scientific reports

Check for updates

OPEN

Normal tissue transcriptional signatures for tumor-type-agnostic phenotype prediction

Corey Weistuch[1], Kevin A. Murgas[2], Jiening Zhu[3], Larry Norton[4], Ken A. Dill[5], Allen R. Tannenbaum[3,6] & Joseph O. Deasy[1]✉

Cancer transcriptional patterns reflect both unique features and shared hallmarks across diverse cancer types, but whether differences in these patterns are sufficient to characterize the full breadth of tumor phenotype heterogeneity remains an open question. We hypothesized that these shared transcriptomic signatures reflect repurposed versions of functional tasks performed by normal tissues. Starting with normal tissue transcriptomic profiles, we use non-negative matrix factorization to derive six distinct transcriptomic phenotypes, called archetypes, which combine to describe both normal tissue patterns and variations across a broad spectrum of malignancies. We show that differential enrichment of these signatures correlates with key tumor characteristics, including overall patient survival and drug sensitivity, independent of clinically actionable DNA alterations. Additionally, we show that in HR+/HER2- breast cancers, metastatic tumors adopt transcriptomic signatures consistent with the invaded tissue. Broadly, our findings suggest that cancer often arrogates normal tissue transcriptomic characteristics as a component of both malignant progression and drug response. This quantitative framework provides a strategy for connecting the diversity of cancer phenotypes and could potentially help manage individual patients.

Keywords Molecular profiling, Metastatic breast cancer, Drug sensitivity prediction, Cancer ecology and evolution, Prognosis

Recent studies have identified conserved phenotype states that predict drug sensitivities and other critical cancer traits across various tumor types[1–6]. These advances have been driven by the adoption of novel mathematical methods in biological research, particularly for characterizing the geometry and clustering of high-dimensional data[4,5,7–10]. Distances from these clusters or "states" can then linked to specific drug sensitivities, metastatic potentials, and other key cancer traits[2,10–13]. Such approaches have shown promise in guiding therapy selection and personalized treatment planning across tumor types, particularly for rare and therapy-resistant malignancies[10].

The immense phenotype diversity of cancer phenotypes poses a significant challenge, with no universally accepted standard for defining and measuring shared tumor states. Developing a standardized atlas of system-level, multi-gene properties could be crucial for monitoring, predicting, and treating these conserved tumor states, thus enabling a more systematic approach to understanding cancer behaviors[14–16]. However, it remains unclear where these states come from and how broadly they can be applied. Therefore, two key needs must be addressed: 1) the identification of signatures with clear biological origins; and 2) a comprehensive study that evaluates how conserved tumor states correlate with drug sensitivities, metastatic patterns, and patient outcomes.

Unlike tumors, normal tissues typically have well-separated gene expression profiles and defined functional roles. Furthermore, many of the transcriptional differences observed across cancers mimic patterns of variability observed in normal tissues, making them an ideal system for contextualizing observed patterns of intertumor and intratumor heterogeneity[16–18]. Leveraging these links could therefore unveil new and general connections between the well-defined functional roles of normal tissues and recurring patterns of tumor heterogeneity.

Here, we use a statistical decomposition method to isolate and interpret transcriptomic motifs utilized both by normal tissues and tumors, referring to these signatures as *normal tissue archetypes*[19]. By restricting our analysis

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA. [2] Department of Biomedical Informatics, Stony Brook University, Stony Brook, USA. [3]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, USA. [4]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, USA. [5]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, USA. [6]Department of Computer Science, Stony Brook University, Stony Brook, USA. ✉email: DeasyJ@mskcc.org

to biological pathways utilized by all tissues, such as glycolysis and DNA repair, the derived archetypes are mapped to unique tasks performed by normal tissues as well as distinct cancer hallmarks[2,9,15]. This employment of a normal tissue reference has the potential to contextualize the origins of conserved transcriptional states across multiple cancer types[20,21]. Specifically, the approach calculates the mix of archetypes within a given cancer cell or bulk tumor in terms of fixed normal tissue archetype signatures, providing a new perspective and a flexible quantitative approach to classify the different possible types of tumor behavior, match them to their most effective treatments, and interpret how they evolve in response to various therapies and through the accumulation of genetic alterations.

In assessing the prognostic potential of normal tissue archetypes identified through our method, we employed a structured case-study approach, utilizing publicly available data from two pan-cancer datasets[22–25] and three datasets focusing on therapy-induced and metastatic adaptation in specific malignancies[26–28]. Our primary aim was to gauge the predictive capability of inferred archetype mixtures in determining the responsiveness of cancer cell lines to distinct chemotherapies, agnostic to mutational status, copy number alterations, and cancer type. Given the recognized divergence in behavior between cancer cell lines and clinical samples, we also scrutinized their ability to predict overall patient survival, differentiate between prognostic cancer subtypes, and anticipate adaptive responses from bulk tumor transcriptomes.

In a broader context, this proof-of-concept study highlights the prognostic significance derived from understanding why tumors commonly adopt normal tissue transcriptional programs distinct from their lineages of origin. The ability to identify and link these programs to drug sensitivities, site-specific metastases, and DNA alterations may pave the way for more tumor-type agnostic personalized cancer therapies.

Methods

Ethical compliance

All analyses used publicly available, de-identified data.

Public datasets

GTEx

We utilized the normal tissue samples from the Genotype-Tissue Expression (GTEx) project to cover the breadth of gene expression space and to provide an enhanced signal for resolving transcriptomic archetypes of both normal and cancerous tissues[29]. GTEx (version 8) gene-level transcripts per million (TPM)-normalized expression data were downloaded from the GTEx Portal[29]. The dataset, consisting of 54 distinct tissues, is one of the most comprehensive resources for studying tissue-specific gene expression. GTEx was established to characterize the tissue-specific determinants of human traits and diseases and provides expression levels for about 44 thousand genes. We focus here on commonly enriched pathways in cancer by utilizing 780 genes from five key cancer-related pathways from the Molecular Signature Database (MSigDB): apoptosis, DNA repair, glycolysis, hypoxia, and oxidative phosphorylation[30,31]. To go beyond current classifications, we filtered out lineage-specific genes already used for cancer classifications, enabling comparisons across cancer types[32]. The above five gene sets ("pathways") were chosen as a precaution against overfitting the 54 median-averaged bulk transcriptomes available in GTEx and because they play particularly frequent roles in cancer[14]). This ensures that our analysis is of relevance to multiple types of cancer. The remaining hallmark pathways, not encompassed by the previously mentioned patterns, included either specific signaling pathways or covered functions that are not ubiquitously expressed across tissues. Finally, to normalize the dataset, we divided the expression of each gene by its standard deviation across all tissues.

Cancer cell line encyclopedia, CCLE

Gene-level TPM-normalized expression from the CCLE (N = 1405) along with matched drug sensitivities ($N = 469$), mutations ($N = 1250$), and copy number alterations ($N = 1387$ were downloaded from the public Dependency Map (DepMap) portal (version 22Q2)[22]. To reduce false positives, only TCGA (The Cancer Genome Atlas) hotspot mutations present in at least five TCGA samples (as reported in CCLE) and ten CCLE samples were retained. Mutation types were not further stratified. Due to the lack of a similar reference for pan-cancer copy number analysis, all whole genome-level copy number alterations reported in CCLE, aside from those on the sex chromosomes, were used. The upper bound for gene-level copy number alterations was set to four in order to remove potential correlational biases from high copy number states. Spearman rank correlations were then computed between the copy number state of each gene and each computed normal archetype score. Finally, region-level correlations were computed by resegmenting the genome into 20000 bins with equal mappability, averaging gene-level correlations within each bin.

Gastrointestinal stromal tumor (GIST) therapy responses

Gene-level TPM-normalized expression data from imatinib-sensitive ($N = 5$) and imatinib-resistant ($N = 5$) GIST patients were downloaded from NCBI GEO accession code GSE155800[26].

Longitudinal study of metastatic breast cancer

Gene-level TPM-normalized expression data from east Asian HR+/HER2- metastatic breast cancer patients before and after treatment with palbociclib plus endocrine therapy ($N = 23$ matched pairs) were downloaded from NCBI GEO accession code GSE186901[27].

The Cancer Genome Atlas, TCGA
Gene-level TPM-normalized expression data from the TCGA within the breast (BRCA, $N = 1111$), colon (COAD, $N = 481$), and pancreatic (PAAD, $N = 178$) cancer cohorts were downloaded via the TCGAbiolinks R package[23–25,33]. Samples were restricted to primary tumors.

Study of site-specific adaptation of metastatic breast cancer
Raw sequencing data from the basal (triple negative) metastatic breast cancer patients used in this study ($N = 7$; $N = 5$ caucasian, $N = 2$ african) are available for controlled access from dbGaP (database of Genotypes and Phenotypes) accession code phs000676.v2.01[28]. Gene-level quantile and TPM-normalized expression data were downloaded from GitHub https://github.com/FaribaRoshanzamir/Metastatic-TNBC/tree/main/data and were additionally processed as documented in[21]. In summary, primary tumors from each patient were used alongside their matched distant metastases to brain ($N = 7$), lung ($N = 6$), liver ($N = 5$), lymph node ($N = 2$), adrenal gland ($N = 2$), and skin ($N = 2$). The samples were then batch-corrected against normal tissue samples from GTEx and primary tumor samples from TCGA corresponding to each metastasis destination. Median expression levels of each metastasis site, normal tissue, and primary tumor were then provided. Due to the lack a of matched normal tissue in GTEx, the sample derived from lymph node metastasis was excluded from our analysis.

N-NMF archetype analysis

Biological data often embody the amalgamation of interconnected constituent parts. Our goal is to isolate the parts mathematically, revealing key structures and hidden patterns. This is commonly accomplished through approximate low-rank matrix and tensor factorizations, including principal component analysis (PCA) and non-negative matrix factorization (NMF). Notably, NMF has played a significant role in recent biological studies, contributing to dimension reduction, discrimination, and clustering[3,4,34]. When the factor weights are normalized (N-NMF), they encode the relative contribution of each part in contributing to the whole[35]. Here we detail our implementation of this factor or "archetype" discovery procedure.

Choosing the optimal number of archetypes
Application of NMF requires predefining the number of archetypes. The optimal number of archetypes was chosen using the profile log-likelihood method[36]. This approach models the unknown number of archetypes as a latent variable that can be directly optimized over. This method was selected due to its simplicity and superior performance compared to alternatives[37]. When training the archetype model on normal tissue transcriptomic data (GTEx), the minimum number of factors was set to $k = 2$ (due to the normalization constraint), with the maximum set at $k = 30$. The optimal number of archetypes, determined by maximizing the resulting profile log-likelihood, was determined to be $k = 6$ (cf. SI Fig. 1A).

Archetype discovery and sample projection
Normalized nonnegative matrix factorization (N-NMF) was employed to establish a low-dimensional representation of gene expression profile variability across normal tissues (GTEx). This general approach seeks to approximate an $N \times M$ data matrix V, representing N gene expression values of M samples, as the product of two low-rank matrices W and H: $V \approx WH$, where W is an $N \times k$ matrix representing coefficients of each gene's contribution to the $k = 6$ archetypes, and H is a $k \times M$ matrix representing the weights of each archetype needed to approximate each gene expression sample (cf. Fig. 1A). In our implementation, these matrices are found by minimizing the Frobenius norm distance (a matrix version of the Euclidean distance) between the data matrix V and its approximation[35]. The optimal solution is obtained through 1000 iterations of the classical Seung-Lee algorithm[38] (see SI Fig. 1B for algorithm convergence curves). Prior to this, the W and H matrices are randomly initialized using a uniform distribution. To avoid degeneracy in the model fitting and ensure that the coefficients of each archetype score sum to 1, the H matrix is column-normalized after each iteration[35]. Subsequently, when calculating archetype scores for new samples, we follow the same procedure but fix the W matrix to its learned values.

Projection method for archetype visualization
To visualize relationships among archetypes, we projected the six-dimensional archetype space onto a regular hexagon. This is done by associating each archetype with one of the six vertices of the hexagon and adding the vectors pointing from the center of the hexagon outward to each vertex, weighted by the archetype scores of each sample. This method ensures that samples near a vertex are those dominated by that archetype, while those along the edges show trade-offs between adjacent archetypes. Furthermore, this method ensures an unbiased distortion of the archetype structure.

Statistical analysis

Statistical analyses were performed using MATLAB 2020b. Associations between archetype scores and all other variables were determined using paired or unpaired Spearman rank correlation t-tests unless otherwise indicated. The Benjamini-Hochberg procedure for multiple comparisons correction was applied when appropriate and as indicated[39]. Associations with $p < 0.05$ were considered significant unless otherwise indicated. Kaplan-Meier survival analyses were performed using MatSurv[40].

Data and code availability

Figure 5C was generated using R version 4.3.3[41]. All other figures were generated using MATLAB version 9.3.3 (2020b)[42]. The source data and MATLAB code that support the findings of this study are available in GitHub with the identifier https://github.com/Corey651/Cancer_Archetypes .

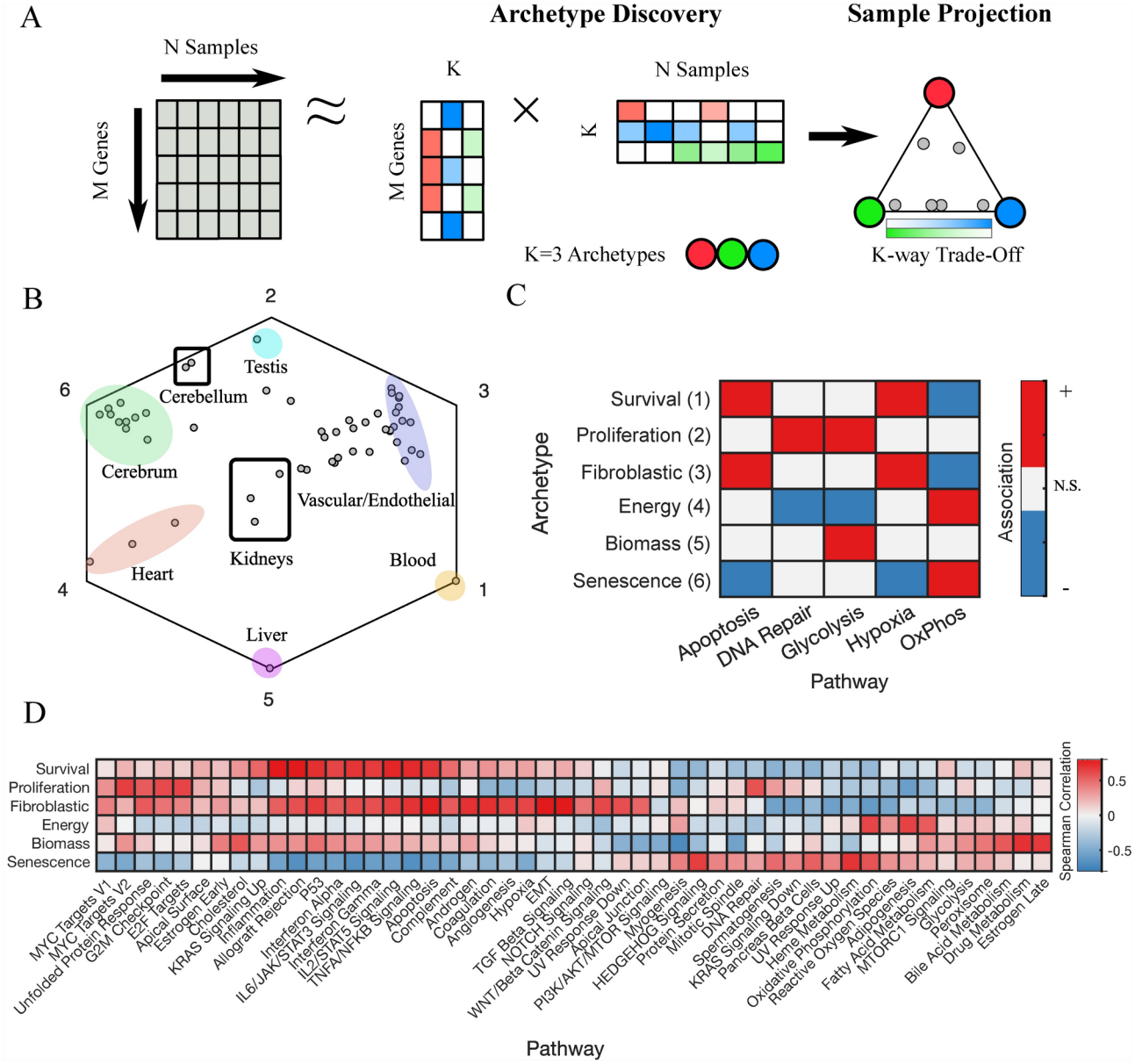

Fig. 1. Normal tissue archetypes as a paradigm to interpret cellular behaviors. (**A**). Flowchart of the analysis pipeline (see Methods: N-NMF archetype analysis). (**B**). Projection of median-averaged normal tissue transcriptomes onto the six archetypes discovered in GTEx ($N = 54$ distinct tissues). The coordinates of each tissue correspond to their degree of similarity to each archetype (the vertices, cf. Methods). Tissue groups enriched for a single archetype are marked by ovals, whereas those achieving a balance of multiple archetypes are marked by squares. (**C**). Heatmap of the significant associations between individual archetypes and the six pathways on which they were trained (Spearman rank permutation test, Benjamini-Hochberg $p < 0.05$). (**D**). Heatmap of the Spearman correlations between individual archetypes and the average expression levels of each MSigDB Hallmark gene group.

Results

Normal tissue diversity signatures mimic oncogenic transcriptional programs

Normalized Nonnegative Matrix Factorization or N-NMF (cf. Methods and Fig. 1A,[35]) was used to identify the most representative archetypes of the normal tissue transcriptomes in the Gene-Tissue Expression Project (GTEx, cf. Methods). The method determined that the data was best represented by six archetypes (cf. SI Fig. 1, Methods), each associated with distinct normal tissue types (cf. Fig. 1B) and biological pathways (cf. Fig. 1C,D). Importantly, each normal tissue was characterized by a mixture of archetypes, although one is typically dominant. Related tissues, such as the constituents of the cerebrum, were seen to have more similar archetype scores, in accordance with expectations. Crucially, the presence of single-tissue archetypes (e.g., testis and liver) and distinct clustering into interpretable tissue classes suggest that our analysis avoids bias towards overrepresented tissue classes.

Although the initial analysis focused on select ubiquitously utilized pathways, Figure 1D illustrates notable Spearman rank correlations among numerous additional Hallmark gene pathways sourced from MSigDB and the individual archetype scores[30,31]. Archetype 1 exhibited enrichment for immune pathways, including interferons, TNFA/NFKB signaling, and apoptosis, suggestive of an anti-apoptotic, immune-evasive phenotype termed "Survival"[43]. Archetype 2 demonstrated enrichment for cell division pathways such as the G2/M checkpoint, MYC targets, and DNA repair, associated with "Proliferation"[44]. Archetype 3, sharing some immune-related pathways with Archetype 1, also displayed enrichment for vascularization (angiogenesis), cell adhesion (apical junction), and cell signaling pathways (NOTCH, TGFB), collectively associated with "Fibroblastic" activity[45]. Archetype 4 was enriched for catabolic or "Energy" metabolism pathways. Archetype 5 was enriched for hormonal (estrogen, cholesterol), drug metabolism, and glycolysis pathways, collectively reflecting known liver functions summarized as "Biomass". Archetype 6 showed enrichment for heme metabolism, oxidative phosphorylation, and HEDGEHOG signaling reflective of a differentiated phenotype. Additionally, archetype 6 exhibited broad negative associations with immune-related pathways, contrasting with archetype 1, reminiscent of "Senescent" cells[46].

Of note, the scores recapitulate known functional trade-offs across normal tissues, such as the elevated oxidative requirements of the heart and cerebrum ("Energy" and "Senescence"), the requirement for glycogen production in the liver ("Biomass"), and the protection, regulation, and manipulation of genetic material in the testis ("Proliferation"). Notably, certain tissues encompass multiple archetypes, thereby striking a balance between various functions. For instance, the cerebellum, akin to the testis[47,48], demonstrates resilience to aging and accumulated DNA damage, while also exhibiting oxygen-demanding characteristics similar to the cerebrum[49]. The kidneys, expressing approximately 70% of the genes in the human body[50], assume intermediate archetype values, reflecting their versatile functional profile.

Crucially, the normal tissue archetype-association enrichment patterns also resemble groups of pathways commonly co-expressed in many different types of cancer[2,9,14]. Specifically, several of the archetypes closely correspond to those previously identified in tumors: Survival → Immune interaction, Proliferation → Cell division, and Fibroblastic → Invasion/tissue remodeling[9]. Energy and Biomass, on the other hand, were somewhat similar to combinations of these existing signatures, whereas Senescence bore no such resemblance. Of note, several chemotherapies induce a treatment-resistant, senescence-like states in tumors, characterized by immunogenic and metabolically-active features related to those of the Senescence archetype[46].

Non-canonical tissue signatures are enriched in cancer cell lines

To determine how well normal tissue archetypes capture shared expression patterns across cancers compared to their respective lineages of origin, we analyzed their distribution across the Cancer Cell Line Encyclopedia or CCLE (N = 1405, Refs.[22,51], cf. Methods). Consistent with typical tumor expression patterns, the cancer cell lines were predominantly enriched for the Proliferation and Senescence archetypes[46] with minor variation across lineages (cf. Fig. 2A). However, while certain cancer types exhibited preferences for specific archetypes, these preferences (with the exception of the liver) did not mirror those of their respective normal tissue lineages (cf. Fig. 2B,C).

Normal tissue signature enrichment influences pan-cancer drug responses

We next investigated the ability of normal tissue archetypes to predict tumor drug sensitivities across a variety of cancers. As illustrated in Fig. 3, transcriptomic signatures derived from normal tissues correlated with drug

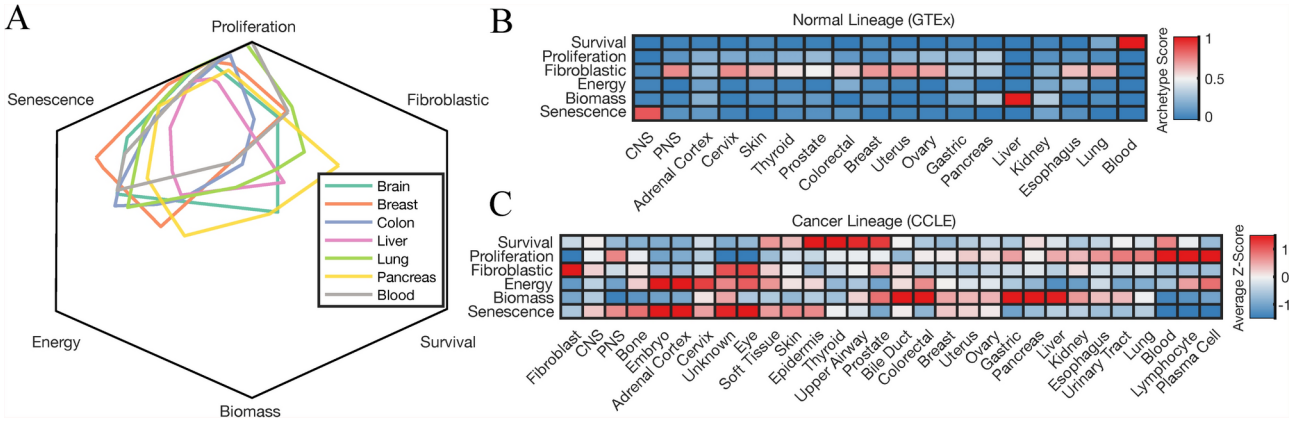

Fig. 2. Variation in normal archetype enrichment across cancer cell line lineages. (**A**). Ranges of archetype values in CCLE, stratified by lineage. Archetype ranges were calculated using convex hull estimation. (**B**). Heatmap of the archetype patterns from normal tissues (GTEx) with corresponding cancer cell line lineages, displaying the first matched tissue type for each lineage. (**C**). Heatmap of the average archetype Z-scores across tumor lineages (CCLE), computed by averaging expression within each lineage and calculating Z-scores relative to the overall data. Z-scores were used to correct for the average archetype trend observed in cancer cell lines (**A**).

sensitivities in both cancer cell lines and patients. Specifically, the drug sensitivities of cancer cell lines – measured by the area under the dose-response curve, referred to as activity area (as defined in[22]) – showed a significant association with at least one archetype for 21 out of 24 anti-cancer drugs profiled by the CCLE, following pooling across cancer types and multiple comparisons correction (cf. Fig. 3A). These associations clustered by drug and aligned with archetype-specific pathway dependencies (cf. Fig. 1D). For example, the Survival archetype, linked to apoptosis, was associated with sensitivity to the anti-apoptosis inhibitor LBW242. The Proliferation archetype, enriched for genes involved in the G2/M checkpoint, correlated with sensitivity to topoisomerase inhibitors (irinotecan, topotecan) and other drugs targeting dividing cells. Finally, the Biomass archetype, linked to estrogen synthesis, was associated with sensitivity to *EGFR* inhibitors (lapatinib and erlotinib).

Proliferation archetype scores were significantly lower in gastrointestinal stromal tumor (GIST) patients resistant to imatinib (cf. Fig. 3B). Furthermore, treatment with palbociclib, known to elicit therapy-induced tumor senescence as a resistance mechanism in various cancer types[52], was associated with significant increases in Senescence archetype scores in HR+/HER2- metastatic breast cancer samples compared to their pre-treatment state (cf. Fig. 3C).

Next, we examined specific drug-archetype correlations for LBW242 (Survival), irinotecan (Proliferation), and lapatinib (Biomass) across cancer lineages (cf. Fig. 3D–F). Although the Survival correlated with LBW242 sensitivity in our global analysis, it was infrequently expressed in the analyzed cancer cell lines (cf. Fig. 3D). In contrast, the Proliferation and Biomass archetypes were significantly associated with increased sensitivity to irinotecan and lapatinib within individual cancer lineages, respectively (cf. Fig. 3E,F). Lineage-specific R^2 values for the irinotecan linear regressions were 0.2199, 0.2152, and 0.1689 for breast, blood, and lung lineages, respectively. Accordingly, lineage-specific R^2 values for the lapatinib linear regressions were 0.2558, 0.0465, and 0.0794 for breast, colorectal, and lung, respectively. Notably, a nonlinear threshold effect was observed for lapatinib, where sensitivity was seen only in colorectal and breast cancer cell lines with Biomass scores above 0.15 (cf. Fig. 3F).

Finally, when averaging activity areas and archetype values across cell lines of the same lineage, we observed associations for LBW242 with the Survival archetype and irinotecan with the Proliferation archetype (cf. Fig. 3G,H). However, no clear trend was observed between lapatinib sensitivity and the Biomass archetype, suggesting that either the averaging procedure obscured the association or tumor-type-specific factors are required to compare lapatinib sensitivities across different malignancies (cf. Fig. 3I).

Normal tissue signature enrichment improves survival prediction and distinguishes site-specific metastases in breast cancer

Recognizing the promising capacity of normal tissue archetypes in predicting and monitoring treatment responses, we sought to assess their broader applicability in interpreting existing treatment guidelines and overall clinical outcomes. This encompassed three specific tasks. Firstly, conducting a comparative analysis between established treatment guidelines and those proposed by our archetypes, thereby illustrating the translational implications of our earlier discoveries. Secondly, evaluating the potential of these signatures for identifying vulnerable patients more effectively than existing stratification schemes. Lastly, characterizing archetype enrichment changes and potential archetype-predicted vulnerabilities associated with metastasis. To accomplish these tasks, we chose breast cancer as a pertinent case study due to its well-established transcriptional subtypes[53], diversity of potential metastatic sites[21,28], and ample available data for study[23,28].

To characterize subtype-specific archetype enrichment patterns, we used RNA-Seq data from breast cancer cell lines (CCLE, $N = 63$) and primary patient samples from The Cancer Genome Atlas (TCGA, $N = 1111$,[23]). Both datasets encompassed standard breast cancer subtypes, with CCLE distinguishing two basal subtypes (A/B) and TCGA distinguishing two luminal subtypes (A/B) along with an additional normal subtype (cf. Fig. 4A,B). Perhaps owing to differences in purity between cell lines and clinical samples, absolute archetype scales differed between CCLE and TCGA. Furthermore, the correlation between breast cancer subtype and individual archetypes was limited, as evidenced by their substantial subtype-intrinsic variance, especially in TCGA (cf. Fig. 4B). Of note, however, both datasets revealed concordant archetype-enrichment trends consistent with existing subtype-specific breast cancer therapies. Basal tumors exhibited the highest median Proliferation scores, and HER2+ tumors showed the highest median Biomass scores across both datasets (cf. Fig. 4A,B). Consistent with these enrichment trends and our findings in Fig. 3A, basal tumors are typically treated with DNA damaging agents, while HER2+ tumors are treated with a combination of trastuzumab and *EGFR* inhibitors including lapatinib. Furthermore, basal A, despite being described as having luminal-like properties[54], showed enrichment for Biomass, suggesting a closer resemblance to HER2+ and potential sensitivity to lapatinib and other EGFR inhibitors (cf. Fig. 4A). Luminal B (LumB), considered and intermediate phenotype between luminal A (LumA) and basal[55], also exhibited an intermediate median Proliferation score in TCGA (cf. Fig. 4B).

To evaluate whether archetype enrichment patterns can be used to predict overall survival in breast cancer, we conducted parallel Kaplan-Meier analyses on subgroups of patients in TCGA stratified by their archetype values (cf. Fig. 4C) and according to their reported breast cancer subtype (cf. Fig. 4D). Normal and HER2+ samples were excluded due to their limited representation and distinct treatment standards, respectively. Following our previous findings, the remaining patients were stratified by their Proliferation archetypes scores. Notably, patients with intermediate Proliferation archetype scores exhibited a significantly poorer prognosis compared to their counterparts (log-rank test, $p = 0.00536$). Furthermore, although patients with luminal B breast cancer showed a similar trend towards worse survival, the results did not reach statistical significance. As additional proofs-of-principle for the use of archetypes to predict overall patient survival, we identified archetype-associated groups with poorer survival in two additional TCGA cohorts: Colon Adenocarcinoma ($N = 481$) and Pancreatic Adenocarcinoma ($N = 178$) ([24,25], cf. SI Fig. 2).

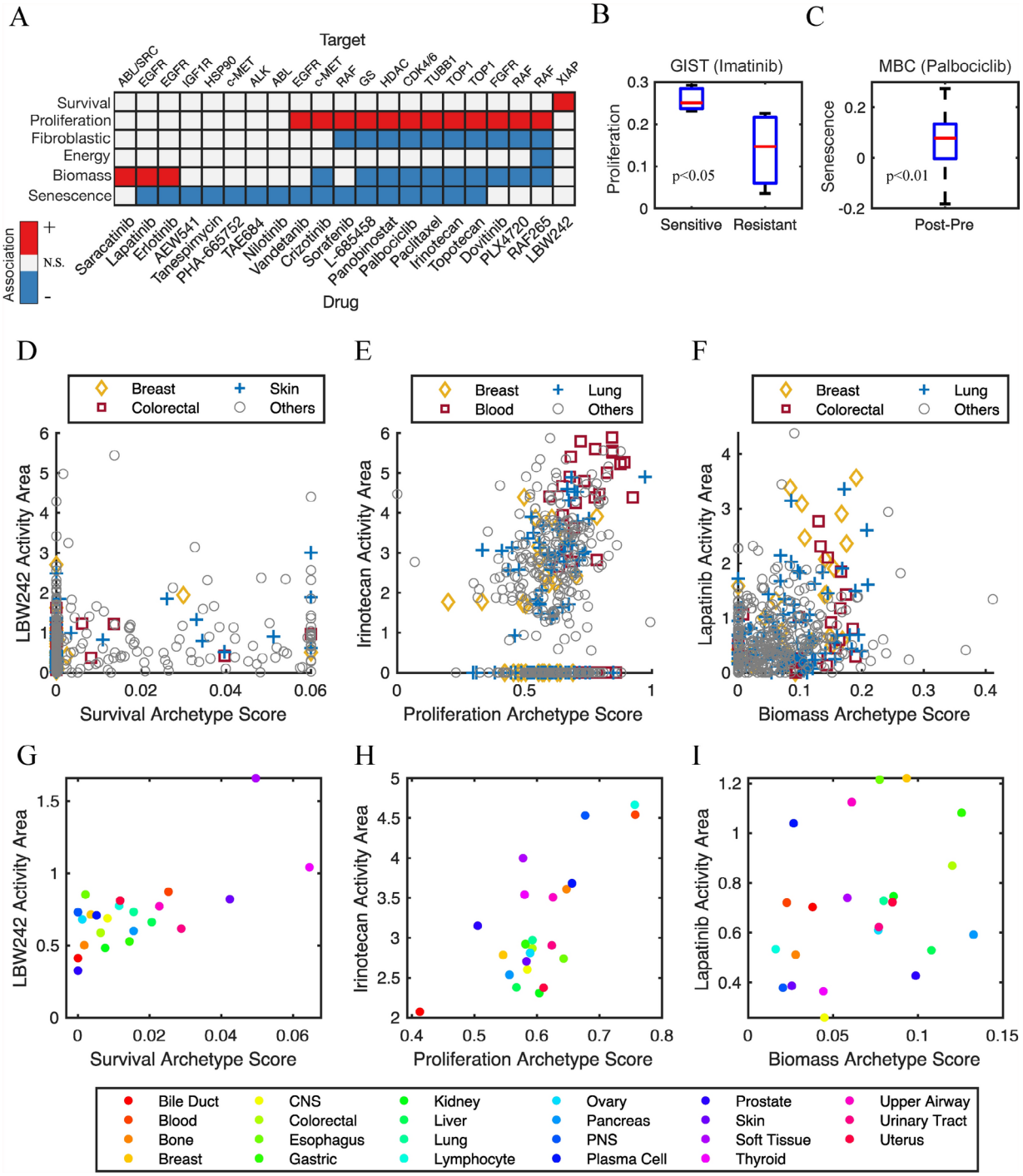

Fig. 3. Associations between normal archetype scores and chemotherapy sensitivities across a range of cancer types. (**A**). Heatmap of the significant associations between individual archetypes and the cancer cell line activity areas of each measured drug (Spearman rank permutation test, Benjamini-Hochberg $p < 0.05$). (**B**). Boxplots comparing the Proliferation archetype scores in imatinib-sensitive and resistant GIST patient samples (t-test, $N = 10$,[26]). (**C**). Boxplot of the change in Senescence archetype scores in HR+/HER2- metastatic breast cancer samples (MBC) before and after treatment with palbociclib (paired t-test, $N = 23$,[27]). (**D**). Scatterplots of the Survival score vs LBW242 activity area, (**E**). Proliferation score vs irinotecan activity area, and (**F**). Biomass score vs lapatinib activity area, stratified by cancer lineage. (**G**). Scatterplots of the Survival score vs LBW242 activity area, (**H**). Proliferation score vs irinotecan activity area, and (**I**). Biomass scores vs lapatinib activity area, averaged by cancer lineage. Cancer lineages shown in (**D**–**F**) were manually selected based on their total number of samples and average archetype expression levels. Cell lines with 0 activity area, indicating incomplete information, were excluded from analysis. To enhance visibility of all data points, regression lines were excluded.

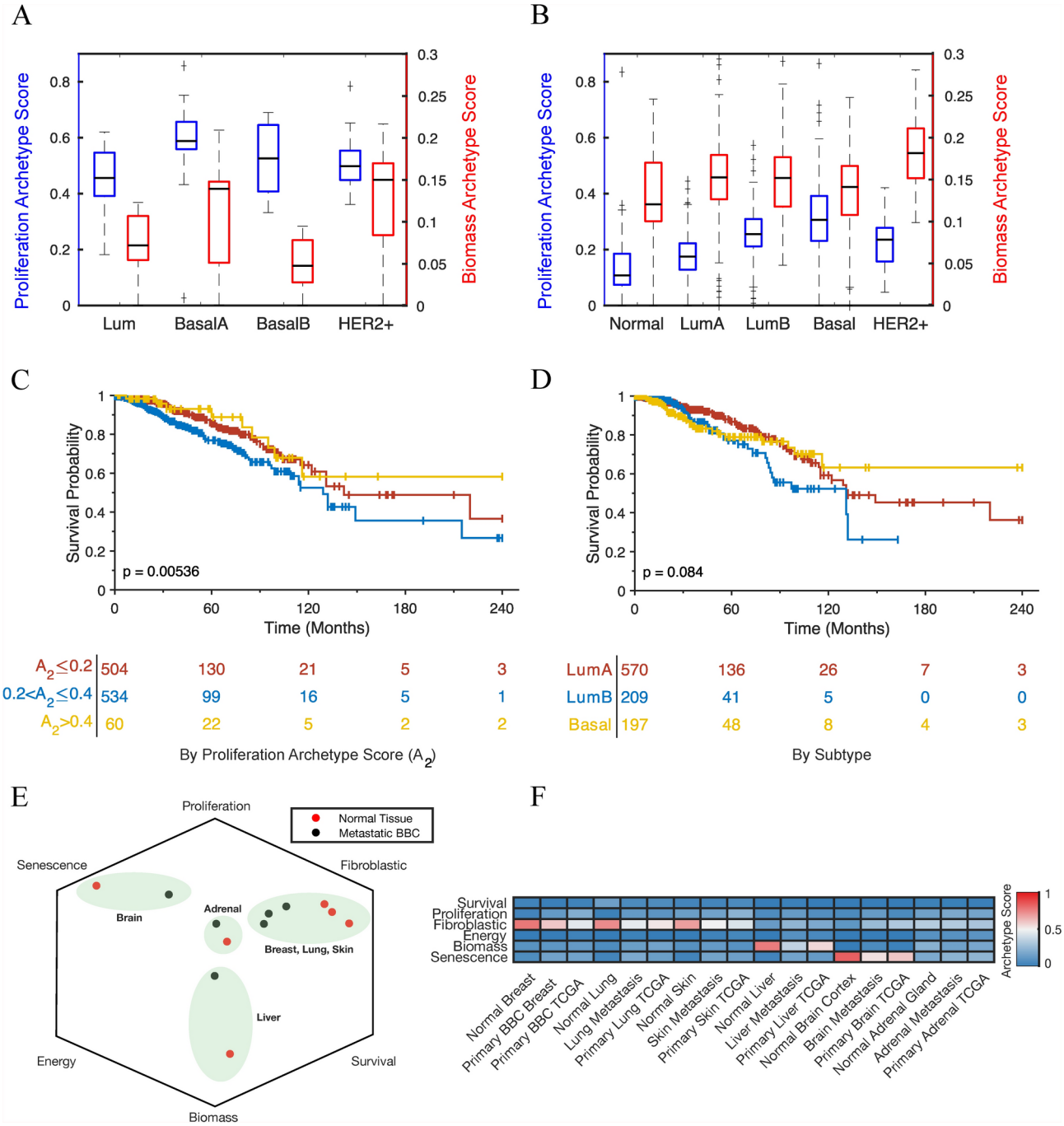

Fig. 4. Associations between normal archetype scores and distinct breast cancer behaviors. (**A**). Boxplots of the Proliferation (blue) and Biomass (red) archetype scores across breast cancer cell lines, stratified by subtype. (**B**). Boxplots of the Proliferation (blue) and Biomass (red) archetype scores across primary breast tumors in TCGA, stratified by subtype. (**C**). Kaplan-Meier survival analysis, stratified by Proliferation archetype score. (**D**). Kaplan-Meier survival analysis, stratified by reported breast cancer subtype. Survival analyses utilized overall survival and RNA-Seq data provided by TCGA. Tables below indicate the number of surviving patients at each time point. Significance was determined using a log-rank test. (**E**). Projection of median-averaged site-specific basal breast metastases (BBC, black) and matched normal tissues from GTEx (orange) projected onto the six normal tissue archetypes. (**F**). Heatmap of the median archetype patterns from breast metastases, normal tissues, and primary tumors associated with each site of metastasis.

Finally, to characterize archetype enrichment patterns in a representative metastasis model, we examined an additional dataset comprising median gene expression profiles of basal breast cancer from primary tumors and five distinct metastatic sites: brain, lung, liver, adrenal gland, and skin ([21,28], cf. Methods). Consistent with[21], breast metastases exhibited archetype enrichment patterns resembling their destination organs more than their primary tumor site (cf. Fig. 4E). However, these distant metastases were less enriched for destination archetypes compared to primary tumors from the same location, indicating an intermediate behavior (cf. Fig. 4F). Nevertheless, these findings, in conjunction with our previous results, could suggest that metastatic acquire treatment sensitivities associated with their destination archetypes.

DNA alterations guide, but do not predict, the onset of non-canonical tissue signatures

We next used CCLE to evaluate the association between tumor archetype scores, copy number alterations (CNA), and mutations (cf. Methods and Fig. 5). Our analysis revealed significant Spearman rank correlations (Benjamini-Hochberg $p < 0.05$) involving 18 out of 73 TCGA considered hotspot mutations and at least one normal tissue archetype (cf. Fig. 5A). Notably, *TP53* mutation was one of the strongest correlates with the Proliferation archetype, aligning with the expected link between *TP53* dysregulation and a shift towards the most commonly enriched tumor archetype across various cancer cell lines (cf. Fig. 2A). However, the effect size was relatively modest, with *TP53* mutation leading to an average increase in the Proliferation archetype score of approximately 0.1 (cf. Fig. 5B).

Our investigation also identified several CNAs significantly associated with normal tissue archetype scores (cf. Fig. 5C). Each archetype was found to be associated with a distinct set of CNAs, similar to our findings regarding mutations. Moreover, multiple regions containing CNAs measured in MSK-IMPACT (a curated DNA sequencing panel of cancer driver genes,[56] were among the top 1% of Spearman correlations (cf. Fig. 5C, inset). This suggests that whole-genome DNA alteration profiles may be able to predict details about tumor archetypes.

To test this hypothesis, we conducted linear regressions using previously identified features, including mutations alone (Fig. 5D), CNAs from MSK-IMPACT alone (Fig. 5E), and mutations combined with CNAs (Fig. 5F). Despite incorporating all 73 hotspot mutations and the top 500 principal components of the CNA data, no regression yielded an R^2 value exceeding 0.60. Consequently, our analyses suggest that normal tissue archetypes offer independent prognostic information about tumors beyond the predictive capacity of the considered genetic features.

Discussion

We introduced a strategy to standardize tumor phenotype quantification across diverse cancers and datasets by identifying conserved transcriptional signatures called archetypes. In doing so, our analysis identified six archetypes – Survival, Proliferation, Fibroblastic, Energy, Biomass, and Senescence – defining common modes of transcriptional variability in both normal tissues and multiple cancers. This universality allows the prediction of disease characteristics, such as drug sensitivities and overall patient survival, across multiple cancers using standardized signatures. Importantly, their association with normal tissue transcriptional modules suggests a connection between site-specific adaptive processes during metastasis and potential therapeutic vulnerabilities. These features, not predictable from mutations and copy number alterations alone, may enhance prognostic capabilities alongside established DNA-alteration-driven personalized frameworks in cancer medicine[57,58].

Site-specific archetype adaptation may present a novel therapeutic vulnerability in metastatic cancers. Consistent with our findings, prior studies indicate that metastatic tumors express genes associated with their target tissues[21,59]. Thus, the dual role of archetypes as markers for such tissue-specific gene expression patterns and predictors of tumor drug sensitivities suggests that site-specific archetype enrichment patterns may highlight vulnerabilities in metastatic tumors to treatments associated with the archetypes of their destination tissues. For instance, breast metastases to the liver, regardless of primary HER2+ status, are anticipated to be responsive to *EGFR* inhibitors such as lapatinib, suggesting a potential untapped therapeutic vulnerability[60]. The question remains whether tumors adapt their archetypes post-metastasis or if cell populations enriched for destination-specific archetypes migrate to those locations[61]. If the latter, interventions could potentially steer primary tumor archetypes, guiding eventual metastases to more manageable sites[62].

Our findings suggest that DNA alterations and archetype enrichment offer distinct yet complementary information. Indeed, recent studies support the notion that DNA alterations alone do not reliably predict transcriptional behaviors or where a cancer will metastasize to[63,64]. Nevertheless, our alteration correlation map (cf. Fig. 5) may aid in interpreting the impact of specific mutations and copy number alterations on tumor phenotype. Conversely, DNA alterations may act, in part, as archetype regulatory elements, steering tumors into specific behaviors and consequently modulating their drug sensitivities. For instance, our analysis associates mutation in *BRAF* with a shift towards the treatment-resistant Senescence archetype, aligning with its role in treatment-resistant cancers[22]. A noteworthy implication is that such alterations might inadvertently sensitize tumors to therapies associated with their new archetype compositions, a phenomenon termed collateral drug sensitivity[65].

The archetype mixture model was trained using normal tissues to maximize signal-to-noise and to compare them with known cellular trade-offs. Furthermore, while we restricted our gene set to five hallmark pathways of broad relevance to cancer, additional archetypes could be contained in the remaining genes. However, even with these restrictions, our archetypes are in broad agreement with known cancer hallmarks[14]. Furthermore, the utilization of N-NMF allows our framework to be easily augmented with additional archetypes and data constraints, allowing our foundation to be used as a starting point for additional development[35].

There are important caveats and considerations for the current analyses. The N-NMF method requires nonnegative data, rendering it unsuitable for application to Z-scores or other transformations containing negative values. Data normalization can also introduce spurious anticorrelations. Thus, when generating

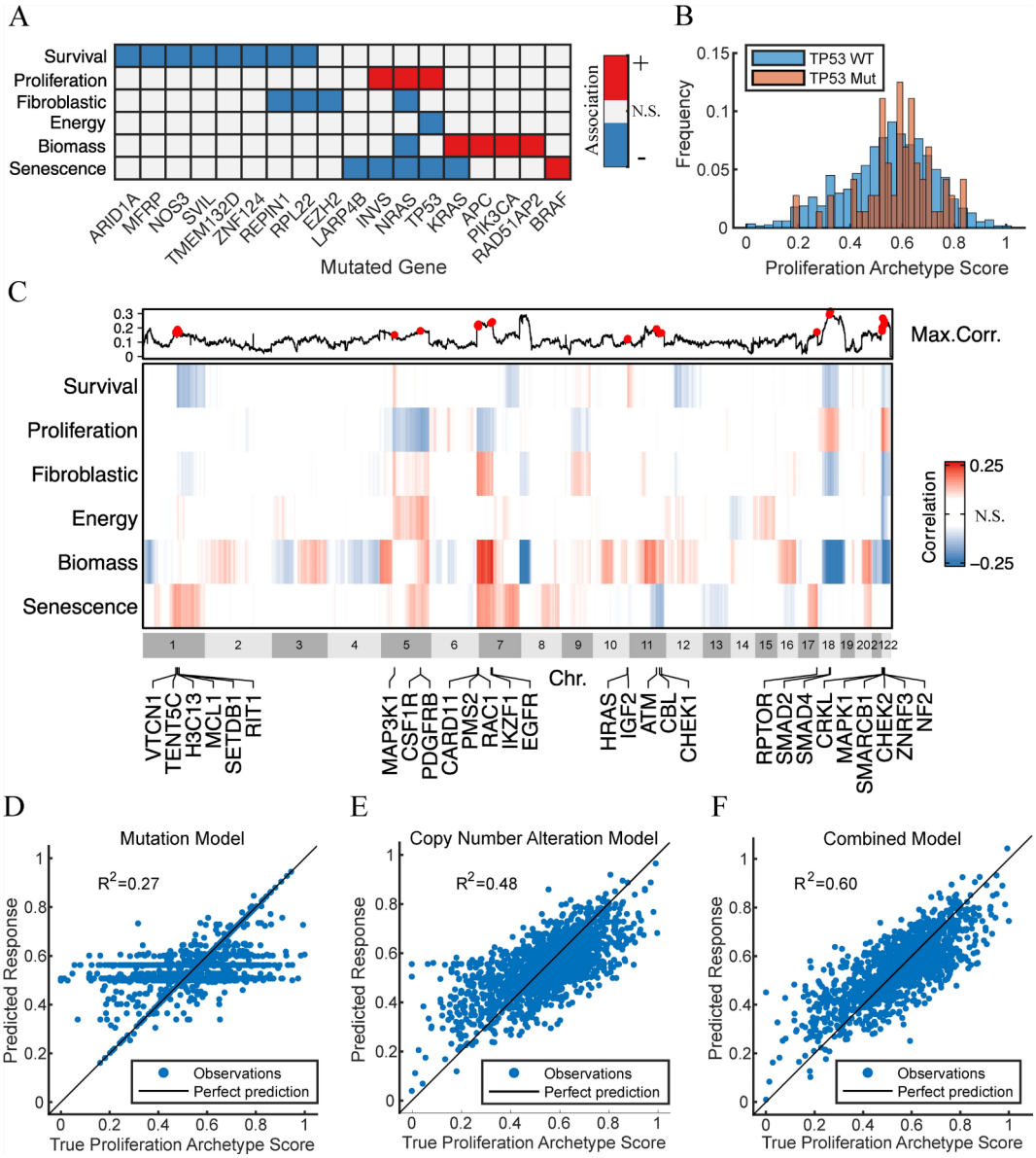

Fig. 5. Associations between normal archetypes, copy number alterations (CNA), and mutations in cancer cell lines. (**A**). Heatmap of the significant associations between individual archetypes and recurrent TCGA hotspot mutations found in CCLE (Spearman rank permutation test, Benjamini-Hochberg $p < 0.05$). (**B**). Histograms of the Proliferation archetype score in cell lines with (orange) and without (blue) a TP53 mutation. (**C**). Heatmap of the significant genome-wide bin-averaged associations between individual archetypes and gene-level CNAs found in CCLE (Spearman rank permutation test, Benjamini-Hochberg $p < 0.05$). The inset above shows the maximum absolute correlation of each bin across the individual archetype scores. Genes labeled and marked in red on the inset were found in both in MSK-IMPACT and in the top 1% of correlations to one of the six archetypes. (**D**). Linear prediction of the Proliferation archetype scores from the 73 recurrent hotspot mutations retained in CCLE. (**E**). Linear prediction of the Proliferation archetype scores from the top 500 principal components of the gene-level CNA data in CCLE. (**F**). Linear prediction of the Proliferation archetype scores from the combined features used in D and E. All regressions were performed using the MATLAB Regression Learner Application. CNA regression analysis was restricted to genes from MSK-IMPACT.

archetypes through N-NMF, it is advisable to compare them with clusters identified in unnormalized data or established biological labels, as we did by mapping the archetypes to distinct tissues. Bulk tumor archetype enrichment patterns may be confounded by stromal tissue infiltration, despite some evidence to the contrary[21]. Furthermore, the significant difference in scale between bulk tumor archetype scores and those from cancer cell lines suggests that current drug sensitivity models cannot be directly applied to bulk tumor samples. The analysis was also biased toward more common cancer types, which may affect the generalizability of the findings. Finally,

a substantial proportion of drug sensitivity cannot be fully explained by applying the same archetype regression models across all tumor types, highlighting the need to also include tumor-specific contextual features.

In summary, six transcriptional signatures identified in normal tissues characterized heterogeneity patterns across a diverse range of human cancers. These signatures predicted both drug sensitivities and prognosis, going beyond what could be anticipated from DNA alterations alone[10,10,63,66]. Notably, these signatures are specifically concentrated in certain sites of basal breast cancer metastases, suggesting a potential new approach for treating metastatic cancers. Ultimately, recognizing the clinical significance of shared transcriptomic behaviors across human cancers may lead to complementary therapies targeting standardized signatures rather than cancer-specific features[10,62].

Received: 28 May 2024; Accepted: 15 October 2024
Published online: 08 November 2024

References

1. Peng, X. et al. Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Rep.***23**(1), 255–269 (2018).
2. Hausser, J. & Alon, U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nat. Rev. Cancer***20**(4), 247–257 (2020).
3. Gavish, A., Tyler, M., Greenwald, A. C., Hoefflin, R., Simkin, D., Tschernichovsky, R., Galili Darnell, N., Somech, E., Barbolin, C., Antman, T. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* 1–9 (2023).
4. Barkley, D. et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.***54**(8), 1192–1201 (2022).
5. Pouryahya, M. et al. Pan-cancer prediction of cell-line drug sensitivity using network-based methods. *Int. J. Mol. Sci.***23**(3), 1074 (2022).
6. Benedetti, E., Liu, E. M., Tang, C., Kuo, F., Buyukozkan, M., Park, T., Park, J., Correa, F., Hakimi, A. A., Intlekofer, A. M. et al. A multimodal atlas of tumour metabolism reveals the architecture of gene–metabolite covariation. *Nat. Metab.* 1–16 (2023).
7. Chifman, J., Pullikuth, A., Chou, J. W., Bedognetti, D. & Miller, L. D. Conservation of immune gene signatures in solid tumors and prognostic implications. *BMC Cancer***16**, 1–17 (2016).
8. Yang, Z., Jones, A., Widschwendter, M. & Teschendorff, A. E. An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol.***16**, 1–15 (2015).
9. Hausser, J. et al. Tumor diversity and the trade-off between universal cancer tasks. *Nat. Commun.***10**(1), 5423 (2019).
10. Mundi, P. S. et al. A transcriptome-based precision oncology platform for patient-therapy alignment in a diverse set of treatment-resistant malignancies. *Cancer Discov.***13**(6), 1386–1407 (2023).
11. Bagaev, A. et al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell***39**(6), 845–865 (2021).
12. Cheng, M. W., Mitra, M. & Coller, H. A. Pan-cancer landscape of epigenetic factor expression predicts tumor outcome. *Commun. Biol.***6**(1), 1138 (2023).
13. Zaorsky, N. G. et al. Pan-cancer analysis of prognostic metastatic phenotypes. *Int. J. Cancer***150**(1), 132–141 (2022).
14. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer the next generation. *Cell***144**(5), 646–674 (2011).
15. Hanahan, D. Hallmarks of cancer: New dimensions. *Cancer Discov.***12**(1), 31–46 (2022).
16. Demicheli, R. & Hrushesky, W. J. M. Reimagining cancer: Moving from the cellular to the tissue level. *Can. Res.***83**(2), 173–180 (2023).
17. Weiße, A. Y., Oyarzún, D. A., Danos, V. & Swain, P. S. Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci.***112**(9), E1038–E1047 (2015).
18. Jacqueline, C. et al. Cancer: A disease at the crossroads of trade-offs. *Evol. Appl.***10**(3), 215–225 (2017).
19. Cutler, A. & Breiman, L. Archetypal analysis. *Technometrics***36**(4), 338–347 (1994).
20. MacCarthy-Morrogh, Lucy & Martin, Paul. The hallmarks of cancer are also the hallmarks of wound healing. *Sci. Signal.***13**(648), eaay8690 (2020).
21. Roshanzamir, F., Robinson, J. L., Cook, D., Karimi-Jafari, M. H. & Nielsen, J. Metastatic triple negative breast cancer adapts its metabolism to destination tissues while retaining key metabolic signatures. *Proc. Natl. Acad. Sci.***119**(35), e2205456119 (2022).
22. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature***483**(7391), 603–607 (2012).
23. Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. 2012 Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, .
24. Network, C. G. A. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature***487**(7407), 330 (2012).
25. Raphael, B. J. et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell***32**(2), 185–203 (2017).
26. Shao, Y. et al. Rp11-616m22. 7 recapitulates imatinib resistance in gastrointestinal stromal tumor. *Mol. Ther.-Nucleic Acids***25**, 264–276 (2021).
27. Park, Y. H. et al. Longitudinal multi-omics study of palbociclib resistance in hr-positive/her2-negative metastatic breast cancer. *Genome Med.***15**(1), 55 (2023).
28. Siegel, M. B. et al. Integrated rna and dna sequencing reveals early drivers of metastatic breast cancer. *J. Clin. Investig.***128**(4), 1371–1383 (2018).
29. GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science***369**(6509), 1318–1330 (2020).
30. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.***102**(43), 15545–15550 (2005).
31. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst.***1**(6), 417–425 (2015).
32. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell***173**(2), 291–304 (2018).
33. Colaprico, A. et al. Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Res.***44**(8), e71–e71 (2016).
34. Kunes, Russell Z., Walle, Thomas, Nawy, Tal. & Pe'er, Dana. Supervised discovery of interpretable gene programs from single-cell data. *bioRxiv*, (2022).
35. Wang, Y.-X. & Zhang, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowl. Data Eng.***25**(6), 1336–1353 (2012).

36. Zhu, M. & Ghodsi, A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Stat. Data Anal.***51**(2), 918–930 (2006).
37. Maisog, J. M. et al. Assessing methods for evaluating the number of components in non-negative matrix factorization. *Mathematics***9**(22), 2840 (2021).
38. Lee, D. D. & Sebastian, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature***401**(6755), 788–791 (1999).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)***57**(1), 289–300 (1995).
40. Creed, J. H., Gerke, T. A. & Berglund, A. E. Matsurv: Survival analysis and visualization in matlab. *J. Open Source Softw.***5**(46), 1830 (2020).
41. Team, R Core. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2024).
42. The MathWorks Inc. Matlab version: 9.9.0 (r2020b), (2020).
43. Yamashita, M. & Passegué, E. Tnf-α coordinates hematopoietic stem cell survival and myeloid regeneration. *Cell Stem Cell***25**(3), 357–372 (2019).
44. Carroll, P. A., Freie, B. W., Mathsyaraja, H. & Eisenman, R. N. The myc transcription factor network: Balancing metabolism, proliferation and oncogenesis. *Front. Med.***12**, 412–425 (2018).
45. Newman, A. C., Nakatsu, M. N., Chou, W., Gershon, P. D. & Hughes, C. C. W. The requirement for fibroblasts in angiogenesis: Fibroblast-derived matrix proteins are essential for endothelial cell lumen formation. *Mol. Biol. Cell***22**(20), 3791–3800 (2011).
46. Wang, L., Lankhorst, L. & Bernards, R. Exploiting senescence for the treatment of cancer. *Nat. Rev. Cancer***22**(6), 340–355 (2022).
47. Azad Kumar, J. et al. Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol. Aging***34**(4), 1199–1209 (2013).
48. Nie, X. et al. Single-cell analysis of human testis aging and correlation with elevated body mass index. *Dev. Cell***57**(9), 1160–1176 (2022).
49. Buckner, R. L. The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. *Neuron***80**(3), 807–815 (2013).
50. Uhlén, M. et al. Tissue-based map of the human proteome. *Science***347**(6220), 1260419 (2015).
51. Tsherniak, A. et al. Defining a cancer dependency map. *Cell***170**(3), 564–576 (2017).
52. Arnedos, M. et al. Modulation of rb phosphorylation and antiproliferative response to palbociclib: the preoperative-palbociclib (pop) randomized clinical trial. *Ann. Oncol.***29**(8), 1755–1762 (2018).
53. Yin, L., Duan, J.-J., Bian, X.-W. & Shi-cang, Yu. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.***22**, 1–13 (2020).
54. Dai, X., Cheng, H., Bai, Z. & Li, J. Breast cancer cell line classification and its relevance with breast tumor subtyping. *J. Cancer***8**(16), 3131 (2017).
55. Creighton, Chad J. The molecular profile of luminal b breast cancer. *Biol. Targets Ther.* 289–297 (2012).
56. Cheng, D. T. et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.***17**(3), 251–264 (2015).
57. Pleasance, E. et al. Whole-genome and transcriptome analysis enhances precision cancer treatment options. *Ann. Oncol.***33**(9), 939–949 (2022).
58. Adam, G. et al. Machine learning approaches to drug response prediction: Challenges and recent progress. *NPJ Precis. Oncol.***4**(1), 19 (2020).
59. Sanghvi, Neel, Calvo-Alcañiz, Camilo, Rajagopal, Padma, Scalera, Stefano, Canu, Valeria, Sinha, Sanju, Schischlik, Fiorella, Wang, Kun, Madan, Sanna, Shulman, Eldad. et al. Charting the transcriptomic landscape of primary and metastatic cancers in relation to their origin and target normal tissues. *bioRxiv*, pages 2023–10, (2023).
60. Rashid, N. S., Grible, J. M., Clevenger, C. V. & Chuck Harrell, J. Breast cancer liver metastasis current and future treatment approaches. *Clin. Exp. Metastasis***38**, 263–277 (2021).
61. Gao, Y. et al. Metastasis organotropism: Redefining the congenial soil. *Dev. Cell***49**(3), 375–391 (2019).
62. Ganesh, K. & Massague, J. Targeting metastatic cancer. *Nat. Med.***27**(1), 34–44 (2021).
63. Califano, A. & Alvarez, M. J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. Rev. Cancer***17**(2), 116–130 (2017).
64. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell***185**(3), 563–575 (2022).
65. Acar, A. et al. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer. *Nat. Commun.***11**(1), 1–14 (2020).
66. Chawla, S. et al. Gene expression based inference of cancer drug sensitivity. *Nat. Commun.***13**(1), 5680 (2022).

Acknowledgements

Research presented here was funded by the following: the Marie-Josée Kravis Fellowship in Quantitative Biology (C.W.), the Laufer Center for Physical and Quantitative Biology (K.A.D.), AFOSR grants FA9550-20-1-0029 and FA9550-23-1-0096 (A.R.T.), NIH grant R01-AG048769 (A.R.T.), a grant from Breast Cancer Research Foundation BCRF-17-193 (J.O.D., L.N., and A.R.T.), Army Research Office grant W911NF2210292 (A.R.T.), and a grant from the Cure Alzheimer's Foundation (A.R.T.).

Author contributions

Conceptualization: C.W.; Methodology: C.W., K.A.M, and J.Z.; Analysis and investigation: C.W. and K.A.M.; Writing: C.W., K.A.M., and J.O.D.; Editing: C.W., K.A.M., J.Z., L.N., K.A.D., A.R.T, and J.O.D.; Supervision: A.R.T. and J.O.D.; Funding acquisition: C.W., L.N., A.R.T, and J.O.D.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76625-1.

Correspondence and requests for materials should be addressed to J.O.D.