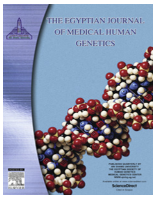
The Egyptian Journal of Medical Human Genetics 18 (2017) 321–328

HOSTED BY

Contents lists available at ScienceDirect

The Egyptian Journal of Medical Human Genetics

journal homepage: www.sciencedirect.com



Original article

Identification of functional SNPs in human *LGALS3* gene by *in silico* analyses

Tarnjeet Kaur [a], Kshema Thakur [b], Jatinder Singh [b], Sukhdev Singh Kamboj [b], Manpreet Kaur [a,*]

[a] *Department of Human Genetics, Guru Nanak Dev University, Amritsar, India*
[b] *Department of Molecular Biology and Biochemistry, Guru Nanak Dev University, Amritsar, India*

ARTICLE INFO

Article history:
Received 29 December 2016
Accepted 10 February 2017
Available online 24 February 2017

Keywords:
Computational analyses
Deleterious variations
Galectin-3
Genetic variants
Single nucleotide polymorphism
Regulatory SNPs

ABSTRACT

Background: Galectin-3 protein, an S-type lectin, is encoded by *LGALS3* gene. It consists of carbohydrate recognition domain (CRD), collagen like tandem repeats of nine amino acids and N-terminal 12-mer peptide. Its serum levels as well as some genetic variants were reported to be involved in various disease conditions like cancer, autoimmune diseases, heart diseases *etc*. Being viewed as an important molecule in biological responses and its association with various diseases, the present study was designed. This is the first *in silico* analyses of *LGALS3*.
Aim: To systematically explore the plausible effects of *LGALS3* genetic variants on structure and functions of galectin-3.
Material and methods: Both sequence based and structure based approaches were adopted for analyses of non-synonymous single nucleotide polymorphisms (nsSNPs). Putative methylation and other post translational modifications were also analyzed using different tools. Muster and Swiss-PDB Viewer were used for modeling of predicted functional variants.
Results: Out of 1130 SNPs reported in dbSNP, only validated SNPs were chosen for analyses. A total of nine nsSNPs which included, 3 of N-terminal region and 6 of CRD encoding region, were found to have deleterious effect as predicted by various softwares. Analyses of regulatory SNPs predicted five functional SNPs in 3′UTR having putative miRNA binding sites and 3 intronic SNPs in potential transcription factor binding sites.
Conclusion: Based on these analyses, the present study suggested that the reported functional SNPs may act as potential targets in genetic association studies.

1. Introduction

Galectin-3, a chimera type member of galectin family, is also being known as a potent immuno-regulator. It is encoded by a single gene *LGALS3* located on chromosome 14q22.3. *LGALS3* gene consists of 6 exons and 5 introns spanning a total of approximately 17 kb of total genome. Exon I and part of exon II encodes untranslated region of the gene, while translation initiation site is located on other part of exon II [1]. Galectin-3 molecule consists of carbohydrate recognition domain (CRD) and N-terminal region including collagen like tandem repeats of nine amino acids Tyro-Pro-Gly-(Pro/Gln)-(Ala/Thr)-(Pro/Ala)-Pro-Gly and 12-mer peptide sequence [2–3]. N-terminal region is encoded by exon III, while CRD is encoded by exon IV, V and VI (Fig. 1). N-terminal region consists of 100–150 amino acids and it directs the secretion of galectin-3 molecule by non-classical pathway. On the other hand, CRD consists of about 135 amino acids and exhibits carbohydrate recognition activity. The biological role of galectin-3 protein is defined by its varied cellular localization. It may be found intracellularly in nucleus or in cytoplasm and extra-cellularly on cell surface or in extracellular space [4–5]. It was reported to be involved in various biological processes such as cellular homeostasis, cell adhesion, angiogenesis, cell signaling, cellular growth, differentiation, cell cycle and apoptosis [6]. Consequently, it was viewed as an important molecule in aberrant immune responses and in tumorigenesis [7–9]. Several single nucleotide gene polymorphisms (SNPs) have been identified in *LGALS3*, harboring intronic, exonic and untranslated regions (UTRs).

SNPs are the simplest form of variations and source of 90% of variations reported in human population [10]. These can be of many types including synonymous SNPs, non-synonymous SNPs

Peer review under responsibility of Ain Shams University.
* Corresponding author.
E-mail address: dr.manpreetdhuna@gmail.com (M. Kaur).

http://dx.doi.org/10.1016/j.ejmhg.2017.02.001
1110-8630/© 2017 Ain Shams University. Production and hosting by Elsevier B.V.

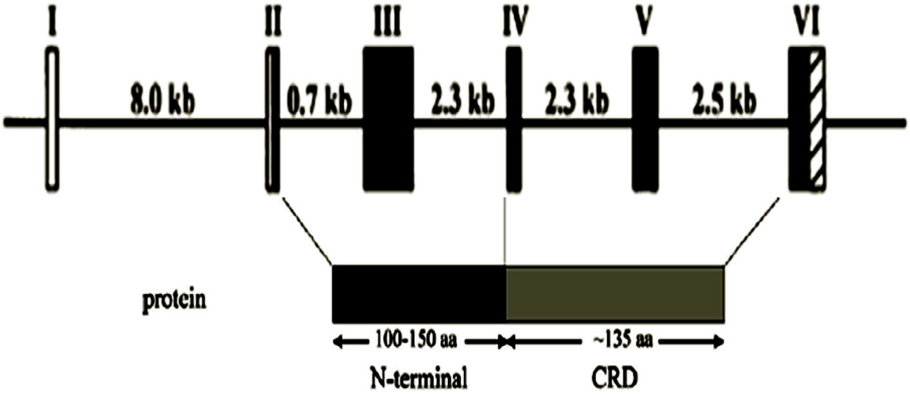


Fig. 1. Gene and protein structure of galectin-3 (adopted from Kadrofske et al., 1998).

(nsSNPs) as well as 3′UTR, 5′UTR and intronic variants. It is likely that nsSNPs play a major role in the functional diversity of encoded proteins and have been linked with many disease conditions [11–12]. These SNPs may affect the protein function by reducing protein solubility or by destabilizing protein structure [13]. The other variants in promoter or intronic regions may affect gene regulation by altering transcription and subsequently translation through altered transcription factor binding sites or splicing sites [14]. Variations in *LGALS3* gene were also found to be associated with various pathological conditions including cancers and autoimmune diseases [15–17]. Genetic analysis of all the SNPs of a particular gene for susceptibility towards any disease is not of much relevance. Moreover, it becomes even more challenging to study all the genetic determinants on large sized population based studies. So, it becomes mandatory to prioritize the functional SNPs, from the plethora of SNPs, to target in genetic and functional studies.

Taking into account all these considerations, the present study was undertaken to explore the possible effects of various *LGALS3* variants on the structure and function of galectin-3 using different computational tools. To the best of our knowledge, this is the first comprehensive and systematic *in silico* analyses of *LGALS3* gene.

2. Material and methods

2.1. Data mining

The data on human *LGALS3* gene was retrieved from Entrez Gene from National Center for Biological Information (NCBI) database. The SNP information (reference sequence ID) and protein sequence (accession number) of the *LGALS3* gene were retrieved from NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp/) and SwissProt (http://expasy.org/) databases respectively.

2.2. Identification of deleterious nsSNPs

To determine the functional impact (deleterious, damaging or neutral), coding nsSNPs were analyzed using five different tools. This ensured the stringency and accuracy of results at the time of screening and only to be likely deleterious variations were further subjected for other analyses like effect on 3D structure, surface accessibility and energy changes using different tools.

SIFT (Sorting intolerant from tolerant) server was used to identify the tolerated and deleterious SNPs. The effect of amino acid substitution (AAS) on protein structure was assessed on the basis of degree of conservation of amino acids using sequence homology (http://siftdna.org/www/SIFT_dbSNP.html). Substitution of an amino acid at each position with probability < 0.05 is predicted to be deleterious and intolerant, while probability $\geq$ 0.05 is considered as tolerant [18]. Out of various reported nsSNPs, a total of 36 nsSNPs were filtered out on the basis of their validation, which were further used as input for SIFT. The input query was submitted in the form of SNP rs IDs.

Polyphen (polymorphism and phenotype) server was used to predict the functional impact of amino acid substitution on protein structure and function based on sequence based characterization (http://genetics.bwh.harvard.edu/pph2/). Prediction outcome was obtained in the form of probability score which classifies the variations as ‘probably damaging’, ‘possibly damaging’ and ‘benign’ [19]. UniProt KB protein accession ID P17931 along with position and name of wild type and variant amino acids of screened nsSNPs were submitted as query.

nsSNP Analyzer, a web based server, was used to predict the disease associated variations based on integrated information on multiple sequence alignment and 3D structure analysis of protein (http://snpanalyzer.utmem.edu/). Additionally, it also provides the structural environment of SNP including solvent polarity, secondary structure and solvent accessibility [20]. Protein sequence along with the list of nsSNPs defining wild type and mutant amino acid residues were submitted as input for analysis.

SNPs & GO web server was used to predict the human disease related mutations (http://snps.biofold.org/snps-and-go). This server was mainly based on support vector machines which can corroborates all the information regarding variations from the existing databases. It annotates variations as deleterious based on information derived from Gene Ontology (GO) Predictor with overall accuracy of 82% [21]. UniProt accession ID of *LGALS3* (P17931) along with name and position of wild type and mutant amino acid was submitted as input for this server.

PANTHER (Protein Analysis Through Evolutionary Relationship) was used to classify the proteins on the basis of their evolutionary relationship, molecular functions and interaction with other proteins (http://www.pantherdb.org/tools.). Analysis was based on substitution position specific evolutionary conservation score, calculated from alignment of different evolutionarily related proteins [22]. Query was submitted as Protein FASTA sequence along with AAS.

Mutpred was used to predict structural and functional changes as a consequence of AAS (http://mutpred.mutdb.org/). These changes were expressed as probabilities of gain or loss of structure and function. It also provided the information regarding specific molecular mechanism responsible for the disease [23]. The input was provided in form of protein FASTA sequence along with amino acid substitution.

2.3. Analyzing the effect of nsSNPs on protein stability, surface accessibility and secondary structure

I-Mutant2.0 is a tool used for prediction of changes in protein stability due to single site mutations under different conditions (http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi). It is a web server based on support vector machine which worked on dataset derived from Protherm, a database of experimental records on protein mutations. It can predict the stability changes in protein with 80% accuracy based on its structure and with 77% of accuracy based on its sequence [24]. The input can be submitted as either in the form of protein sequence or on a structure basis. For the present study, input was submitted in the form of protein FASTA sequence.

NetSurfP server predicts the surface accessibility and secondary structure of amino acids on the basis of their sequence (http://www.cbs.dtu.dk/services/NetSurfP/). The reliability of this prediction method is in the form of Z-score. The Z-score highlights the surface prediction reliability, but not associated with secondary structure [25].

PSIPRED is a secondary structure prediction method based on neural networks (http://globin.bio.warwick.ac.uk/psipred/). Analysis performed by PSIPRED achieved an average Q score of 80% [26]. Protein FASTA sequence was submitted as input query for this tool.

2.4. Prediction of various post translational modifications

Post translational modifications affect a variety of important biological processes including cell signaling, metabolic pathways *etc*. Putative phosphorylation sites were predicted using NetPhos 2.0 (http://www.cbs.dtu.dk/services/NetPhos) and GPS 2.1 (http://gps.biocuckoo.org/) [27–28]. Putative ubiquitylation sites were predicted using the UbPred (www.ubpred.org) and BDM-PUB (bdmpub.biocuckoo.org) programs [29]. Putative sumoylation sites and palmitoylation sites were predicted using respective SUMOsp 2.0 (http://sumosp.biocuckoo.org/) and CSS-Palm softwares programs (http://csspalm.biocuckoo.org/).

2.5. Modeling the phenotypic effects of nsSNPs on protein structure

MUSTER (v1.0), a valuable threading tool, was used for protein structure prediction (http://zhanglab.ccmb.med.umich.edu/MUSTER/). It provided the Z-score and complete full length models by using MODELLER v8.2. Sequence-derived profiles, secondary structures, structured-derived profiles, solvent accessibility, backbone torsion angles and hydrophobic scoring matrix are the six different sources used by MUSTER [30].

I-TASSER creates the full length protein models by excising continuous fragments from threading alignments and further reconstructs them using replica exchanged Monte Carlo simulations (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). The quality of predicted structure is estimated by I-TASSER in the form of confidence score. These are used to predict the quality of the modeling prediction by calculating the distance between two predicted models [31].

Swiss-PDB Viewer (v4.04) was used to generate the mutated models of each of the selected PDB entries for the corresponding amino acid substitutions. Swiss-PDB Viewer allows browsing through a rotamer library to change amino acids. A "mutation tool" was used to replace the native amino acid with a new one. The mutation tool facilitates the replacement of the native amino acid by the "best" rotamer of the new amino acid. The energy minimization for 3D structures was performed by the NOMAD-Ref server. This server uses Gromacs as the default force field for energy minimization based on the methods of steepest descent, conjugate gradient, and L-BFGS [32]. The galectin-3 PDB file named 1KJR was used as input for generating mutated models.

2.6. Predicting molecular effects of regulatory SNPs

PolymiRTS (Polymorphism in miRNA target site) is a web based server used for analyzing the functional impact of genetic polymorphisms in microRNA (miRNA) seed regions and miRNA target sites (http://compbio.uthsc.edu/miRSNP). This program also provided the information regarding miRNA-mRNA binding sites and small INDEL variations in miRNA seed regions and miRNA target sites [33]. The reference sequence ID NM_0023006 for *LGALS3* was submitted as input for this server.

SNPinfo server is a set of web based various selection tools including Gene pipe, Genome pipe, Linkage pipe, Taq SNP, Func Pred, SNPseq, which were used to select functional SNP for genetic association studies (http://snpinfo.niehs.nih.gov/). This algorithm predicts functional impact of both coding and non-coding SNPs based on GWAS and candidate gene information [34]. *LGALS3*, the name of gene was submitted as a query for this program.

RegulomeDB was used to identify putative regulatory potential functional variants (F:/in%20silico/regulome%20db/rs1009977.html). It used high-throughput experimental data sets from ENCODE and other sources, as well as computational predictions and manual annotations to identify functional determinants. These data sources are combined into a powerful tool that scores variants to help to separate functional variants from a large pool and provides a small set of putative sites with testable hypotheses as to their function [35]. The reference sequence IDs of all non-coding SNPs were submitted as query input.

3. Results

As per dbSNP database, a total of 1130 validated and non-validated SNPs were reported for *LGALS3* gene. Out of these, only 610 validated SNPs were investigated further for computational analyses. These included 54 coding SNPs, 2 SNPs harboring 5′UTR, 51 of 5′ near genes, 3 of 3′UTR SNPs, 16 of 3′near genes and 484 were intronic SNPs (Fig. 2). Out of 54 coding SNPs, 16 were found to be synonymous SNPs and 38 were nsSNPs. Among nsSNPs, one was non-sense variant, one frameshift variant and rest others were mis-sense variants.

3.1. Deleterious or disease associated variations and their effects on protein structure, stability and solvent accessibility

Out of 36 mis-sense nsSNPs analyzed by SIFT, 10 variations (5 of N-terminal region and 5 of CRD) were found to be deleterious with

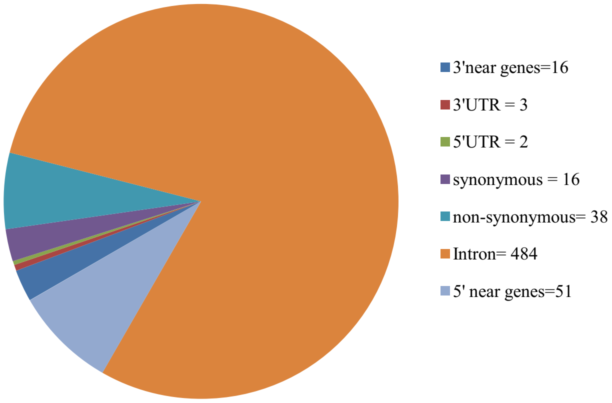


Fig. 2. Pie diagram depicting count of validated SNPs analyzed for computational analysis.

tolerance index < 0.05 and rest all variations were shown to be tolerated with tolerance index ≥ 0.05 (Table 1). The basic amino acids arginine (R), histidine (H) and lysine (K) were the most likely substituted amino acids as deleterious impact of SNPs. Analysis by Polyphen predicted the probably damaging effect of 13 SNPs. Furthermore, only one substitution L11I was found to have possibly damaging effect. Rest all variations showed benign effect on protein. A total of 13 variations were found to be disease associated as predicted by nsSNPAnalyzer. In consonance with Polyphen and SIFT results, seven variations were found to have disease associated effects. However, in contrast to Polyphen and SIFT, neutral effects of G47R and P64H were predicted by nsSNP analyzer. Prediction of functional nsSNPs was also confirmed using SNPs & GO and PANTHER. The variations R162C, R162H, E185G, P191L and R224Q were found to have disease associating effects by both of these tools (Table 1). Possible molecular mechanism by which a deleterious variation associated with disease state was further predicted by Mutpred. Majority of predicted effects included alteration in methylation status and glycosylation, change in solvent accessibility and altered disorderness in secondary structure of protein. Deleterious and disease associated variations along with their predicted effect is given in Table 1. Decrease stability of protein with respective energy changes was observed for all deleterious mutations except G38R and P191L (Table 2). NetSurfP server was used for prediction of surface accessibility and secondary structure of amino acids. Class assignments of 'buried' and 'exposed' were given to amino acids as per their location in protein structure. Out of various nsSNPs submitted, class assignment had been changed for only three mutations including A97P (exposed to buried), L114V (buried to exposed) and E185G (buried to exposed). The other SNPs did not show any alteration in their respective class assignment (Suppl Table 1).

3.2. Prediction of putative methylation, phosphorylation, sumolyation and ubiquitination sites

Besides Mutpred, various other tools were used to identify deleterious variations having putative phosphorylation, ubiquitination *etc.* sites. The four putative phosphorylation sites harboring deleterious variations were predicted by GPS 2.1. These included serine/threonine phosphorylation at 98, 188 and 237 positions. The mutations at these sites resulted in loss of respective phosphorylation. In contrast to this, only one phosphorylation site at position 188, was predicted by Netphos. Amino acid substitution at positions 153 and 183 resulted in significant gain of ubiquitination, while no ubiquitination sites were observed with respective wild type amino acids at these positions as predicted by Ubpred. No sumolyation and palmitoylation sites were observed at wild type amino acids as well as their substituted counterparts in galectin-3 molecule. Different variations harboring deleterious effects as predicted by above said tools as well as by Mutpred are depicted in Fig. 3.

Table 2
Alteration in protein stability due to SNP change.

SNP ID	Amino-acid change	Stability	Free energy change (kcal/mol)
rs568100150	L11I	Decrease	−0.59
rs199980622	G38R	Increase	0.46
rs200440596	P46R	Decrease	−0.66
rs201423229	G47R	Decrease	−0.14
rs4644	P64H	Decrease	−0.53
rs372168966	P67L	Decrease	−0.83
rs370418608	I132V	Decrease	−0.02
rs376506811	R162C	Decrease	−0.85
rs201865041	R162H	Decrease	−2.07
rs542583325	E185G	Decrease	−1.37
rs373019488	P191L	Increase	1.33
rs564378578	R224Q	Decrease	−0.74
rs540554467	L234R	Decrease	−1.74
rs1042869	S237F	Decrease	−1.53
rs150161752	I240T	Decrease	−2.49
rs138668217	Y247H	Decrease	0.10

3.3. Modeling of amino acid substitution effects due to nsSNPs on protein structure, energy minimization

The available structure of *LGALS3* was retrieved from RCSB-Protein data bank with ID 1KJR. The structure accounted only of CRD harboring 113–250 amino acid residues with this ID. The structure of N-terminal region has not been reported in PDB database till date. Only those mutations which were found to have

Table 1
Deleterious and disease associated variations predicted by various softwares.

SNP ID	Amino acid change	SIFT	Polyphen	nsSNP analyzer	SNP &GO	PANTHER	Predicted Effect By Mutpred
rs568100150*	L11I	–	Possibly damaging	Disease	Neutral	–	No significant effect
rs199980622*	G38R	Deleterious	Probably damaging	Disease	Neutral	–	Gain of methylation, loss of relative solvent accessibility
rs200440596*	P46R	Deleterious	Probably damaging	Disease	Neutral	–	Gain of methylation, loss of glycosylation
rs201423229*	G47R	Deleterious	Probably damaging	Neutral	Neutral	–	Gain of methylation, loss of relative solvent accessibility
rs4644*	P64H	Deleterious	Probably damaging	Neutral	Neutral	–	Loss of relative solvent accessibility, loss of glycosylation
rs372168966*	P67L	Deleterious	Probably damaging	Disease	Neutral	–	Loss of relative solvent accessibility, loss of glycosylation
rs370418608	I132V	Deleterious	Benign	Neutral	Neutral	Neutral	No significant effect
rs376506811	R162C	Deleterious	Probably damaging	Disease	Disease	Deleterious	Loss of disorder
rs201865041	R162H	Deleterious	Probably damaging	Disease	Disease	Deleterious	No significant effect
rs542583325	E185G	–	Probably damaging	Disease	Disease	Deleterious	Loss of MoRF binding
rs373019488	P191L	Deleterious	Probably damaging	Disease	Disease	Deleterious	Loss of disorder, loss of methylation
rs564378578	R224Q	–	Probably damaging	Disease	Disease	Deleterious	No significant effect
rs540554467	L234R	–	Probably damaging	Disease	Neutral	Neutral	Gain of disorder, Gain of methylation, Gain of MoRF binding, loss of stability
rs1042869	S237F	Tolerated	Probably damaging	Disease	Neutral	Neutral	Loss of disorder, loss of glycosylation
rs150161752	I240T	Deleterious	Probably damaging	Disease	Neutral	Neutral	Gain of disorder, loss of stability
rs138668217	Y247H	Tolerated	Benign	Disease	Neutral	Neutral	Gain of disorder

* Variations lie in N-terminal domain, other variations lie CRD region; – not determined by software.

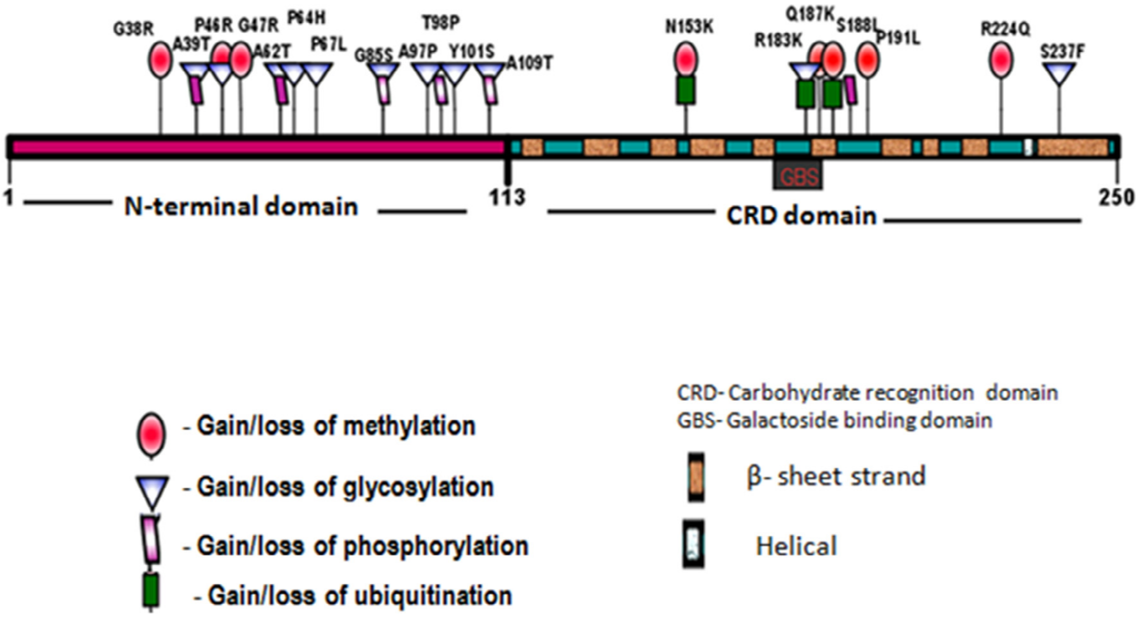

G38R
A39T
P46R
G47R
A62T
P64H
P67L
G85S
A97P
T98P
Y101S
A109T
N153K
R183K
Q187K
S188L
P191L
R224Q
S237F
1
N-terminal domain
113
GBS
CRD domain
250
- Gain/loss of methylation
- Gain/loss of glycosylation
- Gain/loss of phosphorylation
- Gain/loss of ubiquitination
CRD- Carbohydrate recognition domain
GBS- Galactoside binding domain
β- sheet strand
Helical

Fig. 3. Variations harboring putative methylation, glycosylation, phosphorylation and ubiquitination sites (drawn by IBS software).

deleterious effect by at least three tools were modeled to predict the effect of substitution on structure of protein. On this basis, a total of 9 variations were predicted to be deleterious, which involved 3 of N-terminal region and 6 harbored in CRD. An effort was made to predict the complete secondary structure of galectin-3 protein (including N-terminal region and CRD) so as to model the substitution effect for both N-terminal and CRD variations through homology modeling. MUSTER provided the complete full length models along with coverage, sequence identity and Z-score using MODELLER v8.2. All the templates provided by MUSTER had coverage of approximately 0.5 with alignment of only CRD sequence (PDB ID- 1KJR) (Suppl Table 2). However, no template was found for N-terminal sequence. So modeling of amino-acid substitutions was performed only for CRD variations.

Six functional mutations of CRD including rs376506811 (R162C), rs2018650419 (R162H), rs542583325 (E185G), rs373019488 (P191L), rs564378578 (R224Q) and rs150161752 (I240T) were subjected to Swiss-PDB Viewer mutation tool. Significant structural changes were observed in mutated models of R162C and R162H in comparison to their native structure. The variant R162C possessed a significant substitution from a large basic amino acid to small polar amino acid. Molecular simulations analysis revealed that R162C (−9.275 kcal/mol) and R162H (−9.270 kcal/mol) showed accountable difference in their total energy after energy minimization as compared to native structure (−9.564 kcal/mol).

3.4. Prediction of putative miRNA target sites

3′UTR serve as putative target sites for miRNA, an important regulator of gene expression. Single nucleotide change in these regions may either disrupt or create new target sites for miRNA. The SNPs rs61975414, rs1042906, rs1825065 and rs1042918 were predicted to affect the miRNA target sites. Besides this, rs149894805, an INDEL variant was also found to alter the miRNA site. The detailed impact of these SNPs on various miRNAs was listed in Table 3. Two variations rs148982635 and rs200056596 were also found to exist in miRNA seeds which resulted in disruption of their putative target sites.

Table 3
List of SNPs spanning miRNA target sites and seeds.

dbSNP ID	Type of variant	Allele	miR ID	Function class*
Variants in miRNA target sites				
rs61975414 (T/A)	SNP	T	hsa-miR-4727-5p	C
			hsa-miR-4778-5p	
			hsa-miR-5001-3p	
			hsa-miR-6738-3p	
rs1042906 (C/T)	SNP	C	hsa-miR-376a-5p	C
			hsa-miR-4760-5p	
			hsa-miR-8061	
		T		
			hsa-miR-10a-3p	D
rs1042918 (A/C)	SNP	A	hsa-miR-128-3p	D
			hsa-miR-216a-3p	
			hsa-miR-27a-3p	
			hsa-miR-27b-3p	
			hsa-miR-3681-3p	
		C	hsa-miR-513a-5p	C
			hsa-miR-6750-5p	
			hsa-miR-6822-5p	
rs182580267(T/C)	SNP	C	hsa-miR-6757-5p	C
rs149894805(-/AT)	INDEL	AT	hsa-miR-1279	C
			hsa-miR-3672	
			hsa-miR-4461	
			hsa-miR-6864-3p	
Variants in miRNA seeds				
rs148982635(G/A)	SNP	G/A	hsa-miR-4749-3p	D
rs200056596(C/T)	SNP	C/T	hsa-miR-4749-3p	D

* C – site created, D – site disrupted.

3.5. SNPs affecting transcription factor binding site

The non-coding regions including intronic and UTR of *LGALS3* serve as putative binding sites for transcription factors as well as splicing. A single nucleotide change at these positions may alter the binding and subsequently affect transcription or splicing mechanisms. Out of plethora of SNPs tested, 43 SNPs having minor allele frequency ≤ 1% were predicted to affect transcription binding as indicated by their respective RegulomeDB score. Only those SNPs which were having scores ≤ 4 are listed in Table 4. RegulomeDB

Table 4
Functional impact of non-coding variations.

SNP /INDEL	SNPs harboring TFBS*			
	Regulomedb			SNPinfo
	Regulomedb score	Category	Description	
rs12161901	4	Minimal binding evidence	TF© binding+any motif+DNase peak	Y
rs1009977				
rs1009978				
rs8012397				Y
rs17128230				
rs61975412				Y
rs368553467				
rs1047556				
rs57970196				
rs73271727				
rs3832943	3a	Less likely to affect binding	TF binding+any motif+DNase peak	
rs369628908				
rs2075598	2b	Likely to affect binding	TF binding+any motif+DNasefootprint+DNase peak	

* TFBS – Transcription factor binding site, ©TF – Transcription factor.

score of 2b was achieved by only single SNP rs2075598 with likely effect on transcription factor binding. Two SNPs rs3832943 and rs369628908 were found to have 3b score conferring less likely effect on binding. The other 10 SNPs listed in Table 4 were recorded to have score of 4 with minimal binding evidence. The same attribute of SNPs to affect transcription factor binding site (TFBS) was also predicted by SNPinfo. The results indicated that out of 13 functional SNPs predicted by RegulomeDB, only three SNPs i.e. rs12161901, rs8012397 and rs 61975412 were observed to affect TFBS in SNPinfo (Table 4).

4. Discussion

A plethora of SNPs has been distributed in human genome. Due to their association with different diseases, these variations can preferentially act as genetic markers [36]. Studying such a large number of SNPs in case-control association studies offers a great challenge for scientists. Computational analyses provide a major insight to predict those SNPs which can potentially affect the structure and function of the encoded protein in any way. Galectin-3 is an important molecule with well implicated role in various pathologies including cancer and inflammatory diseases [8–9,17]. So, in the present study an effort was made to identify functionally important non-synonymous and regulatory SNPs in human *LGALS3* gene.

A systematic approach including both sequence based and structure based study was undertaken for computational analyses of nsSNPs. Sequence based approach offered an advantage that it is suitable for proteins having closely related members. Based on this approach, both N-terminal and CRD harbored variations of *LGALS3* were investigated. Out of 36 nsSNPs, a total of nine variations were found to be deleterious by at least three tools. Decrease in protein stability due to these variations further potentiates their functional impact. Analysis of these variations by Ensembl indicated that all these variations were found to be expressed in conserved regions. It has been evidenced that conserved regions are biologically very important, so variations in these regions may lead to potential functional changes. The adverse effects of these variations were also observed on protein stability. A couple of variations were also observed to alter methylation, glycosylation and other post translational modifications. Significant gain or loss of methylation had been observed in G38R, P46R, P47R, P191L and L234R variations. DNA methylation is a key regulator in transcription and altered effect of methylation behavior has been implicated in many diseases like cancer, atherosclerosis, aging *etc*. [37–38]. The evidences of hyper-methylation in galectin-3 gene also come from the various studies in cancer [39–41]. Another SNP rs4644 (P64H) was also identified as deleterious by SIFT and Polyphen tools. It is a conserved variation harbored in N-terminal region which involves substitution of C with A, replacing proline with histidine in galectin-3 protein. His^{64} renders galectin-3 molecule susceptible towards cleavage by matrix metallo-proteinases (MMPs) resulting in loss of self association property and alterations in glycan binding properties [42]. This variation was also indicated to result in loss of glycosylation and relative solvent accessibility. Thus, it was predicted to severely affect the biological functions leading to various anomalies. This finding of present study was strongly evidenced from experimental data conducted in our lab as well from perusal of literature indicating genetic association of P64H with rheumatoid arthritis and breast cancer [15,43–44]. As far, no such experimental data has been available in literature for other variations predicted to be deleterious by the present study. So, the results of present study may provide a guidelines for prioritizing the SNPs for further genetic association studies.

Further, sequence based approach was unable to explain the underlying mechanism to study genotype-phenotype relationships. So, another approach based on structure was undertaken to predict the effect of variations on secondary structure, surface accessibility and binding properties of protein. As far, 3-D structure of galectin-3 protein is available only for CRD, not for N-terminal region. So, the structure based predictions were implemented only for variations harboring CRD. β-pleated sheets along with random coils attributed towards secondary structure of galectin-3 protein as predicted by PSIPRED. No change in 2^{0} structures has been observed for R162C, R162H and P191L variations. Furthermore, NetSurfP tool revealed that surface accessibility of these SNPs in secondary structure remains unaffected due to amino acid substitutions. The putative ligand binding site in galectin-3 protein harbored Arg 162 and polymorphism at this position, either R162C or R162H, may affect its binding properties. Taken this in consideration, R162C and R162H were proposed as potential candidate targets for various genetic association studies.

The present study predicted five SNPs in miRNA target site and two SNPs in miRNA seed region. These SNPs may alter miRNA binding site and hence affect subtle gene regulation and it may ultimately lead towards disease susceptibility. The evidence comes from recent experimental studies indicating SNPs in miRNA target sites of *BRCA1*, *TGF-β* genes subsequently contributing towards likelihood of cancer [45–46]. As far, no such experimental data has been reported for *LGALS3* gene. Thus, non-coding SNPs also emerged as important target for genome wide disease association studies affecting miRNA target sites, a novel approach for subtle gene regulation. SNPs in regulatory binding sites may affect the expression of gene and altered expression may lead to pathological conditions. Three SNPs were found to affect TFBS predicted by two different tools independently. The regulatory SNPs of *AKT3, ATF3, VEGFA etc*. were found to play an important role in susceptibility towards cancer [14]. The literature till date documented no such experimental study for regulatory SNPs of galectin-3. Thus, the use of *in silico* SNP prioritization strategies provides an excellent framework for the identification of functional SNPs. As these computational tools are machine based algorithms, so, further validation with experimental investigations and clinical evidences are warranted to complement the findings of the present study.

5. Conclusion

In silico analyses estimated the probability with which coding as well as non-coding variations are likely to have potential impact either on protein structure, function or its regulation. These functional genetic determinants can also act as potential diagnostic and therapeutic targets for different diseases. Some findings of the present study have already been experimentally validated by our lab as well as other research groups. Follow up studies to elucidate the role of other functional SNPs of *LGALS3* gene in the etiology of complex diseases are underway in our lab.

Competing interests

The authors declare no conflicts of interest.

Acknowledgements

This work was supported by the grant of Department of Science & Technology (DST), New Delhi, India [SB/FT/LS-184/2012] and University Grants Commission (UGC), New Delhi, India [F.17-74/2008(SA1)].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejmhg.2017.02.001.

References

[1] Kadrofske MM, Openo KP, Wang JL. The human LGALS3 (galectin-3) gene: determination of the gene structure and functional characterization of the promoter. Arch Biochem Biophys 1998;349:7–20.
[2] Berbís MÁ, André S, Cañada FJ, Pipkorn R, Ippel H, Mayo KH, et al. Peptides derived from human galectin-3 N-terminal tail interact with its carbohydrate recognition domain in a phosphorylation-dependent manner. Biochem Biophys Res Commun 2014;443:126–31.
[3] Byrd JC, Mazurek N, Bresalier RS. Post-translational modification of galectin-3 and its role in biological function. In: Klyosov AA, Traber PG, editors. Galectins and disease implications for targeted therapeutics. American Chemical Society; 2012. p. 137–51.
[4] Dumic J, Dabelic S, Flögel M. Galectin-3: an open-ended story. Biochim Biophys Acta 2006;760:616–35.
[5] Ochieng J, Furtak V, Lukyanov P. Extracellular functions of galectin-3. Glycoconj J 2004;9:527–35.
[6] Krześlak A, Lipińska A. Galectin-3 as a multifunctional protein. Cell Mol Biol Lett 2004;9:305–28.
[7] de Oliveira FL, Gatto M, Bassi N, Luisetto R, Ghirardello A, Punzi L, et al. Galectin-3 in autoimmunity and autoimmune diseases. Exp Biol Med (Maywood) 2015;240:1019–28.
[8] Dondoo TO, Fukumori T, Daizumoto K, Fukawa T, Kohzuki M, Kowada M, et al. Galectin-3 is implicated in tumor progression and resistance to anti-androgen drug through regulation of androgen receptor signaling in prostate cancer. Anticancer Res 2017;37:125–34.
[9] Radosavljevic G, Volarevic V, Jovanovic I, Milovanovic M, Pejnovic N, Arsenijevic N, et al. The roles of Galectin-3 in autoimmunity and tumor progression. Immunol Res 2012;52:100–10.
[10] Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variations. Genomic Res 1998;8:1229–31.
[11] Joshi BB, Koringa PG, Mistry KN, Patel AK, Gang S, Joshi CG. In silico analysis of functional nsSNPs in human TRPC6 gene associated with steroid resistant nephrotic syndrome. Gene 2015;572:8–16.
[12] Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 2007;39:1329–37.
[13] Kucukkal TG, Petukh M, Li L, Alexov E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Curr Opin Struct Biol 2015;32:18–24.
[14] Buroker NE. Regulatory SNPs and transcriptional factor binding sites in ADRBK1, AKT3, ATF3, DIO2, TBXA2R and VEGFA. Transcription 2014;5:e964559.
[15] Kaur T, Sodhi A, Singh J, Arora S, Kamboj SS, Kaur M. Evaluation of galectin-3 genetic variants and its serum levels in Rheumatoid Arthritis in North India. Int J Hum Genet 2015;15:131–8.
[16] Hernández-Romero D, Vílchez JA, Lahoz Á, Romero-Aniorte AI, Jover E, García-Alberola A, et al. Galectin-3 as a marker of interstitial atrial remodelling involved in atrial fibrillation. Sci Rep 2017;7:40378.
[17] Saccon F, Gatto M, Ghirardello A, Iaccarino L, Punzi L, Doria A. Role of galectin-3 in autoimmune and non-autoimmune nephropathies. Autoimmun Rev 2017;16:34–47.
[18] Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 2006;7:61–80.
[19] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[20] Bao L, Zhou M, Cui Y. NsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 2005;33:W480–W4802.
[21] Schwarz DF, Hädicke O, Erdmann J, Ziegler A, Bayer D, Möller S. SNPtoGO: characterizing SNPs by enriched GO terms. Bioinformatics 2008;24:146–8.
[22] Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res 2006;34:W645–50.
[23] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009;25:2744–50.
[24] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–10.
[25] Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;9:51.
[26] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–5.
[27] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Mol Biol 1992;94:1351–62.
[28] Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. GPS, 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 2008;7:1598–608.
[29] Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, et al. Identification, analysis and prediction of protein ubiquitination sites. Proteins 2010;78:365–80.
[30] Wu S, Zhang Y. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 2008;72:547–56.
[31] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12:7–8.
[32] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–23.
[33] Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res 2014;42:D86–91.
[34] Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res 2009;37:W600–5.
[35] Boyle AP. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;22:1790–7.
[36] Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 2002;34:275–305.
[37] Zaina S, Heyn H, Carmona FJ, Varol N, Sayols S, Condom E, et al. DNA methylation map of human atherosclerosis. Circ Cardiovasc Genet 2014;7:692–700.
[38] Robertson KD. DNA methylation and human disease. Nat Rev Genet 2005;6:597–610.
[39] Ruebel KH, Jin L, Qian X, Scheithauer BW, Kovacs K, Nakamura N, et al. Effects of DNA methylation on galectin-3 expression in pituitary tumors. Cancer Res 2005;65:1136–40.
[40] Ahmed H, Cappello F, Rodolico V, Vasta GR. Evidence of heavy methylation in the galectin 3 promoter in early stages of prostate adenocarcinoma: development and validation of a methylated marker for early diagnosis of prostate cancer. Transl Oncol 2009;2:146–56.

[41] Keller S, Angrisano T, Florio E, Pero R, Decaussin-Petrucci M, Troncone G, et al. DNA methylation state of the galectin–3 gene represents a potential new marker of thyroid malignancy. Oncol Lett 2013;6:86–90.

[42] Nangia-Makker P, Raz T, Tait L, Hogan V, Fridman R, Raz A. Galectin-3 cleavage: a novel surrogate marker for matrix metalloproteinase activity in growing breast cancers. Cancer Res 2007;67:11760–8.

[43] Miwa HE, Koba WR, Fine EJ, Giricz O, Kenny PA, Stanley P. Bisected, complex N-glycans and galectins in mouse mammary tumor progression and human breast cancer. Glycobiology 2013;23:1477–90.

[44] Balan V, Nangia-Makker P, Schwartz AG, Jung YS, Tait L, Hogan V, et al. Racial disparity in breast cancer and functional germ line mutation in galectin-3 (rs4644): a pilot study. Cancer Res 2008;68:10045–50.

[45] Quann K, Jing Y, Rigoutsos I. Post-transcriptional regulation of BRCA1 through its coding sequence by the miR-15/107 group of miRNAs. Front Genet 2015;6:242.

[46] Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, et al. Single nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res 2010;70:2789–98.