Happy families are all alike;  
Every unhappy family is unhappy in its own way

Wong Limsoon  
This talk is based on joint work with Wilson Goh
The Anna Karenina Principle

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Translation

• There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest
GETTING THE NULL HYPOTHESIS RIGHT
A seemingly obvious conclusion

- **SNP rs123 is a great biomarker for a disease, based on a prospective study**
  - If rs123 is AA or GG, unlikely to get the disease
  - If rs123 is AG, ~3x higher risk of disease

- **A straightforward $\chi^2$ test. Anything wrong?**

<table>
<thead>
<tr>
<th>SNP</th>
<th>Genotypes</th>
<th>Controls n(%)</th>
<th>Cases n(%)</th>
<th>$\chi^2$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs123</td>
<td>AA</td>
<td>1 0.9%</td>
<td>0 0.0%</td>
<td></td>
<td>4.78E-21$^b$</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>38 35.2%</td>
<td>79 97.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>69 63.9%</td>
<td>2 2.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: SNP, single nucleotide polymorphism.
Careless null hypothesis

- **“Effective” H0**
  - rs123 alleles are identically distributed in the two samples

- **Assumption**
  - Distributions of rs123 alleles in the two samples are identical to the two populations

- **Apparent H0**
  - rs123 alleles are identically distributed in the two populations

- **Apparent H1**
  - rs123 alleles are differently distributed in the two populations
There may be sample bias

- $AG = 38 + 79 = 117$, controls + cases $= 189 \Rightarrow$ population is $\sim 62\%$ AG $\Rightarrow$ population is $>9\%$ AA, unless AA is lethal

- "Big data check" shows AA is non-lethal for this SNP $\Rightarrow$ sample is biased
Time for Exercise #1

- Suppose distributions of rs123 alleles in the two samples are identical to the corresponding populations and the test is significant

- Can we say rs123 mutation causes the disease?

- Hint: Human genetic recombinations take place in large chunks
A seemingly obvious conclusion

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>died</td>
<td>100</td>
<td>165</td>
</tr>
</tbody>
</table>

Treatment A is better

What is happening here?

Women

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>died</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

Men

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>lived</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>died</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>

Treatment B is better
Careless null hypothesis

• **“Effective” H0**
  - Treatment effects are identically distributed in the two samples

• **Assumption**
  - All other factors are equalized in the two samples

• **Apparent H0**
  - Treatment effects are identically distributed in the two populations

• **Apparent H1**
  - Treatment effects are differently distributed in the two populations
A/B sample not equalized in other attributes, e.g. sex

- **Taking A**
  - Men = 100 (63%)
  - Women = 60 (37%)

- **Taking B**
  - Men = 210 (91%)
  - Women = 20 (9%)

<table>
<thead>
<tr>
<th>Women</th>
<th>Overall</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>lived</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>died</td>
<td>100</td>
<td>165</td>
</tr>
</tbody>
</table>
Time for Exercise #2

• Suppose you have tested that a hypothesis is significant in your dataset. What can you do next to increase the depth of your analysis?
In statistical hypothesis testing, the null distribution is the probability distribution of the test statistic when the null hypothesis is true. For example, in an F-test, the null distribution is an F-distribution.

GETTING THE NULL DISTRIBUTION RIGHT
Synthetic lethality

Why interested in synthetic lethality?

Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets

Fig. 7 Two models for pathway-based targeting of synthetic lethal genes $B$ in conjunction with deleted/downregulated genes $A$: a parallel pathways model where targeting $B$ results in disruption of both survival pathways, and b negative feedback-loop model where targeting $B$ shunts of (forward) signals for cell survival.
Synthetic lethal pairs

• **Fact**
  – When a pair of genes is synthetic lethal, mutations of these two genes avoid each other

• **Observation**
  – Mutations in genes (A,B) are seldom observed in the same subjects

• **Conclusion by abduction**
  – Genes (A,B) are synthetic lethal

A seemingly obvious approach based on the hypergeometric test

Mutations of genes (A,B) avoid each other if $P[X \leq |S_{AB}|] \leq 0.05$

Anything wrong with this?
Seems to work fine
Among top ME-genes, GARP score ranks correlate with mutual exclusion ranks

But GARP scores of ME-genes (i.e. have mutually exclusive mutations to BRCA1) are similar to other genes

Really?
A cautionary note

The hypergeometric distribution does not reflect real-world mutations

- **Real-life mutations**
  - Inherited in blocks; those close to each other are correlated
  - Some subjects have more mutations than others, e.g. those with defective DNA-repair genes

⇒ Null distribution is not hypergeometric, binomial, etc.

\[ P[X \leq |S_{AB}|] = 1 - P[X > |S_{AB}|], \]

where \( P[X > |S_{AB}|] \) is computed using the hypergeometric probability mass function for \( X = k > |S_{AB}| \):

\[
P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{S_B} \binom{|S_A|}{k} \binom{|S| - |S_A|}{|S_B| - k} \binom{|S|}{|S_B|}
\]
Real-life example:
Mutations of TP53 and its neighbours

(a) Genomic location of genes close to TP53
(b) CNA profile of genes close to TP53
Time for Exercise #3

- FXR2 is located near TP53
- FXR1 and FXR2 are paralogs that buffer each other’s function
- Do FXR1 and TP53 deletions avoid each other?

![Genetic Alteration Chart]

- Is FXR1 synthetic lethal to TP53?
- Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?
Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down

Fan et al., eLife, 6:e26129, 2017
Gene-selection methods have poor reproducibility

- Low % of overlapping genes from diff microarray expt
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

Contextualizing based on pathways may help

- Each disease phenotype has some underlying cause

- There is some unifying biological theme for genes that are truly associated with a disease subtype

- Uncertainty in selected genes can be reduced by considering biological processes of the genes

- The unifying biological theme is basis for inferring the underlying cause of disease subtype
ORA-Paired

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be a patient
- Let $q_k$ be a normal

- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) - \text{Expr}(g_i,q_k)$

- $H_0$: Pathway $P$ is irrelevant to the diff betw patients and normals, so genes in $P$ behave similarly in patients and normals

$\Rightarrow t$-test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Time for Exercise #4

Which null distribution is appropriate? Why?

ORA-Paired

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be a patient
- Let $q_k$ be a normal

- $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$

$\Rightarrow t$-test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- $t$-distribution with $n \times m$ degrees of freedom
- $t$-distribution with $n + m$ degrees of freedom
- Generate null distribution by gene-label permutation
- Generate null distribution by class-label permutation
Testing the null hypothesis

“Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals”

• By the null hypothesis, a dataset and any of its class-label permutations are exchangeable

⇒ Get null distribution by class-label permutations
  – What happens when sample size is small?

A related cautionary note

<table>
<thead>
<tr>
<th>NN</th>
<th>NN Acc. (%)</th>
<th>Acc. $t_1$-sparse (%)</th>
<th>Acc. $t_2$-sparse (%)</th>
<th>NPAQ r for $t_1$-sparse (%)</th>
<th>NPAQ r for $t_2$-sparse (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCH1</td>
<td>74.00</td>
<td>78.00</td>
<td>81.00</td>
<td>20.31</td>
<td>62.50</td>
</tr>
<tr>
<td>ARCH2</td>
<td>62.00</td>
<td>73.00</td>
<td>78.00</td>
<td>12.50</td>
<td>65.62</td>
</tr>
<tr>
<td>ARCH3</td>
<td>76.00</td>
<td>82.00</td>
<td>83.00</td>
<td>45.31</td>
<td>52.34</td>
</tr>
<tr>
<td>ARCH4</td>
<td>50.00</td>
<td>64.00</td>
<td>72.00</td>
<td>17.19</td>
<td>93.75</td>
</tr>
<tr>
<td>ARCH5</td>
<td>78.00</td>
<td>82.00</td>
<td>83.00</td>
<td>74.22</td>
<td>24.22</td>
</tr>
<tr>
<td>ARCH6</td>
<td>80.00</td>
<td>11.00</td>
<td>87.00</td>
<td>37.50</td>
<td>55.47</td>
</tr>
<tr>
<td>ARCH7</td>
<td>87.00</td>
<td>89.00</td>
<td>89.00</td>
<td>6.25</td>
<td>79.69</td>
</tr>
</tbody>
</table>

Table 2: First and second column refer to the baseline model where we use BNNs with 7 different architectures. The third and fourth represent the accuracies of sparsified models with $t_1 = 0.03, t_2 = 0.05$ sparsification thresholds. The last 2 columns show NPAQ estimates for the difference between each sparsified model and the original model.

Credit: Teodora Baluta
GETTING THE TEST STATISTIC RIGHT
A seemingly obvious conclusion

- A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival
  - Cox’s survival model p-value << 0.05

- A straightforward Cox’s analysis. Anything wrong?
Almost all random signatures also have p-value < 0.05

- What happened?
- Maybe the significant random signatures share some genes with observed signature?
Almost all random signatures sharing no genes with observed signatures also have p-value < 0.05

- What happened?

Goh & Wong, *Drug Discovery Today*, 2018
What is the right null hypothesis?

- A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival
  - Cox’s survival model p-value << 0.05
- A straightforward Cox’s analysis. Anything wrong?

\[ H_0 = \text{the black/red survival curves induced by the observed signature are not different} \]

Almost all random signatures also have p-value < 0.05

- What happened?
- Maybe the significant random signatures share some genes with observed signature?

\[ H_0 = \text{survival curves induced by the observed signature are not different from those induced by random signatures?} \]
What is the right null distribution?

- Generate null samples by permutating sample labels (viz. survival time)
- Null samples are random signatures?

A seemingly obvious conclusion

- A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival
  - Cox’s survival model p-value << 0.05
- A straightforward Cox’s analysis. Anything wrong?

Almost all random signatures also have p-value < 0.05

- What happened?
- Maybe the significant random signatures share some genes with observed signature?
What is the right test statistic?

- Cox’s hazard ratio (HR)
- Cox’s p-value?
- Median $\Delta$HR betw the observed signature and random signatures?
“Excellent health statistics - smokers are less likely to die of age related illness”

SOMETIMES CHANGING PERSPECTIVE HELPS
Almost all random signatures also have p-value < 0.05

- Instead of asking whether a signature is significant, ask what makes a signature (random or otherwise) significant

Venet et al., *PLOS Comput Biol*, 2011
Proliferation is a hallmark of cancer

Hypothesis: 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
Time for Exercise #5

• In the 1st place, how do I know (which) proliferation genes make many random signatures significant?

• Some helpful analytical practices
  – Leverage existing data and knowledge
  – Careful and systematic evaluation of gene sets
  – Rigorous testing against as many published datasets as possible
Leverage background knowledge

- Proliferation is a cancer hallmark

- Good signatures with high diff in p-values or effect size before vs after removing proliferation genes
  - GLINSKY, DAI, RHODES, ABBA, WHITFIELD

- SPS = { 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

- SPS is universally significant on 7 breast cancer datasets
- Random signatures (same size as SPS) are hardly universal, even though they get better p-values than known signatures on some datasets
Time for Exercise #6

- SPS is universally significant on 7 breast cancer datasets
- Random signatures (same size as SPS) are hardly universal, even though they get better p-values than known signatures on some datasets
- Why consider 7 datasets?
SUMMARY
Anna Karenina Principle

• Careless null / alternative hypothesis due to forgotten assumptions
  – Distributions of the feature of interest in the two samples are identical to the two populations
  – Features not of interest are equalized / controlled for in the two samples
  – No other explanation for significance of the test
  – Null distribution models the real world

• These make it easy to reject the carelessly stated null hypothesis and accept an incorrect alternative hypothesis
Avoiding wrong conclusion, Getting deeper insight

• **Check for sampling bias**
  – Are the distributions of the feature of interest in the two samples same as that in the two populations?

• **Check for exceptions**
  – Are there large subpopulations for which the test outcome is opposite?
  – Are there large subpopulations for which the test outcome becomes much more significant?

• **Check for validity of the null distribution etc.**
  – Can you derive it from the null hypothesis?

• **Check on many datasets**