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&*], Tomasz Nowikiewicz[3,4&*], 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], Arkadiusz Piotrowski[1,2*]

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 - Chair of Clinical Pathomorphology, Collegium Medicum UMK, 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.

[&] These authors contributed equally.

[*] Corresponding author:

Email: arkadiusz.piotrowski@gumed.edu.pl (AP) or anna.kostecka@gumed.edu.pl (AK), or tomasz.nowikiewicz@gmail.com (TN)

Abstract

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, including breast-conserving surgery cases. Targeted resequencing of 542 cancer associated genes revealed mosaic 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 ($9x10^{-2}$ to $5.2x10^{-1}$) 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-mosaic (in 10^{-2} to 10^{-4} alleles) hotspot pathogenic variants. Our results raise a question about the oncogenic potential in non-tumor mammary gland tissue of breast-conserving surgery patients.

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, distal to the primary tumor site[13–15]. The most pronounced genetic alterations were identified in 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 for the presence of mosaic somatic pathogenic genetic alterations in breast cancer-related genes in the normal mammary gland of sporadic cancer patients (study overview in the Supplementary Figure S1).

Our study demonstrates that structural chromosomal aberrations and clearly pathogenic point variants in crucial breast cancer driver genes are ubiquitous in 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 glandular tissue (UM) and primary tumor (PT), we included normal mammary gland samples from 26 age-matched individuals that underwent breast reduction surgery and served as the control group (Supplementary Figure S2). Spectrum of CNAs in the studied cohort is presented on Figure 1. Hierarchical clustering revealed two clusters with PT-only and control-only samples and four additional clusters with mixed sample distribution (Supplementary Figure S3). We also carried out cross analysis between the studied sample groups of CNAs type, size and number. The PT samples 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 indicating more heterogeneous nature of the control samples (Supplementary Figure S4).

We identified recurrent chromosomal aberrations in UM samples 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 UM and PT samples, but also in the normal mammary gland tissue of healthy controls[19]. We observed additional events that frequently accompany 8p LOH, in the uninvolved mammary gland: 9p loss and 8q gain. *ERBB2* gains were observed exclusively in PT samples, except for one control mammary gland sample.

Somatic mosaic pathogenic variants in breast cancer driver genes present in the normal mammary gland tissue

We applied targeted DNA sequencing to identify somatic variants in sets of UM, BL and PT samples of 52 individuals diagnosed with sporadic breast cancer to distinguish germline and somatic alterations (Supplementary Table S1, Supplementary Table S2).

Four individuals (4/52, 7.7%) were heterozygous for a constitutional pathogenic variant of a known breast cancer-associated gene, i.e. c.5179A>T (p.Lys1727Ter) and c.181T>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 S3). 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.

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 S4 and S5, 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 (Figure 2). All of these variants except *PIK3CA* c.3140A>G (p.His1047Arg) were detected in BCS patients, in samples from the tissue portion that was not qualified for surgical resection. Below we describe these variants according to their *in-silico* predicted predominating effect on the encoded protein function in the context of cancer.

RAD50

Genome integrity is destabilized in cancer due to the inability to repair DNA lesions and sustained proliferative signaling that promotes the accumulation of genomic changes. DNA damage response (DDR) is a protective mechanism activated upon genotoxic stress. Rad50 is a part of the Mre11-Rad50-Nbs1 complex implicated in early response to DNA double strand breaks[20]. Germline *RAD50* pathogenic variants are associated with an increased risk of breast cancer, with the particular c.2165dup p.Glu723fs variant reported in patients with early-onset familial breast cancer[21,22]. Here we identified this pathogenic variant in a UM sample of P16 individual (Figure 2, Table 2, Supplementary Table S5).

AKT1

Akt1 kinase regulates DNA repair processes and is often abnormally activated in breast tumors. Activating *AKT1* pathogenic variants promote its membrane localization and stimulate proliferation independent of growth receptor signaling. Constitutive activation

of Akt1 negatively regulates DNA repair processes, creating a BRCA1-deficient phenotype without *BRCA1* mutations[23,24]. We detected an *AKT1* missense variant, c.49G>A (p.Glu17Lys) in matched normal mammary gland and Luminal B tumor samples collected from P26 case (Figure 2, Table 2, Supplementary Figure S5). This somatic variant, reported in 3% of primary breast cancers, exclusively in ER-positive, affects the pleckstrin homology domain, alters ligand-binding site, promotes pathological Akt1 membrane association, enhances growth, transforms cells *in vitro* and induces leukemia in mice[25,26]. Notably, the UM sample with the c.49G>A variant had also an excessive load of DNA copy number alterations, reflecting the deleterious effect of this variant on the genome[24] (Supplementary Figure S6).

MAP3K1

Accumulation of DNA lesions is a stress signal that triggers cell death and removes damaged cells. Regulation of apoptosis is complex and involves many protein effectors[27]. MAP3K1, a member of the MAPK kinase family, is involved in the regulation of proliferation, growth and the execution of apoptosis. Loss of MAP3K1 function can render cells insensitive to death-inducing signals[28]. Although *MAP3K1* pathogenic variants predominantly occur in luminal subtypes, reports confirm their presence also in triple-negative tumors[5]. We detected a clearly pathogenic truncating variant, c.2668del (p.Asn891fs) in the uninvolved mammary gland of P15 individual and matched primary tumor sample (Figure 2, Table 2, Supplementary Figure S5).

NCOR1

NCOR1 is one of the most frequently mutated genes in breast tumors and an element of the chromatin remodeling complex that was found to repress ER-mediated transcription[6]. We have previously reported a truncating *NCOR1* variant in the

uninvolved breast tissue[13]. Here we detected a rare c.6715C>A (p.Pro2239Thr) missense variant, not reported in clinical and population databases, but classified as likely pathogenic according to the current recommendations[29] (Figure 2, Table 2, Supplementary Figure S5).

MED12

Mediator subunit 12 (MED12) regulates RNA polymerase II transcription and gene expression[30]. *MED12* pathogenic variants were observed in malignant phyllodes tumors and fibroadenomas with a greater frequency in the latter. According to a model proposed by Pareja *et al*., *MED12* pathogenic variants are early events and when followed by additional activation of oncogenes or inactivation of suppressors can cause malignant progression[31]. We detected a missense c.5983C>T (p.Pro1995Ser), predicted to target the *MED12* catenin-binding site, in the normal glandular tissue of P15 patient. This individual also carried a truncating *MAP3K1* suppressor gene variant in the same UM sample. Under physiological conditions MED12 regulates signal transduction through the Wnt/β-catenin pathway through direct interactions with β-catenin[32]. Altered binding efficiency might impact Wnt/β-catenin transcriptional output (Figure 2, Table 2, Supplementary Figure S5).

CBFB

The core binding factor subunit beta (CBFB) tumor suppressor forms a transcriptional complex with RUNX1 and in the cytoplasm is involved in the regulation of translation[33]. Sequencing of 2433 primary breast tumors revealed a high prevalence of *CBFB* inactivating pathogenic variants that were associated with ER+ and lower grade tumors[7]. We detected a truncating c.207dup (p.Pro70fs) variant in the uninvolved mammary gland of P23 individual (Figure 2, Table 2, Supplementary Figure S5).

TBX3

T-box transcription factor regulates normal mammary gland development. *TBX3* truncating variants have been observed in breast tumors[6,7]. In ductal carcinoma *in situ* (DCIS) high *TBX3* expression is associated with progression from the benign to invasive state[34]. The detected sequence duplication c.963del (p.Ser321fs) targets the DNA-binding domain of TBX3 and is predicted to affect TBX3 function (Figure 2, Table 2, Supplementary Figure S5).

CDH1

E-cadherin encoded by *CDH1* is a tumor suppressor that regulates cell adhesion, prevents invasion and metastasis. *CDH1* truncating variants are common in breast tumors and specifically associated with the lobular histological type[35]. *CDH1* nonsense variant c.1668_1669insT (p.Lys557Ter*) was detected in an uninvolved mammary gland sample and matched primary tumor classified as invasive lobular carcinoma, collected from P18 individual (Figure 2, Table 2, Supplementary Figure S5).

RB1

Somatic driver pathogenic variants of the *RB1* gene are frequently observed in breast malignancies[5,7]. The retinoblastoma tumor suppressor Rb1 coordinates cell cycle via several pathways, its loss deregulates cell cycle progression and is associated with genomic instability[36]. We detected a missense variant c.418A>G (p.Thr140Ala), not previously reported in the clinical or populational databases, but predicted to be likely pathogenic based on *in-silico* analysis (Figure 2, Table 2, Supplementary Figure S5).

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[37,38]. 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 (Figure 2, Supplementary Figure S5). *TP53* tumor suppressor acts as a transcription factor and is frequently inactivated in human malignancies, mostly through loss-of-function *TP53* variants[39–41]. We detected a Ile195Thr hotspot variant in the uninvolved mammary gland that affects the central DNA binding domain (Figure 2, Table 2, Supplementary Figure S5).

To enhance the sensitivity and accuracy of rare variant detection we employed duplex sequencing (Supplementary Figure S7). 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 (Figure 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 mosaic pathogenic variants: c.1093G>A (p.Glu365Lys), c.1358A>G (p.Glu453Gly), c.1633G>A (p.Glu545Lys) c.1634A>C (p.Glu545Ala), c.2164G>A (p.Glu722Lys), c.3140A>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[7,38] (Figure 3, Supplementary Table S6). A screen for *TP53* variants not only confirmed the presence of His168Leu variant, but also revealed additional hotspot variants: c.527G>T (p.Cys176Phe), c.701A>G (p.Tyr234Cys), c.733G>A (p.Gly245Ser), c.745A>T

(p.Arg249Trp), c.818G>A (p.Arg273His), c.839G>C (p.Arg280Thr). Importantly, all these pathogenic variants are located in the central DNA-binding domain indispensable for p53 tumor suppressive function[7,41] (Figure 3, Supplementary Table S6).

Discussion

Post-zygotic changes 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[42]. The mammary gland remains mitotically active during life and under physiological conditions is exposed to DNA-damaging estrogen metabolites[11]. Somatic mosaic 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. In this study we screened for somatic genetic changes in the normal mammary gland tissue of sporadic cancer patients, including tissue sampled in the parts of the breast that were not removed during breast-conserving surgery. We identified widespread genomic structural rearrangements that affect gene dosage and somatic mosaic 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 Figure S8). These variants were present in a considerable percentage of cells suggesting they occurred early 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 mosaic pathogenic variants of main breast tumor 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 (Figure 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 (Figure 1). This facilitated detecting subtle, but noticeable differences in terms of total number and length of all detected CNAs per individual (Supplementary Figure S4). 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[43–45]. 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 drivers of breast malignancies and accordingly the most common changes detected in our study were in the *PIK3CA* gene[5,46]. 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[47]. 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 present already 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 Figure S10, Supplementary Table S7). 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 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[48]. 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 would 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 (⅔ 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). 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 Figure S1. The histological subtypes and tumor tissue content of each PT sample were evaluated by pathologists according to the current American Joint Committee on Cancer guidelines[49]. 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 S1). 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 48h. 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 individuals that underwent breast reduction surgery were used as control samples (Supplementary Figure S2). Genotyping data was analyzed using Nexus Copy Number software version 10.0 (BioDiscovery). Quality control of samples was performed as described previously[14,50]. 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[51].

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 S2). Sequencing libraries were prepared 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/),[52]. Platypus v.0.8.1.1 (https://www.rdm.ox.ac.uk/research/lunter-group/lunter-group/) was used for variant

calling[53]. 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[54].

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), with 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[29]. Based on literature[2,7,39,55,56] we selected 155 breast cancer associated genes that were the primary focus of variant analysis (Supplementary Table S2). Somatic variants presented in Figure 2 and Table 2 were confirmed by Sanger sequencing or High Resolution Melting analysis (Supplementary Figure S5). Lollipop plots with variant demonstration were prepared based on images generated with the Protein paint application[57].

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, as according to standard NGS. The protocols used here are based on the ones described in more detail in Salazar *et al*.[58]

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, 275bp. 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 minutes. Tubes were then placed on a magnet for 5 minutes 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 minutes. 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 minutes, the dissolved beads were allowed to stand at RT for 5 minutes, 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 adaptor 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)[58]. 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 minutes. 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 minutes 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 S8. 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 2 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 on Salazar *et al*., with minor modifications on the post-capture amplification

(Supplementary Table S7)[58]. The biotinylated probes used to target exonic regions of *TP53*, and *PIK3CA* are detailed on Table S8.

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[59]. 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*[60]. 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*.[60]. 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.

Data availability

Raw microarray data are in the process of submission to EGA; https://ega-archive.org/ (accession no. will be available upon publication).

Next generation sequencing and duplex sequencing data are in the process of submission to EGA (accession no. will be available upon publication).

Acknowledgements

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

Competing interests

JPD is cofounder and shareholder in Cray Innovation AB. The remaining authors have declared that no competing interests exist.

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**, (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 postmenopausal 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. 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).

21. Heikkinen, K. *et al.* RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. *Carcinogenesis* **27**, 1593–1599 (2006).

22. Lin, P. H. *et al.* Multiple gene sequencing for risk assessment in patients with early-onset or familial breast cancer. *Oncotarget* **7**, 8310–8320 (2016).

23. Guirouilh-Barbat, J. K., Wilhelm, T. & Lopez, B. S. AKT1/BRCA1 in the control of homologous recombination and genetic stability: the missing link between hereditary and sporadic breast cancers. *Oncotarget* **1**, 691–699 (2010).

24. Plo, I. *et al.* AKT1 inhibits homologous recombination by inducing cytoplasmic retention of BRCA1 and RAD5. *Cancer Res.* **68**, 9404–9412 (2008).

25. Carpten, J. D. *et al.* A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature* **448**, 439–444 (2007).

26. Yi, K. H., Axtmayer, J., Gustin, J. P., Rajpurohit, A. & Lauring, J. Functional analysis of non-hotspot AKT1 mutants found in human breast cancers identifies novel driver mutations: Implications for personalized medicine. *Oncotarget* **4**, 29–34 (2013).

27. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

28. Pham, T. T., Angus, S. P. & Johnson, G. L. MAP3K1: Genomic Alterations in Cancer and Function in Promoting Cell Survival or Apoptosis. *Genes and Cancer* **4**, 419–426 (2013).

29. 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).

30. Chang, H. Y. *et al.* MED12, TERT and RARA in fibroepithelial tumours of the breast. *J. Clin. Pathol.* **73**, 51–56 (2020).

31. Pareja, F. *et al.* Phyllodes tumors with and without fibroadenoma-like areas display distinct genomic features and may evolve through distinct pathways. *npj Breast Cancer* **3**, 1–7 (2017).

32. Kim, S., Xu, X., Hecht, A. & Boyer, T. G. Mediator is a transducer of Wnt/β-catenin signaling. *J. Biol. Chem.* **281**, 14066–14075 (2006).

33. Malik, N. *et al.* The transcription factor CBFB suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* **10**, 1–15 (2019).

34. Krstic, M. *et al.* TBX3 promotes progression of pre-invasive breast cancer cells by inducing EMT and directly up-regulating SLUG. *J. Pathol.* **248**, 191–203 (2019).

35. Christgen, M. *et al.* Lobular breast cancer: Clinical, molecular and morphological characteristics. *Pathol. Res. Pract.* **212**, 583–597 (2016).

36. Witkiewicz, A. K. & Knudsen, E. S. Retinoblastoma tumor suppressor pathway in breast cancer: Prognosis, precision medicine, and therapeutic interventions. *Breast Cancer Res.* **16**, (2014).

37. 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).

38. Martínez-Saéz, O. *et al.* Frequency and spectrum of PIK3CA somatic

mutations in breast cancer. *Breast Cancer Res.* **22**, 1–9 (2020).

39. Vogelstein, B. *et al.* Cancer genome landscapes. *Science (80-.).* **340**, 1546–1558 (2013).

40. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

41. 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).

42. Mustjoki, S. & Young, N. S. Somatic Mutations in “Benign” Disease. *N. Engl. J. Med.* **384**, 2039–2052 (2021).

43. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).

44. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science (80-.).* **370**, 75–82 (2020).

45. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

46. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705.e9 (2018).

47. Soysal, S. D. *et al.* Genetic Alterations in Benign Breast Biopsies of Subsequent Breast Cancer Patients. *Front. Med.* **6**, 1–6 (2019).

48. NCCN Clinical Practice Guidelines in Oncology. Breast Cancer Version 4. 2021. *Natl. Compr. Cancer Netw.* **Version 5.**, (2021).

49. Amin, M.B., Edge, S., Greene, F., Byrd, D.R., Brookland, R.K., Washington, M.K., Gershenwald, J.E., Compton, C.C., Hess, K.R., Sullivan, D.C., Jessup, J.M., Brierley, J.D., Gaspar, L.E., Schilsky, R.L., Balch, C.M., Winchester, D.P., Asare, E.A., Madera, L. R. *AJCC Cancer Staging Manual*. (Springer International Publishing, 2017). doi:978-3-319-40617-6.

50. Rydzanicz, M. *et al.* Variable degree of mosaicism for tetrasomy 18p in phenotypically discordant monozygotic twins—Diagnostic implications. *Mol. Genet. Genomic Med.* **9**, 1–9 (2021).

51. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?e. *J. Classif.* 274–295 (2014) doi:10.1007/s00357-014-9161-z.

52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

53. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

54. 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).

55. Polyak, K. & Metzger Filho, O. SnapShot: Breast Cancer. *Cancer Cell* **22**, 562-562.e1 (2012).

56. Mahdavi, M. *et al.* Hereditary breast cancer; Genetic penetrance and current status with BRCA. *J. Cell. Physiol.* **234**, 5741–5750 (2019).

57. Zhou, X. *et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48**, 4–6 (2015).

58. Salazar, R. *et al.* Discovery of an unusual high number of de novo mutations in sperm of older men using duplex sequencing. *bioRxiv* 2021.04.26.441422 (2021) doi:10.1101/2021.04.26.441422.

59. Stoler, N. *et al.* Family reunion via error correction: an efficient analysis of duplex sequencing data. *BMC Bioinformatics* **21**, 96 (2020).

60. Povysil, G. *et al.* Increased yields of duplex sequencing data by a series of quality control tools. *NAR genomics Bioinforma.* **3**, lqab002–lqab002 (2021).

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

Table 1. Summarized clinicopathological features of sporadic breast cancer patient cohort. 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[49]. 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 S1.

Figure 1.

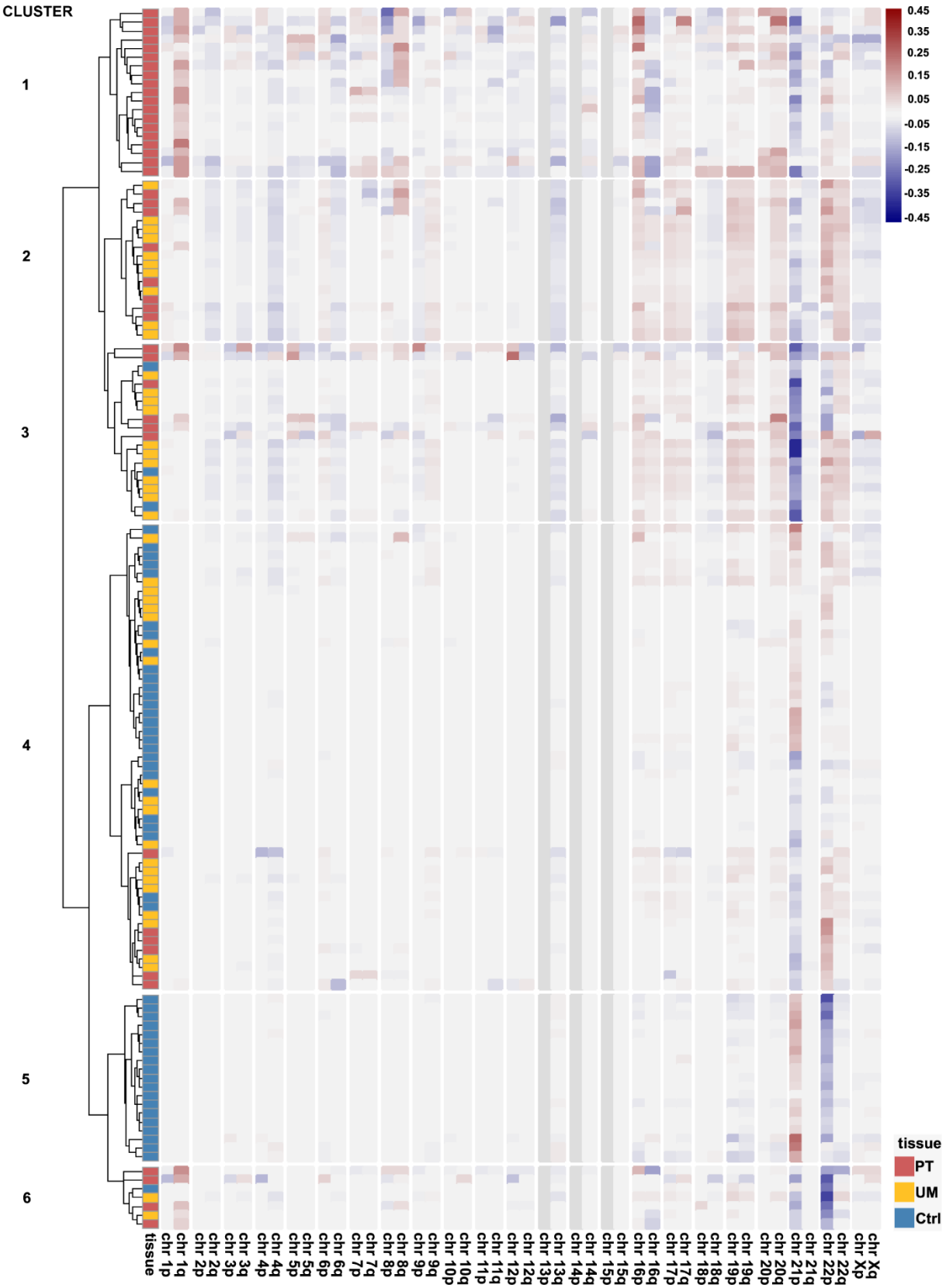

Figure 1. Overview of chromosomal Copy Number Alterations (CNAs) in sporadic breast cancer patients and healthy controls. Chromosomal CNAs were calculated as mean Log R Ratio (LRR) for chromosome arm and normalized to mean LRR of a sample. Results are presented as a heatmap with colors indicating gains (positive LRR values; red) and deletions (negative LRR values; blue). Hierarchical clustering was performed with Ward2 algorithm[51] and identified 6 clusters. Pie charts with proportion of samples within clusters are presented in the Supplementary Figure S3. Ctrl – control cohort mammary gland, UM – uninvolved mammary gland, PT- tumor.

Figure 2.

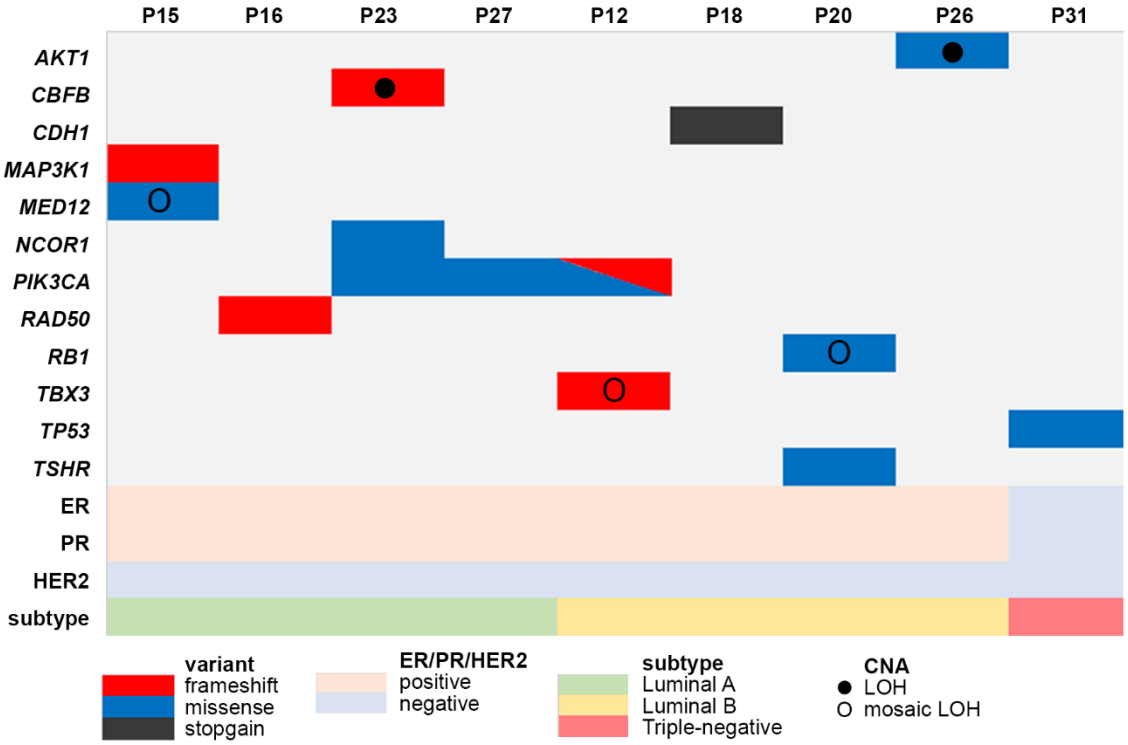

Figure 2. Summary of somatic variants of known breast cancer-associated genes detected in the normal mammary gland tissue. Targeted sequencing revealed somatic variants of known breast cancer-associated genes (rows) present in 9-52% alleles of sporadic breast cancer patients (columns). Information on estrogen receptor (ER), progesterone receptor (PR) and biological subtype of matched primary tumor sample is included. 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.

Table 2. Pathogenicity classification of somatic variants detected in uninvolved mammary gland samples.

ID	Gene	Genomic position[a]	cDNA change (protein change)[b]	ACMG classification[c]	rsID[d]	ClinVar[e]	Tissue allele frequency
P26	*AKT1*	chr14:104780214	c.49G>A (p.Glu17Lys)	Pathogenic	rs121434592	.	0,11
P23	*CBFB*	chr16:67036674	c.207dup (p.Pro70fs)	Pathogenic	.	.	0,15
P18	*CDH1*	chr16:68819382	c.1668_1669insT (p.Lys557Ter)	Pathogenic	.	.	0,1
P15	*MAP3K1*	chr5:56881868	c.2668del (p.Asn891fs)	Pathogenic	.	.	0,09
P23	*MED12*	chrX:71137882	c.5983C>T (p.Pro1995Ser)	Likely pathogenic	.	.	0,15
P15	*NCOR1*	chr17:16040459	c.6715C>A (p.Pro2239Thr)	Likely pathogenic	.	.	0,11
P12	*PIK3CA*	chr3:179234358	c.3203dup (p.Asn1068fs)	Pathogenic	rs587776802	Pathogenic	0,19
P12	*PIK3CA*	chr3:179204536	c.1093G>A (p.Glu365Lys)	Pathogenic	rs1064793732	Pathogenic	0,33
P23	*PIK3CA*	chr3:179203765	c.1035T>A (p.Asn345Lys)	Pathogenic	rs121913284	Likely pathogenic	0,11
P27	*PIK3CA*	chr3:179234297	c.3140A>G (p.His1047Arg)	Pathogenic	rs121913279	Pathogenic	0,11
P16	*RAD50*	chr5:132595759	c.2165dup (p.Glu723fs)	Pathogenic	rs397507178	Pathogenic	0,16
P20	*RB1*	chr13:48345117	c.418A>G (p.Thr140Ala)	Likely pathogenic	.	.	0,11
P12	*TBX3*	chr12:114679572	c.796_797dup (p.Ser266fs)	Pathogenic	.	.	0,18
P31	*TP53*	chr17:7674947	c.584T>C (p.Ile195Thr)	Pathogenic	rs760043106	Likely pathogenic	0,52
P20	*TSHR*	chr14:81068264	c.253A>G (p.Ile85Val)	VUS	.	.	0,13

Table 2. Pathogenicity classification of somatic variants detected in the uninvolved mammary gland samples. 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[29]. (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 S4. Confirmation of somatic variants by Sanger sequencing or High Resolution Melting is provided in the Supplementary Figure S5. VUS – Variant of Unknown Significance.

Figure 3.

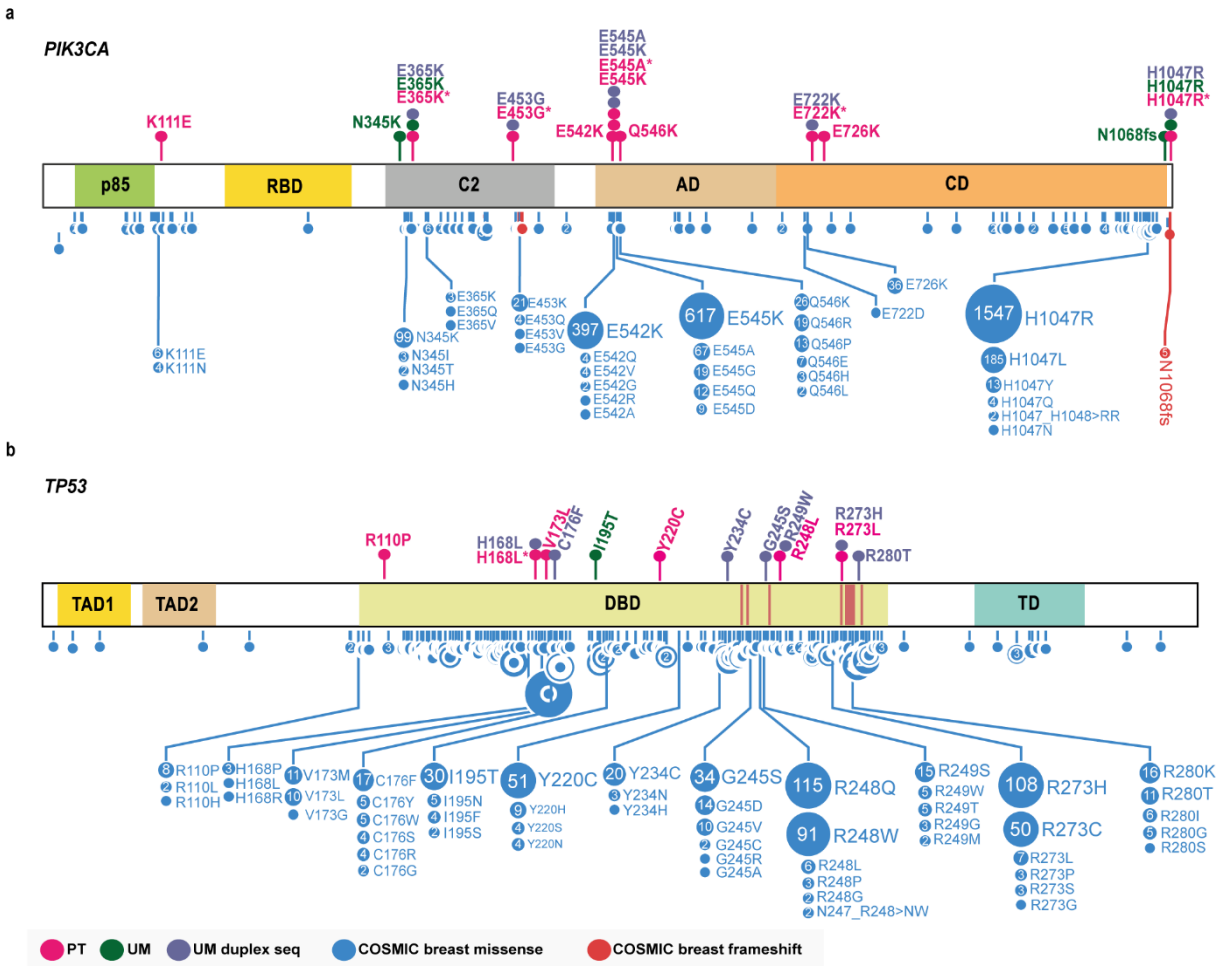

Figure 3. Somatic *PIK3CA/TP53* variants detected by targeted DNA sequencing in the uninvolved mammary gland (UM) and primary tumor (PT) samples of sporadic breast cancer patients. Lollipop plots represent somatic variants of (A) *PIK3CA* and (B) *TP53* genes detected by targeted next generation sequencing (NGS). Upper panel represents variants detected in patient UM and 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[57]. *variants detected by standard NGS in primary tumor samples and selected for duplex sequencing.

Figure 4.

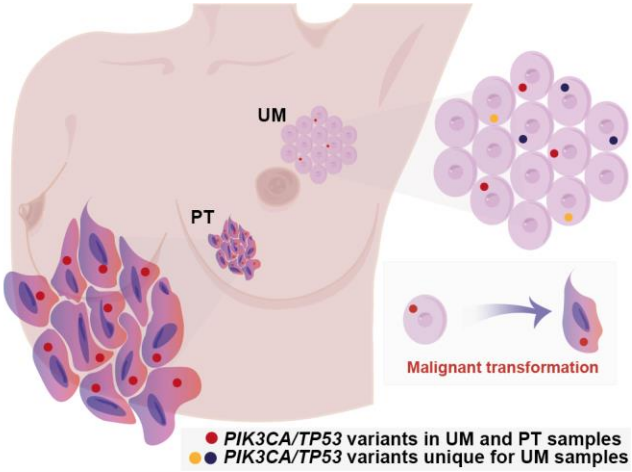

Figure 4. Low-frequency *PIK3CA* and *TP53* variants reside in the seemingly normal breast tissue that is not removed during surgery of sporadic breast cancer patients. We used duplex sequencing to screen for ultra-low frequency variants and detected *PIK3CA* and *TP53* hotspot alterations. The sampled 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.