SPECIMEN FORMAT FOR THESES OF MONTH

Faculty	:	Biosciences
Department	:	Biochemistry, Biotechnology and Bioinformatics
Branch/ Area:	:	Cancer Biology
Sub Subject Heading:	:	-
Candidate's Name	:	Suganya K
Candidate's Address with email	:	17-E, Radhika Avenue EXTN, P.N.Pudur, Coimbatore – 641 041. Email id: sugusweet.k96@gmail.com
Title of the thesis	:	Epidemiological profiling of population-specific risk factors and validation of novel genetic variants and deep learning-driven mitosis detection in breast cancer patients
(i) In Roman Script		-
(ii) In roman Script		-
Nomenclature of Degree:	:	Doctor of Philosophy
Month & Year of Enrolment:	:	January 2021
Month & Year of Registration:	:	January 2021
Month &Year of Submission:	:	May 2024
Month &Year of Award	:	February 2025
Name of Supervisor	:	Dr.S.Sumathi
Designation of Supervisor	:	Professor
Centre/department/school in which research was conducted	:	Biochemistry, Biotechnology and Bioinformatics
University's Name & Address	:	Avinashilingam Institute for Home Science and Higher Education for Women, Bharathi park road, Coimbatore – 641 043

Abstract within 300 words:

Breast cancer is a major global health issue and one of the most common malignancies affecting women globally. Addressing population-specific risk factors, genetic alterations, and disease progression is crucial for developing targeted therapies and minimizing treatment adverse effects. This comprehensive four-phase study analyzed breast cancer risk factors, genetic variants, and disease progression in the Tamil Nadu population. The first phase involved an epidemiological survey of 517 patients, revealing key risk factors such as age (mean 47.40, SD 11), lower education levels, and unemployment. Premenopausal women faced a higher risk ($p<0.001$), and lifestyle factors like reduced physical activity ($p<0.01$), low water intake ($p<0.05$), and shorter sleep duration ($p<0.01$) were significantly associated with breast cancer risk. Additionally, comorbidities ($p<0.01$), tumor location ($p<0.001$), and treatment modalities ($p<0.001$) were associated with breast cancer. In the second phase, whole exome sequencing (WES) of six primary breast cancer samples identified 857 rare genetic variants. Five patients showed cardiomyopathy-associated variants, indicating a need for cardiac evaluation before treatment. Functional analysis revealed that candidate genes played roles in signal transduction, gene regulation, and protein interactions, contributing to cancer progression. Using Sanger sequencing, the third phase validated novel variants MYBPC3: c.2816G>A and PTCH1: c.1889G>A. These genes were linked to the hedgehog and p53 signaling pathways, circadian rhythm, and cardiomyopathy, suggesting a connection between breast cancer and cardiovascular health. In the final phase, convolutional neural networks (CNN) analyzed histopathology images, achieving 96.6% accuracy in mitotic index detection and aiding early cancer progression assessment. Integrating epidemiological, genetic and image analysis approaches open the door for innovative strategies for prevention, treatment, and comprehensive care in the battle against breast cancer. The study's outcome highlights personalized treatment options based on an individual's risk factors and genetic alterations, thereby offering hope for improved patient outcomes and a better quality of life.

i) Major objectives :

- To assess the epidemiological profile of breast cancer-associated risk factors in the Tamil Nadu population - hospital based cohort study.
- To identify rare genetic variants associated with breast cancer by whole exome analysis.
- To validate the identified novel genetic variants using Sanger sequencing.

- Functional and pathway analysis of the rare variants associated with candidate genes.
- To detect mitosis and evaluate breast cancer progression using histopathology images with deep learning methods.

ii) Hypothesis:

Null hypothesis

- There is no significant association between a specific population's epidemiology profile and breast cancer incidence.
- There are no novel genetic variants associated with breast cancer in the specific population.

Alternate hypothesis

- There is a significant association between the epidemiology profile of a specific population and breast cancer incidence.
- There are novel genetic variants associated with breast cancer in a specific population.

iii) Methodology :

3.1 Layout of the study

The study was structured into four phases to accomplish the objectives laid down for the present study.

- Phase I was designed to conduct an epidemiological study to identify and understand the risk factors that may cause breast cancer. This phase involved collecting and analysing data from the hospital based cohort of breast cancer patients to identify patterns, correlations, and potential factors contributing to breast cancer development.
- In Phase II, the study aimed to explore rare genetic variants causing breast cancer by WES. Examining the genetic makeup of breast cancer patients helps to identify specific genetic mutations that might play an important role in the disease in specific-population.
- Phase III involved validating the novel genetic variants in breast cancer patients. This phase focused on confirming the presence and significance of the genetic variants.
- Lastly, the study's Phase IV focused on developing a deep learning-based automated approach for detecting mitosis in breast cancer histopathological images.

iv) Findings:

Breast cancer is a worldwide health problem influenced by a multitude of factors, including population demographics, lifestyle choices, genetics, and environmental factors. These factors lead to genetic alterations and the progression of breast cancer. The incidence, mortality rates, and survival outcomes of breast cancer vary significantly across different regions. Breast cancer incidence is increasing in general and specifically in the Tamil Nadu population because of the sedentary lifestyle and unhealthy food habits. Women are usually diagnosed only during advanced stages because of a lack of awareness. Screening individuals can reduce the burden of breast cancer by enabling early detection and intervention, thereby reducing the need for aggressive treatments and minimizing both side effects and instances of over-diagnosis. Classification of women according to risk factors, genetic mutations, and disease progression can be effective in improving risk-free methods such as regular screening, lifestyle modifications, chemoprevention, and designing targeted therapy.

Epidemiological studies have a significant role in breast cancer research in identifying underlying cause of disease in specific populations. These studies examine various lifestyle and environmental factors, such as diet, physical activity, reproductive history, and exposure to environmental toxins. They help to identify associations between the risk factors and the incidence of breast cancer. These studies also investigate the various genetic risk factors underlying breast cancer including family history of the disease. Family-based epidemiological research can identify familial clusters and potential genetic components of breast cancer cases.

The epidemiological risk factors may cause mutations leading to breast cancer. Exome sequencing is a powerful genetic analysis tool focusing on sequencing the genome's protein-coding regions. This method can identify genetic variants and mutations within the coding genes. Exome sequencing can identify rare genetic variants within specific genes that may increase the risk of developing breast cancer in specific-population. It aids researchers to precisely identify novel candidate genes based on the exome data. Genetic mutations found through exome sequencing can illuminate the biological mechanisms that drive breast cancer development or progression. These insights help to identify genetic alterations based on the epidemiological risk factors in population based studies.

Exome sequencing can identify numerous genetic variants and mutations. However, to confirm the presence of novel variants and mutations with high accuracy, Sanger sequencing is often employed. This validation is a critical step in confirming the specific genetic changes associated with

breast cancer in specific-populations. The genetic information obtained from exome and Sanger sequencing form the basis of designing personalized treatment approaches based on the genetic alterations.

Genetic changes ultimately disrupt the normal functioning of cells, causing them to proliferate abnormally. This abnormal cell growth is a key biomarker used to determine the disease stage. Addressing the proliferation of breast cancer is a crucial consideration in treatment strategies. Specifically, quantifying the mitotic score using histopathology images can aid in the understanding of tumor aggressiveness and proliferation. Histopathology images provide detailed information about the characteristics of breast tumors, including their growth patterns. The high mitotic score associated with specific genetic mutations indicates a more aggressive form of breast cancer. Manual detection of mitosis is time consuming and laborious. Artificial intelligence-based mitosis detection helps to detect more accurately and efficiently. The proliferation score with genetic data and risk factors can guide treatment decisions, potentially leading to personalized and effective therapies.

Overall, the epidemiological studies provide insights into specific populations' breast cancer risk factors. Exome sequencing and Sanger sequencing identifies genetic variants associated with the disease and the rare variants plays a major role in disease progression in specific-population. Mitotic score quantification using AI in histopathology images helps to determine tumor aggressiveness and proliferation. These components may work together to advance the understanding of breast cancer etiology and improve diagnosis and personalized treatment plans based on an individual's genetic profile.

With this background in the current study, we aimed to analyse the risk factors, genetic alterations and disease progression of breast cancer patients. We initiated our investigation by collecting epidemiological data to identify the risk factors prevalent among breast cancer patients. Subsequently, we conducted exome sequencing to delve into the genetic variants associated with the disease. To check the accuracy of our findings, we validated the novel variants through Sanger sequencing. In addition, we performed a convolutional neural network in histopathology images, focusing on the mitotic index as a critical metric for evaluating proliferation. This comprehensive approach allowed us to gain a multifaceted understanding of breast cancer and its determinants.

The study's first phase involved an epidemiological investigation conducted at the Oncology division of Sri Ramakrishna Hospital, Coimbatore after obtaining ethical clearance. The data was collected between January 2021 and May 2023 to identify and understand specific populations' breast cancer risk factors. Data were collected from 517 patients and analyzed using SPSS version 20. The

statistical significance between variables was evaluated using the chi-square test, with a p-value of ≤ 0.05 signifying statistical significance. The 95% confidence interval was utilized to estimate the population range. The study outcome showed significant findings related to demographic characteristics, lifestyle factors, clinical characteristics, and family history.

The detailed analysis of the study findings reveals important insights about the demographic characteristics of breast cancer patients. In our study, we observed that the highest number of affected women were between the age group of 41 to 50, with a mean age of 47.40 (SD 11). Among 517 patients, one male was affected with breast cancer. Breast cancer is rare in males, but hormonal imbalances and family history increases the risk of BC in males. While addressing the socio-economic status, an increased rate of incidence was observed in women with lower levels of education and who were unemployed, and most of them were married. Socio-economic factors such as lower education and unemployment may limit access to healthcare resources, leading to delayed diagnosis and treatment.

On examining the impact of reproductive factors, it was evident that premenopausal women exhibited a higher susceptibility compared to postmenopausal individuals. Women with early menopause before the age of 50 had a higher risk of developing breast cancer. Examination of lifestyle factors showed significant associations between less physical activity, low water intake, and disturbed sleep duration and breast cancer. These observations highlighted the importance of lifestyle changes, including regular exercise, proper hydration, and sufficient sleep in reducing breast cancer risk.

The clinical characteristics of breast cancer patients provided valuable insights into the disease's complexity and formed the basis of treatment planning. The outcome of the survey revealed significant associations between comorbidities and breast cancer. High blood pressure and diabetes are the major comorbid conditions found in patients. The right side of the breast was affected more by the tumor. Coming to the choice of therapy, majority of patients underwent chemotherapy as their first treatment option which resulted in multiple side effects. The findings of family history revealed that a small proportion of breast cancer patients had first-degree relatives diagnosed with breast cancer. The study also showed the major inheritable disease which runs down the family of breast cancer patients was diabetes.

This study comprehensively analysed various factors associated with risk of breast cancer. It reiterates the significance of education, age, menopausal status, lifestyle choices, comorbid conditions, tumor location, and treatment modalities in determining an individual's susceptibility to breast cancer. These findings also suggested the importance of personalized risk assessment in order to facilitate

tailored preventive strategies. Moreover, early detection and access to healthcare, particularly among individuals facing socio-economic challenges, are highlighted as critical components in breast cancer prevention and management.

In the second phase, whole exome sequencing (WES) was done for the representative samples from the survey outcome. Integrating epidemiological studies with WES can thoroughly explain the relationship between risk factors and genetic alterations in disease susceptibility and progression. WES was conducted on six primary breast tumor samples (5-Female, 1-male), including all the breast cancer subtypes such as luminal A, luminal B/HER2-, luminal B/HER2+, and triple-negative breast cancer, and two adjacent normal breast tissue samples. Excellent DNA purity and integrity was seen in all the samples. Whole exome sequencing using the platform Illumina NovaSeq 6000 was done for prepared DNA samples ensuring that their fragment sizes, analyzed using the Tapestation 4150 with D1000 screen tapes, fell within the optimal 350–450 base pair range. Sequencing results consistently maintained a Phred score of 36, attesting to a remarkable 99.9% base call accuracy across all base pairs of the six breast cancer patient samples.

Aligning the exome DNA sequences (FASTQ files) with the human genome (hg38) and the adjacent normal sequence (BC-1 and BC-2) was accomplished using the Burrows–Wheeler Aligner (BWA) package. Subsequently, the VCF files were subjected to comprehensive annotation with ANNOVAR, including details like gene names, transcripts, mutation consequences, and genomic regions (exonic or intronic). Additional functional tools, namely SIFT, Polyphen2, and mutation taster, were employed to assess pathogenicity, aiding in identifying potential disease-causing variants.

The initial analysis revealed an impressive total of 1,01,213 variants across all patients. Post-filtering and annotation identified 857 rare variants in 370 genes, all of which were found to be heterozygous, suggesting a mutation in one of the two gene copies. Notably, these variants were exclusively located in exonic regions and were classified as nonsynonymous single nucleotide variants (SNVs). These SNVs results in amino acid changes that may influence protein structure and function. In BC-4 patient, two specific gene mutations were identified: one synonymous SNV resulting in DNA sequence changes, and another is stop gain mutation leading to premature termination of protein synthesis during translation.

Variant gene analysis showed that cardiomyopathy was the common associated condition among five breast cancer patients. Remarkably, the *TTN* (titin) gene, known for its role in various heart-related conditions, particularly cardiomyopathies, was the common mutation found in five breast cancer cases.

These findings suggest that specific genetic variants may predispose individuals to both breast cancer and cardiomyopathy which are governed by epidemiological risk factors. These variants may influence genes or pathways relevant to both conditions, necessitating potential adjustments in treatment strategies for breast cancer patients who might develop cardiomyopathy. This highlights the importance of comprehensive cardiac assessments before initiating breast cancer treatment.

We also explored non-coding regions, particularly introns, for potential contributions to breast cancer progression. Filtering for intronic variants with an allele frequency of $\leq$0.05 in exome analysis identified 183 variants across six patients, indicating that non-coding regions are also responsible for breast cancer genetics. In our study, we checked for mutations in *BRCA1* and *BRCA2* genes, which are frequently mutated and hotspot genes in breast cancer. We found mutations in *BRCA* genes in all breast cancer patients except BC-2. Discovering mutations in these genes is crucial to identify the risk of breast cancer, implement preventive measures, and develop targeted treatment plans for individuals with a higher genetic tendency for developing breast cancer.

Functional enrichment analysis using FunRich identified rare variants of candidate genes' biological process, molecular function and cellular process. The biological processes of the candidate genes play a crucial role in signal transduction, G-protein coupled receptor signaling pathways, gene expression regulation, synapse assembly, and transcription regulation from RNA polymerase II. The genes are involved in the cellular process of the plasma membrane, and their molecular functions are primarily involved in protein binding. These processes are particularly relevant in the context of breast cancer, as their dysregulation can result in uncontrolled cell growth and proliferation of cancer.

Notably, three new variants were identified among the significant findings. The variants were observed in *MRPL13* c.380T>C (p.Leu127pro), *MYPBC3* c.2816G>A (p.Arg939Gln), and *PTCH1* c.1889G>A (p.Arg630His) and the variant in *PTCH1* gene variant was detected in two different breast cancer patient samples. The new variants had uncertain significance mutation, implying that its impact on disease is not clearly understood till date. These variants did not have the reference ID because it was not reported. We also further checked the databases of dbsnp and VarSome, which are not reported in breast cancer, emphasizing the importance of rare variants playing a role in disease progression in the specific population. The exome sequencing results of breast cancer and the adjacent normal tissue FASTA sequence were deposited in the European Nucleotide Archive database, and the accession number PRJEB74646 (ERP159292) was obtained.

The study revealed the complexity of genetic factors in breast cancer, with identified variants potentially shared with cardiovascular diseases. These insights underscore the need for tailored treatments and comprehensive cardiac evaluations in breast cancer care. The discovery of new variants and their associations with breast cancer provides valuable insights into the disease's complexity.

In the third phase, we did Sanger sequencing to validate the newly identified variants by whole exome sequencing. Sanger sequencing is necessary to validate the novel variants because WES may show few sequencing errors. The variants were observed in *MRPL13, MYPBC3,* and *PTCH1* genes. We designed specific primers for the variants because it was newly reported. We compared the reference genome sequence in NCBI and the flanking sequence in the Ensembl database and identified the altered regions in the sequence. The FASTA sequence was used to design specific primers using Primer3plus software. The forward and reverse primers were subjected to BLAST analysis to confirm their precise targeting of the intended gene. The results showed a 100% match with the gene variant. Following this, *in silico* PCR through the UCSC genome browser was performed to assess the effect of primers targeting specific genomic regions. This involved specifying the genome assembly (GRCh38), chromosome and coordinates to ensure that the primers could generate the desired PCR product.

A touch-down PCR was used to amplify the specific variants, and we confirmed their accuracy through agarose gel electrophoresis. The amplified DNA fragment size closely matched the desired amplicon size designed for the specific primer. This alignment strongly suggested that the PCR product accurately represented the intended DNA region, confirming its suitability for subsequent Sanger sequencing analysis.

Sanger sequencing was then performed for the PCR product of the identified new three variants in breast cancer patients. Variants in the genes of *MRPL13*: c.380T>C (p.Leu127pro) in BC-2 and *PTCH1*: c.1889G>A (p.Arg630His) in BC-6 were not validated, suggesting the absence of mutations in these patients. We successfully validated the presence of *MYBPC3*: c.2816G>A (p.Arg939Gln) and *PTCH1*: c.1889G>A (p.Arg630His) mutations in BC-4 patient. The sequences of *MYBPC3* and *PTCH1,* acquired via Sanger sequencing was subjected to BLAST analysis and verified their alignment with the specific genes of interest. The *MYBPC3*: c.2816G>A variant was observed in the forward primer binding at 131bp and the reverse primer binding at 161bp. In the case of the *PTCH1*: c.1889G>A variant, the forward primer was positioned at 129bp, while the reverse primer binding site was at 152bp. These mutation positions within the primer binding sites were recorded, and chromatograms were used to visualize the sequencing data. *MYBPC3* (Myosin Binding

Protein C) is expressed in cardiac and skeletal muscles. Mutation in *MYBPC3* causes cardiomyopathies. *PTCH1* (Patched 1) is crucial in the hedgehog signaling pathway and is implicated in various cancers. The mutational variants in *MYBPC3* and *PTCH1*, which were validated, were associated with various health conditions, including cardiac health and potential cancer risk. We are the first to confirm these variants in primary breast tumor samples.

Further, we used Network Analyst and KEGG pathway enrichment analysis to map candidate genes' "Interactome network". Specifically, the *MYBPC3* gene showed interactions with 10 genes: *ALK, FBXO32, NCF1, BLNK, CUL1, CAPN1, TNNI3K, TTN, TRIM63,* and *SMURF2*. The *PTCH1* gene displayed interactions with 16 genes: *GLI1, SMO, CCNB1, DNAJA3, SMURF2, IHH, HHIP, SHH, GRK2, SMURF1, SUFU, CDON, GPC3, CDK1, GLI3,* and *GLI2*. Mutations in *MYBPC3* and *PTCH1* may lead to mutations in the interactome genes, disrupting their normal function, as the interactome genes play crucial roles in various cellular processes and signaling pathways.

The *CUL1* gene effectively interacted with *MYBPC3* and involved in multiple cellular pathways, including the circadian rhythm and the hedgehog signaling pathway. These findings tie in with our epidemiological observations of disrupted sleep patterns, and the presence of high blood pressure and diabetes among the patient cohort underscores the importance of addressing circadian rhythm disturbances, as they can significantly affect the overall health, potentially contributing to both cardiomyopathy and breast cancer. In addition, the interacted genes associated with *PTCH1* are primarily involved in the hedgehog signaling pathway. Aberrations in this pathway are reported to initiate breast cancer progression.

The study sheds light on the significance of this genetic alteration in breast cancer and related health conditions in specific population. The confirmation of these novel variants in primary breast tumor samples emphasizes their potential role in disease progression. It provides valuable insights into the mechanisms underlying these health conditions, paving the way for possible new approaches to prevention and personalized treatment based on genetic alteration.

The three phases of the study strongly indicate that risk factors, rare genetic mutations and gene alterations may drive cancer development. Mutated genes affect signaling pathways, cell cycle regulation, and mitosis, causing uncontrolled cell division growth. Tumors with a high mitotic index are associated with a worse prognosis and poorer patient outcomes. Manual detection of mitosis by pathologists can be time-consuming and prone to variability among observers. In hospitals as regular

practice, Ki-67 staining is used to assess cell proliferation, which is expensive and time consuming. So, in the fourth phase of the study, convolutional neural network (CNN) was employed to detect mitosis using breast cancer histopathology images. Mitotic activity can differ significantly between subtypes, further complicating the detection process. This approach aimed to improve diagnosis accuracy and efficiency by providing automation and objectivity, and we used all subtypes of breast cancer images to train the model. A total of 298 histopathology images were used in the study, which included 267 images from the MITOS-ATYPIA-14 dataset and 31 real-time histology images obtained from six breast cancer patients whose DNA samples were subjected to WES and validated by Sanger sequencing. The ground truth images were used to assist in identifying mitosis within these images.

The convolutional neural network architecture used binary classification for breast cancer histopathology images. The architecture included convolutional layers (Conv2D) with ReLU activation and max-pooling layers (MaxPool2D). The model began with a convolutional layer (conv2d_3) configured with 32 filters of size 3x3, followed by a max-pooling layer (max_pooling2d_3) to reduce feature map dimensions. This pattern continued with additional convolutional and max-pooling layers, ultimately ending with a dense output layer for binary classification. To prevent overfitting, a dropout layer was implemented, which randomly sets a fraction of the input units to zero during training. The model's architecture was finalized with a fully connected layer. The model was configured with the Adam optimizer, binary cross-entropy loss function, and accuracy metric. Training encompassed over 25 epochs, utilizing both training and validation data.

The model achieved a remarkable accuracy of 0.96, demonstrating its ability to successfully learn and generalize patterns from training data. The CNN architecture with convolutional layers and max-pooling proved effective in learning hierarchical features. The dropout layer reduced overfitting risks, and the model benefited from ample training data. The Adam optimizer with an appropriately set learning rate facilitated efficient convergence. Training and validation loss and accuracy plots showed that the model effectively learned and generalized without overfitting, making accurate predictions on data. The developed code loads and pre-processes histopathology images and categorizes them based on mitosis scores. We are the first to detect mitosis in real-time histopathology images, which shows the significance of the developed CNN model. This CNN model can potentially be valuable for breast cancer diagnosis and prognosis in the medical field.

The outcome of this comprehensive study which spans four phases contributed valuable insights into breast cancer, its associated risk factors, genetic landscape, and an innovative approach to histopathological image analysis. Age, education, menopausal status, lifestyle factors, comorbid conditions, tumor side, and treatment options have been found to be strongly linked with the incidence of breast cancer. The finding highlights the importance of comprehensive cardiac assessments before initiating breast cancer treatment. Novel variants in *MYBPC3* and *PTCH1* were identified, expanding our understanding of the genetic complexity of breast cancer in specific-population.

Novel genetic variants in *MYBPC3* and *PTCH1* may be valuable biomarkers for more precise diagnosis. Integrating these genetic markers into diagnostic approaches provides clinicians with additional information to guide treatment decisions. The pathway enrichment analysis revealed potential disruptions in cellular processes, focusing on circadian rhythm disturbances, cardiovascular health, and cardiomyopathy. Targeting the pathways associated with *MYBPC3* and *PTCH1* alterations may lead to more effective and selective treatments. Convolutional neural networks (CNN) for mitosis detection in histopathology images proved to be a highly accurate and efficient model. This innovative method holds great promise for enhancing breast cancer analysis, potentially leading to improved patient outcomes.

This multi-phased study has provided a multifaceted understanding of breast cancer, including its potential links to circadian rhythm disruption, cardiac health issues, and genetic mutations in breast cancer. The integration of epidemiological, genetic, and image analysis approaches has opened doors for innovative methods for prevention and treatment strategies. It emphasizes the importance of personalized risk assessment, comprehensive care, and early detection in the battle against breast cancer, providing hope for improved patient outcomes and quality of life. These findings contribute to the greater understanding of the mechanisms of breast cancer and contribute to potential prevention and personalized treatment strategies.

Examiners

Internal Examiner:

Dr.Vasudev R. Thakkar,
Professor
Department of Biochemistry,
BRD School of Biosciences,
Sardar Patel University, Vallabh Vidyanagar-388120, Gujarat, India.

External Examiner:

Dr. Pankaj K. Singh,
Professor,
Department of Oncology Science,
The University of Oklahoma Health Sciences Center in Oklahoma City,
Oklahoma - 73104, United States.