nature communications

Article

https://doi.org/10.1038/s41467-024-54182-5

Digital profiling of gene expression from histology images with linearized attention

Received: 17 January 2024

Accepted: 4 November 2024

Published online: 14 November 2024

Check for updates

Marija Pizurica[1,2,6], Yuanning Zheng[1,6], Francisco Carrillo-Perez[1,6], Humaira Noor[1], Wei Yao[3], Christian Wohlfart[4], Antoaneta Vladimirova[3], Kathleen Marchal[2,7] & Olivier Gevaert[1,5,7]

Cancer is a heterogeneous disease requiring costly genetic profiling for better understanding and management. Recent advances in deep learning have enabled cost-effective predictions of genetic alterations from whole slide images (WSIs). While transformers have driven significant progress in non-medical domains, their application to WSIs lags behind due to high model complexity and limited dataset sizes. Here, we introduce *SEQUOIA*, a linearized transformer model that predicts cancer transcriptomic profiles from WSIs. *SEQUOIA* is developed using 7584 tumor samples across 16 cancer types, with its generalization capacity validated on two independent cohorts comprising 1368 tumors. Accurately predicted genes are associated with key cancer processes, including inflammatory response, cell cycles and metabolism. Further, we demonstrate the value of *SEQUOIA* in stratifying the risk of breast cancer recurrence and in resolving spatial gene expression at loco-regional levels. *SEQUOIA* hence deciphers clinically relevant information from WSIs, opening avenues for personalized cancer management.

Cancer is a dynamic disease characterized by intricate molecular and cellular evolution. Over the course of evolution, cancer becomes more heterogeneous, classified into inter-patient heterogeneity and intra-tumoral heterogeneity[1]. Inter-patient heterogeneity refers to the difference found between patients, which results from patient-specific factors, including germline genetic variations, differences in mutation profiles and environmental factors[2–4]. Comparatively, intra-tumoral heterogeneity describes the co-existence of cell subpopulations carrying different genomics, epigenomics and transcriptomics profiles within the same tissue. The spatial distribution of these cell subpopulations forms a complex ecosystem fostering signaling transduction that drives tumor progression[5,6]. A systematic understanding of cancer heterogeneity presents formidable challenges for effective diagnosis and management.

With the growing interest in precision medicine, molecular profiling has gained significant attention as a critical component of prognostication and treatment planning. In the past decade, the advancement of RNA sequencing (RNA-seq) has enabled the comprehensive measurement of gene expression profiles at both bulk tissue levels and at spatially resolved regional levels[3,5]. The resulting information has deepened our understanding of cancer heterogeneity, leading to the discovery of molecular signatures associated with treatment sensitivity[7–9]. However, incorporating gene expression analysis into clinical practice still represents a challenge. Current methods involve time-consuming and expensive laboratory procedures, limiting the integration of gene expression analysis in routine diagnostics.

With the digitization of histopathology glass slides into Whole-Slide Images (WSIs), unprecedented opportunities arise for cost-efficient analyses of tumor properties. Notably, WSIs are available

[1]Department of Medicine, Stanford Center for Biomedical Informatics Research (BMIR), Stanford University, Stanford, CA 94305, USA. [2]Internet Technology and Data Science Lab (IDLab), Ghent University, Ghent 9052, Belgium. [3]Roche Information Solutions, Roche Diagnostics Corporation, Santa Clara, CA 95050, USA. [4]Roche Diagnostics GmbH, Penzberg 82377, Germany. [5]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA. [6]These authors contributed equally: Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez. [7]These authors jointly supervised this work: Kathleen Marchal, Olivier Gevaert. e-mail: ogevaert@stanford.edu

without additional cost as they are obtained in routine clinical practice for diagnostics. Despite providing only morphological information, WSIs can also reflect the molecular traits of tumors. Over the past decade, machine and deep-learning methods have been developed to extract hidden morphological features from WSIs that are associated with molecular properties[10–22].

Although remarkable progress has been made in this domain, applying state-of-the-art methods to WSIs remains exceedingly challenging. Due to the immense size and resolutions of WSIs, they are first cropped into thousands of smaller tiles. Traditionally, models were developed at 'tile-level', where the model is trained to make predictions for individual tiles[10–20]. Current datasets (e.g., TCGA and CPTAC) provide predominantly bulk RNA-Seq profiling, where only a single gene expression label is available for all tiles within the WSI. However, due to intra-tumor heterogeneity, not all tiles in the WSI carry the same genetic profile. Hence, precise annotations are needed in tile-level workflows to indicate which tiles within the WSI can be used for model training[19]. Gathering these annotations for gene expression post-hoc is time-consuming and imprecise. Moreover, tile-level models cannot capture contextual and hierarchical relationships between multiple tiles of an image.

On the other hand, 'slide-level' workflows have been developed that take into account all tiles in the image at once, with no need for precise annotations. To aggregate information across tiles, HE2RNA utilizes a multilayer perceptron (MLP)[21]. While this approach reached reasonable performance, the MLP can not effectively model the contextual relationships across tiles, thereby limiting its performance. In contrast, tRNAsformer[22] employs a transformer encoder whose self-attention weights allow us to model the contextual inter-tile interactions. However, despite the advanced modeling capabilities of transformers, they are prone to overfitting when the training dataset is small due to the large number of parameters involved in self-attention[23]. In addition, both methods rely on convolutional neural networks (CNNs) as tile feature extractor. These CNNs were pre-trained on the ImageNet dataset, which may not effectively capture the histological information from the images.

To tackle these challenges, we propose *SEQUOIA*, a deep-learning model for **S**lide-based **E**xpression **Qu**antificati**o**n using **Li**nearized **A**ttention. To capture contextualized WSI features, we adapt the parameter-heavy self-attention within the transformer for linearized attention. In addition, we leverage UNI, a foundation model optimized for histological feature extraction[24]. *SEQUOIA* is developed on 7584 tumor samples across sixteen cancer types and validated in two independent cohorts. Further, we establish that the genes with well-predicted expression values are involved in key cancer processes and inform the risk for breast cancer recurrence. Finally, we show how *SEQUOIA* can be used to resolve loco-regional gene expression patterns using two spatial transcriptomics datasets. In conclusion, *SEQUOIA* offers a cost-efficient way to infer and analyze gene expression patterns on a large scale, with potential applications in both research and clinical settings.

Results

SEQUOIA as tool for gene expression prediction from WSIs

We present *SEQUOIA*, a deep-learning model for **S**lide-based **E**xpression **Qu**antificati**o**n using **Li**nearized **A**ttention. ("Methods", Fig. 1a–c). To train and evaluate the model, we utilized WSIs and matched bulk RNA-seq gene expression data of sixteen cancer types available in The Cancer Genome Atlas (TCGA, Supplementary Table 1): (1) bladder urothelial carcinoma (BLCA), (2) breast invasive carcinoma (BRCA), (3) colon adenocarcinoma (COAD), (4) glioblastoma multiforme (GBM), (5) head and neck squamous cell carcinoma (HNSC), (6) kidney renal clear cell carcinoma (KIRC), (7) kidney renal papillary cell carcinoma (KIRP), (8) liver hepatocellular carcinoma (LIHC), (9) lung adenocarcinoma (LUAD), (10) lung squamous cell carcinoma (LUSC), (11) pancreatic adenocarcinoma (PAAD), (12) prostate adenocarcinoma (PRAD), (13) skin cutaneous melanoma (SKCM), (14) stomach adenocarcinoma (STAD), (15) thyroid carcinoma (THCA), (16) uterine corpus endometrial carcinoma (UCEC).

Since histological phenotypes and gene expression profiles vary across cancer types, the model was independently developed and validated 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 allocated for training (of which 10% were used as validation set), while the remaining 20% were reserved for testing (Supplementary Fig. 1). For each gene, we concatenated the predicted expression values of tissues from the test sets and compared them to the ground truth using Pearson's correlation analysis and root mean squared error (RMSE).

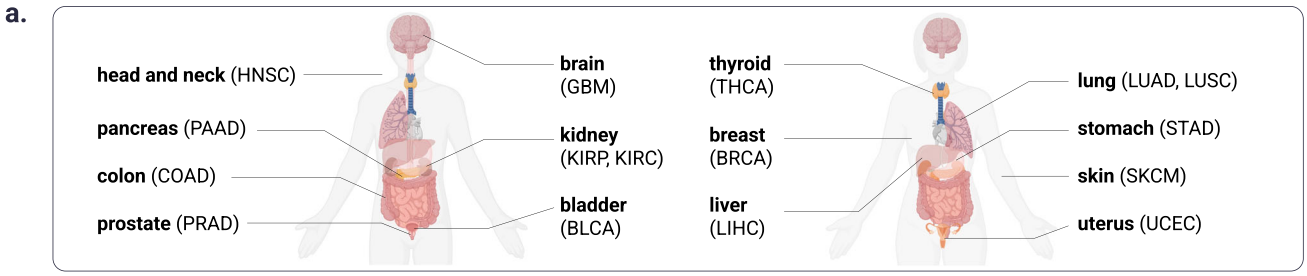
The resulting correlation coefficient and RMSE values were further compared to those obtained with a random, untrained model of the same architecture (see "Methods" for details). To identify genes with significantly well-predicted expression levels, we combined three criteria: (1) the predicted gene expression values must be significantly correlated with the ground truth, with a positive correlation coefficient and the associated P value smaller than 0.05 ($r_1 > 0$ and $p_1 < 0.05$); (2) r_1 must be statistically higher than r_2 ($r_1 > r_2$), as determined by Steiger's Z test, where r_2 represents the correlation coefficient obtained from the random model. For this comparison, we required the raw Steiger P value to be smaller than 0.05 ($p_2 < 0.05$) and the adjusted P value by Benjamini–Hochberg correction smaller than 0.2 ($p_3 < 0.2$); (3) the RMSE values obtained from the trained model must be smaller than those from the random model.

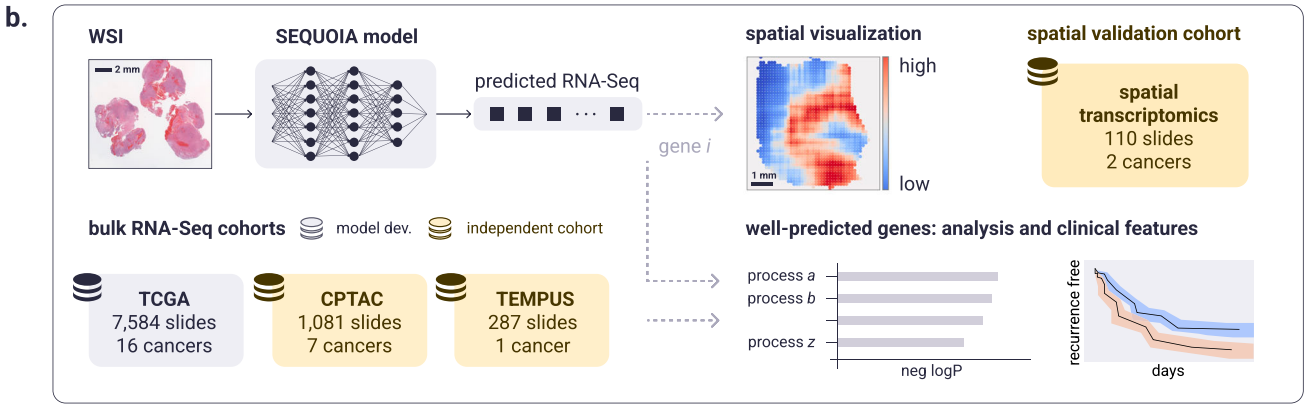
SEQUOIA was able to accurately predict the expression levels of many genes. On average, 15,344 out of 20,820 genes were significantly well predicted across the sixteen cancer types (Fig. 1d and Supplementary Table 3). The number of well-predicted genes was positively correlated with the number of training samples available in each cancer (Supplementary Fig. 2). The highest number ($N = 18{,}878$) of genes was identified in BRCA, the cancer type with the most available slides ($N = 1130$). Further, we identified 18,758 well-predicted genes in THCA ($N = 517$ slides) and 17,623 genes in KIRC ($N = 514$ slides). Comparatively, PAAD had the lowest number of well-predicted genes ($N = 9535$) as well as the lowest number of slides ($N = 202$). To further test the relation between performance and dataset size, we performed different downsamplings (keeping 20%, 30%, 40%) of available data within a single cancer type (BRCA), indeed confirming a consistent trend of decreasing performance across all metrics when the dataset size is reduced (Supplementary Table 6).

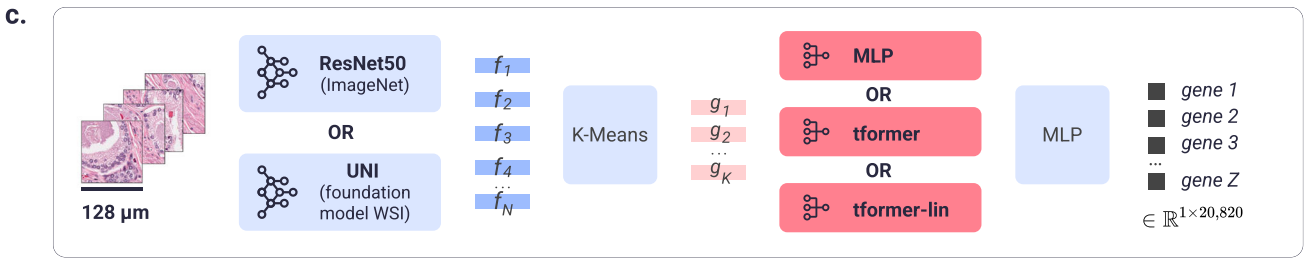
Since the histological appearance of BRCA has been shown to be associated with hormone receptor status[25], we separately assessed the predictive performance in the estrogen receptor (ER) negative and ER positive subtypes. There were 18,139 and 12,241 genes that passed our significant thresholds in the ER positive and negative subtype, respectively. Of these genes, 11,834 genes were significantly well predicted in both subtypes. These results demonstrate the capacity of *SEQUOIA* in predicting gene expression signals specific to breast cancer subtypes.

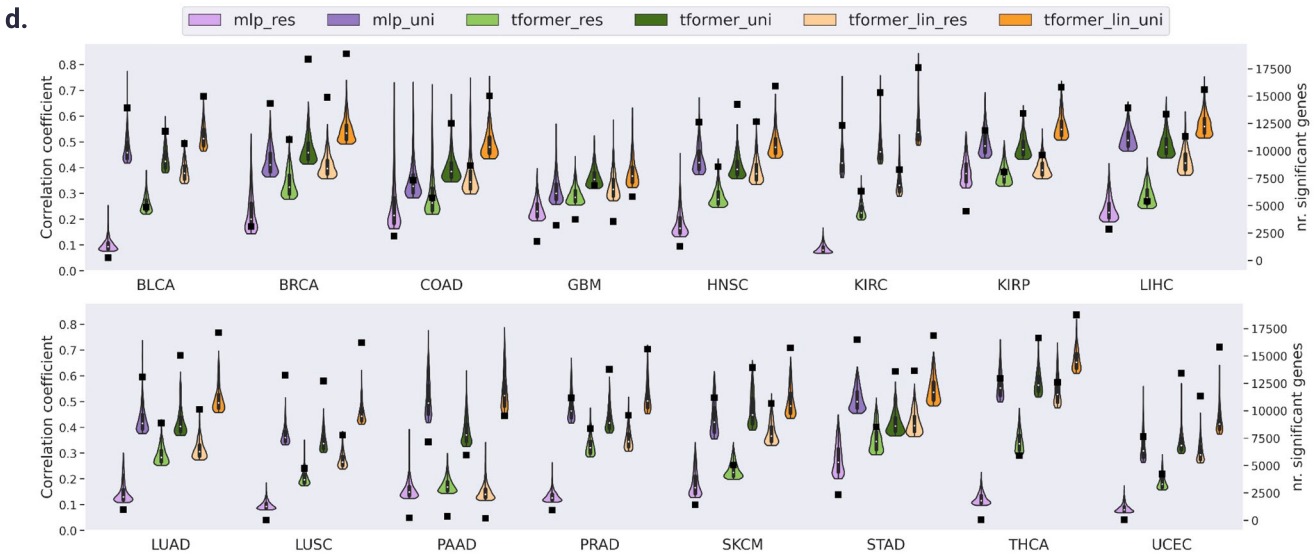
To compare the performance of our model with existing architectures, we thoroughly benchmark the added value of UNI over ResNet-50 pre-trained on ImageNet, and the benefit of using the linearized transformer for tile aggregation instead of a regular transformer (as in tRNAsformer[22]) or MLP layers (as in HE2RNA[21]). Both UNI and improved tile aggregation methods (MLP vs. transformer vs. linearized transformer) independently boost prediction performance by large margins. On average, across cancer types, the number of well-predicted genes (Fig. 1d and Supplementary Table 3) increases by 830% when using UNI instead of ResNet for MLP aggregation, by 210% for transformer aggregation and 155% for linearized transformer aggregation.

https://doi.org/10.1038/s41467-024-54182-5









When considering ResNet features, going for a transformer instead of an MLP increases the number of well-predicted genes by 450%, with an additional increase of 155% when choosing the linearized transformer. Similarly, for UNI features, using a transformer instead of an MLP gives 115% more significant genes, and going for the linearized version gives an additional increase of again a factor 115%. When considering the best feature extractor (UNI), the largest added value of the linearized transformer over the regular version is observed in PAAD, the cancer type with the fewest available training data (increase factor 160%), while the smallest added value occurs in BRCA, for which most slides are available (factor 10%).

The superiority of UNI and the linearized transformer is also demonstrated in the correlation coefficients between predicted gene expression values and ground truth (Fig. 1d, Supplementary Table 4). Using UNI instead of ResNet gives an increase in the correlation coefficient of 250% for MLP, 155% for transformer and 145% for the linearized version. When considering UNI features, the correlation coefficient is similar when using MLP/transformer (0.428/0.419) but

https://doi.org/10.1038/s41467-024-54182-5

Fig. 1 | Overview of the workflow for the *SEQUOIA* model. a Cancer types on which the *SEQUOIA* model is developed and validated. Created with BioRender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en). **b** The model is trained and evaluated using matched WSIs and bulk RNA-Seq data from sixteen cancer types available in the TCGA 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 two spatial transcriptomics datasets[5,44]. Clinical utility is demonstrated by evaluating the model's capacity to predict cancer recurrence. **c** *SEQUOIA* architecture and benchmarked variations. First, *N* tiles are sampled from the WSI. Feature vectors are extracted using either ResNet-50 pre-trained on ImageNet or UNI. We then cluster the feature vectors into *K* clusters, and within-cluster averages result in *K*-aggregated feature vectors. Next, either a Multi Layer Perceptron (MLP), a transformer (`tformer'), or linearized transformer (`tformer-lin') (followed by an MLP) are used to predict gene expression values. **d** Performance benchmarking of *SEQUOIA*. Violin plots illustrate the distribution of Pearson correlation coefficients (left *y* axis) between the predicted and ground truth gene expression values in TCGA test sets. Within each violin plot, a miniature box-and-whisker plot is shown where whiskers bound the min–max values of the data, the bounds of the box represent lower (Q1)/upper (Q3) quartiles, and the central value contains the median value. The top 1000 genes with the highest correlation coefficients obtained from each model are shown. Black squares indicate the absolute number (right *y* axis) of genes with significantly well-predicted expression levels. WSI Whole Slide Image. Source data for **d** are provided in the Source Data File.

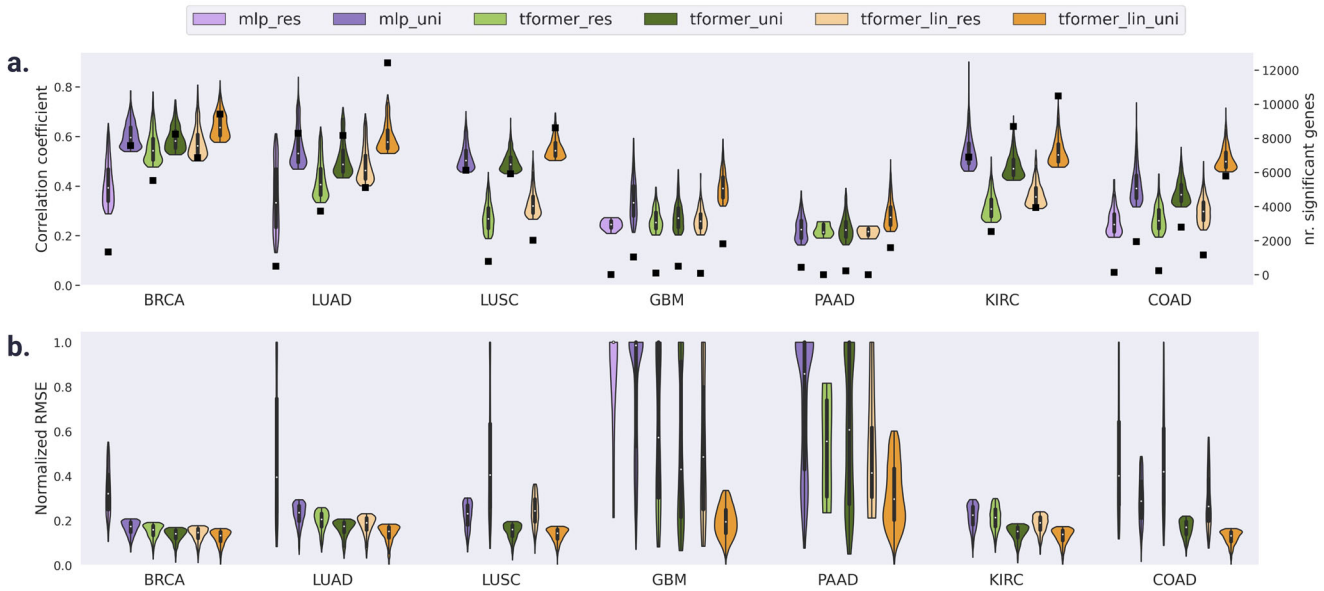


Fig. 2 | Genes that validate both in TCGA test sets and in external cancer cohorts. a Violin plots show the distribution of the Pearson correlation coefficient (left *y* axis) of genes that validate both CPTAC and TCGA test set. Within each violin plot, a miniature box-and-whisker plot is shown where whiskers bound the min–max values of the data, the bounds of the box represent lower (Q1)/upper (Q3) quartiles, and the central value contains the median value. The top 1000 genes with the highest correlation coefficients obtained from each model are shown. Black squares indicate the absolute number (right *y* axis) of genes that validate both TCGA test set and CPTAC. **b** Same as (**a**) for Normalized RMSE. Note that mlp_res did not have any significant genes that overlap between TCGA-CPTAC for LUSC, PAAD, KIRC (Supplementary Table 9) and hence violin plots for these settings do not exist. Source data are provided in the Source Data File.

increases largely when considering the linearized transformer (0.504). The same conclusion can be made by analyzing RMSE (Supplementary Table 5).

SEQUOIA generalizes to independent cohorts

Deep-learning models trained on a specific dataset may be subject to bias due to technical noise, such as stain variations and color range, potentially leading to overfitting and limiting their ability to generalize to other datasets. To test the generalization capacity of *SEQUOIA* and the benchmarked variations, we apply the models developed in the TCGA cohort to the matched cancer type available in the CPTAC (Clinical Proteomic Tumor Analysis Consortium) cohort[3,26–31]. We extend our validation to seven cancers from six tissues available in the CPTAC dataset, including breast, lung, kidney, brain, colon and pancreas (Supplementary Table 2).

The correlation coefficients obtained from *SEQUOIA* are significantly higher (Mann–Whitney $p < 0.0001$) in 6 out of 7 tested cancer types compared to all other combinations (Supplementary Table 7, Supplementary Fig. 3). Only in case of KIRC, the *SEQUOIA* model obtains the same correlation coefficient as a model with UNI features with an MLP aggregation. On average, across the seven cancer types, *SEQUOIA* achieves a correlation coefficient of 0.503, thereby greatly surpassing the second-best model (UNI+MLP), which achieves an average coefficient of 0.463. Similar to findings on TCGA, the best performance (0.636) occurs in BRCA, the cancer type which had the most available training data, while PAAD, the cancer type with the smallest training size, has the worst performance (0.321).

In addition, the RMSE values obtained from *SEQUOIA* were significantly lower (Mann–Whitney $p < 0.0001$) in six out of seven cancer types compared to all other tested model combinations (Supplementary Table 8, Supplementary Fig. 4). Only for PAAD, the RMSE values of *SEQUOIA* are only slightly lower (not statistically significant) than that of the competing model (UNI+transformer). On average, across the cancer types, *SEQUOIA* achieves an RMSE of 0.135, again surpassing the second-best model (UNI+transformer), which achieves an average RMSE of 0.144.

Next, we looked into the overlap between well-predicted genes identified in each specific cancer type from the CPTAC cohort and the well-predicted genes in the TCGA cohort (Fig. 2 and Supplementary Table 9). In terms of number of well-predicted genes, *SEQUOIA* greatly surpasses all other model combinations with 7159 genes validated on average, achieving a 145% increase compared to the second-best

model (UNI+transformer, 4934 genes). Also, in terms of correlation coefficient and RMSE, *SEQUOIA* outperforms all other model combinations (Fig. 2, Supplementary Tables 10, 11). These results indicate a higher generalization capacity of *SEQUOIA* compared to all other model combinations.

Most genes are validated across both TCGA and CPTAC for LUAD (12,422), followed by KIRC (10,477), BRCA (9418), LUSC (8610), COAD (5784), GBM (1816) and PAAD (1589) (Supplementary Table 9). Regarding correlation coefficients among genes in the overlap (Supplementary Table 10), the highest value (0.636) was found for BRCA, followed by LUAD (0.578), LUSC (0.543), KIRC (0.525), COAD (0.498), GBM (0.391) and PAAD (0.274).

To further test the generalization capacity of our models, we extended the validation to a LUAD cohort from Tempus ("Methods", $N = 287$ slides from $N = 249$ patients). This led to the identification of 5851 genes that were well-predicted across all three (TCGA, CPTAC and Tempus) LUAD cohorts.

Pathway-level analysis of the predicted gene expression values

To characterize the biological functions of the well-predicted genes in our model, we carried out gene set analysis. First, we performed gene set variation analysis (GSVA) to assess the predicted gene expression values at the pathway level[32]. Second, we conducted hyper-geometric tests using the well-predicted gene list obtained from each cancer type. For both analyses, we included three categories: (1) gene ontology (biological process), (2) KEGG pathway and (3) cell-type signature. In our GSVA analysis, the average correlation coefficient between the ground truth and predicted pathway scores across TCGA cancer types was 0.53 (range 0.37–0.67) for gene ontology, 0.45 (range 0.19–0.58) for KEGG pathways, and 0.52 (range 0.33–0.52) for cell-type signatures. (Fig. 3a).

For gene ontology, hyper-geometric analysis revealed several common pathways enriched in the well-predicted genes across cancer types, including T cell activation (*CCL2*, *CCR2*, *CCDC88B*), cell-matrix adhesion (*EMP2*, *COL16A1*, *VEGFA*), epithelial-mesenchymal transition (*BMP2*, *PDPN*, *SMAD2*) and response to oxidative stress (*TP53*, *PRDX1*, *VRK2*) (Fig. 3b and Supplementary Data 1). Additionally, some gene sets were enriched in specific cancer types. For instance, in STAD, we identified genes associated with dendritic cell migration (*CXCR1*, *CCL5*, *ALOX5*), B cell homeostasis (*BLK*, *BAX*, *DOCK10*), endothelial cell development (*ICAM1*, *COL15A1*, *MYADM)*, and epithelial-mesenchymal transition (*BCL9L*, *PTEN*, *OVOL2*) (Fig. 3c). Published studies have revealed the critical roles of dendritic cells in promoting the anti-tumoral immunity in gastric cancers, and patients with a low density of tumor-infiltrating dendritic cells had lower survival rates than those with high density[33]. Conversely, higher activation levels of endothelial cell development pathways may reflect enhanced angiogenesis in tumor tissues, which is a known factor that drives the progression and metastasis of gastric cancer[34].

KEGG pathway analysis further revealed the regulatory effects of the well-predicted genes in VEGF signaling (*SPHK1*, *HRAS*, *HSPB1*), HIF-1 signaling (*GAPDH*, *HIF1A*, *VEGFA*), the PD-L1 expression and checkpoint pathway (*CD247*, *CD274*, *MAPK11*), and NF-kappa B signaling (*CXCL12*, *NFKB1*, *PRKCB*) (Fig. 3d, e and Supplementary Data 2).

In addition, we also identified several well-predicted cell-type markers, including those for endothelial cells (*CD69*, *CD93*), CD4 T cell (*CD3E*, *CD4*, *CD48*), M2 macrophage (*CD14*, *CD163*, *CD84*), and B cell (*CD19*, *CD53*, *CD37*) (Supplementary Fig. 5a and Supplementary Data 3). These results indicate the capacity of *SEQUOIA* in capturing tumor microenvironmental features.

Notably, the presented gene ontology and KEGG gene sets were not enriched with the inaccurately predicted genes (i.e., genes that did not pass our significant thresholds) (Supplementary Figs. 5b, c). The functions of inaccurately predicted genes were not strictly related to cancers or not interpretable in the context of the disease (Supplementary Fig. 5d). These results highlight the functional specificity of genes that can be well predicted with *SEQUOIA*.

Further analysis on the genes that were well predicted in both the TCGA and CPTAC cohorts revealed their functions in regulating cell cycle, T cell activation, DNA replication and cell adhesion (Supplementary Figs. 5e, f). These results indicate that the well-predicted genes from *SEQUOIA* were primarily and specifically related to the regulation of cancer development and progression.

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, we next assessed whether these genes have prognostic value. We focused our analysis on breast cancer, in which the highest number of genes ($N = 18,878$) were significantly well-predicted for their expression levels in the TCGA test sets.

The well-predicted genes encompass various published prognostic signatures (Supplementary Data 4)[35–38]. These include 46 out of 70 (66%) genes of the Mammaprint signature, 45 out of 50 (90%) genes of the PAM50 signature, 13 out of 21 (62%) genes of the Oncotype DX signature, all 12 (100%) genes of the EndoPredict signature, 5 out of 7 (71%) genes of the Breast Cancer Index and 3 out of 5 (60%) gene of the Mammostrat signature.

Since the Oncotype DX and MammaPrint signatures are developed on data from RT-qPCR and microarray assays, and their exact mathematical formulas are protected by proprietary license, we next developed an RNA-seq-based gene expression signature that can stratify the risk of breast cancer recurrence. To this end, we fitted a regularized Cox regression model on the ground-truth gene expression values for the genes well predicted by *SEQUOIA*, and the model aims to predict a risk score of recurrence for each patient (see "Methods" for details). We developed the model on the TCGA cohort ($N = 909$ patients) and independently validated it using data from the SCANB ($N = 5034$ patients) and METABRIC ($N = 1933$ patients) cohorts[39,40].

Our analysis led to the identification of a gene expression signature comprising 272 genes significantly associated with recurrence (Fig. 4a–c and Supplementary Data 5). To demonstrate its performance in risk stratification, we first treated the predicted risk scores as a dichotomous variable. Patients within each cohort were divided into a high-risk and a low-risk group based on the median score. Results from our log-rank test demonstrate that the high-risk group had significantly worse prognosis compared to the low-risk group in both the TCGA discovery set ($P < 2e{-}16$) and the two validation sets (SCANB: $P < 2e{-}16$; METABRIC: $P = 0.0001$) (Fig. 4a–c). It is worth noting that the gene expression data from METABRIC were generated using a microarray assay, therefore a lower performance is expected.

Next, we treated the predicted risk score as a continuous variable and evaluated its performance using regression analysis. Since breast cancer subtype is a confounding variable in risk prediction, we incorporated PAM50 molecular subtypes and hormone (i.e., estrogen and progesterone) receptor status as covariates into our analyses. In both validation datasets, the predicted risk score was significantly associated with prognosis: SCANB (HR = 7.86, 0.95 CI = 4.96–12.40, covariate-adjusted $P = 1.50e{-}18$) and METABRIC (HR = 1.68, 0.95 CI = 1.23–2.29, covariate-adjusted $P = 0.001$). Gene ontology analysis of the signature genes revealed their regulatory functions in cell growth (*IGFBP7*, *RGS2TOMM70*), angiogenesis (*GATA2*, *SFRP2*, *RUNX1*), cell-cell adhesion (*EMILIN1*, *EFNB1*, *CD83*), regulation of T cell activation (*SMARCB1*, *IL1RL2*, *HLA-DQB2*) and response to oxidative stress (*IL1A*, *MMP3*, *ERO1A*) (Fig. 4d).

So far, we have developed and validated a gene signature using the ground-truth gene expression values. We then assessed whether utilizing the gene expression values predicted from histology images was sufficient to stratify the risk groups. For each patient in the TCGA

https://doi.org/10.1038/s41467-024-54182-5

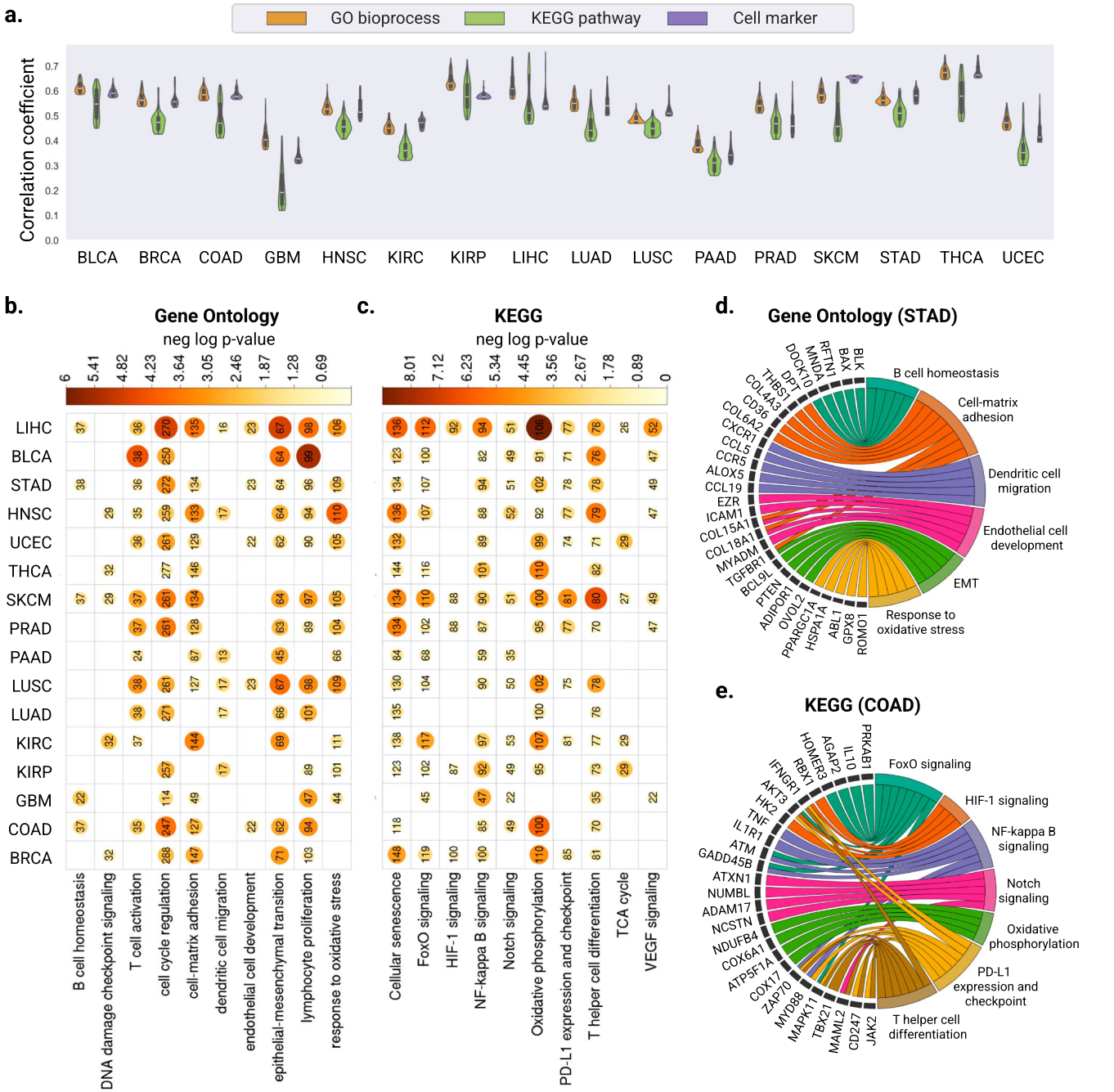


Fig. 3 | Evaluation of gene expression predictions at the pathway level. a Violin plots illustrating the distribution of Pearson correlation coefficients (left *y* axis) between the predicted and ground truth pathway enrichment scores in TCGA test sets. Within each violin plot, a miniature box-and-whisker plot is shown where whiskers bound the min-max values of the data, the bounds of the box represent lower (Q1)/upper (Q3) quartiles, and the central value contains the median value. The top 100 pathways with the highest correlation coefficients obtained from each model are shown. **b**, **c** Heatmap showing the significant *P* values obtained from one-sided hyper-geometric tests in **b** gene ontology and **c** KEGG pathway analysis of the well-predicted 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**, **e** Circos plots showing the **d** biological process enriched with the well-predicted genes in STAD and **e** KEGG pathways in COAD. Gene names are displayed on the left and the corresponding biological processes are shown on the right. Source data for all panels are provided in the Source Data File.

test set, we calculated a risk score using the same risk coefficient in our Cox regression model, but this time replacing the ground-truth gene expression values with the predicted values by *SEQUOIA*. As shown in Fig. 4e, patients assigned with high-risk scores demonstrated significantly shorter recurrence-free survival compared to patients with low-risk scores (Log-rank $P = 0.04$).

To benchmark the predictive performance of this gene expression-based model, we next trained a separate deep-learning model to directly predict recurrence-free survival from histology images. The architecture of this model was identical to that of *SEQUOIA*, except it was trained to predict a risk score for each patient rather than gene expression values (see "Methods" for details). We found that using histology images alone was not able to effectively stratify the patients (Fig. 4f, Log-rank $P = 0.43$). These results demonstrate the potential of *SEQUOIA* in predicting breast cancer recurrence through gene expression prediction.

https://doi.org/10.1038/s41467-024-54182-5

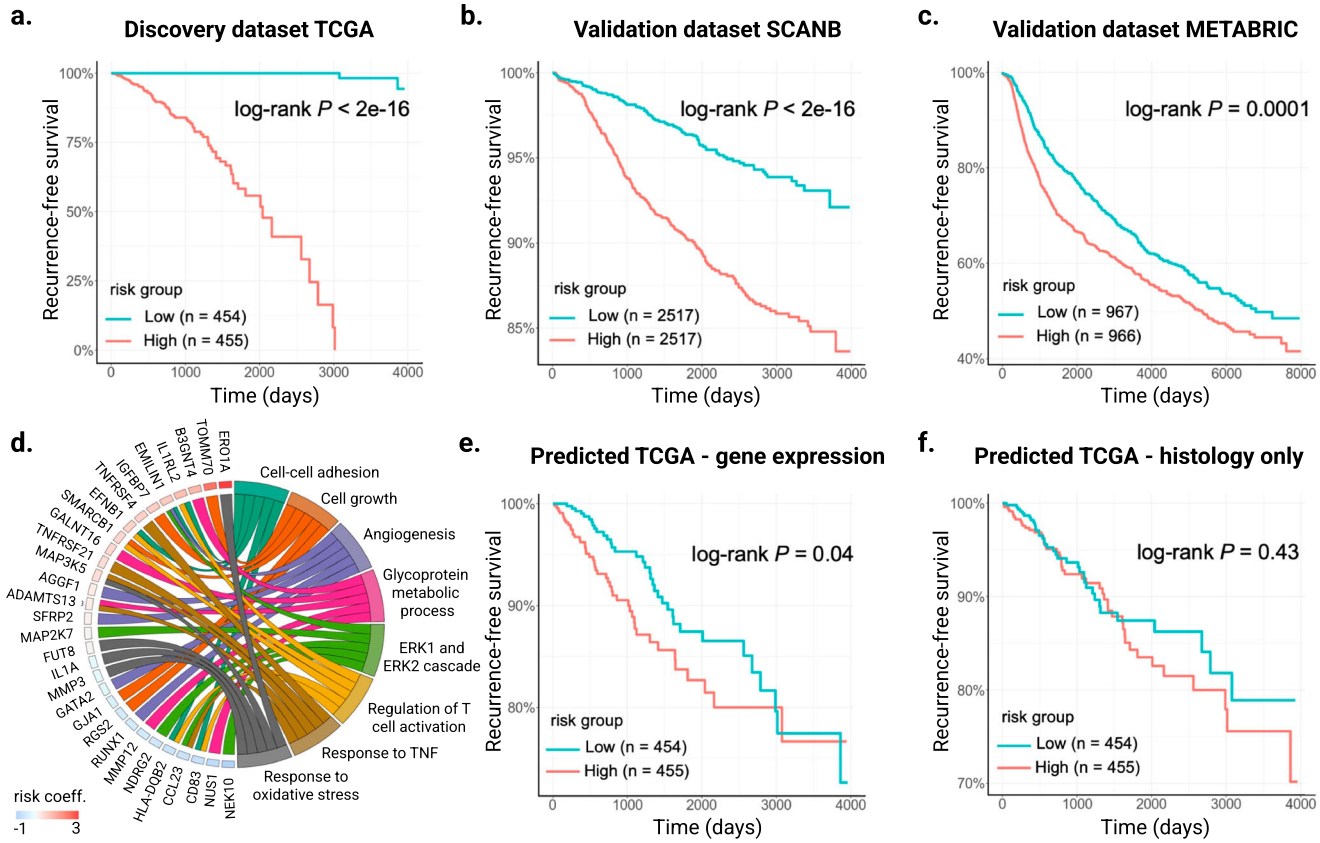


Fig. 4 | Development and validation of a digital gene expression signature for predicting breast cancer recurrence. a Kaplan–Meier curves of recurrence-free survival obtained from the TCGA discovery dataset. Patients were split by the median risk score. **b**, **c** Kaplan–Meier curves of recurrence-free survival in the **b** SCANB and **c** METABRIC validation datasets. **d** Circos plot showing the biological process 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. **e** Kaplan–Meier curves of recurrence-free survival obtained from the predicted gene expression values of the TCGA test set. Patients were split by the median risk score. **f** Kaplan–Meier curves of recurrence-free survival directly predicted from histology images of the TCGA test set. Patients were split by the median risk score. Source data for all panels are provided in the Source Data File.

Tile-level predictions validated with spatial transcriptomics

We have demonstrated the ability of *SEQUOIA* to predict RNA-Seq gene expression values collected from bulk tissues. However, gene expression patterns are known to vary across different tumor regions due to intra-tumoral heterogeneity resulting from uneven spatial distributions of cell phenotypes. Uncovering spatial gene expression patterns can reveal the intricate landscape of tumor architecture and signaling environment, which is known to affect tumor growth, metabolic processes, and resistance to therapy[6,8]. We hence investigated whether our models trained at the slide level can be used to predict gene expression values at loco-regional levels within tumour tissues.

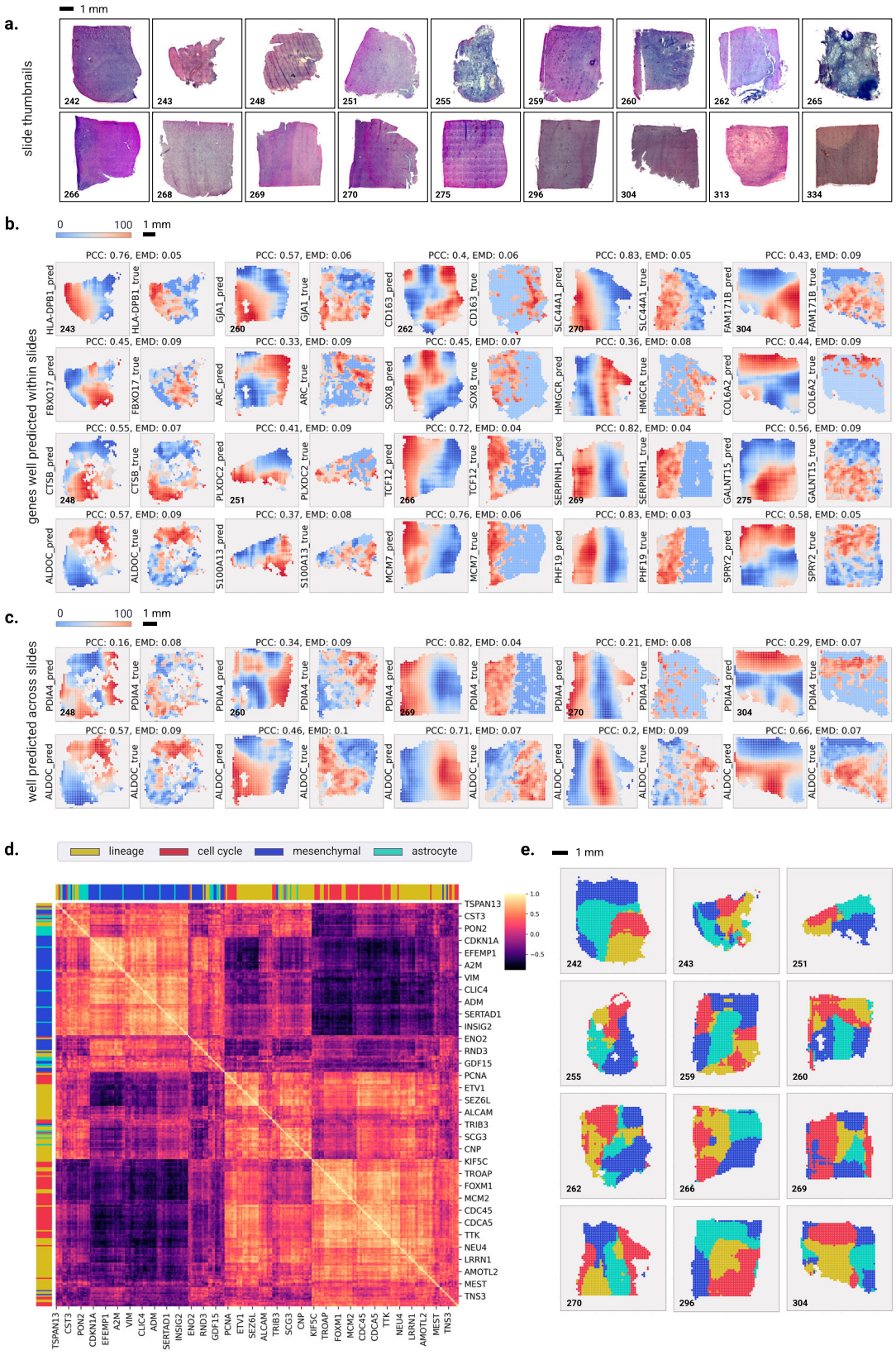
Here, we implemented a sliding-window approach to generate tile-level predictions of gene expression. The histology image was first processed using a sliding window of 10 × 10 tiles starting from the left upper corner, where the dimensions of each tile (128 µm × 128 µm) were consistent with those used for training *SEQUOIA* models. The window size was determined based on *SEQUOIA* architecture, which requires 100 feature vectors as input. For each window, a 100 × 1024 feature vector was extracted and then fed to *SEQUOIA* for generating a prediction. This prediction was then stored for every tile within the window (see "Methods" for details). After processing the entire image, the predicted gene expression for each tile was calculated as the average of the stored values for that tile. A stride of 1 (tile) was chosen for fine-grained analysis.

To validate the prediction, we utilized data from an independent cohort of patients with GBM, which contains matched histology images and spatial transcriptomics data ($N = 54{,}000$ gene expression spots from $N = 18$ patients), providing tile-level ground truth gene expression measurements[5] (Fig. 5a). We focused our analysis on the top 1000 genes for which *SEQUOIA* generated the best-generalized predictions across both the TCGA test set and CPTAC (i.e., genes with the highest Pearson correlation coefficients, PCCs). 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 both the PCC and EMD as evaluation metrics. EMD values are bounded between 0 and 1, with lower values indicating a closer correspondence between predictions and ground truth. Note that PCC is a pixel-level metric, and hence has limitations to evaluate spatial performance. Specifically, if a prediction is shifted with a few pixels, this may heavily impact the PCC, while in reality, this small shift may not be that noticeable. In addition, the spatial slides are smaller and more homogeneous compared to TCGA slides (number of tiles ranging from 250 to 1500 compared to 4000 in TCGA), in which case PCC cannot fully capture true performance. To address these limitations, we included EMD as an additional metric, which takes into account the 2D Euclidean distance between predictions and ground truth (see Methods).

Although the quality of the H&E images in the spatial GBM dataset is considerably lower than on TCGA (Supplementary Fig. 6), *SEQUOIA* was able to achieve good prediction performance across many genes on this spatial dataset (Fig. 5b, c). On average, *SEQUOIA* achieved an EMD of 0.15 (95% CI = 0.149–0.151) across all slides and genes. Well-predicted genes include *COL6A2* (avg. PCC 0.14, EMD 0.14 across slides), *S100A13* (avg. PCC 0.16, EMD 0.14) and *ALDOC* (avg. PCC 0.16,

https://doi.org/10.1038/s41467-024-54182-5



EMD 0.13), each of which have been associated with GBM malignancy and prognosis[41–43]. These results highlight the potential of *SEQUOIA* in predicting spatial gene expression patterns related to GBM malignancy and prognosis.

Since *HE2RNA*[21] has also been shown to be capable of predicting spatial gene expression levels, we benchmarked the predictive performance of *SEQUOIA* with *HE2RNA*. We considered the top 100 genes predicted within each model. Both in terms of EMD and PCC, *SEQUOIA* outperformed HE2RNA (Supplementary Fig. 7, Supplementary Table 12). The median PCC for SEQUOIA was 0.255, almost doubling the value of HE2RNA (0.138). Higher performance was observed in slides with high degrees of spatial variance in gene expression[5,6]. In

Fig. 5 | Spatial visualization of gene expression predicted 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 shown on the left and ground truth on the right. The prediction and ground truth maps were normalized to percentile scores between 0 and 100. Above each pair of prediction and ground truth, we show the Pearson Correlation Coefficient (PCC) and Earth Mover's Distance (EMD) metric. **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. Above each pair of prediction and ground truth, we show the Pearson Correlation Coefficient (PCC) and Earth Mover's Distance (EMD) metric. **d** Heatmap showing the Pearson 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. Source data for panel d are provided in the Source Data File.

the best-performing slide (269), SEQUOIA achieved a median PCC of 0.678 across the 100 genes. Although slides differ significantly in terms of H&E quality, staining and heterogeneity, we also compared the extent to which genes validate across different slides in both models. Hereto, we computed the number of slides where genes are predicted with PCC > 0.1 (Supplementary Table 8). Again, *SEQUOIA* greatly outperformed HE2RNA, with 60 genes validating to more than 8 slides compared to only 11 with HE2RNA.

We then also evaluated the spatial prediction performance of *SEQUOIA* on another independent spatial transcriptomics dataset for breast cancer[44], containing $N = 92$ slides from $N = 48$ unique patients. Again, despite the lower quality of the H&E slides in this cohort compared to TCGA (Supplementary Fig. 9), *SEQUOIA* was able to predict the spatial expression of many genes, thereby outperforming HE2RNA. On average across all slides, the median PCC of the top 100 genes for *SEQUOIA* was 0.168 (compared to 0.154 with HE2RNA) (Supplementary Table 14). In slides with best generalization performance (SPA145, SPA143, SPA146, SPA148), *SEQUOIA* achieved a median PCC > 0.45. Also in terms of EMD, *SEQUOIA* outperformed HE2RNA with a median EMD of 0.103 across the 92 slides, compared to 0.121 with HE2RNA (Supplementary Table 15). Genes that validate the most slides include *YWHAZ, DCN, TMSB10,* which have been related in literature to breast cancer aggressiveness and survival[45–47] (Supplementary Fig. 10, Supplementary Table 16).

Finally, the integrative analysis of single-cell RNA-seq and spatial transcriptomics data in recent studies have revealed that cells sharing the same transcriptional subtype are often co-localized within spatially segmented niches[5,48]. 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)[49]. Spatial correlation analyses showed that genes within the same meta module consistently clustered together, exhibiting similar spatial expression patterns (Fig. 5d, e).

To demonstrate the spatial prediction capacity of *SEQUOIA* in other cancer types, we developed a 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 (WSIs). We combined algorithmic and methodological advancements, followed by thorough analyses of gene functions, clinical relevance, and generalization capacity. Through a comprehensive evaluation of our model in sixteen cancer types across fourteen 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 genetic information from medical images. He et al.[50] developed ST-Net, a convolutional neural network that predicts the expression values of 250 genes from histology images in breast cancer. Their model, however, cannot integrate contextual information across tiles and is trained on individual tiles, which requires high-resolution training labels obtained from spatial transcriptomics assays.

To avoid training on individual tiles, methods are being developed that first extract tile features for the entire slide, which then go through an aggregation function before calculating the output. Graziani et al.[51] perform this aggregation by calculating a weighted average, where the weights are determined by an attention mechanism that determines the importance of tiles in the average. These attention weights are calculated for each tile individually based on the tile's features and, hence, do not take into account contextual relationships. In addition, their strategy requires training a dedicated model for predicting the expression of each individual gene which can lead to computational challenges, particularly when attempting to infer the entire transcriptome. Schmauch et al. developed HE2RNA[21], which contains an MLP aggregation mechanism and is able to predict the entire transcriptome. However, their MLP aggregation also cannot model contextual tile interactions. Then, Alsaafin et al. introduced tRNAsformer, where the aggregation function is implemented with a transformer encoder. The transformer contains self-attention weights that allow us to model these contextual relations across tiles. They apply this for gene expression predictions and subtype classifications in renal cell carcinoma[22]. However, the increased complexity of the transformer (quadratic in terms of input tokens) may lead to overfitting, especially with limited training data[52]. Finally, the mentioned studies use CNNs (ResNet/DenseNet) pre-trained on the ImageNet dataset for tile feature extraction, which does not contain medical images.

To address these shortcomings, we introduce two alternative components in the WSI prediction pipeline. Instead of using a CNN pre-trained on ImageNet, we use the recent advanced UNI[24] foundation model that was specifically developed for WSI feature extraction. In addition, we introduce an alternative tile aggregation method where we linearize the self-attention component in the transformer architecture, enabling contextual representations at linear complexity. We thoroughly benchmark the added value of each component by comparing UNI vs. ResNet (pre-trained on ImageNet) feature extraction across the three aggregation methods mentioned above: MLP aggregation (as in HE2RNA), transformer aggregation (as in tRNAsformer) and our linearized transformer model. The results of our analysis revealed consistent improvements across various types, both when using UNI instead of ResNet, and when using the linearized transformer versus the other two options.

The genes with accurately predicted expression values by *SEQUOIA* were found to be associated with key pathways pertinent to cancer progression. 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 272-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, the integration of gene expression analysis into clinical routines is hindered by the lack of necessary equipment and trained personnel. By leveraging *SEQUOIA*'s predictions, one can gain mechanistic insights linking histopathological phenotypes to molecular characteristics, thereby offering guidance for disease classification, prognostication, and treatment planning.

Understanding spatial topological organization of tumor cells has attracted recent research interest in the field. The advancement of spatial transcriptomics and proteomics technologies has deepened our understanding of the intrinsic signaling environment that drives tumour growth, metastasis and treatment sensitivity. While *SEQUOIA* was trained using bulk RNA gene expression, we demonstrated its potential in predicting gene expression patterns at the loco-regional level. We implemented a technique that enables computational reconstruction of high-resolution spatial gene expression within tumor tissues. The results were validated using two spatial transcriptomics datasets from independent cohorts of patients with glioblastoma and breast adenocarcinoma. Notably, genes with accurate spatial expression predictions include those regulating malignant phenotype and prognosis. Applying such computational method to WSIs can bring significant values 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[19]. In research, this approach enables the cost-efficient exploration of gene expression dynamics at high resolution, which allows us to generate hypotheses about signaling events driving cellular interactions, thereby advancing our understanding of the complex mechanisms underlying cancer progression.

Future efforts will be dedicated to further improving the model's performance by pretraining on large-scale, multi-center data cohorts, exploring the benefit from color normalization, and providing uncertainty measurement using ensemble or bootstrapping methods. The accurate prediction of molecular traits from histology holds immense potential to improve cancer diagnosis and prognosis, provide valuable insights into a tumour's aggressiveness and its molecular characteristics, advance our understanding of cancer heterogeneity, and enable personalized and targeted therapies. To this end, the implementation of AI-based 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 biological functions, clinical relevance, and generalization capacity, our research demonstrates the potential of using transformer-based deep-learning models in predicting high-dimensional gene expression features from whole-slide histology images.

Methods

Patient cohorts and ethics

TCGA. For model training and evaluation, we retrieved anonymized, paraffin-embedded (FFPE) WSIs and matched gene expression data from the publicly available TCGA archive (https://portal.gdc.cancer.gov). We focused on 16 cancer types within this cohort: (1) BLCA, (2) BRCA, (3) COAD, (4) GBM, (5) HNSC, (6) KIRC, (7) KIRP, (8) LIHC, (9) LUAD, (10) LUSC, (11) PAAD, (12) PRAD, (13) SKCM, (14) STAD, (15) THCA, (16) UCEC.

Among these cancer types, seven have independent patient populations available in the CPTAC cohort for external validation (detailed below). Our models were trained on diagnostic slides of tumor tissues, while the adjacent normal tissues were excluded. The number of patients, WSIs and genes in each cancer type are listed in Supplementary Table 1.

CPTAC. For validation, anonymized patient data were retrieved from the publicly available Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohort (https://portal.gdc.cancer.gov). We downloaded matched WSIs and gene expression data from seven cancer types from six tissues, including BRCA, LUAD, LSCC/LUSC, COAD, kidney renal clear cell carcinoma (CCRCC/KIRC), GBM, pancreatic adenocarcinoma (PDA/PAAD). The sample size is described in Supplementary Table 2.

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

Spatial GBM, Spatial BRCA, 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)[5]. Spatial transcriptomics and matched histology images of BRCA were obtained from Jaume et al.[44] (https://doi.org/10.48550/arXiv.2406.16192).

Data of the SCANB and METABRIC breast cancer cohorts were obtained from published studies by Staaf et al.[39] and Curties et al.[40].

Pre-processing of RNA-Seq data

For the 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.

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.

Pre-processing of WSIs

WSIs were acquired in *SVS* format and downsampled to ×20 magnification (0.5 µm px^{-1}). We used the Otsu threshold method to obtain a mask of the tissue, which allows us to omit tiles mostly containing white background[53]. WSIs have much larger dimensions than natural images (usually over 10k × 10k pixels), and therefore cannot be used directly to train machine learning models. To address this challenge, we employed a multiple instance learning (MIL) approach, where each WSI was cropped into non-overlapping tiles of 256 × 256 pixels (128 µm × 128 µm).

In each slide, we randomly selected a maximum of $N = 4000$ tiles, omitting those containing more than 20% background and tiles with low contrast. To obtain a feature representation, we first organized the selected tiles into bags. Then, we either used a ResNet-50 module pre-trained on ImageNet, or UNI[24] to covert each tile into a feature vector (respectively either 1 × 2048 or 1 × 1024 feature dimension).

Then, we used the k-means algorithm to cluster similar tiles of each slide into $K = 100$ clusters. 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. Features of patches within the same cluster were averaged, resulting in a matrix of 100 × 2048 (ResNet) or 100 × 1024 (UNI) vectors that represent the slide.

SEQUOIA architecture and benchmarked variations

As described above, the WSI pre-processing step results in 100 feature vectors which are obtained either with ResNet or UNI. We represent

the dimension of each feature as D (2048 for ResNet, 1024 for UNI). The dimension of the output (number of genes) is defined as $N = 20{,}820$. To combine the 100 feature vectors into a slide-level representation, we compared three approaches described below. In the figures and text, we refer to the first one as 'MLP aggregation', the second as 'transformer aggregation' and the last one as 'linearized transformer aggregation'. The linearized transformer, in combination with UNI features is the *SEQUOIA* architecture.

The first approach to aggregate the 100 feature vectors into a slide-level representation is to feed each feature into a dense neural network (MLP). Then, the transformed feature vectors are combined by an aggregation procedure that is equivalent to a weighted average. We implemented this procedure according to HE2RNA[21].

In the second option, we model the contextual relationships across the 100 features by implementing a transformer encoder (also used in tRNAsformer[23]). The transformer encoder uses self-attention mechanisms to model the contextual relationships across feature vectors, enabling it to determine the relevance of these relationships for slide-level gene expression prediction. Hereto, the feature matrix from each WSI is fed to a transformer encoder, comprising of 6 encoder blocks, 16 attention heads, and a head dimension of 64. After layer normalization, the output is sent to an MLP head with dimension $D \times N$.

As the third option, we implement a linearized version of the self-attention component within the transformer architecture[54]. We can represent the input to the self-attention module as a vector $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ of $T = 100$ feature vectors of dimension D, which are transformed into a hidden representation $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$. To calculate the hidden representation, self-attention calculates a score e_{ij} for each combination of input vectors $\mathbf{x}_i$ with $\mathbf{x}_j$ (quadratic complexity w.r.t the number of input features). Instead, in the linearized version, we compute a summary of the input vectors $\tilde{\mathbf{s}}$ (linear complexity) that is passed to each input vector. Formally, the calculation of the hidden representation is represented in equation (1), with $\oplus$ representing concatenation. The functions f_1, f_2, f_3 are implemented as a dense linear layer (dimension 64) followed by GeLU activation. All other components of the transformer were kept the same as in the regular version, i.e. the number of blocks (6), attention heads (16) and head dimension (64) were kept the same. After layer normalization, the output is sent to an MLP head with the same dimensions as in the regular transformer ($D \times N$).

$$\mathbf{h}_t = f_3(f_2(\mathbf{x}_t) \oplus \tilde{\mathbf{s}}), \quad \tilde{\mathbf{s}} = \frac{1}{T}\sum_{i=1}^{T} f_1(\mathbf{x}_t) \tag{1}$$

Training and evaluation on the TCGA dataset

We trained a dedicated model for each cancer type using data of the TCGA cohort. For training and evaluating the model, we conducted a five-fold cross-validation (Supplementary Fig. 1). In each fold i, the dataset was partitioned at the patient level, allocating 80% for 'global' training and 20% for testing. The model was trained exclusively on the training set and then independently applied to the test set. To determine the optimal stop point for training the model in fold i, we further split the 'global' training set i into 90% for training and 10% for internal validation. We used the Mean Squared Error (MSE) as the loss function during training, with each model being trained for a maximum of 200 epochs.

For early stopping and determining the point for model saving, instead of relying solely on the PCC as described in Schmauch et al.[21], we employed a criterion that considers both MSE and correlation. Specifically, we continued training and saving the model at each optimal MSE point as long as the MSE continued to decrease. Once the MSE stopped improving for a consecutive *patience* interval of 20 epochs, we continued training the model if the correlation had improved in the last *patience* epochs and if the MSE remained below a reasonable threshold (i.e., $MSE < \delta + bestMSE$, *with* $\delta = 0.5$). We then saved the model at the optimal epoch if the correlation had improved (i.e., $corr > best\_corr$).

Throughout this process, we used a fixed learning rate of 1×10^{-3} and a batch size of 16. The model parameters were optimized with the Adam optimizer. For Pearson correlation and RMSE analyses, we concatenated predictions on patients from all test sets i ($i = 1..5$), which allowed us to leverage the predictive strength of the entire cohort.

Identification of significantly well-predicted genes

To assess the performance within the TCGA cohort, we concatenated the predictions of all test sets i ($i = 1..5$). For each gene, the predicted gene expression values were compared to the ground truth using both Pearson's correlation analysis and RMSE. The resulting correlation coefficient and RMSE values were then compared to those obtained with a random, untrained model of the same architecture.

To identify genes with significantly well-predicted expression levels, we combined three criteria: (1) The correlation coefficient (r_1) between ground truth and the predicted gene expression values must be positive and the associated P value (p_1) should be less than 0.05 ($r_1 > 0$ and $p_1 < 0.05$); (2) r_1 must be significantly higher than r_2 ($r_1 > r_2$) as determined by the Steiger's Z test, where r_2 represents the correlation coefficient between ground truth and predicted gene expression values obtained from the random model. We required the raw Steiger P value to be less than 0.05 ($p_2 < 0.05$) and the adjusted P value by Benjamini-Hochberg correction to be less than 0.2 ($p_3 < 0.2$); (3) The RMSE values obtained from the trained model must be smaller than those from the random model ($rmse1 < rmse2$).

For better presentation and interpretation, we calculated a "normalized RMSE value" using a two-step method. First, since the absolute gene expression values varied across different genes, we performed quantile normalization of the RMSE values. For each gene, the RMSE value between the prediction and ground truth was divided by the interquartile range of its absolute expression values across the samples. This normalization ensured that the RMSE values were comparable between different genes. Second, we performed min-max normalization for the quantile-normalized RMSE values across all genes calculated in each specific cancer type. This step scaled the quantile-normalized RMSE values to a range between 0 and 1. Therefore, the final normalized RMSE value is bounded between 0 and 1, with smaller values indicating more accurate predictions.

Gene set analysis

The gene set analysis was performed with the ClusterProfiler R library (version 4.2.1)[55] and GSEApy package (version 1.0.5)[56]. 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. 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. GSVA was performed to assign an enrichment score to each pathway based on the ground truth or predicted gene expression values using the GSEApy package (version 1.0.5).

Identification and validation of the prognostic gene signature

To construct a gene expression model for predicting breast cancer recurrence, we first selected the top 5000 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)[57]. 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 (2):

$$\text{risk score} = \sum_{i=1}^{n} C_i \times \text{Exp}_i \tag{2}$$

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.

Recurrence-free survival prediction from histology images

To predict recurrence-free survival directly from histology images, we changed the dimension of the fully connected layer of the MLP head from $2048 \times num\_genes$ to 2048×1, and a model was trained to predict a risk score for each patient in the TCGA test set. To predict time to recurrence, we applied the Cox proportional hazards model to the feature vector obtained from the transformer encoder. The hazard function was $\lambda(t|Z) = \lambda_0(t)exp(Z \cdot \beta)$, where Z represents the linear feature vector output from the transformer encoder, $\lambda_0(t)$ the baseline hazard function, and β the coefficient weight implemented in the fully connected layer. The model was trained to minimize Cox loss[6]:

$$L(\beta|Z) = -\sum_{i|C_i=1}\left(Z_i\beta - \log\left(\sum_{j|Y_j \geq Y_i} e^{Z_i\beta}\right)\right) \tag{3}$$

where Z_i represents the feature of patient i, Y_i the recurrence-free survival time, and C_i the censored indicator. The model was trained for 50 epochs with a learning rate of 1×10^{-3} and a batch size of 64. To generate a fair comparison, we kept the input tile features from each WSI consistent with the model used for gene expression prediction.

Spatial gene expression prediction at tile level

To predict gene expression levels spatially at tile-level, we implemented a sliding-window method. Starting from the left upper corner of the histology image, we generate a window of 10×10 tiles, equivalent to the number of features used as input for training *SEQUOIA* on the TCGA dataset. Hence, the geometric location of a window can be defined as (x, y), the 2D coordinate of the top-left point of the top-left tile within the window. The window is intially placed at coordinate $(0, 0)$, referred to as $w_{0,0}$. The x coordinate increases when the window moves to the right, and the y coordinate increases when it moves below. The feature vector of dimension 1×2048 is extracted from each tile as described for our pre-processing of WSIs. At each step, the 100×2048 feature vectors of tiles in the window $w_{x,y}$ are fed to the model. The resulting predicted gene expression $g_{w_{x,y}}$ is assigned to all tiles within the window $w_{x,y}$. To resolve the gene expression at single-tile level, we saved the prediction at each individual step. 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 all tiles within the new window. When the window has reached the end of a row, a new window is started at position *stride* below the previous row ($w_{0,y+stride}$). After the window has passed the entire histology image, 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 set $stride = 1$. 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 µm per spot, which is higher than the predicted resolution of 256 µm 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 and 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[58]). 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. In the context of spatial gene expression maps, EMD considers not only the similarity of values between the ground truth and prediction at individual points but also the spatial arrangement and distribution of those values across the tissue. Hence, this metric takes into account the *spatial context* to determine how well the prediction map corresponds to the ground truth. In contrast, a pixel-level metrics like PCC/RMSE calculate the difference between the predicted and ground truth values at each pixel independently, without explicitly considering the spatial relationships between pixels.

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[49]. We included all signature genes from these modules, except for those ($N = 18$ genes) not included in our training process. The neural-progenitor-like and oligodendrocyte-progenitor-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 colour to each row and column in the matrix, indicating the meta module each gene belongs to.

Reporting summary

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

Data availability

Anonymized WSIs, gene expression and clinical data of TCGA cohorts were retrieved from the publicly available Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov). Gene expression data of the 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. Gene expression data and WSIs of the Tempus cohort were obtained through a data transfer agreement with Tempus Labs, Inc. The publicly available spatial transcriptomics data of GBM were acquired from Datadryad using the following accession URL[5]. Spatial transcriptomics and matched histology images of BRCA were obtained from Jaume et al.[44] (https://doi.org/10.48550/arXiv.2406.16192). The RNA-seq data and clinical annotations of the SCANB[39] cohort were obtained from the accession URL, and data from the METABRIC[40] cohort was obtained for cbioportal with accession URL. Source data for all figures/tables in this work are provided as a zipped folder (including main text and supplementary). Each file within the folder is named according to the figure/panel it belongs to. Source data are provided with this paper.

Code availability

Codes for data pre-processing, model training and evaluation were deposited into a public GitHub repository (https://github.com/gevaertlab/sequoia-pub, release tag v1.0.0, https://doi.org/10.5281/zenodo.13821496)[59].

References

1. Hausser, J. & Alon, U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nat. Rev. Cancer* **20**, 247–257 (2020).
2. Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543 (2014).
3. Vasaikar, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049 (2019).
4. Zheng, Y., Luo, L., Lambertz, I. U., Conti, C. J. & Fuchs-Young, R. Early dietary exposures epigenetically program mammary cancer susceptibility through igf1-mediated expansion of the mammary stem cell compartment. *Cells* **11**, 2558 (2022).
5. Ravi, V. M. et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655 (2022).
6. Zheng, Y., Carrillo-Perez, F., Pizurica, M., Heiland, D. H. & Gevaert, O. Spatial cellular architecture predicts prognosis in glioblastoma. *Nat. Commun.* **14**, 4122 (2023).
7. Chawla, S. & Rai, P. Gene expression based inference of cancer drug sensitivity. *Nat. Commun.* **13**, 5680 (2022).
8. Arora, R. & Chandarana, S. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. *Nat. Commun.* **14**, 5029 (2023).
9. Zheng, Y., Jun, J., Brennan, K., Gevaert, O. Epimix is an integrative tool for epigenomic subtyping using DNA methylation. *Cell Rep. Methods* **3**, 100515 (2023)
10. 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*. https://www.biorxiv.org/content/10.1101/064279v9 (2016).
11. Coudray, N. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
12. Kather, J. N. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
13. Kather, J. N. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
14. Bilal, M. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit. Health* **3**, 763–772 (2021).
15. Noorbakhsh, J. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat. Commun.* **11**, 6367 (2020).
16. Fu, Y. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
17. Chen, M. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol.* **4**, 14 (2020).
18. Liao, H. et al. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clin. Transl. Med.* **10**, e102 (2020).
19. Pizurica, M. Whole slide imaging-based prediction of tp53 mutations identifies an aggressive disease phenotype in prostate cancer. *Cancer Res.* **83**, 2970–2984 (2023).
20. Jiang, S., Zanazzi, G. J. & Hassanpour, S. Predicting prognosis and idh mutation status for patients with lower-grade gliomas using whole slide images. *Sci. Rep.* **11**, 16849 (2021).
21. Schmauch, B. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020).
22. 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. *Commun. Biol.* **6**, 304 (2023).
23. Dosovitskiy, A. et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv* https://arxiv.org/abs/2010.11929 (2020).
24. Chen, R. J. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
25. Thennavan, A. et al. Molecular analysis of tcga breast cancer histologic types. *Cell Genomics* **1**, 100067 (2021)
26. Cao, L. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184**, 5031–5052 (2021).
27. Krug, K. et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436–1456 (2020).
28. Wang, L.-B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528 (2021).
29. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225 (2020).
30. Satpathy, S. et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348–4371 (2021).
31. Clark, D. J. et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* **179**, 964–983 (2019).
32. Hänzelmann, S., Castelo, R. & Guinney, J. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinform.* **14**, 1–15 (2013).
33. Kashimura, S. et al. Cd83+ dendritic cells and foxp3+ regulatory t cells in primary lesions and regional lymph nodes are inversely correlated with prognosis of gastric cancer. *Gastric Cancer* **15**, 144–153 (2012).
34. Fuchs, C. S. et al. Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (regard): an international, randomised, multicentre, placebo-controlled, phase 3 trial. *Lancet* **383**, 31–39 (2014).
35. Syed, Y. Y. Oncotype dx breast recurrence score®: a review of its use in early-stage breast cancer. *Mol. Diagn. Ther.* **24**, 621–632 (2020).
36. Slodkowska, E. A. & Ross, J. S. Mammaprint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev. Mol. Diagn.* **9**, 417–422 (2009).

37. Sestak, I. et al. Prognostic value of endopredict in women with hormone receptor–positive, her2-negative invasive lobular breast cancer. *Clin. Cancer Res.* **26**, 4682–4687 (2020).
38. Nielsen, T. O. et al. A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. *Clin. Cancer Res.* **16**, 5222–5232 (2010).
39. Staaf, J. 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**, 94 (2022).
40. Curtis, C. et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
41. Hong, X. et al. Role of col6a2 in malignant progression and temozolomide resistance of glioma. *Cell Signal.* **102**, 110560 (2023).
42. Wang, H., Mao, X., Ye, L., Cheng, H. & Dai, X. The role of the s100 protein family in glioma. *J. Cancer* **13**, 3022 (2022).
43. Chang, Y.-C. et al. Ppar-γ agonists reactivate the aldoc-nr2f1 axis to enhance sensitivity to temozolomide and suppress glioblastoma progression. *Cell Commun. Signal.* **22**, 266 (2024).
44. Jaume, G. et al. Hest-1k: a dataset for spatial transcriptomics and histology image analysis. *arXiv* https://arxiv.org/abs/2406.16192 (2024).
45. Hu, X. et al. Decorin-mediated suppression of tumorigenesis, invasion, and metastasis in inflammatory breast cancer. *Commun. Biol.* **4**, 72 (2021).
46. Mei, J. et al. Ywhaz interacts with daam1 to promote cell migration in breast cancer. *Cell Death Discov.* **7**, 221 (2021).
47. Zhang, X. et al. Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. *Breast Cancer Res.* **19**, 1–15 (2017).
48. Ren, Y. et al. Spatial transcriptomics reveals niche-specific enrichment and vulnerabilities of radial glial stem-like cells in malignant gliomas. *Nat. Commun.* **14**, 1028 (2023).
49. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
50. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
51. Graziani, M. et al. Attention-based interpretable regression of gene expression in histology. In: *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, pp. 44–60 https://arxiv.org/abs/2208.13776 (2022).
52. Variš, D. & Bojar, O. Sequence length is a domain: length-based overfitting in transformer models. *arXiv* https://arxiv.org/abs/2109.07276 (2021).
53. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
54. Parcollet, T., Dalen, R., Zhang, S., Bhattacharya, S. SummaryMixing: a linear-complexity alternative to self-attention for speech recognition and understanding. *arxiv* https://arxiv.org/abs/2307.07421 (2024).
55. Wu, T. et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
56. Fang, Z., Liu, X. & Peltz, G. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics* **39**, 757 (2023).
57. Simon, N., Friedman, J., Tibshirani, R. & Hastie, T. Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
58. Bradski, G. The OpenCV Library. *Dr. Dobb's J.* **120**, 122–125 (2000).
59. Pizurica, M., Carrillo-Perez, F., Zheng, Y. Gevaertlab/sequoia-pub: V1.0.0. https://doi.org/10.5281/zenodo.13821496 (2024).

Acknowledgements

The research reported here was supported by the National Cancer Institute (NCI) under award: 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, a grant from the Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) 1161223N and FWO V467423N. F. Carrillo-Perez was also supported by a predoctoral scholarship from the Fulbright Spanish Commission. The work was further supported by grants of the FWO (3G045620, 3G046318) and UGent BOF (BOF 01J06219, BOF/IOP/2022/045BOF). We are grateful for Roche Information Solutions (RIS) sponsorship, encouragement and support for this project.

Author contributions

M.P., F.C.P., Y.Z., and H.N. performed data pre-processing. M.P. and F.C.P. conceived and developed model architectures. Y.Z. performed validation on independent cohorts and conducted pathway/gene expression analyses. Y.Z. developed and validated the signature for recurrence prediction. M.P. developed and validated spatial prediction. W.Y. and C.W. contributed through discussions and suggestions. O.G., K.M., and A.V. conceived the study. O.G. and K.M. jointly supervised the work. M.P. and Y.Z. wrote the manuscript with contributions and/or revisions from all authors.

Competing interests

W.Y., C.W., and A.V. are employees of F. Hoffmann-La Roche Ltd. The remaining authors have no conflicts of interest to declare.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54182-5.

Correspondence and requests for materials should be addressed to Olivier Gevaert.

Peer review information *Nature Communications* thanks Ali Bashashati, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.