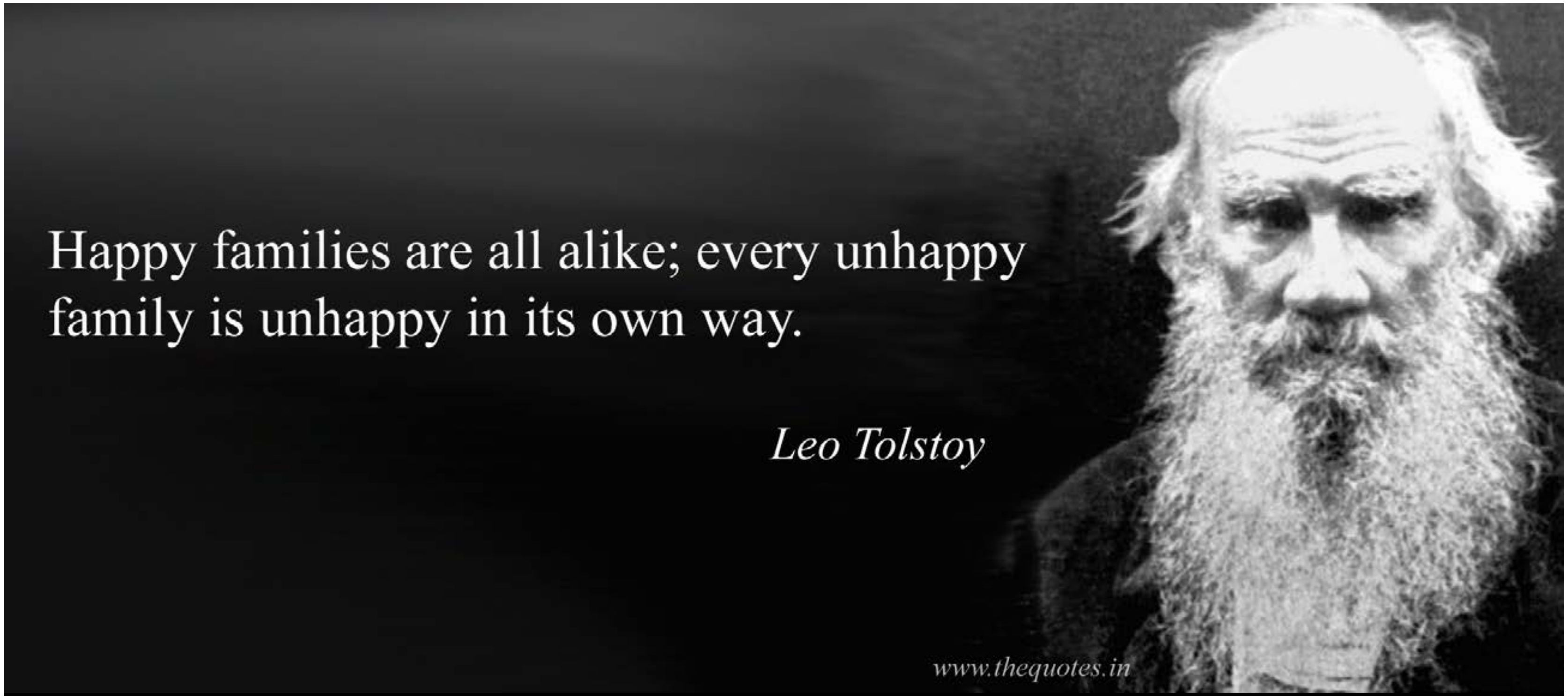
Anna Karenina and the careless null hypothesis in omics data analysis

Wong Limsoon

NUS
National University
of Singapore

Anna Karenina Principle



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

Leo Tolstoy

www.thequotes.in

Translation

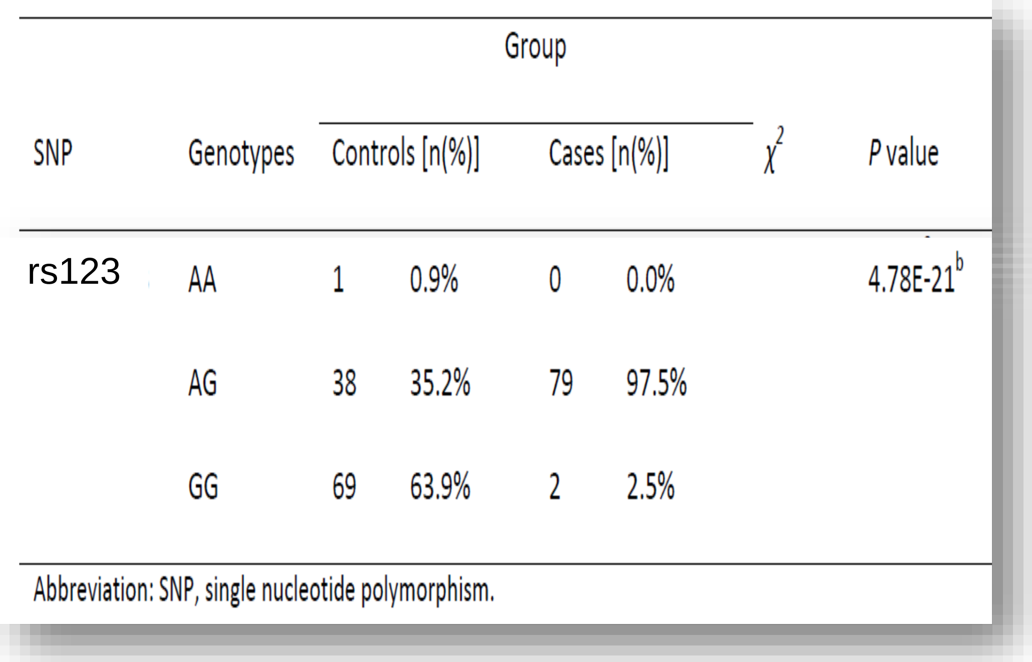
- There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest

A Statistician Responds to a Marriage Proposal
Will you marry me?
I Reject the Null Hypothesis.
Huh? What does that mean?

Reproduced by permission of John Wiley and Sons, Inc
from the book *Statistics from A to Z — Confusing Concepts Clarified.*

GETTING THE NULL HYPOTHESIS RIGHT

Example 1



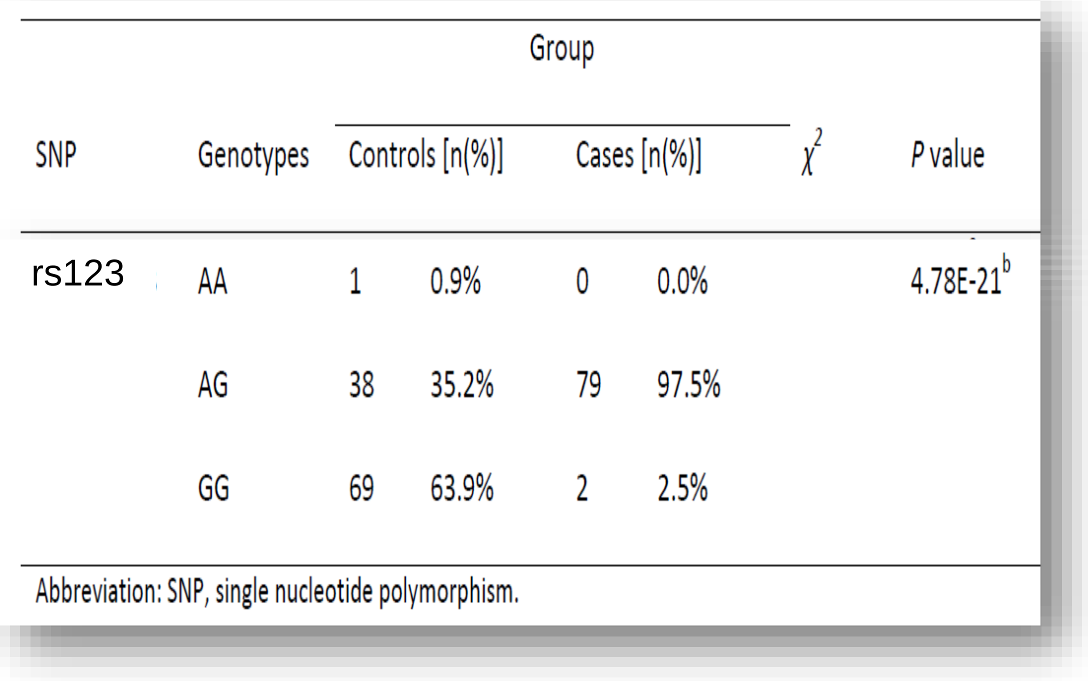
		Group					
SNP	Genotypes	Controls [n(%)]		Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

A seemingly obvious conclusion

- **A scientist claims the SNP rs123 is a great biomarker for a disease**
  - If rs123 is AA or GG, unlikely to get the disease
  - If rs123 is AG, a 3:1 odd of getting the disease
- **A straightforward $\chi 2$ test. Anything more/wrong?**

Discussion #1



		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.

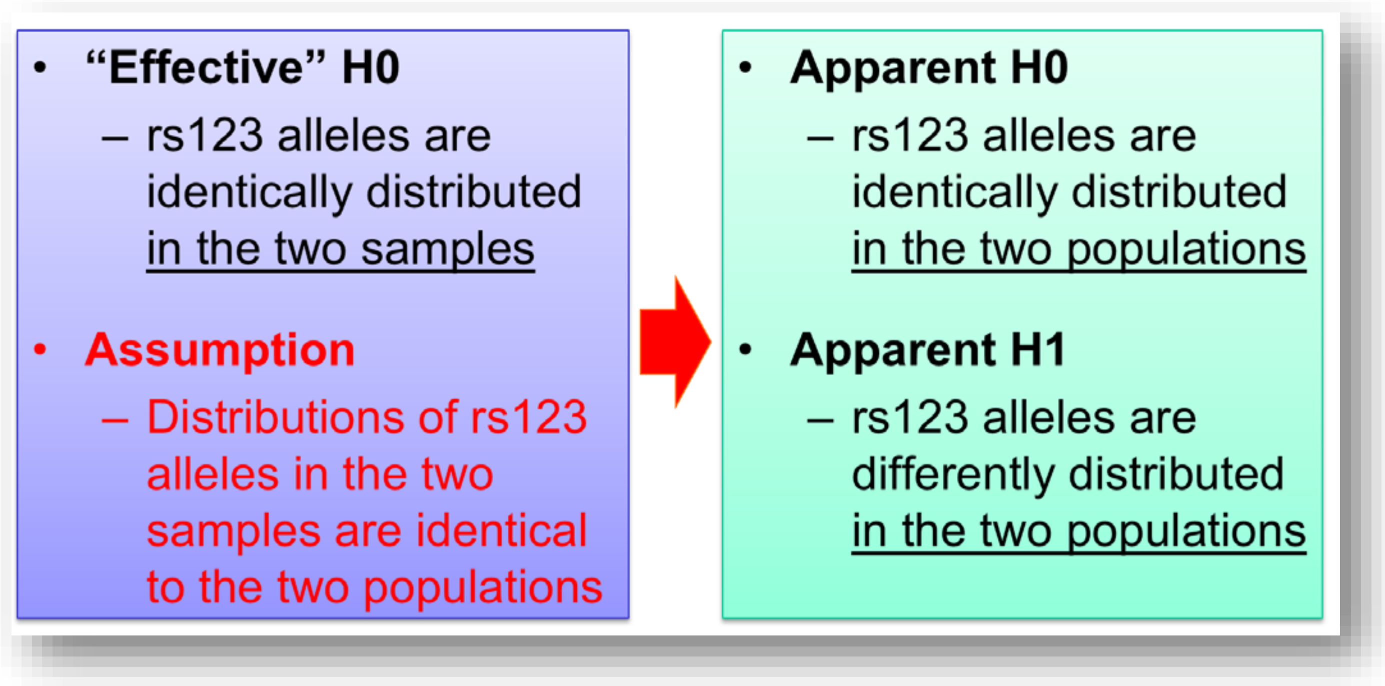
- **What is the null / alternative hypothesis corresponding to this statistical test?**

Careless null hypothesis

- **“Effective” H0**
  - 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 H0**
  - rs123 alleles are identically distributed in the two populations
- **Apparent H1**
  - rs123 alleles are differently distributed in the two populations

Discussion #2



- **“Effective” H0**
  - 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 H0**
  - rs123 alleles are identically distributed in the two populations
- **Apparent H1**
  - rs123 alleles are differently distributed in the two populations

- **The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?**

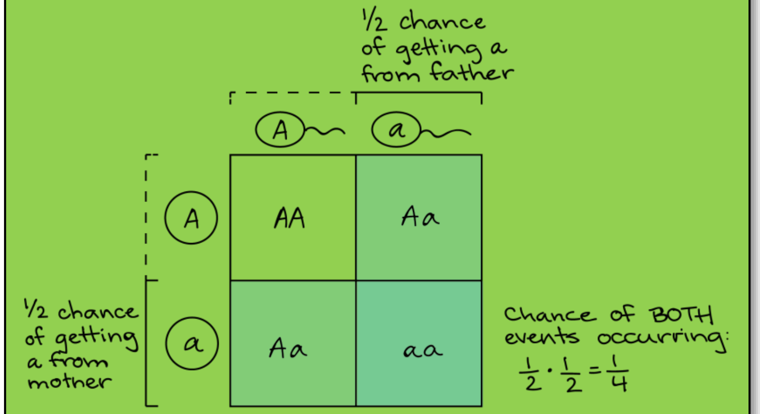
Refined null hypothesis

- **Refined H0**
  - Distributions of rs123 alleles in the two samples are identical to the two populations, **and**
  - rs123 alleles are identically distributed in the two populations

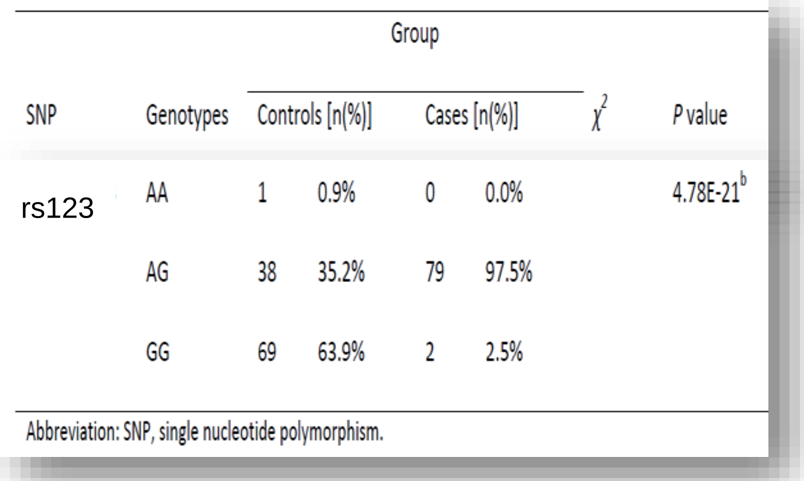
- **Refined H1**
  - Distributions of rs123 alleles in the two samples are different from the two populations, **or**
  - rs123 alleles are differently distributed in the two populations

i.e. sample is biased

Sample bias is revealed by domain logic


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: $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Basic rule of human genetics

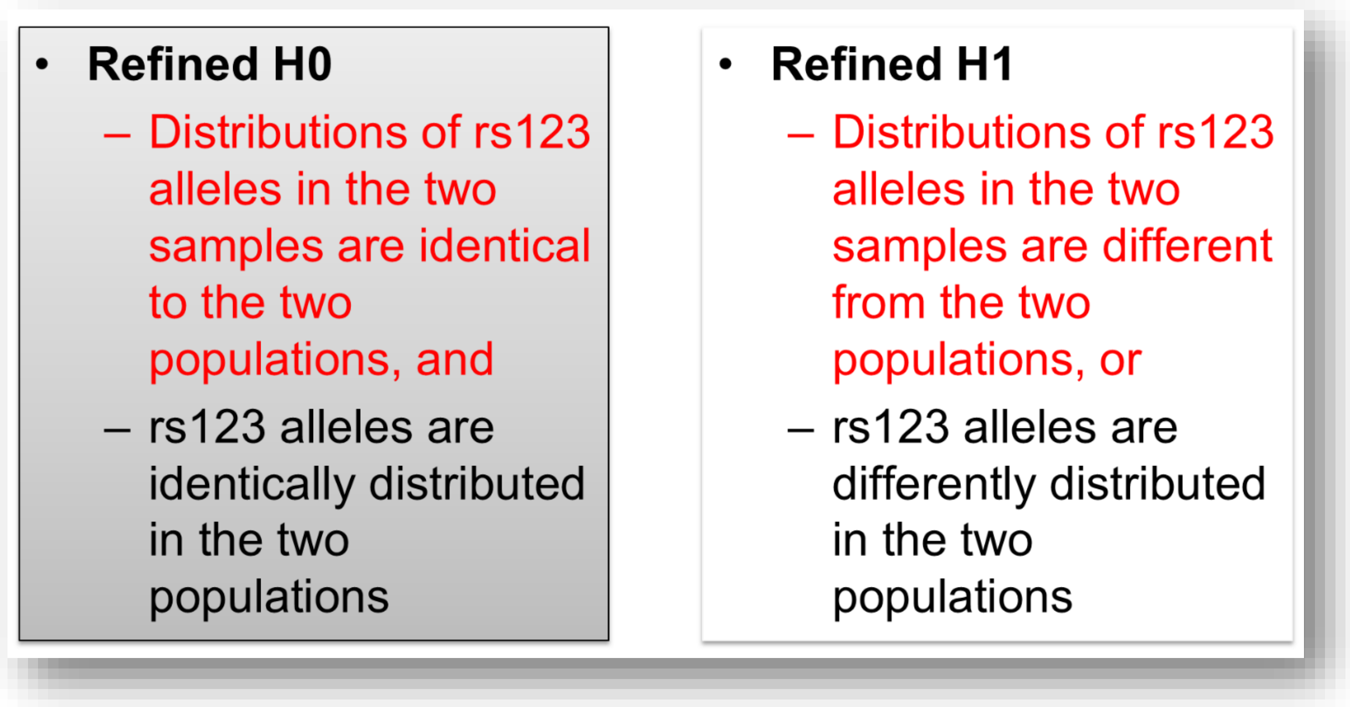


SNP	Genotypes	Group				χ^2	*P* value
		Controls [n(%)]		Cases [n(%)]			
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

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 #3



- **Refined H0**
  - Distributions of rs123 alleles in the two samples are identical to the two populations, and
  - rs123 alleles are identically distributed in the two populations

- **Refined H1**
  - Distributions of rs123 alleles in the two samples are different from the two populations, or
  - rs123 alleles are differently distributed in the two populations

- **Suppose distributions of rs123 alleles in the samples are identical to the populations and the test is significant**
- **Can we say rs123 mutation causes the disease?**

Three types of reasoning

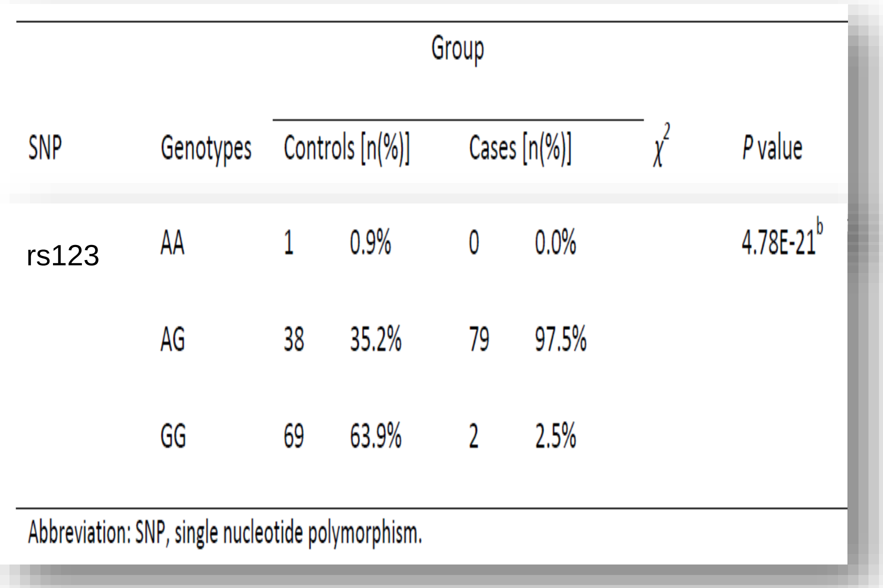
- **Deduction**
  - – All men are mortal
  - – Socrates is a man
  - ⇒Socrates is mortal

- **Induction**
  - – Socrates is a man
  - – Socrates is mortal
  - ⇒All men are mortal, provided there is no counter example

- **Abduction**
  - – All men are mortal
  - – Socrates is mortal
  - ⇒Socrates is a man, provided there is no other explanation of Socrates' mortality

Abduction in action

- **Hypothesis**
  - If rs123 mutation causes disease, the statistical test is significant

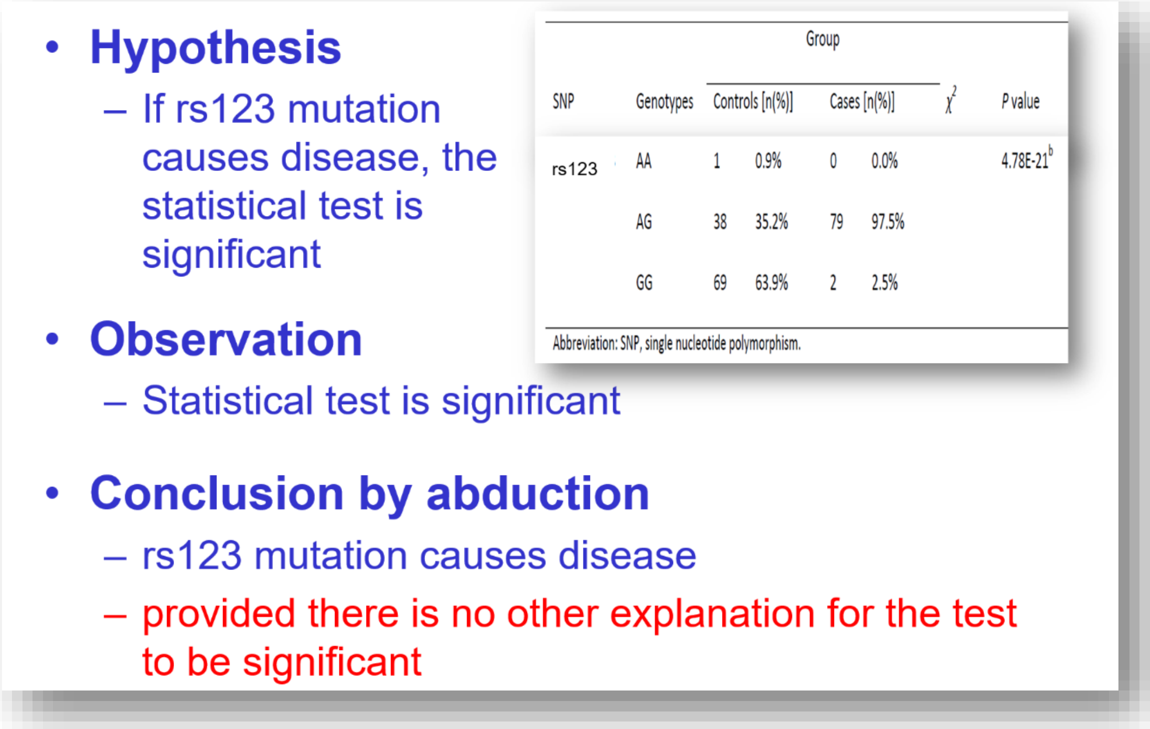


		Group					
SNP	Genotypes	Controls [n(%)]		Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

- **Observation**
  - Statistical test is significant
- **Conclusion by abduction**
  - rs123 mutation causes disease
  - provided there is no other explanation for the test to be significant

Discussion #4



- **Hypothesis**
  - If rs123 mutation causes disease, the statistical test is significant
- **Observation**
  - Statistical test is significant
- **Conclusion by abduction**
  - rs123 mutation causes disease
  - provided there is no other explanation for the test to be significant

		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.

- **How to incorporate "provided there is no other explanation" into the analysis?**

How about this?

- **Choose a sample of Cases and a sample of Controls such that for each stratification p1/p2, the distribution of p1/p2 in Cases is same as the distribution of p1/p2 in Controls**
  - i.e. equalize / control for other factors
- **Then test:**

- **H0**
  - X's alleles are identically distributed in the two samples

- **H1**
  - X's alleles are differently distributed in the two samples

- **This makes the significance of the test independent of other explanations**
- **It does not say "no other explanation"**

Or this?

- Look for another gene X such that

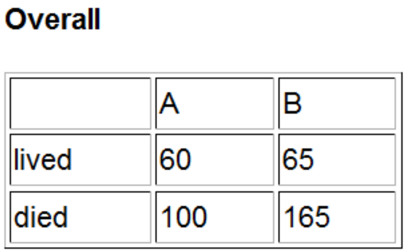
- **H0**
  - Distributions of X’s alleles in the two samples are identical to the two populations, **and**
  - X’s alleles are identically distributed in the two populations

- **H1**
  - Distributions of X’s alleles in the two samples are different from the two populations, **or**
  - X’s alleles are differently distributed in the two populations

- **When the red part of H1 is false, this implies gene X mutation is an alternative explanation for the significance of rs123 mutation and thus the disease. Why?**

Example 2

A seemingly obvious conclusion

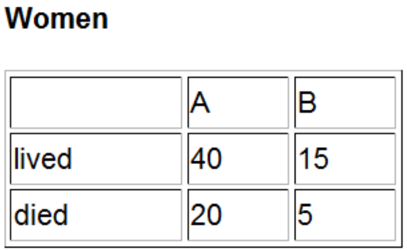


Overall

	A	B
lived	60	65
died	100	165

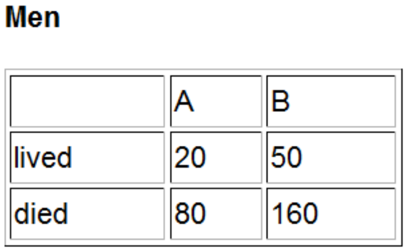
Looks like treatment A is better

What is happening here?



Women

	A	B
lived	40	15
died	20	5



Men

	A	B
lived	20	50
died	80	160

Looks like treatment B is better

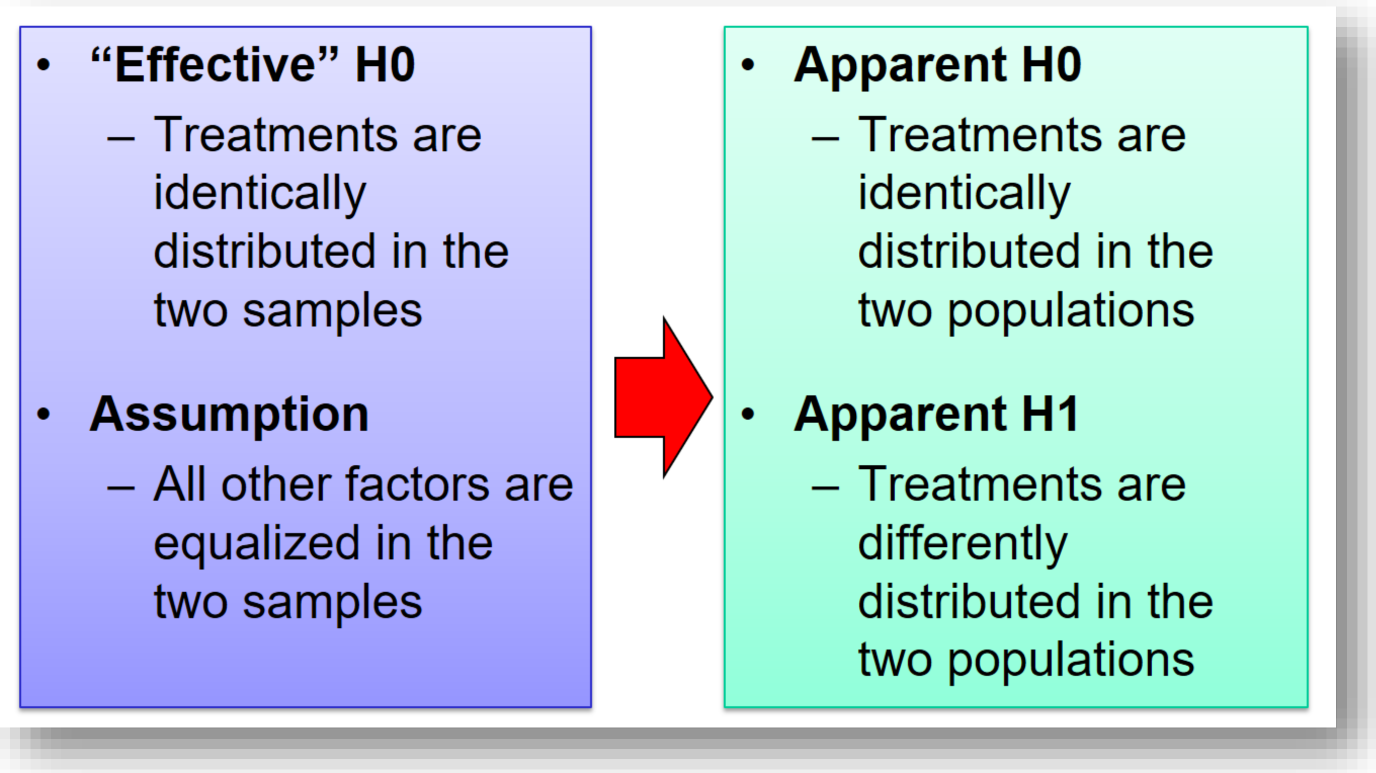
Careless null hypothesis

- **“Effective” H0**
  - Treatments are identically distributed in the two samples
- **Assumption**
  - All other factors are equalized in the two samples

➡

- **Apparent H0**
  - Treatments are identically distributed in the two populations
- **Apparent H1**
  - Treatments are differently distributed in the two populations

Discussion #5



- **“Effective” H0**
  - Treatments are identically distributed in the two samples
- **Assumption**
  - All other factors are equalized in the two samples

- **Apparent H0**
  - Treatments are identically distributed in the two populations
- **Apparent H1**
  - Treatments are differently distributed in the two populations

- **The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?**

Refined null hypothesis

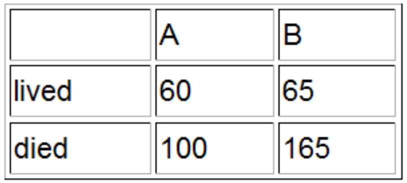
- **Refined H0**
  - All other factors are equalized in the two samples, **and**
  - Treatments are identically distributed in the two samples

- **Refined H1**
  - Some factors are not equalized in the two samples, **or**
  - Treatments are differently distributed in the two populations

- **Any other thing missing?**

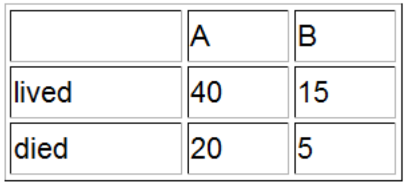
A/B sample not equalized in other attributes, viz. sex

Overall



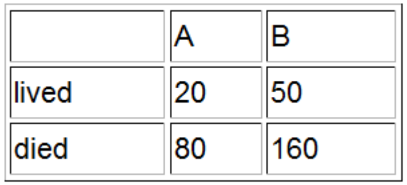
	A	B
lived	60	65
died	100	165

Women



	A	B
lived	40	15
died	20	5

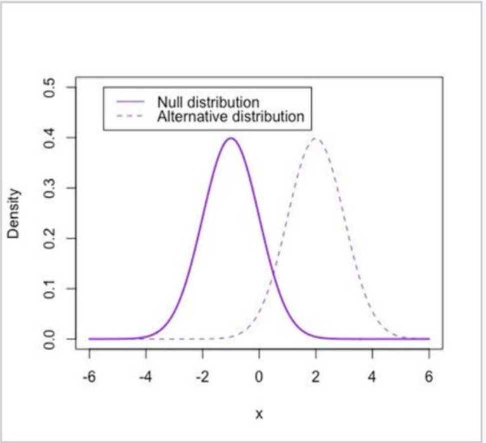
Men



	A	B
lived	20	50
died	80	160

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

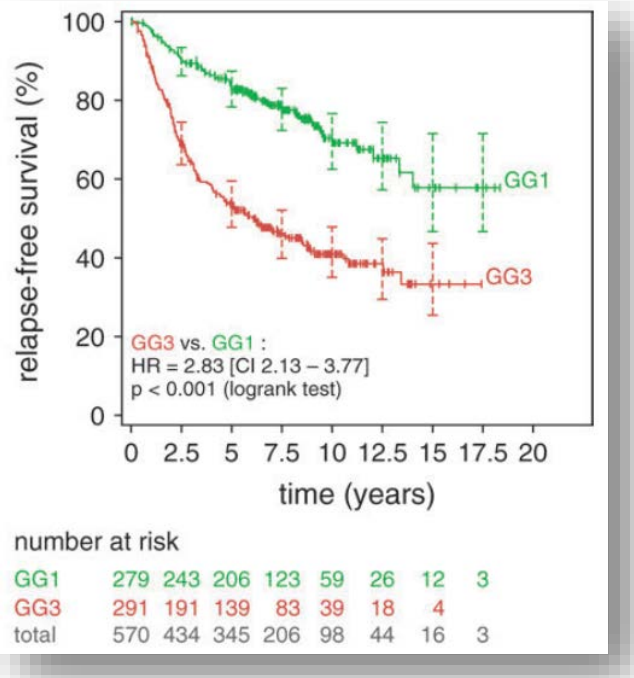
In statistical hypothesis testing, the **null distribution** is the probability **distribution** of the test statistic when the **null** hypothesis is true. For example, in an F-test, the **null distribution** is an F-**distribution**.



Null and alternative distribution

GETTING THE NULL DISTRIBUTION RIGHT

Example 3



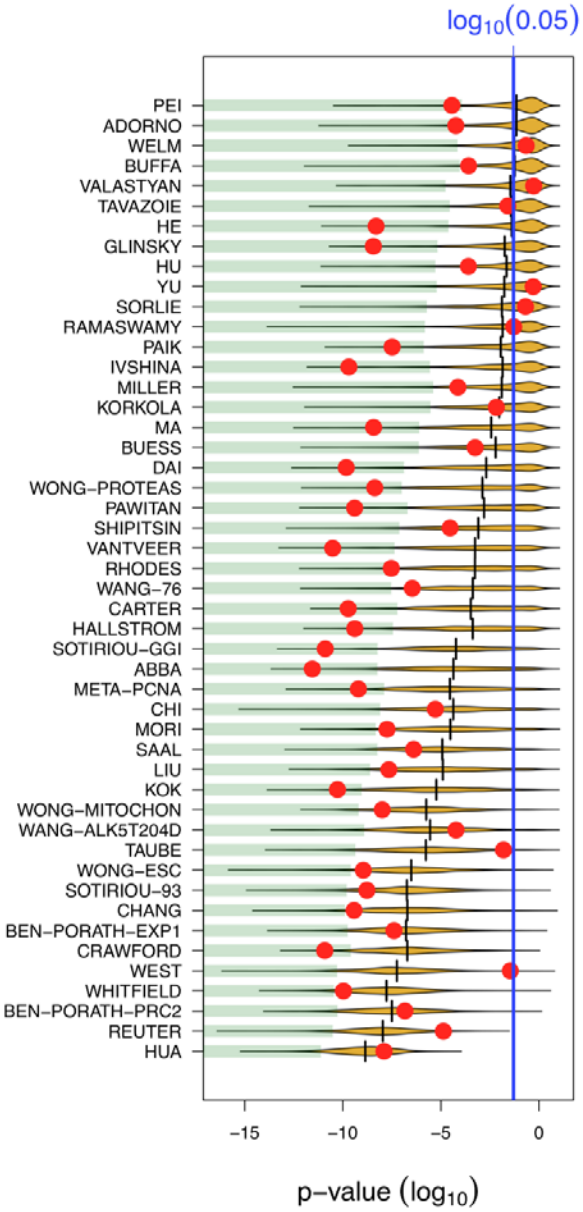
number at risk

GG1	279	243	206	123	59	26	12	3
GG3	291	191	139	83	39	18	4	
total	570	434	345	206	98	44	16	3

A seemingly obvious conclusion

- **A multi-gene signature is claimed as a good biomarker for breast cancer survival**
  - Cox's survival model p-value << 0.05

- **A straightforward Cox's proportional hazard analysis. Anything more/wrong?**

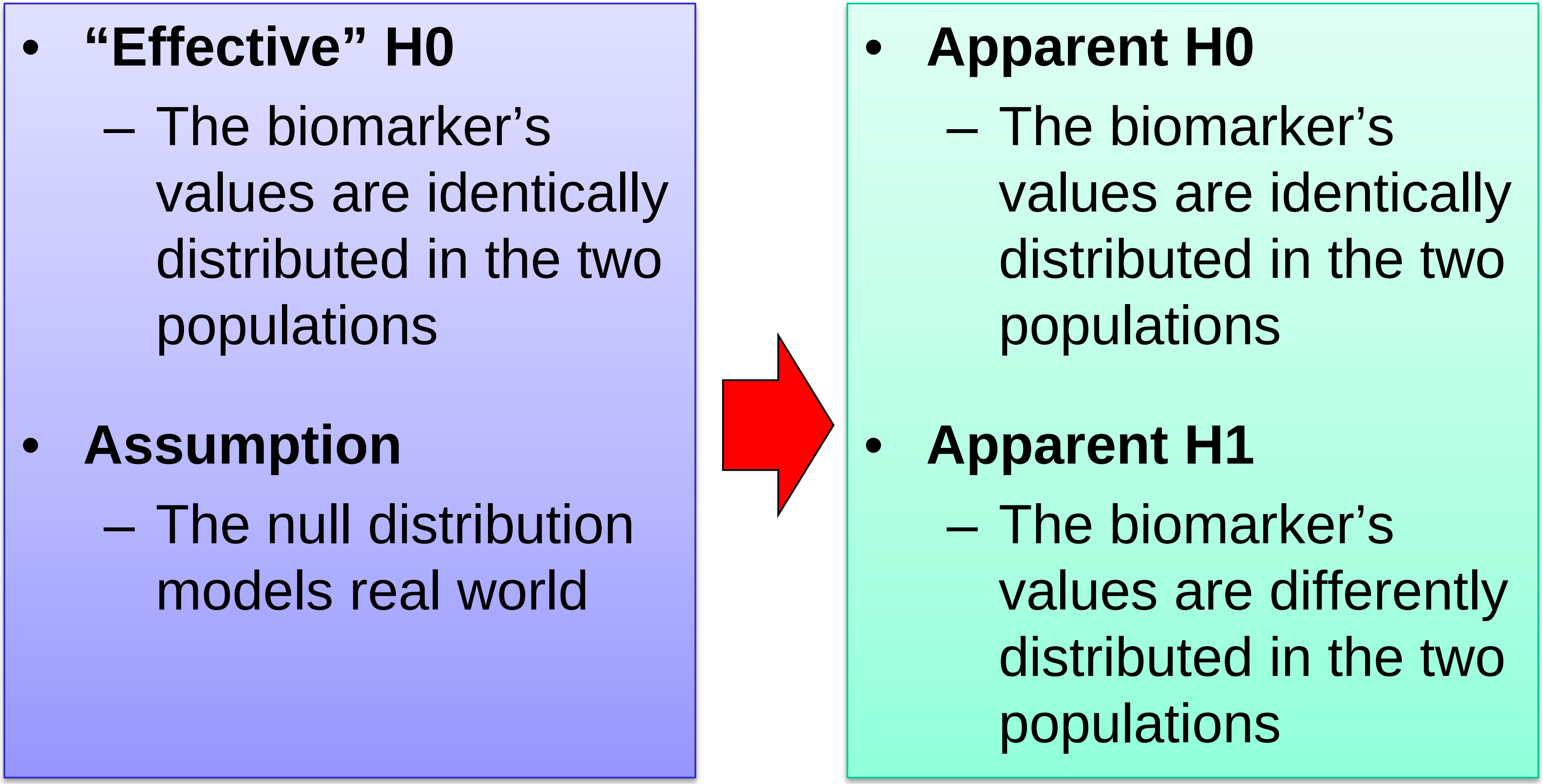


Venet et al., PLOS Comput Biol, 2011

Almost all random signatures also have p-value < 0.05

- **Theoretical null distribution used in Cox's proportion hazard analysis does not match the empirical null distribution**
- **What can we do about this?**

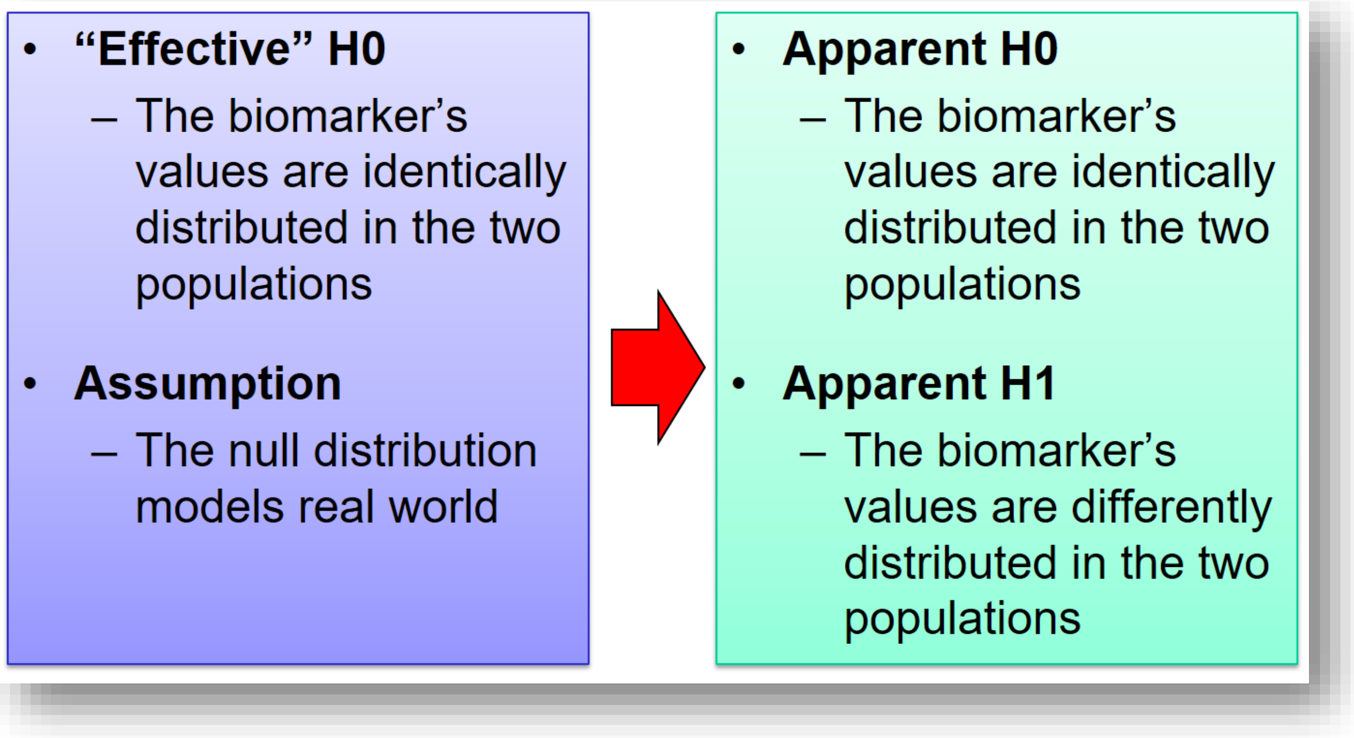
Careless null hypothesis



- **“Effective” H0**
  - The biomarker’s values are identically distributed in the two populations
- **Assumption**
  - The null distribution models real world

- **Apparent H0**
  - The biomarker’s values are identically distributed in the two populations
- **Apparent H1**
  - The biomarker’s values are differently distributed in the two populations

Discussion #6



- **“Effective” H0**
  - The biomarker’s values are identically distributed in the two populations
- **Assumption**
  - The null distribution models real world

- **Apparent H0**
  - The biomarker’s values are identically distributed in the two populations
- **Apparent H1**
  - The biomarker’s values are differently distributed in the two populations

- **The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?**

Refined null hypothesis

- **Refined H0**
  - The biomarker’s values are identically distributed in the two populations, **and**
  - The null distribution models real world

- **Refined**
  - The biomarker’s values are differently distributed in the two populations, **or**
  - The null distribution does not model real world

Example 4

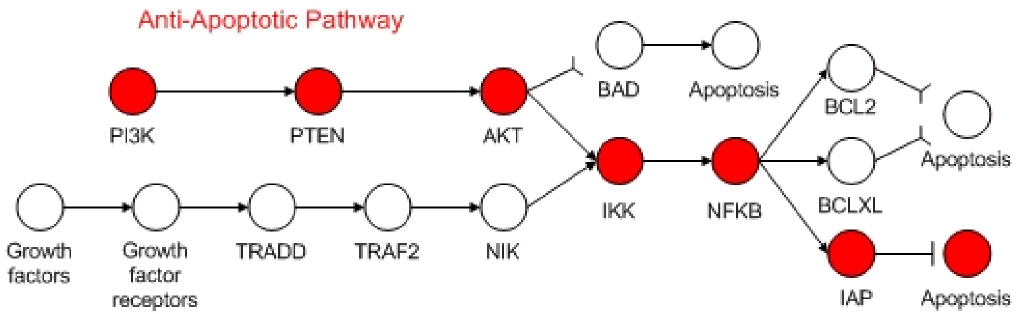
Gene-selection methods have poor reproducibility

- **Low % of overlapping genes from diff expt in general**
  - 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

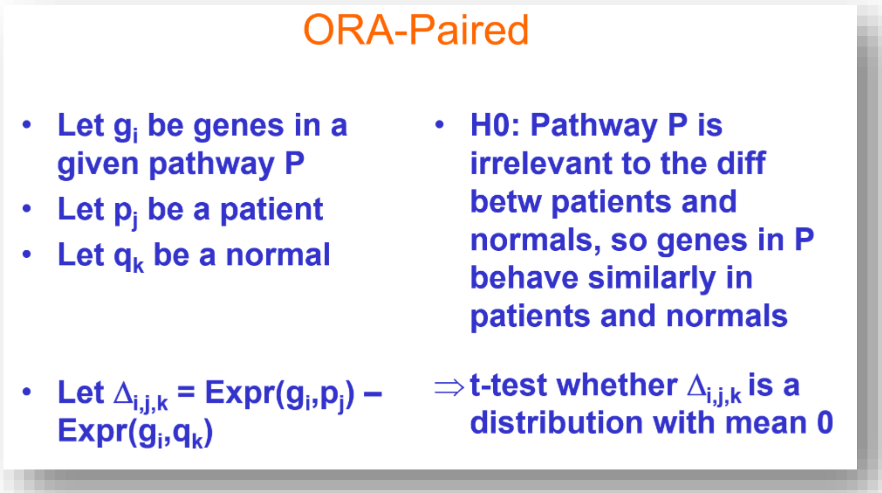
- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) - \text{Expr}(g_i,q_k)$

- 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 #7



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**

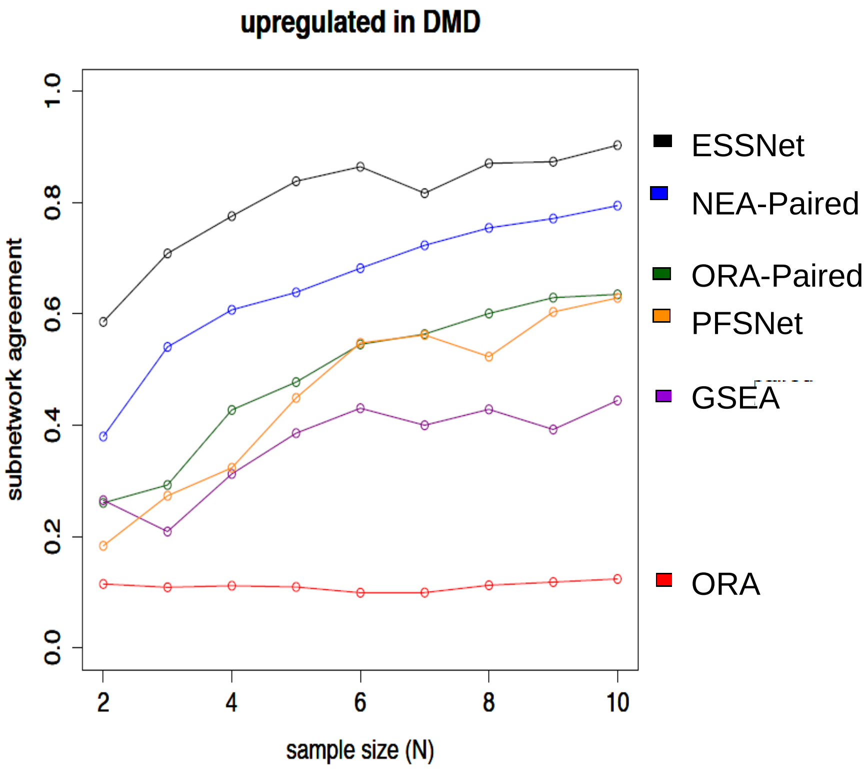
Testing the null hypothesis

"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.

Example 5

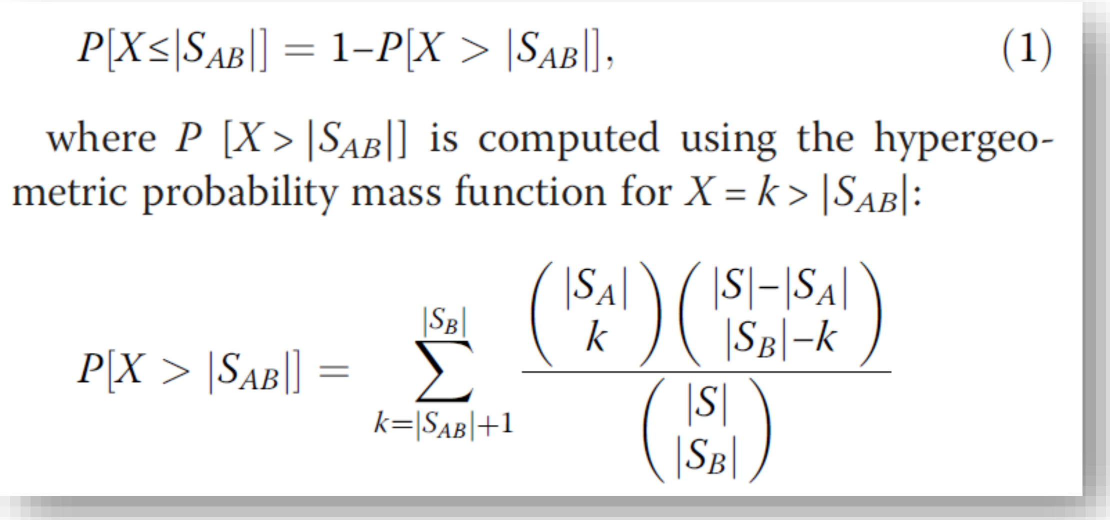
Srihari et al. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. Biology Direct, 10:57, 2015.

Synthetic lethal pairs

- **Fact**
  - When a pair of genes are synthetic lethal, mutations that affect function 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

- **Why interested in synthetic lethality**
  - Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets

Discussion #8



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

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

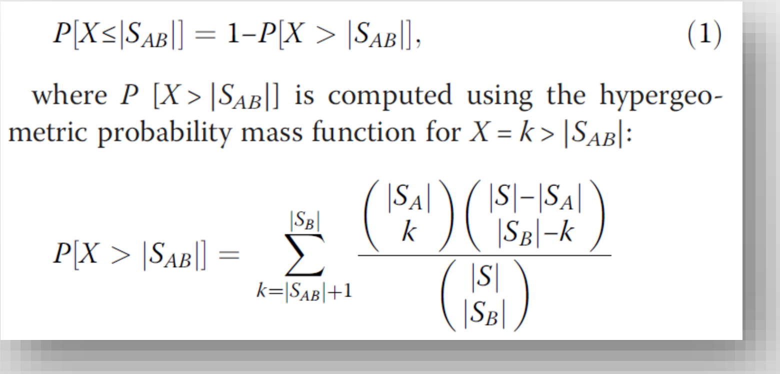
$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

S_{AB} is # of subjects in whom both A and B are mutated

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

- **Anything wrong with this?**

The hypergeometric distribution does not reflect real world mutations



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

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

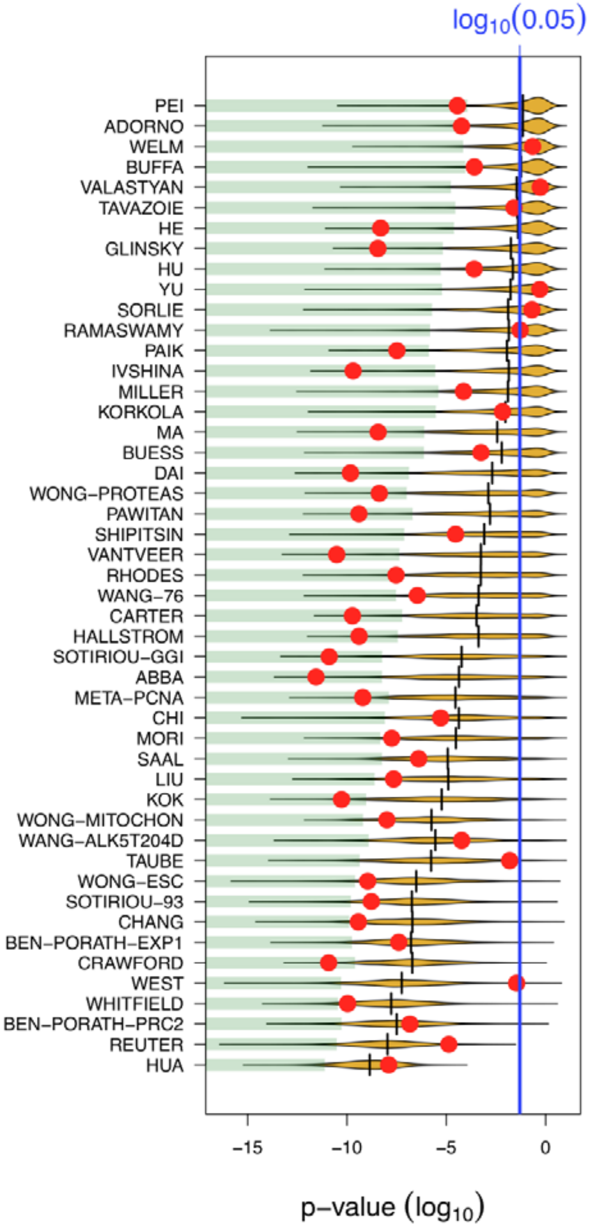
$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

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

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

⇒ **Null distribution is not hypergeometric, binomial, etc.**

SOMETIMES CHANGING PERSPECTIVE HELPS



Venet et al., PLOS Comput Biol, 2011

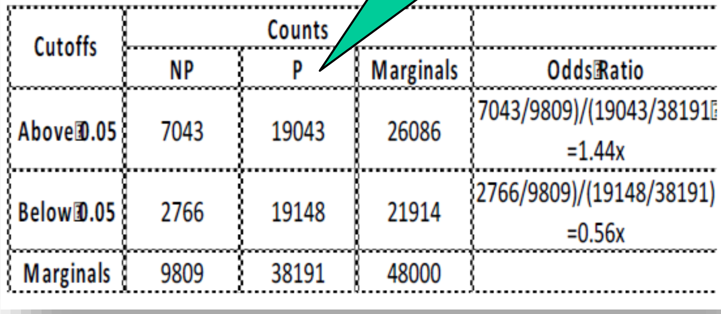
Almost all random signatures also have p-value < 0.05

- **Instead of asking whether a signature is significant, ask what makes a signature (random or otherwise) significant**

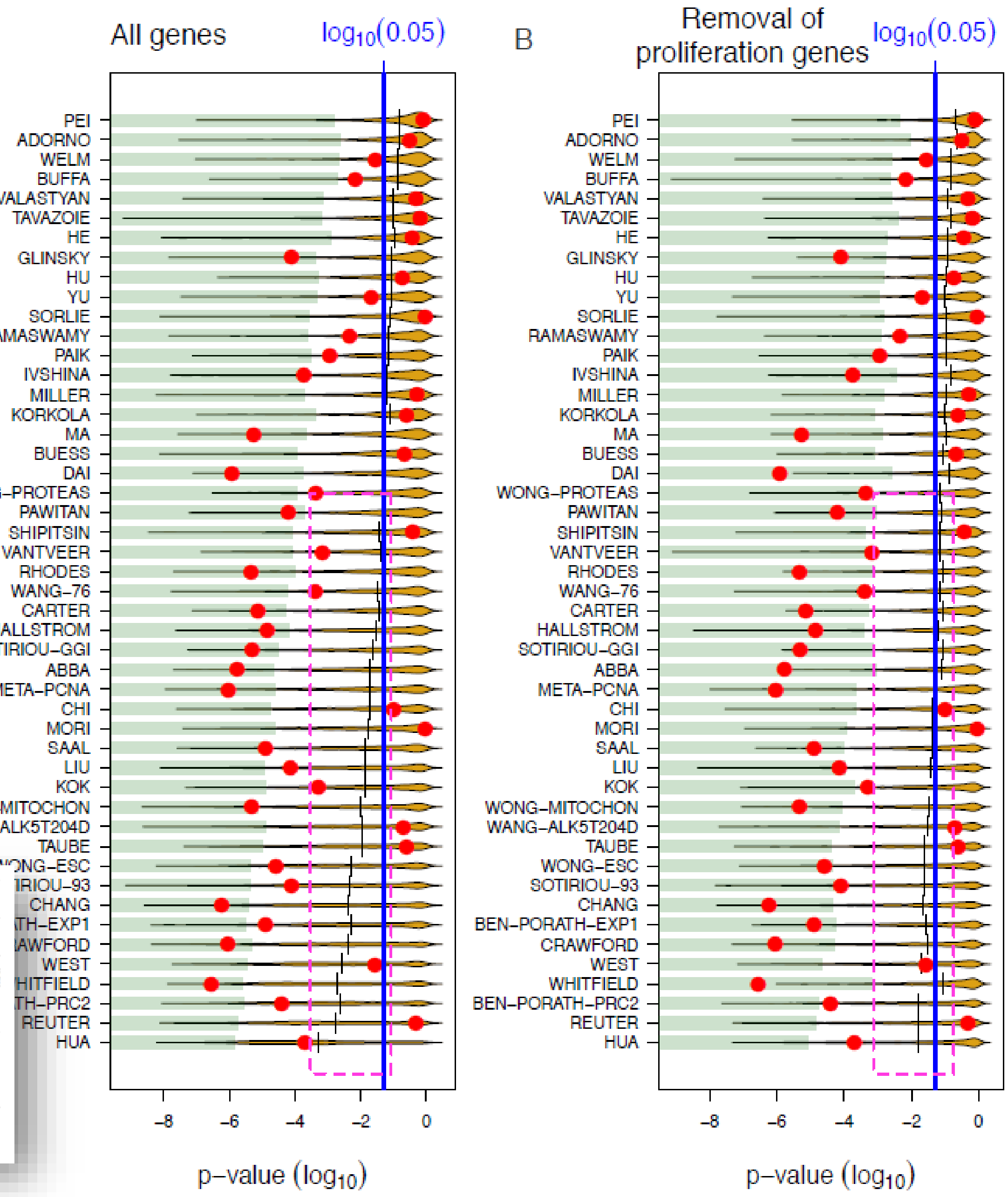
Wilson Goh, private communication, 2017

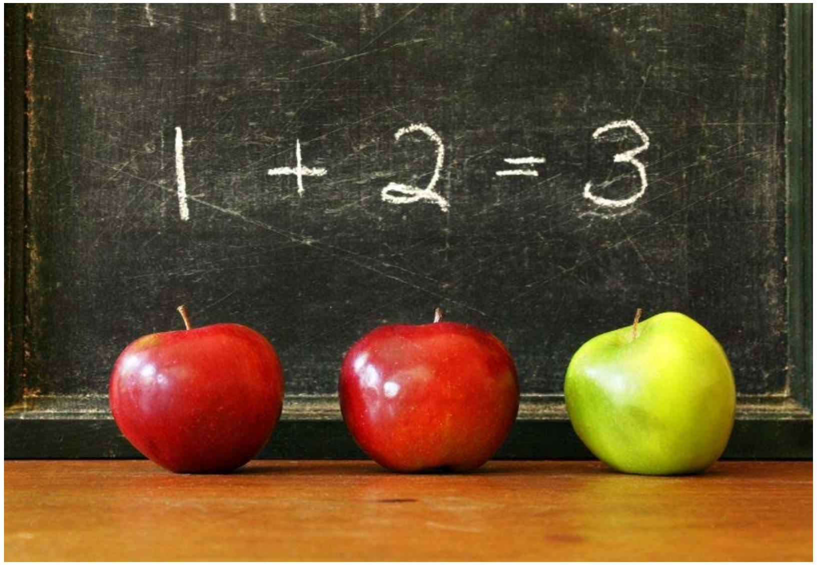
- **Proliferation is a hallmark of cancer**
- **Hypothesis: proliferation-associated genes make a signature significant**

# of random signatures w/ ≥1 prolif gene



Cutoffs	Counts		Marginals	Odds Ratio
	NP	P		
Above 0.05	7043	19043	26086	7043/9809)/(19043/38191 =1.44x
Below 0.05	2766	19148	21914	2766/9809)/(19148/38191) =0.56x
Marginals	9809	38191	48000	





SUMMARY

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**
  - Can you derive it from the null hypothesis?