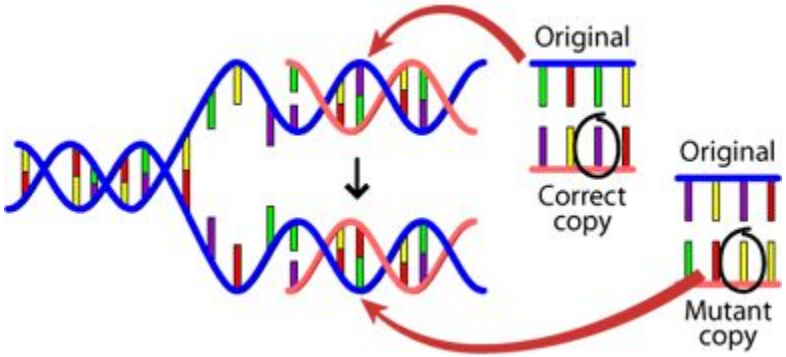
Couple of things...

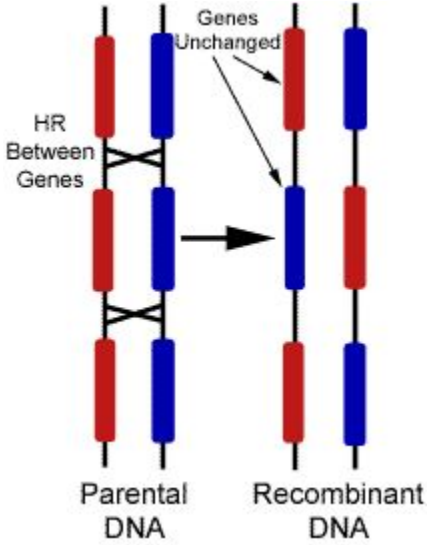
1. **The 8th ISCB Wikipedia Competition** https://en.wikipedia.org/wiki/Wikipedia:ISCB-WP8
    a. Logan DW, Sandal M, Gardner PP, Manske M, Bateman A (2010). "Ten simple rules for editing Wikipedia". PLoS Comput. Biol. 6 (9). doi:10.1371/journal.pcbi.1000941.
    b. https://en.wikipedia.org/wiki/User:Rockpocket/Training
2. **Bioinformatics Contest 2019** https://bioinf.me/en/contest
3. **DREAM Challenges** http://dreamchallenges.org/
    a. Drug-kinase binding prediction
    b. Tumor deconvolution
4. **ENCODE Imputation Challenge** https://www.synapse.org/#!Synapse:syn17083203/wiki/587192

Lecture 10-11: Genetic variation & Quantitative genetics

- Genome-wide association studies
  - Statistical inference, P-values, & Multiple hypothesis testing
  - LD, Regularized linear regression

Genetic variation





Single Nucleotide Polymorphisms (SNPs)

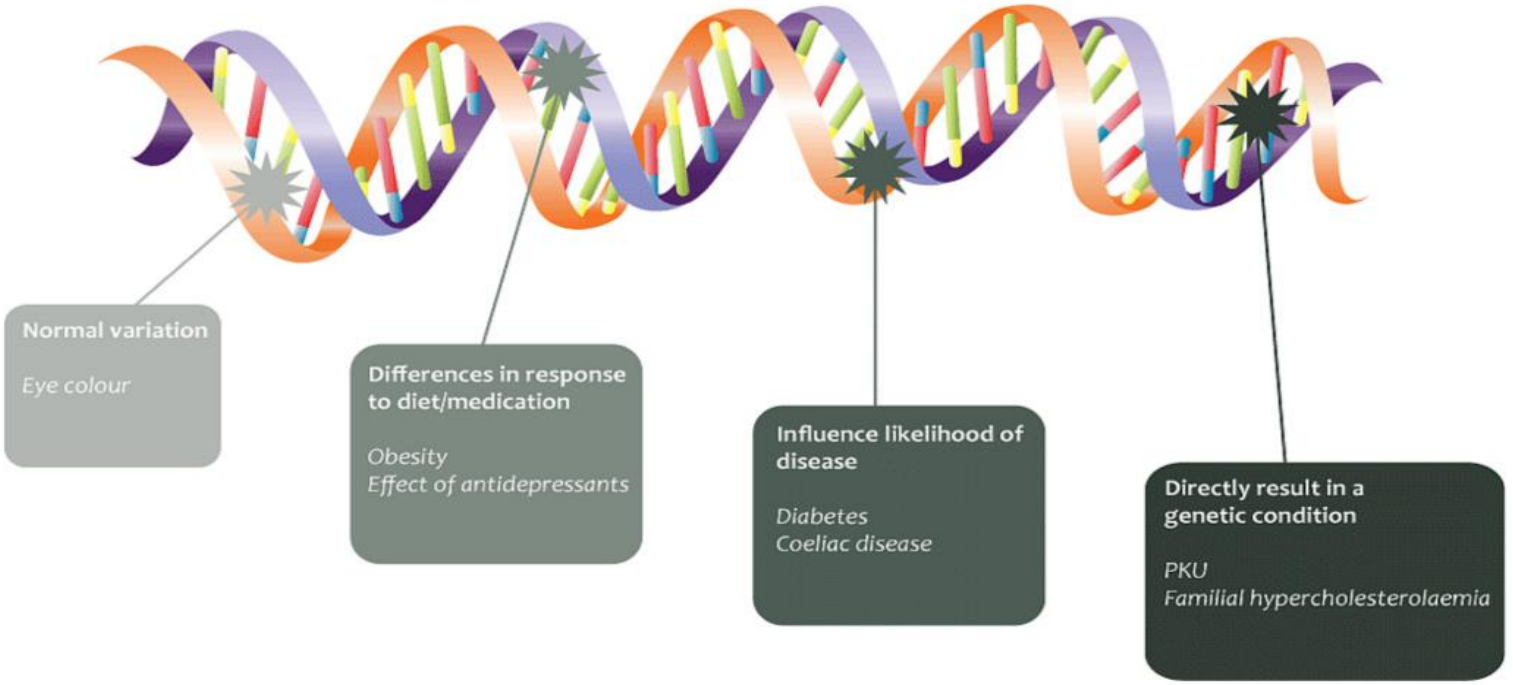
Insertions

Deletions

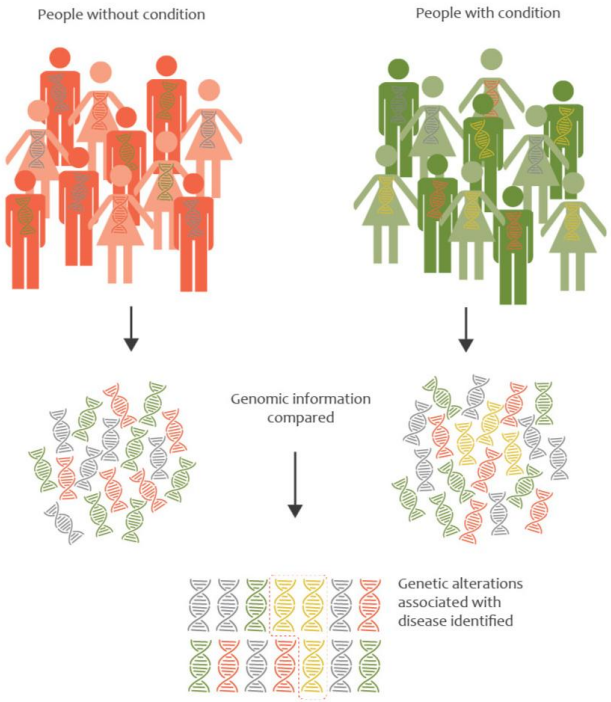
Copy Number Variants (CNVs)

- Duplications & deletions

Genetic variation



Genome-wide Association Study (GWAS)

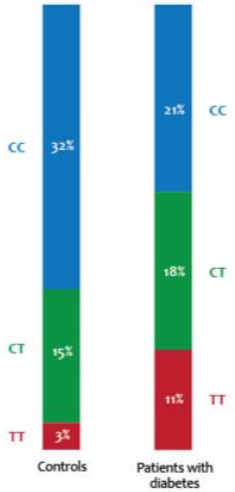


A C/T SNP from a hypothetical GWAS for type 2 diabetes

- Increase in freq of T allele in patients w/ diabetes compared to controls.
- We know where this SNP is on the genome ⟶ study surrounding sequence

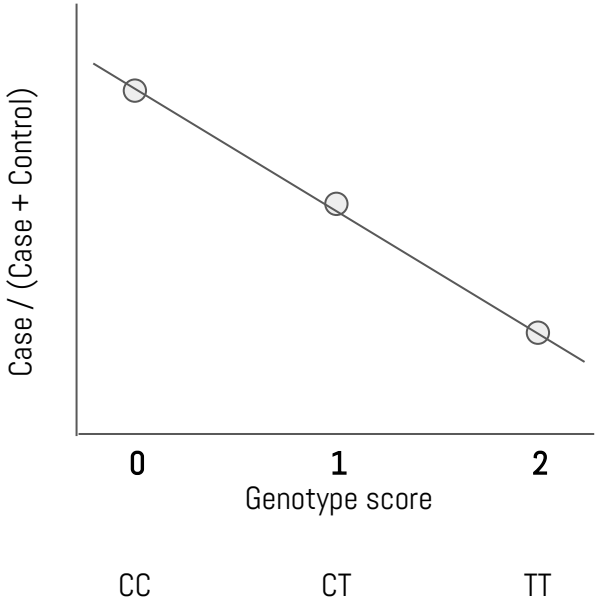
Visscher (2017) AJHG

GWAS – Analysis



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

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

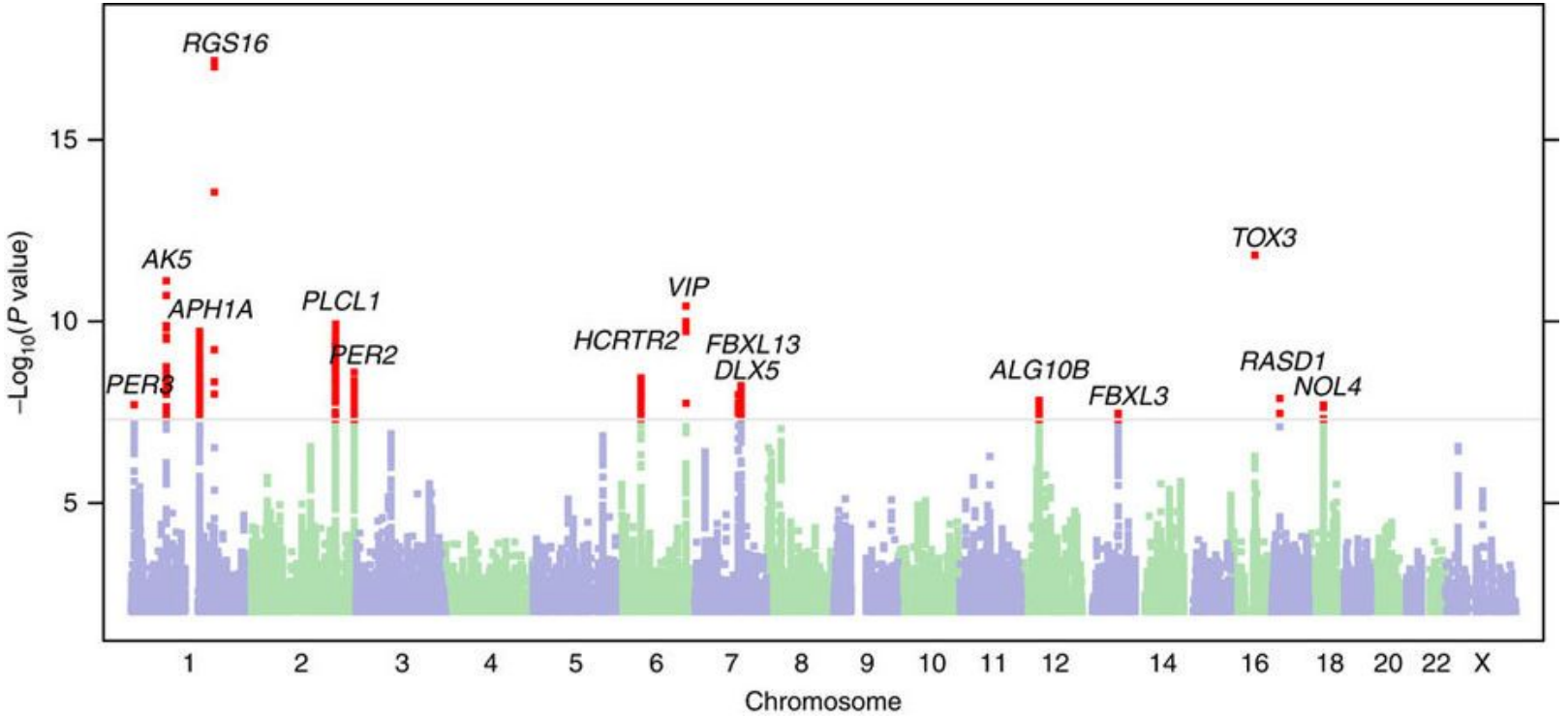


Chi-squared test
Fisher's exact test

Armitage test: testing the linear associations between trait and number of one of the alleles

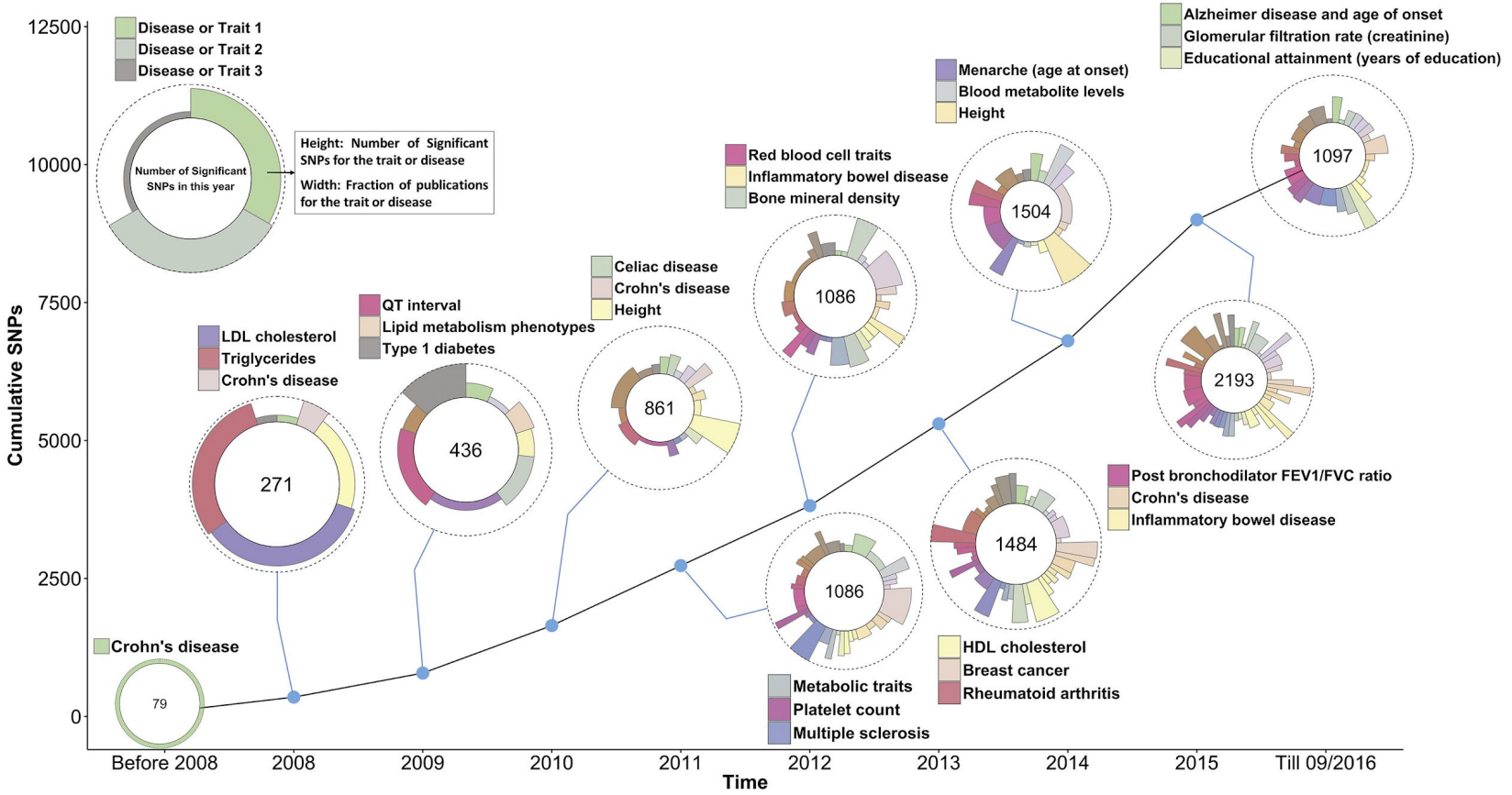
GWAS – Analysis

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



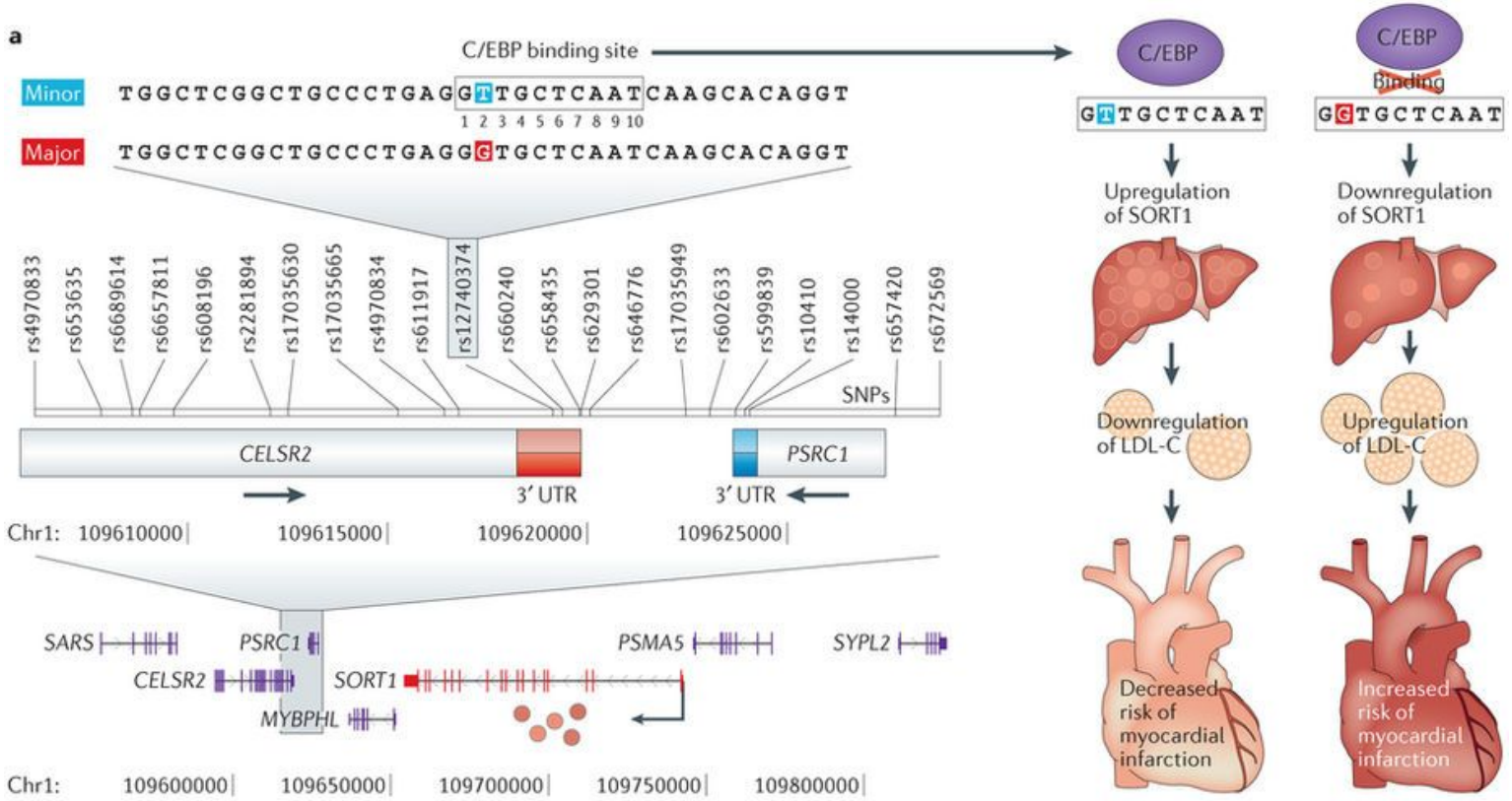
Hu (2016) Nat. Comm.

GWAS – Timeline of discoveries



Visscher (2017) AJHG

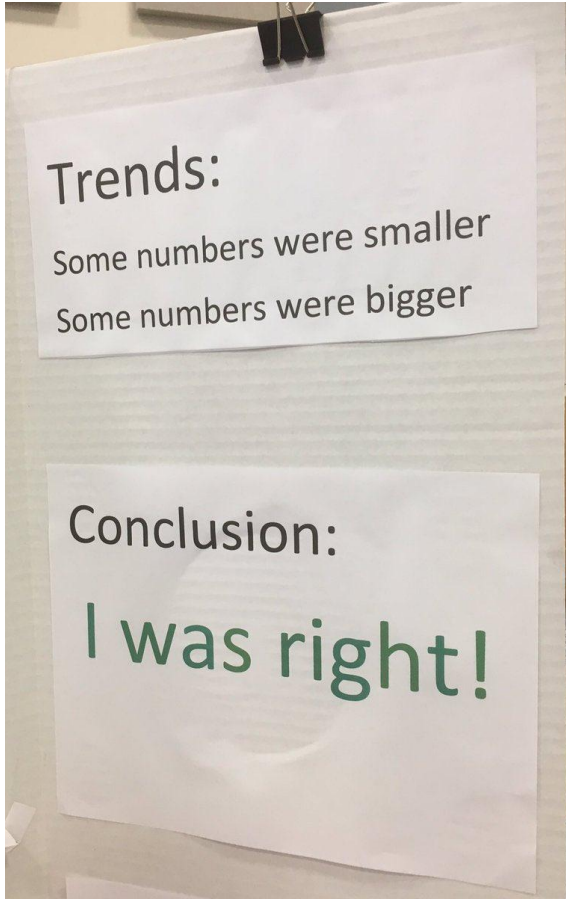
GWAS – Examples



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

Hypothesis testing

- Consider **two competing hypotheses** for a given SNP:
  - Null hypothesis: the frequency of the SNP in the cases is the same as that in controls.
  - Alternative hypothesis: the frequencies are different.
- There's always some difference → Is it statistically significant difference?
  - Calculate a **test statistic** for these measurements, and then determine its p-value.
  - **P-value**: the probability of observing a test statistic that is as extreme or more extreme than the one we have, assuming the null hypothesis is true.



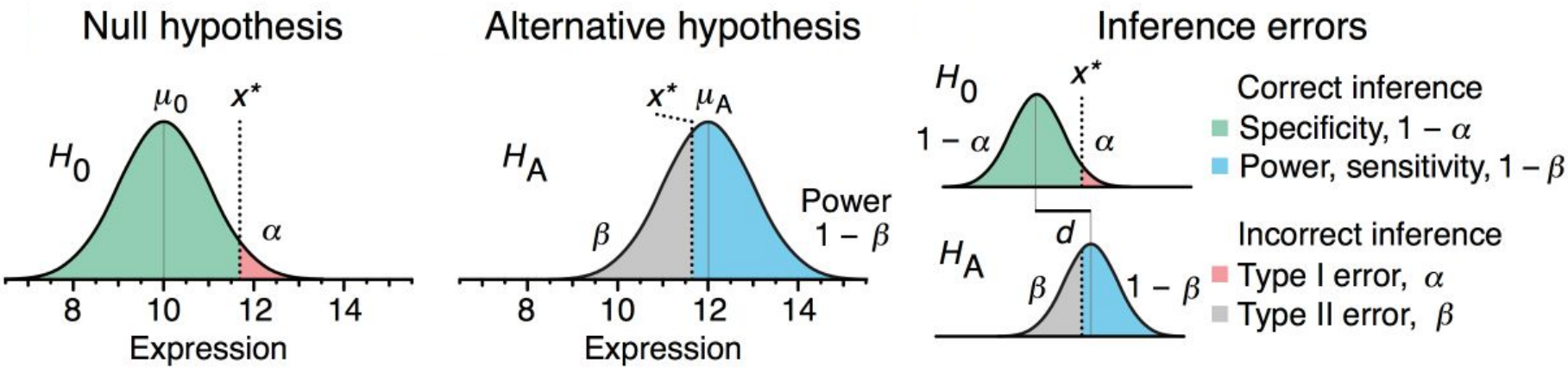
Hypothesis testing

- The p-value is:
  - The amount of evidence that there is an effect?
  - The probability that the observed outcome is important?
  - The probability that the medication is ineffective?

The p-value is the probability that the experiment would have produced the observed outcome (or something more extreme) even if the medication were completely ineffective.

Write code to simulate two distributions and calculate p-values both using a t-test and a permutation test.

Hypothesis testing



David Robinson
@drob

Follow

Remember, mixing up Type I and Type II errors is called a Type III error

Giving mistakes numbers instead of names was a real Type IV error

P-value captures if there is "sufficient" inconsistency with the null hypothesis.

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

Krzywinski & Altman (2013) Nat. Meth.

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.

P-value

Significant or not!

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

The following 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.

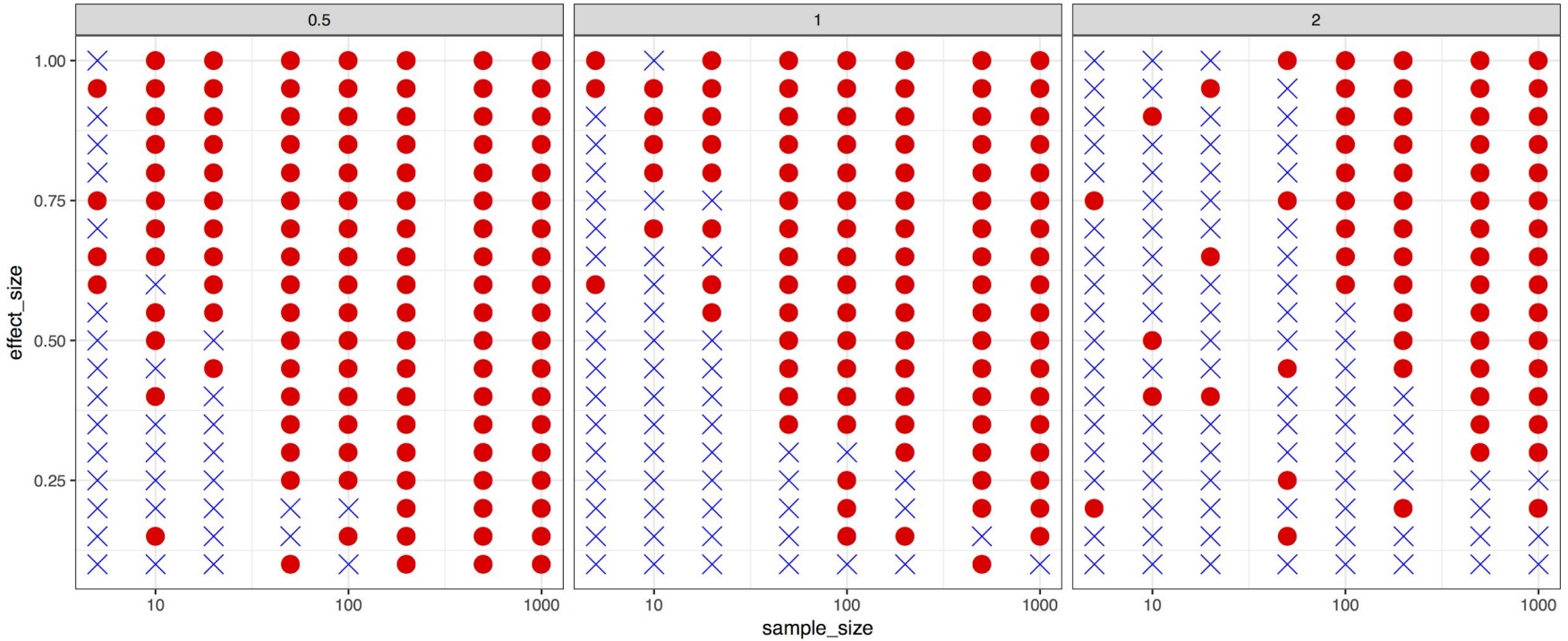
(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable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

Hypothesis testing

- P-values are dependent on:
  a. Size of the effect (effect size)
  b. Sample size
  c. Variance within each group
  d. The underlying experimental design & the null hypothesis (need 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; Neg. Binomial p-value = 0.033.

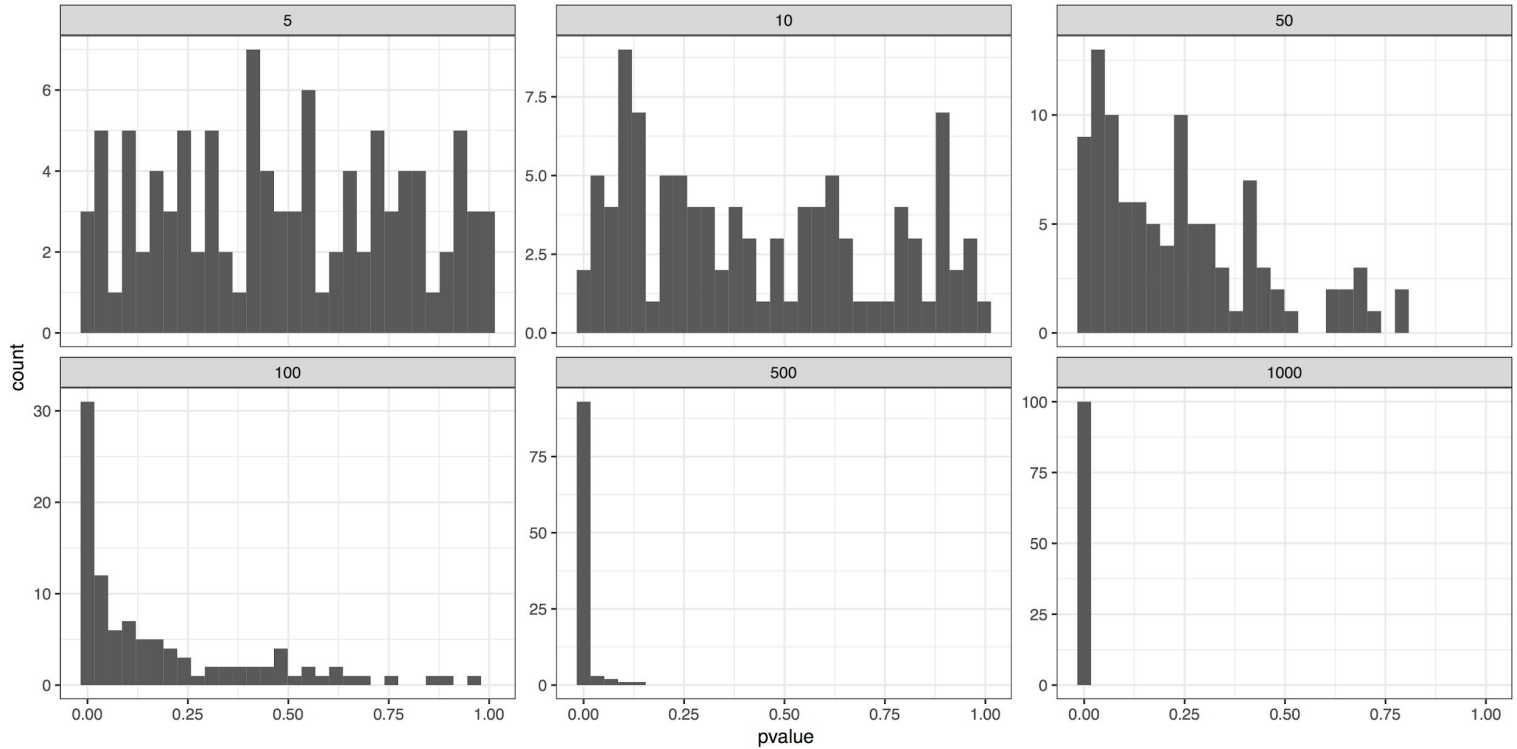
Hypothesis testing

- P-values are dependent on: sample size, effect size, within-group variance

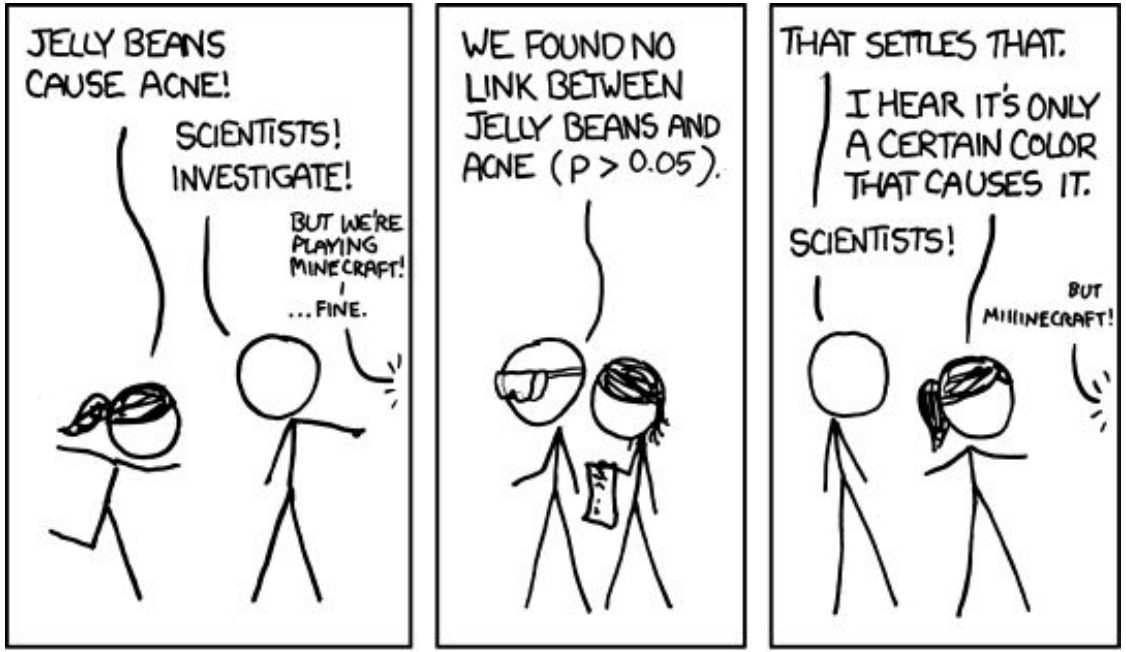


P-value

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

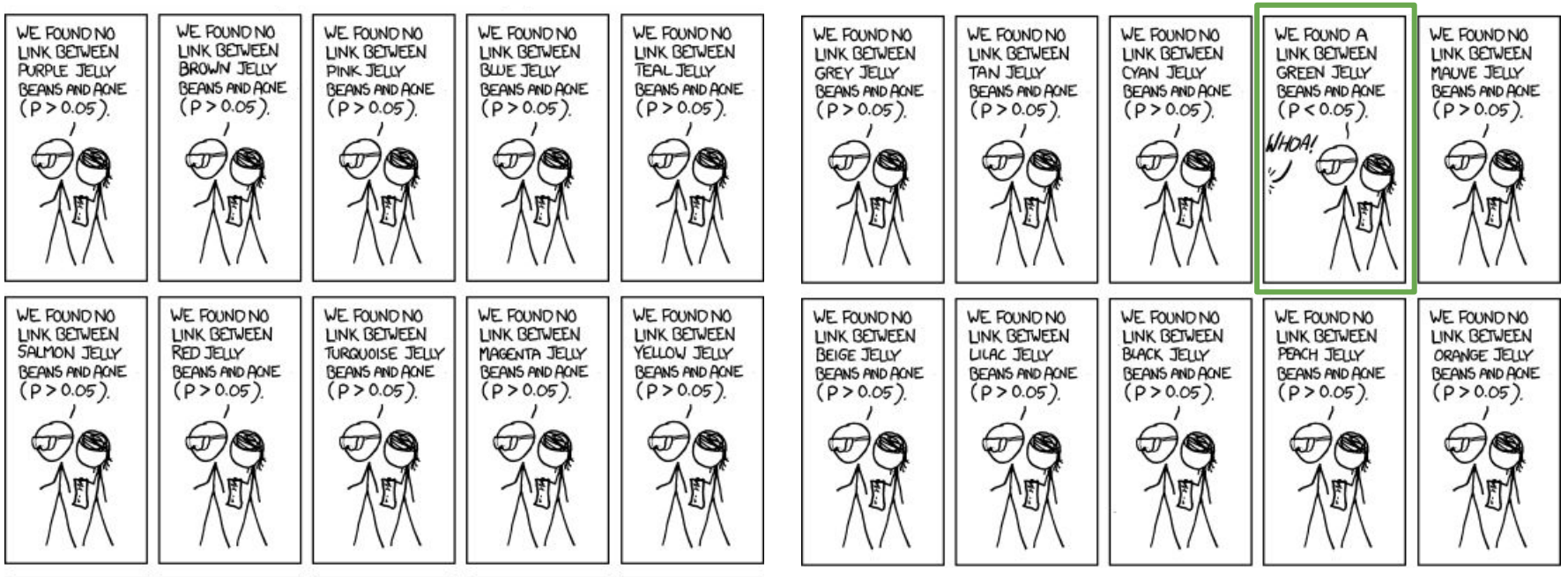


Multiple hypothesis testing



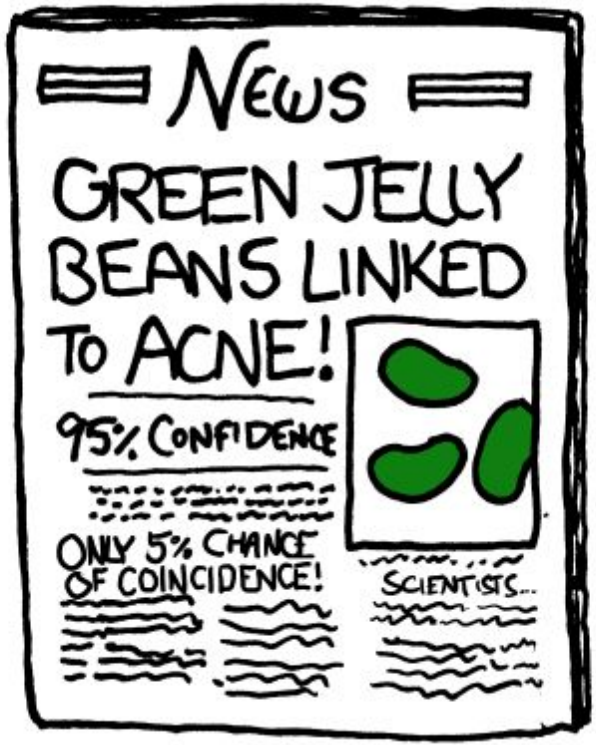
xkcd.com/882/

Multiple hypothesis testing



xkcd.com/882/

Multiple hypothesis testing



- The more inferences are made, the more likely erroneous inferences are to occur.
- Several statistical techniques have been developed to prevent this from happening.
- These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

What is the probability of obtaining at least 1 false positive? Family-wise error rate (FWER)

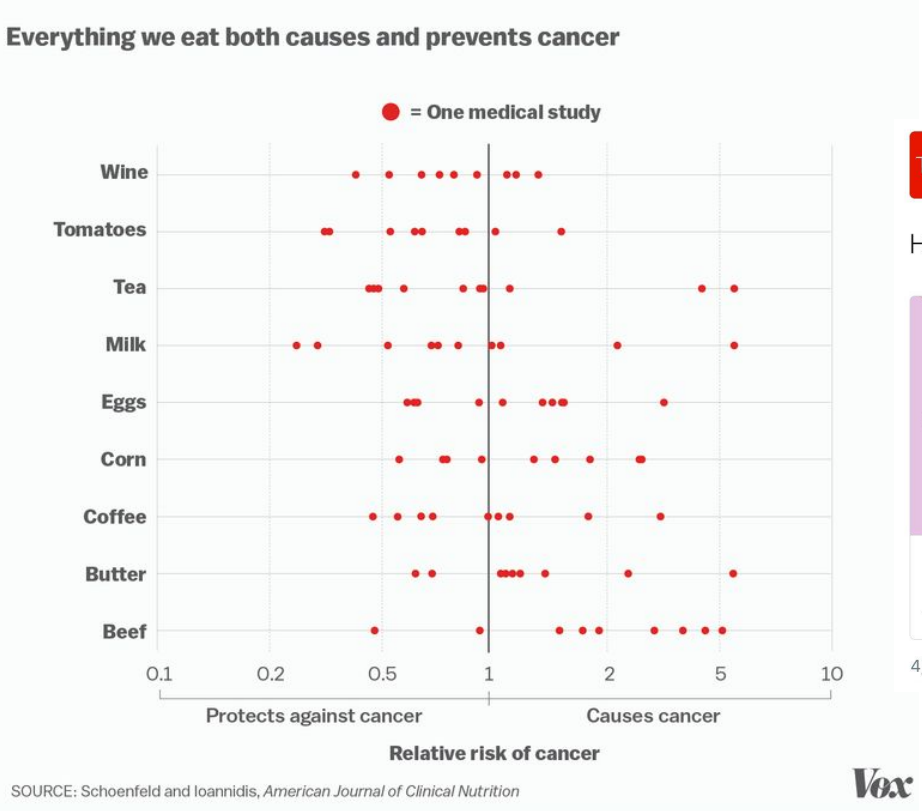
How many of my findings are false? (FDR)

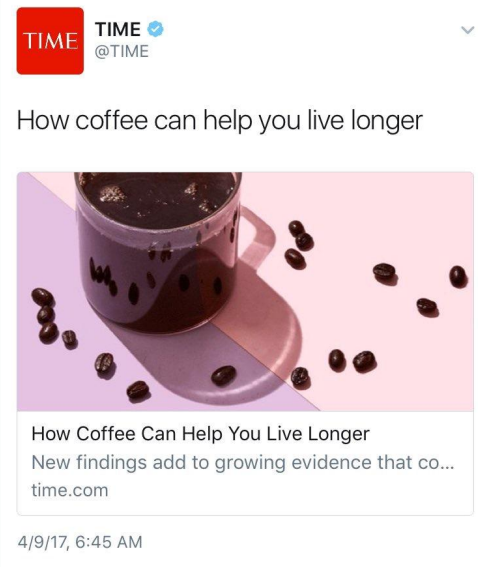
xkcd.com/882/

Multiple hypothesis testing

- FWER (probability of obtaining even 1 false positive) = Pr(#FP ≥ 1)
- False discovery rate (FDR) = E[#FP / #Discoveries]
- Suppose 550 out of 10,000 genes are found to have different expression levels between disease and control samples at $p < 0.05$.
    - If p-value is chosen to control FWER, what is the #FP?
    - If p-value is chosen to control FDR, what is the #FP?

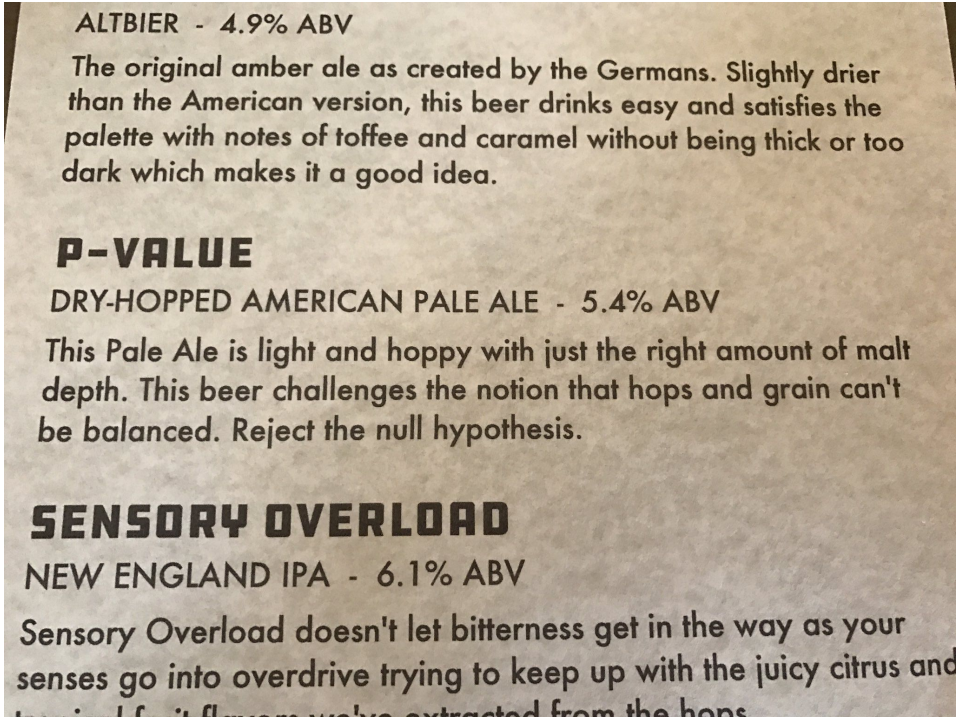
Hypothesis testing



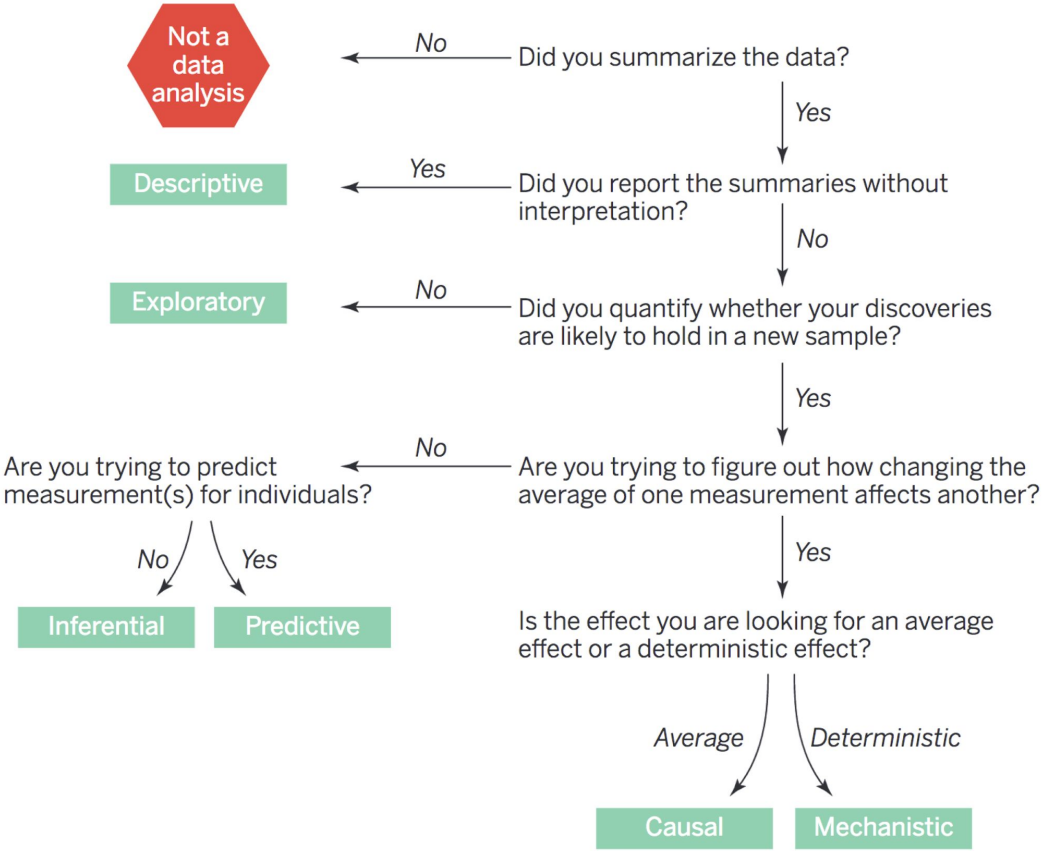


P-value

The 'p' in p-value actually stands for p-otentially interesting!



What is the question?



JT Leek, RD Peng What is the question?

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.

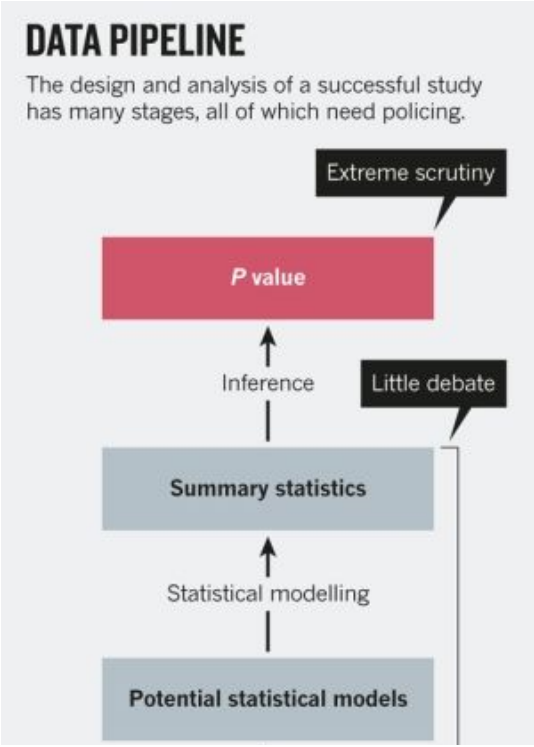
Sanity checks

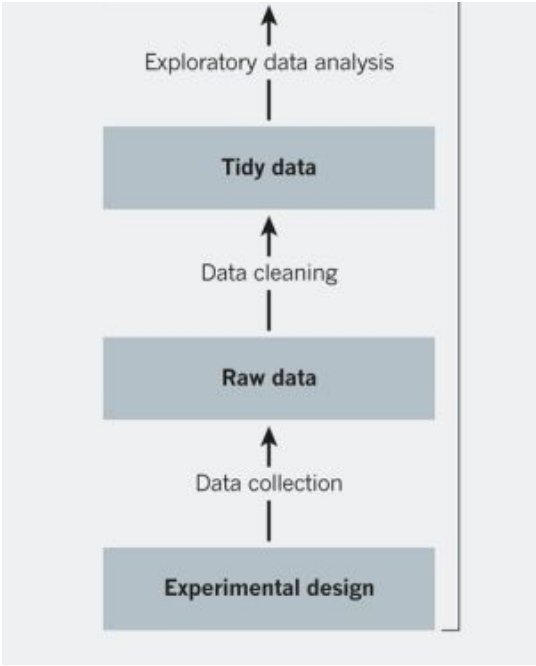
Why are you holding an umbrella?

Because my phone said it's raining!

Hypothesis testing

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

Statistical analysis of genome-wide association

- Description of the problem: cases, features
- Lasso: Regularized linear regression
  - Loss function: L1 vs. L2
  - Regularization (parameter: **λ**)
- Lasso is an example of “feature selection”

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Statistical analysis of genome-wide association

- Solving lasso with the least-angle regression algorithm
- If a non-zero coefficient hits zero, remove it from the active set of predictors and recompute the joint direction.

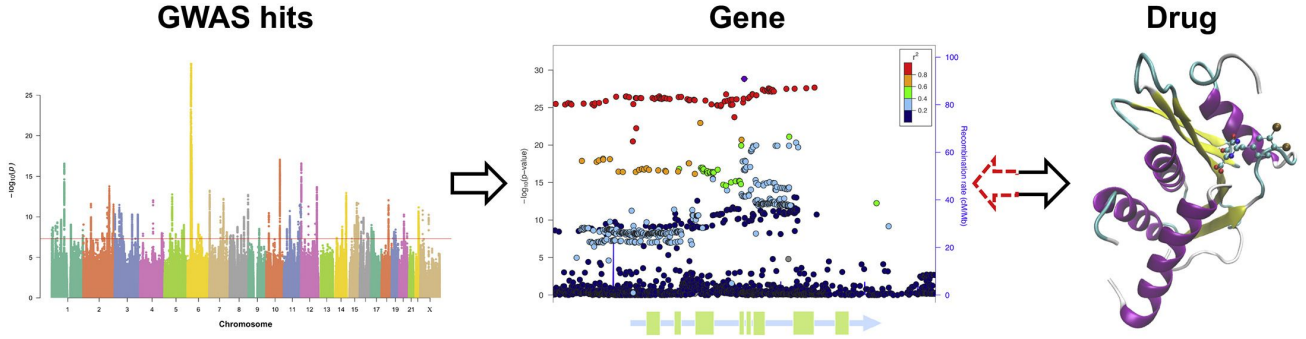
Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.

- If a non-zero coefficient hits zero, remove it from the active set of predictors and recompute the joint direction.

Elements of Statistical Learning

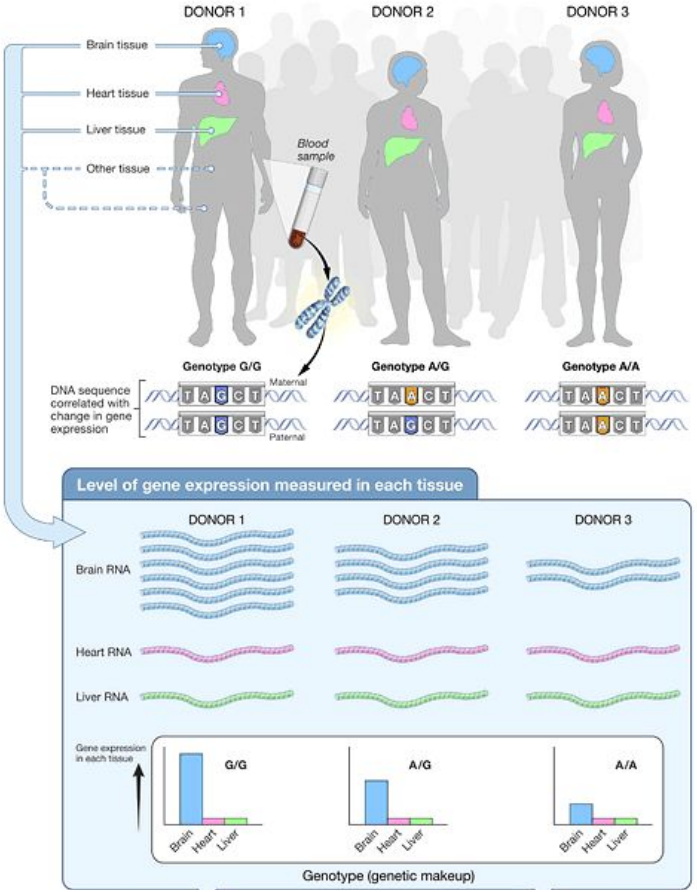
GWAS & drugs

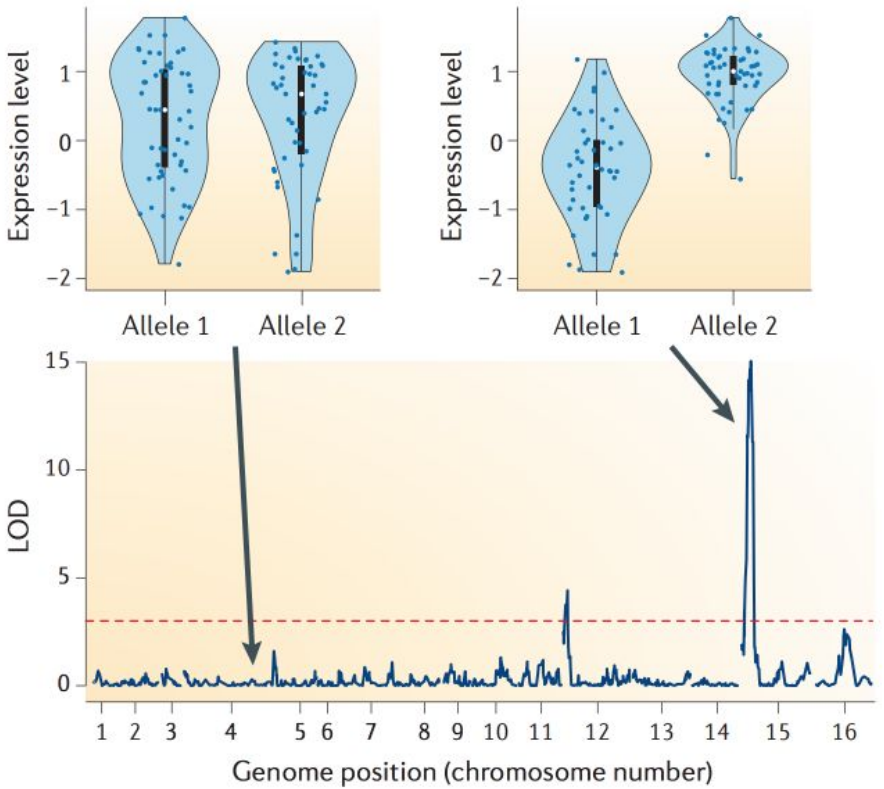


Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*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 – eQTL analysis



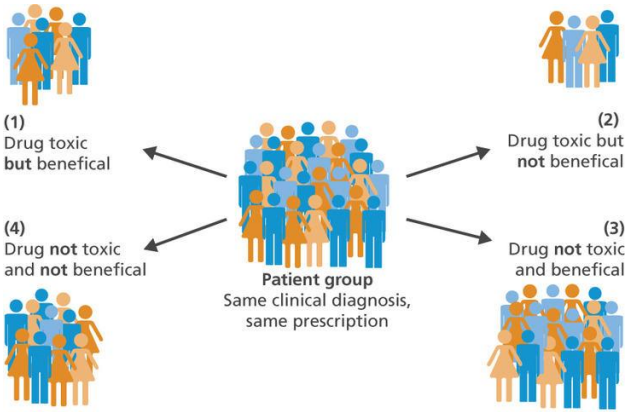


genome.gov

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

GWAS-like approaches

Pharmacogenomics



Clinical diagnosis
Breast cancer (all groups)

↓

Test for biomarker
(protein on cancer cell surface)

Absent:	**Present**: prescribe
do not prescribe Herceptin (not beneficial) (2 & 4)	Herceptin (only beneficial when biomarker present) (1 & 3)

Clinical diagnosis
HIV/AIDS

Drug (Abacavir) is known to be benefical

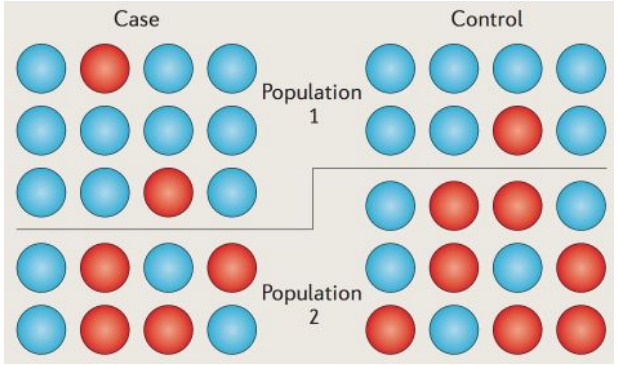
↓

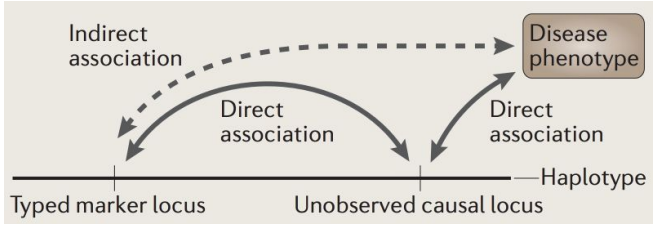
Test for genetic variant

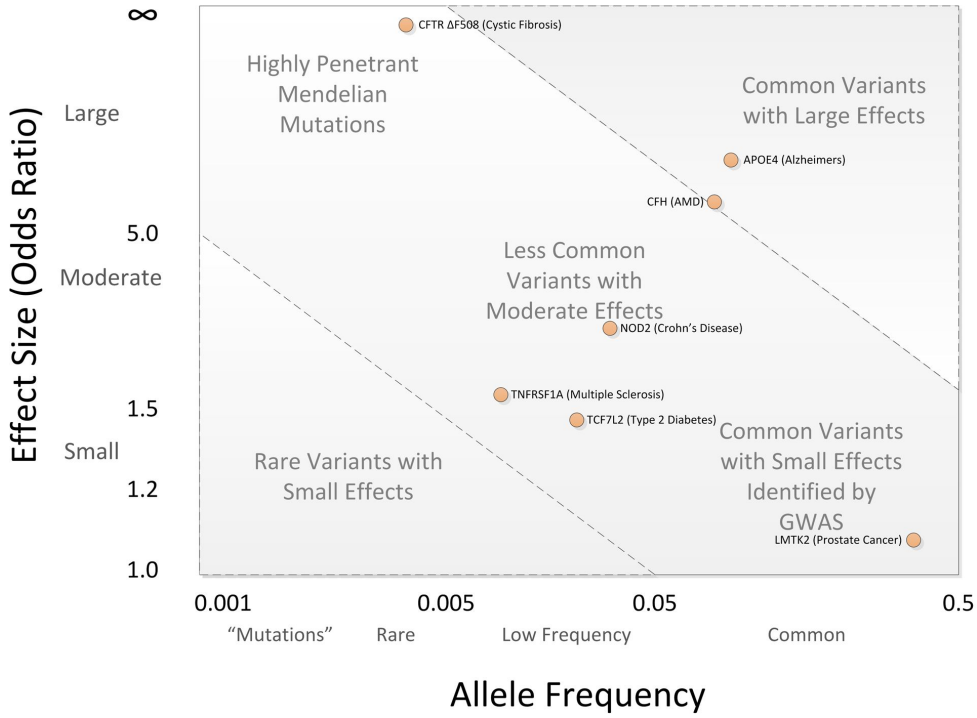
Absent:	**Present**: do
prescribe Abacavir (not toxic) (3)	not prescribe Abacavir (is toxic) (1)

GWAS – Limitations

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







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