Integrated molecular portraits of breast cancer

Yi-Zhou Jiang (yizhoujiang@fudan.edu.cn)
Ding Ma
Xi Jin
Yi Xiao
Ying Yu
Jinxiu Shi
Yi-Fan Zhou
Tong Fu
Cai-Jin Lin
Lei-Jie Dai
Cheng-Lin Liu
Shen Zhao
Guan-Hua Su
Wanwan Hou
Yaqing Liu
Qingwang Chen
Jingcheng Yang
Naixin Zhang
Wen-Juan Zhang
Wei Liu
Weigang Ge
Wen-Tao Yang
Chao You
Yajia Gu
Virginia Kaklamani
François Bertucci
Claire Verschraegen
Anneleen Daemen
Nakul M. Shah
Ting Wang
Tiannan Guo
Leming Shi
Charles M. Perou
Yuanting Zheng (zhengyuanting@fudan.edu.cn)
Wei Huang (huangwei@chgc.sh.cn)
Zhi-Ming Shao (zhimin_shao@yeah.net)
Abstract

The associated article reports multiomic analysis of breast cancers, where we provided a unique public resource that contributes novel insights into the breast cancer, holding promise for further precision treatment approaches. This protocol not only delineates detailed experimental procedures for sample preparation and data generation but also comprehensive analysis methods for genomics, transcriptomics, proteomics, and metabolomics from breast cancer samples. Furthermore, regarding radiomics and pathomics, the protocol describes the extraction of both radiological and pathological features in detail.

Introduction

Reagents

Equipment

Procedure

Evaluation of pathological indicators

Based on the pathological whole slide images (WSIs) that we collected, the stromal tumor-infiltrating lymphocytes (sTILs) were evaluated. According to the recommendations by the International TILs Working Group 1, sTILs were defined as lymphocytes located in the stroma between the tumor nests. The pathological feature underwent independent evaluation by three experienced pathologists who were blinded to the clinical information of the patients. Any disparities among them were resolved through discussion and subsequent consensus.

Sample processing for genomic DNA and total RNA extraction

For the purpose of quality control, we macrodissected fresh frozen tumor tissues and filtered out samples full of stromal tissues (>50% stromal tissue). The presence of tumor cells at a confirmed threshold of 50% or higher was observed in all breast cancer specimens. Genomic DNA was isolated from fresh frozen tissue samples and peripheral blood cells employing the TGuide M24 (Tiangen, Beijing, China). The purity and concentration of the genomic DNA were evaluated by gauging the absorbance at 260 nm (A260) and 280 nm (A280) through a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The extracted DNA was deemed of sufficient purity for subsequent experiments when the A260/A280
ratio fell within the range of 1.6-1.9. Furthermore, total RNA was isolated from tissues preserved in RNAlater solution, utilizing the miRNeasy Mini Kit (Qiagen #217004) following the manufacturer's protocol. The assessment of RNA integrity was performed using an Agilent 4200 Bioanalyzer with RNA ScreenTape (Agilent Inc.), while concentrations were determined using a NanoDrop ND-8000 spectrophotometer (Thermo Fisher Scientific Inc.).

Sample preparation and data generation for RNA sequencing

Libraries were constructed by ribosomal RNA depletion methods. Ribosomal RNA was depleted using a Ribo-off rRNA Depletion Kit (H/M/R) (Vazyme #N406, Vazyme Biotech Co., Ltd., Nanjing, China), and RNA libraries were constructed using a VAHTS Universal V8 RNA-seq Library Prep Kit for Illumina (Vazyme #NR605, Vazyme Biotech Co., Ltd., Nanjing, China). Specifically, the fragmented RNAs were reverse-transcribed into cDNA, and then 3-terminal poly(A) modification was completed. Subsequently, adapters were added to the cDNA. PCR was then used to amplify the libraries. During quality control (QC) of the libraries, Qubit 4.0 (Thermo Fisher Scientific Inc.) and Agilent 2200 Bioanalyzer (Agilent Inc.) were used to detect the concentration and fragment size distribution of the libraries, respectively. The sequencing of the libraries was conducted on Illumina NovaSeq platforms with paired-end reads of 150 bp.

The initial raw Illumina sequence data were demultiplexed and transformed into FASTQ files, subsequently undergoing quantification to identify and eliminate adapter sequences and sequences of low quality. Reference human genome build 38 (https://genome-idx.s3.amazonaws.com/hisat/grch38_snptran.tar.gz) and gene model from Ensembl (http://ftp.ensembl.org/pub/release-93/gtf/homo_sapiens/Homo_sapiens.GRCh38.93.gtf.gz) were used for read mapping and gene quantification. Sample reads were aligned to the hg38 human genome reference through the employment of HISAT2 (v2.1). Subsequently, utilizing StringTie (v1.3.4) and Ballgown (v2.14.149), fragments per kilobase of transcript per million mapped reads (FPKM) were computed. To ensure precision in gene expression values, genes with an FPKM of 0 in over 30% of the samples were excluded from subsequent analyses.

RNA fusion detection

Two callers, STAR-Fusion and Arriba, were employed to call consensus fusion/chimeric events in our samples. The findings from each tool, utilizing both tumor and normal RNA-Seq data, were consolidated into a unified file, followed by meticulous filtering. For an event to be considered a potential RNA fusion, it needed to be detected by both fusion detection tools in at least one sample. We then removed fusions following filtering steps: 1) fusions involving noncoding genes were filtered; 2) fusions detected in normal samples were filtered.
Deconvolution of TME cells from scRNA-seq

We used the large-scale single-cell RNAseq in breast cancer (GSE176078) to deconvolute the bulk transcript profiling by the CIBERSORTx and obtain the proportions of 46 cell subsets, including tumor, stromal, immune, and other multiple cell types. Briefly, the annotation of cells in reference single-cell count matrix followed the cell type identified by Wu et al and all cells were included in our analysis. We then ran CIBERSORTx “cibersortx/fractions” with the following parameters: --single_cell TRUE --G.min 300 --G.max 500 --q.value 0.01 --filter FALSE --k.max 999 --replicates 5 --sampling 0.5 --fraction 0.75 --rmbatchSmode TRUE.

T-cell (TCR) and B-cell receptor (BCR) estimation and calculation of clonality and diversity

To identify the CDR3 sequences of TCRs and BCRs, MiXCR (v3.0.13) was used with optimized parameters to extract repertoires from RNA-seq data. Briefly, paired-end fastq files were analyzed by the “analyze shotgun” command with “--species hs” and “--starting-material rna” and were exported for TCR α (TRA) and β (TRB) and B-cell heavy (IGH) and light (IGL) chain clonotypes. Four chain types of immune receptors (TRA, TRB, IGH, and IGL) inferred by MixCR were imported into VDJtools (v1.2.1) for conversion and statistical analysis. Clonal diversity was calculated through Shannon entropy, and clonality was defined as 1-(Shannon entropy/log2(unique productive clone)) to reflect the clone expansion of T and B cells.

HLA and neoantigen analysis

For neoantigen predictions, POLYSOLVER (v4.1) was used to determine the 4-digit HLA type and mutations in class I HLA genes for each patient. Then, the somatic nonsynonymous SNV and INDEL mutations detected in each sample were used to detect all possible 9-, 10- and 11-mer mutant peptides. Next, we used NetMHCpan (v4.1a) to predict the binding affinities of peptides and POLYSOLVER-inferred HLA alleles. Neoantigens were defined as mutant peptides with IC$_{50}$ < 500 nM and mean mRNA expression>1 (FPKM). HLA alleles were called by POLYSOLVER, and loss of heterozygosity of the HLA alleles (LOHHLA) was analyzed by LOHHLA tools. Tumor and germline BAM files, patient-specific HLA calls, HLA FASTA files, purity and ploidy estimates were used as inputs to LOHHLA. HLA LOH was determined when one of the two alleles of the HLA gene showed a copy number < 1 and p value < 0.05. HLA allele imbalance was defined when two alleles of the HLA gene showed a p value < 0.05.

Sample preparation and data generation for copy number alterations
The OncoScan CNV Assay Kit (Affymetrix, Santa Clara, CA, USA) was used to perform genome-wide copy number analysis according to the manufacturer’s protocol. In summary, 80 ng of DNA from each tumor sample underwent processing. Molecular inversion probes (MIPs) were combined with the sample DNA and annealed at 58°C overnight. The resulting annealed DNA was split into two equal portions and subjected to incubation with AT or GC gap-fill master mixes for ligation. Then, exonuclease treatment was performed to remove the unincorporated, noncircularized MIPs and the remaining genomic template. The process involved the linearization of circularized MIPs using a cleavage enzyme, followed by successive PCR amplifications. The resulting amplified products underwent digestion with HaeIII and Exo enzymes, enabling the isolation of small fragments containing the specific single-nucleotide polymorphism (SNP) genotype, which were subsequently hybridized onto arrays. These arrays underwent a washing and staining procedure facilitated by a GeneChip Fluidics Station 450 (Affymetrix, Santa Clara, CA, USA) and were scanned utilizing a GeneChip Scanner 3000 7G (Affymetrix, Santa Clara, CA, USA). The fluorescence of clusters was measured to produce a DAT file. The values for cluster intensity were automatically computed from DAT files using the GeneChip Command Console software (Affymetrix, Santa Clara, CA, USA), generating a CEL file.

Analysis of SNP array data

An analysis of Affymetrix OncoScan CNV SNP probe assays was performed with Chromosome Analysis Suite (ChAS) v4.1 software (Thermo Fisher Scientific). A copy number reference model file was built by using a reference cohort of DNA from 23 randomly selected white blood cell samples from the mentioned patients and positive control samples from the OncoScan CNV Assay Kit. Probe-level output obtained from the ChAS underwent analysis using ASCAT (v2.4.3) to derive segmented copy number predictions, estimated tumor ploidy and estimated tumor purity results. The ASCAT segments were employed to calculate log2 ratios, achieved by dividing by the total copy number (nAraw + nBraw) with adjustments made for zero values, set to 0.05. These segments were served as the input of GISTIC2.0 (v2.0.22) to investigate the recurrence of copy number variations (CNVs) at the gene level within our sample set. GISTIC2.0 was executed with the following parameters with specific parameters altered from the default settings (-ta 0.2 -td 0.2 -genegistic 1 -smallmem 1 -broad 1 -conf 0.95 -rx 0 –brlen 0.7 -cap 3.5 –armpeel 1 -js 100). Furthermore, a set of adjacent normal tissues from 23 patients was utilized to filter recurrent germline or potential false-positive calls. Based on their segment output, the probes that suggested gain or loss in at least five patients were used with the help of Integrated Genomics Viewer to constitute a CNV file for removing recurrent germline/potential false-positive calls in GISTIC2.0.

Genomic data analysis of whole-exome sequencing data

In the course of our investigation, reads procured from exome sequencing were aligned utilizing the BWA-mem algorithm, with the subsequent BAM files undergoing an essential preprocessing phase with
duplicate marking and base quality score recalibration using version 202010.02 of Sentieon Genomics tools. Sequencing quality was assessed using NGSCheckMate (v1.0.0), FastQ Screen (v0.12.0), FastQC (v0.11.8) and Qualimap (v2.0.0).

**Somatic variant calling**

The identification of somatic mutations employed computational tools such as VarScan2 v2.4.2, with specific parameters detailed (--min-coverage 3 --min-coverage-normal 3 --min-coverage-tumor 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1), along with TNseq and TNscope. For the preliminary outputs from VarScan2, we engaged the processSomatic and somaticFilter modules (--min-coverage 10 --min-reads2 2 2 --min-strands2 1 --min-avg-qual 20 --p-value 0.1) to curate high-fidelity somatic mutations while concurrently mitigating clusters of inaccuracies and SNV calls proximate to indels. TNseq detected and filtered out variants by TNhaplotyper2 (--germline_vcf --pon --algo OrientationBias and --algo ContaminationModel) and TNfilter (--contamination --tumor_segments --orientation_priors). It's noteworthy that both TNseq and TNscope incorporated a panel of normal (PoN) samples, aiming to distinguish anticipated germline variations and potential artifacts. This PoN construct was predicated upon an extensive dataset of 699 normal blood specimens, subsequently yielding two VCF files reflective of mutation sites as discerned by TNseq and TNscope. Moreover, the positional data of the population's germline resource, inclusive of allele frequencies sourced from gnomAD, served a pivotal role in refining the preliminary TNseq findings.

In curating the definitive compilation of variant determinations, a structured biphasic strategy was employed. The initial phase was centered around the elimination of variant calls potentially stemming from sequencing aberrations. This was succeeded by harnessing mutations that achieved consensus across a minimum of two of the three delineated tools for the accurate pinpointing of somatic mutations. Subsequently, a refined filtering mechanism, anchored by bam-readcount v0.8.0 (accessible at https://github.com/genome/bam-readcount), was instituted to mitigate inaccuracies, adhering to set criteria: 1) namely, a variant allele frequency (VAF) not less than 5%; 2) an imposed sequencing depth within the demarcated zone of a minimum of 8; and 3) sequence reads substantiating the variant declaration to be no less than 4. A catalog of cancer driver genes was assembled with: (1) the curated cancer gene list by OncoKB (October 2020); (2) previously published and functionally validated oncogenic driver genes by Bailey, et al; (3) the compendium of mutational cancer driver genes from Integrated OncoGenomics; and (4) genes recorded as oncogenes or tumor suppressor genes (TSGs) by the Cancer Gene Census.

**Germline variant calling**
Pindel v0.2.5b9_3 \(^{21}\) (\(-c\; all\; -x\; 4\; -l\; 0\; -M\; 3\; -J\; hg38_ucsc_centromere.bed\)) and Sentieon DNAseq Haplotyper v202010.02 \(^{9}\) with default parameters were used to identify germline mutations. Only variants with high confidence according to the following criteria were retained: 1) for SNVs, at least 20x coverage, sequencing depth \(\geq\) 5 in the region for the alternative allele, and 20% VAF were needed; 2) for indels, identified by both Haplotyper and Pindel, or Pindel-unique calls with high confidence (at least 30x coverage and 20% VAF).

All the above somatic and germline variant calls were then annotated using both ANNOVAR v24.10.2019 \(^{22}\) and the Ensembl variant effect predictor (VEP v104.0) \(^{23}\).

**MS sample processing and data collection for proteomics**

Proteins were extracted and subjected to denaturation using 30 \(\mu\)L of a lysis buffer composed of 6 M urea, 2 M thiourea, and 100 mM triethylammonium bicarbonate, utilizing 1-2 mg of freshly frozen tissue samples as the substrate, followed by proteolytic digestion using Lys-C (Hualishi, Beijing, China) and trypsin (Hualishi, Beijing, China) assisted by pressure-cycling technology (PCT) \(^{24,25}\). TMTpro 16plex label reagents (Thermo Fisher Scientific, San Jose, USA) were used to label the digested peptides. A common pooled peptide sample was employed as the reference control sample for each TMT batch. The TMT-labelled samples were cleaned with a C18 column (Waters Sep-Pak® Vac 1 cc C18 Cartridge) and fractionated using a Dionex UltiMate3000 HPLC system (Thermo Fisher Scientific, San Jose, USA) \(^{26}\). Peptides were divided into 60 segments and then combined into 30 segments. The peptides, once redissolved, underwent analysis using an liquid chromatography–tandem mass spectrometry (LC–MS/MS) approach via a nanoflow DIONEX UltiMate 3000 RSLCnano System from Thermo Scientific™ (San Jose, USA). This was linked to an Orbitrap Exploris 480 mass spectrometer from the same manufacturer, and it featured a FAIMS Pro™ in DDA mode. The peptide samples were analyzed using an LC gradient of 60 min. The other LC–MS parameters followed a previous publication \(^{26}\): Instrument performance assessment was conducted using a mouse liver digest, while to prevent any carry-over, blank water samples (designated as buffer A) were injected every 4 runs.

**Database search for proteomics**

The mass spectrometric (MS) data were analyzed by Proteome Discoverer (Version 2.4.1.15, Thermo Fisher) using the human protein database downloaded from UniProt (version 15/07/2020, 20368). Two missed trypsin cleavages were allowed. The minimal peptide length was set to be at least 6 residues long. The total peptide quantity was used for normalization. Carbamidomethylation (+57.021 Da) of cysteine was set as a static modification, while oxidation (+15.995 Da) of methionine and acetylation (+42.011 Da) of peptide N-termini were set as variable modifications. Lysine residues and peptide N-termini were tagged with TMTpro (+304.207 Da). The precursor ion mass tolerance was established to 10
ppm, with a corresponding fragment mass tolerance of 0.02 Da. Both unique peptides and razor peptides were used for mapping the best associated master proteins. The master protein abundances were calculated by summation of their associated peptide groups. The false discovery rate (FDR) of the peptide was set to 1% (strict) and 5% (relaxed). The other parameters followed the default setup\(^{27}\). In total, 10864 genes were identified.

### Normalization and quality control of proteomic data

With the primary matrix (samples in columns and proteins in rows) derived from Proteome Discoverer, we performed a series of quality control process to guarantee the high quality and reliability of samples for subsequent analyzes. First, samples with outlier median intensity ratio \((\text{defined as } >\text{mean(intensity ratio)} + 2*\text{IQR(intensity ratio)} \text{ or } <\text{mean(intensity ratio)} - 2*\text{IQR(intensity ratio)})\) were excluded by tumor samples and para-tumor samples respectively. Then, the matrix was log2-transformed and then normalized by column-median. Further, proteins that were absent in over 30% of all remaining samples were excluded. Batch effects were mitigated through the utilization of the "removeBatchEffect" function from the \textit{limma} R package\(^ {28}\). Then effect of batch effect removal were evaluated by principal component analysis (PCA). Then samples were filtered by PCA, where sample out of the 90% confidence eclipses in PCA analysis (plotted with PC1 and PC2 using all included proteins) by tumor samples and para-tumor samples respectively. Finally, duplicated samples and genes were merge by using mean value. We set two batches for replicates to test the reproducitivy of our profiling methods, in which samples showed good correspondence both within the same batch and across different batches.

### Sample selection for metabolomic detection

After sampling for DNA and RNA sequencing, 453 of the remaining samples contained sufficient quantities for metabolic detection, while only 37 (8.2\%) of them were TNBCs. Therefore, we added an additional 58 TNBC samples from our sample bank for metabolic detection with 453 previously collected samples. The PAM50 subtypes of the additional 58 samples were provided in our previous study\(^ {29}\).

### Types of metabolomic detection

In order to acquire a comprehensive landscape of metabolomics, we conducted two types of metabolomic detection: polar metabolites (hydrophilic metabolites) and lipids (lipophilic metabolites)\(^ {30,31}\). These two types metabolomic detection are different in sample processing, detecting methods and data collection. Details of the metabolomic process are listed below.
**MS sample processing and data collection for polar metabolomics**

Quenching and extraction of the sample: An EP tube was filled with twenty-five milligrams of the given sample, followed by the addition of 500 μL extraction solution composed of methanol, acetonitrile, and water in a 2:2:1 ratio. The contents were blended at 35 Hz for 4 minutes and then subjected to sonication in an ice-water bath for 5 minutes. This sequence of blending and sonication was performed thrice. Afterwards, the samples underwent a 1-hour incubation at -40°C, and then they were spun at 13800 (×g) (12000 rpm, R= 8.6 cm) for a span of 15 minutes at a temperature of 4 °C. For QC sample preparation, equivalent portions of the supernatants from each sample were combined.

Chromatography separation: An ultrahigh-performance liquid chromatography (UHPLC) instrument (Vanquish, Thermo Fisher Scientific) linked to a Q Exactive High Field (QE-HFX) mass spectrometer (Orbitrap MS, Thermo) was used for the LC–MS/MS analysis. This utilized a UPLC BEH Amide column with dimensions of 2.1 mm x 100 mm. The mobile phase consisted of 25 mmol/L ammonium acetate and 25 mmol/L ammonia hydroxide in water (pH set at 9.75) (A), and acetonitrile as (B). Samples were injected in 2 μL volumes with the autosampler maintained at a temperature of 4 °C.

Mass spectrometry: We used the Xcalibur software (Thermo) and the QE HFX mass spectrometer to obtain MS and MS/MS spectra. The software constantly reviews the full-scan MS spectrum. The set parameters for the electrospray ionization (ESI) source were: sheath gas at 30 Arb flow rate, Aux gas at 25 Arb, capillary maintained at 350 °C, a full MS resolution of 60,000 and an MS/MS resolution of 7,500, with a collision energy setting of 10/30/60 in the NCE mode. The spray voltages were either 3.6 kV (positive) or -3.2 kV (negative).

Data QC: Internal standards (ISs) and QC samples were used to evaluate instrument variability throughout the process of detection. The median relative standard deviation (RSD) was derived from the ISs introduced to each sample ahead of their injection into the mass spectrometers, serving as a measure of instrument variability. The median RSD value of the ISs in our study was 4.75%.

For QC samples, as described before, an equal volume (10 μL) of each sample was extracted and mixed together. Throughout the detection process, the QC replicate samples were injected every eight samples, and these QC samples were treated independently throughout the process as if they were client study samples. Then, the distributions of QC samples in PCA were plotted to evaluate the instrument and process variability.

Data processing, metabolite identification and data analysis: Raw MS data was transformed into the mzXML format using the ProteoWizard tool (version 3.0.19282) and the R package XCMS (v3.2) was employed for metabolomic processing. This processing encompassed tasks such as peak identification, peak alignment, extracting peak data, correcting the retention time, and integrating peaks. To ensure consistent and reliable metabolomics data, any peaks with an RSD exceeding 30% in the QC samples were eliminated. The peaks that remained were labeled by matching their retention time and mass-to-charge (m/z) ratios with indices in a library, using the R package CAMERA. Subsequently, we acquired
a data matrix that displayed the retention time, m/z, and peak intensities. Peaks that were absent (showing an intensity of 0) in over half the samples were discarded from this matrix. For each metabolite, the missing values were replaced with 50% of the lowest observed value of all detected samples\textsuperscript{34,35}. The area of each peak was then normalized by isotopically labelled ISs for polar metabolomics\textsuperscript{36}. To eradicate any undesired variations within and between batches, every metabolite peak across samples was adjusted employing the locally estimated scatterplot smoothing (LOESS) technique, centered on QC samples\textsuperscript{36}. In brief, a LOESS regression model was crafted, relying on the intensity changes of each metabolite within QC samples. This model then forecasted and amended the intensities for the same metabolite in the subject samples\textsuperscript{36}. In all, 11708 MS features were included for further annotation.

The MS/MS spectra were subsequently examined in an in-house database for polar metabolite annotation according to accurate mass (m/z, ± 5 ppm), retention time, and spectral patterns. A matching score for the MS/MS spectra was computed using the dot-product algorithm, factoring in both fragment information and intensity values\textsuperscript{37}. Metabolites with MS/MS matching scores higher than 0.3 were included in our study. In all, 669 MS/MS features were annotated and included in our study.

In summary, only the peaks with MS/MS names and with MS/MS matching scores higher than 0.3 were included for further analysis.

**MS sample processing and data collection for lipidomics**

Sample quenching and extraction: An EP tube was loaded with twenty milligrams of the selected sample. Sequentially, 200 \( \mu \)L of water followed by 480 \( \mu \)L of the extraction solution (MTBE: MeOH in a 5:1 ratio) were introduced. The mixtures were vortexed for a short 30-second burst, then subjected to a homogenization at 35 Hz for 4 minutes, followed by 5 minutes of sonication in a cold water-ice bath. This cycle of homogenization and sonication was performed a total of three times. Subsequently, the mixtures were allowed to incubate at a chilly -40 °C for an hour, then spun at 900 (\( \times \)g) (3000 rpm, R= 8.6 cm) for 15 min at 4 °C. A portion of 300 \( \mu \)L of the clear supernatant was moved to a new tube. For QC sample preparation, identical portions of supernatant from all samples were amalgamated and the combined solution was dried using a vacuum concentrator at 37 °C. The resultant dry residue was then dissolved in 150 \( \mu \)L of a 50% methanol-dichloromethane mix, followed by 10 minutes of sonication in an ice-water bath. Afterward, the solution was subjected to centrifugation at a speed of 16200 (\( \times \)g) (13000 rpm, R= 8.6 cm) for 15 minutes at 4 °C. A volume of 120 \( \mu \)L of supernatant was carefully transferred into a new glass container, ready for LC/MS analysis.

Chromatography separation: For lipidomics data collection, LC–MS/MS analyzes were executed on a UHPLC system (1290, Agilent Technologies) that was outfitted with a Kinetex C18 column (dimensions: 2.1 x 100 mm, 1.7 \( \mu \)m granularity, from Phenomen). The mobile phase A was a blend of 40% water, 60% acetonitrile, and 10 mmol/L ammonium formate. The mobile phase B comprised 10% acetonitrile and
90% isopropanol, with the addition of 50 mL of 10 mmol/L ammonium formate per 1000 mL of the combined solution. The elution process was segmented as follows: from the start to 12.0 minutes, 40% ~100% of B; 12.0 to 13.5 minutes maintained at 100% B; 13.5 to 13.7 minutes, 100% ~ 40% B; and from 13.7 to 18.0 minutes, a steady 40% B. The column was held at 55 °C during this. Samples were injected in 3 μL volumes, irrespective of whether in positive (pos) or negative (neg) mode, with the autosampler cooled to 4 °C.

Mass spectrometry: Utilizing the Xcalibur 4.0.27 software (Thermo), a QE mass spectrometer operated in DDA mode to capture both MS and MS/MS spectral data. Within this setting, the software perpetually scans and assesses the complete MS spectrum. The conditions for the ESI source were designated as such: sheath gas at a rate of 30 Arb, Aux gas at 10 Arb, capillary temperatures set at 320 °C (for positive) and 300 °C (for negative), with full MS resolutions of 70,000 and MS/MS resolutions of 17,500. The collision energies were set at stages of 15/30/45 in the NCE mode. Spray voltages were either set to 5 kV (when positive) or dropped to -4.5 kV (when negative).

Data QC: The QC process for lipidomics was similar to that in polar metabolomics. The median RSD value of the ISs in our study was 3.41%. The distribution of QC samples in PCA was plotted to evaluate the instrument and process variability.

Data processing, metabolite identification and data analysis: The initial step involved converting the MS raw data files into mzXML format using ProteoWizard software (version 3.0.19282). Subsequently, LipidAnalyzer was employed to process the lipidomics data, encompassing essential data preprocessing tasks such as peak identification, alignment, extraction, retention time correction, and integration of peaks. The filtering of reliable lipid peaks and the normalization of data were similar to those in polar metabolomics. Then, the LipidBlast database was applied for lipid annotation. The MS/MS spectra matching score was also calculated as described in the polar metabolomic analysis section. Generally, 1312 MS/MS peaks were identified for lipidomics.

In summary, only the peaks with MS/MS names and with MS/MS matching scores higher than 0.3 were included for further analysis.

Region of interest (ROI) segmentation

ROIs were delineated with 3D Slicer software (https://www.slicer.org/). ROI delineation was conducted on all slices containing the whole tumor. Two radiologists blinded to the biochemical and pathological indices of each sample performed the ROI delineation independently. Radiologists repeated the ROI drawing and feature extraction twice with an interval of at least 1 month. ROI segmentation was initially performed with DCE-MRI of 60 patients who were selected randomly by the two radiologists to evaluate the intra- and interobserver reproducibility. Intraclass correlation coefficients (ICCs) were employed to
assess the consistency of feature extracted within and between observers. Radiomic features achieved satisfactory intra- and interobserver reproducibility with ICCs > 0.75 38.

During ROI delineation, the peritumoral and intratumoral areas were obtained by expanding the tumor outward with a 5 mm width and then subtracting the tumor area and by shrinking with a 5 mm width, respectively. Tumor and peritumoral regions were merged to constitute the whole tumor region. These delineated regions, namely the tumor, peritumor, intratumor, and whole tumor region, comprised the four sets of ROIs utilized for the subsequent extraction of radiomic features.

**Radiological feature extraction**

In this research, we derived radiomic features from CE-MRI images using the PyRadiomics package V3.0 39. This included shape features, first-order features and texture features. While shape-related features highlight the varying morphologies observed across tumors common to every phase, both the first-order features and texture features were extracted from each specific phase. The first-order features offer insight into the voxel intensity distribution, whereas the texture features draw upon four distinct texture matrices to elucidate the radiological patterns present in the ROI. These matrices are: the gray-level cooccurrence matrix (GLCM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), and gray-level size zone matrix (GLSZM).

**Pathological feature extraction**

We extracted nuclear features and tissue composition features based on the pathological WSIs. For nuclear features, we first used Hover-Net to segment nuclei on WSIs and classified them into tumor, inflammatory, stromal, normal (non-neoplastic epithelial) and dead (necrotic) nuclei40. Then, nuclear morphological features and texture features were extracted for individual tumor cells. Morphological features characterize nuclear shape using descriptors such as nuclear area, elongation, and curvature. Texture features characterize the local pixel distribution pattern within the nuclei, which included the statistics of pixel intensity within the nuclear contour41. Collectively, a total of 18 features were extracted for each individual cell. After this, a series of statistics (mean, standard variation, median, skewness and kurtosis) were calculated to integrated the cell-level features into WSI-level features. Finally, a total of 90 nuclear features were extracted for each WSI. Besides the nuclear features, we computed tissue composition features for each WSI. Our previous study established a tile-level tissue type classifier to classify image tiles from WSIs into five tissue types: tumor, stroma, immune infiltrate, normal gland, and necrosis or hemorrhage42. In current study, we first tessellated all WSIs into 256 × 256 pixel tiles under 20× magnification. Then, our established classifier was applied to all image tiles, and each tile was classified into one of the five tissue types. The percentage of image tiles of each tissue type were calculated and formed the tissue composition features for the WSI.
References


