Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression–Morphology Analysis in Breast Cancer

Yinxi Wang[1], Kimmo Kartasalo[1,2], Philippe Weitz[1], Balazs Acs[3,4], Masi Valkonen[5], Christer Larsson[6], Pekka Ruusuvuori[2,5], Johan Hartman[3,4,7], Liang Zhang[8], Liuliu Pan[8],Kathy Ton[8],Yan Liang [8] and Mattias Rantalainen[1,7]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,Stockholm, Sweden. [2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [3]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. [4]Department of Clinical Pathology and Cytology,Karolinska University Laboratory, Stockholm, Sweden. [5]Institute of Biomedicine,Cancer Research Unit and FICAN West Cancer Centre, University of Turku and Turku University Hospital, Turku, Finland. [6]Division of Translational Cancer Research, Department of Laboratory Medicine, Lund University, Lund, Sweden.[7]MedTechLabs, BioClinicum, Karolinska University Hospital, Solna, Sweden.[8] Nanostring Technologies, Inc, Seattle, USA

Correspondence: Mattias Rantalainen mattias.rantalainen@ki.se

Introduction

Molecular profiling is central in cancer precision medicine but remains costly and is based on tumor average profiles. Morphologic patterns observable in histopathology sections from tumors are determined by the underlying molecular phenotype and therefore have the potential to be exploited for the prediction of molecular phenotypes. We report here the first transcriptome-wide expression–morphology (EMO) analysis in breast cancer, where individual deep convolutional neural networks were optimized and validated for prediction of mRNA expression in 17,695 genes from hematoxylin and eosin–stained whole slide images. Predicted expressions in 9,334 (52.75%) genes were significantly associated with RNA sequencing estimates. We also demonstrated successful prediction of an mRNA-based proliferation score with established clinical value. The results were validated in independent internal and external test datasets. Predicted spatial intratumor variabilities in expression were validated through spatial transcriptomics profiling. These results suggest that EMO provides a cost-efficient and scalable approach to predict both tumor average and intratumor spatial expression from histopathology images.

Significance: Transcriptome-wide expression morphology deep learning analysis enables prediction of mRNA expression and proliferation markers from routine histopathology whole slide images in breast cancer.

Methods

Data Collection

The study consists of female patients with breast cancer from three data sources: Clinseq-BC (N = 270), The Cancer Genome Atlas (TCGA-BC; N = 721), and ABiM (N = 350) as an external validation cohort. Images from Clinseq-BC and TCGA-BC were randomly split into training (N = 558, 56.30%; 4.08 million H&E tiles), tuning(N = 139, 14.03%; 0.97 million H&E tiles), validation (N = 122,12.31%; 0.90 million H&E tiles), and test sets (N = 172, 17.36%; 1.33 million H&E tiles). Transcriptome-wide RNA-seq data representing mRNA expression for a total of 20,477 genes in the reference genome are collected from these samples.

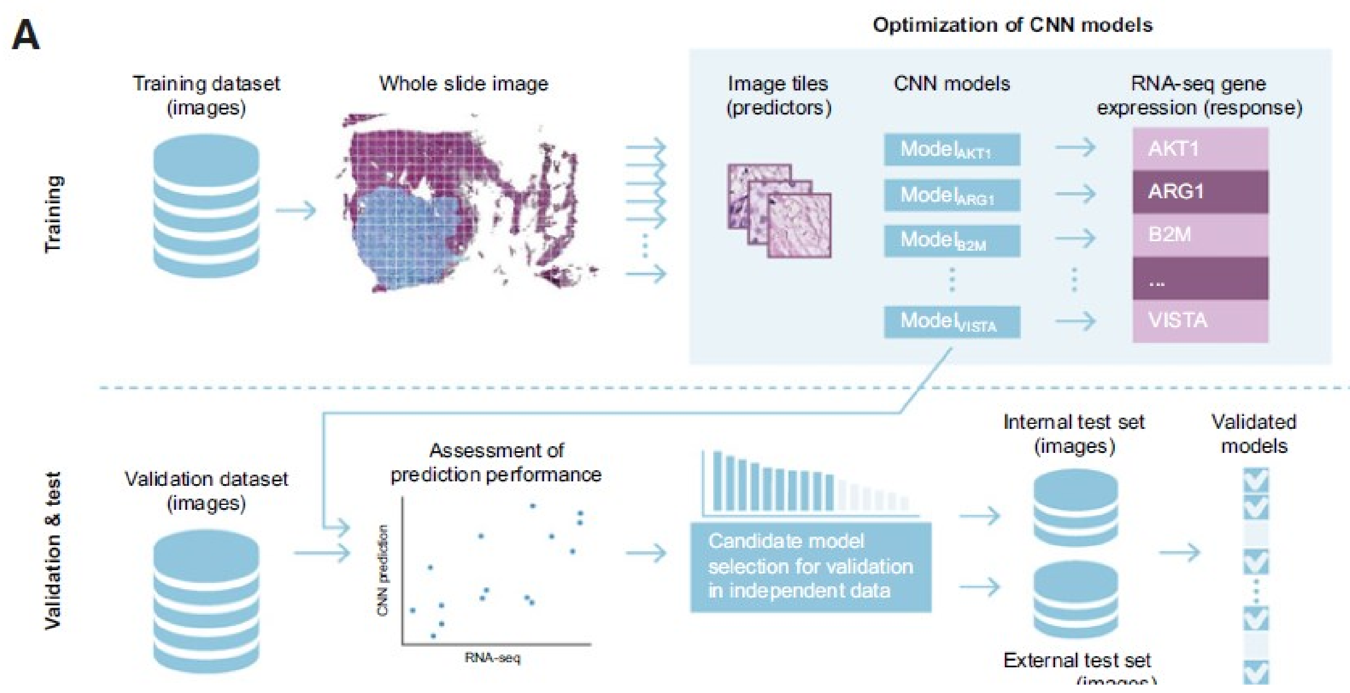
Data Modeling

For each gene, we optimized one deep convolutional neural networks (CNN) model with image tiles as predictors and the sample-level gene expression level obtained from RNA-seq as a response variable.

Model Validation

From an additional independent collection of 168 tumors with both FFPE blocks and WSIs available, 24 tumors were selected for ST profiling using the oncology and immune-oriented gene panel for the GeoMx® DSP platform (GeoMx Immune Pathways Panel, NanoString Technologies).

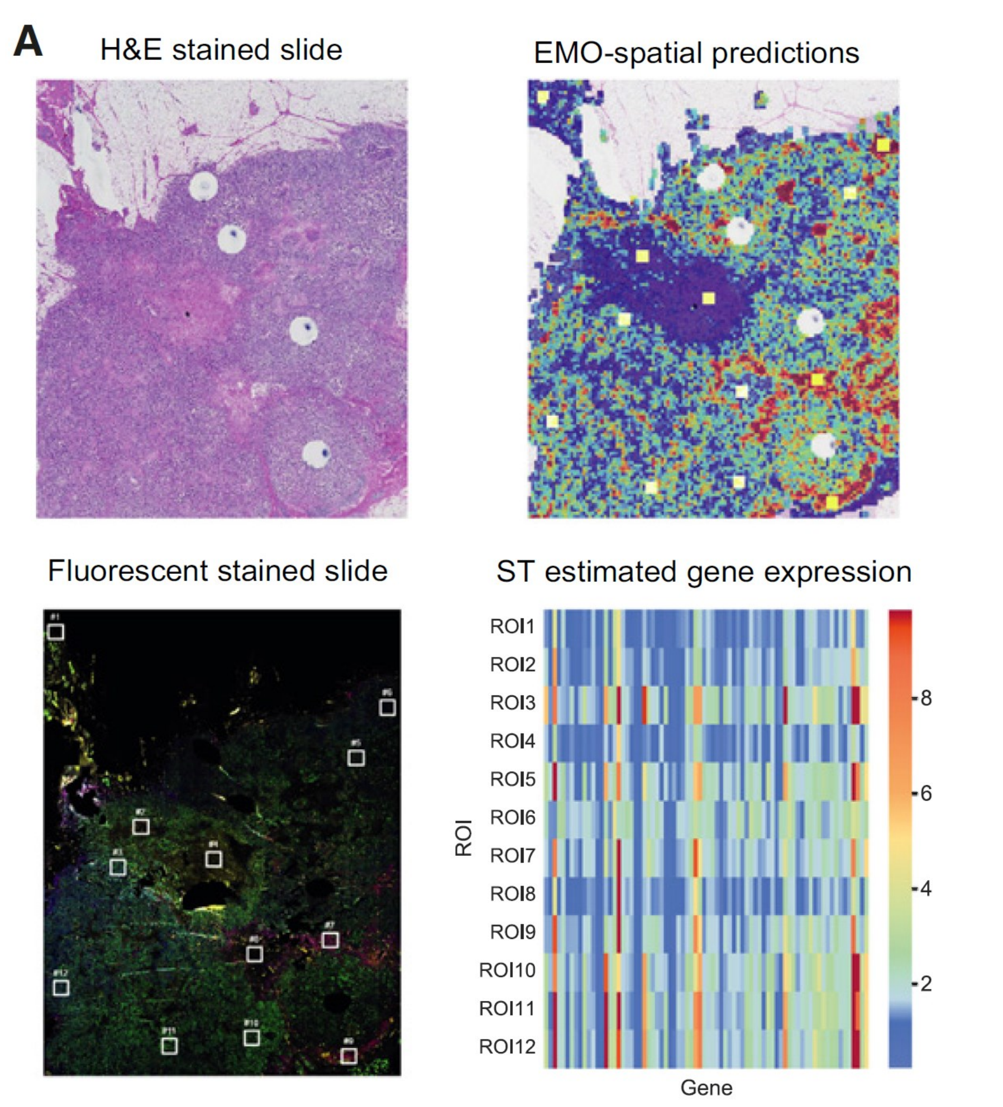
Study Design


A
Optimization of CNN models
Training
Training dataset (images)
Whole slide image
Image tiles (predictors)
CNN models
RNA-seq gene expression (response)
Model_AKT1
Model_ARG1
Model_B2M
Model_VISTA
AKT1
ARG1
B2M
...
VISTA
Validation & test
Validation dataset (images)
Assessment of prediction performance
CNN prediction
RNA-seq
Candidate model selection for validation in independent data
Internal test set (images)
External test set (images)
Validated models

Study design and summary

Statistics for transcriptome-wide predictions. A, Overview of the EMO process. In the training phase, training WSIs (N = 697)were split into image tiles. The tiles (predictors) together with expression levels (response) across the protein coding transcriptome were used to optimize individual deep CNN models (Inception V3) for each gene. All optimized models were then applied to predict expression in WSIs in the validation set (N = 122), association analysis between RNA-seq estimated gene expression values and predicted values was performed, and candidate models were selected for further validation. The validation was performed in the internal (N = 172) and external (N = 350) test sets.

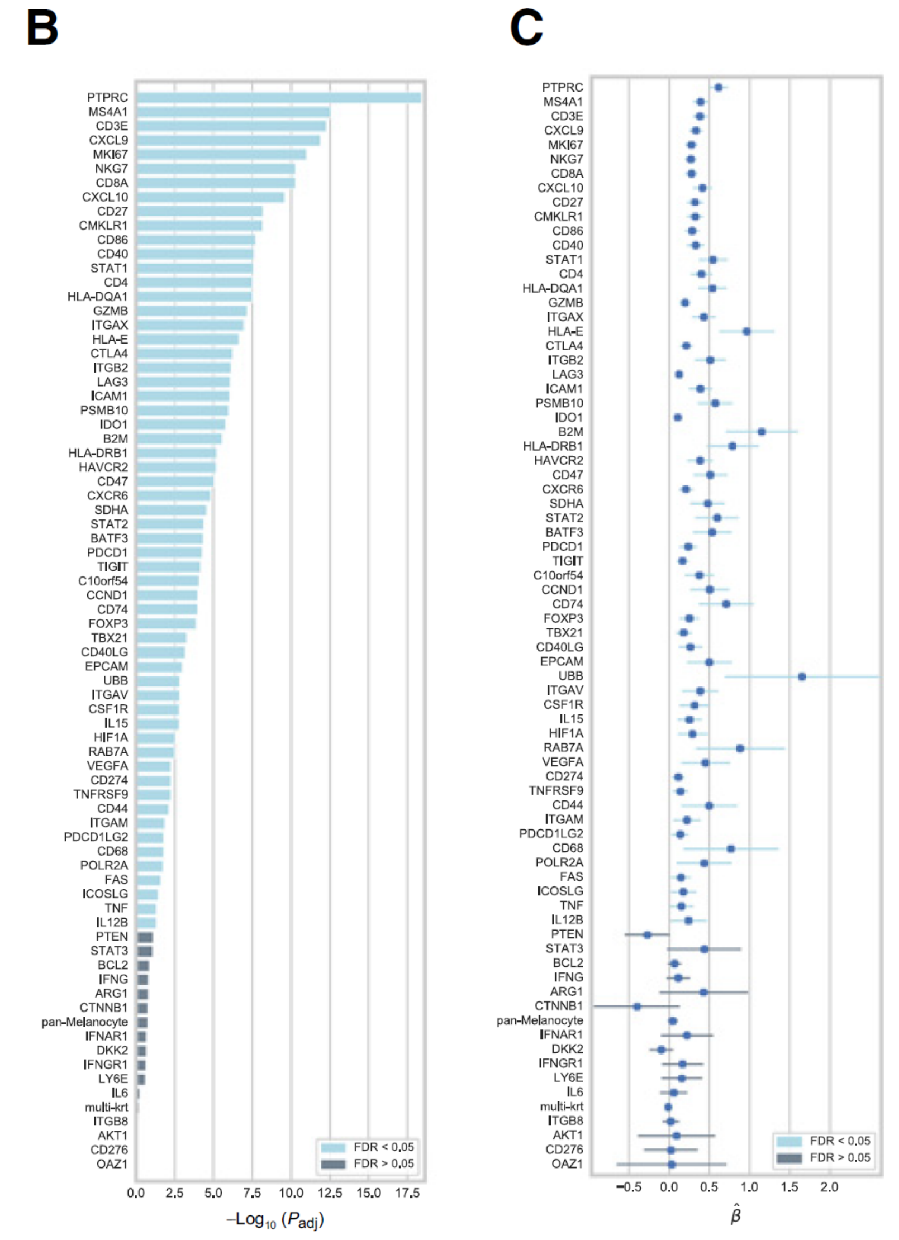
Results



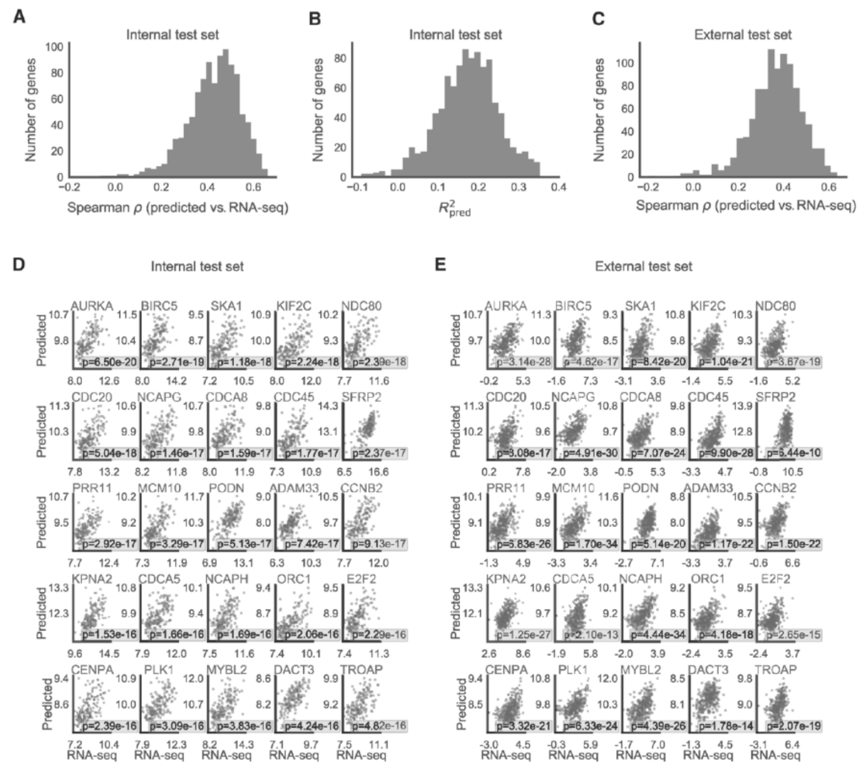
ST validation of spatial expression predictions.

A, Overview of the ST profiling process. For each WSI (top left), optimized CNN models for the genes in the ST gene panel were used to predict spatial (tile-level) expression, visualized as heatmaps. Twelve ROIs (yellow squares) were subsequently manually selected to obtain are a presentative set of regions including low, medium, and high predicted expression across a range of genes (top right). The ROIs from each slide were then manually registered against fluorescently labeled slides from consecutive FFPE sections (bottom left). ST profiling of the ROIs was performed and subsequently used to validate spatial EMO prediction results (bottom right).

B, Bar plot for the ranked _log10(FDR-adjusted P value) for genes from each LME model. Light blue indicates FDR-adjusted P <0.05 (NWSIs = 22). C, Corresponding fixed-effect coefficients and 95% CI related to the EMO prediction for each gene (linear mixed effects model;N_{WSIs} = 22).Wang et al.5122



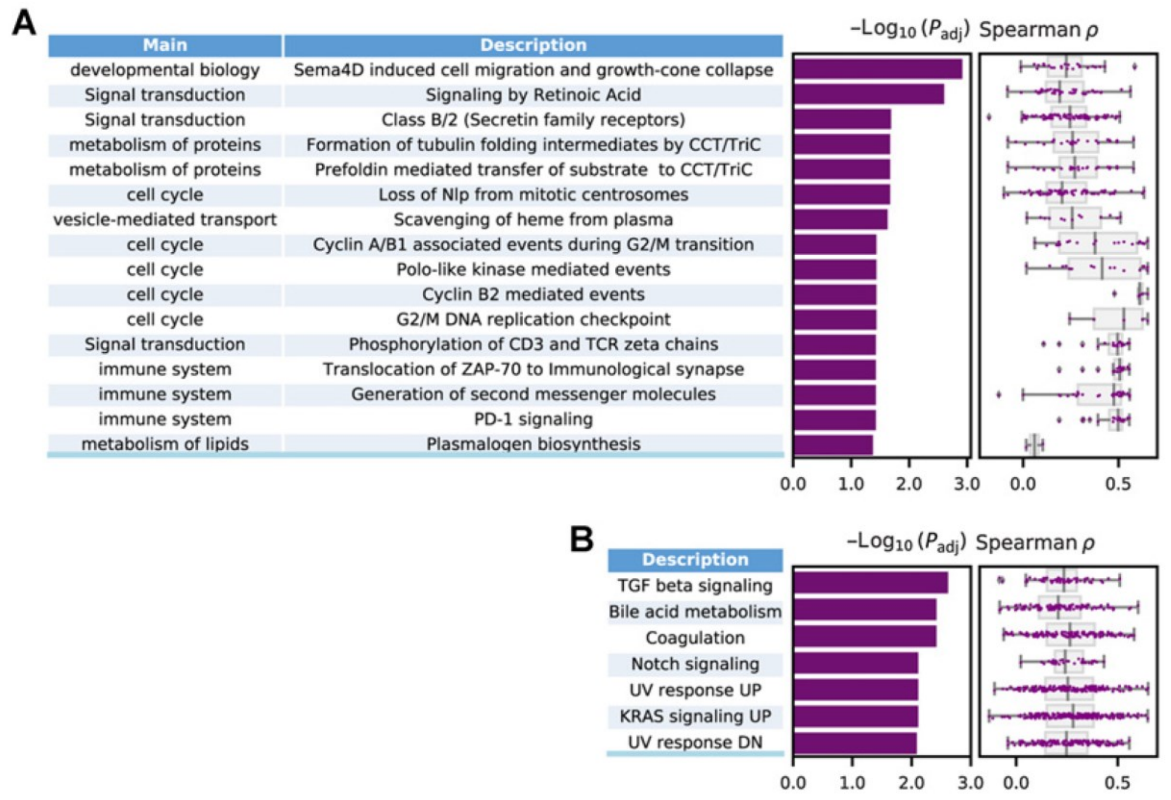
Results



Summary of model performance on test sets

A, Distribution of Spearman's rho in the internal test set. **B,** Distribution of R^2_{pred} in the internal test set (Ngenes = 1,011; onegene with a predicted R^2 <_0.1 was excluded from the figure for clarity). **C**, Distribution of Spearman's rho in the external test set (N_{genes}=995). **D**, Scatter plot of $EMO_{predicted}$ and RNA-seq estimated gene expression values for the 25 top performing genes in the internal test set. **E**, Scatter plot of EMO-predicted and RNA-seq estimated gene expression values for the same 25 genes in the external test set.

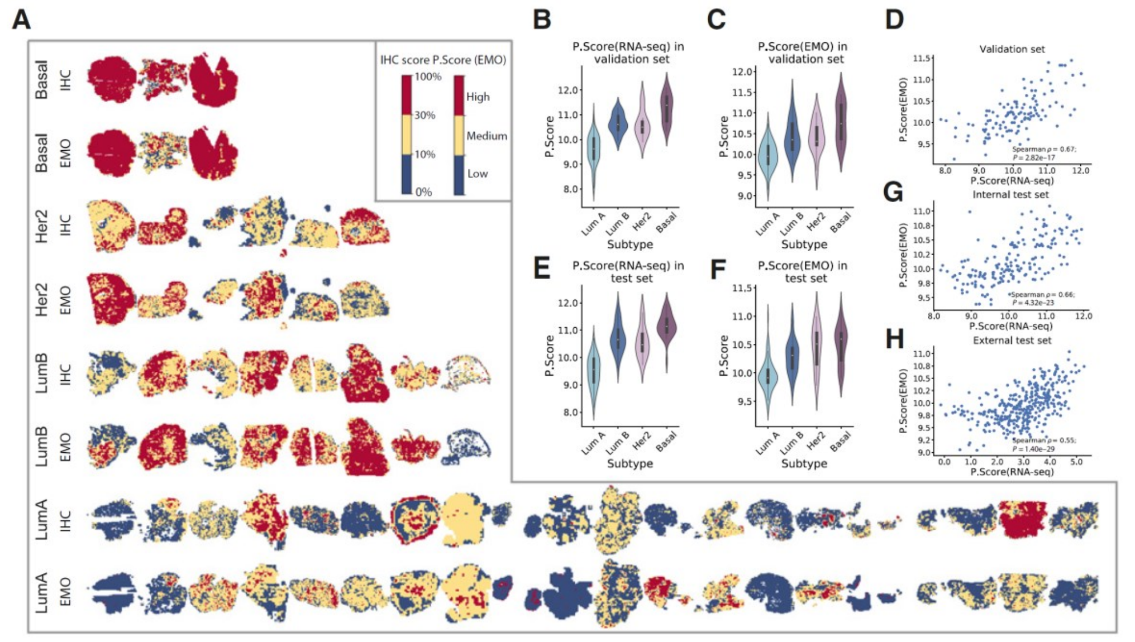
Results



GSEA on whole transcripts

A, Pathway analysis of EMO predictions by GSEA in the Reactome database, revealing 16 significant pathways. The bar plot shows the log-transformed adjusted P values for each pathway, and the boxplot shows the model performance in terms of Spearman's rho between EMO-predicted and RNA-seq expression (validation set) for each gene in each individual pathway. B, GSEA results using the Hallmark gene set, with seven identified pathways.

Results



Proliferation score prediction and validation.

A, Comparison between IHC score and EMO-predicted proliferation score [P.Score(EMO)]) for 37 IHC-HE pairs of tumors in the test set. The IHC-based Ki67 score per tile is indicated in blue (<10%), yellow (≥10% and <30%), and red (≥30%). The color scheme for EMO predictions was chosen based on quantile mapping to the IHC score distribution, with blue, yellow, and red indicating low, medium, and high predicted proliferation levels,respectively. **B,** Distribution of proliferation scores by subtype in the validation set, measured with RNA-seq [P.Score(RNA-seq)]. **C,** Distribution of proliferation scores by subtype in the validation set, predicted by EMO. The distribution of predicted proliferation scores shares similar patterns with RNA-seq measurements, with the basal type exhibiting the highest proliferation level, followed by HER2-enriched (Her2) and luminal B (LumB) subtypes, whereas luminal A (LumA) has the lowest proliferation score. **D,** Scatter plot of RNA-seq–estimated and EMO-predicted proliferation scores in the validation set (N=122).A high correlation between the RNAseq measurements and EMO predictions was observed with a Spearman's rho of 0.67. **E–G,** Corresponds to B–D for the internal test set (N=172). **H,** Scatter plot of RNAseq–estimated and EMO-predicted proliferation scores in the external test set (N = 350).

doi: 10.1158/0008-5472.CAN-21-0482