nature genetics

Article

https://doi.org/10.1038/s41588-023-01507-7

Molecular classification of hormone receptor-positive HER2-negative breast cancer

Received: 19 June 2022

Accepted: 21 August 2023

Published online: 28 September 2023

Check for updates

Xi Jin[1,6], Yi-Fan Zhou[1,6], Ding Ma[1,6], Shen Zhao[1,6], Cai-Jin Lin[1,6], Yi Xiao[1], Tong Fu[1], Cheng-Lin Liu[1], Yi-Yu Chen[1], Wen-Xuan Xiao[1], Ya-Qing Liu[2], Qing-Wang Chen[2], Ying Yu[2], Le-Ming Shi[2,3], Jin-Xiu Shi[4], Wei Huang[4], John F. R. Robertson[5], Yi-Zhou Jiang[1,6,7] ✉ & Zhi-Ming Shao[1,7] ✉

Hormone receptor-positive (HR$^+$)/human epidermal growth factor receptor 2-negative (HER2$^-$) breast cancer is the most prevalent type of breast cancer, in which endocrine therapy resistance and distant relapse remain unmet challenges. Accurate molecular classification is urgently required for guiding precision treatment. We established a large-scale multi-omics cohort of 579 patients with HR$^+$/HER2$^-$ breast cancer and identified the following four molecular subtypes: canonical luminal, immunogenic, proliferative and receptor tyrosine kinase (RTK)-driven. Tumors of these four subtypes showed distinct biological and clinical features, suggesting subtype-specific therapeutic strategies. The RTK-driven subtype was characterized by the activation of the RTK pathways and associated with poor outcomes. The immunogenic subtype had enriched immune cells and could benefit from immune checkpoint therapy. In addition, we developed convolutional neural network models to discriminate these subtypes based on digital pathology for potential clinical translation. The molecular classification provides insights into molecular heterogeneity and highlights the potential for precision treatment of HR$^+$/HER2$^-$ breast cancer.

Breast cancer is the most common female cancer in the world[1]. Recurrence and death due to breast cancer will remain health problems for the foreseeable future. Breast cancers are categorized into different subtypes and treated based on the tumors' hormone receptors (HRs, estrogen receptor (ER) or progesterone receptor (PR)) and human epidermal growth factor receptor 2 (HER2) status. HR-positive/HER2-negative (HR$^+$/HER2$^-$) breast cancer is the most common type, accounting for two-thirds of all breast cancers[2]. Despite high endocrine responsiveness, a persistent risk of distant relapse exists during the 5 years of scheduled endocrine therapy[3,4]. While the recurrence rate is the highest in the first 4–6 years after diagnosis and treatment, the risk never declines to zero, remaining high even after 20–30 years after diagnosis[5]. Challenges to the effective treatment of patients with HR$^+$/HER2$^-$ breast cancer include high intertumoral heterogeneity and resistance to hormonal therapy, which is associated with long-term recurrence risk[5,6]. Improving targeted therapy-based precision medicine

[1]Key Laboratory of Breast Cancer, Department of Breast Surgery, Fudan University Shanghai Cancer Center, Shanghai, China. [2]State Key Laboratory of Genetic Engineering, School of Life Sciences, Human Phenome Institute and Shanghai Cancer Center, Fudan University, Shanghai, China. [3]International Human Phenome Institutes (Shanghai), Shanghai, China. [4]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Shanghai Institute for Biomedical and Pharmaceutical Technologies (SIBPT), Shanghai, China. [5]University of Nottingham, Royal Derby Hospital, Derby, UK. [6]These authors contributed equally: Xi Jin, Yi-Fan Zhou, Ding Ma, Shen Zhao, Cai-Jin Lin, Yi-Zhou Jiang. [7]These authors jointly supervised this work: Yi-Zhou Jiang, Zhi-Ming Shao. ✉e-mail: yizhoujiang@fudan.edu.cn; zhimingshao@fudan.edu.cn

https://doi.org/10.1038/s41588-023-01507-7

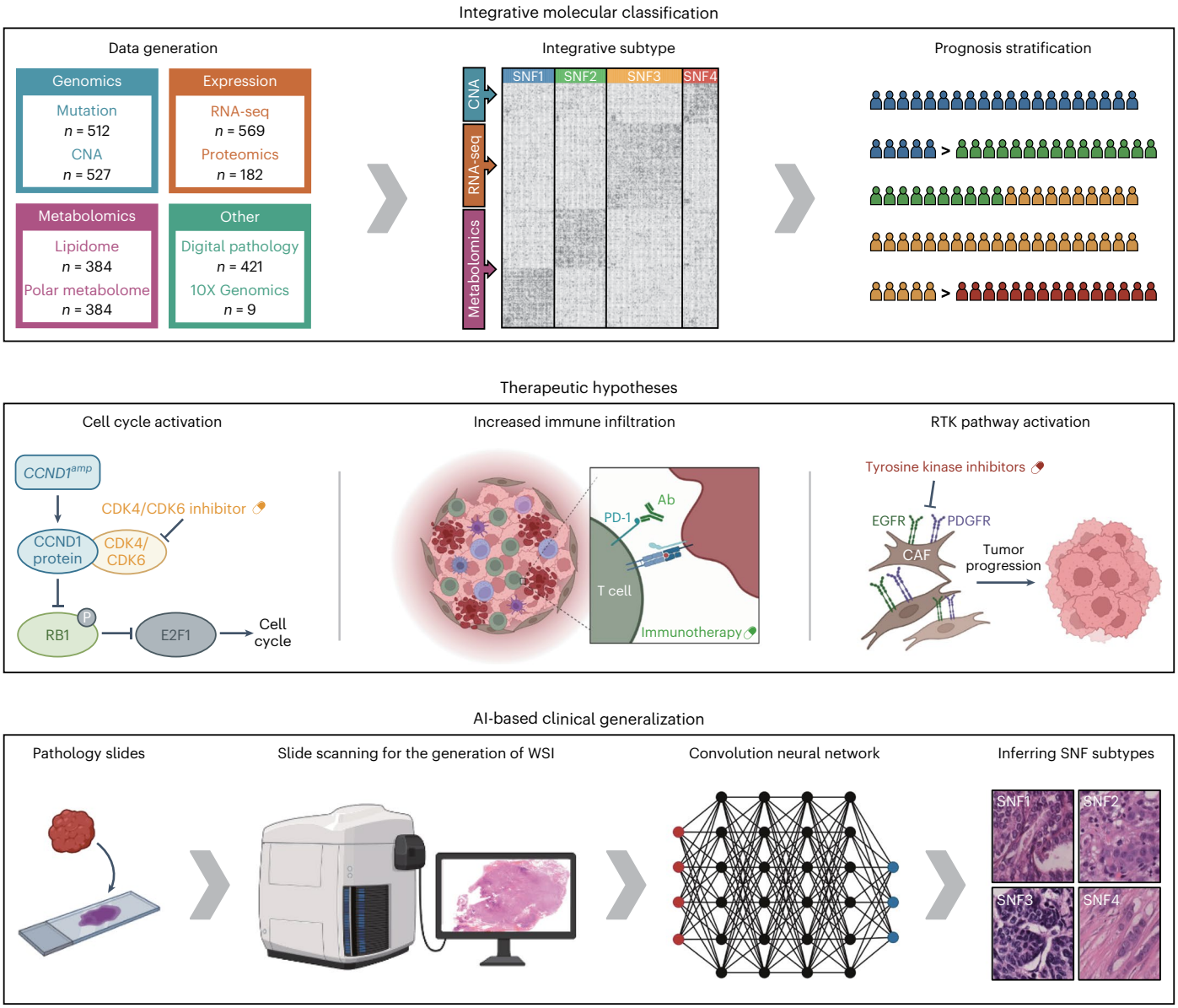


Fig. 1 | Schematic overview of the study. A multi-omics cohort of HR$^+$/HER2$^-$ breast cancer has been constructed for the purpose of multi-omics integrative subtyping. Subsequently, we investigated the molecular features of each subtype and identified therapeutic approaches specific to each subtype. We also devised an AI-based clinically applicable method to infer each subtype. See also Extended Data Fig. 1. AI, artificial intelligence.

strategies might aid in treating HR$^+$/HER2$^-$ breast cancers. Recently, several studies have indicated that adding targeted therapies, such as poly-ADP ribose polymerase (PARP) inhibitors, cyclin-dependent kinase 4 and 6 (CDK4/CDK6) inhibitors and immune checkpoint inhibitors, could improve the survival of patients with HR$^+$/HER2$^-$ breast cancers[7,8]. However, the landscape of druggable targets and the biological relevance of different biomarkers remain uncertain.

Molecular characterization of breast tumors has led to improvements in clinical treatment strategies. For example, the integrated subtyping of breast cancer-based copy number alterations (CNAs) and gene expression in primary tumors from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) has revealed ten integrative clusters (labeled IntClust 1–10)[9]. The Cancer Genome Atlas (TCGA) also identified four major subtypes across all breast cancers using multiplatform clustering[10]. Furthermore, the molecular classification of triple-negative breast cancer has revealed four types and distinct treatment strategies[11]. The benefit of subtyping-based targeted therapy was further validated in a prospective FUTURE clinical trial[12]. Several molecular models have been used in clinical practice for HR$^+$/HER2$^-$ breast cancers[13]. However, the existing clinical staging systems or prognostic models do not accurately reflect biological behavior. To address this issue, we focused on developing a biological characterization that reflected the heterogeneity of HR$^+$/HER2$^-$ breast cancer.

Integrating subtyping through multi-omics has helped accurately profile tumors to identify individual intrinsic biologic characteristics and therapeutic vulnerabilities[14]. Here we performed next-generation DNA and RNA sequencing combined with metabolomics and proteomics to investigate the molecular characterization of HR$^+$/HER2$^-$ breast cancers. In addition, we reported a subtyping system for classifying these tumors into canonical luminal, immunogenic, proliferative and receptor tyrosine kinase (RTK)-driven subtypes. Furthermore, we developed convolutional neural network (CNN) models through deep learning, which can infer subtypes from pathology whole-slide images.

Results

Integrative analyses of HR$^+$/HER2$^-$ breast cancer

We established the large-scale multi-omics cohort of Chinese HR$^+$/HER2$^-$ breast cancer comprising 579 patients. Among them, 569 had RNA sequencing data, 527 had somatic CNA (SCNA) data, 384 had metabolome data (polar metabolomic data and lipidomic data), 512 had whole-exome sequencing (WES) data on primary tumor tissue and paired blood samples, 182 had tandem mass tag (TMT)-based mass spectrometry-quantified protein data and nine had 10X Genomics-based single-cell RNA-sequencing (scRNA-seq) data (Fig. 1, Extended Data Fig. 1a,b and Supplementary Table 1).

To investigate the intrinsic biology of this disease through diverse dimensions, we performed multi-omics clustering for SCNA, mRNA and metabolite abundance using similarity network fusion (SNF) in 351 patients with overlapped datasets (Fig. 1 and Extended Data Fig. 1b). The distribution of PAM50 subtypes was consistent with earlier reports[10], suggesting the accuracy and reliability of our multi-omic data (Extended Data Fig. 1c). Graph showing eigenvalues (*y* axis) of the eigenvectors (*x* axis) from the data graph Laplacian. The greatest eigengap is between the third and fifth eigenvectors, indicating an optimal cluster number of four for SNF cluster analysis, namely, SNF1-4 (Fig. 1 and Extended Data Fig. 1d,e). We found that the SNF classifier combining RNA expression, SCNA and metabolite abundance data can achieve both statistical robustness and a complete delineation of the molecular heterogeneity of HR$^+$/HER2$^-$ breast cancers (Extended Data Fig. 1f). Finally, we further explored the biological role of multi-omics taxonomy, including somatic genomic alteration, SCNA, RNA and metabolomic features, to reveal new therapeutic opportunities (Fig. 1).

The distinct features of each SNF subtype were explored. The association between the SNF subtypes and the PAM50 subtypes is shown in Extended Data Fig. 1g. SNF1 is enriched with PAM50 luminal A subtype, whereas SNF3 mainly comprises the PAM50 luminal B subtype. SNF2 comprised almost all PAM50 HER2-enriched subtype and SNF4 contained a high proportion of the PAM50 normal-like subtype.

SNF1 and SNF3 tumors were almost entirely composed of PAM50 luminal A/PAM50 luminal B subtypes. Over half of the tumors in SNF1 harbored *PIK3CA* mutations (51%), while fewer tumors had *TP53* mutations (11%; Fig. 2a). SNF3 tumors exhibited higher copy number gain frequencies of *CCND1*, *FGFR1* and *MDM2* and more cell-cycle-pathway activations (Fig. 2a,b). Homologous recombination deficiency (HRD) scores were also observed in SNF3 tumors, indicating high genomic instability (Fig. 2a). In addition, SNF1 and SNF3 tumors were enriched for ER pathways (Fig. 2b).

SNF2 and SNF4 tumors contained fewer PAM50 luminal A/PAM50 luminal B subtype tumors and less activation of hormone receptor pathways (Fig. 2b). SNF2 showed enriched adaptive immune response pathways (Fig. 2b). For SNF4 tumors, expression signatures associated with receptor protein kinase and extracellular matrix structural constituents were highly enriched (Fig. 2b).

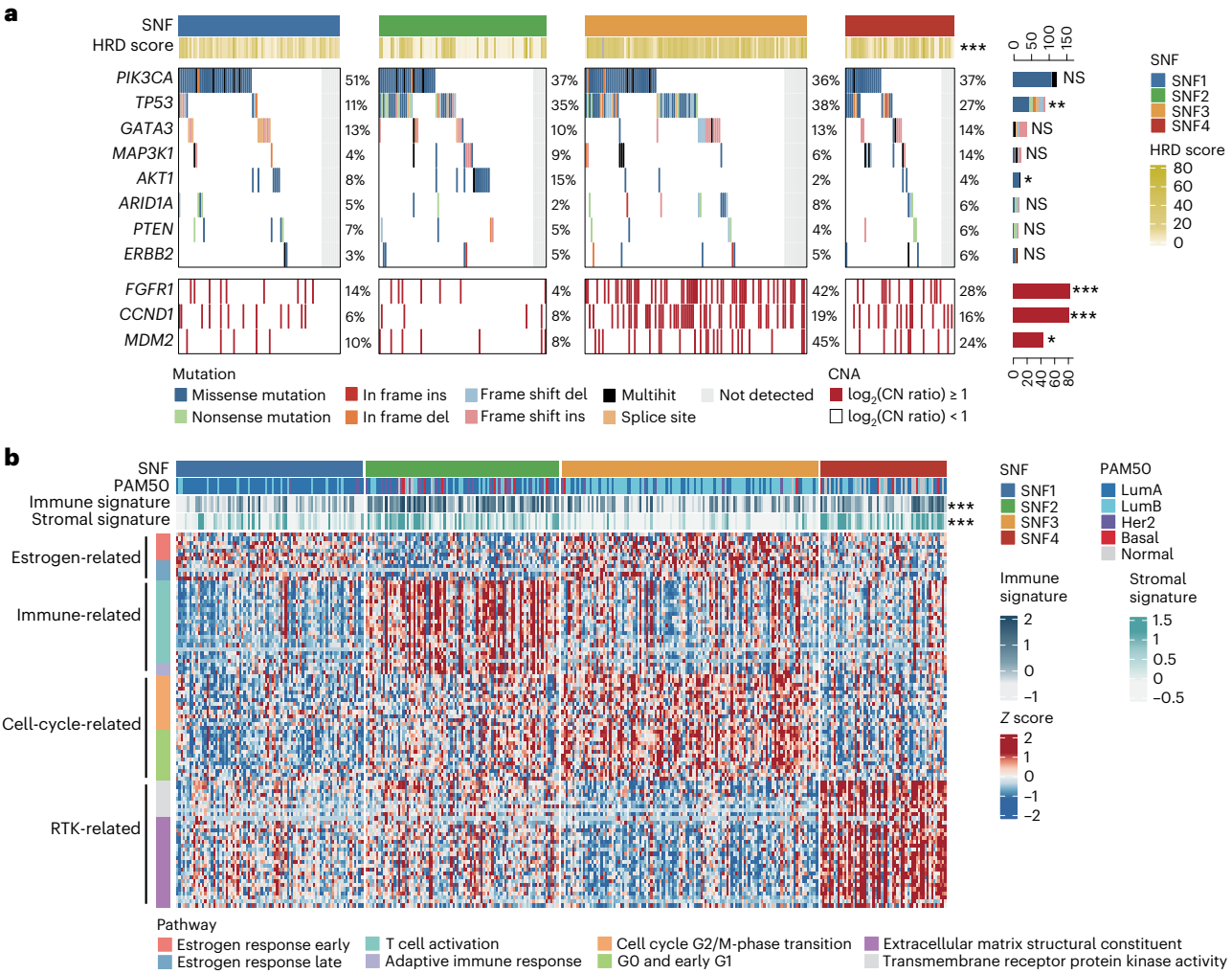
The metabolomic landscape of HR$^+$/HER2$^-$ breast cancer

We next explored the difference in metabolic reprogramming characteristics among the four SNF subtypes. First, the expression distance observed was larger than that within normal samples (Extended Data Fig. 2a). We applied the Euclidean distance to analyze the global divergence in metabolic gene expression within SNF subtypes and their corresponding normal tissues. We found that the SNF3 subtype had the highest level of metabolic dysregulation than other subtypes, while SNF4 was the least metabolically dysregulated subtype (Fig. 2c). This prompted us to further identify the metabolomic features within tumors of different subtypes. We conducted network analyses to explore subtype-specific metabolic genes, polar metabolites and lipids and illustrate the metabolic features among subtypes (Fig. 2d–f and Extended Data Fig. 2b–d). We found that most dysregulated pathways were upregulated in the SNF3 subtype but downregulated in SNF4. In addition, the SNF1 and SNF2 subtypes were characterized by phospholipid metabolism, and high expression of genes involved in phospholipid syntheses, such as *PIP5K1A*[15] and *PIK3CG*[16], was identified.

Furthermore, we attempted to link polar metabolites and lipids to genomic events in an integrated analysis. We analyzed the correlation to explore the connections between genomic alterations and metabolic reprogramming. We investigated the associations among metabolic genes, genomic mutations and copy number variations with the abundance of polar metabolites and lipid subclasses across SNF clusters. Focusing on a curated compilation of 19 known cancer-related genes frequently mutated in HR$^+$/HER2$^-$ breast cancer[17,18], we used a linear regression model (controlling the confounding factors) to assess the associations between somatic mutations and the abundance of metabolites (Extended Data Fig. 3a and Supplementary Table 2). We found that *TP53* mutations were positively associated with the abundance of nucleotides such as deoxyinosine and phosphatidylglycerol (OxPG) (Extended Data Fig. 3b). With respect to CNAs, we found that the copy number values of chromosomal regions (8q23.3 or 1q32.1) containing specific breast cancer oncogenes (such as *TRPS1* and *ADORA1*) were positively or negatively correlated with variant metabolites (nucleotides or amino acids; Extended Data Fig. 3c,d and Supplementary Table 3). Moreover, we also observed the connections between the expression of cell-cycle-related genes and DNA synthesis-related metabolites (Extended Data Fig. 3e and Supplementary Table 4). For instance, we observed a positive correlation of aurora kinase A (*AURKA*) mRNA expression with the abundance of deoxyinosine and a negative correlation of *CCND3* mRNA expression with deoxyuridine monophosphate (dUMP) abundance (Extended Data Fig. 3f). Generally, these analyses might provide insights into the mechanisms driving metabolic reprogramming in HR$^+$/HER2$^-$ breast cancer.

Fig. 2 | Integrated landscape of HR$^+$/HER2$^-$ breast cancer. a, The recurrent somatic mutations identified in tumor samples among the four types. CNA of cancer-related genes. The shown cancer-related genes were all located in significant GISTIC peaks. *P* values were from the two-sided Fisher's exact test. **b**, Heatmaps show differential expression between subtypes. Pathway enrichment analysis highlights differences between subtypes. Estrogen-related pathways, immune response pathways, cell cycle pathways and RTK-related pathways were enriched in the SNF1 (canonical luminal), SNF2 (immunogenic), SNF3 (proliferation) and SNF4 (RTK-driven) subtypes, respectively. *P* values were from the two-sided ANOVA test. **c**, Global differences in metabolic gene expression between SNF subtypes in the luminal cohort. The distribution distances (r.m.s.d.) were calculated between SNF1 (blue), SNF2 (green), SNF3 (yellow), SNF4 (red) tumors and their corresponding normal tissues. The inset shows the average distances between pairs of SNF tissues as a percentage of the average distance between SNF tumors and normal tissues. $P_{(SNF1–SNF2)} = 0.004$, $P_{(SNF1–SNF3)} < 2.2 \times 10^{-16}$, $P_{(SNF1–SNF4)} < 2.2 \times 10^{-16}$, $P_{(SNF2–SNF3)} < 2.2 \times 10^{-16}$, $P_{(SNF2–SNF4)} < 2.2 \times 10^{-16}$ and $P_{(SNF3–SNF3)} < 2.2 \times 10^{-16}$. *P* values were from the two-sided Wilcox rank-sum test and two-sided Kruskal–Wallis test. SNF1, $n = 86$ biologically independent samples; SNF2, $n = 89$ biologically independent samples; SNF3, $n = 118$ biologically independent samples; SNF4, $n = 58$ biologically independent samples, normal samples $n = 11$. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5× IQR, and the data points outside the whisker were outliers. **d**–**f**, Metabolic transcriptomics correlation network based on 836 metabolic genes (**d**), polar metabolomics correlation network based on 669 polar metabolites (**e**) and lipidomics correlation network based on 46 lipid subclasses (**f**) using Spearman correlation >0.4 and FDR < 0.05 cutoff were illustrated. Correlation networks were partitioned and color-coded by a graph-clustering algorithm, and the average quantification of SNF subtypes in the correlation networks was presented. ***FDR < 0.001, **0.001 ≤ FDR < 0.01, *0.01 ≤ FDR < 0.05, NS, FDR ≥ 0.05. See also Extended Data Figs. 2 and 3. ANOVA, analysis of variance; IQR, interquartile range; $\log_2$ (CN ratio), $\log_2$ copy number ratio.

https://doi.org/10.1038/s41588-023-01507-7



https://doi.org/10.1038/s41588-023-01507-7

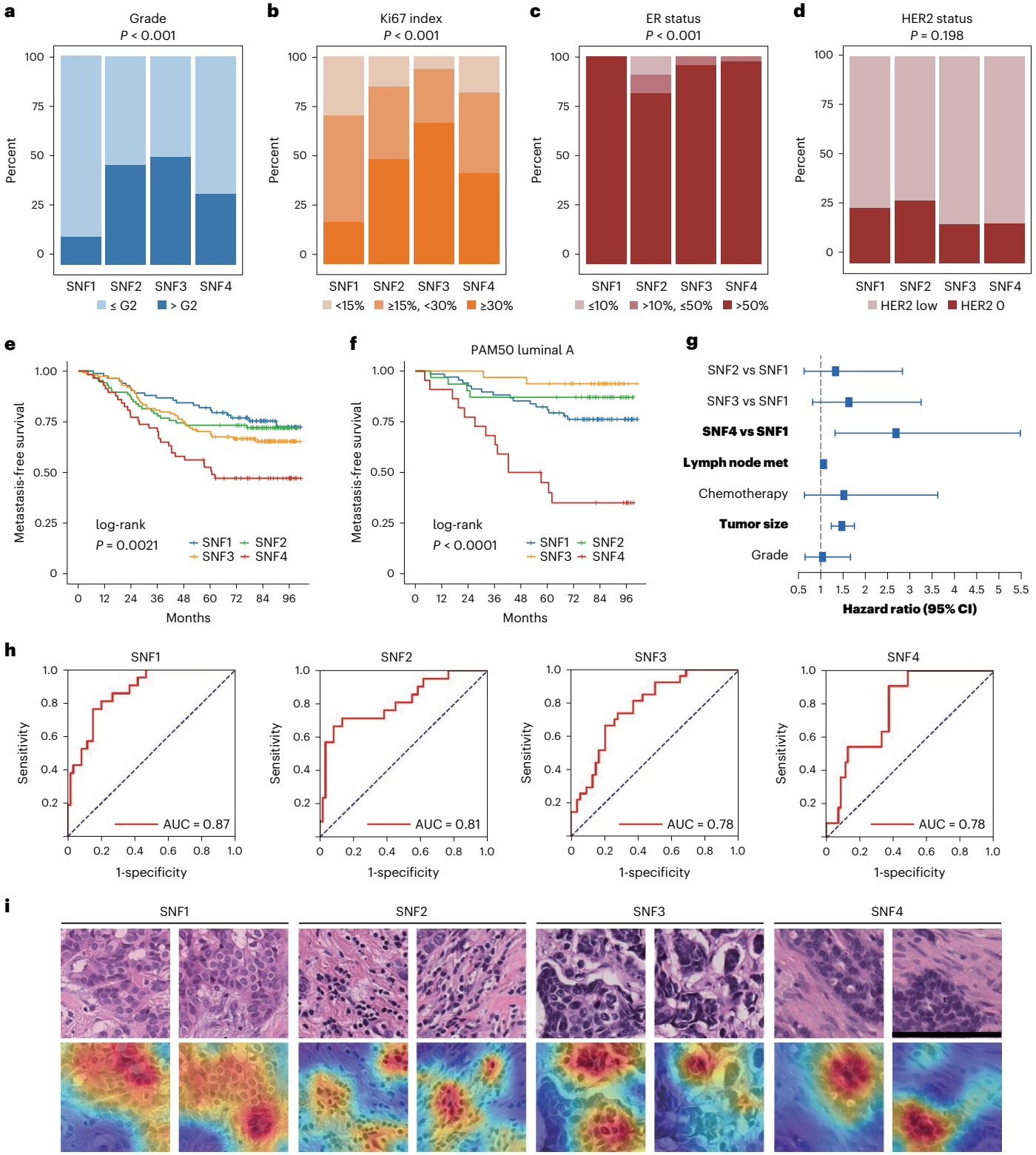
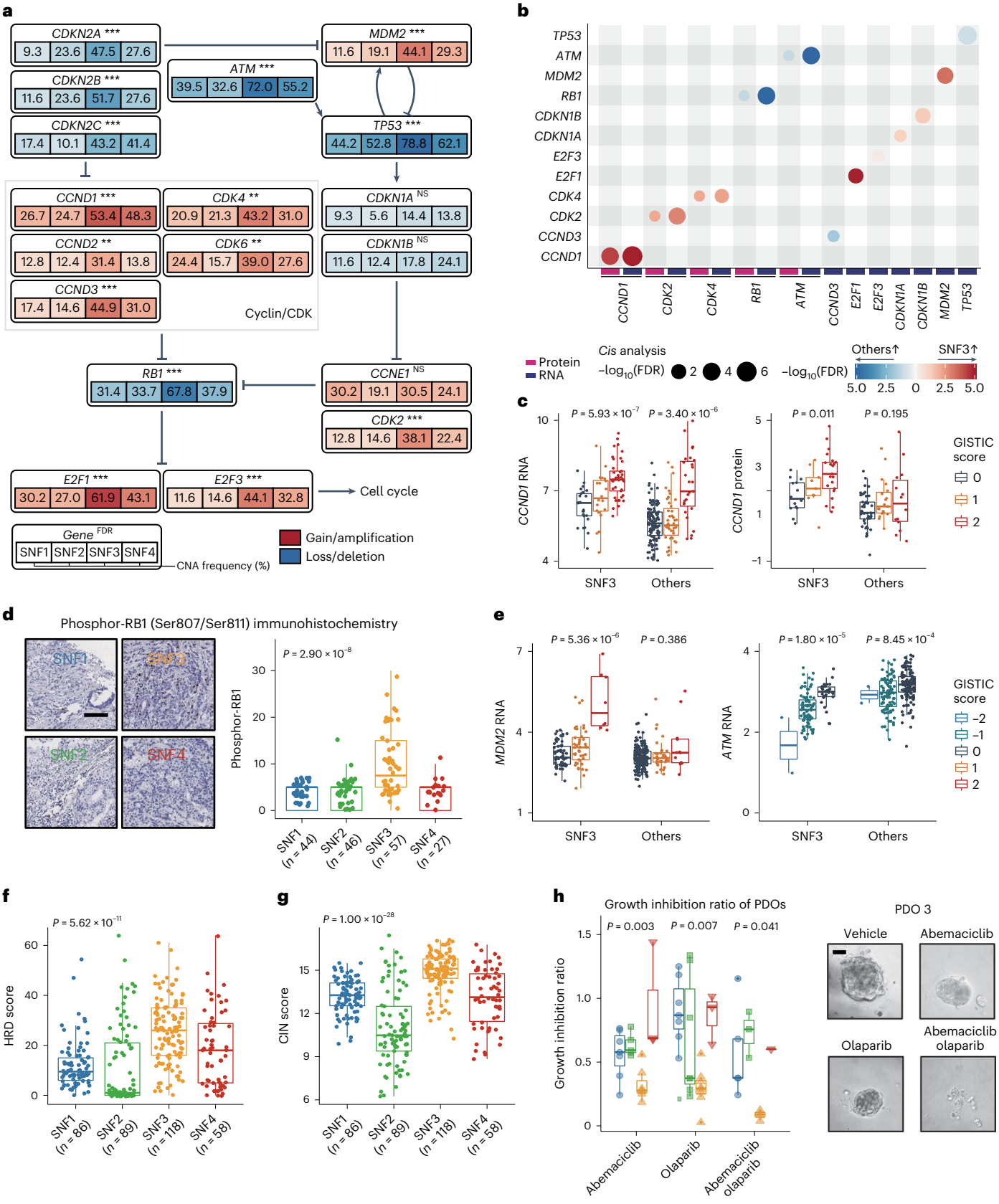


Fig. 3 | Distinct clinical characteristics and pathological patterns of the four SNF subtypes. a–d, Comparison of clinicopathologic characteristics among SNF subtypes. $P_{grade} = 5 \times 10^{-4}$ (**a**), $P_{Ki67\,index} = 5 \times 10^{-4}$ (**b**), $P_{ER\,status} = 1 \times 10^{-4}$ (**c**) and $P_{HER2\,status} = 0.198$ (**d**). *P* values were from the two-sided Fisher's exact test. **e**,**f**, Association of the SNF subtypes with MFS in the overall $HR^{+}/HER2^{-}$ population (**e**) and PAM50 luminal A subtype (**f**). *P* (MFS in PAM50 luminal subtype) = 2.9×10^{-7}. **g**, Forest plot of multivariate Cox regression analysis for MFS adjusting for tumor size, lymph node status, SNF subtypes, chemotherapy and histological grade. The included patients all received endocrine therapy ($n = 296$). The hazard ratios were shown with 95% confidence intervals (CI). Error bar center indicates hazard ratios. SNF2 vs SNF1: hazard ratio = 1.33 (0.63–2.83), $P = 0.458$. SNF3 vs SNF1: hazard ratio = 1.63 (0.82–3.25), $P = 0.166$. SNF4 vs SNF1: hazard ratio = 2.69 (1.32–5.48), $P = 0.006$. Lymph node met: hazard ratio = 1.06 (1.04–1.09), $P = 2.12 \times 10^{-7}$. Chemotherapy: hazard ratio = 1.52 (0.64–3.63), $P = 0.348$. Tumor size: hazard ratio = 1.47 (1.24–1.76), $P = 1.36 \times 10^{-5}$. Grade: hazard ratio = 1.04 (0.65–1.67), $P = 0.871$. Bold font indicates statistical significance. met, metastasis. **h**, ROC curves for using our CNN models to identify the SNF subtypes. **i**, Representative tiles from the cases of each molecular subtype with the highest prediction score. Class activation maps highlight the subtype-specific discriminative subregions. Tile scale bar, 128 µm. All tiles presented were among the top 100 tiles according to the prediction score. See also Extended Data Figs. 4 and 5 and Supplementary Fig. 1.

https://doi.org/10.1038/s41588-023-01507-7



Distinct clinical features among the four SNF subtypes

We further analyzed the clinicopathologic characteristics (Fig. 3a–d, Extended Data Fig. 4a and Supplementary Table 5). SNF3 tumors had a higher Ki67 index and tumor grade, whereas SNF2 tumors had lower ER expression, as measured through immunohistochemical (IHC) staining (Fig. 3a–c). There was no significant difference in HER2 status among SNF subtypes (Fig. 3d). Interestingly, SNF4 had worse distant metastasis-free survival (MFS; $P = 0.0021$) and relapse-free survival ($P = 0.00093$) than the other three subtypes (Fig. 3e and Extended Data Fig. 4b). Especially for patients with PAM50 luminal A,

Fig. 4 | Proteogenomic analysis reveals cell cycle signaling as the target of the SNF3 subtype. a, The CNA frequency of the cell cycle genes across SNF clusters. Two-sided *P* values were obtained from Fisher's exact test and were adjusted using the Benjamini–Hochberg procedure. Significant codes: NS ≥ 0.05; **FDR < 0.01; ***FDR < 0.001. **b**, *Cis* effects of cell cycle genes on RNA and protein levels in SNF3 tumors. The size denotes the *P* values of the *cis* analysis, and the color denotes the *P* values of the comparison of the RNA or protein levels between SNF3 and other clusters. Two-sided *P* values were obtained from the Mann–Whitney–Wilcoxon test and adjusted by the Benjamini–Hochberg procedure. **c**, The mRNA (left) and protein (right) levels of *CCND1* across different GISTIC scores between SNF3 ($n = 89$ in the RNA analysis and $n = 44$ in the protein analysis) and other clusters ($n = 188$ in the RNA analysis and $n = 80$ in the protein analysis). Two-sided *P* values were from the Kruskal–Wallis test. **d**, Immunohistochemical detection of phosphor-RB1 and immunohistochemical staining score quantification among four SNF subtypes. *P* values were from the two-sided Kruskal–Wallis test. Scale bar: 100 μm. **e**, The mRNA levels of *MDM2* (left) and *ATM* (right) across different GISTIC scores between SNF3 ($n = 89$ in the *MDM2* analysis and $n = 111$ in *ATM* analysis) and other clusters ($n = 188$ in the *MDM2* analysis and $n = 220$ in *ATM* analysis). Two-sided *P* values were from the Kruskal–Wallis test. **f**,**g**, Comparison of the HRD (**f**) and CIN (**g**) scores across SNF subtypes. Two-sided *P* values were from the Kruskal–Wallis test. **h**, Results of the cell viability assay testing the efficacy of 0.1 μM abemaciclib (SNF1, $n = 7$; SNF2, $n = 5$; SNF3, $n = 7$ and SNF4, $n = 3$), 1 μM olaparib (SNF1, $n = 7$; SNF2, $n = 7$; SNF3, $n = 8$ and SNF4, $n = 3$) and abemaciclib combined with olaparib on patient-derived organoids (SNF1, $n = 5$; SNF2, $n = 3$; SNF3, $n = 4$ and SNF4, $n = 1$) from different subtypes. Scale bar: 100 μm. *P* values were from the two-sided ANOVA test. In all boxplots, the center lines represent median values; the bounds of the boxplot represent the interquartile ranges; the whiskers show the range of the data. See also Extended Data Fig. 6.

the prognosis of SNF4 was far worse than that of the other subtypes (Fig. 3f and Extended Data Fig. 4c). The application of univariate and multivariate Cox analyses unequivocally confirmed the SNF4 subtype as an independent prognostic indicator (Fig. 3g and Extended Data Fig. 4d). Taken together, we reclassified the $HR^+/HER2^-$ tumors into four subtypes, each characterized by distinct clinicopathological and multi-omics features.

Clinically applicable method for SNF subtype classification

Because the clinical implementation of multi-omics profiling presents challenges arising from its substantial cost, long turnaround time and intricate technological process, cost-effective, fast and convenient approaches are required to extrapolate our subtyping system. Here we used two classifiers, one based on digital pathology data and the other based on transcriptomics data.

For the digital pathology-based method, we asked whether the tumors of the four SNF subtypes differed in pathological patterns and whether they could be distinguished using neural network models based on digital pathology. We used whole-slide images (WSIs) of 243 patients with multi-omics SNF subtyping results. We adopted a deep learning-based pipeline to develop CNN models to identify each of the four subtypes (Supplementary Fig. 1). The area under the curve (AUC) for cross-validation was 0.87 for the SNF1 subtype, 0.81 for the SNF2 subtype, 0.78 for the SNF3 subtype and 0.78 for the SNF4 subtype (Fig. 3h). Based on our developed models, we explored which morphological patterns provided clues for identifying the SNF subtypes by visualizing the representative image tiles. The morphological features of the tiles indicating each SNF subtype could be summarized as follows (Fig. 3i): SNF1, partially conserved the morphology of normal breast gland; SNF2, high immune cell infiltration; SNF3, high abundance of atypical tumor cells; SNF4, enrichment of tumor cell clusters with surrounding fibroblasts. Overall, these results revealed the difference in pathological patterns across the SNF subtypes and indicated the feasibility of differentiating the SNF subtypes based on digital pathology.

For the transcriptomics data-based method, we developed a random forest classifier based on the expression of the highly expressed genes of each SNF subtype (Extended Data Fig. 5a; Methods). We used the same cross-validation cohort in the development of the digital pathology model. The AUC was 0.95 for the SNF1 subtype, 0.93 for the SNF2 subtype, 0.85 for the SNF3 subtype and 0.82 for the SNF4 subtype (Extended Data Fig. 5b and Supplementary Table 6), higher than the CNN models.

We predicted the SNF subtypes using the transcriptomics data-based classifier in TCGA, METABRIC and Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohorts. Of note, all four SNF subtypes existed (Extended Data Fig. 5c–e). Molecular features of the SNF subtypes were preserved in the corresponding subtypes inferred by the classifier.

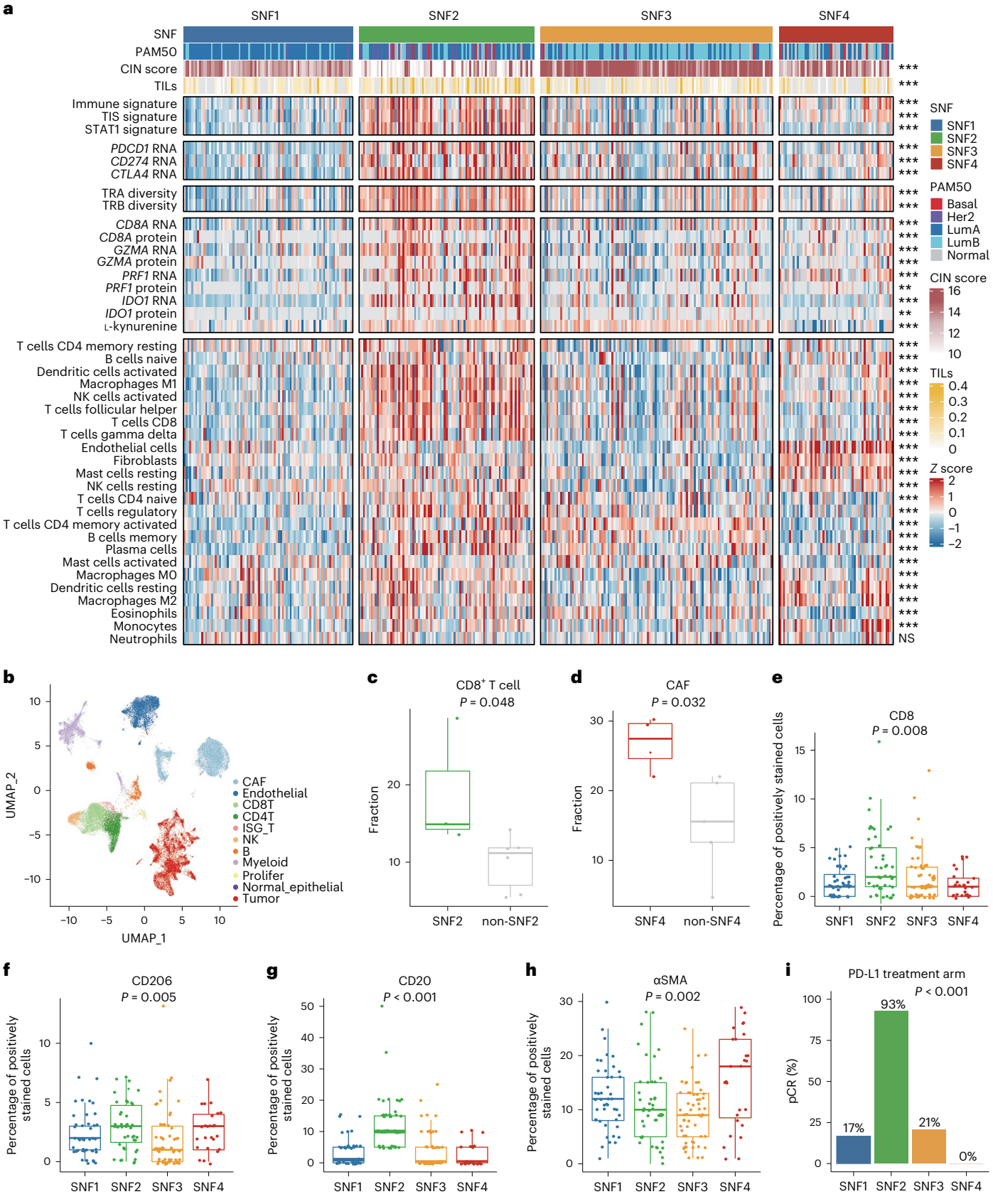
Upregulation of the cell cycle pathway in the SNF3 subtype

Applying CDK4/CDK6 inhibitors could improve the survival of patients with $HR^+/HER2^-$ breast cancer[19–22]. However, a certain proportion of patients with $HR^+/HER2^-$ breast cancer do not respond to CDK4/CDK6 inhibitors[23]. Through genomic analysis, we found a global increase in the CNA frequency of certain cell cycle genes in SNF3 samples (Fig. 4a). We then analyzed the association between these CNAs and RNA or protein expression. We observed a substantial increase in RNA and protein expression for oncogenes such as *CCND1* and a decrease for tumor suppressors such as *RB1* (Fig. 4b), suggesting *cis* activation of cell cycle signaling driven by genomic alterations in SNF3. Specifically, a *cis* effect of *CCND1* between CNA and RNA expression was observed in both SNF3 and other tumors, whereas a *cis* effect between CNA and protein level was observed for SNF3 tumors only (Fig. 4c). Similarly, gene set enrichment analysis (GSEA) demonstrated that cell cycle pathways were activated in SNF3 (Extended Data Fig. 6a), and the mRNA and protein abundance of other cell cycle genes were also increased in SNF3 tumors (Extended Data Fig. 6b–d). To further support this finding, we performed IHC staining for phosphor-RB1 and found that phosphor-RB1 was higher in SNF3 (Fig. 4d).

Fig. 5 | Microenvironment landscape of $HR^+/HER2^-$ breast cancers. **a**, Heatmap showing the tumor-infiltrating lymphocyte score; the expression of the mRNA biomarkers of immune checkpoint blockade; the T cell receptor (*TRA* and *TRB*) diversity; the mRNA and protein abundance of *CD8A*, *GZMA*, *PRF1* and *IDO1*; the abundance of L-kynurenine and the estimated abundance of 24 microenvironment cell types among four SNF subtypes. *P* values were from the two-sided Kruskal–Wallis test or Fisher's exact test. NS ≥ 0.05; **FDR < 0.01; ***FDR < 0.001. **b**, Uniform manifold approximation and projection (UMAP) visualization of 59,479 cells from nine $HR^+/HER2^-$ samples, colored by cell cluster. **c**, Boxplot comparing the proportion of $CD8^+$ T cells between SNF2 ($n = 3$) and non-SNF2 ($n = 6$) samples. *P* values were from the two-sided Wilcoxon test. **d**, Boxplot comparing the proportion of CAFs between SNF4 ($n = 4$) and non-SNF4 ($n = 5$) samples. *P* values were from the two-sided Wilcoxon test. **e**–**h**, Immunohistochemical staining scores of CD8 (**e**), CD206 (**f**), CD20 (**g**) and αSMA (**h**) among the SNF1 ($n = 45$), SNF2 ($n = 47$), SNF3 ($n = 58$) and SNF4 ($n = 27$) subtypes. $P(\text{CD20}) = 1 \times 10^{-4}$. *P* values were from the two-sided Kruskal–Wallis test. **i**, The pathological complete response (pCR) rate of patients with different SNF subtypes in the PD-L1 treatment arm of the I-SPY2 clinical trial. $P = 4.8 \times 10^{-6}$. *P* values were from the two-sided Fisher's exact test. In all violin or boxplots, the median ± interquartile range were indicated. Boxplots in **c**–**h** depict the first and third quartiles as the lower and upper bounds, respectively. The whiskers represent 1.5× IQR, and the center depicts the median. See also Extended Data Figs. 7 and 8 and Supplementary Fig. 2.

https://doi.org/10.1038/s41588-023-01507-7

Cis regulation of cell cycle signaling in SNF3 was also observed for two regulators of G2/M, namely, the G2/M activator *MDM2* and the G2/M inhibitor *ATM* (Fig. 4e and Extended Data Fig. 6e). Moreover, the HRD and chromosomal instability (CIN) scores (Fig. 4f,g) were higher in SNF3 tumors. As the high HRD score correlates with sensitivity to PARP inhibitors[24], we observed the growth of patient-derived organoids (PDOs) from SNF3 tumors was inhibited by a single agent (CDK4/CDK6 inhibitor abemaciclib or PARP inhibitor olaparib) and was further



inhibited by a combination of abemaciclib and olaparib (Fig. 4h and Supplementary Table 7). These results indicated that the SNF3 subtype was enriched for the activation of G1/S and G2/M cell cycle progression.

Microenvironment landscape of HR$^+$/HER2$^-$ breast cancers

The tumor microenvironment (TME) is involved in tumor development[25]. Therefore, we performed a comprehensive analysis of the microenvironmental characteristics among the four SNF subtypes. Interestingly, the abundance of immune cell populations, including particularly cells of the adaptive immune system, assessed through single-sample GSEA using CIBERSORT and Microenvironment Cell Populations-counter (MCP-counter)[26,27], was markedly increased in SNF2 tumors (Fig. 5a and Extended Data Fig. 7a,b). The expression of several cytotoxic factors (*CD8A*, *GZMA* and *PRF1*) and T cell receptor (*TRA* and *TRB*) diversity were higher in SNF2 tumors (Fig. 5a). Through the hematoxylin and eosin (H&E) sections, we observed that tumor-infiltrating lymphocytes (TILs) were more abundant in SNF2 tumors (Fig. 5a). Meanwhile, SNF2 tumors had higher expression levels of several immune checkpoint biomarkers, including the tumor inflammation signature (TIS), *STAT1* signature and *PDCD1* (ref. 7; Fig. 5a and Extended Data Fig. 7c), indicating the potential benefit from ICB with SNF2. However, L-kynurenine and *IDO1* (the rate-limiting enzymes in the kynurenine pathway), which can promote the immune escape of tumors[28], were also higher in the SNF2 subtype (Fig. 5a and Extended Data Fig. 7d), suggesting that they might be immune escape mechanism for those tumors. By contrast, stromal cells, including cancer-associated fibroblasts (CAFs) and endothelial cells, were markedly enriched in SNF4 tumors (Fig. 5a).

We further analyzed the TME characteristics of the SNF2 and SNF4 tumors by scRNA-seq. In total, 59,479 cells from nine tumors were analyzed, and 11 distinct clusters were identified, including B cells, CAFs, CD4$^+$ T cells, CD8$^+$ T cells, interferon-stimulated gene (*ISG*) T cells, natural killer (NK) cells, tumor cells, endothelial cells, myeloid cells, nontumor epithelial cells and proliferating cells (Fig. 5b). Cells were assigned to cell type based on consensus between expression-based clustering, inference of SCNA and annotation by canonical markers (Extended Data Fig. 8a–c). Indeed, we observed a higher proportion of CD8$^+$ T cells in SNF2 samples (Fig. 5c) and CAFs in SNF4 samples (Fig. 5d), which validated the deconvolution results from our bulk RNA-seq data. In addition, we investigated the functional differences in immune cells in patients with different subtypes by single-cell sequencing. CD8$^+$ T cells from patients with SNF2 were enriched in interferon response pathways versus patients with non-SNF2, indicating the activation of the immune response (Extended Data Fig. 8d). Higher levels of the cytotoxic score, *GZMK* and *GNLY* in CD8$^+$ T cells from patients with SNF2 suggested more potent cytotoxicity of CD8$^+$ T cells for these patients (Extended Data Fig. 8e,f). In conclusion, these data indicated that SNF2 tumors had a higher abundance of immune cells and a higher expression of functional signatures of immune activation, especially for CD8$^+$ T cells.

We further performed immunohistochemistry (IHC) to validate the TME features of the four SNF subtypes. The markers for immune infiltration, including CD8, CD206 and CD20, were higher in SNF2 (Fig. 5e–g and Supplementary Fig. 2). The score of α-smooth muscle actin (αSMA), a marker for stromal cells, was markedly higher in SNF4 (Fig. 5h and Supplementary Fig. 2). For external validation, SNF subtyping was extended to samples from I-SPY2, a neoadjuvant platform trial[7]. We found that, in the arm treated with anti-programmed death-ligand 1(PD-L1) inhibitors, the majority of SNF2 samples (93%) achieved pathological complete response (Fig. 5i). Overall, we comprehensively dissected the TME heterogeneity among the SNF subtypes and showed that SNF2 tumors were characterized by immune activation status, potentially supporting the use of immune checkpoint blockade (ICB) for these tumors.

SNF4 subtype-derived CAFs enhance tumor growth

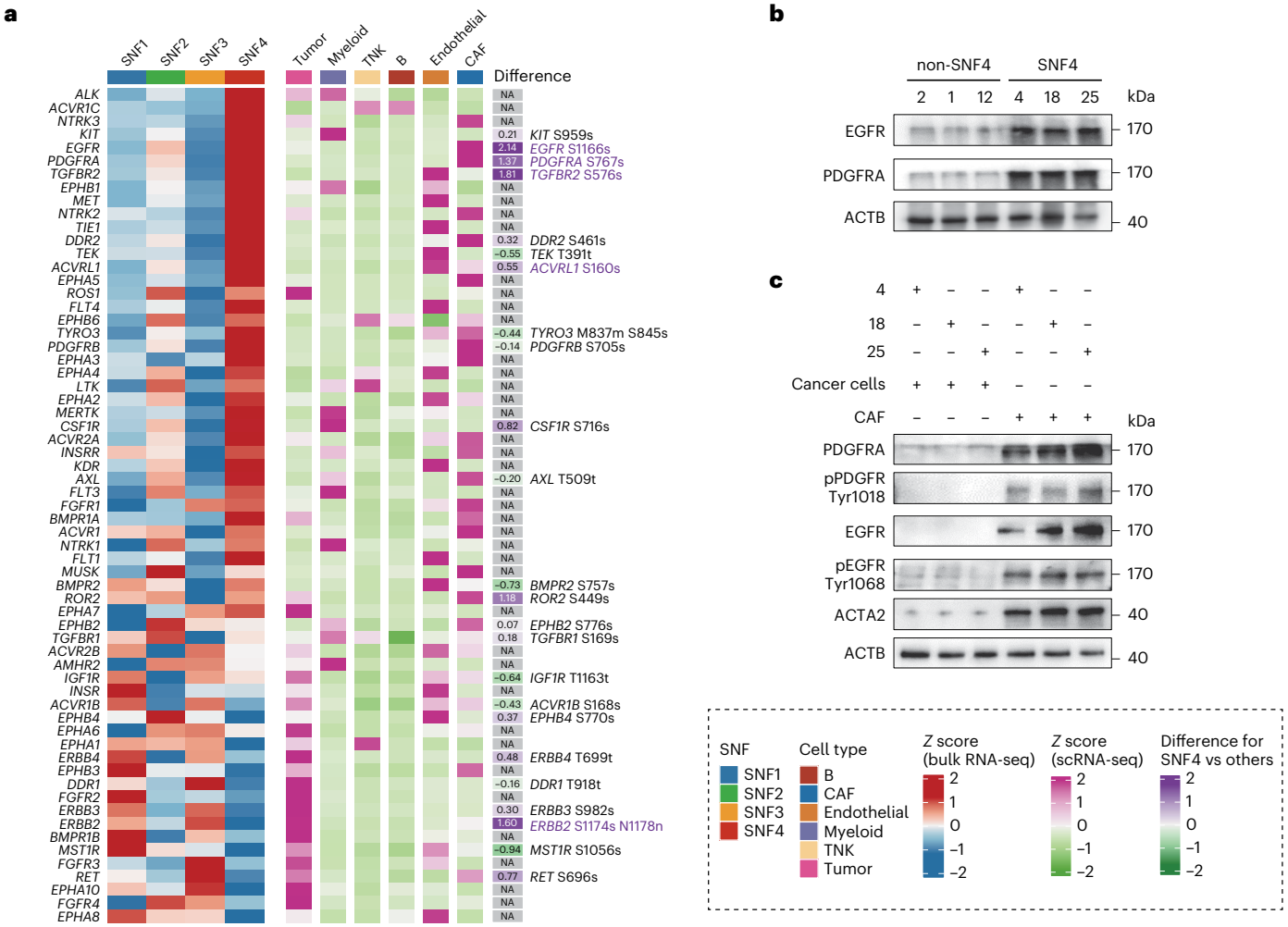
As patients with the SNF4 subtype had the worst outcome among those with HR$^+$/HER2$^-$ breast cancers, we next intended to explore the underlying mechanism underpinning this observation. Using RNA-seq data, we performed GSEA and discovered that the RTK pathway, including several important RTKs, was enriched in patients with SNF4 (Fig. 6a and Extended Data Fig. 9a–d). The RTK pathway is one of the downstream pathways facilitating tumor progression in HR$^+$/HER2$^-$ breast cancers[29]. RTKs can be highly expressed in tumor or stromal cells[30,31]. Given that SNF4 tumors had a higher abundance of CAFs, we reasoned that RTKs might exert their function through enhanced activity in CAFs. Through the integrative analysis of RNA-seq, scRNA-seq and phosphorylated proteomics (CPTAC breast cancer cohort[32]), we observed that the two SNF4 highly expressed RTKs (namely, *EGFR* and *PDGFRA*) were mainly expressed in CAFs and highly phosphorylated (Fig. 6a and Extended Data Fig. 10a,b). Through western blot, we validated that *PDGFRA* and *EGFR* were highly expressed and highly phosphorylated in SNF4 tumor samples, especially in SNF4-derived CAFs (Fig. 6b,c).

CAFs can enhance tumor growth[33–35]. To determine the effects of CAFs on tumor growth of HR$^+$/HER2$^-$ breast cancer, we isolated cancer cells from SNF1 (the better prognosis subtype) or SNF4 (the worse prognosis subtype) tumors and cultured them with conditioned CAF media from SNF4 tumors (Fig. 6d). In addition, we observed that conditioned media (CM) from SNF4-derived CAFs could promote the proliferation of both SNF1- and SNF4-derived cancer cells (Fig. 6e), indicating that CAF abundance might be one of the mechanisms for the aggressiveness of the SNF4 subtype.

It has been reported that phosphorylated *EGFR* and *PDGFRA* can activate the downstream mitogen-activated protein kinase (MAPK) signaling pathway[36]. Correspondingly, in the CPTAC cohort, we observed higher phosphorylation levels of key molecules of the MAPK pathway, such as rapidly accelerated fibrosarcoma (RAF) (*RAF1*) and ERK1/ERK2 (*MAPK1*/*MAPK13*), in SNF4 samples (Extended Data Fig. 10a–c). We used IHC staining for phospho-ERK1/phospho-ERK2 (*MAPK1*/*MAPK3*) to validate the result (Extended Data Fig. 10d). It was reported that the multi-RTK inhibitor sorafenib could effectively target the RAF–MAPK pathway and thus inhibit the activation of *EGFR* and *PDGFRA*[36]. Through the coculture model of cancer cells and CAFs, we further verified that treatment with sorafenib abolished the tumor

Fig. 6 | RTK-driven (SNF4) subtype-derived CAFs could enhance tumor growth and were vulnerable to sorafenib. a, Heatmaps showing bulk RNA-seq expression (left) and scRNA-seq expression (middle) of the list of RTKs among the four SNF subtypes. The right heatmap shows the difference in phosphosite abundance between SNF4 and others. The kinase genes listed in 'GOMF_TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY' (obtained from http://www.gsea-msigdb.org/gsea/msigdb) and a well-organized list of protein kinase genes published previously[54] were included. The phosphosites enriched in SNF4 and detected at $P < 0.05$ were marked in purple. *P* values were from the two-sided *t* test. NA, no phosphosites detected. **b**,**c**, Western blot images showing the expression of *EGFR*, *PDGFRA*, p-*PDGFRA* and p-*EGFR* in SNF4 tumors or non-SNF4 tumors (**b**) and SNF4 tumor cells or SNF4 CAFs (**c**). All western blot experiments were repeated three times with similar results. **d**,**e**, Viability assay testing the tumor growth of SNF1 ($n = 2$) and SNF4 ($n = 2$) patient-derived cancer cells cultured in CAF CM from SNF4 tumors. Scale bar: 100 μm. $P(\text{PDO 5}) = 5.6 \times 10^{-5}$, $P(\text{PDO 4}) = 4.3 \times 10^{-5}$, $P(\text{PDO 18}) = 6.2 \times 10^{-5}$. *P* values were from the two-sided T test. The boxplots show the first and third quartiles as the lower and upper bounds, with the whiskers representing 1.5× IQR and the center indicating the median. **f**,**g**, The efficacy of sorafenib (1 μM) on patient-derived cancer cells cocultured with fibroblast cells. Scale bar: 100 μm. The cancer cells were treated with sorafenib three times. Shown is one of the representative experiments. The experiment was repeated three independent times with similar results. **h**, Schematic diagram showing how CAFs enhance tumor growth of the RTK-driven (SNF4) subtype. See also Extended Data Figs. 9 and 10.

https://doi.org/10.1038/s41588-023-01507-7



growth phenotype induced by the coculture of SNF4 subtype-derived CAFs (Fig. 6f,g). In conclusion, high expression of *EGFR*/*PDGFRA* and the activation of downstream MAPK signaling pathway in CAFs might contribute to the aggressiveness of SNF4 tumors, which also indicates potential therapeutic targets for SNF4 patients (Fig. 6h).

Discussion

Our multi-omic data suggested marked heterogeneity of HR$^+$/HER2$^-$ cancers in mutational, copy number, transcriptional, proteomic, metabolomic and pathological features. Here we successfully divided all HR$^+$/HER2$^-$ populations into the following four clusters: the canonical luminal subtype (SNF1), characterized by the enrichment of *PI3KCA* mutation; the immunogenic subtype (SNF2), enriched in immune cells and *TP53* mutation; the proliferative subtype (SNF3), enriched in cell cycle activation and high CIN and the RTK-driven subtype (SNF4), enriched in RTK pathway signatures. Finally, based on the heterogeneity of driving events, we further proposed precise treatments for each SNF subtype (Supplementary Fig. 3).

The lack of biomarkers guiding targeted therapies is a challenge in HR$^+$/HER2$^-$ breast cancer clinical management. Although endocrine therapies have markedly reduced the recurrence and mortality of ER$^+$ breast cancers[37], 10–41% of patients with operable ER$^+$ breast cancer who are treated with endocrine therapies eventually experience recurrence with metastatic disease[5,38]. Besides, the repertoire of targeted therapies for breast cancer is rapidly expanding, including endocrine therapies, CDK4/CDK6 inhibitors, PI3K/AKT/mTOR pathway inhibitors, RTK pathway inhibitors and immune checkpoint inhibitors[23,39]. To refine precision treatment strategies for HR$^+$/HER2$^-$ breast cancer, a more comprehensive understanding of its molecular heterogeneity is imperative.

To date, a large comprehensive multi-omics dataset of HR$^+$/HER2$^-$ breast cancers is still lacking. Compared with our dataset, previous studies, including METABRIC[9,40], Memorial Sloan Kettering Cancer Center (MSKCC)[41] and The National Cancer Institute's CPTAC[32], have also had their own limitations. Although METABRIC included a large number of HR$^+$/HER2$^-$ patients, the treatment received by patients in METABRIC may not reflect the current treatment standards, as it was based on past clinical practices[9,40]. Similar to METABRIC, the value of the MSKCC cohort was limited by the absence of transcriptomic, metabolomic and proteomic data[41]. Although CPTAC compensated for the lack of multi-omics data in the mentioned studies, its limited sample size, especially in HR$^+$/HER2$^-$ recurrent patients, restricted the ability to conduct a comprehensive exploration of the heterogeneity in HR$^+$/HER2$^-$ breast cancer[32]. Of note, our research provides a valuable data resource, which could extend our biological understanding of HR$^+$/HER2$^-$ breast cancers.

The PAM50 classification is the milestone molecular classification system for all breast cancers[10]. However, it needs refinement for HR$^+$/HER2$^-$ cancers for the following reasons: (1) PAM50 classification could not fully delineate the molecular heterogeneity for HR$^+$/HER2$^-$ cancers[32,42,43]. The main reason was that PAM50 classification was based on mRNA alone, which could not reflect the heterogeneity of CNA[44,45] or metabolomics[46,47]; (2) PAM50 classification could not accurately guide the rapidly expanding targeted therapies for individual patients[23,39]. Therefore, we extracted useful information using our multi-omics data that supplemented the PAM50 classification. For example, SNF1 together with SNF4 helped to subdivide luminal A breast cancers, where SNF1 represented a group of patients who had better prognoses and were most likely to be sensitive to endocrine therapy. Tumors of the SNF3 subtype were potentially sensitive to CDK4/CDK6 inhibitors and PARP inhibitors, further pinpointing a therapeutic strategy for a subgroup of luminal B tumors. Whether the immune-enriched tumor subtype (similar to our SNF2) or stroma-enriched tumor subtype (similar to our SNF4) could be regarded as the independent subtype is still controversial[48–50]. However, as we discovered, the tumors of the SNF2 subtype indicated the application of ICB, while tumors of the SNF4 subtype indicated the potential application of RTK inhibitors. Taken together, our study could help to better inform agent selection for individual HR$^+$/HER2$^-$ patients, which might hopefully increase treatment efficacy.

The application of multi-omics classification in clinical practice is challenging. The processing of multiple high-throughput datasets largely increases the difficulty of establishing a standardized pipeline, and its high cost also increases the financial burden on patients. Digital pathology might be a feasible way to solve these problems due to its convenience and accuracy. Importantly, deep learning-based digital pathology has made our integrative molecular classifications clinically applicable.

Our research has the following limitations. (1) The current conclusion is based on retrospective data analysis. The prospective clinical trials are needed to provide robust evidence for the current conclusions; (2) Further research is needed to examine the efficacy of the new drug across SNF subtypes. The emerging anti-HER2 antibody-drug conjugates (ADCs) are mainly applied to HER2 low-expression breast cancers that are characterized by low expression of the HER2 protein or absence of gene amplification in tumor cells, distinguishing it from HER2$^+$ breast cancer[51]. Some HR$^+$/HER2$^-$ breast cancers are also HER2 low-expression breast cancers. However, there was no significant difference in the prevalence of HER2 low-expression breast cancer among the SNF subtypes. Further investigation on the efficacy of anti-HER2 ADCs in SNF subtypes is important; (3) More experimental models are needed to validate the current phenotype. PDOs are not a suitable model for studying the ICB response due to the absence of immune cells[52,53]. Therefore, we only used SNF2 PDOs as controls for testing the efficacy of the agents that mainly target tumor cells (CDK4/CDK6 inhibitor and PARP inhibitor).

In conclusion, this study, based on a large-scale multi-omics cohort, revealed the heterogeneity of HR$^+$/HER2$^-$ breast cancer from the perspective of its molecular features. We further proposed precision treatment strategies that may pave the way for improved application of precision medicine in HR$^+$/HER2$^-$ breast cancer.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01507-7.

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 7–30 (2020).
2. Huppert, L. A., Gumusay, O., Idossa, D. & Rugo, H. S. Systemic therapy for hormone receptor-positive/human epidermal growth factor receptor 2-negative early stage and metastatic breast cancer. *CA Cancer J. Clin.* **73**, 480–515 (2023).
3. Ma, C. X., Reinert, T., Chmielewska, I. & Ellis, M. J. Mechanisms of aromatase inhibitor resistance. *Nat. Rev. Cancer* **15**, 261–275 (2015).
4. Dowsett, M. et al. Meta-analysis of breast cancer outcomes in adjuvant trials of aromatase inhibitors versus tamoxifen. *J. Clin. Oncol.* **28**, 509–518 (2010).
5. Pan, H. et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N. Engl. J. Med.* **377**, 1836–1846 (2017).
6. Park, Y. H. et al. Patterns of relapse and metastatic spread in HER2-overexpressing breast cancer according to estrogen receptor status. *Cancer Chemother. Pharmacol.* **66**, 507–516 (2010).
7. Pusztai, L. et al. Durvalumab with olaparib and paclitaxel for high-risk HER2-negative stage II/III breast cancer: results from the adaptively randomized I-SPY2 trial. *Cancer Cell* **39**, 989–998 (2021).

8. Nanda, R. et al. Effect of pembrolizumab plus neoadjuvant chemotherapy on pathologic complete response in women with early-stage breast cancer: an analysis of the ongoing phase 2 adaptively randomized I-SPY2 trial. *JAMA Oncol.* **6**, 676–684 (2020).
9. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
10. Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
11. Jiang, Y. Z. et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell* **35**, 428–440 (2019).
12. Jiang, Y. Z. et al. Molecular subtyping and genomic profiling expand precision medicine in refractory metastatic triple-negative breast cancer: the FUTURE trial. *Cell Res.* **31**, 178–186 (2021).
13. Gluz, O. et al. West German Study Group phase III plan B trial: first prospective outcome data for the 21-gene recurrence score assay and concordance of prognostic markers by central and local pathology assessment. *J. Clin. Oncol.* **34**, 2341–2349 (2016).
14. Wang, L. B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528 (2021).
15. East, M. P., Laitinen, T. & Asquith, C. R. M. PIP5K1A: a potential target for cancers with KRAS or TP53 mutations. *Nat. Rev. Drug Discov.* **19**, 436 (2020).
16. Semba, S. et al. Down-regulation of PIK3CG, a catalytic subunit of phosphatidylinositol 3-OH kinase, by CpG hypermethylation in human colorectal carcinoma. *Clin. Cancer Res.* **8**, 3824–3831 (2002).
17. Repana, D. et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.* **20**, 1 (2019).
18. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
19. Johnston, S. R. D. et al. Abemaciclib combined with endocrine therapy for the adjuvant treatment of HR^+, $HER2^-$, node-positive, high-risk, early breast cancer (monarchE). *J. Clin. Oncol.* **38**, 3987–3998 (2020).
20. Sledge, G. W. Jr. et al. The effect of abemaciclib plus fulvestrant on overall survival in hormone receptor-positive, ERBB2-negative breast cancer that progressed on endocrine therapy-MONARCH 2: a randomized clinical trial. *JAMA Oncol.* **6**, 116–124 (2020).
21. Turner, N. C. et al. Overall survival with palbociclib and fulvestrant in advanced breast cancer. *N. Engl. J. Med.* **379**, 1926–1936 (2018).
22. Slamon, D. J. et al. Phase III randomized study of ribociclib and fulvestrant in hormone receptor-positive, human epidermal growth factor receptor 2-negative advanced breast cancer: MONALEESA-3. *J. Clin. Oncol.* **36**, 2465–2472 (2018).
23. Mayer, E. L. et al. Palbociclib with adjuvant endocrine therapy in early breast cancer (PALLAS): interim analysis of a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol.* **22**, 212–222 (2021).
24. Mirza, M. R. et al. Niraparib maintenance therapy in platinum-sensitive, recurrent ovarian cancer. *N. Engl. J. Med.* **375**, 2154–2164 (2016).
25. Maman, S. & Witz, I. P. A history of exploring cancer in context. *Nat. Rev. Cancer* **18**, 359–376 (2018).
26. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
27. Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
28. Bartok, O. et al. Anti-tumour immunity induces aberrant peptide presentation in melanoma. *Nature* **590**, 332–337 (2021).
29. Burstein, H. J. Systemic therapy for estrogen receptor-positive, HER2-negative breast cancer. *N. Engl. J. Med.* **383**, 2557–2570 (2020).
30. Du, Z. & Lovly, C. M. Mechanisms of receptor tyrosine kinase activation in cancer. *Mol. Cancer* **17**, 58 (2018).
31. Östman, A. PDGF receptors in tumor stroma: biological effects and associations with prognosis and response to treatment. *Adv. Drug Deliv. Rev.* **121**, 117–123 (2017).
32. Krug, K. et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436–1456 (2020).
33. Gui, Y. et al. Metastatic breast carcinoma-associated fibroblasts have enhanced protumorigenic properties related to increased IGF2 expression. *Clin. Cancer Res.* **25**, 7229–7242 (2019).
34. Bertero, T. et al. Tumor-stroma mechanics coordinate amino acid availability to sustain tumor growth and malignancy. *Cell Metab.* **29**, 124–140 (2019).
35. Jungwirth, U. et al. Impairment of a distinct cancer-associated fibroblast population limits tumour growth and metastasis. *Nat. Commun.* **12**, 3516 (2021).
36. Perrone, F. et al. PDGFRA, PDGFRB, EGFR, and downstream signaling activation in malignant peripheral nerve sheath tumor. *Neuro Oncol.* **11**, 725–736 (2009).
37. Lin, N. U. & Winer, E. P. Advances in adjuvant endocrine therapy for postmenopausal women. *J. Clin. Oncol.* **26**, 798–805 (2008).
38. Hanker, A. B., Sudhan, D. R. & Arteaga, C. L. Overcoming endocrine resistance in breast cancer. *Cancer Cell* **37**, 496–513 (2020).
39. Loibl, S. et al. Palbociclib for residual high-risk invasive HR-positive and HER2-negative early breast cancer—the Penelope-B trial. *J. Clin. Oncol.* **39**, 1518–1530 (2021).
40. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
41. Razavi, P. et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* **34**, 427–438 (2018).
42. Patten, D. K. et al. Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. *Nat. Med.* **24**, 1469–1480 (2018).
43. Ades, F. et al. Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *J. Clin. Oncol.* **32**, 2794–2803 (2014).
44. Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C. & Perou, C. M. An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat. Genet.* **46**, 1051–1059 (2014).
45. Kim, J. A. et al. Comprehensive functional analysis of the tousled-like kinase 2 frequently amplified in aggressive luminal breast cancers. *Nat. Commun.* **7**, 12991 (2016).
46. Saito, Y. et al. LLGL2 rescues nutrient stress by promoting leucine uptake in ER^+ breast cancer. *Nature* **569**, 275–279 (2019).
47. Golden, E. et al. The oncogene AAMDC links PI3K-AKT-mTOR signaling with metabolic reprograming in estrogen receptor-positive breast cancer. *Nat. Commun.* **12**, 1920 (2021).
48. Huang, C. et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* **39**, 361–379 (2021).
49. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225 (2020).
50. Petralia, F. et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. *Cell* **183**, 1962–1985 (2020).

51. Modi, S. et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N. Engl. J. Med.* **387**, 9–20 (2022).
52. Dijkstra, K. K. et al. Generation of tumor-reactive T cells by co-culture of peripheral blood lymphocytes and tumor organoids. *Cell* **174**, 1586–1598 (2018).
53. Neal, J. T. et al. Organoid modeling of the tumor immune microenvironment. *Cell* **175**, 1972–1988 (2018).
54. Gao, Q. et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238 (2018).

Methods

Specimens and clinical data

The present study entailed a retrospective selection of patients who had been diagnosed with malignant breast cancer. All aspects of this study including the collection of all tissue samples were carried out with approval of the Fudan University Shanghai Cancer Center (FUSCC) Ethics Committee. All study participants provided written informed consent before their inclusion in the study. No participants received compensation. Samples were collected taking into account the quantity and quality of the tissues for analyses as well as the availability of the corresponding clinicopathological data. Our cohort included pretreatment patients treated at FUSCC between January 2013 and December 2014 without intentional selection ($n = 478$). We also selected patients who experienced relapse after surgery between January 2009 and December 2016 ($n = 101$) to profile the patients with relatively high risk. Moreover, the following additional criteria had to be satisfied: (1) patients included in the study had to be diagnosed with unilateral invasive carcinoma exhibiting HR^+ (ER^+ or PR^+) and $HER2^-$ phenotype; (2) pathologic examination of tumor specimens was conducted by the Department of Pathology at FUSCC, where the ER, PR and HER2 status were independently verified by two experienced pathologists using immunochemical analysis and in situ hybridization. ER/PR negativity in IHC testing was defined as having less than 1% positively stained cells, following the guidelines set forth by the American Society of Clinical Oncology/College of American Pathologists[55]; (3) exhibiting no signs of distant metastasis at the time of diagnosis; (4) without receiving any treatments and (5) availability of sufficient frozen tissue samples for further investigation. Patients with inflammatory breast cancer or with breast carcinoma in situ (with or without microinvasion) were deliberately excluded from the study. The clinicopathological features assessed in this study comprised age, tumor size, tumor histologic type, histologic grade, lymph node status, adjuvant therapies administered as well as ER, PR, HER2 and Ki67 status. Disease extent, which was assessed by chest computed tomography, bone scan, abdominal ultrasound, bilateral mammography, breast ultrasound and/or magnetic resonance imaging was recorded.

Follow-up within this cohort of patients was completed on June 30, 2021, and the median length of follow-up was 83.0 months (IQR = 71.0–91.8 months). MFS was defined as the time from the date of surgery to the first distant metastasis detection. Patients without events were censored from the time point of the last follow-up.

Age, sex and other demographic characteristics of human participants providing samples are presented in Supplementary Tables 1 and 7.

Evaluation of pathological indicators

Based on the pathological WSIs that we collected, TILs were evaluated as the area occupied by mononuclear inflammatory cells over total intratumoral stromal area according to the recommendations by the International TILs Working Group[56].

Sample processing for genomic DNA and total RNA extraction

Fresh frozen tumor tissues underwent macrodissection to minimize the potential influence of stromal tissues (<30% stromal tissue), thereby ensuring a confirmed presence of 50% or more tumor cells in all breast cancer specimens. We purified the genomic DNA from fresh frozen samples and peripheral blood cells using TGuide M24 (Tiangen). The assessment of DNA purity and quantity involved measuring the absorbance at 260 nm (A260) and 280 nm (A280) using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). The extracted DNA with an A260/A280 ratio ranging from 1.6 to 1.9 was regarded as pure and suitable for subsequent experimental procedures. According to the manufacturer's instructions, we purified the total RNA from tissues that had been previously stored in RNAlater solution using the miRNeasy Mini Kit (Qiagen, 217004). The integrity of RNA was assessed using an Agilent 4200 Bioanalyzer alongside RNA ScreenTape (Agilent), while concentrations were determined by a NanoDrop ND-8000 spectrophotometer (Thermo Fisher Scientific).

PAM50 classification

PAM50 subtypes were determined based on the PAM50 classifier[57,58]. First, the mRNA-seq data were subsampled to ensure that the proportion of IHC subtypes of the samples aligned with the training set employed for PAM50. Second, normalization was carried out on the fragments per kilobase of transcript per million mapped reads data by adjusting the gene expression level to the median expression calculated from the PAM50 gene set within the IHC balanced subset. Finally, PAM50 typing was conducted.

TME evaluation

The immune and stromal signatures were calculated by the ESTIMATE algorithm (v1.0.13)[59]. TIS and STAT1 signatures were assessed based on the RNA-seq data using the following steps: (1) mean-centering, (2) averaging over genes and (3) z scoring[7]. Gene signatures modified from CIBERSORT[60] and MCP-counter[27] were used for single-sample GSEA ('gsva' function in R package 'GSVA' v1.42 (ref. 61)) to evaluate the abundance of different immune cells in each sample with expression data[62].

GSEA and gene set variation analysis (GSVA)

GSEA was performed to explore enriched pathways and interpret RNA-seq data using predefined gene sets from the Molecular Signatures Database (v.7.1) in GSEA software (v4.0)[63,64]. All basic and advanced fields were set to default.

In the subsection Digital pathology data collection and preprocessing, single-sample GSEA scores were calculated for each sample using the 'gsva' function in the R package 'GSVA' v1.42 (ref. 61). Seven breast cancer-related hallmark gene sets of potential clinical relevance were selected.

REACTOME, Gene Ontology and hallmark gene sets were used in Extended Data Fig. 8d. Gene Ontology molecular function gene sets are shown in Extended Data Fig. 9a,b (obtained from http://www.gsea-msigdb.org/gsea/msigdb).

Estimation of multigene proliferation scores (MGPS)

To calculate MGPS, the mean expression level was calculated for all genes previously identified as exhibiting regulation during the cell cycle[65].

Estimation of HRD score and CIN scores

The HRD score was the sum of the following three independent scores: telomeric allelic imbalance (NtAI) score, loss of heterozygosity (LOH) score and large-scale state transitions (LST) score. All these scores were calculated using copy number data derived from allele-specific copy number analysis of tumors (ASCAT v2.4.3) according to previous studies[66,67]. Briefly, the NtAI score represented the count of subchromosomal regions longer than 11 Mb and exhibited allelic imbalance extending to the telomere. The LOH score was calculated as the number of LOH regions longer than 15 Mb but shorter than the whole chromosome (LOH regions located on chromosome 17 were excluded). The LST score was defined as the number of breakpoints between two chromosomal regions longer than 10 Mb after regions shorter than 3 Mb were smoothed out[66–70]. The CIN score was calculated by summing the squared gene-level GISTIC 2 values[71].

Digital pathology data collection and preprocessing

We aimed to develop a deep learning-based approach to infer the SNF subtypes from pathological WSIs. We collected paraffin-embedded, H&E-stained tumor slides and scanned them using the NanoZoomer S210 digital pathology scanner at ×40 magnification to generate digital WSIs. All of the WSIs and the corresponding pathological reports were rereviewed by pathologists, and the following cases were

excluded: (1) slides or WSIs of poor quality (large areas of debris, folds, pen marks or blurred regions); (2) cases of microinvasive breast cancer. A final dataset of 243 WSIs from 243 cases with available multi-omics SNF classification results was determined for the following analysis.

On each WSI, we drew one or two rectangular regions of interest (ROIs) using ImageScope software (v12.4.0.5043), which included the great majority of invasive cancers and excluded many areas of background and uninformative dragged tissue. Images within the ROIs were tessellated into nonoverlapping 256 × 256-pixel image tiles using MATLAB software (v2021a) and the OpenSlide library[72]. The blurred and noninformative background tiles (on which more than half of the pixels were >210) were discarded.

An intact WSI is composed of several tissue types. For the prediction of the four SNF subtypes, image tiles of appropriate tissue types should be selected for model development. Here a tissue type classifier established in our previous study was applied to all image tiles and classified into five tissue types, including tumor, stroma, immune infiltrates, normal breast gland and necrosis/hemorrhage[73].

Deep-learning model development

One versus rest strategy and threefold cross-validation were adopted to develop the models for identifying each of the four SNF subtypes[73]. In detail, all of the 243 patients with WSIs were divided into three partitions. The distribution of the SNF subtypes was kept balanced across the partitions. For the model development of each SNF subtype, patients belonging to this subtype were assigned a 'positive' label and those of other SNF subtypes were assigned a 'negative' label. A total of 500 image tiles were sampled from each patient's WSI, and all these tiles inherited the label of the corresponding patient. Patient-level threefold cross-validation was used to train and validate the CNN models. That is, the CNN models were trained on the tiles from two partitions and validated on those from the remaining partition. The number of tiles per class in the training set was equalized by downsampling. Because the division was done at the patient level, no tiles from the same patient were used for training and validation at the same time.

Deep-learning-based model development was performed in Python with Pytorch (v1.10.0)[74]. The ResNet-18 pretrained model (https://download.pytorch.org/models/resnet18-5c106cde.pth) was used for fine-tuning. Data augmentation was randomly applied including color jitter with brightness, contrast and saturation. The other hyperparameters were set as follows: loss function, cross-entropy loss; batch size, 256; learning rate, 0.001 and optimizer, Adaptive Moment Estimation (ADAM). During the training process, each tile obtained a prediction score produced by the model. The patient-level scores were calculated by averaging all of the tiles' scores from the corresponding patient and were used for receiver operating characteristics (ROC) analysis. Models were trained for 100 epochs, and the best model was determined according to the AUC in the validation set.

Visualization analysis was performed to reveal the typical pathological features of the tumors from each SNF subtype. First, we displayed and visually characterized the morphologic patterns of the representative image tiles that were from the true positive patients and had the highest prediction score. Then, class activation maps were generated for these tiles to highlight the discriminative subregions that were used by the CNN models to make the prediction[75]. All tiles presented in our study were among the top 100 tiles according to the prediction score.

Inference of SNF subtypes for new patients

Our developed models were applied to the unlabeled new WSIs to infer their SNF subtypes, including 25 patients whose tumor samples were used to develop PDOs. In detail, first, these unlabeled WSIs were collected and preprocessed as in our multi-omics cohort. Then, the four determined models (one for each SNF subtype) were applied to them. After this, each patient obtained a binary classification result from each of the four models. If the patient was classified as 'positive' by only one model, it would be determined as the corresponding SNF subtype. Once the patient was classified as 'positive' by more than one model, the result from the model which had the highest accuracy (measured by AUC) was adopted.

Transcriptomics data-based method for SNF subtype classification

Referring to the methods discussed in ref. 76, we developed a random forest classifier for the probabilistic classification of $HR^+/HER2^-$ tumors into SNF subtypes using transcriptomics data. First, we performed differential expression analysis through the R packages DESeq2 (v1.34) and retrieved the top 100 highly expressed genes of each subtype according to the $\log_2$(fold change) value[77]. Second, we performed collinearity removal to minimize the candidate gene sets. We removed all genes with a mutual Spearman correlation coefficient higher than 0.75. Only the one that showed the highest correlation with the corresponding SNF subtype was retained. Finally, based on the expression of the remaining genes, we developed a random forest classifier and measured its accuracy using ROC analysis.

In vitro viability assay of PDOs

PDO samples were collected freshly from patients who underwent surgery between August 2019 and August 2021 at FUSCC. Overall, 25 $HR^+/HER2^-$ PDO samples were eligible for further experiments. The generation of PDO was conducted as described in a previous study[78]. Briefly, breast cancer tissue was cut into 1–3 mm^3 pieces that were subsequently digested using collagenase (Sigma-Aldrich). The resulting organoids were then cultured in 24-well plates, suspended in basement membrane extract (BME) type 2 (Trevigen). Organoids were diluted to a concentration of 40 organoids per ml in breast cancer organoid medium supplemented with 10% BME. A volume of 25 ml of the organoid suspension was then added to cell-repellent surface black, clear bottom 384-well plates (Greiner Bio-One, 781976-SIN) and cultured for an additional 5–6 d before initiating drug treatments. A volume of 25 ml of organoid suspension was then added to cell-repellent surface black, clear bottom 384-well plates (Greiner Bio-One, 781976-SIN) and cultured for an additional 5–6 d before initiating drug treatments. Organoid cell viability was accessed by a CellTiter-Glo 3D Cell viability assay (Promega, G9683) following the manufacturer's instructions. Compounds including olaparib, abemaciclib, 4-hydroxytamoxifen and sorafenib were purchased from Selleck.

IHC analysis

Paraffin-embedded tissue sections were deparaffinized at 60 °C for 20 min, cleared in xylene and subjected to a series of graded alcohols. H&E staining involved the use of Mayer's hematoxylin (Sigma-Aldrich) and 0.1% sodium bicarbonate, with subsequent counterstaining with eosin Y solution (Sigma-Aldrich). IHC required a heat treatment using saline-sodium citrate buffer at 95–100 °C. Following cooling, the slides were blocked at room temperature using a blocking solution (2% goat serum, 2% BSA and 0.05% Tween in PBS) and incubated with a primary antibody diluted in blocking solution at 4 °C. Endogenous peroxidase activity was quenched using 0.3% H_2O_2. Slides were then incubated with horseradish peroxidase (HRP)-conjugated secondary antibody (Genetech) at room temperature, followed by development with a 3,3′-diaminobenzidine substrate (Genetech). Counterstaining was performed using hematoxylin, and dehydration was carried out using graded alcohol solutions. Positive-staining density was quantified using a computerized imaging system composing a Leica charge-coupled device DFC420 camera connected to a Leica DM IRE2 microscope (Leica Microsystems Imaging Solutions). The densities were determined by counting positive cells on ×10 high-power field of view (~2 mm^2). Phospho-p44/42 MAPK (Erk1/Erk2; Cell Signaling Technology, 4370; 1:200), phosphor-RB1 (Ser807/Ser811; D20B12; Cell

Signaling Technology, 8516; 1:200), CD20 (Abcam, ab78237; 1:100), CD8A (Servicebio, GB12068; 1:1500), αSMA (Servicebio, GB13044; 1:200) and CD206 (Servicebio, GB113497; 1:400) were used. Detection System/Mo&Rb (Genetech, GK6007) was used as secondary antibody.

Isolation, culture and coculture of primary human breast cancer CAFs

Fresh human breast cancer tissues from patients were cut into 1 mm^3 pieces and digested using collagenase (Sigma-Aldrich). The digested cells were plated on a six-well plate in CAF media (DMEM/F12, 20% FBS)[79]. Plates were plated in an incubator, and media were replaced every 2 d.

For the CAF CM experiment[79], CAFs were seeded at 1×10^6 cells per 10 cm plate. Each plate was supplemented with 5 ml of DMEM/F12 medium (Gibco). After culturing for 48 h, the medium was collected, centrifuged and filtered. The CAF CM was obtained by replacing the conventional DMEM/F12 in the standard PDO medium with the CAF cultured media.

For the tumor and CAF coculture model, an experimental model was processed[79]. In brief, first, tumor cells were isolated using flow cytometry as EpCAM-positive cells and cultured as tumorspheres. Then, 2,000 CAFs were plated in 96-well plates. After 24 h, the media was removed and 200 tumorspheres cultured in 100 µl Matrigel (Corning) were added to the well preplated with CAFs and cultured for 48 h before drug treatment.

scRNA-seq data processing

The Cell Ranger software pipeline (v3.1.0) provided by 10X Genomics was used for demultiplexing cellular barcodes, aligning reads to the genome and transcriptome using the STAR aligner, downsampling reads as required to generate normalized aggregate data across samples. This process resulted in a matrix of gene counts versus cells. The unique molecular identifier (UMI) count matrix was further analyzed using the R package Seurat[80] (v4.0.5). The R packages scDblFinder[81] (v1.6.0) and DropletUtils[82] (v1.12.3) were used to distinguish doublets and empty droplets from singlets, respectively. To remove low-quality cells, cells with UMI/gene numbers less than 200 or more than 8,000 were filtered out. We further discarded low-quality cells where >20% of the counts belonged to mitochondrial genes after visual inspection. We performed scRNA on nine patients during the project (September 2021 to January 2022). After applying these quality control (QC) criteria, there were a total of 59,479 single cells remained for further downstream analyses. To obtain the normalized count, we conducted library size normalization in Seurat on our filtered matrix.

Cell clustering, visualization and annotation

The samples from different patients were collected and sequenced on different days, thus introducing batch effects on downstream analysis. Harmony (v0.1.0)[83] was applied to integrate the data from different batches with group.by.vars = orig.ident, plot_convergence = TRUE. The batch-corrected matrix after batch correction was input into Seurat, and the standard Seurat workflow was performed. We used the top 50 principal components with a resolution of 0.6 for Louvain clustering and used uniform manifold approximation and projection for visualization. For cell cluster annotation, we used a computational method named SingleR (v1.6.1)[84], inferred the origin of cells independently and identified cell types with the reference single-cell type transcriptomic dataset 'Human Primary Cell Atlas'[85].

Single-cell copy number calling and cancer cell identification

The R package 'inferCNV' (v1.8.1) was used with devalued parameters to estimate the copy number variations for individual cells. A subset of immune cells was used as a reference to define the baseline copy number profiles, and other cells were used for observation. To identify neoplastic cells from epithelial cells, all the epithelial cells and reference cells used in inferCNV were clustered into 15 subtypes by *k*-means ('kmeans' function in R) based on the copy number profiles of each genomic locus. Epithelial cells clustered with reference cells with fewer CNV alterations were defined as normal cells, whereas epithelial cells clustering individually with more CNV alterations were defined as tumor cells.

Estimation of dysfunction and cytotoxic score

Dysfunction and cytotoxic scores were assigned using the AddModuleScore function in Seurat. The AddModuleScore function in the Seurat package (v4.0.5) was used to calculate gene expression module scores for each cell, using the default parameters. The differences in scores among cell subsets were compared[86].

Western blotting

Cell lysis was conducted using SDS lysis buffer containing 50 mM Tris (pH 8.1), 1 mM EDTA, 1% SDS, 1 mM fresh dithiothreitol, sodium fluoride and leupeptin. Following lysis, the lysates underwent centrifugation at 10,000*g* for 20 min, whereby the supernatants were collected. Protein concentrations in the supernatants were determined using the bicinchoninic acid Protein Assay Kit (Pierce). A total of 30–60 µg of protein was subjected to SDS–polyacrylamide gel electrophoresis for separation, followed by transfer to polyvinylidene difluoride membranes (Millipore). The primary and secondary antibodies, as mentioned above, were used, and the enhanced chemiluminescence substrate used was the SuperSignal West Femto Substrate Trial Kit (Pierce). Epidermal growth factor (EGF) receptor (Cell Signaling Technology, 4267; 1:1,000), phospho-EGF receptor (Tyr1068; Cell Signaling Technology, 3777; 1:1,000), PDGF receptor α (Cell Signaling Technology, 3174; 1:1,000), phospho-PDGF receptor α (Tyr1018; Cell Signaling Technology, 4547; 1:1,000), αSMA (Abcam, ab124964; 1:1,000), β-actin (Proteintech, 66009-1-Ig; 1:20,000) were used. Anti-rabbit IgG, HRP-linked antibodies (Cell Signaling Technology, 7074) were used as secondary antibody for EGF receptor, phospho-EGF receptor (Tyr1068), PDGF receptor α, phospho-PDGF receptor α (Tyr1018) and αSMA. Anti-mouse IgG, HRP-linked antibody (Cell Signaling Technology, 7076) was used as a secondary antibody for β-actin.

Statistics and reproducibility

This study was based on the retrospective cohort of $HR^+/HER2^-$ breast cancers. Our cohort included pretreatment patients treated at FUSCC between January 2013 and December 2014 without intentional selection ($n = 478$). We also selected patients who experienced relapse after surgery between January 2009 and December 2016 ($n = 101$) to profile the patients with relatively high risk. Statistics were performed as described in the legends of Figs. 2a-c, 3a-g, 4, 5a,c-i, 6a, 6e, Extended Figs. 1f-g, 2, 3, 4, 5c-e, 6b-e, 7a, 7c-d, 8e-f, 9, 10a-b, 10d and Methods sections. All experiments including western blot, IHC and viability assay of PDOs were repeated three times with similar results. All uncropped western blots are provided in the Source Data. No statistical method was used to predetermine the sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Inclusion and ethics statement

The research included local researchers throughout the research process. The research is locally relevant. All the roles and responsibilities were agreed upon among collaborators ahead of the research. This research was not severely restricted or prohibited in the setting of the researchers. This study was approved by a local ethics review committee. The research does not result in stigmatization, incrimination, discrimination or otherwise personal risk to participants. We do not take local and regional research relevant to our study into account in citations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The WES data, CNA data, RNA sequencing data and metabolome data for this study have been deposited into the Genome Sequence Archive (GSA) database under accession codes PRJCA017539 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017539). TMT-based mass spectrometry (MS)-quantified protein data have been submitted into iProX (https://www.iprox.cn) under accession codes IPX0006535000. Human Primary Cell Atlas data are obtained from the celldex package (v1.11; https://github.com/LTLA/celldex). The TCGA, METABRIC and CPTAC data were downloaded from the cBioPortal website (https://www.cbioportal.org/). Source data are provided with this paper.

Code availability

All data were analyzed and processed using published software packages whose details are provided and cited either in the Methods section or Supplementary Note. The CNN models and code from this manuscript are available at GitHub (https://github.com/yifanzhou330/SNF) and Zenodo (https://doi.org/10.5281/zenodo.8022438)[87].

References

55. Hammond, M. E. et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Arch. Pathol. Lab. Med.* **134**, e48–e72 (2010).
56. Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).
57. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
58. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
59. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
60. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).
61. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
62. Xiao, Y. et al. Multi-omics profiling reveals distinct microenvironment characterization and suggests immune escape mechanisms of triple-negative breast cancer. *Clin. Cancer Res.* **25**, 5002–5014 (2019).
63. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
64. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
65. Whitfield, M. L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
66. Timms, K. M. et al. Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **16**, 475 (2014).
67. Telli, M. L. et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).
68. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
69. Birkbak, N. J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
70. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
71. Ock, C. Y. et al. Genomic landscape associated with potential response to anti-CTLA-4 treatment in cancers. *Nat. Commun.* **8**, 1050 (2017).
72. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).
73. Zhao, S. et al. Deep learning framework for comprehensive molecular and prognostic stratifications of triple-negative breast cancer. *Fundam. Res.* (2022).
74. Paszke, A., Gross, S., Massa, F., Lerer, A. & Chintala, S. PyTorch: an imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Article 721 (Curran Associates Inc., 2019).
75. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929 (IEEE, 2016).
76. Migliozzi, S. et al. Integrative multi-omics networks identify PKCδ and DNA-PK as master kinases of glioblastoma subtypes and guide targeted cancer therapy. *Nat. Cancer* **4**, 181–202 (2023).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
78. Sachs, N. et al. A living biobank of breast cancer organoids captures disease heterogeneity. *Cell* **172**, 373–386 (2018).
79. Grunwald, B. T. et al. Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell* **184**, 5577–5592 (2021).
80. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
81. McGinnis, C., Murrow, L. & Gartner, Z. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
82. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
83. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
84. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
85. Karlsson, M. et al. A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2021).
86. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
87. Zhou, Y. Molecular classification of hormone receptor-positive HER2-negative breast cancer. *Zenodo* https://doi.org/10.5281/zenodo.8022438 (2023).

Acknowledgements

We are grateful to the patients and their families who contributed to this study. This work was supported by grants from the National Key Research and Development Project of China (2021YFF1201300), the National Natural Science Foundation of China (82341003,

91959207, 92159301, 82272822, 82272704 and 82103039), the Shanghai Key Laboratory of Breast Cancer (12DZ2260100), the Shanghai Hospital Development Center (SHDC) Municipal Project for Developing Emerging and Frontier Technology in Shanghai Hospitals (SHDC12021103), the Program of Shanghai Academic/Technology Research Leader (20XD1421100), the Natural Science Foundation of Shanghai (22ZR1479200 and 23ZR1411800), Youth Talent Program of Shanghai Health Commission (2022YQ012), China Postdoctoral Science Foundation (2022M720790), Shanghai Sailing Program (20YF1408700) and Youth Medical Talents of Shanghai (WJWRC202014). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

X.J., Y.Z.J. and Z.M.S. conceived and designed the study. Y.F.Z., D.M., C.J.L. and C.L.L. performed the proteomics and contributed to the data processing and analyses. X.J. and Y.Z.J. wrote the first draft and organized the figures. S.Z. reviewed the pathological sections and performed the deep-learning-based digital pathology. Y.X. and W.X.X. performed the metabolomics. T.F. performed scRNA-seq. Y.Y.C. carried out IHC experiments and PDO assays. Y.Q.L., Q.W.C., Y.Y., J.X.S., L.M.S. and W.H. performed the WES, OncoScan and RNA sequencing. J.F.R. and Z.M.S. supervised all aspects of the study.

Competing interests

The authors declare no competing interests.

Additional information

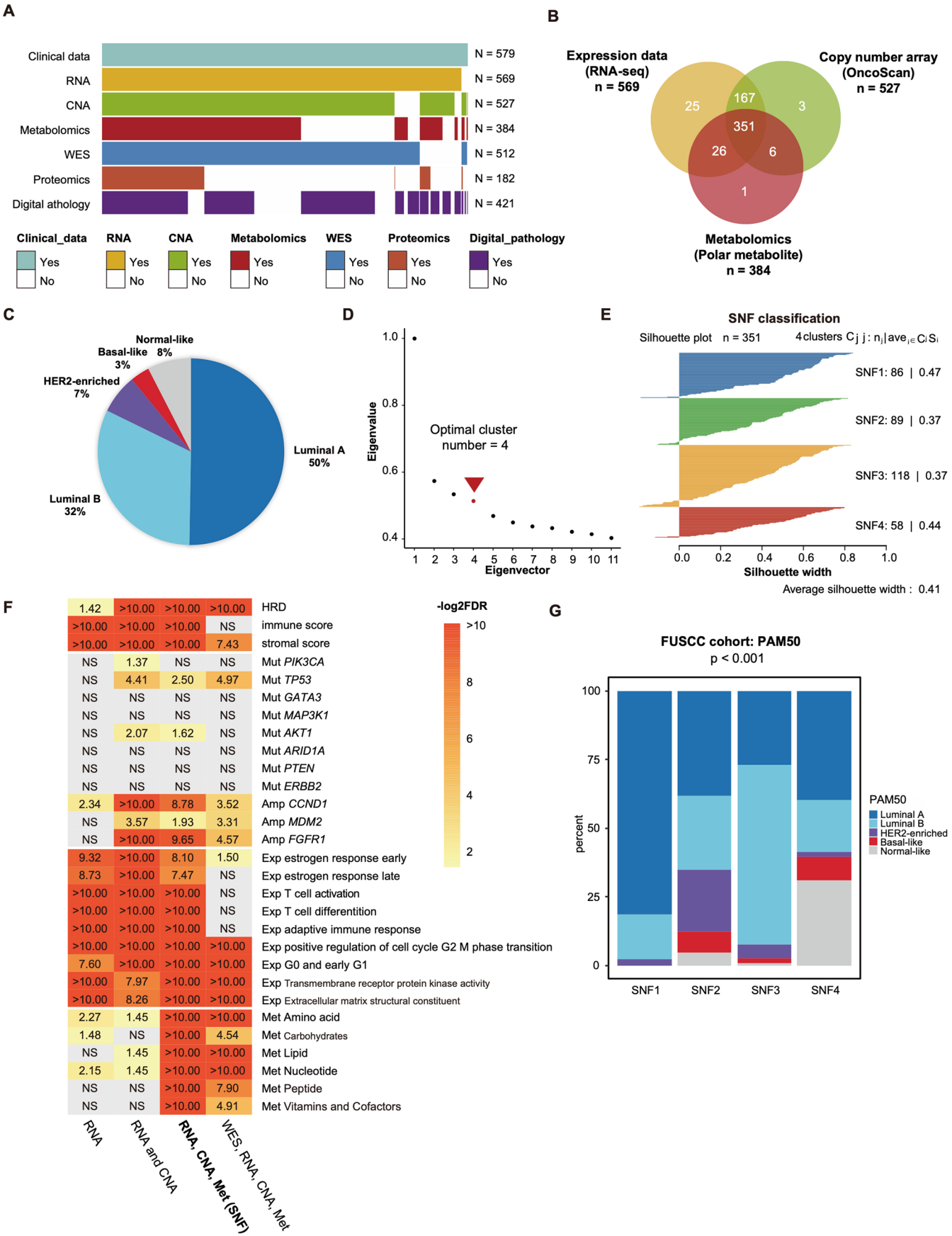
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01507-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01507-7.

Correspondence and requests for materials should be addressed to Yi-Zhou Jiang or Zhi-Ming Shao.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work

https://doi.org/10.1038/s41588-023-01507-7

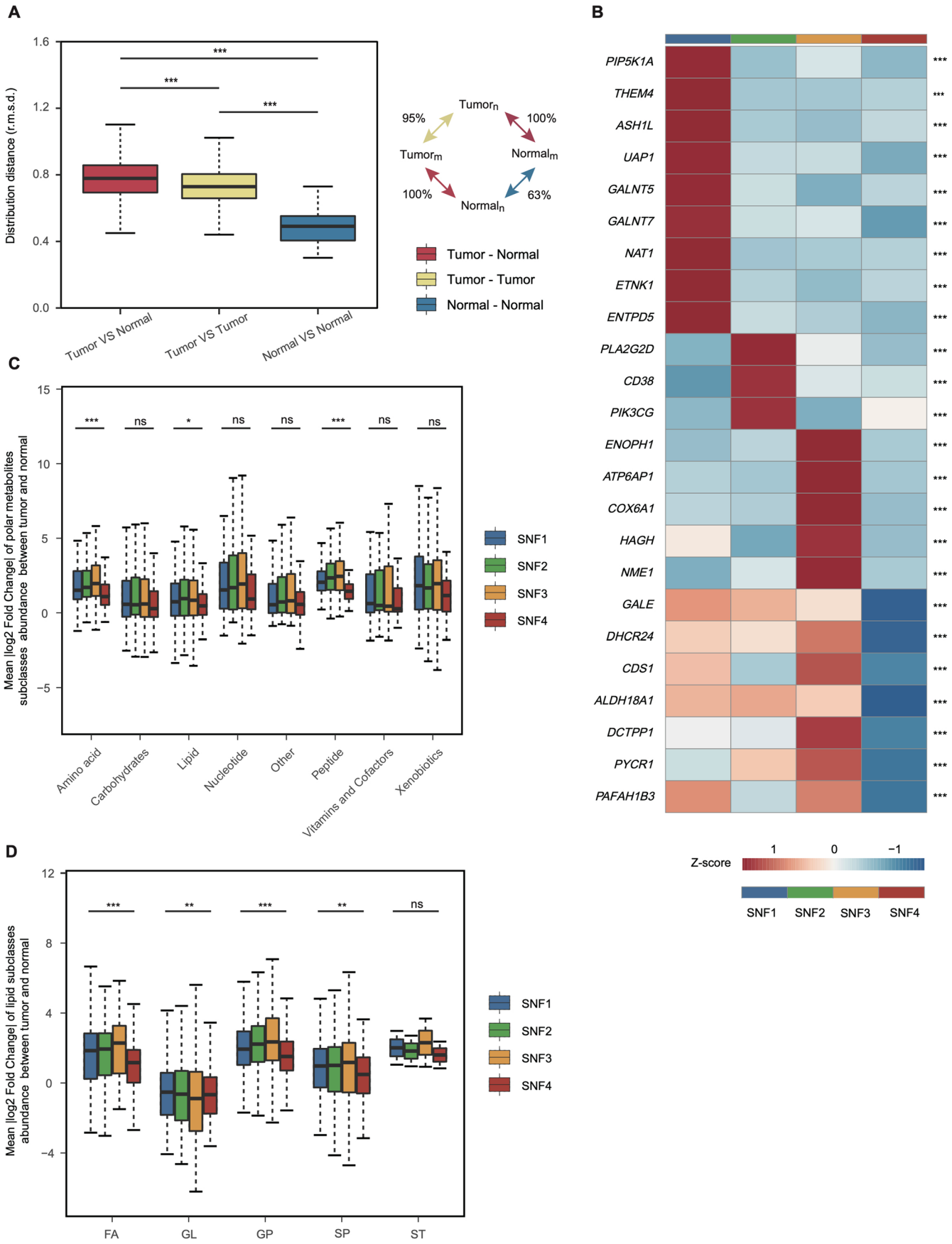


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-023-01507-7

Extended Data Fig. 1 | Landscape of FUSCC HR+/HER2- breast cancer cohort, related to Fig. 1. (**a**, **b**) Schematic overview of multi-omics data acquired for this cohort. (**c**) The proportion of PAM50 subtypes. (**d**) Determination of optimal cluster number. (**e**) Silhouette plot of SNF clustering. (**f**) Summary of adjusted p-value for differences in each multi-omic features among subtypes under different clustering strategies. NS: not significant. Mut: mutation. Amp: amplification. Met: metabolite. HRD: homologous recombination deficiency. Bold font indicates the clustering strategy we used. (**g**) Distribution of PAM50 subtypes among SNF subtypes. P = 5e-04. P values were from the two-sided Fisher's exact test.

https://doi.org/10.1038/s41588-023-01507-7

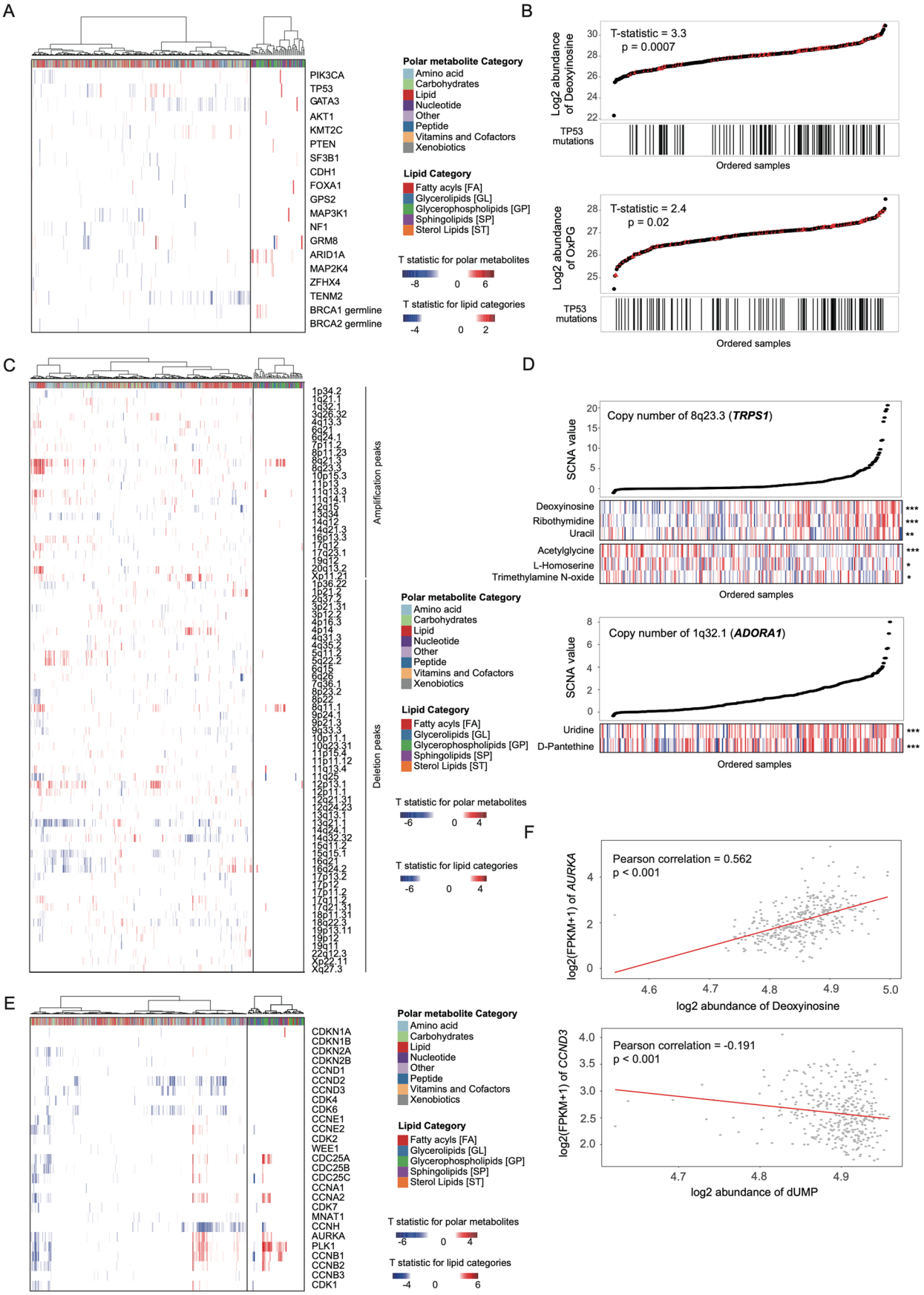


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-023-01507-7

Extended Data Fig. 2 | SNF subtype-specific metabolomic features, related to Fig. 2. (**a**) Global differences in metabolic gene expression between tumors and normal tissues in the luminal cohort. The distribution distances (r.m.s.d) were calculated between tumors and corresponding normal tissues (red), different samples of tumor tissues (yellow), and different samples of normal tissues (blue). The inset shows the average distances between pairs of tissues as a percentage of the average distance between tumors and normal tissues. Tumor samples n = 351, normal samples n = 11. P(T VS N - T VS T) < 2.2e-16, P(T VS N - N VS N) < 2.2e-16, P(T VS T - N VS N) < 2.2e-16. *P* values were from the two-sided Wilcox rank-sum test and Kruskal-Wallis test. (**b**) Heatmap illustrating subtype-specific metabolic genes. *P* values were from the two-sided Kruskal-Wallis test. (**c**) Illustration of subtype-specific polar metabolite subclasses. *P*(Amino acid) = 4.952e-05, *P*(Carbohydrates) = 0.4084, *P*(Lipid) = 0.02426, *P*(Nucleotide) = 0.3861, *P*(Other) = 0.6101, *P*(Peptide) = 0.0002545, *P*(Vitamins and Cofactors) = 0.8432, *P*(Xenobiotics) = 0.7095. *P* values were from the two-sided Kruskal-Wallis test. SNF1 subtype n = 86 biologically independent samples, SNF2 subtype n = 89 biologically independent samples, SNF3 subtype n = 118 biologically independent samples, SNF4 subtype n = 58 biologically independent samples, normal samples n = 11. (**d**) Illustration of subtype-specific lipid subclasses. *P*(FA) = 0.0005, *P*(GL) = 0.008, *P*(GP) = 9.113e-14, *P*(SP) = 0.0005, *P*(ST) = 0.0005. *P* values were from the two-sided Kruskal-Wallis test. SNF1 subtype n = 86 biologically independent samples, SNF2 subtype n = 89 biologically independent samples, SNF3 subtype n = 118 biologically independent samples, SNF4 subtype n = 58 biologically independent samples, normal samples n = 11. FA: Fatty acyls. GL: Glycerolipids. GP: Glycerophospholipids. SP: Sphingolipids. ST: Sterol Lipids. In all boxplots, the center lines represent median values; the bounds of the boxplot represent the interquartile ranges; the whiskers show the range of the data. All *P* values were adjusted using the Benjamini–Hochberg procedure. ***FDR < 0.001; **0.001 ≤ FDR < 0.01; *0.01 ≤ FDR < 0.05; ns, FDR ≥ 0.05.

https://doi.org/10.1038/s41588-023-01507-7

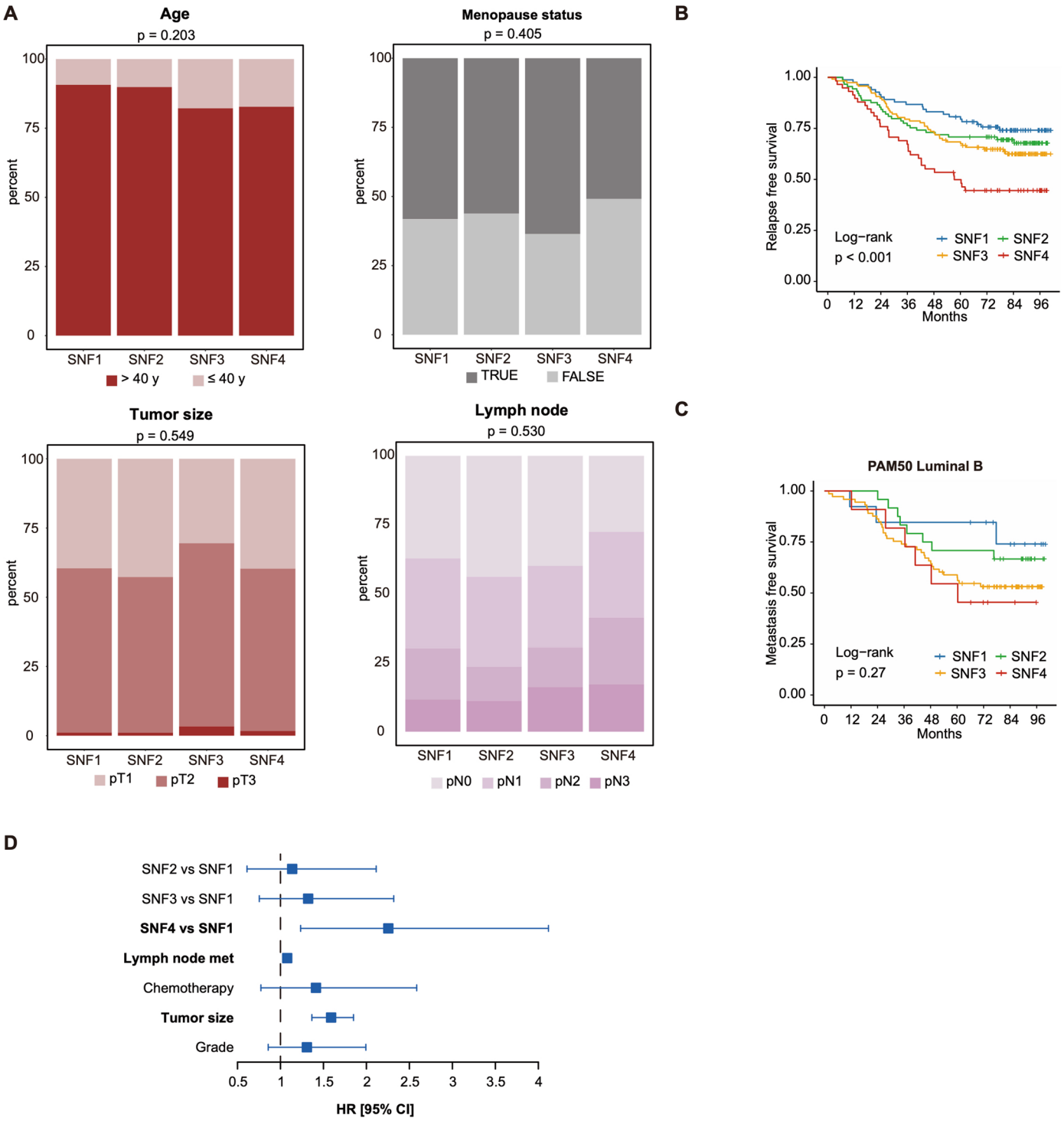


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-023-01507-7

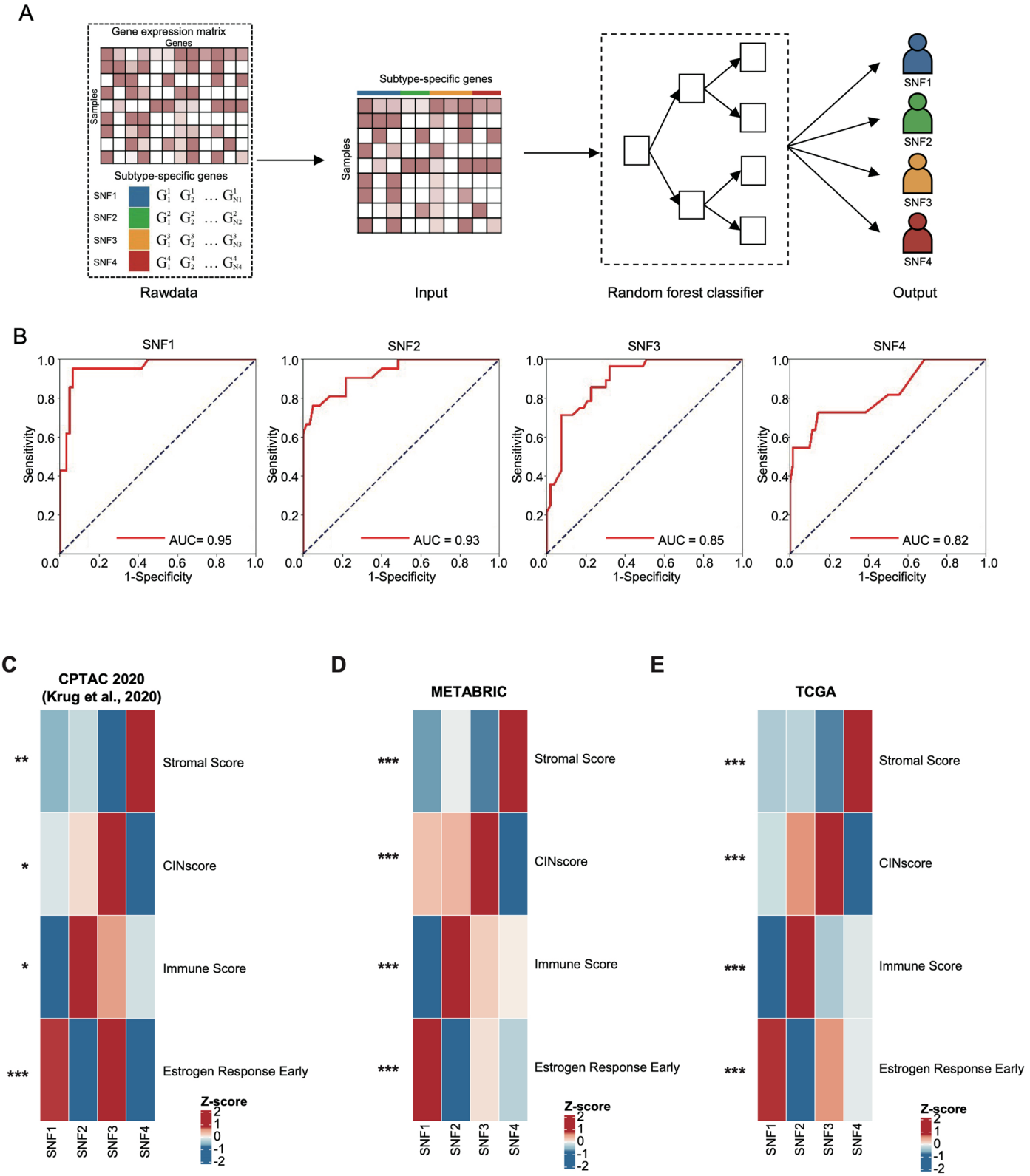
Extended Data Fig. 3 | The associations of polar metabolites and lipids with genomic features, related to Fig. 2. (**a**) Heatmap showing the associations between the abundances of metabolites and the presence of mutations within the indicated genes. The mutations include high frequency somatic mutations (mutated in at least 6% of the cases in at least one SNF subtype) within cancer-related genes and high frequency germline mutations in *BRCA1* and *BRCA2*. T statistics were calculated by a linear regression model that adjusted the cofounding factors. (**b**) Correlations between *TP53* mutations and deoxyinosine (top panel) and OxPG levels (bottom panel). All samples were ordered based on the abundance (y-axis) of deoxyinosine (top panel) or OxPG (bottom panel), and those with *TP53* mutations were highlighted in red and indicated by the corresponding lines displayed on the x-axis. Two-sided T statistics were calculated. (**c**) Heatmap showing the associations between the abundances of metabolites and copy number values of SCNA peaks. T statistics were calculated by a linear regression model that adjusted the cofounding factors. (**d**) Top panel: correlations between the copy number values of 8q23.3 and the abundances of deoxyinosine, ribothymidine, uracil and some amino acids. Bottom panel: correlations between the copy number values of 1q32.1 and the abundances of uridine and D-pantethine. SCNA-related metabolites were shown as lines, and samples were ordered by increasing copy number values. The abundances of the metabolites were illustrated in colors. (**e**) Heatmap showing the correlations between the mRNA expression of cell cycle-related genes (y-axis) and the abundances of metabolites (x-axis). T statistics were calculated by a linear regression model that adjusted the cofounding factors. (**f**) The correlation of deoxyinosine and dUMP abundance with AURKA mRNA expression. P(AURKA-Deoxyinosine) < 2.2e-16, P(CCND3-dUMP) = 3.2e-4. P values were from two-sided Pearson's correlation analysis. ***FDR < 0.001; **0.001 ≤ FDR < 0.01; *0.01 ≤ FDR < 0.05.

https://doi.org/10.1038/s41588-023-01507-7



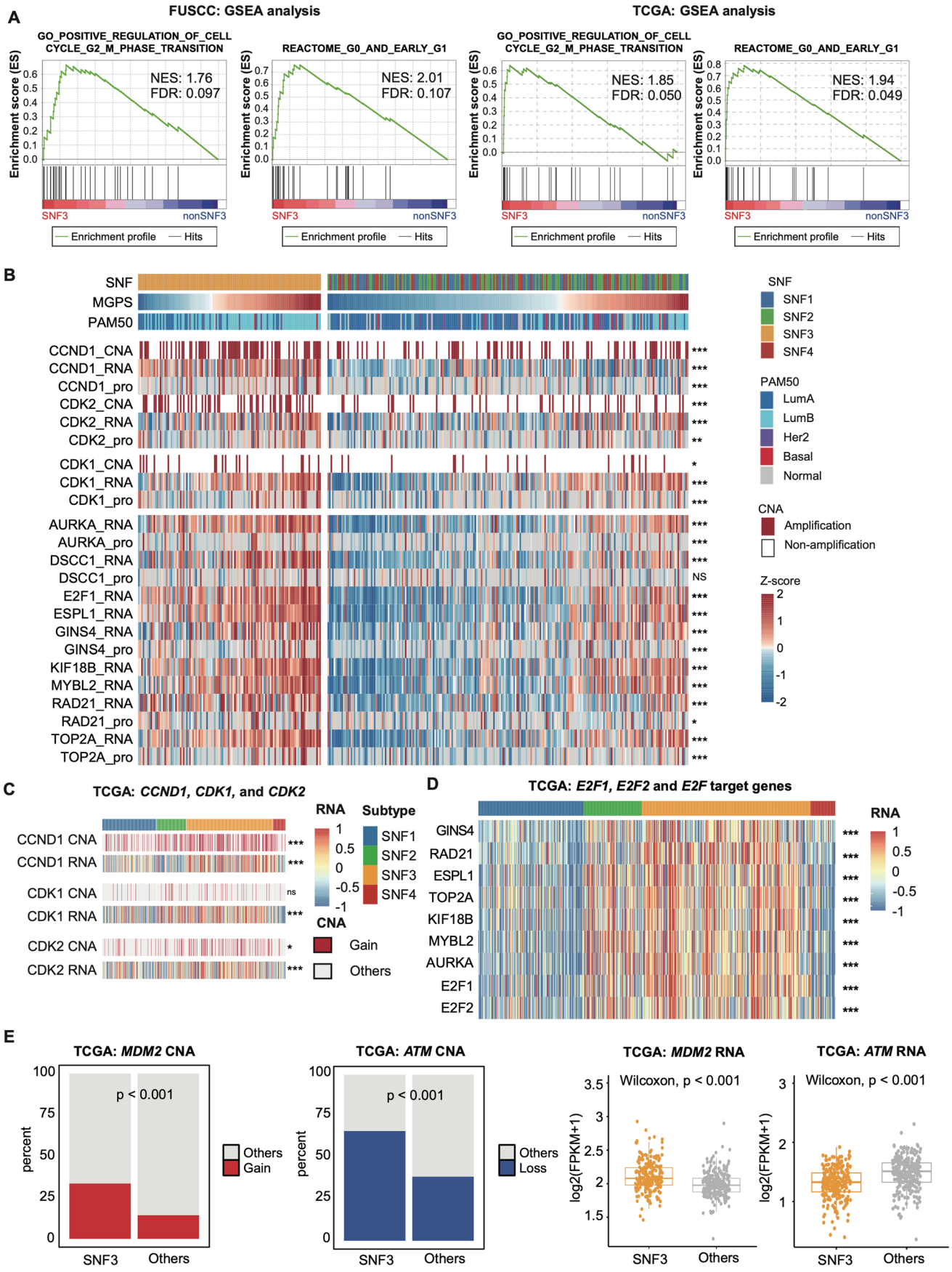
Extended Data Fig. 4 | Extended analysis of clinical status of four subtypes, related to Fig. 3. (**a**) Association of the SNF types with different clinical statuses. *P* values were from the two-sided Fisher's exact test. (**b**) Association of the SNF subtypes with relapse-free survival (RFS). *P* = 9.3e-04. (**c**) Association of the SNF subtypes with metastasis-free survival (MFS) in PAM50 Luminal B patients. (**d**) Forest plot of univariate cox regression analysis for MFS adjusting for tumor size, lymph node status, SNF subtypes, chemotherapy, histological grade. The included patients all received endocrine therapy (n = 296). The hazard ratios (HR) were shown with 95% confidence intervals (CI). Error bar center indicates HR. SNF2vsSNF1: HR = 1.14[0.61–2.11], *P* = 0.687. SNF3vsSNF1: HR = 1.32[0.75–2.32], *P* = 0.330. SNF4vsSNF1: HR = 2.25[1.23–4.12], *P* = 0.008. Lymph node met: HR = 1.08[1.06–1.10], *P* = 1.02e-15. Chemotherapy: HR = 1.41[0.77–2.59], *P* = 0.262. Tumor size: HR = 1.59[1.37–1.85], *P* = 2.17e-09. Grade: HR = 1.31[0.86–1.99], *P* = 0.213. Bold font indicates statistical significance. met: metastasis.

https://doi.org/10.1038/s41588-023-01507-7



Extended Data Fig. 5 | Prediction of SNF subtypes based on the transcriptomics data. (**a**) Workflow for the prediction of SNF subtypes based on the transcriptomics data. (**b**) ROC curves for using the random forest classifier to identify the SNF subtypes. Molecular features of inferred SNF subtypes in (**c**) CPTAC, (**d**) METABRIC and (**e**) TCGA cohort. ***FDR < 0.001; **0.001 ≤ FDR < 0.01; *0.01 ≤ FDR < 0.05; ns: not significant. *P* values were from the two-sided Kruskal-Wallis test.

https://doi.org/10.1038/s41588-023-01507-7

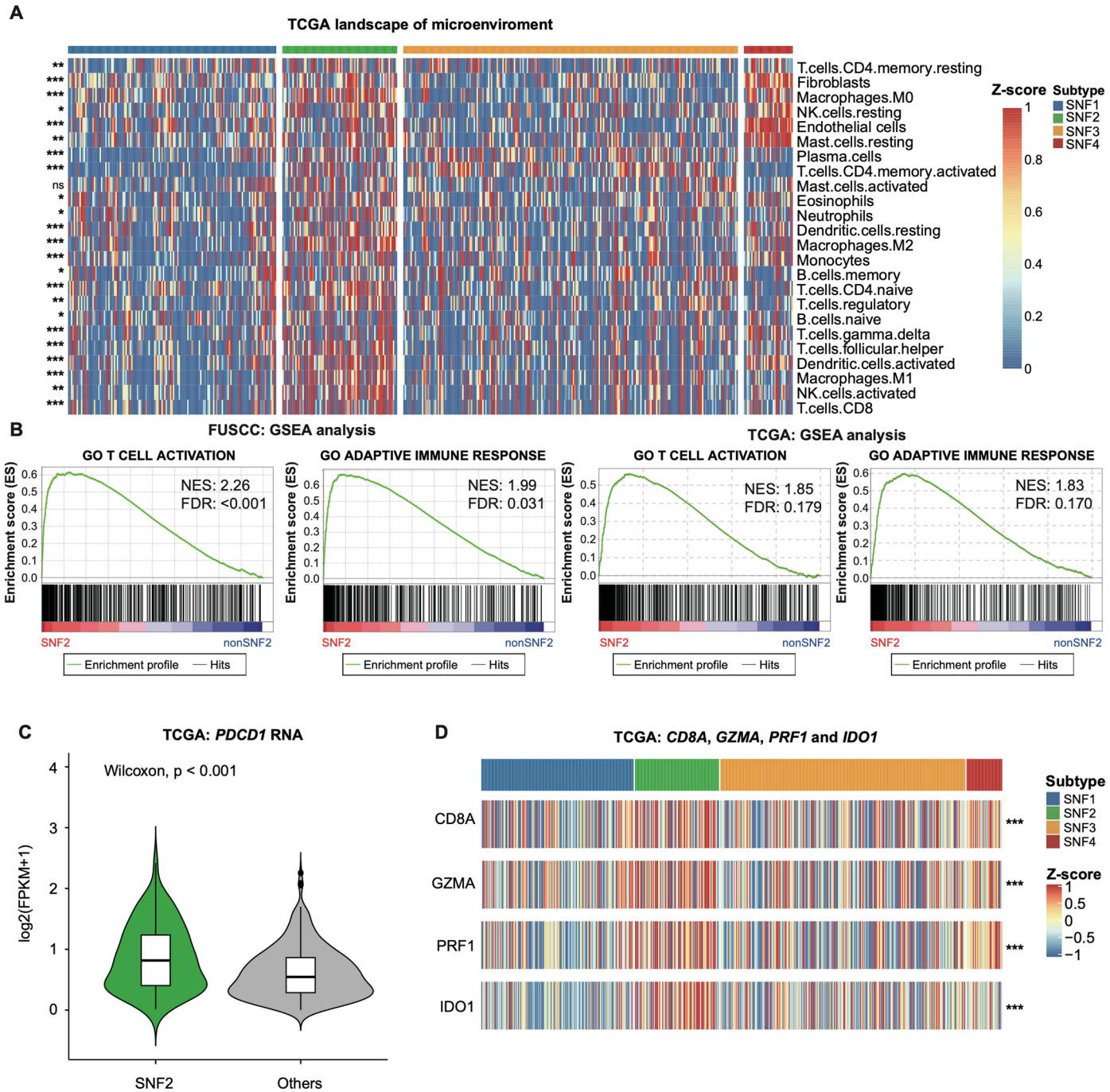


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-023-01507-7

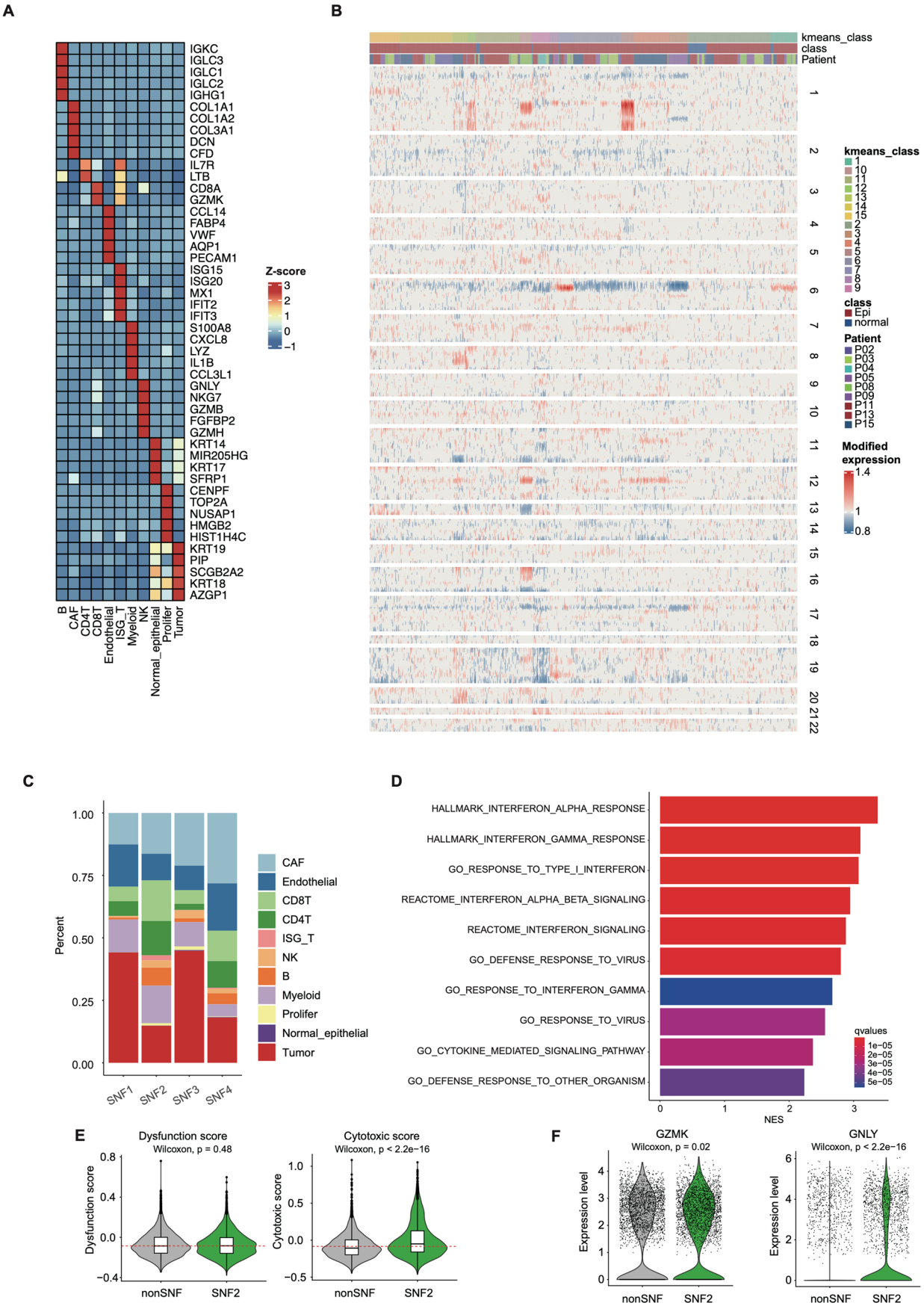
Extended Data Fig. 6 | Extended analysis of SNF3 subtype, related to Fig. 4. (**a**) Representative gene set enrichment analysis plot showing upregulated cell cycle pathway in SNF3 subtype in FUSCC and TCGA cohorts. (**b**) The CNA, mRNA abundance, and protein abundance of *CCND1*, *CDK2*, *CDK1*; the mRNA expression of *E2F1*, *E2F2*, and *E2F* target genes among different subtypes. Copy number amplification was defined as copy number value > log2(4/2). *P* values were from the two-sided ANOVA or Fisher's exact test. MGPS: multi-gene proliferation scores. (**c**) The CNV alteration, and mRNA abundance of *CCND1*, *CDK1*, and *CDK2* among different subtypes in TCGA cohort. Copy number amplification was defined as copy number value > log2(4/2). *P* values were from the two-sided ANOVA or Fisher's exact test. (**d**) Heatmap showing the mRNA expression of *E2F1*, *E2F2*, and *E2F* target genes in TCGA cohort. *P* values were from the two-sided ANOVA test. (**e**) The alteration of two key G2/M cell-cycle regulators (*MDM2* and *ATM*) at the copy number level and mRNA level compared between SNF3 (n = 233) and the other subtypes (n = 259) in TCGA cohort. *P*(*MDM2* CNA) = 7.1e-07, *P*(*ATM* CNA) = 1.2e-09, *P*(*MDM2* RNA) = 2e-11, *P*(*ATM* RNA) = 5.3e-12. *P* values were from the two-sided Wilcoxon or Fisher's exact test. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5xIQR and the data points outside the whisker were outliers. ***FDR < 0.001; **0.001 ≤ FDR < 0.01; *0.01 ≤ FDR < 0.05; NS, FDR ≥ 0.05.

https://doi.org/10.1038/s41588-023-01507-7



Extended Data Fig. 7 | Extended analysis of SNF2 subtype, related to Fig. 5. (**a**) Heatmap showing the estimated abundance of 24 microenvironment cell types among four SNF subtypes in TCGA cohort. *P* values were from the two-sided ANOVA test. (**b**) Representative gene set enrichment analysis plot showing upregulated T cell activation and adaptive immune response in SNF2 subtype in FUSCC and TCGA cohorts. (**c**) Expression of *PDCD1* mRNA expression between SNF2 subtype (n = 80) and other subtypes (n = 412) in TCGA cohort. *P* = 5.6e-05. *P* values were from the two-sided Wilcoxon test. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5xIQR and the data points outside the whisker were outliers. (**d**) The mRNA expression of *CD8A*, *GZMA*, *PRF1*, and *IDO1* between SNF2 subtype and other subtypes in TCGA cohort. *P* values were from the two-sided ANOVA test. ***FDR < 0.001; **0.001 ≤ FDR < 0.01; *0.01 ≤ FDR < 0.05; NS, FDR ≥ 0.05.

https://doi.org/10.1038/s41588-023-01507-7

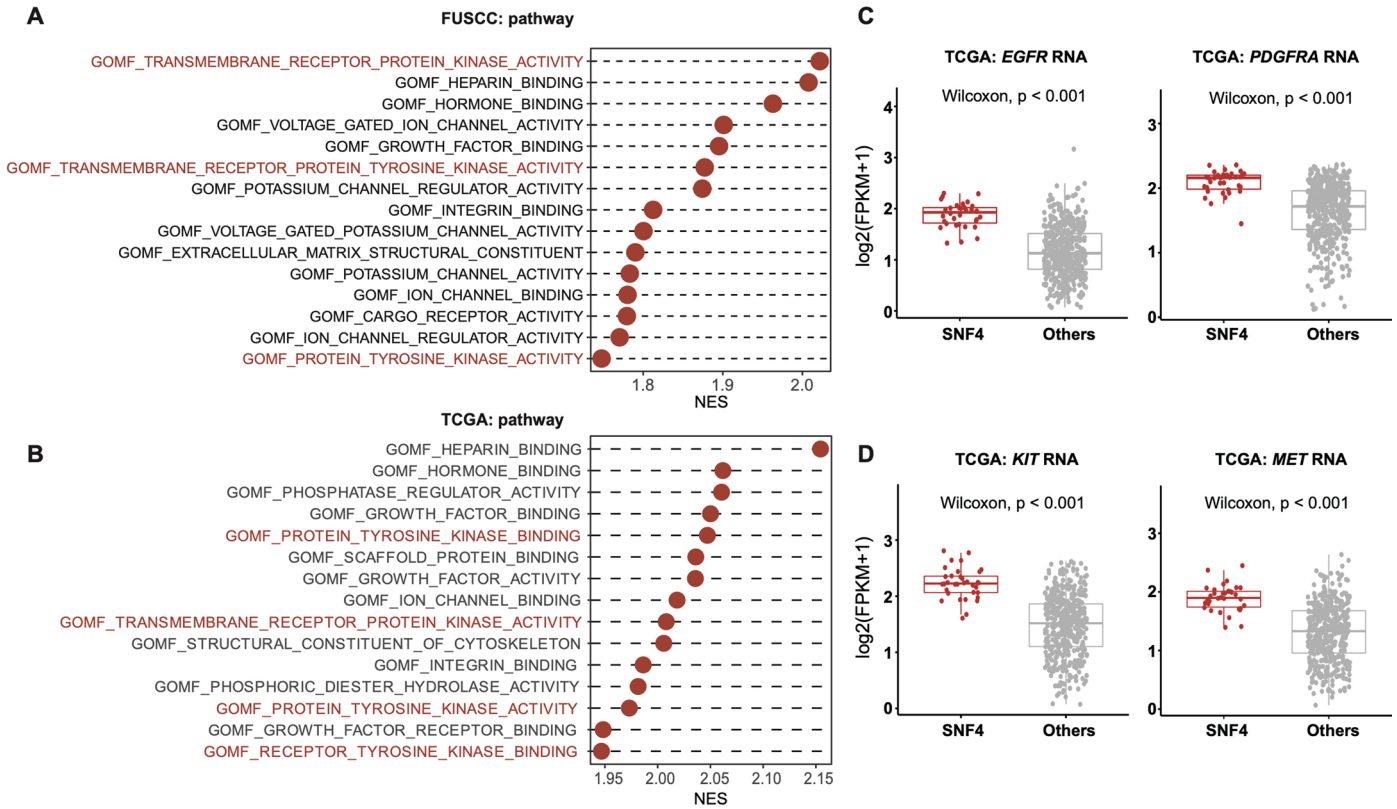


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41588-023-01507-7

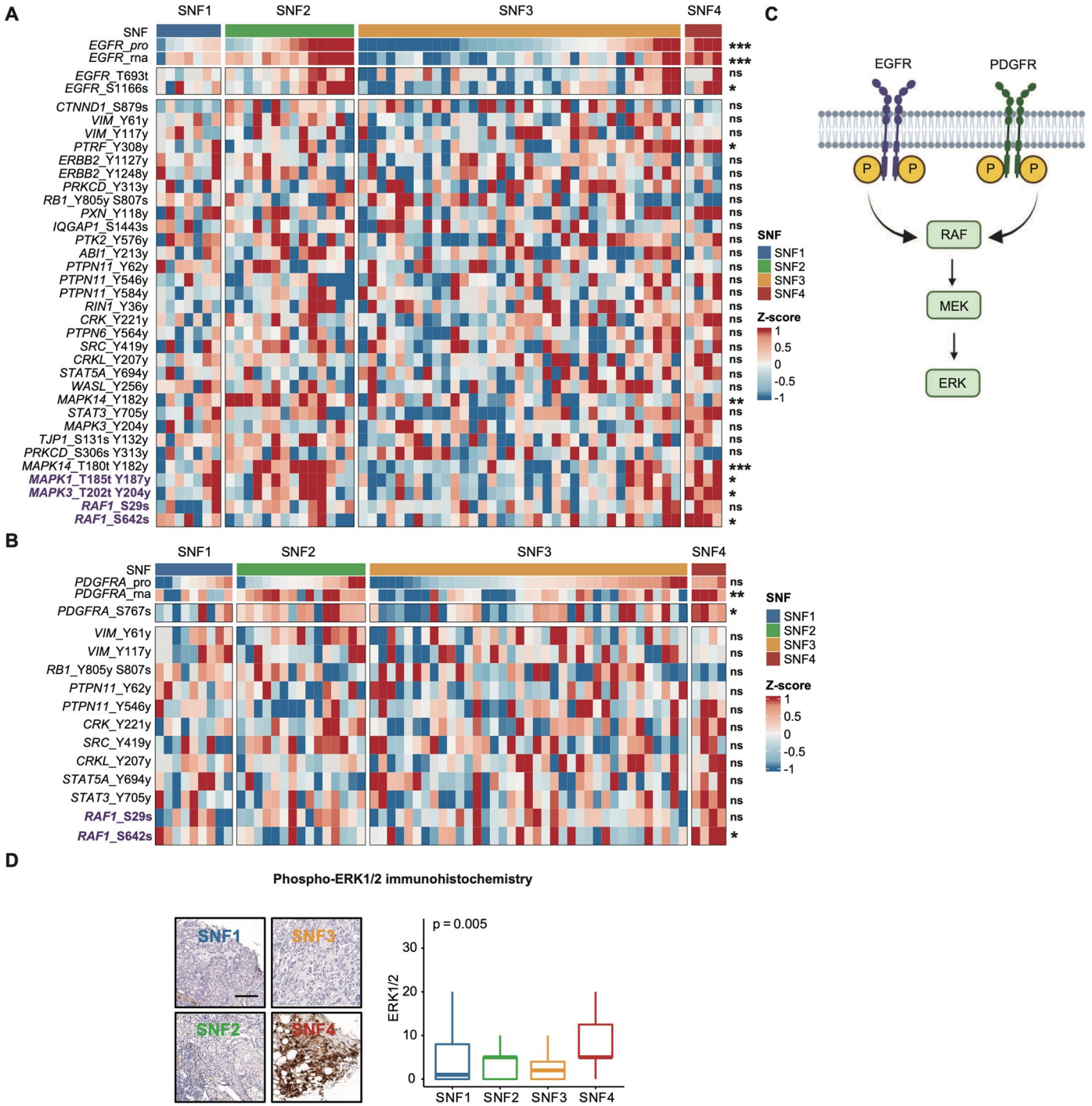
Extended Data Fig. 8 | Cell types detected based on scRNA-seq, related to Fig. 5. (**a**) Heatmap showing the expression of marker genes in the indicated cell types. (**b**) Heatmap showing copy number variations for individual cells (rows) in different genomic segments (column). Sampled immune cells were used as references. (**c**) Distribution of each cell subtype in each SNF subtype. (**d**) GSEA on differentially expressed genes in CD8 + T cell from SNF2 versus non-SNF2 patients for REACTOME, GO and hallmark gene sets. Top 10 pathways enriched in CD8 + T cell from SNF2 samples were shown. (**e**, **f**) Violin plot comparing cytotoxic/dysfunction score (**E**) or cytotoxic-related gene (**F**), *GNLY* and *GZMK*, between CD8 + T cell (n = 6827 cells) from SNF2 (n = 3) versus non-SNF2 (n = 6) patients. *P* values were obtained by two-sided Wilcoxon test. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5xIQR and the data points outside the whisker were outliers.

https://doi.org/10.1038/s41588-023-01507-7



Extended Data Fig. 9 | Extended analysis of SNF4 subtype, related to Fig. 6. (**a**) Cleveland plot showing the top 15 statistically significant pathways (p-value < 0.05, q-value < 0.25) with the highest NES value in FUSCC cohort. Pathways with gene sizes between 50 and 200 have been included. All pathways included were statistically significant. Gene Ontology Molecular Function (GOMF) gene sets were used for GSEA analysis. Receptor tyrosine kinase related pathways were highlighted in red. P-values were calculated by two-sided nonparametric permutation test and adjusted using the Benjamini–Hochberg procedure (q-value). (**b**) Cleveland plot showing the top 15 statistically significant pathways (p-value < 0.05, q-value < 0.25) with the highest NES value in TCGA cohort. Pathways with gene sizes between 50 and 200 have been included. All pathways included were statistically significant. Receptor tyrosine kinase related pathways were highlighted in red. P-values were calculated by two-sided nonparametric permutation test and adjusted by the false discovery rate (q-value). (**c**, **d**) The expression of *EGFR*, *PDGFRA*, *KIT*, and *MET* mRNA level between SNF4 (n = 34) and other subtypes (n = 458) in TCGA cohort. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5xIQR and the data points outside the whisker were outliers. *P*(*EGFR*) = 9.7e-15, *P*(*PDGFRA*) = 3.8e-11, *P*(*KIT*) = 9.4e-15, *P*(*MET*) = 1e-12. *P* values were from the two-sided Wilcoxon test.

https://doi.org/10.1038/s41588-023-01507-7



Extended Data Fig. 10 | Extended analysis of SNF4 subtype, related to Fig. 6. (**A**) Heatmaps showing phosphosite abundance of *EGFR* and their downstream substrates. *P* values were from the two-sided Kruskal-Wallis test without multiple test corrections. (**B**) Heatmaps showing phosphosite abundance of *PDGFRA* and their downstream substrates. *P* values were from the two-sided Kruskal-Wallis test without multiple test corrections. (**C**) Schematic diagram of *PDGFRA/EGFR* and their downstream MAPK signaling pathway. (**D**) Immunohistochemical detection of Phospho-ERK 1/2 and the immunohistochemical staining score quantification among the SNF1 (n = 45), SNF2 (n = 47), SNF3 (n = 55), and SNF4 (n = 27) subtypes. *P* values were from the two-sided Kruskal-Wallis test without multiple test corrections. Scale bar: 100μm. Center line indicates the median, and bounds of box indicate the 25th and 75th percentiles. Whiskers were plotted at 1.5xIQR and the data points outside the whisker were outliers. ***$P < 0.001$; **$0.001 \le P < 0.01$; *$0.01 \le P < 0.05$; NS, $P \ge 0.05$.

nature portfolio

Corresponding author(s): Zhi-Ming Shao, Yi-Zhou Jiang

Last updated by author(s): Jul 20, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For whole exome sequencing and RNA sequencing data, we purified the genomic DNA from fresh frozen samples and peripheral blood cells. For polar metabolomics and lipid metabolomics, we extract metabolites from fresh frozen samples. For digital pathology data, we collected the paraffin-embedded, hematoxylin and eosin-stained tumor slides from the patients . The slides were scanned using the NanoZoomer S210 digital pathology scanner at 40× magnification to generate digital WSIs. For single cell data, pre-treatment biopsy samples from four HR+/HER2- patients were isolated and transported rapidly to the research facility. For further details, please see the Methods section.
Data analysis	The following publicly available tools were used for analyses: GSEA software (v4.0), GSVA (v1.42), ESTIMATE (v1.0.13), ImageScope software (v12.4.0.5043), MATLAB software (v2021a), Pytorch (v1.10.0), MiXCR (v3.0.13), VDJtools (v1.2.1), Chromosome Analysis Suite (ChAS) v4.1, ASCAT v2.4.3, GISTIC2.0 v2.0.22, Sentieon Genomics tools v202010.02, NGSCheckMate (v1.0.0), FastQ Screen (v0.12.0), FastQC (v0.11.8), Qualimap (v2.0.0), VarScan2 v2.4.2, Tnseq (v202010.02), TNscope (v202010.02), StringTie (v1.3.4), Ballgown (v2.14.1), XCMS (v3.2), SNFtool (v2.3.1) , Spectrum (v1.1), Cell Ranger software pipeline (v3.1.0), Seurat (v4.0.5), scDblFinder (v1.6.0), DropletUtils (v1.12.3), Harmony (v0.1.0), SingleR (v1.6.1), inferCNV (v1.8.1), DESeq2 (v1.34), R statistical packages v3.6.1. Details and references can be found within text in the relevant Methods section and Supplementary Note. The CNN models and all relevant codes from this manuscript are available at Github (https://github.com/yifanzhou330/SNF) and Zenodo (https://doi.org/10.5281/zenodo.8022438).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The WES data, CNA data, RNA sequencing data and metabolome data for this study have been deposited into GSA database (https://ngdc.cncb.ac.cn/gsa/) under accession codes PRJCA017539 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017539). TMT-based MS-quantified protein data have been submitted into iProX (https://www.iprox.cn) under accession codes IPX0006535000 . Human Primary Cell Atlas data is obtained from celldex package (v1.11) (https://github.com/LTLA/celldex). The TCGA, METABRIC and CPTAC data were downloaded from the cBioPortal website (https://www.cbioportal.org/). Source data are provided with this paper.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	As breast cancer is extremely rare in males. Only one sample is male and others are all females (n=578)
Population characteristics	Baseline population characteristics of patients with HR+/HER2- breast cancer are detailed in Supplementary Tables 1. The median length of follow-up was 83.0 months
Recruitment	In this retrospective cohort, eligible cases were patients with pathologically confirmed HR+/HER2- breast cancer, confirmed by pathologist with samples available and all clinical data above available. Our cohort included pretreatment patients treated at Fudan University Shanghai Cancer Center between January 2013 and December 2014 without intentional selection (N = 478). We also selected patients who experienced relapse after surgery between January 2009 and December 2016 (N = 101) to profile the patients with relatively high risk. No other sources of significant selection bias were identified.
Ethics oversight	All tissue samples included in the present study were obtained after approval of the research by the FUSCC Ethics Committee, and each patient provided written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample selection was performed retrospectively and determined based on availability of tissue and clinical data. Sample sizes were sufficient for clustering analyses used in this report.
Data exclusions	No data were excluded from the analyses
Replication	Experiments were performed using at least n=3 biological replicates.
Randomization	Because all pretreatment patients with HR+HER2- breast cancer treated at Fudan University Shanghai Cancer Center between January 2013 and December 2014 and patients who experienced relapse after surgery between January 2009 and December 2016 were included in an unbiased fashion, acquisition of primary patient tumor samples was not randomized.
Blinding	Investigators were not blinded to clinical information prior to collection of data, as this was required for the selection of samples with pathologically confirmed HR+/HER2- breast cancer. For data analysis, investigators were not blinded to this clinical information, as this was required to design the analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

For western blot, EGF Receptor (1:1000, Cell Signalling Technology, #4267), Phospho-EGF Receptor (Tyr1068) (1:1000, Cell Signalling Technology, #3777), PDGF Receptor α (1:1000, Cell Signalling Technology, #3174), Phospho-PDGF Receptor α (Tyr1018) (1:1000, Cell Signalling Technology, #4547), αSMA (1:1000, Abcam, #ab124964), β-actin (1:20000, Proteintech, #66009-1-Ig) were used. Anti-rabbit IgG, HRP-linked Antibody (Cell Signaling, #7074) were used as secondary antibody for EGF Receptor, Phospho-EGF Receptor (Tyr1068), PDGF Receptor α, Phospho-PDGF Receptor α (Tyr1018) and αSMA. Anti-Mouse IgG, HRP-linked Antibody (Cell Signaling, #7076) was used as secondary antibody for β-actin.
For immunohistochemistry, phospho-p44/42 MAPK (Erk1/2) (1:200, Cell Signaling Technology, #4370), phosphor-RB1 (Ser807/811) (D20B12) (1:200, Cell Signaling Technology, #8516), CD20 (1:100, Abcam, #ab78237), CD8A (1:1500, Servicebio, #GB12068), αSMA (1:200, Servicebio, #GB13044), and CD206 (1:400, Servicebio, #GB113497) were used. Detection System/Mo&Rb (Gene Tech, #GK6007) was used as secondary antibody.

Validation

#4267 - https://www.cellsignal.com/products/primary-antibodies/egf-receptor-d38b1-xp-rabbit-mab/4267
#3777 - https://www.cellsignal.com/products/primary-antibodies/phospho-egf-receptor-tyr1068-d7a5-xp-rabbit-mab/3777
#3174 - https://www.cellsignal.com/products/primary-antibodies/pdgf-receptor-a-d1e1e-xp-rabbit-mab/3174
#4547 - https://www.cellsignal.com/products/primary-antibodies/phospho-pdgf-receptor-a-tyr1018-antibody/4547
#ab124964 - https://www.abcam.com/products/primary-antibodies/alpha-smooth-muscle-actin-antibody-epr5368-ab124964.html
#66009-1-Ig - https://www.ptgcn.com/products/Pan-Actin-Antibody-66009-1-Ig.htm
#7074 - https://www.cellsignal.com/products/secondary-antibodies/anti-rabbit-igg-hrp-linked-antibody/7074
#7076 - https://www.cellsignal.com/products/secondary-antibodies/anti-mouse-igg-hrp-linked-antibody/7076
#4370 - https://www.cellsignal.com/products/primary-antibodies/phospho-p44-42-mapk-erk1-2-thr202-tyr204-d13-14-4e-xp-rabbit-mab/4370?site-search-type=Products&N=4294956287&Ntt=phospho-p44%2F42+mapk+%28erk1%2F2%29&fromPage=plp
#8516 - https://www.cellsignal.com/products/primary-antibodies/phospho-rb-ser807-811-d20b12-xp-rabbit-mab/8516
#ab78237 - https://www.abcam.com/cd20-antibody-ep459y-ab78237.html
#GB12068 - https://www.servicebio.com/goodsdetail?id=22235
#GB113497 - https://www.servicebio.com/goodsdetail?id=21621
#GB13044 - https://www.servicebio.com/goodsdetail?id=15638
#GK6007 - https://www.genetech.com.cn/goods/goods_detail/757861.html