displayed using dynamically generated pages which describe the relevant information. All headers are defined in the help section and hyperlinked from search pages on this web portal. Moreover, each of SNP ID is also directly hyperlinked to their respective entry in dbSNP database.

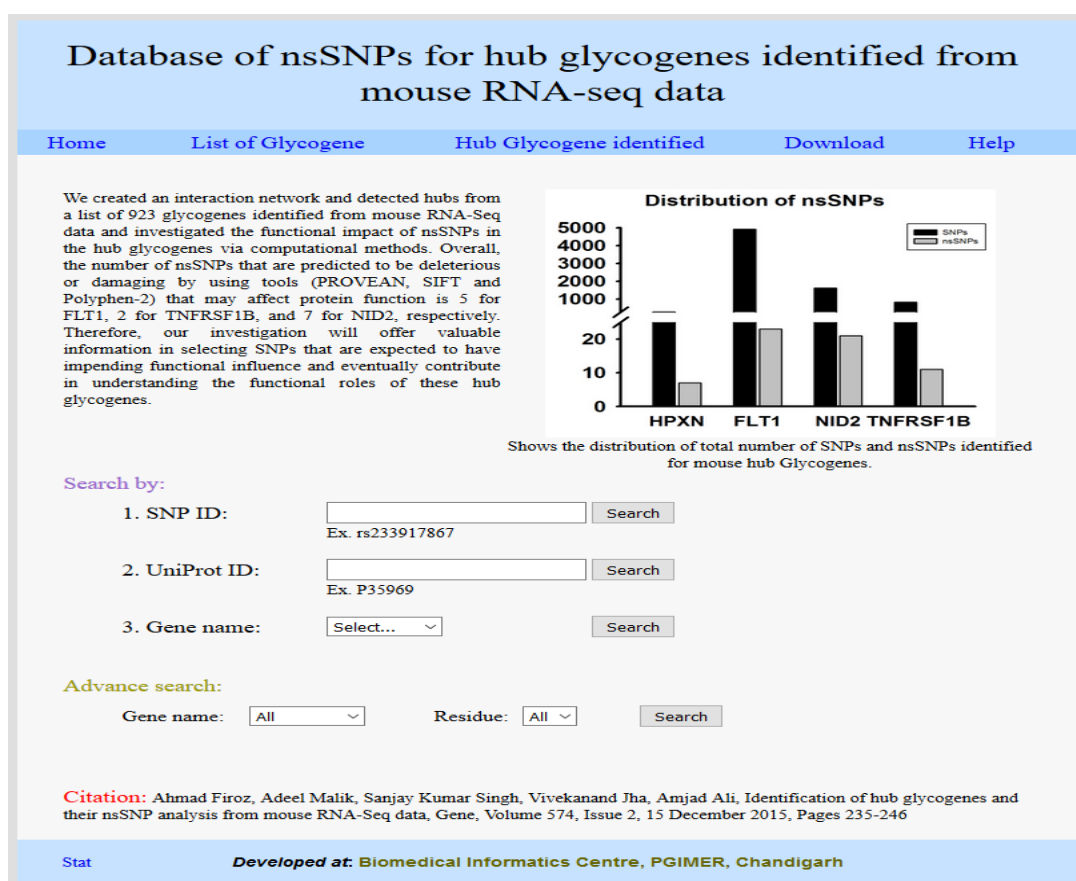


Figure.4.6. Home page of database

Database of nsSNPs for hub glycogenes identified from mouse RNA-seq data

Home

List of Glycogene

Hub Glycogene identified

Download

Help

Click on the AA changes to get the secondary structure & ASA value of normal & mutated by NetSurfP tools.

Download result

SNP ID	UniProt ID	Gene name	AA change	PPH2 Prob	PPH2	PROVEAN Score	PROVEAN Cutoff	SIFT Score	SIFT Cutoff
rs214988215	P35969	FLT1	S80L	0.81	Possibly damaging	-2.053	Neutral	0.11	Tolerated
rs33730929	P35969	FLT1	D85N	0	Benign	0.224	Neutral	0.75	Tolerated
rs233917867	P35969	FLT1	F89L	0.942	Possibly damaging	-3.699	Deleterious	0.1	Tolerated
rs32388970	P35969	FLT1	T97M	0	Benign	-0.714	Neutral	0.08	Tolerated
rs252361901	P35969	FLT1	T113K	0	Benign	-0.904	Neutral	0.08	Tolerated
rs213622875	P35969	FLT1	T150M	0.638	Possibly damaging	-1.885	Neutral	0.02	Affect protein function
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
rs259750200	P35969	FLT1	S945N	0	Benign	-0.214	Neutral	0.76	Tolerated
rs242252147	P35969	FLT1	S970G	0	Benign	2.238	Neutral	0.84	Tolerated
rs260647574	P35969	FLT1	S987N	0.052	Benign	-0.393	Neutral	0.32	Tolerated
rs32389134	P35969	FLT1	Q1164K	0.232	Benign	-2.926	Deleterious	0.02	Affect protein function
rs224361070	P35969	FLT1	P1201S	0.124	Benign	-1.755	Neutral	0.8	Tolerated
rs239789789	P35969	FLT1	H1202Q	0	Benign	0.603	Neutral	NA	Tolerated
rs261438024	P35060	FLT1	H1202R	0	Benign	0.768	Neutral	0.69	Tolerated
rs213904262	P35969	FLT1	V1303M	0.3	Benign	-0.167	Neutral	0.23	Tolerated

Back

Stat

Developed at: Biomedical Informatics Centre, PGIMER, Chandigarh

Figure.4.7. shows prediction of deleterious or damaging from three tools

Database of nsSNPs for hub glycogenes identified from mouse RNA-seq data

HomeList of GlycogeneHub Glycogene identifiedDownloadHelp

Download result

Changes from Normal to SNP

Gene name: FLT1Residue N: AllASA: Class: SS: Search

SNP ID	Residue N	Position	ASA N	Class N	SS N	Residue SNP	SNP	ASA SNP	Class SNP	SS SNP
rs233917867	V	2	23.132	B	E	V	F89L	23.116	B	E
rs233917867	S	3	34.668	E	E	S	F89L	35.699	E	E
rs233917867	C	4	5.658	B	E	C	F89L	5.574	B	E
rs233917867	W	5	87.013	E	E	W	F89L	83.189	E	E
rs233917867	D	6	34.166	B	E	D	F89L	44.152	B	E
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
rs233917867	V	1325	62.804	E	C	V	F89L	62.863	E	C
rs233917867	V	1326	47.632	B	C	V	F89L	45.034	B	C
rs233917867	L	1327	81.681	E	C	L	F89L	80.271	E	C
rs233917867	Y	1328	72.893	E	C	Y	F89L	70.457	E	C
rs233917867	S	1329	46.657	E	C	S	F89L	48.275	E	C
rs233917867	S	1330	61.718	E	C	S	F89L	61.237	E	C
rs233917867	P	1331	51.907	E	C	P	F89L	54.277	E	C
rs233917867	P	1332	74.906	E	C	P	F89L	74.866	E	C
rs233917867	A	1333	86.022	E	C	A	F89L	86.507	E	C

Back

Stat

Developed at: Biomedical Informatics Centre, PGIMER, Chandigarh

Figure.4.8. Shows SS and ASA of predicted SNP

4.8 Discussion

By the progression of bioinformatics during the last ten years, discovery of molecular targets and targeted therapeutics have become essential remedial treatment for various diseases (Dinh *et al.*, 2007). In order to explore the functional roles of all the mouse glycogenes and additionally related genes involved in various biological processes, an interaction network was created and node degree for each gene in the network was calculated. *SLC2A4*, *TNFRSF1B*, *TRFR2* and *UCHL1* were the glycogenes that were considered as hubs (genes with highest node degree) in the network constructed from 923 mouse glycogenes and additional related genes predicted by GeneMANIA. The most significant enriched GO terms within these glycogenes identified by DAVID cluster analysis and GeneMANIA include glycoprotein metabolic process, inflammatory response, extracellular matrix and adhesion related processes (**Table 4.1** and **4.3**). We also created an interaction network of differentially expressed glycogenes (DEGGs) that are involved in similar processes in chapter 3 by using mouse RNA-Seq data (Mortazavi *et al.*, 2008) and identified HPXN, FLT1 and NID2 as potential hubs in chapter 3.

One of the first potential hubs detected in this study is solute carrier family 2, facilitated glucose transporter member 4 (*SLC2A4* or GLUT-4). In muscle and fat cells, GLUT-4 is stimulated by insulin from an intracellular site to the cell exterior, where it aids in decreasing the plasma glucose levels (Bryant *et al.*, 2002). The expression of tumor necrosis factor receptor superfamily member 1B (*TNFRSF1B*) also known as tumor necrosis factor receptor 2 (*TNFR-2*) is greatly controlled and largely limited to the cells of immune system (Grivennikov *et al.*, 2006). *TNFR-2* is released in huge amounts throughout inflammation and acts as functional inhibitor of tumor necrosis factor (*TNF*) (Martin *et al.*, 2014). Transferrin receptor protein 2 is a recognized iron sensor, encoded by *Tfr2* or *TRFR2*, which is expressed in the liver and mutations that affect the protein function lead to iron-overload disease Type III hereditary hemochromatosis (HH) (Worthen *et al.*, 2014). Loss of functional *TRFR2* decreases the production of hepcidin (peptide hormone), which slows down accumulation of iron in various organs and in turn leads to various diseases such as cirrhosis, heart disease, arthritis, and diabetes (Worthen *et al.*, 2014). The final gene (*UCHL1*) that was also detected as a hub codes for one of the first discovered deubiquitinating enzyme Ubiquitin carboxyl-terminal hydrolase isozyme L1 (*UCH-L1*) (Hussain *et al.*, 2013). The protein is expressed only

in neuroendocrine tissues and in certain cancers with poorly understood functional roles (Hussain *et al.*, 2013). Loss of *UCHL1* in mice leads to gracile axonal dystrophy, a lethal neurodegenerative condition (Saigoh *et al.*, 1999), and mutations have been implicated in the pathogenesis of Parkinson's disease (Maraganore *et al.*, 2004; Ragland *et al.*, 2009). Similarly, the hubs (HPXN, FLT1 and NID2) identified in our chapter 3 may possibly play essential roles in regulating various important biological processes as described above.

The functional annotation cluster analysis on non-redundant list of 923 glycogenes was performed by using DAVID. The over-representation of GO terms such as metabolic processes of glycoprotein, glycan and cholesterol; immune response; blood coagulation; organic acid and amino acid transport to name a few suggests the diverse roles that these glycogenes carry out (Janot *et al.*, 2009; Ngamukote *et al.*, 2007; Saravanan *et al.*, 2010). The diverse roles of glycogenes are also supported by KEGG pathway and GeneMANIA enrichment analysis (**Table 4.2 and 4.3**).

Single amino acid mutations within the proteins can considerably alter the stability of a protein structure and in turn may affect its function. Therefore, to understand the functional effects of variations we also investigated the role of nsSNPs that may lead to an amino acid change in the hub proteins. Identifying nsSNPs that disturb protein function and relate to a disease is considered as a defying task in the future (Karchin *et al.*, 2005a; Karchin *et al.*, 2005b). A large number of nsSNPs were identified in this study (**Supplementary Tables 4S3-4S6**) for all the hubs, however, only few had an effect on protein function (**Table 4.4-4.6**).

This could possibly be due to the fact that the effect of many nsSNPs might be neutral because natural selection may have eliminated mutations at critical sites (George Priya Doss *et al.*, 2008).

Structural significance of native and mutant sequence of the hubs was further investigated in this work by sequence based predictions of solvent accessibility and secondary structure analysis. For mutants F89L and T150 in FLT1, K832 (SNP ID: rs233917867 and rs213622875) and K881 (SNP ID: rs213622875) are the only residues that exhibit significant ASA change as well as a secondary structure change from β -sheet to helix or vice versa (**Table 4.7-4.9**). Because of their charged nature, lysine residues are most expected to be present at solvent accessible areas of proteins, thereby making them interesting candidates for further studies on their roles on molecular interactions (Scholten *et al.*, 2006). ASA, thought of as a discriminating characteristic

innsSNPs linked with a disease, have a tendency to occur at buried sites where as benign substitutions are inclined towards solvent accessible sites (Sunyaev *et al.* 2000; Ferrer-Costa *et al.*, 2002). Amino acid residues that form the hydrophobic core of a protein are essential for its stability (George Priya Doss *et al.*, 2008). Polar and charged side chains have greater ASA than nonpolar side chains proposing that establishment of a hydrophobic core is a solid driving force in protein folding (Chan *et al.*, 1990).

The results from our current study proposes that the application of computational methods may offer alternate methods to select target SNPs by understanding the consequence of SNPs on the functional characteristics or molecular phenotype of a protein. Our findings might be useful for additional genotype–phenotype investigations as well as pharmacogenetics studies. Finally, the functional analysis in this study might serve as a worthy model for further exploration in genetically inherited diseases.

References

Bryant, N.J; Govers, R; James, D.E. Regulated transport of the glucose transporter GLUT4. *Nat. Rev. Mol. Cell Biol.* **3**. 267-277, (2002).

Chan, H.S.; Dill, K.A.; Origins of structure in globular proteins. *Proc. Natl. Acad. Sci.* **87**, 6388–6392, (1990).

Dinh, P., Sotiriou, C., Piccart, M.J., 2007 Decc. The evolution of treatment strategies: aiming at the target. *Breast 16* (Suppl. 2), S10–S16.

Ferrer-Costa, C.; Orozco, M.; De la Cruz, X. Characterization of disease-associated single acid polymorphisms in terms of sequence and structure properties. *J. Mol.Biol.* **315**, 771–786, (2002).

George Priya Doss, C.; Rajasekaran, R.; Sudandiradoss, C.; et al., A novel computational and structural analysis of nsSNPs in CFTR gene. *Genom. Med.* **2**. 23–32, (2008).

Grivennikov, S.I.; Kuprash, D.V.; Liu, Z.G.; et al., Intracellular signals and events activated by cytokines of the tumor necrosis factor superfamily: from simple paradigms to complex mechanisms. *Int. Rev. Cytol.* **252**. 129–161, (2006).

Hussain, S.; Feldman, A.L.; Das, C.; et al., Ubiquitin hydrolase UCH-L1 destabilizes mTOR complex 1 by antagonizing DDB1-CUL4- mediated ubiquitination of raptor. *Mol. Cell. Biol.* **33**. 1188–1197, (2013).

Janot, M.; Audfray, A.; Lorient, C.; et al., Glycogenome expression dynamics during mouse C2C12 myoblast differentiation suggests a sequential reorganization of membrane glycoconjugates. *BMC Genomics.* **10**. 483, (2009).

Karchin, R.; Diekhans, M.; Kelly, L.; et al., LS-SNP: large-scale annotation of coding nonsynonymous SNPs based on multiple information sources. *Bioinformatics*. **21**. 2814–2820, (2005b).

Karchin, R.; Kelly, L.; Sali, A. Improving functional annotation of non-synonymous SNPs with information theory. *Pac. Symp. Biocomput.* **10**. 397–408, (2005a).

Magrane, M.; Consortium, U. UniProt Knowledgebase: A Hub of Integrated Protein Data. *Database (Oxford)* p. **bar009**. (2011).

Maraganore, D.M.; Lesnick, T.G.; Elbaz, A.; et al., UCHL1 is a Parkinson's disease susceptibility gene. *Ann. Neurol.* **55**, 512–521, (2004).

Martin, E.M.; Remke, A.; Pfeifer, E.; et al., TNFR2 maintains adequate IL-12 production by dendritic cells in inflammatory responses by regulating endogenous TNF levels. *Innate Immunol.* **20**. 712–720, (2014).

Montejo, J; Zuberi, K.; Rodriguez, H; et al., GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*. **26**. 2927–2928, (2010).

Mortazavi, A.; Williams, B.A.; McCue, K.; et al., Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, **5**. 621–8 (2008).

Ngamukote, S.; Yanagisawa, M.; Ariga, T.; et al., Developmental changes of glycosphingolipids and expression of glycogenes in mouse brains. *J. Neurochem.* **103**. 2327–2341, (2007).

Saigoh, K.; Wang, Y.L.; Suh, J.G.; et al., Intragenic deletion in the gene encoding ubiquitin carboxy-terminal hydrolase in gad mice. *Nat. Genet.* **23**, 47–51, (1999).

Saravanan, C.; Cao, Z.; Head, S.R.; et al., Analysis of differential expression of glycosyltransferases in healing corneas by glycogene microarrays. *Glycobiology*. **20**. 13–23, (2010).

Scholten, A.; Visser, N.F.; Van Den Heuvel, R.H.; et al., Analysis of proteinprotein interaction surfaces using a combination of efficient lysine acetylation and nano LC–MALDI–MS/MS applied to the E9:Im9 bacteriotoxin - immunity protein complex. *J. Am. Soc. Mass Spectrom.* **17**. 983–994, (2006).

Smoot, ME.; Ono, K.; Ruscheinski, J. et al., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. **27**. 431–432, (2011).

Sunyaev, S.; Ramensky, V.; Bork, P. Towards a structural basis of human nonsynonymous single nucleotide polymorphisms. *Trends Genet.* **16**. 198–200, (2000).

Warde-Farley, D; Donaldson, SL; Comes, O.; et al., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**. W214–W220 (2010).

Worthen, C.A.; Enns, C.A.; The role of hepatic transferrin receptor 2 in the regulation of iron homeostasis in the body. *Front. Pharmacol.* **5**, 34, (2014).

Conclusions and Futuristic Aspect

	Contents	Page number
5.1	Introduction	89
5.2	Conclusions from the present work	89
5.3	Futuristic aspects	92
	List of publications	93

Abstract: The results given in chapter 3 and 4 have been concluded, compared and correlated in this chapter. Further studies possible have also been proposed at the end of chapter.

5.1 Introduction

In the present thesis, we have used a systems biology approach to identify the DEGs, the functional enrichment, and identified potential hub genes by constructing glycometabolites interaction networks. We also identified functional modules in these networks and indicate the significance of immune system related processes in liver, and glycosphingolipid metabolic processes in the development of brain. Similarly, we observe that the processes such as cell proliferation, adhesion, and growth are important for muscle development. The findings deduced from the current work are in consistence with the previous studies. Experimental validation will be required to confirm the predictions made in this study that will establish the role of predicted hubs as well as enriched functional processes in these tissues. We also investigated the functional impact of nsSNPs in the identified hub glycometabolites via computational methods. Therefore, our investigation will offer valuable information in selecting SNPs that are expected to have impending functional influence and eventually contribute in understanding the functional roles of these hub glycometabolites. Additionally, we also developed databases for the SNP data that are predicted to be deleterious or damaging

5.2 Conclusions from the present work

- i) We downloaded the precomputed RPKM values for mouse transcriptome that was mapped and quantified previously by RNA-Seq analysis. The role of differentially expressed glycometabolites (DEGs) in three major tissues such as brain, muscle, and liver has been explored using these mouse RNA-seq data, and we identified 579, 501, and 442 DEGs for brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups
- ii) DAVID functional analysis suggested inflammatory response, glycosaminoglycan metabolic process, and protein maturation as the enriched biological processes in BvL579, BvM501, and LvM442, respectively.

iii) A total of 210 pathways were predicted for BvL579, whereas 193 and 198 pathways were predicted for BvM501 and LvM442, respectively. The top 10 KEGG pathways for each of the three categories are shown in **Table 3.3** of chapter 3. From these tables, it can be observed that the majority of the DEGs were found to be associated with important biological processes, many being classified in signaling (e.g., PI3K-Akt and Rap1 signaling) pathways, cytokine-cytokine receptor interaction, or ECM-receptor interaction or being involved in adhesion related functions.

iv) These DEGs were then used to construct three interaction networks by using GeneMANIA, from which we detected potential hub genes such as PEMT and HPXN (BvL579), IGF2 and NID2 (BvM501), and STAT6 and FLT1 (LvM442), having the highest degree.

v) Community analysis results also suggest that the significance of immune system related processes in liver, glycosphingolipid metabolic processes in the development of brain, and the processes such as cell proliferation, adhesion, and growth are important for muscle development

vi) The functional annotation cluster analysis on non-redundant list of 923 glycogenes identified from mouse RNA-Seq data was performed by using DAVID. The over representation of GO terms such as metabolic processes of glycoprotein, glycan and cholesterol; immune response; blood coagulation; organic acid and amino acid transport suggests the diverse roles that these glycogenes. The diverse roles of glycogenes are also supported by KEGG pathway and GeneMANIA enrichment analysis (**Tables 4.2 and 4.3**) of chapter 4.

vii) 222 pathways are predicted by KAAS web server in which these glycogenes may play a functional role. Top 20 KEGG pathways are summarized in (**Table 4.2**) in chapter 4. From these tables, it can be seen that glycogenes are involved in diverse types of biological pathways such as signalling pathways (PI3K–Akt and Rap1 signaling pathway), complement and coagulation cascades, cytokine receptor interaction, pathways in cancer or adhesion related functions (cell adhesion molecules CAMs and focal adhesion)

viii) Network analysis by GeneMANIA for interactions among the glycogene products suggested enrichment of diverse GO terms that may be involved in lysosome, locomotion, extracellular matrix, glycoprotein metabolic process and adhesion related processes (**Table 4.3**) of chapter 4.

ix) Created an interaction network of 923 glycogenes and detected hubs SLC2A4, TNFRSF1B, TRFR2 and UCHL1. The most significant enriched GO terms within these glycogenes identified by DAVID functional cluster analysis and GeneMANIA include glycoprotein metabolic process, inflammatory response, extracellular matrix and adhesion related processes.

x) Investigated the functional impact of nsSNPs in all identified hub glycogenes in chapter 3 & 4 via computational methods. Overall, the number of nsSNPs that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2) that may affect protein function is 4 for FLT1, 2 for TNFRSF1B, and 7 for NID2, respectively.

xi) Residues in the native and mutant proteins were analyzed for solvent accessibility and secondary structure change. Analysis of hubs can help in determining their degree of conservation and understanding their functions in biological processes. For mutants F89L and T150M in FLT1, K832 (SNP ID: rs233917867 and rs213622875) and K881 (SNP ID: rs213622875) are the only residues that exhibit significant ASA change as well as a secondary structure change from β -sheet to helix or vice versa.

xii) Developed databases for the SNP data that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2) and which may in turn affect protein function for hub genes (FLT1, TNFRSF1B and NID2), which is available online at **www.procarb.org/glyco** in the form of a simple database .

xiii) Simple and advance search tools has been provided in the database, where, user can search a specific residue in any gene or all genes (currently the hub genes) available

in this database, and can also filter search results by selecting any field such as Residue N, ASA change, Class change or SS change or any combination for a specific gene.

xiv) It will be worth to obtain additional information regarding the selected SNPs before finalizing the decision. For example, SNPs in protein coding or regulatory regions are more prone to have deleterious effects, and, as such, the potential of triggering a disease is enhanced. Therefore, by using these computational methods and by acquiring additional information about the functionally significant SNPs that are expected to be linked with disease, a significant amount of genotyping and analysis overhead can be reduced. The results from our current study proposes that the application of these computational methods may offer alternate methods to select target SNPs by understanding the consequence of SNPs on the functional characteristics or molecular phenotype of a protein.

5.3 Futuristic aspects

- i) The nsSNPs proposed in this work may be further targeted through experimental methods for understanding structural and functional relationships of hub mutants.
- ii) Our findings might be useful for additional genotype, phenotype investigations as well as pharmacogenetic studies.
- iii) Finally, the functional analysis in this study might serve as a worthy model for further exploration in genetically inherited diseases.

List of publications

1. **Firoz A** , Malik A , Singh SK, Jha V, and **Ali A**. Comparative analysis of glycogene expression in different mouse tissues using RNA-Seq data. **International Journal of Genomics**, vol. 2014, Article ID 837365, 18 pages, 2014. doi:10.1155/2014/837365
2. **Firoz A** , Malik A , Jha V, and **Ali A**. Identification of hub glycogenes and their nsSNP analysis from mouse RNA-Seq data. **GENE**, Vol. 574, Issue 2, 15 December Pages 235–246, 2015

Publications in Conferences

1. **Firoz A**, Singh S, Jha V, **Ali A**, Malik A: NucleoProtDb: a database of experimentally determined three-dimensional structures of protein-adenine/guanine phosphate complexes National Conference on Innovative molecules for Sustainable Future (NCIMSF-2013) October 24th -26th, 2013, Thapar University, Patiala, Punjab, India.
2. **Firoz A**, Malik A , Singh SK, Jha V, and **Ali A**. Computational analysis and identification of hub glycogenes from mouse RNA-Seq data. Accelerating Biology 2016 January 19-21, 2016, Bioinformatics group, Centre for Development of advanced Computing (C-DAC), Pune University campus, Pune , India.

Manuscript in preparation

Firoz A, Malik A , Jha V, and **Ali A**. Database of nsSNPs of glycogenes in for mouse.

Appendix to thesis

Supplementary Table 3S1: List of all KEGG pathways reported for different tissue groups. Each sheet contains the list of KEGG pathways and the number of mapped genes in brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups.

Brain Vs Liver (BvL579)

Pathway Name	Number of mapped Genes
04151 PI3K-Akt signaling pathway	34
04610 Complement and coagulation cascades	34
04142 Lysosome	31
04060 Cytokine-cytokine receptor interaction	23
05200 Pathways in cancer	21
04015 Rap1 signaling pathway	18
04510 Focal adhesion	18
05205 Proteoglycans in cancer	18
04514 Cell adhesion molecules CAMs	16
04976 Bile secretion	16
04014 Ras signaling pathway	14
04512 ECM-receptor interaction	14
04640 Hematopoietic cell lineage	13
05150 Staphylococcus aureus infection	13
04141 Protein processing in endoplasmic reticulum	12
04080 Neuroactive ligand-receptor interaction	12
04810 Regulation of actin cytoskeleton	12
05166 HTLV-I infection	12
02010 ABC transporters	11
04145 Phagosome	11
04360 Axon guidance	11
00564 Glycerophospholipid metabolism	9
00510 N-Glycan biosynthesis	9
04020 Calcium signaling pathway	9
04918 Thyroid hormone synthesis	9
05206 MicroRNAs in cancer	9
04010 MAPK signaling pathway	8
04630 Jak-STAT signaling pathway	8
04144 Endocytosis	8
04915 Estrogen signaling pathway	8
05152 Tuberculosis	8
04668 TNF signaling pathway	7
04612 Antigen processing and presentation	7
04380 Osteoclast differentiation	7

05215 Prostate cancer	7
05222 Small cell lung cancer	7
05322 Systemic lupus erythematosus	7
05323 Rheumatoid arthritis	7
05414 Dilated cardiomyopathy DCM	7
00500 Starch and sucrose metabolism	6
00600 Sphingolipid metabolism	6
00230 Purine metabolism	6
00513 Various types of N-glycan biosynthesis	6
00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin	6
04350 TGF-beta signaling pathway	6
04066 HIF-1 signaling pathway	6
04974 Protein digestion and absorption	6
05203 Viral carcinogenesis	6
05164 Influenza A	6
05168 Herpes simplex infection	6
05146 Amoebiasis	6
05145 Toxoplasmosis	6
00561 Glycerolipid metabolism	5
00565 Ether lipid metabolism	5
00604 Glycosphingolipid biosynthesis - ganglio series	5
04012 ErbB signaling pathway	5
04520 Adherens junction	5
04530 Tight junction	5
04540 Gap junction	5
04911 Insulin secretion	5
04913 Ovarian Steroidogenesis	5
04916 Melanogenesis	5
04972 Pancreatic secretion	5
05202 Transcriptional misregulation in cancers	5
05214 Glioma	5
05410 Hypertrophic cardiomyopathy HCM	5
05412 Arrhythmogenic right ventricular cardiomyopathy ARVC	5
05133 Pertussis	5
05161 Hepatitis B	5
05169 Epstein-Barr virus infection	5
00052 Galactose metabolism	4
00533 Glycosaminoglycan biosynthesis - keratan sulfate	4
00531 Glycosaminoglycan degradation	4
00603 Glycosphingolipid biosynthesis - globo series	4
00511 Other glycan degradation	4
00983 Drug metabolism - other enzymes	4
03013 RNA transport	4

04310 Wnt signaling pathway	4
04340 Hedgehog signaling pathway	4
04390 Hippo signaling pathway	4
04670 Leukocyte transendothelial migration	4
04062 Chemokine signaling pathway	4
04910 Insulin signaling pathway	4
03320 PPAR signaling pathway	4
04970 Salivary secretion	4
04971 Gastric acid secretion	4
04977 Vitamin digestion and absorption	4
04978 Mineral absorption	4
04724 Glutamatergic synapse	4
04727 GABAergic synapse	4
04725 Cholinergic synapse	4
05217 Basal cell carcinoma	4
05218 Melanoma	4
05010 Alzheimer's disease	4
05020 Prion diseases	4
05032 Morphine addiction	4
05416 Viral myocarditis	4
04932 Non-alcoholic fatty liver disease NAFLD	4
05120 Epithelial cell signaling in Helicobacter pylori infection	4
05160 Hepatitis C	4
05144 Malaria	4
00140 Steroid hormone biosynthesis	3
00512 Mucin type O-glycan biosynthesis	3
00514 Other types of O-glycan biosynthesis	3
00563 Glycosylphosphatidylinositol GPI-anchor biosynthesis	3
04210 Apoptosis	3
04620 Toll-like receptor signaling pathway	3
04650 Natural killer cell mediated cytotoxicity	3
04666 Fc gamma R-mediated phagocytosis	3
04920 Adipocytokine signaling pathway	3
04912 GnRH signaling pathway	3
04914 Progesterone-mediated oocyte maturation	3
04270 Vascular smooth muscle contraction	3
04973 Carbohydrate digestion and absorption	3
04975 Fat digestion and absorption	3
04962 Vasopressin-regulated water reabsorption	3
04960 Aldosterone-regulated sodium reabsorption	3
04961 Endocrine and other factor-regulated calcium reabsorption	3
04728 Dopaminergic synapse	3
04722 Neurotrophin signaling pathway	3

05212 Pancreatic cancer	3
05219 Bladder cancer	3
05213 Endometrial cancer	3
05223 Non-small cell lung cancer	3
05034 Alcoholism	3
05110 Vibrio cholerae infection	3
05162 Measles	3
05140 Leishmaniasis	3
00010 Glycolysis / Gluconeogenesis	2
00040 Pentose and glucuronate interconversions	2
00520 Amino sugar and nucleotide sugar metabolism	2
00240 Pyrimidine metabolism	2
00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	2
00601 Glycosphingolipid biosynthesis - lacto and neolacto series	2
00740 Riboflavin metabolism	2
00770 Pantothenate and CoA biosynthesis	2
00830 Retinol metabolism	2
00860 Porphyrin and chlorophyll metabolism	2
00980 Metabolism of xenobiotics by cytochrome P450	2
04330 Notch signaling pathway	2
04064 NF-kappa B signaling pathway	2
04146 Peroxisome	2
04114 Oocyte meiosis	2
04672 Intestinal immune network for IgA production	2
04260 Cardiac muscle contraction	2
04964 Proximal tubule bicarbonate reclamation	2
04723 Retrograde endocannabinoid signaling	2
04742 Taste transduction	2
04320 Dorso-ventral axis formation	2
04713 Circadian entrainment	2
05204 Chemical carcinogenesis	2
05210 Colorectal cancer	2
05221 Acute myeloid leukemia	2
05220 Chronic myeloid leukemia	2
05211 Renal cell carcinoma	2
05320 Autoimmune thyroid disease	2
05330 Allograft rejection	2
05332 Graft-versus-host disease	2
05014 Amyotrophic lateral sclerosis ALS	2
05030 Cocaine addiction	2
05031 Amphetamine addiction	2
04940 Type I diabetes mellitus	2

05131 Shigellosis	2
05100 Bacterial invasion of epithelial cells	2
01200 Carbon metabolism	1
01212 Fatty acid metabolism	1
01230 Biosynthesis of amino acids	1
00030 Pentose phosphate pathway	1
00051 Fructose and mannose metabolism	1
00053 Ascorbate and aldarate metabolism	1
00190 Oxidative phosphorylation	1
00680 Methane metabolism	1
00062 Fatty acid elongation	1
00100 Steroid biosynthesis	1
00590 Arachidonic acid metabolism	1
00330 Arginine and proline metabolism	1
00440 Phosphonate and phosphinate metabolism	1
00480 Glutathione metabolism	1
00760 Nicotinate and nicotinamide metabolism	1
00780 Biotin metabolism	1
00944 Flavone and flavonol biosynthesis	1
00627 Aminobenzoate degradation	1
00363 Bisphenol degradation	1
00982 Drug metabolism - cytochrome P450	1
03050 Proteasome	1
04013 MAPK signaling pathway - fly	1
04391 Hippo signaling pathway -fly	1
04370 VEGF signaling pathway	1
04150 mTOR signaling pathway	1
04110 Cell cycle	1
04115 p53 signaling pathway	1
04621 NOD-like receptor signaling pathway	1
04660 T cell receptor signaling pathway	1
04662 B cell receptor signaling pathway	1
04664 Fc epsilon RI signaling pathway	1
04917 Prolactin signaling pathway	1
04614 Renin-angiotensin system	1
04966 Collecting duct acid secretion	1
04726 Serotonergic synapse	1
04626 Plant-pathogen interaction	1
05216 Thyroid cancer	1
05310 Asthma	1
05321 Inflammatory bowel disease IBD	1
05016 Huntington's disease	1
04930 Type II diabetes mellitus	1

05130 Pathogenic Escherichia coli infection	1
05132 Salmonella infection	1
05134 Legionellosis	1
05142 Chagas disease American trypanosomiasis	1
05143 African trypanosomiasis	1

Brain Versus Muscle (BvM501)

Pathway Name	Number of mapped Genes
04142 Lysosome	29
04151 PI3K-Akt signaling pathway	24
04060 Cytokine-cytokine receptor interaction	19
04514 Cell adhesion molecules CAMs	19
05200 Pathways in cancer	17
04141 Protein processing in endoplasmic reticulum	15
04510 Focal adhesion	15
05166 HTLV-I infection	13
04512 ECM-receptor interaction	12
04610 Complement and coagulation cascades	12
04360 Axon guidance	12
05205 Proteoglycans in cancer	12
04015 Rap1 signaling pathway	11
04080 Neuroactive ligand-receptor interaction	11
00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin	10
05414 Dilated cardiomyopathy DCM	10
00564 Glycerophospholipid metabolism	9
04915 Estrogen signaling pathway	9
04976 Bile secretion	9
04380 Osteoclast differentiation	9
05322 Systemic lupus erythematosus	9
05150 Staphylococcus aureus infection	9
00510 N-Glycan biosynthesis	8
04630 Jak-STAT signaling pathway	8
04020 Calcium signaling pathway	8
04145 Phagosome	8
04918 Thyroid hormone synthesis	8
04916 Melanogenesis	8
05168 Herpes simplex infection	8
00511 Other glycan degradation	7
02010 ABC transporters	7
04810 Regulation of actin cytoskeleton	7
04640 Hematopoietic cell lineage	7
04670 Leukocyte transendothelial migration	7

04974 Protein digestion and absorption	7
05203 Viral carcinogenesis	7
05222 Small cell lung cancer	7
05410 Hypertrophic cardiomyopathy HCM	7
05412 Arrhythmogenic right ventricular cardiomyopathy ARVC	7
05152 Tuberculosis	7
00561 Glycerolipid metabolism	6
00230 Purine metabolism	6
04014 Ras signaling pathway	6
04350 TGF-beta signaling pathway	6
04668 TNF signaling pathway	6
04911 Insulin secretion	6
04724 Glutamatergic synapse	6
04727 GABAergic synapse	6
04725 Cholinergic synapse	6
05202 Transcriptional misregulation in cancers	6
05323 Rheumatoid arthritis	6
05145 Toxoplasmosis	6
00565 Ether lipid metabolism	5
00600 Sphingolipid metabolism	5
00513 Various types of N-glycan biosynthesis	5
00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	5
00601 Glycosphingolipid biosynthesis - lacto and neolacto series	5
00603 Glycosphingolipid biosynthesis - globo series	5
00604 Glycosphingolipid biosynthesis - ganglio series	5
04390 Hippo signaling pathway	5
04144 Endocytosis	5
04540 Gap junction	5
04612 Antigen processing and presentation	5
04913 Ovarian Steroidogenesis	5
04972 Pancreatic secretion	5
05217 Basal cell carcinoma	5
05133 Pertussis	5
05164 Influenza A	5
05161 Hepatitis B	5
05140 Leishmaniasis	5
00514 Other types of O-glycan biosynthesis	4
00533 Glycosaminoglycan biosynthesis - keratan sulfate	4
00531 Glycosaminoglycan degradation	4
00563 Glycosylphosphatidylinositol GPI-anchor biosynthesis	4
04010 MAPK signaling pathway	4
04330 Notch signaling pathway	4

04066 HIF-1 signaling pathway	4
04520 Adherens junction	4
04620 Toll-like receptor signaling pathway	4
04650 Natural killer cell mediated cytotoxicity	4
04062 Chemokine signaling pathway	4
04910 Insulin signaling pathway	4
03320 PPAR signaling pathway	4
04912 GnRH signaling pathway	4
04914 Progesterone-mediated oocyte maturation	4
04270 Vascular smooth muscle contraction	4
04970 Salivary secretion	4
04971 Gastric acid secretion	4
04962 Vasopressin-regulated water reabsorption	4
04961 Endocrine and other factor-regulated calcium reabsorption	4
05215 Prostate cancer	4
05010 Alzheimer's disease	4
05020 Prion diseases	4
05032 Morphine addiction	4
05162 Measles	4
05160 Hepatitis C	4
05146 Amoebiasis	4
05144 Malaria	4
05142 Chagas disease American trypanosomiasis	4
00100 Steroid biosynthesis	3
00512 Mucin type O-glycan biosynthesis	3
04310 Wnt signaling pathway	3
04114 Oocyte meiosis	3
04115 p53 signaling pathway	3
04666 Fc gamma R-mediated phagocytosis	3
04672 Intestinal immune network for IgA production	3
04975 Fat digestion and absorption	3
04977 Vitamin digestion and absorption	3
04728 Dopaminergic synapse	3
04723 Retrograde endocannabinoid signaling	3
04713 Circadian entrainment	3
05034 Alcoholism	3
04940 Type I diabetes mellitus	3
05110 Vibrio cholerae infection	3
05169 Epstein-Barr virus infection	3
01212 Fatty acid metabolism	2
00052 Galactose metabolism	2
00520 Amino sugar and nucleotide sugar metabolism	2
00062 Fatty acid elongation	2

00590 Arachidonic acid metabolism	2
00240 Pyrimidine metabolism	2
00740 Riboflavin metabolism	2
04340 Hedgehog signaling pathway	2
04064 NF-kappa B signaling pathway	2
04530 Tight junction	2
04920 Adipocytokine signaling pathway	2
04260 Cardiac muscle contraction	2
04973 Carbohydrate digestion and absorption	2
04978 Mineral absorption	2
04964 Proximal tubule bicarbonate reclamation	2
04726 Serotonergic synapse	2
04722 Neurotrophin signaling pathway	2
05206 MicroRNAs in cancer	2
05214 Glioma	2
05218 Melanoma	2
05320 Autoimmune thyroid disease	2
05321 Inflammatory bowel disease IBD	2
05330 Allograft rejection	2
05332 Graft-versus-host disease	2
05030 Cocaine addiction	2
05031 Amphetamine addiction	2
05416 Viral myocarditis	2
05120 Epithelial cell signaling in Helicobacter pylori infection	2
01200 Carbon metabolism	1
01230 Biosynthesis of amino acids	1
00010 Glycolysis / Gluconeogenesis	1
00030 Pentose phosphate pathway	1
00051 Fructose and mannose metabolism	1
00500 Starch and sucrose metabolism	1
00562 Inositol phosphate metabolism	1
00190 Oxidative phosphorylation	1
00680 Methane metabolism	1
00140 Steroid hormone biosynthesis	1
01040 Biosynthesis of unsaturated fatty acids	1
00480 Glutathione metabolism	1
00627 Aminobenzoate degradation	1
00363 Bisphenol degradation	1
00983 Drug metabolism - other enzymes	1
03040 Spliceosome	1
03013 RNA transport	1
04012 ErbB signaling pathway	1
04370 VEGF signaling pathway	1

04150 mTOR signaling pathway	1
04146 Peroxisome	1
04210 Apoptosis	1
04621 NOD-like receptor signaling pathway	1
04660 T cell receptor signaling pathway	1
04662 B cell receptor signaling pathway	1
04664 Fc epsilon RI signaling pathway	1
04917 Prolactin signaling pathway	1
04960 Aldosterone-regulated sodium reabsorption	1
04742 Taste transduction	1
04626 Plant-pathogen interaction	1
05210 Colorectal cancer	1
05212 Pancreatic cancer	1
05221 Acute myeloid leukemia	1
05220 Chronic myeloid leukemia	1
05211 Renal cell carcinoma	1
05219 Bladder cancer	1
05213 Endometrial cancer	1
05223 Non-small cell lung cancer	1
05310 Asthma	1
05340 Primary immunodeficiency	1
05012 Parkinson's disease	1
05016 Huntington's disease	1
04930 Type II diabetes mellitus	1
04932 Non-alcoholic fatty liver disease NAFLD	1
05130 Pathogenic Escherichia coli infection	1
05132 Salmonella infection	1
05131 Shigellosis	1
05134 Legionellosis	1
05100 Bacterial invasion of epithelial cells	1
05143 African trypanosomiasis	1

Liver Versus Muscle (LvM442)

Pathway Name	Number of mapped Genes
04142 Lysosome	33
04151 PI3K-Akt signaling pathway	26
04060 Cytokine-cytokine receptor interaction	23
04141 Protein processing in endoplasmic reticulum	17
05205 Proteoglycans in cancer	17
04610 Complement and coagulation cascades	15
04514 Cell adhesion molecules CAMs	14
04510 Focal adhesion	14

04512 ECM-receptor interaction	13
04145 Phagosome	13
05200 Pathways in cancer	13
04630 Jak-STAT signaling pathway	11
05150 Staphylococcus aureus infection	11
05166 HTLV-I infection	11
00510 N-Glycan biosynthesis	10
04014 Ras signaling pathway	10
04015 Rap1 signaling pathway	10
04640 Hematopoietic cell lineage	10
05152 Tuberculosis	9
00564 Glycerophospholipid metabolism	8
04380 Osteoclast differentiation	8
05322 Systemic lupus erythematosus	8
05133 Pertussis	8
05168 Herpes simplex infection	8
05144 Malaria	8
04144 Endocytosis	7
04670 Leukocyte transendothelial migration	7
05164 Influenza A	7
00561 Glycerolipid metabolism	6
00531 Glycosaminoglycan degradation	6
04668 TNF signaling pathway	6
04066 HIF-1 signaling pathway	6
04976 Bile secretion	6
04974 Protein digestion and absorption	6
05202 Transcriptional misregulation in cancers	6
05203 Viral carcinogenesis	6
05160 Hepatitis C	6
05142 Chagas disease American trypanosomiasis	6
00513 Various types of N-glycan biosynthesis	5
00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin	5
04010 MAPK signaling pathway	5
04064 NF-kappa B signaling pathway	5
04020 Calcium signaling pathway	5
04810 Regulation of actin cytoskeleton	5
04650 Natural killer cell mediated cytotoxicity	5
04913 Ovarian Steroidogenesis	5
04614 Renin-angiotensin system	5
04975 Fat digestion and absorption	5
05222 Small cell lung cancer	5
04932 Non-alcoholic fatty liver disease NAFLD	5
05162 Measles	5

05145 Toxoplasmosis	5
00565 Ether lipid metabolism	4
00600 Sphingolipid metabolism	4
00230 Purine metabolism	4
00512 Mucin type O-glycan biosynthesis	4
00604 Glycosphingolipid biosynthesis - ganglio series	4
00511 Other glycan degradation	4
00983 Drug metabolism - other enzymes	4
02010 ABC transporters	4
04310 Wnt signaling pathway	4
04350 TGF-beta signaling pathway	4
04080 Neuroactive ligand-receptor interaction	4
04520 Adherens junction	4
04620 Toll-like receptor signaling pathway	4
04612 Antigen processing and presentation	4
04920 Adipocytokine signaling pathway	4
03320 PPAR signaling pathway	4
04915 Estrogen signaling pathway	4
04977 Vitamin digestion and absorption	4
04360 Axon guidance	4
05206 MicroRNAs in cancer	4
05215 Prostate cancer	4
05323 Rheumatoid arthritis	4
05010 Alzheimer's disease	4
05410 Hypertrophic cardiomyopathy HCM	4
05416 Viral myocarditis	4
05161 Hepatitis B	4
05169 Epstein-Barr virus infection	4
05146 Amoebiasis	4
00100 Steroid biosynthesis	3
00603 Glycosphingolipid biosynthesis - globo series	3
04012 ErbB signaling pathway	3
04390 Hippo signaling pathway	3
04146 Peroxisome	3
04530 Tight junction	3
04666 Fc gamma R-mediated phagocytosis	3
04910 Insulin signaling pathway	3
04918 Thyroid hormone synthesis	3
04270 Vascular smooth muscle contraction	3
05218 Melanoma	3
05020 Prion diseases	3
05412 Arrhythmogenic right ventricular cardiomyopathy ARVC	3
05414 Dilated cardiomyopathy DCM	3

05120 Epithelial cell signaling in Helicobacter pylori infection	3
05100 Bacterial invasion of epithelial cells	3
00030 Pentose phosphate pathway	2
00052 Galactose metabolism	2
00500 Starch and sucrose metabolism	2
00140 Steroid hormone biosynthesis	2
00590 Arachidonic acid metabolism	2
00240 Pyrimidine metabolism	2
00480 Glutathione metabolism	2
00514 Other types of O-glycan biosynthesis	2
00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	2
00533 Glycosaminoglycan biosynthesis - keratan sulfate	2
00563 Glycosylphosphatidylinositol GPI -anchor biosynthesis	2
00740 Riboflavin metabolism	2
00900 Terpenoid backbone biosynthesis	2
00627 Aminobenzoate degradation	2
03013 RNA transport	2
04370 VEGF signaling pathway	2
04210 Apoptosis	2
04672 Intestinal immune network for IgA production	2
04062 Chemokine signaling pathway	2
04916 Melanogenesis	2
04972 Pancreatic secretion	2
04978 Mineral absorption	2
04724 Glutamatergic synapse	2
04727 GABAergic synapse	2
04725 Cholinergic synapse	2
04728 Dopaminergic synapse	2
04722 Neurotrophin signaling pathway	2
05210 Colorectal cancer	2
05214 Glioma	2
05221 Acute myeloid leukemia	2
05220 Chronic myeloid leukemia	2
05211 Renal cell carcinoma	2
05219 Bladder cancer	2
05213 Endometrial cancer	2
05321 Inflammatory bowel disease IBD	2
05014 Amyotrophic lateral sclerosis ALS	2
05132 Salmonella infection	2
05140 Leishmaniasis	2
05143 African trypanosomiasis	2
01200 Carbon metabolism	1

01212 Fatty acid metabolism	1
01230 Biosynthesis of amino acids	1
00010 Glycolysis / Gluconeogenesis	1
00040 Pentose and glucuronate interconversions	1
00051 Fructose and mannose metabolism	1
00562 Inositol phosphate metabolism	1
00680 Methane metabolism	1
00062 Fatty acid elongation	1
01040 Biosynthesis of unsaturated fatty acids	1
00330 Arginine and proline metabolism	1
00440 Phosphonate and phosphinate metabolism	1
00601 Glycosphingolipid biosynthesis - lacto and neolacto series	1
00780 Biotin metabolism	1
00790 Folate biosynthesis	1
00860 Porphyrin and chlorophyll metabolism	1
00944 Flavone and flavonol biosynthesis	1
00232 Caffeine metabolism	1
00363 Bisphenol degradation	1
00980 Metabolism of xenobiotics by cytochrome P450	1
03060 Protein export	1
02020 Two-component system	1
04330 Notch signaling pathway	1
04391 Hippo signaling pathway -fly	1
04150 mTOR signaling pathway	1
04110 Cell cycle	1
04115 p53 signaling pathway	1
04540 Gap junction	1
04621 NOD-like receptor signaling pathway	1
04660 T cell receptor signaling pathway	1
04662 B cell receptor signaling pathway	1
04664 Fc epsilon RI signaling pathway	1
04911 Insulin secretion	1
04912 GnRH signaling pathway	1
04914 Progesterone-mediated oocyte maturation	1
04917 Prolactin signaling pathway	1
04970 Salivary secretion	1
04971 Gastric acid secretion	1
04973 Carbohydrate digestion and absorption	1
04962 Vasopressin-regulated water reabsorption	1
04960 Aldosterone-regulated sodium reabsorption	1
04964 Proximal tubule bicarbonate reclamation	1
04966 Collecting duct acid secretion	1
04726 Serotonergic synapse	1

04626 Plant-pathogen interaction	1
05204 Chemical carcinogenesis	1
05212 Pancreatic cancer	1
05216 Thyroid cancer	1
05217 Basal cell carcinoma	1
05223 Non-small cell lung cancer	1
05320 Autoimmune thyroid disease	1
05330 Allograft rejection	1
05332 Graft-versus-host disease	1
05340 Primary immunodeficiency	1
05016 Huntington's disease	1
05030 Cocaine addiction	1
05031 Amphetamine addiction	1
05034 Alcoholism	1
04940 Type I diabetes mellitus	1
04930 Type II diabetes mellitus	1
05110 Vibrio cholerae infection	1
05131 Shigellosis	1
05134 Legionellosis	1

Supplementary Table 4S1: List of all 923 glycogenes used in this study.

List of 923 glycogenes used in this study								
Aadac	C3	D10Ert438e	Galnt11	Itm2a	Ndst2	Ptprs	Slc43a2	Trak1
Aadacl1	C4b	D10Jhu81e	Galnt2	Itm2b	Neo1	Pttg1ip	Slc44a2	Trf
Abca1	C8a	D11Ert418e	Galnt12	Itm2c	Neu1	Pvrl2	Slc4a2	Trfr2
Abca2	C8b	D19Wsu12e	Galnt14	Jag2	Ngfb	Pvrl3	Slc5a6	Tspan13
Abca8a	C8g	D1Bwg1363e	Ganab	Jam2	Ngrn	Pxdn	Slc6a12	Tspan14
Abca8b	Cacna2d3	D8Ert4354e	Gas6	Jam3	Nid1	Pxmp4	Slc6a13	Tspan15
Abcb11	Cadm1	Dag1	Gc	Kcnc1	Nog	Qpct	Slc6a8	Tspan17
Abcb1a	Cadm3	Dand5	Gcgr	Kcne1l	Nomo1	Qscn6	Slc7a2	Tspan2
Abcb4	Cadm4	Dcbld1	Gcs1	Kcnk3	Nope	Qscn6l1	Slc7a5	Tspan3
Abcc2	Cant1	Dcn	Gdf10	Kdr	Notch1	Ramp2	Slc9a1	Tspan31
Abcc3	Ccdc126	Dgcr2	Gfra1	Klhd7a	Nov	Rbmxt	Slco1b2	Tspan33
Abcc5	Ccdc80	Dirc2	Ghdc	Klkb1	Npc1	Rcn1	Slco2a1	Tspan4
Abcc6	Ccrl1	Dkk3	Ghr	Kng1	Npr1	Rcn3	Slco2b1	Tspan5
Abcd2	Cd151	Dkl1	Gla	Kremen1	Nptn	Reg1	Slit3	Tspan6
Abcd3	Cd164	Dlk1	Glg1	Krt18	Nrg4	Rfng	Smo	Tspan7
Abcg2	Cd1d1	Dll4	Gm2a	Krt8	Nrp1	Rgma	Smoc1	Tspan9
Abcg8	Cd200	Dnajb11	Gnptab	Ktn1	Ntn1	Rgmb	Smoc2	Ttr
Abhd12	Cd300lg	Dnajc16	Gnptg	Lamb1-1	Ntn2l	Rhbdf1	Smpd1	Ttyh2
Abhd14a	Cd302	Dnase1l1	Gns	Lamb2	Ntng2	Rnf128	Smpdl3a	Ttyh3

Ache	Cd320	Dnase2a	Golph4	Lamp2	Nup210	Rnf130	Sorbs1	Tusc3
Acp2	Cd36	Dnase2b	Gpc1	Large	Nup214	Rnf149	Sorl1	Txndc12
Acp5	Cd37	Dnm1l	Gpc4	Lass2	Nup54	Rnf167	Sort1	Uchl1
Acpl2	Cd52	Dpagt1	Gpiap1	Lass5	Nup62	Rpn1	Sparc	Ugt1a1
Acvr1	Cd53	Dpp7	Gpihbp1	Lbp	Nupl2	Rpn2	Sparcl1	Ugt1a10
Acvr1b	Cd55	Ece1	Gpld1	Lbr	Odz3	Rspry1	Spcs3	Ugt1a2
Adam11	Cd59a	Ecm1	Gpr107	Lcat	Olfm1	Rtn4rl1	Spint2	Ugt1a6a
Adam15	Cd5l	Edem1	Gpr124	Lcn2	Olfml3	Ryk	Spnb2	Ugt1a7c
Adam23	Cd63	Edem3	Gpr137	Ldlr	Orm1	Sat1	Srf	Ugt1a9
Adam9	Cd68	Edg1	Gpr146	Ldlrad3	Ostm1	Sbp	Srgn	Ugt2a3
Adamts2	Cd74	Ednrb	Gpr172b	Leprel2	P2rx4	Scap	Srl	Ugt3a2
Adcy2	Cd82	Efemp2	Gpr175	Lfng	P2rxl1	Scara3	Ssr1	Unc5a
Adcy6	Cd99l2	Efna1	Gpr19	Lgals3bp	P2ry5	Scarb1	Ssr2	Unc5b
Adcy9	Cdh5	Egflam	Gpr56	Lgmh	P4ha1	Scarb2	St3gal1	Unc93b1
Admr	Ceacam1	Egfr	Gpr89	Lgr4	P4ha2	Scn1b	St3gal2	Uxs1
Adora1	Ceacam10	Ela1	Gprc5b	Lifr	Pam	Scnn1a	St3gal4	Vasn
Aebp1	Cel	Elane	Gprc5c	Lipa	Pcnx	Scpep1	St3gal6	Vcam1
Afm	Cept1	Emcn	Grn	Lipc	Pcnxl3	Sdc1	St6gal1	Vegfb
Aga	Ces1	Emilin1	Gusb	Lman2	Pcolce	Sdc2	St6galnac4	Vldlr
Ager	Ces1d	Emp2	Gyg	Lman2l	Pcsk4	Sdc3	St6galnac6	Vnn3
Agpat6	Ces3a	Emp3	H13	Lmbrd1	Pcsk9	Sdc4	St7	Vtn
Agpat7	Ces3b	Emr1	H2-Aa	Lmbrd2	Pcyox1	Sdf4	Stab1	Vwa1
Agrn	Cfb	Eng	H2-D1	Lpar1	Pddc1	Sel1l	Steap3	Vwf
Agt	Cfh	Enpep	H2-K1	Lphn1	Pdgfa	Selpl	Stim1	Wfikkn2
Agtr1a	Cfi	Enpp1	H2-Q10	Lpl	Pdgfra	Sema3c	Stt3a	Wfs1
Ahsg	Cfp	Enpp2	H2-Q8	Lrfn3	Pdgfrb	Sema3f	Sulf1	Wnt5b
Akp2	Chga	Enpp4	H2-T23	Lrig1	Pdia4	Sema4a	Sulf2	Xdh
Akt1	Chrnbl	Enpp5	H6pd	Lrp1	Pecam1	Sema4b	Sumf1	Xlkd1
Akt2	Chst12	Entpd2	Hapln4	Lrp5	Perld1	Sema4c	Susd4	Xylt2
Akt3	Clcn3	Entpd5	Hbp1	Lrp6	Pfkl	Sema4g	Sypl	Yipf1
Alg5	Cldnd1	Epb4.1	Hcfc1	Lrpap1	Pfkm	Sema6b	Sypl2	Yipf3
Alg9	Clec11a	Ephb4	Hexb	Lrrc38	Pfkp	Sema6c	Tapbp	Zfyve27
Ambp	Clec2d	Erbb3	Hfe	Lrrc8a	Pgcp	Sema7a	Tapbpl	
Amigo1	Clec4f	Ergic1	Hfe2	Lrrn1	Pglyrp2	Sepp1	Tcta	
Amigo2	Clec4g	Ergic3	Hgfac	Ltbr	Pgs1	Serpina10	Tex2	
Angptl2	Cln3	Erlin1	Hiat1	Lum	Pigo	Serpina11	Tgfa	
Angptl3	Cln5	Erlin2	Hiatl1	Ly6a	Pigr	Serpina1c	Tgfb1	
Angptl4	Clptm1	Ern1	Hist1h2bc	Ly6d	Pigs	Serpina1e	Tgoln1	
Anpep	Clptm1l	Ero1lb	Hmgcr	Ly86	Pigt	Serpina3c	Thbd	
Apcdd1	Clstn1	Es22	Hpn	Lycat	Pkd1	Serpina3k	Thbs4	
Aplp2	Clstn3	Eva1	Hpxn	Lynx1	Pkd2	Serpina3m	Thy1	
Apoc3	Clu	Evc2	Hrb	Lypd1	Pla1a	Serpina3n	Tie1	
Apoc4	Cmklr1	Evi2a	Hs2st1	Lypla3	Pla2g7	Serpina6	Timd2	

Apod	Cnnm2	Ext1	Hs6st1	Lysmd4	Plat	Serpinc1	Tinagl
Apoh	Cnnm3	Extl1	Hsd11b1	M6pr	Pld3	Serpind1	Tkt
App	Cnnm4	Extl2	Hsd17b6	Man1a	Plekhh3	Serpine2	Tm2d1
Aqp4	Cntfr	Extl3	Hsd17b7	Man2a1	Pltp	Serpinf1	Tm2d3
Aqp8	Col11a2	F11r	Hsp90b1	Man2a2	Plvap	Serpinf2	Tm4sf1
Arsa	Col14a1	F12	Hyal1	Man2b1	Plxdc1	Serping1	Tm4sf4
Arts1	Col15a1	F13b	Hyal2	Man2b2	Plxna1	Serpinh1	Tm7sf3
Asah3l	Col18a1	F2	Hyal3	Manba	Plxnb1	Sgce	Tm9sf1
Asahl	Col3a1	F2r	Hyou1	Masp1	Plxnb2	Siae	Tm9sf3
Asgr1	Col4a1	F3	Icam1	Masp2	Plxnc1	Sigirr	Tmco3
Asgr2	Col4a2	F5	Icosl	Mbl1	Plxnd1	Sil1	Tmem104
Atp13a1	Col6a2	Fbln1	Ids	Mbl2	Pmp22	Sirpa	Tmem106b
Atp1b1	Colec12	Fbln2	Ifi30	Mboat5	Pnpla2	Slc10a1	Tmem106c
Atp1b2	Cova1	Fcgr2b	Ifnar1	Mboat7	Pnpla6	Slc12a4	Tmem123
Atp1b3	Cpb2	Fcgr3	Ifnar2	Mbtps1	Pnpla7	Slc12a6	Tmem129
Atp6ap1	Cpd	Fcgrt	Ifngr1	Mcam	Pnpla8	Slc12a7	Tmem132a
Atrn	Cpe	Fcmd	Igfals	Mchr1	Podxl2	Slc12a9	Tmem142a
Atrnl1	Cpn1	Fcna	Igfbp2	Mdga1	Pomt2	Slc13a3	Tmem150
Azgp1	Cpn2	Fetub	Igfbp3	Megf8	Pon1	Slc17a2	Tmem157
B3galnt1	Crb3	Fgb	Igfbp4	Megf9	Pon3	Slc17a5	Tmem158
B3galnt2	Creb3l1	Fgfr3	Igfbp6	Mest	Postn	Slc19a1	Tmem159
B3galt6	Creb3l2	Fgg	Igfbp7	Met	Ppap2a	Slc1a2	Tmem161a
B3gnt1	Creb3l3	Fkbp7	Igsf8	Metap2	Ppap2b	Slc1a3	Tmem167
B4galnt1	Crebl1	Fkbp9	Il15ra	Metrn1	Ppap2c	Slc1a4	Tmem168
B4galt3	Creg1	Fkrp	Il17ra	Mettl9	Ppia	Slc22a1	Tmem16f
B4galt4	Creld1	Flt1	Il17rc	Mfap4	Ppt1	Slc22a5	Tmem16h
B4galt5	Creld2	Flt3l	Il18bp	Mfge8	Pqlc2	Slc23a2	Tmem180
B4galt7	Crim1	Flt4	Il1rap	Mfrp	Prkcsh	Slc24a3	Tmem25
Bace1	Crisp1	Fn1	Il4ra	Mfsd11	Prnp	Slc24a6	Tmem30a
Bcam	Crispld2	Fndc5	Il6ra	Mgat2	Proc	Slc29a1	Tmem39b
Bgn	Crry	Fstl1	Inha	Minpp1	Pros1	Slc29a2	Tmem57
Bmp1	Csf1	Fuca1	Inhbb	Mir16	Prss23	Slc2a4	Tmem59
Bmp2	Csf1r	Fuca2	Inhbc	Mkl2	Prtn3	Slc30a1	Tmem63a
Bmp4	Csf2ra	Furin	Insr	Mmp15	Psap	Slc31a1	Tmem63b
Bmp6	Ctbs	Fut11	Islr	Mmp2	Psma1	Slc33a1	Tmem87a
Bmpr1a	Ctns	Fxyd5	Itfg3	Mmrn2	Ptch1	Slc35c2	Tmprss6
Boc	Ctsa	Fzd4	Itga2b	Mpeg1	Ptdss2	Slc36a1	Tnfrsf19
Bscl2	Ctsc	Fzd8	Itga3	Mpzl1	Ptgds	Slc37a3	Tnfrsf19l
Bsg	Ctsf	Fzd9	Itga5	Mr1	Ptger1	Slc38a2	Tnfrsf1a
Bst2	Ctsh	G6pc	Itga7	Mrc1	Ptgfrn	Slc38a3	Tnfrsf1b
Btd	Ctsk	Gaa	Itgav	Mrgprf	Ptgs1	Slc38a4	Tnfrsf21
C1qa	Ctsl	Gabbr1	Itgb1	Mug2	Pthr1	Slc39a14	Tnrc5
C1qb	Ctso	Gabbr2	Itgb4	Mxra8	Ptplb	Slc39a5	Tor1a

C1qtnf1	Ctss	Galc	Itih2	Myc	Ptpa	Slc39a6	Tor1aip2
C1r	Ctsz	Galnact2	Itih3	Naga	Ptprf	Slc39a9	Tor2a
C1rl	Cxadr	Galns	Itih4	Ncstn	Ptprg	Slc3a2	Tpst1
C1s	Cxcl16	Galnt10	Itih5	Ndst1	Ptprm	Slc40a1	Tpte2

Supplementary Table 4S2: List of all 222 KEGG pathways reported for all the 923 glycogenes, contains the list of KEGG pathways and the number of mapped genes.

Pathway Name	Number of Mapped Genes
04142 Lysosome	49
04151 PI3K-Akt signaling pathway	45
04610 Complement and coagulation cascades	39
04060 Cytokine-cytokine receptor interaction	38
05200 Pathways in cancer	30
04514 Cell adhesion molecules CAMs	28
05205 Proteoglycans in cancer	27
04510 Focal adhesion	26
04141 Protein processing in endoplasmic reticulum	24
04015 Rap1 signaling pathway	23
04512 ECM-receptor interaction	21
04976 Bile secretion	21
05166 HTLV-I infection	21
04145 Phagosome	20
04080 Neuroactive ligand-receptor interaction	18
05150 Staphylococcus aureus infection	18
04014 Ras signaling pathway	17
04360 Axon guidance	17
04810 Regulation of actin cytoskeleton	16
04640 Hematopoietic cell lineage	16
00564 Glycerophospholipid metabolism	14
00510 N-Glycan biosynthesis	14
04630 Jak-STAT signaling pathway	14
04020 Calcium signaling pathway	14
05152 Tuberculosis	14
02010 ABC transporters	13
05322 Systemic lupus erythematosus	13
04380 Osteoclast differentiation	12
05414 Dilated cardiomyopathy DCM	12
00534 Glycosaminoglycan biosynthesis - heparan sulfate / heparin	11
04010 MAPK signaling pathway	11
04144 Endocytosis	11
05206 MicroRNAs in cancer	11
05168 Herpes simplex infection	11
04670 Leukocyte transendothelial migration	10
04915 Estrogen signaling pathway	10

04918 Thyroid hormone synthesis	10
04974 Protein digestion and absorption	10
05222 Small cell lung cancer	10
05410 Hypertrophic cardiomyopathy HCM	10
05144 Malaria	10
00230 Purine metabolism	9
00513 Various types of N-glycan biosynthesis	9
04350 TGF-beta signaling pathway	9
04668 TNF signaling pathway	9
04916 Melanogenesis	9
05202 Transcriptional misregulation in cancers	9
05412 Arrhythmogenic right ventricular cardiomyopathy ARVC	9
05133 Pertussis	9
05164 Influenza A	9
00561 Glycerolipid metabolism	8
00600 Sphingolipid metabolism	8
04390 Hippo signaling pathway	8
04066 HIF-1 signaling pathway	8
04520 Adherens junction	8
04540 Gap junction	8
04612 Antigen processing and presentation	8
04913 Ovarian Steroidogenesis	8
04972 Pancreatic secretion	8
05203 Viral carcinogenesis	8
05215 Prostate cancer	8
05323 Rheumatoid arthritis	8
05146 Amoebiasis	8
05145 Toxoplasmosis	8
00531 Glycosaminoglycan degradation	7
00604 Glycosphingolipid biosynthesis - ganglio series	7
00511 Other glycan degradation	7
04310 Wnt signaling pathway	7
04270 Vascular smooth muscle contraction	7
04724 Glutamatergic synapse	7
05217 Basal cell carcinoma	7
05010 Alzheimer's disease	7
05160 Hepatitis C	7
00500 Starch and sucrose metabolism	6
00565 Ether lipid metabolism	6
00532 Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	6
00603 Glycosphingolipid biosynthesis - globo series	6
04012 ErbB signaling pathway	6
04068 FoxO signaling pathway	6
04530 Tight junction	6
04911 Insulin secretion	6
04910 Insulin signaling pathway	6
04920 Adipocytokine signaling pathway	6
03320 PPAR signaling pathway	6
04970 Salivary secretion	6
04971 Gastric acid secretion	6

04975 Fat digestion and absorption	6
04977 Vitamin digestion and absorption	6
04727 GABAergic synapse	6
04725 Cholinergic synapse	6
05214 Glioma	6
05218 Melanoma	6
05020 Prion diseases	6
05416 Viral myocarditis	6
05120 Epithelial cell signaling in Helicobacter pylori infection	6
05161 Hepatitis B	6
05140 Leishmaniasis	6
05142 Chagas disease American trypanosomiasis	6
00533 Glycosaminoglycan biosynthesis - keratan sulfate	5
00601 Glycosphingolipid biosynthesis - lacto and neolacto series	5
00983 Drug metabolism - other enzymes	5
03013 RNA transport	5
04330 Notch signaling pathway	5
04064 NF-kappa B signaling pathway	5
04620 Toll-like receptor signaling pathway	5
04650 Natural killer cell mediated cytotoxicity	5
04672 Intestinal immune network for IgA production	5
04062 Chemokine signaling pathway	5
04912 GnRH signaling pathway	5
04614 Renin-angiotensin system	5
04978 Mineral absorption	5
05032 Morphine addiction	5
04932 Non-alcoholic fatty liver disease NAFLD	5
05162 Measles	5
05169 Epstein-Barr virus infection	5
00052 Galactose metabolism	4
00140 Steroid hormone biosynthesis	4
00512 Mucin type O-glycan biosynthesis	4
00514 Other types of O-glycan biosynthesis	4
00563 GlycosylphosphatidylinositolGPI-anchor biosynthesis	4
04340 Hedgehog signaling pathway	4
04146 Peroxisome	4
04210 Apoptosis	4
04666 Fc gamma R-mediated phagocytosis	4
04914 Progesterone-mediated oocyte maturation	4
04962 Vasopressin-regulated water reabsorption	4
04961 Endocrine and other factor-regulated calcium reabsorption	4
05219 Bladder cancer	4
05100 Bacterial invasion of epithelial cells	4
00100 Steroid biosynthesis	3
00740 Riboflavin metabolism	3
04114 Oocyte meiosis	3
04115 p53 signaling pathway	3
04260 Cardiac muscle contraction	3
04973 Carbohydrate digestion and absorption	3
04960 Aldosterone-regulated sodium reabsorption	3

04728 Dopaminergic synapse	3
04723 Retrograde endocannabinoid signaling	3
04722 Neurotrophin signaling pathway	3
04713 Circadian entrainment	3
05212 Pancreatic cancer	3
05211 Renal cell carcinoma	3
05213 Endometrial cancer	3
05223 Non-small cell lung cancer	3
05321 Inflammatory bowel disease IBD	3
05014 Amyotrophic lateral sclerosis ALS	3
05034 Alcoholism	3
04940 Type I diabetes mellitus	3
05110 Vibrio cholerae infection	3
01212 Fatty acid metabolism	2
00010 Glycolysis / Gluconeogenesis	2
00030 Pentose phosphate pathway	2
00040 Pentose and glucuronate interconversions	2
00520 Amino sugar and nucleotide sugar metabolism	2
00062 Fatty acid elongation	2
00590 Arachidonic acid metabolism	2
00240 Pyrimidine metabolism	2
00480 Glutathione metabolism	2
00770 Pantothenate and CoA biosynthesis	2
00830 Retinol metabolism	2
00860 Porphyrin and chlorophyll metabolism	2
00900 Terpenoid backbone biosynthesis	2
00627 Aminobenzoate degradation	2
00980 Metabolism of xenobiotics by cytochrome P450	2
04370 VEGF signaling pathway	2
04964 Proximal tubule bicarbonate reclamation	2
04726 Serotonergic synapse	2
04742 Taste transduction	2
04320 Dorso-ventral axis formation	2
05204 Chemical carcinogenesis	2
05210 Colorectal cancer	2
05221 Acute myeloid leukemia	2
05220 Chronic myeloid leukemia	2
05320 Autoimmune thyroid disease	2
05330 Allograft rejection	2
05332 Graft-versus-host disease	2
05030 Cocaine addiction	2
05031 Amphetamine addiction	2
04930 Type II diabetes mellitus	2
05132 Salmonella infection	2
05131 Shigellosis	2
05134 Legionellosis	2
05143 African trypanosomiasis	2
01200 Carbon metabolism	1
01230 Biosynthesis of amino acids	1
00051 Fructose and mannose metabolism	1

00053 Ascorbate and aldarate metabolism	1
00562 Inositol phosphate metabolism	1
00190 Oxidative phosphorylation	1
00680 Methane metabolism	1
01040 Biosynthesis of unsaturated fatty acids	1
00330 Arginine and proline metabolism	1
00440 Phosphonate and phosphinate metabolism	1
00760 Nicotinate and nicotinamide metabolism	1
00780 Biotin metabolism	1
00790 Folate biosynthesis	1
00944 Flavone and flavonol biosynthesis	1
00232 Caffeine metabolism	1
00363 Bisphenol degradation	1
00982 Drug metabolism - cytochrome P450	1
03040 Spliceosome	1
03060 Protein export	1
03050 Proteasome	1
03460 Fanconi anemia pathway	1
02020 Two-component system	1
04013 MAPK signaling pathway - fly	1
04391 Hippo signaling pathway -fly	1
04150 mTOR signaling pathway	1
04110 Cell cycle	1
04621 NOD-like receptor signaling pathway	1
04660 T cell receptor signaling pathway	1
04662 B cell receptor signaling pathway	1
04664 Fc epsilon RI signaling pathway	1
04917 Prolactin signaling pathway	1
04966 Collecting duct acid secretion	1
04626 Plant-pathogen interaction	1
05216 Thyroid cancer	1
05310 Asthma	1
05340 Primary immunodeficiency	1
05012 Parkinson's disease	1
05016 Huntington's disease	1
05130 Pathogenic Escherichia coli infection	1

Supplementary Table 4S3: nsSNPs of HPXN gene in mouse and their prediction by Polyphen2, PROVEAN and SIFT.

Uniprot ID	SNP ID	AA change	PPH2_prob	PPH2	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (Cutoff= < .05)
Q91X72	rs219016026	T28I	0.002	Benign	-0.876	Neutral	0.23	Tolerated
	rs13471517	N30H	0	Benign	1.849	Neutral	0.6	Tolerated
	rs219641236	S182P	0	Benign	-0.394	Neutral	1	Tolerated
	rs252364187	T183A	0	Benign	0.612	Neutral	1	Tolerated
	rs224885217	T262S	0	Benign	1.174	Neutral	1	Tolerated
	rs216873853	P351S	0.001	Benign	0.532	Neutral	0.76	Tolerated
	rs250210929	S421P	0	Benign	0.071	Neutral	0.33	Tolerated

Supplementary Table 4S4: nsSNPs of FLT1 gene in mouse and their prediction by Polyphen2, PROVEAN and SIFT.

Uniprot ID	SNP ID	AA Change	PPH2_Prob	PPH2	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (cutoff= < .05)
P35969	rs214988215	S80L	0.81	Possibly damaging	-2.053	Neutral	0.11	Tolerated
	rs33730929	D85N	0	Benign	0.224	Neutral	0.75	Tolerated
	rs233917867	F89L	0.942	Possibly damaging	-3.699	Deleterious	0.1	Tolerated
	rs32388970	T97M	0	Benign	-0.714	Neutral	0.08	Tolerated
	rs252361901	T113K	0	Benign	-0.904	Neutral	0.08	Tolerated
	rs213622875	T150M	0.638	Possibly damaging	-1.885	Neutral	0.02	Affect protein function
	rs233071760	A277V	0.994	Probably damaging	-1.296	Neutral	0.08	Tolerated
	rs32394398	D284E	0.821	Possibly damaging	-0.069	Neutral	1	Tolerated
	rs232914047	L453F	1	Probably damaging	-3.129	Deleterious	0	Affect protein function
	rs237086847	H476Q	0.311	Benign	-1.271	Neutral	0.64	Tolerated
	rs32392431	Q529R	0	Benign	0.347	Neutral	0.74	Tolerated
	rs222631065	K575N	0.792	Possibly damaging	-0.425	Neutral	0.5	Tolerated

	rs256025183	Y671F	0	Benign	-1.13	Neutral	0.71	Tolerated
	rs257829456	V688I	0.03	Benign	-0.495	Neutral	0.56	Tolerated
	rs241489228	V791I	0	Benign	0.499	Neutral	1	Tolerated
	rs259750200	S945N	0	Benign	-0.214	Neutral	0.76	Tolerated
	rs242252147	S970G	0	Benign	2.238	Neutral	0.84	Tolerated
	rs260647574	S987N	0.052	Benign	-0.393	Neutral	0.32	Tolerated
	rs32389134	Q1164K	0.232	Benign	-2.926	Deleterious	0.02	Affect protein function
	rs224361070	P1201S	0.124	Benign	-1.755	Neutral	0.8	Tolerated
	rs239789789	H1202Q	0	Benign	0.603	Neutral	NA	Tolerated
	rs263438024	H1202R	0	Benign	0.768	Neutral	0.69	Tolerated
	rs213904262	V1303M	0.3	Benign	-0.167	Neutral	0.23	Tolerated

Supplementary Table 4S5: nsSNPs of NID2 gene in mouse and their prediction by Polyphen2, PROVEAN and SIFT.

Uniprot ID	SNP ids	AA Change	PPH2_prob	PPH2	PROVEAN score	PROVEAN (Cutoff=-2.5)	SIFT Score	SIFT (cutoff= < .05)
O88322	rs262873422	P13L	0.93	Possibly damaging	-0.254	Neutral	0.01	Affect protein function
	rs30692168	S14P	0.734	Possibly damaging	-0.536	Neutral	0	Affect protein function
	rs235824951	L146Q	0	Benign	1.12	Neutral	0.17	Tolerated
	rs30689780	V266A	0	Benign	-3.217	Deleterious	0	Affect protein function
	rs223532410	S290T	0.807	Possibly damaging	-0.531	Neutral	0	Affect protein function
	rs254798552	Y317D	0.987	Probably damaging	-2.252	Neutral	0	Affect protein function
	rs241645935	D351G	0.003	Benign	-1.699	Neutral	0	Affect protein function
	rs30689779	V357I	0	Benign	-0.137	Neutral	0	Affect protein function
	rs231633635	D406N	0	Benign	-0.805	Neutral	0	Affect protein function

	rs30689778	R408K	0	Benign	0.589	Neutral	0.92	Tolerated
	rs213611303	V461A	0	Benign	1.033	Neutral	0.83	Tolerated
	rs261439016	T525A	0.983	Probably damaging	-2.642	Deleterious	0.68	Tolerated
	rs217213236	R733S	0	Benign	-1.073	Neutral	0.77	Tolerated
	rs246954764	A743T	0	Benign	0.609	Neutral	0.77	Tolerated
	rs30667034	E810K	0.95	Possibly damaging	-2.586	Deleterious	0.17	Tolerated
	rs255251970	A869T	0.931	Possibly damaging	0.255	Neutral	1	Tolerated
	rs214587624	P872A	0	Benign	0.057	Neutral	0.92	Tolerated
	rs229880960	H911Y	0.008	Benign	-2.252	Neutral	1	Tolerated
	rs224292932	A920T	0.027	Benign	-0.1	Neutral	0.45	Tolerated
	rs30692024	E1242 G	0	Benign	-1.6	Neutral	0.36	Tolerated
	rs212959671	N1341 S	0	Benign	-1.672	Neutral	0.12	Tolerated

Supplementary Table 4S6: nsSNPs of TNFRSF1B gene in mouse and their prediction by Polyphen2, PROVEAN and SIFT.

Uniprot ID	SNP ID	AA Change	PPH2_prob	PPH2	PROVEAN Score	PROVEAN (Cutoff= -2.5)	SIFT Score	SIFT (cutoff= <.05)
P25119	rs13463608	S43L	0	Benign	-1.069	Neutral	0.3	Tolerated
	rs241492411	F91L	0	Benign	1.288	Neutral	1	Tolerated
	rs27627652	T102S	0.001	Benign	1.515	Neutral	0.87	Tolerated
	rs27627654	I108T	0.001	Benign	2.14	Neutral	0.99	Tolerated
	rs261122427	K164R	0	Benign	-1.099	Neutral	0.2	Tolerated
	rs237357155	V218I	0.245	Benign	-0.217	Neutral	0.38	Tolerated
	rs48589646	I283F	0.002	Benign	-0.576	Neutral	0.48	Tolerated
	rs27627670	F360S	0	Benign	1.15	Neutral	0.56	Tolerated
	rs250384859	S396N	0.858	Possibly damaging	-0.651	Neutral	0	Affect protein function
	rs50901502	P431L	0.996	Probably damaging	-2.631	Deleterious	0.01	Affect protein function
	rs27627696	C436Y	0	Benign	0.293	Neutral	0.22	Tolerated

Research Article

Comparative Analysis of Glycogene Expression in Different Mouse Tissues Using RNA-Seq Data

Ahmad Firoz,^{1,2} Adeel Malik,³ Sanjay Kumar Singh,^{2,4} Vivekanand Jha,^{2,5} and Amjad Ali¹

¹ School of Chemistry and Biochemistry, Thapar University, Patiala, Punjab 147004, India

² Biomedical Informatics Center of ICMR, Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh 160012, India

³ School of Biotechnology, Yeungnam University, 280 Daehak-ro, Gyeongsan, Gyeongbuk 712-749, Republic of Korea

⁴ Department of Biotechnology, Thapar University, Patiala, Punjab 147004, India

⁵ Department of Nephrology, Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh 160012, India

Correspondence should be addressed to Ahmad Firoz; ahmadfirozbin@gmail.com and Adeel Malik; adeel@procarb.org

Received 13 May 2014; Accepted 10 June 2014; Published 9 July 2014

Academic Editor: Giulia Piaggio

Copyright © 2014 Ahmad Firoz et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Glycogenes regulate a wide array of biological processes in the development of organisms as well as different diseases such as cancer, primary open-angle glaucoma, and renal dysfunction. The objective of this study was to explore the role of differentially expressed glycogenes (DEGGs) in three major tissues such as brain, muscle, and liver using mouse RNA-seq data, and we identified 579, 501, and 442 DEGGs for brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups. DAVID functional analysis suggested inflammatory response, glycosaminoglycan metabolic process, and protein maturation as the enriched biological processes in BvL579, BvM501, and LvM442, respectively. These DEGGs were then used to construct three interaction networks by using GeneMANIA, from which we detected potential hub genes such as *PEMT* and *HPXN* (BvL579), *IGF2* and *NID2* (BvM501), and *STAT6* and *FLT1* (LvM442), having the highest degree. Additionally, our community analysis results suggest that the significance of immune system related processes in liver, glycosphingolipid metabolic processes in the development of brain, and the processes such as cell proliferation, adhesion, and growth are important for muscle development. Further studies are required to confirm the role of predicted hub genes as well as the significance of biological processes.

1. Introduction

Cell function within an organism has huge variation due to gene expression pattern although all cells in an individual mammal have almost identical DNA [1]. It is important to know how cells and tissues differ in gene expression which is regulated during developmental changes in different tissues, consequently affecting their biological functions. The pattern of glycogene expression is one of important factors that provide clues about many biological functions, developmental changes, and diseases in human, mouse, and tissues of other organisms [2–4]. Glycogene is a gene that is responsible for the glycosylation of proteins, lipids, and proteoglycans and includes genes associated with the synthesis of glycans such as

glycosyltransferase, sugar-nucleotide transporters, and sulfo-transferases [5, 6]. Similarly, the term glycogenome involves all genes that play a role in glycosylation and represents more than 600 genes in the mouse genome [7]. Glycosylation is one of the most common posttranslational modification reactions [8]. The glycogene expression is relatively weak as compared to other molecules; however, different glycogenes are upregulated by specific tissues depending on the local conditions [5].

In order to understand the pathology of various diseases, recently, many studies have targeted and explored the role of glycogenes expression in various diseases [9–12]. Alteration in the glycogene expression is also associated with drug resistance of various types of cancers including pancreatic

cancer epithelial-mesenchymal transition [13], breast cancer [14], and hepatocarcinoma [15]. The change in the glycogene expression is also associated with conjunctival epithelium [2] and primary open-angle glaucoma (POAG) [16]. Moreover, the defects in the normal expression of glycogenes in tonsillar B lymphocytes are also correlated with the proteinuria and renal dysfunction in IgA nephropathy [17, 18].

Several high-throughput microarray studies have been applied to investigate the functional roles of various glycogenes [11, 19, 20]. However, to the best of our knowledge, no study till date has utilized RNA-Seq data in the existing databases to specifically explore the role of glycogenes. Previously, we identified the role of glycogenes in skeletal muscle development in MYOG_{kd} cells by using RNA-Seq data. Therefore, in this work, we intend to investigate the role of glycogenes in various biological processes that are involved in the development of brain, muscle, and liver tissues using the precomputed expression values from mouse RNA-seq data. Functional and pathway analysis of differentially expressed glycogenes (DEGGs) between BvL579, BvM501, and LvM442 were enriched with gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, respectively. Additionally, the interaction networks of DEGGs were constructed to identify the hub genes as well as to detect the functional modules in these networks by using community analysis. We expect that the findings of our study may shed new lights on the roles of glycogenes in the development of these tissues and their pathogenesis.

2. Materials and Methods

2.1. Datasets. The precomputed expression values with no spike for the brain, liver, and muscles C57BL6 mouse tissues were downloaded from WoldLab (<http://woldlab.caltech.edu/~alim/RNAseq/>) for our computational analysis. These precomputed expression values represent the RNA-seq analysis initially performed by Mortazavi et al., using Illumina Genome Analyzer and expression values computed by using ERANGE package [21]. To remove low expressed genes the dataset was filtered by removing genes with RPKM value <1 in all the three tissues.

2.2. Fold Change Analysis. Fold change is often used in analysis of gene expression data of RNAseq experiments, for measuring change in the expression level of a gene and comparing the expression of genes between two sets of arrays, for example, case and control sets. Fold change analysis of genes commonly expressed between two different samples was done to compare the gene expression in glycogenes [22, 23]. Fold change was calculated as the ratio of brain versus liver, brain versus muscle, and liver versus muscle groups. Additionally, to further investigate the role of glycogenes in different tissues, the differentially expressed genes were manually verified in the UniProt database [24] to check whether they represent a glycogene or not. A gene was considered as a glycogene if it was annotated as a “glycoprotein” in the “Keywords” section of the UniProt database.

2.3. Functional Analysis. DAVID (<http://david.abcc.ncifcrf.gov/home.jsp>) functional annotation cluster analysis was performed on the list of differentially expressed glycogenes (DEGGs) with a fold change of ≥ 2 . Only those terms that reported a P value of ≤ 0.05 and count number ≥ 5 genes were selected for analysis. The gene ontology (GO) term biological process (BP) in DAVID was used to categorize enriched biological themes in the list of DEGGs. Pathway mapping was performed using the KEGG Automatic Annotation Server (KAAS) (<http://www.genome.jp/kegg/kaas/>) [25]. The amino acid sequences of these DEGGs were uploaded to the KAAS web server as an input using single-directional best hit (SBH) method to assign orthologs. KAAS offers functional annotation of genes in a genome via a BLAST similarity searches against a manually curated set of ortholog groups in the KEGG GENES database. KAAS assigned a KEGG orthology (KO) number to genes in the data sets, which were mapped to one of KEGG's reference pathways.

2.4. Network Analysis. The functional interactions between DEGGs were analyzed by GeneMANIA webserver [26]. The GO term biological process was used to create the interaction network between the DEGGs and 50 additional genes by using mouse as a source species. The relationship between the genes in the network includes coexpression, physical and genetic interactions, pathways, colocalization, protein domain similarity, and predicted interactions. The network was filtered by removing all the interactions where weights <0.1.

2.5. Identification of Hub Genes. Generally, the biological networks exhibit the scale-free property [27] where only a few nodes in the network have many connections that represent hubs in the network. Hub genes were identified by calculating the node degree distribution [28] by using the NetworkAnalyzer plugin of Cytoscape. A glycogene with the highest degree distribution was considered as a hub in the current network.

2.6. Community Analysis. The functional modules within the network were further detected by using the greedy community-structure detection algorithm, implemented in the Cytoscape plugin GLayer (<http://brainarray.mbni.med.umich.edu/sugang/glayer/>) [29]. GLayer offers Cytoscape users a diverse group of community structure algorithms and graph layout functions for network clustering and structured visualization [29]. To identify the overrepresented biological functions within each cluster, the detected clusters were subjected to functional enrichment analyses by DAVID functional analysis tool. For enrichment analysis, only those communities were analyzed which have at least 10 nodes.

3. Results

3.1. Identification of Differentially Expressed Glycogenes (DEGGs). We downloaded the precomputed RPKM values for mouse transcriptome that was mapped and quantified previously by RNA-Seq analysis [21]. This RNA-Seq data

TABLE 1: Gene expression summary.

(a) Total number of genes in each sample before and after filtering. Each tissue sample consists of 33598 genes that were filtered by removing all those genes that showed RPKM value < 1

Tissue	Total number of genes	Data filtering	Total number of genes after filtering
Brain	33598		15,237
Liver	33598		11,920
Muscle	33598		12,202

(b) Total number of differentially expressed genes between different tissues

Tissue	Genes commonly expressed between different tissues		
	Number of expressed genes	Number of differentially expressed genes (DEGs) (≥ 2 fold change)	Number of differentially expressed glycogenes (DEGGs) (≥ 2 fold change)
Brain versus liver	10,653	5228	579
Brain versus muscle	10,820	5283	501
Liver versus muscle	9,833	3840	442

consists of RNA from adult mouse brain, liver, and skeletal muscle tissues and represents 33598 genes in each tissue sample. The data was further filtered by excluding the genes with RPKM values <1 from the analysis. As a result, 15,237, 11,920, and 12,202 genes were identified from brain, liver, and muscle samples, respectively (Table 1(a)). Then three groups (brain versus liver, brain versus muscle, and liver versus muscle) were created and the number of genes that were common in each group were identified and summarized in Table 1(b). From the table, it can be observed that the number of common genes between brain versus liver is 10653, brain versus muscle is 10820, and liver versus muscle is 9833, respectively. These shared genes in each group were then used to calculate the fold change, which was defined as the ratio of RPKM values of brain versus liver, brain versus muscle, and liver versus muscle groups. In this study, the total fold change of ≥ 2 was considered to classify the differentially expressed genes (DEGs). Based on this definition, there are 5228 DEGs between brain versus liver, 5283 between brain versus muscle, and 3840 between liver versus muscle. The list of DEGs in each group was further filtered by retaining only those genes that were annotated as a “glycoprotein” in the UniProt database. Therefore, 579, 501, and 442 differentially expressed glycogenes (DEGGs) for brain versus liver (BvL579), brain versus muscle (BvM501), and liver versus muscle (LvM442) groups were selected for the final analysis (Table 1(b)). The complete workflow of the analysis is shown in Figure 1.

3.2. Functional Annotation and Pathway Analysis. To classify biological processes that are enriched in the BvL579, BvM5, Annotation Cluster (FAC) tool available in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (<http://david.abcc.ncifcrf.gov/home.jsp>). The GO

term “Biological Process” was used for annotations, and top 10 GO terms having statistically significant *P* values from the resulting functional analysis for each group are listed in Table 2. From this table, it can be seen that the GO terms that are enriched in BvL579 glycogenes represent functions necessary for response to wounding, inflammatory response, and blood coagulation (Table 2(a)). The enriched GO terms in BvM501 include functions related to various metabolic (glycoprotein, glycosaminoglycan, sphingolipid, etc.) processes (Table 2(b)). Similarly, functions related to protein maturation, inflammatory response, and complement activation are enriched in LvM442 samples (Table 2(c)).

In addition to DAVID functional analysis, we also identified the biological pathways of BvL579, BvM501, and LvM442 DEGGs annotated in the present study. FASTA formatted amino acid sequences of DEGGs in these three sets were fed into the KAAS for prediction of distinct pathways. A total of 210 pathways were predicted for BvL579, whereas 193 and 198 pathways were predicted for BvM501 and LvM442, respectively. The top 10 KEGG pathways for each of the three categories are shown in Table 3 and a complete list of all pathways is provided in Supplementary Table S1 (Supplementary Material is available online at <http://dx.doi.org/10.1155/2014/837365>). From these tables, it can be observed that the majority of the DEGGs were found to be associated with important biological processes, many being classified in signaling (e.g., PI3K-Akt and Rap1 signaling) pathways, cytokine-cytokine receptor interaction, or ECM-receptor interaction or being involved in adhesion related functions.

3.3. Interaction Network Construction and Identification of Hub Genes. DEGGs were mapped to the GeneMANIA to investigate how these genes interact with each other and

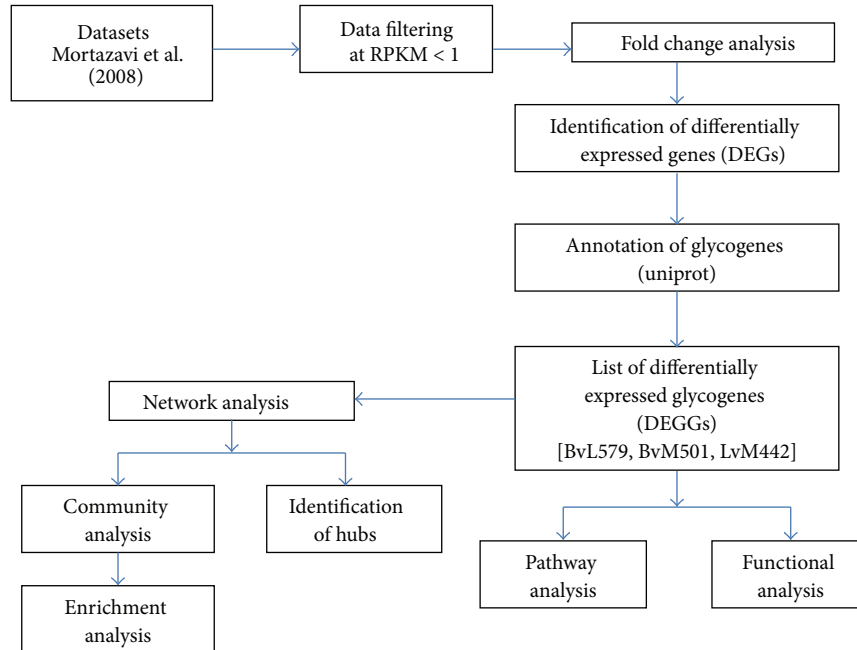


FIGURE 1: Flowchart depicting the overall methodology adopted in this study.

additional genes that are related to a set of query genes by using a very large set of functional interaction data [30]. The genes showing significant interactions with weights higher than 0.1 were selected only for the network analysis. By integrating these relationships, a network between DEGGs and additional related genes was constructed for all the three groups, namely, BvL579, BvM501, and LvM442 (Figure 2). A GeneMANIA network analysis for interactions among the glycogene products suggested enrichment of GO terms related to positive regulation of locomotion, cell motility, and migration for BvL579 (Table 4(a)), as well as LvM442 samples (Table 4(c)). On the contrary, the processes related to extracellular matrix were overrepresented in BvM501 samples (Table 4(b)).

The three interaction networks created from GeneMANIA were then exported to Cytoscape 2.8.2, a bioinformatics package for biological network visualization and data integration [31]. The initial network for BvL579 consists of 395 nodes and 1149 edges which were further filtered to 395 nodes and 1109 edges by removing duplicate edges (Figure 2(a)). In case of BvM501, the initial network consists of 352 nodes and 794 edges, whereas the filtered network has 352 nodes and 748 edges (Figure 2(b)). Similarly, the initial network for LvM442 network has 307 nodes and 695 edges, while the final network is represented by 307 nodes and 653 edges (Figure 2(c)). All the genes in the network are represented by circles and the interactions between them are represented as edges. Additionally, each query gene is shown in red whereas the additional related genes predicted by GeneMANIA are shown in cyan (Figure 2).

The node degree was then calculated for all the nodes in each network by using Network Analyzer plugin of

Cytoscape. The higher the node degree, the more important the gene was, and the gene was denoted as a hub gene. The genes with highest node degree include *PEMT* (BvL579) and *IGF2* (BvM501), having node degree of 29 and 14, respectively. Both *PEMT* and *IGF2* were predicted by GeneMANIA as related genes in their respective networks and do not represent the glycogenes. However, the genes with second highest node degrees of 23 and 13 are represented by glycogenes *HPXN* and *NID2* in BvL579 and BvM501 networks (Figures 3(a) and 3(b)). Additionally, *NID2* was also predicted by GeneMANIA as a related gene in BvM501 network because it was removed from the query list as it showed a fold change of less than 2 in our analysis. In LvM442 network, the top node degree genes are *STAT6*, and the glycogene *FLT1* (Figure 3(c)), both having the node degree of 13.

3.4. Community Analysis and Functional Annotation of the Detected Modules. In order to identify the biologically related genes in the three networks, we employed Fast Greedy community-structure identification algorithm, implemented in the Cytoscape plugin GLayer (<http://brainarray.mbni.med.umich.edu/sugang/glayer/>) [29] to identify functional modules, 31 (Figure 4(a)), 40 (Figure 4(b)), and 31 (Figure 4(c)) clusters were detected for BvL579, BvM501, and LvM442 networks, respectively. However, only those communities were selected for enrichment analyses which have at least 10 nodes. Therefore, based on this criterion, 8, 9, and 8 communities for BvL579, BvM501, and LvM442 networks were finally analyzed for overrepresentation of GO terms.

TABLE 2: Top 10 significantly enriched gene ontology (GO) terms detected by FAC in differentially expressed glycogenes between (a) BvL579, (b) BvM501, and (c) LvM442. Only those terms which reported a P value of ≤ 0.05 and count number ≥ 5 genes were selected for the analysis.

(a) BvL579		
GO Term	Total number of genes	P value
GO:0009611~response to wounding	53	$2.98E - 23$
GO:0002526~acute inflammatory response	23	$4.83E - 16$
GO:0006954~inflammatory response	35	$1.43E - 15$
GO:0051604~protein maturation	23	$2.51E - 14$
GO:0007596~blood coagulation	20	$5.40E - 14$
GO:0050817~coagulation	20	$5.40E - 14$
GO:0007599~hemostasis	20	$7.17E - 14$
GO:0006952~defense response	45	$7.57E - 13$
GO:0050878~regulation of body fluid levels	20	$5.84E - 12$
GO:0016485~protein processing	20	$7.21E - 12$
(b) BvM501		
GO Term	Total number of genes	P value
GO:0009100~glycoprotein metabolic process	30	$1.76E - 18$
GO:0009101~glycoprotein biosynthetic process	23	$8.1E - 14$
GO:0070085~glycosylation	17	$6.09E - 10$
GO:0006486~protein amino acid glycosylation	17	$6.09E - 10$
GO:0043413~biopolymer glycosylation	17	$6.09E - 10$
GO:0030203~glycosaminoglycan metabolic process	11	$3.1E - 08$
GO:0006665~sphingolipid metabolic process	12	$2.79E - 07$
GO:0006643~membrane lipid metabolic process	12	$3.9E - 07$
GO:0006022~aminoglycan metabolic process	11	$4.45E - 07$
GO:0051604~protein maturation	14	$6.1E - 07$
(c) LvM442		
GO Term	Total number of genes	P value
GO:0051604~protein maturation	20	$3.17E - 13$
GO:0016485~protein processing	19	$1.12E - 12$
GO:0009100~glycoprotein metabolic process	23	$1.63E - 12$
GO:0006954~inflammatory response	27	$6.69E - 12$
GO:0009611~response to wounding	33	$1.07E - 11$
GO:0002526~acute inflammatory response	17	$2.53E - 11$
GO:0009101~glycoprotein biosynthetic process	19	$1.14E - 10$
GO:0051605~protein maturation by peptide bond cleavage	15	$1.33E - 10$
GO:0006956~complement activation	11	$4.25E - 09$
GO:0002541~activation of plasma proteins involved in acute inflammatory response	11	$4.25E - 09$

To biologically categorize these clusters, DAVID functional analysis tool was used to classify the genes in each module and observed the enrichment of GO term “Biological Process” in all the selected modules. The enrichment analysis for the modules for three networks is described as follows.

BvL579. The top 10 statistically significant enriched GO terms for DEGs in all 8 clusters for BvL579 community analysis are summarized in Table 5(a). The most statistically significant GO terms that were enriched in cluster 1 were related to immune response, complement activation, and

protein maturation. Clusters 2, 3, and 7 show enrichment for GO terms response to wounding, blood coagulation, and hemostasis. Clusters 2 and 3 also show overrepresentation of terms such as blood vessel development and vasculature development. The GO terms of cluster 4 were mostly related to hair cycle, development of epidermis, and ectoderm. The GO functions of cluster 5 were most related to localization, motility and migration of cells, and regulation of cell proliferation. Cluster 6 was significantly enriched with functions related to homeostasis, transport, and metabolism of cholesterol, and lipid as well as sterol. Cluster 8 showed overrepresentation of functional terms

TABLE 3: Top 10 KEGG pathways for DEGs in each category. (a) BvL579, (b) BvM501, and (c) LvM442.

(a) BvL579	
Pathway name	Number of mapped Genes
04151 PI3K-Akt signaling pathway	34
04610 Complement and coagulation cascades	34
04142 Lysosome	31
04060 Cytokine-cytokine receptor interaction	23
05200 Pathways in cancer	21
04015 Rap1 signaling pathway	18
04510 Focal adhesion	18
05205 Proteoglycans in cancer	18
04514 Cell adhesion molecules CAMs	16
04976 Bile secretion	16
(b) BvM501	
Pathway name	Number of mapped Genes
04142 Lysosome	29
04151 PI3K-Akt signaling pathway	24
04060 Cytokine-cytokine receptor interaction	19
04514 Cell adhesion molecules CAMs	19
05200 Pathways in cancer	17
04141 Protein processing in endoplasmic reticulum	15
04510 Focal adhesion	15
05166 HTLV-I infection	13
04512 ECM-receptor interaction	12
04610 Complement and coagulation cascades	12
(c) LvM442	
Pathway name	Number of mapped Genes
04142 Lysosome	33
04151 PI3K-Akt signaling pathway	26
04060 Cytokine-cytokine receptor interaction	23
04141 Protein processing in endoplasmic reticulum	17
05205 Proteoglycans in cancer	17
04610 Complement and coagulation cascades	15
04514 Cell adhesion molecules CAMs	14
04510 Focal adhesion	14
04512 ECM-receptor interaction	13
04145 Phagosome	13

related to the metabolism of glycosphingolipid, glycolipid, sphingolipid, and ganglioside.

BvM501. The top 10 statistically significant enriched GO terms for DEGs in all 9 clusters for BvM501 community analysis are summarized in (Table 5(b)). Cluster 1 of this network shows overrepresentation of functions related to regulation of cell growth, protein maturation, and immune response. Cluster 2 is enriched in GO terms related to proteoglycan metabolic process as well as the development of blood vessel and vasculature. Cluster 3 is enriched in the functions related to the development of bone, cartilage, and skeletal system. The GO terms that are overrepresented in cluster 4 show enrichment for functions related to adhesion,

such as biological adhesion, cell adhesion, cell-substrate adhesion, and cell-matrix adhesion. The processes related to localization, motility, and migration of cells are also overrepresented in cluster 4. Cluster 5 showed overrepresentation of functional terms related to the metabolism of glycosphingolipid, glycolipid, sphingolipid, and ganglioside. Additionally, the functions related to the metabolism of lipid and oligosaccharides are also enriched in this group. In cluster 6, functions related to homeostasis, remodelling, and regulation of transport are overrepresented. Response to hypoxia and oxygen levels, apoptosis, and metabolic processes such as phospholipid and organophosphate metabolic process are enriched in cluster 7. Similar to cluster 4, the GO terms that are overrepresented in Cluster 9 show enrichment

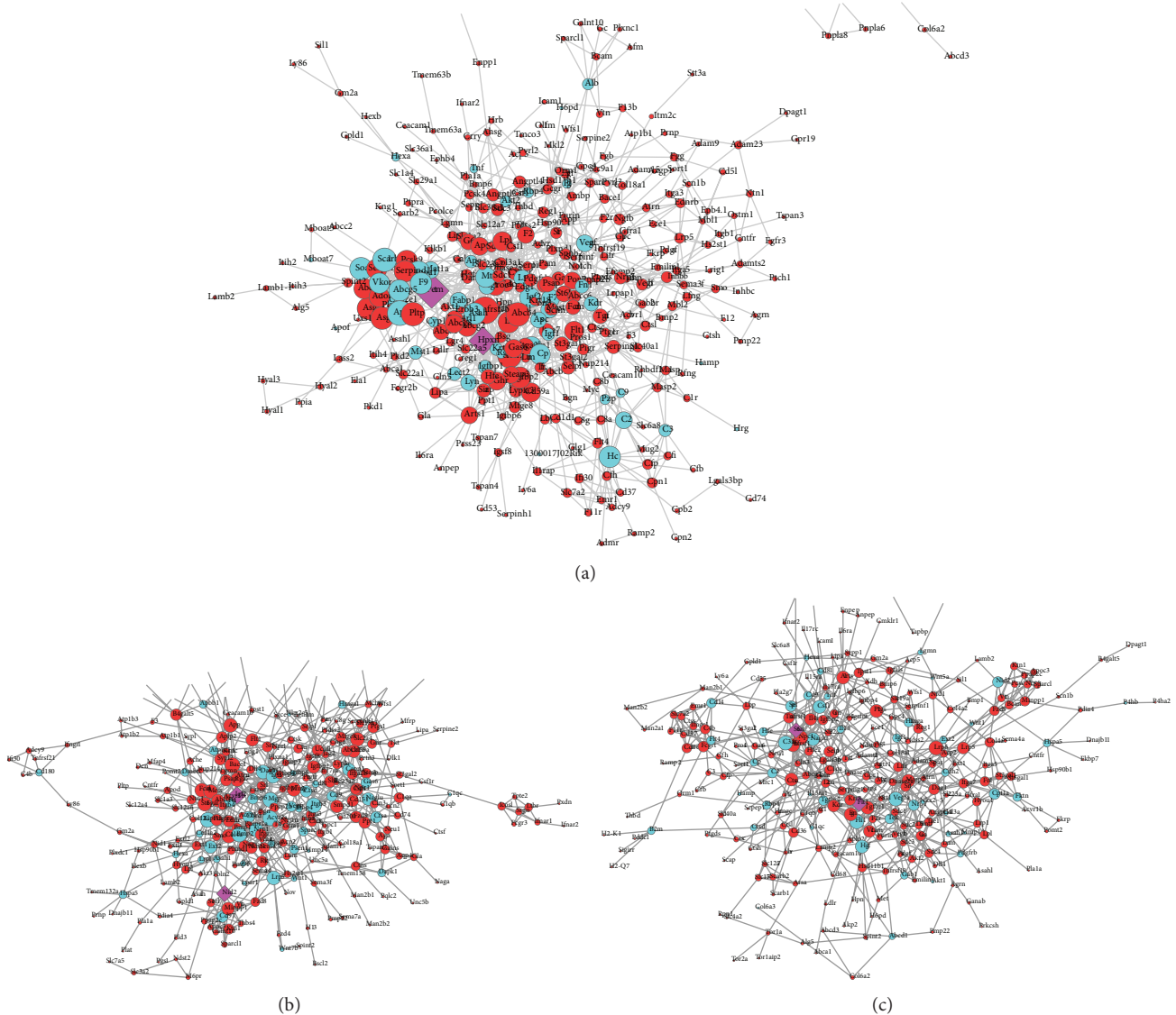


FIGURE 2: Interaction networks between DEGs and additional related genes: (a) BvL579, (b) BvM501, and (c) LvM442. In each of these networks, red indicates a DEGG whereas the GeneMANIA predicted genes are shown in cyan. The top two hubs are shown as purple diamonds. The size of the nodes represents the node degree.

for functions related to biological adhesion. Protein folding and metal ion transport are the other important enriched functions present in this cluster. Cluster 8 did not yield the enrichment of any statistically significant GO category.

LvM442. Table 5(c) summarizes the top 10 statistically significant enriched GO terms for DEGs in all 8 clusters for *LvM442* community analysis. The GO terms related to development and morphogenesis of blood vessels, vasculature, and tube are enriched in cluster 1. Cluster 2 shows enrichment of functions related to various types of responses, such as response to wounding, external stimulus, and defense as well as inflammatory response. Other enriched functional groups in this cluster are regulation of cell growth, proliferation, and

cell activation. One of the overrepresented functional groups in cluster 3 shows enrichment for functions related to biological adhesion, lipid catabolic process, and 4-hydroxyproline metabolic process. Cluster 4 represents the enrichment of functions related to protein folding and bone remodeling and resorption. Cluster 5 shows overrepresentation of GO terms such as regulation of T cell mediated cytotoxicity or immunity, regulation of leukocyte mediated cytotoxicity, and regulation of cell killing. Clusters 6 and 7 are enriched in functions related to immune or inflammatory response and complement activation. Finally, cluster 8 exhibited overrepresentation of signaling pathways such as hepatocyte growth factor receptor signaling and mesenchymal-epithelial cell signaling, in addition to processes concerning MAP kinase activity.

TABLE 4: 10 enriched gene ontology (GO) terms for DEGGs and additional related genes as reported by GeneMANIA for each category (a) BvL579, (b) BvM501, and (c) LvM442.

(a) BvL579		
GO id	Description	<i>q</i> -value
GO:0040017	positive regulation of locomotion	$1.83E - 23$
GO:2000147	positive regulation of cell motility	$1.83E - 23$
GO:0030335	positive regulation of cell migration	$3.40E - 23$
GO:0051272	positive regulation of cellular component movement	$3.40E - 23$
GO:0050817	coagulation	$1.94E - 20$
GO:0010876	lipid localization	$3.80E - 20$
GO:0007596	blood coagulation	$4.54E - 20$
GO:0007599	hemostasis	$7.56E - 20$
GO:0009897	external side of plasma membrane	$2.15E - 18$
GO:0000323	lytic vacuole	$1.05E - 17$
(b) BvM501		
GO id	Description	<i>q</i> -value
GO:0031012	extracellular matrix	$2.43E - 23$
GO:0005578	proteinaceous extracellular matrix	$1.37E - 17$
GO:0009100	glycoprotein metabolic process	$1.99E - 16$
GO:0000323	lytic vacuole	$2.87E - 16$
GO:0005764	lysosome	$2.87E - 16$
GO:0044420	extracellular matrix part	$3.18E - 15$
GO:0005604	basement membrane	$1.20E - 14$
GO:0005773	vacuole	$4.56E - 14$
GO:0030335	positive regulation of cell migration	$5.17E - 12$
GO:0005178	integrin binding	$5.17E - 12$
(c) LvM442		
GO id	Description	<i>q</i> -value
GO:0000323	lytic vacuole	$1.09E - 20$
GO:0005764	lysosome	$1.09E - 20$
GO:0051272	positive regulation of cellular component movement	$8.78E - 20$
GO:0030335	positive regulation of cell migration	$1.04E - 19$
GO:2000147	positive regulation of cell motility	$1.79E - 19$
GO:0040017	positive regulation of locomotion	$6.43E - 19$
GO:0005773	vacuole	$1.70E - 18$
GO:0019838	growth factor binding	$2.80E - 18$
GO:0009100	glycoprotein metabolic process	$2.11E - 15$
GO:0031012	extracellular matrix	$1.48E - 14$

4. Discussion

The current study offers the first thorough insight into the glycogene analysis of brain, muscle, and liver tissues from mouse RNA-Seq data. Understanding the structure and function of these glycogenes is essential for studying the development of various tissues as well as their functional roles. Most of the serum glycoproteins originates from liver suggesting that liver diseases associated with aberrant glycosylation can be reflected by the changes in serum glycoproteins [32]. Recently, the significance of glycomics of central nervous system (CNS) to identify potential glyco-biomarkers in neurological diseases [33] and alterations

in brain glycoproteins resulting from the aging process was shown [34]. Recent studies in the glycomics field also offered insights into the biological significance of the glycome in the pathogenesis of diseases in humans [32]. Additionally, many studies have highlighted the significance of glycoconjugates during skeletal muscle development [7, 23, 35].

With the evolution of bioinformatics during the past decades, molecular target discovery and targeted therapeutics have become a critical remedial treatment for diseases [36]. In this work, we screened DEGGs between brain, muscle, and liver tissues using RNA-Seq data and performed functional and pathway enrichment analysis using DAVID and KEGG tools. The most significant enriched GO terms

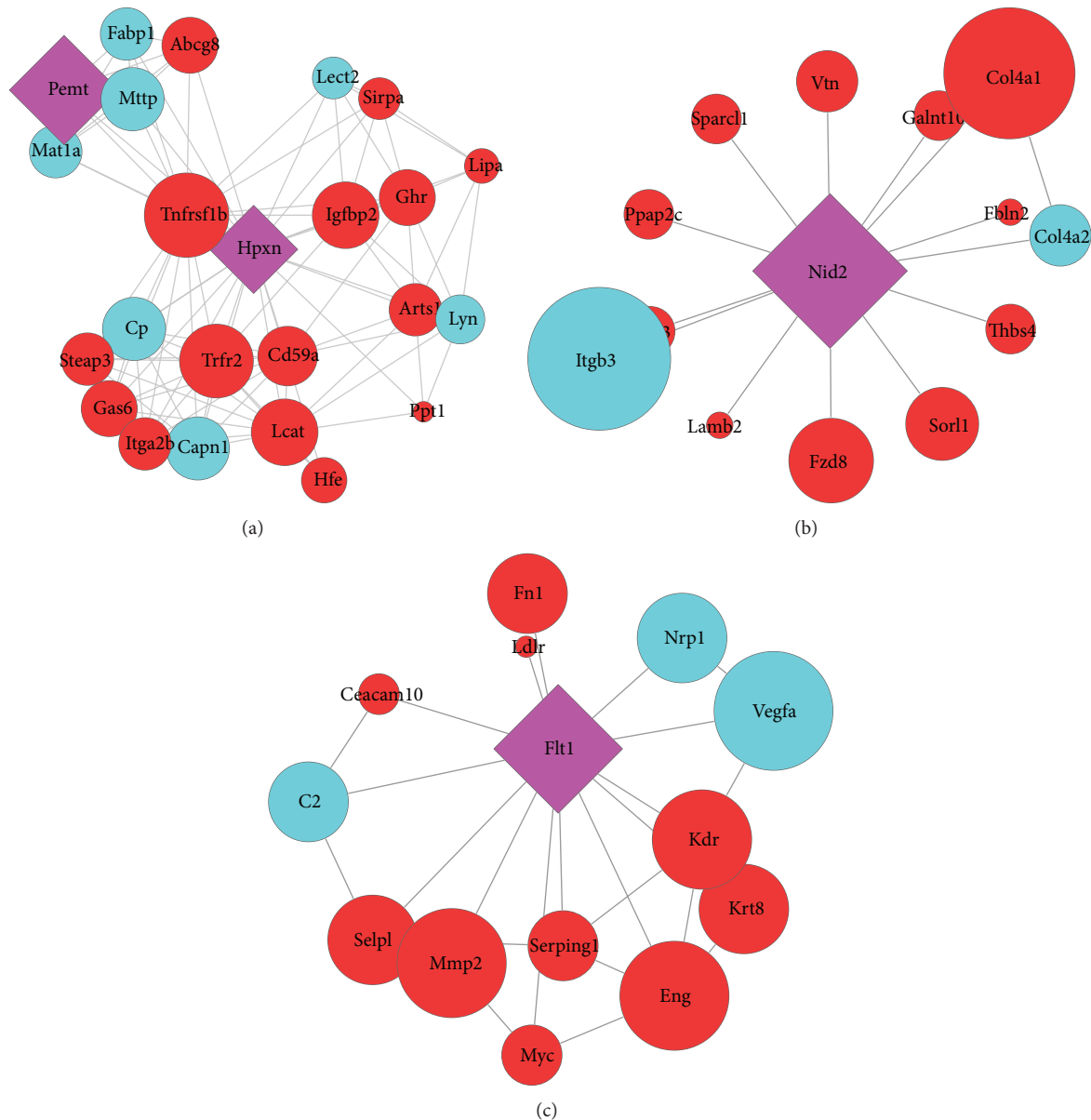
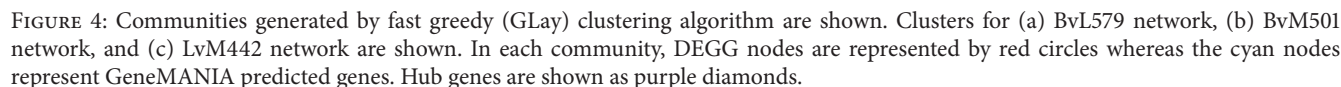


FIGURE 3: Top glyco-genes as hubs and their first neighbors. (a) *HPXN* gene as hubs in BvL579 interaction network. (b) *NID2* as hub gene in BvM501 interaction network. (c) *FLT1* gene as hub in LvM442 interaction network.

for BvL579 are processes related to response to wounding or inflammatory response. These processes are represented by genes that play a prominent roles in innate immunity such as *Mannan-binding lectin serine protease 1* (*MASP1*), *Mannan-binding lectin serine protease 2* (*MASP2*), *Mannose-binding protein A* (*MBL1*), and *Mannose-binding protein C* (*MBL2*) [37, 38]. BvM501 is enriched in functions related to glycoprotein metabolic processes and includes genes that code for various transferases such as N-acetylglucosamine-1-phosphotransferase subunit gamma (*GNPTG*) [39], Beta-galactoside alpha-2,6-sialyltransferase 1 (*ST6GAL1*) [40], and Beta-1,3-galactosyltransferase 6 (*B3GALT6*) [41]. Similarly, the top statistically significant enriched processes in LvM442

are related to protein maturation or processing and are represented by genes such as *Battenin* (*CLN3*) and *Methionine aminopeptidase 2* (*METAP2*). The *CLN3* protein is involved in the late endosomal/lysosomal membrane transport [42], whereas *METAP2* protein catalyzes the hydrolytic cleavage of the N-terminal methionine from newly synthesized polypeptides [43].

Additionally, three interaction networks of DEGs were constructed for BvL579, BvM501, and LvM442 samples, and node degrees of genes were calculated. *HPXN*, *NID2*, and *FLT1* were the glyco-genes with the highest degree in BvL579, BvM501, and LvM442 networks, respectively. However, these genes represent the second highest node degree in their



HPXN (~60 kDa glycoprotein) is mainly synthesized by liver cells [44], secreted to the plasma where it binds free heme or hemin and inhibits its role in free radical reactions [45]. HPXN is also reported to be expressed by the cells of immune systems, ganglionic and photoreceptor cells of

the retina, cells of the peripheral nervous system, and the mesangial cells of kidney [46–48]. Recently, it was shown that HPX protein is also present in various regions of mouse brain [49]. NID2 is homologous to another member of the nidogen family, NID1, and both are found in all basement membranes (BMs) [50]. Both nidogens have a similar distribution in various organs during development; however, in

TABLE 5: Enrichment of GO terms for BvL579, BvM501, and LvM442 cluster analysis.

(a) BvL579

Cluster number	Term	P value
1	GO:0002541~activation of plasma proteins involved in acute inflammatory response	5.02E – 22
	GO:0006956~complement activation	5.02E – 22
	GO:0002526~acute inflammatory response	2.26E – 20
	GO:0006959~humoral immune response	2.25E – 19
	GO:0051604~protein maturation	3.43E – 19
	GO:0051605~protein maturation by peptide bond cleavage	3.19E – 18
	GO:0002253~activation of immune response	3.20E – 18
	GO:0016485~protein processing	6.27E – 18
	GO:0006957~complement activation, alternative pathway	8.91E – 18
	GO:0006954~inflammatory response	1.44E – 17
2	GO:0051094~positive regulation of developmental process	5.33E – 05
	GO:0042592~homeostatic process	4.25E – 04
	GO:0009611~response to wounding	7.29E – 04
	GO:0001568~blood vessel development	9.88E – 04
	GO:0001944~vasculature development	0.001101764
	GO:0050817~coagulation	0.001394123
	GO:0007596~blood coagulation	0.001394123
	GO:0007599~hemostasis	0.001452434
	GO:0050878~regulation of body fluid levels	0.002774109
	GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	0.003100357
3	GO:0050817~coagulation	2.74E – 11
	GO:0007596~blood coagulation	2.74E – 11
	GO:0007599~hemostasis	3.08E – 11
	GO:0050878~regulation of body fluid levels	1.97E – 10
	GO:0042060~wound healing	1.26E – 09
	GO:0009611~response to wounding	5.75E – 08
	GO:0007167~enzyme linked receptor protein signaling pathway	9.33E – 08
	GO:0001568~blood vessel development	5.52E – 07
	GO:0001944~vasculature development	6.64E – 07
	GO:0001525~angiogenesis	4.71E – 05
4	GO:0022404~molting cycle process	2.16E – 04
	GO:0022405~hair cycle process	2.16E – 04
	GO:0001942~hair follicle development	2.16E – 04
	GO:0042633~hair cycle	2.42E – 04
	GO:0042303~molting cycle	2.42E – 04
	GO:0008544~epidermis development	0.003102493
	GO:0007398~ectoderm development	0.00369522
	GO:0050678~regulation of epithelial cell proliferation	0.009595347
	GO:0048732~gland development	0.01092817
	GO:0007173~epidermal growth factor receptor signaling pathway	0.025611634
5	GO:0007166~cell surface receptor linked signal transduction	1.34E – 05
	GO:0016477~cell migration	2.69E – 05
	GO:0042127~regulation of cell proliferation	3.44E – 05
	GO:0051674~localization of cell	4.88E – 05
	GO:0048870~cell motility	4.88E – 05
	GO:0006928~cell motion	6.35E – 05
	GO:0008284~positive regulation of cell proliferation	2.52E – 04
	GO:0007229~integrin-mediated signaling pathway	4.87E – 04
	GO:0040007~growth	6.74E – 04
	GO:0007507~heart development	0.001229375

(a) Continued.

Cluster number	Term	<i>P</i> value
6	GO:0055088~lipid homeostasis	2.16E – 16
	GO:0042632~cholesterol homeostasis	2.64E – 15
	GO:0055092~sterol homeostasis	2.64E – 15
	GO:0030301~cholesterol transport	3.97E – 13
	GO:0015918~sterol transport	3.97E – 13
	GO:0033344~cholesterol efflux	7.23E – 13
	GO:0006869~lipid transport	3.89E – 12
	GO:0008203~cholesterol metabolic process	4.74E – 12
	GO:0010876~lipid localization	8.01E – 12
	GO:0016125~sterol metabolic process	1.02E – 11
7	GO:0009611~response to wounding	3.88E – 07
	GO:0007596~blood coagulation	2.94E – 06
	GO:0050817~coagulation	2.94E – 06
	GO:0007599~hemostasis	3.11E – 06
	GO:0050878~regulation of body fluid levels	7.69E – 06
	GO:0042060~wound healing	1.91E – 05
	GO:0051818~disruption of cells of other organism during symbiotic interaction	3.08E – 05
	GO:0051851~modification by host of symbiont morphology or physiology	3.08E – 05
	GO:0051883~killing of cells in other organism during symbiotic interaction	3.08E – 05
	GO:0031640~killing of cells of another organism	4.30E – 05
8	GO:0006689~ganglioside catabolic process	5.11E – 06
	GO:0046479~glycosphingolipid catabolic process	1.79E – 05
	GO:0019377~glycolipid catabolic process	1.79E – 05
	GO:0001573~ganglioside metabolic process	2.38E – 05
	GO:0046466~membrane lipid catabolic process	1.15E – 04
	GO:0030149~sphingolipid catabolic process	1.15E – 04
	GO:0019915~lipid storage	1.29E – 04
	GO:0006687~glycosphingolipid metabolic process	2.73E – 04
	GO:0006664~glycolipid metabolic process	3.90E – 04
	GO:0050885~neuromuscular process controlling balance	4.98E – 04

(b) BvM501

Cluster number	Term	<i>P</i> value
1	GO:0040008~regulation of growth	3.73E – 08
	GO:0051604~protein maturation	7.27E – 08
	GO:0001558~regulation of cell growth	1.29E – 06
	GO:0016485~protein processing	2.32E – 05
	GO:0016044~membrane organization	7.38E – 05
	GO:0016064~immunoglobulin mediated immune response	9.61E – 05
	GO:0051605~protein maturation by peptide bond cleavage	1.09E – 04
	GO:0019724~B cell mediated immunity	1.09E – 04
	GO:0002684~positive regulation of immune system process	1.29E – 04
	GO:0006958~complement activation, classical pathway	2.00E – 04
2	GO:0009100~glycoprotein metabolic process	2.16E – 06
	GO:0050654~chondroitin sulfate proteoglycan metabolic process	9.34E – 06
	GO:0006029~proteoglycan metabolic process	8.82E – 05
	GO:0048754~branching morphogenesis of a tube	1.02E – 04
	GO:0019800~peptide cross-linking via chondroitin 4-sulfate glycosaminoglycan	1.89E – 04
	GO:0001763~morphogenesis of a branching structure	3.20E – 04
	GO:0001568~blood vessel development	4.29E – 04
	GO:0001944~vasculature development	4.79E – 04
	GO:0030204~chondroitin sulfate metabolic process	6.09E – 04
	GO:0006813~potassium ion transport	8.13E – 04

(b) Continued.

Cluster number	Term	P value
3	GO:0001501~skeletal system development	1.07E – 16
	GO:0030509~BMP signaling pathway	6.34E – 14
	GO:0001503~ossification	4.02E – 12
	GO:0060348~bone development	7.43E – 12
	GO:0030500~regulation of bone mineralization	6.08E – 11
	GO:0007178~transmembrane receptor protein serine/threonine kinase signaling pathway	7.75E – 11
	GO:0070167~regulation of biomineral formation	9.25E – 11
	GO:0007167~enzyme linked receptor protein signaling pathway	1.84E – 10
	GO:0007369~gastrulation	2.22E – 10
	GO:0051216~cartilage development	3.31E – 10
4	GO:0006928~cell motion	9.88E – 05
	GO:0016055~Wnt receptor signaling pathway	2.59E – 04
	GO:0016477~cell migration	4.37E – 04
	GO:0048870~cell motility	7.09E – 04
	GO:0051674~localization of cell	7.09E – 04
	GO:0007160~cell-matrix adhesion	0.001189285
	GO:0009100~glycoprotein metabolic process	0.001250564
	GO:0031589~cell-substrate adhesion	0.001570006
	GO:0007155~cell adhesion	0.005368157
5	GO:0022610~biological adhesion	0.005405426
	GO:0006689~ganglioside catabolic process	1.66E – 05
	GO:0019377~glycolipid catabolic process	5.78E – 05
	GO:0046479~glycosphingolipid catabolic process	5.78E – 05
	GO:0001573~ganglioside metabolic process	7.69E – 05
	GO:0009611~response to wounding	2.01E – 04
	GO:0046466~membrane lipid catabolic process	3.70E – 04
	GO:0030149~sphingolipid catabolic process	3.70E – 04
	GO:0019915~lipid storage	4.16E – 04
	GO:0006687~glycosphingolipid metabolic process	8.77E – 04
6	GO:0009311~oligosaccharide metabolic process	0.001247784
	GO:0042592~homeostatic process	6.03E – 06
	GO:0043270~positive regulation of ion transport	0.001725638
	GO:0051270~regulation of cell motion	0.002725255
	GO:0048771~tissue remodeling	0.003048492
	GO:0001894~tissue homeostasis	0.003842174
	GO:0051050~positive regulation of transport	0.004095371
	GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	0.004147057
	GO:0032869~cellular response to insulin stimulus	0.004461779
	GO:0042127~regulation of cell proliferation	0.004903945
7	GO:0010765~positive regulation of sodium ion transport	0.005901212
	GO:0001666~response to hypoxia	0.006023043
	GO:0070482~response to oxygen levels	0.006648281
	GO:0016042~lipid catabolic process	0.009864913
	GO:0006644~phospholipid metabolic process	0.011805904
	GO:0019637~organophosphate metabolic process	0.013020542
	GO:0006916~anti-apoptosis	0.013774773
	GO:0043066~negative regulation of apoptosis	0.037899056
	GO:0043069~negative regulation of programmed cell death	0.038883095
	GO:0060548~negative regulation of cell death	0.03908111

(b) Continued.

Cluster number	Term	P value
8	NA	NA
	GO:0007160~cell-matrix adhesion	0.001495186
	GO:0031589~cell-substrate adhesion	0.001809131
	GO:0051085~chaperone mediated protein folding requiring cofactor	0.006635187
	GO:0051084~"de novo" posttranslational protein folding	0.008618097
	GO:0007155~cell adhesion	0.00916406
9	GO:0022610~biological adhesion	0.009200307
	GO:0050982~detection of mechanical stimulus	0.010597488
	GO:0006458~"de novo" protein folding	0.010597488
	GO:0007259~JAK-STAT cascade	0.025656487
	GO:0030001~metal ion transport	0.036163834
	GO:0030001~metal ion transport	0.036163834

(c) LvM442

Cluster number	Term	P value
	GO:0001568~blood vessel development	1.06E – 18
	GO:0001944~vasculature development	1.64E – 18
	GO:0048514~blood vessel morphogenesis	8.45E – 16
	GO:0001525~angiogenesis	3.48E – 12
1	GO:0048754~branching morphogenesis of a tube	9.17E – 11
	GO:0007167~enzyme linked receptor protein signaling pathway	5.50E – 10
	GO:0035239~tube morphogenesis	1.14E – 09
	GO:0001763~morphogenesis of a branching structure	1.32E – 09
	GO:0035295~tube development	5.53E – 09
	GO:0001569~patterning of blood vessels	3.29E – 08
	GO:0040008~regulation of growth	4.27E – 07
	GO:0009611~response to wounding	4.18E – 06
	GO:0006954~inflammatory response	3.42E – 05
	GO:0032101~regulation of response to external stimulus	1.89E – 04
2	GO:0048585~negative regulation of response to stimulus	7.82E – 04
	GO:0006952~defense response	0.001399518
	GO:0001558~regulation of cell growth	0.002134607
	GO:0050867~positive regulation of cell activation	0.002784166
	GO:0032102~negative regulation of response to external stimulus	0.004518678
	GO:0008283~cell proliferation	0.004842417
	GO:0007160~cell-matrix adhesion	7.84E – 05
	GO:0031589~cell-substrate adhesion	1.14E – 04
	GO:0044242~cellular lipid catabolic process	9.80E – 04
	GO:0016042~lipid catabolic process	9.93E – 04
3	GO:0051346~negative regulation of hydrolase activity	0.006807864
	GO:0008218~bioluminescence	0.007742186
	GO:0018401~peptidyl-proline hydroxylation to 4-hydroxy-L-proline	0.007742186
	GO:0019471~4-hydroxyproline metabolic process	0.007742186
	GO:0009101~glycoprotein biosynthetic process	0.007780584
	GO:0007155~cell adhesion	0.00843311
	GO:0001501~skeletal system development	7.15E – 04
	GO:0006955~immune response	0.011579
	GO:0051085~chaperone mediated protein folding requiring cofactor	0.013230724
4	GO:0051084~"de novo" posttranslational protein folding	0.01716759
	GO:0045453~bone resorption	0.019783921
	GO:0006458~"de novo" protein folding	0.021089619
	GO:0046849~bone remodeling	0.027593503
	GO:0006952~defense response	0.045862312

(c) Continued.

Cluster number	Term	P value
5	GO:0001916~positive regulation of T cell mediated cytotoxicity	4.08E – 05
	GO:0001914~regulation of T cell mediated cytotoxicity	6.22E – 05
	GO:0002711~positive regulation of T cell mediated immunity	8.81E – 05
	GO:0002474~antigen processing and presentation of peptide antigen via MHC class I	1.53E – 04
	GO:0002709~regulation of T cell mediated immunity	1.72E – 04
	GO:0031343~positive regulation of cell killing	2.84E – 04
	GO:0001912~positive regulation of leukocyte mediated cytotoxicity	2.84E – 04
	GO:0001910~regulation of leukocyte mediated cytotoxicity	4.86E – 04
	GO:0031341~regulation of cell killing	4.86E – 04
	GO:0002821~positive regulation of adaptive immune response	6.25E – 04
6	GO:0002526~acute inflammatory response	4.67E – 14
	GO:0045087~innate immune response	4.65E – 13
	GO:0002684~positive regulation of immune system process	1.53E – 12
	GO:0050778~positive regulation of immune response	3.29E – 12
	GO:0006954~inflammatory response	3.41E – 12
	GO:0002541~activation of plasma proteins involved in acute inflammatory response	4.05E – 12
	GO:0006956~complement activation	4.05E – 12
	GO:0048584~positive regulation of response to stimulus	4.13E – 11
	GO:0006952~defense response	4.75E – 11
	GO:0006959~humoral immune response	5.29E – 11
7	GO:0009611~response to wounding	2.97E – 06
	GO:0006958~complement activation, classical pathway	7.07E – 06
	GO:0002526~acute inflammatory response	7.33E – 06
	GO:0002455~humoral immune response mediated by circulating immunoglobulin	8.69E – 06
	GO:0006956~complement activation	2.20E – 05
	GO:0002541~activation of plasma proteins involved in acute inflammatory response	2.36E – 05
	GO:0016064~immunoglobulin mediated immune response	4.70E – 05
	GO:0019724~B cell mediated immunity	5.25E – 05
	GO:0002449~lymphocyte mediated immunity	1.02E – 04
	GO:0002250~adaptive immune response	1.36E – 04
8	GO:0051450~myoblast proliferation	0.002206368
	GO:0048012~hepatocyte growth factor receptor signaling pathway	0.002206368
	GO:0060665~regulation of branching involved in salivary gland morphogenesis by mesenchymal-epithelial signaling	0.003674845
	GO:0060638~mesenchymal-epithelial cell signaling	0.004408354
	GO:0060693~regulation of branching involved in salivary gland morphogenesis	0.006605966
	GO:0060688~regulation of morphogenesis of a branching structure	0.018252999
	GO:0001889~liver development	0.031208893
	GO:0000187~activation of MAPK activity	0.036205594
	GO:0043406~positive regulation of MAP kinase activity	0.042595839

adult tissues nidogen-2 distribution becomes more confined [51–53]. Additionally, these nidogens show a broad range of interacting partners including other BM proteins such as laminin, collagen IV, and perlecan [51, 53–55]. They are involved in various functions including the regulation of cell attachment [56], neutrophil chemotaxis [57], trophoblast outgrowth [58], angiogenesis [59], osteoblast, and myogenic differentiation [60, 61]. Vascular endothelial growth factors (VEGFs) constitute a family of six polypeptides (VEGF-A, -B, -C, -D, -E, and PlGF) that regulate blood and lymphatic vessel development [62]. VEGF signaling occurs by binding to various cellular receptors such as VEGFR1 (FLT1), VEGFR2 (FLK1), and VEGFR3 (FLT4) [63, 64] and

neuropilin-1 [65] and -2 (NRP1 and NRP2) [66] and heparan sulfate proteoglycans (HSPG) [67]. FLT1 and FLK1 are closely related receptor tyrosine kinases and both share common and specific ligands [68]. FLT1 has weaker kinase activity than FLK1 [68]; however, FLT1 is essential for normal development and angiogenesis as reported in previous FLT1 null mutant mice studies [69–71].

Recognizing the structure and function of biological networks is indispensable for the investigation of biological processes. In this work, we identified 8, 9, and 8 functional modules or communities for BvL579, BvM501, and LvM442 networks using fast greedy algorithm implemented as GLAY [29], plugin for cytoscape. Furthermore, the function of

each module in all the three networks was explored by using functional annotation tool DAVID. Our analysis shows that the modules that are common in BvL579 and LvM442 networks show enrichment for processes related to immune and inflammatory response and response to wounding. Liver is known to play an important role in innate immunity, an important primary line of defense against infection [72]. Additionally, Kupffer cells in the liver are one of the earliest to be affected by bacterial or sterile insults and add to the inflammatory response [73]. This sterile inflammation can be responsible for liver injury; it may also play a role in liver repair [72]. In case of the modules of BvM501 and LvM442 networks, the enrichment of processes related to cell adhesion, proliferation, and regulation of cell growth is observed. The expression of myogenic regulatory factors (MRFs) describes the different stages of skeletal muscle development that includes myoblast proliferation, cell-cycle exit, cell fusion, and the maturation of myotubes to form myofibers [23]. Cell adhesion molecules such as cadherins are glycoproteins that mediate homotypic cell-cell adhesion through their extracellular domain [74], and this cadherin-dependent adhesion is necessary for diverse types of cellular functions [75]. Finally, glycosphingolipid or glycolipid metabolic processes show overrepresentation in modules of BvL579 and BvM501 networks. Glycosphingolipids are universally expressed in all vertebrate cells as well as body fluids, but they are richer in the nervous system [76]. Previous studies have established the role of glycosphingolipids in the development of the brain among various species [77, 78].

5. Conclusion

Overall, in this work we have used a systems biology approach to identify the DEGs, their functional enrichment, and identified potential hub genes by constructing three glycogene interaction networks. We also identified functional modules in these networks and indicate the significance of immune system related processes in liver, and glycosphingolipid metabolic processes in the development of brain. Similarly, we observe that the processes such as cell proliferation, adhesion, and growth are important for muscle development. The findings deduced from the current work are in consistence with the previous studies. Experimental validation will be required to confirm the predictions made in this study that will establish the role of predicted hubs as well as enriched functional processes in these tissues.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Authors would like to acknowledge the support from the Biomedical Informatics Center of ICMR, PGIMER (Chandigarh, India); School of Chemistry and Biochemistry, Thapar University, (Patiala, India); and School of Biotechnology,

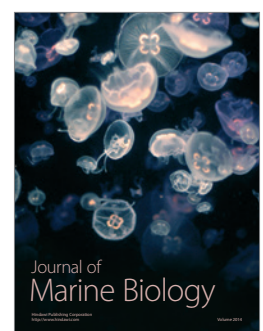
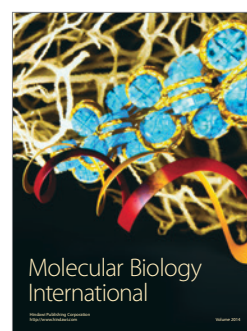
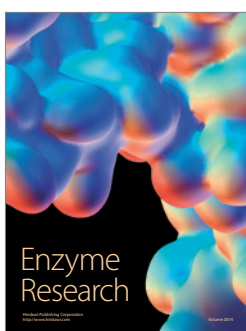
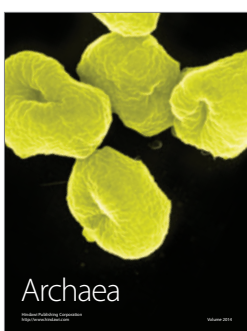
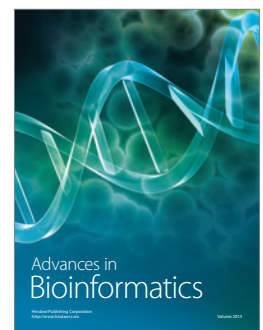
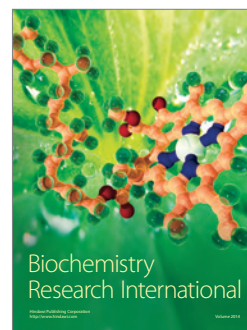
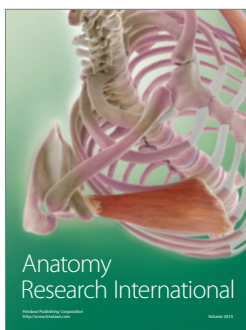
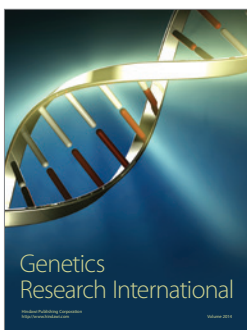
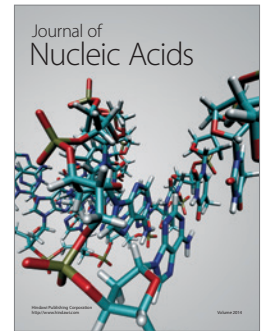
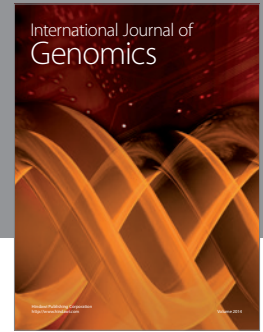
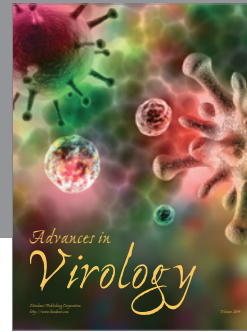
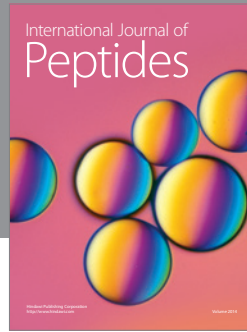
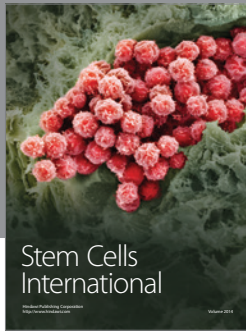
Yeungnam University (Gyeongsan, South Korea) for providing lab facilities to conduct this research.

References

- [1] J. Salzman, H. Jiang, and W. H. Wong, "Statistical modeling of RNA-Seq data," *Statistical Science*, vol. 26, no. 1, pp. 62–83, 2011.
- [2] F. Mantelli, L. Schaffer, R. Dana, S. R. Head, and P. Argüeso, "Glycogene expression in conjunctiva of patients with dry eye: downregulation of Notch signaling," *PLoS ONE*, vol. 50, no. 6, pp. 2666–2672, 2010.
- [3] A. Ishii, S. Hitoshi, I. Fujimoto et al., "Developmental changes in the expression of glycogenes and the content of N-glycans in the mouse cerebral cortex," *Glycobiology*, vol. 17, no. 3, pp. 261–276, 2007.
- [4] Y. Nakayama, N. Nakamura, D. Tsuji, K. Itoh, and A. Kurosaka, "Genetic diseases associated with protein glycosylation disorders in mammals," in *Genetic Disorders*, InTechOpen, 2013.
- [5] <http://www.stellic.com/cn27/pg122.html>.
- [6] N. Taniguchi, A. Suzuki, Y. Ito et al., *Experimental Glycoscience: Glycobiology*, Springer, New York, NY, USA, 2009.
- [7] M. Janot, A. Audfray, C. Lorient, A. Germot, A. Maftah, and F. Dupuy, "Glycogenome expression dynamics during mouse C2C12 myoblast differentiation suggests a sequential reorganization of membrane glycoconjugates," *BMC Genomics*, vol. 10, no. 483, 2009.
- [8] Z. Zhang, J. Sun, L. Hao, C. Liu, H. Ma, and L. Jia, "Modification of glycosylation mediates the invasive properties of murine hepatocarcinoma cell lines to lymph nodes," *PLoS ONE*, vol. 20, no. 8, Article ID e65218, 2013.
- [9] S. Ngamukote, M. Yanagisawa, T. Ariga, S. Ando, and R. K. Yu, "Developmental changes of glycosphingolipids and expression of glycogenes in mouse brains," *Journal of Neurochemistry*, vol. 103, no. 6, pp. 2327–2341, 2007.
- [10] A. Urita, T. Matsushashi, T. Onodera et al., "Alterations of high-mannose type N-glycosylation in human and mouse osteoarthritis cartilage," *Arthritis and Rheumatism*, vol. 63, no. 11, pp. 3428–3438, 2011.
- [11] C. Saravanan, Z. Cao, S. R. Head, and N. Panjwani, "Analysis of differential expression of glycosyltransferases in healing corneas by glycogene microarrays," *Glycobiology*, vol. 20, no. 1, pp. 13–23, 2010.
- [12] C. Saravanan, Z. Cao, S. R. Head, and N. Panjwani, "Detection of differentially expressed wound-healing-related glycogenes in galectin-3-deficient mice," *Investigative Ophthalmology and Visual Science*, vol. 50, no. 12, pp. 5690–5696, 2009.
- [13] K. A. Maupin, A. Sinha, E. Eugster et al., "Glycogene expression alterations associated with pancreatic cancer Epithelial-mesenchymal transition in complementary model systems," *PLoS ONE*, vol. 5, no. 9, Article ID e13002, 2010.
- [14] H. Ma, X. Miao, Q. Ma, W. Zheng, H. Zhou, and L. Jia, "Functional roles of glycogene and N-glycan in multidrug resistance of human breast cancer cells," *IUBMB Life*, vol. 65, no. 5, pp. 409–422, 2013.
- [15] R. Guo, L. Cheng, Y. Zhao et al., "Glycogenes mediate the invasive properties and chemosensitivity of human hepatocarcinoma cells," *The International Journal of Biochemistry and Cell Biology*, vol. 45, no. 2, pp. 347–358, 2013.
- [16] S. Diskin, J. Kumar, Z. Cao et al., "Detection of differentially expressed glycogenes in trabecular meshwork of eyes with

- primary open-angle glaucoma," *Investigative Ophthalmology and Visual Science*, vol. 47, no. 4, pp. 1491–1499, 2006.
- [17] T. Inoue, H. Sugiyama, M. Kitagawa et al., "Abnormalities of glycogenes in tonsillar lymphocytes in IgA nephropathy," *Advances in Otorhinolaryngology*, vol. 72, pp. 71–74, 2011.
 - [18] T. Inoue, H. Sugiyama, Y. Hiki et al., "Differential expression of glycogenes in tonsillar B lymphocytes in association with proteinuria and renal dysfunction in IgA nephropathy," *Clinical Immunology*, vol. 136, no. 3, pp. 447–455, 2010.
 - [19] R. A. Kroes, G. Dawson, and J. R. Moskal, "Focused microarray analysis of glyco-gene expression in human glioblastomas," *Journal of neurochemistry*, vol. 103, supplement 1, pp. 14–24, 2007.
 - [20] Z. Tan, W. Lu, X. Li et al., "Altered N-glycan expression profile in epithelial-to-mesenchymal transition of NMuMG cells revealed by an integrated strategy using mass spectrometry and glycogene and lectin microarray analysis," *Journal of Proteome Research*, vol. 13, no. 6, pp. 2783–2795, 2014.
 - [21] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.
 - [22] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.
 - [23] E. J. Lee, A. Malik, S. Pokharel et al., "Identification of genes differentially expressed in *myogenin* knock-down bovine muscle satellite cells during differentiation through RNA sequencing analysis," *PLoS ONE*, vol. 9, no. 3, Article ID e92447, 2014.
 - [24] <http://www.uniprot.org/>.
 - [25] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. 2, supplement, pp. W182–W185, 2007.
 - [26] J. Montojo, K. Zuberi, H. Rodriguez et al., "GeneMANIA cytoscape plugin: fast gene function predictions on the desktop," *Bioinformatics*, vol. 26, no. 22, Article ID btq562, pp. 2927–2928, 2010.
 - [27] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005.
 - [28] X. R. Wu, Y. Zhu, and Y. Li, "Analyzing protein interaction networks via random graph model," *International Journal of Information Technology*, vol. 11, pp. 125–132, 2005.
 - [29] G. Su, A. Kuchinsky, J. H. Morris, D. J. States, and F. Meng, "GLay: community structure analysis of biological networks," *Bioinformatics*, vol. 26, no. 24, pp. 3135–3137, 2010.
 - [30] D. W. Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, pp. W214–W220, 2010.
 - [31] M. E. Smoot, K. Ono, J. Ruscheinski, P. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.
 - [32] B. Adamczyk, T. Tharmalingam, and P. M. Rudd, "Glycans as cancer biomarkers," *Biochimica et Biophysica Acta—General Subjects*, vol. 1820, no. 9, pp. 1347–1353, 2012.
 - [33] R. Barone, L. Sturiale, A. Palmigiano, M. Zappia, and D. Garozzo, "Glycomics of pediatric and adulthood diseases of the central nervous system," *Journal of Proteomics*, vol. 75, no. 17, pp. 5123–5139, 2012.
 - [34] Y. Sato and T. Endo, "Alteration of brain glycoproteins during aging," *Geriatrics and Gerontology International*, vol. 10, pp. S32–S40, 2010.
 - [35] D. D. W. Cornelison, S. A. Wilcox-Adelman, P. F. Goetinck, H. Rauvala, A. C. Rapraeger, and B. B. Olwin, "Essential and separable roles for Syndecan-3 and Syndecan-4 in skeletal muscle development and regeneration," *Genes and Development*, vol. 18, no. 18, pp. 2231–2236, 2004.
 - [36] P. Dinh, C. Sotiriou, and M. J. Piccart, "The evolution of treatment strategies: aiming at the target," *Breast*, vol. 16, supplement 2, pp. 10–16, 2007.
 - [37] A. J. Nauta, N. Raashou-Jensen, A. Roos et al., "Mannose-binding lectin engagement with late apoptotic and necrotic cells," *European Journal of Immunology*, vol. 33, no. 10, pp. 2853–2863, 2003.
 - [38] M. Møller-Kristensen, S. Thiel, A. Sjöholm, M. Matsushita, and J. C. Jensenius, "Cooperation between MASP-1 and MASP-2 in the generation of C3 convertase through the MBL pathway," *International Immunology*, vol. 19, no. 2, pp. 141–149, 2007.
 - [39] Y. Qian, I. Lee, W. Lee et al., "Functions of the α , β , and γ subunits of UDP-GlcNAc:Lysoosomal enzyme N-acetylglucosamine-1-phosphotransferase," *The Journal of Biological Chemistry*, vol. 285, no. 5, pp. 3360–3370, 2010.
 - [40] B. Kuhn, J. Benz, M. Greif, A. M. Engel, H. Sobek, and M. G. Rudolph, "The structure of human α -2, 6-sialyltransferase reveals the binding mode of complex glycans," *Acta Crystallographica D*, vol. 69, pp. 1826–1838, 2013.
 - [41] X. Bai, D. Zhou, J. R. Brown, B. E. Crawford, T. Hennen, and J. D. Esko, "Biosynthesis of the linkage region of glycosaminoglycans: cloning and activity of galactosyltransferase II, the sixth member of the β 1,3-galactosyltransferase family (β 3GalT6)," *Journal of Biological Chemistry*, vol. 276, no. 51, pp. 48189–48195, 2001.
 - [42] K. Uusi-Rauva, A. Kytälä, R. Van Der Kant et al., "Neuronal ceroid lipofuscinosis protein CLN3 interacts with motor proteins and modifies location of late endosomal compartments," *Cellular and Molecular Life Sciences*, vol. 69, no. 12, pp. 2075–2089, 2012.
 - [43] Q. Xiao, F. Zhang, B. A. Nacev, J. O. Liu, and D. Pei, "Protein N-terminal processing: substrate specificity of *Escherichia coli* and human methionine aminopeptidases," *Biochemistry*, vol. 49, no. 26, pp. 5588–5599, 2010.
 - [44] J. Chen-Roetling, W. Liu, and R. F. Regan, "Hemopexin decreases hemin accumulation and catabolism by neural cells," *Neurochemistry International*, vol. 60, no. 5, pp. 488–494, 2012.
 - [45] E. Tolosano, S. Fagoonee, N. Morello, F. Vinchi, and V. Fiorito, "Heme scavenging and the other facets of hemopexin," *Antioxidants & Redox Signaling*, vol. 12, no. 2, pp. 305–320, 2010.
 - [46] L. Camborieux, N. Bertrand, and J. P. Swerts, "Changes in expression and localization of hemopexin and its transcripts in injured nervous system: a comparison of central and peripheral tissues," *Neuroscience*, vol. 82, no. 4, pp. 1039–1052, 1997.
 - [47] R. C. Hunt, D. M. Hunt, N. Gaur, and A. Smith, "Hemopexin in the human retina: protection of the retina against heme-mediated toxicity," *Journal of Cellular Physiology*, vol. 168, no. 1, pp. 71–80, 1996.
 - [48] J. J. Kapojos, A. van Den Berg, H. van Goor et al., "Production of hemopexin by TNF- α stimulated human mesangial cells," *Kidney International*, vol. 63, no. 5, pp. 1681–1686, 2003.
 - [49] R.-C. Li, S. Saleem, G. Zhen et al., "Hemehemopexin complex attenuates neuronal cell death and stroke damage," *Journal of*

- Cerebral Blood Flow & Metabolism*, vol. 29, no. 5, pp. 953–964, 2009.
- [50] M. Bechtel, M. V. Keller, W. Bloch et al., “Different domains in nidogen-1 and nidogen-2 drive basement membrane formation in skin organotypic cocultures,” *The FASEB Journal*, vol. 26, no. 9, pp. 3637–3648, 2012.
- [51] E. Kohfeldt, T. Sasaki, W. Göhring, and R. Timpl, “Nidogen-2: a new basement membrane protein with diverse binding properties,” *Journal of Molecular Biology*, vol. 282, no. 1, pp. 99–109, 1998.
- [52] M. Murshed, N. Smyth, N. Miosge et al., “The absence of nidogen 1 does not affect murine basement membrane formation,” *Molecular and Cellular Biology*, vol. 20, no. 18, pp. 7007–7012, 2000.
- [53] K. Salmivirta, J. F. Talts, M. Olsson, T. Sasaki, R. Timpl, and P. Ekblom, “Binding of mouse nidogen-2 to basement membrane components and cells and its expression in embryonic and adult tissues suggest complementary functions of the two nidogens,” *Experimental Cell Research*, vol. 279, no. 2, pp. 188–201, 2002.
- [54] J. W. Fox, U. Mayer, R. Nischt et al., “Recombinant nidogen consists of three globular domains and mediates binding of laminin to collagen type IV,” *EMBO Journal*, vol. 10, no. 11, pp. 3137–3146, 1991.
- [55] M. S. P. Ho, K. Böse, S. Mokkapat, R. Nischt, and N. Smyth, “Nidogens-extracellular matrix linker molecules,” *Microscopy Research and Technique*, vol. 71, no. 5, pp. 387–395, 2008.
- [56] S. Chakravarti, M. F. Tam, and A. E. Chung, “The basement membrane glycoprotein entactin promotes cell attachment and binds calcium ions,” *The Journal of Biological Chemistry*, vol. 265, no. 18, pp. 10597–10603, 1990.
- [57] R. M. Senior, H. D. Gresham, G. L. Griffin, E. J. Brown, and A. E. Chung, “Entactin stimulates neutrophil adhesion and chemotaxis through interactions between its Arg-Gly-Asp (RGD) domain and the leukocyte response integrin,” *Journal of Clinical Investigation*, vol. 90, no. 6, pp. 2251–2257, 1992.
- [58] F. D. Yelian, N. A. Edgeworth, L.-J. Dong, A. E. Chung, and D. R. Armant, “Recombinant entactin promotes mouse primary trophoblast cell adhesion and migration through the Arg-Gly-Asp (RGD) recognition sequence,” *Journal of Cell Biology*, vol. 121, no. 4, pp. 923–929, 1993.
- [59] R. F. Nicosia, S. V. Nicosia, and M. Smith, “Vascular endothelial growth factor, platelet-derived growth factor, and insulin-like growth factor-1 promote rat aortic angiogenesis in vitro,” *The American Journal of Pathology*, vol. 145, no. 5, pp. 1023–1029, 1994.
- [60] N. Kimura, T. Toyoshima, T. Kojima, and M. Shimane, “Entactin-2: a new member of basement membrane protein with high homology to entactin/nidogen,” *Experimental Cell Research*, vol. 241, no. 1, pp. 36–45, 1998.
- [61] R. Neu, S. Adams, and B. Munz, “Differential expression of entactin-1/nidogen-1 and entactin-2/nidogen-2 in myogenic differentiation,” *Differentiation*, vol. 74, no. 9–10, pp. 573–582, 2006.
- [62] F. S. Grünwald, A. E. Prota, A. Giese, and K. Ballmer-Hofer, “Structure-function analysis of VEGF receptor activation and the role of coreceptors in angiogenic signaling,” *Biochimica et Biophysica Acta: Proteins and Proteomics*, vol. 1804, no. 3, pp. 567–580, 2010.
- [63] B. I. Terman, M. E. Carrion, E. Kovacs, B. A. Rasmussen, R. L. Eddy, and T. B. Shows, “Identification of a new endothelial cell growth factor receptor tyrosine kinase,” *Oncogene*, vol. 6, no. 9, pp. 1677–1683, 1991.
- [64] M. Shibuya, S. Yamaguchi, A. Yamane et al., “Nucleotide sequence and expression of a novel human receptor-type tyrosine kinase gene (flt) closely related to the fms family,” *Oncogene*, vol. 5, no. 4, pp. 519–524, 1990.
- [65] K. Pajusola, O. Aprelikova, J. Korhonen et al., “FLT4 receptor tyrosine kinase contains seven immunoglobulin-like loops and is expressed in multiple human tissues and cell lines,” *Cancer Research*, vol. 52, no. 20, pp. 5738–5743, 1992.
- [66] G. Neufeld, O. Kessler, and Y. Herzog, “The interaction of neuropilin-1 and neuropilin-2 with tyrosine-kinase receptors for VEGF,” *Advances in Experimental Medicine and Biology*, vol. 5, no. 15, pp. 81–90, 2002.
- [67] S. Tessler, P. Rockwell, D. Hicklin et al., “Heparin modulates the interaction of VEGF165 with soluble and cell associated flk-1 receptors,” *Journal of Biological Chemistry*, vol. 269, no. 17, pp. 12456–12461, 1994.
- [68] J. Nishi, T. Minamino, H. Miyauchi et al., “Vascular endothelial growth factor receptor-1 regulates postnatal angiogenesis through inhibition of the excessive activation of Akt,” *Circulation Research*, vol. 103, no. 3, pp. 261–268, 2008.
- [69] G.-H. Fong, J. Rossant, M. Gertsenstein, and M. L. Breitman, “Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium,” *Nature*, vol. 376, no. 6535, pp. 66–70, 1995.
- [70] G. Fong, L. Zhang, D. Bryce, and J. Peng, “Increased heman-gioblast commitment, not vascular disorganization, is the primary defect in flt-1 knock-out mice,” *Development*, vol. 126, no. 13, pp. 3015–3025, 1999.
- [71] J. B. Kearney, C. A. Ambler, K.-A. Monaco, N. Johnson, R. G. Rapoport, and V. L. Bautch, “Vascular endothelial growth factor receptor Flt-1 negatively regulates developmental blood vessel formation by modulating endothelial cell division,” *Blood*, vol. 99, no. 7, pp. 2397–2407, 2002.
- [72] B. Gao, W. Jeong, and Z. Tian, “Liver: an organ with predominant innate immunity,” *Hepatology*, vol. 47, no. 2, pp. 729–736, 2008.
- [73] V. Bieghs and C. Trautwein, “The innate immune response during liver inflammation and metabolic disease,” *Trends in Immunology*, vol. 34, no. 9, pp. 446–452, 2013.
- [74] S. Charrasse, M. Causeret, F. Comunale, A. Bonet-Kerrache, and C. Gauthier-Rouvière, “Rho GTPases and cadherin-based cell adhesion in skeletal muscle development,” *Journal of Muscle Research and Cell Motility*, vol. 24, no. 4–6, pp. 309–313, 2003.
- [75] K. A. Knudsen, L. Myers, and S. A. McElwee, “A role for the Ca²⁺-dependent adhesion molecule, N-Cadherin, in myoblast interaction during myogenesis,” *Experimental Cell Research*, vol. 188, no. 2, pp. 175–184, 1990.
- [76] R. K. Yu, Y. Nakatani, and M. Yanagisawa, “The role of glycosphingolipid metabolism in the developing brain,” *Journal of Lipid Research*, vol. 50, pp. S440–S445, 2009.
- [77] L. Svennerholm, K. Bostrom, P. Fredman, J. E. Månsson, B. Rosengren, and B. M. Rynmark, “Human brain gangliosides: developmental changes from early fetal stage to advanced age,” *Biochimica et Biophysica Acta*, vol. 1005, no. 2, pp. 109–117, 1989.
- [78] R. K. Yu, L. J. Macala, T. Taki, H. M. Weinfeld, and F. S. Yu, “Developmental changes in ganglioside composition and synthesis in embryonic rat brain,” *Journal of Neurochemistry*, vol. 50, no. 6, pp. 1825–1829, 1988.





Research paper

Identification of hub glycogenes and their nsSNP analysis from mouse RNA-Seq data



Ahmad Firoz^{a,b,*}, Adeel Malik^{c,**}, Sanjay Kumar Singh^b, Vivekanand Jha^{b,d}, Amjad Ali^a

^a School of Chemistry and Biochemistry, Thapar University, Patiala, Punjab 147004, India

^b Biomedical Informatics Center of ICMR, Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh 160012, India

^c Perdana University Centre for Bioinformatics, MARDI Complex, Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia

^d Department of Nephrology, Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh 160012, India

ARTICLE INFO

Article history:

Received 27 January 2015

Received in revised form 23 July 2015

Accepted 6 August 2015

Available online 8 August 2015

Keywords:

Glycogenes

Interaction network

Hubs

nsSNPs

Solvent accessibility

Secondary structure

ABSTRACT

Glycogenes regulate a large number of biological processes such as cancer and development. In this work, we created an interaction network of 923 glycogenes to detect potential hubs from different mouse tissues using RNA-Seq data. DAVID functional cluster analysis revealed enrichment of immune response, glycoprotein and cholesterol metabolic processes. We also explored nsSNPs that may modify the expression and function of identified hubs using computational methods. We observe that the number of nsSNPs predicted by any two methods to affect protein function is 4, 7 and 2 for FLT1, NID2 and TNFRSF1B. Residues in the native and mutant proteins were analyzed for solvent accessibility and secondary structure change. Analysis of hubs can help in determining their degree of conservation and understanding their functions in biological processes. The nsSNPs proposed in this work may be further targeted through experimental methods for understanding structural and functional relationships of hub mutants.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

A large variety of biological processes during the growth and development of organisms in addition to wide array of diseases such as cancer, primary open-angle glaucoma, and renal dysfunction, are regulated by glycogenes (Firoz et al., 2014). Expression pattern of these glycogenes is one of the essential factors that indicate about several biological functions, developmental changes, and diseases in humans and several other organisms (Mantelli et al., 2009; Ishii et al., 2007; Nakayama et al., 2013). One of the most abundant types of

posttranslational modification in biological systems is glycosylation (Zhang et al., 2013) and all genes that play essential roles in glycosylation represent the glycogenome (Janot et al., 2009). The glycogenome comprises of genes involved in glycosylation of proteins, lipids, and proteoglycans and also consists of genes related to synthesis of glycans such as glycosyltransferases, sugar-nucleotide transporters, and sulfotransferases [(Janot et al., 2009; Taniguchi et al., 2009), <http://www.stelic.com/cn27/pg122.html>]. In spite of the fact that all cells in an individual mammal contain nearly duplicate DNA, the function of cells within an organism exhibits great disparity owing to gene expression pattern (Salzman et al., 2011). In general, expression of glycogenes is comparatively weak in comparison to other molecules; yet, some glycogenes are upregulated during particular circumstances [<http://www.stelic.com/cn27/pg122.html>].

Functional roles of glycogenes have been investigated by applying numerous high-throughput microarray studies (Saravanan et al., 2010; Kroes et al., 2007; Tan et al., 2014). Recently, we explored the role of glycogenes in skeletal muscle development in primary bovine MYOG_{kd} muscle satellite cells (Lee et al., 2014) by using RNA-Seq. Additionally, we also used precomputed expression values from mouse RNA-Seq data to identify and understand the roles of differentially expressed glycogenes (DEGGs) in brain, muscle, and liver tissues (Firoz et al., 2014). These studies may help in providing insights that can aid in the detection of genes and their related biological functions and in turn, translate their impact on various biological processes. Therefore, in this work, we created an interaction network of all unique

Abbreviations: nsSNP, Non synonymous single nucleotide polymorphism; RNA-Seq, RNA sequencing; MYOG_{kd}, Myogenin knock-down; DEGGs, Differentially expressed glycogenes; SNPs, Single nucleotide polymorphisms; DAVID, Database for Annotation, Visualization and Integrated Discovery; GO, Gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomics; KAAS, KEGG Automatic Annotation Server; KO, KEGG orthology; SBH, Single-directional best hit; BLAST, Basic Local Alignment Search Tool; SLC2A4, Solute carrier family 2, facilitated glucose transporter member 4; GLUT-4, glucose transporter member 4; TNFRSF1B, Tumor necrosis factor receptor superfamily member 1B; TRFR2, Transferrin receptor protein 2; TNF, Tumor Necrosis Factor; UCHL1, Ubiquitin carboxyl-terminal hydrolase isozyme L1; HPXN, Hemopexin; NID2, Nidogen-2; FLT1, Vascular endothelial growth factor receptor 1; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant; PolyPhen-2, Polymorphism Phenotyping v2; ASA, Accessible surface area; FAC, Functional Annotation Clustering; CAMs, Cell adhesion molecules.

* Correspondence to: A. Firoz, School of Chemistry and Biochemistry, Thapar University, Patiala, Punjab 147004, India.

** Corresponding author.

E-mail addresses: ahmadfirozbin@gmail.com (A. Firoz), adeel@procarb.org (A. Malik).

glycogenes that were recently detected from brain, muscle, and liver tissues of mouse RNA-Seq data (Firoz et al., 2014) and identified key hubs. We also investigated the single nucleotide polymorphisms (SNPs) within these potential hubs which can be recommended as the crucial regions for susceptibility towards altering protein function. The regions proposed in this study can be further targeted through conventional experimental methods such as gene knockout technique, gene targeting, etc., for validating their functional significance in various biological processes.

2. Materials and methods

2.1. Datasets

Recently, we explored the role of differentially expressed glycogenes (DEGGs) and identified hub genes (Firoz et al., 2014) that may play a significant role in various biological processes in the development of brain, muscle, and liver tissues using the precomputed expression values from mouse RNA-Seq data (Mortazavi et al., 2008). In this work, we created a non-redundant list of glycogenes from this RNA-Seq data (Mortazavi et al., 2008) by combining all identified glycogenes of brain, muscle and liver tissues. The data was then filtered by removing any duplicate gene.

2.2. Functional analysis and pathway mapping

Database for Annotation, Visualization and Integrated Discovery (DAVID) [<http://david.abcc.ncifcrf.gov/home.jsp>] functional annotation cluster analysis was carried out on the non-redundant list of glycogenes. Only those clusters that show statistically significant ($p\text{-value} \leq 0.05$) gene ontology (GO) terms were selected for further analysis. The GO term “biological process” (BP) in DAVID was used to classify over-represented biological processes in the non-redundant list of glycogenes. KEGG Automatic Annotation Server (KAAS) [<http://www.genome.jp/kegg/kaas/>] (Moriya et al., 2007) was used for pathway analysis. For pathway mapping the amino acid sequences of glycogenes as input were submitted to KAAS web server by using single-directional best hit (SBH) method to assign orthologs. In KAAS, the functional annotation of genes in a genome is performed by a BLAST similarity search against a manually curated set of ortholog groups in the KEGG GENES database. KEGG orthology (KO) number was assigned by KAAS to glycogenes in the data sets that were mapped to one of KEGG's reference pathways.

2.3. Network construction and identification of hub genes

GeneMANIA cytoscape plugin (Montejo et al., 2010) was used to analyze the functional interactions between glycogenes. To create an interaction network between glycogenes and additional 50 genes the GO term “biological process” and source species “mouse” were used. Connection between genes in the network represent co-expression, physical and genetic interactions, pathways, co-localization, protein domain similarity, and predicted interactions.

Hub genes were identified by calculating the node degree distribution by using NetworkAnalyzer plugin of Cytoscape (Smoot et al., 2011). Top five genes with the highest degree distribution were considered as hubs in the current network.

2.4. Computational analysis of coding non-synonymous SNPs (nsSNPs) of predicted hub genes

Data mining the SNP information for hub genes identified for mouse in the current (SLC2A4, TNFRSF1B, TRFR2, and UCHL1) and our previous study (HPXN, NID2 and FLT1) was retrieved from National Centre for Biotechnology Information (NCBI) database dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP>). Three different programs (PROVEAN, SIFT

and POLYPHEN-2) were used to predict the damaged or deleterious coding nsSNPs.

2.4.1. Prediction of deleterious or damaging coding nsSNPs using PROVEAN Protein and SIFT Sequence tools

PROVEAN (Protein Variation Effect Analyzer) Protein [http://provean.jcvi.org/seq_submit.php] and SIFT (Sorting Intolerant From Tolerant) sequence [http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html] tools were used to provide PROVEAN and SIFT predictions for a list of protein variants. Both PROVEAN and SIFT are software tools that predict whether an amino acid substitution has an impact on the biological function of a protein if the score lies below a certain threshold value. In PROVEAN, the clustering of BLAST hits is carried out and the best 30 clusters of closely linked sequences form the supporting sequence set that is further employed to make the predictions. For every supporting sequence a delta alignment score is computed and then averaged within and across clusters generating final PROVEAN score (Choi et al., 2012). A default score threshold of -2.5 or above is considered to be deleterious whereas anything less than this cut-off score has a neutral effect. On the other hand, SIFT is a multi-step algorithm, and uses sequence homology based method to classify amino acid substitutions. (Kumar et al., 2009; Ng and Henikoff, 2003). These tools accept a list of protein sequence variants as input for the predictions.

2.4.2. Prediction of functional modification of coding nsSNPs by Polymorphism Phenotyping v2 (PolyPhen-2)

The standalone PolyPhen-2 tool was installed for mouse and used to study the possible consequence of nsSNPs on protein structure and function. PolyPhen-2 predicts the possible effect of amino acid replacement on the stability and function of mouse proteins by using structural as well as comparative evolutionary considerations (Adzhubei et al., 2010). For every specific amino acid residue substitution in a protein, PolyPhen-2 mines a range of sequence and structural features of the replacement position and supplies them to a probabilistic classifier (Adzhubei et al., 2010). For this work, we have selected the default option for generating the predictions.

2.5. Prediction of protein surface accessibility and secondary structure by NetSurfP

NetSurfP (ver.1.1) webserver was used to predict protein surface accessibility or accessible surface area (ASA) and secondary structure predictions [<http://www.cbs.dtu.dk/services/NetSurfP>]. In addition to the prediction of ASA and secondary structure from sequence, NetSurfP also predicts the reliability for each prediction simultaneously in the form of a Z-score. The NetSurfP method consists of two sets of neural networks, the primary networks that are trained on sequence profiles and predicted secondary structure, and the secondary networks which exploits the outputs of primary networks as input collectively with sequence profiles to predict the relative surface exposure of the individual amino acid residues (Petersen et al., 2009). The input sequences in fasta format of normal and predicted SNPs were submitted to the server for prediction.

3. Results

We compiled a non-redundant list of 923 glycogenes (Supplementary Table ST1) by combining all glycogenes of brain, muscle and liver tissues from our previous work (Firoz et al., 2014) that we identified using mouse RNA-Seq data (Mortazavi et al., 2008). All these 923 glycogenes are annotated as “glycoproteins” in the UniProt (Magrane and Consortium, 2011) database and were further employed for functional, pathway and network analysis.

Table 1

Top 10 functional annotation clusters by DAVID.

Annotation cluster (enrichment score)	GO term	P-value
1 (12.72)	GO:0009100 ~ glycoprotein metabolic process	2.74E–21
	GO:0009101 ~ glycoprotein biosynthetic process	1.27E–16
	GO:0043413 ~ biopolymer glycosylation	2.24E–12
	GO:0006486 ~ protein amino acid glycosylation	2.24E–12
	GO:0070085 ~ glycosylation	2.24E–12
	GO:0006487 ~ protein amino acid N-linked glycosylation	1.13E–05
2 (12.18)	GO:0009611 ~ response to wounding	2.62E–23
	GO:0051604 ~ protein maturation	1.66E–20
	GO:0002526 ~ acute inflammatory response	7.90E–20
	GO:0016485 ~ protein processing	2.50E–18
	GO:0006954 ~ inflammatory response	4.29E–17
	GO:0051605 ~ protein maturation by peptide bond cleavage	4.32E–16
	GO:0002541 ~ activation of plasma proteins involved in acute inflammatory response	3.44E–15
	GO:0006956 ~ complement activation	3.44E–15
	GO:0006958 ~ complement activation, classical pathway	7.26E–13
	GO:0016064 ~ immunoglobulin mediated immune response	3.54E–12
	GO:0006952 ~ defense response	3.73E–12
	GO:0019724 ~ B cell mediated immunity	6.81E–12
	GO:0002455 ~ humoral immune response mediated by circulating immunoglobulin	1.22E–11
	GO:0006959 ~ humoral immune response	1.69E–11
	GO:0002252 ~ immune effector process	1.92E–11
	GO:0002449 ~ lymphocyte mediated immunity	2.03E–11
	GO:0002460 ~ adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.16E–11
	GO:0002250 ~ adaptive immune response	2.16E–11
	GO:0002443 ~ leukocyte mediated immunity	7.47E–11
	GO:0050778 ~ positive regulation of immune response	1.23E–10
	GO:0002253 ~ activation of immune response	1.75E–09
	GO:0045087 ~ innate immune response	3.26E–09
	GO:0002684 ~ positive regulation of immune system process	6.94E–09
	GO:0048584 ~ positive regulation of response to stimulus	9.42E–09
	GO:0006955 ~ immune response	4.11E–08
	GO:0006957 ~ complement activation, alternative pathway	1.79E–06
3 (10.03)	GO:0030203 ~ glycosaminoglycan metabolic process	9.03E–12
	GO:0006022 ~ aminoglycan metabolic process	8.84E–11
	GO:0005976 ~ polysaccharide metabolic process	9.90E–10
4 (8.88)	GO:0050817 ~ coagulation	2.60E–10
	GO:0007596 ~ blood coagulation	2.60E–10
	GO:0007599 ~ hemostasis	3.39E–10
	GO:0042060 ~ wound healing	7.98E–09
	GO:0050878 ~ regulation of body fluid levels	2.02E–08
5 (6.75)	GO:0001568 ~ blood vessel development	9.69E–09
	GO:0001944 ~ vasculature development	1.81E–08
	GO:0048514 ~ blood vessel morphogenesis	1.70E–06
6 (5.92)	GO:0001525 ~ angiogenesis	3.35E–06
	GO:0046942 ~ carboxylic acid transport	4.10E–09
	GO:0015849 ~ organic acid transport	4.97E–09
	GO:0006865 ~ amino acid transport	4.03E–06
	GO:0015837 ~ amine transport	2.09E–05
7 (5.11)	GO:0015807 ~ L-amino acid transport	0.0014557
	GO:0009311 ~ oligosaccharide metabolic process	1.03E–07
	GO:0006487 ~ protein amino acid N-linked glycosylation	1.13E–05
	GO:0006491 ~ N-glycan processing	3.89E–04
	GO:0006897 ~ endocytosis	1.75E–07
8 (4.98)	GO:0010324 ~ membrane invagination	1.75E–07
	GO:0016044 ~ membrane organization	1.05E–05
	GO:0006909 ~ phagocytosis	1.19E–05
	GO:0016192 ~ vesicle-mediated transport	0.02974658

Table 1 (continued)

Annotation cluster (enrichment score)	GO term	P-value
9 (4.88)	GO:0008203 ~ cholesterol metabolic process	3.38E–06
	GO:0016125 ~ sterol metabolic process	1.09E–05
10 (4.622)	GO:0008202 ~ steroid metabolic process	5.92E–05
	GO:0048002 ~ antigen processing and presentation of peptide antigen	1.84E–10
	GO:0002474 ~ antigen processing and presentation of peptide antigen via MHC class I	1.70E–08
	GO:0019882 ~ antigen processing and presentation	7.94E–08
	GO:0019884 ~ antigen processing and presentation of exogenous antigen	2.88E–05
	GO:0002478 ~ antigen processing and presentation of exogenous peptide antigen	5.87E–05
	GO:0019886 ~ antigen processing and presentation of exogenous peptide antigen via MHC class II	0.00542203
	GO:0002495 ~ antigen processing and presentation of peptide antigen via MHC class II	0.00542203
	GO:0002504 ~ antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	0.01033599
	GO:0042590 ~ antigen processing and presentation of exogenous peptide antigen via MHC class I	0.01972544

3.1. Functional enrichment and pathway analysis of glycogenes

In order to classify enriched biological processes that may be regulated by glycogenes, a non-redundant list was created by pooling all glycogenes from mouse brain, liver and muscle tissues and then removing the duplicates. Functional Annotation Clustering (FAC) tool available in the Database for Annotation, Visualization, and Integrated Discovery (DAVID) was used to classify over-represented biological processes using GO term “Biological Process” for the annotations. The functional annotation cluster analysis of top ten clusters (enrichment score ≥ 4.50) reveals a diverse range of biological processes in which these glycogenes are involved (Table 1). Cluster 1 (enrichment score = 12.72) and 7 (enrichment score = 5.11) mainly consists of genes that are either involved in glycoprotein metabolic process or protein glycosylation. Cluster 2 with an enrichment score of 12.18 is probably the largest cluster among the 10 selected clusters as it consists of maximum number of statistically significant GO terms specifically related to immune response and includes terms such as acute inflammatory response, complement activation, immunoglobulin mediated immune response, etc. The third functional cluster (enrichment score = 10.03) consists of genes related to glycan metabolic processes, whereas cluster 4 (8.88) and cluster 5 (6.75) include genes with known functions in blood coagulation, hemostasis, blood vessel development and blood vessel morphogenesis. Cluster 6 (enrichment score = 5.92) and 8 (enrichment score = 4.98) are either involved in organic acid and amino acid transport, or endocytosis, membrane organization, and vesicle-mediated transport. Cluster 9 (enrichment score = 4.88) has over-represented GO terms related to metabolic processes of cholesterol, sterol or steroid, whereas cluster 10 (enrichment score = 4.62) shows enrichment of processes that are concerned with antigen processing and presentation of peptide antigen.

In addition to DAVID FAC analysis, we attempted to map glycogenes on various biological pathways by submitting FASTA formatted amino acid sequences to KAAS web server that predicted 222 pathways in which these glycogenes may play a functional role. Top 20 KEGG pathways are summarized in Table 2 and a complete list of all mapped pathways is provided in Supplementary Table ST2. From these tables, it can be seen that glycogenes are involved in diverse types of biological pathways such as signaling pathways (PI3K–Akt and Rap1 signaling pathway), complement and coagulation cascades, cytokine–

Table 2
Top 20 KEGG pathways for glycogenes.

Pathway name	No. of mapped genes
04142 Lysosome	49
04151 PI3K-Akt signaling pathway	45
04610 Complement and coagulation cascades	39
04060 Cytokine-cytokine receptor interaction	38
05200 Pathways in cancer	30
04514 Cell adhesion molecules CAMs	28
05205 Proteoglycans in cancer	27
04510 Focal adhesion	26
04141 Protein processing in endoplasmic reticulum	24
04015 Rap1 signaling pathway	23
04512 ECM-receptor interaction	21
04976 Bile secretion	21
05166 HTLV-I infection	21
04145 Phagosome	20
04080 Neuroactive ligand-receptor interaction	18
05150 Staphylococcus aureus infection	18
04014 Ras signaling pathway	17
04360 Axon guidance	17
04810 Regulation of actin cytoskeleton	16
04640 Hematopoietic cell lineage	16

cytokine receptor interaction, pathways in cancer or adhesion related functions (cell adhesion molecules CAMs and focal adhesion) (Fig. 1a-c).

3.2. Analysis of interaction network and hub genes

GeneMANIA plug-in for Cytoscape (Montejo et al., 2010) was used to map glycogenes to probe interactions between these as well as additional genes that may be associated to a list of query genes by using a very large set of functional interaction data (Warde-Farley et al., 2010). Only genes that exhibited noteworthy interactions having weights ≥ 0.1 were selected for the network analysis. Furthermore, these relationships were integrated to construct an interaction network between glycogenes and additional correlated genes (Fig. 2). Network analysis by GeneMANIA for interactions among the glycogene products suggested enrichment of diverse GO terms that may be involved in lysosome, locomotion, extracellular matrix, glycoprotein metabolic process and adhesion related processes (Table 3).

The interaction network was further analyzed in Cytoscape 2.8.2, a bioinformatics package for biological network visualization and data integration (Smoot et al., 2011). All the island nodes, duplicate edges and self-loops were removed from the initial network. The final network consists of 658 nodes (circles) that represent genes and 2114 edges indicating the interactions between them. Moreover, the query glycogenes are shown as red and the genes predicted by GeneMANIA are shown in cyan (Fig. 2).

Network Analyzer plugin of Cytoscape was then used to calculate node degree for all nodes in the network. Nodes with higher node degree were considered as more important genes and were denoted

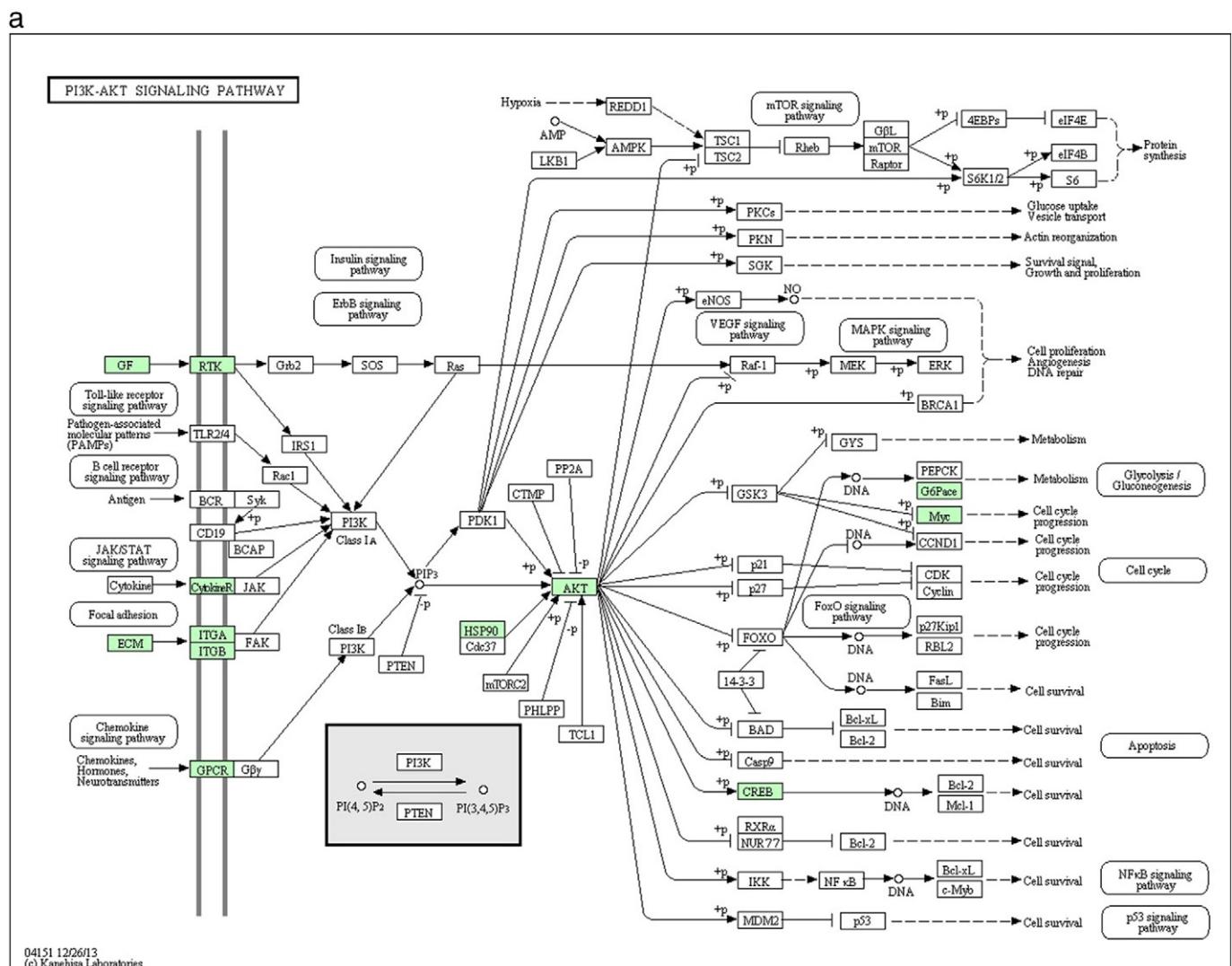


Fig. 1. Biochemical pathway analysis of glycogenes. Representative KEGG pathway for a) PI3K-AKT signaling, b) Cell adhesion molecules, and c) Complement and coagulation cascades.

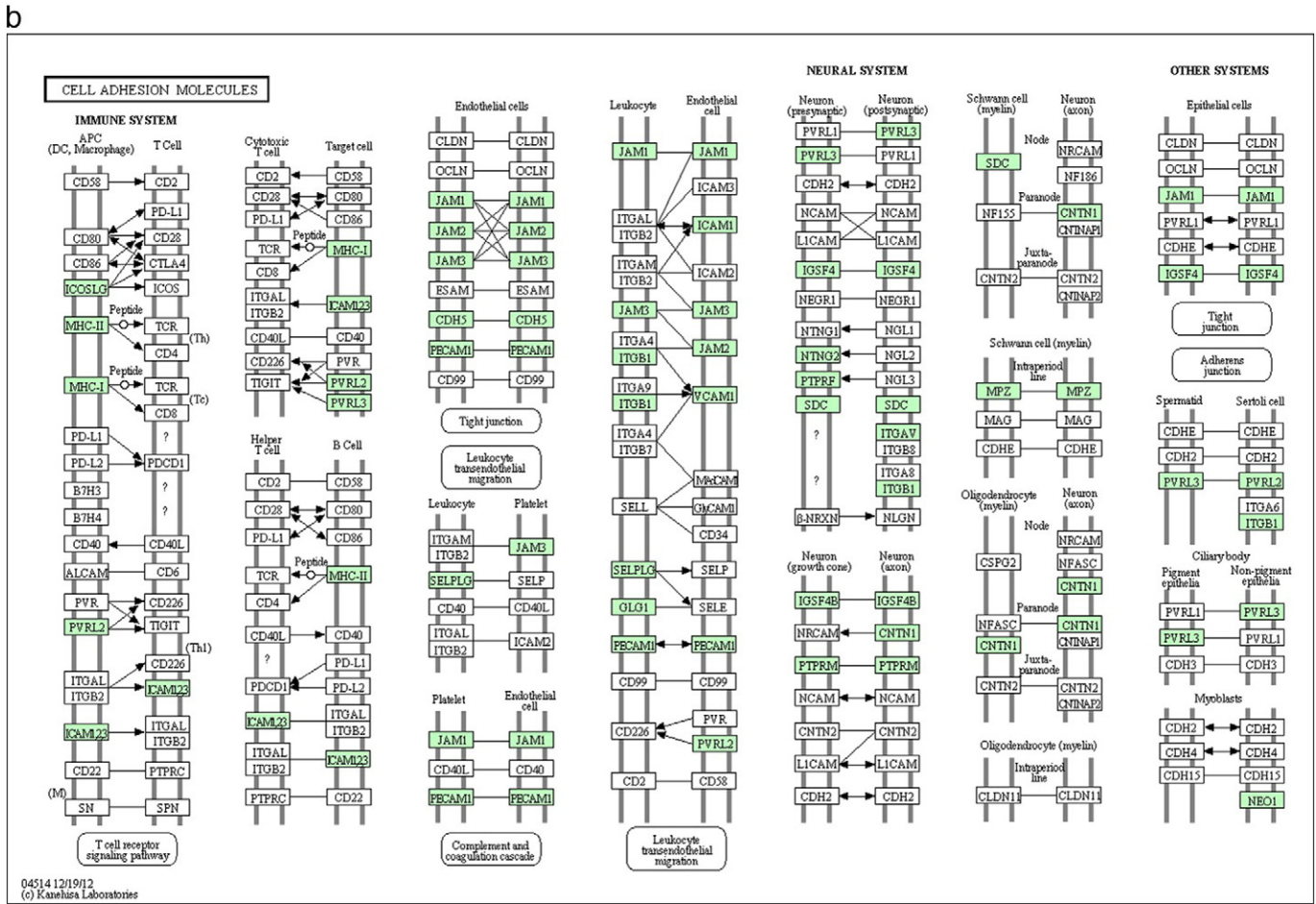


Fig. 1 (continued).

as hub genes. The glycogenes *SLC2A4*, *TNFRSF1B*, *TRFR2* and *UCLH1* showed a node degree of 21 and therefore represent hubs in the current network. Fig. 3 shows the interaction of these hubs (purple nodes) and their first neighbors. These four hub glycogenes were also a part of initial list of input genes submitted to GeneMANIA for predicting interactions.

3.3. SNP analysis

In this work, we have selected coding non synonymous SNPs (nsSNPs) for our computational analysis. Fig. 4 shows the distribution of total number of SNPs and nsSNPs identified for mouse hub genes. It can be observed that HPXN has 278 SNPs and 7 nsSNPs, whereas FLT1 has 4929 SNPs and 23 nsSNPs which are highest among all identified hub genes in both SNPs and nsSNPs. NID2 has 1624 SNPs but has only 21 nsSNPs, whereas TNFRSF1B has 830 SNPs and only 11 nsSNPs. However, no nsSNPs were detected in the remaining three hubs (*SLC2A4*, *UCLH1* and *TRFR2*) for mouse.

3.3.1. Deleterious or damaged nsSNPs by PROVEAN

The PROVEAN protein tool was used for the prediction of damaged or deleterious nsSNPs in the hub genes. The program provides prediction for a protein sequence from any organism. The input to PROVEAN consists of detected coding nsSNPs and protein sequence for hubs, that is, 7 nsSNPs for HPXN (UniProt ID: Q91X72), 23 for FLT1 (UniProt ID: P35969), 21 for NID2 (UniProt ID: O88322) and 11 for TNFRSF1B (UniProt ID: P25119) in mouse. The input for each hub was submitted independently to PROVEAN with default parameters. No deleterious SNP was identified for HPXN by PROVEAN (Supplementary Table ST3).

Among 23 nsSNPs of FLT1 in mouse, 3 were identified by PROVEAN to be deleterious (Table 4). For NID2, 3 out of 21 nsSNPs in mouse were identified to be deleterious (Table 5) whereas, 1 out of 11 nsSNPs were identified to be deleterious for TNFRSF1B (Table 6). All predictions were made by PROVEAN program by using -2.5 as a prediction cut-off.

3.3.2. Predicted nsSNPs that affect protein function by SIFT

The SIFT Sequence tool provides SIFT predictions for a given protein FASTA sequence. The input protein sequences and variants in required format for the same number of nsSNPs and hub genes as mentioned above were submitted with default parameters at http://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html. The amino acids with probabilities < 0.05 are predicted to be deleterious. No SNP was identified for HPXN that would affect protein function by SIFT (Supplementary Table ST3). Among 23 nsSNPs of FLT1 in mouse, 3 were identified to affect protein function (Table 4). For NID2, 8 out of 21 nsSNPs in mouse were identified to affect protein function (Table 5). For TNFRSF1B, 2 out of 11 nsSNPs in mouse were identified to affect protein function (Table 6) by SIFT program.

3.3.3. Damaged nsSNPs by PolyPhen-2

The protein sequences of each hub and their SNP substitution were submitted to standalone polyphen-2 installed in required format for mouse. No damaging SNP was identified for HPXN by PolyPhen-2 (Supplementary Table ST3). 7 out of 23 nsSNPs for FLT1 were identified to be damaging (Table 4). Similarly, 7 out of 21 nsSNPs in mouse were identified as damaging (Table 5) for NID2, and 2 out of 11 nsSNPs in mouse were identified to be damaging for TNFRSF1B (Table 6) by Polyphen-2.

to exposed state and a change in secondary structure from E to C in majority of the cases. The only exception is K832 in which E in the native protein is changed to helix (H) conformation in the mutant protein and also shows the maximum change in ASA as a result of this mutation (Table 7). Similarly, in case of T150M mutation, K832 again shows change from buried to exposed state and change in E to H conformation. Residue K881 on the other hand shows opposite change from exposed to buried state and H to E conformation. However, in L453F and Q1164K, all of the residues that exhibit a significant ASA change show a structural change from/to coiled conformation (Table 7).

It is interesting to note that all residues (Table 8) in NID2 exhibited a change in ASA from exposed to buried state and change from C to H conformation in the mutant protein P13L. The maximum ASA change of 55.14 Å² is shown by T11 for this mutation. The mutations V266A and T525A both show ASA change from buried to exposed state (~20 Å² change) with E to C conformational shift for residues M1183 and E308, respectively. Additionally, E308 also represents the residue with

The protein sequences of each gene in normal and their predicted SNP substitutions were individually uploaded in fasta format to NetSurfP web server for ASA and secondary structure predictions. Only those SNPs that were predicted by at least two software tools (4 SNPs of FLT1, 7 SNPs of NID2 and 2 SNPs of TNFRSF1B) to affect protein function were used for ASA and secondary structure analysis. The predicted highest probabilities of secondary structure from NetSurfP were converted by perl script to one letter code to represent helix (H), β -sheet (E) and coil (C). Additionally, the data was further filtered by selecting only those residues that showed an ASA change of $\geq 10 \text{ \AA}^2$ from buried to exposed state and vice versa, and change in their secondary structure (Tables 7–9). The complete details of ASA and secondary structure analysis are provided separately as supplementary data (Supplementary Tables S17–S19).

All residues (G532, T771, L772, K832, V859, V861 and V1162) in FLT1 having F89L mutation show a change in ASA from a buried state

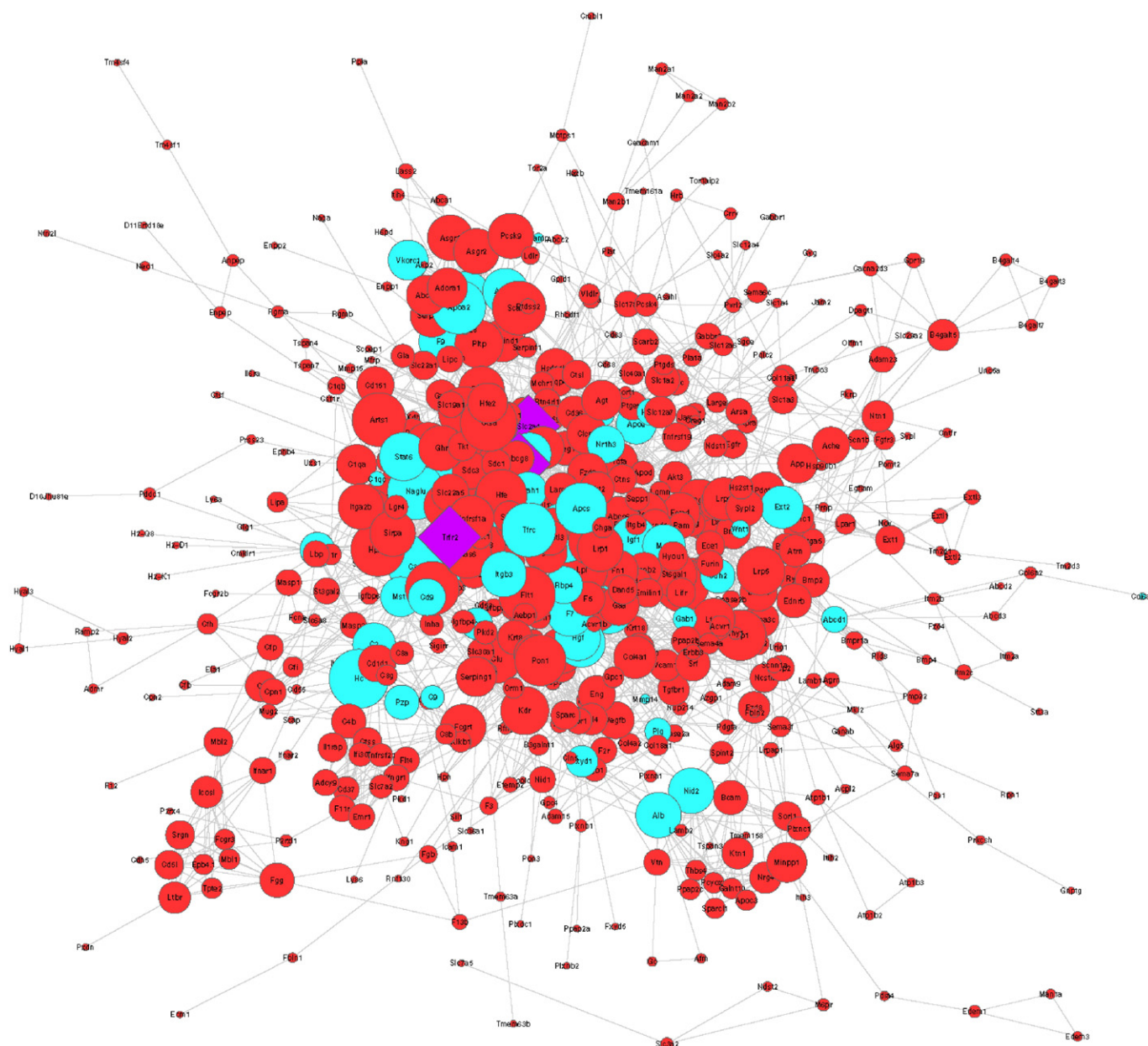


Fig. 2. Interaction networks between glycogenes and additional related genes. A red node indicates a query glycogenes and the genes predicted by GeneMANIA are shown in cyan. The top two hubs are shown as purple diamonds.

highest ASA change from buried to exposed state in mutant protein Y317D with change in E to C conformation (Table 8). No significant change was observed for NID2 protein with S14P, S290T and E810K mutations.

In case of TNFRSF1B, residues Q318 and R368 showed an ASA change from buried to exposed state in the mutant protein P431L (Table 9). Also, R368 represents the residue with maximum change in its ASA as a consequence of this mutation with change in its secondary structure from C to E conformation. No significant data was observed for S396N mutant protein.

Additionally, the SNP data that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2) and which may in turn affect protein function for hub genes (FLT1, TNFRSF1B and NID2) is available online in the form of a simple database available at www.procarb.org/glyco. The web interface uses PHP and JavaScript to allow users to search the database by applying a variety of text-based options like simple search and advanced search. Simple search can be performed by using any of the

three fields such as SNP ID, UniProt ID or Gene name. In advance search there are two fields, “Gene name” and “Residue”. The user can search a specific residue in any gene or all genes (currently the hub genes) available in this database. Both these search methods use a two step approach to display the SNP data. In first step, SNP ID, UniProt ID, Gene name, AA change, PPH2 Prob, PPH2, PROVEAN Score, PROVEAN Cutoff, SIFT Score and SIFT Cutoff will be displayed and in second step by clicking AA change, secondary structure information of particular amino acid change namely SNP ID, Residue N, Position, ASA N, Class N, SS N, Residue SNP, SNP, ASA SNP, Class SNP and SS SNP will be displayed. User can also filter search results by selecting any field such as Residue N, ASA change, Class change or SS change or any combination for a specific gene. Data entries are displayed using dynamically generated pages which describe the relevant information. All headers are defined in the help section and hyperlinked from search pages on this web portal. Moreover, each of SNP ID is also directly hyperlinked to their respective entry in dbSNP database.

Table 3
Top 20 functional categories of glycogenes as additional related genes as predicted by GeneMANIA.

GO id	Description	q-value
GO:0000323	Lytic vacuole	4.75E–29
GO:0005764	Lysosome	4.75E–29
GO:0009897	External side of plasma membrane	3.57E–28
GO:0051272	Positive regulation of cellular component movement	7.39E–26
GO:0040017	Positive regulation of locomotion	8.39E–26
GO:0031012	Extracellular matrix	8.39E–26
GO:2000147	Positive regulation of cell motility	8.39E–26
GO:0030335	Positive regulation of cell migration	1.40E–25
GO:0005773	Vacuole	1.19E–24
GO:0010876	Lipid localization	6.67E–21
GO:0009100	Glycoprotein metabolic process	1.30E–20
GO:0051604	Protein maturation	6.11E–20
GO:0016485	Protein processing	3.97E–18
GO:0001525	Angiogenesis	6.31E–18
GO:0031589	Cell-substrate adhesion	9.23E–18
GO:0019838	Growth factor binding	9.23E–18
GO:0050817	Coagulation	1.75E–17
GO:0007596	Blood coagulation	3.48E–17
GO:0007599	Hemostasis	6.45E–17
GO:0005578	Proteinaceous extracellular matrix	8.65E–17

Distribution of nsSNPs

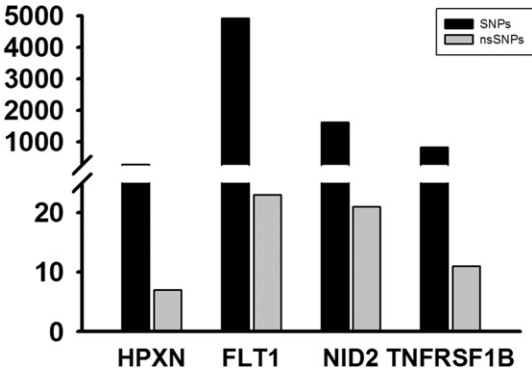


Fig. 4. Distribution of nsSNPs in hubs. Among the hubs identified in this work, SNPs were found for TNFRSF1B only. No SNP was detected for hubs SLC2A4, TRFR2 and UCHL1. HPXN, FLT1 and NID2 were identified as hub glycogenes in our previous work (Firoz et al., 2014).

4. Discussion

Through the progression of bioinformatics during the last ten years, discovery of molecular targets and targeted therapeutics have become

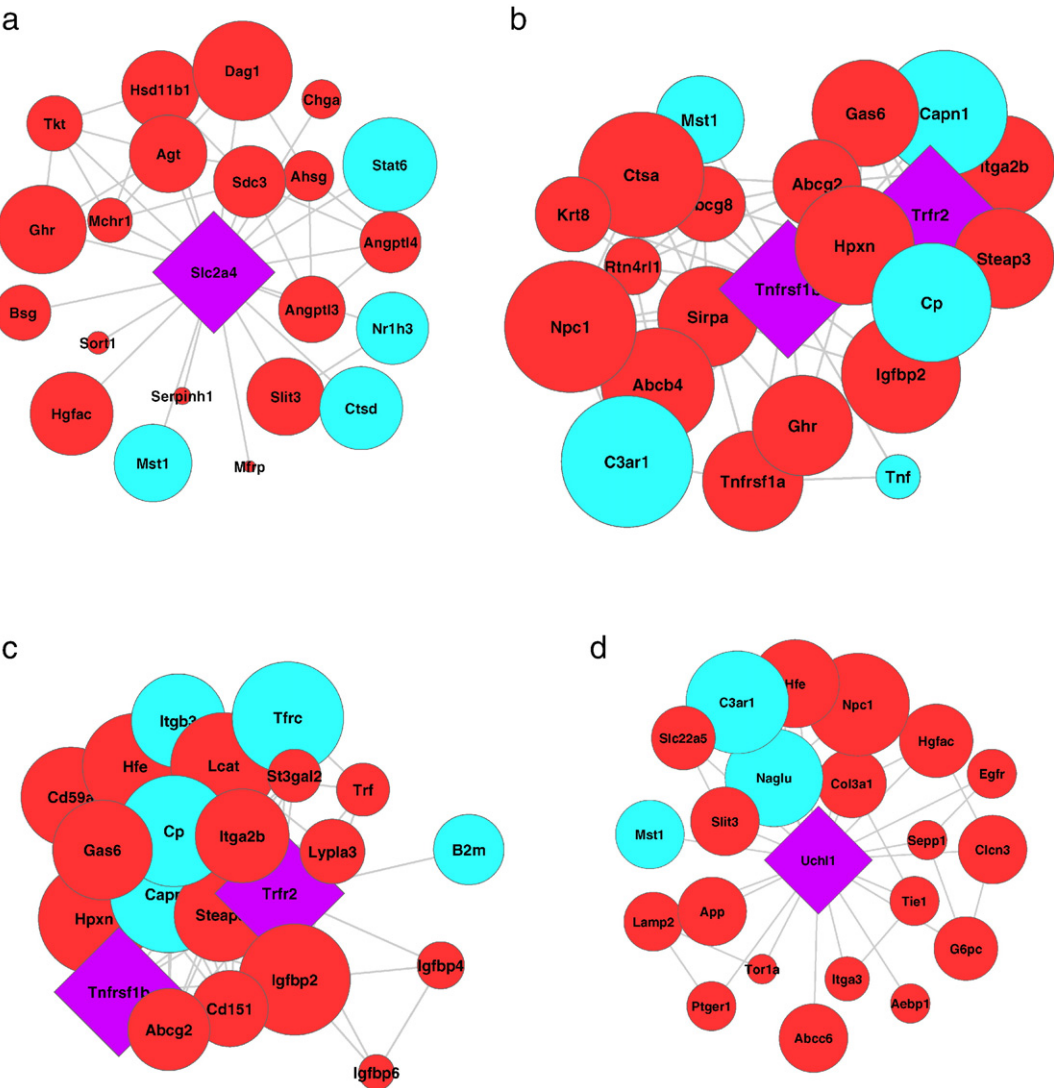


Fig. 3. Top glycogenes as hubs identified in the current work and their first neighbors. A) SLC2A4, B) TNFRSF1B, C) TRFR2 and D) UCHL1. The size of the nodes represents the node degree.

Table 4

nsSNPs of FLT1 (UniProt ID: [P35969](#)) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 3, 3 and 7 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 8 nsSNPs in mouse FLT1 gene out of total 23 were found to be deleterious, damaging or affect protein function and 15 nsSNPs were predicted to be tolerated by these three tools (Supplementary Table ST4). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA change	PROVEAN Score	PROVEAN (cutoff = −2.5)	SIFT score	SIFT (cutoff = <.05)	PPH2_Prob	PPH2
rs214988215	S80L	−2.053	Neutral	0.11	Tolerated	0.81	Possibly damaging
rs233917867	F89L	−3.699	Deleterious	0.1	Tolerated	0.942	Possibly damaging
rs213622875	T150M	−1.885	Neutral	0.02	Affect protein function	0.638	Possibly damaging
rs233071760	A277V	−1.296	Neutral	0.08	Tolerated	0.994	Probably damaging
rs32394398	D284E	−0.069	Neutral	1	Tolerated	0.821	Possibly damaging
rs232914047	L453F	−3.129	Deleterious	0	Affect protein function	1	Probably damaging
rs222631065	K575N	−0.425	Neutral	0.5	Tolerated	0.792	Possibly damaging
rs32389134	Q1164K	−2.926	Deleterious	0.02	Affect protein function	0.232	Benign

Table 5

nsSNPs of NID2 (UniProt ID: [O88322](#)) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 3, 8 and 7 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 11 nsSNPs in mouse NID2 gene out of total 21 were found to be deleterious, damaging or affect protein function and 10 nsSNPs were predicted to be tolerated by these three tools (Supplementary Table ST5). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA change	PROVEAN score	PROVEAN (cutoff = −2.5)	SIFT score	SIFT (Cutoff = <.05)	PPH2_Prob	PPH2
rs262873422	P13L	−0.254	Neutral	0.01	Affect protein function	0.93	Possibly damaging
rs30692168	S14P	−0.536	Neutral	0	Affect protein function	0.734	Possibly damaging
rs30689780	V266A	−3.217	Deleterious	0	Affect protein function	0	Benign
rs223532410	S290T	−0.531	Neutral	0	Affect protein function	0.807	Possibly damaging
rs254798552	Y317D	−2.252	Neutral	0	Affect protein function	0.987	Probably damaging
rs241645935	D351G	−1.699	Neutral	0	Affect protein function	0.003	Benign
rs30689779	V357I	−0.137	Neutral	0	Affect protein function	0	Benign
rs231633635	D406N	−0.805	Neutral	0	Affect protein function	0	Benign
rs261439016	T525A	−2.642	Deleterious	0.68	Tolerated	0.983	Probably damaging
rs30667034	E810K	−2.586	Deleterious	0.17	Tolerated	0.95	Possibly damaging
rs255251970	A869T	0.255	Neutral	1	Tolerated	0.931	Possibly damaging

essential remedial treatment for various diseases (Dinh et al., 2007). In order to explore the functional roles of all the mouse glycogenes and additionally related genes involved in various biological processes, an interaction network was created and node degree for each gene in the network was calculated. *SLC2A4*, *TNFRSF1B*, *TRFR2* and *UCHL1* were the glycogenes that were considered as hubs (genes with highest node degree) in the network constructed from 923 mouse glycogenes and additional related genes predicted by GeneMANIA. The most significant enriched GO terms within these glycogenes identified by DAVID functional cluster analysis and GeneMANIA include glycoprotein metabolic process, inflammatory response, extracellular matrix and adhesion related processes (Tables 1 and 3). Recently, we also created an interaction network of differentially expressed glycogenes (DEGGs) that are involved in similar processes (Firoz et al., 2014) by using mouse RNA-Seq data (Mortazavi et al., 2008) and identified HPXN, FLT1 and NID2 as potential hubs (Firoz et al., 2014).

One of the first potential hubs detected in this study is solute carrier family 2, facilitated glucose transporter member 4 (*SLC2A4* or *GLUT-4*). In muscle and fat cells, *GLUT-4* is stimulated by insulin from an intracellular site to the cell exterior, where it aids in decreasing the plasma glucose levels (Bryant et al., 2002). The expression of tumor necrosis

factor receptor superfamily member 1B (*TNFRSF1B*) also known as tumor necrosis factor receptor 2 (*TNFR-2*) is greatly controlled and largely limited to the cells of immune system (Grivennikov et al., 2006). *TNFR-2* is released in huge amounts throughout inflammation and acts as functional inhibitor of tumor necrosis factor (TNF) (Martin et al., 2014). Transferrin receptor protein 2 is a recognized iron sensor, encoded by *TFR2* or *TRFR2*, which is expressed in the liver and mutations that affect the protein function lead to iron-overload disease Type III hereditary hemochromatosis (HH) (Worthen and Enns, 2014). Loss of functional *TRFR2* decreases the production of hepcidin (peptide hormone), which slows down accumulation of iron in various organs and in turn leads to various diseases such as cirrhosis, heart disease, arthritis, and diabetes (Worthen and Enns, 2014). The final gene (*UCHL1*) that was also detected as a hub codes for one of the first discovered deubiquitinating enzyme Ubiquitin carboxyl-terminal hydrolase isozyme L1 (*UCH-L1*) (Hussain et al., 2013). The protein is expressed only in neuroendocrine tissues and in certain cancers with poorly understood functional roles (Hussain et al., 2013). Loss of *UCHL1* in mice leads to gracile axonal dystrophy, a lethal neurodegenerative condition (Saigoh et al., 1999), and mutations have also been implicated in the pathogenesis of Parkinson's disease (Maraganore et al., 2004; Ragland

Table 6

nsSNPs of *TNFRSF1B* (UniProt ID: [P25119](#)) in mouse that were predicted to be damaging/deleterious/affect protein function by PROVEAN, SIFT and Polyphen2. From the table it can be observed that 1, 2 and 2 nsSNPs were reported to affect protein function by PROVEAN, SIFT and PolyPhen-2, respectively. Overall 2 nsSNPs in mouse *TNFRSF1B* gene out of total 11 were found to be deleterious, damaging or affect protein function and 9 nsSNPs were predicted to be tolerated by these three tools (Supplementary Table ST6). The nsSNPs that were predicted to affect protein function by at least two programs are shown in bold.

SNP ID	AA change	PROVEAN score	PROVEAN (cutoff = −2.5)	SIFT Score	SIFT (Cutoff = <.05)	PPH2_Prob	PPH2
rs250384859	S396N	−0.651	Neutral	0	Affect protein function	0.858	Possibly damaging
rs0901502	P431L	−2.631	Deleterious	0.01	Affect protein function	0.996	Probably damaging

Table 7

Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of $\geq 10 \text{ \AA}^2$, and also show secondary structure change in FLT1 gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP = residues in normal and SNP sequence; ASA_N and ASA_SNP = ASA of residues in normal and SNP sequence; Class_N and Class_SNP = Class of residues in normal and SNP sequence [Class = buried (B) or exposed (E)]; SS_N and SS_SNP = secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change as well as change from E to H or H to E are highlighted in bold.

SNP ID	Residue_N	Position	ASA_N ($\text{\AA}^2$)	Class_N	SS_N	Residue_SNP	SNP	ASA_SNP ($\text{\AA}^2$)	Class_SNP	SS_SNP	ASA-Change ($\text{\AA}^2$)
rs233917867	G	532	20.092	B	E	G	F89L	33.629	E	C	−13.537
	T	771	30.819	B	E	T	F89L	45.507	E	C	−14.688
	L	772	42.278	B	E	L	F89L	60.405	E	C	−18.127
	K	832	57.452	B	E	K	F89L	95.445	E	H	−37.993
	V	859	31.585	B	E	V	F89L	60.589	E	C	−29.004
	V	861	34.921	B	E	V	F89L	67.705	E	C	−32.784
	V	1162	45.096	B	H	V	F89L	55.501	E	C	−10.405
rs213622875	L	772	42.278	B	E	L	T150M	58.134	E	C	−15.856
	K	832	57.452	B	E	K	T150M	95.239	E	H	−37.787
	S	846	40.317	E	C	S	T150M	29.898	B	E	10.419
	V	859	31.585	B	E	V	T150M	58.16	E	C	−26.575
	M	863	62.331	E	E	M	T150M	52.206	B	C	10.125
	L	876	47.057	B	H	L	T150M	72.489	E	C	−25.432
	K	881	104.331	E	H	K	T150M	48.052	B	E	56.279
rs232914047	I	882	63.085	E	C	I	T150M	22.478	B	E	40.607
	V	749	52.535	E	C	V	L453F	29.049	B	E	23.486
	L	828	32.775	B	E	L	L453F	59.947	E	C	−27.172
	K	832	57.452	B	E	K	L453F	71.069	E	C	−13.617
rs32389134	N	891	29.968	B	E	N	L453F	41.314	E	C	−11.346
	L	772	42.278	B	E	L	Q1164K	67.198	E	C	−24.92
	W	774	63.588	B	E	W	Q1164K	86.147	E	C	−22.559
	I	882	63.085	E	C	I	Q1164K	40.904	B	H	22.181
	K	914	69.3	E	C	K	Q1164K	53.585	B	E	15.715
	Y	921	66.76	E	C	Y	Q1164K	54.878	B	H	11.882
	L	922	53.502	E	C	L	Q1164K	40.813	B	H	12.689

et al., 2009). Similarly, the hubs (HPXN, FLT1 and NID2) identified in our previous work (Firoz et al., 2014) may possibly play essential roles in regulating various important biological processes as described above.

The functional annotation cluster analysis on non-redundant list of 923 glycogenes was performed by using DAVID. The over-representation of GO terms such as metabolic processes of glycoprotein, glycan and cholesterol; immune response; blood coagulation; organic acid and amino acid transport to name a few suggests the diverse roles that these glycogenes carry out (Janot et al., 2009; Saravanan et al., 2010; Ngamukote et al., 2007). The diverse roles of glycogenes are also supported by KEGG pathway and GeneMANIA enrichment analysis (Tables 2 and 3).

Single amino acid mutations within the proteins can considerably alter the stability of a protein structure and in turn may affect its function. Therefore, to understand the functional effects of variations we also investigated the role of nsSNPs that may lead to an amino acid change in the hub proteins. Identifying nsSNPs that disturb protein function and relate to a disease is considered as a defying task in the future (Karchin et al., 2005a,b). A large number of nsSNPs were identified in this study (Supplementary Tables ST3–ST6) for almost all the hubs, however, only few had an effect on protein function (Tables 4–6).

This could possibly be due to the fact that the effect of many nsSNPs might be neutral because natural selection may have eliminated mutations at critical sites (George Priya Doss et al., 2008).

Structural significance of native and mutant sequence of the hubs was further investigated in this work by sequence based predictions of solvent accessibility and secondary structure analysis. For mutants F89L and T150 in FLT1, K832 (SNP ID: rs233917867 and rs213622875) and K881 (SNP ID: rs213622875) are the only residues that exhibit significant ASA change as well as a secondary structure change from β -sheet to helix or vice versa [Tables 7–9]. Because of their charged nature, lysine residues are most expected to be present at solvent accessible areas of proteins, thereby making them interesting candidates for further studies on their roles on molecular interactions (Scholten et al., 2006). Solvent accessibility thought of as a discriminating characteristic in nsSNPs linked with a disease, have a tendency to occur at buried sites where as benign substitutions are inclined towards solvent accessible sites (Sunyaev et al., 2000; Ferrer-Costa et al., 2002). Amino acid residues that form the hydrophobic core of a protein are essential for its stability (George Priya Doss et al., 2008). Polar and charged side chains have greater ASA than nonpolar side chains proposing that establishment of a hydrophobic core is a solid driving force in protein folding (Chan and Dill, 1990).

Table 8

Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of $\geq 10 \text{ \AA}^2$, and also secondary structure in NID2 gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP = residues in normal and SNP sequence; ASA_N and ASA_SNP = ASA of residues in normal and SNP sequence; Class_N and Class_SNP = Class of residues in normal and SNP sequence [Class = buried (B) or exposed (E)]; SS_N and SS_SNP = secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change are highlighted in bold.

SNP ID	Residue_N	Position	ASA_N ($\text{\AA}^2$)	Class_N	SS_N	Residue_SNP	SNP	ASA_SNP ($\text{\AA}^2$)	Class_SNP	SS_SNP	ASA-change ($\text{\AA}^2$)
rs262873422	T	11	68.643	E	C	T	P13L	13.496	B	H	55.147
	P	12	50.644	E	C	P	P13L	4.583	B	H	46.061
	P	13	47.778	E	C	L	P13L	10.107	B	H	37.671
	R	27	56.907	E	C	R	P13L	43.144	B	H	13.763
rs30689780	M	1183	47.404	B	E	M	V266A	68.254	E	C	−20.85
rs254798552	E	308	51.624	B	E	E	Y317D	94.949	E	C	−43.325
	M	1183	47.404	B	E	M	Y317D	66.633	E	C	−19.229
	T	1193	35.59	E	E	T	Y317D	24.772	B	C	10.818
rs261439016	E	308	51.624	B	E	E	T525A	70.963	E	C	−19.339

Table 9

Residues that show a change in ASA from buried to exposed state or vice versa with change in ASA of $\geq 10 \text{ \AA}^2$, and also secondary structure in TNFRSF1B gene of mouse due to SNPs predicted by NetSurfP. Here Residue_N and Residue_SNP = residues in normal and SNP sequence; ASA_N and ASA_SNP = ASA of residues in normal and SNP sequence; Class_N and Class_SNP = Class of residues in normal and SNP sequence [Class = buried (B) or exposed (E)]; SS_N and SS_SNP = secondary structure of residue in normal and SNP sequence. Residues with maximum ASA change are highlighted in bold.

SNP ID	Residue_N	Position	ASA_N ($\text{\AA}^2$)	Class_N	SS_N	Residue_SNP	SNP	ASA_SNP ($\text{\AA}^2$)	Class_SNP	SS_SNP	ASA-change ($\text{\AA}^2$)
rs50901502	Q	318	38.149	B	C	Q	P431L	60.224	E	H	–22.075
	R	368	63.273	B	C	R	P431L	93.844	E	E	–30.571
	T	379	49.585	E	C	T	P431L	26.575	B	E	23.01

A large amount of effort is being made in current epidemiology, medicine, and pharmacogenomics which focus on the discovery of genetic variations especially SNPs that are implicated in common and complex diseases (Lee and Shatkay, 2008). Due to the large number of SNPs present in the genomes, prioritizing SNPs on the basis of their possible deleterious functional effects is essential (Bhatti et al., 2006). Since there is lack of experimental data for the majority of SNPs, therefore, it becomes challenging to authenticate their deleterious effects. Many computational resources have been developed and widely used that predict the assumed deleterious effects of these SNPs (Rebeck et al., 2004). In general, these computational tools scrutinize whether the SNP inhabits in functional genomic regions such as exons, splice sites, or transcription regulatory sites, and accordingly predict the possible related functional effects that the SNP may have using a range of machine learning methods (Lee and Shatkay, 2008). However, to comprehensively evaluate the functional implication of SNPs, substantial amount of time and effort is required to apply multiple tools independently, and interpret the prediction results carefully (Lee and Shatkay, 2008). It will be worth to obtain additional information regarding the selected SNPs before finalizing the decision. For example, SNPs in protein coding or regulatory regions are more prone to have deleterious effects, and, as such, the potential of triggering a disease is enhanced. Therefore, by using these computational methods and by acquiring additional information about the functionally significant SNPs that are expected to be linked with disease, a significant amount of genotyping and analysis overhead can be reduced (Lee and Shatkay, 2008).

The results from our current study proposes that the application of these computational methods may offer alternate methods to select target SNPs by understanding the consequence of SNPs on the functional characteristics or molecular phenotype of a protein. Our findings might be useful for additional genotype–phenotype investigations as well as pharmacogenetics studies. Finally, the functional analysis in this study might serve as a worthy model for further exploration in genetically inherited diseases.

5. Conclusion

In this work, we created an interaction network and detected hubs from a list of 923 glyco genes identified from mouse RNA-Seq data and investigated the functional impact of nsSNPs in the hub glyco genes via computational methods. Overall, the number of nsSNPs that are predicted to be deleterious or damaging by using any two tools out of three (PROVEAN, SIFT and Polyphen-2) that may affect protein function is 4 for *FLT1*, 2 for *TNFRSF1B*, and 7 for *NID2*, respectively. Therefore, our investigation will offer valuable information in selecting SNPs that are expected to have impending functional influence and eventually contribute in understanding the functional roles of these hub glyco genes.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.gene.2015.08.012>.

Acknowledgment

Authors acknowledge the support from Biomedical Informatics Center of ICMR, PGIMER (Chandigarh, India); School of Chemistry and Biochemistry, Thapar University, (Patiala, India) and Perdana University

Centre for Bioinformatics, (Selangor, Malaysia) for providing computational resources and support to carry out this research.

References

- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., et al., 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249.
- Bhatti, P., Church, D.M., Rutter, J.L., Struwing, J.P., Sigurdson, A.J., 2006. Oct 15. Candidate single nucleotide polymorphism selection using publicly available tools: a guide for epidemiologists. *Am. J. Epidemiol.* 164 (8), 794–804.
- Bryant, N.J., Govers, R., James, D.E., 2002. Apr. Regulated transport of the glucose transporter GLUT4. *Nat. Rev. Mol. Cell Biol.* 3 (4), 267–277.
- Chan, H.S., Dill, K.A., 1990. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci.* 87, 6388–6392.
- Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and Indels. *PLoS ONE* 7 (10), e46688.
- Dinh, P., Sotiropoulos, C., Piccart, M.J., 2007. Dec. The evolution of treatment strategies: aiming at the target. *Breast* 16 (Suppl. 2), S10–S16.
- Ferrer-Costa, C., Orozco, M., De la Cruz, X., 2002. Characterization of disease-associated single acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* 315, 771–786.
- Firoz, A., Malik, A., Singh, S.K., Jha, V., Ali, A., 2014. Comparative analysis of glycogene expression in different mouse tissues using RNA-Seq data. *Int. J. Genomics* 2014, 1–18 (Article ID 837365).
- George Priya Doss, C., Rajasekaran, R., Sudandiradoss, C., Ramanathan, K., Purohit, R., Sethumadhavan, R., 2008. Jann. A novel computational and structural analysis of nsSNPs in CFTR gene. *Genom. Med.* 2 (1–2), 23–32.
- Grivennikov, S.I., Kuprash, D.V., Liu, Z.G., Nedospasov, S.A., 2006. Intracellular signals and events activated by cytokines of the tumor necrosis factor superfamily: from simple paradigms to complex mechanisms. *Int. Rev. Cytol.* 252, 129–161.
- Hussain, S., Feldman, A.L., Das, C., Ziesmer, S.C., Ansell, S.M., Galaray, P.J., 2013. Marr. Ubiquitin hydrolase UCH-L1 destabilizes mTOR complex 1 by antagonizing DDB1-CUL4-mediated ubiquitination of raptor. *Mol. Cell. Biol.* 33 (6), 1188–1197.
- Ishii, A., Ikeda, T., Hitoshi, S., Fujimoto, I., Torii, T., Sakuma, K., Nakakita, S., Hase, S., Ikenaka, K., 2007. Marr. Developmental changes in the expression of glyco genes and the content of N-glycans in the mouse cerebral cortex. *Glycobiology* 17 (3), 261–276.
- Janot, M., Audfray, A., Lorient, C., Germot, A., Maftah, A., Dupuy, F., 2009. Oct 20. Glycogene expression dynamics during mouse C2C12 myoblast differentiation suggests a sequential reorganization of membrane glycoconjugates. *BMC Genomics* 10, 483.
- Karchin, R., Kelly, L., Sali, A., 2005a. Improving functional annotation of non-synonymous SNPs with information theory. *Pac. Symp. Biocomput.* 10, 397–408.
- Karchin, R., Diekhans, M., Kelly, L., et al., 2005b. LS-SNP: large-scale annotation of coding nonsynonymous SNPs based on multiple information sources. *Bioinformatics* 21 (12), 2814–2820.
- Kroes, R.A., Dawson, G., Moskal, J.R., 2007. Novv. Focused microarray analysis of glyco-gene expression in human glioblastomas. *J. Neurochem.* 103 (Suppl. 1), 14–24.
- Kumar, P., Henikoff, S., Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4, 1073–1081.
- Lee, P.H., Shatkay, H., 2008. Novv. Ranking single nucleotide polymorphisms by potential deleterious effects. *AMIA Ann. Symp. Proc.* 6, 667–671.
- Lee, E.J., Malik, A., Pokharel, S., Ahmad, S., Mir, B.A., et al., 2014. Identification of genes differentially expressed in myogenin knock-down bovine muscle satellite cells during differentiation through RNA sequencing analysis. *PLoS ONE* 9 (3), e92447.
- Magrane, M., Consortium, U., 2011. Mar 29. UniProt Knowledgebase: A Hub of Integrated Protein Data. *Database (Oxford)* p. bar009.
- Mantelli, F., Schaffer, L., Dana, R., Head, S.R., Argüeso, P., 2009. Junn. Glycogene expression in conjunctiva of patients with dry eye: downregulation of Notch signaling. *Invest. Ophthalmol. Vis. Sci.* 50 (6), 2666–2672.
- Maraganore, D.M., Lesnick, T.G., Elbaz, A., Chartier-Harlin, M.C., Gasser, T., Kruger, R., Hattori, N., Mellick, G.D., Quattrone, A., Satoh, J., Toda, T., Wang, J., Ioannidis, J.P., de Andrade, M., Rocca, W.A., 2004. UCHL1 is a Parkinson's disease susceptibility gene. *Ann. Neurol.* 55, 512–521.
- Martin, E.M., Remke, A., Pfeifer, E., Polz, J., Pietryga-Krieger, A., Steffens-Weber, D., Freudenberg, M.A., Mostböck, S., Männel, D.N., 2014. Oct. TNFR2 maintains adequate IL-12 production by dendritic cells in inflammatory responses by regulating endogenous TNF levels. *Innate Immunol.* 20 (7), 712–720.
- Montejo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., et al., 2010. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* 26 (22), 2927–2928.

- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007 Jull. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185 (Web Server issue).
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008 Jull. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5 (7), 621–628.
- Nakayama, Y., Nakamura, N., Tsuji, D., Itoh, K., Kurosaka, A., 2013. Genetic diseases associated with protein glycosylation disorders in mammals. In *Genetic Disorders*, InTechOpen.
- Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31 (13), 3812–3814.
- Ngamukote, S., Yanagisawa, M., Ariga, T., Ando, S., Yu, R.K., 2007 Dec. Developmental changes of glycosphingolipids and expression of glycogenes in mouse brains. *J. Neurochem.* 103 (6), 2327–2341.
- Petersen, B., Petersen, T.N., 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.
- Ragland, M., Hutter, C., Zabetian, C., Edwards, K., 2009. Association between the ubiquitin carboxyl-terminal esterase L1 gene (UCHL1) S18Y variant and Parkinson's disease: a HuGE review and meta-analysis. *Am. J. Epidemiol.* 170, 1344–1357.
- Rebbeck, T.R., Spitz, M., Wu, X., 2004 Aug. Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.* 5 (8), 589–597.
- Saigoh, K., Wang, Y.L., Suh, J.G., Yamanishi, T., Sakai, Y., Kiyosawa, H., Harada, T., Ichihara, N., Wakana, S., Kikuchi, T., Wada, K., 1999. Intragenic deletion in the gene encoding ubiquitin carboxy-terminal hydrolase in gad mice. *Nat. Genet.* 23, 47–51.
- Salzman, J., Jiang, H., Wong, W.H., 2011 Feb. Statistical modeling of RNA-Seq data. *Stat. Sci.* 26 (1).
- Saravanan, C., Cao, Z., Head, S.R., Panjwani, N., 2010 Jann. Analysis of differential expression of glycosyltransferases in healing corneas by glycogene microarrays. *Glycobiology* 20 (1), 13–23.
- Scholten, A., Visser, N.F., Van Den Heuvel, R.H., Heck, A.J., 2006 Jull. Analysis of protein-protein interaction surfaces using a combination of efficient lysine acetylation and nano LC–MALDI–MS/MS applied to the E9:Im9 bacteriotoxin – immunity protein complex. *J. Am. Soc. Mass Spectrom.* 17 (7), 983–994.
- Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T., 2011 Feb 1. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27 (3), 431–432.
- Sunyaev, S., Ramensky, V., Bork, P., 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* 16, 198–200.
- Tan, Z., Lu, W., Li, X., Yang, G., Guo, J., Yu, H., Li, Z., Guan, F., 2014 Jun 6. Altered N-Glycan expression profile in epithelial-to-mesenchymal transition of NMuMG cells revealed by an integrated strategy using mass spectrometry and glycogene and lectin microarray analysis. *J. Proteome Res.* 13 (6), 2783–2795.
- Taniguchi, N., Suzuki, A., Ito, Y., et al., 2009. *Experimental Glycoscience: Glycobiology*. Springer, New York, NY, USA.
- Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G.D., Morris, Q., 2010 Jull. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220 (Web Server issue).
- Worthen, C.A., Enns, C.A., 2014 Mar 6. The role of hepatic transferrin receptor 2 in the regulation of iron homeostasis in the body. *Front. Pharmacol.* 5, 34.
- Zhang, Z., Sun, J., Hao, L., Liu, C., Ma, H., Jia, L., 2013 Jun 20. Modification of glycosylation mediates the invasive properties of murine hepatocarcinoma cell lines to lymph nodes. *PLoS ONE* 8 (6), e65218.