Gene expression analysis: Some lessons for statistical hypothesis testing

WONG Limsoon

NUS
National University of Singapore
National University of Singapore

© Copyright National University of Singapore. All Rights Reserved.

Plan, 12/1/2026

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis

Statistical hypothesis testing

Formulate null hypothesis H0 and alternate hypothesis H1

Devise a test statistic, t(·)

Evaluate t(S) on a sample S

Compare t(S) to the null distribution

If significant, reject H0; otherwise, reject H1

This does not mean we accept H0!

Null distribution is the distribution of t(S0) where S0 ranges over the set of null samples for which H0 holds

Onus of proof: Reject H1 ≠ accept H0

"… a p-value is large doesn't mean that the null hypothesis is true. All a hypothesis test does is measure the strength of evidence against the null hypothesis. That is, we assume the null hypothesis is true until we have enough evidence to reject. Crucially, we never actually claim that the null hypothesis is true - it is just an assumption!"

A pharmaceutical research team constructs a significance test:

H0 – Side effects of new drug X are same as standard drug Y

H1 - Side effects of new drug X are different from standard drug Y

Would you be happy to use the new drug based on a large statistically insignificant p-value?

Taken from https://vsni.co.uk/blogs/never-accept-the-null-hypothesis

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

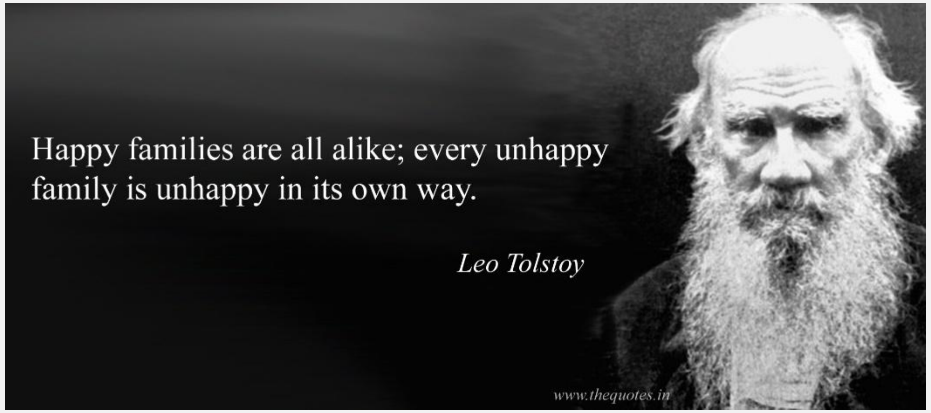
Sample is biased

Null distribution used is inappropriate

Null hypothesis incorrectly stated

Inappropriate expt design

And so on



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

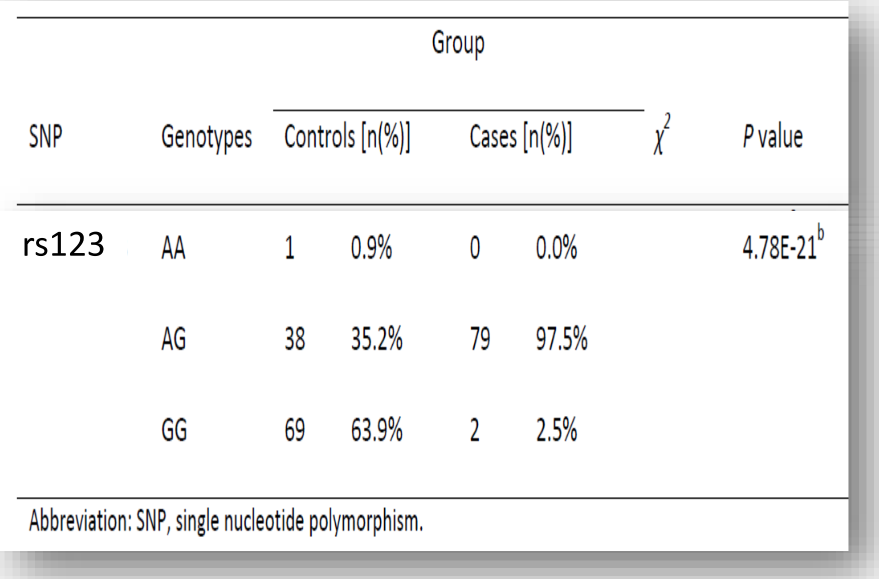
Leo Tolstoy

Exercise

SNP rs123 is good biomarker for a disease:

If rs123 is AA or GG, low risk for disease

If rs123 is AG, high risk for disease

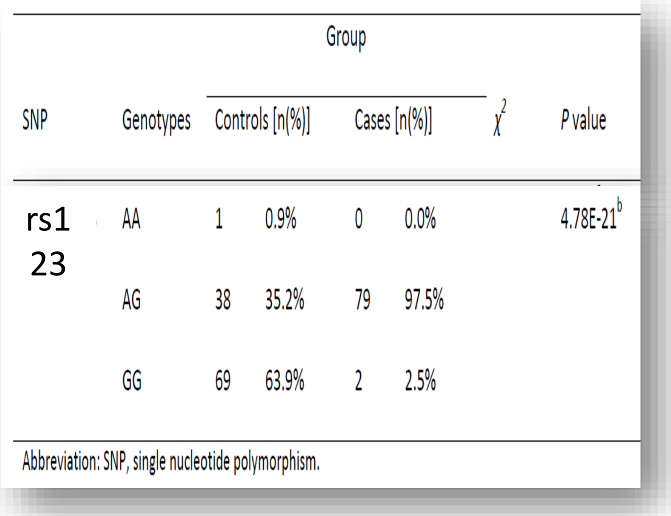


		Group					
SNP	Genotypes	Controls [n(%)]		Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

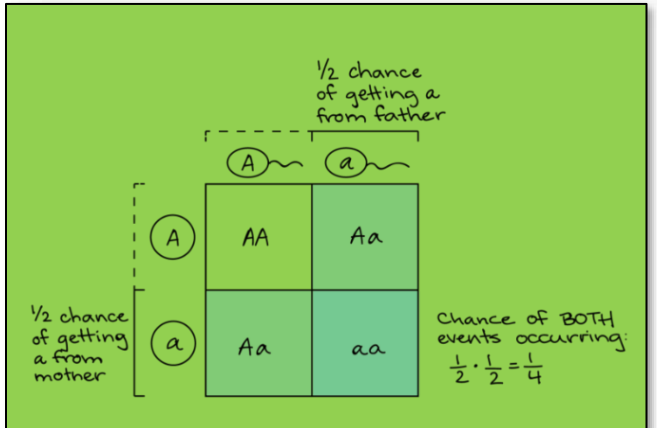
Straightforward $\chi 2$ test. Anything wrong?

Calculations



SNP	Genotypes	Group Controls [n(%)]		Cases [n(%)]		χ^2	P value
rs1 23	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.



½ chance of getting a from father
A
a
A
a
AA
Aa
Aa
aa
½ chance of getting a from mother
Chance of BOTH events occurring: $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Rule of human genetics

AG = 38 + 79 = 117,

Controls + cases = 189

⇒ Population is 117 / 189 = 62% AG

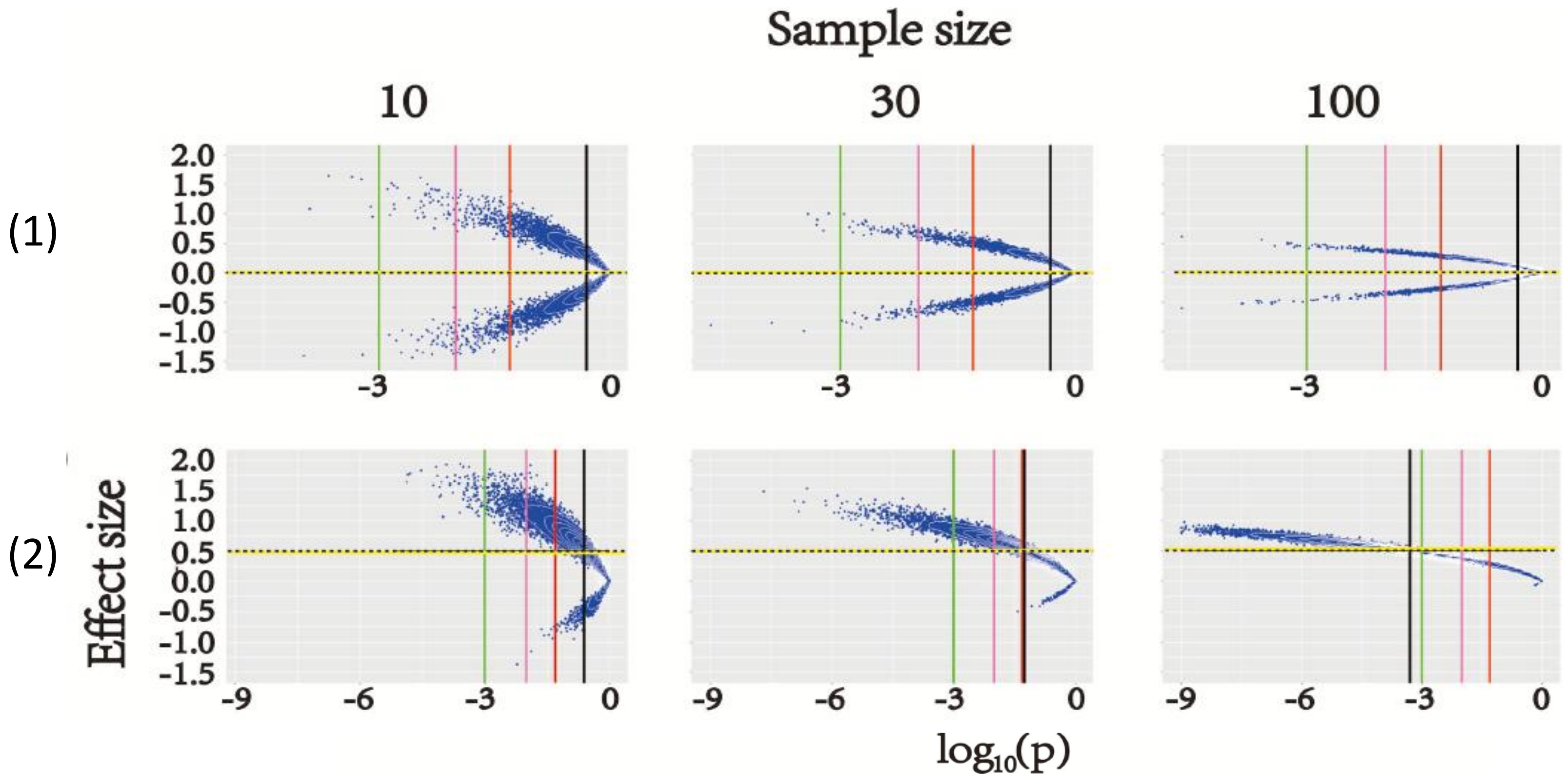
P(AA) = …

Sample may be …

What do you think?

Sampling bias happens often

	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

Careless null hypothesis

H0

rs123 alleles are identically distributed in the two samples

Assumption

Distributions of rs123 alleles in the two samples are resp. identical to the two populations

Apparent H1

rs123 alleles are differently distributed in the two populations

"Actual" H1

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

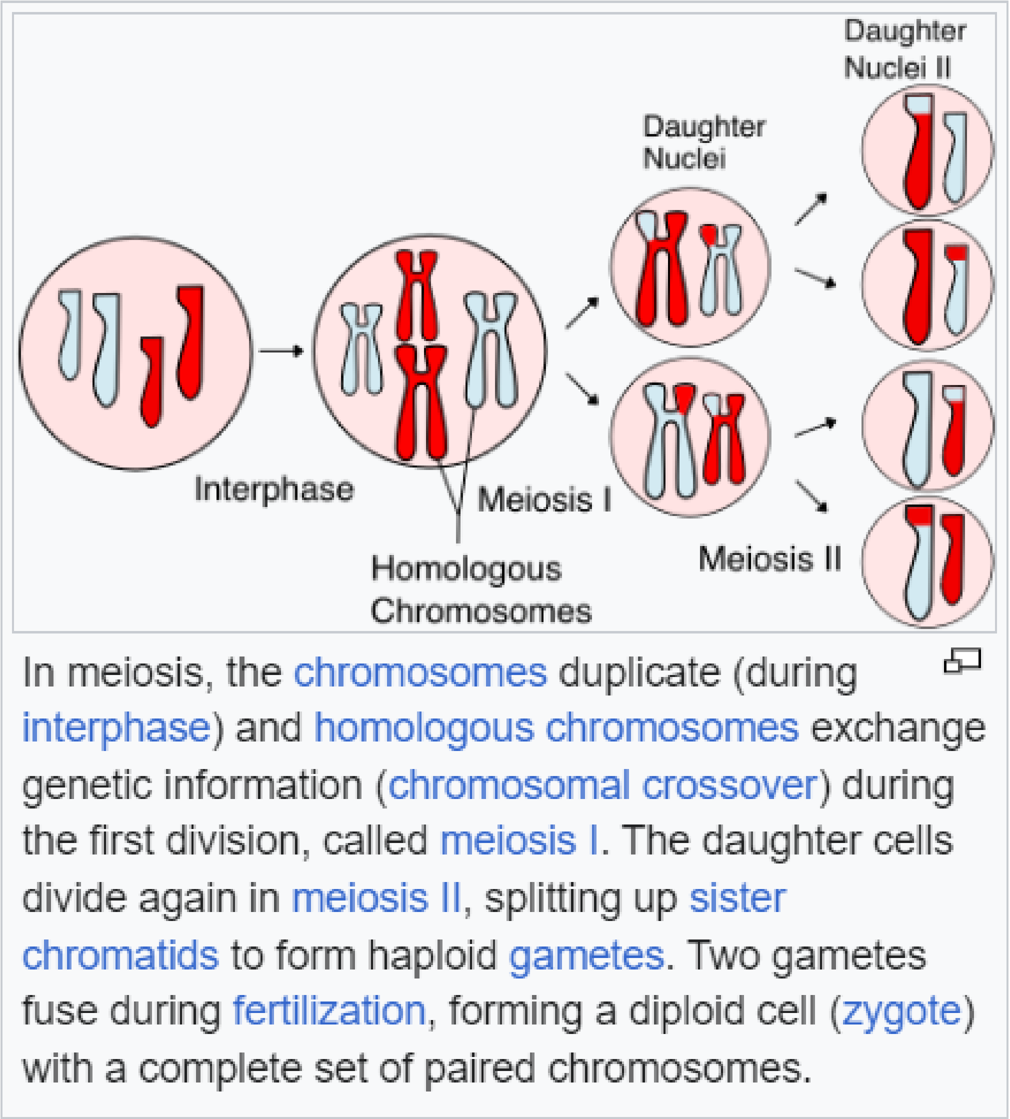
Exercise

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?

Meiosis

Watch this video together, https://youtu.be/BlnUNmfGn7I



In meiosis, the chromosomes duplicate (during interphase) and homologous chromosomes exchange genetic information (chromosomal crossover) during the first division, called meiosis I. The daughter cells divide again in meiosis II, splitting up sister chromatids to form haploid gametes. Two gametes fuse during fertilization, forming a diploid cell (zygote) with a complete set of paired chromosomes.

Image credit: Wikipedia

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

Watch this video together,
https://youtu.be/ZeATszO-6e0

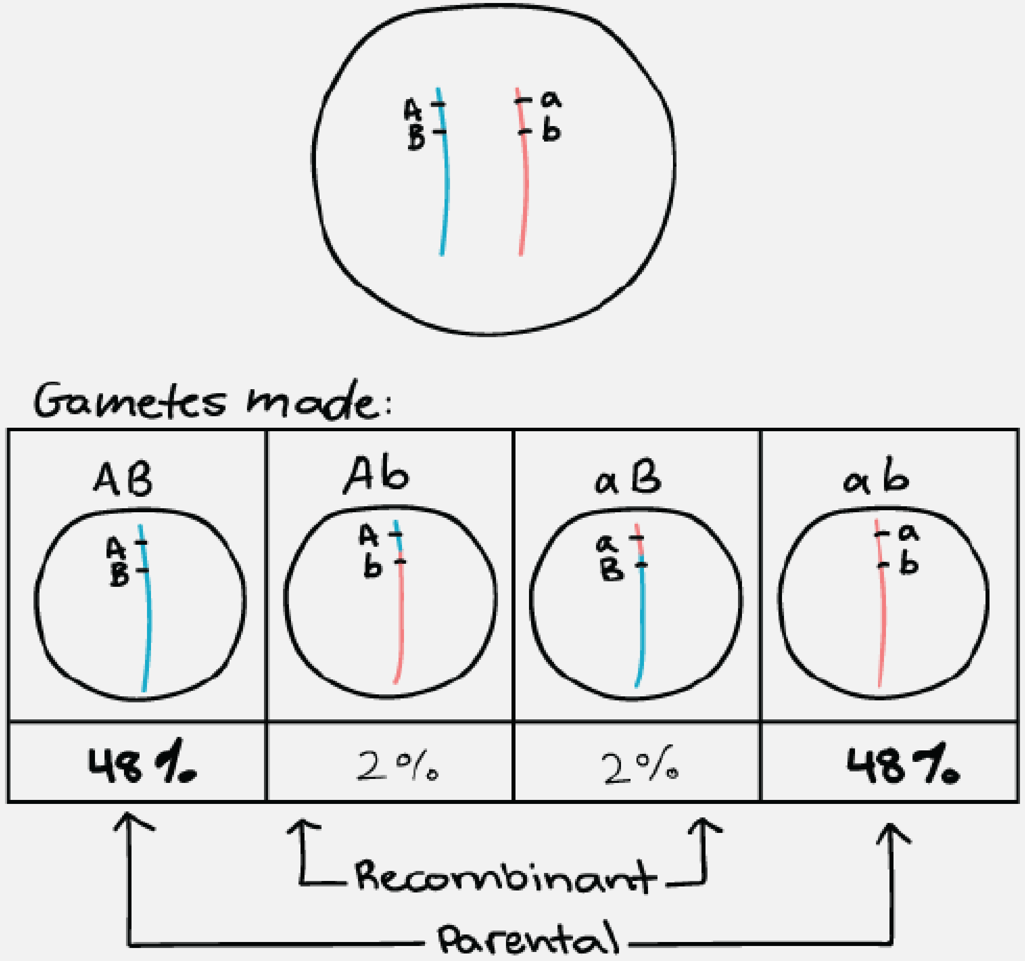
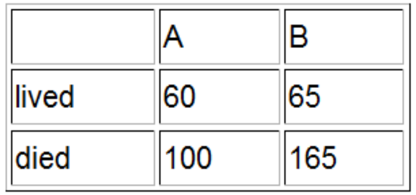


Image credit: Khan Academy

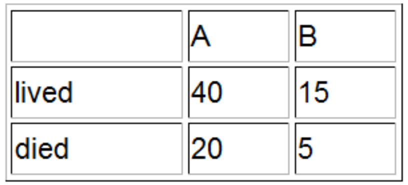
Exercise

Overall



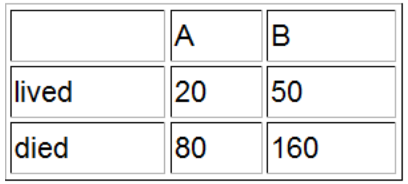
	A	B
lived	60	65
died	100	165

Women



	A	B
lived	40	15
died	20	5

Men



	A	B
lived	20	50
died	80	160

Treatment A is better

Simpson's paradox

Treatment B is better

What is happening here?

A/B sample not …

What do you think?

Careless null hypothesis

H0

Treatment effects are identically distributed in the two samples

Assumption

All other factors are equalized in the two samples

Apparent H1

Treatment effects are differently distributed in the two populations

“Actual” H1

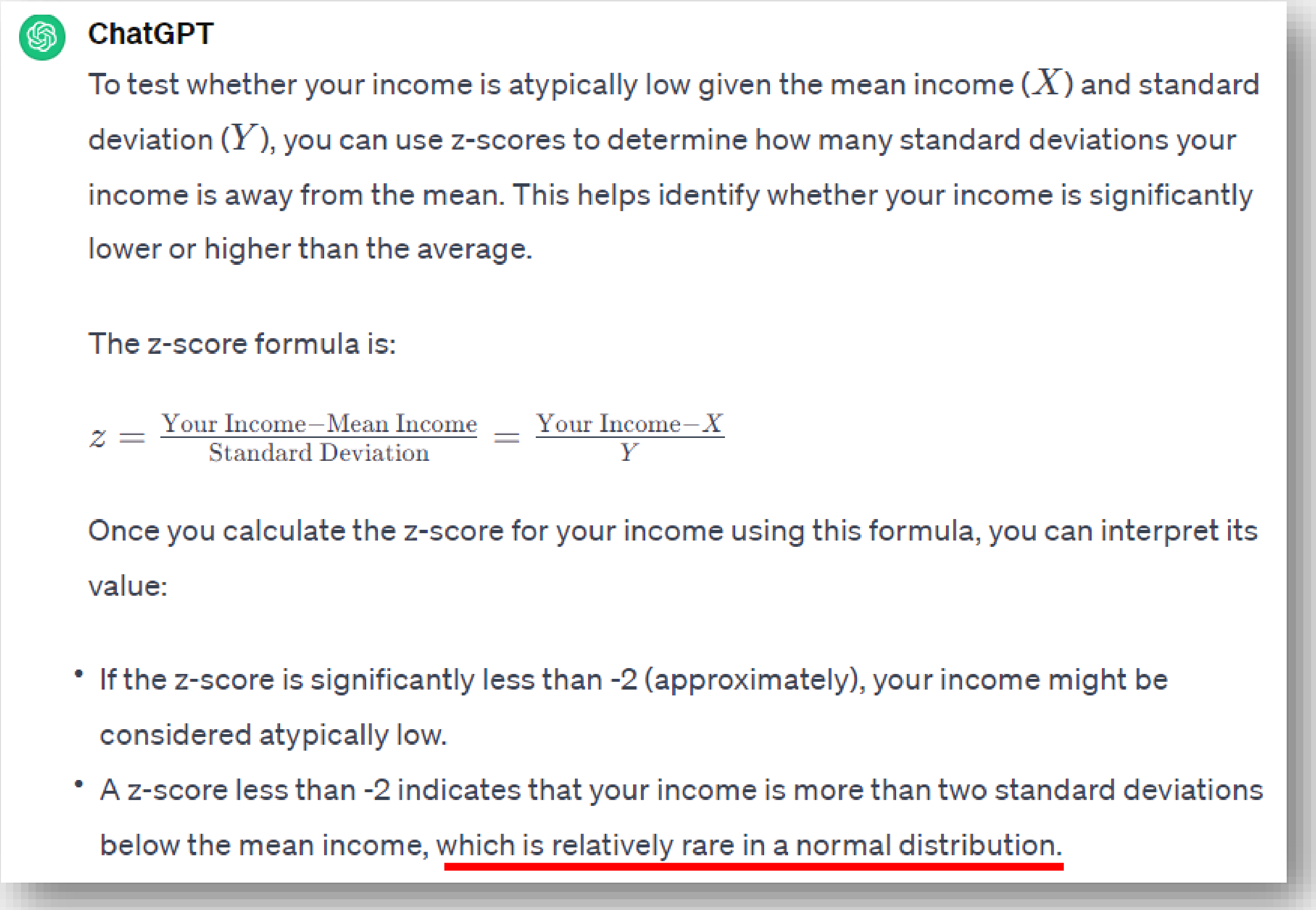
Treatment effects are differently distributed in the two populations OR

Some other factors aren’t equalized in the two samples

Exercise

I asked ChatGPT how to test whether my income is atypically low

What do you think of ChatGPT's response?



ChatGPT

To test whether your income is atypically low given the mean income (X) and standard deviation (Y), you can use z-scores to determine how many standard deviations your income is away from the mean. This helps identify whether your income is significantly lower or higher than the average.

The z-score formula is:

$$z = \frac{\text{Your Income} - \text{Mean Income}}{\text{Standard Deviation}} = \frac{\text{Your Income} - X}{Y}$$

Once you calculate the z-score for your income using this formula, you can interpret its value:

- If the z-score is significantly less than -2 (approximately), your income might be considered atypically low.
- A z-score less than -2 indicates that your income is more than two standard deviations below the mean income, which is relatively rare in a normal distribution.

A nation of unusually poor people, really?

What do you think

Plan, 12/1/2026

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis

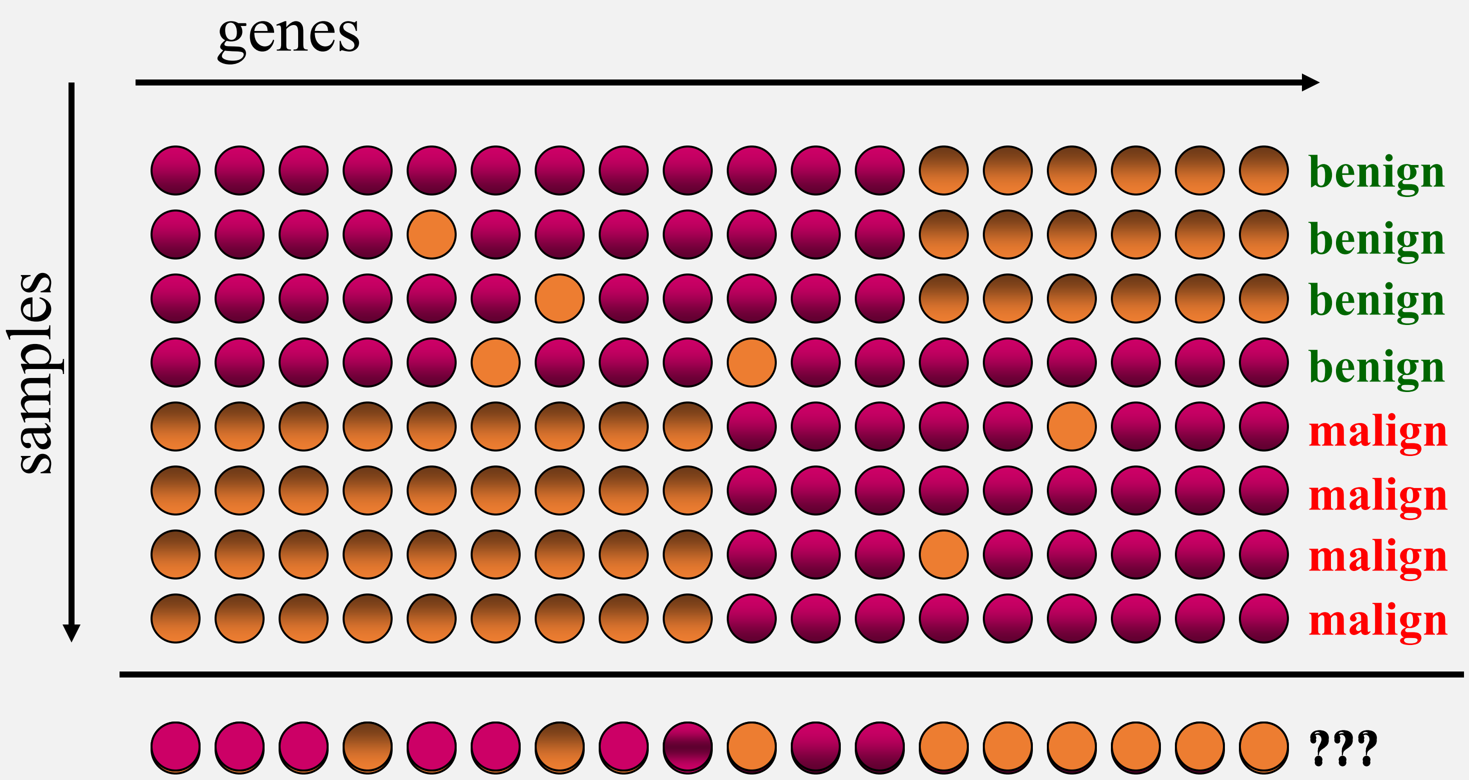
Microarray

Contain large numbers of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers

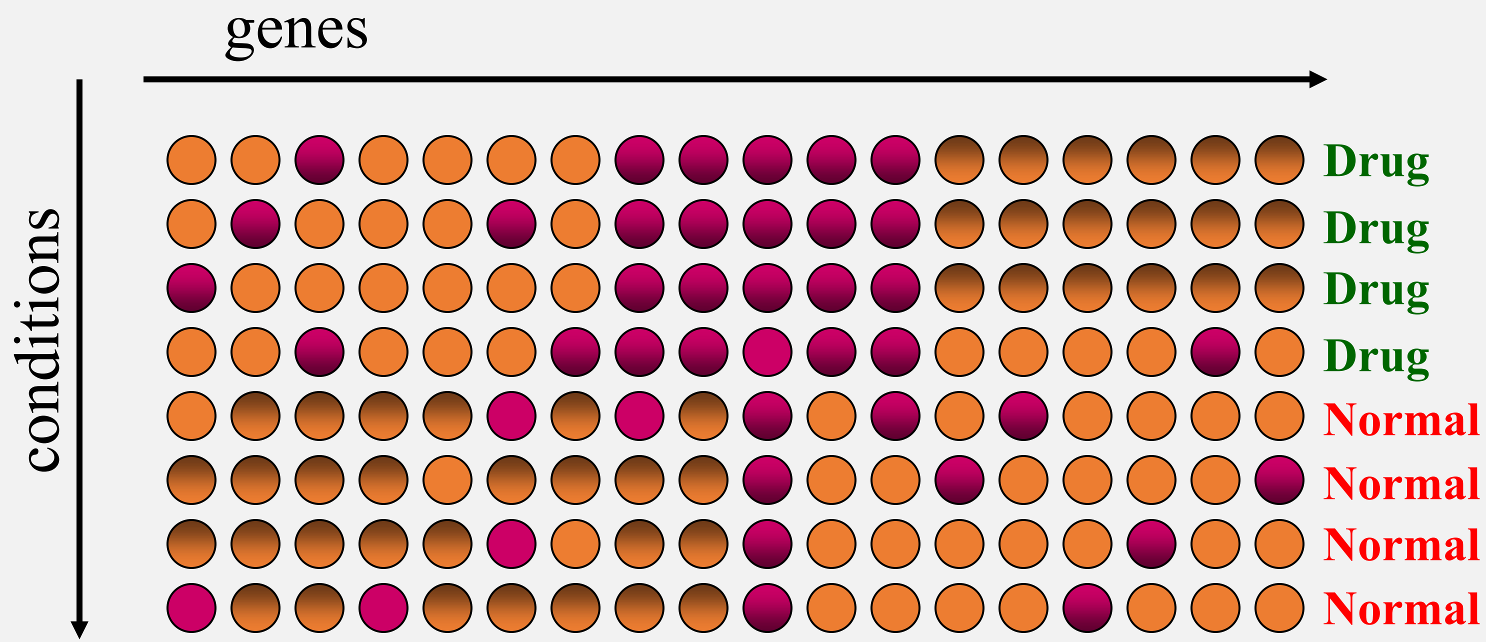
Detect what genes are being expressed or found in a cell of a tissue sample

Measure expression of thousands of genes simultaneously

Application: Disease subtype diagnosis



Application: Drug-action inference



Which group of genes does the drug affect? Why?

Diagnosis using microarrays & machine learning

Gene expression data collection

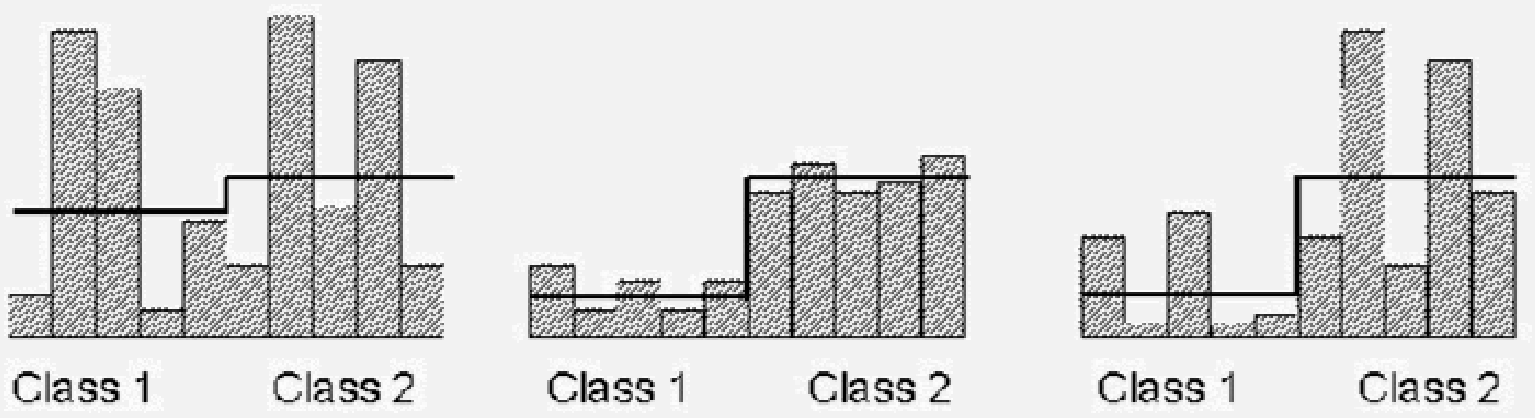
Gene selection using e.g., $\chi2$

Classifier training

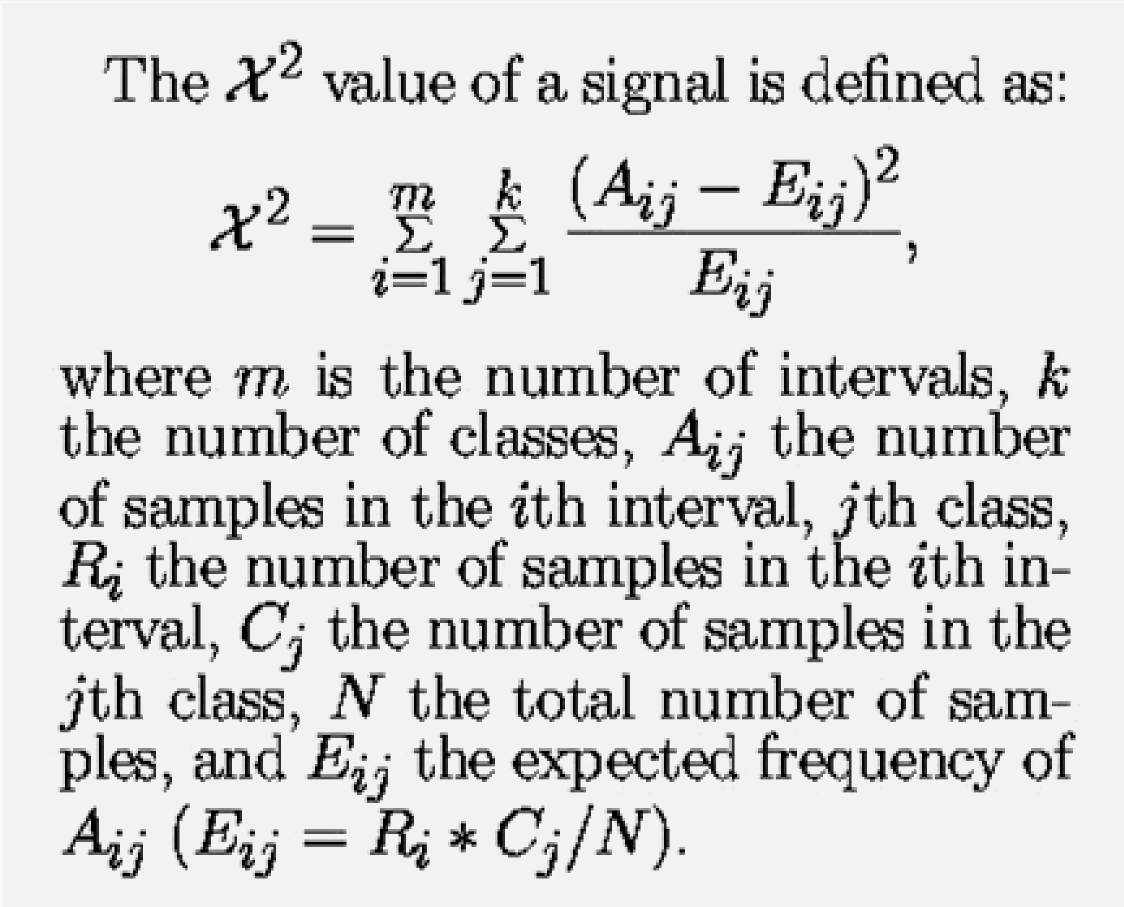
Classifier tuning (optional for some machine learning methods)

Apply classifier for diagnosis of future cases

Gene selection basic idea



Gene selection by $\chi 2$

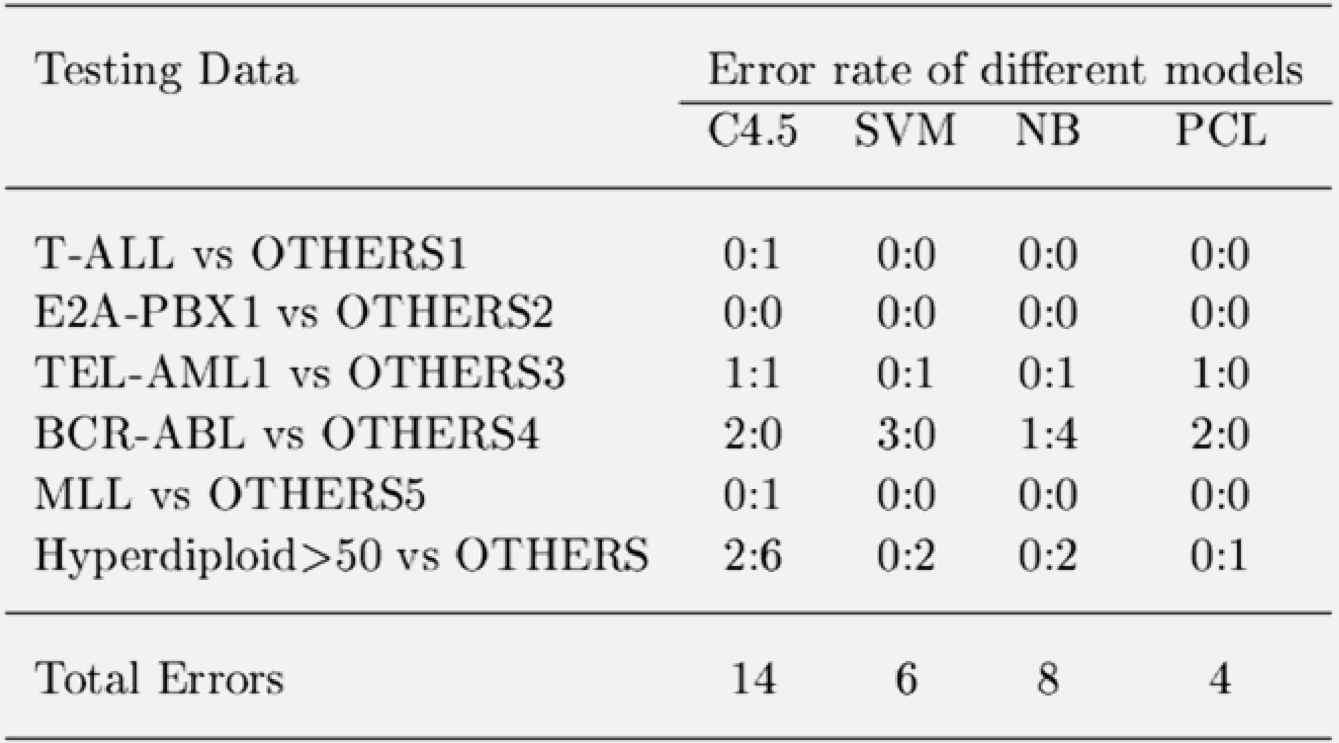


The $\mathcal{X}^2$ value of a signal is defined as:

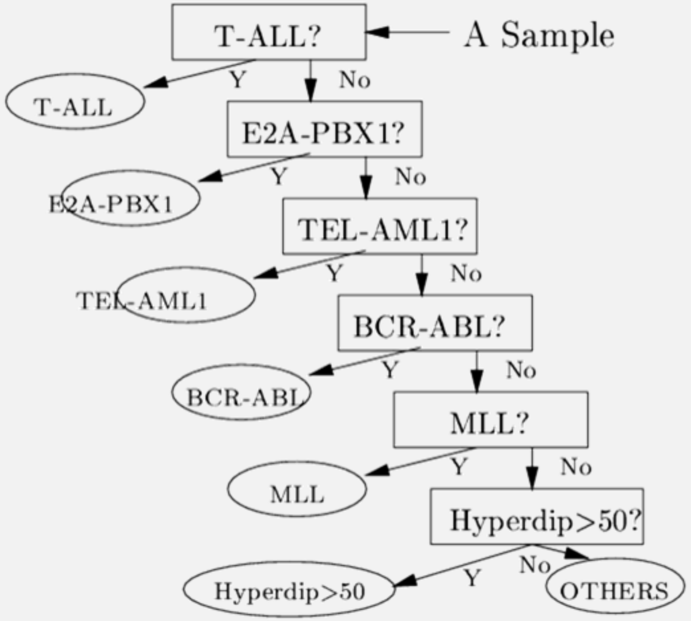
$$\mathcal{X}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} $(E_{ij} = R_i * C_j / N)$.

Performance of various classifiers

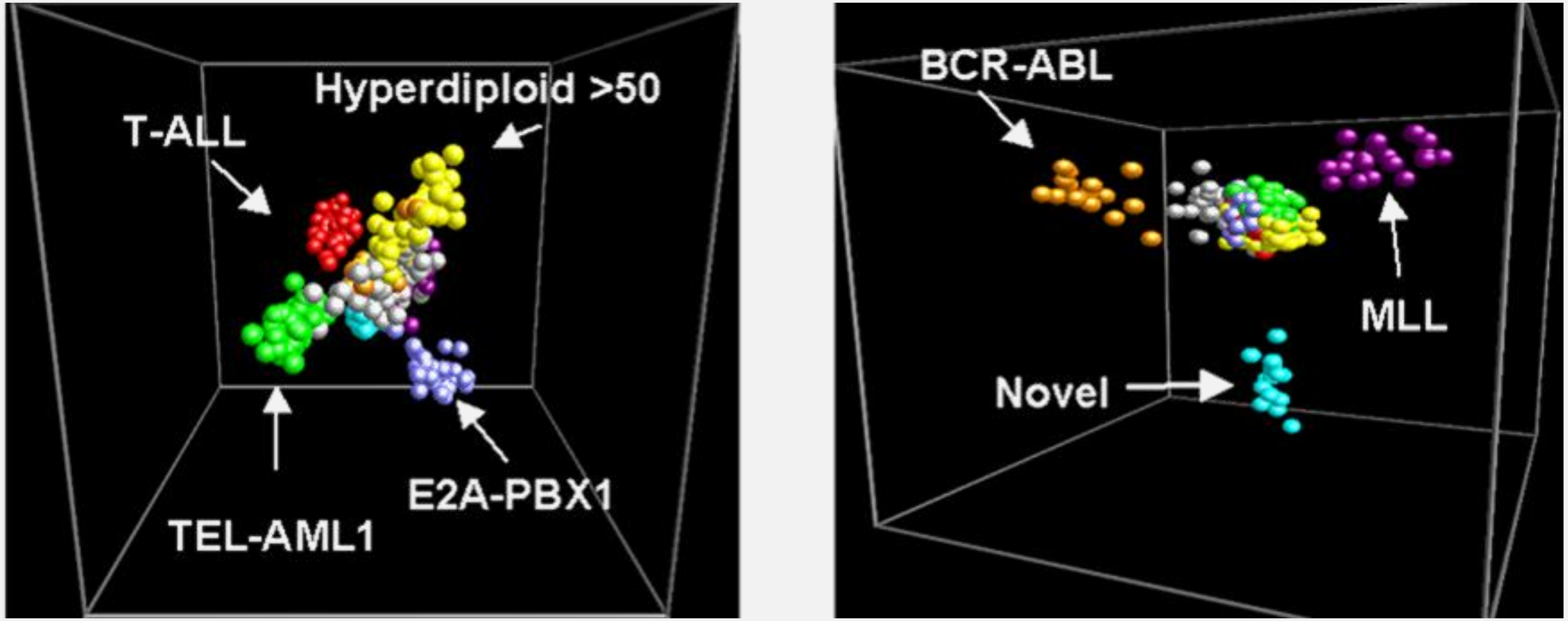


Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4



Classifiers based 20 genes selected by $\chi2$ at each level of the tree

Multidimensional scaling plot for subtype diagnosis



Obtained by performing PCA on the 20 genes chosen for each level

Plan, 12/1/2026

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis

Poor replicability of gene selection

Low % of overlapping genes from diff expt

Prostate cancer

- Lapointe et al, 2004 vs Singh et al, 2002

Lung cancer

- Garber et al, 2001 vs Bhattacharjee et al, 2001

DMD

- Haslett et al, 2002 vs 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

Individual genes

Suppose:

Each gene has 50% chance to be high

You have 3 disease and 3 normal samples

How many genes on a microarray are expected to perfectly correlate to these samples?

Prob(gene is correlated) = $1/2^6$

# of genes on array = 25,000

E(# of correlated genes) = 390

$\Rightarrow$ Many false positives; these cannot be eliminated based on pure statistics!

Group of genes

Suppose:

Each gene has 50% chance to be high

You have 3 disease & 3 normal samples

What's the chance for a group of 5 genes to perfectly correlate to these samples?

When only 1 group is considered, $(1/2^6)^5 << 1/2^6$

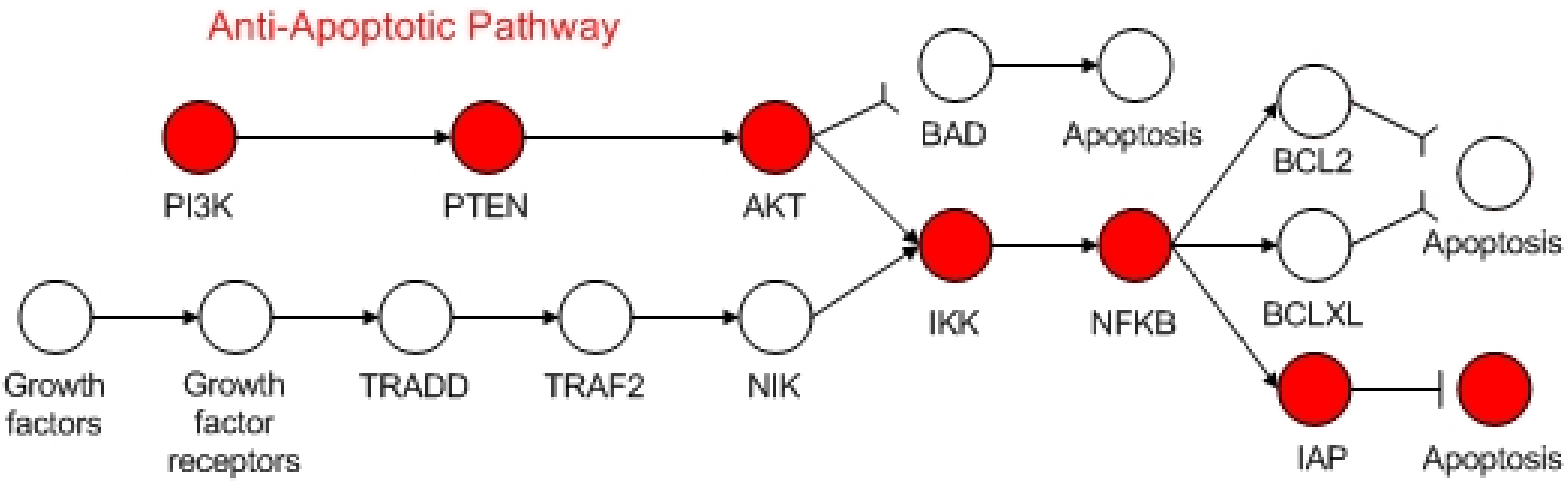
# of groups = ${}^{25000}C_5$

E(# of correlated groups) = ${}^{25000}C_5 * (1/2^6)^5$

$= 7.58 * 10^{10}$

$\Rightarrow$ Even more false positives?

Perhaps no need to consider every group



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 reduced using biological processes

The unifying biological theme is basis for inferring underlying cause of disease subtype

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# ~~of groups = $^{25000}C_5$~~**

⇒ E(~~# of groups of genes correlated) = $^{25000}C_5$ * $(1/2^6)^5$ = $7.58*10^{10}$~~

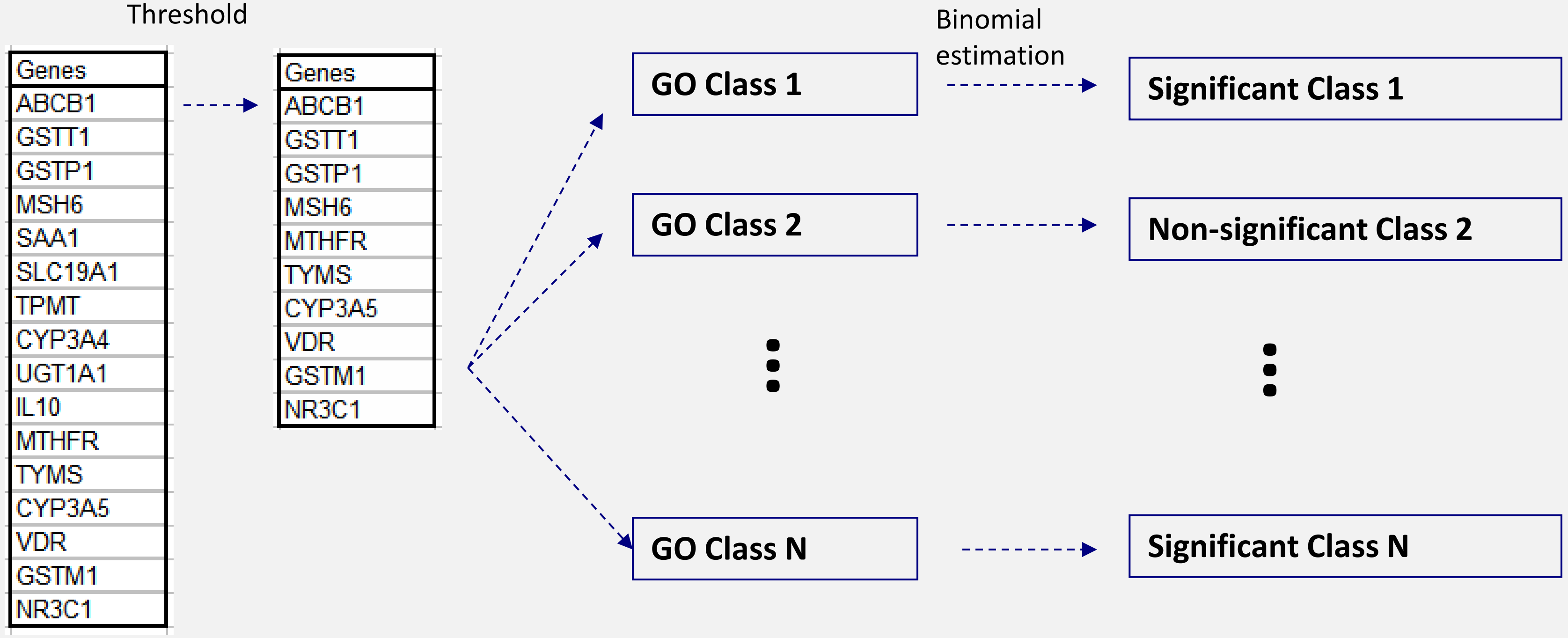
⇒ Even more false positives?

- **Perhaps no need to consider every group**

# of pathways = 1000

E(# of pathways correlated) = 1000 * $(1/2^6)^5$ = $9.3*10^{-7}$

Pathway overlap analysis via Onto-Express (aka ORA)



Threshold

Genes
ABCB1
GSTT1
GSTP1
MSH6
SAA1
SLC19A1
TPMT
CYP3A4
UGT1A1
IL10
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

Genes
ABCB1
GSTT1
GSTP1
MSH6
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

GO Class 1

GO Class 2

⋮

GO Class N

Binomial estimation

Significant Class 1

Non-significant Class 2

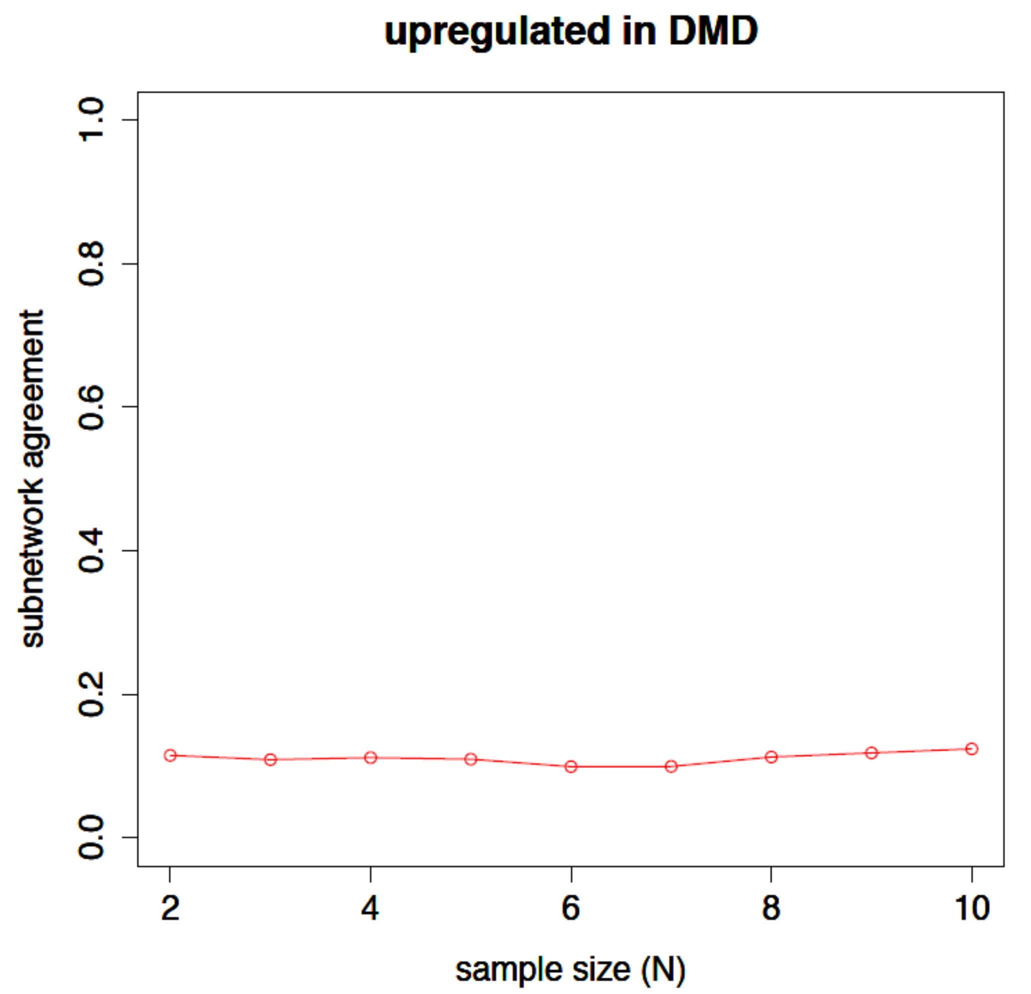
⋮

Significant Class N

ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (e.g., genes whose t-statistic meets the 5% significance threshold of t-test), and checking the significance of the size of the intersection using the hypergeometric test

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

The bewilderment persists… a crisis?



DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

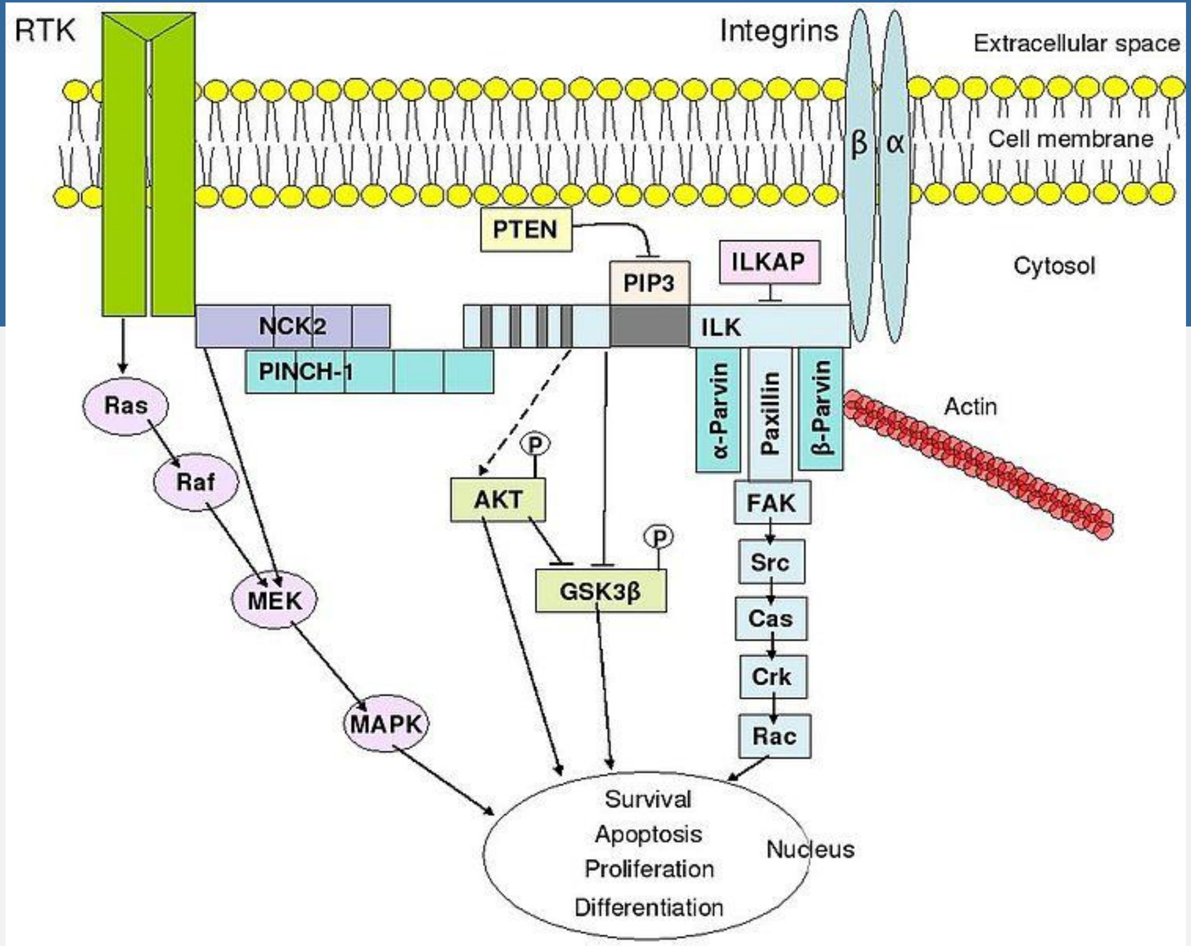
Pathway data
- PathwayAPI, Soh et al., 2010

Exercise

Why does ORA perform so poorly in selecting differentially expressed genes?

Issue #1 with ORA

What do you think?

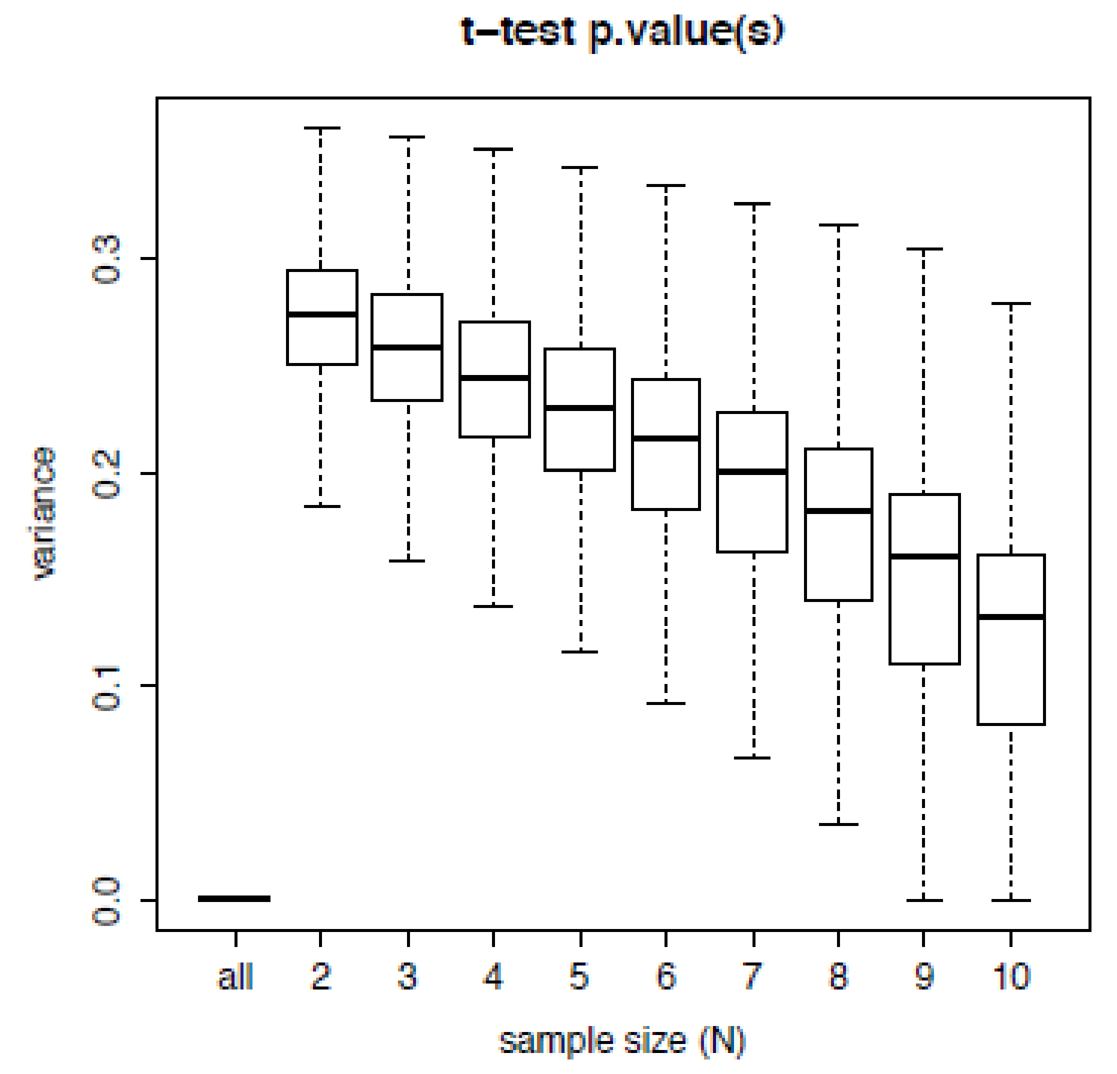


A biological pathway is a chain of actions of molecules in cell leading to a change in cell

⇒ Behavour of genes in a pathway is more coordinated than random ones

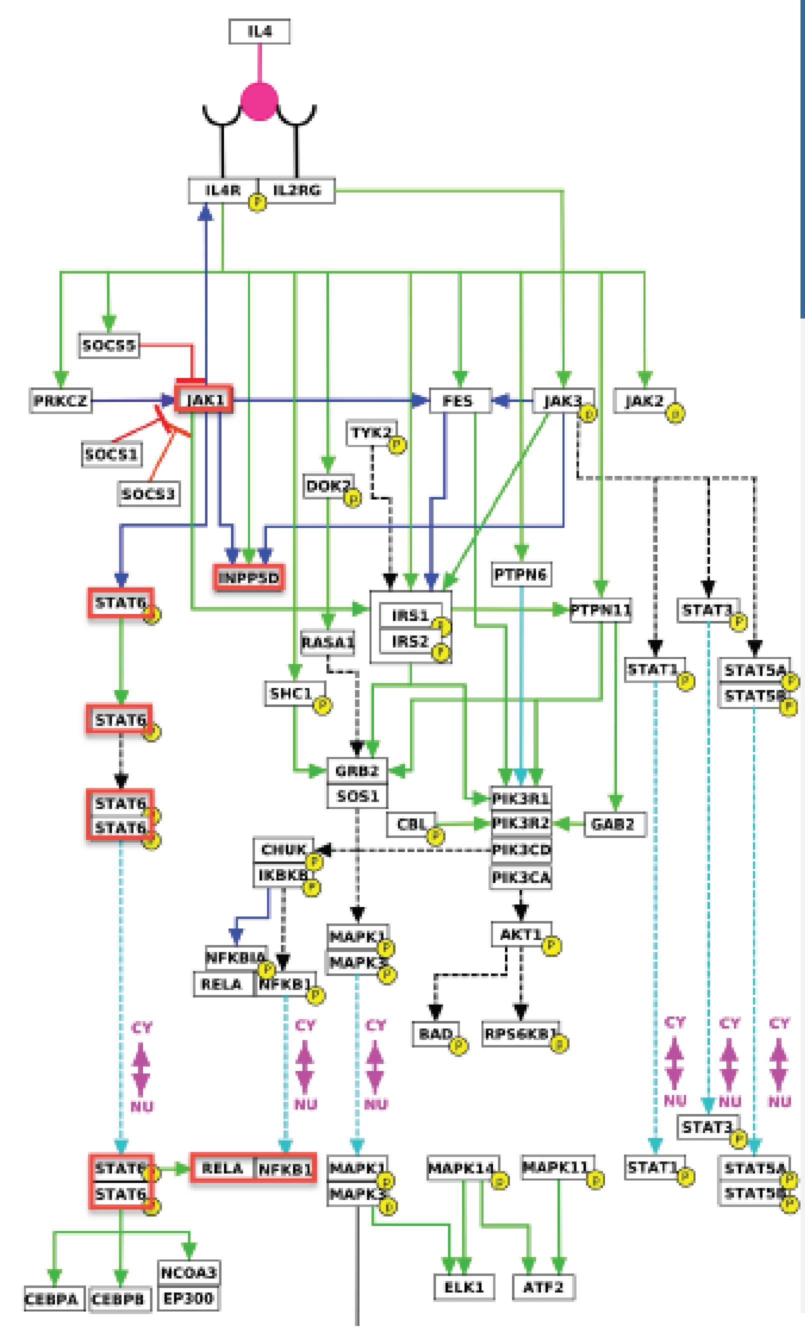
Issue #2 with ORA

What do you think?



Issue #3 with ORA

What do you think?



Exercise

How to solve the issues identified in ORA?

Plan, 26/1/2026

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis

Take-home messages

Statistical hypothesis testing needs careful thought

Right null hypothesis

Right null distribution

Many nuances wrt confounding factors

Make effort to understand the domain

A little domain insight goes a really long way

References

Draghici et al., "Global functional profiling of gene expression", *Genomics,* 81(2):98-104, 2003

Lim et al., “A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small”, *Journal of Bioinformatics and Computational Biology*, 13(4):1550018, 2015

Homework, due 17/1/2026

submit by email to dcswls@nus.edu.sg

Read this paper, focusing on how it proposed to test synthetic-lethal gene pairs:

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

Read the review report (attached below) produced by ChatGPT discuss whether the proposed test is a good one.

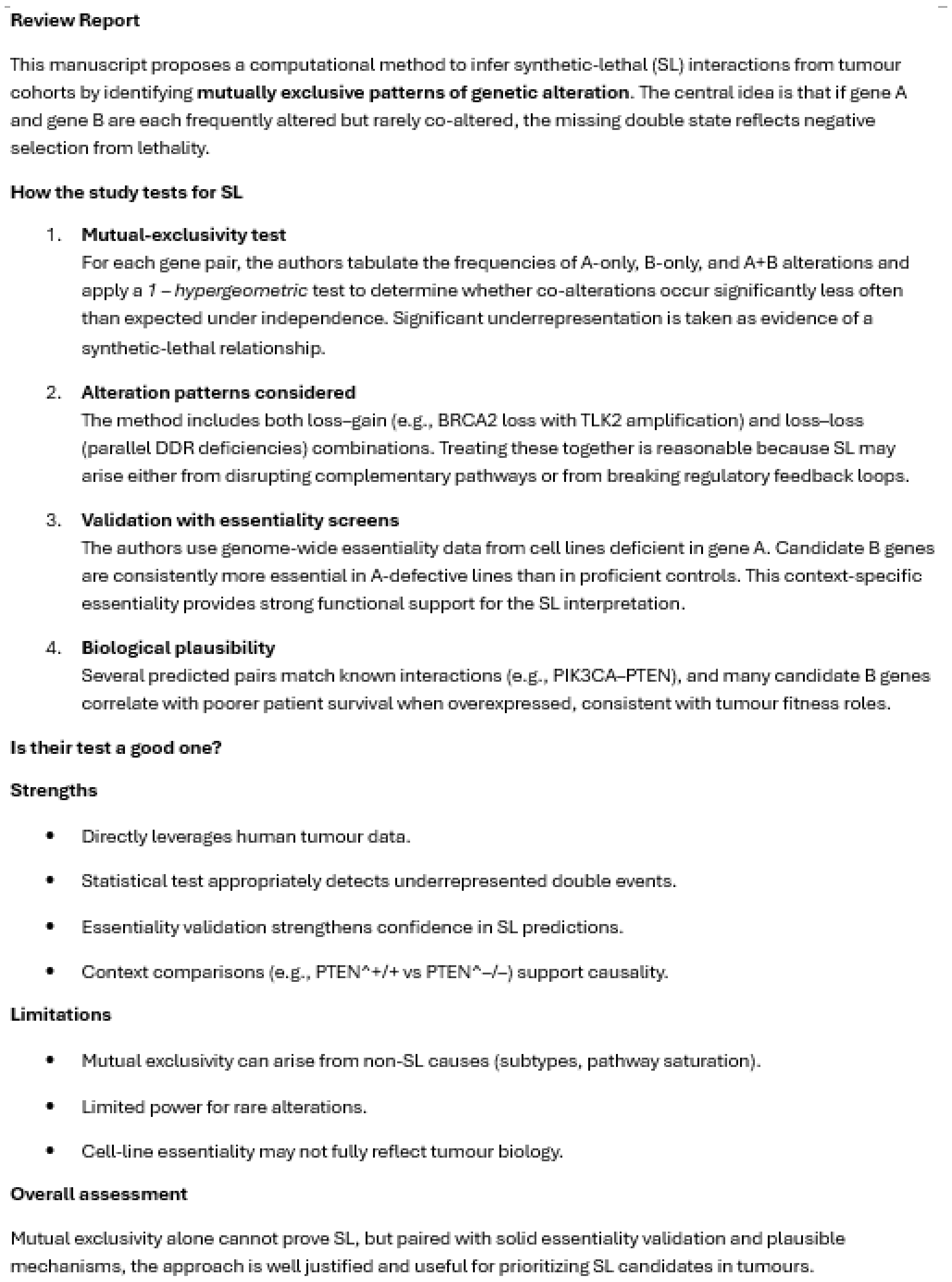
Answer the following questions:

1/ Which parts of the ChatGPT report, if any, do you agree with?

2/ Which parts of the ChatGPT report, if any, do you disagree with? Why?

3/ Are there any major flaws that you think should be pointed out but are missed by ChatGPT?

Make 5-minutes presentation to class on 19/1/2026



Review Report

This manuscript proposes a computational method to infer synthetic-lethal (SL) interactions from tumour cohorts by identifying **mutually exclusive patterns of genetic alteration**. The central idea is that if gene A and gene B are each frequently altered but rarely co-altered, the missing double state reflects negative selection from lethality.

How the study tests for SL

1. **Mutual-exclusivity test**
   For each gene pair, the authors tabulate the frequencies of A-only, B-only, and A+B alterations and apply a *1 – hypergeometric* test to determine whether co-alterations occur significantly less often than expected under independence. Significant underrepresentation is taken as evidence of a synthetic-lethal relationship.
2. **Alteration patterns considered**
   The method includes both loss–gain (e.g., BRCA2 loss with TLK2 amplification) and loss–loss (parallel DDR deficiencies) combinations. Treating these together is reasonable because SL may arise either from disrupting complementary pathways or from breaking regulatory feedback loops.
3. **Validation with essentiality screens**
   The authors use genome-wide essentiality data from cell lines deficient in gene A. Candidate B genes are consistently more essential in A-defective lines than in proficient controls. This context-specific essentiality provides strong functional support for the SL interpretation.
4. **Biological plausibility**
   Several predicted pairs match known interactions (e.g., PIK3CA–PTEN), and many candidate B genes correlate with poorer patient survival when overexpressed, consistent with tumour fitness roles.

Is their test a good one?

Strengths

- Directly leverages human tumour data.
- Statistical test appropriately detects underrepresented double events.
- Essentiality validation strengthens confidence in SL predictions.
- Context comparisons (e.g., PTEN^+/+ vs PTEN^–/–) support causality.

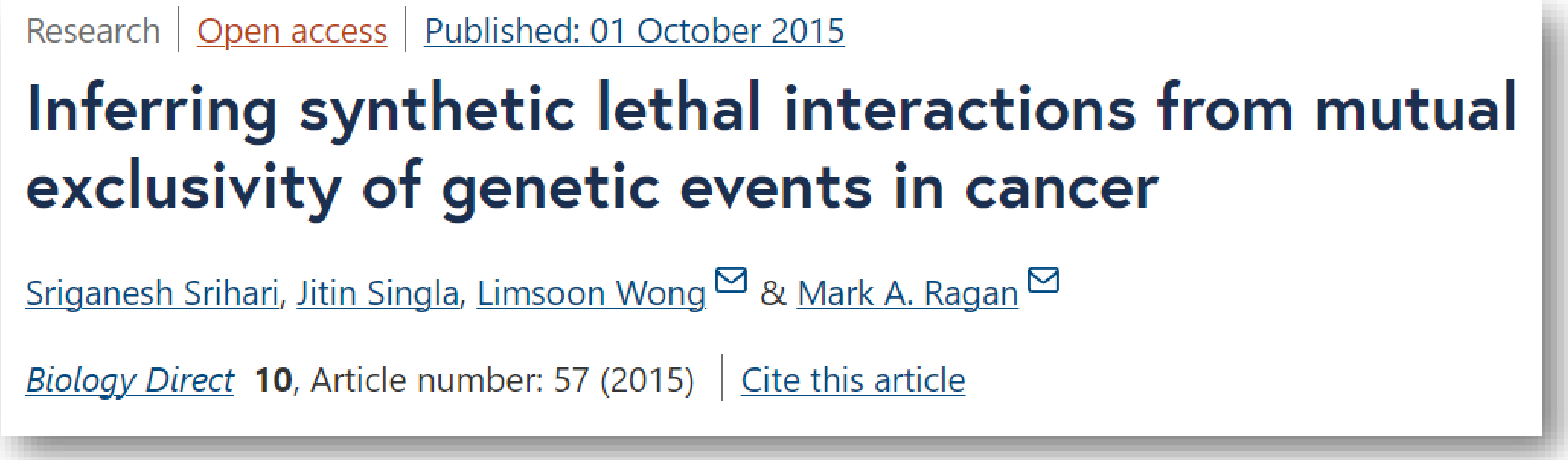
Limitations

- Mutual exclusivity can arise from non-SL causes (subtypes, pathway saturation).
- Limited power for rare alterations.
- Cell-line essentiality may not fully reflect tumour biology.

Overall assessment

Mutual exclusivity alone cannot prove SL, but paired with solid essentiality validation and plausible mechanisms, the approach is well justified and useful for prioritizing SL candidates in tumours.

Presentation & discussion on …



Research | Open access | Published: 01 October 2015

Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer

Sriganesh Srihari, Jitin Singla, Limsoon Wong ✉ & Mark A. Ragan ✉

Biology Direct **10**, Article number: 57 (2015) | Cite this article