4

RESULTS AND DISCUSSION

Breast cancer has emerged as the predominant form of cancer affecting women globally. The prevalence, incidence, and mortality rates linked to breast cancer demonstrate notable disparities both within individual countries and across different regions globally. These differences can be attributed to various factors encompassing reproductive, metabolic, lifestyle, environmental, and occupational influences. Moreover, advancements in medical technology have not only expanded the availability and accessibility of cancer screening and therapy but have also resulted in a higher diagnosis rate for this disease. Despite these advancements, the burden of cancer still significantly impacts low- and middle-income countries (Mohanty *et al*., 2023).

In India, women face a challenge characterized not only by a high fatality rate, but also due to family neglect of healthcare. The incidence of breast cancer in India exceeds the global average, with 178,361 new cases reported in 2020 alone, accounting for 7.9% of total cases. Additionally, 90,408 deaths were attributed to breast cancer in India, accounting for 13% of global mortality from the disease. Among Indian women, breast cancer contributes to 13.5% of new cancer cases and 10.6% of cancer-related deaths (Sung *et al*., 2021). This elevated burden underscores the importance of conducting population-specific studies to address the risk factors associated with breast cancer comprehensively.

Epidemiological studies are crucial in identifying and understanding modifiable risk factors, forming the foundation for preventive strategies (Iacoviello *et al*., 2021). The risk factors of breast cancer often lead to genetic alterations in the body. Understanding the molecular mechanisms behind these alterations is pivotal for developing targeted therapies that can enhance the effectiveness of cancer treatment and improve patient outcomes (Qin, 2019). Incorporating Artificial Intelligence (AI) technologies into clinical practice can significantly enhance the efficiency and accuracy of breast cancer diagnosis, ultimately resulting in better patient outcomes (Pantanowitz *et al*., 2020).

The objectives of the present study was to extensively examine the epidemiological risk factors, genetic alterations, and disease progression in Tamil Nadu population of breast cancer patients. We have done a hospital-based epidemiological survey to analyse the risk factors. We conducted exome sequencing on primary breast tumor samples and validated the novel variants using Sanger sequencing. Patients were categorized based on molecular subtypes, and the study examined the genetic mutations across all breast cancer subtypes. Additionally, AI technology was employed to detect mitotic figures in histopathology images, providing crucial insights for assessing the rate of cell proliferation. The study was designed to be carried out in four phases.

Phase I

4.1 Assessing risk factors of breast cancer patients at the population-specific level: A hospital-based cohort study

An epidemiology study in breast cancer contributes to understanding the disease, risk factors, and its impact on public health (Tollosa *et al.*, 2020). We conducted an epidemiology study from January 2021 to May 2023 at the Oncology division of Sri Ramakrishna Hospital Coimbatore, one of the state's largest health-care centre. Cases included both primary breast cancer patients and those who underwent treatment. The sample size for this research study was determined to be 377 individuals, calculated using the Raosoft sample size calculator. However, 517 individuals were included in the survey, exceeding the intended sample size. The larger sample size of 517 provides more data and potentially increases the statistical association of the study. The collected epidemiology data demonstrated the link between risk factors and breast cancer incidence within the cohort. Findings from this study could provide valuable insights into the relative importance of various risk factors, inform preventive strategies, and contribute to the dissemination knowledge of breast cancer epidemiology.

4.1.1 Demographic characteristics of breast cancer patients

Demographic variables such as gender, age, race or ethnicity, and family history exert considerable influence on an individual's susceptibility to breast cancer. By classifying high-risk people based on demographic characteristics, targeted

interventions can be implemented to increase screening rates, promote early diagnosis, and ensure timely access to appropriate treatment (Gulzar *et al*., 2019).

Breast cancer cases occur in individuals across various age groups, with different proportions. Our survey analysis revealed in individuals aged between 21 to 30 breast cancer was seen in 4.45% (95% CI: 0.02 - 0.05) of the total cases. While breast cancer occurrences are comparatively less frequent in individuals in their 20s, it remains possible for this age group to receive a diagnosis of breast cancer. The percentage of breast cancer incidence cases was found to increased significantly among individuals in their 30s, with the age group 31- 40 accounting for 25.92% (95% CI: 0.22 - 0.30) of the cases. The highest proportion of breast cancer cases was prevalent among individuals aged 41-50, with an average of 35.40% (95% CI: 0.31 - 0.39). The findings suggest that the incidence of breast cancer tends to rise with age, with women in their 40s having a greater probability of developing the disease compared to younger age groups. Among individuals between 51 and 60, 20.50% (95% CI: 0.17 - 0.24) of breast cancer cases were reported. Breast cancer incidence rates tend to remain relatively high among this age range, highlighting the continued risk of developing breast cancer as people get older. Women of age group 61-70, the percentage of breast cancer cases was 11.60% (95% CI: 0.09 - 0.15). The risk of breast cancer remains a significant concern for individuals in their 60s. Lowest incidence of 2.13% (95% CI: 0.01 - 0.03) was observed in women in their 71-80 age. However, this suggests that the risk of breast cancer continues even as individuals enter their 70s and 80s. The age distribution of the population is depicted in **Figure 10A**. The average age of the study population was 47.40 years (SD 11), indicating that individuals had a mean age of 47.40 years, with a standard deviation of 11 years.

Sofi *et al*. (2019) conducted a study in the National Capital Territory of India and reported that the average age at the period of breast cancer diagnosis was 47 ± 10 years. Similar findings were observed by Mohanty *et al*. (2023) in their research conducted in Mumbai, India, where the average age of breast cancer patients was 47 years. Additionally, Singh *et al*. (2021) found that the maximum number of breast cancer cases in Chhattisgarh, India, was observed with a mean age of 48.7 years (SD ± 10.6). Our study observed a maximum number of affected women aged 41 to 50, with a mean

age of 47.40 (SD 11.00). These findings strengthen the evidence that breast cancer tends to occur predominantly in the 40 to 50 years age range in various regions of India.

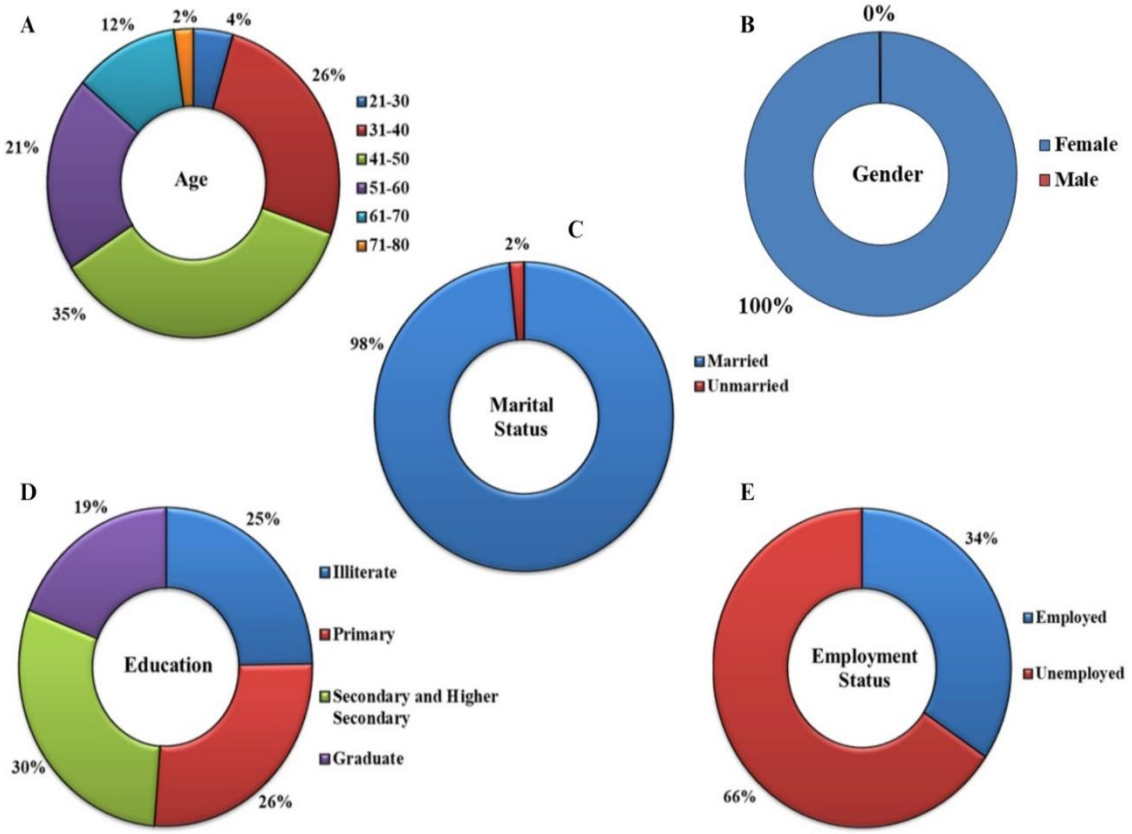
The survey outcome justified that among the 517 patients, one male was diagnosed with breast cancer. According to the results, the study found that females accounted for 99.81% (95% CI: 0.98 - 0.99) of the affected individuals, while males constituted only 0.19% (95% CI: 0.001 - 0.005), given in **Figure 10B**. The result highlights the predominant occurrence of breast cancer in females, with a significantly lower incidence observed in males. Breast cancer is more prevalent among women, but there are few cases reported in men, showing they too are susceptible to the disease. Hormonal imbalances and genetic factors, such as mutations in the *BRCA* genes, are recognized as contributors to elevated breast cancer risk among males (Momenimovahed and Salehiniya, 2019).

According to the research findings, male breast cancer incidence varies across different regions, with the maximum rates in Africa, followed by Europe and America, and the lowest rates in Asia. Studies revealed that a significant proportion of 71.2% of male breast cancer patients were aged 60 years or older (Wang *et al*., 2021). According to Scarpitta *et al*. (2019), the mean age of onset for male breast cancer was reported as 61.3 years. Moreover, they found that in 35 cases of male breast cancer, there was a familial history of the disease. In our cohort research, we also observed a higher prevalence of female breast cancer cases. However, we identified one male breast cancer case; who was 68 years old and had a family history of breast cancer. These findings highlight the importance of considering male breast cancer as a distinct entity and understanding its unique characteristics, including the association with older age and family history.

Marital status is recognized as an important socio-economic determinant that may significantly influence breast cancer outcomes. The data presented in **Figure 10C** reveals that a higher percentage of married individuals (98.45%, 95% CI: 0.97 - 0.99) were affected by breast cancer compared to unmarried individuals (1.55%, 95% CI: 0.01 - 0.03). The statistical analysis revealed that the chi-square (χ^2) value of 56.39 and the

associated p-value of 0.218 indicate no statistical significance between breast cancer incidence and marital status.

Figure 10: Distribution pattern of breast cancer patients based on Socio-demographic profile



Numerous researchers have investigated the link between marital status and breast cancer in a variety of population groups. For instance, a cross-sectional study involving 275 patients found that unmarried women were receiving a late-stage breast cancer diagnosis (Olarewaju *et al*., 2019). Similarly, a population-based analysis of 16,513 patients with metastatic breast cancer revealed that unmarried individuals faced a notably higher risk of mortality compared to married women. Unmarried women had a 15.5% higher probability of breast cancer mortality and a 19.0% increase in overall mortality (Zhu and Lei, 2023). Moreover, a database analysis utilizing Surveillance, Epidemiology, and End Results (SEER) data on 1342 inflammatory breast cancer (IBC) patients highlighted a significantly elevated risk of mortality among unmarried women compared

to married women (Liu *et al*., 2019). However, in our survey, a majority of participants were married, suggesting that marital status might not be a prominent factor influencing breast cancer risk within this specific population.

The relationship between educational status and breast cancer incidence was checked. 24.76% (95% CI: 0.21 - 0.29) of breast cancer cases were observed in illiterate individuals, 26.50% (95% CI: 0.22 - 0.30) in those who had completed primary school level education. 29.60% (95% CI: 0.01 - 0.03) in individuals with secondary and higher secondary education, and 19.14% (95% CI: 0.16 - 0.22) in graduates, is depicted in **Figure 10D**. The chi-square (χ^2) value was 213.20, and the p-value was statistically significant at <0.001, indicating a significant association between education level and incidence of breast cancer. These findings suggest that socioeconomic elements, such as education level, play a significant role in the delayed diagnosis of breast cancer. As per our study findings, as shown in **Figure 10E**, 34.43% (95% CI: 0.30 - 0.38) of individuals affected by breast cancer were employed, while 65.57% (95% CI: 0.61 - 0.69) were unemployed. The survey outcome suggested that unemployed women faced a higher risk of breast cancer.

The key factors contributing to delayed diagnosis and higher mortality in breast cancer cases are lack of health awareness and health education. The survey showed that significant proportions (88.8%) of breast cancer cases were diagnosed only after a delay of more than three months. The result suggests that many women may not be aware of the importance of early detection and may lack knowledge about disease symptoms and screening programs (Gulzar *et al*., 2019). Being unemployed and having a low income was found to be significantly linked to breast cancer incidence. Unemployment, in particular, emerged as a risk factor because of having a low income, which can delay the early detection of breast cancer (Alsolami *et al*., 2019). Hjorth *et al*. (2021) also observed increased mortality rates among women with breast cancer who had primary education, minimum income, and were unemployed. Our survey findings also revealed that most participants were unemployed and had a low education level.

A detailed analysis of the study findings reveals important insights about the demographic characteristics of breast cancer patients. The highest proportions of cases

were observed in the age groups 41-50, suggesting that middle-aged people are more prone to risk of setting breast cancer. A strong association was noted between education level and unemployment, indicating that low education and unemployment may limit access to healthcare resources, which leads to delayed diagnosis of breast cancer.

4.1.2 Association between reproductive variables and breast cancer risk

Reproductive variables play a key role in breast cancer risk due to their influence on hormonal exposure. Factors like menarche age, menopause age, nulliparity, pregnancy, and breastfeeding play crucial roles in determining an individual's risk of breast cancer development. By understanding these reproductive factors, individuals and healthcare professionals can assess an individual's risk profile and make informed decisions regarding preventive measures and regular breast cancer screening (Park *et al*., 2022). The reproductive factors of the patients identified from the cohort are presented in **Table 2**.

Research indicates that age at menarche, which is the onset of a woman's first menstrual period, has been identified as a potential factor associated with the risk of breast cancer. The results of our study identified that among individuals who experienced menarche at the age of 13, approximately 7.17% were affected by breast cancer (95% CI - 0.05 to 0.09). Individuals who had their first menstrual period at the age of 14 had an estimated breast cancer prevalence of approximately 18.02% (95% CI - 0.14 to 0.21). 31.59% of individuals who had menarche at the age of 15 were affected (95% - 0.27 to 0.35). In individuals who had menarche at the age of 16, a higher prevalence of breast cancer was observed, with approximately 43.22% (95%CI - 0.39 to 0.48). Early menarche leads to a longer duration of exposure to estrogen and other hormones, potentially influencing breast cancer risk. The findings suggest that the age of menarche in this cohort did not affect causing breast cancer.

Menstruation-related events play a crucial role as risk factors for breast cancer, with early menarche being a well-established contributor to increased risk, as demonstrated in a meta-analysis involving European women (Song *et al*., 2022). Early menarche is associated with an earlier onset and more frequent ovulatory cycles, leading to extended exposure to elevated ovarian hormone levels, particularly estrogen, for women who experience menarche at a young age.

Table 2: Reproductive characteristics of breast cancer patients

Factors	Number (N)	Percentage (%)	χ^2	p-value	95% CI	
					Lower	Upper
Age at menarche						
13	37	7.17	133.48	0.781	0.05	0.09
14	93	18.02			0.14	0.21
15	163	31.59			0.27	0.35
16	223	43.22			0.39	0.48
Pregnancy						
Nulliparous	3	0.59	100.24	0.418	0	0.012
Parous	505	99.41			0.91	1.08
Age at first childbirth						
<30	411	81.39	100.67	0.406	0.78	0.85
≥30	94	18.61			0.15	0.22
Duration of breastfeeding						
Did not breastfeed	1	0.19	241.00	0.560	0	0.005
1-3 months	6	1.19			0.003	0.021
4-6 months	106	21.0			0.18	0.24
7-12 months	233	46.14			0.41	0.50
≥13 months	159	31.48			0.27	0.35
Abortion						
Yes	177	34.84	49.62	0.448	0.31	0.40
No	331	65.16			0.61	0.69
Menopausal status						
Premenopausal	289	56.01	502.64	***	0.52	0.60
Postmenopausal	227	43.99			0.40	0.48
Age at menopause						
<50	178	78.41	788.46	***	0.73	0.84
≥50	49	21.59			0.16	0.27

***** denotes p < 0.001**

Prolonged exposure to elevated estrogen levels for many years elevates the risk of breast cancer substantially (Liu *et al*., 2019). Case-control studies conducted in Morocco have further substantiated the connection, showing that early menarche (occurring before age 13) is related with a higher risk of breast cancer (Khalis *et al*., 2018). Baset *et al*. (2021) found a notable correlation between early age at menarche and breast cancer development. Their findings indicated that as the age at the onset of menstruation increased, the risk of breast cancer development decreased by approximately 0.83 times. Delaying the onset of menarche may potentially offer a protective effect against breast cancer.

The age of a woman's first full-term pregnancy can impact her risk of breast cancer. An early first pregnancy, particularly before the age of 30, has been linked to a reduced risk of breast cancer. In our cohort study, among the participants, 3 individuals were nulliparous, accounting for 0.59% of the study population (95% CI - 0 to 0.012). 505 parous women were affected by breast cancer, which accounts for 99.41% of the study population, and the 95% CI was 0.91 to 1.08. Among individuals who had their first child at before the age of 30 years, a significant proportion of approximately 81.39% were affected by breast cancer (95% CI - 0.78 to 0.85). For individuals who had their first childbirth at or after the age of 30, the prevalence of breast cancer was observed to be approximately 18.61% (95% CI - 0.15 to 0.22). The study findings indicate that there is no significant implication of breast cancer with pregnancy and childbirth in the cohort.

Numerous studies have repeatedly demonstrated that mothers who breastfeed their children were less prone to breast cancer than those who either did not breastfeed or breastfeed for a shorter duration. In our study, among individuals who did not breastfeed, the prevalence of breast cancer was 0.19% (95% CI - 0 to 0.005). Among those who breastfed for 1-3 months, the percentage of breast cancer incidence was 1.19% (95% CI - 0.003 to 0.021). Among individuals who breastfed for 4-6 months, the prevalence of breast cancer was 21.0% (95% CI - 0.18 to 0.24). A higher proportion was observed at 46.14% who breastfed for 7-12 months (95% CI - 0.41 to 0.50). 31.48% of women breastfed for 13 months were affected (95% CI - 0.27 to 0.35) which was contradictory to other observations. According to the data, there was no significant connection between breastfeeding duration and the risk of breast cancer in this specific cohort. The data indicates that among the studied population, 34.84% of individuals had undergone an

abortion (95% CI - 0.31 to 0.40). 65.16% of individuals reported no incidence of abortion (95% CI - 0.61 to 0.69) again underplaying the role of abortion to breast cancer.

Husby *et al*. (2018) found that an early age at first full-term pregnancies and a higher number of childbirths were linked to a decreased risk of breast cancer, which was consistent with our findings. Moreover, abortions did not appear to influence the risk of breast cancer, which aligns with our reports. Furthermore, women who had their first childbirth at older ages faced an increased risk of breast cancer, although even among those whose first delivery occurred in their 30s, a decrease in breast cancer risk was found. The interval between childbirths also played a role in breast cancer risk. Women with delivery intervals of less than 5 years exhibited a higher risk of breast cancer compared to those with intervals of more than 5 years between childbirths (Park *et al*., 2022). The first full-term pregnancy seemed to have a protective role, triggering maturation of the breast and maximal cellular differentiation, making the breast cells more resistant to carcinogenic effects (Katuwal *et al*., 2019). Antony *et al*. (2018) highlighted that delaying the age of marriage could lead to postponement of the first pregnancy and childbirth, which was considered a risk factor for breast cancer.

Azubuike *et al*. (2022) proposed that being parous or having multiple pregnancies was not notably linked to decreased odds of breast cancer. However, having a history of breastfeeding for a more extended period reduced the chance of developing breast cancer, particularly among younger women and those with ER-negative disease. Furthermore, they observed that advanced age at the full-term pregnancy was strongly associated with an elevated risk of breast cancer, especially among older women and those with ER-negative and triple-negative breast cancer. Moreover, Sankar *et al*. (2022) found that a history of induced abortion was not substantially connected with an elevated risk of breast cancer. They also observed that having a history of abortion did not increase the risk of developing breast cancer. Furthermore, they identified a higher breast cancer risk among women who had not breastfed their children, highlighting breastfeeding as a protective factor. Our study findings are consistent with the existing research, which suggests that breast cancer risk is not significantly influenced by pregnancy, childbirth, breastfeeding, or abortion. In our cohort study, these reproductive factors showed no association and did not influence the breast cancer risk.

The duration of a woman's menstrual cycle, known as the menstrual life span, is linked to breast cancer risk. Women with an extended menstrual life span, including early menstruation onset and late menopause, might have a slightly elevated breast cancer risk (Goldberg *et al*., 2020). In the cohort study of patients who were menstruating, around 56.01% were affected by breast cancer. The 95% confidence interval (CI) for premenopausal women ranges from 0.52 to 0.60. On the other hand, among individuals in postmenopause, approximately 43.99% were affected by breast cancer, with a 95% confidence interval (CI) ranging from 0.40 to 0.48. The chi-square (χ^2) value for the association between menstruation status and breast cancer was 502.64 with a corresponding p-value of <0.001. It signifies a statistically significant association between menopausal status and breast cancer.

The survey findings found that individuals who are menstruating have a higher risk of breast cancer compared to menopausal women. The data suggests that the highest risk of breast cancer incidence (78.41%) was found in women who attained menopause before the age of 50 (95% CI - 0.73 to 0.84). 21.59% (95% CI ranges from 0.16 to 0.27) of women who attained menopause at or after the age of 50 were affected. The chi-squared (χ^2) value of 793.21 and the p-value of <0.001 indicate a significant association between menopause and the risk of breast cancer.

In African American women, the association between menstrual and reproductive characteristics and the risk of hormonal receptor positive breast cancer appeared to be more among those under the age of 50, suggesting a significant influence of hormonal factors on breast cancer development in younger women compared to older individuals (John *et al*., 2020). A study involving 775 participants in Taiwan revealed significant associations between a BMI <24 and premenopausal status with an increased risk of early-onset breast cancer at the age of 40 (Yang *et al*., 2022). In our study, too, we observed a higher impact on premenopausal women, and a substantial portion of the postmenopausal patients were below the age of 50. Notably, we found statistically significant connections between these reproductive factors and breast cancer.

The survey findings revealed that premenopausal women have a notably higher chances of getting breast cancer compared to postmenopausal individuals. Women

experiencing menopause before the age of 50 face a substantially higher risk of breast cancer, whereas those entering menopause at or after 50 have a comparatively lower risk.

4.1.3 Lifestyle factors and their impact on breast cancer

Lifestyle factors such as behavior, habits, and food choices that individuals engage in their daily lives can influence their risk of developing breast cancer. These factors are not inherited or determined by genetics but can be modified or changed to some extent. Several lifestyle factors have been identified as contributors to the development of breast cancer (Chlebowski *et al*., 2020). **Table 3** depicts the lifestyle factors associated with breast cancer.

Consuming a nutritious diet contributes to breast cancer prevention. Diets high in fruits and vegetables, lean proteins, whole grains, and low in saturated fat have been related to a lower risk of breast cancer. On the other hand, a diet high in saturated fats, processed foods, and sugary beverages may increase the risk (Boraka *et al*., 2022). According to the cohort data, 68.28% (95% CI - 0.64 to 0.72) of the surveyed population reported they consumed fresh fruits and vegetables daily, 29.40% consumed fresh fruits and vegetables once a week (95% CI - 0.25 to 0.33) and only 2.32% rarely consumed fresh fruits and vegetables (95% CI - 0.01 to 0.04).

Consuming vegetables is recommended due to their antioxidant properties. An inverse relationship has been observed between vegetable consumption and the likelihood of breast cancer. For instance, a case-control study conducted in Iran revealed that as vegetable intake decreased from daily to weekly or monthly, the odds of breast cancer increased (Marzbani *et al*., 2019). Similarly, the South African breast cancer study found a significant protective association between higher fruit consumption and decreased breast cancer risk among ER-negative and premenopausal women (Jacobs *et al*., 2019). The Women's Health Initiative (WHI) Dietary Modification (DM) randomized clinical trial conducted a long-term follow-up study. It implemented a low-fat dietary pattern that increased the consumption of vegetables, fruits, and grains. Applying this dietary change, along with moderate weight loss, was linked to a notable decrease in the risk of death from breast cancer among postmenopausal women. This finding is based on intention-to-treat comparisons between randomization groups

and involving all 48,835 participants, providing evidence that dietary modifications can have a positive impact on the risk of death from breast cancer among postmenopausal women (Chlebowski *et al*., 2020).

Numerous studies indicate that more intakes of red meat and processed meats (such as bacon, sausage, and deli meats) might correlate with a high risk of breast cancer. In our survey, 83.37% of the patients reported they consumed meat (95% CI - 0.80 to 0.87). 16.63% of the surveyed population reported did not consumed meat (95% CI - 0.13 to 0.20). In a comprehensive cohort study involving 50,884 women; it was observed that there was an association between red meat consumption and an increased risk of invasive breast cancer. Conversely, the consumption of poultry was associated with a decreased risk of developing breast cancer (Lo *et al*., 2020).

The observed increase in breast cancer risk with red meat consumption can be attributed to the presence of pro-carcinogenic elements in red and processed meat. These components include polycyclic aromatic hydrocarbons (PAHs), nitrites (used as additives), heterocyclic aromatic amines (HAA), and N-nitroso compounds (NOCs), which are formed during meat processing and high-temperature cooking. These substances have been associated with direct DNA damage and the development of mammary tumors in humans. The NutriNet-Santé cohort study revealed that the consumption of red meat was associated with an elevated risk of breast cancer (Diallo *et al*., 2018). Another prospective investigation involving 35,372 women demonstrated a higher risk of breast cancer with the consumption of processed meat and total meat (Dunneram *et al*., 2019). Notably, many participants in our hospital cohort study reported regular meat consumption. These findings underscores the importance of dietary choices, particularly the amount of meat consumed and its relation to breast cancer risk.

The consumption of alcohol and the habit of smoking is considered one of the significant lifestyle factors that can impact the risk of developing breast cancer. The survey results showed none consume alcohol nor had smoking history. This indicated that these factors did not play a significant role in our cohort study.

Table 3: Lifestyle factors associated with breast cancer

Factors	Number (N)	Percentage (%)	χ^2	p-value	95% CI	
					Lower	Upper
Consumption of fresh fruits and vegetables						
Daily	353	68.28	96.10	0.535	0.64	0.72
Weekly once	152	29.40			0.25	0.33
Rarely	12	2.32			0.01	0.04
Consumption of meat						
Yes	431	83.37	44.04	0.674	0.80	0.87
No	86	16.63			0.13	0.20
Consumption of alcohol						
Yes	0	0	-	-	-	-
No	517	100				
Smoking						
Yes	0	0	-	-	-	-
No	517	100				
Exercise						
2-3 times a week	16	3.09	75.93	**	0.02	0.04
4-6 times a week	7	1.35			0.003	0.023
Everyday	123	23.79			0.20	0.27
Less than once a week	70	13.54			0.11	0.16
Once a week	22	4.26			0.03	0.06
Rarely/Never	279	53.97			0.50	0.58
Water intake						
1 litre	85	16.44	122.72	*	0.13	0.20
2 litres	129	24.95			0.21	0.29
3 litres	303	58.61			0.54	0.63
Duration of sleep						
5 hours	56	10.83	193.52	**	0.08	0.14
6 hours	154	29.79			0.26	0.34
7 hours	159	30.75			0.27	0.35
8 hours	148	28.63			0.25	0.34
Exposure						
Radiation (e.g., in a factory, laboratory, or medical setting)	0	0	42.88	0.718	-	-
Plastics	0	0			-	-
Agriculture (Pesticides)	101	19.54			0.16	0.23
Control/Mosquito Repellent Chemicals/Dyes	0	0			-	-
Any other exposure (Construction work-cement)	34	6.58			0.05	0.09
No	382	73.88			0.70	0.78

***denotes p < 0.05, **denotes p < 0.01**

A comprehensive analysis of 1101 patients revealed that alcohol usage did not independently contribute to the risk of breast cancer among women in a retrospective study (Liaw *et al*., 2020). Both in the short-term and the longer-term survivorship period, a relatively small percentage of breast cancer survivors (BCSs) (14% to 16%, respectively) did not use alcohol. Adherence to non-smoking and no alcohol consumption slightly increased over the survivorship period. Continued smoking after BC diagnosis is linked to a lower long-term survival rate in comparison to women who give up smoking (Tollosa *et al*., 2020).

Smoking and alcohol consumption have been observed to potentially amplify the metastatic potential of breast cancer cells and stimulate tumor growth. These habits also linked with other existing health conditions that influence survival and could potentially heighten resistance to cancer treatments. In a population-based cohort study comprising 1,926 Black or African American breast cancer survivors, no notable association was found between alcohol consumption, smoking, and breast cancer-specific mortality (Zeinomar *et al*., 2023). The findings of our hospital-based cohort study indicate that alcohol consumption and smoking did not emerge as risk factors in relation to breast cancer-specific mortality.

Physical activity is one of the critical factors and has been associated with a decreased risk of breast cancer. Physical activity exerts its influence through various mechanisms, including alterations in the serum levels of sex hormones, insulin, adipokines, and growth factors. It also impacts molecular processes related to inflammation and the regulation of oxidative stress. The favorable impact of physical activity on cancer development involves a reconfiguration of the tumor microenvironment, influencing metabolism, angiogenesis, aerobic capacity, and the immune response (Orlandella *et al*., 2021).

In our cohort, about 3.09% responded that they exercised 2-3 times a week (95% CI - 0.02 to 0.04). 1.35% reported exercising 4-6 times a week (95% CI - 0.003 to 0.023), whereas 23.79% of the respondents exercised daily (95% CI - 0.20 to 0.27). While 13.54% reported exercising less than once a week (95% CI - 0.11 to 0.16). Around 4.26% of the surveyed population reported exercising once a week (95% CI - 0.03 to 0.06).

Most of the studied population (53.97%) reported exercising rarely or never (95% CI - 0.50 to 0.58). It shows that a significant majority of the population do not engage in regular exercise. The chi-square (χ^2) value of 75.93 and a p-value of 0.008 indicates a statistically significant association between exercise frequency and the risk of developing breast cancer.

In a substantial prospective study involving 47,456 premenopausal and 126,704 postmenopausal women, self-reported levels of physical activity were linked to an approximately 23% decrease in breast cancer risk among premenopausal women and a 17% reduction for postmenopausal women. These findings underscore the protective nature of physical activity against breast cancer risk for women across different life stages (Guo *et al*., 2020). Analyzing a prospective cohort with a median follow-up duration of 23.2 years, it was revealed that substantial physical activity was linked to a 23% lower risk of breast cancer development (Boraka *et al*., 2022).

Through Mendelian randomization analysis, higher genetically predicted levels of accelerometer-measured physical activity was correlated with decreased breast cancer risks. These outcomes suggested that enhancing physical activity could reduce cancer incidence, thus supporting regular physical activity for cancer prevention (Papadimitriou, 2020). Our cohort study also strengthened the above findings, showing a significant statistical association between the lack of exercise and breast cancer risk among most participants and supports the previous findings. The survey findings revealed a statistically significant association between water intake and breast cancer. A χ^2 value of 122.72 and a p-value of 0.046 indicates a significant relationship between water intake and breast cancer.

The influence of water consumption on mitigating breast cancer risk can be attributed to three underlying mechanisms. Firstly, it facilitates a swifter gastrointestinal transit, reducing constipation and rapidly expelling carcinogens from the digestive tract. Secondly, proper cell hydration plays a pivotal role in the carcinogenesis process, impacting cellular division, the expression of oncogenes, gene activity that stimulates cell differentiation, and the prevention of apoptosis. Thirdly, frequent bowel motility, which accelerates the movement of gut contents, has been associated with heightened estrogen

excretion in feces and diminished serum estrogen levels. A retrospective case-control study conducted at the Shaare Zedek Medical Center revealed a potential link between low water intake and an increased risk of breast cancer, particularly among younger individuals (Keren *et al*., 2020). A national prospective study established that a higher intake of total water was correlated with a decreased risk of cancer. Additionally, water supplementation demonstrated promise as a safe and effective approach to reduce fasting plasma glucose levels and lower the risk of diabetes (Zhou *et al*., 2022). Hence, the present study also suggests that more water intake helps to reduce the risk of developing cancer.

Insufficient sleep duration triggers pro-inflammatory and immune-modulatory effects, potentially contributing to the emergence of more aggressive forms of breast tumors. Intricately involved in the circadian rhythm it engages proteins linked to the biological clock that plays essential roles in DNA damage checkpoints. Disruptions in DNA repair can lead to increased cancer progression, genetic instability, and aberrations in chromosome structure (D'cunha *et al*., 2023). The study findings indicated a chi-square (χ^2) value of 193.52 and a p-value of 0.006, providing evidence of a significant association between disturbed sleep duration and the risk of breast cancer.

In a study encompassing a median follow-up of 19.2 years involving 34,350 women aged 40 years and above, a connection was established between short sleep duration and a high risk of breast cancer in Japanese women. This positive correlation was particularly evident among postmenopausal women and those who had many children (Cao *et al*., 2019). Furthermore, among non-Hispanic white women, a notable link between sleep duration and breast cancer stage was detected. Specifically, women with shorter sleep durations exhibited a heightened likelihood of having regional or distant tumors (Soucise *et al*., 2017). In a recent prospective cohort study comprising 10,802 Mexican Americans, sleep duration of less than 6 hours per night was notably linked to an elevated risk of breast cancer (Shen *et al*., 2019). In our study, we similarly identified a significant correlation between disturbed sleep patterns and the risk of breast cancer and this supports many reported studies.

Moving on to environmental and occupational factors have played a pivotal role in elevating the incidence of breast cancer. Among these factors, pesticides pose a significant contamination threat due to their lipid solubility and resistance to biodegradation, leading to widespread environmental pollution and detrimental impacts on human health, particularly in breast cancer induction. The presence of harmful chemicals within pesticides can disrupt hormonal balance, trigger DNA damage, induce tissue inflammation, and silence critical genes, thereby contributing to the development of cancer (Sasikala *et al*., 2023). In the surveyed population, radiation exposure (e.g., in a factory, laboratory, or medical setting) showed no observed cases of breast cancer, indicating a 0% association. Exposure to plastics also showed no observed instances of breast cancer, suggesting a nil association. Agriculture-related exposures, such as pesticides, showed an association with breast cancer. Exposure to mosquito repellent chemicals and dyes showed no association with observed breast cancer cases. We also observed that women construction workers with cement exposure had an incidence percentage of 6.58% (95% CI: 0.05 - 0.09) associated with breast cancer.

Numerous investigations have shown the adverse effects of pesticide usage on the well-being of farmers. A study focusing on the correlation between pesticide-induced genotoxicity and the risk of breast cancer in South Indian women has concluded that individuals engaged in farming and unskilled labor face an elevated risk of breast cancer in comparison to other skilled workers, predominantly due to their frequent exposure to pesticides (Sasikala *et al*., 2023). Similarly, a study conducted in Brazil indicated a higher estimated risk of breast cancer in women exposed to pesticides for a period of 10 or more years (deRezende *et al*., 2023).

A retrospective case-cohort study utilized two extensive national databases: Taiwan's Ministry of Labor's EEW (Especially Exposed Workers) and Taiwan's Cancer Registry. Significantly, an increase in breast cancer risk was noted among female workers exposed to lead, 1,1,2,2-tetrachloroethane, benzene, trichloroethylene/tetrachloroethylene, and asbestos (Chuang *et al*., 2022). Furthermore, 1,1,2,2-tetrachloroethane, recognized as a persistent organic pollutant in the environment, has been linked to promoting obesity, disrupting the epigenomic landscape, instigating enzymatic generation of genotoxic

intermediates leading to elevated reactive oxygen levels, and consequently heightening the risk of breast cancer. Substantive and positive associations were also noted between breast cancer risk, polychlorinated biphenyls (PCBs), and perfluoroalkyl acids (PFAAs). These associations underscore the potential impact of environmental exposure to persistent organic pollutants (POPs) as a contributing factor to increased breast cancer risk among women (Wielsøe *et al*., 2017). The results show that exposure to pesticides and cement may be one of the prominent factors causing breast cancer in the specific cohort.

The findings of the lifestyle factors observed statistically significant relationships between frequencies of exercise, water intake, and sleep duration emphasize the importance of lifestyle modifications in mitigating breast cancer risk, and also occupational factors such as pesticide and cement exposure are causative factors for developing breast cancer. This finding suggests the adoption of regular exercise, proper hydration, and adequate sleep as potential strategies and adopting safety measures to avoid occupational exposure to reduce the likelihood of developing breast cancer.

4.1.4 Clinical manifestations of breast cancer patients

Breast cancer patients can exhibit a range of clinical characteristics that provide valuable information for diagnosis, treatment planning, and prognosis. Some common clinical factors associated with breast cancer patients were studied, including major illness, side of the tumor location, treatments received, side effects, and history of previous surgical operations. By checking the clinical characteristics of patients, potential side effects can be managed, which enhances patient well-being and treatment adherence (Arneja and Brooks, 2021). **Table 4** presents the clinical characteristics of breast cancer patients.

The presence of comorbidities in breast cancer patients was examined. The chi-square (χ^2) value associated with these comorbidities is 78.83, and the corresponding p-value is 0.004. These statistical values suggest a significant association between comorbidities and breast cancer within the study population. The study found that breast cancer patients had a range of comorbidities. Specifically, a substantial proportion of patients had coexisting conditions such as high blood pressure

(14.51%, 95% CI: 0.11 - 0.18), diabetes (11.80%, 95% CI: 0.09 - 0.15), diabetes with high blood pressure (27.47%, 95% CI: 0.24 - 0.31), and diabetes with neurological disorder (12.57%, 95% CI: 0.10 - 0.15). Additionally, few women had asthma (0.19%, 95% CI: 0.001 - 0.005) and thyroid malfunction (2.90%, 95% CI: 0.02 - 0.04). These comorbidities are to be considered in breast cancer management, as they can influence treatment decisions, potential interactions with medications, and overall patient care.

In a cross-sectional examination involving 201 participants, a notable 47% of women reported a history of hypertension, while 13% had benign breast disease, and 12% had diabetes (Moodley *et al*., 2018). In a comprehensive Canadian population-based study encompassing 13,208 participants, an elevated susceptibility to heart failure, ischemic heart disease, diabetes, osteoporosis, depression, and hypothyroidism was observed among women with breast cancer (Ng *et al*., 2019).

Among breast cancer survivors, a web of closely linked conditions, including obesity, dyslipidemia, impaired glucose intolerance, and elevated blood pressure was observed. These comorbidities potentially mediate breast carcinogenesis by activating proto-oncogenes and antiapoptotic transcription factors through pathways linked to estrogen receptors (ER-related pathways) (Arneja and Brooks, 2021). Our study outcome aligned with the conclusions drawn from other documented research and hold statistical significant outcomes with comorbidities and breast cancer.

The study also analyzed the location of the breast tumor in patients. 3.48% (95% CI: 0.02 - 0.05) of cases had tumors on both sides of the breast, 55.13% (95% CI: 0.51 - 0.59) had tumors on the right side, and 41.39% (95% CI: 0.37 - 0.46) had tumors on the left side. These findings reveal a significant association between the location of the tumor and breast cancer. The chi-square (χ^2) value of 184.45 and the p-value <0.001 indicate a highly significant relationship. The results suggest that the side of the breast tumor is a vital factor in breast cancer management, as it plays a role in treatment decisions and patient outcomes but no scientific explanations could clarify the basis of this prevalence.

Table 4: Clinical manifestations of breast cancer patients

Factors	Number (N)	Percentage (%)	χ^2	p-value	95% CI	
					Lower	Upper
Comorbidities						
Blood pressure	75	14.51	78.83	**	0.11	0.18
Diabetes	61	11.80			0.09	0.15
Diabetes and high blood pressure	142	27.47			0.24	0.31
Diabetes and neurological disorder	65	12.57			0.10	0.15
Asthma	1	0.19			0.001	0.005
Thyroid	15	2.90			0.02	0.04
None	158	30.56			0.27	0.35
Primary tumor site						
Both sides	18	3.48	184.45	***	0.02	0.05
Right side	285	55.13			0.51	0.59
Left side	214	41.39			0.37	0.46
Treatment modalities						
Surgery	144	27.85	219.06	***	0.24	0.32
Chemotherapy	238	46.03			0.42	0.50
Radiotherapy	54	10.44			0.08	0.13
Others (Surgery, chemotherapy, radiotherapy)	81	15.68			0.13	0.19
Side effects						
Hair loss	124	23.98	286.47	0.612	0.20	0.28
Nail discoloration	29	5.61			0.04	0.08
Fatigue	35	6.78			0.05	0.09
Joint pain	27	5.22			0.03	0.07
Vomiting	51	9.86			0.07	0.12
Mouth ulcer	28	5.42			0.04	0.07
Multiple side effects	223	43.13			0.39	0.47
Other surgical operation						
Breast mastectomy	118	22.82	299.57	0.399	0.19	0.26
Breast mastectomy and Caesarean section	153	29.59			0.26	0.34
Breast mastectomy and family planning	159	30.75			0.27	0.35
Uterus removal	14	2.71			0.01	0.04
Heart surgery	8	1.55			0.005	0.02
Ovary removal	3	0.58			0	0.01
No	62	12.0			0.03	0.21

****denotes p < 0.01, *** denotes p < 0.001**

In a comprehensive nationwide analysis encompassing 2,423,875 women, the incidence of breast cancer was scrutinized. Interestingly, both left and right breast exhibited nearly equivalent occurrence rates, with a marginal prevalence in the left breast. Among the patients, 50.6% had their primary tumor in the left breast, while 49% presented with cancer in the right breast (Sisti *et al*., 2020). According to Abdou *et al*. (2022), left-sided breast cancer showed high biological aggressiveness compared to its right side. This evaluation was conducted utilizing an extensive patient cohort, and their clinical and pathological features were also analyzed. It was found that left-sided breast tumors possess a more proliferative genomic profile and exhibit lower responses to neoadjuvant chemotherapy.

A study involving 92 patients who underwent surgical treatment for unilateral, primary, invasive breast carcinoma (IBrC) was conducted. 75% of the 20 cases with confirmed nodal metastases had breast cancer on the right side, while 25% had it on the left. Furthermore, this study found that patients with right-sided breast cancer usually had larger primary tumor sizes. The tumor diameter emerged as a predictive factor for disease relapse (Barbara *et al*., 2020). In our study cohort, the right breast was more frequently affected in patients compared to the left side.

Treatment modalities provide valuable insights for optimizing treatment decisions by considering each patient's needs and characteristics. Approximately 27.85% (95% CI: 0.24 - 0.32) of breast cancer patients first underwent breast mastectomy. Around 46.03% (95% CI: 0.42 - 0.50) of patients received chemotherapy as the first part of their breast cancer treatment. 10.44% (95% CI: 0.08 - 0.13) of patients underwent radiotherapy, and 15.68% (95% CI: 0.13 - 0.19) received a combination of surgery, chemotherapy, and radiotherapy. The chi-square (χ^2) value associated with these treatment modalities is 219.06, and the corresponding p-value is <0.001. These statistical values indicate a highly significant relationship between the treatment modalities and breast cancer patients in the study population.

The study also examined the occurrence of side effects in breast cancer patients while receiving cancer treatment. Among the observed side effects, hair loss was the most prevalent, affecting approximately 23.98% (95% CI: 0.20 - 0.28) of patients.

Nail discoloration was 5.61% (95% CI: 0.04 - 0.08), and 6.78% of patients experienced fatigue (95% CI: 0.05 - 0.09). 5.22% had joint pain (95% CI: 0.03 - 0.07), vomiting was reported by 9.86% (95% CI: 0.07 - 0.12) of patients and 5.42% reported mouth ulcer (95% CI: 0.04 - 0.07). Notably, a significant proportion of patients, 43.13% (95% CI: 0.39 - 0.47), experienced multiple side effects concurrently. The results suggest that side effects may vary according to the patient's unique profile.

Chemotherapy is one of the major therapeutic approaches in treating cancer proliferation within the body. The findings of a descriptive cross-sectional study underscore the detrimental consequences of chemotherapy-induced toxicities, including symptoms like nausea, peripheral neuropathy, dysgeusia, myalgia, loss of appetite, and peripheral edema. These toxicities significantly impede the quality of life for breast cancer patients (Prieto-Callejero *et al*., 2020). A retrospective study across tertiary care hospitals examined 353 prescriptions for female patients. Approximately 97.5% of breast cancer patients undergoing chemotherapy experienced adverse effects. Notable adverse effects included alopecia, darkening of nails and fingertips, vascular changes, anemia, neutropenia, left ventricular diastolic dysfunction, and increased serum creatinine levels (Kodati *et al*., 2019). Our study also found chemotherapy as the preferred treatment modality in the patient cohort. Furthermore, the adverse effects observed in our study closely aligned with those reported in other cohort based studies.

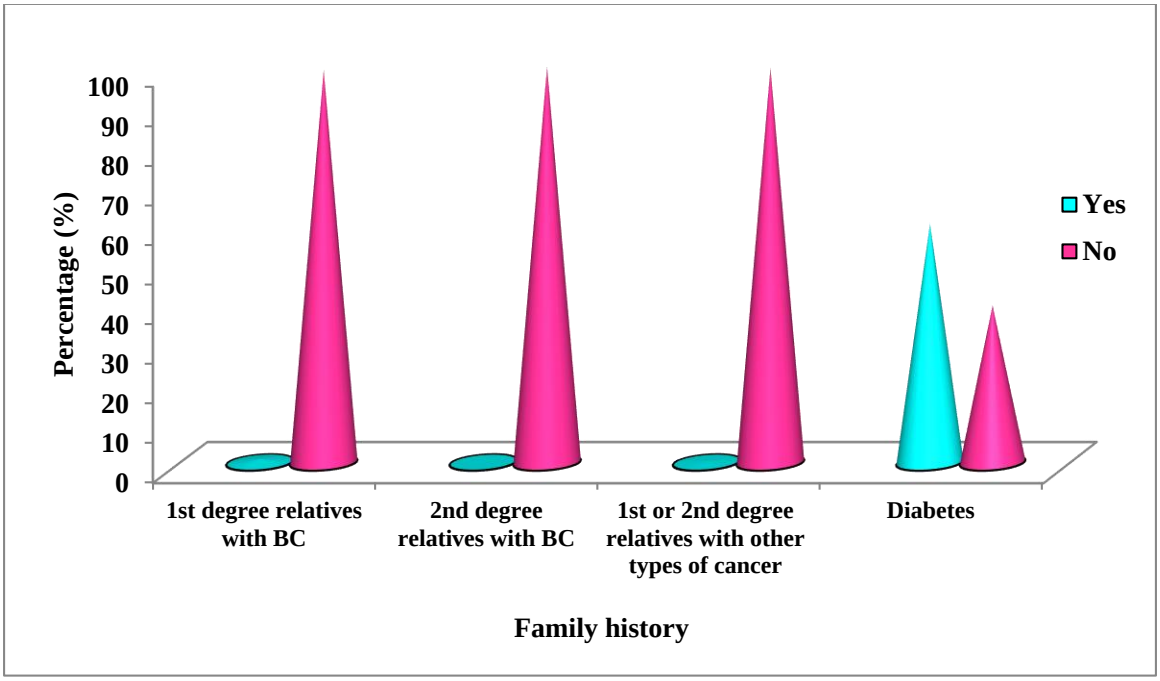
In the survey, we also analysed breast cancer patients who underwent any other surgeries. Approximately 22.82% (95% CI: 0.19 - 0.26) of breast cancer patients underwent a breast mastectomy. A proportion of 29.59% (95% CI: 0.26 - 0.34) of patients had both a breast mastectomy and a Caesarean section. 30.75% (95% CI: 0.27 - 0.35) of patients had a history of breast mastectomy and family planning. 2.71% (95% CI: 0.01 - 0.04) of breast cancer patients had a hysterectomy done. A small proportion 1.55% (95% CI: 0.005 - 0.02) of patients had a history of heart surgery. 0.58% (95% CI: 0 - 0.012) of patients had surgical removal of the ovaries. Approximately 12% (95% CI: 0.03 - 0.21) of breast cancer patients did not have a history of any surgical procedures.

The clinical manifestations of breast cancer patients provided valuable insights into the disease's complexity and management. The research unveiled significant associations between comorbidities and breast cancer, highlighting the importance of considering coexisting conditions in treatment decisions and patient care. The distribution of primary breast tumor location had a highly significant relationship between tumor side and breast cancer incidence. A statistical relationship was observed between treatment modalities and breast cancer patients, emphasizing the need for tailored treatment strategies for individual patient characteristics.

4.1.5 Correlation between family history and characteristics of breast cancer

Assessing the family history of breast cancer patients is crucial in determining the potential genetic component and overall risk assessment. Individuals with a positive family history may be recommended for genetic testing to identify any inherited genetic mutations associated with increased breast cancer risk (Abudawood, 2019). **Figure 11** shows the family history of breast cancer patients.

Figure 11: Family history of breast cancer patients



The survey data provides insights into the prevalence of breast cancer among first-degree relatives of individuals diagnosed with breast cancer. Only few (0.58%, 95%

CI: 0.001 - 0.013) reported where mother too was affected. 0.39% (95% CI: 0.002 - 0.008) of siblings with a history of breast cancer. The majority, 99.03% (95% CI: 0.98 - 0.99), of breast cancer patients did not have any first-degree relatives who were diagnosed with breast cancer. The study found that among breast cancer patients, a small proportion (0.39%, 95% CI: 0.001 - 0.009) had first or second-degree relatives with brain cancer. The majority of breast cancer patients (99.61%, 95% CI: 0.98 - 0.99) did not have any family history of cancer.

Among the study population, approximately 60.35% (95% CI: 0.56 - 0.64) of individuals had a family history of diabetes. Family history serves as a notable risk factor for breast cancer. Data analysis of 10,549 breast cancer patients found that patients without a family history of cancer are diagnosed at more advanced disease stages than those with a family history of cancer (Liu *et al*., 2021). In a cohort of 1101 women diagnosed with breast cancer and confirmed by histopathological examination, 14.4% had at least one family member with breast cancer. Interestingly, a positive family history incidence did not significantly differ across different ethnic groups (Malays, Chinese, and Indians) (Liaw *et al*., 2020). The primary driver of hereditary breast cancer development lies in germline mutations affecting breast cancer susceptibility genes, such as *BRCA1*, *BRCA2, TP53, PTEN, ATM, CHEK2,* and *PPM1D*. These genes contribute to critical DNA repair mechanisms (Mahdavi *et al*., 2019).

While our cohort study did not establish a substantial link between family history and breast cancer, the most prevalent inheritable condition was found to be diabetes. Metabolic changes related to diabetes, notably hyperglycemia and hyperinsulinemia, highlight the significance of glycemic management in diabetic patients with cancer. Poorly controlled blood sugar levels contribute to poorer patient outcomes and reduced overall survival. Hyperglycemia and hyperinsulinemia also impact the tumor microenvironment, affecting temporal tumor heterogeneity. The presence of insulin resistance, hyperinsulinemia, and alterations in the signaling of growth and steroid hormones linked to diabetes can potentially impact the risk of breast neoplasms (Abudawood, 2019).

The extensive examination of various risk factors in this study has yielded crucial insights into the epidemiology of breast cancer within the specific cohort. The demographic profile of breast cancer patients displayed distinctive directions, with the highest occurrence noted among individuals aged 41 to 50 years, underscoring the relevance of age as a potential risk factor. Additionally, education level emerged as a significant determinant linked to breast cancer incidence, emphasizing the impact of education on susceptibility to the disease. This underscores the need for health educational interventions and awareness campaigns targeted at specific demographic groups to enhance early detection and preventive measures. Reproductive factors, particularly menopausal status, exhibited a substantial influence on breast cancer risk. Premenopausal women faced a higher risk compared to postmenopausal individuals, indicating the relevance of hormonal factors in disease development. These findings emphasize the importance of considering reproductive history in assessing breast cancer risk and tailoring preventive strategies accordingly.

Lifestyle factors, including exercise frequency, water intake, and sleep duration, have exhibited significant associations with breast cancer incidence, emphasizing the pivotal role of daily habits in shaping susceptibility to the disease. Meanwhile, occupational factors such as pesticide and cement exposure suggested a potential link to increased risk. These results underscore the impact of daily habits and behaviors on disease risk, providing avenues for lifestyle interventions to reduce susceptibility. Clinical characteristics, such as comorbidities, tumor location, and treatment modalities, revealed intricate associations with breast cancer incidence. The identification of specific comorbidities and their prevalence among breast cancer patients highlights the need for integrated healthcare approaches to manage both cancer and concurrent health conditions. A notable proportion of patients had a family history of diabetes, indicating potential links between genetic factors and certain diseases. The multifaceted analysis of demographic, reproductive, lifestyle, clinical, and familial factors provides a comprehensive understanding of breast cancer epidemiology within the studied population, helping us develop better prevention strategies, early detection, and personalized treatment.

Once we had identified the risk factors, we wanted to know the prevalence of rare variants in representative samples among the selected population. So, we did whole exome sequencing to identify the rare variants. Whole exome sequencing (WES) can provide insights into the genetic components of the epidemiological interactions, helping to elucidate the mechanisms by uncovering specific genetic mutations or variants and the risk factors jointly responsible for disease susceptibility (Staaf *et al*., 2019). Combining these approaches allows researchers and healthcare providers to bridge the gap between population-level trends and personalized genetic insights, ultimately leading to more effective disease prevention, diagnosis, and treatment strategies. Therefore, this study phase was structured to identify genetic variations within specific populations affected by diverse risk factors.

Phase II

4.2 Exome profiling of breast cancer patients

Whole exome sequencing is a high-throughput genomic technique that selectively sequences the exons, the genome's protein-coding regions. These exons encompass a relatively small segment of the complete genome but contain all the genetic variants responsible for causing disease (Ross *et al*., 2020). In this phase of the study, we collected tumor tissue from mastectomy specimens of six primary breast cancer patients. We wanted to identify the genetic alterations before the cancer therapy. So, we selected primary tumor tissue for the study. Primary breast cancer refers to the initial cancer tumor that develops within the breast tissue. Exome profiling has the potential to unveil novel biomarkers associated with breast cancer subtypes or disease progression.

4.2.1 Clinicopathological characteristics of breast cancer patients

The clinicopathology of breast cancer patients involves a comprehensive assessment of both clinical and pathological conditions. The clinicopathological characteristics of six breast cancer patients (BC-1 to BC-6) are given in **Table 5**. The age of the patients was found to be 37 to 74, indicating that breast cancer can affect individuals across different age groups. Four patients had tumors located on the right side of the breast, while two patients had tumors located on the left side. Family history of breast cancer was reported in BC-4 and BC-5, underlining the importance of genetic

factors underlying the cause of breast cancer. All patients were diagnosed with DCIS (Ductal Carcinoma *in situ*), characterized by abnormal glandular epithelial cell growth confined within ducts without breaching the basement membrane. It is converted to invasive ductal carcinoma when the basement membrane is breached (Kalwaniya *et al*., 2023). All six patients had Grade III tumors, suggesting a high level of cell abnormality and aggressiveness.

The TNM (Tumor, Node, Metastasis) staging information for breast cancer patients reveals crucial details about the extent and characteristics of the cancer. The clinical prognostic stage categories are relevant for all individuals diagnosed with primary breast cancer before the initiation of any treatment (Hortobagyi *et al*., 2018). All individuals presented a T2 tumor stage, indicating a tumor size ranging from more than 2 cm to not more than 5 cm. BC-2 patient alone displayed a T1c tumor stage, suggesting a smaller tumor size of 2 cm or less. Patient BC-1 presented regional lymph node involvement (N1) and micro-metastasis (Mi). BC-2 detected no regional lymph node involvement (N0) or distant metastasis (0). Patient BC-3 exhibits advanced regional lymph node involvement (N3a) and distant metastasis (M). BC-4 and BC-5 do not have regional lymph node involvement (N0) or distant metastasis (0). Lastly, BC-6 featured a regional lymph node involvement (N1a) and no distant metastasis (0). This TNM staging data aids in assessing the stage of breast cancer in each patient, informing tailored treatment strategies according to tumor size, lymph node status, and metastatic spread (Amjad *et al*., 2020).

Breast cancer patients can be categorized into different subtypes based on the expression levels of key receptors, including ER, PR, and HER2. BC-1 and BC-6 fall into the Luminal B subtype, characterized by estrogen receptor (ER)-positive, progesterone receptor (PR)-negative, and HER2-positive. Both BC-2 and BC-3 belong to the Triple-Negative subtype, as they exhibit negative receptor status for ER, PR, and HER2. BC-4 is identified as HER2-enriched, with a negative status for ER and PR but positive for HER2. BC-5 falls under the Luminal A subtype, characterized by being ER-positive, PR-positive, and HER2-negative. These subtype distinctions unravel the heterogeneity of breast cancer and play a critical role in determining treatment approaches, considering the specific molecular characteristics of each patient's profile (Johnson *et al.*, 2021).

Table 5: Clinical and pathological features of breast cancer patients

Patients	BC-1	BC-2	BC-3	BC-4	BC-5	BC-6
Age	59	37	53	74	68	46
Gender	Female	Female	Female	Female	Male	Female
Tumor side	Right	Right	Right	Left	Right	Left
Family history	No	No	No	Yes	Yes	No
Histological type	DCIS	DCIS	DCIS	DCIS	DCIS	DCIS
SBR Grade	III	III	III	III	III	III
pT stage	T2	T1c	T2	T2	T2	T2
N stage	N1	N0	N3a	N0	N0	N1a
Metastasis	Mi	0	M	0	0	0
ER status	Positive	Negative	Negative	Negative	Positive	Positive
PR status	Negative	Negative	Negative	Negative	Positive	Positive
HER2 status	Positive	Negative	Negative	Positive	Negative	Positive
Hotspots (Ki-67)	54%	>90%	>90%	90%	65%	62%

One male (BC-5) was affected, highlighting that breast cancer can occur in men. We were interested in checking the genetic alteration in males, so we collected the male tissue also for our study. There were significant similarities in the mutational landscape between male and female breast cancer. Male breast cancer (MBC) cases occur in individuals aged 60 and older, and they often exhibit positive expression of estrogen receptors (ER) and progesterone receptors (PR) (Ben Kridis-Rejeb *et al.*, 2020). Our study reported that a male breast cancer patient was 68 years old and showed a positive status for both ER and PR.

The Ki-67 proliferation index is used in pathology and oncology to assess the rate of cell proliferation in a breast tumor tissue. Patients BC-2 and BC-3 have very high

percentages, over 90%, which means their tumors were growing fast and more aggressive. BC-4 was close behind with 90%, also indicating a high growth rate of cancer cells. On the other hand, BC-1, BC-5, and BC-6 had a lower percentages (54%, 65%, and 62%), suggesting a more moderate level of cell division. The Ki-67 scores indicate the tumors vary in aggressiveness and growth, emphasizing the need to consider these medical details to understand each patient's cancer characteristics. These clinicopathological characteristics provide important information for treatment planning and prognosis, highlighting the heterogeneity of breast cancer and the need for personalized approaches to patient care (Hashmi *et al*., 2019).

4.2.2 Quantification of isolated DNA

DNA is a repository for genetic information and provides insights into the fundamental genetic changes occurring in coding regions, introns, and regulatory elements. These genetic changes help identify cellular function and understand the progression of diseases (Salk *et al*., 2018).

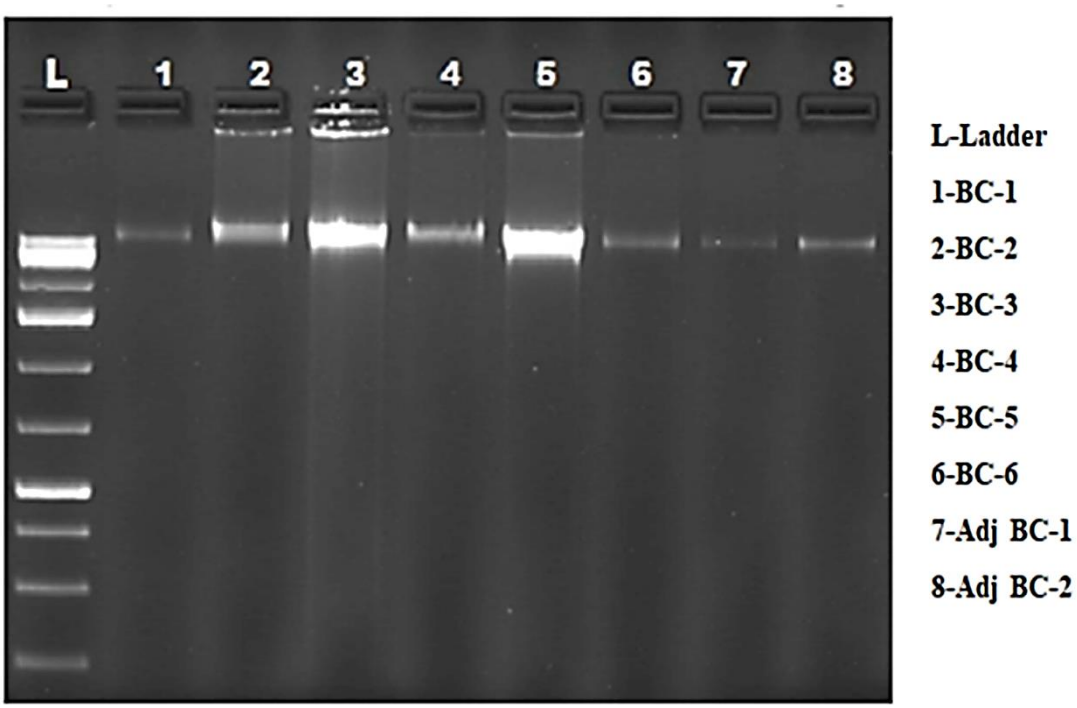
Table 6: Purity and quantity of isolated tumor DNA

Patients	DNA purity (A260/A280)	Concentration (ng/µl)
BC-1	1.83	667.03
BC-2	1.84	94.43
BC-3	1.81	357.99
BC-4	1.82	413.66
BC-5	1.85	177.41
BC-6	1.83	134.77
Adjacent normal tissue (BC-1)	1.80	412.68
Adjacent normal tissue (BC-2)	1.82	234.25

The tumor DNA from the six patients and two adjacent normal tissues DNA of BC-1 and BC-2 were isolated using the QIAamp DNA mini kit. The concentration and purity of the isolated DNA were evaluated using a nanodrop (NanoDrop™ 1000 Spectrophotometer, Thermofisher). The concentration and purity of the isolated DNA is depicted in **Table 6** showing good DNA purity. Ensuring the purity of DNA is essential for the success of exome analysis to ensure the reliability and precision of sequencing data. It minimizes the occurrence of artifacts, improves the efficiency of target enrichment, and optimizes subsequent processes, thereby enhancing the overall success of exome sequencing studies (Oh *et al*., 2020).

Gel electrophoresis helps to confirm the purity of DNA by evaluating the presence of impurities. Additionally, it plays a crucial role in ensuring that DNA fragments fall within the specified size range essential for exome sequencing (Pansare *et al*., 2019). The DNA bands were intact and neither degraded or fragmented as shown in **Figure 12.** The isolated DNA samples had high quality and purity and were further processed for whole exome sequencing.

Figure 12: Gel pattern of breast tumor DNA



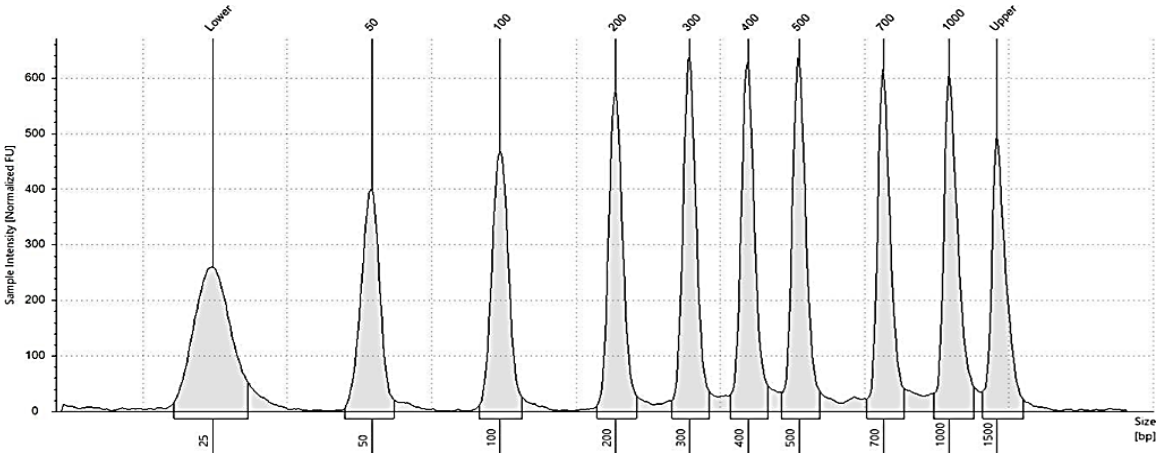
The exome data's fragment size, read quality, coverage, and sequencing metrics were evaluated to ensure data integrity and reliability.

4.2.3 Library quality control

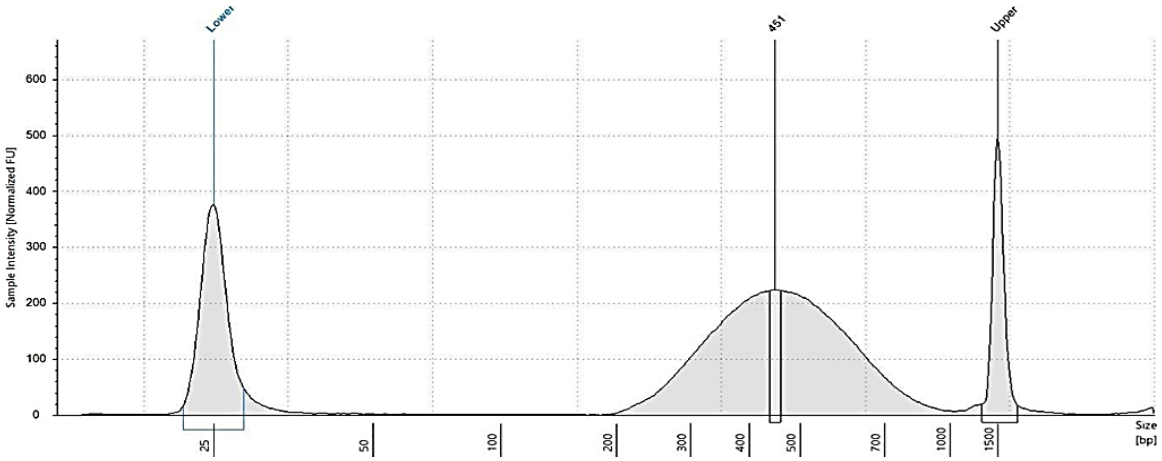
Library preparation is a crucial step in DNA sequencing. DNA fragments were prepared and tagged with specific adapters to enable sequencing during library preparation. Library QC involves assessing the quality and quantity of prepared DNA libraries. Fragment size analysis and quantification are used to ensure the libraries are suitable for sequencing (Hess *et al*., 2020). The insert size analysis of the library was performed using the Tapestation 4150 (Agilent) equipped with high-sensitive D1000 screen tapes, confirmed that the average fragment size for all patient libraries fell within the desired range of 350–450 base pairs. A distinct library peak was observed during the quality control assessment, as illustrated in **Figure 13.** Compared to electronic ladder (automated or electronic system), the library peak facilitated the size distribution of DNA in the patient library.

Figure 13: Library peak

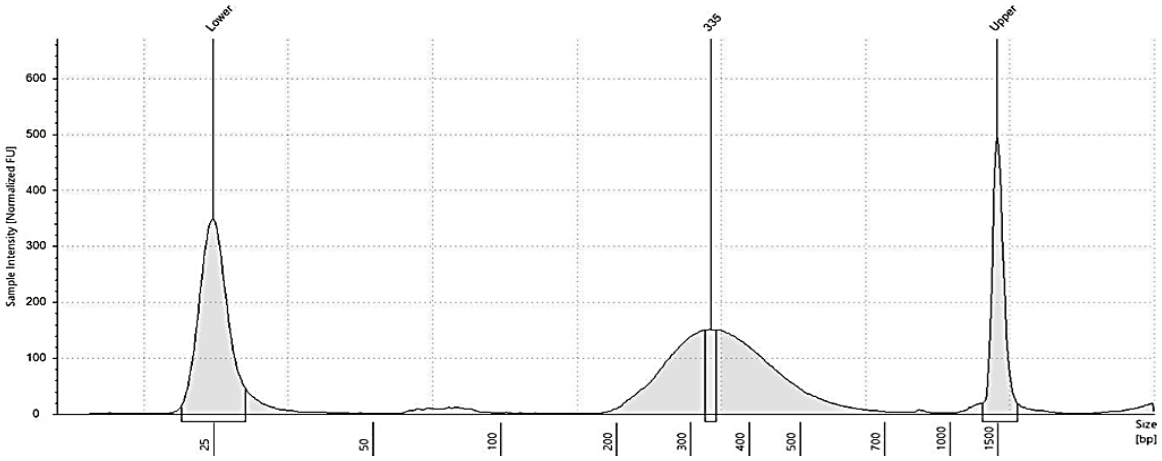
Electronic ladder



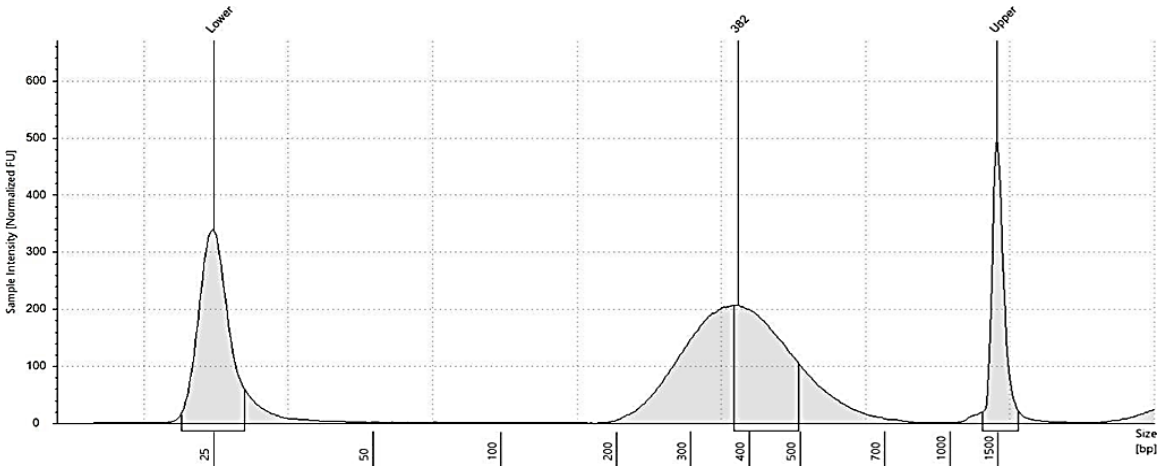
BC-1


Lower
451
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

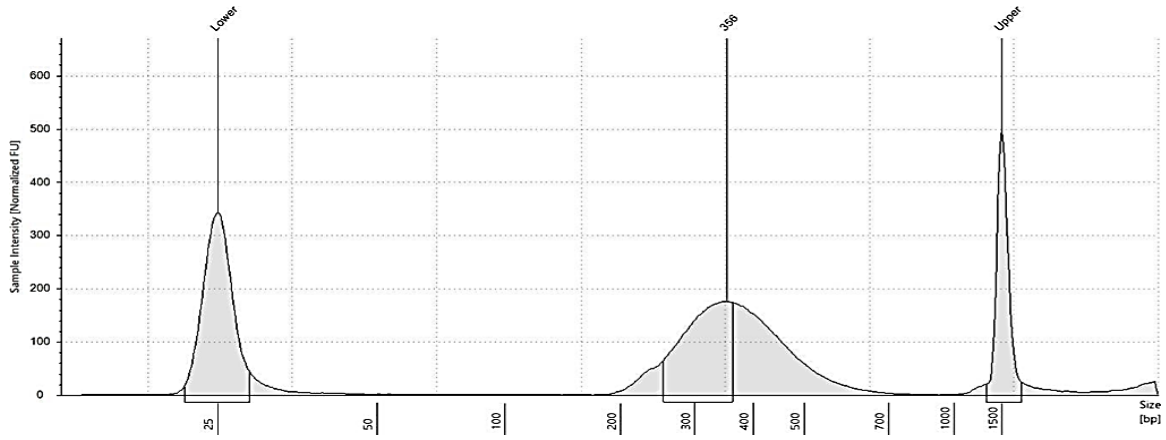
BC-2


Lower
335
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

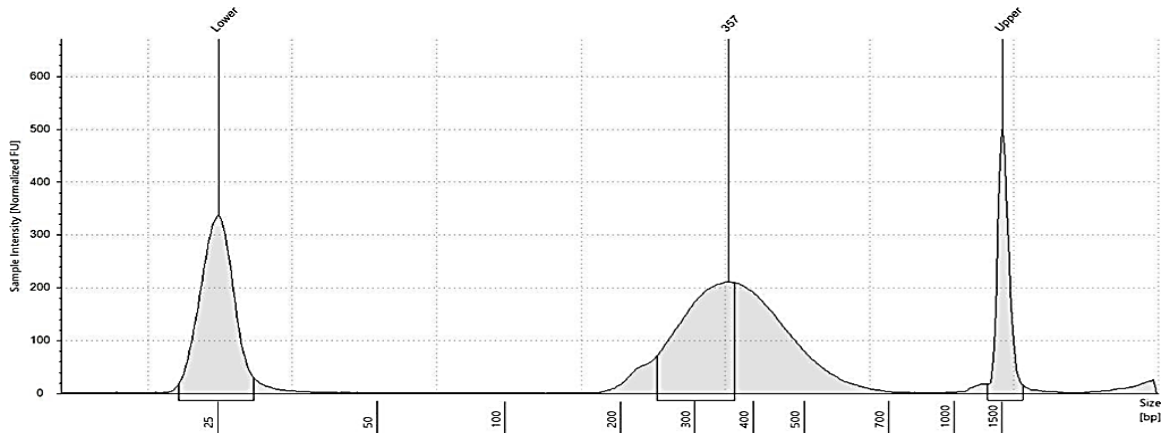
BC-3


Lower
382
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

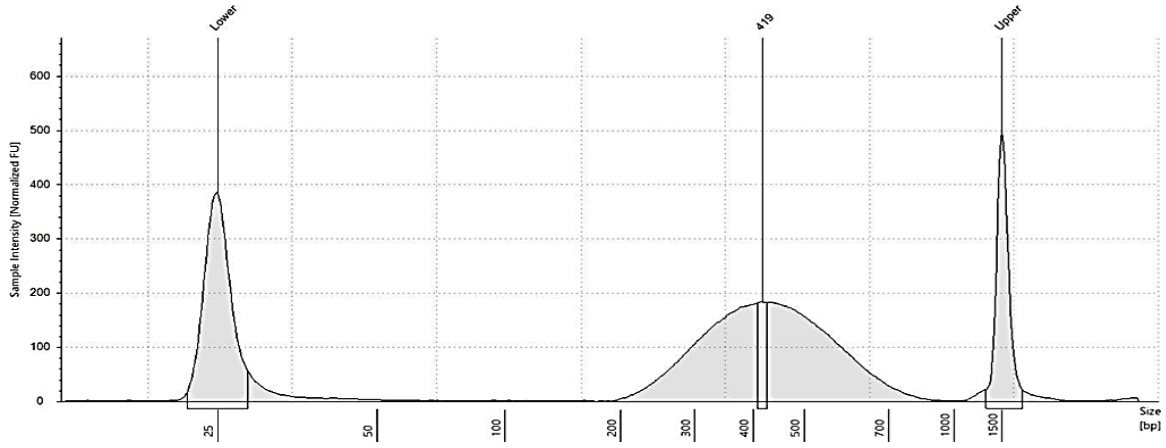
BC-4


Lower
356
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

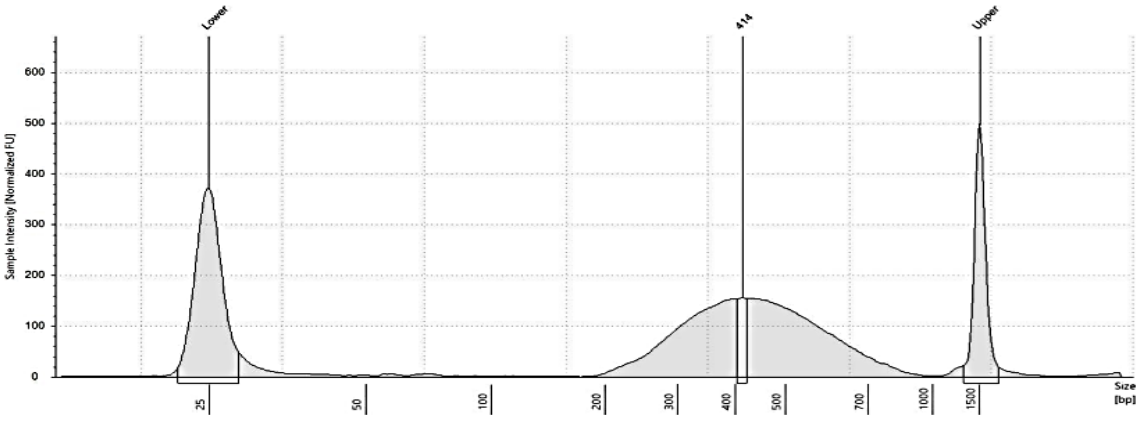
BC-5


Lower
357
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

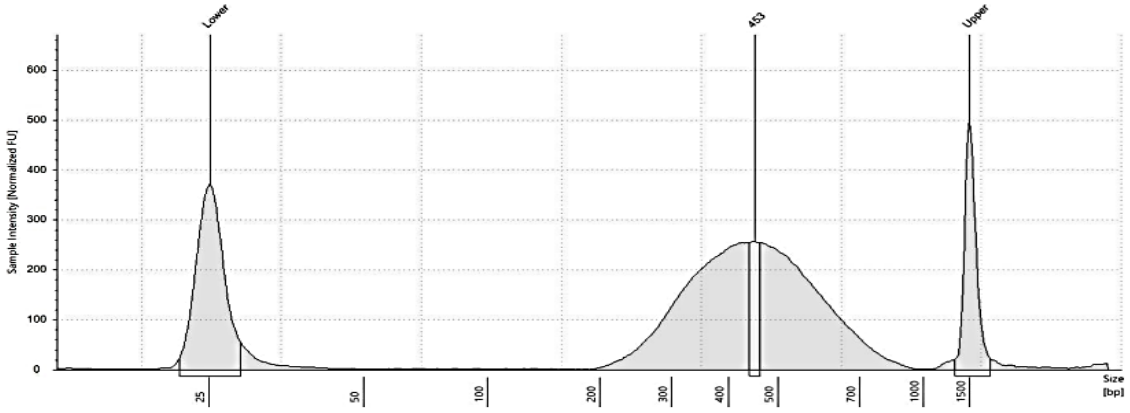
BC-6


Lower
419
Upper
Sample Intensity [Normalized FU]
0
100
200
300
400
500
600
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

Adjacent normal BC-1



Adjacent normal BC-2



Specifically, the insert sizes for patients BC-1 to BC-6 and adjacent normal of BC-1 and BC-2 were determined as follows: 451 bp, 335 bp, 382 bp, 356 bp, 357 bp, 419 bp, 414bp, and 453bp. This quality control step ensured the libraries were prepared with consistent and appropriate insert sizes. Once the DNA libraries passed the library QC, they were loaded onto an Illumina sequencing platform.

4.2.4 Sequencing quality control

Sequencing QC involves verifying the quality of the sequencing run. It includes monitoring parameters like sequencing depth, read quality, and error rates and ensuring correct sequencing functions. Sequencing QC helps to confirm that the actual sequencing process is producing high-quality data. BCL (Base call) files were generated during the sequencing step and contain raw fluorescence intensity data for each sequencing cycle.

These binary files generate high-quality sequence data and were processed further to generate more usable FASTQ files for downstream analysis (Qin, 2019).

Table 7 provides sequencing statistics for six patients (BC-1 to BC-6) and adjacent normal tissue of BC-1 and BC-2 whose exome sequencing was done. The data includes raw and trimmed read counts, raw and trimmed total bases, read length for both Read1 and Read2, and the total data size in giga base pair (Gb) for each patient's sequencing run.

Table 7: Sequencing statistics

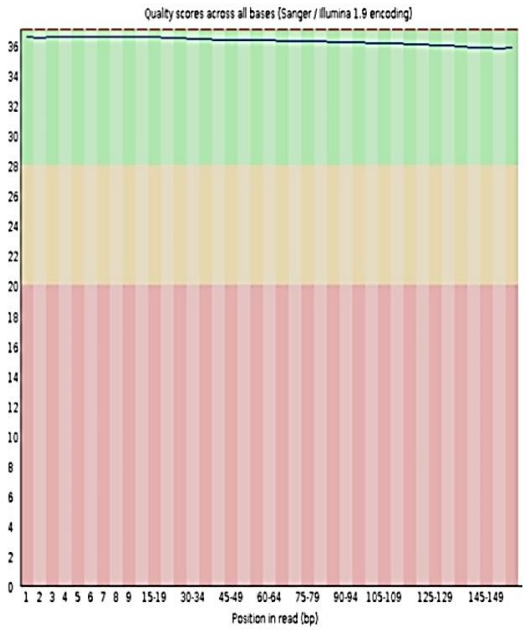
Patients	Raw total reads	Raw total bases	Read1 length	Read2 length	Total data (Gb)	Trimmed total reads	Trimmed total bases	Trimmed read1 length	Trimmed read2 length	Total Data (Gb)
BC-1	62965806	10011563154	159	159	10.011563	59539514	8926351954	149	149	8.926352
BC-2	57046066	9070324494	159	159	9.0703245	55965514	8743537297	156	156	8.7435373
BC-3	102678628	16325901852	159	159	16.325902	100308134	15579256433	155	155	15.579256
BC-4	93864706	14924488254	159	159	14.924488	91502966	14144317016	154	154	14.144317
BC-5	92080028	14640724452	159	159	14.640724	90265720	14109464954	156	156	14.109465
BC-6	87484848	13910090832	159	159	13.910091	85032420	13061808007	153	153	13.061808
Adj BC-1	101803510	16186758090	159	159	16.186758	98318536	14971954956	152	152	14.971955
Adj BC-2	83817132	13326923988	159	159	13.326924	80763388	12313082023	152	152	12.313082

The Phred quality score (Q-score) is a logarithmic scale commonly used in DNA sequencing to represent the accuracy of base calls in sequencing reads. Phred scaling has been extensively adopted for quality metrics in next-generation sequencing bioinformatics, including read mapping and genotype quality assessment. A higher Phred score of above 30 corresponds to a higher confidence level in the accuracy of the base call at a specific position in the read (Larson *et al.*, 2023). The Phred score of 36 was maintained throughout all base pairs in the sequencing read of six breast cancer patients. It corresponds to a base call accuracy of 99.9%, meaning there is a 1 in 1,000 chance of an incorrect base call at that position. It indicates that the sequencing data is of very high quality and is associated with a low error rate. The quality score across all bases for pair-end sequencing is given in **Figure 14**.

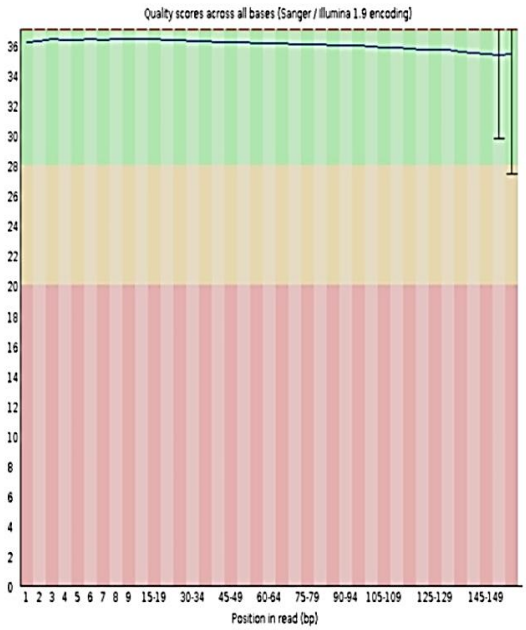
Figure 14: Per base sequence quality

BC-1

Read 1

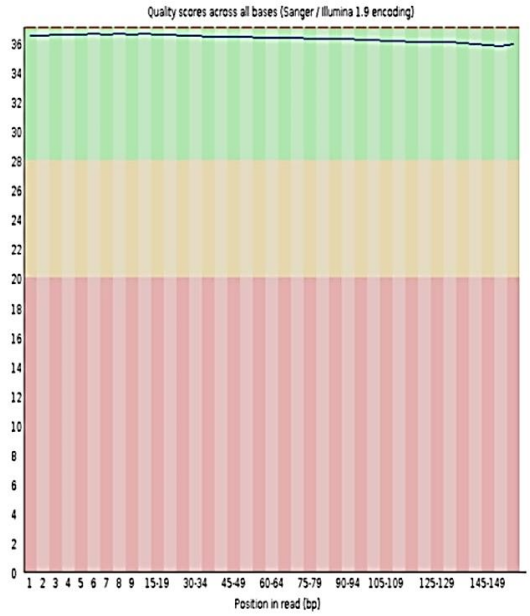


Read 2

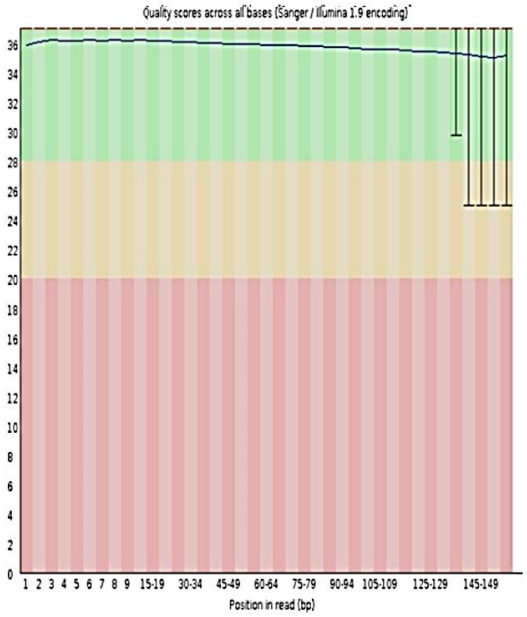


BC-2

Read 1

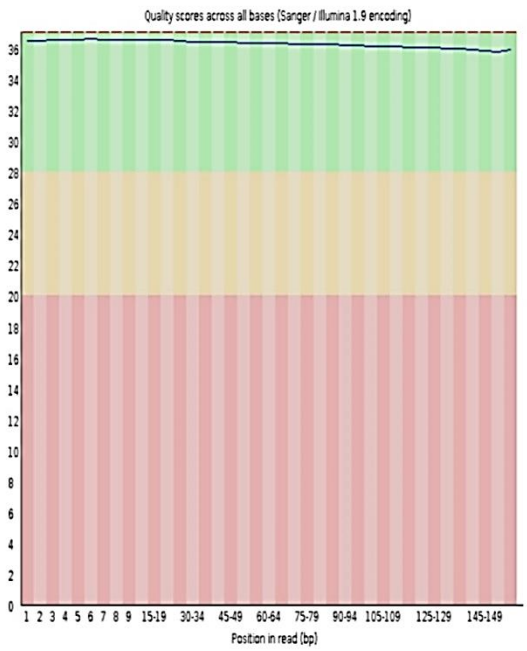


Read 2

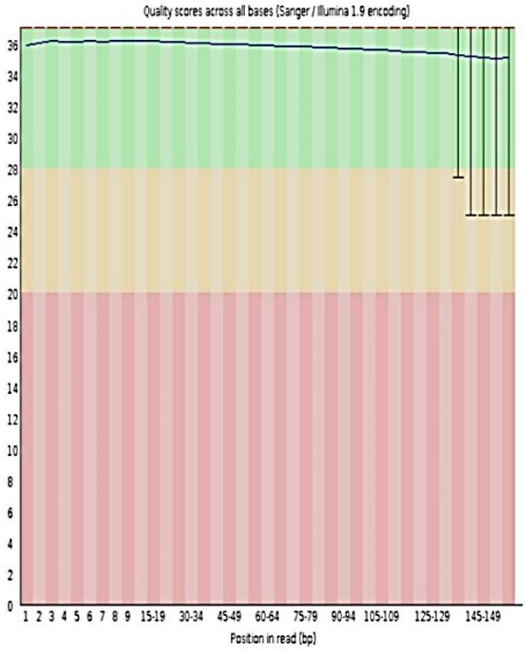


BC-3

Read 1

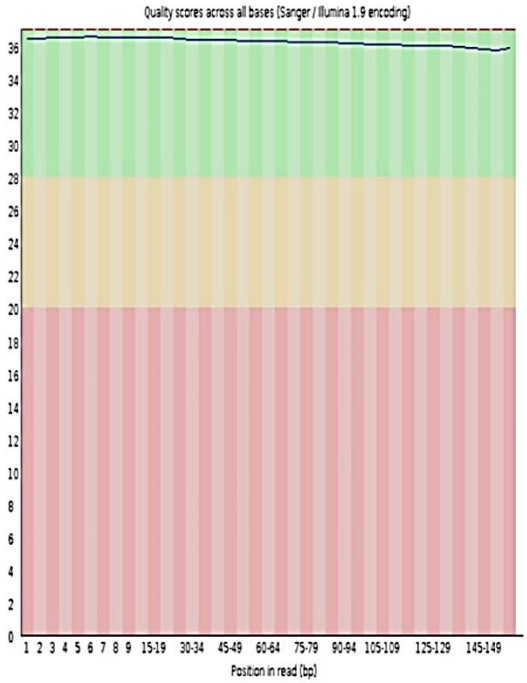


Read 2

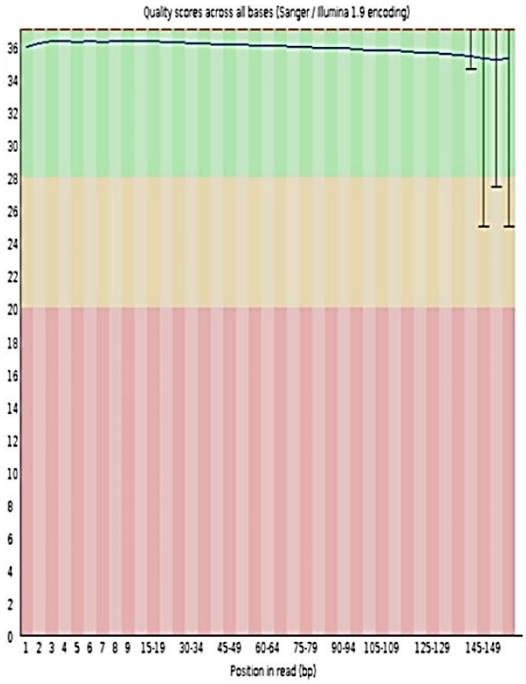


BC-4

Read 1

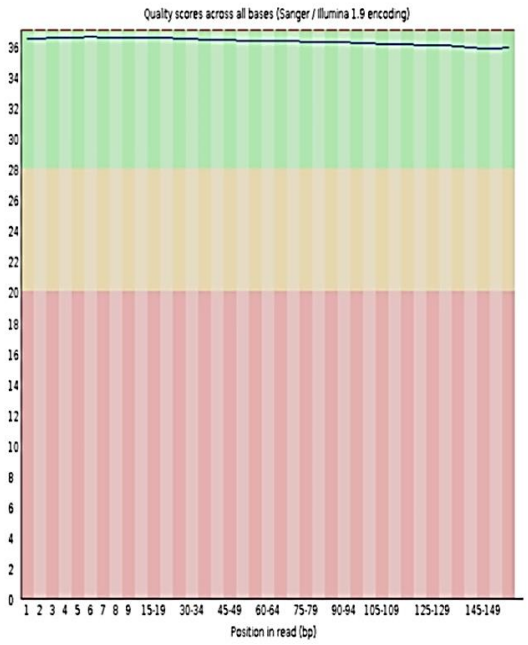


Read 2

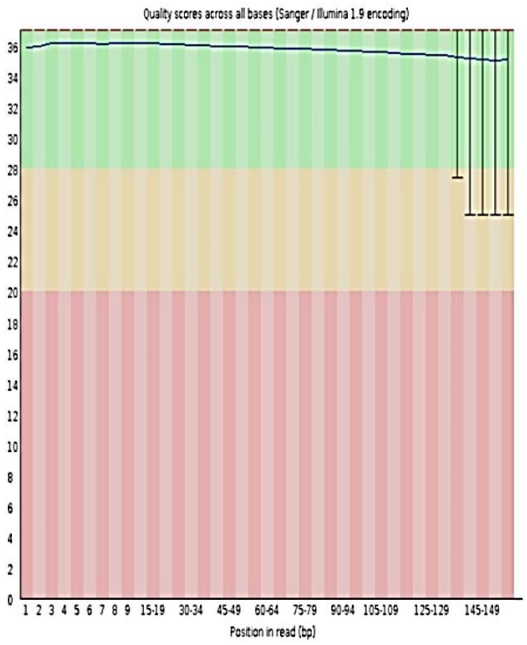


BC-5

Read 1

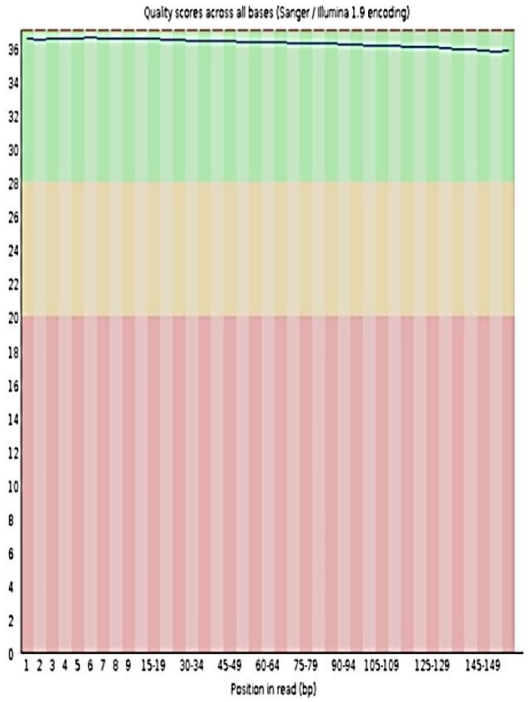


Read 2

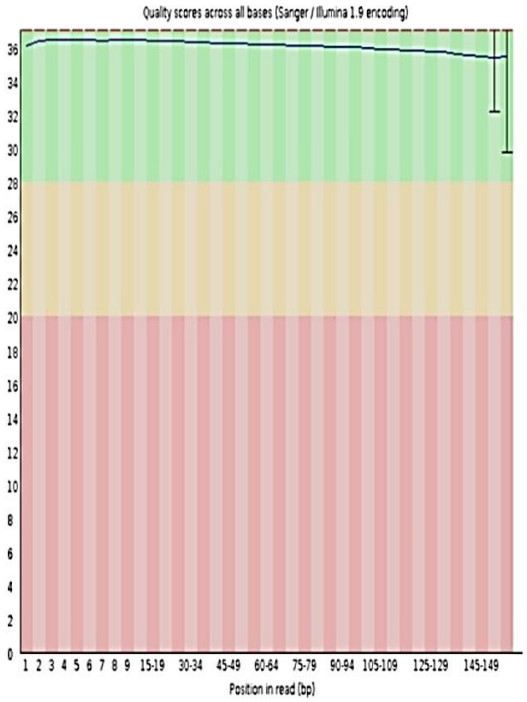


BC-6

Read 1

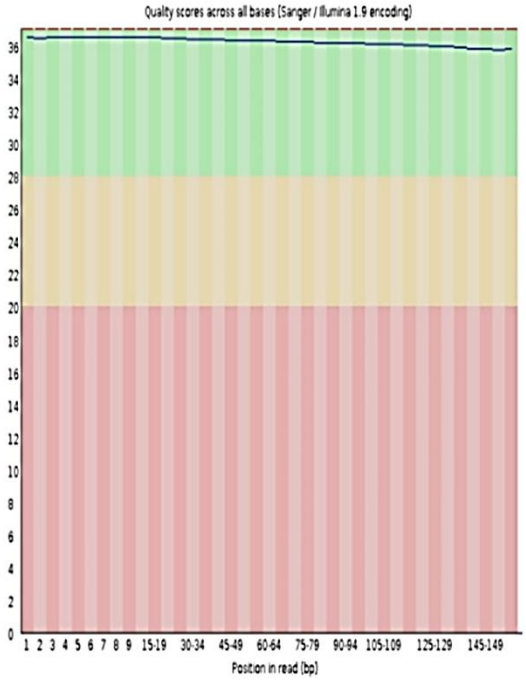


Read 2

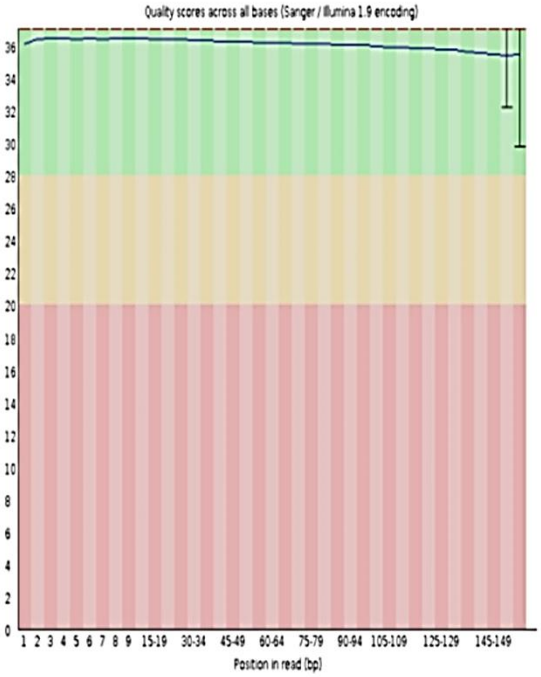


Adjacent normal BC-1

Read 1

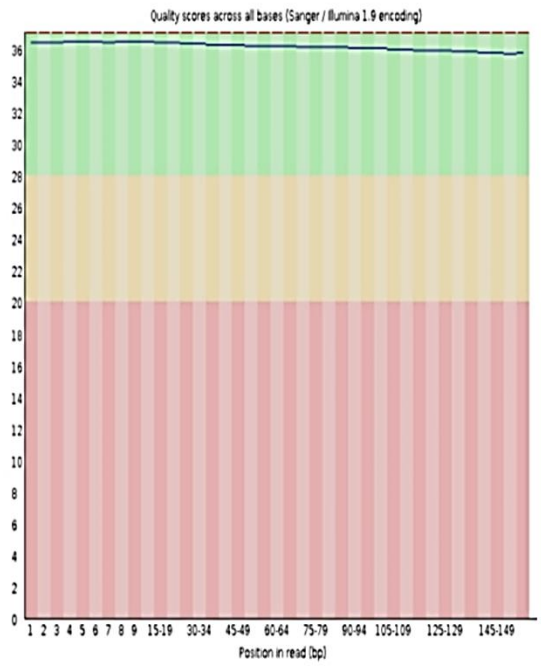


Read 2

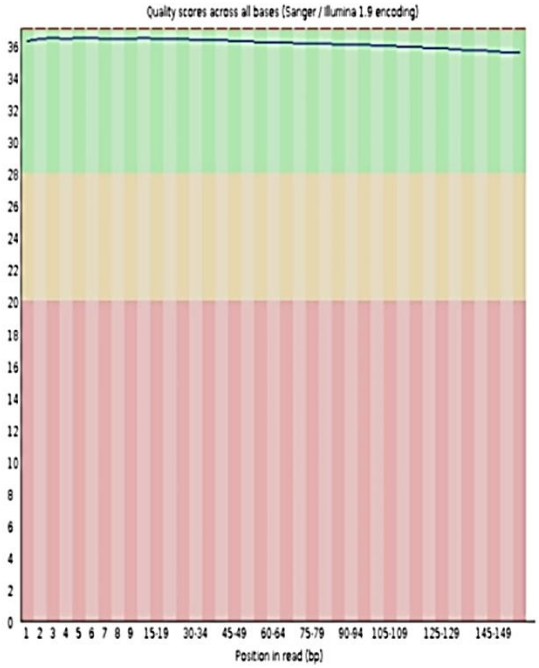


Adjacent normal BC-2

Read 1



Read 2



Sequence quality is vital for variant calling for accurate base calls to detect genetic variants.

4.2.5 Mutation landscape of breast cancer patients

Identifying variants in WES data involves detecting and characterizing genetic variations within the exome, which comprises the genome's protein-coding regions. The exome tumor DNA sequences (FASTQ files) were aligned with the human genome (hg38), and BC-1 and BC-2 sequences were also aligned with the corresponding adjacent normal sequence using the Burrows–Wheeler Aligner (BWA) package for each sample of breast cancer patients. BWA produced BAM files as output because the FASTQ files contained raw sequencing reads. These BAM files included where each read from the FASTQ file aligned in the reference genome (Krøigård *et al*., 2018).

Subsequently, the BAM files were converted into VCF files. Variations between the aligned reads and the reference genome were identified during this conversion process. Variants, including single nucleotide variants (SNVs) and insertions/deletions (indels), were recorded in the VCF format. Each variant in the VCF file are described with details such as genomic position, variant type, quality scores, and allele frequencies (Fernandez-Moya *et al*., 2020).

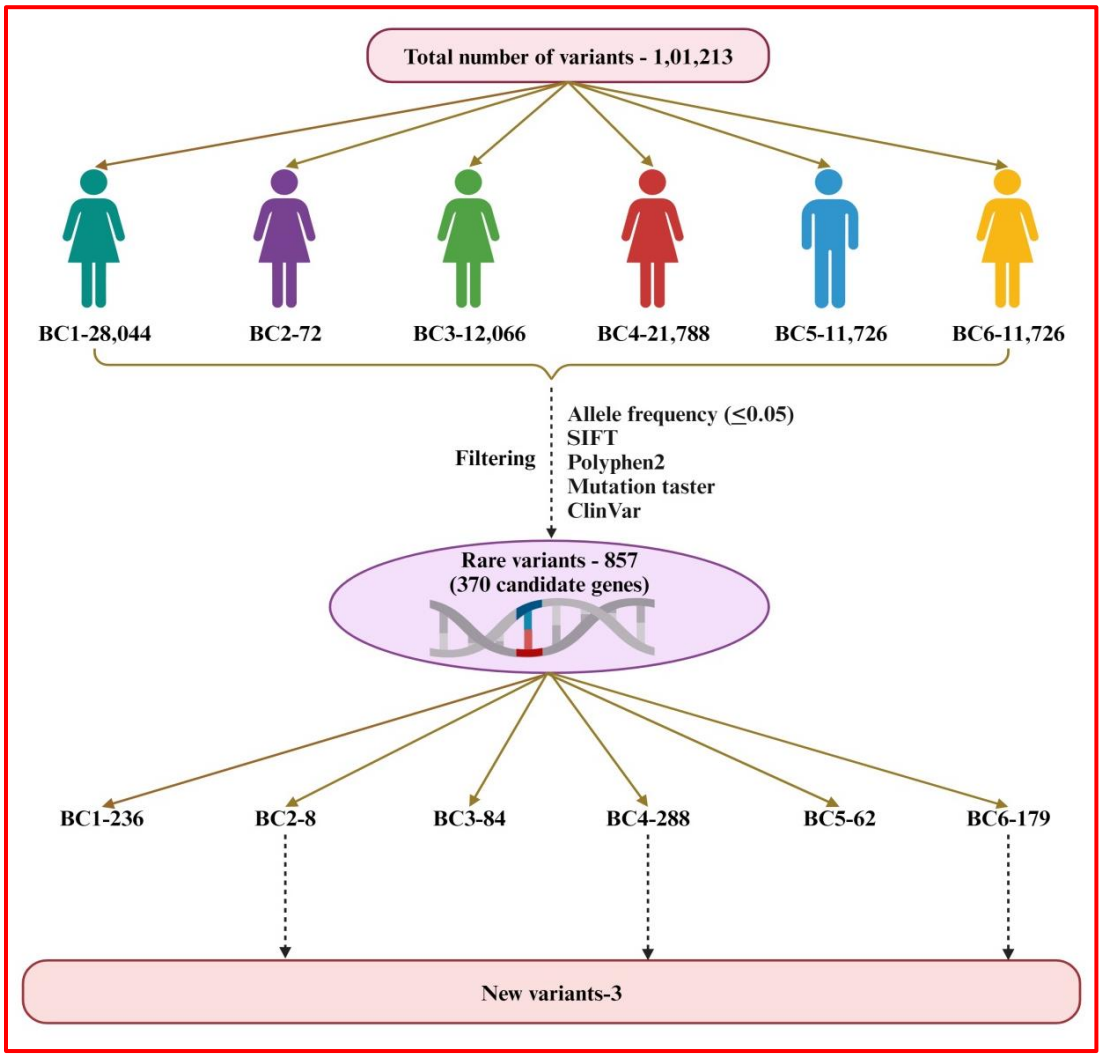
The VCF files were then annotated using ANNOVAR, a versatile annotation tool. ANNOVAR provided extensive annotations for the identified variants. It annotated variants with information about the mutated genes, including gene names, transcripts, and mutation functional consequences. Variants were also annotated based on their genomic regions, such as exonic or intronic locations. Additionally, ANNOVAR provided allele frequency data from various population databases to assess common variants in different populations. The tool also includes pathogenicity predictions, aiding in identifying potential disease-causing variants (Chang *et al*., 2020).

Specifically, a total of 1,01,213 variants were initially detected across all the breast cancer patients in the study. After applying filters such as allele frequency ($\leq$0.05), SIFT, polyphen2, mutation taster, and clinical significance (Benign, Benign/Likely_benign, Likely_benign, Pathogenic/Likely_pathogenic, Uncertain_significance) to annotate the rare variants, a total of 857 significant variants were identified. Exploring

rare variants characterized by low allele frequencies can significantly affect specific diseases. Allele frequencies can vary among different populations. Rare variants may be more prevalent in particular people or ethnic groups. Studying these variants helps to understand population-specific genetic factors that may contribute to disease susceptibility or response to treatment. Incorporating rare variants into genetic studies and clinical assessments can provide a personalized perspective on genetic variability and its implications for health and disease (Momozawa and Mizukami, 2021). So, we focused on rare variants of breast cancer patients in our research.

The flowchart **(Figure 15)** gives the overview of the six breast cancer patient's variants after using subsequent filters. The figure was created with BioRender.

Figure 15: Flowchart depicting steps of identified variants



Comprehensive details of the significant variants of six BC patients include Chromosome (Chr) number, Start and End positions, Reference (Ref) and Alternate (Alt) alleles, Gene name, RefGene annotation, Functional classification in RefGene, Exonic Function in RefGene, Region of the variant, cDNA level change, and Protein level change **(Supplementary Table 1)**.

Among the 857 significant variants identified, all were found to be heterozygous. Heterozygous variants suggest that one of the two copies of a gene is mutated, implying a potential involvement of these genetic alterations in the development and progression of breast cancer. Additionally, all of these variants were located in the exonic region of the genes, which are the coding regions responsible for producing proteins. Mutations in these coding regions can potentially affect the structure and function of proteins, possibly contributing to abnormal cell growth or other characteristics of cancer. Furthermore, all of these variants were classified as nonsynonymous single nucleotide variants (SNVs), resulting in amino acid changes in the protein sequence and potentially affecting protein structure and function. In the context of breast cancer, these changes could influence signaling pathways, DNA repair mechanisms, and cell cycle regulation. In the case of BC-4, two specific gene mutations were identified. One of the mutations in BC-4 occurred as a synonymous single nucleotide variant (SNV) in the gene *SCYL1* in exon 16 (c.2148C>T and c.2199C>T), resulting in genetic changes in the DNA sequence impact gene expression, potentially influencing the cellular environment related to breast cancer. Another mutation in BC-4 was identified as a stop-gain mutation that occurred in exon 25 of the *ULK4* gene (c.1678C>7 and c.2584C>T), resulting in premature termination of protein synthesis during translation. The loss of proper protein function can disrupt normal cellular processes, contributing to the uncontrolled growth of cancer cells.

The nonsynonymous mutations were exclusively observed in cases of DCIS (Ductal Carcinoma *in situ*) and primary tumors. 31,418 nonsynonymous mutations were identified via targeted deep sequencing across the six patients. The study primarily focused on nonsynonymous mutations, as synonymous mutations are generally considered nonfunctional in cancer research (Krøigård *et al*., 2018). In a separate cohort study involving 24 Taiwanese patients, 960 nonsynonymous mutations were identified using whole exome sequencing (WES), emphasizing novel variants within

nonsynonymous mutations (Chang *et al*., 2020). Within our study cohort, which included DCIS and primary tumor samples, 857 significant variants were identified. Notably, 853 variants were non-synonymous mutations. This observation highlights the potential functional relevance of these nonsynonymous mutations in the context of breast cancer.

Similar to female breast cancer (FBC), male breast cancer (MBC) is primarily driven by copy number alterations (CNA). Notably, major genetic factors associated with an elevated risk of MBC include germline mutations in the *BRCA2* and *BRCA1* genes and a family history of the disease. In the Netherlands, male breast cancer patients have been found to carry pathogenic *BRCA2* germline mutations (Moelans *et al*., 2019). Additionally, among Tunisian familial male breast cancer patients, a novel frameshift mutation in the *BRCA2* gene was identified (Ben Kridis-Rejeb *et al*., 2020). In our study, the male breast cancer patient also had a family history of the disease and was found to have a *BRCA2* mutation. The results indicate these changes can significantly affect gene function and contribute to disease development. We also found six BC patients' clinical significance and associated disease **(Supplementary Table 2)**.

Interesting fact, we found that the most common associated disease among five breast cancer patients is cardiomyopathy. The molecular pathways implicated in cardiomyopathy, including abnormal calcium handling and signaling, oxidative stress, and fibrosis, may intersect with pathways involved in breast cancer. Certain diseases and environmental factors that elevate inflammation and oxidative stress levels can serve as risk factors for both breast cancer and cardiovascular disease. Robust epidemiological evidence indicates that reduced physical activity, dietary patterns, environmental exposures, and clinical conditions like diabetes and family history are linked to an elevated risk for both cardiovascular disease and breast cancer (Gulati and Mulvagh, 2018).

The common gene mutation identified among five breast cancer patients is *TTN* (titin). *TTN* encodes titin, the most significant known protein, which plays a crucial role in muscle contraction, including cardiac muscle. Mutations in the *TTN* gene have been linked to a variety of cardiac disorders, including cardiomyopathies. *TTN* gene mutations were the most frequently observed mutations in both primary and metastatic breast cancer cases. The specific role of *TTN* in cancer is unknown. However, several studies

have revealed its direct involvement in cancer-related pathways (Saravia *et al*., 2019). *TTN* also plays a significant role in developing and maintaining chromosomal condensations. *TTN* may localize in chromosomes and serve as a scaffold for the appropriate binding and assembly of other proteins involved in chromosomal condensation. As a result, *TTN* mutations can alter chromosomal condensations and segregations, playing an important role for the onset of cancer. *TTN* mutations have been linked to increased risk of breast cancer, aggressive tumor features, and lower survival rates (Fernandez-Moya *et al*., 2020).

These results suggest that individuals carry specific genetic variants that predispose them to both breast cancer and cardiomyopathy. These variants could affect genes or pathways that play roles in both conditions. If a person with breast cancer develops cardiomyopathy, their treatment plan may need to be adjusted to minimize the risk of further cardiac complications. The finding emphasizes the importance of thorough cardiac evaluation before initiating breast cancer treatment.

Our present study with breast cancer patients, exhibited variants in coding regions of the genes after filters were applied. However, we wanted to explore the potential involvement of non-coding regions, specifically introns, in breast cancer progression. Mutations occurring in introns do not directly affect protein sequences. But introns are large segments of DNA that make up a significant portion of human genes and plays a significant role in influencing gene expression. In the exome analysis of breast cancer, we applied a specific criterion, considering only intronic variants with an allele frequency of $\leq$0.05. The results revealed a total of 183 intronic variants in six breast cancer patients **(Supplementary Table 3)**.

Introns play a role in regulating gene expression through regulatory regions and functional non-coding RNA genes. They also play a direct role through their length and regulation of transcription levels, timing, and splicing. Understanding these intronic variants can provide valuable insights into cancer's complex gene expression regulatory networks (Rigau *et al*., 2019). Considering both exonic and intronic variants is essential to understand the genetic factors influencing breast cancer. The identified intronic variants may be the regulatory mechanisms driving breast cancer progression, offering potential avenues for further research and personalized approaches to treatment.

We also examined the occurrence of mutations in *BRCA1* and *BRCA2* genes, which are recognized as hotspot genes frequently mutated in breast cancer. In our analysis, all breast cancer patients except BC-2 showed mutations in the *BRCA* gene, and the variants identified in the *BRCA* genes were already reported in the dbSNP database **(Supplementary Table 4)**. *BRCA1* and *BRCA2* are vital genes linked to autosomal dominant and highly penetrant forms of breast cancer, encoding Tumor Suppressor Gene (TSG) proteins. *BRCA1* is situated on chromosome 17q, whereas *BRCA2* resides on chromosome 13q, an acrocentric chromosome in males. Variations or mutations in these genes enhance the likelihood of developing breast cancer. The *BRCA1* protein contains 1863 amino acids, while *BRCA2* has 3418 amino acids. These proteins, also known as anti-oncogenes, are essential for repairing damaged DNA and protecting genetic information. If either of these genes is mutated, the cell's ability to repair damaged DNA is disturbed, accumulating genetic changes and mutations, ultimately contributing to cancer development (Armstrong *et al*., 2019).

A mutation in the *BRCA1* gene increases the probability of breast cancer in women by 60% to 80%. Moreover, hereditary mutations in *BRCA2* are discovered in approximately 35% of families with early-onset breast cancer in women (Metcalfe *et al*., 2018). Identifying mutations in these genes is vital for assessing the risk of breast cancer, implementing preventive measures, and developing targeted treatment strategies for individuals with an elevated genetic predisposition to breast cancer. This underscores the significance of gene analysis in informing personalized approaches to breast cancer risk and management.

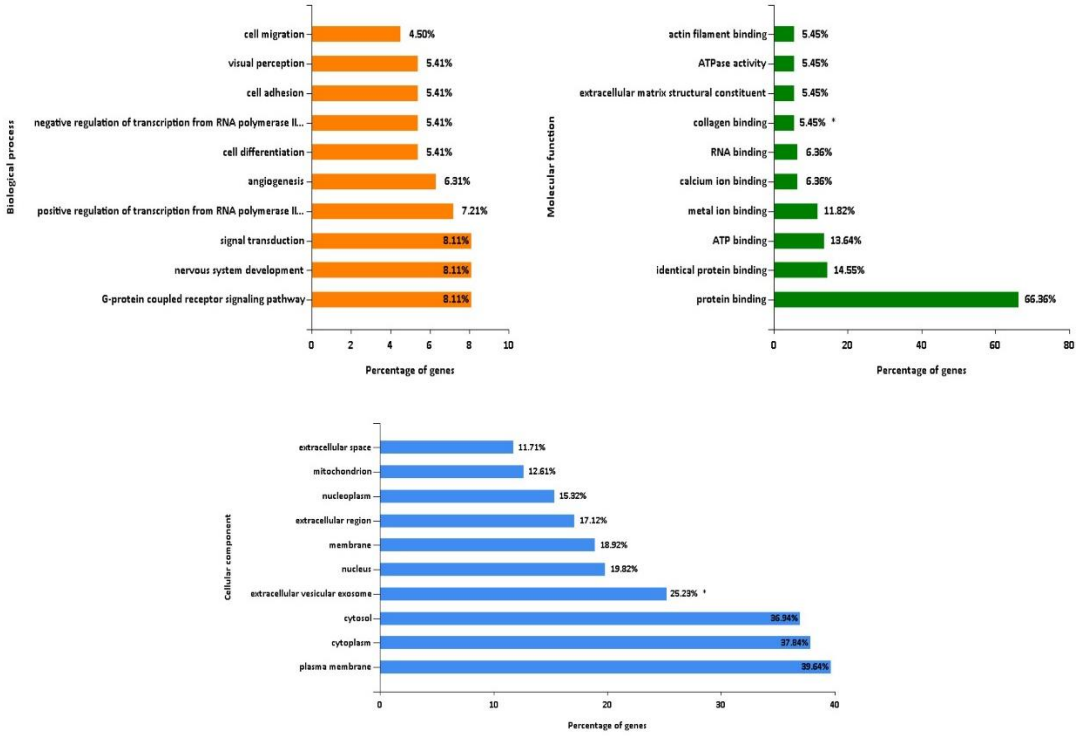
Exome sequencing data for breast cancer and adjacent normal tissue was submitted to the European Nucleotide Archive (ENA) repository, managed by the European Bioinformatics Institute (EBI). ENA acts as a centralized hub for nucleotide sequence data and associated metadata. The primary accession number for the exome data is PRJEB74646, and the secondary accession number is ERP159292.

4.2.6 Functional enrichment analysis of candidate genes of rare variants

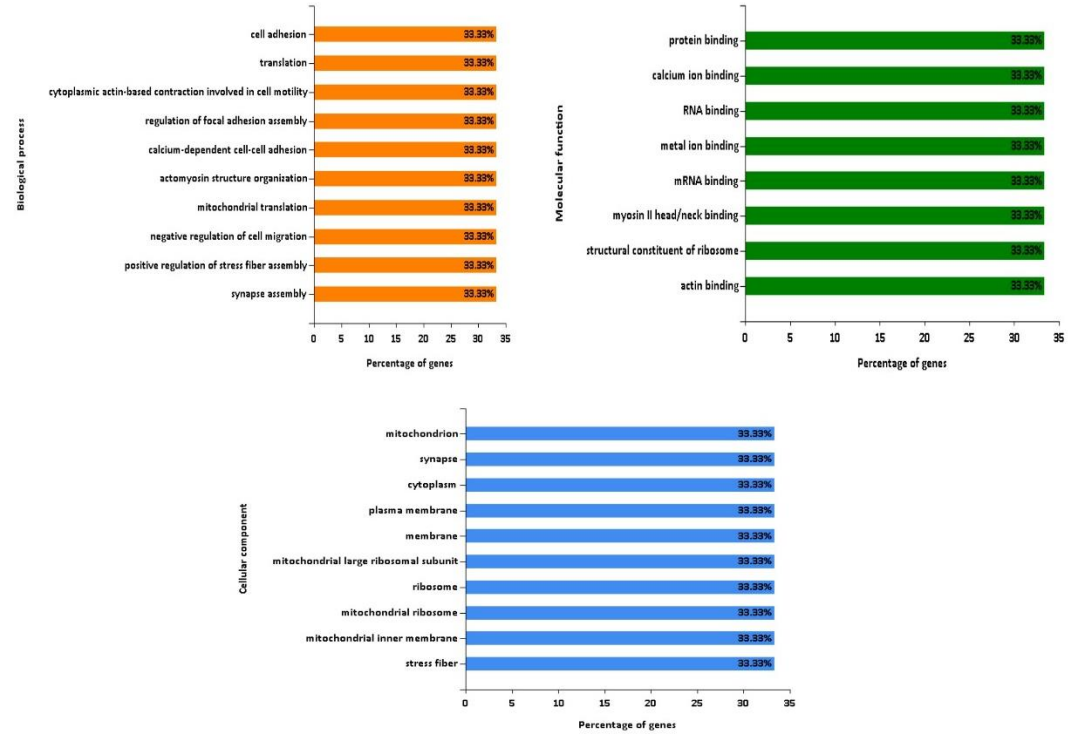
Once the mutations were identified, we also analyzed the gene ontology for 370 genes associated with rare variants. FunRich was used to examine the biological functions, molecular pathways, and cellular processes of the candidate genes. The output of functional enrichment analysis of six breast cancer patients' candidate genes is given in **Figure 16**.

Figure 16: Functional enrichment analysis

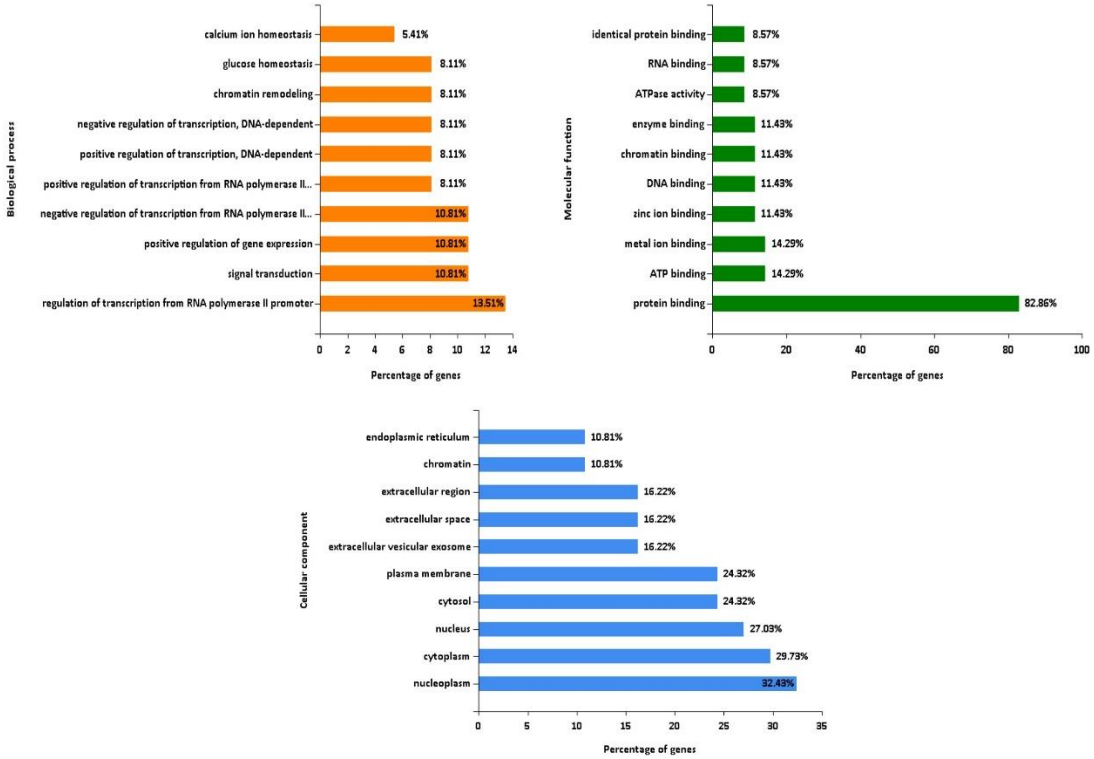
BC-1



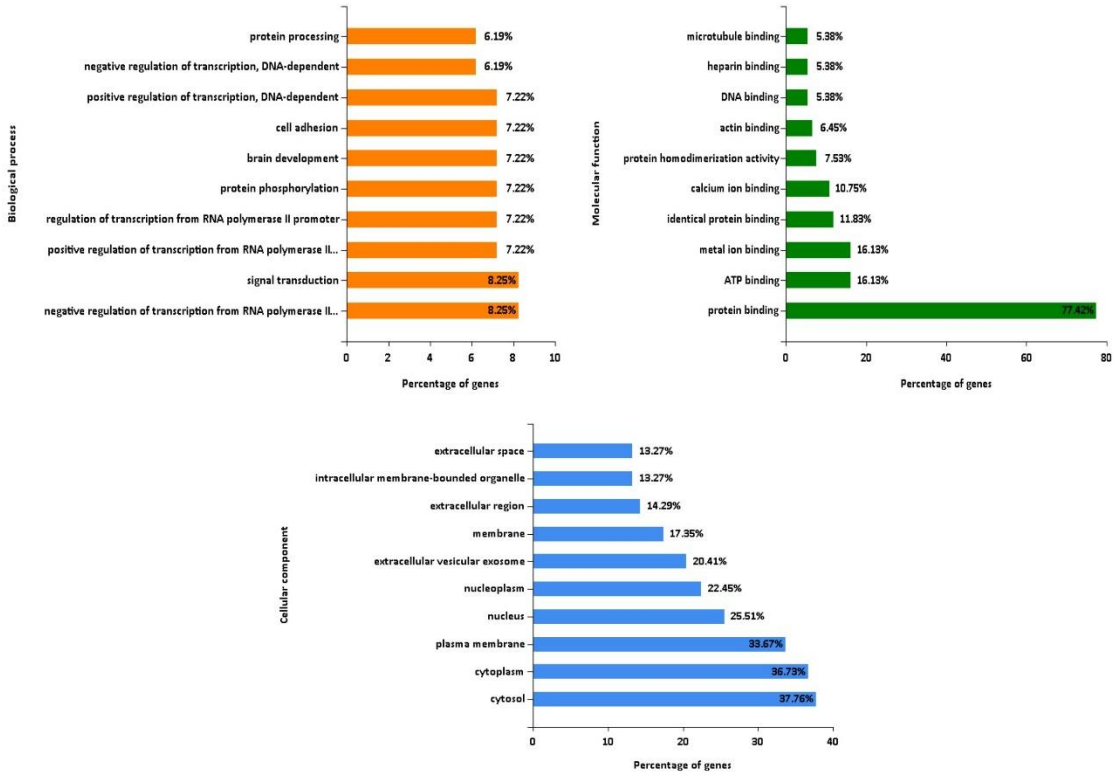
BC-2



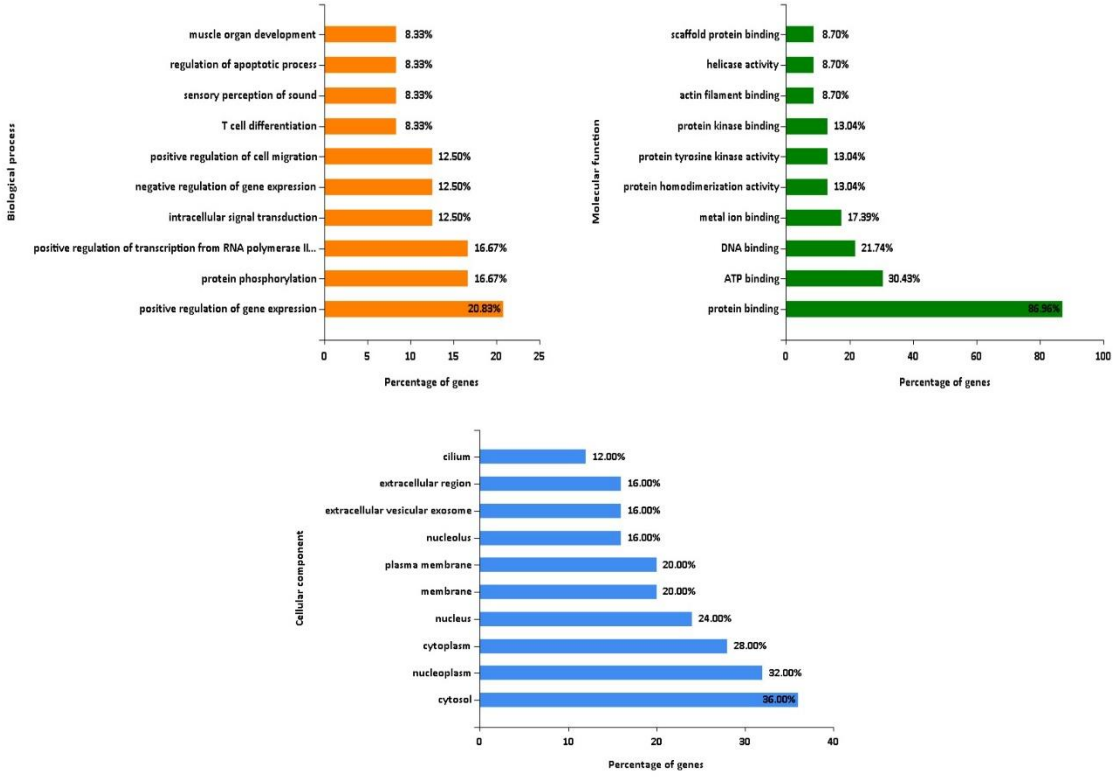
BC-3


Biological process
calcium ion homeostasis 5.41%
glucose homeostasis 8.11%
chromatin remodeling 8.11%
negative regulation of transcription, DNA-dependent 8.11%
positive regulation of transcription, DNA-dependent 8.11%
positive regulation of transcription from RNA polymerase II... 8.11%
negative regulation of transcription from RNA polymerase II... 10.81%
positive regulation of gene expression 10.81%
signal transduction 10.81%
regulation of transcription from RNA polymerase II promoter 13.51%
0 2 4 6 8 10 12 14
Percentage of genes
Molecular function
identical protein binding 8.57%
RNA binding 8.57%
ATPase activity 8.57%
enzyme binding 11.43%
chromatin binding 11.43%
DNA binding 11.43%
zinc ion binding 11.43%
metal ion binding 14.29%
ATP binding 14.29%
protein binding 82.86%
0 20 40 60 80 100
Percentage of genes
Cellular component
endoplasmic reticulum 10.81%
chromatin 10.81%
extracellular region 16.22%
extracellular space 16.22%
extracellular vesicular exosome 16.22%
plasma membrane 24.32%
cytosol 24.32%
nucleus 27.03%
cytoplasm 29.73%
nucleoplasm 32.43%
0 5 10 15 20 25 30 35
Percentage of genes

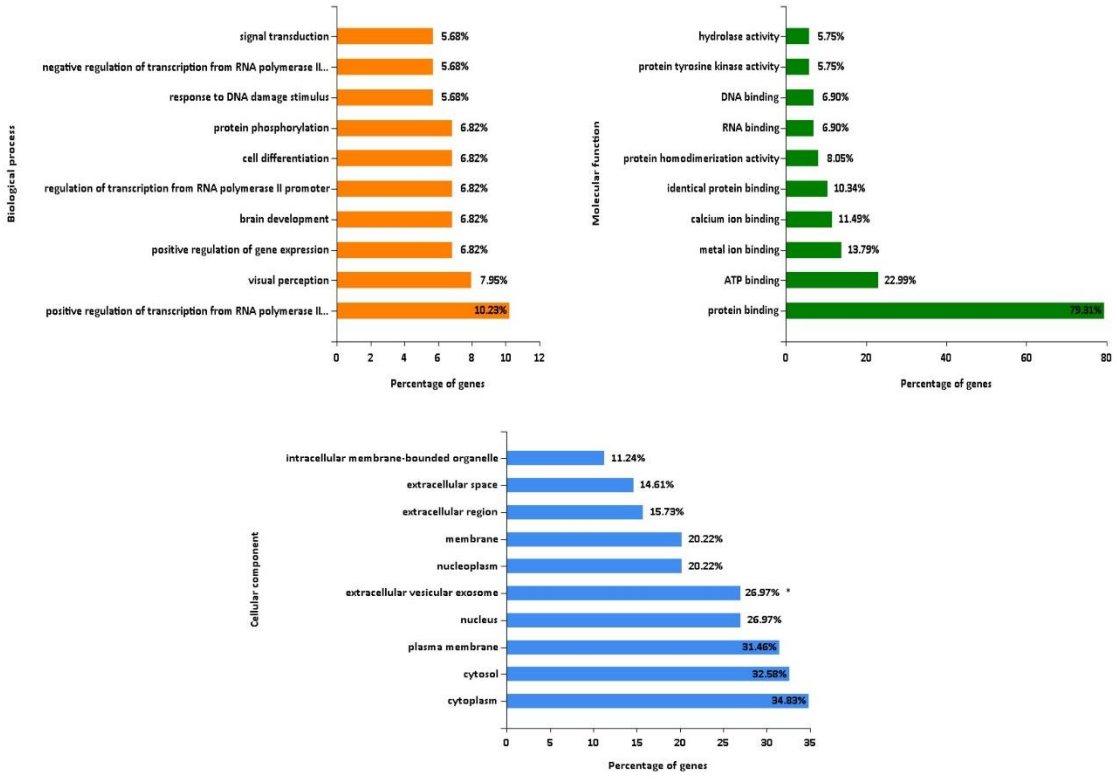
BC-4


Biological process
protein processing 6.19%
negative regulation of transcription, DNA-dependent 6.19%
positive regulation of transcription, DNA-dependent 7.22%
cell adhesion 7.22%
brain development 7.22%
protein phosphorylation 7.22%
regulation of transcription from RNA polymerase II promoter 7.22%
positive regulation of transcription from RNA polymerase II... 7.22%
signal transduction 8.25%
negative regulation of transcription from RNA polymerase II... 8.25%
0 2 4 6 8 10
Percentage of genes
Molecular function
microtubule binding 5.38%
heparin binding 5.38%
DNA binding 5.38%
actin binding 6.45%
protein homodimerization activity 7.53%
calcium ion binding 10.75%
identical protein binding 11.83%
metal ion binding 16.13%
ATP binding 16.13%
protein binding 77.42%
0 20 40 60 80
Percentage of genes
Cellular component
extracellular space 13.27%
intracellular membrane-bounded organelle 13.27%
extracellular region 14.29%
membrane 17.35%
extracellular vesicular exosome 20.41%
nucleoplasm 22.45%
nucleus 25.51%
plasma membrane 33.67%
cytoplasm 36.73%
cytosol 37.76%
0 10 20 30 40
Percentage of genes

BC-5


Biological process
muscle organ development 8.33%
regulation of apoptotic process 8.33%
sensory perception of sound 8.33%
T cell differentiation 8.33%
positive regulation of cell migration 12.50%
negative regulation of gene expression 12.50%
intracellular signal transduction 12.50%
positive regulation of transcription from RNA polymerase II... 16.67%
protein phosphorylation 16.67%
positive regulation of gene expression 20.83%
0 5 10 15 20 25
Percentage of genes
Molecular function
scaffold protein binding 8.70%
helicase activity 8.70%
actin filament binding 8.70%
protein kinase binding 13.04%
protein tyrosine kinase activity 13.04%
protein homodimerization activity 13.04%
metal ion binding 17.39%
DNA binding 21.74%
ATP binding 30.43%
protein binding 86.96%
0 20 40 60 80 100
Percentage of genes
Cellular component
cilium 12.00%
extracellular region 16.00%
extracellular vesicular exosome 16.00%
nucleolus 16.00%
plasma membrane 20.00%
membrane 20.00%
nucleus 24.00%
cytoplasm 28.00%
nucleoplasm 32.00%
cytosol 36.00%
0 10 20 30 40
Percentage of genes

BC-6


Biological process
signal transduction 5.68%
negative regulation of transcription from RNA polymerase II... 5.68%
response to DNA damage stimulus 5.68%
protein phosphorylation 6.82%
cell differentiation 6.82%
regulation of transcription from RNA polymerase II promoter 6.82%
brain development 6.82%
positive regulation of gene expression 6.82%
visual perception 7.95%
positive regulation of transcription from RNA polymerase II... 10.23%
0 2 4 6 8 10 12
Percentage of genes
Molecular function
hydrolase activity 5.75%
protein tyrosine kinase activity 5.75%
DNA binding 6.90%
RNA binding 6.90%
protein homodimerization activity 8.05%
identical protein binding 10.34%
calcium ion binding 11.49%
metal ion binding 13.79%
ATP binding 22.99%
protein binding 79.31%
0 20 40 60 80
Percentage of genes
Cellular component
intracellular membrane-bounded organelle 11.24%
extracellular space 14.61%
extracellular region 15.73%
membrane 20.22%
nucleoplasm 20.22%
extracellular vesicular exosome 26.97%
nucleus 26.97%
plasma membrane 31.46%
cytosol 32.58%
cytoplasm 34.83%
0 5 10 15 20 25 30 35
Percentage of genes

Functional enrichment study showed that the genes are involved in key biological activities, including signal transduction, G-protein coupled receptor signaling pathway, positive regulation of gene expression, synapse assembly, and positive regulation of transcription from RNA polymerase II, which are highly relevant in breast cancer. Signal transduction pathways are crucial in breast cancer. Dysregulation of these pathways can lead to uncontrolled cell growth and proliferation, which are hallmark features of cancer cells. Aberrant signaling, primarily through growth factor receptors, can contribute to breast cancer development (Ortega *et al*., 2020). G-protein-coupled receptors facilitate the transmission of signals from the extracellular environment to the inside of the cell. These receptors can influence cellular responses, including those related to cell survival and proliferation, critical to breast cancer progression (Yu *et al*., 2018). Gene expression control plays a vital in breast cancer. Positive regulation of specific genes can lead to the overexpression of growth-promoting proteins, contributing to tumor growth and metastasis. Synapse assembly may signify the significance of cell-to-cell communication and interactions in cancer progression, including the signaling exchanges between tumor cells and their surrounding microenvironment. Positive regulation of transcription from RNA Polymerase II indicates increased transcriptional activity, potentially leading to the upregulation of genes involved in breast cancer growth and progression (Zhou *et al*., 2019).

The cellular processes of the candidate genes' mainly occurr in the plasma membrane, cytosol, and cytoplasm. Alterations in the plasma membrane can impact cell signaling. Changes in membrane receptors and transporters can affect the response to external stimuli, including growth factors and hormones that influence breast cancer cells. Cytosol and cytoplasm are the cellular parts crucial to various cellular processes, including signal transduction and regulating intracellular pathways. Dysregulation in these areas can contribute to the uncontrolled growth of breast cancer cells (Szlasa *et al*., 2020). The molecular function of candidate genes is mainly involved in protein binding. Protein-protein interactions are pivotal in orchestrating cellular processes. In the context of breast cancer, these interactions can affect the function of critical proteins involved in cell cycle regulation, apoptosis, and metastasis (Hozhabri *et al*., 2022).

Dysregulation of signal transduction pathways, aberrant plasma membrane receptor activity, and altered protein binding events are interconnected processes that contribute to the oncogenic properties of breast cancer cells. These events can lead to uncontrolled growth, survival, and invasion, all of which are hallmarks of cancer.

To summarize, the candidate genes identified in the study and their associated biological processes and molecular functions will provide insights into the molecular mechanisms underlying breast cancer. Abnormalities in these mechanisms can cause the development and progression of breast cancer, making them important targets for further research and potential therapeutic interventions.

Among the significant variants identified, three were identified as new variants that have not been previously reported in breast cancer. These new variants were identified in *MRPL13, MYPBC3,* and *PTCH1* genes. This *PTCH1* gene variant was identified in two different breast cancer patient samples. Upon further investigation using the VarSome database, additional information regarding these variants and their associations with other cancer types was analyzed. *PTCH1* variant was found to be reported in different cancer types, such as endometrial cancer and colorectal cancer. The *MYBPC3* variant seems to be involved in several other cancer types, including esophagogastric cancer, head and neck cancer, endometrial cancer, and myeloproliferative neoplasms. The presence of this specific variant in breast cancer patient's highlights that genetic alteration may be population-specific due to epidemiological risk factors.

The genomic analysis conducted in the current study identifies the molecular landscape of breast cancer, providing valuable insights into the genetic alterations, potential risk factors, and pathways implicated in the disease. Through a meticulous examination of patient samples, we have uncovered numerous genomic variants and mutations, each contributing to the complexity of breast cancer uniquely. The study identified multiple genetic variants, with a particular emphasis on nonsynonymous mutations. Three new variants were identified among the identified variants, representing previously unreported genetic variants in the context of breast cancer. These novel findings underscore the uniqueness of the breast cancer genomic landscape in the studied population. Signal transduction, positive regulation of gene expression, and G-protein

coupled receptor signaling emerged as critical processes. Cellular locations such as the plasma membrane and cytoplasm were notably associated, emphasizing the importance of these environments in breast cancer pathogenesis. Associations with cardiomyopathy show that epidemiological risk factors are common in both cardiovascular health and breast cancer. This study contributes significantly to our understanding of breast cancer at the genomic level, unraveling novel variants, elucidating potential risk factors, and providing a foundation for personalized therapeutic interventions. Integrating genomics into clinical practice holds promise for advancing precision medicine in oncology.

Exome sequencing analysis of phase II provided extensive information from six breast cancer patients, yielding a multitude of genetic variants. Sanger sequencing is vital as a quality control step within the genomic analysis process. It allows for the targeted validation of specific variants of interest that hold clinical or research significance. This selective validation helps mitigate false positives or sequencing errors that occur by exome sequencing (Crossley *et al*., 2020). Integrating exome and Sanger sequencing techniques is essential for ensuring the precision and reliability of genetic variant identification. Particularly in clinical settings, Sanger sequencing confirms the presence of mutations, providing the correct treatment choice. Treatment strategies are customized to match the patient's unique genetic profile. Targeted medicines can be identified by knowing the precise genetic mutations in a patient's tumor. By focusing on specific targets, this focused strategy can reduce needless treatments, minimize side effects, and enhance overall quality of life both during and after treatment.

Phase III

4.3 Mutation profiling of breast cancer patients

Mutation profiling using Sanger sequencing in breast cancer involves systematically analysing genetic alterations within tumor DNA, revealing critical insights into the specific mutations and genomic changes that drive breast cancer development and progression (Hussein *et al*., 2020). From whole exome sequencing results, we were able to identify three new variants that have been detected in different patients: *MRPL13* (c.380T>C) in BC-2 patient, *MYBPC3* (c.2816G>A) in BC-4 patient and *PTCH1* (c.1889G>A) in BC-4 and BC-6 patients. *MRPL13* (Mitochondrial Ribosomal Protein

L13) is integral to mitochondrial ribosomal function, potentially impacting energy production and contributing to many diseases, including cancer. *MYBPC3* (Myosin Binding Protein C), primarily expressed in cardiac and skeletal muscles, is typically linked to cardiomyopathies. *PTCH1* (Patched 1), a crucial player in the hedgehog signaling pathway, is implicated in various cancers, such as basal cell carcinoma and medulloblastoma. Validating the novel variants by Sanger sequencing enables the identification of promising therapeutic targets and developing personalized treatment strategies for breast cancer patients, especially in cases where the choice of treatment becomes very difficult.

4.3.1 Validating variants of uncertain significance

The three variants identified by whole exome sequencing have uncertain significance. The variants of uncertain significance refer to a genetic variant or mutation, and its precise impact on health or disease has not been clearly understood till date (Federici and Soddu, 2020). The three new variants are reported by us for the first time and it do not contain rsid (Reference SNP cluster ID). Detailed information about these variants is given in **Table 8**.

Table 8: Exonic and nonsynonymous SNV of uncertain significance

Patients	Chr	Ref	Alt	Gene. Ref Gene	Region	cDNA level change	Protein level change	Clinvar	avsnp 150 (rsID)
BC-2	chr8	A	G	*MRPL13*	exon5	c.T380C	p.L127p	Uncertain significance	-
BC-4	chr11	C	T	*MYBPC3*	exon26	c.G2816A	p.R939Q	Uncertain significance	-
BC-4	chr9	C	T	*PTCH1*	exon13	c.G1889A	p.R630H	Uncertain significance	-
BC-6	chr9	C	T	*PTCH1*	exon13	c.G1889A	p.R630H	Uncertain significance	-

Since we are reporting the new variants for the first time, no specific primers were available. Hence, we design the specific primers for the new variants. The chromosome sequences for each variant were obtained from the GRCh38.p14 Primary Assembly from NCBI (National Center for Biotechnology Information) database. Flanking or cosmic mutation sequences were acquired from the Ensembl database. Subsequently, the flanking sequences were aligned with the NCBI reference genome to determine the altered region of the variant precisely. The modified region is highlighted in red colour in both the sequences derived from NCBI and Ensembl are given in **Figure 17**.

Figure 17: FASTA sequence of the variants

***MRPL13*: c.380T>C**

NCBI

>NC_000008.11:120419565-120420165 Homo sapiens chromosome 8, GRCh38.p14 Primary Assembly

TTAACAGGGTTTCTGATCATAGCTCCTTGTTCAAAAGCAAGCAGTGATTAAAATTTTTTTTAAAAACTTACTTTTAATT
TTTAGTCCAATGATGGTTCATTTTACATTTTAAAATATTAATAGAATTATCATTTTTTATACATACATATTCATTTAAA
AATACAAAAATATTTCAGAAGTTGACATAAAATTTTTGAAAATAAGTCTGTGGCATTTATTTAAAAAGAAAATGAAACA
TTTAATTGAAAATTTTGGAAAAGCATTGTTACCAATGACCACTGACTTACCTCATCTGGAAAA**[A]**GATGCAACCTTTC
CATCATTGTTCTTCTGTGAAGGTTTTTTGGCAGCATGCCATAAATAGCTAGTTTTACAATCTGAAAGATATACATAGGA
CACAATAAGAAAATTGTGACTTGCGATAGTAAGAAAGCAAACATAATTAGAGATGATGATGAAAATGAGGGGTTTAGAT
TCAACACGAATATAGGAAAGAAATAGACCTAAGTGGATATGTGGAAGCAACAGAAAAGAAACTTTTATTACTTTGTACA
TTACTATTTTCTTTTGGCTCCATTTATTGTTTATTTAGAAAAAGAAAAAA

Ensembl

TTTATCCATGATTCACAGAAGTTTTGCTACTGGTCACATAAAAAGTGAACTAAATTAACACTTTTTCATAGTGAATGAA
ATTATCAACAAGTACCTCAATTTAACAGGGTTTCTGATCATAGCTCCTTGTTCAAAAGCAAGCAGTGATTAAAATTTTT
TTTAAAAACTTACTTTTAATTTTTAGTCCAATGATGGTTCATTTTACATTTTAAAATATTAATAGAATTATCATTTTTT
ATACATACATATTCATTTAAAAATACAAAAATATTTCAGAAGTTGACATAAAATTTTTGAAAATAAGTCTGTGGCATTT
ATTTAAAAAGAAAATGAAACATTTAATTGAAAATTTTGGAAAAGCATTGTTACCAATGACCACTGACTTACCTCATCTG
GAAAA**R**GATGCAACCTTTCCATCATTGTTCTTCTGTGAAGGTTTTTTGGCAGCATGCCATAAATAGCTAGTTTTACAAT
CTGAAAGATATACATAGGACACAATAAGAAAATTGTGACTTGCGATAGTAAGAAAGCAAACATAATTAGAGATGATGAT
GAAAATGAGGGGTTTAGATTCAACACGAATATAGGAAAGAAATAGACCTAAGTGGATATGTGGAAGCAACAGAAAAGAA
ACTTTTATTACTTTGTACATTACTATTTTCTTTTGGCTCCATTTATTGTTTATTTAGAAAAAGAAAAAAGATTAAAAGT
AACCCGTATATGTGCCAATTTTAGATAGAATAATTATCATTTGCACTCATATTTCAATATAATTTGATGTGTGTTAAAG
CTAAACCCACT

***MYBPC3*: c.2816G>A**

NCBI

>NC_000011.10:47334831-47335431 Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly

GGGATTACAGGCTTGAGCCACTGTGCCTGGCCACCCTCTCTGCACTTTTTCCCTAGGCCTAGCGCATGGACGATGGCTC
CAACCCCTCCTGTCTCTGCCCAGCGTTCTGGGCAGAGCATTCTGGGCCTCCCCAACTGTCCCCACCTCCACTGGACACC
AAGGGCCTGGGGTGTCAATGGCGGGTCTTGTGACTGCACAAAGGGGCACTCACGCAGGATCTCCTGCACTGTCACCGGC
TCCGTGGTGGTAACAGGGGCTCCAGGCCCTGCCATATTGTGTGCCCGCACTCGGAAAAGCAGC**[C]**GGGCCCCCGTGGG
CAGGTCCTTCACCAGTATCGATGTGTGCTCTGTCAGCCCCTGCAGGGCAGCCACCCACTCTGAGCCTGGGGGTGGGGAG
GGGGAGGCAAGGCCACAGGCTGTGTCACCACTGACACCCCACTCCCACTGCCCACTCCTCTGATAGGAATCTCCAGGAT
TTAAATATGTTTTTTTAAATTTTTATTTGTTTATTTATTTTTGAGATGGAGTTTCGCTCTTGTCTCCCAGGCTGGAGTG
CAATGGCATGATCTTGGCTCACCGCAACTTCTGCCTCCCGAGTTCAAGTG

Ensembl

TTTTGTATTTTTAGTAGAGACAGGGTTTCACTATGTTGGCCAGGCTGGTCTCAAACTCCTGACCTCAAGTGATCTGCCT
GCCTTGGCCTCCTAAAGTGCTGGGATTACAGGCTTGAGCCACTGTGCCTGGCCACCCTCTCTGCACTTTTTCCCTAGGC
CTAGCGCATGGACGATGGCTCCAACCCCTCCTGTCTCTGCCCAGCGTTCTGGGCAGAGCATTCTGGGCCTCCCCAACTG
TCCCCACCTCCACTGGACACCAAGGGCCTGGGGTGTCAATGGCGGGTCTTGTGACTGCACAAAGGGGCACTCACGCAGG
ATCTCCTGCACTGTCACCGGCTCCGTGGTGGTAACAGGGGCTCCAGGCCCTGCCATATTGTGTGCCCGCACTCGGAAAA
GCAGC**[COSMIC_MUTATION]**GGGCCCCCGTGGGCAGGTCCTTCACCAGTATCGATGTGTGCTCTGTCAGCCCCTGCA
GGGCAGCCACCCACTCTGAGCCTGGGGGTGGGGAGGGGGAGGCAAGGCCACAGGCTGTGTCACCACTGACACCCCACTC
CCACTGCCCACTCCTCTGATAGGAATCTCCAGGATTTAAATATGTTTTTTTAAATTTTTATTTGTTTATTTATTTTTGA
GATGGAGTTTCGCTCTTGTCTCCCAGGCTGGAGTGCAATGGCATGATCTTGGCTCACCGCAACTTCTGCCTCCCGAGTT
CAAGTGATTCTCCTGCCTCAGCCTCCCGCGTAGCTGGGATCACAGGTGCCCACCACCACACCCGGCTAATTTTGTATTT
TTAGTAGAGACGGGGTTTCTCCACGTT

PTCH1: c.1889G>A

NCBI

>NC_000009.12:95468656-95469256 Homo sapiens chromosome 9, GRCh38.p14 Primary Assembly

AAGGAAAAAGAAGAAAAGTAGAAGCAATCTGATGAACTCCAAAGGTTCTGTTATTTTTTTGAAGACAGGAAGAGCCTTA
AGTTGTGGCAGATTACCTTGGCTTTTGGTTTCAAGAGGAAAGGAGCATAGTGCTTCTCAGCAAAAGATGAGAGTGTCCA
CTTCGTACAGGGGGGCTCGAGGCAGTGGAGGCTGGAGTCGGAGAACTGGGAGAGCAGGTCCCTTGTGGAGCTGGTGCTC
TCTGGGCTCTGGCAGCTGAGGGTGTCCTGTGTCACGGTGACGGGCTGCACAGAGATCTCGGAG**[C]**GCGGCTCAGCGGT
GGTGTAGTACACGTGCGTGTGGGGGTCGTACTCCGTGCGGAGCTGGACAGTGGACTGCATGGTAATCTGCGTTTCATGG
GCAAAGCTGTGGCTGCTGTAGGGAGGTGGGGGGCTGTAGCGGGTATTGTCGTGTGTGTCGGTGTAGGCCTGAGGTTCAA
CCTGAATCACTCTGCTGACGCAGGGGCTGAAAGGAGGGGAAACATGTTGCAATGTTATGCTGAAACAGGGAAATGGTGC
TTTCATTCTGCCATTTTTCACTGTGTACGGAGAATACCCATTTTACACAG

Ensembl

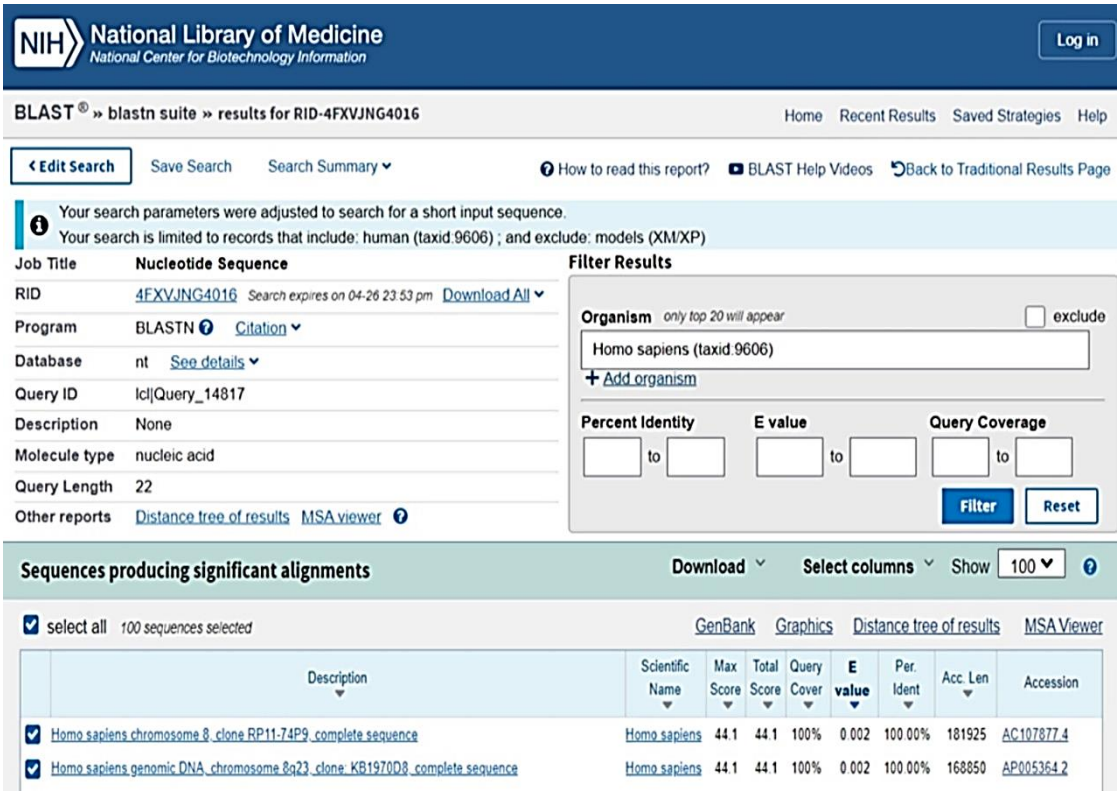
CCCATGGAAGATGACCTCATTAAAATGGCCCTATGCTCTAAAAGCATTGACTTTTGGAGGACTGAAATGTATCATACTT
AAACGAAATTTTTTTTTTTTTAAGGAAAAAGAAGAAAAGTAGAAGCAATCTGATGAACTCCAAAGGTTCTGTTATTTTT
TTGAAGACAGGAAGAGCCTTAAGTTGTGGCAGATTACCTTGGCTTTTGGTTTCAAGAGGAAAGGAGCATAGTGCTTCTC
AGCAAAAGATGAGAGTGTCCACTTCGTACAGGGGGGCTCGAGGCAGTGGAGGCTGGAGTCGGAGAACTGGGAGAGCAGG
TCCCTTGTGGAGCTGGTGCTCTCTGGGCTCTGGCAGCTGAGGGTGTCCTGTGTCACGGTGACGGGCTGCACAGAGATCT
CGGAG**[COSMIC_MUTATION]**GCGGCTCAGCGGTGGTGTAGTACACGTGCGTGTGGGGGTCGTACTCCGTGCGGAGCT
GGACAGTGGACTGCATGGTAATCTGCGTTTCATGGGCAAAGCTGTGGCTGCTGTAGGGAGGTGGGGGGCTGTAGCGGGT
ATTGTCGTGTGTGTCGGTGTAGGCCTGAGGTTCAACCTGAATCACTCTGCTGACGCAGGGGCTGAAAGGAGGGGAAACA
TGTTGCAATGTTATGCTGAAACAGGGAAATGGTGCTTTCATTCTGCCATTTTTCACTGTGTACGGAGAATACCCATTTT
ACACAGCTCTGACTTAGGAGAGAAACATCACATTGCTGAATTCCAGAAACATCGCCTGGTTGATAACATCACCTGGTTC
ATCGCCTGGTGAACTAGATAAAAGTCA

The FASTA sequences of the specific variants were utilized for designing primers. The primer design process was conducted using the Primer3plus software. The forward and reverse primers for the three variants were subjected to a nucleotide BLAST analysis, and the primers were confirmed to target the desired gene of interest specifically. The nucleotide BLAST of the three variants is given in **Figure 18**.

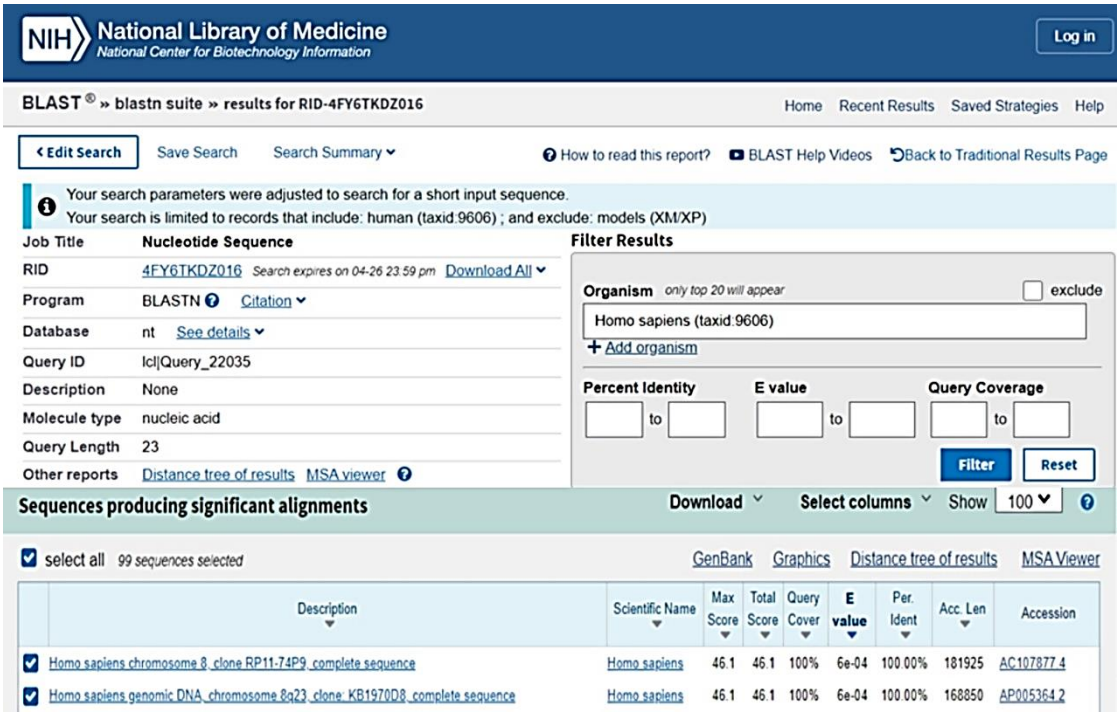
Figure 18: Nucleotide BLAST for variants

***MRPL13*: c.380T>C**

Forward Primer

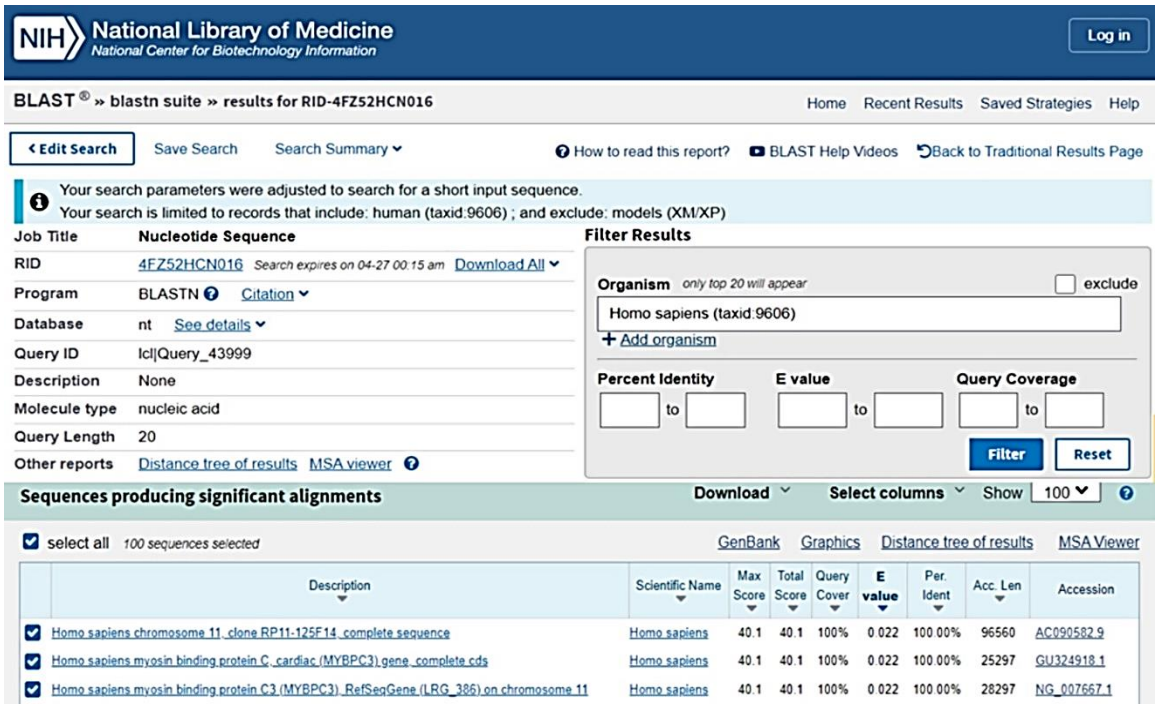


Reverse primer

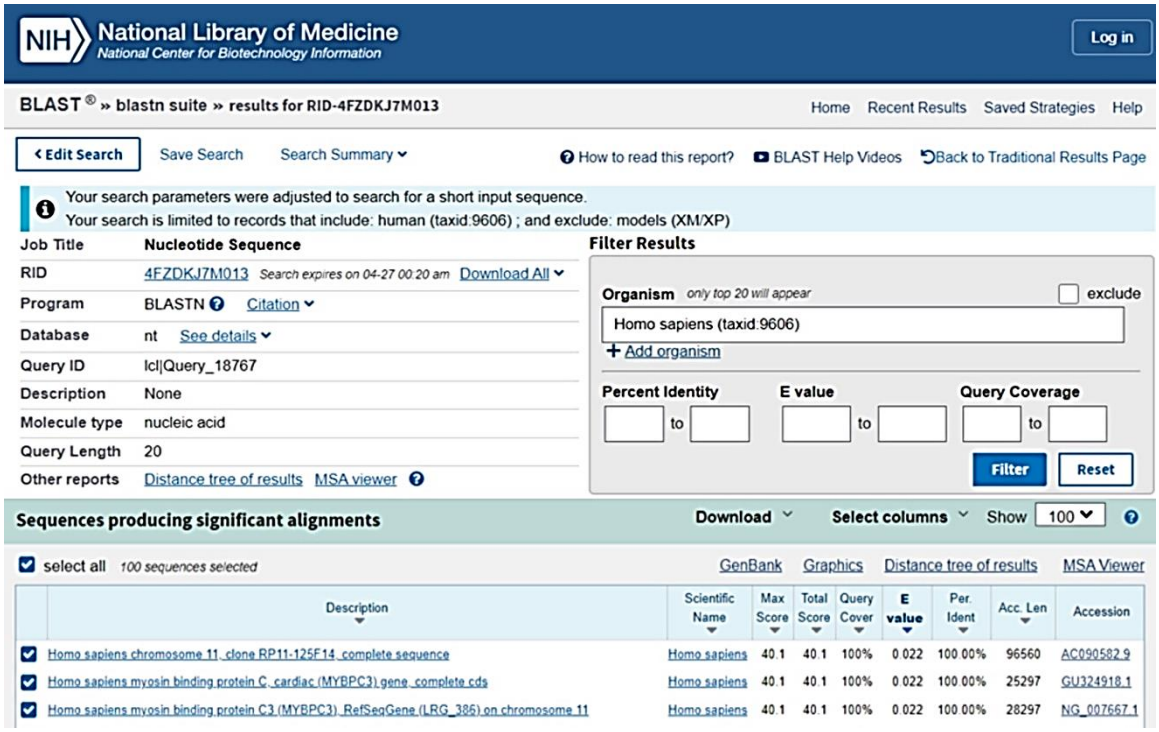


MYBPC3: c.2816G>A

Forward primer

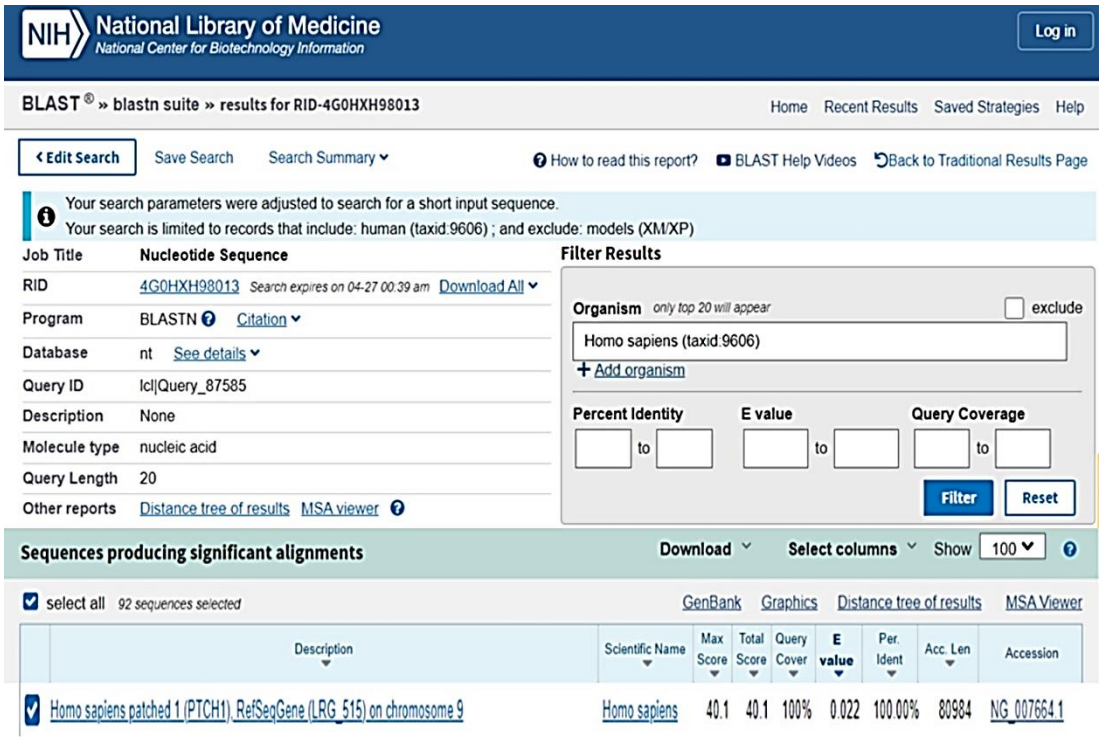


Reverse primer

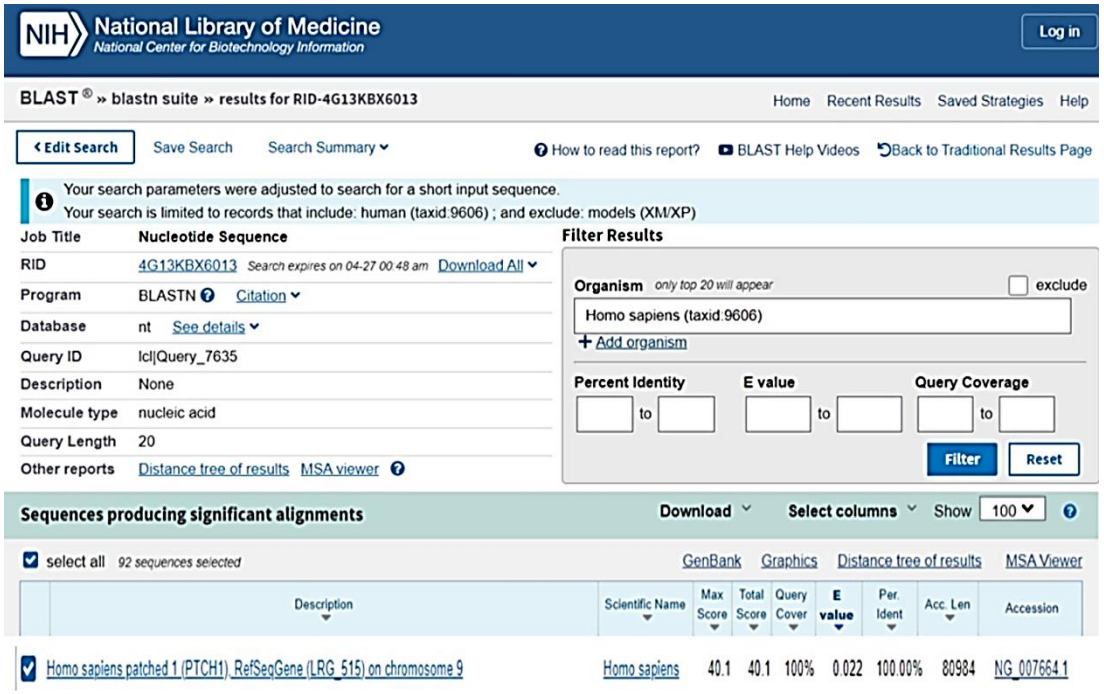


PTCH1: c.1889G>A

Forward primer



Reverse primer



The nucleotide BLAST results indicated that the designed primers exhibited a 100% match with the gene variant. Then, UCSC (University of California, Santa Cruz) *in silico* PCR was done for preliminary assessment of primer design and to verify whether specific genomic regions could be effectively targeted for experimental PCR amplification. This involved specifying the genome assembly (GRCh38), chromosome, and the coordinates of the target region. The tool confirmed that the established forward and reverse primers for the three variants could generate a PCR product, and the expected product sizes are shown in **Figure 19**.

Figure 19: *In silico* PCR

***MRPL13*: c.380T>C**

UCSC In-Silico PCR

>chr8:120419589+120419995 407bp CCTTGTTCAAAAGCAAGCAGTG GCTTTCTTACTATCGCAAGTCAC
CCTTGTTCAAAAGCAAGCAGTGattaaaatttttttttaaaaacttacttt
taatttttagtccaatgatggttcattttacattttaaaatattaataga
attatcatttttttatacatacatattcatttaaaaatacaaaaatatttc
agaagttgacataaaatttttgaaaataagtctgtggcatttatttaaaa
agaaaatgaaacatttaattgaaaattttggaaaagcattgttaccaatg
accactgacttacctcatctggaaaaagatgcaacctttccatcattgtt
cttctgtgaaggtttttttggcagcatgccataaatagctagttttacaat
ctgaaagatatacataggacacaataagaaaattGTGACTTGCGATAGTA
AGAAAGC

***MYBPC3*: c.2816G>A**

UCSC In-Silico PCR

>chr11:47334972+47335287 316bp CACCTCCACTGGACACCAAG ATCAGAGGAGTGGGCAGTGG
CACCTCCACTGGACACCAAGggcctggggtgtcaatggcgggtcttgtga
ctgcacaaaggggcactcacgcaggatctcctgcactgtcaccggctccg
tggtggtaacaggggctccaggccctgccatattgtgtgcccgcactcgg
aaaagcagccgggccccgtgggcaggtccttcaccagtatcgatgtgtg
ctctgtcagcccctgcagggcagccacccactctgagcctgggggtgggg
agggggaggcaaggccacaggctgtgtcaccactgacaccccactcCCAC
TGCCCACTCCTCTGAT

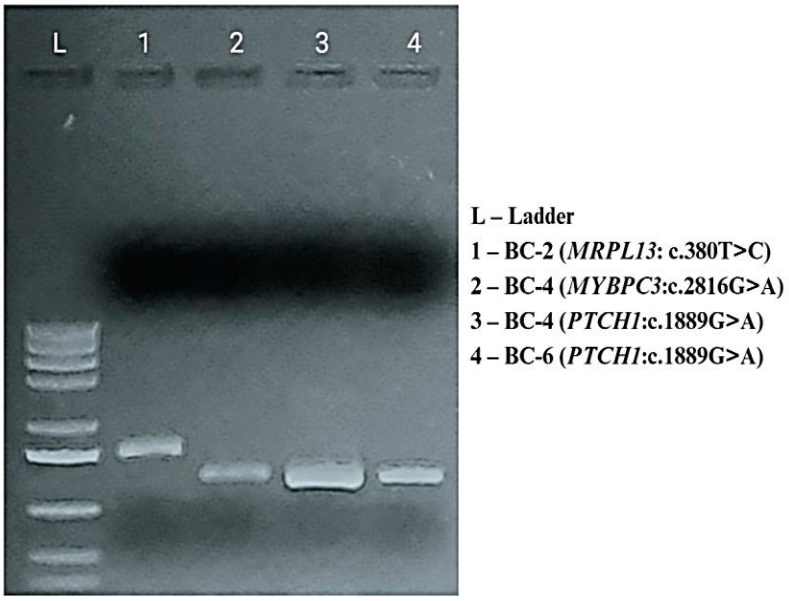
***PTCH1*: c.1889G>A**

UCSC In-Silico PCR

>chr9:95468807+95469115 309bp GTGTCCACTTCGTACAGGGG CCTACACCGACACACACGAC
GTGTCCACTTCGTACAGGGGggctcgaggcagtggaggctggagtcggag
aactgggagagcaggtcccttgtggagctggtgctctctgggctctggca
gctgagggtgtcctgtgtcacggtgacgggctgcacagagatctcggagc
gcggctcagcggtggtgtagtacacgtgcgtgtgggggtcgtactccgtg
cggagctggacagtggactgcatggtaatctgcgtttcatgggcaaagct
gtggctgctgtagggaggtgggggctgtagcgggtattGTCGTGTGTGT
CGGTGTAGG

A touchdown PCR was employed to selectively amplify the variants utilizing gene-specific primers. The utilization of touchdown PCR aimed to enhance both sensitivity and specificity in amplifying the transcript variants, along with optimizing the overall PCR yield (Adamopoulos *et al*., 2021). The touchdown PCR successfully amplified the desired DNA fragments corresponding to the target region of interest. Subsequent gel electrophoresis of the PCR product, compared to the ladder DNA, confirmed that the size of the amplified DNA fragment closely matched the anticipated amplicon size as designed for the primer given in **Figure 20**. This alignment between the amplicon and expected size strongly indicates that the PCR product accurately represents the intended DNA region and is suitable for further Sanger sequencing analysis.

Figure 20: Gel electrophoresis of the PCR product



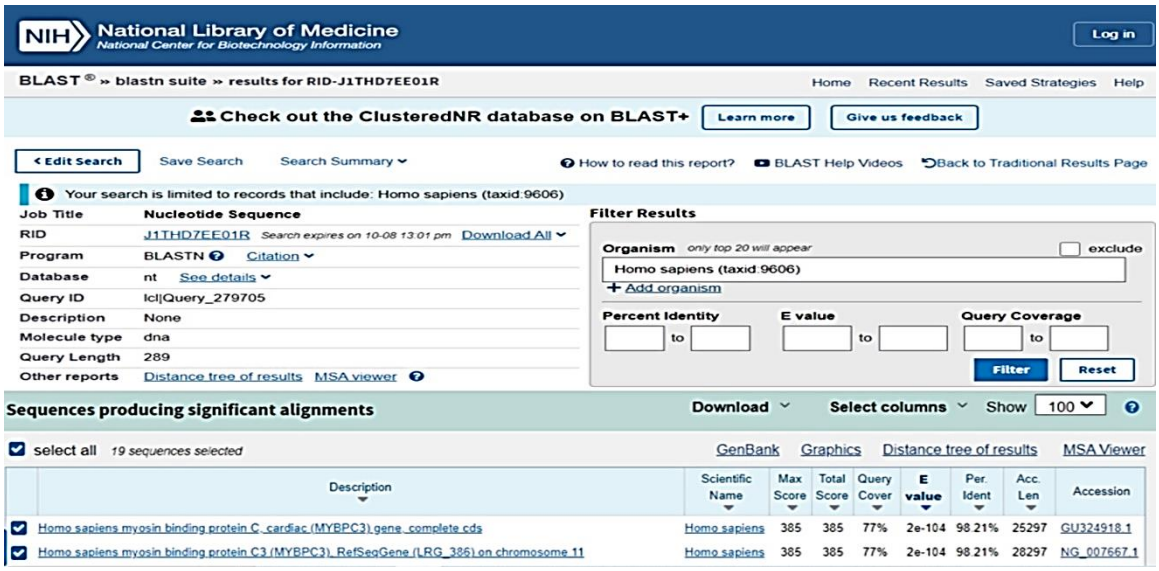
Once the PCR product was retrieved, we proceeded with Sanger sequencing for the newly identified variants in breast cancer patients. It was determined that the *MRPL13*: c.380T>C (p.Leu127pro) variant in BC-2 patient and *PTCH1*: c.1889G>A (p.Arg630His) variant in BC-6 patient was not validated through sequencing, implying a low frequency or absence of variants in the patients. We have successfully validated the *MYBPC3*: c.2816G>A (p.Arg939Gln) and *PTCH1*: c.1889G>A (p.Arg630His) mutations in the BC-4 patient through Sanger sequencing, which confirms the presence of these specific mutations in the tumor DNA. The sequences of *MYBPC3* and *PTCH1*, which

were obtained through Sanger sequencing, were subjected to BLAST analysis to confirm their alignment with the respective genes of interest. The results of the BLAST analysis are presented in **Figure 21**. The sequences aligned with the genes *MRPL13*, *MYBPC3*, and *PTCH1* and confirm the accuracy of the sequencing data.

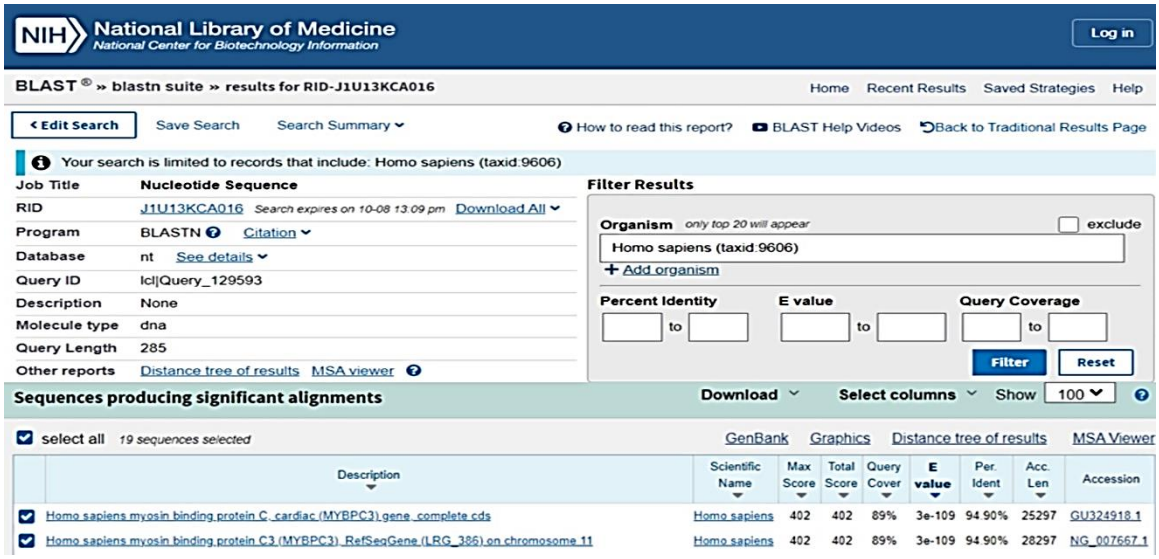
Figure 21: Nucleotide BLAST for Sanger sequence

***MYBPC3*: c.2816G>A**

Forward primer

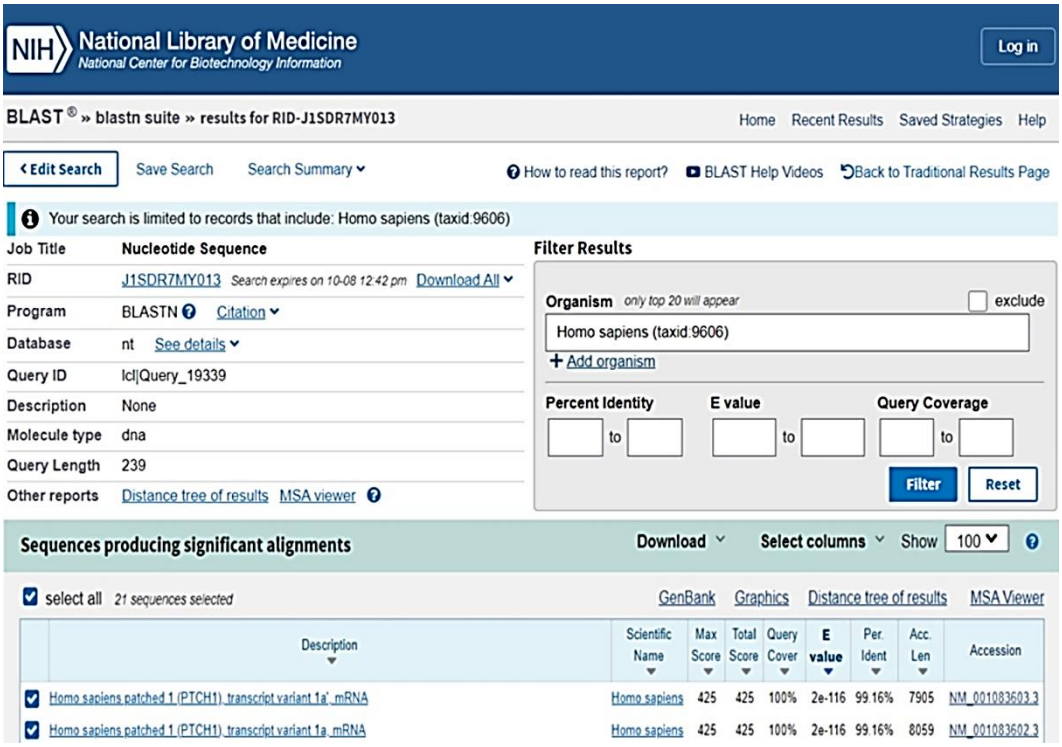


Reverse primer

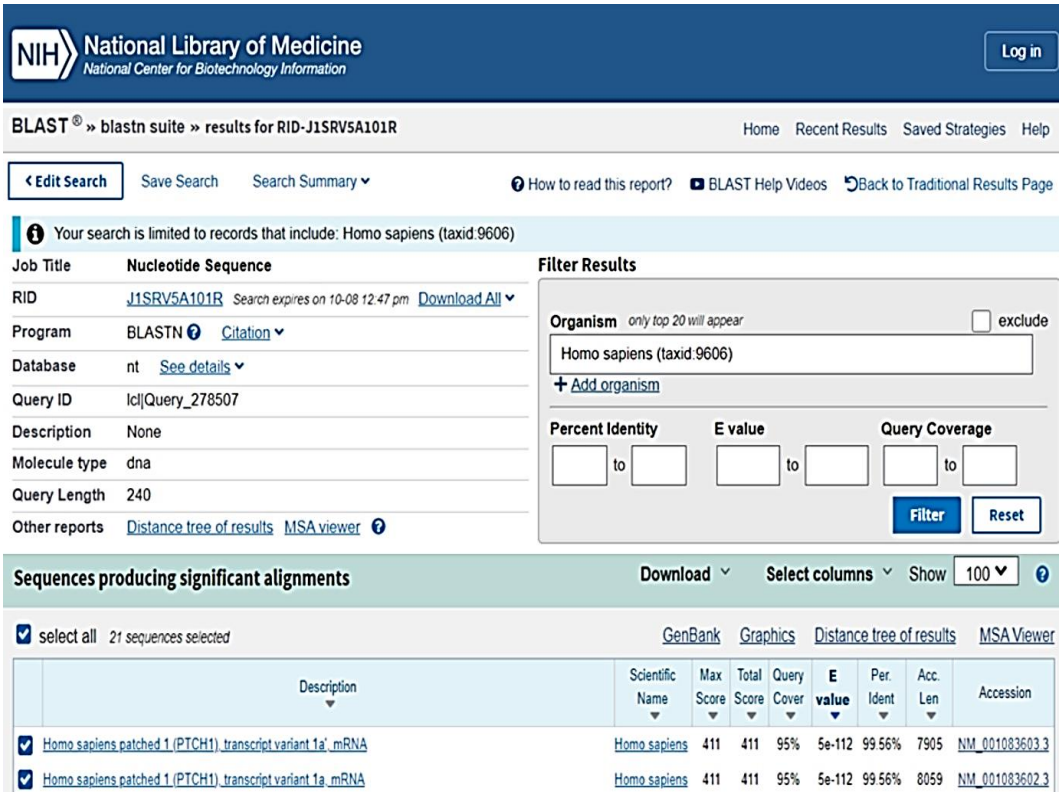


PTCH1: c.1889G>A

Forward primer

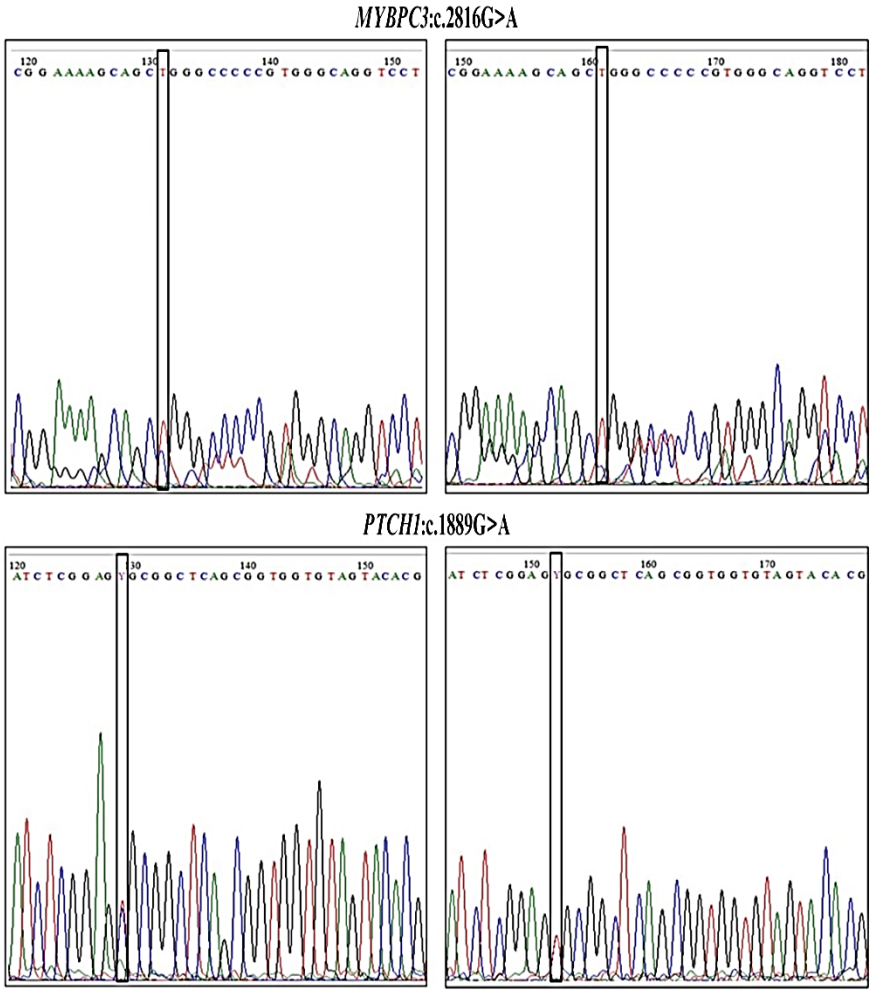


Reverse primer



The *MYBPC3*: c.2816G>A variant in BC-4 patient confirms its presence with a nucleotide change from G to A at position 2816 and an amino acid change from R to Q (Arginine to Glutamine) at position 939. This validation occurred within the forward primer binding site at 131bp and the reverse primer binding site at 161bp. Additionally, the *PTCH1*: c.1889G>A variant in BC-4 patient was also validated, indicating its presence with nucleotide change from G to A at position 1889 and amino acid change from R to H (Arginine to Histidine) at position 630. The variant forward primer was at the 129bp, while the reverse primer binding site was located at the 152bp. The positions of these mutations within primer binding sites were noted, and chromatograms provided visual representations of the sequencing data, showing distinct peaks corresponding to specific nucleotides, which is given in **Figure 22**.

Figure 22: Confirmation of novel variants by Sanger sequencing



The *MYBPC3* gene encodes cardiac myosin-binding protein C (cMyBP-C), a critical protein that regulates the function of both cardiac and skeletal muscles. Adult human muscle contains three MyBP-C isoforms that are closely related: the slow skeletal isoform expressed by the *MYBPC1* gene on chromosome 12q23.3, the fast skeletal isoform produced by the *MYBPC2* gene on chromosome 19q33.3, and the cardiac isoform, cMyBP-C. Notably, cMyBP-C has a unique structure composed of eight immunoglobulin-like and three fibronectin-like domains. Positioned within the cross-bridge-bearing zone (C region) of A bands in striated muscle, cMyBP-C is an optimal platform for orchestrating intracellular signaling processes (Tudurachi *et al*., 2023).

In a healthy heart, sarcomeres are meticulously organized; ensuring muscle fibres are coordinated for contraction and relaxation. MyBP-C plays a pivotal role in this process by interacting with other sarcomeric proteins, finely regulating the timing and strength of muscle contractions. However, mutations in the *MYBPC3* gene can lead to the production of abnormal MyBP-C proteins and reduce the levels of functional MyBP-C. These mutations result in various disruptions, including the disorganization and irregular spacing of sarcomere components, as mutant MyBP-C may not correctly integrate into the sarcomere structure (Martin *et al*., 2022). Furthermore, MyBP-C's regulation of myosin cross-bridges, crucial for generating muscle contractions during each heartbeat, can be compromised by *MYBPC3* mutations, is resulting in hypercontractility of the heart muscle. Additionally, abnormal MyBP-C proteins can stiffen the heart muscle, diminishing its capacity to relax between contractions (Toepfer *et al*., 2019).

Overall, mutations within the *MYBPC3* gene can give rise to a range of cardiac conditions, primarily hypertrophic cardiomyopathy (HCM). This genetic condition is defined by the thickening of the heart muscle, which may hinder the heart's capacity to pump blood efficiently. In severe instances, HCM may lead to complications such as heart failure, arrhythmias, and an elevated risk of sudden cardiac events (Marziliano *et al*., 2021). Despite the significant role of cMyBP-C mutations in HCM, the precise molecular mechanisms underlying these mutations remain a subject of ongoing research and investigation (Tudurachi *et al*., 2023).

The *PTCH1* gene encodes the Patched homolog 1 (*PTCH1*) protein, a critical Hedgehog (Hh) signaling pathway component. This pathway is highly conserved and pivotal in normal embryonic development. It has also been involved in various aspects of mammary gland development, including induction, ductal architecture formation, and lactation differentiation. The disruption of the Hh pathway is closely linked to the onset and progression of breast cancer and other cancers. The *PTCH1* protein is a complex 12-pass transmembrane receptor characterized by two large extracellular loops and two significant intracellular loops (Wang *et al*., 2019). *PTCH1* primary function within the Hh pathway is to act as a tumor suppressor by inhibiting the activity of another transmembrane protein known as Smoothened (SMO). However, *PTCH1* activity is modulated by binding to Hh ligands, which include Sonic Hedgehog (SHH), Indian Hedgehog (IHH), and Desert Hedgehog (DHH) (Bhateja *et al*., 2019). When *PTCH1* binds to these Hh ligands, it undergoes a conformational change, releasing its inhibition on SMO. This event triggers downstream signaling events, resulting in the post-translational processing of *GLI* (glioma-associated oncogene homolog) transcription factors. In mammals, three *GLI* proteins have been identified: *GLI1, GLI2,* and *GLI3*. *GLI1* and *GLI2* generally serve as transcriptional activators, whereas *GLI3* acts as a transcriptional repressor (Jeng *et al*., 2020).

One of the well-known consequences of aberrations in the Hh pathway, such as mutations in *PTCH1*, is the development of oncogenic processes, notably seen in basal cell carcinoma. In addition to genetic mutations, other mechanisms can lead to the aberrant activation of the Hh pathway in various cancers. These include the over-expression of Hh ligands, which can result in autocrine and paracrine signaling. Such dysregulation has been observed in lung cancer, prostate cancer, colorectal cancer, breast cancer, and malignant melanoma, contributing to tumor growth and progression (Nakase *et al*., 2020).

Some genes play dual roles in cellular functions, contributing to cancer development and cardiac health. Specific signaling pathways and regulatory mechanisms may be involved in tumor suppression and cardiac function. Mutations affecting these shared pathways could lead to alterations in both cancer-related and cardiac genes (Libby and Kobold, 2019). Shared environmental factors, such as lifestyle choices or

exposures, could contribute to the development of mutations in both tumor suppressor and cardiac-related genes. Certain risk factors, such as hormonal imbalances or inflammation, can contribute to the occurrence of both breast cancer and heart problems (Gulati and Mulvagh, 2018). Our hospital cohort epidemiology study revealed a heightened risk of breast cancer among premenopausal women, emphasizing the pivotal role of hormonal factors in disease development. Notable associations were observed between breast cancer incidence and lifestyle factors, including exercise frequency, water intake, and sleep duration. Additionally, a prevalence of comorbidities such as high blood pressure and diabetes were identified in a substantial number of patients. These shared risk factors may contribute to the confluence of breast cancer and cardiac issues in affected individuals.

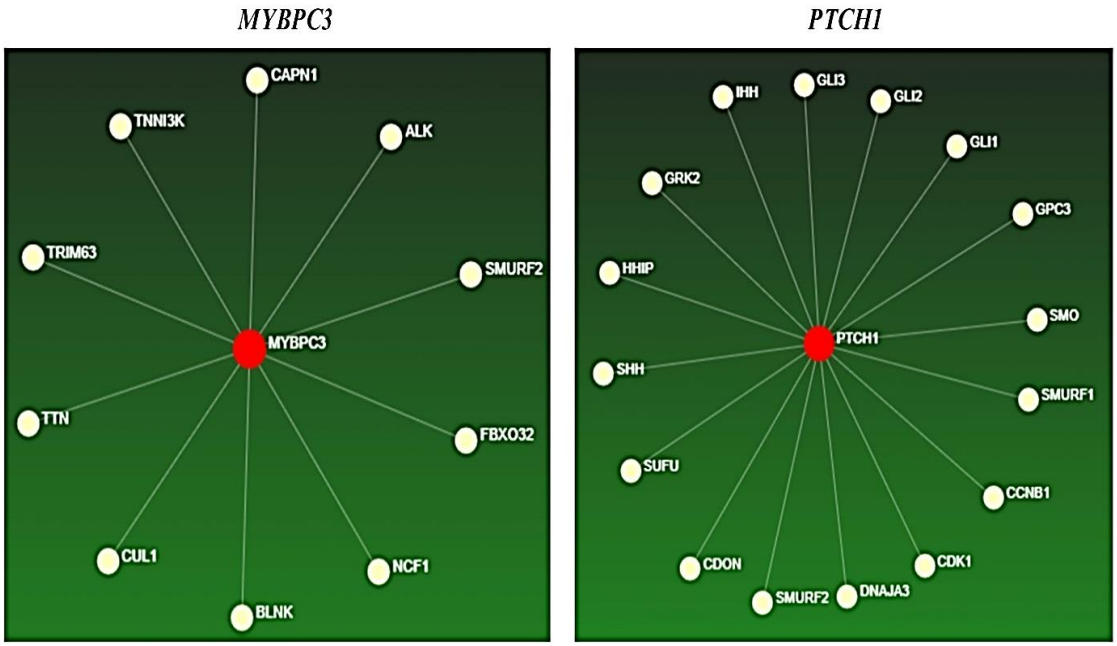
The successful validation of *MYBPC3* and *PTCH1* variants in BC-4 patient was precisely identified and characterized genetic mutations linked to diverse conditions, encompassing cardiac health and potential cancer susceptibility. We are the first to confirm the presence of cardiomyopathy-related and tumor-suppressor gene variants in the primary breast tumor sample. Our findings mark a significant advancement in the field, paving the way for comprehensive and personalized approaches for the diagnosis and treatment strategies for breast cancer patients who also have cardiomyopathy-related genetic alterations.

4.3.2 Interactome network mapping and KEGG pathway analysis

The "Interactome network" mapping of novel candidate genes was done to identify and catalog the interactions among various pathways. KEGG pathway enrichment analysis was also conducted on the genes within the network using Network Analyst. In the network visualization, red nodes represent the novel candidate genes, and interacted genes are depicted as yellow nodes. Specifically, the *MYBPC3* gene interacted with 10 genes: *ALK, FBXO32, NCF1, BLNK, CUL1, CAPN1, TNNI3K, TTN*, *TRIM63*, and *SMURF2*. *PTCH1* gene interacted with 16 genes: *GLI1, SMO, CCNB1, DNAJA3*, *SMURF2, IHH, HHIP, SHH, GRK2, SMURF1, SUFU, CDON*, *GPC3, CDK1*, *GLI3*, and *GLI2* as illustrated in **Figure 23**. The alterations in *MYBPC3* and *PTCH1* can lead to mutations in these interacted genes and disrupt the interacted gene's normal function.

These genetic alterations may contribute to disease progression because the interacted genes are involved in various cellular processes and signaling pathways, which are given in **Table 9**.

Figure 23: Interactome network mapping



The *CUL1* gene shows good interaction with *MYBPC3*, and this interaction is implicated in various cellular pathways, including the hedgehog signaling pathway, TGF-beta signaling pathway, ubiquitin-mediated proteolysis, protein processing in the endoplasmic reticulum, and circadian rhythm.

In the epidemiological study, it was observed that patients with breast cancer experience disturbed sleep cycles, leading to circadian rhythm disruption. Exome analysis too showed that all breast cancer patients suffered from cardiomyopathy, which may also be attributed to one of the effects of circadian rhythm disruption. This was further authenticated by our observation of pathway analysis suggests that the disruption of the *CUL1* gene, potentially due to *MYBPC3* mutation, may lead to disturbances in the circadian rhythm.

Circadian rhythms are vital in regulating genes associated with cardiac metabolism and function. These rhythms coordinate fluctuations in heart rate (HR) and blood pressure (BP), which typically rises during the daytime and decline during the

nighttime. Environmental factors, including mental stress, physical activity, nutrition, temperature, and aging, influence circadian variations in HR and BP. Moreover, the disruptions in these circadian rhythms have been associated with various heart conditions, including hypertrophic cardiomyopathy, heart failure, and stroke (Rabinovich-Nikitin *et al*., 2022). The presence of high blood pressure in the BC-4 patient, along with the observation that a substantial portion of surveyed patients also showed elevated blood pressure levels, underscores the connection between circadian rhythm disturbances and hypertension among the patient cohort.

Circadian rhythms are recognized for their regulatory role in the expression of genes linked to breast cancer, potentially influencing critical cellular processes such as the regulation of the cell cycle, apoptosis, and DNA repair mechanisms. Significantly, disruptions in sleep patterns and circadian rhythms can induce metabolic changes and immune system suppressions, which give rise to a spectrum of health issues, including diabetes, obesity, cardiovascular diseases, and cancer (Lin and Farkas, 2018). In the context of our epidemiology study our patient cohort survey also revealed the prevalence of diabetes. These disturbances in circadian rhythms can profoundly affect the overall health of breast cancer patients, exacerbating multiple health conditions. The candidate genes of six breast cancer patients identified through exome analysis are intricately involved in specific cellular processes, and disruptions in circadian clock function can significantly impact cellular processes, contributing to the hallmark characteristics of cancer. The disturbance in circadian rhythm may, in turn, play a role in the occurrence of both cardiomyopathy and breast cancer.

The interaction of genes associated with *PTCH1* was predominantly involved in the hedgehog signaling pathway. Interestingly, two interacted genes, namely *CUL1* and *SMURF2*, were linked to *MYBPC3*, which also played a significant role in the hedgehog signaling pathway. This pathway holds critical importance in various cellular processes, encompassing developmental regulation, tissue repair, and cell differentiation. Abnormal activation of the hedgehog signaling pathway has been related to the progression of breast cancer. It contributes to several critical aspects of tumorigenesis, including enhanced tumor growth, increased survival of cancer cells, and the facilitation of metastasis. Notably, this pathway can influence cancer stem cell populations, initiating tumors and rendering them resistant to therapeutic interventions (Fattahi *et al*., 2018).

The hedgehog pathway exhibits a capacity for intricate cross-talk with other signaling pathways implicated in cancer, such as the Wnt pathway, Notch pathway, and growth factor signaling. This interplay can further contribute to the complexity of the development and proliferation of cancer. The involvement of the hedgehog pathway in breast cancer is complex and may vary among different subtypes of breast cancer (Bhateja *et al*., 2019). Therefore, hedgehog pathway inhibitors as a therapeutic approach may be tailored to specific patient profiles and tumor characteristics.

The occurrence of cardiomyopathy as a pre-existing condition in primary breast cancer patients is a critical consideration in treatment planning. Understanding this pre-existing cardiac health issue is essential for tailoring breast cancer treatment strategies to minimize the risk of worsening cardiomyopathy. Deciphering the presence of these mutations may play a pivotal role for gaining insights into the mechanisms underlying these health conditions. We will be better equipped to make informed decisions about patient management and pave the way for more effective, tailored, and creative approaches to preventing and treating breast cancer.

Table 9: KEGG pathway analysis

MYBPC3

S.No	Gene	KEGG Pathway
1.	*CUL1*	Hedgehog signaling pathway, TGF-beta signaling pathway, Ubiquitin mediated proteolysis, Protein processing in the endoplasmic reticulum, Circadian rhythm
2.	*SMURF2*	Hedgehog signaling pathway, TGF-beta signaling pathway, Ubiquitin mediated proteolysis
3.	*TTN*	Hypertrophic cardiomyopathy (HCM), Dilated cardiomyopathy
4.	*BLNK*	Osteoclast differentiation, Primary immunodeficiency
5.	*NUF1*	Osteoclast differentiation
6.	*CAPN1*	Protein processing in the endoplasmic reticulum
7.	*ALK*	Non-small cell lung cancer

PTCH1

S.No	Gene	KEGG Pathway
1.	*GLI1*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer, cAMP signaling pathway
2.	*SMO*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer, Proteoglycans in cancer, Axon guidance
3.	*SMURF2*	Hedgehog signaling pathway, Endocytosis, TGF-beta signaling pathway, Ubiquitin mediated proteolysis
4.	*IHH*	Hedgehog signaling pathway, Proteoglycans in cancer
5.	*HHIP*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer, cAMP signaling pathway
6.	*SHH*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer, Proteoglycans in cancer, Axon guidance
7.	*GRK2*	Hedgehog signaling pathway, Endocytosis
8.	*SMURF1*	Hedgehog signaling pathway, Endocytosis, TGF-beta signaling pathway, Ubiquitin mediated proteolysis
9.	*SUFU*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer
10.	*CDON*	Hedgehog signaling pathway
11.	*GLI3*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer, cAMP signaling pathway
12.	*GLI2*	Hedgehog signaling pathway, Basal cell carcinoma, Pathways in cancer
13.	*GPC3*	Proteoglycans in cancer
14.	*CCNB1*	p53 signaling pathway, Progesterone-mediated oocyte maturation, Cell cycle, Oocyte meiosis, Cellular senescence
15.	*CDK1*	p53 signaling pathway, Progesterone-mediated oocyte maturation, Cell cycle, Oocyte meiosis, Cellular senescence

The results of epidemiological analysis, exome sequencing, and Sanger sequencing strongly revealed that risk factors, rare genetic mutations, and alterations in the coding regions of genes may drive cancer development. The mutated genes affect signal transduction, G-protein coupled receptor signaling pathways, gene expression regulation, synapse assembly, transcription regulation, cell cycle regulation, and the regulation of mitosis, which causes uncontrolled cell division. Detecting mutations associated with mitosis can provide insights into the mechanisms underlying abnormal cell division in cancer. The candidate genes' cellular processes are associated with the plasma membrane. During mitosis, the plasma membrane changes to accommodate the division of the cell into two daughter cells. This involves complex rearrangements to distribute the genetic material between the daughter cells properly. The plasma membrane plays a role in controlling the movement of molecules and ions during mitosis. The accumulation of genetic mutations in the plasma membrane transforms normal cells into cancerous ones. The molecular function of candidate genes plays a role in protein binding. Various proteins, such as chromosome condensation, spindle formation, and cytokinesis, orchestrate the cell division events. Dysregulation or mutations in these genes can lead to improper protein interactions, disrupting mitosis. The dysregulation of biological processes and pathways can disrupt the control of cell cycle checkpoints, leading to uncontrolled cell growth and proliferation of cancer cells.

Tumors with a high mitotic index are associated with a worse prognosis, suggesting a higher likelihood of metastasis and poorer patient outcomes. Understanding mitosis and its dysregulation in cancer is crucial for developing targeted therapies and drugs to inhibit aberrant cell division (Alom *et al*., 2020). Integrating whole exome sequencing analysis with histopathological data, such as mitotic index, can provide a more comprehensive understanding of the genetic and cellular factors driving cancer progression. This integration can improve cancer classification, prognosis prediction, and treatment choice.

4.4 AI-driven mitotic cell detection in breast cancer histopathological images

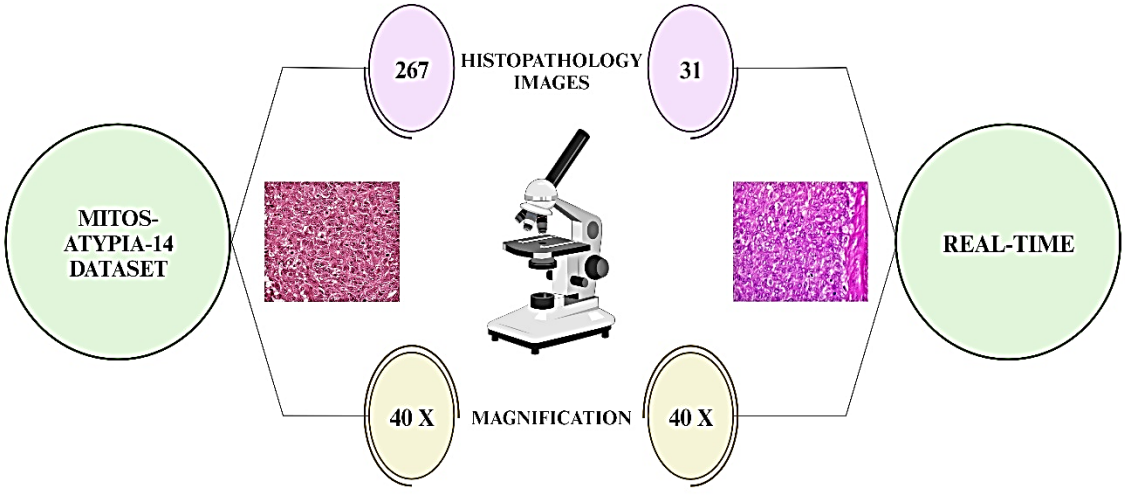
Pathologists manually analyse histopathology images using high-resolution microscopes to identify mitotic cells, which can be difficult due to the small differences

between normal cells and mitosis. Furthermore, this procedure is tedious, time-consuming, and labor-intensive. So, in the hospital, Ki-67 is done to assess the proliferation, which is expensive. AI-based techniques that efficiently detect mitosis in histopathology images were developed to overcome this challenge. Several deep-learning techniques demonstrate low computational cost (Mahmood *et al*., 2020). Using CNN (convolutional neural network) to detect mitosis in histopathology images offers automation and objectivity, which can improve the accuracy of diagnosis. CNNs learn and detect complex patterns and features in images, making them suitable for identifying mitotic figures. Accurate mitosis detection can contribute to the era of personalized medicine. Detecting mitosis is a significant application of artificial intelligence in medical image analysis (Sigirci *et al*., 2022). Hence, we were interested in developing a model to count the abnormal cell division accurately with the help of deep learning.

4.4.1 Enhancing mitosis detection accuracy in histopathology images using CNN (Convolutional Neural Network)

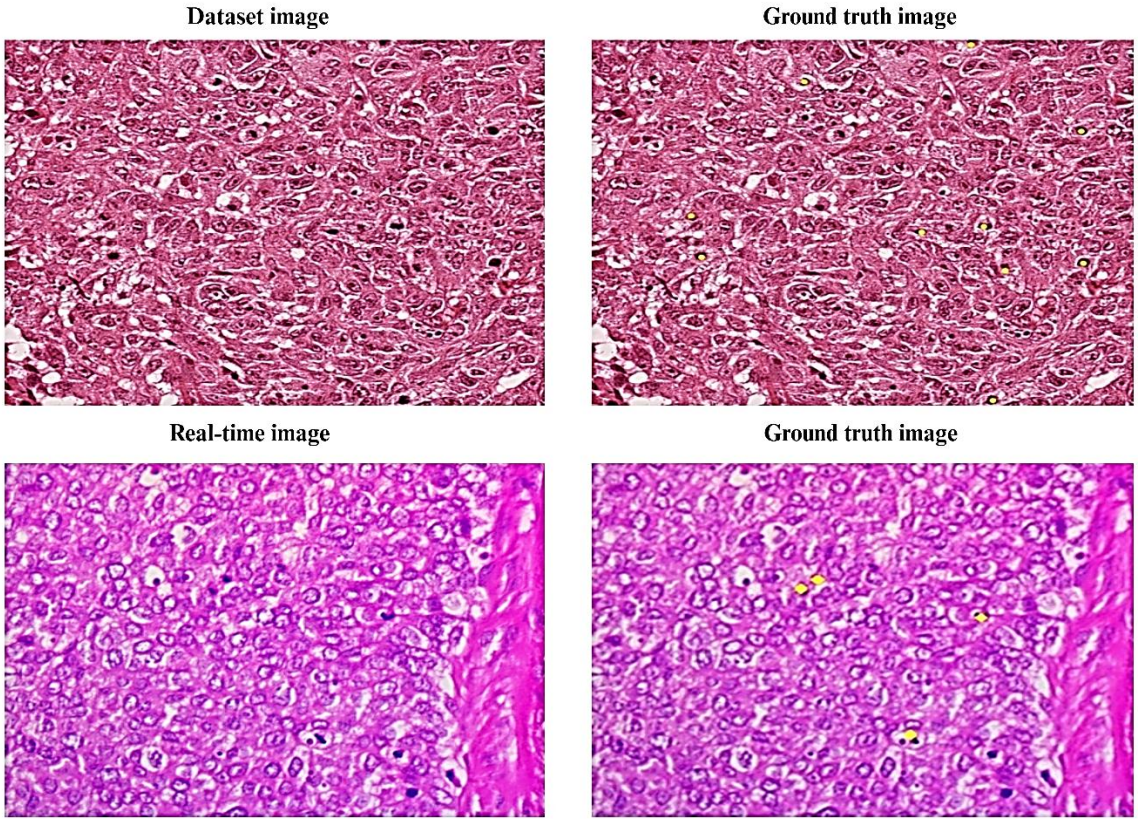
298 histopathology images were used for training and testing the model. 267 images were obtained from the MITOS-ATYPIA-14 dataset, containing 636 mitosis stages, and 31 real-time patient histology images containing 44 mitotic stages were obtained from the Department of Histopathology, Ramakrishna Hospital, Coimbatore. Details regarding the histopathology images used are given in **Figure 24**. The figure was created with BioRender.

Figure 24: Details of the used histopathology images



The identification of mitosis was facilitated by utilizing the ground truth images. **Figure 25** depicts the breast cancer histopathology images of both dataset and real-time patient images.

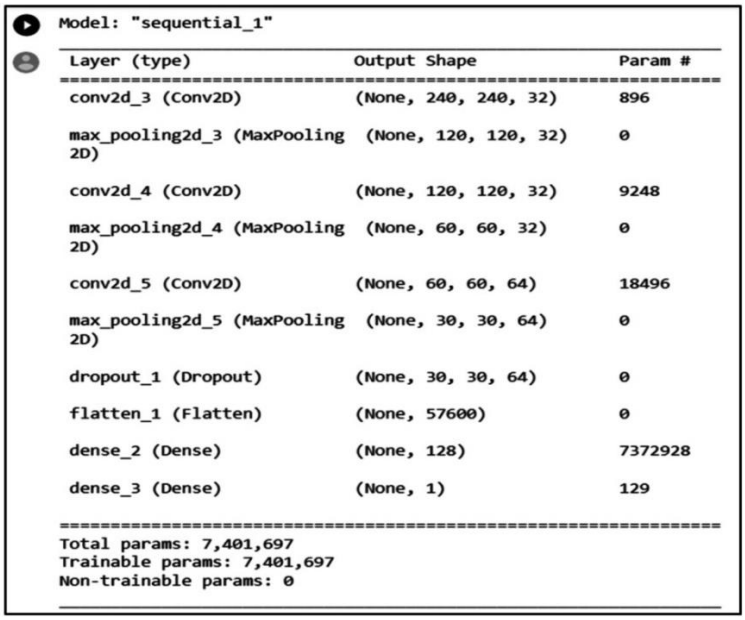
Figure 25: Histopathology images of breast cancer



Neural network architecture was designed for binary classification in breast cancer histopathology images. A Sequential model was created, incorporating Convolutional layers (Conv2D) with ReLU activation and max-pooling layers (MaxPool2D). The first layer, conv2d_3 (Conv2D), was configured as a convolutional layer with 32 filters, each measuring 3x3. It could process input data with None, 240, 240, 32, signifying its ability to handle images sized at 240x240 with 32 channels, which could represent color channels or feature maps. Following conv2d_3, max_pooling2d_3 (MaxPooling2D) was utilized to decrease the dimensions of the feature maps by selecting the maximum value within each 2x2 region. This operation transformed the output shape to None, 120, 120, 32.

Subsequently, conv2d_4 (Conv2D) was introduced as another convolutional layer equipped with 32 filters of size 3x3. This layer received input from the previous layer, resulting in an output shape of None, 120, 120, 32. After convolutional layer conv2d_4, max_pooling2d_4 (MaxPooling2D) was applied, further reducing the dimensions to None, 60, 60, 32. Next, conv2d_5 (Conv2D) was deployed as a subsequent convolutional layer featuring 64 filters with dimensions of 3x3. The outcome of this layer was an output shape of None, 60, 60, 64. Continuing with the architectural pattern, max_pooling2d_5 (MaxPooling2D) was utilized, transforming the dimensions to None, 30, 30, 64. A dropout layer (dropout_1) was incorporated to address the risk of overfitting. During training, this layer randomly sets a fraction of input units to 0 in each update.

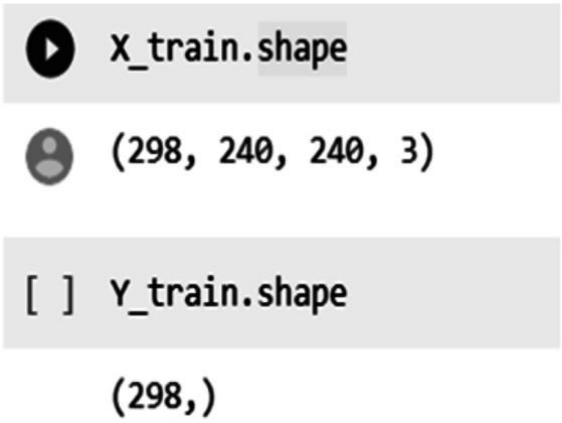
Figure 26: Output during the training phase



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_3 (Conv2D)           (None, 240, 240, 32)      896

 max_pooling2d_3 (MaxPooling  (None, 120, 120, 32)     0
 2D)

 conv2d_4 (Conv2D)           (None, 120, 120, 32)      9248

 max_pooling2d_4 (MaxPooling  (None, 60, 60, 32)       0
 2D)

 conv2d_5 (Conv2D)           (None, 60, 60, 64)        18496

 max_pooling2d_5 (MaxPooling  (None, 30, 30, 64)       0
 2D)

 dropout_1 (Dropout)         (None, 30, 30, 64)        0

 flatten_1 (Flatten)         (None, 57600)             0

 dense_2 (Dense)             (None, 128)               7372928

 dense_3 (Dense)             (None, 1)                 129

=================================================================
Total params: 7,401,697
Trainable params: 7,401,697
Non-trainable params: 0
_________________________________________________________________
```

Flatten_1 (Flatten) layer reshapes the 3D tensor into a 1D vector, resulting in a shape of None, 57600. This flattened vector was a suitable input for the subsequent fully connected layers. In line with this architectural progression, dense_2 (Dense) marked the integration of a fully connected (dense) layer with 128 units or neurons. The architecture culminated with dense_3 (Dense), a dense output layer featuring a single unit. This layer was designed for binary classification tasks. These architectural components collectively constituted the neural network model, contributing to its capacity for processing and classifying data. The parameters trained for the sequential model is illustrated in **Figure 26**.

Figure 27: Training of CNN model




```
history = model.fit(x_train,y_train,epochs = 25, validation_data = (x_val, y_val))

Epoch 1/25
8/8 [==============================] - 25s 3s/step - loss: 1.5091 - accuracy: 0.5630 - val_loss: 0.6422 - val_accuracy: 0.6000
Epoch 2/25
8/8 [==============================] - 20s 3s/step - loss: 0.7065 - accuracy: 0.4916 - val_loss: 0.6807 - val_accuracy: 0.6000
Epoch 3/25
8/8 [==============================] - 18s 2s/step - loss: 0.6841 - accuracy: 0.4916 - val_loss: 0.6468 - val_accuracy: 0.6000
Epoch 4/25
8/8 [==============================] - 18s 2s/step - loss: 0.6801 - accuracy: 0.5252 - val_loss: 0.6363 - val_accuracy: 0.7667
Epoch 5/25
8/8 [==============================] - 20s 3s/step - loss: 0.6339 - accuracy: 0.6681 - val_loss: 0.5552 - val_accuracy: 0.6000
Epoch 6/25
8/8 [==============================] - 18s 2s/step - loss: 0.5913 - accuracy: 0.6513 - val_loss: 0.4784 - val_accuracy: 0.8500
Epoch 7/25
8/8 [==============================] - 21s 3s/step - loss: 0.5159 - accuracy: 0.7395 - val_loss: 0.3946 - val_accuracy: 0.8667
Epoch 8/25
8/8 [==============================] - 18s 2s/step - loss: 0.4407 - accuracy: 0.8235 - val_loss: 0.4429 - val_accuracy: 0.8000
Epoch 9/25
8/8 [==============================] - 19s 2s/step - loss: 0.3566 - accuracy: 0.8782 - val_loss: 0.2753 - val_accuracy: 0.9000
Epoch 10/25
8/8 [==============================] - 20s 3s/step - loss: 0.3046 - accuracy: 0.8739 - val_loss: 0.1952 - val_accuracy: 0.9500
Epoch 11/25
8/8 [==============================] - 19s 2s/step - loss: 0.2880 - accuracy: 0.8613 - val_loss: 0.1914 - val_accuracy: 0.9333
Epoch 12/25
8/8 [==============================] - 21s 3s/step - loss: 0.2028 - accuracy: 0.9328 - val_loss: 0.1645 - val_accuracy: 0.9667
Epoch 13/25
8/8 [==============================] - 18s 2s/step - loss: 0.1749 - accuracy: 0.9328 - val_loss: 0.1724 - val_accuracy: 0.9333
Epoch 14/25
8/8 [==============================] - 20s 2s/step - loss: 0.1519 - accuracy: 0.9580 - val_loss: 0.1315 - val_accuracy: 0.9667
Epoch 15/25
8/8 [==============================] - 19s 2s/step - loss: 0.1552 - accuracy: 0.9286 - val_loss: 0.1383 - val_accuracy: 0.9500
Epoch 16/25
8/8 [==============================] - 18s 2s/step - loss: 0.1482 - accuracy: 0.9328 - val_loss: 0.1607 - val_accuracy: 0.9000
Epoch 17/25
8/8 [==============================] - 19s 2s/step - loss: 0.1877 - accuracy: 0.9118 - val_loss: 0.5310 - val_accuracy: 0.7833
Epoch 18/25
8/8 [==============================] - 18s 2s/step - loss: 0.2786 - accuracy: 0.8782 - val_loss: 0.1700 - val_accuracy: 0.9000
Epoch 19/25
8/8 [==============================] - 21s 3s/step - loss: 0.3446 - accuracy: 0.8403 - val_loss: 0.4002 - val_accuracy: 0.8667
Epoch 20/25
8/8 [==============================] - 18s 2s/step - loss: 0.2536 - accuracy: 0.9034 - val_loss: 0.2098 - val_accuracy: 0.9000
Epoch 21/25
8/8 [==============================] - 19s 2s/step - loss: 0.2065 - accuracy: 0.8950 - val_loss: 0.2372 - val_accuracy: 0.8667
Epoch 22/25
8/8 [==============================] - 20s 2s/step - loss: 0.1989 - accuracy: 0.8866 - val_loss: 0.1730 - val_accuracy: 0.8833
Epoch 23/25
8/8 [==============================] - 19s 2s/step - loss: 0.1419 - accuracy: 0.9496 - val_loss: 0.3414 - val_accuracy: 0.8500
Epoch 24/25
8/8 [==============================] - 20s 3s/step - loss: 0.1450 - accuracy: 0.9370 - val_loss: 0.1193 - val_accuracy: 0.9500
Epoch 25/25
8/8 [==============================] - 18s 2s/step - loss: 0.1166 - accuracy: 0.9664 - val_loss: 0.1068 - val_accuracy: 0.9667
```

The model was compiled using the Adam optimizer, binary cross-entropy loss function, and accuracy metric. The training process involved providing the training data (x_train, y_train) and specifying 25 epochs. Validation data (x_val, y_val) was included to evaluate the model's performance during training. **Figure 27** displays the training of the CNN model.

The model achieved a high accuracy of 0.96, which suggests it has successfully learned to recognize and generalize patterns present within the training data. The selected CNN architecture, featuring convolutional layers and max-pooling, proved effective in learning hierarchical features from the images. Integrating a dropout layer prevented overfitting by reducing the risk of neuron co-adaptation during training. The ample training data enabled the model to acquire diverse features and generalize effectively to various samples. Using the Adam optimizer with an appropriately set learning rate facilitated efficient model convergence. Additionally, two plots were generated to analyze the progress of training.

Figure 28: Loss analysis of the CNN model

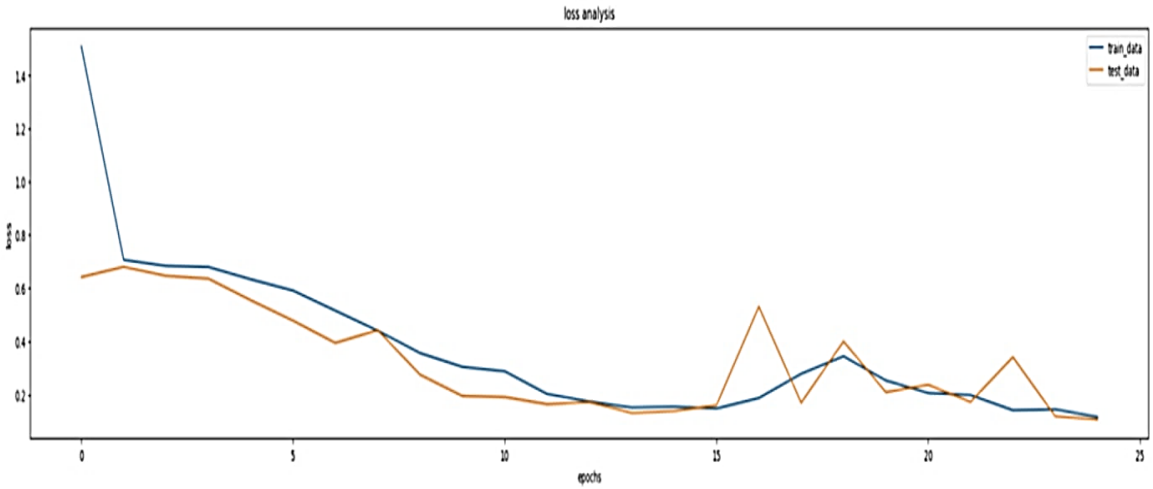


Figure 28 shows the change in loss (binary cross-entropy) over epochs for both the training and testing data. The training loss and testing loss exhibited a decreasing trend and followed similar patterns, suggesting that the model is learning effectively and generalizing well. This means the model is balanced and can make accurate predictions based on data.

Figure 29: CNN model accuracy improvement

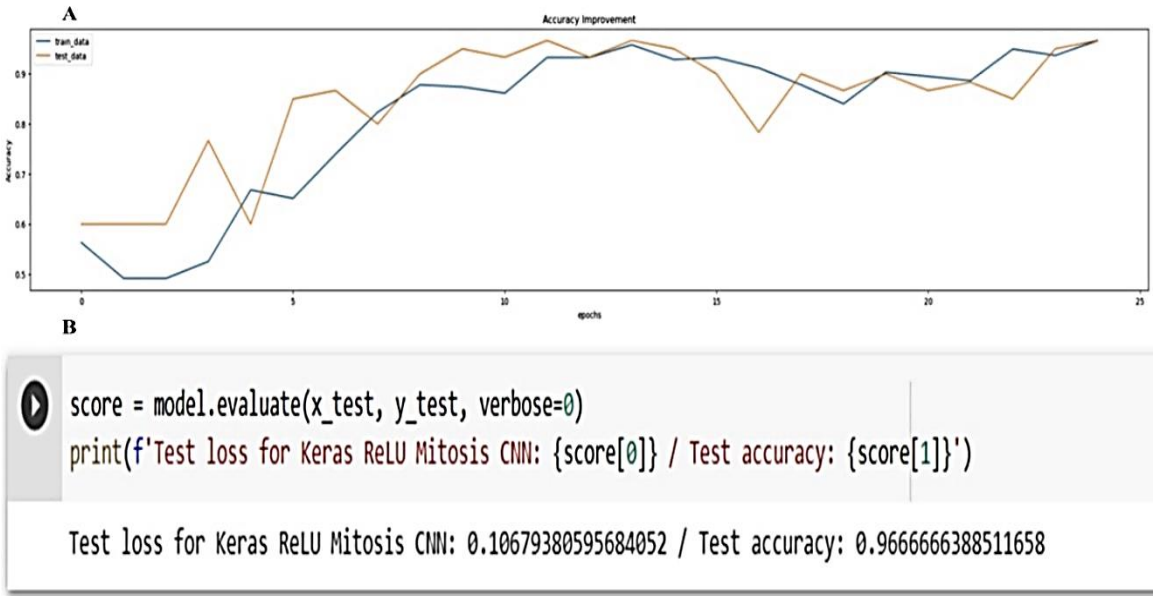
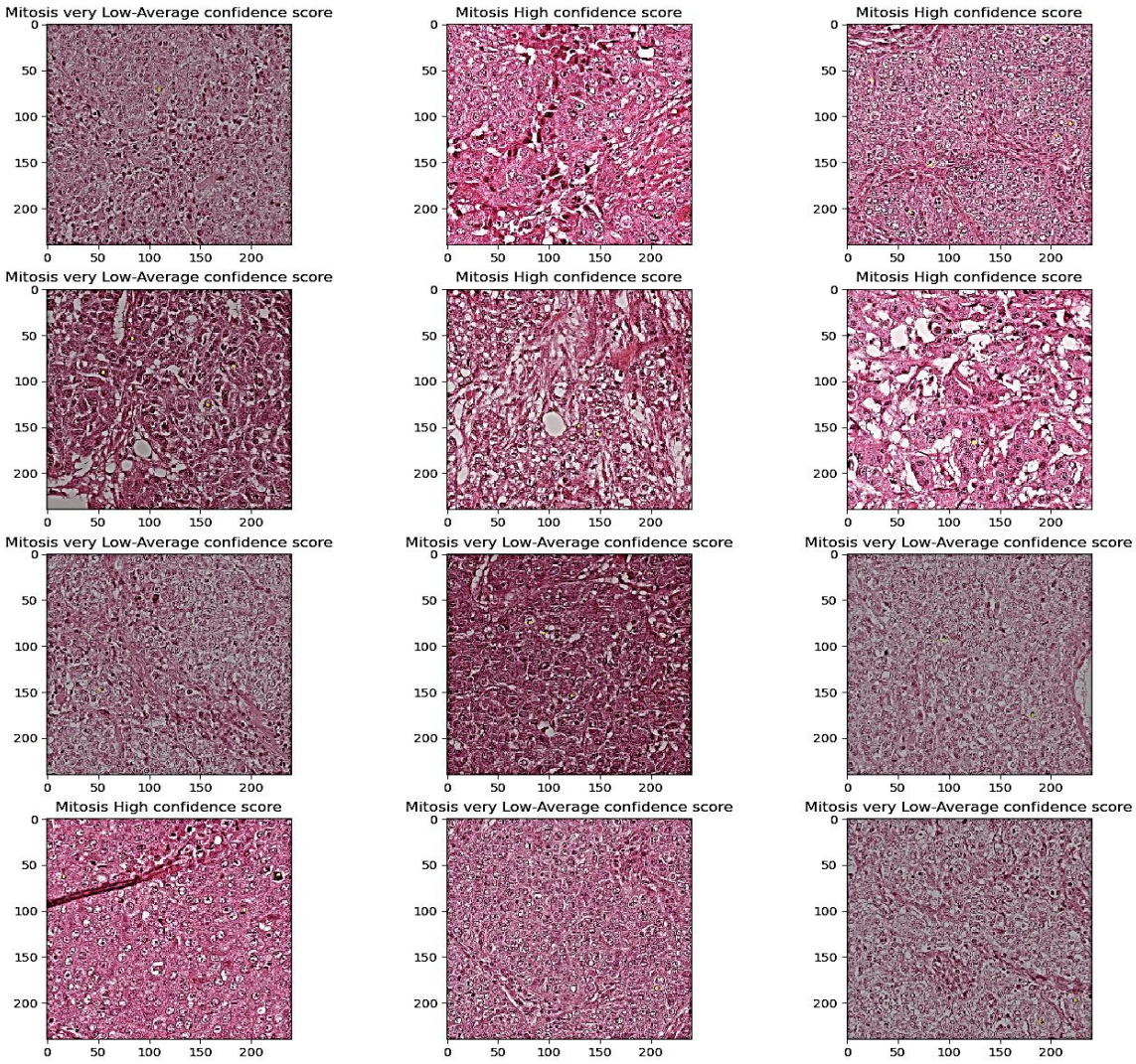


Figure 29 shows the change in accuracy over epochs for both the training and testing data. The training and validation data is increasing and closely tracked, suggesting that the model is improving its ability to correctly classify mitosis in histopathology image data. Again, this indicates good generalization and learning. The successful convergence of both loss and accuracy for training and validation data suggests that the CNN model effectively learns the underlying patterns in histopathology images.

The accuracy of 0.96 (96%) confirms that the model is making accurate predictions on the testing data, and the detection is visualized in **Figure 30**. The developed code loads and pre-processes histopathology images and categorizes them based on mitosis scores. The constructed sequential CNN model for mitosis detection effectively trains the model and analyses the training process using loss and accuracy graphs. The developed CNN model using real-time images helps pathologists provide improved diagnoses and treatment planning, and it is also cost-effective and less time-consuming.

Figure 30: Mitosis detection



The proposed model was compared with already reported works, and our results demonstrated more effective mitosis detection, which is summarized in **Table 10**.

Recent advancements in deep-learning techniques, particularly the integration of Faster R-CNN and deep CNNs, have significantly narrowed the gap between human experts and computers in detecting mitotic cells during various stages of cell division. Demonstrating robust performance on dataset ICPR 2012, leverages Faster R-CNN and Resnet-50 for feature extraction and exhibits promising generalization capabilities for detecting lesions in histology images. The technique effectively accommodates mitotic

cell size variations with the feature-extraction network and anchor-box selection within Faster R-CNN. Using Resnet-50, which preserves mitotic cell information through skip connections, an accuracy of 0.858 was achieved. This research holds practical value for pathologists and cancer researchers working with histopathological images (Mahmood *et al*., 2020).

Table 10: Overview of the existing research work of mitosis detection in breast cancer histopathology images

Author	Year	Doi	Accuracy
Mahmood *et al*.	2020	10.3390/jcm9030749	0.858
Sigirci *et al*.	2022	https://doi.org/10.1007/s11042-021-10539-2	0.869
Alom *et al*.	2020	10.1109/ACCESS.2020.2983995	0.878
Rao,	2018	https://doi.org/10.48550/arXiv.1807.01788	0.955
Lei *et al*.	2020	10.1109/JBHI.2020.3027566	0.851
Developed method	**2023**		**0.966**

Distinct morphological changes in cellular structures result in considerable disparities between mitosis and non-mitotic cells in histopathological images. During the mitotic stage, the nuclear membrane breaks, and chromatin strands move to the cell's core area. These changes also lead to distinct variations in pixel density. Mitotic and non-mitotic cell classification can benefit from statistical, shape-based, and textural descriptors. Recent trends incorporate deep learning-based methods for feature extraction along with traditional algorithms. Results were assessed using the ICPR 2014 dataset, containing approximately 748 mitotic cells from 10 histopathological slides. The deep learning technique obtained a high F-measure score (86.97%) (Sigirci *et al*., 2022).

Accurate detection and counting of mitotic cells significantly impact breast cancer prognosis. There's a high demand for automated mitotic cell detection methods in clinical settings. MitosisNet, an end-to-end multitask learning system, achieves this criterion by including segmentation, detection, and classification models. The system uses

segmentation and detection models to locate mitotic reference regions, followed by a classification model to validate these regions. Evaluation of MITOSIS 2012, MITOSIS 2014, and Case Western Reserve University (CWRU) datasets demonstrated peak accuracy with F-measure scores reaching approximately 87.8% (Alom *et al*., 2020).

Rao (2018) introduced an innovative adaptation of the Faster R-CNN architecture designed specifically for efficient and precise detection of mitotic figures in breast cancer histopathological images. This novel two-stage top-down multi-scale region proposal generation process demonstrates superior performance in mitotic figure detection. With an exceptional F-measure score of 0.955, this approach highlights the superiority of learned features over manually crafted ones in detecting mitotic figures.

A novel approach employs an attention-guided multi-branch convolutional neural network to identify potential mitotic cells in histological sections for automatic screening. Deep convolutional neural networks extract high-level mitotic features, further refined using spatial attention modules to improve feature learning. Multi-branch classification subnets are employed to screen mitotic candidates. Remarkably, this approach yields exceptional results on the ICPR 2012 mitosis detection dataset, achieving an impressive F1 score of 0.851 (Lei *et al*., 2020).

The developed deep learning technique has dramatically improved the accuracy and efficiency of mitosis detection in histopathological images. Various approaches have been explored, combining convolutional neural networks (CNNs) with innovative methods to tackle the challenges posed by mitotic cell detection. The results of the developed method showed consistent improvement and higher accuracy compared to other similar works reported. Only the dataset images were used to train the model in all the similar works reported. Still, in our model, we used both MITOS-ATYPIA-14 dataset and real-time patient images, which shows the significance of our approach using deep learning. Detecting more abnormal cells in real-time images underscores the critical role of abnormal cell division in cancer development. Understanding the molecular alterations associated with mitosis provides targets for therapeutic intervention and potential biomarkers for diagnosis and prognosis. Targeting specific molecular pathways involved

in aberrant mitotic processes may help to develop more effective treatments that disrupt cancer cell division and inhibit tumor progression.

By tailoring treatments to each patient's risk factors, genetic alterations, and disease progression, personalized medicine ensures that interventions are precisely matched to the specific needs of the individual. This approach minimizes the risk of adverse effects and maximizes treatment efficacy, ultimately contributes to better outcomes and an enhanced quality of life for breast cancer patients.

The findings of the present study are summarized, and the conclusions from the results are outlined in the next chapter.