Gene 574 (2015) 235–246

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

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.

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; MYOGkd, Myogenin knock-down; DEGGs, Differentially expressed glycogense; 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).

http://dx.doi.org/10.1016/j.gene.2015.08.012
0378-1119/

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-value ≤ 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 (Montojo 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.