

Supplementary Information

Multi-omics profiling uncovers immune-molecular clusters with distinct chemo-immunotherapeutic vulnerabilities in a mouse model of triple-negative breast cancer

Olivier Castellonet^{1#}, Jean Monatte^{1#}, Nathan Corvaisier^{1§}, Abdessamad El Kaoutari^{1§}, Muge Kaya², Lorène Ferreira², Stephane Audebert¹, Luc Camoin¹, Alexandre de Nonneville¹, Anthony Gonçalves¹, Jean-Paul Borg^{1,3}, Paula Michea¹, Flavio Maina^{1,4*}, Fabienne Lamballe^{1,4*}

¹ Aix Marseille Univ, CNRS, INSERM, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille (CRCM), Marseille, France.

² Aix Marseille Univ, Experimental Histopathology ICEP Platform, CNRS, INSERM, Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille (CRCM), Marseille, France.

³ Institut Universitaire de France (IUF).

⁴ Aix Marseille Univ, Turing Center for Living Systems, Marseille, France

These authors share first authorship

§ These authors contributed equally as second authors

* Corresponding authors: flavio.maina@univ-amu.fr, fabienne.lamballe@inserm.fr

Table of contents

SUPPLEMENTARY MATERIAL AND METHODS.....	2
SUPPLEMENTARY FIGURE LEGENDS.....	18
SUPPLEMENTARY TABLE LEGENDS.....	26
SUPPLEMENTARY REFERENCES	27

SUPPLEMENTARY MATERIAL AND METHODS

Mouse models.

The MMTV-R26^{Met} mouse model was generated as previously reported¹. Briefly, the R26^{stopMet} mice (officially designated as *Gt(ROSA)26Sor^{tm1(Actb-Met)Fmai}*, RRID : MGI:5052130) which carry a conditional *Met* transgene at the *Rosa26* locus, were crossed with the MMTV-*Cre* line (*B6129-Tgn(MMTV-Cre)4Mam*, RRID:IMSR_JAX:003553). Unlike previous studies, these parental mouse lines were maintained on pure genetic backgrounds: *129S2/SvPasOrlRj* (RRID:MGI:6430818) for the R26^{stopMet} line and *C57BL/6Jrj* (RRID:IMSR_RJ:C57BL-6JRj) for the MMTV-*Cre* mice. This breeding strategy was designed to generate MMTV-R26^{Met} F1 transgenic mice with a precise mixed genetic background (50% *129S2/SvPasOrlRj*, 50% *C57BL/6Jrj*). Genotyping of the resultant mice was performed by PCR analysis of genomic DNA, following established protocols¹. To enhance the incidence of mammary gland tumors, MMTV-R26^{Met} mice were continuously bred under controlled conditions. Tumor formation occurred in 54- to 87-week-old multiparous females, consistent with our earlier findings¹.

For syngeneic tumor models, primary tumors were dissected and a single-cell suspension was prepared as detailed in the tissue processing section. The cell suspension was adjusted to a concentration of 1×10^6 cells per 100 μ L, in a 1:1 mixture of DMEM (Dulbecco's Modified Eagle's Medium, ThermoFisher Scientific) and Matrigel (Corning, #356237). For serial tumor passaging, tumors were mechanically dissociated to single cells and reseeded with an approximate 100-fold dilution at each generation, thereby minimizing co-transfer of stromal or immune cells. F1 mice generated by crossing *129S2/SvPasOrlRj* with *C57BL/6Jrj* mice, referred to as syngeneic mice, were anesthetized and orthotopically injected with 100 μ L of the cell mixture into the fourth mammary fat pad using a 28-gauge insulin syringe. Tumor volume, measured twice per week with a caliper, was determined using the formula: $(L \times W^2)/2$ where L represents length and W represents width. To estimate the tumor growth rate in mm³ per day, the following formula was used: $(V_{last} - V_{first})/\Delta T$ where V_{last} is the tumor

volume at the last measurement (mm^3), V_{first} is the volume at the first measurement, and ΔT is the time interval between the two measurements (days). The experiment was concluded when tumors reached an ethical maximum volume of 1500 mm^3 . All tumors were collected at this standardized volume to ensure comparable tumor size for subsequent immunoprofiling, minimizing potential bias due to differences in tumor growth stage. The resulting tumors were harvested, dissociated, and re-implanted into new recipient mice using the same protocol for primary tumors. This procedure was repeated for up to 7 generations. For each generation, a portion of the dissociated cells was cryopreserved in FBS supplemented with 10% DMSO and stored in liquid nitrogen to establish a biobank.

In vivo therapeutic evaluation.

Freshly dissociated tumor cells were orthotopically re-implanted into the mammary fat pad on day 0, as described above. For T37 tumors, 1.3×10^6 cells were injected; for T39 tumors, 1×10^6 cells. When tumors reached $50\text{--}100 \text{ mm}^3$, mice were randomized into treatment groups. Epirubicin (TargetMol: #T69093) was administered intraperitoneally (IP) at 10 mg/kg ($100 \mu\text{L}$) once per week for 3 weeks. Anti-PD-1 antibody (Ichorbio: #ICH1132, RRID:AB_2921498; TargetMol: #T78269, RRID:AB_3714837) was given IP at $200 \mu\text{g}$ ($100 \mu\text{L}$) twice weekly for 3 weeks. Control mice received $100 \mu\text{L}$ sterile 0.9% NaCl, IP. Tumor volume was measured twice weekly using a caliper, and mouse body weight was recorded at the same frequency. Mice were euthanized when tumors reached 1500 mm^3 and tumors were subsequently harvested for analysis.

Animal care and ethical compliance.

All procedures involving animals were conducted in accordance with the European Community Council Directive of 22 September 2010 on the protection of animals used for experimental purposes (2010/63/EU). Experimental protocols adhered to guidelines set by the institutional Ethical Committee for animal research (Comité d'éthique pour l'expérimentation animale – Comité d'éthique de

Marseille) and complied with French law under agreement numbers E1305521 and E1305504, delivered by the “Ministère de l’Enseignement Supérieur et de la Recherche”. Mice were housed in a dedicated pathogen-free facility with a controlled light/dark cycle, fed ad libitum in environmentally enriched cages following Safe Complete Care Competence (SAFE A04). The project received authorization under APAFIS numbers #8214-2016121417291352.v5, #33371-2021100612164824 v6, and #44186-202307131711163 v4, issued by the “Ministère de l’Enseignement Supérieur et de la Recherche”.

Histological analyses.

For histopathological analysis of mouse tissues, control mammary glands and dissected tumors were fixed in 4% paraformaldehyde (PFA, Sigma) for 48 hours, followed by paraffin embedding. Sections (4 µm) were then stained with hematoxylin, eosin, and saffron (HES) according to the standard protocol of the core histological facility of the « Centre de Recherche en Cancérologie de Marseille » (CRCM) or used for immunohistochemistry (IHC). The histological subtype of each tumor was classified by a pathologist according to the WHO Classification of Breast Tumors (2019), and the histological grade was assessed using the Nottingham histological grading system (Elston–Ellis modification of the Scarff–Bloom–Richardson system). All IHC analyses were conducted using the DISCOVERY ULTRA Ventana automated system. For all IHC experiments, sections were preheated at 62°C overnight (o/n) and then deparaffinized using EZ Prep (Ventana: #950-102) in the automated device. Antigen retrieval was performed by incubating the sections with cell-conditioning solution 1 (pH 7.8) or 2 (pH 6.5) (Ventana: #950-124 or #760-107) for 8 min at 95 °C, depending on the antibodies used (Table S7). Sections were then incubated with the primary antibody at the dilutions and for the incubation times specified in Table S7. The signal was developed by incubating with OmniMap anti-rabbit HRP (Table S7) for 16 min, H₂O₂ for 4 min and then ChromoMap DAB for 8 min (Ventana: #760-159). Finally, the sections were counterstained with Hematoxylin II (Ventana: #790-2208). The slides were mounted and then imaged

using the NanoZoomer slide scanner. Image analysis, such as quantification of positive cells, was performed using QuPath software (RRID:SCR_018257) and its positive cell detection module.

TNBC tumor microarray (TMA) analysis.

Immunohistochemical staining was performed on sections from the IPC TNBC tumor microarray (TMA) by the experimental pathology platform ICEP following standardized protocols. Slides were scanned using a Hamamatsu S60 whole-slide scanner. The evaluations of PD-1, CD3, CD68, and MPO immunohistochemical stainings were performed by a pathologist blinded to patient clinical and survival data. For each marker, the percentages of positive cells were quantified separately in the total tumoral stromal cells and in the total tumoral cells, in accordance with recommendations distinguishing the two compartments². Only the immune cells located within the borders of invasive tumors and within the invasive edges were included in the evaluations. The immune cell infiltrations which were located outside the tumor boundaries (e.g., adjacent normal tissue or DCIS) were excluded. Areas of necrosis or fibrosis were also excluded, and immune-positive fractions located within necrotic regions were not counted.

The average total tumoral and stromal cell numbers were used to evaluate the percentage of positive cells in the whole tumoral area, not focusing on hot spots. The quantifications of these areas were performed using QuPath³, and results are provided in Table S5. For each marker, a global score was calculated as the mean percentage of positive cells relatively to the areas of the tumoral and tumoral stroma compartments. A normalized abundance z-score was then computed by scaling each marker's proportion of positive cells to the total abundance of the three immune populations analyzed (macrophages, neutrophils, lymphocytes).

Tissue processing.

Tumors and control mammary glands were dissected and minced into small fragments, followed by enzymatic digestion for 20-35 min at 37°C. The digestion solution consisted of type IV collagenase (2 mg/ml, Sigma), hyaluronidase (1 mg/ml, Sigma), and DNase I (20 µg/ml, Roche) in DMEM/F12 (Dulbecco's modified Eagle's media/F12, 1/1, ThermoFisher Scientific) supplemented with 10% fetal bovine serum (FBS, Sigma) and penicillin–streptomycin (P/S, 100 U/ml / 0.1 mg/ml, ThermoFisher Scientific). After digestion, samples were centrifuged, and the pellet was incubated on ice for 1 min with red blood cell lysis buffer (ThermoFisher Scientific). They were then washed in PBS and further incubated for 5 min at 37°C with trypsin (0.05%, Gibco). Subsequently, the cells were resuspended in DMEM/F12 + 10% FBS + P/S, and filtered through a 40µm nylon cell strainer to remove aggregates. The resulting single-cell suspension was ready for subsequent analyses, such as transcriptomic, proteomic, flow cytometry studies as well as for grafting into syngeneic mice.

Flow cytometry.

Tumors and control mammary glands were digested as described above, and single cells were rinsed in a solution composed of 2% FBS and 2mM EDTA in PBS. Cells were then treated with anti-mouse CD16/CD32 monoclonal antibodies (LifeTechnologies, 14-061-82, RRID:AB_467133) diluted in PBS, and incubated for 15 min on ice in the dark. After washing, they were incubated with fluorochrome-conjugated antibodies (Table S8) for 25 min in the dark, at dilutions optimized by titration. For live/dead cell discrimination, DAPI (Sigma-Aldrich, D9542) was added directly to the cell suspension at a final concentration of 0.1 µg/mL. Finally, cells were resuspended at 1×10^5 - 1.5×10^6 in PBS and acquired on the spectral cytometer Cytex Aurora (5 lasers 16UV-16V-14B-10YG-8R configuration) from the Flow Cytometry Core of the CRCM. Analysis and compensation were carried out using the Cytex SpectroFlo software (version 3.2.1, RRID:SCR_025494). Single-color controls were included for quality control. Both supervised and unsupervised data analyses were performed using OMIQ. Heatmap

visualization of immune cell subset abundance was generated with row-wise min-max normalization. Hierarchical clustering was then performed on the normalized data using Euclidean distance and Ward's linkage on column vectors (samples).

Whole exome sequencing.

Tumor samples and non-cancerous control spleens were snap frozen and sent to Novogene Ltd for DNA extraction and whole exome sequencing (WES). Briefly, genomic DNA was sheared into fragments (180-280 bp), repaired, A-tailed, and ligated with Illumina adapters. After PCR amplification, size selection, and purification, libraries were hybridized with biotin-labeled probes to capture gene exons. Non-hybridized fragments were removed, and the captured libraries were enriched via PCR. Quantification and size distribution of the libraries were assessed using Qubit, real-time PCR, and a bioanalyzer. Libraries were pooled based on concentration and sequenced on Illumina platforms using the PE150 strategy. Sequencing data were filtered to remove contaminants and low-quality reads. The Burrows-Wheeler Aligner (BWA, RRID:SCR_010910) was used for alignment to the mouse reference genome (mm10), followed by sorting with Sambamba (RRID:SCR_024328) and duplicate marking with Picard (RRID:SCR_006525). Final BAM files were indexed and used to compute coverage and depth metrics for quality assessment and downstream analyses.

Somatic variant calling was performed using VarScan 2.0.2 (VarScan somatic, RRID:SCR_006849), which compares tumor and matched normal BAM files to detect somatic single nucleotide variants (SNVs) and small insertions or deletions (indels). Variants were filtered based on p-value (<0.05) and read coverage. To further refine the variant calls, the VarScan "processSomatic" function was employed to classify variants into germline, "Loss of Heterozygosity" (LOH), and somatic categories. Annovar (RRID:SCR_012821) was employed for comprehensive annotation of identified variants against reference databases, including RefSeq (RRID:SCR_003496), dbSNP (RRID:SCR_002338). Functional consequences of variants were predicted using Annovar's functional prediction tools.

Subsequently, the VariantAnnotation R package was utilized to predict amino acid alterations resulting from coding variants. For the tumor mutation analysis, only somatic variants were retained, excluding known polymorphisms, silent variants, and variants located in non-coding regions.

Somatic copy number variation (CNVs) analysis was performed using VarScan “copynumber” function to identify genomic regions with gain or loss in tumor samples compared to the control. Paired tumor-control BAM files were subjected to CNV analysis. The raw CNV calls from VarScan2 were further processed with the “copyCaller” function, which performs GC-correction and filters the raw CNV calls to identify high-confidence CNV events. The final output for each tumor sample consisted of a file containing high-confidence CNV calls, detailing the genomic region, copy number change (gain/loss), and other relevant information. Identified CNV segments were annotated against the mouse genome (EnsDb.Mmusculus.v79) to determine gene-level copy number alterations. The CNV value at the gene level was assigned based on the CNV fragment with the highest absolute $\log_2$ ratio observed within the coding region. CNV burden was determined by counting the number of genes with a CNV value of $|\log_2FC| \geq 0.5$ (to define high-confidence CNVs) and dividing this value by the total number of genes, as previously described⁴. Graphical representations were generated using the Maftools R package (RRID:SCR_024519) and custom R scripts. Whole-exome sequencing data have been deposited in the Sequence Read Archive (SRA, RRID:SCR_001370) under accession number PRJNA1311023.

Sample preparation and mass spectrometry (MS).

Cell pellets were recovered with RIPA buffer in proportion to the number of recovered cells to obtain approximately 1ug/ul, based on the assumption that 1×10^6 cells represent 50ug. A theoretical maximum of 20ug of protein was collected for further processing. For samples containing less than 20ug of protein, the entire sample was diluted with lithium dodecyl sulfate (LDS) sample buffer to a final volume of 20ul. Proteins extracts were loaded on NuPAGE™ 4–12% Bis–Tris acrylamide gels according to the manufacturer’s instructions (Life Technologies). The electrophoresis was stopped as

soon as the proteins had stacked in a single band. Protein-containing bands were stained with Imperial Blue (Pierce), cut from the gel, and digested with high sequencing grade trypsin (Promega) before mass spectrometry analysis.

Briefly, the gel pieces were washed and destained using few steps of 100mM ammonium bicarbonate (NH_4HCO_3). Destained gel pieces were shrunk with 100 mM ammonium bicarbonate in 50% acetonitrile and dried at room temperature (RT). Protein spots were then rehydrated with 10mM DTT in 25 mM ammonium bicarbonate (pH 8.0) for 45 min at 56°C. This solution was replaced by 55 mM iodoacetamide in 25 mM ammonium bicarbonate (pH 8.0), and the gel pieces were incubated for 30 min at RT in the dark. They were then washed twice in 25 mM ammonium bicarbonate and finally shrunk by incubation for 5 min with 25 mM ammonium bicarbonate in 50% acetonitrile. The resulting alkylated gel pieces were dried at RT. The dried gel pieces were reswollen by incubation in 25 mM ammonium bicarbonate (pH 8.0) supplemented with 12.5 ng/ μl trypsin (Promega) for 1h at 4°C, then incubated overnight at 37°C. Peptides were harvested by collecting the initial digestion solution and carrying out two extractions; first in 5% formic acid, followed by 5% formic acid in 60% acetonitrile. Pooled extracts were dried using a centrifugal vacuum system. Samples were reconstituted in 0.1% Trifluoro acetic acid (TFA) and 4% acetonitrile before mass spectrometry using an Orbitrap Fusion Lumos Tribrid Mass Spectrometer (ThermoFisher Scientific, San Jose, CA) online with an Ultimate 3000RSLCnano chromatography system (ThermoFisher Scientific, Sunnyvale, CA). Peptides were separated at 40°C using a two-step linear gradient (4–20% acetonitrile/ H_2O ; 0.1% formic acid for 110 min and 20–32% acetonitrile/ H_2O ; 0.1% formic acid for 10 min). An EASY-Spray nanosource was used for peptide ionization (2,200 V, 275°C). MS was conducted in data-independent acquisition (DIA) mode. Full MS scans were acquired within an m/z range of 375–1,500 at a resolution of 120,000 at m/z 200, with the automatic gain control (AGC) set at 4.0×10^5 with a 50 ms maximum injection time. MS2 spectra were acquired in the Orbitrap with a resolution of 30,000, covering a mass range of 200–1800 m/z , after isolation of parent ion in the quadrupole and fragmentation in the higher-energy collisional dissociation (HCD) cell under collision Energy of 30%. The DIA parent ion range was divided

into 40 windows (16 Da wide) between 400 to 1000 m/z and into 10 windows (50 Da wide) from 1000 to 1500 m/z.

For protein identification and quantification, relative intensity-based label-free quantification (LFQ) was processed using the DIA-NN 1.8.2 beta27 algorithm⁵. Raw files were searched against the mouse UniProt database (55,260 entries, 2023/04/03), implemented with a contaminant database⁶ and a bovine database for common serum proteins. The following parameters were used for searches: (i) trypsin allowing cleavage before proline; (ii) one missed cleavage allowed; (iii) cysteine carbamidomethylation (+57.02146) as a fixed modification and methionine oxidation (+15.99491) and N-terminal acetylation (+42.0106) as variable modifications; (iv) a maximum of 1 variable modification per peptide allowed; and (v) a minimum peptide length of 7 amino acids and a maximum of 30 amino acids. The “match between runs” option was enabled. The precursor false discovery was set to 1%. DIA-NN parameters were configured in double-pass mode for Neural Network classifier, Robust LC High precision for quantification strategy and RT-dependent mode for Cross-run normalization. The library was generated using Smart profiling setup. MS1 and MS2 mass accuracy was automatically calculated with precursor charge fixed between 2 and 4.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD067482. The main output file from DIA-NN was further filtered at 1% FDR, and LFQ intensity was calculated using our DIAGui package at a 1% q-value (<https://github.com/marseille-proteomique/DIAGui>⁷). Statistical analysis was carried out with the Perseus software (version 1.6.15.0), with proteins pre-filtered to remove contaminants and bovine proteins⁸.

Human TNBC samples, IPC cohort.

For comparison with murine tumors, we analyzed a local cohort of patients with TNBC treated at the Institut Paoli-Calmettes (IPC, Marseille, France). All patients were treatment-naïve at the time of

diagnosis and underwent initial tumor sampling before any systemic or local therapy. Ki67 index evaluation was performed in the institutional pathology laboratory according to routine procedures. Clinical and pathological data were fully anonymized prior to analysis. This part of the study was approved by the institutional review board and local ethics committee of IPC (MIB-TNBC-IPC 2023-040).

Publicly available RNA sequencing and proteomic human data analyses.

Publicly available RNA sequencing (RNAseq) and proteomic data from human breast cancer patients were analyzed. Clinical, histological, transcriptomic, and genetic data were extracted from The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) databases using the cBioPortal web tool, or from the GSE96058 dataset⁹. Patients were classified into TNBC and non-TNBC groups based on the presence or absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 status as recorded in their clinical data. Normal breast tissue samples were selected as controls for comparative analysis. Differential expression analysis of the transcriptome was performed to compare BC patients (TNBC, n=99; non-TNBC, n=544) with normal breast tissue (n=106). This analysis was conducted using the TCGAanalyze R package (RRID:SCR_003193). The differentially expressed genes (DEGs) identified in this analysis were then compared to those from *MMTV-R26^{Met}* tumors obtained from proteomic data using the Spearman correlation test.

For the human TNBC proteomic data, information was extracted from the study by Asleh and coll¹⁰. Centroids of the proteome profiles for each of the four TNBC clusters identified in their study were calculated and correlated with the proteomic signatures of *MMTV-R26^{Met}* tumors using the Spearman correlation test.

Human immune cell inference from bulk RNAseq data.

The GSE96058 dataset, which includes RNAseq profiles from 3,273 human BC samples, was used to infer immune cell composition. Immune cell proportions were estimated using CIBERSORT¹¹ (RRID:SCR_016955) with the LM22 signature matrix, allowing the deconvolution of 22 immune cell subsets from bulk RNAseq data. BC subtypes were assigned according to ER, PR, and HER2 status as predicted by multigene classifiers (MGC). The frequencies of the inferred immune cell populations were then transferred to RStudio for hierarchical clustering analysis. For interpretability and to reduce dimensionality, related immune subsets were grouped into broader categories as follows: macrophages (M0, M1, and M2), CD4⁺ TLs (naive, memory resting, memory activated, and follicular helper), BLs (naive, memory, and plasma cells), NK cells (resting and activated), and DCs (resting and activated). These aggregated immune populations were subsequently used for further analyses.

Single-Cell RNA sequencing analysis.

MMTV-R26^{Met} syngeneic tumors were dissected at passage 7 (S34, S35, S37, and S39) and dissociated as described in the tissue processing section. Cell counting and viability assessment using trypan blue revealed that all samples exhibited viability above 80%, so no sorting was performed. The cell concentration was then adjusted to approximately 1,000 cells/ μ L for library preparation. Single-cell RNA sequencing libraries were generated using the 10x Genomics Chromium Single Cell 3' Library and Gel Bead Kit v3.1, following the manufacturer's protocol. Specifically, 12,000 cells per sample were loaded onto a Chromium Controller to encapsulate individual cells into Gel Beads in Emulsion (GEMs). After cell lysis within the GEMs, mRNA was reverse transcribed to produce cDNA, which was then amplified by PCR. This cDNA was subsequently used for sequencing library preparation. The libraries were sequenced on Illumina NextSeq 500 and NextSeq 2000 sequencers, targeting a depth of approximately 60,000 reads per cell. The library generation, validation, and sequencing were performed at the "Transcriptomique et Génomique Marseille Luminy (TGML)" platform. Raw sequencing data were processed using Cell Ranger (v6.1.1, 10x Genomics, RRID:SCR_017344). The

cellranger mkfastq command was employed to demultiplex the raw base call files into FASTQ files, followed by alignment to the mouse reference genome (mm10) using the “cellranger count” command, which also produced gene-cell expression matrices for each sample. Quality Control (QC) was performed using multiple approaches to ensure high-quality cell and gene selection. Ambient RNA contamination and cell barcode calling were initially assessed using the knee detection method implemented in Cell Ranger. Barcode rank plots were inspected to confirm appropriate cell versus background separation. Mitochondrial and ribosomal genes were identified based on gene symbols, where mitochondrial genes were defined by the prefix “mt-” and ribosomal genes by the prefix “Rpl” or “Rps,” using pattern matching on the gene annotation metadata. The scater package (v1.26.0, RRID:SCR_015954)¹² was then used to calculate per-cell quality metrics with “perCellQCMetrics” function, including the percentage of mitochondrial and ribosomal gene expression per cell. Outlier cells were identified and filtered based on library size, number of detected genes, and mitochondrial and ribosomal gene content using the “isOutlier” function from the scater package. This method flags cells that deviate beyond three median absolute deviations (MADs) from the median. Cells identified as outliers were excluded from downstream analyses to ensure data quality. Doublet detection was performed with the DoubletFinder package (v2.0.3, RRID:SCR_018771)¹³, assuming an expected doublet rate of 10%, and clusters with more than 80% predicted doublets were removed. Cell cycle phase assignment was conducted using the “cyclone” function from the scanr package (v1.26.0, RRID:SCR_016944)¹⁴ with the mouse cell cycle marker dataset, categorizing cells into G1, S, and G2/M phases. These QC filters were applied using the SEURAT R package (v5, RRID:SCR_007322)¹⁵, and cells meeting all the criteria were considered high-confidence cells for downstream analysis. Raw sequencing data processing, mappings, and QC were performed at the CRCM Integrative Bioinformatics (Cibi) facility.

The gene expression matrices from the four individual samples (S34, S35, S37, S39) were normalized using the “NormalizeData” function in Seurat. The data were log-normalized, and the top 2,000 highly variable genes (HVGs) were identified using the “FindVariableFeatures” function. The data were then

scaled using the “ScaleData” function, with regression on the number of UMI counts per cell. To integrate the four samples, we used the canonical correlation analysis (CCA)-based integration method in Seurat. The “FindIntegrationAnchors” function was used to identify anchors between datasets, followed by the “IntegrateData” function to create a combined, batch-corrected dataset. The integrated dataset was then subjected to principal component analysis (PCA) to reduce the dimensionality of the data. The top 14 principal components (PCs) were selected based on the elbow plot and used for downstream analysis.

Clustering was performed using the Louvain algorithm, via the “FindClusters” function of Seurat with the resolution set to 1.2 (default 0.8) to achieve optimal delineation of transcriptionally distinct cell populations. Clusters were visualized using t-distributed stochastic neighbor embedding (t-SNE) with the RunTSNE function. Marker genes for each cluster were identified using the “FindAllMarkers” function in Seurat, applying the Wilcoxon rank-sum test. Genes with a log fold change greater than 1 and adjusted p-value < 0.05 were considered significant.

For cell type annotation, high-confidence tumor cells were identified using Single Cell Variational Aneuploidy analysis (SCEVAN)¹⁶, which detects copy number variations indicative of tumor cells. For annotating other cell types, particularly immune cells, we used the Annotation of Cell Types (ACT, RRID:SCR_009033)¹⁷ web server, a tool for manual cell type annotation based on known gene expression signatures. Stringent criteria for tumor cells may exclude cells with low-level alterations, which could explain the high proportion of epithelial cells. Each cluster was annotated by comparing the expression of canonical marker genes against reference signatures, ensuring accurate identification of cell types such as immune cells, fibroblasts, endothelial cells, and epithelial cells. A comprehensive annotation table was generated to map each cluster to a specific cell type, based on the identified signatures and expression of canonical marker genes. All visualizations, including t-SNE plots, feature plots, and violin plots, were generated using the Seurat R package and customized using ggplot2. Raw and processed data are available in the Gene Expression Omnibus (GEO, RRID:SCR_005012) under accession number GSE306716.

RNAseq data analysis.

For bulk RNAseq analysis, raw paired-end reads generated from four *MMTV-R26^{Met}* mouse tumors underwent quality assessment using FastQC. Transcript-level quantification was performed with Salmon (v1.10.2) in quasi-mapping mode with the *validateMappings* enabled to improve alignment accuracy. Quantification was carried out using a Salmon index generated from the mm10 mouse reference transcriptome. Salmon's automatic library-type detection and integrated bias-correction modules (sequence-specific, GC bias, and positional bias) were applied to optimize expression estimates, resulting in transcript-level TPM (transcript per million) and raw count values. Transcript abundances were imported into R and summarized to gene-level abundance using tximport (1.36.1) with length-scaled normalization to ensure compatibility with downstream statistical analyses. All samples were processed jointly using identical quantification parameters to ensure analytical consistency across the dataset.

Immune deconvolution and immune classification of tumors.

Immune cell composition was inferred using CIBERSORT through the deconvo_tme framework, applied separately to RNAseq and microarray datasets. For RNAseq samples, normalized expression matrices were scaled and processed with CIBERSORT using 500 permutations (arrays = FALSE), while microarray data were processed using CIBERSORT in microarray mode (arrays = TRUE, platform = "affymetrix"), ensuring compatibility with probe-level intensity distributions. Resulting cell-type fractions were combined with CIBERSORT quality-control metrics (correlation coefficient, P-value), and only samples with significant p-value < 0.05 were considered for downstream immune classification. Immune fractions were aggregated into four major lineage categories: neutrophils, macrophages, dendritic cells, and lymphocytes. Lineage scores were z-normalized across samples. Enrichment thresholds were defined as the 80th percentile of each lineage's z-score distribution, and tumors were assigned to the

immune class corresponding to the lineage with the strongest enrichment signal. Samples failing QC or lacking a dominant lineage signature were labelled as unclassified.

Statistical analyses.

All statistical analyses of experimental data were performed using GraphPad Prism software (RRID:SCR_002798). Details of the statistical tests applied and the corresponding P values are detailed in the figure legends. A P value > 0.05 was considered as non-significant (ns); P < 0.05 (*); P < 0.01 (**); P < 0.001 (***); P < 0.0001 (****). For two-sided comparisons, data with normal distribution were analyzed using an unpaired two-tailed Student's t-test, while non-parametric data were assessed using the Wilcoxon test. For multiple comparisons, ANOVA followed by Tukey's post hoc test was used for parametric data, whereas the Kruskal-Wallis test followed by Dunn's post hoc test was applied to non-parametric data. For in vivo experiments, tumor growth curves were compared using two-way ANOVA with repeated measures and appropriate post hoc tests. Kaplan–Meier survival analyses were conducted to evaluate overall survival or time-to-endpoint (tumor volume $\geq 1500 \text{ mm}^3$) across treatment groups. Differences between survival curves were assessed using the log-rank (Mantel–Cox) test.

For analyses involving human transcriptomic datasets, multiple visualization and analytical approaches were applied. Alluvial diagrams were generated using the *ggalluvial* package to visualize transitions between the immune classes defined in this study and major TNBC classifiers (PAM50, Lehmann TNBCtype-6, Bareche, Burstein2). Subtype assignments from each framework were merged into a unified long-format table to construct multi-stratum flow diagrams summarizing concordance and divergence across classification systems.

Heatmaps were generated with the *ComplexHeatmap* package to display lineage-level immune z-scores (neutrophils, macrophages, dendritic cells, lymphocytes), derived from aggregated and QC-filtered CIBERSORT output. Samples were hierarchically clustered, and immune class annotations were

overlaid to highlight concordance between computational immune classification and underlying immune composition.

Survival analyses were conducted to assess the prognostic relevance of the immune classes. Overall survival was evaluated using Kaplan-Meier estimates, and differences between groups were assessed using the log-rank test. Multivariable Cox proportional hazards models were fitted to determine whether immune class provided independent prognostic value after adjustment for key clinicopathological variables. Hazard ratios with 95% confidence intervals were summarized using forest plots, and proportional hazards assumptions were verified.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Correlation analyses between average gene expression in the *MMTV-R26^{Met}* mouse model and six independent TNBC patient cohorts. Scatterplots depict the correlation between average gene expression levels in *MMTV-R26^{Met}* tumors and human TNBC samples across the indicated immune-related gene signatures. Each panel displays the Pearson correlation coefficient (R) and corresponding p-value. A linear regression line is included to illustrate the best fit for each dataset. Rows correspond to the six independent human TNBC cohorts (GSE31448, GSE76124, GSE199633, GSE194040, GSE96058, TCGA), and columns represent the six immune pathways analyzed: allograft rejection, complement, IL6-JAK-STAT3 signaling, inflammatory response, interferon alpha response, and interferon gamma response.

Figure S2. Histological and molecular characterization of the *MMTV-R26^{Met}* primary tumors. A. Representative images of histopathological and immunohistochemical analysis of the *MMTV-R26^{Met}* tumors (n=8). Images show hematoxylin/eosin/saffron (HES) staining, as well as expression of estrogen- (ER), progesterone receptors (PR), and HER2. The histological subtype of each tumor is indicated on the left. Carcinoma NST: carcinoma of no Special Type. Note that the T35 tumor expresses HER2. The necrotic areas and keratinization are indicated by asterisks and arrowheads, respectively. The inset in T36 shows the presence of ghost cells. The HES staining images of T37, T38, and T39 are identical to those shown in Figure 2B. Ki67 immunostaining illustrates the proliferation status of the tumors, whereas CD45 immunolabeling reveals the presence of immune cells. NST: no special type. Scale bars: for HES, ER, PR, HER2: 350 μ m, and for Ki67 and CD45: 50 μ m. **B.** Sections of a mammary gland from a virgin female and a 4T-1 tumor were used as positive and negative controls for ER and PR immunostaining, respectively. HES staining is also shown.

Figure S3. Genomic and molecular characterization of *MMTV-R26^{Met}* primary tumors and proliferative index among human TNBC tumors. **A.** Copy number variations (CNVs) in the *ErbB2* genomic region were analyzed across the six *MMTV-R26^{Met}* primary tumors (T34–T40). The x-axis represents the chromosomal position along a portion of chromosome 17, whereas the y-axis shows the CNV log₂ ratio, with red and blue indicating copy number increases and decreases, respectively. No significant variation was observed in the *ErbB2* gene region, indicating stable copy number profiles across all samples. **B.** Immunostaining of *MMTV-R26^{Met}* primary tumors with anti-human MET antibodies against the C-terminal region, to detect the expression of the MET transgene. Scale bar: top panel: 100 µm, bottom panel: 25µm. **C.** Scatter plot illustrating the percentage of Ki67-positive cells in tumors from 103 TNBC patients from a cohort at the Institut Paoli Calmettes Hospital. Each dot represents a single tumor, with Ki67 positivity quantified by immunohistochemistry and plotted on the y-axis.

Figure S4. Immune cell composition in *MMTV-R26^{Met}* primary tumors compared to control mammary glands. **A.** Gating strategy employed for spectral cytometry studies. **B.** UMAP projection of CD45⁺ (dark red) and CD45⁻ (light red) cells in control mammary glands from age-matched non-tumor bearing mice. **C.** Proportions of immune cell types among CD45⁺ cells in tumors (T) and control mammary glands from tumor-free, age-matched mice (C) showing no significant differences between the groups. Each primary tumor is indicated with a specific color. **D.** Proportion of the specified DC types among all DCs. **E.** Scheme depicting all sample groups used for the flow cytometry analysis: tumors (T), contralateral mammary glands from tumor-bearing mice (CL), and mammary glands from tumor-free, age-matched mice (C). **F.** Proportions of immune cell types in the three sample groups, as defined above. **G.** Proportions of PD1⁺ cells among TCD8⁺ (top panel) and TCD4⁺ cells (bottom panel) in the three sample groups (C, CL, T). **H.** UMAP projection of 20,000 manually gated CD45⁺ cells from the eight *MMTV-R26^{Met}* tumors. Statistical analysis: Mann–Whitney test; *p < 0.05, **p < 0.01.

Figure S5. Immune infiltration patterns across BC subtypes and immune correlations in human TNBC patients. **A.** Box plot showing the absolute immune score (composite measure of immune cell infiltration) inferred by CIBERSORT across BC subtypes: TNBC, Luminal A (Lum_A), Luminal B (Lum_B), and HER2-enriched (HER2+). Statistical significance was assessed using t-test. **B.** Correlation heatmap of immune cell abundances and PD-1 expression inferred by CIBERSORT. Pearson correlation coefficients are color-scaled: blue = negative correlations; red = positive correlation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. **C.** Forest plot summarizing the multivariable Cox regression analysis evaluating the association between immune class and overall survival, adjusted for age, chemotherapy, lymph node status, and Ki67 expression. Hazard ratios (HR), 95% confidence intervals (CI), and p-values (p) are shown. The lymphocyte-enriched class served as the reference group. **D.** Elbow plot illustrating the proportion of variance explained across increasing values of k in k -means clustering (related to Figure 4D). The inflection point indicates the optimal number of clusters ($k=3$). **E.** Heatmap showing Pearson correlation coefficients comparing the percentage of PD-1⁺, CD3⁺, MPO⁺, and CD68⁺ cells between tumoral and stromal compartments. Blue = negative correlations; red = positive correlations. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure S6. Identification of four distinct immune clusters in human TNBC patients, survival association, and relationship with previously established TNBC classification, using the GSE199633 cohort. **A.** Heatmap showing the relative proportions of immune cell populations in TNBC patients (n=97) inferred by CIBERSORT. Hierarchical clustering was based on T cell abundance (Lympho-score), macrophages (Macro-score), neutrophils (Neutro-score), and DCs (DC-score). Row-scaled Z-score: blue = low abundance; red = high abundance. **B.** Kaplan-Meier analysis of overall survival of TNBC patients stratified by immune-enriched subgroups. Log-rank test: * $P < 0.05$; ** $P < 0.01$. **C.** Alluvial diagram illustrating the relationship between immune clusters and previously established TNBC molecular

classification (PAM50 intrinsic subtypes, LehmannTNBCtype-6, Bareche, and Burstein). Each ribbon represents an individual patient, and ribbon flows depict how samples classified under one system correspond to classifications in the others. **D.** Forest plot summarizing multivariable Cox regression evaluating the association of immune class with overall survival, adjusted for age, chemotherapy, radiotherapy, and type of surgery. Hazard ratios (HR), 95% confidence intervals (CI), and p-values (p) are shown. The lymphocyte-enriched tumors served as the reference group.

Figure S7. Identification of four distinct immune clusters in human TNBC patients, survival association, and relationship with previously established TNBC classification, using the GSE31448 cohort. **A.** Heatmap showing the relative proportions of immune cell populations in TNBC patients (n=101) inferred by CIBERSORT. Hierarchical clustering was based on T cell abundance (Lympho-score), macrophages (Macro-score), neutrophils (Neutro-score), and DCs (DC-score). Row-scaled Z-score: blue abundance = low; red = high abundance. **B.** Kaplan-Meier analysis of overall survival of TNBC patients stratified by immune-enriched subgroups. Log-rank test: *P < 0.05; ***P < 0.001. **C.** Alluvial diagram illustrating the relationship between immune clusters and previously established TNBC molecular classification (PAM50 intrinsic subtypes, LehmannTNBCtype-6, Bareche, and Burstein). Each ribbon represents an individual patient, and ribbon flows depict how samples classified under one system correspond to classifications in the others. **D.** Forest plot summarizing multivariable Cox regression evaluating the association of immune class with overall survival adjusted for age, Ki67, grade, and molecular subtypes. Hazard ratios (HR), 95% confidence intervals (CI), and p-values (p) are shown. Lymphocyte-enriched tumors served as the reference group.

Figure S8. Differential protein expression and clustering of *MMTV-R26^{Met}* primary tumors. **A.** Volcano plot displaying differentially expressed proteins in the eight *MMTV-R26^{Met}* primary tumors compared to controls. The x-axis shows the log₂FC in protein expression, and the y-axis shows the -log₁₀ adjusted

p-value (Wilcoxon test with Bonferroni correction). A threshold of $\pm 1.5 \log_2FC$ and an adjusted p-value > 0.05 was used to identify upregulated (red) or downregulated (blue) proteins in tumors. Grey points correspond to not significantly differentially expressed proteins. **B.** Heatmap and hierarchical clustering of *MMTV-R26^{Met}* primary tumors based on protein expression profiles. The heatmap depicts the Spearman correlation coefficient between samples, with color intensity indicating the strength of the correlation. Tumors are grouped into four clusters indicated at the bottom of the heatmap. **C.** FACS analysis of CD44/CD24 expression across the *MMTV-R26^{Met}* primary tumors. The bar graph depicts the distribution of CD44/CD24 phenotypes, indicating the percentage of CD45⁻ live cells expressing different combinations of both markers. **D.** Hierarchical clustering of *MMTV-R26^{Met}* tumors based on the abundance (expressed as percentage) of cell types with different combinations of CD24 and CD44 markers. Heatmap column colors: high (red) and low (white) abundance.

Figure S9. Genomic landscape of mutations in human TNBC and *MMTV-R26^{Met}* primary tumors. A. Scatter plot depicting the mutation frequencies (expressed as percentages) of various genes commonly mutated in TNBC and non-TNBC patients, based on TCGA database. Oncogenes are marked in orange, tumor suppressors in green, and genes with dual roles in purple. This annotation has been done using the OncoKB™ database. **B.** Bar chart depicting the classification of detected variants in *MMTV-R26^{Met}* tumors, including missense, frameshift (insertion and deletion), and nonsense mutations. **C.** Variant type distribution, highlighting the predominance of single nucleotide polymorphisms (SNP) and insertions/deletions (indels, INS) in the *MMTV-R26^{Met}* dataset. **D.** Mutation spectrum analysis showing the percentage of different nucleotide substitutions (C>T, T>C, C>A, etc.) on the left, and the relative proportions of transitions (Ti) versus transversions (Tv) on the right. **E.** Pathway analysis displaying the fraction of affected pathways (left panel) and of affected samples (right panel) for the indicated key signaling pathways. **F.** Distribution of PAM50+ Claudin-low molecular subtypes within different $\log_2$ expression levels of the *KAT6B* gene (ranging from low to high) in the METABRIC dataset. Results indicate that low *KAT6B* expression is associated with the Claudin-low subtype. **G.** PTEN protein levels

expressed as log₂ LFQ (Label-Free Quantification) intensity for each tumor determined through mass spectrometry analysis. "ND" indicates that PTEN was not detected in the corresponding sample. **H.** Top 10 amplified and deleted cancer-related genes (filtered using OncoKB™ database). The bar plots show the log₂ ratio CNV for each gene, with amplified (in red) and deleted (in blue) genes. **I.** Scatter plot showing the log₂ CNV ratio for *Met* in *MMTV-R26^{Met}* primary tumors relative to controls, highlighting amplification (AMP, red) or deletion (DEL, blue). **J.** CNVs for *Tpr53* (*Trp53*) and several of its effector genes across *MMTV-R26^{Met}* primary tumors.

Figure S10. Immunohistochemical analysis of *MMTV-R26^{Met}* primary and syngeneic tumors. A-B. IHC staining for HER2, ER, PR (A), and Ki67 (B) in the primary *MMTV-R26^{Met}* tumors and associated syngeneic tumors after 4 (P4) and 7 passages (P7). Representative images of T34, T35, T37, and T39 primary tumors are the same as in Figure S2A. Image size: 1000x1000 μm.

Figure S11. Maintenance of molecular and immune infiltration characteristics in *MMTV-R26^{Met}* syngeneic tumors across passages. A. FACS analysis of CD44/CD24 expression in syngeneic tumors (S) across different passages (P), with data from the corresponding primary tumor (Prim). The bar graphs depict the distribution of CD44/CD24 phenotypes, representing the percentage of CD45⁻ live cells expressing different combinations of CD44 and CD24 markers. **B.** Heatmap of protein expression levels derived from mass spectrometry analysis of CD45-negative cells from primary (T) and their corresponding syngeneic (S) tumors after 4 passages. Note the high similarity between T39 and its corresponding syngeneic S39 tumors, highlighted by the yellow box. Protein expression levels are represented as row z-scores, with rows corresponding to individual proteins and columns to different samples. **C.** Immunostaining using anti-CD45 antibodies on primary *MMTV-R26^{Met}* tumors and their corresponding syngeneic tumors at passages 4 (P4) and 7 (P7). Representative images of primary tumors T34, T35, T37, and T39 are reused from Figure S2. Scale bar: 50 μm. **D.** FACS analysis of the

percentage of CD45⁺ live cells in syngeneic tumors (S) across different passages (P) with data from the corresponding primary tumors (Prim). **E.** Immune infiltrate composition in *MMTV-R26^{Met}* tumors derived from either freshly dissociated (F) or cryopreserved (C) cells. Proportional distribution of major immune cell subsets among live CD45⁺ cells in four representative syngeneic tumors (S34, S35, S37, S39).

Figure S12. Clustering and cell type annotation based on scRNA-seq of *MMTV-R26^{Met}* syngeneic tumors. **A.** t-SNE plot of integrated scRNA-seq data from the indicated tumors. Cells are color-coded according to the sample of origin, showing an efficient integration from different tumors. **B.** t-SNE plot displaying clusters identified at a resolution of 1.2, with cells color-coded by their respective cluster assignments. A total of 29 clusters are identified. **C.** t-SNE plot (top) showing the distribution of high-confidence normal and tumor cells, along with cells that were filtered out due to insufficient features detected by the malignant cell inference algorithm (SCEVAN), across the integrated dataset. The bar graph (bottom) quantifies the percentage of tumor cells present in each cluster, revealing that certain clusters are predominantly composed of tumor cells. **D.** Feature (top) and violin (bottom) plots showing the expression levels of the indicated markers used to identify major cell types in the dataset: *Ptprc* (CD45, immune cell marker), *Epcam* (epithelial cell marker), *Pecam1* (endothelial cell marker), *Dcn* (Decorin, fibroblast marker). The violin plots further illustrate the expression distribution of these markers across the identified clusters, confirming the cell type identity of each cluster. **E.** Table summarizing the annotation of each cluster based on marker gene expression. The table links each cluster number to a specific cell type indicated in the table. **F.** Dot plot showing the expression of stromal cell markers across the indicated different stromal cell types identified in the dataset. The size of the dots indicates the percentage of cells expressing each marker, whereas the color intensity represents the average expression level.

Figure S13. Immunostaining of S37 and S39 tumors and mice body weight monitoring. A. Immunostaining using anti-Ly6G, F4/80 and PD-1 antibodies on S37 and S39 tumors from the biobank. Representative images. Scale bar: 100 μ m. **B-C.** Body weight evolution of S37 (B) and S39 (C) tumor-bearing mice throughout the procedure. No significant changes were observed, indicating that the drug doses were well tolerated.

SUPPLEMENTARY TABLE LEGENDS

Table S1. *MMTV-R26^{Met}* tumor samples used for bulk RNAseq analysis.

Table S2. Available human TNBC cohorts used in this study.

Table S3. Histopathological characterization of *MMTV-R26^{Met}* tumors (primary and syngeneic).

Table S4. Clinico-pathological characteristics of the TNBC samples belonging to the GSE96058 cohort.

Table S5. TMA (Tumor microarray) analysis

Table S6. Specific gene mutations identified in *MMTV-R26^{Met}* primary tumors. List of mutations detected in *MMTV-R26^{Met}* primary tumors, their predicted functional relevance (according to OncoKB™ and Cancer Hotspots databases), the tumors in which they were identified, and supporting literature references^{18–30}.

Table S7. Antibodies used for immunohistochemistry.

Table S8. Antibodies used for spectral cytometry analysis.

SUPPLEMENTARY REFERENCES

1. Lamballe, F. *et al.* Modeling Heterogeneity of Triple-Negative Breast Cancer Uncovers a Novel Combinatorial Treatment Overcoming Primary Drug Resistance. *Adv. Sci.* **8**, 2003049 (2021).
2. Denkert, C. *et al.* Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28**, 105–113 (2010).
3. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
4. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* **109**, 647–668 (2022).
5. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
6. Frankenfield, A. M., Ni, J., Ahmed, M. & Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* **21**, 2104–2113 (2022).
7. Gerault, M.-A., Camoin, L. & Granjeaud, S. DIAgui: a Shiny application to process the output from DIA-NN. *Bioinforma. Adv.* **4**, vbae001 (2024).
8. Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol. Biol. Clifton NJ* **1711**, 133–148 (2018).
9. Brueffer, C. *et al.* Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precis. Oncol.* **2**, PO.17.00135 (2018).
10. Asleh, K. *et al.* Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. *Nat. Commun.* **13**, 896 (2022).

11. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
12. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
13. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329–337.e4 (2019).
14. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
15. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
16. De Falco, A., Caruso, F., Su, X.-D., Iavarone, A. & Ceccarelli, M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat. Commun.* **14**, 1074 (2023).
17. Quan, F. *et al.* Annotation of cell types (ACT): a convenient web server for cell type annotation. *Genome Med.* **15**, 91 (2023).
18. Adamus, A. *et al.* GCSH antisense regulation determines breast cancer cells' viability. *Sci. Rep.* **8**, 15399 (2018).
19. An, H.-X., Beckmann, M. W., Reifemberger, G., Bender, H. G. & Niederacher, D. Gene Amplification and Overexpression of CDK4 in Sporadic Breast Carcinomas Is Associated with High Tumor Cell Proliferation. *Am. J. Pathol.* **154**, 113–118 (1999).
20. Yang, C. *et al.* Acquired CDK6 amplification promotes breast cancer resistance to CDK4/6 inhibitors and loss of ER signaling and dependence. *Oncogene* **36**, 2255–2264 (2017).
21. Hayes, M. J., Thomas, D., Emmons, A., Giordano, T. J. & Kleer, C. G. Genetic Changes of Wnt Pathway Genes Are Common Events in Metaplastic Carcinomas of the Breast. *Clin. Cancer Res.* **14**, 4038–4044 (2008).

22. Légaré, S. *et al.* The Estrogen Receptor Cofactor SPEN Functions as a Tumor Suppressor and Candidate Biomarker of Drug Responsiveness in Hormone-Dependent Breast Cancers. *Cancer Res.* **75**, 4351–4363 (2015).
23. Karlin, K. L. *et al.* The Oncogenic STP Axis Promotes Triple-Negative Breast Cancer via Degradation of the REST Tumor Suppressor. *Cell Rep.* **9**, 1318–1332 (2014).
24. Wagoner, M. P. *et al.* The transcription factor REST is lost in aggressive breast cancer. *PLoS Genet.* **6**, 1–12 (2010).
25. Furth, N. *et al.* LATS1 and LATS2 suppress breast cancer progression by maintaining cell identity and metabolic state. *Life Sci. Alliance* **1**, e201800171 (2018).
26. Tinsley, E., Bredin, P., Toomey, S., Hennessy, B. T. & Furney, S. J. KMT2C and KMT2D aberrations in breast cancer. *Trends Cancer* **10**, 519–530 (2024).
27. Vellichirammal, N. N. *et al.* The mutational landscape of a US Midwestern breast cancer cohort reveals subtype-specific cancer drivers and prognostic markers. *Hum. Genomics* **17**, 64 (2023).
28. Dreyer, C. A., VanderVorst, K., Free, S., Rowson-Hodel, A. & Carraway, K. L. The role of membrane mucin MUC4 in breast cancer metastasis. *Endocr. Relat. Cancer* **29**, R17–R32 (2021).
29. Guha, M. *et al.* Aggressive triple negative breast cancers have unique molecular signature on the basis of mitochondrial genetic and functional defects. *Biochim. Biophys. Acta Mol. Basis Dis.* **1864**, 1060–1071 (2018).
30. Laia, S.-R. *et al.* KAT6B Is a Tumor Suppressor Histone H3 Lysine 23 Acetyltransferase Undergoing Genomic Loss in Small Cell Lung Cancer. *Cancer Res.* **75**, 3936–3945 (2015).