博士論文

Spatial transcriptome analysis reveals heterogeneity of cells in cancer

(がんの空間トランスクリプトーム解析が明らかにするがん組織の多様性**)**

久世　裕太

Table of Contents

1 Introduction

Ductal carcinoma *in situ* (DCIS) is a non-invasive breast cancer that is considered a precursor of invasive ductal carcinoma (IDC) (1). This is because of the following assumption about breast cancer progression: first, ductal hyperproliferation causes carcinoma *in situ*, which then becomes invasive and finally metastatic (1) (Fig. 1). Previously, DCIS specimens were considered rare; however, with the development of diagnostic technology in recent years using mammography, the number of cases has also increased to approximately 20% among all breast cancers (2-4).

If all DCIS were true precursors of IDC, the diagnostic rate of IDC should decrease with the increase in the diagnostic rate of DCIS, but IDC rates have not decreased (2). Additionally, previous studies showed that when patients with DCIS were followed up without treatment, the rate of progression to IDC was 10%–50% (5-11), suggesting that approximately half of patients with DCIS may not require treatment. Thus, DCIS may be a mixture of high-risk DCIS, a true precursor of IDC requiring resection, and true low-risk DCIS, which does not require treatment. Therefore, there is concern that the current treatment may be excessive.

The clinical classification of DCIS involves the following three groups: 1) High-grade DCIS with a high risk of relapse as IDC with standard treatment alone (high-risk DCIS). 2) Intermediate-grade DCIS with a high risk of relapse requiring standard treatment. 3) low-grade DCIS with a low risk of relapse, which may not relapse without treatment (low-risk DCIS). Several clinical trials are being undertaken globally to establish the criteria for diagnosing low-risk DCIS (12-14). However, this classification remains unsatisfactory because the current criteria use only clinicopathological factors such as nuclear grade, tumor size, presence of comedo necrosis, age, hormone receptor, and human epidermal growth factor receptor 2 (HER2) expression status (15-19). Furthermore, a global standard is absent (20), and some of these clinicopathological evaluations differ among pathologists. Therefore,

besides pathological evaluations, molecular evaluations are also required.

Although previous studies on the DCIS genomic landscape have suggested that several genes (*PIK3CA*, *TP53*, and *GATA3*) were frequently mutated in DCIS (20, 21), it is unclear how these contribute to the progression of IDC.

In this study, I analyzed 21 whole-exome sequencing and 72 deep target sequencing datasets to identify genomic risk factors during DCIS progression. Next, I analyzed spatial transcriptome data to reveal the molecular effect of the identified risk factor and the heterogeneity of DCIS.

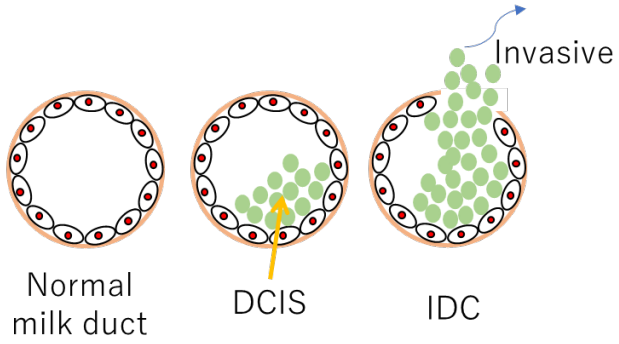


Fig. 1 Schematic representation of the progression from DCIS to IDC.

Ductal carcinoma in situ (DCIS) first occurs in normal breast ducts, which then becomes invasive and finally metastatic. : IDC= invasive ductal carcinoma

2 Contributions

This research was conducted in collaboration with Dr. Satoi Nagasawa. I performed the informatics analysis part. Specimens and clinical information were provided by St. Marianna Medical University through Dr. Satoi Nagasawa. Experiments and sample preparation were performed by Dr. Nagasawa and the technical staff. Pathology and clinical interpretation were performed by Dr. Satoi Nagasawa.

3 Material and Method

3.1 Ethics approval

This study was approved by the Clinical Ethics Committee of St. Marianna University (approval number: 2297-i103) and the University of Tokyo (approval number: 22-234). The analyses were performed as per ethical regulations.

3.2 Exome sequencing sample preparations

First, I used formalin-fixed, paraffin-embedded (FFPE) tissue samples, which were prepared from 21 patients with DCIS and 4 patients with IDC (relapse). The tumor area was separated from the corresponding normal tissue and extracted from the surrounding stroma with minimal contamination via microdissection. DNA was extracted from microdissected tissues using QIAamp DNA Mini and Micro kits (Qiagen, Hilden, Germany). Next, the DNA libraries were prepared using the SureSelect XT HS Kit (Agilent Technologies, Santa Clara, CA) and SureSelect Human All Exon v7 (Agilent Technologies, Santa Clara, CA) following the manufacturer's instructions. The libraries were sequenced using the Hiseq3000 platform (Illumina, San Diego, CA) with 100-bp paired-end reads.

3.3 Target sequencing sample preparations

After obtaining FFPE tissue samples from 72 patients with DCIS, tumor areas were extracted from the normal tissue via microdissection. Owing to limited specimens, only tumor samples were used for analysis. DNA was extracted from microdissected tissues using the QIAamp DNA Mini and Micro kits (Qiagen, Hilden, Germany). Next, deep target-sequencing libraries were prepared using the SureSelect XT HS Kit (Agilent Technologies, Santa Clara, CA) and SureSelect Custom DNA Target Enrichment Probes (Agilent Technologies, Santa Clara, CA),

designed to capture 180 genes associated with breast cancer, following the manufacturer's instructions. The libraries were sequenced using the Hiseq3000 platform (Illumina, San Diego, CA) with 100-bp paired-end reads.

3.4 Exome and Target sequencing data analysis

The sequencing data were aligned to the human genome (hg19) using Burrows–Wheeler Aligner Mem(22) (ver. 0.7.17). Duplications were removed using Picard tools (ver. 2.18.25) (23). Somatic mutations in tumor-normal pair mode or tumor-only mode were detected using GATK Mutect2 (24) (ver. 4.0.12-0) and subsequently annotated using ANNOVAR (25). Filter-passed mutations were extracted from the Mutect2 results, except for those registered in dbSNP, EXAC, ESP, and 1000 Genome (excluding COSMIC); filtered out at the normal tag > 0, tumor variant tag < 2 for exome-data and variant allele frequency (VAF) < 3%, depth < 10, tlod score < 10 for target sequence data using a custom Perl script; and manually checked by Integrative Genomics Viewer (IGV) (26).

3.5 Spatial transcriptome sample preparations

After surgical removal, each sample was immediately embedded at an optimal cutting temperature (OCT) compound (TissueTek Sakura) using a −80°C cryomold. Samples embedded in the OCT compound were thinly sliced at 10-um thickness (Leica CM3050 S) to obtain the sections. These sections were used for library preparation according to the Visium Spatial Gene Expression User Guide (10x genomic, Pleasanton, CA). Based on the results of the tissue optimization time course experiment, tissue permeabilization time was set at 6 minutes. The libraries were sequenced via the Novaseq6000 platform (Illumina, San Diego, CA) with 28/98-bp paired-end reads.

3.6 Spatial transcriptome analysis

The sequencing data were analyzed using Space Ranger (ver. 1.0.0), which performed human genome (hg38) mapping, gene expression quantification, and histological imaging data alignment.

3.7 Single-cell DNA-sequencing sample preparations

Tissue biopsies from patients with DCIS were obtained during surgical procedures. These tissue samples were used for library preparation using the Chromium Single-Cell DNA Reagent Kit (10x genomic, Pleasanton, CA). The libraries were sequenced via the Novaseq6000 platform (Illumina, San Diego, CA) with 100-bp paired-end reads.

3.8 Single-cell DNA-sequencing analysis

The sequencing data were analyzed using Cell Ranger DNA (ver. 1.1.0) with a default parameter, which performed human genome (hg38) mapping, correcting GC bias by fitting to the quadratic function that minimizes the entropy of the read count histogram per 20 kbp, CNV calling with a 20-kb bin size, and report generation. Noisy cells and ploidy < 1.9 or 2.0 < ploidy (Her2 > 8 are an exception) cells were excluded from the results of Cell Ranger DNA. CNV heatmap was plotted using the R pheatmap package and the Manhattan distance. The hierarchical clustering was computed using the Ward's minimum variance method. Properly clustered numbers were estimated by an elbow plot of the sum of squared errors of prediction.

3.9 Case A analysis

Spots with mutations were identified using a custom Perl script. Differential expression genes (DEGs) were identified using R TCC (27) packages and filtered out at FDR > 0.01. Pathway enrichment analysis was performed by Metascape (28) using the top 100 genes in the FDR

value , and target pathways were selected to the MsigDB Hallmark (HALLMARK) (12) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets.

3.10 Case B analysis

One hundred and eighty-eight spots, morphologically determined as DCIS, were manually selected. Hierarchical clustering was performed based on Spearman's distance and Ward's method. The detection of DEG and enrichment analysis involved the same methods as in Case A. DEGs were identified by TCC packages, while Metascape (HALLMARK and KEGG) was used for pathway enrichment analysis using the top 100 genes in the FDR values.

3.11 Case C analysis

The spots were clustered using k-means clustering (k = 2) by Space Ranger. The detected clusters overlapped with morphological cancer and stromal cells. TCC packages were used to identify 2354 DEGs between the two clusters. WGCNA packages were used to group DEGs upregulated in cancer using a co-expression pattern (29). The power (β) parameter was estimated to be 3 by the pickSoftThreshold function, with a fit value exceeding RsquaredCut 0.8. The two identified gene modules were used for pathway enrichment analysis by Metascape (HALLMARK and KEGG). The single-cell DNA analysis method has been described in an earlier section.

3.12 Statistical analyses

Statistical analyses were performed using the R package. Fisher's exact tests were employed for comparisons among categorical variables, whereas Student's t-test was used to compare continuous variables and ordered categorical variables.

4 Results

4.1 Exome sequencing

To identify potential genomic risk factors for invasive cancer relapse, mutation patterns obtained from exome sequencing data of 21 DCIS tumors and normal samples were compared among relapsed and nonrelapsed patients. Table 1 shows the clinical information of the 21 samples used in this analysis, which includes the patient's age, expression of estrogen receptor (ER) and progesterone receptor (PgR) determined using immunohistochemistry (IHC), human epidermal growth factor receptor 2 (HER2) amplification, subtypes of breast cancer, and relapse status. The median follow-up period was 5.0 years (2.5–7.0 years) (Table 2), with 9 of 21 cases relapsing. In addition, relapse specimens were collected and sequenced in four of the nine relapsed cases. Radiation therapy (RT) or adjuvant therapy after surgery is known to reduce the risk of relapse (30). In these datasets, 14 patients were treated with RT or adjuvant therapy. However, 8 of 14 patients relapsed regardless of receiving either treatment (Table 2).

Exome sequencing data were obtained at an average depth of 108× for tumor samples and 107× for normal samples. The duplication rates were 20%–40%, which was reasonable for samples derived from FFPE tissues（Fig. 2). The average number of detected mutations was an 30. The VAF of many detected mutations was less than 10% (Fig. 3).

Mutation analysis results revealed that *GATA3* and *PIK3CA* were the most frequently mutated genes in all 21 cases. *GATA3* mutations were detected in 4 out of 21 cases (19%), and *PIK3CA* mutations were detected in 5 out of 21 (24%) cases (Fig. 4a). Therefore, I focused on these two genes.

For *PIK3CA* mutations, all five cases were confirmed as nonrelapse (Fig. 4b), suggesting that a *PIK3CA* mutation is associated with a lower risk of relapse. For *GATA3* mutations, IDC relapse was confirmed to three out of four cases, with *GATA3* mutations being identified more frequently in relapse cases. To evaluate the association between relapse and *GATA3* mutations, I

calculated the odds ratio (OR), which was 5.5 (95% CI 0.45–154.3). Although the OR was high, it was not statistically significant.

To investigate the correlation between *GATA3* mutations and relapse, the data obtained from relapsed tissue samples were analyzed. Relapse data in *GATA3* mutation cases were obtained from two of the four patients. Next, the mutation patterns between the primary and relapse cases were compared, and the presence of *GATA3* mutations in relapse was identified (Fig. 5). Similarly, clonality analysis using PyClone confirmed the presence of a population with *GATA3* mutations after relapse (Fig. 6). This suggests that *GATA3* mutations are associated with relapse. Furthermore, two out of four patients with *GATA3* mutations had a VAF of <10%. The VAF in *GATA3* mutations increased from 8% in the primary tumor to 39% in the relapse (Fig. 6), suggesting the existence of heterogeneity in the primary tumor.

Reports reveal that mutations in the zinc-finger domain of *GATA3* can result in the loss of DNA-binding ability, leading to a reduction in the expression of *PGR* and an increase in the expression of genes associated with invasion, such as EMT (31). Additionally, *GATA3* ChIP-seq datasets of T47D and MCF-7 breast cancer cell lines from ENCODE (32, 33) (ENCODE4 v1.6.1) databases (ENCSR000BMX and ENCSR423RTK) were used. While MCF-7 harbors the *GATA3* mutation (34), both cell lines show peaks on *PGR* promoter regions (Fig. 7). Therefore, the *PGR* gene is considered the target of *GATA3*. Although in MCF-7, a peak was detected on the *PGR* promoter, it is reported that *PGR* expression is decreased in three-dimensional cultures (35). Therefore, the *GATA3* mutation causes aberrations in *PGR* expression in MCF-7. Therefore, in my dataset, I confirmed the association between the PgR protein expression and mutations in 16 ER-positive samples. Seven samples showed low PgR expressions, and all four samples with *GATA3* mutations also showed low PgR expressions (Table 3). Interestingly, all *GATA3* mutations were frameshift insertions and located on the zinc-finger domain, which is an important domain related to the DNA-binding

ability of *GATA3*. These results suggest that in my dataset, as well as in previous studies, *GATA3* mutations are associated with invasive recurrence.

Overall, these results strongly suggest that *GATA3* mutations are associated with relapse. Therefore, to validate *GATA3* as a potential high-risk factor and *PIK3CA* as a potential low-risk factor, I performed targeted sequencing.

Table 1. Clinical information of the 21 whole-exome samples.

Age		
	>=45	9
	<45	12
ER stats		
	Neg	5
	pos	16
PgR stats		
	Neg	9
	pos	12
HER2 amp		
	Neg	13
	pos	8
Subtype of Breast Cancer		
	Triple Pos	5
	Luminal A	9
	Luminal B	2
	HER2	3
	Triple Neg	2
Relapse		
	Yes	9
	No	11

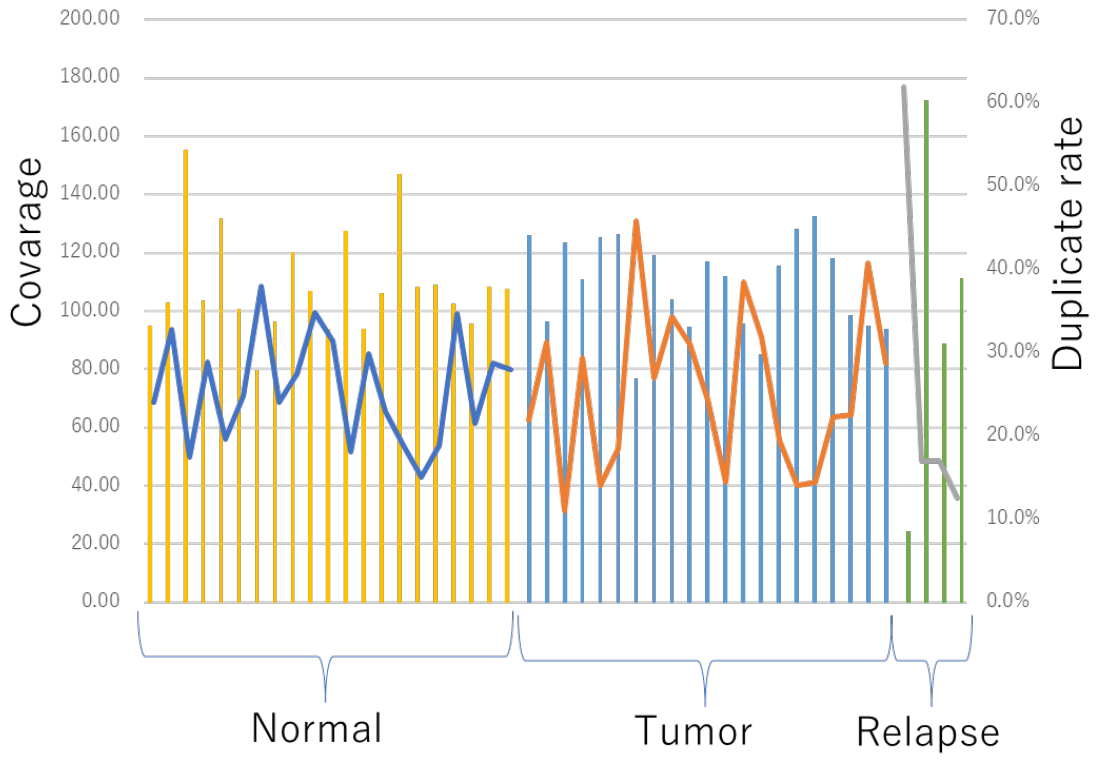


Fig. 2 Sequencing statistics of exome sequencing samples.

Bar plots of sequencing depth (left y-axis); yellow, normal sample; blue, tumor sample; green, relapse sample. Line plots of duplication rates (right y-axis); blue, normal sample; orange, tumor sample; gray, relapse sample.

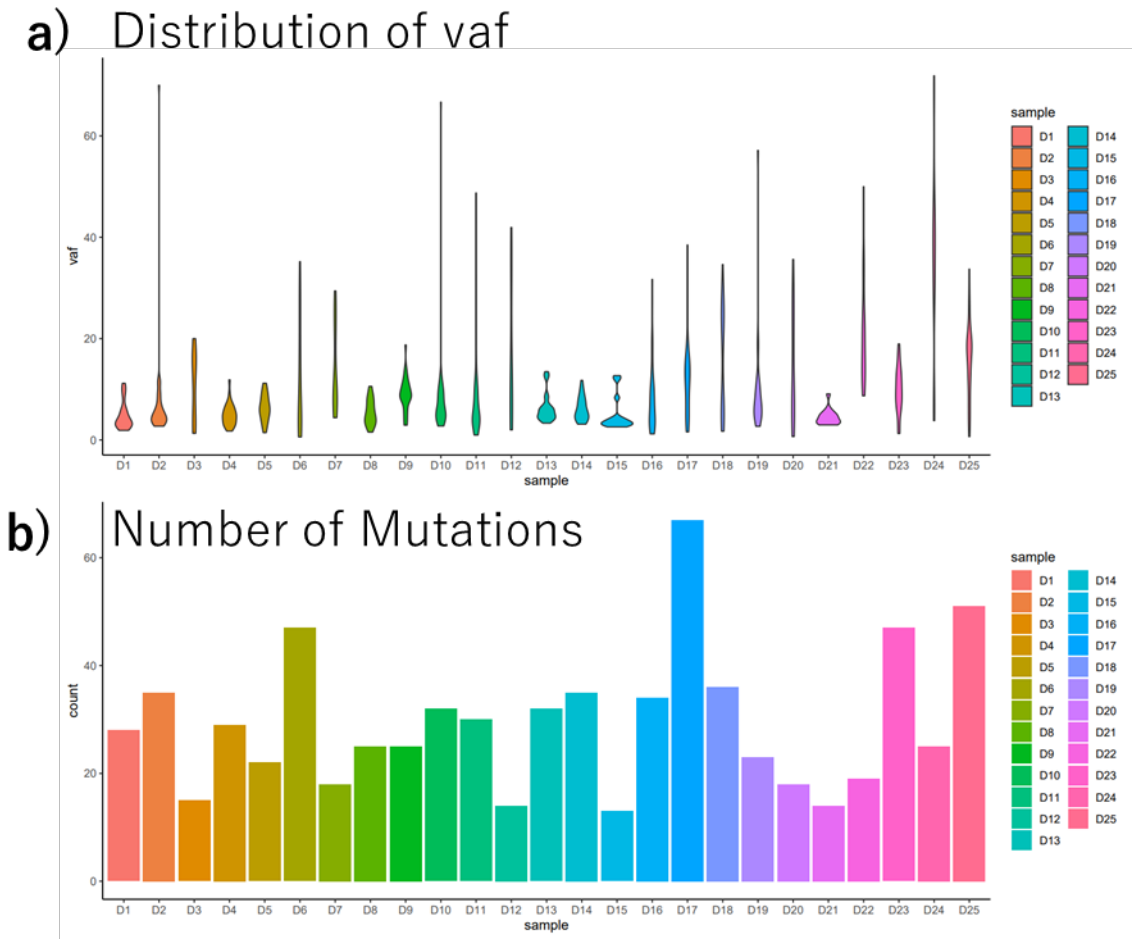


Fig. 3 Mutation frequency and variant allele frequency (VAF).

a). Violin plot of VAF of detected mutations in each patient. The x-axis represents the sample name. The y-axis represents the VAF. D1–21 indicate primary specimens, and D22–25 indicate relapse specimens. b). Mutation frequency in each patient.

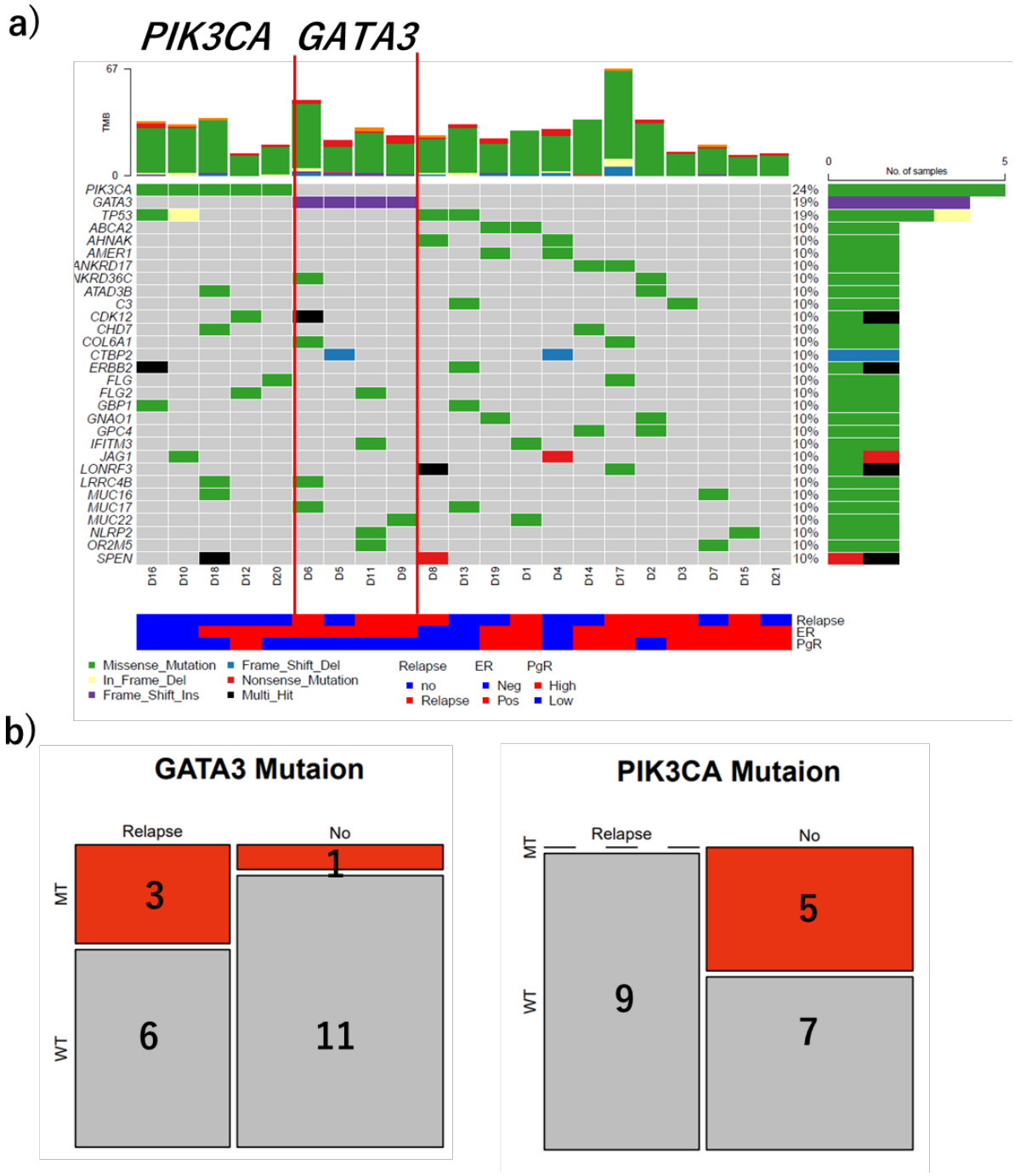


Fig. 4 Mutation profile of 21 patients with DCIS and genomic risk factors for relapse.

a). Oncoplot of somatic mutations (SNV, InDels) in 21 patients with DCIS. Top panel: mutation frequency in each patient. Middle panel: mutation profile of the top 30 most frequently mutated genes. Green, missense mutation; yellow, in-frameshift deletion; purple, frameshift insertion; blue, frameshift deletion; red, nonsense mutation; black, multihit. Right panel: percentage of patients with each gene mutation. Bottom panel: clinical information for

each sample. Relapse stats: red, relapse; blue, no relapse. ER stats: red, positive; blue, negative. PgR stats: red, high; blue, low. b). A mosaic plot of cancer relapse cases with gene mutation background. Red, number of patients with gene mutations; gray, number of nongene-mutated patients. Left panel: *GATA3* mutations. Right panel: *PIK3CA* mutations.

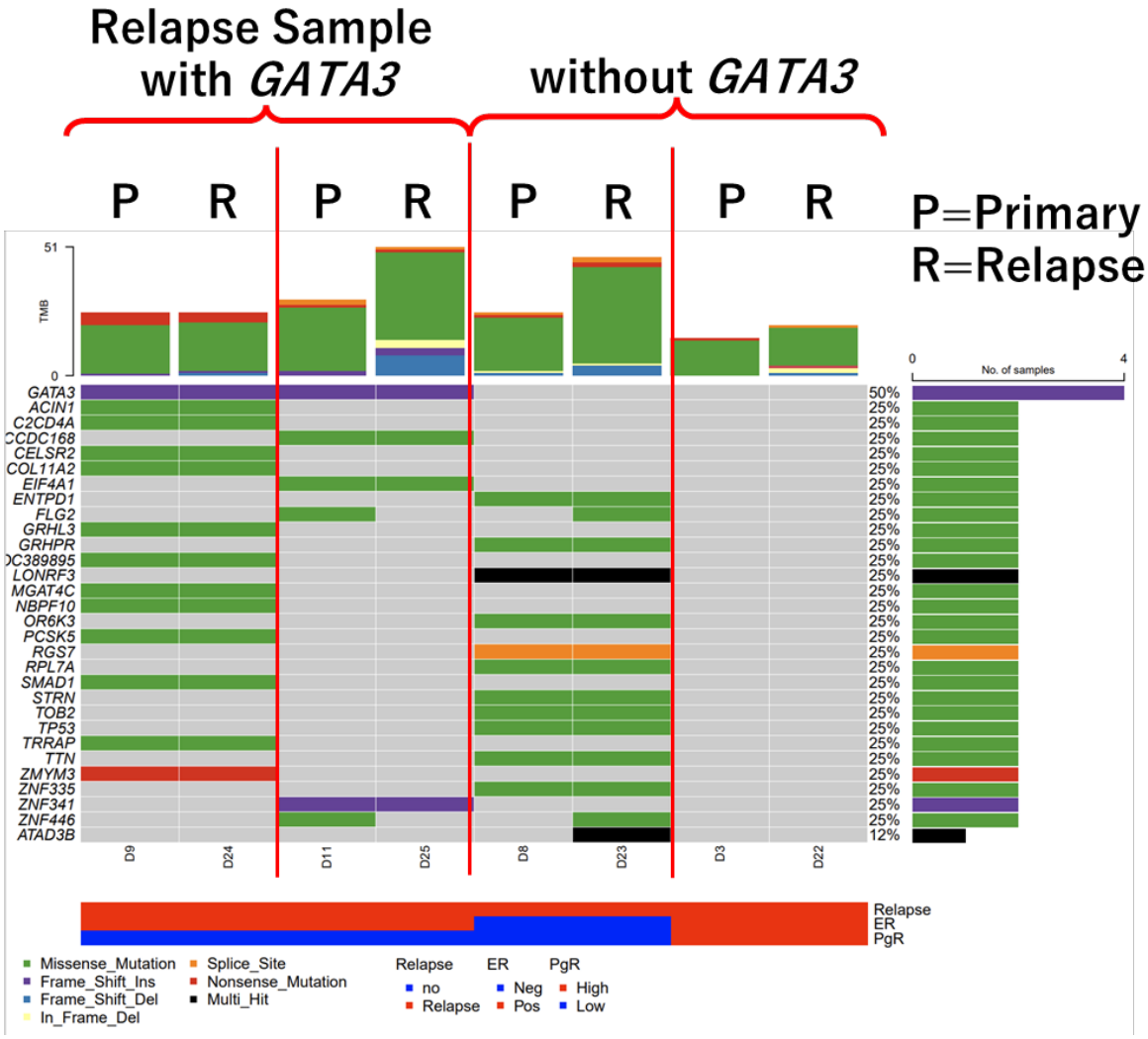


Fig. 5 Mutation profile of the relapse samples

Oncoplot of four patients (eight specimens) experiencing a relapse. P, primary specimen; R, relapse specimen. Left: two patients (four specimens) harbored *GATA3* mutations in primary and relapse cases.

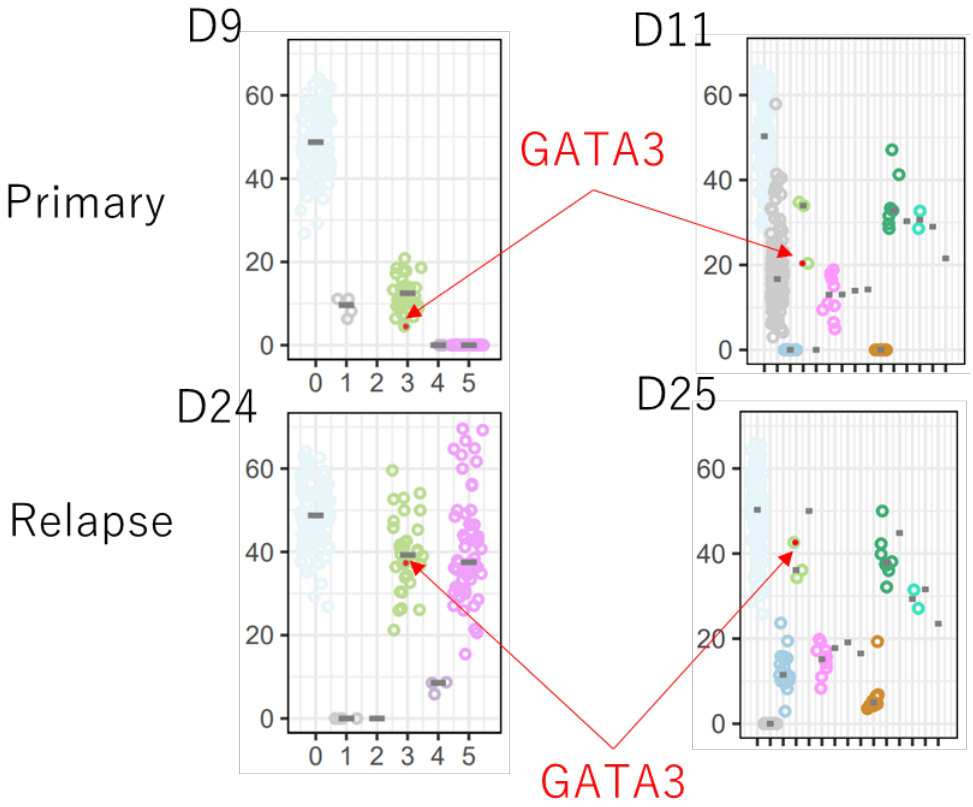


Fig. 6 Clonality changes in relapse samples with *GATA3* mutations.

Jitter plot of variant's VAF clustered using PyClone. The top two panels are primary specimens, whereas the bottom two panels are relapse specimens with *GATA3* mutations belonging to two patients. In the left panel, the patient's *GATA3* mutation is in Cluster 3, with primary VAF at 9% and 39% in relapse. In the right panel, the patient's *GATA3* mutation is in Cluster 3, with primary VAF at 22% and 45% in relapse.

Table 2. Detailed clinical information of the 21 whole-exome samples.

WES_No	Relapse	RFS_year	Radiotherapy	Adjuvant	*GATA3*	*PIK3CA*
1D	1	3.9				
10D	0	6.0	TRUE			H1047R
11D	1	5.8	TRUE		N334fs	
12D	0	6.0				H1047R
13D	0	6.0	TRUE			
14D	0	5.8	TRUE			
15D	1	4.2	TRUE			
16D	0	5.8				E545K
17D	1	4.5		TAM		
18D	0	5.0		TAM		H1047R
19D	0	5.6				
2D	1	3.3	TRUE	TAM		
20D	0	5.4				P366R
21D	0	4.5	TRUE			
3D	1	3.2		LET→TAM		
4D	0	7.0	TRUE			
5D	0	2.5		TAM	S437fs	
6D	1	4.1			W329fs	
7D	0	5.1				
8D	1	2.7	TRUE			
9D	1	3.6	TRUE	TAM	S408fs	

Table 3. Association between *GATA3* mutations and PgR expression in ER-positive cases.

	PgR low	PgR high	total
***GATA3* MT**	4	0	4
***GATA3* WT**	3	9	12
	7	9	16

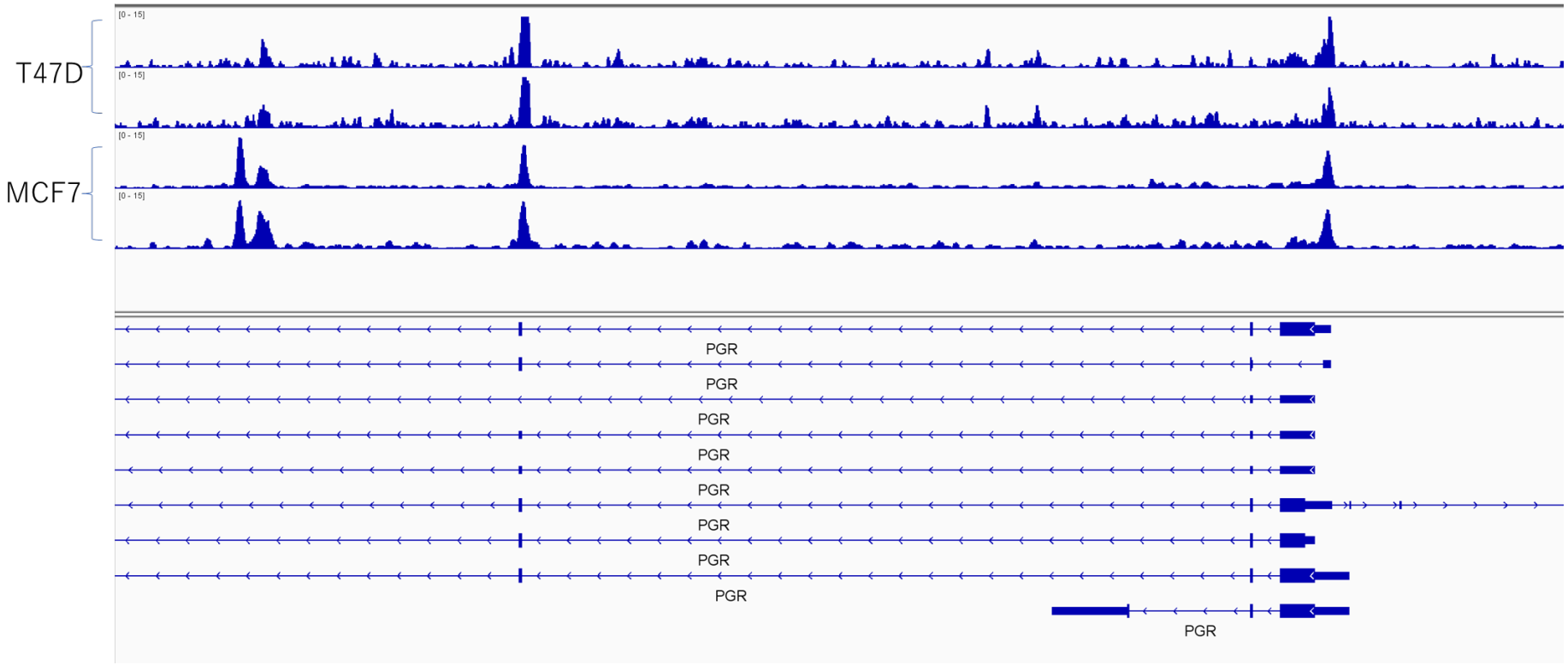


Fig. 7 IGV screenshot of *GATA3* Chip-seq results around *PGR* transcription start sites.

The top two panels indicate the result of the T47D cell line, whereas the following two panels indicate the result of the MCF7 cell line. In both cell lines, a peak was detected on the *PGR* promoter region, indicating the region to be the binding site of *GATA3*.

4.2 Target sequencing results

To validate the association between relapse and mutations in *GATA3* and *PIK3CA*, deep-targeted sequencing of 72 DCIS patient specimens was conducted. Table 5 lists the 180 target genes of deep targeted sequencing. Table 4 presents the clinical information of the analyzed samples. The median follow-up period was 5.5 years (0.5–9.6 years) (Table 6), during which 9 of the 72 cases relapsed. The relapse rate in this dataset is approximately 10% (Table 4). Previous study reported relapse in 14.6% of cases at a follow-up of 11 years (36), which was consistent with the relapse rate observed in this study.

Target sequencing data were obtained at an average depth of 611× for the tumor samples. The duplication rates were 60%–90%, which is reasonable for samples derived from FFPE tissues (Fig. 8). The sequencing depth of several samples was very low, which was taken care of during analysis.

GATA3 and *PIK3CA* mutations were detected in 40/72 and 36/72 patients, respectively (Fig. 9a). Eight of nine relapsed cases were identified to harbor *GATA3* mutations. The association between relapse and *GATA3* mutations was evaluated, and the OR was found to be 7.5 (95% CI 1.01–174.7), indicating a significantly increased risk of relapse (Fig. 9b). *PIK3CA* mutations were detected in three relapsed cases. However, the presence of this mutation was associated with a lower risk of recurrence. Additionally, two of the three relapsed cases with *PIK3CA* mutations had *GATA3* mutations. No other somatic mutations were found to be associated with relapse (Fig. 11).

However, genome analysis showed that several nonrelapsed samples also harbored *GATA3* mutations. The median follow-up time for nonrelapsed patients was 5.5 years and the relapse time in relapse patients was 5.6 years, with no difference. However, the progression of DCIS may take more than 10 years. Because patients analyzed in this study were only followed up for

a median of 5.5 years, patients with *GATA3* mutations may have concerns about a high risk of relapse, which should be taken care of during prognosis.

GATA3 is one of the fundamental transcription factors that maintains normal luminal epithelial differentiation. Therefore, loss of *GATA3* can cause luminal cell proliferation, detachment from the membrane, and eventually metastasis, which can be concerning (37). Furthermore, *GATA3* inhibits invasion and metastasis of breast cancer, indicating that the loss of *GATA3* promotes tumor progression (38-41). Most frequently, these mutations are frameshift mutations in exons 5 and 6, which are loss-of-function mutations. In this study dataset, 28 of 40 *GATA3* mutations were detected in the zinc-finger domain or exon 6 region (Fig. 10) and may be causing loss of function.

PIK3CA is known to play a pivotal role in the catalytic subunit of the phosphatidylinositol 3-kinase (PI3K) signaling pathway, which is associated with tumorigenesis (42). *PIK3CA* mutations were thought to promote the kinase activity of *PIK3CA* and contribute to tumorigenesis. *PIK3CA* H1047R and E545K are known as hotspot and gain-of-function mutations. In my target sequencing datasets, $PIK3CA^{H1047}$ mutations were the most frequent. Nevertheless, *PIK3CA* mutations suggest a low-risk factor. The patients harboring *PIK3CA* mutations might have different effects from those of the previous studies.

Targeted sequencing analysis revealed that DCIS patients with *GATA3* mutations are at a higher risk of cancer progression, whereas those with *PIK3CA* mutations are at a lower risk. Furthermore, it suggests the heterogeneity of DCIS. Therefore, a spatial transcriptome analysis was conducted to reveal the heterogeneity of DCIS and the molecular mechanisms underlying its risk factors.

Table 4. Clinical information of the 72 target sequencing samples.

Age		
	>=45	**47**
	<45	**25**
ER stats		
	Neg	**20**
	Pos	**52**
PgR stats		
	Neg	**23**
	pos	**49**
HER2 amp		
	Neg	**49**
	pos	**32**
Subtype of Breast Cancer		
	Triple Pos	**33**
	Luminal A	**19**
	Luminal B	**5**
	HER2	**15**
Relapse		
	Yes	**9**
	No	**63**

Table 5. Gene list for target sequencing.

CDK12	*CELSR2*	*COL6A1*	*CUL3*	*EGFR*	*ERBB2*
FBXO22	*FLT3*	*GNAS*	*HRAS*	*INO80*	*KDM6A*
KRAS	*MAP2K2*	*MDM4*	*MSH2*	*NF1*	*NRAS*
NTRK3	*PCSK5*	*PIK3R2*	*PTCH1*	*RAD54L*	*RGS7*
RYR3	*SLC26A4*	*SMARCB1*	*SSC5D*	*TESK1*	*TTC16*
ABL1	*AKT3*	*ARID1A*	*AXL*	*BRCA1*	*CCDC168*
CDK4	*CHD7*	*CREBBP*	*DCHS1*	*EIF4A1*	*ERBB3*
FGFR1	*FOXA1*	*GRHL3*	*IDH1*	*JAK1*	*KDR*
LOC389895	*MAP2K4*	*MED12*	*MTOR*	*NFE2L2*	*NRG1*
OK7K1P	*PDGFRA*	*POLD1*	*PTEN*	*RAF1*	*RHOA*
SCN4B	*SLC2A12*	*SMO*	*STAT3*	*TOB2*	*TTN*
ACIN1	*ALK*	*ARID2*	*BAP1*	*BRCA2*	*CCND1*
CDK6	*CHEK2*	*CRKL*	*DDR2*	*ENO1*	*ERBB4*
FGFR2	*GATA3*	*GRHPR*	*IDH2*	*JAK2*	*KEAP1*
LONRF3	*MAP3K1*	*MERTK*	*MYC*	*NOTCH1*	*NT5C2*
OR6K3	*PDGFRB*	*PPL*	*RAC1*	*RB1*	*ROS1*
SETBP1	*SMAD1*	*SOX9*	*STK11*	*TP53*	*TYRO3*
ACTN4	*APC*	*ARR3*	*BARD1*	*C2CD4A*	*CD274*
CDKN1B	*CLCA2*	*CTCF*	*DNMT3A*	*EP300*	*ESR1*
FGFR3	*GNA11*	*HIRIP3*	*IGF1R*	*JAK3*	*KIT*
LTK	*MAP3K4*	*MET*	*MYCN*	*NOTCH2*	*NTRK1*
PALB2	*PIK3CA*	*PPP2R1A*	*RAC2*	*RBMXL3*	*RUNX1*
SETD2	*SMAD4*	*SPEN*	*STRA13*	*TRRAP*	*VHL*

AKT1	*AR*	*ATM*	*BCL2L11*	*CBFB*	*CDH1*
CDKN2A	*COL11A2*	*CTNNB1*	*ECEL1*	*EPHB4*	*EZH2*
FGFR4	*GNAQ*	*HMGCS2*	*IGF2*	*KDM5C*	*KMT2A*
MAP2K1	*MDM2*	*MLH1*	*MYH13*	*NOTCH3*	*NTRK2*
PBRM1	*PIK3R1*	*PRKCI*	*RAD51C*	*RET*	*RUVBL1*
SF3B1	*SMARCA4*	*SPTB*	*TBX3*	*TSC1*	*ZNF341*

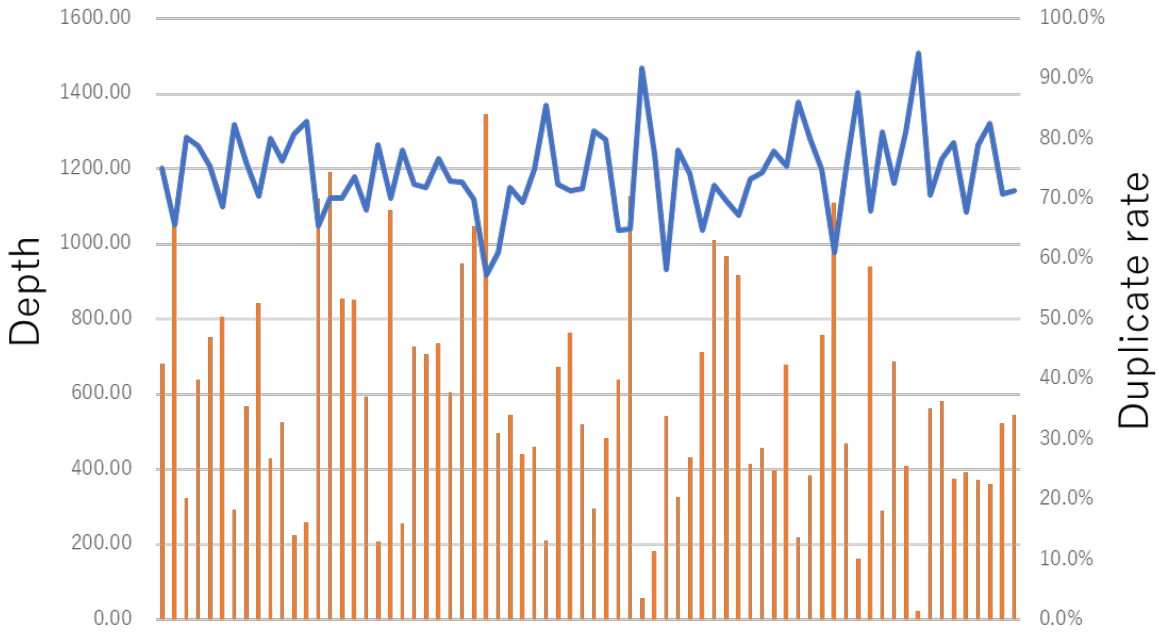


Fig. 8 Sequencing statistics of 72 target sequencing samples.

Bar plots indicate sequencing depth (left y-axis). Line plots indicate duplication rates (right y-axis).

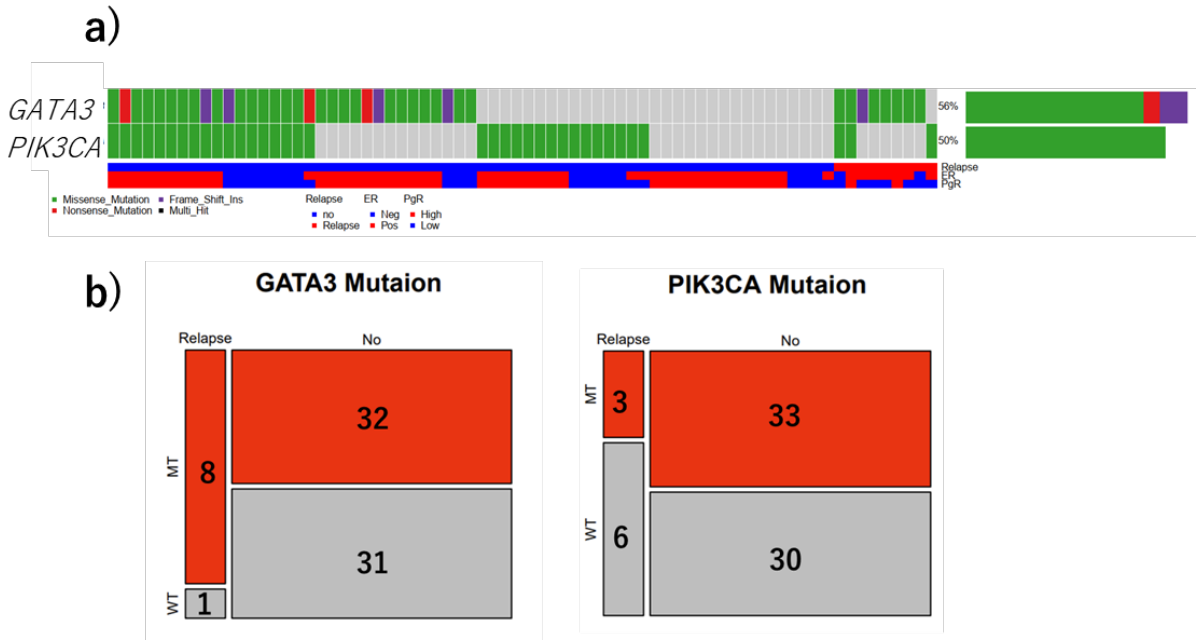


Fig. 9 *GATA3* and *PIK3CA* mutations in 72 patients with DCIS.

a). Oncoplot of *GATA3*/*PIK3CA* mutations (SNV, InDels) in 72 patients with DCIS. Top panel: mutation profile of *GATA3*/*PIK3CA*. Green, missense mutation; purple, frameshift insertion; red, nonsense mutation; black, multihit. Right panel: Percentage of patients with each gene mutation. Bottom panel: clinical information of each sample. Relapse stats: red, relapse; blue, no-relapse. ER stats: red, positive; blue, negative. PgR stats: red, high; blue, low.

b). A mosaic plot of cancer relapse cases with gene mutation backgrounds. Red, number of patients with gene mutations; blue, number of nongene-mutated patients. Left panel: *GATA3* mutations; right panel: *PIK3CA* mutations.

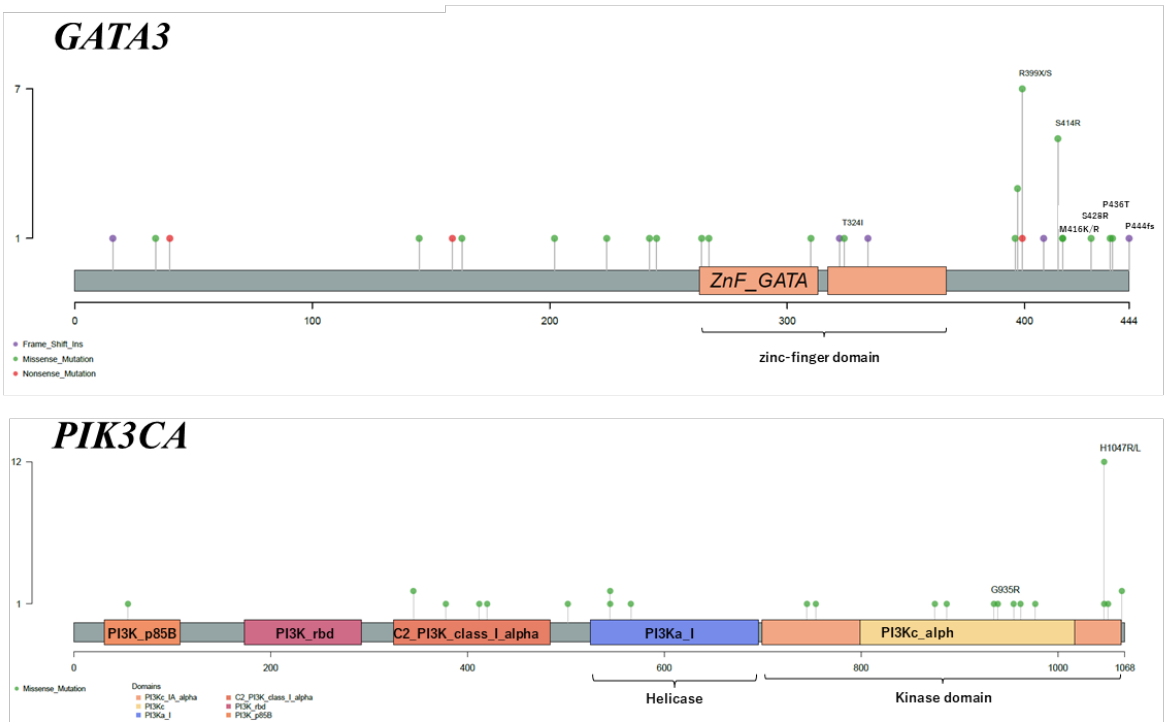


Fig. 10 Lollipop plots of *GATA3* and *PIK3CA* mutations.

Top panel: lollipop plot of *GATA3* mutations. The mutations detected by target sequencing were plotted, and the relapse sample mutations were highlighted. All mutations in relapse samples are located after the zinc-finger domain. Bottom panel: lollipop plot of *PIK3CA* mutations. All mutations in relapse samples are located in the kinase domain region.

Table 6. Detailed clinical information of 72 deep target sequencing samples.

Sample ID	Relapse	RFS_year	radiotherapy	Adjuvant	GATA3	PIK3CA
V_002	0	6.1	TRUE			p.H1047R
V_006	0	7.0				p.H1047R
V_015	0	6.9	TRUE		p.S34F	p.H1047R
V_020	0	5.0	TRUE			
V_073	0	6.0	TRUE	TAM		
V_074	0	5.1	TRUE			
V_075	0	5.1	TRUE		p.T245P	
V_076	0	5.0	TRUE		p.P224L	
V_077	0	5.2				p.E545K
V_078	0	5.0	TRUE		p.Q40X	p.H1065Q
V_079	0	4.0	TRUE		p.L397H	
V_080	0	5.0			p.S414R	
V_081	0	5.3	TRUE	TAM	p.R399X	
V_082	0	4.9	TRUE	TAM		p.H1047R
V_083	0	4.4	TRUE	TAM		
V_084	0	5.5	TRUE		p.A396T	p.H1047R
V_085	0	1.0	TRUE	TAM		p.H1047R
V_086	0	5.0				p.F977L
V_087	0	5.0			p.S414R	p.R502Q
V_088	0	4.9		TAM		
V_089	0	4.9			p.R399S	p.H1047R
V_090	0	4.9	TRUE			
V_091	0	2.6				
V_092	0	3.5				
V_093	0	5.2	TRUE	TAM		
V_094	0	4.9		TAM		
V_095	0	5.0		TAM		
V_096	0	7.7		TAM	p.N334fs	
V_097	0	8.8	TRUE		p.S414R	
V_098	0	8.1	TRUE			p.H1047R
V_099	0	7.8			p.R202C	p.W1051L
V_100	0	7.2	TRUE		p.R399S	p.V955I
V_101	0	6.2		ANA	p.G242S	p.C378Y
V_102	0	5.3	TRUE		p.A310V	
V_103	0	6.0			p.Q322fs	p.E545A
V_104	0	6.0			p.S414R	
V_105	0	6.1	TRUE		p.L397H	p.R412Q
V_106	0	5.6	TRUE	TAM	p.P163L	
V_107	0	5.0				p.G887E
V_108	0	5.6	TRUE			
V_109	0	5.1	TRUE	TAM	p.C264R	
V_110	0	9.6			p.P16fs	p.P566S
V_111	0	8.6	TRUE		p.R399S	p.D939Y
V_112	0	8.7	TRUE		p.S437P	p.H875N
V_113	0	7.1	TRUE		p.R399S	p.I962K
V_114	0	7.9	TRUE			
V_115	0	5.0			p.S408fs	
V_116	0	7.0				p.N345K
V_117	0	6.1				p.H1047R
V_118	0	7.1	TRUE		p.H145Y	p.H1065Q
V_119	0	6.4			p.C267W	
V_120	0	7.0	TRUE		p.L397H	
V_121	0	6.5			p.R399S	p.H1047R
V_122	0	5.1	TRUE			
V_123	0	1.5	TRUE		p.R399S	p.P754L
V_124	0	0.5				p.C420R
V_125	0	6.0				p.E545K
V_126	0	5.0	TRUE			
V_127	0	5.1	TRUE			p.M745I
V_132	0	6.8	TRUE	ANA		p.N345K
V_134	0	6.0	TRUE		p.K159X	p.K55R
V_137	0	6.1	TRUE	TAM		
V_144	0	7.1	TRUE	ANA		p.H1047R
V_145	1	5.7	TRUE	Lue,TAM	p.G444fs	
V_147	1	6.2	TRUE	ANATAM	p.R399S	
V_149	1	4.6			p.M416K	
V_151	1	1.2			p.M416R	
V_153	1	7.0	TRUE		p.P436T	
V_157	1	5.1		CEDOCTmab	p.S428R	p.H1047R
V_158	1	2.8	TRUE		p.T324I	p.H1047L
V_160	1	6.0	TRUE		p.S414R	
V_162	1	7.5				p.G935R

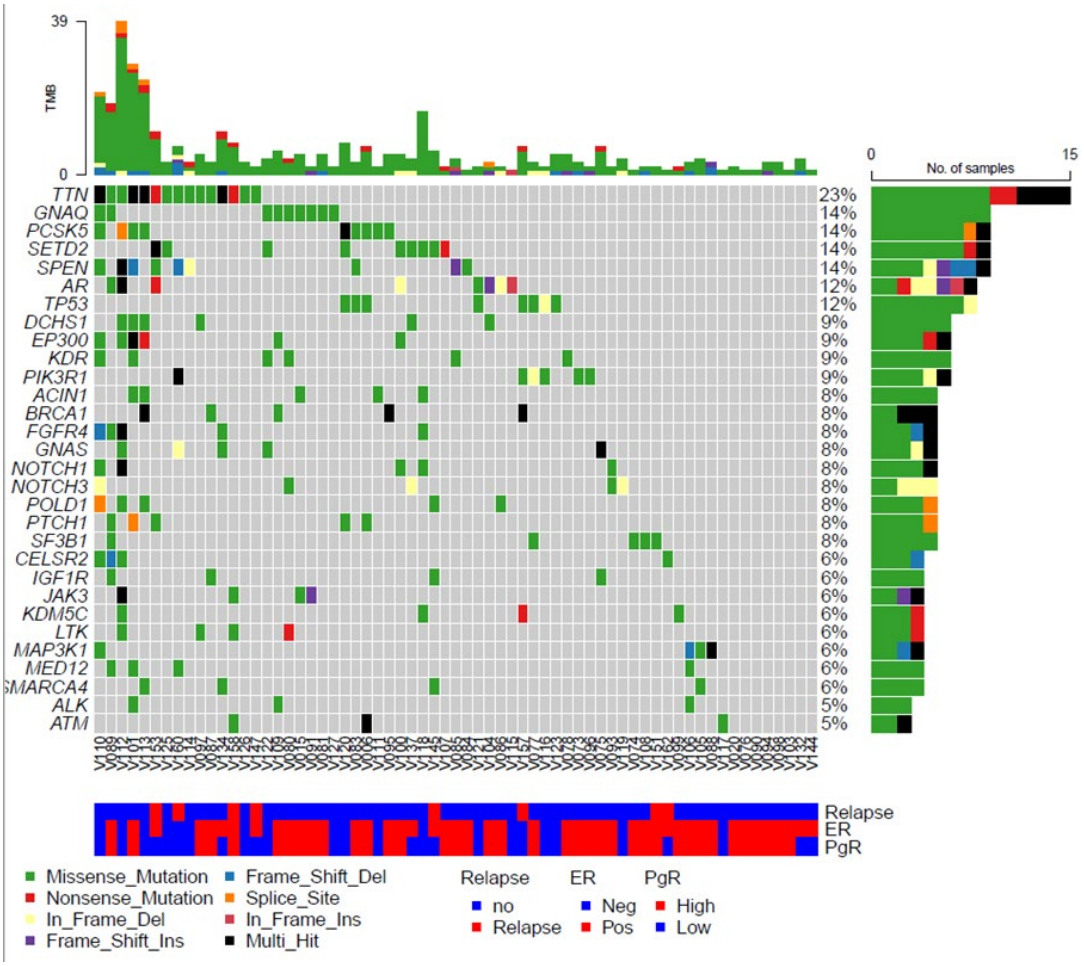


Fig. 11 Mutation profile of 72 patients with DCIS.

Oncoplot of somatic mutations (SNV, InDels excluding *GATA3*/*PIK3CA*) in 72 patients with DCIS. Top panel: mutation frequency in each patient. Middle panel: mutation profile of top 30 most frequently mutated genes. Green; missense mutation; yellow, in-frameshift deletion; purple, frameshift insertion; blue, frameshift deletion; orange, splice site mutation; red, nonsense mutation; black: multihit. Right panel: percentage of patients with each gene mutation. Bottom panel: clinical information of each sample. Relapse stats: red, relapse; blue, no-relapse. ER stats: red, positive; blue, negative. PgR stats: red, high; blue, low.

4.3 Summary of the spatial transcriptome analysis of three samples.

Table 7 presents the clinical information of the three patients with DCIS used for spatial transcriptome analysis. Hematoxylin and eosin (HE)-stained images are shown in Fig. 12.

I conducted a genome sequencing analysis and confirmed the mutations in the genomic risk factor genes (*GATA3*/*PIK3CA*). In Cases A and B, deep target sequencing was conducted, whereas in Case C, whole-exome sequencing was performed. Table 8 shows the sequence statistics of genome sequences. Genome analysis revealed that Case A harbored both *GATA3* and *PIK3CA* mutations, whereas Cases B and C harbored only the *PIK3CA* mutation.

Summary of genomic mutations and clinical information

・Case A harbors a *GATA3* mutation and predicts a high risk of invasion, with microinvasion observed.

・Case B harbors only *PIK3CA* mutations and no *GATA3* mutations, with invasion observed.

・Case C harbors only a *PIK3CA* mutation with no invasion.

Table 9 shows the sequence statistics of the spatial transcriptome. Sequence data of about 700 million to 1.2 billion reads were obtained. The number of detected unique molecular identifiers (UMIs) was 7,469–44,987, and the number of detected genes was 2,928–8,745. The number of detected genes was low only in Case A, which may be due to the large area occupied by stromal cells. However, in the cancer area, sufficient data were obtained. The results of the analysis for each case were then described.

Table 7. Clinical information of the three patients used for spatial transcriptome analysis

	Case A	**Case B**	**Case C**
Age	41	56	43
Management	Mastectomy	Mastectomy	Mastectomy
Follow-up (year)	5.5	4.8	3.2
ER status	Positive (70%)	Positive (90%)	Positive (90%)
PgR status	Positive (90%)	Negative (0%)	Positive (90%)
HER2 amplification (FISH)	3+ (n.a.)	2+ (1.14)	2+ (1.1)
Nuclear atypia *in situ*	Intermediate	Intermediate	Low
Comedo necrosis	Positive	positive	Negative
Tumor size (cm)	1.2	6.5	2.9
GATA3	+	-	-
PIK3CA	+	+	+

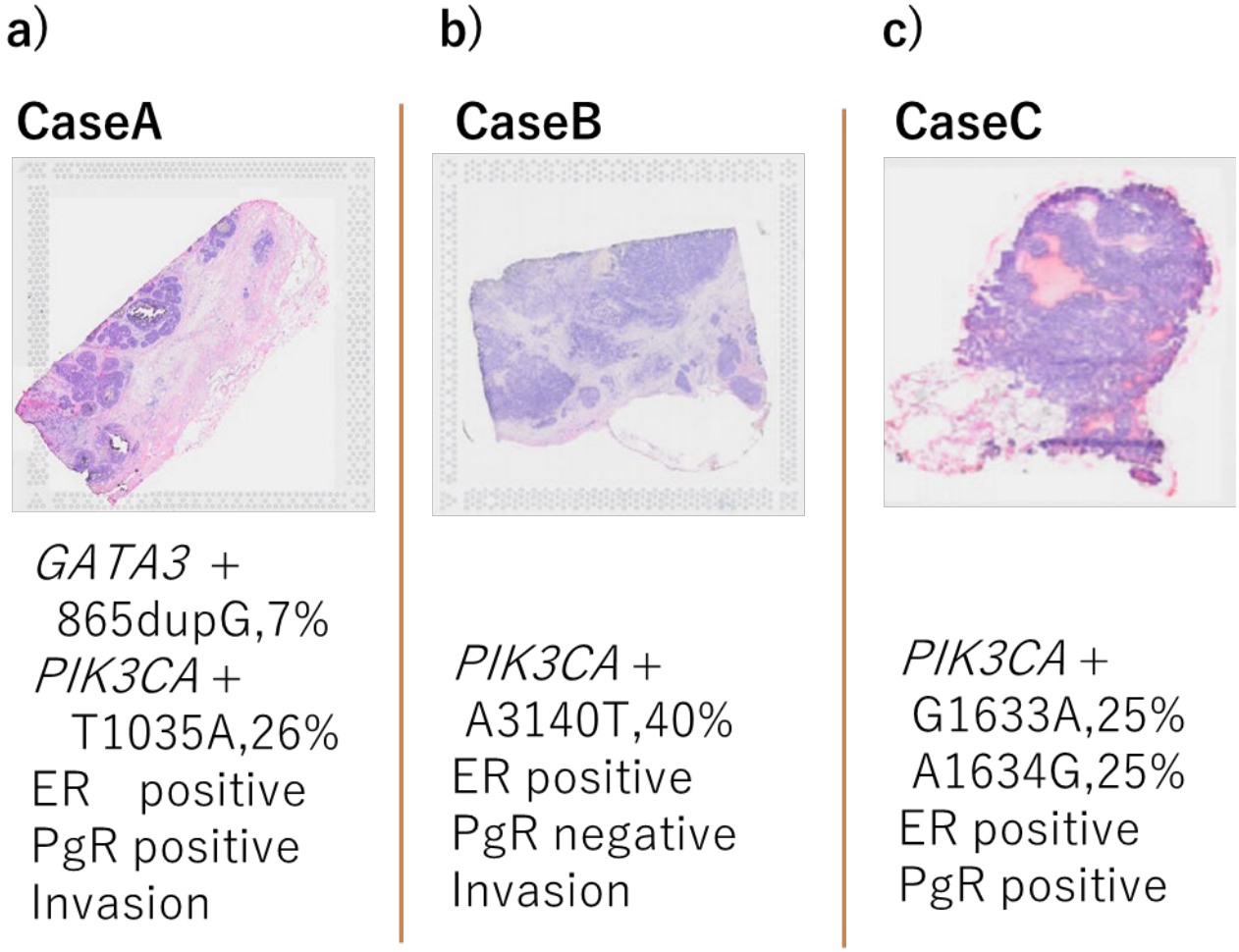


Fig. 12 Hematoxylin and eosin (HE)-stained images of the three specimens.

HE-stained images of each sample. a). Case A harbored *GATA3* and *PIK3CA* mutations, with the detection of microinvasion; ER, positive; PgR, positive. b). Case B harbored *PIK3CA* mutations, with invasion detected; ER, positive; PgR, negative. c). Case C harbored *PIK3CA* mutations, with no invasions detected; ER, positive; PgR, positive.

Table 8. Genome sequencing statistics of the three patients with DCIS.

Case	Data type	Raw fragment	Mapped fragment		Duplication rate	After the removal of duplication	ON target rate	Mean voverage
Case A	Target-seq	56,104,374	53,607,083	96%	42.7%	30,876,061	74%	2391.41
Case B	Target-seq	117,018,611	108,965,586	93%	61.4%	42,281,449	69%	3030.61
Case C tumor	Exome-sequencing	73,361,320	60,006,129	82%	9.4%	54,688,417	71%	105.62
Case C normal	Exome-sequencing	75,284,727	62,938,934	84%	9.7%	57,207,814	69%	111.25

Table 9. Spatial transcriptome sequencing statistics of three patients with DCIS.

Sample ID	Number of spots	Median genes	Number of reads	Saturation	Mapped to genome	Mapped to intergenic	Mapped to intronic	Mapped to exonic	Mapped to transcriptome	Reads in spots	Median UMI*
Case A	2,403	2,928	700,133,428	86%	82%	3%	4%	71%	67%	92%	7,469
Case B	2,479	8,745	1,259,458,106	85%	91%	3%	2%	81%	77%	86%	44,987
Case C	594	8,687	1,171,436,207	93%	87%	9%	18%	47%	44%	72%	44,586

*UMI = unique molecular identifier

4.4 Case A results

This patient was preoperatively diagnosed with DCIS, but postoperative pathology confirmed microinvasion. Genomic analysis revealed that this patient had both *GATA3* and *PIK3CA* mutations. As expected from the genome analysis results, invasion was confirmed. The VAF of the *PIK3CA* mutation was 24% and that of the *GATA3* mutation was 7%, showing that the VAF of *GATA3* was lower than that of *PIK3CA*.

Spatial transcriptome analysis was conducted to reveal the heterogeneity of DCIS in this patient. At first, I manually extracted 422 pathological DCIS spots from all other tissue spots. Next, to examine the spots harboring *GATA3* mutations in 422 DCIS spots, I extracted reads harboring the *GATA3* mutation from bam file using a custom Perl script and confirm the spot barcode of extracted reads. As a result, 46 spots were identified with *GATA3* mutations (Fig. 13a). The sequencing depth of all *GATA3* showed no difference between mutated and wild-type spots (Fig. 13e). Additionally, I confirmed the depth of the *GATA3*-mutated locus (chr10:8106041_G_insertion). The number of spots with reads detected on the mutated locus was found to be 142. The distributions of reads in the *GATA3*-mutated spots were similar to the wild-type ones (Fig. 13c). Therefore, almost no spatial bias existed. Additionally, all identified *GATA3* mutations were in the same genomic locus, with the VAF of *GATA3* being 7%, suggesting that the *GATA3* mutation was acquired relatively late during the development of the DCIS, and was thus considered to be derived from the same clone.

To examine the molecular effect of *GATA3* mutations, I compared 46 spots with and 476 spots without *GATA3* mutations. *PGR* expression was significantly decreased in spots with *GATA3* mutations ($p = 0.029$) (Fig. 14b), suggesting that the genes associated with invasion were activated, as reported in previous studies. Subsequently, R's TCC package was used to identify DEGs, resulting in the detection of 1462 DEGs (Fig. 14c).

The pathway enrichment analysis results of the two groups using the top 100 FDR value genes confirmed that 14 pathways were enriched in *GATA3* mutated spots (Fig. 14d). Some of these pathways were related to invasion, such as epithelial–mesenchymal transition (EMT) and angiogenesis. Additionally, several key genes related to EMT and angiogenic expression were upregulated in *GATA3*-mutated spots (Fig. 14e, f). *VIM* expression was also upregulated when *GATA3* was mutated in luminal subtype breast cancer (31). In contrast, spots without *GATA3* mutations showed no enrichment of pathways associated with invasion and estrogen response but detected tight junctions, suggesting low malignancy (Fig. 14d, g).

These results suggest that in Case A, the invasion was caused by the decreased expression of *PGR* along with the activation of genes related to EMT and angiogenesis in the spot where the *GATA3* mutation occurred.

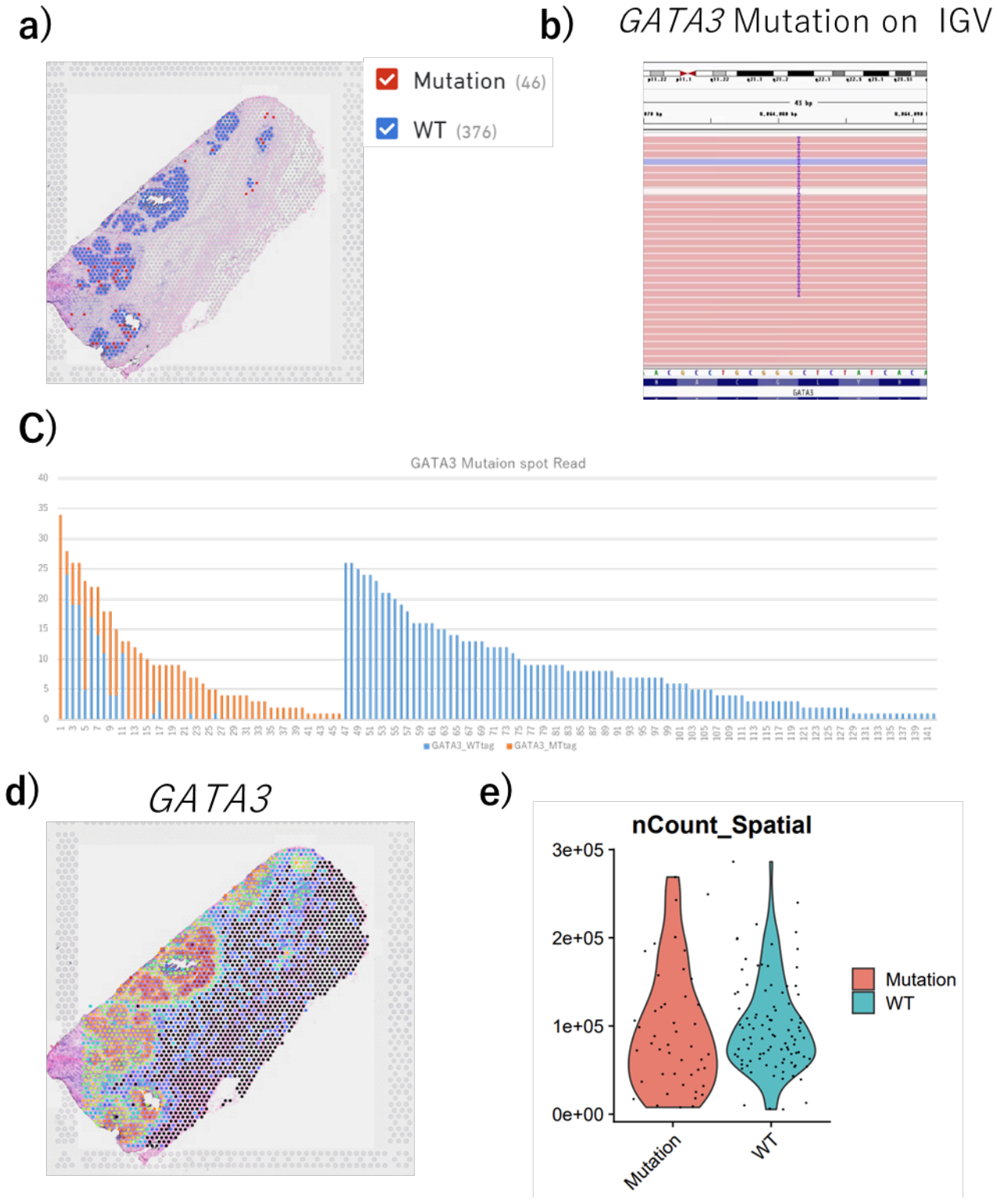


Fig. 13 Detection of spots with *GATA3* mutations.

a). Hematoxylin and eosin (HE)-stained images showed 46 *GATA3* mutated spots (red) and 376 wild-type spots (blue). b). The reads harboring mutations are shown on IGV. c). Read

distribution of *GATA3* mutation locus: orange, mutations read; blue, wild type read. d). HE-stained images showing *GATA3* expression. e). A violin plot of all *GATA3* reads. *GATA3* mutated spots are indicated in red and wild-type spots in blue. Both samples show sufficient *GATA3* expression.

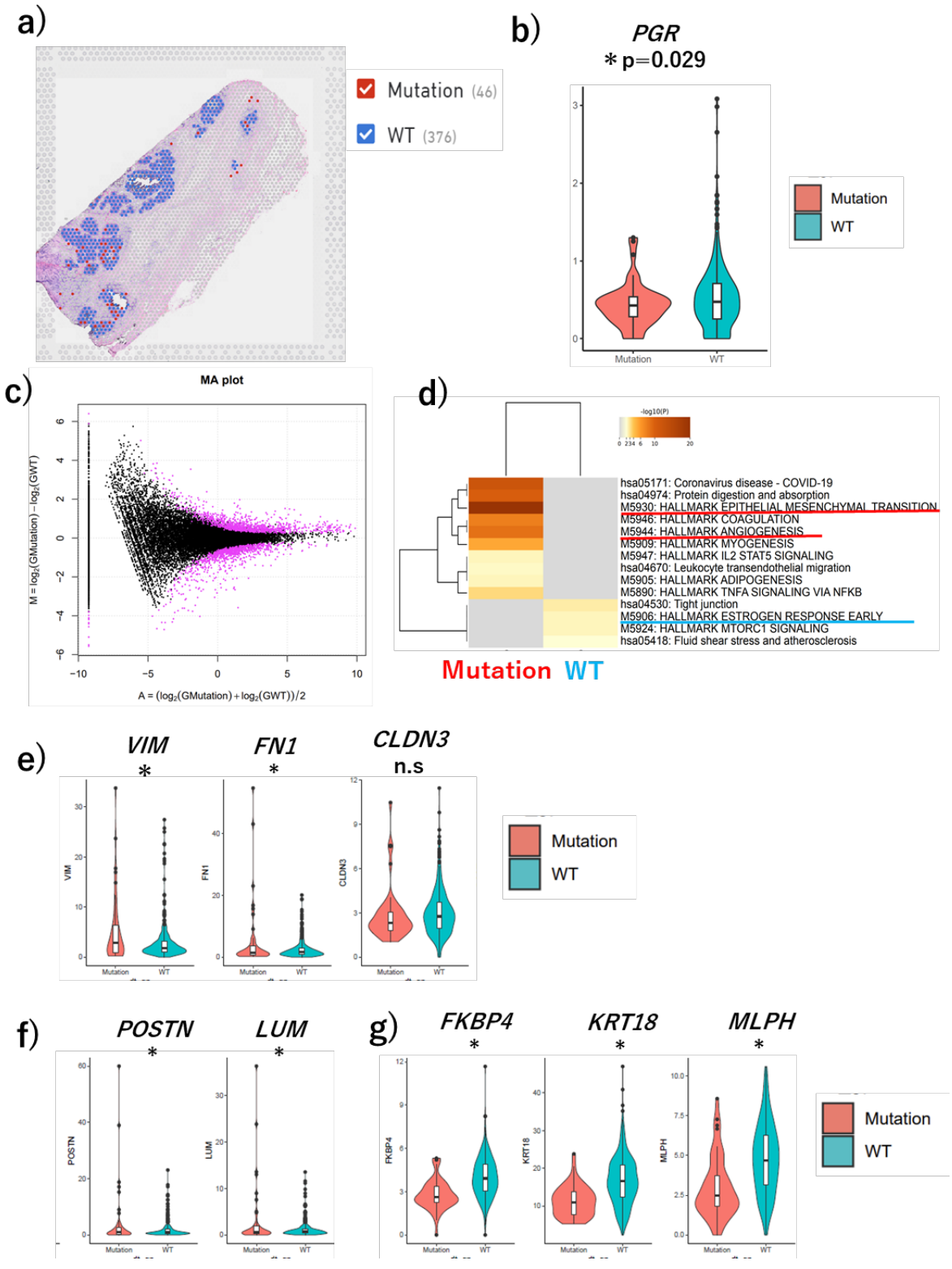


Fig. 14 Spatial transcriptome analysis of Case A harboring a *GATA3* mutation.

a). Visualization of spots with or without *GATA3* mutation on hematoxylin and eosin-stained images. Red, *GATA3* mutated 46 spots; blue, non-mutated 376 spots. b). Violin plot of *PGR*

expression at each spot. Decreased *PGR* expression was observed in the spots with *GATA3* mutations ($p = 0.029$; one-sided *t*-test). c). The MA plot of gene expression fold change between spots with or without *GATA3* mutations. The pink dot is FDR < 0.01. d). Result of enrichment analysis by Metascape. e), f), and g). Violin plots of key genes in each pathway. *$p < 0.05$ (one-sided), n.s.: not significant. e). Violin plot of key genes in the EMT pathway (*VIM*, *FN1*, and *CLDN3*). All genes showed increased expression in spots with *GATA3* mutations. f). Violin plot of key genes (*POSTN* and *LUM*) in the angiogenesis pathway. All genes showed increased expression in spots with *GATA3* mutations. g). Violin plots of key genes (*FKBP4*, *KRT18*, and *MLPH*) in the estrogen response pathway. All genes showed increased expression in spots without a *GATA3* mutation.

4.5 Case B results

This patient did not harbor *GATA3* mutations but showed only *PIK3CA* mutations and was thus expected to have a good prognosis based on genomic analysis. However, postoperative pathology showed invasion. Immunohistochemical analysis confirmed that the specimen was ER-positive and PgR-negative. Similarly, the gene expression of *PGR* and *ESR1* was found to be *ESR1*-positive/*PGR*-negative, which was consistent with the immunohistochemical staining results (Fig. 15a).

To examine the heterogeneity of the DCIS region, 188 spots, morphologically determined as DCIS, were manually extracted (Fig. 15b). Hierarchical clustering was performed on the gene expression data of the extracted regions, which was further divided into three subclusters (Fig. 15c). Next, the spatial and morphological information of the three identified clusters was confirmed. Cluster 1 (colored blue) overlapped at the center of the milk ducts, where DCIS was limited to the cells and did not show invasiveness yet (Fig. 16a). Importantly, Cluster 2 (colored red) was found outside the ducts of DCIS cells and was about to show invasiveness (Fig. 16a). These results suggest that Cluster 2 was an early stage of DCIS with acquired invasive potential. Gene expression was compared between Clusters 1 and 2, and 2766 DEGs were determined (Fig. 16c). Pathway analysis of the top 100 FDR values for each DEG using Metascape revealed significant enrichment in 20 pathways. Cluster 2 showed enrichment of pathways related to invasion, such as EMT and angiogenesis, which were also confirmed in the mutation spot, as in Case A (Fig. 16d). Similarly, several key genes related to EMT and angiogenesis were also upregulated, as in Case A (Fig. 16e, f). In contrast, Cluster 1 showed no enrichment of pathways associated with invasion, suggesting low malignancy (Fig. 16d). Interestingly, *GATA3* expression was significantly decreased in Cluster 2 (Fig. 16b), suggesting that although Case B did not harbor the *GATA3* mutation, aberration of *GATA3* expression activated genes involved in invasion and resulted in invasion,

as displayed in Case A.

To investigate the cause of the decreased *GATA3* expression, transcription factors (TFs) that regulate *GATA3* in TFlink (43) (https://tflink.net/), and a transcription factor database were further explored. As a result, 360 registered genes were found. Of these, 16 TF revealed only small-scale evidence. Among the TF genes with evidence, *MYB* expression showed a trend of decreased expression in Cluster 2 (nonsignificant) (Fig. 17). *MYB* is a transcription factor reported to target *GATA3* (44). Furthermore, it has been reported that high *MYB* expression suppresses breast cancer metastasis into the lung (45). These findings suggested that in Case B, decreased *MYB* expression caused aberrations in *GATA3* expression, eventually leading to invasion.

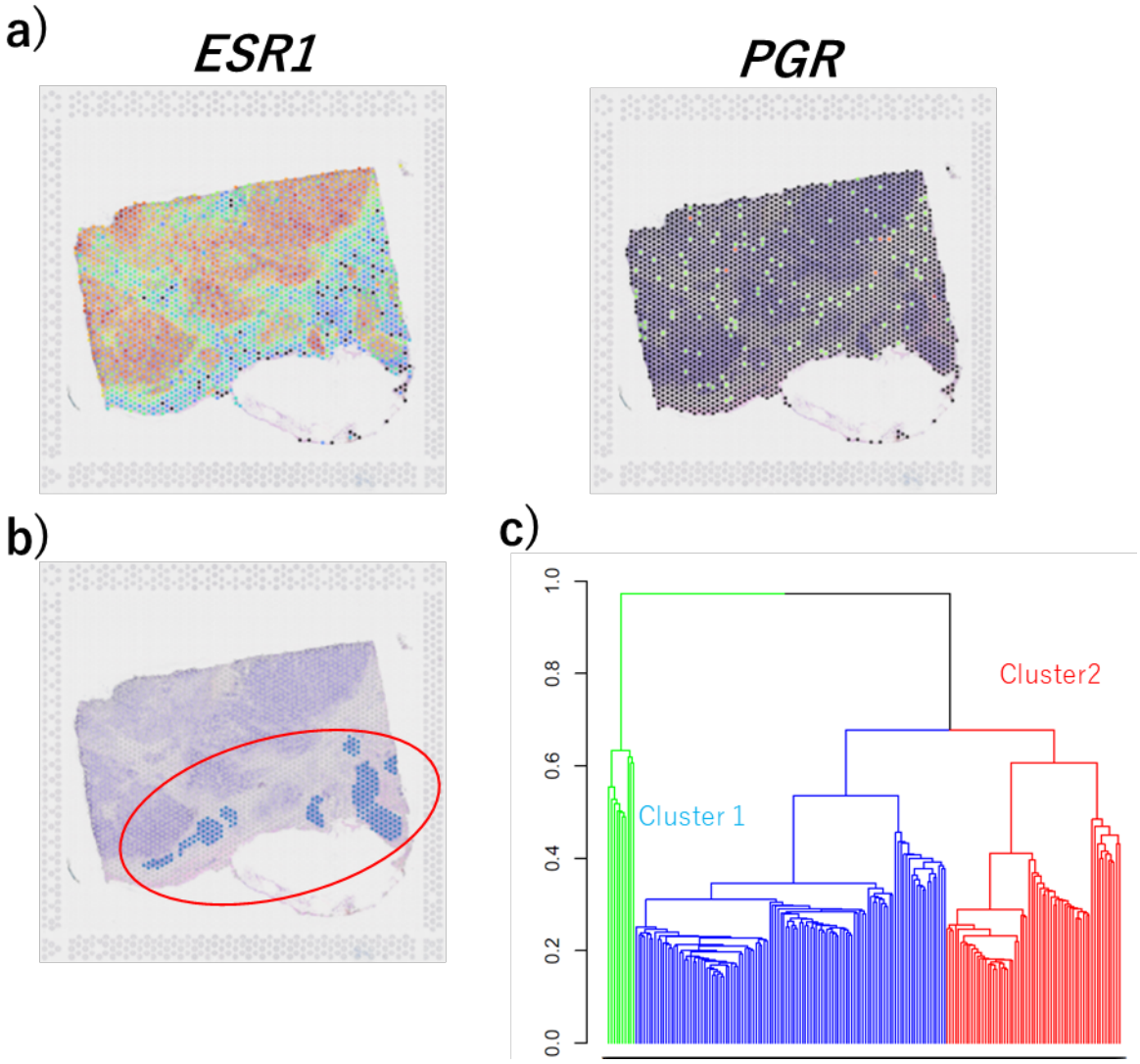


Fig. 15 Hematoxylin and eosin (HE)-stained images and heterogeneity of DCIS in Case B.

a). HE-stained images showing gene expression of *ESR1* (left) and *PGR* (right). Consistent with the IHC results, *ESR1* was positive, and *PGR* was negative. b). HE-stained images showing selected pathological DCIS spots (blue). c) Dendrogram of the gene expression profile of 188 DCIS spots clustered by hierarchical clustering. Blue: Cluster 1, 114 spots; red: Cluster 2, 64 spots; green: Cluster 3, 10 spots.

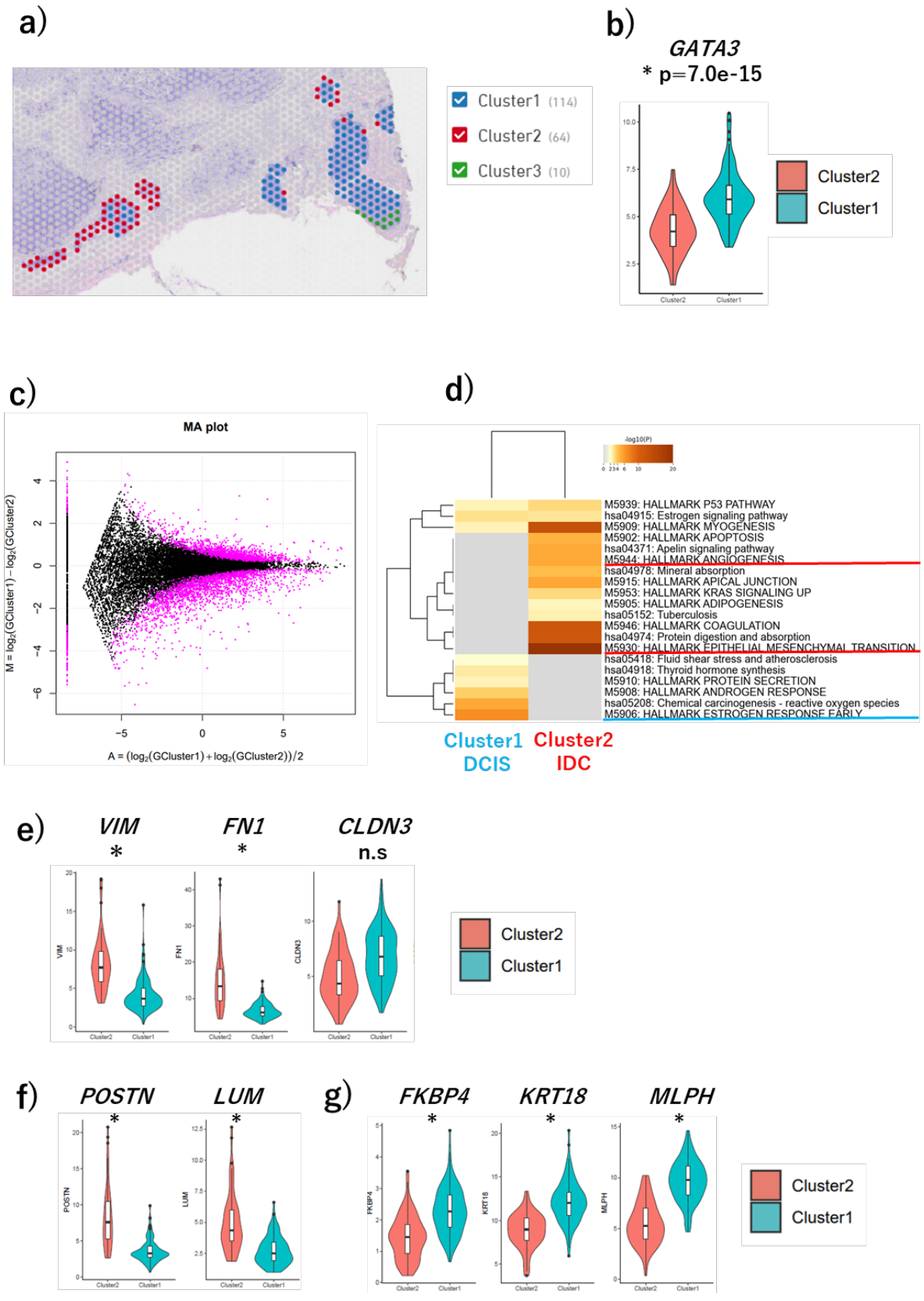


Fig. 16 Spatial transcriptome analysis of DCIS subclusters of Case B.

a). Visualization of DCIS subclusters on hematoxylin and eosin (HE)-stained images. Blue:

Cluster 1, located on DCIS spots; red: Cluster 2, located on the invasion spots; green: Cluster 3, located on DCIS. b). Violin plot of *GATA3* expression at each cluster. Decreasing *GATA3* expression was observed in the spots at Cluster 2 (P = 7.0e-15; one-sided t-test). c). The MA plot of gene expression fold change between Clusters 1 and 2. The pink dot is FDR < 0.01. d). Result of enrichment analysis by Metascape. e), f), and g) Violin plots of key genes in each pathway. *P < 0.05 (one-sided), n.s.: not significant. e). Violin plot of key genes (*VIM*, *FN1*, and *CLDN3*) in the EMT pathway. All genes showed increased expression in Cluster 2. f). Violin plot of key genes (*POSTN* and *LUM*) in the angiogenesis pathway. All genes showed increased expression in Cluster 2. g). Violin plots of key genes (*FKBP4*, *KRT18*, and *MLPH*) in the estrogen response pathway.

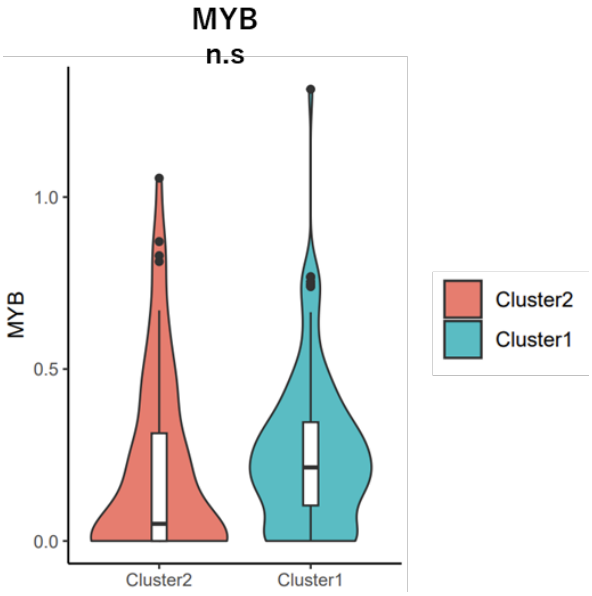


Fig. 17 Violin plots of *MYB* expression.

Expression difference in *MYB* between Clusters 1 and 2. Although there was no significance, *MYB* tended to show decreased expression in Cluster 2.

4.6 Case C results

Pathology revealed no invasion in this patient, who had two *PIK3CA* mutations but no *GATA3* mutation (Fig. 12c).

Spatial transcriptome analysis using nonhierarchical clustering revealed that this specimen consisted of two expression clusters. The spot locations of these two clusters almost overlapped with pathologically determined cancer and stromal spots (Fig. 18a). The tumor cluster's expression pattern resembled that of monoclonals. To reveal the pathway activity in cancer spots, I performed differential expression analysis using the R TCC packages and obtained 347 upregulated genes in cancer spots. To cluster genes correlating with each other, I further conducted gene expression correlation network analysis using WGCNA and obtained two gene modules. The results of the pathway enrichment analysis used in these two modules revealed that 12 and 3 pathways were enriched in modules 1 and 2, respectively. However, pathways related to invasion, such as those detected in Cases A and B, were not identified. These results suggest that this cancer had not acquired invasion ability and that the specimen was truly low-risk DCIS.

Additionally, to confirm that other miner clones did not exist, I conducted single-cell copy number analysis using single-cell DNA sequencing data. Here, to compare clonality, I analyzed the other two DCIS specimens (Cases D and E). Because specimens for spatial transcriptome analysis were not obtained from Cases D and E, only single-cell DNA sequencing was performed. Table 10 shows the sequencing statistics of single-cell DNA sequencing.

Cases D and E showed polyclonal copy number profiles. In Case D, three clones were detected, i.e., 10q−, 1q+,1q+10q−. This patient had not relapsed (Fig. 20). In Case E, three clones were detected, i.e., 1q+ and 16q− clones, chrX- clone, and HER2 amplification clone (Fig. 21). As HER2 amplification is associated with an aggressive phenotype (46), this patient

showed polyclonal cancer contents and predicted a poor prognosis; therefore, relapse was confirmed.

In Case C, the obtained sequencing data was 696 million, the number of detected cells was 722 cells, and the median effective read per MB was 208 reads. After obtaining sufficient data, I estimated genomic copy number variations, which revealed that 64 of 722 cells were normal cells without copy number aberrations, whereas 548 were cancer cells with aberrations of **1q+ and 16q−** (Fig. 19). Other cells were low-quality, noisy cells, so these were filtered out. Additionally, previous studies identified aberrations of 1q+ and 16q− and detected them in low-risk DCIS (47-49). These results differed from those of the malignant lesions (Fig. 21). Overall, Case C was truly low-risk DCIS with no detected malignant cells.

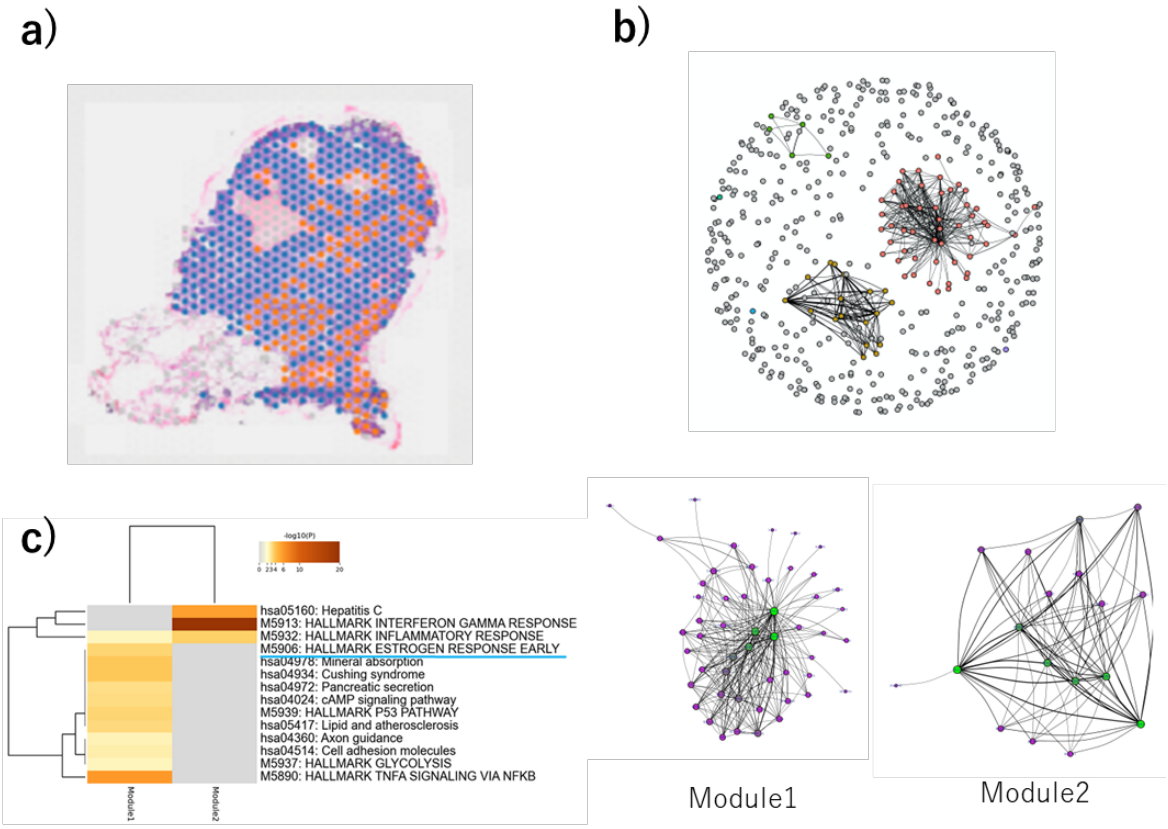


Fig. 18 Spatial transcriptome analysis of Case C.

a). Clustered spots visualized on hematoxylin and eosin-stained images. Blue indicates cancer cell spots, and orange indicates stromal cell spots. b). Gene co-expression network. The top panels represent gene network graphs for all 508 DEGs. Of 508 DEGs, 434 had no edges. The bottom left represents module 1, which consists of 51 genes. The bottom right represents module 2, which consists of 18 genes. c). Result of enrichment analysis by Metascape.

Table 10. Sequencing statistics of single-cell DNA sequencing.

Sample	Clinical Risk	Relapse	*GATA3* mutation	*PIK3CA* mutation	Total reads	Reads from cells	Total mapped reads in cells	Estimated number of cells	Median ploidy	Median effective reads per MB	Mean mapped reads (rmdup)
Case C	Low Risk	-	-	✓	696,077,270	583,228,454	477,374,763	722	2.01	208	661,183
Case D	Low Risk	-	-	-	712,651,128	592,273,894	444,443,131	196	1.99	693	2,267,566
Case E	High Risk	✓	-	-	590,064,788	374,924,333	484,253,052	216	2	527	1,735,760

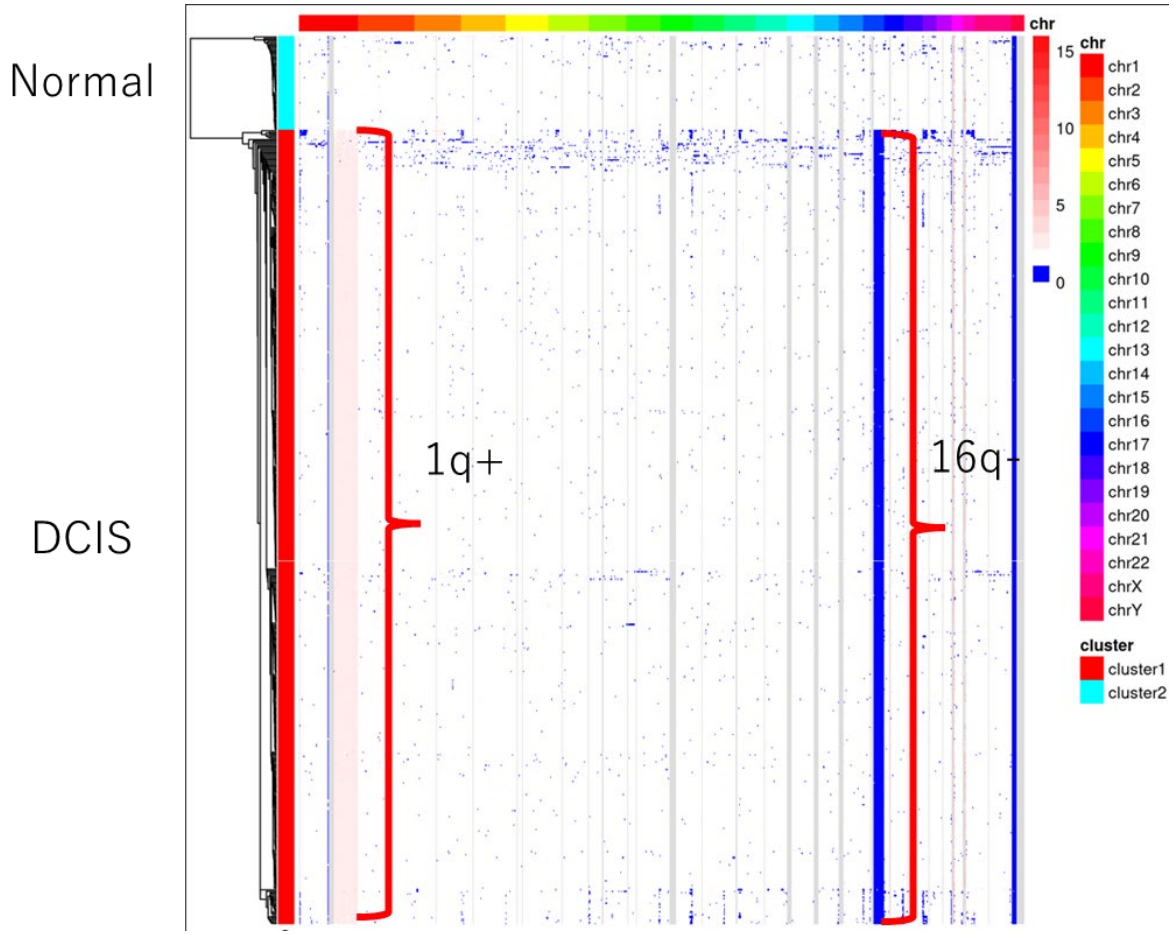
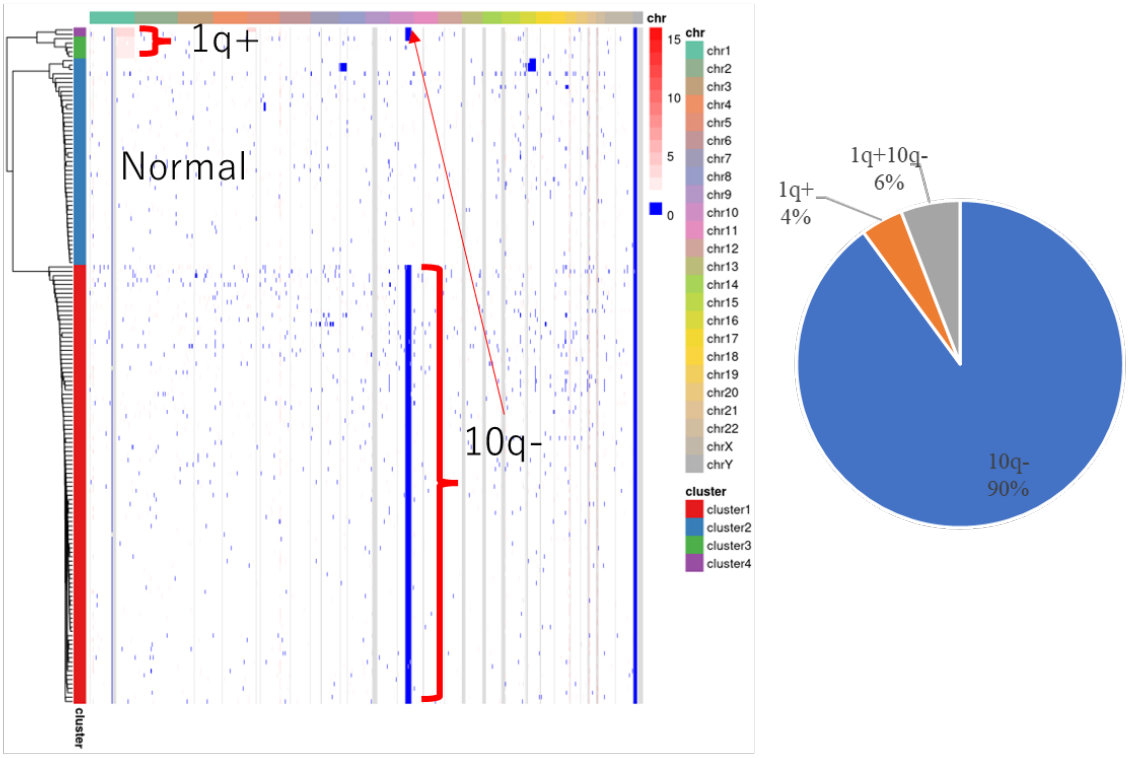


Fig. 19 Single-cell copy number profile of Case C

Heatmap of the genomic copy number profile of Case C. Blue indicates the loss of genomes, and red indicates gains on the heatmap. Two clusters were observed using hierarchical clustering. Rows indicate each cell, and columns indicate each genomic locus. The red row annotation represents DCIS cells with 1q+ and 16q− genomic aberrations. The blue row annotation represents normal cells. The clonality of DCIS cells is monoclonal.

a)

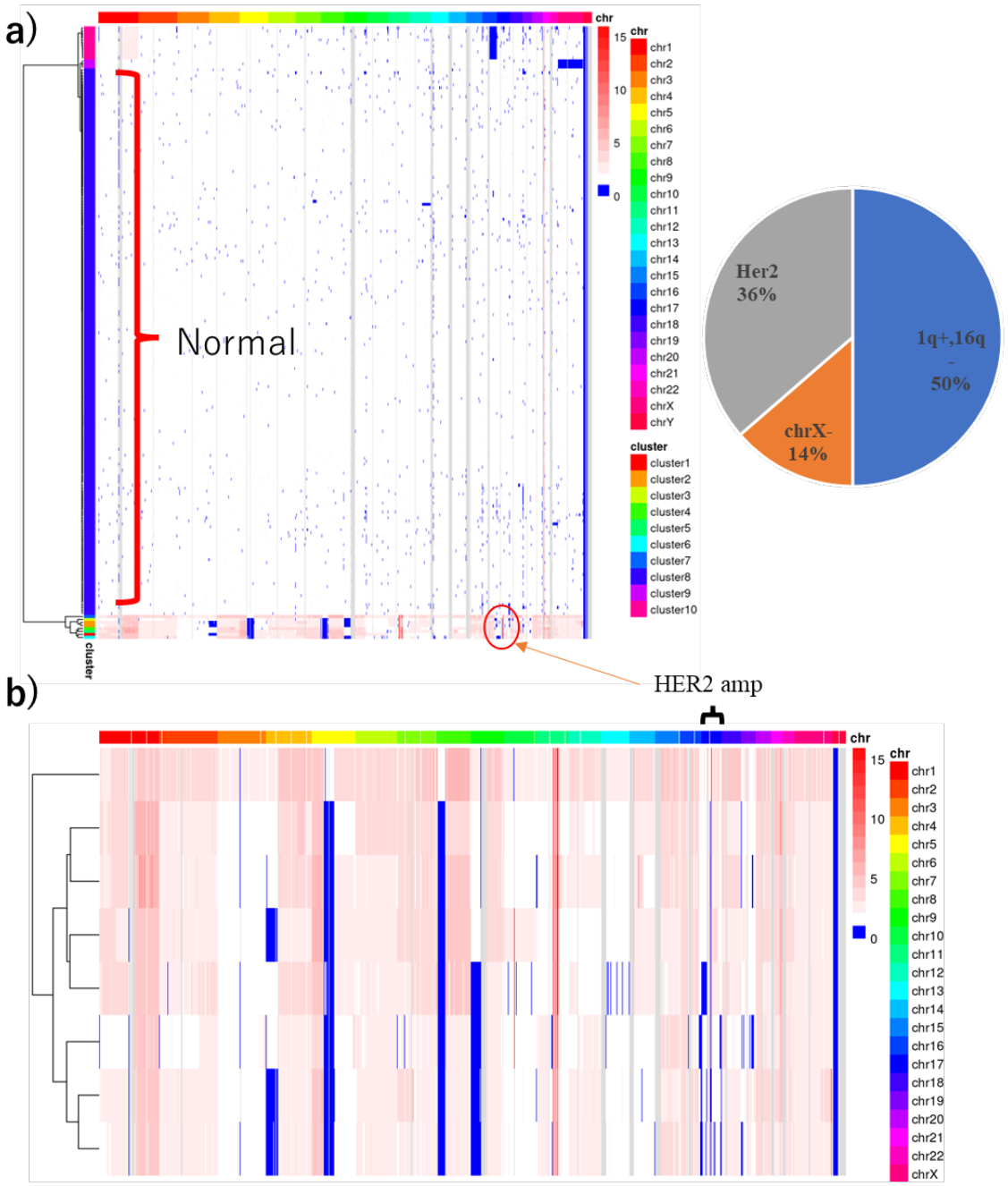


b)

Normal	10q-	1q+	1q+10q-
47	107	5	2

Fig. 20 Single-cell copy number profile of Case D.

Heatmap of the genomic copy number profile of Case D. Blue indicates loss of genome, and red indicates gain on the heatmap. Two clusters were observed using hierarchical clustering. Rows indicate each cell, and columns indicate each genomic locus. The red row annotation represents DCIS cells with 1q+ and 16q− genomic aberrations. The blue rows annotation represents normal cells. The clonality of DCIS cells is monoclonal.



Normal	1q+,16q-	chrX-	Her2
183	11	3	6

Fig. 21 Single-cell copy number profile of Case E.

Heatmap of the genomic copy number profile of Case E. Blue indicates loss of genome, and red indicates gain on the heatmap. Three clusters were observed by hierarchical clustering, with

183 normal cells (Cluster 8, blue), 11 cells with 1q+ and 16q− (Cluster 10, pink), 3 cells with ChrX- (Cluster 9, purple), and 8 cells with Her2 amplification (Cluster 1–7). Picart shows the composition of the abnormal cells. This specimen was found to be polyclonal. b) The heatmap shows only HER2 amplification cells, with these cells confirming many copy number aberrations. c) The table represents the number of cells in each cluster.

5 Discussion

5.1 PIK3CA mutation tends to have a negative association with relapse

PIK3CA mutations tend to exhibit a negative association with the relapse of invasive cancer. *PIK3CA* is one of the most frequently mutated genes in various cancers, including breast cancer (50, 51). The oncogenic mechanism of *PIK3CA* has been reported to be associated with many cancer hallmark pathways (52). However, recently, *PIK3CA* mutations were reported in the kinase domain and the absence of copy number gains in DCIS was found to be protective against progression (20).

The malignancy of mammalian tumor cells harboring the $PIK3CA^{H1047}$ mutation was reported to be influenced by its origin cell (53-55). When the $PIK3CA^{H1047}$ mutation was introduced to basal cells, tumor progression was slowed down and less aggressive than in luminal cells (53, 55). Additionally, tumor cells harboring the $PIK3CA^{H1047R}$ mutation were reported to have acquired multipotency in lineage-restricted basal and luminal unipotent progenitors. Therefore, the tumor subtypes detected by existing marker gene expression (such as luminal or basal) may differ from those of the origin cell. In my datasets, *PIK3CA* was observed to be a low-risk factor for DCIS due to basal cells being the origin of tumor cells harboring the *PIK3CA* mutation.

5.2 Association between invasion and aberration of GATA3 expression.

The pathway related to the invasion was activated in Case A, which harbored *GATA3* mutations. However, the same pathway was also activated in Case B, indicating the absence of *GATA3* mutation. In Case B, a decrease in *GATA3* expression was identified; hence, an alteration of epigenetic conditions may have caused the aberration of a *GATA3* expression. Furthermore, *PGR* was not expressed in Case B (Fig. 15b); thus, it was assumed that in this case, acquiring invasion capacity was not dependent on *PGR*. Previous studies have shown

that *GATA3* expression is correlated with E-cadherin expression. Additionally, it was reported that when the expression of E-cadherin decreased, the expression of N-cadherin and vimentin associated with invasion also decreased (56). In Case B, E-cadherin may have contributed to the acquired invasion capacity.

Besides, microglandular adenosis, a rare neoplasm of the mammary gland, was reported to have aberrant hypermethylation of CpG sites within *GATA3* as well as the loss of *GATA3* protein expression (57). Epigenetic factors may be important for DCIS progression. Therefore, in future analysis, it is essential to compare the methylation sites on *GATA3* between invasive and DCIS cells.

5.3 Applications in clinical practice

In this study, *GATA3* was identified as a high-risk factor for DCIS malignancy. In clinical practice, once a patient shows a suspicious abnormality on mammography screening, their lesions will be further evaluated using needle biopsy (58). Genomic testing of *GATA3* is not always possible during DCIS diagnosis. However, some of these lesions extracted during a biopsy are available for the genomic testing of *GATA3*.

In the recent standard treatment of DCIS, breast-conserving surgery may be adopted as the treatment of choice in some low-risk DCIS patients (59, 60). Adjuvant therapy is also recommended. Because surgical resection has been performed on these patients, lesions are available for genomic testing. Combining *GATA3* mutation information with pathological diagnostic criteria may lead to more optimal treatment strategies. Therefore, the information on the *GATA3* mutation can be helpful for clinical application.

5.4 Limitation of Visium resolution

In this study, cancer cell progression was shown to directly acquire invasive capacity using

spatial transcriptome analysis. The Visium platform (10x genomics) is an excellent technology used to obtain spatial transcriptome data, but the resolution is insufficient. Here, the spot size was 55 μm, and the number of cells captured in a single spot was generally 1–10 cells. Additionally, the distance between the centers of the spots was 100 μm. Thus, several cells could not be analyzed. Xenium platforms have been released to resolve the insufficient resolution of Visium, which is an *in situ* hybridization technology, and surprisingly, resolutions are at the level of a single cell (61). However, because Xenium postgame can measure only 300–400 genes currently, the integrated analysis of multiplatform data is important for future studies.

6 Conclusion

In this study, the *GATA3* mutation was identified as a potential marker for DCIS and for classifying DCIS. Additionally, spatial transcriptome analysis directly revealed that *GATA3* abnormalities promoted malignant transformation. The novelty of this study includes using spatial transcriptome analysis to reveal that *GATA3* mutations cause DCIS malignancies. Furthermore, even in the absence of *GATA3* mutations, abnormal expression suggests the possibility of malignant transformation. This spatial transcriptome analysis method can be applied to other cancer types to reveal their diversity.

7 Data availability

All sequencing data and pathological images from spatial transcriptome analysis have been deposited in the DNA Data Bank of Japan under the accession number JGAS00000000202.

8 Acknowledgments

I would like to express my gratitude to Professor Suzuki Yutaka for all his constructive comments, warm encouragement, and teaching principles as a Scientific Researcher. I would like to express my gratitude to Dr. Nagasawa for his provision of analytical specimens, clinical insights, pathological information, and many other thought-provoking suggestions. I would like to express my gratitude to Professor Suzuki Ayako and Professor Masahide Seki for their insightful comments and suggestions. I would like to express my thanks and gratitude to everyone at Suzuki Laboratory for their support and encouragement. The super-computing resource was provided by the Human Genome Center, the Institute of Medical Science, and the University of Tokyo (http://sc.hgc.jp/shirokane.html).

9 References

1. Polyak K. Is breast tumor progression really linear? Clin Cancer Res. 2008;14(2):339-41. doi: 10.1158/1078-0432.ccr-07-2188. PubMed PMID: 19064970.

2. Gorringe KL, Fox SB, Page DL, Dupont WD, Rogers LW, Landenberger M. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis

Intraductal carcinoma of the breast: follow-up after biopsy only. Front Oncol. 2017;7(4):248. Epub 20171023. doi: 10.3389/fonc.2017.00248

10.1002/1097-0142(19820215)49:4<751::aid-cncr2820490426>3.0.co;2-y. PubMed PMID: 6275978; PubMed Central PMCID: PMC5660056.

3. Kerlikowske K. Epidemiology of ductal carcinoma in situ. J Natl Cancer Inst Monogr. 2010;2010(41):139-41. doi: 10.1093/jncimonographs/lgq027. PubMed PMID: 29109942; PubMed Central PMCID: PMC5161058.

4. Kuerer HM, Albarracin CT, Yang WT, Cardiff RD, Brewster AM, Symmans WF, et al. Ductal carcinoma in situ: state of the science and roadmap to advance the field. J Clin Oncol. 2009;27(2):279-88. Epub 20081208. doi: 10.1200/jco.2008.18.3103. PubMed PMID: 20956818.

5. Welch HG, Black WC. Using autopsy series to estimate the disease "reservoir" for ductal carcinoma in situ of the breast: how much more breast cancer can we find? Ann Intern Med. 1997;127(11):1023-8. doi: 10.7326/0003-4819-127-11-199712010-00014. PubMed PMID: 9412284.

6. Page DL, Dupont WD, Rogers LW, Jensen RA, Schuyler PA. Continued local recurrence of carcinoma 15-25 years after a diagnosis of low grade ductal carcinoma in situ of the breast treated only by biopsy. Cancer. 1995;76(7):1197-200. doi: 10.1002/1097-0142(19951001)76:7<1197::aid-cncr2820760715>3.0.co;2-0. PubMed PMID: 8630897.

7. Page DL, Dupont WD, Rogers LW, Landenberger M. Intraductal carcinoma of the breast: follow-up after biopsy only. Cancer. 1982;49(4):751-8. doi:

10.1002/1097-0142(19820215)49:4<751::aid-cncr2820490426>3.0.co;2-y. PubMed PMID: 6275978.

8. Collins LC, Tamimi RM, Baer HJ, Connolly JL, Colditz GA, Schnitt SJ. Outcome of patients with ductal carcinoma in situ untreated after diagnostic biopsy: results from the Nurses' Health Study. Cancer. 2005;103(9):1778-84. doi: 10.1002/cncr.20979. PubMed PMID: 15770688.

9. Erbas B, Provenzano E, Armes J, Gertig D. The natural history of ductal carcinoma in situ of the breast: a review. Breast Cancer Res Treat. 2006;97(2):135-44. Epub 20051201. doi: 10.1007/s10549-005-9101-z. PubMed PMID: 18223205.

10. Sagara Y, Mallory MA, Wong S, Aydogan F, DeSantis S, Barry WT, et al. Survival Benefit of Breast Surgery for Low-Grade Ductal Carcinoma In Situ: A Population-Based Cohort Study. JAMA Surg. 2015;150(8):739-45. doi: 10.1001/jamasurg.2015.0876. PubMed PMID: 26266975.

11. Morrow M, Katz SJ. Addressing Overtreatment in DCIS: What Should Physicians Do Now? J Natl Cancer Inst. 2015;107(12):djv290. Epub 20150930. doi: 10.1093/jnci/djv290. PubMed PMID: 26424780.

12. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection Do LORIS Trial Eligibility Criteria Identify a Ductal Carcinoma In Situ Patient Population at Low Risk of Upgrade to Invasive Carcinoma? Cell Syst. 2015;1(6):417-25. Epub 20160512. doi: 10.1016/j.cels.2015.12.004. PubMed PMID: 27172775; PubMed Central PMCID: PMC4707969

PMC5070657.

13. Elshof LE, Tryfonidis K, Slaets L, van Leeuwen-Stok AE, Skinner VP, Dif N, et al. Feasibility of a prospective, randomised, open-label, international multicentre, phase III, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ - The LORD study. Eur J Cancer. 2015;51(12):1497-510. Epub 20150526. doi: 10.1016/j.ejca.2015.05.008. PubMed PMID: 26039049.

14. Youngwirth LM, Boughey JC, Hwang ES. Surgery versus monitoring and endocrine therapy

for low-risk DCIS: The COMET Trial. Bull Am Coll Surg. 2017;102(1):62-3. PubMed PMID: 29470143.

15. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-10. doi: 10.1111/j.1365-2559.1991.tb00229.x. PubMed PMID: 18242514.

16. Cutler SJ, Black MM, Friedell GH, Vidone RA, Goldenberg IS. Prognostic factors in cancer of the female breast. II. Reproducibility of histopathologic classification. Cancer. 1966;19(1):75-82. doi: 10.1002/1097-0142(196601)19:1<75::aid-cncr2820190108>3.0.co;2-4.

17. Silverstein MJ. The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast. Am J Surg. 2003;186(4):337-43. doi: 10.1016/s0002-9610(03)00265-4. PubMed PMID: 15016963.

18. Le Doussal V, Tubiana-Hulin M, Friedman S, Hacene K, Spyratos F, Brunet M. Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. Cancer. 1989;64(9):1914-21. doi: 10.1002/1097-0142(19891101)64:9<1914::aid-cncr2820640926>3.0.co;2-g. PubMed PMID: 26025767.

19. The Consensus Conference Committee anb. Consensus Conference on the Classification of Ductal Carcinoma In Situ. The Breast Journal. 1997;3(6):360-4. doi: https://doi.org/10.1111/j.1524-4741.1997.tb00194.x.

20. Lin CY, Vennam S, Purington N, Lin E, Varma S, Han S, et al. Genomic landscape of ductal carcinoma in situ and association with progression. Breast Cancer Res Treat. 2019;178(2):307-16. Epub 20190817. doi: 10.1007/s10549-019-05401-x. PubMed PMID: 32587399; PubMed Central PMCID: PMC6800639.

21. Pang JB, Savas P, Fellowes AP, Mir Arnau G, Kader T, Vedururu R, et al. Breast ductal

carcinoma in situ carry mutational driver events representative of invasive breast cancer. Mod Pathol. 2017;30(7):952-63. Epub 20170324. doi: 10.1038/modpathol.2017.21. PubMed PMID: 28581515.

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 20090518. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.

23. Institute B. Picard Tools: Broad Institute, GitHub repository; Accessed: 2018/02/21; version 2.17.8 [cited 2018]. Available from: http://broadinstitute.github.io/picard/.

24. van der Auwera G, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra: O'Reilly Media, Incorporated; 2020.

25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. Epub 20100703. doi: 10.1093/nar/gkq603. PubMed PMID: 20601685; PubMed Central PMCID: PMC2938201.

26. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 29. United States2011. p. 24-6.

27. Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: an R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics. 2013;14:219. Epub 20130709. doi: 10.1186/1471-2105-14-219. PubMed PMID: 24567146; PubMed Central PMCID: PMC3716788.

28. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523. Epub 20190403. doi: 10.1038/s41467-019-09234-6. PubMed PMID: 31065110; PubMed Central PMCID: PMC6447622.

29. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. Epub 20081229. doi: 10.1186/1471-2105-9-559. PubMed PMID: 19411068; PubMed Central PMCID: PMC2631488.

30. Farante G, Toesca A, Magnoni F, Lissidini G, Vila J, Mastropasqua M, et al. Advances and

controversies in management of breast ductal carcinoma in situ (DCIS). Eur J Surg Oncol. 2022;48(4):736-41. Epub 20211109. doi: 10.1016/j.ejso.2021.10.030. PubMed PMID: 34772587.

31. Takaku M, Grimm SA, Roberts JD, Chrysovergis K, Bennett BD, Myers P, et al. *GATA3* zinc finger 2 mutations reprogram the breast cancer transcriptional network. Nat Commun. 2018;9(1):1059. Epub 20180313. doi: 10.1038/s41467-018-03478-4. PubMed PMID: 30923225; PubMed Central PMCID: PMC5849768.

32. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882-d9. doi: 10.1093/nar/gkz1062. PubMed PMID: 31713622; PubMed Central PMCID: PMC7061942.

33. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PubMed Central PMCID: PMC3439153.

34. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, et al. Mutation of *GATA3* in human breast tumors. Oncogene. 2004;23(46):7669-78. doi: 10.1038/sj.onc.1207966. PubMed PMID: 15361840.

35. Vantangoli MM, Madnick SJ, Huse SM, Weston P, Boekelheide K. MCF-7 Human Breast Cancer Cells Form Differentiated Microtissues in Scaffold-Free Hydrogels. PLoS One. 2015;10(8):e0135426. Epub 20150812. doi: 10.1371/journal.pone.0135426. PubMed PMID: 26267486; PubMed Central PMCID: PMC4534042.

36. Wallis MG, Clements K, Kearins O, Ball G, Macartney J, Lawrence GM. The effect of DCIS grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected DCIS. Br J Cancer. 2012;106(10):1611-7. Epub 20120419. doi: 10.1038/bjc.2012.151. PubMed PMID: 22516949; PubMed Central PMCID: PMC3349181.

37. Kouros-Mehr H, Slorach EM, Sternlicht MD, Werb Z. *GATA3* maintains the differentiation of the luminal cell fate in the mammary gland. Cell. 2006;127(5):1041-55. doi: 10.1016/j.cell.2006.09.048.

PubMed PMID: 17129787; PubMed Central PMCID: PMC2646406.

38. Pei XH, Bai F, Smith MD, Usary J, Fan C, Pai SY, et al. CDK inhibitor p18(INK4c) is a downstream target of *GATA3* and restrains mammary luminal progenitor cell proliferation and tumorigenesis. Cancer Cell. 2009;15(5):389-401. doi: 10.1016/j.ccr.2009.03.004. PubMed PMID: 19451168; PubMed Central PMCID: PMC2699569.

39. Bai F, Zheng C, Liu X, Chan HL, Liu S, Ma J, et al. Loss of function of *GATA3* induces basal-like mammary tumors. Theranostics. 2022;12(2):720-33. Epub 20220101. doi: 10.7150/thno.65796. PubMed PMID: 34976209; PubMed Central PMCID: PMC8692904.

40. Shahi P, Wang CY, Chou J, Hagerling C, Gonzalez Velozo H, Ruderisch A, et al. *GATA3* targets semaphorin 3B in mammary epithelial cells to suppress breast cancer progression and metastasis. Oncogene. 2017;36(40):5567-75. Epub 20170605. doi: 10.1038/onc.2017.165. PubMed PMID: 28925613; PubMed Central PMCID: PMC5629104.

41. Dydensborg AB, Rose AA, Wilson BJ, Grote D, Paquet M, Giguère V, et al. *GATA3* inhibits breast cancer growth and pulmonary breast cancer metastasis. Oncogene. 2009;28(29):2634-42. Epub 20090601. doi: 10.1038/onc.2009.126. PubMed PMID: 19798694.

42. Karakas B, Bachman KE, Park BH. Mutation of the *PIK3CA* oncogene in human cancers. Br J Cancer. 2006;94(4):455-9. doi: 10.1038/sj.bjc.6602970. PubMed PMID: 16449998; PubMed Central PMCID: PMC2361173.

43. Liska O, Bohár B, Hidas A, Korcsmáros T, Papp B, Fazekas D, et al. TFLink: an integrated gateway to access transcription factor–target gene interactions for multiple species. Database. 2022;2022. doi: 10.1093/database/baac083.

44. Maurice D, Hooper J, Lang G, Weston K. c-Myb regulates lineage choice in developing thymocytes via its target gene *GATA3*. Embo j. 2007;26(15):3629-40. Epub 20070719. doi: 10.1038/sj.emboj.7601801. PubMed PMID: 17641686; PubMed Central PMCID: PMC1949015.

45. Knopfová L, Biglieri E, Volodko N, Masařík M, Hermanová M, Glaus Garzón JF, et al.

Transcription factor c-Myb inhibits breast cancer lung metastasis by suppression of tumor cell seeding. Oncogene. 2018;37(8):1020-30. Epub 20171030. doi: 10.1038/onc.2017.392. PubMed PMID: 29084208; PubMed Central PMCID: PMC6711763.

46. Gajria D, Chandarlapaty S. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. Expert Rev Anticancer Ther. 2011;11(2):263-75. doi: 10.1586/era.10.226. PubMed PMID: 21342044; PubMed Central PMCID: PMC3092522.

47. Pang JM, Gorringe KL, Wong SQ, Dobrovic A, Campbell IG, Fox SB. Appraisal of the technologies and review of the genomic landscape of ductal carcinoma in situ of the breast. Breast Cancer Res. 2015;17(1):80. Epub 20150616. doi: 10.1186/s13058-015-0586-z. PubMed PMID: 26078038; PubMed Central PMCID: PMC4469314.

48. Tsuda H. Gene and chromosomal alterations in sporadic breast cancer: correlation with histopathological features and implications for genesis and progression. Breast Cancer. 2009;16(3):186-201. Epub 20090527. doi: 10.1007/s12282-009-0124-x. PubMed PMID: 19483726.

49. Buerger H, Otterbach F, Simon R, Poremba C, Diallo R, Decker T, et al. Comparative genomic hybridization of ductal carcinoma in situ of the breast-evidence of multiple genetic pathways. J Pathol. 1999;187(4):396-402. doi: 10.1002/(sici)1096-9896(199903)187:4<396::aid-path286>3.0.co;2-l. PubMed PMID: 10398097.

50. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High frequency of mutations of the *PIK3CA* gene in human cancers. Science. 2004;304(5670):554. Epub 20040311. doi: 10.1126/science.1096502. PubMed PMID: 16357129.

51. Madsen RR, Vanhaesebroeck B, Semple RK. Cancer-Associated *PIK3CA* Mutations in Overgrowth Disorders. Trends Mol Med. 2018;24(10):856-70. Epub 20180906. doi: 10.1016/j.molmed.2018.08.003. PubMed PMID: 30197175; PubMed Central PMCID: PMC6185869.

52. Arafeh R, Samuels Y. *PIK3CA* in cancer: The past 30 years. Semin Cancer Biol. 2019;59:36-49. Epub 20190210. doi: 10.1016/j.semcancer.2019.02.002. PubMed PMID: 30742905.

53. Koren S, Reavie L, Couto JP, De Silva D, Stadler MB, Roloff T, et al. *PIK3CA*(H1047R) induces multipotency and multi-lineage mammary tumours. Nature. 2015;525(7567):114-8. Epub 20150812. doi: 10.1038/nature14669. PubMed PMID: 26266985.

54. Van Keymeulen A, Lee MY, Ousset M, Brohée S, Rorive S, Giraddi RR, et al. Reactivation of multipotency by oncogenic *PIK3CA* induces breast tumour heterogeneity. Nature. 2015;525(7567):119-23. Epub 20150812. doi: 10.1038/nature14665. PubMed PMID: 26266985.

55. Gough NR. Inducing tumor heterogeneity with PI3K. Science Signaling. 2015;8(393):ec253-ec. doi: doi:10.1126/scisignal.aad3711.

56. Yan W, Cao QJ, Arenas RB, Bentley B, Shao R. *GATA3* inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. J Biol Chem. 2010;285(18):14042-51. Epub 20100226. doi: 10.1074/jbc.M110.105262. PubMed PMID: 20189993; PubMed Central PMCID: PMC2859565.

57. Radner M, van Luttikhuizen JL, Bartels S, Bublitz J, Grote I, Rieger L, et al. Chromosome 2q gain and epigenetic silencing of *GATA3* in microglandular adenosis of the breast. J Pathol Clin Res. 2021;7(3):220-32. Epub 20201231. doi: 10.1002/cjp2.195. PubMed PMID: 33382535; PubMed Central PMCID: PMC8073017.

58. Matsen CB, Neumayer LA. Breast cancer: a review for the general surgeon. JAMA Surg. 2013;148(10):971-9. doi: 10.1001/jamasurg.2013.3393. PubMed PMID: 23986370.

59. Barrio AV, Van Zee KJ. Controversies in the Treatment of Ductal Carcinoma in Situ. Annu Rev Med. 2017;68:197-211. doi: 10.1146/annurev-med-050715-104920. PubMed PMID: 28099081; PubMed Central PMCID: PMC5532880.

60. Fu F, Gilmore RC, Jacobs LK. Ductal Carcinoma In Situ. Surg Clin North Am. 2018;98(4):725-45. Epub 20180528. doi: 10.1016/j.suc.2018.03.007. PubMed PMID: 30005770.

61. Janesick A, Shelansky R, Gottscho A, Wagner F, Rouault M, Beliakoff G, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ

analysis of FFPE tissue. bioRxiv. 2022:2022.10.06.510405. doi: 10.1101/2022.10.06.510405.