Mutational landscape of pure ductal carcinoma in situ and associations with disease prognosis and response to radiotherapy

Noor Rizvi[1], Eliseos J. Mucaki[1,2], Emily L. Salmini[1], Monica Zhang[2], Sabina Trebinjac[3], Ezra Hahn[3], Lawrence Paszat[3], Sharon Nofech-Mozes[4], Michael T. Hallett[1,5,6], Eileen Rakovitch[3], Vanessa Dumeaux[1,2,5,6]

[1] Department of Biochemistry, Western University; [2] Department of Anatomy and Cell Biology, Western University; [3] Sunnybrook Health Sciences Centre, Department of Radiation Oncology, University of Toronto; [4] Sunnybrook Health Sciences Centre, Department of Pathology, University of Toronto; [5] Department of Oncology, Western University; [6] Centre for Translational Cancer Research, Western University

Abstract

Ductal Carcinoma in Situ (DCIS) management is challenged by the absence of reliable markers predictive of radiotherapy (RT) response, leading to both overtreatment of indolent disease and inadequate treatment for aggressive cases. Through whole-exome sequencing of 147 DCIS cases, we characterized the genomic landscape and identified markers for disease prognosis - specifically the risk of local recurrence (in situ or invasive) within 10 years after diagnosis. Our analysis revealed that pure DCIS is characterized by frequent mutations in genes governing tissue architecture, with established cancer drivers (*PIK3CA*, *TP53*) present at lower frequencies (~10%) than in invasive disease. These early driver mutations, while potentially conferring fitness advantages to pre-malignant cells, lack prognostic value, suggesting they may act as fitness enhancers rather than direct drivers of progression. A subset of younger patients exhibited distinct mutational processes with increased mutational burden, though this was not associated with recurrence risk. We identified twelve mutually exclusive genes significantly associated with early recurrence risk across the entire cohort, functioning in cytoskeleton and vesicle dynamics (*MYO7A*, *STON1*), signal transduction (*NPFFR1*), and DNA/RNA regulation. In RT-treated patients specifically, we identified 27 co-occurring variants in genes controlling cytoskeletal organization (*SORBS1*, *KRT1*), cell polarity (*WWC1*, *PATJ*), and extracellular matrix interactions (*COL5A3*, *RELN*) that were also associated with early local recurrence (within 3 years). Copy number analyses revealed that pure DCIS already harbors molecular subtype-specific patterns characteristic of invasive disease with novel recurrence-associated alterations including gains at 11q11-12 and 5p14, containing genes involved in cell adhesion. These findings suggest that certain DCIS lesions harbor genetic alterations that may compromise tissue homeostasis and create an epithelial

microenvironment conducive to disease relapse, some specifically associated as following RT-induced stress, providing novel biomarkers for treatment optimization in DCIS patients.

Introduction

Ductal Carcinoma in Situ (DCIS) is a non-invasive, non-obligate precursor of invasive breast cancer characterized by clonal proliferation of neoplastic cells confined within the breast ducts[1]. DCIS is primarily detected through screening and diagnostic mammograms due to its characteristic presentation with microcalcifications. The widespread implementation of mammographic screening programs has led to a substantial increase in DCIS detection, with incidence rates rising from four to eleven cases per 100,000 women between 1993 and 2007[2,3]. This trend is expected to continue as screening programs expand to include younger women[4].

There are currently no definitive markers to predict which cases will progress to invasive, life-threatening lesions[5,6]. Consequently, treatment, usually involving breast-conserving surgery (BCS) followed by breast radiation therapy (RT), is recommended for all women diagnosed with DCIS[5]. However, many DCIS would not transit to invasive life-threatening disease even if left untreated[7]. The absence of known markers predictive of a patient benefit to RT results in both over-treatment for indolent lesions and under-treatment for some aggressive DCIS likely to evolve to an invasive state; markers that would allow us to identify such cases would allow for tailored treatment that could include for example full mastectomy, adjuvant systemic treatments for patients who may experience in situ or invasive local recurrence (LR) despite receiving RT[8] or more frequent watch and wait strategies for patients with indolent lesions.

Cancer emerges from the accumulation of genetic aberrations in neoplastic cells and genomic instability[9]. Mutations can disrupt normal cellular processes, leading to uncontrolled cell growth, impaired DNA repair, and other aberrant cellular behaviours that may contribute to cancer development[9]. Driver mutations, a subset of somatic mutations, confer selective advantages to the neoplastic epithelial cells directly by increasing their relative fitness and therefore an increase in abundance of these driver mutations is observed in tumor cell populations[10]. In the context of DCIS, the accumulation of driver mutations may play a role in guiding the transition from DCIS to invasive breast cancer, and may determine whether a patient will benefit or not from RT.

Previous studies have revealed important insights into DCIS progression by highlighting specific mutations and chromosomal alterations that may influence the progression towards invasive disease. Many of these studies however, focus on synchronous DCIS – concurrent presentation of DCIS and invasive ductal carcinoma (IDC). These studies explore the overlap between DCIS and IDC, and recognize the potential differences between them. However these findings might be reflective of a timepoint beyond the evolutionary bottleneck, and the full repertoire of mediators of the transition from DCIS to IDC or associated with a future risk of LR cannot be established from these studies[11].

TP53 and *PIK3CA*[12–18] are perhaps the most frequently observed mutations in breast cancer and are often characterized as likely drivers of tumorigenesis promoting growth and spread of cancer cells. Although this need not necessarily be the case since mutant field clonalization, for example, could equip pre-malignant cells with additional (epi-) genetic mutations conferring fitness advantages and allowing them to expand to ductal regions without directly driving invasive behavior[11]. This suggests that driver mutations, alongside frequently mutated genes, may serve as these early fitness enhancers rather than drivers of progression. Several copy number aberrations (CNAs) have also been identified by these studies[11–16,18–21] including frequent gains at 1q, 8q, 11q, and 17q, and losses at 16q, however the specific associations of these CNVs to DCIS prognosis remain unclear.

To address these critical gaps, we conducted comprehensive exome sequencing analysis of 147 pure DCIS cases, including patients treated with and without radiotherapy, to investigate markers of local recurrence within 10 years of diagnosis. Our study reveals a distinct mutational landscape in pure DCIS and identifies novel genomic alterations associated with tumor grade, molecular subtypes, and patient age. Most importantly, we discovered specific variants and CNAs predictive of local recurrence risk, including genetic markers associated with RT response. These findings provide insights into the genomic determinants of DCIS prognosis and treatment response, establishing a foundation for improved risk stratification and personalized treatment strategies for DCIS patients.

Results

A unique cohort of pure DCIS patients

We assembled the largest cohort of pure DCIS patients treated with BCS, with or without subsequent RT, incorporating comprehensive genomic profiling through whole-exome DNA analysis of primary DCIS tumors and matched normal tissues (n = 147; **Table 1**). The study design was balanced to include half of the patients who experienced an ipsilateral invasive or in-situ local recurrence within a 10 year follow-up cutoff, occurring at a median of 7.4 years and 2.7 years time, respectively. Most tumors were of intermediate to high grade, spanning all five molecular subtypes, with no multifocality and negative margins (**Table 1**).

		Tumor-Normal pairs	
		Without Local Recurrence (n = 75)	**With Local Recurrence (n = 72)**
Time to Recurrence in years Median (Range)	DCIS	NA	2.7 (0.5 – 22.0)
	Invasive	NA	7.4 (0.8 – 20.5)
Radiotherapy (RT)	Yes	41	32
	No	34	40
Clear Margins	Positive	3	7
	Negative	63	56
	Undetermined	9	9
Tumor size (mm), Median (Range)		15 (4 - 47.2)	17 (2 - 90)
Nuclear Grade	Low	1	3
	Moderate	48	39
	High	26	30
Multifocality	Present	19	17
	Absent	32	40
	Undetermined	24	15
Age	<50 Years Old	30	22
	50-60 Years Old	26	22
	>60 Years Old	19	28
PAM50 Subtype (Pearson Correction)	Basal-like	9	12
	Her2-like	8	10
	LumA	22	10
	LumB	7	14
	Normal-Like	15	10

Table 1. Patient and tumor clinical attributes

Mutational landscape of DCIS reveals distinct processes driving high mutational burden in early-onset cases

The mutational load of pure DCIS lesions varied considerably, ranging from 3 to 3,482 variants per sample, with a median of 75 non-synonymous variants. The majority (74.2%) were missense mutations, constituting approximately 52K unique variants identified in ~14K genes.

As expected, most mutations are of C>T type (barplot in **Fig. 1**), a common mutational pattern attributed to the spontaneous deamination of 5-methylcytosine, a process frequently observed in many cancer genomes[22,23]. Spontaneous deamination can be exacerbated in FFPE samples due to DNA damage[24]. FFPE-related mutational artefacts are known to resemble certain COSMIC signatures[25], such as SBS30 and SBS1[26]. During library preparation, formalin-induced DNA lesions are chemically repaired with unrepaired profiles resembling SBS30 and repaired profiles resembling SBS1[26]. While SBS30 was not identified in our dataset, SBS1 was detected in a large number of samples (**Fig. 1)**. SBS1, commonly found in tumor genomes, is difficult to distinguish from repaired FFPE-related artifacts due to their high similarity[23,26]. However, previous research has demonstrated that mutational profiles from repaired FFPE samples closely match true tumor mutational profiles[26], and we expect that our careful processing of mutation calling-pipeline minimizes the impact of these artefacts[24].

Interestingly, a subset of DCIS samples exhibit a significantly higher mutational load, characterized by an increased in frequency of C>G, T>C, and C>A mutations, and enrichment of specific SBS signatures including SBS26, SBS7b, and SBS5 (**Fig. 1**, samples on the right). The causes underlying these mutational signatures remain incompletely understood. However, SBS26 has been linked to impaired mismatch repair and microsatellite instability. These cases were not more likely to have a recurrence, of a higher grade or of a specific molecular subtype but were predominantly found in younger patients (< 50 years).

Not all younger patients had a lesion with high TMB. Overall, younger age at diagnosis (<50 years) was associated with a higher risk of LR (Hazard ratio [HR] = 15.92, 95% CI: 1.94-2066.75, p = 0.005) compared to the middle age group (50-60 years) and showed significantly greater benefit from RT (interaction HR = 0.02, 95% CI: 0.0001-0.33, p = 0.004; likelihood ratio test p = 0.046) but this enhanced RT response was not associated with the higher tumor mutational burden observed in some younger patients.

These findings highlight distinct mutational processes in some early-onset DCIS cases, though these molecular features do not explain the age-dependent RT response, suggesting separate biological mechanisms underlying treatment sensitivity in younger patients found in this study.

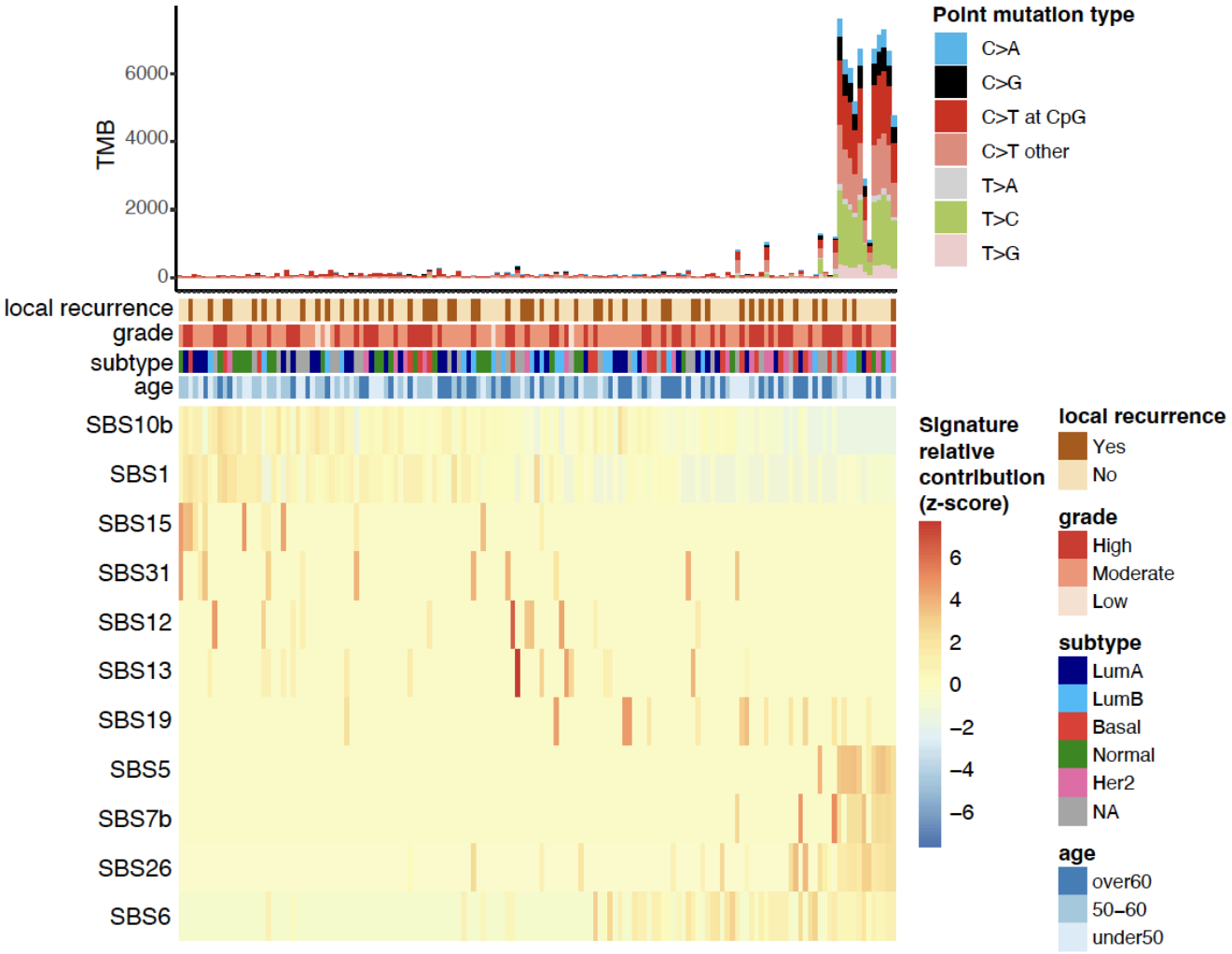

Fig. 1: Mutational Signatures in pure DCIS samples.

Top panel shows tumor mutational burden (TMB) with distribution of point mutation types. Clinical annotations display RT-outcome status (radiotherapy and recurrence), tumor grade, molecular subtype, and patient age. The bottom heatmap depicts the relative contribution across samples of single base substitution (SBS) signatures detected in at least 10 samples (z-score normalized).

Pure DCIS is associated with high frequency of mutations in genes involved cell adhesion, polarity, tissue structure and function

Analysis of pure DCIS revealed distinct patterns of recurrent mutations across multiple genes (**Fig. 2**). *PIK3CA* was the most frequently mutated gene (15% of cases), followed by *FSIP2* and *KIR3DL3* (14%). We also identified functional gene groups among the most frequently mutated genes including several motor genes converting chemical energy to mechanical force (*DNAH12, DNHD1*, and *MYOB15;* 12% each), collagen genes (*COL18A1* and *COL4A3*; 12%

and 10%, respectively), and mucin genes (*MUC3A*, *MUC4*, *MUC22*, and *MUC5AC)* also showed alterations (10-12% of cases each). While mucin genes are typically large and can accumulate mutations by chance, these specific mucin genes were not identified as FLAGS (FrequentLy mutAted GeneS) and therefore not excluded in our analysis (See Methods). Given our stringent rules for mutation calling, these findings suggest that mutations in mucin and other genes controlling epithelial-components including cellular morphology, epithelial function and adhesion are central to the physiopathology of pure DCIS.

We also identified a few mutated genes enriched in distinct clinico-pathological groups including patients with early-onset DCIS (*FILP1L*, *CFAP61*, *FREM1*) or later-onset (*ERBB4*), high-grade lesions (*TP53*) and Her2-enriched subtype (*ASH1L, NAGPA, DMD*) (**Figure 2B**).

Collectively, these findings highlight that pure DCIS harbors frequent mutations in genes governing tissue architecture and cell-cell interactions, suggesting these alterations may be fundamental to DCIS development.

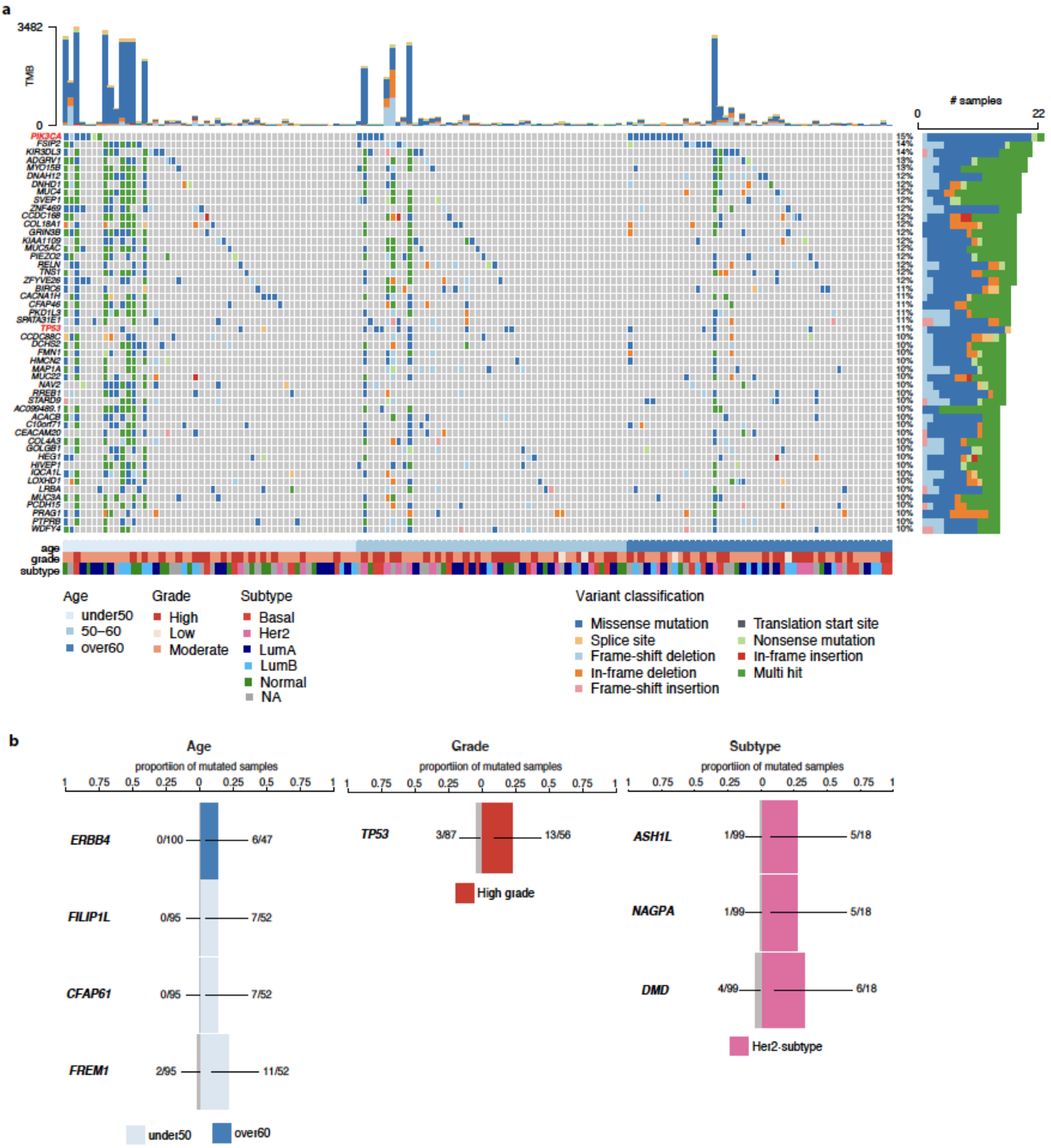

Fig. 2: The most frequently mutated genes and association with clinico-pathological variables in pure DCIS patients.

a The top 50 most frequent non-synonymous small variants identified in pure DCIS lesions. Samples are in columns and variants are color-coded based on their classification. The TMB for each lesion is displayed at the top of the heatmap. Samples are organized by age at diagnosis, with additional clinico-pathological features—grade and subtype—depicted at the

bottom. **b** Mutated genes significantly associated with specific clinico-pathological variables (False discovery rate [FDR] < 0.001). The barplots show the proportion of mutated samples within specific categories: age group (left), high grade (middle), and Her2-subtype (right). Proportions for each category are compared to the proportions of mutated samples in the other respective groups (shown in grey).

Established cancer driver genes are present in pure DCIS but lack prognostic capacity

Mutational analysis identified two significant driver genes in pure DCIS: *PIK3CA* and *TP53* (FDR < 0.05). Both genes were also among the most frequently mutated genes (**Fig. 2a**, red-labelled genes). Out of 44 samples carrying a mutation in at least one of these driver genes, 26 samples harbored at least one potentially actionable alteration as indicated by OncoKB (**Supplementary Fig. 1**).

The mutation spectrum in *PIK3CA* was dominated by the activating H1047R hotspot mutation (>50%) in the kinase domain of exon 21, known to enhance PI3K-mitigated pathway signaling (**Supplementary Fig. 1**)[27,28]. Similarly, *TP53* mutations (70%) clustered in known hotspots within the DNA-binding domain, resulting in the loss of tumor suppression by affecting its ability to bind to DNA (**Supplementary Fig. 1**).

Despite their established roles in cancer progression, none of these driver mutations correlated with increased 10-year local recurrence risk, suggesting that additional factors influence DCIS progression.

Identification of mutations associated with increased local recurrence risk independent of treatment

To identify potential prognostic markers, we analyzed mutated genes associated with the risk of invasive or in-situ local recurrence in the ipsilateral breast occurring between 6 months and 10 years after diagnosis. Our survival analysis identified twelve potential prognostic biomarkers significantly associated with increased 10-year local recurrence risk in DCIS (Firth-corrected cox p-value <0.01; **Fig. 3a-b**). These mutations, occurring in 7-10% of cases, were largely mutually exclusive across molecular subtypes and grades.

The presence of mutations in at least one of these genes was significantly associated with increased recurrence risk (**Fig. 3c,** log-rank p-value p<0.0001), with *MYO7A* and *PDZD8* showing the strongest associations (HR>8.0, **Fig. 3b**). The genes represent diverse cellular functions, clustering in three major functional groups: cytoskeleton and vesicle dynamics (*MYO7A, STON1*, *PDZD8),* signal transduction and receptor activity (*NPFFR1, DERL3 and STON1*), and DNA and RNA regulation (*HIVEP3, STOX1, DNASE2B*) (**Fig. 3c**).

These findings highlight the importance of cytoskeletal reorganization and vesicle dynamics in DCIS prognosis, processes critical for cell motility, enzyme secretion, cytokine production, and

adhesion molecule presentation. Additionally, the identified mutations suggest roles for pathways that may enhance cell proliferation and survival in pure DCIS.

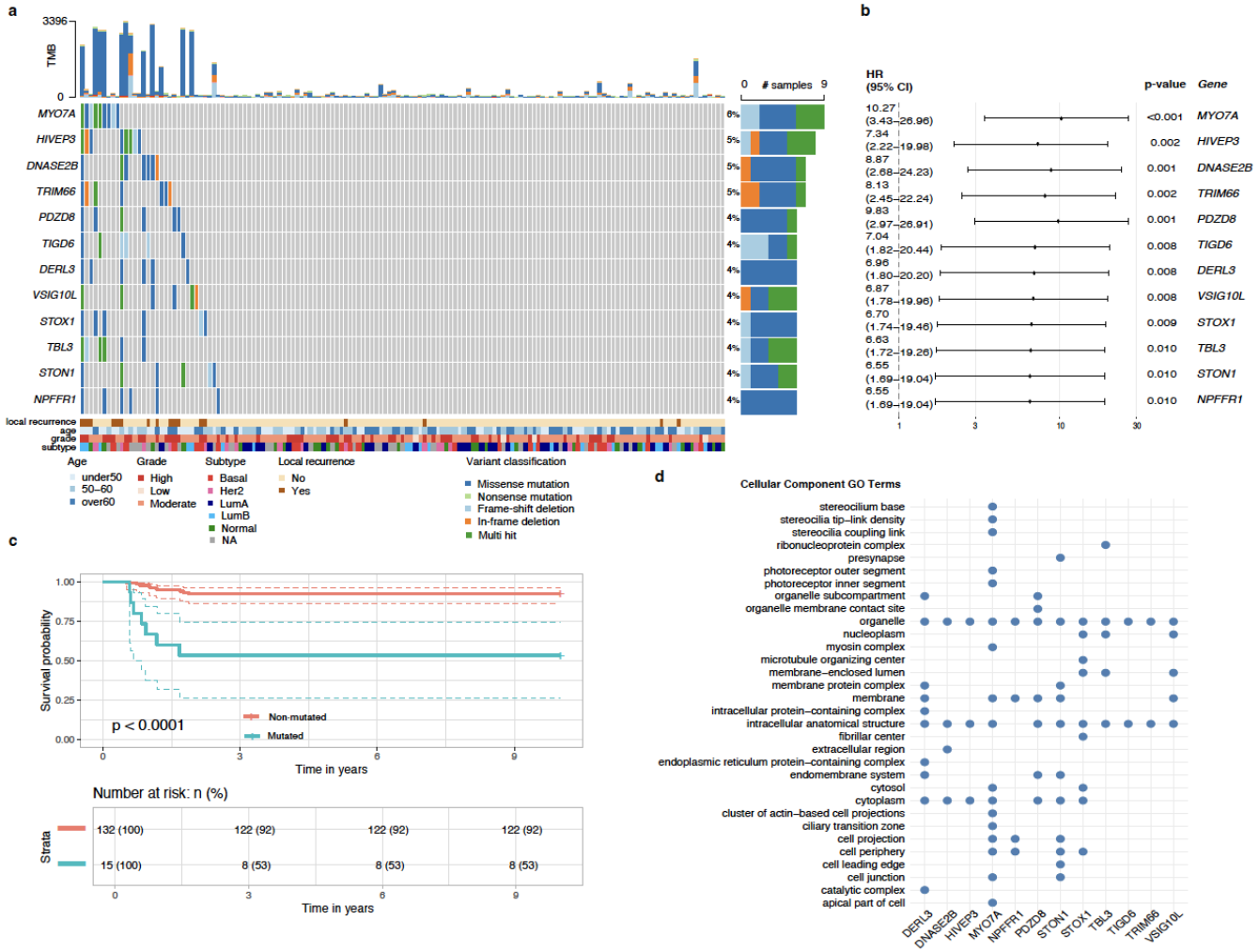

Fig. 3: Gene variants associated with an increased 10-year local recurrence risk in DCIS. **a** Samples are in columns and variants are color-coded based on their classification. Local recurrence status and clinico-pathological characteristics of each lesion are depicted at the bottom. Local recurrence is defined as any recurrence, in situ or invasive, in the ipsilateral breast occurring between 6 months and 10 years after diagnosis. Right-hand side reports the proportion of each variant classification type. **b** Hazard ratio and confidence intervals for each significant mutated gene associated with increased 10-year local recurrence risk (Firth's penalized likelihood Cox regression). **c** Kaplan-Meier analysis of local recurrence-free survival (LRFS) comparing patients with mutations in at least one of the 12 genes (red) versus those without mutations (blue). **d** Cellular component Gene Ontology (GO) terms annotations for each significant gene.

Early-recurring DCIS harbors co-occurring mutations in cell adhesion and ECM-related genes that predict radiotherapy resistance

To identify predictive biomarkers for radiotherapy response, we analyzed mutations associated with 10-year local recurrence as defined above but restricted specifically to patients who received RT. This analysis revealed 27 genes significantly associated with increased risk of recurrence risk despite RT radiotherapy (log-rank test p-value <0.05; **Fig. 4a**). Unlike the prognostic markers, mutations in these genes frequently co-occurred, with at least one mutation present in ~26% of lesions (19/73 patients). Kaplan-Meier analysis demonstrated that most recurrences in mutation-positive cases occurred within 1.5 years after radiotherapy (**Fig. 4b**). Firth's penalized likelihood Cox regression revealed a significant increase in LR risk for lesions with mutations in at least one of these genes in patients treated with RT (HR = 50.96, 95% CI: 6.01-6649.92, p < = 3.24e-05), but not in patients who did not receive RT (p = 0.1). This indicates that the presence of these mutations is specifically associated with an increased risk of recurrence following RT.

Given the frequent co-occurrence of mutations in these genes, we investigated whether the increased recurrence risk might be related to overall TMB. Despite observing several high-TMB lesions clustering among patients with these mutations (**Fig. 4c**), TMB itself showed no significant association with LR risk. Indeed, many lesions with high mutational burden remained recurrence-free during the 10-year follow-up period, suggesting that specific mutations, rather than overall mutation load, drive early recurrence after RT.

We further investigated whether clinico-pathological variables might confound the predictive power of these mutations. The effect of mutations remained significant after adjusting for age (HR = 46.95, 95% CI: 5.41-6161.54, p <0.0001). This result suggests that these mutations represent an independent risk factor for early recurrence following radiotherapy.

Our findings reveal a critical connection between the cell leading edge, apical junction complex, contractile actin filament bundle, ruffle, and clathrin-coated vesicles, highlighting a network of genes that regulate cytoskeletal dynamics, cell morphology and vesicular trafficking (**Fig. 4d**). Key players in this pathway include *SORBS1* (stress fiber formation and actin remodeling at the leading edge), *KRT1* (structural anchoring for actin filaments), *STON1* (vesicle trafficking and clathrin-coated pit formation), *WWC1* (actomyosin tension regulation and apical junction assembly), *PATJ* (apical-basal polarity and actomyosin interaction), and *RELN* (actin filament bundling and ruffle formation). These genes mediate processes essential for epithelial cell adhesion, cytoskeletal remodeling, and vesicular transport in healthy epithelial cells, and mutated forms of these genes play roles in invasion by modifying cell-cell adhesion, migration and epithelial-mesenchymal transition (EMT).

Additionally, we identified a distinct pathway hub associated with fibrillar collagen trimer and banded collagen fibril, where *COL5A3* plays a pivotal role in extracellular matrix (ECM) organization and structural support. Other pathway components include *ABCA2* (lipid

dynamics and ATPase-dependent transport), *SZT2* (GATOR2 and Seh1-associated complex in mTOR regulation), and *CEP12B* (cytoskeletal anchoring via centriolar subdistal appendage).

These findings highlight a multifaceted network that integrates cytoskeletal integrity, cell polarity, and vesicular dynamics with extracellular matrix (ECM) remodeling - these are all processes that may play a critical role in disease relapse after RT. The lack of prognostic value for recurrence in patients treated with BCS alone and not receiving RT underscores that these genes may specifically mediate resistance mechanisms that are activated in response to the oxidative stress, DNA damage, or microenvironmental changes induced by RT.

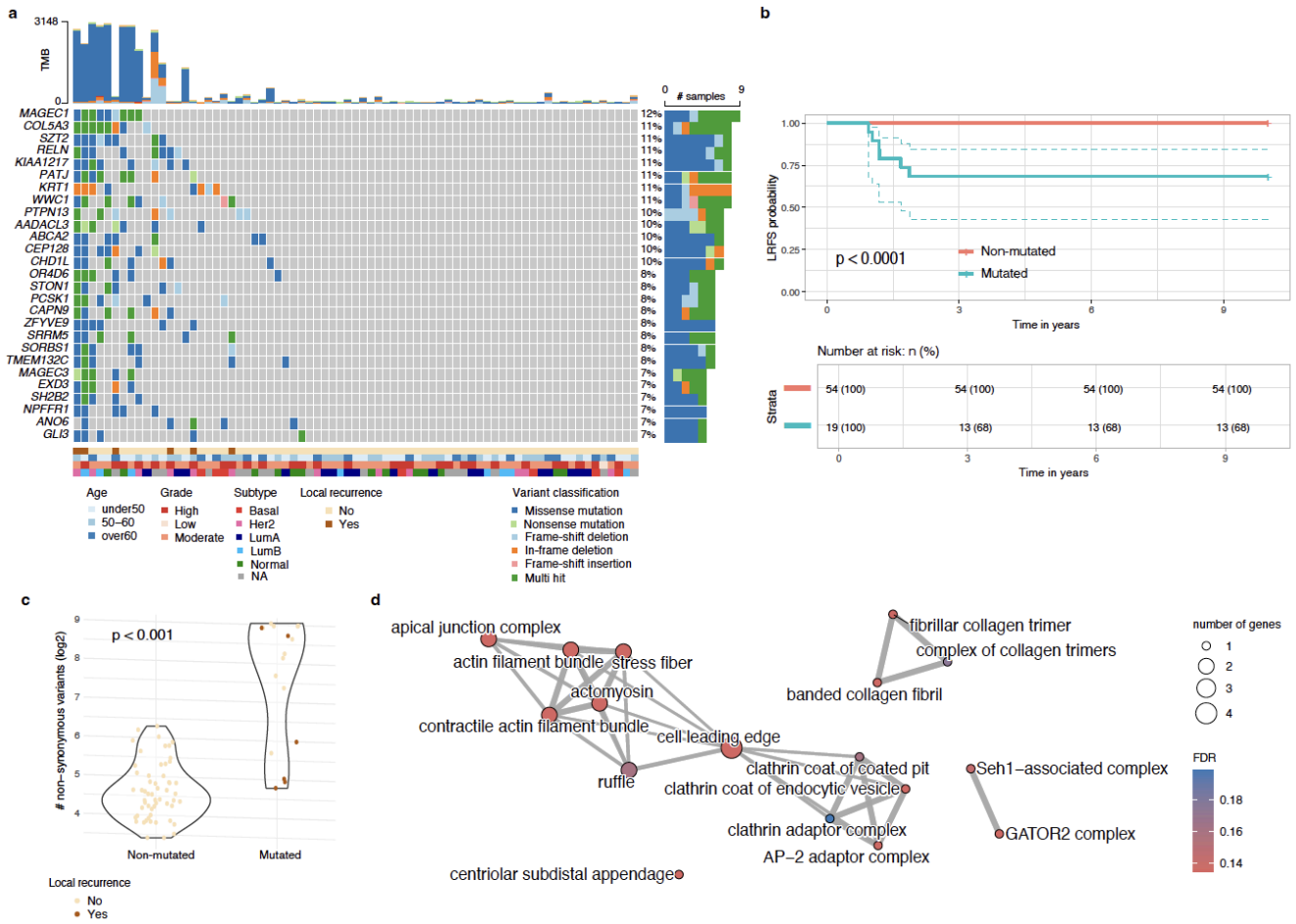

Fig. 4: Mutations associated with radiotherapy resistance in DCIS.
a An oncoplot showing the distribution of mutations in 13 genes significantly associated with local recurrence in radiotherapy-treated patients. Color-coding indicates mutation types; clinical annotations show age, grade, molecular subtype, and recurrence status. **b** Kaplan-Meier analysis of LRFS comparing patients with mutations in at least one of the 13 genes (red) versus those without mutations (blue). **c** Distribution of TMB, shown as number of non-synonymous variants, log2 scale) for lesions with at least one mutation in genes

associated with LRFS colored by recurrence status at 10 years. **d** GO cellular component enrichment network analysis of recurrence-associated genes in radiotherapy-treated DCIS. Network visualization shows enriched GO terms (FDR < 0.2) from genes linked to local recurrence following radiotherapy. Nodes represent individual GO terms with size proportional to gene count, and edges indicate significant semantic similarity between terms. Node color intensity corresponds to enrichment significance.

Copy number alterations display molecular subtype-specific patterns with select genomic regions linked to local recurrence

Analysis of copy number alterations (CNAs) revealed recurrent chromosomal changes similar to those reported in invasive breast cancer (**Fig. 5a**). Significant gains were identified on chromosomal arms 1q, 8q, 16p, 17q, 20p, and 20q, while losses predominantly occurred on 8p, 9p, 11q, 13q, 14q, 16q, and 17p (binomial test, FDR < 0.05).

Global CNA burden varied considerably across samples, with approximately half showing minimal alterations (global CNA score < -0.58; **Fig. 5b**). Basal-like tumors exhibited greater CNA burden, while normal-like tumors showed fewer alterations (**Fig. 5b**, **Supplementary Fig. 2a**). Similarly, high-grade lesions showed a greater CNA burden compared to low-grade lesions (**Supplementary Fig. 2b)**.

We identified distinct CNA patterns across molecular subtypes: basal-like tumors showed enrichment for gains on 8q, 13q, and 19q; LumA tumors frequently exhibited 16q loss; and Her2-enriched tumors showed characteristic 17q12 gains corresponding to the *ERBB2* locus (Fisher's exact test, p < 0.005; **Supplementary Fig. 2c**).

Six specific CNA regions were associated with increased 10-year local recurrence risk after adjustment for grade (Firth's penalized likelihood Cox regression p-value < 0.05; **Fig. 5c**). It includes 17q11 gain enriched in Her2-enriched cases. In contrast, recurrence-associated losses at 15q14 and gains at 11q11-12 (olfactory genes cluster), 5p14 (containing 37 genes including 4 cadherin genes) and 18p11, were observed across different grades and molecular subtypes. These alterations were largely mutually exclusive, except for adjacent cytobands for chromosome 11 gains.

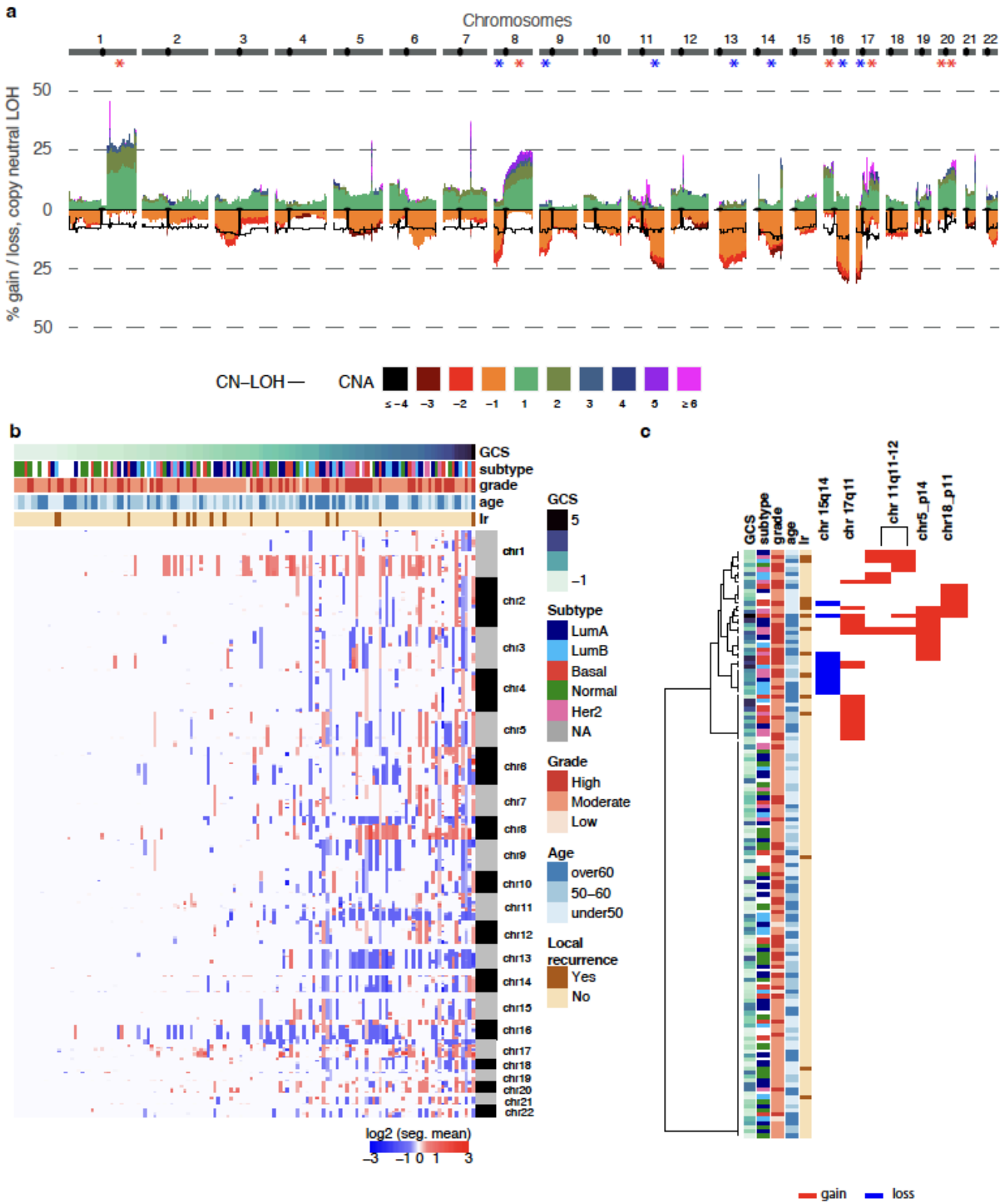

Fig. 5: Frequent copy number alterations in pure DCIS and associations with 10-year LR risk
a Genome-wide frequency of absolute copy number gains and losses across chromosomes 1-22 (1Mb window). The estimated ploidy for each sample is subtracted from the copy number

values of each segment which means that a copy number of 0 is copy number change. The y-axis shows the percentage of samples with each alteration type. Asterisks indicate statistical significance for chromosomal arm alterations **b** Heatmap showing copy number profiles per cytoband across samples (rows), ordered by global CNA score (GCS). Sample annotations include age, grade, molecular subtype and local recurrence status at 10 years. **c** Heatmap of six genomic regions significantly associated with 10-year risk of local recurrence (Firth-corrected cox model p-value < 0.05 adjusted for grade), showing copy number status. Regions present in at least 5 patients are shown, with copy number gains in red and losses in blue. Samples and regions were clustered using Ward's hierarchical clustering with Minkowski distance metric.

Contextualizing our genomic findings with prior profiling studies: consistent early mutational drivers, novel subtype-specific CNA patterns, and predictors of RT response

Most prior DCIS genomic studies analyzed small cohorts (<100 cases) and primarily focused on DCIS cases with concurrent invasive disease (synchronous DCIS) rather than pure DCIS (**Supplementary Table 1**). While recent larger studies, such as those by Strand *et al.* (2022)[21] and Kader et al. (2024)[29], examined hundreds of pure DCIS cases, their reliance on low-pass sequencing or lack of matched normal tissues limits the sensitivity for detecting genomic changes, particularly when working with FFPE samples, which are often the only available option for DCIS studies.

Across studies, *PIK3CA* and *TP53* consistently emerged as the most frequently mutated genes, with mutation rates ranging from 21–55% and 17–52%, respectively (**Supplementary Table 1**). In our cohort, *PIK3CA* and *TP53* were also among the most frequently mutated genes and identified as tumorigenesis drivers in pure DCIS. Mutations in *GATA3* and *PTEN,* which were frequently reported in several studies[14,15,30] , were only detected in a small number of lesions in our cohort (n = 3 for each). Differences across studies likely reflect variations in methodology, including variant-calling pipelines, sample types (e.g., synchronous vs. pure DCIS), cohort composition (e.g., histological grade and ER/HER2 status), and our study's specific focus on local recurrence and radiotherapy response, complicating direct comparisons. Nevertheless, the consistent identification of *PIK3CA* and *TP53* as the most frequently mutated genes across studies and in our cohort reinforces their pivotal roles in the early stages of breast tumorigenesis.

Recurrent CNAs in regions such as 1q, 8q, and 17q gains, as well as 8p, 11q, and 16q losses, were observed in our study, consistent with prior DCIS findings, including those from Strand *et al.* (2022)[21] and Abba *et al.* (2015)[16]. For example, Strand *et al.* (2022) identified 29 recurrent CNAs in DCIS but found no single CNA predictive of recurrence[21]. In contrast, we identified six genomic regions significantly associated with 10-year local recurrence risk, including regions linked to clinicopathological features associated with poor prognosis, such as 5p14 in

high-grade tumors[7,31] and 17q11-12 in Her2-enriched cases[32,33], as well as novel losses at 15q14 and gains at 11q11-12, 5p14 and 18p11 found across grades and subtypes. These differences may reflect the higher resolution of our sequencing approach (100x whole-exome sequencing) compared to the low-pass sequencing used in prior studies. Notably, our study provides valuable new insights into molecular subtype-specific CNA patterns in DCIS. In our cohort, basal-like tumors exhibited gains on 8q, 13q, and 19q; Luminal A tumors displayed 16q loss; and Her2-enriched tumors showed 17q12 gains encompassing *ERBB2*. Aside from amplifications in HER2-positive subtypes, subtype-specific CNA analyses remain poorly represented in the DCIS literature but are well established in invasive breast cancer (see supplementary text). These findings suggest that genetic and molecular aberrations defining subtypes likely arise early and are at least partially established at the DCIS stage.

Finally, while several studies have explored prognostic markers of recurrence, few account for treatment variation, and none specifically examined markers of RT response (**Supplementary Table 1**). Our study uniquely identified genetic alterations within a gene network that integrates cytoskeletal integrity, cell polarity, vesicular dynamics, and ECM remodeling, which are associated with an increased risk of local recurrence within three years following RT. This raises the possibility that impaired tissue integrity responses contribute to resistance mechanisms activated by RT-induced stress, leading to adverse effects in these lesions.

Discussion

Our study provides a comprehensive analysis of pure DCIS, focusing on DNA profiles that may influence disease progression and response to RT. While previous research has primarily examined genomic changes in synchronous DCIS compared to IDC or pure DCIS, our work offers a detailed view of the mutational landscape specific to pure DCIS. Importantly, we identified genomic alterations associated with molecular subtypes and, most critically, early local recurrence—both independent of treatment and specifically in cases following radiotherapy—highlighting potential molecular drivers of treatment outcomes. Our findings also revealed distinct mutational processes in early-onset DCIS with high tumor mutational burden potentially driven by impaired mismatch repair, but these were not associated with prognosis.

While driver genes like *TP53* and *PIK3CA* are critical in the early stages of tumorigenesis, our study found no significant associations between their presence and disease progression or response to treatment. The presence of these mutations highlights their importance in overcoming initial biological constraints; however, once these initial challenges are surpassed, other genetic alterations or microenvironmental factors likely become the driving forces of invasive progression. This is supported by the observation that these mutations are often retained in invasive carcinoma, but may occur at lower prevalence than in hyperplastic and in situ breast lesions[34]. Furthermore, studies have observed different prognostic relevance

depending on tumor clinicopathological characteristics. For instance, Lin et al.[12] observed an inverse association between *PIK3CA* kinase domain mutations in high-grade DCIS tumors and progression, while Silwal-Pandit et al.[35] reported that the prognostic impact of *TP53* mutations varied across molecular subtypes of invasive breast cancer. These findings suggest that *TP53* and *PIK3CA* mutations represent early events in the process of breast epithelial proliferation and tumorigenesis. Their biological roles, however, likely depend on the broader molecular landscape of the tumor, and they are insufficient as standalone drivers of progression to invasive disease. This underscores the complexity of breast cancer evolution, where early driver mutations interact with additional molecular and microenvironmental factors to shape disease trajectory.

Our study identified novel associations between mutations in genes regulating cell adhesion, signaling pathways, and ECM remodeling with an increased risk of invasive or in situ LR in pure DCIS. Specifically, twelve mutated genes were significantly associated with LR across the entire cohort. The genes represent diverse cellular functions, clustering in three major functional groups: cytoskeleton and vesicle dynamics (*MYO7A, STON1, PDZD8),* signal transduction and receptor activity (*NPFFR1*, *DERL3* and *STON1*), and DNA and RNA regulation (*HIVEP3*, *STOX1*, *DNASE2B*). These mutations were mutually exclusive, meaning they rarely co-occurred in the same patient, and each demonstrated independent prognostic value in univariate analyses. This pattern suggests that alterations in any one of these genes may independently contribute to recurrence risk, likely through distinct molecular mechanisms. Mutations in these genes may compromise tissue homeostasis, disrupt epithelial architecture, and facilitate conditions that promote tumor cell escape and recurrence.

When we stratified the cohort based on administration of RT, unique genes at the exception of *STON1* and *NPFFR1* were found to be significantly associated with increased risk of LR within the RT treated group. Notably, mutations in *SORBS1, KRT1*, *STON1*, *WWC1*, *PATJ* and *RELN,* playing a key role in actin dynamics, structural integrity, vesicle trafficking, apical-basal polarity, and cell junction organization, were identified. Mutations in *COL5A3* were also associated with increased risk of LR in RT-treated patients. *COL5A3* regulates the mechanical properties of the ECM, and mutations could further affect the tumor structural support and microenvironment changes in response to RT[36]. These findings build on prior studies linking disruptions in tissue structure and ECM remodeling to DCIS progression[11,16,17,19], while providing novel insights into their specific role in recurrence following RT in pure DCIS. The co-occurrence of these mutations underscores the critical importance of cytoskeletal integrity, polarity maintenance, and ECM interactions to prevent cell detachment and migration under RT-induced stress. Future studies are warranted to investigate how these mutations interact with other molecular pathways and microenvironmental factors to further elucidate their contribution to the adverse effects of RT and to identify potential strategies for mitigating recurrence risk[37–39].

Finally, we identified gains and losses that have been frequently reported in literature, namely gains on 8q[11–13,15,16,29,30], 17q[11,13,16,19–21,29,30], 20q[29,30], and losses on 11q[11,11–13,19,20,29], 16q[11,14,15,21,30].

Interestingly, these alterations reflect patterns similar to those observed in IDC, suggesting that pure DCIS already possess CNA characteristic of invasive cancer including CNAs defining molecular subtypes such as basal, luminal-A or Her2-enriched tumors. While some CNAs associated with LR likely reflect clinical or molecular subtypes with worse prognosis (5p12-15 in high grade and chr 17q12 in Her2-enriched), we also identified aberrations associated with LR independent of clinico-pathological tumor features such as gains at 11q11-12 containing olfactory genes cluster and 5p14 containing 37 genes including four cadherin genes (*CDH10*, *CDH9*, *CDH12*, *CDH18*). Future studies may benefit from focusing on risk stratification methods that go beyond traditional subtype classifications. Furthermore, the presence of LR-associated CNAs at regions linked to cell-cell adhesion, such as gains on 5p14, again indicate the importance of maintaining tissue integrity and cell signaling to prevent local recurrence.

Overall our findings uncover the genomic landscape of pure DCIS and highlight the potential factors that contribute to local recurrence and their role in mediating adverse effects of RT.

Methods

The Ontario DCIS cohort

The Ontario DCIS Cohort was established at the Sunnybrook Health Sciences Center (Toronto, Canada) as a population-based sample of women diagnosed with pure DCIS defined as in situ cancer without any invasive component between 1994 and 2003[40–42]. All patients underwent breast-conserving surgery (BCS), with a subset receiving subsequent RT. Adjuvant endocrine therapy was administered to less than 15% of individuals, while none received systemic chemotherapy or neoadjuvant endocrine therapy. The cohort features comprehensive annotation of clinical annotation and expert pathology review. Previous studies of this cohort have characterized outcomes based on clinical factors including age at diagnosis, pathological features (tumor size and nuclear grade), and treatment modalities[43–46].

Sample selection was prioritized to achieve balanced representation between RT-treated and untreated patients, as well as between patients who experienced invasive or in-situ ipsilateral local recurrence (LR) within ten years post-treatment and those who did not. Tissue cores were obtained from FFPE blocks, sampling DCIS tumors without microinvasion alongside adjacent normal and stromal tissues. DNA and RNA were extracted using the Qiagen AllPrep FFPE DNA/RNA kit (Qiagen). Samples yielding sufficient DNA quantities underwent library construction using the Nextera Flex RNA Exome kit (Illumina) and were sequenced on the NovaSeq6000 platform (100bp paired-end, 100M reads/sample) at the Genome Quebec Innovation Centre (Montreal, Canada). While high-quality sequencing data was obtained for 300 tumor tissues, downstream analyses focused on 147 samples with matched normal profiles (144 normal tissue and 3 stroma non-epithelial samples).

Molecular subtypes were determined using RNA profiles available for a subset of patients (n = 122). PAM50 subtype classification was performed using the `genefu` R package with established centroids. Specifically, normalized expression data of the 50 PAM50 genes were obtained using variance stabilizing transformation implemented in the `DESeq2` R package[47] and compared to subtype-specific centroids using Pearson correlations. Each sample was assigned to the molecular subtype with which it showed the highest correlation coefficient.

Whole-exome sequencing data preprocessing

Raw reads were processed using Trimmomatic[48] (version 0.39) to remove adaptor sequences and low quality bases. Reads were trimmed to retain high-quality sequences, applying quality thresholds of 10 at read ends and 20 within a 4-base sliding window. The remaining paired reads were processed following the GATK4 best practices. Briefly, reads were aligned to the human reference genome (GRCh38, GATK resource bundle) using the Burrows-Wheeler Aligner (BWA)[49]. Post-alignment procedures included sorting, annotating reads with read groups, and marking duplicate reads with Picard. Base quality score recalibration was conducted using GATK4 tools. A recalibration table was generated with the `BaseRecalibrator` function using known variant sites (dbSNP138 for SNPs and Mills and 1000 Genomes for indels), and recalibration was applied with `ApplyBQSR` to adjust base quality scores and correct for systematic technical errors. The process focused on SureSelect Human Exons v7 regions with a 100 bp padding. Finally, properly paired reads were extracted, excluding secondary alignments and low-quality reads, with the resulting files indexed using `Samtools`[50].

Single nucleotide variant & indel calling

For variant calling, we used NeuSomatic[51], a deep learning approach that leverages both tumor and matched normal sequence alignment information, alongside somatic mutation calls from six different approaches: MuTect2, MuSE, VarDict, VarScan2, Strelka2, and SomaticSniper[52–57]. This method was selected because of the low level of agreement between callers in our data - the majority of mutations (88.7%) were identified by only one caller (**Supplementary Fig. 3A**) - an observation consistent with previous studies[58–60]. The number of mutations detected varied significantly across samples, with the minimum identified in any single sample being 1,563 mutations, and the maximum reaching 283,247 mutations (**Supplementary Fig. 3A**).

We used the ensemble extension of NeuSomatic which includes 93 channels to capture features extracted from the six individual methods and 26 additional channels to capture the alignment information in a window of seven bases around the candidate mutation. This results in 119 input channels for each candidate matrix. We used the recommended pre-trained model SEQC-II (SEQC-WGS-Spike model) (trained on 20 whole-genome sequencing replicate pairs with *in silico* somatic mutations of 1%-100% AF, matched with both 95%N and 100%N,

~40x-220x, 5 callers used: MuTect2, Strelka2, MuSE, SomaticSniper, VarDict). After mutation calling, the recommended post-processing was applied to resolve long INDEL sequences. The final NeuSomatic predictions were used for downstream analyses.

We obtained comprehensive genomic information for each variant using the Ensembl Variant Effect Predictor (VEP).[61] This included the effects on gene and protein function, such as consequence types and amino acid changes, variant frequencies in different populations, impact on regulatory regions, and potential associations with diseases and phenotypes. Following annotation with the Ensembl VEP, we used the `vcf2maf` tool to transform VEP-annotated VCF files into the Mutation Annotation Format (MAF). This conversion ensures each variant is uniquely associated with a single gene transcript or isoform, despite the potential for a variant to impact multiple isoforms. Particularly in cases where variants could be classified under different effects — such as a `Missense_Mutation` near a `Splice_Site` — the MAF format forces a singular designation for each variant by leveraging VEP's determinations for canonical isoforms.

We excluded 100 genes commonly mutated in public exome datasets (FLAGS) due to their lower likelihood of disease association[62]. This decision stems from their longer coding regions, which inherently increase mutation probability, and the presence of paralogs that might offset functional loss these mutations could cause[62,63].

High-confidence variants (identified with a probability score of 0.7 or higher) consistently showed higher allele frequencies compared to those categorized as low-quality (with scores between 0.4 and 0.7) and rejected variants (with scores below 0.4) (**Supplementary Fig. 3B**). To reduce potential false positives, we selected high-confidence variants with allele frequency above 0.1.

Mutation patterns and frequencies were visualized using the `oncoplot` function from the `maftools` R package[64], which displays mutation types and frequencies across samples.

Mutational signatures

We performed mutational signature analysis using the COSMIC database of single-base substitution (SBS) signatures[25]. First, the trinucleotide context of each single nucleotide variant was characterized. We then used the `fit_to_signature` function in the `MutationalPatterns` R package (version 3.19) to find the linear combination of COSMIC mutation signatures that most closely reconstructs the mutation spectra for each sample by solving the nonnegative least-squares constraints problem. We used strict refit where the signature with the lowest contribution is removed; refitting is repeated until the cosine similarity between the original and reconstructed profile becomes more than max_delta= 0.004). We selected signatures that contributed to the mutation spectrum of at least 10 samples and plotted their relative contributions using the `pheatmap` R package.

Driver genes

To identify driver genes (i.e., genes under positive selection in cancer), we used the `dNdScv` analysis method[65]. This approach is based on the evaluation of the ratio between synonymous (silent) mutations and non-synonymous (missense) mutations in genes. A higher ratio of non-synonymous to synonymous mutations in a gene indicates positive selection for mutations that may confer a growth advantage to cancer cells, suggesting the potential role of the gene as a driver in tumorigenesis.

The `dNdScv` method estimates the background mutation rate of each gene by combining local information (synonymous mutations within the gene) with global information (variation of mutation rates across genes). This approach controls for the sequence composition of genes and accounts for mutational signatures, providing a more accurate estimation of the expected neutral mutation rate. In particular, the `dNdScv` R package implementation uses trinucleotide context-dependent substitution matrices to mitigate common mutation biases that can affect dN/dS calculations[65].

To visualize and analyze the distribution and nature of mutations in driver genes, we used lollipop plots generated from the cBioPortal MutationMapper tool. These plots provide a comprehensive representation of mutation types and their locations along the protein sequence as well as annotations of predicted functional impact for each mutation from Mutation Assessor (http://mutationassessor.org/, accessed on August 2024), SIFT[66], and PolyPhen-2[67]. To further characterize the identified mutations, we conducted additional investigations incorporating several layers of annotation including likely mutation hotspots as identified by Memorial Sloan Kettering Cancer Hotspots and 3D Hotspots databases[68], and annotation records of therapeutic indication from OncoKB[69], CIVIC[70], and My Cancer Genome (https://www.mycancergenome.org/, accessed on August 2024).

Copy Number Alterations (CNAs)

To investigate copy number alterations, we applied the Allele-Specific Copy Number Analysis of Tumours (ASCAT v3) on our tumor normal pairs estimating tumor purity, ploidy, and allele-specific copy number[71]. The `runAscat` function was executed with default settings optimized for high-throughput exome sequencing data, with the gamma parameter set to 1. After examination of ASCAT sunrise plots, we identified a subset of samples (n=35 samples), for which the initial estimates of tumor purity and ploidy did not align with the regions of highest confidence on the sunrise plots. The `runAscat` was re-run for these samples by manually assigning the aberrant cell fraction (tumor purity) and tumor ploidy parameters corresponding to the regions of highest probability as depicted on the sunrise plots. Overall, eight samples were excluded from further analysis due to poor goodness of fit leaving 139 samples with CNA profiles for downstream analyses. Absolute number gains and losses shared across samples were visualized across whole chromosomal regions using

aCNViewer[73] (window size of 1 Mbp). The estimated ploidy for each sample is subtracted from the copy number values of each segment which means that a copy number of 0 is copy number change. These adjusted windows at base resolution are then plotted into a stacked histogram representing genome-wide absolute copy number and copy neutral variations over all samples in a group.

We applied a re-segmentation approach to adjust for amplitude divergence due to technical variability implemented in CNApp[72] using the default settings (minimum segment length = 100 Kbp, minimum amplitude deviation from segment to zero = 0.16, maximum distance between segments=1 Mb, maximum amplitude deviation between segments = 0.16, and maximum BAF deviation between segments = 0.1). Re-segmented data were then used to calculate the broad, focal and global CNA scores. We then transformed re-segmented data into genomic regions profiles (chromosome arms, cytobands and sub-cytobands) using both focal and broad segments. Length-relative means are computed for each window by considering amplitude values from those segments included in each specific window. Default cutoffs for low-level copy number gains and losses (i.e., |0.2|) were used to infer CNA frequencies.

Survival analyses

We evaluated the association between gene mutations or copy number aberrations in cytobands and 10-year local recurrence-free survival using Firth's penalized likelihood Cox regression which accounts for small sample sizes and rare events. This analysis was conducted using the `coxphf` R package. Aberrations were included in the analysis only if detected in at least five lesions. This analysis was performed across the entire patient cohort. To further investigate genetic alterations associated with response to RT, a stratified analysis was conducted based on treatment groups.

Kaplan-Meier survival curves were used to visualize the results, illustrating event-free survival probabilities over time for patients stratified by mutational status in the specified genes or gene sets.

GO enrichment analyses were performed using the clusterProfiler R package. Entrez gene identifiers were mapped to GO terms using the `org.Hs.eg.db` annotation database. All GO terms with FDR < 0.2 were considered. Semantic similarity between GO terms was calculated using the Wang method implemented in the `pairwise_termsim` function. The enrichment map was visualized using the `emapplot` function which displays the significantly enriched terms.

Third-party studies

A systematic literature search was performed using PubMed to identify previous studies that conducted DNA profiling on pure DCIS or DCIS mixed with invasive lesions. The search strategy included terms related to “ductal carcinoma in situ”, “genetic markers”, “DCIS prognosis”, “DCIS progression”, “DCIS to IDC”, "dcis dna", "dcis prognosis dna markers", "copy number alterations", and “somatic mutations”. Studies were included if they reported genomic analyses of DCIS samples using sequencing or copy number profiling techniques and were published within the last 10 years. Twelve studies met the inclusion criteria, and their key findings were summarized in **Supplementary Table 1**. The review emphasized genetic alterations and pathway dysregulation that may drive DCIS initiation and progression to invasive disease.

Data availability

Owing to the personal, sensitive and inherently identifying nature of raw genomic data, access to raw data and patient metadata is controlled and requires institutional material data transfer agreements (contact person: eileen.rakovitch@sunnybrook.ca).

Code availability

Scripts to reproduce the analyzes performed in this study can be found at https://github.com/dumeaux-lab/dcis-dna_paper.

References

1. Harbeck, N. *et al.* Breast cancer. *Nat. Rev. Dis. Primer* **5**, 1–31 (2019).
2. Sørum, R., Hofvind, S., Skaane, P. & Haldorsen, T. Trends in incidence of ductal carcinoma in situ: The effect of a population-based screening programme. *The Breast* **19**, 499–505 (2010).
3. Gangnon, R. E. *et al.* The contribution of mammography screening to breast cancer incidence trends in the United States: an updated age-period-cohort model. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **24**, 905–912 (2015).
4. Force, U. P. S. T. *et al.* Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **331**, 1918–1930 (2024).
5. Tomlinson-Hansen, S., Khan, M. & Cassaro, S. Breast Ductal Carcinoma in Situ. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2023).
6. Zhang, Y. & Weinberg, R. A. Epithelial-to-mesenchymal transition in cancer: complexity and opportunities. *Front. Med.* **12**, 361–373 (2018).
7. Maxwell, A. J. *et al.* Unresected screen-detected ductal carcinoma in situ: Outcomes of 311

women in the Forget-Me-Not 2 study. *Breast Edinb. Scotl.* **61**, 145–155 (2022).
8. Rakovitch, E. *et al.* Refined estimates of local recurrence risks by DCIS score adjusting for clinicopathological features: a combined analysis of ECOG-ACRIN E5194 and Ontario DCIS cohort studies. *Breast Cancer Res. Treat.* **169**, 359–369 (2018).
9. Andor, N., Maley, C. C. & Ji, H. P. Genomic instability in cancer: Teetering on the limit of tolerance. *Cancer Res.* **77**, 2179–2185 (2017).
10. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
11. Trinh, A. *et al.* Genomic Alterations during the In Situ to Invasive Ductal Breast Carcinoma Transition Shaped by the Immune System. *Mol. Cancer Res.* **19**, 623–635 (2021).
12. Lin, C.-Y. *et al.* Genomic landscape of ductal carcinoma in situ and association with progression. *Breast Cancer Res. Treat.* **178**, 307–316 (2019).
13. Hernandez, L. *et al.* Genomic and mutational profiling of ductal carcinomas in situ and matched adjacent invasive breast cancers reveals intra-tumour genetic heterogeneity and clonal selection. *J. Pathol.* **227**, 42–52 (2012).
14. Nagasawa, S. *et al.* Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. *Commun. Biol.* **4**, 1–13 (2021).
15. Nachmanson, D. *et al.* The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. *Npj Breast Cancer* **8**, 1–13 (2022).
16. Abba, M. C. *et al.* A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. *Cancer Res.* **75**, 3980–3990 (2015).
17. Kaplan, H. G. *et al.* Multi-omic profiling of simultaneous ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Res. Treat.* **205**, 451–464 (2024).
18. Kader, T. *et al.* Atypical ductal hyperplasia is a multipotent precursor of breast carcinoma. *J. Pathol.* path.5262 (2019) doi:10.1002/path.5262.
19. Kim, S. Y. *et al.* Genomic differences between pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast cancer. *Oncotarget* **6**, 7597–7607 (2015).
20. Pareja, F. *et al.* Whole-Exome Sequencing Analysis of the Progression from Non–Low-Grade Ductal Carcinoma *In Situ* to Invasive Ductal Carcinoma. *Clin. Cancer Res.* 1078-0432.CCR-19–2563 (2020) doi:10.1158/1078-0432.CCR-19-2563.
21. Strand, S. H. *et al.* Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of TBCRC 038 and RAHBT cohorts. *Cancer Cell* (2022) doi:10.1016/j.ccell.2022.10.021.
22. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum. Genomics* **4**, 406–410 (2010).
23. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

24. Steiert, T. A. *et al.* A critical spotlight on the paradigms of FFPE-DNA sequencing. *Nucleic Acids Res.* **51**, 7143–7162 (2023).
25. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).
26. Guo, Q. *et al.* The mutational signatures of formalin fixation on the human genome. 2021.03.11.434918 Preprint at https://doi.org/10.1101/2021.03.11.434918 (2021).
27. Yuan, T. & Cantley, L. PI3K pathway alterations in cancer: variations on a theme. *Oncogene* **27**, 5497–5510 (2008).
28. Bader, A. G., Kang, S., Zhao, L. & Vogt, P. K. Oncogenic PI3K deregulates transcription and translation. *Nat. Rev. Cancer* **5**, 921–929 (2005).
29. Kader, T. *et al.* Predictive biomarkers of breast ductal carcinoma in situ may underestimate the risk of recurrence due to de novo ipsilateral breast carcinoma development. 2024.05.19.594731 Preprint at https://doi.org/10.1101/2024.05.19.594731 (2024).
30. Pang, J.-M. B. *et al.* Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* **30**, 952–963 (2017).
31. Groen, E. J. *et al.* Prognostic value of histopathological DCIS features in a large-scale international interrater reliability study. *Breast Cancer Res. Treat.* **183**, 759–770 (2020).
32. Rakovitch, E. *et al.* HER2/neu and Ki-67 expression predict non-invasive recurrence following breast-conserving therapy for ductal carcinoma in situ. *Br. J. Cancer* **106**, 1160–1165 (2012).
33. Miligy, I. M. *et al.* The clinical and biological significance of HER2 over-expression in breast ductal carcinoma in situ: a large study from a single institution. *Br. J. Cancer* **120**, 1075–1082 (2019).
34. Ang, D. C. *et al.* Frequent phosphatidylinositol-3-kinase mutations in proliferative breast lesions. *Mod. Pathol.* **27**, 740–750 (2014).
35. Silwal-Pandit, L., Langerød, A. & Børresen-Dale, A.-L. TP53 Mutations in Breast and Ovarian Cancer. *Cold Spring Harb. Perspect. Med.* **7**, a026252 (2017).
36. Aggarwal, V., Montoya, C. A., Donnenberg, V. S. & Sant, S. Interplay between tumor microenvironment and partial EMT as the driver of tumor progression. *iScience* **24**, 102113 (2021).
37. Deng, Y., Chen, G., Xiao, J. & Deng, H. Role and potential therapeutic strategies of matrix mechanics for optimizing tumor radiotherapy. *Mechanobiol. Med.* **2**, 100037 (2024).
38. Prakash, J. & Shaked, Y. The Interplay between Extracellular Matrix Remodeling and Cancer Therapeutics. *Cancer Discov.* **14**, 1375–1388 (2024).
39. Frascogna, C. *et al.* Role of the mechanical microenvironment on CD-44 expression of breast adenocarcinoma in response to radiotherapy. *Sci. Rep.* **14**, 391 (2024).
40. Rakovitch, E. *et al.* Can we select individuals with low risk ductal carcinoma in situ (DCIS)? A population-based outcomes analysis. *Breast Cancer Res. Treat.* **138**, 581–590 (2013).
41. Rakovitch, E. *et al.* A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. *Breast Cancer*

Res. Treat. **152**, 389–398 (2015).
42. Rakovitch, E. *et al.* Multigene Expression Assay and Benefit of Radiotherapy After Breast Conservation in Ductal Carcinoma in Situ. *J. Natl. Cancer Inst.* **109**, djw256 (2017).
43. Paszat, L., Sutradhar, R., Zhou, L., Nofech-Mozes, S. & Rakovitch, E. Including the Ductal Carcinoma-In-Situ (DCIS) Score in the Development of a Multivariable Prediction Model for Recurrence After Excision of DCIS. *Clin. Breast Cancer* **19**, 35–46 (2019).
44. Klein, J. *et al.* Close or positive resection margins are not associated with an increased risk of chest wall recurrence in women with DCIS treated by mastectomy: a population-based analysis. *SpringerPlus* **4**, 335 (2015).
45. Lalani, N. *et al.* Long-term outcomes of hypofractionation versus conventional radiation therapy after breast-conserving surgery for ductal carcinoma in situ of the breast. *Int. J. Radiat. Oncol. Biol. Phys.* **90**, 1017–1024 (2014).
46. Rakovitch, E. *et al.* DUCHESS: an evaluation of the ductal carcinoma in situ score for decisions on radiotherapy in patients with low/intermediate-risk DCIS. *Breast Cancer Res. Treat.* **188**, 133–139 (2021).
47. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).
49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org* https://arxiv.org/abs/1303.3997v2 (2013).
50. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
51. Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).
52. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. Preprint at https://doi.org/10.1101/861054 (2019).
53. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
54. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
55. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
56. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
57. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311 (2012).
58. Sahraeian, S. M. E. *et al.* Achieving robust somatic mutation detection with deep learning models derived from reference data sets of a cancer sample. *Genome Biol.* **23**, 12 (2022).
59. Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S. & Predeus, A. V. Systematic

benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics* **23**, 155 (2022).
60. de Schaetzen van Brienen, L. *et al.* Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples. *BMC Med. Genomics* **13**, 94 (2020).
61. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
62. Shyr, C. *et al.* FLAGS, frequently mutated genes in public exomes. *BMC Med. Genomics* **7**, 64 (2014).
63. Alfieri, F., Caravagna, G. & Schaefer, M. H. Cancer genomes tolerate deleterious coding mutations through somatic copy number amplifications of wild-type regions. *Nat. Commun.* **14**, 3594 (2023).
64. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
65. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
66. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
67. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
68. Chang, M. T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov.* **8**, 174–183 (2018).
69. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. *Cancer Discov.* **14**, 49–65 (2024).
70. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
71. Lesluyes, T., Tarabichi, M., Haase, K., Demeulemeester, J. & Loo, P. *Robust and Platform-Independent CNA Calling with ASCAT V3*. (2023). doi:10.13140/RG.2.2.18625.12641.
72. Franch-Expósito, S. *et al.* CNApp, a tool for the quantification of copy number alterations and integrative analysis revealing clinical implications. *eLife* **9**, e50267 (2020).
73. Renault, V. *et al.* aCNViewer: Comprehensive genome-wide visualization of absolute copy number and copy neutral variations. *PloS One* **12**, e0189334 (2017).

Acknowledgements

This work was supported by Canadian Institute for Health Research (CIHR) project grant #391682. We thank the Canadian Foundation for Innovation J. Evans Leaders Fund grant #43481 for the support of the computing infrastructure used throughout our analyses.

ER holds LC Campbell Chair for breast cancer research. NR received Breast Cancer Canada and Ontario Graduate Scholarships.

We thank Genome Quebec Innovation Center for RNA processing and sequencing.

Ethics declaration

The authors declare no competing interests.