

Analysing the Pathogenicity of KLF6 Variants in Breast Cancer



By

Zainab Nisar Qureshi

(Registration No: 00000400598)

Department of Biomedicine

Atta-ur-Rahman School of Applied Biosciences (ASAB)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Analysing the Pathogenicity of KLF6 Variants in Breast Cancer



By

Zainab Nisar Qureshi

(Registration No: 00000400598)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in

Biomedicine

Supervisor: Dr. Yasmin Badshah

Co Supervisor: Dr. Maria Shabbir

Department of Biomedicine

Atta-ur-Rahman School of Applied Biosciences (ASAB)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)



FORM TH-4

National University of Sciences & Technology

MS THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: Zainab Nisar Qureshi Reg No. 00000400598 Titled: Analyzing the Pathogenicity of KLF6 Variants in Breast Cancer be accepted in partial fulfillment of the requirements for the award of MS Degree in Healthcare biotechnology degree with (A grade).

Examination Committee Members

- | | |
|------------------------------------|-------------------------------|
| 1. Name: <u>Dr. Peter John</u> | Signature: <u>[Signature]</u> |
| 2. Name: <u>Dr. Rumeza Hanif</u> | Signature: <u>[Signature]</u> |
| 3. Name: <u>Dr. Hashaam Akhtar</u> | Signature: <u>[Signature]</u> |

Co-Supervisor's name: Dr. Maria Shabbir Signature: [Signature]

Supervisor's name: Dr. Yasmin Badshah Signature: [Signature]
Date: 24/7/24

Date: 24/7/24

[Signature]
Dr. Hussain Azeem Wahedi
Head of Department

COUNTERSIGNED

Date: 24/7/24

[Signature]
A/Principal & Dean
Abul-Rahman School of Applied
Sciences (ASAS), NUST Islamabad
Dean/Principal


THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by Ms. Zainab Nisar Qureshi (Registration No. 00000400598), of ASAB has been vetted by undersigned, found complete in all respects as per NUST Statutes/ Regulations/ Masters Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of Master's degree. It is further certified that necessary amendments as point out by GEC members and evaluators of the scholar have also been incorporated in the said thesis.

Signature:  **Dr. Yasmin Badshah**
Assistant Professor
Atta-ur-Rahman School of
Applied Biosciences (ASAB)
NUST Islamabad

Name of Supervisor: Dr. Yasmin Badshah

Date: 07/8/24

Signature (HOD):  **Dr. Hussain Mustatab Wahedi**
Head of Department (HOD)
Department of Biotechnology
Atta-ur-Rahman School of Applied
Biosciences (ASAB), NUST Islamabad

Date: 7/8/24

Signature (Dean/ Principal):  **A/Principal & Dean**
Atta-ur-Rahman School of Applied
Biosciences (ASAB), NUST Islamabad

Date: 7/8/24

AUTHOR'S DECLARATION

I Zainab Nisar Qureshi hereby state that my MS thesis titled “**Analysing the Pathogenicity of KLF6 Variants in Breast Cancer**” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Zainab Nisar Qureshi

Date: 15/7/24

CERTIFICATE FOR PLAGIARISM

It is to confirm that MS thesis entitled titled "Analysing the Pathogenicity of KLF6 Variants in Breast Cancer" of Zainab Nisar Qureshi Reg No. 00000400598 has been examined by me. I undertake that,

1. Thesis has significant new work/knowledge as compared to already elsewhere. No sentence, table, equation, diagram, paragraph, or section has been copied verbatim from previous work except when placed under quotation marks and duly referenced.
2. The work presented is original and own work of the author i.e., there is no plagiarism. No idea, results or work of others have been presented as author's own work.
3. There is no fabrication of data or results such that the research is not accurately represented in the records. The thesis has been checked using Turnitin, a copy of the original report attached and focused within the limits as per HEC plagiarism policy and instruction based on time to time.


Dr. Yasmin Badshah
Assistant Professor
Atta-ur-Rahman School of
Applied Biosciences (ASAB)
NUST Islamabad

(Supervisor) Dr. Yasmin Badshah

Biomedicine

ASAB, NUST

Dedicated To My Parents, Teachers, Siblings & Friends

ACKNOWLEDGMENTS

In the name of Allah, the Most Merciful, the Most Kind. All praises and thanks to Allah Almighty, who has bestowed His countless blessings and gifts upon me. Without His mercy, I would not have been able to complete this thesis.

I would like to express my sincere gratitude to my supervisor, Dr. Yasmin Badshah, and co-supervisor, Dr. Maria Shabbir, for their guidance, motivation, and unwavering support, which made this project possible. Their expertise and encouragement have been invaluable throughout this journey. I extend my heartfelt regards to Dr. Hashaam Akhtar and my seniors, especially Ms. Sameen Zafar, Ms. Amna Hafeez, Ms. Aneela Mustafa for their support and expertise in the project. Their guidance has played a crucial role in shaping this research. I would also like to thank my Research Fellows for their constant support, encouragement, and assistance throughout the project. Lastly, I would like to express my deep gratitude to my parents, my siblings, and all my friends especially, Urva Tariq, Noor ul Huda, Qamar Pirzada, who have made my research journey even more fulfilling and memorable.

In conclusion, I am truly grateful to Allah Almighty and all the individuals mentioned above for their contributions and support. Their presence in my life has made this academic journey rewarding. Thank you.

Zainab Nisar Qureshi

MS Healthcare Biotechnology

2022-2024

TABLE OF CONTENTS

ACKNOWLEDGMENTS	VI
TABLE OF CONTENTS	VII
LIST OF TABLES	X
LIST OF FIGURES	XI
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS	XIII
ABSTRACT	XIV
CHAPTER 1: INTRODUCTION	1
1.1 Breast Anatomy	2
1.2 Breast Cancer	2
1.3 Epidemiology	3
1.4 Classification of Breast Cancer	3
1.4.1 Histological Classification	3
1.4.2 Molecular Classification	3
1.5 Risk Factors and Prevention	4
1.6 Stages of Cancer	4
1.7 Prognosis	5
1.8 Treatment	5
1.9 Cancer	6
CHAPTER 2: LITERATURE REVIEW	8
2.1 History of KLF	8
2.2 KLF6 Structure	9
2.3 KLF6 Protein Structures	11
2.3.1 Conservation	11
2.3.2 Expression	12
2.3.3 Phylogenetic Analysis	12
2.4 Structure of KLFs	13

2.4.1	Zinc Finger Domain	13
2.4.2	Functional Binding Domain	14
2.5	CtBP-Binding Site	14
2.6	Sin3A-Binding Site	15
2.7	KLFs in Cell Proliferation	16
2.7.1	Role of KLF6 in Cell Proliferation	16
2.7.2	Attenuation of CDKN1A by miR-4262	17
2.8	KLFs in Tumour Metastasis	17
2.8.1	EMT	17
2.8.2	Extracellular Matrix (ECM)	18
2.8.3	Invasion	18
2.9	KLFs in Signaling Transduction	18
2.9.1	Hormone Receptor Pathway	18
2.9.2	Treatment	18
2.10	Conclusion	19
CHAPTER 3: MATERIALS AND METHODS		20
3.1	In-Silico Analysis	20
3.2	Data Retrieval and Processing	20
3.3	Structure Prediction	20
3.4	Analysis of Missense SNPs	21
3.5	Unique Missense Variants	22
3.6	Missense Variants Frequency Determination	22
3.7	Protein Stability Analysis	23
3.8	Structural and Functional Analysis	23
3.9	In-Situ Mutagenesis	24
3.10	Molecular Dynamics Simulation	25
3.11	Primer Designing	25
3.12	Experimental Analysis	28
3.12.1	IRB Approval and Sample Collection	28
3.12.2	Inclusion Criteria	28
3.12.3	Exclusion Criteria	28
3.12.4	Genotyping	28
3.12.5	Tetra ARMS-PCR	30

3.12.6 Gel Electrophoresis	31
CHAPTER 4: RESULTS	32
4.1 KLF6 Structure Prediction	32
4.2 Ramachandran Plot	32
4.3 Subcellular Localization	34
4.4 Phylogenetic Tree	35
4.5 1 KLF6 Variant Identification	37
4.6 Protein Stability Analysis	49
4.7 Variant's Effect on Protein's Structure and Function	50
4.8 Project HOPE	52
4.9 RNA FOLD	53
4.9.1 Effects of SNPs on mRNA Secondary Structure	53
4.10 MD Analysis	54
4.11 Laboratory based Experimentation Results	58
4.11.1 Genotype Data of Breast Cancer Patients and Healthy Control Samples	58
CHAPTER 5: DISCUSSION	60
5.1 Prospects	62
CHAPTER 6: CONCLUSION	64
REFERENCES	65

LIST OF TABLES

	Page No.
Table 1.1: List of risk factors (Łukasiewicz et al., 2021).....	4
Table 2.1: List of KLF homologs.....	11
Table 3.1: Pathogenicity prediction of KLF6 SNPs.....	21
Table 3.2: Parameters set for primers.....	26
Table 3.3: Melting temperature for selected SNPs.....	27
Table 4.1: Probability score of KLF6 subcellular localization.....	34
Table 4.2: Pathogenicity table of selected missense variants.....	39
Table 4.3: Missense variants after threshold pathogenicity sorting. (D= Damaging/Deleterious, PD= Probably damaging, LB= Likely benign LD= Likely deleterious, T= Tolerated, DC= Disease causing, H= High, M= Medium, N= Neutral, L= low).....	48
Table 4.4: Protein Stability Analysis.....	49
Table 4.5: MutPred2 results.....	50
Table 4.6: Project HOPE Analysis.....	53
Table 4.7: Genotypic distribution data of KLF6 missense variants in both control and patient samples.....	58

LIST OF FIGURES

	Page No.
Figure 2.1: Different transcripts of KLF6.....	9
Figure 2.2: Structure of KLF6 gene.....	10
Figure 2.3: Splice variants of KLF6 gene.....	11
Figure 2.4: All binding domains of KLF6 gene.....	15
Figure 3.1: In-Silico Mutagenesis is induced using PyMol.....	25
Figure 4.1: Shows the predicted structure using Alpha fold with all three C-terminus domains. B: Shows detailed domains of KLF6 structure.....	33
Figure 4.2: Ramachandran Plot showing the measurements of angles in KLF6.....	33
Figure 4.3: Shows the KLF6 protein's localization route as well as the likelihood score. The path of localization is shown in red.....	35
Figure 4.4: Phylogenetic trees of the KLF6 proteins.....	35
Figure 4.5: Sequence alignment through Clustal Omega.....	36
Figure 4.6: A: Structure of KLF6 via Consurf. B: Sequence alignment using Consurf.....	37
Figure 4.7: Total number of variants in database.....	37
Figure 4.8: Types of all KLF6 variants in each database.....	38
Figure 4.9: Frequency of missense variants per exon.....	39
Figure 4.10: Percentage pathogenicity of missense variants per exons.....	48
Figure 4.11: Amino acids are coloured according to the vibrational entropy change upon mutation. BLUE depicts gain in rigid structure and RED a gain in flexibility....	51
Figure 4.12: Intra-atomic interactions of (A) R222Q and (B) K209E.....	51
Figure 4.13: (A) Mutated from Lysine to Glutamic acid at position 209 (B) Mutated from arginine to glutamine at position 222. Red colour shows conserved regions and black shows uniqueness.....	52
Figure 4.14: MFE value of mRNA Secondary Structures.....	54
Figure 4.15: The RMSD graph depicts structural changes in wildtype and mutant protein over 20ns.....	55
Figure 4.16: The Rg graph depicts structural changes in wildtype and mutant protein over 20ns.....	56
Figure 4.17: The RMSF graph for mutant protein and wild type over 20ns.....	56
Figure 4.18: The Hydrogen bonds over time for mutant protein and wild type over 20ns.....	57

Figure 4.19: The SASA graph for mutant protein and wild type over 20ns.....	57
Figure 4.20: Genotypic association with breast cancer.....	59
Figure 4.21: Allelic association with breast cancer.....	59

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ARMS PCR	Amplification- Refractory Mutation System
CADD	Combined Annotation Dependent Depletion
DNA	Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic Acid
EMT	Epithelial to Mesenchymal transition
HOPE	Have Our Protein Explained
KLF	Krüppel Like Factor
MD	Molecular Dynamics
OR	Odd Ratio
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent accessible surface area
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
TAE	Tris Acetate EDTA
TE	Tris EDTA
UV	Ultraviolet

ABSTRACT

Breast cancer is a heterogenous disease-causing mortality and morbidity in females around globe. For this, the study of KLF6 can prove to be beneficial in treating breast cancer. KLF6 belongs to a family of transcriptional regulators. KLF6 is a tumor suppressor gene and helps in upregulation of apoptosis and downregulation of processes like, invasion, metastasis, and angiogenesis. In this study, computational and experimental analysis will be performed. Data is retrieved from ensemble and checked for pathogenicity using SIFT and POLYPHAN. Functional and structural validation is performed through Alpha-Fold. Mutations are then introduced in the structure using *in silico* mutagenesis via pyMOL. Lastly, the structure and functional analysis will be performed using I-MUTANT, Dynamut, Project HOPE, MutPred2 and MUpro. In experimental analysis, the samples of only females above 30 with confirmed breast cancer cases are collected while all other cases are excluded. Blood-based genotyping is performed, and the DNA is statistically analysed. The expected outcomes will be identification of novel missense variant of KLF6 gene, its association, and a potential biomarker in breast cancer. In experimental analysis, the samples of only females with confirmed breast cancer cases are collected while all other cases are excluded. Blood-based genotyping is performed, and the DNA is statistically analysed. The outcomes showed, significant association of genotype GG and AG was found with breast cancer and no family history was associated with breast cancer.

Keywords: Breast Cancer, SNPs, Bioinformatic tools, Cancer, Genotypic Association

CHAPTER 1: INTRODUCTION

Cancer has always been a threat to human race as it is the second leading cause of death around the globe. According to the data published by American Cancer Society's cancer statistics, the new number of cases in United States tolls up to 1,958,310 and deaths 609,820 in the year 2023. Cancer is difficult to treat which is because it is a diverse disease at the tissue level. It has been reported that men are more prone to acquire prostate, lung and bronchus, colon and rectum, and urinary bladder respectively, while females are mostly diagnosed with breast cancer. Moreover, in children the most prevalent type of cancer includes blood cancer, brain, and lymph nodes respectively.

Cancerous cells occur due to mutations in the DNA which affect the structure and function of the protein causing alteration in the cellular functions. Genetic mutations are caused by different risk factors such as use of chemical compounds, smoking, environmental chemical substances, radiation, bacteria, viruses etc. 7% of cancers are caused by bacteria and viruses. As the cellular relations get disturbed, the cell cycle gets altered resulting in uncontrollable proliferation. The normal cell cycle is maintained by proto-oncogenes, these genes are important for normal functioning. If genetic mutation occurs in the proto-oncogene, it causes disruption of cell cycle which eventually leads to cancer. Moreover, any mutation in tumor suppressor gene causes lack of its ability to function properly resulting in cancer. (Hassanpour & Dehghani, 2017)

Cancer can be prevented by avoiding smoking, getting vaccinated against number of different viral infections such as HBV, HCV as they put the patient at the risk of viral induced cancer. Using sunscreen and not using the artificial tanning products helps

lower the risk of skin cancer. Being active, eating healthy food, no consumption of alcohol helps reduce the risk of cancer.

1.1 Breast Anatomy

Breasts are in the chest on the pectoralis major muscles and the weight of breast are held up by the ligaments to the chest wall. Physiologically, breast consist of glands which lie in the front and functions to produce milk. 15 to 20 lobes circle around and enveloped in fat forms breast. The lobes tell us the shape and size of the breast. The milk producing glands are situated in lobules which together forms lobes. Hormones causes the glands in lobules to produce milk. (Simon & Robb, 2022a)

1.2 Breast Cancer

Breast cancer develops in the breast tissue. Breast cancer is tricky as it occurs almost unknowingly. Majority of females find out their disease during the routine screening of breasts while others may find out by noticing difference in their breast size or shape, discharge from nipples or lumps in breast.

However, pain in breast also known as mastalgia is most common symptom and cannot be neglected and for this physically examining, mammography and tissue biopsy must be done to ensure the risk of breast cancer. The chances to cure breast cancer is to diagnose the disease in its initial stage because as the time passes the tumour starts to metastasize to other parts either lymphatically or haematologically leading to poor prognosis. Hence, routine screening must be done by every female to ensure a healthy life. (Cain et al., 2019) (Narod, 2018)

1.3 Epidemiology

According to Karachi cancer registry, the most frequent cancer among females at the age of 45 is breast cancer. In Pakistan, this is becoming an alarming situation as the risk of breast cancer is continuously increasing, this has been further confirmed by hospital-based studies as females younger than 40 are being diagnosed with breast cancer. Overall, in Asian population, Karachi has maximum number of cases of breast cancer reported. (“Cancer Prevention and Control in Pakistan: Review of Cancer Epidemiology and Challenges,” 2020)

1.4 Classification of Breast Cancer

There are four types of breast cancer which are as follows.

1.4.1 Histological Classification

- Invasive breast cancer (IBC)
- Invasive breast cancer of no special type (NST)
- Invasive lobular carcinoma,
- Tubular, breast cancer
- Mucinous A breast cancer
- Mucinous B breast cancer
- Neuroendocrine breast cancer

1.4.2 Molecular Classification

- Luminal A breast cancer
- Luminal B breast cancer
- HER-2 breast cancer
- Basal like breast cancer (Koboldt et al., 2012; Weigelt et al., 2008)

1.5 Risk Factors and Prevention

Risk factors are listed below in the table.

Table 5.1: List of risk factors (Łukasiewicz et al., 2021).

Non-Modifiable Factors	Modifiable Factors
Female sex	Physical activity
Older age	Obesity
Family history	Smoking
Genetic mutations	Diethylstilbestrol
Ethnicity	Alcohol consumption
Density of breast tissue	Lower intake of supplements (vitamins)
Breast diseases	Increased risk to artificial light
Pregnancy	Consumption of processed food
Breast feeding	Hormonal replacement therapy
Radiation therapy	Risk to chemicals
Menstruation and menopause	Exposure to other drugs

1.6 Stages of Cancer

The staging of breast cancer is firstly determined by doing imaging scrutiny and physical examination before initiating any suggested treatment on the other hand the

stages are known after operating and determining the pathology of the primary tumour and regional lymph nodes. Tumour, Node involvement and metastasis (TNM) classification system is used for determining the staging of breast cancer. With the help of this system scientists have designated stages in accordance with the size of the tumour (T), status of regional lymph node (N), metastasis (M).

1.7 Prognosis

Initial stages of breast cancer have good prognosis. Stage 0 and I has 100% chance of survival for 5 – years while for stage II, III and IV it is 93%, 93%, 72% respectively. When the cancer starts to spread and metastasize affecting other organs, the prognosis gets worse noticeably to only 22% in the next five years. (Simon & Robb, 2022b)

1.8 Treatment

The possible treatments for breast cancer are listed as follows,

- Radiation
- Chemotherapy
- Hormone therapy
- Targeted therapy
- Oral chemotherapy
- Peritumoral lidocaine injection
- Elacestrant
- Antibody drug conjugates
- Surgery (Burguin et al., 2021)

1.9 Cancer

The second most leading cause of death worldwide is due to a generic disease known as cancer. Cancer is characterized as a large group of disease in which the cells of the body grow rapidly, and the immune system can no longer recognize these cells, as to fight against them. Cancer can occur in any part of the body, affecting it, and are also able to spread to the neighbouring organs, this phenomenon is known as metastasis. (Upadhyay, 2021) (Yin et al., 2021). There are six hallmarks of cancer which includes unlimited cell growth, eluding apoptosis, increased angiogenesis, ability to metastasize and evasion from growth inhibiting signals. (Hanahan & Weinberg, 2000)

In the 1980s, it was proved that besides other factors causing cancer such as hormones, sunlight, fungi, alcohol, bacteria, parasites, herbs, pharmaceuticals, salted fish, tobacco and wood dust, viruses also cause cancer. There are eight listed viruses according to the International Agency for Research on Cancer (IARC), which are as follows; Human deficiency virus (HIV), Hepatitis B and C virus, Human herpes virus 8, human papilloma virus, human T-cell lymphotropic virus and Epstein-Barr virus. Apart from the mentioned factors, the American Institute for Cancer and the World Cancer Research Fund have identified more agents which includes, low fibre diets, sedentary life, beta carotene, not breast feeding, processed and red meat, obesity and increased adult height. (Blackadar, 2016)

There have been various treatment options used against cancer such as the use of radiations, surgery, immunotherapy and chemotherapy. Combination of these conventional therapies helps us to halt the altered metabolism and signalling which eventually leads to inhibition of uncontrolled growth of cells and survival of altered

cancerous cells. Cancer can also be coined as an injury that never gets cured. Although we have gathered immense data on different types of cancer making it possible to treat it at an early stage still the ultimate cure of cancer is unknown. (Yin et al., 2021)

CHAPTER 2: LITERATURE REVIEW

2.1 History of KLF

The KLFs were first discovered in fruit flies, *Drosophila melanogaster*. These belong to the DNA – binding proteins, family of zinc finger DNA binding proteins. Mutations were observed in Krüppel like factors by Nüsslein-Volhard and Wieschaus which affected the initial stages of embryogenesis by altering the pattern of anterior and posterior segmentation. (McConnell & Yang, 2010). They were working of understanding what the various genetic determinants of early stages of embryonic development are. These findings led them to win a Nobel Prize in 1995 in the field of Physiology, together with another remarkable scientist Edward B. Lewis.

In humans, the first member of Krüppel like factor identified was EKLF also known as KLF1, it has a crucial role in development and maturation of erythroid. (Nüsslein-volhard & Wieschaus, 1980). An experiment was designed to check the importance of this gene by eliminating KLF1 gene in mice, it was observed that the mice developed β -globin deficiency and became anaemic which puts the embryo at risk to die. (Parkins et al., 1995)

After the identification of KLF1 other members of this family were discovered, totalling to 1-18 members, however, KLF18 is the newest addition to the KLF family it is termed as a pseudogene as it is the result of a gene duplication or retro transposition events. (Pei & Grishin, 2013) The human KLF genes are mammalian homologs of Krüppel like factors found in *D. melanogaster*, which are evolutionarily conserved across species.

2.2 KLF6 Structure

KLFs are zinc finger containing transcription factors, functions in regulating various processes like apoptosis, differentiation, proliferation, and development. Changes in the gene of KLF causes altered function and patho-biology of different diseases in humans such as, cancer, metabolic diseases, and cardio-vascular diseases. KLF6 gene is positioned on the short arm of chromosome 10 (10p15). The gene consists of four exons. There can be different transcripts of KLF6, the known transcripts are 7 but only 3 out of these 7 are known to get translated into proteins: KLF-204, KLF-206 and KLF-207, as seen in the figure 1.

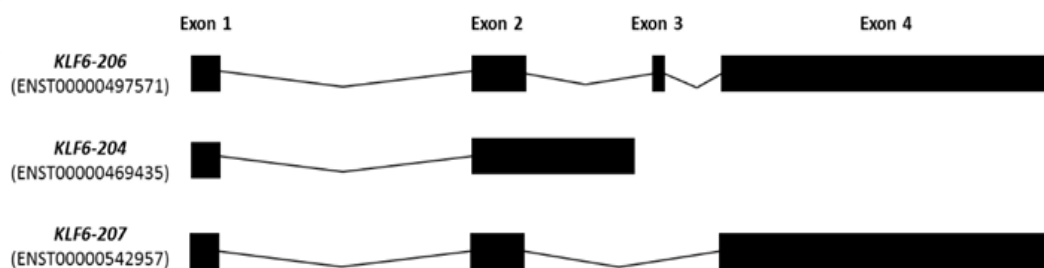


Figure 2.1: Different transcripts of KLF6.

KLF6-206 also known as KLF6 is the primary transcript and is made up of 283 amino acids in a length KLF6 was discovered in placenta where it helps in regulating expression of pregnancy specific glycol protein genes. KLF6 was initially thought to be a core-promotor binding protein. KLF6 is not only just present in placenta but it can be found in liver and leukocytes. (Suzuki et al., 1998) KLF6 has been assigned different names such as ZF9, B-cell derived protein 1 (BCD1), collagen type1 alpha1 (COL1A1), endoglin ENG), core-promotor binding protein (COPEB), suppressor of tumorigenicity12 ST12), transforming growth factor beta 1 (TGFβ1), plasminogen activator urokinase (PLAU).

It has been known that in the gene of KLF6, occurs a germline mutation in the non-coding region with causes guanine to be replaced by adenine (IVS1-27G>A), as this mutation happens to be germline it means it can be transmitted to the offspring. This single nucleotide polymorphism causes splicing of the transcript without a primer, resulting in splice variants of KLF6: KLF6-SV1, KLF6-SV2 and KLF6-SV3. There are different sizes of the splice variants produced, this depends upon the acceptor sites (Figure 2.2).

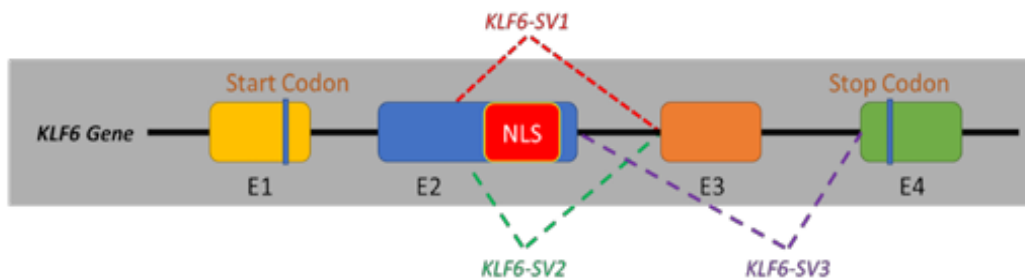


Figure 2.2: Structure of KLF6 gene.

KLF6 gene undergoes a phenomenon known as alternate splicing, this process gives rise to three different splice variants as already mentioned. This alternate splicing causes problems and defect in an individual. The full length KLF gene has all three exons present but when an alternate splicing occurs the Nuclear Localization Sequence (NLS) present at the end of exon 2 gets deleted which results in KLF-SV1 and KLF-SV2. The NLS is the region which guides the KLF gene to get inside the nucleus and function as a transcriptional regulator but as the NLS region gets omitted the KLF gene can no longer enter nucleus rather it stays in the cytoplasm. As the gene is trapped in the cytoplasm, it can no longer function as a transcriptional regulator which eventually results in cancer. Another splice variant KLF-SV3 also contains an SNP which occurs in exon 3 hence, deleting it. The role of Splice variant 3 is still under consideration. The most studied splice variant among all three is KLF6-SV1.

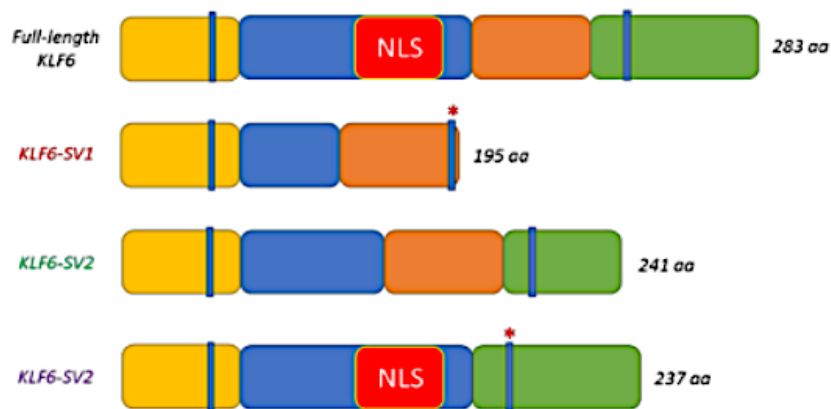


Figure 2.3: Splice variants of KLF6 gene.

2.3 KLF6 Protein Structures

There are portions in the gene of KLF6 which are highly conserved which includes three repeats of Cys2-His2 (C2H2), these are present in the C-terminal of the gene. In the promotor region of the gene are present some motifs such as a GC-box or CACCC, these motifs act as a binding site for the DNA binding zinc-finger repeats. Contrarily, the N-terminal of the KLF6 gene influences with different sets of proteins such as co-repressors, activators, chromatin-modifying enzymes, and transcriptional factors. This concludes that the functional multiformity is controlled by the N-terminal of the KLF6 gene.

2.3.1 Conservation

In mammals, KLF proteins are highly conserved proteins to date ranging from mouse to humans. Scientists have found KLF homologs in different species mentioned in the table.

Table 6.1: List of KLF homologs.

KLF homologs

Gallus gallus (chicken)
Danio rerio (zebrafish)
Xenopus laevis (frog)
Caenorhabditis, elegans genomes – KLF1,2,3

KLF proteins interact with GC box present in promotor and enhancer region, this interaction is possible due to a homology in the carboxy-terminal zinc finger domain. Due to this structural resemblance different KLF proteins bind with their transcriptional targets. For example, KLF 2,4 and 5 in ES cells attaches and turns on Esrrb, Tc11, Fbox15 and Nanog. However, as stated before the N-terminal gives the protein their uniqueness.

2.3.2 *Expression*

KLF proteins have different expression level, some of these proteins are expressed in every tissue example includes KLF 6, 10 and 11 on the other hand KLF1 is only expressed in erythroid cells, KLF2 in lung, KLF4, KLF5 are seen in gastrointestinal tract.

2.3.3 *Phylogenetic Analysis*

KLFs are phylogenetically divided into three separate groups depending on their functional characteristics.

- i. **Group 1:** Transcriptional repressors which binds with the carboxy-terminal binding protein (CtBP). This group includes KLF3, KLF8 and KLF12.
- ii. **Group 2:** Includes transcriptional activators. These includes KLF1, KLF2, KLF4, KLF5, KLF6 and KLF7.

- iii. **Group 3:** Includes transcriptional repressors which engages with Sin3A; a common transcriptional co-repressor. This group includes KLF9, KLF10, KLF11, KLF13, KLF14 and KLF16.
- iv. **KLF15 and KLF17:** these two are less related depending upon no known protein interaction. (Syafuruddin et al., 2020)

2.4 Structure of KLFs

The structure of KLFs is divided into two main domains which are

- Zinc-finger domain
- Functional domain

2.4.1 *Zinc Finger Domain*

Most of the transcription factors possesses a zinc finger domain, the type of present in KLFs is C2H2. In C2H2 type there are two amino acids cysteine and histidine, functions just like a finger and holds the zinc atom in place, the amino acids then condense to form a structure. (Brayer & Segal, 2008) It is safe to state that at the carboxy terminal of KLFs there are present three highly conserved zinc fingers. Location of zinc finger motifs present in KLF are visually represented in figure. There are three zinc finger motifs present with the first two containing 25 amino acids and the last one has 23 amino acids. The DNA sequence is identified by the zinc finger motifs as one zinc finger reads 3 base pairs totalling to 9 base pairs, 3 by each motif. (Nakagawa et al., 2008)

The sites where DNA binds to the transcriptional factor (KLF) has been studied by scientists for a long time and it has been observed that these sites resembled among different KLF proteins. These sites were found rich in GC-sequences with inclination

towards a specific sequence; 5'-CACCC-3', this sequence is identified in a gene called as goblin gene which is easily recognized by KLF1 and others. All the data collected on how KLFs bind to the DNA is being done studying how promotor regions bind and oligonucleotide screens. (Miller & Bieker, 1993) As we know, KLFs must reach to the nucleus to regulate different genes and for this they need nuclear localisation signals. These signals are found in the zinc finger domain, examples include, KLF1, KLF4, KLF8 and KLF11.

2.4.2 Functional Binding Domain

The second domain of KLFs is the functional binding domain. The amino terminal of the KLF is basically the functional terminal as it permits the modifiers, co activators and repressors to bind allowing specificity and diversity. As the binding is specific it helps us to sort out KLF proteins depending upon their function.

2.5 CtBP-Binding Site

When KLF3 was studied it was originally considered as an activator but rather it was later confirmed to be an active repressor. KLF3 has a sequence of 74 amino acids which causes its activity and is present in amino terminal. In yeast, it was seen that the KLF3 engaged itself with the CtBP co-repressor. Consensus sequence PXDLS present in KLF3, KLF8 and KLF12 helps in attaching with the CtBP during this binding process the KLFs activity is turned on and they act as a repressor and inhibits the expression of AP-2 gene.

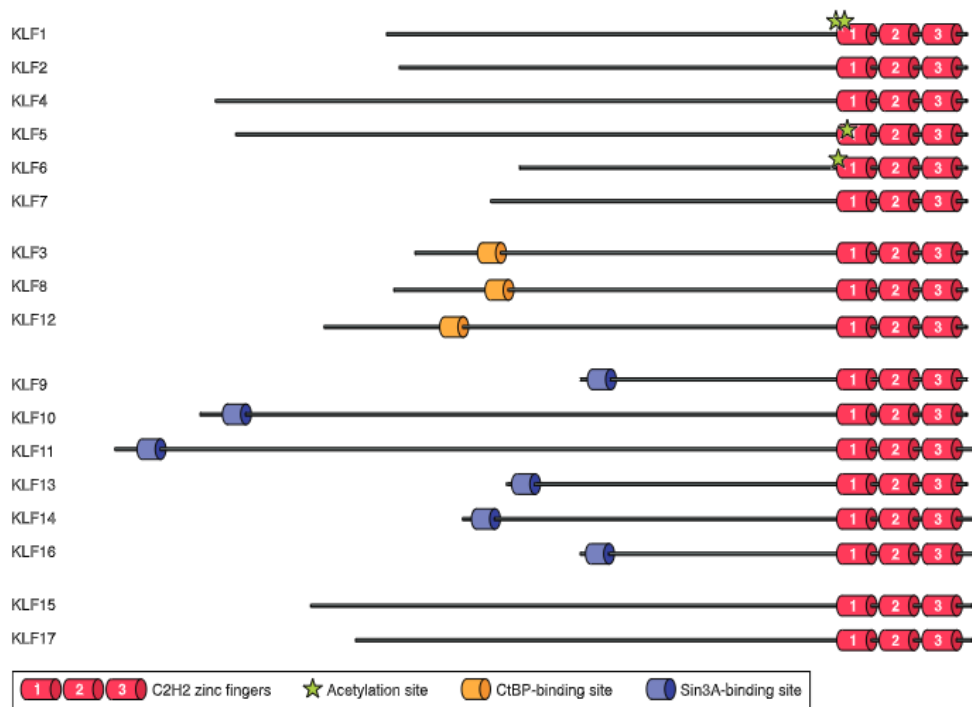


Figure 2.4: All binding domains of KLF6 gene.

2.6 Sin3A-Binding Site

KLF 10 and KLF11 are also transcriptional repressor and due to which the scientists were able to separate three repression sites designated as R1, R2 and R3. These sites are in the amino terminal regions of the KLF proteins. The sin3 binding domain (SID) is present in the R1 domain, this structure has hydrophobic nature meaning it dislikes water, it creates a helical structure ensuring stable interactions with Sin3 proteins, these proteins are histone deacetylase-dependent corepressors. KLF 9,10, 11,13 and 16 attaches to the Sin3A via AA/VXXL, a conserved helical motif. This attachment helps to activate the repressors.

The KLF14 possesses putative Sin3A binding site, hence, no binding between them is yet known. On the contrary, KLF1 does not contain any sin3A binding domain but it recruits Sin3A to function as a repressor, also this interaction is done via carboxy

terminal instead of amino terminal of the zinc finger. (McConnell & Yang, 2010) The function of KLFs in breast cancer are as follows.

2.7 KLFs in Cell Proliferation

Cell proliferation is the key process in the living things as the cell divides it creates more cells, this process is the basis for development, inheritance of organisms, growth, and reproduction. As it has been established that cancer grows rapidly by dividing as due to genetic alterations causes defects in tumour suppressors and regulatory pathways hence, cell proliferation is one of the hallmarks of cancer. However, it is still under consideration how KLFs affect the breast cancer, the full mechanisms are yet to be understood.

KLF6 is vital for cellular differentiation, this was confirmed by experimenting on zebra fish, mice and flies. The results showed that the mice died due to homozygous mutant strain of KLF6. As the KLF6 was unfunctional the yolk sac and haematopoiesis vascularisation had defects causing death of the mice. Moreover, the homozygous deletion of the gene caused defects in cell proliferation and differentiation, KLF6 is also involved in developing the liver. KLF6 gene is also associated with regulation of trophoblast differentiation, differentiating preadipocytes to adipocytes by downregulating the delta like 1 by making it interact with HDAC3.

2.7.1 Role of KLF6 in Cell Proliferation

KLF6 is a tumour suppressor which means it helps in preventing the production or progression of tumour/cancer. KLF6 work by activating a gene known as CDKN1A; this gene codes for a protein called p21; cyclin dependent kinase inhibitor. This protein stops the cell from dividing uncontrollably. KLF6 works as a transcriptional activator

by activating the CDKN1A gene. As this is turned on the uncontrollable proliferation of cells comes to a halt hence, minimizing the risk of developing a tumour.

2.7.2 *Attenuation of CDKN1A by miR-4262*

miR-4262 is a microRNA; small molecules which helps in regulation of gene expression by binding to the mRNA. miR-4262 binds to mRNA and inhibits the activation of CDKN1A resulting in low production of p21 and high rate of cell division leading from a tumour to cancer. To continue with therapeutics, we need to understand all the underlying mechanisms.

2.8 KLFs in Tumour Metastasis

Metastasis is the process in which the cancerous cells start to spread to other parts of the body by adopting either blood circulation or lymphatic pathway. It has been reported multiple times that KLFs are involved in metastasis.

2.8.1 *EMT*

Epithelial to mesenchymal transition occurs when the epithelial cells start to act like mesenchymal cells, they do this by losing cell polarity and cell to cell adhesion; due to loss of E-cadherins, alongside they acquire invasive and migratory properties. E-cadherin proteins can be downregulated by certain known factors such as, Slug, snail, RhoB, TGF- β , TWIST1, IGF-1 and FGF. It has been estimated that KLFs are involved in EMT during breast cancer. KLF6 has a splice variant known as KLF6-SV, this variant is known to enhance the Epithelial to mesenchymal transition by increasing the transcription level of TWIST1. It has been evident that the breast cancer metastasises was increased in mouse models hence, lower chances of survival.

2.8.2 *Extracellular Matrix (ECM)*

The extracellular matrix needs to be broken down during the process of metastasis and this degradation is done by Matrix metalloproteinases. MMPs are regulated by KLFs, for instance, KLF6 causes upregulation of Tissue factor pathway inhibitor-2 (TFPI-2); it is a matrix associated Kunitz inhibitor or causes downregulation of MMPs specifically of zymogen matrix metalloproteinases resulting in sustained EM and no metastasis.

2.8.3 *Invasion*

It has been observed that KLFs play an important role in invasion but any data on KLF6 has not yet be reported. (Zhang et al., 2020)

2.9 KLFs in Signaling Transduction

2.9.1 *Hormone Receptor Pathway*

The family of KLFs play a significant role in the hormone receptor pathway which is crucial for the formation of breast cancer. As much as role of KLFs is crucial it is a very complicated process to understand roles in oestrogen receptors. KLF6 has been involved in signal transduction pathways by downregulating oestrogen receptor – mediated cell growth in ER (+) breast cancer. This was achieved by the reaction of ER α and c-Src, which results in inactivation of c-Src to MAPK signalling pathway.

2.9.2 *Treatment*

KLFs can be a potent target for treating cancer. KLF6 can be enhanced by using Zoledronic acid, as the KLF6 increases it causes apoptosis in MCF7 cells. Along with KLF6, KLF2 is also upregulated.

2.10 Conclusion

KLFs are transcriptional regulators as they perform important role in regulation of cell proliferation, apoptosis, differentiation, and metastasis in breast cancer. We can detect the progress of a breast cancer by checking the levels of KLFs as during the development of breast cancer the levels gets disturbed, which makes them breast cancer markers, these markers then help identify any breast cancer malignancies.

The most important aspect of studying KLFs is to use them in therapeutics, and to consider the possibilities that they might have additional roles than just being a transcriptional regulator. It has been known that some of the KLFs also play their part in the development of breast cancer, but it is still unknown to us whether it has positive or negative role. Until now it is not known how the living beings maintain a balance in tumour repressing and activating KLFs, it might be specific to cells and tissues.

When our body experiences stress it can cause alteration in genes, when such happens with KLFs, other KLF factors get activated to compensate the damage but when the damage is beyond repair other KLFs get activated causing promotion of cancer. Moreover, the different level of expression of KLF factors help determine ups and downs in the relationship of tumour suppressors and oncogenes. Other factors including, genomic instability, mutations causing cancer, tumour microenvironment and cancer specific context also disturbs this balance.

When designing drugs, we need to consider the possibility of off target effects, homology of KLFs, specificity of the drug used, degree of functional redundancy, resistance. In conclusion, we need to have a full grasp in understanding the Krüppel like factors so that we can get started on the therapeutic approach. (Zhang et al., 2020)

CHAPTER 3: MATERIALS AND METHODS

3.1 In-Silico Analysis

Using the in-silico tools the wild type and variant structures of *klf6* were identified. The pathogenicity of missense SNPs of *klf6* were investigated to correlate with breast cancer.

3.2 Data Retrieval and Processing

KLF6 gene sequence was firstly retrieved from ENSEMBL database. FASTA sequence of 283 amino acids was downloaded with the transcript ID ENST00000497571.6. The protein sequence of this gene was also downloaded. Variant table was then downloaded from ENSEMBL. The ENSEMBLE database gives information about predicting protein function, variants of gene and linkage to the disease. (Cunningham et al., 2019). The sequence of amino acid and gene were saved in FASTA format. The data was retrieved from three different databases; like Ensembl (reference assembly GRCh38.p13) (Howe et al., 2021), Catalogue of Somatic Mutations In Cancer (COSMIC) (Forbes et al., 2006) and Genome Aggregation Database (gnomAD) (Koch, 2020). All the data was then sorted into its respected variant such as missense, frameshift, in splice, untranslated region variants etc but, only the missense variants were selected and processed.

3.3 Structure Prediction

The structure of *klf6* gene was predicted by using Alpha Fold, a bioinformatics tool. The FASTA sequence of amino acids was submitted to Alpha fold which resulted a 3-D structure of the protein; this structure was then saved as a .pdb file and was used for further analysis. (Jumper et al., 2021). The predicted structure was then visualized

using PyMol (DeLano, 2002). The FASTA sequence of KLF6 gene was then entered into InterPro. This tool helps to predicting the domains, substrate binding site, motifs, active site of the protein (Blum et al., 2021).

3.4 Analysis of Missense SNPs

Then Detection of the deleterious missense variants was narrowed down, this was done by copying the missense variants from ENSEMBL and all unnecessary columns were removed. Filters are then added to SIFT (Sim et al., 2012), CADD (Kircher et al., 2014), PolyPhen and class REVEL (Ioannidis et al., 2016). These filter the two variants of the KLF6 gene, rs941470359 and rs771063540 which have been found to exhibit a high degree of pathogenicity were subsequently chosen for further investigation through in-silico analysis. The data from ENSEMBLE was then sorted in a separate excel sheet, four columns were generated which included variant ID, Allele, A.A, A.A co-ordinate.

Table 3.1: Pathogenicity prediction of KLF6 SNPs.

Pathogenicity Prediction Tools	Score Range	Interpretation/Classification	Reference
SIFT	0-1	0: Deleterious 1 (or close): Tolerated	(Kumar, Henikoff, & Ng, 2009)
PolyPhen	0-1	0–0.25: Benign 0.25-0.8: Possibly Damaging 0.8-1: Probably Damaging	(Adzhubei, Jordan, & Sunyaev, 2013)
Mutation Assessor	0-1	<0.1: Neutral 0.2-0.4: Low	(Hassan, Shaalan, Dessouky, Abdelnaiem, & ElHefnawi,

		0.5-0.9: Medium >0.9: High	2019)
MetaLR	0-1	<0.5: Tolerated >0.5: Damaging	(Alirezaie, Kernohan, Hartley, Majewski, & Hocking, 2018)
REVEL	0-1	<0.5: Benign >0.5: Diseases Causing	(Ioannidis et al., 2016)
CADD	0-35	≥ 30 : Deleterious	(Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019)

3.5 Unique Missense Variants

All the gene variants of *klf6* were reported from the three databases and graphically represented. The data from different bases was then compared and any redundancy was deleted, and the common variants were treated as one. Moreover, all other variants which were present on a different location were given a number and considered unique variants.

3.6 Missense Variants Frequency Determination

The frequency of missense variants was determined, and a scatter chart represented the frequency of missense variant per amino acid. The location and total number of amino acids within an exon were determined and was graphically represented. The average number of the missense variants was then calculated, and the data was presented in a graph along with the frequency of the variants per exons.

3.7 Protein Stability Analysis

Bioinformatic tools help in determining the stability of protein molecules. By using Mupro, DynaMut and I-Mutant 2.0, the effect of pathogenic missense protein on the structure of klf6 gene can be analyzed. Using I-Mutant, the stability analysis of KLF6 protein was confirmed. The two SNPs which were selected, had their stability role checked on the structure and function of the KLF6 protein. It was considered that a single change in SNP can cause $\Delta\Delta G$ values to change. If the $\Delta\Delta G$ value is lower, it means decrease in the stability of protein. If the value is in between <0.5 and >0.5 , it means that the structure's stability has increased with more flexibility.

MUpro software helps in predicting the impact of SNP on the wild type of structure. This software works by showing the $\Delta\Delta G$ score ranging from -1 to 1, if it is less than 0 shows destabilizing structure. On the contrary, $\Delta\Delta G$ score greater than 0 implies increase in the structural stability. DynaMut helps in determining the effect of SNP on protein structure and evaluating the molecular dynamics and stability of protein. This is achieved through the difference in vibrational entropy (DDS) between wild type and mutant.

3.8 Structural and Functional Analysis

HOPE, MutPred2, and DynaMut were the softwares used to evaluate the effect of amino acid variants on the structure and function of proteins. HOPE helped to understand the physicochemical alterations of the structure based on the size, charge of amino acids, hydrogen-bonding based inter or intra-residual contacts and hydrophobicity. (Dunlavy et al., 2005) The tool named DynaMut was used to identify the fluctuation and distortion in protein because of the missense variants. It was also used to check the effect of mutation on the molecular motion of the protein. In this

regard, this tool uses the Normal Mode Analysis approach which analyzes the vibrational entropy change in the protein because of the single nucleotide polymorphisms (Rodrigues et al., 2018).

Finally, MutPred2, is a software that helps in predicting effect of pathogenic SNPs on proteins and their impact on the molecular activities of proteins. The likelihood that a protein has become pathogenic due to variations in its structure can be calculated using this tool. A score more than 0.5 is considered pathogenic but it may be wrong as there is always a chance of false positive results. If the score obtained is 0.68 and 0.80 the chances of false positive results decrease with 10% and 5% respectively. (Pejaver et al., 2017)

3.9 In-Situ Mutagenesis

Mutagenesis was induced in the normal structure of KLF6 gene. The structure of KLF6 gene was predicted by using Alphafold. The FASTA sequence was submitted to Alphafold after which, the software provided a predicted result. The 3-D structure was then downloaded for further analysis. The 3-D structure was then visualized in PyMol. The original structure was downloaded from Alphafold. In the wild type of structure, mutations were introduced by changing the amino acid.

Through PyMol, the structure of protein was aligned and predicted. The Wizard tool in PyMol was chosen and then “mutagenesis of protein” was selected from the dropdown panel, the wild type amino acid was then changed to the mutated amino acid and then the changes were applied. The structure was changed and was saved in .pdb file. MD simulations were performed on both mutant and wild type klf6 structures.

3.10 Molecular Dynamics Simulation

The molecular dynamic simulations for wild and mutant strain were carried out using GROMACS software (Abraham et al., 2015) on the supercomputer. PuTTY was installed, it helped in accessing the supercomputer with all the data, and WinSCP and the SFTP file transfer protocol acted as a bridge and transferred all the data from PC to the supercomputer. These simulations helped in determining how a missense mutation occurring in the wild type of structure of klf6 causes changes in the structure. So, for both wild and mutant proteins, a 20-nanosecond simulation was done, and different metrics such as root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration, the number of hydrogen bonds, and Solvent accessible surface area (SASA) were observed.

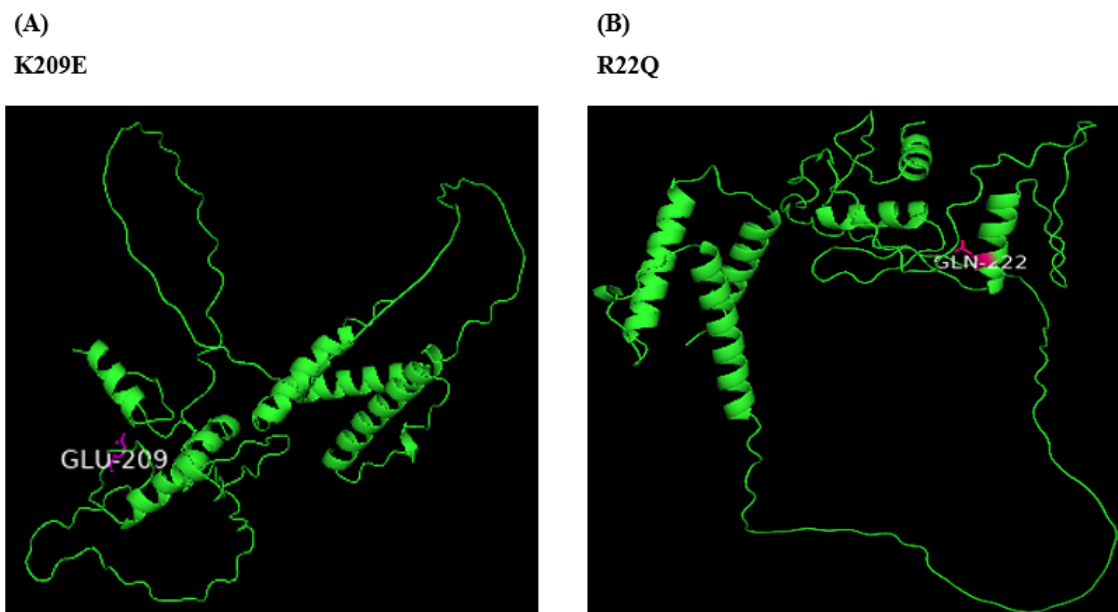


Figure 3.1: In-Silico Mutagenesis is induced using PyMol.

3.11 Primer Designing

Primer 1 was the tool used to computationally generate primers for ARMS PCR (Collins & Ke, 2012). For ARMS PCR four primers, two outer and two inner primers

were designed. Furthermore, through UCSC in silico PCR the primers were validated. The settings of primer 1 tool were set to default except the position of SNP and difference of allele. Next step was to design primer.

This was done by firstly going to Ensemble and downloading RFT file with lines column changed to according to co-ordinates. The location of the SNP was then checked accurately in the RFT file. The desired SNP is highlighted. Primer 1 tool was used to design primers. The sequence was pasted in Primer 1 and primers were then picked. The primers were then validated in Insilco.

Table 7.2: Parameters set for primers.

PARAMETERS SET	
Optimum primer size	22
Maximum primer size	24
Minimum primer size	18
Optimum (inner) product size	200
Maximum (inner) product size	300
Minimum (inner) product size	100
Optimum primer Tm	65
Maximum primer Tm	65
Minimum primer Tm	50
RS ID	rs764830375

Position of SNP	210
Wild type Allele	C
Mutant Allele	T
Strand Position	Reverse strand
Total characters	360

Table 3.3: Melting temperature for selected SNPs.

Primers	Melting Temperature
Forward inner primer (G allele): 190 CCACTTGAAAGCACACCATCG 210 63	63
Reverse inner primer (A allele): 228 CACTGACCTGTGTGCGGCT 210 63	63
Forward outer primer (5' - 3'): 13 GGAACCTTCTCAACTGTGGGGT 34 63	63
Reverse outer primer (5' - 3'): 340 AGTGAGGATTTGTCTGCCCTGA 319 63	63
Product size for G allele: 152 Product size for A allele: 216 Product size of two outer primers: 328	

3.12 Experimental Analysis

3.12.1 IRB Approval and Sample Collection

The first step to sample collection includes approval of Institutional Review Board of ASAB afterwards 100 samples were collected from breast cancer patients and healthy controls each. The protocol was followed by the guidelines set by the ethical review board. A tourniquet was tied on the arm, 3 cm above the visible vein. The area was sterilized using an alcohol swab and 3-5ml of blood was drawn and emptied in EDTA tubes. EDTA tubes (ethylene diamine tetra acetic acid) helps in inhibiting the formation of blood clots. All volunteer individuals involved in this study signed a consent form before giving their blood as samples The history forms were completed for everyone with necessary data, including the patients age, cancer type, tumor grade, receptor subtypes, treatment status, data on pre/post menopausal age, breast feeding, smoking and family history.

3.12.2 Inclusion Criteria

The current study only included female patients above 30 and breast cancer confirmed in them.

3.12.3 Exclusion Criteria

Patients with age less than 30 and co-morbidities such as pregnant, lactating females, females experiencing PCOs were excluded from this study.

3.12.4 Genotyping

The DNA was extracted from the collected blood samples, before extracting the blood samples were stored in refrigerator at room temperature so that the blood

contents get settled. Organic methods for extracting DNA were used. Several different solutions were used while performing Phenol-Chloroform method. Solution A was used first, which was prepared by mixing 10mM Tris (pH 7.5), 0.32 M sucrose and 5mM of MgCl₂ and 1% v/v Triton X-100; Triton was added after autoclaving the rest of the solution. After solution A, solution B was used; 10mM of Tris (pH 7.5), 2mM EDTA (pH 8.0), 400mM NaCl. Solution C is used next which is pure phenol. Lastly, Solution D is used which is made up of chloroform with 24 volumes and isoamyl alcohol of 1 volume. 20% SDS was used which was prepared by mixing 20 grams of SDS in 100mL of water.

750 µl of blood was taken in Eppendorf and same volume of Solution A was added, inverted 5-6 time, placed at room temperature for 10mins and was then centrifuged for 10 mins at 13000 Rpm. Supernatant was then discarded. 750 µl of Solution A was again added and centrifuged, supernatant was discarded. 400 µl Solution B was added and centrifuged at 13000 Rpm for 10 mins, supernatant was discarded. 400 µl of Solution B was added again with 20% SDS and 5 µl Proteinase K and incubated overnight at 37 degrees Celsius.

250 µl of solution C and 250 µl of solution D was mixed in a separate Eppendorf tube and then was added in the solution placed overnight. Centrifuged the tube for 10mins at 13000 Rpm. Two layers were formed and aqueous (upper layer) was transferred into a separate tube. 55 µl of 3M sodium acetate and 500 µl of chilled isopropanol was added, the tubes were inverted until DNA was precipitated. The tube was then again centrifuged at 13000 Rpm for 10mins, supernatant was discarded, and pellet was resuspended in 200 µl of chilled ethanol. The tube was centrifuged again for

8 minutes at 13000 Rpm, let it air dry. After drying was done, DNA was submerged in 200 µl of PCR water.

3.12.5 *Tetra ARMS-PCR*

The newest version of ARMS-PCR is “Tetra ARMS-PCR” and it is considered to be the “Gold Standard”. This technique is used to study point mutations and single nucleotide polymorphism (SNPs). In tetra ARMS-PCR, two set of primers are used, two external and two internal primers. 12 µL of reaction mixture was used for each PCR cycle.

Primers were optimized by running multiple gradient PCRs with different temperatures and concentration. In tetra ARMS PCR, the mixture contained 1 µL of DNA, 1 µL of each four primers (4 µL), 1 µL PCR water (nuclease free) and 6 µL of Solis master mix. This mixture is the complete recipe to run a PCR, it contains dNTPs, PCR buffer, Taq polymerase, magnesium and loading dye. This mixture is prepared in PCR tubes and then centrifuged so that the components mix equally. The PCR included 30 cycles; the first phase is the denaturation in which the template is heated at 95 C for 5 minutes causing the strands to break.

Then comes the annealing phase which lasts for 30 seconds, the primers attach with the single stranded DNA. For the variant rs764830375 the temperature at which the primers were optimized was 62 C. the last phase is the extension, the taq polymerase helps in extending the single stranded DNA which lasts for 7 minutes at 72C. The PCR process is then completed, and the PCR mixture is stored at 4C for further analysis.

3.12.6 *Gel Electrophoresis*

Gel electrophoresis was the next step, this was done to analyze the quality of DNA extracted from the blood samples. Firstly, agarose gel of 50mL was prepared, this was done by mixing 0.5g of agarose in 5mL of 10X TAE buffer and volume was raised to 100mL. Solution was then microwaved for 1-2 minutes to get a clear solution. After the solution cooled down a little, 5 μ l of ethidium bromide was added. The gel mixture prepared was then poured into the gel tank, avoiding formation of bubbles and then comb was placed carefully, and the gel was left to solidify for 30 minutes. For 100ml of gel, 1g of agarose was added to 10ml 10X TAE buffer and raised to 100ml and 5 μ l of ethidium bromide was added. The next step was to load the samples; 3 μ l of loading dye and 5 μ l of sample was mixed and loaded. The gel is then placed in the tank with 1X TAE buffer. The gel was then visualized on a UV-transmitter and the samples were then stored at 4 degrees Celsius for further use.

CHAPTER 4: RESULTS

4.1 KLF6 Structure Prediction

Alpha fold was used to predict the three-dimensional structure of KLF6 protein with the help of residue model confidence score from 0-100. The C-terminal domains show high confidence score (80-91 pDDLT score) showing that the C-terminal domain is highly conserved.

Hence, only the 3 C-terminal alpha helical domains were predicted by Alpha Fold as shown in the figure 4.1 A, while the remaining residues were left unpredicted. Interpro provided labelled domains of KLF6 which are the 3 C-terminal covers 200-229, 230-259 and 260-283, the N-terminal covers 2-201, 2 internal disorder domains cover 106-198, 136-197 amino acids.

While Interpro analysis provided labelled domains of KLF12 protein consist of variable N-terminal domain covers 1 to 224 amino acids, 3 internal disorder domains cover the 79-125, 217-254, 268-314 amino acids respectively. While C-terminal domain consists of 3 conserved zinc finger motifs covering the amino acids number 317 to 399 shown in figure 4.1 B.

4.2 Ramachandran Plot

The validity of the optimized structure predicted from Alpha Fold was done through SAVES, and PROCHECK. The analysis revealed that 75.6% of residues are present in the most favoured regions, 15% of amino acids in additionally allowed, 3.3% amino acids in generously allowed region, 6.1% amino acids in disallowed regions.

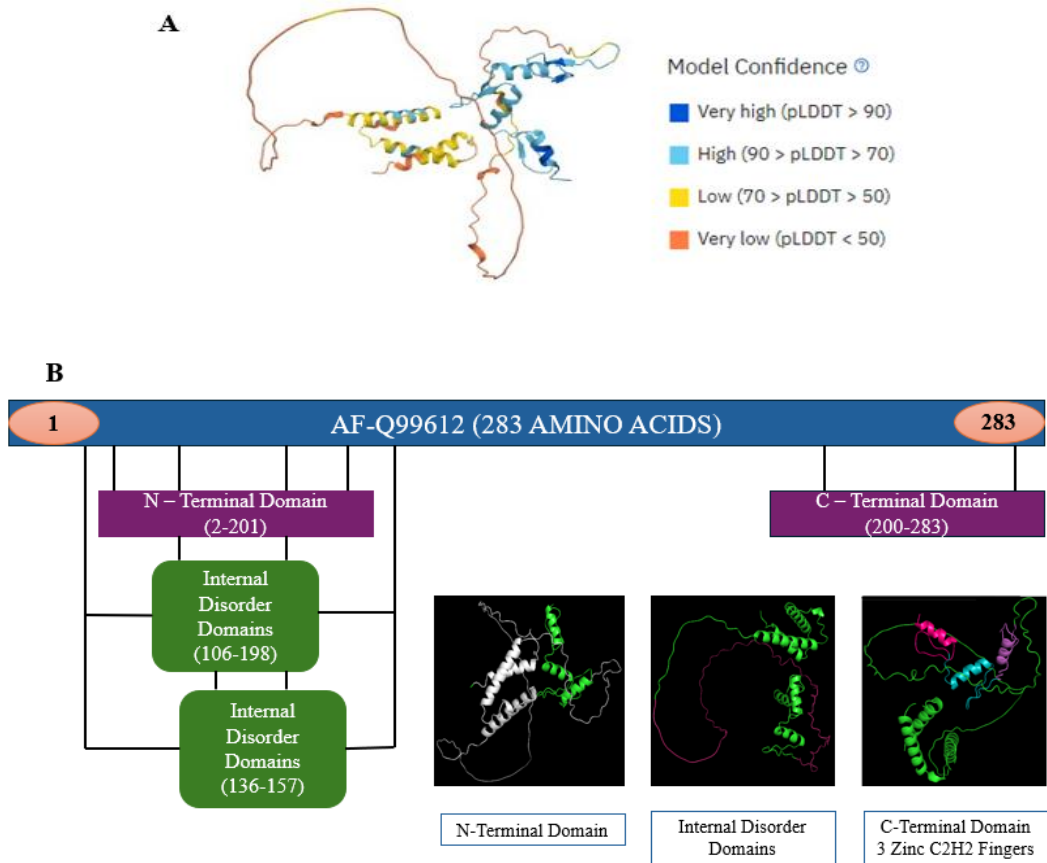


Figure 4.1: Shows the predicted structure using Alpha fold with all three C-terminus domains.
B: Shows detailed domains of KLF6 structure.

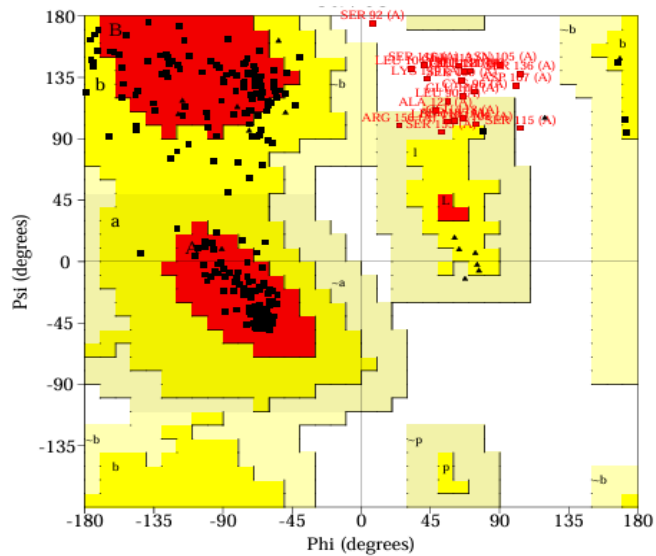


Figure 4.2: Ramachandran Plot showing the measurements of angles in KLF6.

4.3 Subcellular Localization

The subcellular localization of KLF6 protein was done via Deeploc 1.0, the localization is shown through a figure 4.3 and a table 4.1 below. The red line in the figure 4.3 depicts the path that the klf6 protein follows inside a cell. This software provides a probability or a likelihood score meaning where in the cell the protein is found most. The result shown by deeploc 1.0 predicts that the klf6 protein is found highly in nucleus with a probability score of 0.9822 and is soluble with a likelihood of 0.9885.

Table 8.1: Probability score of KLF6 subcellular localization.

Localization	Likelihood	Type	Likelihood
Nucleus	0.9822	Soluble	0.9885
Cytoplasm	0.0168	Membrane	0.0115
Extracellular	0.0007		
Cell membrane	0.0002		
Mitochondria	0		
Golgi apparatus	0		
Lysosomes/Vacuole	0		
Plastid	0		
Endoplasmic reticulum	0		

Peroxisomes	0	
-------------	---	--

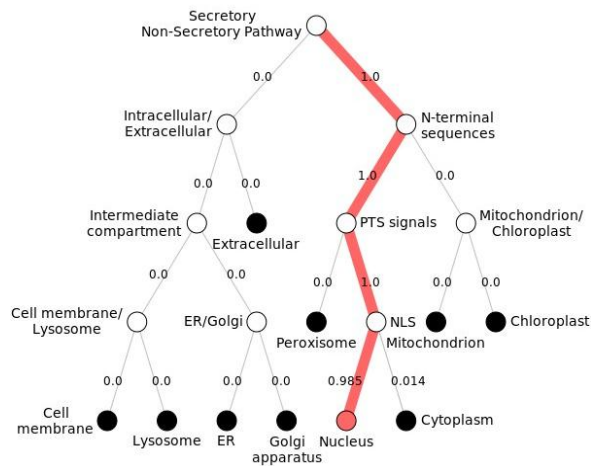


Figure 4.3: Shows the KLF6 protein's localization route as well as the likelihood score. The path of localization is shown in red.

4.4 Phylogenetic Tree

Phylogenetic tree was constructed using two different software tools: Clustal W and Clustal Omega. The tools depict the evolutionary relation between the KLF families.

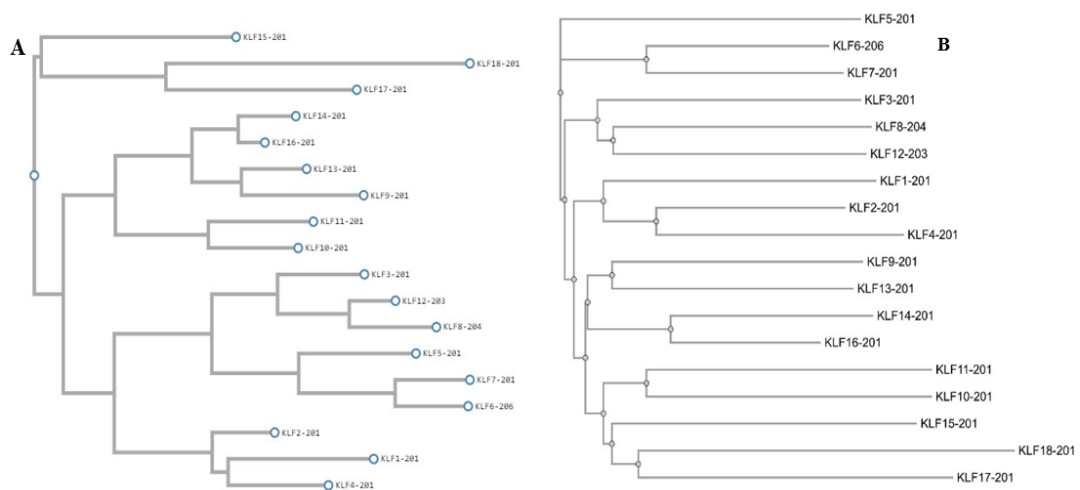


Figure 4.4: Phylogenetic trees of the KLF6 proteins.

The generated tree confirms that the KLF family originated from a common root and has evolved over. The tree shows that klf6 is closely related to klf7 and less related

to klf5. In clustal W, the function “build” of ETE3 3.1.2 (Huerta-Cepas et al., 2016) was used and multiple sequence Alignment was provided for phylogenetic reconstructions. Tree was generated using FastTree v2.1.8 with default parameters (Price et al., 2010).

Furthermore, clustal omega and consurf was used to align the isoform sequences of KLF family which confirms that some regions in all the members of the family are conserved, and some have variations. The C-terminal domain is the highly conserved region while N-terminal region has the most variations this further confirms that the difference in functional activity of the proteins lies in N-terminal as it is varied, and structural properties lies in C-terminus domains due to its conservation over the years.

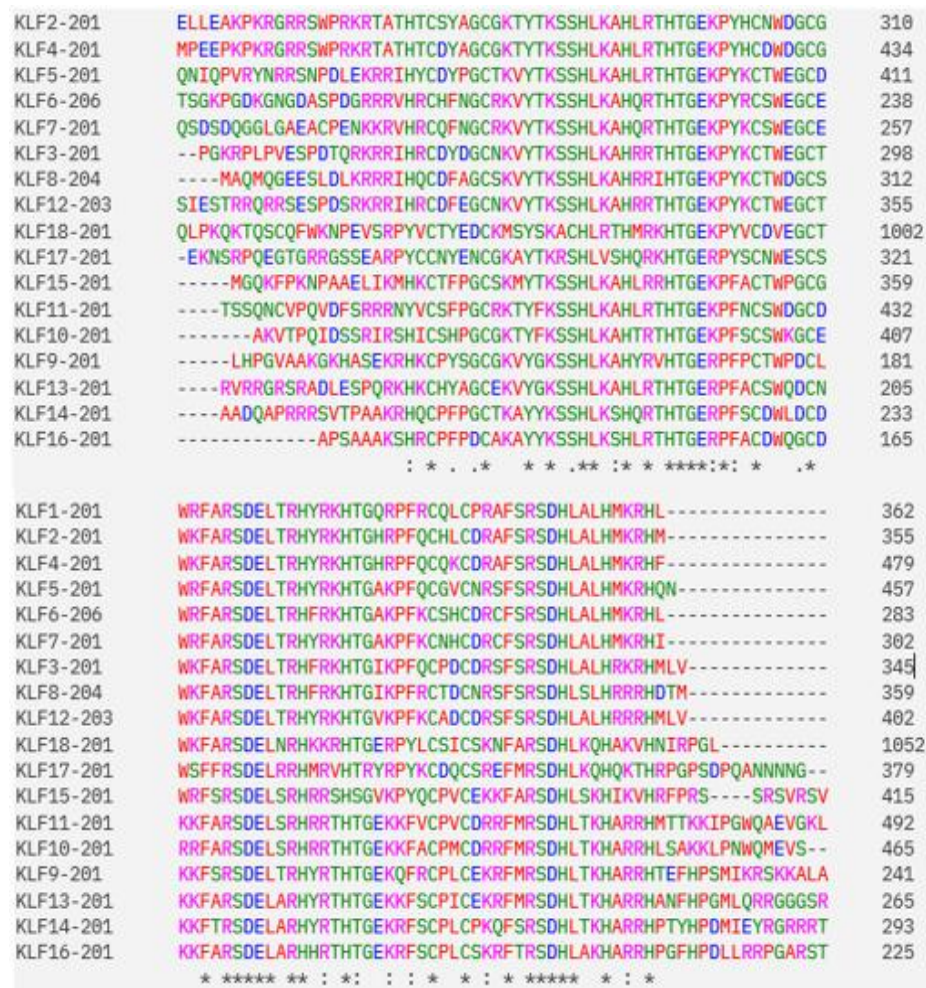


Figure 4.5: Sequence alignment through Clustal Omega.

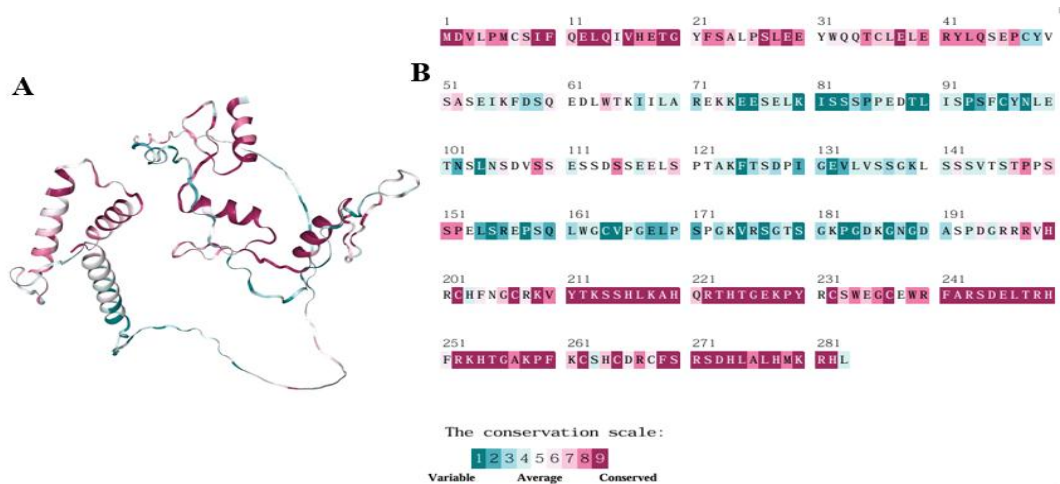


Figure 4.6: A: Structure of KLF6 via Consurf. B: Sequence alignment using Consurf.

4.5 1 KLF6 Variant Identification

Data of KLF6 variants with a total of 469 were retrieved from different databases; Ensembl, genomAD and COSMIC. Data from ensemble included 149 missense, 4 frameshift, 4 non-sense and 37 splice site variants, from genomAD 139 missense, 2 frameshift, 1 non-sense and 17 splice site variants while COSMIC included 84 missense, 10 frameshift, 11 non-sense and 11 splice site variants. Out of 500 variants, 372 were unique and 128 variants were redundant. Unique 372 variants include 345 missense, 15 frameshift, 9 nonsense and 3 spliced variants.

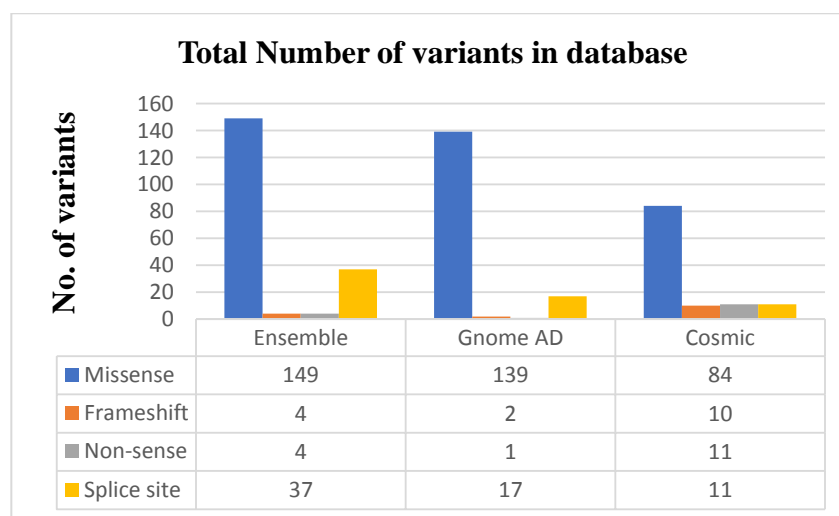


Figure 4.7: Total number of variants in database.

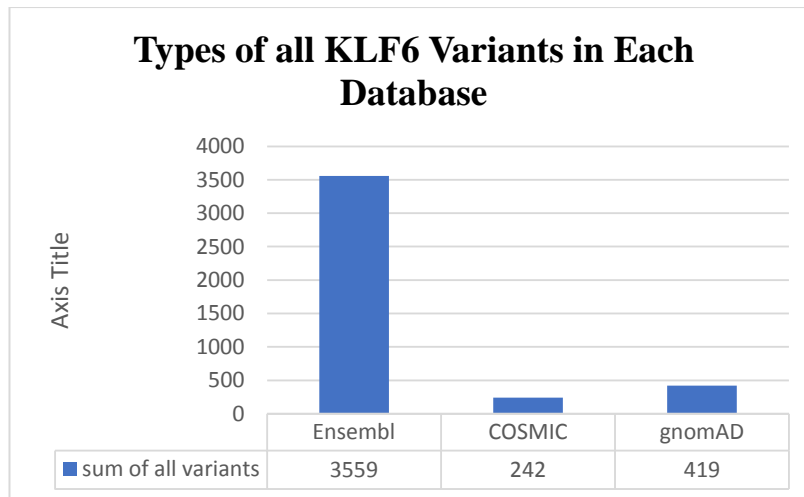


Figure 4.8: Types of all KLF6 variants in each database.

Missense variants were present abundantly among the filtered unique variants, the missense variants were filtered out from the other types of variants. Furthermore, the number of residues per exon was analysed. Klf6 possesses 4 exons and each one of them encodes amino acids. Exon 1, 2, 3, 4 encodes 34, 192, 41 and 16 amino acids with SNP frequency of 7, 133, 7, 2 respectively.

Among the 4 exons the maximum number of amino acids are covered by exon 2 also with the highest number of missense variants observed. The total number of variants present in the exons of klf6 are 5069 SNPs, this data was collected from ensemble. Only missense/non-synonymous variants were filtered out for further processing, a total of 150 non-synonymous SNPs were selected, and pathogenicity was calculated. With the help of pathogenicity calculating tools including SIFT, PolyPhen, CADD, MetaLR, Mutation Assessor and REVEL.

All tools have their own scoring methods hence, a unanimous scoring system was introduced so that each tool can be evaluated on it. For benign, likely benign, tolerated, neutral and low the score given was 0, for medium, possibly damaging was given a

value of 0.5 and for deleterious, likely deleterious, damaging, probably damaging was given a value of 1. The pathogenicity percentage of missense variants was calculated.

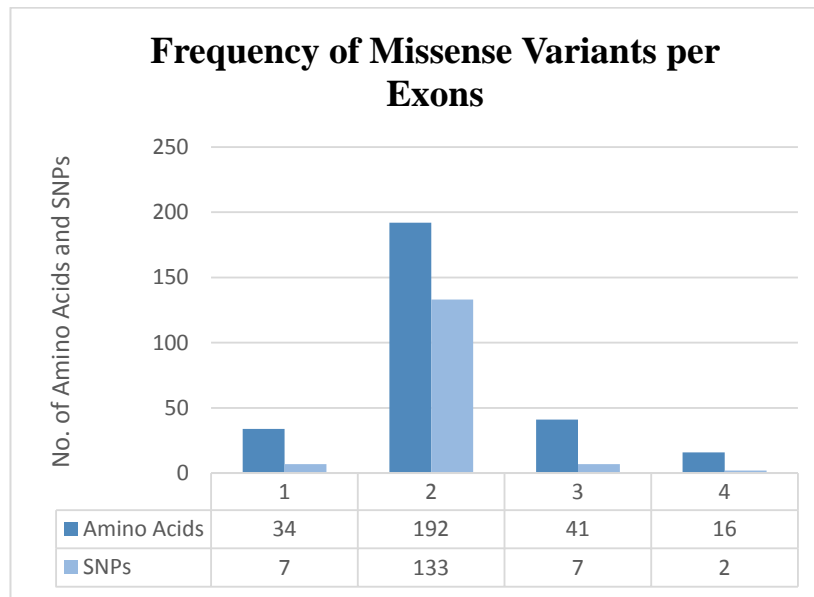


Figure 4.9: Frequency of missense variants per exon.

Table 4.2: Pathogenicity table of selected missense variants.

Variant ID	AA	AA coord	Pathogenicity %
rs1253659402	M/R	6	25
rs1253659402	M/T	6	25
rs1198407776	M/L	6	16.66666667
rs1355084908	Q/R	11	33.33333333
rs780547330	Y/C	21	33.33333333
rs762752613	E/D	29	33.33333333
rs752443335	E/K	30	25

rs776165698	Y/N	49	16.66666667
rs770573018	V/I	50	16.66666667
rs1035370294	S/A	51	25
rs1321018733	A/D	52	16.66666667
rs1390532368	A/T	52	0
rs981973114	E/D	54	8.333333333
rs746555690	D/G	58	0
rs1386659633	D/H	58	33.33333333
rs1424432923	E/A	61	8.333333333
rs748060230	D/G	62	33.33333333
rs758376769	D/N	62	8.333333333
rs778831293	L/V	63	41.66666667
rs778831293	L/M	63	41.66666667
rs121909142	W/R	64	41.66666667
rs1249031510	A/V	70	25
rs1249031510	A/G	70	8.333333333
rs866307815	R/L	71	0
rs866307815	R/Q	71	0

rs1257403042	R/W	71	8.333333333
rs1179625032	E/V	75	8.333333333
rs753610916	E/K	78	0
rs991610969	I/K	81	8.333333333
rs750581486	S/F	82	16.66666667
rs756318202	S/P	82	8.333333333
rs756318202	S/T	82	8.333333333
rs141749814	P/S	86	8.333333333
rs1482239005	E/D	87	16.66666667
rs775004517	D/E	88	8.333333333
rs1203137980	L/P	90	8.333333333
rs1434006255	L/V	90	0
rs1564296185	I/M	91	8.333333333
rs1564296189	I/L	91	8.333333333
rs587778436	S/R	92	0
rs763252895	S/N	92	0
rs1162850438	P/R	93	25
rs1349342383	P/S	93	8.333333333

rs1269937482	N/K	98	0
rs771561652	N/T	98	0
rs747587726	L/V	99	0
rs778728731	E/K	100	8.333333333
rs768589446	S/R	103	16.66666667
rs1329980440	S/N	109	8.333333333
rs1403721472	S/F	112	33.33333333
rs779866252	S/C	113	41.66666667
rs1588344147	S/N	115	41.66666667
rs121909139	S/P	116	25
rs111256842	E/G	117	8.333333333
rs1227606738	E/K	117	16.66666667
rs147965199	L/P	119	41.66666667
rs757426719	L/F	119	41.66666667
rs947615559	S/F	120	41.66666667
rs947615559	S/C	120	41.66666667
rs144914426	T/M	122	41.66666667
rs144914426	T/R	122	16.66666667

rs121909141	A/D	123	8.333333333
rs765619505	K/N	124	0
rs1564296078	T/N	126	25
rs533788452	D/H	128	0
rs533788452	D/N	128	8.333333333
rs1379698490	P/R	129	33.33333333
rs761334451	P/S	129	0
rs371389756	I/T	130	0
rs1432985435	I/V	130	8.333333333
rs1313893989	G/A	131	0
rs1251971839	L/V	134	16.66666667
rs769716317	V/I	135	25
rs184241704	S/R	136	16.66666667
rs121909140	S/L	137	8.333333333
rs121909140	S/W	137	33.33333333
rs1159580959	G/A	138	8.333333333
rs11252089	S/F	142	33.33333333
rs11252089	S/Y	142	33.33333333

rs200767950	S/F	143	16.66666667
rs933440517	S/P	143	16.66666667
rs777889308	T/I	145	8.333333333
rs777889308	T/N	145	25
rs1405050966	S/T	146	16.66666667
rs199602374	T/M	147	41.66666667
rs753240281	P/L	152	41.66666667
rs1457284828	L/P	154	25
rs121909144	S/R	155	25
rs1330721231	S/G	155	8.333333333
rs1039504110	Q/H	160	0
rs537666833	L/P	161	8.333333333
rs537666833	L/Q	161	0
rs1267743907	W/C	162	25
rs943489190	W/S	162	25
rs1200996825	G/S	163	8.333333333
rs375144655	C/S	164	8.333333333
rs375144655	C/Y	164	16.66666667

rs61731927	V/L	165	0
rs61731927	V/M	165	0
rs769806292	P/L	166	0
rs142749289	P/S	166	0
rs201647969	G/R	167	25
rs1411042502	E/K	168	8.333333333
rs121909143	L/P	169	25
rs368266204	L/V	169	8.333333333
rs148536819	S/L	171	8.333333333
rs1157965962	P/S	172	25
rs949283678	K/N	174	33.33333333
rs777779469	V/L	175	8.333333333
rs777779469	V/M	175	8.333333333
rs151112485	R/H	176	33.33333333
rs758505830	R/C	176	33.33333333
rs758505830	R/G	176	8.333333333
rs1241031582	S/R	177	0
rs754239177	G/V	178	0

rs754239177	G/E	178	8.333333333
rs779417302	G/R	178	8.333333333
rs372338890	S/L	180	8.333333333
rs972634245	S/P	180	0
rs751230861	G/C	184	0
rs751230861	G/S	184	0
rs763596063	D/N	185	0
rs1289748441	G/E	187	8.333333333
rs1381298553	G/R	187	25
rs762548312	D/Y	190	25
rs762548312	D/N	190	25
rs1436189659	A/V	191	8.333333333
rs11544695	D/Y	194	41.66666667
rs11544695	D/N	194	33.33333333
rs542676563	G/C	195	25
rs542676563	G/S	195	0
rs748290439	R/Q	201	25
rs772132474	R/W	201	41.66666667

rs111949823	N/K	205	25
rs756516743	G/S	206	58.33333333
rs750705462	R/S	208	25
rs1380328511	R/K	208	25
rs1245944345	K/E	209	83.33333333
rs764830375	R/Q	222	83.33333333
rs200187706	T/M	223	75
rs1360065203	S/L	233	41.66666667
rs771511975	G/R	236	66.66666667
rs747407014	R/C	240	58.33333333
rs773677946	E/K	246	75
rs1170260252	R/Q	252	66.66666667
rs779682182	K/R	258	58.33333333
rs1490355715	S/P	263	41.66666667
rs190182913	A/T	276	41.66666667
rs1381754260	M/L	279	33.33333333

75% pathogenicity filter was applied to filter out the most lethal SNPs that have a high potential of causing deleterious effects to the structure of *klf6*. After applying the

filter 4 SNPs were shortlisted and out of them 2 variants with the highest pathogenicity were selected with rs1245944345 and rs764830375 IDs, both had 83.3% pathogenicity score. Further protein analysis was performed on these two variants.

Table 4.3: Missense variants after threshold pathogenicity sorting. (D= Damaging/Deleterious, PD= Probably damaging, LB= Likely benign LD= Likely deleterious, T= Tolerated, DC= Disease causing, H= High, M= Medium, N= Neutral, L= low)

Variant ID	Conseq. Type	AA	AA coord	SIFT	PolyPhen	CADD	REVEL	MetaLR	Pathogenicity %
rs1245944345	missense variant	K/E	209	D	LD	DC	D	L	83.33333333
rs764830375	missense variant	R/Q	222	D	LD	DC	D	L	83.33333333
rs200187706	missense variant	T/M	223	D	LD	LB	D	M	75
rs773677946	missense variant	E/K	246	D	LD	DC	T	N	75

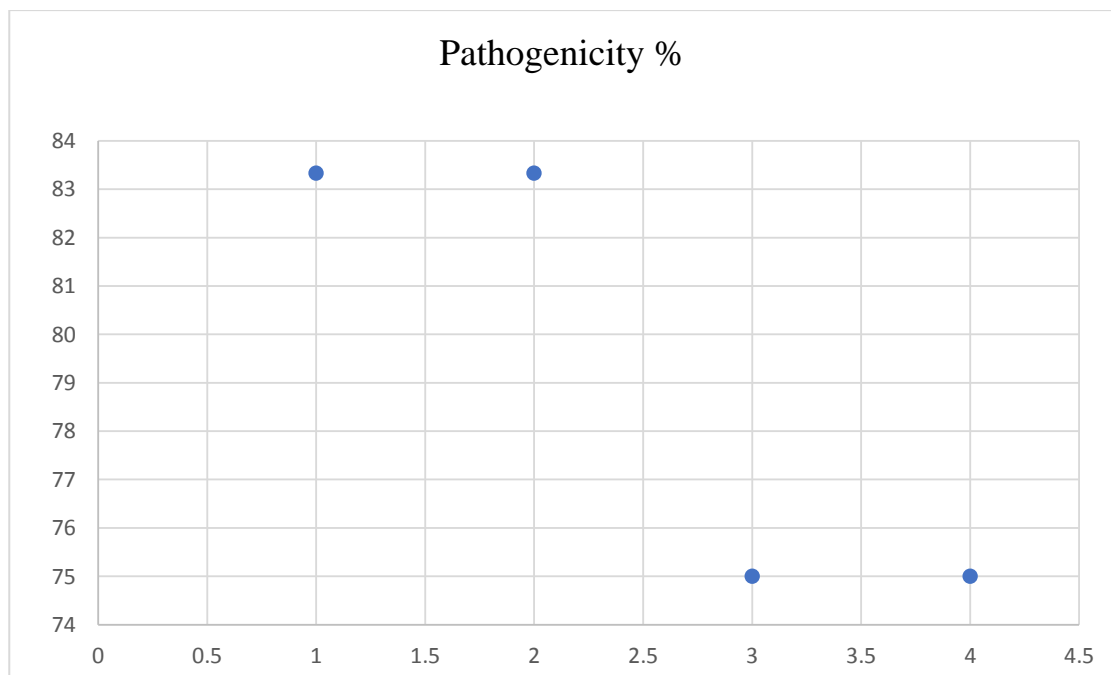


Figure 4.10: Percentage pathogenicity of missense variants per exons.

4.6 Protein Stability Analysis

Protein stability analysis of KLF6 protein was performed by using MuPro, DynaMut2 and MAESTRO WEB. The structural and functional dynamics of the two selected missense variants rs1245944345 and rs764830375 were analysed using the tools. MuPro gives a $\Delta\Delta G$ value which shows the overall protein stability. MuPro calculated the Gibb's free energy for the missense variants which were -0.127 kcal/mol, -0.61 kcal/mol for K209E and R222Q respectively. This indicates destabilization of the missense variants.

DynaMut2 indicated destabilizing stability change for both the variants, showing $\Delta\Delta G$ Stability -0.46 kcal/mol and -0.95 kcal/mol for K209E and R222Q respectively. MAESTRO WEB provided DDG values for selected missense variants which were 1.287 kcal/mol for rs1245944345 with C_{pred} 0.788 and 0.839 with C_{pred} 0.808 for rs764830375. These results show that the selected variants were destabilized due to the mutation as $DDG > 0$ suggests destabilization variation given C_{pred} score 0.0 (not reliable) and 1.0 (highly reliable).

Table 4.4: Protein Stability Analysis.

Tools	rs1245944345		rs764830375	
	DDG Value	Consequence	DDG Value	Consequence
MuPro	-0.127 kcal/mol	Decrease Stability	-0.61 kcal/mol	Decrease Stability
DynaMut 2	-0.46 kcal/mol	Decrease Stability	-0.95 kcal/mol	Decrease Stability
MAESTRO WEB	1.287 kcal/mol	Decrease Stability	0.839 kcal/mol	Decrease Stability

4.7 Variant's Effect on Protein's Structure and Function

The structural and functional changes due to the amino acid sequences were analyzed using MutPred2, HOPE and DynaMut tools. MutPred2 generated data which predicts that the mutation in rs1245944345 at K209E alters the disordered interface, DNA binding and a gain of strand with probability values of 0.55, 0.27, 0.25, respectively. While the mutation in rs764830375 at R222Q causes altered DNA binding and disordered interface with P value of 0.05 and 0.18, respectively.

Table 4.5: MutPred2 results.

rsIDs	Substitution	MutPred2 score	Remarks			
			Type	Property	Property score	P-value
rs1245944345	K209E	0.819	Altered	Disordered Interface	0.55	1.5e-03
			Altered	DNA binding	0.27	0.02
			Gain	Strand	0.25	7.1e-03
rs764830375	R222Q	0.618	Altered	Disordered Interface	0.18	0.05
			Altered	DNA binding	0.16	0.04

DynaMut helps in understanding the protein flexibility because of changes in vibrational entropy energies ($\Delta S_{Vib} ENCoM$). The change in mutation in rs1245944345 at K209E causes increase in flexibility of the structure with a value of $\Delta S_{Vib} ENCoM$ 0.097 kcal.mol⁻¹.K⁻¹. The mutation in rs764830375 at R222Q causes

increase in flexibility of the structure with a value of $\Delta S_{Vib} ENCoM$ 0.327 kcal.mol⁻¹.K⁻¹ (Figure 4.11).

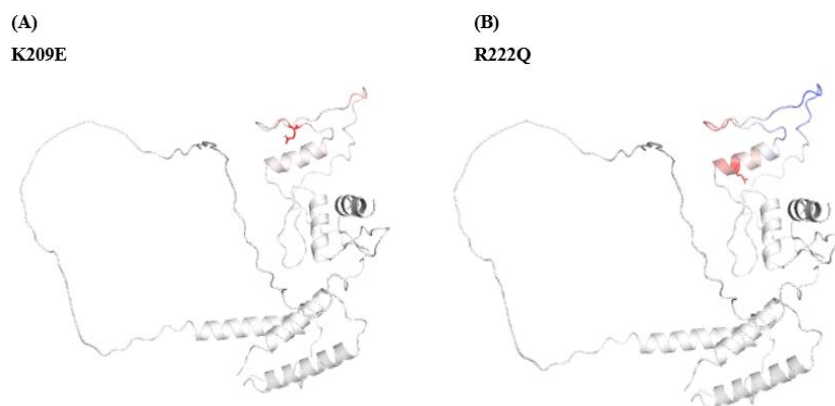


Figure 4.11: Amino acids are coloured according to the vibrational entropy change upon mutation. BLUE depicts gain in rigid structure and RED a gain in flexibility.

Dynamut results depicts that the mutation at position K209E in rs1245944345 and position R222Q in rs764830375, causes changes in inter-atomic interactions, this can be observed as in the wild-type structure there was hydrogen bonding enhancing the stability of the structure while in the mutant structure the hydrogen bonds are lost leading to overall instability in the structure.

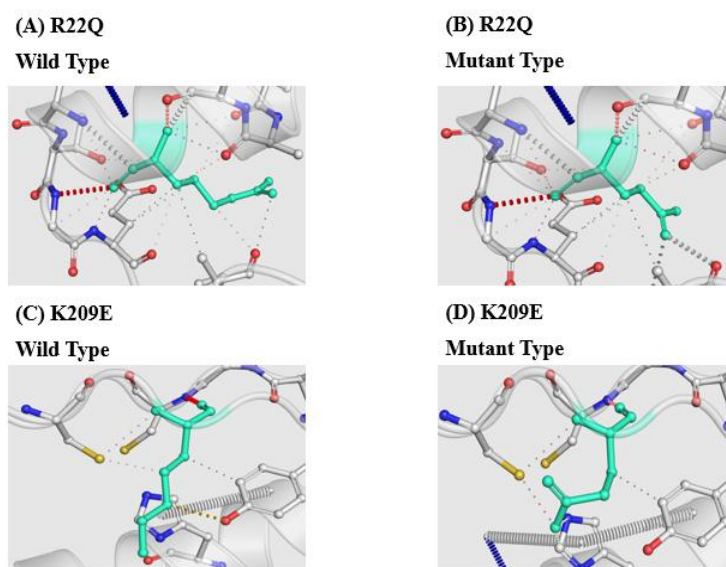


Figure 4.12: Intra-atomic interactions of (A) R222Q and (B) K209E.

4.8 Project HOPE

The Project HOPE analysis revealed that the mutant strain for both the rs IDs was smaller than the wild-type structure. The smaller size causes complications such as repulsion with other protein molecules. The mutant strain of R222Q has a neutral charge and while the mutant K209E has negative charge while the wild-type residue charge of K209E and R222Q was positive. This difference in the charges of wild and mutant residues indicate that loss of interaction of mutant strain with other ligands and protein molecules.

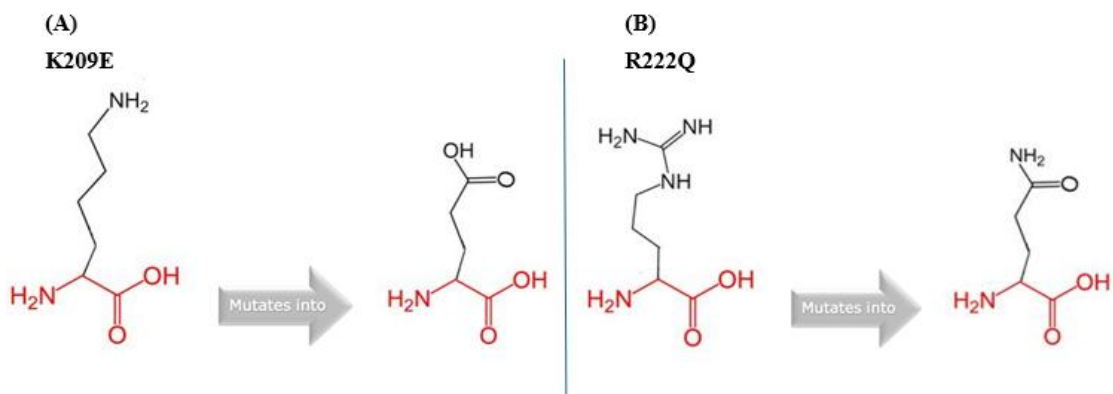


Figure 4.13: (A) Mutated from Lysine to Glutamic acid at position 209 (B) Mutated from arginine to glutamine at position 222. Red colour shows conserved regions and black shows uniqueness.

The mutation in rs1245944345, rs764830375 at K209E and R222Q respectively was found in highly conserved region of the structure. The mutation was similar to the actual residue but mutation at a highly conserved region is damaging as it alters its size causing it to become smaller than normal. The MetaRNN score was calculated by HOPE which was 0.818 and 0.835 suggesting that this mutation is pathogenic. The higher the score between 0-1 the higher the pathogenicity.

Table 4.6: Project HOPE Analysis.

Mutation	Size	Charge	Domain	Conservation	MetaRNN score	Pathogenicity
R222Q	M<W	NEUTRAL	Zinc-finger domain	Highly conserved	0.8352181	Damaging
K209E	M<W	NEGATIVE	Zinc-finger domain	Highly conserved	0.818492	Damaging

4.9 RNA FOLD

4.9.1 *Effects of SNPs on mRNA Secondary Structure*

RNA Fold was used to predict the mRNA secondary structure and the wild type KLF6 structure. The tool calculated minimum free energy (MFE) for both wild type and mutant structures. The mRNA secondary structures were compared with the wild type klf6 structures with significant variations in MFE value.

For rs1245944345 T/C and rs764830375 C/T shows change in MFE indicating that these single nucleotide changes in the structure of wild type have caused changes in the overall mRNA. The MFE for wild type structure of K209E was **-17.10** kcal/mol and for mutant type the MFE decreased to **-20.40** kcal/mol while the R222Q wild type had **-13.10** kcal/mol and for the mutant the MFE increased to **-12.10** kcal/mol.

This indicates that when the mutant K209E in which the nucleotide T has changed to C caused the mutant strain to become more stable with lower MFE value as, lower the MFE value the stable the mRNA structure on the other hand the change in C to T nucleotide in R222Q indicates less stable mRNA structure as the MFE is more positive.

4.10 MD Analysis

The interactions of wild type and mutant protein was performed using the MD analysis for better understanding of the proteins. The simulation ended up in generating different files for 20ns. The simulation results were then plotted in graphs which made the interaction of proteins, and the effect of the mutation has on the protein structure easier to analyse. The analysis of the root means square deviation (RMSD), radius of gyration, solvent accessible surface area (SASA), and the number of hydrogen bonds were used to compare the wild type and mutant proteins.

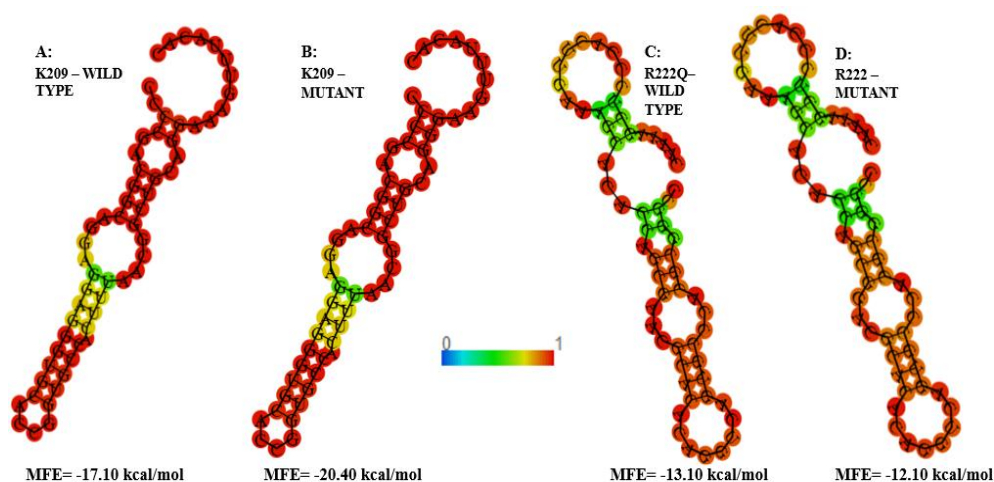


Figure 4.14: MFE value of mRNA Secondary Structures.

The RMSD analysis shows the behaviour of different atoms in a protein structure and how they deviate from their normal position. The figure 4.15 represents that the mutated protein structures do not follow the same pattern as the wild type and deviates significantly. The variations were observed over a period of 20ns and plotted in a graph using RMSD values.

The higher the values of RMSD, lesser stable the structure becomes. The RMSD value of the three structures are almost the same at the beginning but later the RMSD value of wild type structure decreases and then increases stating that the structure went

from being stable to unstable. The mutant strain K209E is comparatively most stable. The RMSD value of R222Q is increased, indicating lower stability.

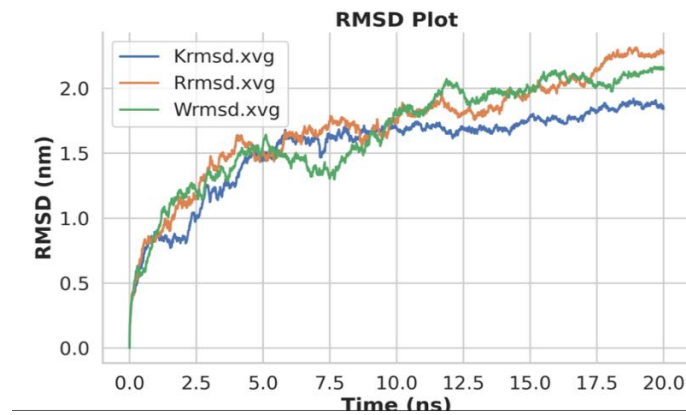


Figure 4.15: The RMSD graph depicts structural changes in wildtype and mutant protein over 20ns.

The radius of gyration (Rg) is the radial distance between all the atoms in a protein and their common axis. Calculating Rg of a protein depicts the protein's folding and compactness over time. If the value of radius of gyration is higher, it means the protein structure is less compact on the contrary if the value is lower it depicts compact structures. As seen in the plotted graph below, the radius of gyration for both structure is starting from 2.7nm and ending at different values. The wild structure has Rg value of 2.1 and mutated type structure has an Rg value of 1.9 at 20ns. This result means that the wild type is less compact than the mutated structure.

The Root Mean Square Fluctuation (RMSF) measures the deviation of individual protein residues from their average positions. Figure 4.16 shows the results of RMSF for both wild and mutant type strains. The root mean square fluctuation of mutant strain R222Q is higher starting at 0.9 and going as high as 1.6 at 120 residues but after this the radius of gyration drops and RMSF increases in wild type structure. The RMSF of K209E is lower than both the structures and remains constant throughout the simulation.

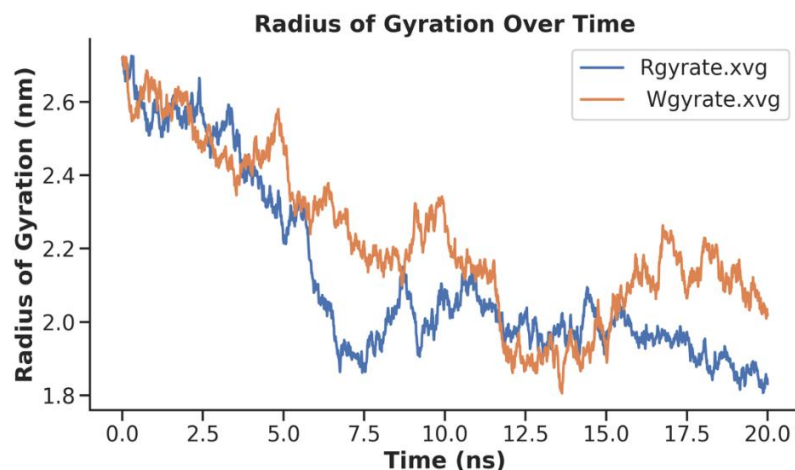


Figure 4.16: The Rg graph depicts structural changes in wildtype and mutant protein over 20ns.

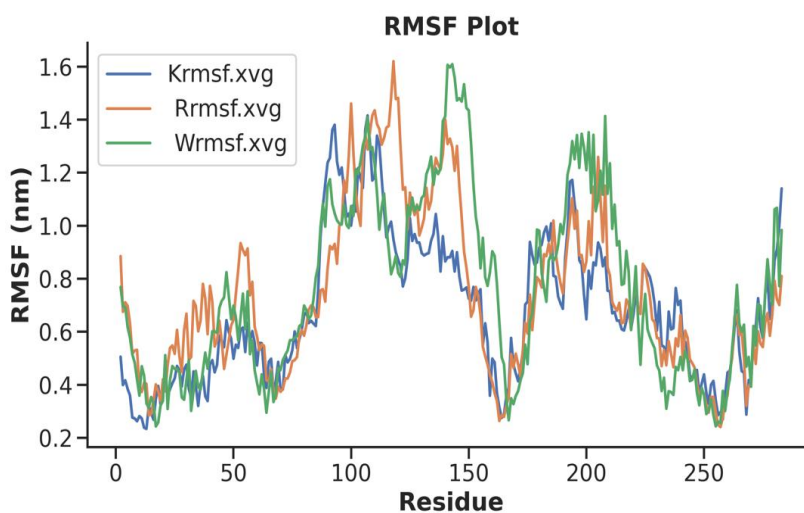


Figure 4.17: The RMSF graph for mutant protein and wild type over 20ns.

Hydrogen bonds are present in a protein's structure and are important in determining the protein interactions and structures. The simulations help in determining how many hydrogen bonds are formed or broken during the time. The figure 4.18 shows that the hydrogen bonds over time have steadily increased from 100 to 160 in mutant strain K209E. The wild and mutant R222Q had lower hydrogen bonds at the beginning but over time the hydrogen bonds increased. The results showed overall increased stability of the structures. The mutant K209E is more stable than wild type structure.

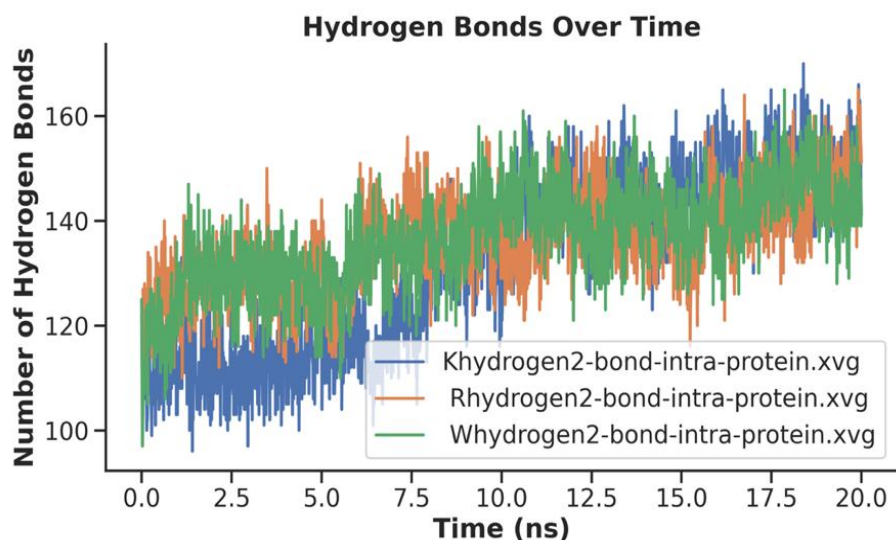


Figure 4.18: The Hydrogen bonds over time for mutant protein and wild type over 20ns.

Solvent accessible surface is (SASA) is a technique in which polarity of surface, exposed or buried amino acids is determined by investigating the surface of the desired protein. As seen in the plotted graph the SASA values for both wild and mutant type are gradually decreasing and there is almost no difference in their end point value this indicates less solubility and more compactness of the structures.

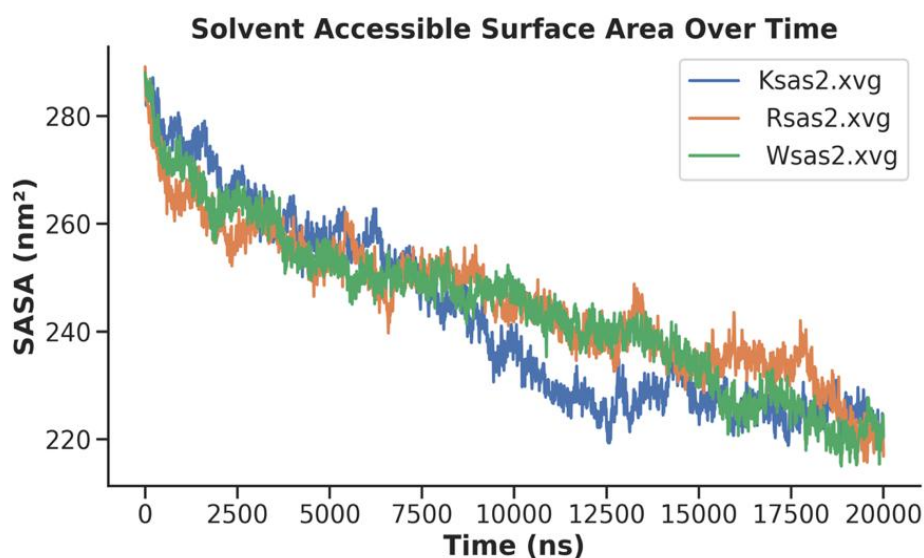


Figure 4.19: The SASA graph for mutant protein and wild type over 20ns.

4.11 Laboratory based Experimentation Results

4.11.1 Genotype Data of Breast Cancer Patients and Healthy Control Samples

The SNP R222Q was analysed in the laboratory, the SNP showed GG, AA and AG alleles were found in both control and diseased samples

Table 4.7: Genotypic distribution data of KLF6 missense variants in both control and patient samples.

Genotype	Patient %	Control %	Odds Ratio	95% CI – Odds Ratio	Relative Risk	95% CI – Relative Risk	P - Value
AA	5.00%	4.00%	1.263	0.3646 to 4.215	1.117	0.5300 to 1.699	>0.005
GG	68.00%	36.00%	3.778	2.068 to 6.622	1.962	1.449 to 2.719	<0.005
AG	27.00%	60.00%	0.2466	0.1381 to 0.4579	0.4804	0.3369 to 0.6648	
G	81.00%	66.00%	2.196	1.075 to 2.321	1.537	1.075 to 2.321	<0.005
A	66.00%	81.00%	0.4553	0.2395 to 0.8575	0.6999	0.5403 to 0.9323	

The result of breast cancer genotyping data revealed that the R222Q SNP is present in diseased as well as healthy control samples in AG and GG form with odds ratio and relative risk of 0.2466, 3.778 and 0.4804, 1.962 respectively, this indicates that this genotype is related to progression and development of breast cancer on the other hand the homozygous AA allele has odds ratio and relative risk lesser than one

indicating no significant results indicating a protective effect against the progression of breast cancer.

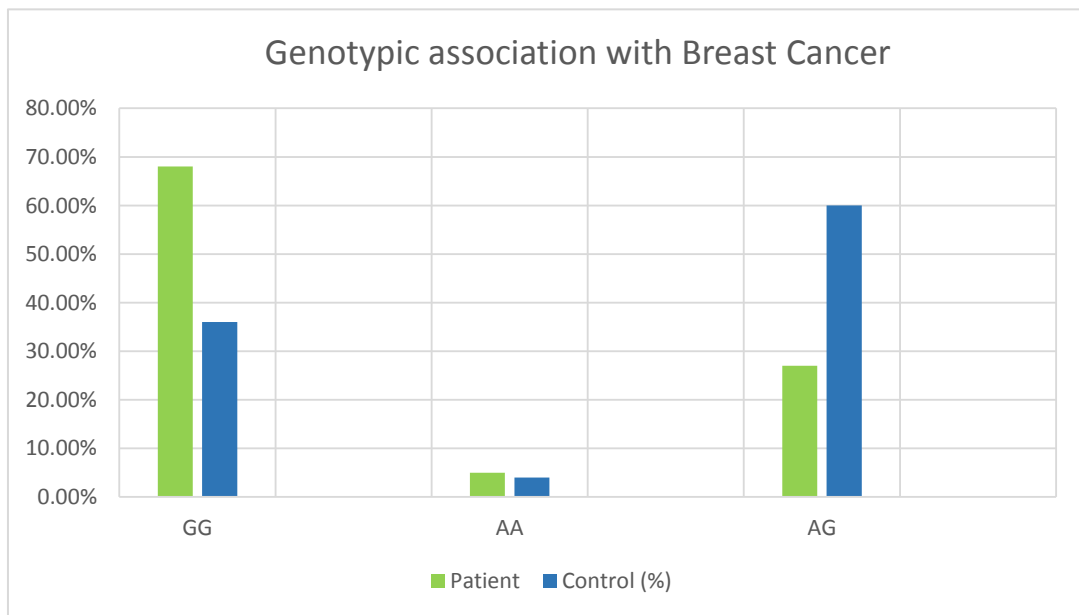


Figure 4.20: Genotypic association with breast cancer.

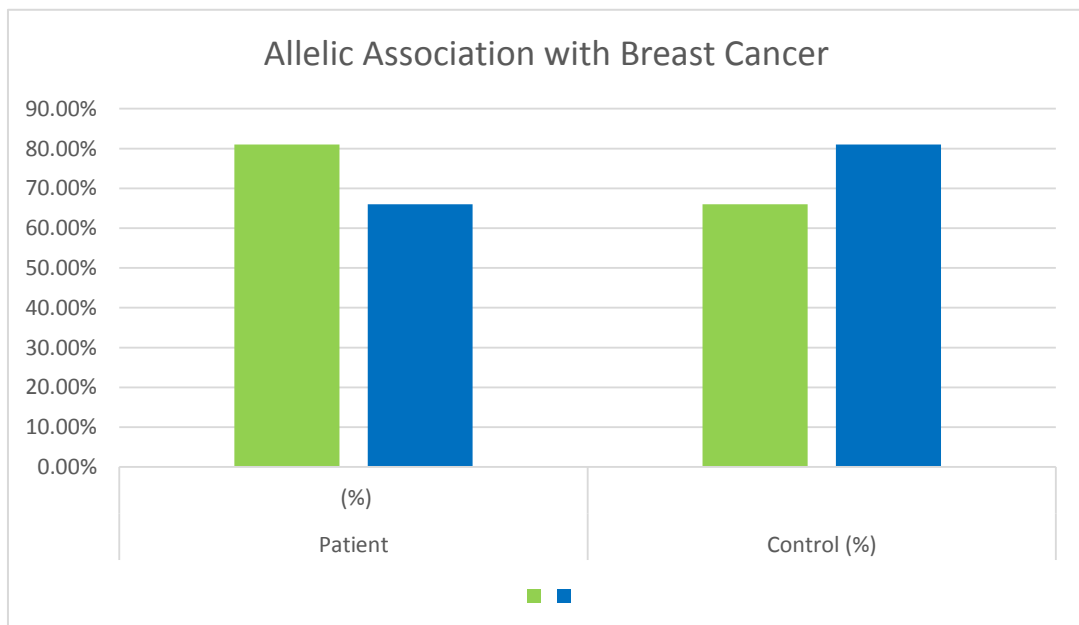


Figure 4.21: Allelic association with breast cancer.

CHAPTER 5: DISCUSSION

Cancer is the second most leading cause of death around the globe and breast cancer is most prevalent in females. Cancer is a generic disease and there have been significant advances in this field, but new research needs to be done to discover better diagnostic and therapeutic approaches. One of the hallmarks of cancer is the dysregulation of genetic factors, one of which is KLF6. This protein is a part of KLF family of tumor repressors and activators, studying the SNPs of KLF6 can be potential in discovering new biomarkers. Deeper understanding is required for which genotyping and analyses of this protein is required.

The primary objective of this study was to identify non-synonymous KLF6 SNPs, structural prediction, its localization in the cell, distinctiveness, molecular dynamics, phylogenetics and evolutionary relationships. For this purpose, genotyping data from blood samples of healthy and diseased individuals were used. this study further aims to explore different and new biomarkers, understanding its correlation with breast cancer and discovering new solutions that can help in the prevention, diagnosis and treatment. Understanding the structure of proteins, helps in decoding its function. As proteins are one of the most important biomolecules, responsible for carrying out different chemical processes inside a cell, its necessary to understand the protein better by determining its structure. (Kuhlman & Bradley, 2019)

The 3D structure of protein can be predicted through various experimental techniques, but computational structure prediction is the considered more accurate (Floudas & bioengineering, 2007). Alphafold was used to predict the structure of KLF6, whose bases work on deep learning algorithms making this tool highly accurate (Varadi et al., 2022). Further understanding of KLF6 structure was done by visualizing domains

using InterPro. KLF6 has a DNA binding domain and transcription activation domain. The DNA binding domain is present at C-terminal, and it possesses three zinc finger domains. The transcription activation domain is present at N-terminal. (Matsumoto et al., 2006)

The KLF6 protein was found to be present largely in the nucleus with a likelihood of 0.9822 and is present in soluble form with a likelihood of 0.9885. Localization of KLF6 protein was determined using the computational tool, DeepLoc 1.0. DeepLoc 1.0 predicts the localization of proteins with the help of the provided protein sequence. (Almagro Armenteros et al., 2017) The behaviour of wild and mutant type was observed via Molecular Dynamics Simulations at 20ns using GROMACS (Van Der Spoel et al., 2005). The results of MD simulations were then interpreted in 5 different graphs, RMSD values showed an ascending trend, indicating more stability of K209E mutant structure than R22Q when seen in comparison.

The RMSF graph depicted increased fluctuation in R222Q at 120 residue indicating flexibility of the molecule as compared to its wild type. The radius of gyration value for the mutant structure is increased as compared to the wild type structure revealing structural compactness. The number of hydrogen bonds increased from initial to final 20ns time interval, making the structures stabilized. SASA values decreased same for wild and mutant type structure depicting lower solubility. There are different variants of KLF6 protein such as missense, splice, nonsense and frameshift. Study of Missense variants was the main goal of this study hence, missense variants were selected from Ensemble Genome Browser (Fernández et al., 2010).

The selected variants were then further filtered for pathogenicity this was done by setting a threshold of 75%. Variants of pathogenicity score of 83.33% were selected as

they showed highest score which was measured using six different tools which included REVEL (Ioannidis et al., 2016), CADD, Meta LR, SIFT (Sim et al., 2012), Mutation Assessor and PolyPhen. All these tools measure the score by amalgamating various genetic variant annotations into a single result also known as C score. (Kircher et al., 2014) The filtered variants were then investigated for their relation to breast cancer. Tetra ARMS PCR (Medrano & De Oliveira, 2014) was used in genotyping study. The results showed that GG and AG were prevalent, associated, and causing the patients predisposed to breast cancer.

5.1 Prospects

This study on missense variants of KLF6 and their association with breast cancer has some significant insights. KLF6 can be a potential target in breast cancer diagnosis as its expression is reduced in p53-mutant breast cancer. (Sun et al., 2020). As evident by the genotyping data, these specific SNPs are present in Pakistan's population which can help in development of personalized therapeutic strategies. Further research involving SNPs will help in deep understanding of how environmental factors and lifestyle affects a person in developing breast cancer. (Seidman et al., 1982)

As KLF6 expression gets reduced in breast cancer, certain drug targets can be identified which can help in boosting its expression. Moreover, functional effects of SNPs can be studied to understand the progression of breast cancer. Genetic counselling is another new technique that can be used in the future to help the patients in making better decisions and opting the right treatment strategies. Deeper research in understanding the SNPs associated with breast cancer can further uncover new biomarkers, better preventive and diagnostic approaches. Longitudinal research and

global collaboration can help accelerate new data and findings which can then be applied practically.

CHAPTER 6: CONCLUSION

Cancer is a disease that has been wrecking havoc across the globe, now it has been established due to extensive research that SNPs are associated with different cancers. For better understanding how cancers occur, its progression, diagnosis and treatment, it is essential to identify new pathogenic single nucleotide polymorphism. In this study, two highly pathogenic variants K209E and R222Q were selected, their structures, domain analysis, cellular localization, evolutionary study and phylogenetic analysis were performed. Molecular dynamics simulations examined the structural flexibility, stability, compactness. Genotyping of KLF6 was performed to determine the link between breast cancer and its genetic variants. The results provided a significant relation of GG and AG genotype with breast cancer which can further help in early detection as it can be used as a biomarker.

REFERENCES

- Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. J. B. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *33*(21), 3387-3395. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. J. S. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *1*, 19-25.
- Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. J. C. p. i. h. g. (2013). Predicting functional effect of human missense Polyphen-2 *76*(1), 7.20. 21-27.20. 41.
- Blackadar, C. B. (2016). Historical review of the causes of cancer. In *World Journal of Clinical Oncology* (Vol. 7, Issue 1, pp. 54–86). Baishideng Publishing Group Co., Limited. <https://doi.org/10.5306/wjco.v7.i1.54>
- Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., . . . Raj, S. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic acids research*, *49*(D1), D344-D354.
- Brayer, K. J., & Segal, D. J. (2008). Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochemistry and Biophysics*, *50*(3), 111–131. <https://doi.org/10.1007/S12013-008-9008-5>
- Burguin, A., Diorio, C., & Durocher, F. (2021). Breast Cancer Treatments: Updates and New Challenges. *Journal of Personalized Medicine*, *11*(8), 808. <https://doi.org/10.3390/JPM11080808>
- Cain, E. H., Saha, A., Harowicz, M. R., Marks, J. R., Marcom, P. K., & Mazurowski, M. A. (2019). Multivariate Machine Learning Models for Prediction of Pathologic Response to Neoadjuvant Therapy in Breast Cancer using MRI features: A Study Using an Independent Validation Set. *Breast Cancer Research and Treatment*, *173*(2), 455. <https://doi.org/10.1007/S10549-018-4990-9>
- Cancer Prevention and Control in Pakistan: Review of Cancer Epidemiology and Challenges. (2020). *Liaquat National Journal of Primary Care*. <https://doi.org/10.37184/lnjpc.2707-3521.1.20>

- Collins, A., & Ke, X. J. T. O. B. J. (2012). Primer1: primer design web service for tetra-primer ARMS-PCR. 6(1).
- DeLano, W. L. (2002). PyMol. In.
- Dunlavy, D. M., O'leary, D. P., Klimov, D., & Thirumalai, D. J. J. o. C. B. (2005). HOPE: A homotopy optimization method for protein structure prediction. 12(10), 1275-1288.
- Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., . . . Futreal, P. J. B. j. o. c. (2006). COSMIC 2005. 94(2), 318-322.
- Floudas, C. A. J. B., & bioengineering. (2007). Computational methods in protein structure prediction. 97(2), 207-213.
- Fernández, X. M., Birney, E. J. V., & Genetics, M. s. H. (2010). Ensembl genome browser. 923-939.
- Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. In *Cell* (Vol. 100).
- Hassanpour, S. H., & Dehghani, M. (2017). Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice*, 4(4), 127-129. <https://doi.org/10.1016/J.JCRPR.2017.07.001>
- Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., . . . Bhai, J. J. N. a. r. (2021). Ensembl 2021. 49(D1), D884-D891.
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., . . . Karyadi, D. J. T. A. J. o. H. G. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. 99(4), 877-885.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . Potapenko, A. J. N. (2021). Highly accurate protein structure prediction with AlphaFold. 596(7873), 583-589.

- Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Alty, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H. J. E., ... Palchik, J. D. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. <https://doi.org/10.1038/nature11412>
- Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. J. N. g. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *46*(3), 310-315.
- Kumar, A., & Purohit, R. (2014). Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs. *PLOS Computational Biology*, *10*(4), e1003318. <https://doi.org/10.1371/journal.pcbi.1003318>
- Kuhlman, B., & Bradley, P. J. N. R. M. C. B. (2019). Advances in protein structure prediction and design. *20*(11), 681-697.
- Łukasiewicz, S., Czezelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review. In *Cancers* (Vol. 13, Issue 17). MDPI. <https://doi.org/10.3390/cancers13174287>
- McConnell, B. B., & Yang, V. W. (2010). Mammalian Krüppel-like factors in health and diseases. *Physiological Reviews*, *90*(4), 1337–1381. <https://doi.org/10.1152/PHYSREV.00058.2009>
- Miller, I. J., & Bieker, J. J. (1993). A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Krüppel family of nuclear proteins. *Molecular and Cellular Biology*, *13*(5), 2776–2786. <https://doi.org/10.1128/MCB.13.5.2776-2786.1993>
- Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., & Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnology*, *26*(1), 101–106. <https://doi.org/10.1038/NBT1374>
- Nüsslein-volhard, C., & Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. *Nature* *1980* *287*:5785, *287*(5785), 795–801. <https://doi.org/10.1038/287795a0>

- Parkins, A. C., Sharpe, A. H., & Orkin, S. H. (1995). Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature*, 375(6529), 318–322. <https://doi.org/10.1038/375318A0>
- Pei, J., & Grishin, N. V. (2013). A New Family of Predicted Krüppel-Like Factor Genes and Pseudogenes in Placental Mammals. *PLOS ONE*, 8(11), e81109. <https://doi.org/10.1371/JOURNAL.PONE.0081109>
- Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H.-J., . . . Iakoucheva, L. M. J. N. c. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. 11(1), 5918.
- Rodrigues, C. H., Pires, D. E., & Ascher, D. B. J. N. a. r. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. 46(W1), W350W355
- Simon, A., & Robb, K. (2022a). Breast Cancer. *Cambridge Handbook of Psychology, Health and Medicine, Second Edition*, 577–580. <https://doi.org/10.1017/CBO9780511543579.131>
- Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. J. N. a. r. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. 40(W1), W452W457.
- Suzuki, T., Yamamoto, T., Kurabayashi, M., Nagai, R., Yazaki, Y., & Horikoshi, M. (1998). Isolation and Initial Characterization of GBF, a Novel DNA-Binding Zinc Finger Protein That Binds to the GC-Rich Binding Sites of the HIV-1 Promoter. *The Journal of Biochemistry*, 124(2), 389–395. <https://doi.org/10.1093/OXFORDJOURNALS.JBCHEM.A022124>
- Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., . . . Zhu, H.-P. J. I. j. o. b. s. (2017). Risk factors and preventions of breast cancer. 13(11), 1387.
- Syafruddin, S. E., Mohtar, M. A., Nazarie, W. F. W. M., & Low, T. Y. (2020). Two sides of the same coin: The roles of KLF6 in physiology and pathophysiology. In *Biomolecules* (Vol. 10, Issue 10, pp. 1–22). MDPI AG. <https://doi.org/10.3390/biom10101378>

- Upadhyay, A. (2021). Cancer: An unknown territory; rethinking before going ahead. In *Genes and Diseases* (Vol. 8, Issue 5, pp. 655–661). Chongqing University. <https://doi.org/10.1016/j.gendis.2020.09.002>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., . . . Laydon, A. J. N. a. r. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. 50(D1), D439-D444.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. J. o. c. c. (2005). GROMACS: fast, flexible, and free. 26(16), 1701-1718.
- Weigelt, B., Horlings, H. M., Kreike, B., Hayes, M. M., Hauptmann, M., Wessels, L. F. A., De Jong, D., Van De Vijver, M. J., Van't Veer, L. J., & Peterse, J. L. (2008). Refinement of breast cancer classification by molecular characterization of histological special types. *The Journal of Pathology*, 216(2), 141–150. <https://doi.org/10.1002/PATH.2407>
- Yin, W., Wang, J., Jiang, L., & James Kang, Y. (2021). Cancer and stem cells. In *Experimental Biology and Medicine* (Vol. 246, Issue 16, pp. 1791–1801). SAGE Publications Inc. <https://doi.org/10.1177/15353702211005390>
- Zhang, J., Li, G., Feng, L., Lu, H., & Wang, X. (2020). Krüppel-like factors in breast cancer: Function, regulation and clinical relevance. In *Biomedicine and Pharmacotherapy* (Vol. 123). Elsevier Masson SAS. <https://doi.org/10.1016/j.biopha.2019.109778>



Handwritten initials 'YS'

Dr. Yasmin Badshah
Assistant Professor
Atsah-Rahman School of
Applied Biotechnologies (ASAB)
KUST Iskandar

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below

Submission author	Zainab Nisar
Assignment title	Quick Submit
Submission title	Analysing the Pathogenicity of KRAS Variants in Breast Cancer
File name	Zainab.docx
File size	541.43K
Page count	32
Word count	9,742
Character count	57,467
Submission date	04 Jul 2024 10:25PM (UTC-0700)
Submission ID	2412716095

Working in Partnership with Turnitin & Blackboard



UNIVERSITY OF
KUALA LUMPUR
APPLIED BIOTECHNOLOGY
SCHOOL OF APPLIED BIOTECHNOLOGIES
ATSAH RAHMAN SCHOOL OF APPLIED BIOTECHNOLOGIES
KUALA LUMPUR UNIVERSITY OF TECHNOLOGY (KUST)

Dr. Yasmin Saadiah
 Assistant Professor
 Amir-ul-Rahman School of
 Applied Biosciences (ASAB)

Analysing the Pathogenicity of KLF6 Variant in Breast Cancer

ORIGINAL REPORT

10%	7%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

REFERENCING

1	Submitted to Higher Education Commission Pakistan <small>Student Paper</small>	2%
2	www.ncbi.nlm.nih.gov <small>Internet Source</small>	1%
3	prc.hec.gov.pk <small>Internet Source</small>	1%
4	www2.mdpi.com <small>Internet Source</small>	1%
5	Jianping Zhang, Guangliang Li, Lifeng Feng, Haiqi Lu, Xian Wang. "Kruppel-like factors in breast cancer: Function, regulation and clinical relevance", Biomedicine & Pharmacotherapy, 2020 <small>Publication</small>	<1%
6	Kanupriya Jha, Amit Kumar, Kartik Bhatnagar, Anupam Patra, Neel Sarovar Bhavesh, Bipin Singh, Sarika Chaudhary. "Modulation of Kruppel-like factors (KLFs) interaction with their binding partners in cancers through	<1%

acetylation and phosphorylation", *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 2023

Publication

7	www.medsci.org Internet Source	<1 %
8	www.frontiersin.org Internet Source	<1 %
9	www.researchsquare.com Internet Source	<1 %
10	Pace, C.. "Tyrosine hydrogen bonds make a large contribution to protein stability", <i>Journal of Molecular Biology</i> , 20010914 Publication	<1 %
11	www.science.gov Internet Source	<1 %
12	www.eurekaselect.com Internet Source	<1 %
13	hdl.handle.net Internet Source	<1 %
14	Submitted to University of Western Sydney Student Paper	<1 %
15	Liu, Xing. "Investigating the Gating Mechanism of NMDAR by Molecular Dynamics Simulation and Machine Learning", State University of New York at Buffalo, 2023	<1 %

Publication		
16	digibug.ugr.es Internet Source	<1 %
17	Updates in Surgery, 2014. Publication	<1 %
18	www.genome.jp Internet Source	<1 %
19	Annalisa Masi, Amina Antonacci, Maria Moccia, Valeria Frisulli et al. "CRISPR-Cas assisted diagnostics: A broad application biosensing approach", TrAC Trends in Analytical Chemistry, 2023 Publication	<1 %
20	Long Tang, Ping Xu, Lingyun Xue, Yian Liu, Ming Yan, Anqi Chen, Shundi Hu, Luhong Wen. "A novel self-attention model based on cosine self-similarity for cancer classification of protein mass spectrometry", International Journal of Mass Spectrometry, 2023 Publication	<1 %
21	Submitted to University of Sheffield Student Paper	<1 %
22	ia600709.us.archive.org Internet Source	<1 %
23	www.coursehero.com Internet Source	<1 %

24	Gupta, Saurabh, Alka Jadaun, Himansu Kumar, Utkarsh Raj, Pritish Kumar Varadwaj, and A.R. Rao. "Exploration of new drug-like inhibitors for serine/threonine protein phosphatase 5 of Plasmodium falciparum: a docking and simulation study", Journal of Biomolecular Structure and Dynamics, 2015. Publication	<1 %
25	pharmacyeducation.fip.org Internet Source	<1 %
26	www.mjms.usm.my Internet Source	<1 %
27	www.wjgnet.com Internet Source	<1 %
28	Bishajit Sarkar, Md Asad Ullah, Yusha Araf. "A systematic and reverse vaccinology approach to design novel subunit vaccines against dengue virus type-1 and human Papillomavirus-16", Informatics in Medicine Unlocked, 2020 Publication	<1 %
29	cms-sc.bupa.co.uk Internet Source	<1 %
30	etd.library.emory.edu Internet Source	<1 %
31	ijper.org	

	Internet Source	<1 %
32	uncnri.org Internet Source	<1 %
33	worldwidescience.org Internet Source	<1 %
34	www.mdpi.com Internet Source	<1 %
35	Nguyen Quoc Khanh Le, Duyen Thi Do, Trinh-Trung-Duong Nguyen, Quynh Anh Le. "A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features", Gene, 2021 Publication	<1 %
36	Xavier Periole. "Molecular dynamics simulations from putative transition states of α -spectrin SH3 domain", Proteins Structure Function and Bioinformatics, 11/15/2007 Publication	<1 %
37	d.docksci.com Internet Source	<1 %
38	encyclopedia.pub Internet Source	<1 %
39	vdoc.pub Internet Source	<1 %

40 www.freepatentsonline.com <1 %
Internet Source

41 Yuchen Zhang, Chongjie Yao, Ziyong Ju, Danli Jiao, Dan Hu, Li Qi, Shimin Liu, Xueqing Wu, Chen Zhao. "Krüppel-like factors in tumors: Key regulators and therapeutic avenues", *Frontiers in Oncology*, 2023 <1 %
Publication

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off