Beyond Tumors: Reduced survival linked to pathogenic *PIK3CA* and *TP53* post-zygotic variants in uninvolved mammary tissue of breast cancer patients with recurrent disease

Maria Andreou[1], Katarzyna Chojnowska[1], Natalia Filipowicz[1], Monika Horbacz[1], Piotr Madanecki[2], Katarzyna Duzowska[1], Urszula Ławrynowicz[1], Hanna Davies[3], Bożena Bruhn-Olszewska[3], Mikołaj Koszyński[1], Kinga Dręże k-Chyła[1], Maciej Jaśkiewicz[1], Marcin Jąkalski[1], Anna Kostecka[1], Marta Drzewiecka-Kłysz[4], Magdalena Nowikiewicz[5], Manuela Las-Jankowska[6,7], Dariusz Bała[6,8], Jacek Hoffman[4], Ewa Śrutek[9,10], Michał Jankowski[6,8], Jerzy Jankau[11], Diana Hodorowicz-Zaniewska[12], Joanna Szpor[13], Łukasz Szylberg[10,14], Wojciech Zegarski[6,8], Tomasz Nowikiewicz[4,6], Patrick G. Buckley[15], Irene Tiemann-Boege[16], Jakub Mieczkowski[1], Magdalena Koczkowska[1], Jan P. Dumanski[1,2,3] Arkadiusz Piotrowski[1,2,*.]

ABSTRACT

Normal-appearing mammary tissue from breast cancer patients can harbor significant genetic alterations. We analyzed DNA variants in both normal-looking tissue and tumors from 77 poor-prognosis patients, 49 patients without prognosis bias, and mammary gland samples from 15 non-cancerous individuals. Whole exome sequencing revealed a higher prevalence of pathogenic post-zygotic variants (29%) in poor-prognosis patients, affecting genes like *AKT1*, *PIK3CA*, *PTEN*, *TBX3*, and *TP53*, compared to 12.5% in those without prognosis criteria (p=0.0008578). Prevailing *PIK3CA* variants were recurrent across patients, while *TP53* variants were present only in those with adverse prognoses. Duplex sequencing detected low-frequency pathogenic *PIK3CA* and *TP53* variants in distant normal tissues of poor-prognosis patients. Disease recurrence significantly reduced survival rates, with poor prognosis patients experiencing higher mortality within 24 months (p=0.0088), further worsened by the presence of pathogenic post-zygotic variants. These findings highlight the importance of genetic monitoring even in normal-appearing mammary tissue.

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

A.P., J.P.D, and M. Koczkowska contributed equally.

Keywords: breast cancer, recurrence, post-zygotic variants, unfavorable outcome, uninvolved mammary gland, mortality

Abbreviations

BL: Whole Peripheral Blood

BCAP: Breast Cancer Adverse Prognoses

BCS: Breast-conserving surgery

BCUS: Brest Cancer Un-Selected

PT: Primary Tumor

RM: Reduction Mammoplasty

SK: Skin

UM: Uninvolved mammary gland

UMD: Uninvolved mammary gland at a greater distance from the primary tumor

WES: Whole Exome Sequencing

[1]3P-Medicine Laboratory, Medical University of Gdańsk, Gdańsk, Poland

[2]Department of Biology and Pharmaceutical Botany, Medical University of Gdańsk, Gdańsk, Poland

[3]Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[4]Department of Clinical Breast Cancer and Reconstructive Surgery, Oncology Center-Prof. Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland

[5]Department of Hepatobiliary and General Surgery, A. Jurasz University Hospital, Bydgoszcz, Poland

[6]Chair of Surgical Oncology, Ludwik Rydygier's Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, 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]Surgical Oncology, Ludwik Rydygier's Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland

[10]Department of Tumor Pathology and Pathomorphology, Oncology Center-Prof Franciszek Łukaszczyk Memorial Hospital, Bydgoszcz, Poland

[11]Department of Plastic Surgery, Medical University of Gdańsk, Gdańsk, Poland

[12]Ist Department of General Surgery, Jagiellonian University Medical College, Breast Unit University Hospital, Krakow, Poland

[13]Department of Pathomorphology, Jagiellonian University Medical College, Kraków, Poland

[14]Department of Clinical Pathomorphology, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland

[15]Genuity Science Genomics Centre, Dublin, Ireland

[16]Institute of Biophysics, Johannes Kepler University, Linz, Austria

INTRODUCTION

Breast cancer accounts for 12.5% of global annual cancer diagnoses, with the incidence rate increasing by 0.5% every year in recent years[1,2] Despite an overall 42% reduction between 1989 and 2021, primarily attributed to increased awareness and early detection, breast cancer still constitutes one of the leading causes of death among women[3]. Notably, stage I, low-risk breast cancer cases still present a 15-20% chance of recurrence even two decades after the initial diagnosis[4]. While 5-10% of breast cancer cases are hereditary, with 25-30% of heritable breast cancer risk attributed to pathogenic variants in genes of high and moderate penetrance[5], the majority of cases are considered sporadic[6,7].

In the pursuit of detecting early indications of tumor formation, preceding diagnosis via imaging or palpation, research has increasingly focused on the normal mammary gland within the affected breast[8]. Concurrently, breast-conserving surgery (BCS), which aims to remove the tumor and preserve the remaining healthy breast tissue, stands as the preferred approach[9,10]. However, mammary gland tissue from breast cancer patients, although appearing normal, has been found to harbor significant genomic and transcriptomic alterations[11–14]. In particular, non-tumorous tissue from patients undergoing BCS shows clearly pathogenic low-level post-zygotic alterations in the *PIK3CA* and *TP53* genes, raising questions regarding its oncogenic potential [15,16].

Nevertheless, the association between post-zygotic alterations in ostensibly normal mammary gland tissue of breast cancer patients and their prognostic implications, and possible utility in precision and personalized medicine, remains unclear. Hence, we screened paired uninvolved mammary gland (UM) and primary tumor (PT) samples of reportedly sporadic breast cancer patients with adverse outcomes within 10 years post-original surgery for the presence of post-zygotic alterations. We compared our findings with a second breast cancer cohort of reportedly sporadic patients recruited without specific prognosis criteria. Additionally, normal mammary gland samples from individuals with no personal or family history of cancer were included as controls (study overview in Supplementary Figure 1).

Here, we demonstrate that pathogenic post-zygotic variants in cancer-associated genes are commonly found in seemingly normal mammary tissue of breast cancer patients with adverse prognoses. These variants are

linked with patient survival, emphasizing the need for comprehensive oncological monitoring to enhance the clinical management of patients.

RESULTS

Study cohorts

We carried out Whole Exome sequencing (WES) on 408 samples from three distinct groups: two cohorts of breast cancer patients with differing prognostic outcomes and a control group composed of individuals who underwent mammoplasty surgery for non-cancer-related reasons. The first cohort included 77 reportedly sporadic breast cancer patients characterized by adverse outcomes. All individuals in this group experienced either recurrent disease, such as local recurrence/metastasis to the breast or secondary organs (n=40), developed a second independent tumor (n=18), or both (n=8), and/or succumbed to the disease (n=45) within the following 10 years (Breast Cancer Adverse Prognoses cohort, BCAP), (Table 1, Supplementary Table 1). The second cohort included 49 individuals from the same ethnic population, also diagnosed with reportedly sporadic breast cancer, but recruited without specific criteria related to prognosis (Breast Cancer Un-Selected cohort, BCUS). Within this group, 5 out of 49 patients experienced recurrence, and 3 of them died within 2 years post-surgery; however, the follow-up period for this cohort was considerably shorter compared to the BCAP cohort (Table 1, Supplementary Table 1). The control group comprised 15 individuals who underwent reduction mammoplasty surgeries and had no personal or familial history of cancer (Reduction Mammoplasty cohort, RM).

There was a significant age difference between BCAP, BCUS, and RM cohorts (Kruskal-Wallis test, p=7.7e-05). Patients in the BCAP cohort (median age: 62, range: 23-85) and the BCUS cohort (median age: 65, range: 37-84) did not show a significant age difference (Kruskal–Wallis H test, p=0.082) (Table 1). However, both cohorts were significantly older than the control group (median age: 42, range: 26-73) (Kruskal–Wallis H test, p=0.000034 and p=0.00036 for BCAP and BCUS respectively). We examined sets of UM and PT samples from all breast cancer patients to identify variants related to breast cancer (Figure

1a). Whole peripheral blood (BL) or skin (SK), if blood was not available, served as a reference sample, and was used to distinguish between post-zygotic and germline variants (Figure 1b). UM samples at a further distance (UMD, 1.5-3 cm from PT, median 2.35 cm) from the corresponding PTs, available for 7 BCAP patients, were collected and included in the downstream confirmatory analysis, however, the UMDs were not included in the WES originally. For the RM cohort, sets of UM and BL samples were included.

	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
Uninvolved mammary gland at greater distance, UMD	7	-
Reference sample	77	49
(whole peripheral blood, BL or skin, SK)		
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

Table 1. Summarized clinicopathological features for breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) and the Breast Cancer Un-Selected (BCUS) cohorts. Matched and primary tumor (PT) and uninvolved mammary gland (UM, ≥1 cm) samples were collected from the 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. Additional UM samples located further away (UMD, 1.5-3 cm from PT, median 2.35 cm) from corresponding PTs, available for 7 BCAP patients, were included. The detailed sampling design is described in Materials and Methods. An overview is also available in Figure 1. *Death status refers to patients who succumbed to the disease (patient with ID BCAP61 died from non-oncological reasons). Detailed clinicopathological information for BCAP and BCUS cohorts is provided in Supplementary Table 1.

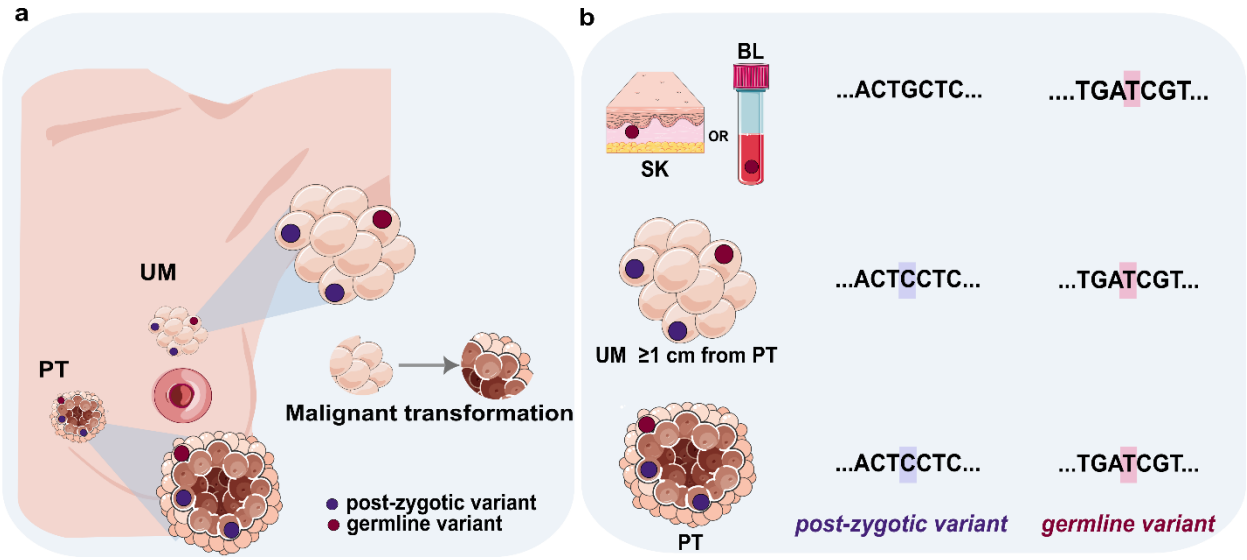

Figure 1. Graphical representation of the sampling design used for variant investigation. a) Sampling design. Pairs of fresh-frozen uninvolved mammary gland (UM) and primary tumor (PT) samples were collected from breast cancer patients with differing prognostic outcomes, i.e. breast cancer patients with adverse prognoses (Breast Cancer Adverse Prognoses, BCAP cohort) and breast cancer patients recruited without specific criteria related to prognoses (Breast Cancer Un-Selected, BCUS cohort) to identify variants according to study criteria (Materials & Methods). UM samples were located at least 1 cm away from the corresponding PT. UM samples were also obtained from individuals who underwent reduction mammoplasty surgeries and had no history of cancer (Reduction Mammoplasty cohort, RM). **b) Investigation of post-zygotic and germline variants.** Matched whole peripheral blood (BL) or skin (SK) samples for every individual were collected as reference samples to distinguish between post-zygotic and germline variants. 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/).

Truncating post-zygotic variants in dosage-sensitive genes predominate in BCAP compared to BCUS and RM cohorts

The details of post-zygotic variants in BCAP, BCUS, and RM cohorts meeting the study's cut-off criteria (Materials & Methods and Supplementary Figure 2) are summarized in Supplementary Table 2. Briefly, variants were filtered based on their annotation within ClinVar or InterVar databases as "pathogenic", "likely pathogenic", "uncertain significance", or "conflicting interpretations of pathogenicity", or their presence in the COSMIC database (cosmic_95coding). Additionally, variants were filtered by their frequency in the general population including only those with a minor allele frequency (MAF) ≤ 0.001 across all gnomAD populations ("popmax") or not noted in the database (gnomAD v2.1.1).

Within the BCAP cohort, 167 distinct variants were identified in UM samples of 41 patients. The corresponding numbers for the BCUS were considerably lower, i.e. 56 variants were identified in the UM samples of 24 patients. The RM cohort presented 10 distinct variants in seven unrelated individuals, but all were either variants of uncertain significance or not reported in the ClinVar/InterVar databases. Variants were further categorized based on their impact on the encoded protein i.e. truncating and non-truncating. Notably, truncating nonsense (n=25) and frameshift (n=12) variants, were exclusive to the BCAP cohort (Supplementary Table 3). These alterations, responsible for transcript elimination via nonsense-mediated mRNA decay[17] were deemed pathogenic, irrespective of their annotation in the ClinVar and InterVar databases. For missense variants in BCAP, BCUS, and RM UM samples, *in-silico* analyses were performed using multiple predictors (Supplementary Table 4).

In the BCAP cohort, 29% (49/167) of the identified variants were deemed pathogenic. This list included truncating variants (n=37), missense variants annotated as pathogenic/likely pathogenic (n=8), and missense variants reported as uncertain significance or pathogenic/likely pathogenic in ClinVar, lacking assertion criteria but showing evidence of pathogenicity according to in-silico analyses (n=4) (Supplementary Tables 3 and 4). In-silico analyses were conducted using the REVEL tool[18], with a threshold set to 0.75 (accessed between June 2022 and August 2022 via ANNOVAR software). In comparison, the BCUS cohort had seven pathogenic variants, representing 12.5% of the total identified variants (n=56). These included variants reported as pathogenic (n=4), and missense variants classified as uncertain significance or pathogenic/likely pathogenic in ClinVar without assertion criteria, showing evidence of pathogenicity according to REVEL (n=3) (Supplementary Table 4). The UM samples from the BCAP cohort were significantly enriched for pathogenic post-zygotic variants (a combination of truncating variants and those classified as pathogenic by ClinVar/REVEL) compared to those from the BCUS cohort (Hypergeometric test, p=0.0008578).

Focusing on post-zygotic variants potentially linked to breast cancer, we concentrated on pathogenic variants as defined by the ClinVar database. Variants annotated as pathogenic or likely pathogenic in

ClinVar, reviewed by expert panels, or submitted by multiple parties without discordance, are listed in Supplementary Table 5 for both BCAP and BCUS cohorts. Missense variants classified as uncertain significance, pathogenic/likely pathogenic without assertion criteria, or with conflicting interpretations in ClinVar with a REVEL score ≥ 0.75 were also included. Truncating variants in known tumor-suppressor genes implicated in breast cancer, such as *KMT2C*[19], *TBX3*[20], and *TP53*[21] were included in the same table even if they were absent from the ClinVar database (Supplementary Table 5).

Several identified variants affected dosage-sensitive genes. These included deleterious variants in the tumor suppressors *KMT2C*[19], *PTEN*[22], *PTCH1*[23], *TBX3*[20], and *TP53*[21] as well as activating variants in oncogenes *AKT1*[24] and *PIK3CA*[25] identified in BCAP UM samples (Supplementary Table 5). Oncogenes such as *SF3B1*[26], *HRAS*[27], and *GNAS*[28], and genes with a dual role in cancer (*RUNX1*[29]) were affected solely in the BCUS cohort, with only the latter two being dosage-sensitive (Supplementary Table 5). *PIK3CA* was the only gene recurrently affected in both BCAP and BCUS cohorts.

Recurrence, coexistence, spatial distribution in the breast, and validation of variants

In the BCAP cohort, pathogenic variants in two driver genes, *PIK3CA* and *TP53*, were predominant across all subtypes of invasive cancer. *PIK3CA*, which encodes the catalytically active p100alpha isoform, is a key regulator of cell proliferation and growth receptor signaling cascades[30]. We detected three distinct pathogenic post-zygotic *PIK3CA* variants in UM samples: c.1624G>A (p.Glu542Lys), c.3140A>G (p.His1047Arg), and c.3140A>T (p.His1047Leu), found in two, four and one unrelated individuals, respectively (Supplementary Table 5). Another *PIK3CA* variant, c.3012G>T (p.Met1004Ile), was found in the UM of a single BCAP cohort individual. However, this variant reported only once as of uncertain significance in the ClinVar database and with a REVEL score of 0.437 suggesting it might be benign, was classified as a variant of uncertain significance due to limited evidence (Supplementary Table 4). *PIK3CA* c.3140A>G (p.His1047Arg) and c.3140A>T (p.His1047Leu) variants, located at the commonly mutated *PIK3CA* site in breast cancer, were also observed in UMs of the BCUS cohort (Figure 2, Supplementary Table 4).

TP53 is the most commonly mutated gene in various human cancers and its wild-type protein's functions are frequently compromised in many types of malignancies[31]. We detected seven *TP53* variants in the normal mammary gland samples of the BCAP cohort, including two recurrent variants: six pathogenic or likely pathogenic (c.151G>T [p.Glu51*], c.329G>C [p.Arg110His], c.637C>T [p.Arg213*], c.711G>A [p.Met237Ile], c.1024C>T [p.Arg342*], and c.1025G>C [p.Arg342Pro]), and one frameshift variant c.227del (p.Ala76Aspfs*47), not previously reported in the ClinVar database (Supplementary Table 5). Importantly, no *TP53* variants were observed in the normal mammary gland samples of the BCUS cohort (Figure 2).

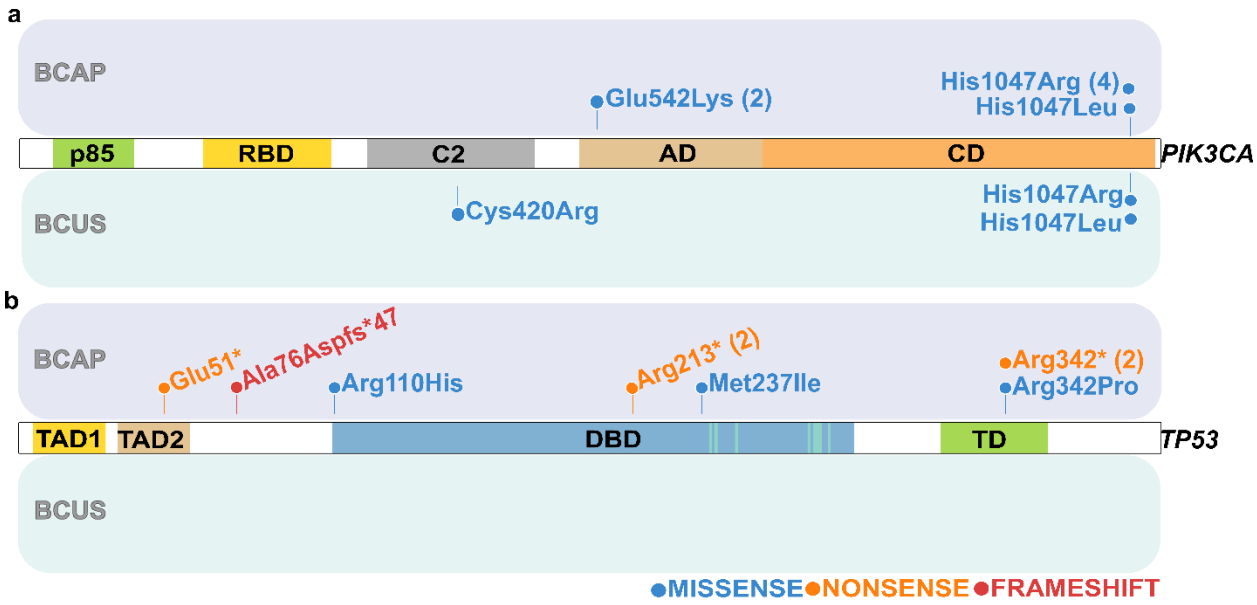

Figure 2. Post-zygotic (a) *PIK3CA* and b) *TP53* variants detected in the uninvolved mammary gland (UM) and primary tumor (PT) samples of breast cancer patients with adverse prognoses and patients recruited without any prognoses bias (BCAP and BCUS cohorts respectively). Lollipop plots represent post-zygotic variants of *PIK3CA* and *TP53* genes detected by Whole Exome Sequencing (WES). The upper panel represents variants detected in UM samples of BCAP patients. The lower panel represents variants detected in UM samples of BCUS patients. *TP53* variants were found exclusively in the BCAP cohort. All identified variants have been reported in the COSMIC database (https://cancer.sanger.ac.uk/cosmic). Detailed descriptions of variants can be found in Supplementary Table 5. 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 green lines, TD - tetramerization domain. Lollipop plots were prepared based on the images generated with the Protein paint application[32]. () indicates the number of patients in whom the variant was identified.

Pathogenic variants in *AKT1*, *PIK3CA*, *PTEN*, *TBX3*, and *TP53* in 16 BCAP patients were selected for validation with independent methods (Table 2, Supplementary Figure 2). The presence of *PIK3CA* variants c.3140A>G (p.His1047Arg) (in three patients), c.1624G>A (p.Glu542Lys) (in two patients) and *TP53* c.711G>A (p.Met237Ile) (in a single patient) was confirmed by Sanger sequencing or High-Resolution Melting in the UM samples of six BCAP patients (Table 2, Supplementary Table 6, Supplementary Figure 3). Notably, the *PIK3CA* c.3140A>G (p.His1047Arg) and *TP53* c.711G>A (p.Met237Ile) variants co-existed in the UM sample of patient BCAP15. Additionally, two pathogenic variants, c.49G>A (p.Glu17Lys) in *AKT1* and c.388C>T (p.Arg130*) in *PTEN*, were identified in the UM samples of three BCAP patients and subsequently confirmed via Sanger sequencing or High-Resolution Melting (Table 2, Supplementary Table 6, Supplementary Figure 3). The presence of the *AKT1* variant was confirmed in the UM samples of two patients and the UMD of one of them, indicating a broad spatial distribution of this variant in that particular patient. Finally, the *TBX3* c.371_372insTGGT (p.Ile125Profs*14) variant was confirmed in the UM sample of a single patient (Table 2, Supplementary Table 6, Supplementary Figure 3).

We further employed duplex sequencing to validate the presence of extremely low-frequency *PIK3CA* and *TP53* variants and examine their tissue/spatial distribution. We selected 11 individuals: BCAP01, BCAP31, BCAP36, BCAP38, BCAP45, BCAP47, BCAP48, BCAP53, BCAP54, BCAP57, and BCAP58 based on the presence of *PIK3CA* and *TP53* variants in proximal UM samples from WES data. UMD samples available for 6 BCAP patients, located at a greater distance from the PTs and not included in the original WES run, were also analyzed (Table 2). Ultra-deep targeted duplex sequencing with a mean coverage of 4789x confirmed the following low-frequency (down to 1.34%) pathogenic variants in *PIK3CA*: c.3140A>G (p.His1047Arg) and c.3140A>T (p.His1047Leu) in the UM samples from two patients, and revealed the presence of c.3140A>G (p. His1047Arg) variant in the UMD samples of another two patients (Table 2, Supplementary Table 7). Furthermore, duplex sequencing confirmed the presence of six *TP53* variants (c.151G>T [p.Glu51*], c.227del [p.Ala76Aspfs*47], c.329G>C [p.Arg110His], c.637C>T

[p.Arg213*], c.1024C>T [p.Arg342*], and c.1025G>C [p.Arg342Pro]) in UM tissues of eight patients, and revealed the presence of c.151G>T (p.Glu51*), c.227del (p.Ala76Aspfs*47), and c.637C>T (p.Arg213*) variants in the paired UMD tissues of four patients (Table 2, Supplementary Table 7). An overview of validated variants in *AKT1*, *PIK3CA*, *PTEN*, *TBX3*, and *TP53* genes along with follow-up information for the corresponding patients is provided in Figure 3.

Gene	Variant[a]	ClinVar[b]	COSMIC ID[c]	AVSNP150[d]	Individual ID	Confirmation
AKT1	c.49G>A (p.Glu17Lys)	Pathogenic	ID=COSV62571334	rs121434592	BCAP32, BCAP66*	SS / HRM
PIK3CA	c.1624G>A (p.Glu542Lys)	Pathogenic	ID=COSV55873227	rs121913273	BCAP56, BCAP45	SS / HRM
PIK3CA	c.3140A>G (p.His1047Arg)	Pathogenic	ID=COSV55873195	rs121913279	BCAP15, BCAP31*, BCAP36, BCAP53*	SS / HRM / DS
PIK3CA	c.3140A>T (p.His1047Leu)	Pathogenic	ID=COSV55873401	rs121913279	BCAP54*	DS
PTEN	c.388C>T (p.Arg130*)	Pathogenic	ID=COSV64288463	rs121909224	BCAP15	SS / HRM
TBX3	c.371_372insTGGT (p.Ile125Profs*14)	n.a.	ID=COSV57471668	n.a.	BCAP44	SS / HRM
TP53	c.151G>T (p.Glu51*)	Pathogenic	ID=COSV52694020	n.a.	BCAP58*	SS / HRM
TP53	c.227del (p.Ala76Aspfs*47)	n.a.	ID=COSV52728465	n.a.	BCAP77	DS
TP53	c.329G>C (p.Arg110His)	Pathogenic	ID=COSV52668419	rs11540654	BCAP45	DS
TP53	c.637C>T (p.Arg213*)	Pathogenic	ID=COSV52665560	rs397516436	BCAP01*, BCAP48*	DS
TP53	c.711G>A (p.Met237Ile)	Pathogenic	ID=COSV52661887	rs587782664	BCAP15	SS / HRM
TP53	c.1024C>T (p.Arg342*)	Pathogenic	ID=COSV52665487	rs730882029	BCAP38, BCAP47	DS
TP53	c.1025G>C (p.Arg342Pro)	Pathogenic/Likely pathogenic	ID=COSV52690857	rs375338359	BCAP57	DS

Table 2. Pathogenic post-zygotic variants within the uninvolved mammary gland (UM) samples of breast cancer patients with adverse prognoses (BCAP cohort), selected for further investigation/validation. Presented variants, identified via Whole Exome sequencing, were corroborated with either Sanger sequencing/High-Resolution Melting, or Duplex sequencing. [a]Variant annotation provided for the basic isoform of the transcript. [b]Pathogenicity classification according to the ClinVar database. [c]ID of the variant in the COSMIC (Cosmic_95 coding) database. [d]rsIDs in dbSNP build 150. Detailed description of selected post-zygotic variants is provided in Supplementary Table 5. Confirmation of post-zygotic variants by Sanger sequencing/High-Resolution Melting, or Duplex sequencing is provided in Supplementary Tables 6 and 7 respectively, and Supplementary Figure 3. SS – Sanger sequencing. HRM – High-Resolution Melting. DS – Duplex sequencing. n.a.- not available. *variants were also detected in the uninvolved mammary gland sample located at a greater distance (UMD) from the primary tumor of selected patients.

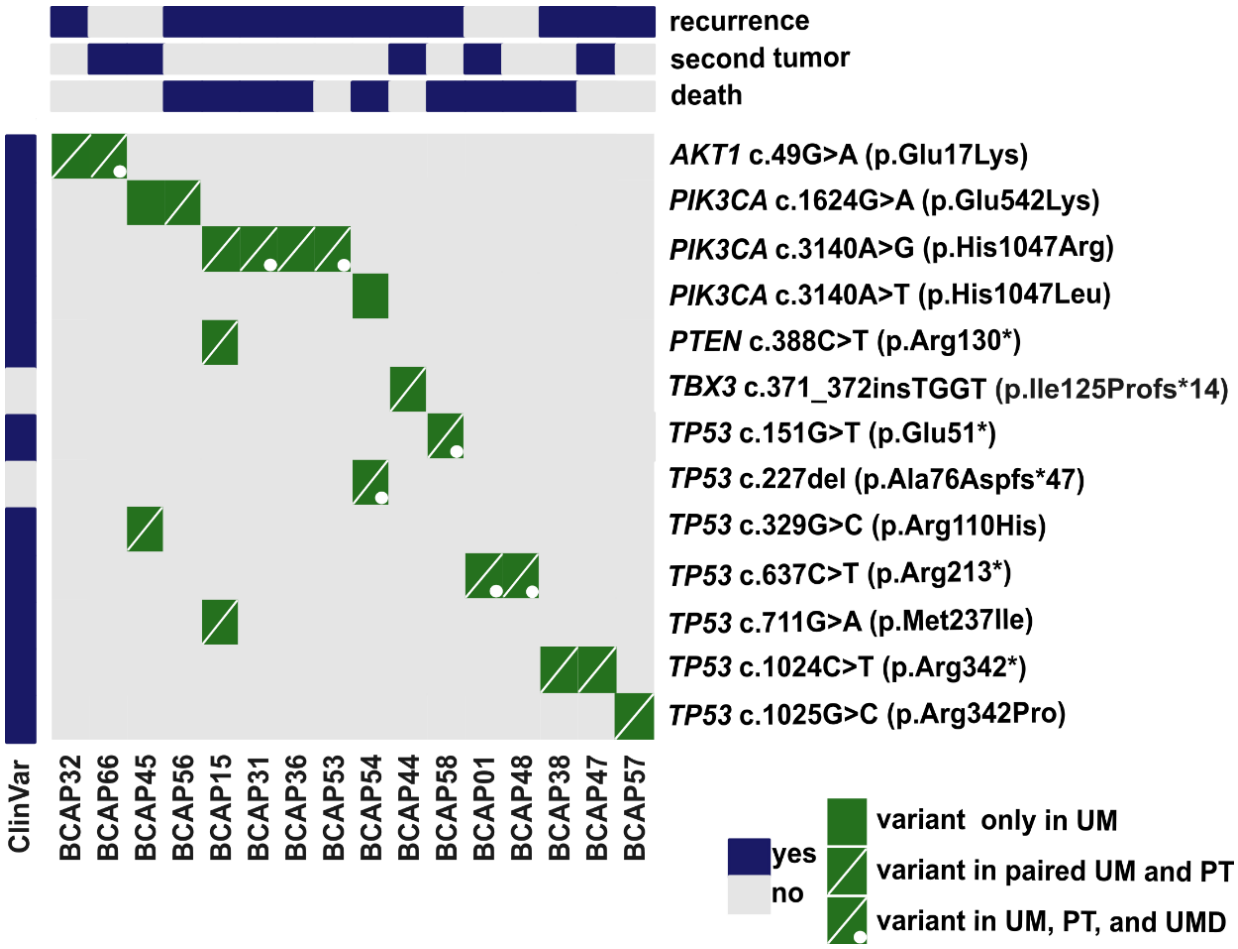

Figure 3. Pathogenic post-zygotic variants in breast cancer patients with adverse prognoses. Variants were confirmed via Sanger sequencing/High-Resolution Melting or Duplex sequencing (Materials and Methods, Table 2, Supplementary Tables 6 and 7, Supplementary Figure 3). Variant presence in the ClinVar database and follow-up information for the corresponding patients are included. Detailed clinicopathological information for presented patients is provided in Supplementary Table 1. A full description of detected variants is provided in Supplementary Table 5. PT – primary tumor. UMD – uninvolved mammary gland at a further distance from corresponding PT.

Spectrum of germline pathogenic variants in the two breast cancer cohorts

All breast cancer cases included in our study were reported as sporadic, however, molecular testing results were not available prior to recruitment. We analyzed BL or SK samples from each participant to screen for pathogenic or likely pathogenic germline variants across all cohorts (Materials & Methods, Supplementary Figure 2). In the BCAP cohort, 14 of 77 individuals (18%) carried germline pathogenic variants in known breast cancer-associated genes[33]. These included c.4186C>T (p.Gln1396*), c.4689C>G (p.Tyr1563*),

c.5179A>T (p.Lys1727*), and c.5266dup (p.Gln1756Profs*74) in *BRCA1*, c.5645C>A (p.Ser1882*), c.6591_6592del (p.Glu2198Asnfs*4), and c.9382C>T (p.Arg3128*) in *BRCA2*, c.172_175del (p.Gln60Argfs*7) and c.1671_1674del (p.Ile558Lysfs*2) in *PALB2*, and c.3233_3236del (p.Lys1079Valfs*28) in *RAD50*. Only *BRCA1* c.5266dup (p.Gln1756Profs*74) and *PALB2* c.172_175del (p.Gln60Argfs*7) were recurrent, observed in four and two unrelated individuals respectively. This incidence surpasses the rates from other studies where up to approximately 10% of reportedly sporadic cases turn out hereditary after molecular testing and likely reflects the aggressive outcomes in these cases[15,21,34]. Four individuals (4/14, 29%) with germline pathogenic variants also carried pathogenic post-zygotic variants in known, curated cancer-related genes[35] in their UM samples. The germline variants in these cases were found in *BRCA1* (four cases) and *RAD50* (one case). The corresponding post-zygotic variants were identified in *PIK3CA* or *TP53*. *BRCA1* c.5266dup (p.Gln1756Profs*74) was the only variant observed in a single patient from the BCUS cohort. In the control group, no individuals were found to carry germline pathogenic or likely pathogenic variants in genes associated with breast cancer. Pathogenic germline variants in high- and moderate-penetrance breast cancer susceptibility genes identified in the BCAP and BCUS cohorts are detailed in Supplementary Table 8.

Pathogenic post-zygotic variants in patients with recurrent disease affect the survival rate

We used Kaplan-Meier plots to evaluate survival probabilities and compare patients with recurrence (n=53) to those without recurrence (n=72) in both the BCAP and BCUS cohorts. Overall, patients with recurrence had significantly lower survival probabilities (log-rank test, p=0.00017) (Supplementary Figure 4a), with a hazard ratio of 2.44 (95% CI: 1.07-5.54, p=0.0337), indicating more than twice the risk of death compared to the non-recurrence group.

Due to the shorter follow-up period for the BCUS cohort (2 years) compared to the BCAP cohort (10 years), we focused on the first 24 months post-diagnosis. During this period, recurrence patients in both cohorts (n=53) had significantly lower survival probabilities compared to non-recurrence patients (n=71) (log-rank test, p<0.0001) (Supplementary Figure 4b), with a hazard ratio of 4.85 (95% CI: 1.4-16.25, p=0.0105),

indicating more than four times the risk of death for the recurrence group. Within the first 24 months, BCAP patients experienced significantly more recurrence events than BCUS patients (Fisher's exact test, p=0.005488). In the BCAP cohort, patients with recurrence (n=48) had lower survival probabilities throughout the follow-up period compared to those without recurrence (n=28) (log-rank test, p=0.015). This pattern was also observed in the first 24 months (n=48 vs. n=27) (log-rank test, p=0.0088) (Supplementary Figures 4c and 4d)

Additionally, we assessed the impact of pathogenic post-zygotic and germline variants on the survival of patients with recurrent disease. Figure 4 displays Kaplan-Meier curves based on the presence and type of pathogenic variants. Survival probabilities differ significantly across studied groups (log-rank test, p=0.024), indicating that the combination of pathogenic post-zygotic variants, pathogenic germline variants, and recurrence status significantly impact patient survival. Patients with pathogenic germline variants (green) had the shortest recurrence-free survival, with most recurrences occurring within the first 60 months. Notably, patients with pathogenic post-zygotic variants in breast cancer-specific genes (blue) also experienced recurrences, though less frequently and over a longer follow-up period, highlighting the impact of these variants on recurrence risk, albeit to a lesser extent than germline variants. Patients without pathogenic germline or postyzygotic variants (yellow) showed intermediate outcomes. In conclusion, while pathogenic germline variants have the most pronounced effect on survival rates, pathogenic post-zygotic variants seem to also play a notable role in influencing recurrence risk and patient outcomes, highlighting their importance in understanding disease progression.

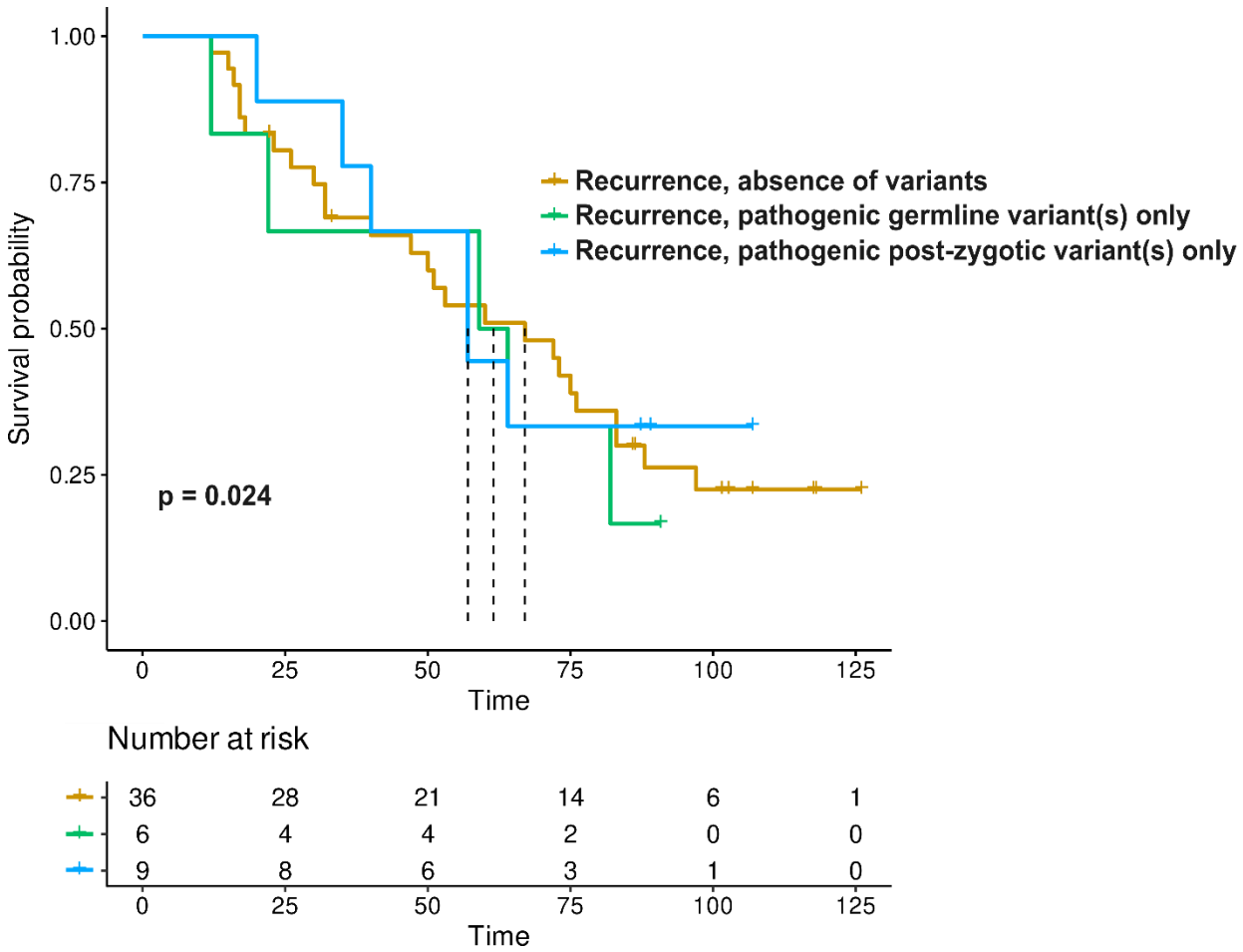

Figure 4. Kaplan-Meier survival curves of breast cancer patients with pathogenic variants and recurrent disease. The curves represent survival probabilities for different groups of patients from the BCAP cohort (breast cancer patients with adverse prognoses) and the BCUS cohort (breast cancer patients 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). Detailed information on recurrence status and patient outcomes can be found in Supplementary Table 1. 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.

DISCUSSION

The early detection and treatment of breast cancer, including its precursors, have shifted research focus from tumors to the normal mammary gland to deepen our understanding of the disease's origins[36]. Genetic and transcriptomic studies have revealed an abundance of pathogenic point variants and other alterations in

critical breast cancer driver genes within the normal mammary gland of patients who have undergone BCS or mastectomy, compared to control tissues[11–15,37].

On the premise that histological normality does not necessarily equate to biological normality, we screened for post-zygotic alterations in two breast cancer cohorts (BCAP and BCUS), which were similar in age (Kruskal–Wallis H test, p=0.082) but differed in survival outcomes. We also included a control group of individuals treated surgically for non-cancer reasons (RM cohort), who were significantly younger than both BCAP and BCUS patients (Kruskal–Wallis H test, p=0.000034 and p=0.00036 respectively). Truncating nonsense and frameshift variants were found exclusively in the UM samples of patients with adverse prognoses (BCAP cohort), while BCUS patients and RM controls showed only missense variants (Supplementary Tables 3 and 4). The UM samples of BCAP patients were significantly enriched for pathogenic post-zygotic variants compared to BCUS patients (Hypergeometric test, p=0.0008578). Several of these variants affected known, curated cancer-related genes[35] i.e. *AKT1*, *KMT2C*, *PIK3CA*, *PTCH1*, *PTEN*, *TBX3*, and *TP53*. These findings suggest that the presence of pathogenic post-zygotic alterations in UM samples could signal a higher risk for aggressive cancer, even before clinical symptoms manifest.

The high prevalence of *PIK3CA* and *TP53* variants in BCAP patients' UM samples highlights the critical roles these genes play in oncogenesis. The *PIK3CA* gene encodes the p110α catalytic subunit of phosphoinositide 3-kinase (PI3K), a key regulator of the PI3K/AKT signaling pathway, essential for cellular growth, proliferation, and survival and has a known oncogene function[30,34]. Missense variants in *PIK3CA*, particularly in the accessory domain (p.Glu542Lys) and the catalytic domain ([p.His1047Arg], [p.His1047Leu]), activate or modify the encoded protein, enhancing its kinase activity (gain-of-function) and driving oncogenic signaling. The recurrence of the p.His1047Arg variant, critical for PIK3CA function, underscores its significance in tumorigenesis. *PIK3CA* variants were common to both BCAP and BCUS cohorts suggesting a fundamental role in breast cancer development. In contrast, *TP53* variants were found exclusively in the BCAP cohort, underscoring the gene's crucial role in maintaining genomic integrity. *TP53* encodes the tumor protein p53, a key tumor suppressor involved in DNA repair, apoptosis, and cell

cycle regulation[21]. Loss-of-function variants in *TP53* can inactivate its tumor-suppressing activity during oncogenesis[38–40]. The observed *TP53* variants included truncating variants (nonsense and frameshift) as well as missense alterations affecting the DNA-binding domain ([p.Arg110His], [p.Met237Ile]) and the tetramerization motif (p.Arg342Pro), highlighting the various ways p53 function can be disrupted, potentially leading to malignancy. The exclusive presence of *TP53* variants in the BCAP cohort may suggest a more aggressive disease phenotype and a poorer prognosis associated with these variants.

The co-occurrence of *PIK3CA* and *TP53* variants in three BCAP patients (BCAP15, BCAP45, BCAP54) suggests a synergistic effect in driving cancer progression. In particular, patient BCAP15 exhibited concurrent pathogenic variants in *PIK3CA* (p.His1047Arg), *TP53* (p.Met237Ile), and *PTEN* (p.Arg130*). *PTEN*, another critical tumor suppressor, negatively regulates the PI3K/AKT pathway[22]. The presence of alterations in all three genes within a single patient emphasizes the complex interplay of multiple oncogenic and tumor-suppressive pathways in breast cancer pathogenesis. This confluence of variants likely contributes to a more aggressive clinical course, underscoring the importance of comprehensive genetic profiling in understanding individual tumor biology.

While post-zygotic variants in *TP53* or *PIK3CA* have been observed in breast tumors, their consequences in normal mammary tissue are less clear. Some studies suggest a benign effect[41,42]. Healthy breast tissue accumulates alterations with age at an accelerating rate[43] and is influenced by hormonal stimuli, undergoing cycles of expansion during puberty, pregnancy, and lactation[44]. Estrogen and its metabolites can cause DNA damage, increasing cellular stress and the risk of genetic alterations and cancer[44]. However, the accumulation of alterations alone doesn't cause cancer; observed tissue-specific patterns and the "ground state" theory suggest that quiescent stem cells with oncogenic variants rarely transform unless activated by developmental, aging, or injury factors, that normally resemble physiological mammary gland conditions[45]. The presence of such variants in ostensibly normal mammary gland tissue suggests they may represent early, pre-cancerous changes. Despite the mammary gland's inherent multi-layer protection system against

clonal expansions, surviving mutant clones can lead to large fields of mutated cells, i.e. field cancerization, thereby increasing cancer risk[46].

A subset of the BCAP cohort (18%, n=14/77) and one patient from the BCUS cohort (2%, n=1/49) carried germline pathogenic variants in breast cancer genes (Supplementary Table 8). Among the BCAP patients with these variants, four had concurrent pathogenic post-zygotic variants in curated cancer genes, whereas 14 had only post-zygotic alterations in curated cancer genes and genes implicated in breast cancer (Supplementary Tables 5 and 8). Despite their differing genetic profiles, all BCAP patients experienced adverse outcomes within ten years of surgery highlighting the impact of these genetic variations on prognosis. The interaction between germline and post-zygotic variants remains unclear, as recent research indicates that the influence of germline variants on tumor behavior can vary significantly based on factors such as penetrance and lineage, with some variants exhibiting minimal or transient effects on tumor development[47].

In the first 24 months post-diagnosis, the BCAP cohort had significantly more recurrence events than the BCUS cohort (Fisher's exact test, $p = 0.005488$). Recurrence was associated with much lower survival probabilities across both cohorts. While pathogenic post-zygotic variants alone did not drastically alter survival rates, their impact became more severe when combined with disease recurrence, suggesting a link to disease aggressiveness. This highlights the importance of comprehensive genetic screening and vigilant monitoring of patients with pathogenic post-zygotic variants in disease-related genes to improve outcomes.

Current diagnostics primarily focus on identifying germline pathogenic variants in known breast cancer-associated genes to assess breast cancer risk and guide personalized therapy[33]. However, over 80% of breast tumors do not have inherited alterations[9]. Our study reveals that pathogenic post-zygotic variants, such as *PIK3CA* and *TP53* alterations, are often found in seemingly normal mammary gland tissue that is not removed during BCS. In some instances, these variants represent distinct clonal populations not present in the corresponding primary tumors (Supplementary Table 5), indicating the independent evolution of cell lineages within the mammary tissue and suggesting they are unlikely to be micrometastases. The timing of

their emergence relative to the primary tumor remains uncertain - whether during early tumor progression or later—remains uncertain, but in both scenarios, they may contribute to recurrence in the breast or metastasis to other organs[48]. Understanding these dynamics is essential for improving diagnostic approaches and tailoring effective treatments.

Our study comes with certain limitations, particularly regarding the notable age differences between the breast cancer cohorts and individuals subjected to reduction mammoplasty surgeries. Recruiting age-matched control individuals poses a challenge, as those opting for cosmetic surgical treatments are typically younger. The incidence of breast cancer diagnosis among younger women is relatively infrequent with only about one out of eight invasive breast cancers being diagnosed in women under the age of 45[1]. Another limitation arises in recruiting healthy control individuals, given that approximately 13% of women are expected to develop invasive breast cancer during their lifetime and the precise onset of carcinogenesis remains unclear[1,2]. Here, control normal mammary glands were sampled from individuals without a personal or familial history of cancer undergoing plastic surgery. Hence, these samples represent the most appropriate, available control samples from a biological standpoint.

Our findings show that post-zygotic genetic variants in genes related to breast cancer are considerably more frequent in normal mammary tissues from breast cancer patients with adverse outcomes compared to the general population of breast cancer patients and controls. These alterations are associated with disease progression, i.e. recurrence, and suggest an elevated risk of aggressive cancer before clinical symptoms emerge, underscoring the need for expanded genetic screening and enhanced surveillance as part of personalized breast cancer management.

MATERIALS & METHODS

Patient recruitment, sample collection, and DNA isolation

We examined two breast cancer cohorts with differing outcomes and a cohort of individuals without a personal or familial history of cancer, serving as controls. Breast cancer patients were recruited either based

on their unfavorable outcomes, with extensive follow-up collected for 10 years post the original surgery (BCAP cohort, n=77), or without any criteria related to prognoses, and follow-up information collected for approximately 2 following years post-original diagnosis (BCUS cohort, n=49). The majority of BCAP and BCUS patients were treated with BCS (n=63 and n=31 respectively) versus mastectomy (n=12 and n=18 respectively) (Supplementary Table 1) (data missing for 2 BCAP patients). All recruited individuals did not receive neoadjuvant therapy. The control cohort consisted of 15 individuals without a history of cancer who underwent reduction mammoplasty surgeries. Samples from all three cohorts were collected by the Oncology Centre in Bydgoszcz, Jagiellonian University Hospital in Cracow, and the University Clinical Centre in Gdańsk, with the necessary ethical approvals and written informed consent from participants. A total of 238 and 147 samples, including PT, UM, UMD, BL, or SK, were collected from the BCAP and BCUS cohorts respectively, and deposited in the biobank of our unit at the Medical University of Gdańsk, along with clinical data, including follow-up information (Table 1, Supplementary Table 1). UM and BL samples were also collected from all RM individuals. UM samples were located at least 1 cm away from the corresponding PT, while UMD samples, available for 7 BCAP patients were located at a greater distance from corresponding PT (median: 2.35 cm, range: 1.5-3 cm,) and were not initially sequenced. All collected samples were frozen at -80°C. Detailed tissue-collecting protocols were previously described by Filipowicz et al[49]. All fragments prepared for molecular analysis were histologically evaluated by expert pathologists to identify tumor fragments (PT) and confirm the normal histology of UM and SK samples. DNA isolation from tissue lysates and whole blood was performed as previously described[49].

Whole exome sequencing, data analysis, variant detection, and validation with independent methods

WES analyses were performed using the Agilent SureSelectXT Human All Exon V7 capture kit for sequencing library construction followed by 150 bp paired-end sequencing on the HiSeq Illumina platform (Illumina, San Diego, CA), outsourced to Macrogen Europe (Amsterdam, The Netherlands). Sequencing coverage was 200x on average, with at least 100x on target.

FASTQ files were inspected and processed with *Trim Galore!* (v0.6.7) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove Illumina-specific adapter sequences and poor-quality reads when necessary. After converting FASTQ files to BAM format and extracting read groups from the raw data, reads were processed using GATK4 Best Practices (v4.0) (https://github.com/gatk-workflows/seq-format-conversion, https://github.com/gatk-workflows/gatk4-data-processing). The reads were mapped to the human genome (hg38) using the BWA-MEM tool (http://bio-bwa.sourceforge.net). Octopus (v0.7.4), in cancer mode, was used for variant calling. The cancer calling model can jointly genotype multiple samples from the same individual, using a reference sample (whole peripheral blood or skin when blood was not available) to distinguish between post-zygotic and germline variants (https://luntergroup.github.io/octopus/). Framesift insertions/deletions, nonsense, and missense variants located in exons were included in the analysis. A random forest filtering approach was implemented to minimize false calls. Variants in reads with poor mapping quality (<30), and variants supported by high-quality bases (≥30) in fewer than five reads were excluded from the analysis. Variants were annotated using ANNOVAR (https://annovar.openbioinformatics.org/) (last updated on 07.07.2020 and accessed between 06.2022-08.2022) and wANNOVAR (https://wannovar.wglab.org/) (last accessed on 21.05.2024), using the MANE SELECT transcript for investigated genes (https://www.ensembl.org/info/genome/genebuild/mane.html). A brief overview of filtering strategies implemented for identifying post-zygotic and germline variants related to breast cancer, within BCAP, BCUS, and RM cohorts and validation experiments for selected post-zygotic variants of BCAP patients is available in Supplementary Figure 2.

Post-zygotic variants

Only variants with sequencing depth ≥ 50 and tissue allele frequency ≥ 0.03 were included in the analysis. Variants were filtered based on their annotation in ClinVar and InterVar databases; variants reported as "pathogenic", "likely pathogenic", "uncertain significance", or "conflicting interpretations of pathogenicity" were included. In parallel, variants described in the COSMIC database (Cosmic_95coding)

were incorporated. Missense variants documented in the COSMIC database, but annotated as "benign" or "likely benign" in ClinVar or InterVar databases, were excluded. Variants in genomic regions with known read-through transcription between adjacent genes, or those spanning multiple genes (e.g., *P2RY11; PPAN-P2RY11* and *KIR2DL1; KIR2DS5; LOC112267881*), were also excluded.

The remaining variants were filtered by their frequency in the general population, retaining only those with a minor allele frequency (MAF) ≤ 0.001 across all gnomAD populations ("popmax") or not listed in gnomAD (v2.1.1) (Supplementary Table 2). Variants were further categorized as truncating (Supplementary Table 3) or missense (Supplementary Table 4). Truncating variants were considered pathogenic, regardless of their annotation in ClinVar or InterVar. For missense variants, in-silico analyses using the REVEL tool[18] (with a threshold score of 0.75) were performed. In summary, truncating variants, variants classified as "pathogenic" or "likely pathogenic" in ClinVar, and missense variants with conflicting interpretations or annotated as "pathogenic" in ClinVar without assertion criteria, but having a REVEL score ≥ 0.75 were deemed pathogenic.

Variants affecting breast cancer-related genes were prioritized for further investigation based primarily on their ClinVar annotation. Variants annotated as "pathogenic" or "likely pathogenic" in ClinVar and missense variants reported in ClinVar as variants of uncertain significance or with conflicting interpretations (or without assertion criteria provided), with a REVEL score of ≥ 0.75 were included. Truncating variants affecting known tumor suppressor genes (i.e. *KMT2C*[35], *TBX3*[20], and *TP53*[21]) were also incorporated even if absent in ClinVar (Supplementary Table 5). UM, UMD, PT, and SK samples of 16 BCAP patients were subjected to further investigation (Sanger sequencing/High-Resolution Melting, Duplex sequencing) to verify or explore the presence of selected variants (Table 2, Supplementary Tables 6 and 7, Supplementary Figure 3).

Germline variants

For germline variant detection, only variants in high- and moderate-penetrance breast cancer susceptibility genes were included in the study, as defined by the NCCN Clinical Practice Guidelines in Oncology[33] (Version 1.2023, September 7, 2022). The list of genes of interest includes: *ATM* (MIM *607585), *BRCA1* (MIM *113705), *BRCA2* (MIM *600185), *BARD1* (MIM *601593), *BRIP1* (MIM *605882), *CHEK2* (MIM *604373), *CDH1* (MIM *192090), *PALB2* (MIM *610355), *PTEN* (MIM *601728), *TP53* (MIM *191170), *NF1* (MIM *613113), *STK11* (MIM *602216), *RAD50* (MIM *604040), *RAD51C* (MIM *602774), *RAD51D* (MIM *602954) and additionally *PIK3CA* (MIM *171834). Variants were filtered based on their frequency in the general population: variants with minor allele frequency (MAF) ≤ 0.01 across all gnomAD populations ("popmax") or not noted in the database (gnomAD v2.1.1) were included. Evidence according to the American College of Medical Genetics and Genomics and the Association for Molecular Pathology recommendations[50] was included to describe all germline pathogenic variants. Specifically, the evaluation of identified *BRCA1* and *BRCA2* variants was performed according to the Evidence-based Network for the Interpretation of Germline Mutation Alleles (ENIGMA) *BRCA1* and *BRCA2* Variant Curation Expert Panel (Version 1.1.0) (Clinical Genome Resource, https://www.clinicalgenome.org/affiliation/50087/, https://cspec.genome.network/cspec/ui/svi/doc/GN092, https://cspec.genome.network/cspec/ui/svi/doc/GN097). Pathogenic germline variants meeting the study's criteria, identified within breast cancer patients of the BCAP and BCUS cohorts are described in Supplementary Table 8.

Duplex sequencing

UM samples from 11 BCAP patients were selected to investigate the presence of low-frequency *PIK3CA* and *TP53* variants beyond the detection limits of Sanger sequencing and High-Resolution Melting and a single, higher-frequency *TP53* variant, i.e. c.151G>T (p.Glu51*), located in a difficult GC-rich region. (Table 2, Supplementary Table 7). Additionally, UMD samples, available for 6 of those patients, were

included to explore further the distribution of selected variants in a more distant seemingly normal mammary tissue. Duplex sequencing was performed as previously described[15,51].

Duplex sequencing data analysis

Raw duplex sequencing data were analyzed using the Snakemake-based Duplex-seq-Pipeline (v1.1.4) (https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline) as previously described[52].

Statistical analysis

All statistical analyses were carried out with in-house developed scripts using R studio version 4.1.2 (2021-11-01). Packages *pheatmap* (version 1.0.12) and *ggplot2* (version 3.4.1) were used for plotting. Statistical significance of differences between two or multiple groups was tested using the Mann–Whitney U test or the Kruskal–Wallis H test respectively. Statistical significance of features between multiple groups was tested with the Hypergeometric test or Fisher's exact test. Hazard Ratios were calculated using the *coxph* function from package *survival* (version 3.5-5). Kaplan-Meier analysis was performed using the *survfit* and *ggsurvplot* functions from *survminer* package (version 0.4.9) and groups were tested with log-rank test. Differences were considered significant at $p < 0.05$.

References

1. American Cancer Society. Breast Cancer Facts & Figures 2022-2024. Atlanta: American Cancer Society, Inc. 2022. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf.

2. American Cancer Society. Cancer Facts & Figures 2023. Atlanta: American Cancer Society; 2023. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

3. *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.*

4. Pan, H. *et al*. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *N. Engl. J. Med.* **377**, 1836–1846 (2017).

5. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **26**, 1291–1299 (2015).

6. Coughlin, S. S. Epidemiology of Breast Cancer in Women. *Adv. Exp. Med. Biol.* **1152**, 9–29 (2019).

7. Kleibl, Z. & Kristensen, V. N. Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management. *Breast Edinb. Scotl.* **28**, 136–144 (2016).

8. 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**, BCBCR.S39384 (2016).

9. Loibl, S., Poortmans, P., Morrow, M., Denkert, C. & Curigliano, G. Breast cancer. *Lancet Lond. Engl.* **397**, 1750–1769 (2021).

10. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019).

11. 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).

12. Aran, D. *et al.* Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* **8**, 1077 (2017).

13. 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).

14. 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).

15. Kostecka, A. *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* **8**, 1–10 (2022).

16. Pipek, O. *et al.* Genomic Landscape of Normal and Breast Cancer Tissues in a Hungarian Pilot Cohort. *Int. J. Mol. Sci.* **24**, 8553 (2023).

17. Brogna, S. & Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* **16**, 107–113 (2009).

18. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

19. Rampias, T. *et al.* The lysine-specific methyltransferase KMT 2C/ MLL 3 regulates DNA repair components in cancer. *EMBO Rep.* **20**, e46821 (2019).

20. Liang, B. *et al.* TBX3 functions as a tumor suppressor downstream of activated CTNNB1 mutants during hepatocarcinogenesis. *J. Hepatol.* **75**, 120–131 (2021).

21. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

22. Luongo, F. *et al*. PTEN Tumor-Suppressor: The Dam of Stemness in Cancer. *Cancers* **11**, 1076 (2019).

23. Gailani, M. R. *et al*. Developmental defects in Gorlin syndrome related to a putative tumor suppressor gene on chromosome 9. *Cell* **69**, 111–117 (1992).

24. Staal, S. P. 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.* **84**, 5034–5037 (1987).

25. Bader, A. G., Kang, S., Zhao, L. & Vogt, P. K. Oncogenic PI3K deregulates transcription and translation. *Nat. Rev. Cancer* **5**, 921–929 (2005).

26. Kesarwani, A. K. *et al*. Cancer-associated SF3B1 mutants recognize otherwise inaccessible cryptic 3' splice sites within RNA secondary structures. *Oncogene* **36**, 1123–1133 (2017).

27. Hobbs, G. A., Der, C. J. & Rossman, K. L. RAS isoforms and mutations in cancer at a glance. *J. Cell Sci.* **129**, 1287–1292 (2016).

28. He, X. *et al*. The G protein α subunit Gαs is a tumor suppressor in Sonic hedgehog-driven medulloblastoma. *Nat. Med.* **20**, 1035–1042 (2014).

29. Otálora-Otálora BA, Henríquez B, López-Kleine L and Rojas A. RUNX family: Oncogenes or tumor suppressors (Review). *Oncol. Rep.* (2019) doi:10.3892/or.2019.7149.

30. Martínez-Sáez, O. *et al*. Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Res.* **22**, 45 (2020).

31. Hainaut, P. & Pfeifer, G. P. Somatic TP53 Mutations in the Era of Genome Sequencing. *Cold Spring Harb. Perspect. Med.* **6**, a026179 (2016).

32. Zhou, X. *et al*. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48**, 4–6 (2016).

33. 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.

34. Pereira, B. *et al*. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).

35. Sondka, Z. *et al*. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

36. Crosby, D. *et al*. Early detection of cancer. *Science* **375**, eaay9040 (2022).

37. 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).

38. 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).

39. Vogelstein, B. *et al*. Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).

40. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium *et al*. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

41. Kandel, R., Li, S.-Q., Ozcelik, H. & Rohan, T. p53 protein accumulation and mutations in normal and benign breast tissue. *Int. J. Cancer* **87**, 73–78 (2000).

42. Soysal, S. D. *et al*. Genetic Alterations in Benign Breast Biopsies of Subsequent Breast Cancer Patients. *Front. Med.* **6**, 166 (2019).

43. Cereser, B. *et al*. The mutational landscape of the adult healthy parous and nulliparous human breast. *Nat. Commun.* **14**, 5136 (2023).

44. Macias, H. & Hinck, L. Mammary gland development. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 533–557 (2012).

45. Jassim, A., Rahrmann, E. P., Simons, B. D. & Gilbertson, R. J. Cancers make their own luck: theories of cancer origins. *Nat. Rev. Cancer* **23**, 710–724 (2023).

46. Ciwinska, M. *et al*. Mechanisms that clear mutations drive field cancerization in mammary tissue. *Nature* **633**, 198–206 (2024).

47. Srinivasan, P. *et al*. The context-specific role of germline pathogenicity in tumorigenesis. *Nat. Genet.* **53**, 1577–1585 (2021).

48. Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta BBA - Rev. Cancer* **1867**, 151–161 (2017).

49. Filipowicz, N. *et al*. Comprehensive cancer-oriented biobanking resource of human samples for studies of post-zygotic genetic variation involved in cancer predisposition. *PLOS ONE* **17**, e0266111 (2022).

50. 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. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

51. 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).

52. Sanchez-Contreras, M. *et al*. A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. *Nucleic Acids Res.* **49**, 11103–11118 (2021).

53. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Ann. Surg. Oncol.* **17**, 1471–1474 (2010).

54. Amin, M. B. *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.* **67**, 93–99 (2017).

55. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).

56. Adzhubei, I. A. *et al*. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

57. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).

58. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinforma. Oxf. Engl.* **31**, 2745–2747 (2015).

59. Davydov, E. V. *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

60. Walker, L. C. *et al*. Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *Am. J. Hum. Genet.* **110**, 1046–1067 (2023).

Data availability

Raw duplex sequencing and WES data are available upon request in the EGA archive, study IDs EGAS50000000538 and EGAS50000000539 respectively.

Acknowledgments

The authors wish to thank Agata Wojdak for her help with administrative activities. 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, University Clinical Centre in Gdańsk and University Hospital in Cracow. We thank Dr. Leszek Kalinowski for the access to selected laboratory facilities.

Contributions

Study design and conception: A.P., M.A.; Clinical data acquisition and interpretation: M.D., M.N., J.J., D.H.Z., J.S., E.S., M.L-J., J.H., D.B., M. Jankowski., W.Z., Ł.S., T.N., N.F., A.M., K.C., K.D-C., U. Ł., K.D., H.D., B. B-O.; Experiments: M.A., K.C., M. Koszyński, M. Jaśkiewicz., U. Ł., K.D., H.D., B. B-O., A.K.; Methodology: I.T-B.; Data analysis and interpretation: A.P., M.A., M. Koczkowska, K.C., P.G.B., P.M., M.H., M. Jąkalski, J.M.; Funding acquisition: A.P., J.P.D., K.C.; Visualization: M.A., M. Jąkalski; Manuscript writing – original: M.A., A.P.; Manuscript writing – review & editing: M.A., A.P., J.P.D., N.F., M. Koczkowska, M.H, P.G.B.; Supervision: A.P., J.P.D., M. Koczkowska. All authors have read and agreed to the published version of the manuscript.

Funding

This work 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) to A.P. and J.P.D. Parts of the study were supported by The Swedish Cancer Society (No. 20 0889 PjF) and Swedish Medical Research Council (No. 2020-02010) to J.P.D., by The National Science Centre Poland Miniatura 4 (Project No. 2020/04/X/NZ2/02084) to K.C., and by the Austrian Science Fund FWF (P30867000) and the European Regional Development Fund (REGGEN ATCZ207) to I.T-B.

Competing interests

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

Ethics statement

Tissue samples and patient histories were provided for this study by the Oncology Centre in Bydgoszcz, Jagiellonian University Hospital in Cracow, 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 Gdańsk (approval number NKBBN/564/2018 with multiple amendments), recruited and enrolled all donors under informed and written consent, collected, and stored all tissue samples.

Supplementary material

Supplementary Figures

Supplementary Figure 1. Graphical representation of the project workflow. Three hundred seventy-eight fresh-frozen samples, including primary tumor (PT), uninvolved mammary gland (UM, ≥1 cm from PT), skin (SK), and whole peripheral blood (BL) were collected from 126 individuals diagnosed with reportedly sporadic breast cancer. UM and BL samples were also collected from 15 individuals who underwent reduction mammoplasty surgeries, serving as controls (Reduction Mammoplasty, RM cohort). Breast cancer patients were stratified into two cohorts based on prognosis: those with adverse prognoses within the first 10 years post-surgery (Breast Cancer Adverse Prognoses, BCAP cohort) and those without specific prognosis-related criteria (Breast Cancer Un-Selected, BCUS cohort). For 7 BCAP individuals, extra UM samples excised further from the PT (UMD, 1.5-3 cm from PT, median 2.35 cm) were also collected raising the total number of samples to 415. Pathologists confirmed normal histology of UM, UMD, and SK samples. DNA was extracted and whole exome sequencing was performed to identify post-zygotic variants in matched UM and PT samples according to study criteria (Supplementary Table 2). BL or SK samples, if BL was not available, were used to distinguish between post-zygotic and germline variants. Post-zygotic and germline variants were filtered according to study criteria (detailed in Materials and Methods). Pathogenic post-zygotic variants in breast cancer-related genes, identified within BCAP and BCUS UM samples were compiled into a single list (Supplementary Table 5), and *PIK3CA* and *TP53* variants were compared between cancer cohorts. Further validation of pathogenic post-zygotic variants in *AKT1*, *PIK3CA*, *PTEN*, *TBX3*, and *TP53* genes was conducted in matching UM and PT samples of the BCAP cohort via Sanger sequencing, High-Resolution Melting, or Duplex sequencing. Selected *AKT1*,

PIK3CA, and *TP53* variants were also identified in the UMD samples of 7 BCAP patients. Hazard ratios were calculated (data not shown), and Kaplan-Meier curves were generated to estimate and visualize survival probabilities over time for different patient groups. 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/).

Supplementary Figure 2. Filtering of post-zygotic and germline variants within the Breast Cancer Adverse Prognoses (BCAP), Breast Cancer Un-Selected (BCUS), and Reduction Mammoplasty (RM) cohorts. Different filtering strategies were applied to identify breast cancer-related post-zygotic and germline variants across 378 UM, PT, BL, or SK samples from 77 BCAP and 49 BCUS patients and paired UM and BL samples from 15 RM individuals (Materials and Methods). Detailed clinicopathological information for the BCAP and BCUS cohorts is summarized in Table 1 and Supplementary Table 1. Post-zygotic variants annotated as "pathogenic," "likely pathogenic," "uncertain significance," or "conflicting interpretations of pathogenicity" in the ClinVar or InterVar databases were included, alongside variants reported in the COSMIC database (Cosmic_95coding). Variants with a minor allele frequency (MAF) ≤ 0.001 in gnomAD (v2.1.1) or not reported in gnomAD were retained for further analysis (Supplementary Table 2). These variants were categorized based on their impact on protein function, as truncating (Supplementary Table 3) or non-truncating (Supplementary Table 4). Truncating variants were classified as pathogenic regardless of ClinVar or InterVar annotation, while missense variants underwent in-silico REVEL[18] analysis (score threshold ≥ 0.75). Truncating variants, variants classified as "pathogenic" or "likely pathogenic" in ClinVar, and missense variants with conflicting interpretations or annotated as pathogenic in ClinVar without assertion criteria but with a REVEL score ≥ 0.75, were deemed pathogenic. ClinVar annotations were prioritized when selecting variants affecting breast cancer-related genes for further investigation. Variants annotated as "pathogenic" or "likely pathogenic" in ClinVar, and missense variants classified as uncertain significance or with conflicting interpretations in ClinVar, with a REVEL score ≥ 0.75, were included. Truncating variants affecting tumor suppressor genes were also considered,

767 even if absent in ClinVar (Supplementary Table 5). Selected BCAP UM, PT, and SK samples were analyzed
768 to verify or further explore pathogenic variants using Sanger sequencing/High-Resolution Melting
769 (Supplementary Table 6, Supplementary Figure 3) and ultra-sensitive Duplex sequencing (Supplementary
770 Table 7). UM samples located further from the corresponding PT (UMD) were included for 7 BCAP
771 patients. For germline variant detection, variants in high- and moderate-penetrance breast cancer
772 susceptibility genes as defined by the NCCN Clinical Practice Guidelines in Oncology[33] (Version 1.2023,
773 September 7, 2022) were included. Variants with a minor allele frequency MAF ≤ 0.01 across all gnomAD
774 populations ("popmax") or not noted in the gnomAD v2.1.1 database were analyzed. Missense variants
775 underwent in-silico analysis using the REVEL tool[18] (threshold set to 0.75). All germline variants were
776 classified according to the American College of Medical Genetics and Genomics and the Association for
777 Molecular Pathology recommendations[50]. Evaluation of identified *BRCA1* and *BRCA2* variants was
778 performed according to the Evidence-based Network for the Interpretation of Germline Mutation Alleles
779 (ENIGMA) *BRCA1* and *BRCA2* Variant Curation Expert Panel (Version 1.1.0) (Clinical Genome Resource,
780 https://www.clinicalgenome.org/affiliation/50087/,
781 https://cspec.genome.network/cspec/ui/svi/doc/GN092,
782 https://cspec.genome.network/cspec/ui/svi/doc/GN097). Full description of germline variants identified in
783 BCAP, BCUS, and RM cohorts is available in Supplementary Table 8. Parts of the figure were drawn by
784 using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative
785 Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

786 **Supplementary Figure 3. Post-zygotic DNA variants detected in the normal mammary gland samples**
787 **of breast cancer patients with adverse outcomes.** The upper part of the lollipop plots presents predicted
788 amino acid change caused by the detected variant. Missense variants are represented by blue dots,
789 frameshift variants by red dots, and nonsense by orange. Lollipop plots were prepared based on images
790 generated with the ProteinPaint application[32]. Middle panels present aligned reads from whole exome
791 sequencing of UM samples with marked variant reads and were prepared based on IGV (Integrative

Genomics Viewer, http://www.broadinstitute.org/igv). Lower panels include variant confirmation by Sanger sequencing or High-Resolution Melting (HRM). UM - uninvolved mammary gland, UMD - uninvolved mammary gland at a greater distance, PT - primary tumor, SK - skin, Control - peripheral blood from an unrelated individual. **a) Presentation of *AKT1* c.49G>A variant**. PH - Protein Kinase B-like pleckstrin homology domain, red lines within the PH domain indicate phosphoinositide binding sites; KD - Catalytic domain of the Serine/Threonine Kinase. **b) Presentation of *PIK3CA* c.1624G>A and c.3140A>G variants**. p85 - p85-binding domain; RBD - Ras-binding domain; C2 - C2 domain; AD - accessory domain; CD - catalytic domain. **c) Presentation of *PTEN* c.388C>T variant.** PTP - protein tyrosine phosphatase-like catalytic domain of phosphatase and tensin homolog, purple lines indicate active catalytic sites; C2 - C2 domain of PTEN tumor suppressor gene). **d) Presentation of *TBX3* c.371_372insTGGT variant**. DBD - DNA-binding domain; TBOX - T-box transcription factor domain. **e) Presentation of *TP53* c.711G>A variant.** TAD1, TAD2 - transcription activation domain 1 and 2; DBD - DNA-binding domain, DNA-binding sites are marked with red lines; TD - tetramerization domain.

Supplementary Figure 4. Kaplan-Meier survival curves for breast cancer patients with recurrent disease and those without. Breast cancer patients with adverse prognoses (BCAP cohort) and breast cancer patients selected without any prognoses bias (BCUS cohort) with recurrent disease had significantly lower survival probabilities compared to those without for **a) the full follow-up period (10 and 2 years for the BCAP and BCUS cohorts respectively)** (log-rank test, p=0.00017) and for **b) the first 24 months post-diagnosis** (log-rank test, p<0.0001). Within the BCAP cohort, the results were similar: **c) for the full follow-up time** (log-rank test, p=0.015) and **d) for the first 24 months after diagnosis** (log-rank test, p=0.0088). The x-axis represents time in months, and the y-axis represents the probability of survival. Censoring events are indicated by vertical ticks on the curves.

Supplementary Tables

Supplementary Table 1. Clinicopathological characteristics of reportedly sporadic breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) and the Breast Cancer Un-

Selected (BCUS) cohorts. Sets of primary tumor (PT), uninvolved mammary gland (UM), and reference (whole peripheral blood, BL or skin, SK) samples were collected from 126 individuals diagnosed with reportedly sporadic breast cancer. Breast cancer patients were recruited according to prognoses, i.e. patients who experienced either recurrent disease, such as local recurrence/metastasis to the breast or secondary organs (n=40), developed a second independent tumor (n=18), or both (n=8), and/or succumbed to the disease (n=46) within the following 10 years (BCAP cohort, n=77), or were recruited without prognoses-related criteria (BCUS cohort, n=49). [a]Cancer TNM stage and [b]tumor grade were assessed according to the American Joint Committee on Cancer guidelines 7th edition[53] (effective from 2010), and 8th edition[54] (effective from 2017) [c]Histological types of breast tumors recognized: invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), neuroendocrine, mucinous, and papillary carcinoma. Comedo refers to comedocarcinoma, a type of pre-invasive breast neoplasia. [d]In situ: ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS). [e]Estrogen receptor (ER), progesterone receptor (PR), and Ki67 proliferation markers scores were used to assign tumors to [f]biological subtypes. [g]BCT- breast-conserving therapy; SLND - sentinel lymph node dissection; MRM (Patey) - Modified Radical Mastectomy (Patey); Q - breast quadrantectomy (included in Breast-conserving therapy procedures); ALND - axillary lymph node dissection; LS - lymphoscintigraphy; SM – Simple Mastectomy. UMD – uninvolved mammary gland farther from PT (Materials and Methods). * The patient died of non-oncological causes. n.a. – not available.

Supplementary Table 2. List of post-zygotic variants that fulfilled the study's criteria, observed within uninvolved mammary gland (UM) samples across all individuals under investigation, stratified across three distinct cohorts. Patients diagnosed with breast cancer were stratified into either the Breast Cancer Adverse Prognoses (BCAP) or the Breast Cancer Un-Selected (BCUS) cohort, based on the presence/absence of criteria related to prognoses, respectively. Individuals, surgically treated for non-cancer reasons comprised the Reduction Mammoplasty (RM) cohort. [a]Genomic position provided for the hg38 assembly. [b]Variant annotation provided for the basic isoform of the transcript. [c]Variant annotation within the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/. [d]Variant annotation within the

InterVar database (https://wintervar.wglab.org/). [e]Variant ID in the Catalogue of Somatic Variants in Cancer database (Cosmic_95database, https://cancer.sanger.ac.uk/cosmic). [f]rsID in dbSNP build 150. [g]Variant allele frequency reported in the Genome Aggregation Database (gnomADv2.1.1; https://gnomad.broadinstitute.org/). [h]Allele frequency of detected post-zygotic variants in studied tissues: PT - primary tumor; UM - uninvolved margin. ANNOVAR Software was used for gene-based annotation (accessed between 06.2022-08.2022). n.a. – not available.

Supplementary Table 3. List of truncating post-zygotic variants that fulfilled the study's criteria, observed within uninvolved mammary gland (UM) samples of patients included in the Breast Cancer Adverse Prognoses (BCAP) cohort. Truncating post-zygotic variants were deemed pathogenic, irrespective of their annotation in ClinVar and InterVar databases. [a]Genomic position provided for the hg38 assembly. [b]Variant annotation provided for the basic isoform of the transcript. [c]Variant annotation within the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). [d]Variant annotation within the InterVar database (https://wintervar.wglab.org/). [e]Variant ID in the Catalogue of Somatic Mutations in Cancer database (Cosmic_95database, https://cancer.sanger.ac.uk/cosmic). [f]rsID in dbSNP build 150. [g]Variant allele frequency reported in the Genome Aggregation Database (gnomADv2.1.1; https://gnomad.broadinstitute.org/). [h]Allele frequency of detected post-zygotic variants in studied tissues: PT - primary tumor; UM - uninvolved margin. ANNOVAR Software was used for gene-based annotation (accessed between 06.2022-08.2022). n.a. – not available.

Supplementary Table 4. List of non-truncating post-zygotic variants that fulfilled the study's criteria, observed within uninvolved mammary gland (UM) tissue across all individuals under investigation, stratified across three distinct cohorts. Patients diagnosed with breast cancer were stratified into either the Breast Cancer Adverse Prognoses (BCAP) or the Breast Cancer Un-Selected (BCUS) cohort, based on the presence/absence of criteria related to prognoses, respectively. Individuals, surgically treated for non-cancer reasons comprised the Reduction Mammoplasty (RM) cohort. [a]Genomic position provided for the hg38 assembly. [b]Variant annotation provided for the basic isoform of the transcript. [c]Variant annotation

within the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). [d]Variant annotation within the InterVar database (https://wintervar.wglab.org/). [e]Variant ID in the Catalogue of Somatic Mutations in Cancer database (Cosmic_95database, https://cancer.sanger.ac.uk/cosmic). [f]rsID in dbSNP build 150. [g]Variant allele frequency reported in the Genome Aggregation Database (gnomADv2.1.1; https://gnomad.broadinstitute.org/). [h]Variant allele frequency reported based on 1000 Genomes Project (https://www.internationalgenome.org/data). [i]Rare Exome Variant Ensemble Learner (REVEL) score[18]. [j]Pathogenicity classification according to Mendelian Clinically Applicable Pathogenicity (M-CAP) Score: (D) - damaging, (T) - tolerated. [k]Sorting Intolerant from Tolerant (SIFT) score: (D) - damaging, (T) – tolerated[55]. [l]Polyphen2 HumDiv and Polyphen2 HumVar[56]. [m]Combined Annotation Dependent Depletion (CADD) phred-like score[57]. [n]Protein Variation Effect Analyzer (PROVEAN) phred-like score: (D) - deleterious, (N) – neutral[58]. [o]Genomic Evolutionary Rate Profiling rejected substitutions (GERP++ RS score[59]. [p]SF, MaxEntScan, and NNSplice splice prediction algorithms were used for in-silico splicing analysis as implemented in Alamut Visual software v1.8.1 (September - October 2023). [q]Allele frequency of detected post-zygotic variants in studied tissues: PT - primary tumor; UM - uninvolved margin. ANNOVAR Software was used for gene-based annotation (accessed between 06.2022-08.2022). SIFT, Polyphen2HDIV, and Polyphen2HVAR were accessed via Alamut Visual software v1.8.1. (September-October 2023). 1000Genomes and CADD scores were acquired via wANNOVAR (March - June 2023). n.a. – not available.

Supplementary Table 5. List of pathogenic post-zygotic variants, according to the study criteria, observed within uninvolved mammary gland (UM) samples across all recruited breast cancer individuals, stratified in two distinct cohorts. Variants were declared as pathogenic based on their annotation in the ClinVar database, REVEL score ≥ 0.75 (for missense cases), and effect on encoded protein (Materials and Methods). Patients diagnosed with breast cancer were stratified into either the Breast Cancer Adverse Prognoses (BCAP) or the Breast Cancer Un-Selected (BCUS) cohort, based on the presence/absence of criteria related to prognoses, respectively. [a]Genomic position provided for the hg38

assembly. [b]Variant annotation provided for the basic isoform of the transcript. [c]Variant annotation within the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). [d]Variant ID in the Catalogue of Somatic Mutations in Cancer database (Cosmic_95database, https://cancer.sanger.ac.uk/cosmic). [e]rsID in dbSNP build 150. [f]Variant allele frequency reported in the Genome Aggregation Database (gnomADv2.1.1; https://gnomad.broadinstitute.org/). [g]Variant allele frequency reported based on 1000 Genomes Project (https://www.internationalgenome.org/data). [h]Rare Exome Variant Ensemble Learner (REVEL) score[18]. [i]Pathogenicity classification according to Mendelian Clinically Applicable Pathogenicity (M-CAP) Score: (D) - damaging, (T) - tolerated. [j]Sorting Intolerant from Tolerant (SIFT) score: (D) - damaging, (T) – tolerated[55]. [k]Polyphen2 HumDiv and Polyphen2 HumVar[56]. [l]Combined Annotation Dependent Depletion (CADD) phred-like score[57]. [m]Protein Variation Effect Analyzer (PROVEAN) phred-like score: (D) - deleterious, (N) – neutral[58]. [n]Genomic Evolutionary Rate Profiling rejected substitutions (GERP++ RS score[59]. [o]SF, MaxEntScan, and NNSplice splice prediction algorithms were used for in-silico splicing analysis as implemented in Alamut Visual software v1.8.1 (September - October 2023). [p]Allele frequency of detected post-zygotic variants in studied tissues: PT - primary tumor; UM - uninvolved margin. ANNOVAR Software was used for gene-based annotation (accessed between 06.2022-08.2022). SIFT, Polyphen2HDIV, and Polyphen2HVAR were accessed via Alamut Visual software v1.8.1. (September - October 2023). 1000Genomes and CADD scores were acquired via wANNOVAR (March - June 2023). n.a. – not available.

Supplementary Table 6. Pathogenic post-zygotic variants in breast cancer-associated genes observed within the uninvolved mammary gland (UM) samples of breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) cohort and validated by Sanger sequencing or High-Resolution Melting. Variants were initially detected in UM and primary tumor (PT) tissues with standard Whole Exome Sequencing (WES). [a]Variant annotation provided for the basic isoform of the transcript. [b]Pathogenicity classification according to the ClinVar database. [c]ID of the variant in the Cosmic_95 coding database. [d]rsIDs in dbSNP build 150. [e]Tissue allele frequency of the detected variants in matched UM and

PT tissues detected with WES. Detailed description of selected post-zygotic variants is provided in Supplementary Table 5. Confirmation of post-zygotic variants by Sanger sequencing or High-Resolution Melting is available in Supplementary Figure 3. *The *AKT1* variant was also confirmed in the UM sample located further from the PT (UMD) of patient BCAP66. n.a. - not available.

Supplementary Table 7. Pathogenic post-zygotic variants in *PIK3CA* and *TP53* genes observed within the uninvolved mammary gland (UM) samples of breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) cohort and validated by Duplex sequencing. Variants were initially detected in UM and primary tumor (PT) tissues with standard Whole Exome Sequencing (WES). UM samples, located further from the PT (UMD), available for 6 individuals, were included in the Duplex sequencing experiment. [a]Variant annotation provided for the basic isoform of the transcript. [b]Variant allele frequency in PT tissues, identified via WES. PT tissues were not included in the Duplex sequencing experiment. *Variants were confirmed via Sanger sequencing/High-Resolution Melting. Detailed description of selected post-zygotic variants is provided in Supplementary Table 5. n.a. – not applicable.

Supplementary Table 8. Pathogenic germline variants in high and moderate penetrance genes for breast cancer according to NCCN Clinical Practice Guidelines in Oncology[33] (Version1.2023, September 7, 2022), detected in breast cancer patients included in the Breast Cancer Adverse Prognoses (BCAP) and the Breast Cancer Un-Selected (BCUS) cohorts. [a]Genomic position provided for the hg38 assembly. [b]Variant annotation provided for the basic isoform of the transcript. [c]Pathogenicity classification according to the current American College of Medical Genetics and Genomics (ACMG) guidelines, classification based on the following criteria, cited directly from Richards et al.[50]. PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease; PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation; PM2 Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium; PP4 Patient's phenotype or family history is highly

specific for a disease with a single genetic etiology; PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation. For *BRCA1* and *BRCA2* variants, evaluation was performed according to the Evidence-based Network for the Interpretation of Germline Mutation Alleles (ENIGMA) *BRCA1* and *BRCA2* Variant Curation Expert Panel[60]. [d]Variant annotation within the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). [e]Variant ID in the Catalogue of Somatic Mutations in Cancer database (Cosmic_95database, https://cancer.sanger.ac.uk/cosmic). [f]rsID in dbSNP build 150. [g]Variant allele frequency reported in the Genome Aggregation Database (gnomADv2.1.1; https://gnomad.broadinstitute.org/). [h]Variant allele frequency reported based on 1000 Genomes Project (https://www.internationalgenome.org/data). [i]Number of entries and variant classification on pathogenicity according to the Leiden Open (source) Variation Database (LOVD) (https://www.lovd.nl/3.0/home) (last accessed on 18.05.2024). [j]Allele frequency of detected germline variants in studied tissues: SK – skin; BL – whole peripheral blood; PT - primary tumor; UM - uninvolved margin. ANNOVAR Software was used for gene-based annotation (accessed between 06.2022-08.2022). 1000Genomes information was acquired via wANNOVAR (as of 21.05.2024). n.a. – not available.