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

Corresponding author: Vanessa Dumeaux (vdumeaux@uwo.ca)

20 **Abstract**

21 **Background**

22 Managing Ductal Carcinoma in Situ (DCIS) remains challenging due to the lack of reliable
23 biomarkers to predict radiotherapy (RT) response, leading to both overtreatment of indolent
24 disease and undertreatment of aggressive cases.

25 **Results**

26 Through whole-exome sequencing of 147 DCIS cases, we characterized the genomic
27 landscape of pure DCIS and identified genetic alterations associated with the risk of
28 recurrence, either in-situ or invasive. DCIS lesions harbored frequent mutations in established
29 cancer drivers (*PIK3CA*, *TP53*) and genes regulating tissue architecture, which likely enhanced
30 pre-invasive cell fitness but lacked prognostic value. A subset of younger patients exhibited
31 distinct mutational processes characterized by high mutational burden, though these were not
32 linked to recurrence risk. Across the cohort, five mostly mutually exclusive genes (*SH2B2*,
33 *PDZD8*, *MYO7A*, *MUCL3*, *DNASE2B*), involved in cell adhesion, membrane organization, and
34 DNA degradation, were significantly associated with 10-year risk of local recurrence. In RT-
35 treated patients, we identified 27 additional mutated genes uniquely associated with
36 recurrence, along with *SH2B2* and *MUCL3*. Most of these genes were involved in cytoskeletal
37 regulation, cell adhesion, and cell-environment interactions. Mutations in metabolic regulators
38 (*MGAM2* and *AADACL3*) and *REV1*, which mediates DNA damage tolerance, may impair cellular
39 responses to RT-induced stress. Notably, we identified distinct genes prognostic for in-situ
40 versus invasive recurrence: nineteen genes predominantly involved in tissue structural
41 maintenance in in-situ relapse, and thirteen genes primarily affecting cell-cycle and genome-

stability pathways in invasive progression. Copy number analyses revealed that pure DCIS exhibits molecular subtype-specific patterns characteristic of invasive disease, with novel alterations associated with recurrence, including three non-adjacent gains and five losses in regions harboring oncogenes, tumor suppressor genes, and genes regulating structural integrity, cell-cell adhesion and interactions.

Conclusions

While *TP53*, *PIK3CA*, and recurrent copy number alterations represent early events in tumorigenesis, they lack prognostic value in pure DCIS, underscoring the need for alternative biomarkers. Our findings identify key genetic alterations associated with local recurrence and RT resistance. We further uncovered distinct molecular programs underlying in-situ versus invasive recurrence, with mutations affecting tissue structural maintenance in in-situ relapse and cell-cycle/genome-stability pathways in invasive progression.

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 identify such cases could enable tailored treatments, such as full mastectomy or adjuvant systemic therapies for patients at risk of in situ or invasive local recurrence (LR) despite receiving RT [8], or more frequent watch-and-wait strategies for those 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]. Previous studies have revealed important insights into DCIS progression by highlighting specific mutations and chromosomal alterations that may influence the progression towards invasive disease [10–17]. 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 [10].

84 *TP53* and *PIK3CA* [11–17] are among the most frequently observed mutations in these studies
85 [11–17] and are often characterized as key drivers of tumorigenesis, promoting growth and
86 spread of cancer cells. However, this may not necessarily be the case, as mutant field
87 clonalization could equip pre-malignant cells with additional (epi-) genetic mutations that
88 confer fitness advantages, allowing them to expand into ductal regions without directly driving
89 invasive behavior [10]. This suggests that driver mutations, alongside frequently mutated
90 genes, may serve as these early fitness enhancers rather than drivers of progression. Several
91 copy number aberrations (CNAs) have also been identified by these studies [10–15,17–20]
92 including frequent gains at 1q, 8q, 11q, and 17q, and losses at 16q, however the specific
93 associations of these CNAs to DCIS prognosis remain unclear.

94 To address these critical gaps, we conducted comprehensive exome sequencing analysis of
95 147 pure DCIS cases, including patients treated with and without RT, to investigate markers of
96 local recurrence within 10 years of diagnosis. Our study characterizes the mutational
97 landscape in pure DCIS and identifies novel genomic alterations associated with tumor grade,
98 molecular subtypes, and patient age. Most importantly, we discovered specific variants and
99 CNAs predictive of local recurrence risk, including genetic markers associated with RT
100 response. These findings provide insights into the genomic determinants of DCIS prognosis
101 and treatment response, establishing a foundation for improved risk stratification and
102 personalized treatment strategies for DCIS patients.

Results

A unique cohort of pure DCIS patients

We assembled a cohort of 147 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 (**Table 1**). The study design was balanced to include at least a third of patients who experienced an ipsilateral invasive or in-situ local recurrence within a 10-year follow-up period and about half received radiotherapy as part of their standard-of-care (**Table 1**). In clinical settings, RT is omitted in some patients with low-risk features of DCIS or due to patient preference. The median time to recurrence was 4.2 years for invasive disease and 2.1 years for in-situ disease. Most tumors were of intermediate to high grade spanning all five molecular subtypes, with normal-like and luminal A subtypes more frequently observed in patients without local recurrence within 10 years (**Table 1**). A minority of tumors exhibited multifocality (24.5%) and positive margins (6.8%). Clinical characteristics were comparable between RT-treated and untreated patients across all variables except age, with women aged 60 years or older less likely to receive RT (**Supplementary Table 1**).

		Without Local Recurrence (n = 96)	**With Local Recurrence (n = 51)**	**p-value[a]**
Type of 10-year Local Recurrence	DCIS	0	24 (47.0%)	
	Invasive	0	27 (53.0%)	
Time to Recurrence in years Median (Range)	DCIS	NA	2.1 (0.5 – 9.7)	
	Invasive	NA	4.2 (0.6 – 9.7)	

Radiotherapy (RT) N (%)	Yes	51 (53.1%)	22 (43.1%)	0.32
	No	45 (46.9%)	29 (56.9%)	
Clear Margins N (%)	Positive	5 (5.2%)	5 (9.8%)	0.57
	Negative	79 (82.3%)	40 (78.4%)	
	Undetermined	12 (12.5%)	6 (11.8%)	
Tumor size (mm), Median (Range)		16.3 (4.0 - 90.0)	21.3 (2.0 - 76.0)	0.05
Nuclear Grade N (%)	Low	2 (2.1%)	2 (3.9%)	0.31
	Moderate	61 (63.5%)	26 (51.0%)	
	High	33 (34.4%)	23 (45.1%)	
Multifocality N (%)	Present	19 (19.8%)	17 (33.3%)	0.14
	Absent	48 (50.0%)	24 (47.1%)	
	Undetermined	29 (30.2%)	10 (19.6%)	
Age N (%)	<50 Years Old	37 (38.5%)	15 (29.4%)	0.11
	50-60 Years Old	34 (35.4%)	14 (27.5%)	
	>60 Years Old	25 (26.1%)	22 (43.1%)	
PAM50 Subtype (Pearson Correction) N (%)	Basal-like	12 (12.5%)	9 (17.6%)	0.01
	Her2-like	9 (9.4%)	9 (17.6%)	
	LumA	27 (28.1%)	5 (9.8%)	
	LumB	9 (9.4%)	12 (23.5%)	
	Normal-Like	20 (20.8%)	5 (9.8%)	
	Undetermined	19 (19.8%)	11 (21.6%)	

[a] Chi-square and t-test statistics for categorical and quantitative variables, respectively

121 **Table 1. Patient and tumor clinical attributes**

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

124 The mutational load of pure DCIS lesions varied considerably, ranging from. 3 to 3,482 non-
125 synonymous variants per sample (median 75) The majority (74.2%) were missense mutations,
126 constituting approximately 52K unique variants identified in ~14K genes. Most of these genes
127 were not frequently mutated, with 2,030 genes (14.5%) harboring non-synonymous aberrations
128 in at least 5 patients.

129 As expected, most mutations are of C>T type (barplot in **Fig. 1**), a common mutational pattern
130 attributed to the spontaneous deamination of 5-methylcytosine, a process frequently observed
131 in many cancer genomes [21,22]. Spontaneous deamination can be exacerbated in FFPE
132 samples due to DNA damage [23]. FFPE-related mutational artefacts are known to resemble
133 certain COSMIC signatures [24], such as SBS30 and SBS1 [25]. During library preparation,
134 formalin-induced DNA lesions are chemically repaired with unrepaired profiles resembling
135 SBS30 and repaired profiles resembling SBS1 [25]. While SBS30 was not identified in our
136 dataset, SBS1 was detected in a large number of samples (**Fig. 1)**. SBS1, commonly found in
137 tumor genomes, is difficult to distinguish from repaired FFPE-related artifacts due to their high
138 similarity [22,25]. However, previous research has demonstrated that mutational profiles from
139 repaired FFPE samples closely match true tumor mutational profiles [25], and we expect that
140 our careful processing of mutation calling-pipeline minimizes the impact of these artefacts
141 [23].

142 Interestingly, a subset of DCIS samples exhibit a significantly higher tumor mutational burden
143 (TMB), characterized by an increase in frequency of C>G, T>C, and C>A mutations, and

144 enrichment of specific SBS signatures including SBS26, SBS7b, and SBS5 (**Fig. 1**, 12 samples

145 on the right). The causes underlying these mutational signatures remain incompletely

146 understood. However, SBS26 has been linked to impaired mismatch repair and microsatellite

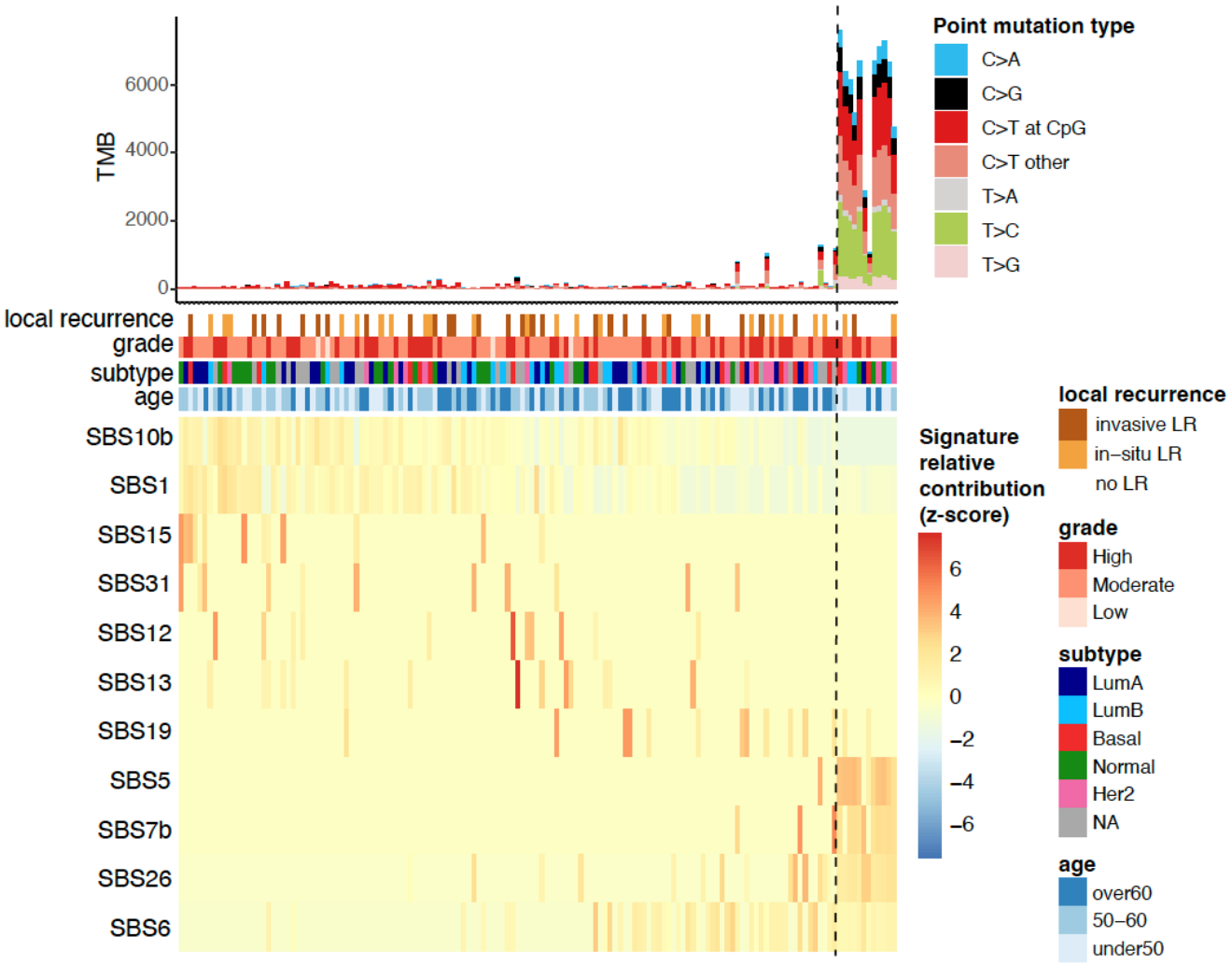
147 instability. These cases were not more likely to have a recurrence, of a higher grade or of a

148 specific molecular subtype but were predominantly found in younger patients (8 out of 12

149 under 50 years; Chi-square test p-value < 0.05).

150 These findings highlight distinct mutational processes in some early-onset DCIS cases, though

151 these molecular features were not associated with prognosis.



152

Fig. 1: Mutational Signatures in pure DCIS samples.

The top panel shows tumor mutational burden (TMB) with distribution of point mutation types. Clinical annotations display 10-year local recurrence outcome, tumor grade, molecular subtype, and patient age. The bottom heatmap depicts the row-scaled relative contribution of single base substitution (SBS) signatures detected in at least 10 samples (z-score). Samples (columns) and signatures (rows) are ordered based on a rank-sum statistic that maximizes the coherence of contribution patterns. The method first splits signatures into two groups using medoid clustering. For each signature, relative contributions are ranked within each group, and the final sample ordering is determined by the average rank-sum across all signatures.

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

175 controlling epithelial-components including cellular morphology, epithelial function and
176 adhesion are central to the physiopathology of pure DCIS.

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

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

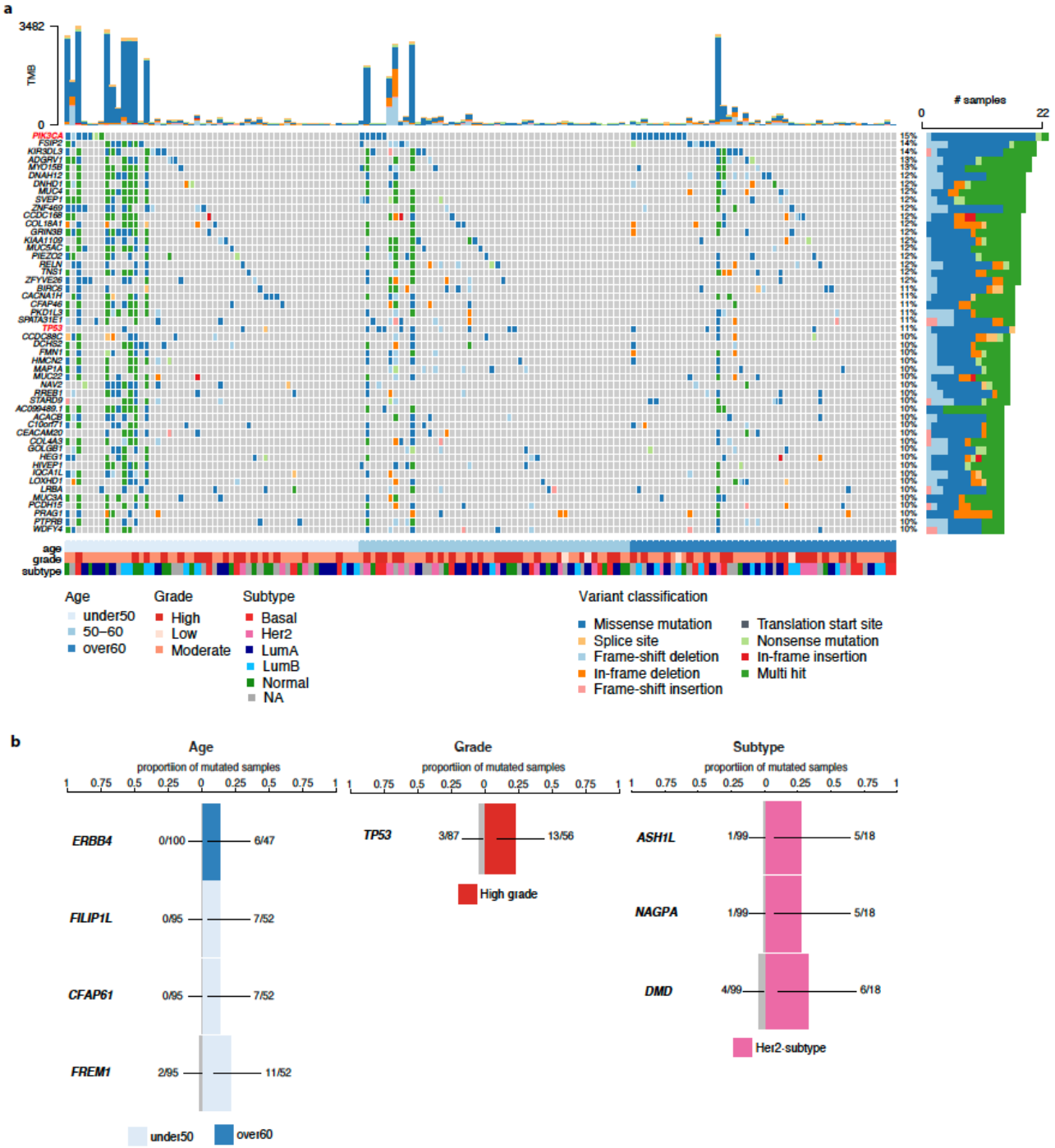


Fig. 2: The most frequently mutated genes and association with clinico-pathological variables in pure DCIS patients.

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

To identify potential cancer driver genes, we analyzed the ratio of non-synonymous to synonymous mutations (dN/dS) across all genes, which can indicate positive selection of mutations that provide growth advantages to cancer cells [26]. This analysis identified two significantly mutated 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). While PIK3CA mutations were not enriched in specific molecular subtypes, TP53 mutations were significantly more frequent in basal-like and HER2-enriched subtypes compared to other subtypes (19% in basal-like [4/21], 33% in HER2-enriched [6/18] vs 5% in other subtypes [4/78], Chi-square test p < 0.005).

Out of 34 samples carrying a mutation in at least one of these driver genes, 15 samples harbored at least one potentially actionable alteration as indicated by OncoKB (**Supplementary**

208 **Fig. 1**). The mutation spectrum in *PIK3CA* was dominated by the activating H1047R hotspot
209 mutation (~50%) in the kinase domain of exon 21, known to enhance PI3K-mitigated pathway
210 signaling (**Supplementary Fig. 1a**) [27,28]. Similarly, *TP53* mutations clustered in known
211 hotspots within the DNA-binding domain, which may result in the loss of tumor suppression by
212 affecting its ability to bind to DNA (**Supplementary Fig. 1b**).

213 Despite their established roles in cancer progression, none of these driver mutations was
214 associated with 10-year local recurrence risk (Firth's penalized likelihood Cox regression p >
215 0.5), suggesting that additional factors influence DCIS prognosis.

216

217 ***Identification of mutations associated with increased local recurrence risk independent of***
218 ***treatment***

219 To identify potential prognostic markers, we analyzed mutated genes associated with the risk
220 of invasive or in-situ local recurrence in the ipsilateral breast occurring between 6 months and
221 10 years after diagnosis. Our survival analysis identified five biomarkers each significantly
222 associated with increased 10-year local recurrence risk (Firth's penalized likelihood Cox
223 regression p-value < 0.01 & permuted p-value < 0.05; **Fig. 3a-b**). These mutations, occurring in
224 4-7% of cases, were largely mutually exclusive and occurred across molecular subtypes and
225 grades.

226 The presence of mutations in at least one of these genes was significantly associated with
227 increased recurrence risk (**Fig. 3c,** log-rank p-value p<0.0001), with *MYO7A* and *PDZD8*
228 showing the strongest associations (Firth's penalized likelihood Cox regression HR>4.4, **Fig.**

229 **3b**). Gene Ontology analysis revealed that most genes are involved in multiple interrelated

230 cellular processes including cell adhesion (*MYO7A*, *SH2B2*), organization and function of the

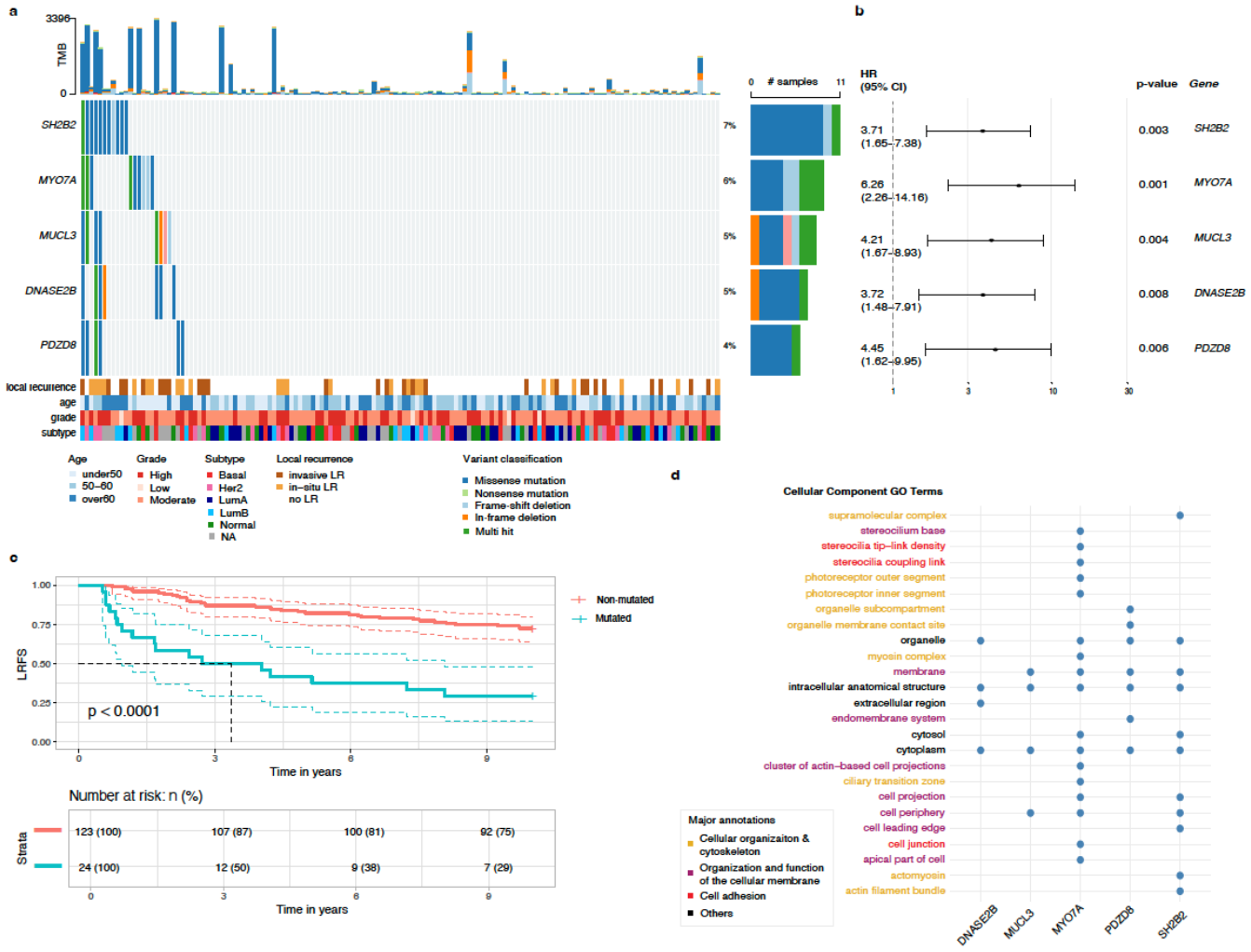
231 cellular membrane (*PDZD8*, *MUCL3*, *MYO7A*), and cellular organization and cytoskeleton

232 (*MYO7A*, *PDZD8*) (**Fig. 3d**). Additionally, *DNASE2B*, a member of the DNase II family of

233 endonucleases, was identified among the significant genes. These findings underscore how

234 cytoskeletal reorganization, changes in cell structure, and compromised cell adhesion might

235 contribute to increased risk of recurrence within 10 years after a DCIS diagnosis.



236

237 **Fig. 3: Gene variants associated with an increased 10-year local recurrence risk in DCIS.**

238 **a** Samples are in columns and variants are color-coded based on their classification. Local
239 recurrence status and clinico-pathological characteristics of each lesion are depicted at the
240 bottom. Local recurrence is defined as any recurrence, in situ or invasive, in the ipsilateral
241 breast occurring between 6 months and 10 years after diagnosis. Right-hand side reports the
242 proportion of each variant classification type. **b** Hazard ratio and confidence intervals for each
243 significant mutated gene associated with increased 10-year local recurrence risk (Firth's
244 penalized likelihood Cox regression). **c** Kaplan-Meier analysis of local recurrence-free survival
245 (LRFS) comparing patients with mutations in at least one of the 5 genes (red) versus those
246 without mutations (blue). **d** Cellular component Gene Ontology (GO) terms annotations for
247 each significant gene.

248

249 ***Mutations in genes governing cytoskeletal organization and membrane dynamics associated***
250 ***with radiotherapy resistance***

251 To identify predictive biomarkers for RT response, we analyzed mutations associated with 10-
252 year local recurrence in a cohort restricted to patients who received RT. This analysis revealed
253 29 genes significantly associated with an increased risk of recurrence (Firth's penalized
254 likelihood Cox regression p-value < 0.05 & permuted p-value < 0.05; **Fig. 4a**). Notably, these
255 mutations often co-occurred, with at least two mutated genes present in approximately 27% of
256 lesions (20/73 patients treated with RT).

257 In RT-treated patients, lesions harboring mutations in at least two of these genes exhibited a
258 markedly increased risk of local recurrence (Firth's penalized likelihood Cox regression HR =
259 4.8, 95% CI: 2.1–11.2, p = 0.0002), with most mutation-positive recurrences occurring within 5

260 years post-RT (**Fig. 4b**). Notably, while mutations in *SH2B2* and *MUCL3* were significantly

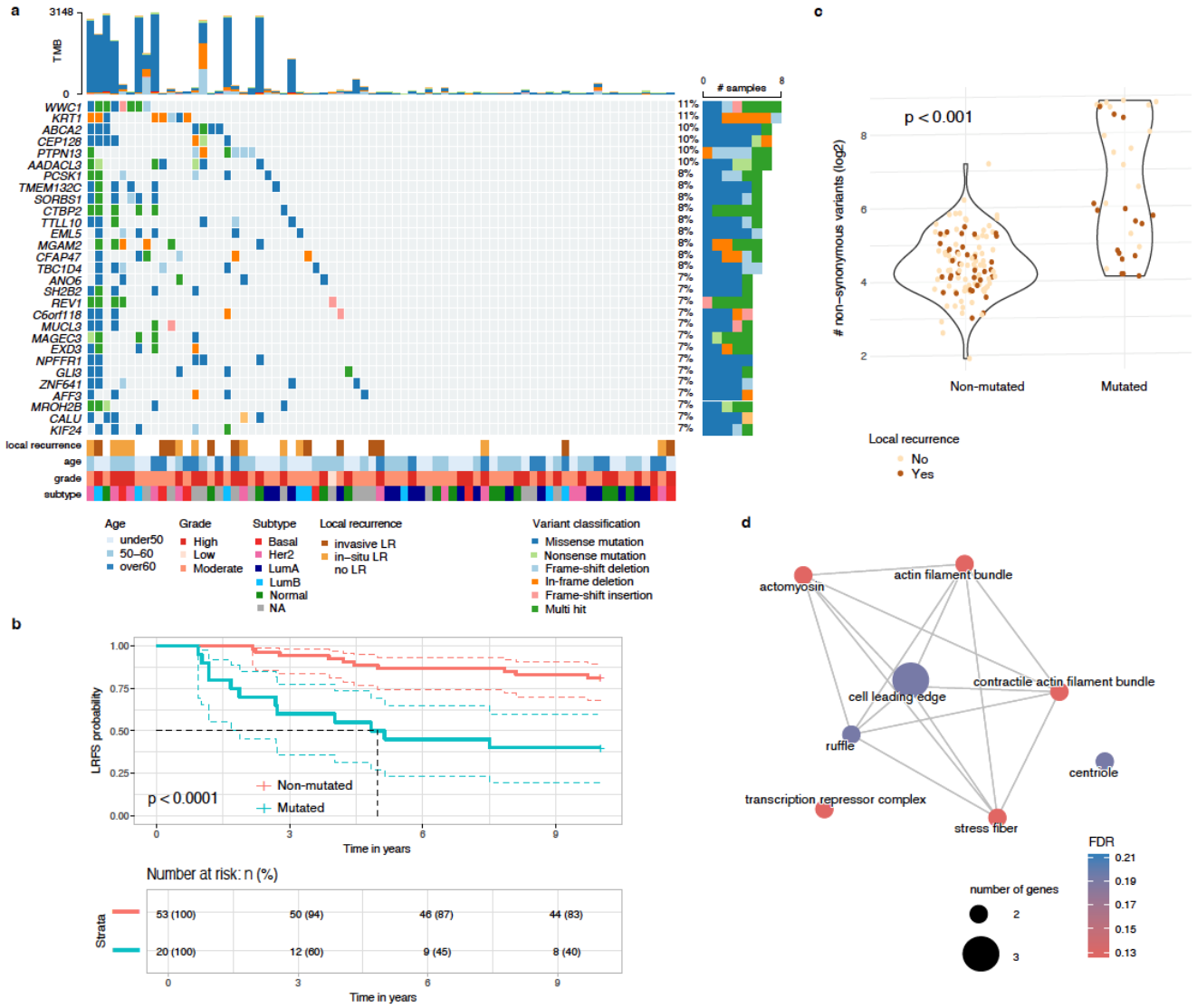
261 associated with prognosis in both the overall cohort and the RT-treated subgroup, an RT-

262 stratified analysis revealed that lesions harboring two mutations in the remaining 26 genes

263 (which were altered in at least five lesions in the no-RT group) were predictive of prognosis

264 only in RT-treated patients ($p < 0.0001$) and not in those who did not receive RT ($p > 0.5$)

265 (**Supplementary Fig. 2**).



266

267 **Fig. 4: Mutations associated with radiotherapy resistance in DCIS.**

268 **a** An oncoplot showing the distribution of mutations in 29 genes significantly associated with
269 local recurrence in RT-treated patients. Color-coding indicates mutation types; clinical
270 annotations show age, grade, molecular subtype, and 10-year recurrence status. **b** Kaplan-
271 Meier analysis of LRFS comparing patients with mutations in at least two of the 29 genes (red)
272 versus those without mutations (blue). **c** Distribution of TMB, shown as number of non-
273 synonymous variants, log2 scale) for lesions with at least one mutation in two genes
274 associated with LRFS colored by recurrence status at 10 years. **d** GO cellular component
275 enrichment network analysis of recurrence-associated genes in RT-treated DCIS. Network
276 visualization shows enriched GO terms (FDR < 0.2) from genes linked to local recurrence
277 following RT. Nodes represent individual GO terms with size proportional to gene count (>2),
278 and edges indicate significant semantic similarity between terms. Node color intensity
279 corresponds to enrichment significance.

280

281 Given the frequent co-occurrence of these mutations, we assessed whether overall TMB might
282 explain the increased recurrence risk. Although several high-TMB lesions were observed
283 among patients with these mutations (**Fig. 4a,c**), TMB itself was not significantly associated
284 with local recurrence risk (Firth's penalized likelihood Cox regression p = 0.8). Indeed, many
285 lesions with high TMB remained recurrence-free over 10 years, suggesting that specific
286 mutations—not overall mutation load—drive LR after RT. We further examined potential
287 confounding by clinico-pathological variables (i.e. age, grade, tumor size, multifocality and
288 subtype) by adjusting our survival model for each factor. The predictive value of these

289 mutations remained significant (Firth's penalized likelihood Cox regression p-values < 0.005),
290 supporting their role as an independent risk factor for recurrence following RT.

291 Functional analysis revealed a network of interrelated cellular functions involved in actin
292 cytoskeleton regulation, cellular polarity, and membrane dynamics—processes essential for
293 invasive behavior (**Fig. 4d**).

294 Alterations in *SH2B2* and *SORBS1*, which modulate stress fiber formation and actin
295 remodeling, and in *PTPN13* and *WWC1*, which affect actomyosin tension at the cell leading
296 edge, suggest disruptions in cytoskeletal dynamics. Additional genes associated with
297 prognosis in RT-treated include *KRT1*, *MUCL3*, *TMEM132C*, *TTLL10*, *EML5*, *CEP128*, *CFAP47*,
298 and *KIF24*, all of which contribute to maintaining cellular architecture, polarity, and adhesion
299 (**Fig. 4a**). In addition, alterations in metabolic regulators—such as *MGAM2* and *AADACL3*—
300 could influence the energetic and biosynthetic demands required in response to RT-induced
301 stress, while *REV1* is implicated in DNA damage tolerance, a critical safeguard following
302 genotoxic stress.

303 Collectively, these findings highlight the pivotal roles of cytoskeletal reorganization, membrane
304 remodeling, metabolic regulation, and DNA repair pathways in the response to RT. The
305 prognostic significance of these mutations in RT-treated patients suggests that these gene
306 alterations may mediate resistance mechanisms triggered by RT-induced stress.

307 ***Differential mutational associations with in-situ versus invasive local recurrence after***
308 ***radiotherapy***

309 To investigate whether individual mutations are preferentially linked to either in-situ or invasive
310 LR, we refitted the gene-based survival models using each LR type as the sole endpoint (the
311 alternative event was censored) and stratified the analyses by RT. For the full cohort we had
312 retained genes with $p < 0.01$ and permutation $p < 0.05$; for the smaller RT-treated group we
313 used $p < 0.05$ and permutation $p < 0.05$. As in the main analysis, many significant associations
314 were found in RT-treated patients, with some overlap with significant genes found in the whole
315 cohort (**Fig. 5a**).

316 Three genes—*MUCL3*, *SH2B2* and *C6orf118*—were associated with outcome in at least three
317 out of six analyses (**Fig. 5a**). In RT-treated tumours, fifteen genes were associated with in-situ
318 LR or with "any LR", and two genes (*CFAP47*, *KRT1*) were associated with invasive or with "any
319 LR" (**Fig. 5a**). Nineteen additional genes were uniquely associated with in-situ LR (**Fig. 5b**),
320 whereas thirteen were unique to invasive LR (**Fig. 5c**) within the RT subgroup. The Kaplan-
321 Meyer analyses demonstrate that patients whose lesions harbor mutations in at least two
322 genes from either the in-situ-specific or invasive-specific gene sets show significantly worse
323 LR-free survival only for the corresponding recurrence type (**Fig. 5d,e**).

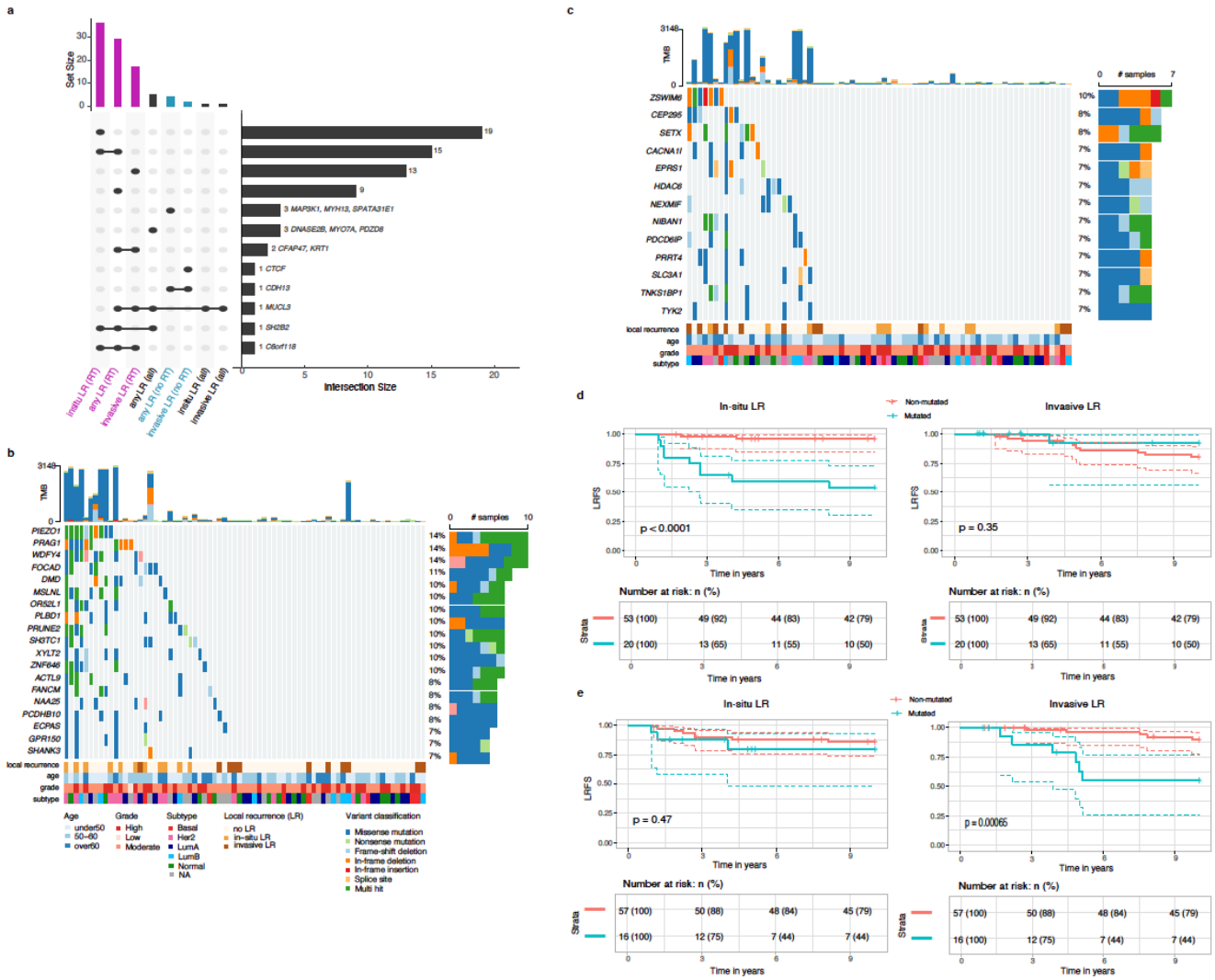


Fig. 5: Differential mutational associations with in-situ versus invasive local recurrence after radiotherapy

a An UpSet plot showing the number of genes significantly associated with risk of invasive, in-situ or any local recurrence (LR) in the whole cohort (all) and in RT-treated (RT) or untreated patients (noRT), with intersections between gene lists indicated by connected dots below and corresponding size of the intersection depicted in the barplot on the right. **b** Oncoplot showing mutations in genes uniquely associated with in-situ LR (n=19) in RT-treated patients. Color-coding indicates mutation types; clinical annotations show age, grade, molecular subtype, and

333 recurrence status. **c** Similar to b but showing mutations in genes uniquely associated with
334 invasive LR (n=13) in RT-treated patients. **d** Kaplan-Meier analysis of in-situ (left) or invasive
335 (right) LR-free survival comparing patients with mutations in at least two of the 19 genes
336 uniquely associated with in-situ LR (red) versus those without mutations (blue). **e** Similar to d
337 but comparing in-situ or invasive LR-free survival between patients with mutations in at least
338 two of the 13 genes uniquely associated with invasive LR versus those without mutations.

339 Functional annotation indicates that mutations uniquely linked to in-situ LR occur mainly in
340 genes that preserve epithelial architecture and mechanosensing: anchoring-junction
341 components that couple cells to neighbouring cells or to the extracellular matrix (*DMD*,
342 *FOCAD*), the stretch-activated channel *PIEZO1*, and scaffolds/adaptors localised to actin-rich
343 membrane projections such as invadopodia (*SHANK3*, *PRAG1*). The set also includes WDFY4, a
344 WD-repeat/FYVE-domain autophagy adaptor implicated in MHC-II antigen presentation.

345 Conversely, mutations associated with invasive LR after RT mapped to genes involved in late
346 cytokinetic abscission (*PDCD6IP*), centriole-to-centrosome maturation (*CEP295*), DNA-damage
347 sensing and repair (*SETX*, *TNKS1BP1*), calcium-regulated motility (*CACNA1I*) and cytokine-
348 dependent inflammatory signalling (*TYK2*).

349 Because stratifying simultaneously by treatment and LR type reduces sample size and event
350 numbers, we cannot exclude the possibility that some endpoint-specific associations—or the
351 absence of others—reflect differences in statistical power. Nevertheless, these findings
352 suggest that distinct biological programs—structural maintenance versus cell-cycle, genome-
353 stability and inflammatory pathways—underlie in-situ and invasive patterns of recurrence after
354 RT.

355

356 ***Copy number alterations display molecular subtype-specific patterns with select genomic***

357 ***regions linked to local recurrence***

358 Analysis of copy number alterations (CNAs) revealed recurrent chromosomal changes similar

359 to those reported in invasive breast cancer (**Fig. 6a**). Significant gains were identified on

360 chromosomal arms 1q, 8q, 16p, 17q, 20p, and 20q, while losses predominantly occurred on 8p,

361 9p, 11q, 13q, 14q, 16q, and 17p (binomial test, FDR < 0.05).

362 Global CNA burden varied considerably across samples, with approximately half showing

363 minimal alterations (global CNA score < -0.58; **Fig. 6b**). Basal-like tumors exhibited greater

364 CNA burden, while normal-like tumors showed fewer alterations (**Fig. 6b**, **Supplementary Fig.**

365 **2a**). Similarly, high-grade lesions showed a greater CNA burden compared to low-grade lesions

366 (**Supplementary Fig. 3b)**.

367 We identified distinct CNA patterns across molecular subtypes: basal-like tumors showed

368 enrichment for gains on 8q, 13q, and 19q; LumA tumors frequently exhibited 16q loss; and

369 Her2-enriched tumors showed characteristic 17q12 gains corresponding to the *ERBB2* locus

370 (Fisher's exact test, p < 0.005; **Supplementary Fig. 3c**).

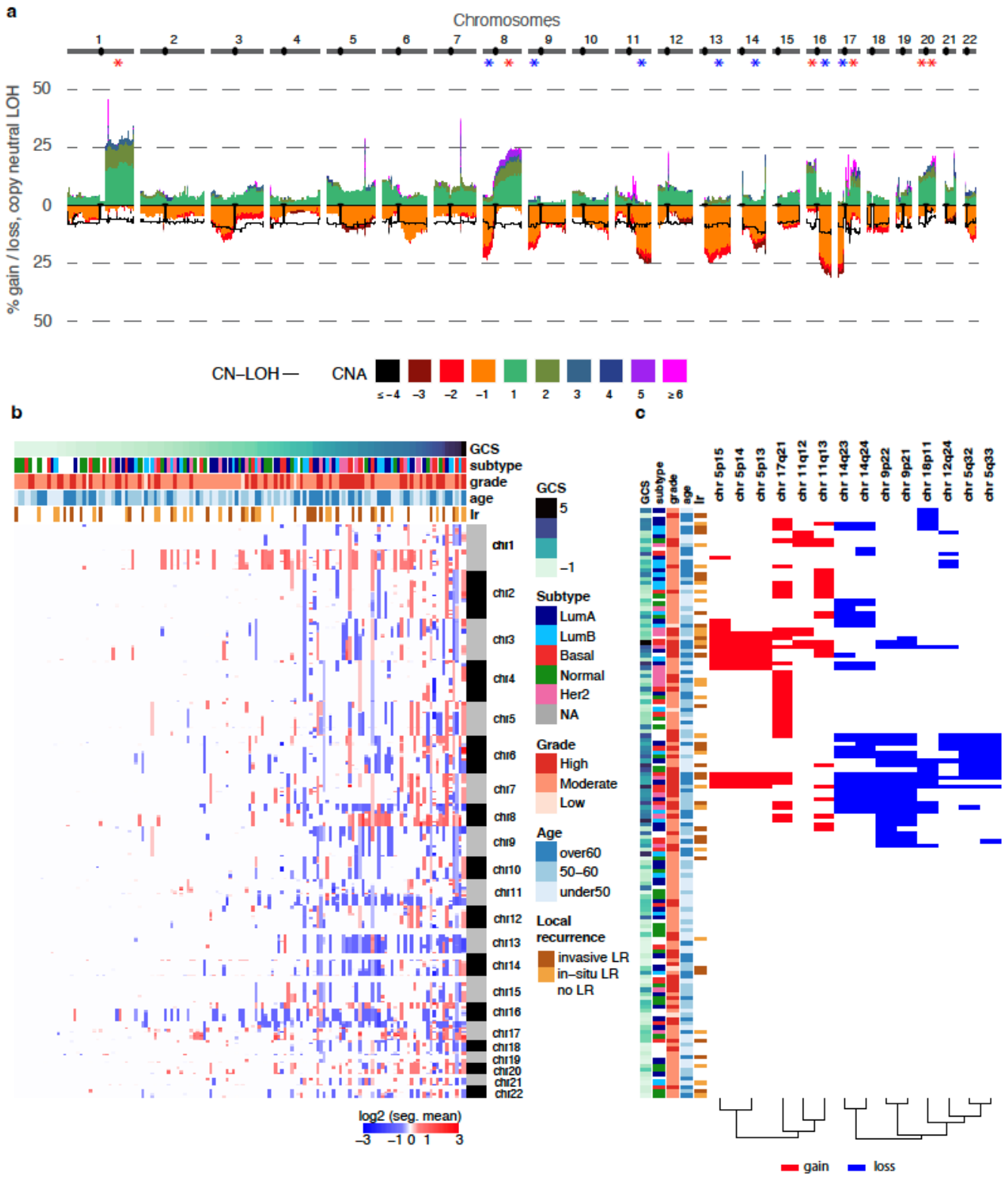


Fig. 6: Frequent copy number alterations in pure DCIS and associations with 10-year LR risk

373 **a** Genome-wide frequency of absolute copy number gains and losses across chromosomes 1-
374 22 (1Mb window). The estimated ploidy for each sample is subtracted from the copy number
375 values of each segment which means that a copy number of 0 is no copy number change. The
376 y-axis shows the percentage of samples with each alteration type. Asterisks indicate statistical
377 significance for chromosomal arm alterations **b** Heatmap showing copy number profiles per
378 cytoband across samples (rows), ordered by global CNA score (GCS). Sample annotations
379 include age, grade, molecular subtype and local recurrence status at 10 years. **c** Heatmap of
380 twenty genomic regions significantly associated with 10-year risk of local recurrence (Firth-
381 corrected cox model p-value < 0.05 adjusted for grade), showing copy number status. CNAs
382 present in at least 5 patients are shown, with copy number gains in red and losses in blue.
383 Samples and regions were clustered using Ward's hierarchical clustering with Minkowski
384 distance metric.

385

386 Fourteen cytoband regions were significantly associated with increased 10-year local
387 recurrence risk after grade adjustment (Firth's penalized likelihood Cox regression p-value &
388 permuted p-values < 0.05; **Fig. 6c**). These included copy number losses in five non-adjacent
389 cytobands and gains in three non-adjacent cytobands. When analyzing invasive local
390 recurrence specifically, four regions were shared (losses in 5q32-33, 9p21, 18p11), with two
391 additional losses identified in 9p13 and 8q11 and one gain in 5p15 (Firth's penalized likelihood
392 Cox regression p-value < 0.01 & permuted p-value < 0.05).

393 Copy number losses could alter critical tumor suppressor functions, as evidenced by the
394 presence of several well-known tumor suppressor genes within these regions. Notable

395 examples include *CDKN2A* at 9p21 (a key regulator of cell cycle progression), *PRDM4* at
396 12q23-24 (involved in cell differentiation and tumorigenesis) [29], *SPARC* at 5q33 (important
397 for extracellular matrix synthesis and modulation of cell shape) [30] and MITOSTATIN at
398 12q24 (a mitochondrial protein with tumor suppressor activity) [31]. Additional loss highlights
399 disruptions in cell adhesion and intercellular contacts as loss of 18p11 may affect DAL-1, a
400 known regulator of cell adhesion and link between cell membrane and cytoskeleton [32,33].

401 Gains were identified in regions harboring oncogenes or potential cancer-related genes. For
402 example, amplification in the 11q region—including 11q13, which contains *CCND1* and *CTTN*—
403 previously associated with breast cancer aggressiveness [34,35]. Additionally, a gain in 5p14
404 includes four cadherin genes (*CDH10*, *CDH9*, *CDH12*, *CDH18*) and *PRDM9*, whose aberrant
405 expression has been associated with genomic instability [36].

406 Collectively, these findings demonstrate that the CNAs associated with 10-year local
407 recurrence not only affect regions harboring established tumor suppressors and oncogenes
408 but also target regions involved in cell structural integrity and cell-cell interactions, potentially
409 contributing to an aggressive and treatment-resistant phenotype.

410

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

413 Most prior DCIS genomic studies analyzed small cohorts (<100 cases) and primarily focused
414 on DCIS cases with concurrent invasive disease (synchronous DCIS) rather than pure DCIS
415 (**Supplementary Table 2**). While recent larger studies, such as those by Strand *et al.* (2022) [20]

416 and Kader et al. (2024) [37], examined hundreds of pure DCIS cases, their reliance on low-pass
417 sequencing or lack of matched normal tissues limits the sensitivity for detecting genomic
418 changes, particularly when working with FFPE samples, which are often the only available
419 option for DCIS studies.

420 Across studies, *PIK3CA* and *TP53* consistently emerged as the most frequently mutated genes
421 in DCIS (and invasive disease), with mutation rates ranging from 21–55% and 17–52%,
422 respectively (**Supplementary Tables 2-3**). In our cohort, *PIK3CA* and *TP53* were also among
423 the most frequently mutated genes and identified as tumorigenesis drivers in pure DCIS.
424 Mutations in *GATA3* and *ERBB2,* which were frequently reported in several studies, were only
425 detected in a small number of lesions in our cohort (n = 3 and 7, respectively, **Supplementary**
426 **Table 3**). Overall, we observe very little overlap between genes reported as frequently mutated
427 across studies (**Supplementary Table 3**)._Differences across studies likely reflect variations in
428 methodology, including variant-calling pipelines, sample types (e.g., synchronous vs. pure
429 DCIS), cohort composition (e.g., histological grade and ER/HER2 status), and our study's
430 specific focus on LR and RT response, complicating direct comparisons. Nevertheless, the
431 consistent identification of *PIK3CA* and *TP53* as the most frequently mutated genes across
432 studies and in our cohort reinforces their pivotal roles in the early stages of breast
433 tumorigenesis.

434 Recurrent CNAs in regions such as 1q, 8q, and 17q gains, as well as 8p, 11q, and 16q losses,
435 were observed in our study, consistent with prior DCIS findings, including those from Strand *et*
436 *al.* (2022) [20] and Abba *et al.* (2015) [15]. For example, Strand *et al.* (2022) identified 29
437 recurrent CNAs in DCIS but found no single CNA predictive of recurrence [20]. In contrast, we

438 identified twenty genomic regions significantly associated with 10-year local recurrence risk,
439 including regions containing known tumor suppressors and oncogenes. These differences
440 may reflect the higher resolution of our sequencing approach (100x whole-exome sequencing)
441 compared to the low-pass sequencing used in prior studies. Our study also provides valuable
442 new insights into molecular subtype-specific CNA patterns in DCIS. In our cohort, basal-like
443 tumors exhibited gains on 8q, 13q, and 19q; Luminal A tumors displayed 16q loss; and Her2-
444 enriched tumors showed 17q12 gains encompassing *ERBB2*. Aside from amplifications in
445 HER2-positive subtypes, subtype-specific CNA analyses remain poorly represented in the DCIS
446 literature but are well established in invasive breast cancer (**see Supplementary Text**). These
447 findings suggest that genetic and molecular aberrations defining subtypes likely arise early
448 and are at least partially established at the DCIS stage.

449 Finally, while several studies have explored prognostic markers of recurrence, few account for
450 treatment variation, and none specifically examined markers of RT response (**Supplementary**
451 **Table 2**). Our study uniquely identified genetic alterations within a gene network that integrates
452 cytoskeletal integrity, cell-cell interactions, cell adhesion, and metabolism which are
453 associated with an increased risk of local recurrence within five years following RT. These
454 findings suggest that these alterations may play a critical role in resistance mechanisms
455 activated by RT-induced stress.

456 **Discussion**

457 Our study provides a comprehensive analysis of pure DCIS, focusing on DNA profiles
458 associated with disease prognosis and response to RT. While previous research has primarily
459 examined genomic changes in synchronous DCIS compared to IDC or pure DCIS, our work

460 offers a detailed view of the mutational landscape specific to pure DCIS and its association
461 with LR. Importantly, we identified genomic alterations associated with molecular subtypes
462 and, most critically, LR—both independent of treatment and specifically in cases following RT—
463 highlighting potential molecular drivers of treatment outcomes. We also observed distinct
464 mutational processes in early-onset DCIS with high tumor mutational burden potentially driven
465 by impaired mismatch repair, but these were not associated with prognosis.

466 While driver genes like *TP53* and *PIK3CA* are critical in the early stages of tumorigenesis, our
467 study did not find any significant associations between their mutations and disease
468 progression or response to treatment. These findings suggest that TP53 and PIK3CA
469 mutations are essential for overcoming initial biological constraints during early tumor
470 development but may become less relevant to progression once these barriers are surpassed.
471 This aligns with prior research showing that these mutations are frequently retained in invasive
472 carcinoma but occur at higher prevalence in hyperplastic and DCIS lesions [38].

473 Furthermore, their prognostic relevance appears to depend on tumor clinicopathological
474 characteristics. For example, Lin et al. [11] observed an inverse association between *PIK3CA*
475 kinase domain mutations in high-grade DCIS tumors and progression, while Silwal-Pandit et al.
476 [39] reported that the prognostic impact of *TP53* mutations varied across molecular subtypes
477 of invasive breast cancer. These findings underscore the complexity of breast cancer
478 progression, where early driver mutations interact with additional genetic and
479 microenvironmental factors to shape disease outcomes.

480 We identified recurrent CNAs frequently reported in the literature, including gains on 8q [10–
481 12,14,15,37,40], 17q [10,12,15,18–20,37,40], 20q [37,40], and losses on 11q [10–12,18,19,37]

482 and 16q [10,13,14,20,40]. These CNAs mirror those observed in invasive ductal carcinoma,
483 suggesting that pure DCIS already harbors genomic features characteristic of invasive
484 cancer—including alterations that define molecular subtypes (e.g., basal, luminal-A, or Her2-
485 enriched tumors; **see Supplementary Text**). Moreover, this observation is consistent with
486 earlier transcriptional and epigenetic studies that show subtype-specific alterations in gene
487 expression and DNA methylation are already evident in DCIS [41,42]. These findings suggest
488 that genetic and molecular aberrations defining subtypes likely emerge early - being at least
489 partially established at the DCIS stage - and support a model of subtype-specific progression
490 from DCIS to invasive breast cancer [41].

491 Importantly, our study also revealed novel associations between mutations in genes regulating
492 cytoskeletal integrity, cell membrane organization and function, cell-cell interactions, and cell
493 adhesion, and the risk of LR in pure DCIS. Across the cohort, five genes were significantly
494 associated with LR, underscoring their potential roles in tumor progression. These mutations
495 were not frequently co-occurring in the same patient, and involved genes maintaining cell
496 adhesion (*MYO7A*, *SH2B2*), organizing and supporting cellular organization, cytoskeleton and
497 membrane function (*PDZD8*, *MUCL3*, *MYO7A*). When the cohort was stratified by RT
498 administration, 27 unique genes - excluding *SH2B2* and *MUCL3* - were significantly associated
499 with increased LR risk in RT-treated patients. Many of these mutations were identified in genes
500 critical for maintaining cellular architecture, polarity, and adhesion, such as *KRT1*, *MUCL3*,
501 *TMEM132C*, *TTLL10*, *EML5*, *CEP128*, *CFAP47*, and *KIF24*, highlighting how disruptions in
502 structural integrity and epithelial cell organization may undermine the effectiveness of RT and
503 contribute to LR. In addition to structural changes, metabolic regulators like *MGAM2* and

504 *AADACL3* may influence the energetic and biosynthetic demands required for cell survival and
505 adaptation under genotoxic conditions. Alterations in *REV1*, a key player in DNA damage
506 tolerance, underscore the importance of effective DNA repair mechanisms following RT. These
507 findings expand upon prior studies linking disruptions in tissue structure and cell adhesion to
508 DCIS progression [43–46], while providing novel insights into their specific roles in recurrence
509 following RT in pure DCIS.

510 Notably, our analysis also revealed distinct biological programs underlying in-situ versus
511 invasive recurrence following RT. Specifically, while some mutations in genes involved in
512 structural maintenance and mechanosensory pathways were uniquely associated with in-situ
513 relapse, invasive progression was characterized by alterations in cell-cycle regulation, genome
514 stability, and inflammatory signaling. Because stratifying by both treatment and LR type
515 reduced our statistical power, these molecular distinctions will require confirmation in larger
516 cohorts.

517 Finally, we uniquely identified losses in cytobands harboring established tumor suppressor
518 genes—such as *CDKN2A* in 9p21 and *PRDM4* in 12q23–24—as well as regions containing
519 genes crucial for cell adhesion and migration, including a cluster of keratin genes in 12q13–14
520 and *DAL-1* in 18p11. In parallel, gains in regions like 11q, which includes oncogenic drivers
521 such as *CCND1* and *CTTN*, suggest that oncogene amplification may further contribute to
522 tumor aggressiveness and ultimately patient prognosis. Gains at 5p14, containing cadherin
523 genes (*CDH10*, *CDH9*, *CDH12*, *CDH18*), again emphasize the importance of maintaining
524 epithelial integrity to prevent LR in pure DCIS patients.

525 Future studies are warranted to investigate how these mutations interact with other molecular
526 pathways and microenvironmental factors to elucidate their contribution to DCIS prognosis
527 and the adverse effects of RT. Such research could help identify strategies to mitigate
528 recurrence risk following RT and improve treatment outcomes [47–49].

529 **Conclusions**

530 Our findings uncover the genomic landscape of pure DCIS and highlight key factors that
531 contribute to LR and mediate the adverse effects of RT. While *TP53* and *PIK3CA* mutations
532 play important roles in early tumorigenesis, they do not predict recurrence, emphasizing the
533 need for alternative biomarkers. We identify distinct mutational processes and genetic
534 alterations that disrupt cytoskeletal integrity, cell-cell interactions and cell adhesion, potentially
535 destabilizing the epithelial tissue environment and contributing to recurrence, particularly
536 following RT-induced stress. Importantly, our data also suggest that in-situ and invasive
537 recurrences after RT may follow divergent molecular trajectories—mutations affecting
538 structural maintenance and mechanosensory pathways were specifically linked to in-situ
539 relapse, whereas alterations in cell-cycle regulation, genome stability, and inflammatory
540 signaling characterized invasive progression._These insights provide a foundation for
541 understanding the genetic basis of DCIS progression and identifying potential molecular
542 drivers of treatment resistance. Future research will be essential to translate these insights
543 into clinical practice, guiding the development of more targeted therapeutic approaches to
544 improve outcomes for patients with DCIS.

545

546 **Methods**

547

548 ***The Ontario DCIS cohort***

549 The Ontario DCIS Cohort was established at the Sunnybrook Health Sciences Center (Toronto,
550 Canada) as a population-based sample of women diagnosed with pure DCIS defined as in situ
551 cancer without any invasive component between 1994 and 2003 [50–52]. All patients
552 underwent BCS, with a subset receiving subsequent RT as part of their standard-of-care.
553 Adjuvant endocrine therapy was administered to less than 15% of individuals, while none
554 received systemic chemotherapy or neoadjuvant endocrine therapy. The cohort features
555 comprehensive annotation of clinical annotation and expert pathology review. Previous studies
556 of this cohort have characterized outcomes based on clinical factors including age at
557 diagnosis, pathological features (tumor size and nuclear grade), and treatment modalities [53–
558 56].

559 Sample selection was prioritized to achieve balanced representation between RT-treated and
560 untreated patients, with equivalent distributions of individuals who did or did not develop an
561 invasive or in-situ ipsilateral recurrence within ten years. Tissue cores were obtained from
562 FFPE blocks, sampling DCIS tumors without microinvasion alongside adjacent normal and
563 stromal tissues. DNA and RNA were extracted using the Qiagen AllPrep FFPE DNA/RNA kit
564 (Qiagen). Samples yielding sufficient DNA quantities underwent library construction using the
565 Agilent SureSelect Human Exome library preparation kit and were sequenced on the
566 NovaSeq6000 platform (100bp paired-end, 100M reads/sample) at the Genome Quebec
567 Innovation Centre (Montreal, Canada). While high-quality sequencing data was obtained for

568 300 tumor tissues, downstream analyses focused on 147 samples with matched normal
569 profiles (144 normal tissue and 3 stroma non-epithelial samples).

570 Molecular subtypes were determined using RNA profiles available for a subset of patients (n =
571 122). Sequencing libraries were prepared using RNA Flex kit (Illumina). Raw reads were
572 processed using `TrimGalore` to remove adapters and low-quality reads [57], followed by
573 alignment to the human genome (Ensembl release v104) using `STAR` [58]. Reads counts were
574 mapped to genomic features using `Rsubread` R package [59] and gene counts were
575 normalized using the variance stabilizing transformation implemented in the `DESeq2` R
576 package [60]. PAM50 subtype classification was performed using the `genefu` R package [61]
577 with established centroids [62]. Specifically, normalized expression data of the 50 PAM50
578 genes were obtained using variance stabilizing transformation implemented in the `DESeq2` R
579 package [60] and compared to subtype-specific centroids using Pearson correlations. Each
580 sample was assigned to the molecular subtype with which it showed the highest correlation
581 coefficient.

582

583 ***Whole-exome sequencing data preprocessing***

584 Raw reads were processed using Trimmomatic [63] (version 0.39) to remove adaptor
585 sequences and low-quality bases. Reads were trimmed to retain high-quality sequences,
586 applying quality thresholds of 10 at read ends and 20 within a 4-base sliding window. The
587 remaining paired reads were processed following the GATK4 best practices. Briefly, reads
588 were aligned to the human reference genome (GRCh38, GATK resource bundle) using the

589 Burrows-Wheeler Aligner (BWA) [64]. Post-alignment procedures included sorting, annotating
590 reads with read groups, and marking duplicate reads with Picard. Base quality score
591 recalibration was conducted using GATK4 tools. A recalibration table was generated with the
592 `BaseRecalibrator` function using known variant sites (dbSNP138 for SNPs and Mills and
593 1000 Genomes for indels), and recalibration was applied with `ApplyBQSR` to adjust base
594 quality scores and correct for systematic technical errors. The process focused on SureSelect
595 Human Exons v7 regions with a 100 bp padding. Finally, properly paired reads were extracted,
596 excluding secondary alignments and low-quality reads, with the resulting files indexed using
597 `Samtools` [65]. After filtration, samples had a mean coverage depth of 152X (range: 52-308X,
598 standard deviation: 42X).

599

600 ***Single nucleotide variant & indel calling***

601 For variant calling, we used NeuSomatic [66], a deep learning approach that leverages both
602 tumor and matched normal sequence alignment information, alongside somatic mutation calls
603 from six different approaches: MuTect2, MuSE, VarDict, VarScan2, Strelka2, and
604 SomaticSniper [67–72]. This method was selected because of the low level of agreement
605 between callers in our data - the majority of mutations (88.7%) were identified by only one
606 caller (**Supplementary Fig. 4a**) - an observation consistent with previous studies [73–75]. The
607 number of mutations detected varied significantly across samples, with the minimum
608 identified in any single sample being 1,563 mutations, and the maximum reaching 283,247
609 mutations (**Supplementary Fig. 4a**).

610 Small-variant calling has become increasingly amenable to deep learning approaches, thanks
611 to the availability of extensive sequencing data and robust benchmarking datasets that cover
612 millions of variants across diverse genomic contexts [76]. A convolutional neural network
613 (CNN) architecture leverages information from sequencing reads and the reference genome in
614 the vicinity of each candidate variant to approximate complex, nonlinear functions, accurately
615 classifying loci as homozygous variant, heterozygous variant or homozygous reference (non-
616 variant) [66].

617 Best practices established using well-characterized somatic reference datasets from the
618 SEQC-II consortium demonstrated that NeuSomatic models achieve robust performance
619 across various sequencing technologies - including both fresh and FFPE DNA inputs - across a
620 range of tumor/normal purities and sequencing coverages, significantly outperforming
621 conventional approaches [73]. Accordingly, we used the ensemble extension of NeuSomatic.
622 This extension combines outputs from six individual variant callers by integrating features
623 from 93 channels and incorporates an additional 26 channels to capture alignment
624 information in a window of seven bases around the candidate mutation, resulting in a total of
625 119 input channels per candidate matrix. We used the recommended pre-trained model SEQC-
626 II (SEQC-WGS-Spike model), which was trained on 20 whole-genome sequencing replicate
627 pairs containing *in silico* somatic mutations with allele frequencies ranging from 1% to 100%
628 (with matched normals at both 95% and 100% purity, and sequencing coverage ranging from
629 ~40x to 220x) using five variant callers used: MuTect2, Strelka2, MuSE, SomaticSniper,
630 VarDict. After mutation calling, the recommended post-processing was applied to resolve long

631 INDEL sequences and the final NeuSomatic predictions were used for all downstream
632 analyses.

633 We obtained comprehensive genomic information for each variant using the Ensembl Variant
634 Effect Predictor (VEP) [77]. This included the effects on gene and protein function, such as
635 consequence types and amino acid changes, variant frequencies in different populations,
636 impact on regulatory regions, and potential associations with diseases and phenotypes.
637 Following annotation with the Ensembl VEP, we used the `vcf2maf` tool to transform VEP-
638 annotated VCF files into the Mutation Annotation Format (MAF). This conversion ensures each
639 variant is uniquely associated with a single gene transcript or isoform, despite the potential for
640 a variant to impact multiple isoforms. Particularly in cases where variants could be classified
641 under different effects — such as a `Missense_Mutation` near a `Splice_Site` — the MAF
642 format forces a singular designation for each variant by leveraging VEP's determinations for
643 canonical isoforms.

644 We excluded 100 genes commonly mutated in public exome datasets (FLAGS) due to their
645 lower likelihood of disease association [78]. This decision stems from their longer coding
646 regions, which inherently increase mutation probability, and the presence of paralogs that
647 might offset functional loss these mutations could cause [78,79].

648 High-confidence variants (identified with a probability score of 0.7 or higher) consistently
649 showed higher allele frequencies compared to those categorized as low-quality (with scores
650 between 0.4 and 0.7) and rejected variants (with scores below 0.4) (**Supplementary Fig. 4B**).
651 To reduce potential false positives, we selected high-confidence variants with allele frequency
652 above 0.1. To further confirm that variants detected were not technical artifacts, we assessed

653 the relationship between tumor mutational burden and sequencing depth (**Supplementary Fig.**
654 **5**). The lack of correlation supports robust variant detection independent of coverage.

655 Mutation patterns and frequencies were visualized using the `oncoplot` function from the
656 `maftools` R package [80], which displays mutation types and frequencies across samples.

657

658 ***Mutational signatures***

659 We performed mutational signature analysis using the COSMIC database of single-base
660 substitution (SBS) signatures [24]. First, the trinucleotide context of each single nucleotide
661 variant was characterized. We then used the `fit_to_signature` function in the
662 `MutationalPatterns` R package (version 3.19) [81] to find the linear combination of
663 COSMIC mutation signatures that most closely reconstructs the mutation spectra for each
664 sample by solving the nonnegative least-squares constraints problem. We used strict refit
665 where the signature with the lowest contribution is removed; refitting is repeated until the
666 cosine similarity between the original and reconstructed profile becomes more than
667 max_delta= 0.004). We selected signatures that contributed to the mutation spectrum of at
668 least 10 samples and plotted their relative contributions using the `pheatmap` R package [82].

669 ***Driver genes***

670 To identify driver genes (i.e., genes under positive selection in cancer), we used the `dNdScv`
671 analysis method [26]. This approach is based on the evaluation of the ratio between
672 synonymous (silent) mutations and non-synonymous (missense) mutations in genes. A higher
673 ratio of non-synonymous to synonymous mutations in a gene indicates positive selection for

674 mutations that may confer a growth advantage to cancer cells, suggesting the potential role of
675 the gene as a driver in tumorigenesis.

676 The `dNdScv` method estimates the background mutation rate of each gene by combining local
677 information (synonymous mutations within the gene) with global information (variation of
678 mutation rates across genes). This approach controls for the sequence composition of genes
679 and accounts for mutational signatures, providing a more accurate estimation of the expected
680 neutral mutation rate. In particular, the `dNdScv` R package [83] implementation uses
681 trinucleotide context-dependent substitution matrices to mitigate common mutation biases
682 that can affect dN/dS calculations [26].

683 To visualize and analyze the distribution and nature of mutations in driver genes, we used
684 lollipop plots generated from the maftools R package [80] and annotation tracks from
685 cBioPortal [84]. These plots provide a representation of mutation types and their locations
686 along the protein sequence as well as annotations including likely mutation hotspots as
687 identified by Memorial Sloan Kettering Cancer Hotspots and 3D Hotspots databases [85], and
688 annotation records of therapeutic indication from OncoKB [86].

689

690 ***Copy Number Alterations (CNAs)***

691 To investigate copy number alterations, we applied the Allele-Specific Copy Number Analysis
692 of Tumours (ASCAT v3) on our tumor normal pairs estimating tumor purity, ploidy, and allele-
693 specific copy number [87]. The `runAscat` function was executed with default settings
694 optimized for high-throughput exome sequencing data, with the gamma parameter set to 1.

695 After examination of ASCAT sunrise plots, we identified a subset of samples (n=35 samples),
696 for which the initial estimates of tumor purity and ploidy did not align with the regions of
697 highest confidence on the sunrise plots. The `runAscat` was re-run for these samples by
698 manually assigning the aberrant cell fraction (tumor purity) and tumor ploidy parameters
699 corresponding to the regions of highest probability as depicted on the sunrise plots. Overall,
700 eight samples were excluded from further analysis due to poor goodness of fit leaving 139
701 samples with CNA profiles for downstream analyses. Absolute number gains and losses
702 shared across samples were visualized across whole chromosomal regions using `aCNViewer`
703 [88] (window size of 1 Mbp). The estimated ploidy for each sample is subtracted from the copy
704 number values of each segment which means that a copy number of 0 is copy number change.
705 These adjusted windows at base resolution are then plotted into a stacked histogram
706 representing genome-wide absolute copy number and copy neutral variations over all samples
707 in a group.

708 We applied a re-segmentation approach to adjust for amplitude divergence due to technical
709 variability implemented in `CNApp` [89] using the default settings (minimum segment length =
710 100 Kbp, minimum amplitude deviation from segment to zero = 0.16, maximum distance
711 between segments=1 Mb, maximum amplitude deviation between segments = 0.16, and
712 maximum BAF deviation between segments = 0.1). Re-segmented data were then used to
713 calculate the broad, focal and global CNA scores. We then transformed re-segmented data
714 into genomic regions profiles (chromosome arms, cytobands and sub-cytobands) using both
715 focal and broad segments. Length-relative means are computed for each window by
716 considering amplitude values from those segments included in each specific window. Default

717 cutoffs for low-level copy number gains and losses (i.e., |0.2|) were used to infer CNA
718 frequencies.
719
720 ***Survival analyses***
721 We evaluated the association between gene mutations or copy number aberrations in
722 cytobands and 10-year local recurrence-free survival using Firth's penalized likelihood Cox
723 regression which accounts for small sample sizes and rare events. This analysis was
724 conducted using the `coxphf` R package [90]. Aberrations were included in the analysis only if
725 detected in at least five lesions (2,030 genes; 303 cytobands). This analysis was performed
726 across the entire patient cohort. To further investigate mutated genes associated with
727 response to RT, a stratified analysis was conducted based on treatment groups.
728 To control for multiple testing in the identification of mutated genes associated with LR, we
729 employed permutation-based testing (1,000 permutations) to establish empirical significance
730 thresholds. We required both a traditional p-value threshold ($p < 0.01$ for full-cohort analyses; p
731 < 0.05 for RT-stratified analyses) and a permuted p-value < 0.05.
732 Kaplan-Meier survival curves were used to visualize the results, illustrating event-free survival
733 probabilities over time for patients stratified by mutational status in the specified genes or
734 gene sets.
735 GO enrichment analyses were performed using the clusterProfiler R package [91]. Entrez gene
736 identifiers were mapped to GO terms using the `org.Hs.eg.db` annotation database [92]. All
737 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 [93]. The enrichment map was visualized using the `emapplot` function [94] which displays the significantly enriched terms with at least 2 genes.

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.

Declarations

Ethics approval and consent to participate

Research ethics board approval was obtained from Sunnybrook Health Sciences Centre, Toronto, Ontario (#2738). This study was facilitated through ICES which is named as a

758 prescribed entity in section 45 of PHIPA (Regulation 329/04, section 18) which allows access
759 and utilization of administrative data for research purposes with a waived requirement for
760 consent.

761 **Consent for publication**

762 Not Applicable

763 **Availability of data and materials**

764 The dataset analyzed in this study is not publicly available due the personal, sensitive, and
765 inherently identifying nature of raw genomic data. Access to raw data and patient metadata
766 may be provided but is controlled and requires institutional material data transfer agreements
767 (contact person: eileen.rakovitch@sunnybrook.ca).

768 The scripts used to reproduce the analyses performed in this study are available in the
769 Dumeaux Lab GitHub Repository https://github.com/dumeaux-lab/dcis-dna_paper.

770 **Competing Interests**

771 The authors declare no competing interests.

772 **Funding**

773 This work was supported by Canadian Institute for Health Research (CIHR) project grant
774 #391682. We thank the Canadian Foundation for Innovation J. Evans Leaders Fund grant
775 #43481 for the support of the computing infrastructure used throughout our analyses.

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

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

Data Acknowledgement

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES, the MOH or MLTC is intended or should be inferred. As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario.

Authors' Contributions

Conceptualization, S.N.M, L.P., E.R., M.T.H and V.D.; Methodology, N.R, E.S, V.D.; Investigation, N.R., E.S., M.Z., V.D.; Software N.R., E.J.M., V.D.; Writing – Original Draft, N.R. and V.D.; Writing – Review & Editing, E.S., M.Z., E.J.M., E.H., S.N.M., S.T., L.P., M.T.H. and E.R. ; Funding Acquisition, E.R., M.T.H, E.H., S.N.M, S.T., L.P. and V.D.; Resources, E.R., S.N.M, S.T., L.P. ; Supervision, V.D.

References

1. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nat Rev Dis Primer. 2019;5:1–31.

799 2. Sørum R, Hofvind S, Skaane P, Haldorsen T. Trends in incidence of ductal carcinoma in situ:
800 The effect of a population-based screening programme. The Breast. 2010;19:499–505.

801 3. Gangnon RE, Sprague BL, Stout NK, Alagoz O, Weedon-Fekjær H, Holford TR, et al. The
802 contribution of mammography screening to breast cancer incidence trends in the United
803 States: an updated age-period-cohort model. Cancer Epidemiol Biomark Prev Publ Am Assoc
804 Cancer Res Cosponsored Am Soc Prev Oncol. 2015;24:905–12.

805 4. Force UPST, Nicholson WK, Silverstein M, Wong JB, Barry MJ, Chelmow D, et al. Screening
806 for Breast Cancer: US Preventive Services Task Force Recommendation Statement. JAMA.
807 2024;331:1918–30.

808 5. Tomlinson-Hansen S, Khan M, Cassaro S. Breast Ductal Carcinoma in Situ. StatPearls
809 [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Nov 16]. Available
810 from: http://www.ncbi.nlm.nih.gov/books/NBK567766/

811 6. Zhang Y, Weinberg RA. Epithelial-to-mesenchymal transition in cancer: complexity and
812 opportunities. Front Med. 2018;12:361–73.

813 7. Maxwell AJ, Hilton B, Clements K, Dodwell D, Dulson-Cox J, Kearins O, et al. Unresected
814 screen-detected ductal carcinoma in situ: Outcomes of 311 women in the Forget-Me-Not 2
815 study. Breast Edinb Scotl. 2022;61:145–55.

816 8. Rakovitch E, Gray R, Baehner FL, Sutradhar R, Crager M, Gu S, et al. Refined estimates of
817 local recurrence risks by DCIS score adjusting for clinicopathological features: a combined
818 analysis of ECOG-ACRIN E5194 and Ontario DCIS cohort studies. Breast Cancer Res Treat.
819 2018;169:359–69.

820 9. Andor N, Maley CC, Ji HP. Genomic instability in cancer: Teetering on the limit of tolerance.
821 Cancer Res. 2017;77:2179–85.

822 10. Trinh A, Gil Del Alcazar CR, Shukla SA, Chin K, Chang YH, Thibault G, et al. Genomic
823 Alterations during the In Situ to Invasive Ductal Breast Carcinoma Transition Shaped by the
824 Immune System. Mol Cancer Res. 2021;19:623–35.

825 11. Lin C-Y, Vennam S, Purington N, Lin E, Varma S, Han S, et al. Genomic landscape of ductal
826 carcinoma in situ and association with progression. Breast Cancer Res Treat. 2019;178:307–
827 16.

828 12. Hernandez L, Wilkerson PM, Lambros MB, Campion-Flora A, Rodrigues DN, Gauthier A, et
829 al. Genomic and mutational profiling of ductal carcinomas in situ and matched adjacent
830 invasive breast cancers reveals intra-tumour genetic heterogeneity and clonal selection. J
831 Pathol. 2012;227:42–52.

832 13. Nagasawa S, Kuze Y, Maeda I, Kojima Y, Motoyoshi A, Onishi T, et al. Genomic profiling
833 reveals heterogeneous populations of ductal carcinoma in situ of the breast. Commun Biol.
834 2021;4:1–13.

835 14. Nachmanson D, Officer A, Mori H, Gordon J, Evans MF, Steward J, et al. The breast pre-
836 cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in
837 situ. Npj Breast Cancer. 2022;8:1–13.

838 15. Abba MC, Gong T, Lu Y, Lee J, Zhong Y, Lacunza E, et al. A Molecular Portrait of High-Grade
839 Ductal Carcinoma *In Situ*. Cancer Res. 2015;75:3980–90.

840 16. Kaplan HG, Dowdell AK, Berry AB, Shimol RB, Robinson FL, Carney CA, et al. Multi-omic
841 profiling of simultaneous ductal carcinoma in situ and invasive breast cancer. Breast Cancer
842 Res Treat. 2024;205:451–64.

843 17. Kader T, Hill P, Zethoven M, Goode DL, Elder K, Thio N, et al. Atypical ductal hyperplasia is a
844 multipotent precursor of breast carcinoma. J Pathol. 2019;path.5262.

845 18. Kim SY, Jung S-H, Kim MS, Baek I-P, Lee SH, Kim T-M, et al. Genomic differences between
846 pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast
847 cancer. Oncotarget. 2015;6:7597–607.

848 19. Pareja F, Brown DN, Lee JY, Da Cruz Paula A, Selenica P, Bi R, et al. Whole-Exome
849 Sequencing Analysis of the Progression from Non–Low-Grade Ductal Carcinoma *In Situ* to
850 Invasive Ductal Carcinoma. Clin Cancer Res. 2020;1078-0432.CCR-19–2563.

851 20. Strand SH, Rivero-Gutiérrez B, Houlahan KE, Seoane JA, King LM, Risom T, et al. Molecular
852 classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis
853 of TBCRC 038 and RAHBT cohorts. Cancer Cell [Internet]. 2022 [cited 2022 Dec 7]; Available
854 from: https://www.sciencedirect.com/science/article/pii/S1535610822005128

855 21. Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. Methylation-mediated
856 deamination of 5-methylcytosine appears to give rise to mutations causing human inherited
857 disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum Genomics. 2010;4:406–
858 10.

859 22. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like
860 mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.

861 23. Steiert TA, Parra G, Gut M, Arnold N, Trotta J-R, Tonda R, et al. A critical spotlight on the
862 paradigms of FFPE-DNA sequencing. Nucleic Acids Res. 2023;51:7143–62.

863 24. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al.
864 Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

865 25. Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. The mutational signatures
866 of formalin fixation on the human genome [Internet]. bioRxiv; 2021 [cited 2024 Dec 6]. p.
867 2021.03.11.434918. Available from:
868 https://www.biorxiv.org/content/10.1101/2021.03.11.434918v1

869 26. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Loo PV, et al. Universal

870 Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171:1029-1041.e21.

871 27. Yuan T, Cantley L. PI3K pathway alterations in cancer: variations on a theme. Oncogene.
872 2008;27:5497–510.

873 28. Bader AG, Kang S, Zhao L, Vogt PK. Oncogenic PI3K deregulates transcription and
874 translation. Nat Rev Cancer. 2005;5:921–9.

875 29. Yang XH, Huang S. PFM1 (PRDM4), a new member of the PR-domain family, maps to a
876 tumor suppressor locus on human chromosome 12q23-q24.1. Genomics. 1999;61:319–25.

877 30. Arnold SA, Brekken RA. SPARC: a matricellular regulator of tumorigenesis. J Cell Commun
878 Signal. 2009;3:255–73.

879 31. Vecchione A, Fassan M, Anesti V, Morrione A, Goldoni S, Baldassarre G, et al. MITOSTATIN,
880 a putative tumor suppressor on chromosome 12q24.1, is downregulated in human bladder and
881 breast cancer. Oncogene. 2009;28:257–69.

882 32. Heller G, Geradts J, Ziegler B, Newsham I, Filipits M, Markis-Ritzinger E-M, et al.
883 Downregulation of TSLC1 and DAL-1 expression occurs frequently in breast cancer. Breast
884 Cancer Res Treat. 2007;103:283–91.

885 33. Kittiniyom K, Gorse KM, Dalbegue F, Lichy JH, Taubenberger JK, Newsham IF. Allelic loss
886 on chromosome band 18p11.3 occurs early and reveals heterogeneity in breast cancer
887 progression. Breast Cancer Res. 2001;3:192–8.

888 34. Schuuring E, Verhoeven E, Mooi WJ, Michalides RJ. Identification and cloning of two
889 overexpressed genes, U21B31/PRAD1 and EMS1, within the amplified chromosome 11q13
890 region in human carcinomas. Oncogene. 1992;7:355–61.

891 35. Ormandy CJ, Musgrove EA, Hui R, Daly RJ, Sutherland RL. Cyclin D1, EMS1 and 11q13
892 Amplification in Breast Cancer. Breast Cancer Res Treat. 2003;78:323–35.

893 36. Houle AA, Gibling H, Lamaze FC, Edgington HA, Soave D, Fave M-J, et al. Aberrant PRDM9
894 expression impacts the pan-cancer genomic landscape. Genome Res. 2018;28:1611–20.

895 37. Kader T, Zethoven M, Mahale S, Saunders H, Tjoeka L, Lehmann R, et al. Predictive
896 biomarkers of breast ductal carcinoma in situ may underestimate the risk of recurrence due to
897 de novo ipsilateral breast carcinoma development [Internet]. bioRxiv; 2024 [cited 2025 Jan 3].
898 p. 2024.05.19.594731. Available from:
899 https://www.biorxiv.org/content/10.1101/2024.05.19.594731v1

900 38. Ang DC, Warrick AL, Shilling A, Beadling C, Corless CL, Troxell ML. Frequent
901 phosphatidylinositol-3-kinase mutations in proliferative breast lesions. Mod Pathol.
902 2014;27:740–50.

903 39. Silwal-Pandit L, Langerød A, Børresen-Dale A-L. TP53 Mutations in Breast and Ovarian

904 Cancer. Cold Spring Harb Perspect Med. 2017;7:a026252.

905 40. Pang J-MB, Savas P, Fellowes AP, Mir Arnau G, Kader T, Vedururu R, et al. Breast ductal
906 carcinoma in situ carry mutational driver events representative of invasive breast cancer. Mod
907 Pathol. 2017;30:952–63.

908 41. Lesurf R, Aure MR, Mørk HH, Vitelli V, Lundgren S, Børresen-Dale A-L, et al. Molecular
909 Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast
910 Cancer. Cell Rep. 2016;16:1166–79.

911 42. Bergholtz H, Lien TG, Swanson DM, Frigessi A, Daidone MG, Tost J, et al. Contrasting DCIS
912 and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions. NPJ Breast
913 Cancer. 2020;6:26.

914 43. Emery LA, Tripathi A, King C, Kavanah M, Mendez J, Stone MD, et al. Early Dysregulation of
915 Cell Adhesion and Extracellular Matrix Pathways in Breast Cancer Progression. Am J Pathol.
916 2009;175:1292–302.

917 44. Glukhova MA, Streuli CH. How integrins control breast biology. Curr Opin Cell Biol.
918 2013;25:633–41.

919 45. Maxwell CA, Benítez J, Gómez-Baldó L, Osorio A, Bonifaci N, Fernández-Ramires R, et al.
920 Interplay between BRCA1 and RHAMM regulates epithelial apicobasal polarization and may
921 influence risk of breast cancer. PLoS Biol. 2011;9:e1001199.

922 46. Qin X, Strand SH, Lee MR, Saraswathibhatla A, van IJzendoorn DGP, Zhu C, et al. Single Cell
923 Expression Analysis of Ductal Carcinoma in Situ Identifies Complex Genotypic-Phenotypic
924 Relationships Altering Epithelial Composition. Cancer Res [Internet]. 2025 [cited 2025 Apr 21];
925 Available from: https://doi.org/10.1158/0008-5472.CAN-24-3023

926 47. Deng Y, Chen G, Xiao J, Deng H. Role and potential therapeutic strategies of matrix
927 mechanics for optimizing tumor radiotherapy. Mechanobiol Med. 2024;2:100037.

928 48. Prakash J, Shaked Y. The Interplay between Extracellular Matrix Remodeling and Cancer
929 Therapeutics. Cancer Discov. 2024;14:1375–88.

930 49. Frascogna C, Mottareale R, La Verde G, Arrichiello C, Muto P, Netti PA, et al. Role of the
931 mechanical microenvironment on CD-44 expression of breast adenocarcinoma in response to
932 radiotherapy. Sci Rep. 2024;14:391.

933 50. Rakovitch E, Nofech-Mozes S, Narod SA, Hanna W, Thiruchelvam D, Saskin R, et al. Can we
934 select individuals with low risk ductal carcinoma in situ (DCIS)? A population-based outcomes
935 analysis. Breast Cancer Res Treat. 2013;138:581–90.

936 51. Rakovitch E, Nofech-Mozes S, Hanna W, Baehner FL, Saskin R, Butler SM, et al. A
937 population-based validation study of the DCIS Score predicting recurrence risk in individuals
938 treated by breast-conserving surgery alone. Breast Cancer Res Treat. 2015;152:389–98.

939 52. Rakovitch E, Nofech-Mozes S, Hanna W, Sutradhar R, Baehner FL, Miller DP, et al. Multigene
940 Expression Assay and Benefit of Radiotherapy After Breast Conservation in Ductal Carcinoma
941 in Situ. J Natl Cancer Inst. 2017;109:djw256.

942 53. Paszat L, Sutradhar R, Zhou L, Nofech-Mozes S, Rakovitch E. Including the Ductal
943 Carcinoma-In-Situ (DCIS) Score in the Development of a Multivariable Prediction Model for
944 Recurrence After Excision of DCIS. Clin Breast Cancer. 2019;19:35–46.

945 54. Klein J, Kong I, Paszat L, Nofech-Mozes S, Hanna W, Thiruchelvam D, et al. Close or
946 positive resection margins are not associated with an increased risk of chest wall recurrence
947 in women with DCIS treated by mastectomy: a population-based analysis. SpringerPlus.
948 2015;4:335.

949 55. Lalani N, Paszat L, Sutradhar R, Thiruchelvam D, Nofech-Mozes S, Hanna W, et al. Long-
950 term outcomes of hypofractionation versus conventional radiation therapy after breast-
951 conserving surgery for ductal carcinoma in situ of the breast. Int J Radiat Oncol Biol Phys.
952 2014;90:1017–24.

953 56. Rakovitch E, Parpia S, Koch A, Grimard L, Soliman H, Stevens C, et al. DUCHESS: an
954 evaluation of the ductal carcinoma in situ score for decisions on radiotherapy in patients with
955 low/intermediate-risk DCIS. Breast Cancer Res Treat. 2021;188:133–9.

956 57. Krueger F. FelixKrueger/TrimGalore [Internet]. 2025 [cited 2025 Jun 2]. Available from:
957 https://github.com/FelixKrueger/TrimGalore

958 58. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
959 universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29:15–21.

960 59. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for
961 alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019;47:e47.

962 60. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol.
963 2010;11:R106.

964 61. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an
965 R/Bioconductor package for computation of gene expression-based signatures in breast
966 cancer. Bioinforma Oxf Engl. 2016;32:1097–9.

967 62. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk
968 Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009;27:1160–7.

969 63. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
970 data. Bioinforma Oxf Engl. 2014;30:2114–20.

971 64. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
972 [Internet]. arXiv.org. 2013 [cited 2024 Apr 8]. Available from:
973 https://arxiv.org/abs/1303.3997v2

974 65. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
975 SAMtools and BCFtools. GigaScience. 2021;10:giab008.

976 66. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional
977 neural networks for accurate somatic mutation detection. Nat Commun. 2019;10:1041.

978 67. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs
979 and Indels with Mutect2 [Internet]. 2019 [cited 2024 Apr 8]. Available from:
980 http://biorxiv.org/lookup/doi/10.1101/861054

981 68. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor
982 heterogeneity using a sample-specific error model improves sensitivity and specificity in
983 mutation calling from sequencing data. Genome Biol. 2016;17:178.

984 69. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a
985 novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic
986 Acids Res. 2016;44:e108.

987 70. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic
988 mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res.
989 2012;22:568–76.

990 71. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and
991 accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4.

992 72. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper:
993 identification of somatic point mutations in whole genome sequencing data. Bioinformatics.
994 2012;28:311.

995 73. Sahraeian SME, Fang LT, Karagiannis K, Moos M, Smith S, Santana-Quintero L, et al.
996 Achieving robust somatic mutation detection with deep learning models derived from
997 reference data sets of a cancer sample. Genome Biol. 2022;23:12.

998 74. Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus AV. Systematic benchmark of
999 state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding
1000 sequence variant discovery. BMC Genomics. 2022;23:155.

1001 75. de Schaetzen van Brienen L, Larmuseau M, Van der Eecken K, De Ryck F, Robbe P, Schuh A,
1002 et al. Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples.
1003 BMC Med Genomics. 2020;13:94.

1004 76. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, et al. Variant calling and
1005 benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023;24:464–
1006 83.

1007 77. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant
1008 Effect Predictor. Genome Biol. 2016;17:122.

1009 78. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW. FLAGS,
1010 frequently mutated genes in public exomes. BMC Med Genomics. 2014;7:64.

1011 79. Alfieri F, Caravagna G, Schaefer MH. Cancer genomes tolerate deleterious coding
1012 mutations through somatic copy number amplifications of wild-type regions. Nat Commun.
1013 2023;14:3594.

1014 80. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and
1015 comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.

1016 81. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-
1017 wide analysis of mutational processes. Genome Med. 2018;10:33.

1018 82. Raivo Kolde. pheatmap: Pretty Heatmaps [Internet]. 2010 [cited 2025 May 20]. p. 1.0.12.
1019 Available from: https://CRAN.R-project.org/package=pheatmap

1020 83. Martincorena I. im3sanger/dndscv [Internet]. 2025 [cited 2025 May 20]. Available from:
1021 https://github.com/im3sanger/dndscv

1022 84. Mutation data transcript annotation [Internet]. [cited 2025 May 20]. Available from:
1023 https://docs.cbioportal.org/mutation-data-transcript-annotation/

1024 85. Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue MTA, Jonsson P, et al.
1025 Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discov. 2018;8:174–83.

1026 86. Suehnholz SP, Nissan MH, Zhang H, Kundra R, Nandakumar S, Lu C, et al. Quantifying the
1027 Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discov.
1028 2024;14:49–65.

1029 87. Lesluyes T, Tarabichi M, Haase K, Demeulemeester J, Loo P. Robust and platform-
1030 independent CNA calling with ASCAT v3. 2023.

1031 88. Renault V, Tost J, Pichon F, Wang-Renault S-F, Letouzé E, Imbeaud S, et al. aCNViewer:
1032 Comprehensive genome-wide visualization of absolute copy number and copy neutral
1033 variations. PloS One. 2017;12:e0189334.

1034 89. Franch-Expósito S, Bassaganyas L, Vila-Casadesús M, Hernández-Illán E, Esteban-Fabró R,
1035 Díaz-Gay M, et al. CNApp, a tool for the quantification of copy number alterations and
1036 integrative analysis revealing clinical implications. Robles-Espinoza CD, Cheah KSE, editors.
1037 eLife. 2020;9:e50267.

1038 90. Heinze G, Ploner M, Jiricka L, Steiner G. coxphf: Cox Regression with Firth's Penalized
1039 Likelihood [Internet]. 2023 [cited 2025 May 20]. Available from: https://cran.r-
1040 project.org/web/packages/coxphf/index.html

1041 91. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological
1042 Themes Among Gene Clusters. OMICS J Integr Biol. 2012;16:284–7.

92. org.Hs.eg.db [Internet]. Bioconductor. [cited 2025 May 20]. Available from: http://bioconductor.org/packages/org.Hs.eg.db/

93. pairwise_termsim: pairwise_termsim in enrichplot: Visualization of Functional Enrichment Result [Internet]. [cited 2025 May 20]. Available from: https://rdrr.io/bioc/enrichplot/man/pairwise_termsim.html

94. emapplot function - RDocumentation [Internet]. [cited 2025 May 20]. Available from: https://www.rdocumentation.org/packages/enrichplot/versions/1.13.1.994/topics/emapplot