Van’t Veer et al. (2002)

Gene expression profiling predicts clinical outcome of breast cancer

**Background**

- Breast cancer pts have different response to treatments
- Strongest predictor (metastases; lymph node grade) fails to classify tumors by clinical behavior
- Chemotherapy reduces risk by 1/3, but 70-80% pts receiving treatment would have survived without it
- No pre-existing diagnostic parameter allows patient-tailored therapy

**Purpose**

- Use microarray analysis to identify “signature” of “poor prognosis” tumors
- Perhaps this diagnostic signature can be used to tailor therapy
- “Signature” would be group of genes whose expression patterns predict disease outcome—proof of prediction would be to use pattern on new set of tumors.

**Methods—tumors**

- 98 primary breast cancers selected [>5 cm (T1 or T2), no axillary metastases (N0), age <55 yrs at Dx]
  - 34 pts with distant metastases within 5 yrs
  - 44 pts continue disease-free for 5 yrs
  - 18 pts with BRCA1 germline mutations
  - 2 pts BRCA1 carriers

**Methods—Sample Prep**

- 5 µg total cellular RNA prepared from each tumor
- Each reverse transcribed with a T7-containing oligo d(T) primer
- The cDNA was transcribed into cRNA using T7 RNA polymerase
- Portion of each cRNA prep was pooled to make a common reference sample
- Cy3- or Cy5-coupled nucleotides added to label cRNA.

**Patient data available in supplemental materials**

<table>
<thead>
<tr>
<th>Sample#</th>
<th>Age</th>
<th>Diameter (mm)</th>
<th>Followup time (yr)</th>
<th>Metastases</th>
<th>Grade</th>
<th>Angioinvasion</th>
<th>ERp</th>
<th>PRp</th>
<th>Lymphocytic Infiltrate</th>
<th>BRCA1 Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
<td>25</td>
<td>12.53</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>80</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>34</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>40</td>
<td>105</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>45</td>
<td>10.66</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>20</td>
<td>11.98</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>100</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>13</td>
<td>11.16</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>80</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>20</td>
<td>10.14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>80</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>84</td>
<td>8</td>
<td>2</td>
<td>88.5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>48</td>
<td>15</td>
<td>10.29</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>60</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>38</td>
<td>15</td>
<td>6.64</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>100</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>22</td>
<td>5.8</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>90</td>
<td>70</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>46</td>
<td>22</td>
<td>7.86</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>41</td>
<td>21</td>
<td>8.17</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>48</td>
<td>10</td>
<td>8.23</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Methods—Microarrays

- Hu25K microarrays represented the 24,479 biological oligonucleotides plus 1,281 control probes.
- Each array was hybridized with one tumor cRNA and one sample of reference pool.
- Each tumor sample was labeled with Cy3 for one array; Cy5 for a second array (reference pool labeled with Cy5 and Cy3, respectively).

Methods—Data

- Each spot (feature) quantified, corrected for background, each array normalized.
- All data downloadable from web site: http://www.rii.com/publications
- Each number represents ratio from specific tumor to pooled reference sample.
- Selected 5,000 genes based on:
  - At least 2-fold different from reference pool
  - P-value less than 0.01 in more than 5 tumors

Methods—Clustering

- Unsupervised, hierarchical clustering applied to selected 5,000 genes over 98 tumors.
- 2-dimensional (clustered genes [columns], then clustered tumors [rows])
Unsupervised, two-dimensional cluster analysis of 98 tumors

Results

• Two distinct groups of tumors emerge
  – In upper group only 34% of pts develop metastases within 5 yrs
  – Lower group 70% metastatic
• Using about 5,000 genes, can produce improved predictor of metastasis.
• Continue to refine analysis by associating gene data with other, known parameters.

Gene clustering partitions BRCA1 status, ER expression, lymphocyte infiltration

Next steps

• 78 pts selected with sporadic (non-BRCA1) tumors
• 5,000 genes selected (significantly regulated in more than 3 tumors out of 78)
• Applied supervised clustering (compared to known metastasis parameter)

Supervised Clustering Method

• selection of discriminating candidate genes by their correlation with the category [metastasis]; (produced 231 genes)
• determination of the optimal set of reporter genes using a leave-one-out cross validation procedure;
• prognostic or diagnostic prediction based on the gene expression of the optimal set of reporter genes

Distribution of Correlation Coefficients (red)

(Compared with random distribution calculated by Monte Carlo permutations that randomize associations with tumor data; blue)
Selection

• Based on distribution of correlation coefficients, and the distributions predicted for randomized correlations, select “greater than 0.3 or less than -0.3” as the discriminator.
• Select 231 genes from real data
• Perform 10,000 Monte Carlo simulations, show distributions of the number of selected data from randomized set

Distribution of numbers of genes selected from randomized set

On average, select 36 genes. This demonstrates a 0.3% probability of selecting 231 genes.

“Leave-one-out” calculations

• took one sample out and used the remaining 77 samples to define a classifier based on the set of 231 discriminating genes
• predicted the outcome of the one sample we left out in the first place
• correlation coefficient is calculated using the selected reporter genes
• repeated this procedure until each of the 78 samples was left out once
• counted in how many cases the predictions were correct and in how many cases the predictions were incorrect

Select best 78 genes

• The combined error rate reaches the minimum when we use 70 marker genes from the top of our candidate list.
• Extended by an even more confusing method to include 8 additional genes.
Supervised Classification

- The 70 genes selected were ordered by correlation with good prognosis group.
- The 78 tumors were sorted by correlation with the average good prognosis group.
- The classifier correctly predicts outcome of disease for 83% of pts (threshold, solid line).

Optimize

- Revised threshold to include no more than 10% of poor prognosis group.
- This threshold resulted in 15 misclassifications (3 out of 34 from the poor group; 12 of 44 of the good group).
- Dashed line...

What genes are in this list?

- Associated with poor prognosis:
  - Cell cycle
  - Invasion
  - Metastasis
  - Angiogenesis
  - Signal transduction

- Genes identified in other studies not in list:
  - Cyclin D1
  - ER-α
  - UPA
  - PAI-1
  - HER2/neu
  - c-myc
Test classifiers

- Obtained 19 new tumor samples from young, lymph node negative tumors.
  - 7 pts metastases-free >5 yrs
  - 12 pts developed metastases <5 yrs
- The 70 gene list predicted outcome in 17/19 pts.

Prediction results

So…what have we got so far?

- Using correlation with metastatic outcome, we’ve whittled down the gene list to 70 genes that provide a good prediction of disease progression.
- This set is a classifier (or "signature").
- We tested classifier on fresh set of tumors and saw 17/19 accuracy.
- Not bad!

Next…

- They did a lot more stuff that’s *almost* understandable.
- “The prediction of the classifier presented in Fig. 2b would indicate that women under 55 years of age who are diagnosed with lymph-node-negative breast cancer that has a poor prognosis signature have a 28-fold odds ratio to develop distant metastasis within 5 years compared with those that have the good prognosis signature.”
- Let’s skip ahead to Fig. 3.

Clustering with known markers

- Unsupervised clustering distinguishes ER-positive from –negative tumors.
38 ER-negative tumors classified by BRCA1 clustering

- 100 optimal BRCA1 reporter genes shown
- Set enriched in lymphocyte-specific genes

Classifiers reduce selection of therapy for "disease-free" group

- Classifier selects similar percentage of poor prognosis group.
- Reduces selection of patients for unnecessary therapy.
- Classifier associated with ER status predicts pts benefiting from adjuvant hormonal therapy.
- Classifier associated with BRCA1 status may improve diagnosis of hereditary breast cancer.
- Genes in poor-prognosis classifier may represent new therapeutic targets.

Value of classifier set

<table>
<thead>
<tr>
<th>Patient group</th>
<th>Total patient group</th>
<th>Metastatic disease at 5 yr (n = 34)</th>
<th>Disease free at 5 yr (n = 44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td></td>
<td>33/34 (97%)</td>
<td>33/44 (75%)</td>
</tr>
<tr>
<td>BI Galen</td>
<td>34/36 (95%)</td>
<td>33/34 (97%)</td>
<td>33/44 (75%)</td>
</tr>
<tr>
<td>NIH</td>
<td></td>
<td>32/34 (94%)</td>
<td>32/44 (73%)</td>
</tr>
<tr>
<td>Prognosis profile</td>
<td>33/36 (92%)</td>
<td>31/34 (91%)</td>
<td>33/44 (75%)</td>
</tr>
</tbody>
</table>

The conventional consensus attends at tumour size <2cm, ER negative, grade 3-5, patient <35 yr.
(1) Either one of the criteria: BI Galen consensus: tumour > 1 cm; NIH consensus.
*Number of tumours having a poor prognosis signature using our microarray profile, defined by the optimized variable (threshold = 0.5)*.
1 Number of tumours having a poor prognosis signature in the group of disease-free patients, when the cross-validated classifier is applied.

Whoa! We made it.

- What have we learned?
  - Gene expression profiling can be used to predict disease progression, selecting pts most appropriate for aggressive therapy.
  - Accurate classifier sets require advanced statistical techniques that may be difficult for biologists (physicians?) to understand.
  - We're going to need to work in teams to solve these advanced problems.
  - Physicians will need to understand these tools!