Disease Association Studies

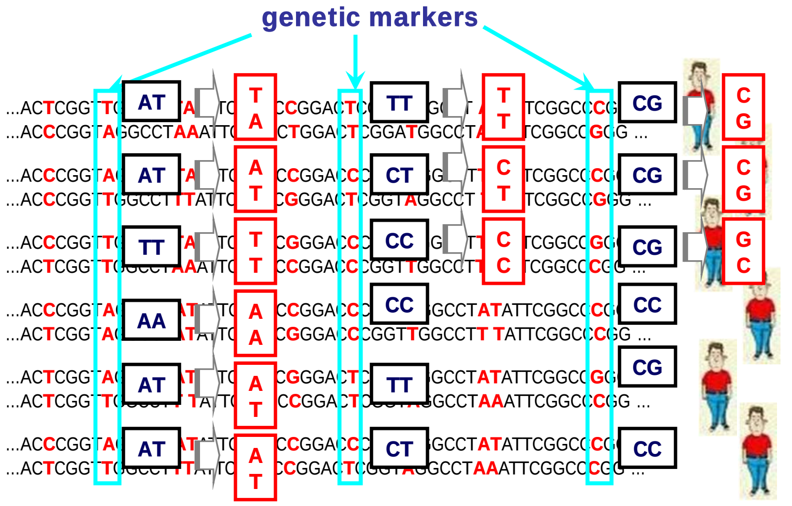
Lectures 7 – Oct 19, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

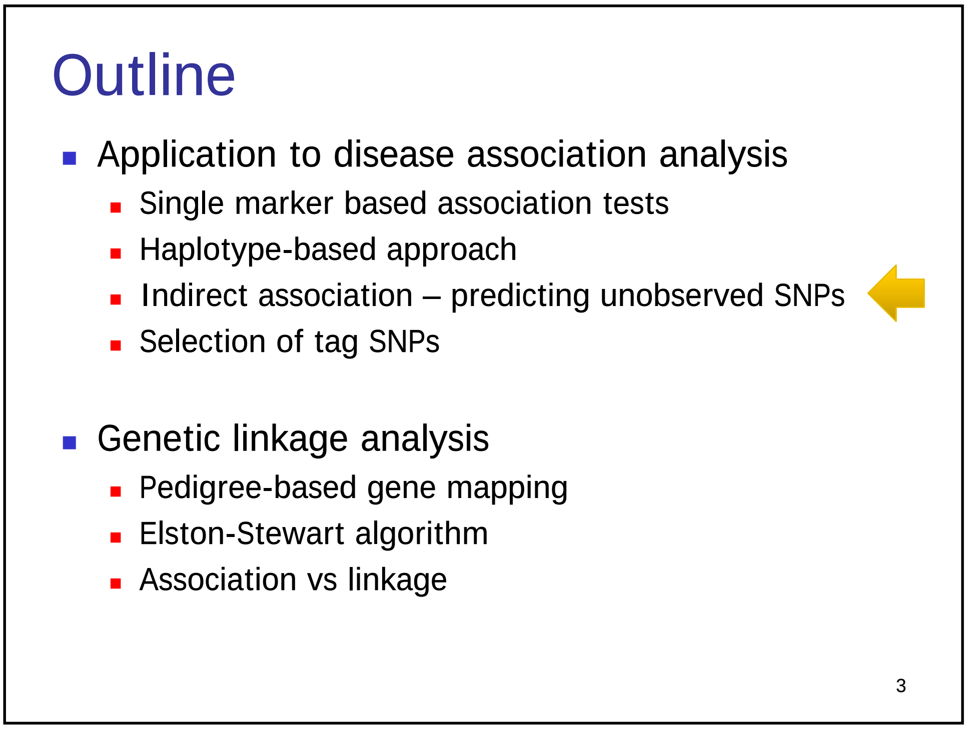
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Last Class ...

- Haplotype reconstruction

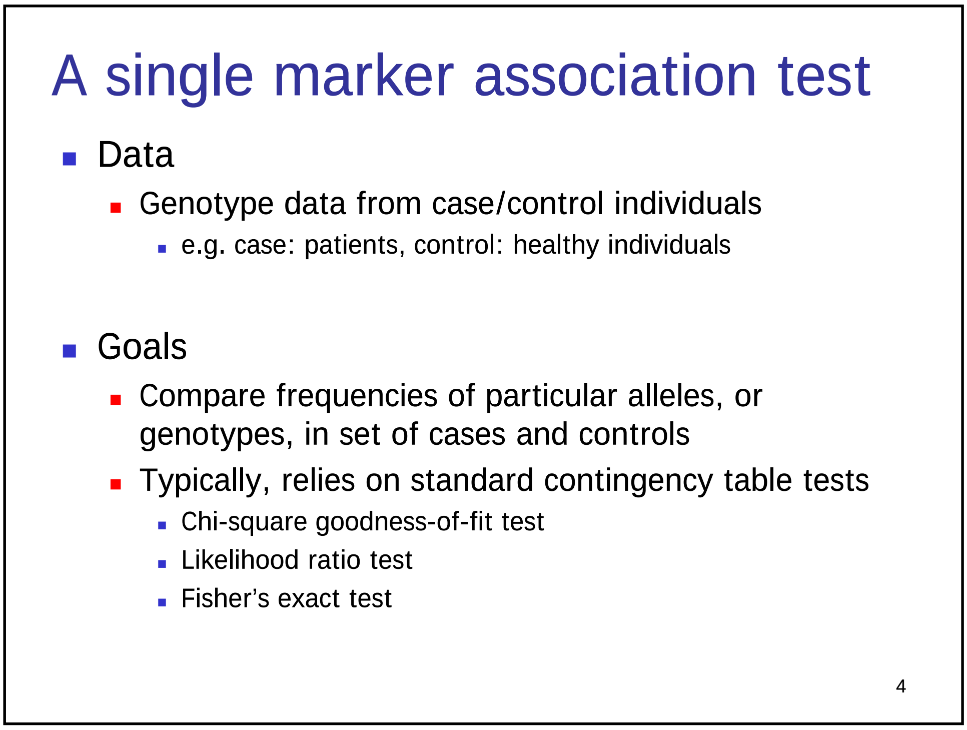


Single nucleotide polymorphism (SNP) [snip] = a variation at a single site in DNA



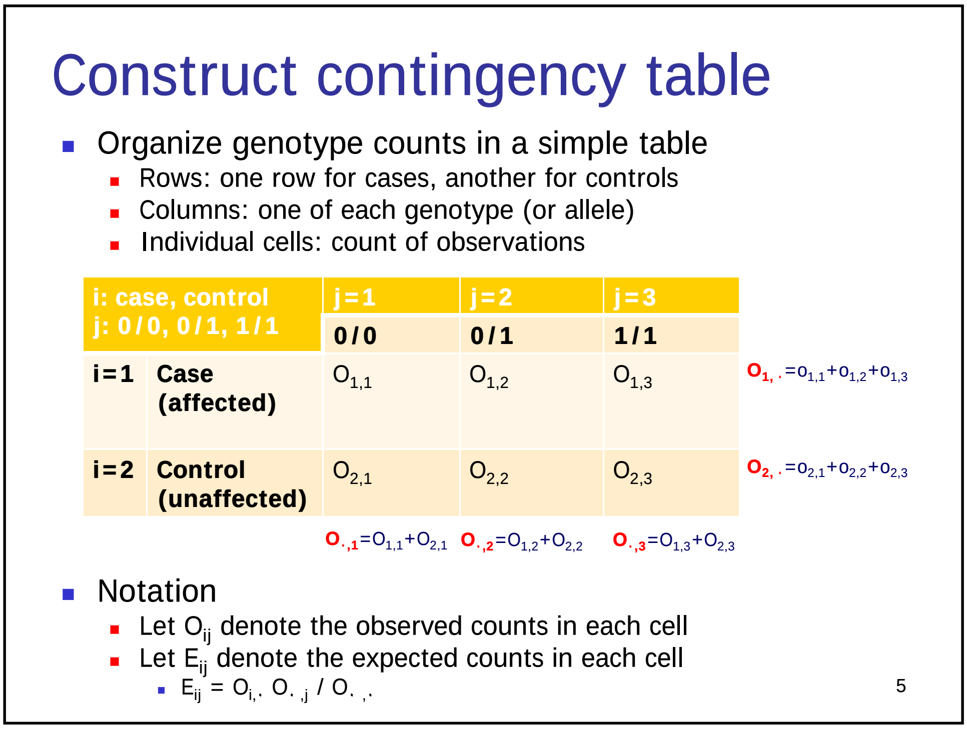
Outline

- Application to disease association analysis
  - Single marker based association tests
  - Haplotype-based approach
  - Indirect association – predicting unobserved SNPs
  - Selection of tag SNPs
- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm
  - Association vs linkage



A single marker association test

- Data
  - Genotype data from case/control individuals
    - e.g. case: patients, control: healthy individuals
- Goals
  - Compare frequencies of particular alleles, or genotypes, in set of cases and controls
  - Typically, relies on standard contingency table tests
    - Chi-square goodness-of-fit test
    - Likelihood ratio test
    - Fisher’s exact test

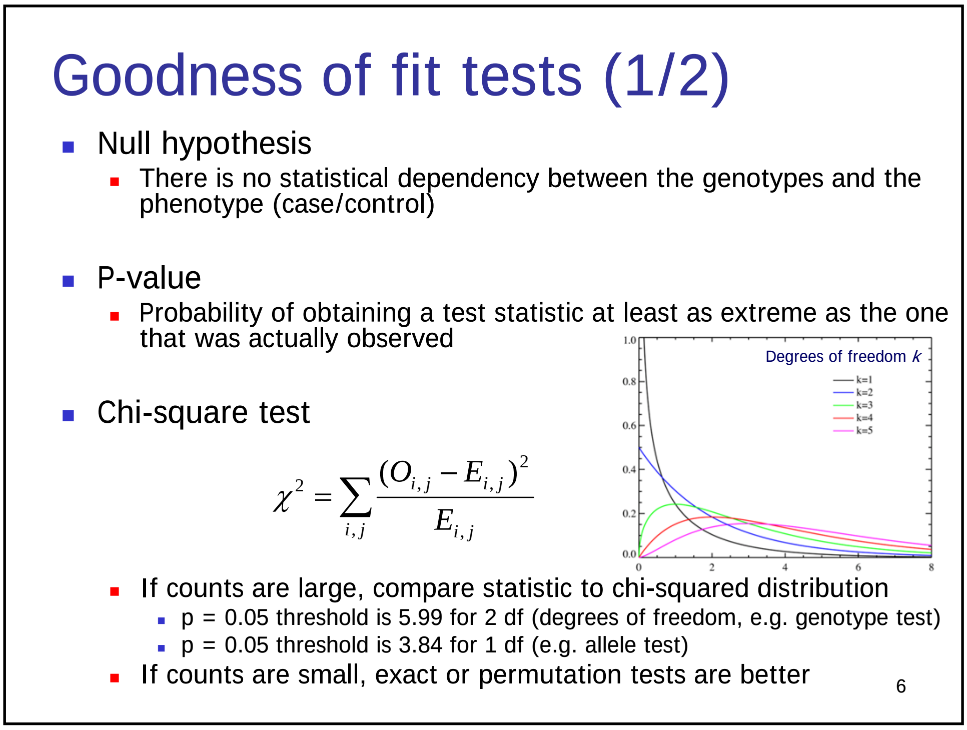


Construct contingency table

- Organize genotype counts in a simple table
  - Rows: one row for cases, another for controls
  - Columns: one of each genotype (or allele)
  - Individual cells: count of observations

i: case, control j: 0/0, 0/1, 1/1		j=1	j=2	j=3	
		0/0	0/1	1/1	
i=1	Case (affected)	$O_{1,1}$	$O_{1,2}$	$O_{1,3}$	$O_{1,.}=o_{1,1}+o_{1,2}+o_{1,3}$
i=2	Control (unaffected)	$O_{2,1}$	$O_{2,2}$	$O_{2,3}$	$O_{2,.}=o_{2,1}+o_{2,2}+o_{2,3}$
		$O_{.,1}=O_{1,1}+O_{2,1}$	$O_{.,2}=O_{1,2}+O_{2,2}$	$O_{.,3}=O_{1,3}+O_{2,3}$	

- Notation
  - Let O_{ij} denote the observed counts in each cell
  - Let E_{ij} denote the expected counts in each cell
    - $E_{ij} = O_{i,.}\ O_{.,j} \,/\, O_{.,.}$



Goodness of fit tests (1/2)

- Null hypothesis
  - There is no statistical dependency between the genotypes and the phenotype (case/control)
- P-value
  - Probability of obtaining a test statistic at least as extreme as the one that was actually observed
- Chi-square test

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Degrees of freedom *k*
k=1
k=2
k=3
k=4
k=5
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8

  - If counts are large, compare statistic to chi-squared distribution
    - p = 0.05 threshold is 5.99 for 2 df (degrees of freedom, e.g. genotype test)
    - p = 0.05 threshold is 3.84 for 1 df (e.g. allele test)
  - If counts are small, exact or permutation tests are better

Goodness of fit tests (2/2)

- Likelihood ratio test
  - The test statistics (usually denoted D) is twice the difference in the log-likelihoods:

$$D = -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right)$$

$$= -2\ln\frac{\prod_{i,j}\left(E_{i,j}/O\right)^{O_{i,j}}}{\prod_{i,j}\left(O_{i,j}/O\right)^{O_{i,j}}} = 2\sum_{i,j} O_{i,j}\ln\frac{O_{i,j}}{E_{i,j}}$$

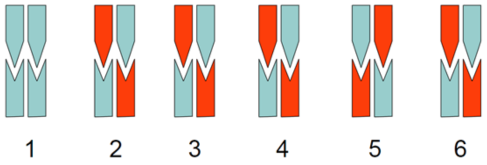
- How about we do this for haplotypes?
  - When does it out-perform the single marker association test?

Haplotype association tests

- Calculate haplotype frequencies in each group
- Find most likely haplotype for each group
- Fill in contingency table to compare haplotypes in the two groups (case, control)
- Not recommended!

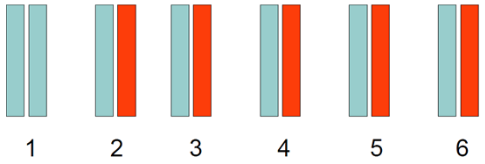
Case genotypes & haplotypes

- Observed case genotypes



  - The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1 …

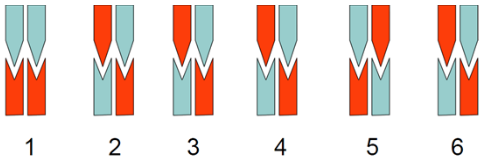
- Inferred case haplotypes



  - This kind of phenomenon will occur with nearly all population based haplotyping methods!

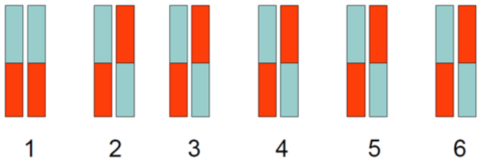
Control genotypes & haplotypes

- Observed control genotypes

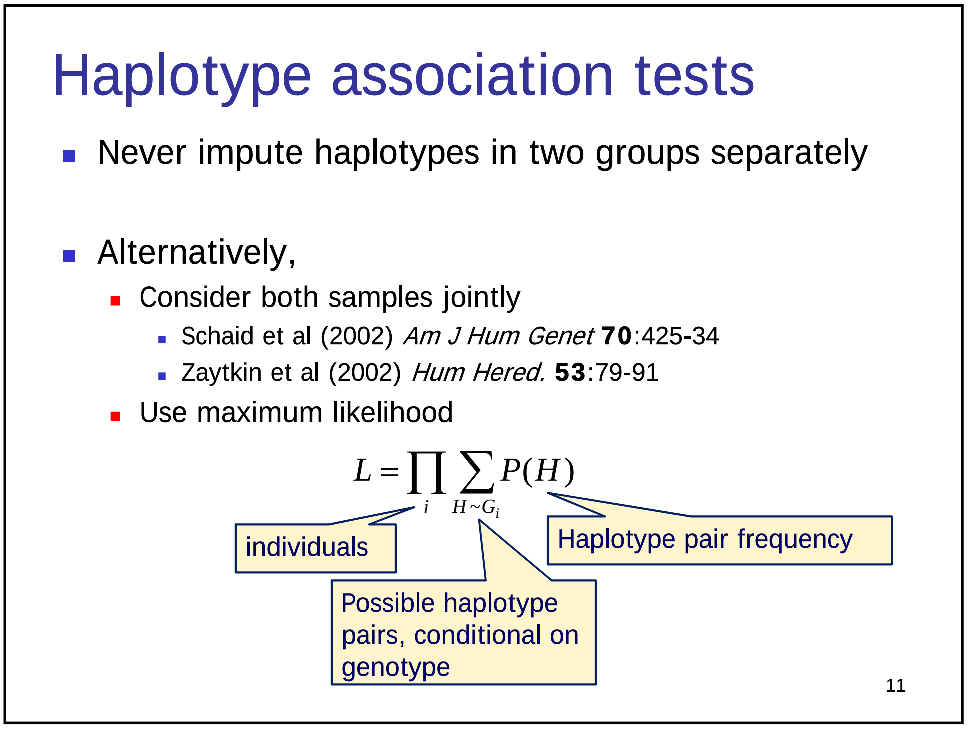


  - Note these are identical, except for the single homozygous individual …

- Inferred case haplotypes



  - Oops… The difference in a single genotype in the original data has been greatly amplified by estimating haplotypes…



Haplotype association tests

- Never impute haplotypes in two groups separately

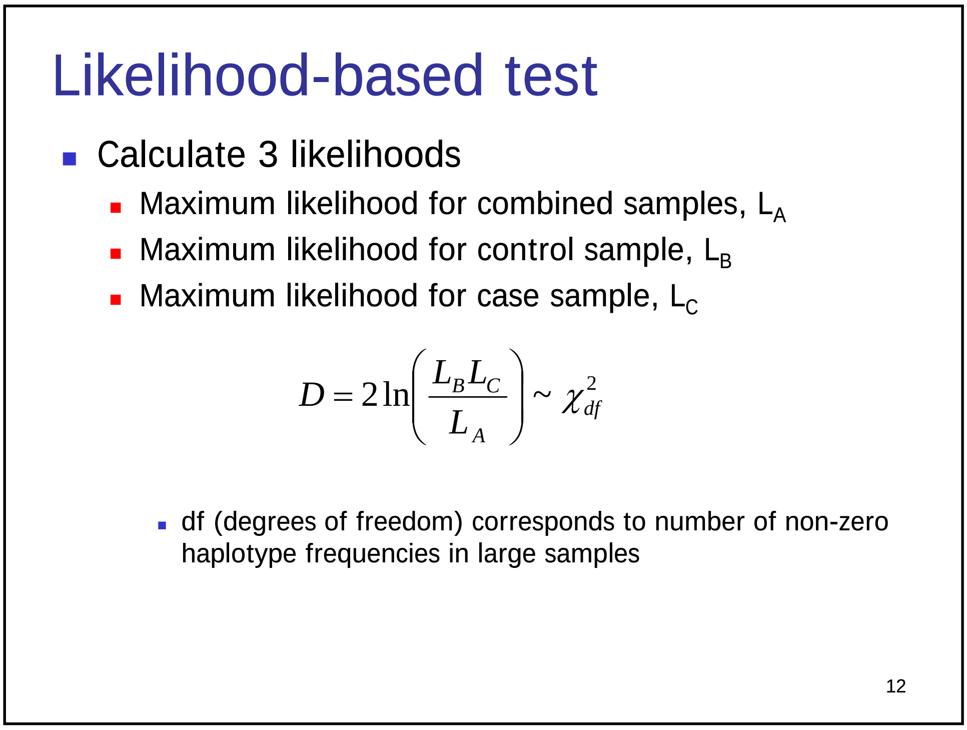
- Alternatively,
  - Consider both samples jointly
    - Schaid et al (2002) *Am J Hum Genet* **70**:425-34
    - Zaytkin et al (2002) *Hum Hered.* **53**:79-91
  - Use maximum likelihood

$$L = \prod_i \sum_{H \sim G_i} P(H)$$

individuals

Possible haplotype pairs, conditional on genotype

Haplotype pair frequency



Likelihood-based test

- Calculate 3 likelihoods
  - Maximum likelihood for combined samples, L_A
  - Maximum likelihood for control sample, L_B
  - Maximum likelihood for case sample, L_C

$$D = 2\ln\left(\frac{L_B L_C}{L_A}\right) \sim \chi^2_{df}$$

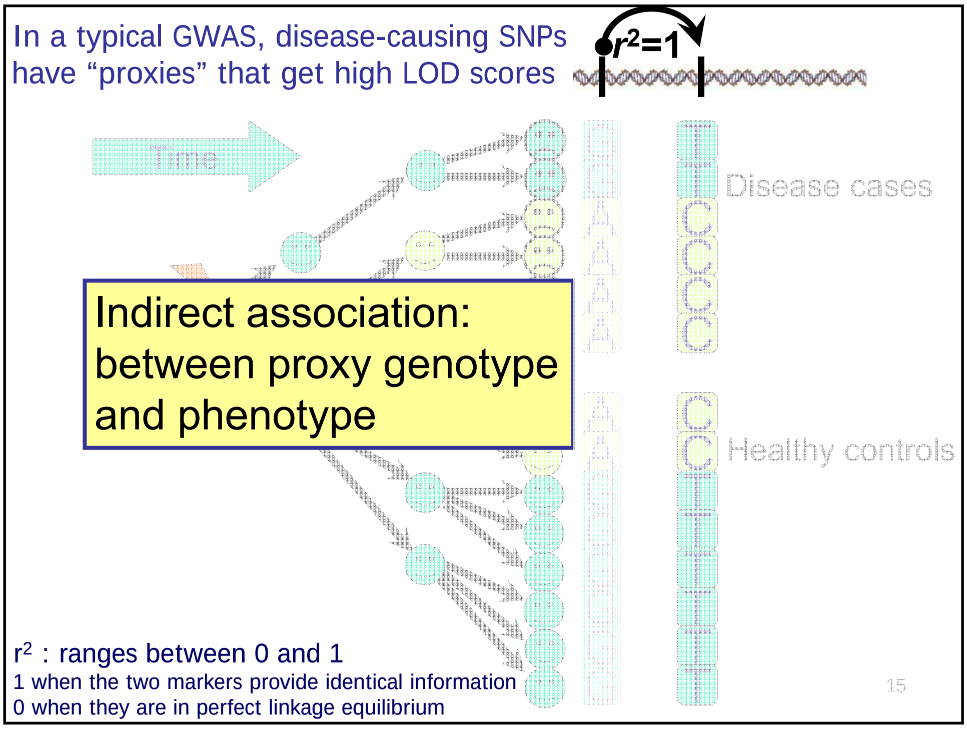
- df (degrees of freedom) corresponds to number of non-zero haplotype frequencies in large samples

Significance in small samples

- In reality sample sizes, it is hard to estimate the number of *df* accurately
- Instead, use a permutation approach to calculate empirical significance levels
- How?

Outline

- Application to disease association analysis
  - Single marker based association tests
  - Haplotype-based approach
  - Indirect association – predicting unobserved SNPs ⇦
  - Selection of tag SNPs
- Genetic linkage analysis
  - Pedigree-based gene mapping
  - Elston-Stewart algorithm
  - Association vs linkage



In a typical GWAS, disease-causing SNPs have “proxies” that get high LOD scores

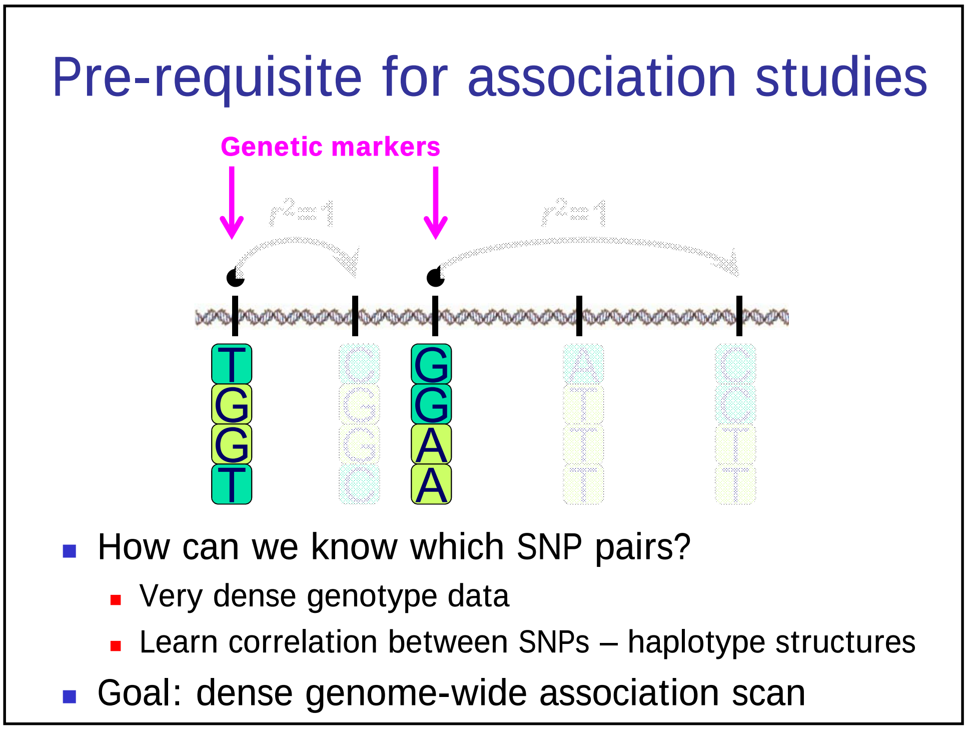
Indirect association: between proxy genotype and phenotype

r^2 : ranges between 0 and 1

1 when the two markers provide identical information

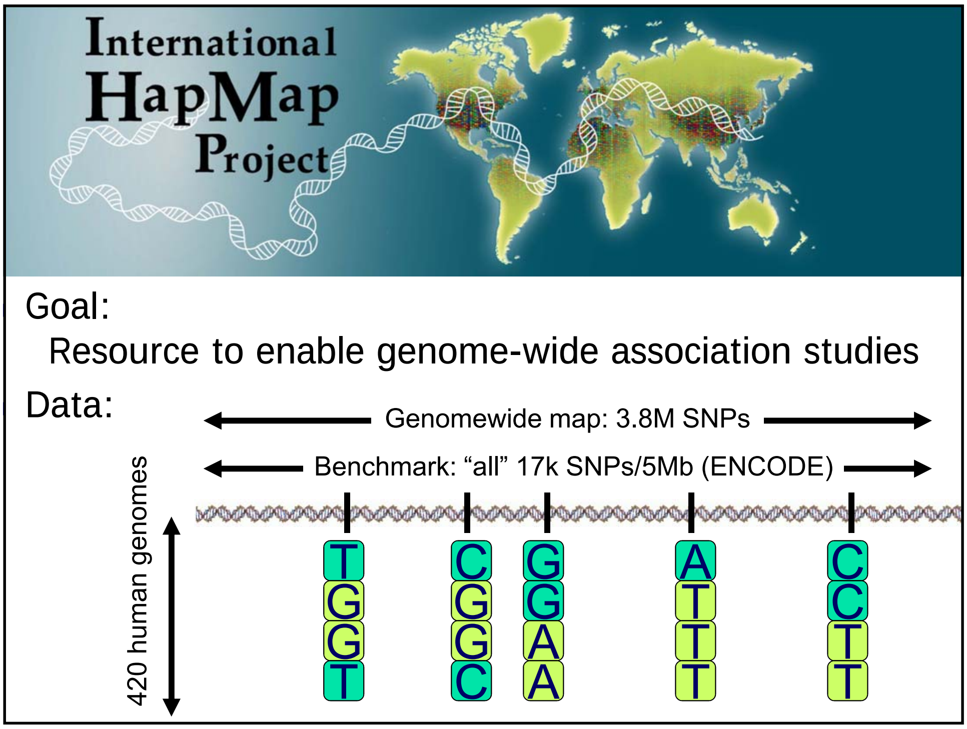
0 when they are in perfect linkage equilibrium

15



Pre-requisite for association studies

- How can we know which SNP pairs?
  - Very dense genotype data
  - Learn correlation between SNPs – haplotype structures
- Goal: dense genome-wide association scan



International HapMap Project

Goal:

Resource to enable genome-wide association studies

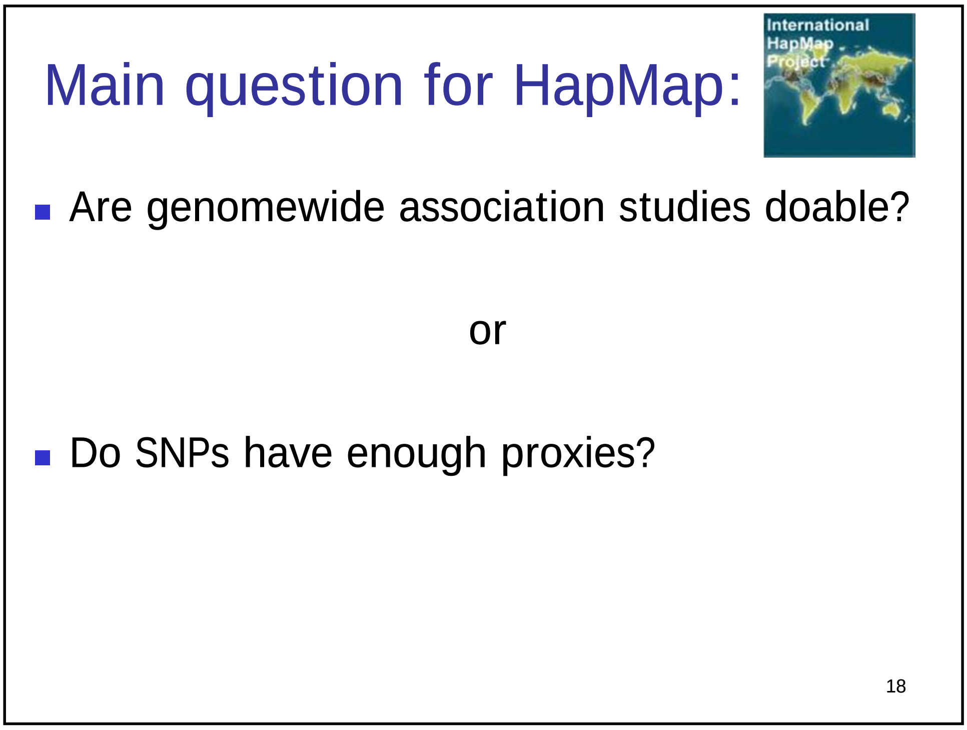
Data:

← Genomewide map: 3.8M SNPs →

← Benchmark: “all” 17k SNPs/5Mb (ENCODE) →

↕ 420 human genomes

T	C	G	A	C
G	G	G	T	C
G	G	A	T	T
T	C	A	T	T



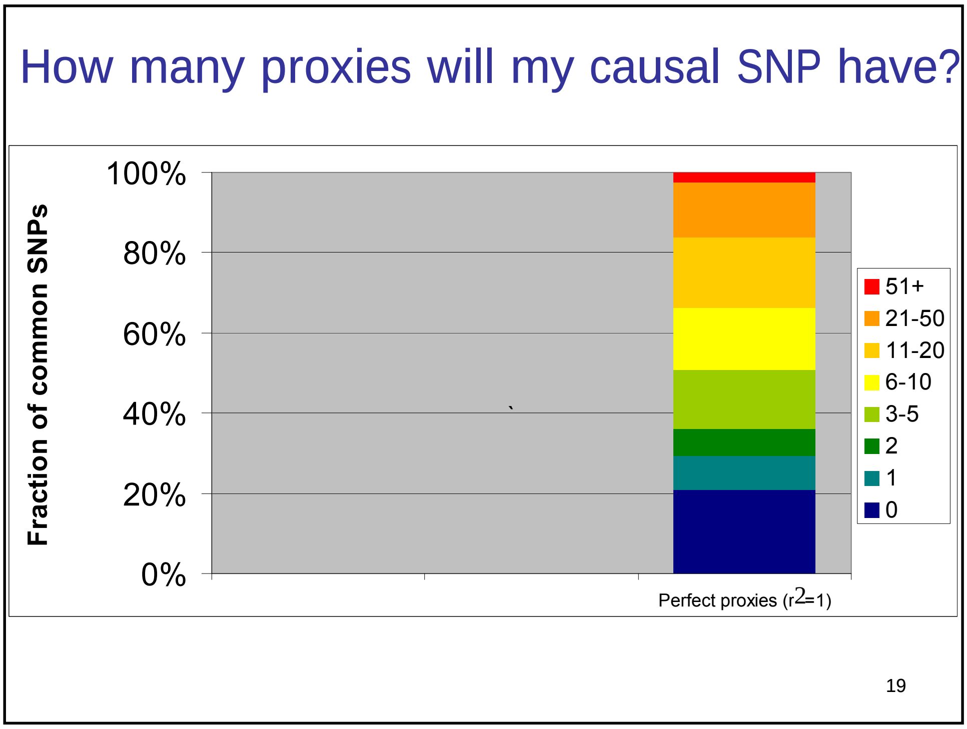
Main question for HapMap:

International
HapMap
Project

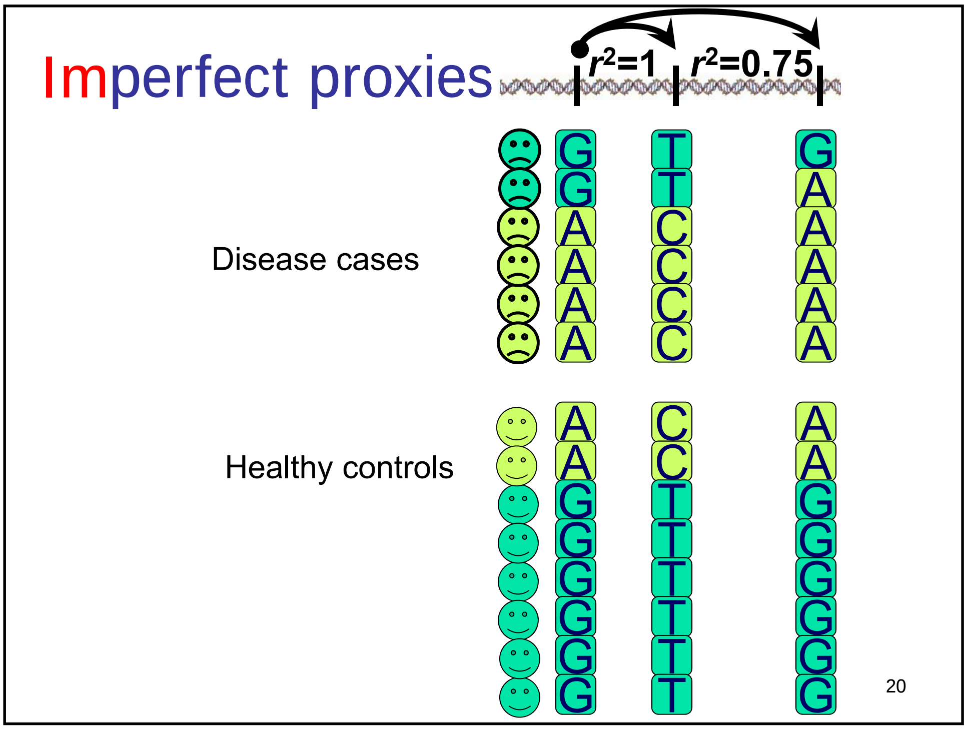
- Are genomewide association studies doable?

or

- Do SNPs have enough proxies?

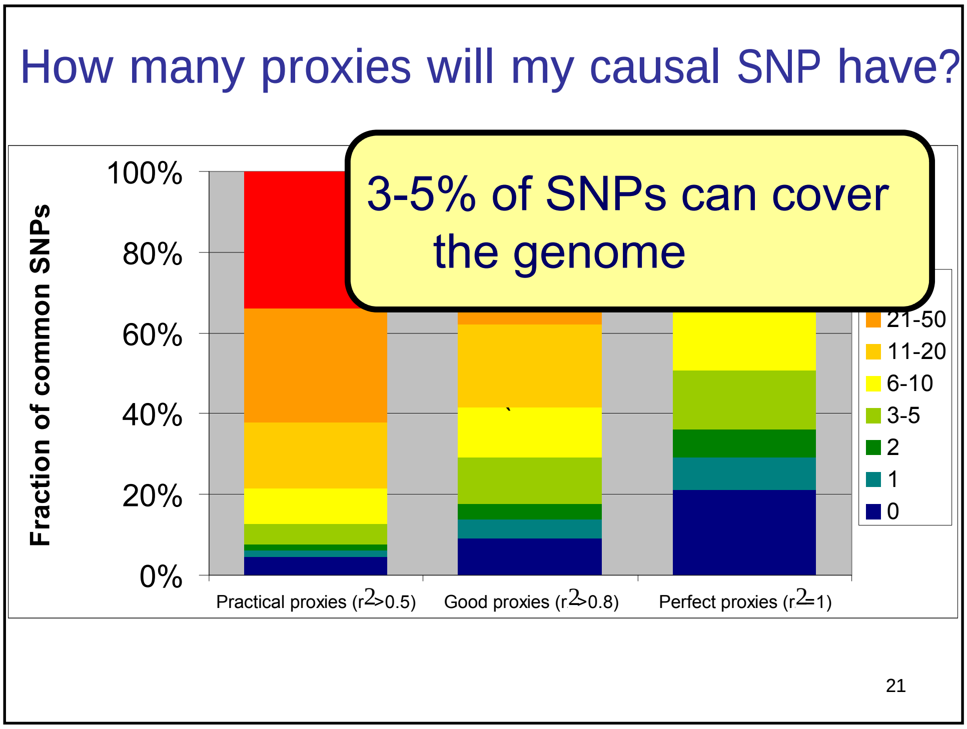


How many proxies will my causal SNP have?



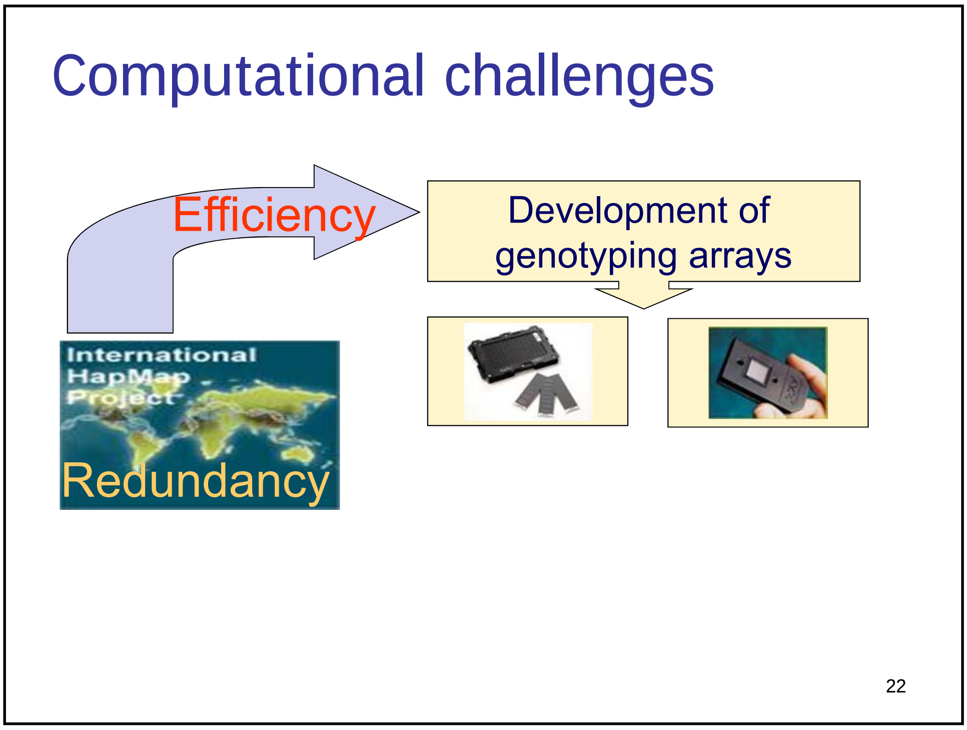
Imperfect proxies

Disease cases	G	T	G
	G	T	A
	A	C	A
	A	C	A
	A	C	A
	A	C	A
Healthy controls	A	C	A
	A	C	A
	G	T	G
	G	T	G
	G	T	G
	G	T	G
	G	T	G
	G	T	G



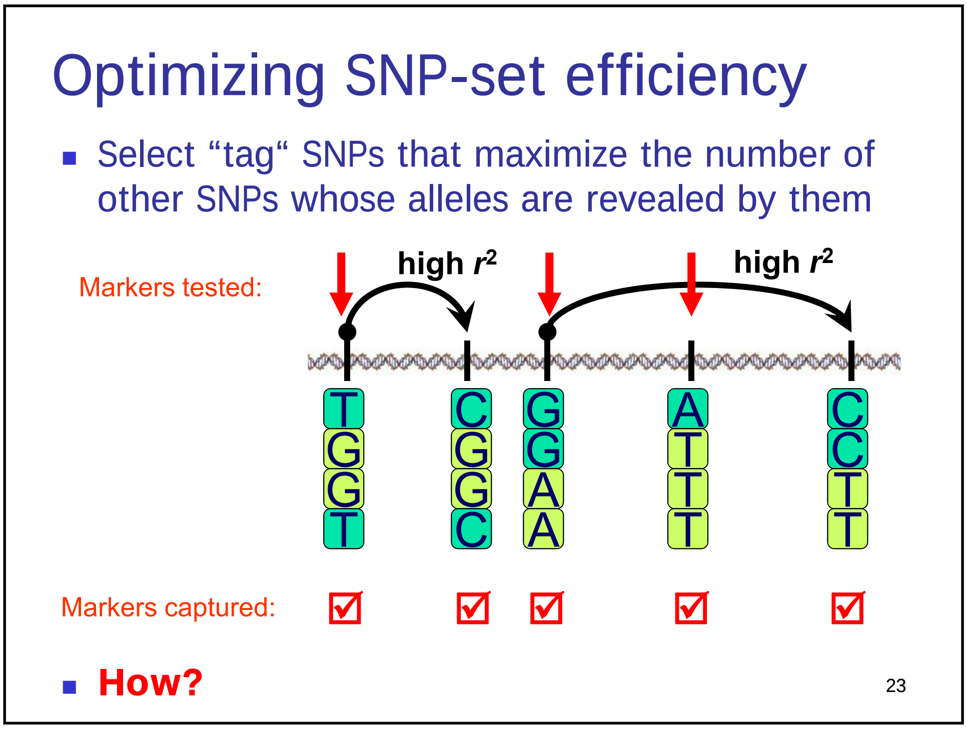
How many proxies will my causal SNP have?

3-5% of SNPs can cover the genome



Computational challenges

Development of genotyping arrays



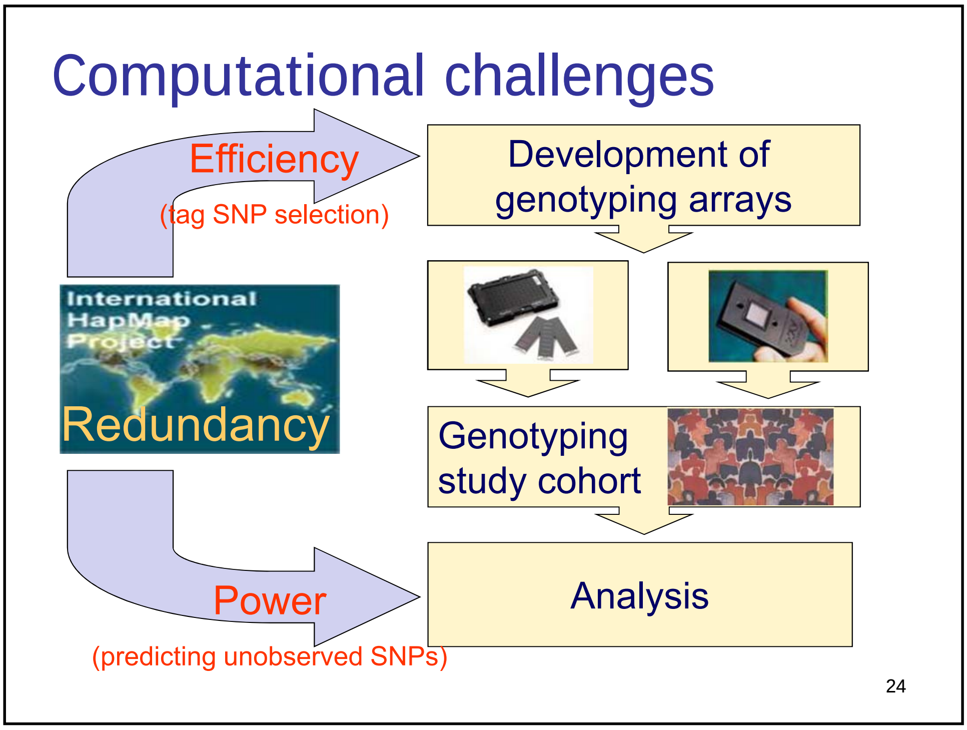
Optimizing SNP-set efficiency

- Select “tag“ SNPs that maximize the number of other SNPs whose alleles are revealed by them

T	C	G	A	C
G	G	G	T	C
G	G	A	T	T
T	C	A	T	T

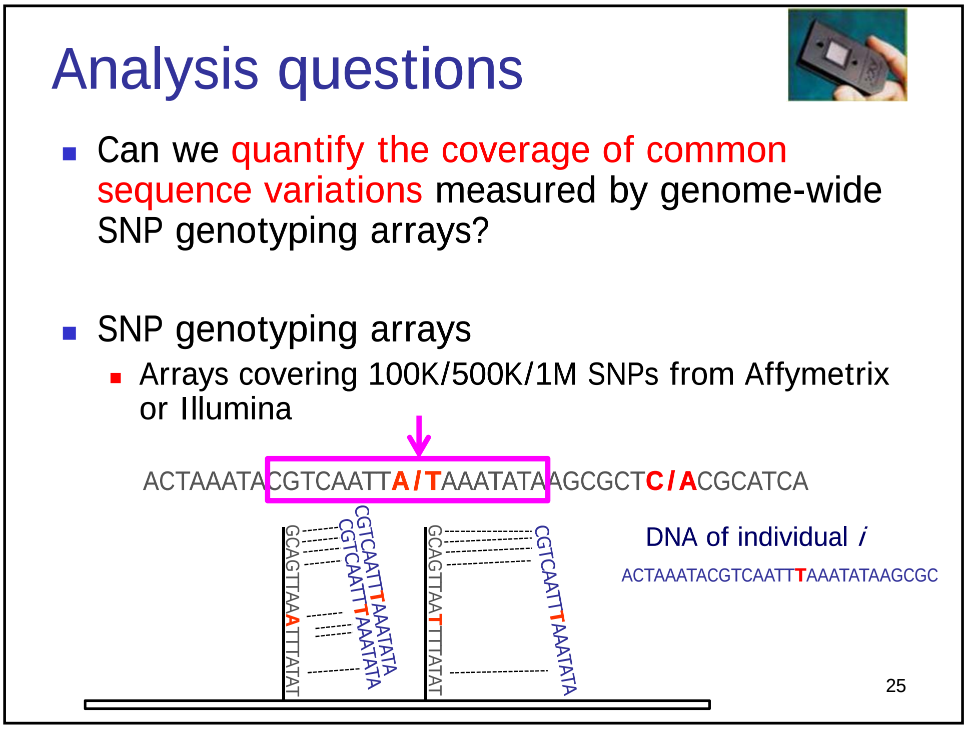
Markers captured: ☑ ☑ ☑ ☑ ☑

- **How?**



Computational challenges

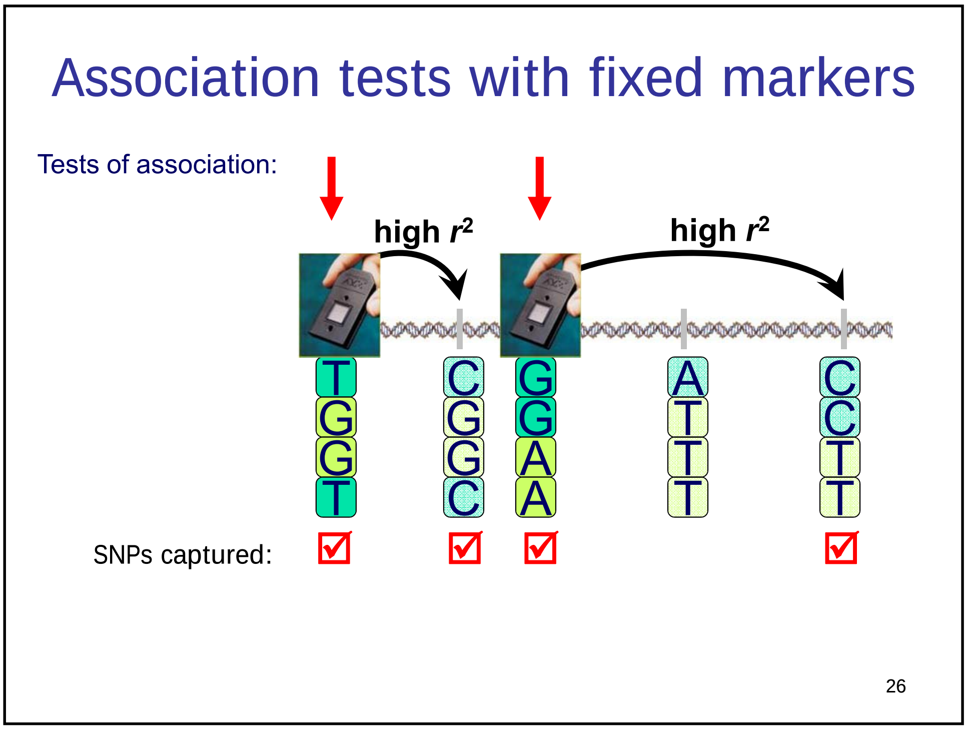
Efficiency
(tag SNP selection)
Development of genotyping arrays
International HapMap Project
Redundancy
Genotyping study cohort
Power
(predicting unobserved SNPs)
Analysis



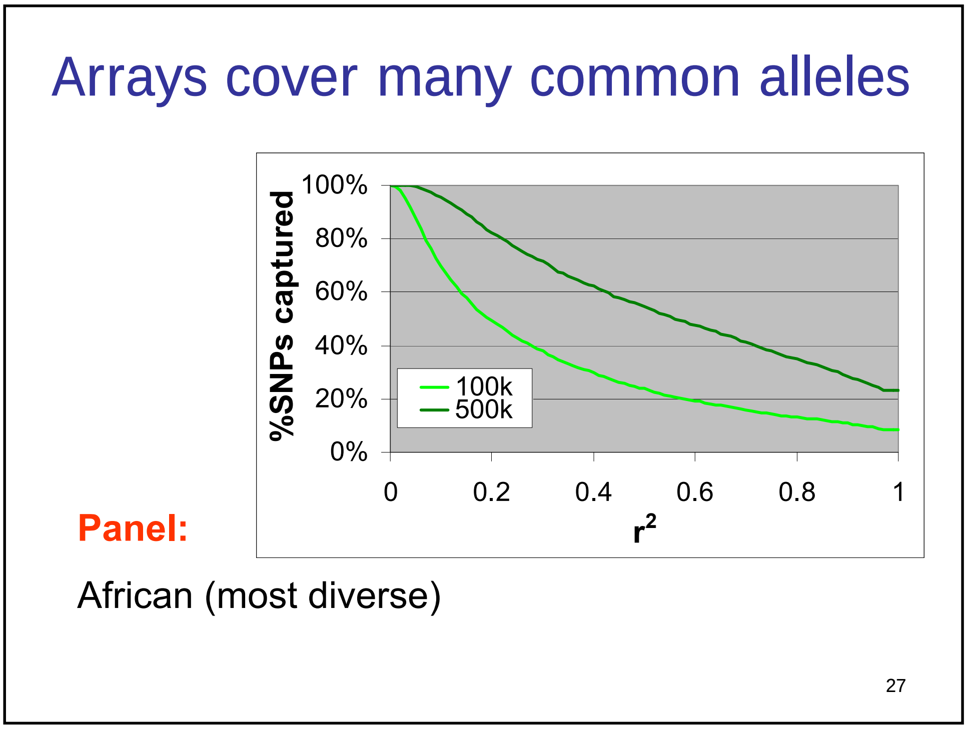
Analysis questions

- Can we quantify the coverage of common sequence variations measured by genome-wide SNP genotyping arrays?

- SNP genotyping arrays
  - Arrays covering 100K/500K/1M SNPs from Affymetrix or Illumina



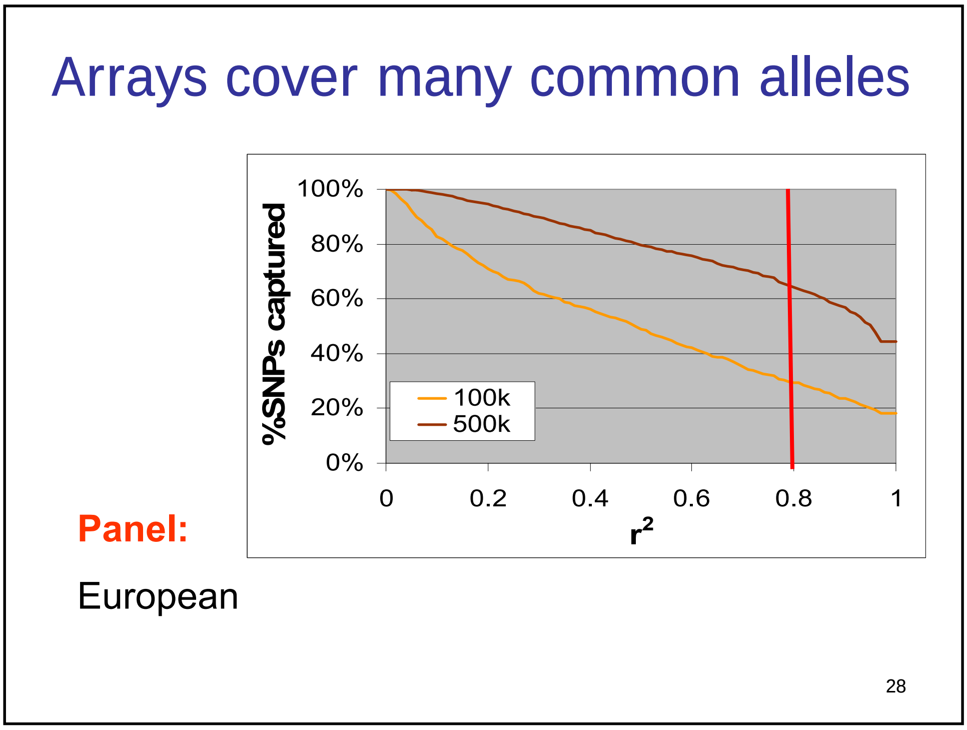
Association tests with fixed markers



Arrays cover many common alleles

Panel:

African (most diverse)



Arrays cover many common alleles

Panel:

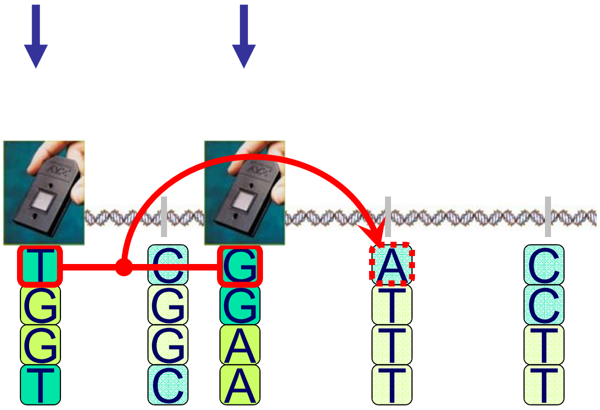
European

Analysis questions

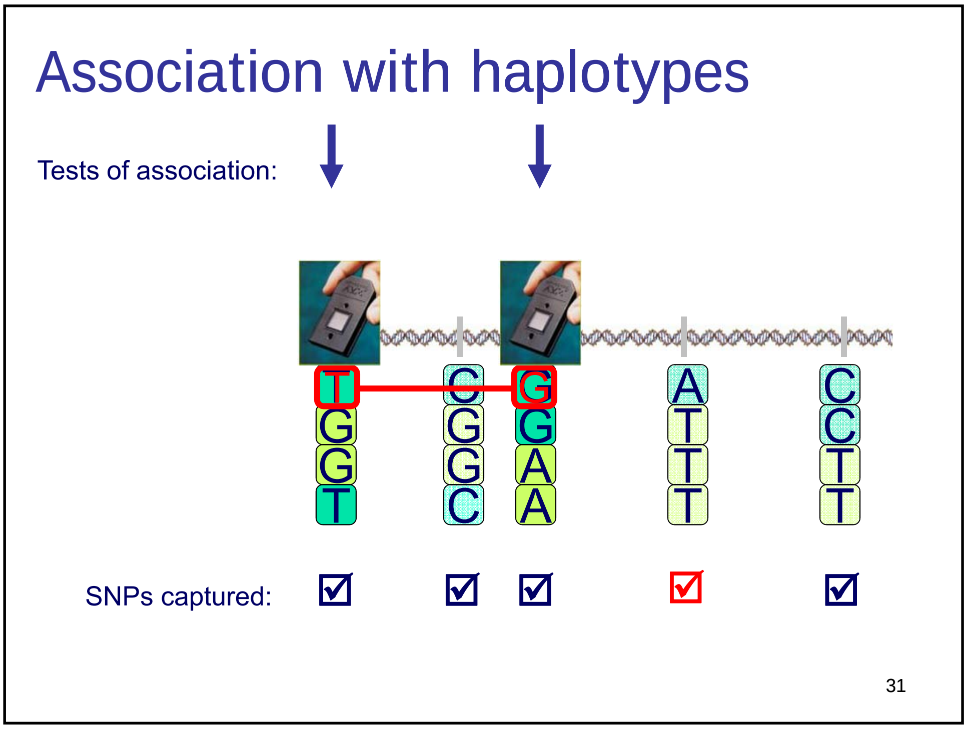
- Can we quantify the coverage of common sequence variations measured by genome-wide SNP genotyping arrays?
- Can we do better?

Association with haplotypes

Tests of association:



SNPs captured: ☑ ☑ ☑ ☑

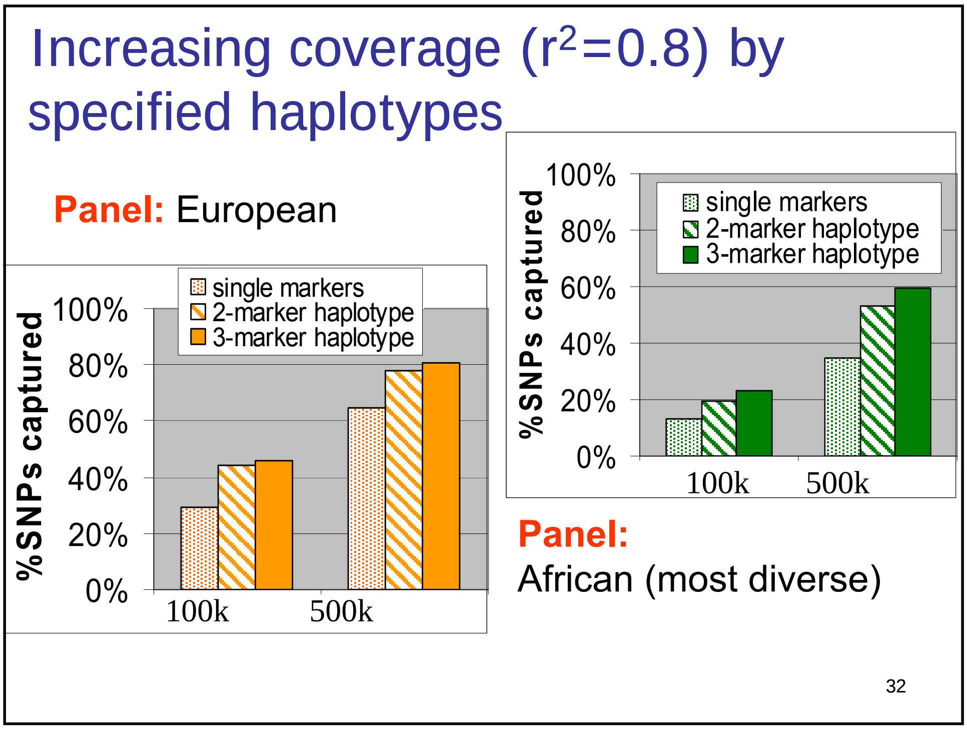


Association with haplotypes

Tests of association:

SNPs captured: ☑ ☑ ☑ ☑ ☑

31

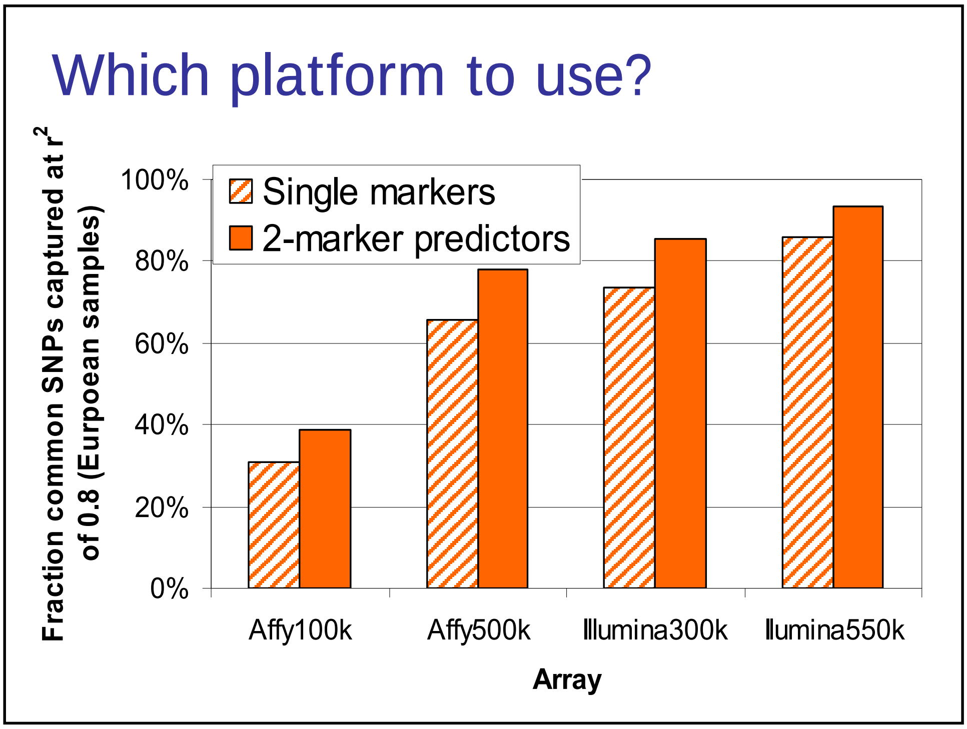


Increasing coverage ($r^2=0.8$) by specified haplotypes

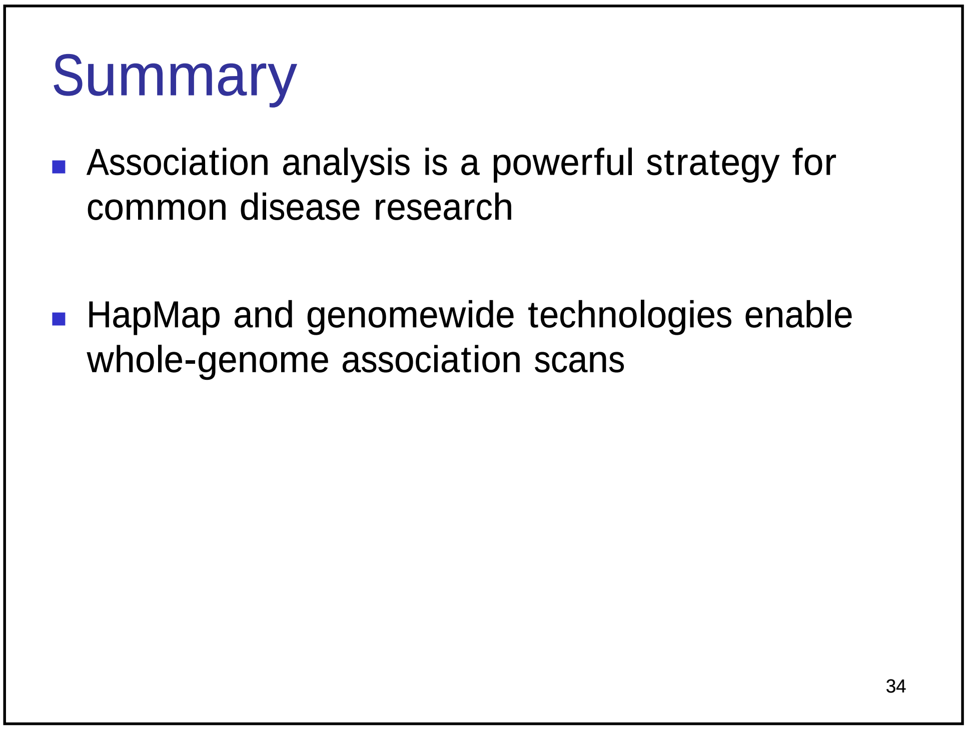
Panel: European

Panel: African (most diverse)

32



Which platform to use?



Summary

- Association analysis is a powerful strategy for common disease research
- HapMap and genomewide technologies enable whole-genome association scans