Supplementary Methods

This document describes in detail the materials and methods in **Trinh *et al*, "Genomic alterations during the in situ to invasive ductal breast carcinoma transition shaped by the immune system".** Further details on WES pipelines can be found at (https://docs.google.com/document/d/1VO2kX_fgfUd0x3mBS9NjLUWGZu794WbTepBel3cBg08/edit#heading=h.yby87l2ztbcj).

Code for downstream analysis and figure generation can be obtained at https://github.com/polyak-lab/matchedDCISIDC

This document describes 6 specific sections:

1. Whole exome sequencing
2. RNA sequencing
3. Cyclic Immunofluorescence
4. Whole exome sequencing characterization pipeline
5. External data sets used
6. Data Analysis

1. Whole exome sequencing

Patient samples were macro-dissected into stromal and epithelial fractions from FFPE (formalin fixed paraffin embedded) tissue-slides, with examples of segmented regions highlighted in **Fig. S1B**. DNA and RNA were extracted from epithelial fractions from matched DCIS and IDC cases and corresponding normal breast using AllPrep DNA/RNA FFPE Kit (Qiagen, Cat. No. 80234). Whole-exome sequencing was performed by the Broad Institute Genomics Services, as described below: 10-100ng of DNA in 50 µL solution was used for library construction was performed as previously described (1). Adapter ligation was performed using palindromic forked adapters with unique dual-indexed molecular barcode sequences to facilitate downstream pooling (Integrated DNA Technologies). With the exception of the palindromic forked adapters, the reagents used for end repair, A-base addition, adapter ligation, and

library enrichment PCR were purchased from KAPA Biosciences in 96-reaction kits. In addition, during the post-enrichment SPRI cleanup, elution volume was reduced to 30 µL to maximize library concentration, and a vortexing step was added to maximize the amount of template eluted. Mean insert size and yield was estimated using KAPA genomic QC assay prior to library capture, which has been shown to be predictors of FFPE sequencing. Hybridization and ICE capture were performed using Illumina's Nextera Rapid Capture Exome Kit and following the manufacturer's suggested protocol, with the following exceptions: first, all libraries within a library construction plate were pooled prior to hybridization. Second, the Midi plate from Illumina's Nextera Rapid Capture Exome Kit was replaced with a skirted PCR plate to facilitate automation. All hybridization and capture steps were automated on the Agilent Bravo liquid handling system.

Library pools were then quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of the adapters. Based on qPCR quantification, libraries were normalized to 2 nM, then denatured using 0.1 N NaOH on the Hamilton Starlet. After denaturation, libraries were diluted to 20 pM using hybridization buffer purchased from Illumina. Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using HiSeq 4000 cluster chemistry and HiSeq 4000 flowcells. Flowcells were sequenced on v1 Sequencing-by-Synthesis chemistry for HiSeq 4000 flowcells. The flowcells are then analyzed using RTA v.1.18.64 or later. Each pool of whole exome libraries was run on paired 76bp runs, reading the dual-indexed sequences to identify molecular indices and sequenced across the number of lanes needed to meet coverage for all libraries in the pool. Sequencing data processing was performed using established best practices pipelines at the Broad Institute: Reads were mapped to the hg19 reference using BWA mem (2,3), duplicate read-pairs were marked using Picard (http://broadinstitute.github.io/picard/), reads were sorted using SortSam, followed by Base Quality Score Recalibration using GATK BSQR (4). Coverage metrics were obtained using GATK DepthOfCoverage (4) and shown in **Supplemental Figure S2A-B**, showing that all samples had at

least 60x average coverage and all samples except for 1DCIS and 2DCIS have 80% of target bases covered at 20x coverage or higher.

2. RNA-sequencing

RNA was extracted from epithelial fractions from matched DCIS and IDC cases and corresponding normal breast using AllPrep DNA/RNA FFPE Kit (Qiagen, Cat. No. 80234). Transcriptome capture and sequencing was performed by the Broad Institute Genomics Services using methods optimized for FFPE. Total RNA is quantified using the Quant-iT RiboGreen RNA Assay Kit and normalized to 100 ng/μl for a total input of 1 μg. If the sample quants lower than 100 ng/μl than a 10 ul direct stamp will be added to the process. Following plating, 2 μL of ERCC controls (using a 1:1,000 dilution) are spiked into each sample. RNA quality is measured using the Caliper LabChip GX system, which is used to calculate a RIN-equivalent RNA Quality Score. The percent of RNA fragments greater than 200 nucleotides are also quantified into a DV200 score. Together, this provides an accurate assessment of RNA quality. Using Illumina's TruSeq RNA Access Library Prep kit, a stranded cDNA library is prepared from isolated RNA, which is then hybridized to a set of DNA oligonucleotide probes to enrich the library for mRNA transcript fragments. This approach is optimized for FFPE tissue, which may not be suitable for poly(A) selection and ribosomal depletion due to degradation. Flowcell cluster amplification and sequencing are performed according to the manufacturer's protocols using HiSeq 2000s. Each run was a 76bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Picard Pipeline which includes de-multiplexing and data aggregation. Alignment is completed using the STAR alignment algorithm (5) against human reference hg19. Transcriptome Capture covers the RefSeq and GENCODE v12 databases to >98%. TPM counts were determined by RSEM (6). Samples were excluded based on issues in read depth (less than 10M reads in the library size) and gene sparsity (less than 15 000 genes have non-0 read counts). This left 6 samples: 8DCIS, 8IDC, 9DCIS, 9IDC, 2IDC and 3IDC for downstream analysis.

3. H&E and Cyclic immunofluorescence

Whole tissue mounts were sectioned to a thickness of 4 μm and stained with hematoxylin and eosin. Expert pathologists (SYP and JJ) have reviewed all cases to classify them as DCIS or IDC. The cyclic IF method was performed as previously described (7). 4-5 μm thick FFPE tissue sections were mounted on charged slides and baked overnight at 55 °C, followed by an hour at 65 °C, then deparaffinized and hydrated through xylenes and graded alcohols. Antigen retrieval was performed in 10 mM citrate buffer, pH 6, for 25 minutes in a programmed pressure cooker, follow by incubation in hot Tris/EDTA buffer, pH9, for 15 minutes (Agilent S2367). Slides were washed in PBS and blocked for 30 minutes in 10% normal goat serum (NGS) and 1% bovine serum albumin (BSA) in PBS. Directly labelled primary antibodies were diluted in 5% NGS and 1% BSA in PBS and slides were incubated overnight at 4°C in a humidity chamber. Slides were washed in PBS, mounted in Slowfade Gold plus DAPI (ThermoFisher, S36938) and imaged on the Zeiss AxioScan.Z1 with a Colibri 7 light source. Following imaging, fluorescence signal was quenched with incubation in 20 mM NaOH and 3% hydrogen peroxide H_2O_2 in PBS, under incandescent light, for 30 minutes. Slides were then washed in PBS and incubated in next round of antibodies for a total of 10 rounds of cyclic IF. See **Supplementary Table S2** for antibodies and catalog numbers.

4. WES characterization

The Getz Lab CGA WES Characterisation pipeline at the Broad Institute (https://docs.google.com/document/d/1VO2kX_fgfUd0x3mBS9NjLUWGZu794WbTepBel3cBg08/edit) was used to call, filter and annotate somatic mutations and copy number variation. This pipeline has been optimized to overcome sample contamination, DNA damage from fixation and filtering false positive calls. ContEst (8) is used to estimate cross sample contamination and DeTiN (9) to overcome potential contamination of tumor cells in normal tissue. Furthermore, oxidative damage to DNA by sample storage (e.g., C>T deamination artefacts due to FFPE preservation) or damage from library preparation are accounted for using PicardMultipleMetrics and Orientation Bias Filter (10). A panel of

8,334 normals (PoN) based on the TCGA cohort (MAF Panel of Normals) (11) is also used to filter false positive germline variants at low coverage regions, and BLAT Realignment Filter (12) is used to realign reads containing variants to ensure unambiguous mapping of that read. The pipeline uses GATK3 CNV pipeline (4) to characterize copy number profiles, whereas variant discovery was performed using a combination of MuTect1 (13), MuTect2 (4) and Strelka (14). Following consensus calling, variants were checked to ensure they pass the OxoG/FFPE Orientation Bias filters, PoN filter, blat realignment filter and have at least 5 reads supporting the variant and VAF of 0.08. This final variant list was then annotated with Oncotator (15) and Variant Effect Predictor (16). Variants in normal exome samples were called using GATK HaplotypeCaller and consolidated using GenotypeGVCFs (4) . Variants were annotated for pathogenicity using ClinVar (17). Ploidy of normal exome samples was estimated using GermlineCNVCaller (4). Normal tissue was from other quadrant for mastectomy specimen and adjacent breast tissue for small lumpectomy specimen.

5. External Data Sets

We have utilized the following data sets in our analysis: The Abba cohort consisting of 29 DCIS for which whole exome sequencing and RNA-sequencing is available (GSE69994) (18). This data set has been profiled in a similar fashion to the recurrence cohort. This cohort has 12 ER^- and 17 ER^+ cases for which exome sequencing has passed quality control. Clinicopathological data is supplied as supplementary data in the relevant publication, and lists patient subtype, pathologist scoring of TILs, ethnicity. No follow up information is available for this cohort. Normal tissue is defined as grossly unremarkable breast parenchyma at least 1 cm away from the DCIS and evaluated histologically. The Lesurf cohort contains 50 DCIS which have been profiled by array CGH and microarray profiling (GSE59248) (19). Patient subtype for this cohort was available through microarray-based PAM50 subtyping, with 15 luminal, 7 basal, 7 $HER2^+$ and 17 normal-like cases. Data was provided pre-processed from the corresponding author. Information on neoantigen load could not be inferred from this cohort. The TCGA cohort (20) has clinical, RSEM-normalized RNA seq and copy number

information available at firebrowse (http://firebrowse.org/). Additionally, neoantigen load, HLA-status and immune signatures have been previously characterized by a number of studies including the pan-cancer immunity study by Thorsson et al (https://gdc.cancer.gov/about-data/publications/panimmune) (21) and immune composition is available on the TIMER cistrome (http://timer.cistrome.org/) (22). We have utilized this preprocessed data in our down-stream analyses.

6. Data Analysis

Copy number analysis: Segmented contiguous chromosomal regions were called using GATK3 CNV in the recurrence and Abba cohorts (13). Array CGH data from Lesurf *et al* (19) were provided pre-processed by the corresponding authors and deletions and gains were defined as having a log fold change of ±0.3. Similarity between matched copy number profiles were assessed by two different metrics: The first calculates the proportion of bases with concordant calls across the exome, normalized by the total exome length. Due to potential contamination from normal tissue, some subclonal gains and losses may be lost using this method. Thus, we also computed the spearman correlation between log2(copy number ratios) across all captured exome regions.

Immune Gene Signatures: Immune signature analysis was performed using single-sample GSEA by the GSVA package (23), which ranks genes by expression value to perform gene set enrichment analysis. The ranked-based approach of this method allows comparisons of samples profiled on different platforms. Gene length normalized counts were used (TPM counts) from the recurrence cohort, Abba cohort and TCGA. Immune signatures were manually curated from the literature. This included gene signatures which have been reported by Thorsson *et al* to classify the TCGA into 6 different immune types (21), alongside activation and inhibitory signatures described by Pardoll *et al* (24), signatures obtained from single-cell profiling in melanoma by Tirosh *et al* (25), immunosuppressive signatures describing checkpoint proteins (Rosenthal and Pardoll) (24,26), and a CD8 T cell specific cytotoxic signature used by Jiang et al (27). These gene-lists are available at

github.com/polyak-lab/matchedDCISIDC/geneLists. Comparison of normalized enrichment scores between specific groups of patients were performed using two methods: Firstly, significance between enrichment scores was using student's t-test and corrected for multiple hypothesis testing. Secondly, a generalized linear model was used to infer the beta-coefficients or contributors of ER status and TILs to a ssGSEA score: These beta values were used to construct heatmaps that illustrate the contribution of a variable (ER status or TILs) to the reported ssGSEA scores.

$$ssGSEA_i = \beta_1 ERstatus + \beta_2 TILs + \varepsilon$$

Inferring immune composition: Proportions of different immune cell populations was inferred from RNA-seq TPM data in the recurrence, Abba and TCGA cohorts. This was performed on the TIMER2 cistrome platform (http://timer.cistrome.org/), which allows concurrent assessment using methods including TIMER (22), CIBERSORT (28), xCELL (29) and EPIC (30) through the immunedeconv (31) package. CIBERSORT (not run in absolute mode) and EPIC infers the relative proportion of a given cell-type, and statistical differences using these methods were assessed using a beta-regression model to account for data bounded between [0,1]. CIBERSORT-ABS (absolute mode), TIMER and xCELL provide enrichment scores, which were compared using Wilcoxon rank sum test. Heatmaps were constructed using beta-values from a generalized linear model, using only significant associations

$$EnrichmentScore = \beta ERstatus + \varepsilon$$

Differential gene expression analysis: Differential gene expression analysis between normal and DCIS in the Abba cohort, and DCIS and IDC in the recurrence cohort was performed using DESeq2 (32). Paired-analysis of DCIS-IDC samples in 8 and 9 was conducted using edgeR, using dispersion estimates calculated from all available RNA samples (33). Fold changes were used to perform Gene Set Enrichment Analysis (GSEA) (34) using the c2 compendium from MSigDB (34,35) through the HTSAnalyzeR2 (36) package. Gene sets with a FDR less than 0.05 were considered significant and visualized in a heatmap.

Genetically altered pathway maps: For each cohort (recurrence, Abba, Lesurf), the number of times each gene was considered aberrated by CNA or pathogenic mutation was counted. These values were used as the input for GSEA using HTSAnalyzeR2 (36) using the c2 compendium (34,35). Pathways with an adjusted *P* <0.1 were considered as significant, and overlapping pathways in the Recurrence, Abba (considering differential gene expression analysis, and genetic aberrated pathways as different entities), Lesurf cohorts and were used to create an enrichment map. The nodes of the resultant enrichment map reflect the number of enriched genes in the listed pathway, and the edge widths indicate the number of genes which are similar between two pathways.

Neoantigen prediction: All non-synonymous mutations were extracted and frame shift mutations were *in silico* translated using SIFT (37) into peptides 8-11 amino acids long. HLA-type inferred from WES using both Polysolver (38) and OptiType (39), showed a 95% concordance in predicted HLAs. Potential neoantigens were queried for predicted binding affinities to their predicted Polysolver HLAs using the NetMHCPan4.0 server (https://services.healthtech.dtu.dk/service.php?NetMHC-4.0) with peptide lengths of 8-11mers. Neoantigens were predicted using NetMHCPan4 (40), and strong and weak binders were defined as having an estimated %rank of 0.5 and 2 binding affinity of <50nM and <200nM respectively. Samtools mpileup (2) was used to determine the number of alternate reads from RNA-seq data at the corresponding mutation sites. Neoantigens were considered to be expressed when at least 2 RNA-seq reads supporting the mutation are present. HLA-types and neoantigens have previously been predicted in the TCGA cohort by Thorsson *et al*, and these results were obtained from GDC (https://gdc.cancer.gov/about-data/publications/panimmune) (21).

Histology Analysis: Computational image analysis was performed firstly by detecting and segmenting cells using Qupath (41), using an estimated cell radius of 4 μm, area between 10-200 μm^2, intensity threshold of 0.1. Extracted shape, intensity and haralick features were used to train a random forest classifier to differentiate between epithelial, stromal, immune cells, red blood cells and other debris in each image separately. Counts for epithelial, stromal and immune cells were exported for downstream analysis and classifications were reviewed by an expert observer.

Cyclic Immunofluorescence: Following collection of images, 16 bit tiffs were exported using the Zeiss Zen software, and images were registered based on DAPI staining using the detectSURFFeatures and automated feature matching functions in Matlab (R2017B). Nuclear and cell segmentation was performed using in-house Java software, using a watershed algorithm applied to DAPI and additional epithelial, immune and stromal marker channels. Mean intensity features were extracted from the nucleus or cytoplasmic region (depending on known localization) for each marker. Background intensity correction was performed by assigning each cell a co-ordinate based on its location within the scanning/imaging window of approximately 1,800x1,800 pixels. The spatstat package (42) was used to smooth the resulting point process pattern using a Gaussian kernel. The smoothed intensity at each square was considered as “background” and subtracted from the raw intensity value. Intensities within a tissue were scaled to a [0, 1] distribution and phenograph (43) was used to classify cells into clusters. Clusters were firstly identified as “microenvironment” or “epithelial” based on spatial location within the image, and expression of CK and CDH1 markers (epithelial) and expression of CD45, SMA, VIM markers (stroma). Clusters which had high intensity values for all markers after reviewing for spatial location were identified as red blood cells or auto-fluorescent regions of necrosis or proteins within milk ducts and were excluded from analysis. Following this, a secondary clustering was performed independently on the epithelial and stromal groups. Clusters were again reviewed manually inspection of features and spatial location, and clusters with similar properties were combined. Specifically, immune cells were identified based on high expression of the following markers: Foxp3 & CD4 (T regs), CD68 (Macrophage), PD1 & Ki67 (PD1), CD4 & CD45 (CD45 T Cells), CD8 & CD45 (CD8 T cells), CD20 & CD45 (B Cells). Tumor cells were primarily separated based on the following markers: Ki67 (proliferating), ER/PR (often co-expressed), basal-like (low CK, high VIM but found clustered amongst other epithelial cells), myoepithelial cells (CK5, CK14, sometimes SMA), and relative expression individual cytokeratins or CDH1 (CK+, CDK8low, CDH1low, CK low groups). Summaries of these expression profiles are shown in **Supplemental Figure S4D**. Differences between subclasses of cells in DCIS and IDC in a patient case was calculated using the proportionality test.

Spatial Analysis: Three different metrics for cellular interaction was considered: The first considered the Euclidean distance between two specific cells of interest and was termed the "interacting fraction". We determined the proportion of cells of class x which are within a given nearest neighbor distance r to a secondary cell class *y*. We define immune-tumor interacting fraction as the y-value of the nearest-neighbor CDF at *r*=10 μm (**Supplementary Fig. S4E**). When multiple immune cell types are present, a null distribution for each type was determined by permuting the immune cell labels 1000 times, and z-scores were reported to allow comparison across samples and calculation of p-values. The second metric considered the k-nearest neighbor metric, where we compute the average distance between a cell type *x* to its k=3 nearest neighbors of class *y*. We have assessed k=3, which makes metrics more robust to misclassified cell types. However, this metric is dependent on the frequency of a particular cell population, and larger distances are expected when measuring distances to a rare cell type. Thus, a r = 50 μm threshold was used to differentiate between close and far cells. Additionally, differences between cumulative distribution functions were assessed using the Kolmogorov-Smirnov test.

Lastly, the Morisita-Horn index for spatial overlap between two cell types of interest (44) was used as a metric which is normalizes for population frequency. We divide the tissue region of interest into *S* squares of size 100μm x 100μm and evaluated the frequency of each cellular population in each grid. The Morisita-Horn index for two cell types *x* and *y* are calculated as:

$$MH = \frac{2\sum_{i=1}^{S} x_i y_i}{\left(\frac{\sum_{i=1}^{S} x_i^2}{X^2} + \frac{\sum_{i=1}^{S} y_i^2}{Y^2}\right) XY}$$

where X and Y are the total counts of the populations of interest, and x_i and y_i refers to the frequency of these cells in grid S_i. To take into account differences in tumor architecture between DCIS and IDC, tiles containing only tumor cells (and are thus not at the stroma-tumor interface) were omitted in this analysis. 95% confidence intervals were obtained by shifting the grid in 5μm increments in both the X and Y directions. This metric provides a global overview of the co-occurrence of two cell types and can be interpreted as a correlation coefficient.

Immune related genomic hotspots: We firstly considered associations between expression of known immune checkpoint proteins or MHC-I genes and their copy number profiles to determine whether CNVs could contribute to over-expression. Within the Abba cohort (RNA-seq and Exome data), and the Lesurf cohort (aCGH and microarray), we calculate the correlation between gene expression and inferred CNVs and considered significant associations with FDR < 0.1. To determine new loci which could be associated with immune regulation, we divided the genome into 606 regions roughly 5MB long, and determined in each region the number of immune-related genes, as defined in InnateDB (45) and ImmPort (46) and the total number of genes. Immune-rich regions were determined by a hypergeometric test passing an FDR of 0.1, yielding 33 regions for consideration. CNVs in patients were categorized into “loss”, “neutral” and “gain” for each region of interest in the Abba and TCGA cohorts. To account for differences in patient stage, a generalized linear model was also used to evaluate associations between a ssGSEA score i and CNVs k=1 to 33:

$$ssGSEA_i = \sum \beta_k CNV_k \quad + \beta_{k+1} Stage + \ \varepsilon$$

This was performed separately for DCIS, ER^+ IDC and ER^- IDC. Significant beta coefficients following p value adjustment were used to construct the association heatmap. Stage takes the discrete factors: 1, 2, 3, 4.

Association between neoantigen frequency and immune signatures: We combined the DCIS cohorts (Abba and recurrence cohorts) and compared it to the TCGA cohort as IDC to determine similarities in neoantigen frequencies. Differences in populations were determined using a proportion test. To determine associations between neoantigens and immune signatures, we used a generalized linear model taking into account all common mutations of interest: For a given signature i, consider

$$ssGSEA_i = \sum \beta_k Gene_{k_{neo}} + \beta_{k+1} Stage + \ \varepsilon$$

Where $Gene_{kneo}$ is an indicator function which equals 1 if the patient is predicted to harbor that neoantigen. Stage takes the discrete factors: 1, 2, 3, 4.

BCR TCR repertoire analysis: CDR3 recombined regions in B cells and T cells were inferred from RNA using MixCR (47). CDR3 reads were filtered to those that have at least 2 supporting reads and have an amino acid sequence length of at least 6. These were searched against the IEDB database (48) of known CDR3 sequences. Due to the low number of reads aligning to TCR sequences, we focused on analyzing similarities in BCR repertoires, showing only clonotypes appearing at a frequency of >1%.

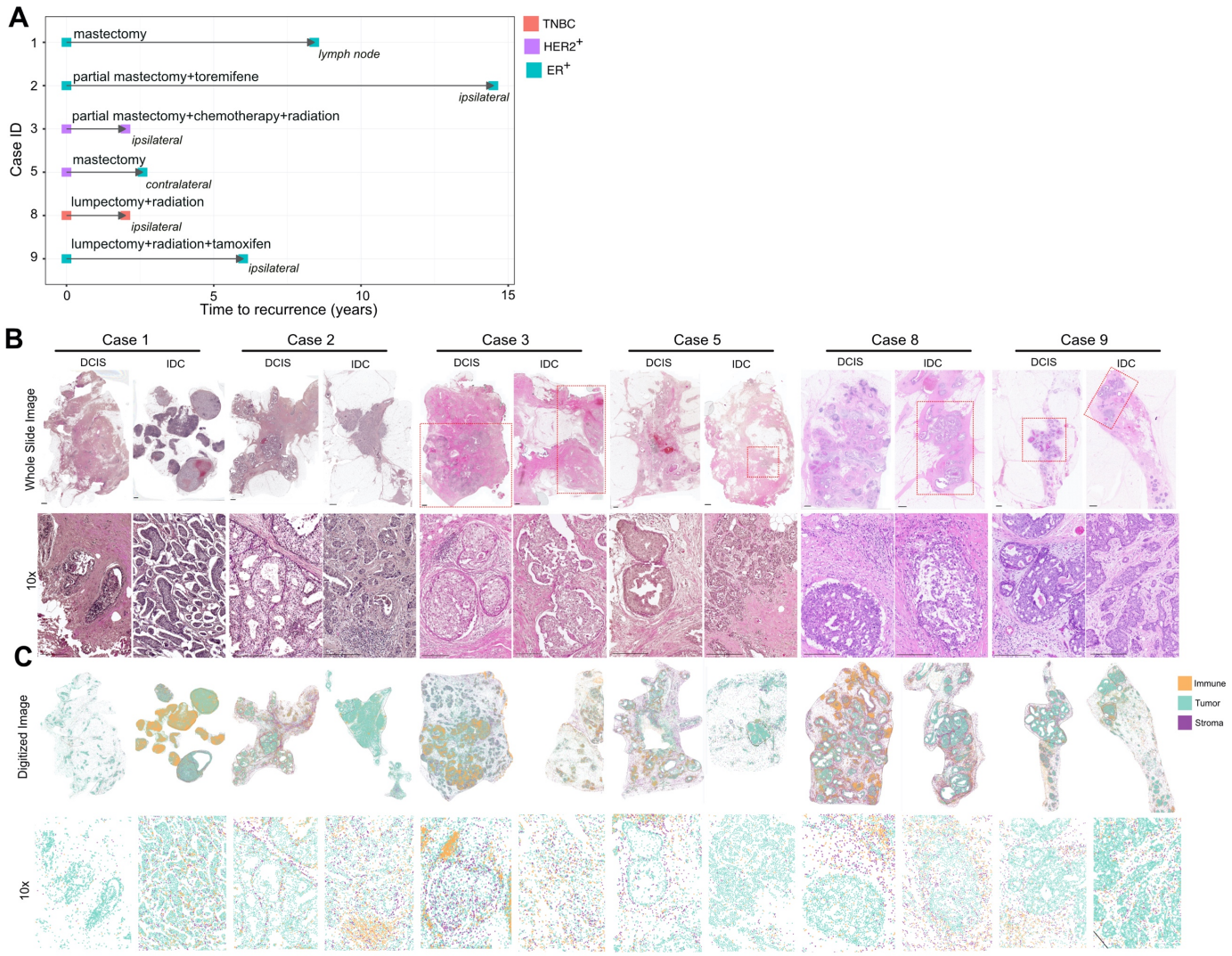
7. References

1. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM*, et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol **2011**;12:R1.
2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **2009**;25:1754-60.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N*, et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**;25:2078-9.
4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **2010**;20:1297-303.
5. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **2013**;29:15-21.
6. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **2011**;12:323.
7. Eng J, Thibault G, Luoh SW, Gray JW, Chang YH, Chin K. Cyclic Multiplexed-Immunofluorescence (cmIF), a Highly Multiplexed Method for Single-Cell Analysis. Methods Mol Biol **2020**;2055:521-62.
8. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics **2011**;27:2601-2.
9. Taylor-Weiner A, Stewart C, Giordano T, Miller M, Rosenberg M, Macbeth A*, et al.* DeTiN: overcoming tumor-in-normal contamination. Nat Methods **2018**;15:531-4.
10. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC*, et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res **2013**;41:e67.
11. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR*, et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **2014**;505:495-501.
12. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res **2002**;12:656-64.
13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C*, et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol **2013**;31:213-9.

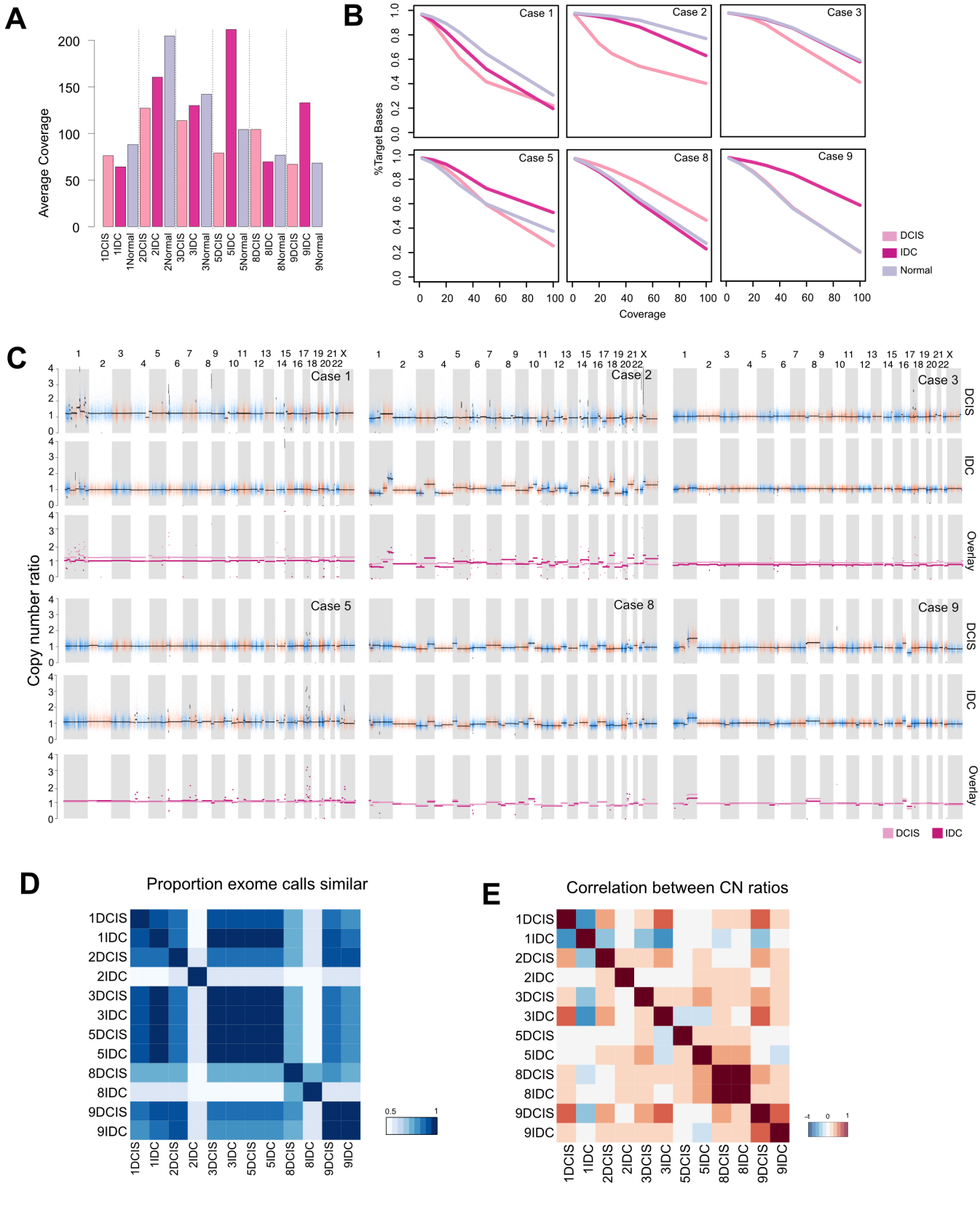
14. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics **2012**;28:1811-7.
15. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G*, et al.* Oncotator: cancer variant annotation tool. Hum Mutat **2015**;36:E2423-9.
16. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A*, et al.* The Ensembl Variant Effect Predictor. Genome Biol **2016**;17:122.
17. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S*, et al.* ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res **2016**;44:D862-8.
18. Abba MC, Gong T, Lu Y, Lee J, Zhong Y, Lacunza E*, et al.* A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. Cancer Res **2015**;75:3980-90.
19. Lesurf R, Aure MR, Mork HH, Vitelli V, Oslo Breast Cancer Research C, Lundgren S*, et al.* Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. Cell Rep **2016**;16:1166-79.
20. TCGA. Comprehensive molecular portraits of human breast tumours. Nature **2012**;490:61-70.
21. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH*, et al.* The Immune Landscape of Cancer. Immunity **2018**;48:812-30 e14.
22. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J*, et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol **2016**;17:174.
23. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics **2013**;14:7.
24. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer **2012**;12:252-64.
25. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science **2016**;352:189-96.
26. Rosenthal R, Cadieux EL, Salgado R, Bakir MA, Moore DA, Hiley CT*, et al.* Neoantigen-directed immune escape in lung cancer evolution. Nature **2019**;567:479-85.
27. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X*, et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med **2018**;24:1550-8.
28. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y*, et al.* Robust enumeration of cell subsets from tissue expression profiles. Nat Methods **2015**;12:453-7.
29. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol **2017**;18:220.
30. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife **2017**;6
31. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH*, et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics **2019**;35:i436-i45.
32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **2014**;15:550.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **2010**;26:139-40.
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **2005**;102:15545-50.
35. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics **2011**;27:1739-40.
36. Wang X, Terfve C, Rose JC, Markowetz F. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. Bioinformatics **2011**;27:879-80.
37. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res **2012**;40:W452-7.

38. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS*, et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol **2015**;33:1152-8.
39. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics **2014**;30:3310-6.
40. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol **2017**;199:3360-8.
41. Bankhead P, Loughrey MB, Fernandez JA, Dombrowski Y, McArt DG, Dunne PD*, et al.* QuPath: Open source software for digital pathology image analysis. Sci Rep **2017**;7:16878.
42. Baddeley A, Turner R. spatstat: An R Package for Analyzing Spatial Point Patterns. 2005 **2005**;12:42.
43. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD*, et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell **2015**;162:184-97.
44. Horn H. Measurement of "Overlap" in Comparative Ecological Studies The American Naturalist **1966**;100:419-24.
45. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R*, et al.* InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. Nucleic Acids Res **2013**;41:D1228-33.
46. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J*, et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data **2018**;5:180015.
47. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV*, et al.* MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods **2015**;12:380-1.
48. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR*, et al.* The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res **2019**;47:D339-D43.

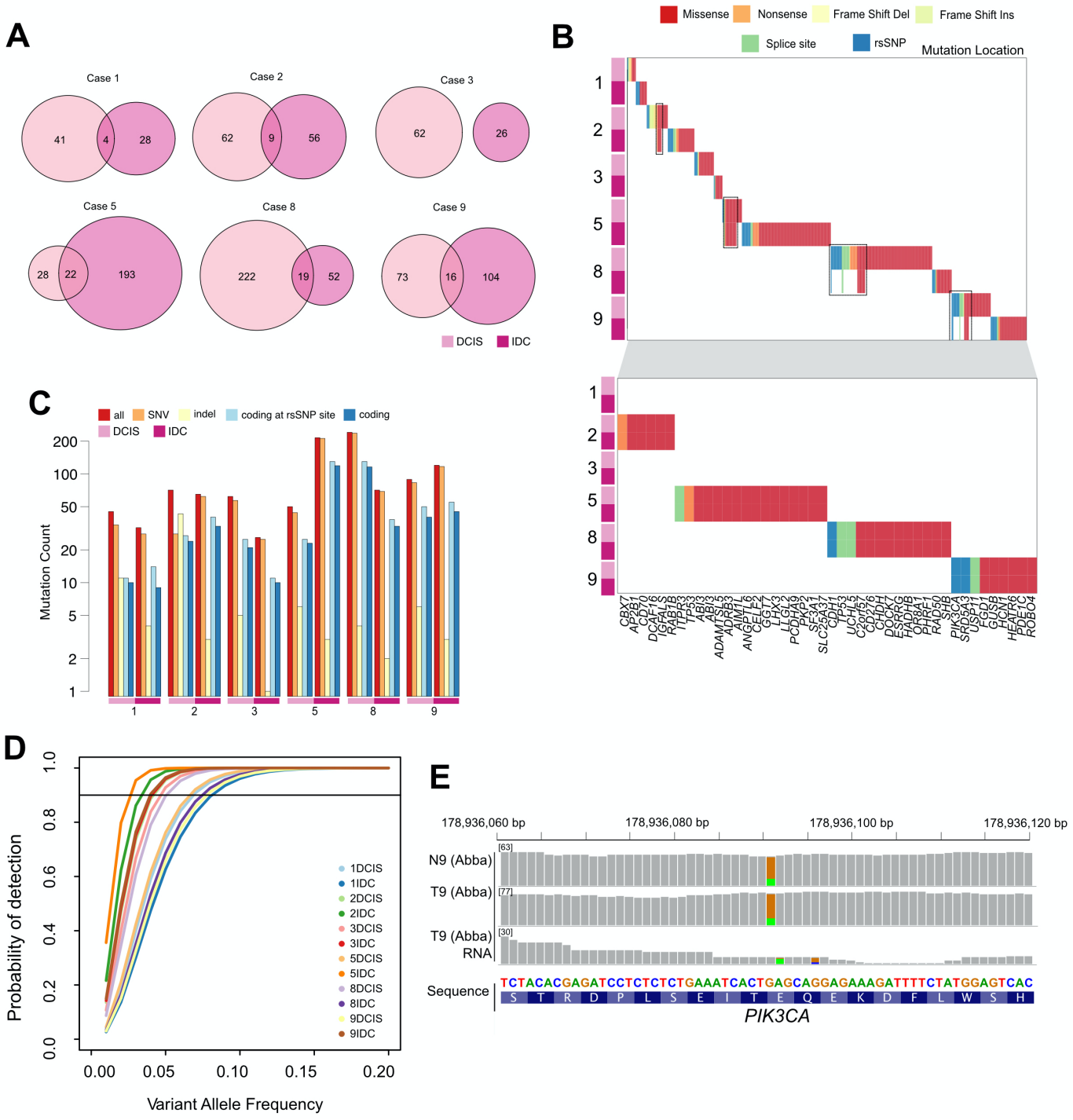
Supplementary Figure S1. Overview of patient cohort. A, Summary of patient cohort. **B,** Whole-slide H&E images of the recurrence cohort. Macrodissected regions for sequencing are indicated by red boxes. Scale bar: 1 mm in Whole Slide Images, 50 µm at 10x magnification. **C,** Digitized summary of the H&E slide showing immune, tumor and stromal cells. Digitally macrodissected regions are outlined in black.



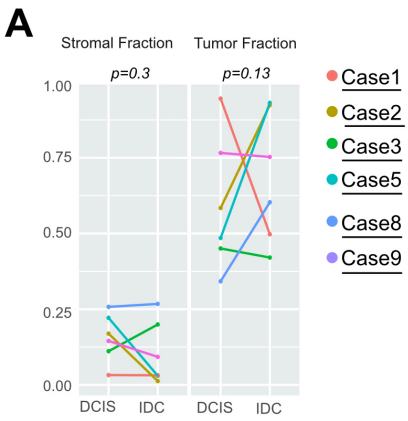
Supplementary Figure S2. Whole exome sequencing of matched DCIS and IDC transition. A, Average coverage inferred using GATK DepthOfCoverage. **B,** Percentage bases sequenced with given coverage. **C,** Copy number ratios of samples generated using GATK. **D,** Heatmap of consistency of CNV calls (loss, gain, neutral) in assessed samples. **E,** Correlation coefficients of inferred log2 copy number ratios between samples.

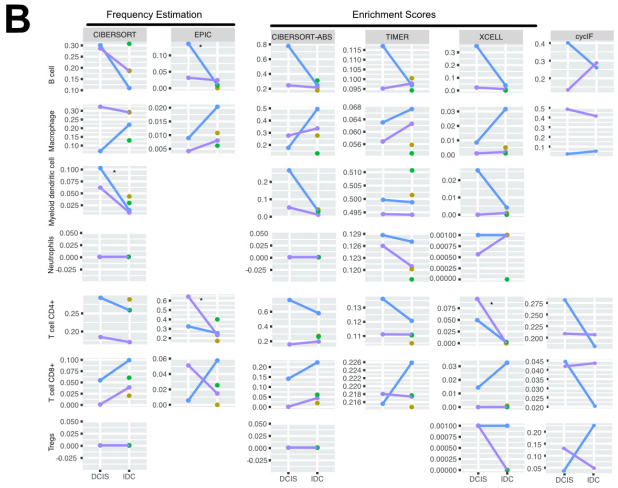


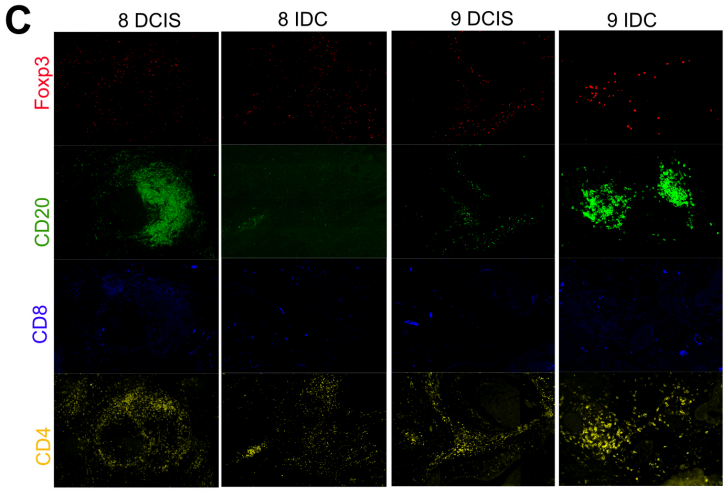
Supplementary Figure S3. Mutational profiling of the DCIS-to-IDC transition. A, Venn diagram of overlap in mutations between samples. **B,** Summary of coding mutations in the recurrence cohort, with zoomed in insert on conserved coding mutations. Each x location indicates a unique genomic site. **C,** Summary of mutational load observed in each patient. **D,** Probability of seeing 2+ reads supporting a variant with increasing VAF given sample coverage estimated in Supplementary Figure S1, using binomial model. **E,** PIK3CA E545K mutation in normal and tumor of immune-cold patient T9.

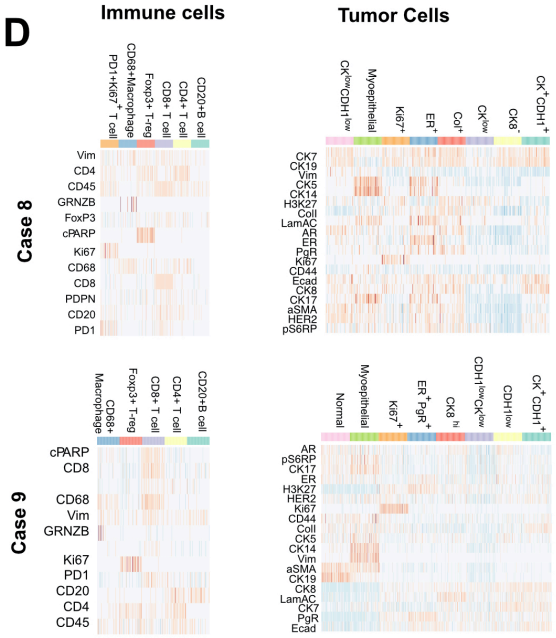


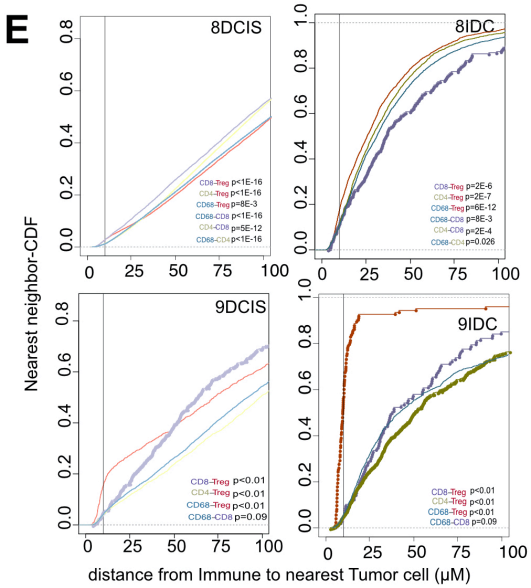
Supplementary Figure S4. Phenotypic immune properties of patient cohort. **A,** Comparison of matched stromal and tumor fractions from H&E images. Differences computed using Wilcoxon rank sum test. **B,** Immune composition of recurrence cohorts inferred using various RNA-deconvolution methods and from cycIF. **FDR*<0.1 using beta-regression model (CIBERSORT, EPIC) or Wilcoxon rank sum test (CIBERSORT-ABSOLUTE, TIMER, XCELL). **C,** Single-channel immunofluorescence panels shown in Figure 4A. **D,** Expression profiles of each immune and tumor cluster from cycIF data. Heatmaps were constructed by randomly sampling 100 cells in each cluster. **E,** Cumulative distribution function of distances from immune cells to nearest tumor cells. Differences in distribution were calculated using the K.S statistic. **F,** Heterogeneity in myoepithelial expression of CK5 and SMA in 9DCIS.

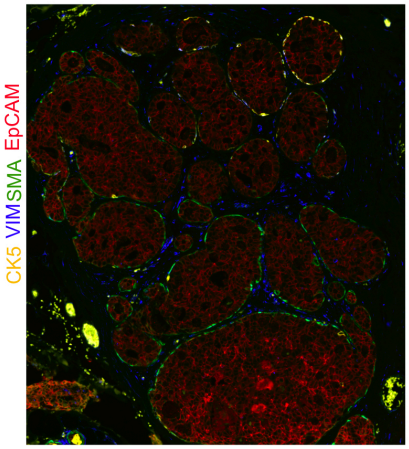












Supplementary Figure S5. Immune-related genomic changes and neoantigen load. A, Copy number profiles of loci bearing MHC-I presenting genes or immune checkpoint genes in the Abba and Lesurf cohorts. **B,** Genome wide view of 33 loci associated with lymphocytic infiltration and frequency of gains and loss in the DCIS cohorts. **C,** Summary of patient predicted HLAs and neoantigen load in the Abba and recurrence cohort. **D,** Mutational frequency compared to number of unique sites for commonly mutated genes in the TCGA cohort. **E,** Summary of the most common HLAs found in the TCGA breast cohort. **F,** Richness of TCR and BCR clonotypes in patient tumor sections and adjacent tissue. Clonotypes supported by at least 2 reads are reported.

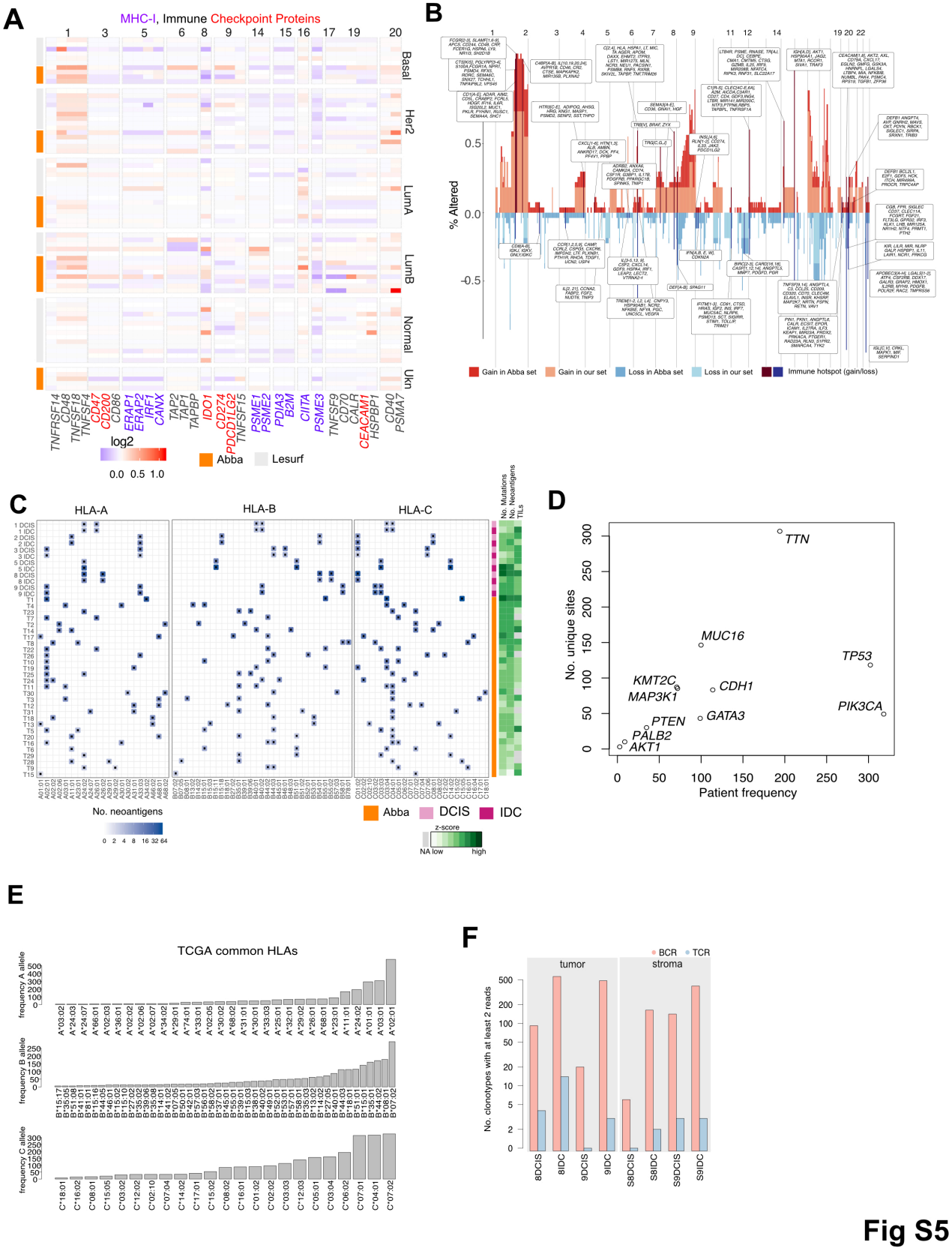


Fig S5

Supplementary Table S1. Clinicopathological characteristics of patient cohort.

Supplementary Table S2. Antibody panel used for cyclicIF.

Supplementary Table S3. Summary of mutations and neoantigen expression in the recurrence cohort (Related to Fig 1C, 6A).

Supplementary Table S4. Summary of altered pathways in DCIS/IDC. (Related to Fig 1F).

Supplementary Table S5. Enriched gene sets between DCIS and IDC (Related to Fig 3A).

Supplementary Table S6. Identified chromosomal regions associated with immune changes (Related to Fig 5).

Supplementary Table S7. Mutational, neoantigen and rsSNP frequency of commonly mutated genes in TCGA and Abba cohorts (related to Fig 6C,E).