Association Mapping

David Evans

Outline

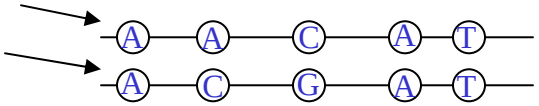
- Association
- Linkage vs association
- HapMap
- Genome-wide Association

Definitions

Locus: *Location* on the genome

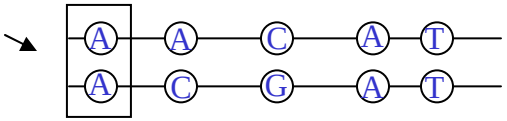
SNP: “Single Nucleotide Polymorphism” a mutation that produces a single base pair change in the DNA sequence

alleles



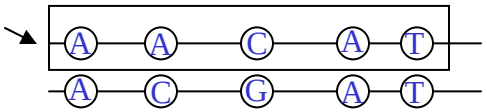
alternate forms of a SNP (mutation)

genotypes



both alleles at a locus form a genotype

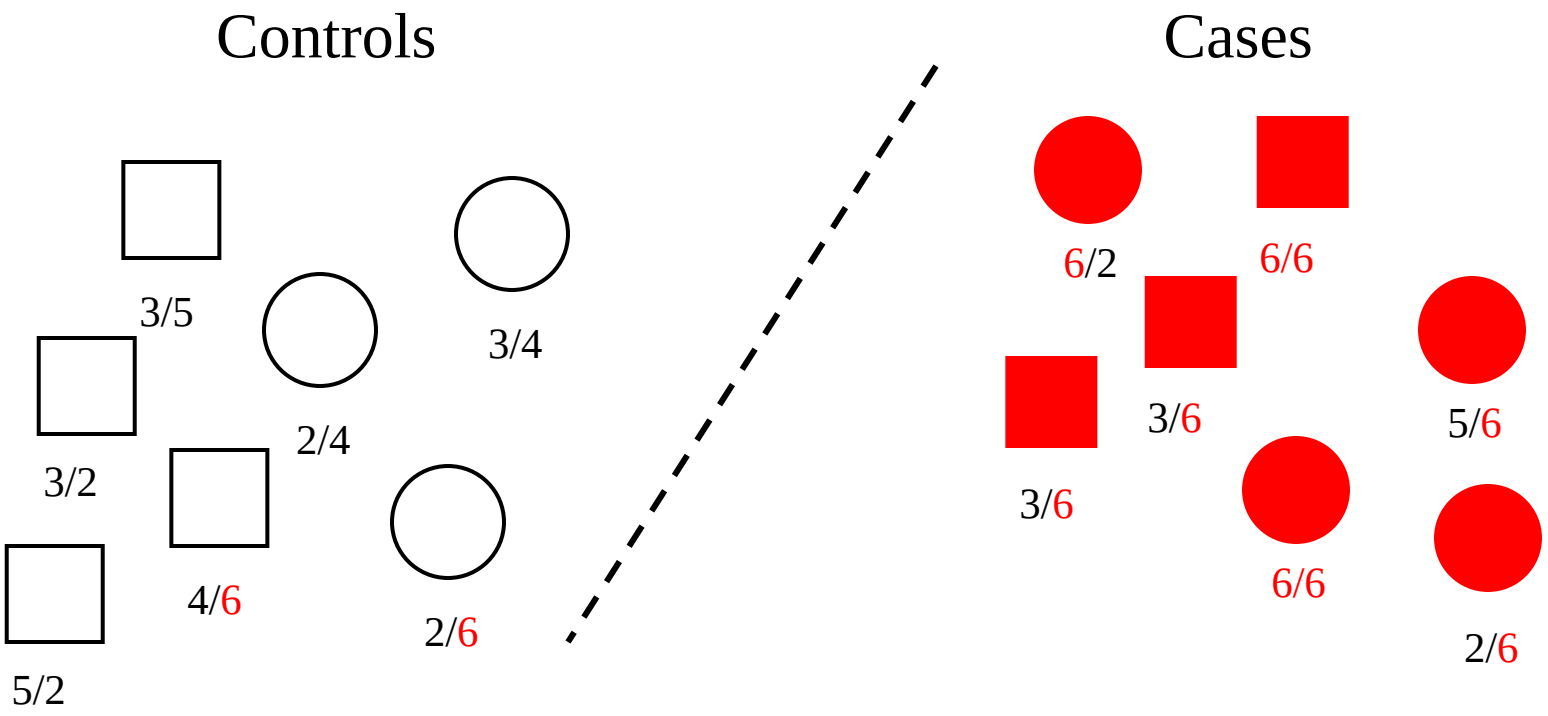
haplotypes



the pattern of alleles on a chromosome

Genetic Association: Correlation between (alleles/genotype/haplotype) and a phenotype of interest.

Genetic Case Control Study



Allele 6 is ‘associated’ with disease

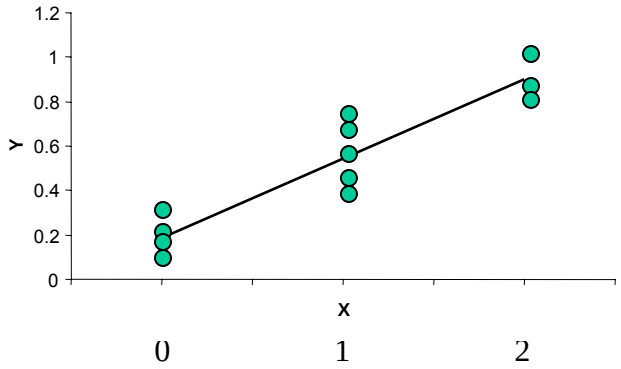
Simple Regression Model of Association

$$Y_i = \alpha + \beta X_i + e_i$$

where

Y_i = trait value for individual i

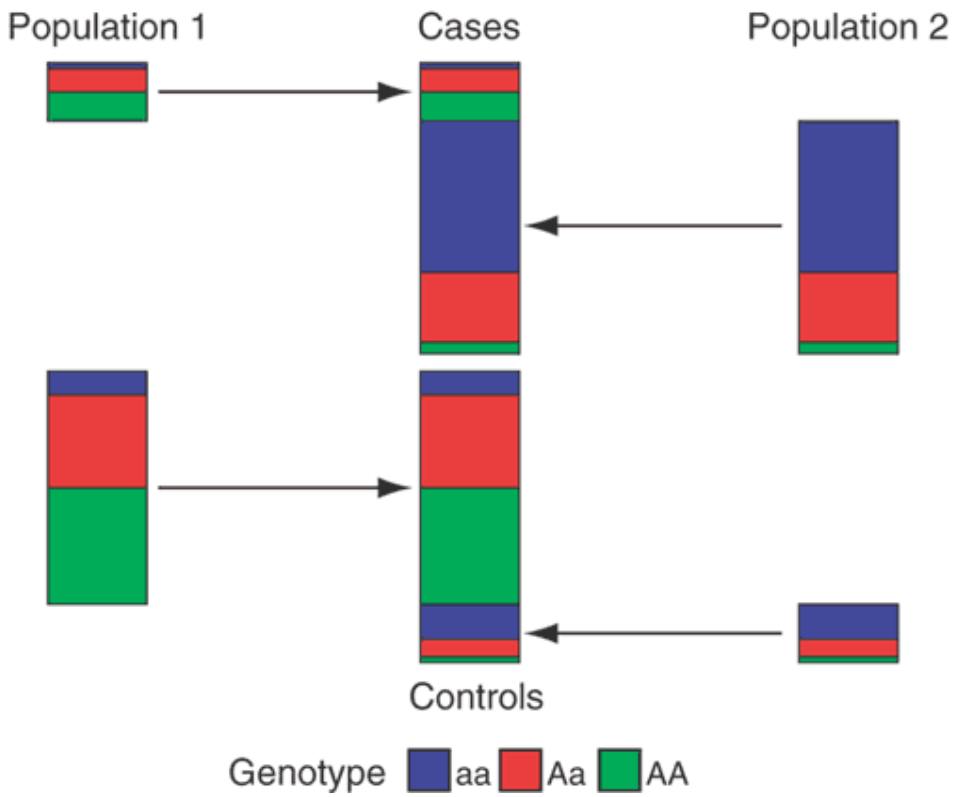
X_i = number of ‘A’ alleles an individual has



Population Stratification

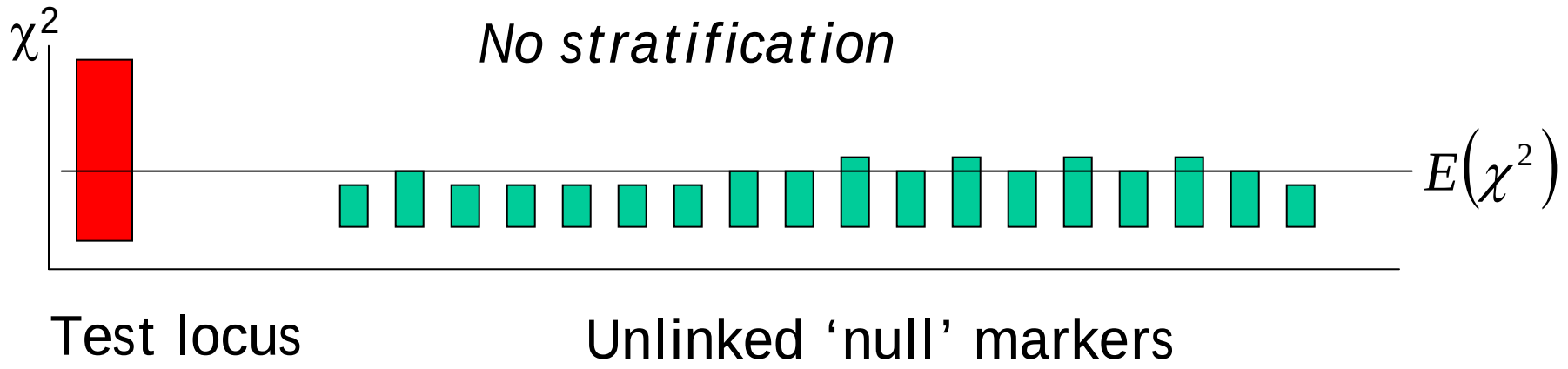
- Imagine a sample of individuals drawn from a population consisting of two distinct subgroups which differ in allele frequency.
- If the prevalence of disease is greater in one sub-population, then this group will be over-represented amongst the cases.
- Any marker which is also of higher frequency in that subgroup will appear to be associated with the disease
- Examples: “Chopsticks” gene, Height in Dutch
- Real world examples perhaps not as obvious, but the possibility of its existence should always be treated seriously (particularly GWA, large sample sizes)

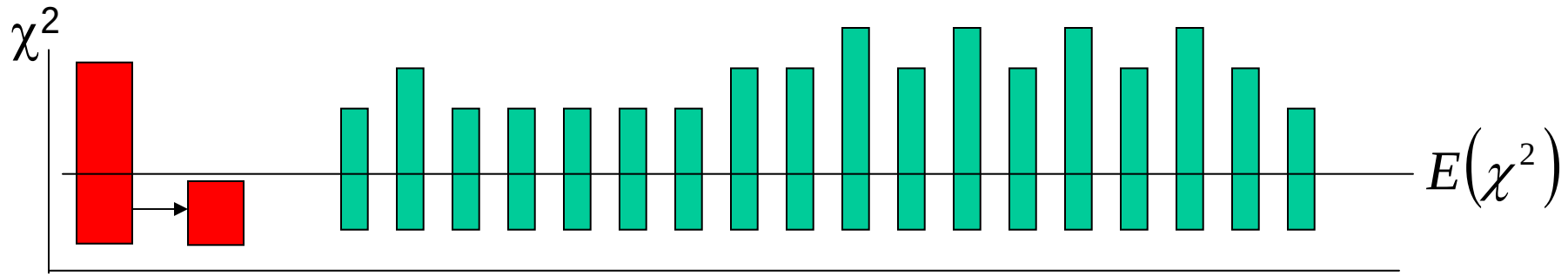
Stratification



Marchini, *Nat Genet.* 2004

Genomic control





Stratification → adjust test statistic

Principal Components Analysis

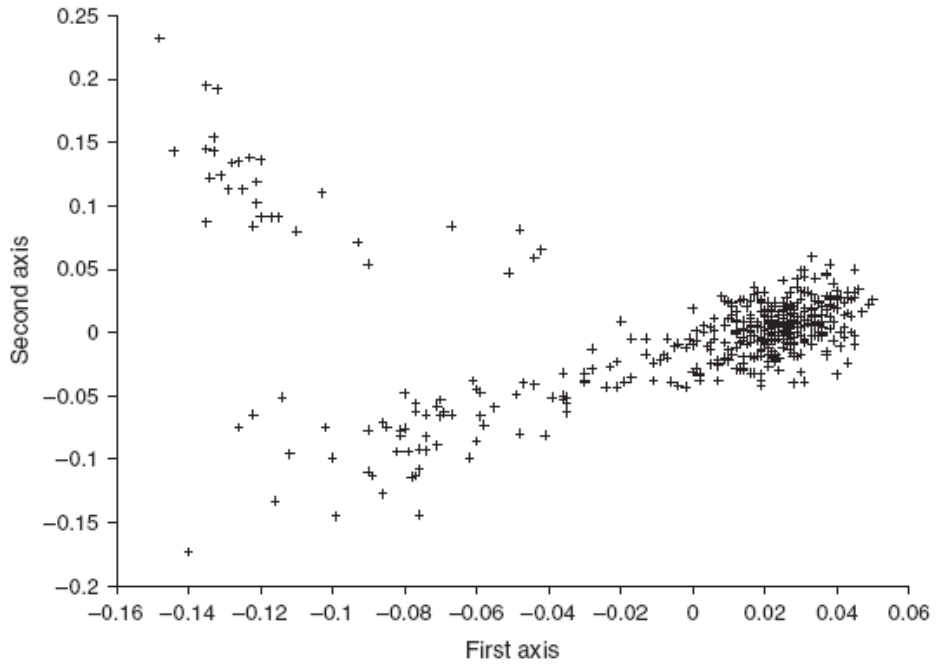
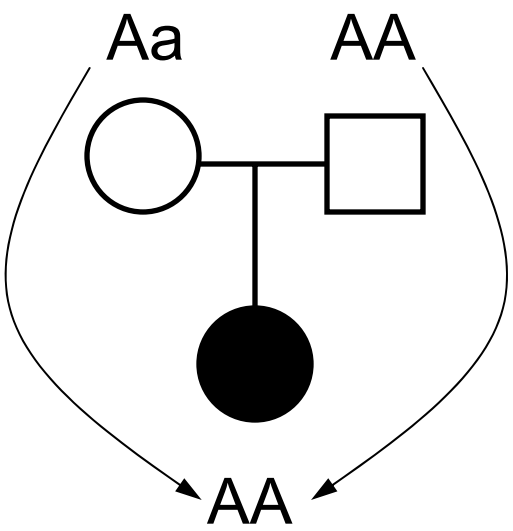


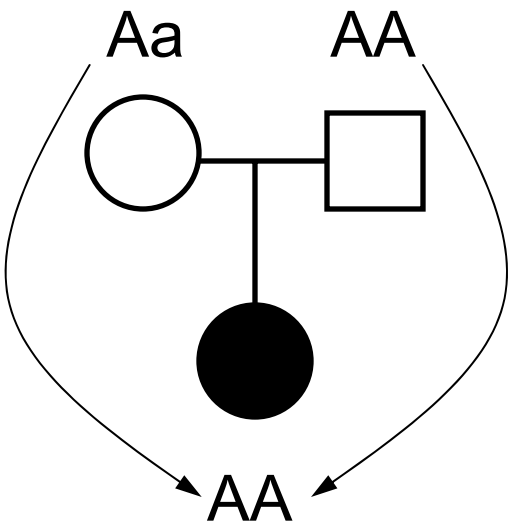
Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

Family Based Tests of Association



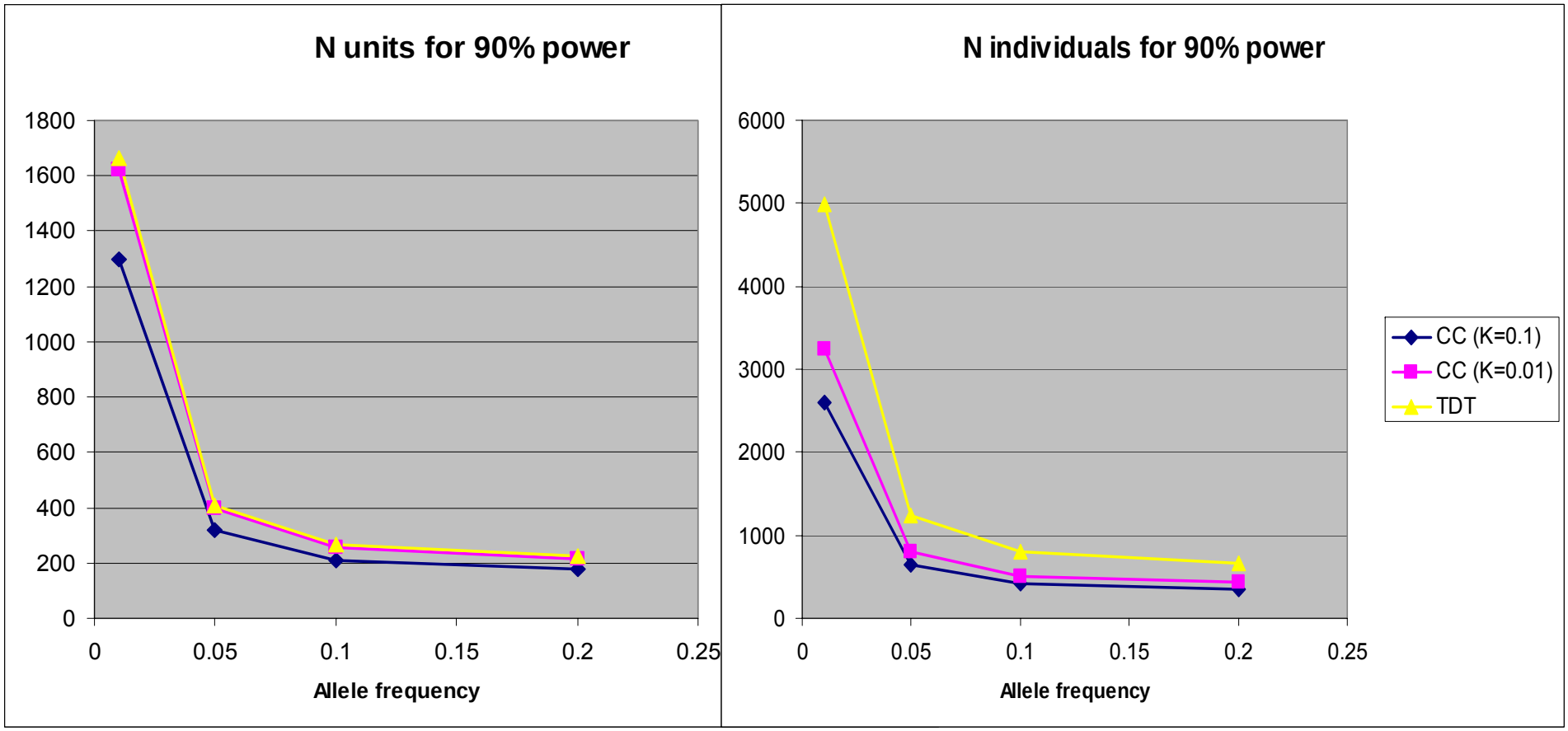
- Rationale: Related individuals have to be from the same population
- Many different family based tests designed to control for substructure (quantitative traits)
- TDT Design

Within Family Tests of Association



- Difficult to gather families
- Difficult to get parents for late onset / psychiatric conditions
- Inefficient for genotyping (particularly GWA)

Case-control versus TDT



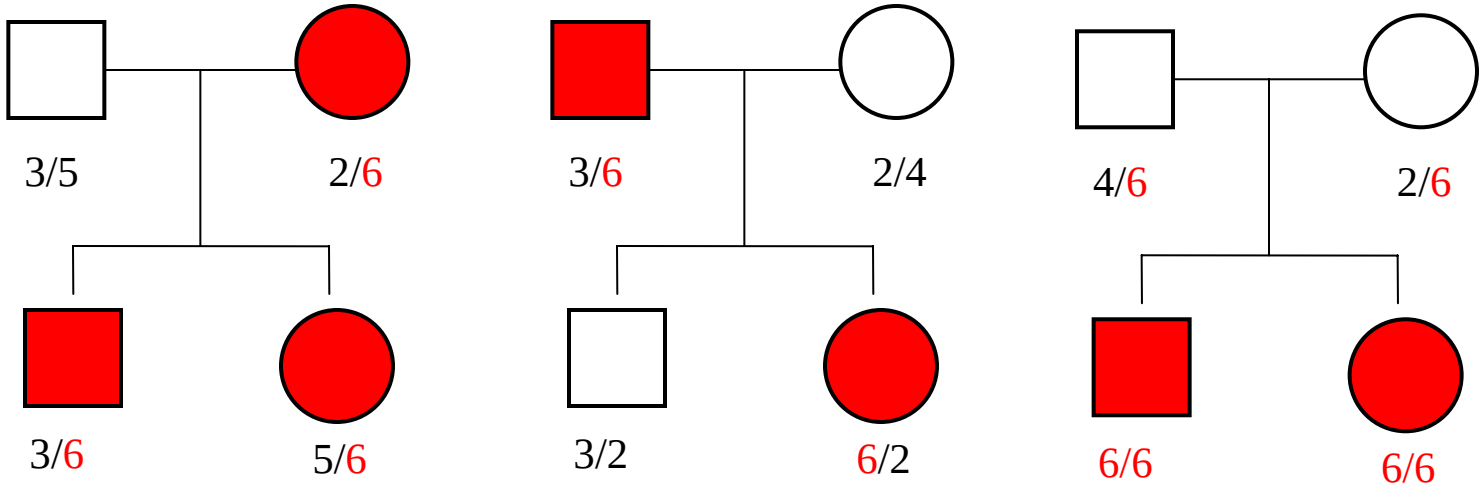
p = 0.1; RAA = RAa = 2

Association Study Designs and Statistical Methods

- Statistical Methods
  - Wide range: from t-test to evolutionary model-based MCMC
  - Principle always same: correlate phenotypic and genotypic variability

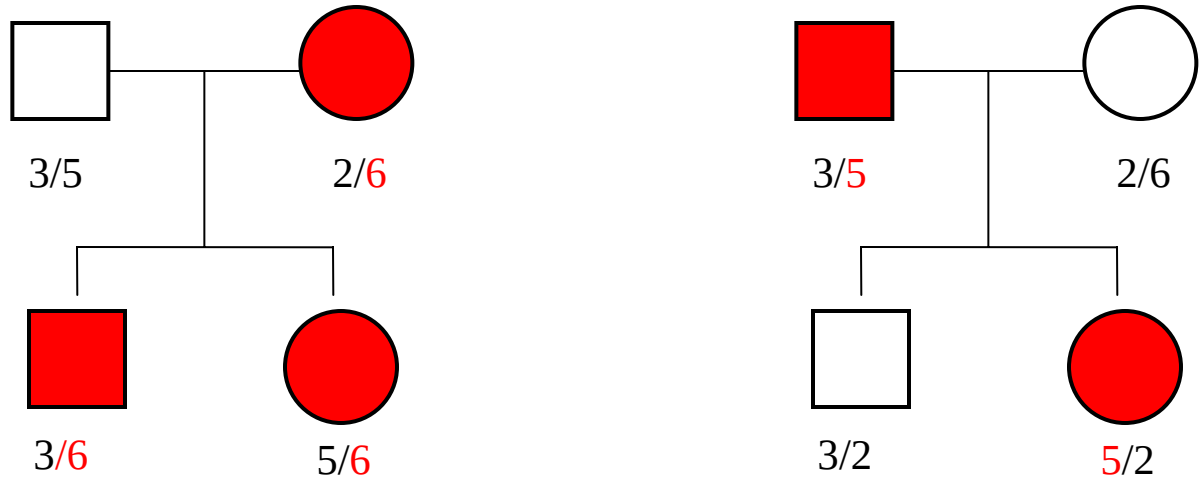
- Designs
  - Family-based
    - Trio (TDT), twins/sib-pairs/extended families (QTDT)
  - Case-control
    - Collections of individuals with disease, matched with sample w/o disease
    - Some ‘case only’ designs

Association (AND Linkage)



All families are ‘linked’ with the marker
Allele 6 is ‘associated’ with disease

Linkage



Both families are ‘linked’ with the marker, but a different allele is involved

Linkage is allelic association WITHIN families

Localization

- Linkage analysis yields broad chromosome regions harbouring many genes
  - Resolution comes from recombination events (meioses) in families assessed
  - ‘Good’ in terms of needing few markers, ‘poor’ in terms of finding specific variants involved
- Association analysis yields fine-scale resolution of genetic variants
  - Resolution comes from ancestral recombination events
  - ‘Good’ in terms of finding specific variants, ‘poor’ in terms of needing many markers

Power of Linkage vs Association

- Association generally has greater power than linkage
  - Linkage based on variances/covariances
  - Association based on means

- Power to detect association depends on:
  - Minor allele frequency
  - Correlation between marker and disease locus (“Linkage Disequilibrium”)
  - Sample Size
  - Alpha level (Number of markers)
  - Statistical test employed

Linkage vs Association

Linkage

1. Family-based
2. Matching/ethnicity generally unimportant
3. Few markers for genome coverage (300-400 microsatellites)
4. Can be weak design
5. Good for initial detection; poor for fine-mapping
6. Powerful for rare variants

Association

1. Families or unrelateds
2. Matching/ethnicity crucial
3. Many markers req for genome coverage (10^5 – 10^6 SNPs)
4. Powerful design
5. Ok for initial detection; good for fine-mapping
6. Powerful for common variants; rare variants generally impossible

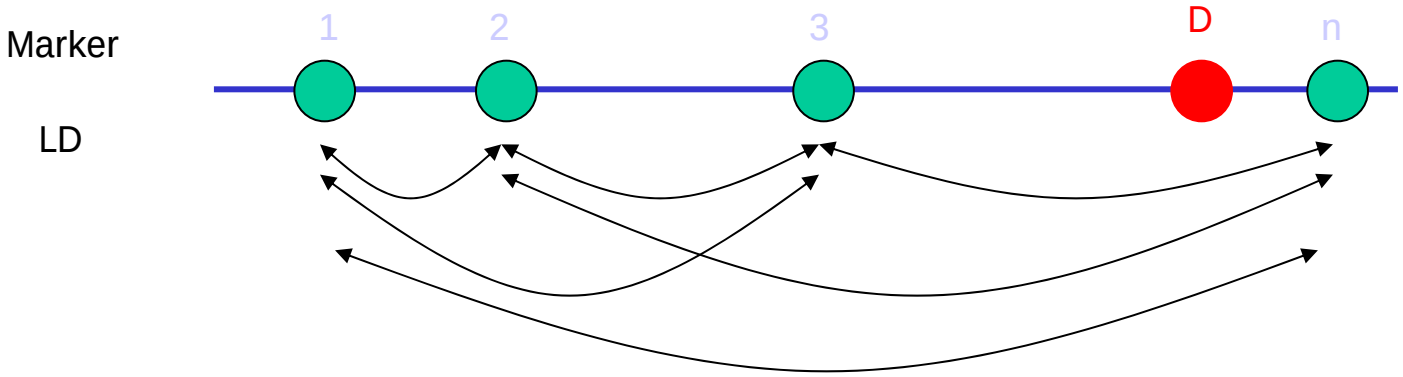
Allelic Association
Three Common Forms

- Direct Association
  - Mutant or ‘susceptible’ polymorphism
  - Allele of interest is itself involved in phenotype

- Indirect Association
  - Allele itself is not involved, but a nearby correlated marker changes phenotype

- Spurious association
  - Apparent association not related to genetic aetiology (most common outcome…)

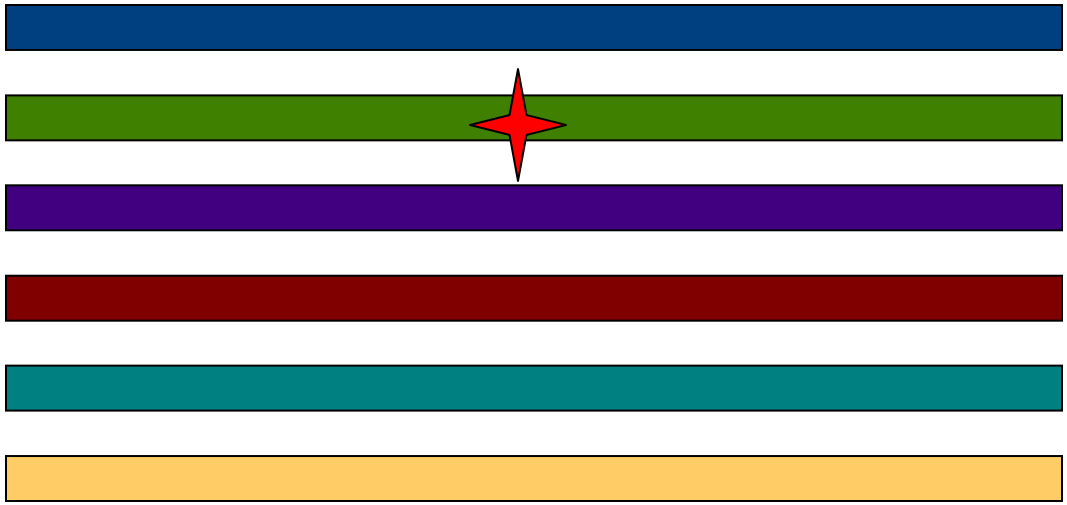
Linkage Disequilibrium & Allelic Association



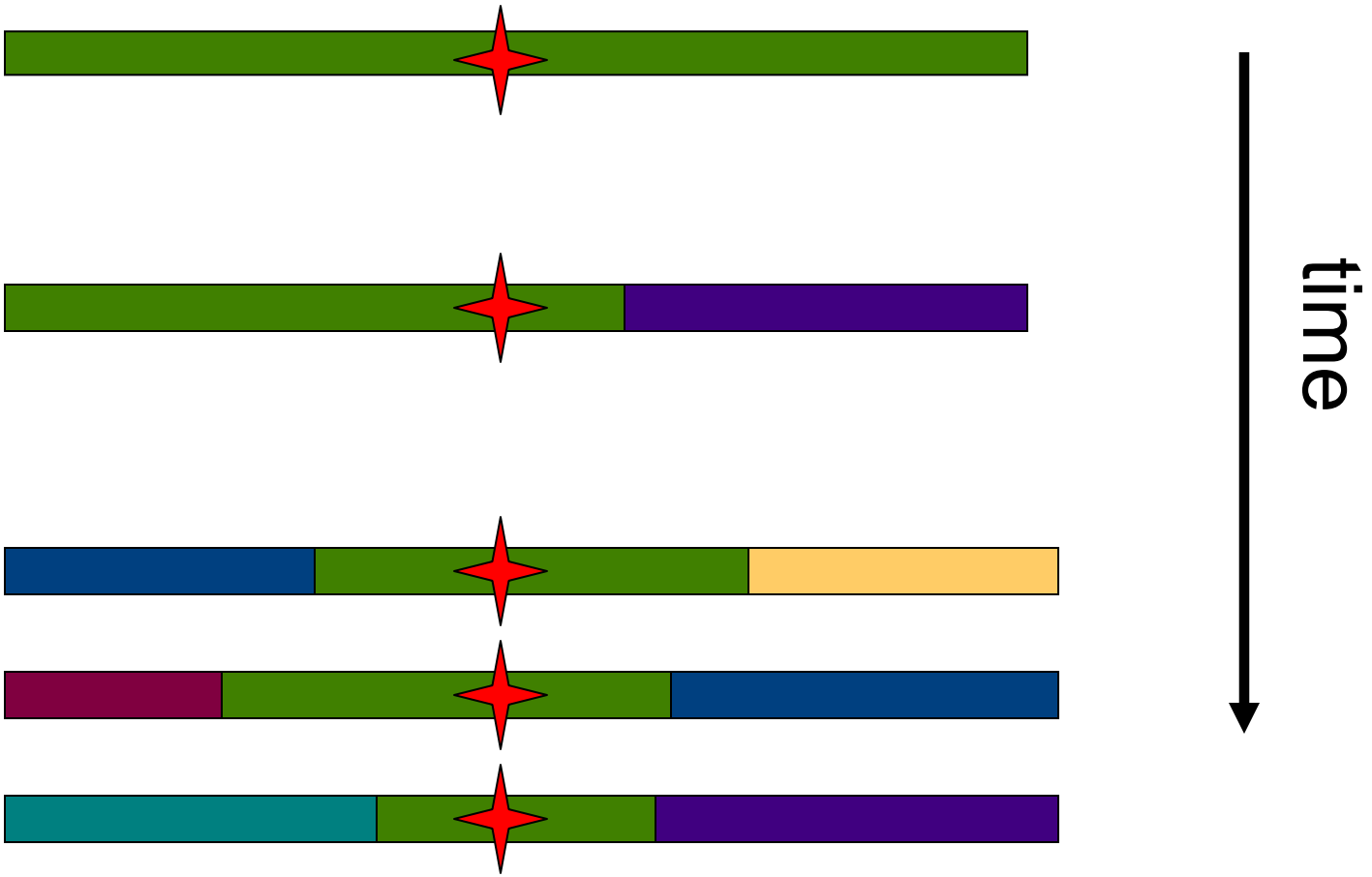
Markers close together on chromosomes are often transmitted together, yielding a non-zero correlation between the alleles. This is *linkage disequilibrium*

It is important for allelic association because it means we don't need to assess the exact aetiological variant, but we see trait-SNP association with a neighbouring variant

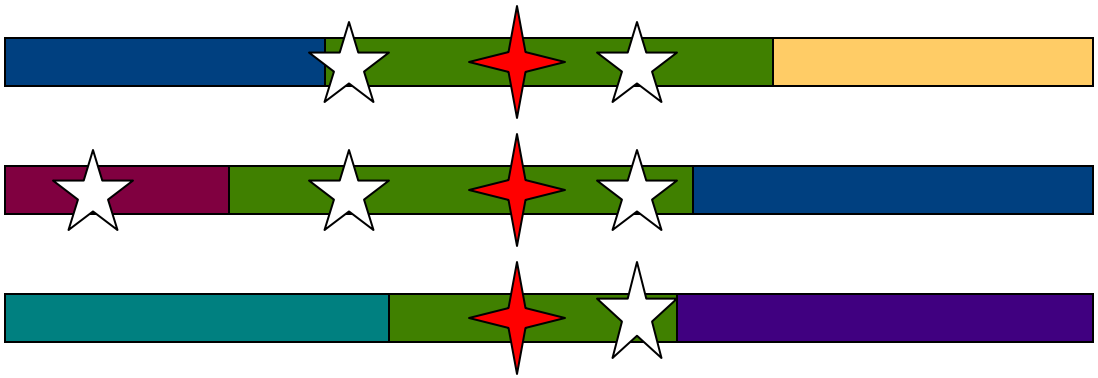
Linkage disequilibrium



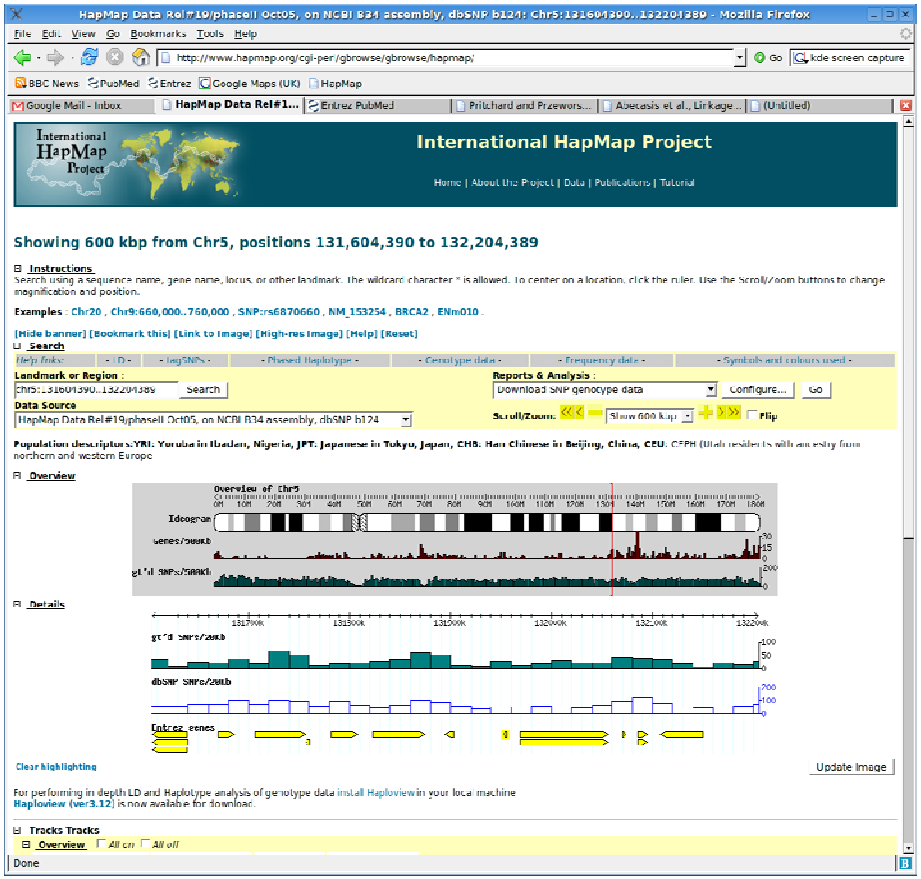
Linkage disequilibrium



Linkage Disequilibrium



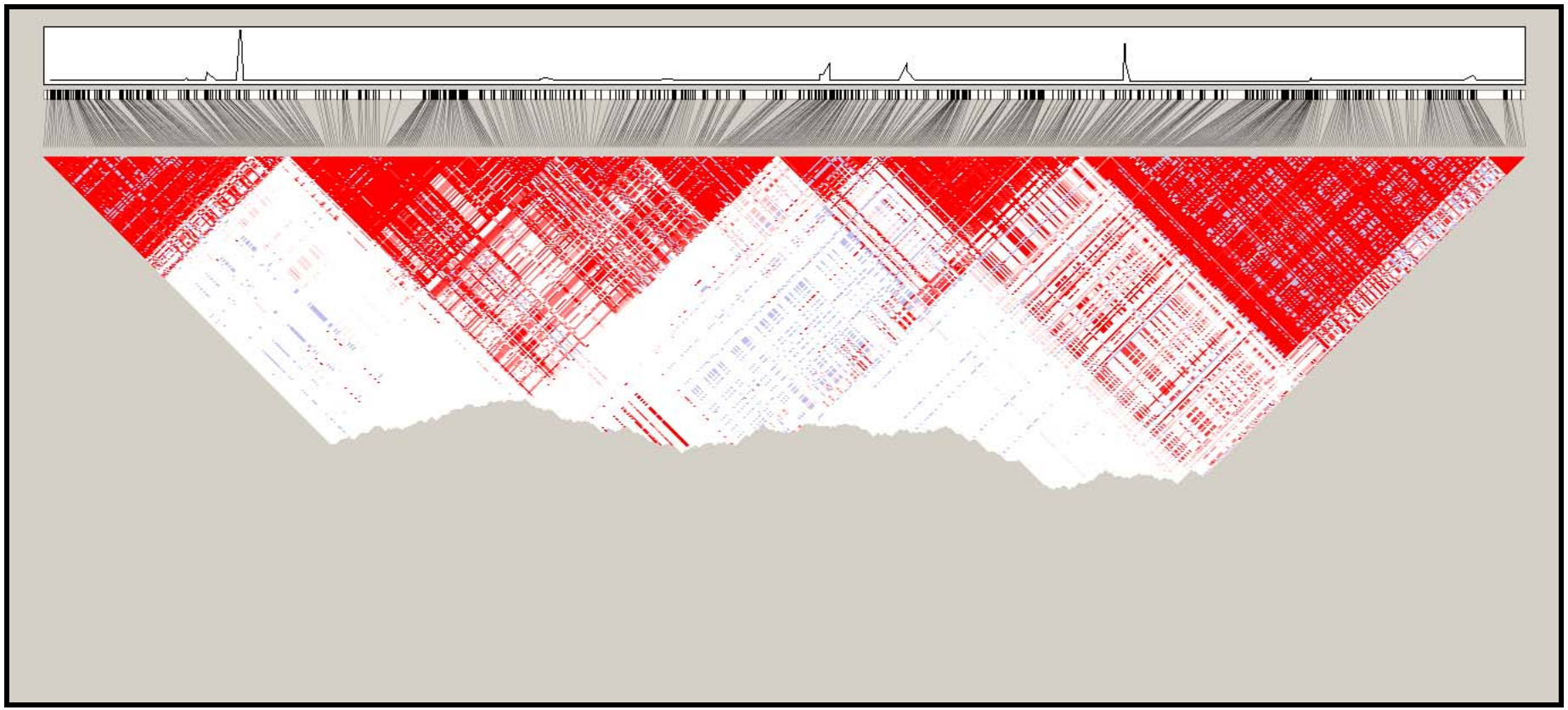
Enabling association studies: HapMap



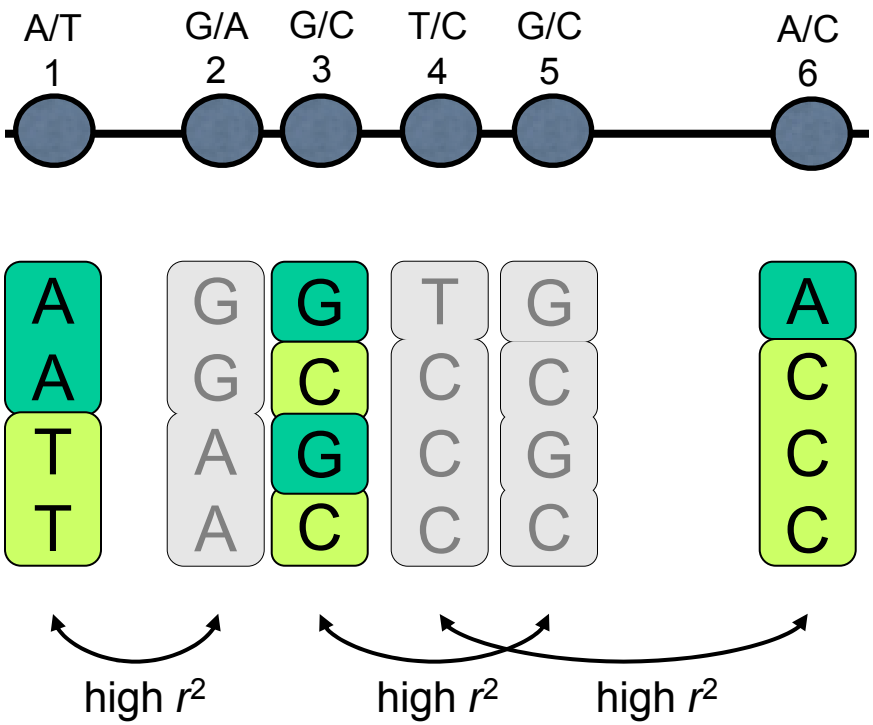
HapMap Strategy

- Rationale: there are ~10 million common SNPs in human genome
  - We can't afford to genotype them all in each association study
  - But maybe we can genotype them once to catalogue the redundancies and use a smaller set of 'tag' SNPs in each association study
- Samples
  - Four populations, 270 indivs total
- Genotyping
  - 5 kb initial density across genome (600K SNPs)
  - Then second phase to ~ 1 kb across genome (4 million)
  - All data in public domain

Visualizing empirical LD



Pairwise tagging



Tags:

SNP 1
SNP 3
SNP 6

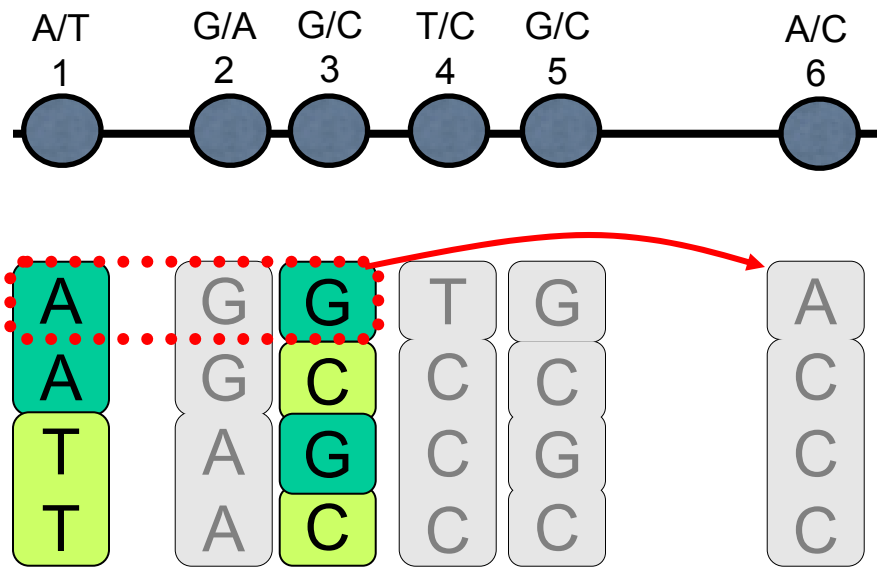
3 in total

Test for association:

SNP 1
SNP 3
SNP 6

Carlson *et al.* (2004) *AJHG* **74**:106

Use of haplotypes can improve genotyping efficiency



Tags:

SNP 1
SNP 3
SNP 6
2 in total
3 in total

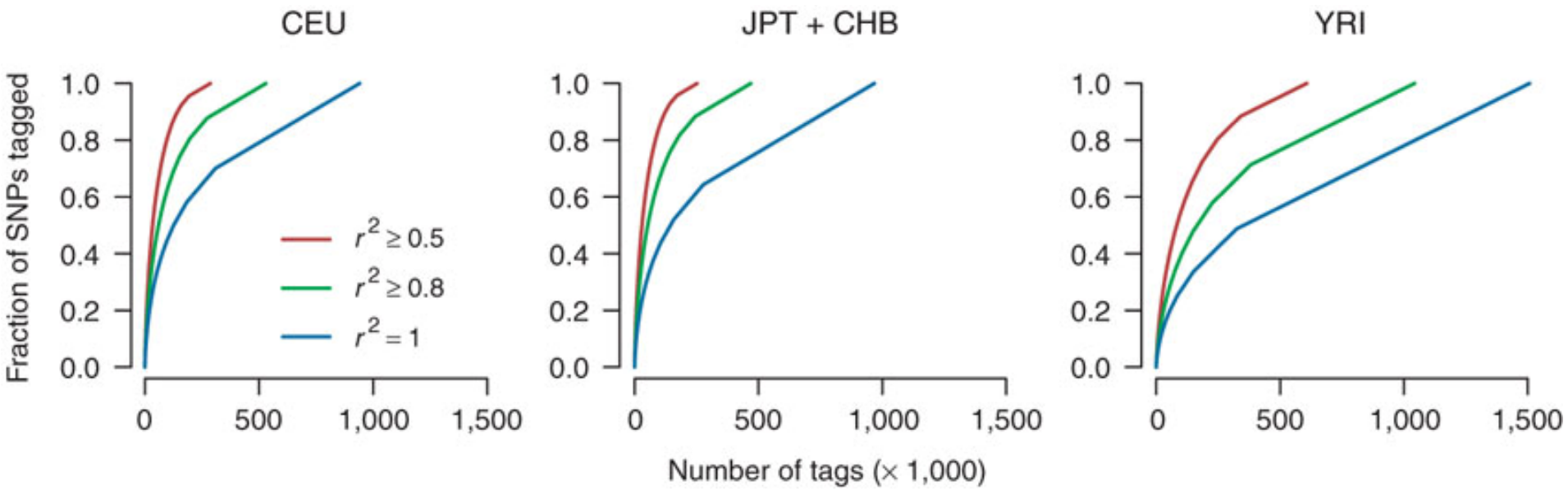
Test for association:

SNP 1 captures 1+2
SNP 3 captures 3+5
SNP 6 captures SNP 4+6

“AG” haplotype

de Bakker *et al.* (2005) *Nat Genet* **37**:1217

Genome-wide tagging coverage



Barrett and Cardon, *Nat Genet* (2006).

Commercial SNP Panels

- Comprise ≈ 100,000 – 1.8 million genetic variants
- Cover up to ~95% of common genetic variants
- Rare variants are not captured well

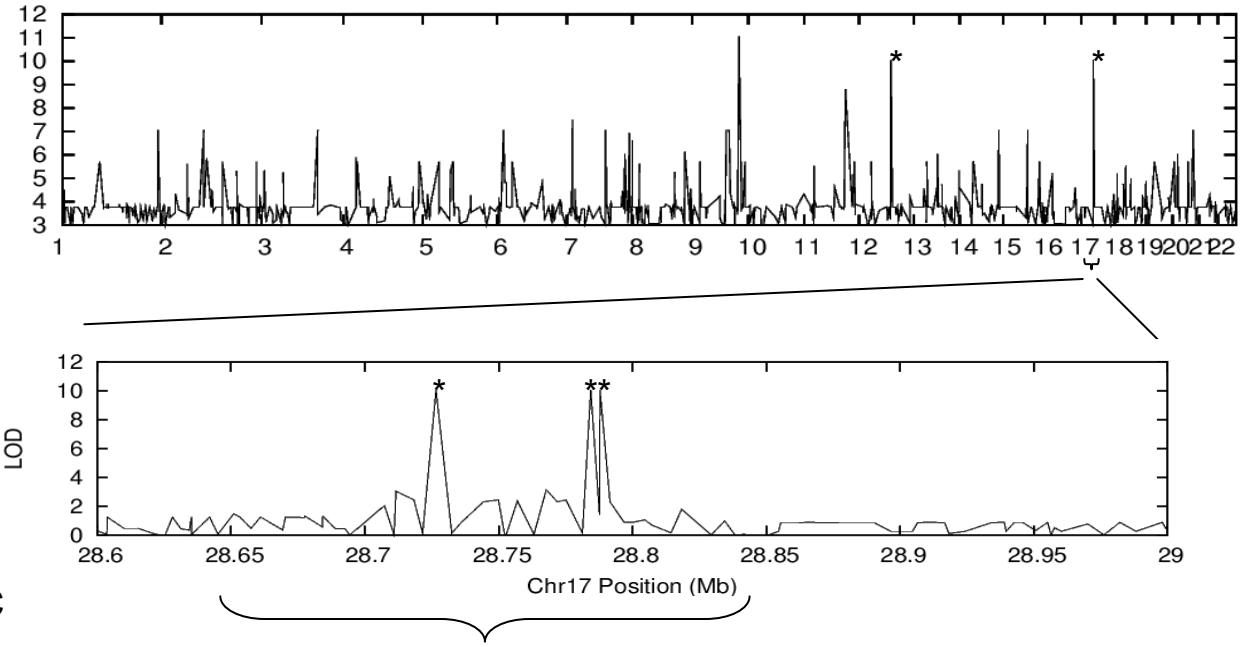
Table 1 Genomic coverage of commercial GWAS products for common SNPs at $r^2 \geq 0.8$, evaluated in Phase II HapMap

		CEU		JPT+CHB		YRI	
	Type	Coverage (%)	Mean r^2	Coverage (%)	Mean r^2	Coverage (%)	Mean r^2
Illumina HumanHap300	Tag	75	0.961	63	0.964	28	0.961
Affymetrix 500K	Random	65	0.975	66	0.974	41	0.971
Affymetrix 111K	Random	31	0.960	31	0.957	15	0.957
Affymetrix 500k + 175K tag	Combination	86	0.975	79	0.978	49	0.973
Illumina Human-1	Gene	26[a]	0.957	28[a]	0.955	12[a]	0.956

Despite the r^2 cutoff of 0.8, the mean r^2 for tagged SNPs is very high; also, 'untagged' SNPs are covered with intermediate values of r^2, providing modest power to detect such alleles (Supplementary Fig. 1).
[a]Coverage estimates for the Human-1 product are underestimates because some of its SNPs were not genotyped in the HapMap project. As these SNPs are largely rare genic SNPs, it is not expected that they would substantially raise coverage of common variation.

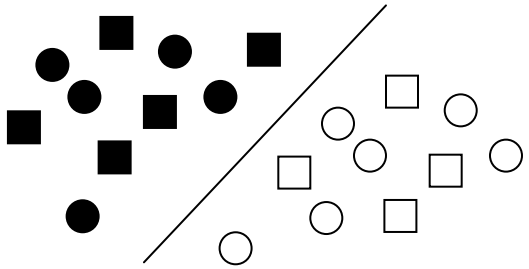
Whole Genome Association

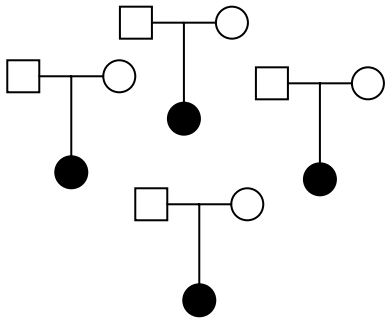
Scan Entire Genome
- 100,000s SNPs



Identify local regions of interest, examine genes, SNP density regulatory regions, etc

Replicate the finding





Programs for performing association analysis

- Mx (Neale)
  - Fully flexible, ordinal data
  - Not ideal for large pedigrees or GWAs
- PLINK (Purcell, Neale, Ferreira)
  - GWA
- Haploview (Barrett)
  - Graphical visualization of LD, tagging, basic tests of association
- MERLIN, QTDT (Abecasis)
  - Association and linkage in families