Identification of Deleterious Single Nucleotide Polymorphism (SNP)s in the Human TBX5 Gene & Prediction of Their Structural & Functional Consequences: An *In Silico* approach

A M U B Mahfuz[1], and Md. Arif Khan[1,2]

[1]Department of Biotechnology & Genetic Engineering, Faculty of Life Science, University of Development Alternative, Dhaka-1209, Bangladesh

[2]Department of Molecular Medicine & Bioinformatics, Faculty of Life Science, University of Development Alternative, Dhaka-1209, Bangladesh

Correspondence should be addressed to arif.khan@bge.uoda.edu.bd

ABSTRACT:

T-box transcription factor 5 (TBX5) gene encodes the transcription factor TBX5 which plays a crucial role in the development of the heart and upper limbs. Alternative splicing resulting in several isoforms regulate the functions of this gene during the developmental process. Damaging single nucleotide variants in this gene alter the structure and disturb the functions of TBX5 and ultimately cause Holt-Oram Syndrome (HOS), an autosomal dominant disease where various congenital malformations of the heart (with or without conduction defects), upper limbs and shoulder girdles are observed. Besides HOS, TBX5 single nucleotide variants can also be associated with Dilated Cardiomyopathy, Atrial Fibrillation, and Tetralogy of Fallot without skeletal deformity.

By exploiting available Single Nucleotide Polymorphism information in dbSNP, this study was designed to identify *in silico* the deleterious TBX5 SNPs, and predict their structural and functional consequences, and alteration of biochemical properties on the candidate protein. For this purpose, various reliable *in silico* analysis tools such as PROVEAN, SIFT, PolyPhen-2, MutPred2, PredictSNP1, PredictSNP2, MetaLR, MetaSVM, REVEL, ConSurf, NetsurfP-2.0, iStable 2.0, Missense3D, UTRdb, MirSNP, and Human Splicing Finder (HSF) have been used. 58 missense substitutions were found damaging by both sequence homology-based tools SIFT (Sorting

Intolerant from Tolerant) and PROVEAN (Protein Variation Effect Analyzer), and structure homology-based tool PolyPhen-2 (Polymorphism Phenotyping-2). Among these 58 substitutions, 13 are already annotated as Pathogenic/Likely Pathogenic in ClinVar database, and so they were excluded. Then, the rest 45 high confidence substitutions were further scrutinized by various disease association predicting meta servers. Next, conservation profile of the native amino acid residues, their surface & solvent accessibility, and stability and structural integrity of the protein upon mutation were assessed. Analysis of 1 stop loss SNP, and 2 nonsense SNPs were done by PredictSNP2. Analysis of SNPs in the UTR region were done using UTRdb and MirSNP, and splice site SNPs were evaluated by Human Splicing Finder (HSF). This study provides a comprehensive list of most deleterious SNPs onTBX5 gene. The results from this study can help in early diagnosis of HOS and in relevant genetic counseling.

Keywords: TBX5, SNP, missense, nonsense, UTR, splice site, Holt-Oram Syndrome, congenital heart disease

Introduction:

Heart is the first organ to develop in the human body and numerous genes and their products orchestrate this process. Among these genes TBX5 is an important one. It is a member of the T-box gene family, the members of which encode transcription factors with a highly conserved DNA binding domain (T-box) of 180-200 amino acids [1]. Members of the T-box gene family have been found from metazoans to human and in humans more than 20 members of this gene family have been reported till now [2,3]. Several human diseases are associated with mutations in the T-box genes, namely DiGeorge Syndrome (TBX1), Ulnar-Mammary Syndrome (TBX3), Small Patella Syndrome (TBX4), Holt-Oram Syndrome(TBX5), Spondylocostal Dysostosis (TBX6), Cousin Syndrome (TBX15), Isolated ACTH Deficiency (TBX19), Congenital Heart Defects (TBX20), X-Linked Cleft Palate with or without Ankyloglossia (TBX22) [4].

According to the recent information retrieved from NCBI and UniProt, TBX5 gene is located on the long arm of chromosome12 at position 24.21 (12q24.21), contains 10 exons and has 3 transcript variants (1, 3 and 4). Transcript variants 1 and 4 encode a polypeptide of 518 amino acid residues and transcript variant 3 encodes a 468 (51-518) amino acid polypeptide with a shortened N-terminus. In TBX5 transcription factors, the T-box spans from amino acid position 58 to 238. The binding site of TBX5 has been identified to be present in the upstream regions of various cardiac-expressed genes, such as cardiac α actin (ACTC), atrial natriuretic factor (ANF), cardiac myosin heavy chain α (MYH6), cardiac myosin heavy chain β (MYH7), myosin light chain 1A (Mlc1a), myosin light chain 1V (Mlc1v) and Nkx2-5, and the consensus binding site for TBX5 corresponds to the first 8 bases from one half of the Brachyury binding 22 palindromic consensus sequence. It was showed *in vitro* that TBX5, when in full length, binds to the target site mostly as a dimer and when truncated, tends to bind mostly as a monomer with reduced dimer formation [5]. It was further showed *in vitro* that they can form a heterodimer complex with Nkx2-5, another important transcription factor for cardiac development in the course of DNA binding [6]. During the early stages of heart development, TBX5 acts mainly as a transcriptional activator of genes responsible for cardiomyocyte maturation and cardiac septa formation, and at the late phase it is necessary for the effective establishment of cardiac conduction system [7]. TBX5 also plays an important role in the very initiation of upper (fore) limbs [8] but the necessity of its presence during limb outgrowth phase remains controversial [8,9].

Holt-Oram Syndrome (HOS) is an autosomal dominantly inherited disease due to mutation in the TBX5 gene, and is considered 100% penetrant with varying expressivity in the anomalies of heart and upper extremity [10]. But incomplete penetrance [11], lack of penetrance [12], somatic mosaicism [12], and probable germinal mosaicism [13] have also been reported. The Majority of the patients suffer from this syndrome due to de novo mutations [14]. Cardiac presentations include atrial septal defect, ventricular septal defect, atrioventricular septal defect, pulmonary atresia/

stenosis, double outlet right ventricle, aortic valve insufficiency/ stenosis, tricuspid valve atresia, mitral valve abnormality, patent ductus arteriosus, tetralogy or pentalogy of Fallot, common arterial truncus, dextrocardia, right aortic arch and other non-specified congenital heart diseases. Upper limb anomalies may include hypoplastic or absent thumb, triphalangeal thumb, accessory/ bifid thumb, aplasia/ hypoplasia of hand and/ or fingers, syndactyly, radial with or without ulnar aplasia or hypoplasia, radio-ulnar synostosis, hypoplasia of humerus, phocomelia (aplasia of humerus) and clavicle abnormalities [15].

Nonsense mutation [16], missense mutation [17], frameshift mutation [16,17], deletion [18], duplication [19] and splice site mutation [20,21,22] of TBX5 gene have been reported to be responsible for HOS. It has been found missense mutations tend to cause more serious cardiac anomalies than mutations that cause a truncated TBX5 protein (nonsense, frameshift, splice site point variants, intragenic deletions, or duplications) [12]. HOS is caused by heterozygous mutation, and to the best of our knowledge only one homozygous missense mutation in the TBX5 gene has been reported till present but only with cardiac involvement (atrioventricular septal defect) and no skeletal involvement [23].

TBX5 missense mutations are also associated with familial [24] and sporadic Dilated Cardiomyopathy [25], and Atrial Fibrillation and Lone Atrial Fibrillation [26]. TBX5 nonsynonymous (missense and nonsense) mutations have also been reported in association with Tetralogy of Fallot without skeletal abnormality [27], and a TBX5 3′UTR variant makes more prone to congenital heart diseases [28]. Considering the above-stated findings, this study was conducted to find out the most deleterious TBX5 SNPs, and analyze their structural and functional outcomes *in silico*. The results from this extensive computational study could guide in early diagnosis and prenatal counseling of HOS in the future. Additionally, there is no full-length experimental structure available from Protein Data Bank, and three computational full-length models of TBX5 were generated using protein 3D structure prediction tools I-TASSER, Phyre2, and Robetta.

Materials and Methods:

Figure 1 depicts the overall workflow adopted in this study.

Retrieval of Sequences and Wild Structures: The nucleotide sequence of the human TBX5 gene was retrieved from National Center for Biotechnology Information (NCBI) 'Gene' database (available at https://www.ncbi.nlm.nih.gov/gene). The amino acid sequence of human TBX5 was downloaded in FASTA format from the UniProt database (UniProt ID: Q99593) (https://www.uniprot.org/uniprot/Q99593). Experimental structures of T-box were collected from Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (https://www.rcsb.org/).

Retrieval of SNPs: The SNP information about human TBX5 gene was collected using NCBI Variation Viewer (https://www.ncbi.nlm.nih.gov/variation/view/). Nonsense, stop lost, splice acceptor, splice donor, and UTR SNPs of TBX5 along with their rs IDs were retrieved from Variation Viewer applying the following filters: dbSNP under Source database, single nucleotide variant under Variant type, and nonsense (stop gained)/ stop lost/ splice acceptor variant/ splice donor variant/ 5 prime UTR variant/ 3 prime UTR variant under Molecular consequence. Information about missense SNPs were downloaded directly from the NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp/).

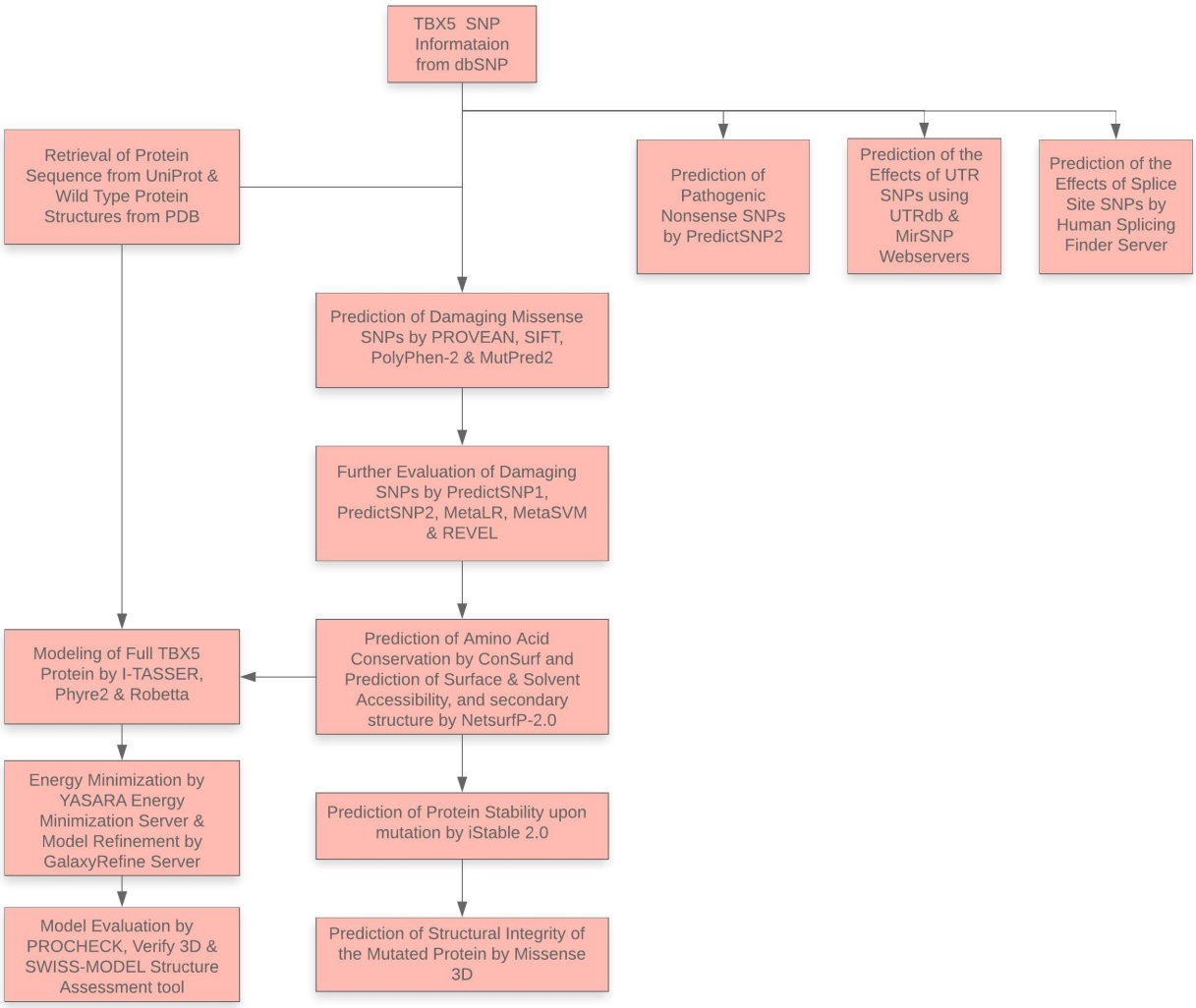


Figure 1: An overview of the steps followed to identify the deleterious TBX5 SNPs and analyze their structural and functional consequences.

A. Identification and Analysis of Damaging Missense SNPs: A total number of 355 missense SNPs were retrieved from dbSNP.

A1. Prediction of Missense SNPs by Sequence-Homology Based Tools: The missense SNPs were first subjected to analysis by sequence-homology based predictor PROVEAN (Human Genome Variants Tool, version 1.1.3), available at http://provean.jcvi.org/genome_submit_2.php?species=human . PROVEAN stands for PROtein Variation Effect ANalyzer. It predicts the effect of single or multiple amino acid substitution or indel on the functions of a protein. In PROVEAN,

BLAST hits with more than 75% global sequence identity are clustered first. The top 30 clusters constitute the supporting sequence set and a delta alignment score is computed for every supporting sequence. These scores are averaged within and across clusters to obtain the final PROVEAN score. An amino acid substitution is predicted 'Deleterious' if the PROVEAN score is below or equal to a predefined threshold value -2.5 and is considered 'Neutral' if the PROVEAN score is above this threshold [29]. In this study, a list of TBX5 gene SNPs defined by chromosomal position, reference allele, and variant allele was used as input for the PROVEAN server.

PROVEAN also provides SIFT score. SIFT (Sorting Intolerant From Tolerant), a web server available at https://sift.bii.a-star.edu.sg presumes that important amino acids are well conserved in a protein family and a replacement in a conserved zone is predicted as 'Damaging' [30]. For the submitted protein sequence, SIFT first compiles a dataset of functionally similar protein sequences by running BLAST searches from UniProt SwissProt, UniProt TrEMBL, or NCBI nr database and then build an alignment from the homologous sequences with the query sequence. Next SIFT algorithm finds the most probable amino acid for each position of the alignment and the calculated probabilities are recorded in a scaled probability matrix (the SIFT score). SIFT scores lie between 0-1, and amino acid substitutions with scores ≤0.05 are predicted to alter protein function and hence considered as 'Damaging' [31].

In this study, 85 SNPs were predicted as 'Deleterious' by PROVEAN and 236 SNPs were predicted as 'Damaging' by SIFT. 80 SNPs were predicted as harmful by both tools.

A2. Prediction of Missense SNPs by Structure-Homology Based Tool PolyPhen-2: PolyPhen-2 (Polymorphism Phenotyping version 2), available at http://genetics.bwh.harvard.edu/pph2/ predicts possible structural and functional impacts on a protein due to an SNP in the coding region. This prediction is made on eight sequence-based and three structure-based predictive features. Two pairs of datasets, viz. HumDiv and HumVar were recruited to train and test two PolyPhen-2 models, HumDiv was developed compiling 3,155 known human Mendelian allelic variants affecting the

function of proteins available from UniProt knowledge base as damaging mutation set and 6,321 differences between human proteins and their closely related mammalian homologous sequences as neutral mutation set. On the other hand, HumVar contains all the 13,032 human disease-causing mutations available in UniProt as damaging mutation set and 8,946 human nsSNPs without annotated involvement in disease as neutral mutation set. For sequence-based features, PolyPhen-2 BLASTs the query sequence against the UniRef100 database, and multiple sequence alignment developed from BLAST hits is used to calculate profile matrix score by PSIC (Position-Specific Independent Counts) software. This score conveys the likelihood of an amino acid's presence in a certain position of a protein sequence. For structure-based features, it BLASTs the query sequence against PDB (Protein Data Bank). It then performs analysis on the selected hits based on DSSP (Dictionary of Secondary Structure in Proteins) parameters and several other calculated parameters, and also considers 3 types of spatial contacts - contacts with heteroatoms, interchain contacts and contacts with functional sites. PolyPhen-2 calculates Naïve Bayes posterior probability of a given mutation's being damaging, and estimates false-positive rate and true-positive rate. It classifies SNPs as 'Probably Damaging', 'Possibly Damaging' or 'Benign' based on pairs of false-positive rate (FPR) thresholds [32 & Overview section in PolyPhen-2 website]. The 80 SNPs that were predicted harmful by both PROVEAN and SIFT were next analyzed by PolyPhen-2 using the batch query option. A list of SNPs defined by chromosome number, chromosomal position, reference allele, and variant allele was used for input. It predicted 58 SNPs as 'Probably Damaging' and these SNPs were considered of high fidelity in disease association. Indeed, this guess came true, and upon searching the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), 12 of these 58 SNPs were found to be already annotated as 'Pathogenic/ Likely Pathogenic' and only 1 is annotated as 'Benign/ Likely Benign'. Excluding these 13 SNPs, 45 SNPs were left for further evaluation.

A3. Prediction of Pathogenic Amino Acid Substitutions by MutPred2: MutPred2, available as both a web server (http://mutpred.mutdb.org/) and standalone software, integrates genetic and

molecular information to predict pathogenicity due to an amino acid substitution in a protein using a machine-learning approach. It has incorporated predictors for more than 50 structural and functional properties of protein which are altered in diseases due to mutations leading to amino acid replacement [33]. It is a bagged ensemble of 30 feed-forward neural networks, and every network is trained on a balanced subset of pathogenic and putatively neutral variants derived from the Human Gene Mutation Database (HGMD), SwissVar, dbSNP, and inter-species pairwise alignments. The final prediction score (MutPred2 score) is achieved by averaging the scores of all networks and the range of the final score is 0 to 1. The cutoff value for predicted pathogenicity is 0.5 and the higher the score, the greater the probability of that amino acid substitution's disease association. TBX5 protein sequence in FASTA format, and wild type and substituted amino acids of 45 high-confidence SNPs with their respective positions were submitted to the MutPred2 server using the default P-value threshold (0.05). Only one substitution (P337H) was found to have a MutPred2 score below 0.5 .

A4. Scrutinization of Predicted Pathogenicity by Meta-predictors: In recent years several meta-predictors have been developed to predict disease-causing SNPs. Such meta-predictors employ various single pathogenicity prediction tools and provide a consensus score gained from different ensemble methods. It is well established that ensemble methods render improved predictive performance than any of its constituent algorithms alone. We decided to crosscheck the predicted 45 pathogenic SNPs against 5 such meta-predictors, namely, PredictSNP1, PredictSNP2, MetaLR, MetaSVM, and REVEL.

i. PredictSNP1: PredictSNP1 (https://loschmidt.chemi.muni.cz/predictsnp1/) incorporates six individual predictors- MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT, and SNAP. PhD-SNP, PolyPhen-2, and SNAP use a decision model trained by different machine-learning algorithms. SIFT predicts on the basis of evolutionary information whereas MAPP also considers physicochemical properties for each amino acid residue besides evolutionary conservation, and

PolyPhen-1 employs an efficient set of empirical rules. These six individual tools have been carefully handpicked from a group of eight high-performance predictors to develop a consensus classifier (PredictSNP1) which outperforms all individual predictors [34].

ii. PredictSNP2: PredictSNP2 (https://loschmidt.chemi.muni.cz/predictsnp2/), like PredictSNP1, generates a consensus score from prediction of five tools- CADD, DANN, FATHMM, FunSeq2, and GWAVA. It can predict for nucleotide variants located in both coding and non-coding regions and returns result based on category-optimal decision thresholds for variants of five distinct categories- (1) regulatory, (2) splicing, (3) missense, (4) synonymous, and (5) nonsense variants [35]. The input was similar to that used for PolyPhen-2. After analysis by PredictSNP2, they were sent for analysis by PredicteSNP1 using the "Send to PredictSNP1" option available on the output page.

iii. MetaSVM and MetaLR: These are two meta-classifiers developed by dbNSFP project *(*https://sites.google.com/site/jpopgen/dbNSFP*)*. They both exert ensemble scores based on the prediction of deleteriousness by 10 component tools (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy & PhyloP) and the maximum frequency observed in the 1000 genomes project populations [36]. The difference between them lies in their adopted approaches- one uses the Support Vector Machine (MetaSVM) algorithm, and the other uses the Logistic Regression (MetaLR) algorithm. They were both accessed through Ensembl Variant Effect Predictor (VEP) (http://www.ensembl.org/Tools/VEP). rsID (Reference SNP ID number)s of the SNPs were used as input and the reference genome chosen was GRCh37.

iv. REVEL: REVEL (Rare Exome Variant Ensemble Learner) is a meta-predictor which integrates 13 tools to predict the probability of disease causation of a missense SNP. The individual tools are- MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons. REVEL adopts the Random Forest algorithm and has been trained using recently discovered pathogenic and rare neutral missense

variants after excluding those which have already been recruited to train its component tools. It is particularly efficient in differentiating pathogenic from rare neutral variants with allele frequencies <0.5%. REVEL scores range between 0 and 1, and the threshold score for pathogenic variants are 0.5 [37]. REVEL scores in this study were obtained alongside MetaSVM and MetaLR scores using Ensembl VEP.

A5. Prediction of Conserved Amino Acid Residues by ConSurf: It is well known that the structural and functional importance of an amino acid residue in a protein is strongly related to its level of conservation. Consurf (https://consurf.tau.ac.il/) ranks evolutionary conservation status of amino acid or nucleic acid positions in a protein or DNA/ RNA molecule respectively. It considers phylogenetic relations between homologous protein sequences for this ranking. It first builds a multiple sequence alignment (MSA) of homologous sequences and a phylogenetic tree is derived from this MSA. Next, ConSurf calculates position-specific conservation scores by an empirical Bayesian method or a maximum likelihood (ML) method [38]. The continuous conservation scores are split into an integer scale of 1 to 9 to depict by a color scheme. The most variable positions (grade 1) are colored turquoise, intermediate conserved positions (grade 5) are colored white, and the most conserved positions (grade 9) are colored maroon. TBX5 protein sequence in FASTA format served as the input. Homology modeling tool MODELLER was chosen for Protein three dimensional structure prediction [39,40].

A6. Surface & Solvent Accessibility and Secondary Structure Prediction by NetSurfP-2.0: Knowing the surface and solvent accessibility of amino acid residues is important to identify the interaction interfaces or active sites in a fully folded protein. Amino acid substitutions in such sites may bring change in binding affinity [41] or disturb catalytic activity if the protein is an enzyme [42]. NetSurfP-2.0 can estimate effectively surface & solvent accessibility, secondary structure, structural disorder, and backbone dihedral angles for amino acid residues. It takes protein sequences

in FASTA format as input and recruits deep neural networks that have been trained on solved protein structures [43]. NetSurfP-2.0 is available at http://www.cbs.dtu.dk/services/NetSurfP/.

A7. Prediction of Stability of the Mutated Protein due to SNPs by iStable 2.0: Amino acid substitutions caused by missense SNPs can alter the stability of the native protein which in turn can affect the functions of the protein and ultimately result in diseases [44]. Many computational tools have been developed to predict stability changes upon mutation. In our study, changes in protein stability due to missense SNPs have been predicted by a meta-classifier, iStable 2.0. It exploits machine learning and comments on the increase or decrease in protein stability due to an amino acid substitution based on the predictions of 3 sequence-based (I-Mutant2.0, MUpro and iPTREE-STAB) and 8 structure-based (I-Mutant2.0, CUPSAT, PoPMuSiC, AUTO-MUTE2.0, SDM, DUET, mCSM, MAESTRO and SDM2) stability prediction tools [45]. A 4-letter PDB code or protein sequence in FASTA format can be used as input, but the structure-based predictor has better performance than the sequence-based predictor. For amino acid substitutions upto F232V, chain A of TBX5 dimer crystal structure containing 1-239 residues (PDB ID: 5BQD) was used. Due to the unavailability of an experimental structure for the rest of the substitutions, they were analyzed using the protein FASTA sequence. iStable 2.0 web server can be found at http://ncblab.nchu.edu.tw/iStable2.

A8. Prediction of the Effects of Amino Acid Substitutions on the Structural Integrity of Protein by Missense3D: The phenotypic aftermath of missense SNPs can be better predicted by analyzing structural stability of the respective wild type and mutant proteins [44]. To analyze the impacts of amino acid substitutions on TBX5 three-dimensional structure, the Missense3D server was used (http://www.sbg.bio.ic.ac.uk/~missense3d/). It takes both experimental structures and predicted protein models as input for analysis and has shown that three-dimensional homology models can be effectively used for such analysis in case of unavailability of a crystal structure. It considers 16 structural features to differentiate disease-associated SNPs from the neutral ones. The

features are- disulfide bond breakage, introduction of a buried proline by the substitution, clash alert (if the mutant structure gains a MolProbity clash score ≥30 and the increase in clash score is >18 compared to the wild type), replacement of a buried hydrophobic residue with a hydrophilic residue, replacement of a buried uncharged residue with a charged residue or vice versa, change in charge type (+/−) of the buried residue, alteration of protein secondary structure, disallowed torsion angles (phi/ psi), breakage of buried H-bond, breakage of a buried salt bridge, expansion or contraction of a cavity or surface pocket volume of ≥ 70Å^3, change between buried and exposed state of the native residue, replacement of proline in the *cis* configuration, and replacement of glycine in a bending curvature. For substitutions upto F232V, 'Position on Protein Sequence' input option was chosen, and UniProt ID (Q99593) and PDB code & chain ID (5BQD; A) along with substitutions were used. For the rest of the substitutions, the energy minimized model generated by Robetta was used [64].

B. Identification and Analysis of Nonsense and Stop Lost SNPs by PredictSNP2: 31 nonsense (stop gained) and 1 stop lost SNPs were retrieved from NCBI Variation Viewer. Among the 31 nonsense SNPs, 29 are already annotated as Pathogenic in ClinVar database. Rest 2 SNPs, rs1393693495 (W514Ter) and rs1555226308 (Q156Ter) were analyzed by PredictSNP2. About PredictSNP2 has been discussed earlier. PredictSNP2 predicted both as 'Deleterious'. Q156Ter (rs1555226308) mutation is of particular importance because it disrupts the highly conserved T-box (amino acid position 58-238). 1 stop lost SNP (rs1394777873) was also analyzed using PredictSNP2 and was predicted to be neutral.

C. Prediction of the effects of UTR SNPs by UTRdb & MirSNP: Untranslated regions (UTR) play a pivotal role in the regulation of gene expression post-transcriptionally. They ascertain nucleo-cytoplasmic transport of mRNA, maintain translational efficiency, subcellular localization, and RNA stability [46]. The important regulators in 5′-UTR are Kozak consensus sequence (ACCAUGG) [Shine-Dalgarno consensus sequence (AGGAGG) in prokaryotes], CpG (5'—C—

phosphate—G—3') sites, uORFs (upstream open reading frames), IRESs (Internal Ribosome Entry Sites), and RBP (Ribosome-binding protein) binding sites. 3′-UTR length, RBP, and miRNA binding sites are the important players in 3′-UTR [47]. Single nucleotide variants in these regions have been found to be associated with many diseases- Thalassemia intermedia [48], atrial septal defect [49], campomelic dysplasia [50], hereditary chronic pancreatitis [51], Marie Unna hereditary hypotrichosis [52], Charcot-Marie-Tooth disease [53], and cerebral amyloid angiopathy [54], to name a few. So, it is important to identify potentially harmful SNPs in the 3′ & 5′-UTR. UTR SNPs that were predicted to affect transcriptional motifs have been retrieved from the UTRsite database (UTRdb). It is a curated database of 5' and 3' untranslated sequences of eukaryotic mRNAs and has been developed by data mining from various primary sources. Here UTR entries are organized in a gene-centric manner [55]. For this study, 'UTRef' (NCBI RefSeq transcripts) as searching database and 'Gene Name' as accession type were chosen during retrieval. UTRdb is freely available at http://utrdb.ba.itb.cnr.it/.

Again, microRNA (miRNA)s have also a role in post-transcriptional regulation of gene expression through mRNA cleavage or post-translational repression [56]. miRNAs are small non-coding RNAs having about 22 nucleotides. miRNA seed region (positions 2-8) is the most important player in finding the complementary sequence in mRNA and binding with it. SNPs in microRNA target sites present in 3′ UTR may create or delete a miRNA target site or change the binding efficiency between miRNA and mRNA. Creation or deletion of a miRNA target site may also affect binding to the neighboring miRNA target sequences. Such SNPs have been reported in several diseases including cancer, psychiatric illness, cardiomyopathy, asthma, and Parkinson's disease [57,58]. In this study, SNPs in predicted miRNA target sites were identified using the MirSNP server (http://bioinfo.bjmu.edu.cn/mirsnp/search/). It can predict whether an SNP in the 3′UTR affects the miRNA target site or not. It categorizes SNPs into four groups, labeled as create, enhance, decrease

or break. The sources of human miRNA and SNP it uses are miRBase18.0 and NCBI dbSNP135 respectively [58]. Gene name was used as input.

D. Predicting the effects of SNPs located in splice sites by HSF (Human Splicing Finder): RNA splicing is necessary for the conversion of precursor mRNA to a mature mRNA. Splicing removes introns and joins exons to make an mRNA ready for translation. Various *cis*-regulatory elements and *trans*-acting elements maintain this splicing process. *Cis*-regulatory elements include 5′ and 3′ consensus splice site sequences in the exon-intron boundaries, and branch point and polypyrimidine tract sequences. 5′ and 3′ consensus splice site sequences are known as splice donor and splice acceptor sites respectively. Auxillary *cis* elements are- exonic and intronic splicing silencer and enhancer sequences. *Trans*-acting elements include spliceosomal small nuclear RNA (snRNA)s and proteins, and various other splicing repressors and activators, many of which are still unknown. Mutations that hamper the controlled process of splicing lead to a good number of diseases [59,60]. SNPs in splice donor and splice acceptor sites may break an established splice site or create a new splice site. They may also activate cryptic sequences (sequences similar to splice sites), affect splicing enhancers and silencers. They may even intercept binding of spliceosomal components by making a conformational change in the mRNA secondary structure [59]. So, it is needless to say the importance of splice site mutations. In this study, 7 splice donor single nucleotide variants and 7 splice acceptor single nucleotide variants were retrieved from dbSNP using NCBI Variation Viewer. Among the 7 donor variants, 2 are annotated as 'Likely-pathogenic', and among the 7 acceptor variants, 1 is annotated as 'Pathogenic' and 3 are annotated as 'Likely-pathogenic' in ClinVar database. Excluding them, the rest 8 SNPs were analyzed employing the Human Splicing Finder (HSF) server available at http://www.umd.be/HSF/. It integrates 12 distinct algorithms to predict the effect of a mutation on splicing signals, and can identify splicing motifs and auxiliary motifs in a given sequence. It can also identify binding sites of the splicing regulator ribonucleoprotein hnRNP A1, and SR (serine-arginine rich) proteins- 9G8 and Tra2-*β*. These

algorithms are established on position weight matrices (PWM), the Maximum Entropy principle, or the Motif Comparison method, and some of these algorithms are borrowed from similar predicting tools. For each algorithmic decision, there is a consensus value threshold and a score variation threshold [61]. For our analysis, 'Analyze a sequence' and 'Paste your own sequence' options were chosen and individual variants with flanking sequences of 100 nucleotides in both 5′ and 3′ direction (retrieved from dbSNP) were used as inputs.

3D Structure prediction for Full TBX5 Protein: Currently there are 4 experimental structures of TBX5 (2X6U, 2X6V, 4SOH & 5BQD) are deposited in the Protein Data Bank (PDB). But, none of them contains the full-length protein. So, we decided to predict the structure of the whole protein by different reliable protein three-dimensional structure prediction tools. Unfortunately, no full-length experimental close or distant homolog structure of TBX5 is currently available and so, 3D structure predictors which adopt template-based modeling like MODELLER [40] failed to generate a model of the whole protein. We used 3 other webtools, namely, I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/), Phyre2 (http://www.sbg.bio.ic.ac.uk/servers/phyre2/html/page.cgi?id=index) and Robetta (https://robetta.bakerlab.org/) which can predict protein tertiary structure by incorporating both template-based and template-free (*de novo* modeling) methods [62,63,64]. Energy minimization of the generated TBX5 models to make them more stable was achieved by applying YASARA force field through the YASARA Energy Minimization Server (http://www.yasara.org/minimizationserver.htm). YASARA force field combines the AMBER (Assisted Model Building with Energy Refinement) all-atom force field equation with multi-dimensional knowledge-based torsional potentials. It engages 42 force field parameters, optimized with random Monte Carlo moves in forcefield parameter space to gain the highest possible accuracy [65,66].

The energy minimized models were further refined by the GalaxyRefine server (http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=REFINE). GalaxyRefine first rebuilds all the side-chain conformations by placing the most suitable rotamers. Next, it performs repeated structure perturbation by side-chain repacking followed by overall structure relaxation by molecular dynamics simulation to find the lowest energy model [67].

Evaluation of TBX5 3D Models: Quality of the 3 predicted models of TBX5 were evaluated by PROCHECK, Verify 3D, and SWISS-MODEL Structure Assessment Tool. PROCHECK examines certain stereochemical and geometrical properties of the query protein structure which are considered good indicators of protein quality and compares them against a set of ideal values obtained from well-refined high-resolution protein structures. It assesses residue-by-residue geometry and overall structure geometry of the query structure [68]. A Ramachandran plot of φ-ψ torsion angles generated by PROCHECK is the most handy quality indicator, and a good quality model is expected to have over 90% residues in the most favored regions. Verify 3D checks the agreement between a protein tertiary structure and its primary structure [69]. It classifies amino acid residues into different structural categories based on their location and environment (alpha, beta, loop, polar, nonpolar etc.) and compares them with the most compatible sequences (derived from experimentally sound structures) for the respective environments in the tertiary structure. PROCHECK and Verify 3D are freely available at https://servicesn.mbi.ucla.edu/PROCHECK/ and at https://servicesn.mbi.ucla.edu/Verify3D/. SWISS-MODEL Structure Assessment Tool (https://swissmodel.expasy.org/assess) performs a detailed structural analysis of the query structure. It provides estimations from MolProbity and QMEAN. These tools evaluate a model quality at both global (i.e. for the entire structure) and local (i.e. per residue) level [70,71]. QMEAN Z-score is an indicator of the “degree of nativeness” of a model [72].

Results:

SNP Information Retrieval: Searching the NCBI Variation Viewer using dbSNP as the source database and single nucleotide variant (SNV) as variant type returned 14413 rsIDs to be SNV. 355 rsIDs were found to be missense variants, 30 rsIDs to be nonsense (stop gained) variants, 1 rsID to be stop lost variant, 210 rsIDs to be synonymous variants, 11300 rsIDs to be intron variants, 821 rsIDs to be noncoding transcript variants, 7 rsIDs to be splice acceptor & 7 rsIDs to be splice donor variants, 689 rsIDs to be 5 prime UTR variants, and 345 rsIDs to be 3 prime variants . On further inspection, 404 missense variants with 355 rsIDs (3 nonsense variants included in rs377649723, rs765204502, and rs903933027, and 11 synonymous variants included in rs186780790, rs377532269, rs533581420, rs374600913, rs753688559, rs762204624, rs771883815, rs777853147, rs972633602, rs973631007 and rs769107196) and 28 nonsense variants with 28 rs IDs (27 stop gained and 1 stop lost- rs1394777873) were found. The reason behind a single rsID having more than one variant is these variants were reported on the same chromosomal location and as per NCBI rule only one rsID is assigned for a chromosomal location [73].

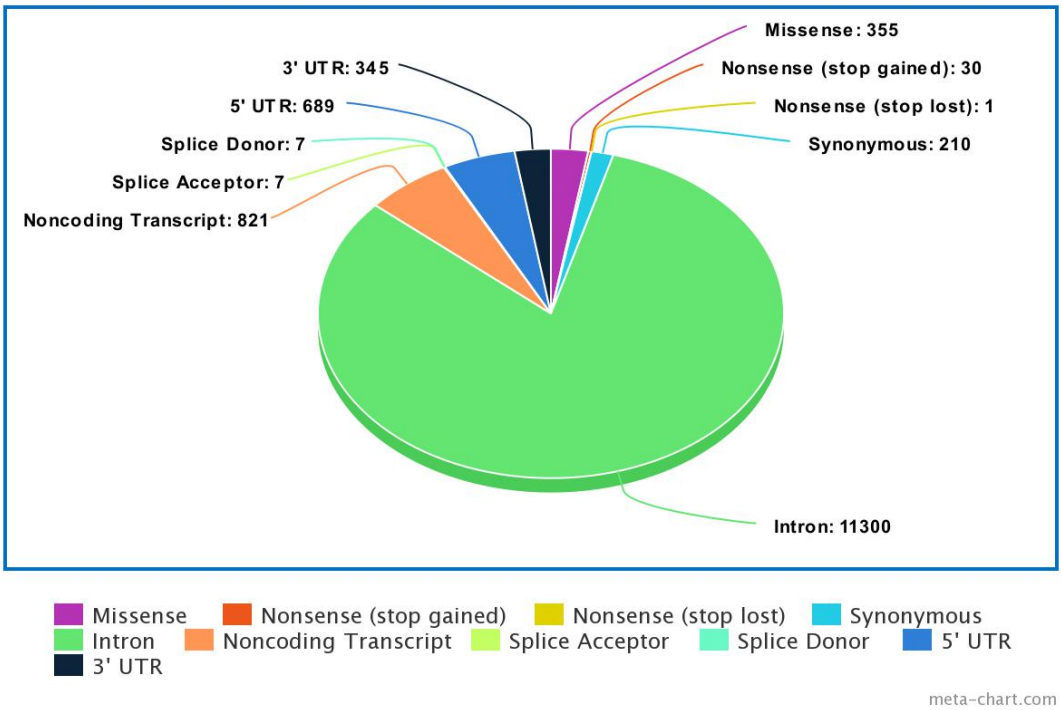


Figure 2 : Pie-chart displaying distribution of SNPs in TBX5 gene.

A. Identification and Analysis of Damaging Missense SNPs:

A1. Prediction of Missense SNPs by Sequence-Homology Based Tools PROVEAN & SIFT: The retrieved 404 missense SNPs were analyzed first using web-based tool PROVEAN. Out of 404 submitted substitutions, PROVEAN predicted 85 as 'Deleterious' and 301 as 'Neutral', and SIFT predicted 236 as 'Damaging' and 150 as'Tolerated'. 242 substitutions were predicted as 'Deleterious' or 'Damaging' by either PROVEAN or SIFT, and 80 substitutions were predicted as 'Deleterious' or 'Damaging' by both tools. Results of rs571328934, rs755664220, rs756902625, rs763178387, rs769107196, rs769895888, rs777370158, rs907398622, rs945043550, rs1263504354, rs1407437475, and rs1489969947 could not be provided by PROVEAN.

A2. Prediction of Missense SNPS by Structure-Homology Based Tool PolyPhen-2: For high fidelity, the 80 SNPs predicted by both PROVEAN and SIFT as 'Deleterious'/ 'Damaging' were next analyzed by PolyPhen-2. Both HumDiv and HumVar dataset trained PolyPhen-2 were used. HumDiv model predicted 58 substitutions as 'probably damaging' (Table-1), 16 substitutions as 'possibly damaging' with the following rsIDs: rs1287994219(V56E), rs777853147(R61S), rs1555226581(W64C), rs1555226577(M74V), rs763219727(A79V), rs515726234(I101F), rs376621016(P108S), rs754145043(V123M), rs1220789160(G125A), rs1274219276(A127S), rs1318693560(M131V), rs771361403(Q156L), rs748766700(K221N), rs1207284425(R249G), rs1565927794(S282F) and rs1565923544(P434L), and 6 substitutions as 'benign' which are rs1327141955(V56L),rs756464860(E62K),rs1457871803(K66I),rs1555226329(V123A), rs533581420(F206L) and rs757039207(K221E). HumVar model predicted 56 substitutions as 'probably damaging' (Table-1), 10 substitutions as 'possibly damaging' which are rs1287994219(V56E), rs1555226581(W64C), rs515726234(I101F), rs376621016(P108S), rs754145043(V123M), rs1220789160(G125A), rs1274219276(A127S), rs771361403(Q156L), rs190881877(V205F) and rs944423586(P337H), and 14 substitutions as 'benign' which are

rs1327141955(V56L), rs777853147(R61S), rs756464860(E62K), rs1457871803(K66I), rs1555226577(M74V), rs763219727(A79V), rs1555226329(V123A), rs1318693560(M131V), rs533581420(F206L), rs748766700(K221N), rs757039207(K221E), rs1207284425(R249G), rs1565927794(S282F), and rs1565923544(P434L). The 58 SNPs that were predicted as 'probably damaging' by HumDiv model were also predicted as 'probably damaging by HumVar model except rs190881877(V205F) and rs944423586(P337H) which were predicted as 'possibly damaging'. Upto this stage, all analysis were conducted *in silico.* Upon searching ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar), among these 58 SNPs, 12 SNPs with 11 rsIDs were found as 'Pathogenic'/'Likely pathogenic'. These are- rs104894381 (G80R), rs1131691460 (F84L), rs1565941579(P85S), rs483353129(R113K), rs863223773(G125R), rs1057519050 (L135R), rs1555225344 (T223M), rs1565935426(K226R), rs1131691932(K226E), rs104894378(R237P), rs104894378(R237Q) and rs104894382(R237W), and 1 was found to be 'Benign'/'Likely Benign' rs77357563(D111Y). These 13 SNPs were excluded and left 45 SNPs were considered for further validation. rs 1411518530 is associated with two nucleotide variants (G>C,T) but they encode the same amino acid (D110E), and can be considered as a single variant.

dbSNP ID	**Codon Change**	**AAS**	**PROVEAN**		**SIFT**		**PolyPhen-2 Prediction**	
			Prediction	**Score**	**Prediction**	**Score**	**Hum Div**	**Hum Var**
rs1322050108	C[T/C]C	L58P	Del	-5.86	Dam	0	PD	PD
rs863223779	A[C/A]G	T72K	Del	-5.31	Dam	0	PD	PD
rs863223779	A[C/T]G	T72M	Del	-5.31	Dam	0	PD	PD
rs1565941576	C[C/T]C	P85L	Del	-9.16	Dam	0	PD	PD
rs775593443	G[T/C]G	V91A	Del	-2.7	Dam	0.001	PD	PD
rs770717773	[G/A]GC	G93S	Del	-5.5	Dam	0.001	PD	PD
rs890394778	G[G/A]C	G93D	Del	-6.41	Dam	0.001	PD	PD
rs890394778	G[G/T]C	G93V	Del	-8.24	Dam	0	PD	PD
rs1203503352	[C/T]CC	P96S	Del	-3.75	Dam	0.024	PD	PD
rs1463838783	T[A/G]C	Y100C	Del	-8.24	Dam	0	PD	PD
rs1463838783	T[A/T]C	Y100F	Del	-3.66	Dam	0.004	PD	PD
rs1350768465	A[T/C]G	M104T	Del	-5.16	Dam	0.038	PD	PD
rs1565941484	[G/A]AC	D105N	Del	-4.59	Dam	0	PD	PD

rs1369673380	G[T/A]A	V107E	Del	-5.51	Dam	0	PD	PD
rs376621016	[C/A]CT	P108T	Del	-3.41	Dam	0.001	PD	PD
rs1411518530	GA[C/G]	D110E	Del	-3.47	Dam	0	PD	PD
rs1411518530	GA[C/A]	D110E	Del	-3.47	Dam	0	PD	PD
rs1428277072	[G/A]AC	D110N	Del	-4.19	Dam	0.001	PD	PD
rs483353129	A[G/T]A	R113I	Del	-7.34	Dam	0	PD	PD
rs757586102	[T/C]AC	Y114H	Del	-4.59	Dam	0	PD	PD
rs1269970792	[T/C]GG	W121R	Del	-12.86	Dam	0	PD	PD
rs1370517694	[G/T]GC	G133C	Del	-4.87	Dam	0.023	PD	PD
rs1178440681	G[T/C]G	V137A	Del	-3.7	Dam	0	PD	PD
rs148715804	GA[C/G]	D140E	Del	-3.7	Dam	0.008	PD	PD
rs1424817435	[C/T]CC	P142S	Del	-7.4	Dam	0	PD	PD
rs374906778	[G/A]CC	A143T	Del	-3.2	Dam	0.011	PD	PD
rs1555226320	[G/T]GG	G145W	Del	-7.4	Dam	0	PD	PD
rs772248871	G[T/A]C	V153D	Del	-6.29	Dam	0	PD	PD
rs1464680326	[C/T]TC	L160F	Del	-3.5	Dam	0.001	PD	PD
rs1183256015	C[A/T]T	H170L	Del	-4.92	Dam	0.001	PD	PD
rs1565939347	[A/G]AT	N174D	Del	-4.79	Dam	0.001	PD	PD
rs1555226019	G[T/G]G	V186G	Del	-6.42	Dam	0	PD	PD
rs557183241	[G/A]TG	V186M	Del	-2.71	Dam	0.006	PD	PD
rs1392668757	[G/T]AT	D189Y	Del	-2.81	Dam	0	PD	PD
rs763004247	[C/G]AC	H204D	Del	-5.74	Dam	0.002	PD	PD
rs190881877	[G/T]TC	V205F	Del	-3.27	Dam	0.001	PD	PoD
rs769113870	GA[G/C]	E208D	Del	-2.78	Dam	0.001	PD	PD
rs987371810	[G/A]CA	A213T	Del	-3.33	Dam	0.039	PD	PD
rs1318021626	A[T/A]C	I222N	Del	-6.66	Dam	0	PD	PD
rs1565935405	[T/G]TT	F232V	Del	-6.7	Dam	0	PD	PD
rs377625550	[A/T]GC	S261C	Del	-2.61	Dam	0	PD	PD
rs1298950685	T[G/A]T	C323Y	Del	-2.95	Dam	0.001	PD	PD
rs944423586	C[C/A]C	P337H	Del	-2.99	Dam	0	PD	PoD
rs377532269	[C/T]GG	R375W	Del	-3.01	Dam	0	PD	PD
rs377649723	TG[G/T]	W401C	Del	-3.14	Dam	0	PD	PD

Table 1: Deliterious SNPs predicted by PROVEAN, SIFT and PolyPhen-2
AAS= Amino Acid Substitution, Del=Deleterious, Dam=Damaging, PD=Probably Damaging & PoD=Possible Damaging

A3. Prediction of Pathogenic Amino Acid Substitutions by MutPred2: MutPred2 considers several molecular characteristics of amino acid residues to predict whether an amino acid substitution is disease-related or neutral in humans. The score it provides is the probability predicted for an amino acid substitution should affect the function of the respective protein or not. The threshold score for pathogenicity prediction is 0.5 and a MutPred2 score ≥ 0.8 can be

considered as a highly confident one. All but one (P337H) substitutions have prediction scores ≥ 0.5. Table-2 provides MutPred2 outcomes.

AAS	MutPred2 score	Molecular mechanisms disrupted with P-values <=0.05	Probability	P-value
L58P	0.951	Altered Stability	0.34	0.0042
		Loss of Helix	0.29	0.01
		Gain of Strand	0.28	0.01
		Altered Metal binding	0.26	0.02
T72K	0.958	Gain of Allosteric site at T72	0.34	0.00056
		Altered Metal binding	0.33	0.0026
		Gain of Acetylation at T72	0.22	0.03
		Loss of Catalytic site at E73	0.17	0.02
		Altered DNA binding	0.15	0.04
T72M	0.922	Gain of Allosteric site at T72	0.29	0.0031
		Altered Metal binding	0.27	0.0074
		Gain of Catalytic site at E73	0.16	0.02
P85L	0.887	Altered Ordered interface	0.32	0.01
		Loss of Allosteric site at R82	0.29	0.0051
		Altered DNA binding	0.28	0.0048
		Loss of Acetylation at K88	0.25	0.01
		Loss of Relative solvent accessibility	0.23	0.05
		Altered Disordered interface	0.19	0.04
		Gain of Catalytic site at R81	0.19	0.01
		Gain of Methylation at K88	0.09	0.05
V91A	0.609	Altered Ordered interface	0.3	0.02
		Gain of Acetylation at K88	0.24	0.02
		Altered DNA binding	0.23	0.01
		Altered Stability	0.16	0.02
		Gain of Methylation at K88	0.1	0.04
G93S	0.865	Altered Ordered interface	0.33	0.0074
		Gain of Relative solvent accessibility	0.32	0.0038
		Altered DNA binding	0.24	0.008
		Loss of Acetylation at K88	0.24	0.02
G93D	0.925	Altered Ordered interface	0.3	0.02
		Altered DNA binding	0.3	0.0015
		Gain of Relative solvent accessibility	0.27	0.02
		Loss of Strand	0.26	0.03
		Loss of Acetylation at K88	0.25	0.01
G93V	0.927	Loss of Relative solvent accessibility	0.31	0.0072
		Altered Ordered interface	0.29	0.02
		Loss of Acetylation at K88	0.25	0.01
		Altered DNA binding	0.24	0.0093
P96S	0.806	Altered Ordered interface	0.33	0.0071
		Gain of Relative solvent accessibility	0.33	0.0025

		Gain of Acetylation at K99	0.25	0.01
		Loss of Allosteric site at Y100	0.2	0.04
		Loss of Methylation at K99	0.1	0.04
Y100C	0.945	Altered Ordered interface	0.36	0.0044
		Loss of Acetylation at K99	0.31	0.0036
		Loss of Relative solvent accessibility	0.28	0.02
		Altered Metal binding	0.24	0.04
		Loss of Allosteric site at D105	0.24	0.02
		Gain of Catalytic site at D105	0.12	0.03
		Loss of Methylation at K99	0.11	0.04
Y100F	0.75	Altered Ordered interface	0.35	0.0052
		Loss of Relative solvent accessibility	0.28	0.02
		Loss of Acetylation at K99	0.28	0.0063
		Altered Metal binding	0.25	0.03
		Loss of Allosteric site at D105	0.24	0.02
		Gain of Catalytic site at D105	0.1	0.04
		Loss of Methylation at K99	0.1	0.04
M104T	0.945	Altered Ordered interface	0.3	0.02
		Gain of Relative solvent accessibility	0.3	0.0086
		Altered Stability	0.27	0.0067
		Altered Metal binding	0.26	0.01
		Gain of Allosteric site at D105	0.24	0.01
		Gain of Acetylation at K99	0.24	0.02
		Gain of Methylation at K99	0.11	0.03
		Gain of Catalytic site at D105	0.1	0.04
D105N	0.913	Altered Metal binding	0.51	0.0036
		Altered Ordered interface	0.29	0.03
		Gain of Strand	0.26	0.04
		Gain of Allosteric site at D105	0.23	0.02
		Gain of Catalytic site at D105	0.1	0.04
V107E	0.934	Altered Metal binding	0.42	0.00098
		Altered Ordered interface	0.32	0.01
		Gain of Allosteric site at D105	0.24	0.01
		Gain of Catalytic site at D105	0.14	0.02
P108T	0.893	Altered Metal binding	0.29	0.005
		Altered Ordered interface	0.28	0.04
		Loss of Allosteric site at D105	0.25	0.02
		Loss of Catalytic site at D105	0.1	0.04
D110E	0.822	Altered Metal binding	0.4	0.0012
		Altered Ordered interface	0.24	0.04
		Gain of Allosteric site at D105	0.23	0.02
		Altered DNA binding	0.17	0.03
		Gain of Catalytic site at D105	0.11	0.03
D110N	0.862	Altered Metal binding	0.47	0.0006
		Altered Ordered interface	0.3	0.02
		Gain of Strand	0.26	0.05

		Loss of Allosteric site at D105	0.22	0.03
		Altered DNA binding	0.16	0.04
		Gain of Catalytic site at D105	0.11	0.04
R113I	0.937	Gain of Relative solvent accessibility	0.36	0.0011
		Altered Metal binding	0.33	0.001
		Altered Ordered interface	0.3	0.02
		Gain of Acetylation at K115	0.19	0.05
		Altered DNA binding	0.17	0.04
		Gain of Methylation at K115	0.09	0.05
Y114H	0.933	Altered Metal binding	0.41	0.0021
		Altered Ordered interface	0.31	0.01
		Loss of Relative solvent accessibility	0.27	0.02
		Altered DNA binding	0.17	0.03
		Altered Stability	0.13	0.03
W121R	0.965	Gain of Intrinsic disorder	0.3	0.04
		Loss of Relative solvent accessibility	0.28	0.02
		Loss of Strand	0.26	0.05
		Loss of Allosteric site at W121	0.21	0.04
		Gain of Methylation at K126	0.16	0.01
G133C	0.763	Gain of Allosteric site at R134	0.34	0.00046
		Altered Metal binding	0.27	0.0072
		Gain of Catalytic site at H138	0.2	0.01
V137A	0.697	Altered Stability	0.41	0.0028
		Loss of Allosteric site at R134	0.26	0.01
		Altered Metal binding	0.24	0.01
		Loss of Catalytic site at D140	0.21	0.01
D140E	0.744	Altered Metal binding	0.3	0.0054
		Gain of Catalytic site at D140	0.24	0.0053
		Loss of Allosteric site at H138	0.23	0.03
P142S	0.747	Loss of Loop	0.27	0.03
		Altered Metal binding	0.26	0.0093
		Loss of Allosteric site at H138	0.24	0.02
		Gain of Catalytic site at D140	0.21	0.0094
A143T	0.541	Altered Metal binding	0.24	0.01
		Loss of Allosteric site at H138	0.23	0.02
		Loss of Catalytic site at D140	0.21	0.01
G145W	0.921	Loss of Loop	0.27	0.04
		Altered Metal binding	0.25	0.01
		Altered Ordered interface	0.25	0.02
		Gain of Catalytic site at D140	0.23	0.0065
V153D	0.942			
L160F	0.833	Gain of Strand	0.26	0.04
		Gain of Allosteric site at N163	0.19	0.05
		Loss of Catalytic site at K159	0.09	0.05
H170L	0.936	Altered Ordered interface	0.29	0.03
		Altered Metal binding	0.28	0.02

		Loss of Allosteric site at H170	0.19	0.05
N174D	0.741	Altered DNA binding	0.29	0.0023
		Altered Ordered interface	0.27	0.05
		Altered Metal binding	0.25	0.0093
		Loss of Allosteric site at H177	0.25	0.02
V186G	0.898	Altered Stability	1	0.00012
		Gain of Intrinsic disorder	0.31	0.04
		Altered Metal binding	0.22	0.02
V186M	0.709	Altered Metal binding	0.22	0.02
		Altered Stability	0.18	0.01
D189Y	0.823	Altered Metal binding	0.23	0.02
H204D	0.902	Altered Ordered interface	0.29	0.02
		Gain of Loop	0.27	0.03
		Altered Transmembrane protein	0.15	0.02
V205F	0.855	Altered Transmembrane protein	0.13	0.02
E208D	0.809	Altered Transmembrane protein	0.14	0.02
A213T	0.805	Altered Ordered interface	0.25	0.02
		Altered Transmembrane protein	0.21	0.0043
		Altered DNA binding	0.16	0.04
		Loss of Catalytic site at Y217	0.15	0.02
		Altered Metal binding	0.13	0.03
I222N	0.943	Altered Ordered interface	0.32	0.0091
		Altered Metal binding	0.24	0.02
		Altered Transmembrane protein	0.21	0.0045
		Gain of Catalytic site at Y217	0.18	0.01
		Altered DNA binding	0.18	0.03
F232V	0.754	Altered Disordered interface	0.37	0.0063
		Loss of Acetylation at K234	0.29	0.0051
		Altered DNA binding	0.17	0.03
		Altered Transmembrane protein	0.12	0.03
		Gain of Proteolytic cleavage at R237	0.12	0.03
S261C	0.5	Loss of ADP-ribosylation at R260	0.23	0.02
		Loss of SUMOylation at K266	0.2	0.04
		Loss of O-linked glycosylation at T262	0.14	0.04
		Loss of Proteolytic cleavage at R260	0.11	0.04
C323Y	0.623	Gain of Phosphorylation at Y321	0.3	0.01
		Loss of SUMOylation at K327	0.26	0.0069
		Gain of Sulfation at Y321	0.03	0.01
P337H	0.255			
R375W	0.676	Loss of Phosphorylation at Y380	0.31	0.02
		Loss of Pyrrolidone carboxylic acid at Q376	0.06	0.03
W401C	0.76	Gain of Intrinsic disorder	0.45	0.0036
		Gain of Loop	0.27	0.03
		Loss of Disulfide linkage at C398	0.1	0.05

Table 2: Effects of amino acid substitutions on the TBX5 protein predicted by MutPred2

A4. Scrutinization of Predicted Pathogenicity by Meta-predictors: The pathogenicity of the aforementioned substitutions were further examined by 5 deleteriousness predicting metaservers whose predictions are based on various ensemble methods. The metaservers used were PredictSNP1, PredictSNP2, MetaLR, MetaSVM, and REVEL. 34 substitutions were predicted 'Damaging' by all metaservers and the rest were predicted 'Damaging' by at least 2 metaservers.

AAS	PredictSNP 1		PredictSNP 2		MetaLR		MetaSVM		REVEL	
	Prediction	Expected Accuracy	Prediction	Expected Accuracy	Prediction	Score	Prediction	Score	Prediction	Score
L58P	D	0.87	D	0.87	D	0.93	D	1.08	D	0.98
T72K	D	0.87	D	0.87	D	0.88	D	1.02	D	0.94
T72M	D	0.87	D	0.87	D	0.83	D	0.86	D	0.95
P85L	D	0.87	D	0.87	D	0.96	D	1.09	D	0.96
V91A	D	0.65	D	0.87	D	0.78	D	0.62	D	0.73
G93S	D	0.87	D	0.87	D	0.96	D	1.10	D	0.97
G93D	D	0.87	D	0.87	D	0.96	D	1.10	D	0.94
G93V	D	0.61	D	0.87	D	0.96	D	1.10	D	0.94
P96S	N	0.63	D	0.87	D	0.86	D	0.87	D	0.83
Y100C	D	0.87	D	0.87	D	0.96	D	1.10	D	0.98
Y100F	D	0.76	D	0.82	D	0.93	D	1.08	D	0.93
M104T	D	0.65	D	0.87	D	0.86	D	0.93	D	0.93
D105N	D	0.87	D	0.87	D	0.89	D	1.01	D	0.87
V107E	D	0.87	D	0.82	D	0.76	D	0.67	D	0.92
P108T	D	0.72	D	0.87	D	0.86	D	0.92	D	0.88
D110E	D	0.87	N	0.65	D	0.83	D	0.76	D	0.72
D110E	D	0.87	N	0.65	D	0.83	D	0.76	D	0.72
D110N	D	0.87	D	0.87	D	0.86	D	0.91	D	0.77
R113I	D	0.87	D	0.82	D	0.89	D	1.00	D	0.95
Y114H	D	0.87	D	0.87	D	0.85	D	0.89	D	0.97
W121R	D	0.87	D	0.87	D	0.97	D	1.08	D	0.98
G133C	N	0.63	D	0.87	D	0.76	D	0.60	D	0.68
V137A	D	0.51	D	0.87	D	0.79	D	0.74	D	0.89
D140E	N	0.65	D	0.87	D	0.79	D	0.65	D	0.74
P142S	D	0.87	D	0.87	D	0.84	D	0.90	D	0.76
A143T	D	0.55	D	0.87	D	0.81	D	0.73	D	0.59
G145W	D	0.87	D	0.87	D	0.98	D	1.04	D	0.89
V153D	D	0.87	D	0.82	D	0.89	D	1.00	D	0.95
L160F	D	0.87	D	0.87	D	0.89	D	1.02	D	0.88
H170L	D	0.72	N	0.63	D	0.78	D	0.75	D	0.96

N174D	D	0.87	D	0.87	D	0.87	D	0.99	D	0.81
V186G	D	0.87	D	0.87	D	0.87	D	0.93	D	0.94
V186M	D	0.61	D	0.87	D	0.86	D	0.86	D	0.85
D189Y	D	0.87	D	0.87	D	0.84	D	0.80	D	0.68
H204D	D	0.87	D	0.87	D	0.80	D	0.69	D	0.88
V205F	D	0.72	D	0.87	D	0.80	D	0.70	D	0.75
E208D	D	0.51	D	0.87	D	0.84	D	0.80	D	0.71
A213T	D	0.55	D	0.87	D	0.85	D	0.71	D	0.72
I222N	D	0.87	D	0.87	D	0.89	D	0.94	D	0.90
F232V	D	0.87	D	0.82	D	0.92	D	1.08	D	0.95
S261C	D	0.55	D	0.82	D	0.71	D	0.54	D	0.88
C323Y	D	0.72	D	0.87	D	0.85	D	0.64	T	0.47
P337H	D	0.61	D	0.87	T	0.20	T	-0.66	T	0.20
R375W	D	0.76	D	0.87	T	0.33	T	-0.32	T	0.38
W401C	D	0.72	D	0.82	T	0.33	T	-0.36	T	0.44

Table 3: Prediction of amino acid substitution and disease relation by metaservers

D=Deleterious/Damaging/Disease, T=Tolerated & N=Neutral

A5. Prediction of Conserved Amino Acid Residues by ConSurf, and Surface & Solvent Accessibility of Residues and TBX5 Secondary Structure by NetSurfP-2.0: ConSurf provides position-specific amino acid conservation scores. It also ranks amino acid residues in a color-coded scale of integers ranging from 1 to 9 where 1 indicates a highly variable and 9 indicates the most conserved residue. NetSurfP-2.0 predicts about surface accessibility (exposed or buried) of amino acids in a given protein, and provides relative and absolute accessible surface area of each residue. It also predicts protein secondary structure. Table 4 summarizes ConSurf and NetSurfP-2.0 outcomes.

		ConSurf			**NetSurfP-2.0**		
Position	**AA**	**Normalized Conservation Score**	**COLOR**	**RESIDUE VARIETY**	**Assigned Class**	**RSA**	**ASA**
58	L	-0.784	8	T,L,M,H	Buried	0.068	12.397
72	T	-1.098	9	T,S	Buried	0.094	13.02
85	P	-0.917	9	G,V,P	Buried	0.094	13.32
91	V	-0.972	9	L,X,V,C,M	Buried	0.062	9.596
93	G	-0.889	8	K,G,R,X	Exposed	0.526	41.429
96	P	-0.715	8	S,A,K,P,E	Exposed	0.369	52.33

100	Y	-1.095	9	Y	Buried	0.068	14.536
104	M	-0.334	6	M,V,I,A,T	Buried	0.009	1.773
105	D	-1.138	9	D	Buried	0.174	25.002
107	V	-1.025	9	I,V	Buried	0.163	25.004
108	P	-0.651	8	A,T,S,Q,P	Buried	0.218	31.002
110	D	-1.138	9	D	Exposed	0.347	50.019
113	R	-1.137	9	R	Exposed	0.277	63.479
114	Y	-1.095	9	Y	Buried	0.064	13.575
121	W	-0.647	8	W,R,C	Buried	0.249	59.799
133	G	-0.794	8	P,R,G,K	Exposed	0.507	39.935
137	V	-0.981	9	P,I,V,L	Exposed	0.292	44.873
140	D	-0.944	9	A,S,D	Exposed	0.559	80.603
142	P	-1.116	9	P	Exposed	0.297	42.08
143	A	-0.875	8	P,V,G,S,A	Buried	0.223	24.613
145	G	-0.901	9	D,G,T	Buried	0.042	3.271
153	V	-0.761	8	I,V,L	Buried	0.041	6.232
160	L	-0.81	8	R,P,T,L	Buried	0.034	6.174
170	H	-0.889	8	H,F,S,Q	Buried	0.165	29.977
174	N	-0.884	8	A,G,S,N,R	Exposed	0.279	40.854
186	V	-0.825	8	A,L,S,M,V,P	Buried	0.046	7.095
189	D	-0.802	8	H,P,A,S,D,T	Exposed	0.44	63.434
204	H	-0.657	8	H,R,F,Q,Y,A	Buried	0.141	25.641
205	V	-0.275	6	W,A,S,L,V,M,I	Exposed	0.44	67.687
208	E	-0.846	8	L,W,P,V,E	Exposed	0.283	49.385
213	A	-0.689	8	G,T,S,A,P,M	Buried	0.028	3.079
222	I	-0.944	9	I,V,N,W	Buried	0.044	8.056
232	F	-0.808	8	N,Y,F,L	Exposed	0.4	80.298
261	S	-0.451	7	N,D,S,T	Exposed	0.424	49.711
323	C	-0.522	7	C,R,Y,L,S,G	Buried	0.247	34.65
337	P	1.6	1	G,T,D,L,S,Q,E,P,A,H	Exposed	0.578	81.958
375	R	-0.748	8	S,Q,G,T,R	Exposed	0.413	94.575
401	W	-1.02	9	W	Exposed	0.309	74.289

Table 4: Prediction by ConSurf and NetSurf P Webtools

RSA=Relative Surface Accessibility & ASA=Absolute Surface Accessibility

ConSurf Results

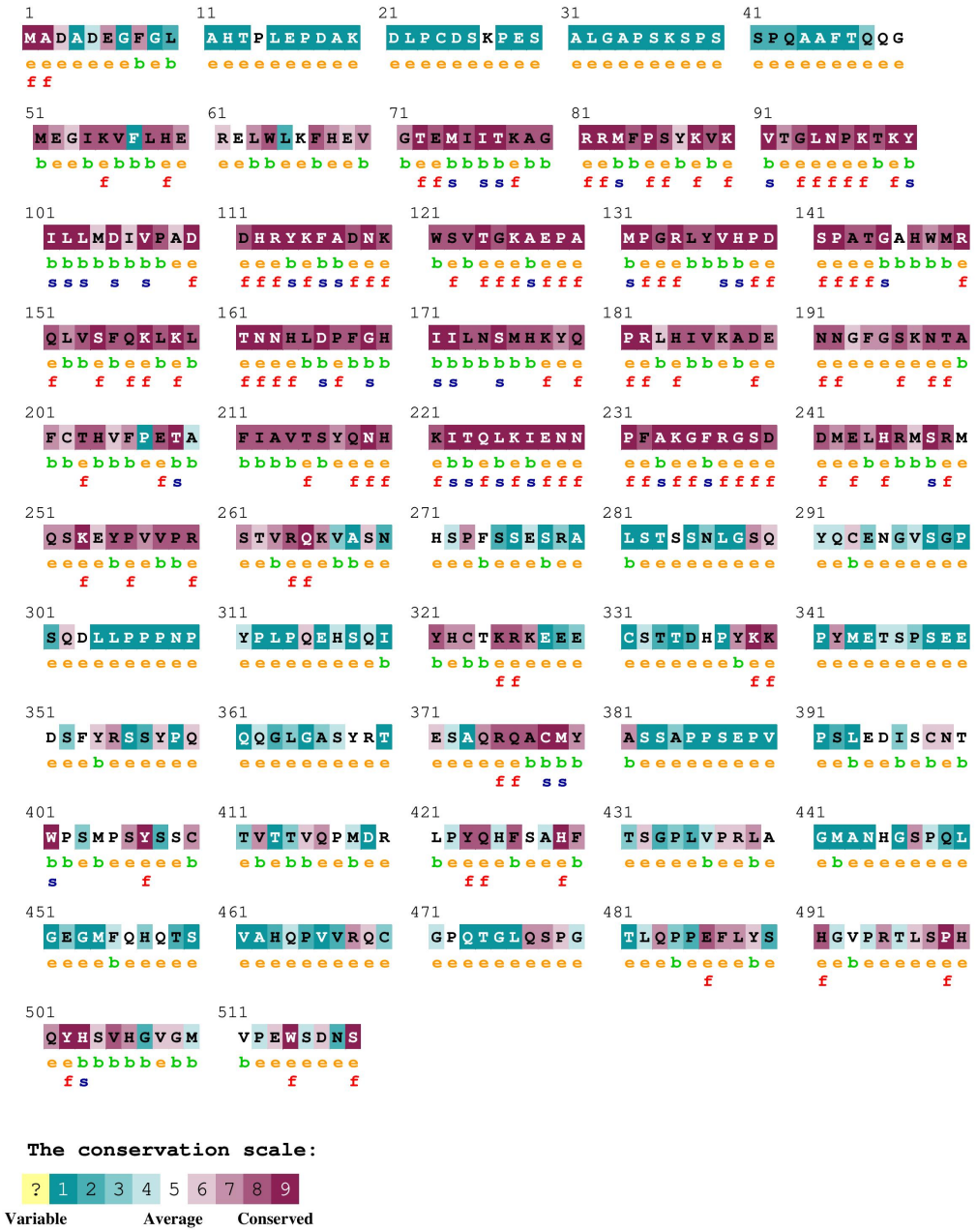


Figure 3:

Detection of Conserved Amino Acids by ConSurf.

e= An exposed residue according to the neural-network algorithm

b= A buried residue according to the neural-network algorithm

f= A predicted functional residue (highly conserved and exposed)

s= A predicted structural residue (highly conserved and buried)

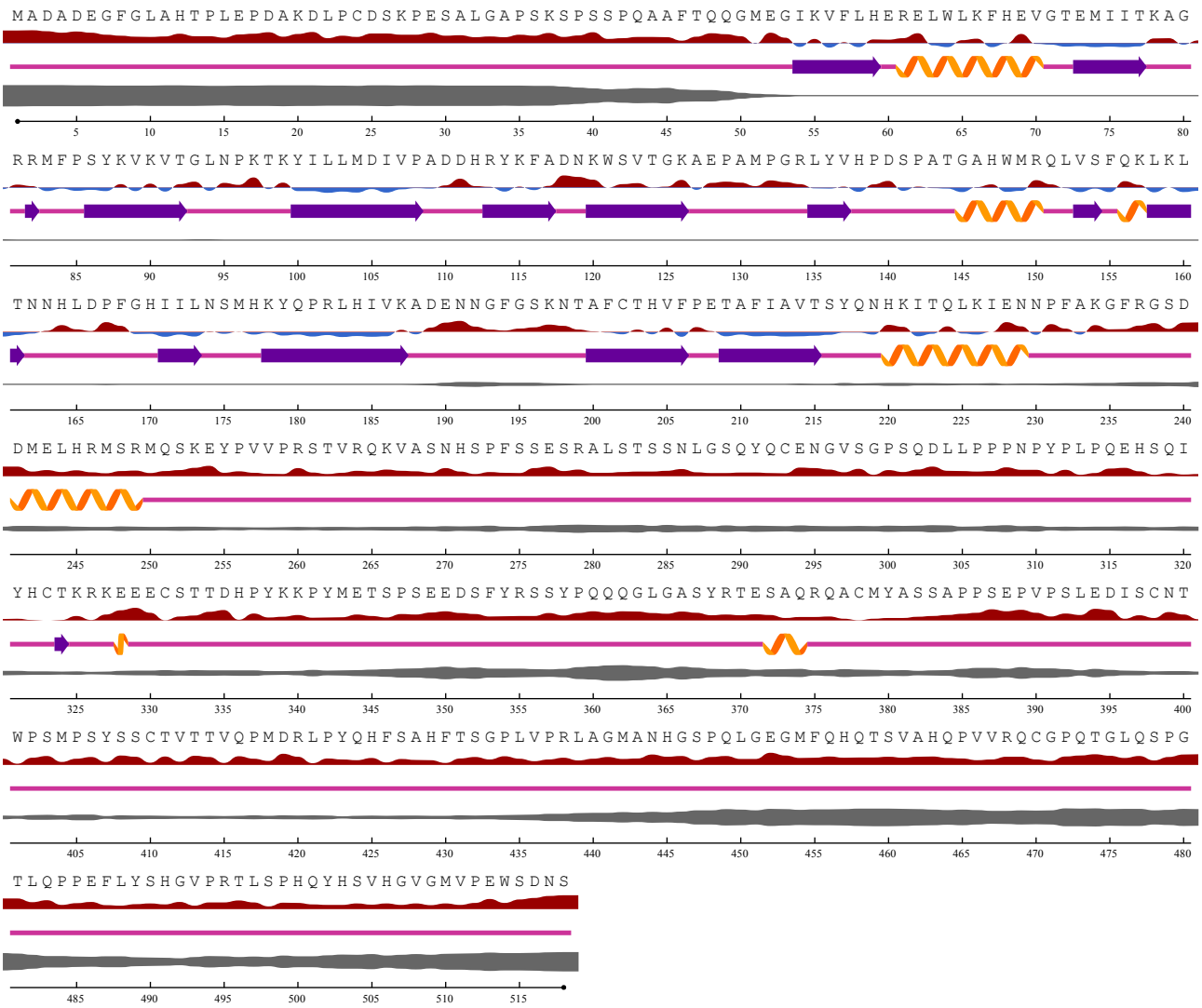


Figure 4: TBX5 Secondary Structure Prediction by NetSurfP-2.0

Relative Surface Accessibility: Red upward elevation is exposed residue and sky blue downward elevation is buried residue, thresholded at 25%.

Secondary Structure: orange spiral= helix, indigo arrow= strand, and pink straight line= coil.

Disorder: black swollen line; thickness of line equals probability of disordered residue.

A7. Prediction of Stability of the Mutated Protein due to SNPs by iStable 2.0: The amino acid substitutions were next subjected to analysis by protein stability prediction webtool iStable 2.0. This webtool consists of 11 sequence- and structure-based prediction tools, and provides decision after integrating all outputs by machine learning approach. For substitutions upto F232V, iStable 2.0 was

run from structure analysis (PDB ID:5BQD, Chain A) and due to unavailability of experimental structure, rest mutations were run from sequence analysis. Table 4 provides iStable 2.0 predictions.

AAS	Confidence Score	Stability
L58P	-1.7689416	Decrease
T72K	-1.0673743	Decrease
T72M	-0.067792356	Increase
P85L	-0.40284556	Increase
V91A	-2.363688	Decrease
G93S	-1.005738	Decrease
G93D	-1.0917563	Decrease
G93V	-1.1831716	Increase
P96S	-1.0492245	Decrease
Y100C	-1.5898333	Decrease
Y100F	-1.0702753	Decrease
M104T	-3.143975	Decrease
D105N	-0.9833634	Decrease
V107E	-1.7875004	Decrease
P108T	-0.82126737	Decrease
D110E	-0.68696284	Decrease
D110E	0.31303716	Decrease
D110N	-0.63278186	Decrease
R113I	-0.0076336265	Increase
Y114H	-2.133617	Decrease
W121R	-1.7060409	Decrease
G133C	-0.7795974	Decrease
V137A	-1.5604656	Decrease
D140E	-0.47431314	Increase
P142S	-1.0563802	Decrease
A143T	-1.4433032	Decrease
G145W	-0.910699	Increase
V153D	-2.2433639	Decrease
L160F	-1.6011777	Decrease
H170L	0.340119	Increase
N174D	-0.6325445	Decrease
V186G	-2.7898633	Decrease
V186M	-0.37593216	Decrease
D189Y	-1.0340729	Decrease
H204D	-1.2200044	Decrease
V205F	-1.4460453	Decrease
E208D	-1.1203609	Decrease
A213T	-1.3542162	Decrease
I222N	-2.3990505	Decrease
F232V	-0.04029095	Decrease

S261C	-0.1880107	Increase
C323Y	-0.45005268	Increase
P337H	-0.93664145	Decrease
R375W	-0.3665071	Increase
W401C	-0.64045143	Decrease

Table 5: Stability Prediction of the Mutated Proteins by iStable 2.0

A8. Prediction of the Effects of Amino Acid Substitutions on the Structural Integrity of Protein by Missense3D: The amino acid substitutions were further analyzed by Missense3D. For the wild type protein structure, it takes input both experimental and predicted 3D structure. Upto F232V substitution, experimental structure (PDB ID:5BQD, Chain A) was used and for the rest of the substitutions, the TBX5 structure predicted by Robetta and energy minimized by YASARA Energy Minimization Server was used. 12 structures were predicted to be 'Damaging'. Table 4 summarizes the output from Missense3D.

AAS	Missense3D prediction	Comment
L58P	Neutral	
T72K	Damaging	Buried charge introduced
T72M	Neutral	
P85L	Damaging	Cis proline replaced
V91A	Neutral	
G93S	Neutral	
G93D	Neutral	
G93V	Neutral	
P96S	Neutral	
Y100C	Neutral	
Y100F	Neutral	
M104T	Neutral	
D105N	Damaging	Buried salt bridge breakage
V107E	Neutral	
P108T	Neutral	
D110E	Neutral	
D110E	Neutral	
D110N	Neutral	
R113I	Neutral	

Y114H	Neutral	
W121R	Damaging	Cavity altered
G133C	Neutral	
V137A	Neutral	
D140E	Neutral	
P142S	Damaging	Cis proline replaced
A143T	Neutral	
G145W	Damaging	Clash alert, Buried Gly replaced
V153D	Damaging	Buried hydropilic residue introduced & Buried charge introduced
L160F	Neutral	
H170L	Neutral	
N174D	Neutral	
V186G	Neutral	
V186M	Neutral	
D189Y	Neutral	
H204D	Neutral	
V205F	Neutral	
E208D	Damaging	Buried H-bond breakage
A213T	Neutral	
I222N	Damaging	Buried hydropilic residue introduced
F232V	Neutral	
S261C	Damaging	Cavity altered (expansion of cavity volume by 113.616 Å^3)
C323Y	Damaging	Clash alert
P337H	Neutral	
R375W	Neutral	
W401C	Damaging	Cavity altered (expansion of cavity volume by 103.68 Å^3)

Table 6: Structural Damage Prediction due to SNPs by Missense3D

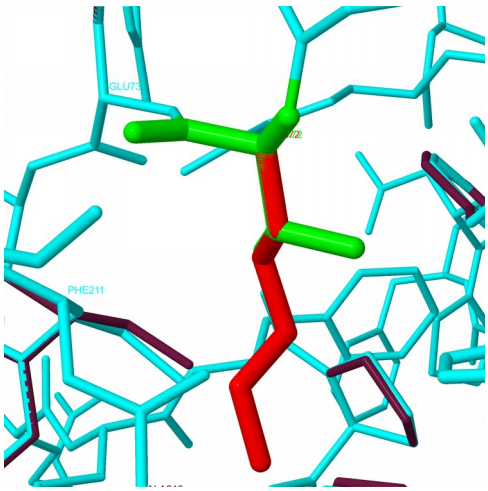


Figure 5.1: Side chain replacement in T72K

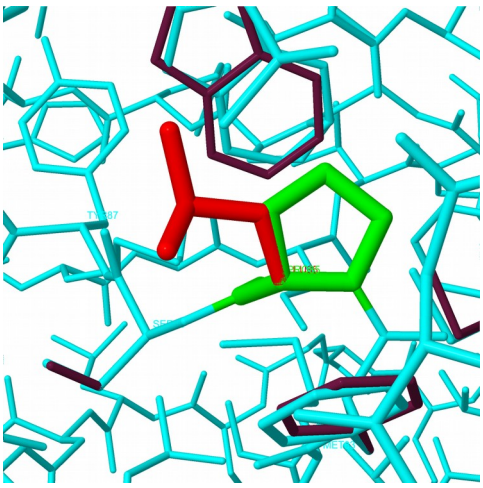


Figure 5.2: Side chain replacement in P85L

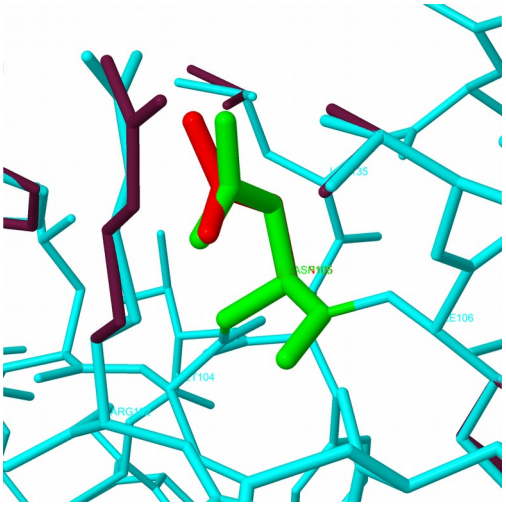


Figure 5.3: Side chain replacement in D105N

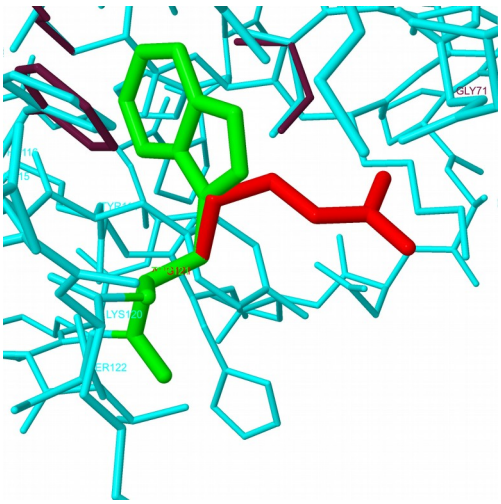


Figure 5.4: Side chain replacement in W121R

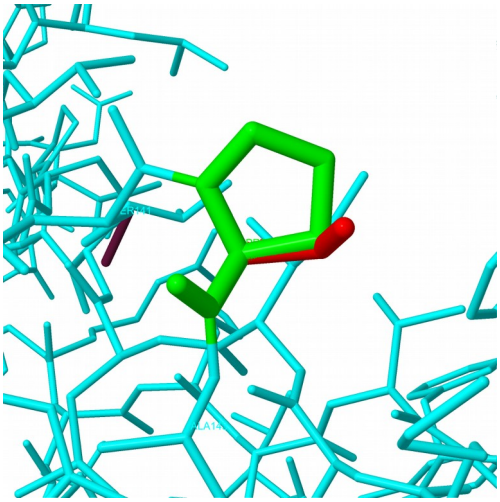


Figure 5.5: Side chain replacement in P142S

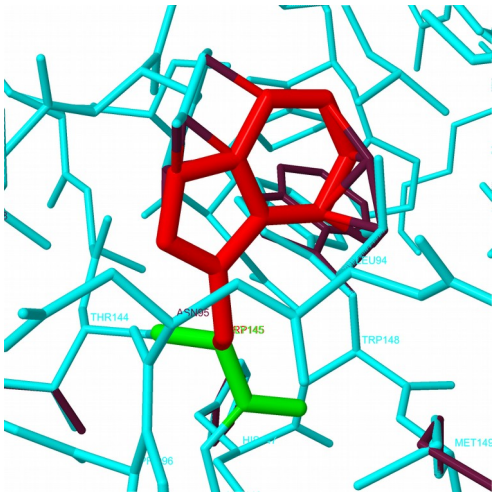


Figure 5.6: Side chain replacement in G145W

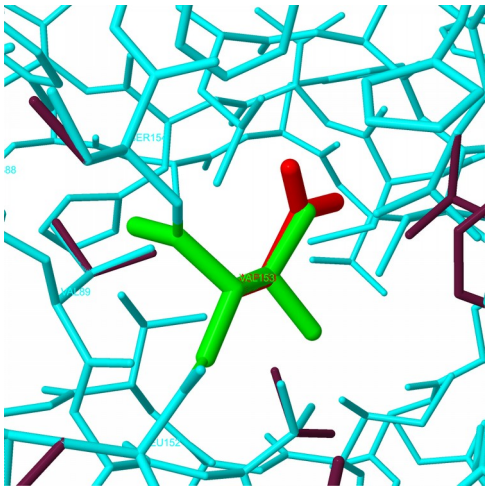


Figure 5.7: Side chain replacement in V153D

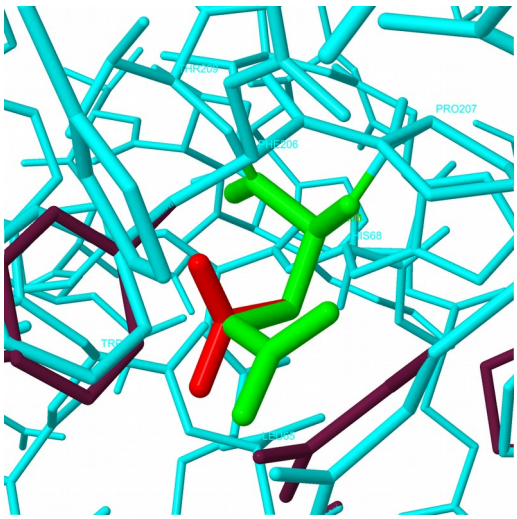


Figure 5.8: Side chain replacement in E208D

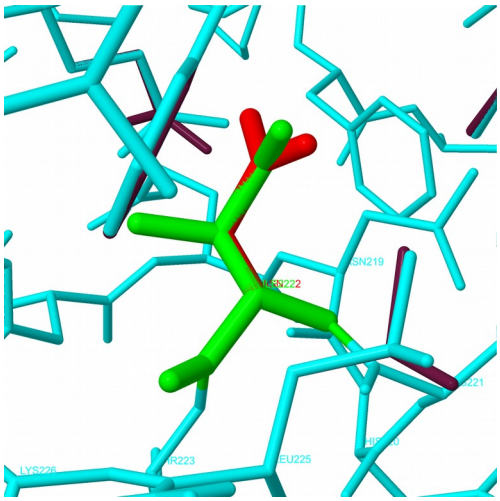


Figure 5.9: Side chain replacement in I222N

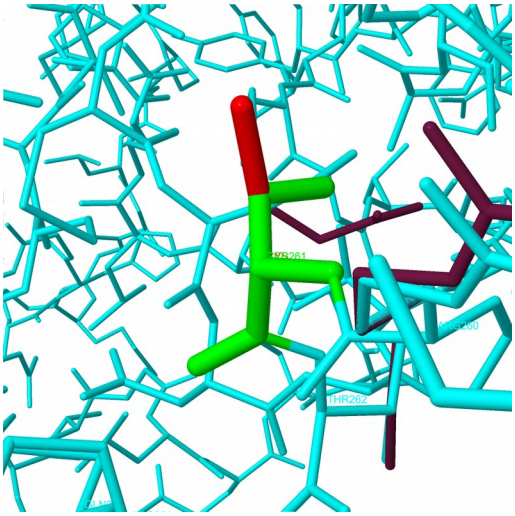


Figure 5.10: Side chain replacement in S261C

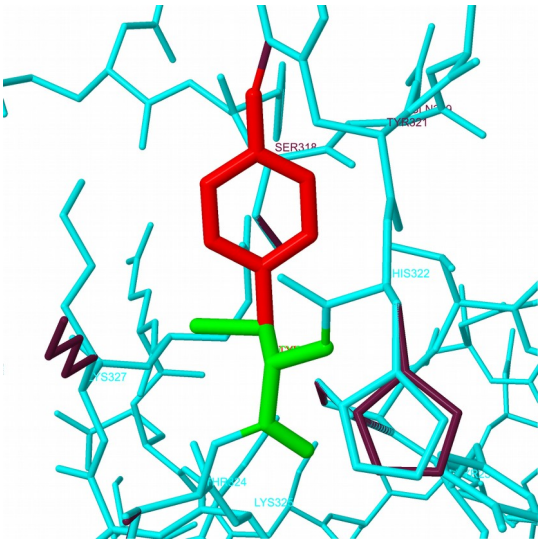


Figure 5.11: Side chain replacement in C323Y

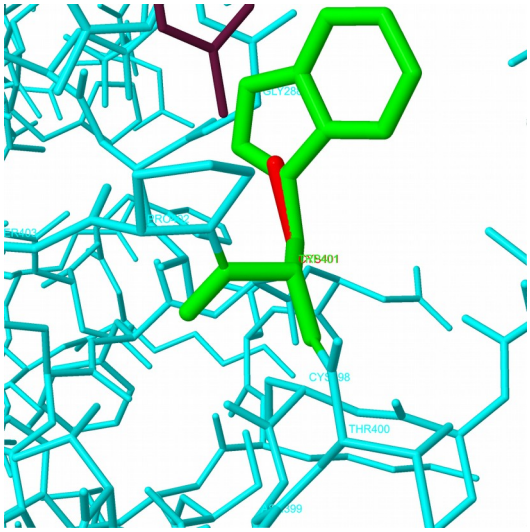


Figure 5.12: Side chain replacement in W401C

Figure 5.1-5.12 are close-up views of the 12 damaging mutations predicted by Missense3D. The wild type chain color is cyan and the mutant chain color is rosewood. Wild type residues are colored in green and the mutant residues are colored in red. These images were generated by Missense3D server.

B. Identification and Analysis of Nonsense and Stop Lost SNPs by PredictSNP2: 2 stop gained (Nonsense) and 1 stop lost SNPs were analyzed by PredictSNP2. It predicted the nonsense mutations as deleterious and the stop lost mutation as neutral.

Variant	PredictSNP2		CADD		DANN		FATHMM		FunSeq2		GWAVA	
	Prediction	Score	Prediction	Score	Prediction	Score	Prediction	Score	Prediction	Score	Prediction	Score
W514Ter	Del	0.6777	Del	46	Del	0.9959	Del	0.9623	Del	5	?	0.39
Q156Ter	Del	1	Del	38	Del	0.9984	Del	0.9167	Del	5	Del	0.5
Ter519Q	Neu	-0.1146	Neu	14.06	Neu	0.758	Del	0.9717	Neu	0	Del	0.31

Table 7: Prediction about Nonsense and Stop Lost SNPs by PredictSNP2
Del= Deleterious & Neu= Neutral

C. Prediction of the effects of UTR SNPs by UTRdb & MirSNP: UTR SNPs that could have an effect on transcriptional motifs were retrieved from the UTRdb server. Three SNPs in the 3′ UTR- rs28730760, rs28730761 and rs883079 were found to be present in polyadenylation (poly A) sites, and hence may result in pathogenicity. UTRdb output returns 6 transcriptional motif matches in 5′ UTR and 3 matches in the 3′ UTR. Table 8 provides information collected from the UTRdb server. 3′ UTR SNPs that might create or break a miRNA target site, or enhance or decrease binding of miRNA to mRNA were predicted by MirSNP server. A total number of 86 SNPs in the 3′ untranslated region of the TBX5 gene were predicted to alter miRNA-mRNA binding. Table 9 shows results from MirSNP server.

Signal Name	UTR Region	Total Match	NCBI Reference Sequence	Genomic Position (NCBI36/hg18)
uORF (Upstream Open Reading Frame)	5'	1	NM_181486.4	Chr12:113328041-113328106
		2	NM_000192.3	Chr12:113330378-113330629 Chr12:113330136-113330216
		2	NM_080717.3	Chr12:113330378-113330629 Chr12:113330136-113330216
IRES (Internal Ribosome Entry Site)	5'	1	NM_000192.3	Chr12:113326087-113330055
CPE (Cytoplasmic polyadenylation element)	3'	1	NM_181486.4	Chr12:113276118-113276332
K-Box	3'	1	NM_181486.5	Chr12:113276848-113276855
PAS (Polyadenylation Signal)	3'	1	NM_181486.6	Chr12:113276118-113276144

Table 8: Transcriptional Motifs Present in TBX5 Untranslated Regions

	Gene	miRNA	SNP	Effect	Allele	Score	Energy	Conservation
1	TBX5	hsa-miR-1256	rs184793497	break	C A	140	-9.41	0.006
2	TBX5	hsa-miR-1294	rs76633287	break	T C	156	-20.21	0.052
3	TBX5	hsa-miR-130a-5p	rs181130921	decrease	G C	146 145	-11.65 -7.2	0.499 0.499
4	TBX5	hsa-miR-130b-5p	rs10850326	decrease	G A	158 148	-16.41 -11.31	0.004 0.007
5	TBX5	hsa-miR-130b-5p	rs112808902	decrease	- AAGAGA	158 148	-16.41 -11.31	0 0.002
6	TBX5	hsa-miR-1323	rs140532076	enhance	T C	151 155	-13 -14.92	0 0
7	TBX5	hsa-miR-141-3p	rs112051831	break	G A	154	-14.34	0
8	TBX5	hsa-miR-144-5p	rs188337192	break	T C	155	-15.19	0

9	TBX5	hsa-miR-144-5p	rs192445542	break	T	155	-15.19	0
					C			
10	TBX5	hsa-miR-147a	rs184793497	create	C			
					A	149	-14.29	0.004
11	TBX5	hsa-miR-200a-3p	rs112051831	break	G	153	-12.15	0
					A			
12	TBX5	hsa-miR-204-5p	rs10850326	break	G	147	-16.09	0.002
					A			
13	TBX5	hsa-miR-21-3p	rs12426660	enhance	T	150	-17.2	0
					C	154	-18.97	0
14	TBX5	hsa-miR-211-3p	rs12426660	break	T	150	-19.42	0.002
					C			
15	TBX5	hsa-miR-211-5p	rs10850326	break	G	147	-14.74	0.002
					A			
16	TBX5	hsa-miR-2355-3p	rs76694710	create	C			
					A	160	-17.09	0.052
17	TBX5	hsa-miR-23a-3p	rs181130921	decrease	G	147	-9.06	0.499
					C	144	-8.22	0.499
18	TBX5	hsa-miR-2467-3p	rs28730760	create	A			
					C	143	-16	0.005
19	TBX5	hsa-miR-29b-1-5p	rs181130921	break	G	159	-23.67	0.237
					C			
20	TBX5	hsa-miR-3119	rs184360838	create	G			
					A	140	-10.91	0.023
21	TBX5	hsa-miR-3134	rs2384410	break	T	140	-11.18	0.02
					C			
22	TBX5	hsa-miR-3148	rs75334514	decrease	T	146	-8.63	0.004
					A	145	-7.44	0.004
23	TBX5	hsa-miR-3148	rs192527148	decrease	T	148	-7.26	0.423
					G	146	-8.63	0.423
24	TBX5	hsa-miR-3154	rs143511878	break	T	163	-27.56	0.853
					C			
25	TBX5	hsa-miR-3160-5p	rs184360838	break	G	169	-28.28	0.033
					A			
26	TBX5	hsa-miR-3170	rs76347803	create	G			
					C	150	-16.34	0.019
27	TBX5	hsa-miR-3170	rs76633287	create	T			
					C	150	-16.34	0.015
28	TBX5	hsa-miR-3170	rs78828439	create	C			
					A	150	-16.34	0.02
29	TBX5	hsa-miR-3180-5p	rs143511878	break	T	146	-19.83	0.298
					C			
30	TBX5	hsa-miR-3183	rs10850326	enhance	G	145	-21.31	0.018
					A	147	-21.26	0.018
31	TBX5	hsa-miR-3183	rs112808902	enhance	-	146	-19	0.042
					AAGAGA	147	-21.26	0

32	TBX5	hsa-miR-3189-3p	rs76694710	create	C			
					A	151	-19.19	0.018
33	TBX5	hsa-miR-3202	rs143511878	create	T			
					C	151	-19.87	0.853
34	TBX5	hsa-miR-330-3p	rs143511878	decrease	T	145	-15.27	0.878
					C	144	-14.11	0.878
35	TBX5	hsa-miR-340-3p	rs28730761	create	A			
					G	148	-16.52	0.254
36	TBX5	hsa-miR-3613-3p	rs147774482	decrease	T	164	-10.03	0.999
					C	156	-10.05	0.999
37	TBX5	hsa-miR-3678-3p	rs28730760	create	A			
					C	146	-15.86	0.005
38	TBX5	hsa-miR-374c-3p	rs183595902	break	T	149	-17.08	0.139
					G			
39	TBX5	hsa-miR-4257	rs28730760	create	A			
					C	156	-18.19	0.005
40	TBX5	hsa-miR-4282	rs75334514	create	T			
					A	140	-2.08	0
41	TBX5	hsa-miR-4286	rs183595902	break	T	144	-15.71	0.466
					G			
42	TBX5	hsa-miR-4287	rs10850326	break	G	140	-18.5	0.001
					A			
43	TBX5	hsa-miR-4311	rs187379531	decrease	G	148	-9.94	0.001
					C	146	-10.85	0.001
44	TBX5	hsa-miR-4427	rs117414057	break	T	150	-10.75	0.004
					C			
45	TBX5	hsa-miR-4446-3p	rs117414057	enhance	T	157	-20.94	0.006
					C	161	-22.89	0.006
46	TBX5	hsa-miR-4455	rs2384410	create	T			
					C	148	-16.7	0.002
47	TBX5	hsa-miR-4635	rs76799455	decrease	T	165	-25.41	0.005
					A	157	-23.67	0.005
48	TBX5	hsa-miR-4647	rs117414057	enhance	T	161	-20.78	0.095
					C	165	-23.09	0.095
49	TBX5	hsa-miR-4656	rs117414057	enhance	T	159	-26.31	0.006
					C	163	-27.88	0.006
50	TBX5	hsa-miR-4685-3p	rs10850326	break	G	147	-17.89	0.001
					A			
51	TBX5	hsa-miR-4688	rs12426660	create	T			
					C	145	-21.04	0.005
52	TBX5	hsa-miR-4697-3p	rs148346089	create	T			
					C	160	-28.56	0.024
53	TBX5	hsa-miR-4711-5p	rs148346089	break	T	143	-19.27	0
					C			
54	TBX5	hsa-miR-4717-3p	rs116382074	break	G	148	-15.92	0.03
					A			

55	TBX5	hsa-miR-4723-3p	rs10850326	decrease	G	155	-22.13	0.018
					A	150	-20.84	0.018
56	TBX5	hsa-miR-4724-5p	rs140532076	break	T	145	-13.87	0
					C			
57	TBX5	hsa-miR-4753-3p	rs10850326	enhance	G	154	-16.7	0.004
					A	166	-22.9	0.002
58	TBX5	hsa-miR-4753-3p	rs112808902	enhance	-	154	-16.7	0
					AAGAGA	166	-22.9	0.001
59	TBX5	hsa-miR-4764-5p	rs2384410	break	T	156	-20.66	0.002
					C			
60	TBX5	hsa-miR-4771	rs12426660	create	T			
					C	160	-17.33	0.23
61	TBX5	hsa-miR-486-3p	rs12426660	create	T			
					C	150	-20.67	0.01
62	TBX5	hsa-miR-494	rs116382074	decrease	G	157	-15.31	0.697
					A	149	-10.83	0.697
63	TBX5	hsa-miR-499a-3p	rs116382074	break	G	153	-21.11	0.271
					A			
64	TBX5	hsa-miR-499b-3p	rs116382074	break	G	146	-12.79	0.271
					A			
65	TBX5	hsa-miR-5006-5p	rs148346089	decrease	T	151	-22.42	0.001
					C	146	-23.1	0.001
66	TBX5	hsa-miR-5007-3p	rs116382074	create	G			
					A	140	-8.37	0.066
67	TBX5	hsa-miR-5010-3p	rs184793497	decrease	C	150	-9.37	0
					A	140	-5.98	0.007
68	TBX5	hsa-miR-5093	rs78058633	create	G			
					A	149	-14.21	0
69	TBX5	hsa-miR-5094	rs148346089	break	T	152	-16.17	0.024
					C			
70	TBX5	hsa-miR-513a-3p	rs75334514	create	T			
					A	153	-8.8	0
71	TBX5	hsa-miR-513c-3p	rs75334514	create	T			
					A	153	-5.75	0
72	TBX5	hsa-miR-5197-5p	rs187379531	create	G			
					C	152	-15.54	0.002
73	TBX5	hsa-miR-545-5p	rs141223729	enhance	T	157	-12.28	1
					C	161	-14.27	1
74	TBX5	hsa-miR-548a-3p	rs140532076	enhance	T	147	-13.58	0
					C	151	-15.87	0
75	TBX5	hsa-miR-548an	rs147774482	create	T			
					C	146	-12.7	0.909
76	TBX5	hsa-miR-548e	rs140532076	enhance	T	154	-9.62	0
					C	158	-11.97	0
77	TBX5	hsa-miR-548f	rs140532076	decrease	T	145	-7.01	0
					C	142	-6.35	0

78	TBX5	hsa-miR-548o-3p	rs140532076	enhance	T	149	-9.53	0
					C	150	-11.15	0
79	TBX5	hsa-miR-5572	rs76694710	create	C			
					A	143	-19.81	0.028
80	TBX5	hsa-miR-5584-5p	rs76799455	break	T	154	-13.2	0.003
					A			
81	TBX5	hsa-miR-576-3p	rs28730760	break	A	159	-18.42	0.003
					C			
82	TBX5	hsa-miR-576-3p	rs112051831	enhance	G	155	-18.63	0
					A	159	-18.42	0
83	TBX5	hsa-miR-599	rs6489956	break	G	145	-9.33	0.001
					A			
84	TBX5	hsa-miR-616-3p	rs188337192	create	T			
					C	157	-15.51	0.001
85	TBX5	hsa-miR-624-5p	rs141223729	enhance	T	157	-19.6	1
					C	161	-21.59	1
86	TBX5	hsa-miR-944	rs192527148	break	T	158	-8.44	0
					G			

Table 9: Functional SNPs in the 3′ UTR predicted by MirSNP

D. Predicting the effects of SNPs located in splice sites by HSF: Human Splicing Finder (HSF) webtool was employed to predict the effects of SNPs located in 5′ and 3′ splice sites. A total number of 11 splice site SNPs were subjected to assessment. All were found to have ‘probably no impact on splicing’ by HSF. Table 10 shows the predictions by HSF.

rsIDs	Alleles	Predicted Signal	Interpretation
	Splice Acceptor Variant		
rs1555226330	C>A	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs1031873727	G>A	ESS Site broken	Alteration of an intronic ESS site,Probably no impact on splicing
rs1031873727	G>T	ESS Site broken	Alteration of an intronic ESS site,Probably no impact on splicing
rs891464399	G>A	New ESE Site	Creation of an intronic ESE site,Probably no impact on splicing

	Splice Donor Variant		
rs958951320	C>T	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs181078973	A>C	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs181078973	A>G	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs374919751	C>G	New ESE Site	Creation of an intronic ESE site,Probably no impact on splicing
rs374919751	C>T	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs1288475235	C>A	No significant splicing motif alteration detected	This mutation has probably no impact on splicing
rs1015550731	T>C	ESS Site broken	Alteration of an intronic ESS site,Probably no impact on splicing

Table 10: Prediction about Splice Site SNPs by HSF webtool

3D Structure prediction for Full TBX5 Protein: Due to the unavailability of TBX5 full-length protein crystal structure in Protein Data Banks, full-length 3D models have been created. 3 web servers (I-TASSER, Phyre2 & Robetta) were employed to predict these models. The models generated by these servers were energy minimized by YASARA Energy Minimization Server. They were next refined by GalaxyRefine server. Figures 5, 6, and 7 are cartoon presentations of the energy minimized and refined models generated by I-TASSER, Phyre2, and Robetta respectively.

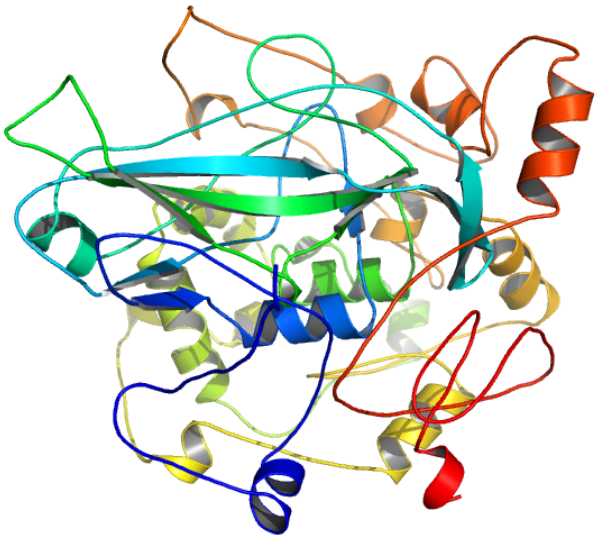


Figure 6: Whole TBX5 Model Predicted by I-TASSER (Energy Minimized and Refined)

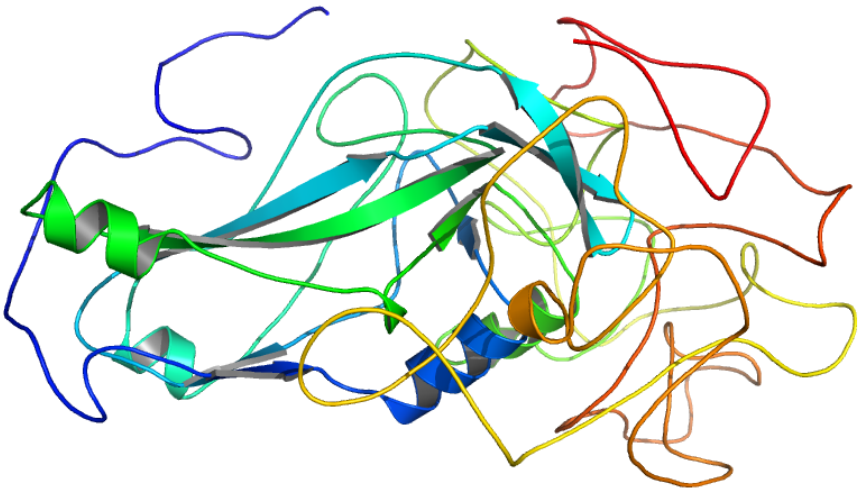


Figure 7: Whole TBX5 Model Predicted by Phyre2 (Energy Minimized and Refined)

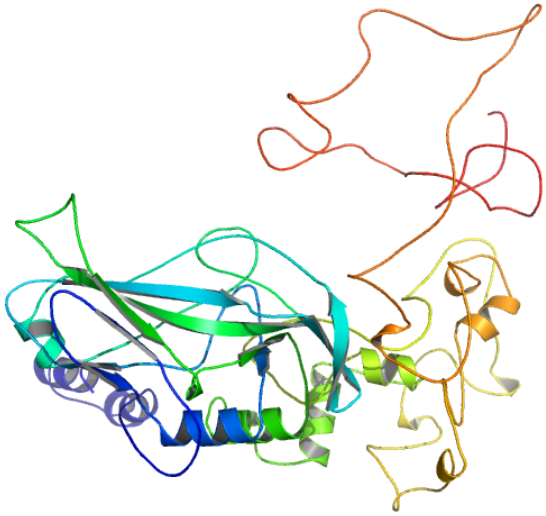


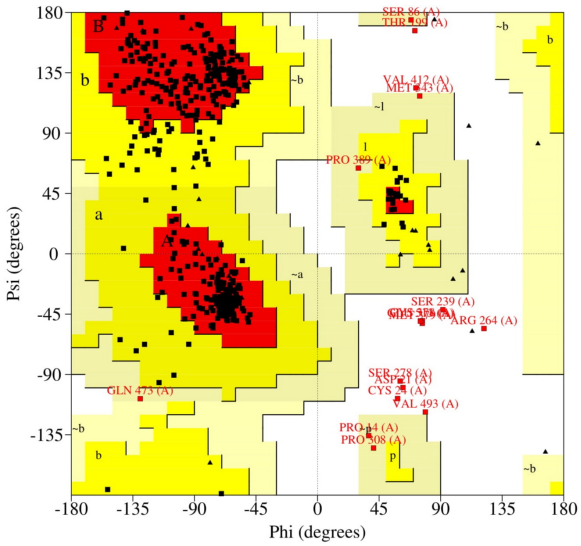
Figure 8: Whole TBX5 Model Predicted by Robetta (Energy Minimized and Refined)

Evaluation of TBX5 3D Models: 3D Model evaluation is indispensable for checking the credibility of the generated model. The models generated in the current study were assessed by PROCHECK, Verify 3D, and SWISS-MODEL Structure Assessment Tool. Ramachandran plot is a good qualitative indicator of protein structure [68]. Ramachandran plots for the created models were calculated by PROCHECK server. A model is considered good by PROCHECK if >90% residues are present in the most favored regions of the Ramachandran plot. Only one model (generated by Robetta and later energy minimized and refined) was found to fulfill this criteria. It has 94.5% of residues in the most favored regions. Verify 3D is another tool which approves a model if 80% or more amino acid residues have a score ≥0.2 in the 3D/1D profile. However, none of the models could fulfill this requirement. MolProbity score and QMEAN Z-score for our models were obtained from the SWISS-MODEL Structure Assessment Tool. A low MolProbity score and a high QMEAN Z-score are preferred. A negative QMEAN Z-score indicates instability in a protein structure [72 & SWISS-MODEL Structure Assessment Help Page]. All of our models have a negative QMEAN Z-

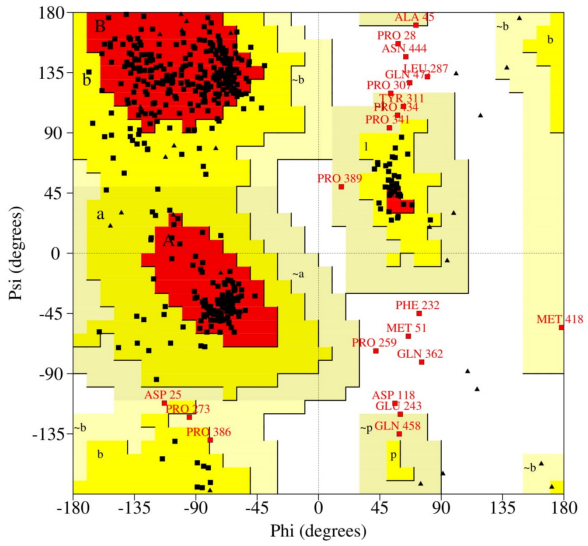
score. This disagreement can be explained by the fact that presently available *ab initio* protein modeling tools predict poorly for proteins having more than 120 residues [74] and for our protein models 279 residues (5BQD-A chain extends from 1-239 residues and the full-length protein has 518 residues) don't have any template. However, among the models, the energy minimized and refined Robetta model has the lowest MolProbity score and the least negative QMEAN Z-score. Taking into consideration Ramachandran plot results, MolProbity scores, and QMEAN Z-scores, it can be decided that Robetta outperformed two other protein structure prediction tools used in this study. Table 11 summarizes the results obtained from PROCHECK, Verify 3D, and SWISS-MODEL Structure Assessment Tool.

Model	**PROCHECK**				**Verify 3D (% residues with averaged 3D-1D score ≥0.2)**	**MolProbity Score**	**QMEAN Z-score**
	Number of residues in most favored regions	**Number of residues in additional allowed regions**	**Number of residues in generously allowed regions**	**Number of residues in disallowed regions**			
I-TASSER	364 (84.1%)	55 (12.7%)	3 (0.7%)	11 (2.5%)	69.69	1.03	-4.68
Phyre2	342 (79.0%)	78 (18.0%)	6 (1.4%)	7 (1.6%)	70.66	1.17	-7.05
Robetta	409 (94.5%)	23 (5.30%)	0 (0.0%)	1 (0.2%)	70.46	0.5	-1.37

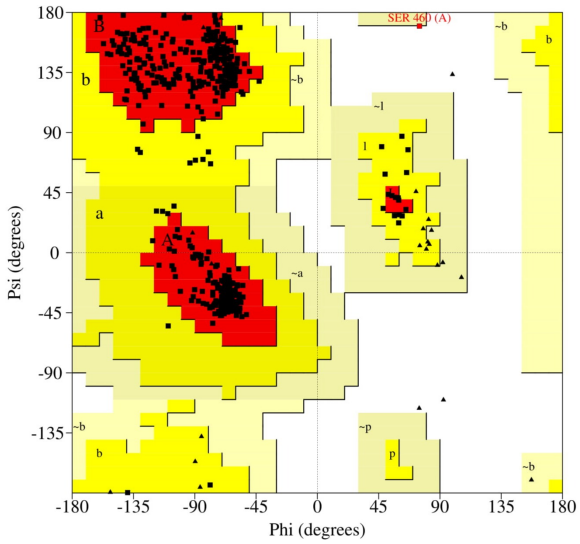
Table 11: Protein 3D Model Evaluation Results from PROCHECK, Verify 3D and SWISS-MODEL Structure Assessment Tool



A



B



C

Figure 9: A, B & C are Ramachandran plots generated by PROCHECK server for energy minimized and refined models produced by I-TASSER, Phyre2, and Robetta respectively.

Red= most favored regions

Yellow= additional allowed regions

Light Yellow= generously allowed regions

White= disallowed regions

Discussion:

In the last two decades, the realm of genetics has been enriched by generous contributions from bioinformatics. No wonder, a good number of quality tools have been developed and improved for pathogenic SNP identification and analysis. The differences between them lie in their algorithms and the datasets that have been used to train them. But no tool is 100% error-free. In our study, we appointed 9 highly dependable prediction tools to identify deleterious SNPs on the TBX5 gene and analyzed their structural and functional consequences with the help of other tools. The 45 deleterious SNPs achieved by concordance among PROVEAN, SIFT and PolyPhen-2 were also found to be damaging by others with a few exceptions. In our opinion, these 45 SNPs are the most important candidates for pathogenicity and among them the 12 SNPs predicted to be damaging by Missense3D deserve special attention.

Conclusion:

TBX5 gene is an important regulator of cardiac development and the gene responsible for Holt-Oram syndrome. In the years past, many mutations including SNPs have been reported in association with this gene. But most of the TBX5 gene SNPs have not been characterized yet. This extensive *in silico* study attempts to find TBX5 SNPs with the most pathogenic potential and is the first of its kind for TBX5 gene. The results obtained from this study could help in future in the study of TBX5 gene related diseases, prenatal diagnosis and counseling, and may potentiate finding cures.

Authors' Contribution: A M U B Mahfuz conceived of and designed this study, performed *in silico* analyses and wrote the manuscript. Md. Arif Khan revised, and made necessary corrections and additions to the manuscript.

Conflicts of Interest: The authors declare that they have no competing interests.

Funding sources: This work was fully self-funded.

References:

1. Papaioannou, V. E. (2014). The T-box gene family: emerging roles in development, stem cells and cancer. Development, 141(20), 3819–3833. DOI: 10.1242/dev.104471

2. Papaioannou, V. E. (2001). T-box genes in development: From hydra to humans. In International Review of Cytology (pp. 1–70). DOI: 10.1016/s0074-7696(01)07002-4

3. Packham, E. A. & Brook, J. D.,(2003). T-box genes in human disorders. Human Molecular Genetics, 12(90001), 37R–44. DOI: 10.1093/hmg/ddg077

4. Ghosh, T. K., Brook, J. D., & Wilsdon, A. (2017). T-box genes in human development and disease. In Current Topics in Developmental Biology (Vol. 122, pp. 383–415). Elsevier. DOI: 10.1016/bs.ctdb.2016.08.006

5. Ghosh, T. K., Packham, E. A., Bonser, A. J., Robinson ,T. E., Cross, S. J., & Brook, J.D. (2001). Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome. Human Molecular Genetics, 10(18), 1983–1994. DOI: 10.1093/hmg/10.18.1983

6. Hiroi, Y., Kudoh, S., Monzen, K., Ikeda, Y., Yazaki, Y., Nagai, R., & Komuro, I. (2001). Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. Nature Genetics, 28(3), 276–280. DOI: 10.1038/90123

7. Steimle, J. D., & Moskowitz, I. P. (2017). Tbx5. In Current Topics in Developmental Biology (Vol. 122, pp. 195–221). Elsevier. DOI: 10.1016/bs.ctdb.2016.08.008

8. Rallis, C., Bruneau, B. G., Del Buono, J., Seidman, C. E., Seidman, J. G., Nissim, S., et al. (2003). Tbx5 is required for forelimb bud formation and continued outgrowth. Development, 130(12), 2741–2751. DOI: 10.1242/dev.00473

9. Hasson, P., Del Buono, J., & Logan, M. P. O. (2007). Tbx5 is dispensable for forelimb outgrowth. Development, 134(1), 85–92. DOI: 10.1242/dev.02622

10. Običan, S., & Maggio, L. (2018). Holt-oram syndrome. In Obstetric Imaging: Fetal Diagnosis and Care (pp. 557-559.e1). Elsevier. DOI: 10.1016/B978-0-323-44548-1.00132-7

11. Boogerd, C. J. J., Dooijes, D., Ilgun, A., Mathijssen, I. B., Hordijk, R., van de Laar, I. M. B. H., Rump, P., Veenstra-Knol, H. E., Moorman, A. F. M., Barnett, P., & Postma, A. V. (2010). Functional analysis of novel TBX5 T-box mutations associated with Holt-Oram syndrome. Cardiovascular Research, 88(1), 130–139. DOI: 10.1093/cvr/cvq178

12. Vanlerberghe, C., Jourdain, A.-S., Ghoumid, J., Frenois, F., Mezel, A., Vaksmann, G., Lenne, B., Delobel, B., Porchet, N., Cormier-Daire, V., Smol, T., Escande, F., Manouvrier-Hanu, S., & Petit, F. (2018). Holt-Oram syndrome: clinical and molecular description of 78 patients with TBX5 variants. European Journal of Human Genetics, 27(3), 360–368. DOI: 10.1038/s41431-018-0303-3

13. Braulke, I., Herzog, S., Thies, U., & Zoll, B. (2008). Holt-Oram syndrome in four half-siblings with unaffected parents: brief clinical report. Clinical Genetics, 39(4), 241–244. DOI: 10.1111/j.1399-0004.1991.tb03021.x

14. Holt-Oram syndrome. The portal for rare diseases and orphan drugs website. https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Expert=392&lng=EN. Accessed April 16, 2020.

15. Barisic, I., Boban, L., Greenlees, R., Garne, E., Wellesley, D., Calzolari, E., et al. (2014). Holt Oram syndrome: a registry-based study in Europe. Orphanet Journal of Rare Diseases, 9(1). DOI: 10.1186/s13023-014-0156-y

16. Yi Li, Q., Newbury-Ecob, R. A., Terrett, J. A., Wilson, D. I., Curtis, A. R. J., Ho Yi, C., Gebuhr, T., Bullen, P. J., Robson, S. C., Strachan, T., Bonnet, D., Lyonnet, S., Young, I. D., Raeburn, J. A., Buckler, A. J., Law, D. J., & Brook, J. D. (1997). Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. Nature Genetics, 15(1), 21–29. DOI: 10.1038/ng0197-21

17. Yang, J., Hu, D., Xia, J., Yang, Y., Ying, B., Hu, J., & Zhou, X. (2000). Three novel TBX5 mutations in Chinese patients with Holt-Oram syndrome. American Journal of Medical Genetics, 92(4), 237–240. DOI: 10.1002/(SICI)1096-8628(20000605)92:4<237::AID-AJMG2>3.0.CO;2-G

18. Borozdin, W., Bravo-Ferrer Acosta, A. M., Seemanova, E., Leipoldt, M., Bamshad, M. J., Unger, S., et al. (2006). Contiguous hemizygous deletion ofTBX5,TBX3, andRBM19 resulting in a combined phenotype of Holt-Oram and ulnar-mammary syndromes. American Journal of Medical Genetics Part A, 140A(17), 1880–1886. DOI: 10.1002/ajmg.a.31340

19. Patel, C., Silcock, L., McMullan, D., Brueton, L., & Cox, H. (2012). TBX5 intragenic duplication: a family with an atypical Holt–Oram syndrome phenotype. European Journal of Human Genetics, 20(8), 863–869. DOI: 10.1038/ejhg.2012.16

20. Guo, Q., Shen, J., Liu, Y., Pu, T., Sun, K., & Chen, S. (2015). Exome sequencing identifies a c. 148-1G>C mutation of TBX5 in a holt-oram family with unusual genotype-phenotype correlations. Cellular Physiology and Biochemistry, 37(3), 1066–1074. DOI: 10.1159/000430232

21. Ríos-Serna, L. J., Díaz-Ordoñez, L., Candelo, E., & Pachajoa, H. (2018). A novel de novo TBX5 mutation in a patient with Holt-Oram syndrome. The Application of Clinical Genetics, Volume 11, 157–162. DOI: 10.2147/TACG.S183418

22. Borozdin, W., Bravo Ferrer Acosta, A. M., Bamshad, M. J., Botzenhart, E. M., Froster, U. G., Lemke, J., Schinzel, A., Spranger, S., McGaughran, J., Wand, D., Chrzanowska, K. H., & Kohlhase, J. (2006). Expanding the spectrum ofTBX5 mutations in Holt-Oram syndrome: Detection of two intragenic deletions by quantitative real time PCR, and report of eight novel point mutations. Human Mutation, 27(9), 975–976. DOI: 10.1002/humu.9449

23. Fayez, A. G., Esmaiel, N. N., El Darsh, A. A., Ashaat, E. A., Kamel, A.K., Shehata, G. M., Ahmed, A. M., & El Ruby, M. O. (2017). Prediction of the functional consequences of a novel homozygous TBX5 variant in isolated AVSD patient. Journal of Innovations in Pharmaceutical and Biological Sciences, 4(3), 97-105.

24. Zhang, X.-L., Qiu, X.-B., Yuan, F., Wang, J., Zhao, C.-M., Li, R.-G., Xu, L., Xu, Y.-J., Shi, H.-Y., Hou, X.-M., Qu, X.-K., Xu, Y.-W., & Yang, Y.-Q. (2015). TBX5 loss-of-function mutation contributes to familial dilated cardiomyopathy. Biochemical and Biophysical Research Communications, 459(1), 166–171. DOI: 10.1016/j.bbrc.2015.02.094

25. Zhou, W., Zhao, L., Jiang, J.-Q., Jiang, W.-F., Yang, Y.-Q., & Qiu, X.-B. (2015). A novel TBX5 loss-of-function mutation associated with sporadic dilated cardiomyopathy. International Journal of Molecular Medicine, 36(1), 282–288. DOI: 10.3892/ijmm.2015.2206

26. Zhang, R., Tian, X., Gao, L., Li, H., Yin, X., Dong, Y., Yang, Y., & Xia, Y. (2016). Common variants in the tbx5 gene associated with atrial fibrillation in a chinese han population. PLOS ONE, 11(8), e0160467. DOI: 10.1371/journal.pone.0160467

27. Baban, A., Postma, A. V., Marini, M., Trocchio, G., Santilli, A., Pelegrini, M., Sirleto, P., Lerone, M., Albanese, S. B., Barnett, P., Boogerd, C. J., Dallapiccola, B., Digilio, M. C., Ravazzolo, R., & Pongiglione, G. (2014). Identification ofTBX5mutations in a series of 94 patients with Tetralogy of Fallot. American Journal of Medical Genetics Part A, 164(12), 3100–3107. DOI: 10.1002/ajmg.a.36783

28. Wang, F., Liu, D., Zhang, R.-R., Yu, L.-W., Zhao, J.-Y., Yang, X.-Y., Jiang, S.-S., Ma, D., Qiao, B., Zhang, F., Jin, L., Gui, Y.-H., & Wang, H.-Y. (2017). A TBX5 3′UTR variant increases the risk of congenital heart disease in the Han Chinese population. Cell Discovery, 3(1), 17026. DOI: 10.1038/celldisc.2017.26

29. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS ONE, 7(10), e46688. DOI: 10.1371/journal.pone.0046688

30. Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. DOI: 10.1093/nar/gkg509

31. Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols, 4(7), 1073–1081. DOI: 10.1038/nprot.2009.86

32. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7(4), 248–249. DOI: 10.1038/nmeth0410-248

33. Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., Mort, M., Cooper, D. N., Sebat, J., Iakoucheva, L. M., Mooney, S. D., & Radivojac, P. (2017). MutPred2: Inferring the molecular and phenotypic impact of amino acid variants. BioRxiv 134981. DOI: 10.1101/134981

34. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., & Damborsky, J. (2014). Predictsnp: Robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Computational Biology, 10(1), e1003440. DOI: 10.1371/journal.pcbi.1003440

35. Bendl, J., Musil, M., Štourač, J., Zendulka, J., Damborský, J., & Brezovský, J. (2016). Predictsnp2: A unified platform for accurately evaluating snp effects by exploiting the different characteristics of variants in distinct genomic regions. PLOS Computational Biology, 12(5), e1004962. DOI: 10.1371/journal.pcbi.1004962

36. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics, 24(8), 2125–2137. DOI: 10.1093/hmg/ddu733

37. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., … Sieh, W.

(2016). Revel: An ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics, 99(4), 877–885. DOI: 10.1016/j.ajhg.2016.08.016

38. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Research, 44(W1), W344–W350. DOI: 10.1093/nar/gkw408

39. Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology, 234(3), 779–815. DOI: 10.1006/jmbi.1993.1626

40.Meier, A., & Söding, J. (2015). Automatic prediction of protein 3d structures by probabilistic multi-template homology modeling. PLOS Computational Biology, 11(10), e1004343. DOI: 10.1371/journal.pcbi.1004343

41. Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., & Ascher, D. B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Progress in Biophysics and Molecular Biology, 128, 3–13. DOI: 10.1016/j.pbiomolbio.2016.10.002

42. van Wijk, R., Rijksen, G., Huizinga, E. G., Nieuwenhuis, H. K., & van Solinge, W. W. (2003). HK Utrecht: Missense mutation in the active site of human hexokinase associated with hexokinase deficiency and severe nonspherocytic hemolytic anemia. Blood, 101(1), 345–347. DOI: 10.1182/blood-2002-06-1851

43. Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., & Marcatili, P. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins: Structure, Function, and Bioinformatics, 87(6), 520–527. DOI: 10.1002/prot.25674

44. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R., & Alexov, E. (2013). Molecular mechanisms of disease-causing missense mutations. Journal of Molecular Biology, 425(21), 3919–3936. DOI: 10.1016/j.jmb.2013.07.014

45. Chen, C.-W., Lin, M.-H., Liao, C.-C., Chang, H.-P., & Chu, Y.-W. (2020). Istable 2. 0: Predicting protein thermal stability changes by integrating various characteristic modules. Computational and Structural Biotechnology Journal, 18, 622–630. DOI: 10.1016/j.csbj.2020.02.021

46. Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. Genome Biology, 3(3), reviews0004.1. DOI: 10.1186/gb-2002-3-3-reviews0004

47. Steri, M., Idda, M. L., Whalen, M. B., & Orrù, V. (2018). Genetic variants in mRNA untranslated regions. Wiley Interdisciplinary Reviews: RNA, 9(4), e1474. DOI: 10.1002/wrna.1474

48. De Angioletti, M., Lacerra, G., Sabato, V., & Carestia, C. (2004). Beta+45 g c: A novel silent beta-thalassaemia mutation, the first in the kozak sequence. British Journal of Haematology, 124(2), 224–231. DOI: 10.1046/j.1365-2141.2003.04754.x

49. Mohan, R. A., van Engelen, K., Stefanovic, S., Barnett, P., Ilgun, A., Baars, M. J. H., Bouma, B. J., Mulder, B. J. M., Christoffels, V. M., & Postma, A. V. (2014). A mutation in the Kozak sequence of GATA4 hampers translation in a family with atrial septal defects. American Journal of Medical Genetics Part A, 164(11), 2732–2738. DOI: 10.1002/ajmg.a.36703

50. von Bohlen, A. E., Böhm, J., Pop, R., Johnson, D. S., Tolmie, J., Stücker, R., Morris-Rosendahl, D., & Scherer, G. (2017). A mutation creating an upstream initiation codon in the SOX9 5′ UTR causes acampomelic campomelic dysplasia. Molecular Genetics & Genomic Medicine, 5(3), 261–268. DOI: 10.1002/mgg3.282

51. Witt, H., Luck, W., Hennies, H. C., Claßen, M., Kage, A., Laß, U., Landt, O., & Becker, M. (2000). Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. Nature Genetics, 25(2), 213–216. DOI: 10.1038/76088

52. Wen, Y., Liu, Y., Xu, Y., Zhao, Y., Hua, R., Wang, K., Sun, M., Li, Y., Yang, S., Zhang, X.-J., Kruse, R., Cichon, S., Betz, R. C., Nöthen, M. M., van Steensel, M. A. M., van Geel, M., Steijlen, P. M., Hohl, D., Huber, M., … Zhang, X. (2009). Loss-of-function mutations of an inhibitory

upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. Nature Genetics, 41(2), 228–233. DOI: 10.1038/ng.276

53. Hudder, A., & Werner, R. (2000). Analysis of a charcot-marie-tooth disease mutation reveals an essential internal ribosome entry site element in the connexin-32 gene. Journal of Biological Chemistry, 275(44), 34586–34591. DOI: 10.1074/jbc.M005199200

54. Nicolas, G., Wallon, D., Goupil, C., Richard, A.-C., Pottier, C., Dorval, V., Sarov-Rivière, M., Riant, F., Hervé, D., Amouyel, P., Guerchet, M., Ndamba-Bandzouzi, B., Mbelesso, P., Dartigues, J.-F., Lambert, J.-C., Preux, P.-M., Frebourg, T., Campion, D., Hannequin, D., … Rovelet-Lecrux, A. (2016). Mutation in the 3'untranslated region of APP as a genetic determinant of cerebral amyloid angiopathy. European Journal of Human Genetics, 24(1), 92–98. DOI: 10.1038/ejhg.2015.61

55. Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., Banfi, S., Gennarino, V. A., Horner, D. S., Pavesi, G., Picardi, E., & Pesole, G. (2010). UTRdb and UTRsite (RELEASE 2010): A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Research, 38(suppl_1), D75–D80. DOI: 10.1093/nar/gkp902

56. Bartel, D. P. (2004). Micrornas. Cell, 116(2), 281–297. DOI: 10.1016/S0092-8674(04)00045-5

57. Moszyńska, A., Gebert, M., Collawn, J. F., & Bartoszewski, R. (2017). SNPs in microRNA target sites and their potential role in human disease. Open Biology, 7(4), 170019. DOI: 10.1098/rsob.170019

58. Liu, C., Zhang, F., Li, T., Lu, M., Wang, L., Yue, W., & Zhang, D. (2012). MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics, 13(1), 661. DOI: 10.1186/1471-2164-13-661

59. Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: Examples, detection, and confirmation. Journal of Applied Genetics, 59(3), 253–268. DOI: 10.1007/s13353-018-0444-7

60. Wang, G.-S., & Cooper, T. A. (2007). Splicing in disease: Disruption of the splicing code and the decoding machinery. Nature Reviews Genetics, 8(10), 749–761. DOI: 10.1038/nrg2164

61. Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. Nucleic Acids Research, 37(9), e67–e67. DOI: 10.1093/nar/gkp215

62. Yang, J., & Zhang, Y. (2015). I-TASSER server: New development for protein structure and function predictions. Nucleic Acids Research, 43(W1), W174–W181. DOI: 10.1093/nar/gkv342

63. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols, 10(6), 845–858. DOI: 10.1038/nprot.2015.053

64. Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., & Baker, D. (2013). High-resolution comparative modeling with RosettaCm. Structure, 21(10), 1735–1742. DOI: 10.1016/j.str.2013.08.005

65. Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., & Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins: Structure, Function, and Bioinformatics, 77(S9), 114–122. DOI: 10.1002/prot.22570

66. Krieger, E., Koraimann, G., & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA-a self-parameterizing force field. Proteins: Structure, Function, and Bioinformatics, 47(3), 393–402. DOI: 10.1002/prot.10104

67. Heo, L., Park, H., & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. Nucleic Acids Research, 41(W1), W384–W388. DOI: 10.1093/nar/gkt458

68. Laskowski, R. A., MacArthur, M. W., & Thornton, J. M. (2012). procheck: Validation of protein-structure coordinates (pp. 684–687). DOI: 10.1107/97809553602060000882

69. Lüthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature, 356(6364), 83–85. DOI: 10.1038/356083a0

70. Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation: PROTEIN SCIENCE.ORG. Protein Science, 27(1), 293–315. DOI: 10.1002/pro.3330

71. Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo—Distance constraints applied on model quality estimation. Bioinformatics, 36(6), 1765–1771. DOI: 10.1093/bioinformatics/btz828

72. Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics, 27(3), 343–350. DOI: 10.1093/bioinformatics/btq662

73. RefSNPs (rs) Defined. SNP FAQ Archive [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK44417/. Accessed April 20,2020.

74. Lee, J., Freddolino, P. L., & Zhang, Y. (2017). Ab initio protein structure prediction. In D. J. Rigden (Ed.), From Protein Structure to Function with Bioinformatics (pp. 3–35). Springer Netherlands. DOI: 10.1007/978-94-024-1069-3_1