Molecular Signatures for Cancer Prognosis

- In high-dimensional settings, strong biological priors can have a huge impact on data modeling.
- Using in-vitro biology, signatures of genes are hypothesized to play a role in cancer prognosis
- The signature is used to score each human cancer sample in a separate study. Hence they are ‘unsupervised’ with respect to our training data.
- These signatures (each one degree of freedom), are then used and compared to traditional prognostic factors in our training data
Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival

Howard Y. Chang\textsuperscript{a,b,c}, Dimitry S. A. Nuyten\textsuperscript{c,d,e}, Julie B. Sneddon\textsuperscript{b}, Trevor Hastie\textsuperscript{f}, Robert Tibshirani\textsuperscript{f}, Therese Sørlie\textsuperscript{b,g}, Hongyue Dai\textsuperscript{h,i}, Yudong D. He\textsuperscript{h,i}, Laura J. van't Veer\textsuperscript{d,j}, Harry Bartelink\textsuperscript{e}, Matt van de Rijn\textsuperscript{3}, Patrick O. Brown\textsuperscript{b,k,l}, and Marc J. van de Vijver\textsuperscript{d,l}

\textsuperscript{a}Program in Epithelial Biology, Departments of \textsuperscript{b}Biochemistry, \textsuperscript{f}Health Research and Policy, and \textsuperscript{i}Pathology, and \textsuperscript{k}Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; Departments of \textsuperscript{d}Diagnostic Oncology and \textsuperscript{e}Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands; \textsuperscript{h}Rosetta Inpharmatics, Seattle, WA 98109; and \textsuperscript{g}Norwegian Radium Hospital, 0310 Oslo, Norway

Contributed by Patrick O. Brown, January 5, 2005

Based on the hypothesis that features of the molecular program of normal wound healing might play an important role in cancer metastasis, we previously identified consistent features in the transcriptional response of normal fibroblasts to serum, and used this “wound-response signature” to reveal links between wound healing and cancer progression in a variety of common epithelial tumors. Here, in a consecutive series of 295 early breast cancer patients, we show that both overall survival and distant metastasis-free survival are markedly diminished in patients whose tumors expressed this wound-response signature compared to tumors that did not express this signature. A gene expression centroid of the wound-response signature provides a basis for prospectively assigning a prognostic score that can be scaled to suit different clinical purposes. The wound-response signature improves risk stratification independently of known clinico-pathologic risk factors and previously established prognostic response” (CSR) genes and their canonical expression pattern in fibroblasts activated with serum, the soluble fraction of clotted blood and an important initiator of wound healing \textit{in vivo}. The CSR genes were chosen to minimize overlap with cell cycle genes, but instead appeared to represent other important processes in wound healing, such as matrix remodeling, cell motility, and angiogenesis, processes that are likely also to contribute to cancer invasion and metastasis. In several common epithelial tumors such as breast, lung, and gastric cancers, expression of the wound-response signature predicted poor overall survival and increased risk of metastasis (10). These initial findings demonstrate the promise of using hypothesis-driven gene expression signatures to provide insights from existing gene expression profiles of cancers. However, as in other methodologies, reproducibility and scales for interpretation need to be evaluated before this strategy can be generally adopted for biologic dis-
Chang et al examined the transcriptional response of normal fibroblasts to serum \textit{in vitro}.

They identified a set of approximately 400 “Core Serum Response” genes that showed a wound response in a subset of the samples.

The set of averages of each gene across samples in the subset gives a profile of up- and down-regulated genes.

Any future sample (from a patient) can be scored for wound signature by computing the correlation of the expression of the corresponding genes with this profile.

They evaluated this signature on an independent sample of 295 breast cancer samples (Netherlands Cancer Institute).
Fig. 1. Performance of a "wound response" gene expression signature in predicting breast cancer progression

Fig. 2. A scalable wound-response signature as a guide for chemotherapy

Fig. 3. Integration of diverse gene expression signatures for risk prediction

Statistical Analysis

• Compared with traditional prognostic factors, using multivariate Cox model — tumor grade and size, lymph-node status, ER status, .... Summary later.

• Examine nature of wound signature score using semi-parametric methods [Hastie & Tibshirani, *Generalized Additive Models*, 1990]

• Using subgroups and scaling of the wound score, we showed that wound signature offers independent prognostic information, and could potentially spare 30% of women from chemotherapy.
Contribution of the wound score to the log-hazard in the proportional hazards model:

\[
\log \lambda(T, X, W) = \log \lambda_0(T) + X^T \beta + f(W),
\]

where \( f(W) \) is modeled by cubic splines (black), or binary (blue).
Hypoxia Signature

lead author Jen-Tsan Ashley Chi, now at Duke MC

- About 250 genes showing response to hypoxia \textit{in vitro} in cultured epithelial cells.

- They saw evidence on Stanford data that tumors with cells showing a strong response to hypoxia were associated with bad outcome.

- Here the signature is obtained by simply averaging the corresponding genes for each cancer patient.
Hypoxia and Wound have low correlation.

70 Gene score is a supervised signature on these data, developed by original authors (van’t Veer et al, Nature 2002).

We focus on Wound and Hypoxia.
Semi-parametric Cox Models

Cox Model: Survival

Log Relative Risk vs. Hypoxia Score (mean)

Cox Model: Time to Recurrence

Log Relative Risk vs. Hypoxia Score (mean)
Kaplan-Meier Curves

Survival

Time to Recurrence

P = 3.1e−06

P = 1.4e−04

Hypoxia Score <0

Hypoxia Score >0
## Marginal and Partial Effects — Survival

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>Marginal</th>
<th>Partial</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angio Invasion</td>
<td>2</td>
<td>11.7</td>
<td>6.1</td>
<td>0.0660602</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>8.8</td>
<td>11.4</td>
<td>0.0013773**</td>
</tr>
<tr>
<td>Diameter</td>
<td>1</td>
<td>12.2</td>
<td>3.1</td>
<td>0.0963076</td>
</tr>
<tr>
<td>Node Status</td>
<td>2</td>
<td>6.9</td>
<td>1.7</td>
<td>0.4621051</td>
</tr>
<tr>
<td>Grade</td>
<td>2</td>
<td>43.7</td>
<td>2.6</td>
<td>0.3142204</td>
</tr>
<tr>
<td>ER Status</td>
<td>1</td>
<td>27.4</td>
<td>8.7</td>
<td>0.0053109**</td>
</tr>
<tr>
<td>Mastectomy</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
<td>0.5943224</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>1</td>
<td>1.1</td>
<td>1.6</td>
<td>0.2341016</td>
</tr>
<tr>
<td>Hormonal</td>
<td>1</td>
<td>1.8</td>
<td>0.1</td>
<td>0.7869375</td>
</tr>
<tr>
<td>Hypoxia</td>
<td>1</td>
<td>22.6</td>
<td>8.6</td>
<td>0.0055803**</td>
</tr>
<tr>
<td>Wound</td>
<td>1</td>
<td>39.9</td>
<td>15.8</td>
<td>0.0001708***</td>
</tr>
</tbody>
</table>

As percent of the total deviance explained (89.5) in Cox model.
Signature Summary

- Wound signature provides an additional 18.7% in prognostic power in the Cox model (when added to all the other factors), and surpasses them all.

- Hypoxia signature adds 9.4%, similar to ER status and surpassed only by Age.

\[
\begin{array}{lcccc}
\text{Df} & \text{Marginal} & \text{Partial} & \text{Partial.add} \\
\text{Traditional} & 12 & 76.4 & 42.2 & 55.2 \\
\text{Signatures} & 2 & 57.8 & 23.6 & 40.8 \\
\end{array}
\]

- Wound and Hypoxia signature account for an additional 40.8% in prognostic power.

- External signatures avoid issues of overfitting associated with supervised signatures.
Pre-validation

• signatures above were constructed from an external source, without knowledge of the training set

• what if a predictor $Z$ was instead derived from training set? Can it be compared on an equal footing to external predictors such as age, gender etc?

• No! $Z$ has an unfair advantage since it has already seen the outcome in the training set.
Pre-validation- continued

- In pre-validation we divide the data into say 10 folds. For each $i$, we construct the predictor $Z_i$ from the other 9 folds that do not contain observation $i$.

- Doing this for all $i$ gives the pre-validated predictor $\tilde{Z}_i$.

- Then it’s fair to compare $\tilde{Z}_i$ to age, gender, etc. on the training set.

regression

outcome

Expression data

pre-validated predictor

clinical predictors

cases

omitted part

logistic regression

cases

predictor

pre-validated genes

cases

predictors

clinical

t~

Expression data

X

y

Expression data

X

y

Expression data

X

y

Expression data

X

y

Expression data

X

y
Schematic of pre-validation process. The cases are divided up into (say) 10 equal-sized groups. The cases in one group are left out, and a microarray predictor is derived from the expression data of the remaining cases. Evaluating this predictor on the left-out cases yields the pre-validated predictor $\tilde{z}$ for those cases. This is done for each of the 10 groups in turn, producing the pre-validated predictor $\tilde{z}$ for all cases. Finally, $\tilde{z}$ can be included along with clinical predictors in a logistic regression model, in order to assess its relative strength in predicting the outcome.