

CANCER

Multiomic integration reveals subtype-specific predictors of neoadjuvant treatment response in breast cancer

Zongchao Mo^{1,2†}, Mei Yang^{3,4†}, Zhihan Zhu^{1†}, Minghao Wang^{1†}, Haoyu Wang¹, Zhanye Zhang¹, Shanshan Lyu⁵, Fangping Xu⁵, Haixia Shang¹, Huan Lin^{6,7}, Zeyan Xu^{6,7}, Suyun Li^{6,7}, Xiaobo Chen^{6,7}, Kun Wang^{4*}, Changhong Liang^{6,7*}, Jiguang Wang^{1,2*}, Zaiyi Liu^{6,7*}

Neoadjuvant therapy has been widely used in breast cancer, but treatment response varies among individuals. We conducted multiomic profiling on tumor samples from 149 Chinese patients with breast cancer across ER⁺HER2⁺, ER⁺HER2⁻, and ER⁻HER2⁻ subtypes, categorizing outcomes as pathologic complete response (pCR; $n = 81$) or residual disease (RD; $n = 68$). We identified distinct molecular features linked to pCR in each subtype: elevated cell proliferation in patients with ER⁻HER2⁻ pCR, higher *CDKN2A* methylation in patients with ER⁻HER2⁻ RD, increased *KIT* methylation in patients with ER⁻HER2⁺ RD, and *MAP4K1* hypermethylation in patients with ER⁺HER2⁺ RD. These findings were subsequently validated in independent datasets. By integrating clinical and multiomic data, we developed MOPCR, a subtype-specific machine learning model that outperformed single-omic approaches in predicting treatment response. MOPCR demonstrated potential generalizability across cohorts and provided preliminary stratification of patient subgroups with higher pCR probability, offering valuable insights for precision cancer management.

INTRODUCTION

As the most common tumor type in women, breast cancer caused 685,000 deaths in 2020 (1). The age-standardized rates of breast cancer in China were estimated as 17.0 to 34.3 cases per 100,000 women (2). Neoadjuvant chemotherapy (NACT) is a widely used method for treating breast cancer before surgery. NACT aims to shrink tumors and make locally advanced or inoperable breast cancer operable, avoiding axillary lymph node dissection and allowing breast-conserving surgery. Another substantial advantage lies in the potential to obtain additional prognostic and predictive data. Given that the residual cancer burden (RCB) is linked to the recurrence risk (3), therapy can be customized according to the treatment response. For example, the KATHERINE trial showed that patients with human epidermal growth factor receptor 2 (HER2)-positive breast cancer and residual disease (RD) after HER2-targeted therapy benefited from trastuzumab emtansine (T-DM1), improving recurrence-free survival (4). While achieving pathologic complete response (pCR) is associated with better event-free survival (5), NACT may delay surgery and increase disease

progression risk (6). Early pCR prediction could mitigate these risks and improve survival. However, the molecular mechanisms underlying NACT response remain unclear, and accurate pCR prediction remains challenging.

Breast cancer exhibits substantial heterogeneity in genomics, transcriptomics, proteomics, and treatment responses, underscoring the need for individualized, multiomic-driven management. The PAM50 model, for instance, associates luminal A tumors with the best relapse-free survival but the lowest pCR rates (7). The Cancer Genome Atlas Program (TCGA)'s multiomic analysis of 825 invasive breast adenocarcinomas identified subtype-specific features, such as *TP53* mutations and *MYC* gains in basal-like tumors, and distinct methylation patterns (8). However, these studies did not focus on NACT, leaving a gap in understanding multiomic features linked to NACT outcomes.

Recent studies highlight the role of clinical, transcriptomic, cellular, and genomic features in NACT response. High Ki67 levels (>20%) correlate with increased pCR rates (9), and gene expression-based predictors have been developed for HER2⁻ breast cancer (10–12). Tumor-infiltrating lymphocytes are also associated with higher pCR rates (13). Despite these insights, most studies examine features in isolation, leaving the predictive potential of integrative analyses underexplored.

Integrative DNA and RNA analyses have improved pCR prediction in HER2⁺ tumors, linking pCR to intrinsic subtypes and tumor microenvironment features (14). Other multiomic approaches integrating DNA, RNA, and pathology images have improved pCR prediction accuracy compared to traditional clinical predictors, emphasizing the importance of data integration from multiple platforms (15). However, these models were primarily developed in Western populations and lack validation in specific subtypes or diverse ethnic groups. This is particularly relevant as different receptor subtypes exhibit varying pCR rates in clinical practice, and studies indicate that Chinese patients with breast cancer have distinct molecular profiles

¹Division of Life Science, Department of Chemical and Biological Engineering, and State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong, SAR, China. ²SIAT-HKUST Joint Laboratory of Cell Evolution and Digital Health, HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen 518000, China. ³Department of Breast Surgery, Zhujiang Hospital, Southern Medical University, Guangzhou 510280, China. ⁴Department of Breast Cancer, Cancer Centre, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China. ⁵Department of Pathology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China. ⁶Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China. ⁷Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China.

*Corresponding author. Email: jgwang@ust.hk (J.W.); wangkun@gdph.org.cn (K.W.); liangchanghong@gdph.org.cn (C.L.); liuzaiyi@gdph.org.cn (Z.L.)

†These authors contributed equally to this work.

compared to Western ones (16, 17). Developing subtype-specific models is essential for enhanced predictive performance, and incorporating non-Western populations could help assess model generalizability. In addition, integrating epigenomics and proteomics—often overlooked in prior research (10, 14, 15)—could provide critical diagnostic and prognostic insights, as demonstrated in other cancers (18–22).

In this study, we performed multiomic profiling of 149 Chinese patients with breast cancer, analyzing somatic mutations, copy number alterations, DNA methylation, gene expression, protein expression, and phosphorylation. Patients were classified as pCR or RD based on postsurgery pathology. After identifying subtype-specific multiomic features associated with treatment response, we developed multi-omics pCR prediction models (MOPCR), a machine learning prototype model integrating clinical and multiomic data to predict pCR. The MOPCR prototype demonstrated promising accuracy and generalizability across independent cohorts, offering insights into treatment outcomes and survival. The model prototype was deployed on an online server at <http://www.wang-lab-hkust.com:8050> as a resource for the research community. To sum up, this study provided a comprehensive multiomic molecular portrait of Chinese breast cancer responders and nonresponders and a computational framework to predict NACT treatment outcomes.

RESULTS

Patient enrollment and treatment outcome evaluation

To study breast cancer treatment outcomes, we enrolled 149 newly diagnosed patients (age range: 23 to 73 years) from the Guangdong Provincial People's Hospital (GDPH) (Fig. 1, A and B). All patients received NACT followed by surgical intervention. HER2 and estrogen receptor (ER) statuses were assessed at diagnosis using immunohistochemistry staining, and the enrolled patients were grouped into three subtypes: ER⁺HER2⁺ ($n = 55$), ER⁺HER2⁺ ($n = 42$), and ER⁺HER2⁺ ($n = 52$). Anti-HER2 treatments were applied to all HER2⁺ patients except for one case. As outlined in fig. S1A, chemotherapy regimens included docetaxel plus carboplatin (TCb; $n = 74$), epirubicin plus cyclophosphamide followed by docetaxel (EC-T; $n = 61$), and docetaxel plus pyrotinib (TY; $n = 7$). Seven additional patients received regimens, including docetaxel alone (T).

To evaluate the treatment outcomes of the NACT, a pathologic examination was performed after the completion of surgical resection to assess whether there were residual tumors after the neoadjuvant treatments (fig. S1B) (23, 24). Patients were classified as pCR if no residual cancer cells were detected in resected tissues or lymph nodes during pathologic evaluation. Those identified as RD were further classified into RCB classes, including RCB I, RCB II, and RCB III (Materials and Methods and fig. S1C). Of 149 patients, 81 patients achieved pCR (26 in ER⁺HER2⁺, 32 in ER⁺HER2⁺, and 23 in ER⁺HER2⁺), while 68 patients had RD (29 in ER⁺HER2⁺, 10 in ER⁺HER2⁺, and 29 in ER⁺HER2⁺) (table S1). We found that pCR is associated with improved overall survival in the ER⁺HER2⁺ and favorable disease-free survival in both the ER⁺HER2⁺ and ER⁺HER2⁺ subtypes (fig. S1, D and E). We next assessed the association between clinical features and pCR using multivariable and univariate logistic regression analysis, identifying significant correlations with several clinical characteristics (Fig. 1C and fig. S2A). The HER2/ER-based subtypes had discriminative pCR rates of 47.3% for ER⁺HER2⁺, 76% for ER⁺HER2⁺, and 44% for ER⁺HER2⁺ ($P = 0.0035$, chi-square test; Fig. 1D). As tumor size

increased, pCR rates decreased from 61% in T1 and T2 to 17% in T3 and T4 tumors (Fig. 1E). Furthermore, the presence of lymph node metastasis was associated with resistance to NACT, resulting in a lower likelihood of pCR, with an odds ratio (OR) of 0.39 [confidence interval (CI): 0.18 to 0.80, $P = 0.0075$, Fisher's exact test, confidence level of 0.95; Fig. 1E]. Moreover, a higher tumor proliferation marker Ki67 value was associated with an increased pCR rate in the ER⁺HER2⁺ subtype ($P = 0.0079$, Wilcoxon rank sum tests; Fig. 1F). These findings align with observations from previous studies involving other cohorts (9, 15, 25–28).

We next compared pCR rates in our cohort to those of the previously reported breast cancer cases who underwent NACT (anti-HER2 therapy and/or chemotherapy). We found that our cohort showed overall higher pCR rates compared to the British breast cancer cohort TransNEO by Sammut *et al.* (15) in ER⁺HER2⁺ ($P = 0.094$) and ER⁺HER2⁺ ($P = 0.040$) subtypes (fig. S2B). A younger patient age in the pCR group reported by Sammut *et al.* (15) was not found in our cohort ($P = 0.108$), although the age distribution was comparable between the two cohorts ($P = 0.873$) (fig. S2C). Subsequently, we integrated our cohort with the TransNEO cohort and performed a multivariable analysis with 243 patients from two cohorts and found that tumor size of T3/T4, lymph node metastasis, and ER-positive status were still significantly associated with RD, while histologic grade III was associated with pCR. Notably, the cohort factor emerged as an independent predictor of RD, with an OR of 3.94 (CI: 1.95 to 8.24, $P = 1.7 \times 10^{-4}$). The confidence level for this analysis was set at 95% (fig. S2D). It has been recently reported that Chinese breast cancer had remarkable genomic disparities from the Caucasian cohort (16, 17). Specifically, the frequency of *PIK3R1* (phosphoinositide-3-kinase regulatory subunit 1) mutation was prevalent in Chinese breast cancer regardless of molecular subtypes (16), while the frequency of *PIK3CA* mutation was reported to be lower in Chinese ER⁺HER2⁺ (17). A multiomic dataset focusing on well-labeled Chinese patients with breast cancer might reveal population-specific disease mechanisms and help to establish alternative models with better performance on East Asian patients.

Multiomic profiling of breast cancer samples

We collected multiomic data from pretreatment tumor biopsies (Fig. 1B and table S1). Specifically, whole-exome sequencing (WES) was performed on 104 tumor-blood pairs to identify somatic single-nucleotide variants (SNVs), small insertions/deletions (INDELs), and somatic copy number variants (CNVs) from cancer genomes. Moreover, whole-genome bisulfite sequencing (WGBS), RNA sequencing (RNA-seq), and liquid chromatography-mass spectrometry were carried out to profile methylome, transcriptome, and proteome/phosphorylome on 145, 140, and 137 tumor samples, respectively. Among all the enrolled patients, 89 had data from all the above-mentioned omic platforms (fig. S2E).

The average sequencing depth of tumor WES data was approximately 600× (fig. S3A). Somatic SNVs and small INDELs were called across all samples using GATK Mutect2 (29), yielding a median load of 74 (ranging from 3 to 1078) nonsynonymous mutations in coding regions (Materials and Methods and table S2). These mutations included previously reported breast cancer driver genes, such as *TP53* ($n = 71$, 75.5%), *PIK3CA* ($n = 32$, 34.0%), *KMT2D* ($n = 8$, 8.5%), *KMT2A* and *KMT2C* ($n = 7$, 7.4%), *GATA3*, *MED13*, and *CSMD1* ($n = 6$, 6%) (Fig. 2A) (30–33). The median variant allele frequency of these driver genes was 0.166 (fig. S3B). Mutational signature analysis

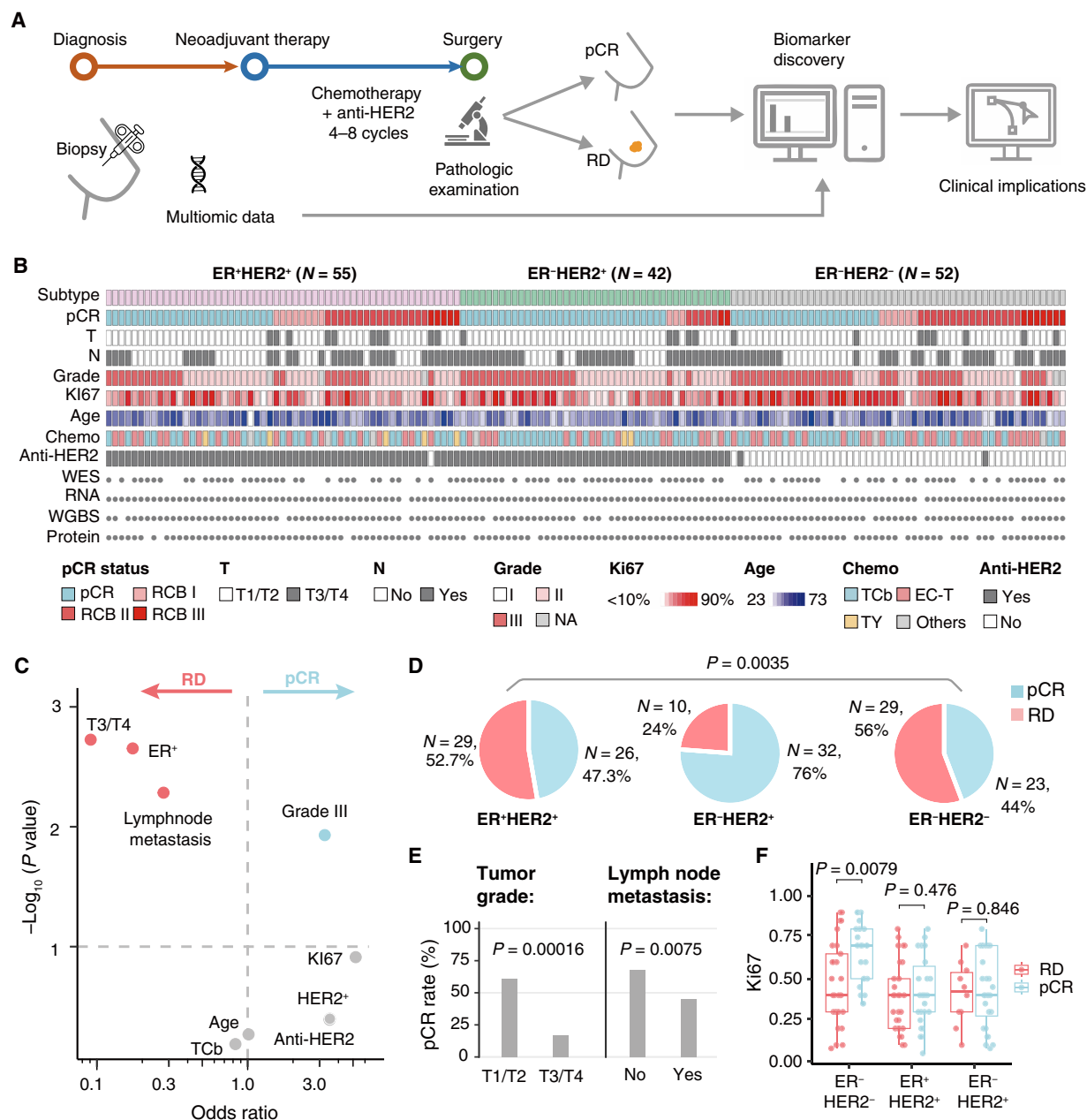


Fig. 1. Summary of the cohort and clinical features associated with response to NACT. (A) Design of this study. Biopsy samples for multiomic data were collected at diagnosis, with treatments administered before surgery. Pathologic examination classified patients into pCR and RD groups. Biomarkers were identified by comparing these groups across different omic data and integrated using machine learning models to predict treatment outcomes. (B) Summary of clinical information and availability of four omic data types used in this study. T, tumor size; N, lymph node metastasis; grade, histological grade; chemo, the combination of chemotherapy; T, docetaxel; Cb, carboplatin; EC, epirubicin plus cyclophosphamide; Y, pyrotinib; anti-HER2, anti-HER2 treatment. (C) Multivariate analysis results by logistic regression, highlighting clinical features with significant independent association with NACT outcome ($P < 0.1$). (D and E) Association between pCR and subtypes (D), clinical tumor grade (E), and lymph node metastasis status (E) at diagnosis. P values were calculated by the chi-square test (D) and Fisher's exact test (E). (F) Association between pCR and Ki67 index among subtypes. P values were calculated by Wilcoxon rank sum tests.

demonstrated a predominance of age-related signature (SBS1 and SBS5) and APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) signature (SBS3 and SBS13), suggesting that these processes drove the most of the observed mutagenesis (Fig. 2B) (34). As expected, SBS1 and SBS5 were positively correlated with patient age at diagnosis (fig. S3C). Notably, the SBS3 signature was

enriched in the ER[−]HER2[−] ($P = 1.1 \times 10^{-5}$) but not the SBS13 (fig. S3D). Further, we inferred the CNV and purity based on the paired WES data and observed no significant difference in tumor purity between pCR and RD (fig. S3E and table S3). The CNV landscape revealed a distinct profile among subtypes (fig. S3F). Afterward, genomic identification of significant targets in cancer 2 (GISTIC2)

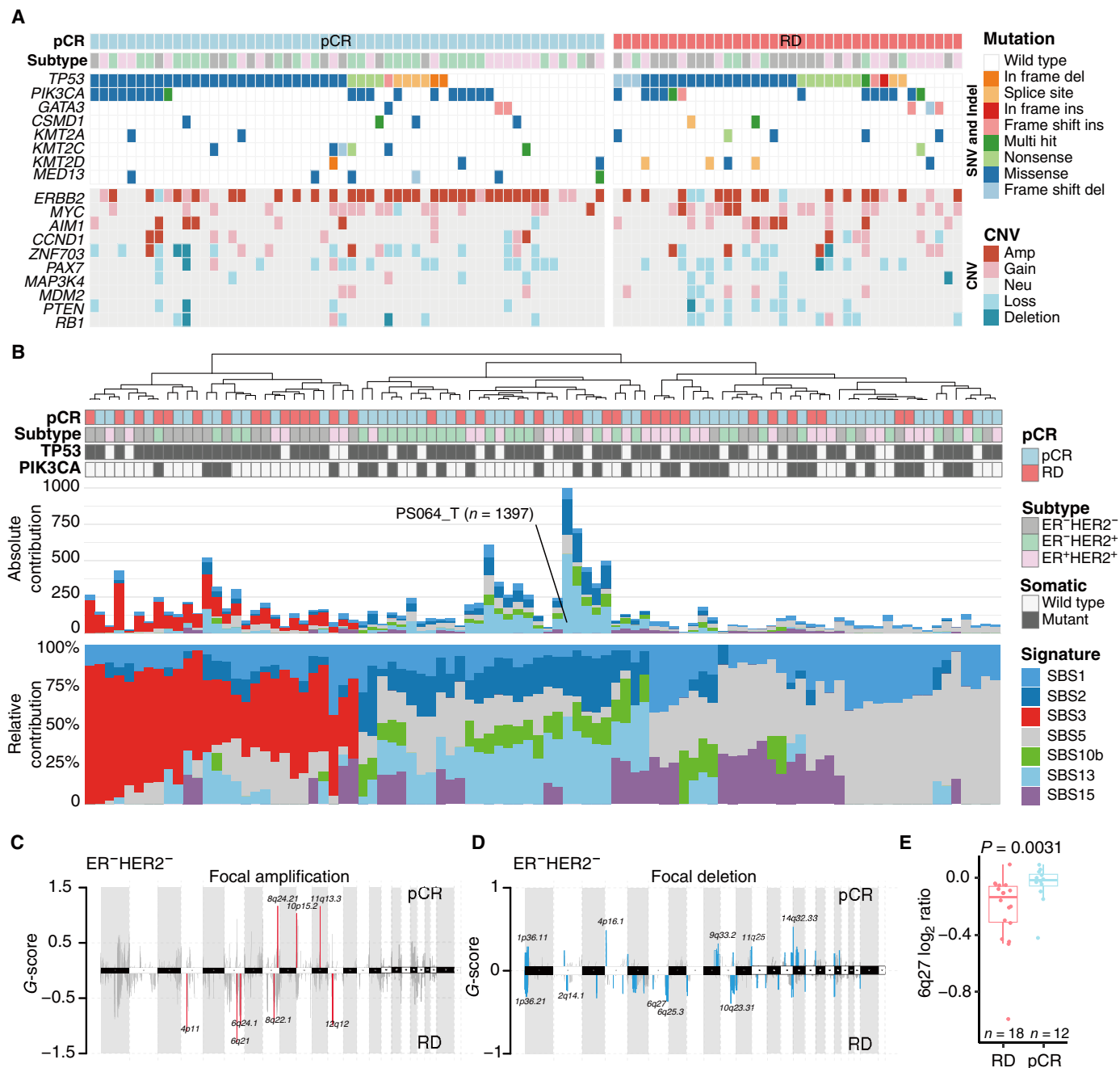


Fig. 2. Somatic mutation signature and landscape. (A) Somatic mutation landscape in pCR and RD groups, respectively. Only 94 samples with available purity by FACETS were shown. (B) Absolute and relative SNV signature distribution in 94 patients. Patients were clustered on the basis of the relative signature contribution. (C and D) Focal amplification and focal deletion comparison between pCR and RD groups in the ER⁻HER2⁻ subtype. (E) Log₂ ratio of cytoband 6q27 in the ER⁻HER2⁻ subtype.

(35) was applied to find recurrent focal and arm-level copy number alterations in this cohort (fig. S3G). Consistent with previous results, *ERBB2*, *MYC*, *AIM1*, and *CCND1* gain/amplification were abundant in the breast cancer cohort (Fig. 2B) (36).

Somatic mutational landscape of pCR versus RD breast cancer genomes

To interrogate whether somatic mutations contribute to treatment response, we compared the frequencies of nonsilent somatic alterations

in the pCR group versus the RD group. No somatic gene SNV/INDEL was significantly associated with pCR after the false discovery rate (FDR) adjustment. For the CNV, by comparing the focal CNV landscape between the pCR and RD groups in each subtype, we found that the 6q27 copy number was significantly lower in the RD group in the ER⁻HER2⁻ (Fig. 2, C to E; Materials and Methods; and fig. S3H). Further inclusion of more samples might verify this observation and enable the identification of more genetic variants that can predict the treatment outcomes.

Methylome comparison between pCR versus RD breast cancer samples

Methylation of gene regulatory regions plays a crucial role in regulating gene expression and may influence treatment response. To investigate the relationship between pCR and the methylation landscape, we analyzed WGBS data from 145 tumor samples, focusing on predefined gene regulatory regions and their association with NACT outcome. Initial analysis revealed no significant differences in methylation levels between pCR and RD groups when subtypes were not stratified (fig. S4A). However, consistent with prior studies, we observed significantly higher methylation levels in promoter and enhancer regions in $HER2^+$ subtypes compared to ER^-HER2^- subtypes (Fig. 3A) (37). When comparing methylation levels between pCR and RD groups within each subtype, we identified elevated methylation in the distal enhancer-like signature region ($P = 0.076$) and the CCCTC-Binding Factor (CTCF)-only region ($P = 0.060$) in the ER^+HER2^+ subtype, although these differences did not reach statistical significance (Fig. 3A). These findings suggest potential subtype-specific methylation patterns that may be associated with treatment response, warranting further investigation (Fig. 3A).

To uncover more specific methylation differences between pCR and RD, we segmented WGBS CpG methylation data into methylation regions for differential methylation analysis (Fig. 3B). To address the high dimensionality of CpG methylation data and reduce false positives, we developed a network-based feature selection method. This approach integrated prior knowledge from biological pathways (38), gene-disease association (39), protein-protein interaction (40), and gene regulation (41) to construct a breast cancer-specific biological network, termed BRCANET (Fig. 3B and Materials and Methods). We further correlated methylation regions with corresponding gene expression to identify methylation-regulated regions, prioritizing those with significant methylation expression anticorrelations and differential methylation between pCR and RD (Fig. 3B and Materials and Methods). Using this framework, we highlighted methylation region candidates within the context of BRCANET.

To explore the relationship between methylation and pCR across molecular subtypes, we conducted differential methylation gene analysis (Materials and Methods). Briefly, genome-wide CpG methylation data were segmented into regions, and the top-ranking differentially methylated region (DMR) for each gene was selected as its representative [differentially methylated gene (DMG)] based on RNA-methylation associations and methylation differences stratified by the NACT outcome (Materials and Methods). When analyzing all subtypes collectively, hypermethylation of the *FGF1* and *SNAIL1* genes was associated with the pCR group, while hypermethylation of the *MAP4K1* and *CDKN2A* genes was observed in the RD group (fig. S4B). Separately in the ER^-HER2^- , ER^-HER2^+ , and ER^+HER2^+ subtypes, we identified 465, 576, and 2242 hypermethylated genes in the pCR group, compared to 1381, 708, and 208 hypermethylated genes in the RD group, respectively. Using the BRCANET framework, we further refined these findings to 37, 32, and 98 hypermethylated genes in the pCR group and 69, 15, and 11 hypermethylated genes in the RD group for the ER^-HER2^- , ER^-HER2^+ , and ER^+HER2^+ subtypes, respectively (Fig. 3C and fig. S4C). Notably, *CDKN2A* and *BIRC5*, involved in the cell cycle, were significantly hypermethylated in ER^-HER2^- RD tumors. *XRCC2* in ER^-HER2^+ and *MAP4K1* and *MFNG* in ER^+HER2^+ were hypermethylated in RD tumors (Fig. 3C). Conversely, *NFIA* (nuclear factor I A), *BRCA1* (BRCA1 DNA repair associated, involved in DNA repair), and *SNAIL1* (snail family transcriptional repressor 1,

involved in epithelial-to-mesenchymal transition) hypermethylation was associated with pCR in ER^-HER2^- tumors; *GADD45A* (growth arrest and DNA damage-inducible α), *NCOR2* (nuclear receptor corepressor 2), *SNAIL1*, and *KIT* (KIT proto-oncogene, receptor tyrosine kinase) hypermethylation in ER^-HER2^+ tumors; and *AQP1* (aquaporin 1), *PPARA* (Peroxisome Proliferator-Activated Receptor α), *SPOP* (speckle type BTB/POZ protein), and *JUP* (junction plakoglobin, a common junctional plaque protein) hypermethylation in ER^+HER2^+ tumors (Fig. 3C). While these genes exhibited significant P values and relatively large methylation differences between pCR and RD, the limited sample size and the lack of FDR correction raise the possibility of high false positives, complicating their selection for downstream analysis. Nonetheless, among these, *CDKN2A* in the ER^-HER2^- subtype exhibited a strong β value difference and significant P value, prompting further investigation.

CDKN2A methylation as a predictor of pCR in triple-negative breast cancer

CDKN2A, a known tumor suppressor gene, is frequently inactivated in various cancers through mechanisms including promoter hypermethylation (42, 43). In our analysis, *CDKN2A* methylation levels were significantly lower in the ER^-HER2^- subtype compared to $HER2^+$ subtypes (fig. S4D), which was also supported by the TCGA cohort (fig. S4E) (8). Notably, higher *CDKN2A* methylation levels were specifically associated with RD in ER^-HER2^- tumors, with no such association observed in other subtypes (fig. S4D). This association persisted when stratified by treatment regimens and RCB classes (fig. S4, F and G). *CDKN2A* methylation exhibited a strong negative correlation with both gene expression (Pearson's correlation coefficient $R = -0.55$, $P = 7.1 \times 10^{-05}$) and protein expression (Pearson's correlation coefficient $R = -0.31$, $P = 0.042$) (fig. S5, A to C). Consistent with these findings, *CDKN2A* expression was lower in patients with RD, particularly in the ER^-HER2^- subtype ($P = 0.0499$) (fig. S5D). When stratified by treatment regimen, reduced *CDKN2A* expression remained evident (fig. S5E).

To validate the association between *CDKN2A* methylation and treatment response, we analyzed six external ER^-HER2^- cohorts with NACT and pCR labels. Given the scarcity of methylation data and the strong negative correlation between *CDKN2A* methylation and expression in our cohort, we focused on *CDKN2A* gene expression as a surrogate marker. In five of the six cohorts—Sammuto2021 (15), Wolf2022 (44), Hatzis2011 (10), Hoogstraat2022 (45), and BrightNess (46)—*CDKN2A* expression was consistently higher in patients with pCR compared to patients with RD, although some RD cases also exhibited elevated *CDKN2A* expression. The Anurag2022 (47) cohort was the exception, showing no significant difference but only a trend (fig. S5, F to K). These findings further support the elevated *CDKN2A* expression in ER^-HER2^- pCR breast cancer.

To further confirm the *CDKN2A* DMR methylation as a predictor of pCR in ER^-HER2^- , we collected an independent cohort of 75 patients with ER^-HER2^- treated with NACT. We designed a region for bisulfite pyrosequencing based on the methylation profiles in the discovery cohort, using sliding windows of 100 bp. This approach enabled us to identify the optimized CpG region that was most strongly associated with pCR in ER^-HER2^- tumors, which was found near one of the transcription start sites (TSSs) of *CDKN2A* (Fig. 3D). Pyrosequencing was then performed on formalin-fixed, paraffin-embedded (FFPE) samples from the 75 additionally collected cases (table S4), and 10 cases with WGBS data were used for

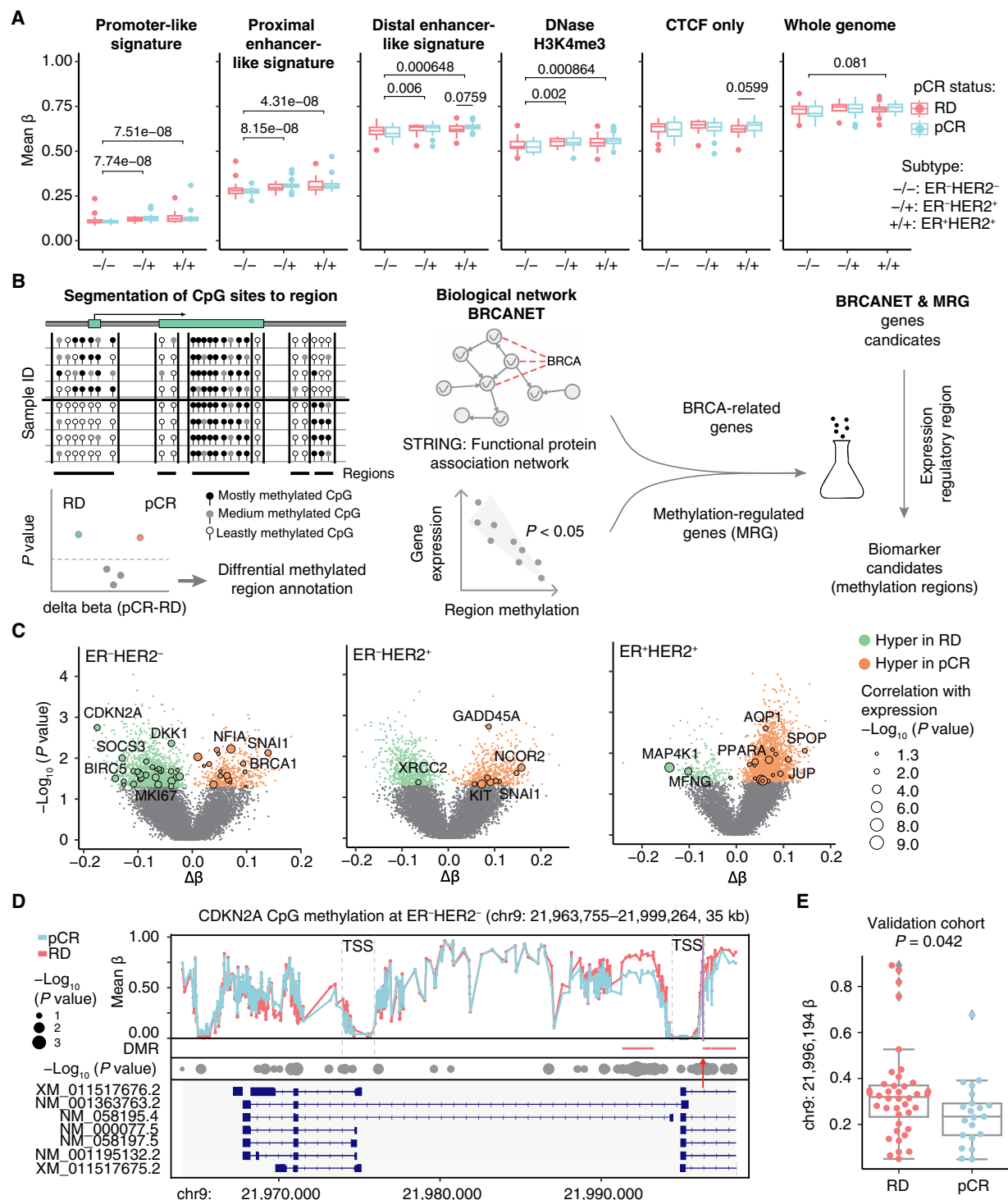


Fig. 3. Methylation features as predictors for pCR. (A) Mean methylation levels in cis-regulatory elements for pCR and RD in each subtype. P values were obtained by comparing pCR versus RD and among subtypes; only $P < 0.1$ is shown. DNase, deoxyribonuclease. (B) Framework for identifying methylation biomarkers. Genome-wide CpG methylation was segmented into regions, and differential methylation regions (DMRs) were identified using rank sum tests. A representative region was selected for each gene to represent the gene by considering methylation differences (pCR versus RD) and associations with RNA expression. Comparison of representative regions helps identify the differentially methylated genes (DMGs). Genes from the BRCANET list were prioritized. DMRs within the DMGs are prioritized as biomarkers for downstream analysis (C) DMGs for pCR versus RD in each subtype ($P < 0.05$). BRCANET genes are outlined in black. $\Delta\beta$ indicates mean methylation differences; larger points reflect more significant associations with gene expression. (D) CpG-level methylation for *CDKN2A* in ER-HER2⁻ (pCR versus RD). Top: Mean β values for each CpG site; dashed lines mark ± 1000 bp from the transcription start site (TSS). The second panel: Top two DMRs identified through segmentation. The third panel: Rank sum test P values for CpG sites ($P < 0.05$); larger points indicate greater significance. RefSeq transcripts are shown at the bottom, with red arrows marking the gene start and end. (E) The distribution of methylation level at chr9: 21,996,194 by pyrosequencing in the independent ER-HER2⁻ validation cohort. P value was calculated by the Wilcoxon rank sum test.

quality control. After quality control based on consistency between WGBS data and pyrosequencing data, as well as sequencing quality (Materials and Methods), one CpG site (chromosome 9: 21,996,194) was selected for downstream analysis (fig. S6, A and B). Despite the small sample size and FFPE samples, we observed a significant association between hypermethylation and RD in the validation cohort ($P = 0.042$; Fig. 3E).

In the ER⁺HER2⁺ subtype, highlighted DMGs such as *GADD45A*, *SNAI1*, and *KIT*, whose differential methylation regions were near TSSs (fig. S7A), did not exhibit significant differences in gene expression in independent cohorts (fig. S7B). However, *KIT* expression was significantly higher in the RD groups in this study cohort (fig. S7B). Upon further examination of the CpG level methylation pattern in this study and the expression in independent external datasets, *MAP4K1* hypermethylation in the RD group of ER⁺HER2⁺ subtypes might be a potential biomarker for NACT outcomes (Fig. 3C, Materials and Methods, and fig. S8A). While *MAP4K1* methylation differed significantly between pCR and RD ($P = 0.017$; fig. S8B) and significantly correlated with its expression ($P = 1.4 \times 10^{-09}$, Spearman $R = -0.49$; fig. S8C), no significant expression difference was observed in our cohort ($P = 0.805$; fig. S8D). Stratified by the treatment regimen, the *MAP4K1* methylation difference remained significant and had the same trend in TCb and EC-T regimens, respectively (fig. S8E), suggesting minimal regimen-driven impact. Higher *MAP4K1* methylation was observed in each RCB class compared to pCR (fig. S8F). In the Hurvitz2020 cohort (48), the expression difference approached significance ($P = 0.059$) in the treatment regimen with the largest sample size (fig. S8G). Furthermore, significantly lower *MAP4K1* expression in patients with RD from the Sammut2021 ($P = 4.7 \times 10^{-4}$) and Wolf2022 ($P = 3.5 \times 10^{-3}$) cohorts supported its association with treatment outcome, aligning with the hypermethylation observed in patients with RD (fig. S8, H and I).

In summary, *CDKN2A* hypermethylation and reduced expression in ER⁺HER2⁺ RD tumors suggest its potential as a predictive biomarker for NACT outcomes in this subtype. *KIT* and *MAP4K1* hypermethylation in patients with RD in the ER⁺HER2⁺ and ER⁺HER2⁺ subtypes could also be a candidate for predicting pCR.

Subtype-associated transcriptomic predictors of chemosensitivity

Using the RNA-seq on 140 patients, we further investigated the association between pCR and gene or pathway expression. Initially, we examined the distribution of gene expression through t-distributed stochastic neighbor embedding (t-SNE) and found that samples belonging to the same subtype showed a trend to cluster together (Fig. 4A). Examination of the t-SNE1 and t-SNE2 component distributions revealed significant differences among the various subtypes (fig. S9A). Stratified by the pCR and RD, only t-SNE2 exhibited a significant difference in the ER⁺HER2⁺ subtype (fig. S9B). Next, to create expression-based clusters linked to clinical phenotypes, we first conducted a three-way analysis of variance (ANOVA) to associate gene expression with HER2 status, ER status, and pCR status, followed by implementing hierarchical clustering using the genes identified by ANOVA (Fig. 4B and fig. S9, C and D).

Next, we conducted differential expression analysis using DESeq2 (49) and identified 1481 and 1431 overexpressed genes in the pCR and RD groups, respectively (fig. S9, E and F). BRCANET highlighted 84 pCR-associated and 45 RD-associated overexpressed genes, including *ERBB2*, which was overexpressed in pCR (fig. S9, E and F). Gene set

enrichment analysis (GSEA) (50) revealed that pCR-associated genes were enriched in proliferation-related pathways (e.g., MTORC1_SIGNALING and E2F_TARGETS) and immune-related pathways (e.g., ALLOGRAFT_REJECTION and INFLAMMATORY_RESPONSE) (fig. S9G). Subtype-specific analysis identified 911, 782, and 584 overexpressed genes in the pCR group and 905, 1056, and 564 in the RD group for ER⁺HER2⁺, ER⁺HER2⁺, and ER⁺HER2⁺ tumors, respectively (Fig. 4C and fig. S10A). Wilcoxon rank sum tests and BRCANET highlighted 59, 38, and 19 pCR-associated overexpressed genes and 34, 39, and 20 RD-associated overexpressed genes in ER⁺HER2⁺, ER⁺HER2⁺, and ER⁺HER2⁺ tumors, respectively (Fig. 4C and fig. S10A).

In ER⁺HER2⁺ tumors, pCR was associated with lower expression of *KLK3* and *AR* but higher expression of proliferation-related genes such as *CDK1* and *MKI67*. In ER⁺HER2⁺ tumors, patients with RD exhibited higher expression of *GFRA3* and *FOXO3*, while in ER⁺HER2⁺ tumors, patients with RD showed elevated expression of *ABCC3*, an ABC transporter gene linked to multidrug resistance (Fig. 4C) (51). Pathway enrichment analysis revealed enrichment of cell cycle pathways (e.g., E2F_TARGETS, G2M_CHECKPOINT, MTORC1_SIGNALING) and immune-related pathways (e.g., MYC_TARGETS_V1 and INTERFERON_ALPHA_RESPONSE and INTERFERON_GAMMA_RESPONSE) across subtypes in pCR groups, with varying levels of significance depending on the subtype (Fig. 4D).

Accordingly, in the single-sample gene set enrichment analysis (ssGSEA) (52), we extracted sample-level pathway expressions as features (table S5). In ER⁺HER2⁺ tumors, we observed that E2F_TARGETS ssGSEA scores were higher in the pCR group (fig. S10B) and strongly correlated with the Ki67 index (fig. S11A), indicating consistency between data from different platforms in our cohort. In addition, we found that overexpression of the E2F_TARGETS pathway was an exclusive pCR predictor for ER⁺HER2⁺ tumors in our cohort, which was also observed in the TransNEO cohort (fig. S11B). In ER⁺HER2⁺ tumors, apical surface and estrogen response pathways showed significantly higher expression in the pCR group, although this trend was not statistically significant in the TransNEO cohort (figs. S10B and S11C).

Overall, transcriptomic analysis revealed elevated expression of proliferation-associated genes and pathways in ER⁺HER2⁺ tumors achieving pCR. The differentially expressed genes and pathways between the pCR and RD groups vary across molecular subtypes, suggesting the presence of potential subtype-specific mechanisms of treatment resistance or sensitivity.

TOP2A as a predictor in triple-negative tumors identified in proteomic analysis

To uncover the posttranscriptional differences between patients with pCR and patients with RD, we analyzed protein and phosphoprotein expression data from 137 individuals. After filtering low-expression proteins and normalizing the data (Materials and Methods), we compared 5473 proteins, identifying 152 overexpressed in the pCR group and 156 in the RD group. BRCANET highlighted 13 pCR-associated proteins [e.g., MCM7 (minichromosome maintenance complex component 7), ERBB2 (erb-B2 receptor tyrosine kinase 2), and S100A8 (S100 calcium binding protein A8)] and 8 RD-associated proteins [e.g., VCAN (versican), PTPRG (protein tyrosine phosphatase receptor type G), and EGFR (epidermal growth factor receptor)] (fig. S12A). Among 1737 phosphoproteins, 42 were overexpressed in the pCR

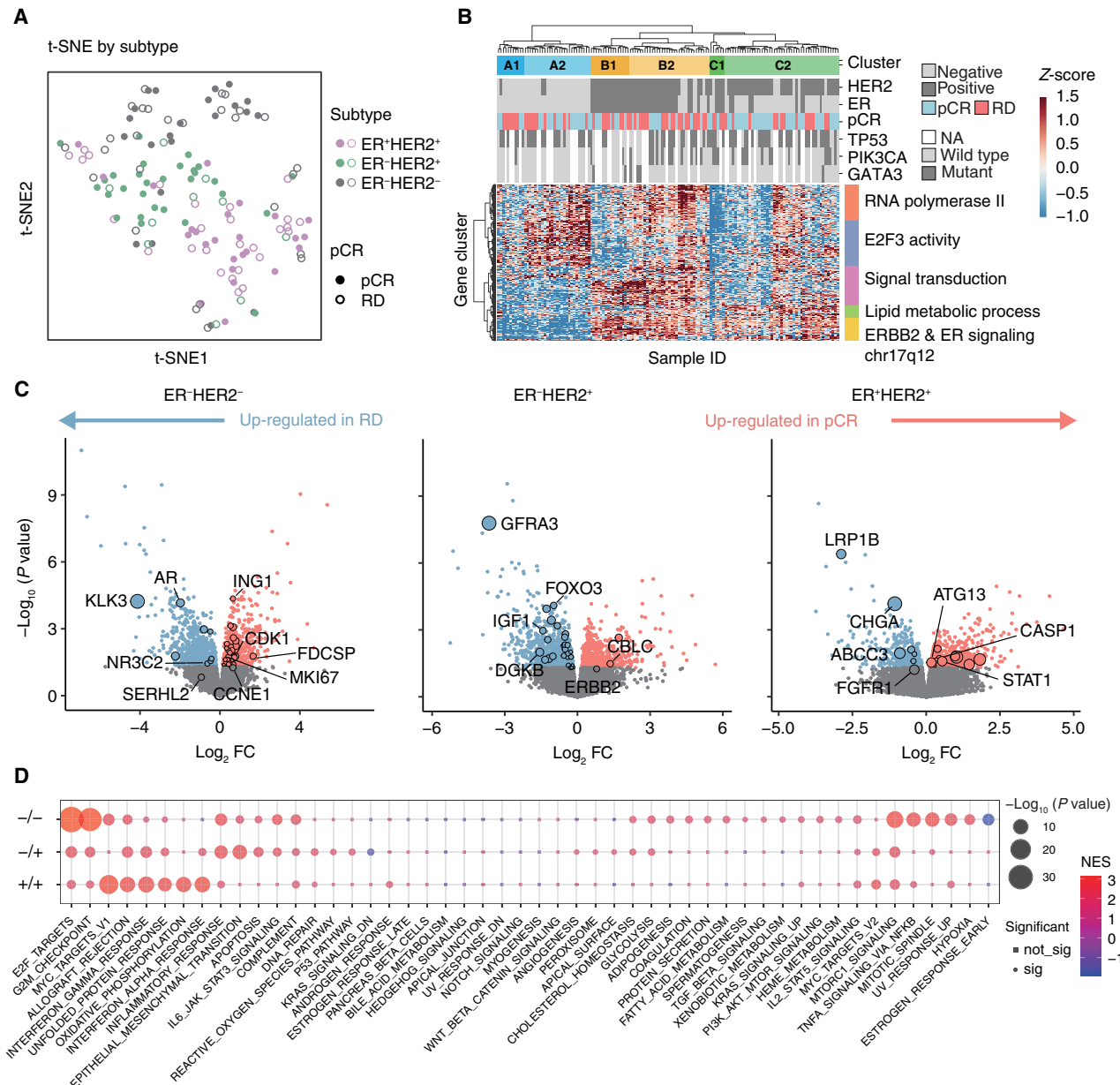


Fig. 4. Transcriptomic features associated with NACT outcome. (A) t-SNE analysis of gene expression. Points are colored by molecular subtypes, with solid points indicating pCR and hollow points representing RD. (B) Analysis of variance (ANOVA)-based clustering of gene expression. The top 100 subtype-associated genes (ANOVA) were used for clustering, with rows showing functional annotations of enriched gene clusters. (C) Differentially expressed genes between pCR and RD in each subtype (DESeq2, $P < 0.05$). BRCANET genes are highlighted with black borders. Dot size represents P values; FC indicates fold change. (D) Gene set enrichment analysis (GSEA) on the hallmark pathway for each subtype. Pathways were clustered by $-\log_{10}$ (adjusted P value). Dot size indicates significance; red denotes pCR enrichment, and blue denotes RD enrichment. Subtypes: $-/-$ (ER $^{-}$ HER2 $^{-}$), $-/+$ (ER $^{-}$ HER2 $^{+}$), and $+/+$ (ER $^{+}$ HER2 $^{+}$). NES, normalized enrichment score.

group [e.g., RPS6KB1 (ribosomal protein S6 kinase B1), FBLN2 (fibulin 2), MCM2 (minichromosome maintenance complex component 2), and ABCC1 (ATP binding cassette subfamily C member 1)], while 45 were overexpressed in the RD group [e.g., HIC1 (HIC ZBTB transcriptional repressor 1), FBXO4 (F-box only protein 4), and CREBBP (CREB binding protein)] (fig. S12B).

Subtype-specific analysis revealed differentially expressed proteins associated with pCR. In ER⁻HER2⁻ tumors, BRCANET highlighted 27 pCR-associated proteins [e.g., MCM7, CDK1 (cyclin dependent kinase 1), PCNA (proliferating cell nuclear antigen), MCM2, and

TOP2A (DNA topoisomerase II alpha)] and 20 RD-associated proteins (e.g., HERC2, a DNA damage response regulator; Fig. 5A and fig. S12, C and D) (53). In ER⁻HER2⁺ tumors, CTR9 and HIF1AN were overexpressed in the pCR group, while STAT5B, essential for breast cancer cell migration, and MSH6, involved in DNA mismatch repair, were elevated in patients with RD. Specifically, MSH6 protein showed a trend of increasing expression corresponding to higher RCB status (Fig. 5A and fig. S12D) (54, 55). In ER⁺HER2⁺ tumors, eight pCR-associated and five RD-associated proteins were highlighted by BRCANET (Fig. 5A and fig. S12C).

For differentially expressed phosphoproteins (DEPPs), ER[−]HER2[−] tumors showed overexpression of cell cycle-related proteins (e.g., TOP2A, MKI67, MCM2, and RPS6KB1) in the pCR group, while patients with RD exhibited higher expression of HSPB1, HIC1 (linked to chromosomal stability), and MTOR (involved in stress responses) (Fig. 5B and fig. S12, E and F) (56, 57). In ER[−]HER2⁺ and ER⁺HER2⁺ tumors, only a few DEPPs were identified, such as FBXO4 and ADAM17 in ER[−]HER2⁺ and SMARCA2 and CTNND1 in patients with ER⁺HER2⁺ RD (Fig. 5B and fig. S12, E and F).

Hallmark pathway analysis using protein and phosphoprotein abundance revealed that patients with ER[−]HER2[−] pCR exhibited higher expression of cell cycle-related pathways (e.g., E2F_TARGETS and G2M_CHECKPOINT), while patients with RD showed elevated adipogenesis and myogenesis pathways, supported by both protein and phosphoprotein data (Fig. 5C). Higher expression of the adipogenesis protein pathway was independently validated in an independent ER[−]HER2[−] cohort (fig. S13A).

TOP2A, a DNA topoisomerase, was significantly up-regulated at both protein and phosphoprotein levels in patients with ER[−]HER2[−] pCR (fig. S13B), with a correlated expression pattern across RNA, protein, and phosphoprotein levels (fig. S13C). Its overexpression has been linked to enhanced anthracycline-based therapy response (58). Given the TOP2A difference supported by multiomics, we investigated whether multiomic integration could help identify additional convergent differences between patients with pCR and patients with RD. We conducted multiomic integration and Hallmark pathway enrichment analysis on significantly differentially expressed genes ($P < 0.1$) across different breast cancer subtypes under pCR and RD conditions (Fig. 5D). Pathway enrichment patterns varied substantially among subtypes. In the ER[−]HER2[−] subtype, both multiomic and single-omic analyses identified significantly enriched pathways ($P < 0.05$), primarily associated with cell cycle regulation (e.g., E2F_TARGETS, G2M_CHECKPOINT, and MITOTIC_SPINDLE). For example, we can observe that genes in the E2F_TARGETS pathway exhibit significant and directionally consistent changes across various single-omic datasets (fig. S13D). In contrast, ER[−]HER2⁺ and ER⁺HER2⁺ tumors showed more scattered pathway enrichment. Nonetheless, with multiomic integration, we observed potential biologically relevant signals, such as the XENOBIOTIC_METABOLISM pathway in ER[−]HER2⁺ tumors, which was not significant in single-omic analyses (Fig. 5D).

An integrated model to predict patient outcome

By comparing patients with pCR and patients with RD across omic layers for each subtype, we identified candidate biomarkers associated with treatment outcomes. The initial prediction of the Chinese breast cancer cohort using the TransNEO machine learning model yielded modest performance [area under the receiver operating characteristic curve (AUC): 0.62 for clinical + RNA versus 0.57 for clinical only; fig. S13E]. To improve predictive power, we sought to develop subtype-specific pCR predictors by integrating multiomic features within a machine learning framework. However, high collinearity between omics and among features (fig. S14, A to C) and the risk of overfitting due to limited sample sizes and numerous features posed challenges for model generalizability. Furthermore, most publicly available datasets primarily contained RNA-seq or array data, limiting the generalizability of models to external validation datasets as well. To address this, we calculated hallmark pathway activity using single-sample GSEA for RNA, protein, and phosphoprotein

data. These pathway-level features, combined with clinical information and highly reproducible methylation markers validated in external datasets (Materials and Methods and figs. S15, A to C, and S16A), were used to train a pCR prediction model, ensuring generalizability to independent validation datasets.

We developed models using the LightGBM algorithm (59) and evaluated the performance of the MOPCR against single-omic models through 30 repeated fivefold cross-validation (Fig. 6A and Materials and Methods). MOPCR consistently outperformed single-omic models across all subtypes (Fig. 6B and fig. S16, B and C). In the ER[−]HER2[−] subtype, MOPCR achieved an average accuracy of 0.82, an AUC of 0.98, and an area under the precision-recall curve of 0.97, significantly surpassing single-omic models. Similar improvements were observed in ER[−]HER2⁺ and ER⁺HER2⁺ subtypes (Fig. 6B and fig. S16, B and C).

To ensure interpretability, we calculated SHapley Additive exPlanations (SHAP) values (60) to assess feature importance in MOPCR (Fig. 6C). Key features for the ER[−]HER2[−] model included MTORC1_SIGNALING, E2F_TARGETS, and UV_RESPONSE_UP protein pathways, Ki67 index, and CDKN2A methylation. In the ER[−]HER2⁺ model, KIT methylation status was the most critical feature, followed by tumor grade, tumor size, and RNA-level pathways including EPITHELIAL_MESENCHYMAL_TRANSITION and INFLAMMATORY_RESPONSE. For the ER⁺HER2⁺ model, the MAP4K1 methylation level was the top feature, followed by RNA-level MYC_TARGETS_V1, phosphoprotein-level UNFOLDED_PROTEIN_RESPONSE, and protein-level HEME_METABOLISM pathways. Notably, methylation features consistently ranked as the top significant predictors in ER[−]HER2⁺ and ER⁺HER2⁺ subtypes (Fig. 6C), underscoring the importance of methylation profiling in treatment outcome prediction. Further validation in larger and more diverse cohorts is needed to confirm these findings and explore the causal relationship between methylation markers and treatment outcomes.

MOPCR demonstrated that integrating CDKN2A methylation and Ki67 index outperformed using a single feature only in the pyrosequencing cohort, underscoring the value of multiomic integration (fig. S16D). To further evaluate robustness and generalizability, we applied MOPCR to publicly available datasets. For datasets lacking methylation data, we imputed methylation levels using a least absolute shrinkage and selection operator (LASSO) model based on RNA data (Fig. 6A and Materials and Methods). In the TransNEO dataset, MOPCR exhibited better or comparable performance with imputed methylation than with RNA as a replacement of methylation (fig. S16, E to G). Across three independent RNA-seq datasets, AUC values ranged from 0.589 to 0.768 in the ER[−]HER2[−] subtype. A similar performance could also be observed for the ER[−]HER2⁺ and ER⁺HER2⁺ subtypes (Fig. 6D). In the BrightNess study, which included three treatment regimens—(A) paclitaxel + carboplatin + veliparib; (B) paclitaxel + carboplatin + placebo; and (C) paclitaxel + placebo + placebo—MOPCR achieved AUC values of 0.61, 0.60, and 0.55 for arms A, B, and C, respectively. The improved performance in arms A and B, which included carboplatin (a DNA replication-targeting drug) and veliparib (a DNA repair-targeting drug), suggests that the MOPCR ER[−]HER2[−] model may be more informative for evaluating treatment regimens combining paclitaxel with other cell cycle-targeting drugs (fig. S17, A and B), although this model could not suggest different treatment regimens.

Of all the independent RNA-seq cohorts subjected to MOPCR to predict the probability of pCR, the ER[−]HER2[−] subtype exhibited a

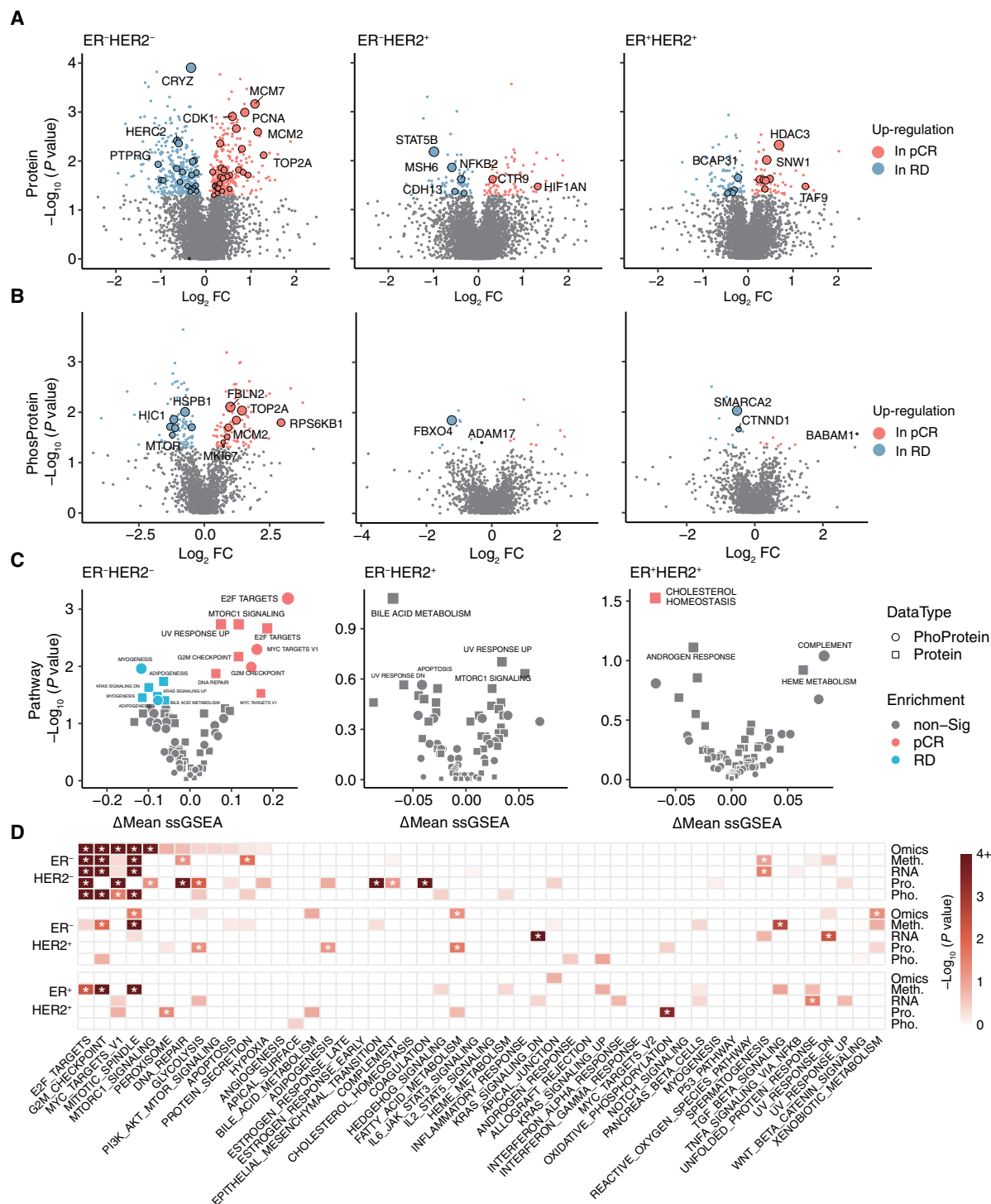


Fig. 5. Proteomic features associated with NACT outcome. (A and B) Differentially expressed proteins (A) and phosphoproteins (B) between pCR and RD in each subtype. *P* values were calculated using the Wilcoxon rank sum test. Significantly differentially expressed proteins and phosphoproteins ($P < 0.05$) within BRCANET are outlined in black. Dot size represents *P* values. FC, fold change. (C) Proteins and phosphoprotein pathway differences between pCR and RD. Pathway activity was assessed using protein and phosphoprotein abundance based on the ssGSEA method. Only significant pathways are shown for the ER⁺HER2⁻ subtype, while the top five pathways are displayed for HER⁺ subtypes. *P* value was calculated using the Wilcoxon rank sum test. (D) Hallmark pathway enrichment analysis of differentially expressed genes in single-omic data (Meth., methylation; RNA, transcription; Pro., protein; Pho., phosphoprotein) and multiomic data between pCR and RD in each subtype using the multiomic integration method. *P* values less than 1×10^{-4} are labeled as 1×10^{-4} , with significant results ($P < 0.05$) marked by the asterisk (*). UV, ultraviolet.

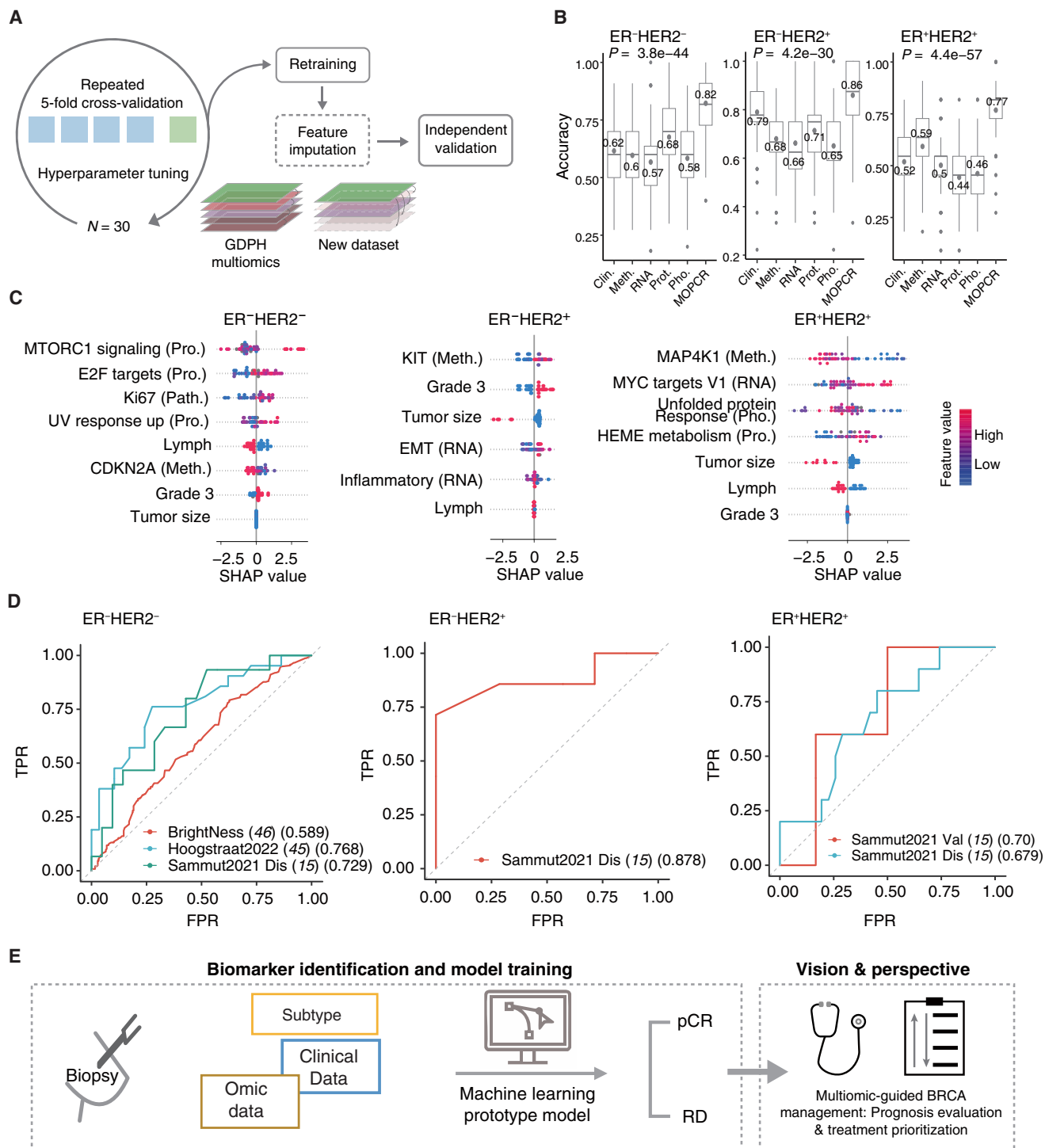


Fig. 6. Machine learning predictor of clinical outcome. (A) Framework for training and validating the machine learning model. External datasets with missing omic features were imputed using multiomic data from this study before independent validation. (B) Cross-validation performance (accuracy) comparison between single-omic models and multiomic pCR prediction models (MOPCR) across three subtypes. *P* values (Wilcoxon rank sum tests) compare MOPCR to aggregated results from other methods. (Clin., clinical features; Meth., methylation; RNA, gene expression; Prot., protein expression; Pho., phosphoprotein expression; MOPCR, multiomic model with feature selection.) The mean accuracy for each model was labeled. (C) SHapley Additive exPlanations (SHAP) values for features in the optimized MOPCR model. The dot color represents the feature value. (Lymph, lymph node metastasis; grade 3, tumor grade 3 at diagnosis; tumor size, T3/T4 at diagnosis; Path., pathologic; Meth., methylation; RNA, gene expression; Prot., protein expression; Pho., phosphoprotein expression.) EMT, EPITHELIAL_MESENCHYMAL_TRANSITION. (D) Receiver operating characteristic curves for independent validation in external datasets for each subtype (15, 45, 46). AUC values are shown in parentheses for each cohort. [FPR, false-positive rate; TPR, true-positive rate; Sammut Dis, discovery dataset from the TransNEO cohort (15); Sammut Val, validation dataset from the TransNEO cohort (15). Only cohorts with >10 cases were included.] (E) Summary, vision, and perspectives of the machine learning predictors in this study.

significantly higher likelihood in the pCR group than the RD group (fig. S17C), although there were instances of high pCR probability scores in the RD group and low pCR probability scores in the pCR group. Future integration of more high-confident biomarkers using larger datasets might help improve the prediction performance. To facilitate the use of MOPCR by the research communities to achieve the long-term vision and perspective of NACT pCR prediction in breast cancer (Fig. 6E), we developed a website for deploying the prototype models, which can be accessed at <http://www.wang-lab-hkust.com:8050>. In sum, although further investigation is still needed, our study showed that the combination of multiomic data and clinical information using machine learning techniques can lead to the informative prediction of pCR.

DISCUSSION

The motivation for predicting the pCR of NACT stemmed from the observed correlation between pCR and improved prognosis (5). In this study, patients achieving pCR generally experienced better outcomes than those with RD, underscoring the importance of predicting NACT response before treatment (fig. S1, D and E). Specifically, patients with RD in the ER[−]HER2[−] subtype showed poorer overall and disease-free survival, while in the ER[−]HER2⁺ subtype, RD did not affect overall survival but was associated with earlier disease progression ($P = 0.024$; fig. S1, D and E). For ER⁺HER2⁺ patients, where pCR is not directly linked to survival, additional data, such as drug sensitivity from NACT, would be invaluable. Nevertheless, accurate pCR prediction could still help optimize treatment regimens and reduce costs for ER⁺HER2⁺ patients, although our current sample size limits this capability.

Accurate pCR prediction could help mitigate NACT limitations by identifying patients unlikely to benefit. By facilitating patient stratification and personalized treatment, pCR prediction is critical for enhancing survival, quality of life, and cost effectiveness.

NACT response is influenced by subtypes, racial factors, and a tumor microenvironment, making prediction challenging (15, 25, 61). To address this, we conducted comprehensive multiomic profiling of a Chinese breast cancer cohort. The model by Sammut *et al.* was trained using all subtypes, prompting us to develop subtype-specific predictors tailored to the Chinese population while ensuring generalizability to external cohorts. We identified subtype-specific biomarkers by comparing pCR and RD groups. In the ER[−]HER2[−] subtype, *CDKN2A* methylation and cell proliferation were associated with pCR and validated in an independent cohort. While cell proliferation and immune-related pathways have been linked to pCR (15, 62, 63), our findings highlight their subtype-specific relevance, suggesting distinct response mechanisms across breast cancer subtypes.

We developed MOPCR, a subtype-specific multiomic model for pCR prediction, which demonstrated promising accuracy and generalizability. While the model identifies patients unlikely to achieve pCR, it has the potential to inform future clinical decision-making (10).

Our study has several limitations. First, the diversity of treatment regimens, resistance mechanisms, and sample size constraints may limit the model's generalizability to larger, more diverse datasets. Second, to avoid overfitting, not all identified biomarkers from each omic layer were included in MOPCR. Third, the limited sample size restricts subtype-specific stratification by treatment regimen, preventing the model from accounting for regimen-specific effects. In addition, in the machine learning-independent validation, the distribution

of some key features in the external datasets was explored to ensure reliability before validation, which means that the validation process is not entirely independent. Last, while our model aims at the long-term vision of application, it remains a research prototype, and further investigation is required before it can be applied in clinical settings.

Future directions include expanding sample sizes to uncover additional resistance mechanisms and improve model performance. Single-cell RNA-seq could provide deeper insights into tumor-infiltrating lymphocytes (64–66). Germline *BRCA1/BRCA2* mutations and structural variants, detectable via long-read sequencing technologies, warrant further investigation (67–70).

In conclusion, our study demonstrates that pCR prediction is enhanced by integrating subtype-specific multiomic features. We identified *CDKN2A* and *MAP4K1* methylation as predictive biomarkers for ER[−]HER2[−] and ER⁺HER2⁺ patients and developed MOPCR, an accessible tool for evaluating breast cancer treatment outcomes. This work also provides a comprehensive multiomic dataset for the community.

MATERIALS AND METHODS

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used (DeepSeek) to (correct grammar errors and polish the manuscript). After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. Typical prompt: Below is an introduction/result/method/discussion section of my manuscript aiming at Science Advances. Please help correct the grammar errors and polish it more concisely, coherently, and logically.

Patients and ethics statements

In this study, we enrolled 149 women with early and locally advanced breast cancer who underwent neoadjuvant treatment according to NCCN (National Comprehensive Cancer Network) guidelines. All participants provided written informed consent before enrollment in the study. Ethical approval for the study was granted by the GDPH review board (research ethics reference: KY-Z-2020-680-02). Molecular subtypes were determined through immunohistochemistry according to the four molecular indicators: ER/PR/HER2/KI67. Eligible patients included those with triple-negative (ER[−]/PR[−]/HER2[−]) and HER2⁺ subtypes for further analysis. All the cases had a complete molecular/digital pathology dataset, received more than 4 cycles of chemotherapy and/or targeted therapy, and underwent efficacy assessment. The response was assessed on completion of NACT and surgery by two pathologists, with confirmation from clinicians. pCR was defined as the absence of any residual cancer cell within the resected surgical specimen, including both breast and axillary lymph node specimens, irrespective of ductal carcinoma in situ (ypT0/ is ypN0). Patients with any residual cancer cells within the breast or lymph nodes were classified as RD. Specifically, one patient showing vascular tumor emboli and with evidence of cancer cells in the adipose tissue was classified as RD. For patients with RD, the RCB class was categorized as RCB I, II, or III, as previously reported (15, 23, 24). In this study, the MD Anderson RCB calculator (www.mdanderson.org/breastcancer_RCB) was used for calculation.

Sequence data processing

Fresh-frozen pretreatment core needle tumor biopsies were collected from these cases using ultrasound guidance and sent to Novogene Co. Ltd. Sequencing data and proteomic data were generated by Novogene Co. Ltd. WES data were aligned to the human hg38 reference genome using BWA v0.7.8-455 (71) and was subject to marking duplication using Picard. WGBS data were aligned to GRCh38_Ensemble_94 human reference genome obtained from ftp://ftp.ensembl.org/pub/release-94/fasta/homo_sapiens/dna/ using BSMAP v2.90 (with parameters: -n 1 -r 0 -p 4 -S 1234) (72). After sorting the data with Samtools (1.3.1), Sambamba was used to mark duplicates using the command `markdup -l 1 -t 10 --sort-buffer-size 16000 --overflow-list-size 10000000`. Samples with an average sequencing depth below 20 were subsequently excluded. The remaining BAM files were then analyzed with `wgbstools` (73) and `Bedtools` (74) to generate a β value matrix comprising 28,292,717 CpG features. Positions with a depth lower than 10 were filtered out. RNA-seq data were aligned to GRCh38_Ensembl_97 human reference genome obtained from ftp://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/ using STAR v2.0.5 (75), and read counts, fragments per kilobase per million (FPKM), and transcripts per million (TPM) values were calculated with RNA-Seq by Expectation Maximization (RSEM) (76). SNP pairing tests were performed to ensure the paired samples from the same individual were matched using NGSCheckMate (77).

Somatic variant calling

Somatic mutations were identified from tumor and normal paired WES BAM files using the Mutect2 pipeline and were annotated using the Functotator embedded in GATK4 (29). Somatic copy numbers and tumor purity were calculated using FACETS with default hg19 version of bam files and parameters (78). Tumor purity was set to not available (NA) if not callable by FACETS. The thresholds of CNAs were 0.26 for gain, 0.80 for amplification, -0.32 for loss, and -0.73 for deletion. Focal copy number alterations were identified using GISTIC2 (35) with hg19 parameters: -genegistic 1 -smallmem 1 -broad 1 -brlen 0.5 -conf 0.90 -armpeel 1 -savegene 1 -gcm extreme.

Focal somatic copy number difference

First, cytobands demonstrating focal amplification and/or deletion from GISTIC2 (35) were aggregated for each subtype, stratified by pCR or RD status. Segment files were annotated using oncotator to extract coding gene-level $\log_2$ ratio. Only genes annotated in more than 80% of samples were kept. The median $\log_2$ ratio of all genes within the relevant cytoband was calculated to represent that cytoband's $\log_2$ ratio. Specifically, the $\log_2$ ratio, which indicates copy number changes, was determined for each individual using tumor alongside matched blood control WES data via the FACETS tool (78). Last, for each cytoband, the difference in the $\log_2$ ratio between the RD and pCR groups was assessed using the rank sum test for each subtype.

Mutation signature analysis

Somatic SNVs within coding regions of 94 samples with callable tumor purity were subjected to SigProfilerExtractor to extract the 96 context mutation matrix and perform de novo identification of the mutation signature (79). Then, the de novo signature was aligned and decomposed to the Catalogue Of Somatic Mutations In Cancer (COSMIC) signature to obtain the contribution of each COSMIC signature in each patient (80).

Methylation of the gene regulatory region

The Encyclopedia of DNA Elements (ENCODE) registry of candidate cis-regulatory elements (CCRE), such as CTCF, trimethylation of histone H3 at lysine 4 (H3K4me3), enhancer domain and enhancer-promoter interaction, and promoter regions (81), were obtained through the UCSC genome browser <https://genome.ucsc.edu/>. Promoter positions were defined as regions located from 1000 bp upstream of the TSSs to 1000 bp downstream of the TSSs. TSSs were obtained from the UCSC Table Browser (82). Methylation levels of defined gene regulatory regions were calculated as the average methylation β value of CpG sites within regions.

Differential methylation analysis

CpG methylation level data were segmented into 1,356,835 regions using the `wgbstools` segment command (73). Regions containing fewer than five CpG sites were excluded. The mean methylation level for each region was calculated per sample using the `wgbstools beta_to_table -c 10` command. DMRs were identified for each subtype using the Wilcoxon rank sum test on the resulting methylation matrix. The genomic annotation of DMRs was performed using the ChIPseeker R package (83) with the TxDb.Hsapiens.UCSC.hg38.knownGene database. Regions within the *CDKN2A* gene were specifically annotated to the *CDKN2A* locus. For segments annotated to a single gene, prioritization was performed using a two-step approach. First, segments were ranked on the basis of the Spearman correlation *P* value between segment methylation levels and corresponding RNA expression across all samples, with a focus on identifying significant negative associations. Subsequently, segments were further prioritized by evaluating the methylation difference between pCR and RD groups using the Wilcoxon rank sum test. Representative segments were selected on the basis of the significance level and mean β value differences between pCR and RD to form DMGs. Only DMGs exhibiting high variance (>50%) across all samples were retained for downstream analysis. Last, for verification, CpG-level information and genomic positions of both the representative DMRs and other segments showing significant methylation differences between pCR and RD, but not selected as representatives, were visualized and examined.

Biological network BRCANET

A list of breast cancer drivers, including genes in the collection named Breast Cancer with ID H00031 in the Kyoto Encyclopedia of Genes and Genomes human diseases database (38), and the elite genes in the collection named Breast Cancer with ID BRS047 in the Malacards database (39) were extracted. The neighbors of these genes from the protein-protein interaction database Search Tool for Retrieval of Interacting Genes/Proteins (STRING) (40) filtered by “experiments scores greater than 700” and the gene regulation database RegNetwork (41) filtered by “organism: human; evidence: experiment; confidence: high” were also extracted as breast cancer-related genes (table S6).

TCGA-independent methylation data

The TCGA breast cancer subtype information was obtained as previously defined (84), and only three receptor subtypes pertinent to this study were kept. The TCGA CpG site-level methylation data sequenced using the Illumina HM450K platform were downloaded from the LinkedOmics website <https://linkedomics.org/login.php> (85). For validation of *CDKN2A* methylation data difference regarding the DMR identified in the ER⁻HER2⁻, CpG site, cg14069088

(hg38, chr9: 21,996,207 to 21,996,209), overlapping with the *CDKN2A* DMR was selected.

Bisulfite pyrosequencing of *CDKN2A* promoter

Two regions, S1 with five CpGs and S2 with six CpGs, were analyzed by bisulfite pyrosequencing using PyroMark Q96/48 ID. For region S1 with sequence to be analyzed: GGYGGGGGATGGGGAAAGAGA-AGTTTGTYGTTTTTTTAGGTTATYGTATYGTTTTA-AGTGGGTGGTYYGTTAGTGTAGGTTTTATTTTTTTT, the forward polymerase chain reaction (PCR) primer was AGGGT-TTAGTTTTGTTTTGGGTTTGT, the reverse PCR primer was ACCCCCTTCTCAATAACTTCCTATTCA, and the sequencing primer was GGTAGTAAAGTTAAGTGATGG. For sequencing region S2 with sequence to be analyzed: YGTAAGAGTGATYGTAAATTT-TATTTTTTTTAAAGGGTTTYGTTTTGTTTTGGGTTTGATY-GAGGTTYGGTAGTAAYGTAAAGTGATGGGG, the forward PCR primer was AGTAAGGTGTTGGGTAAATTTTGAGTGAA, and the reverse PCR primer was ACAAACTTCTCTTCCCCATCC. Ten samples in the discovery cohort were also prepared for bisulfite pyrosequencing to perform quality control, of which nine samples were successfully sequenced for S1 and all samples were successfully sequenced for S2. Sequencing qualities were first used to filter CpG, and sites with more than half of the samples having good quality were kept. Methylation ratios calculated from bisulfite pyrosequencing results were compared with WGBS results, and successful sites were defined as those with Spearman correlation coefficients $R > 0.3$ between the two platforms.

Gene expression-based unsupervised clustering

TPM was log-transformed with base 2 after adding a pseudo-count of 0.01. The three-way ANOVA was performed to select genes for clustering, using the Python package statsmodels v0.12.2. Genes ranked top 100 for each term in ANOVA were used for clustering. Principal components analysis was implemented with the Python package Scipy v1.5.2 (86). Hierarchical clustering was performed using the Ward method and Pearson correlation as the distance metric and implemented with the Python package Scipy v1.5.2 (86).

Proteomic data processing

For the 137 patients who underwent proteomic quantification, proteins or phosphoproteins that were not expressed in more than 90% of samples were filtered out. Expression values were log-transformed with base 2 after adding a pseudo-count of 1. Quantile normalization was applied to the whole cohort to remove batch effects and ensure that all samples share the same distribution of protein/phosphoprotein expression.

Differential gene expression, protein, and phosphoprotein analysis

Differential analyses were conducted between patients with pCR and patients with RD within each subtype. For protein and phosphoprotein data, Wilcoxon rank sum tests were performed using the Python package Scipy v1.5.2, with fold change quantified by reversing the normalized values of proteins and phosphoproteins from log₂-transformed format back to their original scale (86). For proteins, a P value smaller than 0.05 in normalized data was considered significant. For phosphoproteins, the difference in mean values between pCR and RD was assessed, with $P < 0.05$ for normalized data regarded as significant. For gene expression data, DESeq2 (49) was applied to find differentially

expressed genes. GSEA for the preranked gene list derived from DESeq2 was conducted using fgsea (87).

Single-sample gene set enrichment analysis

Gene expression TPM values were adjusted by adding a pseudo-count of 0.01 and then log₂-transformed. After performing a z -score by gene, ssGSEA scores for Hallmark gene sets (88) were calculated for each sample using GSEAPy (52). For public datasets with normalized z -score values, pathway scores were calculated directly for each sample.

Integration analysis of multiomic data

Differential analysis results for methylation, transcription, protein, and phosphoprotein datasets were obtained as described in the above sections. These results were integrated using the ActivePathways R package (v2.0.5) with default parameters (89). Pathway enrichment analysis was conducted for each omic dataset as well as for the integrated multiomic data using the human Hallmark pathway collection (v2024.1). For the directional integration approach, the relationships between omic datasets were modeled as follows: Methylation was negatively correlated with transcription and protein, transcription was positively correlated with protein, and phosphoproteins were assumed to have no directional correlation with the other omic datasets.

Machine learning feature preselection

Feature selection was performed using study data to prioritize high-confidence features and pathway-level features to mitigate the potential model overfitting problem. For clinical features, those significantly associated with pCR in multivariate analysis ($P < 0.05$, logistic regression) were included across all subtypes. The Ki67 feature was significantly associated with pCR in the ER⁺HER2[−] subtype and was retained. For methylation data, only high-confidence DMGs were selected for each subtype to avoid model overfitting. Transcriptomic features were restricted to the top three significant ssGSEA pathways based on GSEA enrichment results to minimize false positives. For protein and phosphoprotein data, the top three ssGSEA hallmark pathways associated with pCR were selected on the basis of the sample-level comparison.

Machine learning model training

All discrete features were one-hot-encoded (tumor grade, 0 for T1/T2 and 1 for T3/T4; lymph node metastasis, 0 for no and 1 for yes). All continuous features were transformed to z -scores on the corresponding subtype. A boosted tree algorithm was used and implemented with LightGBM (59). The hyperparameters, the number of estimators T and learning rate α , were tuned in the 30-times-repeated five-fold cross-validation using Scikit-learn (90). The scope for T was set to {50, 100, 200}, and the scope for α was set to {0.05, 0.1, 0.2}. The optimized models were selected with the highest average AUC in repeated cross-validation. The threshold for feature selection was set to 0.3 for all three subtypes. SHAP values (60) were calculated for the model explanation.

We implemented a systematic feature selection approach during cross-validation to optimize model stability and predictive performance in our multiomic integration framework. The strategy began with the segregation of clinical features, which were maintained as essential baseline predictors, while nonclinical features were subjected to comprehensive combinatorial analysis. We generated an

exhaustive set of feature combinations through power set computation, systematically evaluating each possible combination ranging from individual features to the complete set of nonclinical predictors. Each feature combination was assessed using a rigorous cross-validation framework, where model performance was evaluated on both training and validation sets. The selection criteria prioritized feature combinations that minimized the absolute difference between training and validation set AUC while maintaining validation performance superior to the best single-omic model. To ensure robustness, we tracked the selection frequency of each feature across all cross-validation iterations, ultimately retaining features that demonstrated consistent selection in a certain threshold. This approach effectively balanced the trade-off between model complexity and generalizability while maintaining the biological relevance of selected features.

Feature imputation for independent cohorts

A LASSO-based imputation method was developed to address missing features during independent validation. The GDPH cohort gene expression matrix was used as the initial dataset. After quantile normalization and $\log_2$ transformation, gene filtering was performed on the basis of mean expression values, and the resulting set was intersected with breast cancer-associated genes. This process yielded several hundred candidate genes, whose expression matrix was subsequently used to construct a LASSO cross-validation model. The model was designed to predict various missing target features, such as CDKN2A methylation status. To assess potential overfitting during cross-validation, ~30% of patient samples were randomly allocated to an independent validation cohort. Model performance was evaluated using Spearman correlation coefficients and their corresponding *P* values. For those imputed features whose Spearman correlation is between 0 and 0.1 during independent validation, we manually set them NA.

Statistical analysis

The Wilcoxon rank sum and Student's *t* test were used for quantitative pair-wise comparisons. A Kruskal-Wallis test was used for multiple-group quantitative comparison. A chi-square test was used for comparing categorical variables. Pearson or Spearman methods were used for calculating the correlation. The Kaplan-Meier method was used for survival data comparisons. An AUC value was calculated using pROC (87) and plotROC (88) R packages.

Supplementary Materials

The PDF file includes:

Figs. S1 to S17

Legends for tables S1 to S6

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S6

REFERENCES AND NOTES

- S. Lei, R. Zheng, S. Zhang, S. Wang, R. Chen, K. Sun, H. Zeng, J. Zhou, W. Wei, Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Commun.* **41**, 1183–1194 (2021).
- L. Fan, K. Strasser-Weippl, J.-J. Li, J. St Louis, D. M. Finkelstein, K.-D. Yu, W.-Q. Chen, Z.-M. Shao, P. E. Goss, Breast cancer in China. *Lancet Oncol.* **15**, e279–e289 (2014).
- X. Xu, W. Zhao, C. Liu, Y. Gao, D. Chen, M. Wu, C. Li, X. Wang, X. Song, J. Yu, Z. Liu, Z. Yu, The residual cancer burden index as a valid prognostic indicator in breast cancer after neoadjuvant chemotherapy. *BMC Cancer* **24**, 13 (2024).
- G. von Minckwitz, C.-S. Huang, M. S. Mano, S. Loibl, E. P. Mamounas, M. Untch, N. Wolmark, P. Rastogi, A. Schneeweiss, A. Redondo, H. H. Fischer, W. Jacot, A. K. Conlin, C. Arce-Salinas, I. L. Wapnir, C. Jackisch, M. P. DiGiovanna, P. A. Fasching, J. P. Crown, W. Wülfing, Z. Shao, E. R. Caremoli, H. Wu, L. H. Lam, D. Tesarowski, M. Smitt, H. Douthwaite, S. M. Singel, C. E. Geyer Jr., KATHERINE Investigators, Trastuzumab emtansine for residual invasive HER2-positive breast cancer. *N. Engl. J. Med.* **380**, 617–628 (2019).
- L. M. Spring, G. Fell, A. Arfe, C. Sharma, R. Greenup, K. L. Reynolds, B. L. Smith, B. Alexander, B. Moy, S. J. Isakoff, G. Parmigiani, L. Trippa, A. Bardia, Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: A comprehensive meta-analysis. *Clin. Cancer Res.* **26**, 2838–2848 (2020).
- B. Asselain, W. Barlow, J. Bartlett, J. Bergh, E. Bergsten-Nordström, J. Bliss, F. Boccardo, C. Boddington, J. Bogaerts, G. Bonadonna, R. Bradley, E. Brain, J. Braybrooke, P. Broet, J. Bryant, J. Burrett, D. Cameron, M. Clarke, A. Coates, R. Coleman, R. C. Coombes, C. Correa, J. Costantino, J. Cuzick, D. Danforth, N. Davidson, C. Davies, L. Davies, A. Di Leo, D. Dodwell, M. Dowsett, F. Duane, V. Evans, M. Ewertz, B. Fisher, J. Forbes, L. Ford, J.-C. Gazet, R. Gelber, L. Gettins, L. Gianni, M. Gnant, J. Godwin, A. Goldhirsch, P. Goodwin, R. Gray, D. Hayes, C. Hill, J. Ingle, R. Jagsi, R. Jakesz, S. James, W. Janni, H. Liu, Z. Liu, C. Lohrisch, S. Loibl, L. MacKinnon, A. Makris, E. Mamounas, G. Mannu, M. Martin, S. Mathoulin, L. Mauriac, P. McGale, T. McHugh, P. Morris, H. Mukai, L. Norton, Y. Ohashi, I. Olivetto, S. Paik, H. Pan, R. Peto, M. Piccart, L. Pierce, P. Poortmans, T. Powles, K. Pritchard, J. Ragaz, V. Raina, P. Ravdin, S. Read, M. Regan, J. Robertson, E. Rutgers, S. Scholl, D. Slamon, L. Sölkner, J. Sparano, S. Steinberg, R. Sutcliffe, S. Swain, C. Taylor, A. Tutt, P. Valagussa, C. van de Velde, J. van der Hage, G. Viale, G. von Minckwitz, Y. Wang, Z. Wang, X. Wang, T. Whelan, N. Wilcken, E. Winer, N. Wolmark, W. Wood, M. Zambetti, J. A. Zujewski, Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: Meta-analysis of individual patient data from ten randomised trials. *Lancet Oncol.* **19**, 27–39 (2018).
- J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, P. S. Bernard, Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
- The Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- M. Colleoni, G. Viale, D. Zahrieh, L. Bottiglieri, R. D. Gelber, P. Veronesi, A. Balduzzi, R. Torrisi, A. Luini, M. Intra, S. Dellapasqua, A. Cardillo, R. Ghisini, G. Peruzzotti, A. Goldhirsch, Expression of ER, PgR, HER1, HER2, and response: A study of preoperative chemotherapy. *Ann. Oncol.* **19**, 465–472 (2008).
- C. Hatzis, L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, M. Martin, J. Cotrina, H. Gomez, R. Hubbard, J. I. Chacón, J. Ferrer-Lozano, R. Dyer, M. Buxton, Y. Gong, Y. Wu, N. Ibrahim, E. Andreopoulou, N. T. Ueno, K. Hunt, W. Yang, A. Nazario, A. DeMichele, J. O'Shaughnessy, G. N. Hortobagyi, W. F. Symmans, A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
- A. Prat, V. Guarneri, T. Pascual, F. Brasó-Maristany, E. Sanfeliu, L. Paré, F. Schettini, D. Martínez, P. Jares, G. Griguolo, M. V. Dieci, J. Cortés, A. Llombart-Cussac, B. Conte, M. Marín-Aguilera, N. Chic, J. A. Puig-Butillé, A. Martínez, P. Galván, Y.-H. Tsai, B. González-Farré, A. Mira, A. Vivancos, P. Villagrás, J. S. Parker, P. Conte, C. M. Perou, Development and validation of the new HER2DX assay for predicting pathological response and survival outcome in early-stage HER2-positive breast cancer. *eBioMedicine* **75**, 103801 (2022).
- M. Martín, S. R. Stecklein, O. Gluz, G. Villacampa, M. Monte-Millán, U. Nitz, S. Cobo, M. Christgen, F. Brasó-Maristany, E. L. Álvarez, I. Echavarría, B. Conte, S. Kuemmel, C. Bueno-Muñoz, Y. Jerez, R. Kates, M. Cebollero, C. Kolberg-Liedtke, O. Bueno, J. Á. García-Saenz, F. Moreno, E.-M. Grischke, H. Forstbauer, M. Braun, M. Warm, J. Hackmann, C. Uleer, B. Aktas, C. Schumacher, R. Wuerstleins, M. Graeser, C. Zu Eulenburg, H. H. Kreipe, H. Gómez, T. Massarrah, B. Herrero, L. Paré, U. Bohn, S. López-Tarruella, A. Vivancos, E. Sanfeliu, J. S. Parker, C. M. Perou, P. Villagrás, A. Prat, P. Sharma, N. Harbeck, TNBC-DX genomic test in early-stage triple-negative breast cancer treated with neoadjuvant taxane-based therapy. *Ann. Oncol.* **36**, 158–171 (2025).
- C. Denkert, G. von Minckwitz, S. Darb-Esfahani, B. Lederer, B. I. Heppner, K. E. Weber, J. Budczies, J. Huober, F. Klauschen, J. Furlanetto, W. D. Schmitt, J.-U. Blohmer, T. Karn, B. M. Pfitzner, S. Kümmel, K. Engels, A. Schneeweiss, A. Hartmann, A. Noske, P. A. Fasching, C. Jackisch, M. van Mackelenbergh, P. Sinn, C. Schem, C. Hanusch, M. Untch, S. Loibl, Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **19**, 40–50 (2018).
- M. Tanioka, C. Fan, J. S. Parker, K. A. Hoadley, Z. Hu, Y. Li, T. M. Hyslop, B. N. Pitcher, M. G. Soloway, P. A. Spears, L. N. Henry, S. Toloney, C. T. Dang, I. E. Krop, L. N. Harris, D. A. Berry, E. R. Mardis, E. P. Winer, C. A. Hudis, L. A. Carey, C. M. Perou, Integrated analysis of RNA and DNA from the phase III trial CALGB 40601 identifies predictors of response to trastuzumab-based neoadjuvant chemotherapy in HER2-positive breast cancer. *Clin. Cancer Res.* **24**, 5292–5304 (2018).
- S. J. Sammut, M. Crispin-Ortuzar, S. F. Chin, E. Provenzano, H. A. Bardwell, W. Ma, W. Cope, A. Dariush, S. J. Dawson, J. E. Abraham, J. Dunn, L. Hiller, J. Thomas, D. A. Cameron,

- J. M. S. Bartlett, L. Hayward, P. D. Pharoah, F. Markowitz, O. M. Rueda, H. M. Earl, C. Caldas, Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
16. L. Chen, L. Yang, L. Yao, X.-Y. Kuang, W.-J. Zuo, S. Li, F. Qiao, Y.-R. Liu, Z.-G. Cao, S.-L. Zhou, X.-Y. Zhou, W.-T. Yang, J.-X. Shi, W. Huang, X. Hu, Z.-M. Shao, Characterization of PIK3CA and PIK3R1 somatic mutations in Chinese breast cancer patients. *Nat. Commun.* **9**, 1357 (2018).
17. Y.-Z. Jiang, D. Ma, C. Suo, J. Shi, M. Xue, X. Hu, Y. Xiao, K.-D. Yu, Y.-R. Liu, Y. Yu, Y. Zheng, X. Li, C. Zhang, P. Hu, J. Zhang, Q. Hua, J. Zhang, W. Hou, L. Ren, D. Bao, B. Li, J. Yang, L. Yao, W.-J. Zuo, S. Zhao, Y. Gong, Y.-X. Ren, Y.-X. Zhao, Y.-S. Yang, Z. Niu, Z.-G. Cao, D. G. Stover, C. Verschraegen, V. Kaklamani, A. Daemen, J. R. Benson, K. Takabe, F. Bai, D.-Q. Li, P. Wang, L. Shi, W. Huang, Z.-M. Shao, Genomic and transcriptomic landscape of triple-negative breast cancers: Subtypes and treatment strategies. *Cancer Cell* **35**, 428–440.e5 (2019).
18. J. Kaur, A. Matta, I. Kak, G. Srivastava, J. Assi, I. Leong, I. Witterick, T. J. Colgan, C. MacMillan, K. W. M. Siu, P. G. Walfish, R. Ralhan, S100A7 overexpression is a predictive marker for high risk of malignant transformation in oral dysplasia. *Int. J. Cancer* **134**, 1379–1388 (2014).
19. T. B. M. Schaaïj-Visser, A. P. Graveland, S. Gauci, B. J. M. Braakhuis, M. Buijze, A. J. R. Heck, D. J. Kuik, E. Bloemen, C. R. Leemans, M. Slijper, R. H. Brakenhoff, Differential proteomics identifies protein biomarkers that predict local relapse of head and neck squamous cell carcinomas. *Clin. Cancer Res.* **15**, 7666–7675 (2009).
20. R. Henrique, C. Jerónimo, Molecular detection of prostate cancer: A role for GSTP1 hypermethylation. *Eur. Urol.* **46**, 660–669 (2004).
21. M. P. A. Ebert, M. Tänzer, B. Balluff, E. Burgermeister, A. K. Kretschmar, D. J. Hughes, R. Tetzner, C. Lofton-Day, R. Rosenberg, A. C. Reinacher-Schick, K. Schulmann, A. Tannapfel, R. Hofheinz, C. Röcken, G. Keller, R. Langer, K. Specht, R. Porschen, J. Stöhlmacher-Williams, T. Schuster, P. Ströbel, R. M. Schmid, TFAP2E–DKK4 and chemoresistance in colorectal cancer. *N. Engl. J. Med.* **366**, 44–53 (2024).
22. M. Esteller, J. Garcia-Foncillas, E. Andion, S. N. Goodman, O. F. Hidalgo, V. Vanaclocha, S. B. Baylin, J. G. Herman, Inactivation of the DNA–repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343**, 1350–1354 (2024).
23. W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, A. U. Buzdar, S. E. Singletary, G. N. Hortobagyi, L. Pusztai, Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J. Clin. Oncol.* **25**, 4414–4422 (2007).
24. W. F. Symmans, C. Wei, R. Gould, X. Yu, Y. Zhang, M. Liu, A. Walls, A. Bousamra, M. Ramineni, B. Sinn, K. Hunt, T. A. Buchholz, V. Valero, A. U. Buzdar, W. Yang, A. M. Brewster, S. Moulder, L. Pusztai, C. Hatzis, G. N. Hortobagyi, Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J. Clin. Oncol.* **35**, 1049–1060 (2017).
25. L. A. Carey, E. C. Dees, L. Sawyer, L. Gatti, D. T. Moore, F. Collichio, D. W. Ollila, C. I. Sartor, M. L. Graham, C. M. Perou, The triple negative paradox: Primary tumor chemosensitivity of breast cancer subtypes. *Clin. Cancer Res.* **13**, 2329–2334 (2007).
26. P. Cortazar, L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N. Wolmark, H. Bonnefoi, D. Cameron, L. Gianni, P. Valagussa, S. M. Swain, T. Prowell, S. Loibl, D. L. Wickerham, J. Bogaerts, J. Baselga, C. Perou, G. Blumenthal, J. Blohmer, E. P. Mamounas, J. Bergh, V. Semiglazov, R. Justice, H. Eidtmann, S. Paik, M. Piccart, R. Sridhara, P. A. Fasching, L. Slaets, S. Tang, B. Gerber, C. E. Geyer, R. Pazdur, N. Ditsch, P. Rastogi, W. Eiermann, G. von Minckwitz, Pathological complete response and long-term clinical benefit in breast cancer: The CTNeoBC pooled analysis. *Lancet* **384**, 164–172 (2014).
27. C. Denkert, S. Loibl, B. M. Müller, H. Eidtmann, W. D. Schmitt, W. Eiermann, B. Gerber, H. Tesch, J. Hilfrich, J. Huober, T. Fehm, J. Barinoff, C. Jackisch, J. Prinzler, T. Rüdiger, E. Erbstößer, J. U. Blohmer, J. Budczies, K. M. Mehta, G. von Minckwitz, Ki67 levels as predictive and prognostic parameters in pretherapeutic breast cancer core biopsies: A translational investigation in the neoadjuvant GeparTrio trial. *Ann. Oncol.* **24**, 2786–2793 (2013).
28. B. Fisher, A. Brown, E. Mamounas, S. Wieand, A. Robidoux, R. G. Margoless, A. B. Cruz, E. R. Fisher, D. L. Wickerham, N. Wolmark, A. DeCillis, J. L. Hoehn, A. W. Lees, N. V. Dimitrov, Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: Findings from national surgical adjuvant breast and bowel project B-18. *J. Clin. Oncol.* **15**, 2483–2493 (2024).
29. A. Mckenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytzky, K. Garimella, D. Altschuler, S. Gabriel, M. Daly, M. A. Depristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
30. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendt, J. Kim, B. Reardon, P. K.-S. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortés-Ciriano, D. C. Zhou, W.-W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphaviat, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang, M3 Working Group, Cancer Genome Atlas Research Network, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
31. S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjærde, A. Langerød, M. Ringnér, S.-M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. J. Hooijer, S. J. Jang, D. R. Jones, H.-Y. Kim, T. A. King, S. Krishnamurthy, H. J. Lee, J.-Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O'Meara, I. Pauporté, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodríguez-González, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Thomasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van't Veer, A. Tutt, S. Knappskog, B. K. T. Tan, J. Jonkers, Å. Borg, N. T. Ueno, C. Sotiropoulos, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. M. Martens, A.-L. Børresen-Dale, A. L. Richardson, G. Kong, G. Thomas, M. R. Stratton, Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
32. Y.-Z. Jiang, D. Ma, X. Jin, Y. Xiao, Y. Yu, J. Shi, Y.-F. Zhou, T. Fu, C.-J. Lin, L.-J. Dai, C.-L. Liu, S. Zhao, G.-H. Su, W. Hou, Y. Liu, Q. Chen, J. Yang, N. Zhang, W.-J. Zhang, W. Liu, W. Ge, W.-T. Yang, C. You, Y. Gu, V. Kaklamani, F. Bertucci, C. Verschraegen, A. Daemen, N. M. Shah, T. Wang, T. Guo, L. Shi, C. M. Perou, Y. Zheng, W. Huang, Z.-M. Shao, Integrated multiomic profiling of breast cancer in the Chinese population reveals patient stratification and therapeutic vulnerabilities. *Nat. Cancer* **5**, 673–690 (2024).
33. M. Kamal, D. L. Holliday, E. E. Morrison, V. Speirs, C. Toomes, S. M. Bell, Loss of CSMD1 expression disrupts mammary duct formation while enhancing proliferation, migration and invasion. *Oncol. Rep.* **38**, 283–292 (2017).
34. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, P. Boutros, K. Chan, A. Fujimoto, G. Getz, M. N. Huang, M. Kazanov, M. Lawrence, I. Martincorena, S. Morganella, H. Nakagawa, P. Polak, S. Prokopc, S. A. Roberts, S. G. Rozen, N. Saini, T. Shibata, Y. Shiraishi, M. R. Stratton, B. T. Teh, I. Vázquez-García, F. Yousif, W. Yu, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
35. C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, G. Getz, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
36. C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, S. Aparicio, J. D. Brenton, I. Ellis, D. Huntsman, S. Pinder, L. Murphy, H. Bardwell, Z. Ding, L. Jones, B. Liu, I. Papatheodorou, S. J. Sammut, G. Wishart, S. Chia, K. Gelmon, C. Speers, P. Watson, R. Blamey, A. Green, D. MacMillan, E. Rakha, C. Gillett, A. Grigoriadis, E. De Rinaldis, A. Tutt, M. Parisien, S. Troup, D. Chan, C. Fielding, A. T. Maia, S. McGuire, M. Osborne, S. M. Sayalero, I. Spiteri, J. Hadfield, L. Bell, K. Chow, N. Gale, M. Kovalik, Y. Ng, L. Prentice, S. Tavaré, F. Markowitz, A. Langerød, E. Provenzano, A. Purushotham, A. L. Børresen-Dale, C. Caldas, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
37. X. Chen, J. Zhang, X. Dai, DNA methylation profiles capturing breast cancer heterogeneity. *BMC Genomics* **20**, 823 (2019).
38. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
39. N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. Iny Stein, I. Bahir, F. Belinky, C. P. Morrey, M. Safran, D. Lancet, MalaCards: An integrated compendium for diseases and their annotation. *Database* **2013**, bat018 (2013).
40. C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, P. Bork, STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).
41. Z.-P. Liu, C. Wu, H. Miao, H. Wu, RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, bav095 (2015).
42. J. C. Kim, J. S. Choi, S. A. Roh, D. H. Cho, T. W. Kim, Y. S. Kim, Promoter methylation of specific genes is associated with the phenotype and progression of colorectal adenocarcinomas. *Ann. Surg. Oncol.* **17**, 1767–1776 (2010).
43. A. Ameri, A. Alidoosti, Y. Hosseini, M. Parvin, M. H. Emranpour, F. Taslimi, E. Salehi, P. Fadavi, Prognostic value of promoter hypermethylation of retinoic acid receptor beta (RARβ) and CDKN2 (p16/MTS1) in prostate cancer. *Chin. J. Cancer Res.* **23**, 306–311 (2011).
44. D. M. Wolf, C. Yau, J. Wulfschuh, L. Brown-Swigart, R. I. Gallagher, P. R. E. Lee, Z. Zhu, M. J. Magbanua, R. Sayaman, N. O'Grady, A. Basu, A. Delson, J. P. Coppé, R. Lu, J. Braun, S. M. Asare, L. Sit, J. B. Matthews, J. Perlmutter, N. Hylton, M. C. Liu, P. Pohlmann, W. F. Symmans, H. S. Rugo, C. Isaacs, A. M. DeMichele, D. Yee, D. A. Berry, L. Pusztai, E. F. Petricoin, G. L. Hirst, L. J. Esserman, L. J. van't Veer, Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. *Cancer Cell* **40**, 609–623.e6 (2022).
45. M. Hoogstraet, E. H. Lips, I. Mayayo-Peralta, L. Mulder, P. Kristel, I. van der Heijden, S. Annunziato, M. van Seijen, P. M. Nederlof, G. S. Sonke, W. Zwart, J. Wesseling,

- L. F. A. Wessels, Comprehensive characterization of pre- and post-treatment samples of breast cancer reveal potential mechanisms of chemotherapy resistance. *NPJ Breast Cancer* **8**, 60 (2022).
46. S. Loibl, J. O'Shaughnessy, M. Untch, W. M. Sikov, H. S. Rugo, M. D. McKee, J. Huober, M. Golschan, G. von Minckwitz, D. Maag, D. Sullivan, N. Wolmark, K. McIntyre, J. J. Ponce Lorenzo, O. Metzger Filho, P. Rastogi, W. F. Symmans, X. Liu, C. E. Geyer, Addition of the PARP inhibitor veliparib plus carboplatin or carboplatin alone to standard neoadjuvant chemotherapy in triple-negative breast cancer (BrighTNess): A randomised, phase 3 trial. *Lancet Oncol.* **19**, 497–509 (2018).
 47. M. Anurag, E. J. Jaehnig, K. Krug, J. T. Lei, E. J. Bergstrom, B. J. Kim, T. D. Vashist, A. M. T. Huynh, Y. Dou, X. Gou, C. Huang, Z. Shi, B. Wen, V. Korchina, R. A. Gibbs, D. M. Muzny, H. Doddapaneni, L. E. Dobrolecki, H. Rodriguez, A. I. Robles, T. Hiltke, M. T. Lewis, J. R. Nangia, M. N. Shafae, S. Li, I. S. Hagemann, J. Hoog, B. Lim, C. K. Osborne, D. R. Mani, M. A. Gillette, B. Zhang, G. V. Echeverria, G. Miles, M. F. Rimawi, S. A. Carr, F. O. Ademuyiwa, S. Satpathy, M. J. Ellis, Proteogenomic markers of chemotherapy resistance and response in triple-negative breast cancer. *Cancer Discov.* **12**, 2586–2605 (2022).
 48. S. A. Hurvitz, J. L. Caswell-Jin, K. L. McNamara, J. J. Zoeller, G. R. Bean, R. Dichmann, A. Perez, R. Patel, L. Zehngebot, H. Allen, L. Bosserman, B. DiCarlo, A. Kennedy, A. Giuliano, C. Calfa, D. Molthrop, A. Mani, H.-W. Chen, J. Dering, B. Adams, E. Kotler, M. F. Press, J. S. Brugge, C. Curtis, D. J. Slamon, Pathologic and molecular responses to neoadjuvant trastuzumab and/or lapatinib from a phase II randomized trial in HER2-positive breast cancer (TRIO-US B07). *Nat. Commun.* **11**, 5824 (2020).
 49. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
 50. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
 51. S. A. Balaji, N. Udupa, M. R. Chamallamudi, V. Gupta, A. Rangarajan, Role of the drug transporter ABCC3 in breast cancer chemoresistance. *PLOS ONE* **11**, e0155013 (2016).
 52. Z. Fang, X. Liu, G. Peltz, GSEAPy: A comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
 53. S. Bekker-Jensen, J. R. Danielsen, K. Fugger, I. Gromova, A. Nerstedt, C. Lukas, J. Bartek, J. Lukas, N. Mailand, HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. *Nat. Cell Biol.* **12**, 80–86 (2010).
 54. T. M. Bernaciak, J. Zareno, J. T. Parsons, C. M. Silva, A novel role for signal transducer and activator of transcription 5b (STAT5b) in β 1-integrin-mediated human breast cancer cell migration. *Breast Cancer Res.* **11**, R52 (2009).
 55. G. Yang, S. J. Scherer, S. S. Shell, K. Yang, M. Kim, M. Lipkin, R. Kucherlapati, R. D. Kolodner, W. Edelmann, Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. *Cancer Cell* **6**, 139–150 (2004).
 56. A. Szczepny, K. Carey, L. McKenzie, W. S. N. Jayasekara, F. Rossello, A. Gonzalez-Rajal, A. S. McCaw, D. Popovski, D. Wang, A. J. Sadler, A. Mahar, P. A. Russell, G. Wright, R. A. McCloy, D. J. Garama, D. J. Gough, S. B. Baylin, A. Burgess, J. E. Cain, D. N. Watkins, The tumor suppressor Hic1 maintains chromosomal stability independent of Tp53. *Oncogene* **37**, 1939–1948 (2018).
 57. R. Danesh Pazhooh, P. Rahnamay Farnood, Z. Asemi, L. Mirsaefaei, B. Yousefi, H. Mirzaei, mTOR pathway and DNA damage response: A therapeutic strategy in cancer therapy. *DNA Repair (Amst)*. **104**, 103142 (2021).
 58. J. C. Brase, M. Schmidt, T. Fischbach, H. Sultmann, H. Bojar, H. Koelbl, B. Hellwig, J. Rahnenfuhrer, J. G. Hengstler, M. C. Gehrmann, ERBB2 and TOP2A in breast cancer: A comprehensive analysis of gene amplification, RNA levels, and protein expression and their influence on prognosis and prediction. *Clin. Cancer Res.* **16**, 2391–2401 (2010).
 59. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree" in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), NIPS'17, pp. 3149–3157.
 60. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
 61. B. K. Killelea, V. Q. Yang, S.-Y. Wang, B. Hayse, S. Mougalian, N. R. Horowitz, A. B. Chagpar, L. Pusztai, D. R. Lannin, Racial differences in the use and outcome of neoadjuvant chemotherapy for breast cancer: Results from the national cancer data base. *J. Clin. Oncol.* **33**, 4267–4276 (2015).
 62. R. Yerushalmi, R. Woods, P. M. Ravdin, M. M. Hayes, K. A. Gelmon, Ki67 in breast cancer: Prognostic and predictive potential. *Lancet Oncol.* **11**, 174–183 (2010).
 63. S. E. Stanton, M. L. Disis, Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J. Immunother. Cancer* **4**, 59 (2016).
 64. R. Salgado, C. Denkert, C. Campbell, P. Savas, P. Nuciforo, C. Aura, E. de Azambuja, H. Eidtmann, C. E. Ellis, J. Baselga, M. J. Piccart-Gebhart, S. Michiels, I. Bradbury, C. Sotiriou, S. Loi, Tumor-infiltrating lymphocytes and associations with pathological complete response and event-free survival in HER2-positive early-stage breast cancer treated with lapatinib and trastuzumab: A secondary analysis of the NeoALTTO trial. *JAMA Oncol.* **1**, 448–454 (2015).
 65. G. Bianchini, L. Pusztai, T. Pienkowski, Y.-H. Im, G. V. Bianchi, L.-M. Tseng, M.-C. Liu, A. Lluch, E. Galeota, D. Magazzù, J. de la Haba-Rodriguez, D.-Y. Oh, B. Poirier, J. L. Pedrini, V. Semiglazov, P. Valagussa, L. Gianni, Immune modulation of pathologic complete response after neoadjuvant HER2-directed therapies in the NeoSphere trial. *Ann. Oncol.* **26**, 2429–2436 (2015).
 66. M. Ignatiadis, G. Van den Eynden, S. Roberto, M. Fornili, Y. Bareche, C. Desmedt, F. Rothé, M. Maetens, D. Venet, E. Holgado, V. McNally, A. Kiermaier, H. M. Savage, T. R. Wilson, J. Cortes, A. Schneeweiss, K. Willard-Gallo, E. Biganzoli, C. Sotiriou, Tumor-infiltrating lymphocytes in patients receiving trastuzumab/pertuzumab-based chemotherapy: A TRYPHAEANA substudy. *JNCI J. Natl. Cancer Inst.* **111**, 69–77 (2019).
 67. N. Tung, N. U. Lin, J. Kidd, B. A. Allen, N. Singh, R. J. Wenstrup, A.-R. Hartman, E. P. Winer, J. E. Garber, Frequency of germline mutations in 25 cancer susceptibility genes in a sequential series of patients with breast cancer. *J. Clin. Oncol.* **34**, 1460–1468 (2016).
 68. Y. Li, N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korbel, J. E. Haber, M. Imielinski, PCAWG Structural Variation Working Group, J. Weischenfeldt, R. Beroukhim, P. J. Campbell, PCAWG Consortium, Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
 69. A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillip, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
 70. D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggins, J. A. Schloss, The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
 71. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 72. Y. Xi, W. Li, BSMAP: Whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics* **10**, 232 (2009).
 73. N. Loyfer, J. Magenheimer, A. Peretz, G. Cann, J. Bredno, A. Klochendler, I. Fox-Fisher, S. Shabi-Porat, M. Hecht, T. Pelet, J. Moss, Z. Drawshy, H. Amini, P. Moradi, S. Nagaraju, D. Bauman, D. Shveiky, S. Porat, U. Dior, G. Rivkin, O. Or, N. Hirshoren, E. Carmon, A. Pikarsky, A. Khalailah, G. Zamir, R. Grinbaum, M. Abu Gazala, I. Mizrahi, N. Shussman, A. Korach, O. Wald, U. Izhar, E. Erez, V. Yutkin, Y. Samet, D. Rotnemer Golinkin, K. L. Spalding, H. Druid, P. Arner, A. M. J. Shapiro, M. Grompe, A. Aravanis, O. Venn, A. Jamshidi, R. Shemer, Y. Dor, B. Glaser, T. Kaplan, A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
 74. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
 75. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
 76. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
 77. S. Lee, S. Lee, S. Ouellette, W.-Y. Park, E. A. Lee, P. J. Park, NGSCheckMate: Software for validating sample identity in next-generation sequencing studies within and across data types. *Nucleic Acids Res.* **45**, –e103 (2017).
 78. R. Shen, V. E. Seshan, FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131–e131 (2016).
 79. S. M. A. Islam, M. Diaz-Gay, Y. Wu, M. Barnes, R. Vangara, E. N. Bergstrom, Y. He, M. Vella, J. Wang, J. W. Teague, P. Clapham, S. Moody, S. Senkin, Y. R. Li, L. Riva, T. Zhang, A. J. Gruber, C. D. Steele, B. Otlu, A. Khandekar, A. Abbasi, L. Humphreys, N. Syulyukina, S. W. Brady, B. S. Alexandrov, N. Pillay, J. Zhang, D. J. Adams, I. Martincorena, D. C. Wedge, M. T. Landi, P. Brennan, M. R. Stratton, S. G. Rozen, L. B. Alexandrov, Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
 80. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
 81. The ENCODE Project Consortium, J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shores, J. Adrian, T. Kawi, C. A. Davis, A. Dobin, R. Kaul, J. Halow, E. L. Van Nostrand, P. Freese, D. U. Gorkin, Y. Shen, Y. He, M. Mackiewicz, F. Pauli-Behn, B. A. Williams, A. Mortazavi, C. A. Keller, X.-O. Zhang, S. I. Elhajjaj, J. Huey, D. E. Dickel, V. Snetkova, X. Wei, X. Wang, J. C. Rivera-Mulia, J. Rozowsky, J. Zhang, S. B. Chhetri, J. Zhang, A. Victorson, K. P. White, A. Visel, G. W. Yeo, C. B. Burge, E. Lécuyer, D. M. Gilbert, J. Dekker,

- J. Rinn, E. M. Mendenhall, J. R. Ecker, M. Kellis, R. J. Klein, W. S. Noble, A. Kundaje, R. Guigó, P. J. Farnham, J. M. Cherry, R. M. Myers, B. Ren, B. R. Graveley, M. B. Gerstein, L. A. Pennacchio, M. P. Snyder, B. E. Bernstein, B. Wold, R. C. Hardison, T. R. Gingeras, J. A. Stamatoyannopoulos, Z. Weng, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
82. D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent, The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, 493D–4496D (2004).
83. G. Yu, L.-G. Wang, Q.-Y. He, ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
84. B. D. Lehmann, A. Colaprico, T. C. Silva, J. Chen, H. An, Y. Ban, H. Huang, L. Wang, J. L. James, J. M. Balko, P. I. Gonzalez-Ericsson, M. E. Sanders, B. Zhang, J. A. Pietersen, X. S. Chen, Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes. *Nat. Commun.* **12**, 6276 (2021).
85. S. V. Vasaika, P. Straub, J. Wang, B. Zhang, LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **46**, D956–D963 (2018).
86. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
87. G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, A. Sergushichev, Fast gene set enrichment analysis. bioRxiv 60012 [Preprint] (2021). <https://doi.org/10.1101/060012>.
88. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
89. M. Slobodyanyuk, A. T. Bahcheli, Z. P. Klein, M. Bayati, L. J. Strug, J. Reimand, Directional integration and pathway enrichment analysis for multi-omics data. *Nat. Commun.* **15**, 5690 (2024).
90. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Development Joint Fund of National Natural Science Foundation of China (no. U22A20345 to Z.L.), the Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (no. 2022B1212010011 to Z.L.), the National Key R&D Program of China (no. 2021YFF1201003 to Z.L.), the National Natural Science Foundation of China (grant nos. 82373427 and 82072939 to M.Y.), the Guangzhou Science and Technology Program key projects (grant no. 202206010110 to M.Y.), and the Guangdong Basic and Applied Basic Research Foundation (grant no. 2022A1515010425 to M.Y.). This work was supported in part by RGC CRF grants (CRF C6034-21G, C4008-23W, and T12-705/24R to J.W.), the ITC grant (ITCPD/17-9 to J.W.), and the Seed fund of the Big Data for Bio-Intelligence Laboratory (Z0428 to J.W.). J.W. was also supported by the Padma Harilela Professorship. **Author contributions:** J.W., Z.L., C.L., and K.W. initialized the study; J.W. and Z.L. conceived and supervised the study. J.W., Z.L., and M.Y. obtained the funding. Z.L., K.W., C.L., M.Y., H.L., Z.X., S. Li, and X.C. obtained the patient samples and collected the multiomic data of the GDPH and the pyrosequencing cohorts; Z.L., M.Y., S. Ly., and F.X. collected and curated clinical data and pCR/RCB labels; J.W. supervised the computational analyses; Z. Zhu and Z.M. carried out most of the bioinformatic analyses and visualization; M.W. developed the imputation model and contributed to machine learning; H.W. contributed to Fig. 5D; Z.Zha. contributed to website updates; H.S. contributed to data analysis; Z.M., Z. Zhu, Z.L., and J.W. drafted the original manuscript, with review and editing from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The raw WES data and transcriptomic data have been deposited in the National Genomics Data Center (GSA) under accession code HRA000428, available at <https://ngdc.cncb.ac.cn/gsa-human/browse/HRA000428>. The raw sequencing data can be provided by the GDPH upon scientific review and a completed material transfer agreement through the GSA application system. Requests should be directed to Z.L. (liuzaiyi@gdph.org.cn). The source code used for data analysis and figure generation can be found at <https://github.com/WangLabHKUST/BRCA-pCR> and <https://zenodo.org/records/15673132>. Processed data are accessible through our website at <http://www.wang-lab-hkust.com:8050>. The human hg38 version of the CCRE BED file (926,535 items, last updated 20 May 2020) was downloaded from the UCSC Genome Browser. The TransNEO data were obtained from <https://github.com/cclab-brca/neoadjuvant-therapy-response-predictor> (15). Additional datasets include Hatzis2011 (GSE25055) (10), Wolf2022 (GSE194040) (44), Hoogstraat2022 (GSE191127 and GSE192341) (45), BrightNess (GSE164458) (46), Anurag2022 (47), and Hurvitz2020 (GSE130788) (48). All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Acknowledgments

Funding: This work was supported by the Area Research and Development Program of Guangdong Province, China (no. 2021B0101420006 to Z.L.), the National Science Fund for Distinguished Young Scholars of China (no. 81925023 to Z.L.), the Regional Innovation and

Submitted 4 November 2024

Accepted 2 June 2025

Published 4 July 2025

10.1126/sciadv.adu1521