

# Gene expression analysis: Some lessons for statistical hypothesis testing

WONG Limsoon



National University of Singapore

# Plan, 20/1/2025

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  $H_0$  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, reject  $H_0$ ; otherwise, reject  $H_1$

This does not mean  
we accept  $H_0$ !

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

# Onus of proof: Rejecting $H_1 \neq$ accepting $H_0$

“... 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 significant test:

*$H_0$  – Side effects of new drug  $X$  are same as standard drug  $Y$*

*$H_1$  - 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?

# 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

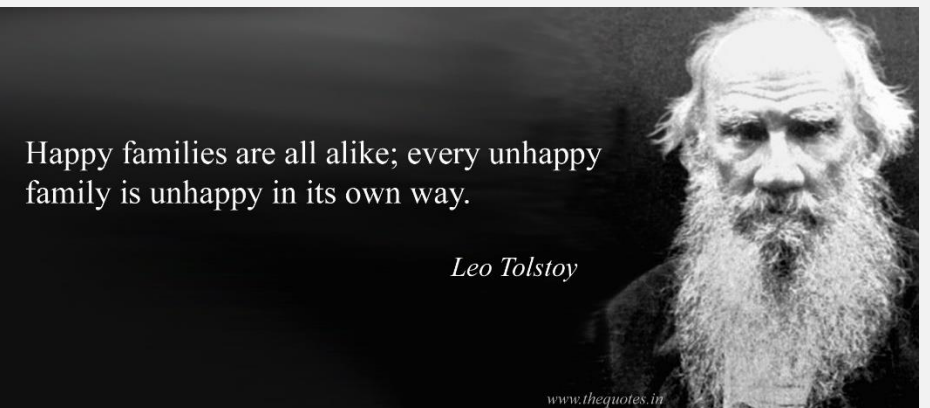
*Sample is biased*

*Null distribution used is inappropriate*

*Null hypothesis incorrectly stated*

*Inappropriate expt design*

*And so on*



# 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*

Straightforward  $\chi^2$  test. Anything wrong?

SNP	Genotypes	Group				$\chi^2$	P value
		Controls [n(%)]		Cases [n(%)]			
rs123	AA	1	0.9%	0	0.0%	4.78E-21 <sup>b</sup>	
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		
Abbreviation: SNP, single nucleotide polymorphism.							

# Calculations

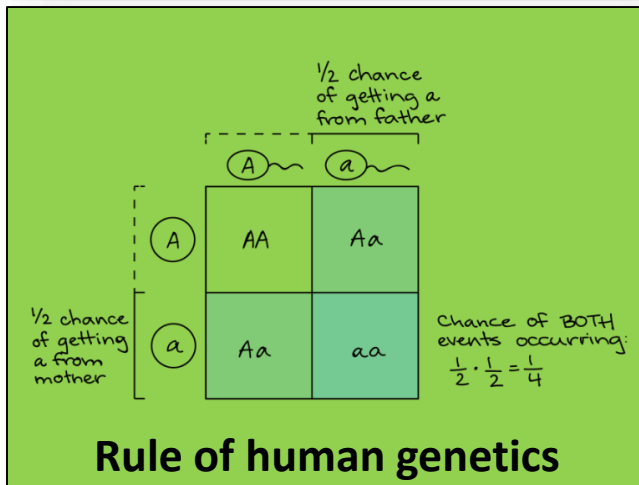
SNP	Genotypes	Group				$\chi^2$	P value
		Controls [n(%)]		Cases [n(%)]			
rs123	AA	1	0.9%	0	0.0%	4.78E-21 <sup>b</sup>	
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

$$AG = 38 + 79 = 117,$$

$$\text{Controls} + \text{cases} = 189$$

$$\Rightarrow \text{Population is } 117 / 189 = 62\% \text{ AG}$$



$$P(AA) = \dots$$

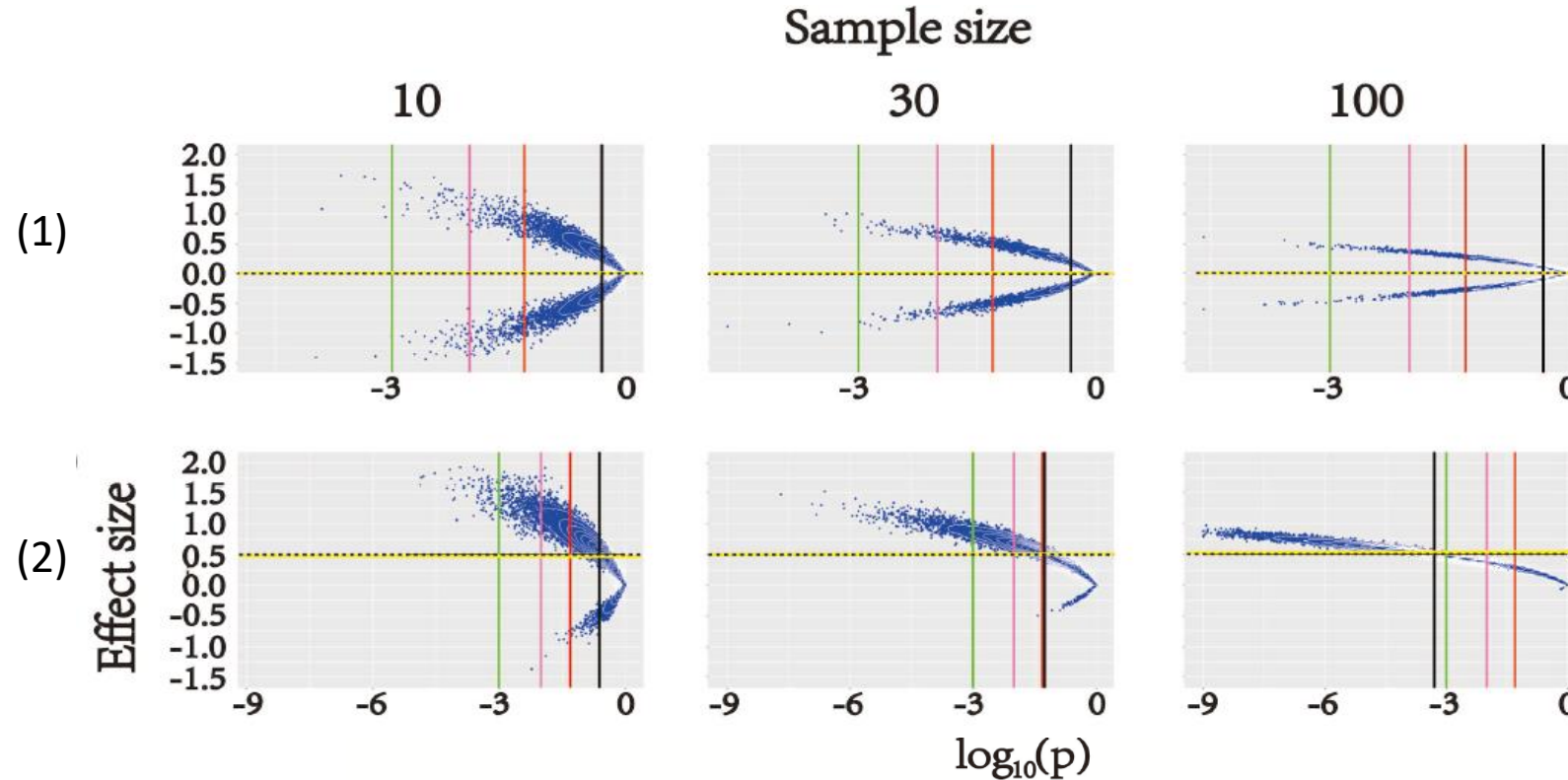
**Sample may be  
biased**

Intentionally left blank



# Sampling bias happens often

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



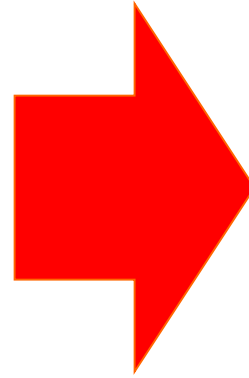
# Careless null hypothesis

## 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

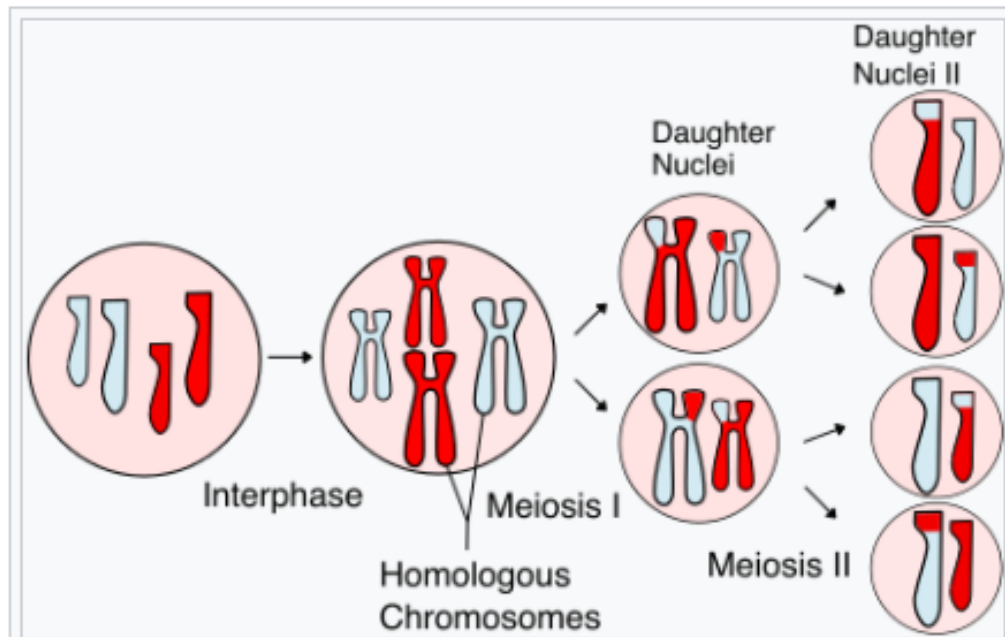
Distribution of rs123 alleles in the two samples are not identical to the two populations

# 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



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

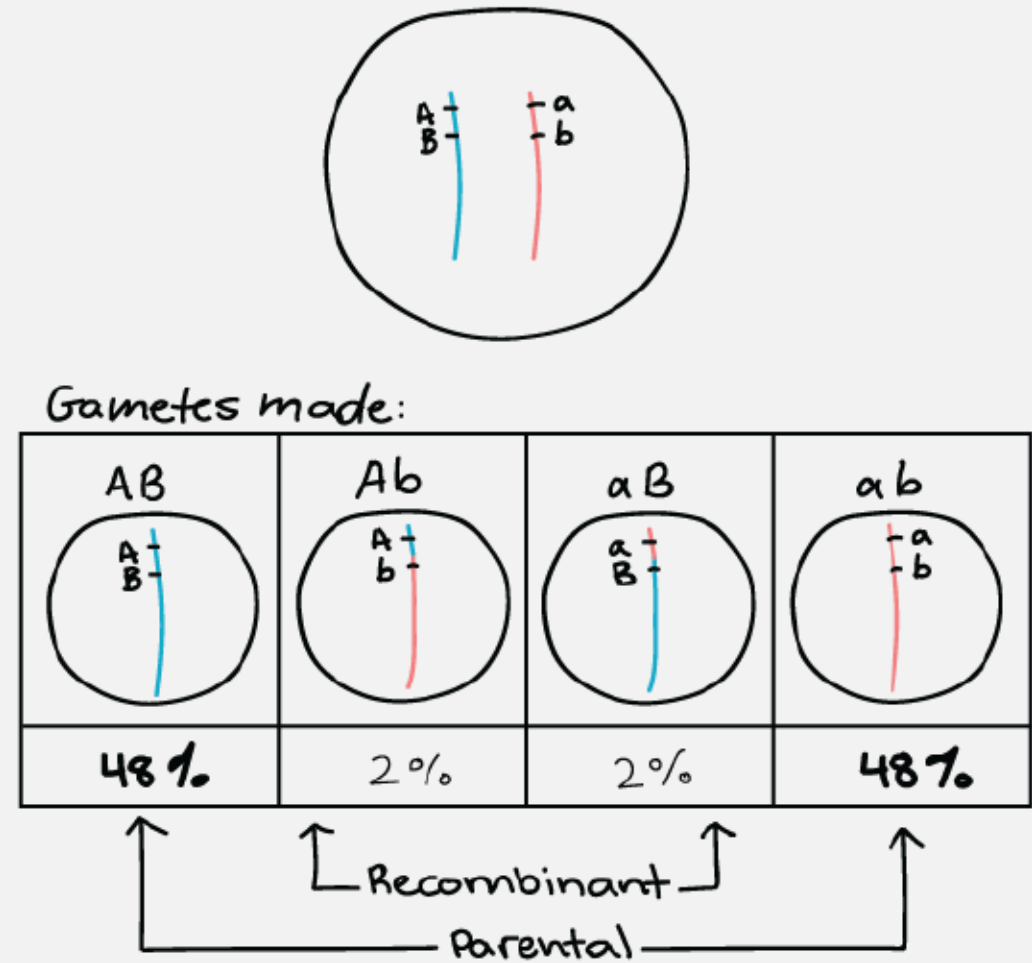


Image credit: Khan Academy

# 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

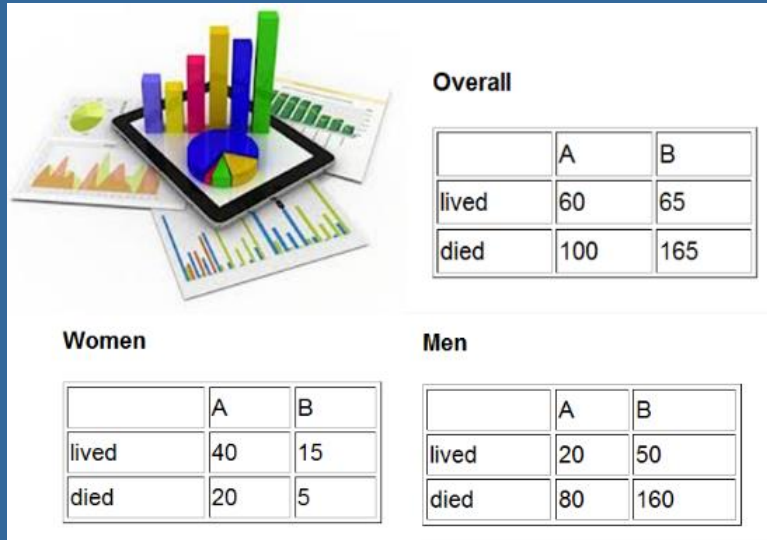
Treatment A is better

Treatment B is better



What is happening here?

# A/B sample not equalized in other attributes, ...



Intentionally left blank

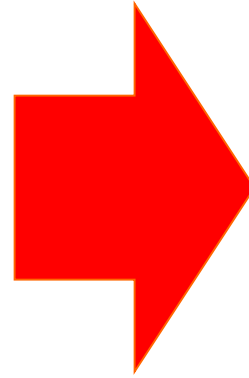
# Careless null hypothesis

## 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?

Intentionally left blank

# | Homework, due 24/1/2025

Read this paper

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](#)

Write a 1-page review report focusing how it tests for synthetic-lethal gene pairs. Discuss whether their test is a good one

# Plan, 20/1/2025

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

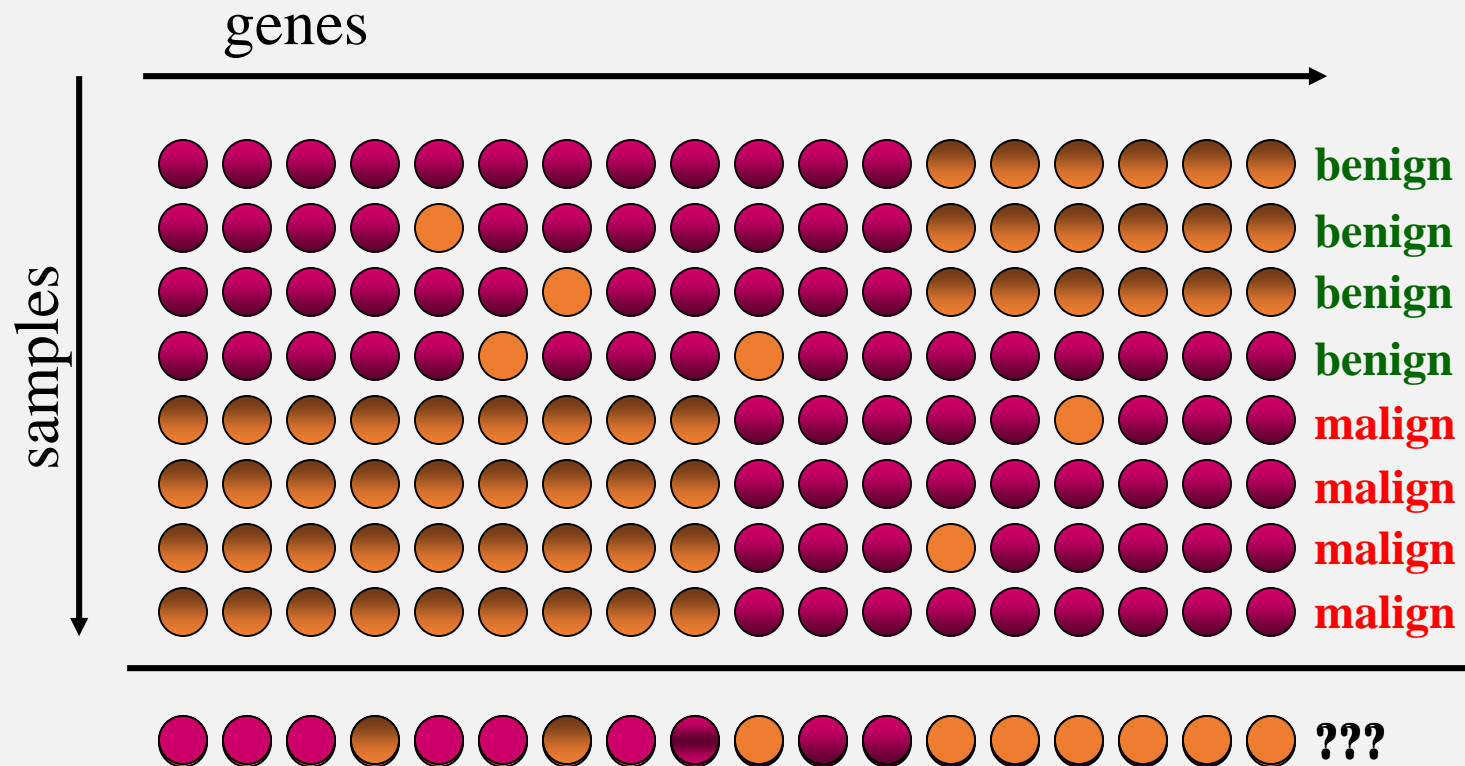
# Microarray

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

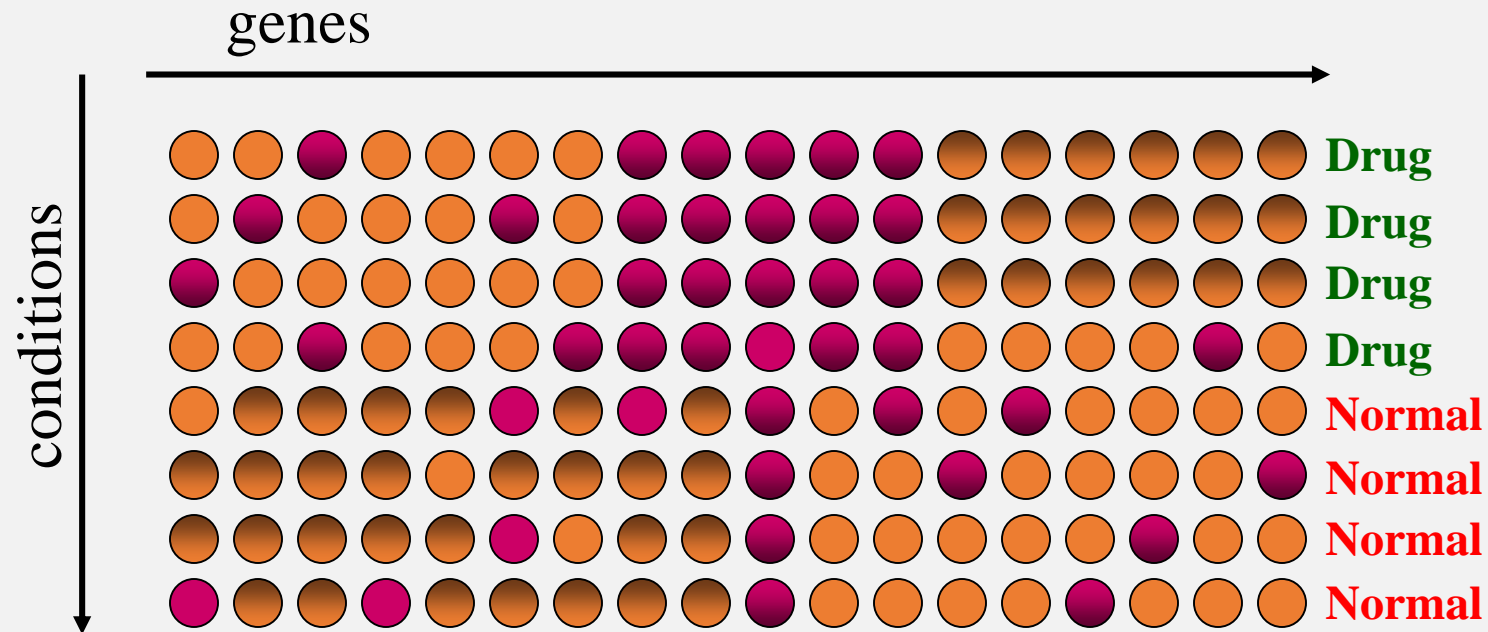
Detect what genes are being expressed or found in a cell of a tissue sample

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.,  $\chi^2$

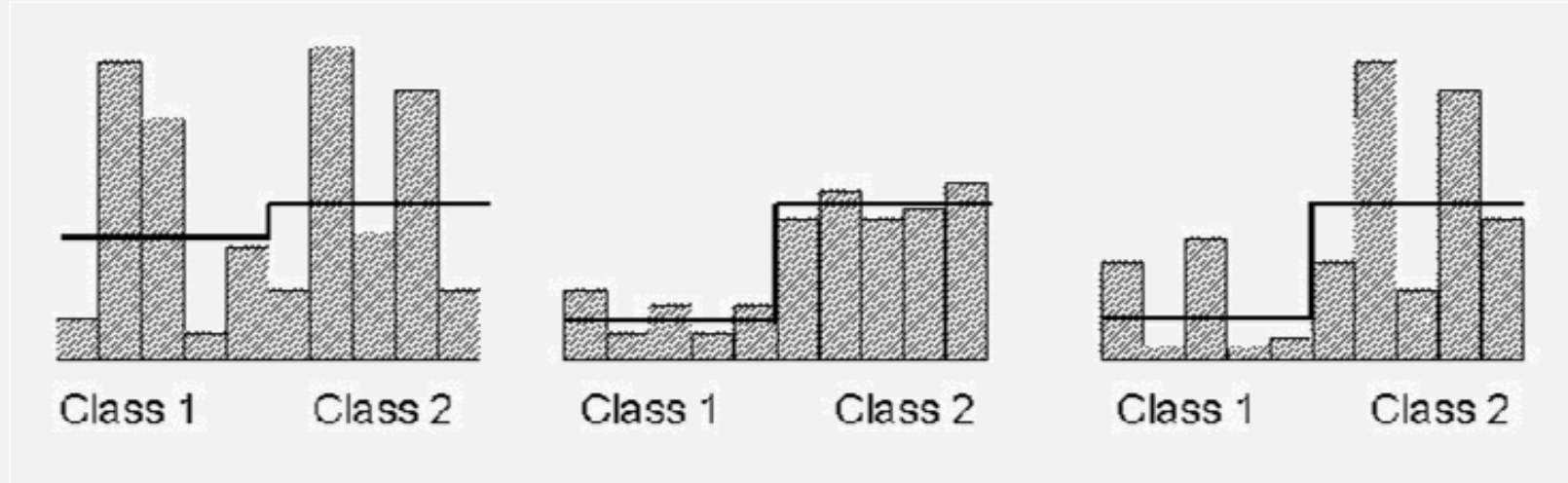
Classifier training

Classifier tuning (optional for some machine learning methods)

Apply classifier for diagnosis of future cases



# Gene selection basic idea



# Gene selection by $\chi^2$

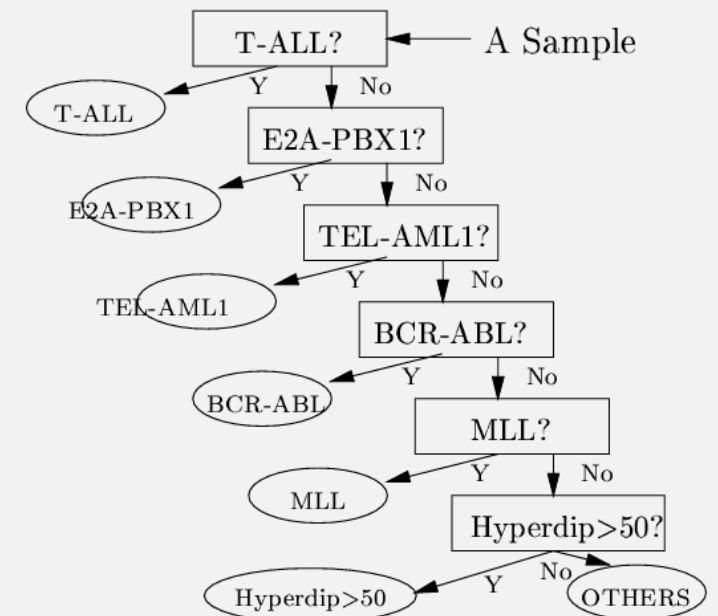
The  $\chi^2$  value of a signal is defined as:

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

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

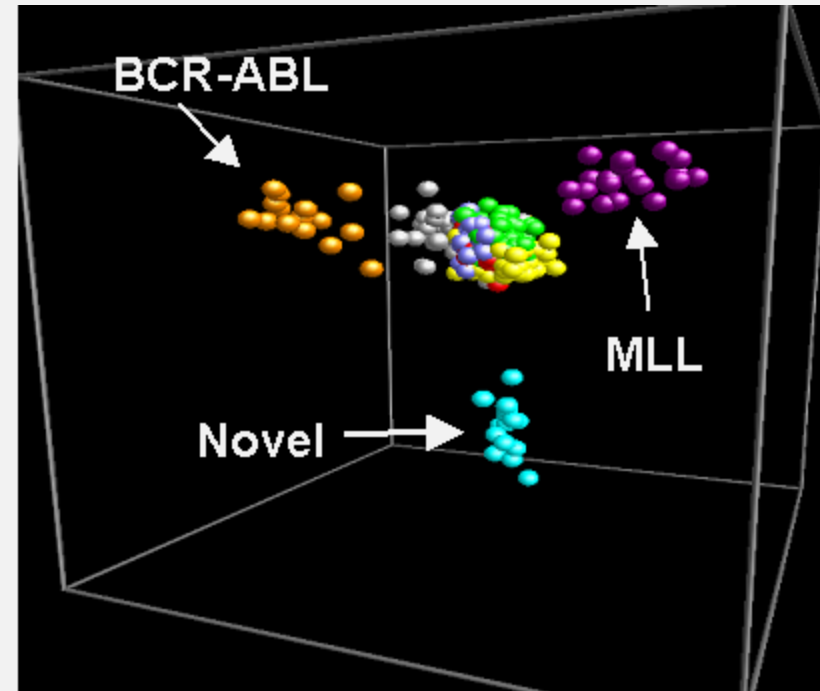
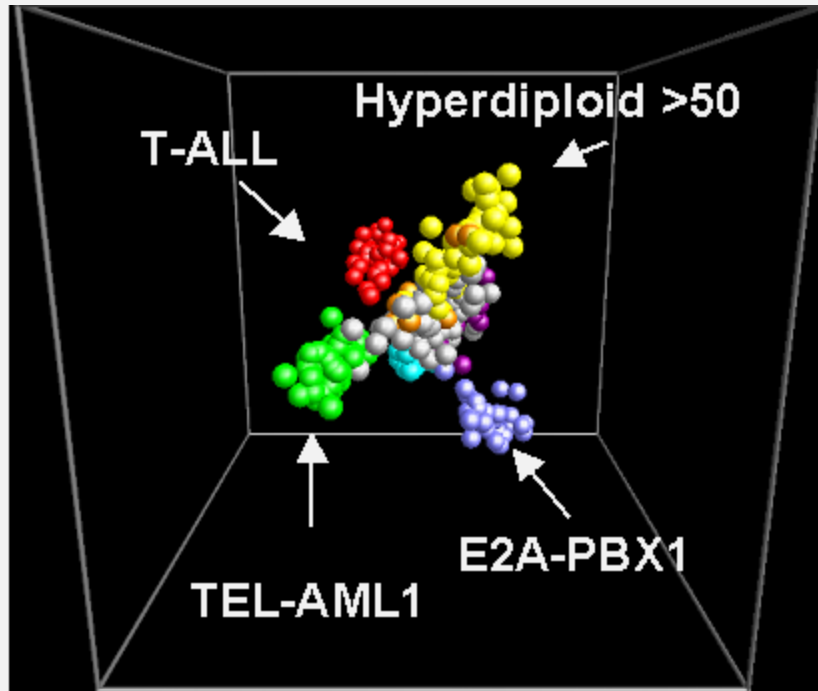
# Performance of various classifiers

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  $\chi^2$  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, 20/1/2025

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?

$$\text{Prob}(\text{gene is correlated}) = 1/2^6$$

$$\# \text{ of genes on array} = 25,000$$

$$E(\# \text{ of correlated genes}) = 390$$

⇒ 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,  $\ll 1/2^6$

# of groups =  $^{25000}C_5$

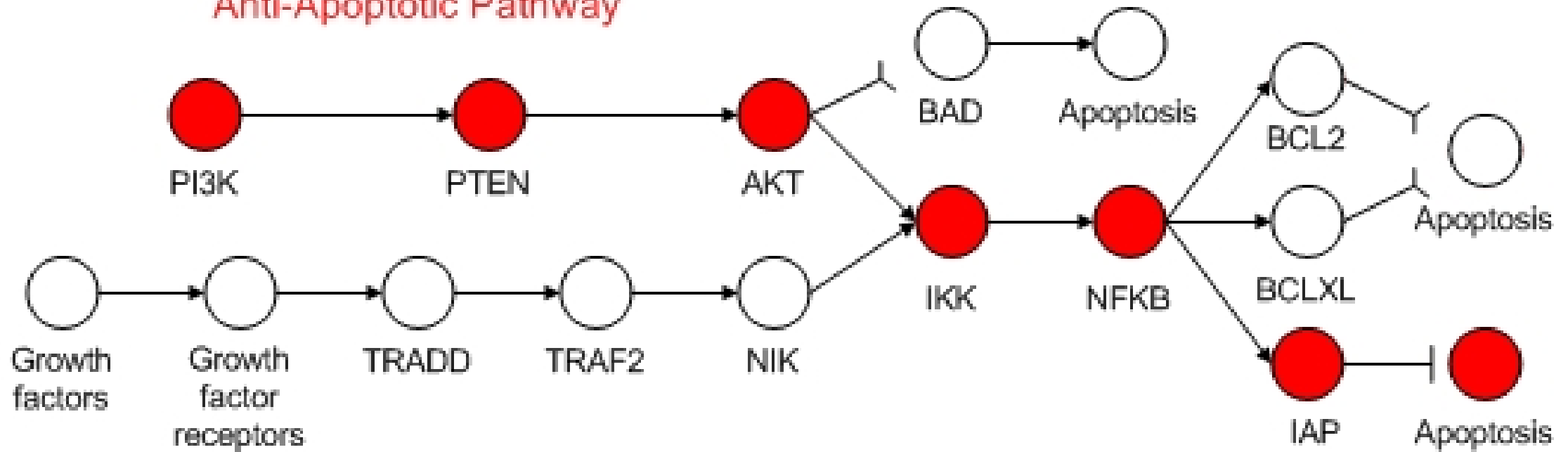
$$\begin{aligned} E(\# \text{ of correlated groups}) &= ^{25000}C_5 * (1/2^6)^5 \\ &= 7.58 * 10^{10} \end{aligned}$$

$\Rightarrow$  Even more false positives?

Perhaps no need to consider every group

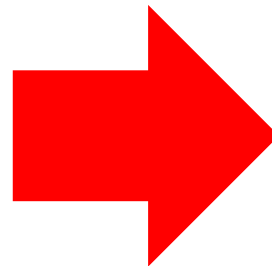


## Anti-Apoptotic Pathway



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

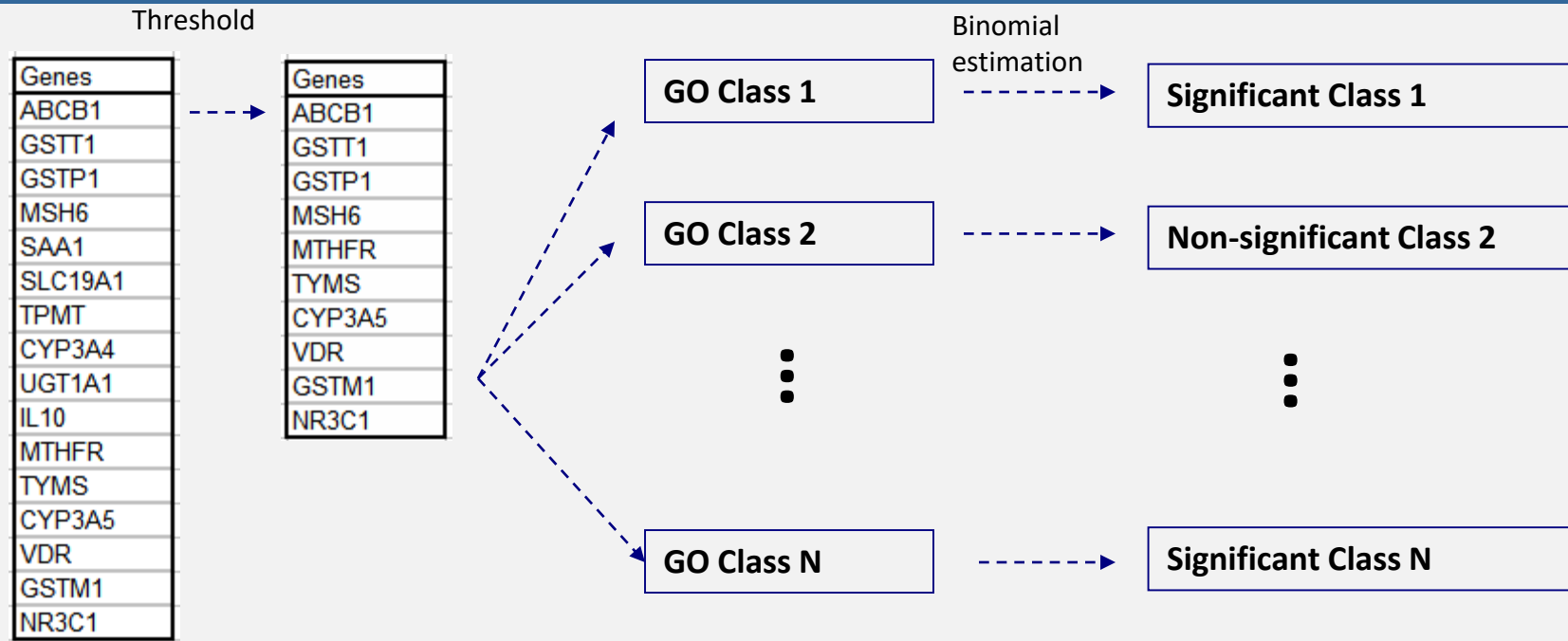
# of pathways = 1000

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

⇒ **Even more false positives?**

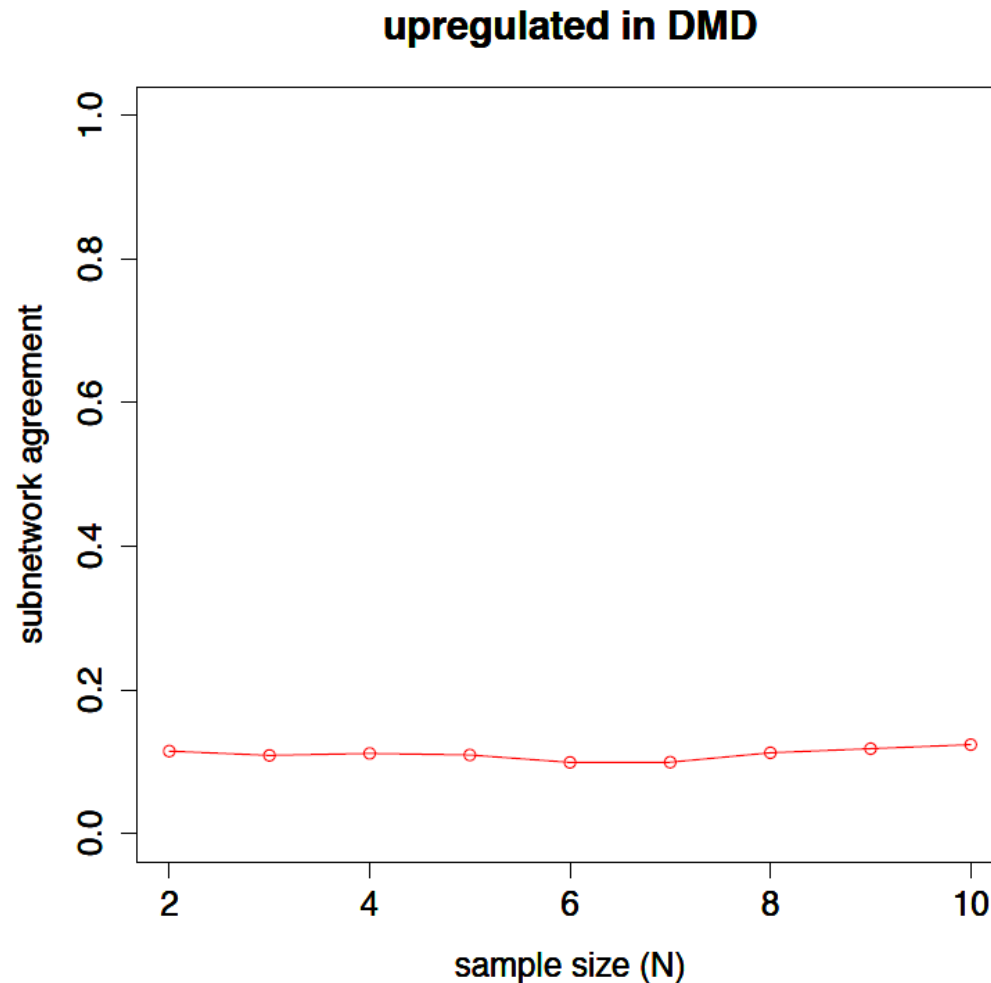
- **Perhaps no need to consider every group**

# Pathway overlap analysis via Onto-Express (aka ORA)



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

# The bewilderment persists... a crisis?



DMD gene expression data

- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data

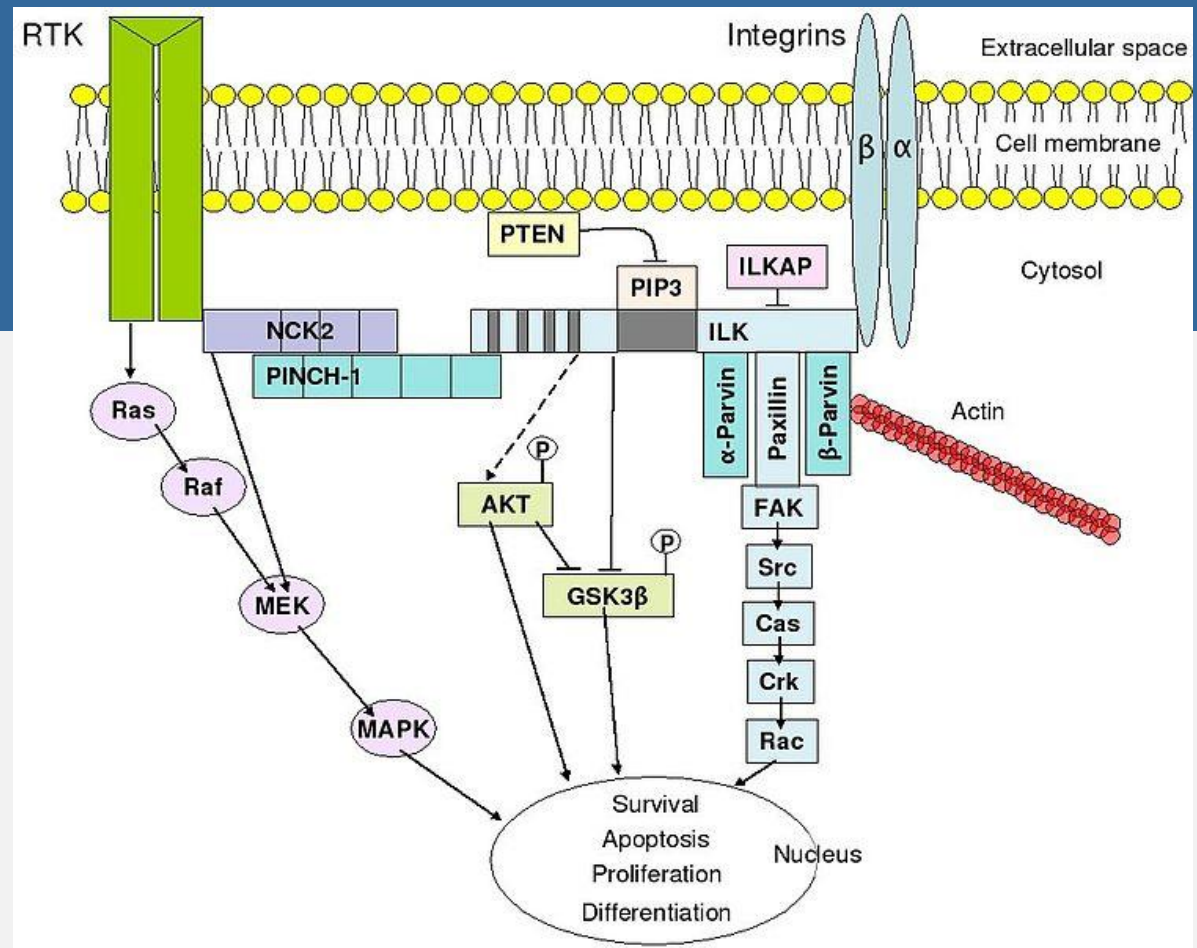
- PathwayAPI, Soh et al., 2010

# Exercise

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

# Issue #1 with ORA

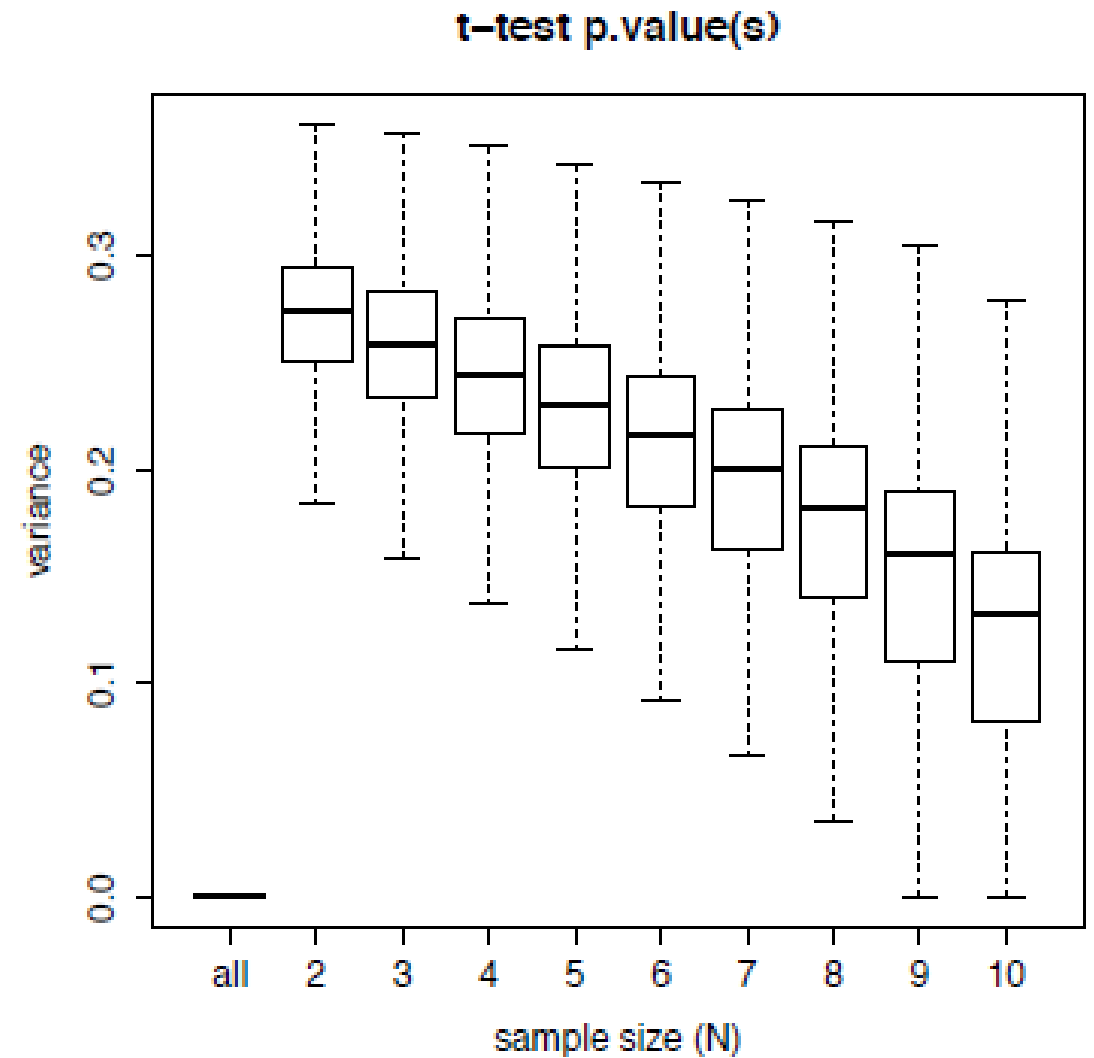
Intentionally left blank



A biological pathway is a chain of actions of molecules in cell leading to a change in cell  
 $\Rightarrow$  Behaviour of genes in a pathway is more coordinated than random ones

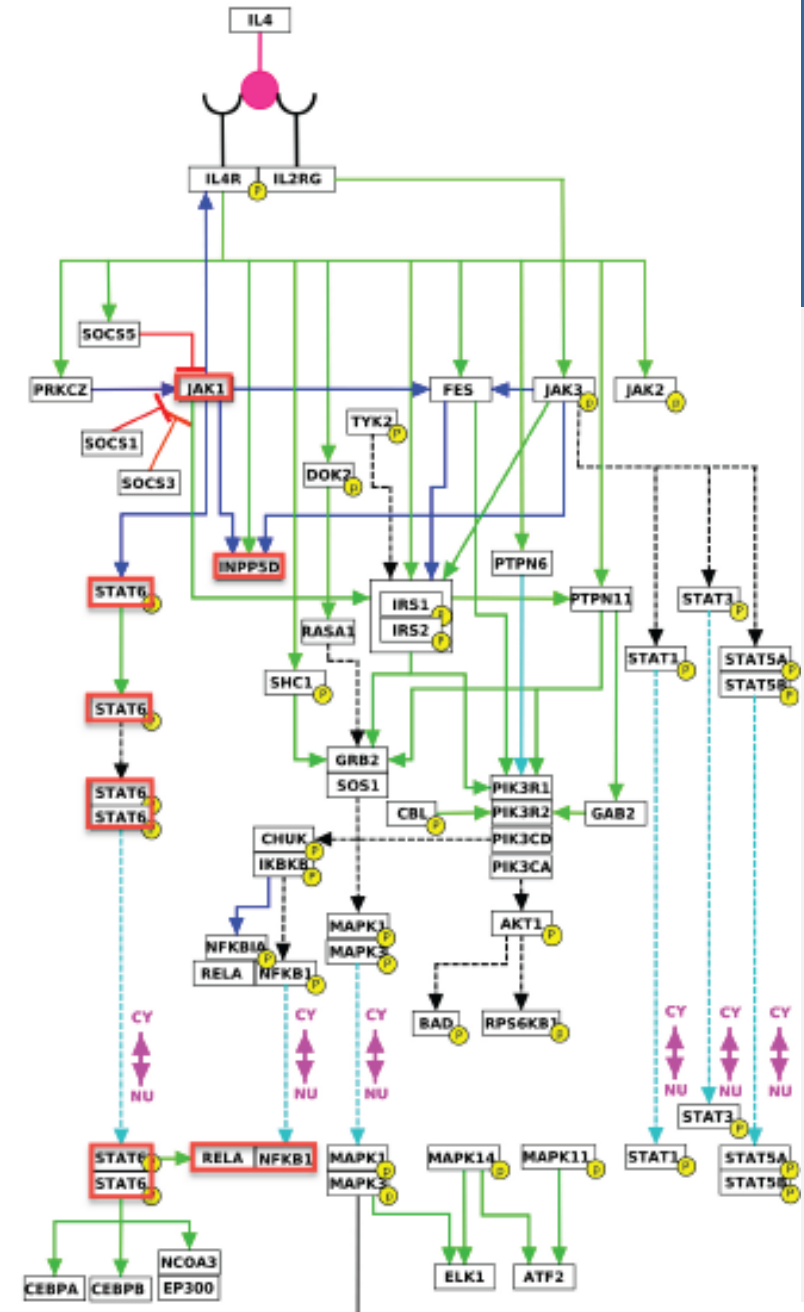
# Issue #2 with ORA

Intentionally left blank



# Issue #3 with ORA

Intentionally left blank





# Plan, 27/1/2025

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

# Exercise

How to solve the issues identified in ORA?

# ORA-Paired: Paired test and new null hypothesis

Intentionally left blank

Issue #1 is solved

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

Issue #2 is solved

*No need pre-determined list of DE genes*

Issue #3 is unsolved

Assume  
absence of  
batch effects

# Exercise

Let  $g_i$  be a gene in a pathway  $P$

Let  $p_j$  be a patient

Let  $q_k$  be a normal

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

Test whether  $\Delta_{i,j,k}$  is a distribution with mean 0

How many  $\Delta_{i,j,k}$  are there?

$|patients| * |normals| * |genes\ in\ P|$

Does this mean sample size now larger?

Does this mean more degrees of freedom?

| Testing the null hypothesis

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

Method #1

*T-test w/ the right degrees of freedom?*

*# normals + # patients – 1*

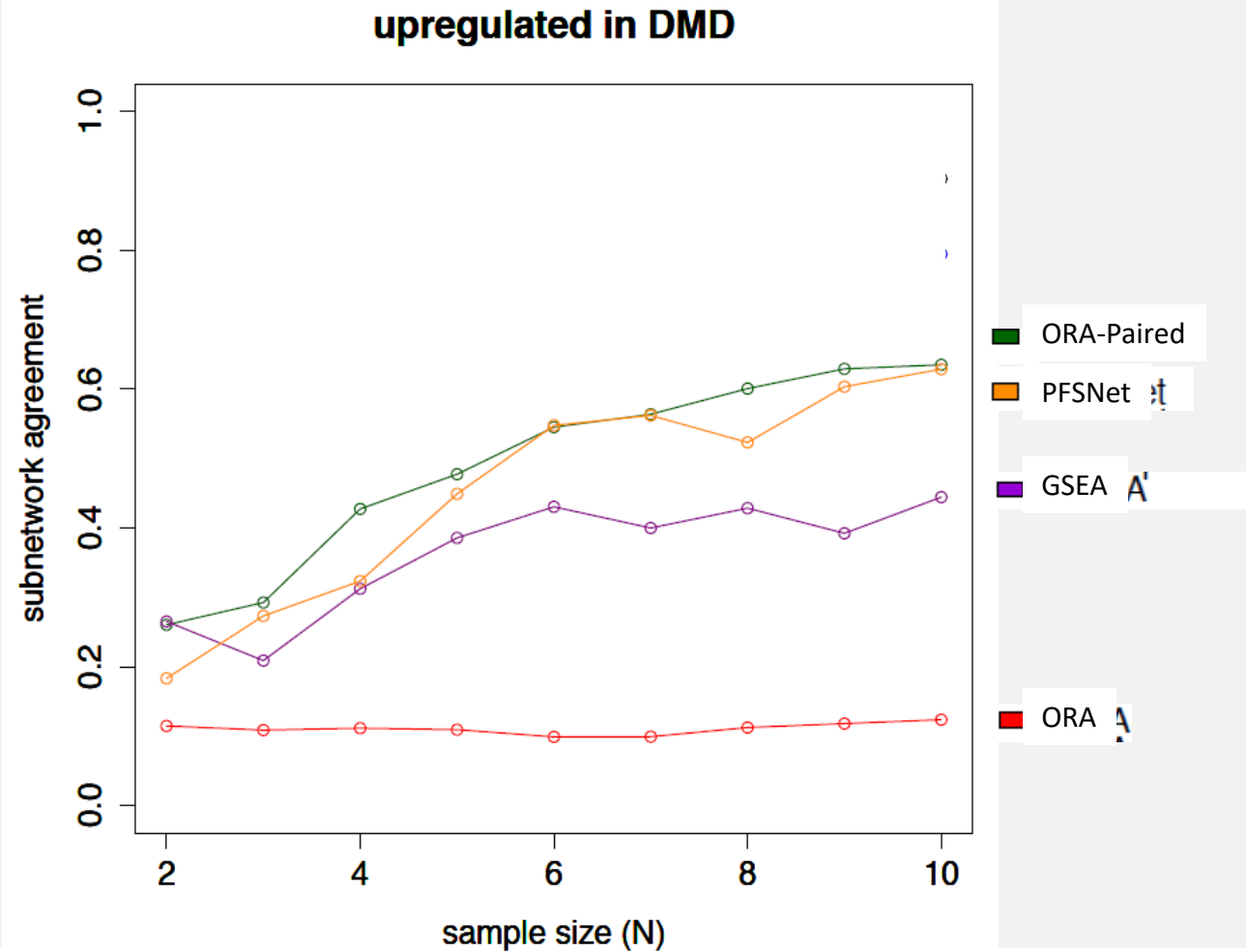
Method #2

*By the null hypothesis, a dataset & its class-label permutations are exchangeable*

*Get null distribution by class-label permutations*

*Only for large-size sample*

# Better, but not super-duper good



# NEA-Paired: Paired test on subnetworks

Given a pathway P

Let each node and its  
immediate neighbourhood in  
P be a subnetwork

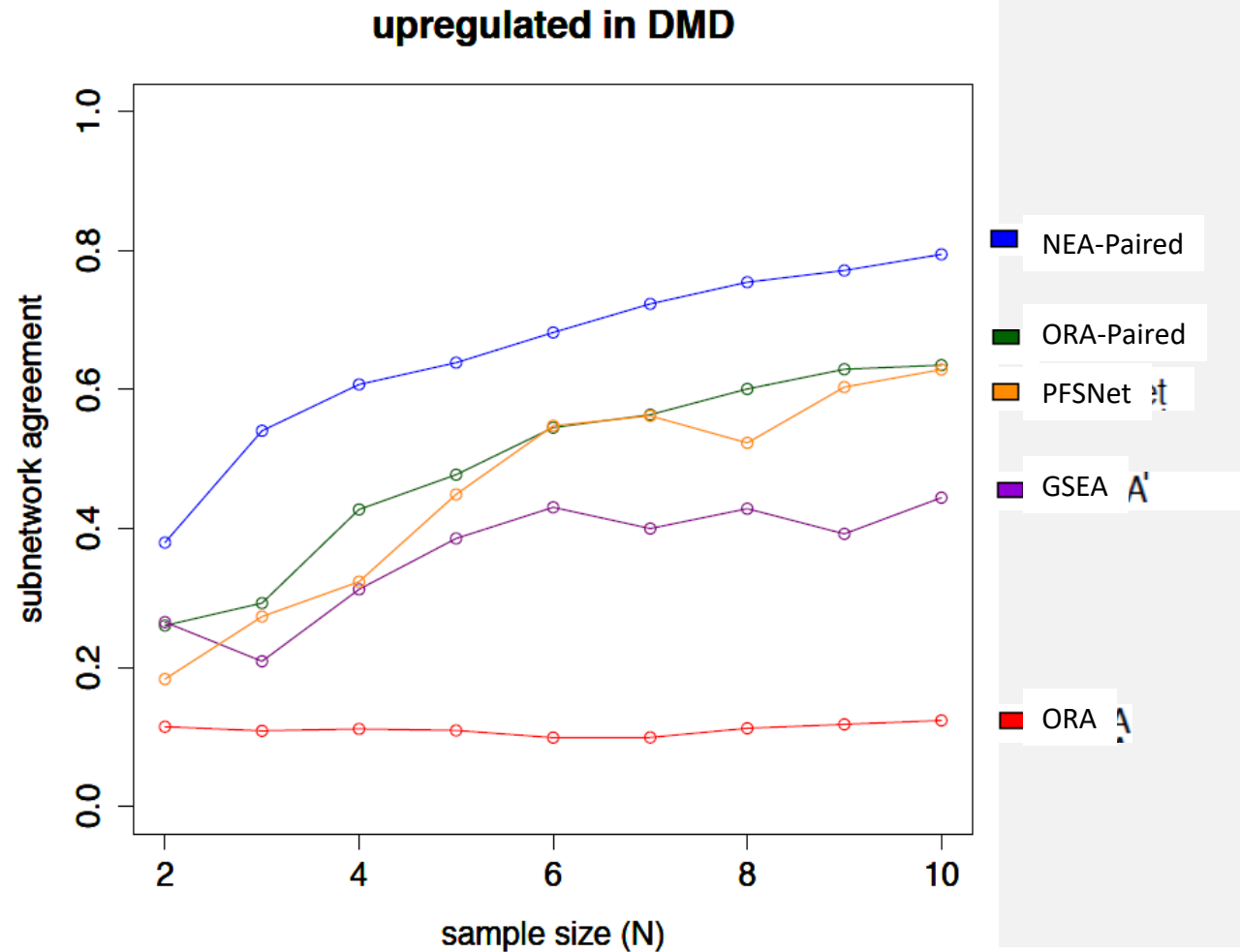
Apply ORA-Paired on each  
subnetwork individually

Issues #1 & #2 are solved as per  
ORA-Paired

Issue #3 is partly solved

*Testing subnetworks instead of  
whole pathways*

# Much better performance





# 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

# | Presentations and 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](#)