In-silico analysis of mutations in ANK1, SPTB, SPTA1, SLC4A1, and EPB4.2 genes responsible for Hereditary Spherocytosis

Bharti Vyas

Jamia Hamdard

Sunil Kumar

CSRM Division, Agriculture Statistics, Indian Institute of Farming Systems Research, Modipuram, Meerut (U.P), India

Garima Singh

Jamia Hamdard

Mymoona Akhter (✉ makhtar@jamiahamdard.ac.in)

Jamia Hamdard

Farhan Jalees Ahmad

Jamia Hamdard

Mohammed Mumtaz Alam

Jamia Hamdard

Mohammed Shaquiquzzaman

Jamia Hamdard

Research Article

Keywords:

Posted Date: May 5th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-464356/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

In-silico analysis of mutations in ANK1, SPTB, SPTA1, SLC4A1, and EPB4.2 genes responsible for Hereditary Spherocytosis

Bharti Vyas¹, Sunil Kumar³, Garima Singh¹, Mymoona Akhter^{2*}, Farhan Jalees Ahmad^{1,2}, Mohammed Mumtaz Alam², Mohammed Shaquiquzzaman²

¹School of Interdisciplinary Studies, Jamia Hamdard, New Delhi, India

²School of Pharmaceutical Education and Research, Jamia Hamdard, New Delh, India

³CSRM Division, Agriculture Statistics, Indian Institute of Farming Systems Research, Modipuram, Meerut (U.P), India

*Correspondence:

Mymoona Akhter

Associate Professor

School of Pharmaceutical Education and Research, Jamia Hamdard, New Delhi– 110062

E-mail: makhtar@jamiahamdard.ac.in.

ABSTRACT

Hereditary spherocytosis (HS) is a rare inherited disorder of red blood cells which are characterized by spherical, doughnut-shaped with increase deformability that lead to the gallstones and splenomegaly. The role of mutation in the genes responsible for the regulation of synthesis of proteins and stucture of RBC is well know studied. It was found that there are five genes whose mutation result in hereditary spherocytosis. Therefore, we aimed to study the consequences of ANK1, EPB4.2, SPTA1, SPTB, and SLC4A1 non-synonymous mutaion by using advanced insilico methods. Studied for nsSNPs using *insilico* techniques including OMIN, clinVar, SIFT, Polyphen, homology modelling. Misssence nsSNP were identified in all the gene selected and their effect on the protein structure, stability and functioning was studies. The result showed that 52 nsSNPs are responsible for the changes in the shape of RBCs. After identifying the nsSNPs the structure of proteins were modelled and their RMSD, relative solvent accessibility, and protein stability were studied. Protein stability analysis revealed significant change in free energy ($\Delta\Delta G$) of the most identified nsSNPs variants. These finding may be helpful for genotype- phenotype research as well as development in

pharmacogenetic studies. Finally, this study unveil a significance of *inslico* methods to figure out highly pathogenic genomic variants affected the structure and functional of HS causing protein

INTRODUCTION

Hereditary spherocytosis (HS) is inherited hemolytic disorder of red blood cells that is clinically and genetically heterogeneous, characterizes by the production of spherocytic red blood cells.

Spherocytosis leads to the reduced deformability of RBCs as a result increase in the rate of destruction of RBC and their removal from circulation by spleen resulting in a decrease in RBC count and ultimately anemic condition [1-4]. Hemolysis leads to increased level of bilirubin and finally to jaundice, gallstones, and splenomegaly [5-8]. Identification of Hereditary spherocytosis in neonates remains largely underacknowledged due to the common occurrence of physiological jaundice [9,10].

It is one of the most common inherited disorder which affects 1 in 2500 to 5000 in Caucasians population. The exact data about the disease in the Indian population needs to be ascertained yet [11].

There are five genes involved in RBC membrane disorder ANK1, SPTB, SPTA1, SLC4A1, and EPB4.2. Ankyrin encoded by the ANK1 gene is involved in the interaction between transmembrane protein to cell skeleton through band3, Erythrocyte membrane protein band 4.2 and spectrins. Band 3 encoded by the SLC4A1 gene is responsible for anionic exchange transporter. Alpha-spectrin encoded by SPTA1 gene play role in the shape and deformability of RBC. Beta-spectrin encoded by SPTB gene is responsible for the cell membrane organization and stability [12]. Erythrocyte membrane protein band 4.2 encoded by EPB4.2 gene plays a major role in regulation of the shape and mechanical properties. In 75% of the cases of HS, autosomal dominant is sufficient to cause the disease [5,13]. Erythrocyte membrane defect commonly have a morphological characteristics such as elliptocytes and spherocytosis, these features can be caused by various genetic and acquired causes.

Single Nucleotide Polymorphisms (SNPs) are the most common form of genetic variation in the human genome sequence. Two types of SNPs are known synonymous SNP (sSNP) and non-synonymous SNP (nsSNP), but the nsSNP also called as missense variants are particularly of important as they result into changes in the translated amino acid residue sequence [14,15]. These

nsSNPs are crucial, as they might be playing a major role in the functional diversity of coded proteins in humans that are responsible for various complex or hereditary diseases [16,17]. nsSNPs may be deleterious or tolerant in nature (18). Deleterious nsSNPs affects the function, interactions, and structure of proteins. Damaging effects may affect the stability of protein structure, alter the protein charge, geometry, translation, inter/intra protein structure, change in gene regulation, transcription and structure integrity of cells etc [19,20].

All these studies support the fact that SNP especially missense SNP is the simplest and most common source of genetic polymorphism in the human genome. Reports have been found that the missense SNPs is responsible for HS that leads to hereditary anemia. This is the first ever study which involve large area of the computational analysis of nsSNPs of Ankyrin, Spectrins (alpha and beta), Band 3 and Erythrocyte membrane protein band 4.2 protein. The ability to better explain the clinical complications of mutation depends on identifying the pathogenic mutations from the genetic correlation. Studying such mutations in vitro and in vivo requires technical expertise and enormous tools and servers of bioinformatics are designed to discover the outcomes of genetic changes on the biophysical characteristics, structure and functional of proteins. Thus, we aimed to analyze the pathogenic variants of the membrane genes under external conditions and to estimate the structure and functional effects of membrane proteins based on gene sequence with non-synonymous mutations by using various computational methods.

Result

Identification and analysis of nsSNPs

The study of literature revealed that five genes namely ANK1, EPB4.2, SLC4A1, SPTB, and SPTA1 are responsible for hereditary spherocytosis (HS). The study involved identification and analysis of nsSNPs responsible for HS. To identify and predict variation in genes responsible for HS, ClinVar and ExAC databases were used. The results obtained revealed 435 variations in all the five genes subjected for the study. Out of the variations identified 301 were related to missense variations, 20 to nonsense variations, 13 to frameshift variations, 93 to UTR (Untranslated Region) variations, and 8 to splice site variations. Further to understand the effect of variations (damaging or tolerate) on prognosis of disease, the result obtained from ClinVar and ExAC database were subjected to analysis

by SIFT tool. The inputs included chromosome number, position of variation and substitution of nucleotide. From SIFT, 77 variants were found to be damaged and were further selected for cross check by using WHES.db module PolyPhen-2 server. From the results 52 variants were found to be damaged in consensus with the SIFT results.

Genewise results are like six variants namely, R1218W, T472I, R446T, N251K, T188M, R669Q present in ANK1, eighteen (D2284N, E2224D, R2141W, R2016C, L1804F, V1698A, R1493W, L1485F, R1330I, R1133P, R891Q, I809V, L260P, L207P, L49F, R45S, R28H, R28C;) in SPTA1, ten (R2064P, A1884V, G1408R, R1359W, R1035W, D536N, R498H, D448N, V350M, W202R) in SPTB, two (R280Q, A112T) in EPB4.2 and sixteen (R870W, P868L, T837A, R646W, S613F, R589H, R589C, L153M, S773P, G771D, R760Q, G701D, R646Q, R602P, R589S, V488H) in SLC4A1 gene. All these variations are considered as most probably damaging variations. The damage variations are given in below table 1.

Prediction of change in the protein stability of the mutant protein, Relative Solvent Accessibility (RSA) and RMSD evaluation

Changes in the protein stability of missense variants were analyzed by using I-Mutant2.0 sever. All the nsSNPs 52 identified from the genes under study were submitted for analysis. The results obtained revealed that nine nsSNP's has increased stability viz. SLC4A1 residue namely V488M (rs28931584), SPTB gene residues V350M (rs141973081), A1884V (rs148337824), R1035W (143827332), W202R (rs121918646) and SPTA1 gene residues R2016C (rs78394850), L1485F (rs34973695), R1133P (rs35733059), L207P (rs121918643) while rest of the nsSNP's showed decrease in their stability. RMSD is vital parameters used to measures the rate of deviation of mutant protein structure from native protein structures and the deflection in the rate of protein structure is directly proportional to RMSD values. The increased RMSD value suggests that deviation in mutant and native protein structure configuration, which implies has extreme effect on protein structure, stability and function. Protein Relative Solvent Accessibility (RSA) is necessary to know the limit to examine the protein stability. The protein RSA was examined by NetsurfP server. Solvent accessible are studies determined the surface area of protein. NetsurfP sever gives the result in two categories, buried (B)

and exposed (E), indicating the low and high accessibility of biomolecules insolvent [22]. I-Mutant prediction result, RMSD calculation data and NetsurfP prediction results are given in table2.

Effect of change of amino acid on membrane protein phenotypic description:

Amino acid point mutations (nsSNPs) may change protein structure and function. This study compares residues that locally change protein three-dimensional structure. Such local conformational changes may impact protein function and may cause disease. Usually, this is more likely for structure changes connected to binding sites and folding. For instance, the disruption of hydrophobic interactions, hydrogen bonding, or the introduction of charged residues into buried sites, or mutations that break beta-sheets often impact phenotype severely and raise the susceptibility for disease [39-41]. Compared results are given in table 3.

Generation of the 3 D structure of membrane proteins

Target of ANK1, EPB42 and SLC4A1 is not experimentally determined completely cover sequence; computational based techniques were used for 3 D structure generate (Figure1). The FASTA sequence of the ANK1 protein (ID: P16157) with 1881 amino acid residues, SLC4A1 protein (ID: P02730) with 911 amino acid residues and EPB42 protein (ID: P16452) with 691 amino acid residues was identified from UniProt database. The template was maximum identity recognized by subjecting the objected sequence to BlastP. The template was validated by subjecting the received sequence to BlastP and represents the resulted maximum E-value. The ID: 4RLV was selected as template of ANK1 protein on the basis of query cover (45%) maximum identity (65%) with target. The ID 4PYG was selected as template of EPB42 protein on the basis of query cover (99%) maximum identity (33%) with target. The ID: 4YZF was selected as template of SLC4A1 protein on the basis of query cover (99 %) maximum identity (99 %) with target but other two proteins having below (35 %) of similarity.

The validation of the model

The 3 D structure of protein model developed was validated by PROCHECK and ProSA for model integrity [43]. Ramachandran plot (Figure 2) of the ANK1 protein represents 80.1% (584 aa) of the total residues in the most favoured region and 14.1% (103 aa) in additionally allowed regions, this indicating good quality of model; but in mutant ANK1 protein represents 76.4/% (557aa) of the total residues in the most favoured region and 18.2 % (133 aa) in additionally allowed regions, this indicating impact on protein model structure after mutation. In SLC4A1 protein analysis 86.5% (526 aa) of the total residues where the most favoured region and 10.9 % (66 aa) in additionally allowed regions, this indicating good quality model but in mutant SLC4A1 protein represents 86.4/% (586aa) of the total residues in the most favoured region and 10.7 % (85aa) in additionally allowed regions, this indicating impact on protein model structure after mutation. In case of EPB42 protein analysis 89.5% (544aa) of the total residues in the most favoured region and 8.1% (49 aa) in additionally allowed regions, indicating good quality model; but in mutant EPB42 protein represents 89.74/% (558aa) of the total residues in the most favoured region and 7.9 % (49 aa) in additionally allowed regions, this indicating impact on protein model structure after mutation. Moreover, ProSA Z- score (dark spot) is 0.33, -3.82, -7.83 which falls within the values range of the know protein determine by X-ray (light blue) and NMR (dark blue).

Superimposition of mutant protein structures

There were five HS related gene namely ANK1, SLC4A1, EPB42, and SPTB. Here required to an examination of mutational study on structure level for damaging mutation of gene those genes have played a vital role in proving the shape to RBC. The individual was compared mutational with native structure. In ANK1 gene the mutations namely (R1218W, T472I, R446T, T188M, N251K, R699Q), SLC4A1 gene mutation namely (R870W, P868L, T837A, S773P, G771D, R760Q, G701D, R646Q, R646W, S613F, R602P, R589H, R589S, R589C, V488M, L153M), EPB42 gene mutation namely (R280Q, A112T), SPTB gene mutation namely (R2064P, A1884V, G1408R, R1359W, R1035W, D536N, R498H, D448N, V350M, W202R) and SPTA1 gene mutation namely (D2284N, E2224D, R2141W, R2016C, L1804F, V1698A, R1493W, L1485F, R1330I, R1133P, R891Q, I809V, L260P, L207P, L49F, R45S, R28H, R28C) were confirmed damage effect and analysis the modelling of

structure by Modeller9.23 and visualization of change amino acid on structure by Pymol tool. In the given structure were compared with the native and mutant, native and mutant residue were highlight by cyan and magenta color using the Pymol tool.

Discussion

The RBC cell membrane disorder leads to a weakening of the cell membrane stability, increase fragility, membrane loss and leading to morphological changes due to irregular shape. This is most rare disorder. The genes responsible to maintain the cell shape of RBCs and their membrane disorders include ANK1 (Ankyrin 1), SLC4A1 (band 3), SPTA1 (α -spectrin), SPTB (β - spectrin), EPB4.2 (protein4.2). RBC cells function due to their unique shape. Studies reported that mutation in membrane proteins altered shape and function of cells, with this affect the protein-protein interaction. Till date various genetic mutations in membrane genes are found with the help of advance techniques. It becomes necessary to aware of the rare causes of inherited RBC disorder which is not easily detected by routine laboratory approaches. It is the best example of compound heterozygosity, in this case clinically insignificant defects of RBC cell genes present in parents but not responsible to caused significant hemolysis in offspring. However, the biological outcome of the alleles is not clearly understood. The relative study of genomic variants responsible for precise clinical data with their molecular approach is more expensive as well as time taking. Whereas, Insilco analysis is useful and advance approach that can help in investigate the selection of pathogenic variants in genetic variants studies and for the structural and functional protein phenotype prediction.

In the present investigation, advanced computational approaches has refined Missense mutations and predicted structure and functional impacts on RBC cell membrane protein. SIFT, clinVar tools were used to screen and identify the most pathogenic missense variants of RBC cell membrane protein. POLYPHEN 2.0 severs helped to fillet the missense mutations depending on their disease relevance property. The functional prediction revealed total 301 missense variants as damaging effect to RBC cell membrane proteins. In regard with commonly found disease relevance effect for the mutations, POLYPHEN 2.0 severs assorted 52 mutations as highly pathogenic. Protein stability effect due to

amino acid change studied by I-mutant and DUET server revealed that all amino acid variants showing DDG values ranging between -0.22 and -3.19Kcal/Mol demonstrated decreased stability of cell membrane protein due to these amino acid variants. These mutations also have induced changes in protein stability and surface accessibility of RBC cell mutant genes which can directly or indirectly affect the intermolecular and intramolecular interaction of amino acids.

Arginine residue is strong basic positively charged amino acid in nature and has ability to form multiple hydrogen bonds and salt bridge. The strong charge of the Arginine residue its location towards the outer hydrophilic surfaces of the proteins. As a result, guanidinium group of arginine is readily accessible for binding to other adjacent molecules with negative charge, e.g., DNA/ RNA complexes [44-46]. Due to this fact the arginines are one of the most common residues involved in the formation of complexes with other biomolecules. It can be seen as playing crucial role in RBC membranes stability and flexibility, as this types of interaction can be elucidate by opposite charge attraction, length and flexibility of side chain, and the potential to produce excellent hydrogen-bonding geometries with other biomolecules like nucleobases and phosphate groups. On the basis of the potential to form these kinds of interaction, it is explicit that, substitution of Arginine with any other amino acid due to mutation could have disturbing effect on cell wall stability and other functions of the cell. There are increasing experimental evidences that suggest that the dysregulation of arginine-modifying enzymes plays pivotal roles in occurrence of various diseases like cancer, inflammatory diseases, neurodegenerative diseases, and other conditions [47,48]. Taking an example of substitution by tryptophan residue, tryptophan is bigger in size, neutral in charge and hydrophobic in nature in comparison to basic and small structure of arginine. Due hydrophobic nature of tryptophan it has high propensity to get buried in protein hydrophobic region and also does not involve in sulphide linkage at C1216 position. This significant deviation of amino acid residues may disturb physical characteristics such hydrogen bonds, active sites of residue and electrostatics charge.

The difference of structural deviations in terms of RMSD between mutant and native type models demonstrated significant deviation in all twenty-three mutated models of RBC cell membrane. The significant deviations' of amino acid residues may disturbs biophysical characteristics such as

hydrogen bonds, active sites of residue and electric charges. This result used to characterize the affect of nsSNP on RBC cell membrane mutant genes and advice that in computational investigation may be useful tool to predict the effect of variants on gene function and expression.

Methods and Material

Disease Associate Gene

Information related to hereditary spherocytosis was collected from the databases Online Mendelian Inheritance In Man (OMIM), Genetic Association Database (GAD), Gene Card^{21,22} and other literature analysis.

Collection of Missense Variants

The dataset related to missence variants in human genome was collected from the clinVar database (<https://www.ncbi.nlm.nih.gov/clinvar/>). All collected SNVs were classified as non-coding and coding based on the nature and condition of the variants. Only Missense variants (nsSNPs) were chosen for computational analyzes because of their ability to disturb the structural deformation of proteins [23].

Identification of nsSNP and their functional effect

Identification of missence SNPs was done by using the Soring Intoleration From Toleration (SIFT) tool. Sorting Intoleration From Toleration (SIFT) is online bioinformatics software which is used to find out the deleterious coding nsSNPs form other SNPs [24]. SIFT (<http://sift.jcvi.org/>) predicts the nsSNPs in submitted chromosome number and position that affect amino acid substitution in effective proteins. It assigns a score for each amino acid residue on range of 0 to 1, if the score is equal or less than 0.05 the nsSNPs is taken as intoleran nsSNP. The results were cross checked by using another online softwere PolyPhen-2 [25]. PolyPhen-2 predicts the functional significance of SNPs. WHESS.db module of PolyPhen-2, is rapid, reliable tool and result are produced quickly after the submitted query as nsSNPs. The outcome can be: probably damaging, possibly damaging or benign, with score for amino acid residue on range of 0 to 1, if the score is 0 to 0.15 the nsSNPs is taken as

benign similarly a score between 0.15-1.0 reflects damaging nsSNPs. Polyphen server was accessed at <http://genetics.bwh.harvard.edu/pph2/index.shtml>[26].

Mutant protein stability prediction

Protein stability analysis of mutant protein was done by I-Mutant2.0 tool [27]. I-Mutant2.0 tool is support vector machine SVM based tool that predicts the change in protein stability upon single point mutation while taking the target protein sequence or structure (PDB formate and Chain) as an input. All nsSNP of hereditary spherocytosis disease were submitted to the I-Mutant2.0 tool to predict the free energy change values. Results obtained are in the form of stability of protein (increased or decreased stability) and Gibbs free energy in the form of DDG values [28]. The decrease protein stability is reflections of increased degradation, miss-folding, and aggregation of the protein and vice-versa. Sequence of proteins ankyrin, α and β spectrins, band2 and Erythrocyte membrane protein band 4.2 were retrieved form uniprot database and further subjected to protein stability prediction analysis using I-Mutant2.0 tool. Protein sequences AC P16157, AC P16452, AC P02730, AC P02549 of alpha-spectrin and AC P11277 of Ankyrin, Erythrocyte membrane protein band 4.2, band 3, α and β respectively spectrin was selected for analysis.

Effect of mutant protein on solvent accessibility

Solvent accessibility of mutant protein was detected by using online NetSurfP server. NetSurfP servers Predicts the solvent accessible surface of protein, by using the FASTA sequence of protein. This sever performs the solvent accessibility on the basis of accessibility of the amino acid residues by the solvent (water) and analysis was obtained in two classes as ether buried or exposed[29,30].

3D modelling of the protein and RMSD calculations

The generation of 3 D structure of Ankyrin, erythrocyte membrane protein was carried out using modeller9.22 [31]. The amino acid sequences of membrane proteins was retrieve from the UniProtKB database in FASTA format, and submitted to BLASTp to identify suitable protein template. The template with (>30%) identity was selected for generating the 3D structure of the target [32-34]. PyMOL was used for calculating the RMSD (Root Mean Square Deviation) values between native structure and mutant four of each protein[35-36].

3D structure validation

The quality of the generated model was evaluated by PROCHECK and Protein Structural analysis (ProSA). PROCHECK a server that relies on Ramachandran plot for structure verification; understands the stereo-chemical quality of the model. ProSA was used to refine and validate the experimental protein structure and modelling [37,38].

Conclusion

This study conducted with online database, computational tool and servers like ClinVar, ExAC, SIFT, PolyPhen, I-Mutant, NetPhos, NetsurfP, Pymol and Modell9.23 etc conclude that list of nsSNP's retrieved from SIFT and PolyPhen, sixteen nsSNP's with SNP ID's (ANK1 (rs146346710), EPB42 (rs121917734), rs104894487), SPTB (rs121918650, rs150698819, rs121918646), SPTA1 (rs138732899, rs41273519, rs35733059, rs36057043, rs121918634, rs121918642), SLC4A1 (rs121912741, rs121912755, rs121912758, rs121912754) were found to be highly damaging and deleterious to the red blood cells protein structure, stability and function and were predicted to be key cause for occurrence of Hereditary spherocytosis. These above predicted deleterious mutation might be result in the altered function of given five genes leading to hereditary spherocytosis. Mutant ANK1, SPTB, SPTA1, EPB42 and SLC4A1 gene are active (expresses) in red blood cells, but also found in brain and muscle cells, these genes are located at cell membrane and bind to other membrane protein which effect on stability and structure of red blood cells. Which was predicted to be the outcome associated with effect of the nsSNP's above on ankyrin, spectrins (alpha and beta), band2 and Erythrocyte membrane protein band 4.2 proteins. Those nsSNPs are causing potential effect on functional and structural of protein. This study is more helpful, important nsSNPs to select for wet lab evaluation and development in potent drug discovery.

Bibliography

1. Tasdemir, P. Linkage Analysis of Hereditary Spherocytosis in Four Generations of a Family with SPTB Gene Deficiency. *Int. J. Hematol. Oncol.* **22**, 85–94 (2012).

2. Llaudet-Planas, E. *et al.* Osmotic gradient ektacytometry: A valuable screening test for hereditary spherocytosis and other red blood cell membrane disorders. *Int. J. Lab. Hematol.* **40**, 94–102 (2018).
3. Bogusławska, D. M., Heger, E. & Sikorski, A. F. [Molecular mechanism of hereditary spherocytosis]. *Pol. Merkur. Lek. Organ Pol. Tow. Lek.* **20**, 112–116 (2006).
4. He, B.-J. *et al.* Molecular Genetic Mechanisms of Hereditary Spherocytosis: Current Perspectives. *Acta Haematol.* **139**, 60–66 (2018).
5. Perrotta, S., Gallagher, P. G. & Mohandas, N. Hereditary spherocytosis. *Lancet Lond. Engl.* **372**, 1411–1426 (2008).
6. Manciu, S., Matei, E. & Trandafir, B. Hereditary Spherocytosis - Diagnosis, Surgical Treatment and Outcomes. A Literature Review. *Chir. Buchar. Rom. 1990* **112**, 110–116 (2017).
7. An, X. & Mohandas, N. Disorders of red cell membrane. *Br. J. Haematol.* **141**, 367–375 (2008).
8. Barcellini, W. *et al.* Hereditary red cell membrane defects: diagnostic and clinical aspects. *Blood Transfus. Trasfus. Sangue* **9**, 274–277 (2011).
9. Christensen, R. D. & Henry, E. Hereditary Spherocytosis in Neonates With Hyperbilirubinemia. *PEDIATRICS* **125**, 120–125 (2010).
10. Agarwal, A. M. Ankyrin Mutations in Hereditary Spherocytosis. *Acta Haematol.* 63–64 (2019)
doi:10.1159/000495339.
11. Kalyan, M., Kanitkar, S., Gaikwad, A. & Kumar, H. Hereditary spherocytosis. *Med. J. Dr Patil Univ.* **7**, 229 (2014).
12. Meng, L.-L. *et al.* Next-generation sequencing identified a novel SPTB frameshift insertion causing hereditary spherocytosis in China. *Ann. Hematol.* **98**, 223–226 (2019).
13. Will, A., Henderson, C. A., Jnah, A. J. & Newberry, D. Hereditary Spherocytosis in the Neonatal Period: A Case Report. *Neonatal Netw. NN* **36**, 280–288 (2017).
14. Theodoratou, E. *et al.* Genome-wide scan of the effect of common nsSNPs on colorectal cancer survival outcome. *Br. J. Cancer* **119**, 988–993 (2018).
15. Amir, M. *et al.* Investigation of deleterious effects of nsSNPs in the POT1 gene: a structural genomics-based approach to understand the mechanism of cancer development. *J. Cell. Biochem.* (2018)
doi:10.1002/jcb.28312.

16. Chandrasekaran, G. *et al.* In silico analysis of the deleterious nsSNPs (missense) in the homeobox domain of human HOXB13 gene responsible for hereditary prostate cancer. *Chem. Biol. Drug Des.* **90**, 188–199 (2017).
17. Dabhi, B. & Mistry, K. N. In silico analysis of single nucleotide polymorphism (SNP) in human TNF- α gene. *Meta Gene* **2**, 586–595 (2014).
18. Banaganapalli, B. *et al.* A Computational Protein Phenotype Prediction Approach to Analyze the Deleterious Mutations of Human MED12 Gene: A COMPUTATIONAL PROTEIN PHENOTYPE PREDICTION APPROACH. *J. Cell. Biochem.* **117**, 2023–2035 (2016).
19. Yadav, P., Chatterjee, A. & Bhattacharjee, A. Identification of deleterious nsSNPs in α , μ , π and θ class of GST family and their influence on protein structure. *Genomics Data* **2**, 66–72 (2014).
20. Masoodi, T. A., Rao Talluri, V., Shaik, N. A., Al-Aama, J. Y. & Hasan, Q. Functional genomics based prioritization of potential nsSNPs in EPHX1, GSTT1, GSTM1 and GSTP1 genes for breast cancer susceptibility studies. *Genomics* **99**, 330–339 (2012).
21. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-798 (2015).
22. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.* **36**, 431–432 (2004).
23. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
24. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
25. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, (2013).
26. Hassan, M. M. *et al.* Bioinformatics Approach for Prediction of Functional Coding/Noncoding Simple Polymorphisms (SNPs/Indels) in Human BRAF Gene. *Adv. Bioinforma.* **2016**, 2632917 (2016).
27. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, W306–W310 (2005).
28. Pucci, F., Bernaerts, K., Kwasigroch, J. M. & Rومان, M. Quantification of biases in predictions of protein stability changes upon mutations. *bioRxiv* (2018) doi:10.1101/308239.

29. Klausen, M. S. *et al.* NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinforma.* (2019) doi:10.1002/prot.25674.
30. Petersen, B., Petersen, T., Andersen, P., Nielsen, M. & Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **9**, 51 (2009).
31. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).
32. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187-191 (2006).
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
34. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
35. Vadija, R. *et al.* Homology modeling and virtual screening studies of FGF-7 protein—a structure-based approach to design new molecules against tumor angiogenesis. *J. Chem. Biol.* **9**, 69–78 (2016).
36. DeLano, W. L. Pymol. An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography. **40**, 82–92 (2002).
37. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, W407-410 (2007).
38. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
39. Xu, J., Baase, W. A., Baldwin, E. & Matthews, B. W. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci. Publ. Protein Soc.* **7**, 158–177 (1998).
40. Gong, S. & Blundell, T. L. Structural and Functional Restraints on the Occurrence of Single Amino Acid Variations in Human Proteins. *PLoS ONE* **5**, e9186 (2010).
41. Wang, Z. & Moulton, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270 (2001).
42. Scheiner, S., Kar, T. & Pattanayak, J. Comparison of various types of hydrogen bonds involving aromatic amino acids. *J. Am. Chem. Soc.* **124**, 13257–13264 (2002).
43. Lovell, S. C. *et al.* Structure validation by $\text{C}\alpha$ geometry: ϕ, ψ and $\text{C}\beta$ deviation. *Proteins Struct. Funct. Bioinforma.* **50**, 437–450 (2003).

44. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860–2874 (2001).
45. Calnan, B. J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A. D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252**, 1167–1171 (1991).
46. Hunter, T. Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 2513–2516 (2012).
47. Copeland, R. A., Solomon, M. E. & Richon, V. M. Protein methyltransferases as a target class for drug discovery. *Nat. Rev. Drug Discov.* **8**, 724–732 (2009).
48. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
49. Copeland, R. A., Solomon, M. E. & Richon, V. M. Protein methyltransferases as a target class for drug discovery. *Nat. Rev. Drug Discov.* **8**, 724–732 (2009).

Declarations

Acknowledgement:

The authors are thankful to Department of Biotechnology, Ministry of science & Technology, Government of India for the Bioinformatics infrastructure facility at Jamia Hamdard under BTISNet program. One of the author Ms. Bharti Vyas acknowledges the fellowship obtained from ICMR, New Delhi

Conflict of interest: The other authors have no conflicts of interest to declare.

Tables

Table 1: The list of genes, damaged nsSNPs and affected amino acids their tolerance index, predicted impact and probability score by SIFT and POLYPHEN

GENE NAME	SIFT Prediction				POLYPHEN Prediction	
	nsSNP ID	Amino acid change	Tolerance Index	Predicted Impact	Probability Score	Predicted Impact
ANK1	rs142542271	R1218W	0	DAMAGING	0.987	PROBABLY DAMAGING
	rs150655828	T472I	0	DAMAGING	0.965	PROBABLY DAMAGING
	rs142010751	R446T	0	DAMAGING	0.994	PROBABLY DAMAGING
	rs61735313	N251K	0	DAMAGING	0.999	PROBABLY DAMAGING
	rs146346710	T188M	0	DAMAGING	1	PROBABLY DAMAGING
	rs150032875	R699Q	0	DAMAGING	0.953	PROBABLY DAMAGING
SPTB	rs121918650	R2064P	0	DAMAGING	1	PROBABLY DAMAGING
	rs148337824	A1884V	0	DAMAGING	0.999	PROBABLY DAMAGING
	rs17245552	G1408R	0.01	DAMAGING	0.997	PROBABLY DAMAGING
	rs189656371	R1359W	0.01	DAMAGING	0.773	PROBABLY DAMAGING
	rs143827332	R1035W	0	DAMAGING	0.998	PROBABLY DAMAGING
	rs145675502	D536N	0	DAMAGING	0.969	PROBABLY DAMAGING
	rs140796444	R498H	0	DAMAGING	0.999	PROBABLY DAMAGING
	rs150698819	D448N	0.01	DAMAGING	1	PROBABLY DAMAGING

	rs141973081 rs121918646	V350M W202R	0.01 0	DAMAGING DAMAGING	0.915 1	PROBABLY DAMAGING PROBABLY DAMAGING
SLC4A1	rs28931585 rs121912759 rs121912750 rs121912753 rs121912741 rs121912755 rs121912748 rs121912757 rs121912758 rs121912746 rs121912754 rs121912744 rs121912747 rs121912745 rs28931584 rs145041032	R870W P868L T837A S773P G771D R760Q G701D R646Q R646W S613F R602P R589H R589S R589C V488M L153M	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.02 0	DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING	0.998 0.979 0.989 0.998 1 1 0.999 0.994 1 0.992 1 0.991 0.991 0.999 0.915 0.706	PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING
EPB4.2	rs121917734 rs104894487	R280Q A112T	0 0	DAMAGING DAMAGING	1 1	PROBABLY DAMAGING PROBABLY DAMAGING
SPTA1	rs138732899 rs142775522 rs41273519 rs78394850 rs116959874 rs182430449 rs35237700 rs34973695 rs34214405 rs35733059 rs36057043 rs7547313 rs121918634 rs121918643 rs121918639 rs121918637 rs121918641 rs121918642	D2284N E2224D R2141W R2016C L1804F V1698A R1493W L1485F R1330I R1133P R891Q I809V L260P L207P L49F R45S R28H R28C	0 0.03 0 0.01 0.01 0 0.01 0 0.3 0 0 0.01 0 0 0 0.01 0.01 0	DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING DAMAGING	1 0.993 1 0.999 0.998 0.963 0.995 0.999 0.997 1 1 0.999 1 0.988 0.999 0.946 0.995 1	PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING PROBABLY DAMAGING

Table: 2. List of genes, damaged nsSNPs and affected amino acids their RMSD value, I-Mutant Score, DDG Values, RSA value before and after mutation by PyMol, I-Mutant2.0 and NetsurfP

Gene Name	nsSNP ID	Amino Acid	RMSD	I-Mutant2.0 Result		NetsurfP Result			
				I-Mutant Score	DDG Value (Kcal/mol)	Original class assignment	RSA	Mutant class assignment	RSA
ANK1	rs142542271	R1218Q	0.022	Decrease	0.31	E	0.395	E	0.401
	rs150655828	T472I	0.018	Decrease	-0.64	B	0.050	B	0.049
	rs142010751	R446T	0.022	Decrease	-0.39	E	0.424	E	0.412
	rs61735313	N251K	0.025	Decrease	-0.35	E	0.854	E	0.791
	rs146346710	T188M	0.039	Decrease	-0.95	B	0.043	B	0.041
	rs150032875	R669Q	0.025	Decrease	-2.46	E	0.522	E	0.579
EPB42	rs121917734	R280Q	0.023	Decrease	-1.68	B	0.032	B	0.033
	rs104894487	V112T	0.031	Decrease	-1.43	B	0.201	B	0.211
	rs28931585	R870W	0.034	Decrease	-1.36	B	0.169	B	0.170
	rs121912759	P868L	0.068	Decrease	-0.55	B	0.213	B	0.215
	rs121912750	T837A	0.022	Decrease	-2.94	B	0.218	B	0.207

SLC4A1	rs121912753	S773P	0.100	Decrease	-1.77	B	0.240	B	0.205
	rs121912741	G771D	0.023	Decrease	0.46	E	0.249	E	0.281
	rs121912755	R760Q	0.020	Decrease	-3.19	B	0.168	B	0.153
	rs121912748	G701D	0.011	Decrease	0.23	E	0.438	E	0.481
	rs121912757	R646Q	0.037	Decrease	-1.10	E	0.434	E	0.401
	rs121912758	R646W	0.038	Decrease	-0.54	B	0.434	E	0.420
	rs121912746	S613F	0.030	Decrease	-0.91	B	0.231	B	0.169
	rs121912754	R602P	0.030	Decrease	-2.94	E	0.092	E	0.312
	rs121912747	R589S	0.032	Decrease	-0.21	E	0.418	E	0.411
	rs121912745	R589C	0.016	Increase	-0.58	B	0.418	E	0.400
	rs28931584	V488M	0.033	Decrease	0.71	B	0.226	B	0.216
	rs145041032	L153M	0.030	Decrease	-0.32	B	0.042	B	0.041
SPTB	rs121918650	R2064P	0.065	Decrease	-0.45	E	0.434	E	0.418
	rs148337824	A1884V	0.032	Increase	0.45	E	0.360	E	0.293
	rs17245552	G1408R	0.017	Decrease	-0.22	E	0.555	E	0.558
	rs189656371	R1359W	0.032	Decrease	-0.77	E	0.517	E	0.315
	rs143827332	R1035W	0.039	Increase	-0.24	E	0.269	E	0.572
	rs145675502	D536N	0.039	Decrease	-0.57	E	0.612	B	0.228
	rs140796444	R498H	0.032	Decrease	-1.18	E	0.518	E	0.501
	rs150698819	D448N	0.014	Decrease	-1.94	E	0.522	E	0.563
	rs141973081	V350M	0.042	Decrease	-0.23	E	0.509	E	0.685
	rs121918646	W202R	0.020	Increase	-2.25	B	0.091	E	0.586
SPTA1	rs138732899	D2284N	0.075	Decrease	-0.14	E	0.430	E	0.541
	rs142775522	E2224D	0.012	Decrease	-0.64	E	0.355	E	0.531
	rs41273519	R2141W	0.025	Decrease	0.50	B	0.084	E	0.706
	rs78394850	R2016C	0.035	Increase	0.14	E	0.398	B	0.204
	rs116959874	L1804F	0.025	Decrease	-0.74	B	0.071	B	0.159
	rs182430449	V1698A	0.034	Decrease	0.26	B	0.103	E	0.462
	rs35237700	R1493W	0.029	Decrease	0.36	E	0.491	B	0.221
	rs34973695	L1485F	0.036	Increase	-0.36	B	0.125	B	0.215
	rs34214405	R1330I	0.033	Decrease	-0.34	E	0.473	E	0.682
	rs35733059	R1133P	0.081	Increase	1.55	E	0.403	E	0.474
	rs36057043	R891Q	0.020	Decrease	-1.19	B	0.237	E	0.598
	rs7547313	I809V	0.024	Decrease	-2.11	B	0.229	E	0.420
	rs121918634	L260P	0.088	Decrease	-2.43	E	0.282	E	0.864
	rs121918643	L207P	0.086	Increase	-0.36	B	0.116	E	0.439
	rs121918639	L49F	0.058	Decrease	0.50	B	0.138	E	0.289
	rs121918637	R45S	0.033	Decrease	0.13	B	0.190	E	0.412
	rs121918641	R28H	0.027	Decrease	-1.95	B	0.156	E	0.506
	rs121918642	R28C	0.022	Decrease	-2.08	B	0.156	B	0.184

DDG: free energy value. RMSD: root-mean-square deviation. RSA: Relative Solvent Accessibility.

Table3 : Membranes protein phenotype features prediction

Native amino acid	Mutant Amino acid	Characters
R	W	Tryptophan (mutant) is bigger than Arginine (native). Arginine residue charge was positive and Tryptophan residue charge is neutral. The native residue is more hydrophobic than the mutant residue
T	I	Threonine (native)is hydrophilic molecule and Isoleucine (mutant) residue is more hydrophobic nature.
R	T	Arginine residue charge was positive and Threonine (mutant)is hydrophilic nature
N	K	Lysine (mutant) residue is bigger than the Asparagine (native) residue. It has changed the charge from neutral into positive that can create repulsion between the mutant residue and neighboring residues.
T	M	Threonine (native)is hydrophilic molecule and Methionine is hydrophobic nature and it size bigger than native amino acid
R	Q	Glutamine (mutant) is smaller than Arginine (native). The native residue charge was positive; the mutant residue charge is neutral.
R	P	proline (mutant) is smaller than Arginine (native).The native residue charge was positive, proline residue charge is neutral.
A	V	Both residue are hydrophobic but valine (mutant) have bigger size than alanine
G	R	Arginine (mutant) is bigger than Glycine (native).. The native residue charge was neutral, the mutant residue charge is positive. Glycine is more hydrophobic than Arginine residue
D	N	Aspartic Acid (native) residue charge was negative and Asparagine (mutant) residue charge is neutral. The difference in charge will disturb the ionic interaction
R	H	Both residues are positive charges but Arginine (native) residues is bigger size than Histidine (mutant)
V	M	Methionine (mutant) is bigger than valine (native) residues; Both residues are hydrophobic nature.
W	R	Tryptophan (native) is bigger than Arginine (mutant). Arginine residue charge was positive and Tryptophan residue charge is neutral. The mutant residue is more hydrophobic than the native residue
P	L	proline (mutant) residue charge is neutral and leucine (native) is Non-polar, aliphatic residues
T	A	Threonine (native)is bigger than alanine (mutant). Threonine is hydrophilic molecule but alanine is hydrophobic.
S	P	Serine is polar residues from hydrogen bond and proline is hydrophobic molecules normally buried inside the protein core.
G	D	Glycine is smaller than Aspratic acid (mutant). Aspratic acid residues were negative charge it's from salt bridge.

S	F	phenylalanine (mutant) is bigger than Serine. Serine is polar residues from hydrogen bond but phenylalanine hydrophobic molecules normally buried inside the protein core.
R	C	Arginine residue charge was positive and cysteine (mutant) is hydrophilic molecule. Arginine is bigger than cysteine residues
R	S	Arginine residue charge was positive and serine (mutant) is hydrophilic molecule, Arginine is bigger than serine residues.
L	M	Methionine (mutant) is bigger than leucine (native).
A	T	Threonine (mutant) is bigger than alanine (native). Threonine is hydrophilic molecule but alanine is hydrophobic.
E	D	Both are negative charge molecules but glutamic acid is bigger than aspartic acid
L	F	Phenylalanine (mutant) is bigger than leucine (native).
V	A	Both residue are hydrophobic but valine (native) have bigger size than alanine (mutant)
R	I	Arginine (native) is bigger than isoleucine (native). Arginine residue charge was positive isoleucine is hydrophobic molecules normally buried inside the protein core.

A)

B)

ANK1

Fig.1. (A.) EPB42 in, (I) symbolizes A112T mutation (II) corresponds to R280Q mutation; Likewise, (B) For SLC4A1 protein, (I) characterizes the native structure of SLC4A1 and (II) represents the native structure superimposed to the mutant structure where (II) symbolizes the R589H mutation (III) represents S613F mutation, (IV) represents R646W mutation, (V) represents G701D mutation, (VI) represents G771D mutation, (VII) represents S773P mutation, (C) For ANK1 protein, is the native structure of ANK1 in cartoon representation and in the native structure is superimposed with the mutant structure where (I) corresponds to R446T mutation, (II) represents N251K mutation, (III) represents T188M mutation, (VI) represents R699Q mutation where the native residues are highlighted in cyan colour and the mutants in magenta colours

A) EPB4.2

B) SLC4A1

C) ANK1

Fig2. Stereochemical analysis of membrane proteins A) EPB4.2, B) SLC4A1 and C) ANK1. The red region declares the most favourable area of residues; the yellow region is additionally allowed. The RC plot of EPB42 declares 92 % of residues falling in allowed region, RC plot of SLC4A1 declares 86 % of residues falling in allowed region and RC plot of ANK1 declares 80.1% of residues falling in allowed region.

Figures

Figure 1

(A.) EPB42 in, (I) symbolizes A112T mutation (II) corresponds to R280Q mutation;Likewise, (B) For SLC4A1 protein, (I) characterizes the native structure of SLC4A1 and (II) represents the native structure superimposed to the mutant structure where (II) symbolizes the R589H mutation (III) represents S613F mutation, (IV) represents R646W mutation, (V) represents G701D mutation,(VI) represents G771D mutation, (VII) represents S773P mutation, (C) For ANK1protein, is the native structure of ANK1 in cartoon representation and in the native structure is superimposed with the mutant structure where (I) corresponds to R446T mutation, (II) represents N251K mutation, (III) represents T188M mutation, (VI) represents R699Q mutation where the native residues are highlighted in cyan colour and the mutants in magenta colours