Reflection on Venet et al., PLoS Computational Biology, 2011

Limsoon Wong
Almost all random signatures have p-value < 0.05

And the larger a random signature is, the more likely this happens.

Venet et al., *PLOS Comput Biol*, 2011
A bewilderment

Breast cancer survival signatures are no better than random signatures

Really? Then all cancer survival signature papers should be rejected w/o review!

Is there some way that a true cancer survival signature is better than random signatures?
Perhaps instead of asking whether a signature is significant,

Ask what makes a signature significant.
Proliferation is a hallmark of cancer

Hypothesis a la Venet et al.: Proliferation-associated genes make a signature significant

<table>
<thead>
<tr>
<th>Cutoffs</th>
<th>Counts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NP</td>
<td>P</td>
<td>Marginals</td>
</tr>
<tr>
<td>Above 0.05</td>
<td>7043</td>
<td>19 043</td>
<td>26 086</td>
</tr>
<tr>
<td>Below 0.05</td>
<td>2766</td>
<td>19 148</td>
<td>21 914</td>
</tr>
<tr>
<td>Marginals</td>
<td>9809</td>
<td>38 191</td>
<td>48 000</td>
</tr>
</tbody>
</table>

# of random signatures w/ ≥1 prolif gene
Impact of proliferation genes on reported signatures

P-value of reported signatures, before removing proliferation genes

P-value of reported signatures, after removing proliferation genes
Many random signatures with proliferation genes are not significant

Which proliferation genes make many random signatures significant?
Leverage background knowledge

Proliferation is a cancer hallmark

Good signatures with high diff in p-values before vs after removing proliferation genes:

GLINSKY, DAI, RHODES, ABBA, WHITFIELD

SPS = \{ \text{genes appearing in at least two of these good signatures} \}:

83 genes in total

81 of these are proliferation associated
Systematic evaluation

SPS genes show additive effect

Other proliferation genes don’t
Test on many datasets

For any independent dataset, a random signature has ~50% chance to be significant in it.

How many independent datasets are needed to avoid reporting random signatures as significant?

<table>
<thead>
<tr>
<th>n</th>
<th>(50%)^n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.00%</td>
</tr>
<tr>
<td>2</td>
<td>25.00%</td>
</tr>
<tr>
<td>3</td>
<td>12.50%</td>
</tr>
<tr>
<td>4</td>
<td>6.25%</td>
</tr>
<tr>
<td>5</td>
<td>3.13%</td>
</tr>
<tr>
<td>6</td>
<td>1.60%</td>
</tr>
<tr>
<td>7</td>
<td>0.78%</td>
</tr>
</tbody>
</table>
Test on many datasets

SPS & other true signatures are significant on all 7 breast cancer datasets

Random signatures are hardly universal, despite better p-values than known signatures on some datasets
A theory-practice gap

~50% of random signatures are significant in 1 dataset

Red: Expected # of random signatures significant in n independent dataset

Blue: Observed distribution
Do people take independent validation seriously?
How bad the situation is

Check using Nature Publishing Group journals, https://www.nature.com/search/advanced

# of papers doing x-validation
cross validation classifier, 2019/20 ➔ 1296 results
classifier, 2019/20 ➔ 2566 results

1296 / 2566 = ~50% of papers that pertain to classifier construction published in NPG journals in 2019 to 2020 uses x-validation as part of evaluation methodology
How bad the situation is

I inspected the first 40 papers among the 1296 results in the “cross-validation, 2019 to 2020” set

20 / 40 = 50% uses cross-validation, and no clear independent datasets

⇒ 50% * 50% = 25% of NPG papers in 2019 to 2020 pertaining to classifier construction may have dodgy results?!
Closing remarks

Bewilderment: Breast cancer survival signatures are no better than random signatures

Enlightenment: SPS genes

Cautionary note 1: Need to validate on many independent data sets

Cautionary note 2: Some independent data sets are not as independent as you think

Goh & Wong. Why breast cancer signatures are no better than random signatures explained. *Drug Discovery Today*, 23(11):1818-1823, 2018