Digital profiling of cancer transcriptomes from histology images with grouped vision attention

Yuanning Zheng[1,*], Marija Pizurica[1,2,*], Francisco Carrillo-Perez[1,*], Humaira Noor[1], Wei Yao[3], Christian Wohlfart[4], Kathleen Marchal[2], Antoaneta Vladimirova[3], Olivier Gevaert[1,5,#]

[1]Department of Medicine, Stanford Center for Biomedical Informatics Research (BMIR), Stanford University, Stanford, 94305, USA.

[2]Internet technology and Data science Lab (IDLab), Ghent University, Technologiepark-Zwijnaarde 126, Ghent, 9052, Gent, Belgium.

[3]Roche Information Solutions, Inc., Santa Clara, CA.

[4]Roche Diagnostics GmbH, Penzberg, Germany.

[5]Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA.

[*]These authors contributed equally to this work.

[#]Corresponding author: ogevaert@stanford.edu.

Abstract

Cancer is a heterogeneous disease that demands precise molecular profiling for better understanding and management. RNA-sequencing has emerged as a potent tool to unravel the transcriptional heterogeneity. However, large-scale characterization of cancer transcriptomes is hindered by the limitations of costs and tissue accessibility. Here, we develop ***SEQUOIA***, a deep learning model employing a transformer architecture to predict cancer transcriptomes from whole-slide histology images. We pre-train the model using data from 2,242 normal tissues, and the model is fine-tuned and evaluated in 4,218 tumor samples across nine cancer types. The results are further validated across two independent cohorts compromising 1,305 tumors. The highest performance was observed in cancers from breast, kidney and lung, where ***SEQUOIA*** accurately predicted 13,798, 10,922 and 9,735 genes, respectively. The well predicted genes are associated with the regulation of inflammatory response, cell cycles and hypoxia-related metabolic pathways. Leveraging the well predicted genes, we develop a digital signature to predict the risk of recurrence in breast cancer. While the model is trained at the tissue-level, we showcase its potential in predicting spatial gene expression patterns using spatial transcriptomics datasets. ***SEQUOIA*** deciphers clinically relevant gene expression patterns from histology images, opening avenues for improved cancer management and personalized therapies.

Introduction

Estimates from the World Health Organization (WHO) in 2019 [1] revealed cancer as the primary or secondary cause of death before the age of 70 years in 112 out of 183 countries, and third or fourth leading cause of death in another 23 countries. Its multifaceted nature characterized by diverse subtypes, intricate molecular profiles and both inter- and intra-patient heterogeneity presents formidable challenges for effective diagnosis and treatment.

With the growing interest in personalized and precision medicine, molecular profiling has gained significant attention as a critical component of prognostication and treatment planning for various cancer types [2]. In the past decades, the advancement of RNA sequencing (RNA-seq) has deepened our understanding of inter-patient heterogeneity, providing a comprehensive view of gene expression patterns and biological processes within tumors. This information has led to the discovery of molecular signatures specific to different cancer types or subtypes, enabling more accurate diagnosis. [3–5].

However, deriving gene expression profiles from tissue samples is still a bottleneck in the workflow of clinical practice. Current methods involve time-consuming and expensive laboratory procedures, limiting the widespread integration of gene expression analysis in routine diagnostics. This is especially the case for cancers with high degree of intra-tumoral heterogeneity, where the analysis of multiple tumor regions is needed to determine the subtype and malignant status.

With the digitization of histopathology glass slides to Whole Slide Images (WSIs), unprecedented opportunities arise for cost-efficient analyses of tumor properties. Namely, WSIs are available without additional cost as they are obtained in routine clinical practice for diagnostic purposes. Despite providing only morphological information, WSIs may also reflect the molecular status of the tumor. In recent studies, deep learning-based computational methods were used to extract hidden morphological features from WSIs that associate with molecular properties, such as aneuploidies, genetic alterations and expression signatures of cancer infiltrating immune cells [6–17]. Hence, deep learning models predicting gene expression from WSI offer a cost-efficient way to infer and analyze gene expression patterns on a large scale, with potential applications in both research and clinical settings.

Although remarkable progress has been made in computer vision on medical images, the application of state-of-the-art methods on WSIs remains exceedingly challenging due to their immense resolutions and the presence of noisy labeling. To enable the application of feature extraction techniques (e.g., ResNet), the WSI is first divided into thousands of smaller tiles. Notably, slide-level bulk sequencing labels may correspond to only a small portion of the WSI, which may not fully capture the molecular heterogeneity of the entire tissue. The lack of fine-grained annotations for training tile-level models [6, 9, 12, 13, 16, 18] poses significant challenges in extracting relevant morphological information that is associated with gene expression.

Here, we present *SEQUOIA*, a deep learning model for **S**lide-based **E**xpression **Qu**antification using gr**o**uped V**i**sion **A**ttention. We employ the concept of multiple instance learning (MIL) [19], where training instances (i.e. tiles) are organized into bags and the slide-level label is assigned to the bag as a whole [20, 21]. To generate contextualized representations of image features, we utilize the transformer architecture, and a model is trained to automatically derive which tiles from the bag are relevant for the slide-level prediction. Furthermore, we harness the power of transfer learning, where the weight parameters are pre-trained using data from normal tissues. The model is fine-tuned and evaluated using data from 4,218 tumor samples across nine cancer types, and the results were further validated in two independent cohorts. Our analysis reveals the model's capacity in accurately predicting gene expression governing cancer progression, recurrence, and therapy resistance. Finally, we demonstrate the clinical utility of our model in predicting cancer recurrence and unveiling spatial gene expression patterns.

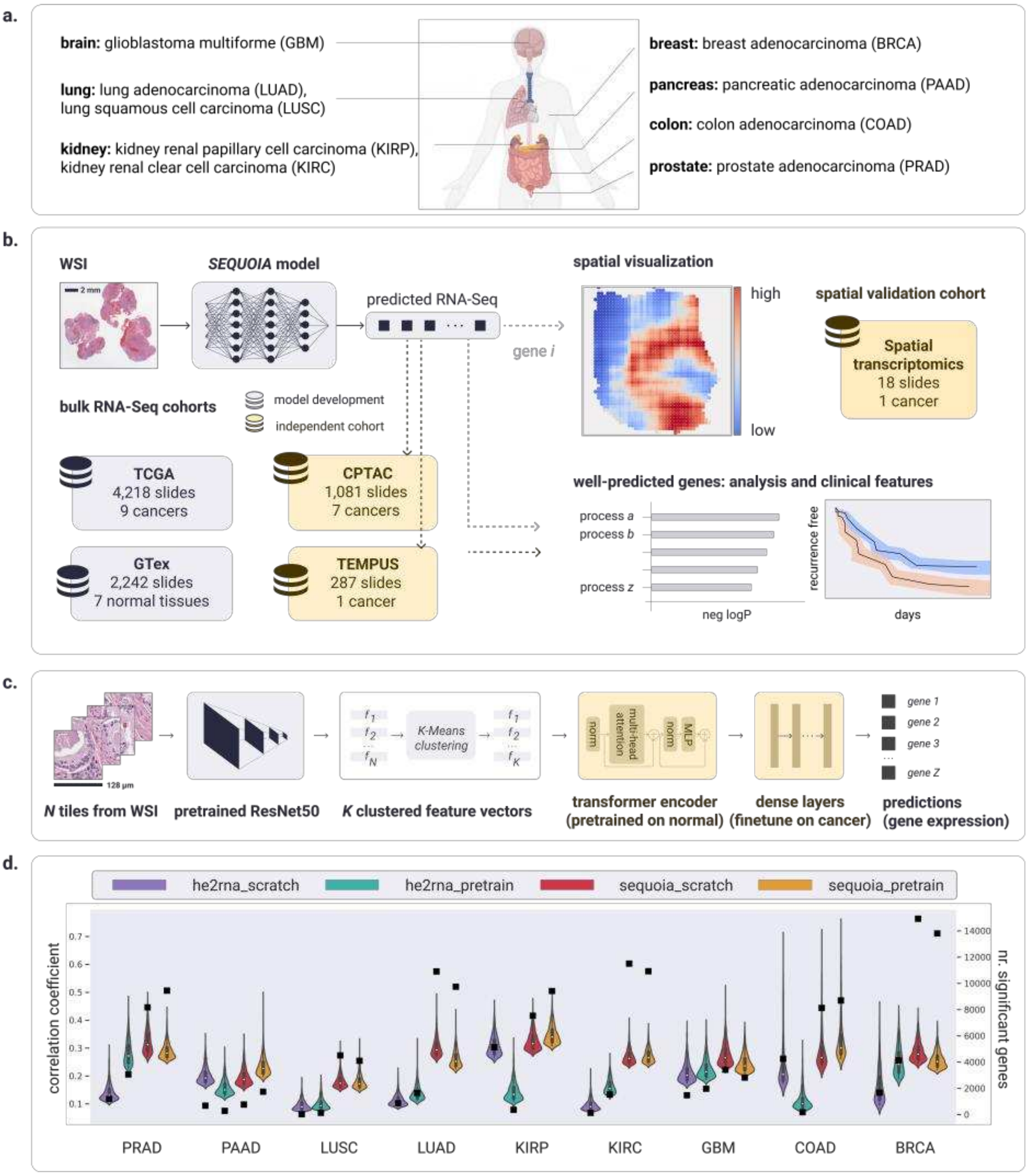


Fig. 1: **Overview of workflow for the** *SEQUOIA* **model.** a) Cancer types on which the *SEQUOIA* model is developed and validated. The panel is created with BioRender.com. b) The model is trained and evaluated using matched WSIs and bulk RNA-Seq data from nine cancer types available in the TCGA database. To pre-train the transformer encoder, we use data of normal tissues from the GTEx database. The model is independently validated using data from the CPTAC and Tempus cohorts. Apart from predicting tissue-level gene expression, we integrate a spatial prediction technique that elucidates region-level gene expression patterns within tumor tissues, validated using spatial transcriptomics datasets [22]. Clinical utility is demonstrated by evaluating the model's capacity to predict cancer recurrence. c) *SEQUOIA* architecture. First, N tiles are sampled from the WSI, and a feature vector is extracted from each tile using a pre-trained ResNet-50 module. We then cluster the feature vectors into K clusters, and an average feature vector is obtained from each cluster, resulting in K aggregated feature vectors. Next, a transformer encoder and dense layers translate the obtained K feature vectors to the predicted gene expression values. d) Performance of *SEQUOIA* compared to *HE2RNA*. For both architectures, we show the performance when trained from scratch and when finetuning from a model pre-trained on normal tissues. Violin plots illustrate the distribution of Pearson correlation coefficients (left y axis) between the predicted and ground truth gene expression values in the TCGA test sets. The top 1,000 genes with the highest correlation coefficients in each architecture are included. Black squares indicate the absolute number (right y axis) of significantly accurately predicted genes across each cancer.

Results

SEQUOIA as tool for gene expression prediction from WSIs

We present *SEQUOIA*, a deep learning model for **S**lide-based **E**xpression **Qu**antification using gr**o**uped v**i**sion **A**ttention ("Methods", Figures 1a-b-c). To train and evaluate the model, we utilized WSIs and RNA-seq gene expression data of nine cancer types available in The Cancer Genome Atlas (TCGA): (1) prostate adenocarcinoma (PRAD), (2) pancreatic adenocarcinoma (PAAD), (3) lung adenocarcinoma (LUAD), (4) lung squamous cell carcinoma (LUSC), (5) kidney renal papillary cell carcinoma (KIRP), (6) kidney renal clear cell carcinoma (KIRC), (7) glioblastoma multiforme (GBM), (8) colon adenocarcinoma (COAD), and (9) breast invasive carcinoma (BRCA).

Since the tumor architecture and gene expression profiles differ between cancer types, the model was developed and validated independently in each cancer type. To evaluate the model, we carried out five-fold cross-validation. In each iteration, slides from 80% of the patients were used for training and internal validation, and 20% for testing. To assess the prediction performance, we combined Pearson's correlation analysis and Root Mean Squared Error (RMSE) ("Methods"). A gene is defined "well-predicted" in case of statistical significant, positive correlation between the ground truth and predicted gene expression values across the test cohorts. In addition, we required the Pearson's correlation coefficient for a well-predicted gene to be significantly higher than the coefficient obtained with an untrained model with the same architecture.

When trained from scratch, *SEQUOIA* was able to generate many well-predicted genes. On average, 7,756 out of 25,749 genes were well predicted across the nine cancer types (Supplementary Table A4, Figure 1d). Overall, the number of well-predicted genes was positively correlated with the number of available training samples available in each cancer (Supplementary Table A1). The highest number ($N = 14,915$) of genes was identified in BRCA, the cancer type with the most available slides ($N = 1,053$ slides). Further, we identified 11,505 well-predicted genes in KIRC ($N = 515$ slides) and 10,900 genes in LUAD ($N = 536$ slides). Comparatively, PAAD and GBM had the lowest number of well-predicted genes as well as the lowest number of slides (PAAD: $N = 757$ genes from $N = 195$ slides; GBM: $N = 3,413$ genes from $N = 236$ slides).

The correlation observed between the number of accurately predicted genes and the size of the available training data suggests that the model can reach higher performance if a larger training dataset was incorporated. Since published studies have revealed advantages of pre-training transformer models on data of the same modality [23], we took steps to pre-train the weights of each model using WSIs and RNA-seq data from normal tissues in the GTex cohort [24]. We found that finetuning a model pre-trained on normal tissues increased the number of well-predicted genes in four cancer types (i.e., PAAD, KIRP, PRAD , COAD), in which a relatively small numbers ($N <= 450$) of tumor slides were available (Figure 1d and Supplementary Table A4). On average, *SEQUOIA* significantly predicted 7,851 out of 25,749 genes across the nine cancer types. As with training from scratch, the highest number of well-predicted genes were found in BRCA (13,798 genes) and least in PAAD (1,737 genes).

Since the histological appearance of BRCA has been shown to be associated with hormone receptor status [25], we separately assessed the performance in the estrogen receptor (ER) negative and ER positive BRCA subtypes. In the ER positive subtype, we identified 8,517 well-predicted genes, and in ER negative subtype 3,840 genes were well-predicted (Supplementary Figure A3a). Of these genes, 2,103 genes were significantly predicted in both subtypes. These results demonstrate the capacity of *SEQUOIA* in predicting gene expression signals specific to breast cancer subtypes.

To compare the performance of our model with existing architectures, we benchmarked our results with the *HE2RNA* [26] model. *SEQUOIA* outperformed *HE2RNA* in all cancer types, irrespective of using the pre-trained models or training from scratch (Supplementary Table A4). Notably, in BRCA, the *SEQUOIA* model identified three times more genes (13,798 genes versus 4,117 genes) than the *HE2RNA* model. Further, in LUAD and KIRC, the number of well-predicted genes was respectively six and seven times higher for *SEQUOIA*. The cancer type with the smallest factor of increase (×1.4) was GBM, where *SEQUOIA* significantly predicted 2,820 genes as compared to 1,963 genes using *HE2RNA*.

Finally, to compare the quality of predictions, we compared the correlation coefficients between the prediction and ground truth between the $SEQUOIA$ and $HE2RNA$ model (Figure 1d, Supplementary Table A5). We found that $SEQUOIA$ significantly outperformed $HE2RNA$ in all cancer types (Supplementary Figure A2, Whitney U test, $P < 0.0001$).

Characterization of the accurately predicted genes

In our subsequent analysis, we focused on results obtained from the pre-trained $SEQUOIA$ models. The well-predicted genes include protein-coding genes, long non-coding RNAs (lncRNAs) and micro-RNAs (miRNAs). On average, over 90% of the genes are protein-coding genes (Supplementary Figure A3b). To characterize their biological functions, we carried out gene set enrichment analysis in each individual cancer type, and three different gene sets were considered: (1) gene ontology, (2) KEGG pathway and (3) cell-type signatures. Gene ontology analysis revealed several common pathways that were significantly well predicted across cancer types, including lymphocyte proliferation (e.g., *ARG1*, *HLA-DRB1*, *IL10*), T cell activation (e.g., *CCL2*, *CCR2*, *CCDC88B*), cell-matrix adhesion (e.g., *ECM2*, *FN1*, *EMP2*) and response to oxidative stress (e.g., *TP53*, *PRDX1*, *VRK2*) (Figure 2a and Supplementary Data 1).

In addition, some gene sets were found in specific cancer types. In GBM (Figure 2b), we identified genes associated with macrophage migration (*CCL2*, *EDN2*, *CKLF*), protein kinase B signaling (*TNF*, *ADTRP*, *SETX*), and endothelial cell development (*ROCK1*, *IKBKB*, *TNFRSF1A*). Similarly, in LUSC (Supplementary Figure A3c), we identified genes associated with collagen biosynthetic process (*CREB3L1*, *WNT4*, *TGFB1*), natural killer cell differentiation (*PTPRC*, *PIK3CD*, *KAT7*), and histone H2A acetylation (*ACTL6A*, *MEAF6*, *DMAP1*). Gene sets identified in other cancer types are listed in Supplementary Data 1.

Furthermore, KEGG pathway analysis revealed that the well-predicted genes are involved in the PD-L1 expression and check point pathway (*CD247*, *CD274*, *CD14*), NF-kappa B signaling (*CXCL12*, *SYK*, *PRKCB*), HIF-1 signaling (*GAPDH*, *HIF1A*, *VEGFA*), and p53 signaling (*TP53I3, PTEN*, *CDK4*) (Figures 2c-d and Supplementary Data 2). Additionally, we identified several cell-type signatures, including cell-type markers for endothelial cells (*CD69*, *CD93*), CD4 T cell (*CD3E*, *CD4*, *CD48*), M2 macrophage (*CD14*, *CD163*, *CD84*), and B cell (*CD19*, *CD53*, *CD37*) (Figure 2e and Supplementary Data 3). Overall, these results highlight the critical biological functions of the accurately predicted genes in regulating cell cycles, inflammation and hypoxia response.

We next extended our functional analyses to the well predicted lncRNA genes (”Methods”). We focused on the lncRNAs that were significantly predicted in at least three cancer types, resulting in a set of 449 unique lncRNAs (Supplementary Figure A3d). Gene set analysis revealed that they are involved in the regulation of T cell activation (*LINC00528*, *LINC00861*, *LINC02195*), stem cell maintenance (*FGD5-AS1*, *MIR100HG*, *TRBV11-2*), epigenetic regulation (*LINC01089*, *FGD5-AS1*, *PSMA3-AS1*) and chromatin remodeling (*FGD5-AS1*, *LINC01355*, *LINC01857*) (Figure 2f, Supplementary Figure A3e and Supplementary Data 4). Collectively, these results demonstrate the capacity of $SEQUOIA$ in predicting the expression of both protein-coding genes and lncRNAs.

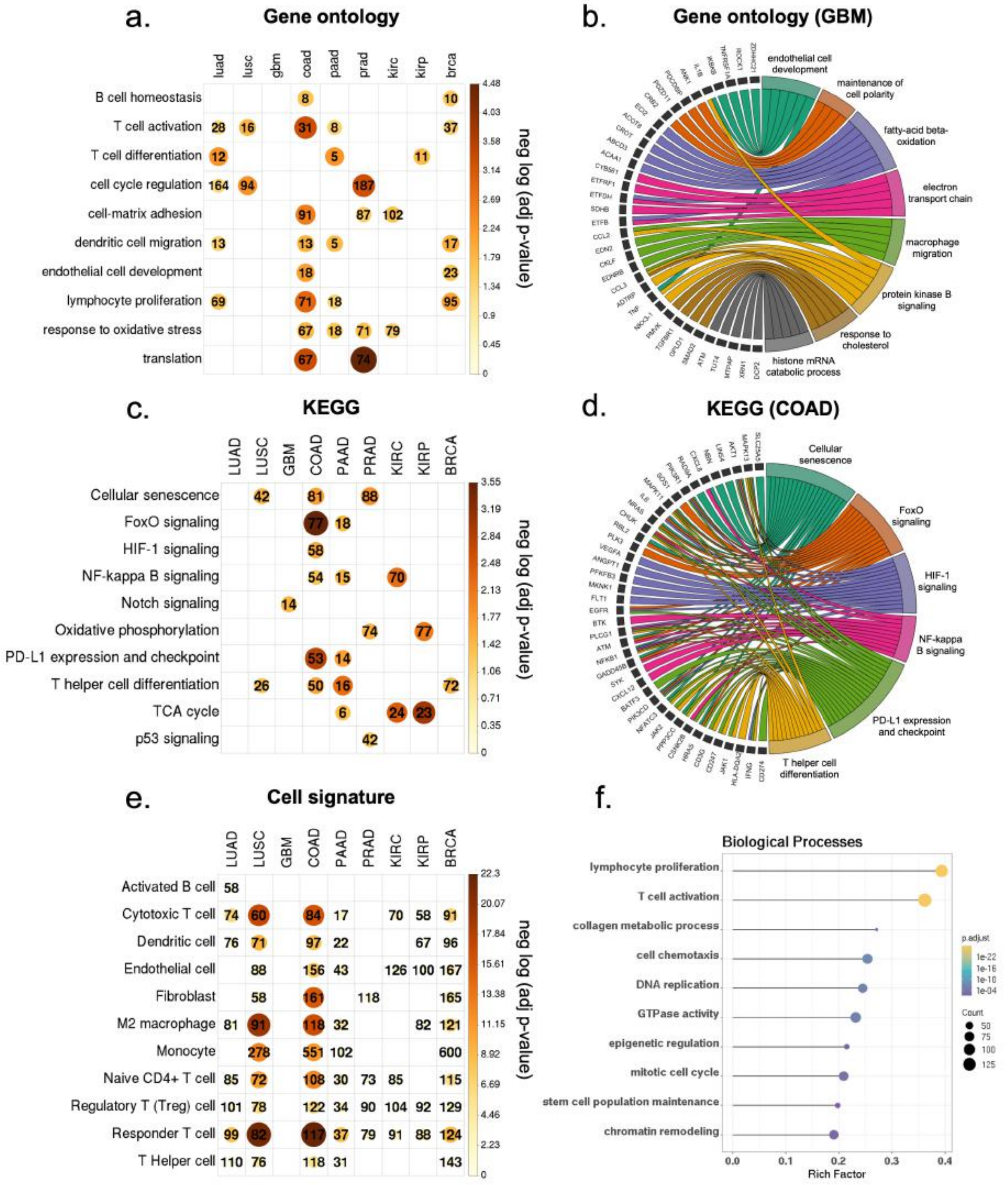


Fig. 2: **Biological functions of the well-predicted genes.** a) Heatmap showing the significant P values from the gene ontology analysis of the well-predicted genes in each cancer type. Color and size of the circles represent the negative log-transformed P values. Integers represent the absolute gene count in each category, and non-significant categories are left in blank. b) Circos plot showing the enriched biological processes associated with the well-predicted genes in GBM. Gene names are displayed on the left and the corresponding biological processes are shown on the right. c) Heatmap showing the significant P values of the KEGG pathways across cancer types. Color and size of the circles represent the negative log-transformed P values. Integers represent the absolute gene count in each category, and non-significant categories are left in blank. d) Circos plot showing the KEGG pathways associated with the well-predicted genes in COAD. Gene names are displayed on the left and the corresponding pathways on the right. e) Heatmap showing the significant P values for the enrichment of cell-type signatures across cancer types. Color and size of the circles represent the negative log-transformed P values. Integers represent the absolute gene count in each category, and non-significant categories are left in blank. f) Top enriched biological processes associated with the well-predicted lncRNA genes. Sizes of the circles represent the number of genes in each biological process, and colors represent significant P values from the enrichment analysis. P values were adjusted for multiple testing using the Benjamini–Hochberg method.

SEQUOIA generalizes to independent cohorts

To test the generalization capacity of *SEQUOIA*, we applied the models developed in each cancer with the TCGA cohort to the matched cancer type in the CPTAC cohort [27–33]. We extended our validation to cancers from six tissues, including breast, lung, kidney, brain, colon and pancreas. Since the cohort size in CPTAC is smaller compared to the TCGA (Supplementary Table A2), it is expected that fewer genes can pass our significance threshold. Despite this limitation, we were able to validate many well-predicted genes (Figure 3a). In BRCA, we validated 10,082 genes in the CPTAC cohort that overlapped with the predictions from the TCGA test cohort. This accounted for 73% of all significant genes ($N =$ 13,798 genes) identified in the TCGA cohort. Additionally, we identified 4,437 (41%) genes in KIRC, 4,226 (49%) genes in COAD, 4,060 (42%) genes in LUAD, 3,030 (74%) genes in LUSC, 1,517 (54 %) genes in GBM and 743 (43%) genes in PAAD.

Gene ontology analysis of the overlapping genes between the TCGA and CPTAC cohorts revealed their key functions in regulating cell proliferation, inflammatory response and tumor growth. Specifically, they are associated with the regulation of cell cycle, B cell proliferation, T cell differentiation and angiogenesis (Figure 3b). Furthermore, KEGG pathway analysis showed that the overlapping genes are associated with cell adhesion, NF-kappa B signaling, T cell receptor signaling and p53 signaling pathway (Figure 3c). To benchmark the generalization capacity of our model to existing architectures, we compared the prediction results to those from the *HE2RNA* model. Although *HE2RNA* identified numerous significant genes in both the TCGA and CPTAC cohorts, only a limited overlap of genes was identified between the two cohorts (Supplementary Table A6).

To further test the generalization capacity, we extended the validation to a lung adenocaricnoma (LUAD) cohort from Tempus ("Methods", $N = 287$ slides from $N = 249$ patients). This led to the identification of 763 genes that were well-predicated across all three (TCGA, CPTAC and Tempus) cohorts for lung cancer patients. Functional analysis of these genes revealed their regulatory functions in inflammatory response (*ITGAL*, *CYBB*, *PTGER4*), angiogenesis (*VEGFD*, *TSPAN12*, *EMP2*), ERK signaling (*PRKCZ*, *PDGFA*, *FGF10*), and Wnt signaling pathway (*WIF1*, *SFRP2*, *NKD2*) (Figure 3d). These results demonstrate the generalization capacity of *SEQUOIA* in predicting gene expression values across independent cohorts.

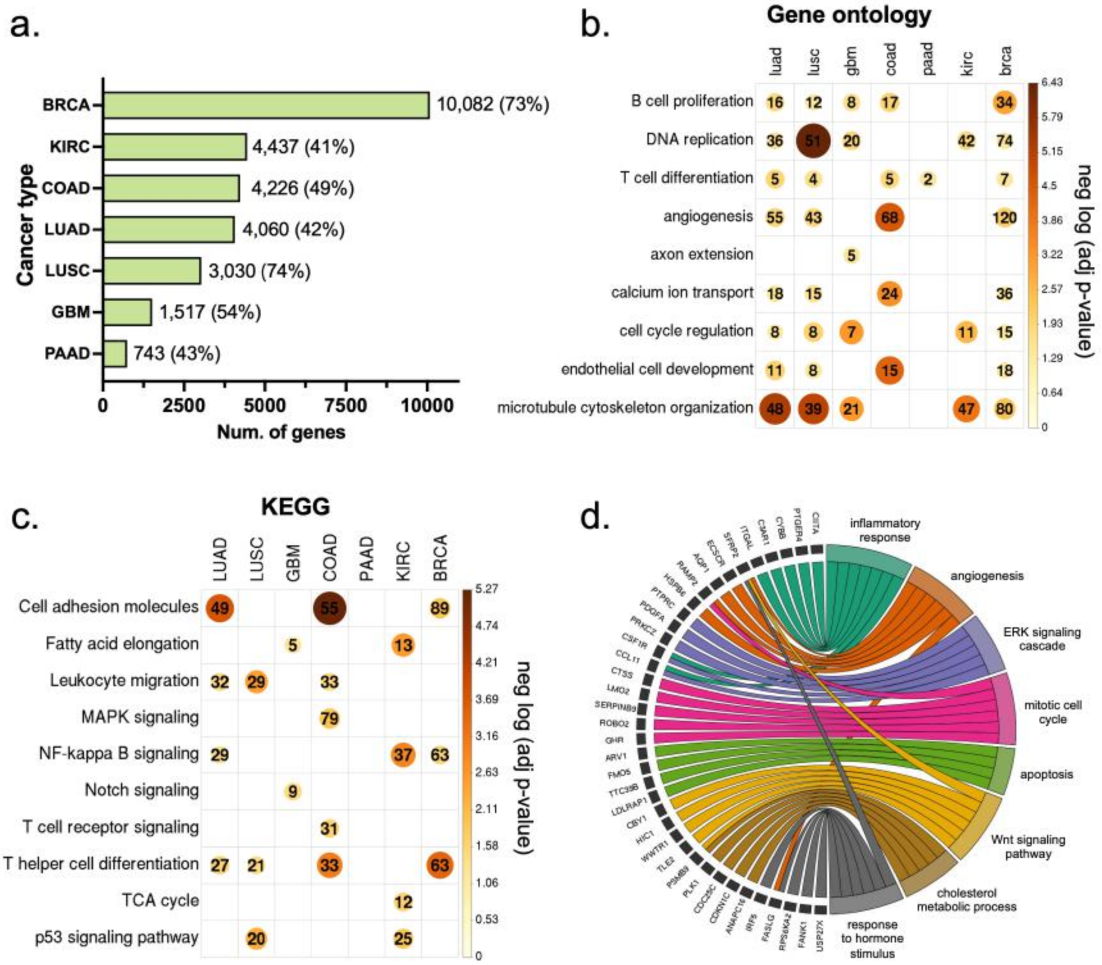


Fig. 3: **Characterization of the genes validated in external cancer cohorts**. a) The number of genes validated in the CPTAC cohort from each cancer type. Percentages enclosed within the parenthesis indicate the proportion of significant genes obtained from the TCGA cohort. b) Heatmap showing the significant P values from the gene ontology analysis of the validated genes. Color and size of the circles represent the negative log-transformed P values. Integers represent the absolute gene count in each category, and non-significant categories are left in blank. c) Heatmap showing the significant P values from the KEGG analysis of the validated genes. Color and size of the circles represent the negative log-transformed P values. Integers represent the absolute gene count in each category, and non-significant categories are left in blank. d) Circos plot showing the enriched biological processes associated with the validated genes in lung adenocarcinoma.P values were adjusted for multiple testing using the Benjamini–Hochberg method.

A digital signature for breast cancer recurrence prediction

Given that *SEQUOIA* was able to predict the transcriptional activity of genes involved in key cancer-related pathways (Figures 2 and 3), we next assessed whether these genes have prognostic value. We focused our analysis on breast cancer, for which the highest number of genes ($N = 13{,}798$) was accurately predicted.

The genes accurately predicted from *SEQUOIA* encompass various published prognostic signatures (Supplementary Data 5)[34–37]. These include 47 out of 50 (94%) genes of the PAM50 signature, all 12 (100%) genes of the EndoPredict signature, 13 out of 21 (62%) genes of the Oncotype DX signature, 47 out of 70 (67%) genes of the Mammaprint signature, 6 out of 7 (86%) genes of the Breast Cancer Index and 4 out of 5 (80%) gene of the Mammostrat signature. These results highlight the capacity of *SEQUOIA* in predicting prognosis-associated gene expression in breast cancer.

Next, we sought to develop a multi-gene signature that can stratify the risk of breast cancer recurrence, leveraging the accurately predicted genes from $SEQUOIA$. We fitted a regularized Cox regression model on the ground-truth gene expression values, where the model aims at predicting a risk score of recurrence (”Methods”). High risk scores indicate a greater likelihood of recurrence. The model was developed on the TCGA cohort (N = 858 patients) and further validated using data from two independent cohorts: (1) ”SCANB” (N = 5,034 patients) [38] and (2) ”METABRIC” (N = 2,262 patients) [39].

Our analysis led to the identification of a 50-gene signature significantly associated with recurrence (Figures 4a-c and Supplementary Data 6). To assess its performance, we first treated the predicted risk score as a continuous variable. Results from univariate Cox regression analyses (Figures 4a-b and Supplementary Figure A3f) showed that the predicted risk scores were significantly associated with recurrence-free survival: TCGA (HR = 2.24, .95CI = 1.99-2.52, $P < 2e - 16$), SCANB (HR = 1.31, .95CI =1.25-1.39, $P < 2e - 16$), and METABRIC (HR = 1.21, .95CI = 1.08-1.36, $P = 7.8e - 07$). To further assess the model, we treated the predicted risk score as a dichotomous variable. Patients within each cohort were divided into a high-risk and a low-risk group based on the median risk score (Figures 4a-b). Results from the log-rank test demonstrate that the high-risk group had significantly worse prognosis compared to the low risk group: TCGA ($P < 2e - 16$), SCANB ($P = 4e - 13$), and METABRIC ($P = 8e - 05$).

To assess whether breast cancer subtype was a confounding variable in risk prediction, we incorporated the PAM50 molecular subtypes and hormone (estrogen and progesterone) receptor status as covariates into our Cox regression analyses. We found that the predicted risk score was still significantly associated with prognosis after including these covariates: TCGA (HR = 2.22, .95CI =1.96-2.52, $P < 2e - 16$), SCANB (HR = 1.26, .95CI =1.20-1.33, $P < 2e - 16$), METABRIC (HR = 1.22, .95CI = 1.08-1.39, $P = 7.6e - 05$).

Gene ontology analysis (Figure 4c) revealed the regulatory functions of the signature genes in cell apoptosis (*MSX1*, *PTPN2*, *BCL2A1*), cell-cycle phase transition (*NEK10*, *CDK14*, *RAD51C*, *PLK2*), inflammatory response (*IL21*, *PTGS2*, *PPP2R3C*), cytokine production (*CLEC9A*, *CLEC6A*), cellular metabolic process (*PDE8B*, *ADA2*, *NEU3*, *SUMO2*, *ACAT2*), ERK signaling cascade (*THPO*, *PTPN2*), cell-fate specification (*POU5F1*, *SIX2*), cell membrane transportation (*CACNA1E*, *PTGS2*, *CACNG4*, *PLK2*), and cell migration (*MDGA1*, *HTR6*).

So far, we have developed and validated a 50-gene signature using the ground-truth gene expression values. We then tested whether utilizing the gene expression values from histology images alone was sufficient to stratify the risk groups. For each patient, we calculated a risk score using the same risk coefficient from our signature model, but this time replacing the ground-truth gene expression values with the predicted values. As shown in Figure 4d, patients assigned with high risk scores demonstrated significantly worse prognosis compared to patients with low risk scores (Cox regression: HR = 1.58, .95Cl = 1.16-2.15, P = 0.006; Log-rank test: P = 0.008). These results indicate that $SEQUOIA$ can accurately predict the expression of genes associated with breast cancer recurrence.

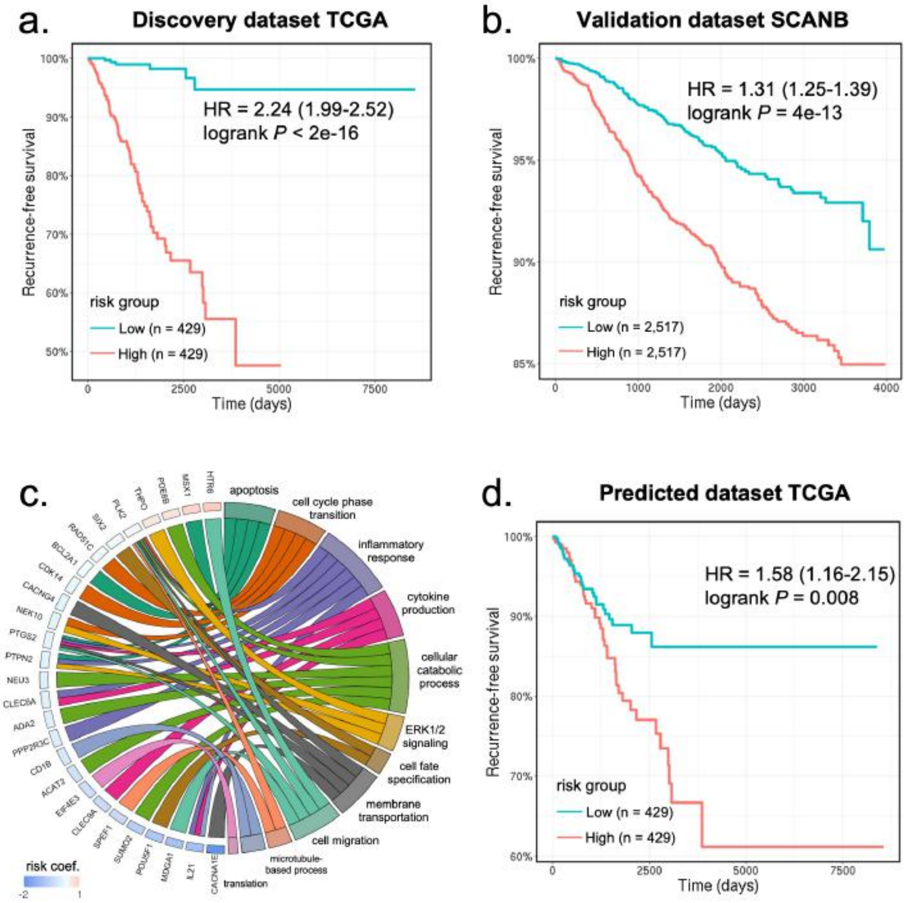


Fig. 4: **Development and validation of a digital signature for predicting breast cancer recurrence**. a) Kaplan-Meier curves of recurrence-free survival obtained from the TCGA discovery dataset. Patients were split by the median risk score. b) Kaplan-Meier curves of recurrence-free survival in the SCANB validation dataset. c) Circos plot showing the biological processes associated with the prognostic gene signature. Gene names and the associated risk coefficients are shown on the left and the corresponding biological processes are shown on the right. d) Kaplan-Meier curves of recurrence-free survival obtained from the predicted gene expression values in the TCGA dataset. Patients were split by the median risk score. HR: hazard ratio.

Tile-level predictions validated with spatial transcriptomics

So far, we have demonstrated the ability of *SEQUOIA* to accurately predict RNA-Seq gene expression values at the tissue (i.e., bulk) level. However, gene expression patterns are known to vary across different tumor regions due to intra-tumoral heterogeneity of cell-type compositions. Uncovering spatial gene expression patterns can reveal the intricate landscape of tumor architecture and the microenvironment, which is known to affect tumor growth, metabolic processes, and resistance to therapy [40, 41]. We hence investigated whether our models trained at the slide-level can be used to predict gene expression values at the region level within tumor tissues.

Here, we implemented a sliding window method to generate tile-level predictions of gene expression ("Methods"). To validate the prediction, we utilized a cohort of eighteen glioblastoma (GBM) patients, which contains matched histology images and spatial transcriptomics data, providing tile-level ground truth gene expression measurements [22] (Figure 5a). We focused our analysis on the top 500 genes for which *SEQUOIA* generated the best predictions on the TCGA test set (i.e. genes with the highest Pearson correlation coefficients). For each of these genes, we generated a spatial heatmap illustrating their expression values across the slide. To quantitatively assess the prediction performance, we used

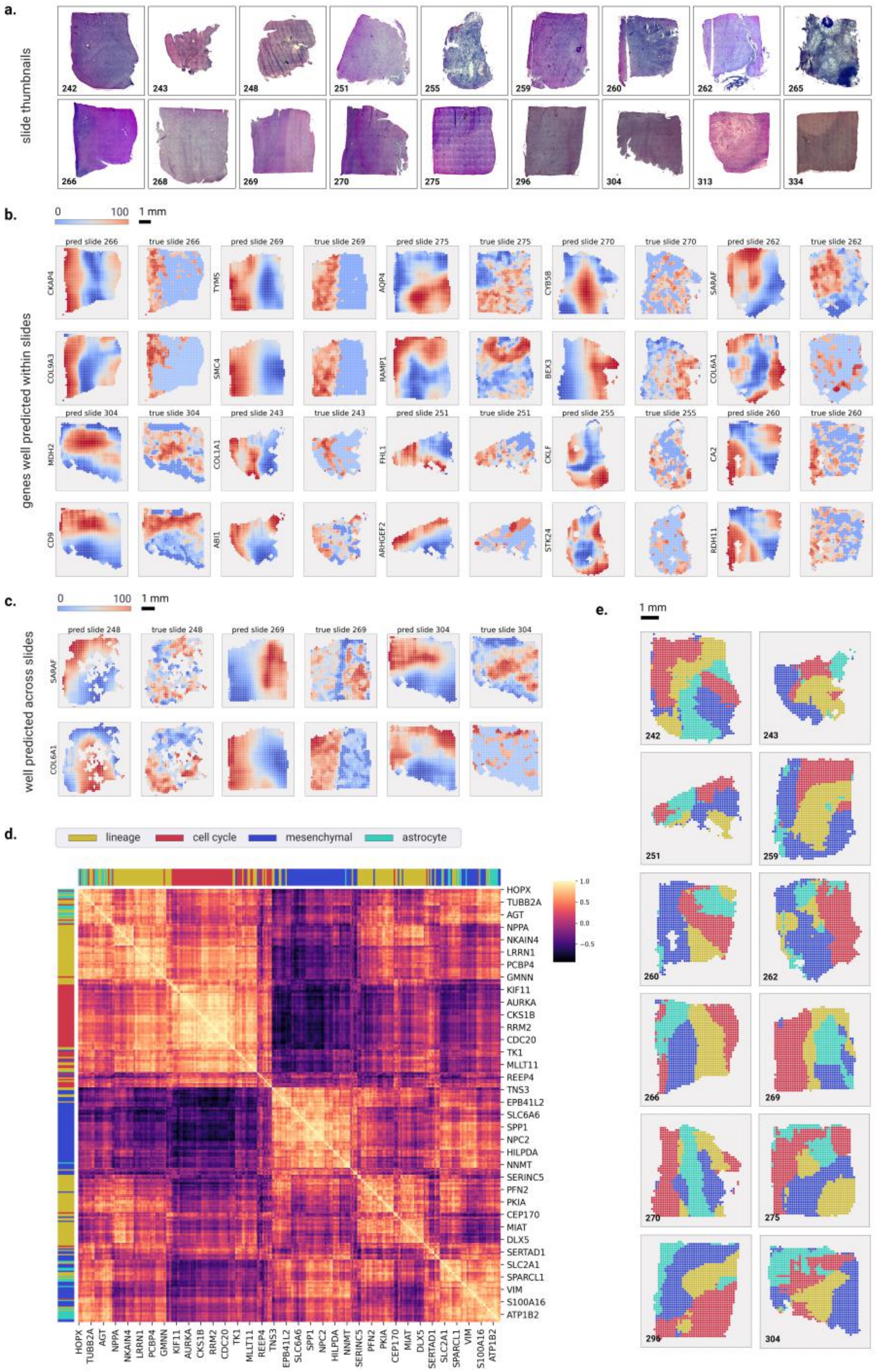

a.
slide thumbnails
242
243
248
251
255
259
260
262
265
266
268
269
270
275
296
304
313
334
b.
0
100
1 mm
genes well predicted within slides
pred slide 266
true slide 266
pred slide 269
true slide 269
pred slide 275
true slide 275
pred slide 270
true slide 270
pred slide 262
true slide 262
CKAP4
TYMS
AQP4
CYB5B
SARAF
COL9A3
SMC4
RAMP1
BEX3
COL6A1
pred slide 304
true slide 304
pred slide 243
true slide 243
pred slide 251
true slide 251
pred slide 255
true slide 255
pred slide 260
true slide 260
MDH2
COL1A1
FHL1
CKLF
CA2
CD9
ABI1
ARHGEF2
STK24
RDH11
c.
0
100
1 mm
well predicted across slides
pred slide 248
true slide 248
pred slide 269
true slide 269
pred slide 304
true slide 304
SARAF
COL6A1
d.
lineage
cell cycle
mesenchymal
astrocyte
1.0
0.5
0.0
−0.5
HOPX
TUBB2A
AGT
NPPA
NKAIN4
LRRN1
PCBP4
GMNN
KIF11
AURKA
CKS1B
RRM2
CDC20
TK1
MLLT11
REEP4
TNS3
EPB41L2
SLC6A6
SPP1
NPC2
HILPDA
NNMT
SERINC5
PFN2
PKIA
CEP170
MIAT
DLX5
SERTAD1
SLC2A1
SPARCL1
VIM
S100A16
ATP1B2
e.
1 mm
242
243
251
259
260
262
266
269
270
275
296
304

Fig. 5: **Spatial visualization of gene expression predicated at the tile level**. a) Whole Slide Image thumbnails from the validation cohort. b) Examples of genes that are well-predicted spatially within slides, with predicted spatial gene expression on the left and ground truth on the right. The prediction and ground truth maps were normalized to percentile scores between 0-100. c) Examples of genes that are spatially well-predicted across several slides. Each row shows the prediction map (on the left) and ground truth (on the right) for a particular gene across four slides. d) Heatmap showing the correlation coefficients of meta-gene modules that define the transcriptional subtype and proliferation state of GBM cells. e) Spatial organization of the predicted transcriptional subtypes within different slides. Transcriptional subtypes were assigned based on the meta-gene module showing the highest prediction values

the Earth Mover's Distance (EMD) as an evaluation metric ("Methods"). EMD values are bounded between 0 and 1, with lower values indicating a closer correspondence between predictions and ground truth. On average, *SEQUOIA* achieved an EMD of 0.15 (.95CI = 0.148-0.152) across all slides and genes (Supplementary Table A7). Higher performance was observed in slides with high degrees of spatial variance in gene expression [22, 41].

Notably, *SEQUOIA* generated accurate spatial predictions for genes that hold significant relevance to GBM malignancy and prognosis. For instance, the *COL6A1* and *COL9A3* genes are highly expressed in the mesenchymal subtype of GBM, a subtype associated with unfavorable prognoses [42, 43]. In our spatial predictions, we observed a median EMD of 0.11 across all slides for both *COL6A1* and *COL9A3* (Figures 5b-c). Furthermore, the *CKAP4* gene, with and EMD of 0.12 has been shown to mediate the growth, migration, and invasion of GBM cells [44]. These results highlight the potential of *SEQUOIA* in accurately predicting spatial gene expression patterns related to GBM malignancy and prognosis.

The utilization of single-cell RNA-seq and spatial transcriptomics assays in recent studies has revealed that cells sharing the same transcriptional subtype are often co-localized within spatially segmented niches [22, 45]. To investigate whether *SEQUOIA* captured true biological signals that reflect underlying tissue compositions, we assessed spatial co-expression patterns of functionally related genes. We considered four previously established meta-gene modules governing the transcriptional subtype and proliferation state of GBM cells: (1) 'lineage development' (124 genes), (2) 'cell cycle' (70 genes), (3) 'mesenchymal-like' (92 genes) and (4) 'astrocyte-like' (37 genes) [46]. Spatial correlation analyses showed that genes within the same meta module consistently clustered together, exhibiting similar spatial expression patterns (Figures 5d-e). To demonstrate the spatial prediction capacity of *SEQUOIA* in other cancer types, we developed an user-friendly, interactive web application (https://sequoia.stanford.edu) where users can explore the spatial heatmap for genes predicted in the TCGA cohorts. These results demonstrate the potential of *SEQUOIA* in resolving spatial cellular architectures within heterogeneous tumor tissues.

Discussion

Transcriptomic analysis of tumor tissues holds immense promise in advancing personalized diagnosis and outcome predictions. In this study, we presented *SEQUOIA*, a deep learning model for predicting RNA-seq gene expression data from Whole Slide Images. We combined algorithmic and methodological advancements, followed by thorough analyses of gene functions, clinical relevance, and generalizability. Through a comprehensive evaluation of our model in nine cancer types across seven tissues, we demonstrated the value of *SEQUOIA* in predicting clinically relevant gene expression patterns.

Over the past decade, deep learning has revolutionized cancer diagnosis. Published studies have demonstrated the potential of deep neural networks in extracting intricate patterns from medical images. He et al. developed ST-Net, a convolutional neural network that predicts the expression values of 250 genes from histology images in breast cancer [47]. Their model however is trained on individual tiles, which does not integrate contextual information across tiles and requires high-resolution training labels obtained from spatial transcriptomic assays. To model contextual information, Graziani et al. incorporate an attention mechanism into their model for gene expression prediction. However, this strategy requires training a dedicated model for predicting the expression of each individual gene [48]. While this approach reaches publishable performance, it can lead to computational challenges, particularly when attempting to infer the entire transcriptome. A recent study by Alsaafin et al. utilized transformer modules to extract latent representation of WSIs. Their model was tested in renal cell carcinoma for gene expression prediction and subtype classification tasks [3].

To demonstrate the advantages of our model, we compared it with $HE2RNA$ [26], a recent model for whole transcriptome prediction from WSIs. The results of our analysis revealed consistent improvements across various cancer types when using $SEQUOIA$ in comparison to $HE2RNA$. A key factor driving this performance boost lies in the attention-based mechanism leveraged by $SEQUOIA$, which enables effective integration of information between tiles. In contrast, $HE2RNA$ treated each tile as an independent entity, limiting its ability to capture the contextual relationships present in the data. However, the increased complexity of the transformer-based model can also lead to overfitting, particularly when confronted with limited training data. To address this challenge, we pre-trained the weight parameters of the model using data from normal tissues. We found that the pre-training regimen improved the performance, especially in case of small training datasets.

The genes accurately predicted by $SEQUOIA$ were associated with key pathways pertinent to cancer. Among these were genes involved in regulating cell cycles, inflammation, angiogenesis, and hypoxia response. Additionally, the model effectively captured cell-type markers, including those for endothelial cells, CD4 T cells, M2 macrophages and B cells. Building upon the well predicted genes, we developed a 50-gene signature that predicts the risk of breast cancer recurrence. Although the gene expression signature was developed on ground-truth gene expression values, we demonstrated its utility in patient stratification by just using the predicted gene expression. Despite the decreasing costs for transcriptomics sequencing, many hospitals still lack the necessary equipment and trained personnel to conduct a comprehensive analyses. However, by harnessing $SEQUOIA$'s predictions, it becomes possible to swiftly examine transcriptomics profiles from routinely obtained whole slide images, thereby streamlining the diagnostic process and significantly cutting down on expenses.

While $SEQUOIA$ was trained using bulk RNA gene expression, we demonstrated its potential in predicting gene expression patterns at a local level. We evaluated the accuracy of the predicated high-resolution spatial maps on an independent cohort of eighteen patients with GBM. We showed a number of genes for which $SEQUOIA$ was able to generate accurate spatial maps, for genes with good visualization within as well as across slides. Several of these genes have been identified as prognostic for GBM, related to aggressive phenotype and/or as potential therapeutic targets. Such high-resolution spatial prediction of gene expression on WSIs can bring significant value to both clinical and research settings. In the clinic, it can aid in identifying specific regions within a heterogeneous tumor that require sequencing (hence ensuring the accurate detection of biomarkers and preventing the omission of critical lesions [16]). In research, this approach enables the cost-efficient exploration of gene expression dynamics at high resolution which allows to generate hypotheses about spatial co-occurrences and interactions between genes, thereby advancing our understanding of the complex mechanisms underlying cancer progression.

In the future, the accurate prediction of gene expression from whole slide images holds immense potential for enhancing diagnosis and prognosis for cancer. Furthermore, the predicted gene expression and associated pathways can provide valuable insights into a tumor's aggressiveness and its molecular characteristics, thereby enabling personalized and targeted therapies. Once clinically validated and further improved (e.g. by training on larger cohorts and spatial transcriptomics cohorts), the implementation of such predictive models has the potential to streamline medical processes, save costs and improve efficiency by rapidly identifying actionable information from image-based data.

In conclusion, by combining algorithmic advancements with thorough analyses of gene prediction, clinical relevance, survival prediction, and generalizability, our research offers a comprehensive understanding of the potential applications of gene expression prediction from WSIs.

Methods

Patient cohorts and ethics

TCGA

For model training, anonymized patient data were retrieved from the publicly available 'The Cancer Genome Atlas' (TCGA) archive (available at https://portal.gdc.cancer. gov). We used paraffin-embedded (FFPE) whole slide images (WSIs) and matched gene expression data of nine cancer types, including prostate adenocarcinoma (PRAD), pancreatic adenocarcinoma (PAAD), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), colon adenocarcinoma (COAD), and breast adenocarcinoma (BRCA). The number of patients, WSIs and genes used for training each cancer type is listed in Supplementary Table A1.

CPTAC

For validation, we used the publicly available patient data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohort (https://portal.gdc.cancer. gov). We retrieved matched WSIs and gene expression data from seven cancer types, including breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LSCC/LUSC), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (CCRCC/KIRC), glioblastoma multiforme (GBM), pancreatic adenocarcinoma (PDA/PAAD). The sample size is described in Supplementary Table A2.

Tempus

For an additional validation, we utilized matched WSIs and RNA-seq data ($N = 287$ slides from $N = 249$ patients) of lung adenocarcinoma (LUAD). The data were obtained through a data transfer agreement with Tempus Labs, Inc.

GTex

For pre-training the models, we used WSIs and gene expression data from six normal tissues (i.e., brain, colon, kidney, lung, pancreas, prostate). Data were obtained from The Genotype-Tissue Expression (GTEx) project (https://gtexportal.org), and the sample size is described in Supplementary Table A3

Spatial GBM, SCANB, METABRIC

Spatial transcriptomic data and matched histology images of GBM were obtained from a published study by Ravi et al. (https://datadryad.org/stash/dataset/doi:10.5061/dryad.h70rxwdmj) [22]. Data of the SCANB and METABRIC breast cancer cohorts were obtained from published studies by Staaf et al.[38] and Curties et al.[39].

Preprocessing of RNA-Seq data

For training and validation of our models, we used FPKM-UQ normalized gene expression values. Since the gene expression values span several orders of magnitude and our model was trained using the Mean Squared Error loss function, the training process may introduce bias to genes with large gene expression values. To overcome this potential bias, we performed log2 transformation ($v \to log_2(v + 1)$) of the gene expression values.

For pre-training the weight of the models, we obtained the RNA-seq data of normal tissues from the GTEx data portal (https://gtexportal.org/home/datasets). We performed the same log2 transformation ($v \to log_2(v + 1)$) of the gene expression values. Since during the pre-training phase, we combined data from all tissue types, we performed a z-score normalization of the gene expression values in each individual tissue type, and the normalized gene expression matrices were concatenated across the tissue types.

We focused our analysis on three gene categories: (1) protein-coding genes, (2) micro-RNAs (miRNAs) and (3) long non-coding RNAs (lncRNAs). On average, the protein-coding genes account for 85% of all the analyzed genes.

Preprocessing of Whole Slide Images

Whole-slide images (WSIs) were acquired in SVS format and downsampled to 20× magnification ($0.5\mu m$ px^{-1}). We used the Otsu threshold method to obtain a mask of the tissue, which allows to omit tiles mostly containing white background [49]. WSIs have much larger dimensions than natural images (usually over $10k \times 10k$ pixels), and therefore cannot be used directly to train machine learning models. Thus, as commonly done in WSI analysis [12, 13], we randomly sampled non-overlapping tiles to train models. We chose a maximum of $N = 4000$ random non-overlapping tiles of 256×256 pixels (at $0.5\mu m$ px^{-1}), omitting those containing more than 20% background and tiles with low contrast.

To obtain a representation at slide-level, we decided to follow the super-tile methodology proposed by Schmauch et al. [26]. First, we used a pre-trained ResNet-50 (pre-trained on ImageNet) to obtain a feature representation of each tile. Then, we used the k-means algorithm to cluster similar tiles into $K = 100$ clusters per slide. Each cluster contains tiles with similar morphological features, where cluster A may represent tiles that mostly contain tumor cells, cluster B may contain tiles with mostly connective tissue and so on. Finally, the corresponding 100 cluster means are obtained which each represent a super-tile. This leaves a matrix of 100 × 2048 super-tile feature vectors which represent the slide.

SEQUOIA architecture

SEQUOIA is inspired by the vanilla Vision Transformer (ViT) architecture [50] which extrapolates the Transformer architecture from the natural language processing (NLP) domain to computer vision [51]. For a ViT, an image is divided in patches of 16 × 16 px, representing “tokens” of the image. Feature vectors are then extracted from these patches (by linear projection in the ViT). Then, they are fed to a transformer encoder, which outputs a new representation of the input and forwards it to a multi-layer perceptron (MLP) head that makes the final prediction.

In our work, the division of an image into small patches from the original ViT corresponds to the WSI being divided into super-tiles. The feature vectors of the super-tiles used as input to the transformer encoder are the 100 cluster means of dimension 2048 described above (‘Preprocessing of Whole Slide Images’). Importantly, the transformer encoder allows to model relationships across super-tiles before deciding whether they are relevant for the slide-level prediction.

Specifically, our model takes as input a 100×2048 feature matrix (for the 100 super-tile, i.e. cluster mean feature vectors). This is fed to the transformer encoder which contains 6 encoder blocks, 16 attention heads, and a head dimension of 64. After layer normalization, the output is sent to an MLP layer with dimension $2048 \times num\_genes$, with num_genes the number of genes available to predict in each cancer type (see Supplementary Table A1).

Pre-training on normal tissue

We performed pre-training of the *SEQUOIA* and *HE2RNA* models using normal tissues corresponding to the tested cancer types. We combined all normal tissues for the pre-training step, and the resulting model was afterwards finetuned for each specific cancer type.

Notably, during our preliminary experiments, we observed that incorporating breast normal tissue into pre-training led to an overall decrease in performance compared to the model pre-trained on all other normal tissues without the breast. This decrease can likely be attributed to the known differences in tissue composition between normal breast tissues and breast cancer. While the normal breast mainly comprises adipocytes, the tumors consist of transformed epithelial cells. Consequently, we excluded the normal breast tissue from the pre-training process, resulting in the following used tissues for pre-training: lung, brain, kidney, pancreas, prostate, colon.

The model was trained using the Mean Squared Error loss function for 200 epochs with early stopping (early stop if the loss did not decrease for a *patience* of 100 epochs), and batch size 16. Model parameters were optimized with the Adam optimizer with learning rate 3×10^{-3}. For training the model, we considered 19,198 genes for which gene expression levels are available across all normal tissues (Supplementary Table A3).

Training details

After pre-training the model on normal tissues, we finetuned a dedicated model for the nine cancer types we considered from TCGA. Hereto, the transformer encoder was initialized with the weights from the pre-trained model and a new prediction head (a layer norm and a linear layer) was trained (see Figure 1).

Regarding splitting the TCGA data for model training (for each cancer type), we used the same approach as Schmauch et al. [26] which is a five fold cross-validation. Specifically, five folds are made each of which consist of a ‘global’ train and test set (samples from the same patients always restricted to the same set). In each fold i, the ‘global’ train set is further split into a train (90%) and validation (10%)

set. The validation set i is used to determine the optimal point to stop training model i, which is then evaluated on test set i. Afterwards, predictions on patients from test sets i ($i = 1..5$) are concatenated before calculating performance measures (e.g. Pearson correlation between predicted gene expression and ground truth expression across patients). The reason for concatenating the patients across these test sets (instead of the conventional approach where one held-out test set is chosen and the rest split for cross-validation) is because each test set i is too small to determine statistical significance of correlation across patients within that set.

We used the Mean Squared Error (MSE) as loss function for training the model and we trained each model for a maximum of 200 epochs. Instead of only considering the Pearson correlation coefficient as performance metric for determining the optimal point for stopping training and saving model checkpoints (as in Schmauch et al. [26]), we considered a criterion that takes into account both MSE and correlation. Namely, while the MSE decreases, we continue training and saving model weights on optimal MSE. Once the MSE stops improving, we continue training if in the last *patience* epochs, correlation has improved and if there has been a reasonable MSE (i.e. $MSE < \delta + bestMSE,\ with\ \delta = 0.5$). We then save model weights in an epoch if correlation has improved (i.e. $corr > best\_corr$).

We used a fixed learning rate of 1×10^{-3} and batch size of 16 for model training. The model parameters were optimized with the Adam optimizer.

Identification of well-predicted genes

To identify a significantly well-predicted gene, we used Pearson's correlation analysis to compare the ground truth gene expression values versus the predicted values. The resulting correlation coefficient and p-value was further compared to those obtained with a random, untrained model with the same architecture. We combined three criteria to select significant genes: (1) The correlation coefficient (r_1) between the ground truth and the predicted gene expression values across the validation cohorts must be positive ($r_1 > 0$) and the P value (p_1) should be smaller than 0.05 ($p_1 < 0.05$); (2) The correlation coefficient r_1 must be greater than r_2 ($r_1 > r_2$), where r_2 represents the coefficient between the ground truth and predicted gene expression values obtained from the random model; (3) r_1 must be significantly higher than r_2 as determined by the Steiger's Z test. We required the raw P value to be smaller than 0.05 and the adjusted P value by Benjamini-Hochberg correction to be smaller than 0.2.

Gene set analysis

The gene set analysis was performed with the ClusterProfiler R library (version 4.2.1) [52] and GSEApy package (version 1.0.5) [53]. Biological processes from gene ontology and cell-type signatures were obtained from the MSigDB database (https://www.gsea-msigdb.org/gsea). KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway annotations were obtained from the KEGG database (https://www.genome.jp/kegg/catalog/org_list.html). The enrichment analysis was performed with hyper-geometric testing, and the P values were corrected with the Benjamini-Hochberg procedure. To generate heatmaps of the P values, we aggregated gene sets with high similarities (e.g., "regulation of T cell proliferation" and "positive regulation of T cell proliferation"), and the average P values were shown.

Identification and validation of the prognostic gene signature

To construct a gene expression model for predicting breast cancer recurrence, we first selected the top 5,000 well-predicted protein-coding genes from the TCGA-BRCA cohort as potential candidates. Then, we performed LASSO Cox regression model analysis with the 'glmnet' R package (version 4.1)[54]. The penalized Cox regression model with LASSO penalty was used to achieve shrinkage and variable selection simultaneously. The optimal value of the penalty parameter λ was determined through a five-fold cross-validation.

Utilizing the optimal λ value, we curated a list of prognostic genes, each associated with a coefficient (i.e., hazard ratio) that was not equal to zero. The risk score was derived by performing a linear combination of the expression levels of the selected genes, with each expression level being weighted by its associated coefficient, as described by the equation 1:

$$risk\ score = \sum_{i=1}^{n} C_i \times Exp_i \tag{1}$$

where C_i represents the coefficient of a gene and Exp_i its expression value.

The patients in each dataset were split into a low-risk and a high-risk group according to the median risk score. Finally, the Kaplan–Meier estimator and the log-rank test were performed to assess the difference in recurrence-free survival between the low-risk and high-risk groups.

Spatial visualization of predicted gene expression on tile level

To visualize predicted gene expression spatially on tile-level, we implemented a sliding-window method. Starting from the left upper corner of the WSI, consider a window of 10 × 10 tiles. For clarity, we refer to the location of the window by the (x, y) coordinate of the left upper tile in the window. Hence, the window in the left upper corner has coordinate $(0, 0)$ and is referred to as $w_{0,0}$ (x, y axes defined as in image processing, origin in left upper corner and x-axis increases when moving to the right, y-axis increasing when moving below).

The 100 × 2048 feature vectors of tiles in the window $w_{x,y}$ are fed to the model (each individual tile feature vector serves as a 'cluster mean vector'). The resulting predicted gene expression $g_{w_{x,y}}$ is saved for all tiles in the window. Then, the window is moved *stride* number of tiles to the right ($w_{x+stride,y}$), and the predicted gene expression is again saved for each tile in the window. When the window has reached the end of a row ($x + stride + 10$ equals the width of the image), a new window is started at position *stride* below the previous row ($w_{0,y+stride}$). After the window has passed the entire WSI, the prediction for each tile is calculated as the average of all values that were saved for that tile when it was part of a window $w_{x,y}$. In our implementation, we chose $stride = 1$ (larger strides require less compute time but are less fine-grained).

For comparison of the predicted spatial gene expression with the spatial transcriptomics measurement in the ground truth, we resampled the ground truth resolution to match the predicted resolution. Namely, the ground truth resolution was $55\mu m$ per spot which is higher than the predicted resolution of $256\mu m$ per spot. Hence, we compared each spot in the prediction with the average of the four nearest spots in the ground truth (nearest in terms of smallest Euclidean distance between the x,y coordinates of the spots). We also performed median filtering on the ground truth map to remove noise (window size 3×3) and we only considered genes with $>= 10$ unique measured values in the spatial ground truth map (to avoid incorporating noisy measurements). Finally, we converted both the predicted and ground truth values to normalized percentile scores between 0-100.

Earth Mover's Distance

For a quantitative evaluation of the spatial visualization capabilities of the model, we used the two dimensional Earth Mover's Distance (EMD) (implemented with the $cv2.EMD$ function from opencv-python [55]). Intuitively, the metric captures the minimum amount of 'work' required to transform one distribution into the other. Often the two distributions are informally described as different ways of piling up earth/dirt, and the 'work' to transform one distribution into another is defined as the amount of dirt multiplied by the distance (Euclidean distance in our case) over which it is moved. Hence, this metric takes into account the *spatial context* to determine how well the prediction map corresponds to the ground truth. This is in contrast to pixel-level metrics which only take into account how correct a certain pixel is irrelevant of its 2D location and context (e.g. calculating for each pixel the Mean Squared Error between prediction and ground truth).

Spatial correlation analysis of GBM signature genes

To assess whether genes exhibiting similar spatial expression patterns are functionally related, we used four recurrent meta-gene modules governing the transcriptional subtype and proliferation state of GBM cells as discovered from a published single-cell RNA-seq study [46]. We included all signature genes from these modules, except for those ($N = 18$ genes) not included in our training process. The neural-progenitor-like (NPC-like) and oligodendrocyte-progenitor-like (OPC-like) modules were combined into one group, namely 'lineage development', which includes a total of 124 genes. Further, gene modules regulating G1/S and G2/M phase transitions ($N = 70$ gens) were combined into a 'cell-cycle' module. Finally, the 'mesenchymal-like' ($N = 92$ genes) and 'astrocyte-like' ($N = 39$ genes) modules were included as separate groups.

To assess spatial co-expression patterns, we determined the similarity of spatial prediction maps for each pairwise combination of genes ($N = 325$ genes in total). This was accomplished by first flattening the tile-level predictions into two 1D arrays and then computing the Pearson correlation between them. This process was repeated in each slide, and the resulting correlation matrices were averaged across all eighteen slides. The spatial correlation matrix was clustered using hierarchical clustering to reveal genes that exhibit similar spatial expression patterns. We further assigned a color to each row and column in the matrix indicating the meta-module each gene belongs to.

Code availability

Codes for data pre-processing, model training and evaluation were deposited into a public GitHub repository (https://github.com/gevaertlab/sequoia-pub).

Data availability

Anonymized WSIs, gene expression and clinical data of the The Cancer Genome Atlas' (TCGA) cohorts were retrieved from the publicly available Genomic Data Commons (GDC) portal (https://portal.gdc.cancer. gov). Gene expression data of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohort were downloaded from GDC portal (https://portal.gdc.cancer. gov) and WSIs were obtained from the Cancer Image Archive with the accession URL (https://www.cancerimagingarchive.net/collections). Gene expression data and WSIs of the Tempus cohort were obtained through a data transfer agreement with Tempus Labs, Inc. Publicly available gene expression data and WSIs of The Genotype-Tissue Expression (GTEx) project were retrieved with the accession URL (https://gtexportal.org). The publicly available spatial transcriptomics data of GBM were acquired from Datadryad using the following accession URL (https://doi.org/10.5061/dryad.h70rxwdmj) [22]. The RNA-seq data and clinical annotations of the SCANB cohort were obtained from the accession URL (https://data.mendeley.com/datasets/yzxtxn4nmd/3), and data of the METABRIC cohort was obtained for cbioportal with accession URL (https://www.cbioportal.org/study/summary?id=brca_metabric).

Acknowledgments

Research reported here was further supported by the National Cancer Institute (NCI) under awards: R01 CA260271. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. In addition, M. Pizurica was supported by a Fellowship of the Belgian American Educational Foundation and a grant from FWO 1161223N. F. Carrillo-Perez was also supported by a predoctoral scholarship from the Fulbright Spanish Commission.

Appendix A Supplementary tables and figures

Table A1: Number of genes, Whole Slide Images and unique number of patients used from each cancer type in TCGA.

TCGA project	**number genes**	**number WSIs**	**number patients**
TCGA-BRCA	25,761	1,053	993
TCGA-LUAD	25,812	536	473
TCGA-LUSC	26,443	510	476
TCGA-GBM	26,530	236	101
TCGA-PRAD	25,587	448	401
TCGA-PAAD	26,172	195	169
TCGA-KIRP	25,141	275	251
TCGA-KIRC	26,653	515	509
TCGA-COAD	23,645	450	442
total		4,218	3,815

Table A2: Number of genes, Whole Slide Images and unique number of patients used from each cancer type in CPTAC.

CPTAC project	**number genes**	**number WSIs**	**number patients**
CPTAC-BRCA	25,761	106	106
CPTAC-CCRCC	26,653	302	211
CPTAC-COAD	23,645	103	103
CPTAC-GBM	26,530	94	94
CPTAC-LUAD	25,812	222	222
CPTAC-LSCC	26,443	109	108
CPTAC-PDA	26,172	146	146
total		1,081	989

Table A3: Number of genes, Whole Slide Images and unique number of patients used from each cancer type in GTex.

GTex project	**number genes**	**number WSIs**	**number patients**
GTex-Breast	19,198	440	440
GTex-Brain	19,198	238	238
GTex-Colon	19,198	405	405
GTex-Kidney	19,198	65	65
GTex-Lung	19,198	530	530
GTex-Pancreas	19,198	325	325
GTex-Prostate	19,198	239	239
total		2,242	861

Table A4: Number of genes significantly well predicted in the TCGA test sets. N shows the total number of slides available for each cancer type.

	he2rna_scratch	he2rna_pretrain	sequoia_scratch	sequoia_pretrain
BRCA (N = 1053)	1,672	4,117	14,915	13,798
LUAD (N = 536)	856	1,646	10,900	9,735
KIRC (N = 515)	115	1,531	11,505	10,922
LUSC (N = 510)	18	127	4,498	4,099
COAD (N = 450)	4,250	174	8,125	8,687
PRAD (N = 448)	1,187	3,060	8,162	9,449
KIRP (N = 275)	5,135	362	7,532	9,413
GBM (N = 236)	1,462	1,963	3,413	2,820
PAAD (N = 195)	670	280	757	1,737

Table A5: Median correlation coefficient between prediction and ground truth in TCGA test set for top 1000 genes within each model. Top genes defined as genes with highest correlation coefficient for each model type.

	he2rna_scratch	he2rna_pretrain	sequoia_scratch	sequoia_pretrain
BRCA	0.143	0.242	0.278	0.251
LUAD	0.107	0.132	0.293	0.252
KIRC	0.090	0.156	0.263	0.266
LUSC	0.089	0.094	0.174	0.172
COAD	0.205	0.101	0.266	0.300
PRAD	0.130	0.272	0.312	0.282
KIRP	0.301	0.134	0.316	0.340
GBM	0.203	0.215	0.266	0.237
PAAD	0.193	0.150	0.187	0.227

Table A6: Number of genes validated in the CPTAC cohort using the $HE2RNA$ model versus the $SEQUOIA$ model.

Cancer	**Abbreviation**	**he2rna_pretrain**	**sequoia_pretrain**
Breast invasive carcinoma	BRCA	18 (0.4%)	10,082 (73%)
Lung adenocarcinoma	LUAD	2 (0.1%)	4,060 (42%)
Lung squamous cell carcinoma	LUSC (LSCC)	0 (0.0%)	3,030 (74%)
Colon adenocarcinoma	COAD	0 (0.0%)	4,226 (49%)
Kidney renal papillary cell carcinoma	KIRC (CCRCC)	3 (0.8%)	4,437 (41%)
Glioblastoma multiforme	GBM	6 (0.3%)	1,517 (54%)
Pancreatic adenocarcinoma	PAAD (PDA)	3 (1.1%)	743 (43%)

Table A7: Median Earth Mover's Distance between prediction and ground truth for top 500 genes from TCGA test set evaluated on different slides in spatial validation cohort.

slide ID	median EMD	slide ID	median EMD
242	0.116	266	0.096
243	0.152	268	0.238
248	0.170	269	0.128
251	0.171	270	0.134
255	0.113	275	0.132
259	0.123	296	0.138
260	0.131	304	0.121
262	0.126	313	0.179
265	0.174	334	0.197

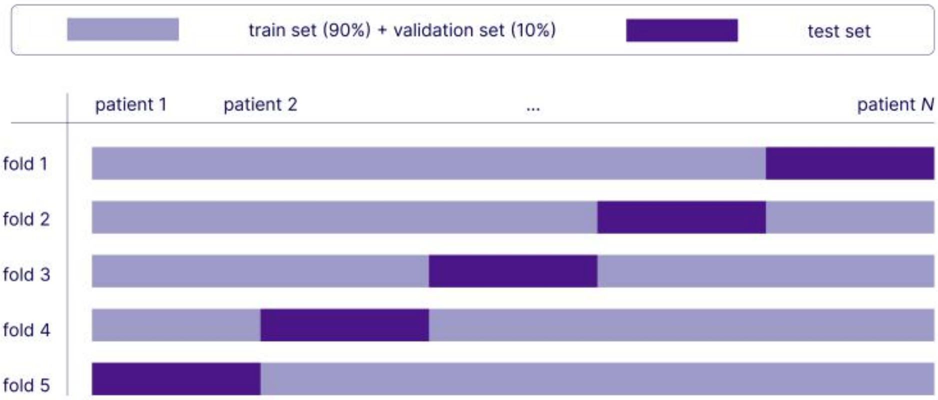
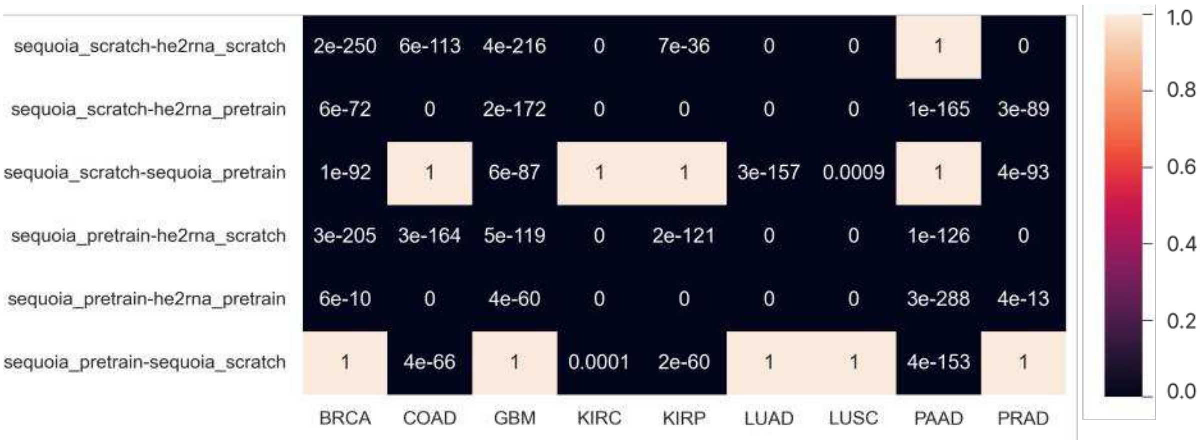


Fig. A1: **Data splitting.** First, five folds are made each of which consist of a 'global' train and test set. The 'global' train set is further split into a train (90%) and validation (10%) set. In each fold i, validation set i is used to determine the optimal point to stop training model i, which is then evaluated on test set i. Afterwards, predictions on patients from test sets i ($i = 1..5$) are concatenated before calculating performance measures (e.g. Pearson correlation between predicted gene expression and ground truth expression across patients).



	BRCA	COAD	GBM	KIRC	KIRP	LUAD	LUSC	PAAD	PRAD
sequoia_scratch-he2rna_scratch	2e-250	6e-113	4e-216	0	7e-36	0	0	1	0
sequoia_scratch-he2rna_pretrain	6e-72	0	2e-172	0	0	0	0	1e-165	3e-89
sequoia_scratch-sequoia_pretrain	1e-92	1	6e-87	1	1	3e-157	0.0009	1	4e-93
sequoia_pretrain-he2rna_scratch	3e-205	3e-164	5e-119	0	2e-121	0	0	1e-126	0
sequoia_pretrain-he2rna_pretrain	6e-10	0	4e-60	0	0	0	0	3e-288	4e-13
sequoia_pretrain-sequoia_scratch	1	4e-66	1	0.0001	2e-60	1	1	4e-153	1

Fig. A2: **Statistical comparison of distributions of correlation coefficient for top 1,000 genes for each model**. For each pairwise model comparison, the P value of the Mann-Whitney U test for testing whether the distribution of the correlation coefficient for model x is larger than for model y, formatted on the left axis as x-y. Mann-Whitney U test calculated with scipy.stats in python with alternative='greater'.

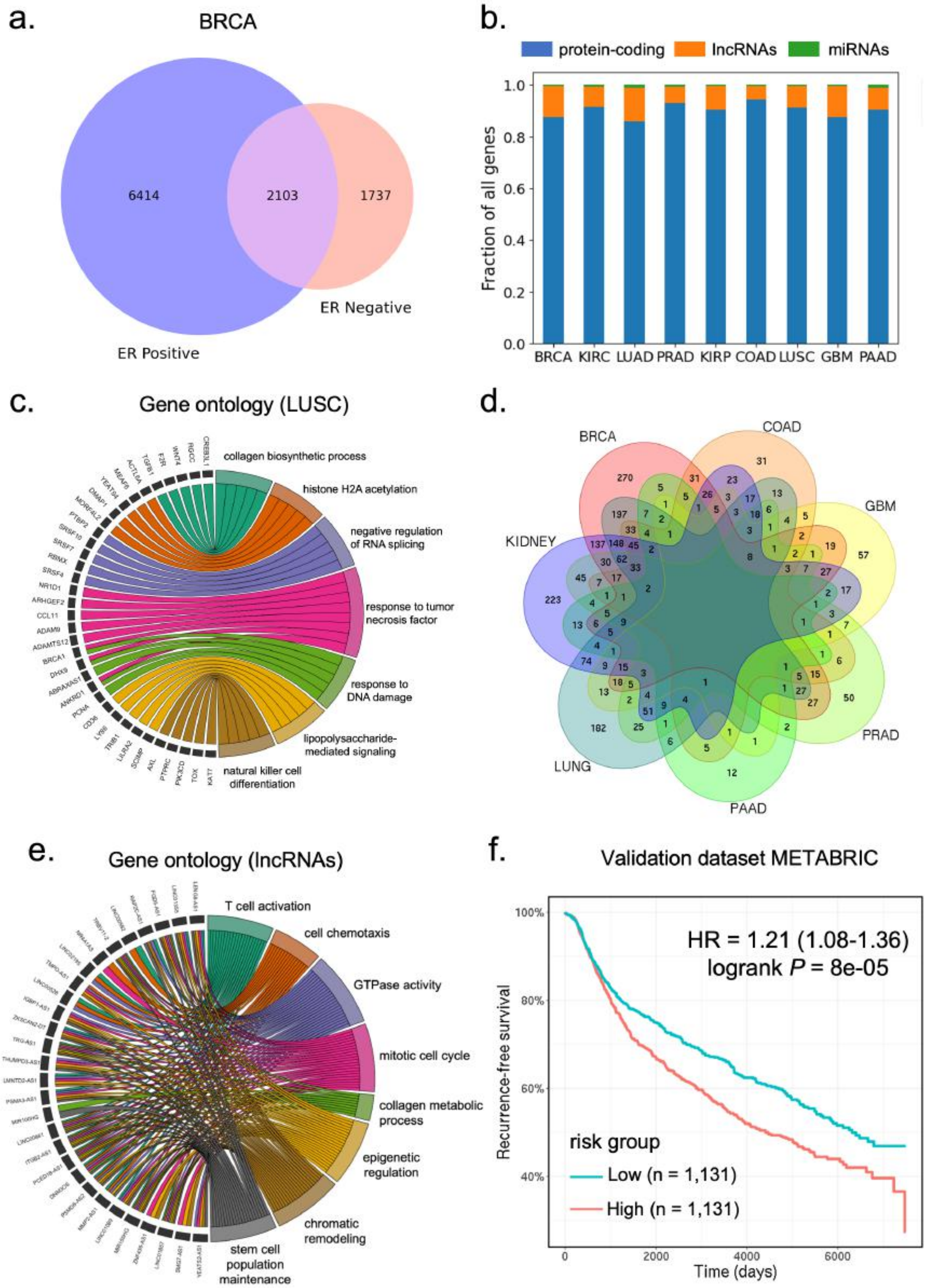


Fig. A3: **Characterization of the well-predicted genes.** a) Venn diagram showing the number of well predicted genes in the estrogen-receptor (ER) positive and ER negative breast cancer. b) The proportion of protein-coding genes, miRNAs and lncRNAs among the well predicted genes from each cancer type. c) Circos plot showing the biological processes associated with the well predicted genes in LUSC. d) Venn diagram showing the number of well predicted lncRNAs in each cancer type. Genes from the two lung cancer subypes (LUAD and LUSC) were combined, and same for the two kidney cancer subtypes (KIRC and KIRP). e) Circos plot showing the enriched biological processes associated with the well predicated lncRNAs. f) Kaplan-Meier curves of recurrence-free survival in the METABRIC validation dataset (n = 2,262 patients). Patients were split by the median risk score. HR: hazard ratio.

References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)

[2] Patil, P.D., Hobbs, B., Pennell, N.A.: The promise and challenges of deep learning models for automated histopathologic classification and mutation prediction in lung cancer. Journal of thoracic disease **11**(2), 369 (2019)

[3] Alsaafin, A., Safarpoor, A., Sikaroudi, M., Hipp, J.D., Tizhoosh, H.: Learning to predict rna sequence expressions from whole slide images with applications for search and classification. Communications Biology **6**(1), 304 (2023)

[4] Segal, E., Friedman, N., Kaminski, N., Regev, A., Koller, D.: From signatures to models: understanding cancer using microarrays. Nature genetics **37**(Suppl 6), 38–45 (2005)

[5] Zheng, Y., Jun, J., Brennan, K., Gevaert, O.: Epimix is an integrative tool for epigenomic subtyping using dna methylation. Cell Reports Methods, 100515 (2023)

[6] Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., Mao, Q., Yu, H., Cai, X.: Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. NPJ precision oncology **4**(1), 14 (2020)

[7] Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine **24**(10), 1559–1567 (2018)

[8] Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., *et al.*: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature medicine **25**(7), 1054–1056 (2019)

[9] Liao, H., Long, Y., Han, R., Wang, W., Xu, L., Liao, M., Zhang, Z., Wu, Z., Shang, X., Li, X., et al.: Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. Clinical and translational medicine **10**(2) (2020)

[10] Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.M.: Novel deep learning algorithm predicts the status of molecular pathways and key mutations in colorectal cancer from routine histology images. MedRxiv, 2021–01 (2021)

[11] Noorbakhsh, J., Farahmand, S., Foroughi Pour, A., Namburi, S., Caruana, D., Rimm, D., Soltanieh-Ha, M., Zarringhalam, K., Chuang, J.H.: Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. Nature communications **11**(1), 6367 (2020)

[12] Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M.: Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature cancer **1**(8), 800–810 (2020)

[13] Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., *et al.*: Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer **1**(8), 789–799 (2020)

[14] Jiang, S., Zanazzi, G.J., Hassanpour, S.: Predicting prognosis and idh mutation status for patients with lower-grade gliomas using whole slide images. Scientific reports **11**(1), 16849 (2021)

[15] Zheng, H., Momeni, A., Cedoz, P.-L., Vogel, H., Gevaert, O.: Whole slide images reflect dna methylation patterns of human tumors. NPJ genomic medicine **5**(1), 11 (2020)

[16] Pizurica, M., Larmuseau, M., Eecken, K., Brienen, L., Carrillo-Perez, F., Isphording, S., Lumen, N., Van Dorpe, J., Ost, P., Verbeke, S., Gevaert, O., Marchal, K.: Whole slide imaging-based prediction

of tp53 mutations identifies an aggressive disease phenotype in prostate cancer. Cancer Research, 22 (2023)

[17] Steyaert, S., Qiu, Y.L., Zheng, Y., Mukherjee, P., Vogel, H., Gevaert, O.: Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. Communications Medicine **3**(1), 44 (2023)

[18] Schaumberg, A.J., Rubin, M.A., Fuchs, T.J.: H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer. BioRxiv, 064279 (2016)

[19] Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)

[20] Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F.: Ai-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021)

[21] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[22] Ravi, V.M., Will, P., Kueckelhaus, J., Sun, N., Joseph, K., Salié, H., Vollmer, L., Kuliesiute, U., Ehr, J., Benotmane, J.K., *et al.*: Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. Cancer Cell **40**(6), 639–655 (2022)

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929 (2010)

[24] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.*: The genotype-tissue expression (gtex) project. Nature genetics **45**(6), 580–585 (2013)

[25] Thennavan, A., Beca, F., Xia, Y., Garcia-Recio, S., Allison, K., Collins, L.C., Gary, M.T., Chen, Y.-Y., Schnitt, S.J., Hoadley, K.A., et al.: Molecular analysis of tcga breast cancer histologic types. Cell genomics **1**(3) (2021)

[26] Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., *et al.*: A deep learning model to predict rna-seq expression of tumours from whole slide images. Nature communications **11**(1), 3877 (2020)

[27] Cao, L., Huang, C., Zhou, D.C., Hu, Y., Lih, T.M., Savage, S.R., Krug, K., Clark, D.J., Schnaubelt, M., Chen, L., *et al.*: Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell **184**(19), 5031–5052 (2021)

[28] Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., *et al.*: Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell **183**(5), 1436–1456 (2020)

[29] Wang, L.-B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., *et al.*: Proteogenomic and metabolomic characterization of human glioblastoma. Cancer cell **39**(4), 509–528 (2021)

[30] Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., *et al.*: Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell **182**(1), 200–225 (2020)

[31] Satpathy, S., Krug, K., Beltran, P.M.J., Savage, S.R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M.H., Avanessian, S.C., *et al.*: A proteogenomic portrait of lung squamous cell carcinoma. Cell **184**(16), 4348–4371 (2021)

[32] Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi,

Z., Arshad, O.A., *et al.*: Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell **177**(4), 1035–1049 (2019)

[33] Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., Veiga Leprevost, F., Reva, B., Lih, T.-S.M., Chang, H.-Y., *et al.*: Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell **179**(4), 964–983 (2019)

[34] Syed, Y.Y.: Oncotype dx breast recurrence score®: a review of its use in early-stage breast cancer. Molecular diagnosis & therapy **24**, 621–632 (2020)

[35] Slodkowska, E.A., Ross, J.S.: Mammaprint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert review of molecular diagnostics **9**(5), 417–422 (2009)

[36] Sestak, I., Filipits, M., Buus, R., Rudas, M., Balic, M., Knauer, M., Kronenwett, R., Fitzal, F., Cuzick, J., Gnant, M., *et al.*: Prognostic value of endopredict in women with hormone receptor–positive, her2-negative invasive lobular breast cancer. Clinical Cancer Research **26**(17), 4682–4687 (2020)

[37] Nielsen, T.O., Parker, J.S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S.R., Snider, J., Stijleman, I.J., Reed, J., *et al.*: A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. Clinical cancer research **16**(21), 5222–5232 (2010)

[38] Staaf, J., Häkkinen, J., Hegardt, C., Saal, L.H., Kimbung, S., Hedenfalk, I., Lien, T., Sørlie, T., Naume, B., Russnes, H., *et al.*: Rna sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer. NPJ breast cancer **8**(1), 94 (2022)

[39] Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.*: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346–352 (2012)

[40] Arora, R., Cao, C., Kumar, M., Sinha, S., Chanda, A., McNeil, R., Samuel, D., Arora, R.K., Matthews, T.W., Chandarana, S., *et al.*: Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. Nature Communications **14**(1), 5029 (2023)

[41] Zheng, Y., Carrillo-Perez, F., Pizurica, M., Heiland, D.H., Gevaert, O.: Spatial cellular architecture predicts prognosis in glioblastoma. Nature Communications **14**(1), 4122 (2023)

[42] Lin, H., Yang, Y., Hou, C., Zheng, J., Lv, G., Mao, R., Xu, P., Chen, S., Zhou, Y., Wang, P., *et al.*: Identification of col6a1 as the key gene associated with antivascular endothelial growth factor therapy in glioblastoma multiforme. Genetic testing and molecular biomarkers **25**(5), 334–345 (2021)

[43] Comba, A., Faisal, S.M., Dunn, P.J., Argento, A.E., Hollon, T.C., Al-Holou, W.N., Varela, M.L., Zamler, D.B., Quass, G.L., Apostolides, P.F., *et al.*: Spatiotemporal analysis of glioma heterogeneity reveals col1a1 as an actionable target to disrupt tumor progression. Nature communications **13**(1), 3606 (2022)

[44] Xu, K., Zhang, K., Ma, J., Yang, Q., Yang, G., Zong, T., Wang, G., Yan, B., Shengxia, J., Chen, C., *et al.*: Ckap4-mediated activation of foxm1 via phosphorylation pathways regulates malignant behavior of glioblastoma cells. Translational Oncology **29**, 101628 (2023)

[45] Ren, Y., Huang, Z., Zhou, L., Xiao, P., Song, J., He, P., Xie, C., Zhou, R., Li, M., Dong, X., *et al.*: Spatial transcriptomics reveals niche-specific enrichment and vulnerabilities of radial glial stem-like cells in malignant gliomas. Nature Communications **14**(1), 1028 (2023)

[46] Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., *et al.*: An integrative model of cellular states, plasticity, and genetics

for glioblastoma. Cell **178**(4), 835–849 (2019)

[47] He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)

[48] Graziani, M., Marini, N., Deutschmann, N., Janakarajan, N., Müller, H., Martínez, M.R.: Attention-based interpretable regression of gene expression in histology. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, pp. 44–60 (2022). Springer

[49] Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)

[50] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[52] Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al.: clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. The innovation **2**(3) (2021)

[53] Fang, Z., Liu, X., Peltz, G.: Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics **39**(1), 757 (2023)

[54] Simon, N., Friedman, J., Tibshirani, R., Hastie, T.: Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software **39**(5), 1–13 (2011) https://doi.org/10.18637/jss.v039.i05

[55] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)