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	1,130	1,059
TCGA-LUAD	25,812	536	473
TCGA-LUSC	26,443	510	476
TCGA-GBM	26,530	237	102
TCGA-PRAD	25,587	448	401
TCGA-PAAD	26,172	202	176
TCGA-KIRP	25,141	299	275
TCGA-KIRC	26,653	514	508
TCGA-COAD	23,645	455	447
total		4,331	3,917

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 normal tissue in GTex.

GTex project	number genes	number WSIs	number patients
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		1,802	1,802

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 = 1130)	1,429	2,902	12,239	11,069
LUAD (N = 536)	413	601	9,747	8,759
KIRC (N = 514)	0	66	6,985	10,086
LUSC (N = 510)	22	6	5,269	4,919
COAD (N = 455)	2,348	464	6,090	7,740
PRAD (N = 448)	973	2,124	8,895	10,825
KIRP (N = 299)	626	336	8,884	7,905
GBM (N = 237)	983	1,303	4,208	2,498
PAAD (N = 202)	314	387	418	3,378

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	0.120	0.200	0.327	0.282
LUAD	0.116	0.117	0.284	0.269
KIRC	0.090	0.108	0.229	0.300
LUSC	0.097	0.088	0.199	0.206
COAD	0.223	0.141	0.268	0.310
PRAD	0.128	0.205	0.321	0.303
KIRP	0.164	0.135	0.368	0.333
GBM	0.188	0.205	0.289	0.242
PAAD	0.169	0.164	0.173	0.265

Table A6: Median RMSE values between prediction and ground truth in TCGA test set within each model.

	he2rna_scratch	he2rna_pretrain	sequoia_scratch	sequoia_pretrain
BRCA	0.559	0.476	0.456	0.449
LUAD	0.543	0.488	0.471	0.452
KIRC	0.576	0.457	0.379	0.386
LUSC	0.635	0.524	0.451	0.456
COAD	0.613	0.555	0.422	0.413
PRAD	0.693	0.388	0.352	0.348
KIRP	0.702	0.472	0.432	0.445
GBM	0.727	0.492	0.389	0.420
PAAD	0.906	0.695	0.376	0.395

Table A7: 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	11 (0.4%)	8,587 (78%)
Lung squamous cell carcinoma	LUSC (LSCC)	0 (0.0%)	3,560 (72%)
Kidney renal clear cell carcinoma	KIRC (CCRCC)	0 (0.0%)	7,259 (72%)
Lung adenocarcinoma	LUAD	1 (0.1%)	5,330 (61%)
Glioblastoma multiforme	GBM	0 (0.0%)	1,464 (59%)
Colon adenocarcinoma	COAD	1 (0.2%)	3,941 (51%)
Pancreatic adenocarcinoma	PAAD (PDA)	1 (0.3%)	1,177 (35%)

Table A8: 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

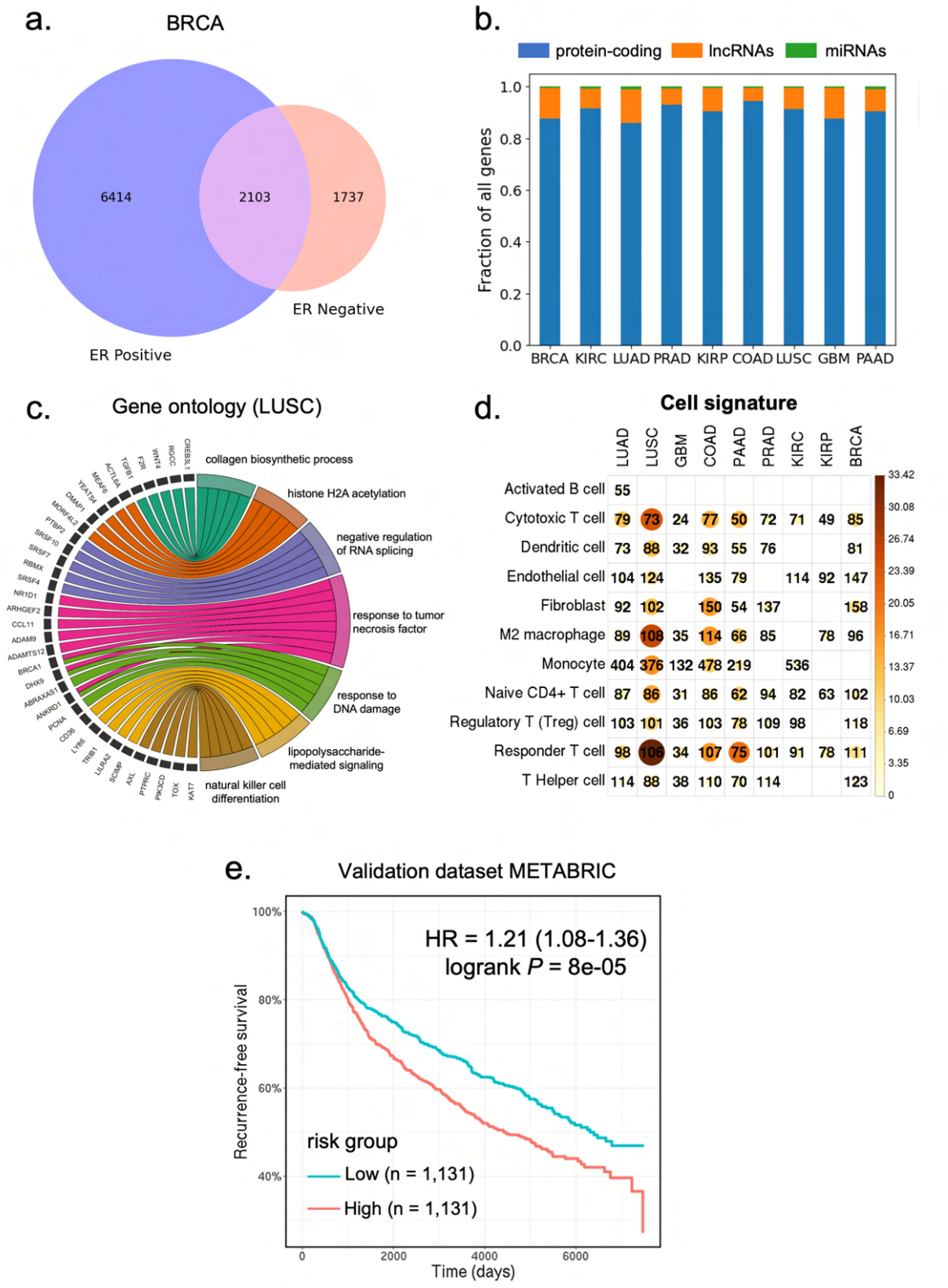

Fig. A1: **Characterization of the well-predicted genes.** a) Venn diagram showing the number of well predicted genes in the estrogen-receptor (ER) positive and ER negative breast cancer. b) The proportion of protein-coding genes, miRNAs and lncRNAs among the well predicted genes from each cancer type. c) Circos plot showing the biological processes associated with the well predicted genes in LUSC. d) 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. P values were adjusted for multiple testing using the Benjamini–Hochberg method. e) 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.

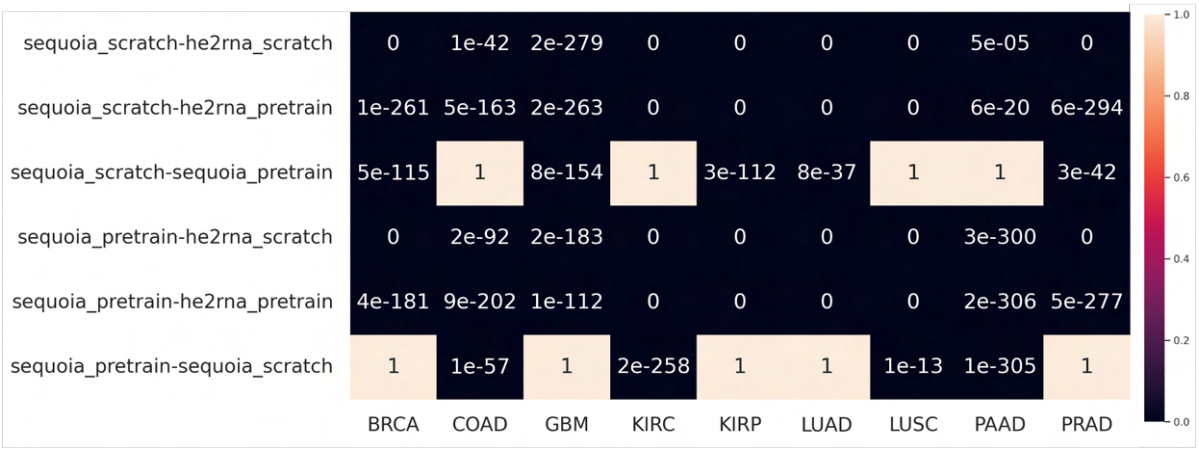

Fig. A2: ***P* values for testing the distributions of correlation coefficients for the top 1,000 most accurately predicted genes obtained from each model**. A P value is calculated for each pairwise comparison using a one-sided Mann-Whitney U test for the hypothesis that model x is larger than model y, formatted on the left axis as x-y.

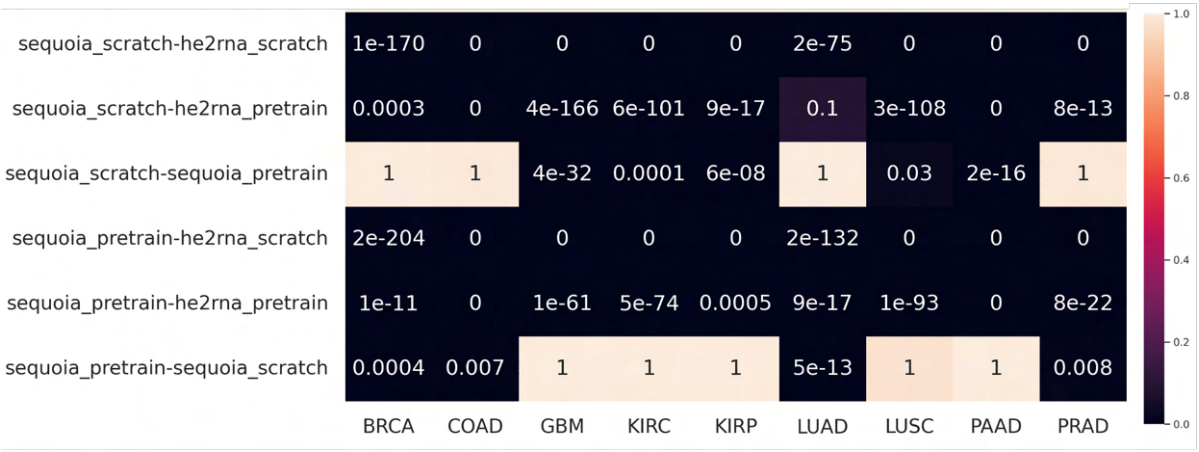

Fig. A3: ***P* values for testing the distributions of RMSE values between each of the two models.** A P value was calculated for each pairwise comparison using a one-sided Mann-Whitney U test for the hypothesis that model x is smaller than model y, formatted on the left axis as x-y.

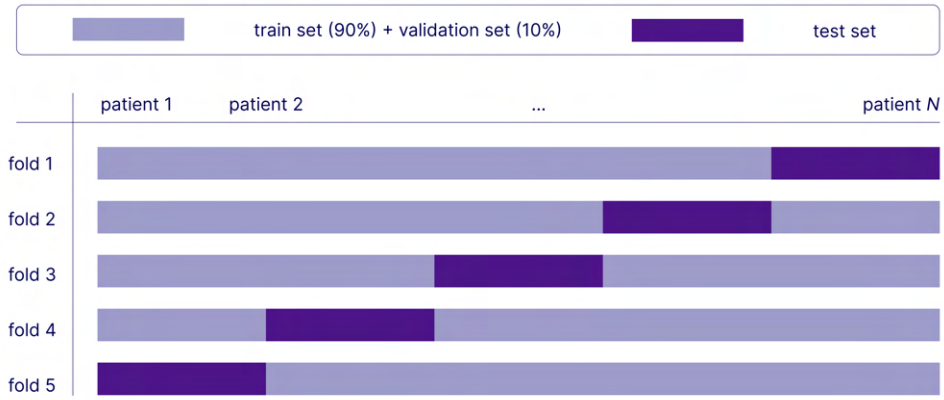

Fig. A4: **Data splitting.** First, five folds are made each of which consist of a 'global' train and test set. The 'global' train set is further split into a train (90%) and validation (10%) set. In each fold i, validation set i is used to determine the optimal point to stop training model i, which 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).

References

[1] Hausser, J., Alon, U.: Tumour heterogeneity and the evolutionary trade-offs of cancer. Nature Reviews Cancer **20**(4), 247–257 (2020)

[2] Network, C.G.A.R., *et al.*: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543 (2014)

[3] 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)

[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**(16), 2558 (2022)

[5] 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)

[6] 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)

[7] Chawla, S., Rockstroh, A., Lehman, M., Ratther, E., Jain, A., Anand, A., Gupta, A., Bhattacharya, N., Poonia, S., Rai, P., *et al.*: Gene expression based inference of cancer drug sensitivity. Nature communications **13**(1), 5680 (2022)

[8] 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)

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

[10] 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)

[11] 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)

[12] 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)

[13] Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.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. The Lancet Digital Health **3**(12), 763–772 (2021)

[14] 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)

[15] 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)

[16] 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)

[17] 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)

[18] 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)

[19] 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)

[20] 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)

[21] 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)

[22] 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)

[23] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)

[24] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)

[25] 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)

[26] 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)

[27] 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)

[28] 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)

[29] 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)

[30] Hänzelmann, S., Castelo, R., Guinney, J.: Gsva: gene set variation analysis for microarray and rna-seq data. BMC bioinformatics **14**, 1–15 (2013)

[31] 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)

[32] 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)

[33] 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)

[34] 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)

[35] 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)

[36] 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)

[37] 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)

[38] 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)

[39] 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)

[40] 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)

[41] 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)

[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] Variš, D., Bojar, O.: Sequence length is a domain: Length-based overfitting in transformer models. arXiv preprint arXiv:2109.07276 (2021)

[50] 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)

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

[52] 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)

[53] 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)

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

[55] 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

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