MEDICAL UNIVERSITY OF GDANSK

FACULTY OF PHARMACY

Molecular Classification of Uninvolved Mammary Tissue: A Potential Tool for Early Breast Cancer Diagnosis and Recurrence Risk Assessment

Molekularna klasyfikacja niezajętej tkanki sutka: potencjalne narzędzie do wczesnej diagnostyki raka piersi i oceny ryzyka wznowy

Maria Andreou, M.Sc.

Ph.D. supervisor:

Prof. dr hab. Arkadiusz Piotrowski

Doctoral dissertation conducted at 3P-Medicine Laboratory, under the International Research Agenda program granted by the Foundation for Polish Science at the Medical University of Gdansk

GDAŃSK 2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Arkadiusz Piotrowski, for their unwavering support, inspiration, and invaluable guidance throughout every stage of this work. Your profound knowledge and dedication to science have been a tremendous source of motivation for me. Your mentorship has not only shaped this research but also played a crucial role in my development as a scientist.

I am sincerely grateful to Dr. Natalia Filipowicz for her steadfast support both in and out of the lab and for her insightful and thorough reviews of my work.

My heartfelt appreciation extends to the staff at the 3P-Medicine Laboratory at the Medical University of Gdańsk for their indispensable assistance, and expertise, and for fostering such a welcoming and collaborative environment.

I would especially like to thank Dr. Marcin Jąkalski for the fruitful cooperation and exchange of ideas, which greatly enriched this project. I am particularly thankful for the opportunity to collaborate to build a paper together—a rewarding and inspiring experience I deeply value.

I am also grateful to Prof. Jan P. Dumanski, Dr. Jakub Mieczkowski, and Dr. Magdalena Koczkowska for their valuable reviews and insights into this work, significantly enhancing its quality.

To my colleagues, and most importantly my friends, Katarzyna Bisewska, Dr. Katarzyna Chojnowska, Katarzyna Duzowska, Dr. Ulana Juhas, Dr. Anna Kostecka, Urszula Ławrynowicz, Dr. Magdalena Wójcik and Ayse Yigit: thank you for the endless conversations, unwavering support, and for simply being there when it mattered most.

A special thank you goes to my family for their unconditional love, encouragement, and belief in me through every challenge and success. To my dearest friends Chrysanthi, Elisavet and Petros, thank you for standing by me, supporting me throughout this journey, and always lifting my spirits.

Lastly, I am profoundly grateful to all the sample donors and clinicians whose invaluable contributions made this research possible. Without their involvement, this work could not have been accomplished.

This work was supported by the Foundation for Polish Science under the International Research Agenda Program, financed by the Smart Growth Operational Programme 2014 2020 (Grant Agreement No. MAB/2018/6).

This thesis is based on the papers listed below and unpublished findings (manuscript under review, preprint). The listed papers are referred to within the text by Roman numerals I, II, and III.

Paper I:

Filipowicz N., Drężek K., Horbacz M., Wojdak A., Szymanowski J., Rychlicka-Buniowska E., Juhas U., Duzowska K., Nowikiewicz T., Stańkowska W., Chojnowska K., **Andreou M.**, Ławrynowicz U., Wójcik M., Davies H., Śrutek E., Bieńkowski M., Milian-Ciesielska K., Zdrenka M., Ambicka A., Przewoźnik M., Harazin-Lechowska A., Adamczyk A., Kowalski J., Bała D., Wiśniewski D., Tkaczyński K., Kamecki K., Drzewiecka M., Wroński P., Siekiera J., Ratnicka I., Jankau J., Wierzba K., Skokowski J., Połom K., Przydacz M., Bełch Ł., Chłosta P., Matuszewski M., Okoń K., Rostkowska O., Hellmann A., Sasim K., Remiszewski P., Sierżęga M., Hać S., Kobiela J., Kaska Ł., Jankowski M., Hodorowicz-Zaniewska D., Jaszczyński J., Zegarski W., Makarewicz W., Pęksa R., Szpor J., Ryś J., Szylberg Ł., Piotrowski A., Dumanski J. P. Comprehensive cancer-oriented biobanking resource of human samples for studies of post zygotic genetic variation involved in cancer predisposition. Plos one. 2022 Apr 7;17(4):e0266111 doi: 10.1371/journal.pone.0266111. Impact Factor: 3.700. MNiSW score: 100.000.**Paper II:**

Kostecka A., Nowikiewicz T., Olszewski P., Koczkowska M., Horbacz M., Heinzl M., **Andreou M.**, Salazar R., Mair T., Madanecki P., Gucwa M., Davies H., Skokowski J., Buckley P. G., Pęksa R., Śrutek E., Szylberg Ł., Hartman J., Jankowski M., Zegarski W., Tiemann – Boege I, Dumanski J. P., Piotrowski A. High prevalence of somatic *PIK3CA* and *TP53* pathogenic variants in the normal mammary gland tissue of sporadic breast cancer patients revealed by duplex sequencing. NPJ breast cancer, 2022 8(1), 76. doi: 10.1038/s41523-022-00443-9. Impact Factor: 5.900. MNiSW score: 40.000.

Paper III:

Andreou M., Jąkalski M., Duzowska K., Filipowicz N., Kostecka A., Davies H., Horbacz M., Ławrynowicz U., Chojnowska K., Bruhn-Olszewska B., Jankau J., Śrutek E., Las-Jankowska M., Bała D., Hoffman J., Hartman J., Pęksa R., Skokowski J., Jankowski M., Szylberg Ł., Maniewski M., Zegarski W., Nowikiewicz M., Nowikiewicz T., Dumanski J.P., Mieczkowski K., Piotrowski A. Prelude to malignancy: A gene expression signature in normal mammary gland from breast cancer patients suggests pre-tumorous alterations and is associated with adverse outcomes. International journal of cancer, 2024 155(9), 1616–1628. doi:

10.1002/ijc.35050. Impact Factor: 5.700. MNiSW score: 140.000. *Maria Andreou and Marcin Jąkalski have contributed equally to this study.*

Unpublished findings, manuscript under review (preprint):

Andreou M., Chojnowska K., Filipowicz N., Horbacz M., Madanecki P., Duzowska K., Ławrynowicz U., Davies H., Bruhn-Olszewska B., Koszyński M., Drężek-Chyła K., Jaśkiewicz M., Jąkalski M., Kostecka A., Drzewiecka-Kłysz M., Nowikiewicz M., Las-Jankowska M., Bała D., Hoffman J., Śrutek E., Jankowski M., Jankau J., Hodorowicz-Zaniewska D., Szpor J., Szylberg Ł., Zegarski W., Nowikiewicz T., Buckley P.G., Tiemann-Boege I., Mieczkowski J., Koczkowska M., Dumanski J.P., Piotrowski A. Beyond Tumors: Reduced survival linked to pathogenic *PIK3CA* and *TP53* post-zygotic variants in the uninvolved breast tissue of recurrent cancer patients. medRxiv 2024.10.04.24313634; doi: https://doi.org/10.1101/2024.10.04.24313634.

TABLE OF CONTENTS

SUMMARY IN ENGLISH

This doctoral thesis investigates the molecular mechanisms underlying breast cancer progression, focusing on genetic and transcriptomic alterations in histologically normal mammary tissues. Unlike prior studies that focus primarily on tumors, this work examines non-tumorous tissues to uncover early molecular changes that could serve as preclinical indicators of breast cancer or predictors of recurrence risk. We analyzed data from 184 breast cancer patients, recruited either with or without criteria related to prognosis, and 94 control individuals undergoing reduction mammoplasty. Samples collected included 267 uninvolved margin tissues at varying distances from the primary lesions, 184 skin or whole blood samples as reference, and 184 primary tumors, originating from cancer patients (Papers II, III, and manuscript under review). In addition to uninvolved margin tissues, 41 skin or whole blood samples were collected from control individuals to investigate somatic mosaicism (Papers II and manuscript under review).

The research is grounded in the concept of field cancerization, which suggests that ostensibly normal tissues surrounding tumors may harbor genetic and transcriptomic changes that predispose them to malignancy. Leveraging cutting-edge techniques, we investigated structural chromosomal alterations, post-zygotic variants, and transcriptomic changes, with high sensitivity and specificity.

The work began with establishing a comprehensive biobank of mammary tissues, which addressed critical challenges such as tissue heterogeneity and the difficulty of obtaining matched control samples. Rigorous histological validation, ensured accurate classification of non-tumorous tissues, distinguishing them from micrometastases or other malignant regions. This biobank provided a high-quality and reliable foundation for downstream analyses, overcoming variability in sample quality (Paper I).

The next phase focused on the genetic profiling of non-tumorous tissues. Employing single nucleotide polymorphism arrays, whole exome sequencing, and ultra-sensitive duplex sequencing, we identified structural chromosomal alterations and low-frequency pathogenic post-zygotic variants, such as in *AKT1, PIK3CA, and TP53* which are typically associated with cancer but were also detected in histologically normal tissues. These findings challenge the conventional understanding of non-tumorous tissues as passive bystanders, highlighting their active role in early tumorigenic processes (Paper II).

Building on these genetic insights, RNA-seq-based transcriptomic profiling was performed to identify gene expression patterns in non-tumorous tissues at various distances from the primary lesions of breast cancer patients with adverse outcomes. Advanced bioinformatics tools, including pathway enrichment and survival analyses, revealed a distinct molecular signature involving keratins, adhesion proteins, oncogenes, and tumor suppressors present in histologically normal mammary gland samples of breast cancer patients who experienced recurrent disease, metastasis, secondary tumors, or death within a 10-year follow-up period. Key pathways such as cell adhesion, hormone signaling, and immune regulation were identified as critical players in early tumorigenic processes. These molecular signatures were correlated with patient survival data, demonstrating their utility in predicting recurrence risk (Paper III).

Lastly, we focused on pathogenic post-zygotic variants in non-tumorous tissues from the same breast cancer cohort. An abundance of these variants was observed in the normal mammary gland of patients with poor prognoses, often affecting genes known to drive tumor progression (i.e. *ATK1, PIK3CA, PTEN, TBX3, TP53*) (unpublished findings, manuscript under review). These variants appeared to worsen patient survival, especially in patients with recurrent disease. These findings emphasize the importance of analyzing non-tumorous tissues as they harbor alterations strongly associated with aggressive cancer phenotypes and poorer survival outcomes. By integrating molecular findings with survival metrics, the study demonstrated the clinical relevance of these alterations, which were both detectable and associated with clinical outcomes.

This research addressed critical methodological challenges, including the identification of truly non-tumorous tissues through rigorous histological verification, and the detection of early transcriptomic alterations and low-frequency variants using ultra-sensitive techniques. The inclusion of control samples from reduction mammoplasty surgeries provided a baseline for distinguishing malignant tissues from normal ones with early molecular changes.

Collectively, these findings challenge the assumption that histological markers fully capture underlying molecular aberrations. They confirm the existence of an intermediate state, where microscopically normal tissues harbor alterations driving tumor initiation and metastasis. Uninvolved mammary tissues are shown to play an active role in cancer progression through early tumorigenic processes.

By combining genetic and transcriptomic analyses with long-term clinical data, this thesis offers a comprehensive understanding of how molecular changes drive tumor initiation and recurrence. These findings have significant implications for early detection, recurrence risk assessment, and personalized treatment strategies, paving the way for future studies to improve breast cancer outcomes through earlier, more precise interventions.

Keywords: breast cancer, uninvolved margin, mammary gland, transcriptomic alterations, post-zygotic variants, single nucleotide polymorphisms, copy number variations, somatic mosaicism, mortality, recurrence, poor prognosis.

SUMMARY IN POLISH

Niniejsza rozprawa doktorska bada molekularne mechanizmy leżące u podstaw progresji raka piersi, koncentrując się na zmianach genetycznych i transkryptomicznych w histologicznie prawidłowych tkankach gruczołu sutkowego. W przeciwieństwie do wcześniejszych badań skupiających się głównie na guzach, w niniejszej pracy wykonano analizę tkanek niezmienionych nowotworowo w celu wykrycia wczesnych zmian molekularnych, które mogą stanowić wskaźniki przedkliniczne raka piersi lub predyktory ryzyka nawrotu. Przeanalizowano dane pochodzące od 184 pacjentek z rakiem piersi, rekrutowanych zarówno z kryteriami dotyczącymi rokowania, jak i bez nich, oraz 94 osób kontrolnych poddanych redukcyjnej mammoplastyce. Pobranie próbek obejmowało 267 tkanek z marginesu wolnego od zmian nowotworowych w różnej odległości od pierwotnych zmian, 184 próbki skóry lub krwi obwodowej jako materiał odniesienia oraz 184 guzy pierwotne pochodzące od pacjentek z rakiem piersi (Artykuły II, III i IV). Dodatkowo od 41 osób kontrolnych pobrano próbki skóry lub pełnej krwi w celu zbadania mozaikowatości somatycznej (Artykuły II i IV).

Badania opierają się na koncepcji pola nowotworzenia (ang. field cancerization), która sugeruje, że pozornie normalne tkanki otaczające guzy mogą zawierać zmiany genetyczne i transkryptomiczne predysponujące je do transformacji nowotworowej. Wykorzystując nowoczesne techniki, zbadano strukturalne zmiany chromosomalne, warianty postzygotyczne oraz zmiany transkryptomiczne z wysoką czułością i swoistością.

Praca rozpoczęła się od utworzenia kompleksowego biobanku tkanek gruczołu sutkowego, który rozwiązywał kluczowe wyzwania, takie jak heterogeniczność tkanek i trudności w uzyskaniu dopasowanych próbek kontrolnych. Rygorystyczna walidacja histologiczna zapewniła dokładną klasyfikację tkanek niezmienionych nowotworowo, pozwalając na ich odróżnienie od mikroprzerzutów lub innych zmian złośliwych. Ten biobank stanowił wysokiej jakości, niezawodną bazę do dalszych analiz, niwelując zmienność jakości próbek (Artykuł I).

Kolejny etap dotyczył profilowania genetycznego tkanek niezmienionych nowotworowo. Z wykorzystaniem mikromacierzy SNP, sekwencjonowania egzomu i ultraczułego sekwencjonowania dupleksowego zidentyfikowano strukturalne zmiany chromosomalne i patogenne warianty postzygotyczne o niskiej częstości, w genach takich jak *AKT1*, *PIK3CA* i *TP53*,

które są zwykle związane z rakiem, ale zostały również wykryte w histologicznie prawidłowych tkankach. Odkrycia te podważają tradycyjne rozumienie tkanek niezmienionych nowotworowo jako „biernych obserwatorów", wskazując na ich aktywną rolę we wczesnych procesach nowotworzenia (Artykuł II).

Na podstawie wniosków z genetycznej części badania przeprowadzono profilowanie transkryptomiczne oparte na RNA-seq w celu identyfikacji wzorców ekspresji genów w tkankach nietumorowych w różnej odległości od pierwotnych zmian u pacjentek z rakiem piersi i niekorzystnym rokowaniem. Zaawansowane narzędzia bioinformatyczne, w tym analizy wzbogacenia ścieżek i przeżycia, wykazały unikatową sygnaturę molekularną obejmującą keratyny, białka adhezyjne, onkogeny i geny supresorowe obecną w histologicznie prawidłowych próbkach gruczołu sutkowego pacjentek z nawrotem choroby, przerzutami, wtórnymi guzami lub zgonem w ciągu 10-letniej obserwacji. Kluczowe szlaki, takie jak adhezja komórkowa, sygnalizacja hormonalna i regulacja immunologiczna, zostały zidentyfikowane jako istotne w procesach wczesnego nowotworzenia. Podpisy molekularne skorelowano z danymi dotyczącymi przeżycia, co potwierdziło ich przydatność w przewidywaniu ryzyka nawrotu (Artykuł III).

Na koniec skupiono się na patogennych wariantach postzygotycznych w tkankach niezmienionych nowotworowo tej samej kohorty pacjentek z rakiem piersi. Zaobserwowano mnogość tych wariantów w normalnych tkankach gruczołu sutkowego u pacjentek z niekorzystnymi rokowaniami, które często dotyczyły genów związanych z progresją nowotworu (tj. *AKT1*, *PIK3CA*, *PTEN*, *TBX3*, *TP53*) (Artykuł IV). Warianty te były związane z gorszym przeżyciem pacjentek, zwłaszcza z nawrotem choroby. Te odkrycia podkreślają znaczenie analizy tkanek niezmienionych nowotworowo, które zawierają zmiany silnie związane z agresywnymi fenotypami raka i gorszymi wynikami przeżycia.

Wykonane badania były związane z istotnymi wyzwaniami metodologicznymi, w tym identyfikacją niezmienionych nowotworowo tkanek poprzez rygorystyczną weryfikację histologiczną oraz wykrywaniem wczesnych zmian transkryptomicznych i wariantów postzygotycznych o niskiej częstości za pomocą ultraczułych technik sekwencjonowania. Włączenie próbek kontrolnych od osób poddanych redukcyjnej mammoplastyce dostarczyło

istotnego punktu odniesienia dla rozróżnienia tkanek złośliwych od histologicznie prawidłowych tkanek z wczesnymi zmianami molekularnymi.

Wyniki badań podważają założenie, że markery histologiczne w pełni odzwierciedlają molekularne aberracje. Potwierdzają one istnienie „stanu pośredniego," w którym mikroskopowo prawidłowe tkanki zawierają zmiany napędzające inicjację nowotworu i przerzuty.

Integrując analizy genetyczne i transkryptomiczne z danymi klinicznymi, rozprawa ta oferuje kompleksowe zrozumienie, w jaki sposób zmiany molekularne napędzają inicjację nowotworu i ryzyko nawrotu. Odkrycia te mają istotne implikacje dla wczesnego wykrywania raka piersi, oceny ryzyka nawrotu oraz opracowywania spersonalizowanych strategii leczenia, torując drogę przyszłym badaniom mającym na celu poprawę wyników leczenia poprzez wcześniejsze i bardziej precyzyjne interwencje terapeutyczne.

Słowa kluczowe: rak piersi, niezajęty margines, gruczoł sutkowy, zmiany transkryptomiczne, warianty postzygotyczne, polimorfizmy pojedynczego nukleotydu, zmiany liczby kopii, mozaicyzm somatyczny, śmiertelność, nawrót, złe rokowanie.

I. INTRODUCTION

1.1. Global Burden of Cancer

Cancer remains a major global health challenge, accounting for approximately one in six deaths worldwide. In 2022, there were around 20 million new cancer cases and 10 million cancer-related deaths, underscoring the persistent threat despite significant research advances (**Figure 1a**). As the global population continues to grow and age, the cancer burden is expected to rise, with an estimated 35 million new cases projected by 2050—a 75% increase from 2022 levels. This increase is driven by lifestyle factors such as sedentary behavior, unhealthy diets, and exposure to harmful substances, including tobacco smoke and environmental pollutants. Improved diagnostic technologies have also led to higher detection rates, particularly in previously underreported regions.

A small number of cancer types account for the majority of cases and deaths. According to GLOBOCAN 2022, the ten most common cancers represented about 64% of all new diagnoses and 70% of cancer-related deaths globally. Lung cancer, responsible for almost 2.5 million new cases (12.4% of all cancers globally), was the most frequently diagnosed, followed by female breast cancer (11.6%), colorectal cancer (9.6%), prostate cancer (7.3%), and stomach cancer (4.9%) (**Figure 1a**). The number of new cancer cases is projected to rise to over 35 million by 2050, driven by population growth and aging[1].

1.2. Breast Cancer Incidence, Mortality, and Risk Factors

In 2022, breast cancer was the second most common cancer globally, with 2.3 million new cases, making up 11.6% of all cancer diagnoses. It was the fourth leading cause of cancer death worldwide, with 666,000 fatalities. Among women, breast cancer was the most frequently diagnosed and the leading cause of cancer death in 157 countries for incidence and 112 countries for mortality, accounting for nearly one in four cancer cases and one in six cancer deaths globally (**Figure 1b**)[1].

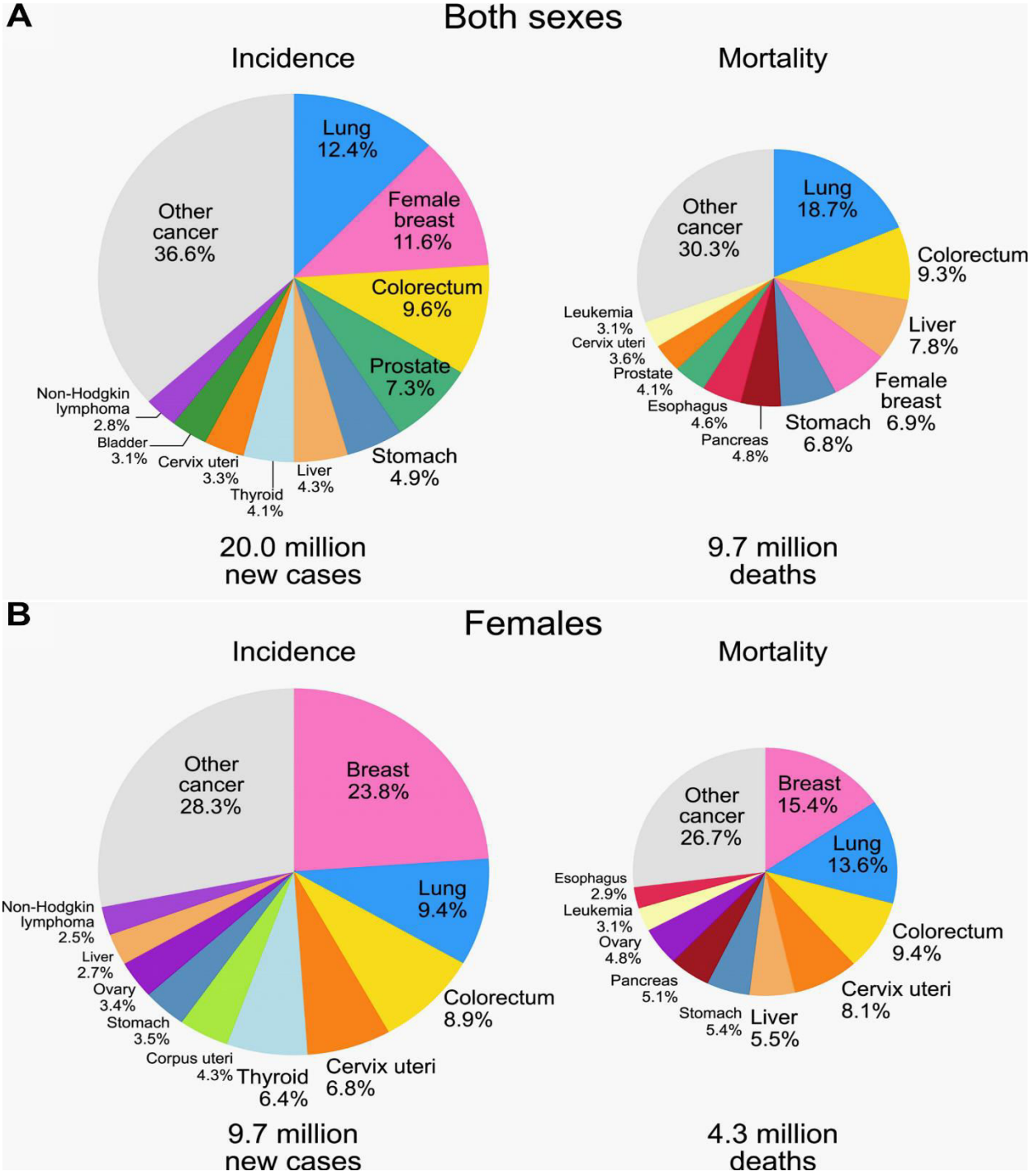

Figure 1. *The pie charts illustrate the proportion of each primary cancer type among all diagnoses and the proportion of each cancer type among all cancer-related deaths in 2022 for A: both sexes and B: females. For each sex, the area of the pie chart reflects the proportion of the total number of cases or deaths; nonmelanoma skin cancers (excluding basal cell carcinoma) are included in the other category. Figure adapted from Bray et al. (2024), CA Cancer J Clin. [1].*

Both modifiable and non-modifiable factors influence breast cancer risk. Approximately 30% of cases are associated with lifestyle factors, such as obesity, physical inactivity, and alcohol consumption, which can be mitigated through dietary adjustments, increased physical activity, and reduced alcohol intake [2]. Non-modifiable risk factors include age, family history, and genetic predisposition. However, the majority of breast cancer cases are classified as sporadic, occurring without a known genetic link or family history. Conversely, only 5-10% of cases are hereditary, with 25-30% attributable to pathogenic variants in genes like *ATM, BRCA1, BRCA2, CHEK2, PALB2, PTEN,* and *TP53* [3–5]. Reproductive factors such as late age at first birth, nulliparity, early menarche, and late menopause are also associated with an elevated risk.

1.3. Breast Cancer Classification and Staging

Breast cancer is a heterogeneous disease with a variety of histological types and molecular subtypes, each with unique clinical presentations, prognoses, and treatment approaches. The two most prevalent types are Invasive Ductal Carcinoma (IDC), which originates in the milk ducts and spreads to surrounding breast tissue, and Invasive Lobular Carcinoma (ILC), which begins in the lobules. Less common invasive forms include Medullary Carcinoma; Mucinous (Colloid) Carcinoma, characterized by mucus-producing cells; and Tubular Carcinoma, a less aggressive IDC subtype. Papillary Carcinoma, though rare, generally occurs in older women and is often associated with a favorable prognosis. Breast cancer can be broadly categorized into non-invasive and invasive types. Non-invasive breast cancer, such as Ductal Carcinoma In Situ (DCIS), remains confined within the milk ducts, whereas invasive breast cancers like IDC and ILC extend beyond their original site [6].

Molecular classification based on hormone receptors (estrogen and progesterone) and HER2 status has been pivotal in guiding personalized treatment approaches. High-throughput genomics and transcriptomics have further elucidated the molecular complexity of breast tumors, leading to the identification of four clinically meaningful subtypes: Luminal A, Luminal B, HER2-enriched, and Basal-like or Triple-Negative Breast Cancer (TNBC) [7,8]. Accurate staging, based on the American Joint Committee on Cancer (AJCC) system, is essential for determining appropriate treatment and predicting outcomes, with stages ranging from Stage 0 (DCIS) to Stage IV (distant metastases) [9].

1.4. Physiology of the human mammary gland

Exploring the heterogeneity of breast cancer and its varying prognoses requires a detailed knowledge of the normal development and physiology of the mammary gland. The human mammary gland is a complex and dynamic organ integral to the female reproductive system. Its development undergoes distinct changes during the embryonic stage, puberty, and lactation phases (**Figure 2**).

Much of what is known about these processes comes from studies using model organisms, particularly mice, which offer valuable insights into mammary gland biology and its relevance to human breast cancer. Mammary gland development begins during the embryonic stage with the formation of bilateral mammary lines at around embryonic day 10.5 (E10.5) in mice. These lines develop into mammary placodes by E11.5, which then invaginate to form mammary buds. These buds continue to develop into a rudimentary ductal tree, which remains quiescent until puberty. During this stage, critical signaling pathways and interactions with the surrounding mesenchyme guide the early formation of the mammary gland. At puberty, the mammary gland undergoes extensive ductal growth and branching, driven by hormonal signals, particularly estrogen. The Terminal End Buds (TEBs) are the key structures involved in this process. These TEBs, located at the tips of growing ducts, are responsible for the elongation and bifurcation of ducts into the mammary fat pad, leading to the formation of a complex ductal network. The process of ductal morphogenesis during puberty establishes the basic architecture of the mammary gland. During pregnancy, the mammary gland experiences further development, characterized by the proliferation and differentiation of alveolar structures. Under the influence of hormones such as progesterone and prolactin, these alveolar units prepare the gland for lactation. The gland undergoes significant expansion, forming mature alveoli capable of milk production. This stage is crucial for the gland's function in feeding offspring. Following lactation, the gland enters a phase of involution, where the alveolar structures regress, and the gland returns to a state similar to the pre-pregnant phase, although some changes persist. These developmental stages highlight the dynamic nature of the mammary gland. Understanding these processes is essential for grasping the underlying mechanisms of mammary gland biology and their implications for breast cancer, particularly in identifying the origins of different breast cancer subtypes [10–13].

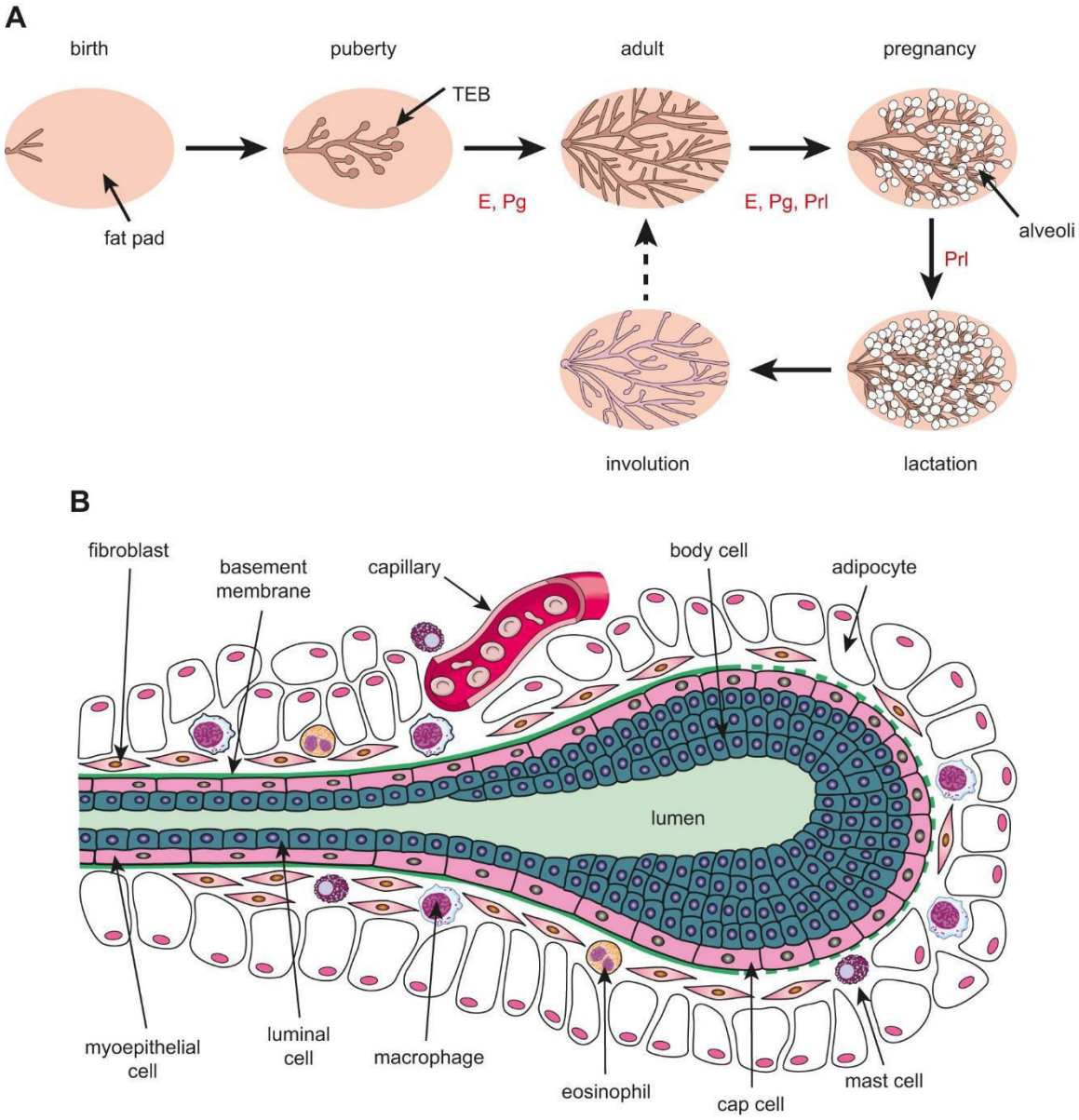

Figure 2. Diagram of postnatal mammary gland development. A: in the postnatal animal, the early mammary gland grows in an allometric fashion and remains relatively dormant until the onset of puberty. At this stage, dramatic morphogenesis occurs, largely under the control of Estrogen (E). In the young adult, Progesterone (Pg) regulates side-branching, while in pregnancy, the steroid hormones E, Pg, and Prolactin (Prl) exert roles in expansion of the alveolar units. In the late stages of pregnancy and during lactation, the peptide hormone Prl plays a key role in establishing the secretory state. After lactation, the gland involutes and returns to a resting state. B: representation of a terminal end bud in a pubertal mouse mammary gland. Figure reproduced from Fu et al. (2020), Physiol Rev. [10].

1.5. Screening and Treatment

Early detection and accurate diagnosis of breast cancer are critical for effective management and improved survival outcomes. Conventional diagnostic methods, such as mammography, ultrasound, and biopsy, remain the cornerstone of breast cancer detection. However, these methods are not without limitations. Mammography, for instance, has long been the standard for breast cancer screening, yet it is prone to errors due to its reliance on manual interpretation, which can lead to variability in identifying and assessing masses. The accuracy of mammography can also be affected by the expertise of radiologists and their workload, particularly in resource-limited regions where access to specialized training and technology may be restricted. To overcome these challenges, recent advancements in genomic biomarkers and deep learning algorithms have significantly enhanced diagnostic accuracy and personalized risk assessment. These technologies enable more precise identification of cancerous cells and allow for better differentiation between benign and malignant tumors, reducing the likelihood of false positives and negatives. Despite these technological improvements, the search for even more reliable and non-invasive imaging modalities continues. Researchers are actively exploring innovative imaging techniques such as microwave imaging, ultrasound tomography, breast tomosynthesis, and contrast-enhanced digital mammography. These emerging methods hold the potential to provide more accurate detection by improving image resolution and contrast, which are crucial for identifying small or dense breast tumors. However, they also present challenges, including high costs, radiation exposure, and limited accessibility, which may hinder widespread adoption. Magnetic Resonance Imaging (MRI), known for its ability to detect small lesions that mammography might miss, is another tool in the diagnostic arsenal. However, MRI's low specificity can lead to overdiagnosis, which may result in unnecessary treatments and increased patient anxiety. Positron Emission Tomography (PET) is highly effective in visualizing tumor spread and assessing therapy response, yet it requires expensive, specialized equipment, making it less accessible, particularly in low-resource settings [14,15].

As breast cancer screening technology continues to evolve, the development of more accurate, cost-effective, and widely accessible methods is imperative. Such advancements will play a crucial role in improving early detection rates, thereby reducing the global burden of breast cancer and enhancing patient outcomes. In addition to technological innovations, a deeper understanding of the seemingly normal mammary gland and the earliest molecular and cellular alterations that

precede cancer development is essential. By studying these early changes, researchers can identify biomarkers and patterns indicative of the initial stages of cancer, potentially allowing for diagnosis before the disease becomes clinically apparent. This proactive approach could revolutionize breast cancer screening, enabling interventions at the very onset of malignancy, ultimately improving survival rates and reducing the need for more aggressive treatments.

Breast cancer treatment has severely shifted towards Breast-Conserving Surgery (BCS) over mastectomy in recent years, reflecting global trends and advancements in medical care. This preference for less invasive options prioritizes both oncological outcomes and aesthetic considerations, driven by the integration of improved detection methods, advancements in radiation therapy, and a stronger emphasis on shared decision-making between patients and healthcare providers. The evolution of treatment approaches has allowed patients to play a more active role in their care, as they carefully weigh the benefits and risks of BCS against those of mastectomy. This decision-making process is influenced by a range of factors, including the stage and aggressiveness of cancer, patient preferences, and the availability of effective adjuvant therapies, such as chemotherapy, hormonal therapy, and targeted therapies [16,17].

For women at increased risk of breast cancer, particularly those with a strong family history or genetic predisposition, Contralateral Prophylactic Mastectomy (CPM) presents a preventive option aimed at reducing the risk of developing cancer in the opposite breast. However, CPM is not without its challenges; it carries potential psychological and physical implications, making it a complex and highly individualized decision. The consideration of CPM must involve a thorough evaluation of the patient’s personal risk factors, the psychological impact of living with the fear of recurrence, and the potential consequences of surgery on the patient's quality of life. As a result, the decision to undergo CPM requires careful deliberation, often involving genetic counseling, psychological support, and a detailed discussion of the expected outcomes [18].

Breast cancer treatment, now prioritizing breast-conserving surgery, underscores the importance of early detection and understanding of initial tissue changes in identifying cancer early, enabling less invasive treatments and improving outcomes.

II. LITERATURE REVIEW

2.1. The Origin of Breast Cancer and Cancer Evolution Theories

Cancer, rather than a static disease, represents a continually evolving entity shaped by complex genetic and environmental interactions. At the core of this dynamic process lies Intra-Tumor Heterogeneity (ITH) - the diversity of genetic and phenotypic profiles within a single tumor and its metastases - which poses a significant challenge for developing universally effective treatments [19]. To explain ITH, several theories of cancer origin have emerged, including the Cancer Stem Cell (CSC) theory [20] and the Clonal evolution theory [21], each offering unique insights into tumor progression.

The CSC theory suggests that tumors originate from a rare subset of self-renewing cells that can differentiate into various CSC and non-CSC subpopulations, contributing to tumor complexity and treatment resistance. Initially observed in hematopoietic cancers and later in solid tumors like breast and brain cancers, this model depicts CSCs as sitting at the top of a hierarchical tumor structure. CSCs divide asymmetrically, generating both new CSCs and non-CSC cells, the latter comprising the tumor bulk but contributing less to its growth. CSCs' resilience is linked to high recurrence and therapy resistance, as non-CSCs can revert to CSCs, fueling aggressive tumor behavior[22].

Conversely, the Clonal evolution theory, proposed by Nowell in 1976, posits that tumors stem from a single cell accumulating mutations over time, creating increasingly aggressive and diverse subpopulations [23]. Clonal evolution can follow two paths: Linear Evolution (LE), with sequential mutation accumulation, and Branched Evolution (BE), where distinct mutations diversify the tumor. Although LE is less common in advanced cancers, BE is widely applicable to breast cancer and is supported by findings of subclonal driver mutations and convergent evolution, where different lineages acquire the same driver mutations, leading to parallel expansions.

The clonal evolution theory is often associated with the stochastic theory, but the two differ subtly. The stochastic theory suggests that any cancer cell can acquire mutations that drive tumor growth, emphasizing random mutations over predefined CSCs. Thus, ITH results from random genetic variations and environmental influences, affecting treatment responses [24]. Tumor evolution in humans, however, is challenging to study due to ethical constraints, so researchers often infer

evolutionary history from single time-point samples through phylogenetic methods, though these may miss key intermediates [25].

Key components in understanding tumor evolution models include the definition of a clone (a group of tumor cells sharing a similar genotype and mutational profile) and a subclone (a subset of tumor cells that have diverged from the main lineage and acquired additional mutations) [26]. Fitness refers to a tumor cell's ability to survive, proliferate, and propagate its genotype within the tumor. Driver mutations confer a fitness advantage, increasing the prevalence of certain clones, while passenger mutations do not affect fitness. Clonal expansion occurs when a genotype with increased fitness becomes more prevalent in the tumor mass, and a selective sweep happens when a highly fit genotype outcompetes all other clones [27].

Key concepts in tumor evolution include clones (cells sharing a genotype) and subclones (cells diverging from the main lineage with additional mutations) [26]. Fitness, the cell's capacity to proliferate and spread, is enhanced by driver mutations that increase clone prevalence. Clonal expansion occurs when a genotype with increased fitness becomes more prevalent in the tumor mass, and a selective sweep happens when a highly fit genotype outcompetes all other clones [27]. Next-Generation Sequencing (NGS) techniques, reveal genetic profiles and evolutionary trees, highlighting patterns of clonal divergence and subclonal mutations in breast cancer [25].

The recently proposed Punctuated Evolution (PE) model diverges from the gradual alteration accumulation seen in BE and LE models. Instead, it suggests that significant genomic changes occur in short bursts early in tumor development, generating high ITH upfront. Following these bursts, dominant clones stabilize, forming most of the tumor mass and contributing to a relatively stable structure thereafter. Analogous to the "Punctuated Equilibrium" in species evolution, this model implies that tumors may be "pre-programmed" to become aggressive or therapy-resistant [28,29].

Evidence supporting the PE model has been particularly observed in DNA copy number aberrations and chromosomal rearrangements. Early studies, such as those on "firestorms" in breast cancer, described localized amplifications on single chromosome arms correlated with aggressive disease [30,31]. Similarly, "chromothripsis," involving massive chromosomal rearrangements occurring in a single event, has been observed in bone, colorectal, and prostate cancers. These rearrangements

result in highly branched phylogenetic trees with few intermediates, supporting the PE model [32–35].

Interestingly, these models of tumor evolution may not be static but can transition over time or coexist for different mutation types. LE, though less frequent in advanced cancers, may occur in the early stages, transitioning to BE as the tumor grows. In breast cancer, early-stage tumors may follow different patterns than advanced ones, with early, rapid bursts of copy number alterations stabilizing over time, while point mutations evolve gradually [36]. Single-cell sequencing and longitudinal studies suggest that Copy Number Alterations (CNAs) and point mutations follow distinct evolutionary paths, further illustrating cancer's complexity [25].

Understanding these tumor evolution principles, especially ITH, is crucial for cancer treatment. Each tumor's unique evolutionary path underscores the importance of personalized medicine. Therapies tailored to the molecular profile of each patient's cancer can better target diverse tumor subpopulations, requiring combination therapies to simultaneously target multiple pathways and reduce resistance. Adaptive therapy strategies, which continuously monitor tumor changes and adjust treatments, are vital for durable responses.

In conclusion, a comprehensive understanding of cancer's evolutionary trajectory is essential for developing effective treatment strategies. While the clonal evolution theory, with its linear and branching models, provides a foundational framework, the cancer stem cell theory and the punctuated evolution model introduce additional layers of complexity. These theories emphasize the dynamic and heterogeneous nature of tumor progression. The interplay between these models, particularly the role of ITH, underscores the need for a nuanced understanding of tumor evolution. This knowledge is vital for devising personalized and adaptive therapies that can effectively target the diverse subpopulations within a tumor and improve patient outcomes. As our grasp of tumor evolution continues to advance, we can anticipate the development of more targeted and effective cancer treatments that accommodate the complexity and dynamism of this challenging disease.

2.2. Evolving Perspectives on Tumor Origins

The debate over tumor origin has shifted from single-cell models to multifaceted approaches. The "cancer field effect" or "field cancerization," first proposed by D.P. Slaughter in 1953, views carcinogenesis as a stepwise genetic process, where an initial mutation gives a cell a proliferative advantage, creating a field of clonal cells [37]. This concept led to investigations of seemingly

normal tissue surrounding lesions, suggesting that these tissues may harbor mutations that, under certain conditions, could trigger cancer initiation or recurrence.

As research has evolved, the focus in breast cancer studies has shifted from the single-cell origin model to more nuanced perspectives emphasizing hormonal exposure, tissue microenvironment, and gene-environment interactions. Prolonged estrogen exposure, beginning at menarche and continuing through menopause, increases susceptibility to breast cancer through mechanisms such as DNA damage and cellular stress responses, even in women considered to be at "normal risk" [11]. Additional factors, including age, parity, and inherited pathogenic genetic variants—particularly in the *BRCA1* and *BRCA2* genes—significantly influence mammary tissue composition, with specific changes observed in immune and epithelial cells [38]. The etiologic field theory developed more than 60 years after the original concept, further expands on these ideas, emphasizing the abnormal tissue microenvironment's role at all stages of tumor development. This theory challenges the notion that markers solely indicate neoplasia, suggesting instead that they may reflect broader environmental changes, including contributions from non-transformed cells and the extracellular matrix to cancer progression [39].

Recent advances reveal that breast cancer may arise from lineage-restricted progenitor cells rather than multipotent stem cells, especially during puberty and pregnancy [10]. This suggests that specific progenitor cells may drive particular breast cancer subtypes, including aggressive forms like triple-negative breast cancer. Targeting these progenitor cells could improve treatment outcomes by addressing the unique molecular profiles underlying these subtypes.

The evolving understanding of cancer - from single-cell to multi-stage, gene-environment models - reflects the complexity of tumorigenesis. Insights into progenitor cells and the tumor microenvironment are shaping more precise therapies. Continued research is crucial to prevent, diagnose, and treat cancer effectively, leveraging an increasingly nuanced understanding of cancer's origins and progression.

2.3. Premalignant Changes and Transcriptomic Alterations in Cancer-Adjacent Tissues

Research on cancer initiation increasingly highlights the role of seemingly normal tissues adjacent to tumors in early disease development. While studies have traditionally centered on genetic and molecular changes within tumors, emerging evidence reveals that histologically normal yet genetically altered tissues surrounding tumors may also contribute to cancer progression.

Numerous studies identify transcriptomic changes, genetic mutations, and epigenetic alterations in nearby mammary gland tissues, suggesting a precancerous state even in non-tumorous cells [40].

Challenging the traditional single-cell origin model, Nishimura et al. propose that breast tumors often arise from multiple founder cells [41]. This hypothesis is supported by the presence of cancer-related clones in non-cancerous tissues, indicating that precancerous cells may exist before detectable lesions appear. The study emphasizes estrogen's role in mutation accumulation pre-menopause, as well as the significant contribution of localized microenvironmental and epigenetic changes to tumor development.

Transcriptomic studies further underscore the presence of significant molecular alterations in histologically normal tissue adjacent to tumors. Graham et al. found altered gene expression related to inflammation, cell cycle regulation, and DNA repair in these tissues, supporting the "field cancerization" concept and hinting that such alterations could be early markers of cancer risk [42]. Roman-Perez and colleagues identified subtypes in adjacent normal tissues, including an "Active" subtype linked to poorer survival in estrogen receptor-positive patients due to TWIST1 overexpression and claudin-low features [43]. Huang et al., using data from The Cancer Genome Atlas, demonstrated that these normal-adjacent tissues provide valuable prognostic insights potentially distinct from those of the tumor microenvironment [44]. Similarly, Aran et al. identified a unique gene expression signature in adjacent tissues across multiple cancer types, suggesting that tumors actively shape their surroundings to support invasion and metastasis [45].

Recent studies by Gadaleta et al. and Morla-Barcelo et al. further illuminate molecular changes in peritumoral tissues. Gadaleta’s team identified four transcriptomic subtypes in adjacent normal tissues that provide prognostic value, while Morla-Barcelo et al. found upregulation of genes associated with inflammation, cell cycle, and extracellular matrix remodeling in estrogen receptor-positive tumors [46,47]. Additionally, Sverchkova et al. investigated immune-related gene expression in adjacent tissues, identifying Human Leukocyte Antigens (HLA) genotyping as a promising biomarker for immunotherapy stratification, particularly in triple-negative breast cancer [48]. Lastly, Lau et al. categorized peri-tumoral samples into clusters based on immune and cellular compositions, finding a pro-inflammatory, adipose-enriched cluster linked to poorer survival and a myofibroblast and adaptive immune-enriched cluster associated with better outcomes. The study suggests that mammographic breast density may influence peri-tumoral subtypes and patient

prognosis [49]. Together, these findings underscore the importance of studying cancer-adjacent tissues for early cancer detection and risk assessment. Identifying transcriptomic and molecular alterations in adjacent tissues offers new avenues for early intervention and tailored surveillance, advancing personalized cancer treatment strategies.

2.4. Copy Number Alterations and Somatic Mutations in the Normal Mammary Gland

Somatic mosaicism, resulting mainly from post-zygotic mutations, significantly contributes to genetic diversity within tissues. These mutations, which include simple nucleotide variants and structural changes, are crucial for processes such as immune diversification and neuronal complexity. However, they are also linked to diseases like cancer, cardiovascular disease, and Alzheimer's, particularly in aging populations [50].

Clonal expansions driven by somatic mutations are prevalent in both cancerous and normal tissues, increasing with age and exposure to environmental factors like UV light, smoking, and inflammation. Mutated clones accumulate across various organs, remodeling tissues and potentially influencing cancer and other diseases, including cardiovascular conditions, autoimmune disorders, and infections. Many driver mutations arise early in life but may take decades to lead to cancer, highlighting a slow, multi-stage progression to carcinogenesis. Clonal selection mechanisms vary between normal and cancerous tissues; for instance, mutations in *NOTCH1* can suppress tumor growth in the esophagus (**Figure 3**) [51]. Understanding these clones may aid in early diagnosis and prevention, leveraging clonal expansions to manage cancer risks and age-related diseases.

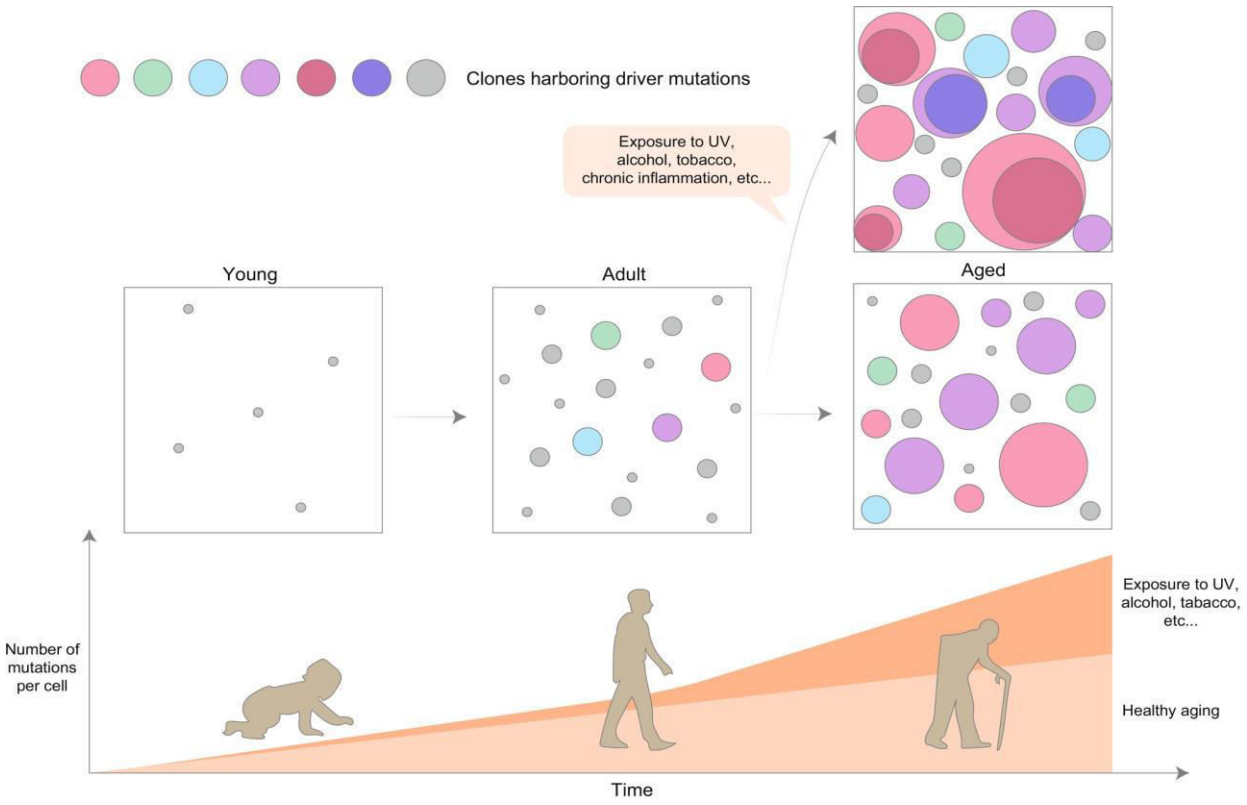

Figure 3. Mutation accumulation and clonal expansion in normal tissues. The patchwork plot shows the spread of clones harboring driver mutations in normal tissues. With aging, somatic mutations accumulate in cells, clones with driver mutations expand, and tissues undergo remodeling. Exposure to environmental factors, such as UV rays (skin), alcohol consumption, and smoking (esophagus), increases the mutation accumulation rate and promotes the expansion of mutant clones. Figure reproduced from Maeda and Kakiuchi (2024), Cancer Sci. [51]

It has been recently proposed that metabolic factors such as obesity and diabetes, along with treatments like metformin, significantly impact the expansion of *PIK3CA* mutant clones in normal tissues [52]. Conditions that activate the PI3K-mTOR pathway enhance the competitive fitness of these oncogenic mutants, facilitating their clonal expansion even in non-cancerous tissues. Conversely, metabolism-modulating interventions like metformin can reduce the fitness advantage of these mutants, suggesting potential strategies for cancer prevention by targeting metabolic pathways.

In breast cancer, chromosomal CNAs—including focal deletions, amplifications, and aneuploidy—serve as essential biomarkers for diagnosis, treatment planning, and patient stratification. These CNAs contribute to cancer heterogeneity and therapy resistance, offering

prognostic insights beyond traditional histology. For example, amplification of chromosome arm 1q is particularly valuable for prognosis in TNBC [53].

Significantly, structural genetic aberrations are found not only in tumor tissues but also in cancer-free breast tissues of patients with sporadic breast cancer, affecting nearly 40% of cases. Aberrations like *ERBB2* gene gains observed far from primary tumors suggest a widespread and progressive field cancerization process, supporting the idea that genetically altered cells can precede tumor formation [50].

Additionally, approximately 10% of uninvolved glandular tissue cells from breast cancer patients display CNAs, indicating a predisposition to genomic instability that may contribute to early cancer initiation. This finding has profound implications for early detection and risk management, especially regarding radiotherapy [54]. Pereira et al. emphasize the importance of integrating CNA profiling with gene mutation analysis for breast cancer classification, identifying several mutation-driver genes relevant to targeted therapy, such as *FOXO3* and *AGTR2* [55].

Recent research has expanded our knowledge of breast cancer mutations. For instance, Nik-Zainal et al. sequenced whole genomes from 560 breast cancers and identified millions of base substitutions, small indels, and genomic rearrangements, underscoring the need for continued exploration of cancer's genomic landscape [56]. In another study, Li et al. reported frequent somatic *TP53-PIK3CA* co-mutations in Chinese breast cancer patients, correlating with poorer survival [57].

Oh and Sung (2020) found that somatic mutations, including those in *PIK3CA*, also appear in histologically normal tissues near cancerous areas. Often matching mutations in the adjacent tumors, these findings suggest that nearby "normal" tissues may harbor early tumorigenic events [58].

Advances in mutation detection have further revealed modest variation in somatic mutation rates across cell types, indicating that division-independent mutational processes may play significant roles in somatic mutation [59]. A study from Hungary observed frequent mutations in *TP53*, *PIK3CA*, and *KMT2C*, reinforcing their significance in breast cancer [60].

In recent work, Rockweiler et al. (2023) utilized a multi-tissue atlas to examine post-zygotic mutations across 54 tissue types from 948 donors. Their study shows that mutations vary

significantly across tissues and are influenced by age and tissue type, with prenatal mutations often being more deleterious. This research sheds light on the potential for using post-zygotic mutations in diagnostics, particularly for cancer risk assessment [61].

2.5. Clinical Implications and Limitations

The discovery of early molecular changes—including transcriptomic alterations, pathogenic point mutations, and chromosomal CNAs—in normal tissues adjacent to tumors holds considerable potential for advancing cancer diagnosis, risk assessment, and personalized treatment strategies. By identifying these early alterations, researchers hope to achieve more precise patient stratification and to develop targeted therapies that improve patient outcomes. However, the clinical application of these findings faces several challenges. Small cohort sizes, patient heterogeneity, and inconsistent definitions of "normal" tissue affect the robustness and generalization of results. Furthermore, a lack of healthy control samples and reliance on mastectomy-derived tissue samples limit the applicability of these insights. The variability in detection methods and the risk of tumor contamination further complicate data interpretation. Addressing these issues will be essential for realizing the clinical potential of early molecular alterations in improving cancer care.

2.6. Challenges in Defining and Controlling Histologically Normal Tissue in Breast Cancer Research

Studying adjacent normal tissue holds great promise, yet several challenges must be resolved. Defining "histologically normal" mammary tissue accurately is one of the primary hurdles. To ensure samples are uncontaminated by tumor cells, researchers must adhere to meticulous protocols, selecting samples from regions distant from the primary tumor and, ideally, from separate breast lobes. Independent pathologists should evaluate these samples to confirm their normalcy. Terms such as "uninvolved margin (UM)" or "uninvolved tissue" are commonly used to describe non-tumorous tissue with no visible cancer signs [50,54].

Selecting suitable control samples poses additional complexities. Ideally, controls would consist of individuals without a personal or familial cancer history, but achieving age-matching is challenging since younger patients typically undergo cosmetic surgeries rather than cancer treatments. Tissue from reduction or prophylactic mastectomies can be used as controls, but these may not always be completely free of cancer risk. Breast cancer is relatively uncommon in younger

women, with only one in eight invasive cases diagnosed in women under 45[62]. Moreover, breast cancer's high lifetime incidence—approximately 13% of women will develop the disease—complicates efforts to identify truly "healthy" control samples [62,63].

III. AIMS

To investigate the molecular mechanisms underlying breast cancer development and progression, this doctoral research addresses critical gaps in sample collection, detection methodologies, and molecular profiling. **The overarching objective is to advance the understanding of early molecular alterations, their prevalence, and their clinical relevance in reportedly sporadic breast cancer.** This study employs a multi-faceted approach: first, the establishment of unique and comprehensive biobanking protocols (**Aim I**); second, genetic analyses of histologically verified, non-tumorous mammary tissues from breast cancer patients, with the inclusion of control samples from mammoplasty patients (**Aim II**). Furthermore, transcriptomic and genetic investigations were conducted in histologically verified, non-tumorous mammary tissues from breast cancer patients with adverse outcomes, incorporating comparisons with breast cancer patients recruited without any criteria related to prognosis and control samples, to assess the clinical implications of these findings (**Aims III and IV**).

Aim I of this work (Paper I, Filipowicz N. et al.) laid the foundation by developing a comprehensive biobank of **histologically controlled, non-tumorous mammary gland samples** collected from various distances from primary lesions. The biobanking protocols, which also included tumor, blood, and skin samples, were designed to mitigate challenges related to small cohort sizes, patient heterogeneity, and the difficulty of defining histologically normal tissue. These carefully curated samples are pivotal to ensuring the reliability and quality of "omics" studies in cancer research.

Building upon this resource, **Aim II** (Paper II, Kostecka A. et al.) employed **ultra-high sensitivity techniques** to identify subtle molecular changes, such as structural rearrangements and pathogenic post-zygotic genetic variants, in **breast cancer-related genes** within the normal mammary glands of sporadic cancer patients. This study addresses limitations in current detection methods, offering insights into early molecular alterations that may precede tumor formation.

In **Aim III** (Paper III, Andreou M. & Jąkalski M. et al.), **transcriptomic profiling** of histologically controlled non-tumorous tissues and primary tumors was performed, with a focus on patients with **adverse outcomes**. Using a custom gene panel targeting genes associated with breast cancer dissemination and metastasis, this study assessed the **clinical relevance of transcriptomic**

alterations in non-tumorous tissues by comparing patient samples with those from individuals undergoing reduction mammoplasty. This analysis also aimed to address issues of sample contamination and improve the selection of appropriate controls for breast cancer research.

Finally, **Aim IV** (unpublished findings, manuscript under review) investigated the prevalence of **post-zygotic and germline variants** in paired histologically controlled non-tumorous mammary tissues and primary tumors from sporadic breast cancer patients with adverse outcomes, as well as from patients without prognosis-specific criteria and control individuals. This study aimed to evaluate the correlation of pathogenic **post-zygotic variants** with clinical outcomes, such as recurrence and mortality, contributing to improved **patient stratification and treatment strategies** based on genomic alterations.

Together, this research addresses complementary aspects of somatic mosaicism, molecular alterations, and patient outcomes, providing a robust framework for advancing breast cancer research and refining clinical approaches to diagnosis and treatment.

IV. MATERIALS AND METHODS

4.1. Cohorts: Patient Selection and Samples Studied

Access to well-characterized, histologically validated non-tumorous samples, reference tissues (e.g., blood or skin), and control samples, combined with comprehensive clinical follow-up information, is crucial for investigating the oncogenic potential of seemingly normal tissues. Given the diversity of the research papers discussed in this thesis, multiple cohorts were indispensable. The primary cohorts analyzed are described below:

4.1.1. Biobank (Paper I)

This biobank encompasses samples from five cancer types known for high incidence and/or often fatal outcomes: breast (933 donors), colorectal (383 donors), prostate (221 donors), bladder (81 donors), and exocrine pancreas carcinomas (15 donors), as well as metachronic metastases of colorectal cancer to the liver (14 donors). Additionally, samples from 64 healthy male donors were included in studies on the Loss Of the Y chromosome (LOY). The recruitment took place across five clinical centers in Poland: the Oncology Center in Bydgoszcz, the National Institute of Oncology in Cracow, the University Clinical Centre in Gdańsk, the University Hospital in Cracow, and Specialist Hospital in Koscierzyna. Sample collection was approved by the Independent Bioethics Committee for Research at the Medical University of Gdańsk, and written informed consent was obtained from all participants. Stringent inclusion criteria were enforced, especially for breast cancer patients undergoing mastectomy or BCS, who were required to be free from neoadjuvant therapy. The biobank at the end of 202, after two years of collecting, included 1,711 patients and controls, totaling 23,226 samples. On average, 74 donors and 1,010 samples were added monthly over nearly two years. Notably, 40% of samples are from macroscopically healthy cancer-adjacent tissues, and 12% are from tumors, adding significant value for studies on cancer predisposition.

Sample collection protocols were designed through collaboration among molecular teams, surgeons, and pathologists. For each diagnosis, the standard sample set included 1-2 Primary Tumor fragments (PT), 1-12 UM specimens from various distances from the PT, 1-4 Whole Blood (WB) samples (1.5 ml each), and 1-2 blood Plasma (BP) samples (1-1.5 ml each) for future proteomic studies. Each tissue fragment was split into two parts: one is fresh-frozen at -80°C, and the other is fixed in formalin, embedded in paraffin, and processed for Hematoxylin and Eosin

(H&E) staining. For breast and colorectal cancers, local lymph node metastases were also collected if identifiable. Each sample is verified with histopathological reports. Additionally, uninvolved margin and skin samples are collected to establish organoids and primary cell cultures.

4.1.2. Breast Cancer Patients Diagnosed with Sporadic Breast Cancer (Paper II)

This cohort includes 52 patients diagnosed with sporadic breast cancer who did not receive neoadjuvant therapy. The focus was on patients undergoing breast-conserving surgery (BCS), who constituted two-thirds of the cohort. A total of 204 samples were collected, including UM, PT, Skin (SK), and Peripheral Blood (BL), from the Oncology Centre in Bydgoszcz and the University Clinical Centre in Gdańsk, with the approval of the bioethics committee at the Medical University of Gdańsk (MUG). Written informed consent was obtained from all participants. The histological subtypes and tumor tissue content of each PT sample were evaluated by pathologists according to the respective AJCC guidelines [9], and tumor samples with less than 50% neoplastic cell content were excluded. The normal mammary gland was sampled from the opposite quadrant relative to the primary tumor site, maintaining a mandatory distance of at least 3 cm to exclude potential contamination by residual tumor cells. Pathologists also evaluated these tissue samples to confirm normal histology. All normal mammary gland samples from patients who underwent breast-conserving surgery were derived from the tissue that remained intact after the surgery.

4.1.3. Reportedly Sporadic Breast Cancer Patients Selected Based on Unfavorable Prognosis (Papers III and IV)

Breast cancer patients with unfavorable outcomes and extensive clinical follow-up data collected for up to 10 years post-surgery were recruited as part of a large biobanking effort between 2012 and 2018 (n=497). Criteria included disease recurrence, additional tumors, and/or death. None of the recruited individuals received neoadjuvant therapy, and all breast cancer cases were reported as sporadic. Samples collected included PT, UM (both distal [UMD, 1.5–5 cm] and proximal [UMP, at least 1 cm away from PT]), and SK, from the Oncology Centre in Bydgoszcz. These samples were stored at −80°C, and tumor presence and normal histology of uninvolved margins and skin samples were confirmed microscopically. Histological subtypes were assessed according to AJCC guidelines [9,64].

4.1.3.1. Transcriptome Study (Paper III)

This study focused on 83 breast cancer patients with unfavorable outcomes. Most patients underwent BCS (n=68), with fewer undergoing mastectomy (n=13). A total of 242 samples, including PT, UMD, and UMP, were analyzed after excluding outliers. Samples from two distinct tumor sites (PT1 and PT2) were included for two patients with multifocal primary tumors.

4.1.3.2. Variant Study (unpublished findings, manuscript under review)

The cohort for this study consisted of 77 breast cancer patients with unfavorable outcomes, primarily undergoing BCS (n=63) (Breast Cancer Adverse Prognoses cohort, BCAP cohort). A total of 231 samples, including matched PT, UMP (referred to as UM), and SK, were analyzed, with some UM samples taken at a further distance from the PT (UMD) included in later analyses. SK samples, as blood samples were unavailable, served as references to differentiate between post-zygotic and germline variants due to the unavailability of blood samples.

The cohorts for the Transcriptome Study (Paper III) and the Variant Study (unpublished findings, manuscript under review) partially overlapped, collectively including a total of 83 breast cancer patients with unfavorable prognoses. These patients were analyzed across the two studies to investigate distinct molecular aspects of breast cancer progression.

4.1.4. Reportedly Sporadic Breast Cancer Patients Selected Without Prognosis Criteria (unpublished findings, manuscript under review)

The BCUS (Breast Cancer Un-Selected) cohort comprised 49 sporadic breast cancer patients recruited without specific prognosis-related criteria, representing 5.25% of the total 933 breast cancer donors in the biobank. Most patients underwent BCS (n=31) rather than mastectomy (n=18). Among these 49 patients, 5 experienced recurrences, and 3 died within two years post-surgery; however, the follow-up period for this cohort (approximately 2 years) was considerably shorter than that of the breast cancer patients with adverse outcomes. A total of 147 samples, including PT, UM, and BL, were analyzed. UM samples were collected at least 1 cm away from the PT, and their normal histology was confirmed by two independent pathologists.

4.1.5. Control Patients

Paper II: Normal mammary gland samples from 26 age-matched women undergoing breast reduction surgery, with no history of cancer, served as controls. Histological evaluations confirmed tissue normalcy by two independent pathologists.

Paper III: Fifty-three individuals undergoing breast reduction surgery, with no personal or familial history of cancer, served as controls (CTRL). Samples were collected at the Karolinska Institute and the University Clinical Centre in Gdańsk, and histology was confirmed by dedicated pathologists.

Unpublished findings (manuscript under review): Fifteen individuals undergoing breast reduction surgery, with no personal or familial history of cancer, formed the Reduction Mammoplasty (RM) cohort and served as controls. Paired normal UM and BL samples were collected at the University Clinical Centre in Gdańsk, with histological evaluations confirming tissue normalcy.

4.2. Technologies

4.2.1. SNP Arrays (Paper II)

Single Nucleotide Polymorphism (SNP) arrays, developed in the 1990s, are used for genotyping and detecting genetic variations, including SNPs and Copy Number Variations (CNVs) across the genome. DNA is fragmented and labeled with fluorescent dyes before being hybridized onto a chip with probes designed for specific SNP sites. Post-hybridization, the chip is scanned to collect fluorescence intensity data, which is then analyzed to determine SNP genotypes. The intensity of the signal also helps identify CNVs, indicating heterozygous or homozygous alleles. SNP arrays are valued for their high-throughput, cost-effective, and reliable nature in genetic research [65]. In Paper II, SNP arrays identified recurrent genetic aberrations in paired PT and UM samples from breast cancer patients, as well as normal mammary tissue from age-matched controls.

4.2.2. Targeted DNA Sequencing (Paper II).

Although SNP arrays focus on the most common genetic variants, they capture only a small, preselected subset of all potential variations. Targeted DNA sequencing focuses on specific regions of the genome, such as genes or loci known to be involved in diseases, making it an efficient method for identifying relevant genetic variations. This technique involves using capture probes or primers to selectively enrich and amplify targeted genomic regions before sequencing. It provides higher depth of sequencing and sensitivity compared to whole genome sequencing, detecting low-frequency alterations and somatic variants with greater accuracy. This approach is particularly useful for studying genes with known disease associations, enabling rapid and accurate genetic analysis for diagnostic, prognostic, and therapeutic purposes [66,67]. In Paper II, targeted DNA

sequencing was employed to identify post-zygotic and germline variants in UM, BL, and PT samples from sporadic breast cancer patients.

4.2.3. Targeted RNA Sequencing (Paper III)

Targeted RNA Sequencing (RNA-seq) involves selectively sequencing specific RNA transcripts of interest rather than the entire transcriptome. This method uses capture probes or primers to enrich and amplify targeted RNA regions before sequencing. Total RNA or Messenger RNA (mRNA) is reverse-transcribed into Complementary DNA (cDNA), which is then enriched for specific genes and sequenced. This approach allows for high sensitivity in detecting low-abundance transcripts and rare isoforms, while also being cost-effective and reducing computational demands compared to whole transcriptome sequencing. Targeted RNA-seq is especially useful for studying gene expression and transcript variants associated with specific diseases or biological processes. However, because it concentrates on a predefined set of genes, targeted RNA-seq may introduce bias and potentially overlook important transcripts outside the target regions, limiting the discovery of novel transcripts and alternative splicing events [68,69]. In Paper III, a customized RNA panel was used to differentiate malignant from non-malignant breast samples and to identify a pre-tumorous state in normal mammary gland tissue.

4.2.4. Whole Exome Sequencing (WES) (unpublished findings, manuscript under review)

Targeted DNA sequencing is an efficient and cost-effective method for investigating genetic alterations, but it is limited to specific regions of the genome rather than covering the entire set of coding regions. In contrast, Whole Exome Sequencing (WES) targets the coding regions of the genome, focusing on exons where many disease-causing variants are found. DNA is fragmented, hybridized with probes specific to exonic regions, and sequenced to generate millions of short reads. These reads are aligned to a reference genome, and variants are identified and filtered to differentiate between germline and somatic origins. Variants are then annotated to assess their functional impact and relevance to disease, with comparisons made to genetic databases for interpretation. WES is cost-effective compared to whole genome sequencing and provides valuable insights into genetic disorders, potential therapeutic targets, and disease mechanisms. However, by focusing on exonic regions, WES may miss structural variants, large deletions, or duplications. Additionally, sophisticated bioinformatics tools are required for accurate variant calling and

interpretation, which can complicate data analysis [70,71]. Here, WES detected pathogenic germline and low-frequency post-zygotic variants in UM samples from breast cancer patients.

4.2.5. Duplex Sequencing (Papers II and unpublished findings, manuscript under review)

WES provides a comprehensive view of the protein-coding regions of the genome and is a cost-effective approach. However, it may miss low-frequency mutations due to higher error rates and the lack of strand-specific error correction, often requiring follow-up studies to verify rare variants. Duplex sequencing is an advanced method that reduces sequencing errors by independently analyzing both original DNA strands. DNA fragments are tagged with Unique Molecular Identifiers (UMIs) at both ends. These tagged fragments are sequenced, and reads are paired based on UMIs to create consensus sequences for each strand. This method enhances error correction and improves variant detection, particularly for rare mutations and low-frequency variants. Duplex sequencing is especially valuable in cancer research for detecting low-frequency somatic variants. Nevertheless, the technique is more complex and costly due to the additional steps of adapter ligation, high-depth paired-end sequencing, and the need for sophisticated bioinformatics for error correction [72,73]. In Paper II and the manuscript under review (unpublished findings), duplex sequencing was employed to identify low-level subclonal variants in selected genes.

4.3. Challenges and Solutions

Exploring the oncogenic potential of seemingly normal mammary gland tissue from breast cancer patients across multiple patient cohorts and control groups involves several significant challenges. These challenges affect the validity and reliability of the findings and include variability in sample collection protocols, the need for robust and age-matched control cohorts, and the sensitivity of the utilized methods. This section discusses these challenges and the methodological solutions implemented to address them, thereby enhancing the rigor and clarity of the studies.

4.3.1. Different Sample Collection Protocols

Breast cancer patients were also recruited using various protocols established prior to the biobanking project, leading to differences in the collection of uninvolved mammary gland samples. These protocols varied in terms of the distances from the primary tumor at which samples were obtained, resulting in a diverse array of tissue samples. This variability introduces complexity into the analysis, as different sampling distances can affect the interpretation of results. To mitigate this issue, we provided detailed explanations and illustrations of these differences in the “Cohorts:

Patient Selection and Samples Studied" section and in each publication. Despite this variability, the diversity in sampling distances might enhance the robustness of our findings, as alterations in transcriptomic profiles and genetic variants were observed not only in cancer-adjacent tissue (1 cm) but also at greater distances from the primary lesions (Papers II and III, and manuscript under review). This broader sampling range helps capture a more comprehensive picture of potential oncogenic changes.

4.3.2. Necessity for Robust, Age-Matched Control Cohorts

Controls for our studies were individuals without a personal or familial history of cancer. However, these control individuals were not always age-matched with the breast cancer patients (Paper III and manuscript under review). Age-matching controls are challenging because individuals opting for cosmetic surgical procedures, who usually serve as controls, are typically younger. Breast cancer diagnosis is relatively rare in younger women, with only about one in eight invasive breast cancers diagnosed in women under the age of 45 [62]. Furthermore, recruiting healthy controls is complicated by the high lifetime risk of breast cancer, with approximately 13% of women expected to develop the disease, and the exact onset of carcinogenesis remaining uncertain [62,63]. Therefore, in Papers II and III, and the manuscript under review, normal mammary glands were sampled from individuals without cancer history undergoing plastic surgery, providing the most appropriate available control samples from a biological perspective.

4.3.3. Effectiveness and Sensitivity of Utilized Methods

To investigate transcriptomic alterations among UM, PT, and CTRL samples, a custom RNA sequencing panel comprising 634 genes associated with breast cancer and related processes such as epithelial-to-mesenchymal transition, cell death, and apoptosis was employed in Paper III. This panel also included genes from the AIMS and PAM50 predictors (74,75), which classify breast tumors into molecular subtypes. Using a gene panel focused on a predefined set of genes introduces potential bias and may not capture the complete transcriptomic landscape. This limitation could result in missing important pathways involved in the disease. To address this issue, we validated the custom RNA sequencing panel's effectiveness by comparing it to external datasets of full transcriptome and custom RNA-seq panel data from the same cohort of 18 breast cancer patients, collected and processed similarly to the main dataset. This benchmarking confirmed the panel's

ability to capture critical transcriptomic information, ensuring it consistently produces valid results across different breast cancer samples.

Targeted DNA sequencing (Paper II) and WES (manuscript under review) were initially used to identify post-zygotic variants in paired UM and PT samples from breast cancer patients, as well as UM samples from control individuals. BL or SK samples were used as reference samples to differentiate between post-zygotic and germline variants. However, detecting low-frequency variants in heterogeneous UM and PT samples with standard NGS methods presents challenges due to sequencing depths (typically 100-200x for WES or 500-1000x for targeted DNA sequencing) and the inherent error rates of these technologies (approximately 0.1-1%). This can lead to false positives and complicate the identification of true low-frequency variants [66,74].

To address these challenges, selected variant cases were validated using independent methods such as Sanger sequencing and High-Resolution Melting (HRM). Sanger sequencing, or dideoxy sequencing, is known for its accuracy and is suitable for sequencing individual genes and validating variants. It involves DNA synthesis with chain-terminating dideoxynucleotides, resulting in fragments of varying lengths that are separated by capillary electrophoresis and identified based on fluorescence emission [75]. While Sanger sequencing is renowned for its precision, it is relatively slow and labor-intensive compared to modern high-throughput sequencing methods. Moreover, its sensitivity is limited: detecting variants with a Variant Allele Frequency (VAF) of less than 10% is challenging, and identifying variants with a VAF of less than 5% is virtually impossible. HRM is a post-PCR technique that identifies variations based on DNA melting behavior, offering high sensitivity and cost-effectiveness. Differences in DNA sequences cause variations in melting temperature (Tm), which are detected by the HRM analysis, allowing for the identification of mutations, polymorphisms, and epigenetic differences. HRM is advantageous due to its high sensitivity, specificity, and cost-effectiveness, as well as its ability to analyze multiple samples quickly without the need for labeled probes [76]. However, HRM is an indirect method and requires high-quality DNA and expertise to interpret the results accurately.

To further address these challenges duplex sequencing (Papers II and manuscript under review) was employed. This ultra-deep sequencing approach significantly increases coverage (up to thousands of times), enhancing sensitivity for detecting low-frequency variants. Duplex sequencing reduces sequencing errors by independently tracking both original strands of the DNA

molecule. True variants will appear on both strands, allowing for accurate differentiation from sequencing errors [72,73]. Duplex sequencing confirmed the presence of previously identified low-frequency pathogenic variants and revealed new extremely low-frequency pathogenic variants in UM samples.

V. SUMMARIES OF PUBLICATIONS

5.1. Paper I – Filipowicz et al.

5.1.1. Introduction

Combining multiple samples from a large number of distinct patients and control individuals, along with comprehensive, long-term clinical follow-up data, can significantly improve translational research. More than 90% of cancer cases are not attributed to inherited genetic alterations; instead, the accumulation of post-zygotic variants occurring after fertilization has been hypothesized to contribute to cancer predisposition [77–79]. Additionally, the mosaic loss of chromosome Y in the leukocytes of aging men—representing the most common post-zygotic variant in blood samples—is associated with earlier mortality and morbidity, including multiple cancer diagnoses [80–83]. The collection of histologically controlled non-tumorous tissues and blood samples, in addition to tumor samples from patients and unrelated healthy individuals, is crucial for genetic and proteomic analyses. This publication outlines the development of a specialized biobank designed to support cancer research by providing high-quality human tissue and blood samples, along with detailed patient questionnaires for comprehensive data. This biobanking effort aims to systematically explore the contribution of post-zygotic genetic variations in normal tissues to cancer predisposition.

5.1.2. Results and Discussion

Samples were collected from patients with breast, colorectal, prostate, bladder, and pancreatic cancers, as well as healthy male controls for loss-of-chromosome-Y studies. Recruitment occurred across five clinical centers over nearly two years, resulting in 1,711 donors and 23,226 samples. Breast carcinoma was the predominant diagnosis, with 933 donors treated either with mastectomy or breast-conserving surgery, and detailed demographic and clinical information was collected. A substantial portion of samples came from normal tissue margins at various distances from the corresponding primary tumors, providing a unique resource for cancer research. The collection process, involving detailed pathology and histopathology, required about 2,800 working hours.

Standardized protocols ensured the quality and reliability of these samples, which are essential for reproducible and accurate genetic studies. By offering a rich repository of biological materials, the biobank aims to facilitate the identification of new biomarkers and therapeutic targets, ultimately advancing personalized cancer treatment and improving patient outcomes. Dedicated software (MABData1 and MABData2) was designed and implemented to support the decentralized biobanking approach, allowing efficient data management and ensuring compliance with data safety standards. This resource supports diverse "omics" studies and has the potential to enhance the understanding of genetic variations involved in cancer predisposition.

5.2. Paper II – Kostecka A. et al.

5.2.1. Introduction

Breast cancer, which affects 24% of women globally and is a leading cause of cancer-related female deaths, mostly arises without inherited mutations in high-penetrance genes like *BRCA1* or *BRCA2* (85-90% of cases) [3,4,84]. High-throughput genomics has classified breast cancer into four subtypes and identified somatic driver mutations in key genes such as *PIK3CA* and *TP53* [7,8,32,55,56]. Traditionally, these mutations were studied in tumors, overlooking the mutational landscape in normal mammary tissue, which is hormonally stimulated and prone to DNA damage [11,85–87]. This study screened for subclonal somatic pathogenic alterations in the normal mammary gland tissue of sporadic cancer patients, particularly post-BCS. The findings reveal frequent structural chromosomal aberrations and pathogenic point variants in crucial breast cancer genes in the histologically normal tissue left after BCS. These genetic alterations in preserved normal tissue suggest a link with recurrence risk and implications for future treatment, highlighting the need for thorough genetic screening in breast cancer management.

5.2.2. Results and Discussion

A total of 204 UM, PT, SK, and BL samples were collected from 52 reportedly sporadic breast cancer patients, treated mostly with BCS. Normal mammary gland samples were also collected from 26 age-matched control individuals undergoing breast reduction surgeries. SNP arrays were implemented to analyze chromosomal rearrangements and to detect DNA CNAs and Loss Of Heterozygosity (LOH). Hierarchical clustering revealed distinct PT-only and control-only clusters, with PTs exhibiting significant differences in CNAs (Wilcoxon test, $p = 0.0094$). Surprisingly, control samples showed greater heterogeneity. Recurrent chromosomal aberrations, such as losses

at 1p, 16p11.2, and 9p21.3, and gains at 3q25.3, 4q13.1, 8q, and 20q, were identified in UMs. LOH at chromosome 8p, associated with poor breast cancer outcomes [88], was present in UMs, PTs, and controls. *ERBB2* gains were detected exclusively in PTs, except in one control sample. Targeted DNA sequencing of UM, BL, and PT samples from breast cancer patients found heterozygous constitutional pathogenic variants in 7.7% (4/52) of cases. After excluding individuals with germline pathogenic variants, the analysis focused on 48 sporadic breast cancer patients, identifying 15 somatic pathogenic variants in the normal mammary gland tissue of 19% (9/48) of patients. These variants affected genes related to tumor suppression, oncogenesis, cell death regulation, DNA repair, translation, gene expression, and chromatin remodeling. Ultra-deep duplex sequencing was implemented to enhance the sensitivity and accuracy of rare variant detection. *PIK3CA* and *TP53* were prominent driver genes, with 6 hotspot *PIK3CA* variants and 7 hotspot *TP53* variants in total, revealed either with targeted DNA sequencing or duplex sequencing. *PIK3CA* and *TP53* variants, prevalent in tumors [7,35], were detected at lower levels in normal tissue, suggesting potential secondary tumor sites. These findings underscore the importance of profiling normal tissue to elucidate disease origins, potentially enhancing treatment and clinical management. The study advocates for genetic and clinical surveillance of sporadic breast cancer patients post-surgery to improve personalized care.

5.3. Paper III – Andreou M. & Jąkalski M. et al.

5.3.1. Introduction

Breast cancer is a major health issue, ranking as the most common cancer globally in 2020, with 2.26 million cases, surpassing lung cancer incidence [89,90]. Enhanced mammographic screening and extensive educational efforts have facilitated early detection, aiding in the identification of breast carcinomas during asymptomatic phases. BCS is favored for its tissue preservation benefits, yet recurrence rates post-surgery remain substantial, implicating residual disease or alterations in unexcised mammary gland tissue [16,17,91]. Current therapeutic decisions rely on tumor and resection margin analyses, but emerging research underscores the prognostic potential of normal tissue [40,50,54,92]. Unlike prior studies focusing on cancerous tissues of patients selected without any criteria related to prognoses [42,45,46], this study aimed to investigate the transcriptomic landscape of uninvolved mammary gland tissue at various distances from the primary lesion in patients with adverse prognoses. Distinct gene expression patterns distinguish malignant from non-malignant tissues, and a potentially pre-tumorigenic environment emerges in apparently normal

tissue, associating with smaller tumors and poorer outcomes. The study underscores the significance of incorporating normal tissue analysis into breast cancer research for improved prognostication and therapeutic strategies.

5.3.2. Results and Discussion

The transcriptomic profiles of 242 PT and UM samples collected proximal (UMP) and distal (UMD) to the PT from 83 breast cancer patients who experienced unfavorable outcomes were analyzed. Patients included suffered from disease recurrence and/or the presence of a second, independent tumor and/or succumbed to the disease within 10 years post-original surgery. CTRL samples from 53 individuals undergoing reduction mammoplasty surgeries without a history of cancer were used as a reference group. Two independent pathologists examined tissue samples to identify cancerous areas in PT samples and confirm the normal histology of UMs and CTRLs. A custom panel comprising 634 genes associated with breast cancer progression and metastasis was utilized for expression profiling. The custom RNA-sequencing panel's ability to capture comprehensive information representative of the entire mammary tissue was validated using external datasets from 18 breast cancer patients. The results highlight a clear distinction between malignant (PT) and non-malignant (UMP, UMD, CTRL) tissues through Principal Component Analysis (PCA), revealing significant differences in expression profiles. Differential expression analysis showed the largest deregulation of genes when comparing PT to all non-malignant tissues, with fewer Differentially Expressed Genes (DEGs) when PTs were compared with controls or UMs separately. Functional annotation of DEGs linked to cancer-related pathways indicated aggressive tumor profiles and poor outcomes. PCA further revealed heterogeneity within non-malignant samples, with a subset of UMs forming a distinct group. AIMS and PAM50 gene expression-based classifiers, originally developed and used on full-blown tumors, corroborated the histopathological evaluation for all CTRL samples, while some UMs exhibited tumor-like features according to PAM50. Hierarchical clustering revealed four distinct clusters, with Cluster 4, enriched with UMs, exhibiting unique attributes and a down-regulated gene signature. This signature, named KAOS (for Keratins-Adhesion-Oncogenes-Suppresors), featured key cellular components encoding keratins, CDH1, CDH3, and EPCAM cell adhesion proteins, matrix metallopeptidases, oncogenes, tumor suppressors, along with crucial genes (*FOXA1*, *RAB25*, *NRG1*, *SPDEF*, *TRIM29*, and *GABRP*) having dual roles in cancer. Furthermore, Cluster 4 was significantly associated with clinical outcomes, showing smaller tumor sizes ($p=0.033$, Mann-Whitney U test), higher age

(p=0.025, Mann-Whitney U test), HER2-positive status (p=0.004265, Fisher's test), and a higher death status (p=0.04493345 and p=0.01512627, Fisher's test for UMD and UMP, respectively). Enrichment analyses showed deregulated pathways in Cluster 4, including PPAR signaling, regulation of lipolysis in adipocytes, and estrogen pathways. These findings suggest the presence of a pre-tumorigenic environment within histologically normal mammary tissue, indicating potential prognostic value and implications for patient management and personalized care.

5.4. Manuscript under review, unpublished findings

5.4.1. Introduction

Breast cancer represents 12.5% of global cancer diagnoses, with incidence rates rising by 0.5% annually [62,63]. Although awareness and early detection have led to a 42% reduction in mortality from 1989 to 2021, breast cancer remains a leading cause of death among women, with even low-risk stage I cases exhibiting a 15-20% recurrence risk after two decades [93,94]. While 5-10% of cases are hereditary, most are sporadic [3–5]. Recent research has shifted focus to the normal mammary gland tissue for early molecular detection of tumors, revealing that even histologically normal tissue from breast cancer patients often contains significant genomic alterations, particularly in the *PIK3CA* and *TP53* genes [40,45,46,50,54,60,92]. However, the clinical relevance of post-zygotic alterations in histologically normal mammary gland tissue of breast cancer patients remains unclear. This study screened UM and PT samples from sporadic breast cancer patients with poor outcomes and found that pathogenic post-zygotic variants in cancer-associated genes are prevalent in normal mammary tissue. These variants correlate with patient survival, highlighting the importance of molecular screening for better clinical management.

5.4.2. Results and Discussion

The genetic profiles of 378 samples of PT, UM, and BL or SK tissue from reportedly sporadic breast cancer patients were analyzed. Patients were stratified into two cohorts: 77 patients with adverse outcomes (BCAP cohort) and 49 patients without specific prognosis-related criteria (BCUS cohort), all from the same ethnic population. The BCAP group had poor outcomes, with patients experiencing recurrence or metastasis (n=40), developing a second tumor (n=18), or both (n=8), and/or succumbing to the disease (n=45) within 10 years (Table 1). Additionally, UM and BL samples from 15 individuals undergoing mammoplasty for non-cancer-related reasons served as a control group. Two pathologists confirmed cancerous areas in PT samples and verified the

normal histology of UM and SK tissues. Post-zygotic variants were filtered based on their truncating nature (nonsense and frameshift), annotation in the ClinVar/InterVar databases ("pathogenic", "likely pathogenic", "uncertain significance", or "conflicting interpretations of pathogenicity"), presence in the COSMIC database, and minor allele frequency (MAF); variants with MAF ≤ 0.001 across all gnomAD populations ("popmax") or not noted in the database (gnomAD v2.1.1) were included.

Table 1. Summarized clinicopathological features of breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) and the Breast Cancer Un-Selected (BCUS) cohorts.

	BCAP cohort	**BCUS cohort**
Number of individuals	77	49
Age (median, range)	62, 23-85	65, 37-84 p value = 0.082
Collected samples	238	147
Primary Tumor, PT	77	49
Uninvolved mammary gland, UM	77	49
Distal fragment of uninvolved mammary gland, UMD	7	-
Reference sample (whole peripheral blood, BL or skin, SK)	77	49
Histology		
Invasive ductal carcinoma, IDC	59	40
Invasive lobular carcinoma, ILC	3	4
IDC - ILC	6	1
other	9	4
Receptors		
Estrogen, ER (positive / negative / not available)	57 / 20	43 / 5 / 1
Progesterone, PR (positive / negative / not available)	43 / 34	44 / 4 / 1
HER2 (positive / negative / not available)	16 / 56 / 5	5 / 43 / 1
Subtype		

Luminal A	14	22
Luminal B	37	21
HER-2 enriched	9	2
Triple-negative breast cancer, TNBC	11	1
Not available	6	3
Follow-up information		
Recurrence (yes / no)	50 / 27	5 / 44
Second cancer (yes / no)	26 / 51	0 / 49
Death* (yes / no)	45 / 31	3 / 46

*Matched primary tumor (PT) and uninvolved mammary gland (UM, ≥1 cm) samples were collected from two breast cancer cohorts, i.e. 77 individuals characterized with adverse outcomes (BCAP cohort) and 49 individuals recruited without any pre-selection criteria related to prognosis (BCUS cohort). Whole peripheral blood (BL) or skin (SK) samples (if BL was not available) were collected as reference samples to distinguish between post-zygotic and germline variants. Distal UM samples (UMD, 1.5-3 cm from PT, median 2.35 cm), available for 7 BCAP patients, were included. *Death status refers to patients who succumbed to the disease (patient with ID BCAP61 died from non-oncological reasons).*

A total of 167 variants were identified in UM samples from 41 BCAP patients, compared to 56 variants in 24 BCUS patients and 10 variants in 7 RM individuals. Truncating nonsense and frameshift mutations (n=37) were exclusive to BCAP, many affecting tumor suppressor genes such as *KMT2C* [95], *PTEN* [96], *TBX3* [97], and *TP53* [7]. Missense variants were further evaluated using the REVEL score [98] (threshold 0.75) to predict pathogenicity. In the BCAP cohort, 29% (49/167) of variants were classified as pathogenic, including truncating (n=37) and missense variants (n=12) with in-silico evidence of pathogenicity (REVEL score ≥ 0.75). Notably, 24% of pathogenic BCAP variants were detected only in UM samples, absent from corresponding tumors. The BCUS cohort exhibited only 7 pathogenic variants (13%), with 43% of them exclusive to UM samples, though significantly fewer than in BCAP (Hypergeometric test, p=0.0008578).

Several of the identified pathogenic variants affected dosage-sensitive genes in BCAP, such as tumor suppressors (*KMT2C* [95], *PTEN* [96], *TBX3* [97], and *TP53* [7]*)* and oncogenes (*PIK3CA* [99], *AKT1* [100]). *PIK3CA* variants were present in both BCAP and BCUS, while *TP53* variants

(seven, including two recurrent variants) were exclusive to BCAP, suggesting a stronger role in breast cancer initiation. In contrast, BCUS samples contained pathogenic variants in genes like *SF3B1* [101], *HRAS* [102], *GNAS* [103], and *RUNX1* [104], with only the latter two being dosage-sensitive.

Duplex sequencing detected low-frequency (low as 1.34%) pathogenic *PIK3CA* and *TP53* variants in the UM samples of poor-prognosis patients. These variants, linked to aggressive cancer progression, were often observed exclusively in UM samples rather than in the primary tumor, indicating potential early tumorigenic processes. Notably, some BCAP patients had multiple pathogenic variants across cancer-related genes[105], suggesting potential synergistic effects contributing to disease severity. Selected pathogenic or likely pathogenic variants in the BCAP cohort were further validated using Sanger sequencing or High-Resolution Melting. An overview of identified pathogenic, likely pathogenic variants or variants of uncertain significance or conflicting interpretations with in-silico evidence of pathogenicity (REVEL score $\geq$ 0.75), identified in BCAP and BCUS cohorts is provided in Table 2.

Table 2. Pathogenic, likely pathogenic variants, and variants of uncertain significance or conflicting interpretations of pathogenicity with evidence for pathogenicity according to REVEL score (≥0.75), identified via whole exome sequencing in breast cancer patients according to the study's criteria.

Gene	Variant[a]	ClinVar[b]	(MAF) gnomAD[c]	REVEL score[d]	COSMIC ID[e]	AVSNP150[f]	Individual ID and UM sample VAF[g]	Cohort
AKT1	c.49G>A (p.Glu17Lys)	Pathogenic	0.000008872	0.51	ID=COSV62571334	rs121434592	BCAP32 (0.6%), BCAP66* (0.7%)	BCAP
CNOT9	c.259T>C (p.Ser87Pro)	Likely Pathogenic	n.a.	0.737	ID=COSV55299564	rs1057519956	BCAP26 (0.4%)	BCAP
ERBB2	c.2264T>C (p.Leu755Ser)	Likely Pathogenic	n.a.	0.86	ID=COSV54062780	rs121913470	BCAP53 (0.7%)	BCAP
GATA4	c.1078G>A (p.Glu360Lys)	Uncertain Significance	0.0001	0.757	ID=COSV100632768	rs368489876	BCAP47 (19%)	BCAP
GNAS	c.680A>G (p.Gln227Arg)	Pathogenic	n.a.	0.888	ID=COSV55671120	rs121913494	BCUS32 (0.3%)	BCUS
HRAS	c.182A>T (p.Gln61Leu)	Uncertain Significance	n.a.	0.839	ID=COSV54236656	rs121913233	BCUS45 (0.9%)	BCUS
KMT2C	c.10279C>T (p.Gln3427*)	n.a.	n.a.	not applicable	ID=COSV51484133	n.a.	BCAP20 (0.9%)	BCAP
MKI67	c.4991_4992del (p.Thr1664Argfs*7)	Pathogenic	0.0003	not applicable	ID=COSV64072397	rs145960091	BCAP03 (0.3%)	BCAP
PIK3CA	c.1258T>C (p.Cys420Arg)	Pathogenic	n.a.	0.788	ID=COSV55874020	rs121913272	BCUS45 (0.9%)	BCUS
PIK3CA	c.1624G>A (p.Glu542Lys)	Pathogenic	n.a.	0.439	ID=COSV55873227	rs121913273	BCAP56 (0.7%), BCAP45 (12%)	BCAP
PIK3CA	c.3140A>G (p.His1047Arg)	Pathogenic	0.00000891	0.455	ID=COSV55873195	rs121913279	BCAP15 (0.08%), BCAP31* (19%), BCAP36 (0.3%), BCAP53* (0.7%), BCUS39 (28%)	BCAP, BCUS
PIK3CA	c.3140A>T (p.His1047Leu)	Pathogenic	0.00000891	0.359	ID=COSV55873401	rs121913279	BCAP54* (0.5%), BCUS49 (0.6%)	BCAP, BCUS
POMGNT1	c.1814G>A (p.Arg605His)	Pathogenic/Likely Pathogenic	0.00001776	0.871	ID=COSV64340932	rs267606962	BCAP31 (10%)	BCAP
PTCH1	c.2714G>A (p.Arg905His)	Conflicting interpretations	0.00003266	0.881	ID=COSV59488865	rs764310195	BCAP20 (11%)	BCAP
PTEN	c.388C>T (p.Arg130*)	Pathogenic	0.00003266	not applicable	ID=COSV64288463	rs121909224	BCAP15 (0.7%)	BCAP
RUNX1	c.497G>A (p.Arg166Gln)	Pathogenic	n.a.	0.962	ID=COSV55867644	rs1060499616	BCUS47 (0.5%)	BCUS
SF3B1	c.1996A>G (p.Lys666Glu)	Likely Pathogenic	n.a.	0.685	ID=COSV59206172	rs754688962	BCUS48 (18%)	BCUS
TBX3	c.371_372insTGGT (p.Ile125Profs*14)	n.a.	n.a.	not applicable	ID=COSV57471668	n.a.	BCAP44 (12%)	BCAP
TP53	c.151G>T (p.Glu51*)	Pathogenic	n.a.	not applicable	ID=COSV52694020	n.a.	BCAP58* (16%)	BCAP
TP53	c.227del (p.Ala76Aspfs*47)	n.a.	n.a.	not applicable	ID=COSV52728465	n.a.	BCAP54 (0.5%)	BCAP
TP53	c.329G>C (p.Arg110His)	Pathogenic	n.a.	0.593	ID=COSV52668419	rs11540654	BCAP45 (0.8%)	BCAP

TP53	c.637C>T (p.Arg213*)	Pathogenic	n.a.	not applicable	ID=COSV52665560	rs397516436	BCAP01* (0.8%), BCAP48* (0.6%)	BCAP
TP53	c.711G>A (p.Met237Ile)	Pathogenic	0.00005437	0.923	ID=COSV52661887	rs587782664	BCAP15 (0.8%)	BCAP
TP53	c.1024C>T (p.Arg342*)	Pathogenic	n.a.	not applicable	ID=COSV52665487	rs730882029	BCAP38 (0.5%), BCAP47 (0.7%)	BCAP
TP53	c.1025G>C (p.Arg342Pro)	Pathogenic/Likely Pathogenic	n.a.	not applicable	ID=COSV52690857	rs375338359	BCAP57 (0.6%)	BCAP

*[a]Variant annotation provided for the basic isoform of the transcript. [b]Pathogenicity classification according to the ClinVar database. [c]Minor allele frequency (MAF) across all gnomAD populations (gnomAD v2.1.1). [d]REVEL score. [e]ID of the variant in the COSMIC (Cosmic_95 coding) database. [f]rsIDs in dbSNP build 150. [g]Individual ID and Variant Allele Frequency (VAF) for UM samples. BCAP – Breast Cancer Adverse Prognosis, BCUS – Breast Cancer Un-Selected, n.a.- not available. *variants were also detected in selected patients' distal uninvolved mammary gland sample (UMD).*

All breast cancer cases in this study were classified as sporadic based on family history, though genetic testing results were unavailable before recruitment. To assess germline pathogenic variants, BL or SK samples were analyzed across cohorts. In the BCAP cohort, 14 of 77 individuals (18%) carried pathogenic variants in known high or moderate penetrance breast cancer genes [106]. These included *BRCA1* (c.4186C>T [p.Gln1396*], c.4689C>G [p.Tyr1563*], c.5179A>T [p.Lys1727*], c.5266dup [p.Gln1756Profs74]), *BRCA2* (c.5645C>A [p.Ser1882], c.6591_6592del [p.Glu2198Asnfs4], c.9382C>T [p.Arg3128]), *PALB2* (c.172_175del [p.Gln60Argfs7], c.1671_1674del [p.Ile558Lysfs2]), and *RAD50* (c.3233_3236del [p.Lys1079Valfs28]). *BRCA1* c.5266dup (p.Gln1756Profs74) and *PALB2* c.172_175del (p.Gln60Argfs*7) were recurrent, observed in four and two unrelated individuals, respectively. The 18% mutation frequency in BCAP surpasses previous reports (~10%) and may reflect the aggressive nature of these cases [7,55,92].

Four BCAP individuals (29%) with germline pathogenic variants also harbored post-zygotic pathogenic variants in known cancer-related genes in their UM samples. The germline variants in these cases were found in *BRCA1* (four cases) and *RAD50* (one case), while the corresponding post-zygotic variants were identified in *PIK3CA* or *TP53*. In the BCUS cohort, only a single patient carried a pathogenic *BRCA1* variant (c.5266dup [p.Gln1756Profs*74]). No germline pathogenic or likely pathogenic variants in breast cancer-associated genes were detected in the control group.

Kaplan-Meier survival analysis revealed that patients with recurrence (n=53) had significantly lower survival probabilities compared to those without recurrence (n=72) across the BCAP and BCUS cohorts (log-rank test, p=0.00017), with a hazard ratio of 2.44 (95% CI: 1.07-5.54, p=0.0337), indicating more than twice the risk of death. Given the shorter follow-up period for BCUS (2 years) versus BCAP (10 years), early outcomes were assessed within the first 24 months post-diagnosis. During this period, recurrence patients (n=53) had significantly lower survival probabilities than non-recurrence patients (n=71) (log-rank test, p<0.0001), with a hazard ratio of 4.85 (95% CI: 1.4-16.25, p=0.0105), suggesting over four times the risk of death. BCAP patients had significantly more recurrence events than BCUS patients within this timeframe (Fisher's exact test, p=0.005488). Within the BCAP cohort, recurrence patients (n=48) had lower survival probabilities throughout the follow-up period compared to non-recurrence patients (n=28) (log-rank test, p=0.015), with a similar trend observed in the first 24 months (log-rank test, p=0.0088).

Survival probabilities differed significantly across groups (log-rank test, p=0.024), indicating that the presence and type of pathogenic variants (germline or post-zygotic), along with recurrence status, strongly influence patient outcomes (Figure 4). Patients with pathogenic germline variants (green) had the shortest recurrence-free survival, with most recurrences occurring within the first 60 months. In contrast, patients with pathogenic post-zygotic variants in breast cancer-specific genes (blue) experienced recurrences less frequently and over a longer follow-up period, suggesting a moderate but significant effect on recurrence risk. Patients without pathogenic germline or post-zygotic variants (yellow) showed intermediate survival outcomes.

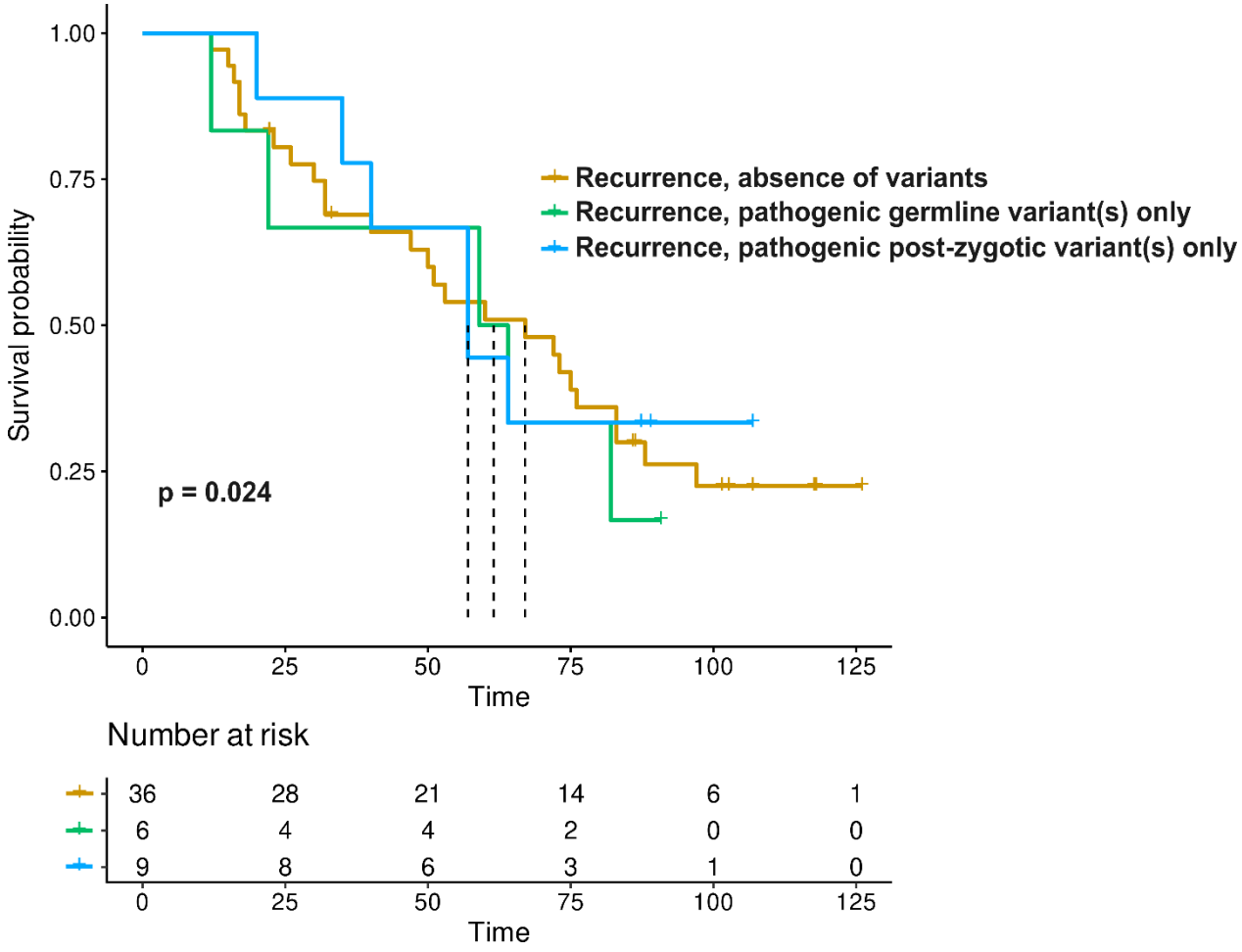

Figure 4. Kaplan-Meier survival curves of breast cancer patients with pathogenic variants and recurrent disease. The curves represent survival probabilities for different groups of patients from the BCAP cohort (breast cancer patients with adverse prognoses) and the BCUS cohort (breast cancer patients without specific prognosis criteria), stratified by the presence of recurrent disease and/or pathogenic germline or post-zygotic variants in breast cancer-specific

genes. Survival time was measured from the date of diagnosis to death or the end of the follow-up period (10 years for BCAP and 2 years for BCUS). The x-axis represents time in months, and the y-axis represents the probability of survival. Vertical ticks on the curves indicate censored death events.

PIK3CA and *TP53* variants co-occurred in three BCAP patients, suggesting a synergistic role in cancer progression. Notably, a single BCAP patient had concurrent pathogenic variants in *PIK3CA*, *TP53*, and *PTEN*, highlighting the complex interplay of oncogenic and tumor-suppressive pathways. These combined variants likely contribute to a more aggressive disease course, underscoring the need for comprehensive genetic profiling. Post-zygotic variants in *TP53* and *PIK3CA* have been observed in breast tumors, but their effects on normal mammary tissue are less clear [107,108]. These alterations, which accumulate with age and hormonal changes [11,109], may represent early pre-cancerous changes that could lead to cancer if activated by factors like aging or injury. All BCAP patients experienced adverse outcomes within 10 years, indicating the significant impact of these genetic variations on prognosis.

While current diagnostics focus on germline variants [106], our study shows that post-zygotic variants, like those in *PIK3CA* and *TP53*, are often found in normal tissue after breast-conserving surgery, with allele frequencies ranging from 0.03 to 0.28, suggesting they may contribute to recurrence or metastasis. Our findings show that pathogenic post-zygotic variants in breast cancer-associated genes are more prevalent in normal mammary tissues of patients with adverse outcomes, such as recurrence or metastasis, compared to those without specific prognosis criteria or control individuals. Monitoring these patients for nearly a decade allowed us to directly link these variants to clinical outcomes. These alterations were strongly associated with disease progression, particularly recurrence, indicating an increased risk of aggressive cancer before clinical symptoms appear. This underscores the need for expanded genetic screening and enhanced surveillance to improve personalized management, especially for patients with poor prognoses

VI. CONCLUSIONS

Based on the findings from the four studies encompassed in this doctoral work, the following conclusions can be drawn:

6.1. Paper I – Filipowicz N. et al.

- The establishment of a comprehensive biobank of histologically controlled, non-tumorous mammary gland samples, alongside tumor, blood, and skin samples, was successfully achieved.
- This resource addressed significant challenges such as patient heterogeneity and small cohort sizes, creating a foundation for high-quality "omics" studies.
- By defining stringent sampling protocols, the study ensured reliable differentiation between normal and pathological tissue, improving the reproducibility of cancer research.

6.2. Paper II – Kostecka A. et al.

- Ultra-high sensitivity methods successfully identified structural rearrangements and somatic pathogenic variants in breast cancer-related genes within histologically normal mammary tissue.
- Subtle molecular alterations, including low-frequency pathogenic variants in genes such as *PIK3CA* and *TP53*, were detected, highlighting the potential role of normal tissue in cancer progression.
- These findings underscore the importance of genetic profiling in apparently normal tissues for improved understanding of early oncogenic changes.

6.3. Paper III – Andreou M. & Jąkalski M. et al.

- Transcriptomic profiling of histologically normal mammary tissues revealed distinct expression patterns associated with poor clinical outcomes, such as smaller tumors and HER2-positive status.
- The study identified a potential pre-tumorigenic environment in non-tumorous tissues, emphasizing its clinical relevance for prognostication.
- The KAOS gene signature was defined, offering potential biomarkers for identifying early-stage cancer risks and refining patient management strategies.

6.4. Unpublished findings, manuscript under review (preprint).

- Pathogenic post-zygotic variants were detected in non-tumorous tissues, correlating strongly with adverse clinical outcomes, including recurrence, metastasis, and mortality.
- Key genes such as *PIK3CA*, *TP53*, and *AKT1* emerged as central to early tumorigenic processes, with their variants linked to aggressive disease progression and poorer survival outcomes in patients with poor prognoses.
- This study demonstrated that such pathogenic variants are more frequently observed in patients with adverse outcomes than those without specific prognosis-related criteria, underscoring their value as potential prognostic biomarkers.
- The findings support the use of molecular screening of normal mammary tissues to identify high-risk patients, enabling targeted interventions and improving clinical management for those at greatest risk of recurrence or mortality.

6.5. General Conclusions

- This work advances the understanding of somatic mosaicism in sporadic breast cancer, demonstrating that histologically normal mammary gland tissue harbors molecular changes that are clinically relevant, associating a gene expression signature and deleterious post-zygotic variants with tumor size, increased mortality and survival.
- The findings emphasize the need for high-sensitivity detection methods, comprehensive sampling protocols, and the integration of normal tissue analysis into routine cancer diagnostics and research, potentially through the use of advanced molecular technologies and systematic tissue sampling approaches.
- These studies provide a robust foundation for future efforts to refine patient stratification, prognosis, and personalized treatment strategies by leveraging early molecular alterations in normal tissues, supported by advanced data analysis techniques and the identification of potential diagnostic markers.

These conclusions collectively illustrate the successful fulfillment of the research aims and the significant contributions of this work to breast cancer research.

6.6. Future Perspectives

Advancing high-sensitivity detection methods is crucial to identifying subtle cellular changes that might exist well before tumor detection by conventional techniques. State-of-the-art sequencing,

single-cell analysis, and duplex sequencing technologies provide the precision needed to detect early molecular changes with higher accuracy at lower thresholds. These technologies offer a deeper understanding of tumorigenesis and reveal a more nuanced view of the genomic landscape within normal mammary tissues adjacent to tumors.

Studying these alterations across patients with varied clinical outcomes is essential to linking changes in normal tissue with patient prognosis. Stratifying patients by the presence or absence of molecular alterations could reveal correlations with disease progression, treatment efficacy, and survival, enabling tailored treatment strategies that consider the molecular profile of normal tissue rather than focusing solely on the tumor.

Longitudinal studies that track these molecular changes over time in patients with differing outcomes are also invaluable. Such studies could identify early biomarkers for relapse or resistance, providing clinicians with actionable insights for timely intervention. This proactive approach could improve relapse prediction and inform adjuvant therapy decisions, potentially enhancing long-term outcomes for breast cancer patients.

In conclusion, advancing sensitive detection methods and rigorously investigating molecular alterations in diverse patient groups are essential for translating these findings into clinical practice. These efforts hold promise for refining cancer diagnostics, enhancing treatment precision, and ultimately improving patient care through more targeted, effective therapeutic options.

VII. BIBLIOGRAPHY

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74:229–63. https://doi.org/10.3322/caac.21834.

[2] Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, McCullough ML, Patel AV, Ma J, Soerjomataram I, Flanders WD, Brawley OW, Gapstur SM, Jemal A. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin 2018;68:31–54. https://doi.org/10.3322/caac.21440.

[3] Coughlin SS. Epidemiology of Breast Cancer in Women. Adv Exp Med Biol 2019;1152:9–29. https://doi.org/10.1007/978-3-030-20301-6_2.

[4] Kleibl Z, Kristensen VN. Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management. Breast Edinb Scotl 2016;28:136–44. https://doi.org/10.1016/j.breast.2016.05.006.

[5] Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol Off J Eur Soc Med Oncol 2015;26:1291–9. https://doi.org/10.1093/annonc/mdv022.

[6] Makki J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clin Med Insights Pathol 2015;8:CPath.S31563. https://doi.org/10.4137/CPath.S31563.

[7] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70. https://doi.org/10.1038/nature11412.

[8] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Van De Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale A-L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci 2001;98:10869–74. https://doi.org/10.1073/pnas.191367098.

[9] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. CA Cancer J Clin 2017;67:93–9. https://doi.org/10.3322/caac.21388.

[10] Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem Cells and the Differentiation Hierarchy in Mammary Gland Development. Physiol Rev 2020;100:489–523. https://doi.org/10.1152/physrev.00040.2018.

[11] Macias H, Hinck L. Mammary gland development. Wiley Interdiscip Rev Dev Biol 2012;1:533–57. https://doi.org/10.1002/wdev.35.

[12] Muschler J, Streuli CH. Cell-Matrix Interactions in Mammary Gland Development and Breast Cancer. Cold Spring Harb Perspect Biol 2010;2:a003202–a003202. https://doi.org/10.1101/cshperspect.a003202.

[13] Hassiotou F, Geddes D. Anatomy of the human mammary gland: Current status of knowledge. Clin Anat 2013;26:29–48. https://doi.org/10.1002/ca.22165.

[14] Moloney BM, O'Loughlin D, Abd Elwahab S, Kerin MJ. Breast Cancer Detection—A Synopsis of Conventional Modalities and the Potential Role of Microwave Imaging. Diagnostics 2020;10:103. https://doi.org/10.3390/diagnostics10020103.

[15] Wang L. Early Diagnosis of Breast Cancer. Sensors 2017;17:1572. https://doi.org/10.3390/s17071572.

[16] Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. Lancet Lond Engl 2021;397:1750–69. https://doi.org/10.1016/S0140-6736(20)32381-3.

[17] Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA 2019;321:288–300. https://doi.org/10.1001/jama.2018.19323.

[18] Scheepens JCC, Veer LV 'T, Esserman L, Belkora J, Mukhtar RA. Contralateral prophylactic mastectomy: A narrative review of the evidence and acceptability. The Breast 2021;56:61–9. https://doi.org/10.1016/j.breast.2021.02.003.

[19] Beca F, Polyak K. Intratumor Heterogeneity in Breast Cancer. In: Stearns V, editor. Nov. Biomark. Contin. Breast Cancer, vol. 882, Cham: Springer International Publishing; 2016, p. 169–89. https://doi.org/10.1007/978-3-319-22909-6_7.

[20] Meacham CE, Morrison SJ. Tumour heterogeneity and cancer cell plasticity. Nature 2013;501:328–37. https://doi.org/10.1038/nature12624.

[21] Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481:306–13. https://doi.org/10.1038/nature10762.

[22] Esparza-López J, Escobar-Arriaga E, Soto-Germes S, Ibarra-Sánchez MDJ. Breast Cancer Intra-Tumor Heterogeneity: One Tumor, Different Entities. Rev Investig Cl�nica 2017;69:96. https://doi.org/10.24875/RIC.17002177.

[23] Nowell PC. The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. Science 1976;194:23–8. https://doi.org/10.1126/science.959840.

[24] Hanson FB, Tier C. A stochastic model of tumor growth. Math Biosci 1982;61:73–100. https://doi.org/10.1016/0025-5564(82)90097-9.

[25] Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta BBA - Rev Cancer 2017;1867:151–61. https://doi.org/10.1016/j.bbcan.2017.01.003.

[26] Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, Caldas C, Califano A, Doherty M, Elsner M, Esteller M, Fitzgerald R, Korbel JO, Lichter P, Mason CE, Navin N, Pe'er D, Polyak K, Roberts CWM, Siu L, Snyder A, Stower H, Swanton C, Verhaak RGW, Zenklusen JC, Zuber J, Zucman-Rossi J. Toward understanding and exploiting tumor heterogeneity. Nat Med 2015;21:846–53. https://doi.org/10.1038/nm.3915.

[27] Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer 2006;6:924–35. https://doi.org/10.1038/nrc2013.

[28] Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, Curtis C. A Big Bang model of human colorectal tumor growth. Nat Genet 2015;47:209–16. https://doi.org/10.1038/ng.3214.

[29] Gould SJ, Eldredge N. Punctuated equilibrium comes of age. Nature 1993;366:223–7. https://doi.org/10.1038/366223a0.

[30] Hicks J, Muthuswamy L, Krasnitz A, Navin N, Riggs M, Grubor V, Esposito D, Alexander J, Troge J, Wigler M, Maner S, Lundin P, Zetterberg A. High-Resolution ROMA CGH and FISH Analysis of Aneuploid and Diploid Breast Tumors. Cold Spring Harb Symp Quant Biol 2005;70:51–63. https://doi.org/10.1101/sqb.2005.70.055.

[31] Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leibu E, Esposito D, Alexander J, Troge J, Grubor V, Yoon S, Wigler M, Ye K, Børresen-Dale A-L, Naume B, Schlicting E, Norton L, Hägerström T, Skoog L, Auer G, Månér S, Lundin P, Zetterberg A. Novel patterns

of genome rearrangement and their association with survival in breast cancer. Genome Res 2006;16:1465–79. https://doi.org/10.1101/gr.5460106.

[32] Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin M-L, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 2011;144:27–40. https://doi.org/10.1016/j.cell.2010.11.055.

[33] Zhang C-Z, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. Chromothripsis from DNA damage in micronuclei. Nature 2015;522:179–84. https://doi.org/10.1038/nature14493.

[34] Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, Vermaat JS, Van Roosmalen MJ, Van Lieshout S, Nijman IJ, Roessingh W, Van 'T Slot R, Van De Belt J, Guryev V, Koudijs M, Voest E, Cuppen E. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol 2011;12:R103. https://doi.org/10.1186/gb-2011-12-10-r103.

[35] Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA. The genomic complexity of primary human prostate cancer. Nature 2011;470:214–20. https://doi.org/10.1038/nature09744.

[36] Newburger DE, Kashef-Haghighi D, Weng Z, Salari R, Sweeney RT, Brunner AL, Zhu SX, Guo X, Varma S, Troxell ML, West RB, Batzoglou S, Sidow A. Genome evolution during progression to breast cancer. Genome Res 2013;23:1097–108. https://doi.org/10.1101/gr.151670.112.

[37] Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer 1953;6:963–8. https://doi.org/10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q.

[38] Reed AD, Pensa S, Steif A, Stenning J, Kunz DJ, Porter LJ, Hua K, He P, Twigger A-J, Siu AJQ, Kania K, Barrow-McGee R, Goulding I, Gomm JJ, Speirs V, Jones JL, Marioni JC, Khaled WT. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. Nat Genet 2024;56:652–62. https://doi.org/10.1038/s41588-024-01688-9.

[39] Lochhead P, Chan AT, Nishihara R, Fuchs CS, Beck AH, Giovannucci E, Ogino S. Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. Mod Pathol 2015;28:14–29. https://doi.org/10.1038/modpathol.2014.81.

[40] Danforth DN. Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. Breast Cancer Basic Clin Res 2016;10:BCBCR.S39384. https://doi.org/10.4137/BCBCR.S39384.

[41] Nishimura T, Kakiuchi N, Yoshida K, Sakurai T, Kataoka TR, Kondoh E, Chigusa Y, Kawai M, Sawada M, Inoue T, Takeuchi Y, Maeda H, Baba S, Shiozawa Y, Saiki R, Nakagawa MM, Nannya Y, Ochi Y, Hirano T, Nakagawa T, Inagaki-Kawata Y, Aoki K, Hirata M, Nanki K, Matano M, Saito M, Suzuki E, Takada M, Kawashima M, Kawaguchi K, Chiba K, Shiraishi Y, Takita J, Miyano S, Mandai M, Sato T, Takeuchi K, Haga H, Toi M, Ogawa S. Evolutionary histories of breast cancer and related clones. Nature 2023;620:607–14. https://doi.org/10.1038/s41586-023-06333-9.

[42] Graham K, de las Morenas A, Tripathi A, King C, Kavanah M, Mendez J, Stone M, Slama J, Miller M, Antoine G, Willers H, Sebastiani P, Rosenberg CL. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. Br J Cancer 2010;102:1284–93. https://doi.org/10.1038/sj.bjc.6605576.

[43] Román-Pérez E, Casbas-Hernández P, Pirone JR, Rein J, Carey LA, Lubet RA, Mani SA, Amos KD, Troester MA. Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients. Breast Cancer Res 2012;14:R51. https://doi.org/10.1186/bcr3152.

[44] Huang X, Stern DF, Zhao H. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival – Evidence from TCGA Pan-Cancer Data. Sci Rep 2016;6:20567. https://doi.org/10.1038/srep20567.

[45] Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun 2017;8:1077. https://doi.org/10.1038/s41467-017-01027-z.

[46] Gadaleta E, Fourgoux P, Pirró S, Thorn GJ, Nelan R, Ironside A, Rajeeve V, Cutillas PR, Lobley AE, Wang J, Gea E, Ross-Adams H, Bessant C, Lemoine NR, Jones LJ, Chelala C. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. NPJ Breast Cancer 2020;6:38. https://doi.org/10.1038/s41523-020-00182-9.

[47] Morla-Barcelo PM, Laguna-Macarrilla D, Cordoba O, Matheu G, Oliver J, Roca P, Nadal-Serrano M, Sastre-Serra J. Unraveling malignant phenotype of peritumoral tissue: transcriptomic insights into early-stage breast cancer. Breast Cancer Res 2024;26:89. https://doi.org/10.1186/s13058-024-01837-2.

[48] Sverchkova A, Burkholz S, Rubsamen R, Stratford R, Clancy T. Integrative HLA typing of tumor and adjacent normal tissue can reveal insights into the tumor immune response. BMC Med Genomics 2024;17:37. https://doi.org/10.1186/s12920-024-01808-8.

[49] Lau HSH, Tan VKM, Tan BKT, Sim Y, Quist J, Thike AA, Tan PH, Pervaiz S, Grigoriadis A, Sabapathy K. Adipose-enriched peri-tumoral stroma, in contrast to myofibroblast-enriched stroma, prognosticates poorer survival in breast cancers. Npj Breast Cancer 2023;9:84. https://doi.org/10.1038/s41523-023-00590-7.

[50] Forsberg LA, Rasi C, Pekar G, Davies H, Piotrowski A, Absher D, Razzaghian HR, Ambicka A, Halaszka K, Przewoźnik M, Kruczak A, Mandava G, Pasupulati S, Hacker J, Prakash KR, Dasari RC, Lau J, Penagos-Tafurt N, Olofsson HM, Hallberg G, Skotnicki P, Mituś J, Skokowski J, Jankowski M, Śrutek E, Zegarski W, Tiensuu Janson E, Ryś J, Tot T, Dumanski JP. Signatures of post-zygotic structural genetic aberrations in the cells of histologically normal breast tissue that can predispose to sporadic breast cancer. Genome Res 2015;25:1521–35. https://doi.org/10.1101/gr.187823.114.

[51] Maeda H, Kakiuchi N. Clonal expansion in normal tissues. Cancer Sci 2024;115:2117–24. https://doi.org/10.1111/cas.16183.

[52] Herms A, Colom B, Piedrafita G, Kalogeropoulou A, Banerjee U, King C, Abby E, Murai K, Caseda I, Fernandez-Antoran D, Ong SH, Hall MWJ, Bryant C, Sood RK, Fowler JC, Pol A, Frezza C, Vanhaesebroeck B, Jones PH. Organismal metabolism regulates the expansion of

oncogenic PIK3CA mutant clones in normal esophagus. Nat Genet 2024;56:2144–57. https://doi.org/10.1038/s41588-024-01891-8.

[53] Shahrouzi P, Forouz F, Mathelier A, Kristensen VN, Duijf PHG. Copy number alterations: a catastrophic orchestration of the breast cancer genome. Trends Mol Med 2024;30:750–64. https://doi.org/10.1016/j.molmed.2024.04.017.

[54] Ronowicz A, Janaszak-Jasiecka A, Skokowski J, Madanecki P, Bartoszewski R, Bałut M, Seroczyńska B, Kochan K, Bogdan A, Butkus M, Pęksa R, Ratajska M, Kuźniacka A, Wasąg B, Gucwa M, Krzyżanowski M, Jaśkiewicz J, Jankowski Z, Forsberg L, Ochocka JR, Limon J, Crowley MR, Buckley PG, Messiaen L, Dumanski JP, Piotrowski A. Concurrent DNA Copy-Number Alterations and Mutations in Genes Related to Maintenance of Genome Stability in Uninvolved Mammary Glandular Tissue from Breast Cancer Patients. Hum Mutat 2015;36:1088–99. https://doi.org/10.1002/humu.22845.

[55] Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut S-J, Tsui DWY, Liu B, Dawson S-J, Abraham J, Northen H, Peden JF, Mukherjee A, Turashvili G, Green AR, McKinney S, Oloumi A, Shah S, Rosenfeld N, Murphy L, Bentley DR, Ellis IO, Purushotham A, Pinder SE, Børresen-Dale A-L, Earl HM, Pharoah PD, Ross MT, Aparicio S, Caldas C. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun 2016;7:11479. https://doi.org/10.1038/ncomms11479.

[56] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjærde OC, Langerød A, Ringnér M, Ahn S-M, Boyault S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GKJ, Jang SJ, Jones DR, Kim H-Y, King TA, Krishnamurthy S, Lee HJ, Lee J-Y, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporté I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodríguez-González FG, Romieu G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, Van Den Eynden GG, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, Veer LV, Tutt A, Knappskog S, Tan BKT, Jonkers J, Borg Å, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, Van De Vijver MJ, Martens JWM, Børresen-Dale A-L, Richardson AL, Kong G, Thomas G, Stratton

MR. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016;534:47–54. https://doi.org/10.1038/nature17676.

[57] Li G, Guo X, Chen M, Tang L, Jiang H, Day JX, Xie Y, Peng L, Xu X, Li J, Wang S, Xiao Z, Dai L, Wang J. Prevalence and spectrum of AKT1, PIK3CA, PTEN and TP53 somatic mutations in Chinese breast cancer patients. PLOS ONE 2018;13:e0203495. https://doi.org/10.1371/journal.pone.0203495.

[58] Oh J-H, Sung CO. Comprehensive characteristics of somatic mutations in the normal tissues of patients with cancer and existence of somatic mutant clones linked to cancer development. J Med Genet 2021;58:433–41. https://doi.org/10.1136/jmedgenet-2020-106905.

[59] Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, Kwa EJ, Lee-Six H, Cagan A, Coorens THH, Chapman MS, Olafsson S, Leonard S, Jones D, Machado HE, Davies M, Øbro NF, Mahubani KT, Allinson K, Gerstung M, Saeb-Parsy K, Kent DG, Laurenti E, Stratton MR, Rahbari R, Campbell PJ, Osborne RJ, Martincorena I. Somatic mutation landscapes at single-molecule resolution. Nature 2021;593:405–10. https://doi.org/10.1038/s41586-021-03477-4.

[60] Pipek O, Alpár D, Rusz O, Bödör C, Udvarnoki Z, Medgyes-Horváth A, Csabai I, Szállási Z, Madaras L, Kahán Z, Cserni G, Kővári B, Kulka J, Tőkés AM. Genomic Landscape of Normal and Breast Cancer Tissues in a Hungarian Pilot Cohort. Int J Mol Sci 2023;24:8553. https://doi.org/10.3390/ijms24108553.

[61] Rockweiler NB, Ramu A, Nagirnaja L, Wong WH, Noordam MJ, Drubin CW, Huang N, Miller B, Todres EZ, Vigh-Conrad KA, Zito A, Small KS, Ardlie KG, Cohen BA, Conrad DF. The origins and functional effects of postzygotic mutations throughout the human life span. Science 2023;380:eabn7113. https://doi.org/10.1126/science.abn7113.

[62] Breast Cancer Facts & Figures 2022-2024 n.d.

[63] Cancer Facts & Figures 2023 1930.

[64] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. Ann Surg Oncol 2010;17:1471–4. https://doi.org/10.1245/s10434-010-0985-4.

[65] LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 2009;37:4181–93. https://doi.org/10.1093/nar/gkp552.

[66] Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;7:111–8. https://doi.org/10.1038/nmeth.1419.

[67] Mertes F, ElSharawy A, Sauer S, Van Helvoort JMLM, Van Der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 2011;10:374–86. https://doi.org/10.1093/bfgp/elr033.

[68] Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 2011;12:87–98. https://doi.org/10.1038/nrg2934.

[69] Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol 2009;10:R115. https://doi.org/10.1186/gb-2009-10-10-r115.

[70] Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272–6. https://doi.org/10.1038/nature08250.

[71] Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522–7. https://doi.org/10.1038/ng.2007.42.

[72] Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci 2012;109:14508–13. https://doi.org/10.1073/pnas.1208715109.

[73] Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen J-C, Risques R-A, Loeb LA. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 2014;9:2586–606. https://doi.org/10.1038/nprot.2014.170.

[74] Biesecker LG, Green RC. Diagnostic Clinical Genome and Exome Sequencing. N Engl J Med 2014;370:2418–25. https://doi.org/10.1056/NEJMra1312543.

[75] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 1977;74:5463–7. https://doi.org/10.1073/pnas.74.12.5463.

[76] Słomka M, Sobalska-Kwapis M, Wachulec M, Bartosz G, Strapagiel D. High Resolution Melting (HRM) for High-Throughput Genotyping—Limitations and Caveats in Practical Case Studies. Int J Mol Sci 2017;18:2316. https://doi.org/10.3390/ijms18112316.

[77] Frank C, Fallah M, Sundquist J, Hemminki A, Hemminki K. Population Landscape of Familial Cancer. Sci Rep 2015;5:12891. https://doi.org/10.1038/srep12891.

[78] Hemminki K, Sundquist J, Bermejo JL. How common is familial cancer? Ann Oncol Off J Eur Soc Med Oncol 2008;19:163–7. https://doi.org/10.1093/annonc/mdm414.

[79] Scheuner MT, McNeel TS, Freedman AN. Population prevalence of familial cancer and common hereditary cancer syndromes. The 2005 California Health Interview Survey. Genet Med Off J Am Coll Med Genet 2010;12:726–35. https://doi.org/10.1097/GIM.0b013e3181f30e9e.

[80] Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, Diaz de Ståhl T, Zaghlool A, Giedraitis V, Lannfelt L, Score J, Cross NCP, Absher D, Janson ET, Lindgren CM, Morris AP, Ingelsson E, Lind L, Dumanski JP. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet 2014;46:624–8. https://doi.org/10.1038/ng.2966.

[81] Dumanski JP, Rasi C, Lönn M, Davies H, Ingelsson M, Giedraitis V, Lannfelt L, Magnusson PKE, Lindgren CM, Morris AP, Cesarini D, Johannesson M, Tiensuu Janson E, Lind L, Pedersen NL, Ingelsson E, Forsberg LA. Mutagenesis. Smoking is associated with mosaic loss of chromosome Y. Science 2015;347:81–3. https://doi.org/10.1126/science.1262092.

[82] Dumanski JP, Lambert J-C, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, Lindgren CM, Campion D, Dufouil C, European Alzheimer's Disease Initiative Investigators, Pasquier F, Amouyel P, Lannfelt L, Ingelsson M, Kilander L, Lind L, Forsberg LA. Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. Am J Hum Genet 2016;98:1208–19. https://doi.org/10.1016/j.ajhg.2016.05.014.

[83] Dumanski JP, Halvardson J, Davies H, Rychlicka-Buniowska E, Mattisson J, Moghadam BT, Nagy N, Węglarczyk K, Bukowska-Strakova K, Danielsson M, Olszewski P, Piotrowski A, Oerton E, Ambicka A, Przewoźnik M, Bełch Ł, Grodzicki T, Chłosta PL, Imreh S, Giedraitis V, Kilander L, Nordlund J, Ameur A, Gyllensten U, Johansson Å, Józkowicz A, Siedlar M, Klich-Rączka A, Jaszczyński J, Enroth S, Baran J, Ingelsson M, Perry JRB, Ryś J, Forsberg

LA. Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. Cell Mol Life Sci CMLS 2021;78:4019–33. https://doi.org/10.1007/s00018-021-03822-w.

[84] Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. Lancet Glob Health 2020;8:e1027–37. https://doi.org/10.1016/S2214-109X(20)30215-1.

[85] Dall GV, Britt KL. Estrogen Effects on the Mammary Gland in Early and Late Life and Breast Cancer Risk. Front Oncol 2017;7:110. https://doi.org/10.3389/fonc.2017.00110.

[86] Almeida M, Soares M, Fonseca-Moutinho J, Ramalhinho AC, Breitenfeld L. Influence of Estrogenic Metabolic Pathway Genes Polymorphisms on Postmenopausal Breast Cancer Risk. Pharm Basel Switz 2021;14:94. https://doi.org/10.3390/ph14020094.

[87] Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. N Engl J Med 2006;354:270–82. https://doi.org/10.1056/NEJMra050776.

[88] Cai Y, Crowther J, Pastor T, Abbasi Asbagh L, Baietti MF, De Troyer M, Vazquez I, Talebi A, Renzi F, Dehairs J, Swinnen JV, Sablina AA. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. Cancer Cell 2016;29:751–66. https://doi.org/10.1016/j.ccell.2016.04.003.

[89] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209–49. https://doi.org/10.3322/caac.21660.

[90] Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. Br J Radiol 2022;95:20211033. https://doi.org/10.1259/bjr.20211033.

[91] Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S, McGale P, Correa C, Taylor C, Arriagada R, Clarke M, Cutter D, Davies C, Ewertz M, Godwin J, Gray R, Pierce L, Whelan T, Wang Y, Peto R. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet Lond Engl 2011;378:1707–16. https://doi.org/10.1016/S0140-6736(11)61629-2.

[92] Kostecka A, Nowikiewicz T, Olszewski P, Koczkowska M, Horbacz M, Heinzl M, Andreou M, Salazar R, Mair T, Madanecki P, Gucwa M, Davies H, Skokowski J, Buckley PG, Pęksa

R, Śrutek E, Szylberg Ł, Hartman J, Jankowski M, Zegarski W, Tiemann-Boege I, Dumanski JP, Piotrowski A. High prevalence of somatic PIK3CA and TP53 pathogenic variants in the normal mammary gland tissue of sporadic breast cancer patients revealed by duplex sequencing. Npj Breast Cancer 2022;8:1–10. https://doi.org/10.1038/s41523-022-00443-9.

[93] Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2018, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2018/, based on November 2020 SEER data submission, posted to the SEER web site, April 2021. n.d.

[94] Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P, Peto R, Pritchard KI, Bergh J, Dowsett M, Hayes DF, EBCTCG. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. N Engl J Med 2017;377:1836–46. https://doi.org/10.1056/NEJMoa1701830.

[95] Rampias T, Karagiannis D, Avgeris M, Polyzos A, Kokkalis A, Kanaki Z, Kousidou E, Tzetis M, Kanavakis E, Stravodimos K, Manola KN, Pantelias GE, Scorilas A, Klinakis A. The lysine-specific methyltransferase KMT 2C/ MLL 3 regulates DNA repair components in cancer. EMBO Rep 2019;20:e46821. https://doi.org/10.15252/embr.201846821.

[96] Luongo F, Colonna F, Calapà F, Vitale S, Fiori ME, De Maria R. PTEN Tumor-Suppressor: The Dam of Stemness in Cancer. Cancers 2019;11:1076. https://doi.org/10.3390/cancers11081076.

[97] Liang B, Zhou Y, Qian M, Xu M, Wang J, Zhang Y, Song X, Wang H, Lin S, Ren C, Monga SP, Wang B, Evert M, Chen Y, Chen X, Huang Z, Calvisi DF, Chen X. TBX3 functions as a tumor suppressor downstream of activated CTNNB1 mutants during hepatocarcinogenesis. J Hepatol 2021;75:120–31. https://doi.org/10.1016/j.jhep.2021.01.044.

[98] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh C-L, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. REVEL: An

Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 2016;99:877–85. https://doi.org/10.1016/j.ajhg.2016.08.016.

[99] Bader AG, Kang S, Zhao L, Vogt PK. Oncogenic PI3K deregulates transcription and translation. Nat Rev Cancer 2005;5:921–9. https://doi.org/10.1038/nrc1753.

[100] Staal SP. Molecular cloning of the akt oncogene and its human homologues AKT1 and AKT2: amplification of AKT1 in a primary human gastric adenocarcinoma. Proc Natl Acad Sci U S A 1987;84:5034–7. https://doi.org/10.1073/pnas.84.14.5034.

[101] Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, Pillai MM. Cancer-associated SF3B1 mutants recognize otherwise inaccessible cryptic 3' splice sites within RNA secondary structures. Oncogene 2017;36:1123–33. https://doi.org/10.1038/onc.2016.279.

[102] Hobbs GA, Der CJ, Rossman KL. RAS isoforms and mutations in cancer at a glance. J Cell Sci 2016;129:1287–92. https://doi.org/10.1242/jcs.182873.

[103] He X, Zhang L, Chen Y, Remke M, Shih D, Lu F, Wang H, Deng Y, Yu Y, Xia Y, Wu X, Ramaswamy V, Hu T, Wang F, Zhou W, Burns DK, Kim SH, Kool M, Pfister SM, Weinstein LS, Pomeroy SL, Gilbertson RJ, Rubin JB, Hou Y, Wechsler-Reya R, Taylor MD, Lu QR. The G protein α subunit Gαs is a tumor suppressor in Sonic hedgehog-driven medulloblastoma. Nat Med 2014;20:1035–42. https://doi.org/10.1038/nm.3666.

[104] Otálora-Otálora B, Henríquez B, López-Kleine L, Rojas A. RUNX family: Oncogenes or tumor suppressors (Review). Oncol Rep 2019. https://doi.org/10.3892/or.2019.7149.

[105] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 2018;18:696–705. https://doi.org/10.1038/s41568-018-0060-1.

[106] NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic Version 1.2023 — September 7, 2022 NCCN.org n.d.

[107] Kandel R, Li S-Q, Ozcelik H, Rohan T. p53 protein accumulation and mutations in normal and benign breast tissue. Int J Cancer 2000;87:73–8. https://doi.org/10.1002/1097-0215(20000701)87:1<73::AID-IJC11>3.0.CO;2-U.

[108] Soysal SD, Ng CKY, Costa L, Weber WP, Paradiso V, Piscuoglio S, Muenst S. Genetic Alterations in Benign Breast Biopsies of Subsequent Breast Cancer Patients. Front Med 2019;6:166. https://doi.org/10.3389/fmed.2019.00166.

[109] Cereser B, Yiu A, Tabassum N, Del Bel Belluz L, Zagorac S, Ancheta KRZ, Zhong R, Miere C, Jeffries-Jones AR, Moderau N, Werner B, Stebbing J. The mutational landscape of the adult healthy parous and nulliparous human breast. Nat Commun 2023;14:5136. https://doi.org/10.1038/s41467-023-40608-z.

VIII. LIST OF FIGURES WITH FIGURE LEGENDS

IX. LIST OF TABLES WITH TABLE LEGENDS

X. LIST OF ABBREVIATIONS

AJCC: American Joint Committee On Cancer

BCAP: Breast Cancer Adverse Prognoses

BCS: Breast-Conserving Surgery

BCUS: Breast Cancer Un-Selected

BE: Branched Evolution

BL: Peripheral Blood

BP: Plasma

cDNA: Complementary DNA

CNAs: Copy Number Alterations

CNVs: Copy Number Variations

CPM: Contralateral Prophylactic Mastectomy

CSC: Cancer Stem Cell

CTRL: Control

DCIS: Ductal Carcinoma In Situ

DEGs: Differentially Expressed Genes

E: Estrogen

H&E: Hematoxylin and Eosin

HLA: Human Leukocyte Antigens

HRM: High-Resolution Melting

IDC: Invasive Ductal Carcinoma

ILC: Invasive Lobular Carcinoma

LE: Linear Evolution

LOH: Loss Of Heterozygosity

LOY: Loss of the Y chromosome

MRI: Magnetic Resonance Imaging

mRNA: Messenger RNA

MUG: Medical University of Gdańsk

NGS: Next-Generation Sequencing

PCA: Principal Component Analysis

PE: Punctuated Evolution

PET: Positron Emission Tomography

Pg: Progesterone

Prl: Prolactin

PT: Primary Tumor

RM: Reduction Mammoplasty

RNA-seq: RNA sequencing

SK: Skin

SNP: Single Nucleotide Polymorphism

TEBs: Terminal End Buds

TNBC: Triple-Negative Breast Cancer

UM: Uninvolved Margin / Uninvolved Mammary gland

UMD: UM Distal from PT

UMIs: Unique Molecular Identifiers

UMP: UM Proximal from PT

VAF: Variant Allele Frequency

WB: Whole Blood

WES: Whole Exome Sequencing

XI. STATEMENT OF AUTHOR CONTRIBUTIONS

I, Maria Andreou, declare my contributions to the following three publications and preprint (unpublished findings, manuscript under review), which are included in this thesis.

Paper I – Filipowicz N. et al.

Contributions:

- Investigation.
- Writing – review & editing.

Paper II – Kostecka A. et al.

Contributions:

- Experiments.

Paper III – Andreou M. & Jąkalski M. et al

**Maria Andreou and Marcin Jąkalski have contributed equally to this work*

Contributions:

- Conceptualization.
- Data curation.
- Investigation.
- Visualization.
- Interpretation.
- Article writing-original.
- Article writing – review, editing, and acquisition of additional data for review.

This statement accurately reflects my contributions to these publications and their role in my thesis.

Unpublished findings, manuscript under review (preprint)

Contributions:

- Study design and conception.
- Experiments.
- Data analysis and interpretation.

- Visualization.
- Manuscript writing-original.
- Manuscript writing – review and editing.

Maria Andreou

XII. PUBLICATIONS

Paper I

PLOS ONE

RESEARCH ARTICLE

Comprehensive cancer-oriented biobanking resource of human samples for studies of post-zygotic genetic variation involved in cancer predisposition

Natalia Filipowicz[1]*, Kinga Drężek[1], Monika Horbacz[1], Agata Wojdak[1], Jakub Szymanowski[1,2], Edyta Rychlicka-Buniowska[1], Ulana Juhas[1], Katarzyna Duzowska[1], Tomasz Nowikiewicz[3,4], Wiktoria Stańkowska[1], Katarzyna Chojnowska[1], Maria Andreou[1], Urszula Ławrynowicz[1], Magdalena Wójcik[1], Hanna Davies[5], Ewa Śrutek[4,6], Michał Bieńkowski[7], Katarzyna Milian-Ciesielska[8], Marek Zdrenka[6], Aleksandra Ambicka[9], Marcin Przewoźnik[9], Agnieszka Harazin-Lechowska[9], Agnieszka Adamczyk[9], Jacek Kowalski[7], Dariusz Bała[4,10], Dorian Wiśniewski[10], Karol Tkaczyński[10], Krzysztof Kamecki[11], Marta Drzewiecka[3], Paweł Wroński[11], Jerzy Siekiera[11], Izabela Ratnicka[12], Jerzy Jankau[12], Karol Wierzba[13], Jarosław Skokowski[14,15], Karol Połom[14], Mikołaj Przydacz[16], Łukasz Belch[16], Piotr Chłosta[16], Marcin Matuszewski[17], Krzysztof Okoń[8], Olga Rostkowska[18], Andrzej Hellmann[18], Karol Sasim[19], Piotr Remiszewski[18], Marek Sierżęga[20], Stanisław Hać[18], Jarosław Kobiela[18], Łukasz Kaska[18], Michał Jankowski[4,10], Diana Hodorowicz-Zaniewska[20], Janusz Jaszczyński[21], Wojciech Zegarski[4,10], Wojciech Makarewicz[14,22], Rafał Pęksa[7], Joanna Szpor[8], Janusz Ryś[9], Łukasz Szylberg[6,23], Arkadiusz Piotrowski[1,24], Jan P. Dumanski[1,5,24]*

1 3P-Medicine Laboratory, Medical University of Gdańsk, Gdańsk, Poland, **2** Bioenit Jakub Szymanowski, Gdańsk, Poland, **3** Department of Breast Cancer and Reconstructive Surgery, Oncology Center—Prof. Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland, **4** Surgical Oncology, Ludwik Rydygier's Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland, **5** Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden, **6** Department of Tumor Pathology and Pathomorphology, Oncology Center—Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland, **7** Department of Pathomorphology, Medical University of Gdańsk, Gdańsk, Poland, **8** Department of Pathomorphology, Jagiellonian University Medical College, Kraków, Poland, **9** Department of Tumor Pathology, Maria Skłodowska-Curie National Research Institute of Oncology, Kraków, Poland, **10** Department of Surgical Oncology, Oncology Center—Prof. Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland, **11** Department of Urology, Oncology Center—Prof. Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland, **12** Department of Plastic Surgery, Medical University of Gdańsk, Gdańsk, Poland, **13** Department of Internal Medicine, Connective Tissue Diseases and Geriatrics, Medical University of Gdańsk, Gdańsk, Poland, **14** Department of Surgical Oncology, Medical University of Gdańsk, Gdańsk, Poland, **15** Department of Medical Laboratory Diagnostics-Biobank, Medical University of Gdańsk, Gdańsk, Poland, **16** Department of Urology, Jagiellonian University Medical College, Kraków, Poland, **17** Department and Clinic of Urology, Medical University of Gdańsk, Gdańsk, Poland, **18** Department of General, Endocrine and Transplant Surgery, Medical University of Gdańsk, Gdańsk, Poland, **19** Clinic of Urology and Oncological Urology, Specialist Hospital of Kościerzyna, Kościerzyna, Poland, **20** Department of General, Oncological, and Gastrointestinal Surgery, Jagiellonian University Medical College, Kraków, Poland, **21** Department of Urology, Maria Skłodowska-Curie National Research Institute of Oncology, Kraków, Poland, **22** Clinic of General and Oncological Surgery, Specialist Hospital of Kościerzyna, Kościerzyna, Poland, **23** Department of Clinical Pathomorphology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland, **24** Department of Biology and Pharmaceutical Botany, Medical University of Gdańsk, Gdańsk, Poland

* natalia.filipowicz@gumed.edu.pl (NF); jan.dumanski@igp.uu.se (JPD)

OPEN ACCESS

Citation: Filipowicz N, Drężek K, Horbacz M, Wojdak A, Szymanowski J, Rychlicka-Buniowska E, et al. (2022) Comprehensive cancer-oriented biobanking resource of human samples for studies of post-zygotic genetic variation involved in cancer predisposition. PLoS ONE 17(4): e0266111. https://doi.org/10.1371/journal.pone.0266111

Editor: Isaac Yi Kim, Yale University School of Medicine, UNITED STATES

Received: August 29, 2021

Accepted: March 14, 2022

Published: April 7, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0266111

Data Availability Statement: We are not able to provide the access to our internal database MABData 2, as this is only available locally for the

authorized researchers of our unit. However, we have prepared a minimal anonymized dataset in excel file with all the donors and clinical/medical data that were included in the paper. We provide this information as a supporting information file (S1 Table).

Funding: This study was sponsored by the Foundation for Polish Science (FNP) under the International Research Agendas Program to J.P.D. and A.P., co-financed by the European Union under the European Regional Development Fund. Our biobank also received financing via the "Excellence Initiative - Research University" program from Medical University of Gdansk. This project obtained further partial funding from The Swedish Cancer Society and Swedish Medical Research Council to J.P.D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: J.P.D. is cofounder and shareholder in Cray Innovation AB. The remaining authors have declared that no competing interests exist.

Abstract

The progress in translational cancer research relies on access to well-characterized samples from a representative number of patients and controls. The rationale behind our biobanking are explorations of post-zygotic pathogenic gene variants, especially in non-tumoral tissue, which might predispose to cancers. The targeted diagnoses are carcinomas of the breast (via mastectomy or breast conserving surgery), colon and rectum, prostate, and urinary bladder (via cystectomy or transurethral resection), exocrine pancreatic carcinoma as well as metastases of colorectal cancer to the liver. The choice was based on the high incidence of these cancers and/or frequent fatal outcome. We also collect age-matched normal controls. Our still ongoing collection originates from five clinical centers and after nearly 2-year cooperation reached 1711 patients and controls, yielding a total of 23226 independent samples, with an average of 74 donors and 1010 samples collected per month. The predominant diagnosis is breast carcinoma, with 933 donors, followed by colorectal carcinoma (383 donors), prostate carcinoma (221 donors), bladder carcinoma (81 donors), exocrine pancreatic carcinoma (15 donors) and metachronous colorectal cancer metastases to liver (14 donors). Forty percent of the total sample count originates from macroscopically healthy cancer-neighboring tissue, while contribution from tumors is 12%, which adds to the uniqueness of our collection for cancer predisposition studies. Moreover, we developed two program packages, enabling registration of patients, clinical data and samples at the participating hospitals as well as the central system of sample/data management at coordinating center. The approach used by us may serve as a model for dispersed biobanking from multiple satellite hospitals. Our biobanking resource ought to stimulate research into genetic mechanisms underlying the development of common cancers. It will allow all available "-omics" approaches on DNA-, RNA-, protein- and tissue levels to be applied. The collected samples can be made available to other research groups.

Introduction

One of the prerequisites for translational research is availability of well-characterized samples of different types from patients suffering from various diseases. Another requirement is the access to comprehensive and long-term follow-up clinical records for patients, which is important for the correlations between molecular findings and medical parameters. The third condition is a broad participation of patients treated at hospitals as well as control subjects, via their donation of samples to research projects. When these conditions are met, progress can be made towards 3P Medicine, i.e. preventive, personalized and precision.

Cancer is generally defined as a genetic disease, but the frequency of germline cancer-predisposing mutations vary considerably between different tumors and these inherited mutations are responsible for less than 10% of all cancers [1–3]. The remaining >90% of cancers arise as a result of mutations acquired during lifetime in normal somatic cells and the bulk of all cancers occurs late in life. Studies of cancer genomes have contributed to numerous discoveries of mutations that drive cancer growth. However, a fully developed tumor is often clonally heterogeneous and represents a late stage of the disease. This might restrict description of mutations that are occurring very early and initiate tumor development.

Post-zygotic (or somatic) mutations (PZM) in histologically normal human cells from various organs that develop cancer have increasingly been suggested over the past decade as a major source of cancer driving mutations, but this field is still poorly explored [4]. Examples can be given for breast cancer [5–7], normal skin and other normal tissues [8,9], colon cancer [10], urinary bladder cancer [11], aberrant clonal expansions in peripheral blood of healthy subjects (also known as clonal hematopoiesis of indeterminate potential) [12–14] and esophageal cancer [15]. In the latter study, the prevalence of cancer driving mutations was higher in normal epithelium than in esophageal cancers. Furthermore, our biobanking project has been influenced by our interest in analysis of mosaic Loss of chromosome Y (LOY) in blood. It has been noted for over 50 years that chromosome Y is frequently lost in the leukocytes of aging men [16,17]. Recent epidemiological investigations show that LOY in leukocytes, representing lack of nearly 2% of the haploid nuclear genome, is associated with earlier mortality and morbidity from many diseases in men, including multiple common forms of cancer [13,18,19]. Moreover, LOY is the most common post-zygotic (somatic) mutation from analyses of bulk DNA and single-cells from peripheral blood [20].

Thus, comprehensive cancer-oriented biobanking requires sampling of not only tumor tissues but also normal fragments from the affected organ and other control tissue(s) that is not directly involved in the disease process. We use the term "uninvolved margin (UM) or tissue" that refers to histologically controlled non-tumorous tissue, which is located at various distances from the site of primary tumor [5,6]. Collection of blood and plasma (liquid biopsy) is also crucial for future genetic and proteomic analyses. We report here the results of biobanking activities for five common cancer diagnoses that have been ongoing at five major clinical cancer centers in four cities of Poland for a period of more than 23 months.

Materials and methods

Diagnoses, logistics and collection protocols

We selected five diagnoses, i.e. breast-, colorectal-, prostate-, bladder- and exocrine pancreas carcinomas, as well as metachronic metastases of colorectal cancer to liver. The choice was based on the high incidence of these diagnoses and/or often fatal outcome of the disease. We established collaboration with five clinical centers in Poland: Oncology Center in Bydgoszcz; National Institute of Oncology in Cracow; University Clinical Centre in Gdansk, University Hospital in Cracow and Specialist Hospital in Koscierzyna. The Pathology Departments at each of these centers were equipped with small -80˚C freezers (80-liter volume, model ULTF 80, Arctico), used for temporary storage of samples prior to shipment to 3P-Medicine Laboratory in Gdansk. The freezers were accompanied by a laptop computer and rack reader for 2D Data-Matrix coded tubes (2 ml, Micronic), used for registration of samples with help of newly developed "MABData1" biobanking software (see below). Dispatch of samples on dry-ice and the corresponding documents from each hospital to 3P-Medicine Laboratory in Gdansk takes place approximately quarterly according to a well-defined procedure.

Figs 1 and 2 show types and number of independent samples collected for two common diagnoses such as breast- and prostate cancer. Breast cancer samples were obtained during mastectomy or breast conserving therapy (BCT) surgeries. Fig 3 show similar outlines of sample collection from patients operated for colorectal cancer (resection of primary tumor and metachronic metastasis of colorectal cancer to liver), bladder cancer (treated by either cystectomy or Trans-Urethral Resection of Bladder Tumor [TURBT]) and exocrine pancreas cancer, respectively. For each diagnosis, the compulsory set of samples include: 1–2 primary tumor fragments (PT); 1–12 specimens of uninvolved margin (UM) composed of macroscopically normal tissue collected at various distances from PT; 1–4 samples of whole blood (WB) (1.5 ml

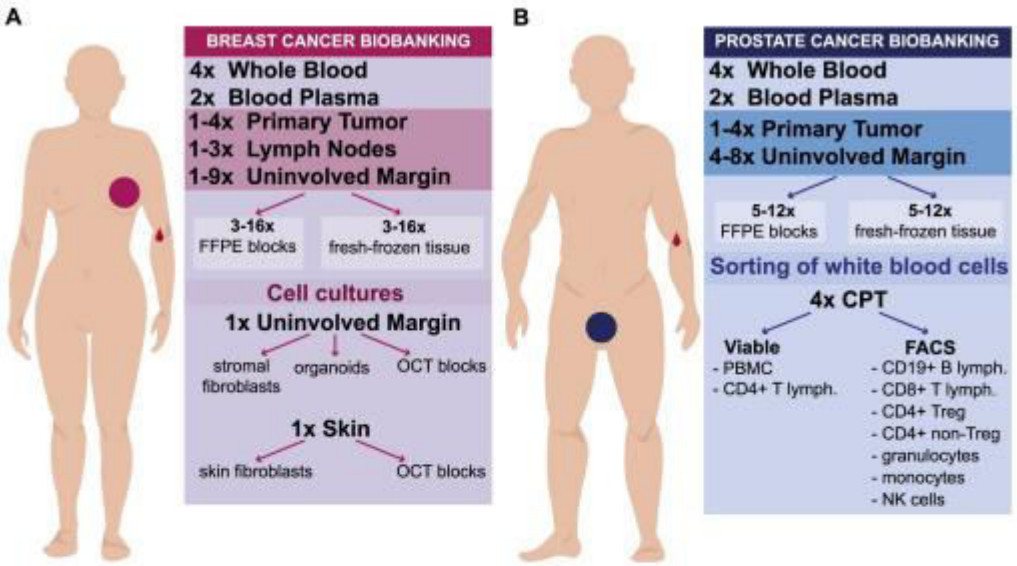

Fig 1. A summary of samples collected for two common cancer diagnoses. (A) Collection for breast carcinoma patients. (B) Collection for prostate carcinoma patients. FFPE, Formalin-Fixed Paraffin-Embedded blocks; OCT, Optimal Cutting Temperature compound for fresh-frozen tissue; PBMC, Peripheral Blood Mononuclear Cells; CPT, Cell Preparation Tube with Sodium Heparin (BD Bioscience) for separation of granulocyte- and PBMC-fraction of white blood cells; FACS, Fluorescent Activated Cell Sorting; lymph., lymphocytes; Treg, T-regulatory lymphocytes; NK, Natural Killer cells.

https://doi.org/10.1371/journal.pone.0266111.g001

each) as the normal reference tissue; and 1–2 samples blood plasma (BP) (1–1.5 ml each) for future proteomic studies. Whenever available, for breast and colorectal carcinoma, local metastases to lymph node(s) (LN) were also collected, but only when they were clearly identifiable and large enough on gross examination. The volumes of samples from solid tissues ranged between 0.005 cm^3 to 1 cm^3. The tissues were collected according to the well-defined protocols. After macro-sectioning of the resected organ, small tissue fragments were selected and excised for biobanking. Subsequently, each fragment was cut in half: one portion was placed into a cryovial and fresh-frozen at -80˚C, while the other one was fixed in formalin, embedded in paraffin and underwent standard processing, sectioning and H&E staining (FFPE). The latter FFPE tissue sectioning was done along the cut surface closest to the fresh-frozen biobanked piece of tissue, so that the FFPE section is as much as possible representative for the tissue in the frozen specimen. Therefore, despite the degree of uncertainty/discrepancy in the macroscopic assessment in some situations (i.e. for prostatic adenocarcinoma, multifocal breast cancer, pancreatic adenocarcinoma with coexistent chronic pancreatitis, tumors after neoadjuvant therapy), every single biobanked tissue fragment (both tumor and normal) has its matching FFPE tissue that undergoes pathological verification of the actual tumor content.

The dispersed nature and scale of our biobanking required development of unified sample collection protocols, using well-defined clinical criteria for patient inclusion (Table 1). These protocols were developed in close collaboration between the molecularly-oriented team, surgeons involved in patient recruitment and treatment and pathologists collecting samples. In planning for material collection, we relied on our previous experience from studies of breast

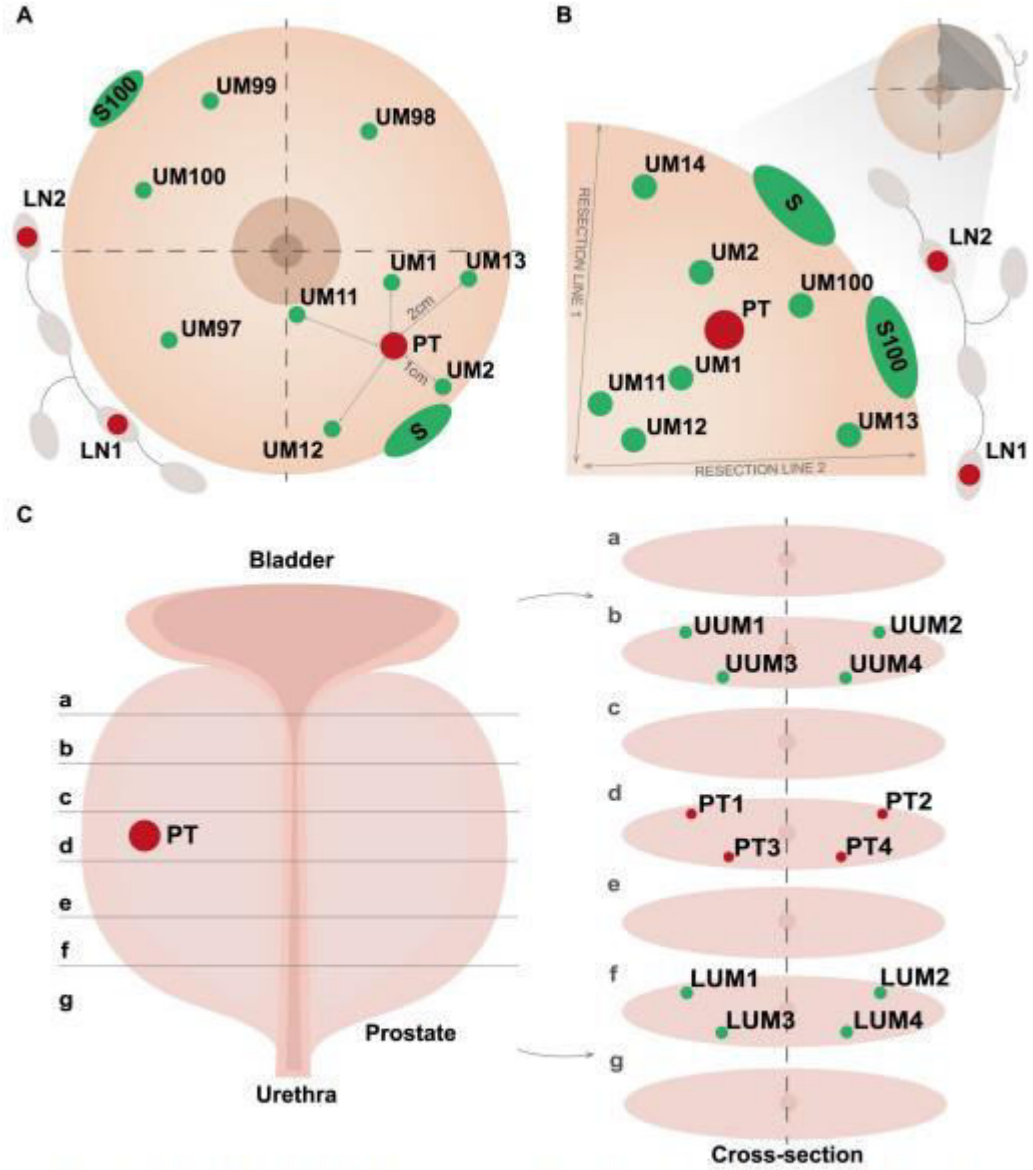

Fig 2. An illustration of sample collection protocols for breast- and prostate cancer. (A) Procedure for breast carcinoma samples treated with mastectomy. (B) Procedure for breast carcinoma patients treated with Breast Conservative Therapy (BCT). The distances in centimeters between samples of primary tumor and normal

tissue are illustrated in panel A with solid lines. (C) Protocol for prostatectomy with detailed scheme of sample collection in different cross-sections (a–g). Abbreviations: UM, uninvolved margin composed of macroscopically normal tissue; PT, primary tumor; S, skin; LUM, lower uninvolved margin; UUM, upper uninvolved margin; LN, regional lymph node. Detailed description of particular fragments for the protocols is given in the Materials and Method section.

https://doi.org/10.1371/journal.pone.0266111.g002

cancer [5,6,21]. In order to assure good quality of RNA/DNA extracted from the collected material, the general condition for all collected samples was a standardized time of usually 60–75 minutes (and maximum 2 hours) between tumor/organ resection and the moment of -80-degrees freezing for specimens dissected by the pathologist. It did not apply for the material used for Fluorescent Activated Cells Sorting (FACS) of leukocytes in the context of Loss Of chromosome Y (LOY) in male colorectal- and prostate cancer study as well as for material used for establishment of primary cell cultures (see below).

Tissue fragments dissected from mastectomy (Fig 2A) specimen include 1–4 PTs (depending on the tumor size and multifocality) and 9 UMs. Samples UM1 and UM2 located ~1 cm from PT (and presumably in the same lobe of the breast) towards the nipple and outer part, respectively. UMs 11–13 located ~2 cm from PT in the same quadrant, UMs 97 and 98 in two adjacent quadrants, UM99 in opposite quadrant, and S (skin) sample close to resection margin. BCT protocol (Fig 2B) involves collection of 1–2 PTs (depending on the tumor size), 7 UMs and 1 S. UMs 11–14 are collected in the vicinity of both resection lines; two of them (UMs 11 and 12) are closer to the nipple, while the other two (UMs 13 and 14) are further away from the center of the breast. Sample UM1 is taken between the nipple and PT, UM2 between PT and outer part of the breast. All the latter UMs are presumed to be located within the same quadrant and possibly the same lobe, in which the tumor is localized. In both breast cancer procedures, local metastases (LN) are also collected, when possible. Moreover, for both breast cancer surgical procedures, S100 and UM100 are dissected by surgeons directly at the operation theatre to a tube with sterile medium with antibiotics and sent to Gdansk within 24 hours to establish organoids and primary cell cultures (see below).

Due to the usually multifocal growth of prostate adenocarcinoma (constituting up to 95% of all prostate cancers) within this organ and frequent problems to macroscopically assess its exact location, the whole gland is sliced from base to the apex (Fig 2C). Four samples are taken with the punch in the slice that likely include the tumor: PT1 peripheral, right lobe; PT2 peripheral, left lobe; PT3 periurethral, right lobe; PT4 periurethral, left lobe. Analogically, four fragments of potentially unaffected tissue are collected in the slice located towards the base (UUM1–4) and apex (LUM1–4). A total of twelve tissue fragments are biobanked, each followed by the pathological report on the actual tumor content in the matching FFPE sample.

For colorectal cancer, up to 4 PTs (labeled PT1 and PT2 in case of multifocality, or A and B when two fragments are resected) and 8 UMs of mucosa (without muscular and serosa layer) were preserved: UMs 1–3 collected 1 cm away from the margin of PT, UMs 11–13 with 2 cm distance from the tumor, UM98 and 99 at least 5 cm from PT, with UM99 as the most distant one, located as far as possible from PT (Fig 3A). For metachronic metastasis (MT) of colorectal cancer to liver two fragments MT1 and MT11 (in the center and margin of the metastatic tumor respectively) and maximum two UM (close—UM1 and distant—UM99) are collected (Fig 3B). A similar pattern of samples is applicable for radical cystectomy (Fig 3C), while a unique protocol for TURBT involves one PT and two UM samples (Fig 3D). Collection for exocrine pancreatic cancer involves one PT and 1–3 UMs (depending on the type of resection and size of the material that is available) located about 1 cm from the margin of PT, while the UM99 is being the most remote fragment from PT (Fig 3E and 3F).

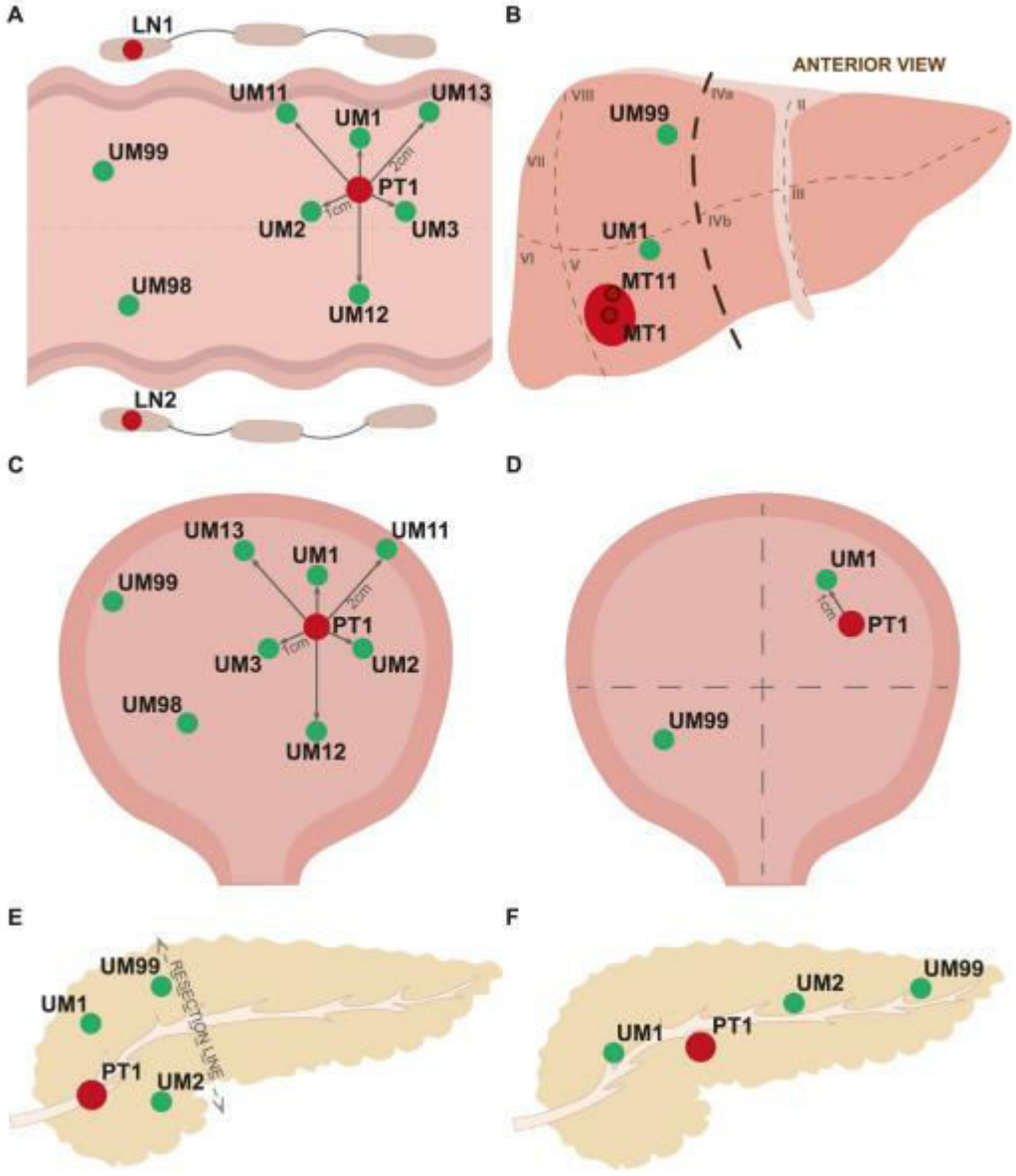
A
LN1
UM11
UM1
UM13
UM99
2cm
PT1
1cm
UM2
UM3
UM98
UM12
LN2
B
ANTERIOR VIEW
VIII
IVa
II
UM99
VII
III
IVb
UM1
VI
V
MT11
MT1
C
UM13
UM1
UM11
UM99
2cm
PT1
1cm
UM3
UM2
UM98
UM12
D
UM1
1cm
PT1
UM99
E
UM99
RESECTION LINE
UM1
PT1
UM2
F
UM2
UM99
UM1
PT1

Fig 3. An illustration of sample collection protocols for colorectal carcinoma and metastases of colorectal cancer to liver, urinary bladder- and exocrine pancreas carcinomas. (A) Scheme of sample collection for colorectal carcinoma. (B) Protocol of samples collection for metastases of colorectal cancer to liver. (C) Protocol of sample collection for urinary bladder after cystectomy. (D) Collection of samples for transurethral resection of tumor (TURBT). (E) Scheme of sample collection for the surgical removal of pancreas head. (F) Scheme of sample collection for the total pancreatomy. Primary tumors in all panels are drawn in red and samples of normal tissues in green. Abbreviations: UM, uninvolved margin composed of macroscopically normal tissue; PT, primary tumor; LN, regional lymph node; the lines show distances in centimeters from primary tumor.

https://doi.org/10.1371/journal.pone.0266111.g003

Cell cultures for breast cancer- and sorted leukocytes for Loss Of chromosome Y (LOY) project

In addition to the standard deep-frozen samples for biobanking, we also gathered additional unique material demanding dedicated procedures. Primary cultures from skin and uninvolved non-tumorous glandular tissue fragment of breast cancer patients (Fig 2A and 2B, S100 and UM100) yielding skin and stromal fibroblasts as well as organoids were initiated for a subset of cases using in house designed primary cell-culture protocols. For the same subset of patients, OCT blocks from skin and uninvolved margin were also prepared as a possible material for future spatial transcriptomics analysis.

In order to further study the association between LOY in blood and prostate- as well as male colorectal cancer, a total volume of 36 ml of peripheral blood was used for leukocyte sorting using FACS, as described previously [20]. Blood processing involved viable freezing of 2–4 million of peripheral blood mononuclear cells (PBMCs) and 0.1–4 million of $CD4^+$ cells for single cell analysis. The sorted fractions included: $CD19^+$ B cells, $CD8^+$ T cells, $CD4^+$ T-regulatory (Treg) and $CD4^+$ non-Treg cells, granulocytes, monocytes and natural killer (NK) cells for further DNA and RNA analysis. We also collected samples from whole blood and buccal swabs for males without cancer or Alzheimer disease diagnosis that were age-matched to the above-mentioned cohorts of prostate- and male colorectal cancer patients. These subjects will serve as controls and their blood leukocytes were sorted and preserved according to the above-mentioned scheme.

Development of dedicated IT solutions

The computerization and semi-automation of the sample collection was implemented from the very beginning at each collection site. This facilitated the process of preserving a very

Table 1. Inclusion and exclusion criteria for each diagnosis.

Diagnosis	*Inclusion criteria*
Breast cancer	BCT with/without neoadjuvant therapy Unilateral or bilateral mastectomy with/without neoadjuvant therapy
Colorectal cancer	Resection of uni- or multifocal primary tumor with/without neoadjuvant therapy
Liver metastasis	Resection of uni- or multifocal metachronous tumor with/without perioperative therapy
Prostate cancer	Prostatectomy with/without neoadjuvant therapy
Bladder cancer	TURBT with/without neoadjuvant therapy Radical cystectomy with/without neoadjuvant therapy
Pancreatic cancer	Uni- or multifocal *Adenocarcinoma* (exocrine cancer) * Pancreaticoduodenectomy ** without neoadjuvant therapy Total pancreatectomy without neoadjuvant therapy
Control group	Age ≥ 65 y.o. without oncological and Alzheimer Disease in clinical history

BCT–Breast Conservative Therapy; metachronic tumor stands for the secondary tumor that arose more than 6 months after diagnosis of first malignancy; TURBT–Transurethral Resection of Bladder Tumor, * —exclusion criteria: Preoperative neoadjuvant therapy, Endocrine cancer; ** —Pancreaticoduodenectomy (Whipple procedure)–operation performed to remove the cancerous head of the pancreas.

https://doi.org/10.1371/journal.pone.0266111.t001

high number of donors, reaching 100 per month without the risk of mix-up of samples. Each Pathology Department at the partner hospital was provided with a PC and a dedicated MABData1 software designed for that purpose. This program package has a simple, user-friendly web browser-based interface enabling registration of patients/samples with a set of clinical data (excluding personal information for safety reasons) and automated registration of tubes containing unique 2D codes. It also allows introduction of medical follow-up information at a later stage. All data is being synchronized every 5 minutes with MABServer in Gdansk using safe Advanced Encryption Standard (AES). The main goal of developing dedicated software for donor and sample registration at satellite hospitals was to introduce a user-friendly system that allows fast data synchronization with the central server, independence from network access and proper function also in an unstable internet connection environment. MABServer software is a proxy system (located at the 3P-Medicine Laboratory at the Medical University of Gdansk) that holds data from all hospitals for further use. Moreover, this system was developed also for safety reasons; in case of any MABData1 computer failure all data is kept at MABServer, which is included in the central backup schedule at the 3P-Medicine Laboratory. The management of samples and data locally is being assisted with MABData2 system, allowing donor and sample pseudonymization, registration of original samples from hospitals and numerous types of derivative specimens, together with many additional parameters, adding extra attachments, simple and advanced searching and data exporting, as well as inventory of samples. MABData2 is a complete, stand-alone biobank management system that supports not only current biobanking project, but is also the main software solution, connecting data from other projects and collections for further cross-searches. The documentation received on paper is pseudonymized, scanned and uploaded to the MABData2 system, using implementation of Optical Character Recognition (OCR) algorithms inserted into the database.

Bioethical approval

All procedures for sample collection were approved by the Independent Bioethics Committee for Research at the Medical University of Gdansk (approval number NKBBN/564/2018 with multiple amendments). This approval is valid for collection of samples at multiple collaborating hospitals. Since our collection of samples is still ongoing, our initial ethical approval provides us with the possibility to extend the scope of biobanking for additional diagnoses, after an amendment of the application. Written informed consent was obtained from all the patients prior to surgery. All procedures were performed in accordance with the relevant national and international laws and guidelines as well as in compliance with European Union General Data Protection Regulation (EU GDPR).

Other laboratory procedures

DNA extraction from frozen solid tissues was performed using standard phenol/chloroform method with several in-house modifications (depending on the tissue type) and lysis buffer with SDS (0.5% SDS, 60mM Tris pH 7.9, 5 mM EDTA) or sarcosine and SDS (2% sarcosine, 0.5% SDS, 50mM Tris pH 7.9, 10 mM EDTA); DNA from whole blood was acquired using QIAamp DNA Blood Midi Kit (Qiagen) or QuickGene DNA whole blood kit S (Kurabo) with QuickGene-Mini 480 instrument (Kurabo), DNA from sorted cells was extracted either using sarcosine lysis buffer (1% sarcosine, 10 mM Tris-HCl, 50mM NaCl, 10 mM EDTA) followed by ethanol precipitation (<500 000 cells) or with QIAamp DNA Mini Kit (Qiagen) (>500 000 cells). All laboratory processes using our samples were carried out according to standard protocols including: sorting of leukocytes on FACS machines, establishing primary cultures,

extraction of DNA and RNA for various downstream applications; e.g. droplet digital PCR (ddPCR), construction of NGS libraries for targeted DNA/RNA sequencing.

Results

Rationale

The main rationale and aims of our biobanking program are systematic explorations PZMs especially in normal tissues, which is an understudied field of cancer research. We also incorporate material for studies of mosaic LOY in males that might predispose to various diseases, with a particular focus on cancer. The selected diagnoses cover sporadic breast-, colorectal-, prostate-, bladder- and pancreatic carcinomas and represent consecutively collected patients, affected by common and often fatal diseases. The setup of our biobank should allow a wide range of "omics"- and other methods to be applied in studies of the collected clinical material.

General statistics

The first patient was registered in the database on June 5, 2019 and statistics described here are up to May 12, 2021. Our collection originating from five clinical cancer centers, after nearly 2-years of cooperation, reached in total 1963 donors. However, 1711 of these were collected effectively (i.e. with all tissue types from the collection protocols, described above in Materials and Methods as compulsory set) yielding 23226 independent samples (Fig 4A and 4B). This results with an average of 74 donors and 1010 samples collected per month. Incomplete sampling affected 13% of all donors and was caused by several factors, such as small size of tumor resulting in insufficient amount of tissue material, prolonged surgery time and inability to prepare material by pathologists the same day, as well as unresectability of the tumor. The predominant diagnosis was breast carcinoma (933 donors), followed by colorectal- (383 donors), prostate- (221 donors), urinary bladder- (81 donors), exocrine pancreas carcinomas (15 donors) and metachronous metastases of colorectal cancer to liver (14 donors). We also collected blood and buccal swabs from 64 healthy male control subjects that were age matched for the cohort of males with colorectal- and prostate carcinoma (Table 2). The average age of male oncological patients was 67 years (1 SD ±9, range 33–93 years), while for female oncological donors it was 62 years (1 SD ±13, 24–92 years) and for healthy control subjects 71 years (1 SD ±5, 61–91 years).

Fig 4C–4F shows the distribution of ICD-10 codes within particular diagnoses: C50 (malignant neoplasm of breast), C18—C21 (malignant neoplasm of colon, rectosigmoid junction, rectum, anus and anal canal), C67 (malignant neoplasm of bladder), and C25 (malignant neoplasm of pancreas). Within colorectal diagnoses clear predominance of rectal cancers (C20, 99 cases) is notable, followed by tumors of caecum (C18.0, 57 donors) and sigmoid colon (C18.7, 53 cases). In bladder cancer patients, the affected sites include trigone and lateral wall of the bladder (C67.0 and C67.2 and 34 and 25 donors, respectively). The vast majority of pancreatic tumors are located in the body of the organ (C25.0, 12 patients). In addition to ICD-10 diagnoses, we also gathered other clinical information regarding the donors recruited to the project, which is summarized in Table 3. It covers basic clinical data: imaging results for the patient, type of surgery, dates of initial diagnosis and treatment, blood count, as well as full histopathological report with data for microscopic examination of all resected tissue fragments. Moreover, we collected information from each donor via medical questionnaires covering oncological history in the family, chronic illnesses and smoking habits.

The uniqueness of our collection is illustrated by the number of fragments dissected from macroscopically normal cancer-neighboring margin of tissue (named UM, LUM or UUM) (Figs 1 and 2), which accounts for nearly 40% of the total sample count, while the contribution from

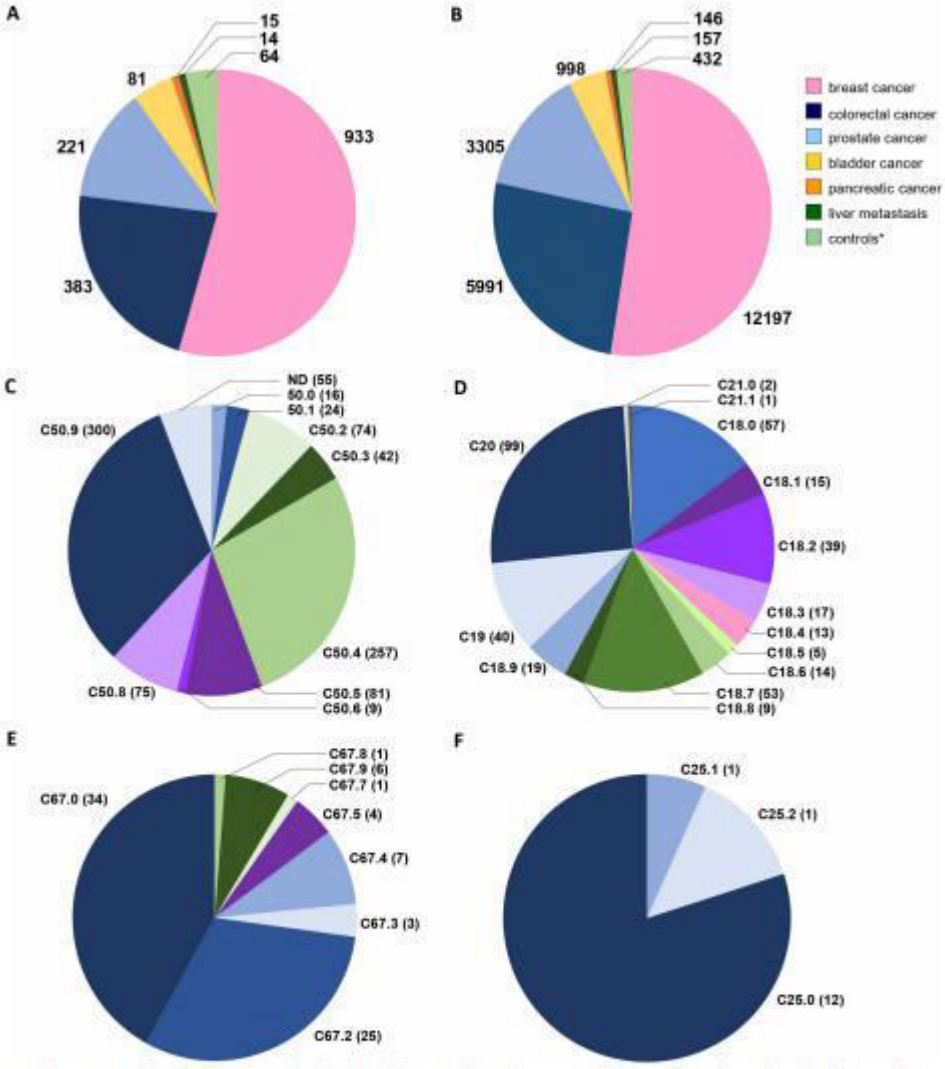

Fig 4. The statistics of donors and samples collected in five collaborating hospitals; status as of May 12, 2021. (A) The total number of donors with compulsory set of samples (as described in Materials and Methods-section and shown in Figs 1 and 2). (B) The sum of all samples collected from recruited donors. The numbers for donors and samples are divided for six different cancer diagnoses and controls (*). The control (*) category represents a healthy male cohort ≥ 65-year-old recruited as controls for patients with prostate- and colorectal cancer, used in the Loss of Y Chromosome (LOY) project. (C–F) show distribution of diagnoses according to International Classification of Diseases (ICD-10, World Health Organization) for breast, colorectal, bladder and pancreatic cancer patients, respectively. Abbreviations: C50, Malignant neoplasm of breast; C50.0, Nipple and areola; C50.1, Central portion of breast; C50.2, Upper-inner quadrant of breast; C50.3, Lower-inner quadrant of breast; C50.4, Upper-outer quadrant of breast; C50.5, Lower-outer quadrant of

breast; C50.6, Axillary tail of breast; C50.8, Overlapping lesion of breast; C50.9, Breast, unspecified; C18, Malignant neoplasm of colon; C18.0, Caecum, fleocaecal valve; C18.1, Appendix; C18.2, Ascending colon; C18.3, Hepatic flexure; C18.4, Transverse colon; C18.5, Splenic flexure; C18.6, Descending colon; C18.7, Sigmoid colon, Sigmoid (flexure); C18.8, Overlapping lesion of colon;C18.9, Colon, unspecified, Large intestine, unspecified; C19, Malignant neoplasm of rectosigmoid junction, including colon with rectum, rectosigmoid colon; C20, Malignant neoplasm of rectum, Including rectal ampulla; C21, Malignant neoplasm of anus and anal canal; C21.0, Anus, unspecified, excluding anal margin and perianal skin; C21.1, Anal canal, Anal sphincter; C67, Malignant neoplasm of bladder; C67.0, Trigone of bladder; C67.2, Lateral wall of bladder; C67.3, Anterior wall of bladder; C67.4, Posterior wall of bladder; C67.5, Bladder neck, Internal urethral orifice; C67.7, Urachus; C67.8, Overlapping lesion of bladder; C67.9, Bladder, unspecified; C25, Malignant neoplasm of pancreas; C25.0, Head of pancreas; C25.1, Body of pancreas; C25.2, Tail of pancreas. ND–not yet defined due to temporary lack of medical documentation.

https://doi.org/10.1371/journal.pone.0266111.g004

tumor samples is 12%. The remaining sample represent blood, plasma, skin and sorted leukocytes using FACS. Furthermore, whenever possible we also collected material from local metastases to lymph nodes, for a total of 117 and 80 donors with breast and colorectal cancer, respectively (1% of all samples). The implemented sampling procedures are demanding in terms of time needed for material preparation. The estimated average time devoted by the Pathology Departments is 98 minutes per patient to fulfill the requirements of our protocols (ranging from 60 minutes for liver metastasis resection to 155 minutes for prostatectomy), which include patient registration, blood processing for collection of tubes with peripheral blood and plasma, tissue dissection by pathologists, preparation of FFPE blocks/slides, and histopathological assessment. Thus, the collection from 1711 donors sums up to about 2800 working hours or about 70 working weeks with full time effort. This time represents only part of the entire project, as we do not include time spent on the following tasks: recruitment of patients by surgeons, collecting

Table 2. A summary of donors and cancer diagnoses included in the collection (status as of May 12, 2021).

Diagnosis	*Sex distribution*	*Average age*	*Average samples (range)*	*Clinical information*
Breast cancer (ICD-10 C50)	F– 99% (n = 921) M– 1% (n = 12)	60 y ± 13 64 y ± 8	13 (7–32)	• Mastectomy– 42% (n = 391) • BCT– 53% (n = 494) • Non-specified* – 5% (n = 48)
Colorectal cancer (ICD-10 C18 –C21)	F– 45% (n = 174) M –55% (n = 209)	66 y ± 12 67 y ± 10	16 (7–27)	
Liver metastasis (colorectal cancer) (ICD-10 C78.7)	M– 86% (n = 11) F– 14% (n = 3)	65 y ± 11 66 ± 4	11 (7–15)	
Prostate cancer (ICD-10 C61)	M– 100% (n = 221)	65 y ± 7	15 (7–19)	Gleason score: • 3+2–0.5% (n = 1) • 3+3–16% (n = 35) • 3+4–28% (n = 61) • 4+3–28% (n = 61) • 4+4–2% (n = 5) • 4+4+5–0.5% (n = 2) • 4+5–1% (n = 3) • Non-specified* – 24% (n = 53)
Bladder cancer (ICD-10 C67)	F– 21% (n = 17) M– 79% (n = 64)	68 y ± 8 69 y ± 9	12 (7–17)	• TURBT– 47% (n = 38) • Cystectomy– 53% (n = 43)
Pancreatic cancer (exocrine, ICD-10 C25)	Female– 47% (n = 7) Male– 53% (n = 8)	68 y ± 6 66 y ± 9	10 (8–11)	
Control group**	Male– 100% (n = 64)	71 ± 5	7 (6–12)	

F–female; M–male; *—information not yet available and incorporated in our database; y–years; BCT–Breast Conservative Therapy. ICD codes: C50, Malignant neoplasm of breast; C18, Malignant neoplasm of colon; C19, Malignant neoplasm of rectosigmoid junction; C20, Malignant neoplasm of rectum; C21, Malignant neoplasm of anus and anal canal; C78.7, Secondary malignant neoplasm of liver and intrahepatic bile duct; C61, Malignant neoplasm of prostate; C67, Malignant neoplasm of bladder; C25, Malignant neoplasm of pancreas; TURBT–Transurethral Resection of Bladder Tumor.

**—Healthy male cohort ≥ 65 years old recruited as controls for the male patients with prostate and colorectal cancer for whom the white blood cell fractions were sorted by FACS to study loss of chromosome Y.

https://doi.org/10.1371/journal.pone.0266111.t002

Table 3. Type of data collected from medical records of patients.

Document	*Information*
Registration form in the hospital	• Clinical data available from the hospitals: CT, PET, MRI, RTG, USG, colonoscopy, mammography, urography, cystoscopy, scintigraphy; • Date of first diagnosis and treatment; • Type of surgery (radical cystectomy/TURBT, mastectomy/BCT, type of pancreatectomy).
Histopathological report	• Histopathological type of cancer; • Microscopic description of tumor/non-tumoral tissue; • Grade of cancer; • pTNM stage; • Gleason score (ICD-10 C61); • ER, PR, HER2 status (ICD-10 C50); • Ki-67 (ICD-10 C50); • Size of resected organ.
Patient questionnaire	• Smoking status; • Chronic illnesses; • Family history–oncological treatment, Alzheimer Disease.
Complete blood count	• Red Blood Count; • White Blood Count; • Platelets.

CT–Computer Tomography; PET–Positron Emission Tomography; MRI–Magnetic Resonance Imaging; RTG–radiography; USG–ultrasonography. ICD codes: C50, malignant neoplasm of breast; C61, malignant neoplasm of prostate; TURBT, Transurethral Resection of Bladder Tumor; BCT, Breast Conservative Therapy; ER, Estrogen Receptor status; PR, Progesterone Receptor status; HER2, status of the human epidermal growth factor receptor; Ki-67, marker of proliferation Ki-67.

https://doi.org/10.1371/journal.pone.0266111.t003

informed consents and filling of questionnaires, local transport within each hospital and from five partner hospitals to Gdansk, as well as acquisition of material by the 3P-Medicine Laboratory in Gdansk. Furthermore, in the time devoted to development and testing of software (MAB-Data1 and MABData2) we do not count preparation of common protocols for sample collection as well as preparation of formal agreements with the participating hospitals.

Another important aspect of our biobanking approach is the histopathological verification of the tissue fragments that were macroscopically presumed to represent either tumor or non-tumor fragment and were later verified by analysis of FFPE sections. We performed such comparison for prostate cancer and colorectal cancer. As suspected, prostate cancer represents a challenge due to the lack of clear macroscopic demarcation of the tumor. The calculation on the representative cohort of 100 prostate cancer patients showed that out of 400 specimens collected presumably as tumorous tissue, 263 (65.7%) showed the presence of tumor cells. For the similar sample size of 100 colorectal cancer patients, 205 out of 205 (100%) presumed tumor samples were confirmed as containing tumor tissue. It is important to mention that each fragment selected for further molecular examination is verified with the histopathological report.

Analysis of smoking status, which is included in the medical questionnaire (Fig 5), revealed that nearly 50% of female cancer patients declared themselves as non-smokers and 31% as past-smokers. Corresponding numbers of non-smokers among males is clearly lower (28%) and past-smokers is much higher (54%). Similar numbers of females and males declared themselves as present smokers (16% and 14%, respectively), and this is also valid for control subjects (19%).

Specific diagnoses

The summary of donors and samples for all diagnoses is shown in Table 2. Breast cancer is the most common type of cancer diagnosed world-wide in 2020 [22] (https://www.who.int/news-

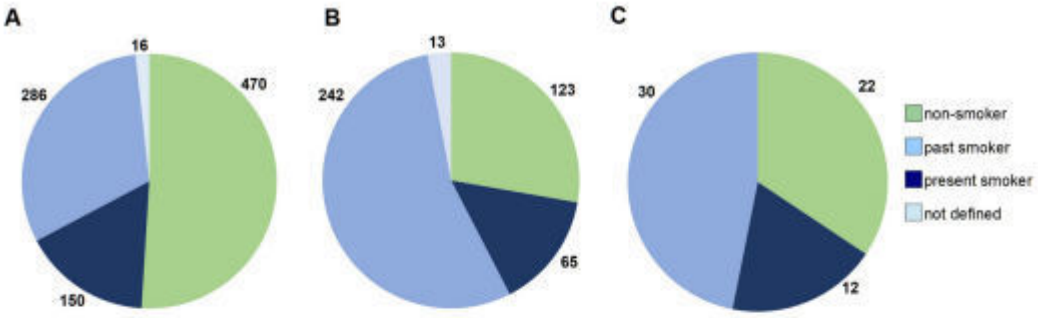

Fig 5. Smoking status declared by donors in the questionnaire. (A) Female cohort. (B) Male cohort. (C) Control male group, as described in Materials and Methods.

https://doi.org/10.1371/journal.pone.0266111.g005

room/fact-sheets/detail/cancer) and it is also well represented in our collection, accounting for 55% of all donors, with obvious predominance of women and only 1% of males with breast cancer. The average age for females undergoing mastectomy or BCT is 60 years (1 SD ±13), and the corresponding number for males is 64 years (1 SD ±8). Counting all breast surgeries together, 53% are BCTs, and 42% are mastectomies, remaining 5% are not defined due to the temporary lack of medical documentation. Colorectal cancer is the second most common diagnosis in the collection (22% of donors), with the predominance of male patients (55% versus 45% of females), with the average age of initial surgical treatment similar for both sexes (66 versus 67 years). Due to the usually large extent of the resection of the colon, we reached the highest average sample count (n = 16) among all diagnoses. Prostate cancer in our assembly is diagnosed and treated on average at the age of 65 years, at a medium-grade stage: Gleason score 3+4 (27% of patients) and 4+3 (27%), low medium-grade: Gleason score 6 (15%) and rarely at a high-grade, Gleason score ≥8 (3.5%). The collection procedure for prostatectomy is the most complicated of all diagnoses and the average sample count for prostate cancer is the second highest (n = 15). Organ sparing treatment using TURBT, was applied for 47% of all gathered bladder cancer donors, which adds to the uniqueness of our collection. The frequency of bladder cancer among males is much higher than for females (79% versus 21%), with a similar average age of onset for both sexes (69 and 68 years, respectively). Exocrine pancreas cancer, which is frequently unresectable and the most fatal disease in our collection, stands for below 1% of all our donors and has a similar distribution of age and sex.

Well-defined recruitment criteria for volunteers as healthy controls for the LOY project (Table 1) resulted in a homogeneous group of male subjects, with an average age of 71 years and average sample count of seven. The starting material is peripheral blood for preparation of viable PBMCs and $CD4^+$ T-cells and sorting of seven populations of leukocytes using FACS. Additionally, buccal swabs are collected as reference material for non-mesoderm-derived tissue. The same procedure for sorting of leukocytes, as above for controls, was applied for 20 prostate and 28 colorectal cancer patients.

Quality of derivative samples

The collected samples have already been used for subsequent research activities and a wide variety of downstream applications. Hence, we optimized and implemented several laboratory protocols that include DNA and RNA preparation and aliquoting. We present here the

concentrations and quality of DNA extracted from solid tissue (PT and UMs), blood and the pellets of FACS sorted cells (Fig 6). The amount of tissue used for DNA preparation differs widely for various diagnoses and is lowest for bladder and prostate samples (starting from 2 mg of fresh frozen tissue), where the collected fragments usually do not exceed 8 mm^3. However, the quality, integrity (DIN) and spectroscopic purity of DNAs were high for all samples (Fig 6B and 6C) with the exception of A260/230 ratio for DNAs from sorted cell pellets, especially from the subpopulations with low cell count after sorting (below 150 000 cells) (Fig 6C). This quality is, however, still sufficient to run ddPCR analysis.

Dedicated software for dispersed biobanking

In a typical research project oriented on sample/donor collection, software solutions are built in central models as one system case with a single pipe for data input. In our project, we created a dedicated solution for dispersed biobanking at satellite hospitals. Each partner hospital is actually an independent biobank, equipped with both hardware and software solutions allowing sample/data registration as well as storage, including a possibility of *a posteriori* data verification and supplementation. This developed system and communication methods bring data safety standards that correspond to ISO-270001 requirements. The most important features are included in MABData2 software. This package is not only the tool dedicated for biobanking collection, administration and management, but also allows for customized data search, connecting different types of data and parameters describing samples and donors. Both MABData1 and MABData2 packages are owned by the Medical University of Gdansk, which has full license- and copyrights. Therefore, this in-house developed software allows independence for future development and for any number of users.

Discussion

One of the major aims of our biobanking effort is to gather a representative collection of samples from multiple common sporadic cancer diagnoses, which would primarily allow large-scale studies of early predisposing PZMs that initiate tumor development in normal cells from

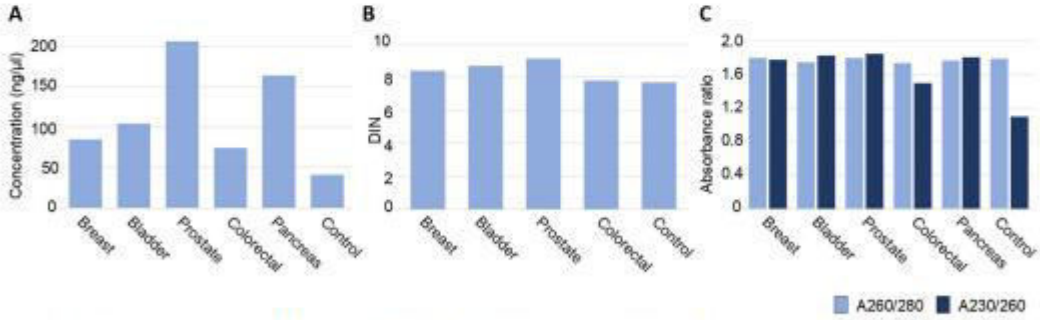

Fig 6. DNA quality isolated from collected tissues. (A) Concentration of DNA (ng/µl) obtained from blood and tissues from donors per diagnoses, measured with the fluorometric method (Qubit Fluorometric Quantification and/or Agilent TapeStation System). (B) DNA Integrity Number (DIN) for DNA obtained from blood and tissue in four diagnoses measured by Genomic DNA ScreenTape Analysis (Agilent). (C) DNA quality measured by UV-Vis spectroscopic method for DNA obtained from blood and tissues from donors with five diagnoses. The number of samples for each diagnosis that was used for calculations are: 88 for breast cancer; 202 for colorectal cancer; 3092 for prostate cancer; 189 for bladder cancer; 12 for pancreas cancer; and 62 for controls. The amount of frozen tissue used for DNA extractions range as follows: Breast cancer 15–60 mg; bladder cancer 2–19 mg; prostate cancer 6–43 mg; colorectal cancer 15–33 mg; pancreas cancer 12–21 mg; and controls (sorted leukocytes) $0.05x10^6$ – $1x10^6$ cells.

https://doi.org/10.1371/journal.pone.0266111.g006

the affected organ. Consequently, the largest number of samples are derived from histologically verified non-tumorous tissue, which is located at various distances from the site of primary tumor. The design of our biobanking is based on our previous experience from studies of multifocal breast cancer [5,6,21]. Furthermore, availability of material from local metastases to lymph nodes for breast and colorectal cancer as well as distant metastases of colorectal cancer to the liver allows for assessment of tumorous tissue at different stages of disease. Moreover, we collect at least one additional reference tissue that is not related to tumor development; usually blood or skin (Figs 1 and 2). This reference material is important whenever a possibly pathogenic variant is found, in order to confidently exclude that it could represent a germline mutation. The information regarding oncological family history gathered from the patient questionnaire might also be helpful in such cases. Collection of multiple tubes of plasma (liquid biopsy) from donors suffering from cancer and from healthy controls is also crucial for future genetic and proteomic analyses. The highest number of independent samples per single donor is thirty-two (Table 2). Thus, according to the official BBMRI-ERIC directory (https://directory.bbmri-eric.eu/#/), our collection is comprehensive and represents the largest assembly of samples oriented towards studies of early events in cancer development. This collection will allow essentially all available "-omics" approaches on DNA-, RNA-, protein- and tissue levels as well as many other methodological approaches to be applied. Our biobanking project also assumes acquisition of long-term follow-up (3 to 10 years after treatment) for recruited patients, which will be an important added value. As mentioned above, our biobanking effort is still ongoing and we have an opportunity to modify collection protocols and include other cancer diagnoses. The collected biological material and clinical data can be made available for other investigators after a request to both corresponding authors. The letter should outline the aim, number/type of requested samples and methodology of the proposed collaborative project. Such scientific cooperation can be established based on bilateral scientific cooperation agreement.

The number of recruited donors already provides us a good perspective regarding statistics of the four most frequent diagnoses in our study. For the breast carcinoma, a clear predominance (32%) of C50.9 diagnosis (Breast, unspecified) is apparent together with the relative high incidence of ICD 50.8 (Malignant neoplasm of overlapping sites of breast), which suggest unclear location or lack of the record in the patient documentation. This is followed by the carcinomas located in upper outer (ICD 50.4–27%), lower outer (ICD 50.5–9%), and upper inner (ICD 50.2–8%) quadrants of the breast. This trend reflects the distribution observed in other large-scale study [23]. In the case of colorectal cancer, the location of tumor is a crucial factor determining molecular type, disease progression, prognosis, treatment and outcome [24]. In our collection 38% of cases are right-sided colon cancers (RSCC/proximal): ICD-10 from C18.0 to C18.5 and 61% represent left-sided colon cancers (LSCC/distal): C18.6 to C19, C20 (Fig 4D). The distribution is in accordance with a larger study [25]. RSCCs are generally higher in females and have worse overall prognosis—in our study it is comparable for both sexes (21% females vs 19% males), the opposite trend is observed in LSCC (25% for females and 35% in this study). Such a representation and distribution of donors and samples in the collection will enable the examination of molecular heterogeneity, etiology and progression of these cancers.

For breast- and urinary bladder cancer patients, we collected samples derived from two different surgical procedures. Breast cancer operations can be performed using either mastectomy or BCT and the frequency of these two surgeries in our material are 42% and 53%, respectively. The tumors removed via BCT are diagnosed at earlier stage of development, are typically smaller and these patients usually do not show suspicion of multifocal tumor development. The trend in our cohort reflects the worldwide shift in surgical approach

from mastectomy to BCT [26]. Bladder cancer patients are qualified for either cystectomy or sparing therapy using (TURBT) and their respective frequencies of these treatments in our collection are 53% and 47%. Hence, we have an opportunity to molecularly study these two common cancers in patients with different grades of severity of the disease, which represents an added value.

It is well known that males have a higher incidence and mortality from most sex-unspecific cancers, which is largely unexplained by known risk factors [27,28]. This substantial cancer-related sex-disparity appears to be a neglected aspect of cancer research. Our collection of donors for bladder- and colorectal cancer confirms this. Bladder cancer is much more frequent among males than females (79 versus 21), with a similar average age of onset for both sexes (69 and 68 years, respectively) (Table 2). Colorectal cancer also shows male predominance (55% males vs. 45% females), although the difference is less pronounced. Smoking habits among males in Poland might be, at least partially, responsible for these differences. As shown in Fig 5, 47% and 68% oncological female and male patients, respectively, declared themselves as past- or present smokers, and there is a proven correlation between cancer incidence, smoking habits and LOY for men [18,29,30]. Our ongoing effort to obtain sorted subpopulations of leukocytes using FACS from males treated for prostate and colorectal cancer will help to study and possibly provide further support for this hypothesis.

Supporting information

S1 Table. Pseudonymized list of patients that were included in the statistics presented in the manuscript with the information on diagnosis, age, sex, ICD10, type of surgery where applicable, number of original samples and the status of smoking declared in the medical questionnaire.
(XLSX)

Acknowledgments

We thank all the patients and volunteer controls for acceptance to participate, sample contribution and information provided in the medical questionnaire. We thank Dr. Leszek Kalinowski for use of the temporary office, laboratory and freezer space as well as access to other laboratory facilities. We acknowledge Drs. Darek Kędra, Marco Günthel, and Paweł Olszewski for consultations regarding laboratory and bioinformatic procedures. We also thank physicians and nurses involved in the patient recruitment process, collaborating technicians, diagnosticians and pathologists from: Oncology Center—Prof. Franciszek Łukaszczyk Memorial Hospital in Bydgoszcz (Jowita Nowaczewska, Katarzyna Krzysiak, Anetta Słupicka, Mateusz Matusiak); Maria Skłodowska-Curie National Research Institute of Oncology in Kraków (Justyna Wajda, Dorota Lech, Kaja Majchrzyk); University Clinical Centre in Gdańsk (Wojciech Biernat, Wojciech Polom, Jakub Gondek, Tomasz Cwaliński, Grażyna Stęplewska, Elżbieta Wierszyło, Elżbieta Pietruszka, Grażyna Dombrowska, Paweł Górny, Małgorzata Derwis, Justyna Pietruszewska, Irena Pellowska, Michał Kunc, Aleksandra Korwat, Ewa Miłoszewska, Aleksandra Kaczor, Dawid Foltynowski); University Hospital in Cracow (Weronika Natkaniec, Magdalena Smolik, Małgorzata Molus, Jadwiga Galek, Hieronim Strojniak, Adrian Głownia, Anna Janas, Iwona Zawadzka, Monika Cala, Joanna Ciężarek, Shymko Anzhela, Izabela Pabisz-Zarębska); and Specialist Hospital in Kościerzyna (Atanasiu Apostolis, Barbara Koenner, Michał Kujach, Renata Knuth). We thank Drs. Daniil Sarkisyan and Eva Tiensuu Janson for critical review of the manuscript.

Author Contributions

Conceptualization: Natalia Filipowicz, Agata Wojdak, Jakub Szymanowski, Jarosław Skokowski, Arkadiusz Piotrowski, Jan P. Dumanski.

Data curation: Natalia Filipowicz, Kinga Dręże k, Jakub Szymanowski, Jan P. Dumanski.

Formal analysis: Natalia Filipowicz, Jakub Szymanowski.

Funding acquisition: Arkadiusz Piotrowski, Jan P. Dumanski.

Investigation: Natalia Filipowicz, Kinga Dręże k, Monika Horbacz, Edyta Rychlicka-Buniowska, Ulana Juhas, Katarzyna Duzowska, Wiktoria Stańkowska, Katarzyna Chojnowska, Maria Andreou, Urszula Ławrynowicz, Magdalena Wójcik.

Methodology: Natalia Filipowicz, Monika Horbacz, Agata Wojdak, Edyta Rychlicka-Buniowska, Ulana Juhas, Tomasz Nowikiewicz, Katarzyna Chojnowska, Hanna Davies, Ewa Śrutek, Michał Bieńkowski, Jarosław Skokowski, Krzysztof Okoń, Stanisław Hać, Łukasz Kaska, Michał Jankowski, Diana Hodorowicz-Zaniewska, Rafał Pęksa, Joanna Szpor, Janusz Ryś, Łukasz Szylberg, Arkadiusz Piotrowski, Jan P. Dumanski.

Project administration: Natalia Filipowicz, Agata Wojdak, Arkadiusz Piotrowski, Jan P. Dumanski.

Resources: Jakub Szymanowski, Tomasz Nowikiewicz, Ewa Śrutek, Michał Bieńkowski, Katarzyna Milian-Ciesielska, Marek Zdrenka, Aleksandra Ambicka, Marcin Przewoźnik, Agnieszka Harazin-Lechowska, Agnieszka Adamczyk, Jacek Kowalski, Dariusz Bała, Dorian Wiśniewski, Karol Tkaczyński, Krzysztof Kamecki, Marta Drzewiecka, Paweł Wroński, Jerzy Siekiera, Izabela Ratnicka, Jerzy Jankau, Karol Wierzba, Jarosław Skokowski, Karol Połom, Mikołaj Przydacz, Łukasz Bełch, Piotr Chłosta, Marcin Matuszewski, Krzysztof Okoń, Olga Rostkowska, Andrzej Hellmann, Karol Sasim, Piotr Remiszewski, Marek Sierżęga, Stanisław Hać, Jarosław Kobiela, Łukasz Kaska, Michał Jankowski, Diana Hodorowicz-Zaniewska, Janusz Jaszczyński, Wojciech Zegarski, Wojciech Makarewicz, Rafał Pęksa, Joanna Szpor, Janusz Ryś, Łukasz Szylberg, Jan P. Dumanski.

Software: Jakub Szymanowski.

Supervision: Natalia Filipowicz, Arkadiusz Piotrowski, Jan P. Dumanski.

Visualization: Natalia Filipowicz, Monika Horbacz, Jan P. Dumanski.

Writing – original draft: Natalia Filipowicz, Monika Horbacz, Jakub Szymanowski, Jan P. Dumanski.

Writing – review & editing: Natalia Filipowicz, Kinga Dręże k, Monika Horbacz, Agata Wojdak, Jakub Szymanowski, Edyta Rychlicka-Buniowska, Ulana Juhas, Katarzyna Duzowska, Tomasz Nowikiewicz, Wiktoria Stańkowska, Katarzyna Chojnowska, Maria Andreou, Urszula Ławrynowicz, Magdalena Wójcik, Hanna Davies, Ewa Śrutek, Michał Bieńkowski, Katarzyna Milian-Ciesielska, Marek Zdrenka, Aleksandra Ambicka, Marcin Przewoźnik, Agnieszka Harazin-Lechowska, Agnieszka Adamczyk, Jacek Kowalski, Dariusz Bała, Dorian Wiśniewski, Karol Tkaczyński, Krzysztof Kamecki, Marta Drzewiecka, Paweł Wroński, Jerzy Siekiera, Izabela Ratnicka, Jerzy Jankau, Karol Wierzba, Jarosław Skokowski, Karol Połom, Mikołaj Przydacz, Łukasz Bełch, Piotr Chłosta, Marcin Matuszewski, Krzysztof Okoń, Olga Rostkowska, Andrzej Hellmann, Karol Sasim, Piotr Remiszewski, Marek Sierżęga, Stanisław Hać, Jarosław Kobiela, Łukasz Kaska, Michał Jankowski, Diana Hodorowicz-Zaniewska, Janusz Jaszczyński, Wojciech Zegarski, Wojciech Makarewicz,

Rafał Pęksa, Joanna Szpor, Janusz Ryś, Łukasz Szylberg, Arkadiusz Piotrowski, Jan P. Dumanski.

References

1. Frank C, Fallah M, Sundquist J, Hemminki A, Hemminki K. Population Landscape of Familial Cancer. Sci Rep. 2015; 5:12891. Epub 2015/08/11. https://doi.org/10.1038/srep12891 PMID: 26256549; PubMed Central PMCID: PMC4530455.
2. Hemminki K, Sundquist J, Bermejo JL. How common is familial cancer? Ann Oncol. 2008; 19(1):163–7. Epub 2007/09/07. https://doi.org/10.1093/annonc/mdm414 PMID: 17804474.
3. Scheuner MT, McNeel TS, Freedman AN. Population prevalence of familial cancer and common hereditary cancer syndromes. The 2005 California Health Interview Survey. Genet Med. 2010; 12(11):726–35. Epub 2010/10/06. https://doi.org/10.1097/GIM.0b013e3181f30e9e PMID: 20921897.
4. Forsberg LA, Gisselsson D, Dumanski JP. Mosaicism in health and disease—clones picking up speed. Nat Rev Genet. 2017; 18(2):128–42. Epub Dec 12. https://doi.org/10.1038/nrg.2016.145 PMID: 27941868.
5. Ronowicz A, Janaszak-Jasiecka A, Skokowski J, Madanecki P, Bartoszewski R, Balut M, et al. Concurrent DNA Copy-Number Alterations and Mutations in Genes Related to Maintenance of Genome Stability in Uninvolved Mammary Glandular Tissue from Breast Cancer Patients. Hum Mutat. 2015; 36 (11):1088–99. Epub 2015/07/30. https://doi.org/10.1002/humu.22845 PMID: 26219265.
6. Forsberg LA, Rasi C, Pekar G, Davies H, Piotrowski A, Absher D, et al. Signatures of post-zygotic structural genetic aberrations in the cells of histologically normal breast tissue that can predispose to sporadic breast cancer. Genome Res. 2015; 25(10):1521–35. Epub 2015/10/03. https://doi.org/10.1101/gr.187823.114 [pii]. PMID: 26430163; PubMed Central PMCID: PMC4579338.
7. Gadaleta E, Fourgoux P, Pirro S, Thorn GJ, Nelan R, Ironside A, et al. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. NPJ Breast Cancer. 2020; 6:38. Epub 2020/09/05. https://doi.org/10.1038/s41523-020-00182-9 PMID: 32885042; PubMed Central PMCID: PMC7442642.
8. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015; 348(6237):880–6. https://doi.org/10.1126/science.aaa6806 PMID: 25999502; PubMed Central PMCID: PMC4471149.
9. Yizhak K, Aguet F, Kim J, Hess JM, Kubler K, Grimsby J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science. 2019; 364(6444). Epub 2019/06/07. https://doi.org/10.1126/science.aaw0726 PMID: 31171663; PubMed Central PMCID: PMC7350423.
10. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019; 574(7779):532–7. Epub 2019/10/28. https://doi.org/10.1038/s41586-019-1672-7 PMID: 31645730.
11. Lawson ARJ, Abascal F, Coorens THH, Hooks Y, O'Neill L, Latimer C, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science. 2020; 370(6512):75–82. https://doi.org/10.1126/science.aba8347 PMID: 33004514
12. Forsberg LA, Rasi C, Razzaghian H, Pakalapati G, Waite L, Stanton Thilbeault K, et al. Age-related somatic structural changes in the nuclear genome of human blood cells. Am J Hum Genet. 2012; 90 (2):217–28. https://doi.org/10.1016/j.ajhg.2011.12.009 PMID: 22305530; PubMed Central PMCID: PMC3276669.
13. Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nature Genetics. 2014; 46(6):624–8. Epub 2014/04/30. https://doi.org/10.1038/ng.2966 [pii]. PMID: 24777449.
14. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med. 2014; 371(26):2488–98. https://doi.org/10.1056/NEJMoa1408617 PMID: 25426837; PubMed Central PMCID: PMC4306669.
15. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018; 362(6417):911–7. Epub 2018/10/20. https://doi.org/10.1126/science.aau3879 PMID: 30337457; PubMed Central PMCID: PMC6298579.
16. Jacobs PA, Brunton M, Court Brown WM, Doll R, Goldstein H. Change of human chromosome count distribution with age: evidence for a sex differences. Nature. 1963; 197:1080–1. Epub 1963/03/16. https://doi.org/10.1038/1971080a0 PMID: 13964326.
17. Pierre RV, Hoagland HC. Age-associated aneuploidy: loss of Y chromosome from human bone marrow cells with aging. Cancer. 1972; 30(4):889–94. Epub 1972/10/01. https://doi.org/10.1002/1097-0142(197210)30:4<889::aid-cncr2820300405>3.0.co;2-1 PMID: 4116908

18. Dumanski JP, Rasi C, Lonn M, Davies H, Ingelsson M, Giedraitis V, et al. Smoking is associated with mosaic loss of chromosome Y. Science. 2015; 347(6217):81–3. Epub 2014/12/06. doi: 1262092 [pii] science.1262092 [pii] https://doi.org/10.1126/science.1262092 PMID: 25477213.

19. Dumanski JP, Lambert JC, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, et al. Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. Am J Hum Genet. 2016; 98(6):1208–19. https://doi.org/10.1016/j.ajhg.2016.05.014 PMID: 27231129; PubMed Central PMCID: PMC4908225.

20. Dumanski J, Halvardson J, Davies H, Rychlicka-Buniowska E, Mattisson J, Torabi Moghadam B, et al. Immune cells lacking Y chromosome show dysregulation of autosomal gene expression. Cell Mol Life Sci. 2021. https://www.biorxiv.org/content/10.1101/673459v2 https://doi.org/10.1007/s00018-021-03822-w PMID: 33837451.

21. Pekar G, Davies H, Lukacs AP, Forsberg L, Hellberg D, Dumanski J, et al. Biobanking multifocal breast carcinomas: sample adequacy with regard to histology and DNA content. Histopathology. 2016; 68 (3):411–21. https://doi.org/10.1111/his.12758 PMID: 26083274.

22. Heer E, Harper A, Escandor N, Sung H, McCormack V, Fidler-Benaoudia MM. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. Lancet Glob Health. 2020; 8(8):e1027–e37. Epub 2020/07/28. https://doi.org/10.1016/S2214-109X(20)30215-1 PMID: 32710860.

23. Yu C, Mitchell JK. Non-randomness of the anatomical distribution of tumors. Cancer Converg. 2017; 1 (1):4. Epub 2017/01/01. https://doi.org/10.1186/s41236-017-0006-7 PMID: 29623957; PubMed Central PMCID: PMC5876694.

24. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–7. Epub 2012/07/20. https://doi.org/10.1038/nature11252 PMID: 22810696; PubMed Central PMCID: PMC3401966.

25. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature medicine. 2015; 21(11):1350–6. Epub 2015/10/13. https://doi.org/10.1038/nm.3967 PMID: 26457759; PubMed Central PMCID: PMC4636487.

26. Margenthaler JA, Dietz JR, Chatterjee A. The Landmark Series:Breast Conservation Trials (including oncoplastic breast surgery). Annals of Surgical Oncology. 2021; 28(4):2120–7. https://doi.org/10.1245/s10434-020-09534-y PMID: 33521897

27. Cook MB, McGlynn KA, Devesa SS, Freedman ND, Anderson WF. Sex disparities in cancer mortality and survival. Cancer Epidemiol Biomark Prev. 2011; 20(8):1629–37. Epub 2011/07/14. https://doi.org/10.1158/1055-9965.EPI-11-0246 PMID: 21750167; PubMed Central PMCID: PMC3153584.

28. Edgren G, Liang L, Adami HO, Chang ET. Enigmatic sex disparities in cancer incidence. Eur J Epidemiol. 2012; 27(3):187–96. Epub 2012/01/04. https://doi.org/10.1007/s10654-011-9647-5 PMID: 22212865.

29. Loftfield E, Zhou W, Yeager M, Chanock SJ, Freedman ND, Machiela MJ. Mosaic Y Loss Is Moderately Associated with Solid Tumor Risk. Cancer Res. 2019; 79(3):461–6. https://doi.org/10.1158/0008-5472.CAN-18-2566 PMID: 30610122; PubMed Central PMCID: PMC6359954.

30. Loftfield E, Zhou W, Graubard BI, Yeager M, Chanock SJ, Freedman ND, et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. Sci Rep. 2018; 8(1):12316. https://doi.org/10.1038/s41598-018-30759-1 PMID: 30120341.

Paper II

npj | breast cancer

www.nature.com/npjbcancer

ARTICLE OPEN

Check for updates

High prevalence of somatic *PIK3CA* and *TP53* pathogenic variants in the normal mammary gland tissue of sporadic breast cancer patients revealed by duplex sequencing

Anna Kostecka[1,2,15], Tomasz Nowikiewicz[3,4,15], Paweł Olszewski[2], Magdalena Koczkowska[1,2], Monika Horbacz[2], Monika Heinzl[5], Maria Andreou[2], Renato Salazar[5], Theresa Mair[5], Piotr Madanecki[1], Magdalena Gucwa[1], Hanna Davies[6], Jarosław Skokowski[7], Patrick G. Buckley[8], Rafał Pęksa[9], Ewa Śrutek[3], Łukasz Szylberg[10,11], Johan Hartman[12,13,14], Michał Jankowski[3], Wojciech Zegarski[3], Irene Tiemann-Boege[5], Jan P. Dumanski[2,6] and Arkadiusz Piotrowski[1,2]

The mammary gland undergoes hormonally stimulated cycles of proliferation, lactation, and involution. We hypothesized that these factors increase the mutational burden in glandular tissue and may explain high cancer incidence rate in the general population, and recurrent disease. Hence, we investigated the DNA sequence variants in the normal mammary gland, tumor, and peripheral blood from 52 reportedly sporadic breast cancer patients. Targeted resequencing of 542 cancer-associated genes revealed subclonal somatic pathogenic variants of: *PIK3CA*, *TP53*, *AKT1*, *MAP3K1*, *CDH1*, *RB1*, *NCOR1*, *MED12*, *CBFB*, *TBX3*, and *TSHR* in the normal mammary gland at considerable allelic frequencies (9×10^{-2}– 5.2×10^{-1}), indicating clonal expansion. Further evaluation of the frequently damaged *PIK3CA* and *TP53* genes by ultra-sensitive duplex sequencing demonstrated a diversified picture of multiple low-level subclonal (in 10^{-2}–10^{-4} alleles) hotspot pathogenic variants. Our results raise a question about the oncogenic potential in non-tumorous mammary gland tissue of breast-conserving surgery patients.

npj Breast Cancer (2022)8:76 ; https://doi.org/10.1038/s41523-022-00443-9

INTRODUCTION

Breast cancer affects 24% of women worldwide and is the leading cause of cancer-related deaths in women[1]. Most breast cancer cases (85–90%) are not associated with inherited mutations of high penetrance genes, such as *BRCA1* (MIM *113705) or *BRCA2* (MIM *600185)[2,3]. High throughput genomics technologies have highlighted the molecular complexity of breast tumors which has led to the molecular classification of four clinically meaningful subtypes: Luminal A, Luminal B, HER2-enriched and basal-like[4,5]. Large cohort studies of breast tumor samples identified somatic driver mutations in key breast cancer-associated genes, such as *PIK3CA* (MIM *171834), *TP53* (MIM *191170), *MAP3K1* (MIM *600982), *CDH1* (MIM *192090), *AKT1* (MIM *164730), *CBFB* (MIM *121360), *TBX3* (MIM *601621), *RB1* (MIM *614041)[6–8]. To date, the identification of somatic driver pathogenic variants has been inferred only from tumors, without providing information on the mutational landscape and allelic frequencies of specific variants in the tissue of cancer origin, i.e., normal tissue of the mammary gland. This is highly relevant as under physiological conditions mammary gland tissue is mitotically stimulated by hormones and undergoes cycles of intense proliferation and remodeling during puberty, pregnancy, and lactation[9]. During life, the mammary gland is exposed to estrogen and its metabolites that damage DNA by single- and double-strand breaks, mutations or, the formation of depurinating adducts[10–12]. These stress conditions can promote the accumulation of post-zygotic, somatic genetic alterations that create the risk of malignant transformation. Indeed, several studies, including ours, have identified such changes in the uninvolved mammary gland of breast cancer patients that is defined as histologically normal glandular tissue, distant from the primary tumor site[13–15]. The most pronounced genetic alterations were identified in the normal tissue from mastectomy patients that per se did not have direct clinical implications, as this affected tissue was removed completely during surgery, but might suggest an increased mutational load in the second breast. At the same time, current clinical management of breast cancer includes breast-conserving surgery (BCS) - removing the tumor and sparing normal breast tissue as one of the recommended treatments[16,17]. The presumed presence of pathogenic genetic alterations in the seemingly normal mammary gland tissue that is not removed during BCS might create a risk of recurrence and can affect future treatment.

Hence, we aimed to screen at unprecedented sensitivity for the presence of subclonal somatic pathogenic genetic alterations in breast cancer-related genes in the normal mammary gland of sporadic cancer patients (study overview in the Supplementary Fig. 1).

Our study demonstrates that structural chromosomal aberrations and clearly pathogenic point variants in crucial breast cancer

[1]Faculty of Pharmacy, Medical University of Gdansk, Gdansk, Poland. [2]3P Medicine Lab, Medical University of Gdansk, Gdansk, Poland. [3]Department of Surgical Oncology, Ludwik Rydygier's Collegium Medicum UMK, Bydgoszcz, Poland. [4]Department of Breast Cancer and Reconstructive Surgery, Prof. F. Lukaszczyk Oncology Center, Bydgoszcz, Poland. [5]Institute of Biophysics, Johannes Kepler University, Linz, Austria. [6]Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [7]Department of Surgical Oncology, Medical University of Gdansk, Gdansk, Poland. [8]Genuity Science Genomics Centre, Dublin, Ireland. [9]Department of Patomorphology, Medical University of Gdansk, Gdansk, Poland. [10]Department of Tumor Pathology, Prof. F. Lukaszczyk Oncology Center, Bydgoszcz, Poland. [11]Department of Perinatology, Gynaecology and Gynaecologic, Oncology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, Bydgoszcz, Poland. [12]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden. [13]Department of Pathology, Karolinska University Hospital, Stockholm, Sweden. [14]MedTech Labs, Bioclinicum, Karolinska University Hospital, Stockholm, Sweden. [15]These authors contributed equally: Anna Kostecka, Tomasz Nowikiewicz. [✉]email: anna.kostecka@gumed.edu.pl; tomasz.nowikiewicz@gmail.com; arkadiusz.piotrowski@gumed.edu.pl

Published in partnership with the Breast Cancer Research Foundation

driver genes are frequent in the normal mammary glandular tissue that remains after breast-conserving surgery.

RESULTS

Patterns of chromosomal aberrations

We carried out analysis of chromosomal rearrangements with SNP arrays to detect DNA copy number alterations (CNAs) as well as copy number neutral loss-of-heterozygosity events via mitotic recombination. In addition to matched samples of normal uninvolved mammary gland (UM) and primary tumor (PT), we included normal mammary gland samples from 26 age-matched women that underwent breast reduction surgery and served as the control group (Supplementary Fig. 2). Spectrum of CNAs in the studied cohort is presented on Fig. 1. Hierarchical clustering revealed two clusters with PT-only and control-only samples and four additional clusters with mixed sample distribution (Supplementary Fig. 3). We also carried out cross analysis of CNAs type, size and number between the studied sample groups. The PTs stand out in this comparison (Wilcoxon test, $p = 0{,}0094$), with slight differences between normal mammary tissue from breast cancer patients and the control cohort. Nonetheless, per individual basis, total number of CNAs, the number of gains, the size of deletions, and size of CNAs in general were the discriminating features between the normal mammary tissue from breast cancer patients and the control cohort, surprisingly suggesting more heterogeneous nature of the control samples (Supplementary Fig. 4).

We identified recurrent chromosomal aberrations in UMs from sporadic breast cancer patients, such as loss of 1p, 16p11.2, and 9p21.3, and 3q25.3, 4q13.1, 8q, and 20q gains, in line with previous studies[5,18]. Presence of loss of heterozygosity (LOH) at chromosome 8p, associated with poor outcome in breast cancer, was observed in matched UMs and PTs, but also in the normal mammary gland tissue of healthy controls[19]. We observed additional events that frequently accompany 8p LOH, in the UMs: 9p loss and 8q gain. *ERBB2* gains were observed exclusively in PT samples, except for one control mammary gland sample.

Subclonal somatic pathogenic variants in breast cancer driver genes present in the normal mammary gland tissue

We applied targeted DNA sequencing to identify variants in sets of UM, BL, and PT samples of 52 individuals diagnosed with sporadic breast cancer to distinguish germline and post-zygotic mutations (Supplementary Table 1, Supplementary Table 2).

Four individuals (4/52, 7.7%) were heterozygous for a constitutional pathogenic variant of a known breast cancer-associated gene, i.e. c.5179 A > T (p.Lys1727Ter) and c.181 T > G (p.Cys61Gly) in the *BRCA1* gene, c.509_510del (p.Arg170fs) and c.354del (p.Thr119fs) in the *PALB2* and *RAD50* genes, respectively (Supplementary Table 3). These results correspond to similar rates from other studies where up to 10% of reportedly sporadic cases turns out hereditary after molecular testing[5,7]. Individuals with germline pathogenic variants were excluded from further analysis, resulting in a total of 48 clearly sporadic breast cancer patients. Constitutional variants of breast cancer-associated genes are listed in the Supplementary Table 3.

The summary of somatic variants fulfilling the cut-off criteria detected in known breast cancer-associated and candidate breast cancer-associated genes is provided in Supplementary Tables 4 and 5, respectively. We identified 15 somatic pathogenic, likely pathogenic variants or variants of uncertain significance with predicted deleterious effect on the encoded protein in the normal mammary gland tissue of 19% (9/48) of patients (Fig. 2). The affected genes are tumor suppressors (*TP53*[5], *RB1*[20], *CDH1*[21]), oncogenes (*PIK3CA*[22]), regulate cell death (*MAP3K1*[23]), DNA repair (*AKT1*[24], *RAD50*[25]), translation (*CBFB*[26]), gene expression (*MED12*[27], *TSHR*[28]) and chromatin remodeling (*NCOR1*[6]). A detailed description of these genes in the context of breast cancer is provided in Supplementary Tables 6, 7 and Supplementary Fig. 8. All of these variants except *PIK3CA* c.3140 A > G (p.His1047Arg) were detected in BCS patients, in samples from the tissue portion that was not qualified for surgical resection.

Heterogeneity of *PIK3CA* and *TP53* pathogenic variants revealed in the normal mammary gland tissue

Two driver genes dominate across all subtypes of invasive breast cancer: *PIK3CA* and *TP53*[5]. *PIK3CA* encodes the catalytically active p100alpha isoform that regulates cell proliferation and growth receptor signaling cascade. Activating *PIK3CA* point variants are the most prevalent in breast tumors and were confirmed to lead to malignant transformation[22,29]. We detected four hotspot *PIK3CA* somatic variants in the uninvolved mammary gland, all of them have been described in the COSMIC database and reported in breast tumors (Fig. 2, Table 2, Supplementary Fig. 5). *TP53* tumor suppressor acts as a transcription factor and is frequently inactivated in human malignancies, mostly through loss-of-function *TP53* variants[30–32]. We detected an Ile195Thr hotspot variant in the uninvolved mammary gland that affects the central DNA-binding domain (Fig. 2, Table 2, Supplementary Fig. 5).

To enhance the sensitivity and accuracy of rare variant detection, we employed duplex sequencing (Supplementary Fig. 7). We selected four individuals: P10, P28, P51, and P52 based on the presence of *PIK3CA* and *TP53* hotspot variants in PT samples according to standard NGS data (Fig. 3) and screened for variants in the normal mammary gland samples with high sensitivity duplex NGS sequencing. Ultra-deep targeted duplex sequencing of *PIK3CA* detected low-level subclonal pathogenic variants: c.1093 G > A (p.Glu365Lys), c.1358 A > G (p.Glu453Gly), c.1633G > A (p.Glu545Lys) c.1634A > C (p.Glu545Ala), c.2164 G > A (p.Glu722Lys), c.3140 A > G (p.His1047Arg), in the uninvolved mammary gland samples of three individuals. The detected variants were located in the known *PIK3CA* hotspot regions, reported in breast tumors in the COSMIC database and functionally confirmed to affect PIK3CA function[7,22] (Fig. 3, Supplementary Table 8). A screen for *TP53* variants not only confirmed the presence of His168Leu variant, but also revealed additional hotspot variants: c.527 G > T (p.Cys176Phe), c.701 A > G (p.Tyr234Cys), c.733 G > A (p.Gly245Ser), c.745 A > T (p.Arg249Trp), c.818 G > A (p.Arg273His), c.839 G > C (p.Arg280Thr). Importantly, all these pathogenic variants are located in the central DNA-binding domain indispensable for p53 tumor-suppressive function[7,32] (Fig. 3, Supplementary Table 8).

DISCUSSION

Post-zygotic variations contribute to the genetic heterogeneity of an individual, which is reflected in a mosaic pattern of genetic alterations in all cells that make up the human body[33]. The mammary gland remains mitotically active during life and under physiological conditions is exposed to DNA-damaging estrogen metabolites[11]. Subclonal somatic genetic changes acquired during life pose a risk of cancer development. Hence, we hypothesized that these factors can increase the mutational burden in the mammary gland. Other studies have reported the presence of genomic and transcriptomic changes in the normal mammary gland, and suggested that histological normalcy does not exclude pathological biological changes[34–36]. However, these studies have been carried out on normal mammary tissue obtained from mastectomies or cancer-adjacent samples, hence the clinical relevance of the these findings was limited. In this study, we screened for somatic genetic changes in the normal mammary gland tissue of sporadic cancer patients, including tissue biopsies from the parts of the breast that normally would not have been removed during breast-conserving surgery. We identified

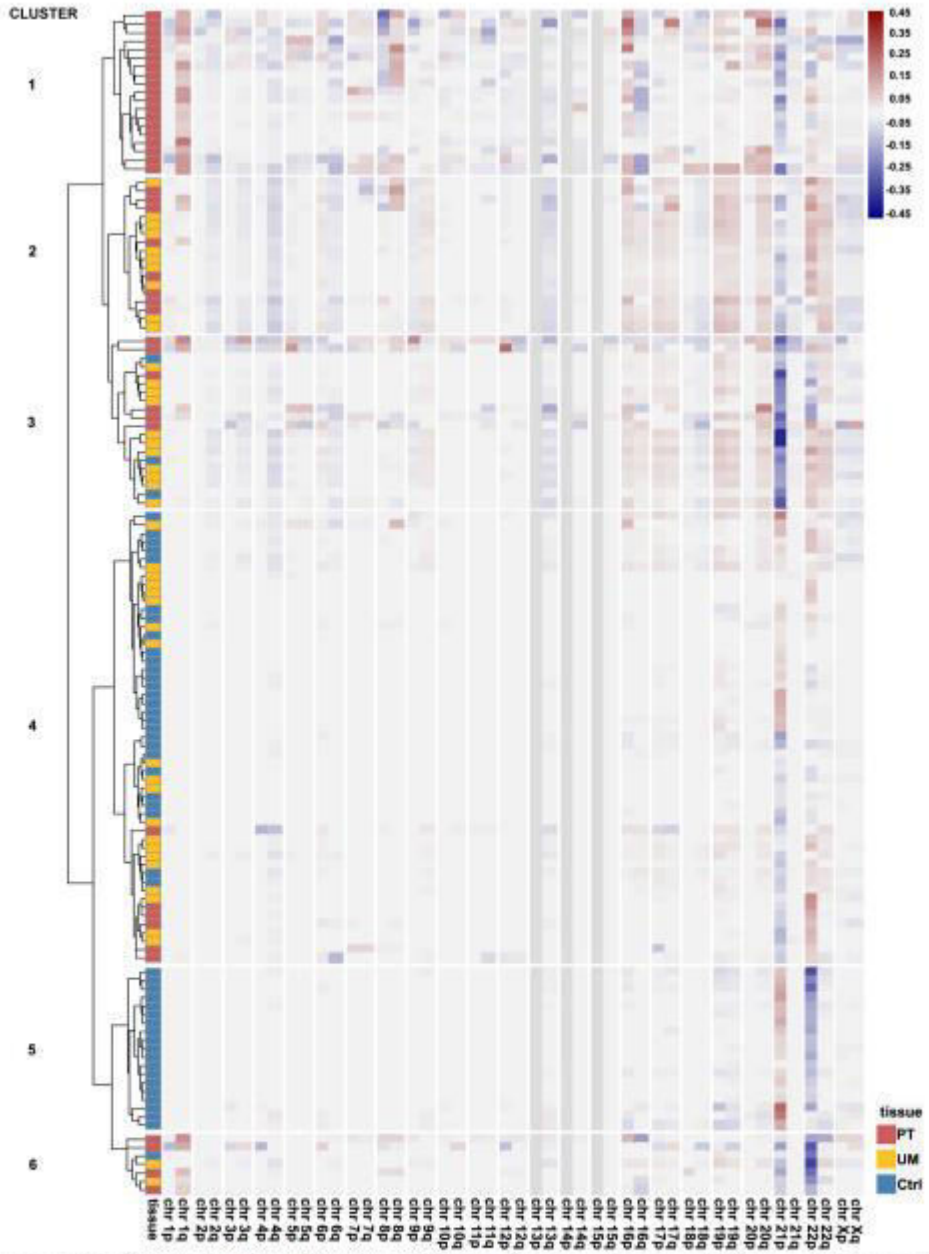

Fig. 1 Summary of Copy Number Alterations (CNAs) detected in the studied cohort. Chromosomal CNAs were calculated as mean Log R Ratio (LRR) for chromosome arm and normalized to mean LRR of a sample. Results are presented as a heatmap with colors indicating gains (positive LRR values; red) and deletions (negative LRR values; blue). Hierarchical clustering was performed with Ward2 algorithm[43] and identified six clusters. Pie charts with proportion of samples within clusters are presented in the Supplementary Fig. 3. Ctrl control cohort mammary gland, UM uninvolved mammary gland, PT tumor.

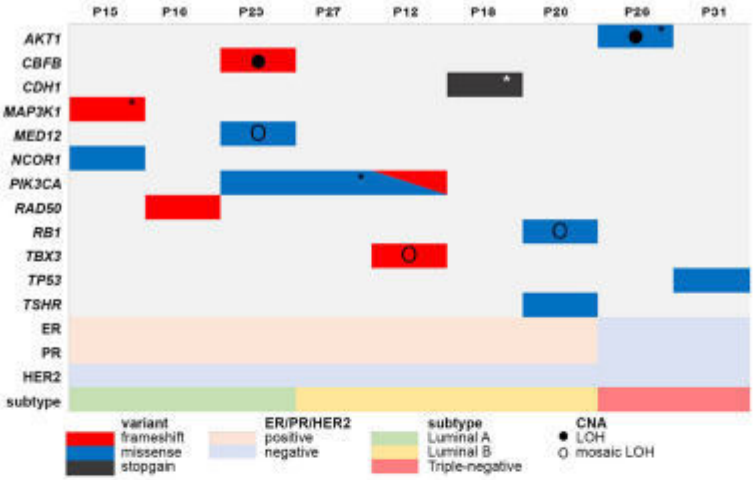

Fig. 2 Somatic variants detected in the uninvolved mammary gland (UM). Targeted sequencing revealed somatic variants of known breast cancer-associated genes (rows) present in 9–52% alleles in the UM of sporadic breast cancer patients (columns). Information on estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and biological subtype of matched primary tumor sample is included. *Variants detected in matched PT sample. CNA Copy Number Alteration status based on SNP arrays. LOH loss of heterozygosity. Description of detected variants, including genomic position and pathogenicity classification is provided in Table 2.

widespread genomic structural rearrangements that affect gene dosage and somatic subclonal sequence variants of known breast cancer-associated genes that control proliferation, cell death, metastasis, and genome integrity: *PIK3CA*, *TP53*, *AKT1*, *MAP3K1*, *CDH1*, *RB1*, *NCOR1*, *MED12*, *CBFB*, *TBX3*, and *TSHR* (Supplementary Fig. 8). These variants were present in a considerable percentage of cells, suggesting they occurred earlier in the mammary gland development or the carrier cells gained growth advantage and underwent clonal expansion. Further, ultra-sensitive duplex sequencing revealed heterogenous mosaic landscape of low-level subclonal pathogenic variants of main breast cancer drivers: *PIK3CA* and *TP53* in the normal mammary gland tissue. Notably, the setup of these variants was markedly different between tumor and normal mammary tissue from the same individuals which is suggestive of multiple, independent mutational events that occurred in the mammary gland (Fig. 4).

In parallel to sequence variants, we identified recurrent CNAs in the mammary gland of breast cancer patients, but also in the age-matched control group (Fig. 1). This facilitated detecting subtle, but noticeable differences in terms of total number and length of all detected CNAs per individual (Supplementary Fig. 4). Both groups: breast cancer and control were age-matched and therefore the mammary gland tissue was exposed to cycles of estrogen for comparable time and that can explain the accumulation of copy number alterations in both cohorts.

The most important finding from this part of our study is that the normal mammary tissue from cancer patients showed DNA copy number alterations as well as evidence of copy number neutral loss-of-heterozygosity. These genomic alterations in concert with damaging sequence variants recapitulate alternative routes of gene inactivation that are typically observed in the malignant tumors, but not in the benign tissue. In this context, our study demonstrates that normal tissue profiling provides direct information on the very origin of the disease and may improve the choice of treatment as well as may aid in further clinical management of the affected individuals[37–39]. This is in contrast to typical molecular profiling studies that rely on limited retrospective information inferred from the tumors.

The *PIK3CA* and *TP53* genes are the leading oncogenic mutations of breast malignancies and accordingly the most common changes detected in our study were in the *PIK3CA* gene[5,40]. Soysal et al. screened for somatic variants in benign biopsies of patients that subsequently developed breast cancer. *PIK3CA* and *TP53* variants were the most prevalent changes in tumor samples, but not detected in benign biopsies, possibly due to limited sensitivity of standard massively parallel sequencing for rare variant detection[41]. To overcome this limitation, we implemented duplex sequencing technology to detect *PIK3CA* and *TP53* variants in the normal mammary gland samples at very low frequency. In the uninvolved mammary gland tissue, we detected known hotspot pathogenic variants that might activate *PIK3CA* kinase or target DNA-binding domain of *TP53* tumor suppressor, disabling its function.

We confirmed that these variants observed in tumor samples were already present in the normal glandular tissue as well, albeit at lower levels compared to the corresponding tumors. Strikingly these changes were accompanied in the same samples by other *PIK3CA* and *TP53* pathogenic variants, present in the normal tissue, but not in the corresponding tumors. This may suggest the existence of potential sites of secondary tumor formation. Notably, the majority of somatic pathogenic variants, including these *PIK3CA* and *TP53* hotspot alterations, occurred in the normal mammary gland samples not removed during breast-conserving surgery, not from radical mastectomy patients.

At the same time *PIK3CA* and *TP53* variant spectra in the normal glandular tissue were more similar to the ones reported in cancer-oriented database (COSMIC) than those in general population (gnomAD), suggesting that the studied UM tissues reflect the repertoire of somatic variants seen in tumor samples (Supplementary Fig. 9, Supplementary Fig. 10, Supplementary Table 9). However, given the limited number of four individuals included in duplex sequencing analysis, these conclusions should be interpreted with caution. Further studies on a larger well-characterized cohort of sporadic breast cancer patients are needed for understanding how specific variants arise and expand during life. Nevertheless, we demonstrate here that ultra-sensitive duplex

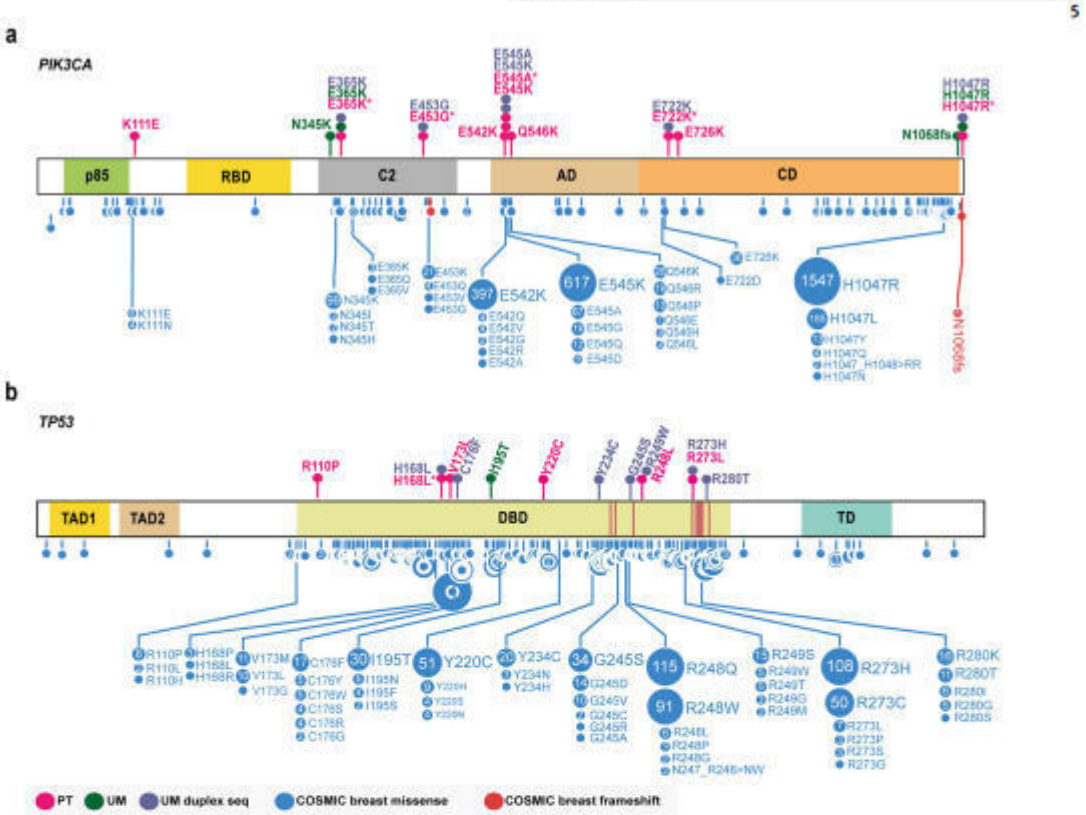

Fig. 3 Somatic *PIK3CA* and *TP53* variants detected in the uninvolved mammary gland (UM) and primary tumor (PT) samples. Lollipop plots represent somatic variants of (**a**) *PIK3CA* and (**b**) *TP53* genes detected by targeted next-generation sequencing (NGS). Upper panel represents variants detected in patient uninvolved mammary gland (UM) and tumor (PT) samples. All somatic variants detected according to the standard NGS and pathogenic/likely pathogenic variants detected by duplex sequencing in UM samples are included. Lower panel is a summary of somatic variants detected in breast tumors reported in the COSMIC database (https://cancer.sanger.ac.uk/cosmic). p85 p85-binding domain, RBD Ras-binding domain, C2 C2 domain, AD accessory domain, CD catalytic domain. TAD1, TAD2 transcription activation domain 1 and 2, DBD DNA-binding domain, DNA-binding sites are marked with red lines, TD tetramerization domain. Lollipop plots were prepared based on the images generated with the Protein paint application[52]. *Variants detected by standard NGS in primary tumor samples and selected for duplex sequencing.

sequencing approach might be beneficial to detect very low-level frequency somatic mosaicism in different tissue samples, with its potential clinical implications in terms of molecular diagnostics and prognosis.

After surgical intervention, breast cancer patients remain under clinical surveillance with recommended yearly mammogram and physical examination every 3–4 months for the first two years after surgery[42]. The current diagnostic approach has been focused mainly on the identification of constitutional pathogenic variants in known breast cancer-associated genes to catch early these individuals who are in a higher risk of breast cancer development and/or to whom the personalized targeted therapy could be offered. However, over 80% of all breast cancer cases are not associated with inherited changes[17].

Our results demonstrate a complex landscape of mutational burden in the seemingly normal mammary glandular tissue and indicate an oncogenic potential of the tissue not removed during surgery. This study provides a rationale for thorough genetic and clinical surveillance of sporadic breast cancer patients that underwent breast-conserving surgery. Including molecular evaluation of the normal glandular tissue of sporadic breast cancer patients could be beneficial for personalized patient care.

METHODS

Patient samples and DNA isolation

We analyzed samples from 52 patients diagnosed with reportedly sporadic breast cancer with an emphasis on breast-conserving surgery (2/3 of the patients studied) and who did not receive neoadjuvant therapy. Altogether a total of 204 uninvolved mammary gland (UM), primary tumor (PT), skin (SK), and peripheral blood (BL) samples were collected via the Oncology Centre in Bydgoszcz and the University Clinical Centre in Gdansk, with the approval of bioethics committee at Medical University of Gdansk (MUG). We have obtained written informed consent from all participants. PT, UM, SK, and BL samples from each patient were collected and stored in −80 °C upon DNA isolation. The overview of sample processing workflow is presented in the Supplementary Fig. 1. The histological subtypes and tumor tissue content of each PT sample were evaluated by pathologists according to the current American Joint Committee on Cancer

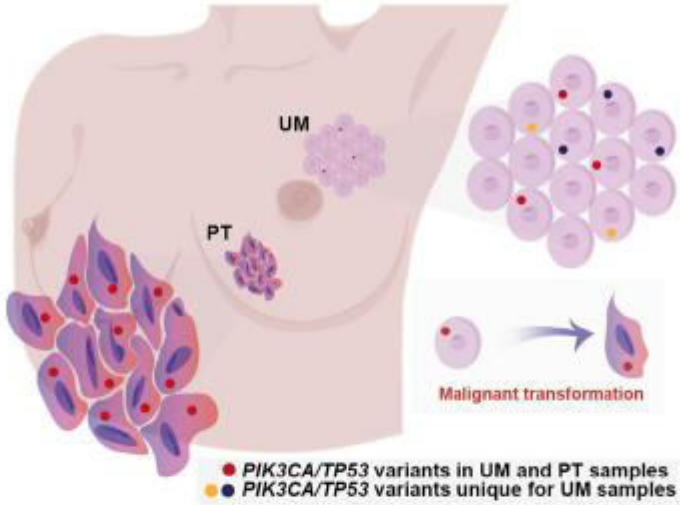

Fig. 4 Oncogenic potential of the normal mammary tissue. We used duplex sequencing to screen for ultra-low frequency variants and detected *PIK3CA* and *TP53* hotspot alterations. The sampled normal mammary gland tissue is referred to as uninvolved glandular tissue and was not removed during surgical resection of the tumor mass. Detected variants might alter the function of the main breast cancer drivers: activate *PIK3CA* oncogene and impair *TP53* tumor suppressor DNA-binding capacity. The presence of these changes implicates an oncogenic potential of the uninvolved mammary gland tissue and emphasizes the importance of thorough monitoring of sporadic breast cancer patients that underwent breast-conserving surgery.

guidelines[43]. Tumor samples with less than 50% of neoplastic cell content were excluded. The normal mammary gland was sampled preferably from the opposite quadrant relative to the primary tumor site, with a mandatory cut-off criterion of at least 3 cm in each case, to exclude potential contamination with residual tumor cells. These tissue samples were also evaluated by pathologists to confirm normal histology (Table 1, Supplementary Table 1). All normal mammary gland samples from patients who underwent breast-conserving surgery were derived from the portion of tissue that remained intact in the patient body after breast-conserving surgery. Solid tissues were homogenized in a lysis buffer, then Proteinase K was added and samples were incubated at 55 °C for 48 h. DNA isolation from UM, PT, and SK tissue lysates was performed by phenol–chloroform extraction as previously described[13]. Blood DNA extraction was performed with the QIAamp DNA Blood Mini Kit according to the manufacturer's protocol (Qiagen, Germantown, MD).

Copy number alteration detection

SNP array genotyping was performed for UM and PT samples on an Illumina Infinium Global Screening Array, according to the manufacturer's recommendations (Illumina, San Diego, CA). SNP genotyping data from mammary gland tissues of 26 age-matched women that underwent breast reduction surgery were used as control samples (Supplementary Fig. 2). Genotyping data was analyzed using Nexus Copy Number software version 10.0 (BioDiscovery). Quality control of samples was performed as described previously[14,44]. Briefly, samples with Log R Ratio (LRR) sd > 0.2 were flagged as poor quality and excluded from the analysis. The analysis was performed with default settings except that significance threshold for Copy Number Alterations (CNA) calling was decreased to $5*10^{-13}$ (default $5*10^{-7}$), minimal number of probes per segment was increased to 10 (default 3), gain threshold was set to 0.49 and 0.14 which corresponds to approximately 40% and 10% change for a high gain and gain respectively (the default is 0.41 and 0.06 for a high gain and gain), the loss threshold was set to −0.16 and −0.74 what corresponds to approximately −10% and −40% change for a loss and high loss respectively (the default is −0.09 and −1.1 for a loss and high loss). Hierarchical clustering was performed using the Ward2 algorithm[45].

Statistical analysis

All statistical analyses were carried out using R version 3.6.2 and package *stats*. Packages *pheatmap* and *ggpubr* were used for plotting. Statistical significance of differences between two groups was tested using the Mann–Whitney U test. Differences were considered significant at a two-sided $p < 0.05$.

Targeted DNA resequencing

Targeted DNA sequencing panel was designed with Roche NimbleDesign online tool (Roche, https://hyperdesign.com/). The panel included exons with +/- 50 kbp flanking regions of 542 genes selected based on in-house database and literature research (Supplementary Table 2). Sequencing libraries were prepared for sets of UM, BL, and PT samples with the capture-based Roche SeqCap EZ system according to the manufacturer's protocol (Roche, Pleasanton, CA), followed by 150 bp paired-end sequencing performed on Illumina NextSeq550 and MiniSeq instruments (Illumina, San Diego, CA). Sequencing read alignment to the human reference genome (hg38) was performed with the Burrows–Wheeler transform aligner (http://bio-bwa.sourceforge.net/)[46]. Platypus v.0.8.1.1 (https://www.rdm.ox.ac.uk/research/lunter-group/lunter-group/) was used for variant calling[47]. Variants with poor mapping quality (<30), variants supported by high-quality bases (≥30) in fewer than five reads, and variants outside the targeted regions were excluded from analysis. Variants were annotated with VarAFT (version 2.17-2) software[48].

For variant selection, only variants with sequencing depth ≥ 30 and tissue allele frequency ≥ 0.07 were included in the analysis. All truncating variants were included. For non-truncating variants, the following criteria were used: variants were filtered by their clinical significance as reported in the ClinVar database (as of June 2021), variants classified as Pathogenic, Likely Pathogenic, Conflicting interpretations of pathogenicity, risk factor, and drug response were included in the study. The remaining non-truncating variants were included based on their frequency in the general population: variants with minor allele frequency (MAF) ≤ 0.001 across all gnomAD populations ("popmax") or not noted in the database were included. For in silico splicing analysis splice prediction algorithms, i.e. SSF, MaxEntScan, and NNSplice, embedded in Alamut Visual software (version 2.14) were used. Variants described in this study were classified according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology recommendations[49]. Based on

Table 1. Summarized clinicopathological features of sporadic breast cancer patient cohort.

Number of individuals	52
Collected samples:	204
UM	52
PT	52
BL	52
SK	48
Age (median/range)	45/ 28–60
Histology	
IDC	44
ILC	4
IDC-ILC	1
other	3
Receptors	
ER (positive/negative)	46/6
PR (positive/negative)	46/6
HER2 (positive/negative)	5/47
Subtype	
Luminal A	22
Luminal B	24
HER2-enriched	2
Triple-negative	4

Uninvolved mammary gland tissue (UM), primary tumor (PT), skin (SK), and peripheral blood (BL) samples were collected from 52 individuals diagnosed with reportedly sporadic breast cancer. Histological evaluation of tumor samples was performed according to the current American Joint Committee on Cancer guidelines[42]. PT samples were classified as Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC), mixed (ICD-ILC) or other. Estrogen (ER), progesterone (PR), and ERBB2 (HER2) receptors were evaluated based on immunostaining or immunostaining and FISH (HER2). Biological subtypes were assigned based on ER/PR/HER2 and Ki67 status. Detailed clinicopathological information is provided in the Supplementary Table 1.

literature[2,7,30,50,51] we selected 155 breast cancer-associated genes that were the primary focus of variant analysis (Supplementary Table 2). Somatic variants presented in Fig. 2 and Table 2 were confirmed by Sanger sequencing or High Resolution Melting analysis (Supplementary Fig. 5). Lollipop plots with variant demonstration were prepared based on images generated with the Protein paint application[52].

Duplex sequencing

UM, PT, BL, and SK samples of four individuals (P10, P28, P51, and P52) were selected for detection of variants by duplex sequencing based on the presence of *PIK3CA* or *TP53* hotspot variants in PT, but not UM tissue, according to standard NGS. The protocols used here are based on the ones described in more detail in Salazar et al.[53].

Random DNA shearing and size selection. DNA was ultrasonicated for 10 min at ≤10 °C using a Bandelin Sonorex Super RK 102 H Ultrasonic bath ending up with a fragment size distribution of, on average, 275 bp. A double-size selection was performed using Sera-Mag Select beads (Cytiva) in order to exclude fragments outside a range of 100-400 bp. The size selection was performed in 50 µl of sonicated DNA (2 µg), 20 µl 10x CutSmart buffer (NEB), 47.6 µl PCR grade water with 0.7 volumes beads. The reaction was mixed by pipetting thoroughly and incubated at room temperature (RT) for 10 min. Tubes were then placed on a magnet for 5 min and 190 µl of supernatant was transferred to a fresh tube. Next, 2.5 volumes of beads in total considering the initial bead solution was added to the solution and mixed by pipetting. The mixture was incubated at RT for 10 min. Tubes were placed on a magnet and supernatant was discarded. The beads were washed twice with 80% ethanol, air dried at room temperature and 23 µl of PCR grade water was added to resuspend by pipetting. After incubating at RT for 5 min, the dissolved beads were allowed to stand at RT for 5 min, placed on a magnet and the clear supernatant containing the size-selected DNA was transferred to a new tube.

End-repair, A-tailing, adapter ligation, and bead purification. Size selected genomic DNA was end-repaired and A-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs) according to the manufacturer's instructions followed by adapter ligation with the NEBNext® Ultra™ II Ligation Module (New England Biolabs) following the manufacturer's instructions. The adapters ligated to the A-tailed DNA were synthesized as previously described (Adapter 2)[53]. The ligation reaction was then purified using 1.2 volumes of Sera-Mag Select beads (Cytiva). A total of 96.5 µl sample was thoroughly mixed with 115.8 µl beads by pipetting and incubated at RT for 10 min. Tubes were placed on a magnet and the supernatant was discarded. The beads were washed twice with 80% ethanol. Next, the beads were dried at room temperature and 23 µl of PCR grade water was added to resuspend by pipetting. After incubating the dissolved beads at RT for 5 min they were placed on a magnet and the clear supernatant containing the purified DNA was transferred to a fresh tube.

Pre-capture amplification. Ligated fragments were amplified with KAPA HiFi HotStart ReadyMix PCR Kit (KAPA Biosystems). Reaction components, primer sequences, and cycling conditions are listed in the Supplementary Table 10. For libraries with input DNA higher than 240 ng, two parallel reactions were prepared and pooled in the end, just before purification. The first step of amplification was 6 or 12 cycles of single primer extensions followed by the addition of the primer NEBNext Universal and a standard PCR amplification of 2 cycles. PCR products were purified with 1.2 volumes Sera-Mag Select beads as described above, followed by two rounds of targeted capture steps to enrich the templates of interest.

Targeted captures and post-capture amplification. Two rounds of targeted captures followed by PCR amplification were performed as described in Salazar et al., with minor modifications on the post-capture amplification (Supplementary Table 10)[53]. The biotinylated probes used to target exonic regions of *TP53*, and *PIK3CA* are detailed on Supplementary Table 10.

Duplex sequencing data analysis

FastQ files were analyzed with Galaxy platform (available on a private server provided by the Medical University of Gdansk) and first processed by the tool *Trim Galore!* to trim Illumina-specific adapter sequences including the barcode and spacer sequence at the 3' end of the raw reads. Next, the reads were analyzed according to a duplex sequencing (DS) specific pipeline that includes an error correction tool[54]. After creating the duplex consensus sequence (DCS), a trimming step of 5 nucleotides from both 5' and 3' end was included. The trimmed consensus sequences were then aligned by *BWA-MEM* and *BamLeftAlignIndels* to the human genome assembly hg38. To avoid false-positive variants that would occur within any partial adapter sequences and barcodes at the 3' end of the consensus sequence and were not removed by the first adapter trimming step, the tool *clipOverlap* from the package BamUtil was applied. Variant calling was then performed by the variant caller *LoFreq*. Finally, the variants (substitutions only) were further inspected and assigned to tiers using the *Variant Analyzer*[55]. Variants with DCS coverage below 500 and variants outside the probe regions were discarded from our analysis and only Tier 1 variants were kept, together with Tier 2 that were detected more than once. For more details on this analysis see Povysil et al.[55]. The full Galaxy workflow is publicly available: https://usegalaxy.org/u/jku-itb-lab/w/gdansk-paper---galaxy-workflow.

The variant frequency was calculated by dividing the number of DCS calling the variant by the DCS coverage at the position of the variant within the library it was detected. The variant frequency was calculated by the count for each alteration type (e.g. A > C) divided by the frequency of the sequenced reference allele (e.g., frequency of A's in the reference sequence multiplied by the sum of the mean DCS coverage for that library). The relative count is the count for each variant type divided by the sum of all occurring variants within the tissue.

Table 2. Pathogenicity classification of somatic variants detected in the uninvolved mammary gland (UM) samples.

ID	Gene	Genomic position[a]	cDNA change (protein change)[b]	ACMG classification[c]	rsID[d]	ClinVar[e]	UM allele frequency[f]	PT allele frequency[f]
P26	*AKT1*	chr14:104780214	c.49 G > A (p.Glu17Lys)	Pathogenic	rs121434592	–	0,11	0,36
P23	*CBFB*	chr16:67036674	c.207dup (p.Pro70fs)	Pathogenic	–	–	0,15	not detected
P18	*CDH1*	chr16:68819382	c.1668_1669insT (p.Lys557Ter)	Pathogenic	–	–	0,1	0,17
P15	*MAP3K1*	chr5:56881868	c.2668del (p.Asn891fs)	Pathogenic	–	–	0,09	0,15
P23	*MED12*	chrX:71137882	c.5983 C > T (p.Pro1995Ser)	Likely pathogenic	–	–	0,15	not detected
P15	*NCOR1*	chr17:16040459	c.6715 C > A (p.Pro2239Thr)	Likely pathogenic	–	–	0,11	not detected
P12	*PIK3CA*	chr3:179234358	c.3203dup (p.Asn1068fs)	Pathogenic	rs587776802	Pathogenic	0,19	no data[g]
P12	*PIK3CA*	chr3:179204536	c.1093 G > A (p.Glu365Lys)	Pathogenic	rs1064793732	Pathogenic	0,33	no data[g]
P23	*PIK3CA*	chr3:179203765	c.1035 T > A (p.Asn345Lys)	Pathogenic	rs121913284	Likely pathogenic	0,11	not detected
P27	*PIK3CA*	chr3:179234297	c.3140 A > G (p.His1047Arg)	Pathogenic	rs121913279	Pathogenic	0,11	0,11
P16	*RALDS0*	chr5:132595759	c.2165dup (p.Glu723fs)	Pathogenic	rs397507178	Pathogenic	0,16	not detected
P20	*RB1*	chr13:48345117	c.418 A > G (p.Thr140Ala)	Likely pathogenic	–	–	0,11	not detected
P12	*TBX3*	chr12:114679572	c.796_797dup (p.Ser266fs)	Pathogenic	–	–	0,18	not detected
P31	*TP53*	chr17:7674947	c.584 T > C (p.Ile195Thr)	Pathogenic	rs760043106	Likely pathogenic	0,52	no data[g]
P20	*TSHR*	chr14:81068264	c.253 A > G (p.Ile85Val)	VUS	–	–	0,13	not detected

Targeted DNA sequencing identified somatic DNA variants of known breast cancer-associated genes in the uninvolved mammary gland tissue of sporadic breast cancer patients.
[a] Genomic position according to the hg38 sequence assembly.
[b] Variant annotation provided for the basic isoform of the transcript.
[c] Pathogenicity classification according to the current ACMG guidelines[40].
[d] rsIDs in dbSNP build 152.
[e] Variant pathogenicity classification according to the ClinVar database. Detailed description of somatic variants detected in UM samples is provided in the Supplementary Table 4.
[f] Tissue allele frequency of the detected variants in matched UM and PT tissue specimens.
[g] PT sample was not available.
Confirmation of somatic variants by Sanger sequencing or high-resolution melting is provided in the Supplementary Fig. 5. VUS Variant of Unknown Significance.

DATA AVAILABILITY

Raw microarray, NGS and duplex sequencing data are available upon request in the EGA archive, study ID EGAS00001005698.

Received: 29 September 2021; Accepted: 10 June 2022;
Published online: 29 June 2022

REFERENCES

1. Heer, E. et al. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *Lancet Glob. Heal* **8**, e1027–e1037 (2020).
2. Coughlin, S. S. Epidemiology of breast cancer in women. *Adv. Exp. Med. Biol.* **1152**, 9–29 (2019).
3. Kleibl, Z. & Kristensen, V. N. Women at high risk of breast cancer: molecular characteristics, clinical presentation and management. *Breast* **28**, 136–144 (2016).
4. Sorlie, T. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS* **98**, 10869–10874 (2001).
5. Koboldt, D. C. et al. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
6. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
7. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
8. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
9. Macias, H. & Hinck, L. Mammary gland development. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 533–557 (2012).
10. Dall, G. V. & Britt, K. L. Estrogen effects on the mammary gland in early and late life and breast cancer risk. *Front. Oncol.* **7**, 1–10 (2017).
11. Almeida, M., Soares, M., Fonseca-Moutinho, J., Ramalhinho, A. C. & Breitenfeld, L. Influence of estrogenic metabolic pathway genes polymorphisms on post-menopausal breast cancer risk. *Pharmaceuticals* **14**, 1–9 (2021).
12. Yager, J. D. & Davidson, N. E. Estrogen carcinogenesis in breast cancer. *N. Engl. J. Med.* **354**, 270–282 (2006).
13. Ronowicz, A. et al. Concurrent DNA copy-number alterations and mutations in genes related to maintenance of genome stability in uninvolved mammary glandular tissue from breast cancer patients. *Hum. Mutat.* **36**, 1088–1099 (2015).
14. Forsberg, L. A. et al. Signatures of post-zygotic structural genetic aberrations in the cells of histologically normal breast tissue that can predispose to sporadic breast cancer. *Genome Res.* **25**, 1521–1535 (2015).
15. Danforth, D. N. Genomic changes in normal breast tissue in women at normal risk or at high risk for breast cancer. *Breast Cancer Basic Clin. Res.* **10**, 109–146 (2016).
16. Waks, A. G. & Winer, E. P. Breast cancer treatment: a review. *JAMA -J. Am. Med. Assoc.* **321**, 288–300 (2019).
17. Loibl, S., Poortmans, P., Morrow, M., Denkert, C. & Curigliano, G. Breast cancer. *Lancet* **397**, 1750–1769 (2021).
18. Parris, T. Z. et al. Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma. *Clin. Cancer Res.* **16**, 3860–3874 (2010).
19. Cai, Y. et al. Loss of chromosome 8p governs tumor progression and drug response by altering lipid metabolism. *Cancer Cell* **29**, 751–766 (2016).
20. Witkiewicz, A. K. & Knudsen, E. S. Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine, and therapeutic interventions. *Breast Cancer Res.* **16**, 207 (2014).
21. Christgen, M. et al. Lobular breast cancer: clinical, molecular and morphological characteristics. *Pathol. Res. Pract.* **212**, 583–597 (2016).
22. Martínez-Saéz, O. et al. Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Res.* **22**, 1–9 (2020).
23. Pham, T. T., Angus, S. P. & Johnson, G. L. MAP3K1: Genomic alterations in cancer and function in promoting cell survival or apoptosis. *Genes Cancer* **4**, 419–426 (2013).
24. Plo, I. et al. AKT1 inhibits homologous recombination by inducing cytoplasmic retention of BRCA1 and RAD5. *Cancer Res.* **68**, 9404–9412 (2008).
25. Fagan-Solis, K. D. et al. A P53-independent DNA damage response suppresses oncogenic proliferation and genome instability. *Cell Rep.* **30**, 1385–1399.e7 (2020).
26. Malik, N. et al. The transcription factor CBFB suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* **10**, 1–15 (2019).
27. Chang, H. Y. et al. MED12, TERT and RARA in fibroepithelial tumours of the breast. *J. Clin. Pathol.* **73**, 51–56 (2020).
28. Liu, Y. C., Yeh, C. T. & Lin, K. H. Molecular functions of thyroid hormone signaling in regulation of cancer progression and anti-apoptosis. *Int. J. Mol. Sci.* **20**, 1–27 (2019).
29. Thorpe, L. M., Yuzugullu, H. & Zhao, J. J. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer* **15**, 7–24 (2015).
30. Vogelstein, B. et al. Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).
31. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
32. Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death Differ.* **25**, 154–160 (2018).
33. Mustjoki, S. & Young, N. S. Somatic mutations in "benign" disease. *N. Engl. J. Med.* **384**, 2039–2052 (2021).
34. Gadaleta, E. et al. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. *npj Breast Cancer* **6**, 38 (2020).
35. Aran, D. et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* **8**, 1–13 (2017).
36. Troester, M. A. et al. DNA defects, epigenetics, and gene expression in cancer-adjacent breast: a study from the cancer genome atlas. *npj Breast Cancer* **2**, 16007 (2016).
37. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
38. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
39. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
40. Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705.e9 (2018).
41. Soysal, S. D. et al. Genetic alterations in benign breast biopsies of subsequent breast cancer patients. *Front. Med.* **6**, 1–6 (2019).
42. Gradishar, W. J. et al. NCCN clinical practice guidelines in Oncology. Breast Cancer Version 4. 2021. *Natl. Compr. Cancer Netw.* **16**, 310–320 (2021).
43. Amin, M. B., et al. *AJCC Cancer Staging Manual* (Springer International Publishing, 2017).
44. Rydzanicz, M. et al. Variable degree of mosaicism for tetrasomy 18p in phenotypically discordant monozygotic twins—diagnostic implications. *Mol. Genet. Genom. Med.* **9**, 1–9 (2021).
45. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
47. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
48. Desvignes, J. P. et al. VarAFT: A variant annotation and filtration system for human next generation sequencing data. *Nucleic Acids Res* **46**, W545–W553 (2018).
49. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
50. Polyak, K. & Metzger Filho, O. SnapShot: breast cancer. *Cancer Cell* **22**, 562–562.e1 (2012).
51. Mahdavi, M. et al. Hereditary breast cancer; genetic penetrance and current status with BRCA. *J. Cell. Physiol.* **234**, 5741–5750 (2019).
52. Zhou, X. et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48**, 4–6 (2015).
53. Salazar, R. et al. Discovery of an unusually high number of de novo mutations in sperm of older men using duplex sequencing. *Genome Res.* **32**, 499–511 (2022).
54. Stoler, N. et al. Family reunion via error correction: an efficient analysis of duplex sequencing data. *BMC Bioinform.* **21**, 96 (2020).
55. Povysil, G. et al. Increased yields of duplex sequencing data by a series of quality control tools. *NAR Genom Bioinform.* **3**, lqab002 (2021).

ACKNOWLEDGEMENTS

This work was supported by the National Science Center, Poland grant (award no. UMO-2015/19/B/NZ2/03216) to A.P. and partially funded by the Foundation for Polish Science (FNP) under the International Research Agendas Program (grant number MAB/2018/6) to J.P.D. and A.P., co-financed by the European Union under the European Regional Development Fund.

AUTHOR CONTRIBUTIONS

Study design and conception: A.P., I.T.-B., A.K. Sample collection and preparation: T.N., M.G., H.D., J.S., E.Ś., R.P., M.J., Ł.S., W.Z., J.H. Experiments: A.K., M.H., R.S., M.A., T.M. Data analysis and interpretation: A.K., P.O., A.P., M.K., M.H., I.T.-B. Manuscript writing: A.K., A.P.,

M.K., I.T.-B., J.P.D. All authors have read and approved the manuscript. A.K. and T.N. contributed equally.

COMPETING INTERESTS

The authors declare no competing financial interests, but the following competing non-financial interests have been declared: J.P.D. is cofounder and shareholder in Cray Innovation AB.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41523-022-00443-9.

Correspondence and requests for materials should be addressed to Anna Kostecka, Tomasz Nowikiewicz or Arkadiusz Piotrowski.

Paper III

Received: 19 October 2023 | Revised: 15 April 2024 | Accepted: 22 April 2024
DOI: 10.1002/ijc.35050

RESEARCH ARTICLE
Cancer Therapy and Prevention

IJC INTERNATIONAL JOURNAL of CANCER | UICC

Prelude to malignancy: A gene expression signature in normal mammary gland from breast cancer patients suggests pre-tumorous alterations and is associated with adverse outcomes

Maria Andreou[1] | Marcin Jąkalski[1] | Katarzyna Duzowska[1] | Natalia Filipowicz[1] | Anna Kostecka[1] | Hanna Davies[2] | Monika Horbacz[1] | Urszula Ławrynowicz[1] | Katarzyna Chojnowska[1] | Bożena Bruhn-Olszewska[2] | Jerzy Jankau[3] | Ewa Śrutek[4,5] | Manuela Las-Jankowska[6,7] | Dariusz Bała[6,8] | Jacek Hoffman[9] | Johan Hartman[10,11,12] | Rafał Pęksa[13] | Jarosław Skokowski[14] | Michał Jankowski[6,8] | Łukasz Szylberg[5,15] | Mateusz Maniewski[5] | Wojciech Zegarski[6,8] | Magdalena Nowikiewicz[16] | Tomasz Nowikiewicz[6,9] | Jan P. Dumanski[1,2,17] | Jakub Mieczkowski[1] | Arkadiusz Piotrowski[1,17]

Correspondence
Arkadiusz Piotrowski, 3P-Medicine Laboratory, Medical University of Gdańsk, Gdańsk, Poland; Department of Biology and Pharmaceutical Botany, Medical University of Gdańsk, Gdańsk, Poland.
Email: arkadiusz.piotrowski@gumed.edu.pl

Funding information
Swedish Medical Research Council, Grant/Award Number: 2020-02010; Cancerfonden, Grant/Award Number: 200889PjF; Fundacja na rzecz Nauki Polskiej, Grant/Award Number: MAB/2018/6

Abstract

Despite advances in early detection and treatment strategies, breast cancer recurrence and mortality remain a significant health issue. Recent insights suggest the prognostic potential of microscopically healthy mammary gland, in the vicinity of the breast lesion. Nonetheless, a comprehensive understanding of the gene expression profiles in these tissues and their relationship to patient outcomes remain missing. Furthermore, the increasing trend towards breast-conserving surgery may inadvertently lead to the retention of existing cancer-predisposing mutations within the normal mammary gland. This study assessed the transcriptomic profiles of 242 samples from 83 breast cancer patients with unfavorable outcomes, including paired uninvolved mammary gland samples collected at varying distances from primary lesions. As a reference, control samples from 53 mammoplasty individuals without cancer history were studied. A custom panel of 634 genes linked to breast cancer progression and metastasis was employed for expression profiling, followed by whole-transcriptome verification experiments and statistical analyses to discern molecular signatures and their clinical relevance. A distinct gene expression signature was

Jan P. Dumanski, Jakub Mieczkowski, and Arkadiusz Piotrowski contributed equally to senior authorship.

Maria Andreou and Marcin Jąkalski have contributed equally to this study.

For affiliations refer to page 1626

identified in uninvolved mammary gland samples, featuring key cellular components encoding keratins, CDH1, CDH3, EPCAM cell adhesion proteins, matrix metallopeptidases, oncogenes, tumor suppressors, along with crucial genes (*FOXA1*, *RAB25*, *NRG1*, *SPDEF*, *TRIM29*, and *GABRP*) having dual roles in cancer. Enrichment analyses revealed disruptions in epithelial integrity, cell adhesion, and estrogen signaling. This signature, named KAOS for Keratin-Adhesion-Oncogenes-Suppressors, was significantly associated with reduced tumor size but increased mortality rates. Integrating molecular assessment of non-malignant mammary tissue into disease management could enhance survival prediction and facilitate personalized patient care.

KEYWORDS
breast cancer, gene signature, mortality, unfavorable outcome, uninvolved margin

What's New?

Patients who undergo breast-conserving surgery for breast cancer have a significant risk of recurrence if the healthy tissue left behind harbors cancer-predisposing alterations. Here, the authors use a custom-made gene panel, to potentially identify the risk of recurrence by detecting such alterations. The authors investigated the gene expression in tumor samples, paired non-cancerous tissue samples, and healthy mammary tissue samples from control individuals. The gene expression signature that emerged was associated with disruptions in estrogen signaling, cell adhesion, and epithelial integrity, as well as increased mortality.

1 | INTRODUCTION

Breast cancer remains a pervasive global health concern and represents the most prevalent malignancy worldwide, surpassing lung cancer with 2.26 million reported incidents in 2020.[1,2] Improved mammographic screening and widespread educational initiatives, resulting in increased self-monitoring, have facilitated the early detection of breast carcinomas in asymptomatic stages. Consequently, breast-conserving surgery (BCS) has become increasingly prominent as a favored treatment approach, involving the removal of the tumor, while minimizing the removal of healthy tissue, thereby preserving a substantial portion of the breast.[3,4] However, despite histopathologically negative surgical margins, suggesting a complete tumor excision during BCS, a considerable proportion of patients experience recurrence rates as high as 19.3% in those receiving radiotherapy and 35% in patients subjected to BCS alone,[5] while administration of adjuvant chemotherapy reportedly reduces the recurrence rate by one-third 10 years post-surgery.[6] The underlying cause of recurrence, whether it is due to undetected residual disease or the development of additional changes within the unexcised mammary gland remains unclear.

Currently, decisions regarding therapeutic management primarily rely on pathological examination and genetic tests performed solely on fragments originating from tumors as well as resection margins (mammary gland tissue in the tumor perimeter excised during surgery). However, emerging evidence suggests that the normal mammary gland tissue surrounding the cancerous lesion holds the promise of prognostic value.[7–10] Notably, the inclusion of normal, cancer-adjacent tissue samples in study designs significantly enhances the accuracy of overall survival predictions compared to relying solely on tumor data.[11] While previous studies have investigated paired normal and cancerous tissue samples,[12–15] the association of transcriptomic landscape of normal, uninvolved mammary gland tissue, located at a greater distance from the primary tumor and remaining in the patient's body following BCS, with unfavorable patient outcomes has not been thoroughly investigated. Furthermore, the limited availability of an adequate number of control samples in study designs posed challenges in interpreting findings. Taking these factors into account, our study aimed to investigate a unique cohort of breast cancer patients characterized by adverse prognoses, with comprehensive follow-up data that extended to nearly a decade after their initial surgeries. We employed targeted RNA sequencing, to analyze the transcriptomic profiles of primary tumor and paired proximal and distal uninvolved mammary gland samples, as well as mammary glandular tissue samples from control individuals without any personal and familial history of cancer.

Our custom RNA-seq panel effectively discriminates between malignant (primary tumor) and non-malignant (uninvolved margin and control) tissues, by elucidating unique gene expression patterns for each group. Notably, our study uncovers the existence of a pre-tumorigenic microenvironment within the seemingly normal mammary gland tissue, displaying an association with smaller tumor size and higher patient mortality.

2 | MATERIALS AND METHODS

2.1 | Patient recruitment, sample collection, and RNA isolation

We analyzed specimens obtained from 83 individuals who had been diagnosed with breast cancer including 68 who underwent breast-conserving surgery and 13 with mastectomies (data missing for 2 patients). All recruited individuals did not receive neoadjuvant therapy and were characterized by the presence of recurrent disease (metastasis to the breast or secondary organs) and/or the appearance of a second independent tumor and/or death in the following 10 years (Table 1; Additional File S1, Table S1). For two individuals, two distinct samples from multifocal primary tumor (described as PT1 and PT2) were obtained. Fifty-three individuals subjected to breast reduction surgery without personal and familial history of cancer were recruited as controls (Table 1; Additional File S1, Table S2). A graphical representation of the project workflow can be found in Figure 1A. A total of 295 samples, including primary tumor (PT), uninvolved mammary gland distal from (UMD, 1.5–5 cm), and proximal to the primary tumor (UMP, at least 1 cm from the primary tumor and always in shorter distance than UMD), as well as normal mammary gland from control individuals (CTRL), were collected in the Oncology Centre in Bydgoszcz, the University Clinical Centre in Gdańsk, and Karolinska Institute and deposited, along with clinical data and follow-up information, in biobank of our unit at the Medical University of Gdansk.[16] PT, UM, and CTRL samples collected were frozen at −80°C. Detailed sampling design is presented in Figure 1B. All fragments prepared for molecular analysis were microscopically evaluated to identify tumor fragments and confirm normal histology of uninvolved margins and controls. Based on histopathological evaluation samples with >10%–15% immune infiltration were excluded from the study. RNA was extracted from tissues using the RNeasy Mini according to the original protocol with two modifications: (a) 1-bromo-3-chloropropane was used instead of chloroform to prevent foaming and emulsification and (b) the elution was carried out with 90 µL of water for PT and 30 µL of water for UM and CTRL samples, followed by repeated elution with the entire volume of the original eluate (Qiagen, Germantown, MD). RNA concentration and quality were determined using Agilent TapeStation (Agilent Technologies).

2.2 | Targeted RNA sequencing

The targeted RNA sequencing panel, designed with the Roche NimbleDesign online tool (Roche, now HyperDesign, https://hyperdesign.com/#/), covered 7229 regions with a total length of 1,243,523 bp. The panel includes 634 genes selected from literature research (Additional File S1, Table S3). The genes have been associated with breast cancer and processes related to its dissemination and metastasis, such as epithelial-to-mesenchymal transition, cell death, and apoptosis. Furthermore, the panel incorporated genes from the AIMS and PAM50 predictors, originally developed to classify breast tumors into five distinct subtypes: Luminal A, Luminal B, HER2 enriched, basal-like, and normal-like.[17,18] Sequencing libraries were prepared using the KAPA RNA HyperPrep kit (Illumina Platforms, KR1350-v2.16) and the BRAVO NGS workstation (Agilent) with the dedicated automatization protocol (KAPA RNA Hyperprep kit KR1350-v.1.16). Hybridization was carried out with SeqCap RNA Choice Probes using the KAPA

TABLE 1 Summarized clinicopathological characteristics of breast cancer patient and control cohort.

Number of individuals	136
Breast cancer (BC) patients	83
Controls (CTRL)	53
Age (median/range)	
BC patients	62 (23–85)
CTRL	44 (18–76)
	p value = 1.324e−09
Collected samples (BC patients)	242
Primary tumor, PT	79
Uninvolved margin distal from PT, UMD	81
Uninvolved margin proximal to PT, UMP	82
Histology (BC patients)	
Invasive ductal carcinoma, IDC	63
Invasive lobular carcinoma, ILC	4
IDC–ILC	7
Other	9
Receptors (BC patients)	
Estrogen, ER (positive/negative)	62/21
Progesterone, PR (positive/negative)	48/35
HER2 (positive/negative)	17/61
Subtype (BC patients)	
Luminal A	16
Luminal B	40
HER-2 enriched	9
Triple-negative	12
Not available	6
Follow-up information (BC patients)	
Recurrence (yes/no)	50/33
Second cancer (yes/no)	31/52
Death (yes/no)	50/33

Note: PT, UMD, and UMP samples were collected from 83 individuals diagnosed with breast cancer. CTRL samples were collected from 53 individuals without any personal and familial history of cancer. Histological evaluation was performed to identify tumor samples and confirm the normal histology of uninvolved margin and control samples. PT samples were classified as Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC), mixed (ICD-ILC), or other. Estrogen (ER), progesterone (PR), and ERBB2 (HER2) receptors were evaluated based on immunostaining. Biological subtypes were assigned based on ER/PR/HER2 and Ki67 status. Recurrent disease was reported for 50 patients, the presence of a second, independent tumor was confirmed for 31 patients, and 50 patients died by the time of last contact. Detailed clinicopathological information for breast cancer patients is provided in Additional File S1, Table S1.

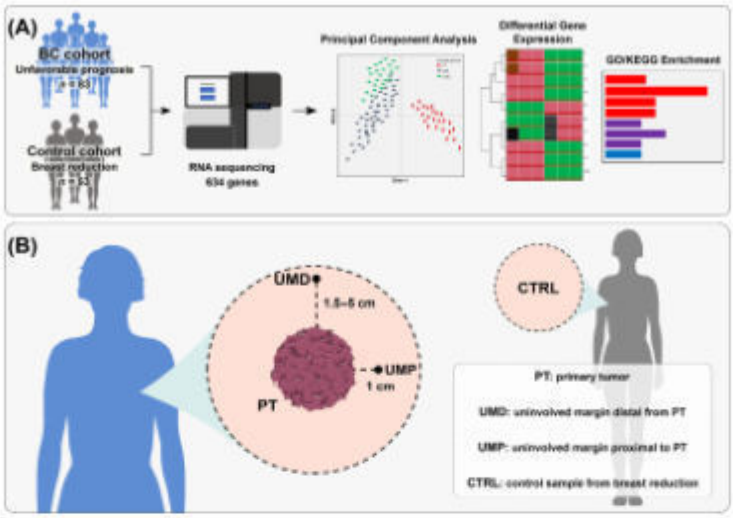

FIGURE 1 (A) Graphical representation of the project workflow. A total of 295 fresh-frozen primary tumor (PT), uninvolved mammary gland excised proximal (UMP, >1 cm) and distal (UMP, 1.5–5 cm) from the PT, and control samples (CTRL) were collected from 83 individuals diagnosed with breast cancer and 53 individuals subjected to breast reduction surgeries, respectively. After RNA extraction and library construction, targeted RNA sequencing was performed using a customized panel including genes previously associated with breast cancer. Bioinformatics analysis implementing standard tools was used to investigate expression patterns in PT, UM, and CTRL samples, as well as associations with follow-up clinical information. (B) Detailed sampling design. Tissue blocks were collected from PT, UMP, and UMD samples from breast cancer patients with unfavorable outcomes. UMP was always collected at a smaller distance than UMD from the PT. Tissue samples were evaluated by two independent pathologists to confirm the normal histology of uninvolved margin (UM) and control (CTRL) samples and identify tumor areas in breast cancer cases. Control (CTRL) mammary gland samples were collected from individuals subjected to breast reduction surgeries without personal and familial history of cancer. Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

HyperCapture Reagent and Bead kit (v2, Roche Sequencing Solutions, Inc.) according to SeqCap RNA Enrichment System User's Guide (v.1.0) with slight modifications. Component A was replaced with formaldehyde, and the Multiplex Hybridization Enhancing Oligo Pool was replaced with Universal Blocking Oligos (UBO). Next, the cDNA libraries were quantified using the KAPA Library Quantification kit (KR0405-v11.20, Kapa Biosystems, Woburn, MA). Paired-end reads of 150 bp were generated using TruSeq RNA Access sequencing chemistry on HiSeq X instrument (Illumina, San Diego, CA) by an external service provider (Macrogen Europe, Amsterdam, The Netherlands). The sequencing coverage and quality statistics for each sample are summarized in Additional File S1, Table S4.

2.3 | Data analysis

2.3.1 | Processing of sequencing data from the targeted panel

The raw RNA-seq data were first subjected to quality check using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, version 0.11.9), followed by adapter trimming with BBDuk from the BBTools package (https://sourceforge.net/projects/bbmap/, version 38.36) with the following set of parameters: ktrim = r, k = 23, mink = 11, hdist = 1, minlen = 70, tpe, tbo. The processed reads were then mapped against the reference human genome (hg38, GENCODE version 35) using STAR (version 2.7.3a).[19] The generated ReadsPerGene.out.tab files (obtained through -quantMode GeneCounts parameter) were used to extract raw read counts that mapped to the annotated genes. Subsequently, the merged raw read matrix was generated with a custom R script and further processed with the edgeR (version 3.36.0).[20] First, the genes were filtered to keep only those whose expression was at least 1 count per million (CPM) in at least one sample. Next, the processed gene expression matrix was normalized using the TMM method in edgeR.[21]

Principal Component Analyses (PCA) were performed to identify outlier samples using the R package FactoMineR (version 2.4).[22] This was carried out in several rounds: for all sample types separately (controls, tumor samples, etc.) and all samples merged.

2.3.2 | Cancer subtype prediction

PAM50 and AIMS classifiers were applied to normalized gene expression matrices using the R package genefu (version 2.26).[17,18,23] The

samples were classified into one of the following categories: Normal, LumA, LumB, Her2, or Basal.

2.3.3 | Sample clustering and differential expression analyses

Gene expression heatmaps were generated with the pheatmap R library (version 1.0.12). Sample clusters were identified using the built-in hierarchical clustering function of pheatmap (default parameters), which uses the Euclidean distance as the similarity measure and the complete linkage method. The number of sample clusters was set to four.

EdgeR was used to identify differentially expressed genes (DEGs) using the Quasi-Likelihood F-test (QLF) with a significance threshold set to 0.05 (False Discovery Rate, FDR) (19). Differential expression (DE) analyses were performed in two modes: first, to perform pairwise sample type comparisons; second, to additionally include the information on cluster membership of the samples. We included age as a covariate in the DE analyses due to the significant age difference between controls (44, 18–76) and cancer patients (62, 23–85) (Wilcoxon test). The identified sets of differentially expressed genes were subjected to enrichment analyses of Gene Ontology (GO) terms and KEGG pathways using ClusterProfiler.[24]

2.3.4 | External RNA-seq datasets

External bulk RNA-seq datasets were used to corroborate our findings. These datasets originated from other breast cancer studies in our lab. The sample collection and processing methods were consistent with those described in this study. Their FASTQ files were processed from scratch using the same tools as described above. Data integration was done with the ComBat_seq function from the sva R library (version 3.42.0, default parameters) to adjust for the effect of different data batches.[25,26]

3 | RESULTS

3.1 | Clear delineation between malignant and non-malignant tissues

Principal component analysis (PCA) of all samples, using the normalized expression profiles of panel genes, revealed distinct differences between malignant (PT) and non-malignant (UM + CTRL) samples, identified by the first principal component (Figure 2A). Fourteen outliers were identified and excluded from downstream analyses, leaving 295 samples (53 controls, 163 margins, and 79 tumor samples). DE analysis, which used non-malignant samples as a baseline, showed the largest set of deregulated genes (FDR ≤0.05, log-fold change of ≥1) when comparing PT against all non-malignant tissues (CTRLs, UMs) (Figure 2B). The number of differentially expressed genes decreased when comparing PT tissues with CTRLs or UMs separately. Interestingly, a relatively small number of differentially expressed genes was found between UMP and UMD, suggesting similar expression profiles and minor effects regardless of their physical distance from the primary tumor.

Examination of the functional annotation of differentially expressed genes (DEGs) identified, as expected, enrichment of gene ontology terms and pathways previously associated with cancer (Figure S1). The observed sets of enriched biological terms and pathways among the primary tumor profiles, such as those related to proliferation and cell cycle, reflected the aggressiveness of those cancers and the unfavorable outcome of this cohort.

The second principal component of the PCA of all samples showed heterogeneity within the non-malignant group (Figure 2A). CTRL samples formed a relatively homogeneous population, distinct from uninvolved margins (UMs), while UM samples were more variable and dispersed over a broader area. To investigate this further, we performed PCA solely on UM and CTRL samples. This revealed a subset of UMs forming a distinct population (Figure S2A). Notably, the list of the top 50 genes that accounted for the variability in the first principal component (y-axis), distinguishing between UM and CTRL samples, included genes related to epithelial matrix structure and organization (Figure S2B).

3.2 | UM tissues display abnormal features according to PAM50 gene classifier

AIMS and PAM50 predictors were applied to all samples to corroborate the histopathological classification. Both tools validated and classified all samples from reduction mammoplasty surgeries (CTRL) as normal-like. In tumor fragments, AIMS and PAM50 classifiers tend to agree more in assigning the basal-like subtype to PT samples, rather than Luminal A, Luminal B, or HER2-enriched (AIMS vs. histopathological classification accuracy: 70% (Luminal A), 74% (Luminal B), 65% (HER2-enriched), and 88% (Basal-like)/PAM50 versus histopathological classification accuracy: 0 (Luminal A), 66% (Luminal B), 70% (HER2-enriched), and 87% (Basal-like)) (Additional File S1, Table S5). However, we noticed some disagreement in UM samples; while AIMS classified most as normal-like, PAM50 assigned ~40% of the UM samples to tumor-like subtypes. This discrepancy could potentially be explained by the different number of genes incorporated and the different principles used by each tool. At the same time, PAM50 probability scores for individual samples indicated that the classification of UMs often balanced between the normal-like and tumor-like subtypes suggesting the presence of features deviating from the "normal" state (Additional File S1, Table S5).

3.3 | A distinct cluster emerges within uninvolved margin tissues

Through hierarchical clustering of all samples in our dataset, using the expression profiles of panel genes, we identified four distinct clusters

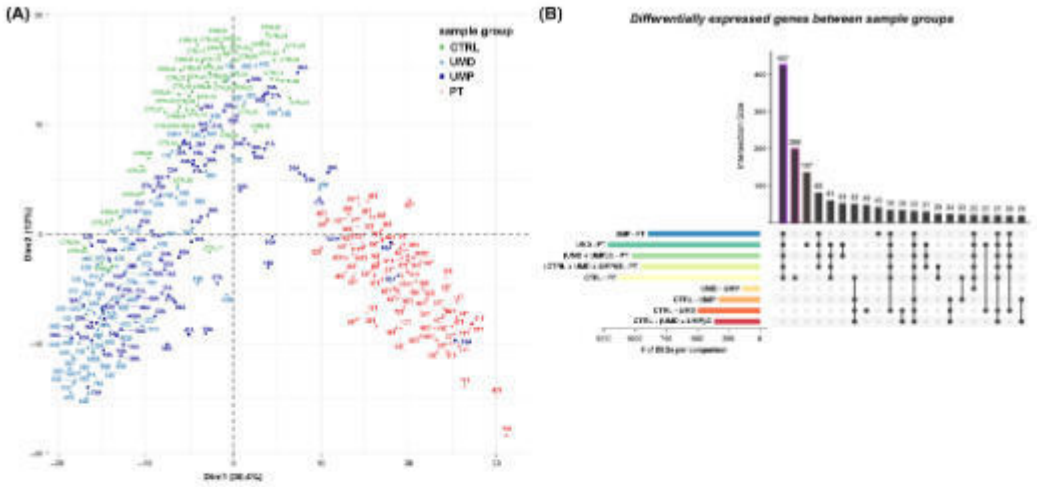

FIGURE 2 (A) Principal Component Analysis (PCA) of primary tumor (PT), uninvolved margin (UM), and control (CTRL) samples. PCA was performed based on the expression of panel genes. UM samples include uninvolved margin proximal to primary tumor (UMP) and uninvolved margin distal from the primary tumor (UMD). This analysis illustrates the broad dispersion of breast cancer and morphologically normal tissue samples across the main principal axes. Each point represents the orientation of a sample projected into the transcriptional space, color-coded to indicate its group membership. Tumor profiles primarily aggregate in a distinct quadrant of the transcriptional space, whereas UM and CTRL tissues occupy two separate quadrants. The first PC represents the maximum variance direction in the data (30.4%), which corresponds to the differences between primary tumors versus all other samples. The second PC primarily reflects differences between UM and CTRL profiles, with proximal and distal profiles displaying a broad spread across the main principal axes. The PCA plot was generated with the fviz_pca_ind function from the R package factoextra (version 1.x0.7). (B) Differential gene expression analysis across primary tumor (PT), uninvolved margin (UM), and control (CTRL) samples. UM samples include uninvolved margin proximal to primary tumor (UMP) and uninvolved margin distal from the primary tumor (UMD). The highest number of DEGs (false discovery rate [FDR] ≤0.05 and a log-fold change of ≥1; QLF test) is observed when comparing PT versus all non-malignant tissues, while UMP and UMD share few DEGs indicating overall similarity between their expression profiles. The variable contribution plots were generated with the fviz_contrib function.

(Figure 3). Two of these clusters, referred to as Cluster 1 and Cluster 2, were predominantly populated by PT samples. Clusters 1 and 2 appeared to be formed according to PT molecular subtypes, as determined by histopathological evaluations combined with AIMS and PAM50 predictors. Cluster 1 ($n = 25$) predominantly contained HER2-enriched and basal-like tumors, whereas Cluster 2 ($n = 57$) was mainly composed of luminal tumors. The remaining two clusters included non-malignant samples. CTRL samples, barring two exceptions, populated Cluster 3 ($n = 145$), while UM samples dispersed between Clusters 3 and 4 ($n = 68$). We further sought to investigate why UM samples split between these two clusters, while CTRL samples largely coalesced within Cluster 3, despite both originating from the same tissue type. The notable difference in age between breast cancer patients (UM samples) and individuals subjected to reduction mammoplasty surgeries (CTRL samples) could potentially be an element adding to the situation, although we included age as a covariate in DE analysis.

A subsequent DE analysis that incorporated cluster assignment along with the sample type, revealed that the highest number of differentially expressed genes was observed between CTRL samples in Cluster 3 and UM samples in Cluster 4 (FDR ≤0.05, log-fold change of ≥1). The second-highest number of DEGs appeared when comparing UM profiles between Clusters 3 and 4 (Figure S3).

Remarkably, the top down-regulated genes in UM tissues in Cluster 4 included keratins: *KRT14*, *KRT15*, *KRT17*, *KRT6B*, *KRT5*, *KRT7*, *KRT19*, cell adhesion-related genes: *CDH1*, *CDH3*, *EPCAM*, and a matrix metallopeptidase *MMP7*. This list also comprised transcription factors *FOXI1*, *FOXA1*—tumor suppressor or candidate tumor suppressor genes, dual-role genes *RAB25*, *NRG1*, *SPDEF*, *TRIM29*, and the *GABRP* gene previously associated with breast cancer metastatic potential. A selection of these genes is presented in Figure 4A. The statistical significance of these findings persisted ($p < .05$, Quasi-Likelihood *F*-test—QLF) when including the sample group information and even under multiple comparison scenarios (UMs in Cluster 4 vs. UMs in Cluster 3, UMPs in Cluster 4 vs. UMPs in Cluster 3, UMDs in Cluster 4 vs. UMDs in Cluster 3) (Figure 4B). These genes exhibited a bimodal expression pattern in both types of UMs, best explained by the split of UM samples between Clusters 3 and 4. They form a distinct signature, hereby named as KAOS signature for Keratin-Adhesion-Oncogenes-Suppressors.

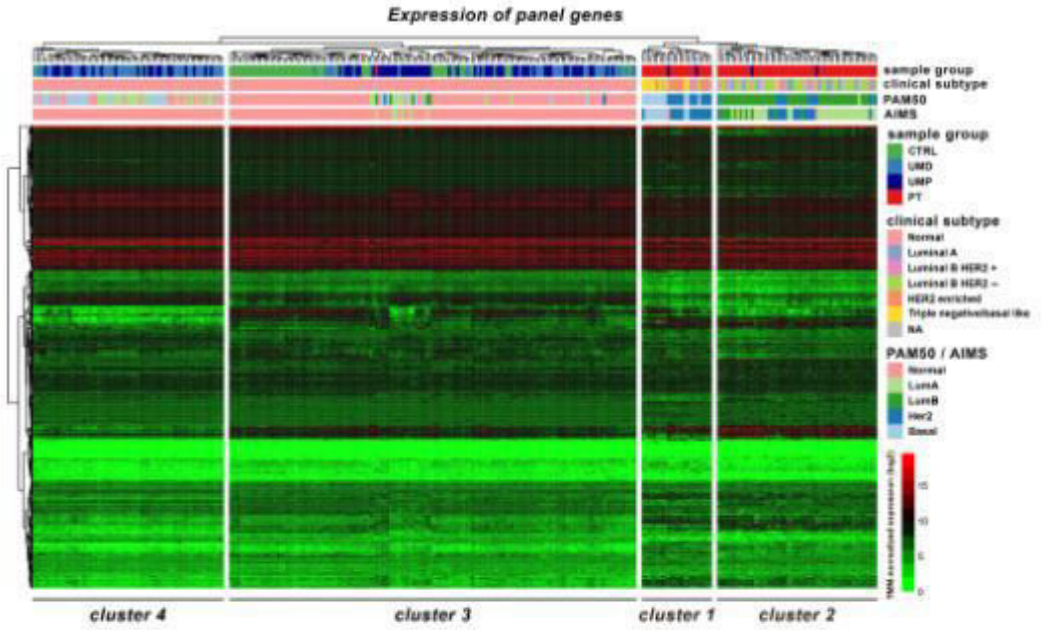

FIGURE 3 Hierarchical clustering of primary tumor (PT), uninvolved margin (UM), and control (CTRL) samples. Clustering performed using the expression profiles of panel genes reveals four distinct clusters. Cluster 1 ($n = 25$) primarily contains HER2-enriched and basal-like tumors, whereas Cluster 2 ($n = 57$) mainly includes luminal tumors. The remaining two clusters accommodate non-malignant samples. CTRL samples, with the exception of two cases, are mainly clustered in Cluster 3 ($n = 145$), while UM samples are distributed across Clusters 3 and 4 ($n = 68$). The clinical subtype assigned via histopathological examination is illustrated for all primary tumors. Additionally, subtype information, assigned via AIMS and PAM50 predictors is included for all samples in the dataset (Additional File S1, Table S5).

3.4 | Cluster 4 is distinguished from other non-malignant profiles through enrichment analysis

Enrichment analyses were conducted to identify disrupted GO terms and KEGG pathways in samples located in Cluster 4 (hypergeometric test, FDR <0.05) (Additional File S1, Tables S6 and S7). Analyses revealed that primarily epithelial/stem cell developmental processes, the estrogen signaling pathway, and the cell adhesion molecules pathway were up-regulated in Cluster 3 UMs relative to Cluster 4 UMs. In contrast, the "Regulation of Lipolysis in Adipocytes" and the "PPAR Signaling Pathway" were significantly down-regulated in Cluster 3 UMs compared to Cluster 4 UMs (Figure 5A). Notably, all above-described observations remained consistent when performing multiple comparisons between UM and CTRL samples located in Clusters 3 and 4 (Figure 5B). Re-clustering of samples based on the expression of targeted genes involved in metabolic-related pathways distinguishingly separated malignant samples (PT) from the non-malignant samples (UM, CTRL) when clustering the full dataset for genes included in the "PPAR signaling pathway." UM and CTRL samples originally grouped into Clusters 3 and 4 were again separated in the new clusters, however, new Cluster 2 comprising mostly CTRL samples, also included several PTs (Figure S4). Re-clustering the full dataset for genes included in the "Regulation of lipolysis in adipocytes" pathway disagreed with the original clustering of samples (Figure S5). It should be noted here that the customized RNA-sequencing panel's ability to efficiently capture crucial information, representative of the full mammary tissue transcriptome, was validated by analyzing two distinct external datasets (full transcriptome—custom RNA-seq panel) of PT and paired UM samples from the same 18 breast cancer patients that originated from other breast cancer study in our lab (Figure S6).

3.5 | Cluster 4 significantly associates with patients' clinical outcome

Cluster 4 shared a greater degree of similarity with Cluster 3, populated by CTRLs and UMs, than Clusters 1 and 2, which were dominated by malignant profiles (Figure S7). Cluster 4 was significantly enriched with UM samples ($p = 1.98e\text{-}08$, Fisher's test), encompassing 23% of all samples and 40% of total UM samples, representing 41% of patients. Furthermore, it was significantly enriched ($p = 7.62e\text{-}05$, Fisher's test) with samples that were classified by

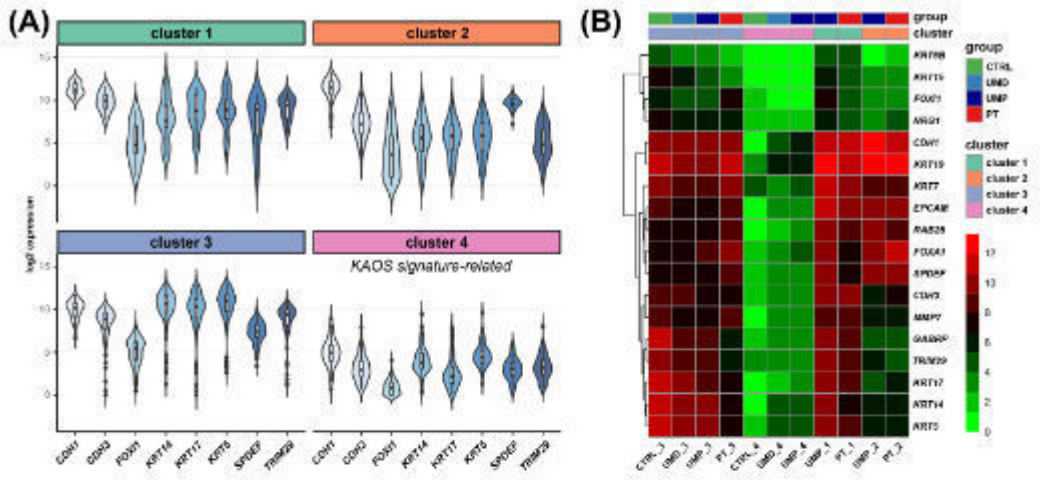

FIGURE 4 (A) Violin/box plots illustrating the expression profiles of genes included in the identified signature in identified Clusters 1, 2, 3, and 4. Low expression of selected genes is noted in Cluster 4 compared to Cluster 3. (B) Heatmap with average expression of genes included in the identified signature in Clusters 1, 2, 3, and 4. Further stratification, including sample group, that is, primary tumor (PT), uninvolved margin proximal (UMP) and distal (UMD) from the PT, and control (CTRL) highlights a relative down-regulation of gene expression in UMP, UMD, and CTRL samples located in Cluster 4 compared to those in Cluster 3.

PAM50 as one of the breast cancer molecular subtypes (when comparing within UMs only). This association was significant for all UMs, as well as for both distal and proximal UMs ($p = 1.02e\ 08$ and $p = .00256$, respectively). However, the impact of each individual's stage (i.e., stratification of advanced disease based on patient's history and physical examination findings supplemented by imaging and pathology data[27–29]) was not significant in terms of sample clustering (Pearson's Chi-squared test, $p = .6774$). Specifically, stage was not a factor contributing to the assignment of UM samples into Clusters 3 or 4.

Identification of a distinct patient group, having both UM samples (proximal and distal) in one cluster, allowed us to execute comprehensive comparisons using follow-up information collected for each patient. Patients with both UMs in Cluster 4, as opposed to Cluster 3, exhibited smaller tumor sizes as measured by ultrasonography and pathological examination ($p = .0013$ and $p = .033$, respectively, Mann–Whitney U test) and were older ($p = .025$, Mann–Whitney U test). Cluster 4 membership was also associated with tumor HER2 positive status ($p = .004265$, Fisher's test). Upon restricting our analysis to patients with only one UM sample assigned to either Cluster 3 or 4, it was observed that Cluster 4 had a notable overrepresentation of patients with a positive death status ($p = .04493345$ and $p = .01512627$, Fisher's test for UMD and UMP, respectively). Finally, when performing another comparison, for patients with a strictly defined UMs clustering pattern, that is, UMD in Cluster 3 and UMP in Cluster 4, a substantial link to patient death status emerged, contrasting with the patients who had the reverse UM assignment (UMD in Cluster 4 and UMP in Cluster 3) ($p = .001396$). These findings indicate that the spatial information of uninvolved mammary tissue, obtained by samples at different distances from the primary tumor, could reveal different pieces of information of the patient's clinical picture.

4 | DISCUSSION

This study takes an important step toward understanding the properties of microscopically normal, uninvolved mammary tissue in breast cancer patients with unfavorable outcomes. Our results provide new insight into the concept of biological abnormality in histologically normal mammary gland tissue, removed at different distances from the primary tumor.[11–13,30]

We identified a distinct subset of histologically non-cancerous uninvolved mammary (UM) samples, referred to as Cluster 4, which demonstrates unique attributes in contrast to other non-cancerous UM samples and, importantly, to mammary gland samples collected from individuals without cancer (CTRL). Notably, Cluster 4 is significantly enriched with UMs (both proximal and distal) categorized as tumor-like by PAM50. This likely suggests the presence of feature characteristics in Cluster 4 UMs divergent from the "normal" state.

Furthermore, a distinct gene signature present in Cluster 4 UMs, named as KAOS signature, involving the down-regulation of genes participating in various processes, supports the above-mentioned thesis. Genes comprising this signature can be grouped into two main categories; (a) cell adhesion and structural support genes, and (b) transcription factors and tumor suppressors/oncogenes. The first

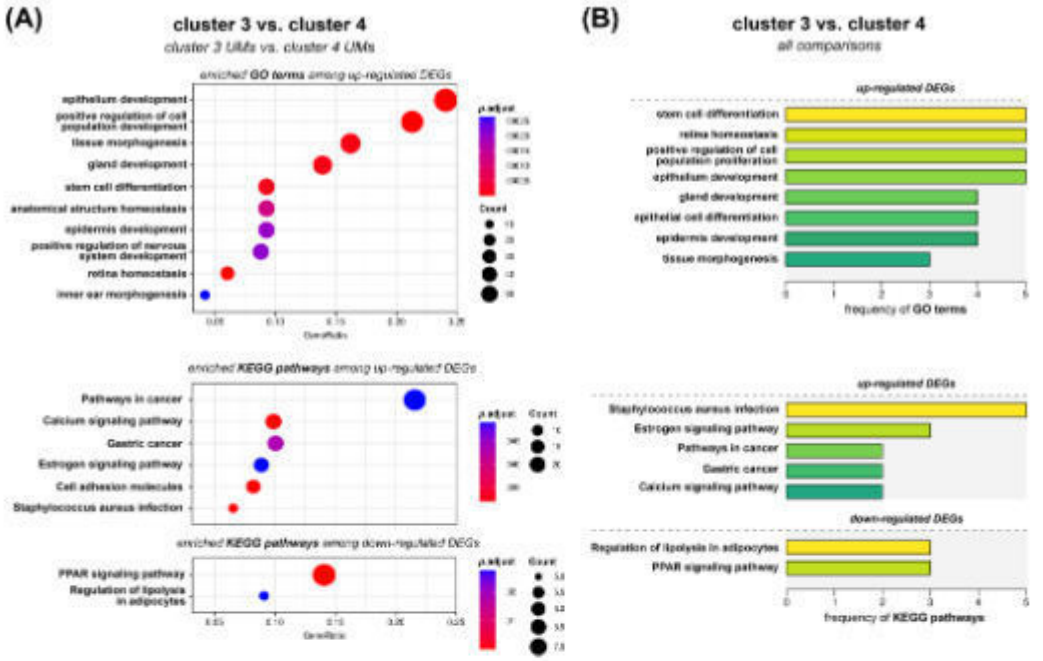

FIGURE 5 (A) Enrichment analysis across uninvolved margin (UM) samples in Clusters 3 and 4. Analysis performed for differentially expressed genes (DEGs) identifies Gene Ontology (GO) terms and KEGG pathways. A maximum of 10 significantly enriched GO terms/KEGG pathways are presented here. The DEG sets were filtered to retain only those demonstrating a log fold change (logFC) of ≥1, with an adjusted *p*-value (*p* adj) of <.05. (B) Frequency of enriched features across multiple comparisons involving Clusters 3 and 4. Comparisons include UM_3 versus UM_4, UMP_3 versus UMP_4, UMD_3 versus UMD_4, and CTRL_3 versus UMs_4 (full list of DE tests is available in the Additional File S1, Tables S6 and S7). GO terms/KEGG pathways with a frequency of at least 3 among the conducted comparisons are presented. The Estrogen signaling pathway, the PPAR signaling pathway, and the Regulation of lipolysis in adipocytes pathway consistently appear enriched among DEGs upregulated in Cluster 4 UM samples upon multiple comparisons.

group included genes encoding for keratins (*KRT14*, *KRT15*, *KRT17*, *KRT6B*, *KRT5*, *KRT7*, *KRT19*), metallopeptidases (*MMP7*), and cell adhesion molecules (*CDH1*, *CDH3*, *EPCAM*). Motility keratins *KRT5*, *KRT14*, and *KRT17*, have been previously implicated in "subtype switching," that is, switching of molecular subtype between lung and pleura metastases versus the primary breast tumor,[31] as well as deemed essential for the tumorigenic potential and migration of a basal-like breast cancer cell line along with *GABRP*.[32] Furthermore, *KRT19*, *KRT7*, and *KRT15* are individually linked to breast cancer, especially associated with metastatic potential and poor patient outcomes.[33–35] Diminished cytokeratin, cell adhesion-related (*CDH1*, *CDH3*, *EPCAM*) and matrix metallopeptidase (*MMP7*) gene expression indicates the presence of alterations linked to invasiveness and signifies disruptions in the cytoskeleton, pointing to loss of adhesion and epithelial tissue integrity. Subsequently, these observations point to epithelial-to-mesenchymal transition (EMT), an otherwise normal process during which epithelial cells acquire migratory and invasive properties, observed also in tumor metastasis.[36,37] Nonetheless, we did not observe concomitant up-regulation of mesenchymal markers. This suggests the potential presence of a partial/hybrid EMT.[38] The second group contained candidate oncogenes and candidate tumor suppressor genes in breast cancer (*RAB25*, *NRG1*, *SPDEF*), reportedly having dual-functioning roles associated with estrogen (ER) status.[39–41] Additionally, this group included the tumor suppressor *TRIM29* gene, whose depletion has been linked with preneoplastic changes such as loss of polarity and increased migration and invasion in non-tumorigenic breast cells,[42] as well as alteration of keratin expression to enhance cell invasion in squamous cell carcinoma.[43] Finally, members of the forkhead box transcription family (*FOXA1*, *FOXI1*) previously associated with breast cancer,[44,45] were part of this group. The observed down-regulation of genes with dual roles in cancer, both established and candidate tumor suppressor genes, as well as transcription factors, underscores the disturbed environment in the Cluster 4 UM samples.

Interestingly, Cluster 4 was further characterized by the down-regulation of the estrogen signaling pathway and the up-regulation of the "Regulation of lipolysis in adipocytes" and the "PPAR signaling" pathways. Deregulation of the latter, an upstream effector of fatty

acid oxidation,[46] suggests a link to metabolic imbalance. Disruption of metabolic-related processes has been observed in patients with altered PPAR signaling, reinforcing the established role of PPARs in lipid transport, fatty acid oxidation, and their involvement in crosstalk with other lipogenic pathways.[47] Intriguingly, PPARs can share common ligands with estrogen receptors (ERs) and both have contrasting regulatory effects on the PIK3K/AKT signaling pathway, which influences breast cancer cell survival and proliferation.[48]

We observed a significant association between cluster membership and less favorable patient outcomes, as patients with UM samples in Cluster 4 were associated with a positive death status, and also with smaller size tumors. The latter finding is intriguing, although not straightforward to explain. Patients enrolled in this study, barring two exceptions, experienced mortality primarily attributed to breast cancer itself, disease recurrence, or the emergence of secondary tumors. However, the presence of comorbidities (undocumented here) such as hypertension, cardiovascular disease (CVD), and type 2 diabetes can affect the progression of the disease, complicate treatment, and influence the patient's health outcome.[49] Although we did not find a direct significant association between cluster membership (for UM samples) and recurrence or secondary tumor events in corresponding patients, our findings could likely reflect the overall systemic aggressiveness of the disease. The concept of untransformed cells dissociating from the original diseased organ, disseminating via the vascular system, and incorporating into parts of otherwise normal-appearing organs to seed metastases, has been recently raised again through the work of Rahrmann et al.[50] The disturbed tissue microenvironment found in Cluster 4 UMs could potentially facilitate the homing of these untransformed cells early on, thus assisting in the spread of cancer throughout the body and ultimately leading to death.

The etiologic field theory,[51] a different perspective on the field effect theory initially proposed by Slaughter's group in 1953,[52] supports the above-mentioned thesis. This concept embraces tumor–host and gene–environment interactions and highlights the existence of an abnormal tissue microenvironment present within microscopically normal tissue that can influence every stage of tumor development. Importantly, the etiologic field effect concept challenges the notion that markers exclusively indicate neoplasia. Instead, it suggests that these markers may represent environmental changes, including the potential contribution of non-transformed cells and extracellular matrices to neoplastic evolution. A continuous model, involving multiple stages, favoring the acquisition of alterations, might be a better representation of a realistic tumorigenic process.

Our findings present an alternative perspective to a previous study, which postulated that histologically normal tissue adjacent to breast cancer exhibits only minimal gene expression changes compared to breast reduction tissue.[53] According to that study, these differences in gene expression primarily represented individual tissue- and patient-specific variability, rather than any associations with the patient's clinical picture. However, it is worth noting that our study differed in terms of patient selection, as we focused on patients with adverse outcomes, including the presence of recurrence, the emergence of a second independent tumor, or mortality, as the principal inclusion criteria. Furthermore, the determination of tumor adjacency varied across these two studies. Consequently, our findings indicate the development of a pre-tumoral, change-favoring environment, a feature characteristic of patients with a higher risk of recurrence and a decreased survival rate.

While our study offers valuable insights into the molecular changes occurring within the uninvolved mammary gland of breast cancer patients with unfavorable outcomes, it does come with certain limitations. We understand that we may not have captured the complete spectrum of molecular changes happening within these tissues, since we only focused on aberrations at the gene expression level. Secondly, we did not include samples from metastatic sites in our study design. The incorporation of samples from secondary lesions would have allowed for a thorough assessment and comparison of gene expression profiles among primary tumor sites, surrounding normal-appearing gland tissue, and metastatic sites. In this context, examining the tumor microenvironment, specifically focusing on alterations in stromal and immune cells, might have provided valuable insights. However, the procurement of the corresponding samples presents significant challenges. Finally, the transition from a cross-sectional to a longitudinal study design might have allowed us to better track the evolution of the tissues over time, their contribution to cancer progression and metastasis as well as the influence of external factors, such as the patient's lifestyle and environmental factors.

Nevertheless, the significant link between the clustering pattern and patient death status implies a potential prognostic value, suggesting that the spatial distribution of uninvolved mammary tissue could hold crucial information about breast cancer outcomes.

Our study highlights the potential presence of a pre-tumorigenic environment, within the ostensibly normal mammary gland tissue, promoting changes that are closely linked to patient mortality. The aberrant gene expression profiles of uninvolved mammary tissue intriguingly exhibit tumor-like characteristics as shown by the PAM50 predictor, marked by dysregulation of crucial pathways such as estrogen and PPAR signaling.

It remains to be determined whether these observed alterations stem from the nearby tumor's influence or signify the independent emergence of early pre-tumorous conditions facilitated by a perturbed environment. The strong association of Cluster 4 characteristics with mortality, but not directly with recurrence, may suggest these features are more indicative of the disease's systemic aggressiveness than of its potential to re-emerge.

This study offers an indication for comprehensive monitoring of breast cancer patients with recurrence or secondary tumor events. Integrating molecular assessments of non-malignant mammary tissue into disease management strategies could enhance personalized patient care, including improved survival prediction.

AUTHOR CONTRIBUTIONS

Conceptualization: Arkadiusz Piotrowski, Jan P. Dumanski, Natalia Filipowicz, Maria Andreou. Resources: Jan P. Dumanski, Tomasz Nowikiewicz, Wojciech Zegarski, Łukasz Szylberg, Michał Jankowski, Jerzy

Jankau, Ewa Śrutek, Manuela Las-Jankowska, Dariusz Bała, Jacek Hoffman, Magdalena Nowikiewicz, Rafał Pęksa, Jarosław Skokowski, Johan Hartman, Hanna Davies, Bożena Bruhn-Olszewska. Data Curation: Natalia Filipowicz, Arkadiusz Piotrowski, Maria Andreou, Tomasz Nowikiewicz, Katarzyna Duzowska. Investigation: Maria Andreou, Natalia Filipowicz, Katarzyna Duzowska, Urszula Ławrynowicz, Katarzyna Chojnowska, Monika Horbacz. Methodology: Anna Kostecka, Natalia Filipowicz, Monika Horbacz. Data analysis: Marcin Jąkalski, Jakub Mieczkowski Visualization: Marcin Jąkalski, Maria Andreou. Interpretation: Maria Andreou, Arkadiusz Piotrowski, Marcin Jąkalski, Jakub Mieczkowski, Anna Kostecka. Article writing-original: Maria Andreou, Marcin Jąkalski, Arkadiusz Piotrowski. Article writing – review, editing, and acquisition of additional data for review: A.M., Marcin Jąkalski, Arkadiusz Piotrowski, Jan P. Dumanski, Jakub Mieczkowski, Natalia Filipowicz, Anna Kostecka, Monika Horbacz, Hanna Davies, Mateusz Maniewski, Tomasz Nowikiewicz, Wojciech Zegarski, Łukasz Szylberg, Michał Jankowski, Jerzy Jankau, Ewa Śrutek, Manuela Las-Jankowska, Dariusz Bała, Jacek Hoffman, Magdalena Nowikiewicz, Rafał Pęksa, Jarosław Skokowski, Johan Hartman, Bożena Bruhn-Olszewska, Katarzyna Duzowska, Urszula Ławrynowicz, Katarzyna Chojnowska. All authors have read and approved the article. Maria Andreou and Marcin Jąkalski have contributed equally to this work. Supervision: Jan P. Dumanski, Jakub Mieczkowski, Arkadiusz Piotrowski. The work reported in the article has been performed by the authors, unless clearly specified in the text.

AFFILIATIONS

[1]3P-Medicine Laboratory, Medical University of Gdańsk, Gdańsk, Poland
[2]Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[3]Department of Plastic Surgery, Medical University of Gdańsk, Gdańsk, Poland
[4]Department of Surgical Oncology, Ludwik Rydygier's Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland
[5]Department of Tumor Pathology and Pathomorphology, Oncology Center-Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland
[6]Chair of Surgical Oncology, Ludwik Rydygier's Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland
[7]Department of Clinical Oncology, Oncology Center-Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland
[8]Department of Surgical Oncology, Oncology Center-Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland
[9]Department of Clinical Breast Cancer and Reconstructive Surgery, Oncology Center-Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland
[10]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden
[11]Department of Pathology, Karolinska University Hospital, Stockholm, Sweden
[12]MedTech Labs, Bioclinicum, Karolinska University Hospital, Stockholm, Sweden
[13]Department of Pathomorphology, Medical University of Gdańsk, Gdańsk, Poland
[14]Academy of Applied Medical and Social Science, Elbląg, Poland
[15]Department of Obstetrics, Gynaecology and Oncology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, Bydgoszcz, Poland
[16]Department of Hepatobiliary and General Surgery, Antoni Jurasz University Hospital, Bydgoszcz, Poland
[17]Department of Biology and Pharmaceutical Botany, Medical University of Gdańsk, Gdańsk, Poland

ACKNOWLEDGMENTS

The authors wish to thank Agata Wojdak and Kinga Drężek for their help with administrative and wet lab activities, respectively. We also would like to thank all the patients and volunteer control individuals for acceptance to participate in the study and sample contribution; hospital staff involved in the patient recruitment process in Oncology Center—Prof. Franciszek Łukaszczyk Memorial Hospital in Bydgoszcz and University Clinical Centre in Gdańsk. We thank Dr. Leszek Kalinowski for the access to selected laboratory facilities.

FUNDING INFORMATION

This research was supported by the Foundation for Polish Science under the International Research Agendas Program financed from the Smart Growth Operational Program 2014–2020 (Grant Agreement No. MAB/2018/6) and by The Swedish Cancer Society (No. 20 0889 PjF) and Swedish Medical Research Council (No. 2020-02010) to J. P.D.

CONFLICT OF INTEREST STATEMENT

Jan P. Dumanski is cofounder and shareholder in Cray Innovation AB. Jakub Mieczkowski is a co-founder and shareholder of Genegoggle sp. z o.o. The remaining authors have declared that no competing interests exist.

DATA AVAILABILITY STATEMENT

The bulk RNA-seq data generated in this study are available from the European Genome-Phenome Archive (EGA, https://ega-archive.org/) under ID EGAS50000000011 accession number. All other primary data analyzed and presented in this study are located in the Supplementary files attached to this article. Unless otherwise stated, all analyses were conducted in R (version 4.1.2). The code used for data processing is available on GitHub: https://github.com/jakalsj3/Breast_cancer_KAOS. Further information is available from the corresponding author upon request.

ETHICS STATEMENT

Tissue samples and patient histories were provided for this study by the Oncology Centre in Bydgoszcz and the University Clinical Centre in Gdańsk, who, under a research protocol approved by the Bioethical Committee at the Collegium Medicum, Nicolaus Copernicus University in Toruń (approval number KB509/2010) and by the Independent Bioethics Committee for Research at the Medical

University of Gdansk (approval number NKBBN/564/2018 with multiple amendments), recruited and enrolled all donors under informed and written consent, collected, and stored all tissue samples.

ORCID

Arkadiusz Piotrowski https://orcid.org/0000-0002-0823-0607

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660
2. Wilkinson L, Gathani T. Understanding breast cancer as a global health concern. *Br J Radiol*. 2022;95(1130):20211033. doi:10.1259/bjr.20211033
3. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet*. 2021;397(10286):1750-1769. doi:10.1016/S0140-6736(20)32381-3
4. Waks AG, Winer EP. Breast cancer treatment: a review. *Jama*. 2019;321(3):288-300. doi:10.1001/jama.2018.19323
5. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S, McGale P, et al. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. *Lancet*. 2011;378(9804):1707-1716. doi:10.1016/S0140-6736(11)61629-2
6. Early Breast Cancer Trialists' Collaborative Group (Ebctcg). Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. *Lancet*. 2012;379(9814):432-444. doi:10.1016/S0140-6736(11)61625-5
7. Danforth DN. Genomic changes in normal breast tissue in women at normal risk or at high risk for breast cancer. *Breast Cancer Basic Clin Res*. 2016;10:109-146. doi:10.4137/BCBCR.S39384
8. Kostecka A, Nowikiewicz T, Olszewski P, et al. High prevalence of somatic PIK3CA and TP53 pathogenic variants in the normal mammary gland tissue of sporadic breast cancer patients revealed by duplex sequencing. *npj Breast Cancer*. 2022;8(1):1-10. doi:10.1038/s41523-022-00443-9
9. Ronowicz A, Janaszak-Jasiecka A, Skokowski J, et al. Concurrent DNA copy-number alterations and mutations in genes related to maintenance of genome stability in uninvolved mammary glandular tissue from breast cancer patients. *Hum Mutat*. 2015;36(11):1088-1099. doi:10.1002/humu.22845
10. Forsberg LA, Rasi C, Pekar G, et al. Signatures of post-zygotic structural genetic aberrations in the cells of histologically normal breast tissue that can predispose to sporadic breast cancer. *Genome Res*. 2015;25(10):1521-1535. doi:10.1101/gr.187823.114
11. Huang X, Stern DF, Zhao H. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival – evidence from TCGA pan-cancer data. *Sci Rep*. 2016;6(1):20567. doi:10.1038/srep20567
12. Gadaleta E, Fourgoux P, Pirró S, et al. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. *npj Breast Cancer*. 2020;6:38. doi:10.1038/s41523-020-00182-9
13. Aran D, Camarda R, Odegaard J, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*. 2017;8(1):1077. doi:10.1038/s41467-017-01027-z
14. Tripathi A, King C, De La Morenas A, et al. Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int J Cancer*. 2007;122(7):1557-1566. doi:10.1002/ijc.23267
15. Graham K, de Las Morenas A, Tripathi A, et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br J Cancer*. 2010;102(8):1284-1293. doi:10.1038/sj.bjc.6605576
16. Filipowicz N, Drężek K, Horbacz M, et al. Comprehensive cancer-oriented biobanking resource of human samples for studies of post-zygotic genetic variation involved in cancer predisposition. *PLoS One*. 2022;17(4):e0266111. doi:10.1371/journal.pone.0266111
17. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370
18. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst*. 2015;107(1):357. doi:10.1093/jnci/dju357
19. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616
21. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25
22. Lê S, Josse J, Husson F. FactoMineR: an *R* package for multivariate analysis. *J Stat Softw*. 2008;25(1):1-18. doi:10.18637/jss.v025.i01
23. Gendoo DMA, Ratanasirigulchai N, Schröder MS, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2016;32(7):1097-1099. doi:10.1093/bioinformatics/btv693
24. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118
25. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883. doi:10.1093/bioinformatics/bts034
26. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. 2020;2(3):lqaa078. doi:10.1093/nargab/lqaa078
27. Amin MB, Greene FL, Edge SB, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99. doi:10.3322/caac.21388
28. Giuliano AE, Connolly JL, Edge SB, et al. Breast cancer-major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin*. 2017;67(4):290-303. doi:10.3322/caac.21393
29. Giuliano AE, Edge SB, Hortobagyi GN. Eighth edition of the AJCC cancer staging manual: breast cancer. *Ann Surg Oncol*. 2018;25(7):1783-1785. doi:10.1245/s10434-018-6486-6
30. Román-Pérez E, Casbas-Hernández P, Pirone JR, et al. Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients. *Breast Cancer Res*. 2012;14(2):R51. doi:10.1186/bcr3152
31. Klebe M, Fremd C, Kriegsmann M, et al. Frequent molecular subtype switching and gene expression alterations in lung and pleural metastasis from luminal A-type breast cancer. *JCO Precis Oncol*. 2020;4:848-859. doi:10.1200/PO.19.00337
32. Sizemore GM, Sizemore ST, Seachrist DD, Keri RA. GABA (A) receptor pi (GABRP) stimulates basal-like breast cancer cell migration through activation of extracellular-regulated kinase 1/2 (ERK1/2). *J Biol Chem*. 2014;289(35):24102-24113. doi:10.1074/jbc.M114.593582
33. Mi L, Liang N, Sun H. A comprehensive analysis of KRT19 combined with immune infiltration to predict breast cancer prognosis. *Genes*. 2022;13(10):1838. doi:10.3390/genes13101838

34. Statz E, Jorns JM. Cytokeratin 7, GATA3, and SOX-10 is a comprehensive panel in diagnosing triple negative breast cancer brain metastases. *Int J Surg Pathol*. 2021;29(5):470-474. doi:10.1177/1066896921990717
35. Zhong P, Shu R, Wu H, Liu Z, Shen X, Hu Y. Low KRT15 expression is associated with poor prognosis in patients with breast invasive carcinoma. *Exp Ther Med*. 2021;21(4):305. doi:10.3892/etm.2021.9736
36. Leggett SE, Hruska AM, Guo M, Wong IY. The epithelial-mesenchymal transition and the cytoskeleton in bioengineered systems. *Cell Commun Signal*. 2021;19(1):32. doi:10.1186/s12964-021-00713-2
37. Nieto MA, Huang RYJ, Jackson RA, Thiery JP. EMT: 2016. *Cell*. 2016;166(1):21-45. doi:10.1016/j.cell.2016.06.028
38. Yamashita N, Tokunaga E, Iimori M, et al. Epithelial paradox: clinical significance of coexpression of E-cadherin and vimentin with regard to invasion and metastasis of breast cancer. *Clin Breast Cancer*. 2018;18(5):e1003-e1009. doi:10.1016/j.clbc.2018.02.002
39. Cheng JM, Volik L, Janaki DKM, Vyakaranam S, Ran S, Rao KA. Tumor suppressor function of Rab25 in triple-negative breast cancer. *Int J Cancer*. 2010;126(12):2799-2812. doi:10.1002/ijc.24900
40. Chua YL, Ito Y, Pole JCM, et al. The NRG1 gene is frequently silenced by methylation in breast cancers and is a strong candidate for the 8p tumour suppressor gene. *Oncogene*. 2009;28(46):4041-4052. doi:10.1038/onc.2009.259
41. Ye T, Li J, Feng J, et al. The subtype-specific molecular function of SPDEF in breast cancer and insights into prognostic significance. *J Cell Mol Med*. 2021;25(15):7307-7320. doi:10.1111/jcmm.16760
42. Liu J, Welm B, Boucher KM, Ebbert MTW, Bernard PS. TRIM29 functions as a tumor suppressor in nontumorigenic breast cells and invasive ER+ breast cancer. *Am J Pathol*. 2012;180(2):839-847. doi:10.1016/j.ajpath.2011.10.020
43. Yanagi T, Watanabe M, Hata H, et al. Loss of TRIM29 alters keratin distribution to promote cell invasion in squamous cell carcinoma. *Cancer Res*. 2018;78(24):6795-6806. doi:10.1158/0008-5472.CAN-18-1495
44. Seachrist DD, Anstine LJ, Keri RA. FOXA1: a pioneer of nuclear receptor action in breast cancer. *Cancer*. 2021;13(20):5205. doi:10.3390/cancers13205205
45. Onodera Y, Takagi K, Neoi Y, et al. Forkhead box I1 in breast carcinoma as a potent prognostic factor. *Acta Histochem Cytochem*. 2021;54(4):123-130. doi:10.1267/ahc.21-00034
46. Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. *Nat Rev Cancer*. 2013;13(4):227-232. doi:10.1038/nrc3483
47. Kersten S, Desvergne B, Wahli W. Roles of PPARs in health and disease. *Nature*. 2000;405(6785):421-424. doi:10.1038/35013000
48. Bonofiglio D, Gabriele S, Aquila S, et al. Estrogen receptor alpha binds to peroxisome proliferator-activated receptor response element and negatively interferes with peroxisome proliferator-activated receptor gamma signaling in breast cancer cells. *Clin Cancer Res*. 2005;11(17):6139-6147. doi:10.1158/1078-0432.CCR-04-2453
49. Connor AE, Schmaltz CL, Jackson-Thompson J, Visvanathan K. Comorbidities and the risk of cardiovascular disease mortality among racially diverse patients with breast cancer. *Cancer*. 2021;127(15):2614-2622. doi:10.1002/cncr.33530
50. Rahrmann EP, Shorthouse D, Jassim A, et al. The NALCN channel regulates metastasis and nonmalignant cell dissemination. *Nat Genet*. 2022;54(12):1827-1838. doi:10.1038/s41588-022-01182-0
51. Lochhead P, Chan AT, Nishihara R, et al. Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. *Mod Pathol*. 2015;28(1):14-29. doi:10.1038/modpathol.2014.81
52. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*. 1953;6(5):963-968. doi:10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q
53. Finak G, Sadekova S, Pepin F, et al. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res*. 2006;8(5):R58. doi:10.1186/bcr1608

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Andreou M, Jąkalski M, Duzowska K, et al. Prelude to malignancy: A gene expression signature in normal mammary gland from breast cancer patients suggests pre-tumorous alterations and is associated with adverse outcomes. *Int J Cancer*. 2024;155(9):1616-1628. doi:10.1002/ijc.35050