Uncovering Molecular Insights in Breast Cancer Prognostics Utilizing Foundation Model and Digital Pathology

AACR 2025
Presentation #2433

Shlomit Strulov Shachar[1], Tal Etan[1], Didi Feldman[1], Ariella Deutsch-Lukatsky[1], Lee Ben-Itzhak[2], Inbar Gilboa[3], Inbal Gazy[3], Ron Sinai[3], Addie Dvir[4], Reva Basho[5], Amir Sonnenblick[1]*, Dov Hershkovitz[2]*

[1] Oncology Division, Tel Aviv Sourasky Medical Center, Israel; [2] Institute of Pathology, Tel Aviv Sourasky Medical Center, Israel; [3] Imagene Ltd., Israel; [4] Imagene Inc., USA; [5] Ellison Medical Institute, USA. *Equal contributors

Background

Breast cancer (BC) is a leading cause of cancer-related mortality among women. There is a need to discover novel biomarkers that more accurately predict patient prognosis and recurrence risk, which in turn can inform treatment strategies and improve outcomes.

Here, we developed an AI model for the prediction of early BC distant disease recurrence risk directly from hematoxylin and eosin (H&E)-stained pathology slide images. We then applied it on an external cohort from TCGA to extract insights from gene expression data.

Methods

Model development

- Whole-slide images (WSIs) of H&E-stained tissue from breast cancer cases were collected from Sourasky Medical Center.
- The training cohort included HR+ HER2− cases with either early distant recurrence (within 5 years of diagnosis; n=19 (6.6%)) or without any recurrences and at least 6 years of follow-up (table 1) .
- Embeddings were extracted from the WSIs using a version of Imagene's foundation model, CanvOI[1]. A 5-fold cross-validation approach was then used to develop the early distant recurrence risk model.

Biomarker Discovery

- TCGA breast cancer WSIs (n = 1038) were analysed with the established model, assigning a recurrence prediction score for each sample.
- Upper and lower thresholds were applied to the prediction scores, categorizing roughly (due to rounding adjustments) the top and bottom 10% of samples as AI high-risk (n=100) and AI low-risk (n=124) respectively.
- Differential gene expression analysis was performed (using bulk RNA expression data) between the groups.

Table 1. Training cohort characteristics

Characteristic	Total (n = 287)
Female, no. (%)	287 (100)
Age	
Median yr (range)	61 (25-75)
≤50 yr, no. (%)	53 (18.47)
Nodal status, n (%)†	
N0	207 (72.13)
N1	80 (27.87)

† Pathological stage

Table 2. TCGA cohort characteristics

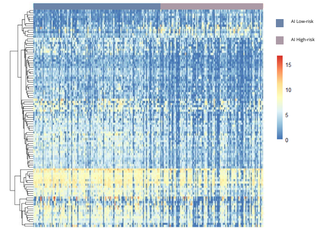
Characteristic	Total (n = 1038)
Female, no. (%)	1026 (98.84)
Age	
Median yr (range)	59 (26-90)
≤50 yr, no. (%)	310 (29.87)
Nodal status, n (%)	
N0	507 (48.84)
N1	348 (33.53)
N2	111 (10.69)
N3	72 (6.94)
Receptors status, n (%)†	
HR+ HER2-	575 (55.39)
HER2 positive	161 (15.51)
Triple Negative	148 (14.26)
N/A*	154 (14.84)

* HER2 status and/or HR status is equivocal/unknown
† Pathological stage

Results

Analysis of the TCGA breast cancer AI high- vs. AI low-risk groups identified 299 differentially expressed genes (DEGs) defined according to $|\text{Log}_2$ fold change$|>1$ and adjusted p-value <0.05.

Figure 1. Expression Heatmap of Top 100 DEGs



* Ranked by absolute fold change

Figure 2. Differential expression analysis for AI high- vs. AI low-risk

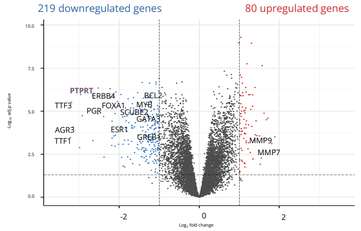
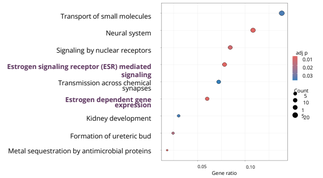
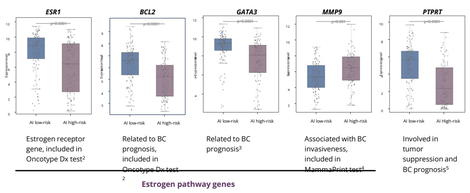


Figure 3. Pathway enrichment analysis



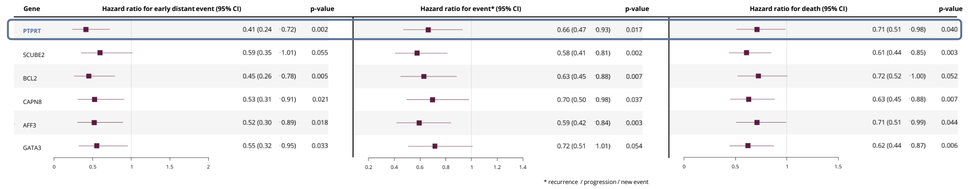
Reactome pathway enrichment analysis revealed significant dysregulation of estrogen signaling pathways (adjusted p<0.05), known to be associated with BC outcomes.

Figure 4. Selected examples of differentially expressed genes



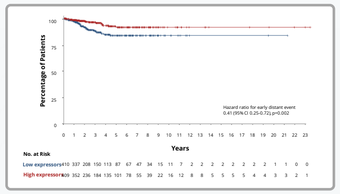
Shown here are expression levels of estrogen pathway genes, known to be associated with breast cancer outcomes and recurrence, in AI low- and high-risk groups.
Also included is *PTPRT*, a less-studied gene, and a potential biomarker for BC recurrence.

Figure 5. Early distant event free survival, event free survival and overall survival of selected DEGs



- Genes shown include the top 20 DEGs and estrogen receptor–related genes with statistically significant or borderline associations ($p < 0.05$) across all three outcomes, sorted by their average p-value. For each gene, patients were divided into high and low expression groups based on the median expression value.
- An HR < 1 indicates that lower gene expression is associated with increased risk of event.
- ***Notably, reduced expression of PTPRT was consistently associated with poorer outcomes across all evaluated outcomes,*** *with the lowest average p-value among all analyzed genes.*

Figure 6. *PTPRT* expression is associated with early distant event free survival



Kaplan–Meier curves for TCGA samples with early distant event data demonstrate clear separation between low (blue) and high (red) *PTPRT* expression groups (stratified by the median expression level).

Results Summary

Using an AI classification model, we identified *PTPRT* as a potential biomarker of interest for breast cancer recurrence. Although the role of *PTPRT* in breast cancer is not well characterized, our findings are consistent with existing evidence linking lower expression of *PTPRT* to shorter overall and recurrence-free survival[5]. Further studies, including analyses across various HER2 and HR groups, are warranted to clarify its role in breast cancer recurrence.

Discussion

We present a proof-of-concept study demonstrating that an AI-based recurrence model trained on clinical H&E-stained slides can be effectively applied to an external cohort to stratify it into AI low- and high-risk groups.

References

1. Zalach J. *et al.* arXiv, 2024. 2409.02885
2. Sparano J. and Paik S. J. *Clin. Oncol.* 2008, 26(5):721-728
3. Yoon N.K. *et al. Hum. Pathol.* 2010, 41(21):1794-1801
4. Tian S. *et al. Biomark. Insights*. 2010, 5:129-138
5. Li L. *et al. Biomed. Res. Int.* 2021, 3301402

Acknowledgements

The results here are based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

Correspondence: addie@imagene-ai.com