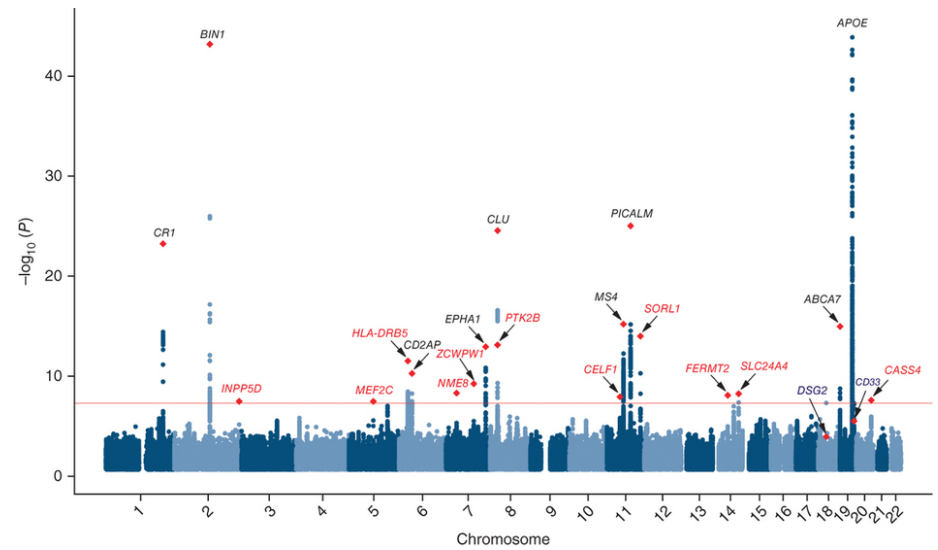


# CS 68: Bioinformatics

Prof. Sara Mathieson  
Spring 2018  
Swarthmore College



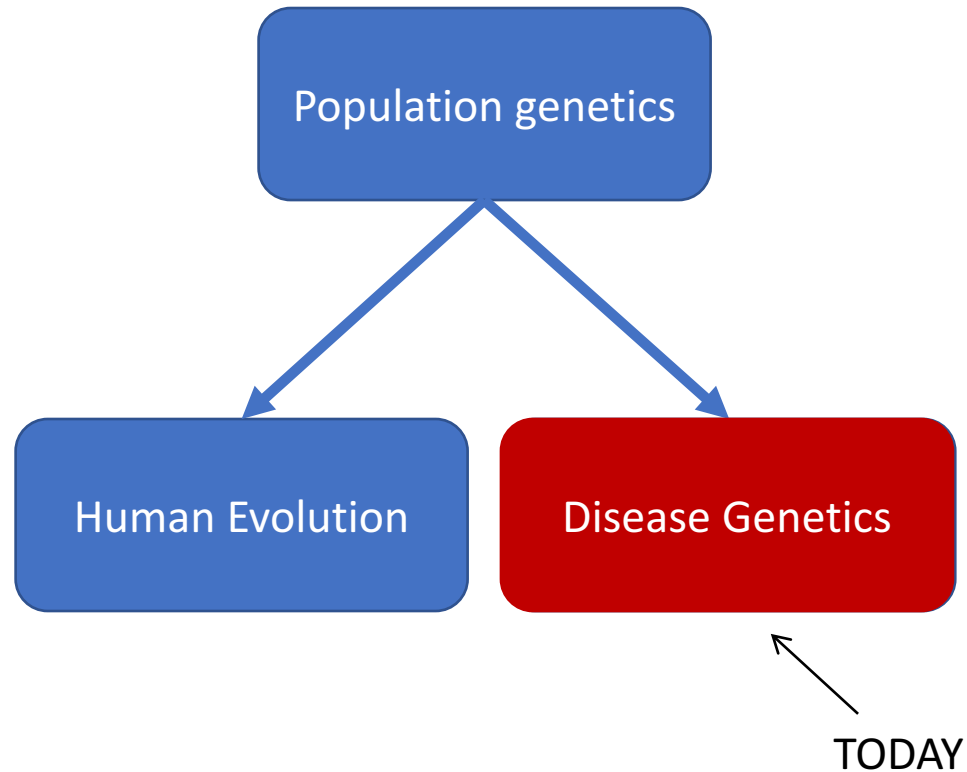
# Outline: April 23

- Disease analysis in computational biology
- Genome-Wide Association Studies (GWAS)
- Impact of population structure

## Notes:

- Project proposal due TONIGHT
- Office hours TODAY 3-5pm
- Midterm 2 in-lab on Thursday

# Applications of genetic sequencing and method development (in humans)



# Human vs nonhuman genetics

## Nonhuman

Can do experiments

Small sample sizes

Large effects

Can easily choose phenotypes



## Human

Have to use natural variation

Large sample sizes ( $n=1,000,000$ )

Large and small effects

Medical phenotypes usually involve complex biology



“Effect” meaning effect on the phenotype (i.e. the physical manifestation of a trait)



# What is the point?

Two big goals of human genetics:

**GWAS**

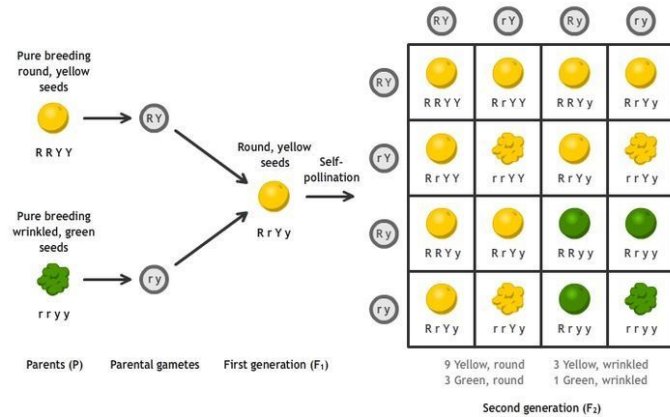
Goal 1: Identify genetic variants (mutations, alleles) that are associated with phenotype, particularly disease

Goal 2: Understand the biological mechanisms through which those variants act.

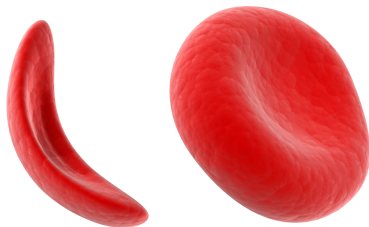
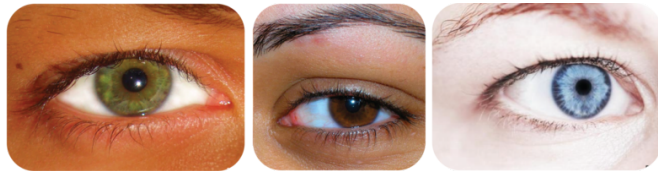
**Hard!**

# What are we looking for?

## Mendelian traits



© 2005-2011 The University of Waikato | www.biotechlearn.org.nz



Thalassemia  
Fragile X  
Tay-Sachs  
Haemophilia

## Complex traits



Type II Diabetes

Pigmentation

Schizophrenia

Anxiety

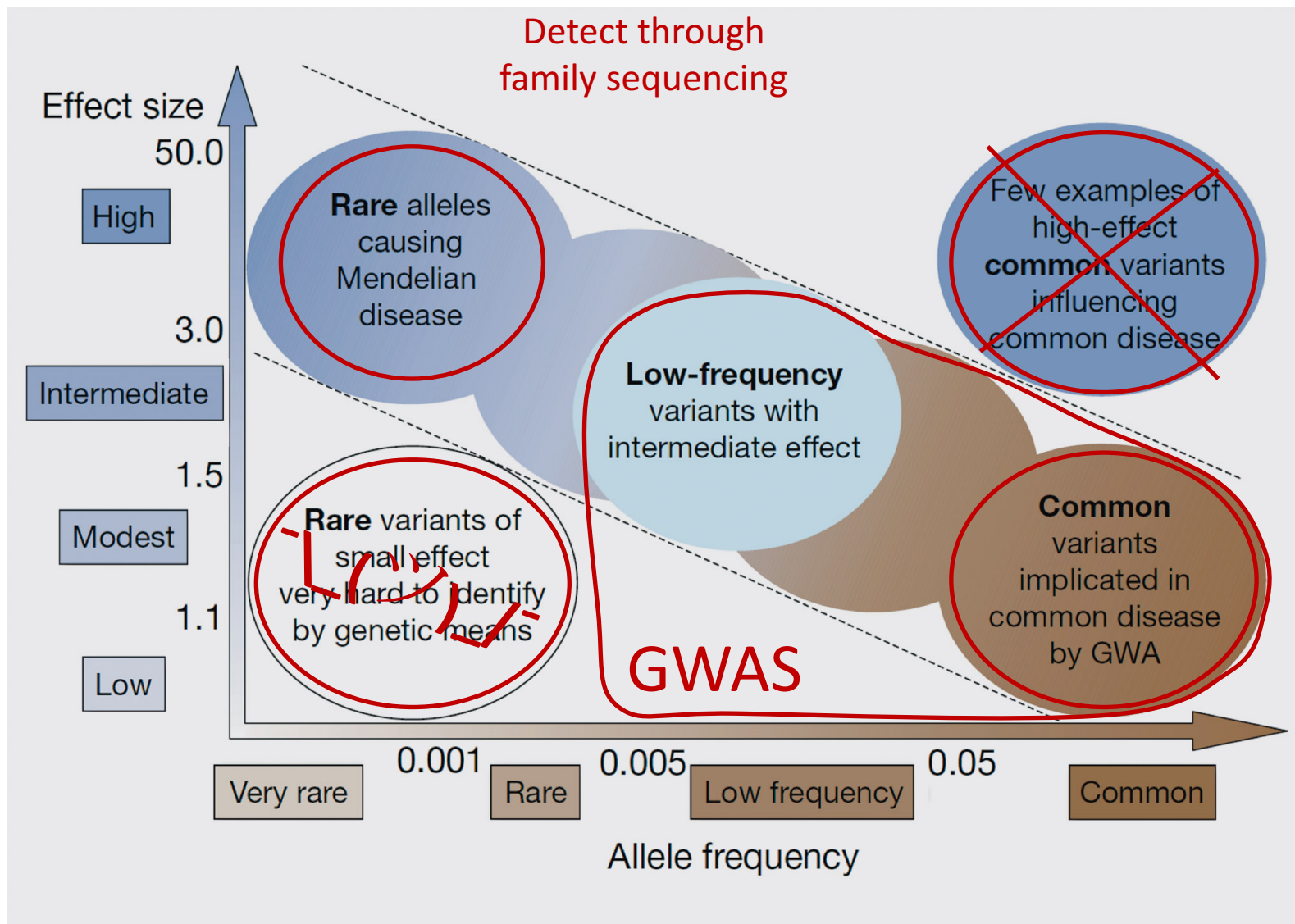
Heart disease

BMI

Cancer susceptibility

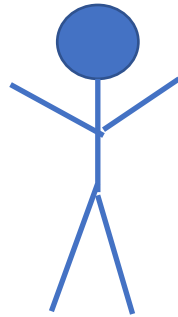
Cholesterol

# What are we looking for?



# ~~Genome-wide~~ Association Studies

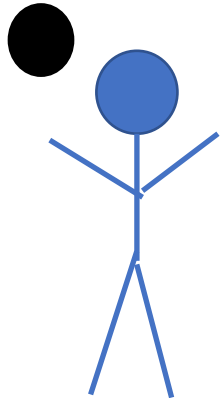
Does not carry  
variant



→ Low risk of disease

Hypothesis

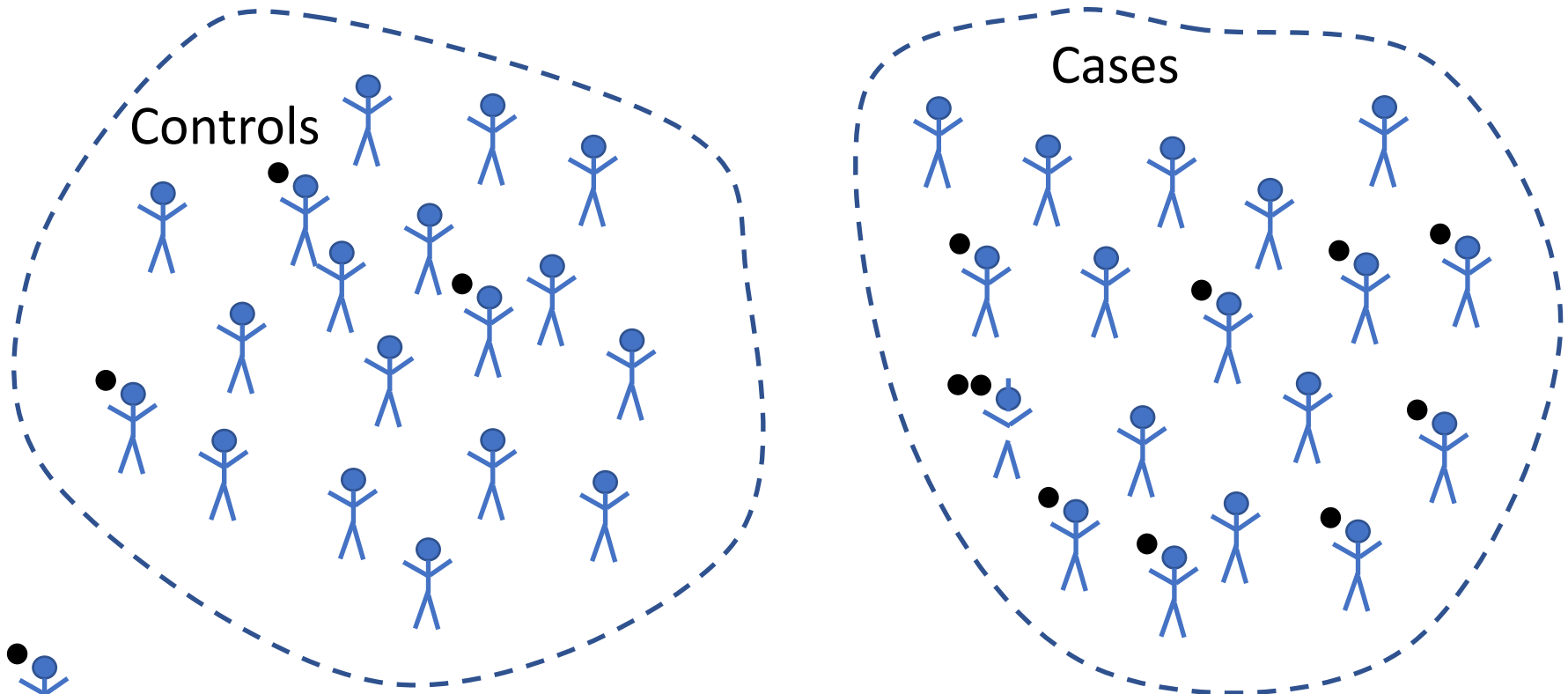
Carries variant




→ High risk of disease



# Test hypothesis: Case-control study

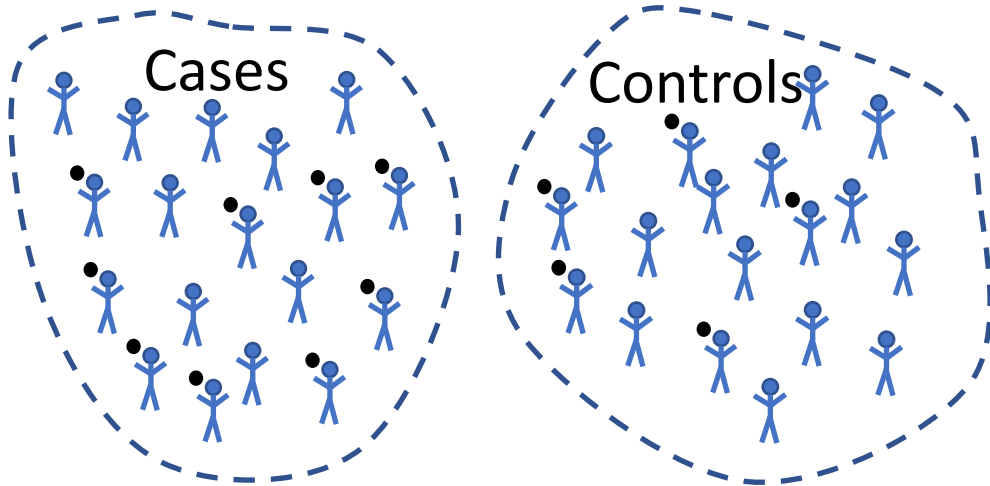


 Has variant

 Doesn't have variant

	Cases	Controls
Has variant	9	3
No variant	8	14

# P-value measures non-randomness

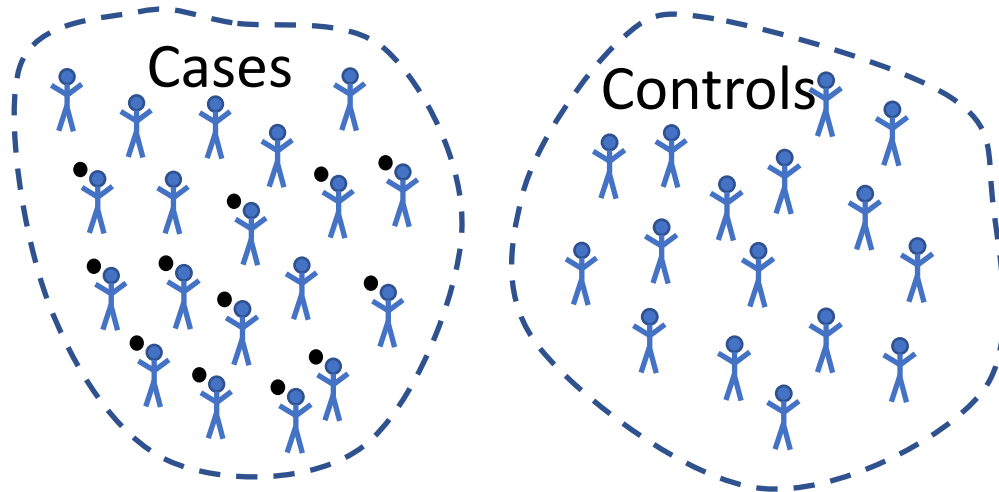


$P=1$

Variant is equally common in cases and controls.

$P=0.05$

Variant is much more common in one group (here cases).



$P=0.05$  means that there is a 1 in 20 (5%) chance of seeing a more extreme result, if the variant is not actually associated with the trait.

# P-values: is this result significant?

	Cases	Controls	TOTAL
Has variant	9	3	12
No variant	8	14	22
TOTAL	17	17	34

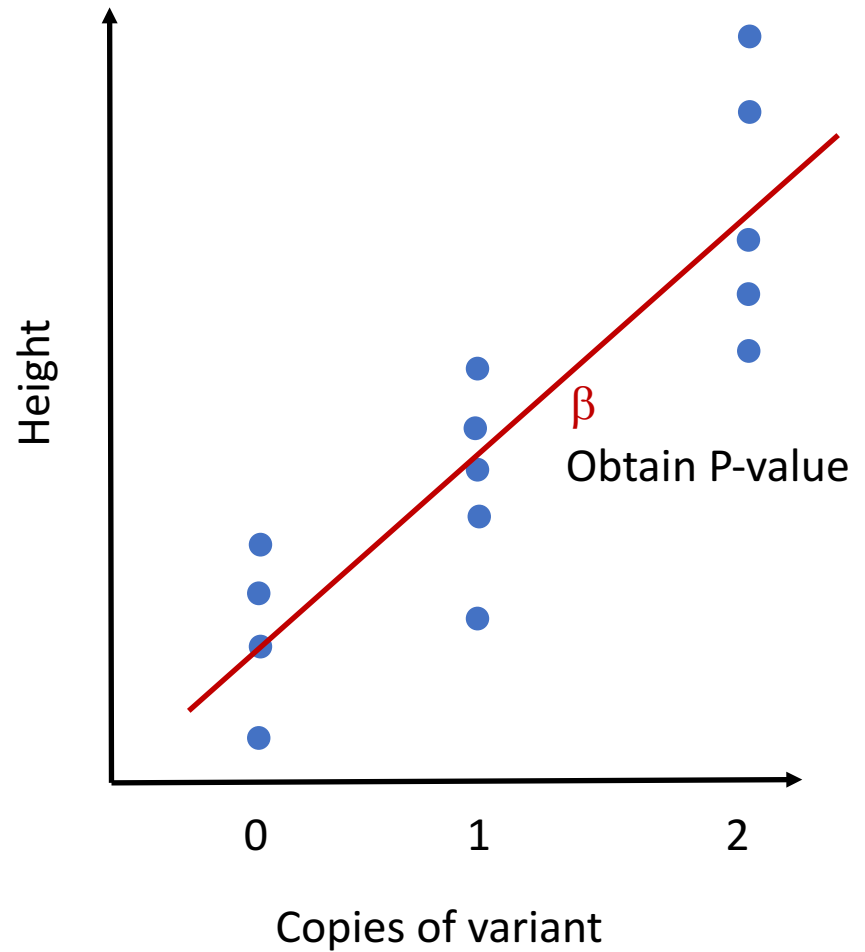
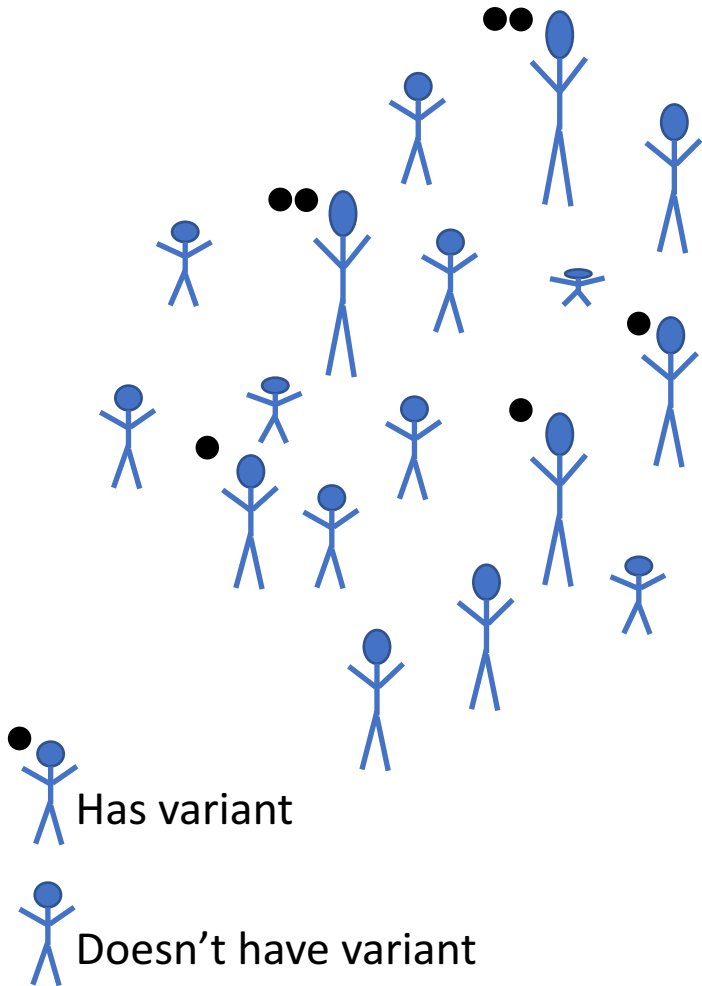
- Expected number of cases with variant =  $17 \cdot 12 / 34 = 6$
- Expected number of controls with variant =  $17 \cdot 12 / 34 = 6$
- Expected number of cases without variant =  $17 \cdot 22 / 34 = 11$
- Expected number of controls without variant =  $17 \cdot 22 / 34 = 11$

$$\begin{aligned}\text{Compute a } \chi^2 \text{ statistic} &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(9-6)^2}{6} + \frac{(3-6)^2}{6} + \frac{(8-11)^2}{11} + \frac{(14-11)^2}{11} \\ &= 4.636\end{aligned}$$

Yes, at a 0.05  
significance level

Is this significant? P=0.0313 [R code: `1-pchisq(4.636, df=1)`]

# Continuous (“quantitative”) traits



# Association Studies

Sometimes called “Candidate gene studies”

Two problems:

1) Need to know which genes/variants to look at *a priori*

Solution: Test lots of variants in the whole genome (“genome-wide”)

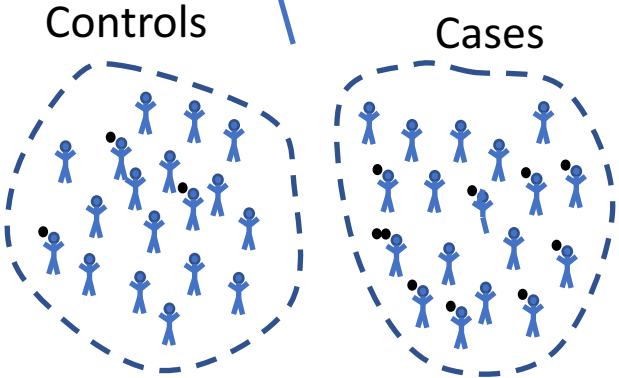
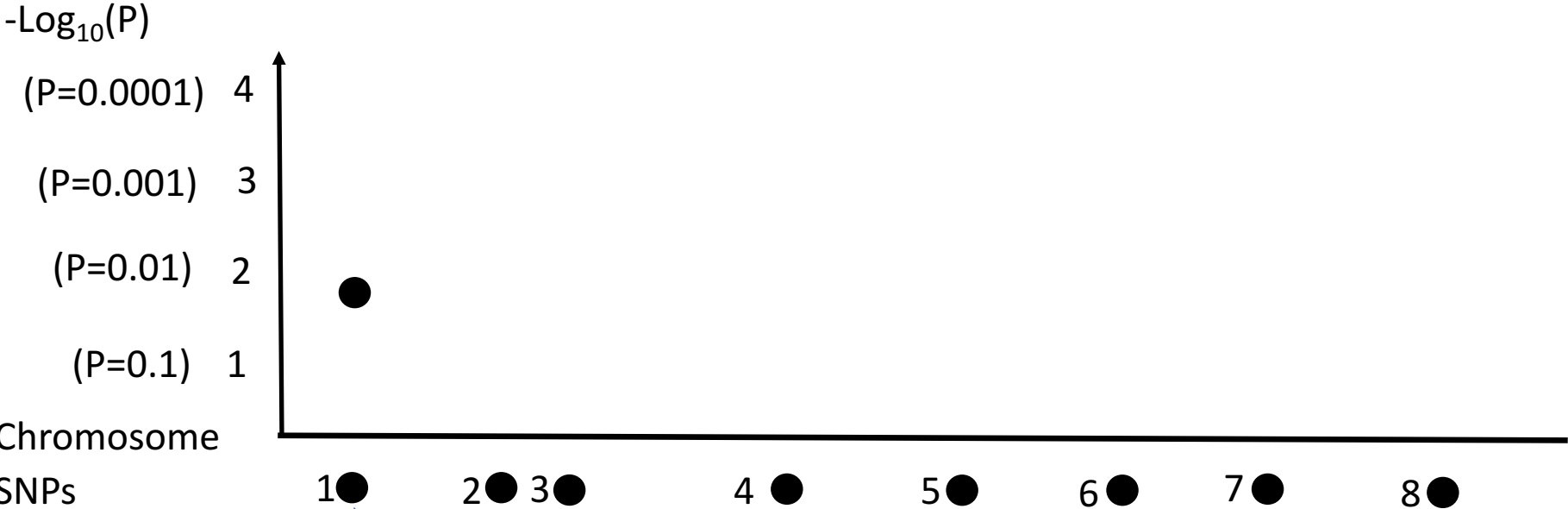
2) Confounded by population structure

Solution: If you test the whole genome, most of the variants will not be associated with the trait. So use those to measure and correct structure

# Genome-wide Association Studies

- Lots of people. Number of people depends on the effect size. Most GWAS today have  $n=10,000-1,000,000$ .
- Genome-wide data. Usually SNP-array data. Typically 100,000-1,000,000 SNPs across the genome
- A phenotype. Anything! GWAS have been carried out for 3,357 traits.

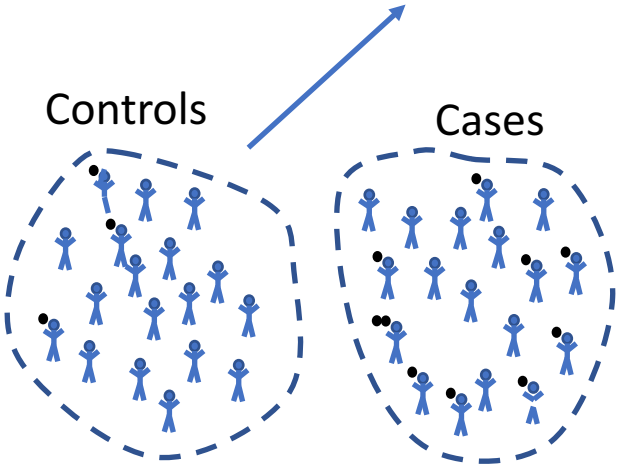
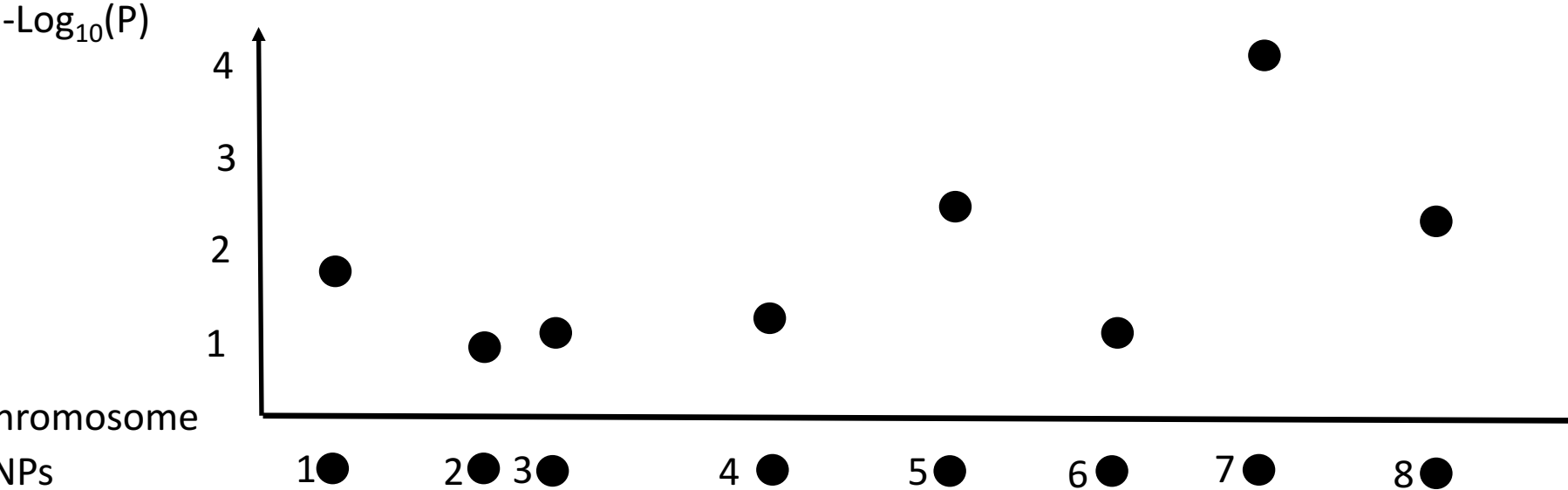
# Genome-wide Association Studies



	Cases	Controls
Has variant	9	3
No variant	8	14

$P=0.03$

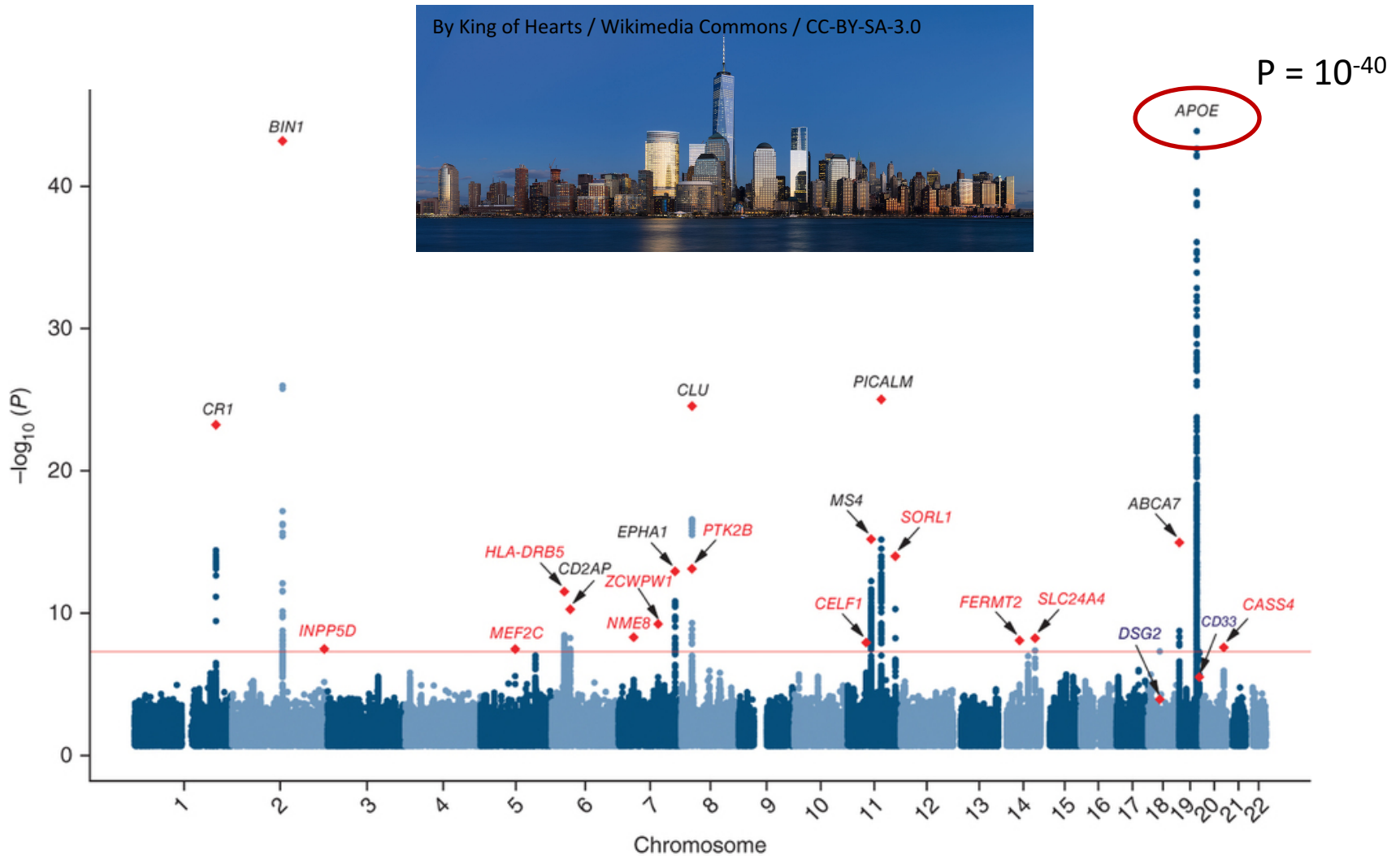
# Genome-wide Association Studies



	Cases	Controls
Has variant	6	6
No variant	8	12

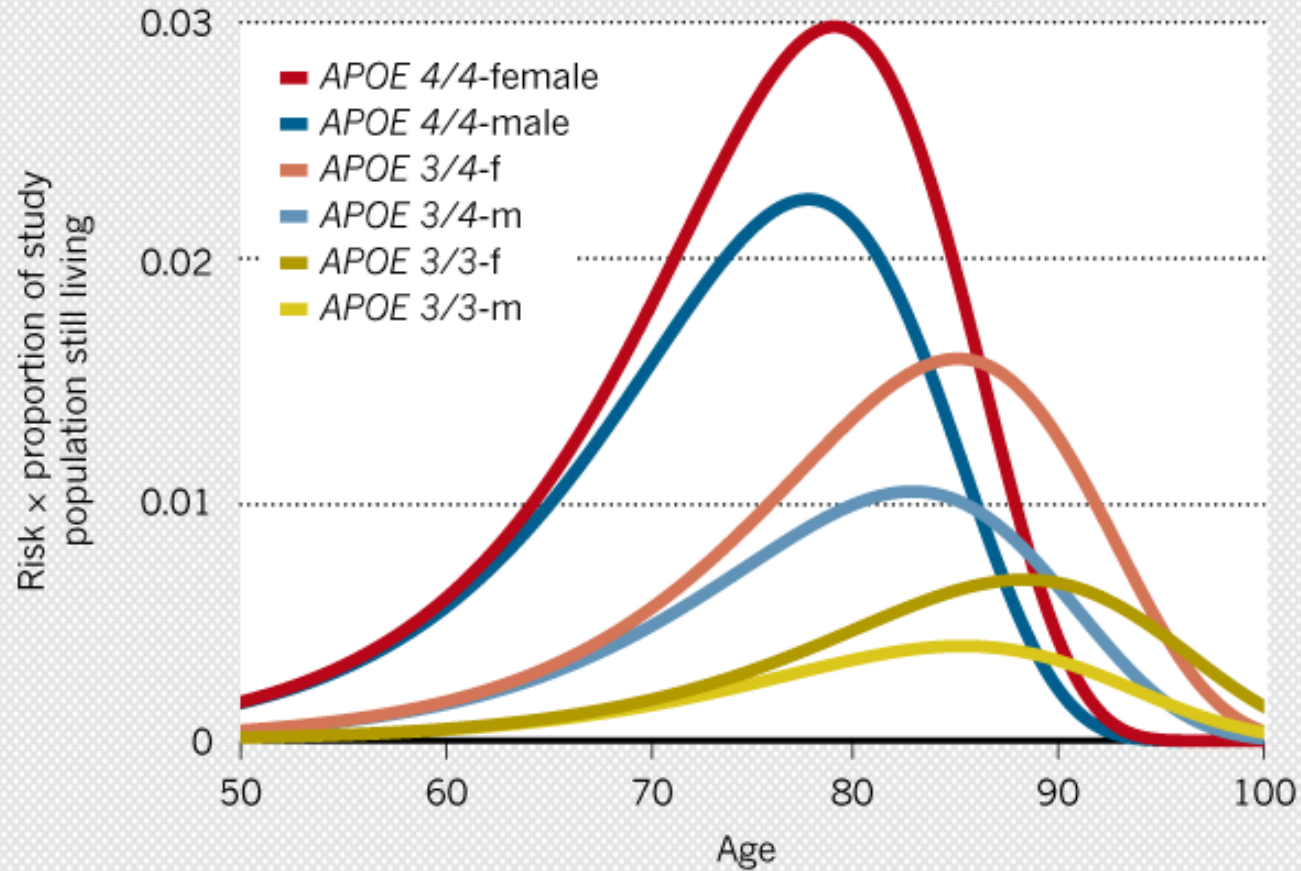


# Manhattan plot



# RISKY INHERITANCE

People who carry the gene variant *APOE4* tend to develop Alzheimer's at a younger age than those with two copies of *APOE3*.



# Example of association tests in industry



OUR SERVICES ▾

HOW IT WORKS ▾

REPORTS

STORIES

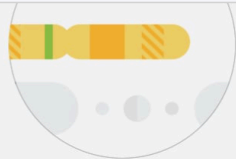
SHOP



SIGN IN

REGISTER KIT

HELP ▾



SCANDINAVIAN  
34.5%



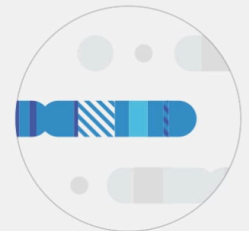
What is your DNA story?  
75+ reports on health, traits and ancestry.

shop now

A trademark of Ziff Davis, LLC Used under license.  
Reprinted with permission. © 2018 Ziff Davis, LLC. All Rights Reserved.



SWEET VS SALTY  
PREFERENCE



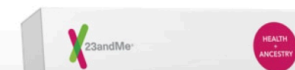
LACTOSE  
INTOLERANCE

Ancestry Service



RECOMMENDED

Health + Ancestry Service

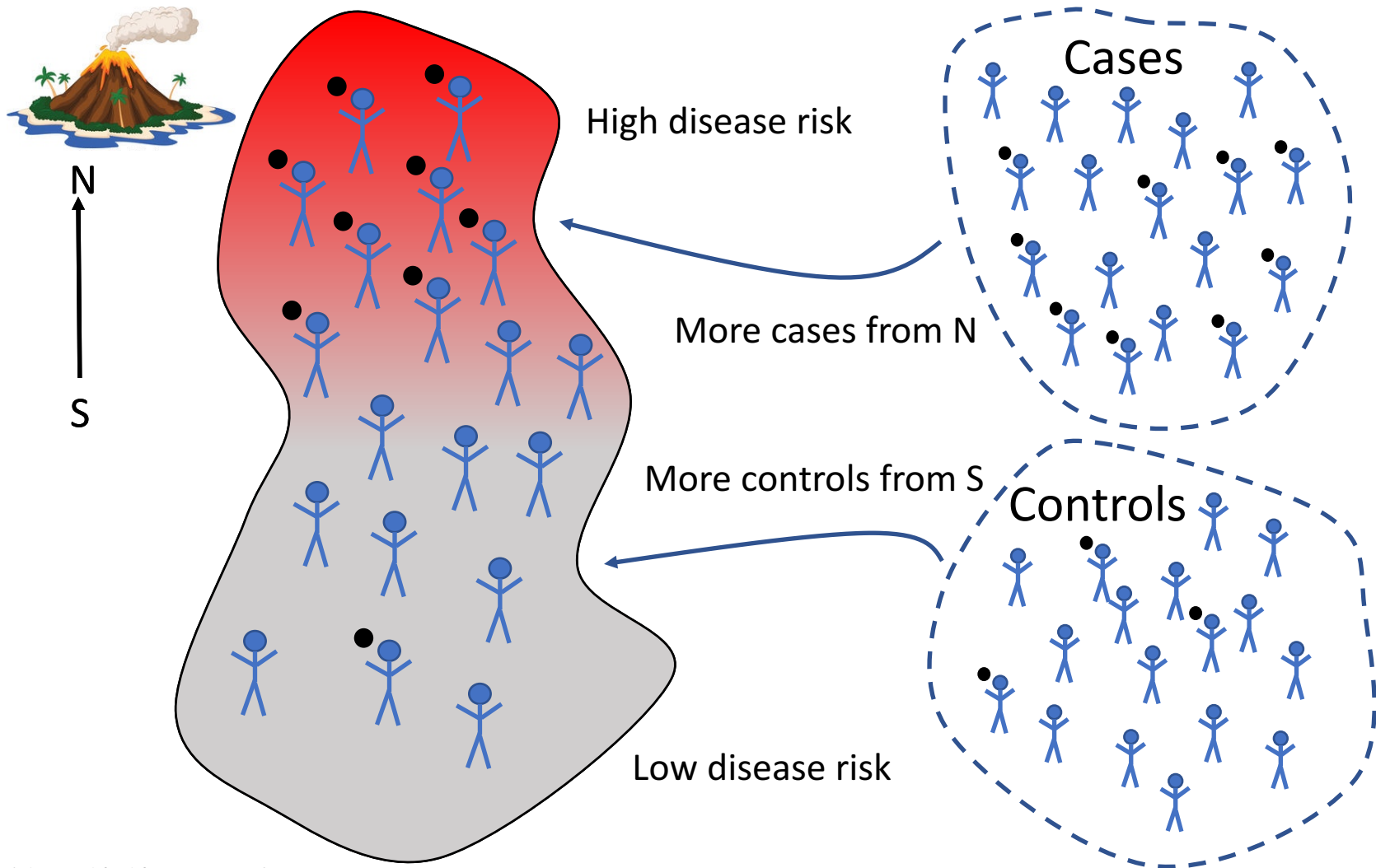


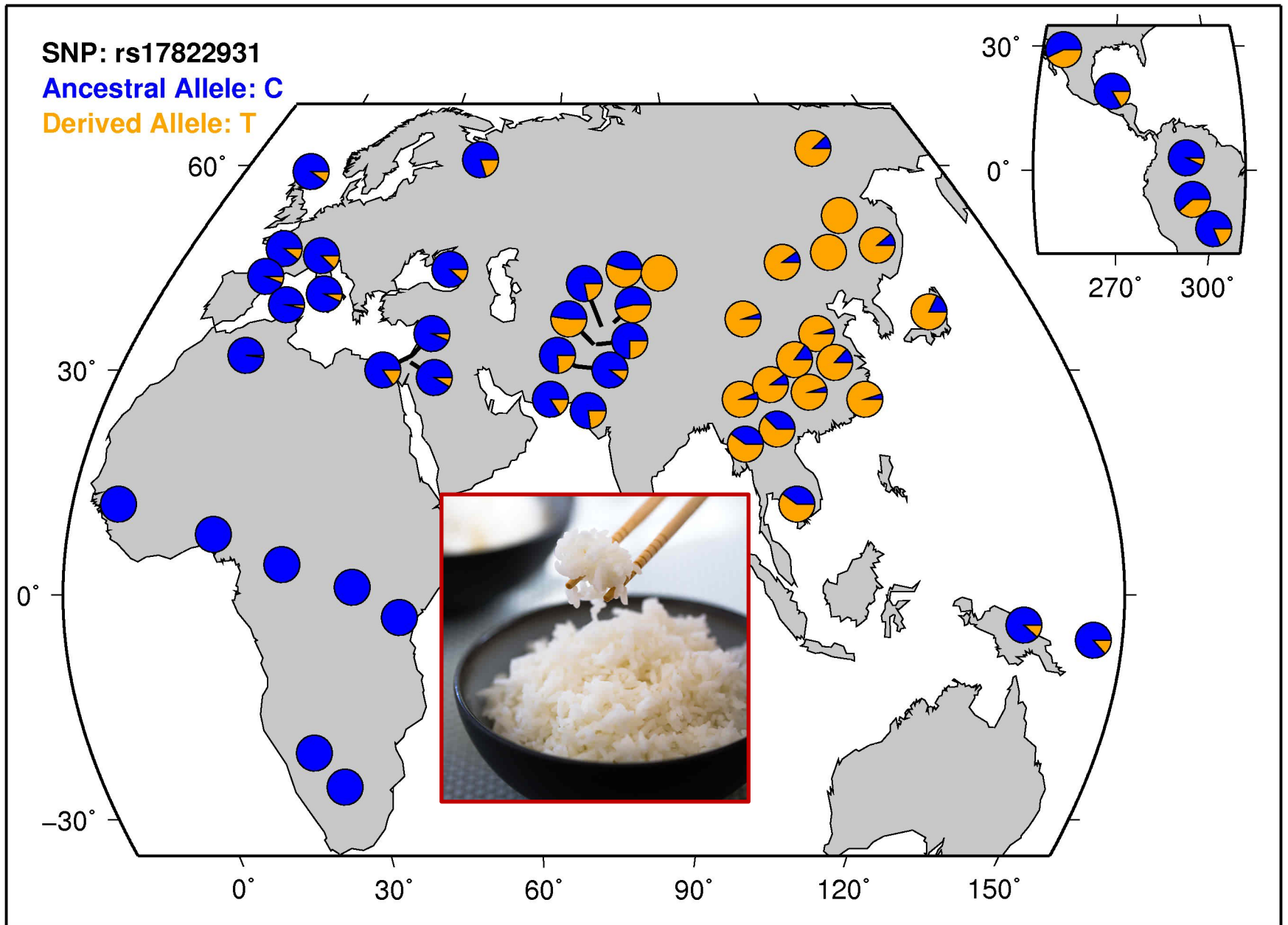
NOW WITH

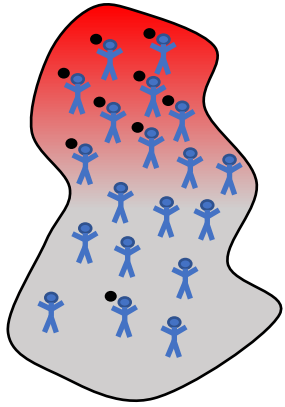
23andme.com

Impact of population structure  
and “genome-wide” testing

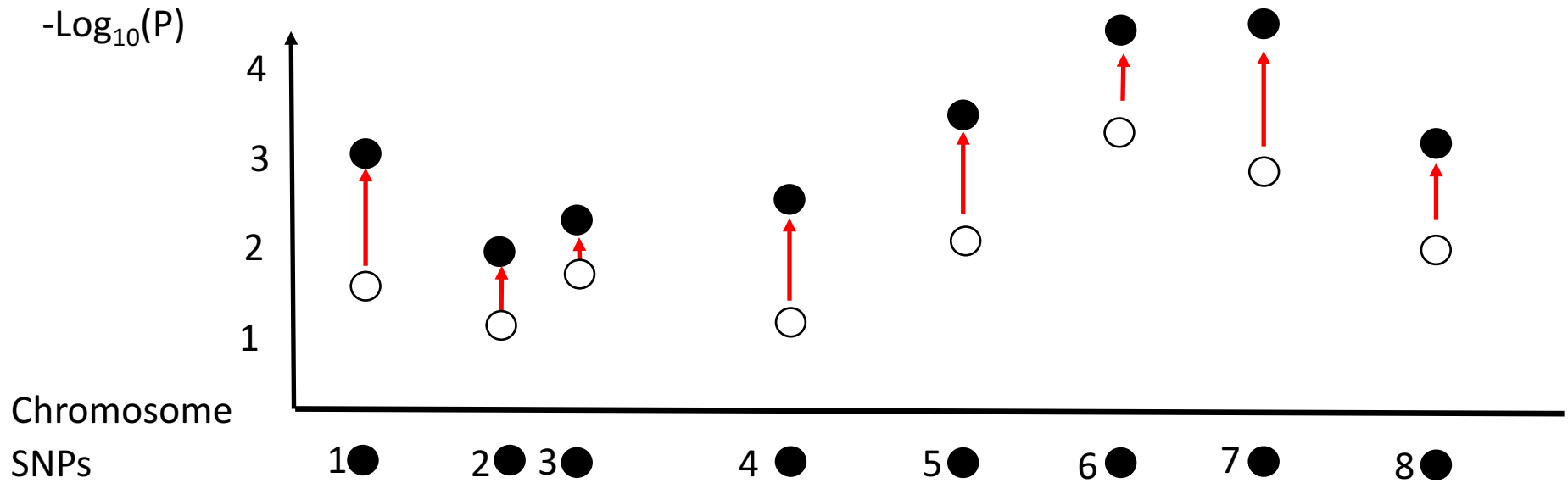
# Population structure







# Population structure



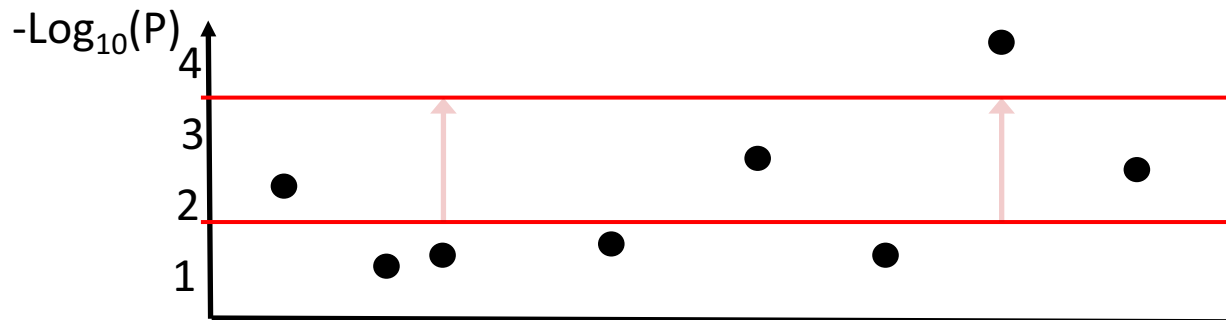
# Multiple testing

$P < 0.05$  means that there is less than a 5% chance that the result happens by chance.

	Cases	Controls
Has variant	9	3
No variant	8	14

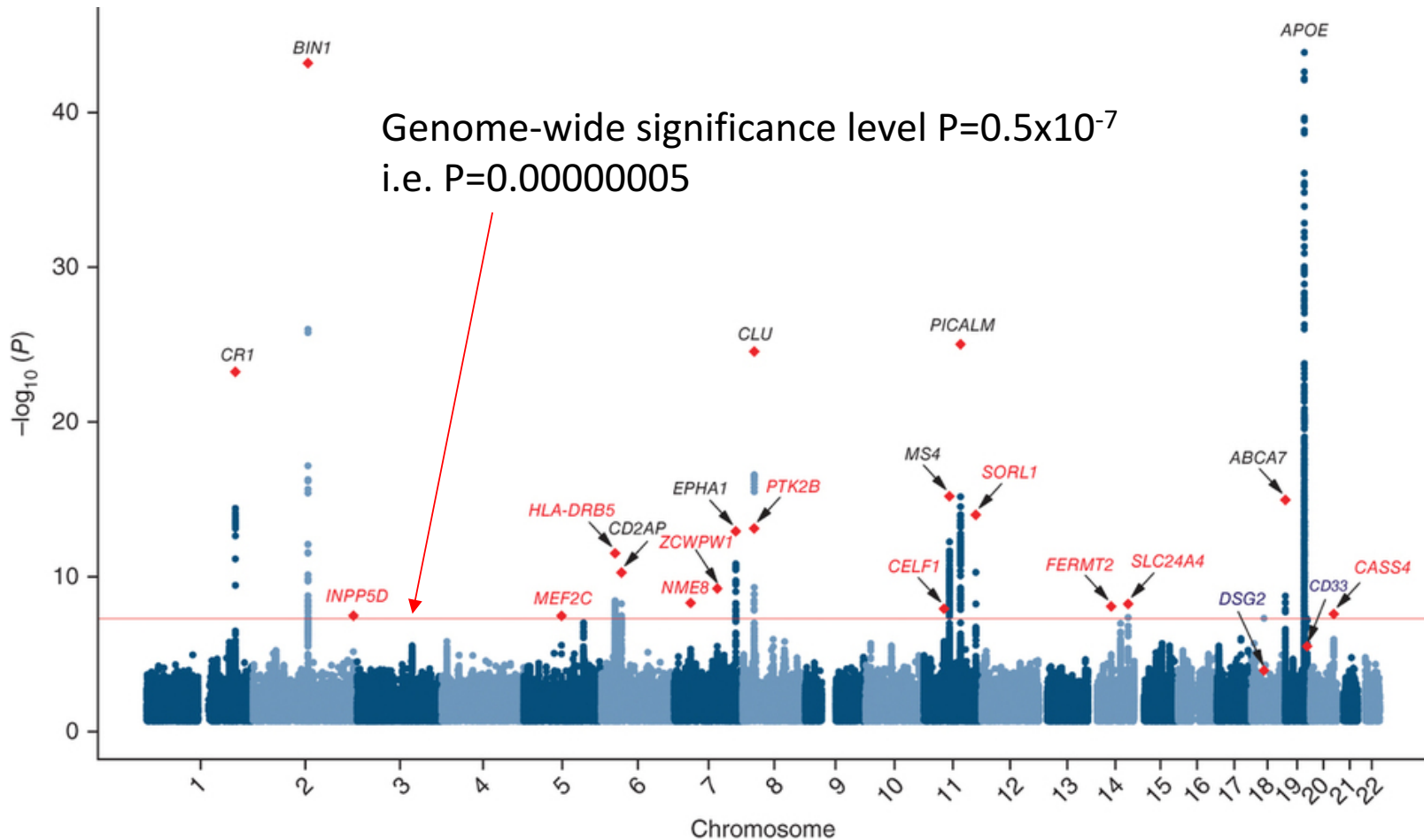
$P=0.03$

But if you try lots of tests, then the chance that one of them is significant is high  
So we need to only look at things that are extremely significant



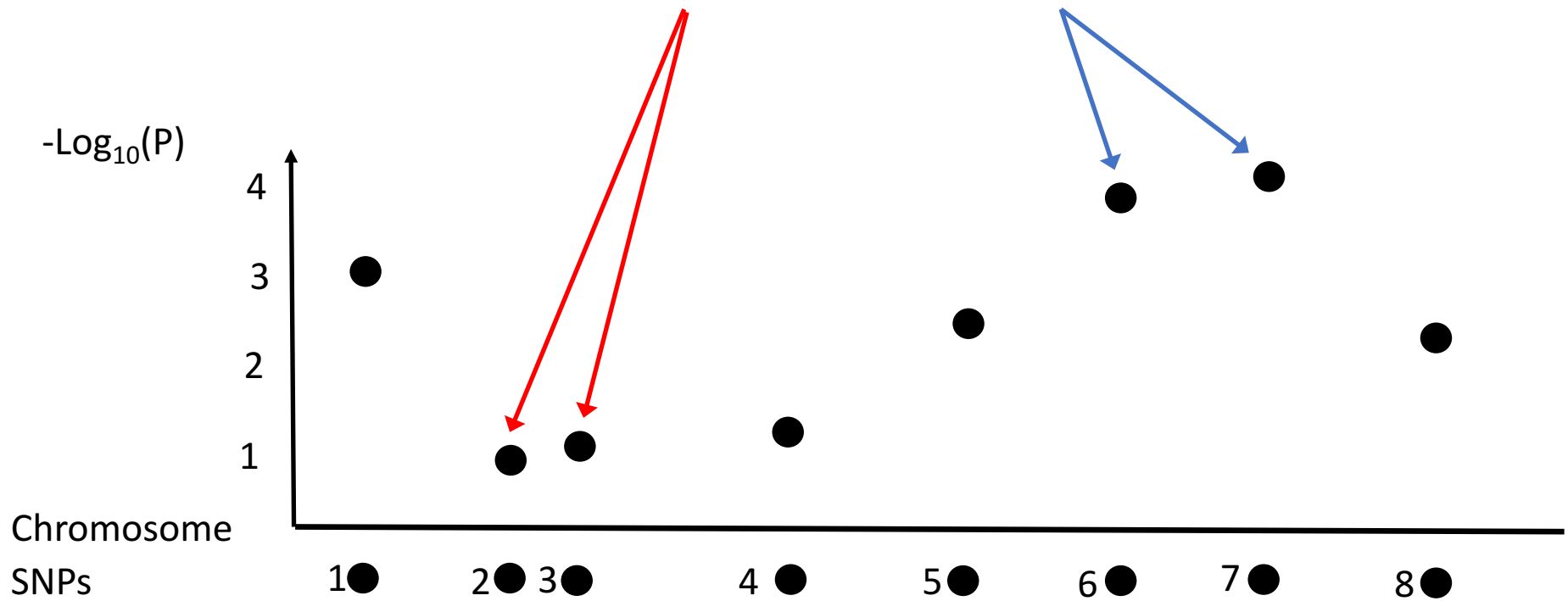


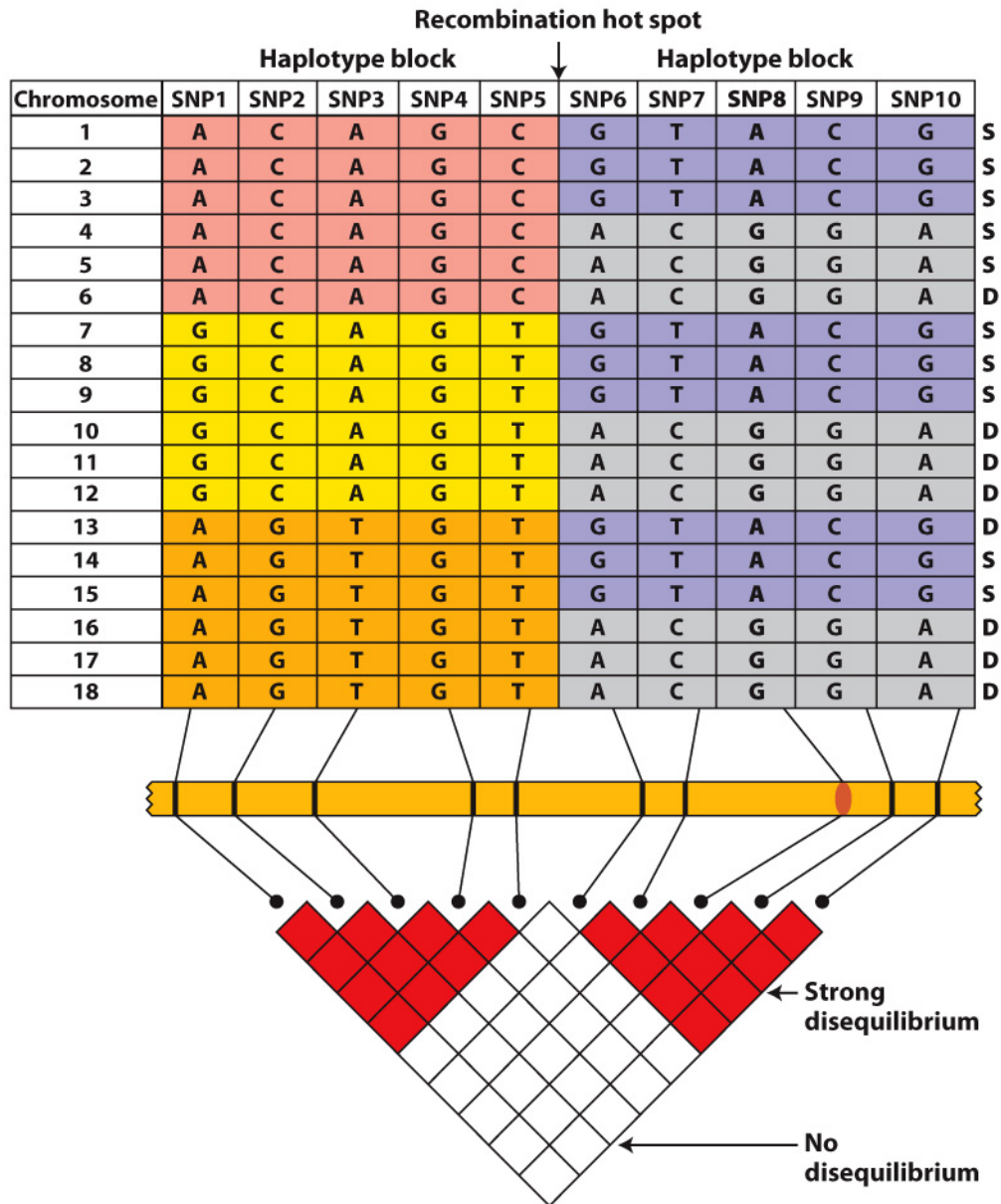
# Multiple testing



# Linkage

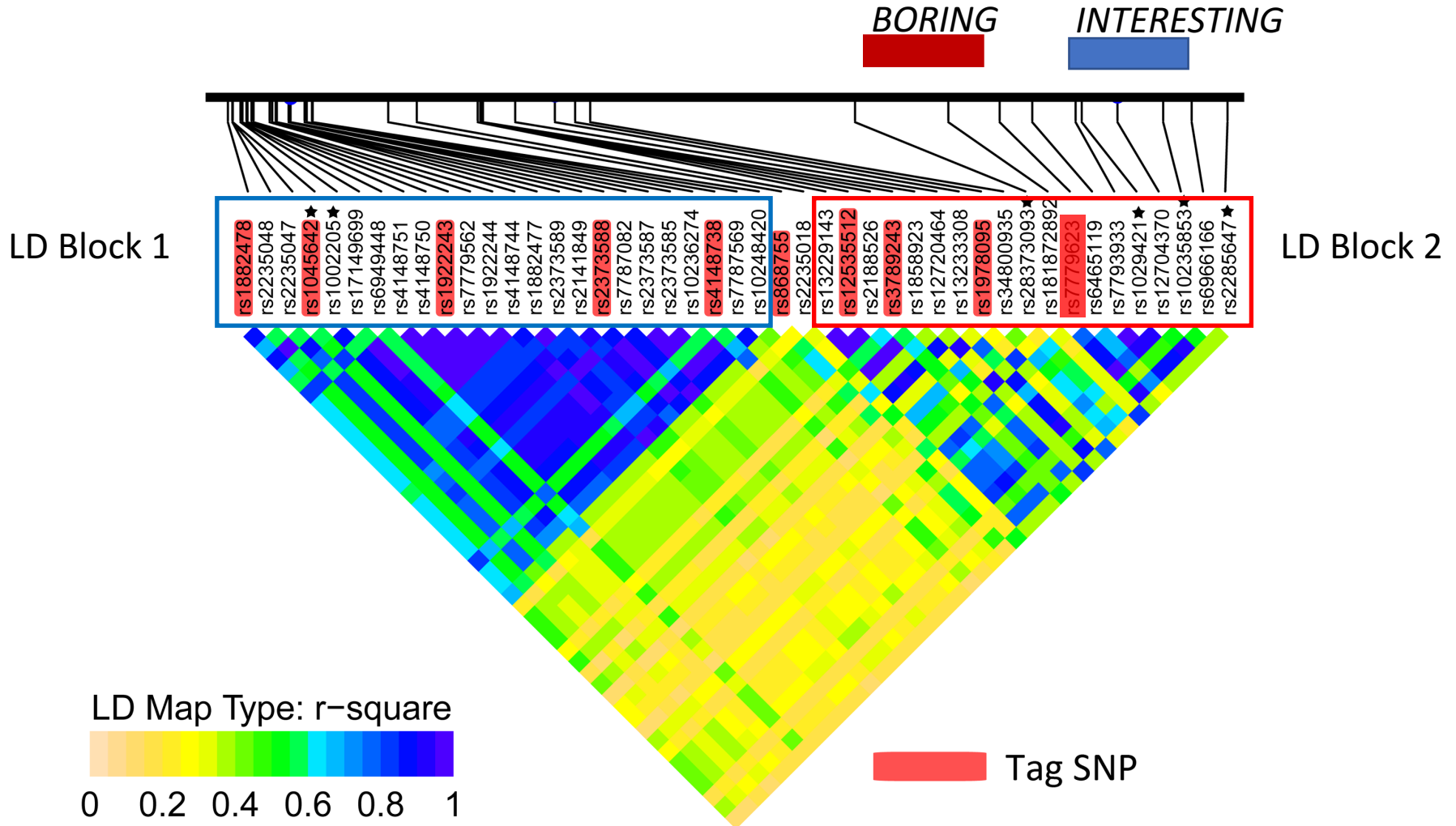
SNPs that are close together tend to behave similarly, not broken up by recombination!



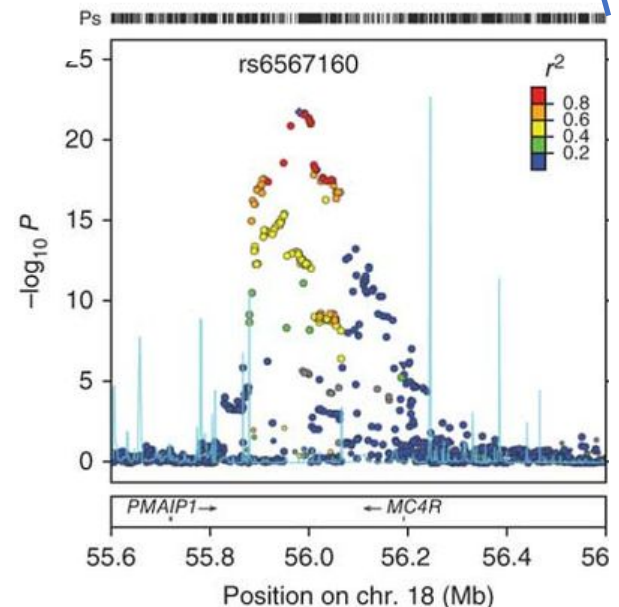
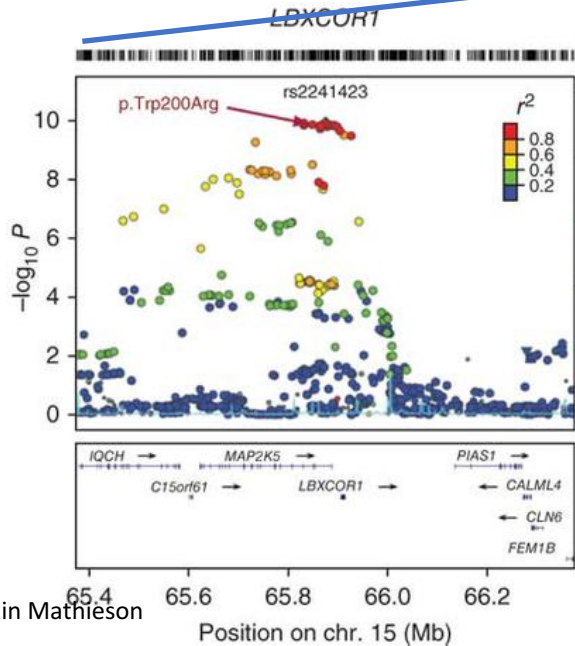
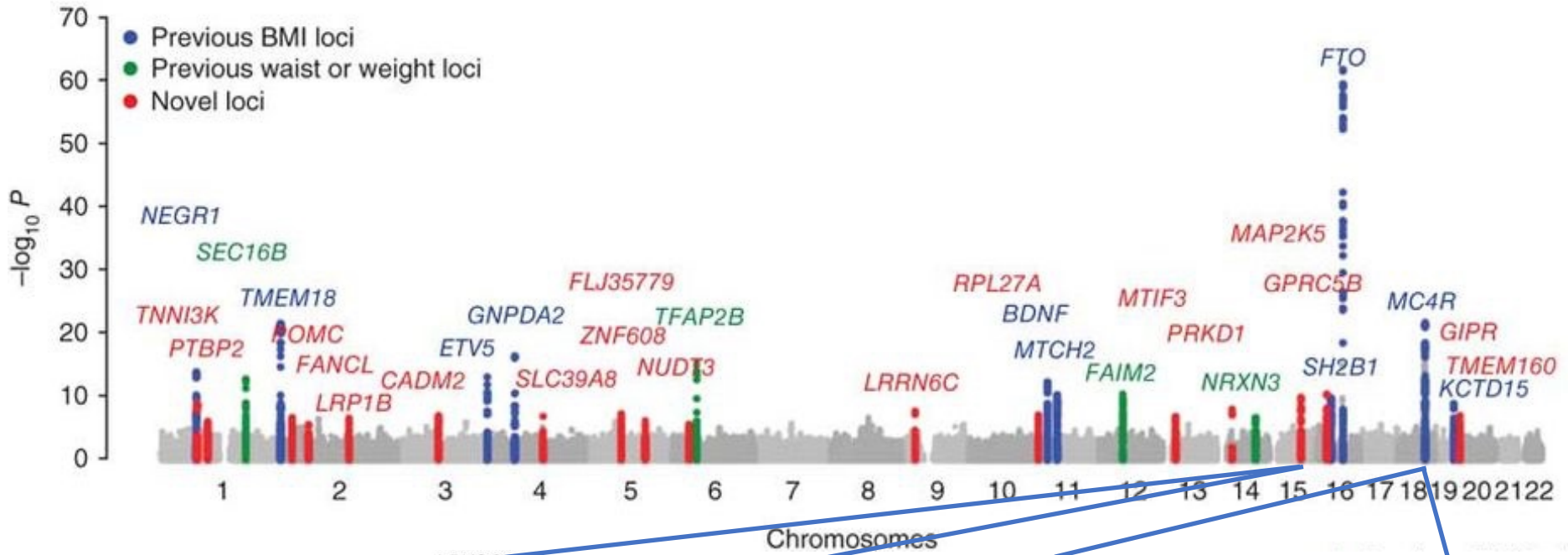


**Figure 19-17**  
*Introduction to Genetic Analysis*, Eleventh Edition  
 © 2015 W. H. Freeman and Company

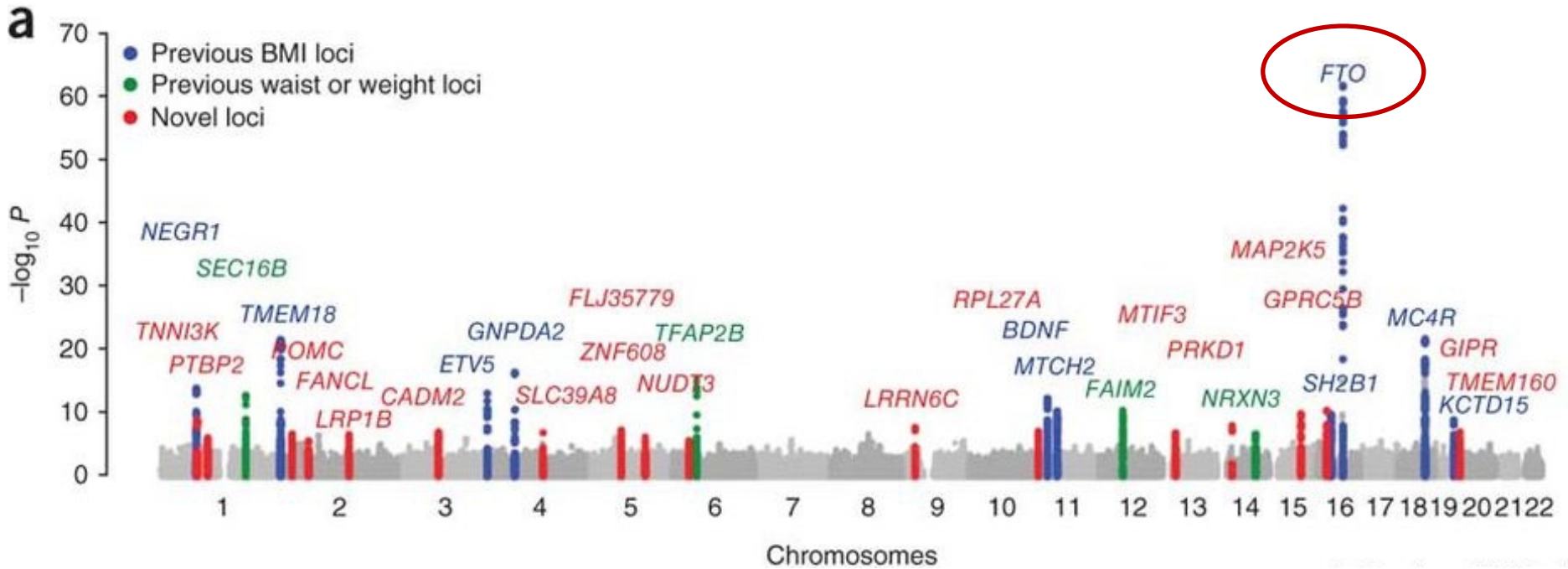
# Linkage blocks and tag SNPs



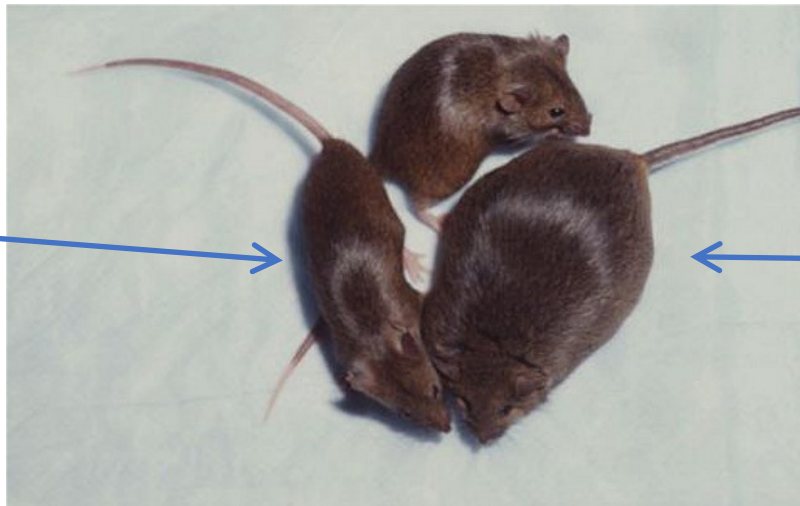
# Fine-Mapping



# Function



Wild-type mouse



Mouse with extra Copies of *f<sub>to</sub>*

Actually, this particular story is more complicated....

# Fine-Mapping

Once we find an association in a linkage block, how do we identify which specific variant is affecting the trait. “What is the causal variant?”

- Sequence the whole region so that we can find all variants, not just the tag SNPs
- Use functional information – e.g. information about which variants affect gene expression or protein function
- Use prior information about what genes are likely to be associated with a trait (but now we are back to step 1)

What have we learned from GWAS?



# What is the point?

Two big goals of human genetics:

GWAS

Goal 1: Identify genetic variants (mutations, alleles) that are associated with phenotype, particularly disease

Goal 2: Understand the biological mechanisms through which those variants act.

Hard!

Before we start: How do we know that *any* genetic variants that are associated with phenotype?



Slide: modified from Iain Mathieson



# Estimating narrow-sense heritability ( $h^2$ )

Broad-sense heritability:  $H^2 = \text{Var}(\text{Genotype } G)/\text{Var}(\text{Phenotype } P)$

Where:  $P = \text{Genotype } (G) + \text{Environment } (E)$

In other words, what fraction of variation can be explained by genetics?

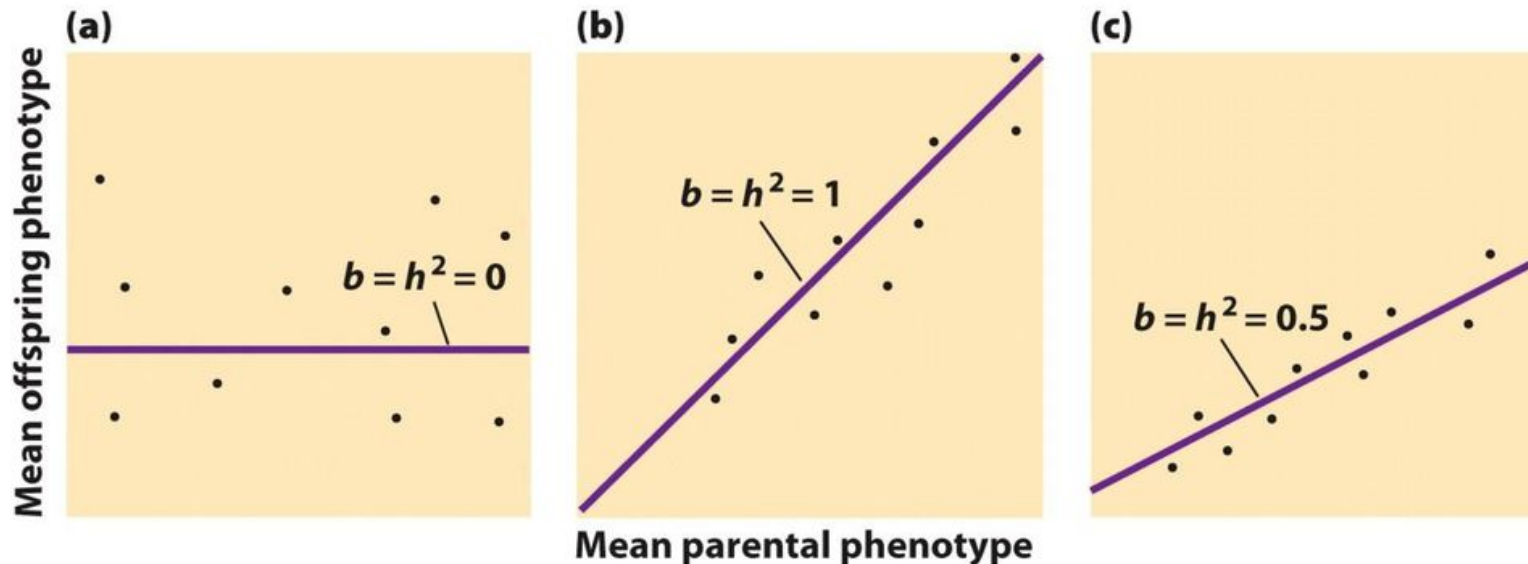
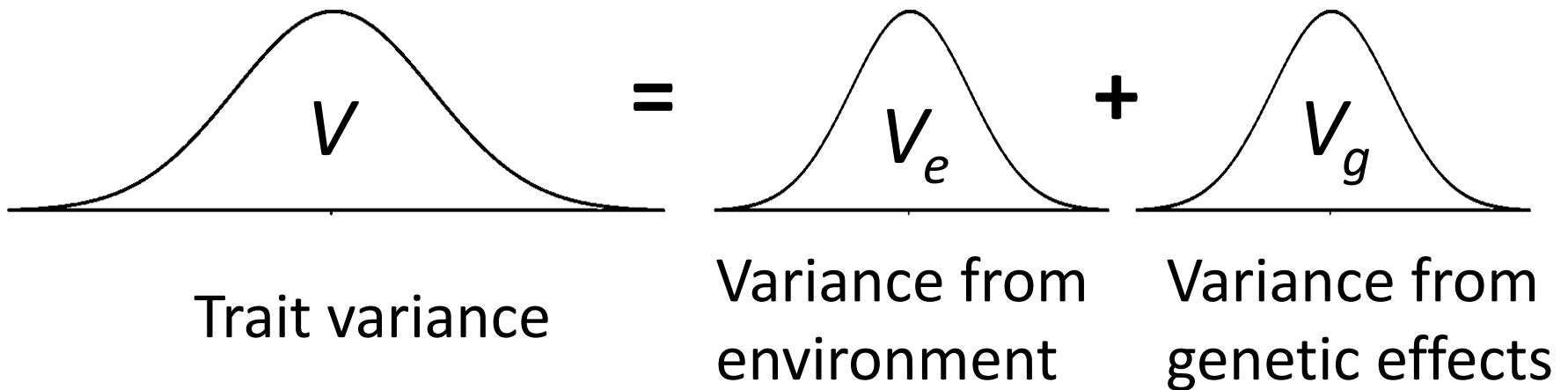


Figure 24.17  
*Genetics: A Conceptual Approach, Fifth Edition*  
© 2014 W. H. Freeman and Company

# [Narrow-sense] Heritability

**What proportion of the variance in a trait is explained by additive genetic effects?**



Does not include: Recessive/dominance effects (included in "broad sense" heritability), gene-environment interactions

# Heritability ( $h^2$ )

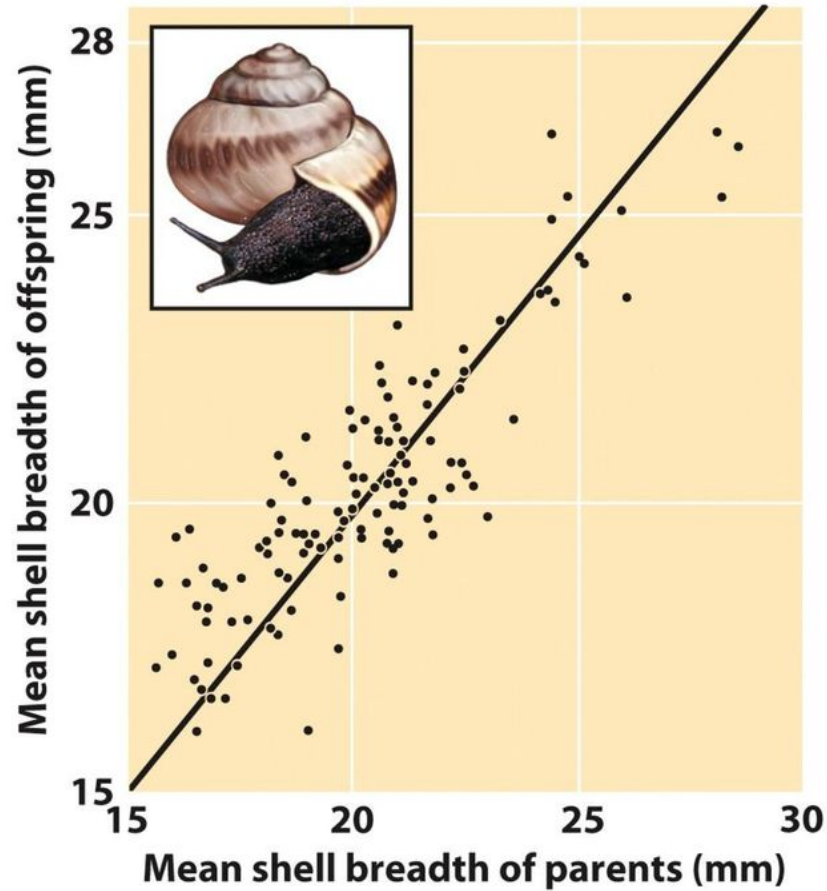
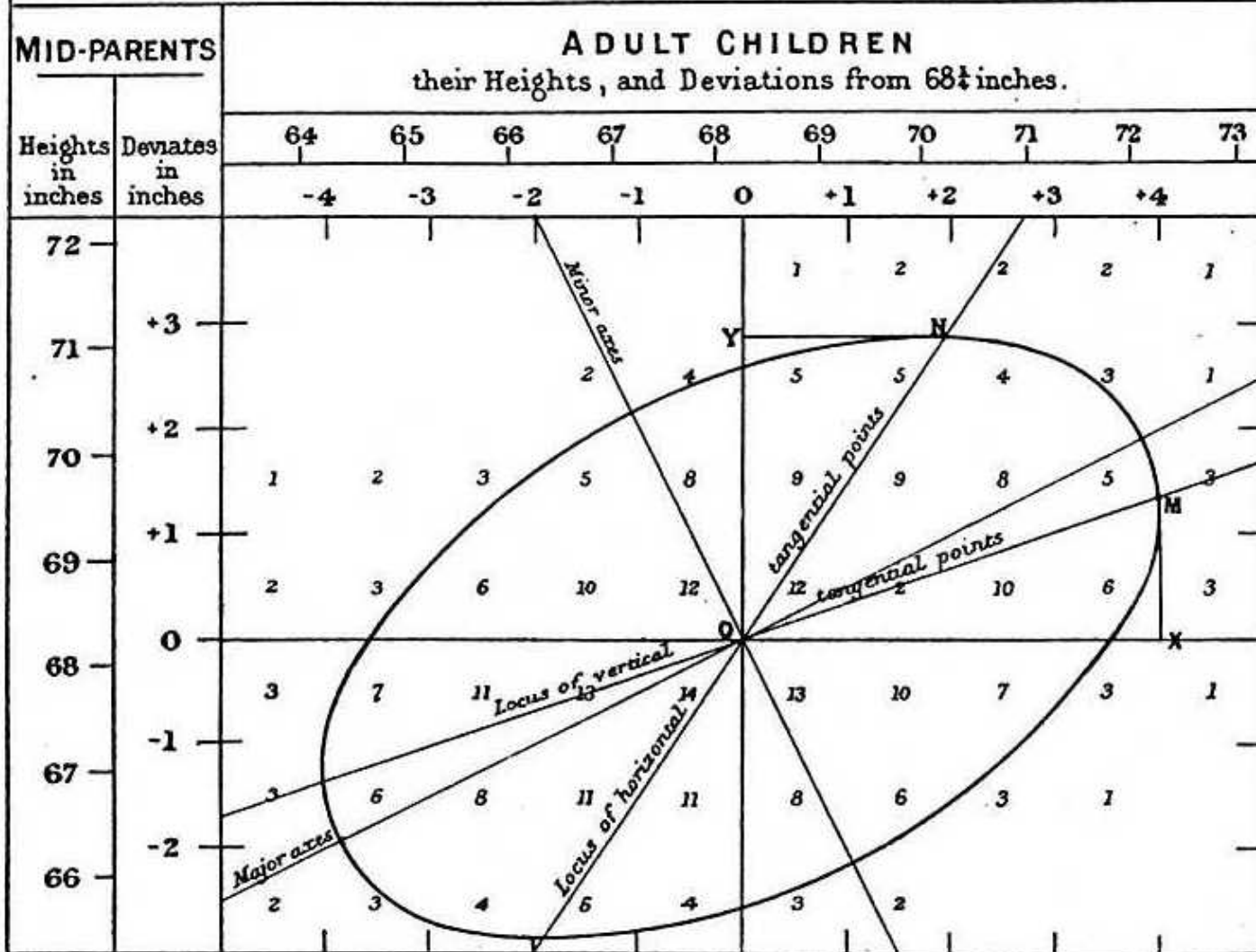


Figure 24.18  
*Genetics: A Conceptual Approach*, Fifth Edition  
© 2014 W. H. Freeman and Company

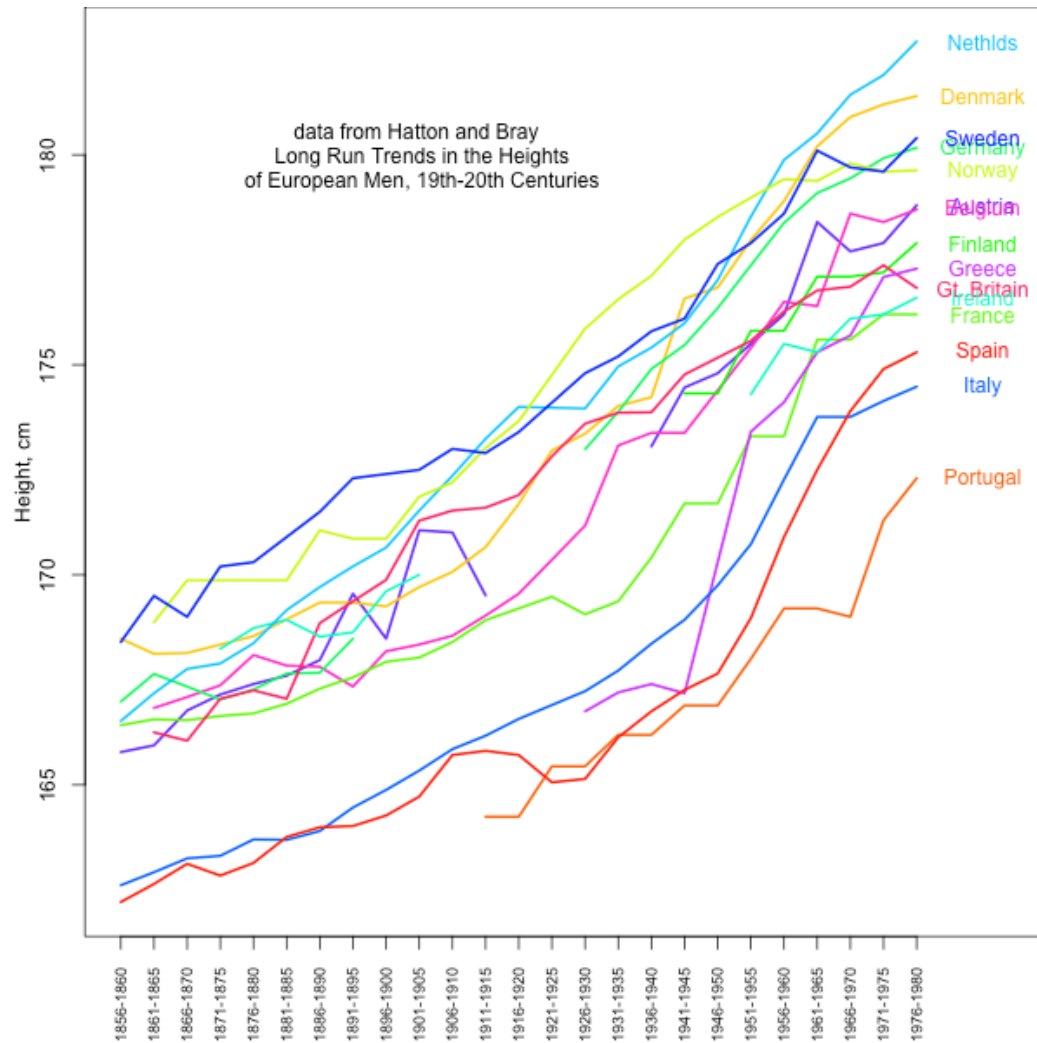
# DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1.08)



Galton 1886

# High heritability does not mean low environmental effect



Height in European men over time (by Graham Coop)



# Heritability

Heritability typically estimated by comparing relatives. Particularly *twin studies* comparing monozygotic and dizygotic twins.

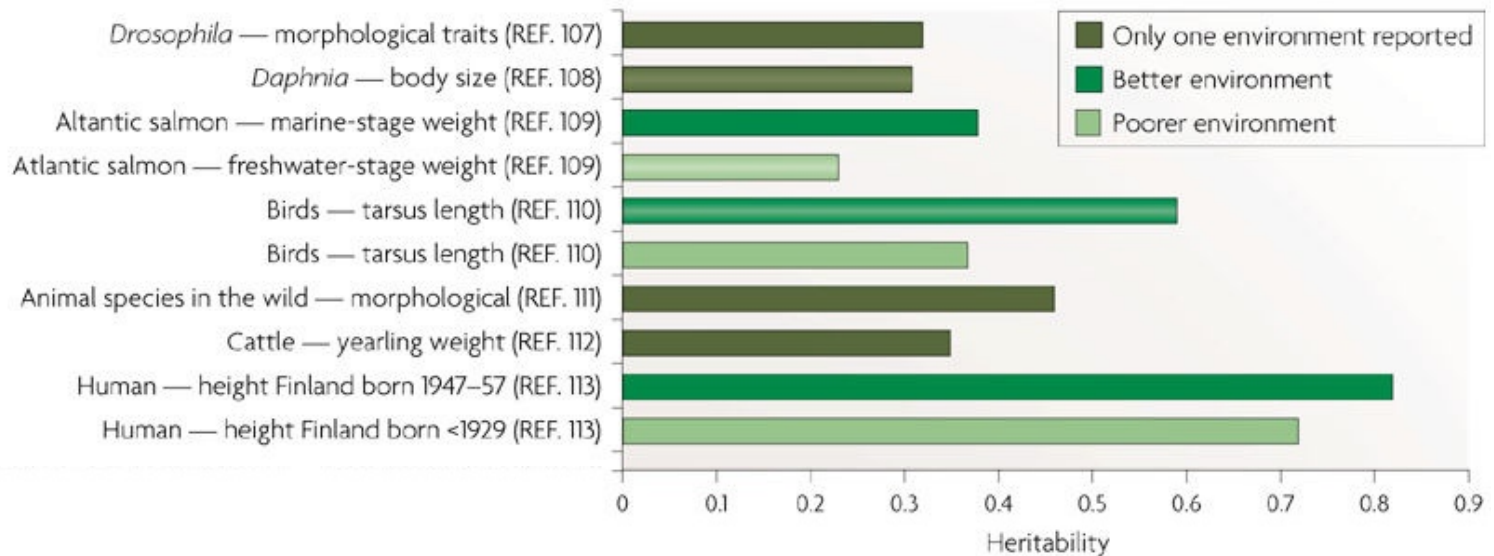
Heritability estimates:

Height : 0.7-0.8

BMI : 0.4-0.8

IQ : 0.4-0.7

## Morphological traits



# Heritability and response to selection

Highly heritable traits can be selected for.

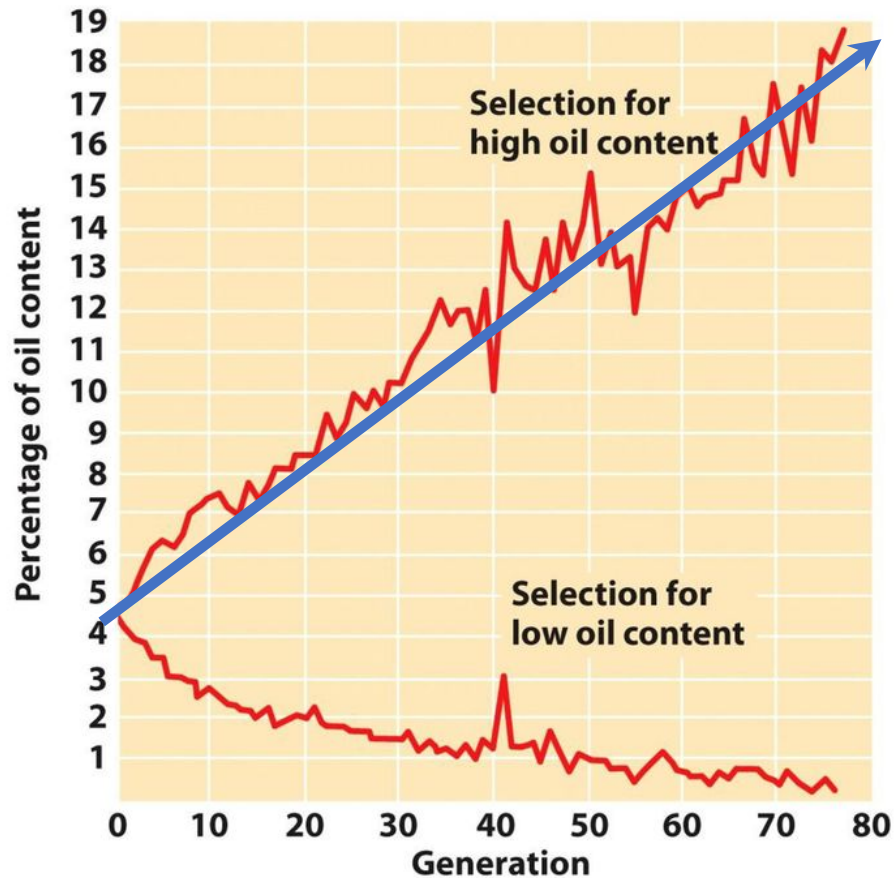
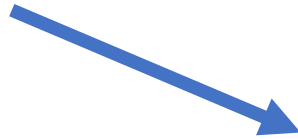
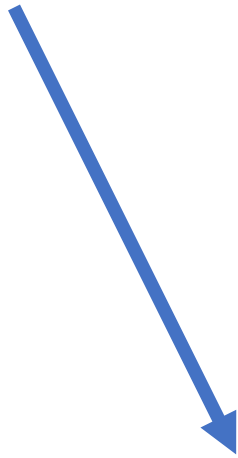
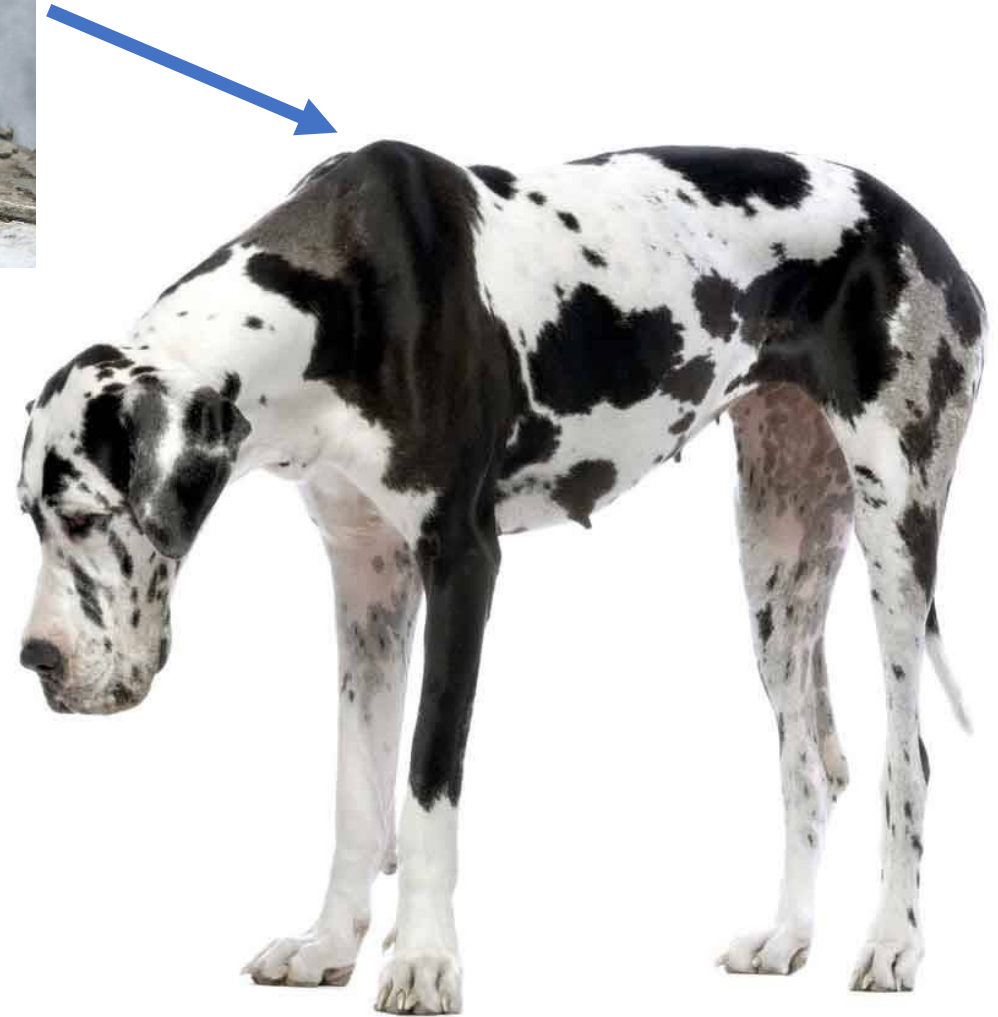
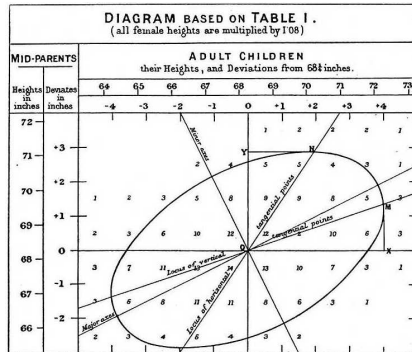


Figure 24.22  
*Genetics: A Conceptual Approach, Fifth Edition*  
© 2014 W. H. Freeman and Company

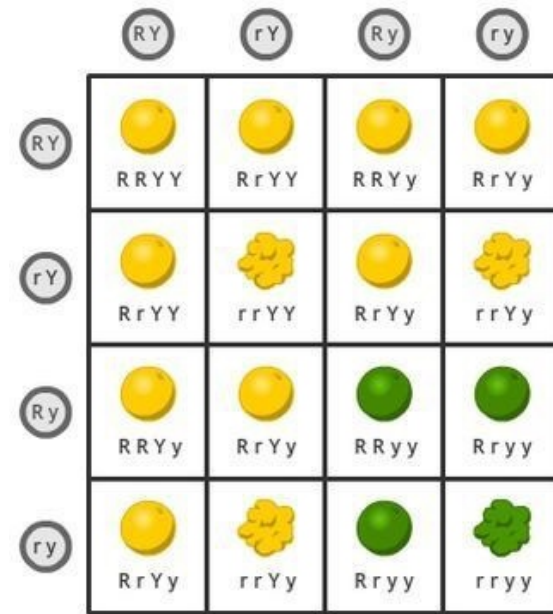
# Artificial Selection works!



# How to reconcile with Mendelian inheritance?

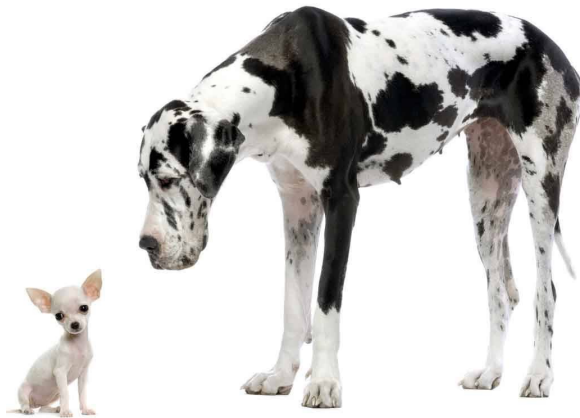


?



9 Yellow, round  
3 Green, round  
3 Yellow, wrinkled  
1 Green, wrinkled

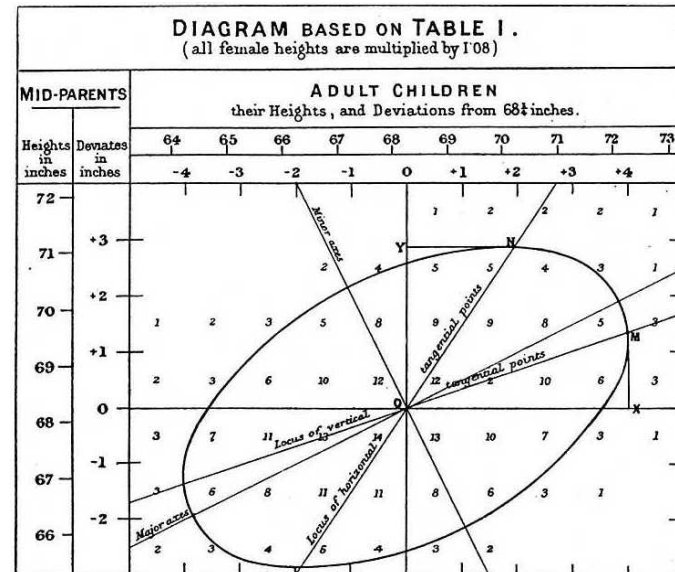
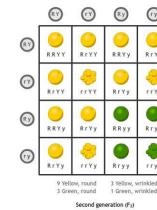
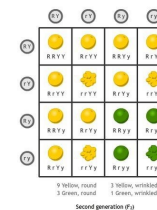
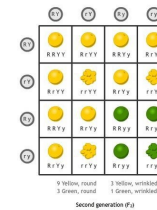
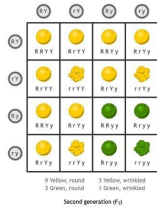
Second generation (F<sub>2</sub>)



“Blending inheritance”

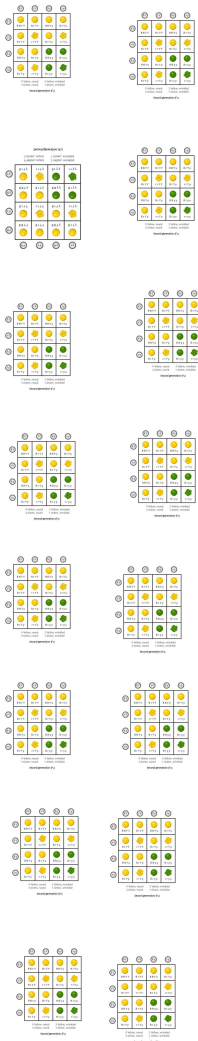
Mendelian inheritance

# More than one locus?

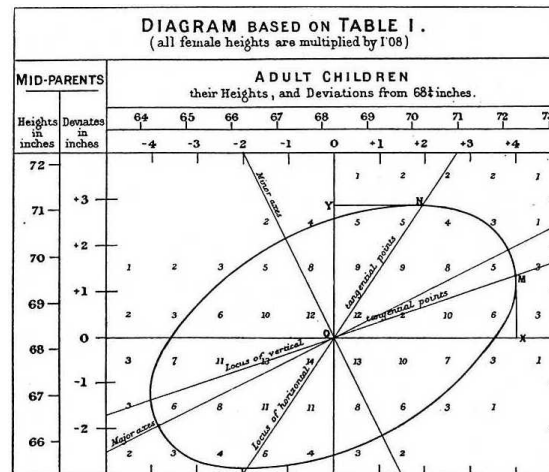


Many Mendelian factors affecting the trait?

# The infinitesimal model 1918



Fisher



XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. By R. A. Fisher, B.A. Communicated by Professor J. ARTHUR THOMSON. (With Four Figures in Text.)

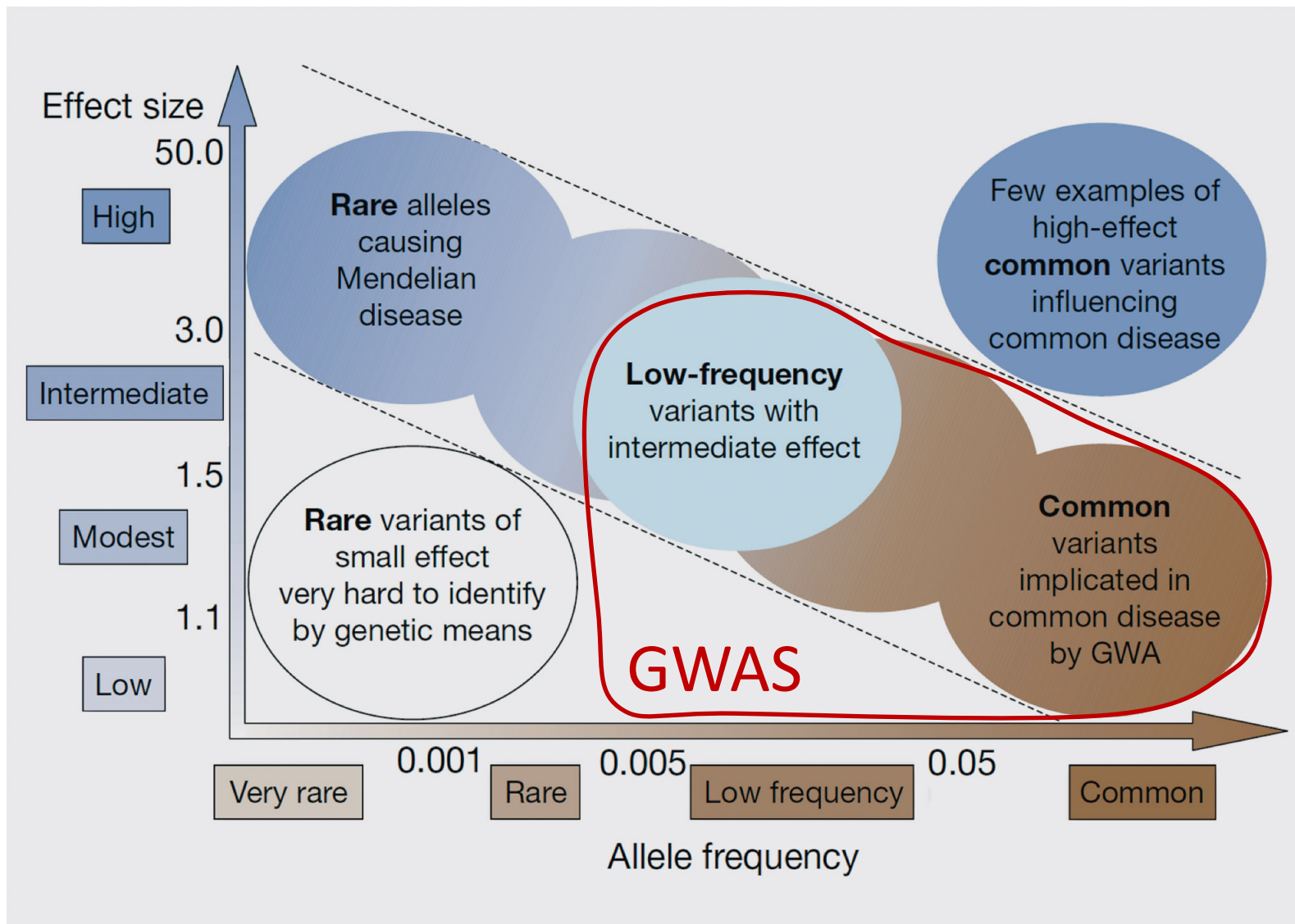
(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

	PAGE		PAGE
1. The superposition of factors distributed independently . . . . .	402	15. Homogamy and multiple allelomorphism . . . . .	416
2. Phase frequency in each array . . . . .	402	16. Coupling . . . . .	418
3. Parental regression . . . . .	403	17. Theories of marital correlation; ancestral correlations . . . . .	419
4. Dominance deviations . . . . .	403	18. Ancestral correlations (second and third theories) . . . . .	421
5. Correlation for parent; genetic correlations . . . . .	404	19. Numerical values of association . . . . .	421
6. Fraternal correlation . . . . .	405	20. Fraternal correlation . . . . .	422
7. Correlations for other relatives . . . . .	406	21. Numerical values for environment and dominance ratios; analysis of variance . . . . .	423
8. Epistasy . . . . .	408	22. Other relatives . . . . .	424
9. Assortative mating . . . . .	410	23. Numerical values (third theory) . . . . .	425
10. Frequency of phases . . . . .	410	24. Comparison of results . . . . .	427
11. Association of factors . . . . .	411	25. Interpretation of dominance ratio (diagrams) . . . . .	428
12. Conditions of equilibrium . . . . .	412	26. Summary . . . . .	432
13. Nature of association . . . . .	413		
14. Multiple allelomorphism . . . . .	415		

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this scheme. It is hoped that in this way it will be possible to make a more exact analysis of the causes of human variability. The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations  $\sigma_1$  and  $\sigma_2$ , it is found that the distribution, when both causes act together, has a standard deviation  $\sqrt{\sigma_1^2 + \sigma_2^2}$ . It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the **Variance** of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce. It

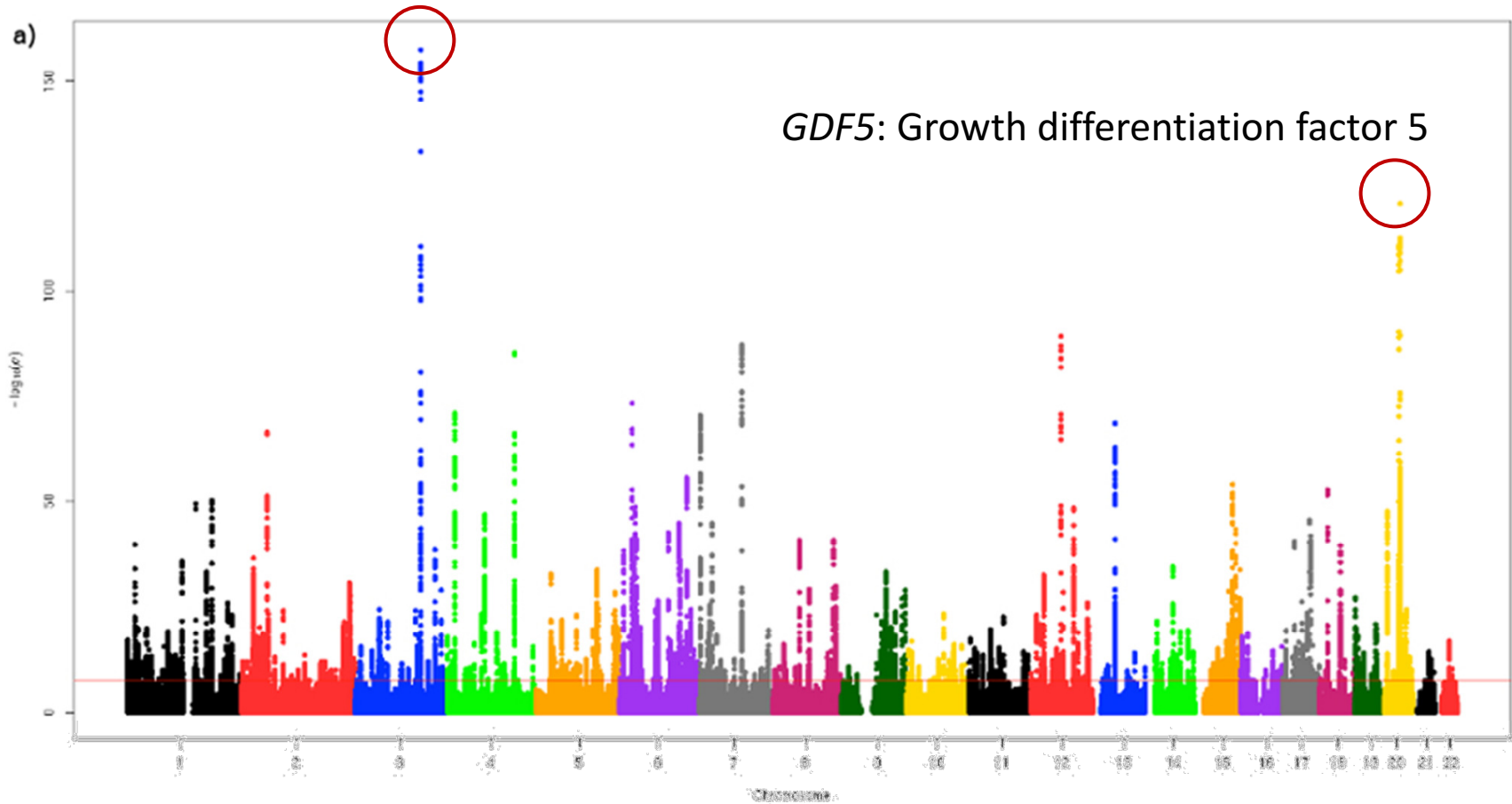
# Fast-forward 100 years: GWAS



# Fast-forward 100 years: GWAS in height

*ZBTB38*: Zinc Finger And BTB Domain Containing 38

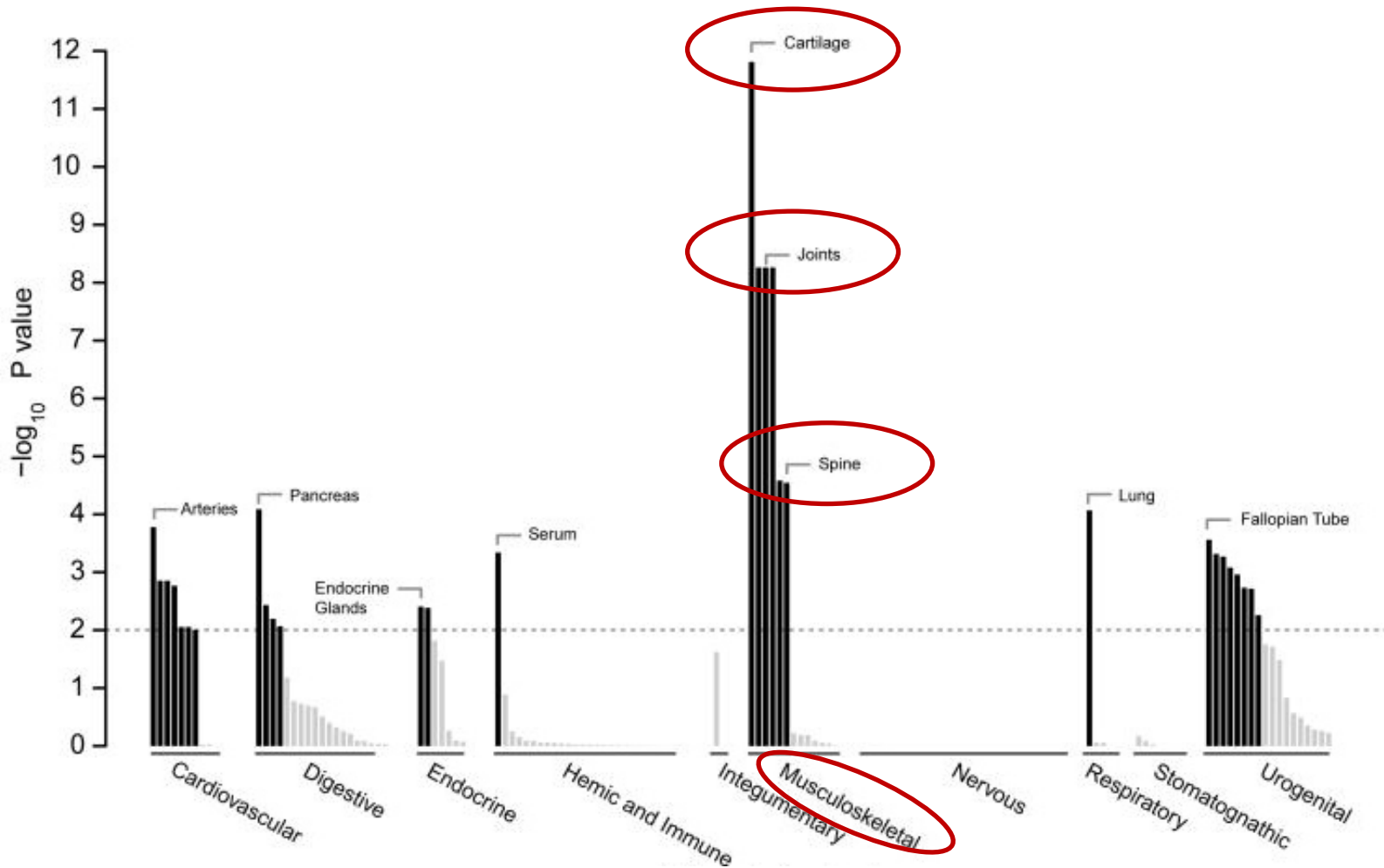
*GDF5*: Growth differentiation factor 5



697 independent SNPs significantly associated with height – Wood et al. 2014  
Together explain about 15% of the phenotypic variance

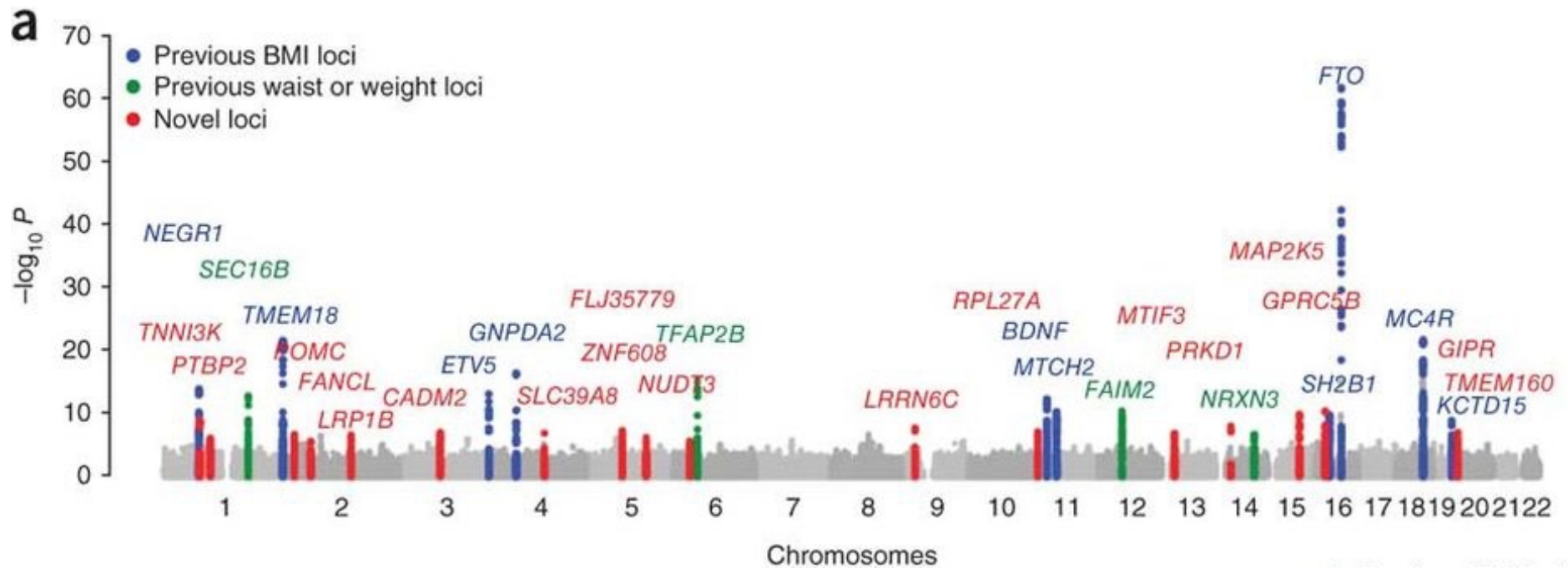


# Height-associated variants enriched in relevant genes



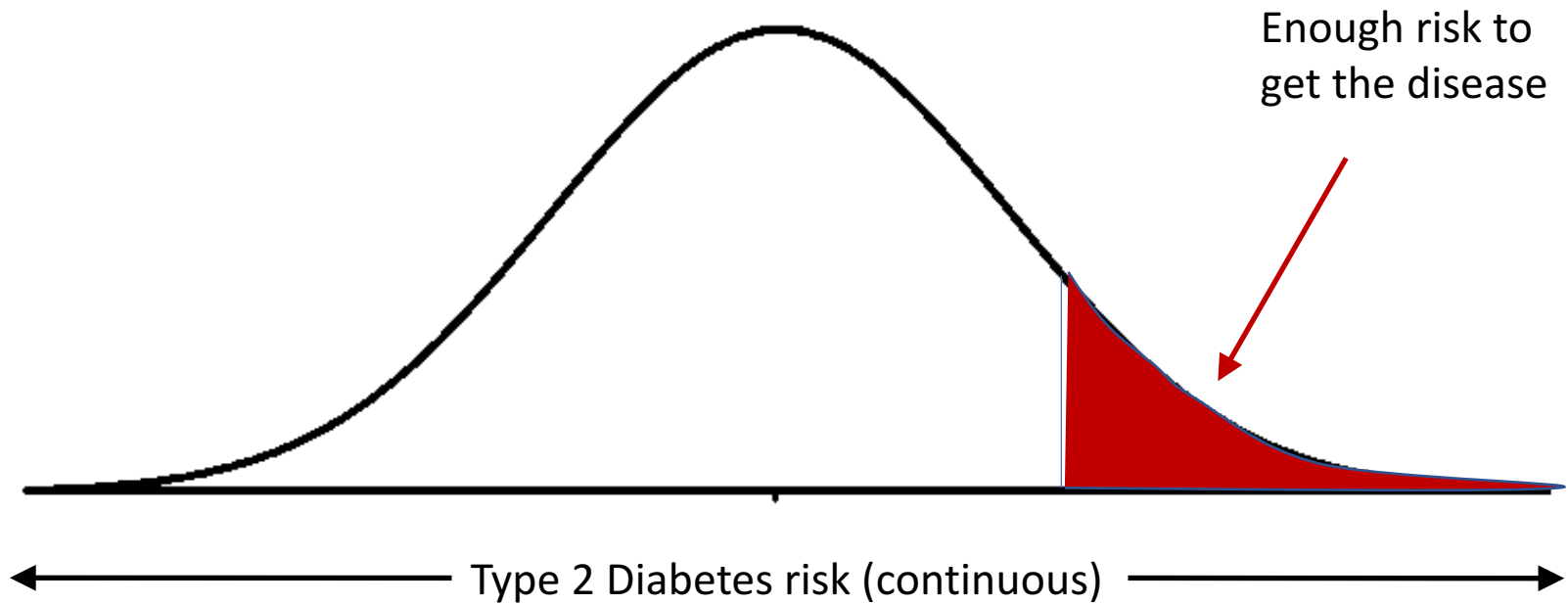
Wood et al. 2014

# BMI GWAS



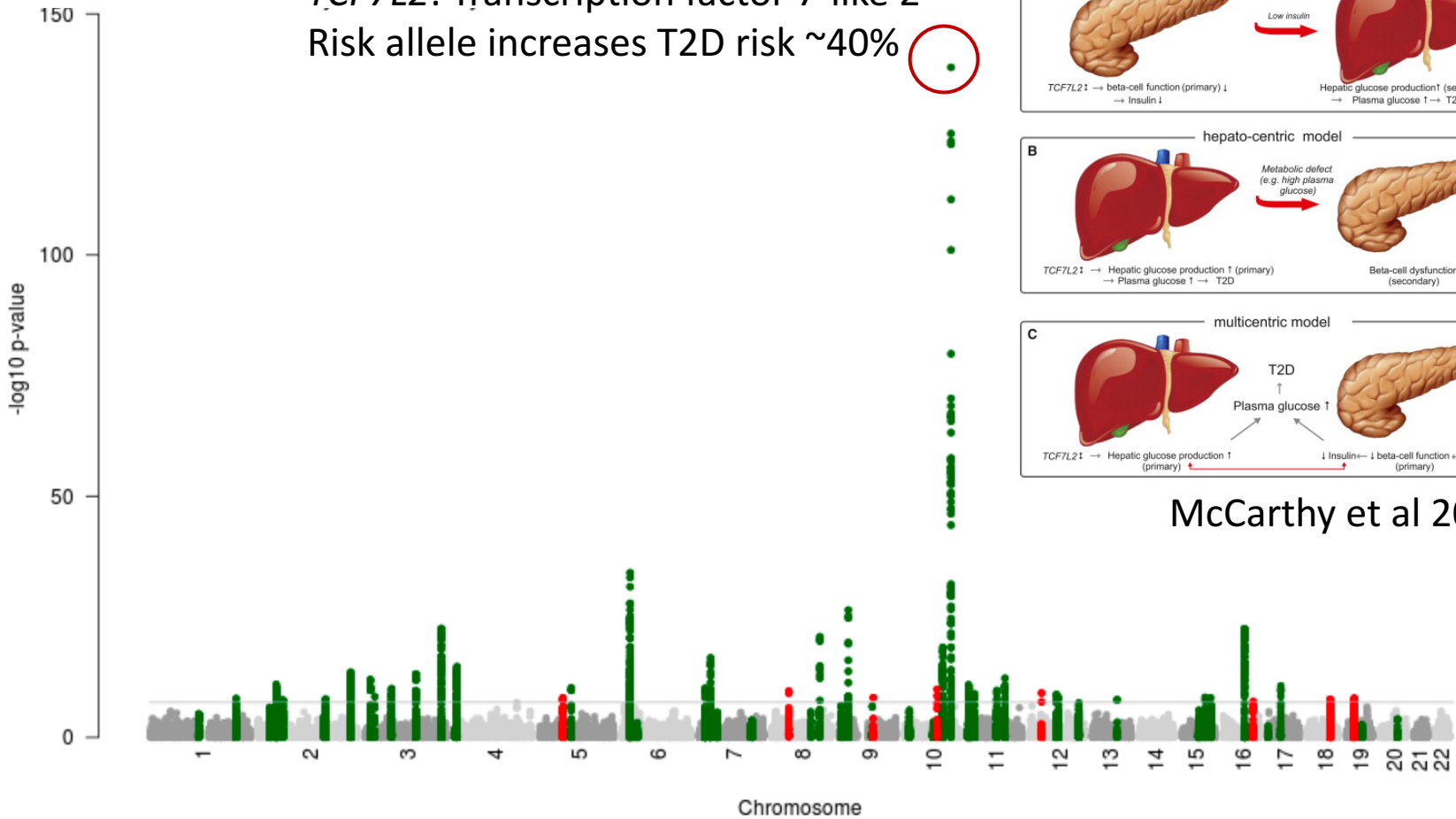
32 independent SNPs explain 1.45% of the variance in BMI – Speliotes et al. 2010

# The liability threshold model



# Type 2 Diabetes GWAS

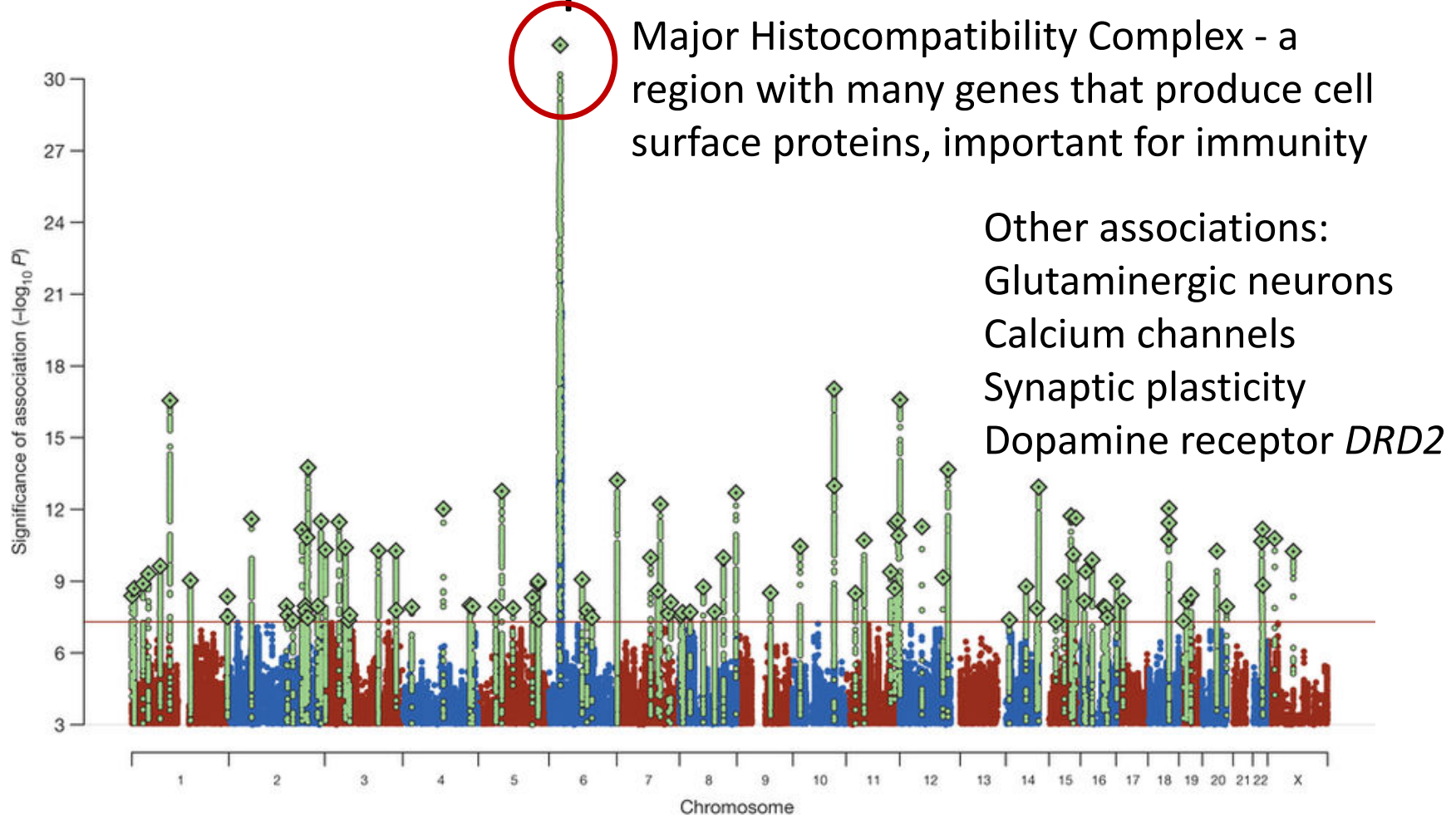
*TCF7L2*: Transcription factor 7-like 2  
Risk allele increases T2D risk ~40%



McCarthy et al 2013

63 independent loci explain 5.7% of the variance – Morris et al. 2012

# Schizophrenia GWAS



108 independent loci explain 3.4% of the variance – Ripke et al. 2014

# Missing Heritability?

NEWS FEATURE PERSONAL GENOMES

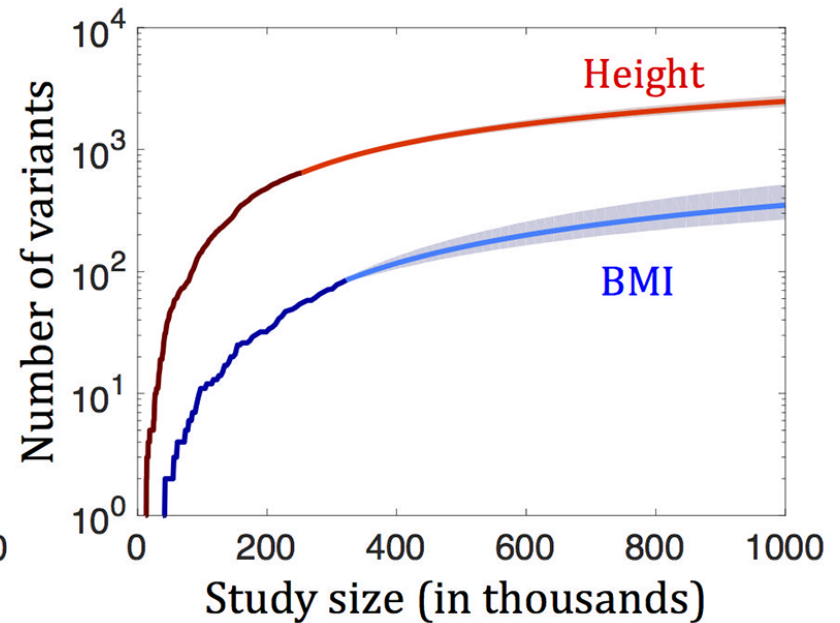
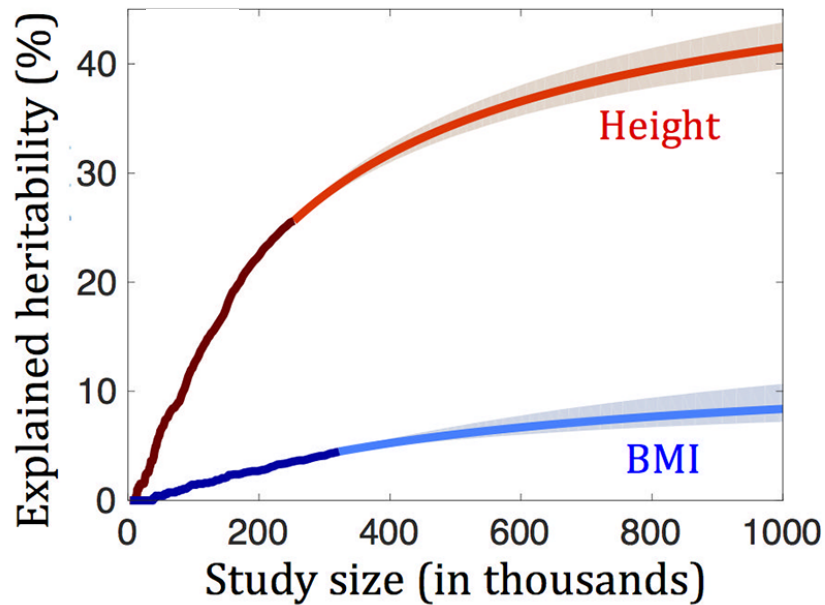
NATURE | Vol 456 | 6 November 2008



## The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

The bigger the sample size, the more variants you find



Simons & Sella 2018

# Missing Heritability?

“Missing heritability” is not really missing

Mostly just hidden in very small effects that GWAS are not big enough to detect

May be some hidden in epistatic effects or gene-environment interactions

Heritability estimates might be a bit too high



# How successful have GWAS been?

Twelve years.

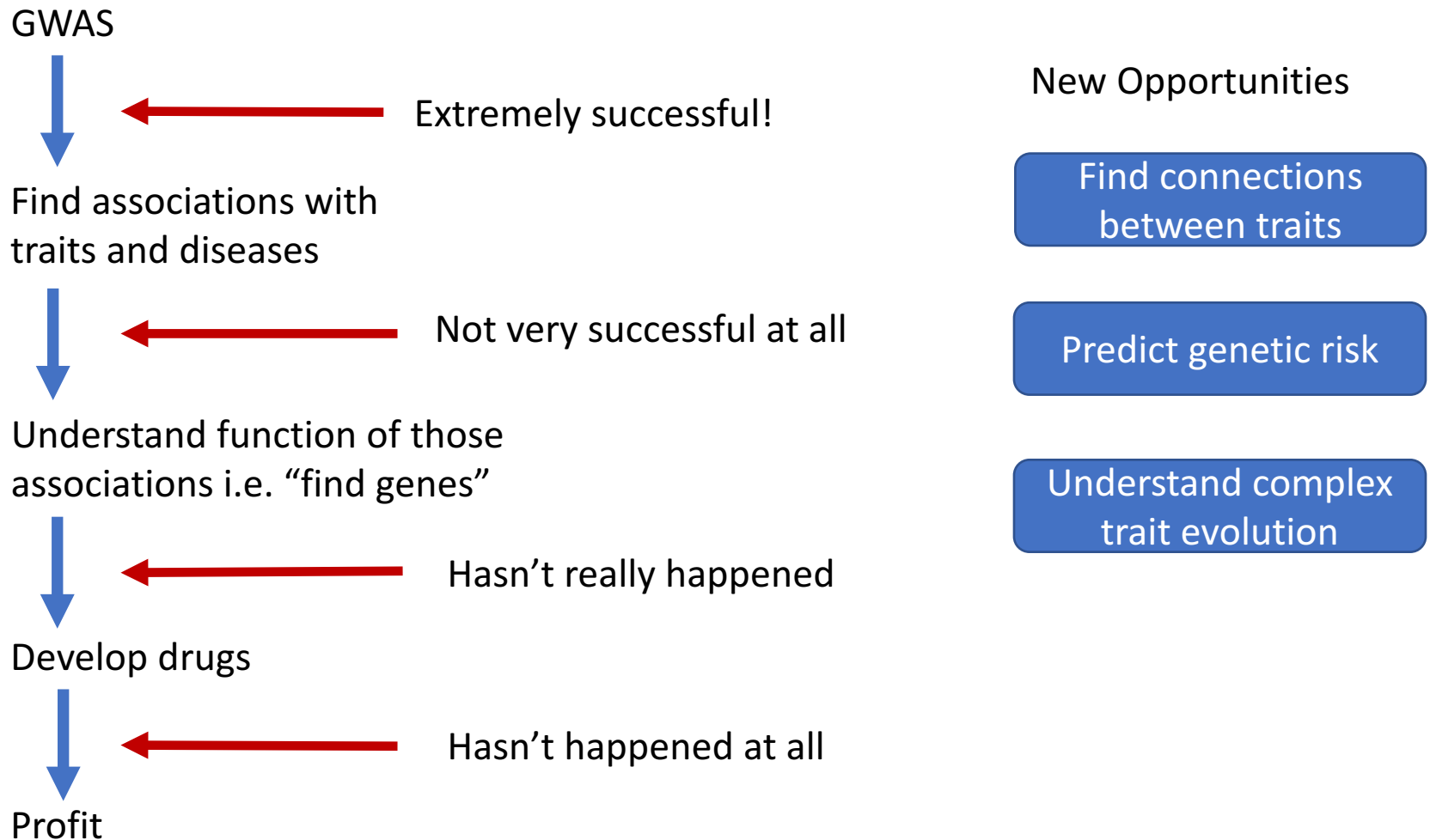
Thousands of studies

Tens of thousands of researchers

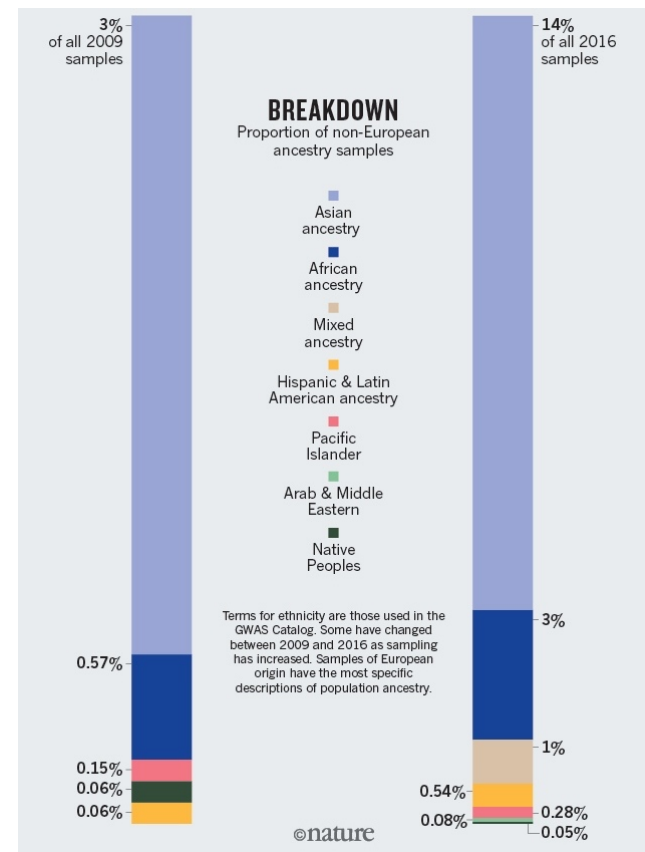
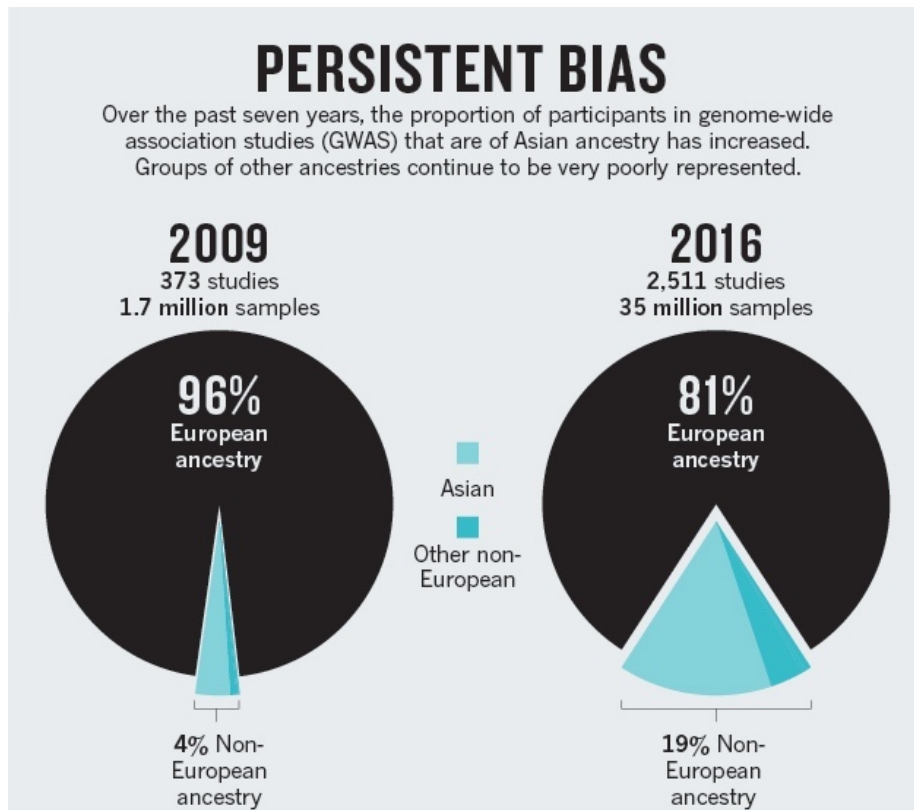
Tens of millions of patient-participants

Billions (?) of dollars

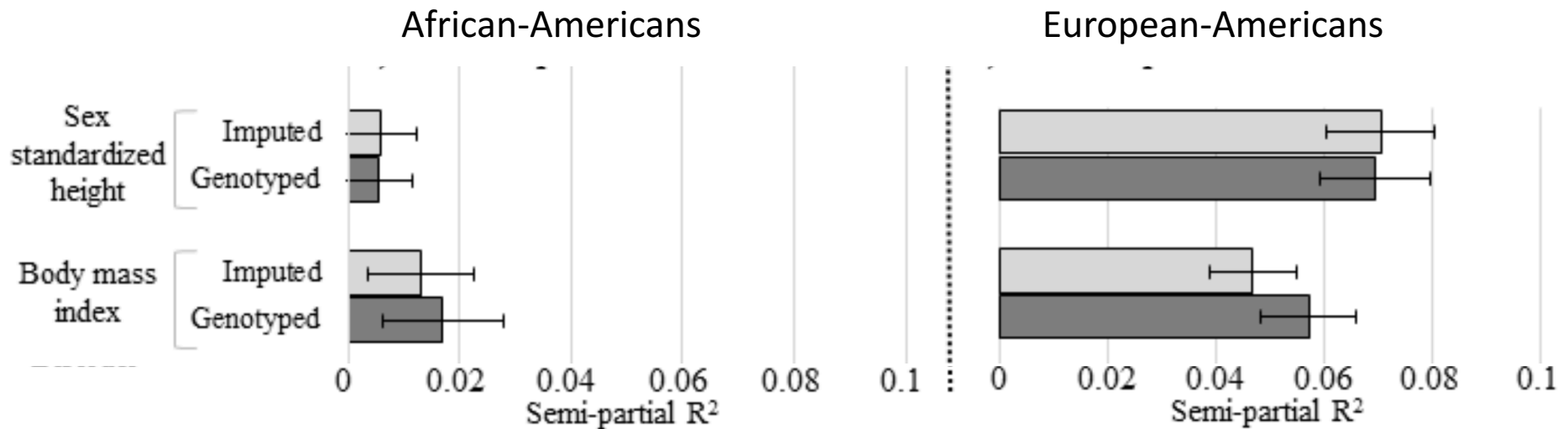
# How successful have GWAS been?



# Almost all GWAS are carried out in European-Ancestry populations



# European GWAS results do not translate to non-European ancestry populations



Ware et al 2018

# Summary

Genome-wide association studies:

Map common/low frequency variants associated with traits/disease

The bigger the sample size (more people) the smaller the effects you can detect

Do not tell us anything about function

Need to be extremely careful about population structure and multiple testing

