Structural, functional and molecular dynamics analysis of *cathepsin B* gene SNPs associated with tropical calcific pancreatitis, a rare disease of tropics

Garima Singh [1] , **Sri Krishna jayadev Magani** [1] , **Rinku Sharma** [1] , **Basharat Bhat** [1] , **Ashish Shrivastava** [1] , **Madhusudhan Chinthakindi** [2] , **Ashutosh Singh** [Corresp. 1]

[1] Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, GREATER NOIDA, UTTAR PRADESH, India

[2] Department of Surgical Gastroenterology, Osmania General Hospital, Hyderabad, India

Corresponding Author: Ashutosh Singh
Email address: ashutosh.singh@snu.edu.in

Tropical Calcific Pancreatitis (TCP) is a neglected juvenile form of chronic non-alcoholic pancreatitis. *Cathepsin B* (CTSB), a lysosomal protease involved in the cellular degradation process, has recently been studied as a potential candidate gene in the pathogenesis of TCP. According to *Cathepsin B* hypothesis, mutated CTSB can lead to premature intracellular activation of trypsinogen, a key regulatory mechanism in pancreatitis. So far, CTSB mutations have been studied in pancreatitis and neurodegenerative disorders, but little is known about the structural and functional effect of variants in CTSB. In this study, we investigated the effect of single nucleotide variants (SNVs) specifically associated with TCP, using molecular dynamics and simulation algorithms. There were two non-synonymous variants (L26V and S53G) of CTSB, located in the propeptide region. We tried to predict the effect of these variants on structure and function using multiple algorithms: SIFT, Polyphen2, PANTHER, SDM sever, i-Mutant2.0 suite, mCSM algorithm, and Vadar. Further, using databases like miRdbSNP, PolymiRTS, and miRNASNP, two SNPs in the 3'UTR region were predicted to affect the miRNA binding sites. Structural mutated models of nsSNP mutants (L26V and S53G) were prepared by MODELLER v9.15 and evaluated using TM-Align, Verify 3D, ProSA and Ramachandran plot. The 3D mutated structures were simulated using GROMACS 5.0 to predict the impact of these SNPs on protein stability. The results from *in silico* analysis and molecular dynamics simulations suggested that these variants in the propeptide region of *Cathepsin B* could lead to structural and functional changes in the protein and thus could be pathogenic. Hence, the structural and functional analysis results have given interim conclusions that these variants can have a deleterious effect in TCP pathogenesis, either uniquely or in combination with other mutations. Thus, it could be extrapolated that *Cathepsin B* gene can be screened in samples from all TCP patients in future, to decipher the distribution of variants in patients.

1   **Structural, functional and molecular dynamics analysis of _Cathepsin B_ gene**

2   **SNPs associated with tropical calcific pancreatitis, a rare disease of tropics.**

3   Garima Singh[1], Shri Krishna jayadev Magani[1], Rinku Sharma[1], Basharat Bhat[1], Ashish

4   Shrivastava[1], Madhusudhan Chinthakindi[2] and Ashutosh Singh[1*]

5   [1] School of Natural Sciences, Department of Life Sciences, Shiv Nadar University, Greater

6   Noida, UP, India

7   [2] Department of Surgical Gastroenterology, Osmania General Hospital, Hyderabad, India.

8   * Corresponding Authors (ashutosh.bio@gmail.com)

9   **Abstract**

10   Tropical Calcific Pancreatitis (TCP) is a neglected juvenile form of chronic non-alcoholic

11   pancreatitis. _Cathepsin B_ (CTSB), a lysosomal protease involved in the cellular degradation

12   process, has recently been studied as a potential candidate gene in the pathogenesis of TCP.

13   According to _Cathepsin B_ hypothesis, mutated CTSB can lead to premature intracellular

14   activation of trypsinogen, a key regulatory mechanism in pancreatitis. So far, CTSB mutations

15   have been studied in pancreatitis and neurodegenerative disorders, but little is known about the

16   structural and functional effect of variants in CTSB. In this study, we investigated the effect of

17   single nucleotide variants (SNVs) specifically associated with TCP, using molecular dynamics

18   and simulation algorithms. There were two non-synonymous variants (L26V and S53G) of

19   CTSB, located in the propeptide region. We tried to predict the effect of these variants on

20   structure and function using multiple algorithms: SIFT, Polyphen2, PANTHER, SDM sever, i-

21   Mutant2.0 suite, mCSM algorithm, and Vadar. Further, using databases like miRdbSNP,

22   PolymiRTS, and miRNASNP, two SNPs in the 3'UTR region were predicted to affect the

23 miRNA binding sites. Structural mutated models of nsSNP mutants (L26V and S53G) were

24 prepared by MODELLER v9.15 and evaluated using TM-Align, Verify 3D, ProSA and

25 Ramachandran plot. The 3D mutated structures were simulated using GROMACS 5.0 to predict

26 the impact of these SNPs on protein stability. The results from *in silico* analysis and molecular

27 dynamics simulations suggested that these variants in the propeptide region of *Cathepsin B* could

28 lead to structural and functional changes in the protein and thus could be pathogenic. Hence, the

29 structural and functional analysis results have given interim conclusions that these variants can

30 have a deleterious effect in TCP pathogenesis, either uniquely or in combination with other

31 mutations. Thus, it could be extrapolated that *Cathepsin B* gene can be screened in samples from

32 all TCP patients in future, to decipher the distribution of variants in patients.

33 **Introduction**

34 Pancreatitis is a multifactorial, heterogeneous disease with enigmatic etiologies. It is an

35 inflammatory condition leading to morphological changes in the pancreas, causing–pain and

36 functional abnormalities. Alcohol(1), malnutrition(2), gallstones(3), familial clustering (4) and

37 sometimes severe infections (5) have been observed to be significant causes of pancreatitis.

38 Pancreatitis is broadly classified (6) as acute and chronic. Tropical Calcific Pancreatitis (TCP)

39 (7) is a juvenile form of chronic calcific non-alcoholic pancreatitis. It is a form of Idiopathic

40 Chronic Pancreatitis (ICP), mostly reported in developing tropical countries. The phenotypic

41 heterogeneity (8) includes abdominal pain, ductal dilation, large pancreatic calculi, and

42 pancreatic atrophy. The genetic heterogeneity related to TCP is still unexplored. Fibrocalculous

43 pancreatic diabetes (FCPD) (9), a unique form of diabetes, is the unique secondary feature of

44 TCP. TCP progresses gradually to FCPD and then at a later age, TCP patient suffers from

45 pancreatic cancer (10).

46   The pathophysiology of the pancreas is composed of an exocrine gland which is responsible for

47   digesting food and an endocrine gland critical for glucose homeostasis. Trypsinogen, cathepsins,

48   serine proteases, calcium sensing receptors are some of the essential genes for pancreatic

49   function regulation. According to the trypsin-centred theory of pancreatitis, trypsinogen, a key

50   zymogen in pancreatic juice and a key regulator of digestion, exhibits premature activation in

51   acinar cells of the pancreas during pancreatitis. This aberrant activation of trypsinogen leads to

52   activation of other zymogens in the pancreas itself, thereby resulting in inflammation and

53   autodigestion of pancreas. Although TCP is a distinct form of pancreatitis without a known

54   cause, what remains undebated is the initial step during initiation of TCP, which is the premature

55   activation of trypsinogen in the pancreas itself. The mortality rate in TCP is as high as 17%, and

56   patients majorly die because of pancreatic cancer at a later stage (10).

57   Recently, we have built a database (11), mutTCPdb, which is a comprehensive database, giving

58   details about the genes and variants predicted to be associated with TCP until now. Activity of

59   trypsin inside the pancreas is the primary critical factor in pathogenesis of TCP, and all the risk

60   genes predicted to date, are known to regulate trypsin activity like chymotrypsin C (*CTRC*),

61   cystic fibrosis transmembrane conductance regulator (*CFTR*), serine protease inhibitor Kazal-

62   type I (*SPINK1*) and *Cathepsin B* (*CTSB*). According to "*Cathepsin B* hypothesis", *CTSB* plays

63   an essential role in the premature activation of trypsinogen in the pancreas, primarily due to

64   colocalization of *Cathepsin B* and zymogens (12). The precise rationale behind this

65   colocalization is yet unknown. The reason could be aberrant trafficking mechanism of

66   procathepsin B, due to mutations in procathepsin B or deleterious mutations in the molecules

67   associated with the trafficking of procathepsin B, in a diseased state.

68 A research article in 2006, described two missense mutations in CTSB (L26V and S53G),

69 identified in TCP patients from Asian Institute of Gastroenterology, Hyderabad (India) (13). The

70 minor allele frequency (MAF) of variants L26V and S53G in TCP patients were 0.46 and 0.09

71 respectively. Also in 2008, an article suggested that coexistence of variants in transcription factor

72 7-like 2 (*TCF7L2*), *SPINK1* and *CTSB* (L26V), might lead to exocrine damage in TCP and

73 determine the onset of FCPD (14). The analysis in this paper (14) was performed with TCP

74 patients and control population from Dravidian and Indo-European ethnicities. There is another

75 article which described a missense mutation (p. Q334P) in cathepsin B gene discovered in

76 chronic pancreatitis patients but not in TCP patients (15). An article in 2014 illustrated no

77 association of L26V mutation with TCP (16). In this paper, statistical significance tests have

78 indicated the lack of association of L26V mutation with TCP, but this mutation was observed in

79 7 out of 150 TCP patients. Hence, the association of this mutation with TCP cannot be

80 completely disregarded (16). Although researchers have identified SNVs in CTSB gene

81 observed in TCP patients, lacunae lie in the information about the functional effect of these

82 SNVs in the pathogenesis of TCP.

83 Human *Cathepsin B* (catB, E.C 3.4.22.1) is a lysosomal cysteine protease which is involved in

84 several cellular processes like protein degradation, extracellular matrix degradation, regulatory

85 mechanisms, cell death, autophagy and antigen representation (17). It belongs to papain

86 superfamily and acts both as an endopeptidase and as an exopeptidase. *Cathepsin B* is

87 synthesized as an inactive proenzyme (*cathepsin B*) and is activated by other proteases and by

88 autocatalytic processes (18). *Procathepsin B* (length of protein = 339aa) has an N-terminus

89 propeptide of 62 amino acid length from Arg-Lys (18-79 residues). Signal sequence (1-17

90 residues) and post-translational glycosylation modification (19) targets *cathepsin B* to

91 endosomes/lysosomes (20) via mannose-6-phosphate receptor pathway. Propeptide exhibits an

92 essential role in the processing and maturation of *cathepsin B*. It acts as (a) a scaffold for

93 catalytic domain during protein folding, (b) involved in intracellular trafficking of *cathepsin B* to

94 lysosome after N-terminal glycosylation and phosphorylation and (c) as a high-affinity reversible

95 inhibitor for the premature activation of zymogen. The crystal structure of procathepsin B [PDB

96 ID: 3PBH] has a propeptide region [ArP1 to LysP62], and main chain [Leu1 to Asp254] enzyme

97 residues (21). The main chain has two domains (R and L domains) with active site residues

98 Cys29 and His199, located at the interdomain cleft. The propeptide siting in the active site cleft

99 is in reverse direction to that of the substrate, thus suggesting its role as an inhibitor. The

100 structure has an "occluding loop" [Ile105 to Pro126] which has an alternate conformation in

101 propeptide and in mature enzyme. The occluding loop is lifted above in procathepsin B*, while it*

102 is tightly packed in the active enzyme, thus exposing the active sites in *Cathepsin B*.

103 Procathepsin B is activated by other proteases like *cathepsin D* and, also by autoactivation. The

104 potential intermolecular cleavage site identified in *cathepsin B* is CystP42-GlyP43.–At low pH,

105 acidic residues at propeptide surface destabilize—propeptide secondary structure, resulting in

106 distortion of hydrophilic and hydrophobic interaction with mature region of protein.

107 Subsequently, intermolecular cleavage takes place, and propeptide gets completely dissociated

108 from mature enzyme. Thus, autoactivation is a bimolecular process. Once CTSB gets activated,

109 it activates trypsinogen (22). Mutations affect different regions of *Cathepsin B* protein but how

110 these variants affect the function of *Cathepsin B* is yet to be studied.

111 Since CTSB plays a cardinal role in premature trypsinogen activation, therefore in the present

112 study, we decided to analyze computationally the functional and structural effect of the missense

113 variants identified in the previous study (13), in order to determine the clinical significance of

114   these mutations in TCP pathogenesis. We predicted the effect of these coding variants in the

115   propeptide region of cathepsin B, using various in silico algorithms. Also, we predicted that

116   variants present in 3'UTR region (noncoding) in cathepsin B are associated with miRNA binding

117   sites, and hence they could be significant. Evidential results from the structural and functional

118   analysis of SNVs in *Cathepsin B* have implicated the potential role of these variants in the

119   pathogenesis of tropical calcific pancreatitis. This study is the first attempt to structurally and

120   functionally characterize the variants found in human *Cathepsin B* protein screened in TCP

121   patients.

Materials and Methods

123   **Data curation**

124   The single nucleotide polymorphisms (SNPs) in CTSB gene associated with TCP was extracted

125   from an article published in 2006 (13). In this article, researchers have done direct exome

126   sequencing of *CTSB* gene, taking samples from 25 controls and 51 TCP patients, and further

127   replicating the sequencing in 130 controls and 89 TCP patients from the same cohort, in order to

128   ensure their results. In the current study, we have mapped the SNPs extracted from the literature

129   on current human genome assembly, GRCh38.7. The mRNA accession number, NM_147782.2,

130   and protein accession number, NP_680092.1 of gene *Cathepsin B* (CTSB), was used in our

131   computational analysis. The current data about these SNVs were retrieved from human SNP

132   database, dbSNPbuild150. The workflow for the computational analysis performed to decipher

133   the significance of SNPs is depicted in **Figure 1.**

134   **[Insert Figure 1]**

135   **Sequence retrieval and Alignment**

136   The sequence of *Cathepsin B* (CTSB) was retrieved from UniProt database- P07858

137   (CATB_HUMAN). The non-synonymous variants (L26V and S53G) were manually inserted in

138   the wild-type protein sequence for further analysis.

139

140

141   **Non-synonymous SNP Analysis**

142   The functional effect of mutations was predicted using the following algorithms: SIFT (Sorting

143   Intolerant from Tolerant) (23), PolyPhen-2 (Polymorphism Phenotyping v2) (24) and

144   PANTHER (25). SIFT predicts whether the non-synonymous coding mutation affects protein

145   function or not, based on sequence homology and physical properties of amino acids. SIFT

146   calculates median conservation value for each amino acid position and thus measures the

147   diversity of sequence. Finally, it gives the score which is the normalized probability of an amino

148   acid change. Score of less than 0.05 are deleterious substitutions. PolyPhen-2 predicts the impact

149   of substitution on structure and function of a protein, after annotating the substitutions and

150   finally building conservation profiles. The prediction algorithm of PolyPhen-2 calculates Naïve

151   Bayes posterior probability about the damaging effect of mutation and gives prediction

152   sensitivity scores also. PolyPhen-2 annotates the substitution as "Possibly damaging", "Probably

153   damaging" or "Benign", based on their scores. PANTHER predicts the functional effect of

154   coding mutation based on "evolutionary preservation" metric of a given substitution and

155   calculates preservation time- position-specific evolutionary preservation (PESP). Longer the

156   PESP time, the more likely that substitution will have a deleterious effect. All these softwares:

157   SIFT, PolyPhen-2, and PANTHER, were based on evolutionary conservation-based algorithms.

158   Another tool, ProtParam (26) was used to calculate the hydropathicity or the GRAVY (grand

159   average of hydropathicity) score (27) of mutated procathepsin B sequences. Hydrogen bond

160   length and the rotational angles of main chain hydrogen bonds is a significant descriptor to study

161   the conformation and dynamics of a protein. Therefore, to calculate the altered hydrogen

162   bonding patterns in the mutated three-dimensional procathepsin B structures, Vadar v1.8

163   program (28) was used. This program calculates the H-bond distances in the main chain, side

164   chain and the bond angle between main chain residues.

165   Further, the stability of the mutated models was calculated by SDM (Site Directed Mutator)

166   server (29), I-Mutant suite (30) and mCSM (Mutation Cutoff Scanning Matrix Calculation) (31)

167   webservers. These algorithms calculate the difference in change in Gibbs free energy ($\Delta\Delta G$).

168   SDM server is used to calculate the difference in thermal stability of wild-type protein structure

169   and mutated protein structure, using constrained environment specific substitution tables

170   (ESSTs). I-Mutant suite is a support vector machine-based algorithm which predicts the protein

171   stability upon mutation, by taking datasets from ProTherm (32) database. This server also

172   calculates the change in Gibbs free energy ($\Delta\Delta G$) between wild-type and mutant protein

173   structures. mCSM server is used to predict the effect of mutations in proteins using graph-based

174   signatures. It predicts the effect of single point mutation on protein stability by extracting data

175   sets about various thermodynamic parameters from the ProTherm database and calculates change

176   in Gibbs free energy ($\Delta\Delta G$) between wild-type and mutant protein structures. Altogether, these

177   three algorithms calculate $\Delta\Delta G$ for a protein on mutation.

178   The change in Gibbs free energy ($\Delta\Delta G$) is as follows:

179    $\Delta G = \Delta H - T\Delta S$

180    $\Delta\Delta G = \Delta G_W - \Delta G_M$

181   ΔG = Change in Gibbs free energy of a system from unfavourable to favourable condition

182   ΔH = Change in enthalpy of the system

183   ΔS = Change in entropy of the system, T = Temperature of the system

184    $\Delta\Delta G$ = Value of free energy stability change of a protein upon mutation

185   ΔG_W = Change in Gibbs free energy of the wild-type protein from unfavourable to favourable

186   conditions

187   ΔG_m = Change in Gibbs free energy of mutant from unfavourable to favourable conditions.

188    $\Delta\Delta G > 0$ = Increase protein stability upon mutation

189    $\Delta\Delta G < 0$ = Decrease protein stability upon mutation.

190   Additionally, we have also done multiple sequence alignment (MSA) of procathepsin B protein

191   using sequences from 8 model organisms along with *Homo sapiens* (NP_680092.1): *Mus musculus*

192   (NP_031824.1), *Sus scrofa* (NP_001090927.1), *Macaca mulatta* (NP_001181828.1), *Rattus*

193   *norvegicus* (NP_072119.2), *Ovis aries* (NP_001295516.1), *Danio rerio* (NP_998501.1), *Bos*

194   *taurus (*NP_776456.1). The MSA was performed using Clustal Omega program (33)

195

Homology modeling

197 The two non-synonymous SNPs (L26V and S53G) were modelled to analyze the structural effect

198 of variants on protein. The position specific iterated blast program (PSI-BLAST) with protein

199 databank database (PDB) and default advanced settings, was used to find the template for

200 homology modelling. MODELLER 9.15 (34) was used to build mutated models of procathepsin

201 B. The best predicted models according to the lowest value of DOPE score (Discrete Optimized

202 Protein Energy), was used for further evaluation and analysis. The predicted 3D mutated models,

203 L26V and S53G, were evaluated for their quality by using Verify-3D (35) and ProSA (Protein

204 Structure Analysis) servers (36).   Verify-3D examines the correctness of 3D-structures by

205 comparing the 3D structure to the 1D structure. If 3D-1D score of each amino acid is $\geq$ 1, then

206 the model is correct. ProSA analysis the correctness of theoretical models by calculating the Z-

207 score of the input structure. ProSA considers C-alpha atoms of protein structure and calculates Z-

208 scores based on the similarity of crystal and NMR structures of the same size. If the Z-score for a

209 model is negative, then it is a model with minimum or no errors. After these models pass the

210 respective thresholds, the Ramachandran plot was evaluated (37). TM-score (38) and root mean

211 square deviation (RMSD) values of mutant structures (L26V and S53G), were calculated with

212 respect to wild-type by using TM-Align web server (39).

Molecular dynamics simulation

214 The molecular dynamic simulation was performed with Gromacs-5.0 package (40) on the native

215 (PDB ID: 3PBH) and mutant structures (S53G and L26V). This computational investigation was

216 done with a viewpoint to examine if these single nucleotide variants might lead to changes in

217 surface properties or distort the protein orientation. The protein molecule was solvated in a

218 dodecahedron box with SPC216 water molecules at 1.5 Å marginal radiuses. The system was

219 made neutral by adding 7 Na+ (Sodium ions) because the initial charge of the system is -7.

220 Subsequently, the molecular system was subjected to steepest distance energy minimization until

221 reaching the criterion of 1000kJ/mol (the minimization is converged when maximum force is

222 less than 1000 kJ/mol) with OPLS-all atom force field (40). Berendson temperature coupling

223 method (41) was used to regulate the temperature inside the box at 300k. Isotropic pressure

224 coupling was performed using Parinello-Rahman method (42), and the pressure of the system

225 was maintained at 1 bar. LINCS algorithm (43) was used to treat bond lengths including H-

226 bonds. Van der Waals and Coulomb interactions were truncated at 1nm, and Particle Mesh

227 Ewald method (44) was used to compute electrostatic interaction. Finally, the simulation was

228 performed for 35ns. The structural deviations between native and mutated structures were

229 subjected to comparative analysis by computing RMSD (Root mean square deviation) and

230 RMSF (Root mean square fluctuation). The trajectories were analyzed, and finally, protein

231 compactness was studied by calculating the radius of gyration (Rg). The secondary structure

232 analysis of wild-type and simulated mutant structures was also done using the do-dssp program

233 of Gromacs.

234

235 **Analysis of SNPs in UTR region**

236 5'UTR and 3'UTR regions in a gene play a crucial role in regulating gene expression at the post-

237 translational level. UTRs regulate the exit of mRNAs from the nucleus, translation efficiency,

238   sub-cellular localization and mRNA stability (45). The effect of SNPs in UTR regions was

239   analyzed using databases like1. miRdSNP (46), 2. PolymiRTS (47) and 3. miRNASNP (48).

240   **Results**

241   **Data curation**

242   The SNPs associated with TCP were extracted from literature and is tabulated in Table 1. The

243   SNPs were further categorized according to their type. There were total of 23 SNPs in CTSB

244   gene found to be associated with TCP. The non-coding region included 20 SNPs (1 deletion, 6 in

245   5'UTR, 2 in 3'UTR and 11 in introns). Coding region had 2 missense variants and 1 synonymous

246   variant (**Table 1).**

247   **[Insert Table 1]**

248

249   **Sequence retrieval and Alignment**

250   The protein sequence (NP_680092.1) was retrieved from the UniProt database, and the desired

251   variants were manually inserted in the sequence **(Figure 2)**. The UniProt ID for the procathepsin

252   B sequence is P07858 (CATB_HUMAN). Mutant 1 (L26V) where residue L (Leucine) is

253   substituted with residue V (Valine). Mutant 2 (S53G) where residue S (Serine) is substituted

254   with residue G (Glycine).

255   **[Insert Figure 2]**

256   **Homology modeling**

257   The 3D-structure of mutant (L26V and S53G) proteins was predicted after template searching by

258   PSI-BLAST. The protein sequence (NP_680092.1) of procathepsin B was used as a query, and

259   the resulting templates were then filtered. Finally, the X-ray crystal structure of human

260   procathepsin B (PDB ID: 3PBH) with 2.5 Å resolution (Sequence identity: 100% and Query

261   coverage: 93%) was used as a template for homology modeling. The DOPE scores, TM-scores,

262   and RMSD of the predicted best models by Modeller 9.15 are shown in **Table 2**. The

263   stereochemical properties of the mutated procathepsin B structures were evaluated using the

264   Ramachandran plot from RAMPAGE. The plot defines the amino acids in favoured, allowed and

265   outlier regions in the mutated structures as well as in the wild-type *Cathepsin B* structures (**Table**

266   **3).** Verify-3D did structure validation of predicted models, and it was observed that 99.05%

267   amino acids had average 3D-1D protein score in a 21 residue sliding window >=0.2 for L26V

268   mutated model and 97.16% amino acids had average 3D-1D protein score in a 21 residue sliding

269   window >=0.2 in S53G mutated model. Additionally, the ProSA web server was also used to

270   evaluate the quality of predicted 3D mutated models. The Z-score (by ProSA webserver) of

271   L26V model was -7.32 and of S53G model was -7.47, which were within the acceptable range of

272   X-ray and NMR studies. The interaction energy analyzed by ProSA tool was negative for

273   maximum residues in L26V, and S53G predicted models, in a sliding window of 10 and 40

274   respectively. Since the mutations were present in the propeptide region of procathepsin B, only

275   the mutated propeptide structures are shown in **Figure 3** and **Figure 4.**

276   **[Insert Table 2]**

277   **[Insert Table 3]**

278   **[Insert Figure 3]**

279 **[Insert Figure 4]**

280 **Non-Synonymous SNP Analysis**

281 The functional effect of the mutations predicted by using algorithms described in the "Methods"

282 section are tabulated in **Table 4**. Analysis by SIFT and PANTHER suggests that S53G and L26V

283 mutations can have damaging effects. The GRAVY scores of wild-type WT was -0.470, and the

284 mutants (L26V and S53G) was -0.469 for both. Thus, it could be concluded that there is no

285 significant effect of mutations on hydropathicity of the protein. The comparative analysis of

286 hydrogen bond lengths and the rotational angles between WT and mutants were calculated using

287 Vadar v1.5 server, at 10 different regions of protein, playing imperative role in the functioning of

288 procathepsin B. Remarkable differences in the H-bond lengths and Bond angles between WT

289 and mutants (S53G and L26V) were observed as represented in **Figure 5**. Hence, it could be

290 interpreted that these mutations alter the binding between residues in mutated structures.,

291 The change in Gibbs free energy ($\Delta\Delta G$) of the mutated structures calculated by SDM server, I-

292 Mutant 2.0, and mCSM webservers indicatedestabilization of mutated proteins (**Table 5**).

293 Additionally, the results of MSA revealed that leucine at 26th position and Serine at 53rd position

294 are conserved residues. Extrapolating these results indicate that any alteration in conserved

295 residues will affect the structure and function of the protein (**Figure 6**). Altogether, L26V and

296 S53G mutations were predicted to have a deleterious effect on the structure and function of the

297 protein, through non-synonymous SNP analysis algorithms.

298 **[Insert Table 4]**

299 **[Insert Table 5]**

300   **[Insert Figure 5]**

301   **[Insert Figure 6]**

302   **Analysis of SNPs in UTR region.**

303   UTRs play an essential role in mRNA processing during post-translational mechanism. Hence,

304   the SNPs in the UTR region can significantly affect the functionality of UTRs, provided they

305   affect the miRNA binding sites. The 3'UTR region is essential for microRNA (miRNA) binding

306   which can lead to degradation or transcriptional suppression of mRNA and thus, can further

307   affect the downstream processing. The databases miRdbSNP, PolymiRTS, miRNASNP, were

308   used to predict the significance of SNPs in the 3'UTR region (**Table 6**). The two SNPs present in

309   3'UTR region, rs709821 and rs8898, were predicted to be present in miRNA binding sites and

310   therefore are significant. The SNP, rs8898 was predicted to create a new miRNA site, while

311   rs709821 disrupt a non-conserved miRNA site as predicted by PolymiRTS database. The two

312   miRNAs targeting CTSB gene having both non-coding variants, hsa-miR-96 and hsa-miR-1271,

313   are pancreas specific miRNAs, deciphered from the miRNet database (49).

314   **Molecular dynamics simulations.**

315   The comparative analysis of trajectories by calculating RMSD, RMSF and Radius of gyration

316   after MD simulation of 35ns, for both native and mutants, was performed. Interestingly, it was

317   observed from RMSD of backbone residues, that both mutated structures (S53G and L26V) were

318   conspicuously deviated from the native structure (PDB ID: 3PBH). To infer effect of mutations

319   on the dynamic behaviour of each residue, RMSF of Cα-atoms was calculated. There was

320   fluctuation observed in mutated protein structures as compared to the native structure. The

321 protein compactness was determined by the radius of gyration (Rg). It was observed that Rg of

322 mutated structures were distinctly fluctuated as compared to native structure, throughout the

323 simulation **(Figure 7)**. The secondary structure analysis of wild-type and simulated mutant

324 structures, by do-dssp program, implicated that mutations had caused deviation to the protein

325 **(Figure 8)**.

326 **[Insert Figure 7]**

327 **[Insert Figure 8]**

328 **Discussion**

329 Tropical Calcific Pancreatitis (TCP) has distinct morphological characteristics with undefined

330 etiology. The propeptide region of procathepsin B (Arg1 -Lys 62) i.e. the N-terminal part inhibits

331 the activity of *Cathepsin B* in the pancreas, thereby regulating its premature activation, also act

332 as a scaffold for protein folding and as a chaperone for endosome/lysosomal trafficking. L26V

333 and S53G are the two missense variants observed in the propeptide region of *Cathepsin B* protein

334 in TCP patients. The *in-silico* SNP analysis of the mutated protein sequences, resulted in

335 alteration of secondary structure, thereby predicting an adverse-folding effect on the protein. The

336 phenotypic effect of the mutations was calculated using sequence analysis algorithms, and it was

337 observed that at least two algorithms indicated a deleterious effect of these mutations on protein

338 functionality. The free energy ($\Delta\Delta$G) calculations of mutated proteins structures, by various

339 algorithms, indicated that mutations are destabilizing. Further, the comparative analysis of H-

340 bond distances between mutated and native 3D-structure of procathepsin B provided a–unique

341 information about the structural characteristics of motifs around main chain H-bonds which are

342   altered in mutant protein structures, thereby affecting the function of the protein (50).

343   Additionally, MD simulation of mutated and native protein structures indicated that the

344   mutations distinctly deviate the structural conformation of procathepsin B, thereby having a

345   deleterious effect on–downstream signalling mechanism. Thus, the structural and functional

346   analysis of mutated procathepsin B predicts the significance of these mutations in the propeptide

347   region of *Cathepsin B*. Hence, we could extrapolate from these results *(in silico* analysis of the

348   mutated structures and sequences) that both mutations (L26V and S53G) have a deleterious

349   effect on structure and function of protein. These results will provide a lead towards designing

350   the experimental research strategy on the mutations involved in the pathogenesis of TCP to

351   understand the disease etiopathogenesis.

352   **Conclusion**

353   The dearth of information about the etiopathogenesis of tropical calcific pancreatitis was the

354   driving force for this study. The literature only has the information about the SNVs in *cathepsin*

355   *B* gene associated with TCP and lacks the crucial theories about the relative effects of these

356   SNVs in the pathogenesis of TCP. In this study, we predicted the structural and functional effect

357   of *cathepsin B* SNVs which were identified in TCP patients in previous studies. The predicted

358   deleterious effect of these SNVs is a lead towards developing biomarkers and therapeutics for

359   TCP. Further studies in this direction will help in defining the pathophysiology of TCP, which is

360   still a conundrum.

361

362   **Acknowledgements**

363 AS is thankful to Shiv Nadar University for providing all the necessary support to perform this

364 study.

365

366

367

368

369

370

371

References:

373 1.   Pandol SJ, Gorelick FS, Gerloff A, Lugea A. Alcohol abuse, endoplasmic reticulum
374 stress and pancreatitis. Dig Dis.2010;28:776–82.

375 2.   Witt H, Bhatia E.Genetic aspects of tropical calcific pancreatitis. Rev Endocr Metab
376 Disord.2008; 9(3):213–26.

377 3.   Levy P, Dominguez-Munoz E, Imrie C, Lohr M, Maisonneuve P. Epidemiology of
378 chronic pancreatitis: burden of the disease and consequences.United Eur Gastroenterol J.2014;
379 2:345–54.

380 4.   Kereszturi É, Szmola R, Kukor Z, Simon P, Ulrich Weiss F, Lerch MM, Sahin-Tóth M.
381 Hereditary pancreatitis caused by mutation-induced misfolding of human cationic trypsinogen: a
382 novel disease mechanism. Human mutation. 2009;30(4):575-82..

383 5.   Zhang Y-F, Deng H-L, Fu J, Zhang Y, Wei J-Q. Pancreatitis in hand-foot-And-mouth
384 disease caused by enterovirus 71. World J Gastroenterol.2016; 22(6):2149–52.

385 6.   Sarner M, Cotton PB. Classification of pancreatitis. Gut. 1984; 25:756–9.

386 7.   Barman KK, Premalatha G, Mohan V.Tropical chronic pancreatitis.Postgrad Med
387 J.2003;79:606–15

388 8.   Paliwal S, Bhaskar S, Chandak GR. Genetic and phenotypic heterogeneity in tropical
389 calcific pancreatitis. World J Gastroenterol. 2014; 20(46):17314–23.

390   9. Hassan Z, Mohan V, Ali L, Allotey R, Barakat K, Faruque MO, Deepa R, McDermott MF,
391   Jackson AE, Cassell P, Curtis D. SPINK1 is a susceptibility gene for fibrocalculous pancreatic
392   diabetes in subjects from the Indian subcontinent. The American Journal of Human Genetics.
393   2002 Oct 1;71(4):964-8.

394   10. Midha S, Khajuria R, Shastri S, Kabra M, Garg PK. Idiopathic chronic pancreatitis in India:
395   phenotypic characterisation and strong genetic susceptibility due to SPINK1 and CFTR gene
396   mutations. Gut. 2010 Jun 1;59(6):800-7.

397   11. Singh G, Bhat B, Jayadev MS, Madhusudhan C, Singh A. mutTCPdb: a comprehensive
398   database for genomic variants of a tropical country neglected disease—tropical calcific
399   pancreatitis. Database. 2018;2018.

400   12. Lerch MM, Halangk W. Human pancreatitis and the role of *Cathepsin B*. Gut. 2006;
401   55(9):1228-30.

402   13. Mahurkar S, Idris MM, Reddy DN, Bhaskar S, Rao GV, Thomas V, Singh L, Chandak GR.
403   Association of cathepsin B gene polymorphisms with tropical calcific pancreatitis. Gut.
404   2006;55(9):1270-5.

405   14. Mahurkar S, Bhaskar S, Reddy DN, Prakash S, Rao GV, Singh SP, Thomas V, Chandak GR.
406   TCF7L2 gene polymorphisms do not predict susceptibility to diabetes in tropical calcific
407   pancreatitis but may interact with SPINK1 and CTSB mutations in predicting diabetes. BMC
408   medical genetics. 2008;9(1):80.

409   15. Xiao Y, Yuan W, Yu B, Guo Y, Xu X, Wang X, Yu Y, Gong B, Xu C. Targeted gene next-
410   generation sequencing in Chinese children with chronic pancreatitis and acute recurrent
411   pancreatitis. The Journal of pediatrics. 2017;191:158-63.

412   16. Singh S, Choudhuri G, Agarwal S. Frequency of CFTR, SPINK1, and cathepsin B gene
413   mutation in North Indian population: connections between genetics and clinical data. The
414   Scientific World Journal. 2014;2014.

415   17.   Olson O.C, Joyce J.A. Cysteine cathepsin proteases: regulators of cancer progression and
416   therapeutic response. Nature Reviews Cancer. 2015; 15(12):712-29.

417   18.   Pungerčar JR, Caglič D, Sajid M, Dolinar M, Vasiljeva O, Požgan U, Turk D, Bogyo M,
418   Turk V, Turk B. Autocatalytic processing of procathepsin B is triggered by proenzyme activity.
419   The FEBS journal. 2009;276(3):660-8.

420   19.   Katunuma N. Posttranslational processing and modification of cathepsins and cystatins. J
421   Signal Transduct. 2010; 375345.

422   20.   Ghosh P, Dahms NM, Kornfeld S. Mannose 6-phosphate receptors: new twists in the tale.
423   Nature reviews Molecular cell biology. 2003; 4(3):202-13.

424 21.   Podobnik M, Kuhelj R, Turk V, Turk D. Crystal structure of the wild-type human
425 procathepsin B at 2.5 Å resolution reveals the native active site of a papain-like cysteine protease
426 zymogen. Journal of molecular biology. 1997; 271(5):774-88.

427 22. Halangk W, Lerch MM, Brandt-Nedelev B, Roth W, Ruthenbuerger M, Reinheckel T,
428 Domschke W, Lippert H, Peters C, Deussing J. Role of cathepsin B in intracellular trypsinogen
429 activation and the onset of acute pancreatitis. The Journal of clinical investigation. 2000
430 ;106(6):773-81.

431 23. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic
432 acids research. 2003; 31(13):3812-4.

433 24. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS,
434 Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods.
435 2010;7(4):248.

436 25. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with
437 the PANTHER classification system. Nature protocols. 2013; 8(8):1551-66.

438 26. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-
439 classical and leaderless protein secretion. Protein Engineering Design and Selection.
440 2004;17(4):349-56.

441 27. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein.
442 Journal of molecular biology. 1982;157(1):105-32.

443 28. Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS. VADAR: a
444 web server for quantitative evaluation of protein structure quality. Nucleic acids research.
445 2003;31(13):3316-9.

446 29. Worth CL, Preissner R, Blundell TL. SDM—a server for predicting effects of mutations on
447 protein stability and malfunction. Nucleic acids research.2011.

448 30. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation
449 from the protein sequence or structure. Nucleic acids research. 2005; 33 (Web Server issue):
450 W306-W310.

451 31. Douglas E. V. Pires, David B. Ascher and Tom L.Blundell. mCSM: predicting the effects of
452 mutations in proteins using graph-based signatures.Bioinformatics. 2014; 30(3): 335–342.

453 32. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0:
454 thermodynamic database for proteins and mutants. Nucleic acids research. 2004 Jan
455 1;32(suppl_1):D120-1.

456 33. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H,
457 Remmert M, Söding J, Thompson JD. Fast, scalable generation of high-quality protein multiple
458 sequence alignments using Clustal Omega. Molecular systems biology. 2011;7(1):539.

459 34. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints.
460 Journal of molecular biology. 1993;234(3):779-815.

461 35. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: assessment of protein models with three-
462 dimensional profiles. Methods in enzymology. 1997; 277:396.

463 36. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in
464 three-dimensional structures of proteins. Nucleic acids research 35(suppl 2). 2007; W407-W10.

465 37. S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G.Prisant, J.S.
466 Richardson and D.C. Richardson. Structure validation by Calpha geometry: phi,psi and Cbeta
467 deviation. Proteins: Structure, Function & Genetics. 2002; 50: 437-450.

468 38. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-
469 score. Nucleic acids research. 2005; 33(7):2302-9.

470 39. H.J.C.Berendsen, D.van der Spoel,R.van Drunen. GROMACS: A message-passing parallel
471 molecular dynamics implementation.Computer Physics Communications. 1995; 91:43-56.

472 40. Kutzner C, Páll S, Fechner M, Esztermann A, de Groot BL, Grubmüller H. Best bang for
473 your buck: GPU nodes for GROMACS biomolecular simulations. Journal of computational
474 chemistry. 2015;36(26):1990-2008.

475 41. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics
476 with coupling to an external bath. J Chem Phys. 1984;81:3684–90.

477 42. Martoňák R, Laio A, Parrinello M. Predicting crystal structures: the Parrinello-Rahman
478 method revisited. Physical review letters. 2003;90(7):075503.

479 43. Hess B, Bekker H, Berendsen HJ, Fraaije JG. LINCS: a linear constraint solver for molecular
480 simulations. Journal of computational chemistry. 1997;18(12):1463-72.

481 44. Darden T, York D, Pedersen L. Particle mesh Ewald: an N$\square$log(N) method for Ewald sums
482 in large systems. J Chem Phys 1993;98:10089–92.

483 45. Flavio Mignone, Carmela Gissi, Sabino Liuni and Graziano Pesole. Untranslated regions of
484 mRNAs.Genome Biology. 2002; 3(3).

485 46. Andrew E. Bruno, Li Li, James L. Kalabus, Yuzhuo Pan, Aiming Yu, Zihua Hu miRdSNP: a
486 database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes.
487 BMC Genomics.2012; 13(1):44.

488 47. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in
489 microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids
490 Res.2014; 42(D1):D86-D91.

491 48. Jing Gong, Yin Tong, Hong-Mei Zhang, An-Yuan Guo. miRNASNP: a database of miRNA
492 related SNPs and their effects on miRNA function. BMC Bioinformatics.2012; 13:A2.

493 49. Fan Y, Siklenka K, Arora SK, Ribeiro P, Kimmins S, Xia J. miRNet-dissecting miRNA-
494 target interactions and functional associations through network-based visual analysis. Nucleic
495 acids research. 2016 Apr 21;44(W1):W135-41.

496 50. Penner RC, Andersen ES, Jensen JL, Kantcheva AK, Bublitz M, Nissen P, Rasmussen AM,
497 Svane KL, Hammer B, Rezazadegan R, Nielsen NC. Hydrogen bond rotations as a uniform
498 structural tool for analyzing protein architecture. Nature communications. 2014; 17;5.

499

Figure 1

Workflow to identify the potential effect of SNPs

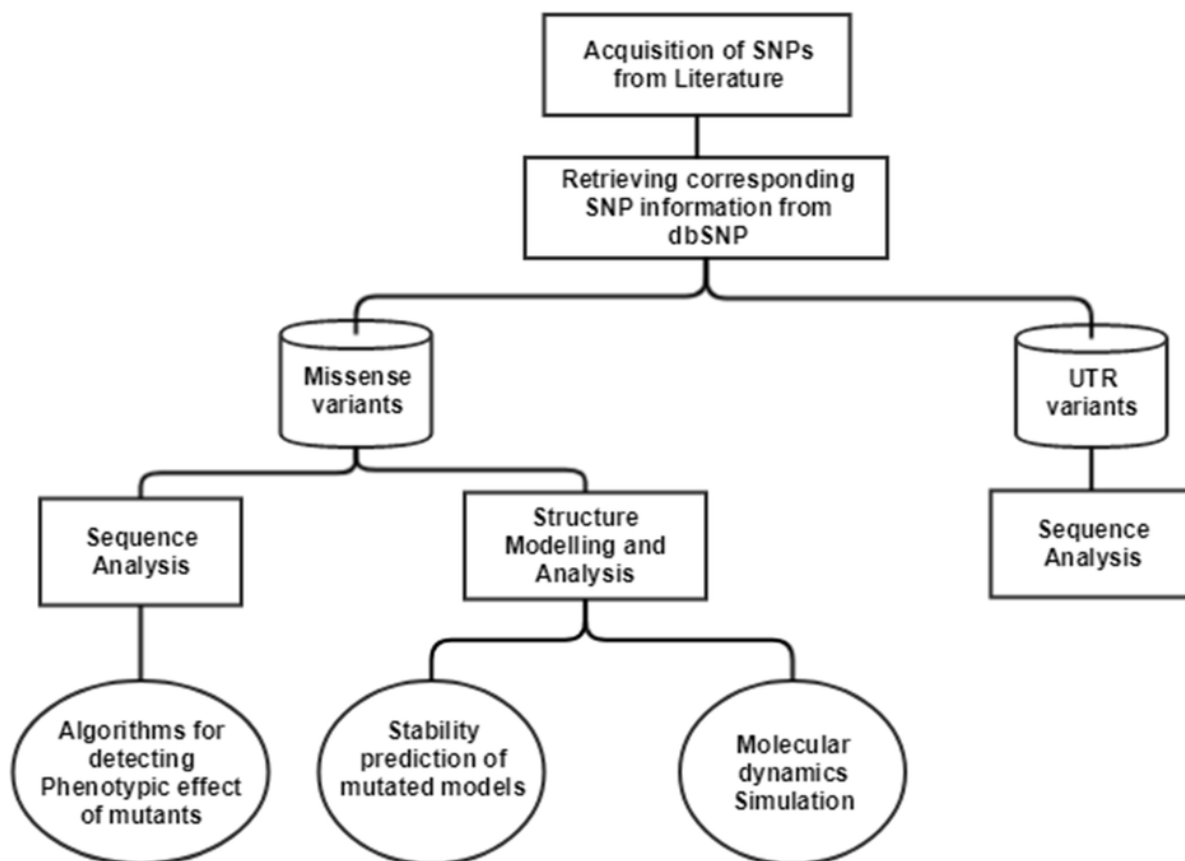



Figure 2

Fasta alignment of *procathepsin B protein* retrieved from Uniprot database

(A) Fasta sequence of wild-type *procathepsin B* (NP_680092.1, Isoform 1) retrieved from Uniprot database (ID : P07858). The wild type amino acids which were mutated in TCP patients, are highlighted in red

(A). >sp|P07858|CATB_HUMAN Cathepsin B OS=Homo sapiens GN=CTSB PE=1 SV=3 -
Native
MWQLWASLCCLLVLANARSRPSFHPLSDELVNYVNKRNTTWQAGHNFYNVDMSY
LKRLCGTFLGGPKPPQRVMFTEDLKLPASFDAREQWPQCPTIKEIRDQGSCGSCWAF
GAVEAISDRICIHTNAHVSVEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGL
VSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGEGDTPKCSKICEPGYSPTYKQDKHY
GYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVTGEMMGGHAIR
ILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQYW
EKI

Figure 3

Mutated (orange color) and wild-type (black color) propeptide models are shown

The mutation, L26V is shown in sticks, which is equivalent to LEU9P-VAL10 in PDB files. Visualization and numbering is done using PyMOL tool. Note: Numbering of amino acids in wild type PDB (3PBH) file differ to that of mutated models because propeptide and peptide regions are numbered separately in the published wild-type PDB file.

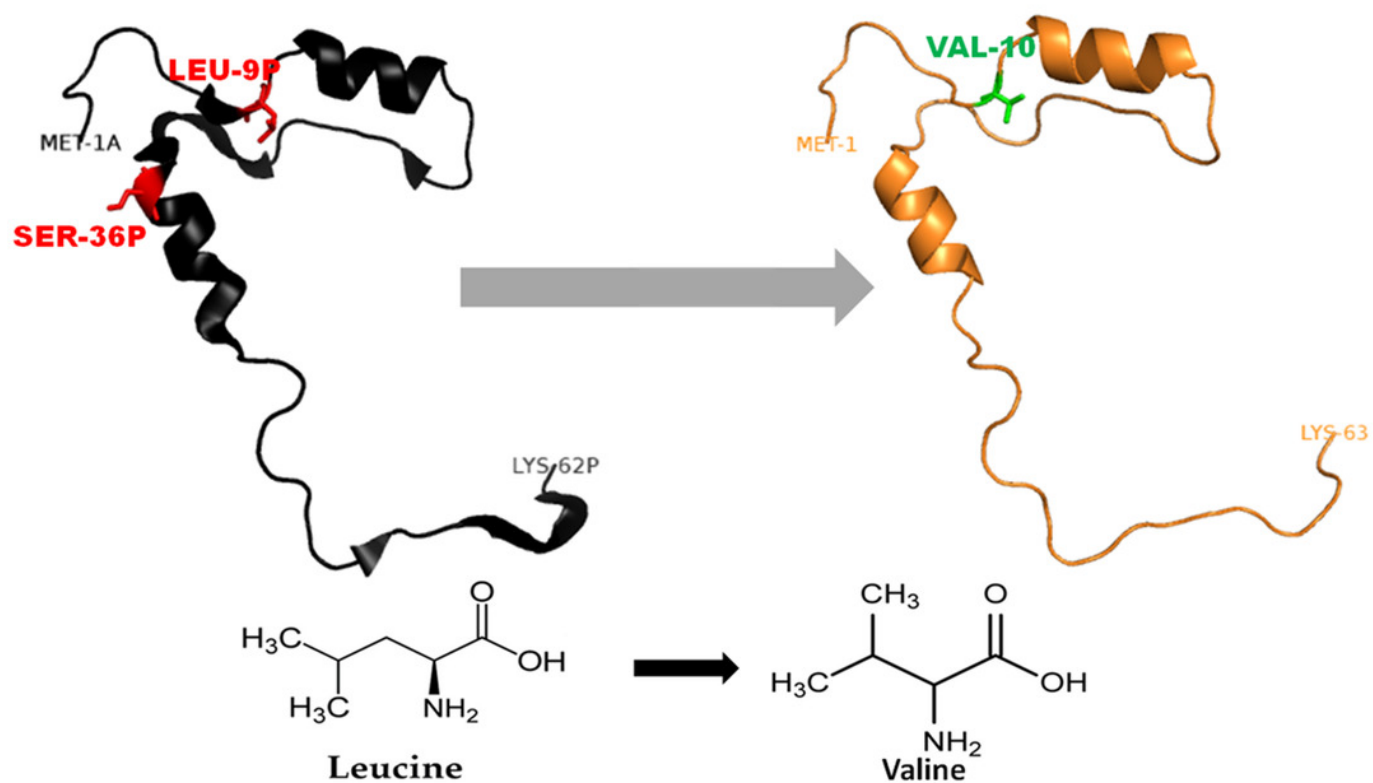



Figure 4

Mutated (Blue color) and wild-type (black color) propeptide models are shown.

The mutation, S53G is shown insticks as SER36-GLY37 in PDB files. Visualization and numbering is done using PyMOL tool. Note: Numbering of amino acids in wild type PDB (3PBH) file differ to that of mutated models because propeptide and peptide regions are numbered separately in the published wild-type PDB file.

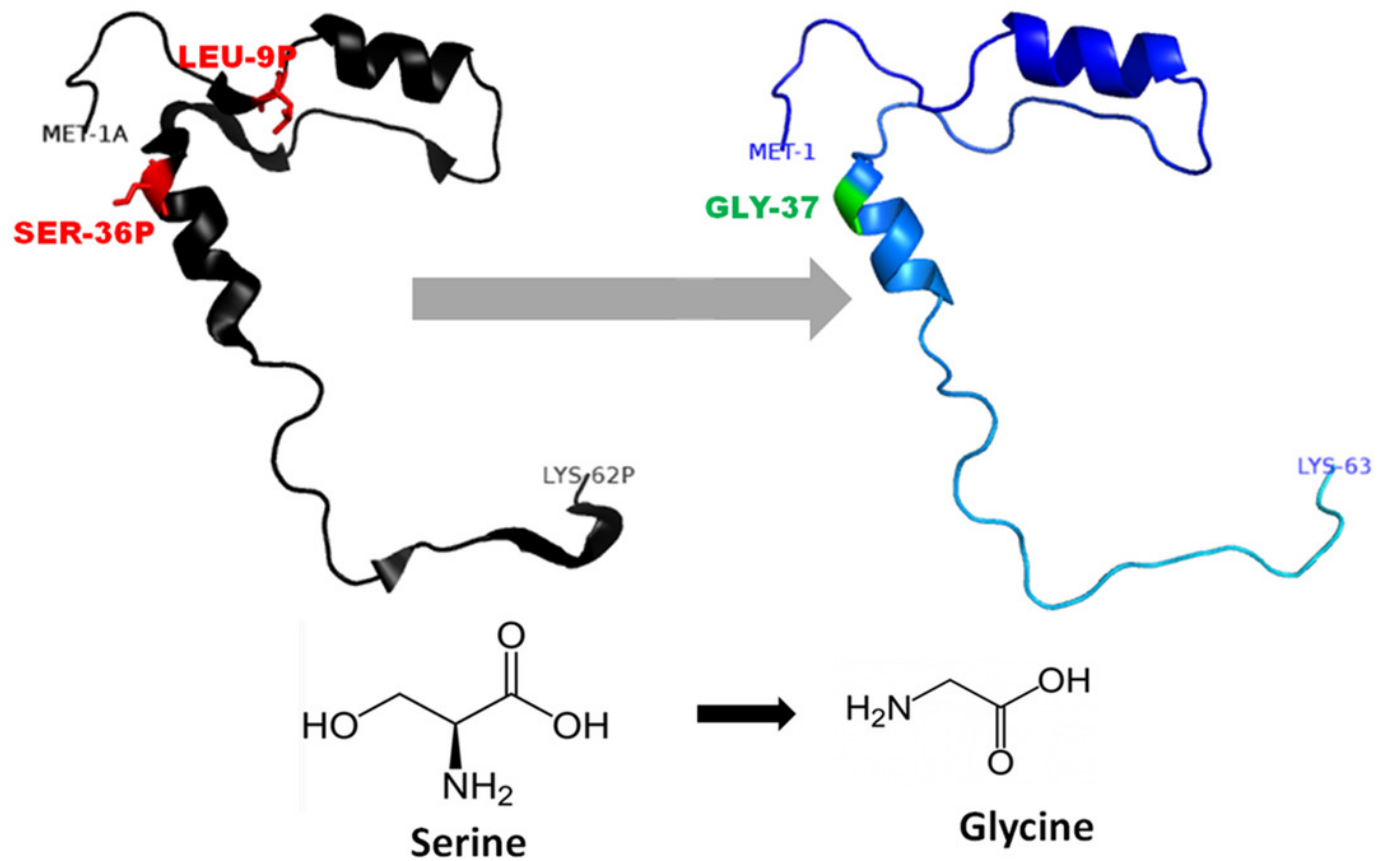



Figure 5

The comparative analysis of H-bond length between wild-type (WT) *procathepsin B* protein (PDB ID: 3PBH) and

A. Mutated structure (L26V) having a mutation at 26th amino acid (Leucine to Valine) in propeptide region **B**. Mutated structure (S53G) having a mutation at 53rd position (Serine to Glycine) in propeptide region. The colour key ranges from 1.5Å to 3.5Å with red as strong H-bonding and blue as weak H-bonding. Wide spaces indicate the absence of H-bonds at that position.

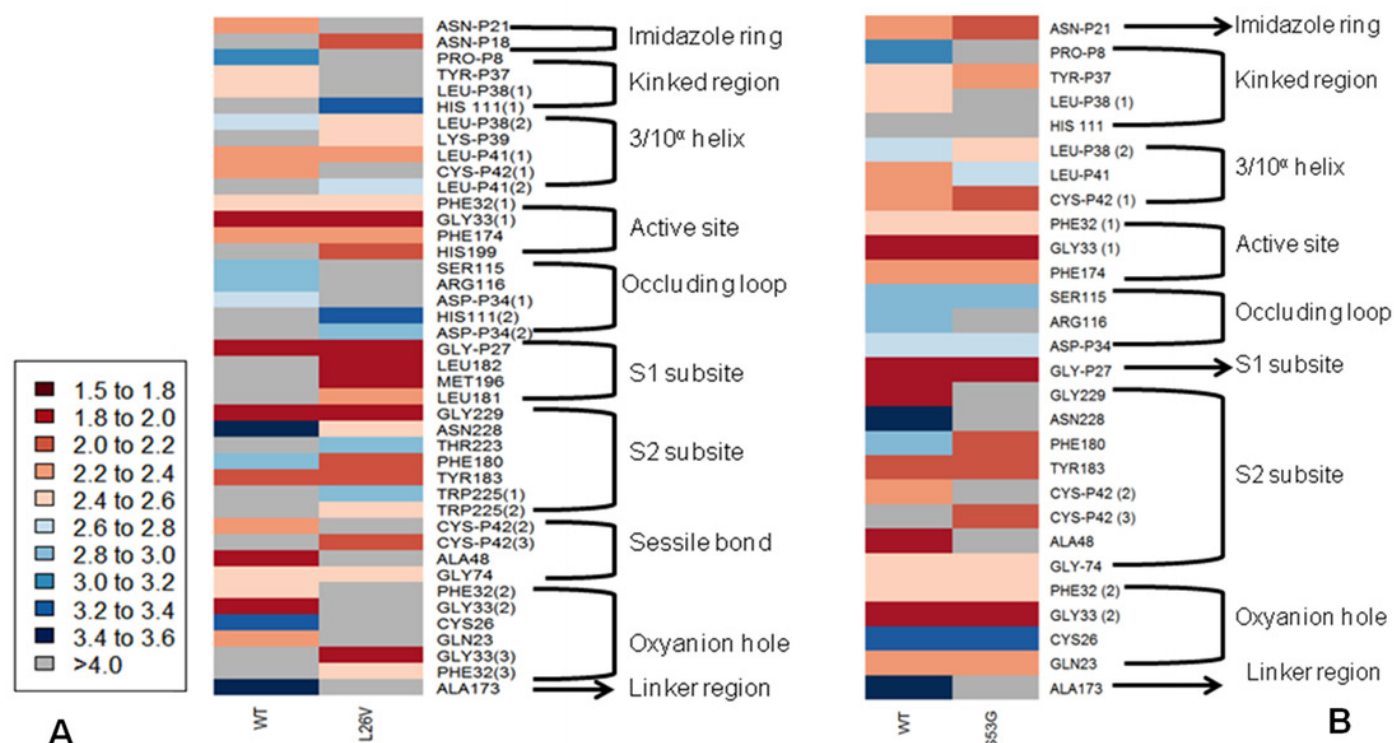



Figure 6

Multiple sequence allignment

(**A**). MSA of wild type procathepsin B protein (NP_680092.1). (**B.)** MSA of mutated protein sequence with mutation S53G. (**C**) MSA of mutated protein sequence with mutation L26V.

Figure 7

MD Simulation

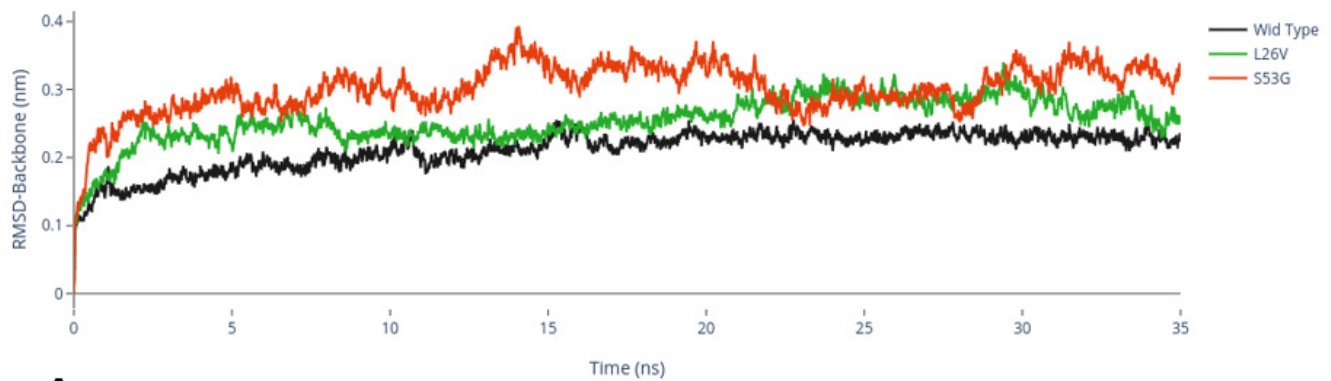
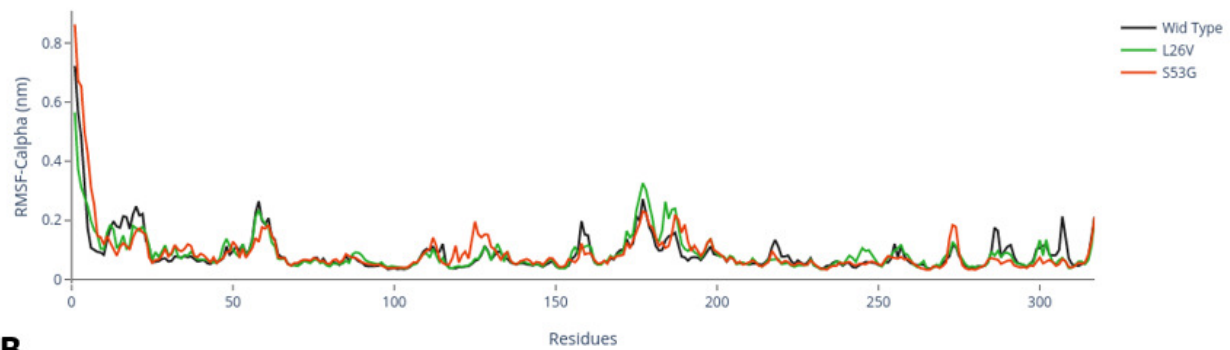
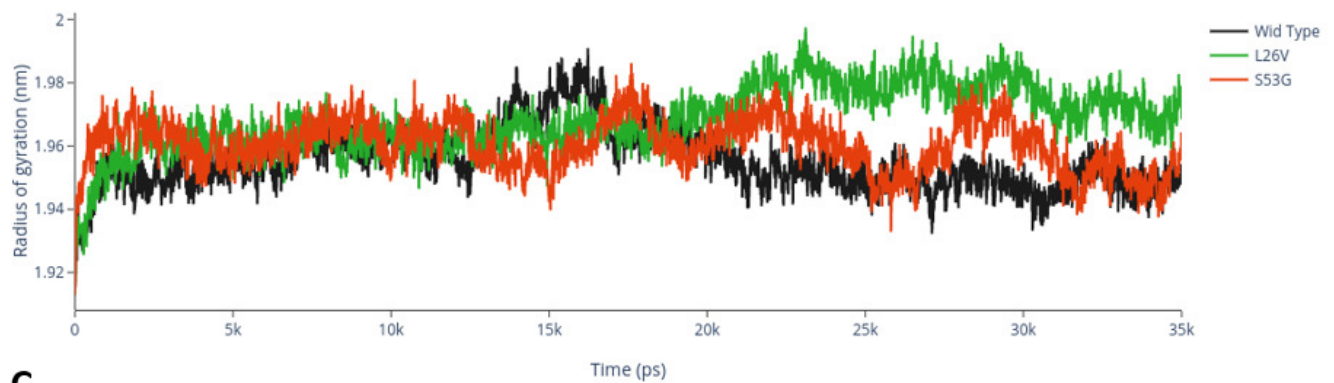
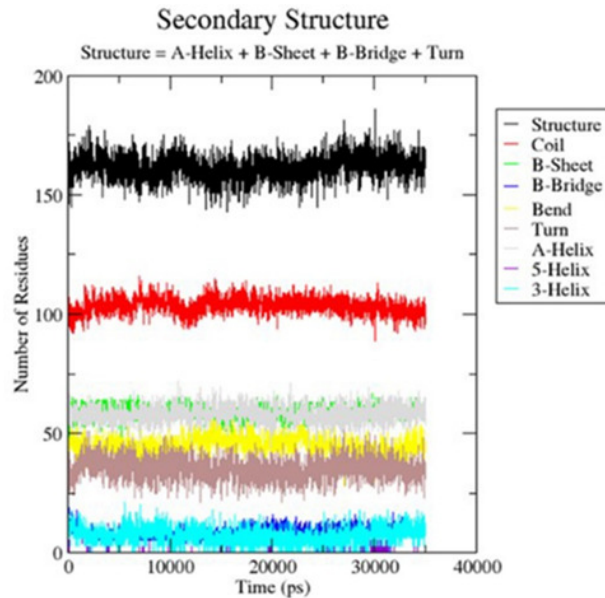
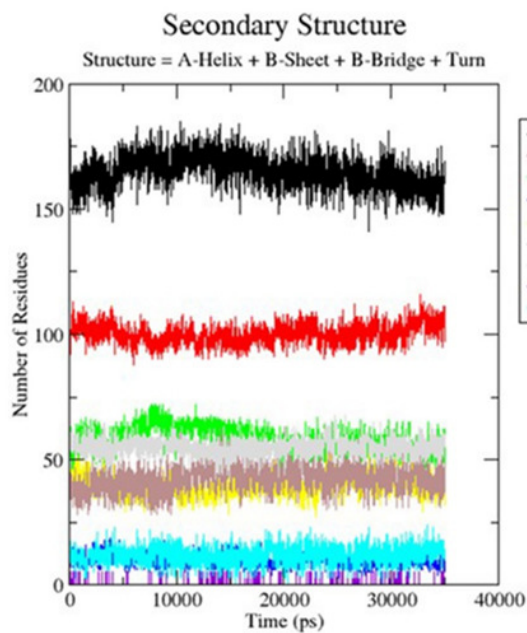
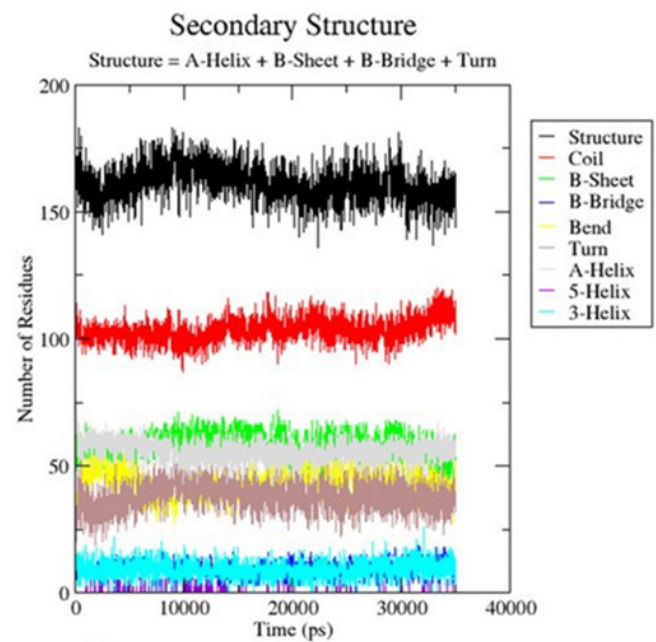
(A) Root mean square deviation (RMSD)-Backbone: Comparative analysis of RMSD-backbone between Wild-type and both mutated structures.**(B)**. **Root mean square Fluctuation (RMSF)**-C-alpha: Comparative analysis RMSF between mutated (s53g and l26v) and native protein structures. **(C)**. **Radius of gyration (Rg)**: Comparative analysis of Rg between wild-type protein (s53g, l26v) and native protein structures.

A

B

C

Figure 8

Secondary structure analysis:

The graphs depicts the secondary structure analysis using do-dssp program of **(A) Wild type protein structure** - 3PBH **(B). Mutated structure 1:** L26V and **(C) Mutated protein structure 2**: S53G

A   **3PBH-WildType protein**

B   **L26V**

C   **S53G**

The Single Nucleotide Polymorphisms in cathepsin B protein mined from literature (PMID: 16492714).

The SNP information is with respect to Ref Seq sequence ID: NT_077531.5 and dbSNP Build 150.

Ref rsID/ss ID	Position	Type	CDS position(relative to CDS start)	CDS Allele change	Protein Position	Residue change
-	Exon 1(5'UTR)	Non coding	14609	C>A	-	-
-	Intron1(5'UTR)	Non coding	14520	G>C	-	-
-	Intron1(5'UTR)	Non coding	14453	G>A	-	-
rs1293311	Intron1(5'UTR)	Non coding	14425	C>A	-	-
rs2645415	Intron1(5'UTR)	Non coding	11083	T>C	-	-
-	Exon 2(5'UTR)	Non coding	10927	C>G	-	-
rs4292649(rs12338)	Exon 3	Non-synonymous coding(Missense)	76	C>G	26	L>V (Leu> Val)
rs1293293(rs1122182)	Intron 3	Non-coding	335	A>T	-	-
-	Intron 3	Non-coding	394	G>A	-	-
rs1293292	Intron 3	Non-coding	595	C>T	-	-
rs1293291	Intron 3	Non-coding	663	T>C	-	-
rs1803250	Exon 4	Non-synonymous coding(Missense)	790*	A>G	53	S>G (Ser>Gly)
rs2272766	Intron 5	Non-coding	2609	C>T	-	
rs13332	Exon 6	Synonymous coding	4383*	A>C	140	T>T (Thr>Thr)
-	Intron 6	Non-coding	4451	G>C	-	-
rs1736090	Intron 6	Non-coding	4735	A>G	-	-
rs1692819	Intron 7	Non-coding	5516	C>T	-	-
rs2294139	Intron 7	Non-coding	5522	C>A	-	-
rs3215434	Intron 7	Non-coding(Deletion)	5581-5582		-	-
-	Intron 7	Non-coding	5622	C>G	-	-
rs2294138	Intron 8	Non-coding	5825	G>A	-	-
rs709821	Exon 11(3'UTR)	Non-coding	8370*	C>G	-	-
rs8898	Exon 11(3'UTR)	Non-coding	8422*	A>G	-	-

Quality assessment scores after modelling protein structures:

DOPE scores after homology modelling by Modeller 9.15 of mutants (L26V and S53G) and the structure alignment scores (TM-score and RMSD) of the CTSB mutant models with wild-type, 3PBH structure.

Predicted mutant model	DOPE Score	TM-score	RMSD
	Modeller 9.15	TM-Align	
L26V	-34105.02344	0.99941	0.16
S53G	-34285-52344	0.99973	0.17

The Ramachandran plot analysis of mutated models:

The table enlists the analysis from Ramachandran plot for each of the mutated protein structures (L26V and S53G)

Models	Favoured region	Allowed region	Outlier region
L26V	92.7%	5.1%	2.2%
S53G	91.7%	6.0%	2.0%

Prediction of functional effect of mutations by using different algorithms:

The table enlists the scores from SIFT, Polyphen-2 and PANTHER for each mutated protein sequence: L26V and S53G.

rsID	Allele change	AA change	SIFT		Polyphen 2		PANTHER	
			Score	Prediction	Score	Prediction	Score	Prediction
rs4292649(rs12338)	C>G	L26V	0	Affect protein function	0.01	Benign	1629	Probably damaging
rs1803250	A>G	S53G	0.02	Affect protein function	0.06	Benign	750	Probably damaging

Prediction of protein (*procathepsin B*) stability upon mutation:

The table enlists the change in Gibbs free energy (ΔΔG) in kcal/mol. ΔΔG>0 indicates stabilization while ΔΔG <0 indicates destabilization.

Algorithm	S53G	L26V	Effect
SDM	-.174	-0.94	Destabilizing
I-Mutant 2.0	-1.48	-1.93	Destabilizing
mCSM	-1.063	-1.638	Destabilizing

The SNPs in 3'UTR region of CTSB protein:

The table enlists the predicted miRNAs targeting the CTSB gene sequence having 3'UTR SNVs.

rsID	Region	Allele change	miRdSNP	PolymiRTS	miRNASNP
rs709821	UTR-3	C>G	hsa-miR-186	-	-
			hsa-miR-339-5p	-	-
			hsa-miR-7	-	-
			hsa-miR-214	-	-
			hsa-miR-431	-	-
			hsa-miR-186	-	-
			hsa-miR-320a	-	-
			hsa-miR-320d	-	-
			hsa-miR-320c	-	-
			hsa-miR-320b	-	-
			hsa-miR-96	-	-
			hsa-miR-1271	-	-
rs8898	UTR-3	A>G	hsa-miR-339-5p	hsa-miR-10a-5p	hsa-miR-10a-5p
			hsa-miR-186	hsa-miR-10b-5p	hsa-miR-10b-5p
			hsa-miR-7	hsa-miR-339-5p	hsa-miR-339-5p
			hsa-miR-214	hsa-miR-4421	
			hsa-miR-431	hsa-miR-5699-3p	
			hsa-miR-186	hsa-miR-6747-3p	
			hsa-miR-320a	hsa-miR-6752-3p	
			hsa-miR-320d		
			hsa-miR-320c		
			hsa-miR-320b		

			hsa-miR-96		
			hsa-miR-1271		