Genomic and Precision Medicine

Week 4: Methods for Dissecting the Genetic Basis of Complex Diseases

Jeanette McCarthy,
MPH, PhD

UCSF Medical Genetics

Robert Nussbaum, MD

University of California
San Francisco

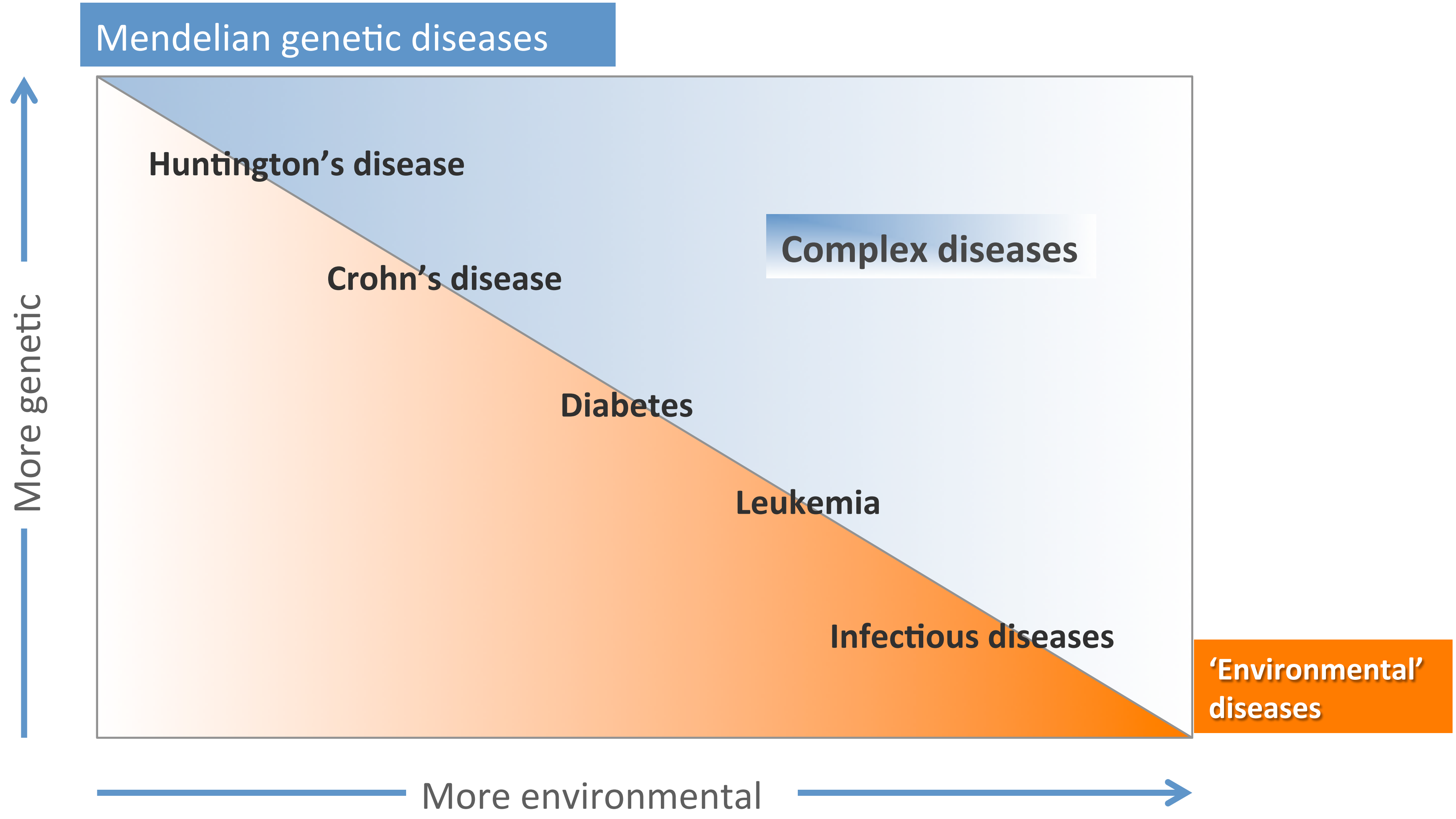
advancing health worldwide™

The Lecture

- **MODULE 1**: Background
  - Mendelian vs. complex diseases
  - How do we know a trait has a genetic component
- **MODULE 2**: GWAS methods
  - Genotyping
  - Study designs
  - Confounding and bias
- **MODULE 3**: GWAS analysis
  - Significance testing
- **MODULE 4**: GWAS interpretation
  - Measures of effect
  - External validity
- **MODULE 5**: What do we know about the genetics of common, complex diseases?

MODULE 1: Background

Spectrum of genetic disease



Mendelian genetic diseases

Huntington's disease

Crohn's disease

Complex diseases

Diabetes

Leukemia

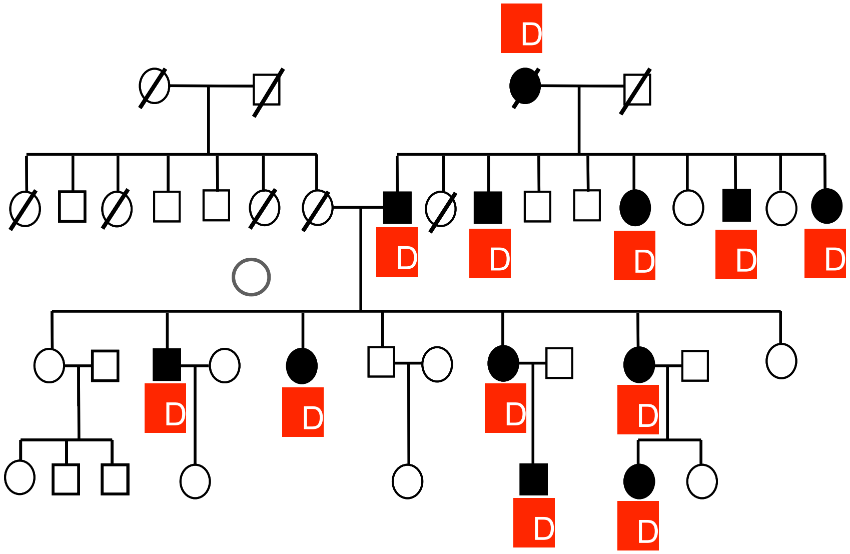
Infectious diseases

'Environmental' diseases

More genetic

More environmental

Mendelian vs. complex diseases

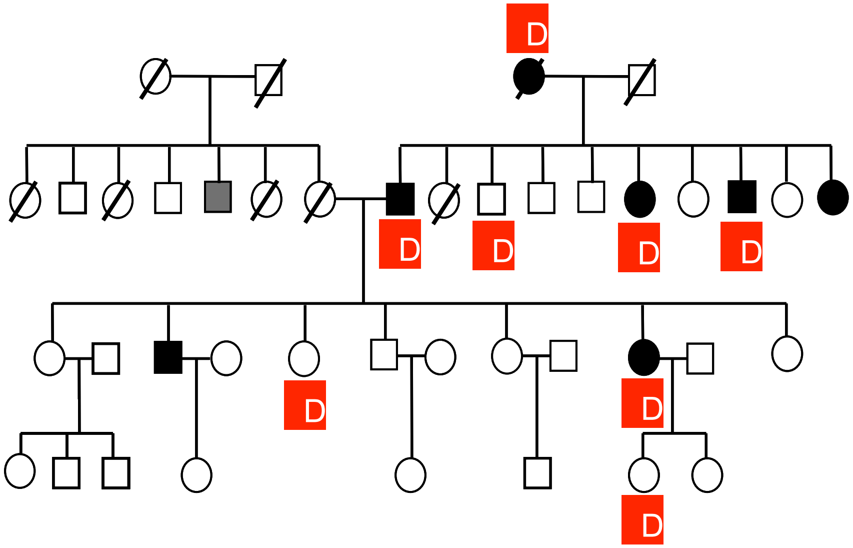


Mendelian

Clear inheritance pattern
(dominant, recessive, etc.)
High penetrance

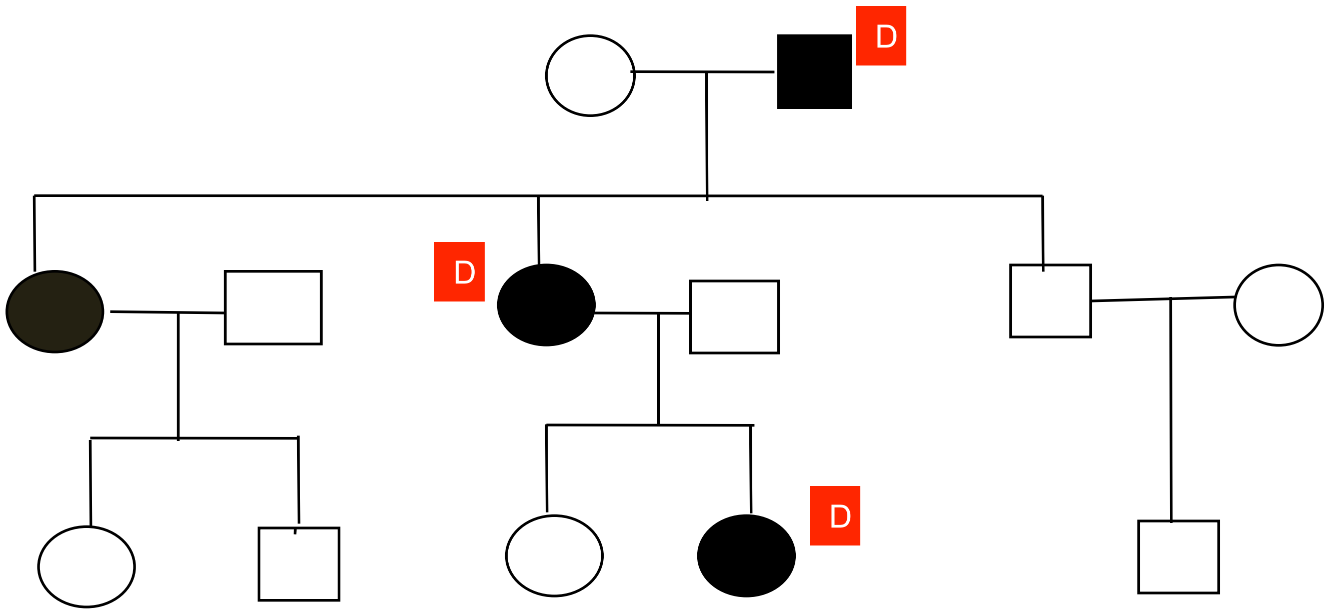
Complex

No clear inheritance pattern
Why????



Phenocopies

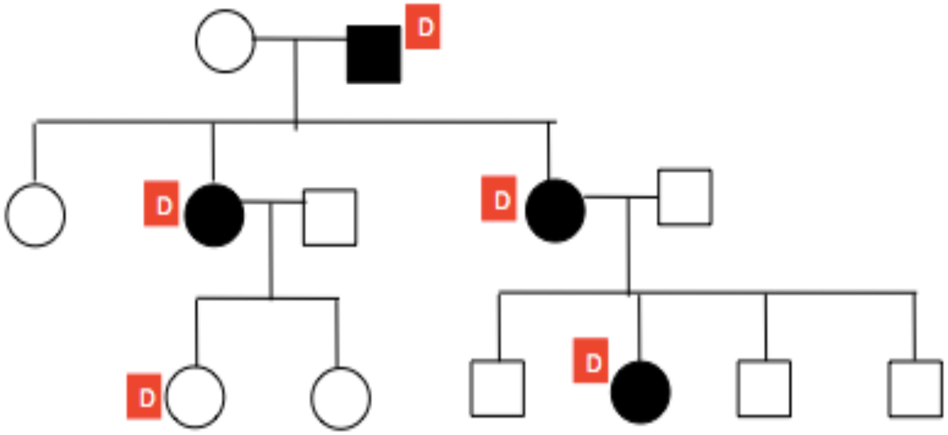
Non-genetic form of disease that is indistinguishable at the clinical level from genetic form of disease



Only 10% of breast cancer is thought to be genetic

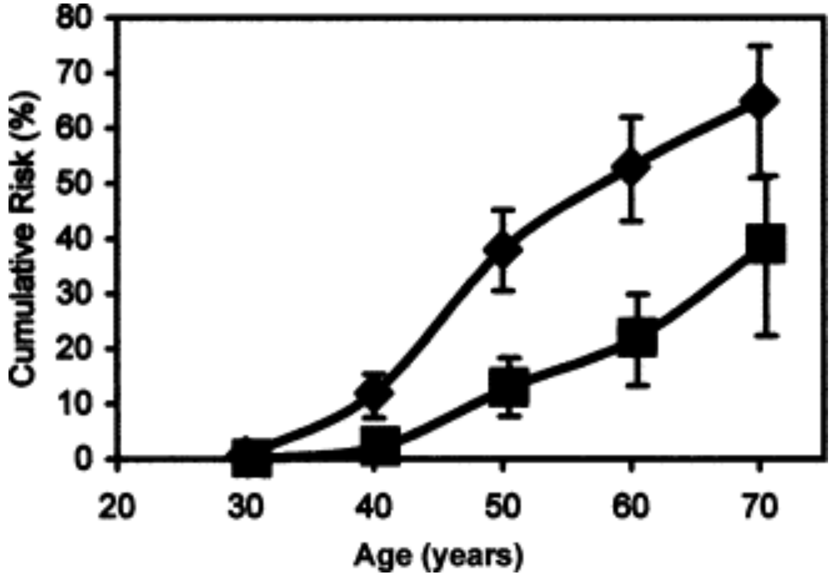
Incomplete penetrance

Not all genetically susceptible people develop disease



Penetrance =

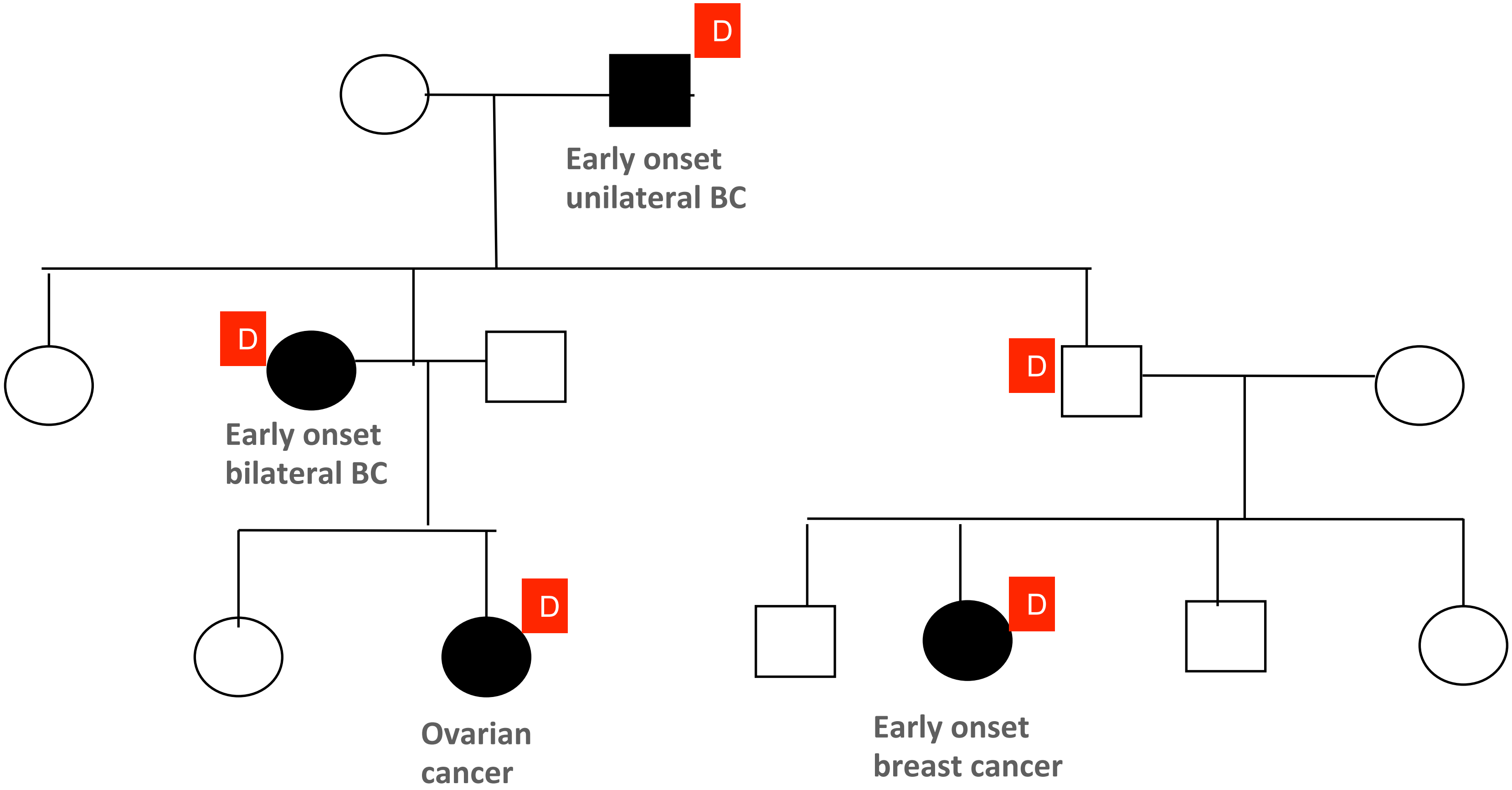
Probability of disease in mutation carriers



Cumulative risk of breast (♦) and ovarian (▪) cancer in BRCA1-mutation carriers.

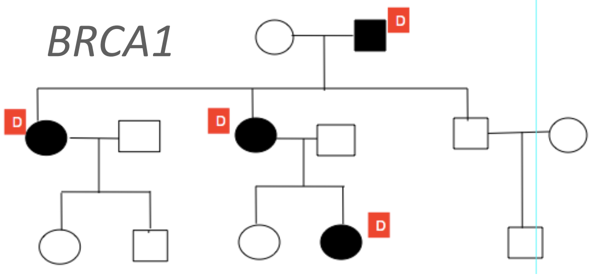
Variable expressivity

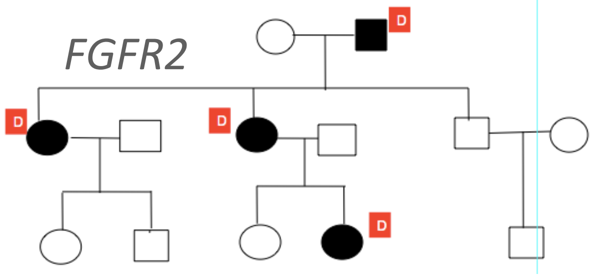
Same genetic factor causes multiple phenotypes

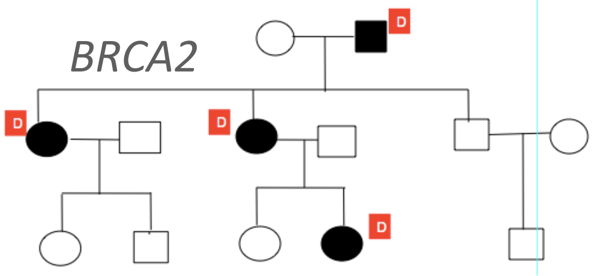


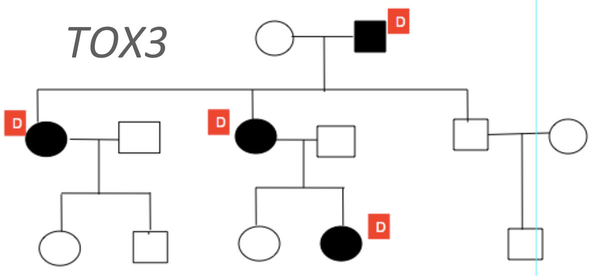
Genetic heterogeneity

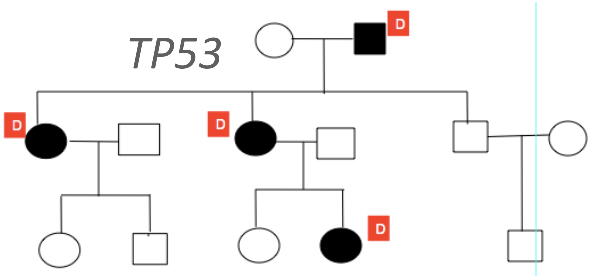
Mutations in different genes can lead to same disease

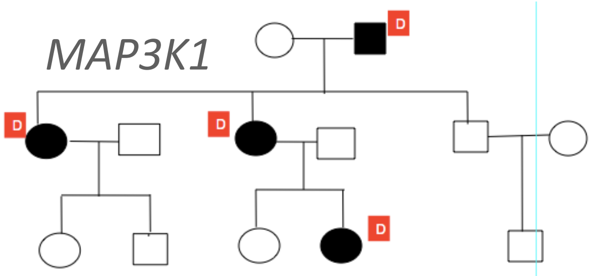












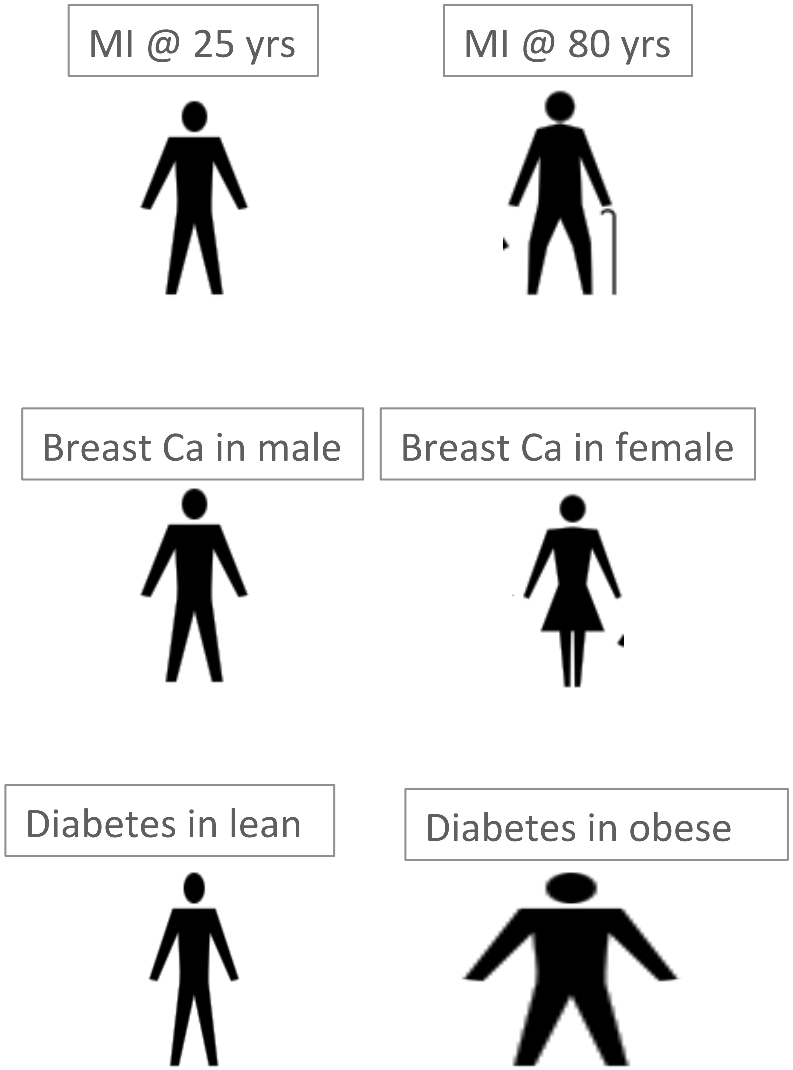
UCSF

Complex vs Mendelian traits

- Mendelian
  - Typically rare diseases (<1% prevalence) with single cause that is genetic
  - High penetrance
- Complex
  - Usually common diseases with multiple causes, both genetic and non-genetic
  - Low penetrance

Symptoms suggestive of a genetic condition

- Earlier age at onset of disease than expected
- Condition in the less often affected sex
- Family history with multiple generations affected
- Disease in the absence of known risk factors



MI @ 25 yrs

MI @ 80 yrs

Breast Ca in male

Breast Ca in female

Diabetes in lean

Diabetes in obese

Human genetic approaches for finding disease genes

	Candidate gene	Genome-wide
Populations	***Best for complex diseases***	
	√ Genetic association	√ GWAS
Families	***Best for Mendelian diseases***	
		√ Linkage analysis √ NGS for rare variants

Question

Which of the following is NOT a characteristic of complex diseases?

A. Genetic heterogeneity
B. Mendelian inheritance
C. Variable expressivity
D. Reduced penetrance

Answer

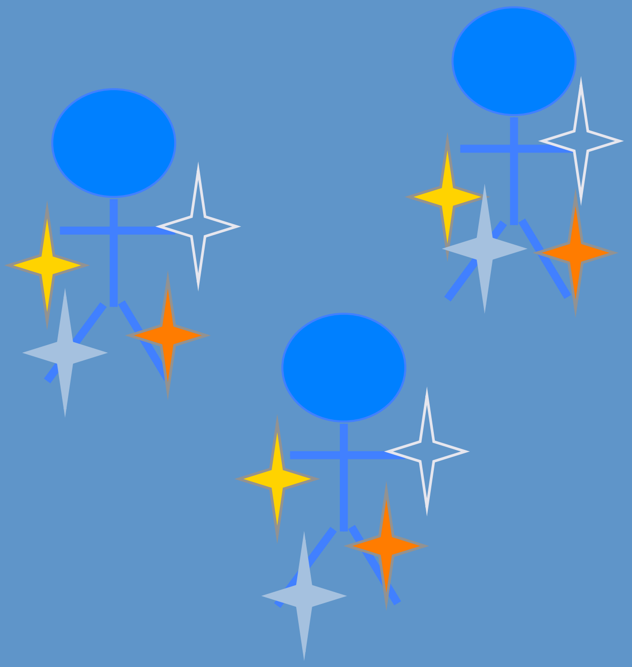
B. Mendelian inheritance

Complex traits are characterized by a departure from Mendelian patterns of inheritance

MODULE 2: Genome-wide association study methods

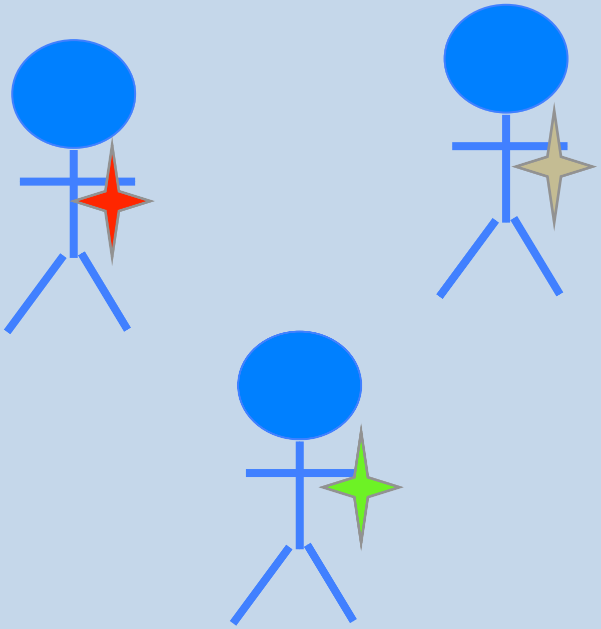
Theory behind GWAS strategy

Common disease – common variant



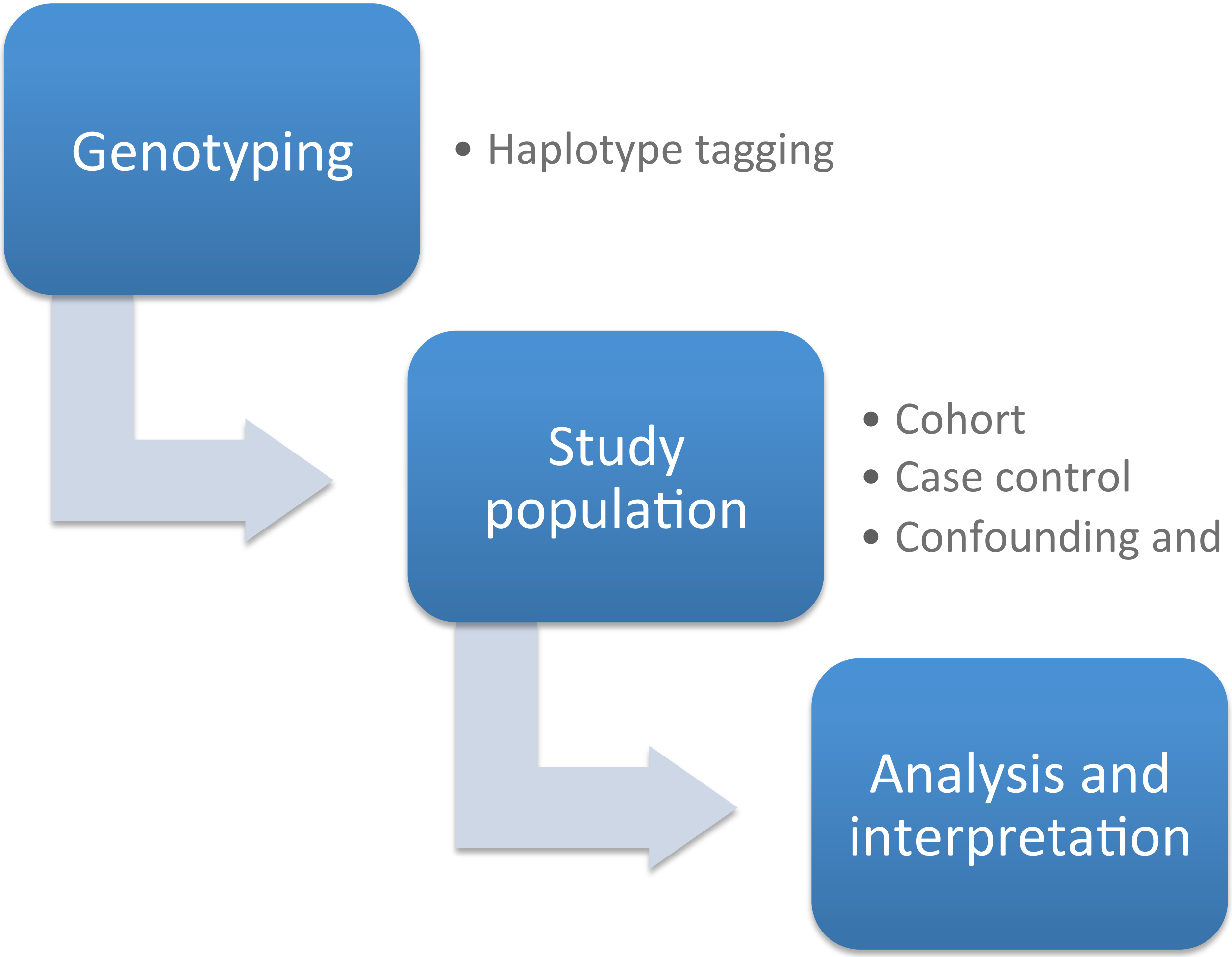
Cumulative effect of many common, low penetrance variants

Common disease – rare variant



Different single, rare, high penetrance variants

GWAS approach



Genotyping

- Haplotype tagging

Study population

- Cohort
- Case control
- Confounding and bias

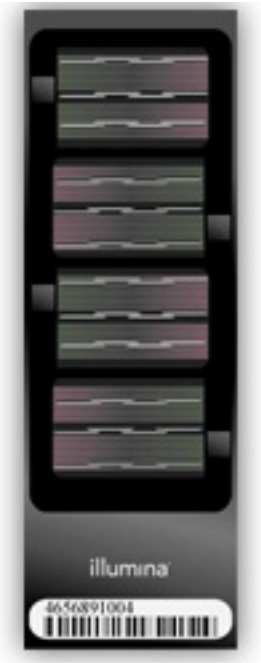
Analysis and interpretation

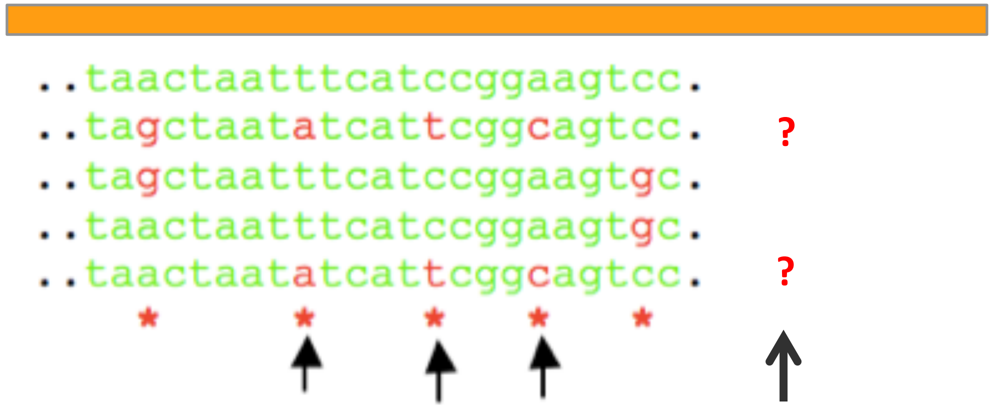
- Significance testing
- Measures of effect
- External validity

Genotyping Platform

Genotyping arrays/SNP chips

- 1,000,000 SNPs in one experiment
- Direct and indirect capture of ‘all’ common variants by using ‘tag’ SNPs





```
..taactaatttcatccggaagtcc.
..tagctaatatcattcggcagtcc.   ?
..tagctaatttcatccggaagtgc.
..taactaatttcatccggaagtgc.
..taactaatatcattcggcagtcc.   ?
```

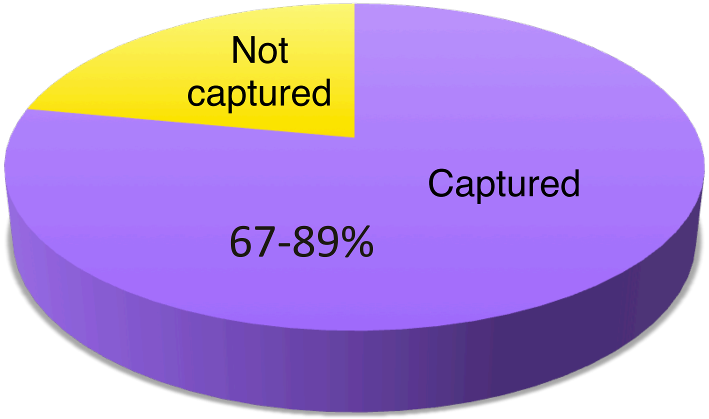
Genotyping any ONE of these four captures all

An association with a tag SNP helps define the region (block) harboring the causal variant

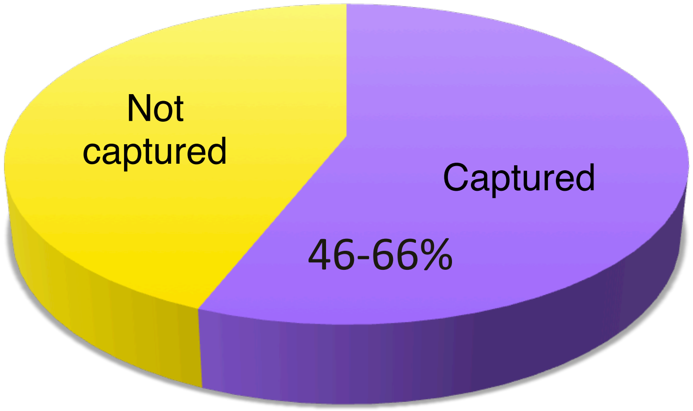
Genomic coverage of SNP chips

- How well do these chips capture common variants?

% of common variants captured by 1M SNP chip in Europeans/Asians



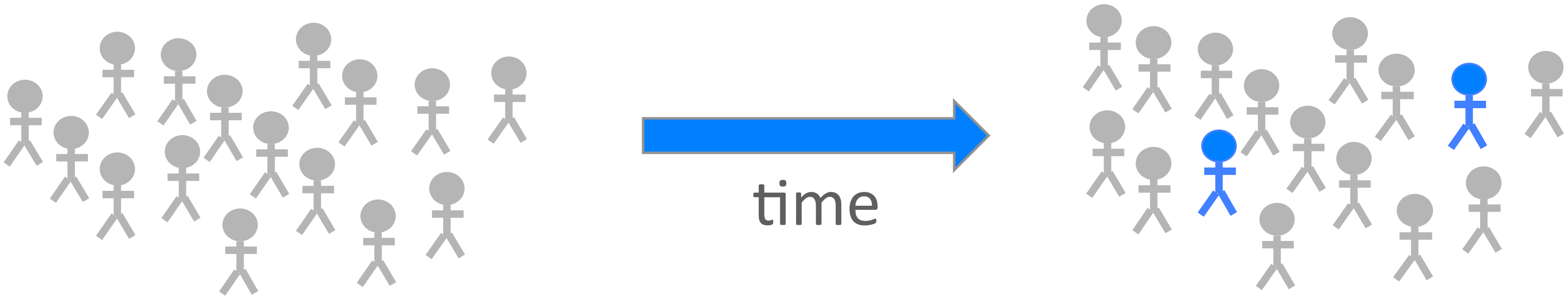
% of common variants captured by 1M SNP chip in African ancestries



Study Designs

- Common observational studies
  - Cohort
  - Case-control
- Common biases
  - Confounding
  - Misclassification bias

Cohort studies



time

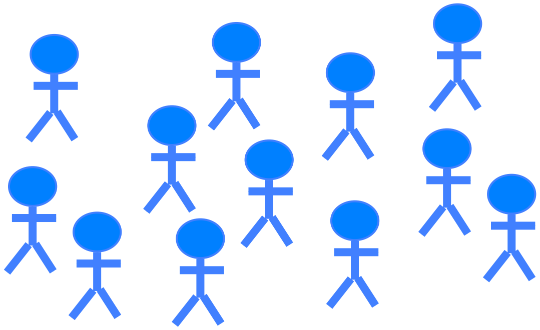
Disease-free

Who has the disease?
Who has the genetic variant?

Drawbacks

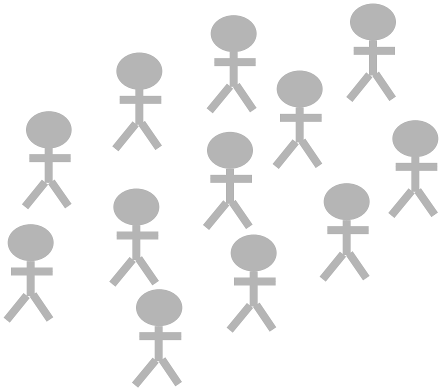
- Need to be large for rare diseases
- Need to follow a long time for diseases with long latency

Case-control studies



With disease

How many have gene variant?



Disease-free

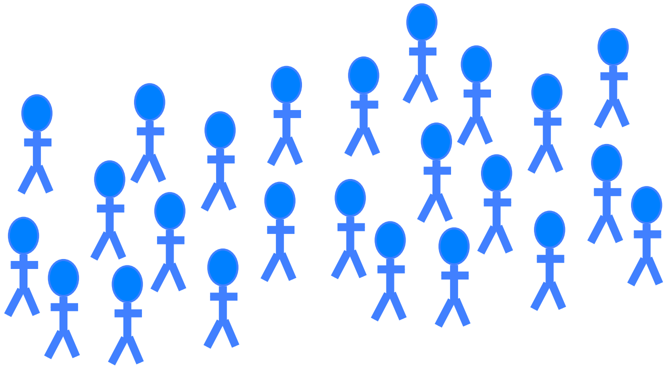
How many have gene variant?

Drawbacks

- Prone to confounding and other biases

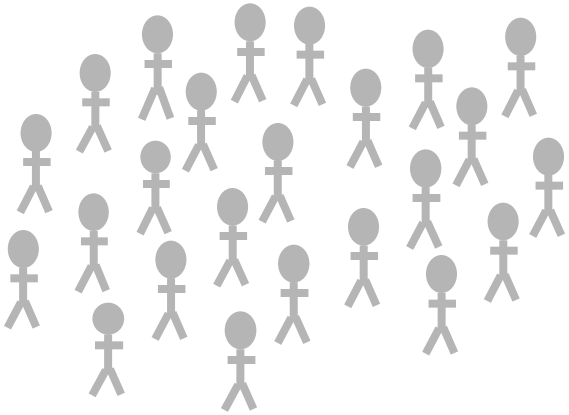
Confounding

Diabetic cases



✓ Genetic variant more prevalent

Non-diabetic controls

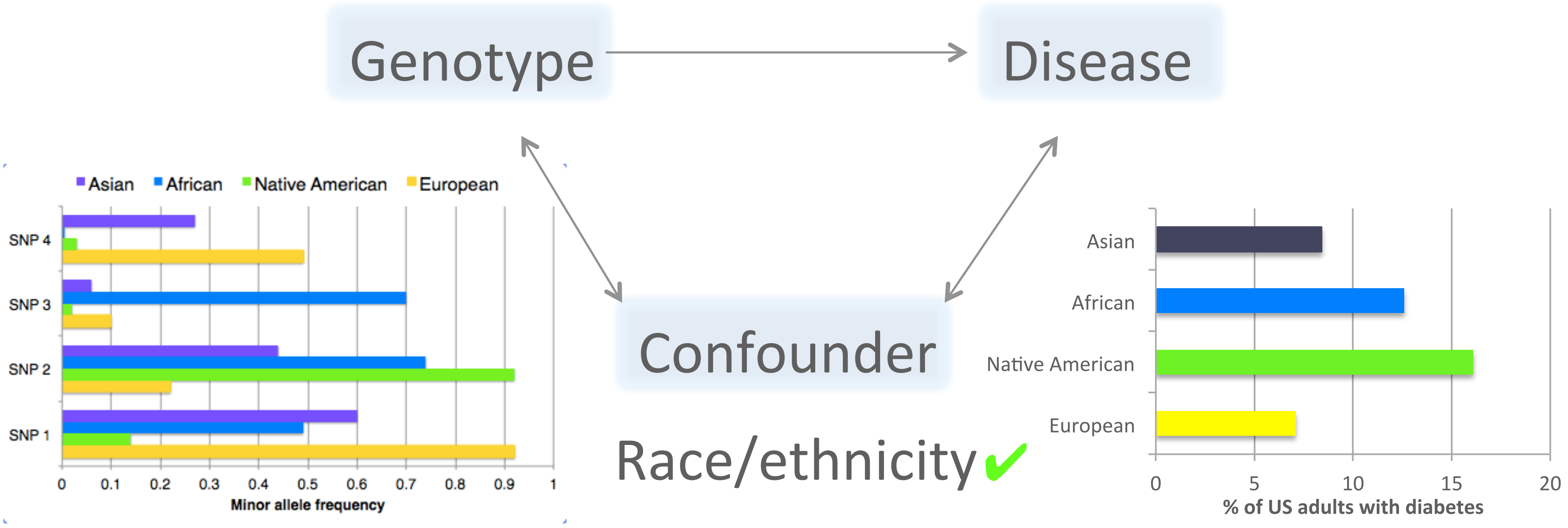


✓ Genetic variant less prevalent

How else might these two groups systematically differ?

Latino	Non-Latino
Smoker	Non-smoker
Obese	Non-obese

Race is a common confounder in GWAS (aka population stratification)


Genotype

Disease

Confounder

Race/ethnicity ✓

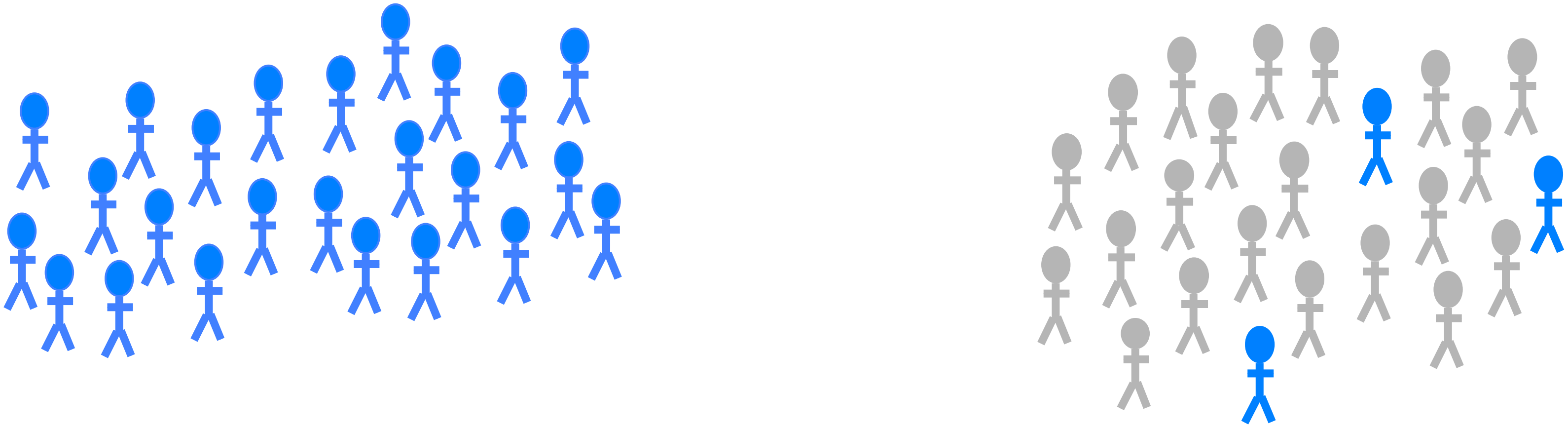
- *Can lead to false positive or false negative associations*
- *Must be controlled in design or analysis*

Misclassification bias

- Some cases erroneously classified as controls

Cases

Controls



How does this happen?

Misclassification bias (cont'd)

Genotypes assigned incorrectly

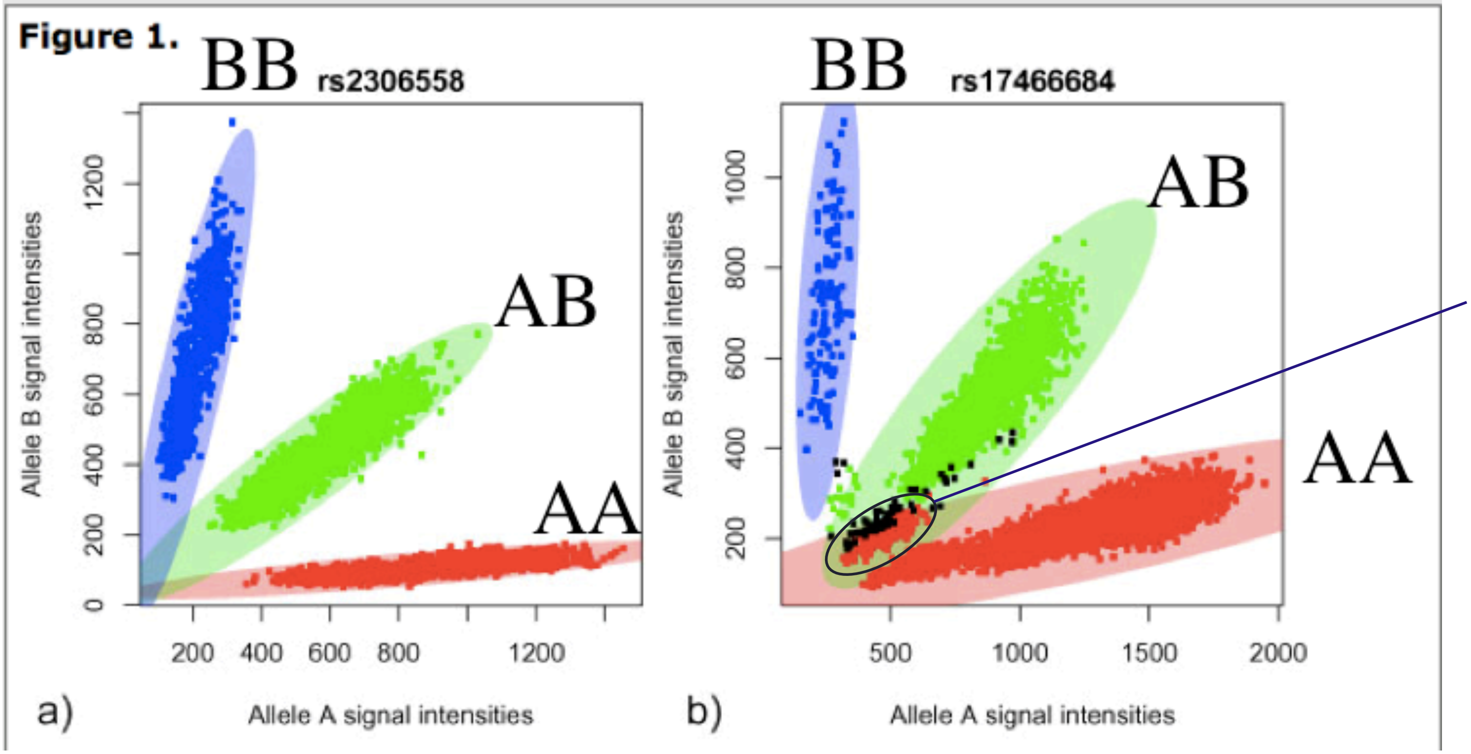


Figure 1.

True AB (green) erroneously called AA (red) or missing (black)

Shillert et. al. BMC Proceedings 2009 3(Suppl 7): S58

Effect of misclassification bias

Randomly distributed

- E.g. misclassification of disease irrespective of genotype
- E.g. genotyping error equally as likely in cases and controls
- False negative (bias toward the null)

Differentially distributed

- E.g. misclassification of disease in one genotype vs another
- E.g. genotyping error occurs in controls but not cases
- False negative or false positive

Question

Which study design is more prone to confounding and bias?

A. Case-control

B. Cohort

Answer

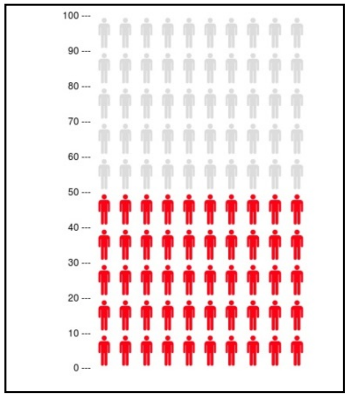
A. Case-control

Case-control studies are more prone to confounding and bias than cohort studies because cases and controls are often difficult to match on important variables

MODULE 3: Genome-wide association study — analysis

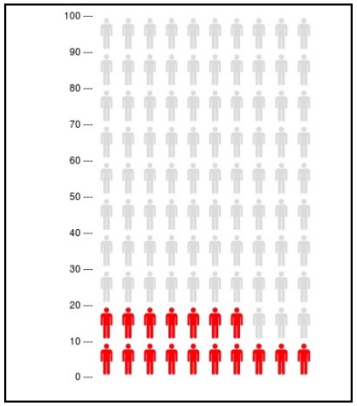
For a given SNP, how many people carry the variant allele?

With disease



50% carry variant

Without disease

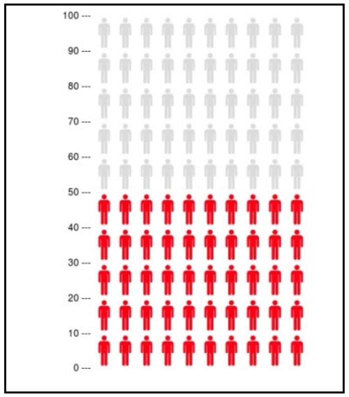


17% carry variant

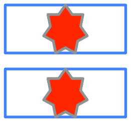
- Statistical test to compare the proportion of diseased and non-diseased individuals with the variant allele

Need to account for fact that humans have 2 copies of each gene

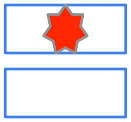
With disease



50% carry variant

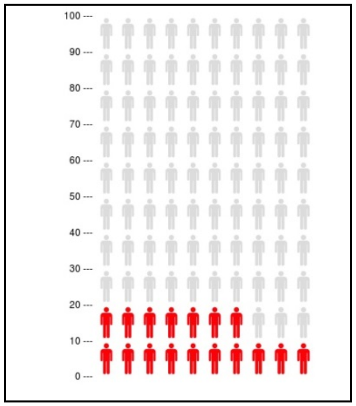


25% have 2 copies

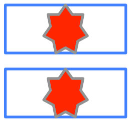


25% have 1 copy

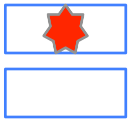
Without disease



17% carry variant



3% have 2 copies

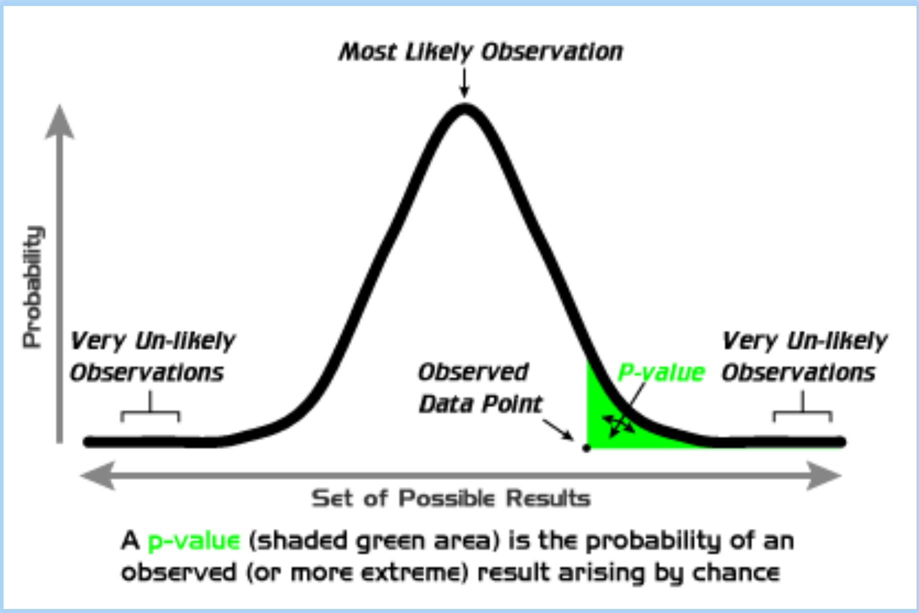


14% have 1 copy

A statistical test tells us how likely the results are true

- Compare proportion of diseased/non-diseased with zero, one or two copies of variant allele

	With disease	Without disease
2 copies	25 (25%)	26 (3%)
1 copy	25 (25%)	124 (14%)
0 copies	50 (50%)	750 (83%)



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

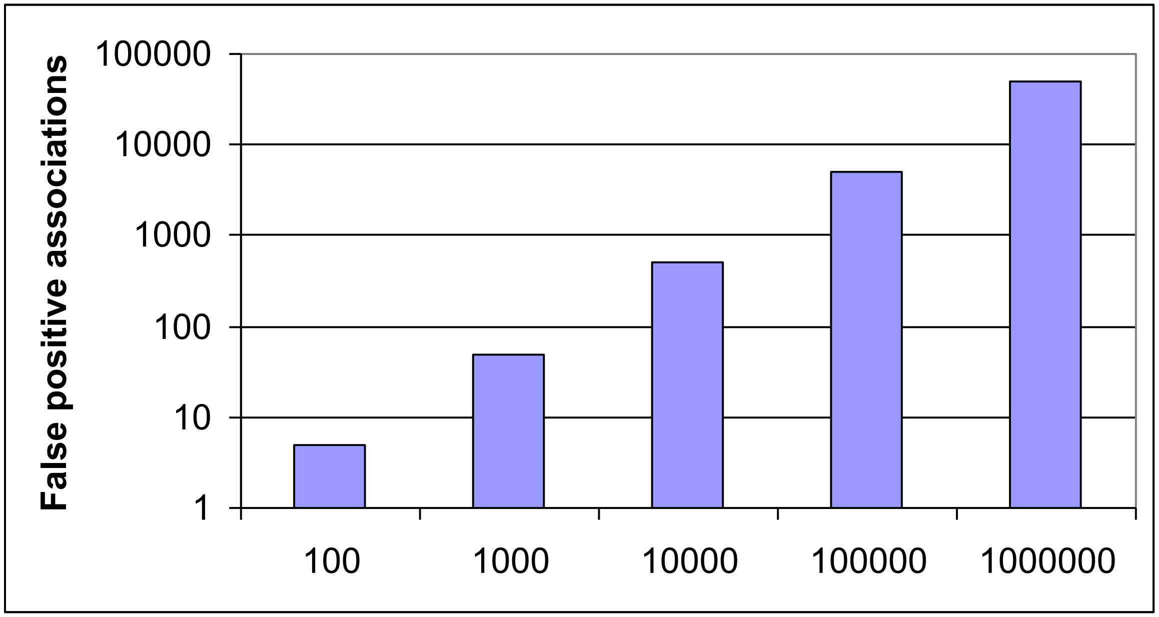
- Statistical test: Armitage trend test (1 d.f.)

Hypothesis testing and p values

- Statistical test tells you whether the difference in allele distribution between the two groups is likely to be due to chance or not

- What does a $p<0.05$ mean?
- <5% probability that the observation is due to chance (i.e. a false positive)
- This association is ‘statistically significant’

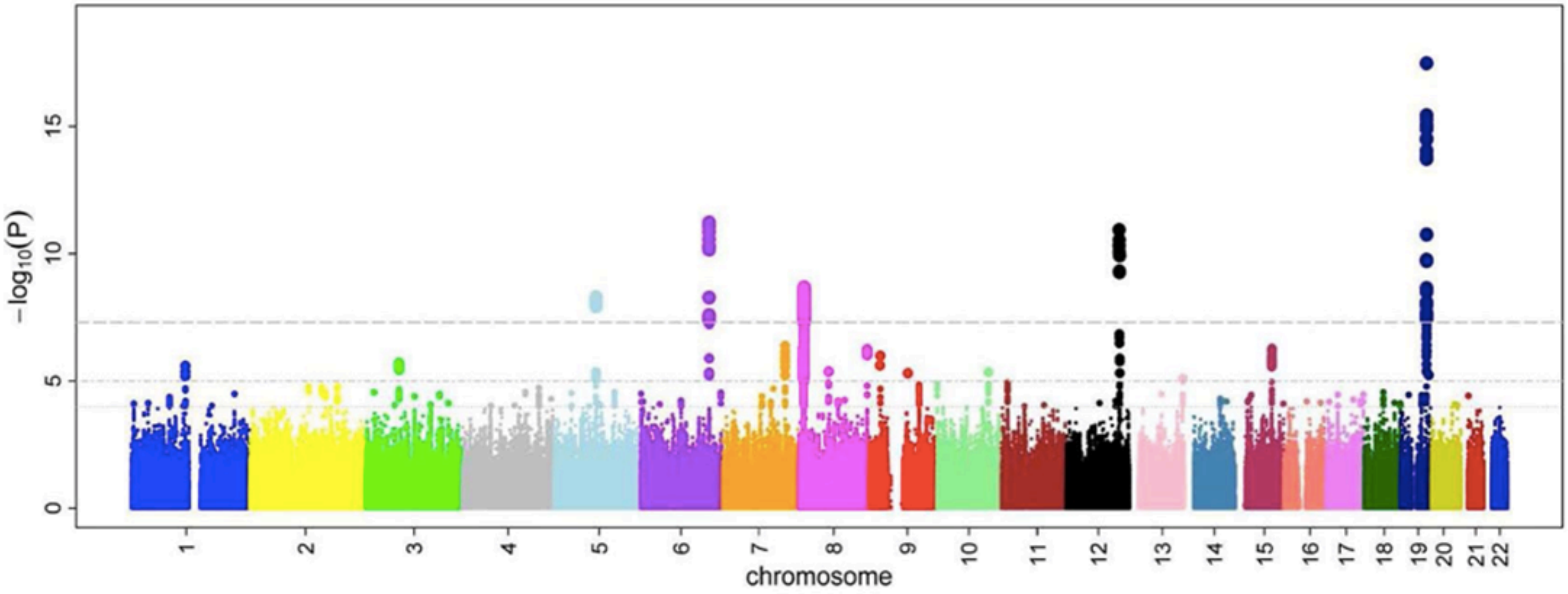
Correction for multiple testing



- For 1M tests, by chance alone we expect to see 50,000 'significant' associations at p<0.05

- p<.05 not stringent enough in this situation
- Genome-wide significance ≈p<0.00000005 (5x10^{-8})

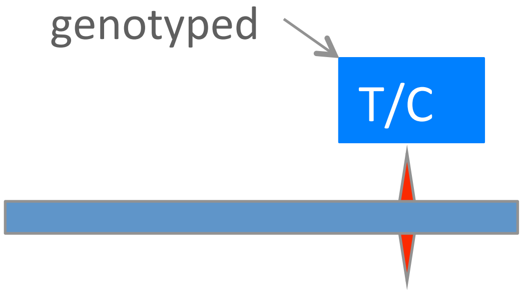
Manhattan plot showing genome-association with early microvascular disease

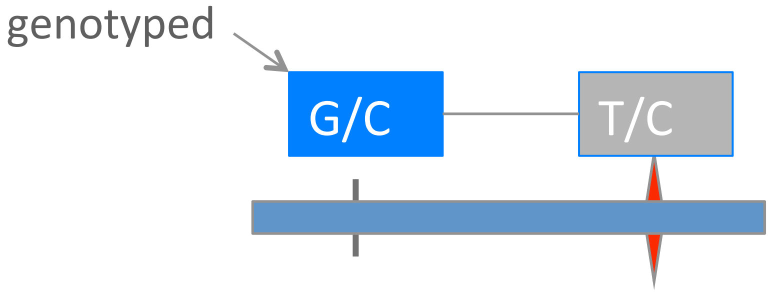


Ikram MK et al (2010), PLoS Genetics

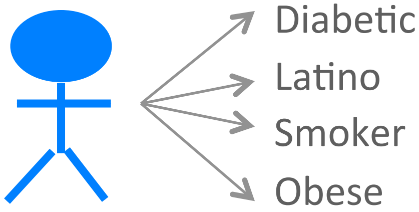
Reasons for association

- True association (true positive)
  - Causal variant (direct)
  - Linked to causal variant (indirect)

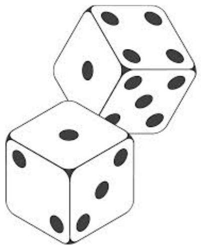




- False association (false positive)
  - Spurious (confounding/bias)
  - Chance



Association with wrong trait

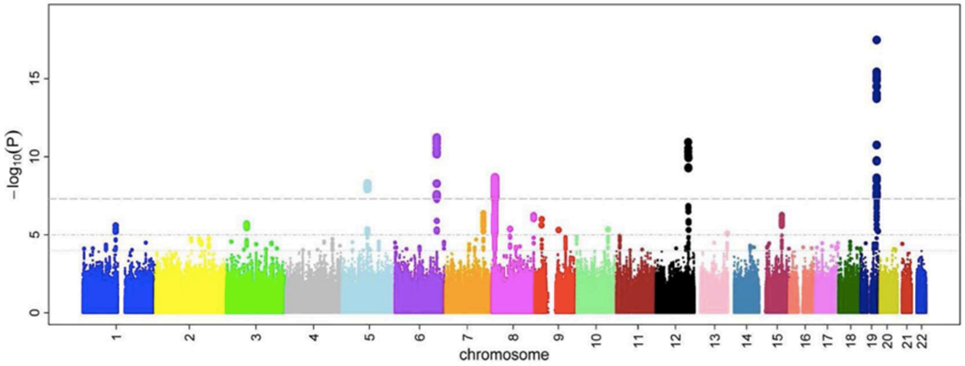


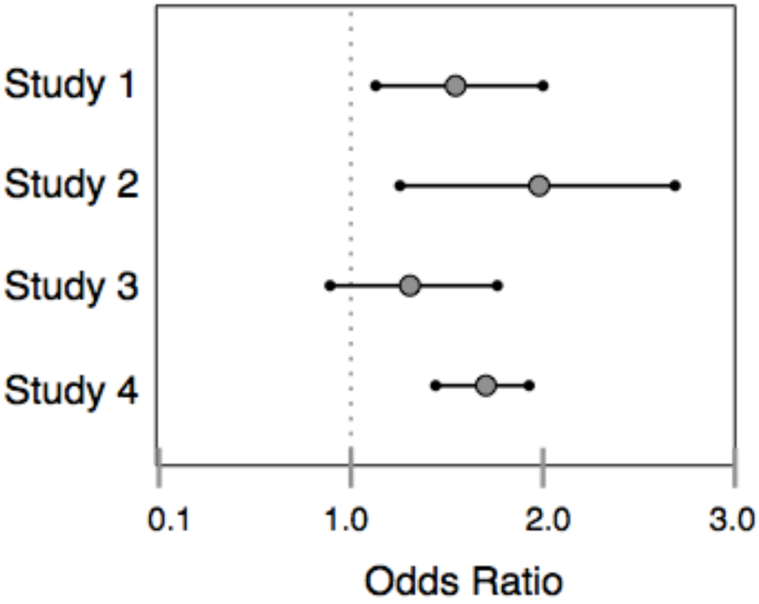
Properties of a valid association

✓ Not due to chance

✓ Free of bias

✓ Reproducible





Question

The role of P values in GWAS is to:

A. Guard against confounding and bias

B. Guard against chance associations

C. Both

Answer

B. Guard against chance associations.

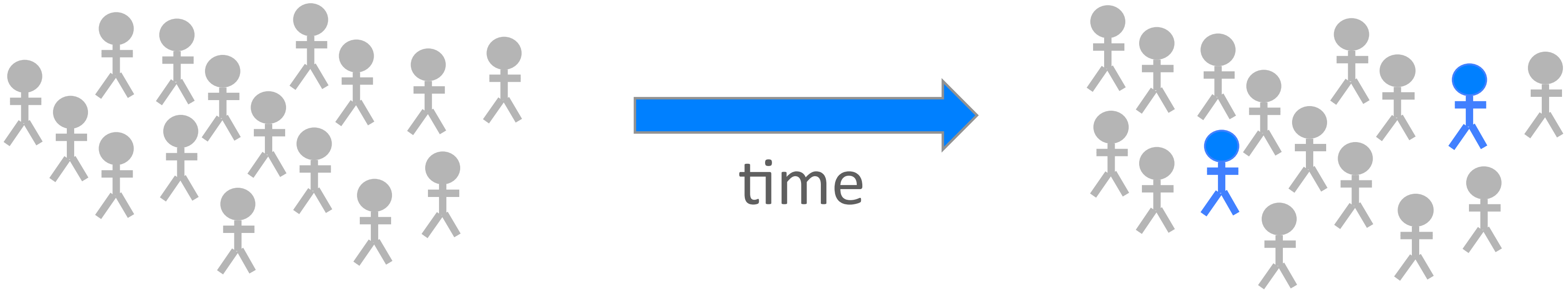
You can have a statistically significant result that is still confounded or otherwise biased.

Epidemiologically-sound study design is the best guard against bias and confounding.

MODULE 3: Genome-wide association study — interpretation

Calculation of risk

- Risk= incidence of disease
- Can be calculated from cohort studies



Disease-free
(n=1005)

How many get disease?
105/1005 = 0.10

Risk of disease is 10%

Calculation of risk for each genotype

	D+	D-	Total	Risk	Interpretation
All	105	900	1005	105/1005=0.10	10% risk of disease
TT	15	84	99	15/99= 0.15	15% risk of disease
TC	46	383	429	46/429= 0.11	11% risk of disease
CC	44	433	477	44/477= 0.09	9% risk of disease

Each is an absolute risk, conveying the likelihood of developing disease if you have a specific genotype

Calculation of a relative risk

Relative risk = ratio of two risks

Measures the 'effect' of the variant on risk of disease

	Absolute risk	Relative Risk
TT	0.15	0.15/0.09 = 1.7
TC	0.11	0.11/0.09 = 1.2
CC	0.09	1.0 (reference)

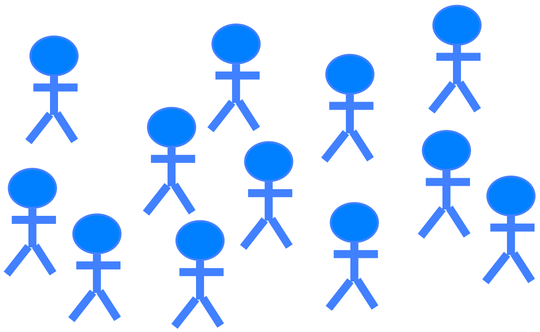
1.7-fold increased risk of disease

70% increased risk of disease

1.2-fold increased risk of disease

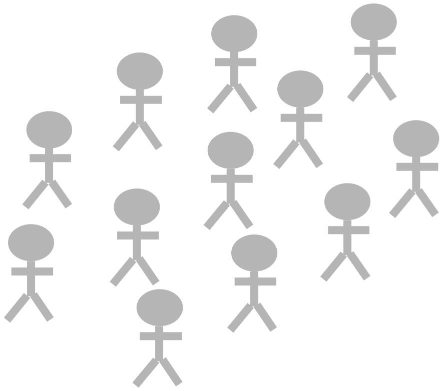
20% increased risk of disease

Can we calculate risk (incidence) of disease from a case-control study???



Cases (with disease)

(n=500)



Controls (disease-free)

(n=500)

NO!

# of cases in study is pre-selected
500/1000 ≠ disease incidence

We CAN calculate the ODDS of disease

Odds = disease (cases) : no disease (controls)

	Cases	Controls	Total	Odds of disease
All	500	500	1000	500/500=1.0
TT	160	108	268	160/108=1.5
TC	160	121	281	160/121=1.3
CC	180	271	451	180/271=0.7

Odds of 1.0 = 50:50 chance of disease

Odds >1 = chance of disease greater than no disease

Odds <1 = chance of disease less than no disease

Calculation of an odds ratio

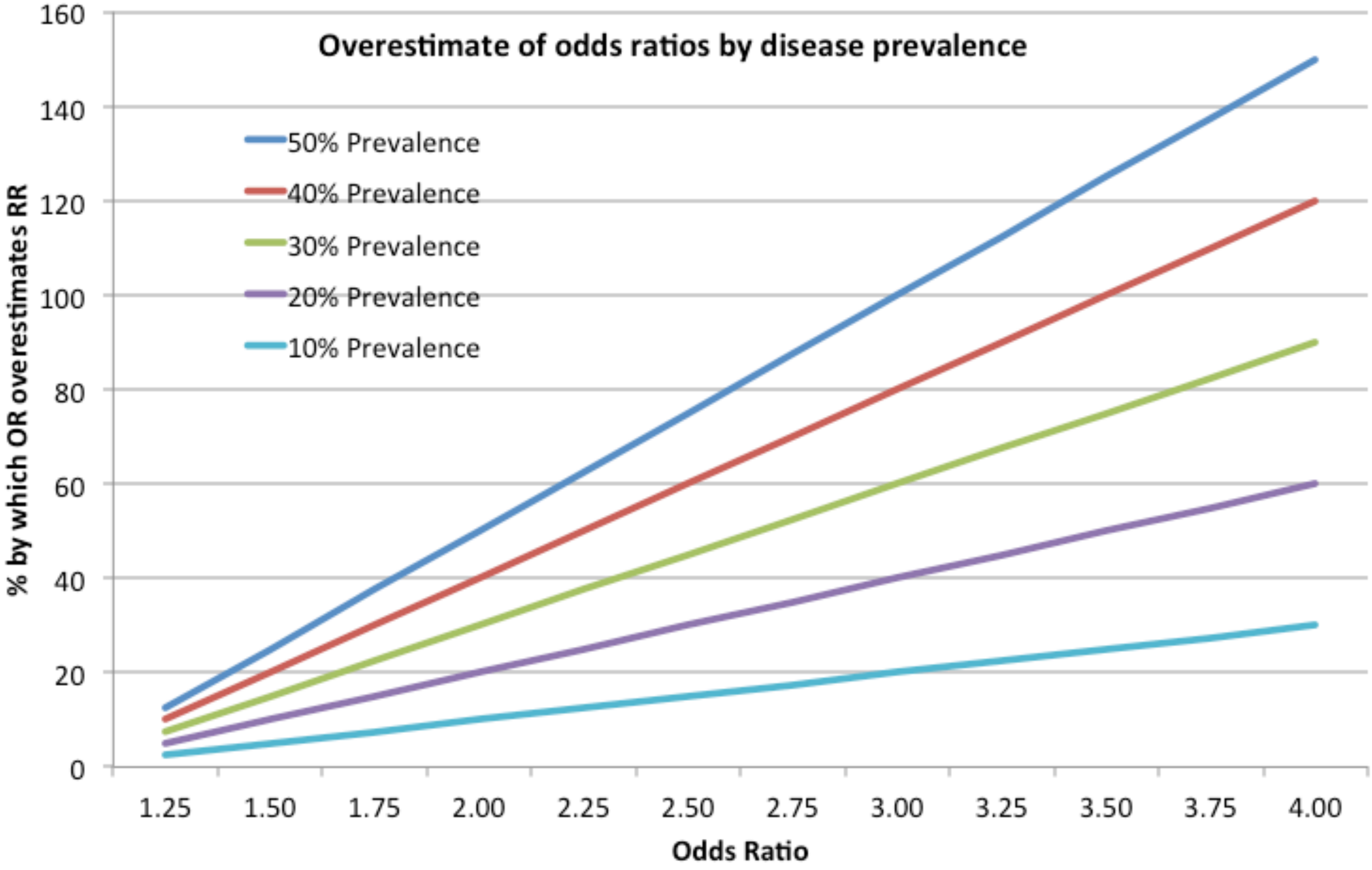
Odds ratio = ratio of two odds

	Odds	Odds ratio (OR)
TT	1.5	1.5/0.7 = 2.1
TC	1.3	1.3/0.7 = 1.9
CC	0.7	1.0 (ref.)

2.1-fold increased odds of disease

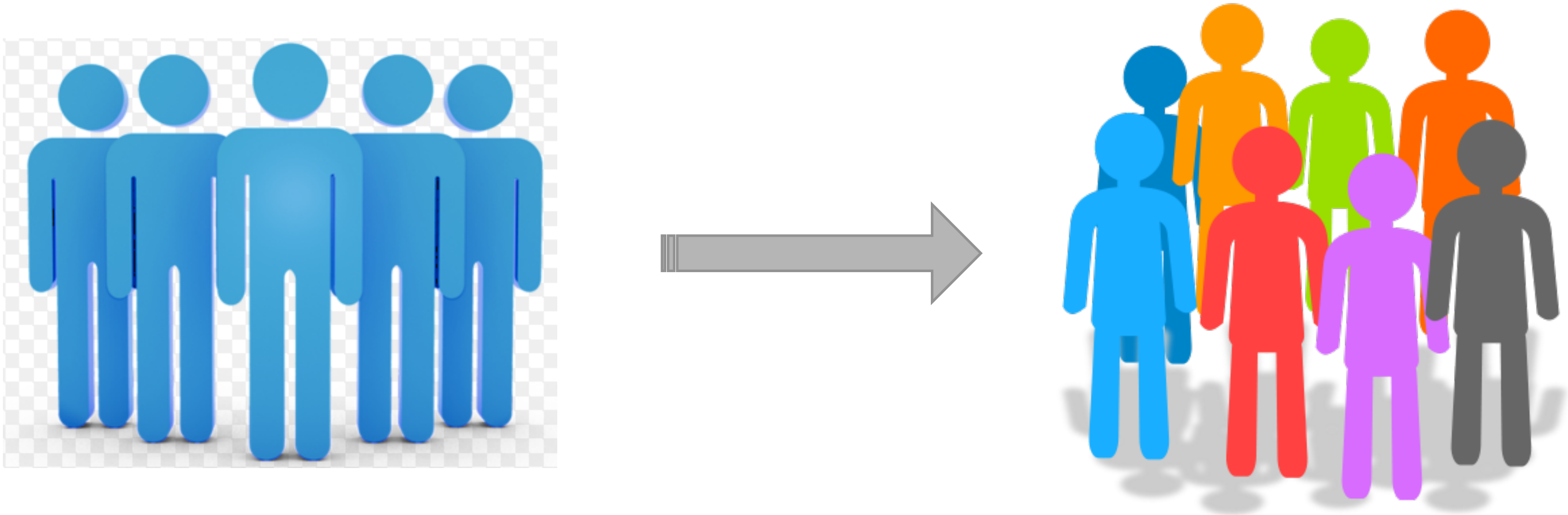
1.9-fold increased odds of disease

Odds ratios overestimate relative risks



Generalizability (external validity)

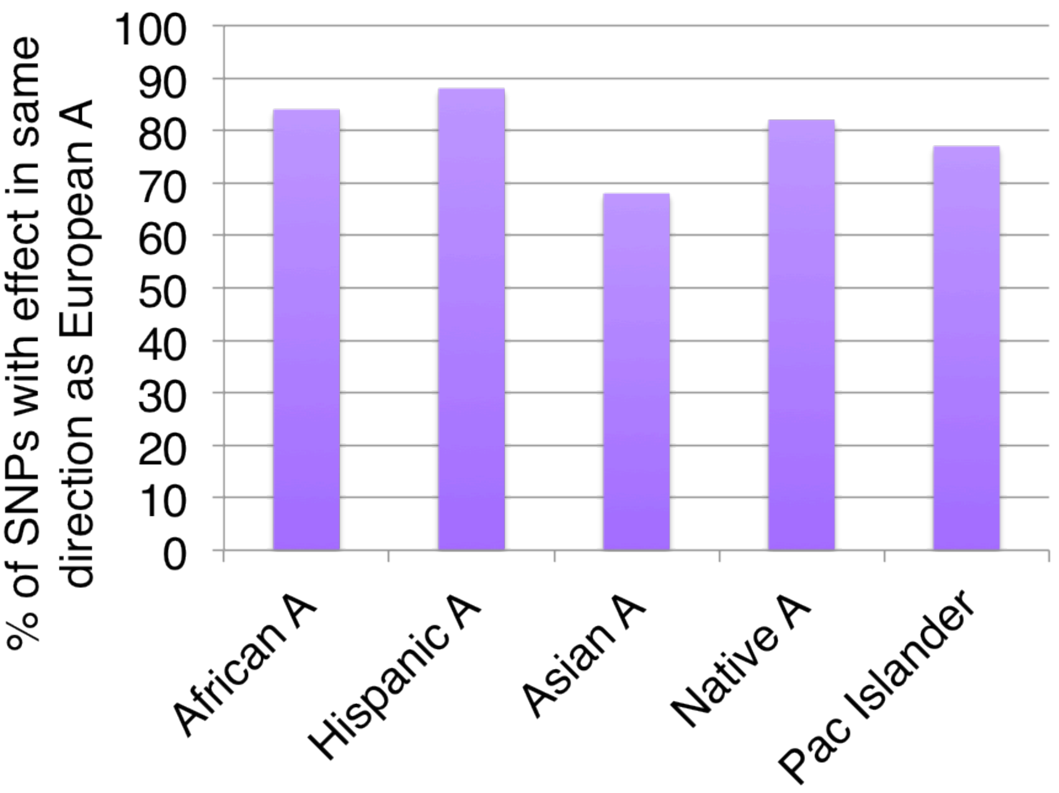
- How well does the study population represent the general population to which the results are being applied?



Study population

General population

Generalizability of GWAS results across race



Most GWAS done in Europeans

Most associations generalize from European to non-European populations, but effect sizes usually differ, especially for African Americans.

PLOS BIOLOGY TENTH ANNIVERSARY

UCSF

Carlson CS, et al. (2013) PLoS Biol 11(9): e1001661. doi:10.1371/journal.pbio.1001661

Question

A relative risk can be measured directly from which study design(s)?

A. Case – control

B. Cohort

C. Both

Answer

B. Relative risks can be calculated directly from cohort studies, not case control studies.

Odds ratios can be calculated from case-control studies.

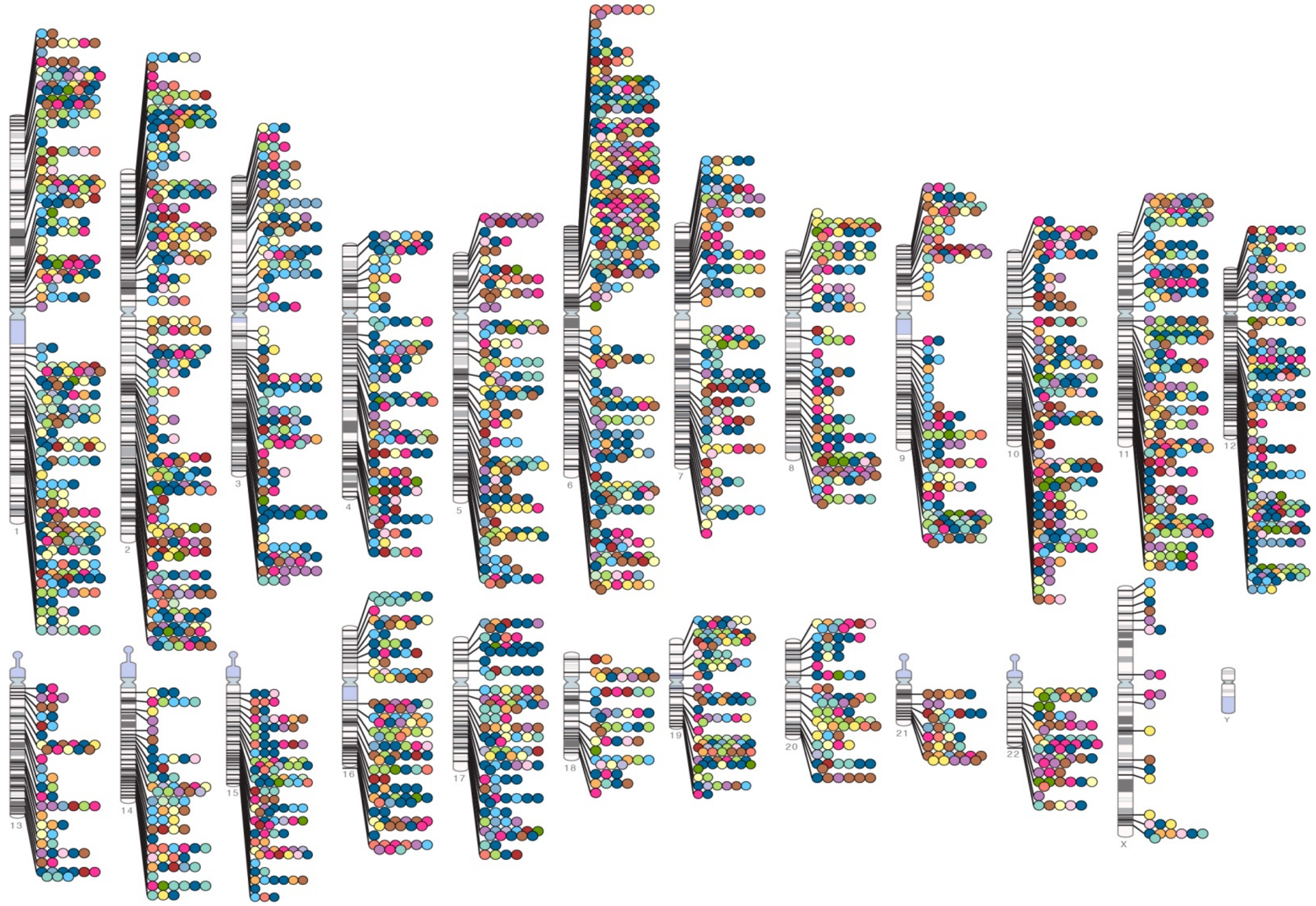
Odds ratios and relative risks are not the same thing, especially for common diseases where ORs overestimate RRs.

MODULE 5: What do we know about the genetics of common, complex diseases?

Published Genome-Wide Associations through 12/2012

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

National Human Genome Research Institute

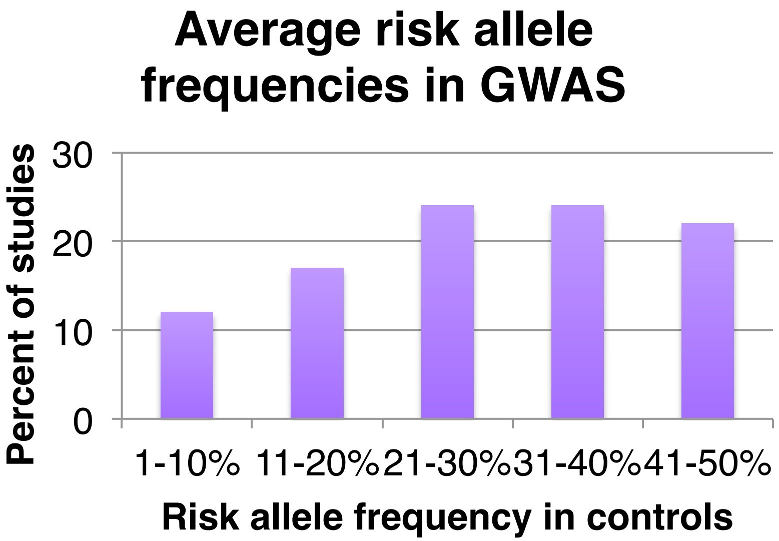


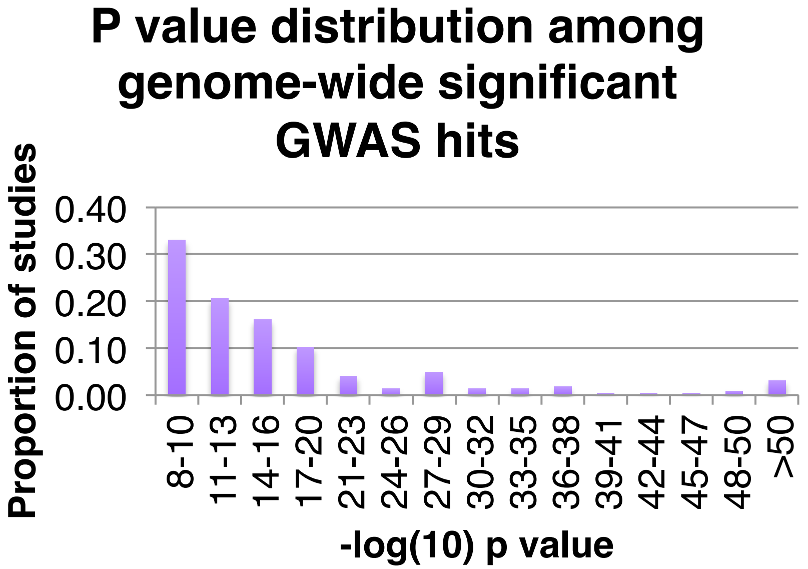
- ● Digestive system disease
- ● Cardiovascular disease
- ● Metabolic disease
- ● Immune system disease
- ● Nervous system disease
- ● Liver enzyme measurement
- ● Lipid or lipoprotein measurement
- ● Inflammatory marker measurement
- ● Hematological measurement
- ● Body measurement
- ● Cardiovascular measurement
- ● Other measurement
- ● Response to drug
- ● Biological process
- ● Cancer
- ● Other disease
- ● Other trait

As of 03/13/14, the catalog includes 1836 publications and 12756 SNPs.

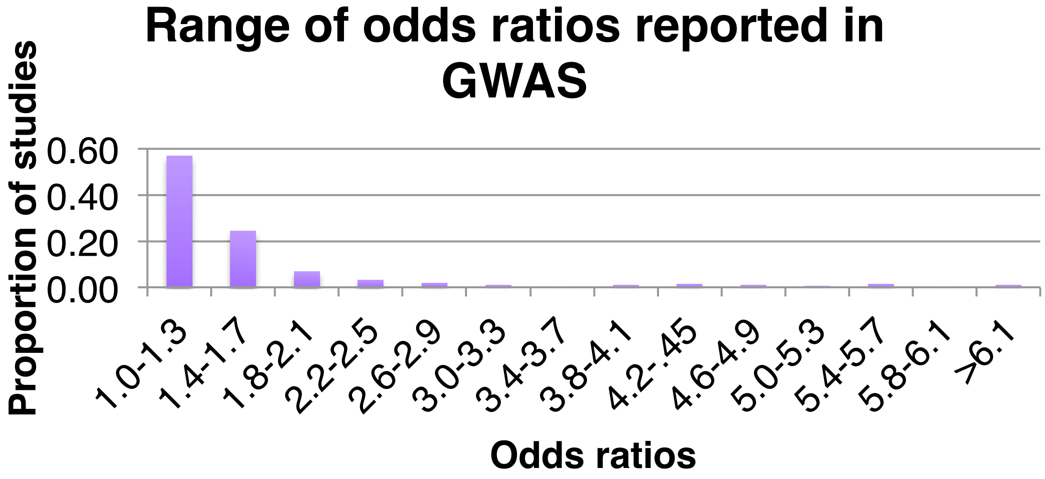
EMBL-EBI

What are we finding?





- Highly significant associations
- Common SNPs with weak effects… i.e. small increased risk, not diagnostic



Many SNPs for each disease/trait

Disease/trait	# GWAS loci	% heritability explained
Type 1 diabetes	41	~60%
Fetal hemoglobin	3	~50%
Macular degeneration	3	~50%
Type 2 diabetes	39	20-25%
Crohn’s disease	71	20-25%
LDL/HDL levels	95	20-25%
Height	180	~12%

GWAS SNPs explain only a fraction of the heritability

From Lander, ES. Nature 2011, 470: 187-197

Limitations of GWAS…what we're missing

- Common SNPs not tagged well
- Rare variants
- Other types of variants (CNV, etc)
- Epistatic effects (gene-gene interaction)
- Effects of gene*environment interaction

THE END