J Biosci (2021)46:61
DOI: 10.1007/s12038-021-00184-0

Multiple allelic associations from genes involved in energy metabolism were identified in celiac disease

Sandilya Bhagavatula[1], Pratibha Banerjee[1], Ajit Sood[2], Vandana Midha[3], B. K. Thelma[4] and Sabyasachi Senapati[1]*

[1]*Immunogenomics Laboratory, Department of Human Genetics and Molecular Medicine, School of Health Sciences, Central University of Punjab, Bathinda, Punjab, India*

[2]*Department of Gastroenterology, Dayanand Medical College and Hospital, Ludhiana, Punjab, India*

[3]*Department of Medicine, Dayanand Medical College and Hospital, Ludhiana, Punjab, India*

[4]*Department of Genetics, University of Delhi South Campus, New Delhi, India*

**Corresponding author (Email, sabyasachi1012@gmail.com)*

MS received 12 September 2020; accepted 28 May 2021

Energy metabolism is a critical factor that influences disease pathogenesis. Recent high-throughput genomic studies have enabled us to look into disease biology with greater details. Celiac disease (CD) is an inflammatory autoimmune disease where ~60 non-HLA genes were identified which in conjunction with HLA genes explain ~55% of the disease heritability. In this study we aimed to identify susceptibility energy metabolism genes and investigate their role in CD. We re-analysed published Immunochip genotyping data, which were originally analysed for CD association studies in north Indian and Dutch population. 269 energy metabolism genes were tested. Meta-analysis was done for the identified SNPs. To validate the functional implications of identified markers and/or genes, *in silico* functional annotation was performed. Six SNPs were identified in north Indians, of which three markers from two loci were replicated in Dutch. rs2071592 (P_{Meta}=5.01e−75) and rs2251824 (P_{Meta}=1.87e−14) from *ATP6V1G2-NFKBIL1-DDX39B* locus and rs4947331 (P_{Meta}= 9.85e−13) from *NEU1* locus were found significantly associated. Identified genes are key regulators of cellular energy metabolism and associated with several immune mediated diseases. *In silico* functional annotation showed significant biological relevance of these novel markers and genes. FDI approved therapeutics against *ATP6V1G2* and *NEU1* are currently in use to treat chronic and inflammatory diseases. This study identified two pathogenic loci, originally involved in energy metabolism. Extensive investigation showed their synergistic role in CD pathogenesis by promoting immune mediated enteric inflammation. Proposed CD pathogenesis model in this study needs to be tested through tissue-on-chip and *in vivo* methods to ensure its translational application.

Keywords. Association; celiac disease; Dutch; energy metabolism; immunochip; meta-analysis; north Indians; replication

Abbreviations: ATP6V1G2, ATPase H+ transporting V1 subunit G2; CD, Celiac disease; DDX39B, DExD-box helicase 39B; ECM, Extracellular matrix; EGFR, Epidermal growth factor receptor; eQTL, Expression quantitative trait locus; ERC, Elastin receptor complex; GPCR, G-protein-coupled receptor; GWAS, Genome-wide association study; HLA, Human leucocyte antigen; IL1B, Interleukin 1 beta; KEGG, Kyoto Encyclopedia for Genes and Genomes; LPS, Lipopolysaccharide; MAF, Minor allele frequency; MHC, Major histocompatibility complex; mTOR, Mammalian target for rapamycin; NEU1, Neuraminidase 1; NFKB, Nuclear factor kappa B; OR, Odds ratio; PPI, Protein–protein interaction; SNP, Single nucleotide polymorphism; TLR, Toll-like receptor; TNF-a, Tumor necrosis factor alpha; TREX, Transcription export complex; tTG2, Tissue transglutaminase.

Supplementary Information: The online version contains supplementary material available at https://doi.org/10.1007/s12038-021-00184-0.

http://www.ias.ac.in/jbiosci
Published online: 23 June 2021

1. Introduction

Celiac Disease (CD) is a common inflammatory disease of the gastrointestinal tract, triggered as an autoimmune reaction towards ingested gluten in genetically susceptible individuals. It is primarily characterized by malabsorption, inflammation and atrophy of gastrointestinal walls and small intestinal villi upon exposure to dietary gluten (Lebwohl and Rubio-Tapia. 2021). Complete exclusion of gluten has been proven to aid in the treatment of Celiac Disease (Gujral *et al*. 2012; Parzanese *et al*. 2017).

Deamidation of glutenin and gliadin peptides of gluten enhances their ability to bind to the HLA-DQ2 or HLA-DQ8 pockets on the macrophages in the lamina propria, where they are identified by the T-cell receptors as antigens (Biesiekierski 2017). This leads to humoral and cell mediated immune responses in the duodenum (Parzanese *et al*. 2017). There is a lack of complete explanation for the entire mechanism of celiac disease, as only 40% of the CD cases could be explained by the known HLA-DQ2/DQ8 alleles. Multiple genomic studies and meta-analyses have identified more than 60 non-HLA genes in CD which explain additional 10-12% of CD (Dieli-Crimi *et al*. 2015; Parzanese *et al*. 2017; Trynka *et al*. 2010). Several of the non-HLA genes have been reported to be linked to intercellular communication such as chemokine activity, interleukin production and antigen reception.

In recent studies functionally relevant mutations and polymorphisms in multiple genes have been identified to cause metabolic dysfunction leading to imbalances in energy homeostasis in many common diseases such as colorectal cancer, non-small cell lung cancer, prostate cancer, myocardial infarction, and multiple neurologic and metabolic disorders (Crous-Bou *et al*. 2012; DeBerardinis and Thompson 2012; Lee *et al*. 2016; Montén *et al*. 2015; Murtola *et al*. 2015). Recent evidence shows that many energy metabolism genes such as Glutaminase (GLS), Neurocalcin Delta (NCALD), and Insulin Receptor (INSR) have differential expressions relevant to nutrient signalling and energy homeostasis in CD pathogenesis. Redox signalling and oxidative stress was also found to significantly affect gastrointestinal mucosal health, which is a critical signature in CD (Bhattacharyya *et al*. 2014; Pérez *et al*. 2017). Endoplasmic reticulum-resident protein 57 (Erp57), a member of protein disulfide isomerise family containing thioredoxin motifs majorly implicated in protein folding of the major histocompatibility complex, has been found to oxidise human transglutaminase 2 (TG2), a key enzyme in gluten metabolism (Yi *et al*. 2018). GWAS studies have also pointed out multiple genes implicated in energy metabolism processes namely GLB1, PFKFB3, TH, and TREH, to be associated directly with CD. This sufficiently establishes the direct involvement of energy metabolic genes in several diseases including in CD (Trynka *et al*. 2011; Östensson *et al*. 2013; Garner *et al*. 2014).

Hence, we hypothesised that there are additional novel genes involved in energy metabolism that are associated with CD. In this study we looked into the published CD association dataset in north Indians and ethnically distinct European (Dutch) cohorts to identify any novel genes involved in energy metabolism. The datasets were generated using Illumina Immunochip platform, enriched for around 200 known immune disease associated loci and has been used extensively for fine mapping studies (Trynka *et al*. 2011). Study findings would contribute in basic understanding of CD, and could have translational value in personalized and predictive medicine.

2. Methods

2.1 *Selection of genes for the study*

An exhaustive list of genes involved in energy metabolism was re-constructed from the public database Kyoto Encyclopedia of Genes and Genomes – Pathway database (KEGG pathway database). Study genes were selected from the energy metabolic pathways leading to production of energy such as carbohydrate, lipid, amino acid, and other glycan metabolism pathways; and the energy homeostasis and energy transferring pathways involved in energy homeostasis and energy metabolism, irrespective of their tissue of expression, cellular localization, and mechanism of action. Alternate gene names and gene IDs were confirmed from NCBI-Gene and Ensembl. UniProt and GeneCards databases were used to reconfirm and authenticate their cellular and molecular functions.

2.2 *Genotype data*

A previously published association data of north Indian CD association (using Illumina Immunochip) was available in-house (Senapati *et al*. 2015, 2016). Details of the informed consent and ethical approval for the original study were mentioned elsewhere (Trynka *et al*.

2011; Senapati *et al*. 2015, 2016). Required, ethical approval (CUPB/CC/16/931) was also taken from Institutional Ethics Committee of Central University of Punjab to perform the data analyses included in this article. To perform replication, Dutch CD association statistics was used. Dutch genotype data for CD was obtained from Prof. Cisca Wijmenga's laboratory, University Medical Center, Groningen, The Netherlands (as part of Coeliac Disease Immunochip Consortium). Allelic association summary statistics was used in the current study. Detailed characteristics of study subjects and genotype quality are available elsewhere (Trynka *et al*. 2011; Senapati *et al*. 2015 2016).

Illumina Immunochip annotation file (hg19) was used to retrieve genetic variants (SNPs) mapped against the list of selected genes mentioned above. Genomic coordinates of all the selected markers were updated from hg19 to hg38 using LiftOver tool in UCSC genome browser. SNP rsIDs of each variant were also updated using 1000 Genomes genotype data available in Ensembl.

2.3 *Association statistics*

For each of the index SNPs, reference/minor alleles, minor allele frequency (MAF), allelic association p-value and odds ratio (OR) with 95% CI were retrieved from available CD association data. For further analyses we selected all the SNPs with association p-value < 0.01, having MAF $> 10\%$ to obtain approximately 90% power to detect true associations. Much relaxed p-values were taken into consideration in order to retain large number of genes involved in the molecular pathways or biological processes that might be involved in CD pathogenesis. Genotype data from Dutch CD dataset was used to perform cross-ethnic replication of all the markers identified in north Indians. Conventional 5% level of significance was used to determine the replication of each of these SNPs. Further, to improve the study power, eliminate type-II errors, meta-analysis was performed to estimate the combined association of these markers. PLINK 1.9 (Purcell *et al*. 2007) whole genome association analysis tool was used to perform meta-analysis.

To explore the role of the identified genes in human diseases, cross-disease association was checked by scanning all the published GWAS reports. NHGRI-EBI GWAS catalogue (*https://www.ebi.ac.uk/gwas/*), an open source up-to date public database, was used to retrieve the relevant data.

2.4 *In silico functional annotation of identified genetic variants*

Physical and molecular annotations of the associated markers were done as per RefSeq references and UniProtKB/SwissProt respectively. Regulatory significance of the SNPs were obtained from HaploRegv4.1 (Ward and Kellis 2012), RegulomeDB (Boyle *et al*. 2012) and ENCODE (ENCODE project consortium 2012) database. SNPs were evaluated for their presumed contribution based upon their genomic location in epigenetic modifications, transcription factor binding, chromatin remodelling, any histone modifications (H3K4me1/3 and/or H3K27ac), and CTCF binding. From ENCODE, tissue specific epigenetic regulation data was evaluated for T-cell, B-cell and small intestine only, which are directly involved in the CD pathogenesis.

To further understand their functional importance, GTEx Portal v7 (GTEx Consortium 2013) was used to obtain tissue specific eQTL values to evaluate the influence of the identified SNPs on the regulation of expression on the nearby genes. As CD is a systemic disease that primarily affects small intestine, eQTL was evaluated for whole blood and small intestine only. We limit our tissue specific expression regulation potential of identified SNPs within blood cells (whole blood or T- and B-cells) and small intestine to investigate directly involved tissues in CD.

2.5 *Identification of molecular interactions and networks*

Significance of identified genes in biological processes were estimated by evaluating molecular networks and protein–protein interactions (PPIs). The analyses were performed using NetworkAnalyst 3.0, a tool based on Walktrap algorithms (Xia *et al*. 2015), capable of extracting and curating data available from multiple sources such as KEGG (Kanehisa and Goto 2000), Reactome (Matthews *et al*. 2009), Gene Ontology (Ashburner *et al*. 2000), and PANTHER (Thomas *et al*. 2003) databases. All the genes shortlisted in our study were used to investigate their functional implications, network enrichment, generic protein–protein interaction, and tissue specific co-expression. Inspection of the generated interactive gene-gene, and protein–protein network data was done by analysing the level of significance (adjusted P-values). Pathways reconstructed from the data taken from KEGG, Reactome, Gene Ontology (GO:

Table 1. Association statistics, in north Indians, replication status in dutch and meta-analysis results of six CD associated SNPs identified in this study. Positional annotation of the SNPs was recorded from RefSeq and dbSNP

				North Indian			Dutch			Meta-analysis	
Markers	Chr_bp (hg38)	Gene name	Minor Allele	MAF	P-value	Indian OR	MAF	P-Value	OR	P_{Meta}	OR
rs2071592	6:31547563	*ATP6V1G2-NFKBIL1-DDX39B*	T	0.29	6.09e-26	2.92 (2.39-3.56)	0.32	7.89e-51	2.94 (2.56-3.39)	5.01e-75	2.93
rs2251824	6:31544080		A	0.24	0.002	0.69 (0.55-0.87)	0.17	2.49e-13	0.49 (0.41-0.60)	1.87e-14	0.57
rs4947331	6:31853264	*NEU1*	T	0.13	0.0009	0.59 (0.43-0.80)	0.11	1.14e-10	0.45 (0.36-0.58)	9.85e-13	0.50
rs9267577	6:31845997	*C6orf48*	T	0.11	0.002	0.60 (0.43-0.83)	0.18	NA	NA	NA	NA
rs6068799	20:54130130	*CYP24A1*	G	0.19	0.003	0.69 (0.54-0.88)	0.10	0.90	0.99 (0.81-1.21)	0.30	0.83
rs17637745	4:42879180	*RN7SKP82*	T	0.11	0.01	0.67 (0.49-0.92)	0.02	NA	NA	NA	NA

Association statistics for rs9267577 and rs17637745 was unavailable in Dutch dataset.
MAF (Minor Allele Frequency),OR (Odds Ratio).

Biological Processes, GO: Molecular Function, GO: Cellular Components) and PANTHER (Biological Processes, Molecular Function, Cellular Components) were used in the analysis. Generic PPI data was generated using data from IMEx Interactome (Orchard *et al*. 2012). Tissue specific co-expression analysis was considered in specific tissues (whole blood or T- cells and B-cells only, and small intestine) upon considering that pathologically important tissues to have much significant modulations in the co-expression and interaction patterns. Adjusted p-values of the above data, which are generated on the basis of permutation analysis and corrections, were considered for identification of molecular interactions which were significant.

2.6 *Drug target analysis*

All the significantly associated genes that were replicated in our study were scrutinized for their possible interactions with any known drugs. DrugBank v5.1.2 database was utilized to identify known drugs or therapeutic compounds having proven role to act on these gene products (Wishart *et al*. 2018).

3. Results

3.1 *Association statistics and meta-analysis*

Immunochip dense genotyping association statistics obtained was used to study the association for 269 genes involved in energy metabolism. 1440 genetic variants from 269 gene were screened for the assessment of their associations. Association dataset for north Indian CD cohort comprised of 1227 individuals, which included 497 cases and 736 controls (Senapati *et al*. 2015 and 2016). Only six SNPs were found significantly associated with CD in north Indian patients ($p \leq 0.01$) (table 1). All of them were found independent from each other after pair-wise LD estimation ($r^2 \leq 0.2$). Two independent markers were identified from *ATP6V1G2-NFKB1L1-DDX39B* locus. One markers rs2071592 from *ATP6V1G2-NFKB1L1* was identified as statistically significant with p=6.09e−26, and odds ratio = 2.92 (2.39–3.56). This marker located in the first intron of *NFKB1L1* and nearly 2kb 5′ of *ATP6V1G2*. Second association signal rs2251824 (p=0.002) located within *ATP6V1G2-DDX39B* long non coding RNA. Other associated genes include *NEU1* (rs4947331, p=0.0009), *C6orf48*

(rs9267577, p=0.002), *CYP24A1* (rs6068799, p=0.003) and *RN7SKP82* (rs17637745, p=0.01). Apart from rs6068799 and rs17637745, all other markers located within extended MHC region of chr:6, however, pairwise linkage disequilibrium estimation (r^2 values) confirmed absence of detectable linkage disequilibrium (r^2<0.02) between any of these SNPs within the MHC region with any reported CD associated HLA alleles (DQ2 or DQ8) or their proxy SNPs. Cross-ethnic comparison was performed using Dutch CD cohort association data. Published Dutch dataset consists of 1150 CD patients and 1173 control samples. Three SNPs namely, rs2071592 (p=8.89e−51), rs4947331 (p=1.14e−10) and rs2251824 (p=2.49e−13) were found replicating in Dutch dataset. Comparative linkage disequilibrium analysis at region between rs2071592 and rs2251824 showed absence of any LD (r^2) in both north Indian and Dutch populations. Substantial differences in haplotypic sequence and frequencies were observed between these populations (figure 1). These observations established their independent association with CD. rs6068799 from *CYP24A1* was unable to replicate and markers from *C6orf48* and *RN7SKP82* were not available in the Dutch dataset. Three SNPs rs2071592 (P_{Meta}=5.01-e−75) and rs2251824 (P_{Meta}=1.87e−14) and rs4947331 (P_{Meta}= 9.85e−13) were found statistically significant after meta-analysis. Allelic directionality for all the identified SNPs was same in both in north Indians and Dutch (table 1).

All the identified genes were reported previously for their association with several diseases, including immune mediated diseases, inflammatory diseases, neuropsychiatric conditions, cancers, etc. (supplementary table 1). Multiple SNPs from each of these genes were found associated with these diseases. These genes were known to induce tissue inflammation and thus may pose considerable risk independent of HLA genes.

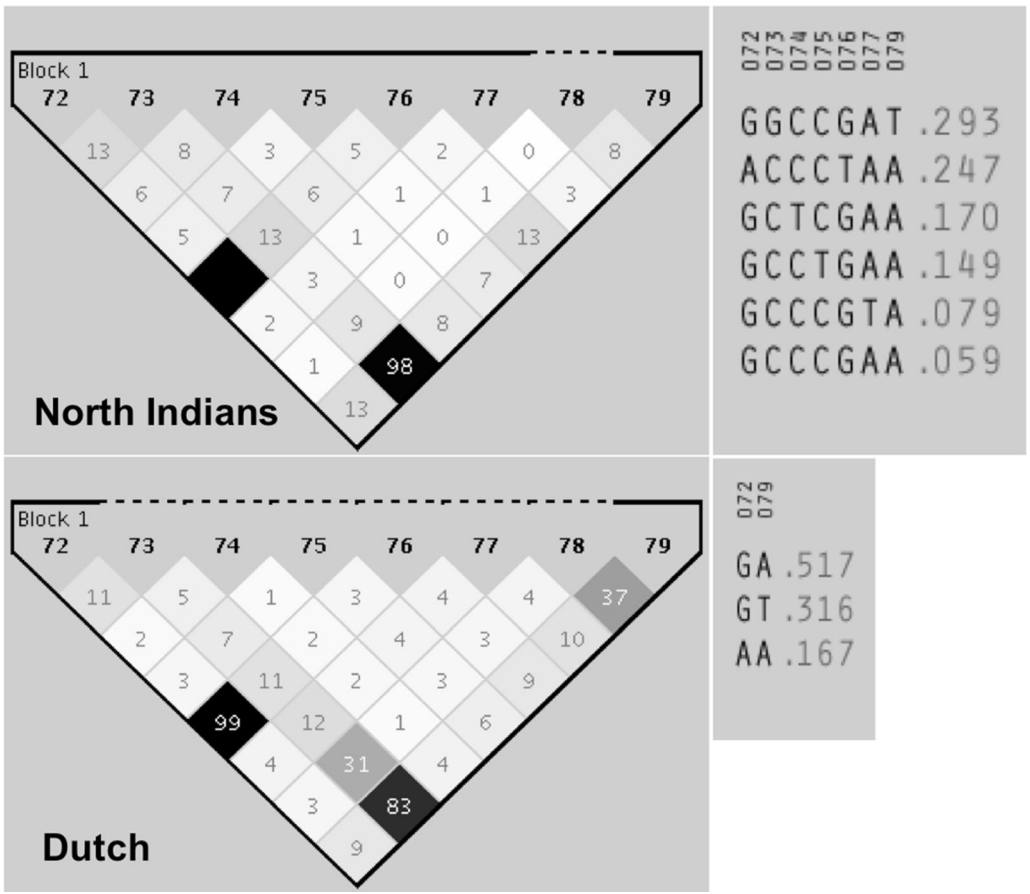

Figure 1. Linkage disequilibrium plots showing LD background in the genomic region between rs2071592 and rs2251824 at *ATP6V1G2-NFKBIL1-DDX39B*. It shows lack of LD at this locus in both north Indians and Dutch populations. Different haplotypic arrangements between these two populations were also observed.

3.2 *Functional annotation of the associated SNPs*

Among the top four significantly associated SNPs (rs2071592, rs4947331, rs2251824, rs9267577), localization of the variant is either intronic or very near to the UTRs (5′ and 3′), overlapping the regulatory regions that control transcript variants, and mRNA stability respectively (table 1 and supplementary table 2). Markers localized near the UTRs fall within a window of 4.5 kb from the transcription start site (TSS). While all the identified genes have relevant functional expressions in whole blood, only some of them have relevant expression levels lymph nodes, and small intestine as well (supplementary table 2). Additionally, tissue specific epigenetic signatures of each variant in T-cells, B-cells, and small intestine, identify significantly strong DNase I hypersensitivity, histone modifications (H3K4me1/3 and H3K23ac), and CTCF binding (zinc finger domain containing transcription factors) for both rs2071592 and rs2251824 from *ATP6V1G2-NFKBIL1-DDX39B* (supplementary table 3). Tissue specific eQTL data was screened from GTEx portal for each SNP and relevant data was available for three SNPs. Genes were identified to be expressed in at least one of the tissues relevant in prognosis of the disease (whole blood and small intestine). These SNPs were found to influence the expression of *DDX39B*, *HLA-DRB6*, *HLA-DRB5*, *MICB*, *HLA-DRB9*, *HLA-DQB2*, *HLA-DQA2*, *HCG27*, *C4B*, *LY6G5B*, and *HLA-DQB1-AS1*in whole blood and *HLA-DRB6* in small intestine (supplementary table **2**).

3.3 *Pathway enrichment analysis*

All of the genes identified (table 1), and genes which are in strong eQTL were tested together for their pathway enrichment from different databases. Resulting pathways were screened for association following multiple corrected significance levels ($p_{adj} \leq 0.01$). Multiple significant cellular pathways were identified with significant p-values such as: Antigen processing and presentation of peptide and polysaccharide antigens via MHC-II (p_{adj}=3.03e−5), Activation of immune response (p_{adj}=7.2e−5), PD-1 signalling (p_{adj}=7.5e−4), Translocation of Zap-70 to immunological synapse (p_{adj}=0.0006), etc. Associated genes were enriched for cellular components such as Vacuolar membrane (p_{adj}=2.69e−5), Vacuolar part (p_{adj}=6.06e−5) and Lysosomal membrane (p_{adj}=1.61-e−4) (supplementary table 4).

3.4 *Co-expression in target tissues*

Co-expressed genes in specific tissues were assessed for their enrichment (p_{adj}<0.001) in different tissue specific pathways. Pathways identified to consist of co-expressed genes in whole blood include Formation of a pool of free 40S subunits (p_{adj}=8.82e−47), Nonsense Mediated Decay Independent of the Exon Junction Complex (p_{adj}=6.59e−46), GTP hydrolysis and joining of the 60S ribosomal subunit (p_{adj}=6.59e−46), Metabolism of RNA (p_{adj}=4.23e−37), Structural constituent of ribosome (p_{adj}=2.88e−47) amongst several others. On the other hand, pathways of co-expressed genes in small intestine include Cell Adhesion Molecules (p_{adj}=1.84e−9), PD-1 signalling (p_{adj}=1.69e−8), Protein targeting to membrane (p_{adj}=1.68e−9), Antigen processing and presentation (p_{adj}=1.11e−10) amidst the other key regulators in the intestine (supplementary table 4).

3.5 *Identification of potential drug targets*

DrugBank database was screened to identify known drugs against the proteins translated by the genes identified in this study. Six approved drugs/therapeutic compounds were found to act on products of *ATP6V1G2* and *NEU1*. All these compounds are inhibitor in function and currently used to treat several diseases (supplementary table 5).

4. Discussion

Present study confirmed the significant association of three SNPs from two novel loci *ATP6V1G2-NFKBIL1-DDX39B* and *NEU1*. Suggestive associations were also observed for *C6orf48*, *CYP24A1* and *RN7SKP82*, which were identified in north Indians but failed to get replicated in Dutch dataset (table 1). These genes were found to be functionally involved in key regulatory mechanisms in energy metabolism. Further, functional annotation uncovered their cumulative role in CD pathogenesis.

Majority of the significant SNPs are located near the extended MHC region, however, none of them share any linkage with the same or other *HLA* markers associated with CD. This indicates their independent association to the disease. Additionally, many of these genes were reported as associated with several human traits and diseases, as identified from the reports listed in the GWAS catalogue, such as different Carcinomas,

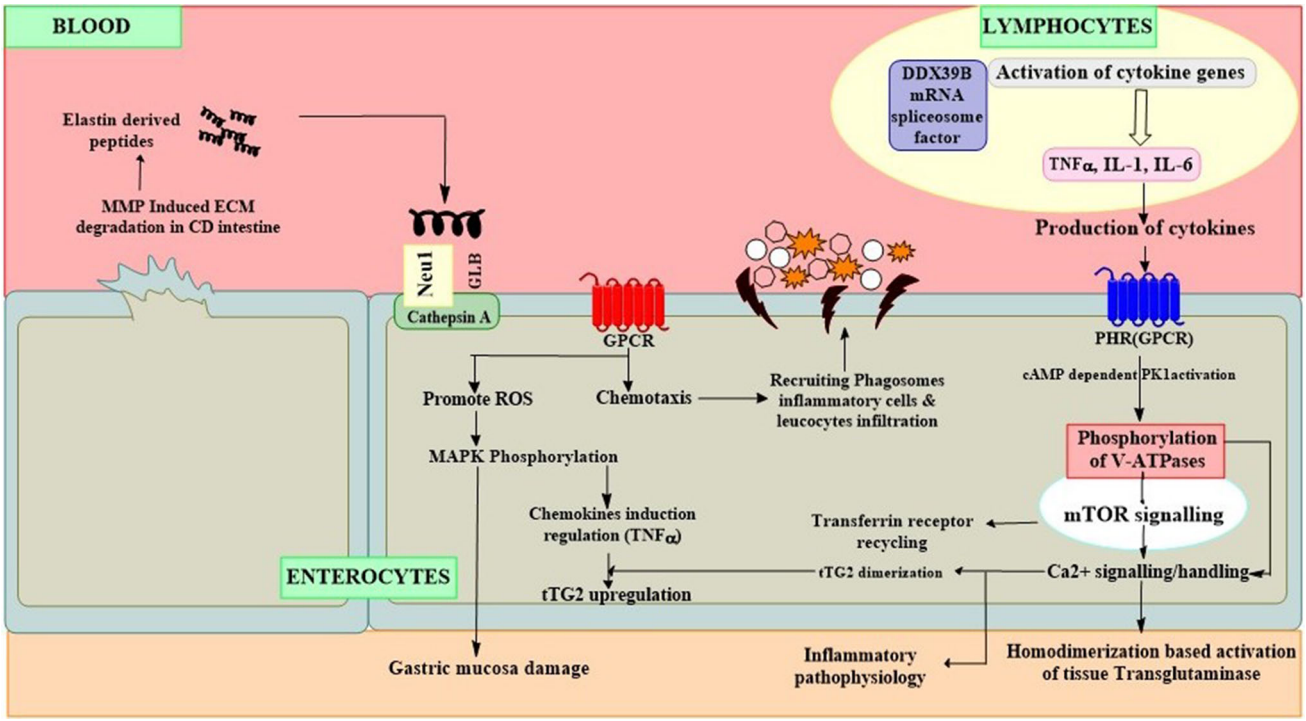

Figure 2. Synergistic mechanism of function of the novel genes identified as associated with celiac disease pathogenesis – Suggested Model. Matrix Metallo-Proteases induce the degradation of Extracellular Matrix (ECM) of the CD intestine. Degraded ECM leads to the formation of elastin derived peptides which are identified by the Elastin Receptor Complex (ERC) on the surface of enterocytes. ERC consists of a membrane bound NEU1, GLB1 and Cathepsin A, of which GLB1 identifies and binds to the peptides. This leads to ERC signalling, which leads to GPCR signal dependent Reactive Oxygen Species (ROS) upregulation and chemotaxis. Chemotaxis recruits the inflammatory cells, while ROS promotion leads to MAPK phosphorylation. MAPK phosphorylation leads to TNFα induction and upregulation which mediates Tissue Transglutaminase 2 (tTG2) upregulation. TNFα induction also leads to Gastric Mucosa Damage via inflammatory cascades. Simultaneously, DDX39B is expressed in lymphocytes, where it is an important factor in the transcription of cytokine genes, and is hence involved in their upregulation. Produced cytokines signal GPCR receptors on the surface of the enterocytes, which phosphorylate V-ATPases on the endosomes via cAMP dependent PK1 activation. Activated V-ATPases cause downstream mTOR signalling and leads to modifications in Ca^{2+} handling in the cell. G2 subunit of V-ATPase also directly interacts with calcium in the cell. Calcium is an important factor for tTG2 homodimerization based activation.

Multiple sclerosis, Arthritis, Systemic lupus erythromatosus, Mysositis, Atopic dermatitis, Diabetic kidney disease etc (supplementary table 1), indicating their roles in inflammation and related pathways. These cross disease association of these genes suggests: (i) their possible phenotypic and functional significance in human disease, (ii) presence of multiple pathways shared among them, and (iii) independent functional implication of these genes despite their presence in extended MHC region. This could be utilized in understanding more about the disease and help in its management and in the designing of new therapeutic drugs.

The three statistically significant SNPs rs2071592, rs2251824 and rs4947331 are located near the regulatory portions of the genes (within 4.5 kb window around 3′ and 5′ UTRs) (supplementary table 2) indicating their possible influence on the expression of the gene. *ATP6V1G2-NFKBIL1-DDX39B* (rs2071592, rs2251824) is the most significant locus identified in this study. Proteins coded by these genes (vacuolar type ATPase V1 subunit G2, NFKB inhibitor like 1 and DExD-box helicase 39B) have major roles in maintaining cellular nutrient and energy homeostasis (Atzei *et al*. 2010; Awasthi *et al*. 2018; Collins and Forgac 2018). Another significantly associated gene *NEU1* codes for protein neuraminidase 1 maintains cellular physiology and affects inflammatory functions (Chen *et al*. 1997, 2000; Haxho *et al*. 2018; Sieve *et al*. 2018). rs2071592 from *ATP6V1G2-NFKBIL1-DDX39B* locus

falls within the promoter sequence of *ATP6V1G2* and first intron of *NFKBIL1* while rs2251824 from the same locus falls within the promoter region of *DDX39B* and within second and third intron of read-through *ATP6V1G2-DDX39B* transcript targeted by nonsense mediated decay. Both of them affect the epigenetic modifications, and transcription machinery binding in T-cells, B-cells and small intestine (supplementary tables 2 and 3). This SNP was also found to significantly influence expression of *DDX39B* in whole blood tissue. *DDX39B* is a DEAD box family of RNA-dependent ATPases, which mediate ATP hydrolysis during pre-mRNA splicing. It is a major component of the transcription export complex (TREX), and is an essential splicing factor required for association of U2 small nuclear ribonucleoprotein with pre-mRNA (Fleckner *et al*. 1997). TREX binds to 5' of mRNA and enables export through TAP/NFX1 pathways. This promotes translation through regulation of pre-ribosomal RNA levels (Awasthi *et al*. 2018). Genes transcribed are involved in production of cytokines, and inflammatory proteins which help in recruiting the phagosome into the site of inflammation (Allcock *et al*. 2001; Soosanabadi *et al*. 2015).

Inflammatory signals chemokines/cytokines acts through vacuolar ATPases or V-type ATPases coded by *ATP6V1G2* to induce mTORC1 dependent signalling cascade (Sun-Wada and Wada 2015; Pamarthy *et al*. 2018). mTOR signalling affects calcium handling of the cell. Calcium channels are associates with the G2 subunit of the ATPase structure, thereby affecting the muscle contractions and has been identified to cause structural remodelling in the intestine in the pathology of Crohns' disease (Gao and Hosey 2000; Shea-Donohue *et al*. 2012). Calcium levels of the cell in regulation of tissue transglutaminase, and its link to celiac disease pathogenesis is also known (Agardh *et al*. 2005). Similar mechanisms may be involved in the pathology of the CD. V-ATPases are also involved in EGFR/ErbB receptor induced mTOR activation via a clatherin dependent endosomal/lysosomal protein derivative pathway, which regulates Transferrin endocytosis downstream, which playing a major role in Gliadin transport (Xu *et al*. 2012).

NFKBIL1 is a divergent member of the IκB family of proteins of NF-κB signalling pathway, mediators of tissue inflammation and destruction (Okamoto *et al*. 2003). It functions to negatively regulate the transcriptional activation of NF-κB target genes thereby enabling the inflammatory pathways. It is well known for being highly associated with rheumatoid arthritis (RA), an autoimmune inflammatory disease of the joints (Greetham *et al*. 2007) and in other chronic inflammatory diseases such as idiopathic inflammatory myopathy (IIM), an autoimmune muscle wastage disease, and takayasu arteritis, an inflammatory vasculitis of large blood vessles (Shibata *et al*. 2006; Chinoy *et al*. 2012). Some studies have also identified its involvement in mRNA processing, RNA splicing, and translational regulation in RA (Greetham *et al*. 2007).

Another associated marker rs4947331 localized nearly 4.4 kb downstream of *NEU1*, which codes for Neuraminidase 1. This marker was found to influence acetylation and DHS1 activity in small intestinal tissue (supplementary table 3). *NEU1* involved as a catalyst in the removal of N-acetyl neuraminic acid moieties from various glycoproteins and glycolipids and localized in the lysosomal lumen and lysosomal membrane. Extracellular matrix degradation triggers elastin derived peptides (EP), which bind to elastin receptor complex (ERC) containing *NEU1*. ERC signals GPCR thereby targeting production of ROS and multiple chemotaxic factors downstream (Hirsch and Ghigo 2014). It has been identified to working conjunction with IL-1β and lipopolysaccharide (LPS) to induce a pro-inflammatory cascade via TNF-α and IL-1β in monocytes and macrophages in patients with atherosclerosis (Sieve *et al*. 2018). Elastin derived peptides also induce Monocyte migration dependent upon NEU1 activity, and aids in leukocyte infiltration in the same disease (Gayral *et al*. 2014). Additionally, identification of an antigen via TLR recruits Neu1 from the inside of the cell to the surface, removing the sialic acid binding immunoglobulin type lectin (siglec moieties) from TLRs disrupting their interaction and activating the receptors and triggering an immune response against the antigen, and hence may influence CD pathogenesis in the same way (Chen *et al*. 2014).

Together, these three genes (*ATP6V1G2*, *DDX39B* and *NEU1*) can be said to be contributing towards the pathogenesis of CD, initiating from the production of cytokines and chemotaxic factors by T- and B-cells to ECM degradation based - *NEU1* dependent induction at the site of action. The recruited macrophages could additionally influence the processing and presentation of the identified antigenic peptides, the gliadin molecules. DDX39B induced production of cytokines in lymphocytes may further induce phosphorylation of V-type ATPase based signalling cascades leading to variance in calcium signalling which is identified to be associated with tissue transglutaminase activity, further contributing and controlling the physiological ailments of the disease quantitatively (figure 2). Pathway enrichment, PPI and tissue specific co-expression

analysis supports the above mentioned synergistic model of function of the identified genes. However, genotype-phenotype correlations need to be established through functional studies to establish this model. Notably, several approved drugs and therapeutic compounds are available against two of the identified genes namely, ATP6V1G2 and NEU1, which could further be tested in CD in controlled trials.

5. Conclusion

ATP6V1G2-NFKBIL1-DDX39B and *NEU1* are the two statistically significant loci identified in this study. Though these genes localized within extended MHC loci, prominent and independent association has been reported for several diseases and human traits. *In silico* functional analyses performed in this study and previously published reports indicate these genes directly interacting with several proteins and act synergistically to promote CD pathogenesis. In future, as therapeutics are already available against *ATP6V1G2* and *NEU1*, they could be tested for their effectiveness in treating CD in a controlled trial.

Acknowledgements

We acknowledge the contribution of Prof. Cisca Wijmenga, University Medical Center Groningen, The Netherlands in data generation, and Dr. Gosia Trynka, Wellcome Sanger Institute, UK in Immunochip data analysis. DST-FIST support (SR/FST/LS-I/2017/49-C) to the Department of Human Genetics & Molecular Medicine is acknowledged.

Funding

We acknowledge financial supports from Department of Science and Technology-Science and Research Board (DST-SERB), Govt. of India (#ECR/2016/001660), University Grants Commission, New Delhi, India (30-4/2014-BSR) to SS. PB is supported by the junior research fellowship from ICMR (33/6/2019-TF/Rare/BMS) and DST Inspire Fellowship [DST/INSPIRE Fellowship/2019/IF190501].

Declarations

Ethical approval and consent to participate Required ethical approval was obtained from Central University of Punjab before conducting this study.

Consent for publication Informed consent was obtained from all the study subjects originally included in the study.

References

Agardh D, Roth B, Lernmark A and Stenberg P 2005 Calcium activation of tissue transglutaminase in radioligand binding and enzyme-linked autoantibody immunoassays in childhood celiac disease. *Clin. Chim. Acta.* **358** 95–103

Allcock RJ, Williams JH and Price P 2001 The central MHC gene, BAT1, may encode a protein that down-regulates cytokine production. *Genes Cells* **6** 487–494

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al*. 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25** 25–29

Atzei P, Gargan S, Curran N and Moynagh PN 2010 Cactin targets the MHC class III protein IkappaB-like (IkappaBL) and inhibits NF-kappaB and interferon-regulatory factor signaling pathways. *J. Biol. Chem.* **285** 36804–36817

Awasthi S, Chakrapani B, Mahesh A, Chavali PL, Chavali S, *et al*. 2018 DDX39B promotes translation through regulation of pre-ribosomal RNA levels. *RNA Biol.* **15** 1157–1166

Bhattacharyya A, Chattopadhyay R, Mitra S and Crowe SE 2014 Oxidative stress: an essential factor in the pathogenesis of gastrointestinal mucosal diseases. *Physiol. Rev.* **94** 329–354

Biesiekierski JR 2017 What is gluten? *J. Gastroenterol. Hepatol.* **32** 78–81

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, *et al*. 2012 Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22** 1790–1797

Chen GY, Brown NK, Wu W, Khedri Z, Yu H, *et al* 2014 Broad and direct interaction between TLR and Siglec families of pattern recognition receptors and its regulation by Neu1. *Elife* **3** e04066

Chen XP, Ding X and Daynes RA 2000 Ganglioside control over IL-4 priming and cytokine production in activated T cells. *Cytokine* **12** 972–985

Chen XP, Enioutina EY and Daynes RA 1997 The control of IL-4 gene expression in activated murine T lymphocytes: a novel role for neu-1 sialidase. *J. Immunol.* **158** 3070–3080

Chinoy H, Li CK, Platt H, Fertig N, Varsani H, *et al*. 2012 Genetic association study of NF-κB genes in UK Caucasian adult and juvenile onset idiopathic inflammatory myopathy. *Rheumatology* **51** 794–799

Collins MP and Forgac M 2018 Regulation of V-ATPase assembly in nutrient sensing and function of V-ATPases in breast cancer metastasis. *Front. Physiol.* **9** 902

Crous-Bou M, Rennert G, Salazar R, Rodriguez-Moranta F, Rennert HS, *et al*. 2012 Genetic polymorphisms in fatty

acid metabolism genes and colorectal cancer. *Mutagenesis* **27** 169–176

DeBerardinis RJ and Thompson CB 2012 Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* **148** 1132–1144

Dieli-Crimi R, Cénit MC and Núñez C 2015 The genetics of celiac disease: A comprehensive review of clinical implications. *J. Autoimmun.* **64** 26–41

ENCODE Project Consortium 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74

Fleckner J, Zhang M, Valcárcel J and Green MR 1997 U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev.* **11** 1864–1872

Gao T and Hosey MM 2000 Association of L-type calcium channels with a vacuolar H(+)-ATPase G2 subunit. *Biochem. Biophys. Res. Commun.* **277** 611–616

Garner C, Ahn R, Ding YC, Steele L, Stoven S, *et al* 2014 Genome-wide association study of celiac disease in North America confirms FRMD4B as new celiac locus. *PLoS One* **9** e101428

Gayral S, Garnotel R, Castaing-Berthou A, Blaise S, Fougerat A, *et al*. 2014 Elastin-derived peptides potentiate atherosclerosis through the immune Neu1-PI3Kγ pathway. *Cardiovasc. Res.* **102** 118–127

Greetham D, Ellis CD, Mewar D, Fearon U, an Ultaigh SN, , *et al*. 2007 Functional characterization of NF-kappaB inhibitor-like protein 1 (NFkappaBIL1), a candidate susceptibility gene for rheumatoid arthritis. *Hum. Mol. Genet.* **16** 3027–3036

Gujral N, Freeman HJ and Thomson AB 2012 Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World J. Gastroenterol.* **18** 6036–6059

Haxho F, Haq S and Szewczuk MR 2018 Biased G protein-coupled receptor agonism mediates Neu1 sialidase and matrix metalloproteinase-9 crosstalk to induce transactivation of insulin receptor signaling. *Cell Signal.* **43** 71–84

Hirsch E and Ghigo A 2014 Elastin degradation and ensuing inflammation as emerging keys to atherosclerosis. *Cardiovasc. Res.* **102** 1–2

Kanehisa M and Goto S 2000 KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** 27–30

Lebwohl B and Rubio-Tapia A 2021 Epidemiology, presentation, and diagnosis of celiac disease. *Gastroenterology* **160** 63–75

Lee SY, Jin CC, Choi JE, Hong MJ, Jung DK, *et al*. 2016 Genetic polymorphisms in glycolytic pathway are associated with the prognosis of patients with early stage non-small cell lung cancer. *Sci Rep.* **6** 35603

GTEx Consortium 2013 The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45** 580–585. *https://doi.org/10.1038/ng.2653*

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, *et al*. 2009 Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37** D619–D622

Montén C, Gudjonsdottir AH, Browaldh L, Arnell H, Nilsson S, *et al*. 2015 Genes involved in muscle contractility and nutrient signaling pathways within celiac disease risk loci show differential mRNA expression. *BMC Med. Genet.* **16** 44

Murtola TJ, Wahlfors T, Haring A, Taari K, Stenman UH, *et al*. 2015 Polymorphisms of genes involved in glucose and energy metabolic pathways and prostate cancer: interplay with metformin. *Eur. Urol.* **68** 1089–1097

Okamoto K, Makino S, Yoshikawa Y, Takaki A, Nagatsuka Y, *et al*. 2003 Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am. J. Hum. Genet.* **72** 303–312

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, *et al*. 2012 Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9** 345–350

Östensson M, Montén C, Bacelis J, Gudjonsdottir AH, Adamovic S, *et al* 2013 A possible mechanism behind autoimmune disorders discovered by genome-wide linkage and association analysis in celiac disease. *PLoS One* **8** e70174

Pamarthy S, Kulshrestha A, Katara GK and Beaman KD 2018 The curious case of vacuolar ATPase: regulation of signaling pathways. *Mol. Cancer* **17** 41

Parzanese I, Qehajaj D, Patrinicola F, Aralica M, Chiriva-Internati M, *et al*. 2017 Celiac disease: From pathophysiology to treatment. *World J. Gastrointest. Pathophysiol.* **8** 27–38

Pérez S, Taléns-Visconti R, Rius-Pérez S, Finamor I and Sastre J 2017 Redox signaling in the gastrointestinal tract. *Free Radic. Biol. Med.* **104** 75–103. *https://doi.org/10.1016/j.freeradbiomed.2016.12.048*

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, *et al*. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575

Senapati S, Gutierrez-Achury J, Sood A, Midha V, Szperl A, *et al*. 2015 Evaluation of European coeliac disease risk variants in a north Indian population. *Eur. J. Hum. Genet.* **23** 530–535

Senapati S, Sood A, Midha V, Sood N, Sharma S, *et al*. 2016 Shared and unique common genetic determinants between pediatric and adult celiac disease. *BMC Med. Genomics* **9** 44

Shea-Donohue T, Notari L, Sun R and Zhao A 2012 Mechanisms of smooth muscle responses to inflammation. *Neurogastroenterol. Motil.* **24** 802–811. *https://doi.org/10.1111/j.1365-2982.2012.01986.x*

Shibata H, Yasunami M, Obuchi N, Takahashi M, Kobayashi Y, *et al*. 2006 Direct determination of single nucleotide polymorphism haplotype of NFKBIL1

promoter polymorphism by DNA conformation analysis and its application to association study of chronic inflammatory diseases. *Hum Immunol.* **67** 363–373

Sieve I, Ricke-Hoch M, Kasten M, Battmer K, Stapel B, *et al*. 2018 A positive feedback loop between IL-1β, LPS and NEU1 may promote atherosclerosis by enhancing a pro-inflammatory state in monocytes and macrophages. *Vascul. Pharmacol.* **103–105** 16–28

Soosanabadi M, Bayat H, Kamali K, Saliminejad K, Banan M, *et al*. 2015 Association study of IL-4 -590 C/T and DDX39B -22 G/C polymorphisms with the risk of late-onset Alzheimer's disease in Iranian population. *Curr. Aging Sci.* **8** 276–281

Sun-Wada GH and Wada Y 2015 Role of vacuolar-type proton ATPase in signal transduction. *Biochim. Biophys. Acta.* **1847** 1166–1172

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, *et al*. 2003 PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13** 2129–2141

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, *et al*. 2011 Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43** 1193–1201

Trynka G, Wijmenga C and van Heel DA 2010 A genetic perspective on coeliac disease. *Trends Mol. Med.* **16** 537–550

Ward LD and Kellis M 2012 HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40** D930–D934

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, *et al*. 2018 DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46** D1074–D1082

Xia J, Gill EE and Hancock RE 2015 NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* **10** 823–844

Xu Y, Parmar A, Roux E, Balbis A, Dumas V, *et al*. 2012 Epidermal growth factor-induced vacuolar (H+)-atpase assembly: a role in signaling via mTORC1 activation. *J. Biol. Chem.* **287** 26409–26422

Yi MC, Melkonian AV, Ousey JA and Khosla C 2018 Endoplasmic reticulum-resident protein 57 (ERp57) oxidatively inactivates human transglutaminase 2. *J. Biol. Chem.* **293** 2640–2649. https://doi.org/10.1074/jbc.RA117.001382

Communicated by Ullas Kolthur-Seetharam