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], Balázs Ács[3,4], Masi Valkonen[5], Christer Larsson[6], Pekka Ruusuvuori[2,5], Johan Hartman[3,4,7], and Mattias Rantalainen[1,7]

ABSTRACT

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 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.

Introduction

Microscopic morphologic patterns observable in stained tumor tissue are routinely characterized by pathologists to classify and diagnose cancers. General morphology is assessed using hematoxylin and eosin (H&E) staining, while IHC enables semiquantitative assessment of specific markers. However, cancer is a genetic disease where somatic alterations and their interactions with other phenotypic factors and the tumor microenvironment give rise to a complex and dynamic molecular phenotype. Profiling of, for example, somatic DNA alterations, RNA expression, or protein abundances provide a comprehensive characterization of tumors. In breast cancer, the molecular phenotype defined by the mRNA expression profile contains prognostic information (1–4) and defines the intrinsic molecular subtypes (5, 6). Furthermore, mRNA expression profiling also reveals information about cell proliferation, which has clinical value as a prognostic marker and potentially as a predictor of response to systemic therapy (7, 8). Compared with routine pathology, molecular profiling represents a more comprehensive characterization of the individual tumor (9), providing information relevant for precision medicine (10), and information that can contribute to the discovery of novel therapeutic targets and diagnostic markers.

Intratumor heterogeneity is a key contributing factor to emerging treatment resistance, or reduced efficacy of treatment, which is caused by either subclonality or as a consequence of plasticity in the dynamic molecular phenotype of a tumor (11, 12). Tumor evolution and subclonality can be inferred from genetic data. However, the more comprehensive phenotype defined by the mRNA expression profile, and other dynamic molecular phenotypes, is generally acquired from a bulk average mRNA pool where intratumor variability is lost. Single-cell RNA sequencing (RNA-seq; refs. 13, 14) enables profiling of thousands of individual cells, providing unique information to characterize intratumor heterogeneity (15). Although techniques for single-cell sequencing now are mainstream, it remains challenging on primary human samples as fresh samples typically are required. Spatial transcriptomic (ST) profiling (16, 17) is another emerging technology enabling characterization of intratumor heterogeneity, but it is still technically demanding, expensive, and offers low resolution both spatially and in terms of the number of genes that can be detected.

Computational pathology, driven by deep learning–based artificial intelligence applied on digital whole slide images (WSI), has recently emerged and demonstrated human pathologist level performance in cancer detection and classification (18, 19). Deep convolutional neural networks (CNN) have also been applied for prediction of molecular phenotypes from routine formalin-fixed paraffin-embedded (FFPE) H&E-stained sections (19–23). More importantly, this approach also

[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.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Y. Wang and K. Kartasalo contributed equally to this article as co–first authors.

P. Ruusuvuori, J. Hartman, and M. Rantalainen contributed equally to this article as co–senior authors.

Corresponding Author: Mattias Rantalainen, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-171 77 Stockholm, Sweden. Phone: 46-0-8-5248-0000, ext. 2465; E-mail: mattias.rantalainen@ki.se

Cancer Res 2021;81:5115–26

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

enables inference of spatial heterogeneity. For example, He and colleagues trained models on WSIs using spatially measured gene expression value of 250 genes and identified 102 genes that can be successfully predicted (21); In two other studies, features extracted in a pan-cancer setting were applied for prediction of molecular phenotypes (22, 23).

To date, no studies have performed genome-wide and disease-specific analyses. Previously reported studies have not optimized gene-specific models, and comprehensive validation in fully independent data, which is required to establish generalizability, has not been reported.

Here we report the first transcriptome-wide expression–morphology (EMO) analysis in breast cancer using large-scale deep learning and routine H&E WSIs for prediction of mRNA expressions. The study is comprehensive in that individual models were optimized for each gene across the mRNA transcriptome. The results were validated in a fully independent external patient cohort at the gene level. Furthermore, we demonstrate that our CNN models enable prediction of spatial expression patterns, which were validated in independent tumors using ST profiling. Finally, we applied and validated EMO for prediction of an established multivariate proliferation score, demonstrating a clinically relevant application.

Materials and 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 (Supplementary Table S1). For Clinseq-BC and ABiM, H&E-stained FFPE histopathology slides were scanned in-house with a Hamamatsu Nanozoomer XR (Hamamatsu Photonics) at ×40 magnification (0.226 μm/pixel). WSIs from TCGA-BC were downloaded from https://portal.gdc.cancer.gov/. WSIs in TCGA-BC that were scanned at 20× were excluded to ensure image quality. One WSI image was included from each individual. For patients in Clinseq-BC where the same slide had been rescanned, the most recently scanned was used. For the ABiM cohort, when a patient had multiple WSIs, we first chose the one that was selected to perform IHC biomarker analysis in the routine clinical workflow (i.e., the piece that the clinical pathologist indicated as most relevant); if no WSIs had such indication, we chose the one with the largest predicted tumorous area. All included patients have corresponding RNA-seq data available for analysis.

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). The internal and external test sets remained untouched during the model development and training phase and were used only once for final evaluation of model performance at the end of the project.

Data preparation

Image data preprocessing

Each WSIs were tiled into image tiles of 598 × 598 pixels (271 × 271 μm). Sharpness was evaluated for all tiles as a quality assurance step. Next, color normalization (24) was performed to adjust for staining differences across institutions and scanners. Tumor detection and segmentation was then applied to segment invasive cancer regions for subsequent analyses (see Supplementary Materials and Methods).

RNA-seq data preparation

We collected transcriptome-wide RNA-seq data representing mRNA expression for a total of 20,477 genes in the reference genome. For Clinseq-BC and ABiM, RNA-seq, preprocessing, and normalization were performed as described previously (25, 26). For TCGA-BC, RNA-seq data were downloaded from http://cancergenome.nih.gov/ and were preprocessed in the same way as Clinseq-BC (25). Only patients with both RNA-seq and WSIs available were included in the study. In total, 19,112 genes had non-zero gene expression variance. In addition, we hypothesized that genes with close to zero variance are less informative for EMO and the potential for extracting relevant morphologic features gradually decreases with diminishing gene expression variance. The benefit of getting meaningful results is further limited after considering the computational cost of training each model. Hence, we chose to only include genes with a variance larger than 0.01 in further analysis, which resulted in 17, 695 genes as the final training targets.

As Clinseq-BC and TCGA-BC were merged together, to reduce batch effects associated with the RNA-seq, the RNA-seq data from Clinseq-BC were normalized to have median value equal to TCGA-BC. In brief, we first calculate the median expression level of each gene for both TCGA-BC and Clinseq-BC data sources. Only data from the training and validation sets were included in this step. Next, the differences between median values of these two data sources were calculated. Finally, the Clinseq-BC expression values were normalized by genewise addition of the offsets computed in the previous step. TCGA-BC data remained unchanged.

During the testing phase, RNA-seq data in the test sets from Clinseq-BC, TCGA-BC, and ABiM cohorts were all median normalized using the same procedure with TCGA-BC (training and validation data) as a reference.

EMO analysis

Model optimization

For each gene, we optimized one CNN model with image tiles as predictors and the sample-level gene expression level obtained from RNA-seq as a response variable. Inception V3 (27) architecture, modified by replacing the last layer with one neuron and a linear activation, was employed to build a regression model. Tiles from the training and tuning set were used to optimize the model. We employed the Adam (28) optimizer with the mean squared error loss function and default parameters as follows: learning rate $= 1 \times 10^{-6}$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $\varepsilon =$ None and decay $= 0$. Random 90° rotations and flips of the tiles were applied as data augmentation.

We used a minibatch of 32 image tiles per step and ran the optimization for 150 steps per epoch. We sampled 313 mini-batches from the tuning set to assess the validation loss on each epoch, and used early stopping with a minimum change in loss of 0.003 and a patience of 80 epochs, continuing optimization until the early stopping criterion was met or a maximum number of 500 epochs (75,000 steps) had been completed. From each optimization run, we stored the models from the 10 epochs resulting in the best performance on the tuning tiles and 10 models from randomly selected epochs. Depending on when the early stopping criterion was met, the optimization runs took approximately 12 to 70 hours on a single GPU.

Model validation

For each gene, the model with the lowest loss recorded on the tuning set was applied on the combined Clinseq-BC and TCGA-BC validation set. The mean of all tile-level predictions across one slide was used to obtain a patient-level prediction. To evaluate model performance, we calculated the Spearman correlation between the predicted patient-level gene expression values and those measured using RNA-seq across the validation set. The associated *P* values were adjusted for multiple testing using the Benjamini–Hochberg approach (29). In an attempt to measure the proportion of variance that could be predicted using the CNN models, R^2 score was also calculated.

Model selection for testing

To further validate the generalizability of the CNNs, we selected the subset of genes (i.e., models) with predicted R^2_{pred} higher than 0.2 and adjusted *P* value from Spearman correlation lower than 0.001 according to the performance on the validation set. In total, 1,011 genes were brought forward for model testing on the internal test set. Out of these, 995 genes could be matched in the ABiM study (accession no. GSE81538), and were included in the evaluation on the external test data. As the scale is dataset dependent and the values of RNA-seq data can vary due to differences in protocols and profiling platforms, the R^2_{pred} score was not calculated in the external ABiM test set.

The tile-level predictions were postprocessed as described in "Model validation." Bonferroni correction was applied to account for multiple testing.

Gene set enrichment analysis

With the aim of understanding if genes associated with particular molecular mechanisms were enriched among the genes that were predicted well, we conducted a pathway analysis to identify enriched pathways based on the results on the validation set. To do this, instead of arbitrarily selecting a cutoff threshold for assigning significance among all available genes, and performing analysis based on the subset of genes, we considered a rank-based algorithm (30) to avoid potential bias in such selection.

In brief, the 17,695 genes ranked by P_{adjusted} value from the Spearman correlation analysis were used as input. We followed the procedures described in ref. 30 and conducted the analysis with the "SetRank" R package and the Reactome (31) as well as the Hallmark (32) pathway databases for pathway annotations.

For the pathway analysis of transcripts with nonsignificant predictions, since model performance cannot be applied as a ranking criteria, we adopted a set-based approach using the FUMA platform (33) with Reactome and Hallmark gene sets.

ST analysis

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). The 24 slides were selected to have predicted (by the CNN models) spatially varying expression levels, assessed by visual inspection, across a number of genes (*BCL2, CD4, GZMB, HIF1A, HLA-DQA1, ITGB2*, and *VEGFA*). These genes represent a diverse set from the panel in that they belong to different pathways and were also among the best performing genes ($R^2 > 0.15$, $P_{\text{adjusted}} < 0.0001$) in the validation set (EMO-average). Selecting slides exhibiting spatial variability ensured that intratumor variability existed. Regions of interest (ROI; 600 μm × 600 μm) were manually selected from the H&E-stained WSIs based on EMO-spatial predictions, to cover a range of predicted gene expression values across a variety of genes. For each tumor, two consecutive sections were produced. The first section was stained with H&E and used to generate a routine WSI (used for prediction of expression level in the EMO-spatial workflow). The second section was used for ST profiling in a standard workflow for the GeoMX DSP platform. This slide was stained with four fluorescent stains targeting PanCK, SMA, CD45, and DNA to outline general morphologic structures (H&E stains were not an option on the platform). Manual registration of the selected ROIs from the first section (H&E-stained slide and associated EMO-spatial predictions) and the second section was performed, to mark corresponding locations on the second consecutive section for ST profiling. Two slides were damaged during fluorescent staining and discarded, resulting in 22 slides remaining for ST analysis. Finally, gene expression values within each ROI were quantified by the GeoMx DSP platform by counting the unique indexing oligos assigned to each target with the NanoString nCounter instrument. Gene expression values were normalized by dividing each value with the average expression levels across six negative controls, to account for any nonspecific binding, and subsequently $\log_2$ transformed before further analysis.

To estimate the EMO-spatial predictions, we calculated the mean of tile-level predictions within each ROI per gene. The gene panel consists of 84 RNA probes (see Supplementary Table S2 for full list of genes), of which, six served as negative controls, and two (*CCL5* and *PECAM1*) had a variance of gene expression lower than 0.001 and were therefore excluded prior to any further analyses. Furthermore, the probe named "multi-krt" include probes against a group of genes: *KRT18, KRT6B, KRT6C, KRT6A, KRT19, KRT17, KRT7, KRT10*, and *KRT14*; and the probe named "pan-Melanocyte" contains probes against *SOX10, PMEL*, and *S100B*. The predicted gene expression values for these two targets were calculated by summing the predictions for the respective sets of genes.

We then measured the performance of the CNN models by first comparing the predictions with ST-measured gene expression using a linear mixed effect (LME) model with results displayed with a bar plot and a line plot. The model was fitted (maximum likelihood) with the log-transformed ST estimated expression as response, the EMO-spatial prediction as a fixed effect and slide ID as a random effect (accounting for variability between slides). A likelihood ratio test was applied to test the significance of the fixed effect parameter. *P* values were adjusted for multiple testing using the Benjamini–Hochberg method, and FDR < 0.05 was considered significant. Furthermore, for each slide and gene, Spearman ρ was also calculated, between EMO-spatial predictions and ST expression estimates, across 12 ROIs per tumor. The empirical distribution of Spearman ρ estimates across the 22 individuals, and for each gene, was summarized as boxplots.

Proliferation score analysis

To compute the proliferation score, we adopted a proliferation signature that consists of 11 genes from the PAM50 gene panel (34). The proliferation score was defined as the mean expression of these 11 genes.

The association between tile-level EMO-spatial proliferation scores and IHC-stained Ki67 scores was examined using a LME model, with the log-transformed Ki67 score as response, the EMO-spatial prediction as a fixed effect and slide ID as a random effect. Moreover, the slide-level EMO-average predictions were compared with the slide-level estimated proliferation score (P.Score) from RNA-seq data as well as to the log-transformed Ki67 scores in terms of Spearman correlation. The details relating to scoring of IHC slides (Supplementary Table S3), HE-IHC image registration, and analyses of intratumor

spatial proliferation patterns are documented in the Supplementary Materials and Methods.

Software and hardware

All image preprocessing steps were conducted with Python (v. 3.6) packages, including scikit-image (v. 1.14.2), OpenCV (v. 3.4.1), OpenSlide (v.3.4.1 and API v. 1.1.1). Training of CNN models was carried out using Keras (v. 2.2.4) with Tensorflow (35) backend (v. 1.12). Color normalization was performed using Python, adapted from StainTools (https://github.com/Peter554/StainTools) and "Staining Unmixing and Normalization in Python" (https://github.com/schaugf/HEnorm_python). Spearman correlation was calculated using the SciPy package (v. 1.2.0) in Python, R^2 was calculated with Python package scikit-learn (v. 0.20.2). Statistical testing, multiple testing correction fitting were performed with the statsmodels Python package (v. 0.9.0). LME models were fitted using R (3.6.3) with R package "lme4" and "lmerTest." Model training and predictions were run on the GPU partition of the Puhti compute cluster (CSC IT Center for Science), consisting of 80 compute nodes. Each node is equipped with two 20-core Xeon Gold 6230 CPUs (Intel), 384 GB of memory, 4 T V100 32 GB GPU accelerators (Nvidia) and 3.6 TB of local NVME storage. The GPUs were running Nvidia driver version 440.33.01.

Because models for different genes are fully independent of each other, transcriptome-wide training and prediction represent "an embarrassingly parallel problem." We therefore ran each model as a separate computation job on a single V100 GPU, and automated job submission through the SLURM scheduler system, resulting in 50 to 300 models being trained in parallel at any given time over a period of several months. At the beginning of each computation run, the input image tiles were copied from the central file system to the local NVME disk to avoid I/O bottlenecks due to the large number of parallel runs relying on the same data. In parallel with the GPU computation, mini-batches were prepared using multi-threading on two CPU cores to maintain an in-memory data buffer equal in size to two mini-batches.

Results

Study overview

We performed a transcriptome-wide EMO analysis, where individual deep CNN models were optimized separately for each mRNA transcript. RNA-seq was used to quantify the expression of 20,477 individual genes (see Materials and Methods for details). In total, 991 patients (7.28 million H&E tiles) from two studies [TCGA breast cancer (9) and Clinseq-breast (25)], each with one WSI, were included and split into training ($N = 697$, 70.33%), validation ($N = 122$, 12.31%) and internal test sets ($N = 172$, 17.36%) prior to model optimization and validation. The preprocessing of WSIs included segmentation of tissue and invasive cancer, tiling of WSIs into tiles of 598 × 598 pixels (271 μm × 271 μm), quality control for image sharpness, and color normalization to adjust for variations in stains and scanners (see Materials and Methods). In the training set, 17,695 genes remained after excluding transcripts with low variance (see Materials and Methods). For each of the transcripts, a deep CNN model (Inception V3; Supplementary Table S4; ref. 27) was optimized to predict normalized gene expression using images. Models were trained in parallel on a high-performance compute cluster (CSC), with the transcriptome-wide analysis requiring approximately 300,000 GPU hours. The tile-level predictions of each slide were averaged to obtain slide-level predicted expression (EMO-average), which was compared with gene expression measured by RNA-seq. The optimized models were subsequently applied and evaluated in validation and test sets (**Fig. 1A**).

Breast cancer RNA expression can be predicted by deep CNN models from routine histopathology images

In the validation set, of 17,695 genes, the predicted expression of 9,334 (52.75%) genes was significantly correlated with expression levels measured by RNA-seq (Spearman correlation, FDR-adjusted $P < 0.05$; **Fig. 1B** and **C**). Next, we assessed the proportion of variance predicted: 1,026 (5.80%) genes showed a coefficient of determination (R^2_{pred}) higher than 0.2, and 222 (1.25%) and 26 (0.15%) genes had R^2_{pred} higher than 0.3 and 0.4, respectively (**Fig. 1D**). Furthermore, we observed that genes with higher variance had a slightly better prediction performance compared with those with lower expression variance (Supplementary Fig. S1A and S1B).

To establish whether the predicted expression levels were associated with corresponding routine clinical (protein) biomarkers, we visualized the RNA-seq estimated expression, and the EMO-predicted expression by IHC status for each clinical routine marker (ER, PR, HER2, and Ki67). These markers are assessed in the clinic by IHC and are widely adopted to classify breast cancer cases. As shown in the violin plots, patients with a positive status of ER, PR, and HER2 tend to have higher RNA-seq expression levels of the corresponding transcripts. Similarly, a higher level of *MKI67* gene expression is associated with a high histologic grade (**Fig. 1E–H**). The same trends hold for all the model predicted markers except for *ERBB2*, which encodes the HER2 protein, and could not be predicted by the CNN model (**Fig. 1I–L**).

Taken together, these results indicate that morphologic patterns in histopathology images can be learned by deep CNN models and used to predict gene expression for a substantial proportion of genes across the transcriptome.

Validation of gene-specific predictions of expression in independent datasets

To assess the generalizability of the approach, 1,011 genes with $R^2_{\text{pred}} > 0.2$ and FDR-adjusted $P < 0.001$ in the validation set (Supplementary Table S5) were brought forward for validation in the internal ($N = 172$) and external test sets [ABiM study (26), $N = 350$]. A total of 876 (86.65%) genes had a significant association between predicted (EMO-average) and observed (RNA-seq) expression (Bonferroni-adjusted $P < 0.05$, Spearman correlation; **Fig. 2A**). A total of 479 of these genes had an $R^2_{\text{pred}} > 0.2$ (**Fig. 2B**) in the internal test set. A total of 908 (91.26%) genes were successfully validated in the external test set [Bonferroni-adjusted P value (Spearman correlation) <0.05; **Fig. 2C**]. The estimated correlation coefficients (Spearman ρ) between EMO-average prediction and RNA-seq across the 1,011 genes had a high concordance between the validation, internal, and external test sets (Supplementary Fig. S1C–S1E), indicating similar levels of prediction performance across datasets. Concordance between EMO-average predicted and RNA-seq estimated gene expression for the 25 genes with the best prediction performances in the internal test set, ranked by P value (Spearman correlation), are visualized in **Fig. 2D**, with the corresponding results in the external test set in **Fig. 2E**.

EMO prediction performance of transcripts in established gene panels

Next, we assessed the prediction performance for genes belonging to established breast cancer biomarker panels based on gene expression, including Oncotype DX, Prosigna/PAM50, and Endopredict

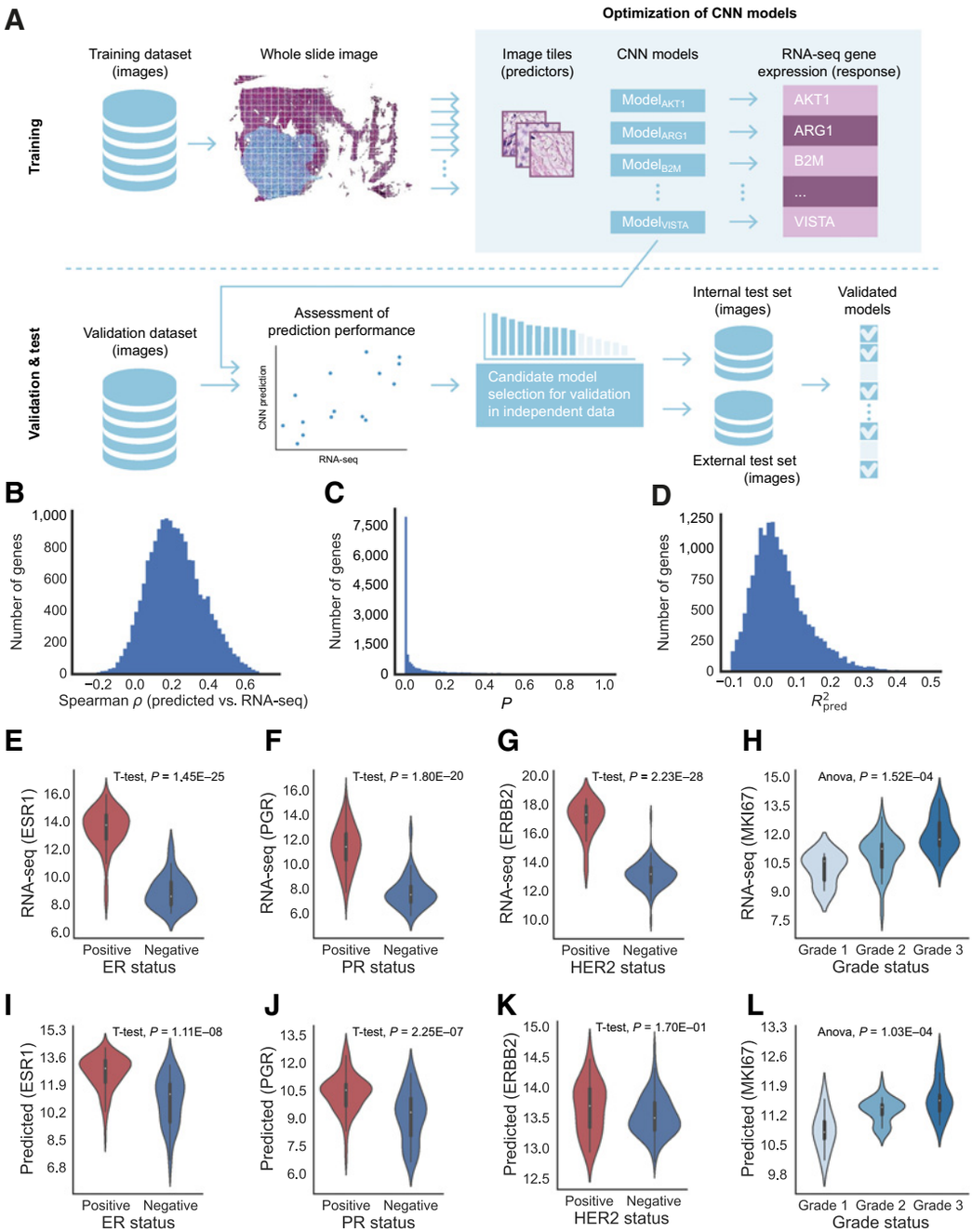

Figure 1.

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. **B,** Histogram describing the empirical distribution of predicted R^2 in the validation set (458 genes with a predicted $R^2 < -0.1$ were excluded from the figure for clarity). **C,** Histogram of the empirical distribution of Spearman ρ between EMO predictions and RNA-seq in the validation set. **D,** Histogram of the P values related to **C**. **E–H,** Distribution of RNA-seq expression values for routine biomarkers (*ESR1, PGR, ERBB2*, and *MKI67*), with respect to clinical status (IHC) of protein expression for the corresponding proteins encoded by each gene. **I–L,** Corresponding distribution of model-predicted gene expression values for each one of the clinical routine markers.

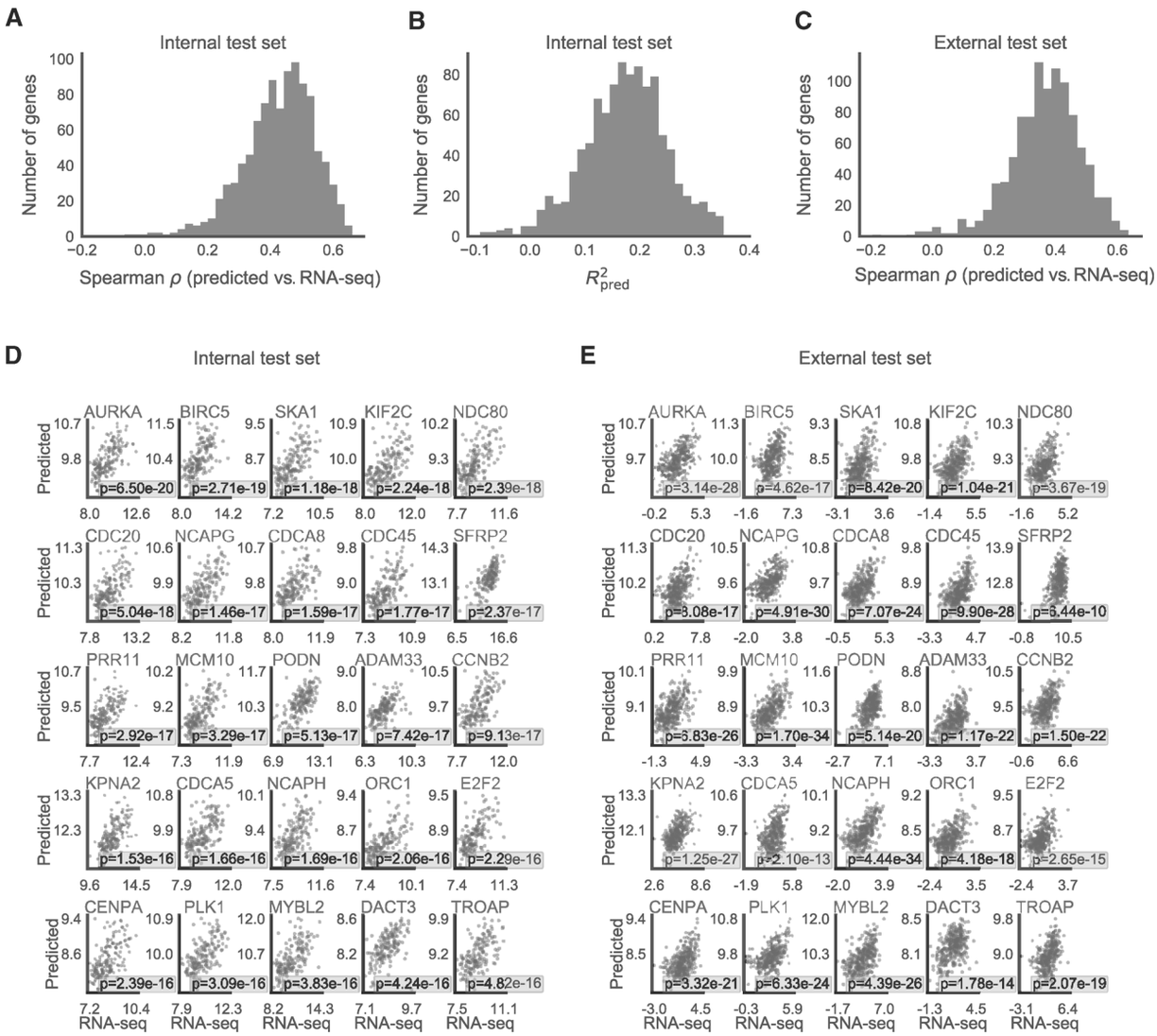

Figure 2.

Summary of model performance on test sets. **A,** Distribution of Spearman ρ in the internal test set. **B,** Distribution of R^2_{pred} in the internal test set ($N_{genes} = 1{,}011$; one gene with a predicted $R^2 < -0.1$ was excluded from the figure for clarity). **C,** Distribution of Spearman ρ 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.

(Supplementary Table S6). For the EndoPredict panel, only three (*BIRC5, IL6ST*, and *UBE2C*) of 12 genes were brought forward for external validation (see Materials and Methods), and could be validated in internal and external test sets. For Oncoptype DX, 10 of 21 genes were brought forward for validation and were successfully validated in the test sets. For Prosigna/PAM50, all genes were evaluated in the test sets and 29 of 50 were successfully validated.

To explore whether the predicted expression of genes in the PAM50 panel shared trends in coexpression patterns with the RNA-seq estimated expression, we performed a cluster analysis. Transcripts and patients were first clustered by the RNA-seq data and visualized as a heatmap of expression values (Supplementary Fig. S2A). Next, the corresponding visualization was generated based on the EMO-predicted gene expression values, using the same order of transcripts and patients (Supplementary Fig. S2B) to allow direct comparison with the RNA-seq data in panel A. Interestingly, there is indeed substantial concordance between the two heatmaps, suggesting a reasonable similarity in the expression patterns for a large number of genes. The same procedure and results were replicated in the external test set (Supplementary Fig. S2C and S2D).

Next, we applied consensus clustering (four clusters; ref. 36) in the RNA-seq and EMO-average prediction (PAM50) datasets separately and estimated the similarity between the two clusterings of patients using adjusted Rand index (37). The adjusted Rand index was 0.25 [95% confidence interval (CI) = (0.23–0.27), permutation-based $P < 0.001$] and 0.20 [95%CI = (0.19–0.22), permutation-based $P < 0.001$] for internal and external test sets, respectively, indicating significant similarity between the two clusterings of patients. We also assessed the

similarity between clusterings and intrinsic subtype labels. The adjusted Rand index was 0.1 for EMO-average prediction [95%CI = (0.08–0.13), permutation-based $P < 0.001$], and 0.27 for RNA-seq [95% CI = (0.25–0.29), permutation-based $P < 0.001$]. We also investigated the similarity between clusterings of genes (PAM50) between RNA-seq and EMO-average prediction (see Supplementary Materials and Methods; Supplementary Fig. S3A–S3H), with the adjusted Rand index estimated to 0.73 (permutation-based $P < 0.001$) and 0.71 (permutation-based $P < 0.001$) for internal test set and external test set, respectively. Finally, we assessed differential gene expression between ER^+ and ER^- patients in RNA-seq and EMO-average prediction data. We fitted linear fixed effects models for each gene in the set of 1,011 genes that were significantly predicted by the EMO approach in the validation set, with the expression as the response variable and ER status, age, HER2 status, tumor size, and lymph node status as covariates. In the internal test set, 584 genes were differentially expressed (FDR-adjusted $P < 0.05$) with respect to ER status in RNA-seq, 514 in EMO-average prediction, and 431 of these were common [FDR-adjusted $P < 0.05$ and same sign of the ER status–related coefficient ($\beta_{hat,\ ER}$)] between RNA-seq and EMO-average prediction (Supplementary Table S7). In the external test set, 760 of 995 genes were differentially expressed (FDR-adjusted $P < 0.05$) between ER^+ and ER^- patients in RNA-seq, 801 in EMO-average prediction, and 701 of these were common (FDR-adjusted $P < 0.05$ and same sign of $\beta_{hat,\ ER}$) between RNA-seq and EMO-average prediction (Supplementary Table S8).

Gene set enrichment analysis identified cancer-associated molecular pathways

To determine whether genes involved in particular molecular mechanisms or processes were enriched in the set of transcripts that could be predicted from histopathology images, we conducted a gene set enrichment analysis (GSEA; ref. 38) across all 17,695 genes (30) using the Reactome database. A total of 16 pathways (**Fig. 3A**) were significantly enriched (FDR-adjusted $P < 0.05$); a majority of these have previously been found to be associated with breast cancer. The functional classes of the significant gene sets included angiogenesis, cell proliferation, cell cycle, apoptosis, signal transduction, metabolism, and immune system. Among the enriched pathways, the "Sema4D induced cell migration and growth-cone collapse" had the strongest association. Sema4D has previously been reported to be overexpressed in breast cancer (39). Sema4D, together with the small GTPase Rho gene family (i.e., RhoA, RhoB, RhoC), which are encoded by genes that belong to the same pathway and were well predicted by the CNN model, are associated with tumor angiogenesis (40). Ranking second was "Signaling by Retinoic Acid." Retinoic acid has been reported as being associated with downregulating genes that relate to breast cancer cell proliferation and upregulating proapoptotic genes, which induces cell death (41). These events are also associated with morphologic changes. In addition, four pathways relating to cell cycle were also identified, including, "cyclin A/B1-associated events during G_2–M transition." Genes that were predicted well and belong to this pathway (*CCNA2, CCNB1*) encode cyclin A2 and B1, respectively. These proteins have been reported to be associated with breast cancer histologic grade (42) and prognosis (43, 44). Another well-predicted gene, *CDK1*, is a member of the enriched pathway "G_2–M DNA replication checkpoint," and the protein it encodes is associated with prognosis in patients with breast cancer (45). Moreover, in the pathway relating to "Loss of Nlp from mitotic centrosomes," Nlp (ninein-like protein) has been recently recognized as an oncogenic protein, whose centrosomal localization and stability could be disturbed in case of *BRCA1* mutations, and eventually lead to abnormal mitotic

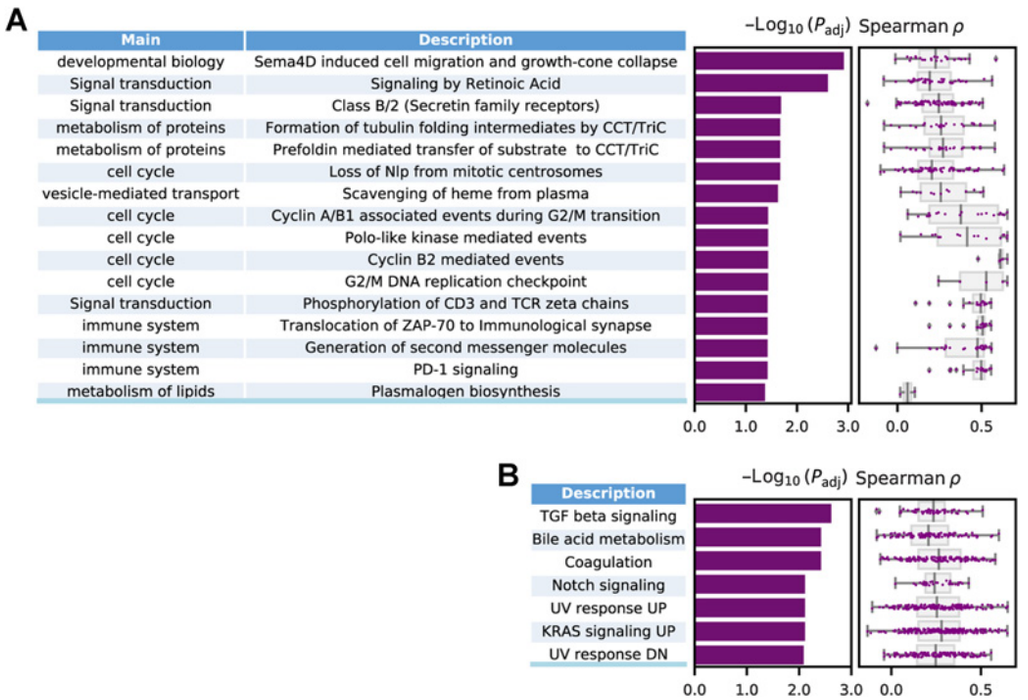

Figure 3.

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 ρ 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.

progression as well as tumorigenesis (46, 47). The remaining significant pathways are involved in biological processes such as "signal transduction," "metabolism of proteins," "cell cycle," and "immune system."

The GSEA was also performed in the HALLMARK gene sets (**Fig. 3B**), which represents a smaller and more curated catalog of gene sets relating to biological functions and processes. This analysis identified seven significant pathways, including those associated with tumor growth and invasion ("TGFβ signaling," "Notch signaling," and "KRAS signaling UP") and metabolism (Bile acid metabolism).

Furthermore, to explore whether there were particular biological mechanisms that were not possible to predict from histopathology images, we performed pathway analysis using only transcripts with nonsignificant EMO predictions ($P_{adjusted} > 0.05$) and with a variance larger than the median variance across all genes (4,184 transcripts). The results of the analysis are included in Supplementary Tables S9 and S10.

Spatial gene expression variability can be predicted by CNN models

Next, we validated EMO predictions of intratumor expression variability (EMO-spatial) by performing ST analysis. Expressions of 76 genes across 12 ROIs in 22 tumors (FFPE sections from independent sets of tumors, 264 ROIs in total) were measured using the Nanostring GeoMX DSP platform (**Fig. 4A**) and compared with EMO-spatial predictions. To ascertain whether intratumor heterogeneity in expression could be predicted, we assessed the association between EMO-spatial predictions and ST measurements, using LME models fitted for each gene across all ROIs and slides, with the ST expression as response, EMO-spatial prediction as a fixed effect, and the slide ID included as random effect to account for slide-level systematic variability. Spatial predictions of 59 genes (77.63%) were significantly associated with ST estimated expression levels (FDR-adjusted $P < 0.05$, likelihood ratio test; **Fig. 4B** and **C**; see also Supplementary Fig. S4 for gene-level within-slide estimates of Spearman correlations between EMO-spatial and ST expression; Supplementary Fig. S5 for examples of prediction results across the 22 WSIs). Among the ten genes with the most significant association between ST estimates and EMO-spatial predictions, three genes could be found in the T-cell receptor pathway (*CD3E, CD8A*, and *CD27*) and three genes in the cytokine and chemokine signaling pathway (*CXCL9, CXCL10*, and *CMKLR1*), other genes were found in the total immune (*PTPRC*), B cells (*MS4A1*), proliferation (*MKI67*), and cytotoxicity (*NKG7*) pathways. Taken together, these results indicate that EMO-spatial

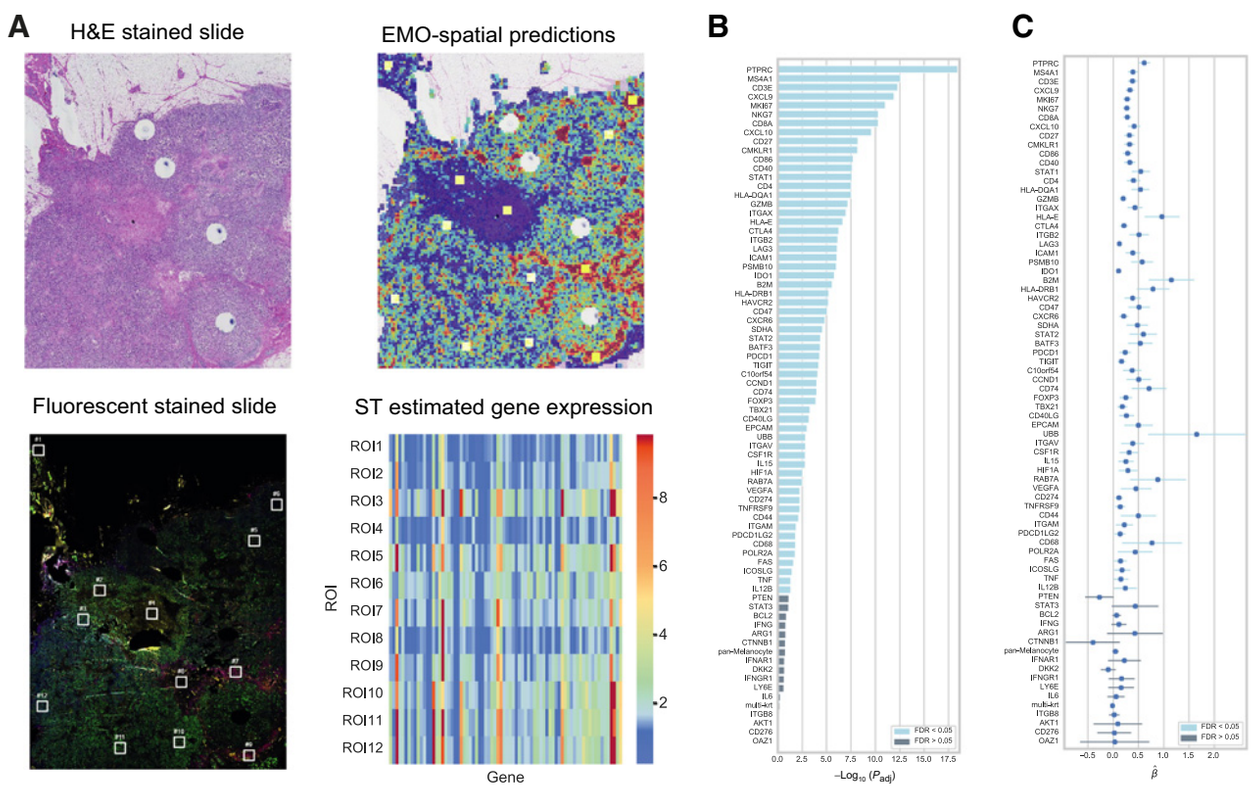

Figure 4.

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 a representative 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 $-\log_{10}$(FDR-adjusted P value) for genes from each LME model. Light blue indicates FDR-adjusted $P < 0.05$ ($N_{WSIs} = 22$). **C,** Corresponding fixed-effect coefficients and 95% CI related to the EMO prediction for each gene (linear mixed effects model; $N_{WSIs} = 22$).

prediction offers a methodology that can enable exploration of intratumor gene expression heterogeneity based on routine H&E-stained sections.

Prediction of a gene expression–based proliferation score from WSIs

To examine whether the EMO model can predict a mRNA expression-based proliferation score directly from WSIs of H&E-stained tissue, we compared the estimated proliferation score (P.Score) from EMO-average predictions with RNA-seq data as well as with IHC-based Ki67 score (**Fig. 5A**). We used a subset of 11 genes in the PAM50 gene panel whose expression is associated with cell proliferation (34). The 11 genes were successfully predicted by EMO-average models ($R^2_{\text{pred}} > 0.2$, FDR-adjusted $P < 0.001$; Supplementary Table S11). The proliferation score was calculated as the average of the 11 gene expression levels (see Materials and Methods for further details).

We then compared the distribution of measured proliferation score [P.Score(RNA-seq)] and EMO prediction [P.Score(EMO)], with respect to each intrinsic molecular subtype, because it is well-known that there is a difference in proliferation across the subtypes (48, 49). As is shown in **Fig. 5B** and **C**, the distribution of predicted proliferation scores was concordant with RNA-seq estimates in the validation set. Luminal A had the lowest score indicating a low rate of cell proliferation, whereas luminal B, HER2-enriched, and basal-like tumors showed an increasing trend in proliferation rate, which is associated with the higher number of mitoses and inferior outcome for patients with these subtypes. The slide-level proliferation scores derived from the EMO-average predictions in the validation set were highly correlated with those of RNA-seq (Spearman ρ of 0.67, $P = 2.82\text{e-}17$; **Fig. 5D**). The results were confirmed in the internal test set (**Fig. 5E–G**; Spearman ρ of 0.66, $P = 4.32\text{e-}23$) and the external test set (**Fig. 5H**; Spearman ρ of 0.55, $P = 1.40\text{e-}29$).

Next, we assessed to what extent P.Score(EMO), based on tile-level predictions (EMO-spatial), enables prediction of intratumor spatial variability of proliferation in comparison with an orthogonal assay (IHC-based Ki67 score, see Materials and Methods). On a general level, as expected, tumors belonging to the luminal A subgroup generally have lower levels of both IHC Ki67 score and P.Score(EMO). In comparison, other subtypes are more proliferative, and here the majority of these tumors and tumor regions had high Ki67 scores as well as high P.Score(EMO). When comparing slide-level IHC Ki67 score with P.Score (EMO), we observed a positive correlation (Spearman $\rho = 0.61$, $P = 5.12\text{e-}5$; Supplementary Fig. S6). In terms of intratumor heterogeneity, the P.Score(EMO) shared a high similarity with the spatial estimates of the IHC Ki67 score. The spatial association between IHC estimated Ki67 and P.Score(EMO) was also confirmed by statistical analysis (LME model, $P < 2\text{e-}16$).

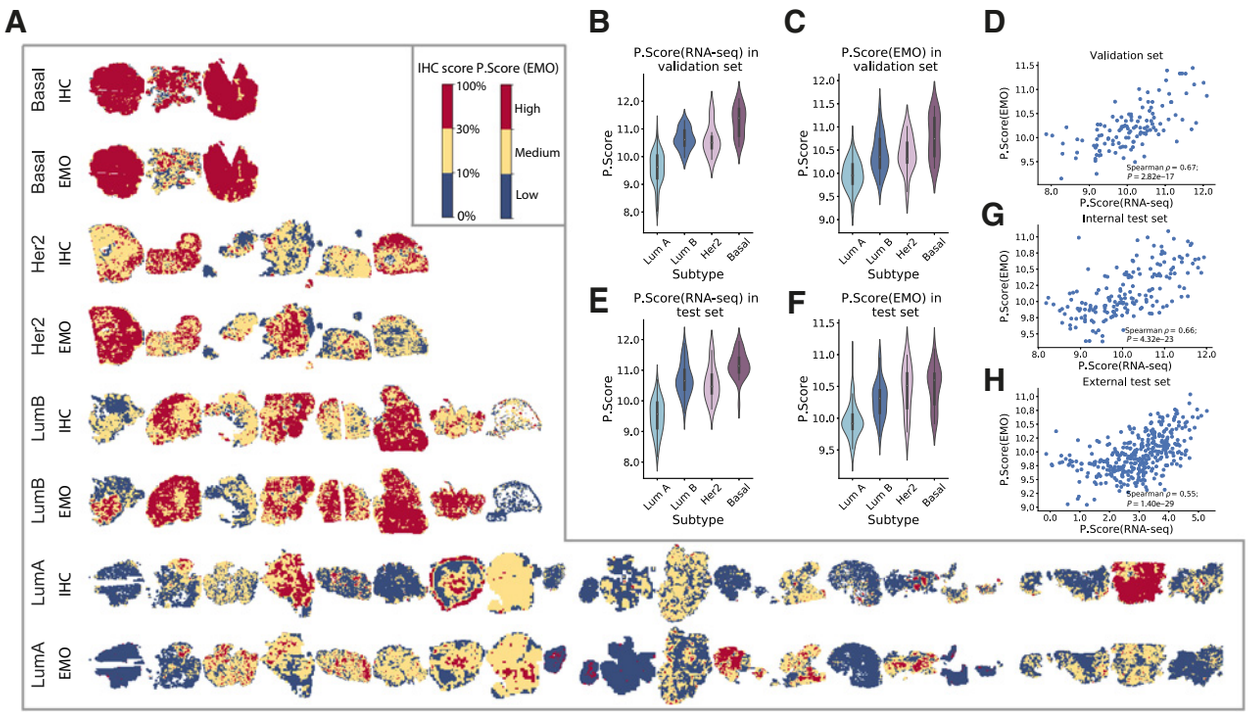

Figure 5.

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 RNA-seq measurements and EMO predictions was observed with a Spearman ρ of 0.67. **E–G,** Corresponds to **B–D** for the internal test set ($N = 172$). **H,** Scatter plot of RNA-seq–estimated and EMO-predicted proliferation scores in the external test set ($N = 350$).

Discussion

We have performed the first reported transcriptome-wide expression–morphology study in breast cancer based on individually optimized gene-level models. Tumor-level prediction results were validated in a completely independent external cohort, and spatial expression predictions were validated in independent tumors by ST profiling. A total of 17,695 gene-specific CNN models were optimized for prediction of gene expression; out of these, 9,334 had a significant association between EMO-average prediction and RNA-seq estimates. Of 1,011 genes brought forward for final validation, prediction performance could be confirmed for 86.65% and 91.26% in the internal and external test data, respectively. We further assessed similarity between clustering of patients and genes based on RNA-seq estimates and EMO predictions and found that there were significant similarities between the clusterings; however, the clusterings were not identical. We also demonstrated that the predicted spatial variabilities in gene expression generated by our approach were significantly associated with ST profiling in 59 of 76 genes. The validation by ST technique suggested that deep CNN models enable characterization of intratumor heterogeneity in RNA expression.

Through GSEA, we found several pathways enriched for genes that could be successfully predicted, which are also associated with breast cancer–related molecular mechanisms. This further supports the hypothesis that morphology is associated with gene expression patterns, and that morphology can be used to predict cancer-related gene expression patterns across numerous genes and pathways.

It has previously been demonstrated that gene expression–based proliferation scores are prognostic and provide treatment predictive information (49, 50). However, only a small fraction of patients with breast cancer have access to expensive molecular profiling, while IHC of Ki67 remains the de facto standard clinical proliferation marker in many countries, despite many problems with reliability (51, 52). Here we wanted to explore whether it is possible to predict an expression-based proliferation score directly from WSIs of H&E-stained tissue. We observed high concordance with the RNA-seq–based proliferation scores, not only with respect to correlations, but also in the trend across intrinsic subtypes of breast cancer. We also demonstrated that spatially resolved predictions of the proliferation score were associated with Ki67 score by IHC staining.

A higher level of Ki67 index has been used to distinguish luminal B subtype from luminal A breast cancers (53), to identify higher risk of disease recurrence (54) and is associated with response to adjuvant chemotherapy (55). However, the clinical utility of Ki67 has been largely impeded by the unsatisfactory interobserver or intraobserver variations (56). Proliferation scores, on the other hand, are computed as a function of expression levels of a set of genes and provide a more reliable measurement of tumor growth rate while avoiding the limitations associated with IHC-based Ki67 analysis. Previous results also showed that an expression-based proliferation score outperformed the Ki67 index in predicting relapse-free survival and disease-specific survival (34). In computational pathology, attempts have been made to predict proliferation scores directly from WSIs (57), where the best performing method achieved a Spearman ρ of 0.62, while no spatially resolved predictions were attempted. In this study, we observed a slightly improved performance with a Spearman ρ of 0.66 and 0.67 on validation and test sets, respectively. In addition, we demonstrated a strong and significant association (Spearman $\rho = 0.61$) between slide-level IHC Ki67 score and the predicted proliferation score [P.Score (EMO)]. We note that a strong association between mRNA expression and protein expression is not expected in general due to different temporal scales in half-life of these two types of molecules and differences in regulation. Nevertheless, we observed a relatively high level of intratumor spatial coexpression between IHC Ki67 and the predicted proliferation score [P.Score (EMO)]. Our results, together with previous evidence, indicate that CNNs enable objective and reproducible estimation of proliferation scores, and provide information of direct clinical value (7, 58).

To date, three studies with the objective of predicting gene expression phenotypes from histopathology images have been reported previously (21–23). However, these studies have substantial limitations in one or more aspects: (i) the number of genes analyzed (250 genes) and sample size (23 patients; ref. 21); (ii) extensive use of transfer learning, that is, a single-global CNN model for prediction of all phenotypes rather than optimization of gene-specific models (22, 23); (iii) the use of a pan-cancer approach (22, 23), where a single model is used across a range of cancer diseases, which by design will lead to models optimized to capture morphologies shared across the majority of diseases included; (iv) lack of independent external validation cohorts (22, 23), or validation in very small datasets (two tumors; ref. 21); or (v) lack of validation by orthogonal experimental techniques and spatial expression predictions (22, 23).

By developing models for a single cancer disease (breast cancer) and by optimizing individual deep CNN models for each gene, we avoid several strong assumptions made in previously reported studies (21–23) that are unlikely to hold. Pan-cancer models assume shared morphologies across cancer diseases, which provides a fundamental limitation given the broad range of morphologic characteristics observable in different cancer types. Strong reliance on transfer learning across genes represents another fundamental limitation that is likely to constrain the ability to develop models that are effective for modelling more specific relationships between morphology and gene expression.

This study is limited with respect to the size of the training dataset, and it is expected that with more training data, the prediction performance could improve further. Spatially resolved data for model optimization also has the potential to improve model performance in the future. One previously reported study has applied that approach, however, their training dataset was limited to only 23 tumors and 250 genes (21). In our study, the ST validation was limited to a panel of 76 genes, which was dictated by the availability of FFPE compatible ST profiling gene panel at the time of the study. Furthermore, models in this study were trained with tiles at a fixed resolution level and tile size for all transcripts. The prediction performances could potentially be further improved with individually optimized image scale for each transcript, or by implementing a multiscale modeling approach. In the context of tile-based models, it is also implicitly assumed that there is a large enough perceptive field and resolution to capture both microscopic and macroscopic details.

Prediction of gene expression from routine H&E WSIs has the potential to impact both clinical diagnostics as well as cancer research. Prediction of molecular phenotypes can enable cost-effective precision medicine, either by direct predictions of key markers, or as a way to prioritize which patients are likely to benefit from comprehensive but costly molecular profiling. In the research domain, cost-effective predictions of expression will enable large-scale epidemiologic studies that include gene expression phenotypes as exposures. Spatial prediction of gene expression provides a complement to single-cell sequencing and ST profiling for studies of intratumor heterogeneity and tumor microenvironment, and enables studies at a substantially larger scale compared with what is possible by direct molecular profiling. The results from this study are promising and we expect that our approach

will also work well for application in other cancer diseases, and for prediction of other types of molecular phenotypes, such as somatic mutations, copy-number alterations, epigenetic factors, metabolite or protein abundances. In the current study, we also evaluated to what extent transcripts included in commercial gene expression assays could be predicted, and found that around half of these genes can be predicted, defined by a significant association between the predicted and experimental expression estimates. Some of these genes have a relatively high correlation with the RNA-seq estimates, but for other transcripts the association is not as strong, which may limit their use. Furthermore, it is important to realize that at this point in time, and based on this study alone, it would be premature to suggest gene expression assays for clinical use could be replaced by image analysis. However, it would be of interest in future studies to compare prognostic performance between established gene expression assays and expression predicted by CNN models from histopathology images. Our findings suggest that deep learning–based image analysis for prediction of the tumor average expression of a substantial number of transcripts is possible and feasible. We also applied the CNN-based analysis to successfully predict a clinically relevant expression-based proliferation score. However, more importantly, we demonstrated and experimentally validated that spatial gene expression predictions can be used to characterize intratumor expression heterogeneity.

Authors' Disclosures

Y. Wang reports personal fees from Stratipath AB outside the submitted work. B. Ács reports grants from The Swedish Society for Medical Research (Svenska Sällskapet för Medicinsk Forsknings—SSMF) outside the submitted work. P. Ruusuvuori reports grants from Academy of Finland, ERA PerMed JTC2020, and Cancer Foundation Finland and other support from CSC Centre for Scientific Computing during the conduct of the study. J. Hartman reports grants from Swedish Cancer Fund, Medtech Labs, Swedish Breast Cancer Association, and Stockholm Cancer Society during the conduct of the study; personal fees from Roche, Pfizer, Merck, MSD, and Eli Lilly, grants from Cepheid, and grants and personal fees from Novartis outside the submitted work; and is cofounder of and shareholder in Stratipath AB. M. Rantalainen reports grants from Swedish Research Council, Swedish Cancer Society, Swedish e-Science Research Centre (SeRC), ERA PerMed (through Swedish Research Council), Karolinska Institutet (Cancer Research KI; StratCan), and MedTechLabs during the conduct of the study, and is cofounder of and shareholder in Stratipath AB. No disclosures were reported by the other authors.

Authors' Contributions

Y. Wang: Data curation, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing, scientific computing. **K. Kartasalo:** Validation, investigation, methodology, writing–original draft, writing–review and editing, scientific computing. **P. Weitz:** IHC-HE image registration, drafting the method for the corresponding sections. **B. Ács:** IHC scoring, drafting the method for the corresponding sections. **M. Valkonen:** Investigation, scientific computing. **C. Larsson:** Resources. **P. Ruusuvuori:** Resources, supervision, writing–original draft, writing–review and editing, directed the project. **J. Hartman:** Conceptualization, resources, directed the study. **M. Rantalainen:** Resources, conceptualization, supervision, funding acquisition, methodology, writing–original draft, writing–review and editing, directed the project.

Acknowledgments

This project was supported by funding from the Swedish Research Council under the frame of ERA PerMed (ERAPERMED2019–224—ABCAP; M. Rantalainen), Swedish Research Council (M. Rantalainen, J. Hartman), Swedish Cancer Society (M. Rantalainen, J. Hartman), Karolinska Institutet (Cancer Research KI; StratCan; M. Rantalainen, J. Hartman), MedTechLabs (M. Rantalainen, J. Hartman), Swedish e-science Research Centre (SeRC)—eCPC (M. Rantalainen), Stockholm Region (J. Hartman), Stockholm Cancer Society (J. Hartman), Swedish Breast Cancer Association (J. Hartman), Academy of Finland (326463, 341967, and 335976; P. Ruusuvuori), Academy of Finland Center of Excellence programme (312043; P. Ruusuvuori), Cancer Foundation Finland (P. Ruusuvuori), ERA PerMed ABCAP (P. Ruusuvuori), CSC—IT Center for Science (Finland; Grand Challenge pilot project AI-EMO, 2001568; P. Ruusuvuori), Tampere University graduate school (K. Kartasalo), and University of Turku Graduate School UTUGS and Turku University Foundation (M. Valkonen). The authors would like to acknowledge the patients, clinicians, and hospital staff participating in the SCAN-B study; the staff at the central SCAN-B laboratory at Division of Oncology, Lund University; the Swedish National Breast Cancer Quality Registry (NKBC); Regional Cancer Center South; and the South Swedish Breast Cancer Group (SSBCG). They also thank Johan Vallon-Christersson (Lund University) for help in preparing these data. The authors thank Duong Nguyen Thuy Tran and other personnel that have been contributing to slide scanning operations in the Rantalainen group/CHIME project at Karolinska Institute. They thank TCGA Research Network, https://www.cancer.gov/tcga, for providing access to part of the data used in this study. The authors thank The Swedish Society for Medical Research (Svenska Sällskapet för Medicinsk Forsknings—SSMF) for a postdoctoral grant and the Hungarian Society of Senology for supporting B. Ács.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 12, 2021; revised April 30, 2021; accepted July 28, 2021; published first August 2, 2021.

References

1. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective validation of a 21-gene expression assay in breast cancer. N Engl J Med 2015;373:2005–14.
2. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics 2015;8:54.
3. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
4. van de Vijver MJ, van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
5. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003;100:8418–23.
6. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.
7. Beresford MJ, Wilson GD, Makris A. Measuring proliferation in breast cancer: practicalities and applications. Breast Cancer Res 2006;8:216.
8. Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, Mac Grogan G, et al. A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. Clin Cancer Res 2008;14:1744–52.
9. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
10. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793–5.
11. Caiado F, Silva-Santos B, Norell H. Intra-tumour heterogeneity - going beyond genetics. FEBS J 2016;283:2245–58.
12. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15:81–94.
13. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82.
14. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–30.
15. Rantalainen M. Application of single-cell sequencing in human cancer. Brief Funct Genomics 2018;17:273–82.
16. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat Biotechnol 2020;38:586–99.

17. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016;353:78–82.
18. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol 2020;21:222–32.
19. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.
20. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. BioRxiv 064279 [Preprint]. 2018. Available from: https://doi.org/10.1101/064279.
21. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng 2020;4:827–34.
22. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer 2020;1:800–10.
23. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun 2020;11:3877.
24. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In Proceedings of 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2009 Jun 28–Jul 1; Boston, MA. New York: IEEE; 2009.
25. Wang M, Klevebring D, Lindberg J, Czene K, Grönberg H, Rantalainen M. Determining breast cancer histological grade from RNA-sequencing data. Breast Cancer Res 2016;18:48.
26. Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative. JCO Precis Oncol 2018;2:PO.17.00135.
27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv:1512.00567 [cs.CV] [Preprint]. 2015. Available from: https://arxiv.org/abs/1512.00567.
28. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980v5 [cs.LG] [Preprint]. 2014. Available from: https://arxiv.org/abs/1412.6980v5.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol 1995;57:289–300.
30. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics 2017;18:151.
31. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503.
32. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database hallmark gene set collection. Cell Syst 2015;1:417–25.
33. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8:1826.
34. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, 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 2010;16:5222–32.
35. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous systems. 2015.
36. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
37. Hubert L, Arabie P. Comparing partitions. J Classification 1985;2:193–218.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
39. Ch'ng ES, Kumanogoh A. Roles of Sema4D and Plexin-B1 in tumor progression. Mol Cancer 2010;9:251.
40. Jiang H, Chen C, Sun Q, Wu J, Qiu L, Gao C, et al. The role of semaphorin 4D in tumor development and angiogenesis in human breast cancer. Onco Targets Ther 2016;9:5737–50.
41. Hua S, Kittler R, White KP. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. Cell 2009;137:1259–71.
42. Boström P, Söderström M, Palokangas T, Vahlberg T, Collan Y, Carpen O, et al. Analysis of cyclins A, B1, D1 and E in breast cancer in relation to tumour grade and other prognostic factors. BMC Res Notes 2009;2:140.
43. Poikonen P, Sjöström J, Amini RM, Villman K, Ahlgren J, Blomqvist C. Cyclin A as a marker for prognosis and chemotherapy response in advanced breast cancer. Br J Cancer 2005;93:515–9.
44. Agarwal R, Gonzalez-Angulo AM, Myhre S, Carey M, Lee JS, Overgaard J, et al. Integrative analysis of cyclin protein levels identifies cyclin b1 as a classifier and predictor of outcomes in breast cancer. Clin Cancer Res 2009;15:3654–62.
45. Kim SJ, Nakayama S, Shimazu K, Tamaki Y, Akazawa K, Tsukamoto F, et al. Recurrence risk score based on the specific activity of CDK1 and CDK2 predicts response to neoadjuvant paclitaxel followed by 5-fluorouracil, epirubicin and cyclophosphamide in breast cancers. Ann Oncol 2012;23:891–7.
46. Jin S, Gao H, Mazzacurati L, Wang Y, Fan W, Chen Q, et al. BRCA1 interaction of centrosomal protein Nlp is required for successful mitotic progression. J Biol Chem 2009;284:22970–7.
47. Shao S, Liu R, Wang Y, Song Y, Zuo L, Xue L, et al. Centrosomal Nlp is an oncogenic protein that is gene-amplified in human tumors and causes spontaneous tumorigenesis in transgenic mice. J Clin Invest 2010;120:498–507.
48. Heng YJ, Lester SC, Tse GMK, Factor RE, Allison KH, Collins LC, et al. The molecular basis of breast cancer pathological phenotypes. J Pathol 2017;241:375–91.
49. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 2008;10:R65.
50. Martín M, Prat A, Rodríguez-Lescure A, Caballero R, Ebbert MTW, Munárriz B, et al. PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer. Breast Cancer Res Treat 2013;138:457–66.
51. Polley M-YC, Leung SCY, McShane LM, Gao D, Hugh JC, Mastropasqua MG, et al. An international Ki67 reproducibility study. J Natl Cancer Inst 2013;105:1897–906.
52. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer Working Group. J Natl Cancer Inst 2011;103:1656–64.
53. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, et al. Tailoring therapies–improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol 2015;26:1533–46.
54. Andre F, Arnedos M, Goubar A, Ghouadni A, Delaloge S. Ki67–no evidence for its use in node-positive breast cancer. Nat Rev Clin Oncol 2015;12:296–301.
55. Criscitiello C, Disalvatore D, De Laurentiis M, Gelao L, Fumagalli L, Locatelli M, et al. High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in Luminal B HER2 negative and node-positive breast cancer. Breast 2014;23:69–75.
56. Penault-Llorca F, Radosevic-Robin N. Ki67 assessment in breast cancer: an update. Pathology 2017;49:166–71.
57. Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med Image Anal 2019;54:111–21.
58. Stover DG, Coloff JL, Barry WT, Brugge JS, Winer EP, Selfors LM. The role of proliferation in determining response to neoadjuvant chemotherapy in breast cancer: a gene expression-based meta-analysis. Clin Cancer Res 2016;22:6039–50.