Complex traits: what to believe?

Joel N. Hirschhorn, MD, PhD
Children's Hospital/Harvard Medical School
Whitehead/MIT Center for Genome Research

SNPs, patterns of variation, and complex traits

- Introduction
- Common genetic variation and disease
- Methods for finding variants for complex traits
- Interpreting genetic studies
  - Association
  - Linkage
  - Resequencing

What could we learn?

SNPs, patterns of variation, and complex traits

- Introduction
- Common genetic variation and disease
- Methods for finding variants for complex traits
- Interpreting genetic studies
  - Association
  - Linkage
  - Resequencing
- What could we learn?

Many common diseases have genetic components...

Diseases

Bipolar disorder

Stroke

Heart attack

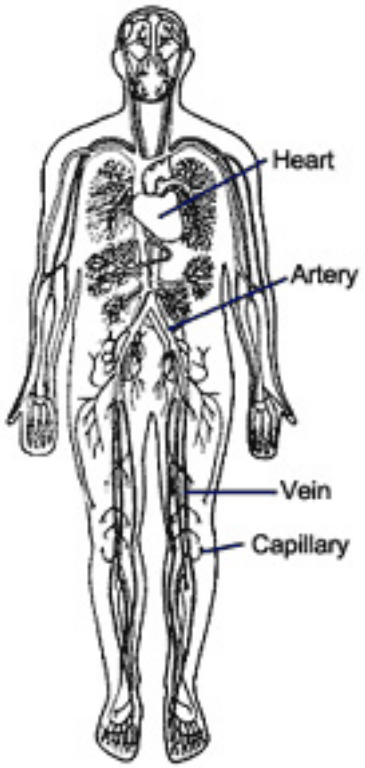
Breast cancer

Diabetes

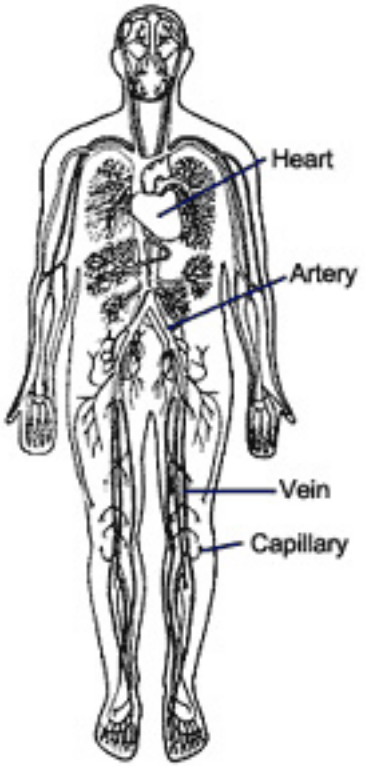
Inflammatory bowel disease

Prostate cancer

Arthritis



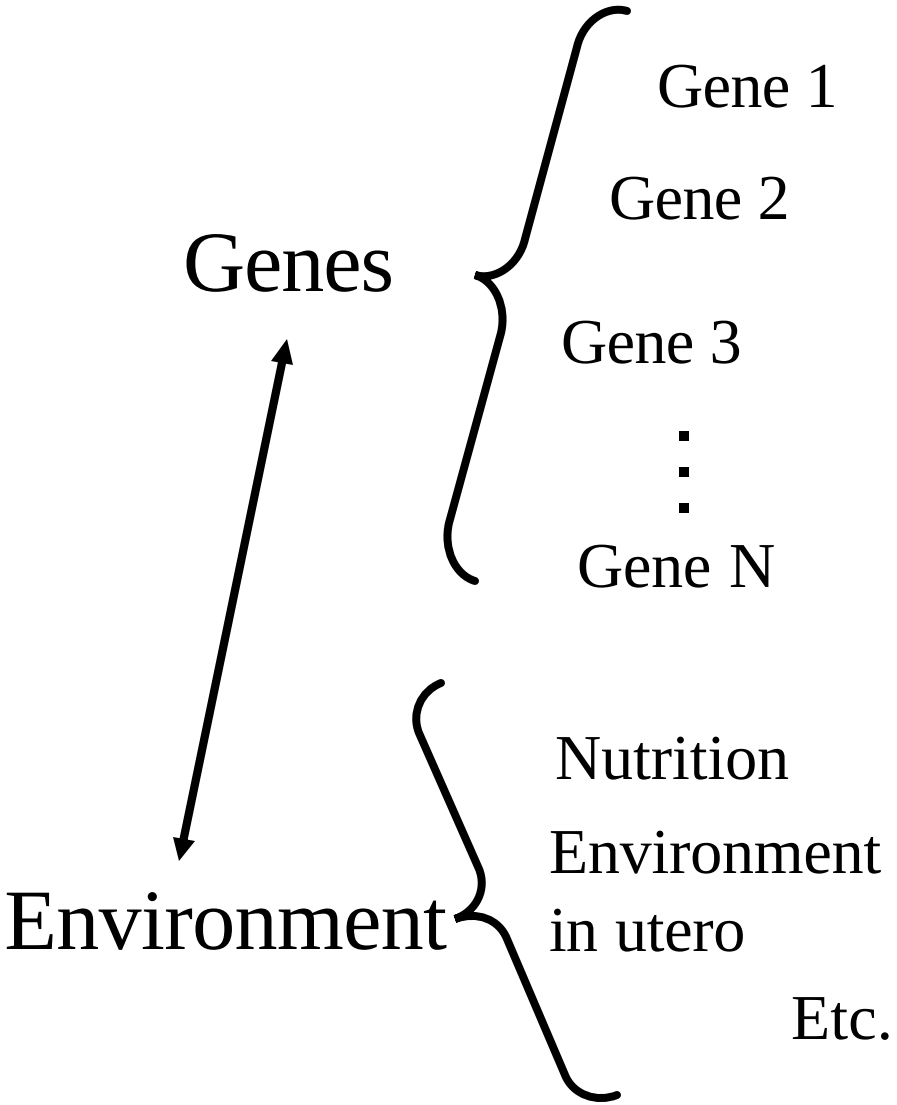
…as do many quantitative traits...



Quantitative Traits

- Height
- Blood pressure
- Insulin secretion
- Weight
- Waist-hip ratio
- Timing of Puberty
- Bone density

…but the genetic architecture is usually complex



Goal: Connect genotypic variation with phenotypic variation

Inherited DNA sequence variation **?** → Variation in phenotypes

Associating inherited (DNA) variation with biological variation

- Each person's genome is slightly different
- Some differences alter biological function

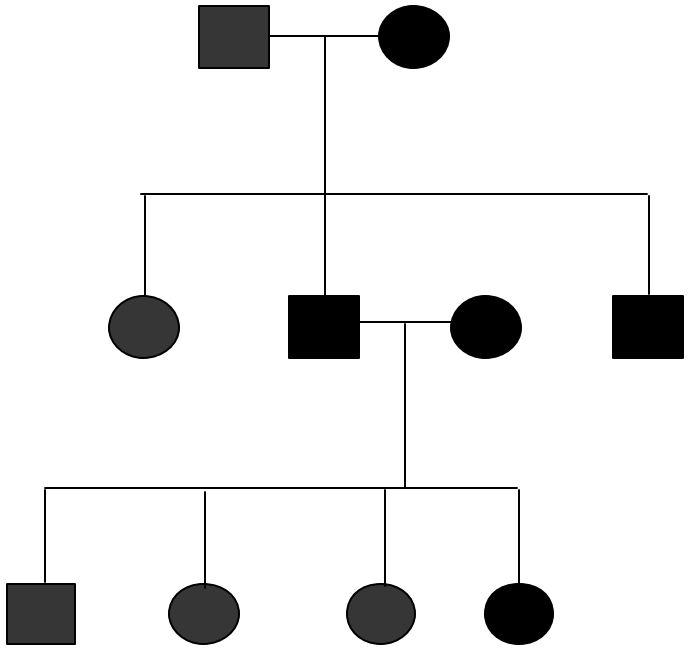
- Which differences matter?

How do we know genetics plays a role?

Twin studies

- Identical (monozygotic) twins are more similar than fraternal twins (dizygotic)
- Example: type 2 diabetes
  - MZ twins: >80% concordant
  - DZ twins: 30-50% concordant

How do we know genetics plays a role?



Family studies

- Risk to siblings and other relatives is greater than in the general population

- Example: type 2 diabetes
  - Risk to siblings: 30%
  - Population risk: 5-10%

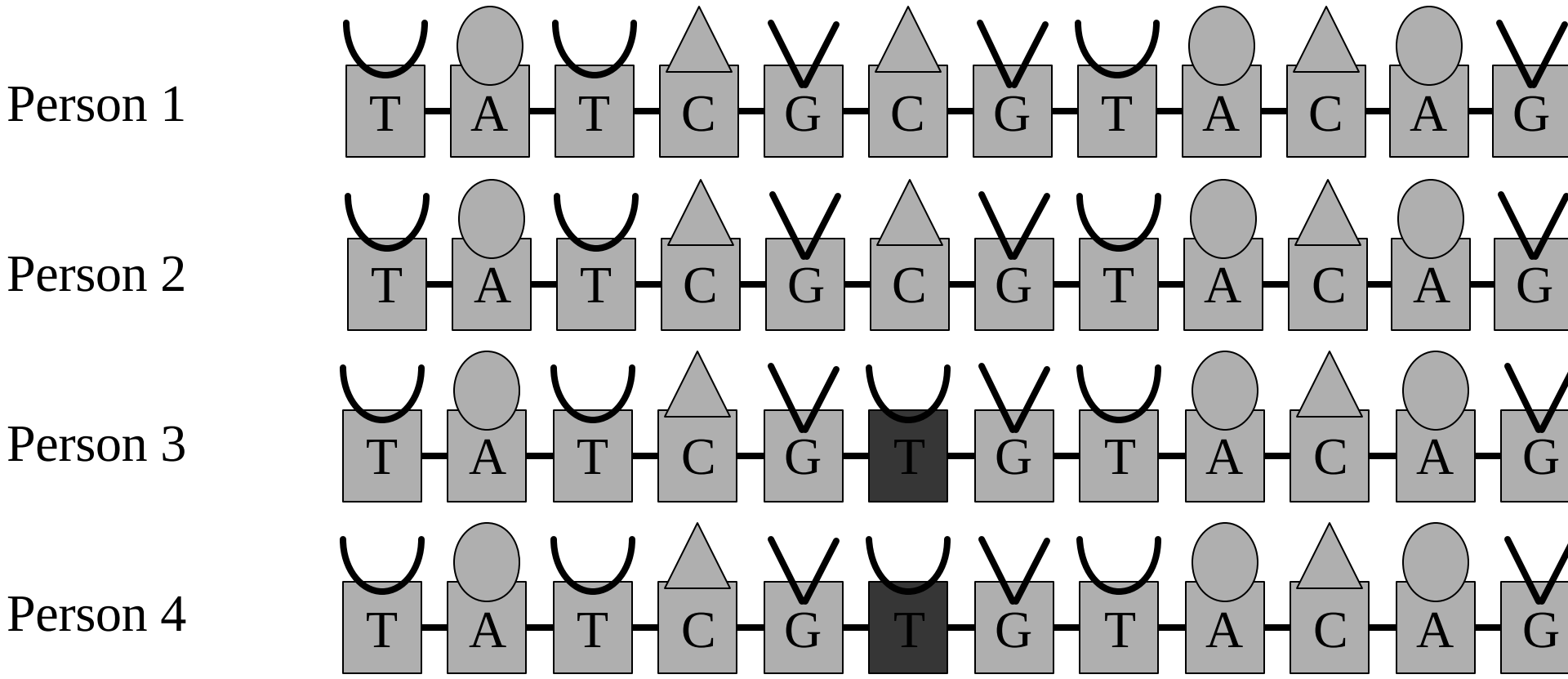
SNPs, patterns of variation, and complex traits

- Introduction
- Common genetic variation and disease
- Methods for finding variants for complex traits
- Interpreting genetic studies
  - Association
  - Linkage
  - Resequencing
- Approaches for the present and future
  - Haplotypes and linkage disequilibrium
- What could we learn?

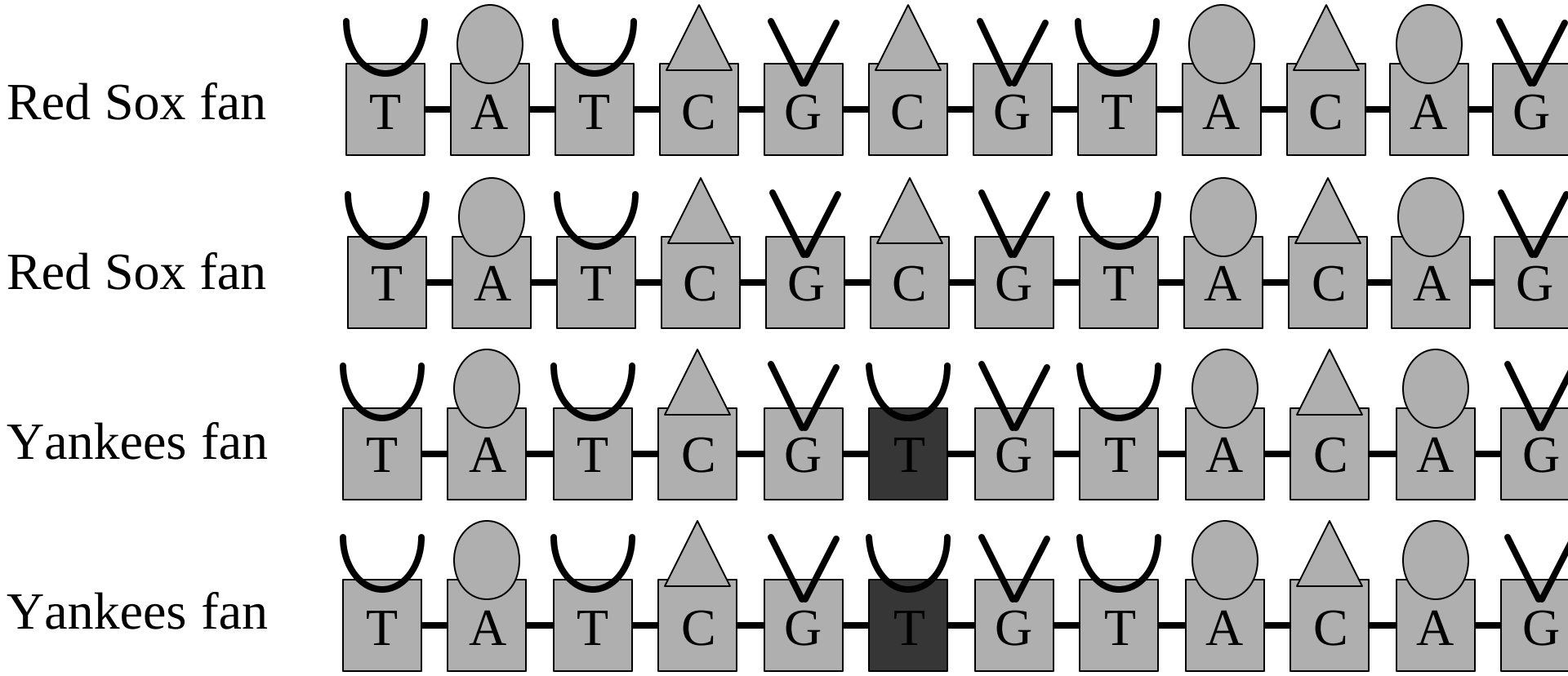

CCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGAC
TGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAC
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAAC
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCA
GTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACAT
TACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACAC
TCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAG
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAAC
GTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTA
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAG
CGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAG
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAG
CGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATA
ACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACA
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACA
TCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATAT
CTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACAT
GTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTA
TCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTA


CCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGAC
TGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAC
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAAC
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCA
GTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACAT
TACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACAC
TCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAG
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAAC
GTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTA
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAG
CGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAG
GATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATC
CTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACAC
ACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAG
CGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATA
ACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACA
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACAC
TCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCAT
TCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATAT
CTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACAT
GTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTA
TCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTA

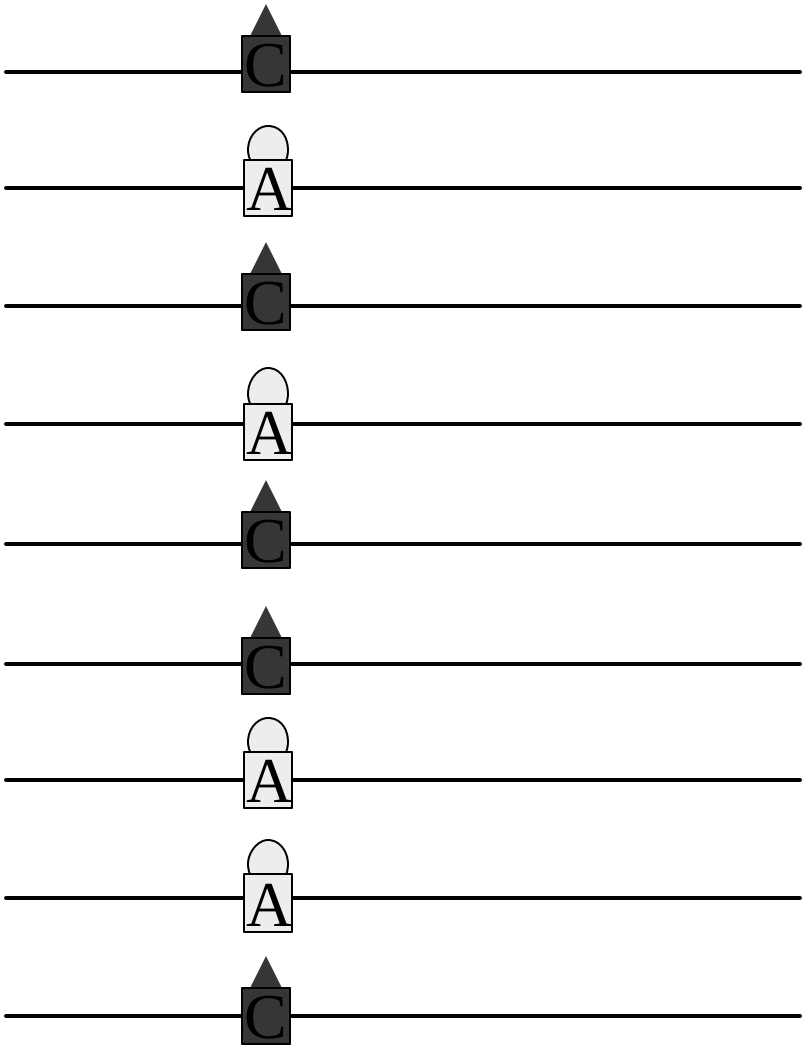
Most variants change a single DNA letter: single nucleotide polymorphism ("SNP")



Most variants change a single DNA letter: single nucleotide polymorphism (“SNP”)



Human variation and common variants



Shared, common variation
is the rule
(90% of heterozygosity)

Common disease-common variant hypothesis

- Most variation is evolutionarily neutral

- Most of this neutral variation is due to common variants

- Traits under negative selection will be largely due to rare variants
  - Pritchard et al., 2002
- Traits not under negative selection will be at least partly explained by common variants
  - Reich and Lander 2002

Cataloging common variation

- 10 million common SNPs (>1%)
- > 6 million are in databases

Please refer to UCSC SNP browser website at http://genome.ucsc.edu/

How to use these tools to find (common) disease alleles?

- Study every (common) variant?
  - Unbiased, genome-wide search
  - Not currently practical
- Need to select genes and variants to study

SNPs, patterns of variation, and complex traits

- Introduction
- Common genetic variation and disease
- Methods for finding variants for complex traits
- Interpreting genetic studies
  - Association
  - Linkage
  - Resequencing
- What could we learn?

Selecting genes and variants

Linkage: Narrow search to a small chromosomal region

- Affected relatives co-inherit markers in a region more often than expected by chance
- Monogenic disorders: successful
- Multigenic disorders: less successful

Association: Choose and test common variants in genes

- Candidate genes
- Well-suited to common alleles of modest penetrance

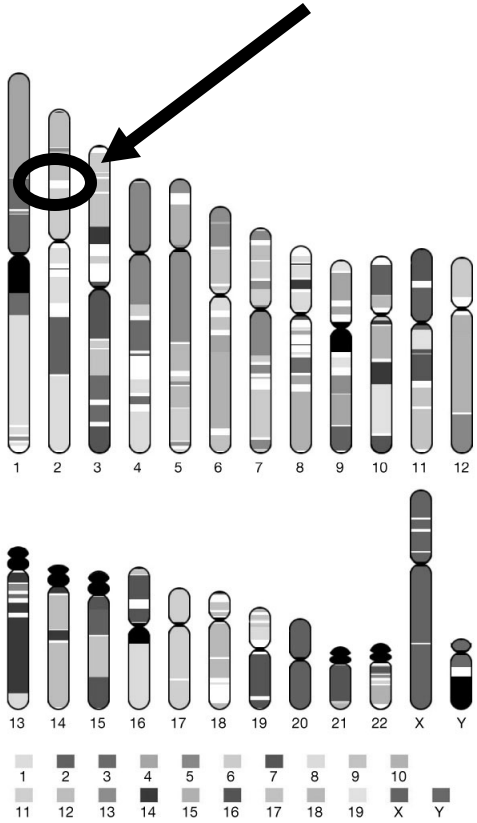
Association: Find and test rare variants in genes

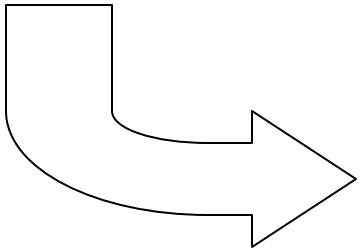
- Candidate genes
- Resequencing to find rare variants
- Very expensive

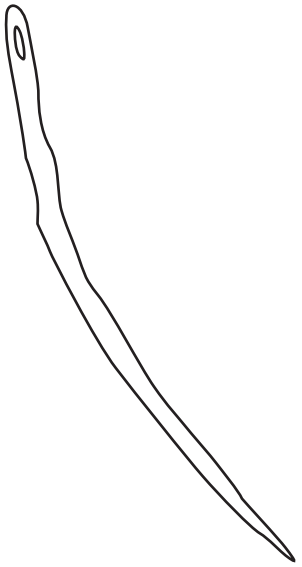
Finding variants that affect complex traits

Search the whole genome

Guess where to look





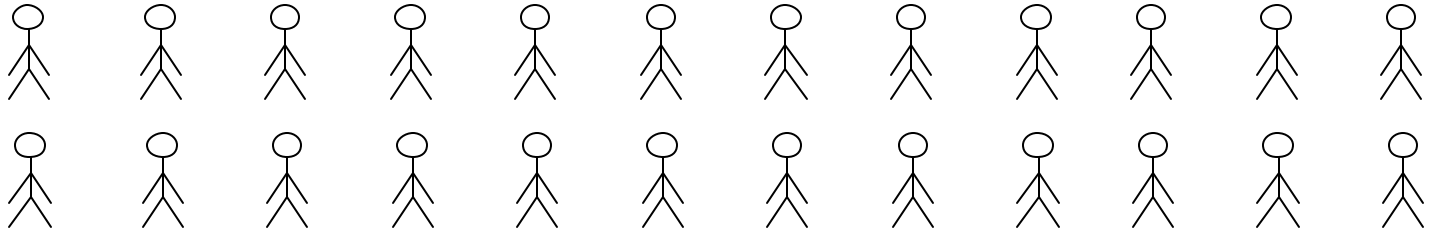


Linkage analysis

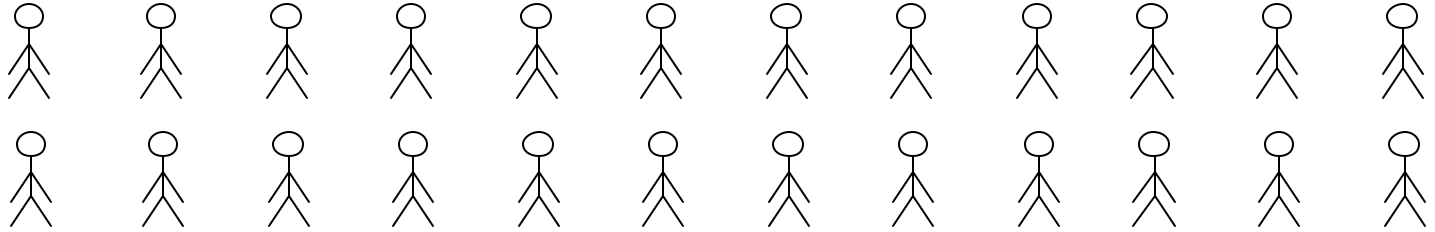
Candidate gene studies

Association studies to find disease alleles

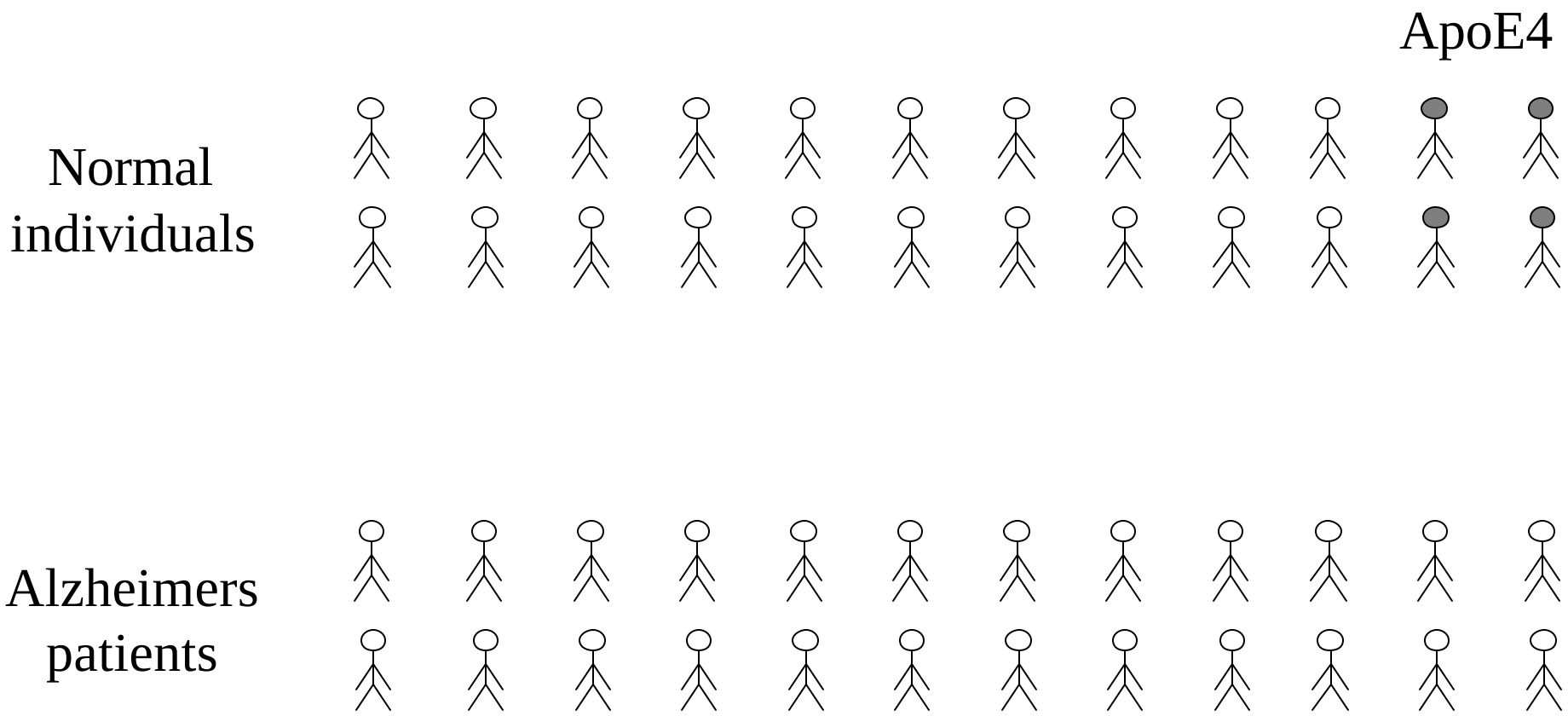
Normal individuals



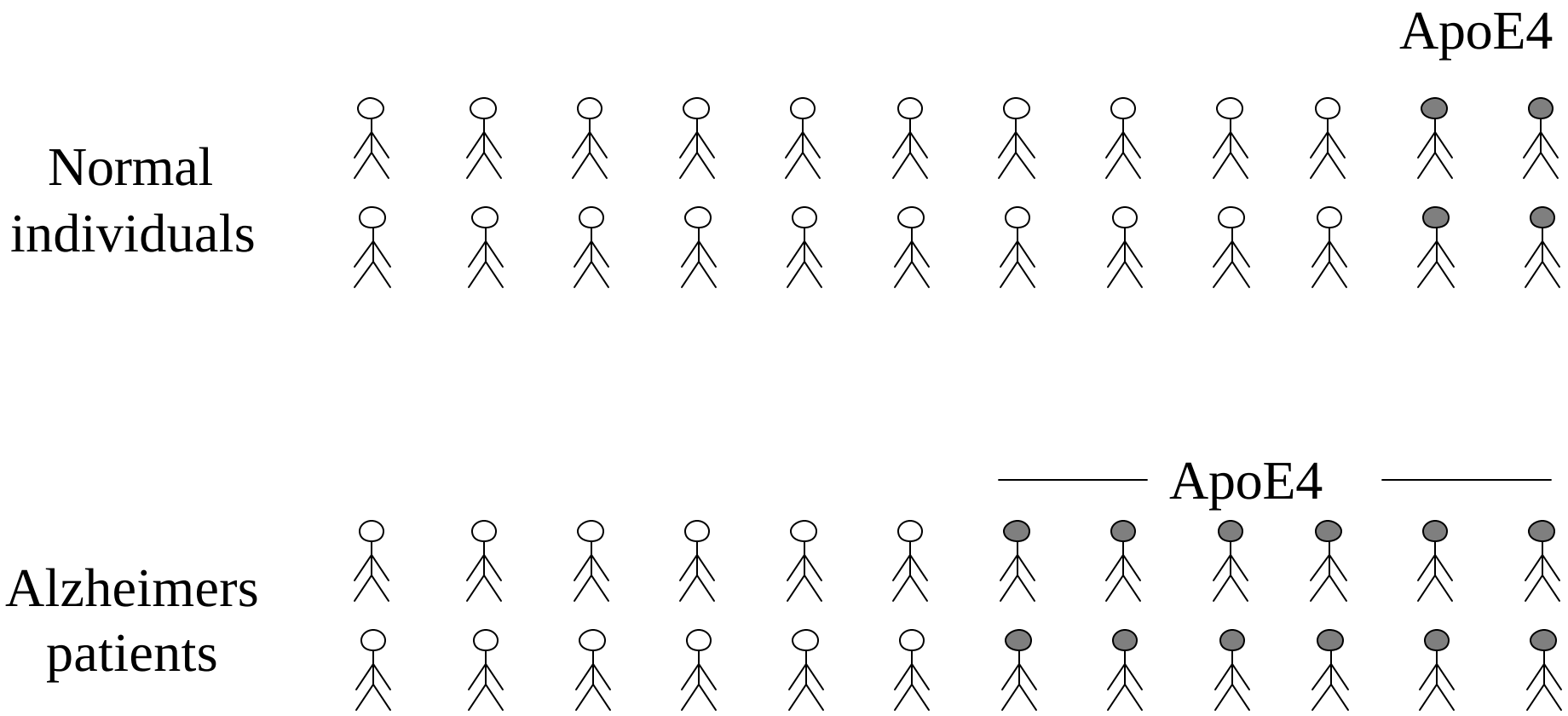
Alzheimers patients



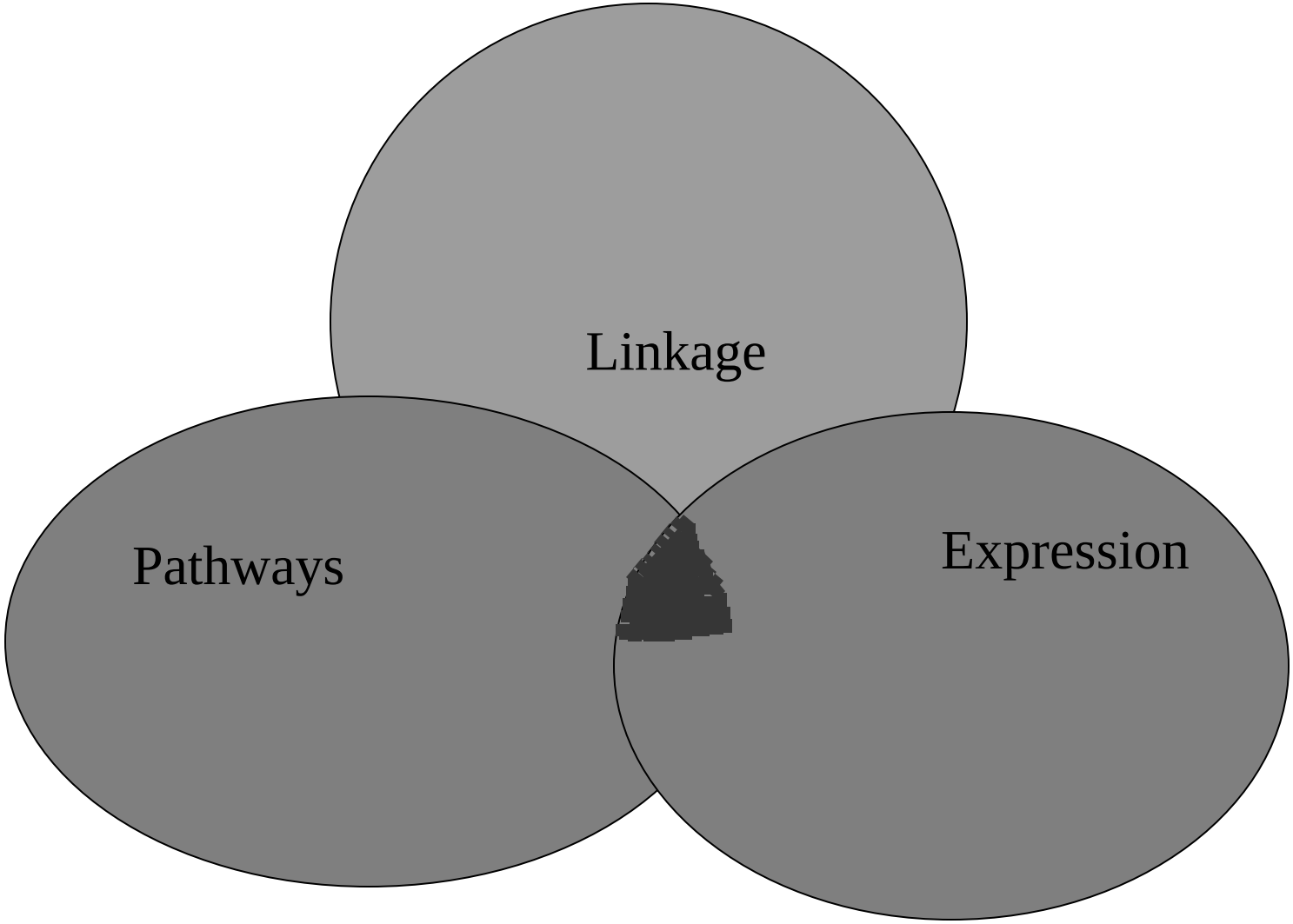
Association studies to find disease alleles



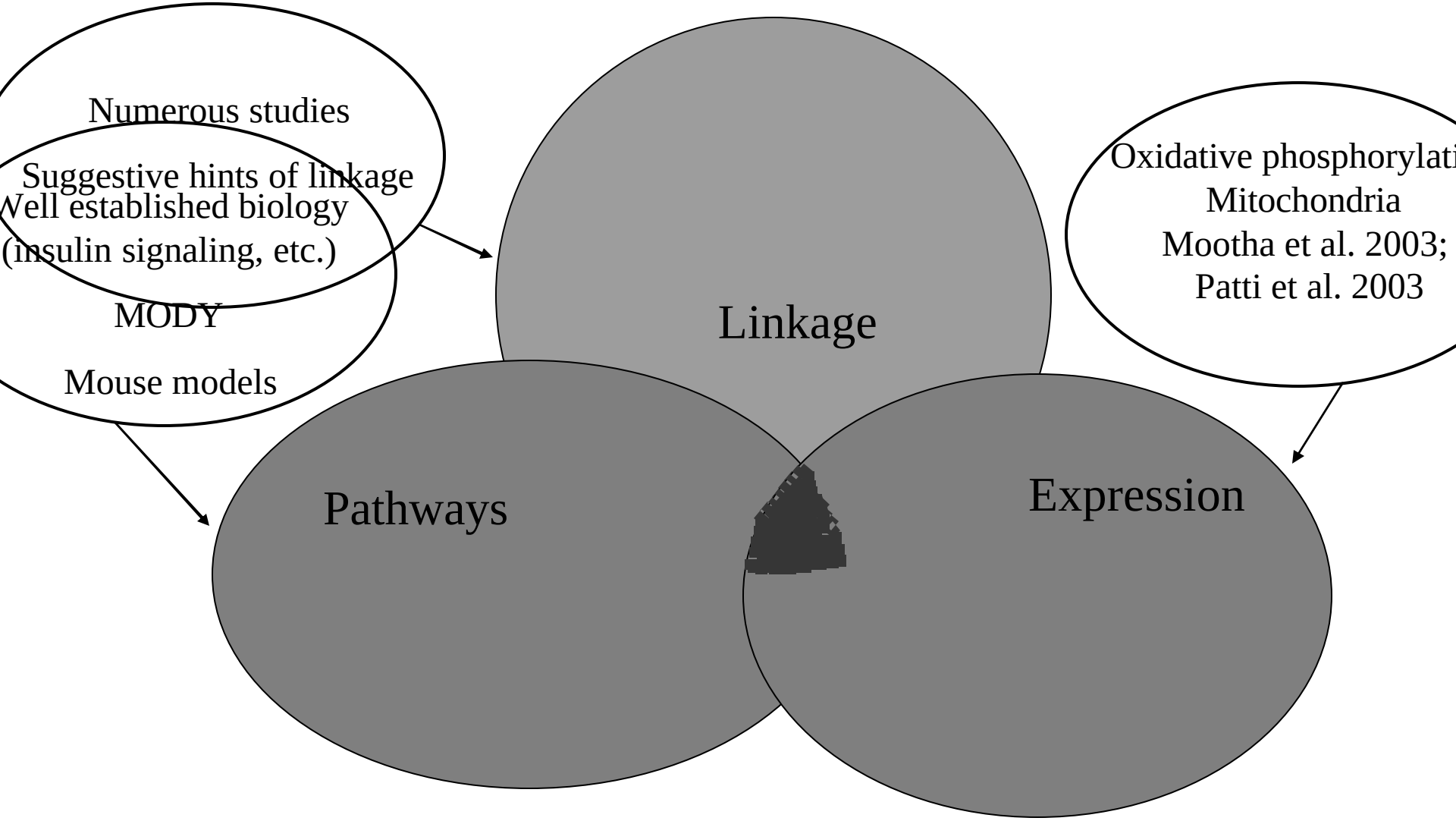
Association studies to find disease alleles



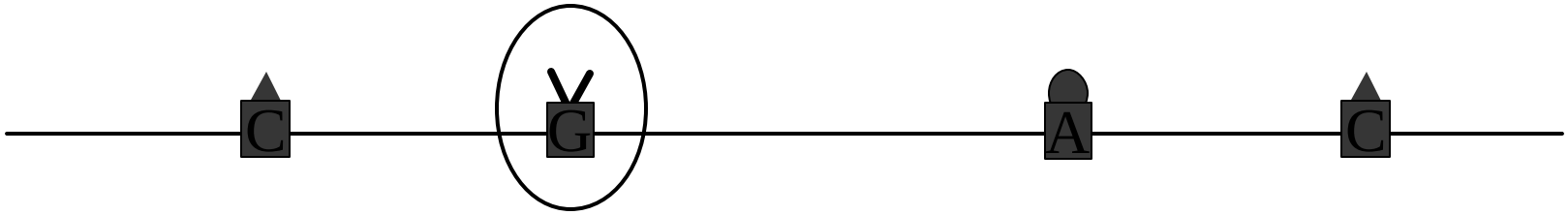
Association studies: which genes?



Type 2 diabetes: which genes?



Association studies: which variants?



Ideally, causal variant available and genotyped

Maximal power

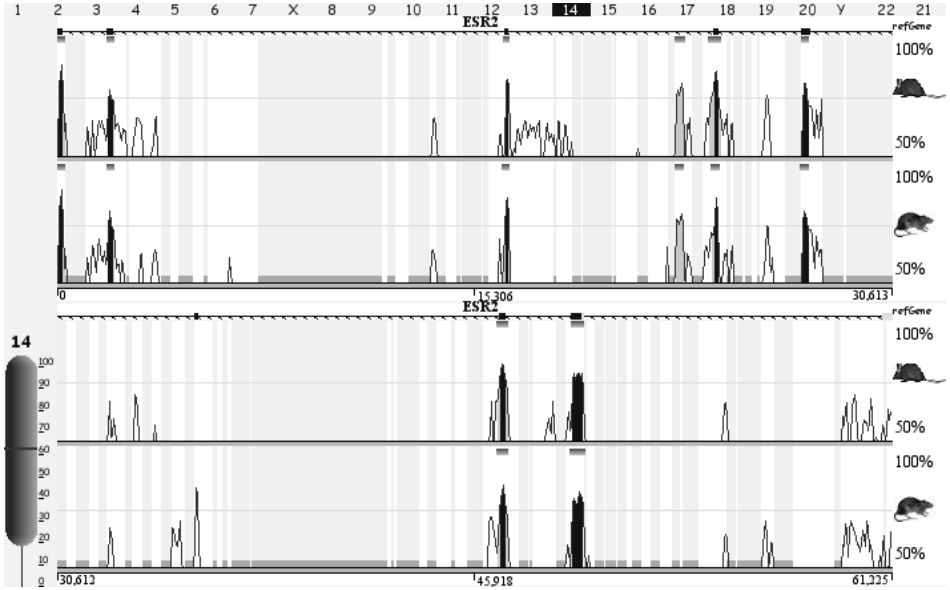
- marker tested is perfectly correlated with causal variant

Finding putative functional variants

- Missense variants
  - Easy to recognize
  - Many are mildly deleterious
  - Can group together variants (rare variant model)

Finding putative functional variants

- Regulatory variants
  - Hard to recognize
  - May be enriched in evolutionarily conserved noncoding regions (ECRs)

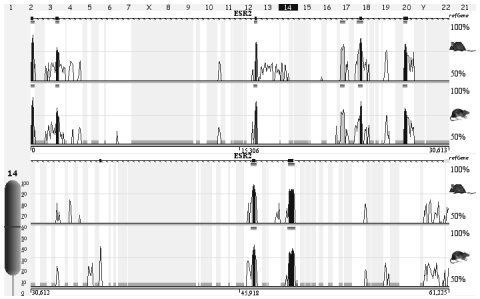


http://ecrbrowser.dcode.org
Lawrence Livermore
Eddy Rubin group

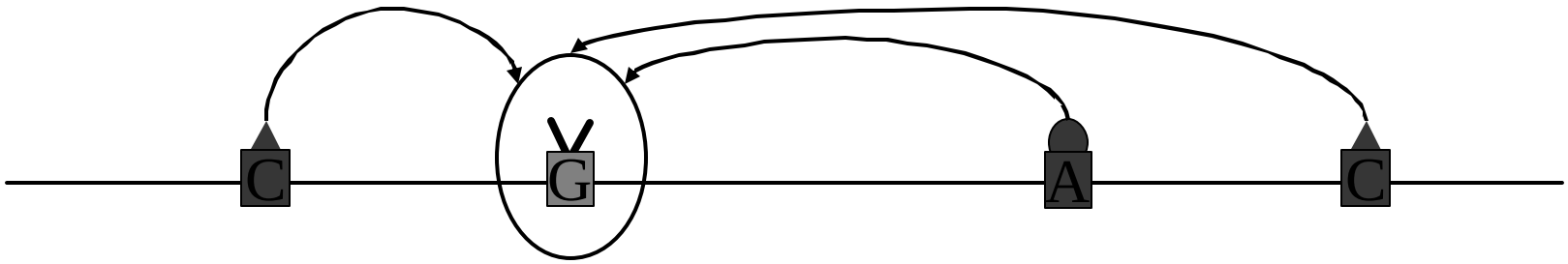
Resequencing to discover variants

DNA samples

Resequence target regions (expensive) → Identify SNPs (still not automated) →



An association might be indirect, so we should understand correlation between variants…

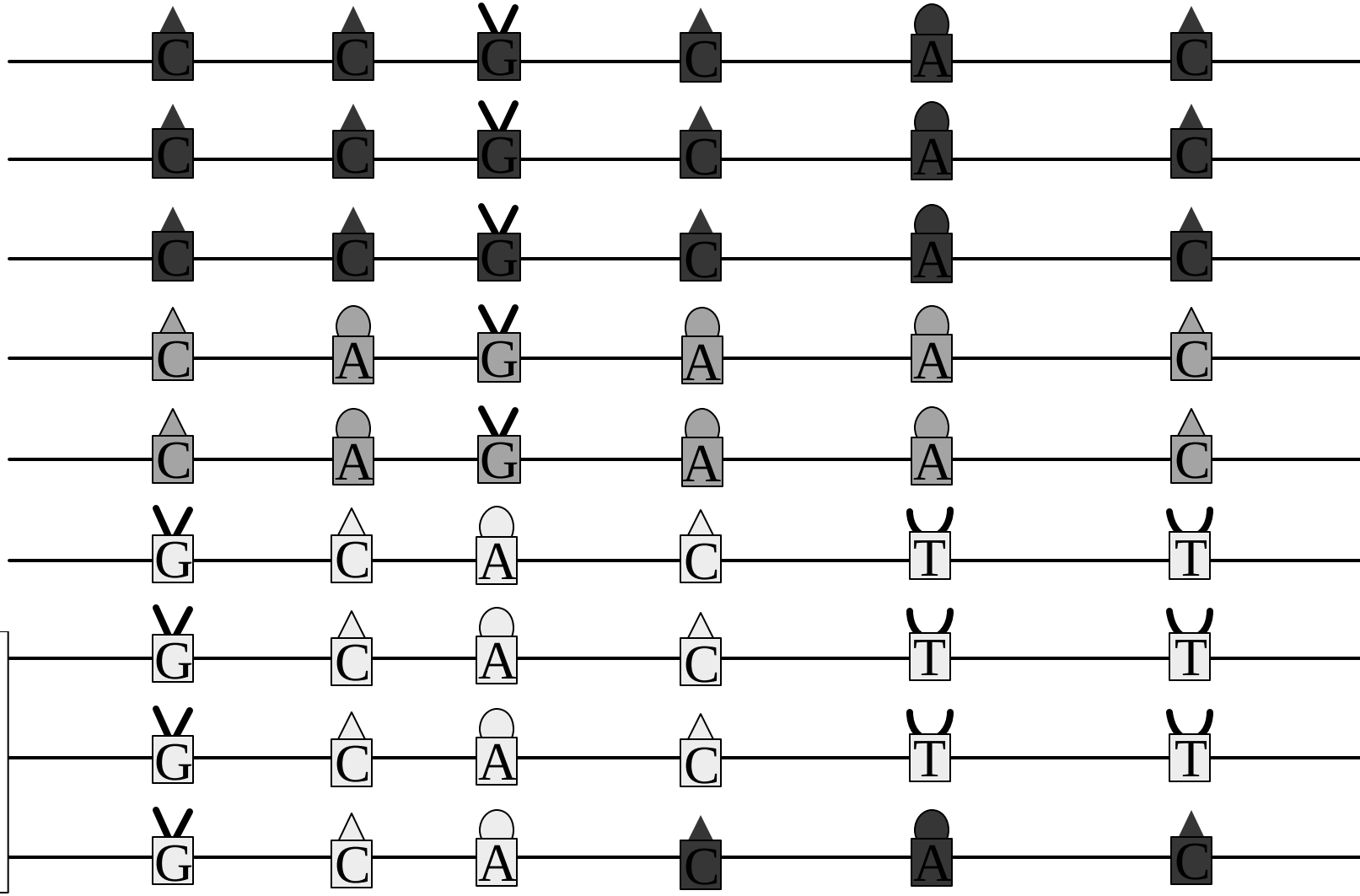


Causal variant not genotyped

Effect of causal variant inferred by genotyping neighboring SNPs

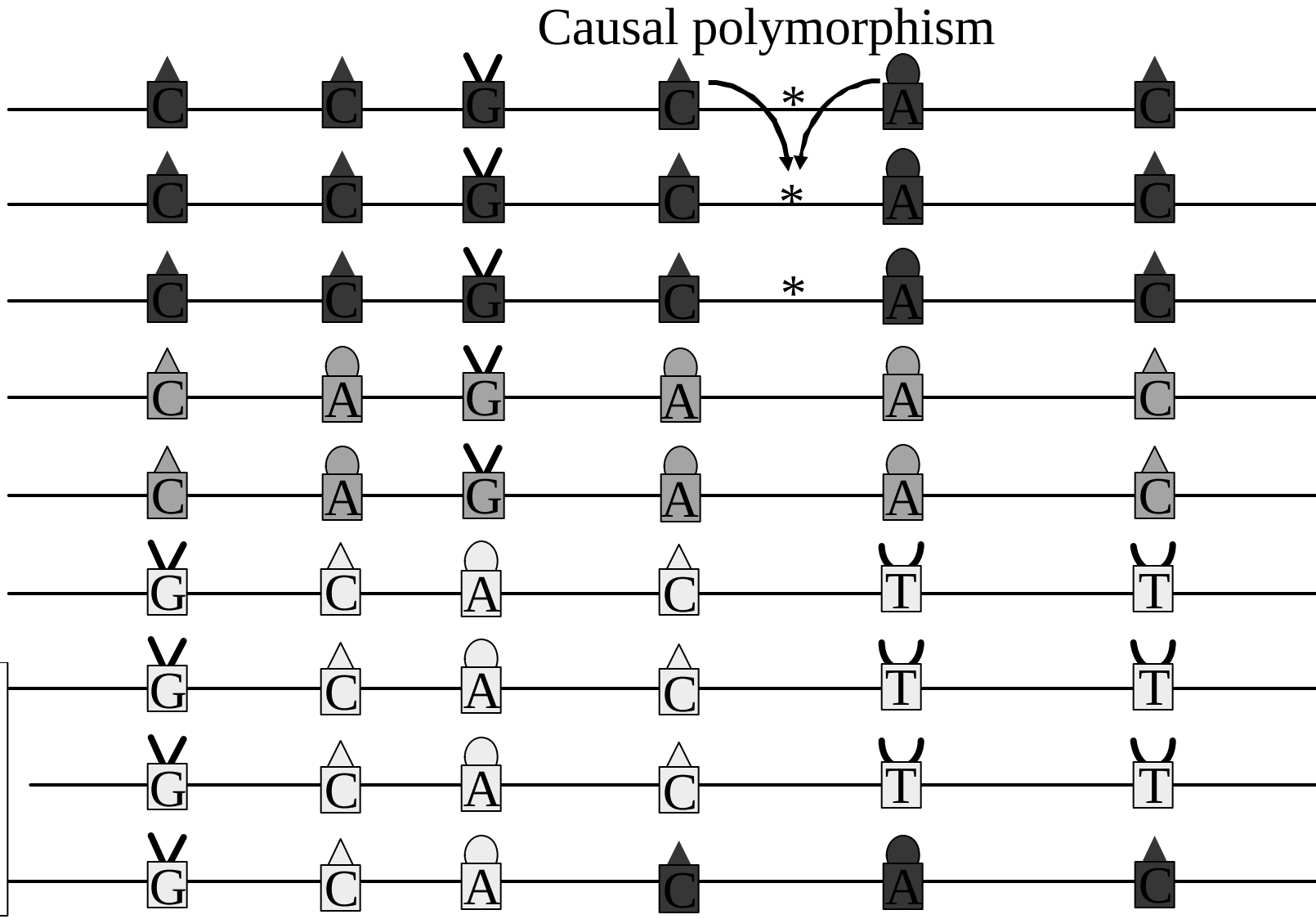
Neighbors must be correlated (in linkage disequilibrium) with causal variant

Haplotypes: patterns of variation at multiple markers (SNPs)



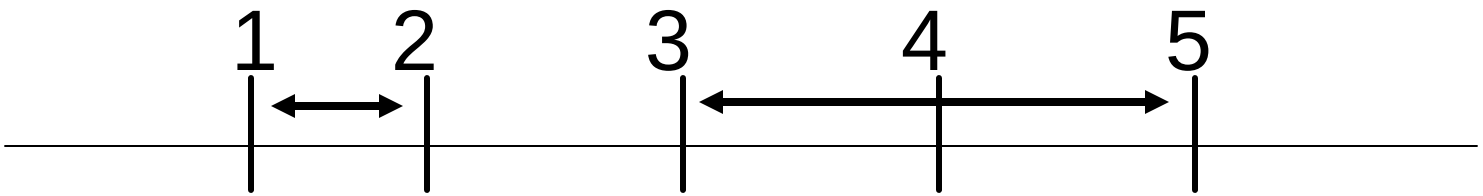
Gabriel et al. Science 2002
Daly et al. Nat Genet 2001

Using linkage disequilibrium (LD) to detect unknown variants



Gabriel et al. Science 2002
Daly et al. Nat Genet 2001

Measuring linkage disequilibrium (D’)



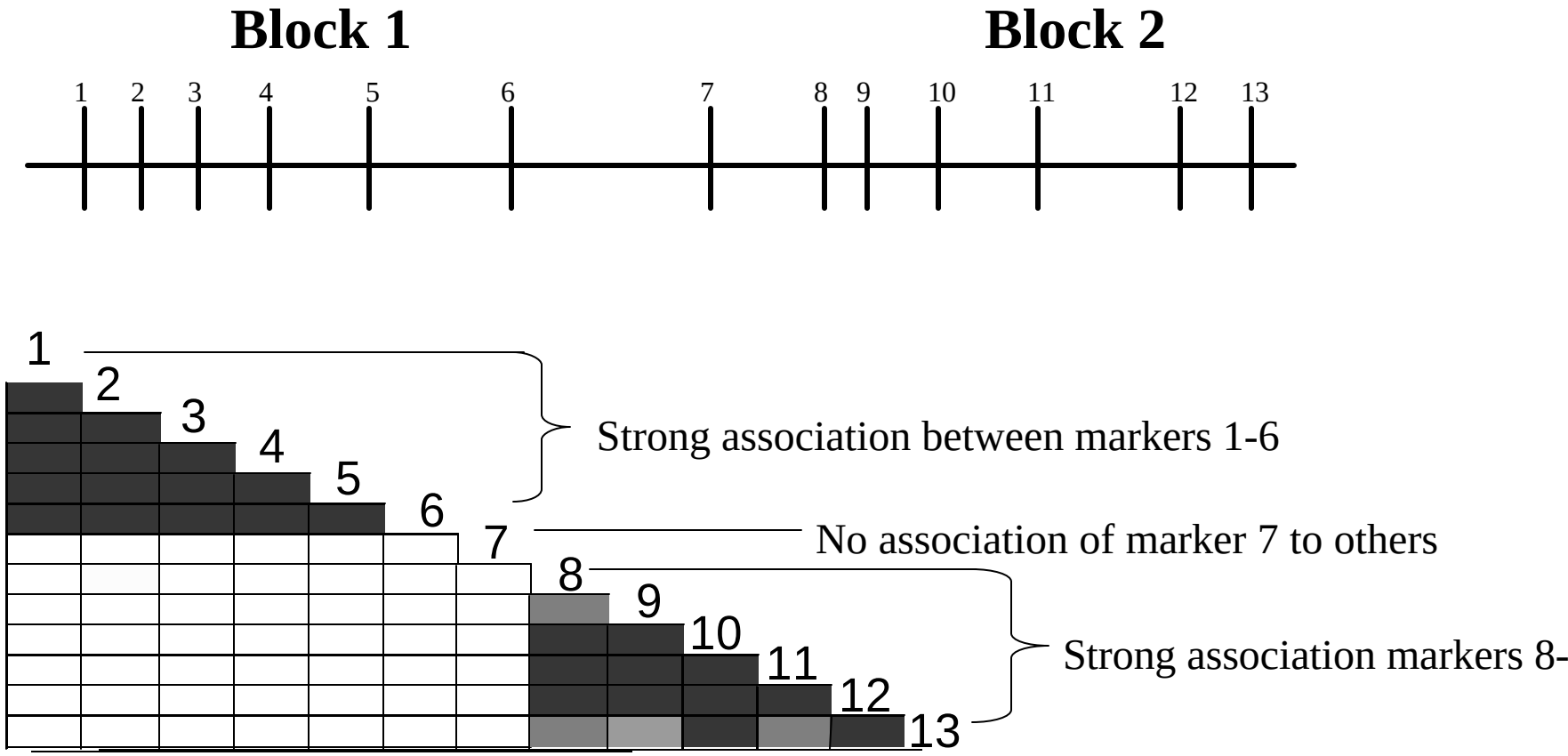
D’ value

LOD score in favor of LD

Red means LD that is strong and significant

1				
1.00 5.59	2			
1.00 22.21	1.00 5.59	3		
1.00 21.48	1.00 7.19	1.00 21.48	4	
0.54 1.21	0.19 0.20	0.54 1.21	0.45 0.97	5

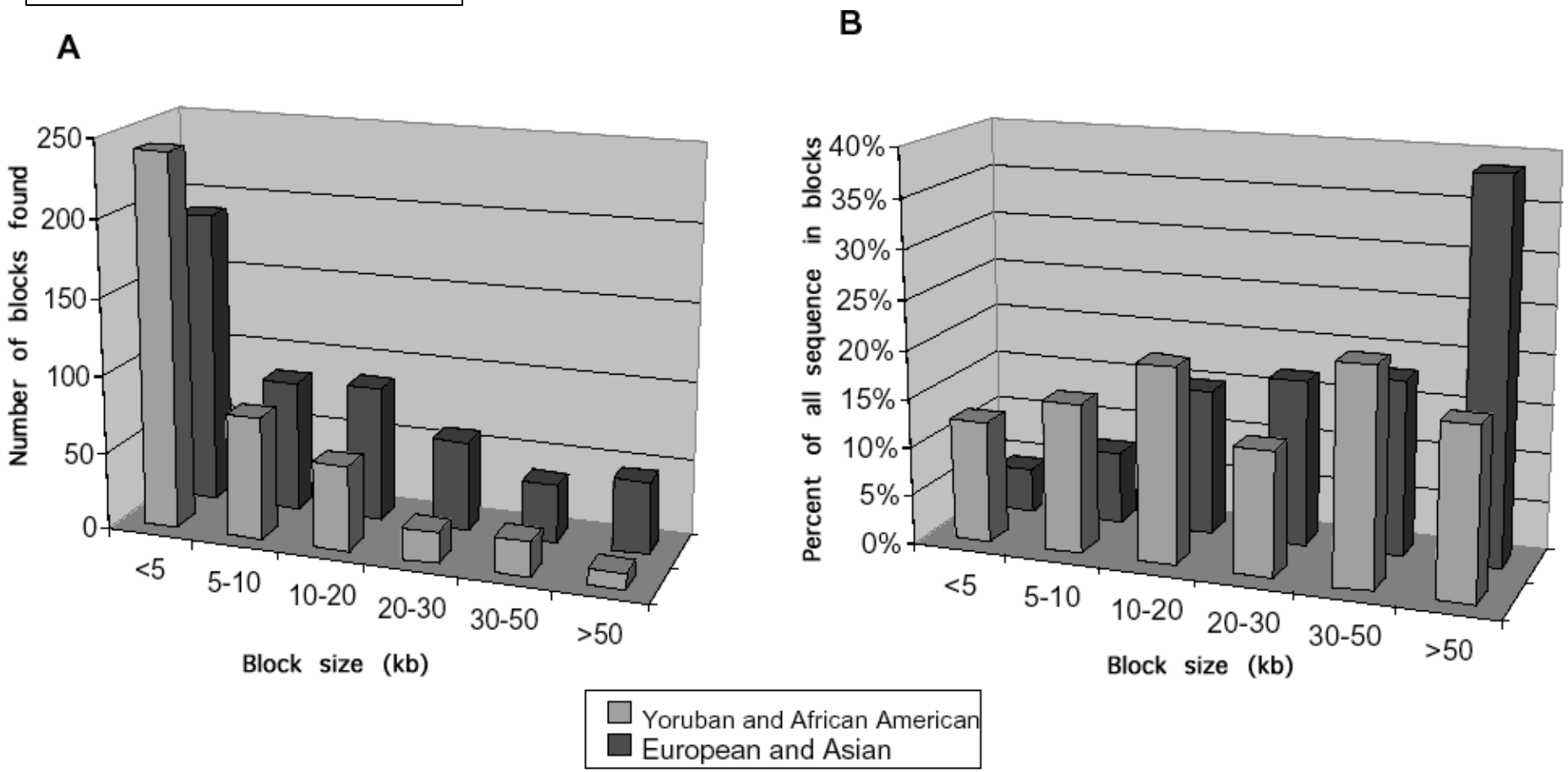
“Blocks” of linkage disequilibrium



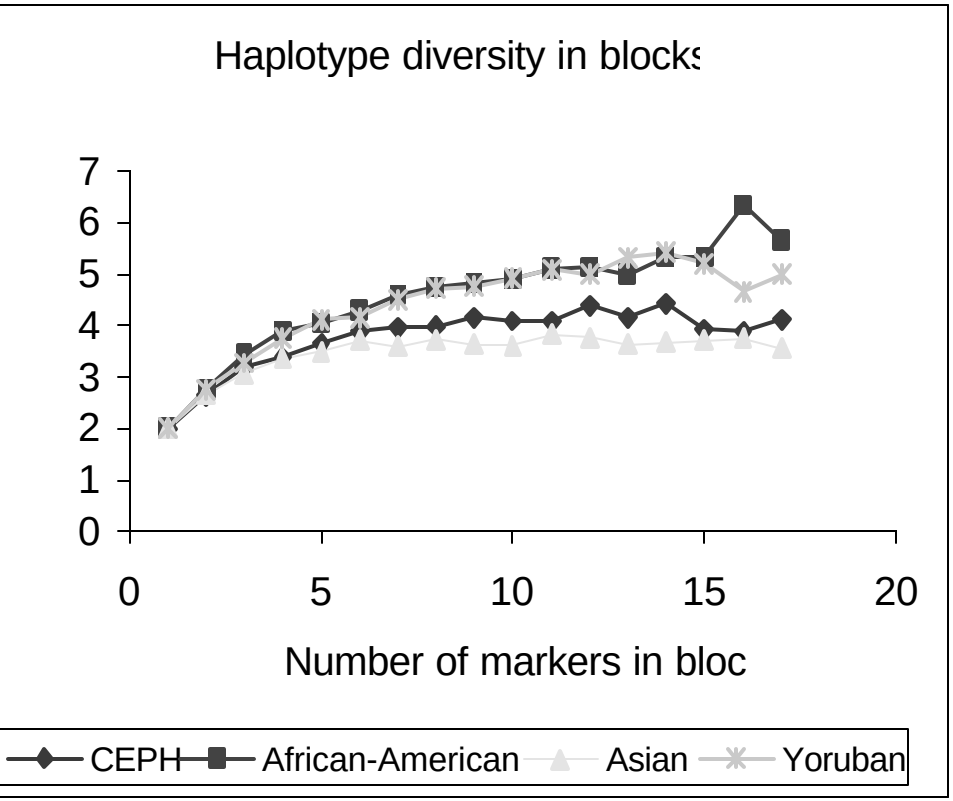
Daly et al., Nature Genetics 2001; Gabriel et al., Science, 2002

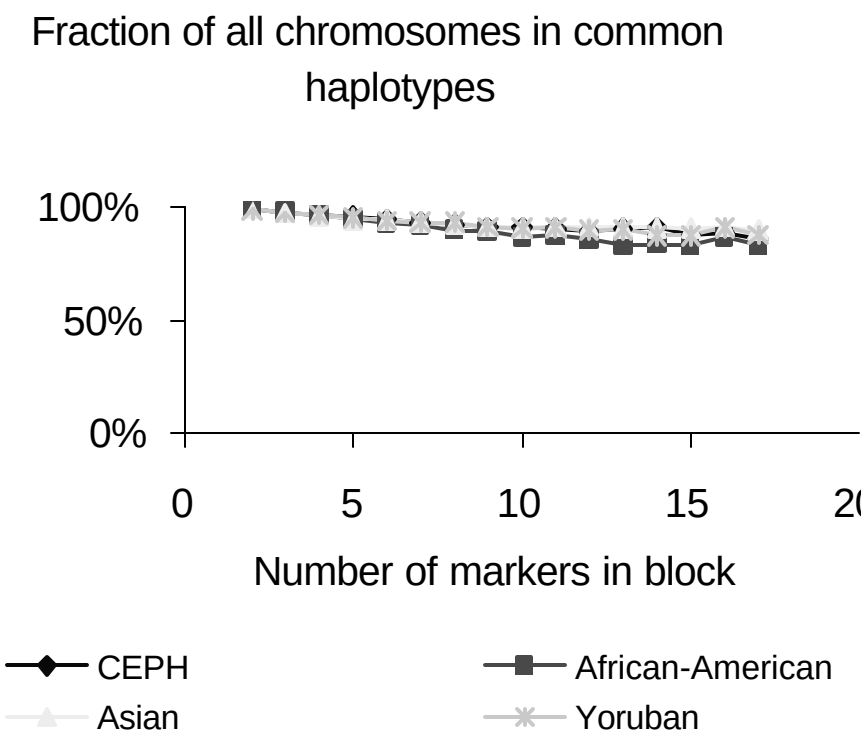
Distribution of sizes of haplotype blocks

Gabriel et al. 2002



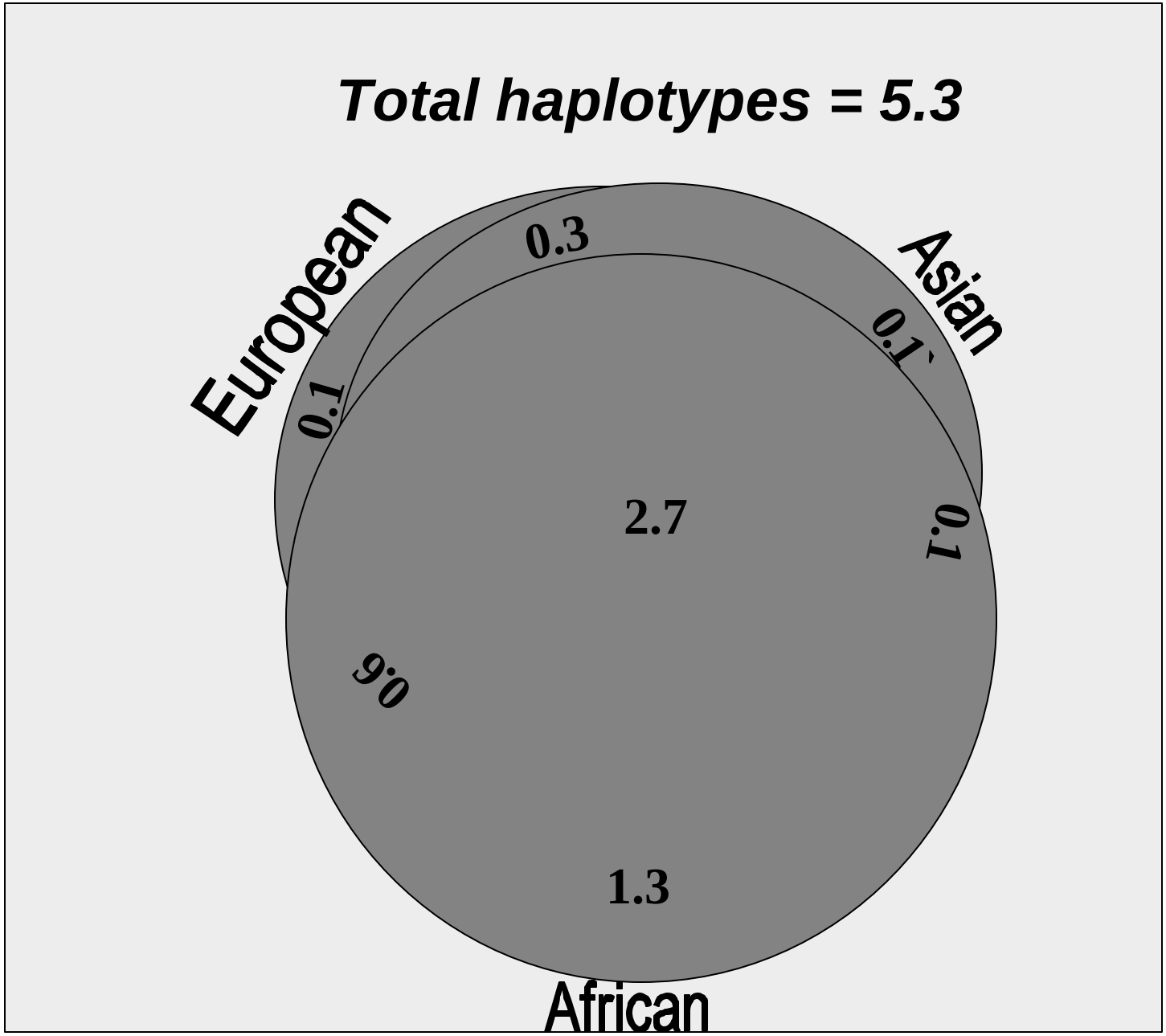
Within blocks, only a few common haplotypes explain 90% of chromosomes in each sample





4-5 common haplotypes

˜ 90% of all chromosomes


Total haplotypes = 5.3
European
Asian
African
0.3
0.1
0.1
2.7
0.1
0.6
1.3

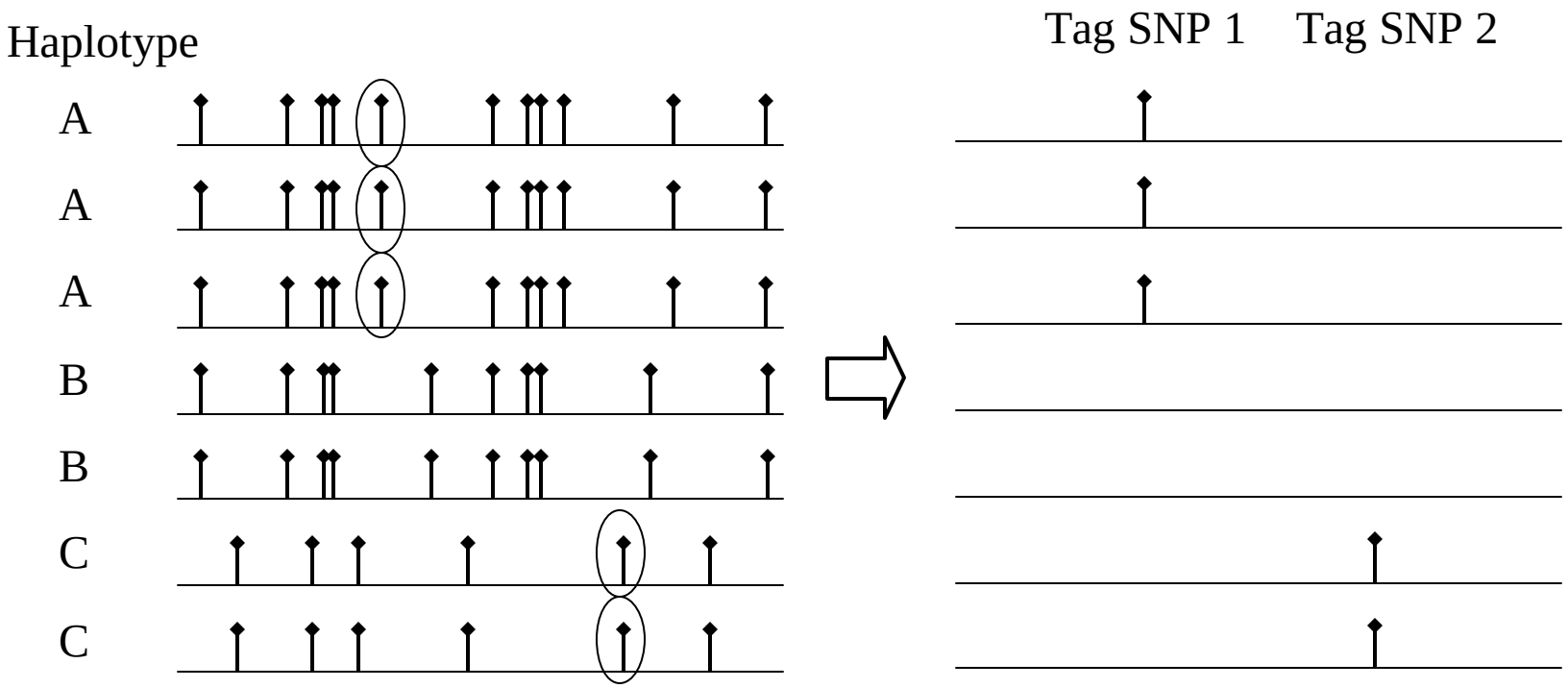
Biological and demographic forces contribute to shaping haplotype blocks

Please refer to
Jeffreys AJ, et. al.Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001 Oct;29(2):217-22.

“Hotspots” of recombination

Human demographic history

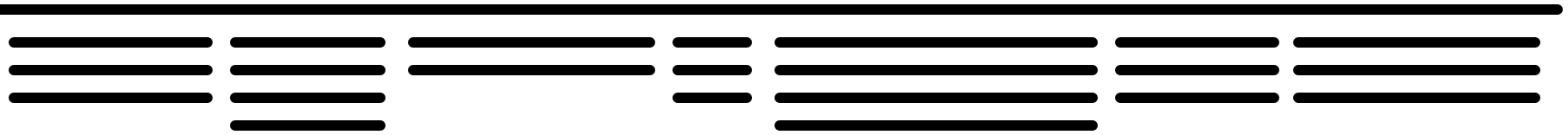
Using tag SNPs to capture common variation



By typing an adequate density of SNPs, one can identify tags that capture the vast majority of common variation in a region

Johnson et al., Nature Genetics 2001; Gabriel et al., 2002; Stram et al. 2003; others

Haplotype Map of Human Genome



Goals:

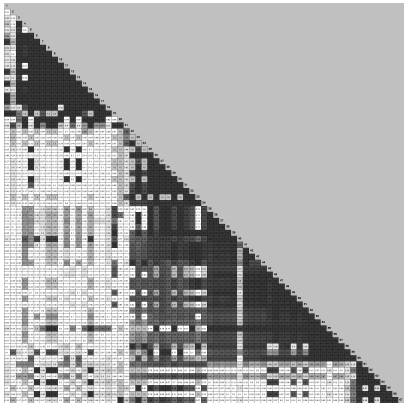
- **Define haplotype “blocks” across the genome**
- **Identify reference set of SNPs: “tag” each haplotype**
- **Enable unbiased, genome-wide association studies**

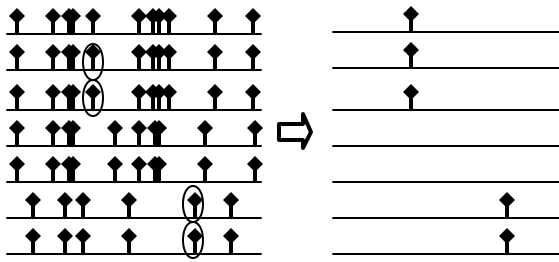
www.hapmap.org; see Nature 2993 426:789-96

Approach to LD-based association studies

SNPs from database

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.





Genotype SNPs in reference panels

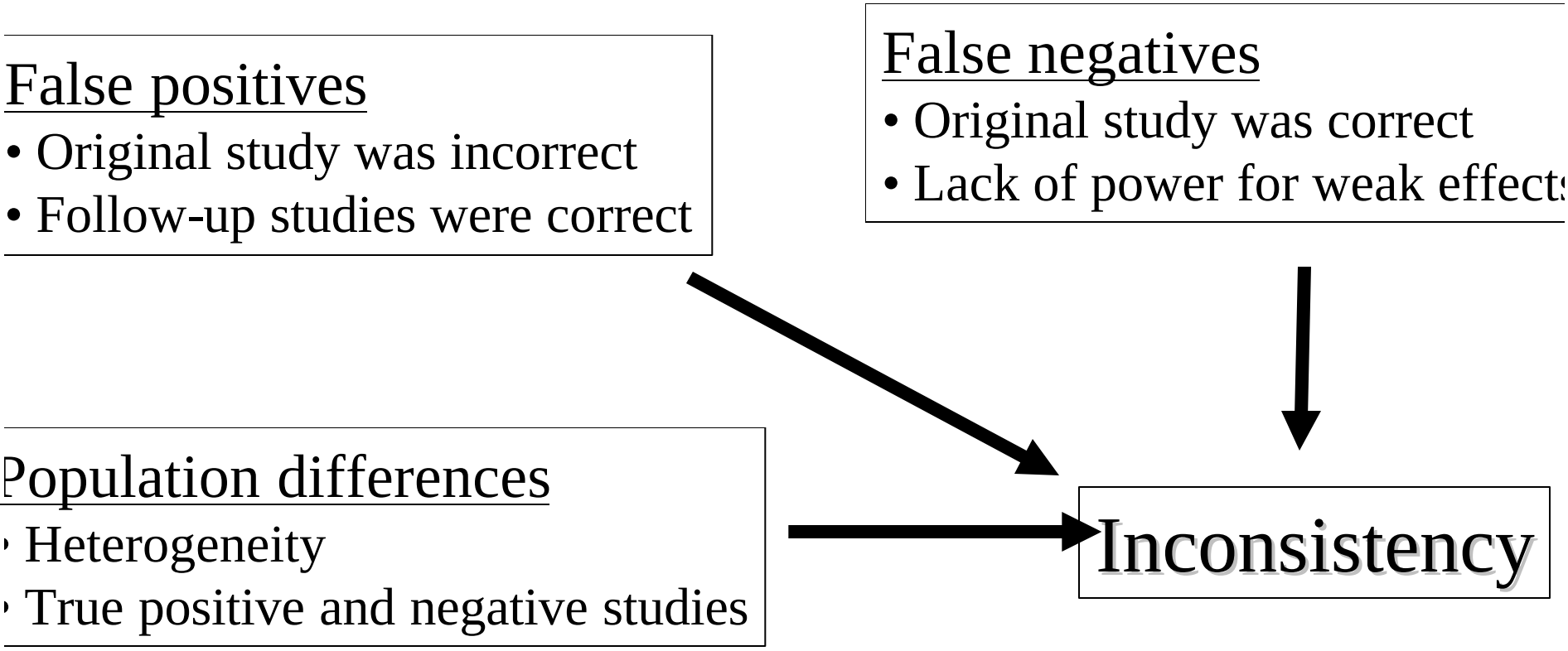
Measure LD, determine haplotype and select tag SNPs

SNPs, patterns of variation, and complex traits

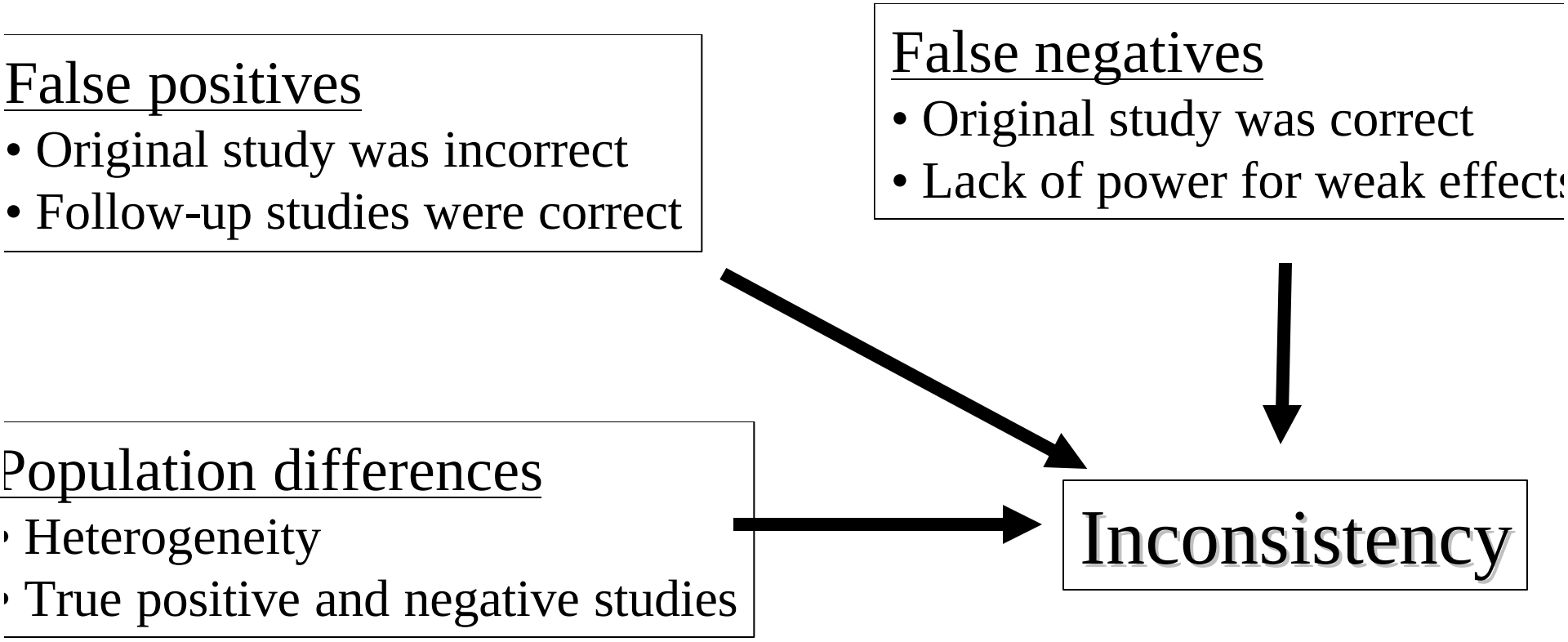
- Introduction
- Common genetic variation and disease
- Methods for finding variants for complex traits
- Interpreting genetic studies
  - Association
  - Linkage
  - Resequencing
- What could we learn?

Association studies are powerful but problematic

Most reported associations have not been consistently reproduced



What explains the lack of reproducibility?



Review of association studies

603 associations of polymorphisms and disease

166 studied in at least three populations

Only six seen in =75% of studies

Hirschhorn et al., Genetics in Medicine, 2002

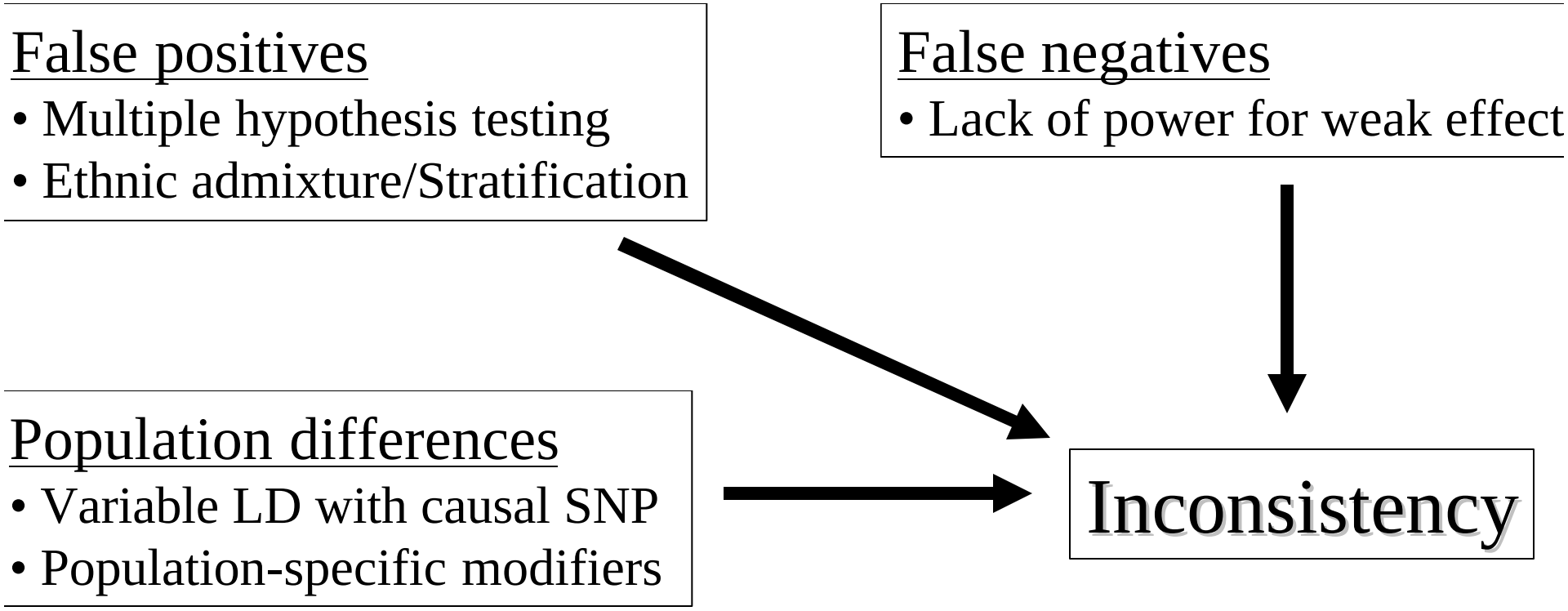
Highly consistently reproducible associations

Gene	Polymorphism	Disease
APOE	epsilon 4	Alzheimer's Disease
CCR5	delta32	HIV infection/AIDS
CTLA4	T17A	Graves' Disease
F5	R506Q	Deep Venous Thrombosis
INS	VNTR	Type 1 Diabetes
PRNP	M129V	Creutzfeld-Jacob Disease

What about the other 160?

91/160 seen at least one more time

What explains the lack of reproducibility?



Meta-analysis of association studies

- Selected 25 inconsistent associations with diallelic markers
  - Bipolar disease (2)
  - Schizophrenia (6)
  - Type 2 diabetes (9)
  - Random (8)

301 studies,
excluding original positive reports

If no true associations:
expect 5% to have $P < 0.05$
1% to have $P < 0.01$, etc.

Lohmueller et al., Nature Genetics 2003

Rate of replication for 25 inconsistent associations

Large excess of significant follow-up studies

- 20% of 301 studies had $P < 0.05$ (vs. 5% expected, $P < 10^{-14}$)
- Most (47/59) were in same direction as original report
- Replications were clustered among 11 of the 25 associations

Publication bias - can it explain excess replications?

Rate of replication for 25 inconsistent associations

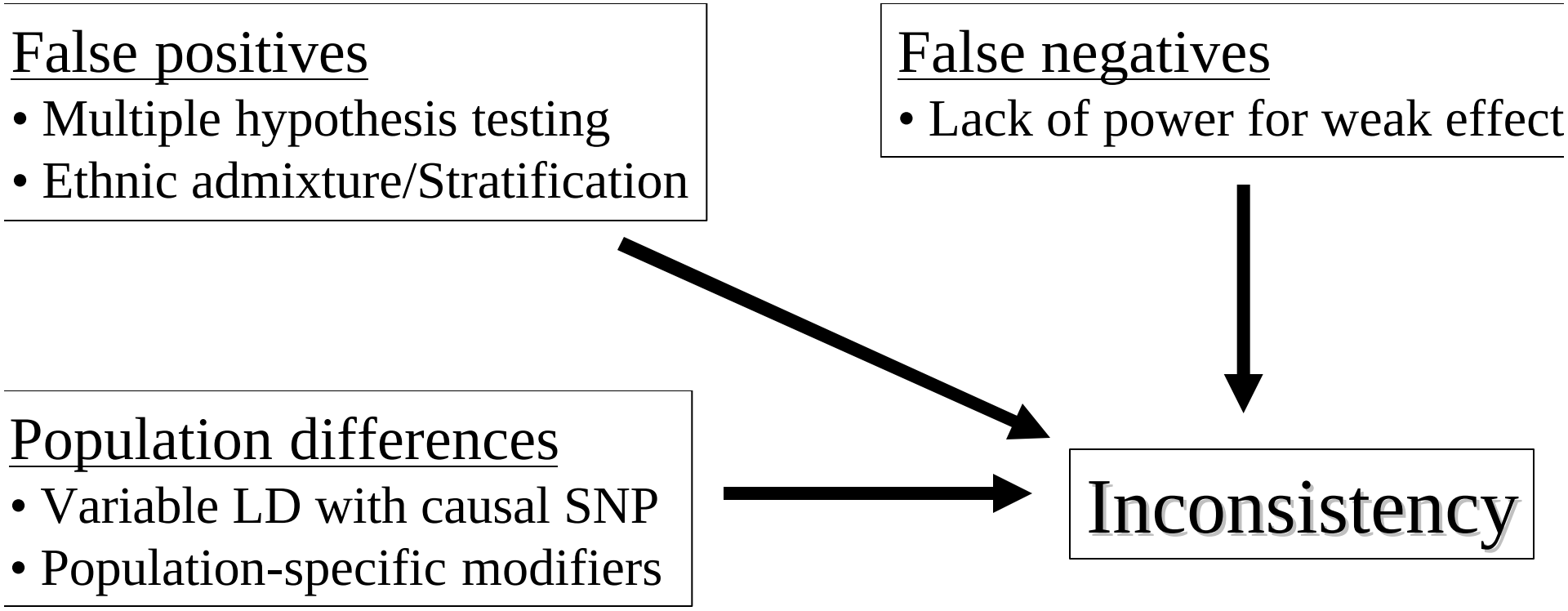
Large excess of significant follow-up studies

- 20% of 301 studies had $P < 0.05$ (vs. 5% expected, $P < 10^{-14}$)
- Most (47/59) were in same direction as original report
- Replications were clustered among 11 of the 25 associations

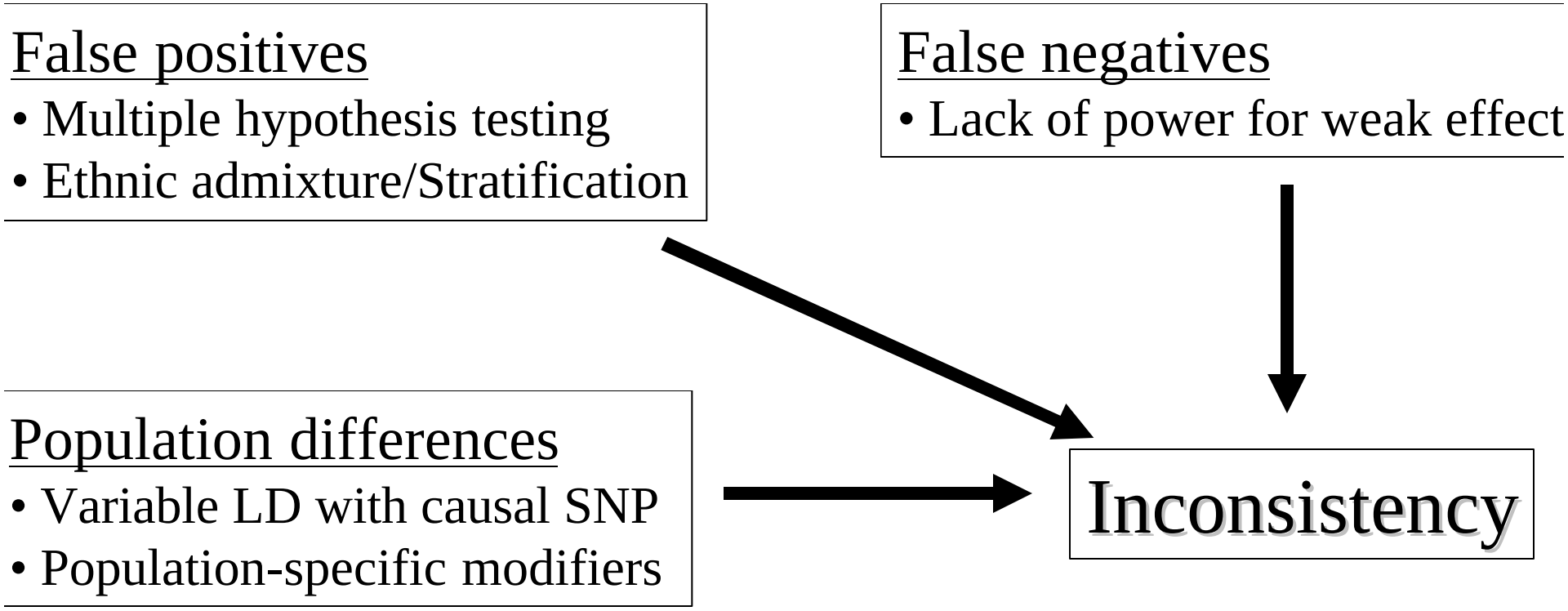
Probably not publication bias

- Requires postulating 40-80 unpublished studies/association

What explains the lack of reproducibility?

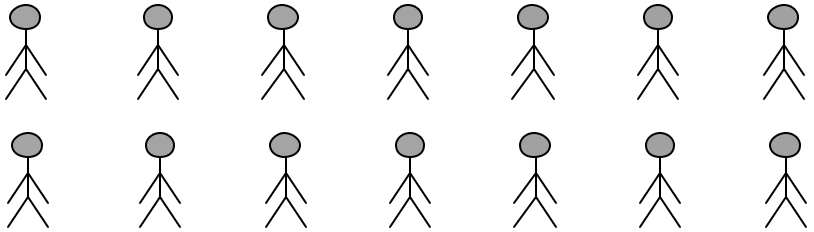


What explains the lack of reproducibility?

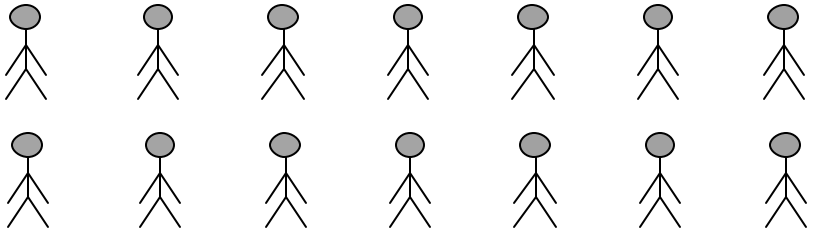


Ethnic admixture and population stratification

Cases



Controls

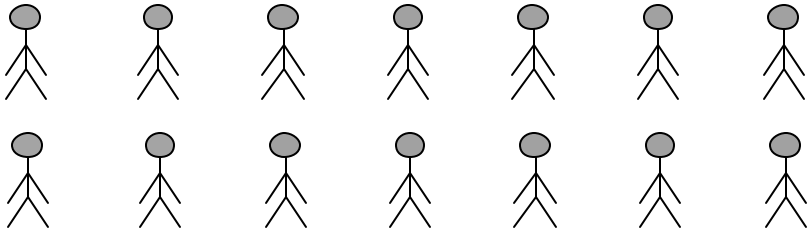


Well-matched

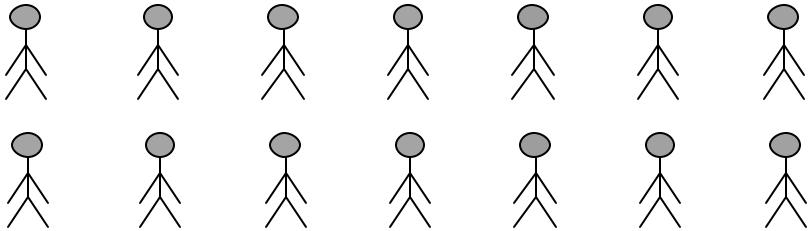
No stratification

Ethnic admixture and population stratification

Cases



Controls



Poorly matched

Stratification present

Assessing and controlling for stratification

- Family-based tests of association
  - TDT
  - Sib-based tests (SDT, PDT, Sib-TDT)
  - FBAT
- Genomic control
  - Type many random markers
  - Determine frequency of false positive associations
  - Use genotype data to match cases and controls

Spielman et al. 1993; Spielman and Ewens 1998; Martin et al 2000; Horvath et al. 2001; Pritchard and Rosenberg 1999; Pritchard et al. 2000; Devlin and Roeder, 1999; Reich and Goldstein, 2001

Rate of replication for 25 inconsistent associations

Large excess of significant follow-up studies

- 19% of 298 studies had $P < 0.05$ (vs. 5% expected, $P < 10^{-14}$)
- Most (45/56) were in same direction as original report
- Replications were clustered among 11 of the 25 associations

Probably not publication bias

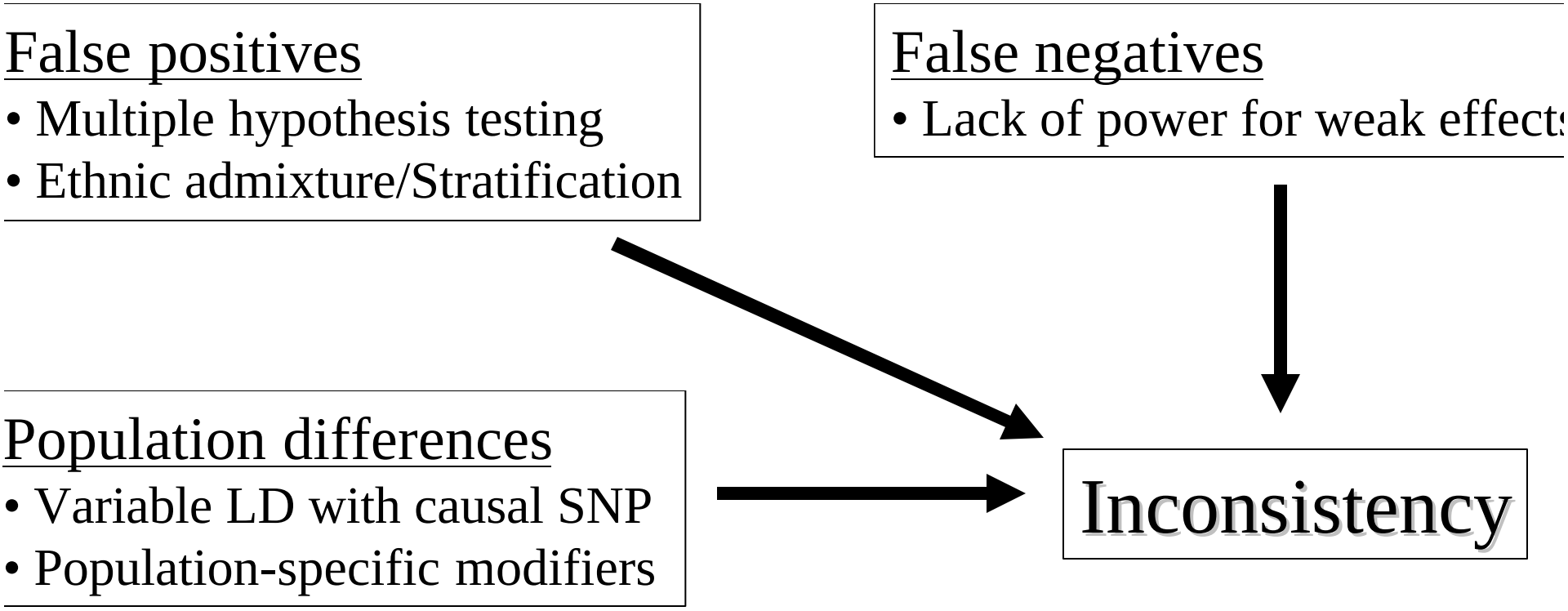
- Requires postulating 40-80 unpublished studies/association

Probably not population stratification/admixture

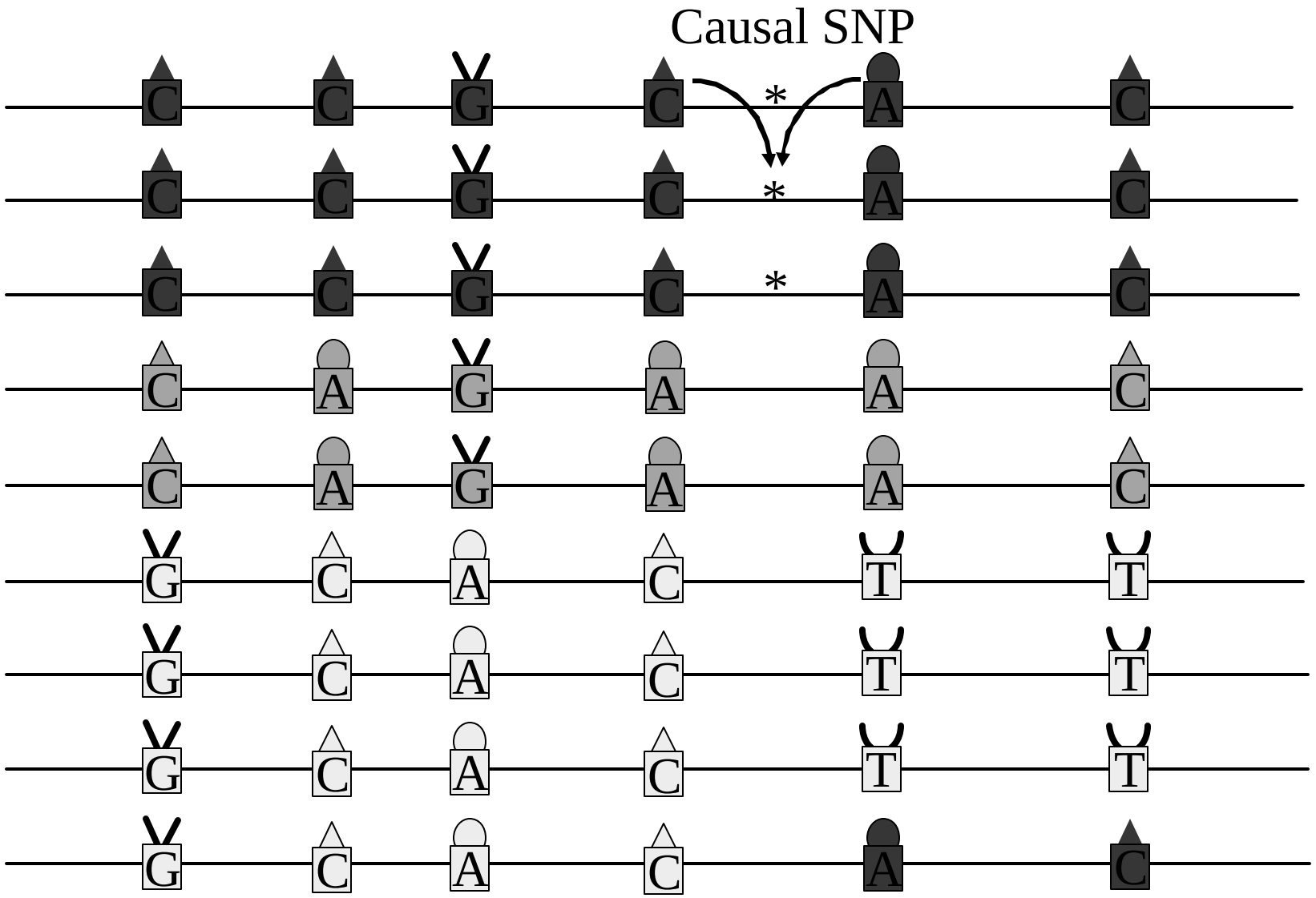
- Family-based controls and/or seen in multiple ethnic groups

Association studies are powerful but problematic

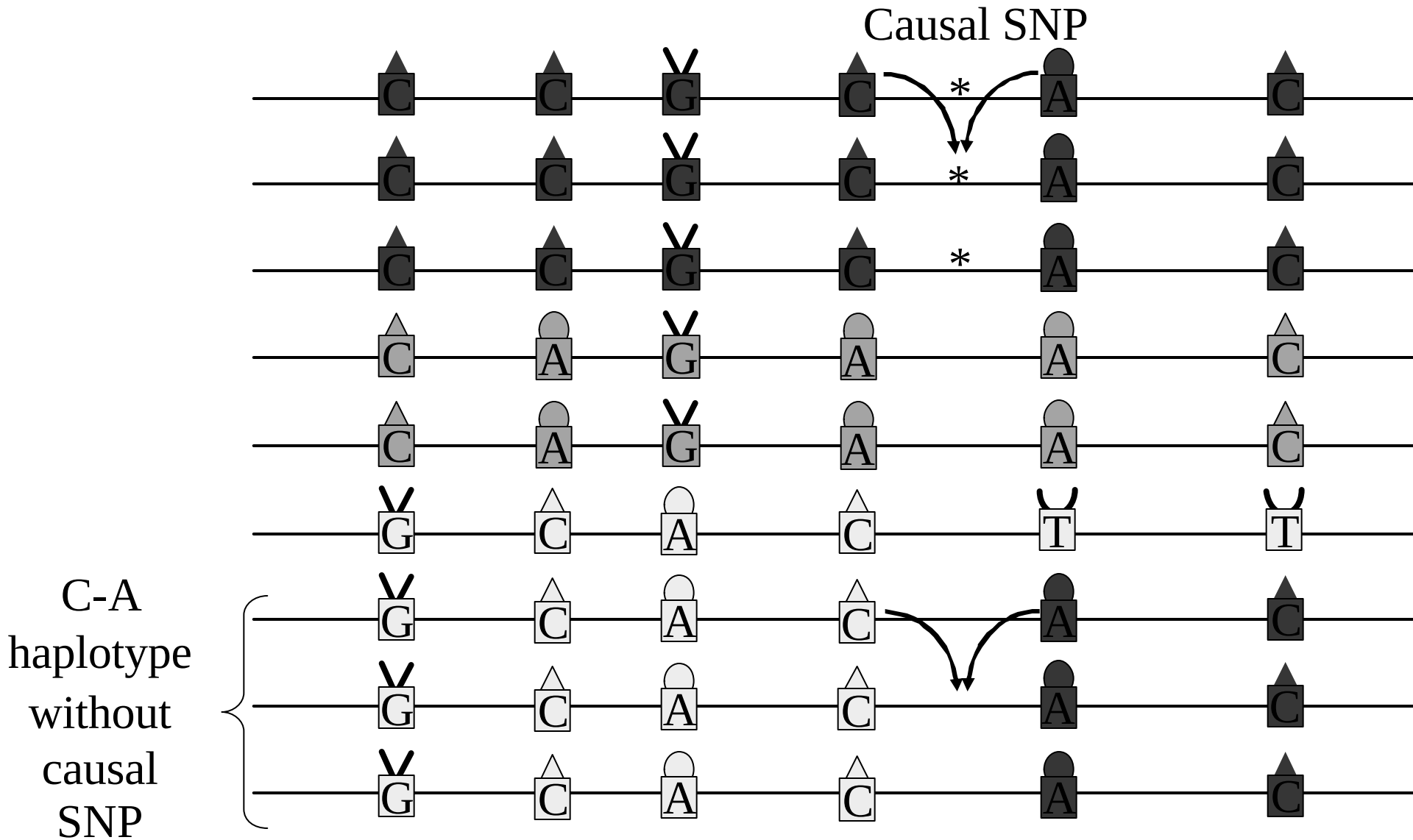
Most reported associations have not been consistently reproduced



Using linkage disequilibrium (LD) to detect unknown variants



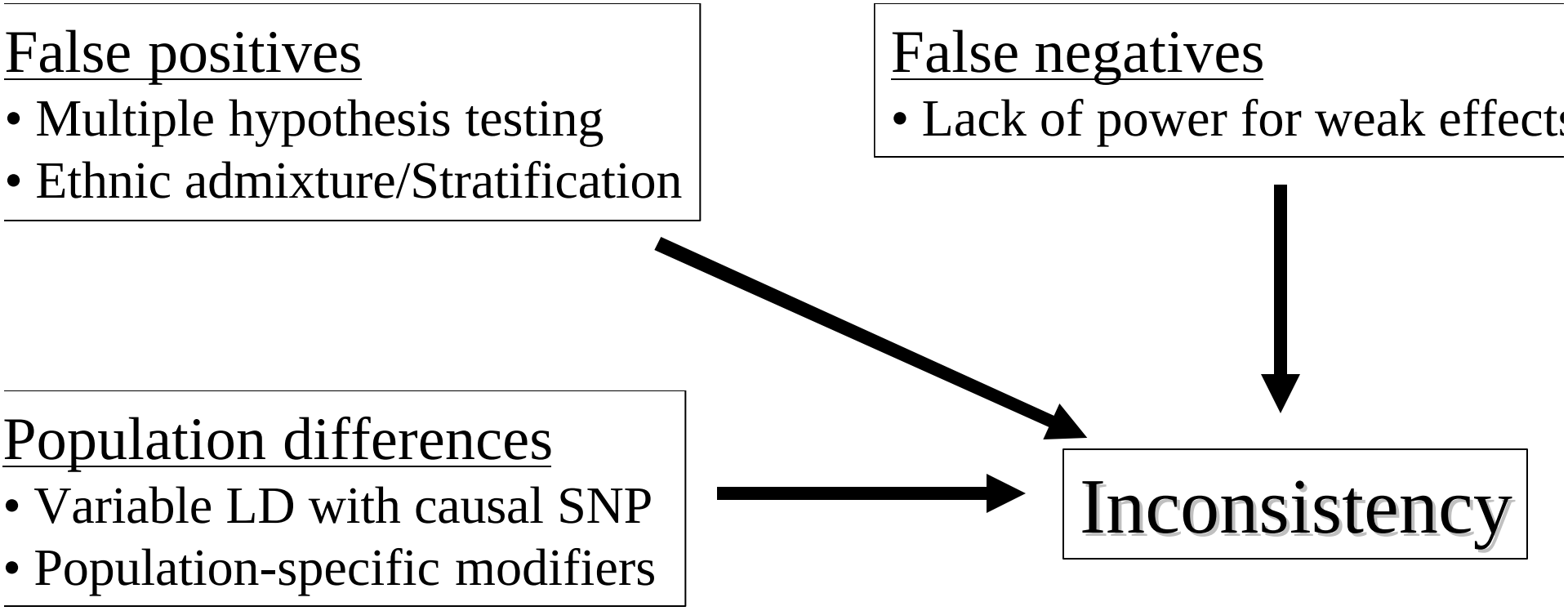
Different patterns of LD can yield different strength signals



Determining the LD patterns around associated SNPs may be critic

Association studies are powerful but problematic

Most reported associations have not been consistently reproduced



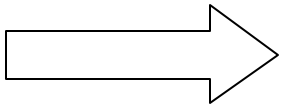
Modest effects and lack of power cause inconsistency

Diabetes

Cancer Epidemiology Biomarkers & Prevention

The American Journal of Human Genetics

Nature genetics



8/25 associations replicate

All eight increase risk by less than 2-fold

Pool all data for 25 associations

Lohmueller et al., Nature Genetics, 2003

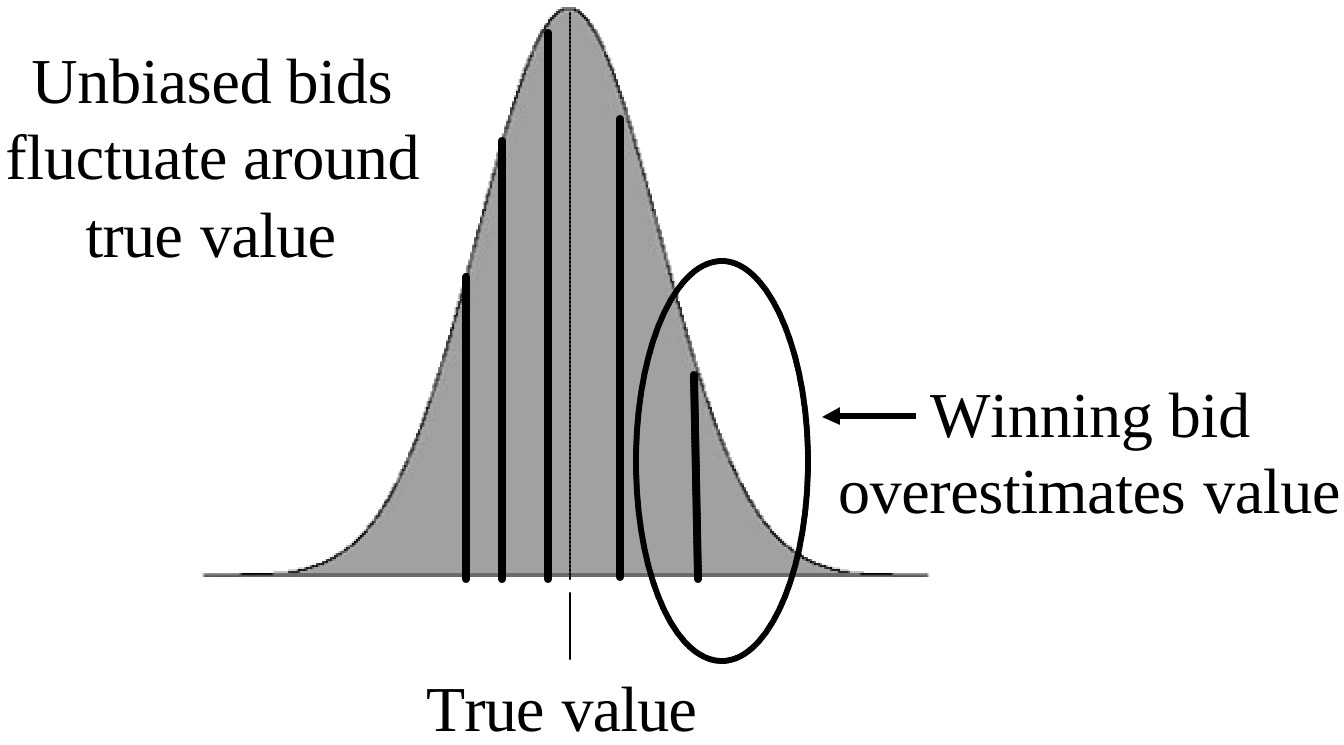
First positive reports are unreliable estimators

24/25 first positive reports overestimated the genetic effect

Consistent with “winner’s curse”?

“Winner’s curse”

Best described for auction theory



Winner's curse and association studies

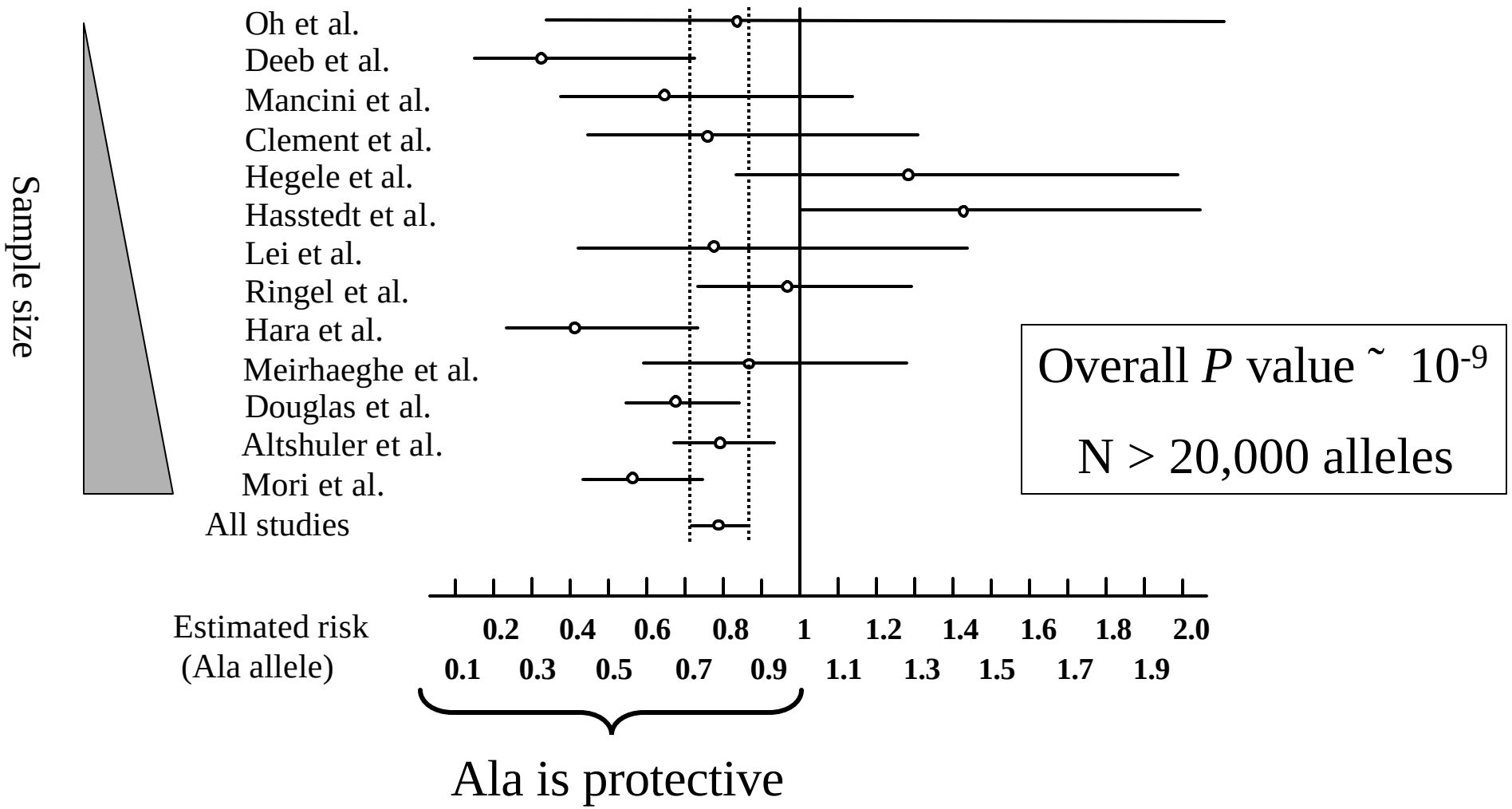
- In association studies, first positive report is equivalent to winning bid

- 23/25 associations consistent with winner's curse

Meta-analysis of association studies

- A sizable fraction (but less than half) of reported associations are likely correct
- Genetic effects are generally modest
  - Beware the winner's curse
- Large study sizes are needed to detect these reliably

Example: PPARγ Pro12Ala and diabetes



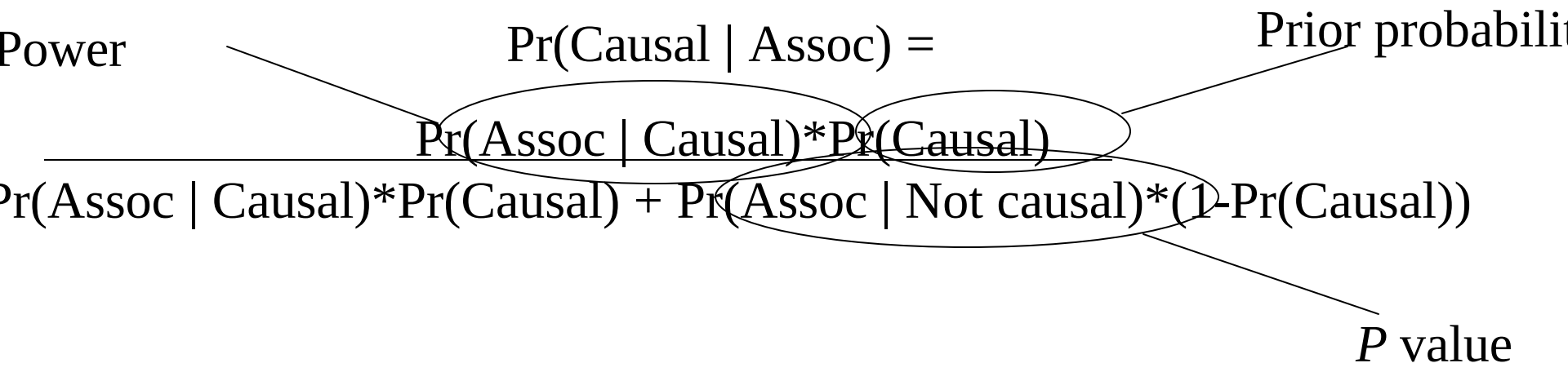
Should we believe association study results?

Initial skepticism is warranted

Replication, especially with low *p* values, is encouragin

Large sample sizes are crucial

Applying Bayes' theorem to association studies



Pr(Causal) = probability variant is causal

Pr(Assoc) = probability of observing an association

We observe associations, and we are interested in Pr (Causal | Assoc), which is the probability of the variant being causal given the data we observe

What are the prior probabilities?

- Random variants:
- About 600,000 independent common variants
- At least a few will be causal
- Prior probability = 1/10,000 - 1/100,000

What are the prior probabilities?

Candidate genes:

300 candidate genes * 12 independent variants/gene =
3,600 candidate variants

Assume half of all causal variants are in candidate genes

Prior probability = 1/100 - 1/1,000

What are the prior probabilities?

Positional candidate genes (linkage):

About 100 candidate genes * 12 variants/gene = 1,200 candidate variants

Only one causal gene

Prior probability = 1/1,000

Positional candidates (genes under linkage peaks)

are about as plausible as other candidate genes

Bayes' Theorem in action

Type of variant	Prior probability	*P* value	Posterior probability
Great candidate	0.01	0.05	0.14
Typical candidate	0.001	0.05	0.015
Positional candidate	0.001	0.05	0.015
Random gene	0.0001	0.05	0.0015

A single *P* value of 0.05 is probably, or nearly certainly, a false association

Bayes' Theorem in action

Type of variant	Prior probability	*P* value	Posterior probability
Great candidate	0.01	4×10^{-4}	0.95
Typical candidate	0.001	4×10^{-5}	0.95
Positional candidate	0.001	4×10^{-5}	0.95
Random gene	0.0001	4×10^{-6}	0.95

Low *P* values are required for higher degrees of certainty

Conclusions

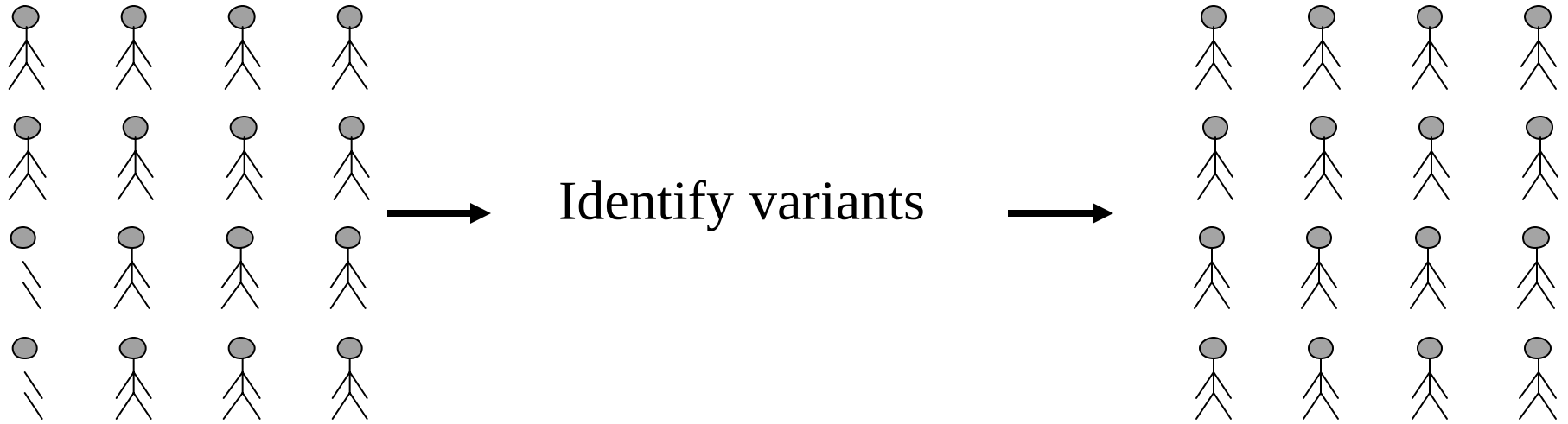
- Most reported associations are likely false
- Some will turn out to be correct
- Previous evidence of association is relevant if:
  - *P* values are low ($< 10^{-3}$ in the best case)
  - Associations are replicated, or
  - There is a very good reason for plausibility
- Genes under linkage peaks are more or less equivalent to other candidate genes

Similar issues arise in linkage studies

- Most regions of linkage not reproduced

- Why?
  - Population-specific differences
  - False positives (although this is better understood)
  - Lack of power and expected statistical variation

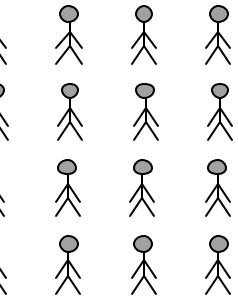
What about rare variant association studies?



Resequence gene
in affected individuals

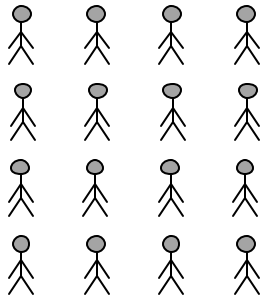
Genotype in unaffected
individuals

A possible resequencing association study



esequence gene X
in 200
abetic individuals

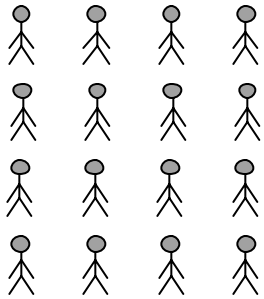
Identify variants:
10 rare missense variants



Type 200 healthy individu
Variants not seen at all

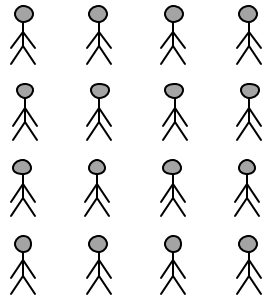
Rare missense variants in gene X cause diabetes!

A possible resequencing association study



Resequence gene X
in 200
diabetic individuals

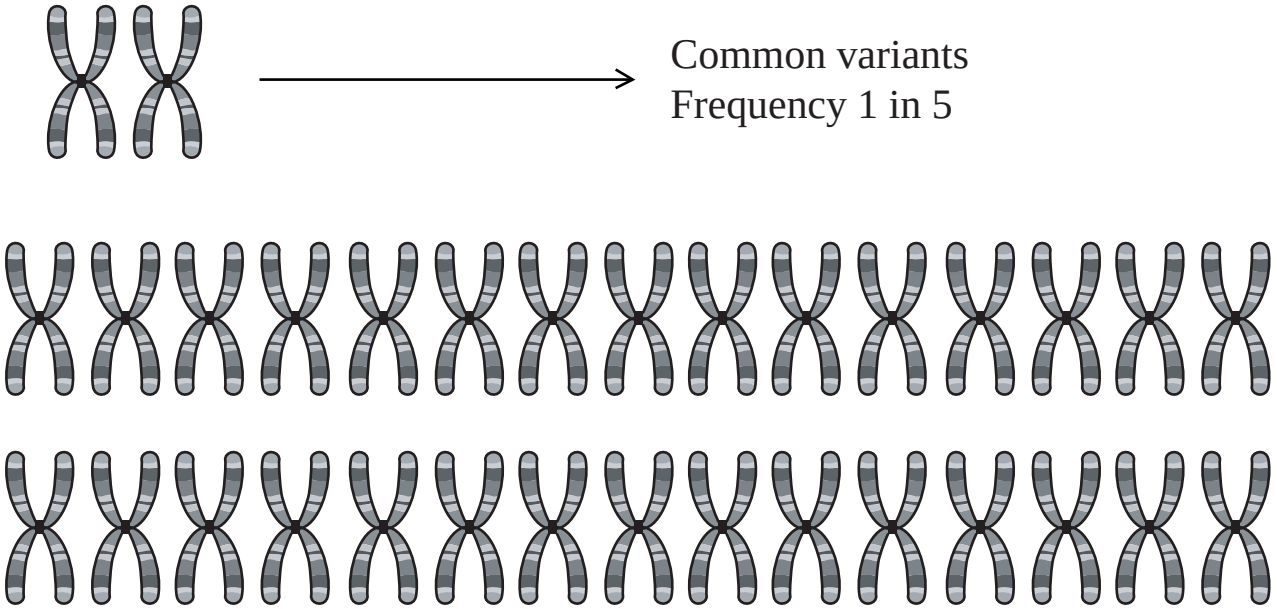
Identify variants:
10 rare missense variants



Type 200 healthy individual
Variants not seen at all!

Rare missense variants in gene X make you root for the Red Sox!

Expected allele frequency depends on depth of resequencing



Common variants
Frequency 1 in 5

Rare variants
Frequency 1/10,00

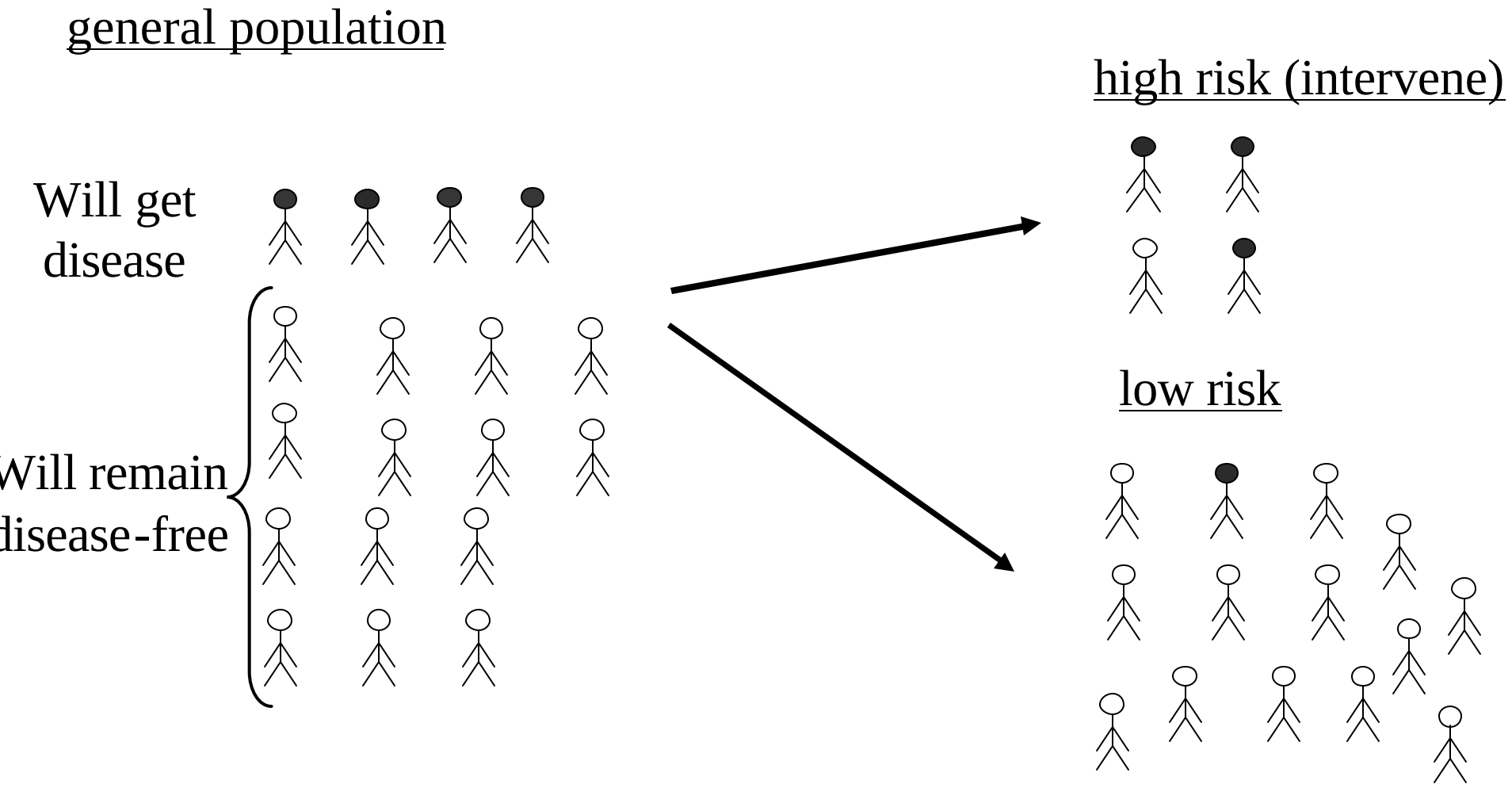
Don’t get fooled again…

- Controls must be resequenced with equal vigor!
- Rare variants must be grouped for analysis, BEFORE knowing the association study results

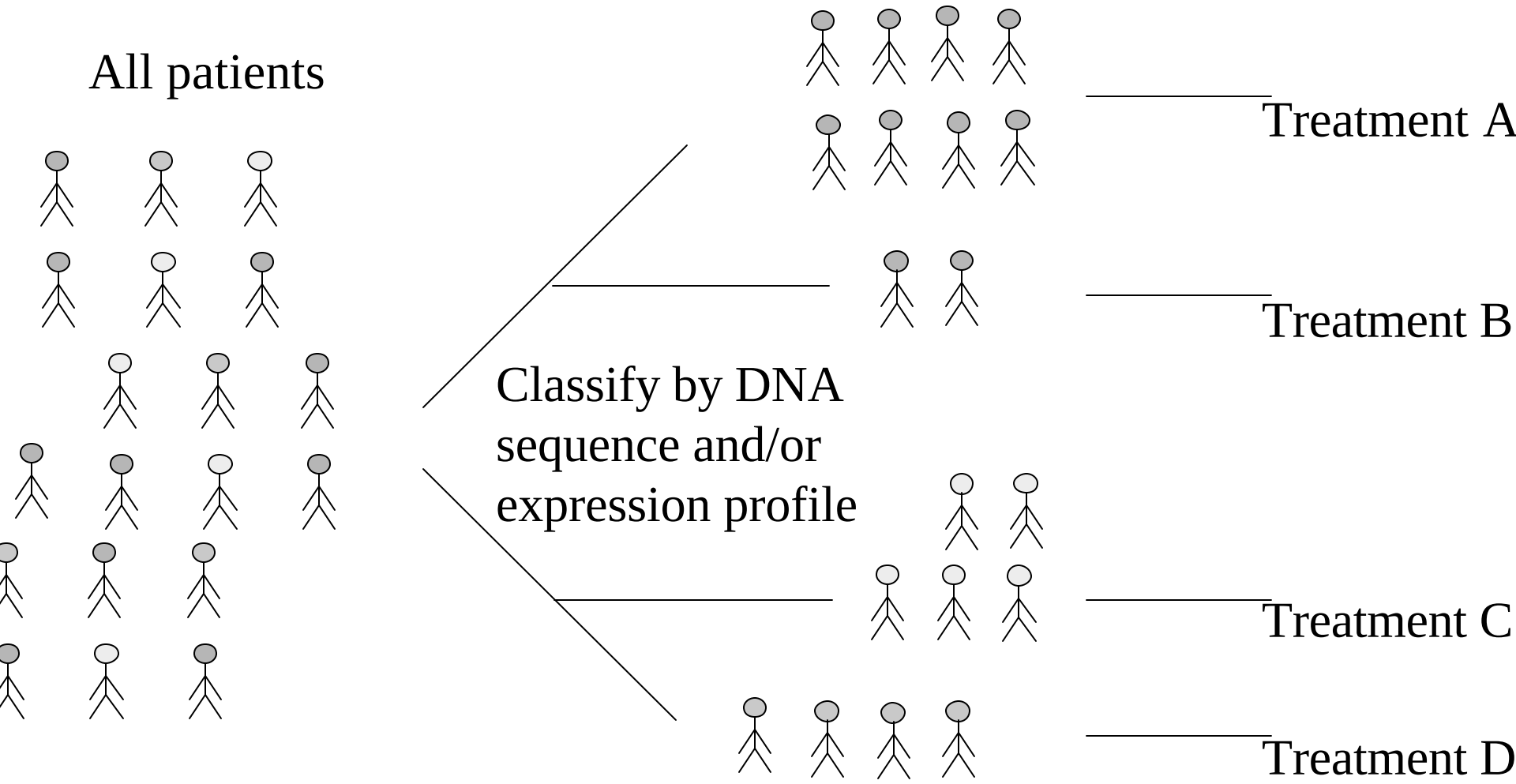
SNPs, patterns of variation, and complex traits

- Introduction
- Patterns of human genetic variation and disease
- Finding variants for complex traits
  - Linkage
  - Association
- Interpreting genetic studies
- What could we learn?

Prediction/Prevention

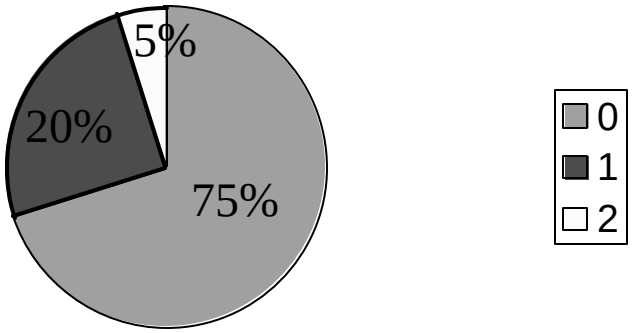


Reclassification to guide therapy

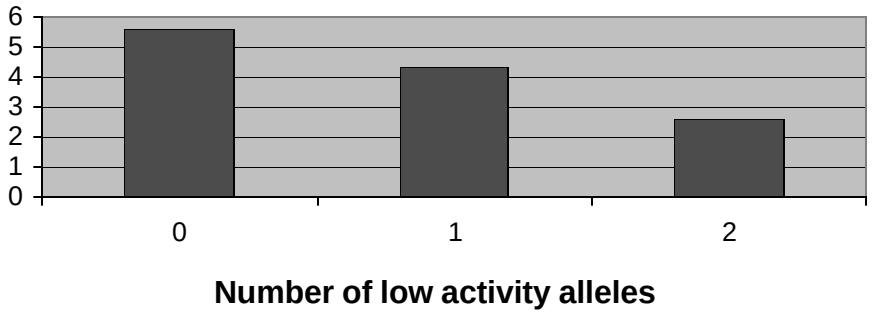


CYP2C9 and Warfarin

Prevalence of low activity alleles



Dosage and low activity alleles



Two common low activity alleles

2 alleles = 6x risk of serious complications

Higashi et al. JAMA 2002; Aithal et al. Lancet 1999

Genetic risk factors identify therapeutic targets

Sulfonylurea:
$K_{ir}6.2$ E23K

Thiazoladinedion
PPARγ P12A

Goal: Connect genotypic variation with phenotypic variation

Potential difficulties

- Privacy concerns
  - Insurance discrimination
- Improper interpretation of “predictive” information
  - Misguided interventions
  - Psychological impacts
- Impact on reproductive choices
- Interaction with concepts of race and ethnicity
- Genetics of performance

Acknowledgements

David Altshuler
Stacey Gabriel
Mark Daly
Steve Schaffner
Noel Burtt
Leif Groop
Cecilia Lindgren
Vamsi Mootha
Kirk Lohmueller
Leigh Pearce
Eric Lander

The SNP Consortium
The Human Genome Project
The Human Haplotype Map Project