Inferring Spatial Genomic Expression profiles from H&E-Stained Histology Using Weakly Supervised Deep Learning in Breast Cancer

Shweta S. Chavan[1], PhD; Chao Feng[1], PhD; Hassan Muhammad[1], PhD; Hirak S. Basu[1], PhD; Wei Huang[1,2], MD; Rajat Roy[1], MS; George Wilding[1], MD; Gordon Mills[3], MD, PhD; Shivaani Kummar,[1,3] MD, FACP

[1]Pathomiq Inc., Cupertino, CA, USA, [2]Department of Pathology, University of Wisconsin-Madison, Madison, WI, USA, [3]Knight Cancer Institute, Oregon Health Science University, Portland, OR, USA

Introduction

- Genomic markers such as ESR1, PGR, EGFR, MKI67, and FOXC1, along with clinically validated multigene signatures like Oncotype DX, MammaPrint, Prosigna, and EndoPredict, are widely used to inform prognosis and guide therapeutic decisions in breast cancer.
- However, the widespread use of these gene expression-based assays is often constrained by their high cost and limited accessibility .
- Moreover signatures such as OncotypeDX are not performed spatially, potentially diluting signal. Conklin, M. W., & Keely, P. J. (2012)
- To address this challenge, we developed an AI model capable of inferring the spatial distribution of gene expression directly from routine digitized H&E-stained whole slide images (WSIs).

Methodology

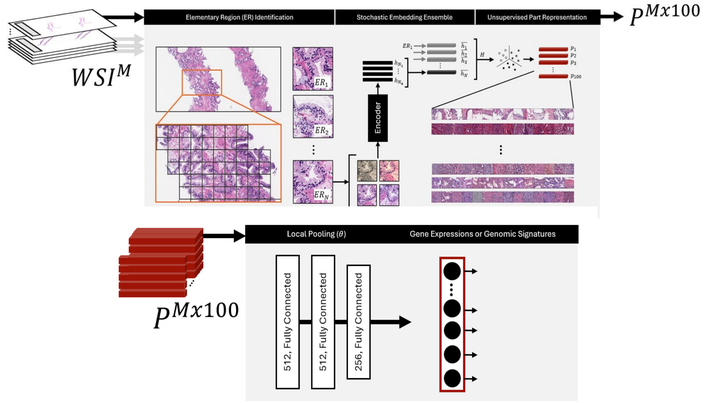


Figure 1: (Top) Each WSI is converted to a "part representation," a high dimensional representation of unique phenotypes present in the WSI. **(Bottom)** Part representations are used as input data into an MLP to predict gene expressions and signatures.

- We trained a weakly supervised multiple instance learning model using a publicly available dataset (TCGA), containing paired bulk RNA-Seq and H&E WSIs across multiple disease types.
- M WSIs were divided in elementary regions (ER) and these ER were converted to a high dimensional vector (h) using a foundation model, following a test-time augmentation protocol. Afterwards, ERs within a slide were clustered to generate phenotype-level part representations (P). MIL aggregation was used to predict slide-level expressions. Attention weights were visualized to simulate an expression heatmap.
- Although trained solely on bulk tissue-level expression data, the model accurately reconstructs spatial gene expression patterns, offering a scalable and cost-effective alternative to spatial transcriptomics.
- This pan-cancer model was trained to predict expression levels of thousands of gene expressions, including known breast cancer-relevant genes such as ESR1, PGR, FOXA1, AURKA, BIRC5, MELK, MYBL2, PLK1, CDC20, CCNE1, and those in Mammaprint, and OncotypeDx.
- Once trained, the model was predicted on a held-out validation dataset of breast cancer WSIs to measure slide-level prediction performance using Pearson's correlation.
- Spatial predictions were validated on held-out WSI datasets with ground truth data.

Results

- Our model accurately predicted the expression of over 2,500 genes with a Pearson correlation greater than 0.6.
- Prediction of Oncotype signature has an AUC of 0.83
- Notably, breast cancer relevant genes such as ESR1, FOXA1, and AURKA showed particularly high correlation between predicted expression values and ground truth bulk RNA-seq measurements.
- Among these, we qualitatively compared the predicted spatial expression patterns of PD-L1, CD163, and CD8 with corresponding IHC-stained tissue sections. The regions of high predicted expression were aligned with areas of high marker density in the IHC images.
- Additionally, comparison with spatial transcriptomic data from 10x Genomics demonstrated strong concordance between our model's spatial predictions and the ground truth measurements, particularly for breast cancer-relevant genes.

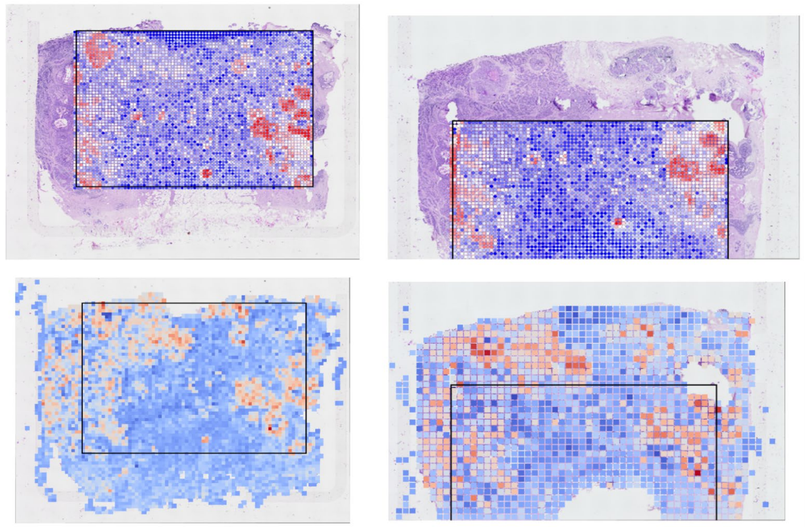


Figure 2: (Top) Spatial Expression of ESR1 based on single cell RNA-Seq data (Bottom) Prediction of ESR1 expression on WSI using AI.

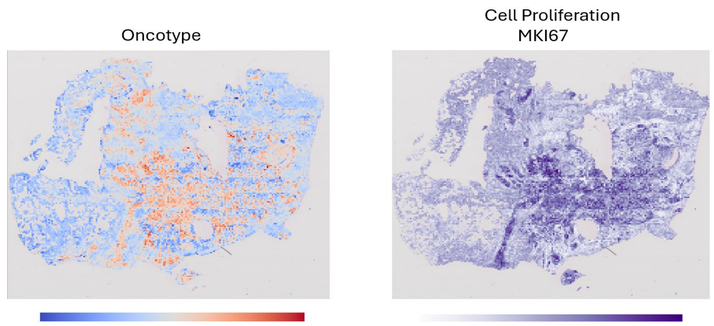


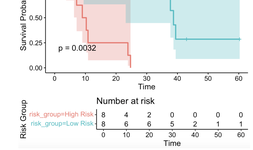
Figure 3: Spatial prediction of ODX correlates with spatial prediction of MKI67, a key marker in the proliferation signature of ODX.

Results

Marker	Performance
SKA1	0.71 (Corr)
CDC20	0.68 (Corr)
FANCA	0.68 (Corr)
KIF2C	0.67 (Corr)
MELK	0.67 (Corr)
FOXC1	0.57 (Corr)
TRMT9B	0.50 (Corr)
PEBP4	0.51 (Corr)
GIPR	0.47 (Corr)
Mammaprint	0.74 (Corr)
Oncotype DX	0.83 (AUC)
Internally Developed ER Hormone Marker	0.28 HR (p-value 0.001)
Internally Developed ER Chemo Marker	2.42 HR (p-value 0.014)

Table 1: Prediction accuracy of selected gene expressions and genomic signatures.

- In a late stage TNBC cohort we showed that AI based gene expression is predictive of response to Olaparib and Durvalumab
- GIPR, PEBP4, TRMT9B, and FOXC1 among 15 other genes were predicted to be associated with PFS



Conclusion

- Our results demonstrate that routine H&E-stained whole slide images can be leveraged to accurately infer both bulk and spatial gene expression using a weakly supervised AI model.
- This capability enables large-scale, retrospective spatial analysis of archival tissue and opens new avenues for understanding tumor heterogeneity in breast cancer using standard pathology slides.
- By providing gene-level insights without the need for additional staining or molecular assays, this approach may augment current histopathological evaluation and support more informed clinical research, risk stratification, and treatment planning.

Funding & COI

This research was funded by PATHOMIQ Inc.

Contact

Hassan Muhammad
PATHOMIQ Inc
Email: hassan@pathomiq.com
Website: www.pathomiq.com

This presentation is the intellectual property of the author/presenter. Contact them at hassan@pathomiq.com for permission to reprint and/or distribute.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com