From bewilderment to enlightenment: Logic in cancer research

Wong Limsoon

Based on joint work with

Wilson Goh and Sriganesh Srihari

About Limsoon

Position

Kwan-Im-Thong-Hood-Cho-Temple Chair Professor, Dept of Computer Science, NUS

Research

database systems & theory, knowledge discovery, bioinformatics & computational biology

Honours

- **ACM Fellow**
- **FEER Asian Innovation Gold Award 2003**
- **ICDT Test of Time Award 2014**

Two bewilderments

Breast cancer survival signatures are no better than random signatures

Mutation mutual exclusivity are not associated with synthetic lethality

And maybe
some enlightenment at the end….

Some NUS numbers

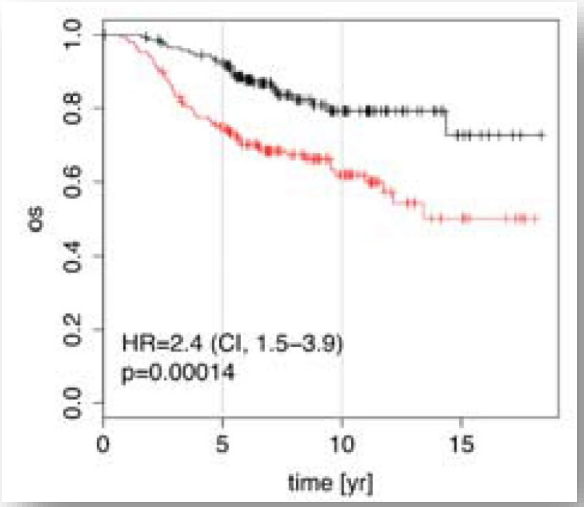
- **3 campuses**
  - Kent Ridge, Bukit Timah, & Outram
- **150 hectares**
- **13 undergrad schools**
- **4 graduate schools**
- **28k undergrads**
- **10k grad students**
- **2.4k faculty**
- **3.5k research staff**
- **5.4k other staff**

NUS National University of Singapore

Story #1

BREAST CANCER-SURVIVAL SIGNATURES

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

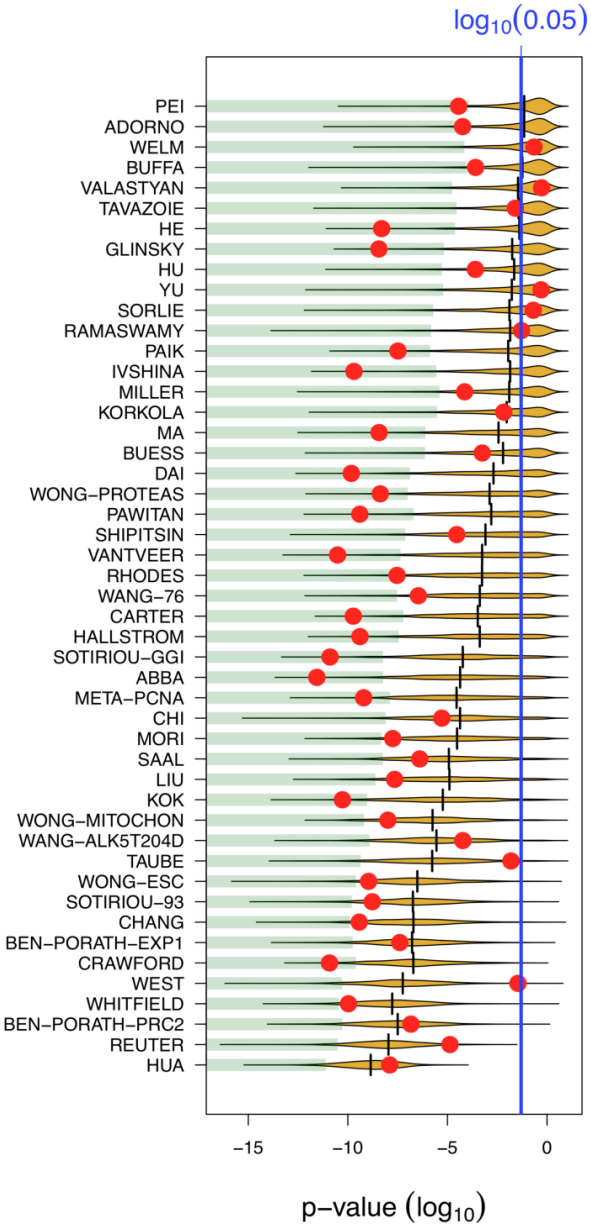


A seemingly obvious conclusion

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

In fact, almost all random signatures also have p-value < 0.05;

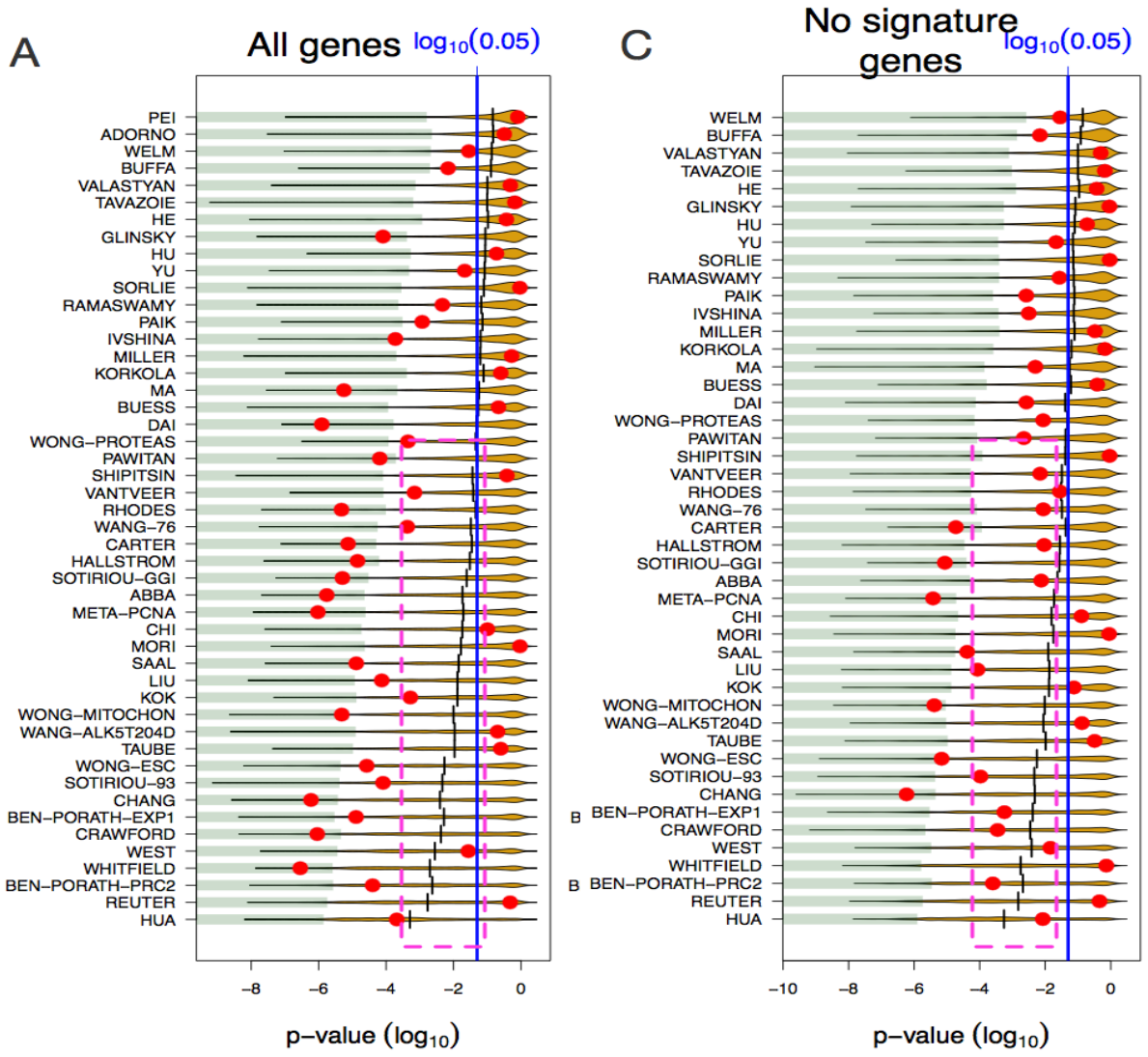
And the larger a random signature is, the more likely this happens

Goh & Wong, Why breast cancer signatures are no better than random signatures explained. *Drug Discovery Today*, 2018

NUS National University of Singapore

Maybe significant random signatures share genes with reported signatures?

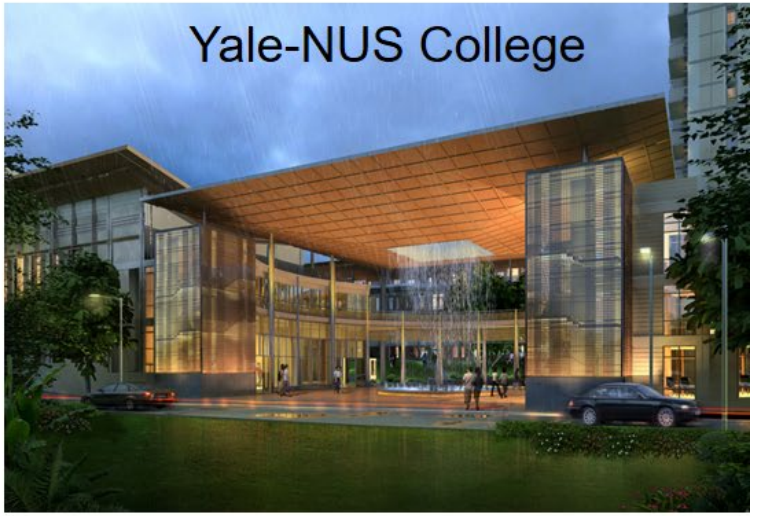
Not quite…



NUS National University of Singapore

Perhaps instead of asking whether a signature is significant, ask what makes a signature significant

NUS
National University of Singapore



Proliferation is a hallmark of cancer

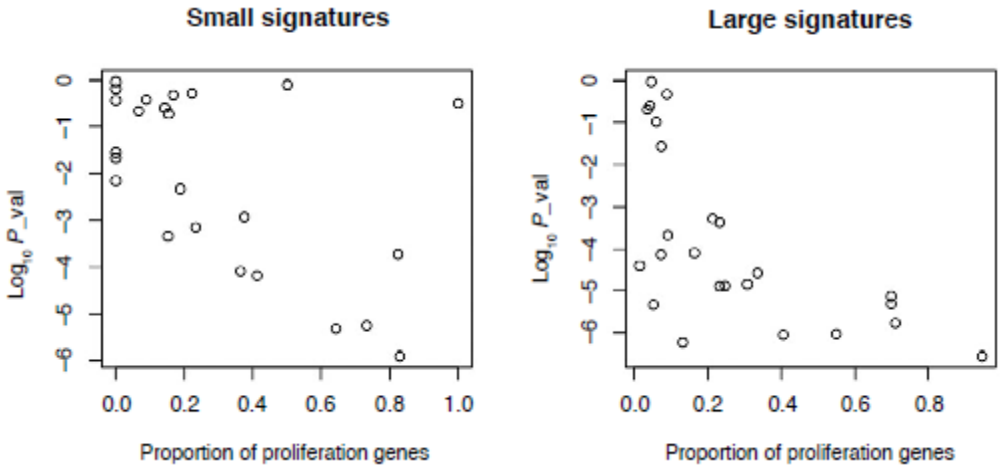
NUS National University of Singapore

Hypothesis: Proliferation-associated genes make a signature significant

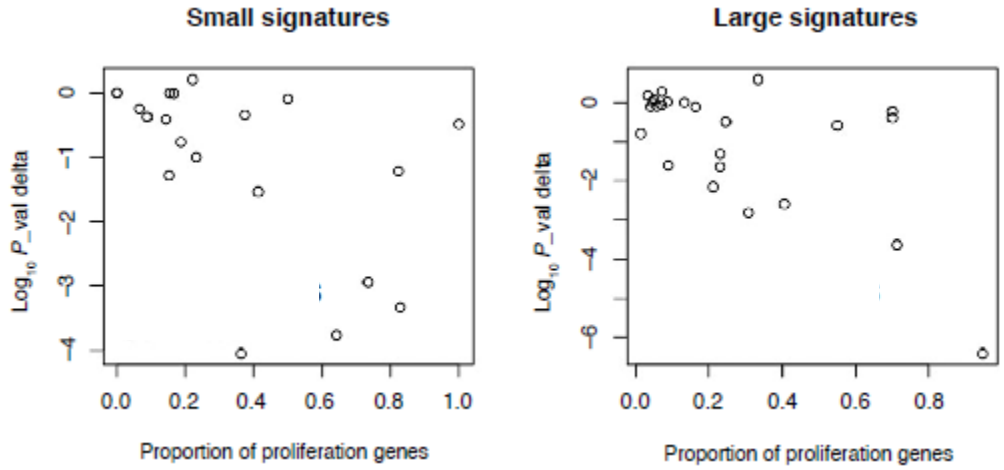
# of random signatures w/ ≥1 prolif gene

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



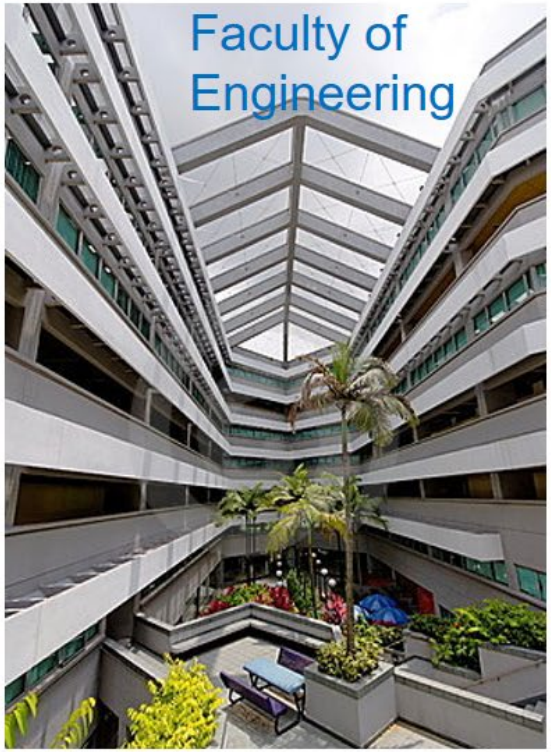
P-value of reported signatures, before removing proliferation genes



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?



NUS
National University
of Singapore

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:

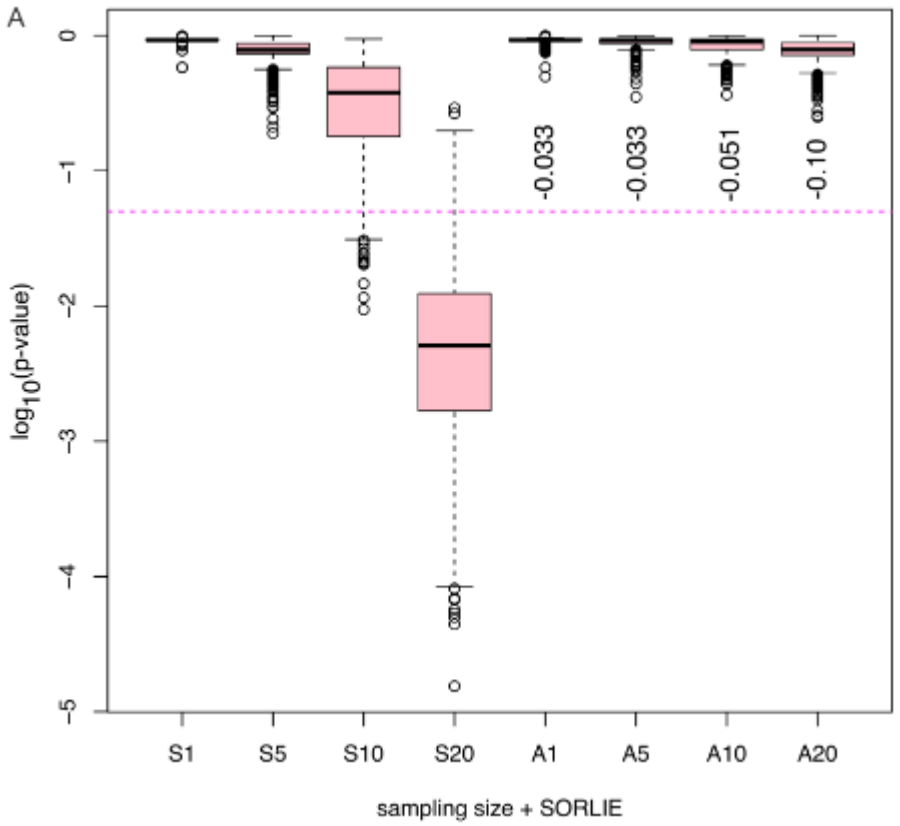
GLINSKY, DAI, RHODES, ABBA, WHITFIELD

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

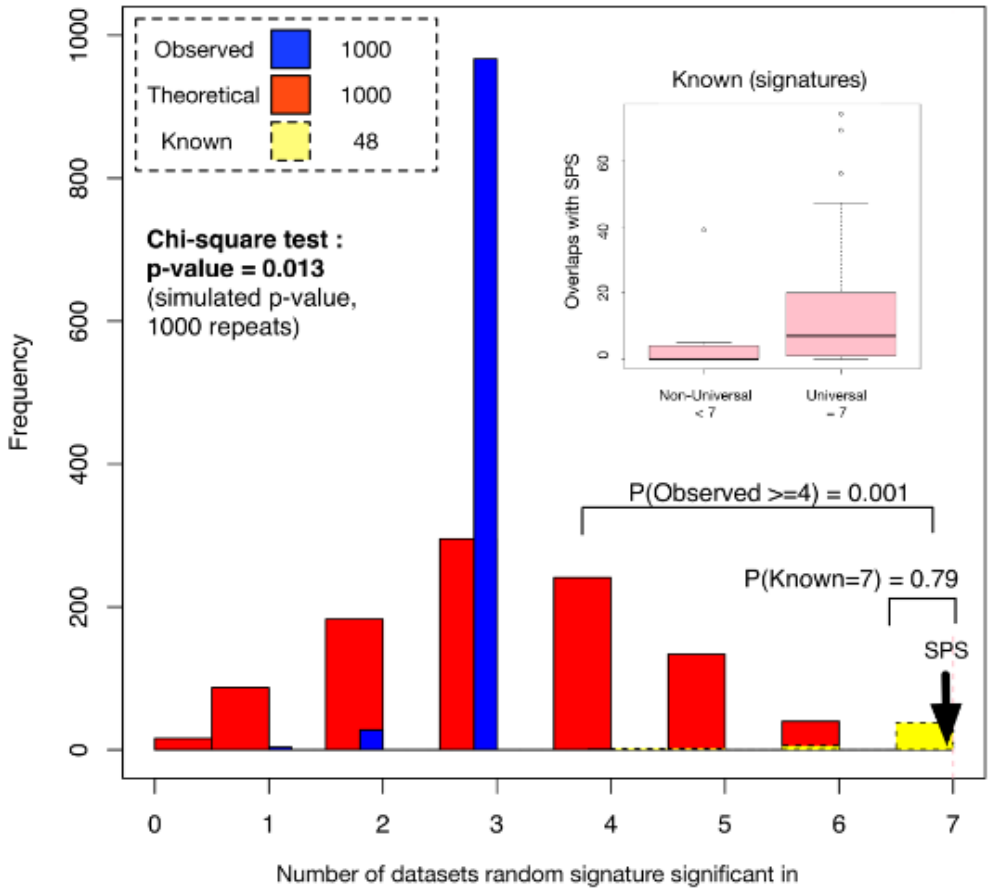
Test on many datasets

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?

n	$(50\%)^n$
1	50.00%
2	25.00%
3	12.50%
4	6.25%
5	3.13%
6	1.60%
7	0.78%

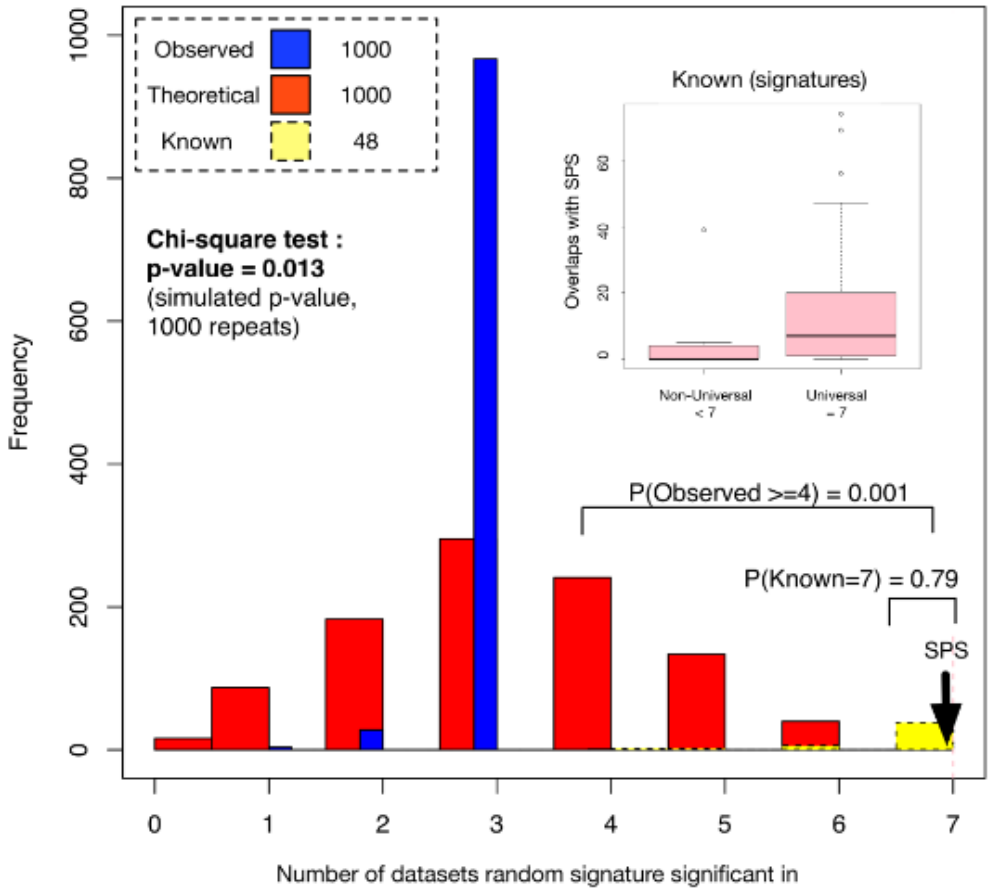
Test on many datasets



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

A theory-practice gap



~50% of random signatures are significant in 1 dataset

Red histogram is expected # of random signatures significant in n independent dataset (according to bionomial distribution)

Blue histogram is observed distribution

A related 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

NUS
National University
of Singapore

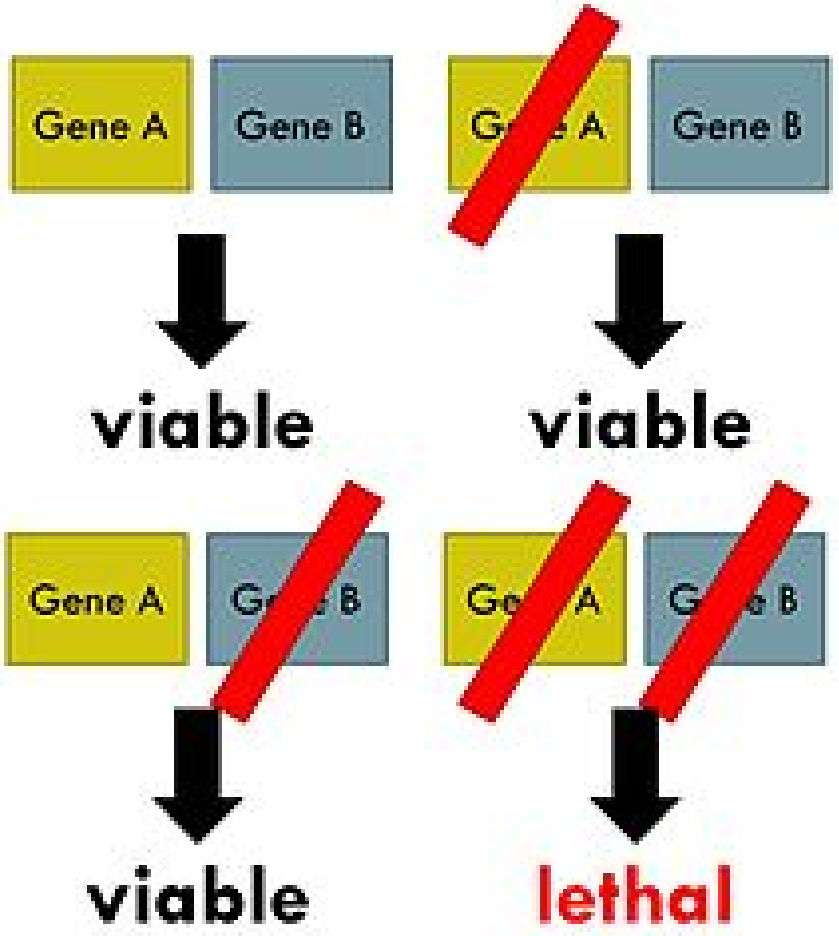
Alice Lee
Centre for
Nursing
Studies

NUS National University of Singapore

Story #2

SYNTHETIC LETHALS

Synthetic lethality



Some models of synthetic lethality

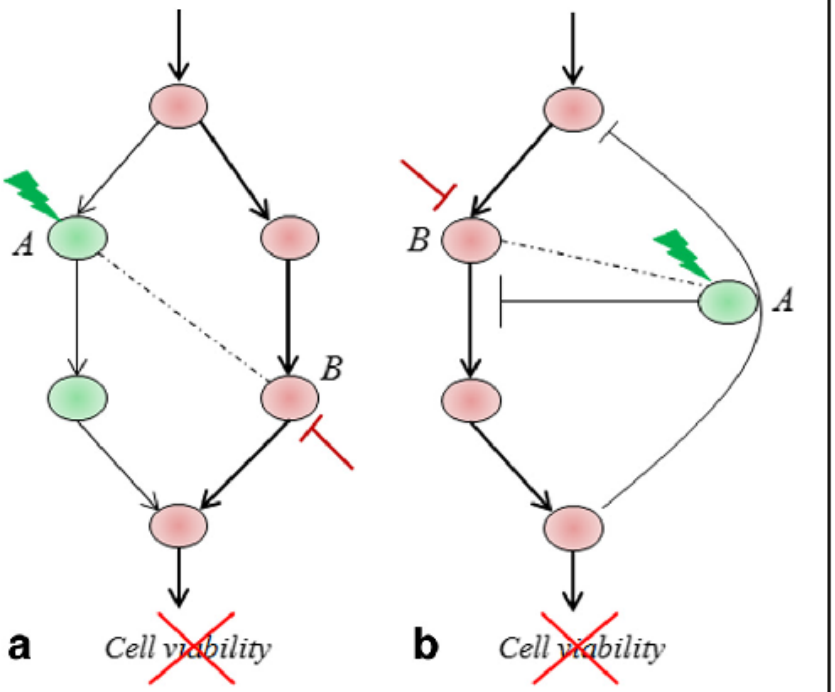


Fig. 7 Two models for pathway-based targeting of synthetic lethal genes *B* in conjunction with deleted/downregulated genes *A*: **a** parallel pathways model where targeting *B* results in disruption of both survival pathways, and **b** negative feedback-loop model where targeting *B* shunts of (forward)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

Why interested in synthetic lethality?

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

Synthetic lethality implies mutual exclusivity

Fact/postulate:

When a pair of genes is synthetic lethal, mutations of these two genes avoid each other

Observation:

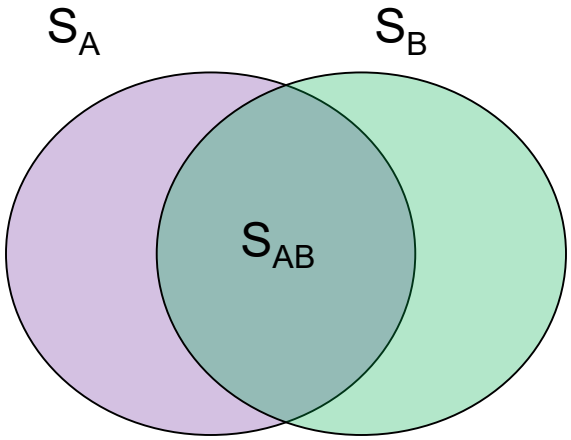
Mutations in genes (A,B) are seldom observed in the same subjects

Conclusion by abduction:

Genes (A,B) are synthetic lethal

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

A seemingly obvious approach based on hypergeometric test



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

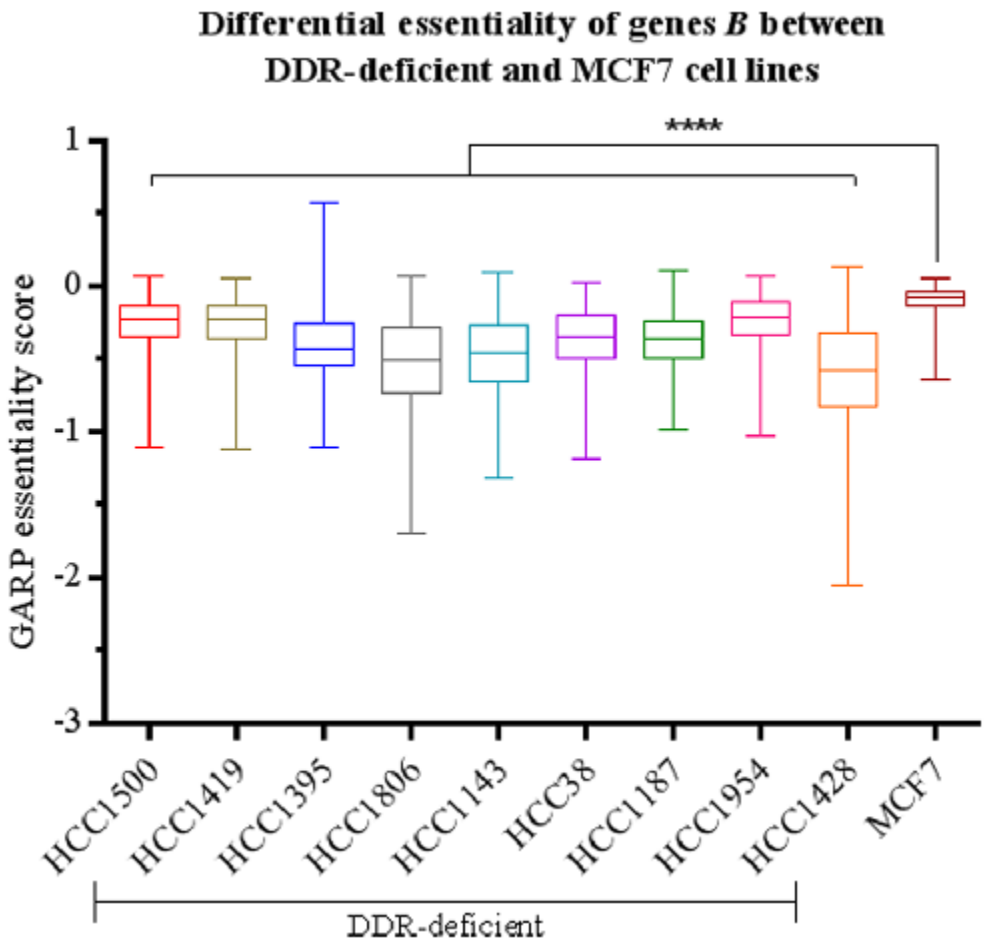
where $P[X > |S_{AB}|]$ is computed using the hypergeometric probability mass function for $X = k > |S_{AB}|$:

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

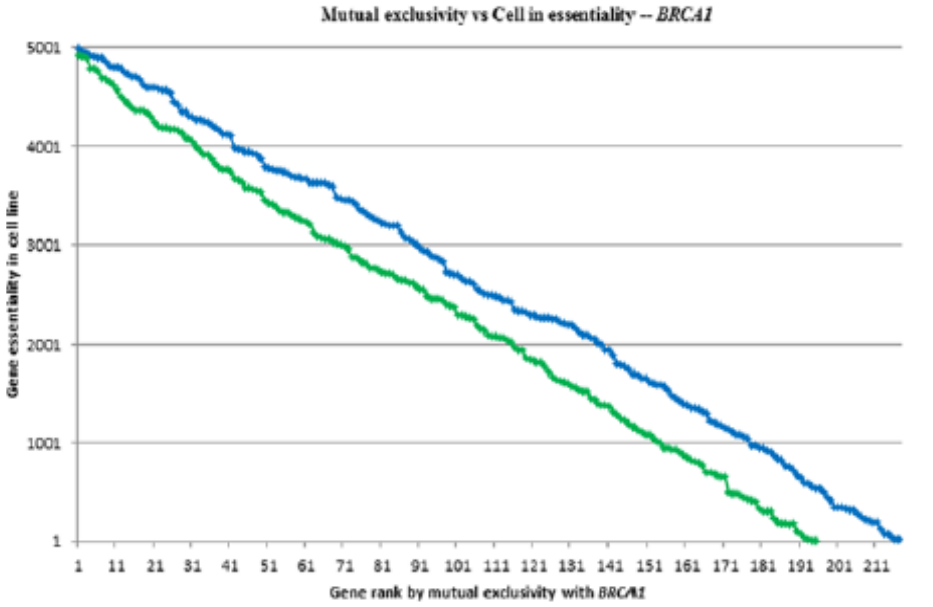
Mutations of genes (A,B) avoid each other if $P[X \leq S_{AB}] \leq 0.05$

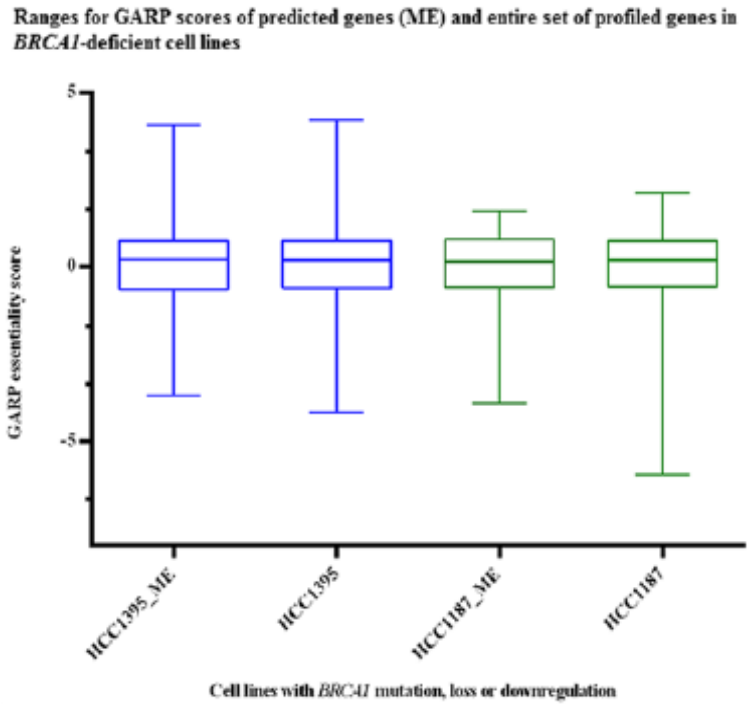
Anything wrong with this?

Seems to work fine



Really?





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


HEAT RETURN POINT

The hypergeometric distribution does not reflect real-world mutations

NUS National University of Singapore

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

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

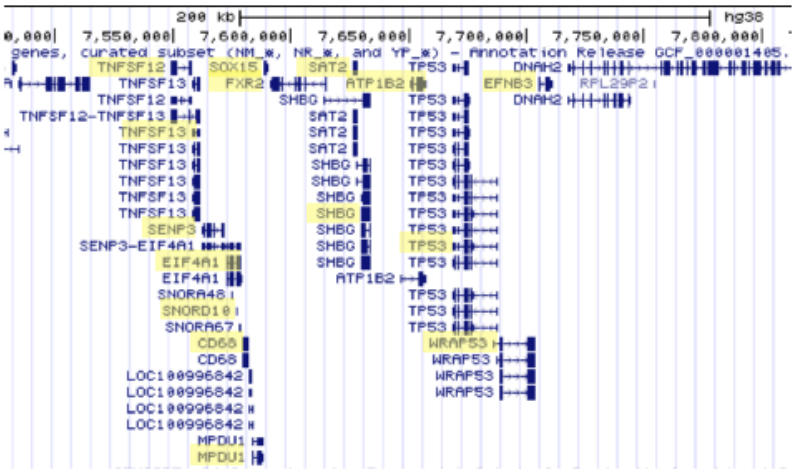
Mutations have equal chance to appear in a subject

Real-life mutations are inherited in blocks; those close to each other are correlated

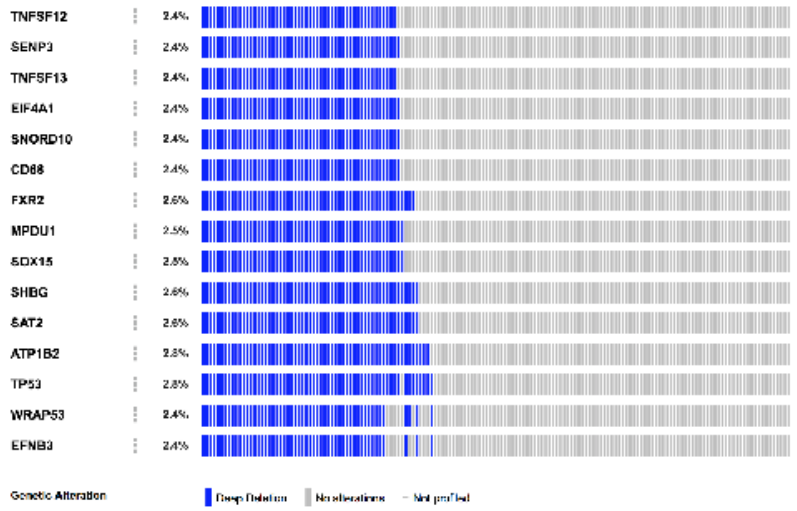
Some subjects have more mutations than others

⇒ Null distribution is not hypergeometric, binomial, etc.

Real-life example: Mutations of TP53 and its neighbours



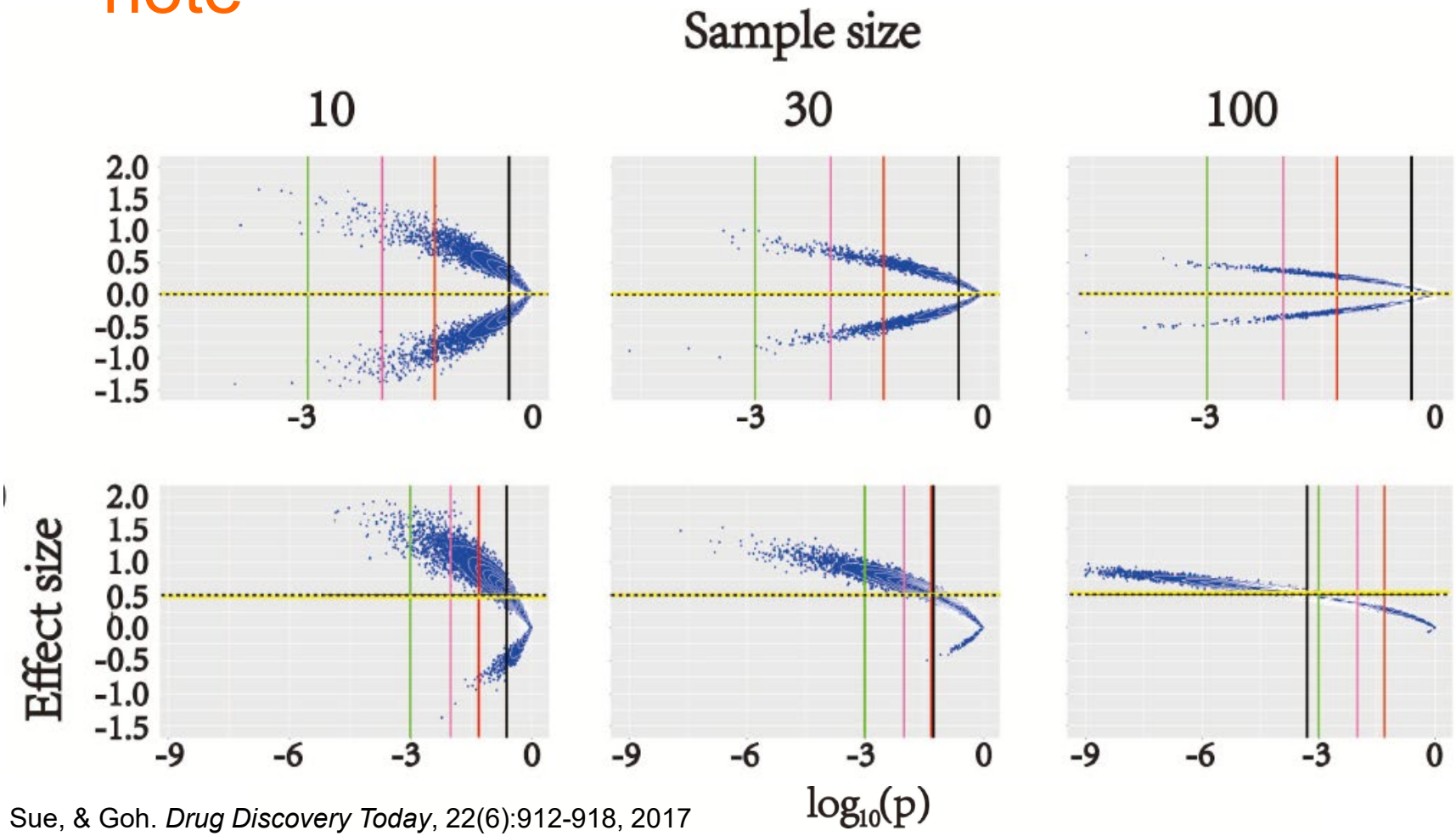
(a) Genomic location of genes close to TP53



(b) CNA profile of genes close to TP53

A related cautionary note

Scenario	Distribution		Mean		Standard deviation		Sample size		
	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

Solution?

Group genes into genomic clusters

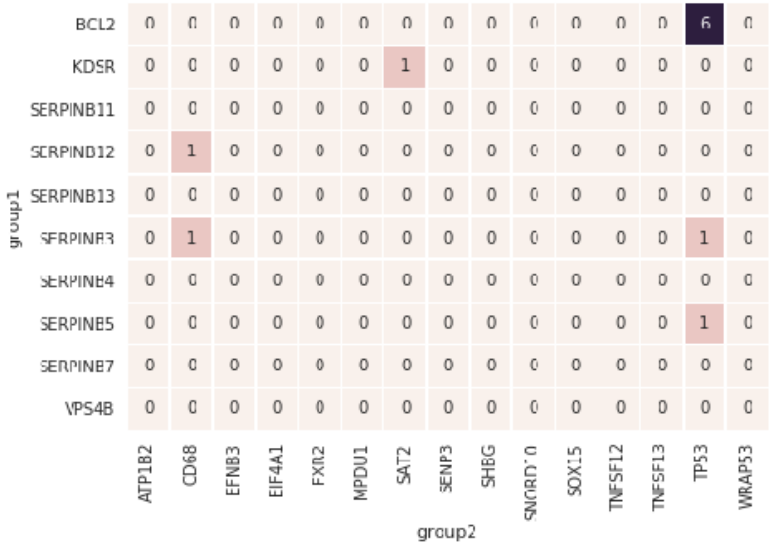
Test genes in far-apart genomic clusters for mutually exclusive mutations

Mutually exclusive clusters should contain synthetic-lethal & collateral-lethal gene pairs

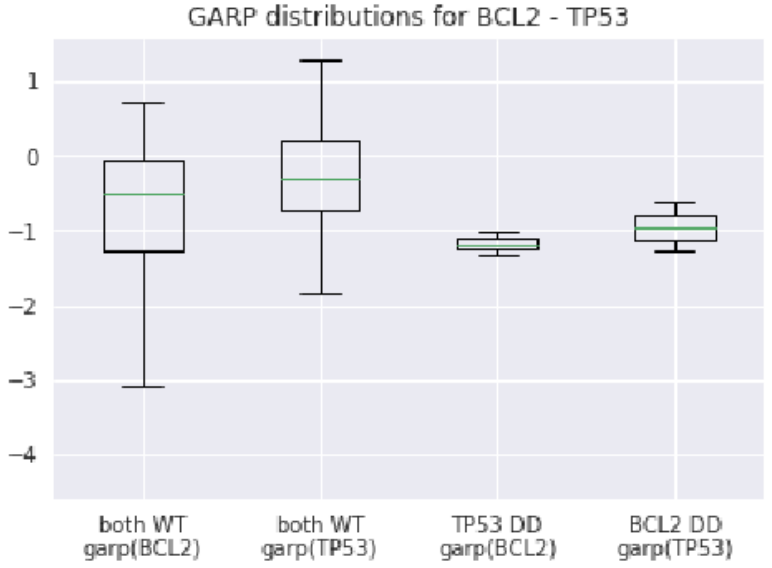
Prediction of collateral / synthetic lethality partners of 7 DDR genes

Gene	HDMI	ME Pairs	PW Confirmed
PTEN	0.040	71	53
TP53	0.025	51	36
BRCA2	0.030	34	22
ATM	0.020	26	19
CDH1	0.025	33	17
RB1	0.040	52	37
MSH3	0.025	18	11

A example of synthetic lethality: TP53-BCL2

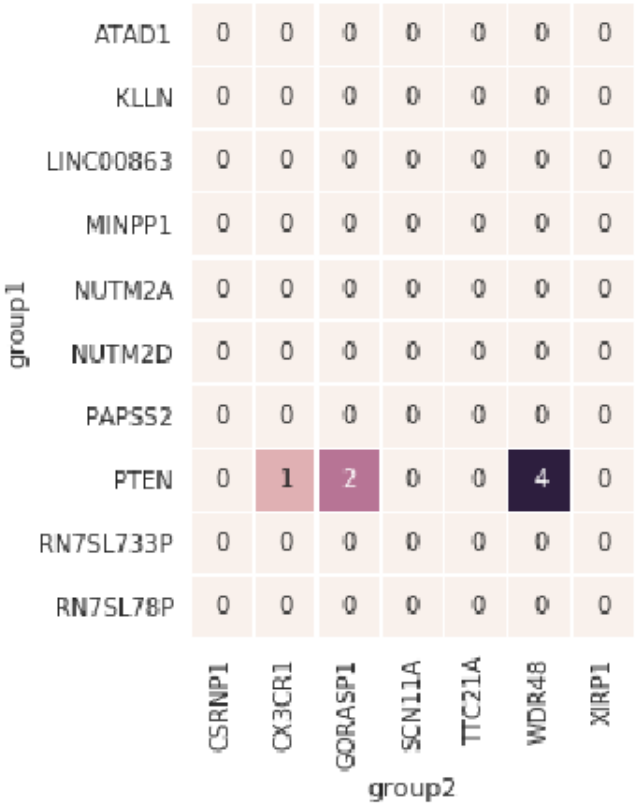


(a) Pathway heatmaps of the TP53 group and the BCL2 group.

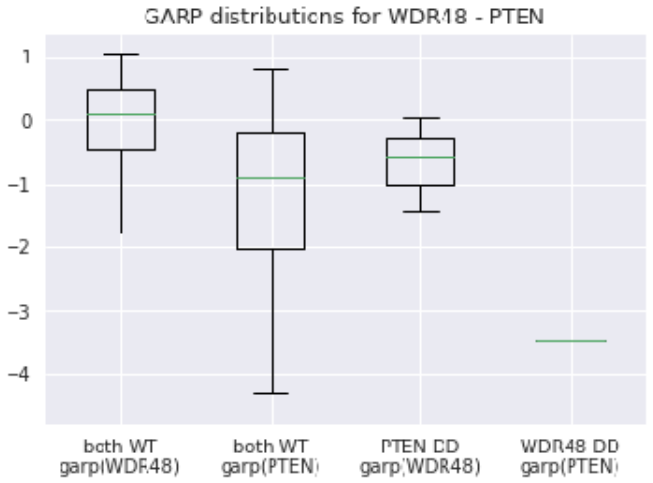


(b) GARP score distribution of the TP53 and BCL2 genes.

Another example: PTEN-WDR48



(a) Pathway heatmaps of the PTEN group and the WDR48 group.



(b) GARP score distribution of the PTEN and WDR48 genes.

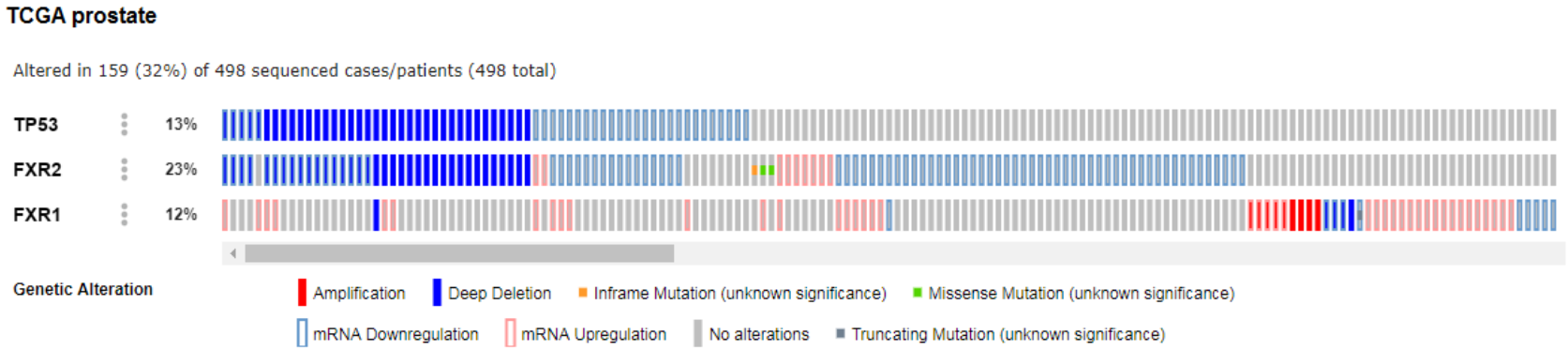
An example of collateral lethality

NUS National University of Singapore

FXR2 is located near TP53

FXR1 and FXR2 are paralogs that buffer each other's function

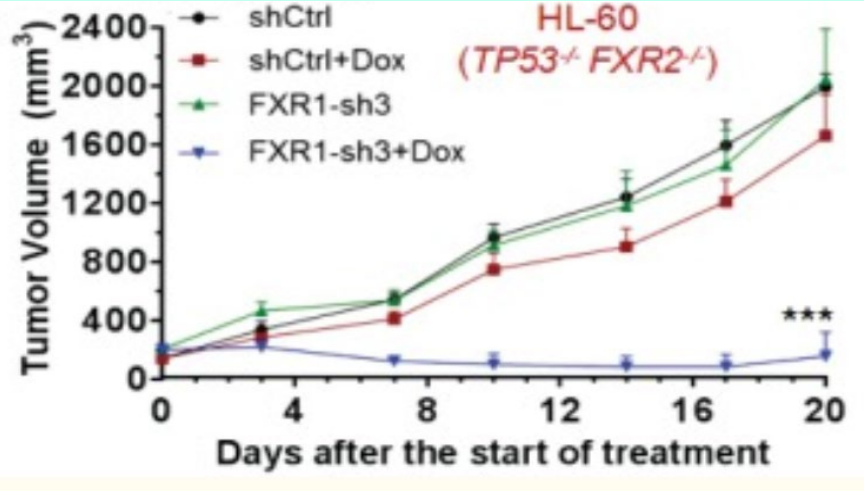
Do FXR1 and TP53 deletions avoid each other?



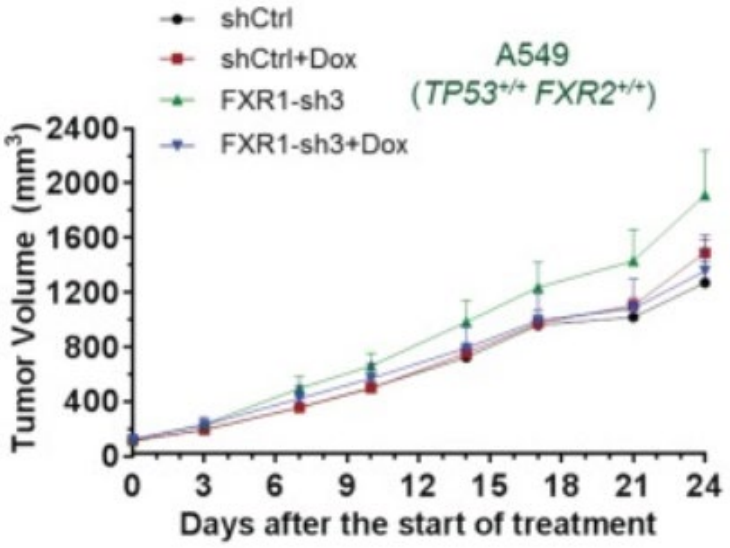
Is FXR1 synthetic lethal to TP53?

Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?

Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down



Fan et al., eLife, 6:e26129, 2017



NUS
National University of Singapore

Prince George's Park

NUS National University of Singapore

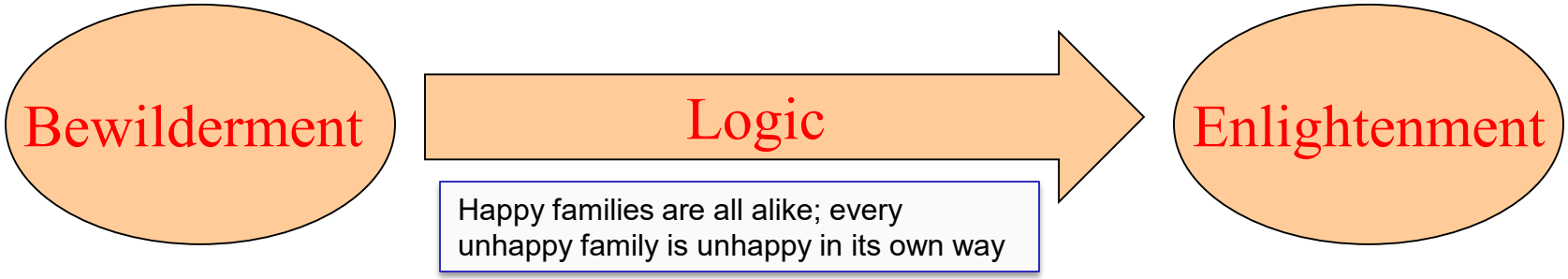
SUMMARY

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

Enlightenment: **SPS genes**

Bewilderment: **Mutation mutual exclusivity are not associated with synthetic lethality**

Enlightenment: **Collateral lethality**



References

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

Wilson Wen Bin Goh, Limsoon Wong. **Turning straw into gold: Building robustness into gene signature inference**. *Drug Discovery Today*, in press.

Sriganesh Srihari, Jitin Singla, Limsoon Wong, Mark Ragan. **Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer**. *Biology Direct*, 10:57, 2015.

NUS National University of Singapore

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