In silico identification of single nucleotide variations at CpG sites regulating CpG island existence and size

Nivas Shyamala1, Chaitra Lava Kongettira3, Kaushik Puranam3, Keerthi Kupsal1, Ramanjaneyulu Kummari1, Chiranjeevi Padala1,2 & Surekha Rani Hanumanth1

Genetic and epigenetic modifications of genes involved in the key regulatory pathways play a significant role in the pathophysiology and progression of multifactorial diseases. The present study is an attempt to identify single nucleotide variations (SNVs) at CpG sites of promoters of ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1 and TIMP1 genes influencing CpG island (CGI) existence and size associated with the pathophysiology of Diabetes mellitus, Coronary artery disease and Cancers. Promoter sequences located between −2000 to +2000 bp were retrieved from the EPDnew database and predicted the CpG island using MethPrimer. Further, SNVs at CpG sites were accessed from NCBI, Ensembl while transcription factor (TF) binding sites were accessed using AliBaba2.1. CGI existence and size were determined for each SNV at CpG site with respect to wild type and variant allele by MethPrimer. A total of 200 SNVs at CpG sites were analyzed from the promoters of ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1 and TIMP1 genes. Of these, only 17 (8.5%) SNVs were found to influence the loss of CGI while 70 (35%) SNVs were found to reduce the size of CGI. It has also been found that 59% (10) of CGI abolishing SNVs are showing differences in binding of TFs. The findings of the study suggest that the candidate SNVs at CpG sites regulating CGI existence and size might influence the DNA methylation status and expression of genes involved in molecular pathways associated with several diseases. The insights of the present study may pave the way for new experimental studies to undertake challenges in DNA methylation, gene expression and protein assays.

Multifactorial diseases like Diabetes mellitus (DM), Coronary artery disease (CAD) and Cancers are the top leading causes of death worldwide1. Globally, understanding of underlying mechanisms and prevention of these diseases with different strategies are potential challenges for researchers in medicine2. These diseases are influenced by common risk factors such as family history, smoking, obesity, insufficient physical activity, etc3. Studies suggest that besides these conventional risk factors, genetic and epigenetic modifications of certain genes also play a significant role in pathophysiology and progression of these diseases4–6.

Evidences suggest that epigenetic modifications regulate the genome structure and expression pattern of genes7,8. These mechanisms include DNA methylation, histone modification and non-coding RNAs regulation, which can be inherited from one generation to the next9. DNA methylation is a common molecular alteration at CpG sites of DNA sequence which is influenced by genetic and environmental factors. DNA methylation in various cell types regulate the expression of genes and shows an association with the pathophysiology of diseases10–13.

DNA methylation at CpG sites is an enzymatic reaction catalysed and maintained by DNA methyltransferase (DNMT) family in particular DNMT3A, 3B and DNMT114. DNMTs convert cytosine to 5-methylcytosine by adding methyl group at CpG dinucleotide sites of CpG islands (CGIs). CGIs are typically located at the regulatory regions, predominantly in promoters and are 500-1500 bp long15,16. Commonly, transcriptional activity of promoter depends on the binding efficiency of RNA polymerase II and transcription factors (TF) to the core

1Present address: Department of Genetics and Biotechnology, University College of Science, Osmania University, Hyderabad 500007, Telangana State, India. 2Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad 500046, Telangana State, India. *email: surekharanih@gmail.com
promoter\textsuperscript{17}. Studies suggested that the methylation of cytosines in a promoter DNA suppresses the rate of transcription, reduces the mRNA copy number and ultimately affects the protein synthesis\textsuperscript{18–20}. 

Initially, genes under the study ACAT1\textsuperscript{21–22}, APOB\textsuperscript{23–24}, APOE\textsuperscript{25–27}, CYBA\textsuperscript{28–30}, FAS\textsuperscript{30–31}, FLT1\textsuperscript{32–33}, KSR2\textsuperscript{34}, LDLR\textsuperscript{35–36}, MMP9\textsuperscript{37}, PCSK9\textsuperscript{38–39}, PHOXA2\textsuperscript{40–42}, REST\textsuperscript{43–44}, SH2B3\textsuperscript{45–47}, SORT1\textsuperscript{48–49} and TIMP1\textsuperscript{50–52} were selected which were found to be involved in several key regulatory pathways associated with the pathology of DM, CAD and Cancers (Supplementary Table 1). These genes and gene products enormously involve in various pathways: ACAT1, PCSK9 & SORT1 in cholesterol homeostasis; APOB, APOE & LDLR in lipid metabolism; CYBA, KSR2 & PHOXA2 in oxidative stress; FAS, REST & SORT1 in apoptosis; FLT1 & SH2B3 in inflammation and angiogenesis; MMP9 & TIMP1 in maintenance of extracellular matrix and vascular smooth muscle cells.

Studies suggest that the single nucleotide variations (SNVs) located at promoter, exonic & intronic regions of these genes regulate the expression, alternative splicing of mRNA, structural conformation of proteins, etc\textsuperscript{28,30,31,36,53}. Moreover, these genes were found to have genome-wide significant loci for risk of multifactorial diseases in various populations. In addition, epigenetic studies have suggested that the DNA methylation of factors binding at enhancer or silencer region and miRNA binding at 3'UTR region\textsuperscript{67–70}. The SNVs at CpG sites were accessed from National Center for Biotechnology Information (NCBI) and Ensembl. NCBI and GeneCards contain 4806 promoters from various species like Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae, etc\textsuperscript{71}.

Initially, genes under the study ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOXA2, REST, SH2B3, SORT1 and TIMP1 were selected which were found to be involved in several key regulatory pathways associated with the pathology of DM, CAD and Cancers (Supplementary Table 1). These genes and gene products enormously involve in various pathways: ACAT1, PCSK9 & SORT1 in cholesterol homeostasis; APOB, APOE & LDLR in lipid metabolism; CYBA, KSR2 & PHOXA2 in oxidative stress; FAS, REST & SORT1 in apoptosis; FLT1 & SH2B3 in inflammation and angiogenesis; MMP9 & TIMP1 in maintenance of extracellular matrix and vascular smooth muscle cells.

Studies suggest that the single nucleotide variations (SNVs) located at promoter, exonic & intronic regions of these genes regulate the expression, alternative splicing of mRNA, structural conformation of proteins, etc\textsuperscript{28,30,31,36,53}. Moreover, these genes were found to have genome-wide significant loci for risk of multifactorial diseases in various populations. In addition, epigenetic studies have suggested that the DNA methylation of factors binding at enhancer or silencer region and miRNA binding at 3'UTR region\textsuperscript{67–70}. The SNVs at CpG sites were accessed from National Center for Biotechnology Information (NCBI) and Ensembl. NCBI and GeneCards contain 4806 promoters from various species like Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae, etc\textsuperscript{71}.

The studies of promoter\textsuperscript{17} suggest that the methylation of cytosines in a promoter DNA suppresses the rate of transcription, reduces the mRNA copy number and ultimately affects the protein synthesis\textsuperscript{18–20}. 

Initially, genes under the study ACAT1\textsuperscript{21–22}, APOB\textsuperscript{23–24}, APOE\textsuperscript{25–27}, CYBA\textsuperscript{28–30}, FAS\textsuperscript{30–31}, FLT1\textsuperscript{32–33}, KSR2\textsuperscript{34}, LDLR\textsuperscript{35–36}, MMP9\textsuperscript{37}, PCSK9\textsuperscript{38–39}, PHOXA2\textsuperscript{40–42}, REST\textsuperscript{43–44}, SH2B3\textsuperscript{45–47}, SORT1\textsuperscript{48–49} and TIMP1\textsuperscript{50–52} were selected which were found to be involved in several key regulatory pathways associated with the pathology of DM, CAD and Cancers (Supplementary Table 1). These genes and gene products enormously involve in various pathways: ACAT1, PCSK9 & SORT1 in cholesterol homeostasis; APOB, APOE & LDLR in lipid metabolism; CYBA, KSR2 & PHOXA2 in oxidative stress; FAS, REST & SORT1 in apoptosis; FLT1 & SH2B3 in inflammation and angiogenesis; MMP9 & TIMP1 in maintenance of extracellular matrix and vascular smooth muscle cells.

Materials and methods

**Study design.** The detailed study design is presented in Fig. 1.

**Promoter sequence retrieval.** Promoter sequences located between −2000 to +2000 bp were retrieved from Eukaryotic promoter database (EPD) new to check the CpG island status of genes under the study. EPD new allows access to several databases of experimentally validated promoters and published articles of model organisms. EPD new contains 4806 promoters from various species like Homo sapiens, Mus musculus, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Saccharomyces cerevisiae, etc\textsuperscript{71}.

**Prediction of CpG Islands.** CpG islands (CGIs) in promoter sequence of genes under the study were predicted using MethPrimer v1.1 beta. CGI existence and size were determined for each single nucleotide variation at CpG site with respect to wild type and variant allele. MethPrimer predicts potential CGIs in the input promoter DNA sequence and design sequence specific primers for Methylation-Specific PCR and Bisulfite- Sequencing PCR. The output results are presented in graphical view for predicted CpG island and in text format for PCR primers\textsuperscript{72}. The criteria used for gain and loss of CGI prediction is Island size > 100bp, GC percent > 50.0, ratio of Obs/Exp no of CpG dinucleotides > 0.60\textsuperscript{73}.

**Selection of SNVs at CpG sites.** CpG sites were identified from the results of MethPrimer and the SNVs at CpG sites were accessed from National Center for Biotechnology Information (NCBI) and Ensembl. NCBI and Ensembl are widely used genome browsers in global scientific community. The browsers were developed with the data of genomic regions, genes, SNVs and gene variants across genomes, genetic variation, genotypes, etc. The tools visualize DNA sequence and their respective annotated genetic variations to identify the SNVs at CpG sites in CpG islands\textsuperscript{74,75}.

**Transcription factor binding site prediction.** AliiBaba2.1 tool was used for the prediction of transcription factor binding sites in wild type and variant alleles of SNVs at CpG sites. It is an online tool to identify transcription factors and their respective binding sites for the input DNA sequence by constructing matrices on
the fly from TRANSFAC 4.0 sites. AliBaba tool has significantly higher sensitivity and sensitivity/specificity ratio than other current approaches76.

**Co-expression prediction.** APOE, CYBA, FAS, LDLR, MMP9, PCSK9, PHOX2A, SH2B3 and TIMP1 genes were analysed to know the other co-expressing, physically interacting, co-localizing and key biological pathway related genes using GeneMANIA. GeneMANIA is a potent database of almost 2300 networks with 600 million interactions covering upto 164,000 genes in model organisms and provide genomic, proteomic, and gene function data. It is an effective approach to predict the function of input single gene/multiple gene queries physically interacting proteins, co-expressing and co-localizing genes, genetic interactions, shared protein domains and pathways77,78.

Layouts generated by GeneMANIA web server have nodes and edges. Nodes represent gene and its products, while edges represent co-expression interaction and weight of each edge implies the evidence of co-functionality data source.

**Gene ontology enrichment analysis.** Gene ontology (GO) enrichment analysis of genes (ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1, TIMP1) was performed using Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 online tool (https://david.ncifcrf.gov/home.jsp). The GO terms were classified into three categories: biological process (BP), cellular component (CC) and molecular function (MF) with significant p value of <0.05. Further, GO term enrichment analysis was used to annotate the disease class and functional clustering of genes under the study.

**Results**

Promoter sequence of ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1 and TIMP1 genes were analysed for the prediction of CpG islands and have observed CpG islands for all the genes (Fig. 2A, B). Further, the existence and sizes of CGI for wild type and variant alleles of all the CpG SNVs were analyzed. In addition, transcription factors binding to both the wild type and variant alleles of CpG SNVs abolishing CGI were predicted.

A total of 200 SNVs at CpG sites were studied for ACAT1 (10), APOB (3), APOE (1), CYBA (7), FAS (12), FLT1 (6), KSR2 (31), LDLR (16), MMP9 (28), PCSK9 (8), PHOX2A (22), REST (5), SH2B3 (29), SORT1 (16) and TIMP1 (6) genes. Of these, 17 (8.5%) candidate SNVs abolished the CpG islands existence and 70 (35%) SNVs potentially decreased the CpG islands size in various genes (Table 1). The percentage of abolished CGIs and change in size of CGIs of all genes are represented in Table 1 and Fig. 3.

**CpG SNVs abolishing and reducing sizes of CGI.** APOE gene has a single SNV rs769448 at CpG site, its variant allele has lost the entire 112 bp CGI. Among the 16 CpG SNVs studied in 2 CGIs (island 1:138 bp,
Figure 2. CpG islands prediction in promoter sequence of genes. (A) ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR. (B) MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1, TIMP1. The figure consists:

- input sequence to predict the CpG islands and to design bisulfite/methylation specific PCR primers,
- CpG island region.
SNVs at CpG sites abolishing the CGIs of LDLR, MMP9, SH2B3, TIMP1 and APOE 1 genes were analysed to predict the difference in binding of transcription factors (TF) at the site of variation. As represented in Table 2, we have observed that SNVs 4 in LDLR, 2 in MPP9, 1 in SH2B3, 2 in TIMP1 and 1 in APOE genes have shown a difference in binding of TFs.

To the 4 SNVs of LDLR gene that abolished CGI, TFs binding site prediction has shown that rs1026272027 wild type allele has a binding site for C/EBPαp and variant allele has a binding site for C/EBPβet. For rs887608252, C/T, rs1006494933, G/A and rs1024897634, C/T SNVs, there were no TF binding sites for their wild type alleles, but their variant alleles have binding sites for C/EBPα, GATA-1 & Oct-1 and Oct-1 TFs respectively.

Likewise, 2 SNVs abolishing CGIs in MPP9 gene have shown the difference in binding of TFs, rs370018925 wild type allele has no binding site for any TF whereas variant allele is bound by Sp1 transcription factor. Though the rs1014494202 has Sp1 binding site for wild type allele, variant allele has an additional binding site for BRF-1 transcription factor.

For rs922413124 in SH2B3 gene, there was a binding site for Sp1 in wild type allele, but it is abolished in variant allele. Similarly, APOE rs7694448 has binding site for Sp1 transcription factor but its variant allele is lacking a site for binding of any transcription factor.

Furthermore, 2 SNVs that abolished CGIs in TIMP1 gene has shown that the wild type alleles of rs779329701 and rs376836551 has binding sites for Egr-1 and Sp1 transcription factors while variant alleles have binding sites for NF-1 and N-Myc transcription factors respectively.

GeneMANIA co-expression network revealed that APOE, LDLR, MPP9, SH2B3 and TIMP1 genes might regulate the expression of several other genes. Single gene queries have shown that APOE gene influencing the expression of APOC3, APOA1, APOB, LIPC, LDLR influences LCN2, TIMP1; MPP9 influences LIPC, MPP1, LCN2, SH2B3 influences VLDLR, LDLRAP1, TGFβ1, KIT; TIMP1 influences VLDLR, LDLR, MMP1, MPP9, MMP3, LCN2, SH2B3 genes (Fig. 4A–E). While multi gene queries interestingly displayed that APOE, LDLR, MPP9, SH2B3 and TIMP1 genes expression are associated with each other (Fig. 5). GeneMANIA consolidated networks revealed that the APOE, LDLR, MPP9, SH2B3, TIMP1 genes are involved in various signaling pathways. It has been shown that APOE & LDLR genes are involved in lipid and lipoprotein metabolisms, while MPP9 and TIMP1 genes are significantly modulating the degradation of extracellular matrix. In addition, these genes show an internal correlation in their co-expression network (Supplementary Fig. 1).

The gene ontology enrichment analysis of the genes set is shown in Fig. 6. The top 10 GO terms of biological process (BP), cellular component (CC), molecular function (MF) and disease class analyses in genes were sorted by p-value or gene count. According to the BP analysis, the GO term pathways were mainly associated with the cholesterol biosynthesis, metabolism and homeostasis, regulation of apoptosis, receptor mediated endocytosis, etc (Fig. 6A). For the CC analysis, the GO terms of these genes were mainly located and enriched in the plasma membrane, extracellular exosomes and space, golgi apparatus, etc (Fig. 6B). In the MF analysis, 15 genes were mainly enriched and associated with binding activity and transporter activity particularly protein binding, metal ion binding, identical protein binding, low-density lipoprotein particle receptor binding, cholesterol transporter activity, etc (Fig. 6C).

The GO terms disease class analysis of these genes revealed that the genes are associated with metabolic diseases, neurological diseases, cardiovascular diseases, cancers, etc (Fig. 6D). Later, functional annotation clustering of these genes was performed and functional chart of cluster with highest gene enrichment score (3.17) is shown in Fig. 6E. Out of the 15 genes APOB, APOE, LDLR, PCSK9, SORT1 genes are associated with golgi complex, early endosome, cholesterol metabolism, etc (Supplementary data 1).

**Discussion**

The multifactorial diseases like diabetes mellitus, coronary artery disease and cancers are leading cause for morbidity and mortality worldwide. Genetic and epigenetic modifications are also recognized as significant risk factors for the pathophysiology of these diseases. Studies reported that epigenetic modifications play a crucial role in cell differentiation at embryonic development. Besides, environmental factors and age affect the DNA methylation and demethylation patterns in mammals. The methylation patterns of promoter DNA depends upon the presence of CpG sites, CpG islands existence and their respective size in the promoter region. Genetic
| S. No. | CpG island and size (bp) | Single nucleotide variations (SNVs) (rs number; variation) | CpG coordinates on chromosome | CpG island status with change in CpG island size (bp) |
|-------|--------------------------|---------------------------------------------------------|-----------------------------|--------------------------------------------------|
| Gene  | Acetyl-Coenzyme A acetyltransferase 1 (ACAT1) | | | |
| 1     | Island,341 | rs539426263,C/A* | chr1:108121278 | Present | Present | 339 |
| 2     | rs376263077,G/C | chr1:108121289 | Present | Present | 341 |
| 3     | rs376263077,G/T* | Present | Present | 339 |
| 4     | rs979540931,C>G* | chr1:108121307 | Present | Present | 339 |
| 5     | rs51761017,C>A* | chr1:108121313 | Present | Present | 339 |
| 6     | rs1191223847,G>T* | chr1:108121314 | Present | Present | 339 |
| 7     | rs1294688280,C>T | chr1:108121367–108121378 | Present | Present | 341 |
| 8     | rs1294688280,G>A | Present | Present | 341 |
| 9     | rs1246409549,C>T | chr1:108121403 | Present | Present | 341 |
| 10    | rs1197006182,G>A* | chr1:108121404 | Present | Present | 341 |
| Gene  | Apolipoprotein B (APOB) | | | |
| 11    | Island,344 | rs745633995,G/A* | chr2:21044088 | Present | Present | 340 |
| 12    | rs956977643,C>T* | chr2:21044082 | Present | Present | 343 |
| 13    | rs9734345426,C/A | chr2:21044076 | Present | Present | 344 |
| Gene  | Apolipoprotein E (APOE) | | | |
| 14    | Island,112 | rs7694483,C/T** | chr19:44906322 | Present | Abolished | 0 |
| Gene  | Cytochrome b-245 alpha chain (CYBA) | | | |
| 15    | Island,136 | rs1021215371,C>T* | chr16:88651087 | Present | Present | 135 |
| 16    | rs544939962,G/A* | chr16:88651070 | Present | Present | 135 |
| 17    | rs760194356,C>T* | chr16:88651047 | Present | Present | 135 |
| 18    | rs785100422,G/T* | chr16:88651064 | Present | Present | 135 |
| 19    | rs760194355,C>T | chr16:88651047 | Present | Present | 136 |
| 20    | rs7603344376,G/A | chr16:88651046 | Present | Present | 136 |
| 21    | rs9734345426,C/A | chr16:88651027 | Present | Present | 136 |
| Gene  | Factor associated suicide death receptor (FAS) | | | |
| 22    | Island,1,199 | rs752145197,G/C* | chr10:88990538 | Present | Present | 190 |
| 23    | rs755644207,C/T* | chr10:88990539 | Present | Present | 177 |
| 24    | rs868047456,G/A* | chr10:88990540 | Present | Present | 191 |
| 25    | rs77766435,C/A* | chr10:88990541 | Present | Present | 190 |
| 26    | rs533623533,G/A* | chr10:88990542 | Present | Present | 191 |
| 27    | rs9658677,G/A | chr10:88990582 | Present | Present | 199 |
| 28    | rs902017811,C/A* | chr10:88990595 | Present | Present | 128 |
| 29    | rs1021894100,C>T* | chr10:88990642 | Present | Present | 128 |
| 30    | rs769222278,G/C* | chr10:88990643 | Present | Present | 128 |
| 31    | rs777396229,C/A* | chr10:88990656 | Present | Present | 128 |
| 32    | rs904814298,G/C* | chr10:88990657 | Present | Present | 128 |
| 33    | rs557366318,G/A* | chr10:88990715 | Present | Present | 184 |
| Gene  | Fms related tyrosine kinase 1 (FLT1) | | | |
| 34    | Island,1,211 | rs930109277,G/C | chr13:28495711 | Present | Present | 211 |
| 35    | rs61763160,C/T* | chr13:28495681 | Present | Present | 199 |
| 36    | rs1024357361,G/A* | chr13:28495655 | Present | Present | 198 |
| 37    | rs779832391,G/A* | chr13:28495524 | Present | Present | 188 |
| 38    | Island,2,204 | rs1028125144,C/G | chr13:28495300 | Present | Present | 188 |
| 39    | rs990306653,G/T | chr13:28495276 | Present | Present | 188 |
| Gene  | Kinase suppressor of ras 2 (KSR2) | | | |
| 40    | Island,838 | rs7490418,C/T* | chr12:117969559 | Present | Present | 803 |
| 41    | rs962883023,G/A* | chr12:117969543 | Present | Present | 804 |
| 42    | rs1010334904,G/C | chr12:117969521 | Present | Present | 838 |
| 43    | rs891647346,G/T,A—T | chr12:117969518 | Present | Present | 838 |
| 44    | rs522195962,G/C | chr12:117969510 | Present | Present | 838 |
| 45    | rs1829660335,G/A | chr12:117969500 | Present | Present | 838 |
| 46    | rs939805222,GGCGGAGGCGGCGAC GCTCCTC*C | chr12:117969450–117969478 | Present | Present | 817 |
| 47    | rs1011133176,C/T | chr12:117969464 | Present | Present | 838 |
| 48    | rs114278232,G/A | chr12:117969418 | Present | Present | 838 |
| 49    | rs52320001,C/G | chr12:117969394 | Present | Present | 838 |
| 50    | rs7490907,G/C,A—C | chr12:117969393 | Present | Present | 838 |
| 51    | rs103436188,G/C | chr12:117969386 | Present | Present | 838 |
| 52    | rs931660247,C/A | chr12:117969367 | Present | Present | 838 |

Continued
| S. No. | CpG island and size (bp) | Single nucleotide variations (SNVs) (rs number; variation) | CpG coordinates on chromosome | CpG island status with Wild type allele | Variant allele | Change in CpG island size (bp) |
|--------|-------------------------|-----------------------------------------------------------|-----------------------------|----------------------------------------|----------------|-------------------------------|
| 53     | rs89888603G/C         | chr12:117969341 Present Present | 838                         |                                        |                |                               |
| 54     | rs54381965C/T        | chr12:117969330 Present Present | 838                         |                                        |                |                               |
| 55     | rs971314425G/A      | chr12:117969329 Present Present | 838                         |                                        |                |                               |
| 56     | rs908447922TCCCGCCGCAGCAGT         | chr12:117969312—117969327 Present Present | 824                         |                                        |                |                               |
| 57     | rs92780337G/A      | chr12:117969310 Present Present | 838                         |                                        |                |                               |
| 58     | rs90876827G/C      | chr12:117969287 Present Present | 838                         |                                        |                |                               |
| 59     | rs10220899G/C       | chr12:117969287 Present Present | 838                         |                                        |                |                               |
| 60     | rs95496228G/C      | chr12:117969273 Present Present | 838                         |                                        |                |                               |
| 61     | rs95614421G/G      | chr12:117969268 Present Present | 838                         |                                        |                |                               |
| 62     | rs89034883G/A      | chr12:117969244 Present Present | 838                         |                                        |                |                               |
| 63     | rs55770395G/TCTT      | chr12:117969236 Present Present | 838                         |                                        |                |                               |
| 64     | rs99992906G/T      | chr12:117969228 Present Present | 838                         |                                        |                |                               |
| 65     | rs8862144G/T       | chr12:117969152 Present Present | 838                         |                                        |                |                               |
| 66     | rs10572182G/C     | chr12:117969151 Present Present | 838                         |                                        |                |                               |
| 67     | rs55742283G/T      | chr12:117969140 Present Present | 838                         |                                        |                |                               |
| 68     | rs53489307G/TAGAG     | chr12:117969130 Present Present | 838                         |                                        |                |                               |
| 69     | rs74031469G/C     | chr12:117969128 Present Present | 838                         |                                        |                |                               |
| 70     | rs908113700G/C     | chr12:117969116 Present Present | 838                         |                                        |                |                               |

Gene  Low density lipoprotein receptor (LDLR)

| Island 1;138 | rs531870546G/G | chr19:11087615 Present Present | 138                         |                                        |                |                               |
| 71     | rs54367681G/A/T*   | chr19:11087616 Present Present | 136                         |                                        |                |                               |
| 72     | rs10267202G/T**    | chr19:11087638 Present Present | 0                           |                                        |                |                               |
| 73     | rs88760182G/T**   | chr19:11087645 Present Present | 0                           |                                        |                |                               |
| 74     | rs10664949G/A/*    | chr19:11087646 Present Present | 0                           |                                        |                |                               |
| 75     | rs53249368G/A/*   | chr19:11087670 Present Present | 0                           |                                        |                |                               |
| 76     | rs10248976G/T**    | chr19:11087677 Present Present | 0                           |                                        |                |                               |
| 77     | rs10383990G/C       | chr19:11087733 Present Present | 108                         |                                        |                |                               |
| 78     | rs89933107G/A     | chr19:11087734 Present Present | 108                         |                                        |                |                               |
| 79     | rs37179807G/T**    | chr19:11087737 Present Present | 108                         |                                        |                |                               |
| 80     | rs10467793G/C     | chr19:11087738 Present Present | 138                         |                                        |                |                               |
| 81     | Island 2;167 | rs57471391G/C     | chr19:11089227 Present Present | 167                         |                                        |                |                               |
| 82     | rs17249134G/T     | chr19:11089281 Present Present | 167                         |                                        |                |                               |
| 83     | rs17249141G/T*    | chr19:11089332 Present Present | 152                         |                                        |                |                               |
| 84     | rs54995837G/T*    | chr19:11089343 Present Present | 152                         |                                        |                |                               |
| 85     | rs18201767G/C/A*  | chr19:11089347 Present Present | 152                         |                                        |                |                               |

Gene  Matrix metalloproteinase 9 (MMP9)

| Island 1;172 | rs13962047G/A/T—A/* or C/A/T—T/* | chr20:46009987 Present Abolished | 0                           |                                        |                |                               |
| 86     | rs37018925G/T**   | chr20:46009908 Present Abolished | 0                           |                                        |                |                               |
| 87     | rs20106999G/A/*    | chr20:46009909 Present Abolished | 0                           |                                        |                |                               |
| 88     | rs10144942G/T**   | chr20:46009936 Present Abolished | 0                           |                                        |                |                               |
| 89     | rs14671929G/A/*    | chr20:46009937 Present Abolished | 0                           |                                        |                |                               |
| 90     | rs200849957C/G/T—G or C/G/T—T | chr20:46009970 Present Present | 172                         |                                        |                |                               |
| 91     | rs1805308G/A      | chr20:46009971 Present Present | 172                         |                                        |                |                               |
| 92     | rs10236008G/C/T   | chr20:46009976 Present Present | 172                         |                                        |                |                               |
| 93     | rs14369540G/A/A—A or T | chr20:46009977 Present Present | 172                         |                                        |                |                               |
| 94     | rs45482493G/T     | chr20:46009991 Present Present | 172                         |                                        |                |                               |
| 95     | rs37725182G/C/A   | chr20:46010100 Present Present | 172                         |                                        |                |                               |
| 96     | rs14035254G/T     | chr20:46010200 Present Present | 172                         |                                        |                |                               |
| 97     | Island 2;205 | rs62336901G/T*    | chr20:46010433 Present Present | 137                         |                                        |                |                               |
| 98     | rs67597004G/A*    | chr20:46010475 Present Present | 135                         |                                        |                |                               |
| 99     | rs75762462G/C*    | chr20:46010497 Present Present | 134                         |                                        |                |                               |
| 100    | rs49347450G/T*    | chr20:46010509 Present Present | 134                         |                                        |                |                               |
| 101    | rs20063734G/T*    | chr20:46010511 Present Present | 134                         |                                        |                |                               |
| 102    | rs75454847G/T*    | chr20:46010515 Present Present | 150                         |                                        |                |                               |
| 103    | rs47572481G/A/*   | chr20:46010529 Present Present | 149                         |                                        |                |                               |
| 104    | rs77647714G/A*    | chr20:46010539 Present Present | 150                         |                                        |                |                               |
| 105    | rs20190213G/C/G/T—G or C/G/T—T | chr20:46010558 Present Present | 149                         |                                        |                |                               |
| 106    | rs67959655G/A*    | chr20:46010561 Present Present | 149                         |                                        |                |                               |
| 107    | rs538990326G/C/A  | chr20:46010569 Present Present | 205                         |                                        |                |                               |

Continued
| S. No. | CpG island and size (bp) | Single nucleotide variations (SNVs) (rs number; variation) | CpG coordinates on chromosome | CpG island status with Change in CpG island size (bp) |
|--------|--------------------------|-----------------------------------------------------------|-------------------------------|-----------------------------------------------------|
| 110    | rs77380909(G/A)         | chr20:46010628 Present                                    | Present                        | 205                                                  |
| 111    | rs202214757(C/A)        | chr20:46010629 Present                                    | Present                        | 205                                                  |
| 112    | rs183834856(G/A)        | chr20:46010630 Present                                    | Present                        | 205                                                  |
| 113    | rs94503896(C/A)         | chr20:46010639 Present                                    | Present                        | 205                                                  |
| 114    | rs201044639(G/A)        | chr20:46010640 Present                                    | Present                        | 205                                                  |
|        | Gene                     | Proprotein convertase subtilisin/kexin type 9 (PCSK9)     |                               |                                                      |
| 115    | Island:494              | rs911797628(G>T) Present                                | chr1:55039338 Present          | Present                                              | 464  |
| 116    | rs97696811(G>A)         | chr1:55039389 Present                                    | Present                        | 464                                                  |
| 117    | rs371053631(G>T)        | chr1:55039390 Present                                    | Present                        | 464                                                  |
| 118    | rs976397913(G/A)        | chr1:55039391 Present                                    | Present                        | 464                                                  |
| 119    | rs665975999(C>T)        | chr1:55039416 Present                                    | Present                        | 491                                                  |
| 120    | rs884737926(G>T)        | chr1:55039402 Present                                    | Present                        | 491                                                  |
| 121    | rs188274059(C/T)        | chr1:55039516 Present                                    | Present                        | 491                                                  |
| 122    | rs745962138(G/A)        | chr1:55039517 Present                                    | Present                        | 491                                                  |
|        | Gene                     | Paired like homeobox 2a (PHOX2A)                         |                               |                                                      |
| 123    | Island:964              | rs946255361(G/A)* Present                               | chr11:72244638 Present         | Present                                              | 880  |
| 124    | rs985354082(C/G)        | chr11:72244600 Present                                    | Present                        | 964                                                  |
| 125    | rs562101265(C/A)*       | chr11:72244597 Present                                    | Present                        | 879                                                  |
| 126    | rs54309053(G/A)*        | chr11:72244598 Present                                    | Present                        | 880                                                  |
| 127    | rs919731268(G>T)*       | chr11:72244574 Present                                    | Present                        | 880                                                  |
| 128    | rs973070136(G/C)        | chr11:72244535 Present                                    | Present                        | 964                                                  |
| 129    | rs947059499(S/G)        | chr11:72244511 Present                                    | Present                        | 964                                                  |
| 130    | rs1021783886(G/A)       | chr11:72244510 Present                                    | Present                        | 964                                                  |
| 131    | rs1010395824(G/A)       | chr11:72244507 Present                                    | Present                        | 964                                                  |
| 132    | rs990410699(G/C)        | chr11:72244571 Present                                    | Present                        | 964                                                  |
| 133    | rs950032571(C/T)*       | chr11:72244555 Present                                    | Present                        | 964                                                  |
| 134    | rs533723835(G/A)*       | chr11:72244322 Present                                    | Present                        | 390,571                                              |
| 135    | rs1021105224(G/A)*      | chr11:72244319 Present                                    | Present                        | 390,571                                              |
| 136    | rs10198844836(G/A)*     | chr11:72244305 Present                                    | Present                        | 390,571                                              |
| 137    | rs89904293(G/C)         | chr11:72244293 Present                                    | Present                        | 964                                                  |
| 138    | rs91770863(G/C)*        | chr11:72244248 Present                                    | Present                        | 964                                                  |
| 139    | rs937911897(C/T)        | chr11:72244236 Present                                    | Present                        | 964                                                  |
| 140    | rs978754333(G/C)        | chr11:72244197 Present                                    | Present                        | 964                                                  |
| 141    | rs992030944(G/A)        | chr11:72244196 Present                                    | Present                        | 964                                                  |
| 142    | rs951619630(G/C)        | chr11:72244194 Present                                    | Present                        | 964                                                  |
| 143    | rs1019771178(C/T)*      | chr11:72244193 Present                                    | Present                        | 964                                                  |
| 144    | rs1008498233(G/T)       | chr11:72244187 Present                                    | Present                        | 964                                                  |
|        | Gene                     | RE1 silencing transcription factor (REST)                |                               |                                                      |
| 145    | Island:298              | rs964635804(G/A)* Present                               | chr4:56907734 Present          | Present                                              | 291  |
| 146    | rs98221493(G/C)         | chr4:56907790 Present                                    | Present                        | 298                                                  |
| 147    | rs928222537(G/C)        | chr4:56907803 Present                                    | Present                        | 298                                                  |
| 148    | rs938247687(G/A)        | chr4:56907809 Present                                    | Present                        | 298                                                  |
| 149    | rs1047872882(G/GGC/GGT)** | chr4:56907870–56907874 Present                | Present                        | 304                                                  |
|        | Gene                     | SH2B adaptor protein 3 (SH2B3)                          |                               |                                                      |
| 150    | Island:2,214            | rs960136772(G/A)* Present                               | chr12:111405136 Present       | Present                                              | 150  |
| 151    | rs53845017(C/T)**       | chr12:111405235 Present                                  | Present                        | 0                                                    |
| 152    | rs922413526(G/A)**      | chr12:111405236 Present                                  | Present                        | 0                                                    |
| 153    | rs999730506(G/A)*       | chr12:111405248 Present                                  | Present                        | 114                                                  |
| 154    | rs741737025(C/T)        | chr12:111405270 Present                                  | Present                        | 214                                                  |
| 155    | Island:2,796            | rs421259159(G/A) Present                                 | chr12:111405355 Present       | Present                                              | 778,754 |
| 156    | rs1029968061(C/T)*      | chr12:11140569 Present                                  | Present                        | 778                                                  |
| 157    | rs1042427385(G/A)       | chr12:111405693 Present                                  | Present                        | 796                                                  |
| 158    | rs673087655(G/C)        | chr12:111405694 Present                                  | Present                        | 796                                                  |
| 159    | rs899785538(C/A)        | chr12:111405709 Present                                  | Present                        | 796                                                  |
| 160    | rs75390213(G/A)         | chr12:111405712 Present                                  | Present                        | 796                                                  |
| 161    | rs43838140(G/A)         | chr12:111405728 Present                                  | Present                        | 796                                                  |
| 162    | rs982567306(G/T)        | chr12:111405743 Present                                  | Present                        | 796                                                  |
| 163    | rs151395998(C/A)        | chr12:111405750 Present                                  | Present                        | 796                                                  |
| 164    | rs1029498594(G/A)       | chr12:111405764 Present                                  | Present                        | 796                                                  |

Continued
variants and epigenetic modifications of CGIs at promoter regions autonomously have a great impact on the regulation of gene expression.

The genes selected for the study are influencing the various pathways such as lipid metabolism and cholesterol homeostasis (ACAT1, APOB, APOE, LDLR, PCSK9, SORT1), oxidative stress (CYBA, KSR2, PHOX2A), apoptosis (FAS, REST, SORT1), inflammation & angiogenesis (FLT1, SH2B3), maintenance of extracellular matrix and vascular smooth muscle cells (MMP9 & TIMP1). Elucidation of gene expression regulating mechanisms have a significant role in understanding the pathogenesis and risk prediction of several diseases21–28,30–38,40–51.

Accumulating evidences have shown that the genetic variants of the APOE, LDLR, SH2B3, TIMP1, MMP9 genes were found to have an impact on risk of the diseases like diabetes, coronary artery disease, acute lymphoblastic leukemia, cancer, lung cancer, etc21,36,45,52,81–87.

Dayeh, T. A. et al., have reported that CpG SNVs are associated with differential DNA methylation and gene expression in human pancreatic islets in type 2 diabetes88. Hawkins, N. J. et al., and Rapkins, R. W. et al., studied the association of O6-methylguanine-DNA methyltransferase (MGMT) gene rs16906252 polymorphism with DNA methylation and reported that the individuals with MGMT rs16906252 T-allele has 5.5 folds and 2.64 folds highly methylated than C-allele individuals in colorectal cancer and glioblastoma patients respectively67,68. Another study on effect of RAD50 gene DNase I hypersensitive site? (RH57) region rs2240032 polymorphism on DNA methylation has shown that, it is significantly affecting the Sq31 locus IL13 gene promoter DNA methylation.

| S. No. | CpG island and size (bp) | Single nucleotide variations (SNVs) | CpG coordinates on chromosome | CpG island status with | Wild type allele | Variant allele | Change in CpG island size (bp) |
|--------|--------------------------|-----------------------------------|-----------------------------|------------------------|----------------|----------------|-----------------------------|
| 165    | rs974278790C/A/T—A or C/A/T—T | chr1:114105774                   | Present                      | Present                | 796            |                |
| 166    | rs523267698G/T            | chr1:114105775                   | Present                      | Present                | 796            |                |
| 167    | rs1015689151G/A           | chr1:114105795                   | Present                      | Present                | 796            |                |
| 168    | rs917942737G/C           | chr1:114105807                   | Present                      | Present                | 796            |                |
| 169    | rs560612237C/T          | chr1:114105823                   | Present                      | Present                | 796            |                |
| 170    | rs1005740439G/C         | chr1:114105854                   | Present                      | Present                | 796            |                |
| 171    | rs1054243996C/T         | chr1:114105879                   | Present                      | Present                | 796            |                |
| 172    | rs808066828G/C         | chr1:114105889                   | Present                      | Present                | 796            |                |
| 173    | rs1015267150G/T        | chr1:114105900                   | Present                      | Present                | 796            |                |
| 174    | rs962487794C/T         | chr1:114105903                   | Present                      | Present                | 796            |                |
| 175    | rs686119397G/C/T—C or G/C/T—T | chr1:114105908                   | Present                      | Present                | 796            |                |
| 176    | rs1033875297C/T       | chr1:114105929                   | Present                      | Present                | 796            |                |
| 177    | rs959781377G/C        | chr1:114105930                   | Present                      | Present                | 796            |                |
| 178    | rs992435354G/A        | chr1:114105940                   | Present                      | Present                | 796            |                |

Table 1. Single nucleotide variations (SNVs) at CpG sites associated with loss or change in the size of CpG island. **indicates the SNVs abolish CpG island, *indicates the SNVs change CpG island size; rs: reference sequence

variants and epigenetic modifications of CGIs at promoter regions autonomously have a great impact on the regulation of gene expression.

The genes selected for the study are influencing the various pathways such as lipid metabolism and cholesterol homeostasis (ACAT1, APOB, APOE, LDLR, PCSK9, SORT1), oxidative stress (CYBA, KSR2, PHOX2A), apoptosis (FAS, REST, SORT1), inflammation & angiogenesis (FLT1, SH2B3), maintenance of extracellular matrix and vascular smooth muscle cells (MMP9 & TIMP1). Elucidation of gene expression regulating mechanisms have a significant role in understanding the pathogenesis and risk prediction of several diseases21–28,30–38,40–51.

Accumulating evidences have shown that the genetic variants of the APOE, LDLR, SH2B3, TIMP1, MMP9 genes were found to have an impact on risk of the diseases like diabetes, coronary artery disease, acute lymphoblastic leukemia, cancer, lung cancer, etc21,36,45,52,81–87.

Dayeh, T. A. et al., have reported that CpG SNVs are associated with differential DNA methylation and gene expression in human pancreatic islets in type 2 diabetes88. Hawkins, N. J. et al., and Rapkins, R. W. et al., studied the association of O6-methylguanine-DNA methyltransferase (MGMT) gene rs16906252 polymorphism with DNA methylation and reported that the individuals with MGMT rs16906252 T-allele has 5.5 folds and 2.64 folds highly methylated than C-allele individuals in colorectal cancer and glioblastoma patients respectively67,68. Another study on effect of RAD50 gene DNase I hypersensitive site? (RH57) region rs2240032 polymorphism on DNA methylation has shown that, it is significantly affecting the Sq31 locus IL13 gene promoter DNA methylation.
To date, there are very limited studies reported on the effect of single nucleotide variations at CpG sites on CpG island existence, size and their respective methylation status. Furthermore, Palumbo, D. et al., reported that the methylation variability depends upon the CpG cluster density such as high density regions showing low levels of CpG methylation variability, while intermediate density and low density regions have increasingly higher levels of CpG methylation.

Study by Zhou, D. et al., identified 9,42,429 loci for CpG SNPs from HapMap phase II and observed that 51.9% were CpG gain-SNPs and 47.9% were CpG-loss-SNPs and his successive studies on tumor tissues of colon cancer have shown that CpG-loss-SNPs are lowering the methylation in tumor tissues and inferred that the SNPs at CpG sites are significantly associated with traits in cancers. In addition, Wang, Z. et al., identified novel functional CpG-SNPs by conditional false discovery rate (cFDR) analysis from statistical data of two large GWAS of type 2 DM and CAD. Among them, 13 CpG-SNPs of DM, 15 CpG-SNPs of CAD have a significant methylation quantitative trait locus effect and increased susceptibility to disease.

In view of the above, the present study has been designed to analyze the impact of single nucleotide variations at CpG sites in promoter CpG islands of ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1 and TIMP1 genes.

![Figure 3](https://www.nature.com/scientificreports/)  
**Figure 3.** Single nucleotide variations showing influence on CGIs status & size for ACAT1, APOB, APOE, CYBA, FAS, FLT1, KSR2, LDLR, MMP9, PCSK9, PHOX2A, REST, SH2B3, SORT1 and TIMP1 genes.

### Table 2. Transcription factors associated with the single nucleotide variations (SNVs) abolishing CGIs.

| Gene                     | Single nucleotide variations (rs number; variation) | Transcription factors |
|--------------------------|----------------------------------------------------|-----------------------|
| Low density lipoprotein receptor (LDLR) | rs1026272072,G/T* | C/EBPαp, C/EBPβet |
|                          | rs88768252,G/T* | No TF                |
|                          | rs1006494933,G/A* | No TF, GATA-1, Oct-1 |
|                          | rs532491368,G/A | No TF                |
| Matrix metalloproteinase 9 (MMP9) | rs139620474,C/A/T | No TF                |
|                          | rs370014825,G/T* | No TF, Sp1           |
|                          | rs201069991,G/A | No TF                |
|                          | rs1014494202,C/T* | Sp1, BRF-1           |
|                          | rs146719297,G/A | Sp1                  |
| SH2B adaptor protein 3 (SH2B3) | rs538445017,C/T | Tra-1                |
|                          | rs922413124,G/A* | Sp1                  |
| Tissue inhibitor of metalloproteinase 1 (TIMP1) | rs779329701,G/A* | Fgr-1, NF-1          |
|                          | rs993047389,G/A | Sp1                  |
|                          | rs376386551,C/T* | Sp1, N-Myc           |
|                          | rs926004266,G/A | Sp1                  |
| Apolipoprotein E (APOE) | rs7694488,C/T* | Sp1                  |

*change in transcription factor binding; No TF: No transcription factor
Figure 4. Concentric bipartites by GeneMANIA represents co-expression networks of A. APOE B. LDLR C. MMP9 D. SH2B3 E. TIMP1 genes.
Figure 5. Linear bipartite by GeneMANIA represents Co-expression networks of multi gene queries for APOE, LDLR, MMP9, SH2B3 and TIMP1.
Figure 6. Gene ontology (GO) annotation. The top 10 GO terms in each category. (A) Biological process. (B) Cellular component. (C) Molecular function. (D) Disease class. (E) Functional annotation clustering.
induction in cancers. A study reported that methylation of APOE is significantly lower in men with coronary heart disease than healthy control men and is inversely proportional to APOE plasma levels. Thus, it is considered that the DNA methylation is a potential factor for regulation of APOE gene expression. In the present study, we have observed that APOE rs769448 has abolishing the CGI existence that might influence the methylation pattern and further may regulate the gene expression. The GO enrichment analysis has shown that the APOE gene is a key regulator in the cholesterol metabolism and transportation contributing to the initiation and progression of multiple diseases.

Similarly, Low density lipoprotein receptor (LDLR) gene codes a cell surface LDL receptor protein mediating endocytosis of LDL particles regulate cholesterol levels. Evidences suggest that elevated circulating cholesterol levels are involved in the coronary artery disease, cancer growth promotion and progression. Ghose, S. et al. reported that LDLR gene undergoes hypomethylation and induces an increased expression which subsequently decreases the LDL levels and reduces the risk of CAD. In the present study, we have observed that 31% of CpG SNVs abolished the CGI existence and ~ 44% decreased the size of CGI. The abolishment and reduced CGI size, decreases the possibility of methylation and inversely increases the gene expression. The increased gene expression associates with decreased LDL-cholesterol levels and lead to reduced risk of diseases.

Furthermore, Src homology 2-B adaptor protein 3 (SH2B3) plays a critical role in haematoopoiesis and acts as a negative regulator of several tyrosine kinases and cytokine signaling. SH2B3 was associated with diseases like atherosclerosis and thrombosis, cancers, diabetes, etc. A recent study on Celiac disease (CeD) revealed that the expression of SH2B3 is influenced by the methylation and it is reported that hypomethylation is associated with higher expression of the genes in CeD patients than controls. The methylated DNA sequence is showing differences in binding of regulatory elements to control the expression of gene at mRNA level. The present study investigations have shown SH2B3 gene promoter has 7% CGI abolishing SNVs besides 17% size reducing SNVs. The differences in CGI existence, binding of transcription factors and CGI size influences the methylation pattern to regulate the expression. According to gene ontology disease class term SH2B3 is playing a significant role in metabolic, cardiovascular and immune diseases.

In recent years, there is a growing interest on matrix metalloproteinase (MMP) family to understand their significant association with various disease pathophysiologies such as cancers, CAD and DM. MMP9 and Tissue inhibitors of metalloproteinases 1 (TIMP1) were known to be associated with the risk of cardiovascular disease and several cancers. A study on MMP9 promoter methylation suggested that serum circulating levels were inversely associated with methylation level in Diabetic nephropathy patients. MMP9 demethylation increases its serum circulating levels that might be accompanying with the incidence and prognosis of diabetic nephropathy. Tissue inhibitors of metalloproteinases (TIMPs) are inhibitors of the MMPs involved in extracellular matrix degradation. In chronic periodontitis, TIMP1 promoter methylation positively correlated with severity of the disease. In another study, DNA methylation in TIMP3 gene contributed to its lower expression and eventually lead to metastasis of oral cancer. In the present analysis, ~ 18% of MMP9 and ~ 67% of TIMP1 CpG SNVs have shown for the loss of CGIs, further 57% of MMP9 and 33% of TIMP1 CpG SNVs reduced the size of CGI. GO enrichment analysis of MMP9, TIMP1 revealed that these two genes are playing a significant role in metabolic, neurological, cardiovascular diseases and cancers. Altogether, abolishment and reduction of CGI size, differential binding of TFs could influence their gene expression in ECM remodelling and degradation which can further mediate the pathological conditions of various diseases.

Further, 50% of ACAT1, ~ 67% of APOB, 57% of CYBA, ~ 92% of FAS, 50% of FLT1, ~ 13% of KSR2, ~ 44% of LDLR, ~ 36% of MMP9, 50% of PCSK9, 36% of PHOX2A, 40% of REST, ~ 14% of SH2B3, ~ 13% of SORT1 and 33% of TIMP1 SNVs are altering the size of CGIs. Among all the 200 SNVs in the genes under study, we have observed that approximately 9% of SNVs at CpG site are abolishing the existence of CpG island; whereas 35% are decreasing the size of CGIs. Consequently, loss of CGI & decreased CGI size leads to the intermittent and asymmetrical DNA methylation pattern of gene which can regulate the expression of genes by affecting binding of transcription factors to the promoter.

The findings of the study suggest that the SNVs at CpG sites in the promoter region regulating CGI existence and size might influence the DNA methylation status and expression of genes that take part in molecular pathways associated with multifactorial diseases like diabetes mellitus, cardiovascular diseases, cancers, etc. The insights of the present study may pave the way for new experimental studies to undertake challenges in DNA methylation, gene expression and protein assays.

Limitations
A primary limitation of the study is that this is an in silico study, designed to know the impact of single nucleotide variations at CpG sites on CpG island existence, size and their respective DNA methylation pattern and gene expression. Another limitation of the study is that the genes are randomly selected from the various pathways to test the hypothesis. Therefore, the predicted results should be essentially validated using experimental analyses such as genotyping, DNA methylation and their subsequent gene expression assays for further correlation with disease phenotypes.

Data availability
The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 3 June 2021; Accepted: 3 January 2022
Published online: 04 March 2022
43. Forouzanfar, M. H. et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1659–1724 (2016).

44. Brookwell, T. F. et al. A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121**, 837–848 (2005).

45. Hong, L. et al. Role of SH2B3 R262W gene polymorphism and risk of coronary heart disease A PRISMA-compliant meta-analysis. *Medicine (United States)* **97**, 1–7 (2018).

46. Clinical, E. Leak. Res. **72**, 67–70 (2018).

47. Ey, L. et al. The carriage of the type 1 diabetes-associated R262W variant of human LNK correlates with increased proliferation of peripheral blood monocytes in diabetic patients. *Pediatr. Diabetes* **12**, 127–132. https://doi.org/10.1111/j.1399-5448.2010.00656.x (2011).

48. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

49. Biscetti, F. et al. Sortilin levels are associated with peripheral arterial disease in type 2 diabetic subjects. *Cardiovasc. Diabet.** 18**, 1–8 (2019).

50. Roselli, S. et al. Sortilin is associated with breast cancer aggressiveness and contributes to tumor cell adhesion and invasion. *Oncotarget* **6**, 10473–10486 (2015).

51. Meng, C. et al. TIMP-1: a novel serum biomarker for the diagnosis of colorectal cancer: A meta-analysis. *PLoS ONE* **13**, 1–15 (2018).

52. Cheng, G. et al. Higher levels of TIMP-1 expression are associated with a poor prognosis in triple-negative breast cancer. *Mol. Cancer* **15**, 1–13 (2016).

53. McPherson, R. & Tybaaer-Hansen, A. Genetics of coronary artery disease. *Circ. Res.** 118**, 564–578 (2016).

54. Liang, Y. et al. Homocysteine-mediated cholesterol efflux via ABCA1 and ACAT1 DNA methylation in THP-1 monocyte-derived foam cells. *Acta Biochim. Biophys. Sin. (Shanghai)* **45**, 220–228 (2013).

55. Hedman, Å. K. et al. Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. *Circ. Cardiovasc. Genet.* **10**, 1–13 (2017).

56. Song, Y. et al. DNA methylation-mediated silencing of FLT1 in parthenogenetic porcine placentas. *Placenta* **58**, 86–89 (2017).

57. Huang, Y. S., Zhi, Y. F. & Wang, S. R. Hypermethylation of estrogen receptor-a gene in atheromatosis patients and its correlation with homocysteine. *Pathophysiology* **16**, 259–265 (2009).

58. Chicoine, É. et al. Evidence for the role of promoter methylation in the regulation of MMP-9 gene expression. *Biochem. Biophys. Res. Commun.* **297**, 765–772 (2002).

59. Lohoff, F. W. et al. Methylicom profiling and replication implicates deregulation of PCSK9 in alcohol use disorder. *Mol. Psychiatry* https://doi.org/10.1038/mp.2017.168 (2018).

60. González-Mundo, I. et al. DNA methylation of the RE-1 silencing transcription factor in peripheral blood mononuclear cells and expression of antioxidant enzyme in patients with late-onset Alzheimer disease. *Exp. Gerontol.* **136**, 110951 (2020).

61. Cielo, D. et al. Combined analysis of methylation and gene expression profiles in separate compartments of small bowel mucosa identified celiac disease patients' signatures. *Sci. Rep.* **9**, 1–12 (2019).

62. Sung, H. Y. et al. aberrant promoter hypomethylation of sortilin 1: A moyamoya disease biomarker. *J. Stroke* **20**, 350–361 (2018).

63. Lu, X., Lu, J., Teng, W., Zhao, C. & Ye, X. Quantitative evaluation of MMP-9 and TIMP-1 promoter methylation in chronic periodontitis. *DNA Cell Biol.* **37**, 168–173 (2018).

64. Zhou, D. et al. Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. *Oncotarget* https://doi.org/10.18632/oncotarget.36580 (2015).

65. Wang, Z. et al. Identification of novel functional CpG-SNPs associated with type 2 diabetes and coronary artery disease. *Mol. Genet. Genomics* **295**, 607–619 (2020).

66. Samy, M. D., Yavorski, J. M., Mauro, J. A. & Blanck, G. Impact of SNPs on CpG Islands in the MYC and HRAS oncogenes and in a wide variety of tumor suppressor genes: A multi-cancer approach. *Cell Cycle* **15**, 1572–1578 (2016).

67. Hawkins, N. J. et al. MGMT methylation is associated primarily with the germline CT SNP (rs16906252) in colorectal cancer and normal colonic mucosa. *Mod. Pathol.* **22**, 1588–1599 (2009).

68. Rapkins, R. W. et al. The MGMT promoter SNP rs16906252 is a risk factor for MGMT methylation in glioblastoma and is predictive of response to temozolomide. *Neuro. Oncol.* **17**, 1589–1598 (2015).

69. Schiec, M. et al. A polymorphism in the TH2 locus control region is associated with changes in DNA methylation and gene expression. *Allergy Eur. J. Allergy Clin. Immunol.* **69**, 1171–1180 (2014).

70. Vohra, M., Sharma, A. R., Prabhu, B. N. & Rai, P. S. SNPs in sites for DNA methylation, transcription factor binding, and miRNA targets leading to allelic-specific gene expression and contributing to complex disease risk: A systematic review. *Public Health Genomics* **23**, 155–170 (2021).

71. Drees, R., Ambrosini, G., Groux, R., Perier, R. C. & Bucher, P. The eukaryotic promoter database in its 30th year: Focus on non-vertebrate organisms. *Nucleic Acids Res.* **45**, D51–D55 (2017).

72. Li, L. C. & Daihi, R. MethPrimer: Designing primers for methylation PCRs. *Bioinformatics* **18**, 1427–1432 (2001).

73. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).

74. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 Jul 21]. Available from: https://www.ncbi.nlm.nih.gov/gene/ (2020).

75. Hunt, S. E. et al. Ensembl variation resources. *Database (Oxford)*. **2018**, 1–12 (2018).

76. Grabe, N. AliBaba2: Context specific identification of transcription factor binding sites. In *Silico Biol.* **2**, 1–15 (2002).

77. Montejo, J., Zuberi, K., Rodriguez, H., Badger, G. D. & Morris, Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Res* **3**, 1–7 (2014).

78. Franz, M. et al. GeneMANIA update 2018. *Nucleic Acids Res.* **46**, W60–W64 (2018).

79. Messerschmidt, D. M., Knowles, B. & Solter, D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.* **28**, 812–828 (2014).

80. Ornella Affinit et al., 2016 Modeling DNA methylation by analyzing the individual configurations of single molecules_Enhanced Reader.pdf.

81. Lai, C. Y. et al. Association of tissue inhibitor of metalloproteinase-1 genotypes with lung cancer risk in Taiwan. *Anticancer Res.* **36**, 155–160 (2016).

82. Gomez-Delgado, F. et al. Apolipoprotein E genetic variants interact with Mediterranean diet to modulate postprandial hypertriglyceridemia in coronary heart disease patients: CORDIOPREV study. *Eur. J. Clin. Invest.* **49**, 1–9 (2019).

83. Grettarsdottir, S. et al. A Splice Region Variant in LDLR Lowers Non-High Density Lipoprotein Cholesterol and Protects against Coronary Artery Disease. *PLoS Genet.* **11**, 1–20 (2015).

84. Li, Y. Y. et al. Matrix metalloproteinase-9 gene -1562C>T gene polymorphism and coronary artery disease in the Chinese Han population: A meta-analysis of 5468 subjects. *Front. Physiol.* **7**, 1–10 (2016).

85. Weng, Y. et al. The association between extracellular matrix metalloproteinase inducer polymorphisms and coronary heart disease: A potential way to predict disease. *DNA Cell Biol.* **39**, 244–254 (2020).

86. Kral, B. G. & Becker, L. C. Genetics of Coronary Disease. *Transil. Res. Coron. Artery Dis. Pathophysiol. to Treat* **81–101 (2016), https://doi.org/10.1016/B978-0-12-002385-3.00008-5.
87. Rybakowski, J. K. Matrix metalloproteinase-9 (MMP9)—A mediating enzyme in cardiovascular disease, cancer, and neuropsychiatric disorders. *Cardiomyocyte*. **2009**, 1–7 (2009).
88. Dayeh, T. A. *et al.* Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia* **56**, 1036–1046 (2013).
89. Palumbo, D., Affinito, O., Monticelli, A. & Cocozza, S. DNA Methylation variability among individuals is related to CpGs cluster density and evolutionary signatures. *BMC Genomics* **19**, 1–9 (2018).
90. Vogel, T. *et al.* Apoliprotein E: A potent inhibitor of endothelial and tumor cell proliferation. *J. Cell. Biochem.* **54**, 299–308 (1994).
91. Hossain, T. *et al.* Matrix metalloproteinases in human. *HHS Public Access*. *Physiol. Behav.* **176**, 139–148 (2015).
92. Cross, Sarah J. *et al.* Matrix metalloproteinases in human. *HHS Public Access*. *Physiol. Behav.* **176**, 100–106 (2016).
93. Yue, S. *et al.* Cholesteryl ester accumulation induced by PTEN loss and PI3K/AKT activation underlies human prostate cancer aggressiveness. *Cell Metab.* **19**, 393–406 (2014).
94. Ghose, S. *et al.* Investigating Coronary Artery Disease methylome through targeted bisulfite sequencing. *Gene* **721**, 141407 (2019).
95. Wang, W. *et al.* Matrix metalloproteinases in human. *HHS Public Access*. *Physiol. Behav.* **119**, 1–25 (2017).
96. Maslah, N., Cassinat, B., Verger, E., Kiladjian, J. J. & Velazquez, L. The role of LNK/SH2B3 genetic alterations in myeloproliferative neoplasms and other hematological disorders. *Leukemia* **31**, 1661–1670 (2017).
97. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes The Wellcome Trust Case Control Consortium. *Nat. Genet.* **39**, 857–864 (2007).
98. Mittal, B., Mishra, A., Srivastava, A., Kumar, S. & Garg, N. Matrix metalloproteinases in coronary artery disease. *Adv. Clin. Chem.** 64**, 1–72 (2014).
99. Wu, Z. S. *et al.* Prognostic significance of MMP-9 and TIMP-1 serum and tissue expression in breast cancer. *Int. J. Cancer* **122**, 2050–2056 (2008).
100. Olivares-Urbano, M. A. *et al.* Matrix metalloproteinases and TIMPs as prognostic biomarkers in breast cancer patients treated with radiotherapy: A pilot study. *J. Cell. Mol. Med.* **24**, 139–148 (2020).
101. Yukawa, N. *et al.* Impact of plasma tissue inhibitor of matrix metalloproteinase-1 on long-term survival in patients with colorectal cancer. *Oncology* **72**, 205–208 (2008).
102. Pludowski, P. *et al.* The role of EBV in thyroid disease. *Endokrynol. Pol.* **64**, 319–327 (2013).
103. Su, C. W. *et al.* Loss of TIMP3 by promoter methylation of Sp1 binding site promotes oral cancer metastasis. *Cell Death Dis.* **10**, 1–17 (2019).

**Acknowledgements**
This work was supported by University Grants Commission (UGC), New Delhi, India: Basic Science Research (BSR) Fellowship (FNo. 4-1/2006/BSR/5-78/2007/BSR/2013-2014/03), Indian Council of Medical Research (ICMR)-Senior Research Fellowship (FNo. 3/1/2(10)/CVD/2019-NCD-II), Osmania University—Department of Science & Technology-Promotion of University Research and Scientific Excellence (OU DST PURSE-II Programme (C-DST-PURSE-II/23/2017)).

**Author contributions**
N.S.: Conceptualization, Methodology, Data Curation, Formal analysis, Writing—Original Draft, Visualization; C.L.K.: Methodology, Formal analysis, Data Curation; K.P.: Formal analysis, Validation; K.K.: Methodology, Validation; R.K.: Methodology, Formal analysis, Data Curation, Writing—Original Draft, Visualization; S.R.H.: Conceptualization, Methodology, Data Curation, Formal analysis, Writing—Original Draft, Visualization; K.P.: Formal analysis, Validation; S.R.H.: Conceptualization, Methodology, Data Curation, Formal analysis, Writing—Original Draft, Visualization; N.S.: Conceptualization, Methodology, Data Curation, Formal analysis, Writing—Original Draft, Visualization; C.L.K.: Methodology, Formal analysis, Data Curation; K.P.: Formal analysis, Validation; K.K.: Methodology, Validation; R.K.: Methodology, Validation; C.P.: Bioinformatics analysis, Revision; S.R.H.: Conceptualization, Supervision and Writing.

**Competing interests**
The authors declare no competing interests.

**Additional information**
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05198-8.

**Correspondence** and requests for materials should be addressed to S.H.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022