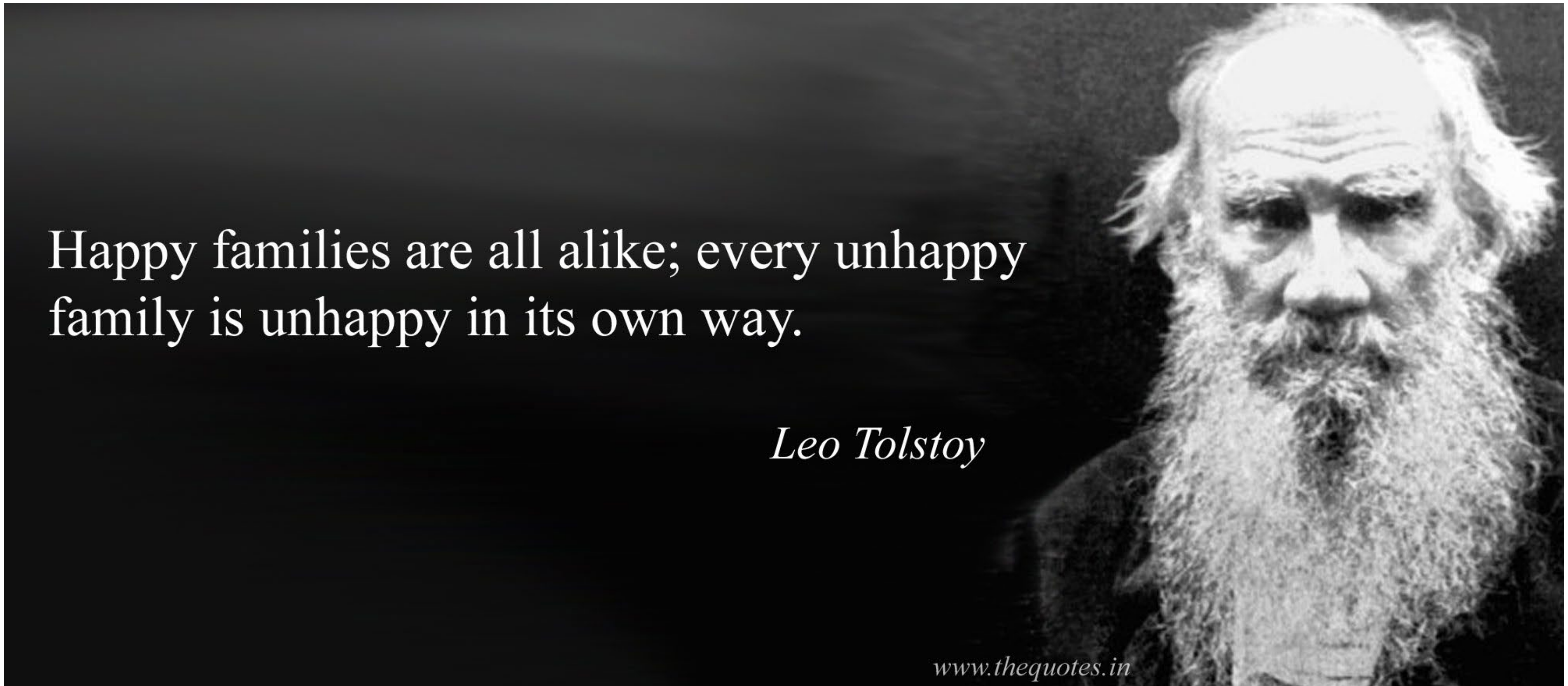
Dealing with confounders in omics data analysis

Limsoon Wong

This talk is based on joint work with Wilson Goh

National University of Singapore

The Anna Karenina Principle



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

NUS National University of Singapore

SNP	Genotypes	Controls [n(%)]		Cases [n(%)]		χ^2	*P* value
		Group					
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

A seemingly obvious conclusion

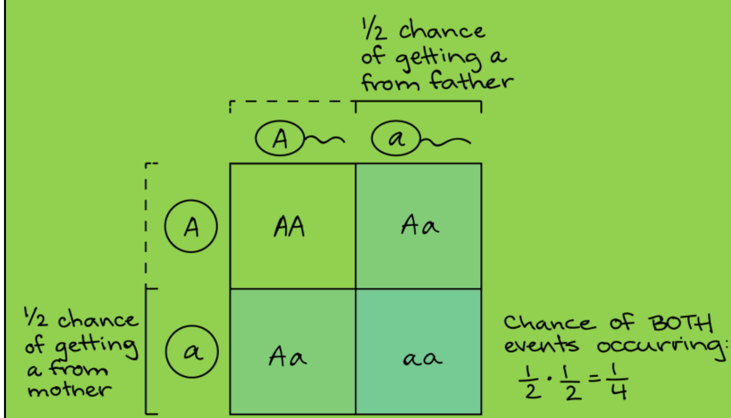
- **SNP rs123 is a great biomarker for a disease, based on a prospective study**
  - If rs123 is AA or GG, unlikely to get the disease
  - If rs123 is AG, ~3x higher risk of disease
- **A straightforward χ2 test. Anything wrong?**

Careless null hypothesis

- **"Effective" H0**
  - rs123 alleles are identically distributed <u>in the two samples</u>
- **Assumption**
  - Distributions of rs123 alleles in the two samples are identical to the two populations

- **Apparent H0**
  - rs123 alleles are identically distributed <u>in the two populations</u>
- **Apparent H1**
  - rs123 alleles are differently distributed <u>in the two populations</u>

There may be sample bias



Basic rule of human genetics

SNP	Genotypes	Group				χ^2	P value
		Controls [n(%)]		Cases [n(%)]			
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

- **AG = 38 + 79 = 117, controls + cases = 189 ⇒ population is ~62% AG ⇒ population is >9% AA, unless AA is lethal**
- **"Big data check" shows AA is non-lethal for this SNP ⇒ sample is biased**

Discussion

- **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**

NUS
National University of Singapore

Some NUS numbers

- **3 campuses**
  - Kent Ridge, Bukit Timah, & Outram
- **150 hectares**
- **13 undergrad schools**
- **4 graduate schools**

- **28k undergrads**
- **10k grad students**
- **2.4k faculty**
- **3.5k research staff**
- **5.4k other staff**

A seemingly obvious conclusion

Overall

	A	B
lived	60	65
died	100	165

Treatment A is better

What is happening here?

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

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. gender

Overall

	A	B
lived	60	65
died	100	165

Women

	A	B
lived	40	15
died	20	5

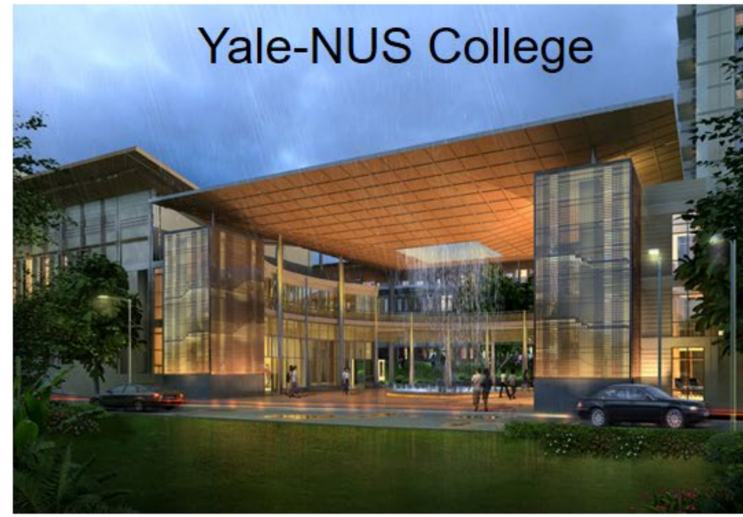
Men

	A	B
lived	20	50
died	80	160

- **Taking A**
  - Men = 100 (63%)
  - Women = 60 (37%)
- **Taking B**
  - Men = 210 (91%)
  - Women = 20 (9%)

NUS
National University of Singapore

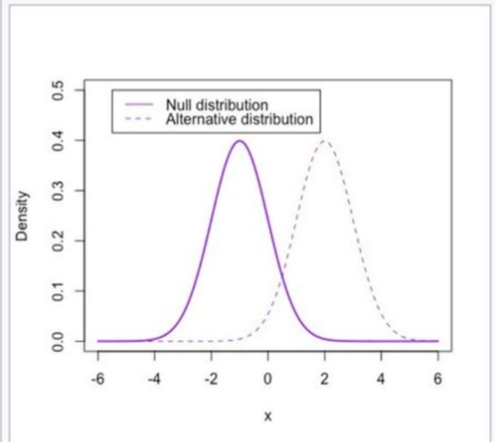
Faculty of Arts & Social Sciences


Yale-NUS College


Yong Siew Toh Conservatory of Music


Faculty of Law

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**.



Null and alternative distribution

GETTING THE NULL DISTRIBUTION RIGHT

Synthetic lethality

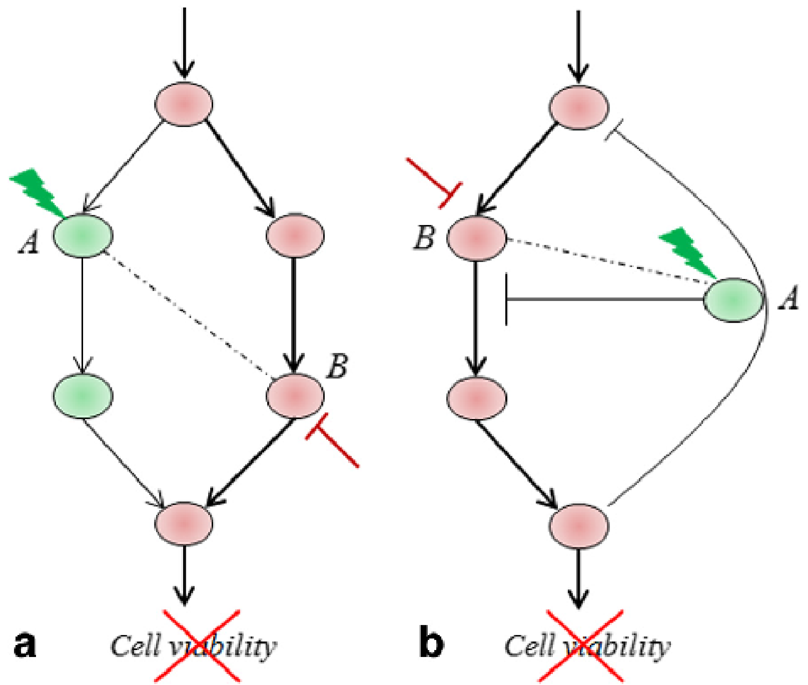


Fig. 7 Two models for pathway-based targeting of synthetic lethal genes *B* in conjunction with deleted/downregulated genes *A*: **a** parallel pathways model where targeting *B* results in disruption of both survival pathways, and **b** negative feedback-loop model where targeting *B* shunts of (forward)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

Why interested in synthetic lethality?

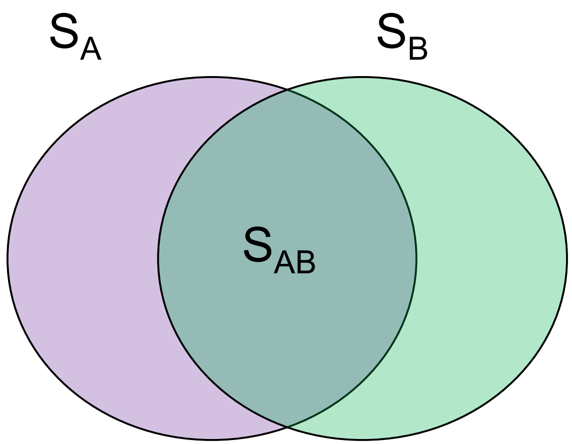
Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets

Srihari et al. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biology Direct*, 10:57, 2015.

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



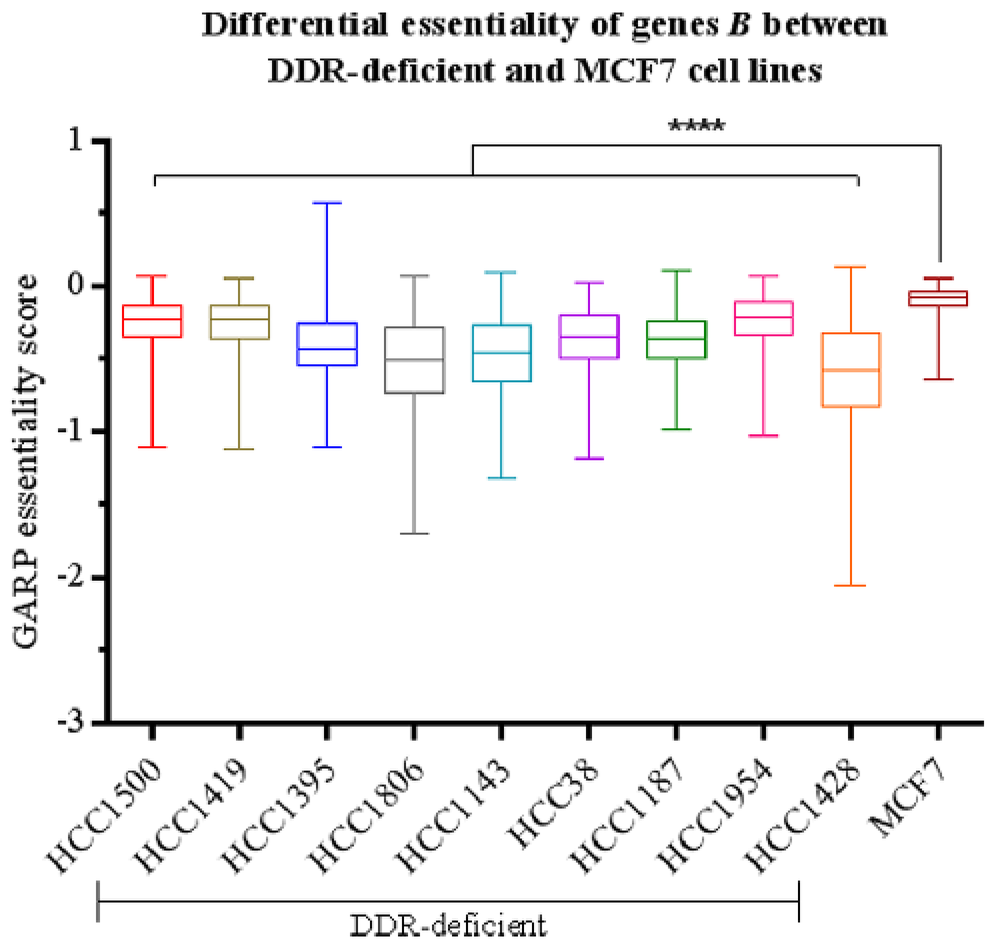
$$P[X \leq |S_{AB}|] = 1 - P[X > |S_{AB}|], \quad (1)$$

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

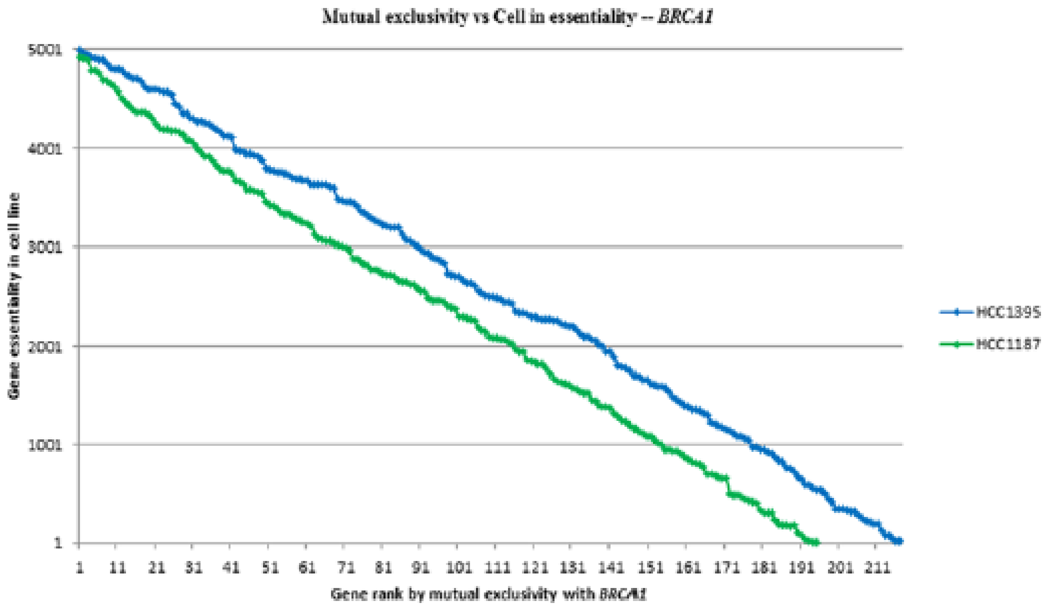
$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

- **Mutations of genes (A,B) avoid each other if $P[X \leq S_{AB}] \leq 0.05$**
- **Anything wrong with this?**

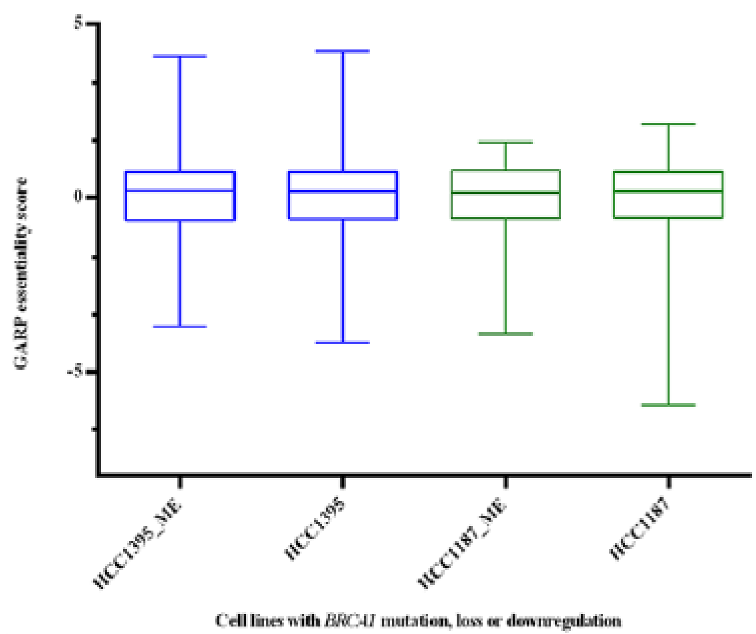
Seems to work fine



Really?



Ranges for GARP scores of predicted genes (ME) and entire set of profiled genes in *BRCA1*-deficient cell lines



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

The hypergeometric distribution does not reflect real-world mutations

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

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

$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

- **The Hypergeometric distribution assumes**
  - Mutations are independent
  - Mutations have equal chance to appear in a subject

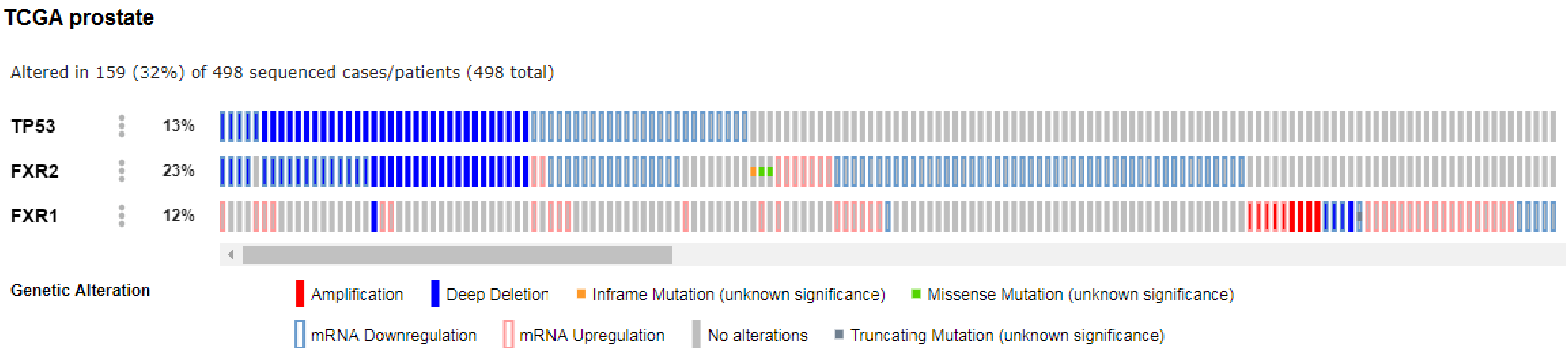
- **Real-life mutations**
  - Inherited in blocks; those close to each other are correlated
  - Some subjects have more mutations than others, e.g. those with defective DNA-repair genes

$\Rightarrow$ **Null distribution is not hypergeometric, binomial, etc.**

Discussion

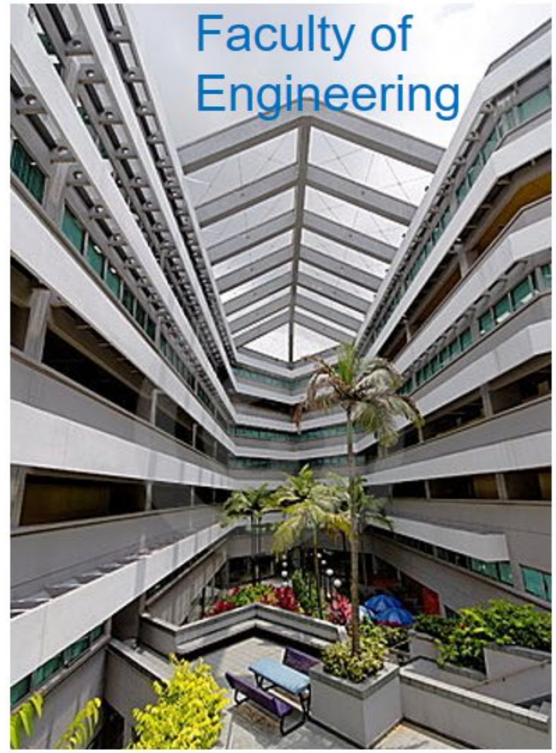
- **FXR2 is located near TP53**
- **FXR1 and FXR2 are paralogs that buffer each other's function**

- **Do FXR1 and TP53 deletions avoid each other?**



- **Is FXR1 synthetic lethal to TP53?**
- **Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?**

NUS
National University of Singapore



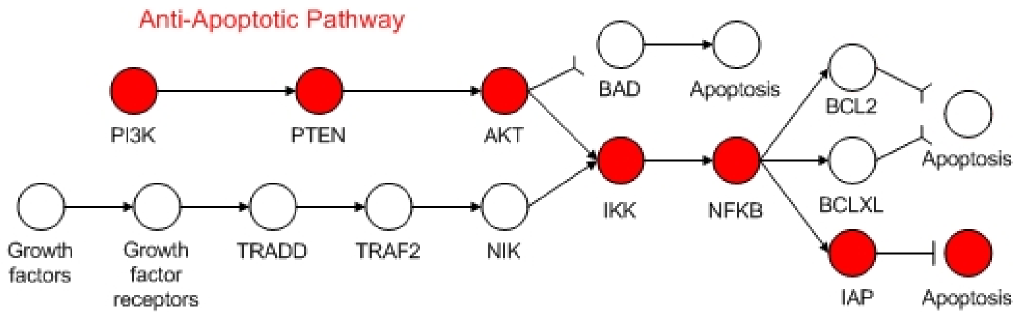
Gene-selection methods have poor reproducibility

- **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**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, *Bioinformatics*, 2009

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

- **H0: 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**

Lim et al., *JBCB*, 13(4):1550018, 2015.

Discussion

ORA-Paired

- Let g_i be genes in a given pathway P
- Let p_j be a patient
- Let q_k be a normal
- H0: Pathway P is irrelevant to the diff betw patients and normals, so genes in P behave similarly in patients and normals
- Let $\Delta_{i,j,k} = Expr(g_i,p_j) - Expr(g_i,q_k)$

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

Which null distribution is appropriate? Why?

- t-distribution with n*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

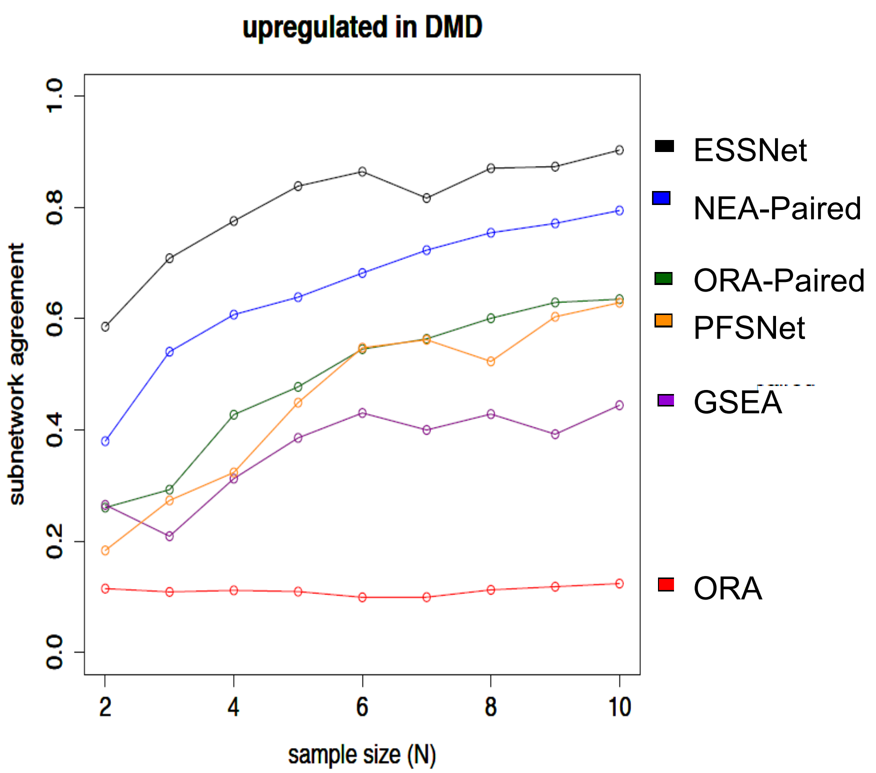
Testing the null hypothesis

"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?



Lim et al., *JBCB*, 13(4):1550018, 2015.

NUS
National University
of Singapore


Yong Loo Lin School of Medicine

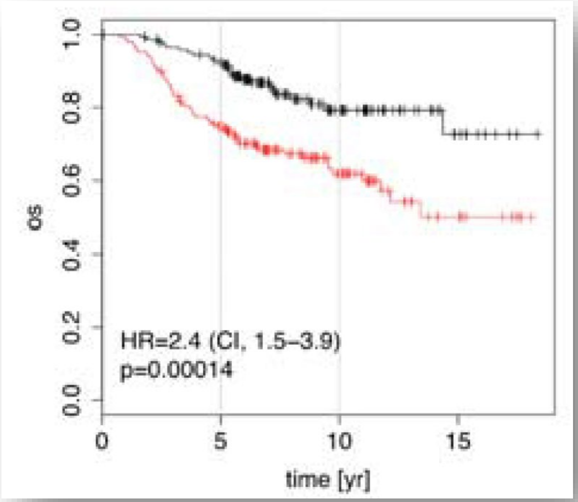

Duke-NUS Graduate Medical School


Alice Lee Centre for Nursing Studies

Faculty of Dentistry

GETTING THE TEST STATISTIC RIGHT

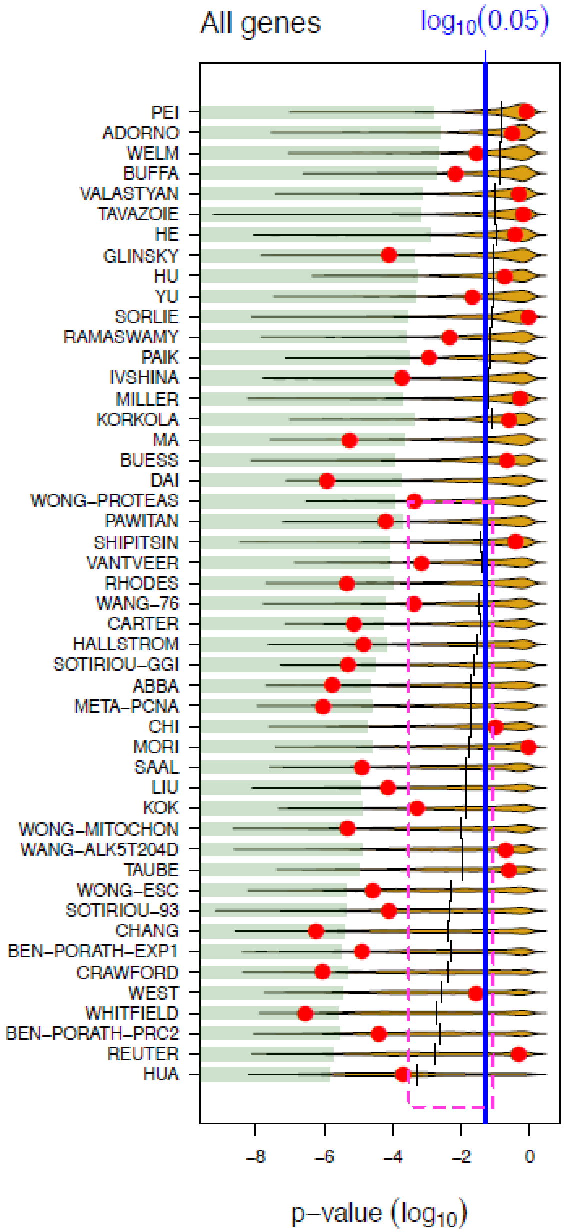
Venet et al., *PLOS Comput Biol*, 2011



A seemingly obvious conclusion

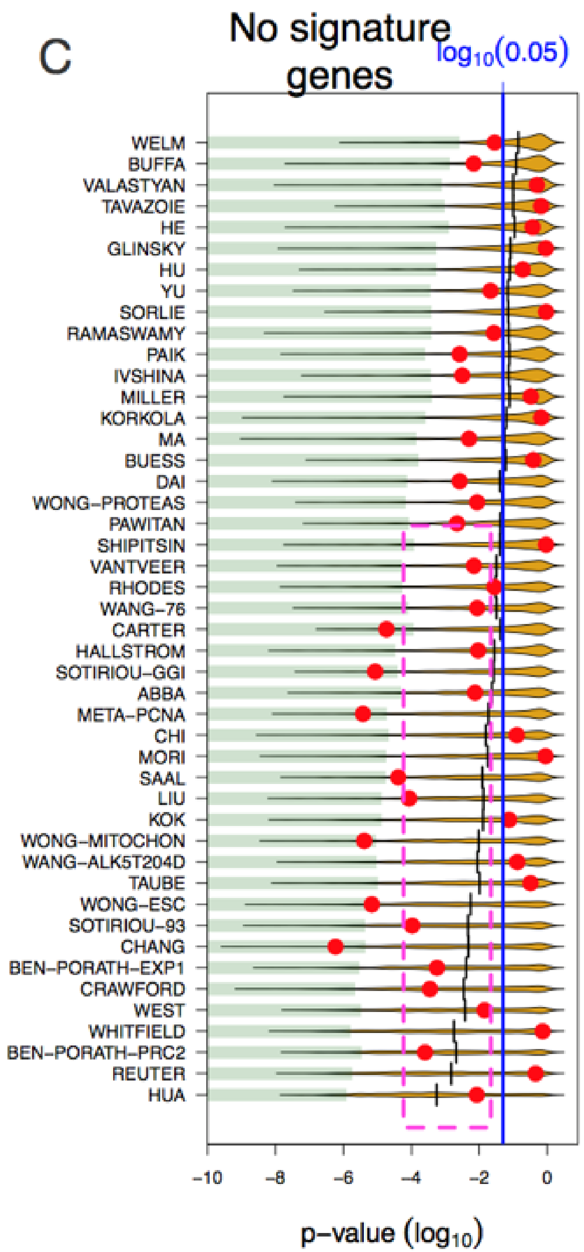
- **A multi-gene signature (social defeat in mice) is claimed as a good biomarker for breast cancer survival**
  - Cox's survival model p-value << 0.05
- **A straightforward Cox's analysis. Anything wrong?**

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



Almost all random signatures also have p-value < 0.05

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

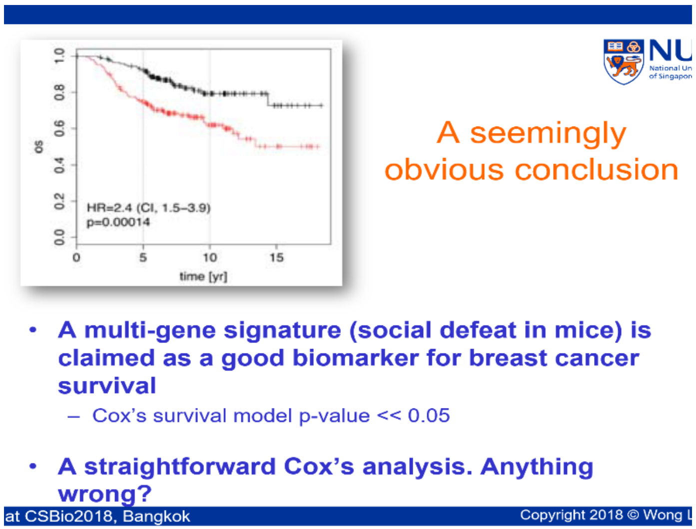


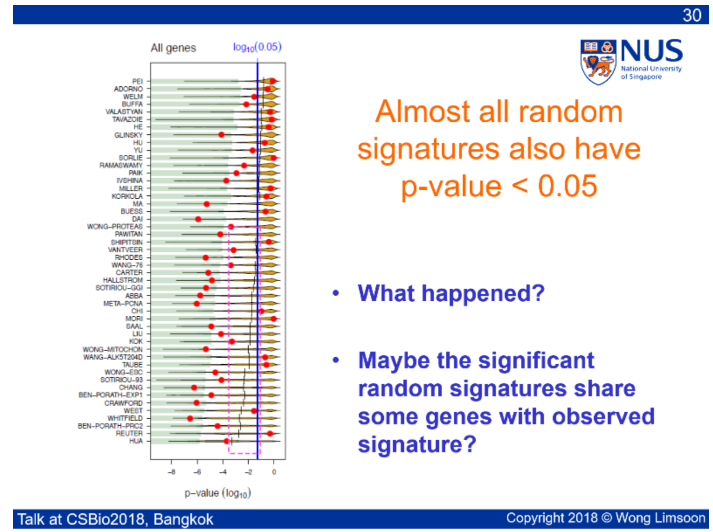
Almost all random signatures sharing no genes with observed signatures also have p-value < 0.05

- **What happened?**

Goh & Wong, *Drug Discovery Today, 2018*

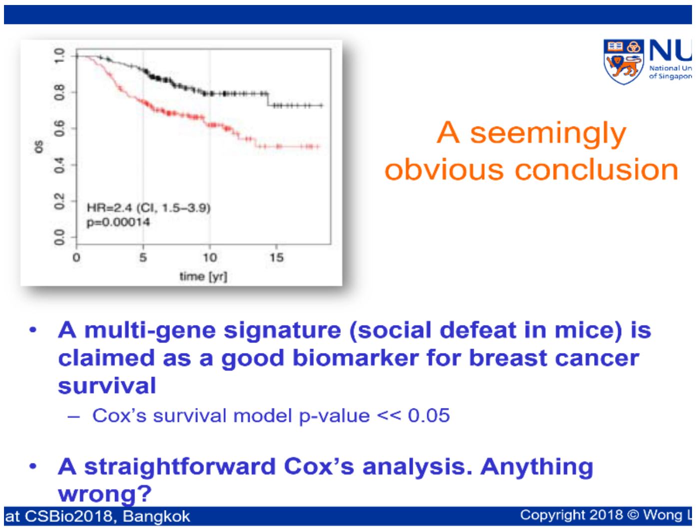
What is the right null hypothesis?

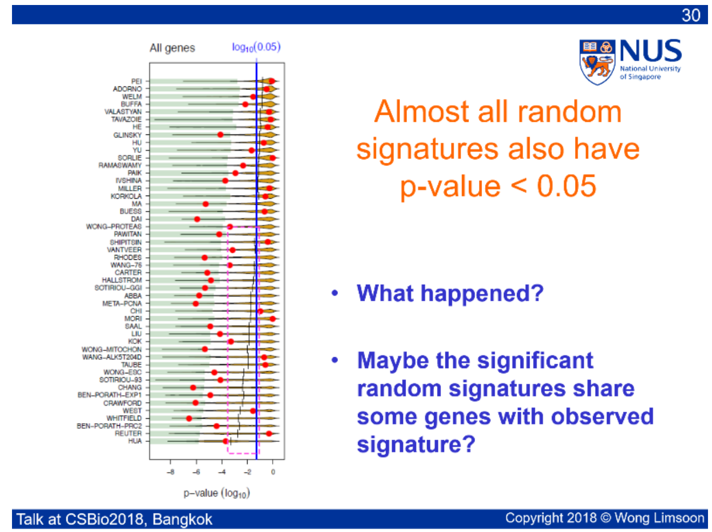




- **H0 = the black/red survival curves induced by the observed signature are not different**
- **H0 = survival curves induced by the observed signature are not different from those induced by random signatures?**

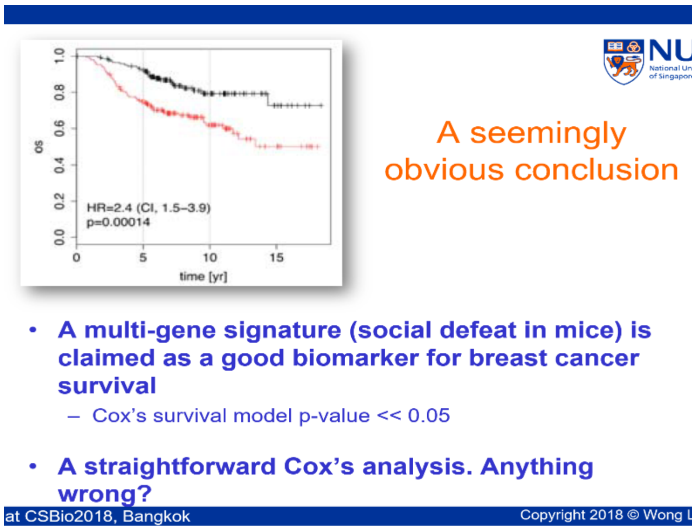
What is the right null distribution?

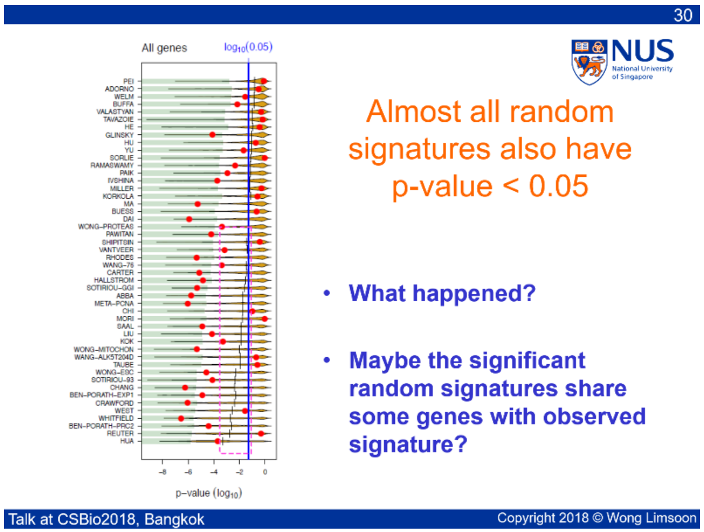




- **Generate null samples by permutating sample labels (viz. survival time)**
- **Null samples are random signatures?**

What is the right test statistic?





- **Cox's hazard ratio (HR)**

- **Cox's p-value?**
- **Median ΔHR betw the observed signature and random signatures?**

NUS
National University of Singapore

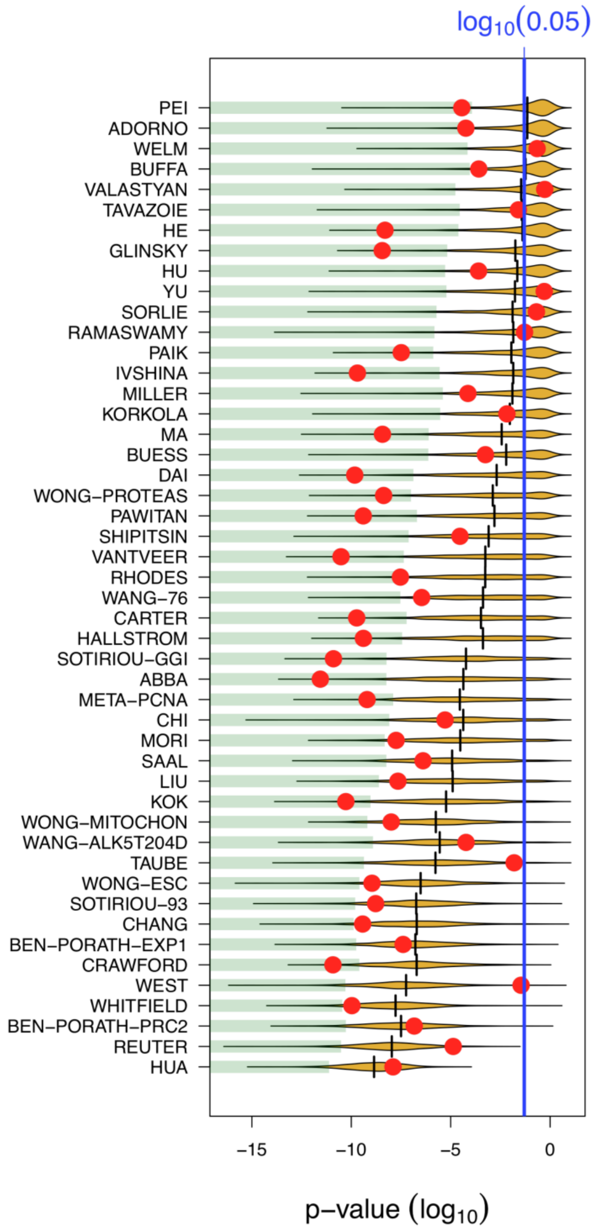
LKC
Natural History Museum

U-Town

Centre for the Arts

Prince George's Park

SOMETIMES CHANGING PERSPECTIVE HELPS



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

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**

Proliferation is a hallmark of cancer

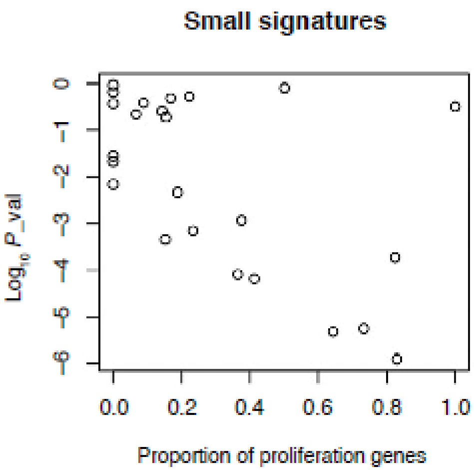
Hypothesis: Proliferation-associated genes make a signature significant

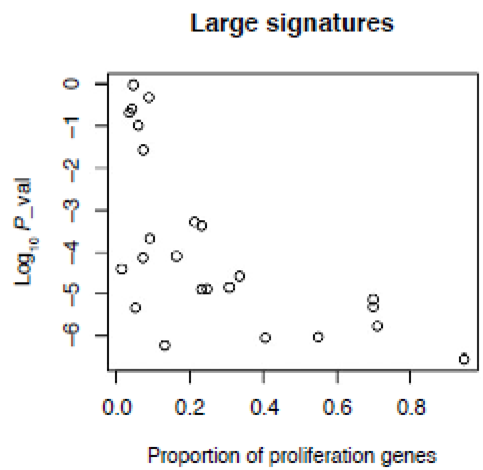
of random signatures w/ ≥1 prolif gene

Cutoffs	Counts		
	NP	P	Marginals
Above 0.05	7043	19 043	26 086
Below 0.05	2766	19 148	21 914
Marginals	9809	38 191	48 000

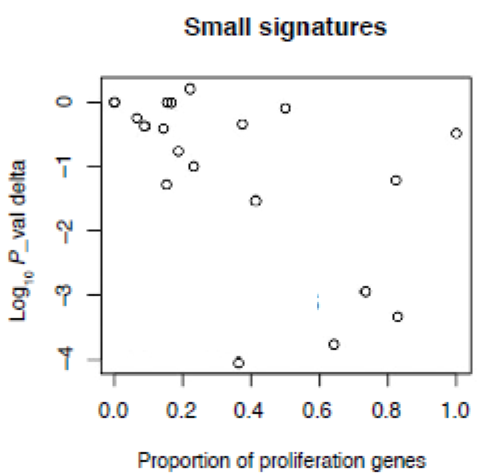
Impact of proliferation genes on reported signatures

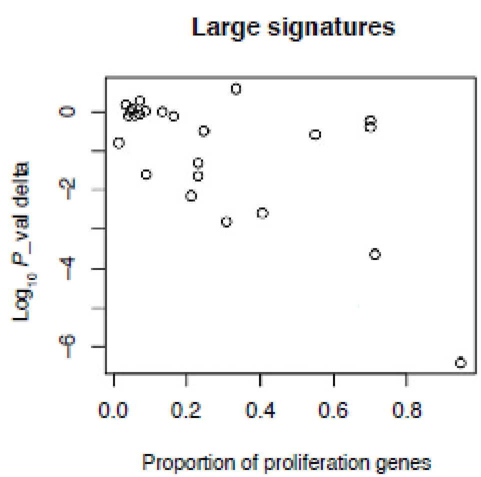
NUS National University of Singapore





P-value of reported signatures, before removing proliferation genes





P-value of reported signatures, after removing proliferation genes

Discussion

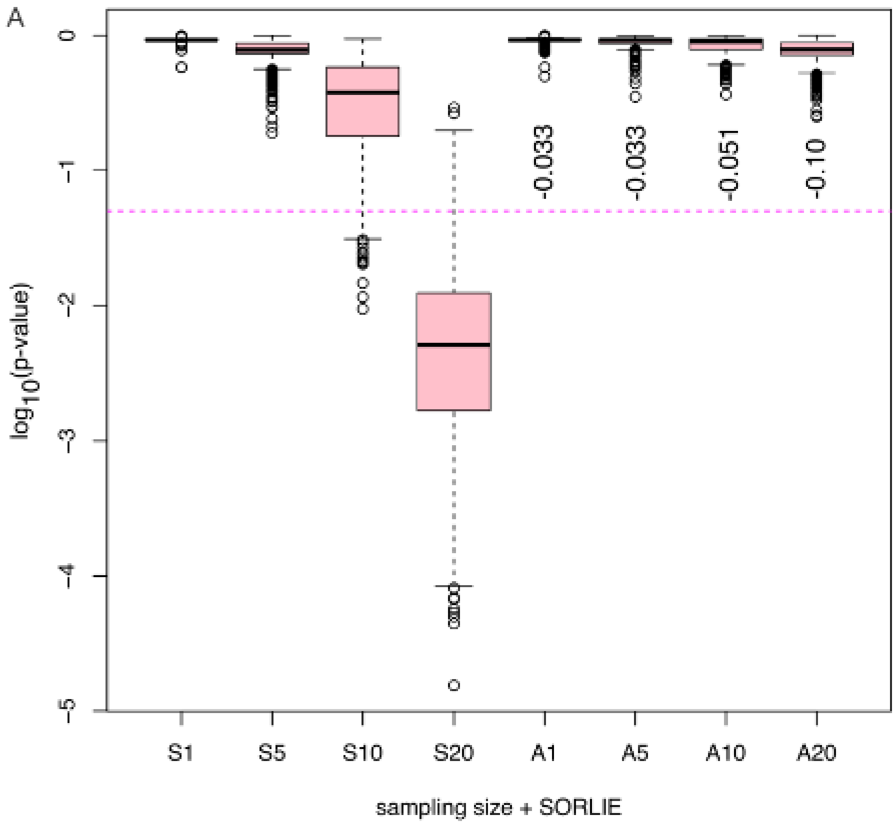
- **Many proliferation genes do not make random signatures significant. How do I know which proliferation genes make many random signatures significant?**

- **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

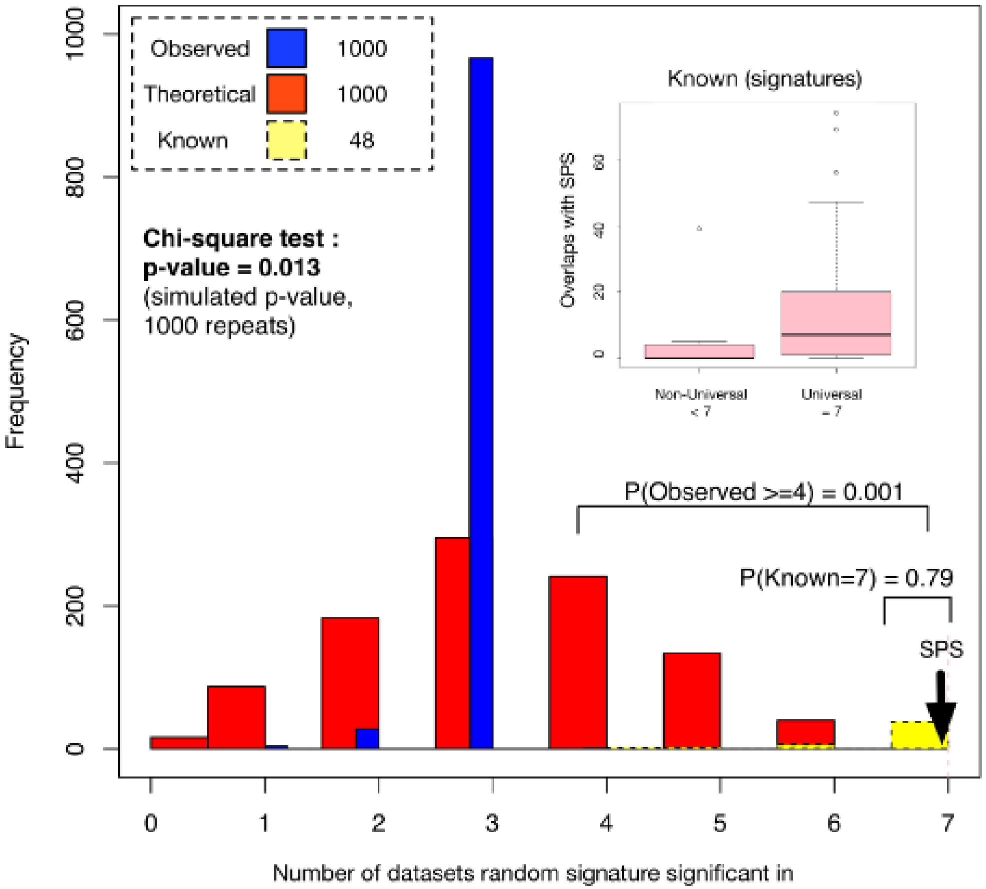
- **Proliferation is a cancer hallmark**
- **Good signatures with high diff in p-values before vs after removing proliferation genes**
  - GLINSKY, DAI, RHODES, ABBA, WHITFIELD
- **SPS = { genes appearing in at least two of these good signatures }**
  - 83 genes in total
  - 81 of these are proliferation associated

Systematic evaluation



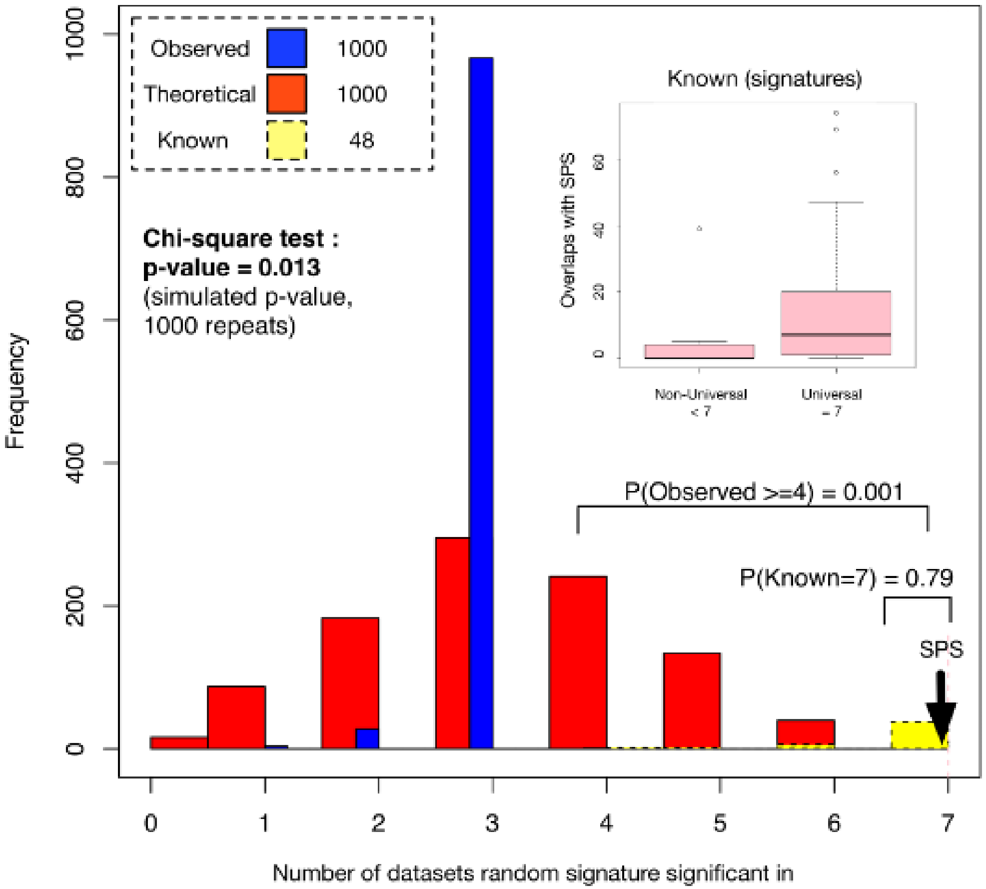
- **SPS genes show additive effect, other proliferation genes don’t**

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**

Discussion

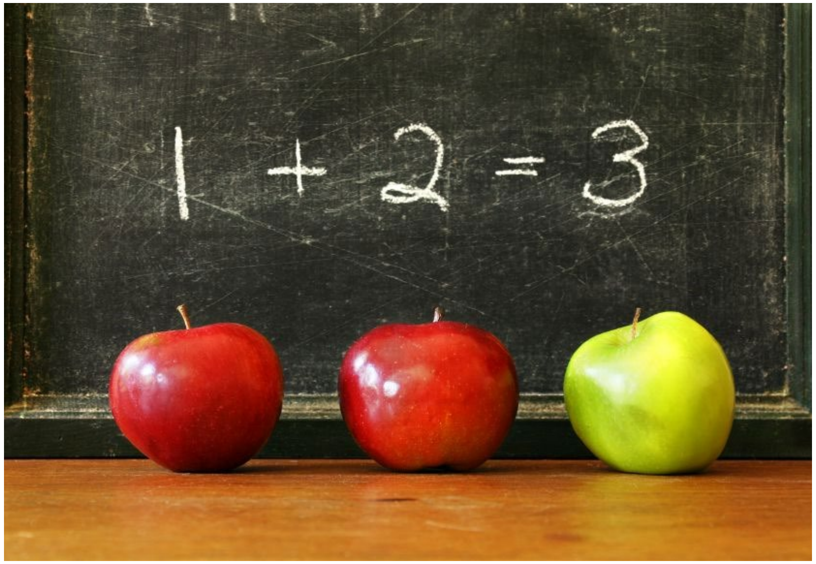


- **How many independent datasets are needed to avoid reporting random signatures as significant?**

- **What might explain the diff betw the observed (blue) and the theoretical (red) distributions?**




TRAY RETURN POINT



SUMMARY & CAUTIONARY NOTES

Anna Karenina Principle

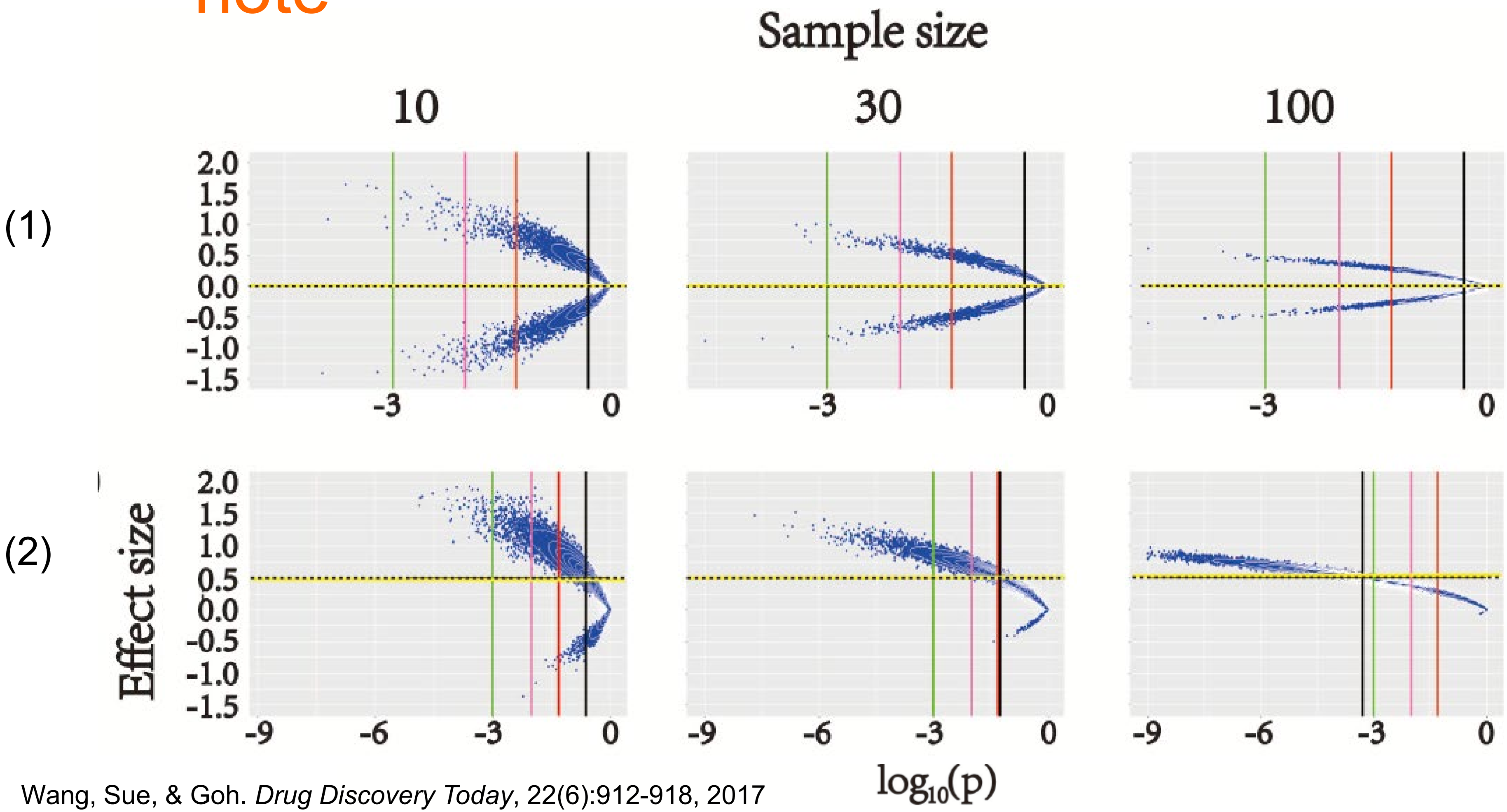
- **Careless null / alternative hypothesis due to forgotten assumptions**
  - Distributions of the feature of interest in the two samples are identical to the two populations
  - Features not of interest are equalized / controlled for in the two samples
  - No other explanation for significance of the 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**

A cautionary note

	Distribution		Mean		Standard deviation		Sample size		
Scenario	A	B	A	B	A	B			
(1)	Normal	Normal	0	0	1	1	10	30	100
(2)	Normal	Normal	0	0.5	1	1	10	30	100



Wang, Sue, & Goh. *Drug Discovery Today*, 22(6):912-918, 2017

Another cautionary note

NN	NN Acc. (%)	Acc. t_1-sparse (%)	Acc. t_2-sparse (%)	NPAQ r for t_1-sparse (%)	NPAQ r for t_2-sparse (%)
$ARCH_1$	74.00	78.00	81.00	20.31	62.50
$ARCH_2$	62.00	73.00	78.00	12.50	65.62
$ARCH_3$	76.00	82.00	83.00	45.31	52.34
$ARCH_4$	50.00	64.00	72.00	17.19	93.75
$ARCH_5$	78.00	82.00	83.00	74.22	24.22
$ARCH_6$	80.00	11.00	87.00	37.50	55.47
$ARCH_7$	87.00	89.00	89.00	6.25	79.69

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 orignal model.

Credit: Teodora Baluta

PhD program at NUS Graduate School of Integrative Sciences and Engineering, http://ngs.nus.edu.sg/graduate_programme.html

PhD program at NUS School of Computing, http://comp.nus.edu.sg/programmes/pg/phdcs