Partial Least Squares Regression

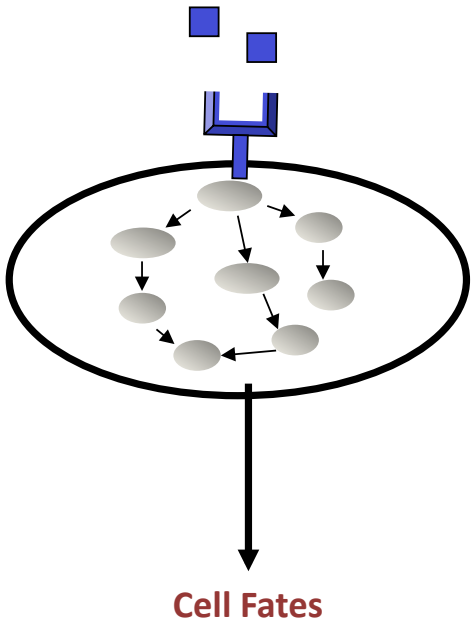
Aaron Meyer

Outline

- Administrative Issues
- Principal Components Regression
- Partial Least Squares Regression
- Some Examples
- Implementation

Adapted from slides by Pam Kreeger.

Common Challenge: Cue/Signal/Response Relationships



Cues

(e.g. growth factors, steroids, ECM)

Signals

(e.g. phosphorylation, transcription)

Due to time scale and multiple components, difficult to model with mechanistic detail

Response

(e.g., proliferation, apoptosis, migration, differentiation)

Many Methods for Relating a Signal to Response

Say we have some measurement from cells and how they respond:

$$[1, 2, 1.5, 5, 6, 7] \sim [5, 10, 7, 24, 31, 35]$$

From the variation we can see that:

- Low signal is correlated with low response
- High signal is correlated with high response

If we can find a quantitative correlation between the input and output, we can predict new outcomes for measurements we haven't yet seen.

Challenges with Univariate Relationships

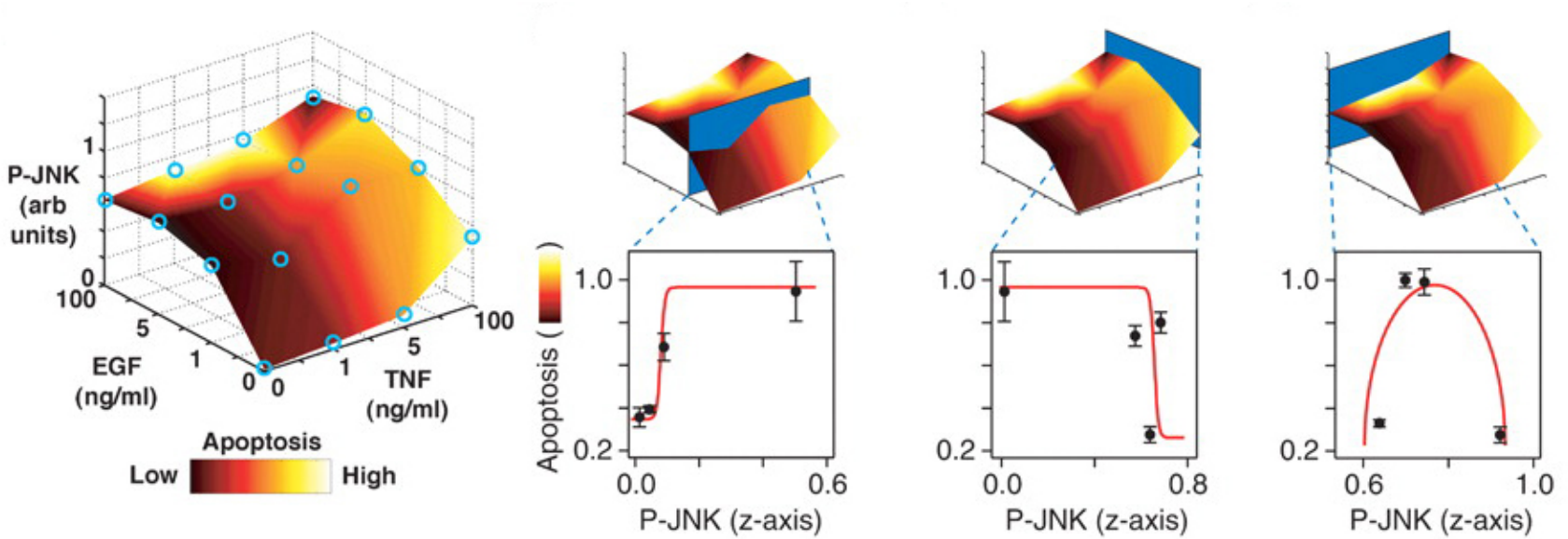


Figure: Janes, et al. Science, 2005

- The relationship between JNK activation and apoptosis appears to be highly context-dependent
  - Univariate relationships are often insufficient
  - Cells respond to an environment with multiple factors present

Notes about Methods Today

- Both methods are supervised learning methods, however have a number of distinct properties from others we will discuss.
- Learning about PLS is more difficult than it should be, partly because papers describing it span areas of chemistry, economics, medicine and statistics, with little agreement on terminology and notation.
- These methods will show one example of where the model and algorithm are quite distinct—there are multiple algorithms for calculating a PLSR model.

Multi-Linear Regression (MLR)

In biology we often have multiple signals and multiple responses that were measured:

$$y_1 = a_1 x_1 + b_1 x_2 + e_1$$

$$y_2 = a_2 x_1 + b_2 x_2 + e_2$$

This can be written more concisely in matrix notation as:

$$Y = XB + E$$

Where Y is a $n \times p$ matrix and X is a $n \times m$ matrix; minimizing E and solving for B:

$$B = (X^t X)^{-1} X^t Y$$

Underdetermined Systems

If n observations and m variables:

- $m < n$: no exact solution, least-squares solution possible
- $m = n$: one solution
- $m > n$: no unique solution unless we delete independent variables since $X^t X$ cannot be inverted
  - $m > n$ **is often the case in systems biology!**

If a model is underdetermined with multiple solutions, there are two general approches we can take:

- Regularization: We can use other information we know to focus on one answer
- Sampling: We can look at all possible models

Regularization

Today we will use regularization.

- We will assume the larger variation in the data is more meaningful.
- Therefore, we will assume that smaller changes are less important.
- This is a choice that must be correct for the relevant biological question at hand.

Principal Components Regression (PCR)

One solution - use the concepts from PCA to reduce dimensionality.

First step: **Simply apply PCA!**

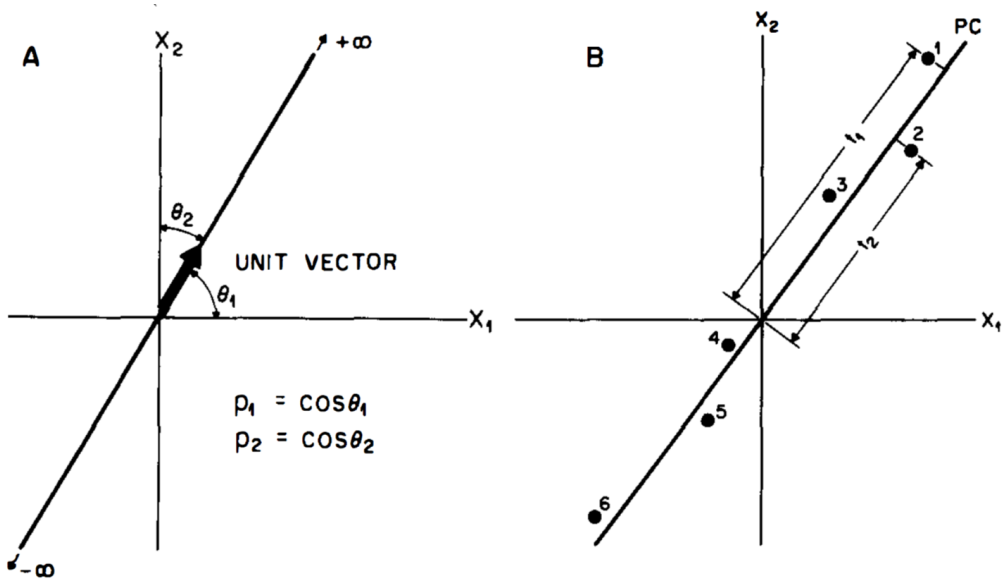


Figure: Geladi *Analytica Chimica Acta* 1986

Dimensionality goes from m to N_{comp}.

Principal Components Regression (PCR)

1) Decompose X matrix (scores T, loadings P, residuals E)

$$X = TP^T + E$$

2) Regress Y against the scores (Scores describe observations – by using them we link X and Y for each observation)

$$Y = TB + E$$

Challenge

How might we determine the number of components using our prediction?

Potential Problem

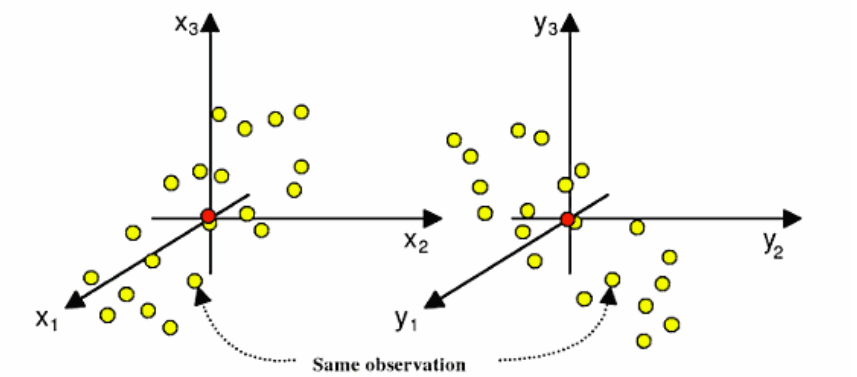
- PCs for the X matrix do not necessarily capture X-variation that is important for Y
  - So later PCs are going to be more important to regression
- Example: the first components capture signaling that is related to another cell fate, while the signals that co-vary for this particular y are buried in later components

How might we handle this differently?

PLSR

PLSR = partial least squares regression
OR projection to latent structures

Data has values in both X and Y spaces for each observation



Find PCs for both matrices (while emphasizing the parts of **X** that correlate with **Y**) – will use NIPALs algorithm to construct the principal components.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^t + \mathbf{F}$$

scores loadings residuals

Eriksson, et al. Multi- and Megavariate Data Analysis 2006

PLSR - NIPALs with Scores Exchanged

Steps for each component (h)

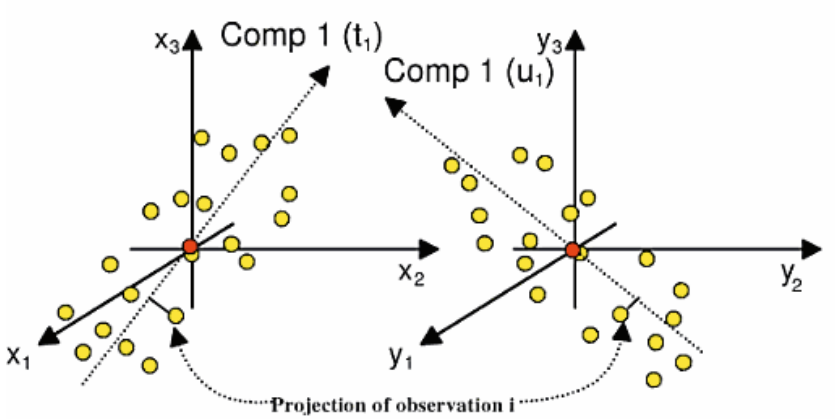
1) Find scores for **Y** ($\mathbf{u_h}$)
2) Use $\mathbf{u_h}$ to find the loadings for **X** ($\mathbf{p_h}$)
3) Use $\mathbf{p_h}$ to find scores for **X** ($\mathbf{t_h}$)
4) Use $\mathbf{t_h}$ to find **Y** loadings ($\mathbf{q_h}$)
5) Use $\mathbf{q_h}$ to calculate $\mathbf{u_h}$

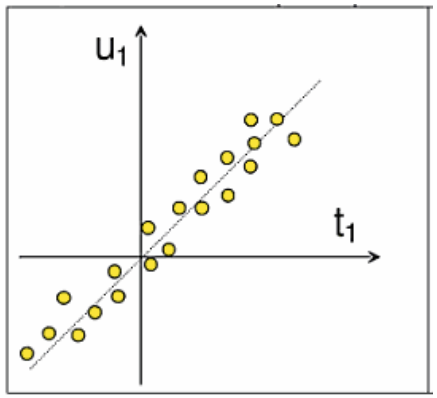
Repeat until get convergence
The scores vectors are related by:

$$\mathbf{u_h} = b_h \mathbf{t_h} \ (\mathbf{U=TB})$$

This allows us to relate **X** and **Y**:

$$\mathbf{Y = TBQ^t + F}$$



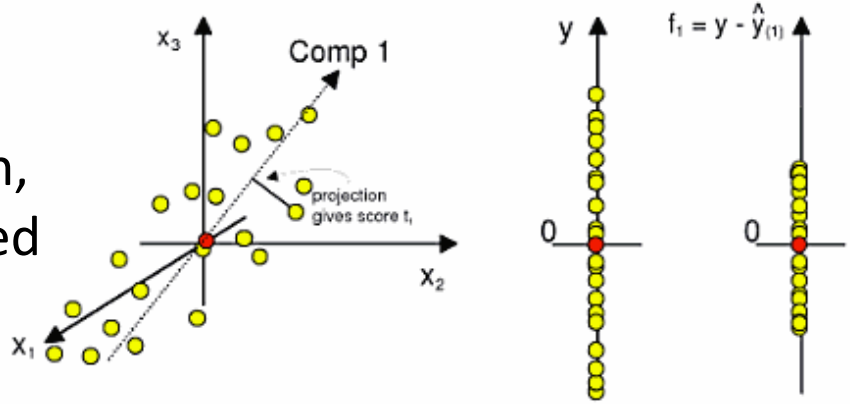


Eriksson, et al. Multi- and Megavariate Data Analysis 2006

PLSR - NIPALs with Scores Exchanged

By forcing the **X** and **Y** matrices to swap scores vectors we rotate the principal components toward the independent variables that link most strongly to the dependent variables.

The first component still captures the most information, and what is in PC1 is subtracted before PC2 is calculated.



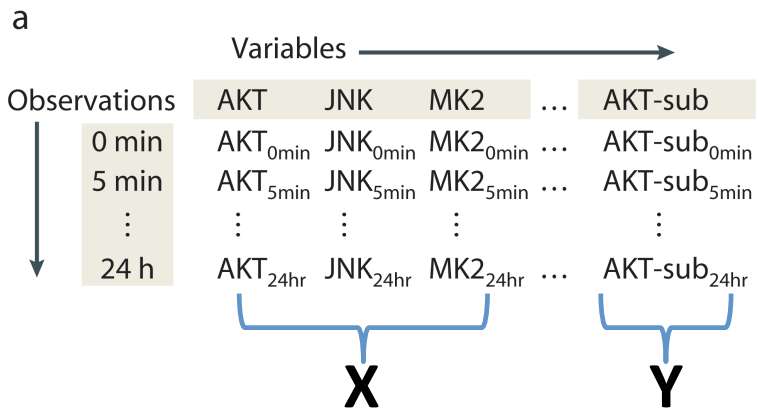
Note:

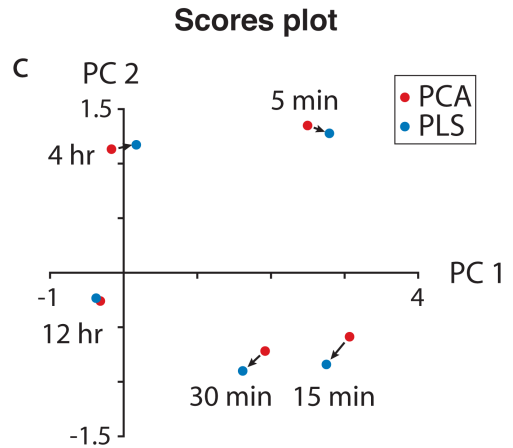
To obtain orthogonal components, **p** must be replaced with weights (**w**) in the NIPALS algorithm. See Geladi, *Anal Chim* 1986 for more detail.

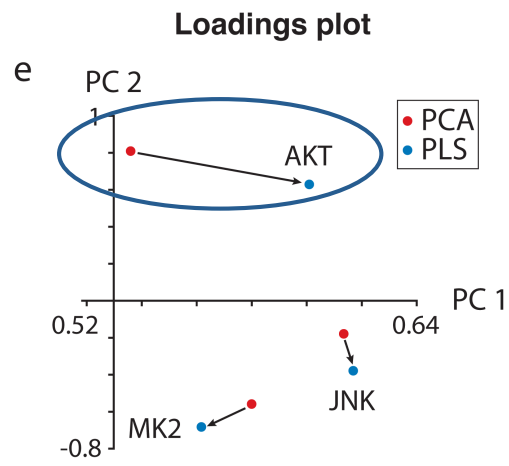
Data is mean-centered for PLSR. Unit variance scaling can also be applied if the magnitudes of **X** values are not considered important.

Eriksson, et al. Multi- and Megavariate Data Analysis 2006

Components in PLSR and PCA Differ







Compare 2 models:

1) PCA on the **X** matrix
2) PLSR of the **X** and **Y** matrix

For example, AKT has a larger loading in PC1 in PLSR than in PCA

Janes *Nat Rev MCB* 2006

Determining the Number of Components

The optimal model will have enough components to accurately fit data and be predictive, but remain simple enough for interpretation. Additionally, the model is subject to over-fitting constraints.

Three metrics are used to evaluate the utility of adding a new component (a):

R^2X: sum of squares for the variation in the **X** matrix

$$R^2X = 1 - \frac{\Sigma(X_{model,a} - X_{obs})^2}{\Sigma(X_{obs}^2)}$$

R^2Y: sum of squares for the variation in the **Y** matrix

$$R^2Y = 1 - \frac{\Sigma(Y_{model,a} - Y_{obs})^2}{\Sigma(Y_{obs}^2)}$$

Q^2Y: fraction of the total variation in the **Y** matrix that can be predicted

$$Q^2Y = [1.0 - \Pi(PRESS/SS)_a]$$

PRESS = Prediction Error Sum of Squares

1) Remove an individual data element (i,k)
2) Fit model
3) Predict the element i,k that was withheld

$$(observed_{i,k} - predicted_{i,k})^2$$

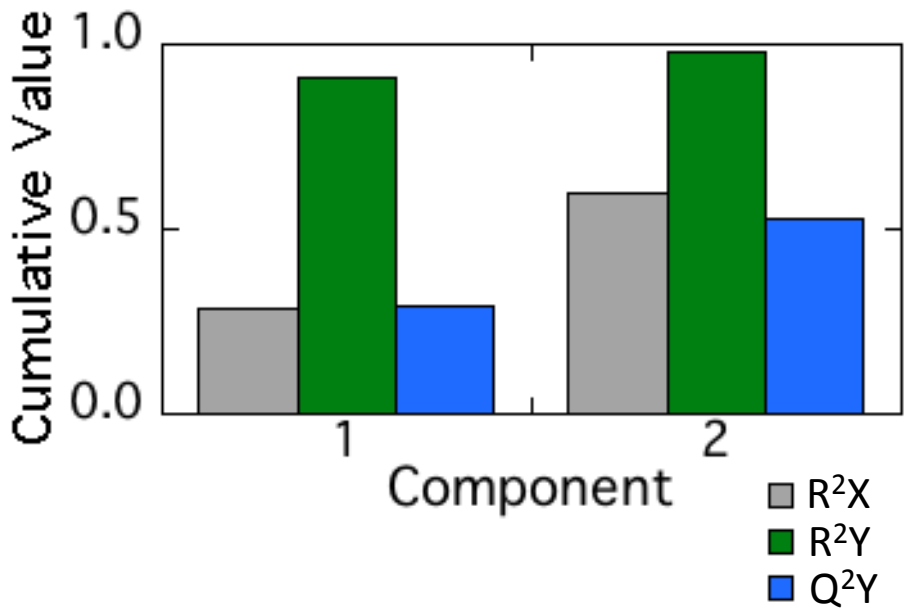
4) Repeat until each element has been withheld once and only once

Determining the Number of Components

Each component contributes to these metrics – we evaluate those contributions and the cumulative value to determine if adding a new component is beneficial (Q^2Y is prioritized in this evaluation).

With each new component, evaluate the change to the cumulative Q^2Y

- Q^2Y increases significantly (>0.05), keep the component and evaluate the effect of adding another component
- Q^2Y goes down or has minimal change, stop model at the previous component



Variants of PLSR

Discriminant PLSR

Tensor PLSR

Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks

Michael J. Lee,[1,2] Albert S. Ye,[2,3] Alexandra K. Gardino,[1,2] Anne Margriet Heijink,[1] Peter K. Sorger,[2,4] Gavin MacBeath,[2,4] and Michael B. Yaffe[1,2,*]

[1]Departments of Biology and Biological Engineering, David H. Koch Institute for Integrative Cancer Research

[2]Cell Decision Processes Center

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

[4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

*Correspondence: myaffe@mit.edu

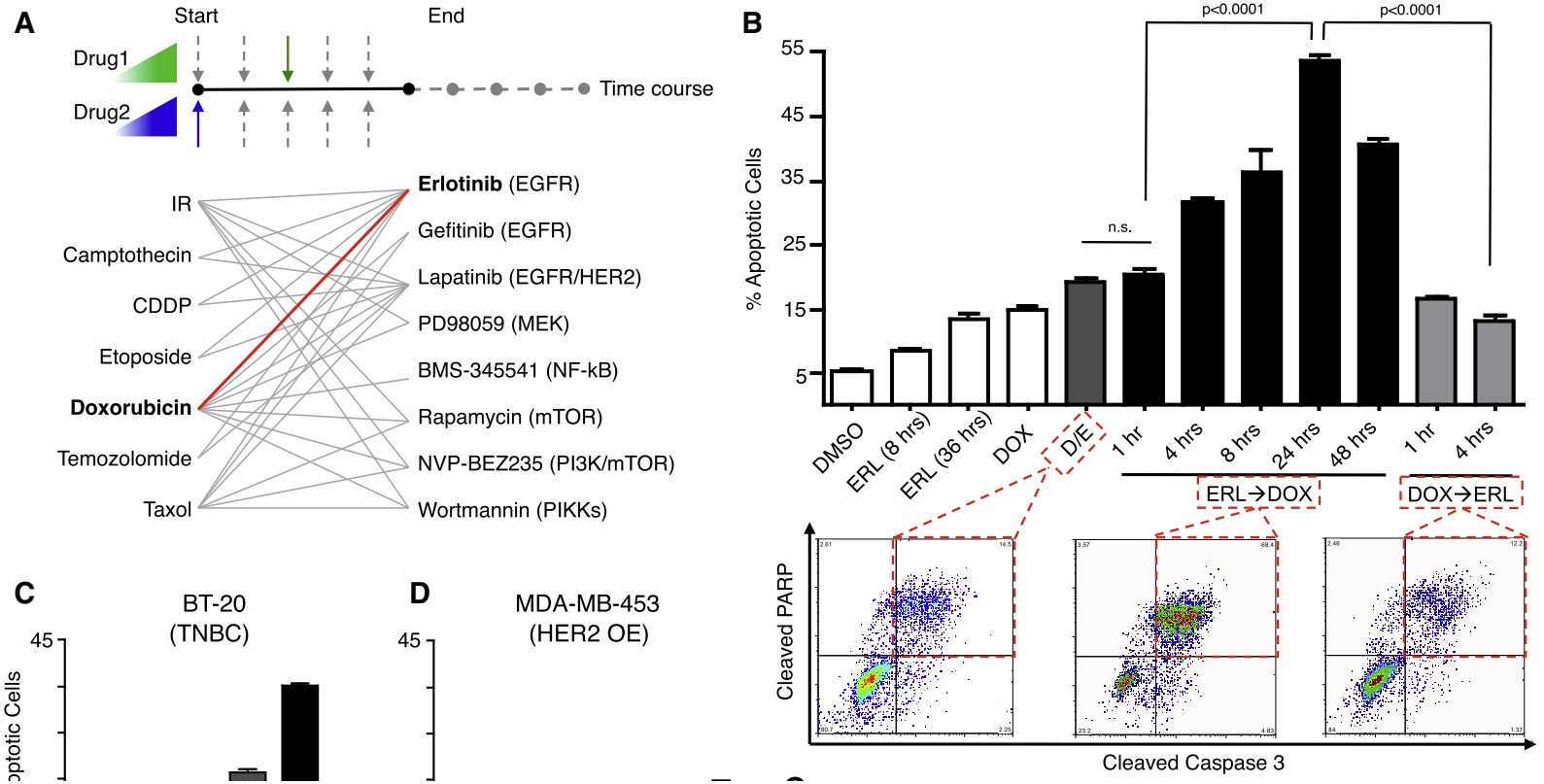
DOI 10.1016/j.cell.2012.03.031

SUMMARY

Crosstalk and complexity within signaling pathways and their perturbation by oncogenes limit component-by-component approaches to understanding human disease. Network analysis of how normal and oncogenic signaling can be rewired by drugs may provide opportunities to target tumors with high specificity and efficacy. Using targeted inhibition of oncogenic signaling pathways, combined with DNA-damaging chemotherapy, we report that time-staggered EGFR inhibition, but not simultaneous coadministration, dramatically sensitizes

cell death (Harper and Elledge, 2007). The DDR is highly interconnected with other progrowth and prodeath signaling networks, which function together to control cell fate in a nonlinear fashion due to multiple levels of feedback and crosstalk. Thus, it is difficult to predict a priori how multiple, often conflicting signals will be processed by the cell, particularly by malignant cells in which regulatory networks often exist in atypical forms. Predicting the efficacy of treatment and the optimal design of combination therapy will require a detailed understanding of how the DDR and other molecular signals are integrated and processed, how processing is altered by genetic perturbations commonly found in tumors, and how networks can be ''rewired'' using drugs individually and in combination (Sachs et al., 2005).

Application



Application

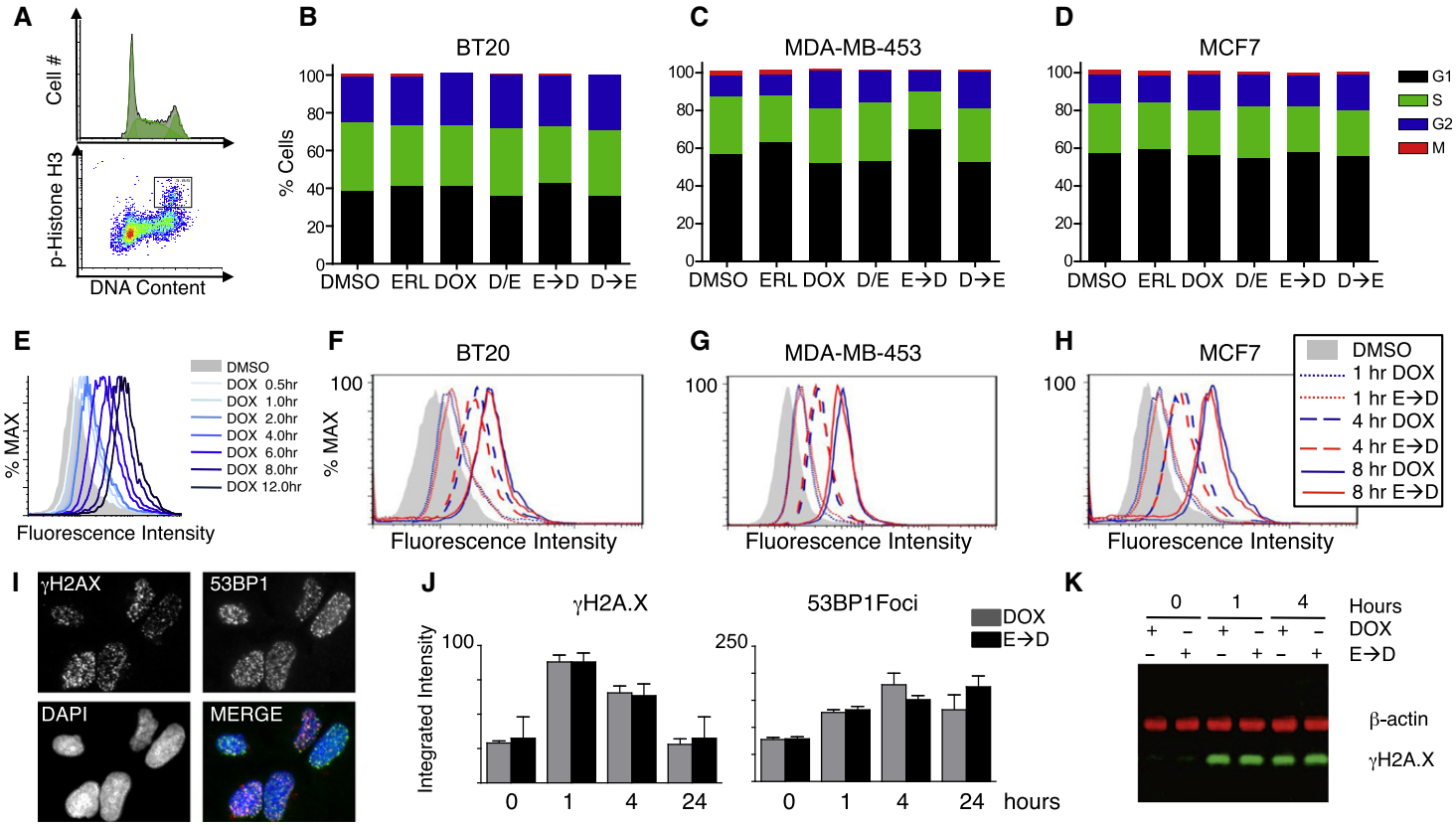


Figure 2. Prolonged Treatment with Erlotinib Does Not Change Cell-Cycle Profile, Doxorubicin Influx/Efflux, or the Level of DNA Damage

(A–D) Quantitative cell-cycle analysis. DNA content and the percentage of mitotic cells were measured by FACS. (A) Example FACS plots from untreated BT-20 cells. (B–D) Cell-cycle stage quantified from three experiments, each performed in duplicate. Cells were treated as in Figure 1, and data were collected at 6, 8, 12, 24, and 48 hr after DOX treatment. 8 hr data shown for each cell type.

(E–H) Doxorubicin retention measured by flow cytometry. (E) Sample time course of BT-20 cells treated with 10 μM DOX for the indicated times. (F–H) Cells treated with doxorubicin or pretreated with erlotinib for 24 hr prior to DOX (E→D). Cells were collected at 1, 4, or 8 hr after DOX exposure as indicated, and internal doxorubicin fluorescence was measured.

(I and J) Quantitative microscopy of the early DNA double-stranded break response. (I) Example image of cells treated with DOX for 1 hr and stained for γH2AX, 53BP1, or nuclear content (DAPI). (J) Integrated intensity per nucleus of γH2AX and 53BP1 foci was measured in BT-20 cells after the indicated treatments and times. Mean values ± SD from triplicate experiments shown.

(K) Western blot analysis of γH2AX in BT-20 cells. β-actin shown as a loading control.

Application

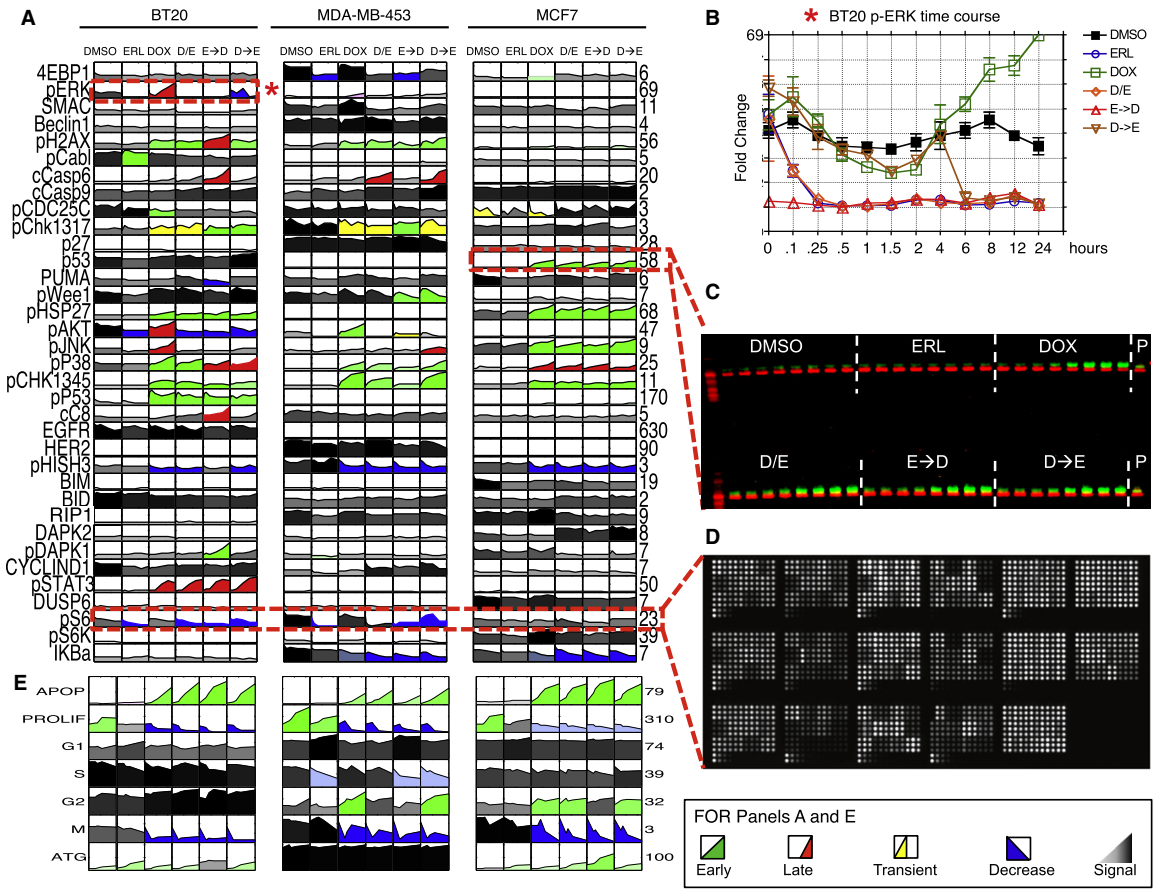


Figure 4. A Systems-Level Signal-Response Data Set Collected Using a Variety of High-Throughput Techniques

(A–D) (A) The complete signaling data set for three different breast cancer subtypes following combined EGFR inhibition and genotoxic chemotherapy treatments as in Figure 1. Each box represents an 8 or 12 point time course of biological triplicate experiments. Time course plots are colored by response profile, with early sustained increases in signal colored green, late sustained increases colored red, and transient increases colored yellow. Decreases in signal are colored blue. Signals that are not significantly changed by treatment are shaded gray to black with darkness reflecting signal strength. Numbers to the right of each plot report fold change across all conditions and/or cells. (B) Sample detailed signaling time course from (A), highlighted by dashed box and asterisk, showing p-ERK activation in BT-20 cells. Mean values ±SD of three experiments are shown. (C) Forty-eight-sample western blots analyzed using two-color infrared detection. Each gel contained an antibody-specific positive control (P) for blot-to-blot normalization. The example shown is one of three gels for total p53 in MCF7 cells (p53 in green; β-actin in red). (D) Reverse-phase protein lysate microarrays were used to analyze targets of interest when array-compatible antibodies were available. The slide shown contains ~2,500 lysate spots (experimental and technical triplicates of all of our experimental samples, and control samples used for antibody calibration), probed for phospho-S6.

(E) The complete cellular response data set, colored as in (A).

Application

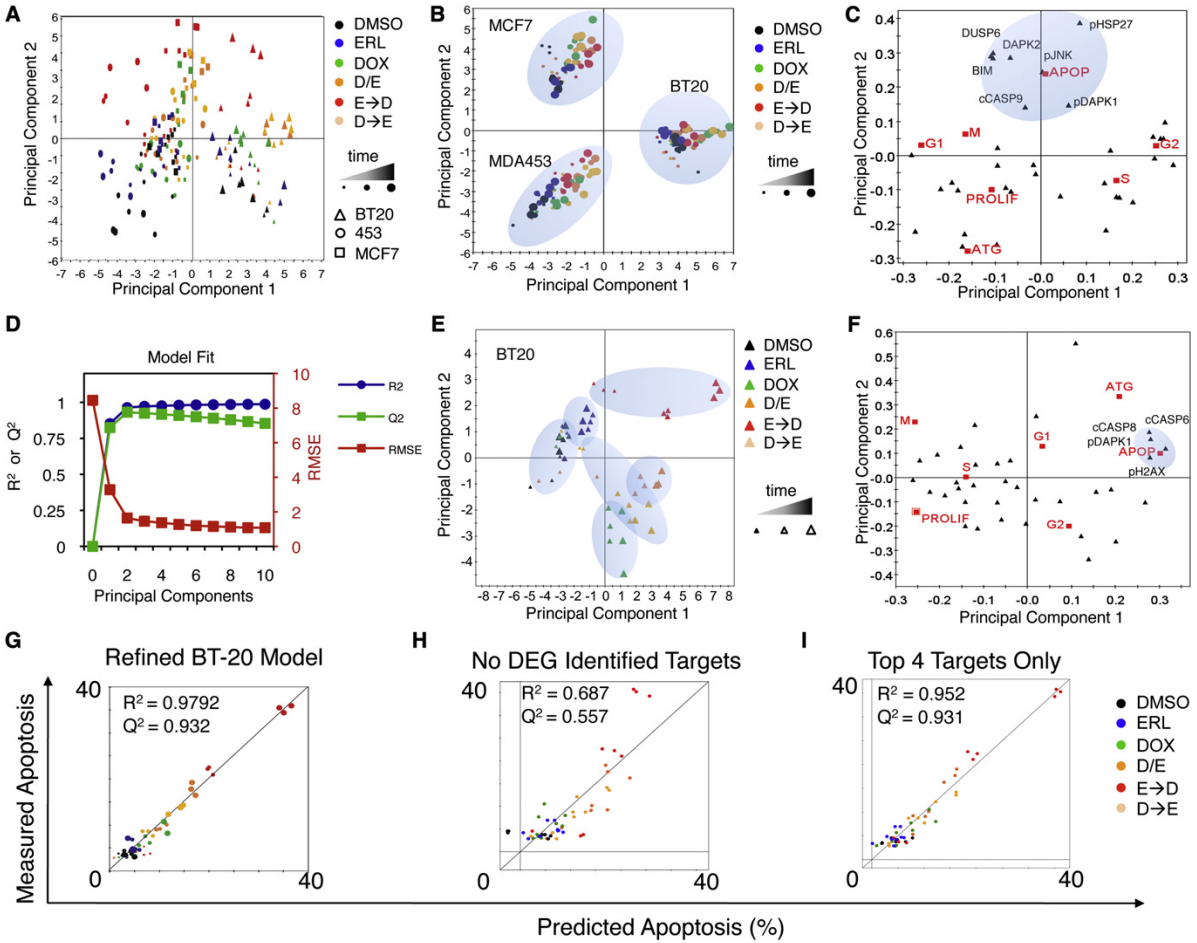


Figure 5. A PLS Model Accurately Predicts Phenotypic Responses from Time-Resolved Molecular Signals

(A) Principal components analysis of covariation between signals. Scores plot represents an aggregate measure of the signaling response for each cell type under each treatment condition at a specified time, as indicated by the colors and symbols in the legend.

(B and C) Scores and loadings for a PLS model. (B) Scores calculated and plotted as in (A), except the principal components now reflect covariation between signals and responses. (C) PLS loadings plotted for specific signals and responses projected into principal component space.

(D–I) BT-20 cell line-specific model calibration. (D) R^2, Q^2, and RMSE for BT-20 models built with increasing numbers of principal components. (E and F) Scores and loadings plots, respectively, for a two-component model of BT-20 cells. (G–I) Apoptosis as measured by flow cytometry or as predicted by our model using jack-knife cross-validation. R^2 reports model fit, and Q^2 reports model prediction accuracy. (G) Final refined model of apoptosis in BT-20. (H) BT-20 model minus targets identified as DEGs in microarray analysis. (I) Model using only the top four signals: c-caspase-8, c-caspase-6, p-DAPK1, and pH2AX.

Application

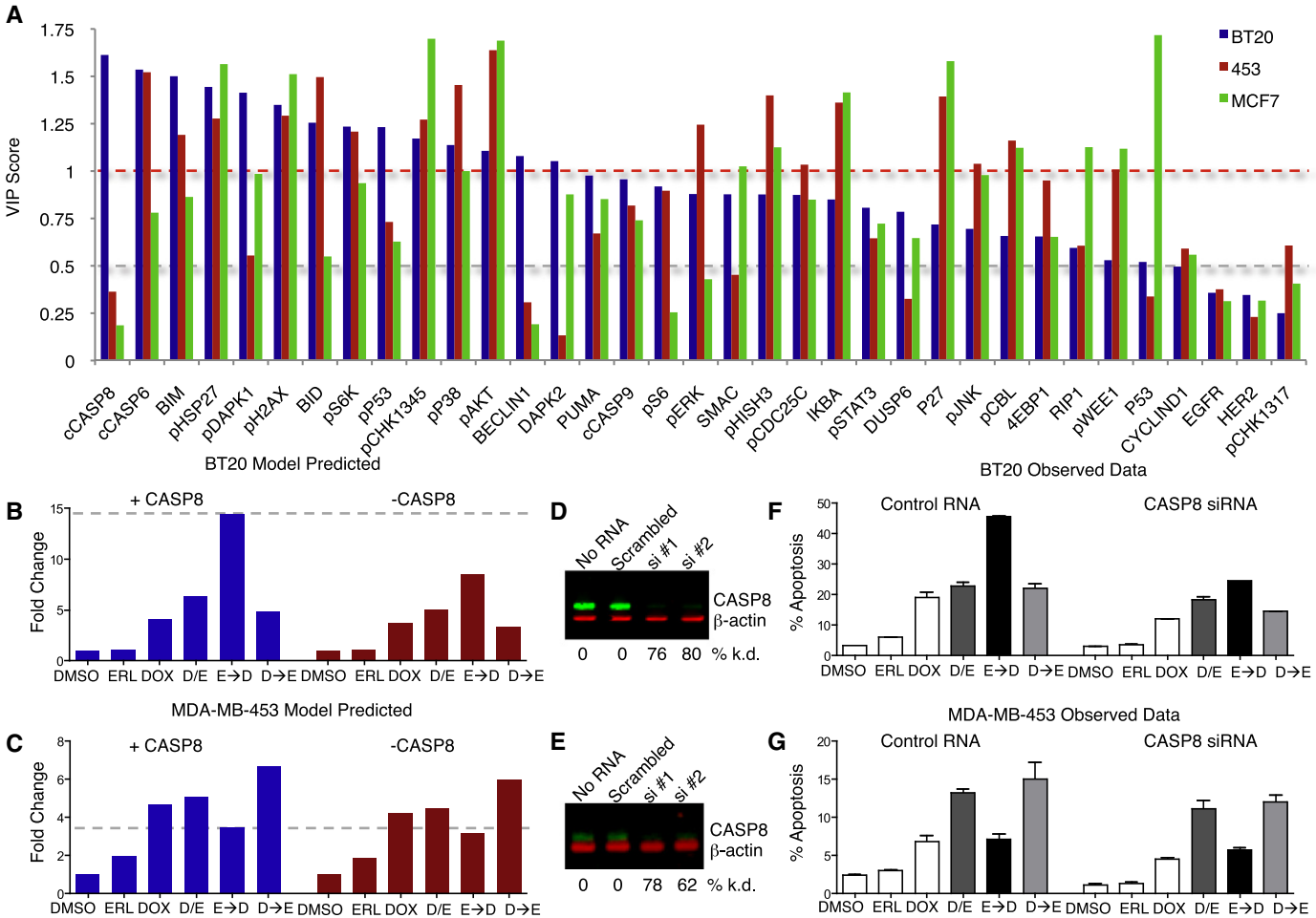


Figure 6. Enhanced Sensitivity to Doxorubicin Is Mediated by Caspase-8 Activation

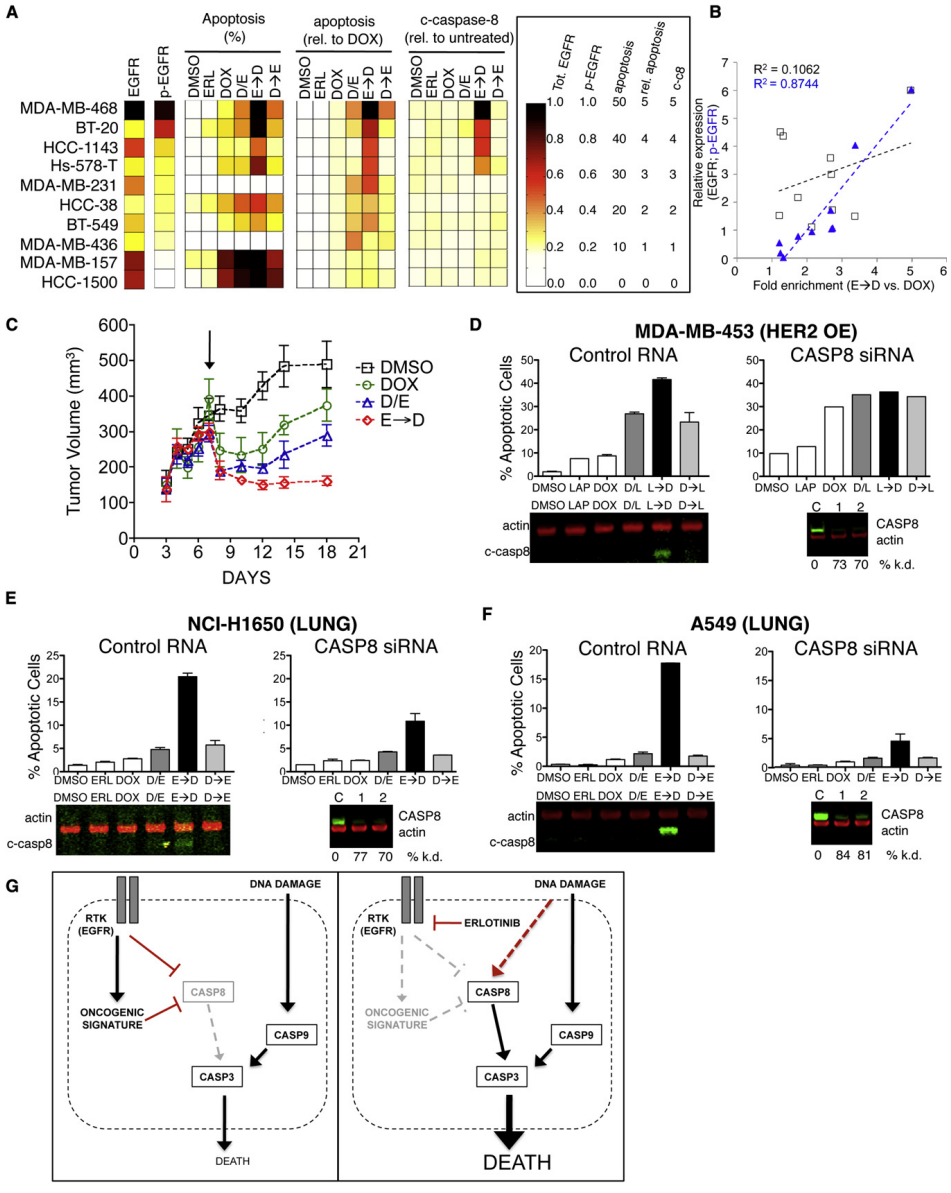
(A) VIP scores for predicting apoptosis plotted for each cell line-specific PLS model. VIP score >1 indicates important *x* variables that predict *y* responses, whereas signals with VIP scores <0.5 indicate unimportant *x* variables.

(B and C) Model-generated predictions of apoptosis with (blue) or without (red) caspase-8 activation 8 hr after the indicated treatments in BT-20 (B) and 453 (C).

(D and E) Western blot verifying caspase-8 knockdown in BT-20 (D) and 453 (E).

(F and G) Measured apoptosis 8 hr after the indicated treatment in cells expressing control RNA or caspase-8 siRNA. (F) BT-20. (G) 453. In both (F) and (G), apoptotic values represent mean response ±SD from both siRNAs, each in duplicate.

Application



Practical Notes

PCR

- sklearn does not implement PCR directly
- Can be applied by chaining `sklearn.decomposition.PCA` and `sklearn.linear_model.LinearRegression`
- See: https://scikit-learn.org/stable/auto_examples/plot_digits_pipe.html

PLSR

- `sklearn.cross_decomposition.PLSRegression`
  - Uses `M.fit(X, Y)` to train
  - Can use `M.predict(X)` to get new predictions
  - `PLSRegression(n_components=3)` to set number of components on setup
  - Or `M.n_components = 3` after setup

https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html

Summary

PLSR

- Maximizes the covariance
- Takes into account both the dependent (Y) and independent (X) data

PCR

- Uses PCA as initial decomp. step, then is just normal linear regression
- Maximizes the variance explained of the independent (X) data

Summary

Interpreting PLSR

- R^2X, R^2Y, Q^2Y (maximum value of 1)
- Using Q2Y to determine number of components
  Scores/loadings
- DModY (lower = better predicton)
- VIP (>1 indicates important)
- **Ultimately, these metrics are seconary to whether a model works upon crossvalidation**

Summary

- PLSR is amazingly well at prediction
  - This is **incredibly** powerful
- Interpreting **WHY** PLSR predicts something can be very challenging