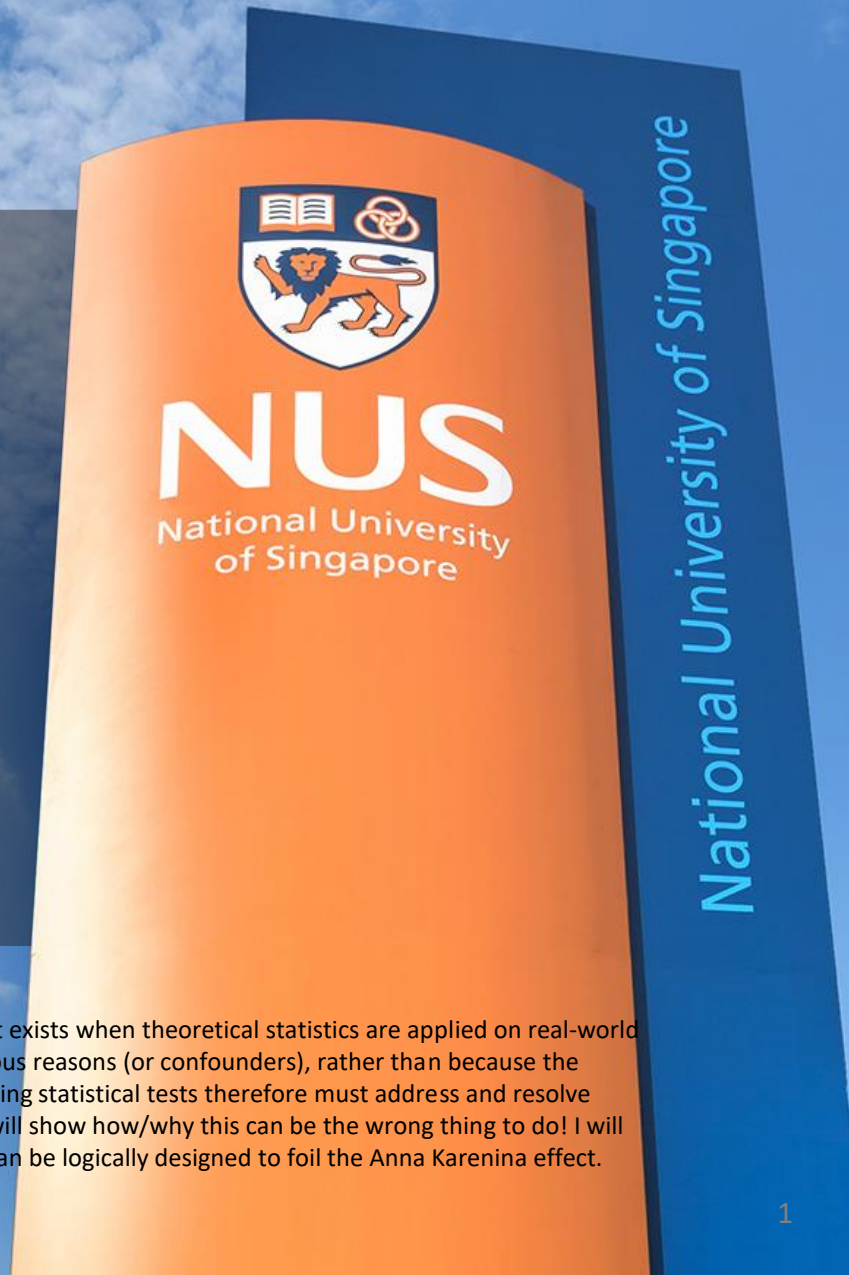
Anna Karenina Principle

Wong Limsoon



Outline: The Anna Karenina effect is a manifestation of the theory–practice gap that exists when theoretical statistics are applied on real-world data. It derives from the situation where the null hypothesis is rejected for extraneous reasons (or confounders), rather than because the alternative hypothesis is relevant to the disease phenotype. The mechanics of applying statistical tests therefore must address and resolve confounders. It is inadequate to simply rely on manipulating the P-value; indeed, I will show how/why this can be the wrong thing to do! I will discuss some mechanistic elements with real-life examples, and suggest how they can be logically designed to foil the Anna Karenina effect.

Hypothesis testing

Steps of hypothesis testing

Formulate null H_0 and alternate hypothesis H_1

Devise a test statistic, $t(\cdot)$

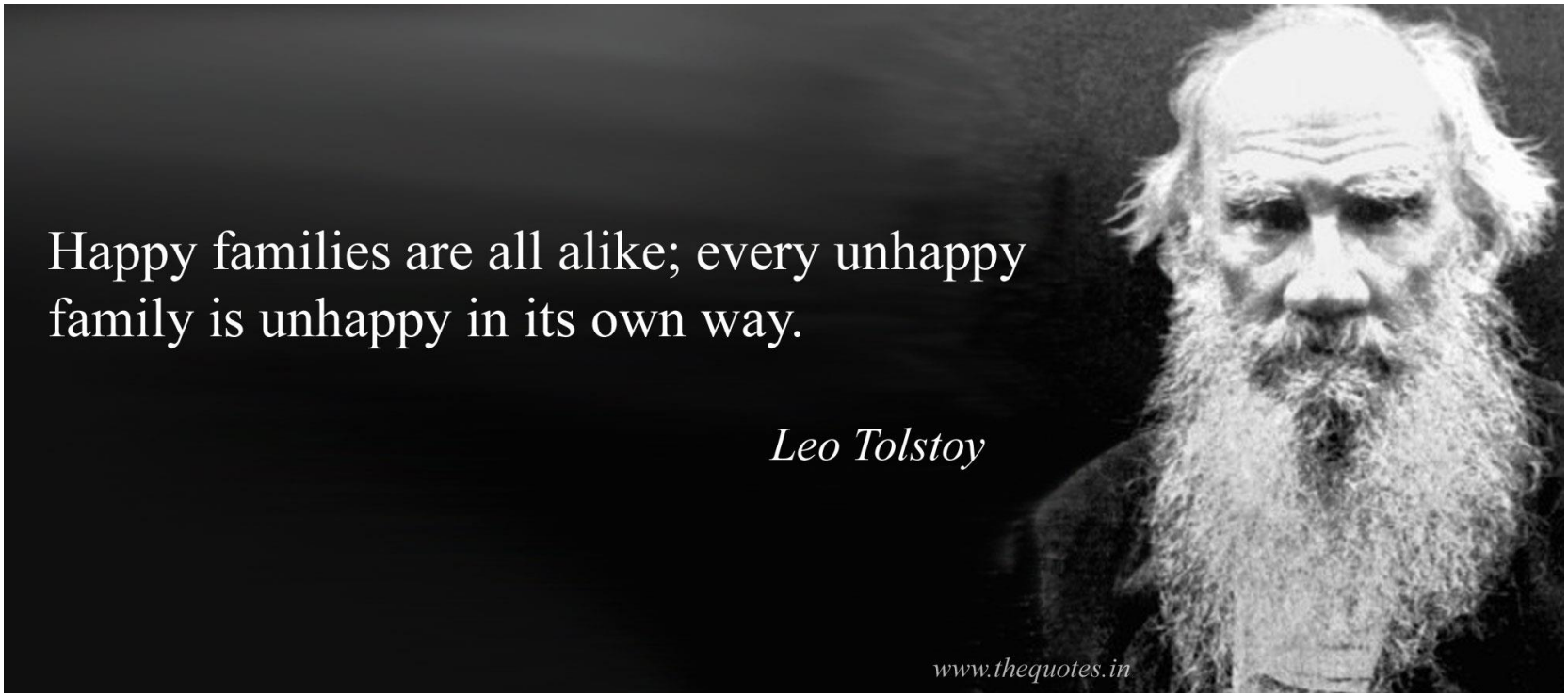
Evaluate $t(S)$ on a sample S

Compare $t(S)$ to the null distribution

If significant, accept H_1; otherwise, accept H_0

Null distribution is the distribution of $t(S_0)$ where S_0 ranges over the set of null samples S_0 for which H_0 holds

Anna Karenina



Anna Karenina Principle

There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest

Sample is biased

Null distribution used is inappropriate

Null / alternative hypothesis incorrectly stated

Inappropriate expt design

And so on

Biased sample

Exercise #1

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.

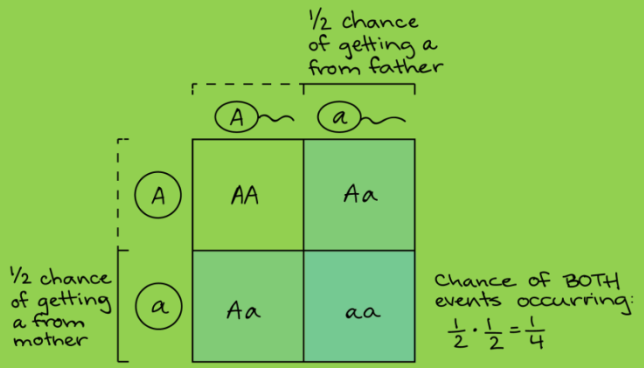
SNP rs123 is a great biomarker for a disease, based on a prospective study

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?

There may be sample bias



Basic rule of human genetics

Intentionally left blank

Careless null hypothesis

"Effective" H_0

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 H_1

rs123 alleles are differently distributed in the two populations

"Effective" H_1

rs123 alleles are differently distributed in the two populations OR

Distribution of rs123 alleles in the two samples are not identical to the two populations

Exercise #2

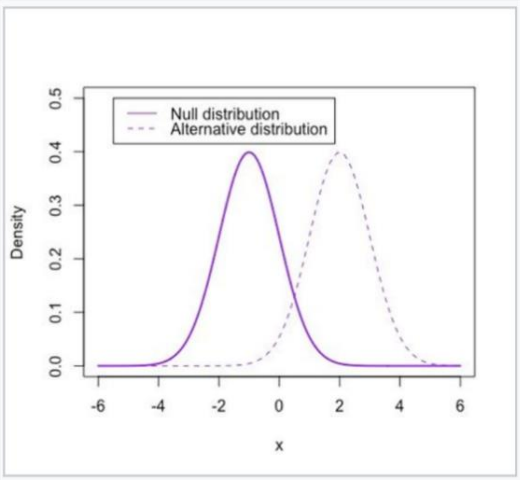
Suppose distributions of rs123 alleles in the two samples are identical to the corresponding populations and the test is significant

Can we say rs123 mutation causes the disease?

When two genes are close together, this is what happens during meiosis

Intentionally left blank

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

Inappropriate null distribution

Synthetic lethality

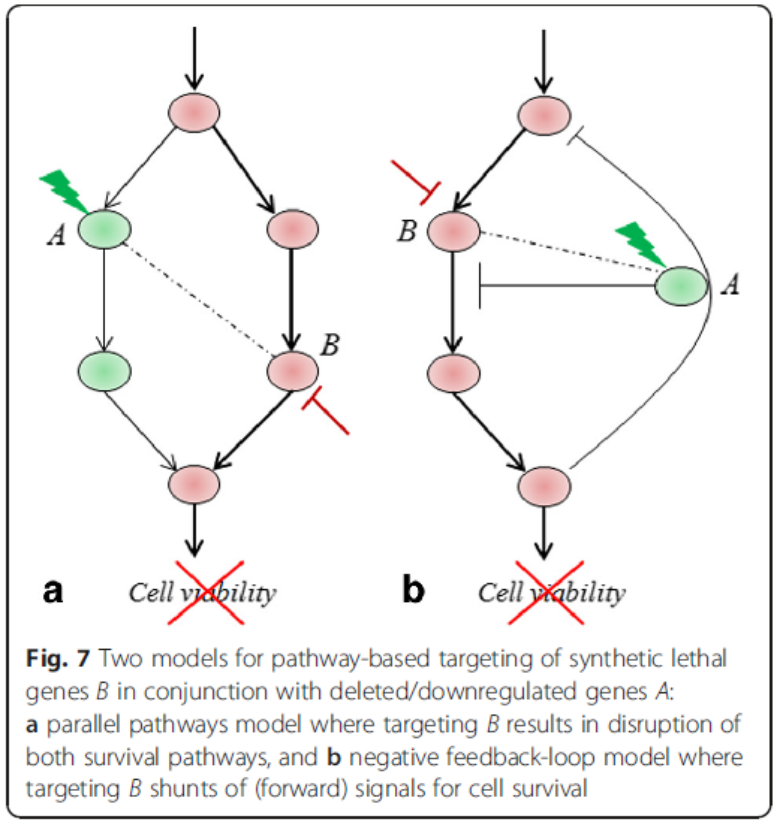


Fig. 7 Two models for pathway-based targeting of synthetic lethal genes *B* in conjunction with deleted/downregulated genes *A*: **a** parallel pathways model where targeting *B* results in disruption of both survival pathways, and **b** negative feedback-loop model where targeting *B* shunts of (forward)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

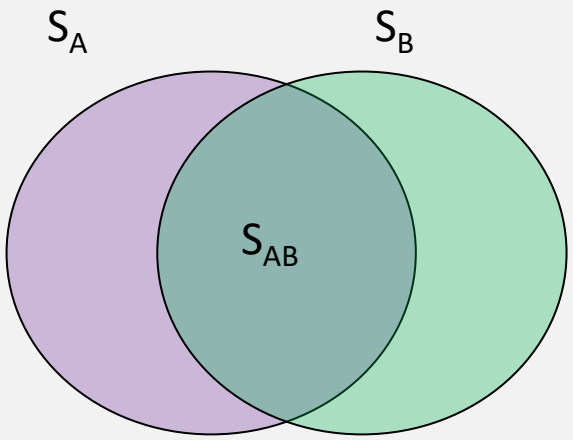
Observation:

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

Conclusion by abduction:

Genes (A,B) are synthetic lethal

Exercise #3



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

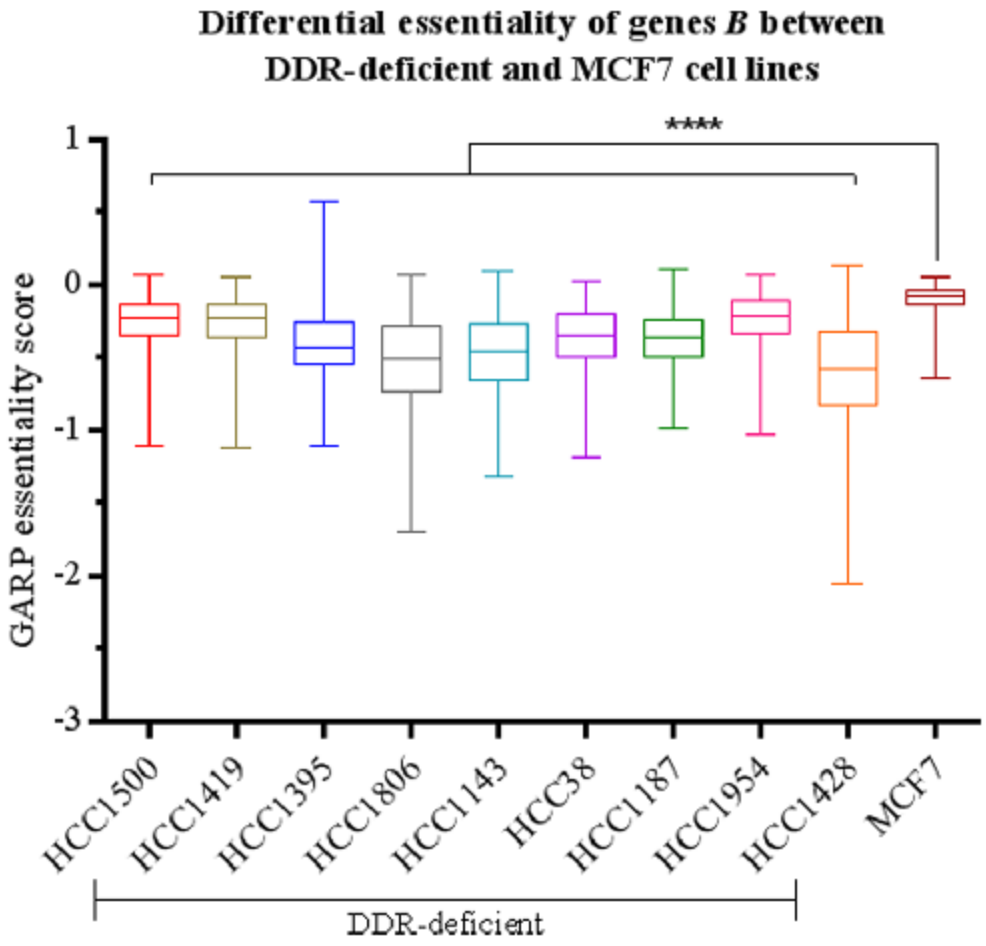
where $P[X > |S_{AB}|]$ is computed using the hypergeometric probability mass function for $X = k > |S_{AB}|$:

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

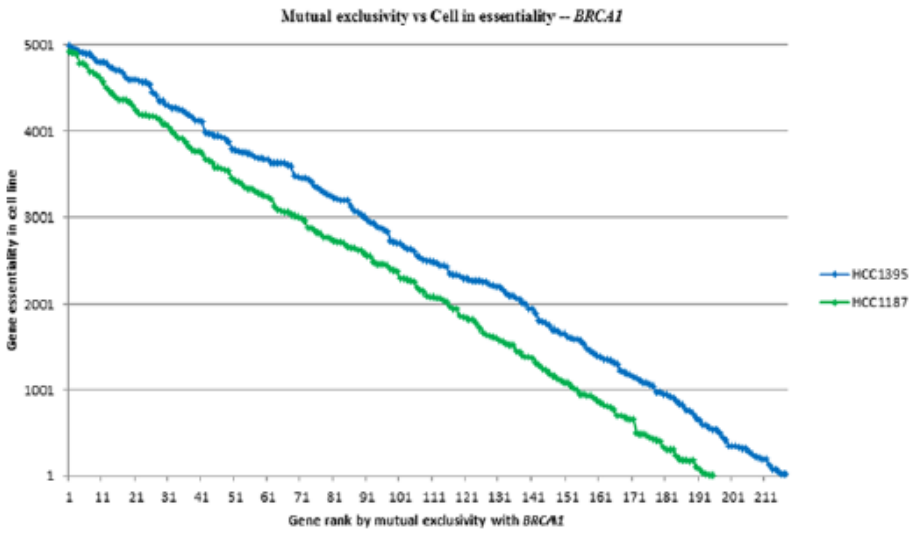
Mutations of genes (A,B) avoid each other if $P[X \leq S_{AB}] \leq 0.05$

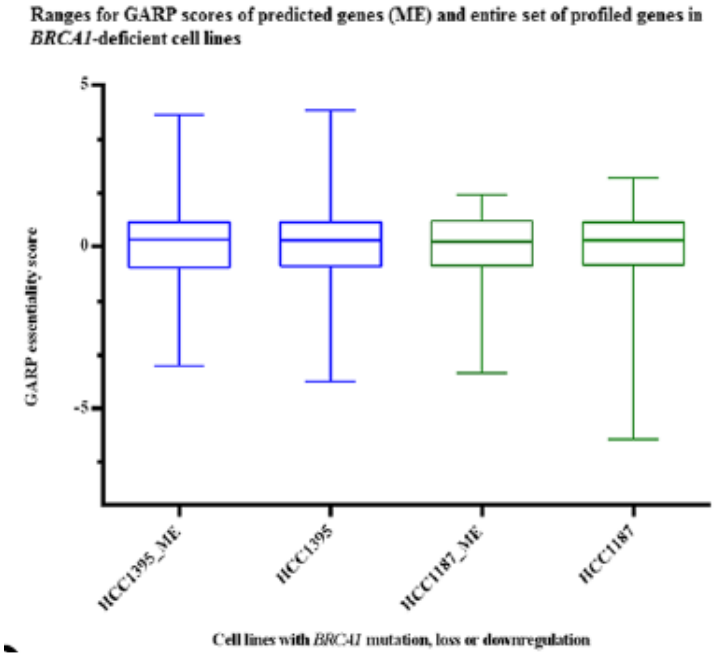
Anything wrong with this?

Seems to work fine



What is happening?

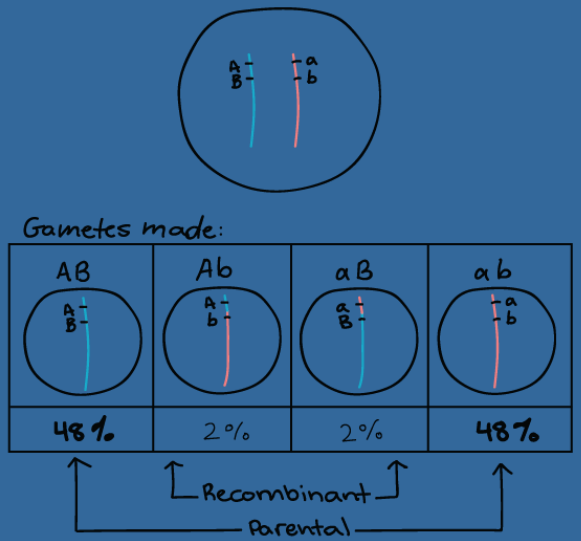




Among top ME-genes, GARP score ranks correlate with mutual exclusion ranks

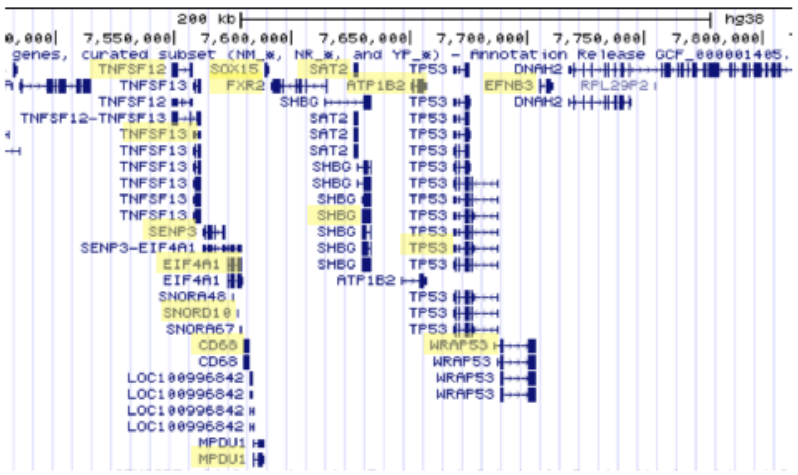
But GARP scores of ME-genes (i.e. have mutually exclusive mutations to BRCA1) are like other genes

Hyper-geometric distribution doesn't reflect real mutations

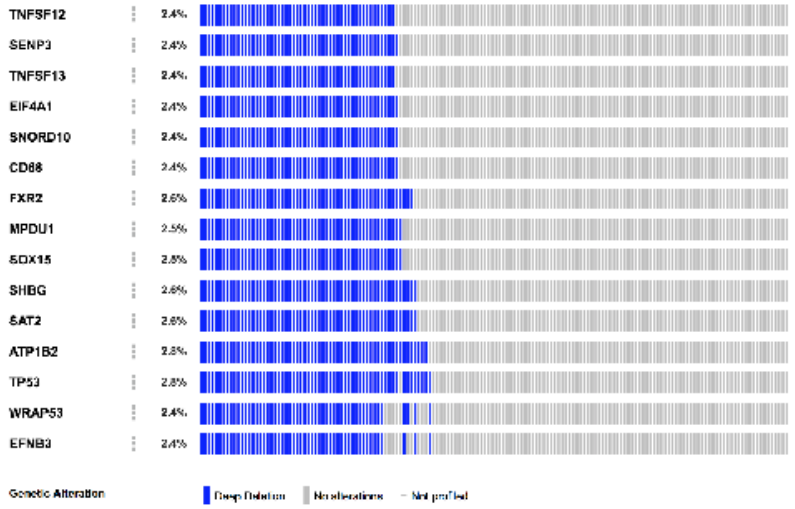


Intentionally left blank

Real-life example: Mutations of TP53 and its neighbours



(a) Genomic location of genes close to TP53

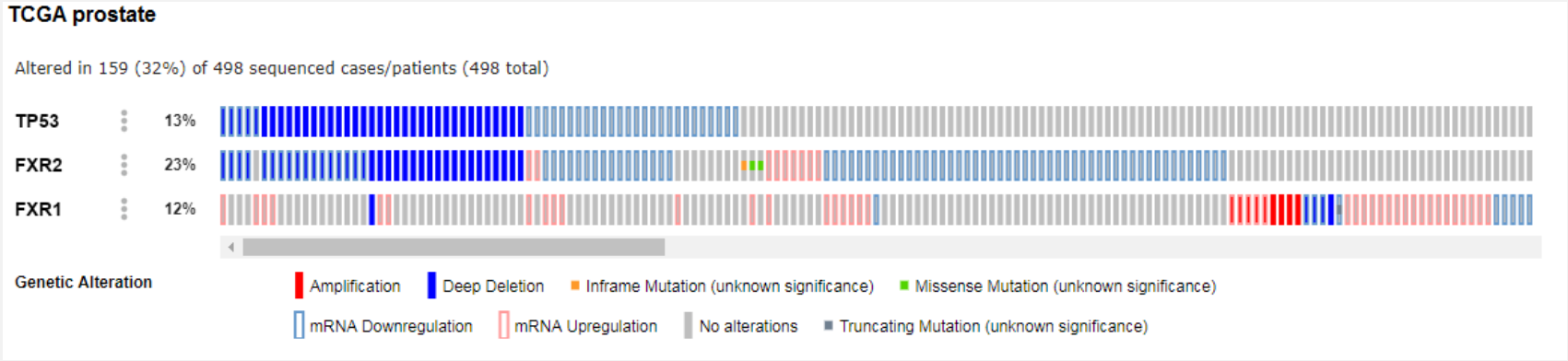


(b) CNA profile of genes close to TP53

Exercise #4

FXR2 is located near TP53

FXR1 and FXR2 buffer each other's function



Is FXR1 synthetic lethal to TP53?

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

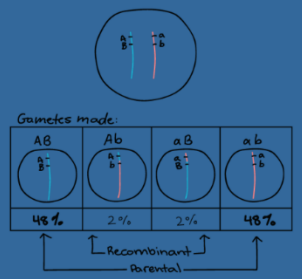
Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down

Intentionally left blank

Exercise #5

Propose some possible solutions to this problem

Hyper-geometric distribution doesn't reflect real mutations



Hypergeometric distribution

Mutations are independent

Mutations 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

Inappropriate experiment design

Exercise #6

Overall

	A	B
lived	60	65
died	100	165

Treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Treatment B is better

What is happening here?

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

Intentionally left blank

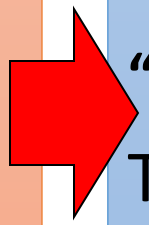
Careless null hypothesis

"Effective" H_0

Treatment effects are identically distributed in the two samples

Assumption

All other factors are equalized in the two samples



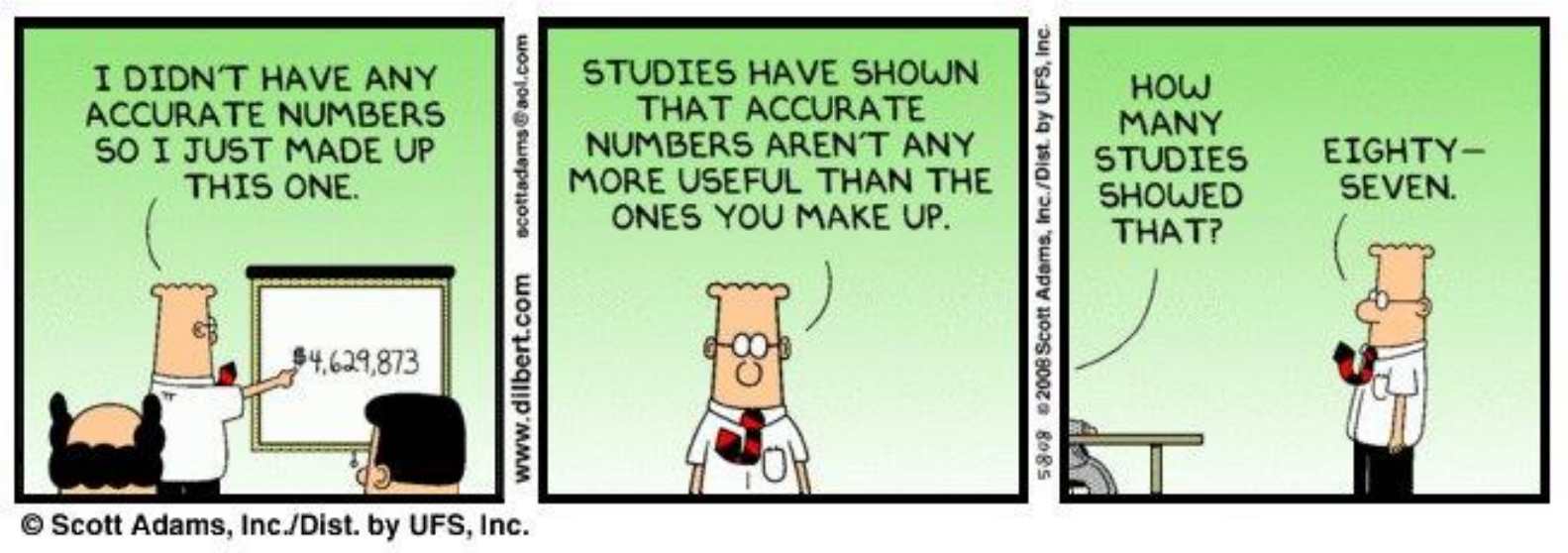
Apparent H_1

Treatment effects are differently distributed in the two populations

"Effective" H_1

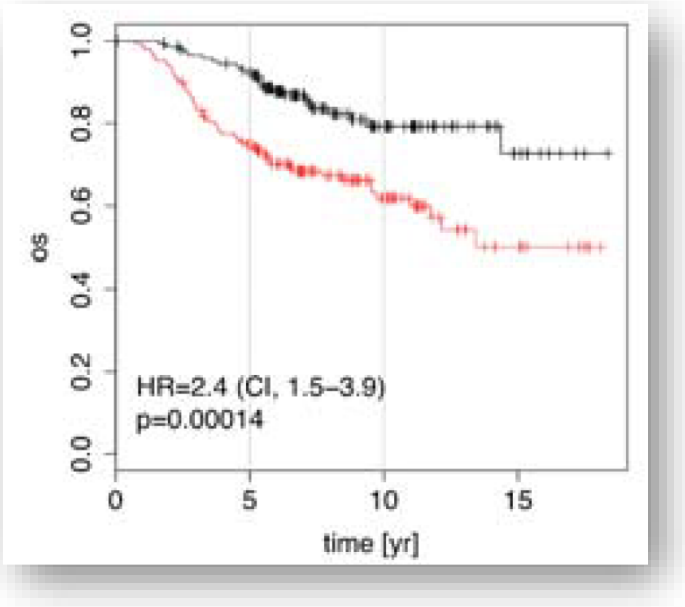
Treatment effects are differently distributed in the two populations OR

Some other factors aren't equalized in the two samples



Confounders abound

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



A seemingly obvious conclusion

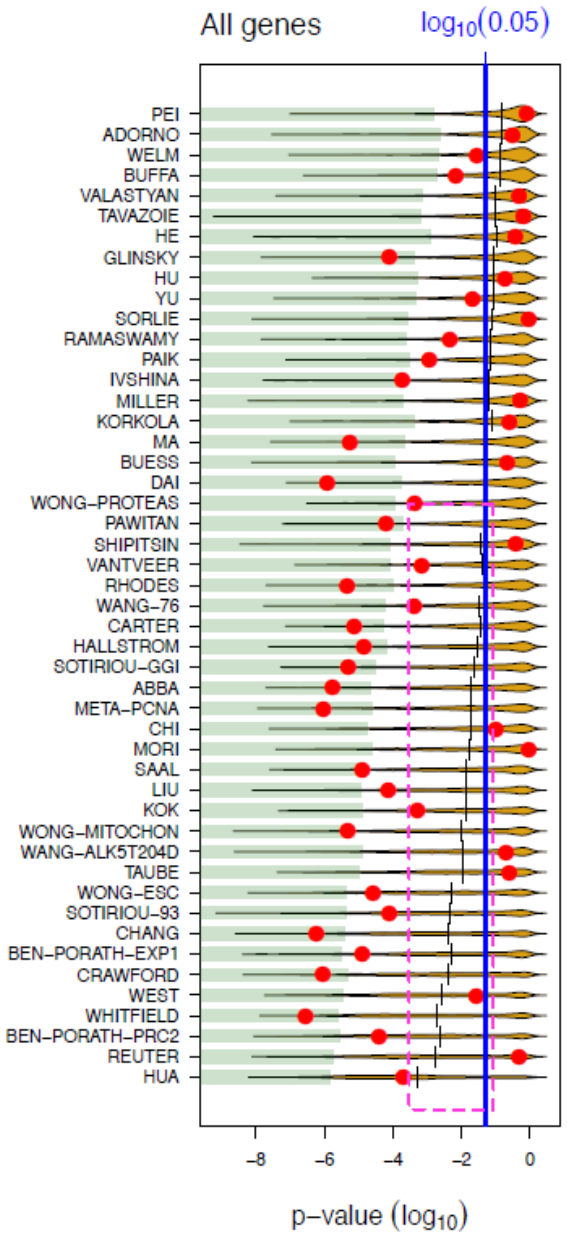
A multi-gene signature (social defeat in mice) good as a biomarker for breast cancer survival

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

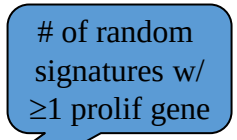
Venet et al., *PLOS Comput Biol*, 2011



What makes random signatures significant?

Proliferation is a hallmark of cancer

Hypothesis: Proliferation-associated genes make a signature significant

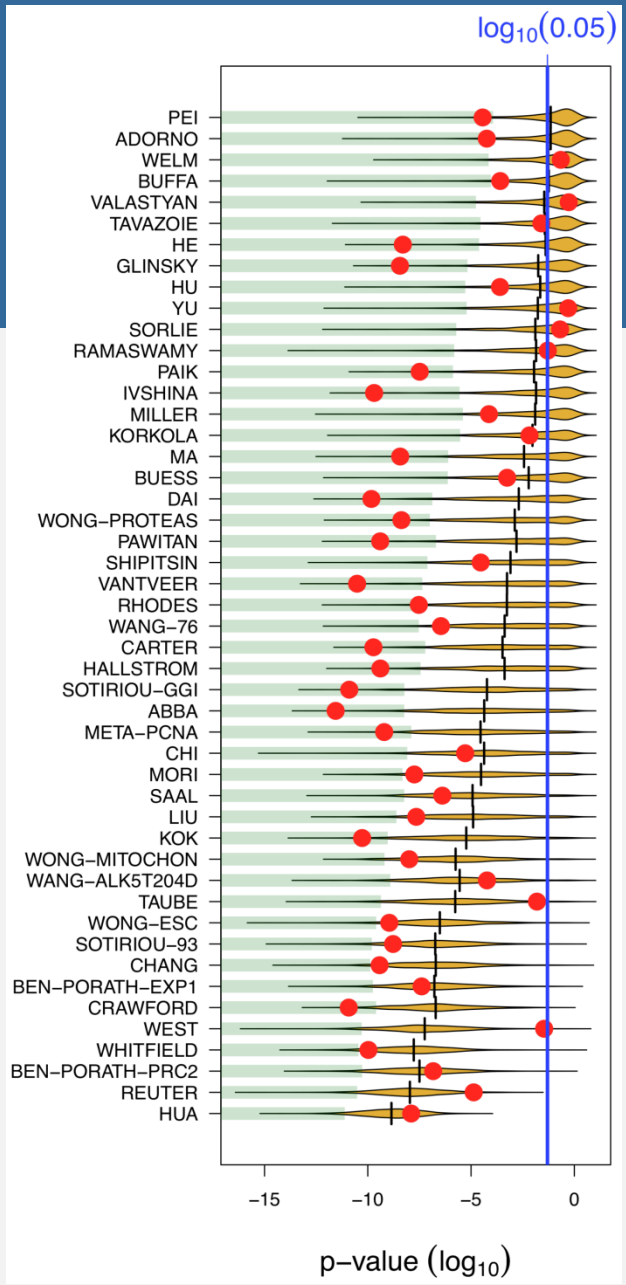


Cutoffs	Counts		
	NP	P	Marginals
Above 0.05	7043	19 043	26 086
Below 0.05	2766	19 148	21 914
Marginals	9809	38 191	48 000

Exercise #7

40-50% of random signatures have p-value << 0.05

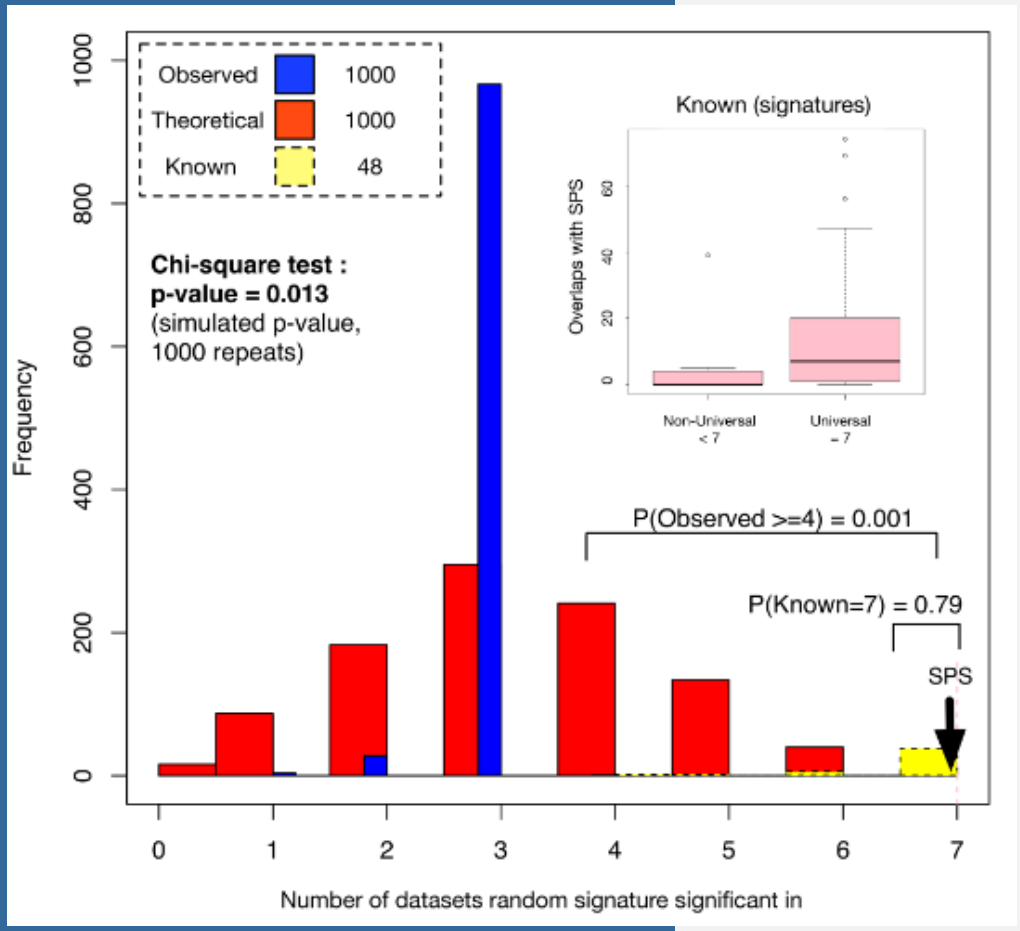
How to get rid of them?



An engineer's solution

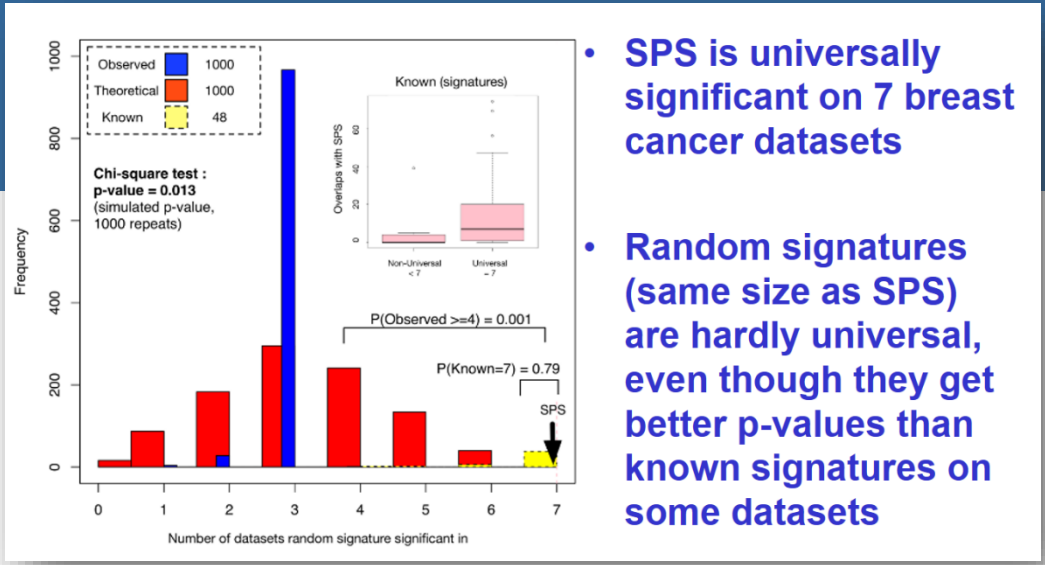
Intentionally left blank

Test on many datasets



Intentionally left blank

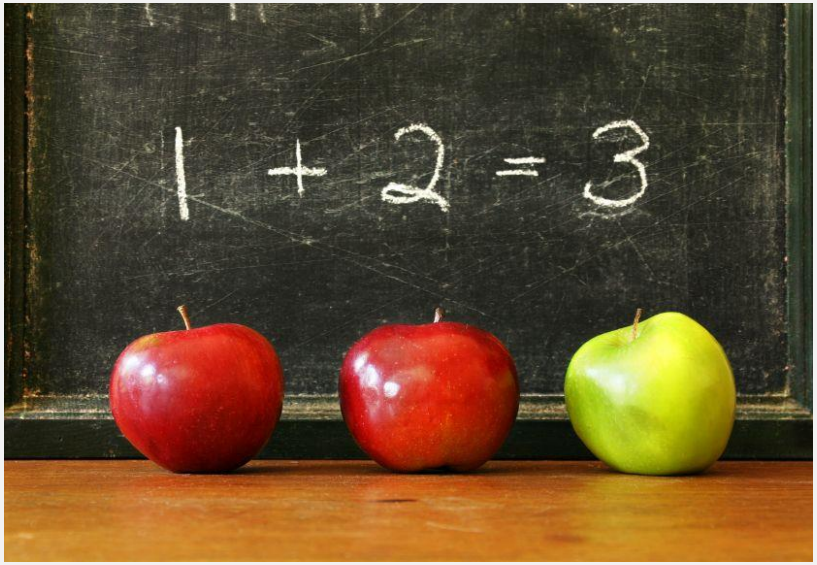
Exercise #8



- **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**

The red bars show the theoretical binomial distribution on expected # of random signatures that should be significant on n datasets

What do you think is happening here?


1 + 2 = 3

What have we learned?

When a statistical test is significant, think again!

Sample is biased

Null distribution used is inappropriate

Null / alternative hypothesis incorrectly stated

Inappropriate expt design

Confounders are aplenty

“Independent” test data are not as independent as you think

References

Goh & Wong. Dealing with confounders in –omics analysis. *TIBTECH*, 36(5):488-498, 2018

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

Pinoli et al. Identifying collateral and synthetic lethal vulnerabilities within the DNA-damage response. *BMC Bioinformatics*, 22:250, 2021

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

Goh & Wong. Turning straw into gold: Building robustness into gene signature inference. *Drug Discovery Today*, 24(1):31-36, 2019

Ho et al. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, 1(8):100129, 2020

Projects

Project 1. Vanderbilt Study: GRE score and PhD performance

Moneta-Koehler et al. PLOS ONE 12(1):e0166742, 2017

What are the main claims of this study? Can you find some analysis/methodological bugs that might invalidate some of these claims?

Project 2. Lung cancer and Doppelgangers

Coudray et al. Nature Medicine 24:1559-1567, 2018

What are the main claims of this study? Can you find some analysis/methodological bugs that might invalidate some of these claims?

Project 3. Protein function and Twilight Zone

Seo et al. Bioinformatics 34(13):254-262, 2018

What are the claims of this study? Can you find some analysis/methodological bugs that might cast doubts on these claims?

And when a statistical test is not significant, it may not be insignificant

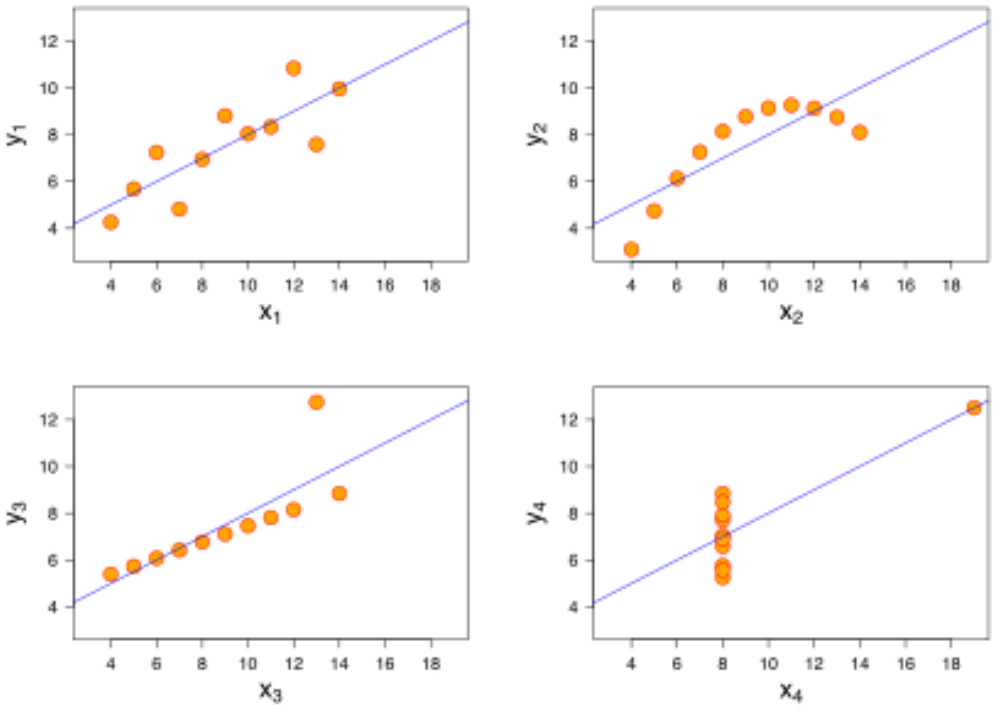
Anscombe's quartet

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

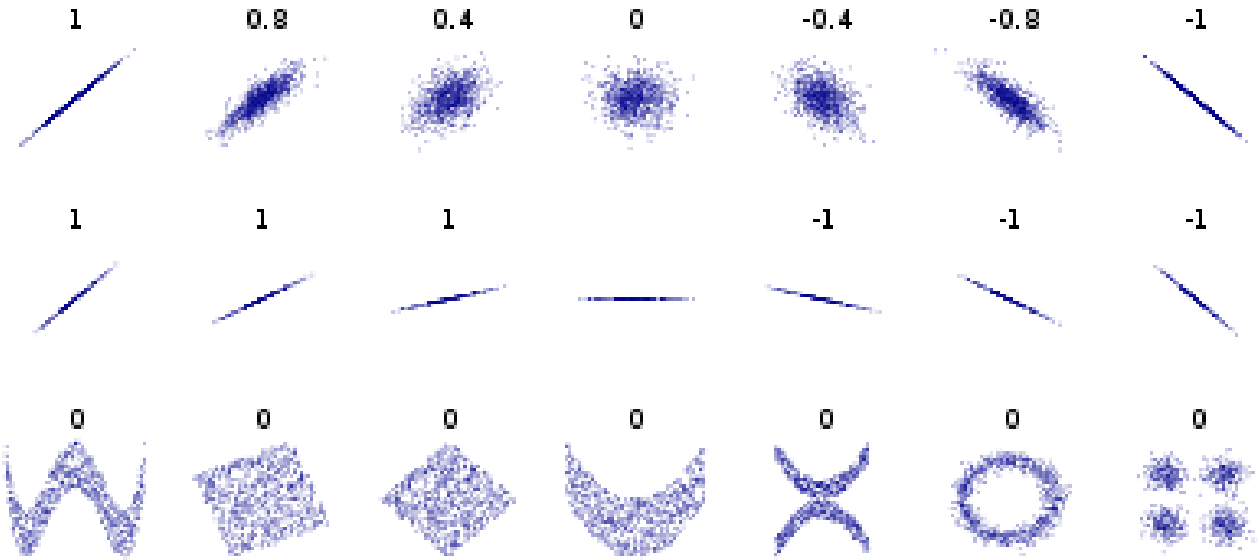
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

The R^2 = 0.67 of the correlation line is good

Is there really x-y correlation in each of these cases?



Correlation and association

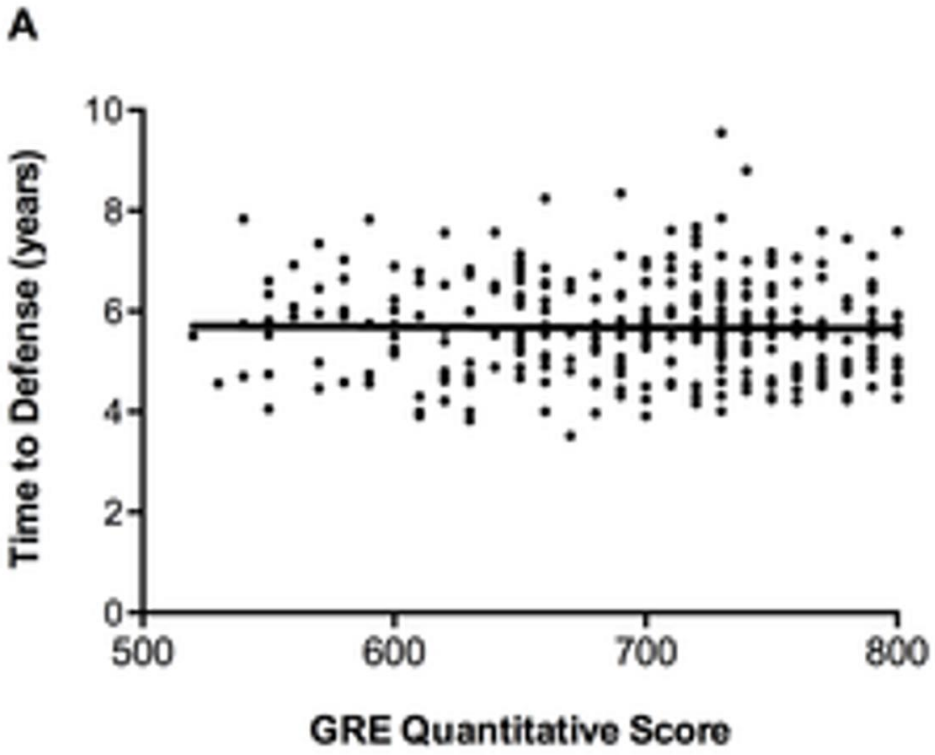


Association - any relationship betw 2 variables

Correlation – a linear relationship betw variables

Vanderbilt study

Some studies suggest no correlation betw GRE and PhD outcomes (e.g. passing the PhD on time)

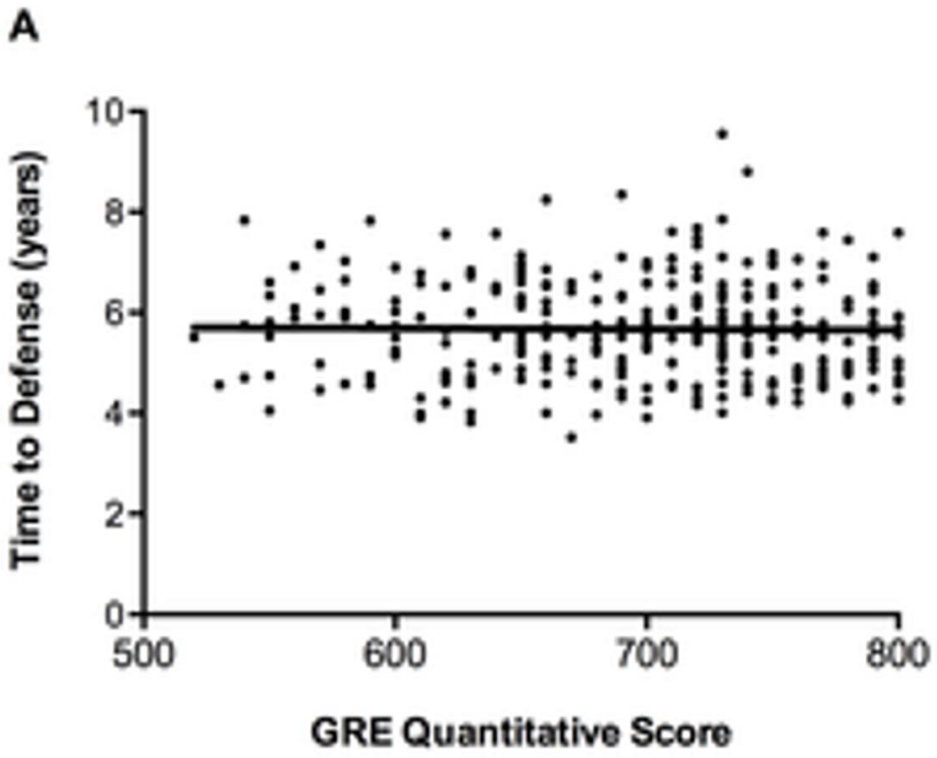


Moneta-Koehler et al, PLOS ONE 12(1):e0166742, 2017

Exercise #9

Is there a relationship betw time-to-defense and GRE scores?

Explain your answer



Moneta-Koehler et al, PLOS ONE 12(1):e0166742, 2017

How to participate?

WEB

1. Connect to www.wooclap.com/EOKTYT
2. You can participate

wooclap

80 %

2


www.wooclap.com/EOKTYT
Is there a relationship between time-to-defense and GRE score...
No: Regression line is flat
Click on the projected screen to start the question
wooclap
80 %
1

Observation

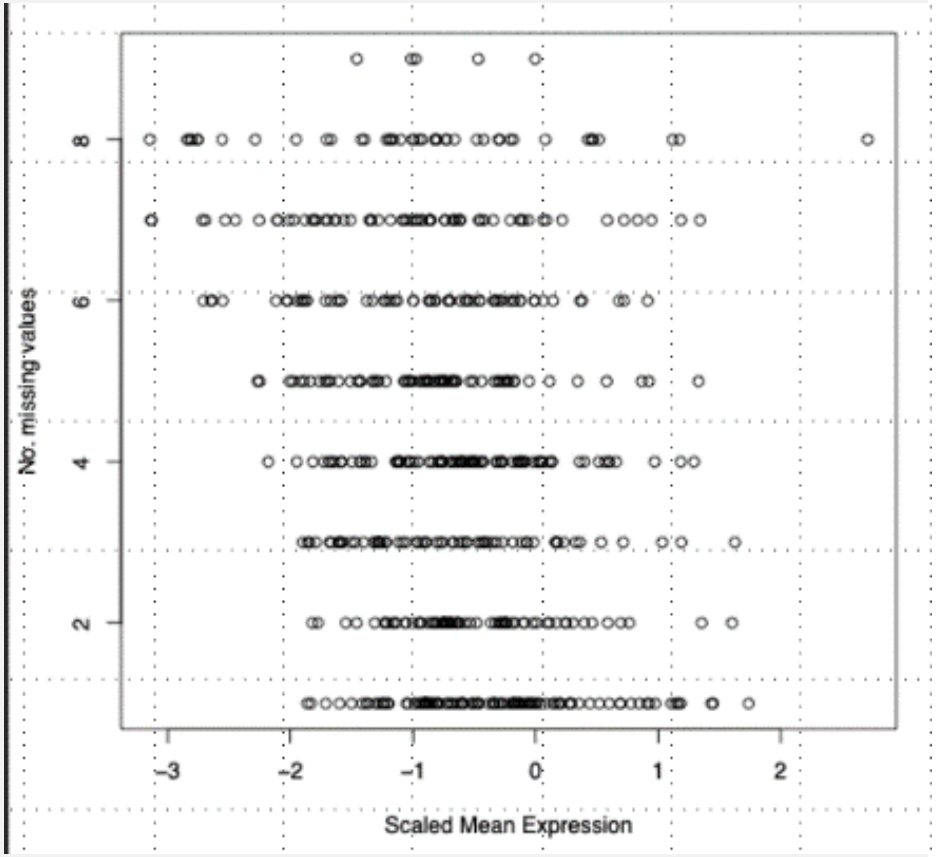
Intentionally left blank

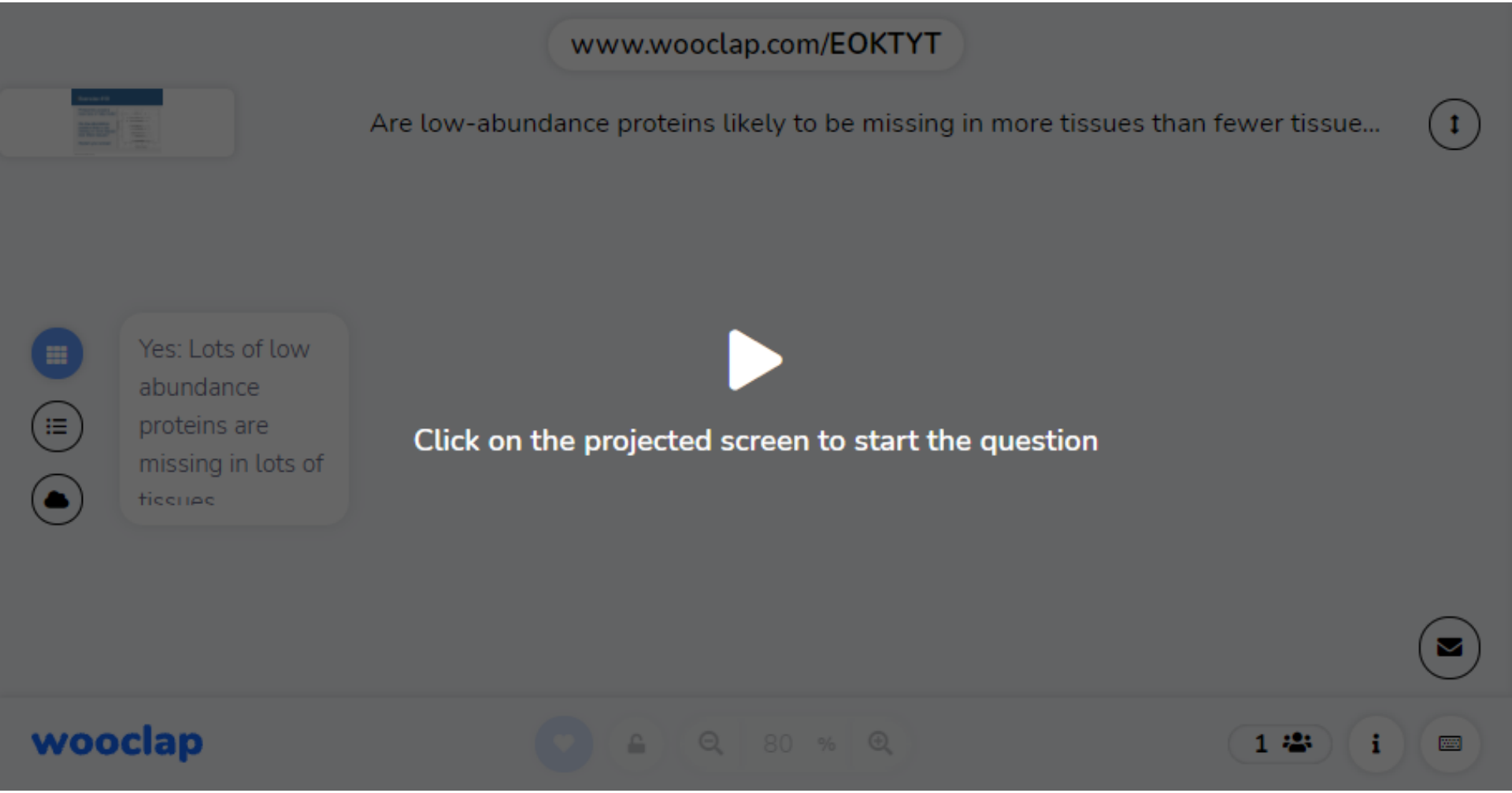
Exercise #10

Proteomics screens have lots of “data holes”

Are low-abundance proteins likely to be missing in more tissues than fewer tissues?

Explain your answer




www.wooclap.com/EOKTYT
Are low-abundance proteins likely to be missing in more tissues than fewer tissue...
Yes: Lots of low abundance proteins are missing in lots of
Click on the projected screen to start the question
wooclap
80 %
1

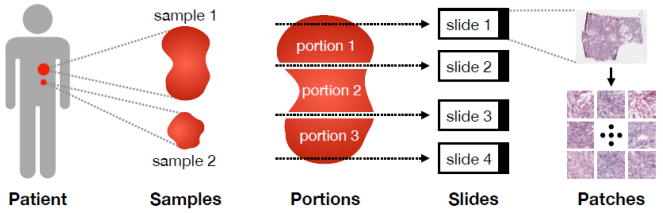
Putting in the median lines

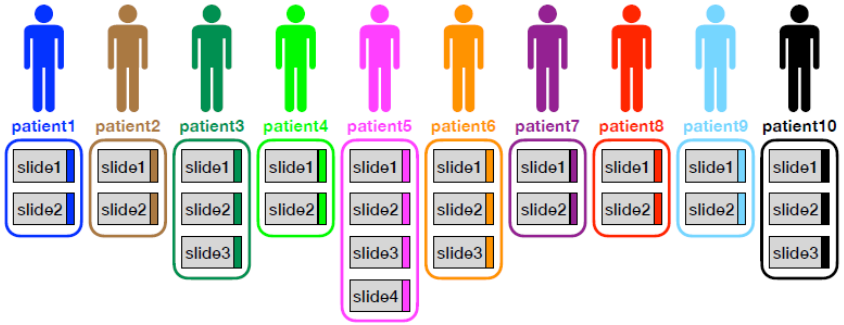
Intentionally left blank

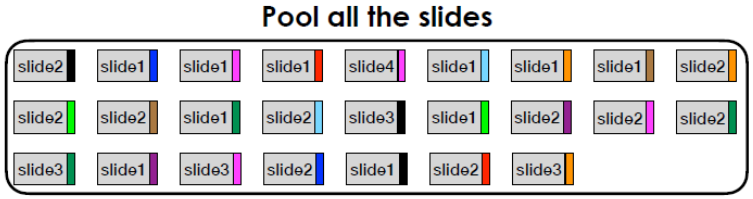
Good result may not be real result

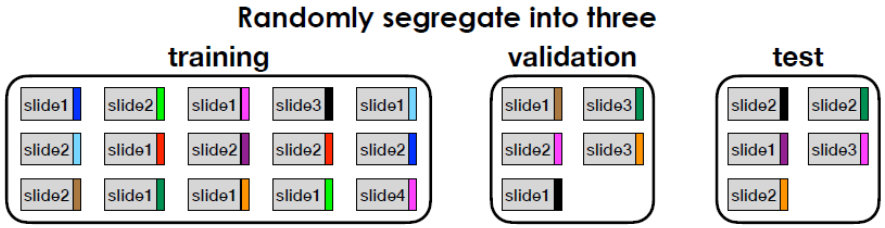
Deep learning from histopath images for Lung cancer diagnosis

Coudrayet al, *Nat Med* 24:1559-1567, 2018









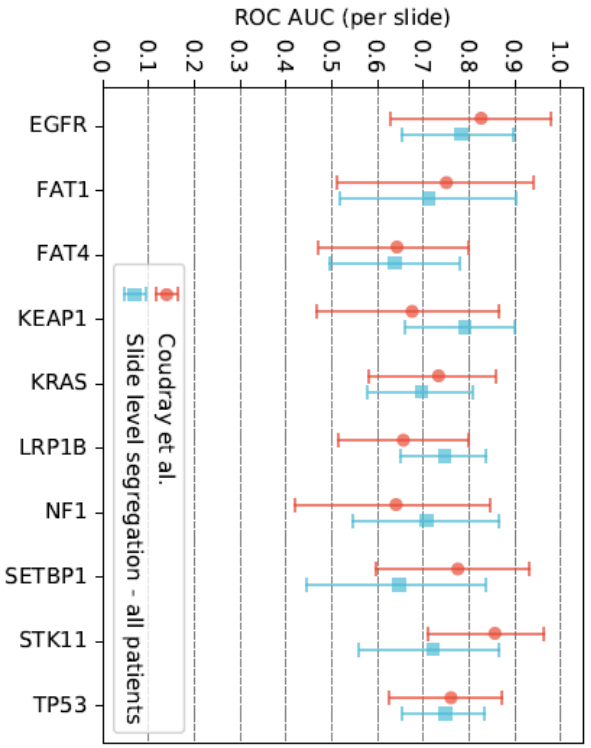


Image credit: Mustafa Umit Oner

Exercise #11

Coudray et al. report exciting results that common mutations in lung cancers can be predicted from histopath images using deep learning

Is this claim sound based purely on their results?

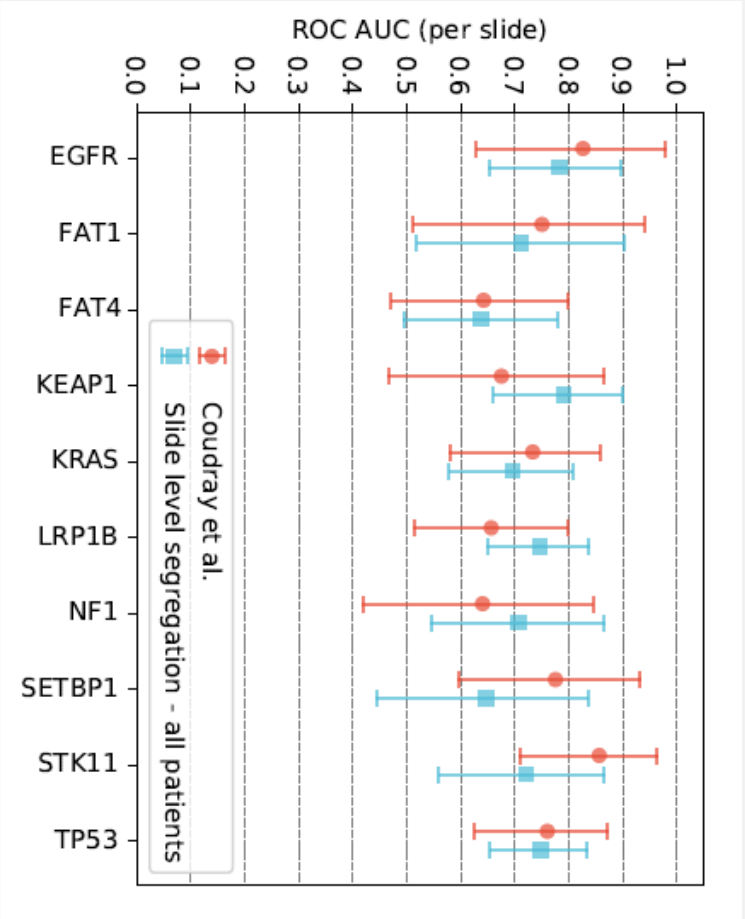


Image credit: Mustafa Umit Oner

Doppelgangers

Intensionally left blank

Protein function prediction

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
```

Seq similarity to known proteins high ⇒ easy

Seq similarity is low (~30%) ⇒ error prone

Seq similarity is very low ⇒ really hard

Exercise #12

DeepFam is a deep learning classifier for predicting the function class of unknown proteins

Will it work well in real deployment?

Dataset	COG-500-1074	COG-250-1796	COG-100-2892
DeepFam	**95.40**	**94.08**	91.40
pHMM	91.75	91.78	**91.67**
3-mer LR	85.59	81.15	75.44
Protvec LR	47.34	41.76	37.05

Bold indicates the best performance for each dataset.

Seo et al. Bioinformatics 34(13):i254-i262, 2018

Out-of-class proteins

Intentionally left blank

Doppelgangers

Intentionally left blank