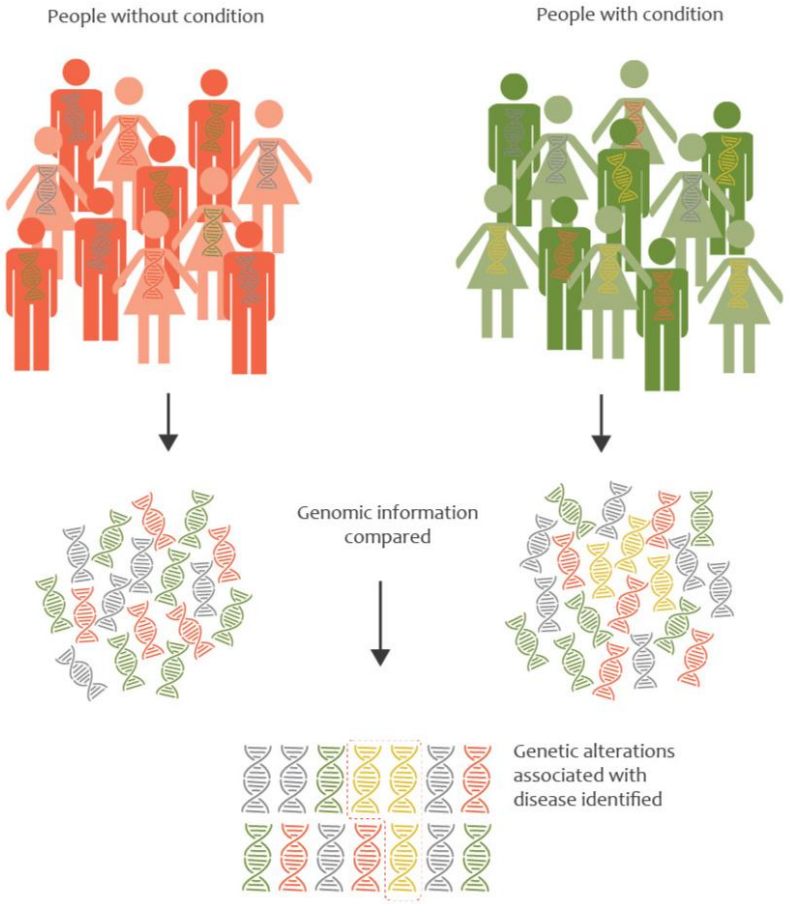
Quantitative genetics

- Genome-wide association studies
  - Complex traits
  - Statistical inference, P-values, & Multiple hypothesis testing
  - Regularized linear regression
  - Polygenic risk score

Complex traits and diseases

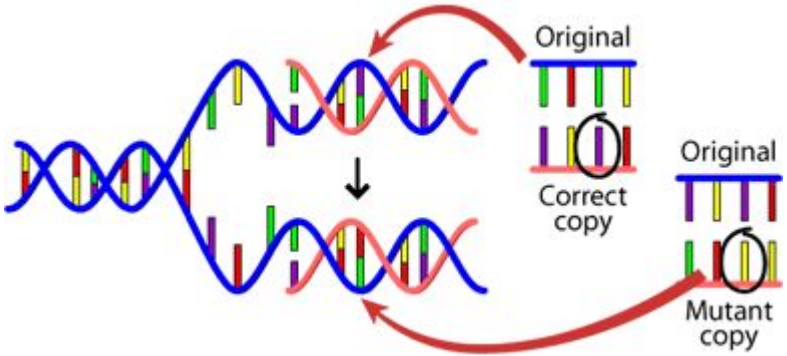


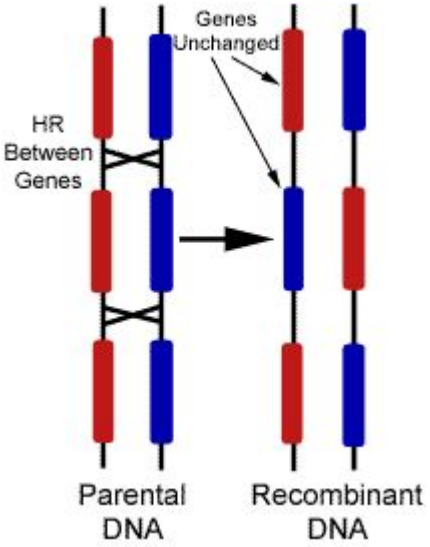
What factors contribute to a particular trait or the risk of getting a particular disease?

- Genetic factors (numerous)
- Other biological factors: e.g. age, sex
- Environmental factors: e.g. geography, nutrition, toxins
- Interaction between genome and environment
  - Phenotypic Variation = G + E + GxE

How do you quantify how much the genome actually contributes?

Genetic variation





Single Nucleotide Polymorphisms (SNPs)

Small Insertions

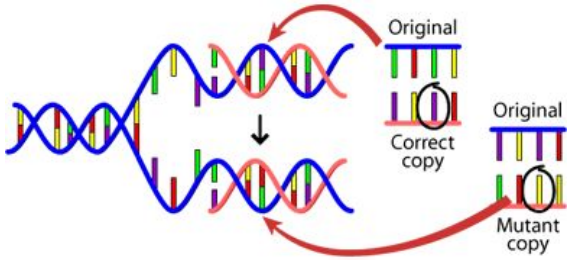
Small Deletions

Copy Number Variants (CNVs)

- Duplications & deletions

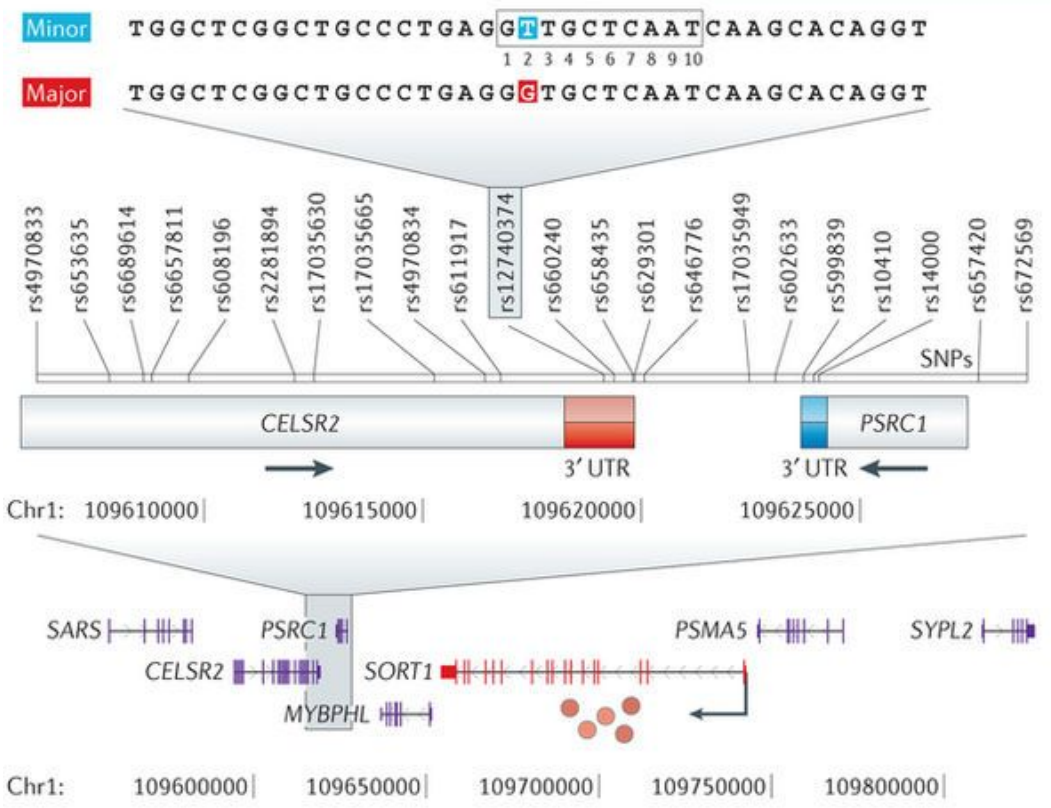
Still expensive to sequence entire genome.

Single Nucleotide Polymorphisms (SNPs)



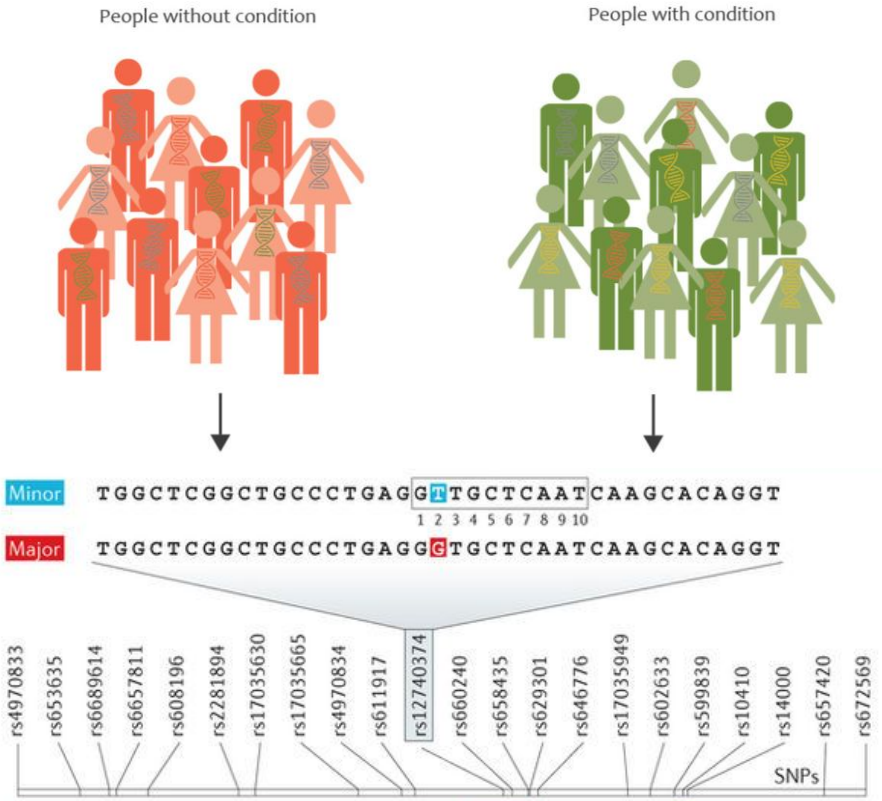
Focus only on commons SNPs that might contribute to variation:

About 5–10 million SNPs; a small part of the human genome.



Albert & Kruglyak (2015) Nat. Rev. Genet.

Genome-wide Association Study (GWAS)

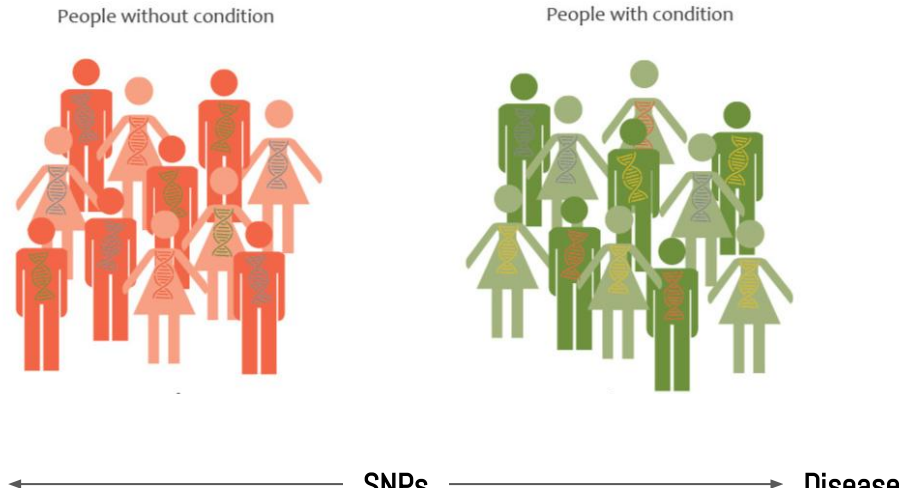


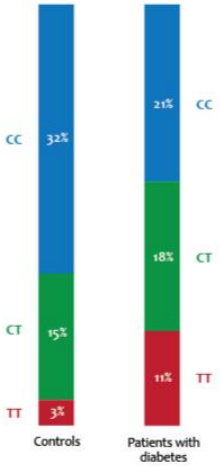
Single Nucleotide Polymorphisms (SNPs)

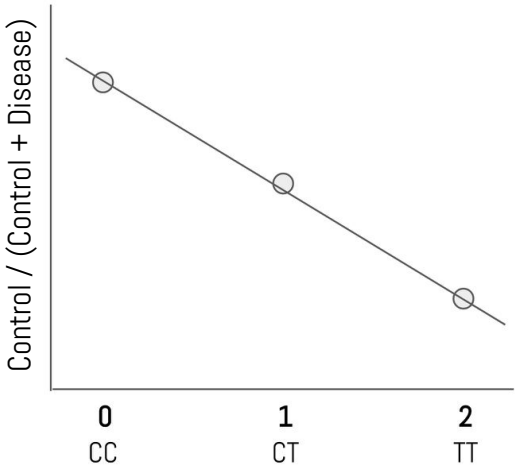
SNP array: a small chip that has DNA probes complementary to regions in the genome that have SNPs.

Visscher (2017) AJHG; @genomicsedu

Genome-wide Association Study (GWAS)







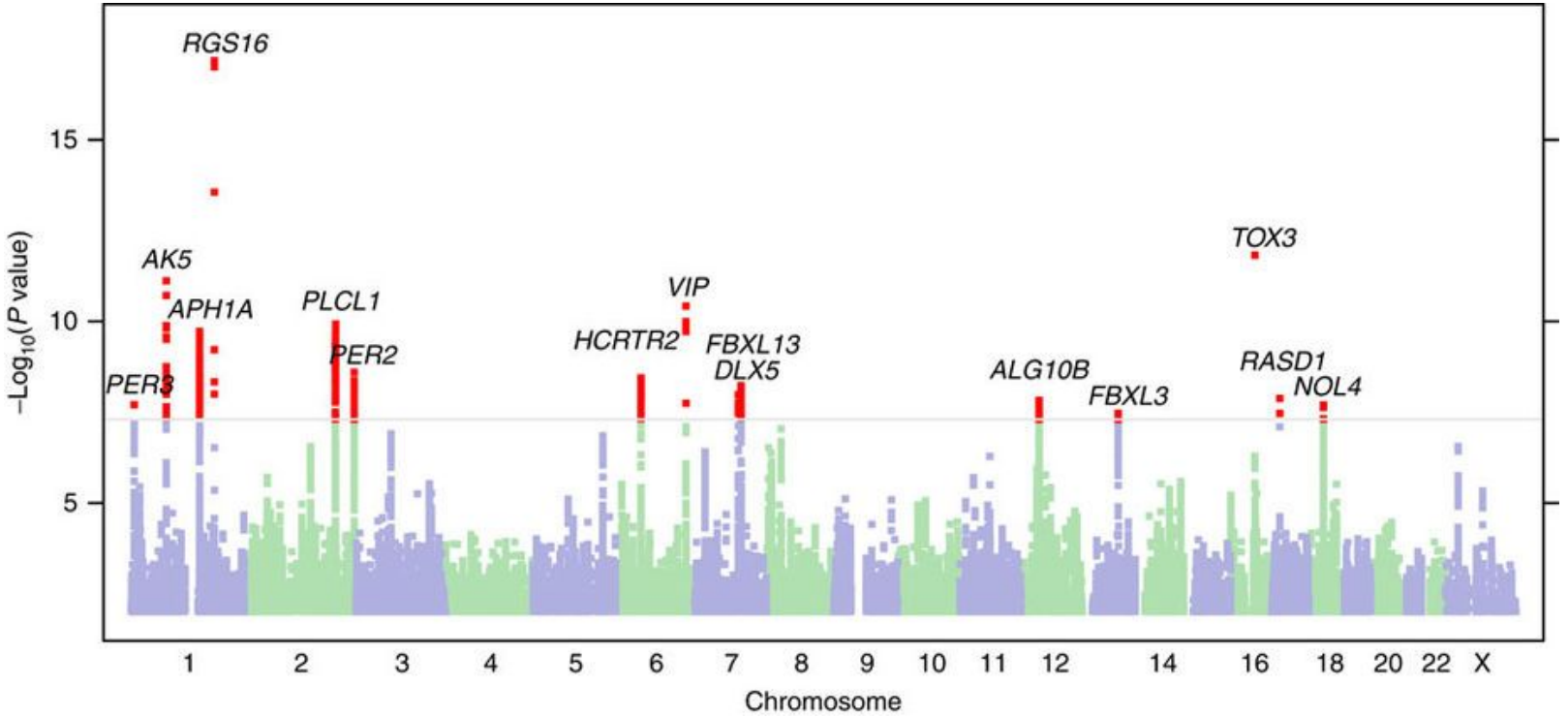
A C/T SNP from a GWAS for type 2 diabetes

- Increase in freq of T allele in individuals with diabetes compared to controls.
- We know where this SNP is on the genome → study surrounding sequence

Visscher (2017) AJHG; @genomicsedu

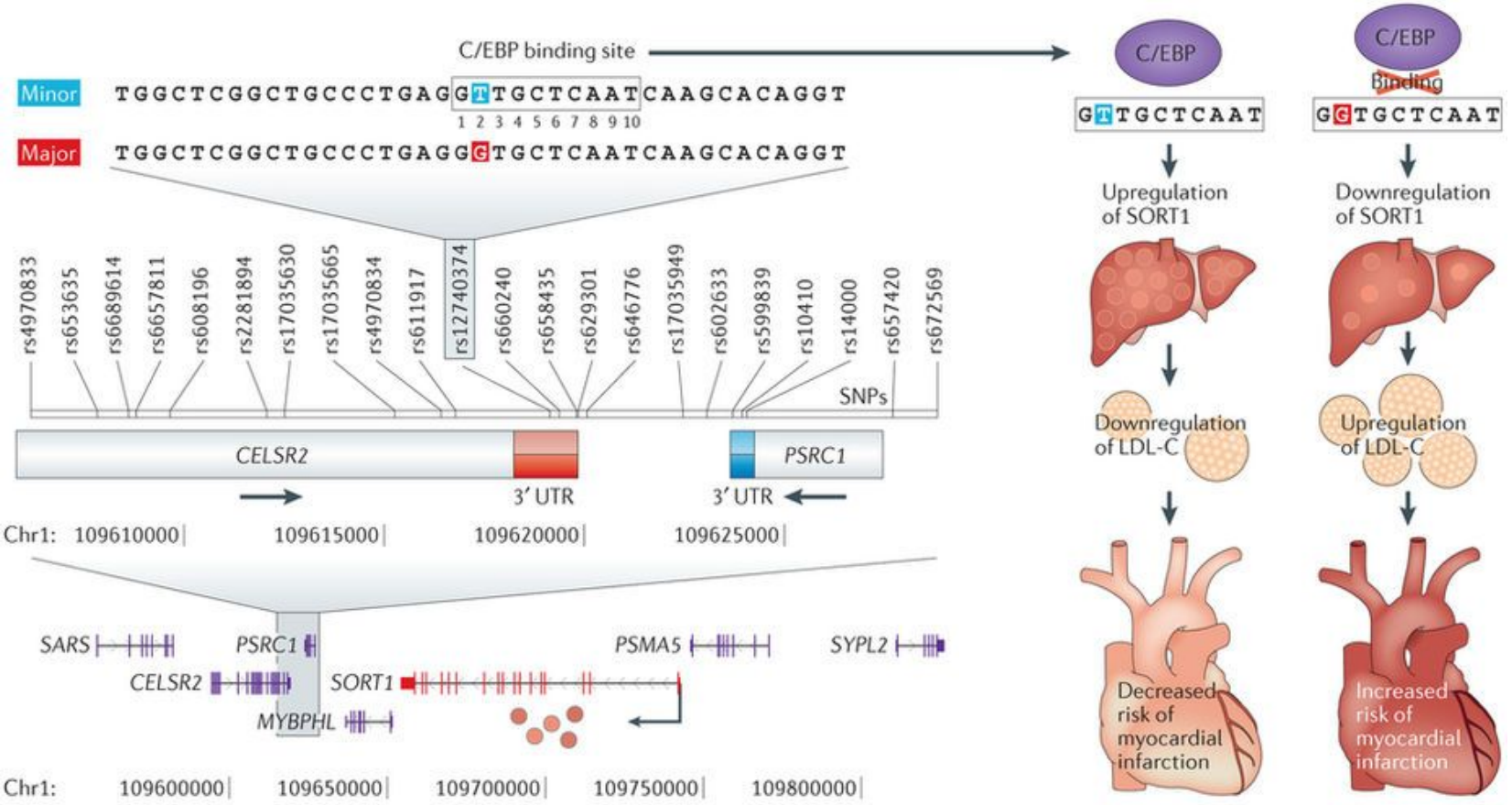
Results of a GWAS

GWAS of 89,283 individuals identifies genetic variants associated with… being a morning person!



Hu (2016) Nat. Comm.

GWAS – Examples



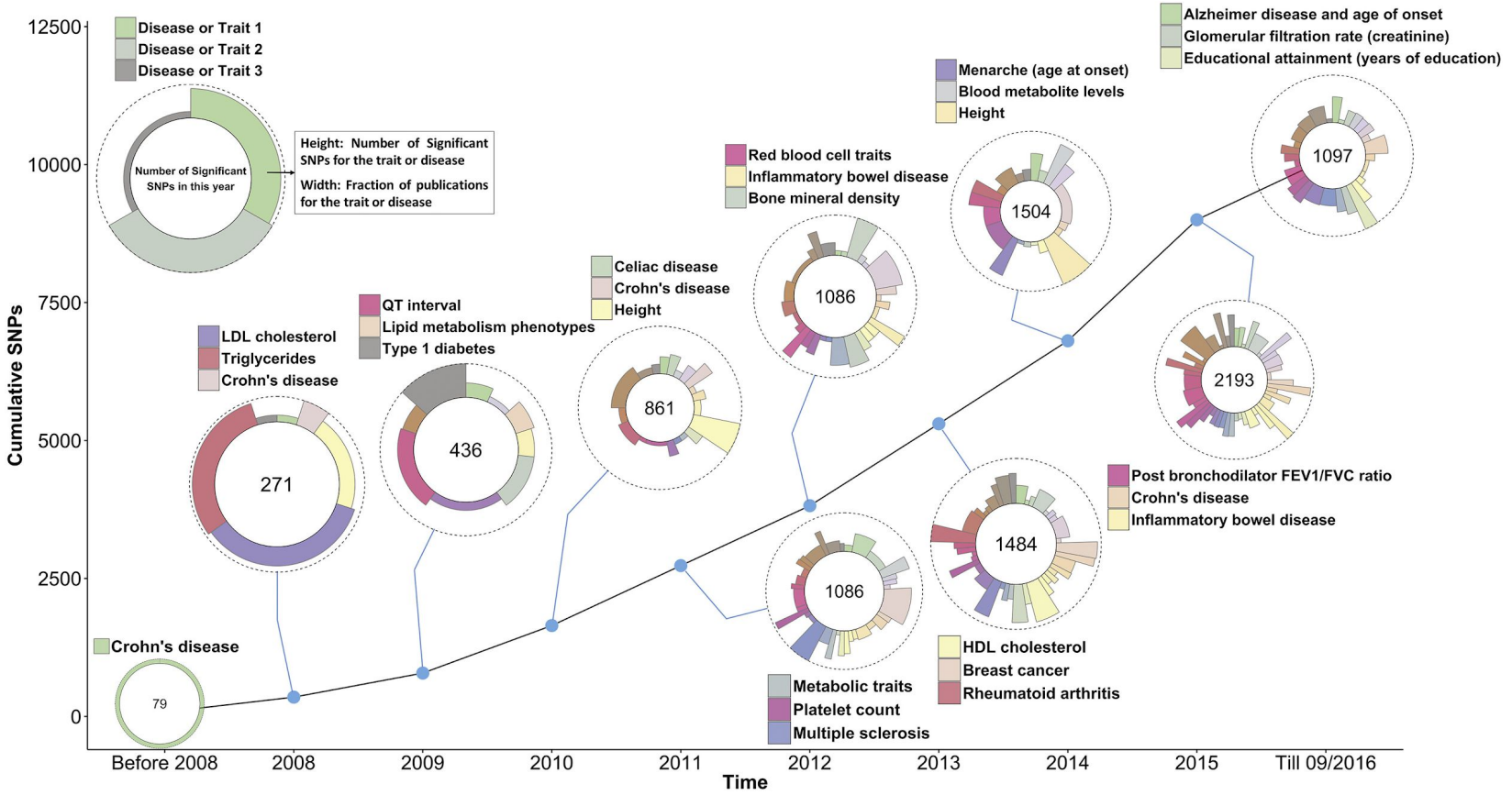
Albert & Kruglyak (2015) Nat. Rev. Genet.

GWAS – Examples

Variation in the **nicotinic receptor** leads to higher levels of **lung cancer** in the developed world.

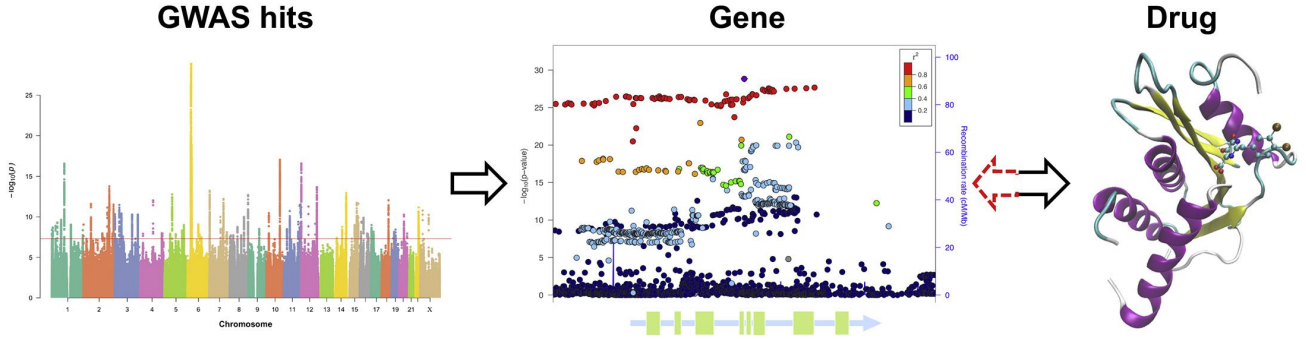
- This is *not* because the nicotinic receptor is directly involved in the molecular aspects of lung cancer.
  - Rather, these variants make people get a bigger hit from nicotine.
  - I.e. *if* they start smoking, they are less likely to *stop* smoking (more smoke exposure).
- So, this variant is causally involved in lung cancer.
  - I.e. if one has the variant, their odds of developing cancer are fundamentally higher.
- However, the mechanism will not be clear if we didn't know about nicotine from other studies.
- Smoking exposure is the **main cause** & this variant in the nicotine receptor is a **modifier**.

GWAS – Timeline of discoveries



Visscher (2017) AJHG

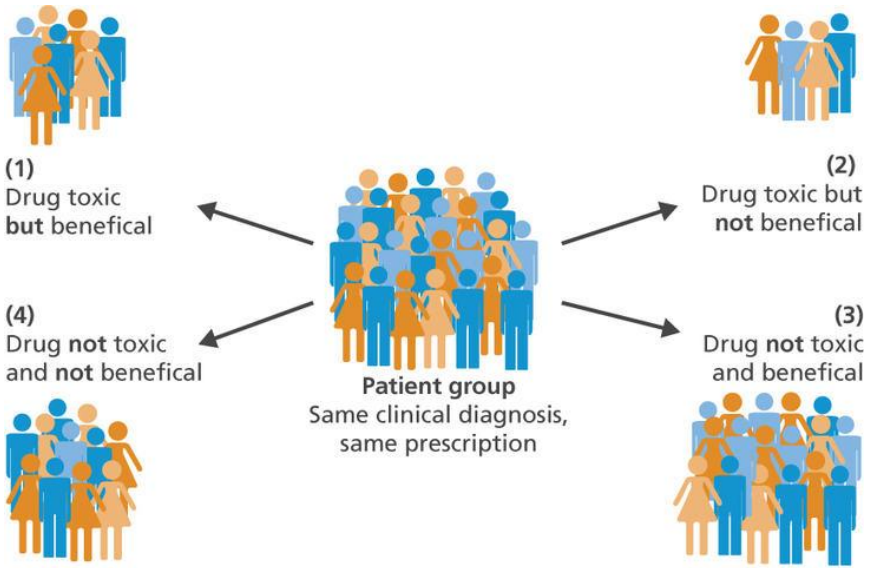
GWAS to drugs

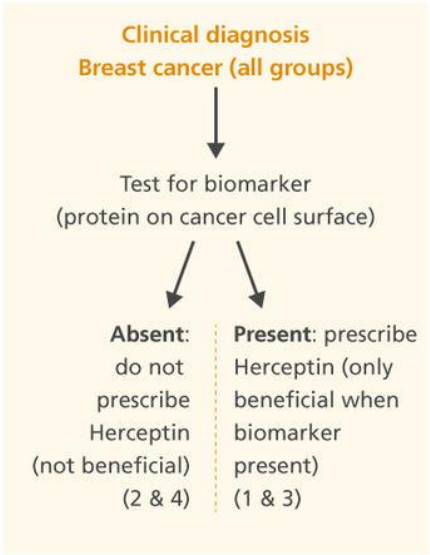


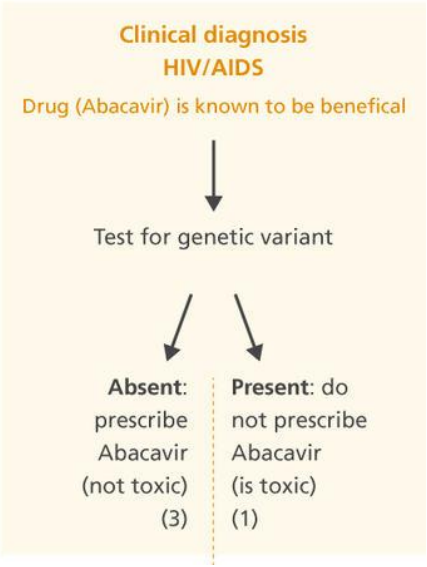
Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

Visscher (2017) AJHG

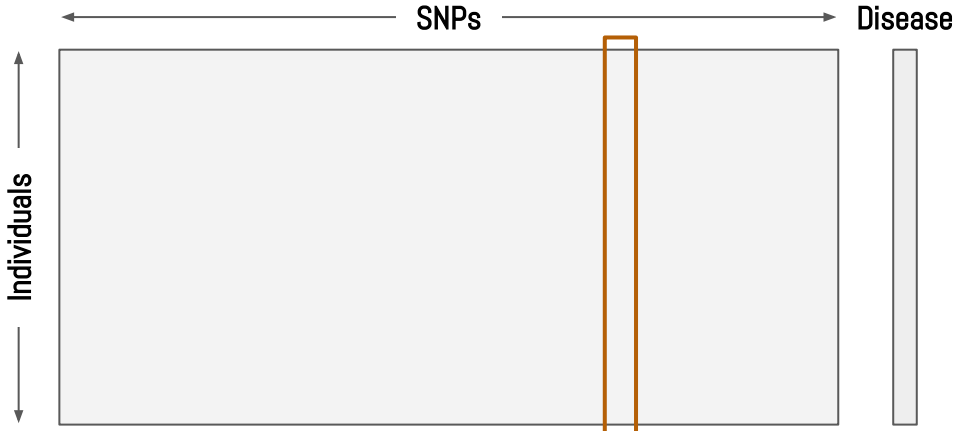
GWAS-like approaches – Pharmacogenomics

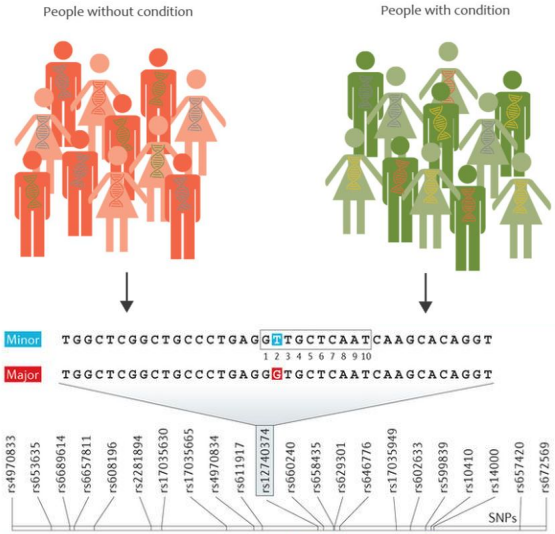


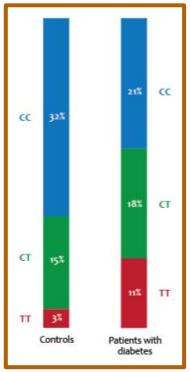




Statistical hypothesis testing for GWAS







	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

Pearson's χ^2 test

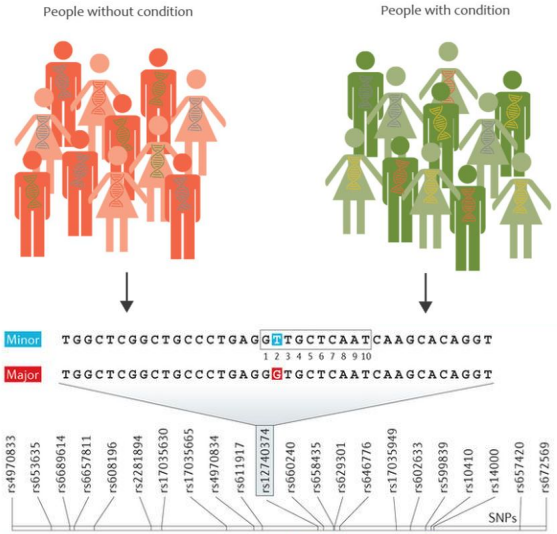
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

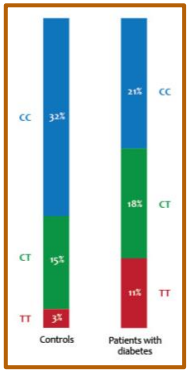
Statistical hypothesis testing for GWAS

Consider two competing hypotheses for a given SNP:

- **Null hypothesis**: the SNP in not associated with the phenotype.
- **Alternative hypothesis**: the SNP is associated with the phenotype.

There's always some difference ⟶ Is it significant difference?





	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

Pearson's χ^2 test

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Statistical hypothesis testing for GWAS

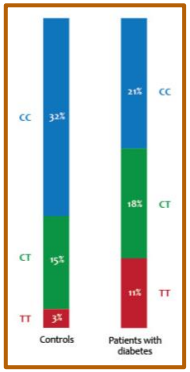
Consider two competing hypotheses for a given SNP:

- **Null hypothesis**: the SNP in not associated with the phenotype.
- **Alternative hypothesis**: the SNP is associated with the phenotype.

How is this question typically answered?

There's always some difference ⟶ Is it significant difference?

Perform hypothesis testing & calculate a p-value



	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0\cdot}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1\cdot}$
Totals	$O_{\cdot 0}$	$O_{\cdot 1}$	$O_{\cdot 2}$	S

Pearson's χ^2 test

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

What is the P-value?

Hypothesis: If **a SNP** is **associated** with a **trait/disease**.

Given this context, which of the following defines **what is P-value**?

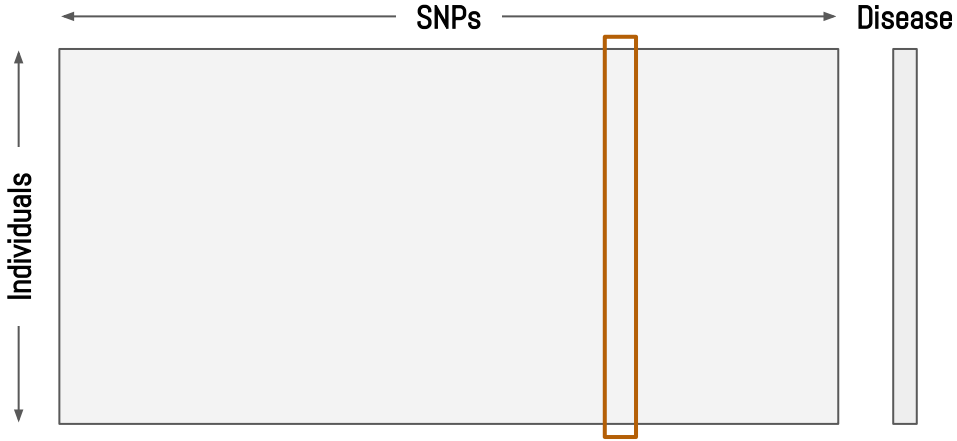
1. The **amount of evidence** that the SNP is associated with the disease.
2. The **strength of the effect** of the SNP on the disease.
3. One minus the probability that the **outcome** of the study **is important**.
4. One minus the probability that the result can be **replicated**.
5. The probability that the selected **SNP is not associated** with the disease.

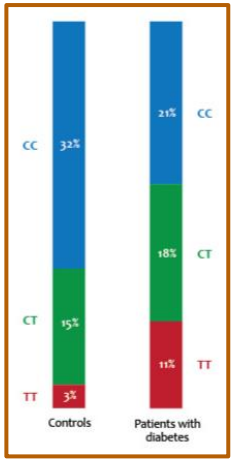
How unlikely is the observed outcome?

There is always some difference between the groups.

I.e, the χ^2 is going to be > 0.

But, how unlikely is the observed difference?





	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0\cdot}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1\cdot}$
Totals	$O_{\cdot 0}$	$O_{\cdot 1}$	$O_{\cdot 2}$	S

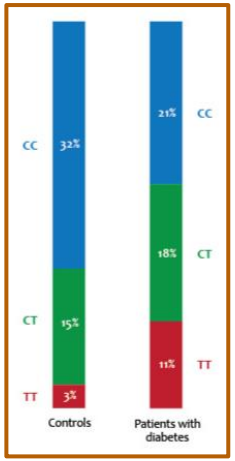
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

How unlikely is the observed outcome?

Null hypothesis: The SNP is not associated with the disease.

- Even if the SNP is not associated, we'll observe as much or more difference in the SNP frequencies across the disease groups. (I.e., equal or higher χ^2.)
- Irrespective of the group each individual is a part of, we'll observe as much or more difference b/w the observed and expected frequencies.

1. Calculate the real test statistic.
2. Repeat the following 100,000 times to set up the null hypothesis for this test statistic:
   - Randomly assign individuals to groups.
   - Record the test statistic of the permuted assignments.
3. Calculate the p-value of the real test statistic. [**How?**]



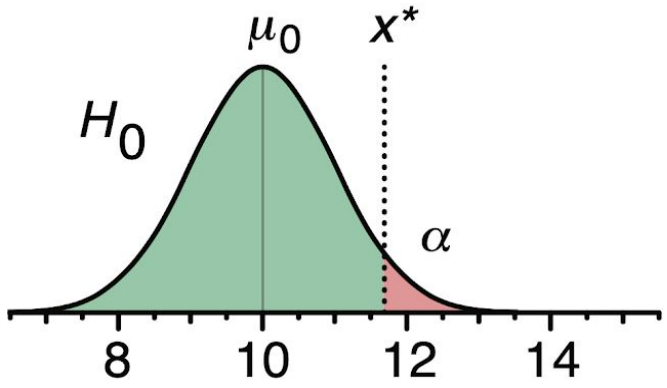
	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0\cdot}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1\cdot}$
Totals	$O_{\cdot 0}$	$O_{\cdot 1}$	$O_{\cdot 2}$	S

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

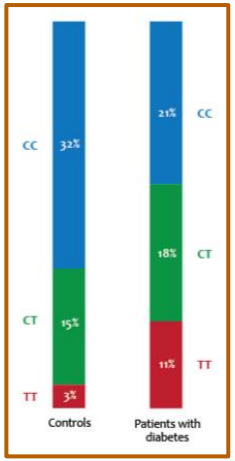
How to calculate the P-value?

Null hypothesis: The SNP is not associated with the disease.

- Even if the SNP is not associated, we'll observe as much or more difference in the SNP frequencies across the disease groups. (I.e., equal or higher χ^2.)
- Irrespective of the group each individual is a part of, we'll observe as much or more difference b/w the observed and expected frequencies.



Null distribution of the test statistic



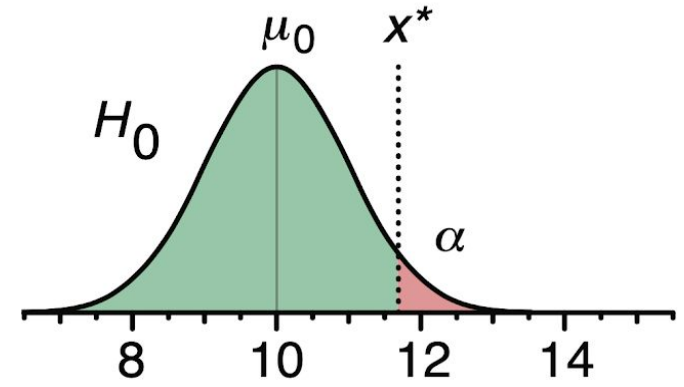
	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0\cdot}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1\cdot}$
Totals	$O_{\cdot 0}$	$O_{\cdot 1}$	$O_{\cdot 2}$	S

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

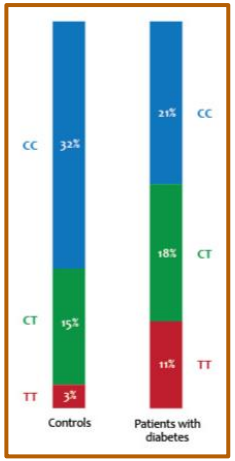
How to calculate the P-value?

The p-value is the probability that the study would have produced the observed outcome — or something more extreme — even if the SNP is not associated with the disease.

The p-value is the area under the null distribution corresponding to outcome equal to or more extreme than the observed statistic.



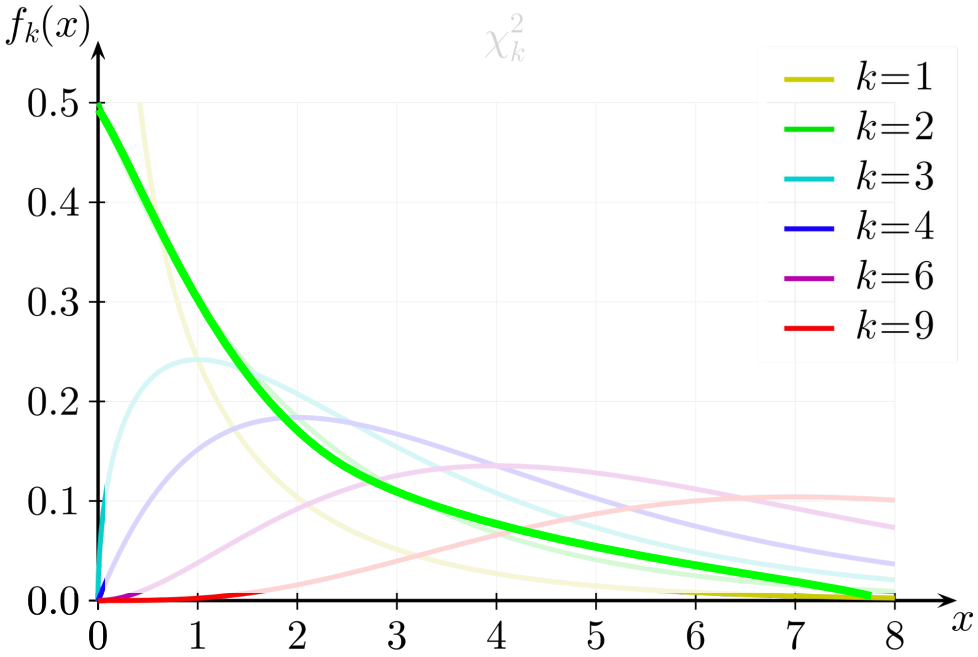
Null distribution of the test statistic



	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0\cdot}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1\cdot}$
Totals	$O_{\cdot 0}$	$O_{\cdot 1}$	$O_{\cdot 2}$	S

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

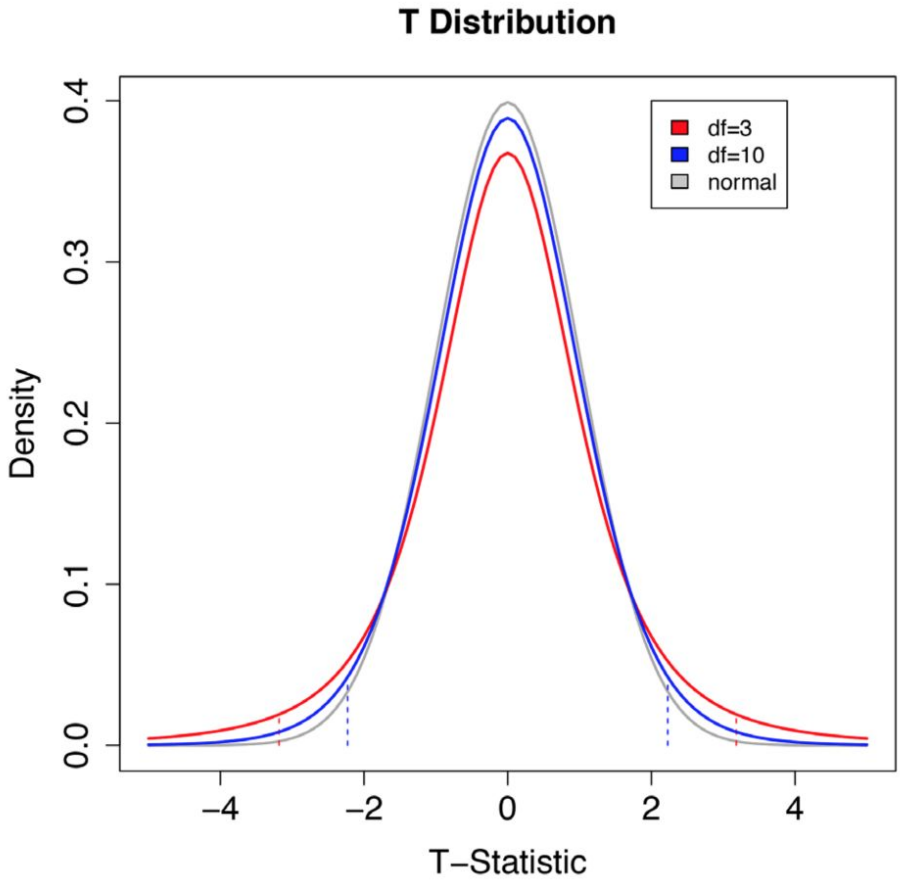
How to calculate the P-value?



The p-value is the area under the null distribution corresponding to outcome equal to or more extreme than the observed statistic.

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

How to calculate the P-value?



The p-value is the area under the null distribution corresponding to outcome equal to or more extreme than the observed statistic.

Student's one-sample test

$$t = \frac{\bar{x} - \mu_0}{\text{SEM}}$$

Welch's two-sample test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

P-value - History

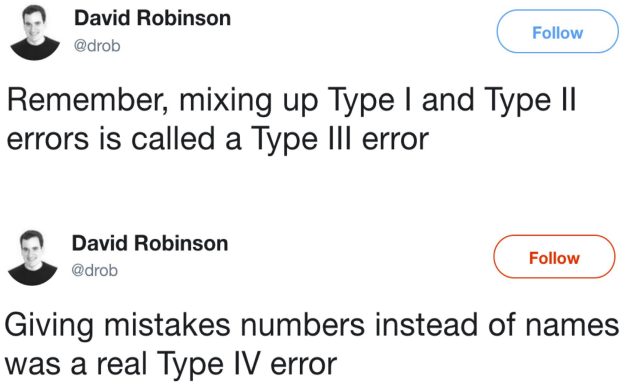
- Fisher (1920s):
  - Informal method to help interpret the data along with prior experience, domain knowledge, size of the effect, etc.

- Neyman & Pearson:
  - Control false positive rate at α, set by the experimenter based on what can be tolerated.
  - Formulate null and alternative hypothesis.
  - Reject null when $p < \alpha$.
    - The threshold $\alpha = 0.05$ is merely a convention.

Type I & type II errors

Choosing $p < \alpha$ controls Type I error at α.

- Type I error: False-positive rate (α)
- Type II error: False-negative rate (β)
- Remember the story of the boy that cried wolf!

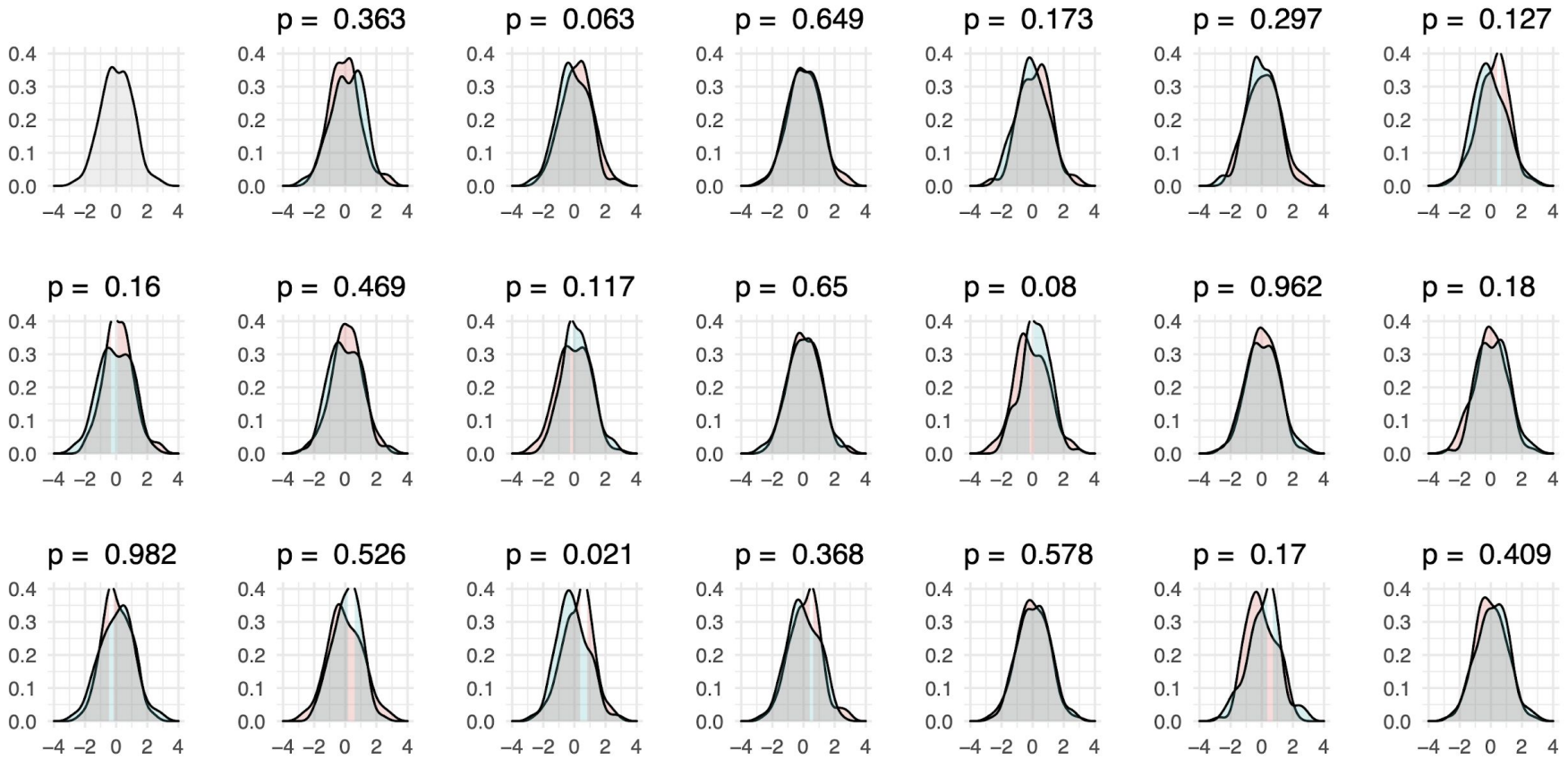


P-value depends on multiple factors

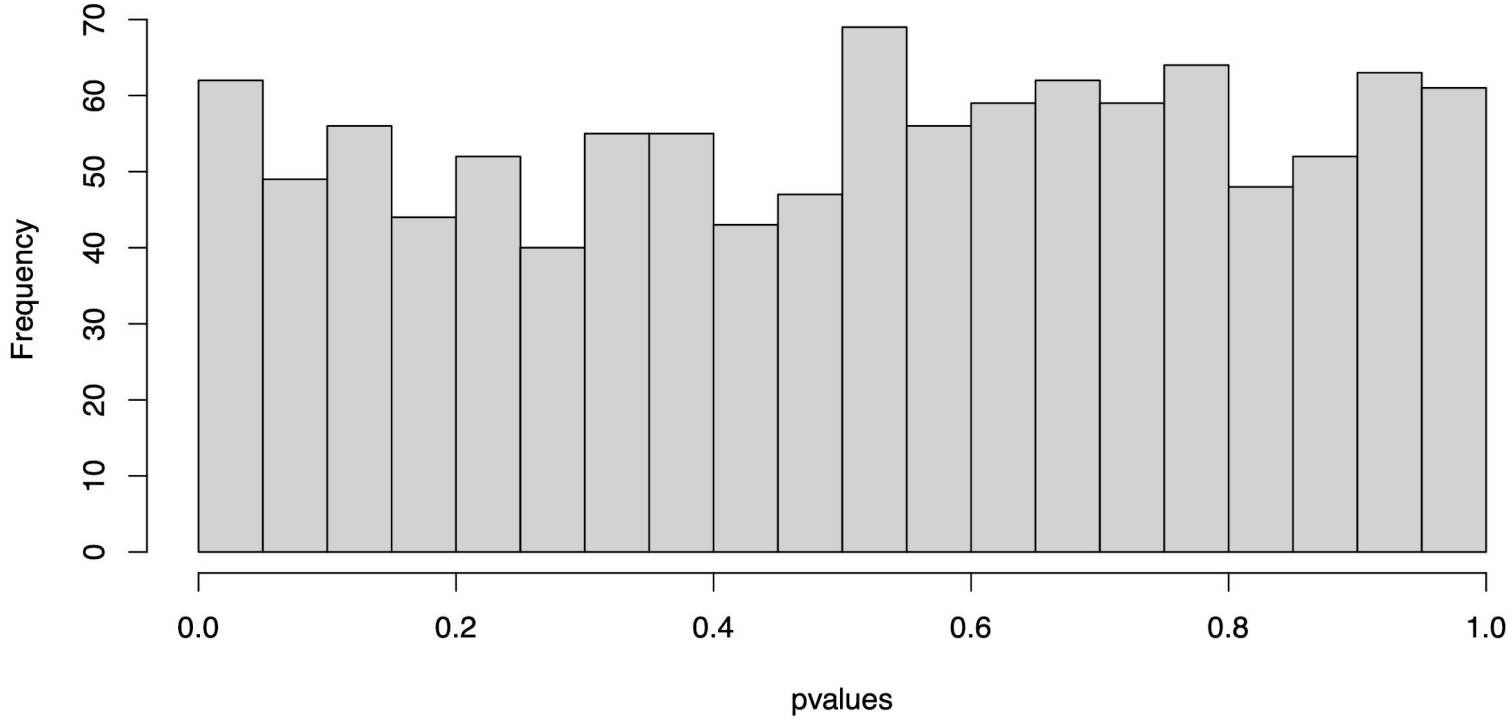
P-values are dependent on:

- Size of the effect (effect size)
- Variance within each group
- Sample size
- The underlying experimental design & the null hypothesis (need not always be random chance).
    - Conversely, two completely different experiments can give same data but end up very different p-values.
        - 3 out of 9: Binomial p-value = 0.073
        - 3 out of 9: Neg. Binomial p-value = 0.033.

Distribution of p-values under the null hypothesis

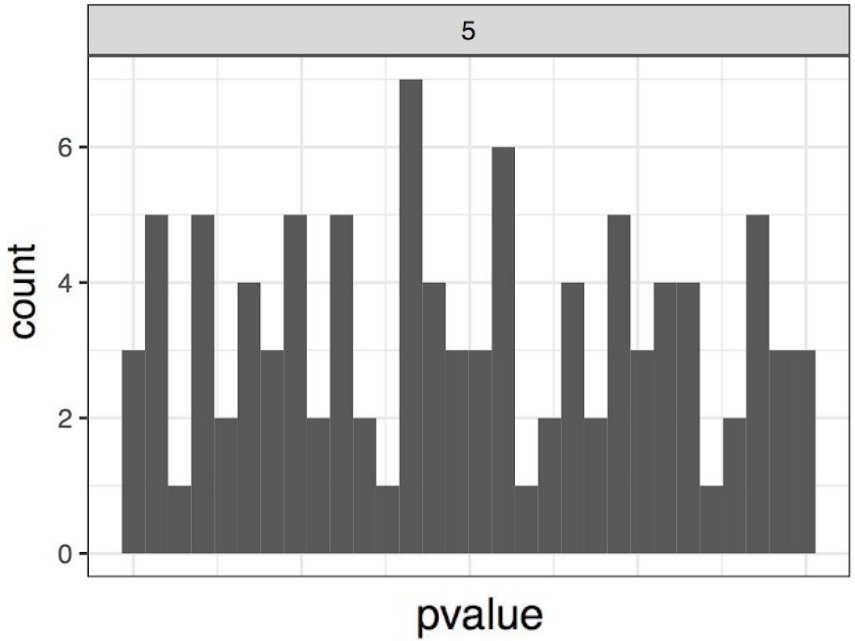


Distribution of p-values under the null hypothesis



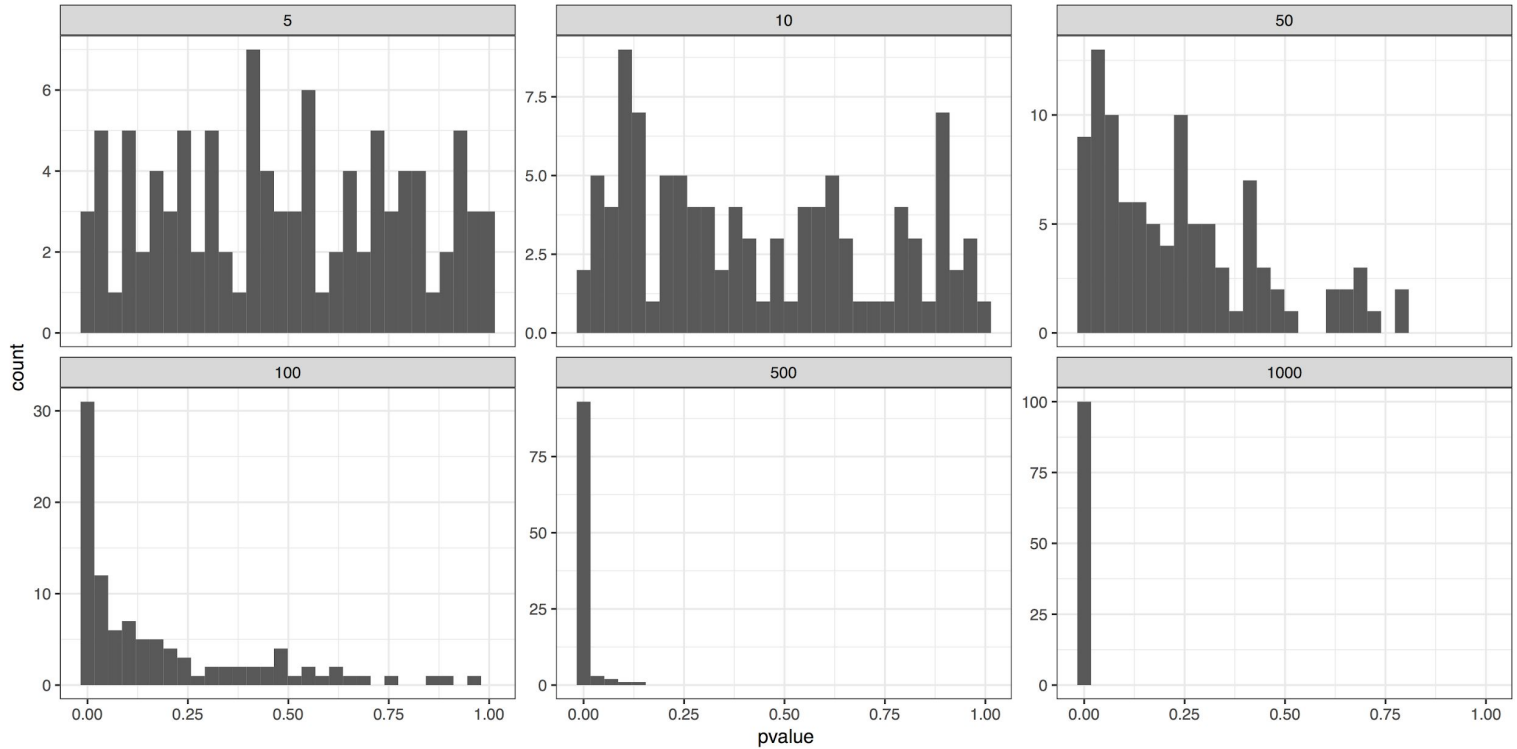
P-value depends on multiple factors

- P-values are dependent on: sample_size (effect_size = 0.25, std_deviation = 1)



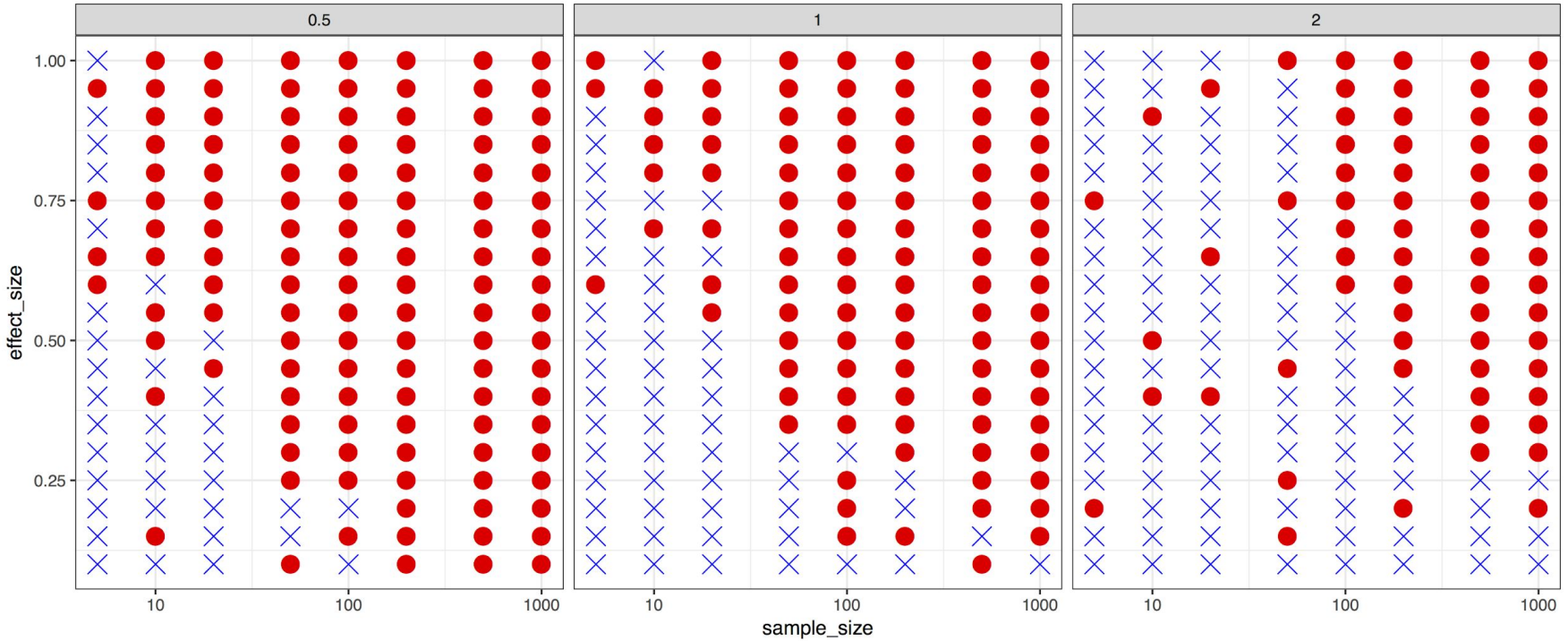
P-value depends on multiple factors

- P-values are dependent on: sample_size (effect_size = 0.25, std_deviation = 1)



P-value depends on multiple factors

- P-values are dependent on: sample_size, effect_size, within-group variance



P-value – Significant or not?

This list is culled from peer-reviewed journal articles in which:

a) the authors set themselves the threshold of 0.05 for significance,

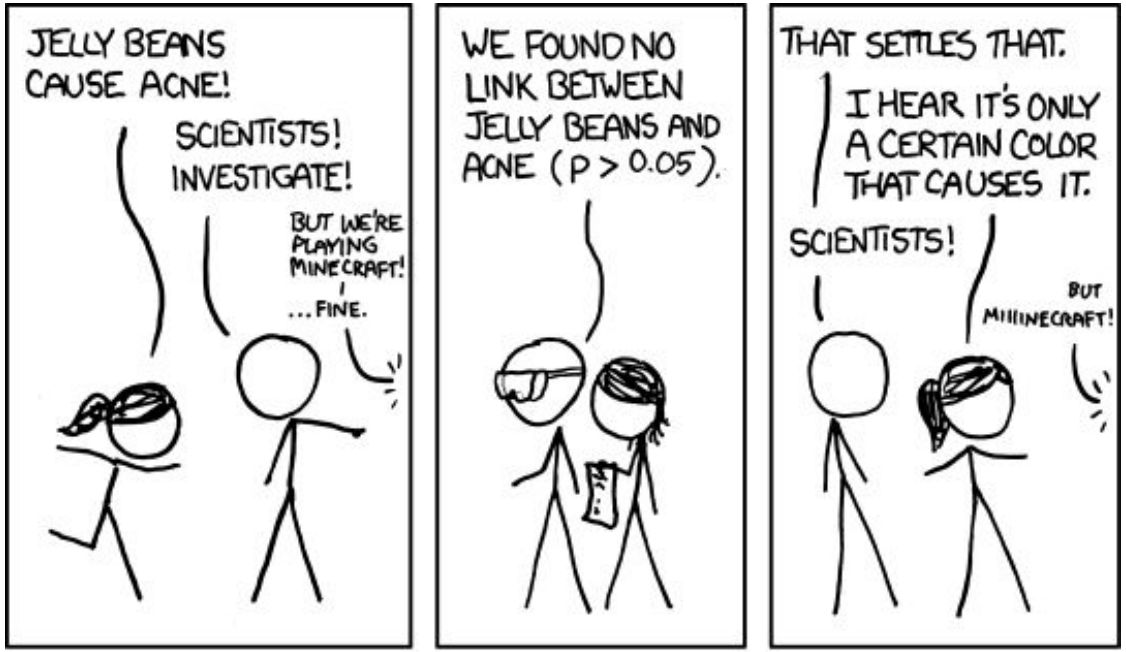
b) failed to achieve that threshold value for p and

c) described it in such a way as to make it seem more interesting.

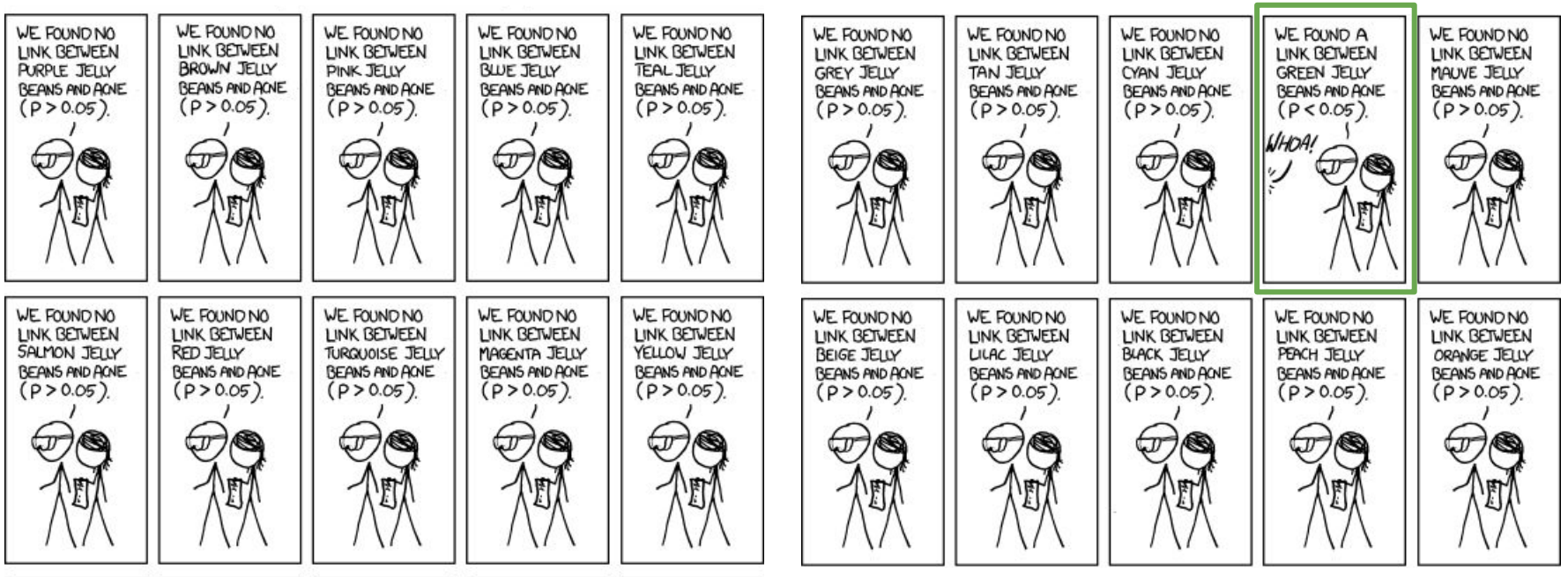
(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable trend toward significance (p=0.069)
a decreasing trend (p=0.09)
a definite trend (p=0.08)
a distinct trend toward significance (p=0.07)
a favorable trend (p=0.09)

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

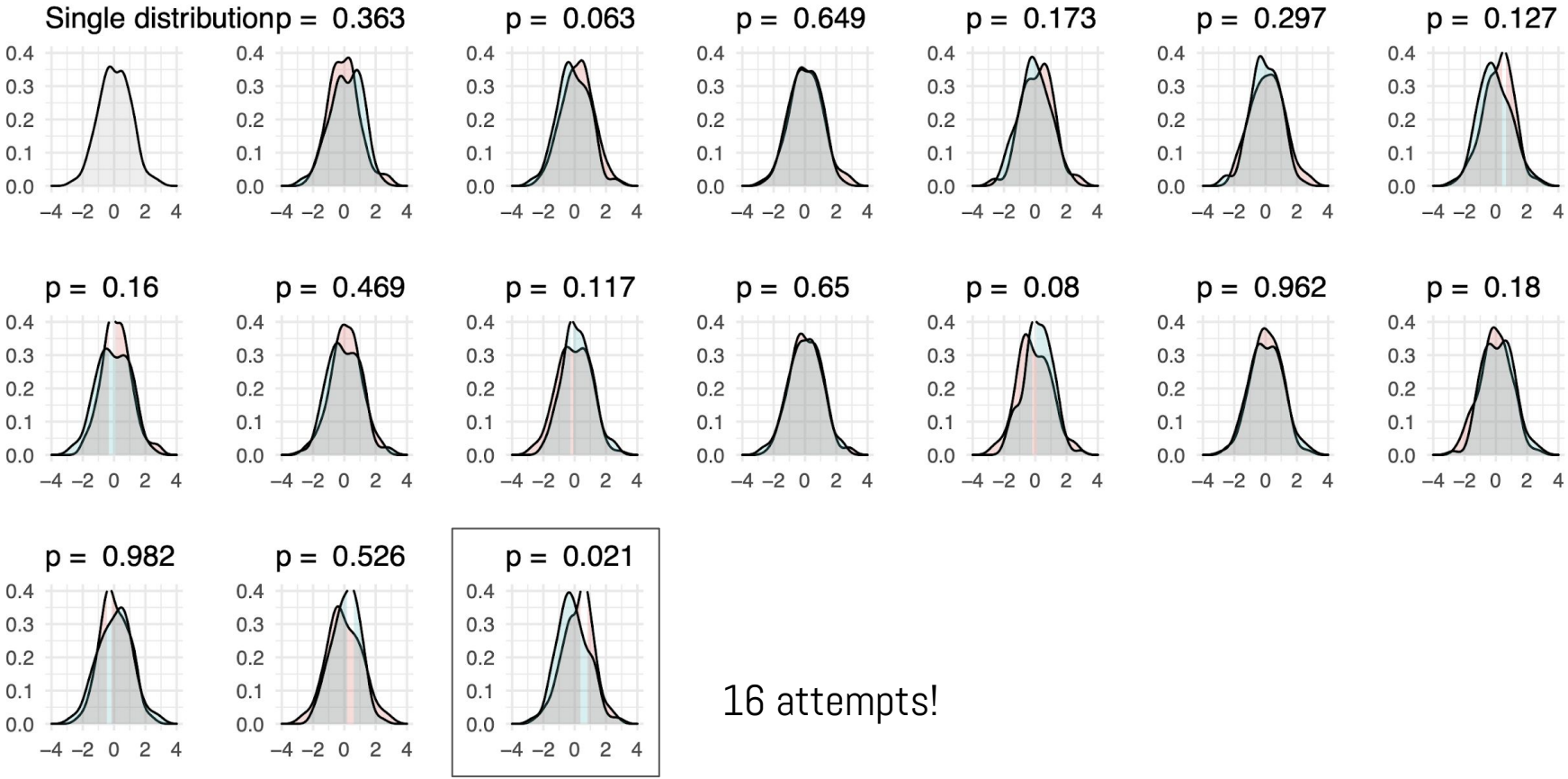
Multiple hypothesis testing



Multiple hypothesis testing

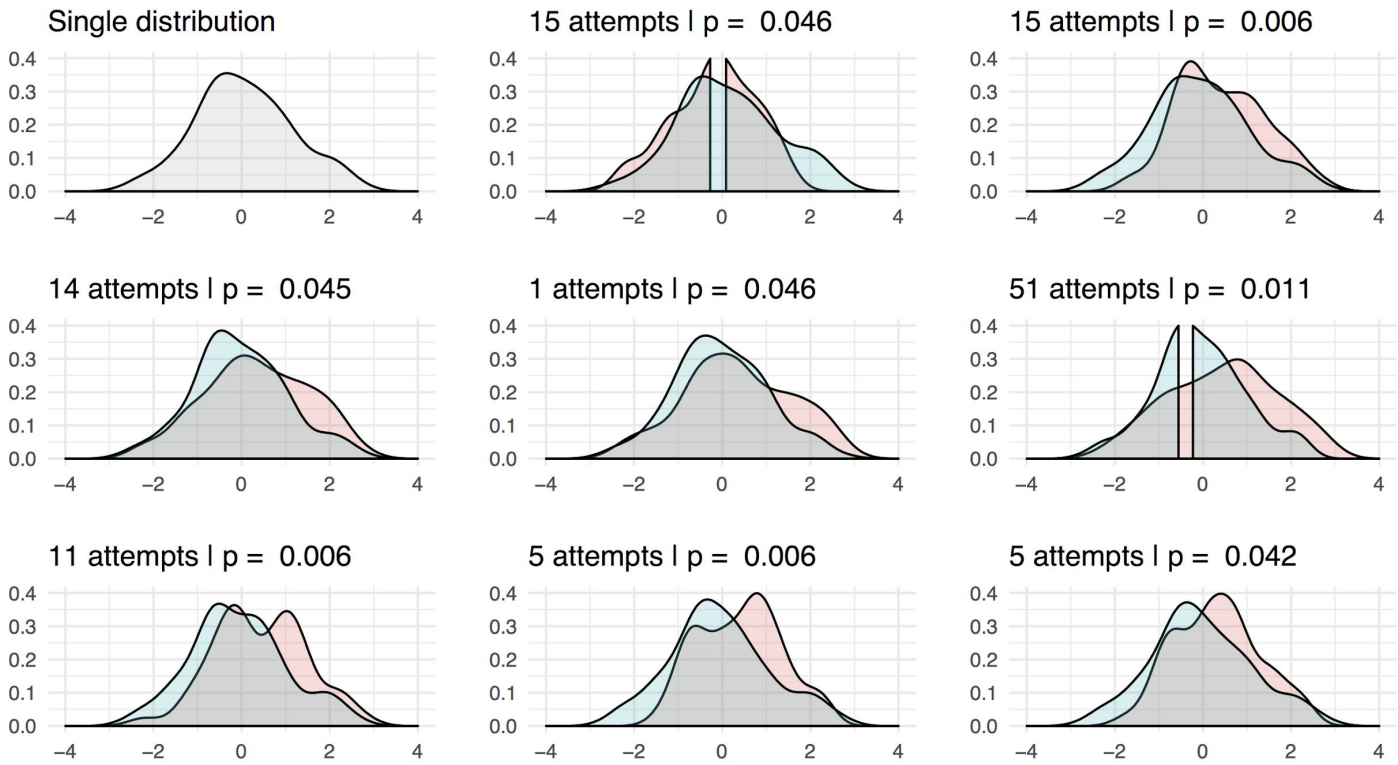


Multiple hypothesis testing



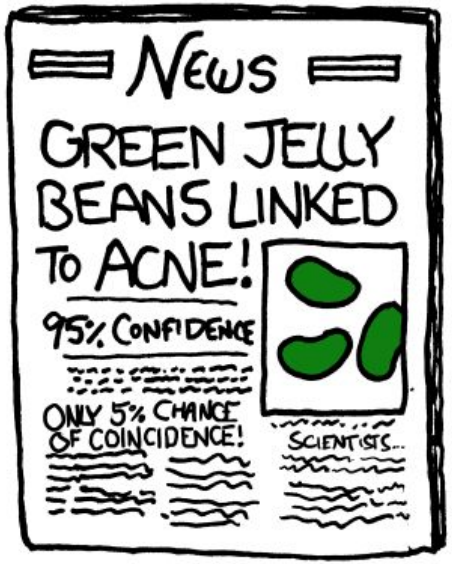
16 attempts!

Multiple hypothesis testing



"When a measure become a target, it ceases to be a good measure" – Strathern's Law

Multiple hypothesis testing



The more inferences are made, the more likely erroneous inferences are to occur.

Let α be the Type 1 error rate for a statistical test.

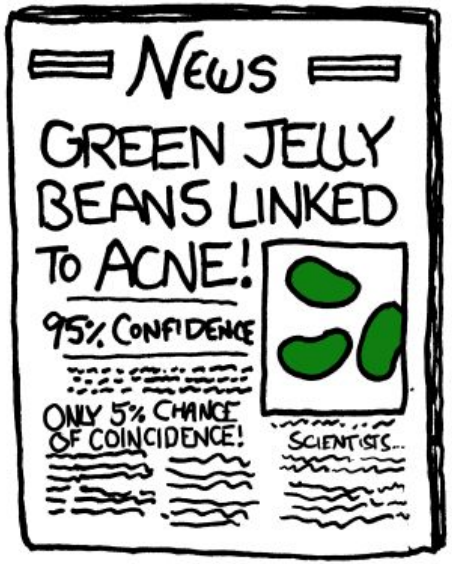
If the test is performed n times, what is the experimental-wise error rate α'? (Same as: What is the probability of obtaining at least 1 FP?)

$\alpha' = 1 - (1 - \alpha)^n$ (Check for $\alpha = 0.05$ & $n = 5$.)

The result may not be that significant even if its p-value $< \alpha$.

To solve this problem, the nominal p-value need to be corrected/adjusted.

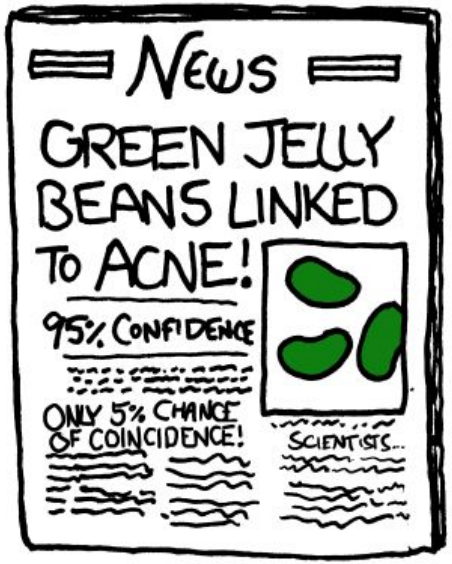
Correcting for multiple hypothesis testing



Controlling for **Family-wise Error Rate**
(FWER: the probability of at least 1 FP):

- Bonferroni correction:
  - $p'_i = p_i * n$ (permutation test)
- Permutation test:
  - Permute the data K times, each time calculate minimum p-value
  - p'_i = #{min_pvalue < p_i} / K

Correcting for multiple hypothesis testing



Controlling for **False Discovery Rate**
(FDR: proportion of FP among all significant hypotheses):

- Benjamini-Hochberg correction:
  - $p'_i = p_i * (n / i)$

FDR & FPR

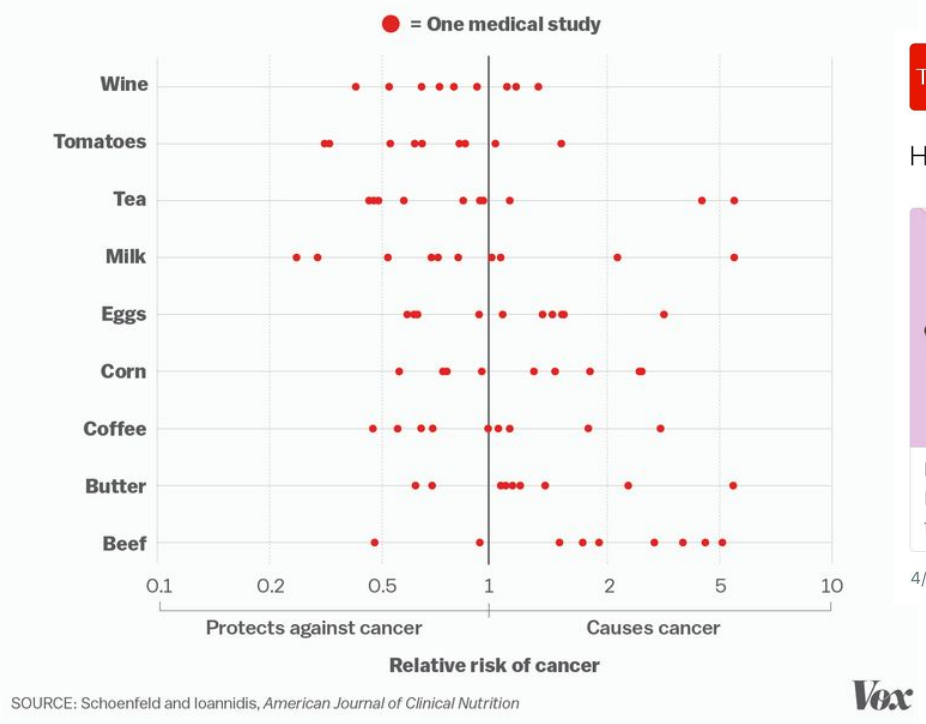
False Positive Rate: fraction of real –ves that are declared +ve.

False Discovery Rate: fraction of declared +ves that are –ve.

Expressing these quantities as conditional probabilities.

Multiple hypothesis testing

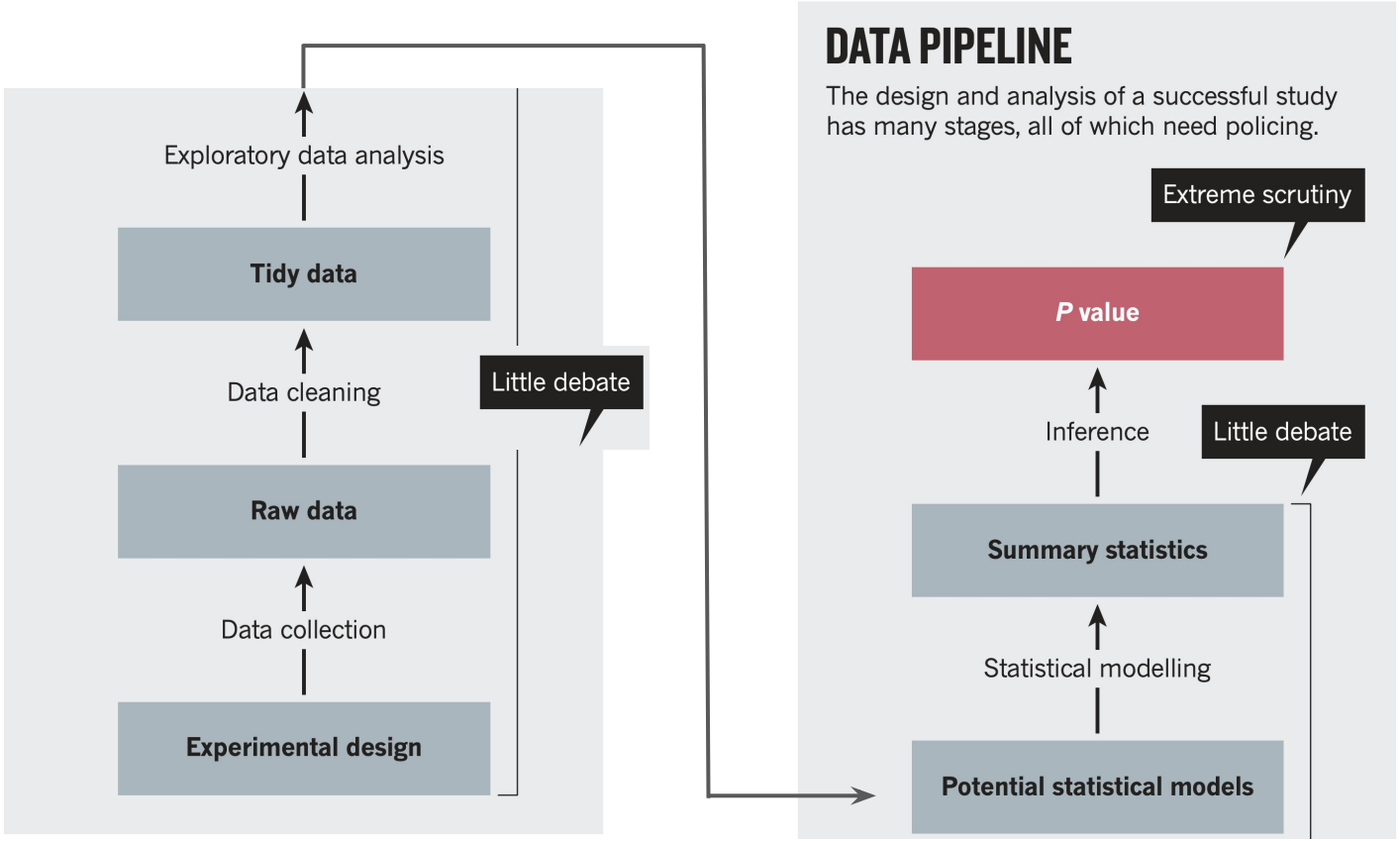
Publication bias (studies with nonsignificant results have lower publication rates)



Questionable research practices

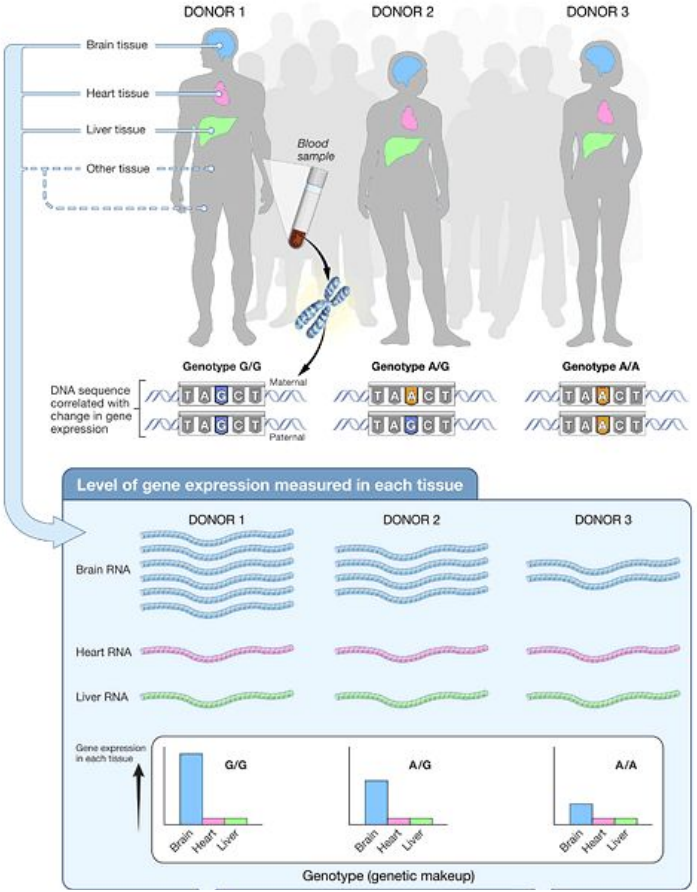
- Exclusively using p-values to determine the relevance and sanity of the results of a statistical test.
- Analyzing the data until the desired results are found.
- Collecting more data to reach smaller p-values.
- Trying many hypothesis until one of them gives a low p-value, and reporting just that final result.

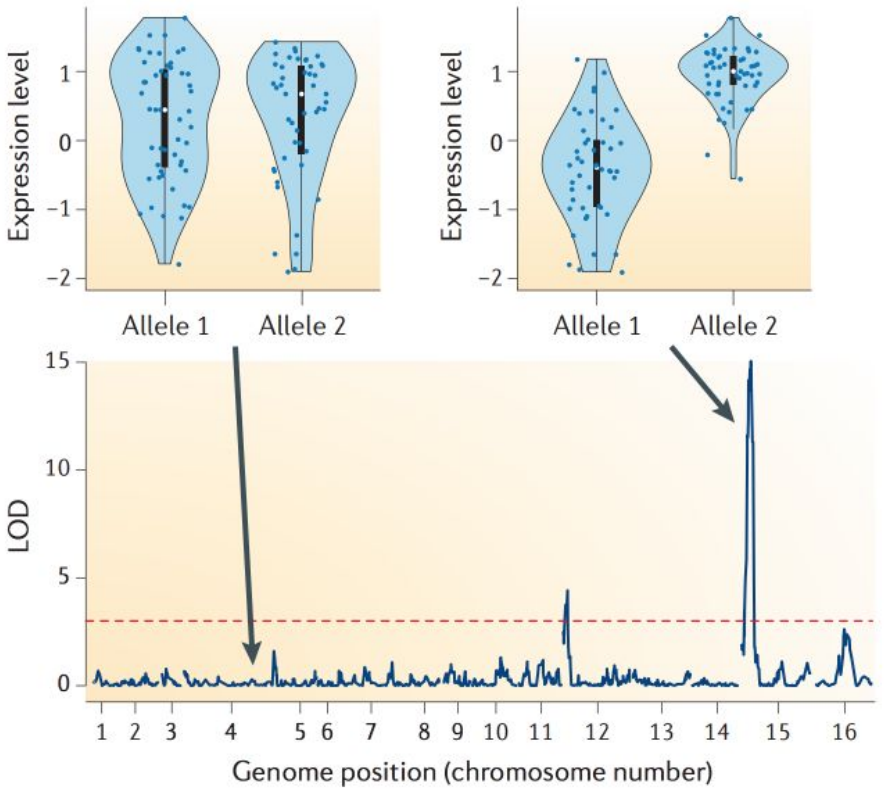
P-values are just the tip of the iceberg!



JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

GWAS-like approaches – eQTL analysis

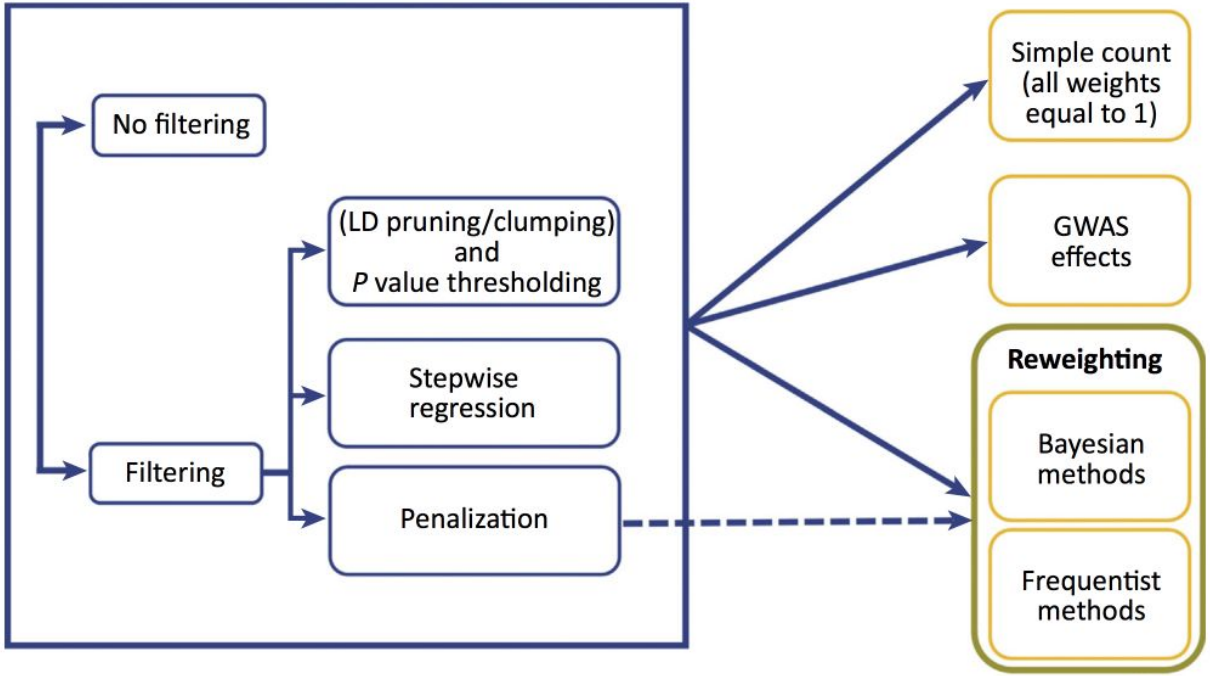




genome.gov
Albert & Kruglyak (2015) Nat. Rev. Genet.

Polygenic risk score

A weighted sum of the number of risk alleles carried by an individual, where the risk alleles and their weights are defined by the loci and their measured effects as detected by GWAS.

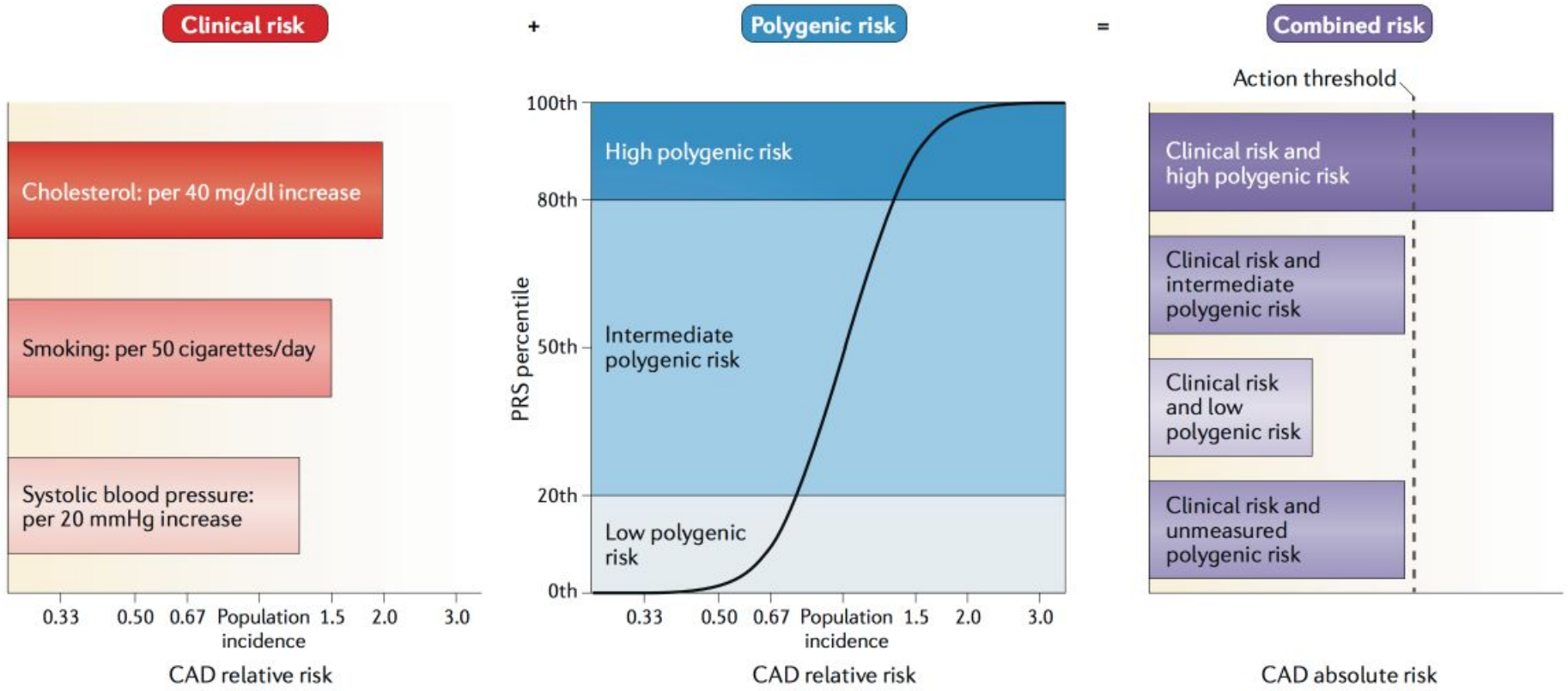


$$PRS_i = \sum_{j \in SNPs} d_{ij}$$

$$PRS_i = \sum_{j \in SNPs} \beta_j d_{ij}$$

Chasioti (2019) Trends. Genet.

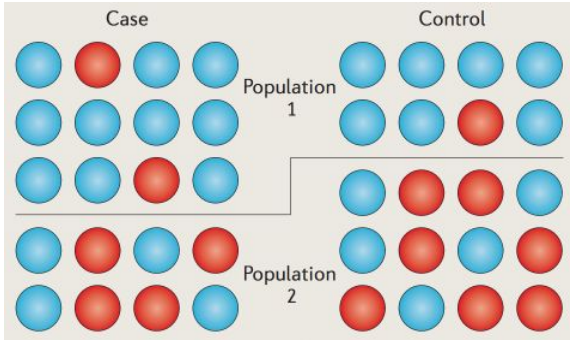
Polygenic risk score

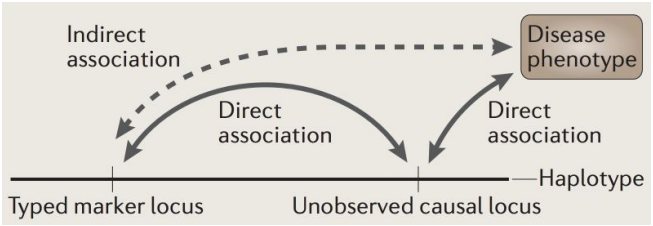


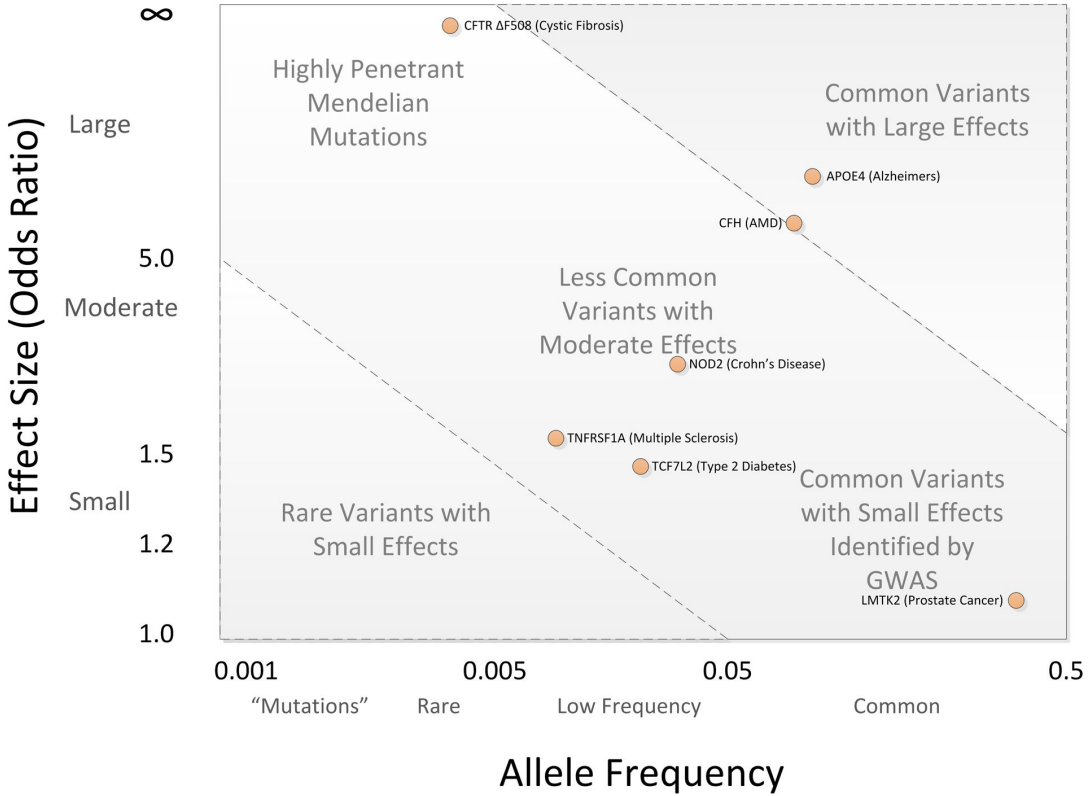
Torkamani (2018) Nat. Rev. Genet.

GWAS – Limitations

- Population structure
- Allele frequency & effect size
- Epistasis
- Identification of causal variant







Bush & Moore (2012) PLoS Comp. Biol.
Balding(2006) Nat. Rev. Genet.