# In Silico Genetics Revealing 5 Mutations in *CEBPA* Gene Associated With Acute Myeloid Leukemia

Cancer Informatics
Volume 18: 1–18
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935119870817

Mujahed I Mustafa[1], Zainab O Mohammed[2], Naseem S Murshed[1], Nafisa M Elfadol[1], Abdelrahman H Abdelmoneim[1] and Mohamed A Hassan[1]

[1]Department of Biotechnology, Africa City of Technology, Khartoum North, Sudan. [2]Department of Haematology, Ribat University Hospital, Khartoum, Sudan.

## ABSTRACT

**BACKGROUND:** Acute myeloid leukemia (AML) is an extremely heterogeneous malignant disorder; AML has been reported as one of the main causes of death in children. The objective of this work was to classify the most deleterious mutation in CCAAT/enhancer-binding protein-alpha (*CEBPA*) and to predict their influence on the functional, structural, and expression levels by various Bioinformatics analysis tools.

**METHODS:** The single nucleotide polymorphisms (SNPs) were claimed from the National Center for Biotechnology Information (NCBI) database and then submitted into various functional analysis tools, which were done to predict the influence of each SNP, followed by structural analysis of modeled protein followed by predicting the mutation effect on energy stability; the most damaging mutations were chosen for additional investigation by Mutation3D, Project hope, ConSurf, BioEdit, and UCSF Chimera tools.

**RESULTS:** A total of 5 mutations out of 248 were likely to be responsible for the structural and functional variations in CEBPA protein, whereas in the 3′-untranslated region (3′-UTR) the result showed that among 350 SNPs in the 3′-UTR of *CEBPA* gene, about 11 SNPs were predicted. Among these 11 SNPs, 65 alleles disrupted a conserved miRNA site and 22 derived alleles created a new site of miRNA.

**CONCLUSIONS:** In this study, the impact of functional mutations in the CEBPA gene was investigated through different bioinformatics analysis techniques, which determined that R339W, R288P, N292S, N292T, and D63N are pathogenic mutations that have a possible functional and structural influence, therefore, could be used as genetic biomarkers and may assist in genetic studies with a special consideration of the large heterogeneity of AML.

**KEYWORDS:** Acute myeloid leukemia, malignant disease, CEBPA, bioinformatics analysis, genetic biomarkers

**RECEIVED:** July 13, 2019. **ACCEPTED:** July 30, 2019.

**TYPE:** Original Research

**FUNDING:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**DECLARATION OF CONFLICTING INTEREST:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Mujahed I Mustafa, Department of Biotechnology, Africa City of Technology, Al Siteen St, Khartoum North, Khartoum 79371, Sudan.
Email: mujahedimustafa@gmail.com

## Introduction

Acute myeloid leukemia (AML) is an extremely heterogeneous malignant disorder; in recent years, AML developed so rapidly by affecting children and adults, and hence it has been reported as one of the main causes of death in children.[1-5] It is the most common acute leukemia in adults,[6-8] with a frequency of more than 20 000 cases per year in the United States alone.[6] It is characterized by genetic alterations in hematopoietic ancestor cells that change usual mechanisms of self-replicating.[2,9] AML is commonly triggered by mutation in CCAAT/enhancer-binding protein-alpha (*CEBPA*) gene.[10-14] The *CEBPA* is a transcription element that affects immune cell density and diversity.[15,16] Most patients with AML who have *CEBPA* alterations simultaneously transport double mutations,[17-19] nevertheless, different mutations have been reported[20-24]; some studies have been reported which claimed some related factors besides *CEBPA* mutation, such as smoking, alcohol, and exposure to solvents and agrochemicals may cause AML,[25-27] but no evidence of publication bias. Other genes have been reported which cause AML such as *fms-related tyrosine kinase 3* (*FLT3-ITD*) and *nucleophosmin 1* (*NMP1*), which assisted to improve person diagnosis; furthermore, these mutant molecules characterize as a potential target for molecular therapies.[6,28-31] Sometimes, patients with chronic lymphocytic leukemia can develop AML,[32] whereas in rare cases patients with AML can develop esophageal cancer.[33]

Stem cell transplantation treatment is related to the result of treatment for patients with cytogenetically usual AML.[34] Nevertheless, the advantage of the transplant is exclusive to subcategory of patients with *CEBPA* mutations alone[31,34]; in spite of this hopeful recent evolution, the outcomes of patients with AML remain insufficient, with more than 50% of the patients eventually dying from this devastating disease. The purpose of this study is to classify functional mutations located in the coding region of *CEBPA* gene using in silico analysis.

Disease-causing single nucleotide polymorphisms (SNPs) are frequently found to arise at evolutionarily conserved regions; these have a key role at structural and functional levels of the protein. The capability to calculate whether a particular SNP is deleterious or not is very important for the prognosis of disorder.[35-45] The practice of translational bioinformatics has solid influence on the identification of candidate SNPs and can

contribute in pharmacogenomics by identifying high-risk SNP mutation contributing to drug response as well as developing novel therapeutic elements for this deadly disease.[46-54] This is the first silico analysis in coding and non-coding regions of *CEBPA* gene that prioritized SNPs to be used as diagnostic markers with a special consideration of the large heterogeneity of AML among different populations.

## Materials and Methods

### *Data mining*

The polymorphic data of *CEBPA* gene were claimed from National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/), and the reference sequence of human protein was collected from UniProt[55] (https://www.uniprot.org/).

### *Functional analysis*

*Sorting Intolerant From Tolerant.* It is the first in silico functional analysis that calculates whether an amino acid alteration affects protein function or not. Sorting Intolerant From Tolerant (SIFT) scores $< 0.05$ are expected to be damaging altered amino acid, otherwise it is considered to be tolerant.[56] It is available at https://sift.bii.a-star.edu.sg/.

*PolyPhen-2.* It is a trained machine learning to predict whether an amino acid replacement affects protein function and structure or not, by calculating position-specific independent count (PSIC) for each SNP at a time. There are 2 outputs whether probably damaging (values are more frequently 1) and possibly damaging or benign (values range from 0 to 0.95).[57] It is available at http://genetics.bwh.harvard.edu/pph2/.

*PROVEAN.* It is an online in silico functional analysis tool that calculates whether an amino acid replacement has an influence on the organic function of a protein stranded on the alignment-based score. If the PROVEAN score $\leqslant -2.5$, the protein variant is expected to have a "deleterious" effect, whereas if the PROVEAN score is $>-2.5$, the variant is expected to have a "neutral" effect.[58] It is available at http://provean.jcvi.org/index.php.

*SNAP2.* It is a trained functional analysis tool that differentiates between effect and neutral SNPs by taking various features into validation. SNAP2 got an accuracy of 83%, which has 2 expectations: effect (positive score) or neutral (negative score). It is considered an important and substantial enhancement over other methods. It is available at https://rostlab.org/services/snap2web/.

*SNPs&GO.* It is a trained machine learning based on the technique to precisely calculate the deleterious associated alterations from protein sequence. SNPs&GO collects in a unique framework information derived from protein sequence, evolutionary information, and function as coded in the Gene Ontology terms and underperforms other available predictive methods (PhD-SNP and PANTHER).[59] It is available at http://snps.biofold.org/snps-and-go/snps-and-go.html.

*PMut.* It is a web-based tool for the explanation of SNP alternates on proteins, which allows the rapid and precise calculation (80%) of the compulsive features of each SNP grounded on the practice of neural networks.[60] It is accessible at http://mmb.irbbarcelona.org/PMut.

### *Stability analysis*

*I-Mutant 3.0.* I-Mutant is a support vector machine (SVM)-based tool. I-Mutant predicts whether the protein mutation stabilizes or destabilizes the protein structure by calculating free energy change by coupling predictions with the energy-based FOLD-X tool.[61] It is available at http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi.

*MUpro.* It is a structural analysis online tool for the calculation of protein stability variations in arbitrary SNPs. The value of the energy change is expected, and assurance mark between –1 and 1 for evaluating the assurance of the expectation is calculated. A score of $<0$ means the mutant decreases the protein stability; conversely, a score of $>0$ means the mutant increases the protein stability.[62] It is available at http://mupro.proteomics.ics.uci.edu/.

### *3-Dimensional clustering analysis*

*Mutation3D.* It is a functional calculation and visualization online tool for investigating the 3-dimensional (3D) plan of amino acid alterations in protein models and structures.[63] It is available at http://mutation3d.org.

### *Biophysical validation*

*Project HOPE.* It is a web server to search protein 3D structures by bringing together structural information from several sources such as UniProt database. The main aims for the submissions in Project HOPE are to analyze and confirm the results that we obtained earlier. It is available at http://www.cmbi.ru.nl/hope.

### *Conservational analysis*

*BioEdit.* It is a software package proposed to stream a distinct program that can run nearly any sequence operation as well as a few basic alignment investigations. It is available for download at http://www.mbio.ncsu.edu/bioedit/bioedit.html.

*ConSurf server.* It is a web server that offers evolutionary conservation summaries for proteins of known structure in the protein data bank. ConSurf spots the parallel amino acid sequences and runs multialignment methods. The conserved

amino acid across species flags its position using specific algorithm.[64] It is available at http://consurf.tau.ac.il/.

### *3D structural analysis*

*RaptorX.* The 3D structure of human CEBPA protein is not available in the Protein Data Bank. Hence, RaptorX was used to make a 3D structural model for wild-type CEBPA. RaptorX is a web server predicting structure property of a protein sequence without using any templates.[65] It is available at http://raptorx.uchicago.edu/.

*UCSF Chimera.* It is a visualization analysis program of 3D structure prototype, docking analysis, and so many related analyses. A predicted model was created by RaptorX to visualize and compare the amino acid alterations using UCSF Chimera.[66] UCSF Chimera 1.8 is free for download at http://www.cgl.ucsf.edu/chimera/.

### *GeneMANIA*

It is a method to know protein function prediction integrating multiple genomics and proteomics data sources to make inferences about the function of unknown proteins.[67] It is available at http://www.genemania.org/.

### *Variant Effect Predictor*

The Ensembl Variant Effect Predictor (VEP) software provides tools and methods for a systematic approach to annotate and aid prioritization of variants in both large-scale sequencing projects and smaller analysis studies.[68] It is available at http://www.ensembl.org/vep.

### *PolymiRTS server*

It is a server for investigating functional SNPs in 3′-untranslated region (3′-UTR) of *CEBPA* gene that may change miRNA binding on target sites, resulting in different functional consequences.[69] It is available at http://compbio.uthsc.edu/miRSNP/.

## Results

The total number of SNPs in different regions of *CEBPA* gene was retrieved from NCBI. The distribution of non-synonymous single nucleotide polymorphisms (nsSNPs) in coding and non-coding regions of *CEBPA* gene contained 248 nsSNPs, with 350 SNPs in the 3′-UTR and 11 in the 5′-untranslated region (5′-UTR; Figure 1).

A total of 248 missense mutations were retrieved from the database of single nucleotide polymorphism (dbSNP)/NCBI database, and these SNPs were submitted into different functional analysis tools such as SIFT, polymorphism phenotyping v2 (PolyPhen-2), PROVEAN, and SNAP2, respectively. Sorting Intolerant From Tolerant server predicted 28

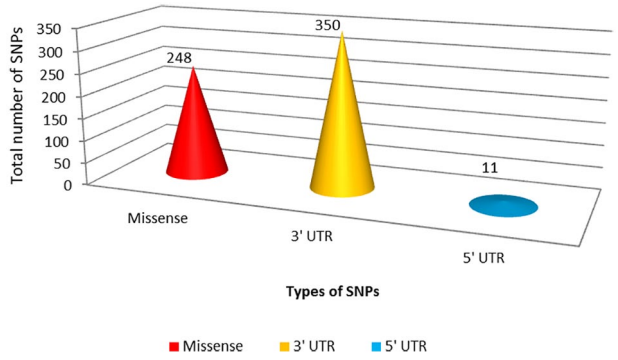


**Figure 1.** The distribution of SNPs in coding and in non-coding regions of *CEBPA* gene. SNPs indicates single nucleotide polymorphisms.

deleterious SNPs, PolyPhen-2 predicted 85 damaging SNPs (29 were possibly damaging and 56 were probably damaging to protein), PROVEAN represented 34 deleterious SNPs, whereas in SNAP2 we filtered the triple-positive deleterious SNPs from the previous 3 analysis tools, out of 53 SNPs there were 19 predicted deleterious SNPs by SNAP2. Table 1 represents the Quad-positive of deleterious SNPs after filtrations, the number decreased rapidly to 19 SNPs, after submitting them into SNPs&GO and PhD-SNP, PMut and PANTHER, respectively, to run more investigation on these SNPs and their effect on the functional level. The triple positive in the 3 tools was 5 disease-associated SNPs (Table 2). Finally, we submitted them to I-mutant 3.0 and MUpro, respectively, to investigate their effect on structural level. The 2 online tools revealed that all 5 mutations predicted a dramatic decrease in the protein stability, except for 2 SNPs (N292T and D63N) that were predicted by I-Mutant to increase the stability of the protein (Table 3).

Single nucleotide polymorphisms in 3′-UTR of *CEBPA* gene were submitted as batch to PolymiRTS server. The result shows that among 350 SNPs in the 3′-UTR of *CEBPA* gene, about 11 SNPs were predicted, namely, rs116528776, rs113670631, rs34017519, rs146104564, rs187516157, rs2376497, rs192371350, rs1049969, rs41367646, rs184965384, and rs187751931; among these 11 SNPs, 65 alleles disrupted a conserved miRNA site and 22 derived alleles created a new site of miRNA (Table 6).

## Discussion

In vitro mutagenesis, functional and characterization studies, is an unwieldy task regarding workload, time, and fees. For these reasons, bioinformatics analysis is an appropriate, rapid, low-cost, and dependable approach to enhance our understanding of how mutations could disturb the protein structure and function.[53,54] Disease-causing SNPs are commonly found to arise at evolutionarily conserved regions. Those have a key role at structural and functional levels of the protein[36,38]; therefore, our focus was dedicated to the coding region, which unmasked 5 mutations in *CEBPA* gene using different sequence and

**Table 1.** Affect or damaging nsSNPs associated predicted by various software.

| DBSNP RS# | SUB | SIFT PREDICTION | SCORE | POLYPHEN PREDICTION | SCORE | PROVEAN PREDICTION | SCORE | SNAP2 PREDICTION | SCORE |
|---|---|---|---|---|---|---|---|---|---|
| rs1358286265 | G355V | AFFECT | 0 | Damaging | 1 | Deleterious | −7.083 | Effect | 48 |
| rs1402033817 | G355S | AFFECT | 0 | Damaging | 1 | Deleterious | −4.501 | Effect | 47 |
| rs1439202716 | R343C | AFFECT | 0 | Damaging | 1 | Deleterious | −6.702 | Effect | 43 |
| rs1455027551 | R339W | AFFECT | 0 | Damaging | 1 | Deleterious | −7.798 | Effect | 74 |
| rs758726582 | R333C | AFFECT | 0 | Damaging | 1 | Deleterious | −7.409 | Effect | 14 |
| rs1422138876 | V328G | AFFECT | 0 | Damaging | 1 | Deleterious | −6.823 | Effect | 51 |
| rs1306818311 | R327W | AFFECT | 0 | Damaging | 1 | Deleterious | −6.468 | Effect | 47 |
| rs781549846 | R325C | AFFECT | 0 | Damaging | 1 | Deleterious | −6.324 | Effect | 19 |
| rs1013241741 | S299C | AFFECT | 0 | Damaging | 1 | Deleterious | −4.866 | Effect | 23 |
| rs1391793930 | K298R | AFFECT | 0 | Damaging | 1 | Deleterious | −2.924 | Effect | 31 |
| rs1392203731 | K298Q | AFFECT | 0 | Damaging | 1 | Deleterious | −3.899 | Effect | 34 |
| rs776590829 | N292T | AFFECT | 0 | Damaging | 1 | Deleterious | −5.788 | Effect | 35 |
| – | N292S | AFFECT | 0 | Damaging | 1 | Deleterious | −4.907 | Effect | 29 |
| rs1064794962 | R288P | AFFECT | 0 | Damaging | 1 | Deleterious | −6.636 | Effect | 66 |
| rs376856647 | E284D | AFFECT | 0 | Damaging | 1 | Deleterious | −2.694 | Effect | 44 |
| rs1196766447 | E284A | AFFECT | 0 | Damaging | 1 | Deleterious | −5.588 | Effect | 18 |
| rs1352573347 | P233H | AFFECT | 0 | Damaging | 1 | Deleterious | −3.924 | Effect | 37 |
| rs1267025311 | R156W | AFFECT | 0 | Damaging | 1 | Deleterious | −2.983 | Effect | 71 |
| rs1452063514 | D63N | AFFECT | 0 | Damaging | 1 | Deleterious | −2.992 | Effect | 72 |

Abbreviations: dbSNP, database of single nucleotide polymorphism; nsSNPs, non-synonymous single nucleotide polymorphisms; SIFT, Sorting Intolerant From Tolerant; SUB, substitution.

structure-based algorithms (Figure 2). The SNPs that have been found in this study could be used in prognostics of disease, because identification of *CEBPA* status in AML has a major clinical importance, allowing relapse risk to be stratified properly for post-remission treatment.[70,71]

All these SNPs (D63N, R288P, N292T, N292S, and R339W) were retrieved from the dbSNPs/NCBI database as untested and all were found to be pathogenic mutations.

At the functional level analysis, our results showed that all these nsSNP substitutions (D63N, R288P, N292T, N292S, and R339W) were classified as highly pathogenic mutations (Table 1). The analysis of different SNPs on the protein structure can disturb interactions with other molecules, MUpro results showed a decrease in stability for all these SNPs (D63N, R288P, N292T, N292S, and R339W), whereas I-Mutant results showed a decrease in stability for these SNPs (R288P, N292S, and R339W), thus suggesting that these mutations could directly or indirectly destabilize the amino acid interactions triggering functional deviations of protein to some point.

*CEBPA* offers information for building a protein termed CCAAT enhancer-binding protein alpha. It's a transcription factor (TF) and its performance is a malignant suppressor, which means it is complicated in cellular mechanisms and could help to prevent the cells from developing and dividing too swiftly or in an uncontrolled mode and that is the principle of cancer.[24,72] We also achieved analysis by Mutation3D server, all SNPs in red (R288P, N292S, and N292T) are clustered mutation, significantly, such mutation clusters are commonly associated with human cancers,[73] whereas SNPs in blue (R339W) and gray (D63N) are covered and uncovered mutations, respectively (Figure 3).

Project HOPE server was used to submit the most damaging SNPs (R288P): interestingly, proline interrupts an α-helix when not positioned at 1 of the first 3 positions of that helix. If this happened, a major impact on the protein structure could occur (Figure 4). In this study, we also observed that only 1 SNP (D63N), the residue predicted to be mutated, is evolutionarily conserved across species, and this may increase the

**Table 2.** Pathogenic nsSNPs associated variations predicted by various software.

| SUB | SNPS&GO PREDICTION | RI | PROBABILITY | PANTHER PREDICTION | RI | PROBABILITY | PhD-SNP PREDICTION | RI | PROBABILITY | PMUT PREDICTION | PROBABILITY |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R339W | Disease | 1 | .539 | Disease | 9 | .949 | Disease | 2 | .615 | Disease | .85 (91%) |
| N292T | Disease | 5 | .723 | Disease | 10 | .985 | Disease | 4 | .749 | Disease | .85 (91%) |
| N292S | Disease | 5 | .751 | Disease | 10 | .98 | Disease | 5 | .767 | Disease | .85 (91%) |
| R288P | Disease | 4 | .704 | Disease | 7 | .86 | Disease | 7 | .872 | Disease | .85 (91%) |
| D63N | Disease | 4 | .717 | Disease | 9 | .94 | Disease | 4 | .697 | Disease | .67 (85%) |

Abbreviations: nsSNPs, non-synonymous single nucleotide polymorphisms; RI, reliability index; SUB, substitution.

possibility of altered transcriptional and cell cycle regulation (Figure 5).

The 3D protein structure analysis enables mapping of amino acid substitutions and, therefore, RaptorX was used to make a 3D structure model for CEBPA protein (Figure 6) to support and match the results acquired from different computational tools, UCSF Chimera serves this purpose (Figures 7 to 11), show the differences between native and mutant amino acids, in the green and red boxes the schematic structures of the native amino acids (in the left side), and the mutant ones (in the right side). The backbone, which is the same for each amino acid, is colored red and the side chain, unique for each amino acid, is colored black, the 3D wide-type residues colored green and mutant ones colored red, whereas the protein is colored cyan.

In Figure 7, D63N shows the native amino acid (aspartic acid) and the mutant one (asparagine) at position 63; the mutated residue is located on the surface of a domain with unknown function; the residue was not found to be in contact with other domains of which the function is known within the used structure; however, contact with other molecules or domains is still possible and might be affected by this mutation.

In Figure 8, R288P shows close-up angle of the native amino acid (arginine) and the mutant one (proline) at position 288; the mutated residue is located in a domain that is important for the activity of the protein and in contact with residues in another domain, and it is possible that this interaction is important for the correct function of the protein. The mutation can affect this interaction and as such affect protein function; the mutation introduces an amino acid with different properties, which can disturb this domain and abolish its function, the charge of the wild-type residue is lost by this mutation, which can cause loss of interactions with other molecules; the mutant residue is smaller than the wild-type residue; and this will cause a possible loss of external interactions.

In Figure 9, N292S shows the schematic structures of the original amino acid (asparagine) and the mutant one (serine) at position 292; each amino acid has its own specific size, charge, and hydrophobicity value. The original wild-type residue and newly introduced mutant residue often differ in these properties; the mutant residue is more hydrophobic than the wild-type residue; the mutant residue is smaller than the wild-type residue; and this will cause a possible loss of external interactions.

In Figure 10, N292T shows close-up angle of the native amino acid (asparagine) and the mutant one (threonine) at position 292; the mutated residue is located in a domain that is important for the activity of the protein and in contact with residues in another domain. It is possible that this interaction is important for the correct function of the protein. The mutation can affect this interaction and as such affect protein

**Table 3.** Structural investigation calculated using I-Mutant 3.0 and MUpro.

| DBSNP RS# | SUB | I-MUTANT PREDICTION | RI | DDG VALUE PREDICTION | MUPRO PREDICTION | SCORE |
|---|---|---|---|---|---|---|
| rs1455027551 | R339W | Decrease | 3 | −0.21 | Decrease | −0.24675 |
| rs776590829 | N292T | Increase | 2 | −0.35 | Decrease | −1.48124 |
| – | N292S | Decrease | 6 | −0.73 | Decrease | −1.46979 |
| rs1064794962 | R288P | Decrease | 5 | −0.73 | Decrease | −0.66005 |
| rs1452063514 | D63N | Increase | 1 | −0.2 | Decrease | −1.22302 |

Abbreviations: dbSNP, database of single nucleotide polymorphism; RI, reliability index; SUB, substitution.
DDG value: free energy changes value.

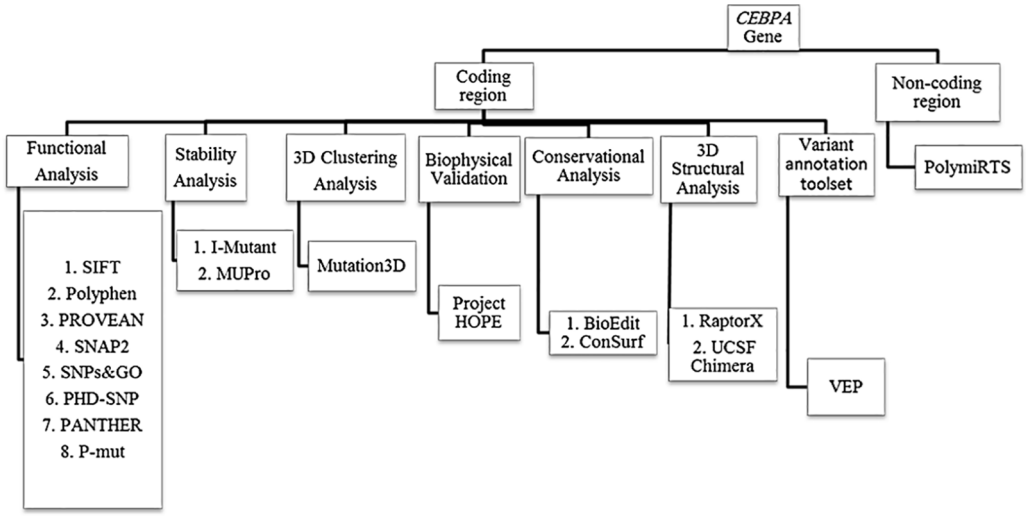


**Figure 2.** Descriptive workflow of softwares used in SNP analysis. SNP indicates single nucleotide polymorphism.

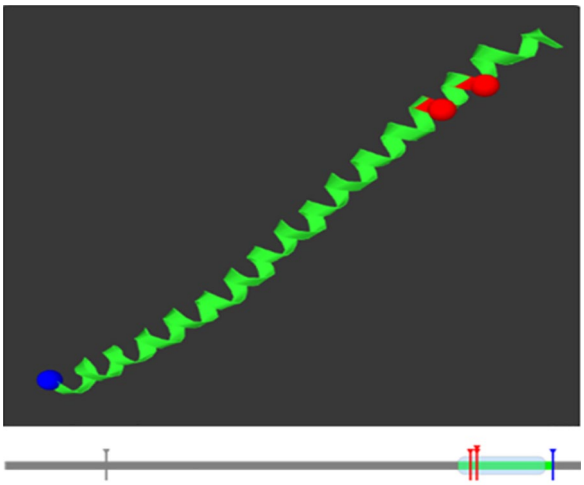

**Figure 3.** Structural simulations for mutant residues in CEBPA protein, demonstrated by Mutation3D.

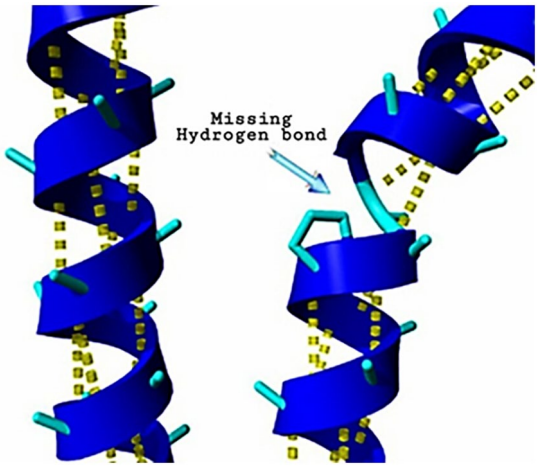


**Figure 4.** The mutant residue located in an α-helix.

function; the mutant residue is smaller than the wild-type residue; and this will cause a possible loss of external interactions.

In Figure 11, R339W shows the schematic structures of the original amino acid (arginine) and the mutant one (tryptophan) at position 339; the residue is located on the surface of the protein; mutation of this residue can disturb interactions with other molecules or other parts of the protein; and the charge of the wild-type residue (positive) is lost by this mutation. This can cause loss of interactions with other molecules. The mutant residue is more hydrophobic than the wild-type residue, which can disturb this domain and abolish its function.

We also used ConSurf web server; the nsSNPs that are shown by black boxes located in highly conserved regions and predicted to cause structural and functional impacts on CEBPA protein (Figure 12).

GeneMANIA revealed strong functional associations that *CEBPA* gene had observed with transforming growth factor beta (*TGFB1*) and tumor necrosis factor (*TNF*) genes (Figure 13). Besides, weak interactions with less confidence have been observed for prolactin regulatory element binding (*PREB*) and early B cell factor 1 (*EBF1*) genes. The genes co-expressed with, sharing similar protein domain, or contributed to achieve similar function are shown in Tables 4 and 5.

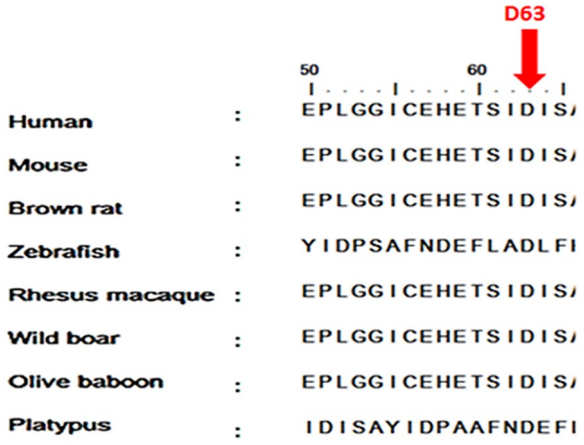


**Figure 5.** Alignments of 8 amino acid sequences of CEBPA representing that the residues predicted to be mutated are evolutionarily conserved across species. Sequences alignment was done by BioEdit (v7.2.5).

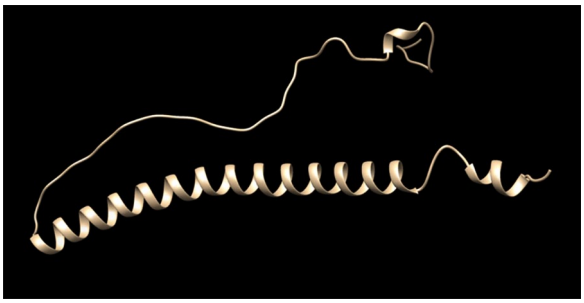

**Figure 6.** The 3D structure of the CEBPA protein model was generated by using RaptorX; it could not generate the 3D structure of all amino acid positions; therefore, the model was done from positions 52 to 358, due to the lack of information. 3D indicates 3-dimensional.

The VEP annotates variants using a wide range of reference data, including transcripts, regulatory regions, and frequencies from previously observed variants, citations, clinical significance information, and predictions of biophysical consequences of variants, and that is what makes VEP to give accurate results; as far as we know, the only limitation is that VEP annotates each input variant independently, without considering the potential compound effects of combining alternate alleles across multiple variant loci,[68] and this is the reason why we could not predict the consequences of N292S mutation, whereas the predicted variant consequences are shown in Table 6. VEP reported regulatory consequences for many variants, including 5 variants within a coding region, 6 variants within a non-coding region, 10 variants within upstream gene, 6 variants within downstream gene, 4 variants within non-coding transcript exon, and 1 variant within transcription factor binding site (TFBS); in conclusion, mutations within a coding region affect the protein function, whereas regulatory variants within non-coding genomic regions can greatly affect disease and could be involved in the specific recruitment or sequestration of spliceosome factors and RNA-binding proteins (RBPs)[74,75]; the SNPs in the upstream, downstream, 5′-, and 3′-UTRs might affect transcription or translation process[76]; whereas alteration at TFBS has many consequences, such as variants within a TFBS differentially influence its TF-binding affinity; another consequence that could affect TFBS is that multiple variants in the promoter regions can "transform" an existing binding site of a particular TF into a site for another TF, even from a different TF family.[77] Figure 14 illustrates the summary pie charts and statistics.

Single nucleotide polymorphisms in 3′-UTR of *CEBPA* gene were submitted as batch to PolymiRTS server. The result showed that 11 SNPs may affect microRNA binding sites. As an example, rs2376497 SNP containing (D) allele had 8 microRNA sites (miRSite) as target binding site that can disrupt a conserved miRNA and (C) alleles had 5 miRSites that

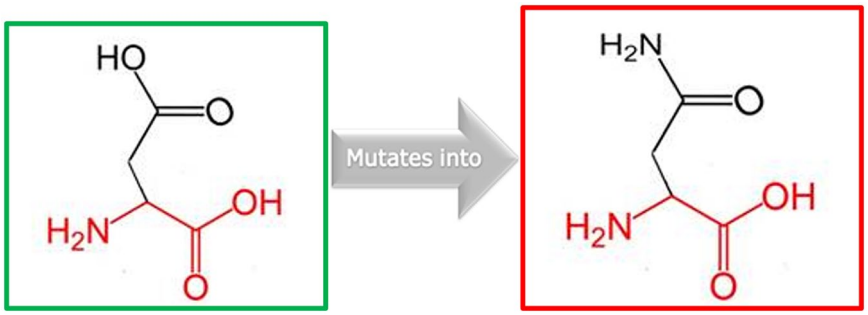


**Figure 7.** D63N: aspartate (green box) changes to asparagine (red box) at position 63.

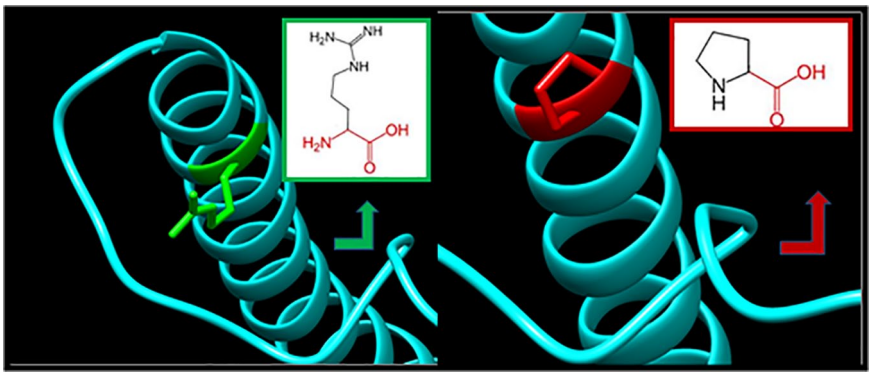

**Figure 8.** R288P: the amino acid arginine changes to proline at position 288; illustration was done by UCSF Chimera (v 1.8.) and project HOPE.

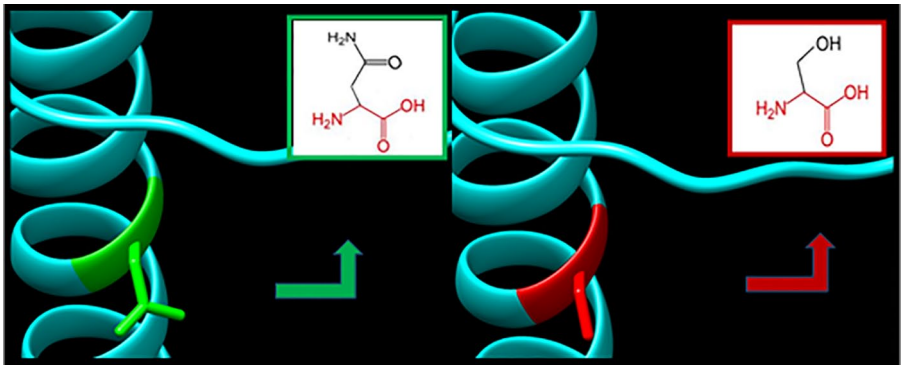

**Figure 9.** N292S: the amino acid asparagine changes to serine at position 292; illustration was done by UCSF Chimera (v 1.8.) and project HOPE.

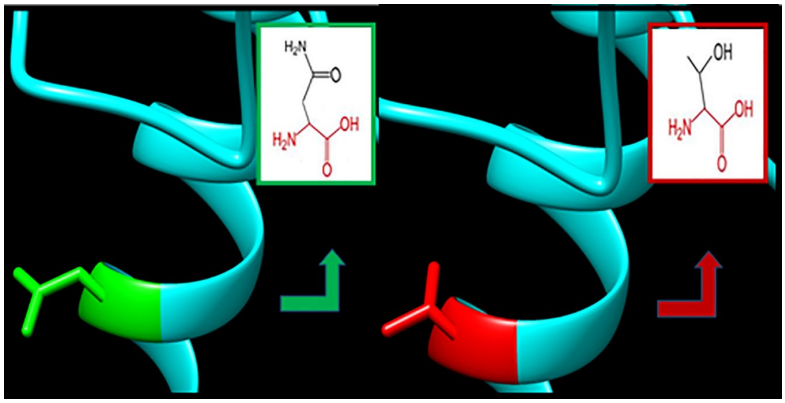

**Figure 10.** N292T: the amino acid asparagine changes to threonine at position 292; illustration was done by UCSF Chimera (v 1.8.) and project HOPE.

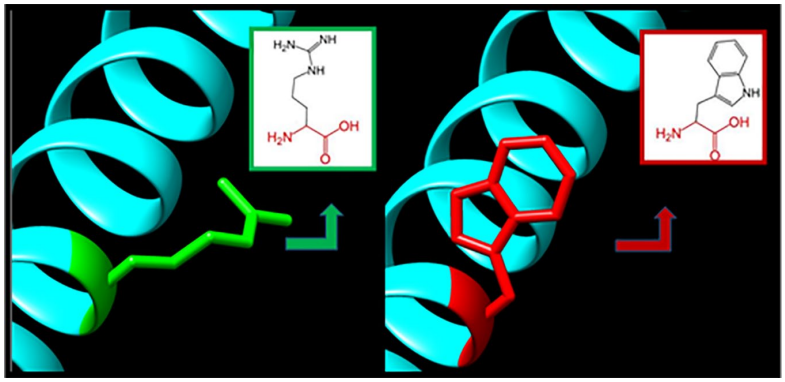

**Figure 11.** R339W: the amino acid arginine changes to tryptophan at position 339; illustration was done by UCSF Chimera (v 1.8.) and project HOPE.

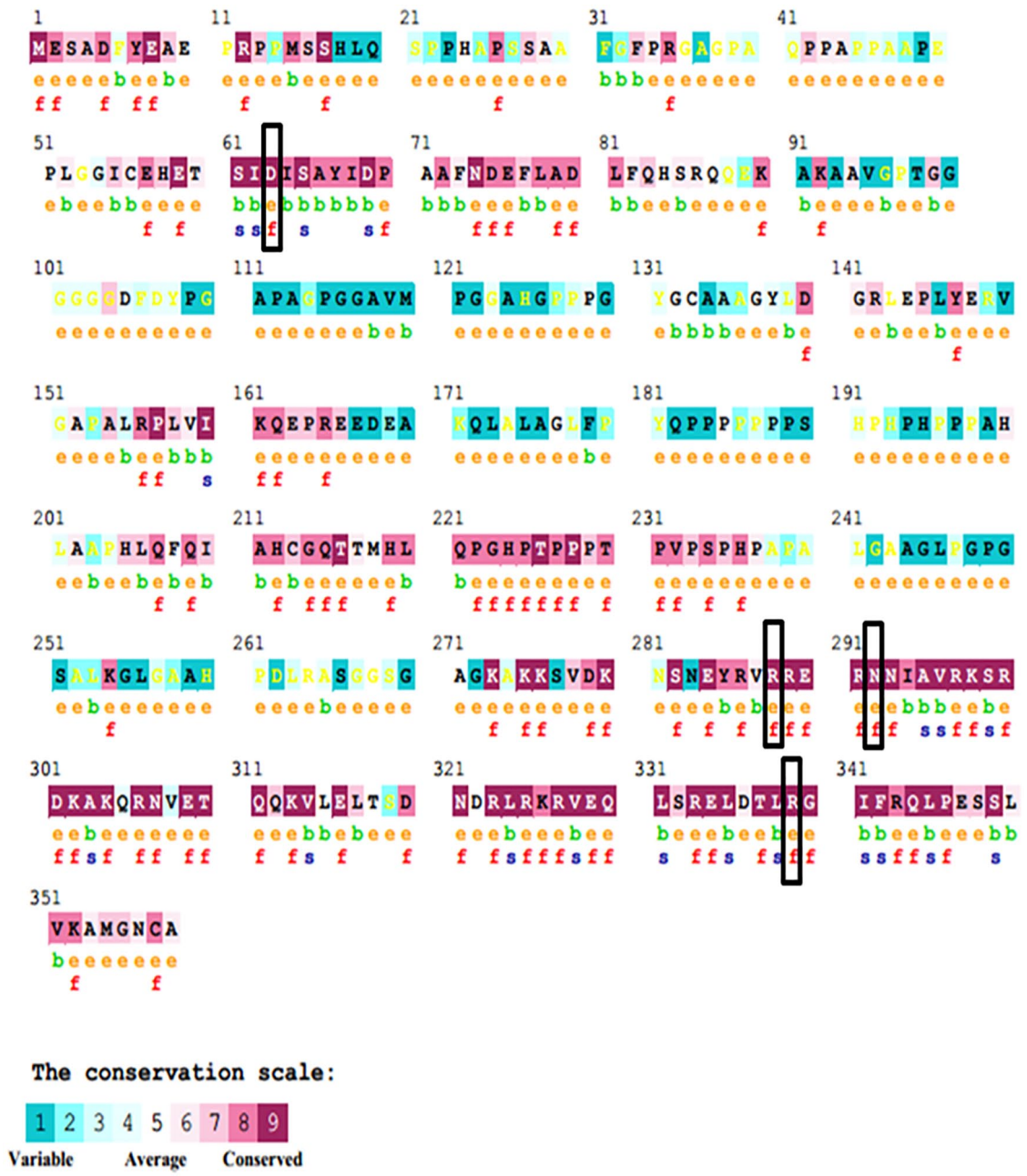


**Figure 12.** The conserved amino acids across species in CEBPA protein were determined using ConSurf. e: exposed residues according to the neural-network algorithm are indicated in orange letters. b: residues predicted to be buried are demonstrated via green letters. f: predicted functional residues (highly conserved and exposed) are indicated with red letters. s: predicted structural residues (highly conserved and buried) are demonstrated in blue letters. I: insufficient data—the calculation for this site performed in less than 10% of the sequences is demonstrated in yellow letters.

disrupt a conserved miRSite. Table 7 demonstrates the SNPs predicted by PolymiRTS to induce disruption or formation of miRSite.

The limitations of this study are that it focuses on coding and 3′-UTRs using different numbers of tools of silico analysis; yet there are number of genes responsible for AML although AML is frequently triggered by mutation in *CEBPA* gene[10-13]; in general, it is likely to achieve that computational approach remains as an accurate way to make a rapid analysis regarding the expected effect of mutations; nevertheless, the more factors that are taken into account, the more accurate the prediction will be. To take the best advantage of bioinformatics analysis, different computational tools could be used, trying to cover the major aspects influencing protein structure and function, Mutation Taster,[78] SNPdryad,[79] and ACES (a machine learning toolbox for clustering analysis and visualization).[80] The 5′-UTRs have not been analyzed in this study; these SNPs are likely to affect the level of gene expression; the impact of SNPs at the 5′-UTRs can be predicted by using some of the RNA assessment tools, such as PreTIS.[81]

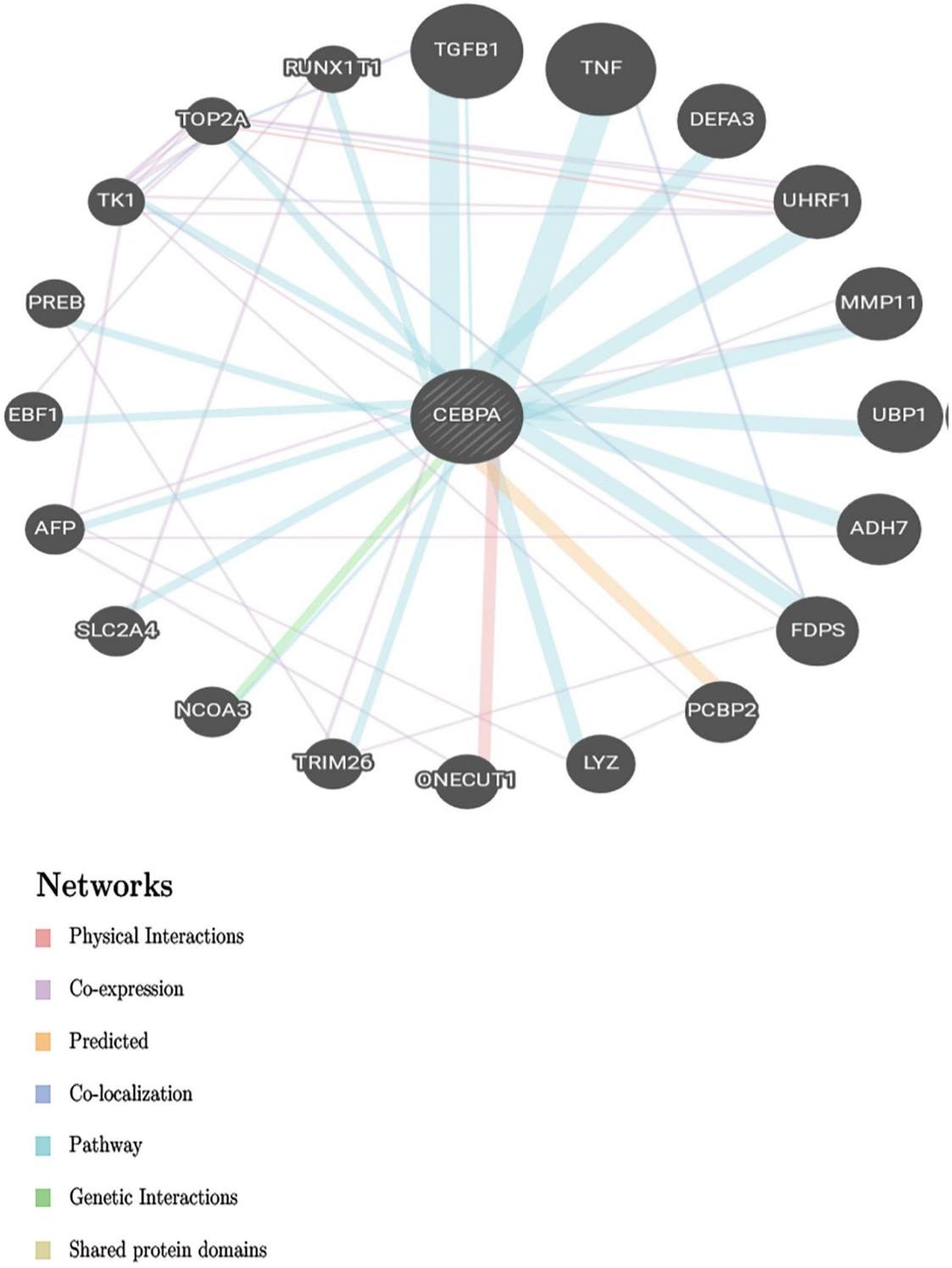


**Figure 13.** Interaction between *CEBPA* and its related genes.

This study is the first in silico analysis while all other previous studies were next-generation sequencing (NGS) analysis, in vitro analysis, and in vivo analysis[14,71,82,83]; also, it is the first computational analysis, which revealed that 5 SNPs were identified as highly deleterious in the coding region, whereas 11 SNPs were detected to be damaging in the 3′-UTR, and therefore, may be used as diagnostic markers for AML and might create an ideal target for cancer therapy. These outcomes in combination with all earlier discoveries make AML a model for understanding the philosophies of cancer development.[84,85] Finally, clinical techniques are recommended to support these findings.

**Table 4.** The *CEBPA* gene functions and its appearance in network and genome.

| FUNCTION | FDR | GENES IN NETWORK | GENES IN GENOME |
|---|---|---|---|
| Regulation of multiorganism process | 0.375211393 | 4 | 216 |
| Response to bacterium | 0.375211393 | 4 | 167 |
| Golgi lumen | 0.375211393 | 3 | 84 |
| Viral genome replication | 0.375211393 | 3 | 64 |
| Negative regulation of multiorganism process | 0.375211393 | 3 | 79 |
| Negative regulation of cytokine production | 0.698836286 | 3 | 111 |
| Negative regulation of fat cell differentiation | 0.698836286 | 2 | 22 |
| Regulation of viral process | 0.828689471 | 3 | 134 |
| Protein import into nucleus, translocation | 0.828689471 | 2 | 26 |
| Lipopolysaccharide-mediated signaling pathway | 0.829247433 | 2 | 30 |

Abbreviation: FDR, false discovery rate.
FDR is greater than or equal to the probability that this is a false positive.

**Table 5.** The gene co-expression, shared domain, and interaction with CEBPA gene network.

| GENE 1 | GENE 2 | WEIGHT | NETWORK GROUP |
|---|---|---|---|
| *TK1* | *PCBP2* | 0.00530806 | Co-expression |
| *RUNX1T1* | *SLC2A4* | 0.02216304 | Co-expression |
| *TRIM26* | *CEBPA* | 0.01914615 | Co-expression |
| *TOP2A* | *TK1* | 0.00946418 | Co-expression |
| *AFP* | *CEBPA* | 0.01361087 | Co-expression |
| *TK1* | *AFP* | 0.02290919 | Co-expression |
| *MMP11* | *CEBPA* | 0.00309695 | Co-expression |
| *AFP* | *ADH7* | 0.00763638 | Co-expression |
| *TK1* | *FDPS* | 0.00380695 | Co-expression |
| *TOP2A* | *TK1* | 0.00417671 | Co-expression |
| *TOP2A* | *TK1* | 0.004765 | Co-expression |
| *TOP2A* | *TK1* | 0.00795774 | Co-expression |
| *TOP2A* | *UHRF1* | 0.00966428 | Co-expression |
| *TOP2A* | *TK1* | 0.00797961 | Co-expression |
| *MMP11* | *CEBPA* | 0.00799819 | Co-expression |
| *TOP2A* | *TK1* | 0.00777507 | Co-expression |
| *TK1* | *UHRF1* | 0.01041377 | Co-expression |
| *TOP2A* | *UHRF1* | 0.01174308 | Co-expression |
| *LYZ* | *PCBP2* | 0.01229207 | Co-expression |
| *AFP* | *LYZ* | 0.0059269 | Co-expression |
| *AFP* | *ONECUT1* | 0.01915715 | Co-expression |
| *PREB* | *TRIM26* | 0.00768956 | Co-expression |
| *TRIM26* | *FDPS* | 0.01174667 | Co-expression |
| *RUNX1T1* | *EBF1* | 0.00970602 | Co-expression |
| *TK1* | *UHRF1* | 0.00909379 | Co-expression |

*(Continued)*

**Table 5.** (Continued)

| GENE 1 | GENE 2 | WEIGHT | NETWORK GROUP |
|---|---|---|---|
| *TOP2A* | *UHRF1* | 0.01288607 | Co-expression |
| *TOP2A* | *TK1* | 0.00770565 | Co-expression |
| *FDPS* | *TNF* | 0.00918617 | Co-localization |
| *TOP2A* | *TGFB1* | 0.00832004 | Co-localization |
| *TOP2A* | *FDPS* | 0.0065701 | Co-localization |
| *TOP2A* | *TK1* | 0.00696822 | Co-localization |
| *NCOA3* | *CEBPA* | 0.29423913 | Genetic interactions |
| *TGFB1* | *CEBPA* | 0.00700947 | Pathway |
| *DEFA3* | *CEBPA* | 0.17928578 | Pathway |
| *UHRF1* | *CEBPA* | 0.17928578 | Pathway |
| *MMP11* | *CEBPA* | 0.17928578 | Pathway |
| *UBP1* | *CEBPA* | 0.17928578 | Pathway |
| *ADH7* | *CEBPA* | 0.17928578 | Pathway |
| *FDPS* | *CEBPA* | 0.17928578 | Pathway |
| *TRIM26* | *CEBPA* | 0.07640404 | Pathway |
| *AFP* | *CEBPA* | 0.07640404 | Pathway |
| *PREB* | *CEBPA* | 0.07618967 | Pathway |
| *TK1* | *CEBPA* | 0.07600352 | Pathway |
| *TOP2A* | *CEBPA* | 0.07955996 | Pathway |
| *RUNX1T1* | *CEBPA* | 0.07140907 | Pathway |
| *TGFB1* | *CEBPA* | 0.3443214 | Pathway |
| *TNF* | *CEBPA* | 0.3443214 | Pathway |
| *NCOA3* | *CEBPA* | 0.01949818 | Pathway |
| *SLC2A4* | *CEBPA* | 0.10226673 | Pathway |
| *EBF1* | *CEBPA* | 0.08983881 | Pathway |
| *LYZ* | *CEBPA* | 0.3277232 | Pathway |
| *TOP2A* | *UHRF1* | 0.02391628 | Physical interactions |
| *ONECUT1* | *CEBPA* | 1 | Physical interactions |
| *PCBP2* | *CEBPA* | 0.6985996 | Physical interactions |

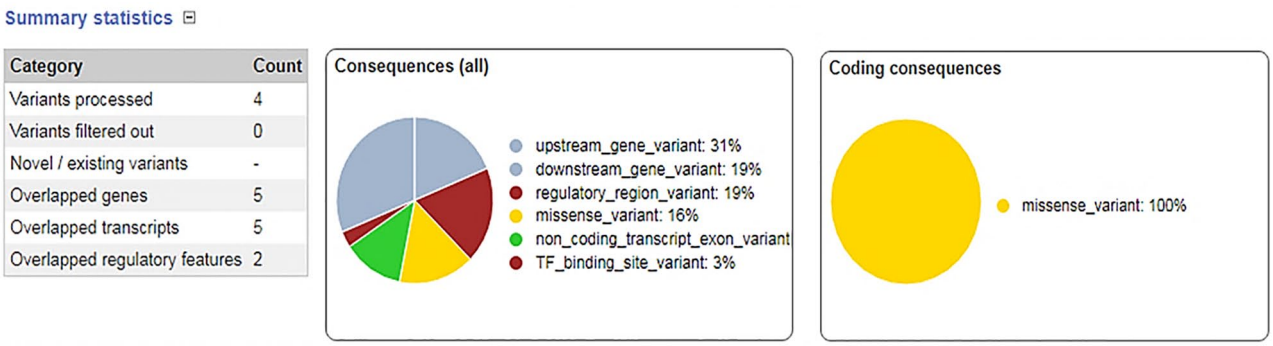


**Figure 14.** Summary pie charts and statistics.

**Table 6.** Shows variant consequences, transcripts, and regulatory features by VEP tool.

| UPLOADED VARIATION | LOCATION | ALLELE | CONSEQUENCE | SYMBOL | GENE | FEATURE | PROTEIN POSITION | AMINO ACIDS |
|---|---|---|---|---|---|---|---|---|
| rs1455027551 | 19:33301400-33301400 | A | Missense variant | CEBPA | ENSG00000245848 | ENST00000498907.3 | 339 | R/W |
| – | – | A | Non-coding transcript exon variant | AC008738.5 | ENSG00000267727 | ENST00000587312.1 | – | – |
| – | – | A | Downstream gene variant | AC008738.3 | ENSG00000267580 | ENST00000589932.1 | – | – |
| – | – | A | Upstream gene variant | CEBPA-DT | ENSG00000267296 | ENST00000592982.2 | – | – |
| – | – | A | Upstream gene variant | AC008738.2 | ENSG00000267130 | ENST00000593041.1 | – | – |
| – | – | A | Regulatory region variant | – | – | ENSR00000588478 | – | – |
| rs776590829 | 19:33301540-33301540 | C | Missense variant | CEBPA | ENSG00000245848 | ENST00000498907.3 | 292 | N/S |
| – | – | G | missense variant | CEBPA | ENSG00000245848 | ENST00000498907.3 | 292 | N/T |
| – | – | C | Non-coding transcript exon variant | AC008738.5 | ENSG00000267727 | ENST00000587312.1 | – | – |
| – | – | G | Non-coding transcript exon variant | AC008738.5 | ENSG00000267727 | ENST00000587312.1 | – | – |
| – | – | C | Downstream gene variant | AC008738.3 | ENSG00000267580 | ENST00000589932.1 | – | – |
| – | – | G | Downstream gene variant | AC008738.3 | ENSG00000267580 | ENST00000589932.1 | – | – |
| – | – | C | Upstream gene variant | CEBPA-DT | ENSG00000267296 | ENST00000592982.2 | – | – |
| – | – | G | Upstream gene variant | CEBPA-DT | ENSG00000267296 | ENST00000592982.2 | – | – |
| – | – | C | Upstream gene variant | AC008738.2 | ENSG00000267130 | ENST00000593041.1 | – | – |
| – | – | G | Upstream gene variant | AC008738.2 | ENSG00000267130 | ENST00000593041.1 | – | – |
| – | – | C | Regulatory region variant | – | – | ENSR00000588478 | – | – |
| – | – | G | Regulatory region variant | – | – | ENSR00000588478 | – | – |
| rs1064794962 | 19:33301552-33301552 | G | Missense variant | CEBPA | ENSG00000245848 | ENST00000498907.3 | 288 | R/P |

*(Continued)*

**Table 6.** (Continued)

| UPLOADED VARIATION | LOCATION | ALLELE | CONSEQUENCE | SYMBOL | GENE | FEATURE | PROTEIN POSITION | AMINO ACIDS |
|---|---|---|---|---|---|---|---|---|
| – | – | G | Non-coding transcript exon variant | AC008738.5 | ENSG00000267727 | ENST00000587312.1 | – | – |
| – | – | G | Downstream gene variant | AC008738.3 | ENSG00000267580 | ENST00000589932.1 | – | – |
| – | – | G | Upstream gene variant | CEBPA-DT | ENSG00000267296 | ENST00000592982.2 | – | – |
| – | – | G | Upstream gene variant | AC008738.2 | ENSG00000267130 | ENST00000593041.1 | – | – |
| – | – | G | Regulatory region variant | – | – | ENSR00000588478 | – | – |
| rs1452063514 | 19:33302228-33302228 | T | Missense variant | CEBPA | ENSG00000245848 | ENST00000498907.3 | 63 | D/N |
| – | – | T | Downstream gene variant | AC008738.5 | ENSG00000267727 | ENST00000587312.1 | – | – |
| – | – | T | Downstream gene variant | AC008738.3 | ENSG00000267580 | ENST00000589932.1 | – | – |
| – | – | T | Upstream gene variant | CEBPA-DT | ENSG00000267296 | ENST00000592982.2 | – | – |
| – | – | T | Upstream gene variant | AC008738.2 | ENSG00000267130 | ENST00000593041.1 | – | – |
| – | – | T | Regulatory region variant | – | – | ENSR00000588478 | – | – |
| – | – | T | Regulatory region variant | – | – | ENSR00000588491 | – | – |
| – | – | T | TF binding site variant | – | – | ENSM00524410754 | – | – |

Abbreviations: TF, transcription factor; VEP, Variant Effect Predictor.

**Table 7.** SNPs and INDELs in miRNA target sites in *CEBPA* gene.

| LOCATION | dbSNP ID | miR ID | CONSERVATION | miRSITE | FUNCTION CLASS | CONTEXT + SCORE CHANGE |
|---|---|---|---|---|---|---|
| 33791117 | rs116528776 | hsa-miR-548aa | 6 | tcctttTGGTTTT | D | −0.032 |
| | | hsa-miR-548ap-3p | 6 | tcctttTGGTTTT | D | −0.032 |
| | | hsa-miR-548t-3p | 6 | tcctttTGGTTTT | D | −0.032 |
| | | hsa-miR-186-3p | 9 | tcCTTTGGGtttt | C | −0.11 |
| | | hsa-miR-548as-3p | 6 | tcctttGGGTTTT | C | −0.079 |
| 33791158 | rs113670631 | hsa-miR-5192 | 7 | ACTCTCCgtcggc | D | −0.103 |
| | | hsa-miR-5739 | 7 | aCTCTCCGtcggc | D | −0.234 |
| 33791211 | rs34017519 | hsa-miR-548aa | 15 | tttattTGGTTTT | D | −0.12 |
| | | hsa-miR-548ap-3p | 15 | tttattTGGTTTT | D | −0.12 |
| | | hsa-miR-548t-3p | 15 | tttattTGGTTTT | D | −0.12 |
| | | hsa-miR-548at-3p | 15 | tttattCGGTTTT | C | −0.179 |
| | | hsa-miR-548ay-3p | 15 | tttattCGGTTTT | C | −0.169 |
| 33791239 | rs146104564 | hsa-miR-3939 | 21 | aatgTGCGCGTct | D | −0.344 |
| 33791250 | rs187516157 | hsa-miR-130a-5p | 11 | tgtgcAATGTGAA | D | −0.06 |
| | | hsa-miR-23a-3p | 11 | tgtgcAATGTGAA | D | −0.06 |
| | | hsa-miR-23b-3p | 11 | tgtgcAATGTGAA | D | −0.06 |
| | | hsa-miR-23c | 11 | tgtgcAATGTGAA | D | −0.039 |
| | | hsa-miR-25-3p | 12 | tGTGCAATgtgaa | D | −0.092 |
| | | hsa-miR-32-5p | 12 | tGTGCAATgtgaa | D | −0.102 |
| | | hsa-miR-363-3p | 12 | tGTGCAATgtgaa | D | −0.092 |
| | | hsa-miR-367-3p | 12 | tGTGCAATgtgaa | D | −0.111 |
| | | hsa-miR-92a-3p | 12 | tGTGCAATgtgaa | D | −0.083 |
| | | hsa-miR-92b-3p | 12 | tGTGCAATgtgaa | D | −0.074 |
| | | hsa-miR-513b-5p | 15 | tgtgcaTTGTGAA | C | 0.006 |
| 33791649 | rs2376497 | hsa-miR-1233-5p | 4 | acgcctCTCCCAC | D | −0.082 |
| | | hsa-miR-30c-1-3p | 4 | acgccTCTCCCAc | D | 0.01 |
| | | hsa-miR-30c-2-3p | 4 | acgccTCTCCCAc | D | 0.01 |
| | | hsa-miR-6731-5p | 3 | acgcCTCTCCCAc | D | −0.143 |
| | | hsa-miR-6778-5p | 4 | acgcctCTCCCAC | D | −0.101 |
| | | hsa-miR-6788-5p | 4 | acgccTCTCCCAc | D | −0.018 |
| | | hsa-miR-6878-5p | 4 | acgccTCTCCCAc | D | −0.04 |
| | | hsa-miR-8085 | 3 | acgcCTCTCCCAc | D | −0.143 |
| | | hsa-miR-3153 | 3 | acgcCTTTCCCAc | C | −0.046 |
| | | hsa-miR-4484 | 2 | aCGCCTTTcccac | C | −0.073 |
| | | hsa-miR-4668-5p | 4 | acgccTTTCCCAc | C | 0.033 |
| | | hsa-miR-6733-5p | 3 | acgcCTTTCCCAc | C | −0.067 |

*(Continued)*

**Table 7.** (Continued)

| LOCATION | dbSNP ID | miR ID | CONSERVATION | miRSITE | FUNCTION CLASS | CONTEXT + SCORE CHANGE |
|---|---|---|---|---|---|---|
| | | hsa-miR-6739-5p | 3 | acgcCTTTCCCAc | C | −0.078 |
| 33791665 | rs192371350 | | | | | |
| | | hsa-miR-6508-5p | 21 | gagggTTTCTAGt | C | 0.048 |
| | | hsa-miR-8067 | 21 | gagggTTTCTAGt | C | 0.058 |
| 33791807 | rs1049969 | hsa-miR-1233-3p | 2 | aggaggAGGGCTC | D | −0.167 |
| | | hsa-miR-4290 | 2 | aggaGGAGGGCtc | D | −0.177 |
| | | hsa-miR-4667-3p | 2 | aggAGGAGGGctc | D | −0.035 |
| | | hsa-miR-4687-5p | 2 | aggagGAGGGCTc | D | −0.12 |
| | | hsa-miR-5193 | 2 | agGAGGAGGgctc | D | −0.083 |
| | | hsa-miR-660-3p | 9 | AGGAGGAgggctc | D | −0.143 |
| | | hsa-miR-1225-3p | 2 | aggaggGGGGCTC | C | −0.204 |
| | | hsa-miR-3943 | 2 | aggagGGGGGCTc | C | −0.163 |
| | | hsa-miR-6887-3p | 3 | aGGAGGGGggctc | C | −0.117 |
| 33791883 | rs41367646 | hsa-miR-519d-5p | 9 | taTTTGGAGgttt | D | −0.115 |
| | | hsa-miR-3671 | 7 | TATTTGAAggttt | C | −0.051 |
| | | hsa-miR-607 | 9 | tATTTGAAggttt | C | 0.016 |
| 33792082 | rs184965384 | hsa-miR-2467-3p | 4 | cccttCCTCTGCg | D | −0.162 |
| | | hsa-miR-3125 | 6 | cccTTCCTCTgcg | D | −0.003 |
| | | hsa-miR-3202 | 4 | CCCTTCCtctgcg | D | −0.13 |
| | | hsa-miR-3916 | 6 | cccTTCCTCTgcg | D | 0.016 |
| | | hsa-miR-4476 | 7 | cCCTTCCTctgcg | D | −0.096 |
| | | hsa-miR-6847-5p | 6 | ccctTCCTCTGcg | D | −0.109 |
| | | hsa-miR-6859-5p | 6 | cccTTCCTCTgcg | D | 0.006 |
| | | hsa-miR-6876-5p | 7 | cCCTTCCTctgcg | D | −0.059 |
| | | hsa-miR-298 | 4 | cccttCTTCTGCg | C | −0.146 |
| | | hsa-miR-3154 | 4 | CCCTTCTtctgcg | C | −0.131 |
| | | hsa-miR-3185 | 5 | ccCTTCTTCtgcg | C | 0.017 |
| 33792099 | rs187751931 | hsa-miR-3186-3p | 7 | tgctCCGCGTGtc | D | −0.314 |
| | | hsa-miR-151a-5p | 7 | tgCTCCTCGtgtc | C | −0.204 |
| | | hsa-miR-151b | 7 | tgCTCCTCGtgtc | C | −0.204 |

Abbreviations: miRSite, microRNA site; SNPs, single nucleotide polymorphisms.
D: the derived allele disrupts a conserved miRNA site (ancestral allele with support ⩾ 2). C: the derived allele creates a new miRNA site.

## Conclusions

In this study, the impact of functional mutations in the CEBPA gene was investigated through different bioinformatics analysis techniques, which determined that R339W, R288P, N292S, N292T, and D63N are pathogenic mutations, which have a possible functional influence, and therefore, can be used as diagnostic markers and may assist in genetic studies with a special consideration of the large heterogeneity of AML among different populations. In addition, this study draws attention to 11 SNPs that were identified to be deleterious in the 3′-UTR.

## Acknowledgements

The authors wish to acknowledge the enthusiastic cooperation of Africa City of Technology, Sudan.

## Author Contributions

MIM and ZOM helped in data curation, methodology, also conceptualized the data, formal analysed the data, illustrated, validated and wrote the manuscript and also helped in drafting the original manuscript. NSM, NME, and AHA helped in data curation, methodology of the data, and also formal analysed the data. MAH conceptualized and validated the manuscript, helped in reviewing & editing, project administration of the data, and also supervised the manuscript.

## Data Availability

All data underlying the results are available as part of the article, and no additional source data are required.

## ORCID iD

Mujahed I Mustafa https://orcid.org/0000-0001-6893-0536

## REFERENCES

1. Chaudhury S, O'Connor C, Canete A, et al. Age-specific biological and molecular profiling distinguishes paediatric from adult acute myeloid leukaemias. *Nat Commun*. 2018;9:5280.
2. Liew E, Owen C. Familial myelodysplastic syndromes: a review of the literature. *Haematologica*. 2011;96:1536-1542.
3. Short NJ, Rytting ME, Cortes JE. Acute myeloid leukaemia. *Lancet (London, England)*. 2018;392:593-606.
4. Falini B, Martelli MP. Impact of genomics in the clinical management of patients with cytogenetically normal acute myeloid leukemia. *Best Pract Res Clin Haematol*. 2015;28:90-97.
5. Masetti R, Castelli I, Astolfi A, et al. Genomic complexity and dynamics of clonal evolution in childhood acute myeloid leukemia studied with whole-exome sequencing. *Oncotarget*. 2016;7:56746-56757.
6. De Kouchkovsky I, Abdul-Hay M. Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer J*. 2016;6:e441.
7. Hirsch CM, Przychodzen BP, Radivoyevitch T, et al. Molecular features of early onset adult myelodysplastic syndrome. *Haematologica*. 2017;102:1028-1034.
8. Tomizawa D. Acute leukemia in adolescents and young adults. *Rinsho ketsueki*. 2017;58:2160-2167.
9. Dohner K, Dohner H. Molecular characterization of acute myeloid leukemia. *Haematologica*. 2008;93:976-982.
10. Akin DF, Oner DA, Kurekci E, Akar N. Determination of CEBPA mutations by next generation sequencing in pediatric acute leukemia. *Bratisl lek listy*. 2018;119:366-372.
11. Green CL, Koo KK, Hills RK, Burnett AK, Linch DC, Gale RE. Prognostic significance of CEBPA mutations in a large cohort of younger adult patients with acute myeloid leukemia: impact of double CEBPA mutations and the interaction with FLT3 and NPM1 mutations. *J Clin Oncol*. 2010;28:2739-2747.
12. Steffen B, Muller-Tidow C, Schwable J, Berdel WE, Serve H. The molecular pathogenesis of acute myeloid leukemia. *Crit Rev Oncol Hematol*. 2005;56:195-221.
13. Kato N. Analysis of leukemogenesis induced by C/EBPalpha mutations. *Rinsho ketsueki*. 2011;52:320-328.
14. Mueller BU, Pabst T. C/EBPalpha and the pathophysiology of acute myeloid leukemia. *Curr Opin Hematol*. 2006;13:7-14.
15. Leroy H, Roumier C, Huyghe P, Biggio V, Fenaux P, Preudhomme C. CEBPA point mutations in hematological malignancies. *Leukemia*. 2005;19:329-334.
16. Kasakura K, Takahashi K, Itoh T, et al. C/EBPalpha controls mast cell function. *FEBS Lett*. 2014;588:4645-4653.
17. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood*. 2009;113:3088-3091.
18. Pabst T, Mueller BU. Complexity of CEBPA dysregulation in human acute myeloid leukemia. *Clin Cancer Res*. 2009;15:5303-5307.
19. van Vliet MH, Burgmer P, de Quartel L, et al. Detection of CEBPA double mutants in acute myeloid leukemia using a custom gene expression array. *Genet Test Mol Biomarkers*. 2013;17:395-400.
20. Schwieger M, Lohler J, Fischer M, Herwig U, Tenen DG, Stocking C. A dominant-negative mutant of C/EBPalpha, associated with acute myeloid leukemias, inhibits differentiation of myeloid and erythroid progenitors of man but not mouse. *Blood*. 2004;103:2744-2752.
21. Asou H, Gombart AF, Takeuchi S, et al. Establishment of the acute myeloid leukemia cell line Kasumi-6 from a patient with a dominant-negative mutation in the DNA-binding region of the C/EBPalpha gene. *Genes Chromosomes Cancer*. 2003;36:167-174.
22. Schuster MB, Porse BT. C/EBPalpha in leukemogenesis: identity and origin of the leukemia-initiating cell. *Biofactors*. 2009;35:227-231.
23. Shiba N, Yoshida K, Shiraishi Y, et al. Whole-exome sequencing reveals the spectrum of gene mutations and the clonal evolution patterns in paediatric acute myeloid leukaemia. *Br J Haematol*. 2016;175:476-489.
24. Vinhas R, Tolmatcheva A, Canto R, et al. A novel mutation in CEBPA gene in a patient with acute myeloid leukemia. *Leuk Lymphoma*. 2016;57:711-713.
25. Du Y, Fryzek J, Sekeres MA, Taioli E. Smoking and alcohol intake as risk factors for myelodysplastic syndromes (MDS). *Leuk Res*. 2010;34:1-5.
26. Avgerinou C, Giannezi I, Theodoropoulou S, et al. Occupational, dietary, and other risk factors for myelodysplastic syndromes in Western Greece. *Hematology (Amsterdam, Netherlands)*. 2017;22:419-429.
27. Strom SS, Velez-Bravo V, Estey EH. Epidemiology of myelodysplastic syndromes. *Semin Hematol*. 2008;45:8-13.
28. Holme H, Hossain U, Kirwan M, Walne A, Vulliamy T, Dokal I. Marked genetic heterogeneity in familial myelodysplasia/acute myeloid leukaemia. *Br J Haematol*. 2012;158:242-248.
29. Duployez N, Lejeune S, Renneville A, Preudhomme C. Myelodysplastic syndromes and acute leukemia with genetic predispositions: a new challenge for hematologists. *Expert Rev Hematol*. 2016;9:1189-1202.
30. Gaidzik V, Dohner K. Prognostic implications of gene mutations in acute myeloid leukemia with normal cytogenetics. *Semin Oncol*. 2008;35:346-355.
31. Bienz M, Ludwig M, Leibundgut EO, et al. Risk assessment in patients with acute myeloid leukemia and a normal karyotype. *Clin Cancer Res*. 2005;11:1416-1424.
32. Ito S, Fujiwara SI, Mashima K, et al. Development of acute myeloid leukemia in patients with untreated chronic lymphocytic leukemia. *Ann Hematol*. 2017;96:719-724.
33. Tanaka M, Ogasawara H, Nakagawa S, et al. Effective treatment of a case of acute myeloid leukemia with advanced esophageal cancer. *Gan to Kagaku Ryoho*. 2016;43:1405-1408.
34. Schlenk RF, Dohner K, Krauter J, et al. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *N Engl J Med*. 2008;358:1909-1918.
35. Kucukkal TG, Yang Y, Chapman SC, Cao W, Alexov E. Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics. *Int J Mol Sci*. 2014;15:9670-9717.
36. Khan IA, Mort M, Buckland PR, O'Donovan MC, Cooper DN, Chuzhanova NA. In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity. *In Silico Biol*. 2006;6:23-34.
37. Shaw G. Polymorphism and single nucleotide polymorphisms (SNPs). *BJU Int*. 2013;112:664-665.
38. Qu HQ, Lawrence SG, Guo F, Majewski J, Polychronakos C. Strand bias in complementary single-nucleotide polymorphisms of transcribed human sequences: evidence for functional effects of synonymous polymorphisms. *BMC Genom*. 2006;7:213.
39. Qiu P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem Biophys Res Commun*. 2003;309:495-501.
40. Rost B. Twilight zone of protein sequence alignments. *Protein Eng*. 1999;12:85-94.
41. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11:863-874.
42. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol*. 2002;322:891-901.
43. Pei J, Grishin NV. Combining evolutionary and structural information for local protein structure prediction. *Proteins*. 2004;56:782-794.
44. Yue P, Moult J. Identification and analysis of deleterious human SNPs. *J Mol Biol*. 2006;356:1263-1274.
45. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics (Oxford, England)*. 2006;22:2729-2734.

46. Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet*. 2007;52:871-880.
47. Chaudhary R, Singh B, Kumar M, et al. Role of single nucleotide polymorphisms in pharmacogenomics and their association with human diseases. *Drug Metab Rev*. 2015;47:281-290.
48. Katara P. Single nucleotide polymorphism and its dynamics for pharmacogenomics. *Interdiscip Sci*. 2014;6:85-92.
49. Alwi ZB. The use of SNPs in pharmacogenomics studies. *Malays J Med Sci*. 2005;12:4-12.
50. Ahn TJ, Park K, Son DS, et al. Selecting SNPs for pharmacogenomic association study. *Int J Data Min Bioinform*. 2012;6:521-534.
51. McCarthy JJ, Hilfiker R. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat Biotechnol*. 2000;18:505-508.
52. Lonetti A, Fontana MC, Martinelli G, Iacobucci I. Single nucleotide polymorphisms as genomic markers for high-throughput pharmacogenomic studies. *Methods Mol Biol (Clifton, N.J.)*. 2016;1368:143-159.
53. Tenenbaum JD. Translational bioinformatics: past, present, and future. *Genomics Proteomics Bioinformatics*. 2016;14:31-41.
54. Vamathevan J, Birney E. A review of recent advances in translational bioinformatics: bridges from biology to medicine. *Yearb Med Inform*. 2017;26:178-187.
55. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2017;45: D158-D169.
56. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40:W452-457.
57. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics*. 2011;12:S3.
58. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*. 2012;7:e46688.
59. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat*. 2009;30:1237-1244.
60. Lopez-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpi JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res*. 2017;45:W222-W228.
61. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 2005;33:W306-W310.
62. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*. 2006;62:1125-1132.
63. Meyer MJ, Lapcevic R, Romero AE, et al. mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum Mutat*. 2016;37:447-456.
64. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. 2016;44:W344-350.
65. Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res*. 2016;44:W430-W435.
66. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera —a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25: 1605-1612.
67. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38:W214-W220.
68. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17:122.
69. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res*. 2014;42:D86-D91.
70. Matsuo H, Kajihara M, Tomizawa D, et al. Prognostic implications of CEBPA mutations in pediatric acute myeloid leukemia: a report from the Japanese Pediatric Leukemia/Lymphoma Study Group. *Blood Cancer J*. 2014;4:e226.
71. Tawana K, Rio-Machin A, Preudhomme C, Fitzgibbon J. Familial CEBPA-mutated acute myeloid leukemia. *Semin Hematol*. 2017;54:87-93.
72. Mannelli F, Ponziani V, Bencini S, et al. CEBPA-double-mutated acute myeloid leukemia displays a unique phenotypic profile: a reliable screening method and insight into biological features. *Haematologica*. 2017;102:529-540.
73. Roberts SA, Gordenin DA. Clustered and genome-wide transient mutagenesis in human cancers: hypermutation without permanent mutators or loss of fitness. *Bioessays*. 2014;36:382-393.
74. Rojano E, Seoane P, Ranea JAG, Perkins JR. Regulatory variants: from detection to predicting impact [published online ahead of print June 8,2018]. *Brief Bioinform*. doi:10.1093/bib/bby039.
75. Dhamija S, Menon MB. Non-coding transcript variants of protein-coding genes—what are they good for? *RNA Biol*. 2018;15:1025-1031.
76. Tatarinova TV, Chekalin E, Nikolsky Y, et al. Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep*. 2016;6:35730.
77. Cheng SJ, Jiang S, Shi FY, Ding Y, Gao G. Systematic identification and annotation of multiple-variant compound effects at transcription factor binding sites in human genome. *J Genet Genomics*. 2018;45:373-379.
78. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*. 2010;7:575-576.
79. Wong KC, Zhang Z. SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. *Bioinformatics (Oxford, England)*. 2014;30:1112-1119.
80. Gao J, Sundstrom G, Moghadam BT, Zamani N, Grabherr MG. ACES: a machine learning toolbox for clustering analysis and visualization. *BMC Genom*. 2018;19:964.
81. Reuter K, Biehl A, Koch L, Helms V. PreTIS: a tool to predict non-canonical 5' UTR translational initiation sites in human and mouse. *Plos Comput Biol*. 2016;12:e1005170.
82. Harada H, Harada Y. Recent advances in myelodysplastic syndromes: molecular pathogenesis and its implications for targeted therapies. *Cancer Sci*. 2015;106: 329-336.
83. Pituch-Noworolska A. Biological properties and sensitivity to induction therapy of differentiated cells expressing atypical immunophenotype in acute leukemia of children. *Folia Med Cracov*. 2001;42:5-80.
84. Grove CS, Vassiliou GS. Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer? *Dis Model Mech*. 2014;7:941-951.
85. Haider M, Duncavage EJ, Afaneh KF, Bejar R, List AF. New insight into the biology, risk stratification, and targeted treatment of myelodysplastic syndromes. *Am Soc Clin Oncol Educ Book*. 2017;37:480-494.