Linking signaling pathways to transcriptional programs in breast cancer

HATICE U. OSMANBEYOGLU
RAPHAEL PELOSSOF
JACQUELINE F. BROMBERG
CHRISTINA S. LESLIE1

The Problem (2014)

Cancer process:

- Cancer cells acquire genetic and epigenetic alterations that often target signal transduction pathways
- These alterations lead to dysregulation of oncogenic signal transduction pathways
- In turn, this alters downstream transcriptional programs.

Problem/Motivation:

- **Deciphering signaling pathways** that are deregulated in a given tumor in order to personalize therapy is a major goal.
- Much effort devoted to cataloging somatic alterations across large sets of tumors and mapping them to cellular pathways
- These projects have generated massive repositories of tumor mRNA data, giving a complex readout of the transcriptional changes downstream from altered signaling pathways.
- Unable to translate the mutational landscape of a tumor into a **usable model of affected pathways**.
- Unable to use mutational status to accurately **predict response to targeted therapies**.
- Numerous methods attempt to deduce aberrant signaling pathways in tumors from **mRNA data alone**.
- But these pathway analysis approaches remain **qualitative and imprecise**.

Recent Developments

- Advent of proteomic methods has the potential to provide a systematic map of critical signaling pathways that are altered in cancer.
- Recently, TCGA project has added RPPA profiling for a panel of proteins and phosphoproteins. Reverse-phase protein microarrays (RPPAs) are a medium-throughput technology to analyze the expression levels of a protein or phosphoprotein across many samples at once.
- Quantitative profiling of proteins in tumor tissues using RPPA presents many technical challenges: Antibody validation, Variability in tissue handling & Intra-tumoral heterogeneity.
- This gives rise to noisy measurements of the activity of signaling proteins.

The Idea

- Link upstream signaling to downstream transcriptional response
- Do so by exploiting Reverse Phase Protein Array (RPPA) and mRNA expression data
- Model views RPPA data as a noisy readout of the activity of signaling pathways;
  - Oncogenic signaling pathways converge on a set of Transcription Factors (TFs)
  - TF's dysregulated activity in turn alters the mRNA expression levels of TF target genes.
- Created an algorithm called **Affinity Regression** to learn an interaction matrix between Upstream signal transduction proteins and Downstream transcription factors (TFs) to explain target gene expression
- Use TF binding site prediction to determine the set of TFs that potentially regulate each gene.
- The trained model can then be used in multiple ways:
  - Given a tumor sample's protein expression profile, we can predict the TF activity.
  - Given a tumor sample's gene expression profile, we can infer the signaling protein activity.

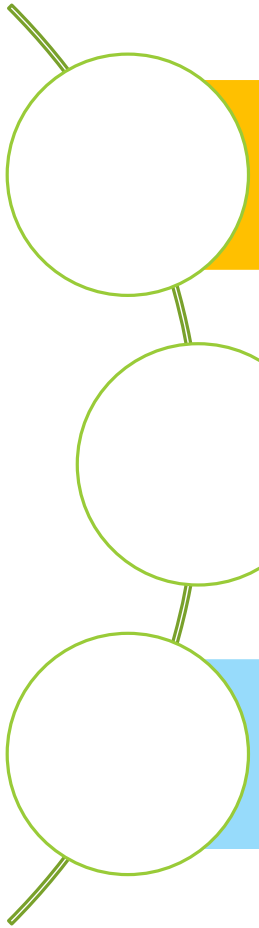
Summary of Results

- Applied approach to 397 breast cancer profiles from TCGA for which both RPPA and mRNA data are available
- Used Affinity Regression:

1. To infer the deregulated signaling pathways that drive expression changes in distinct breast cancer subtypes
2. To leverage the tumor model to predict drug sensitivity using breast cancer cell line mRNA and drug response data
3. To predict survival within the heterogeneous ER+, Luminal A subtype.

Breast Cancer

- Breast cancer has been categorized into three basic therapeutic groups.
- Within the ER+ category, gene expression profiling studies (PAM50) have identified two subtypes within ER-positive breast cancers, Luminal A and Luminal B.
- Although patients with Luminal A cancers have the best prognosis, these tumors are heterogeneous, and there exist few markers that predict recurrence and survival.



G1: Basal-like or triple-negative breast cancers

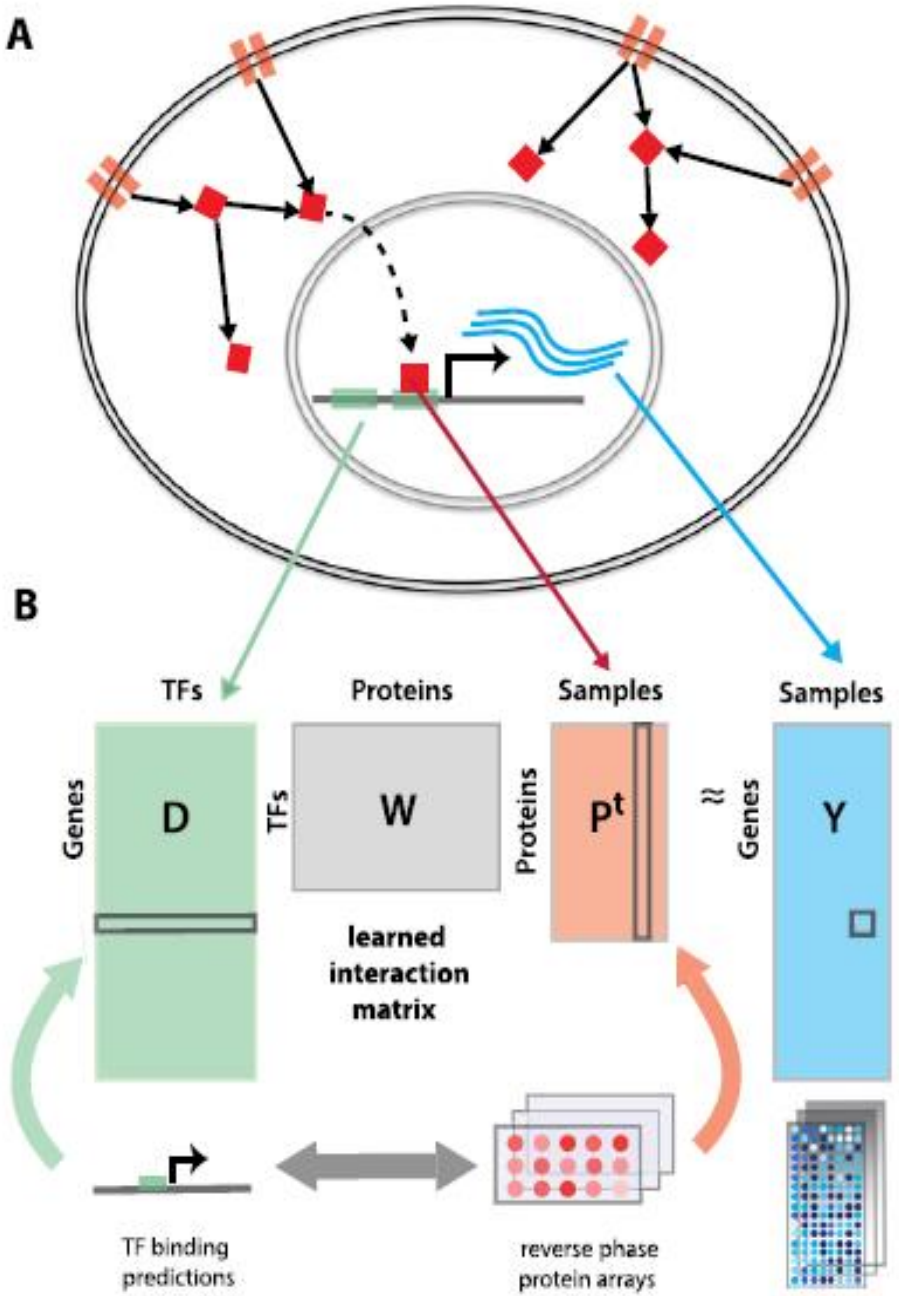
- TNBCs, lacking expression of the estrogen receptor [ER], progesterone receptor [PR], and HER2,
- Characterized by a poor prognosis and no specific targeted therapies

G2: HER2 (ERBB2) amplified

- Associated with relatively poor prognosis if untreated
- With significant clinical benefit from anti-HER2-therapy

G3: Estrogen Receptor-positive (Luminal)

- Characterized by a relatively good prognosis and response to targeted hormonal therapies.



Affinity Regression

- **Matrix Y:** Data set of N genes from M tumor samples; Y = NxM matrix of mean-centered log gene expression profiles (Microarray data)
- **Matrix-D:** Using TF binding site prediction in gene promoters, we define a matrix D = NxQ, where each row represents a gene and each column is a binary vector representing the target genes of a TF. (Motif data from MSigDB TRANSFAC v7.4)
- **Matrix-P:** P = MxS of tumor sample (phospho) protein attributes where each row represents a tumor sample and each column represents mean-centered RPPA expression levels of a signaling protein across tumor samples.
- **Matrix-W:** Transcription Factors to Proteins mapping (To be Learned)
- Bilinear regression using: $\mathbf{D * W * P^T + \varepsilon = Y}$

Discussion

The W matrix represents an interaction between TFs and Proteins. In this study, they have learned the W from tumor samples. What are the implications of this?

Should W have instead been learned from normal samples? Would a W learned from normal samples have been significantly different?

Would W be different for different types of tumor cells (different types of cancer)? What about different stages?

Is there a notion of a universal (true) W? Can that W actually be learned from normal samples? What are the challenges?

Is the W that is learned from tumor samples meant to be an approximation of true W? Or is the W learned from tumor samples meant to be different from true W and reflective of the fact that these cells are cancerous and reflective of the specific type of cancer?

Affinity Regression Outperforms Nearest Neighbor for Gene Expression Prediction on Held-out Samples

Experiment #1

- Evaluated approach on a data set of BRCA tumors from TCGA where both genome-wide mRNA expression data and RPPA measurements for 164 proteins/phosphoproteins are available.
- Trained model on equal numbers of samples for each subtype (n = 48x4 = 192).
- For Motif data, used binding site predictions for 230 TFs in the promoter regions from MSigDB
- Use the learned Affinity Regression model
  - D = 4029 Genes x 230 TFs
  - W = 230 TFs x 164 proteins/phosphoproteins.
  - P^T = 164 proteins/phosphoproteins x 192 samples
  - $Y = DxWxP^T$
- For statistical evaluation, computed the mean Spearman rank correlation between predicted and measured gene expression profiles on held-out samples using six-fold cross-validation.
- Compared results with a Nearest Neighbor approach, where neighbors are chosen based on similarity of protein expression profiles
- To further validate the performance, also examined an independent test set of 205 TCGA samples.

Performance vs NN Baseline

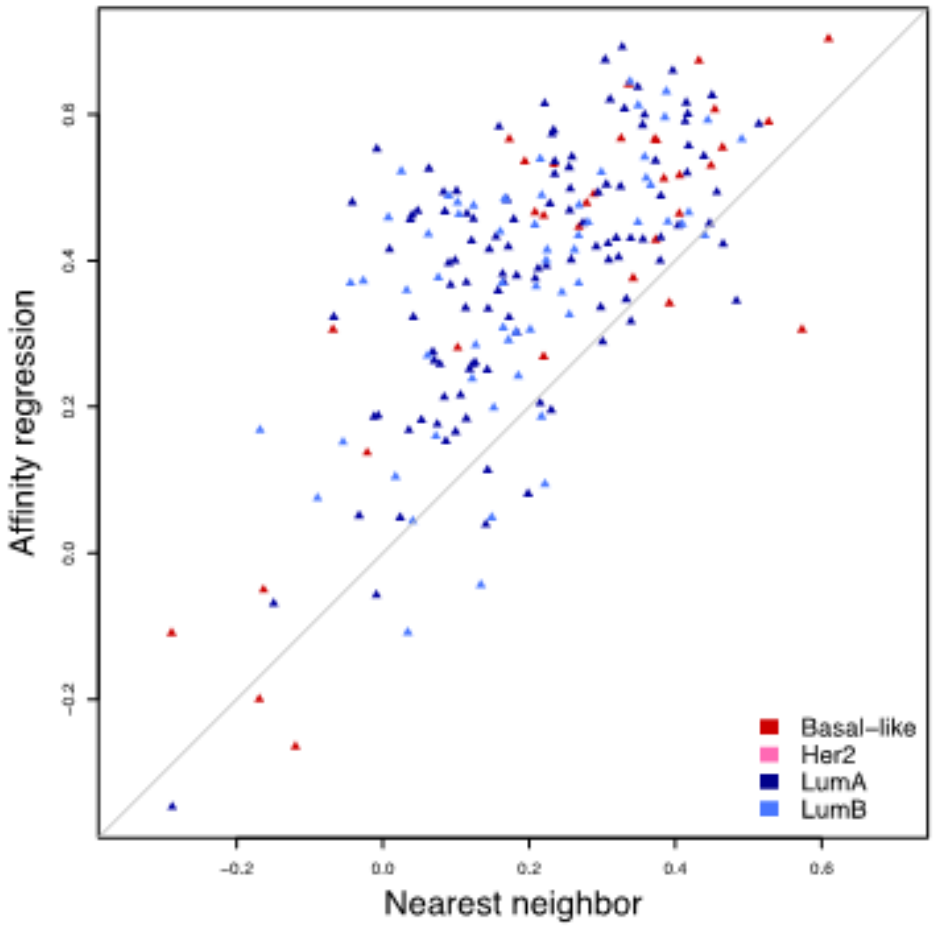


Figure S1. Performance of the trained affinity regression model on an independent test set of TCGA samples, compared to nearest neighbor.

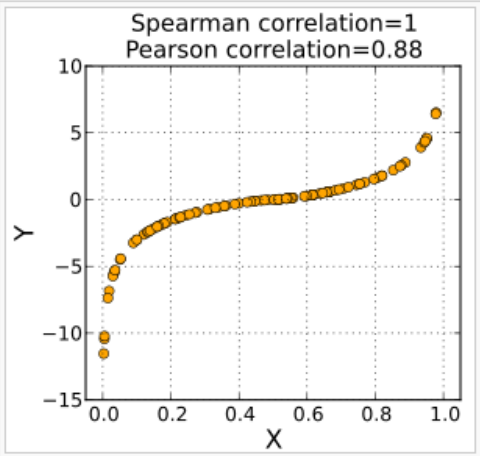
Plot showing **Spearman correlations** between predicted and actual gene expression changes relative to a median ref.

Claim: Affinity Regression outperforms the baseline of NN. Therefore, the model explains a meaningful part of the dysregulation of gene expression in breast cancer based on the ability to predict gene expression variation across tumors on held-out tumor samples.

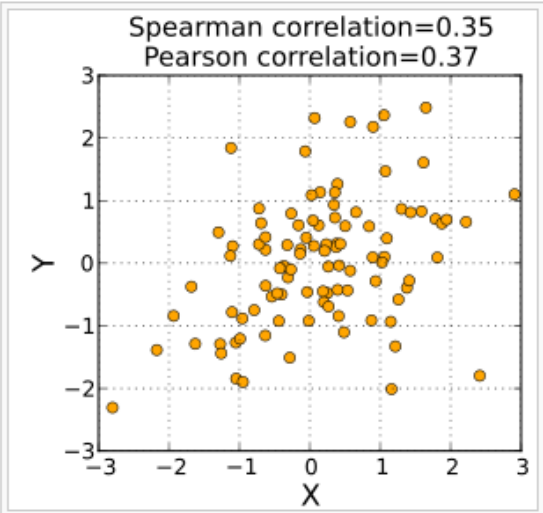
Critique:

1. Is Nearest Neighbor the right baseline?
2. How good are Spearman correlation scores of 0.41 (training sample) and 0.39 (test sample)?

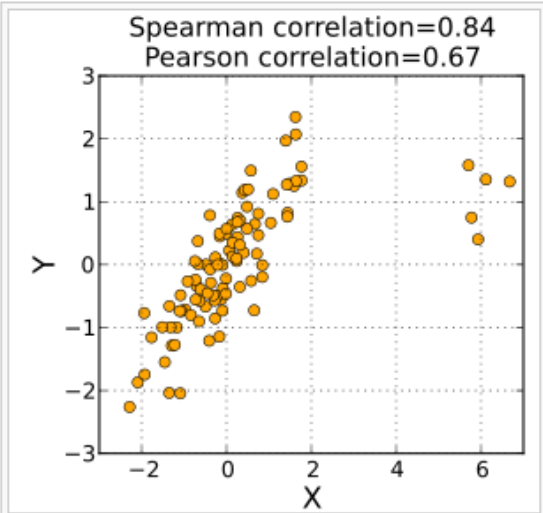
Spearman Correlation



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data-points with greater x-values than that of a given data-point will have greater y-values as well. In contrast, this does not give a perfect Pearson correlation.



When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

Spearman Correlation

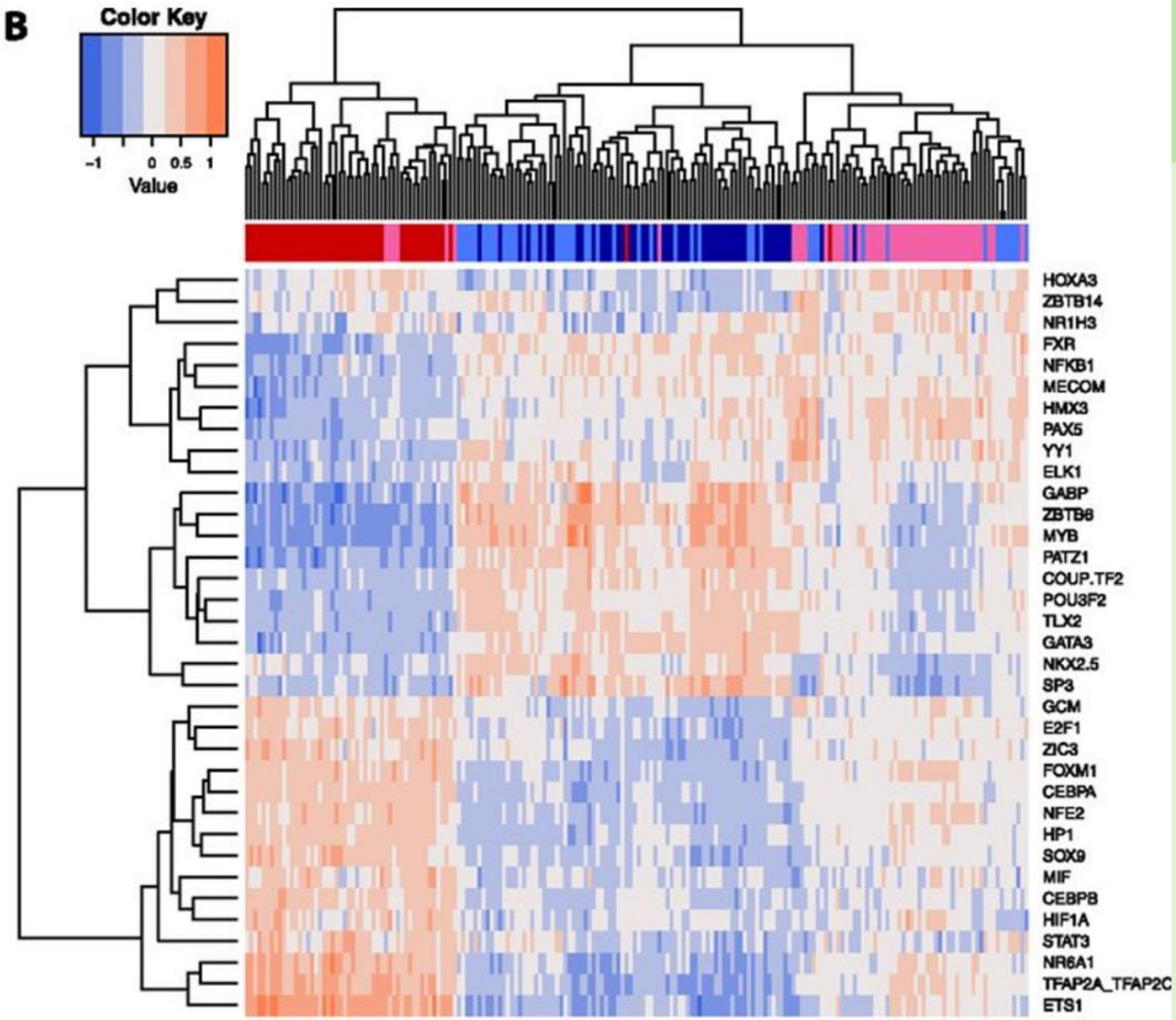
Assesses how well the relationship between two variables can be described using a monotonic function.

Critique: In reality, a Spearman Correlation of 0.35 to 0.45 is only indicative of an elliptical distribution (similar to the middle picture). The Spearman Correlation for the training samples was 0.41 and test samples was 0.39.

Affinity Regression Largely Captures Previously Defined Transcriptomic Subtypes

Experiment #2A: Identify Active TFs for each Tumor Sample

- **Objective:** Examine whether the model reflects the existing PAM50 expression-based breast cancer subtype classifications.
- **Process:**
  - Mapped its protein expression profile P^T through the learned interaction matrix by WxP^T to obtain a weight vector over TFs
  - All training examples (n = 192) were used to learn the model.
- **Results:**
  - Hierarchical clustering of inferred TF activity of tumor samples (WP^T) largely recovered the distinction between the three major subtypes
  - Adjusted Rand Index 0.615 for three-way clustering & ARI of 0.449 for four-way clustering
  - Basal-like samples were well separated from other subtypes.
- **Claim:** The model largely captures previously defined transcriptomic subtypes.



Unsupervised Hierarchical Clustering

Figure S2. Performance of Affinity Regression using data from the TCPA RPPA data set.

Key Takeaway: Hierarchical clustering of inferred TF activities recovers major transcriptional subtypes.

Critique:

1. Per their own admission, LumA and LumB were not well separated (error rate of 40-60%).
2. Even the Her2 cluster seems to have an error of more than 25%.
3. The heat maps seem to have an intensity predominantly between -0.50 and 0.50 (quite low).

Experiment #2B: Identify the Activity of Signaling Proteins for each Tumor Sample

- **Objective:** Examine whether the model reflects the existing PAM50 expression-based breast cancer subtype classifications.
- **Process:**
  - Mapped the expression profiles through the motif hit matrix and our learned model by $Y^T DW$
  - This gives a weight vector over (phospho) proteins for each sample.
- **Results:**
  - Hierarchical clustering of samples by inferred protein activity also recovered the distinction between breast cancer subtypes
  - Adjusted Rand Index 0.58 for three-way clustering & ARI of 0.435 for four-way Clustering
  - Using just the RPPA values, the ARI was 0.289 for four-way clustering
- **Claim:** The model largely captures previously defined transcriptomic subtypes.

Adjusted Rand Index

Adjusted Rand Index is a measure of the similarity between two data clusterings.

Definition: Given a set of n elements S, and two partitions of S: X (a partition of S into r subsets) and Y (a partition of S into s subsets),

- TP, the number of pairs of elements in S that are in the same set in X and in the same set in Y
- TN, the number of pairs of elements in S that are in different sets in X and in different sets in Y
- FP, the number of pairs of elements in S that are in the same set in X and in different sets in Y
- FN, the number of pairs of elements in S that are in different sets in X and in the same set in Y

The Rand index (R) is: R = {TP+TN}/{TP+TN+FP+FN}

Rand index: Perfectly random clustering returns the minimum score of 0, perfect clustering returns the maximum score of 1.

Adjusted Rand index: A variation of the Rand index which takes into account the fact that random chance will cause some objects to occupy the same clusters, so the Rand Index will never actually be zero. ARI can return a value between -1 and +1.

$$AR = \frac{2(TP * TN - FP * FN)}{(TP + TN)(TN + FP) + (TP + FN)(FN + FP)}$$

Affinity Regression Identifies Subtype-Specific TFs & Signaling Proteins Associated with Expression Changes

Experiment #3

- Assessed TF-subtype associations using a Mann-Whitney U-test to compare inferred TF activity between pairs of transcriptional subtypes or groups of subtypes
- Also assessed differences in inferred protein activity across the clinically relevant transcriptional subtypes
- Tested three pairwise comparisons for each TF: (1) basal-like vs. HER2, Luminal A, Luminal B; (2) HER2 vs. Luminal A, Luminal B; and (3) Luminal A vs. Luminal B.
- **Results:**
  - Found more associations using model than with TF mRNA expression levels directly
  - Basal-like-specific TF regulators include ETS1, CEBPB, NFATC4, HMG and SOX9, HMX3 and ZBTB14 for HER2 and SMAD4, NKX2-1 and FOXA1 for Luminal-A
  - Similar results for protein activity

Mann-Whitney U-Test

- Also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test (WRS), or Wilcoxon–Mann–Whitney test
- Is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other.
- *Switch to Wiki Example*

Experiment #3: Results

Table 1. Transcription factors that show significant subtype specificity

TF	Basal-like vs. HER2, Luminal	HER2 vs. Luminal	LumA vs. LumB	Subtype specificity
SOX9	0.07	0.61	0.74	Basal-like
HMG	0.08	0.69	0.83	Basal-like
NFATC4	0.09	0.85	0.86	Basal-like
ETS1	0.07	0.54	0.86	Basal-like
CEBPB	$<10^{-3}$	0.85	1	Basal-like
ZBTB14	0.95	0.76	0.1	HER2
HMX3	$<10^{-3}$	0.85	0.97	HER2
SMAD4	0.51	0.85	$<10^{-3}$	LumA
VSX2	0.97	0.86	0	LumA
NKX2-2	0.55	0.58	0.06	LumA
TTF1	0.92	0.81	0.06	LumA
FOXA1	0.86	0.81	0.09	LumA
IRF2	0.49	0.54	0.1	LumA
GTF2I	0.08	0.65	0.71	LumA/B
GATA3	0.1	0.54	0.86	LumA/B
MEIS1	0.61	$<10^{-3}$	0.94	LumA/B
FXR	$<10^{-3}$	0.85	0.97	LumA/B
IRF10	0.1	0.54	0.97	LumA/B
FOXF1	0.05	0.85	0.97	LumA/B
MECOM	$<10^{-3}$	0.95	0.97	LumA/B/HER2

FDR adjusted $P < 0.1$.

Table 2. Proteins/phosphoproteins that show significant subtype specificity

Protein	Basal-like vs. HER2, Luminal	HER2 vs. Luminal	LumA vs. LumB	Subtype specificity
KIT	$<10^{-3}$	0.45	0.27	Basal-like
CCNE1	$<10^{-3}$	0.06	0.27	Basal-like
MSH2	$<10^{-3}$	0.04	0.16	Basal-like
CHEK2	$<10^{-3}$	0.14	0.03	Basal-like
CDH3	$<10^{-3}$	0.03	0.92	Basal-like
MSH6	0.01	0.14	0.12	Basal-like
WWTR1 (pS89)	0.01	1	0.54	Basal-like
PTGS2	0.01	0.03	0.99	Basal-like
KDR	0.02	1	0.54	Basal-like
CTNNB1	0.02	1	0.27	Basal-like
CCNB1	0.02	0.28	$<10^{-3}$	Basal-like
STAT5A	0.02	0.06	0.7	Basal-like
MAPK14 (pT180)	0.06	0.72	0.71	Basal-like
RB1 (pS807)	0.08	0.33	0.93	Basal-like
ERBB2 (pY1248)	$<10^{-3}$	$<10^{-3}$	0.76	HER2
ERBB2	$<10^{-3}$	$<10^{-3}$	0.99	HER2
AKT1/AKT2/AKT3 (pS473)	0.03	0.09	0.7	HER2
AKT1/AKT2/AKT3	$<10^{-3}$	0.33	0.52	HER2/LumA
RPS6 (pS235)	0.03	0.81	0.05	HER2/LumB
AR	$<10^{-3}$	0.03	0.99	HER2/LumA/B
PDK1 (pS241)	0.01	0.22	0.16	LumA
PGR	0.13	$<10^{-3}$	0.09	LumA
PEA15	0.44	0.79	0.09	LumA
IGFBP2	0.09	0.49	0.29	LumB
RPS6 (pS240)	0.12	0.74	$<10^{-3}$	LumB
INPP4B	$<10^{-3}$	0.01	0.76	LumA/B
ESR1	$<10^{-3}$	0.04	0.54	LumA/B
GATA3	$<10^{-3}$	$<10^{-3}$	0.97	LumA/B
FN1	0.02	0.06	0.98	LumA/B
CAV1	0.02	0.05	0.52	LumA/B
CCND1	0.02	$<10^{-3}$	0.99	LumA/B
BCL2	0.1	0.04	0.99	LumA/B

FDR adjusted $P < 0.1$.

Inferred Protein Activity In Breast Cancer Cell Lines Can Be Used To Predict Drug Response

Experiment #4A

Tested whether our affinity regression model—trained on paired mRNA and RPPA data from breast cancer tumors—could be used to:

- Infer protein signaling activity in breast cancer cell lines from their mRNA expression profiles alone
- Whether these inferred protein signatures were useful for predicting drug sensitivity.

Used previously published gene expression data for 35 breast cancer cell lines with corresponding drug response data for 77 drugs quantified by growth inhibition (GI50)

Found that 45 out of 74 (61%) of the drugs produced variable responses across the cell lines (standard deviation of log-transformed GI50 across cell lines greater than 0.5)

Restricted analysis to these drugs. Out of 45 cell lines, 28 were luminal (ER+), and 15 of those were ERBB2-amplified.

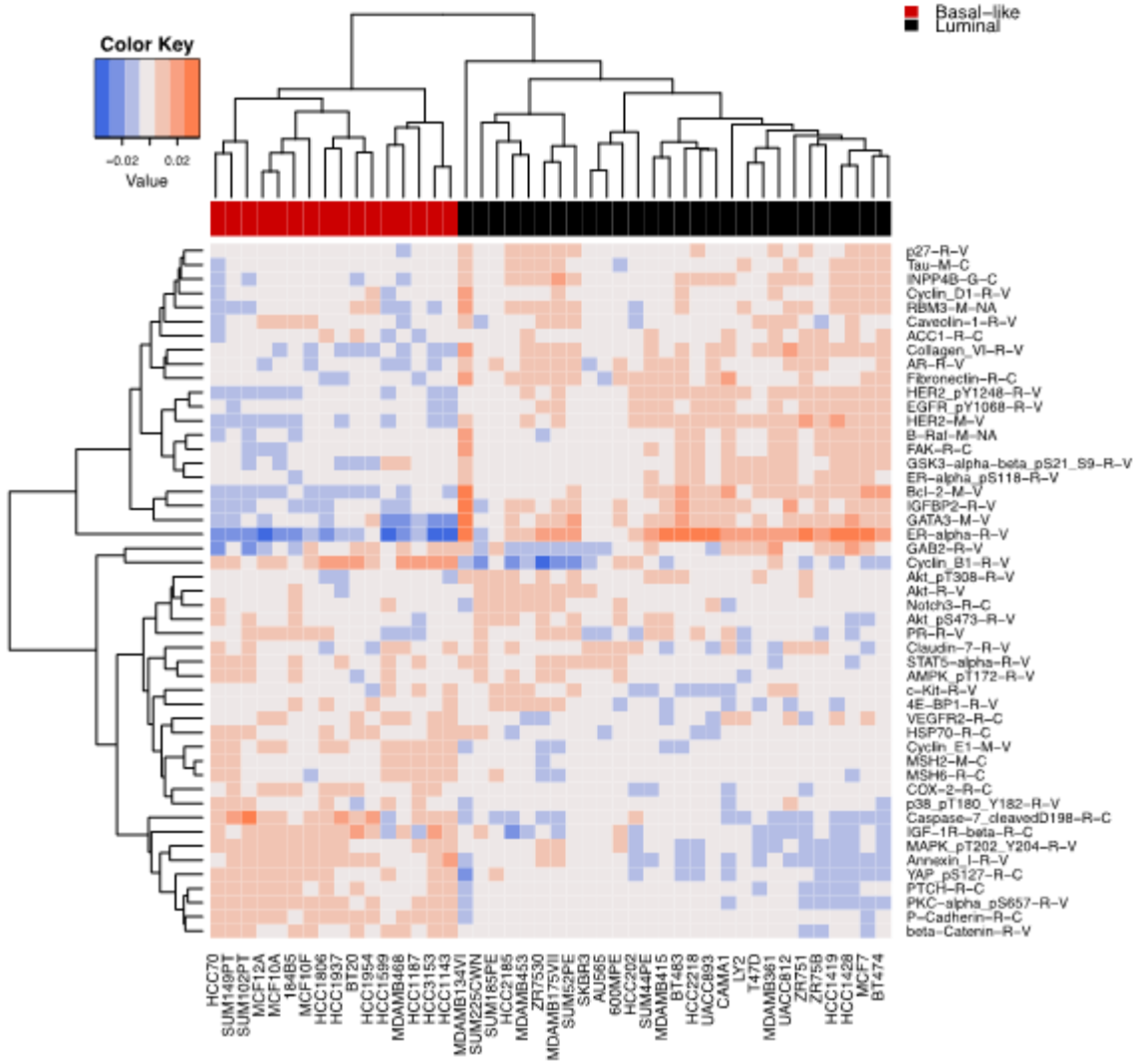
Experiment #4A

Used the TCGA-trained affinity regression model to infer protein activity profiles for individual cell lines (Y^TDW)

Applied unsupervised hierarchical clustering to these profiles, and confirmed that this clustering discriminated between basal-like and luminal subtypes for the breast cancer cell lines

In contrast, mapping the cell lines through randomized versions of the interaction matrix W did not correctly recover Basal-like vs. Luminal subtypes

Indicates that the model—and not only the initial mRNA expression profiles of the breast cancer cell lines—was crucial for segregating cell lines by subtype



Experiment #4A

Figure S8. Clustering of breast cancer cell lines by inferred protein activities.

Unsupervised hierarchical clustering of breast cancer cell lines by their inferred protein activity, using the TCGA affinity regression model, correctly distinguishes between Basal-like and Luminal cell lines.

Experiment #4B

Objective: Explore possible associations between inferred protein activity and drug response

Method: Computed Spearman rank correlations between (inferred) protein activity and drug GI50 for each (phospho) protein-drug pair over cell lines. Followed this up with clustering. To confirm the findings of clustering analysis, for each pair of drugs, asked whether ridge regression models trained to predict one drug's response would generalize to predict the other drug's response.

Results:

- Figure shows the two-way clustering of drugs and proteins by these pairwise Spearman rank correlations; drugs are clustered into groups according to the protein activities that correlate with their response.
- Several drugs with similar mechanisms of action or affecting a common signaling pathway clustered together.
- Several drugs commonly used in combination for the treatment of breast cancer were often found to cluster together in our analysis.
- Results align with several clinical trials demonstrating that inhibiting multiple targets that regulate cancer growth is more effective than monotherapy
- Results of transfer learning exercise from ridge regression models found similar relationships between drug sensitivities

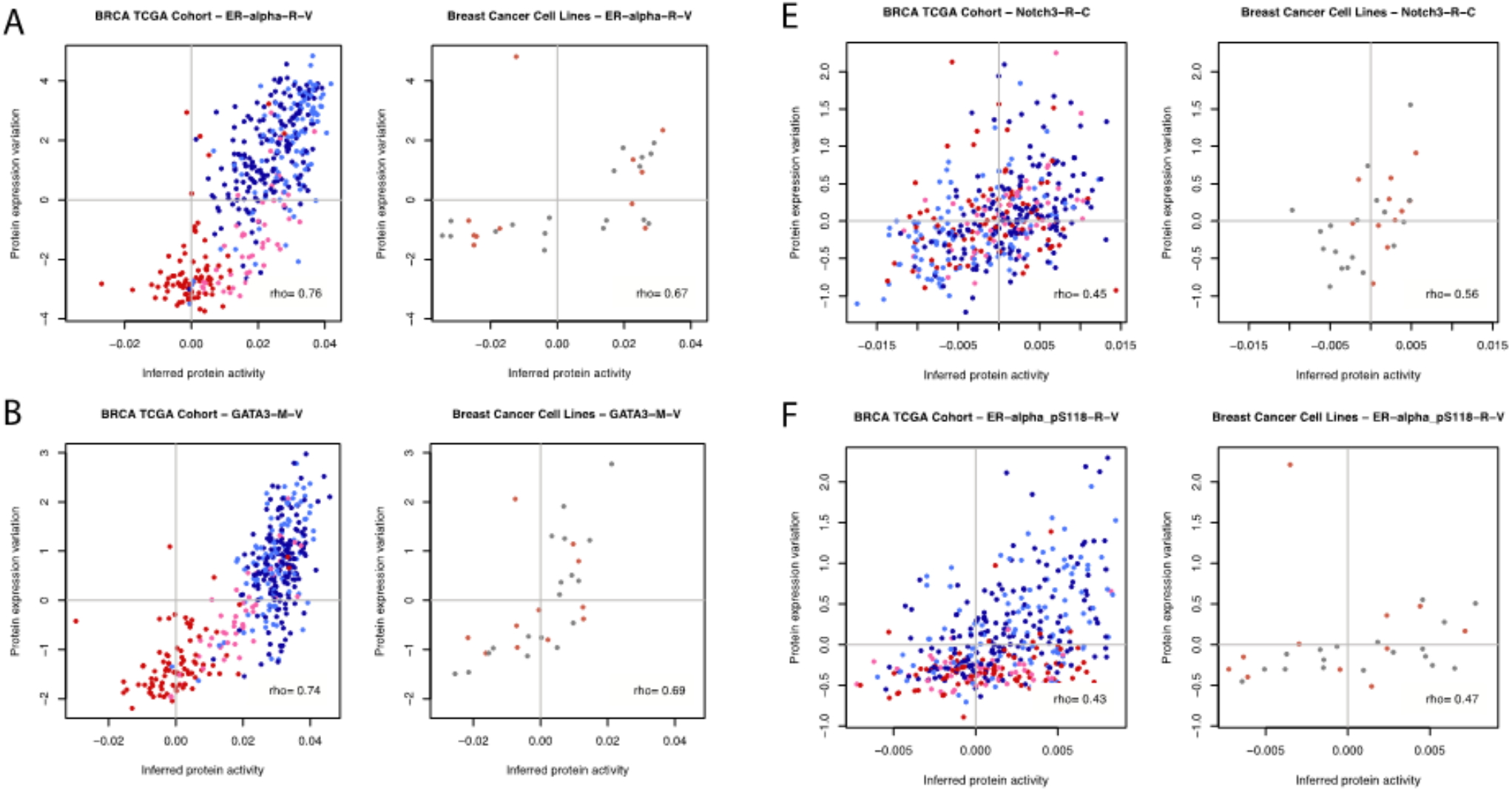


Figure S9. Correlation of inferred and measured protein activities. Correlation of inferred protein activity and measured protein variation across tumors (left) and breast cancer cell lines (right); for cell line data, we predict protein activity using the TCGA affinity regression model and measure protein expression using the TCPA resource. (Basal-like (red), HER2 (pink), LumA (dark blue), LumB (light blue), for tumors; luminal (black) and basal (coral) for breast cancer cell lines.)

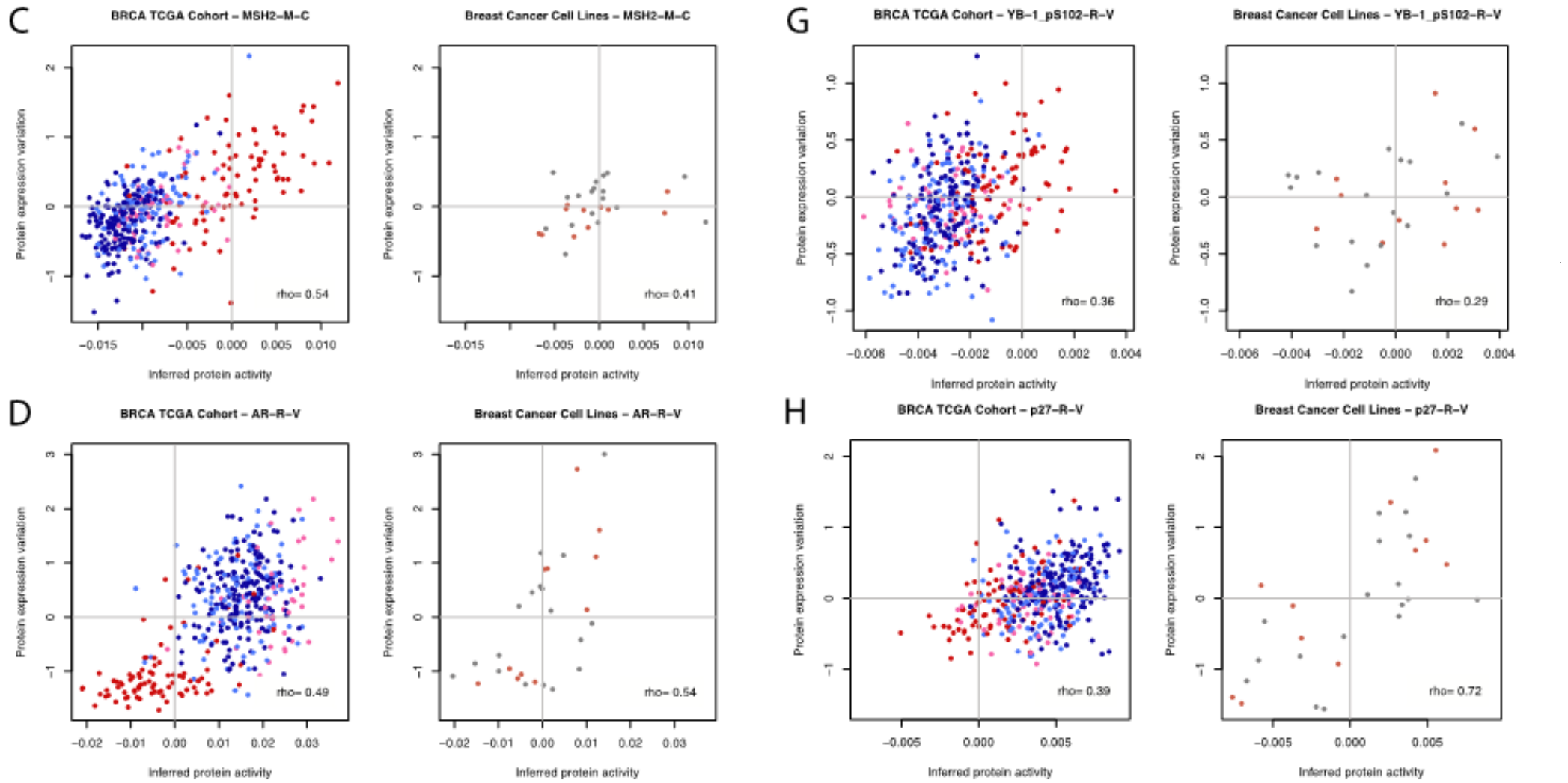


Figure S9. Correlation of inferred and measured protein activities. Correlation of inferred protein activity and measured protein variation across tumors (left) and breast cancer cell lines (right); for cell line data, we predict protein activity using the TCGA affinity regression model and measure protein expression using the TCPA resource. (Basal-like (red), HER2 (pink), LumA (dark blue), LumB (light blue), for tumors; luminal (black) and basal (coral) for breast cancer cell lines.)

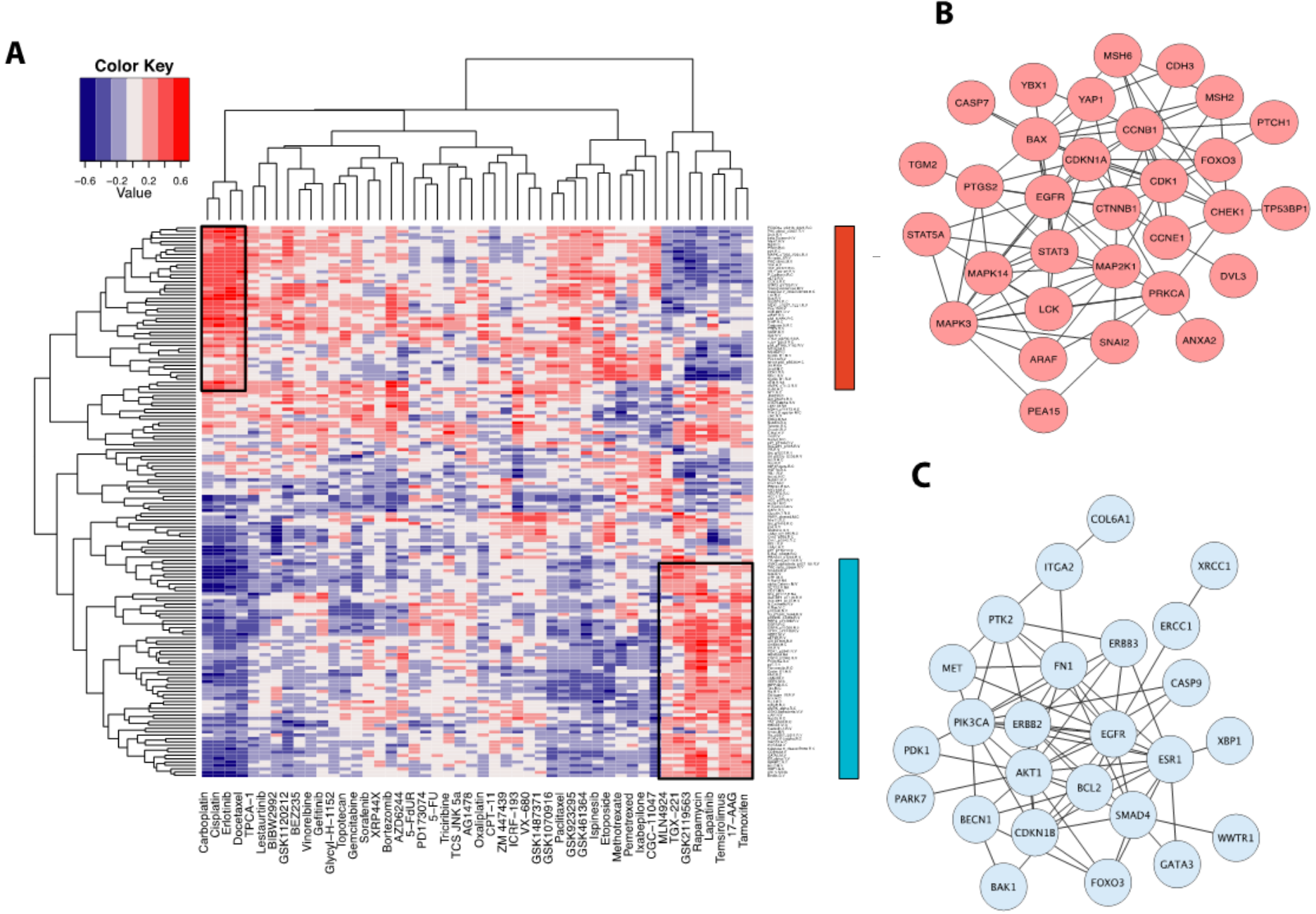
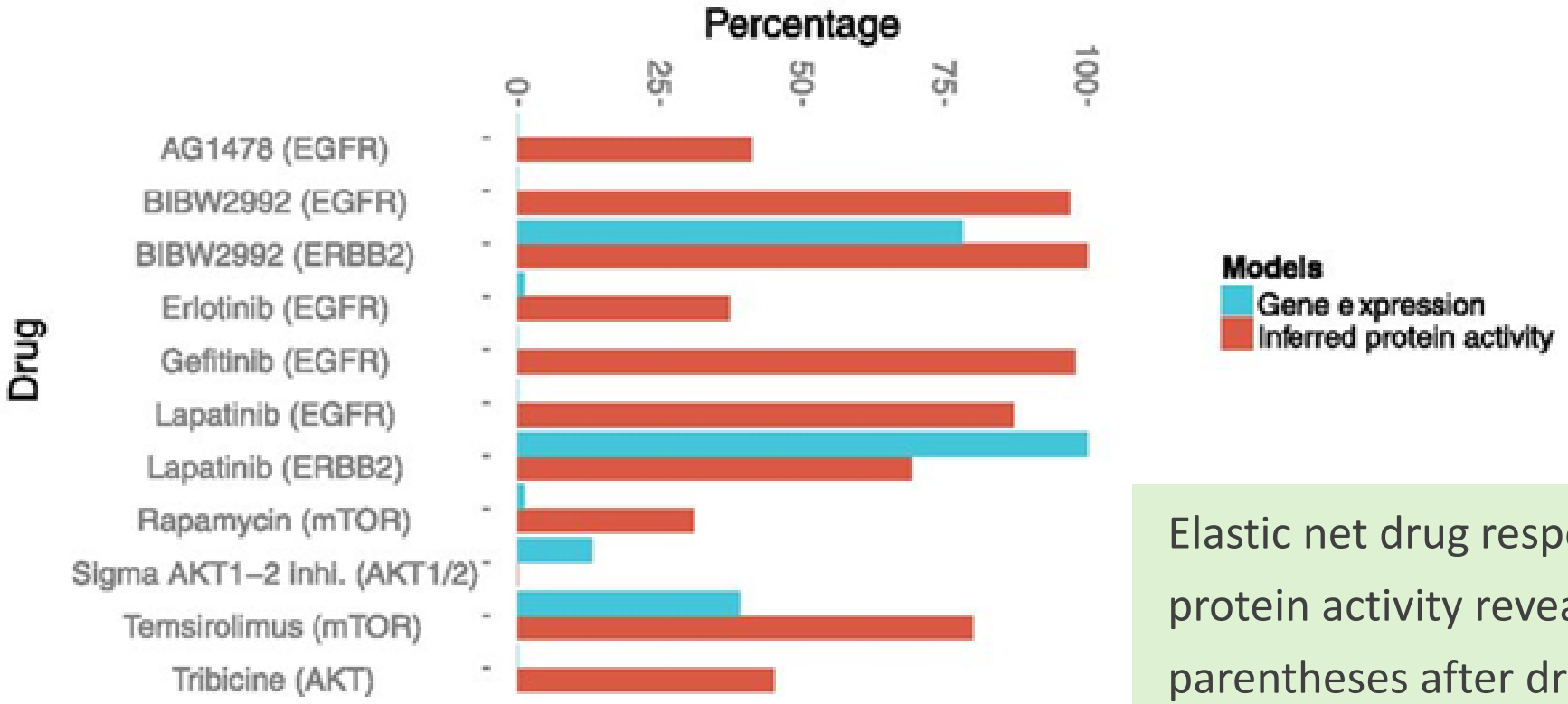


Figure S10. Correlation of inferred protein activities with drug responses in breast cancer cell lines.

(A) Heatmap revealing correlations between inferred protein activities of cell lines (rows) and drug responses (columns). Identified two clusters of drugs from unsupervised analysis (corresponding targets given in parentheses): a group consisting mostly of cytotoxic drugs including Carboplatin, Cisplatin, and Docetaxel, but also Erlotinib (EGFR), shown in (B); and a group of targeted therapies including Tamoxifen (ESR1), 17-AAG (HSP90), Temsirolimus (mTOR), Rapamycin (mTOR), Lapatinib (EGFR, ERBB2), and GSK2119563 (PIK3CA), shown in (C).

(B) and (C) Interaction maps using the STRING resource are constructed for proteins whose inferred activities are highly correlated with drug sensitivity for group (B) and (C).

TGCA affinity regression model infers signaling activity in breast cancer cell lines and predicts drug sensitivity



Elastic net drug response models built from inferred protein activity reveal drug targets (shown in parentheses after drug name) more often than models built using gene expression.

Experiment 4B

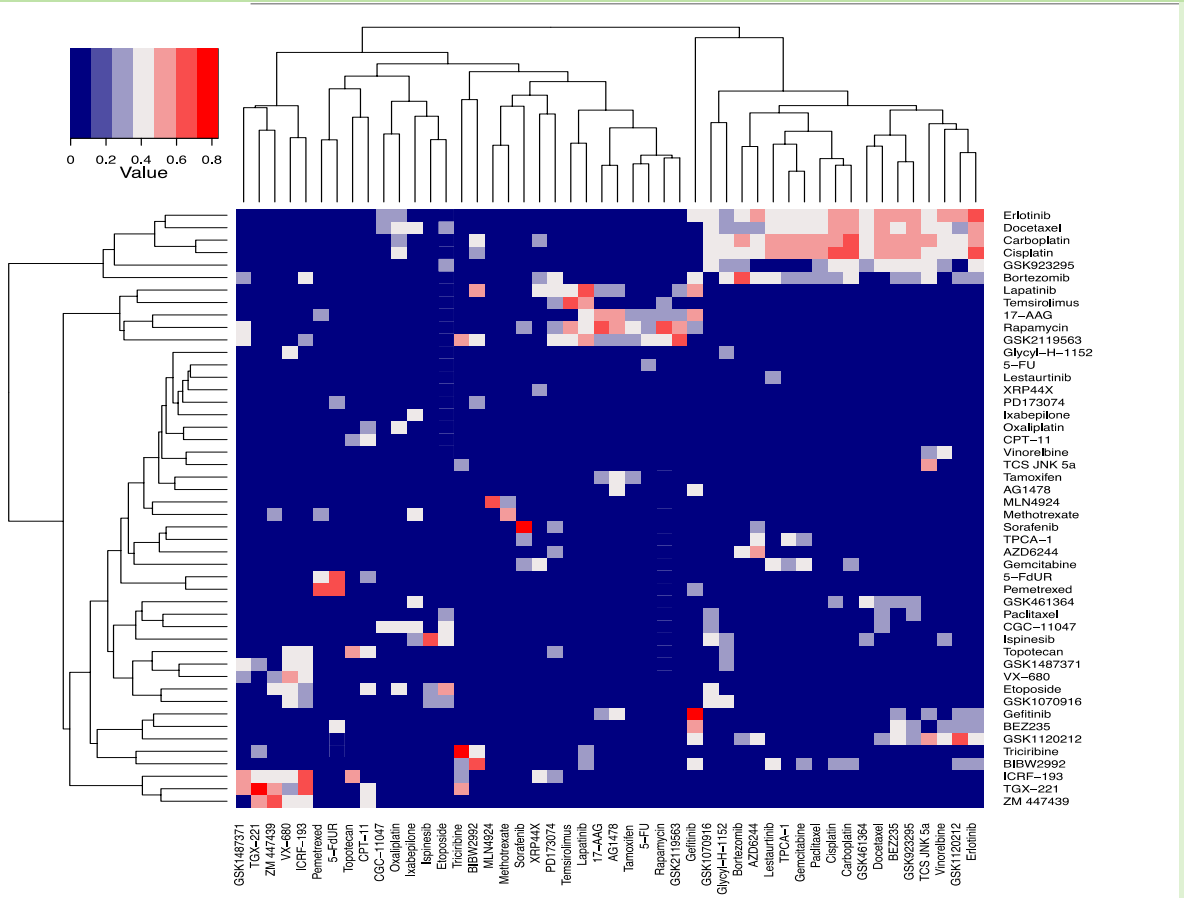


Figure S11. Transfer learning for drug response models.

Prediction performance of elastic net models for each drug (shown in columns) predicting drug response for all drugs (shown in rows); performance reported as Spearman correlations, with values below 0.3 set to 0.

Experiment #4C

Objective: Learn Predictive Signatures to Drug Response

Method: Trained an elastic net regression model for each drug separately using inferred protein activities as input features and log-transformed GI50 values as output values. As a baseline comparison method, used mRNA expression profiles as input features.

Results:

- The drug response signatures associated with inferred protein activities were more likely to include the drug target:
  - The protein activity drug signatures contained the drug target at least 10% of the time for 11 out of 14 targeted drugs (79%)
  - The mRNA drug signature contained the drug target at least 10 times in 100 iterations of training for only 4 out of 14 targeted drugs (28%)
- Use of inferred protein activities as features incurs some loss in prediction accuracy compared to mRNA features, perhaps due in part to the difference between tumor and cell line data.

Inferred Protein Activity of Luminal A Cohort Predicts Survival

Experiment #5

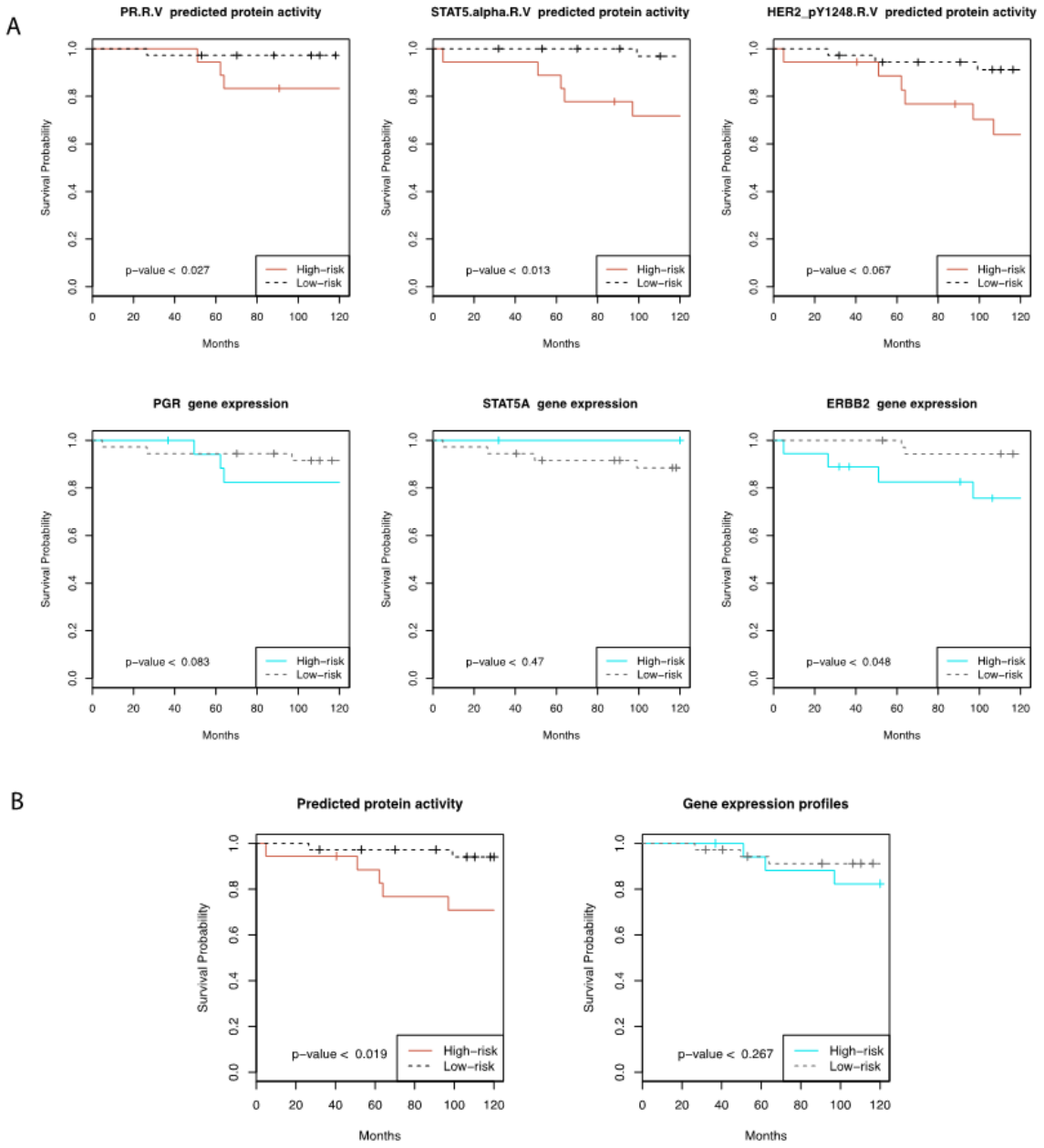
Objective: Determine whether inferred protein activities based on model could predict survival in patients with Luminal A breast cancers.

Method:

- Used the METABRIC cohort, which consists of a discovery set and validation set (n = 465 and 254 Luminal A tumors, respectively) with mRNA expression profiles and long-term clinical follow-up.
- Used the TCGA-trained affinity regression model to infer protein activity profiles of Luminal A samples in the METABRIC cohort (Y^TDW).
- Using the inferred protein activity, first identified proteins with univariate Cox $P < 0.001$ on the discovery set.
- Associations were tested by predicting the risk for each patient in the validation set using the univariate models and performing Kaplan-Meier survival analysis
- Built multivariate stepwise Cox regression models using the predicted protein activity and the gene expression profiles of the RPPA proteins on the discovery set.

Experiment #5: Results

- Univariate survival analysis: (1) For PGR and STAT5A, associated high protein activity with better overall survival (2) For high ERBB2 and phosphorylated ERBB2 (pY1248) showed a worse prognosis.
- Univariate models built from inferred protein activity can predict survival in the validation cohort but not models built from the gene expression levels of those proteins.
- Multivariate models: For the validation cohort, the model trained with inferred protein activities can predict survival but not the model trained on gene expression profiles corresponding to RPPA-profiled proteins.
- Further confirmed that the multivariate and most of the univariate survival results generalized to Luminal A patients in two other cohorts, TRANSBIG and NKI.
- **Claim:** Inferred protein activity from model can predict survival for Luminal A cohort



Experiment #5

Figure S15. Inferred protein activity predicts survival in patients with Luminal A breast cancers (TRANSBIG).

- Using inferred protein activity, a prognostic signature for overall survival was trained on the METABRIC discovery set. Kaplan–Meier survival curves reveals higher- versus lower-risk patients on the TRANSBIG datasets (Desmedt et al. 2007) using inferred protein activity (top panels) but not the corresponding gene expression (bottom panels) with (A) univariate Cox models for PGR, STAT5A and ERBB2 and (B) multivariate Cox models.