Association Mapping

Lon Cardon

University of Oxford
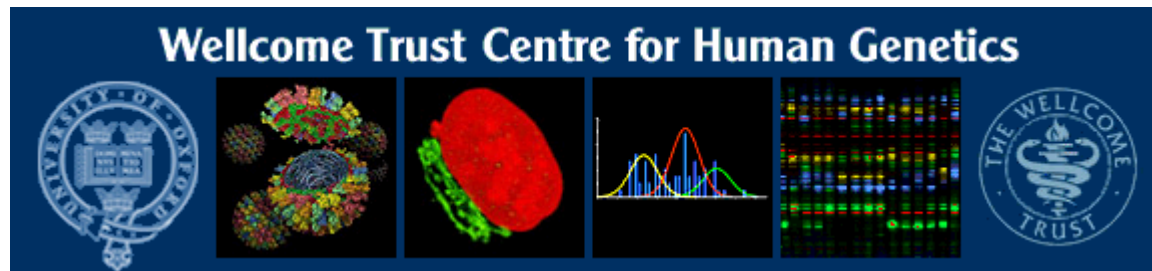
Wellcome Trust Centre for Human Genetics

Outline

1. Association and linkage
2. Association and linkage disequilibrium
3. History and track record of association studies
4. Challenges
5. Example

Outline

1. **Association and linkage**
2. Association and linkage disequilibrium
3. History and track record
4. Challenges
5. Example

Association Studies

Simplest design possible
> Correlate phenotype with genotype

Candidate genes for specific diseases
> common practice in medicine/genetics

Pharmacogenetics
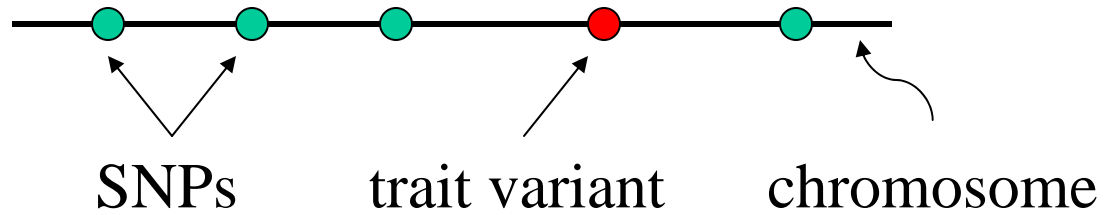> genotyping clinically relevant samples (toxicity vs efficacy)

Positional cloning
> recent popular design for human complex traits

Genome-wide association
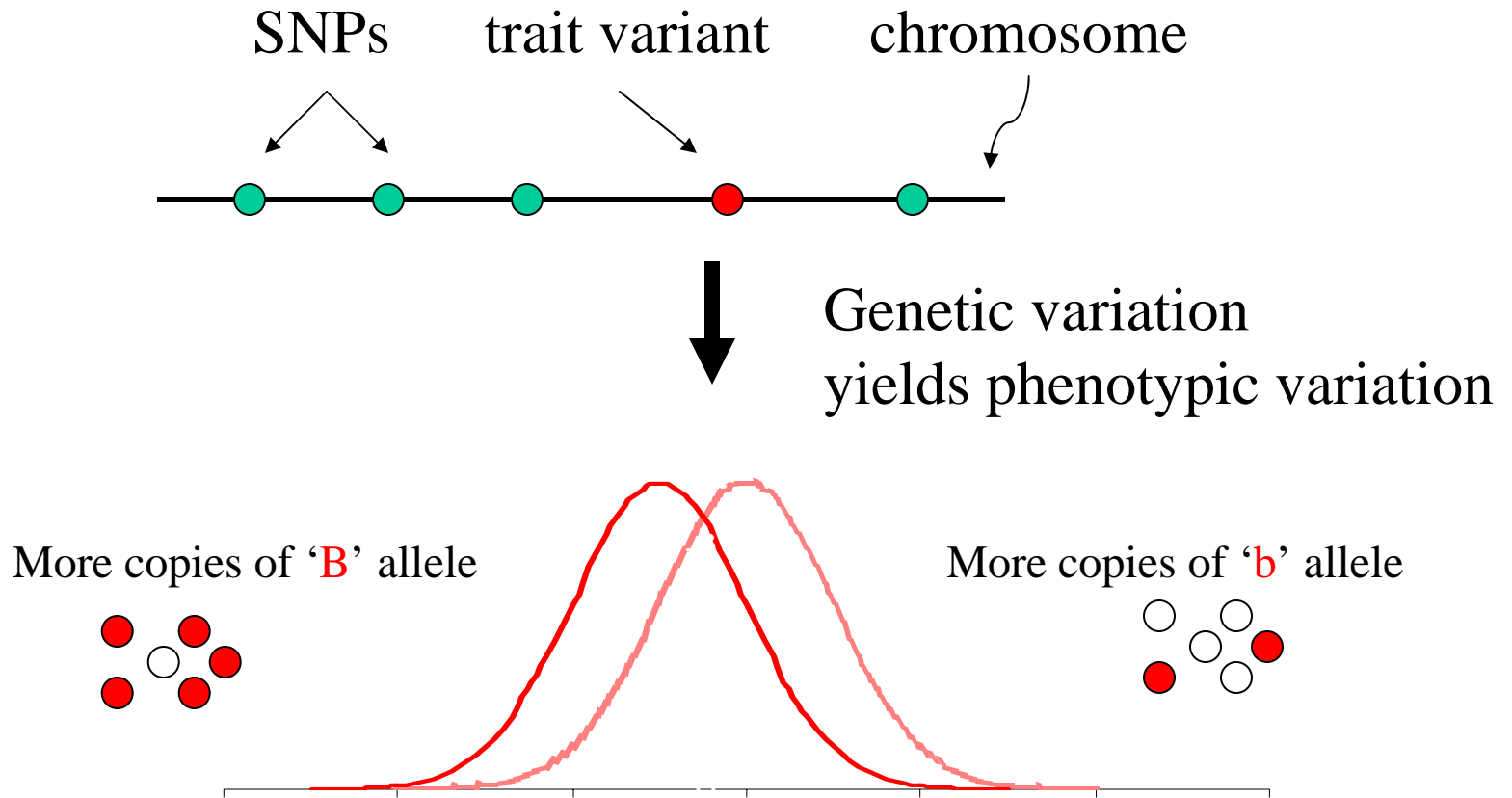> with millions available SNPs, can search whole genome exhaustively
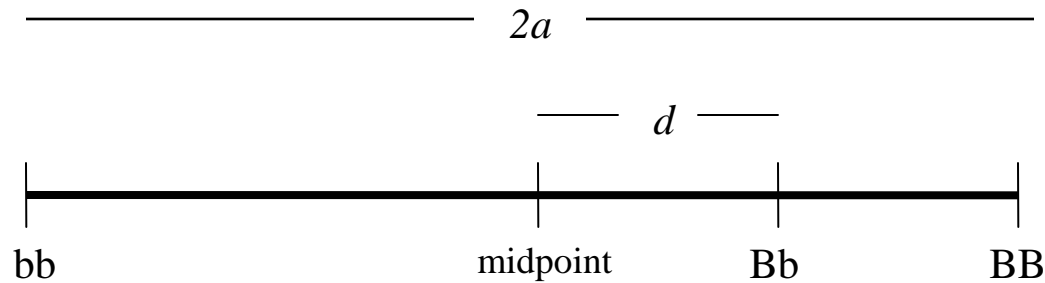
Definitions

SNPs   trait variant   chromosome

Population Data

	Affection	Trait$_1$...Trait$_n$	
haplotypes	A	10.3	75.66
genotypes	A	9.9	-99
alleles	U	15.8	101.22

Allelic Association

Biometrical Model

Genotype	Genetic Value
BB	a
Bb	d
bb	$-a$

$$V_a \text{ (QTL)} = 2pqa^2 \text{ (no dominance)}$$

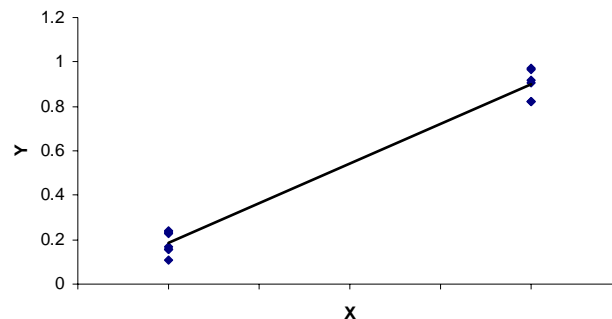# Simplest Regression Model of Association

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

where

$Y_i =$ trait value for individual i

$X_i =$ 1 if allele individual i has allele 'A'

         0 otherwise

i.e., test of mean differences between 'A' and 'not-A' individuals

Association Study Designs and Statistical Methods

- **Designs**
  - Family-based
    - Trio (TDT), sib-pairs/extended families (QTDT)

  - Case-control
    - Collections of individuals with disease, matched with sample w/o disease
    - Some 'case only' designs

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

Linear Model of Association
(Fulker et al, *AJHG*, 1999)

Biometrical basis

$$y_{ij} = G_{ij} + g_{ij} + e_{ij};$$

$$G_{ij} = \begin{cases} a & \text{if } \text{genotype}_{ij} = \text{BB} \\ d & \text{if } \text{genotype}_{ij} = \text{Bb} \\ -a & \text{if } \text{genotype}_{ij} = \text{bb} \end{cases}$$

Variance model (linkage)

$$Cov(y_{ij}, y_{ik} / \pi_{ijk}) = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } i = j \\ \sigma_a^2 f(\pi_{ikj}) + \dfrac{1}{2}\sigma_g^2 & \text{if } i \neq j \end{cases}$$

π_{ijk} = proportion of alleles shared ibd at marker
σ_a^2 = additive genetic variance parameter
σ_g^2 = polygenic (residual) variance parameter
σ_e^2 = environmental (residual) variance parameter
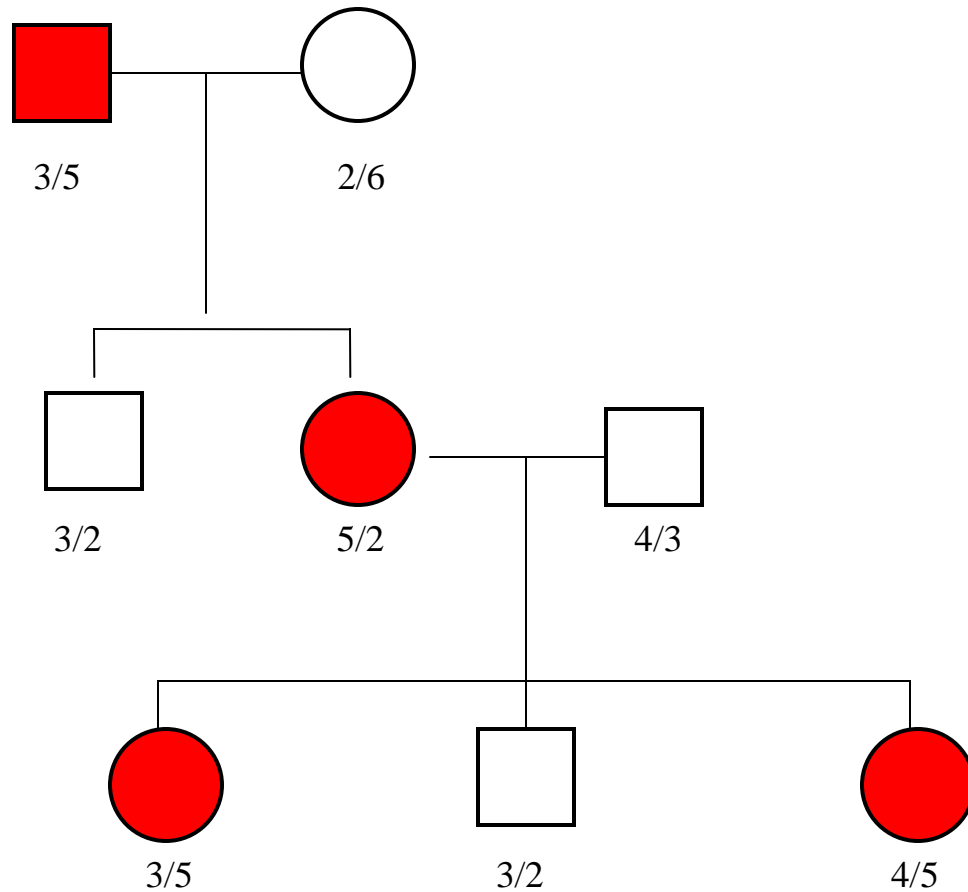
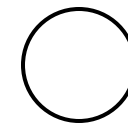## Linear model (association)

$$\mu_{ij} = \alpha + \beta X_{ij}$$

Likelihood

$$\log(L) = c - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{\Omega}_i| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{\mu}_i)'\mathbf{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{\mu}_i)$$

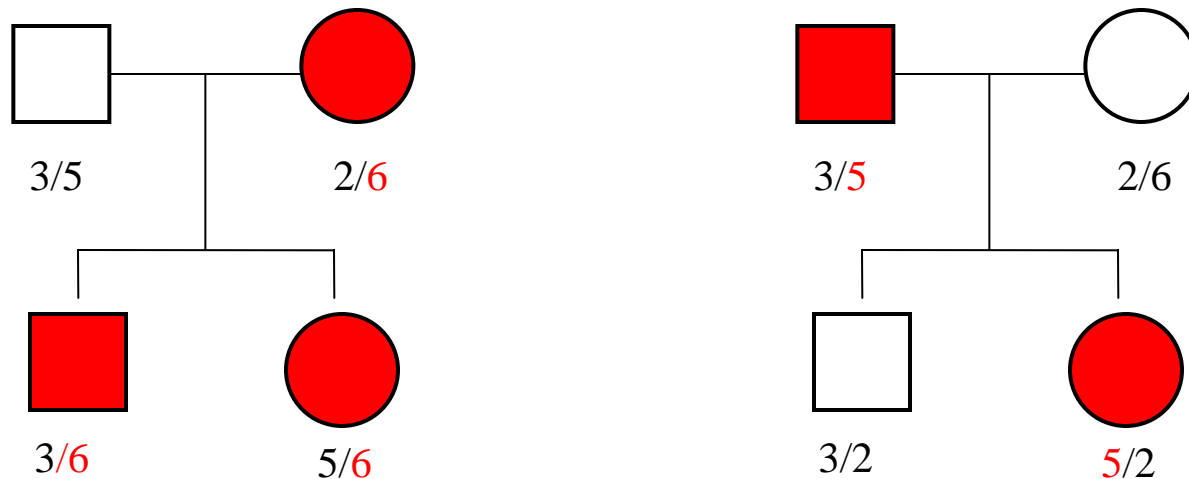# Linkage: Allelic association WITHIN FAMILIES

affected

Allele coded by CA copies
2 = CACA
6 = CACACACACACA

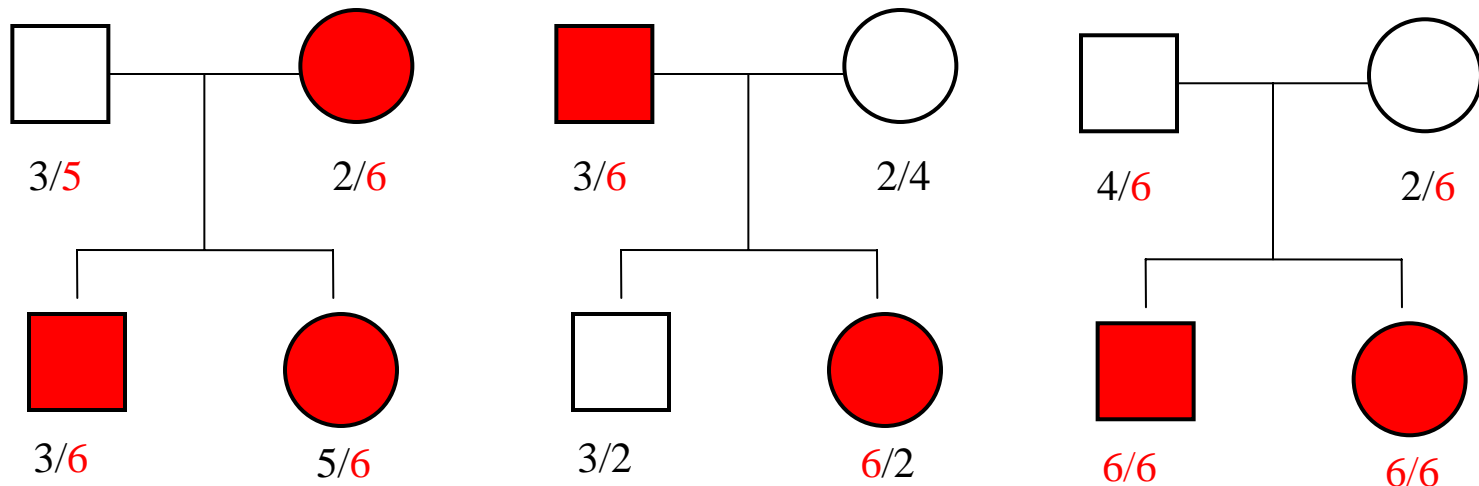Disease linked to '5' allele in dominant inheritance

Allelic Association:
Extension of linkage to the population

3/5   2/6   3/5   2/6

3/6   5/6   3/2   5/2

Both families are 'linked' with the marker, but a different allele is involved
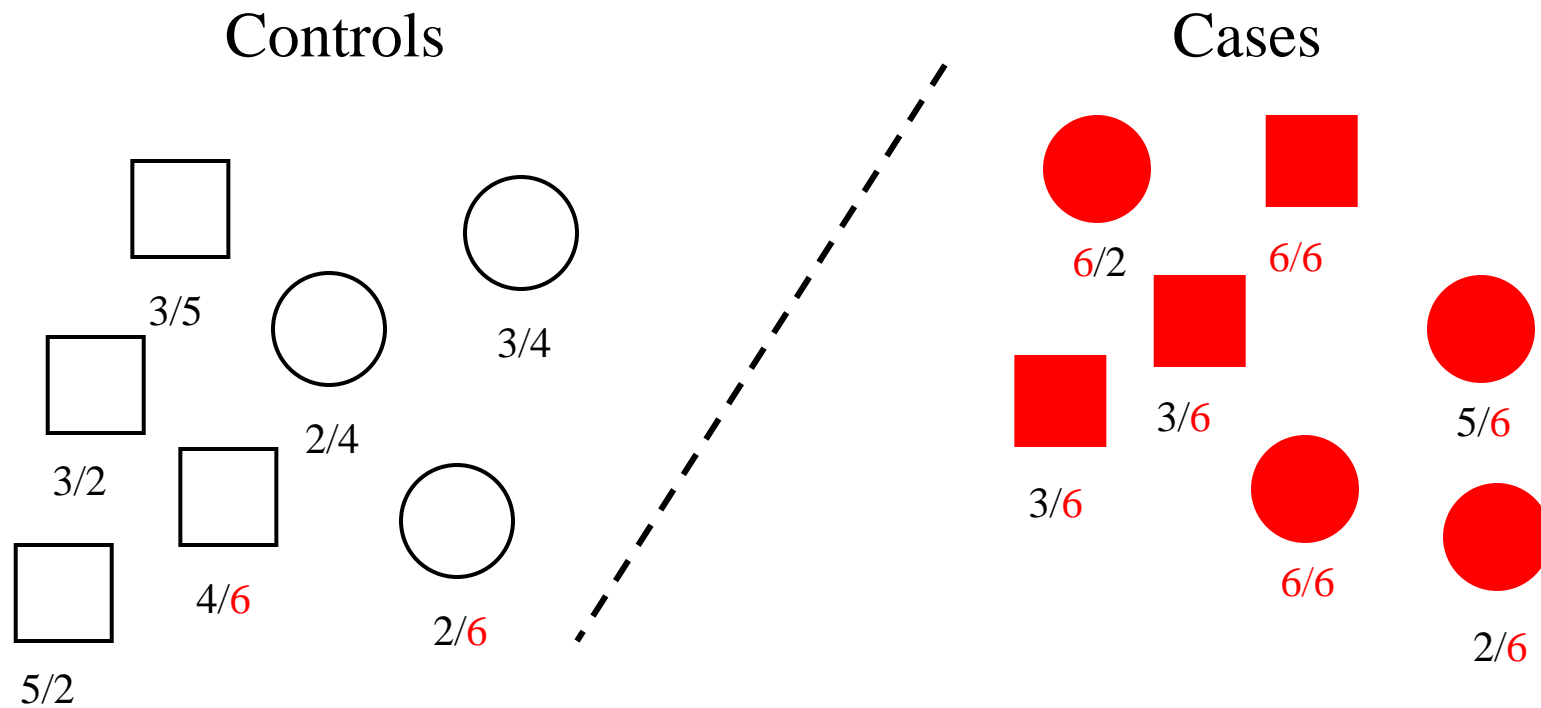
Allelic Association
Extension of linkage to the population

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

Allelic Association

Controls   Cases

3/5

3/4

2/4

3/2

4/6

2/6

5/2

6/2   6/6

3/6

3/6   5/6

6/6

2/6

Allele 6 is 'associated' with disease

Power of Linkage vs Association

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

  - See lectures by Ben Neale (linkage power), Shaun Purcell (assoc power)

Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease

Jean-Pierre Hugot*†‡, Mathias Chamaillard*†, Habib Zouali*,
Suzanne Lesage*, Jean-Pierre Cézard‡, Jacques Belaiche§,
Sven Almer‖, Curt Tysk¶, Colm A. O'Morain#, Miquel Gassull☆,
Vibeke Binder**, Yigael Finkel††, Antoine Cortot‡‡,
Robert Modigliani§§, Pierre Laurent-Puig†, Corine Gower-Rousseau‡‡,
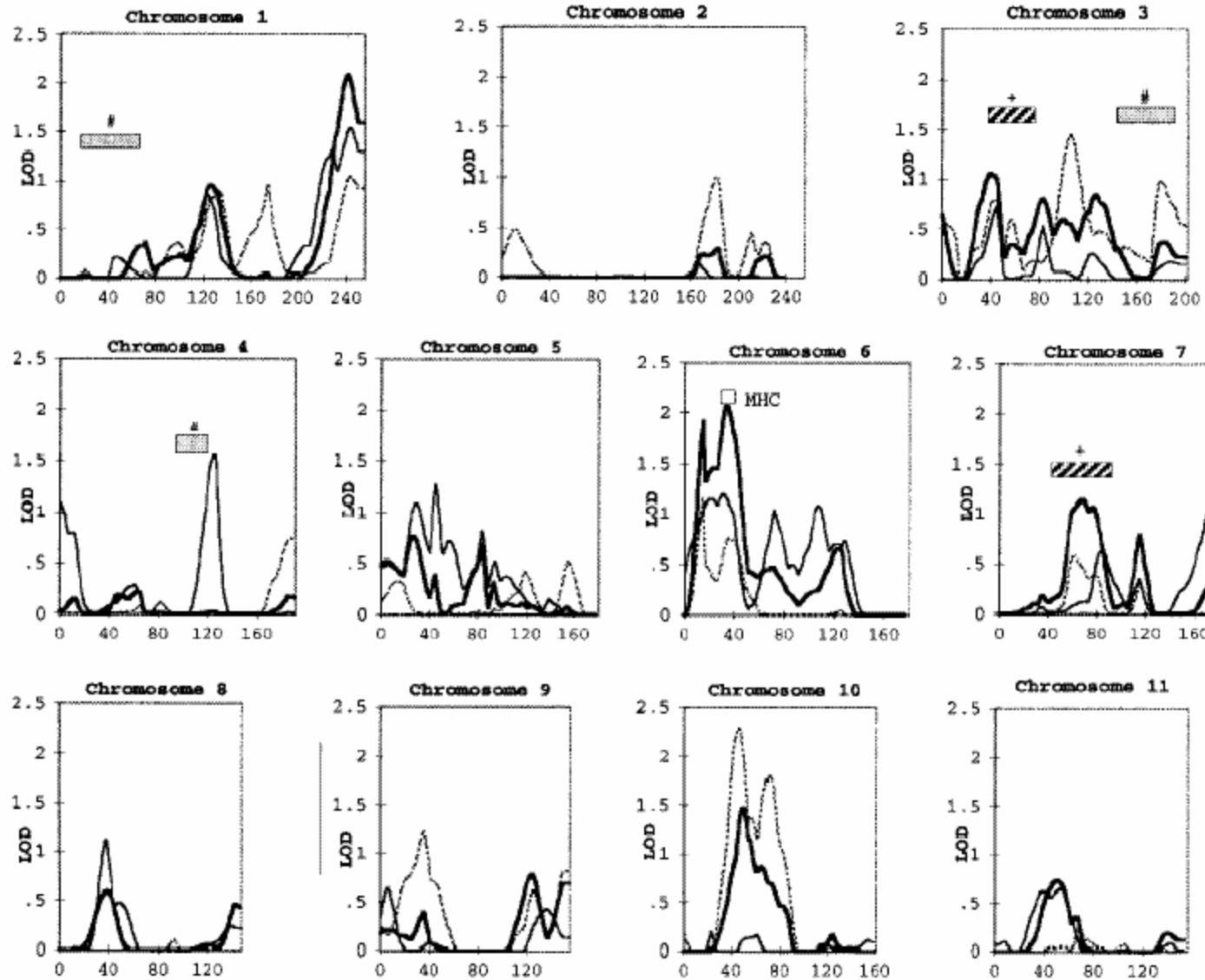Jeanne Macry‖‖, Jean-Frédéric Colombel‡‡, Mourad Sahbatou*
& Gilles Thomas*†¶¶

First (unequivocal)
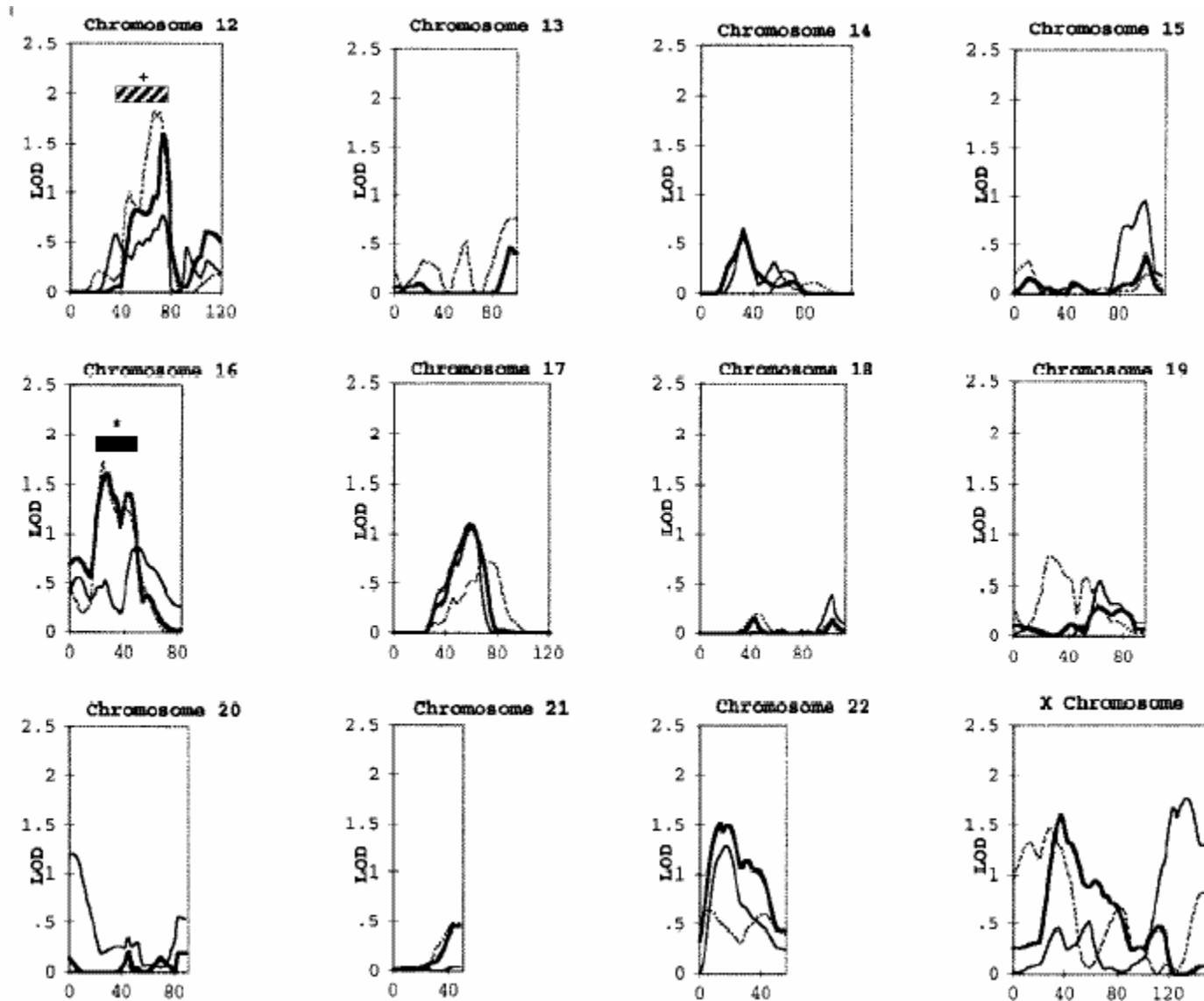positional cloning of a
complex disease QTL !
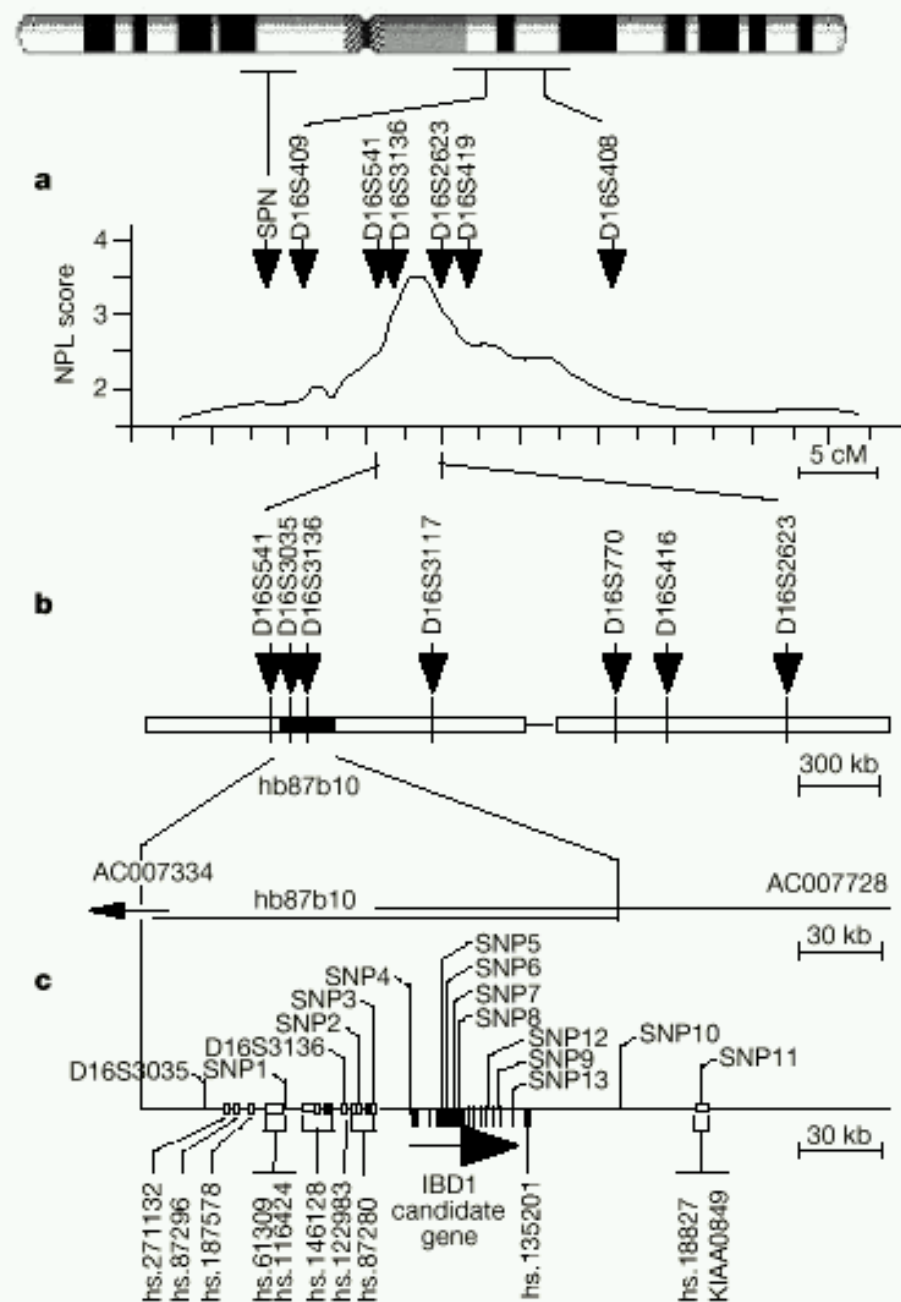
Inflammatory Bowel Disease Genome Screen
Satsangi et al, Nat Genet 1996

Inflammatory Bowel Disease Genome Screen

NOD2 Association Results Stronger than Linkage Evidence

- Analysis strategy: same families, same individuals as linkage, but now know mutations. Were the effects there all along?

- TDT

TABLE VI. *NOD2* Haplotype Associations With CD

Haplotype[a]	All families n = 294 CD trios[b]	
Pro268/Arg702/Gly908/Leu1007	71 TR : 109 NT	
Pro268Ser/Arg702/Gly908/Leu1007	50 TR : 70 NT	
Pro268Ser/**Arg702Trp**/Gly908/Leu1007	43 TR : 21 NT	$P = 0.01$
Pro268Ser/Arg702/Gly908/**Leu1007fsinsC**	41 TR : 6 NT	$P = 0.000007$
Pro268Ser/Arg702/**Gly908Arg**/Leu1007	9 TR : 9 NT	$P = $ ns

[a]Minor variants shown in bold type.
[b]*P*-values shown for over-transmitted haplotypes. Transmission numbers are not identical to Table V, due to haplotype ambiguity and occasional PCR failure.

- Case-control

Genotype Rel Risk = 58.9, $p < 10^{-8}$

Same CD cases vs 229 controls

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

Linkage Resolution
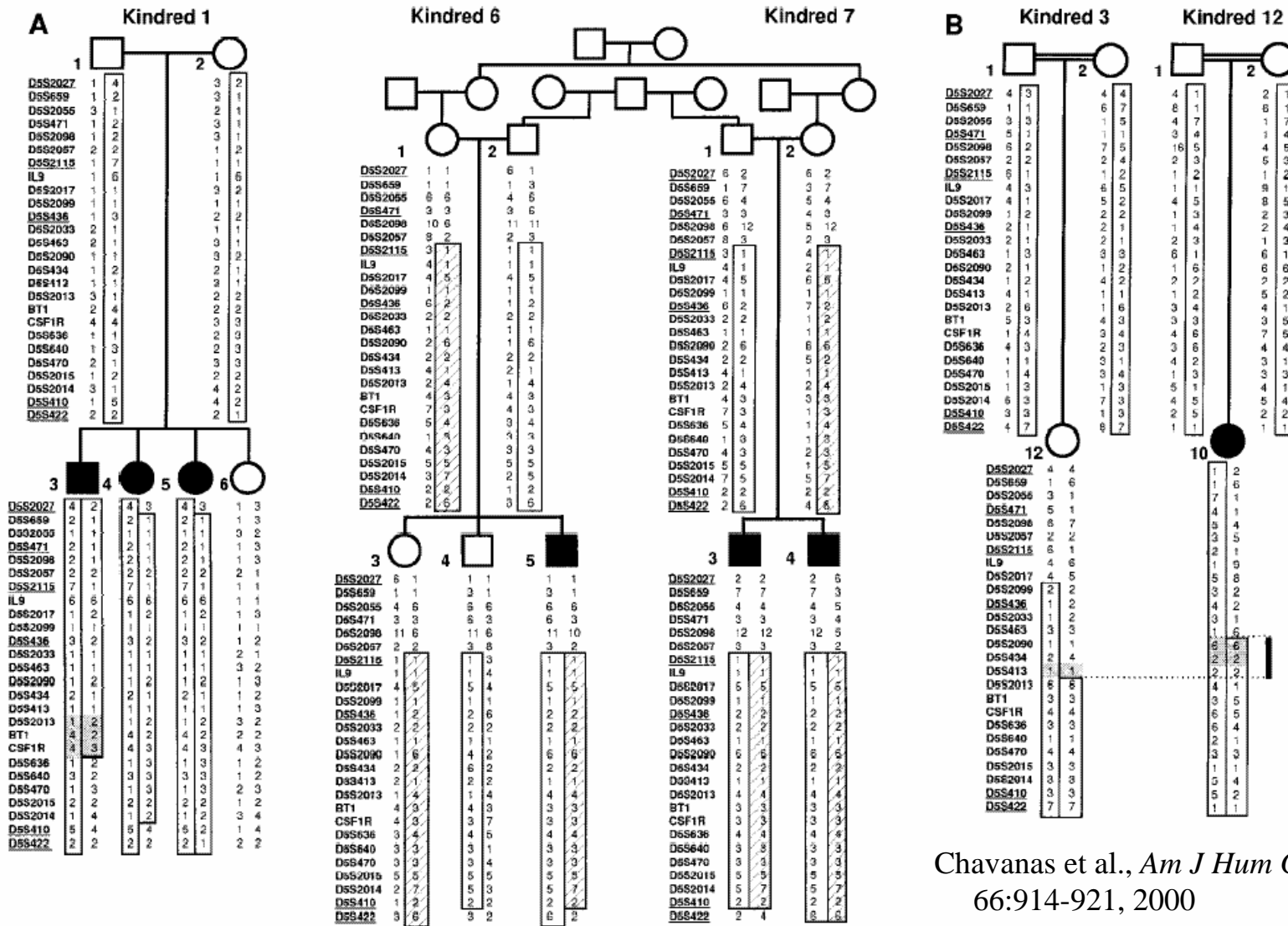
Chavanas et al., *Am J Hum Genet*, 66:914-921, 2000

Association – identical ancestral origin

Generation I - a disease-causing mutation occurs on a chromosome

Generation II - about 50% of the children receive the mutation and a surrounding chromosomal segment from the mutated founder

Generation III - the lengths of the segments originating from the mutated founder chromosome are shorter than or equal to those in GII.

Generation n - very short segments around the mutated locus conserved

Linkage vs Association

Linkage

1. Family-based

2. Matching/ethnicity generally unimportant

3. Few markers for genome coverage (300-400 STRs)

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. Poor for initial detection; good for fine-mapping

6. Powerful for common variants; rare variants generally impossible

Outline

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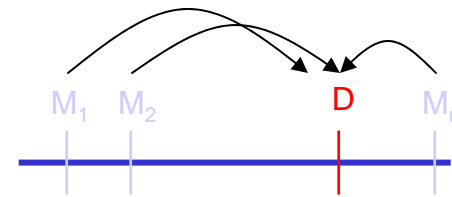 - Apparent association not related to genetic aetiology (most common outcome…)

Indirect and Direct Allelic Association

Direct Association

D

*

Measure disease relevance (*) directly, ignoring correlated markers nearby

Indirect Association & LD

M_1 M_2 D M_n

Assess trait effects on D via correlated markers (M_i) rather than susceptibility/etiologic variants.

Semantic distinction between
Linkage Disequilibrium: correlation between (any) markers in population
Allelic Association: correlation between marker allele and trait

How far apart can markers be to detect association?
Expected decay of linkage disequilibrium
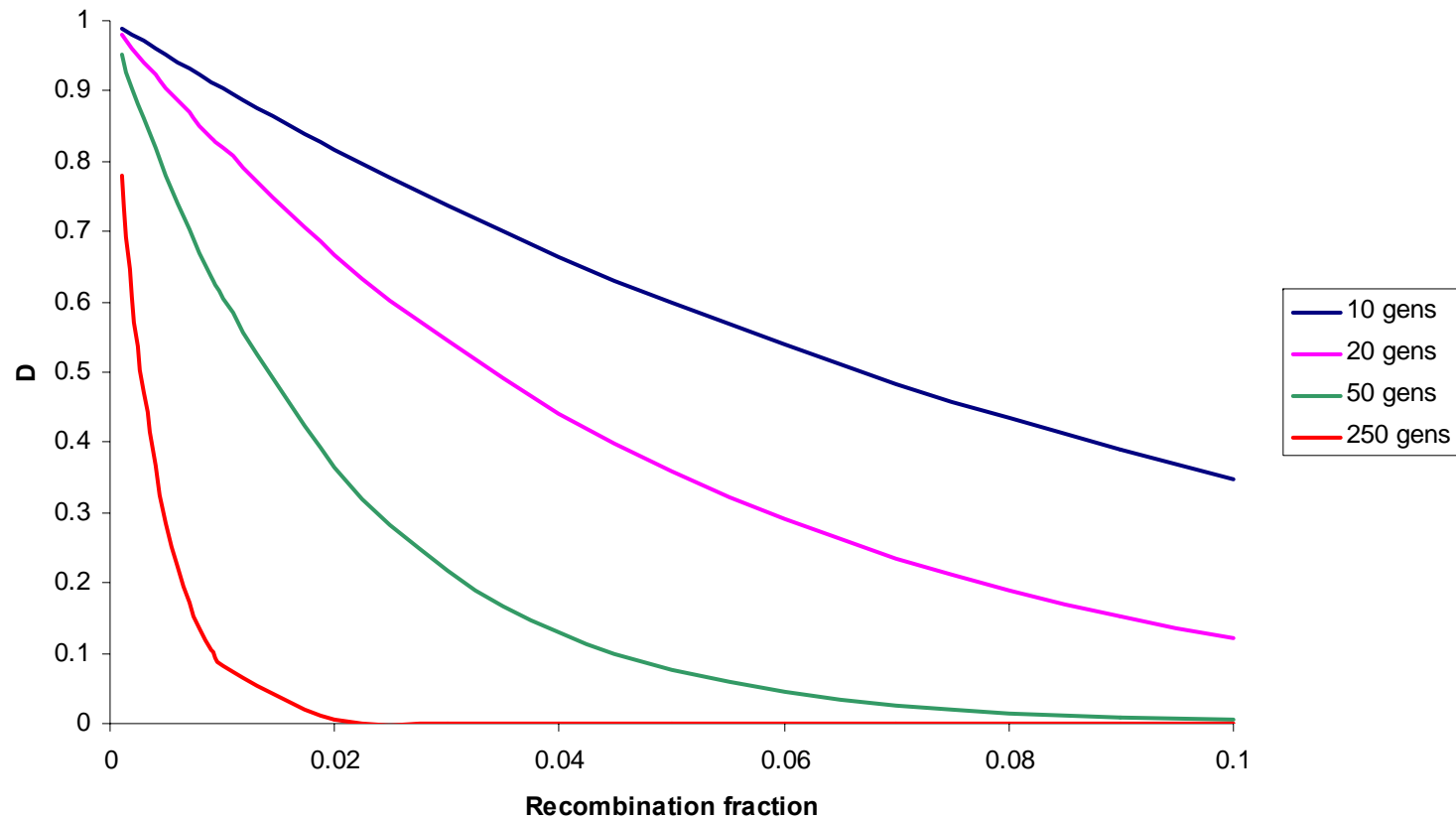
$$D_t = (1 - \theta)^t D_0$$
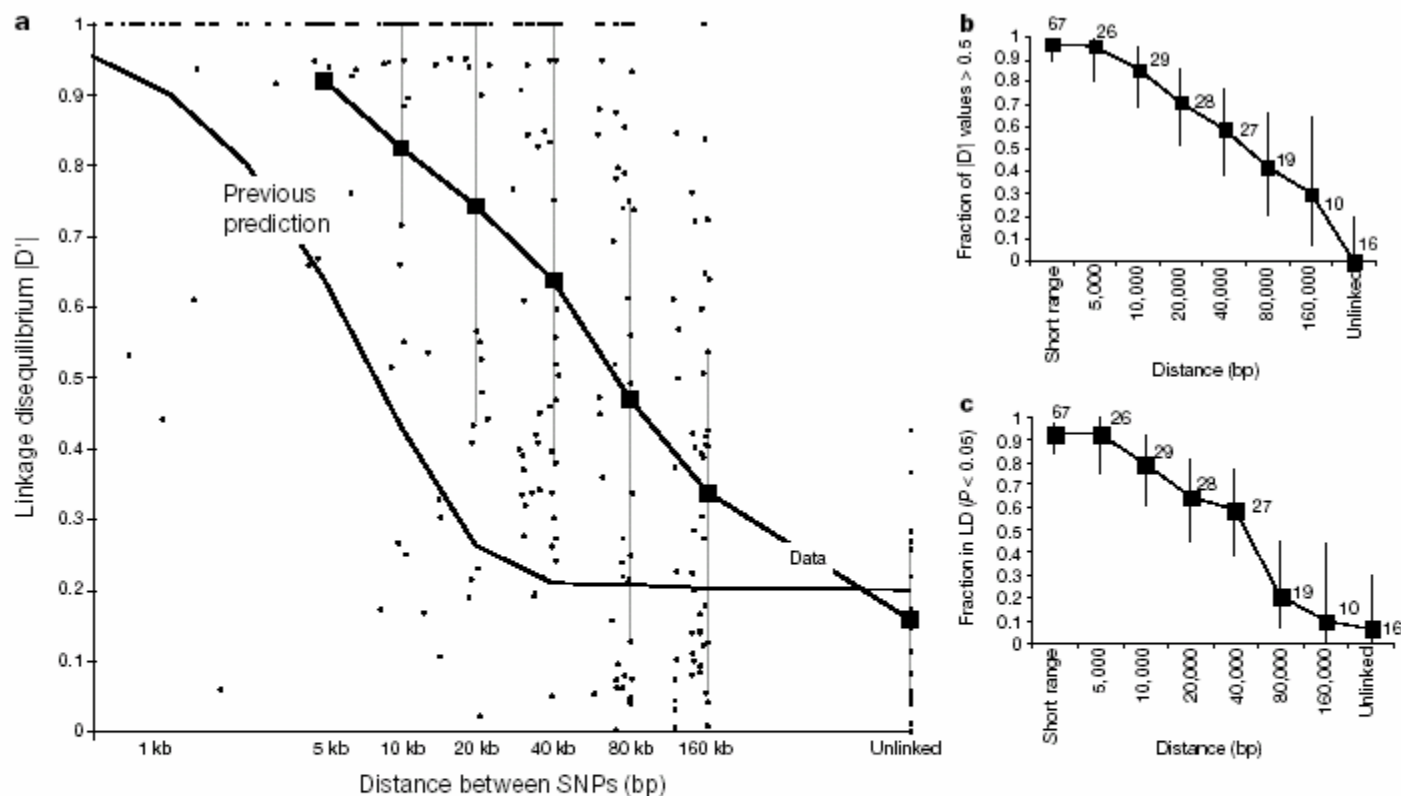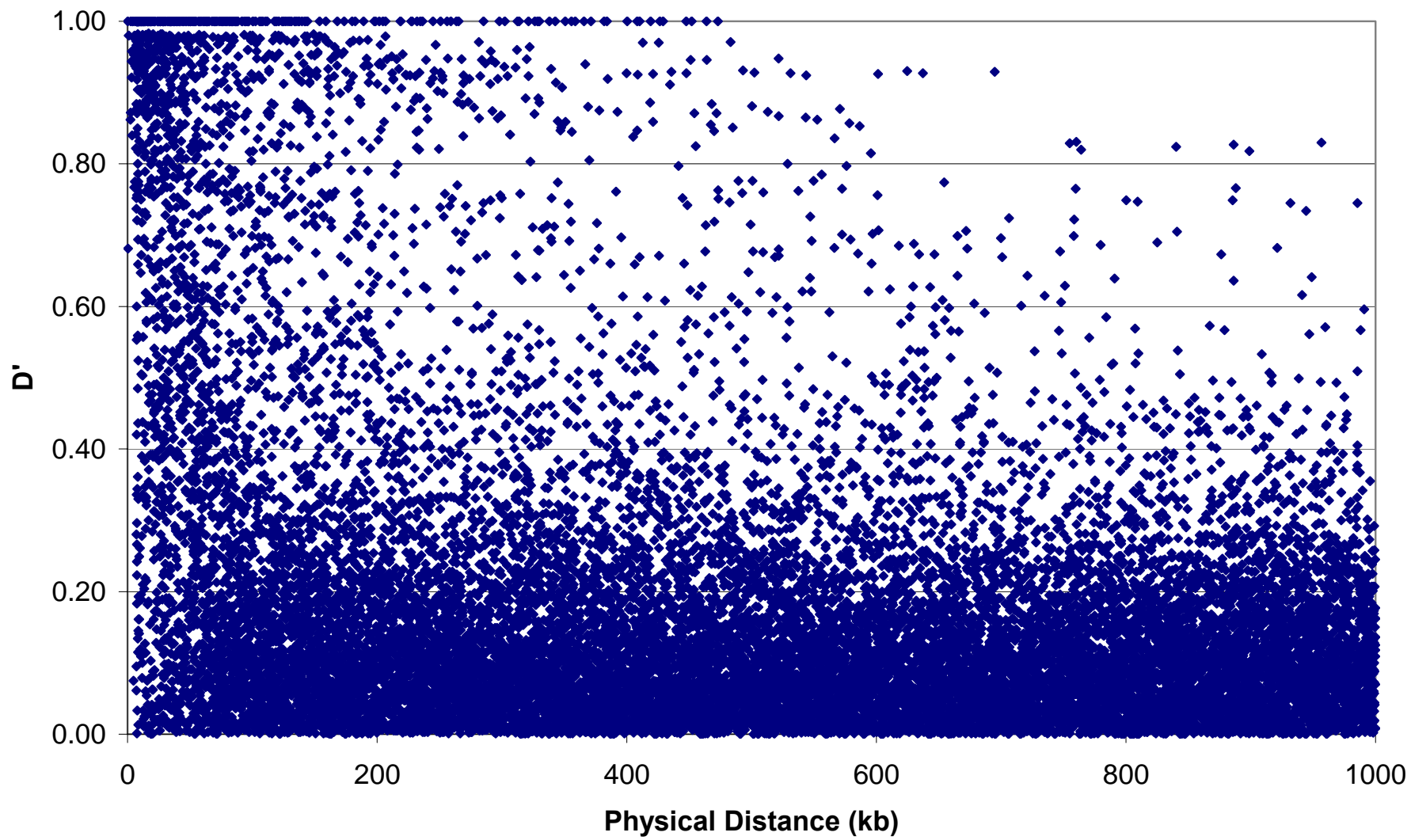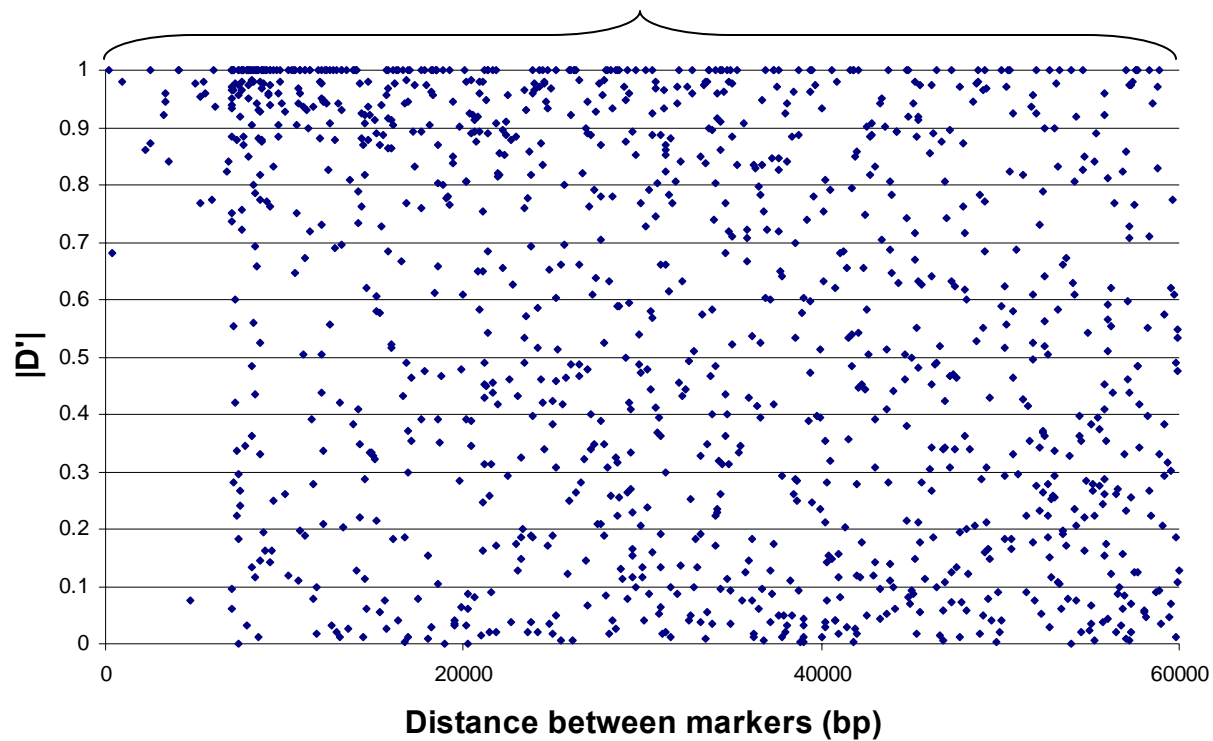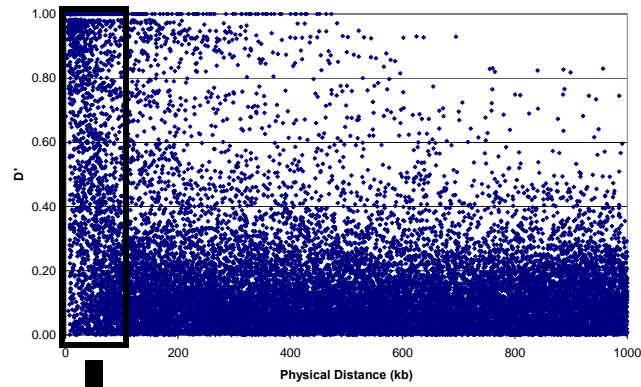
Decay of Linkage Disequilibrium

Figure 1 LD versus physical distance between SNPs. For each distance from the core SNP (Table 1), we chose the SNP with the largest number of copies of the minor allele for comparison to SNPs at other distances. At a given distance, all comparisons are independent. **a**, Average ID'I values for each distance separation ('Data'; dotted lines indicate the 25th and 75th percentiles), compared with a prediction[2] based on simulations (see Methods). ID'I values for shorter physical distances were calculated by looking within contiguously sequenced stretches of DNA containing at least two SNPs, and picking the two with the most minor alleles. Unlinked marker comparisons are obtained by comparing SNPs in the 40-kb bin in each row of Table 1 to those in the next row. **b**, **c**, Fraction of ID'I values greater than 0.5 (**b**) and proportion of significant ($P < 0.05$) associations (**c**) between two SNPs separated by a given distance (as assessed by a likelihood ratio test[10]). Bars indicate 95% central confidence intervals. The number of data points used to make the calculations are shown.

Variability in Pairwise LD on Chromosome 22

Variability in LD overwhelms the mean: D'

Average Levels of LD along chromosomes

CEPH
W.Eur
Estonian

Chr22

q11.21 q11.22 q11.23 q12.1 q12.2 q12.3 q13.1 q13.2 q13.31 q13.32 q13.33

Dawson et al
Nature 2002
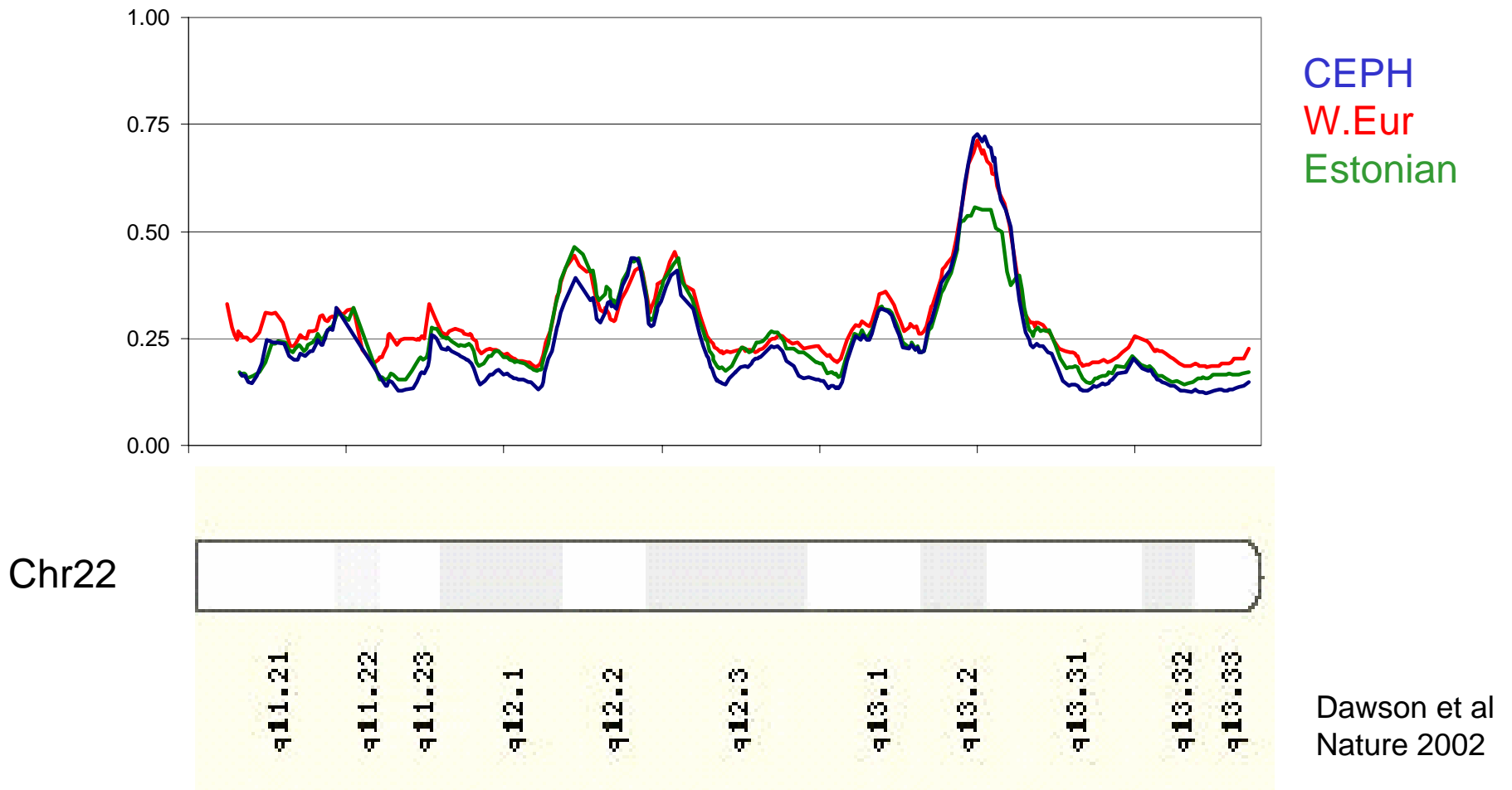
Characterizing Patterns of Linkage Disequilibrium

Average LD decay vs physical distance

Mean trends along chromosomes
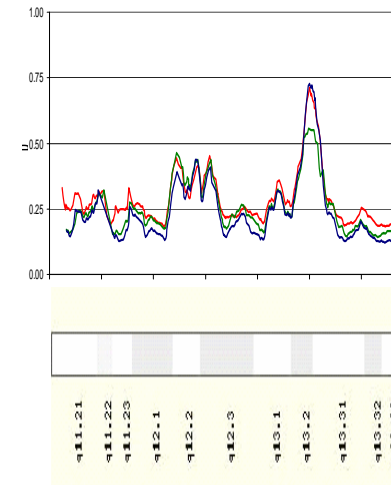
Haplotype Blocks
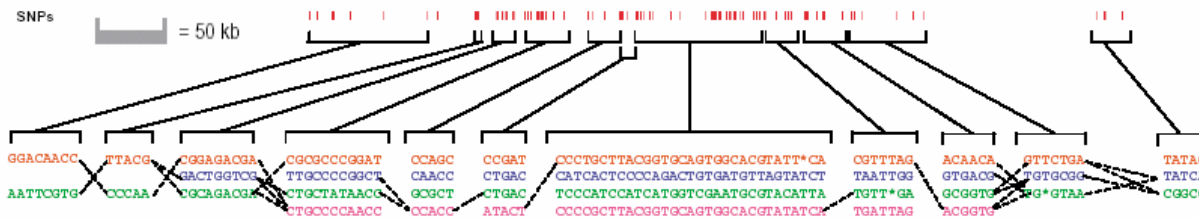
Linkage Disequilibrium Maps & Allelic Association
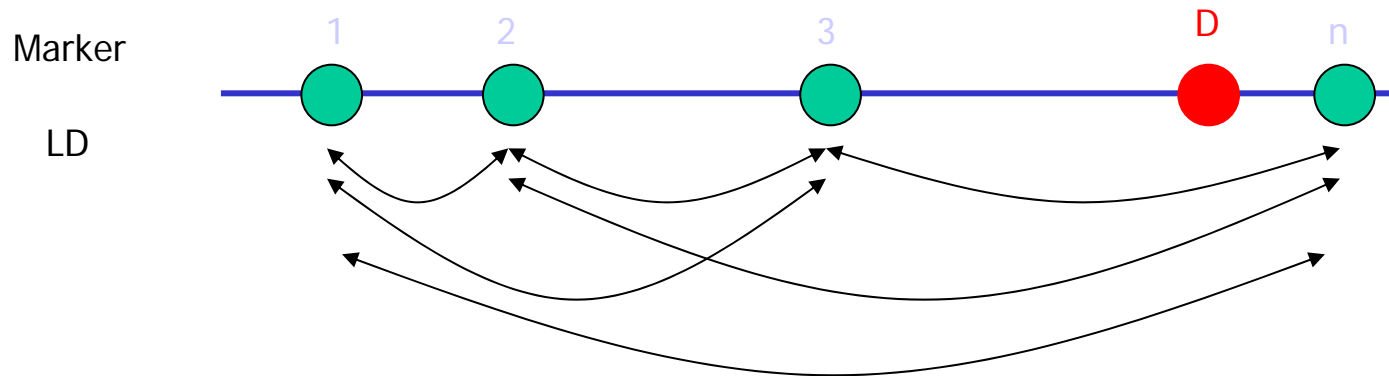
Primary Aim of LD maps: Use relationships amongst background markers (M_1, M_2, M_3, …M_n) to learn *something* about D for association studies

Something =   * Efficient association study design by reduced genotyping
                    * Predict approx location (fine-map) disease loci
                    * Assess complexity of local regions
                    * Attempt to quantify/predict underlying (unobserved)
                      patterns
                    ...

LD Patterns and Allelic Association

Type 1 diabetes and Insulin VNTR
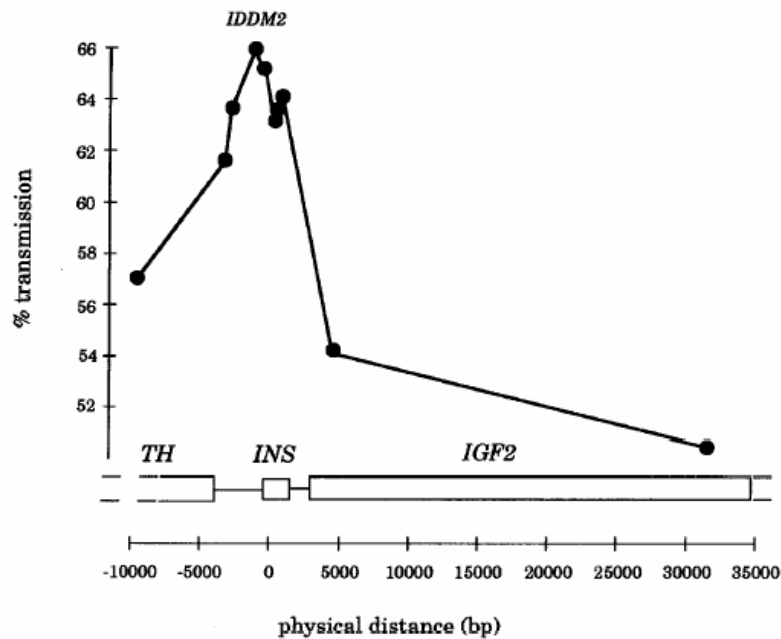
Bennett & Todd, *Ann Rev Genet*, 1996

Alzheimers and ApoE4
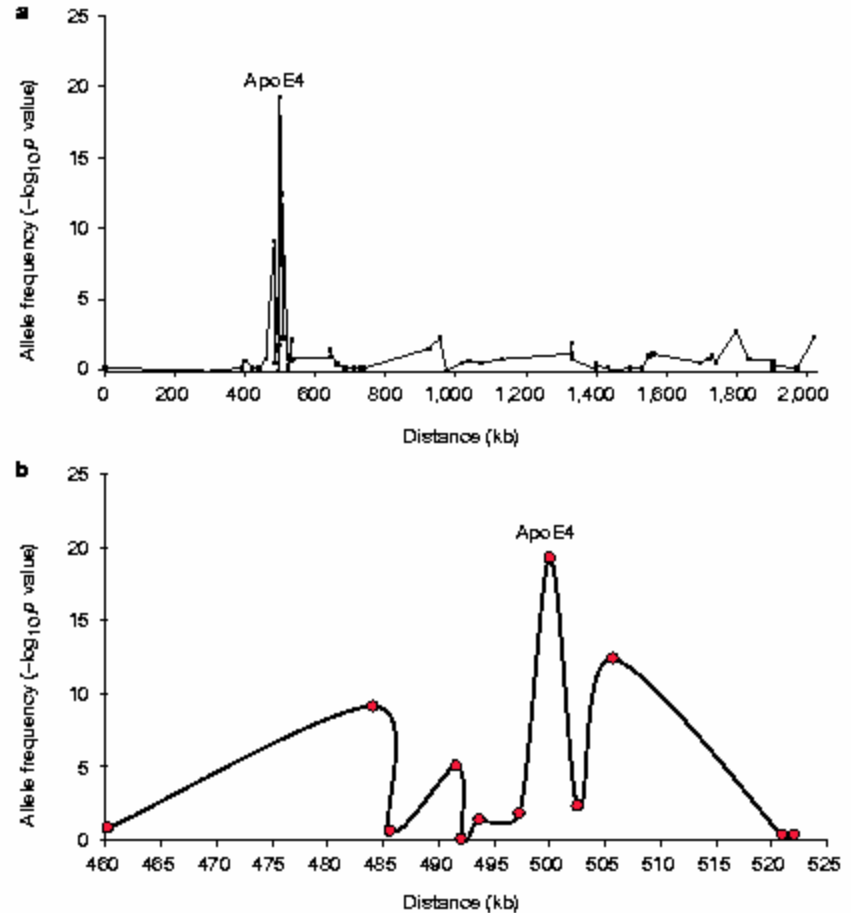
Roses, *Nature* 2000

The International HapMap Project

The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

....

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.

Building Haplotype Maps for Gene-finding

1. Human Genome Project
→ Good for consensus,
not good for individual
differences

Sept 01   Feb 02   April 04   Oct 04

2. Identify genetic variants
→ Anonymous with respect to
traits.

April 1999 – Dec 01

3. Assay genetic variants
→ Verify polymorphisms,
catalogue correlations
amongst sites
→ Anonymous with respect to
traits

Oct 2002 - present

HapMap Strategy

- Samples
  - Four populations, small samples
- Genotyping
  - 5 kb initial density across genome (600K markers)
  - Subsequent focus on low LD regions
  - Recent NIH RFA for deeper coverage

David Evans to discuss further

Hapmap validating millions of SNPs.
Are they the right SNPs?

Distribution of allele frequencies in public markers is <u>biased</u> toward common alleles

Expected frequency in population

Frequency of public markers
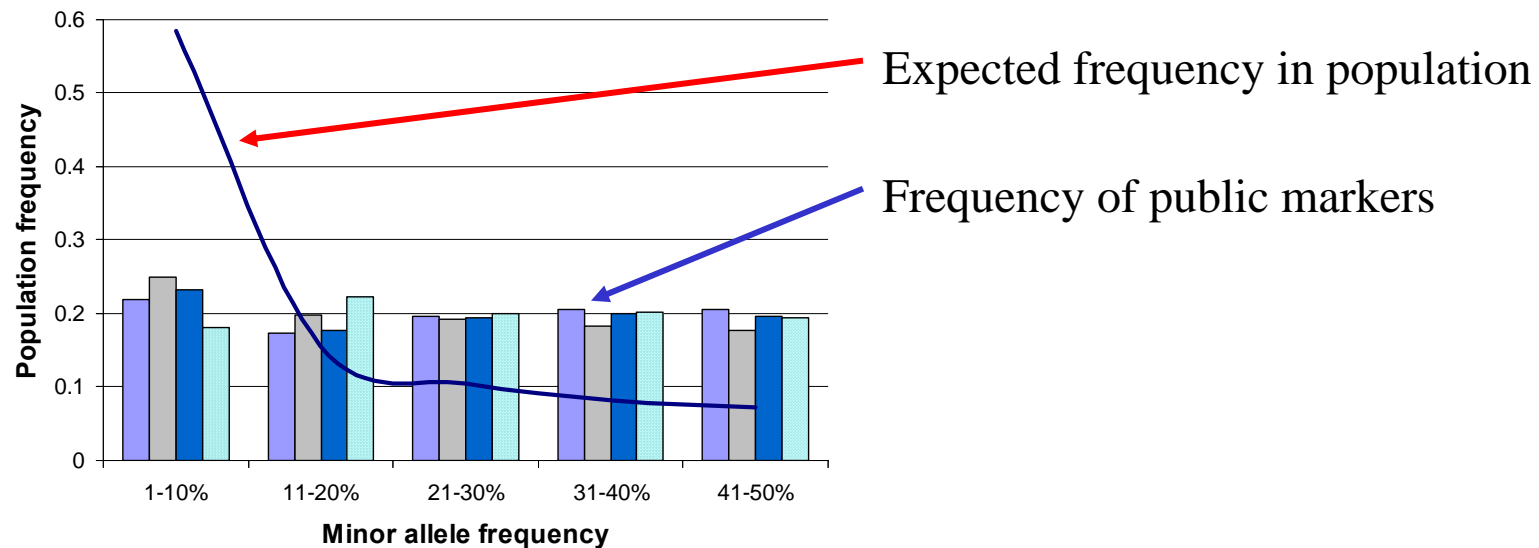
Phillips et al. Nat Genet 2003

Common-Disease Common-Variant Hypothesis

Common genes (alleles) contribute to inherited differences in common disease

Given recent human expansion, most variation is due to old mutations that have since become common rather than newer rare mutations.

Highly contentious debate in complex trait field

Common-Disease/Common-Variant

For

Against

Table I

Summary of allelic heterogeneity in support of the common disease/common variant or multiallele/multilocus hypotheses

(a) Common disease/common variant hypothesis

| Disease type	Locus	Allele	Trait	Frequency	Effect	Comments
Cardiovascular	APOE	ε4	Alzheimer disease	0.10–0.15 (Caucasian)	Early onset	Allele present in primates and all world populations; possible interaction with dietary fats; may account for 20% of Alzheimer disease
			Age-related macular degeneration	0.10–0.15	Decreased risk	Well-established protective effect on age-related macular degeneration
			Cardiovascular disease	0.10–0.15	Increased risk	Accounts for 10–16% of plasma cholesterol variance (western populations); increases risk of cardiovascular disease (odds ratio approximately 1.5)
	F5	R506Q	Venous thrombosis	0.02–0.08	Increased risk	Carriers have around 10% lifetime risk for significant venous thrombosis
Metabolic/ nutritional	PPARG	P12A	Type 2 diabetes mellitus	0.85 (Caucasian)	Increased risk	Relative risk 1.25
	CAPN10	Haplotypes 112 and 121	Type 2 diabetes mellitus	0.03–0.29 (low to high risk populations)	Increased risk in 121/112 haplotype heterozygotes	Complex risk haplotypes that may include several SNPs, including CAPN10-g.4852G/A (UCSNP-43)
	HFE	C282Y	Haemochromatosis	0.05 (Caucasian)	Around 40% risk for homozygotes	High frequency in Caucasians, low in Asiatics (suggesting admixture), so it may be a recent mutation (less than 50,000 years ago)
Cancer	ELAC2	S217L and A541T	Prostate cancer	0.30 and 0.04	Increased risk	Odds ratio 2.4–3.1
	BRCA2	N372H	Breast cancer	0.22–0.29 (Caucasian)	Increased risk	Relative risk = 1.31 for HH compared to NN genotypes
Infectious/ inflammatory	MHC class I	HLA-B*2702, 04, 05	Ankylosing spondylitis	0.09 (Caucasian)	Increased risk	Odds ratio approximately 170, mechanism unclear; also associated with reactive arthritis and uveitis; about 2% of B27-positive carriers develop ankylosing spondylitis
	MHC class II	DQB1*0302-DRB1*0401/DQB1*0201-DRB1*03	Type 1 diabetes mellitus	0.05 (European)	Increased risk	Around 10% of heterozygotes for these high risk haplotypes develop type 1 diabetes mellitus; relative risk approximately 20
	IL12B	3' UTR allele 1	Type 1 diabetes mellitus	0.79 (Caucasian)	Increased risk	Interaction with HLA; increased expression of IL12B in vitro
	G6PD	A− (V68M/N126D)	G6PD deficiency	Approximately 0.20 (West African)	Decreased risk of severe malaria	High allele frequency proposed to be due to balancing selection
	HBB	HbC (E6K)	Anaemia (homozygotes)	0.09 (West African)	Decreased risk of severe malaria	High allele frequency proposed to be due to balancing selection
	CCR5	Δ32-CCR5	HIV-1 transmission	0.09 (Caucasian)	Decreased HIV-1 transmission	Recent origin - estimated approximately 700 years ago [13]
Developmental	PDGFRA	Promoter H1/H2α haplotypes	Neural tube defect	0.23 (Caucasian)	Increased risk for sporadic neural tube defect	At least six polymorphic sites within each haplotype

Table I (continued)

(b) Multilocus/multiallele hypothesis

Disease type	Locus	Allele	Trait	Frequency	Effect	Comments
Cardiovascular	LDLR	> 735 alleles	Coronary artery disease	All rare, except in isolate or founder populations	Increased risk of coronary artery disease	
	APOB	> 24 alleles	Coronary artery disease	R3500Q 0.002, remainder rare	Increased risk of coronary artery disease	Single common R3500Q allele
Cancer	BRCA1	> 483 alleles	Familial breast-ovarian cancer	All rare, except in isolate or founder populations	Increased risk	
	BRCA2	> 404 alleles	Familial breast cancer	All rare, except in isolate or founder populations	Increased risk	Common N372H allele (frequency approximately 0.25) with relative risk 1.31
	MLH1	> 143 alleles	Hereditary non-polyposis colorectal cancer (HNPCC)	All rare	Increased risk	
	MSH2	> 108 alleles	Hereditary non-polyposis colorectal cancer (HNPCC)	All rare	Increased risk	
	P53	> 144 alleles	Multiple cancers	All rare	Increased risk	
Neurosensory	ABCA4	> 350 alleles	Stargardt disease, retinitis pigmentosa	Most rare, G863A allele approximately 0.014 (Europeans)	Increased risk	
	RHO	> 88 alleles	Retinitis pigmentosa, congenital stationary night blindness	All rare	Increased risk	
	GJB2	> 45 alleles	Non-syndromic deafness	Most rare, 30delG allele around 0.015 (Europeans)	Increased risk	30delG absent from non-European populations
Metabolic/ nutritional	CFTR	> 963 alleles	Cystic fibrosis	Most rare, ΔF508 accounts for approximately 70% of cystic fibrosis alleles in Caucasians	Increased risk	ΔF508 allele recent - estimated to have arisen 3,000 years ago [14]

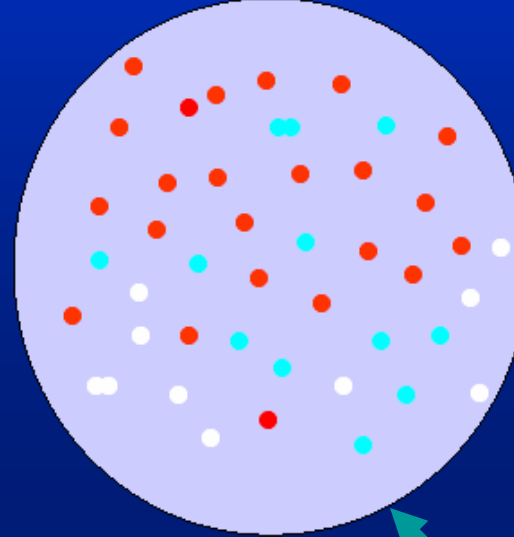Data are from the Online Mendelian Inheritance in Man database [30].
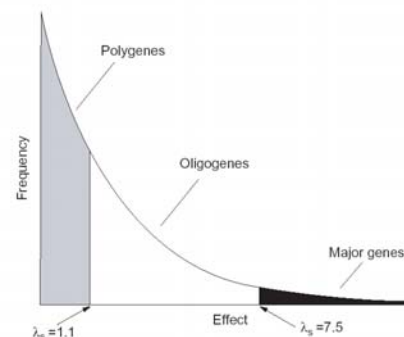
Wright & Hastie, Genome Biol 2001
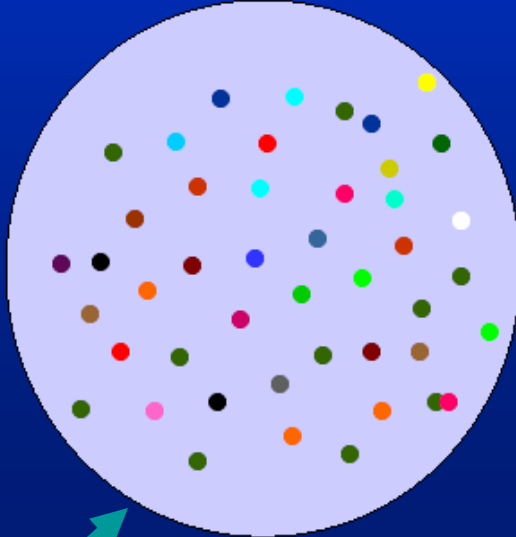
Taken from Joel Hirschorn presentation, www.chip.org

If this scenario, association studies will not work

If this scenario, properly designed association studies can work

The International HapMap Project

The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.

Deliverables: Sets of haplotype tagging SNPs

Haplotype Tagging for Efficient Genotyping

(a) Haplotype blocks

Block 1   Block 2   Block 3   Block 4   Block 5

(b) Mosaic chromosomes in population

Block   1   2   3   4   5   Cardon & Abecasis, TIG 2003

TRENDS in Genetics
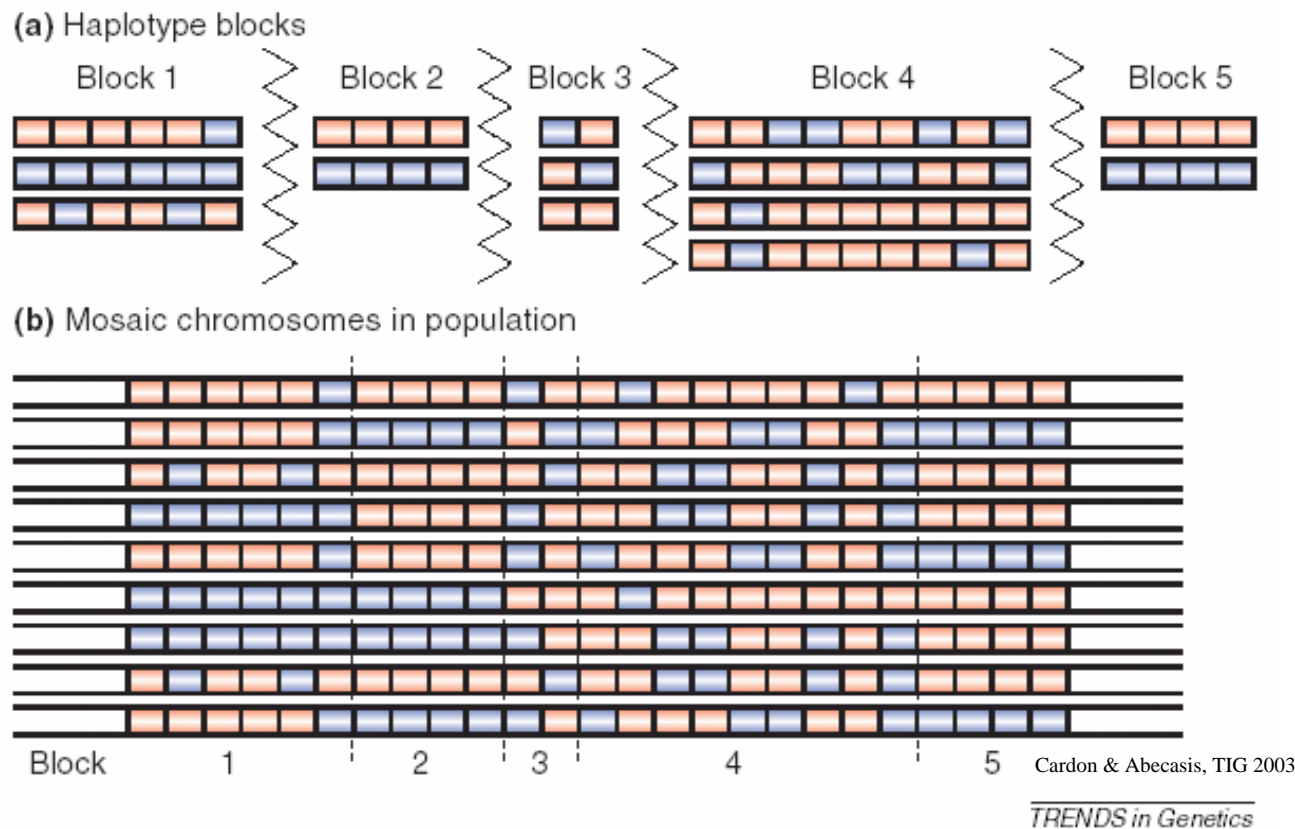
- Some genetic variants within haplotype blocks give redundant information

- A subset of variants, 'htSNPs', can be used to 'tag' the conserved haplotypes with little loss of information (Johnson et al., *Nat Genet*, 2001)

- ... Initial detection of htSNPs should facilitate future genetic association studies

Summary of Role of Linkage Disequilibrium on Association Studies

- **Marker characterization is becoming extensive and genotyping throughput is high**

- **Tagging studies will yield panels for immediate use**
  - Need to be clear about assumptions/aims of each panel

- **Density of eventual Hapmap probably cover much of genome in high LD, but not all**

Challenges

- Just having more markers doesn't mean that success rate will improve
- Expectations of association success via LD are too high. Hyperbole!
- **Need to show that this information can work in trait context**

Outline

Association Studies: Track Record

- Pubmed: Mar 2005. "Genetic association" gives 20,096 hits

- Q: How many are real?

- A: < 1%
  - Claims of "replicated genetic association" → 183 hits (0.9%)
  - Claims of "validated genetic association" → 80 hits (0.3%)

Association Study Outcomes

Reported p-values from association studies in *Am J Med Genet* or *Psychiatric Genet* 1997
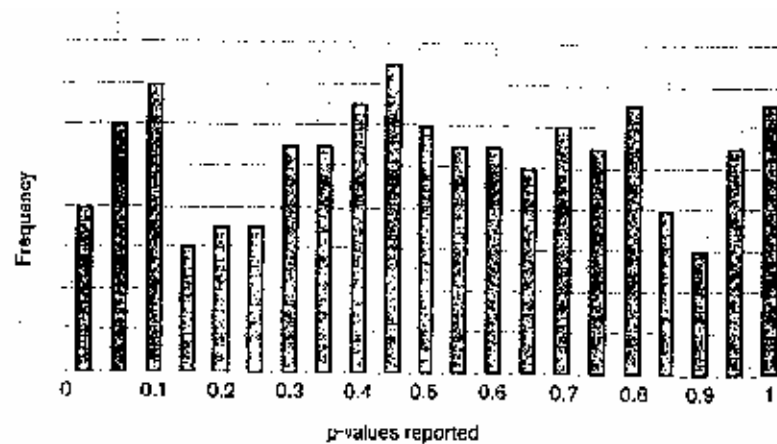
Terwilliger & Weiss, *Curr Opin Biotech*, 9:578-594, 1998

Why limited success with association studies?

1. Small sample sizes → results overinterpreted

2. Phenotypes are complex and not measured well. Candidate genes thus difficult to choose

3. Allelic/genotypic contributions are complex. Even true associations difficult to see.

4. Population stratification has led clouded true/false positives

Sample Size Matters

PPARγ and NIDDM

ACE and MI

Figure 3: Meta-analysis of published studies of the association between DD genotype and myocardial infarction
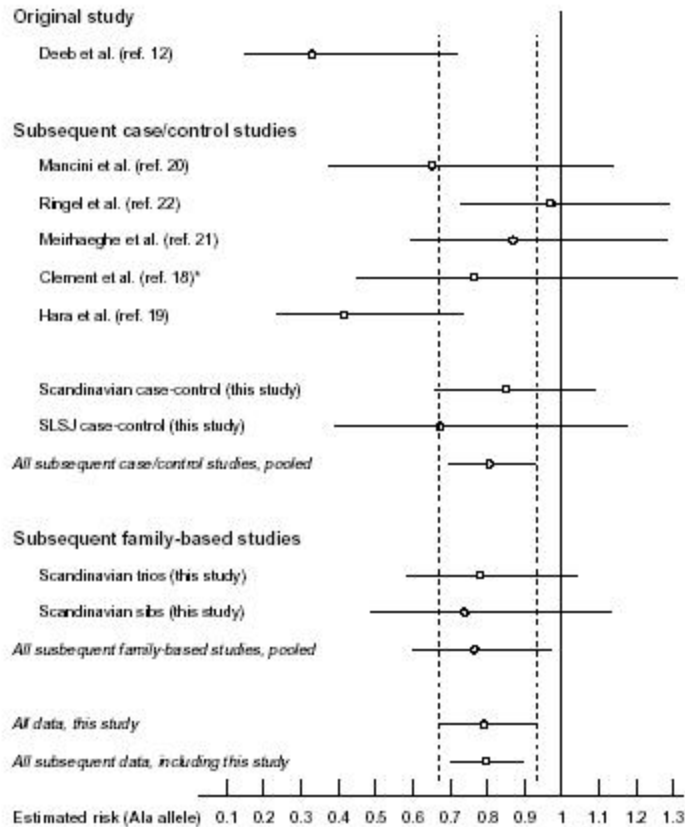
Altshuler et al *Nat Genet* 2000
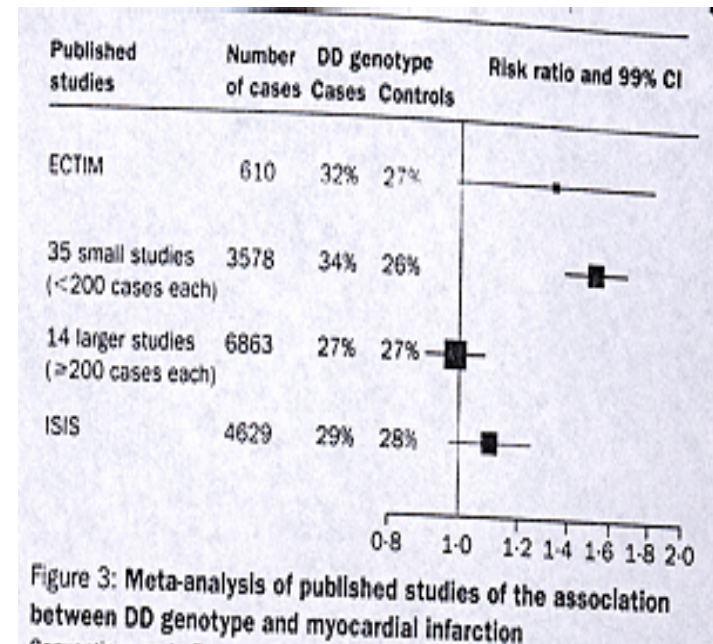
Keavney et al *Lancet* 2000

Phenotypes are Complex

Weiss & Terwilliger, Nat Genet, 2000

Many Forms of Heterogeneity

(a) Model 1 : allelic homogeneity

(b) Model 2: allelic heterogeneity

(c) Model 3: multiple mutations in multiple genes
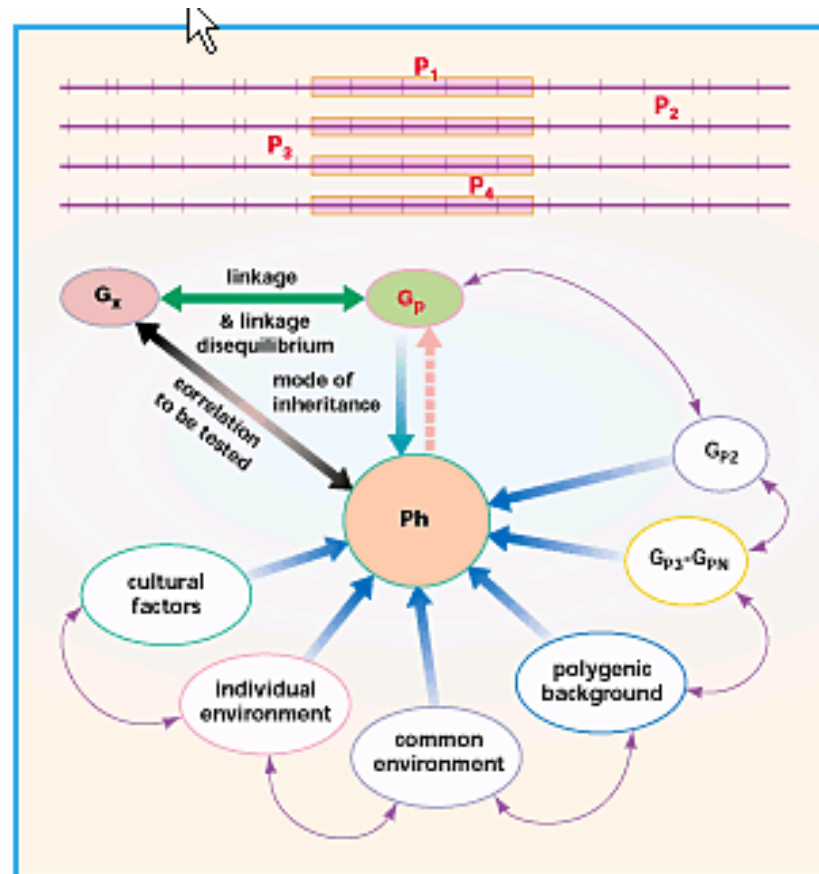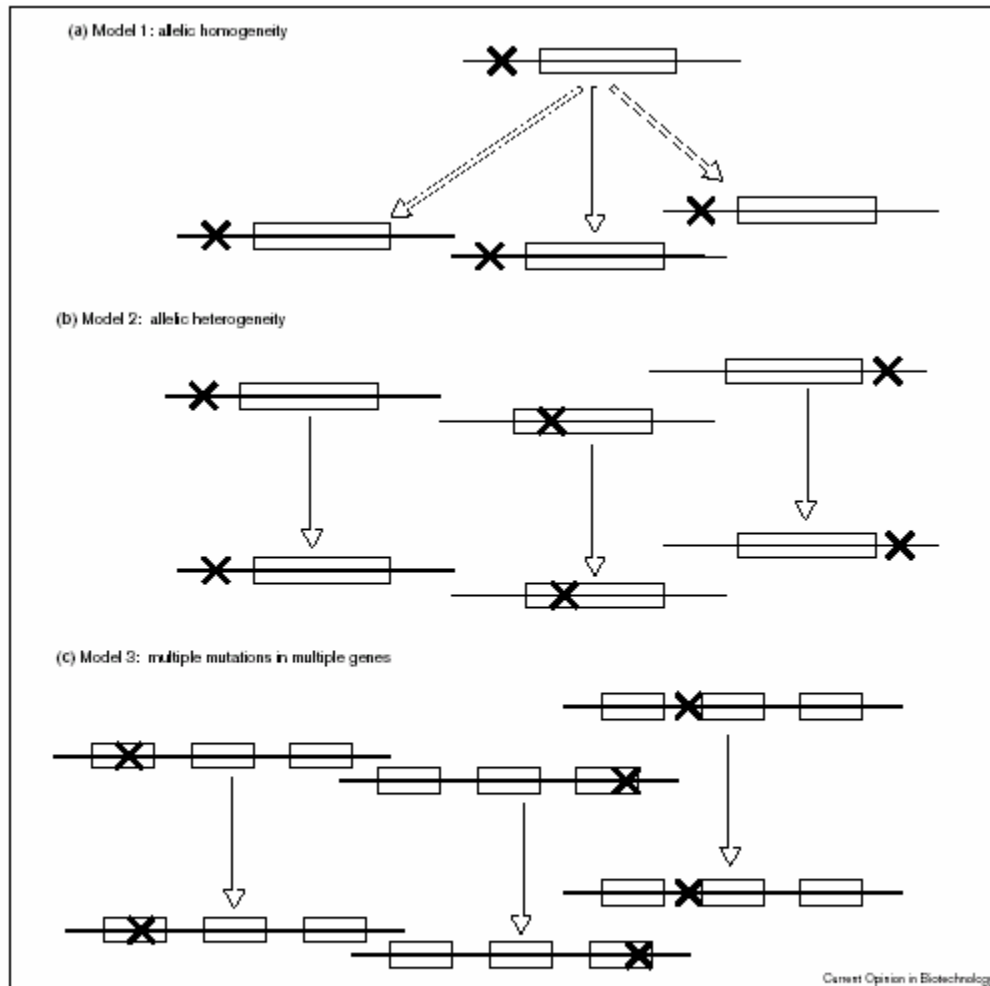
Current Opinion in Biotechnology

Three simple models for the allelic complexity of genetic disease are shown. (a) In Model 1, all disease-predisposing alleles at a given locus are identical by descent in the population – having derived from some common ancestor. In this situation, there is expected to be a conserved haplotype around the disease allele, which is shared by all carriers in the population many generations later. (b) Model 2 shows the case of allelic heterogeneity, in which multiple different allelic variants can each predispose to the phenotype. Thus among individuals with one of these 'D' alleles, there will be an assortment of haplotype backgrounds. The more heterogeneity, the less LD. (c) Model 3 shows the situation for multiple 'D' alleles in different genes. These genes may be linked (as shown) or unlinked.

Terwilliger & Weiss, Curr Opin Biotechnol, 1998

Main Blame

Why do association studies have such a spotted history in human genetics?

Blame: Population stratification

Analysis of mixed samples having different allele frequencies is a primary concern in human genetics, as it leads to false evidence for allelic association.
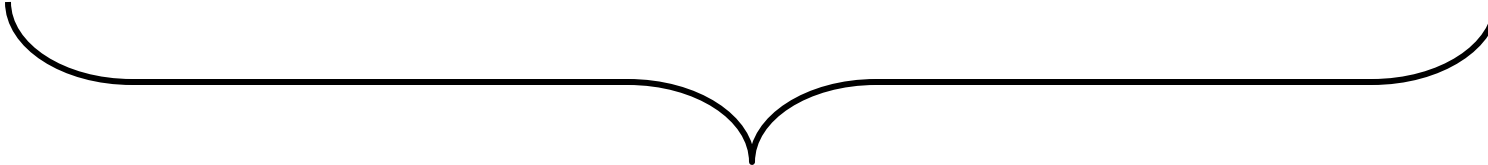
Population Stratification

- Leads to spurious association

- Requirements:
  - Group differences in allele frequencies AND
  - Group differences in outcome

- In epidemiology, this is a classic matching problem, with genetics as a confounding variable

Most oft-cited reason for lack of association replication

Population Stratification

Sample 'A'

	M	m	Freq.
Affected	50	50	.10
Unaffected	450	450	.90
	.50	.50	

χ^2_1 is n.s.

$+$

Sample 'B'

	M	m	Freq.
Affected	1	9	.01
Unaffected	99	891	.99
	.10	.90	

χ^2_1 is n.s.

	M	m	Freq.
Affected	51	59	.055
Unaffected	549	1341	.945
	.30	.70	

$$\chi^2_1 = 14.84, p < 0.001$$

Spurious Association

Population Stratification: Real Example

Full heritage American Indian Population

$\text{Gm}^{3;5,13,14}$ **+**   **-**

      ~1%   **~99%**

(NIDDM Prevalence ≈ 40%)

Caucasian Population

$\text{Gm}^{3;5,13,14}$ **+**   **-**

      ~66%   **~34%**

(NIDDM Prevalence ≈ 15%)

Study without knowledge of genetic background:

$\text{Gm}^{3;5,13,14}$ haplotype	Cases	Controls
+	7.8%	29.0%
-	92.2%	71.0%

OR=0.27
95%CI=0.18 to 0.40

Proportion with NIDDM by heritage and marker status

Index of Indian Heritage	$\text{Gm}^{3;5,13,14}$ haplotype	
	+	-
0	17.8%	19.9%
4	28.3%	28.8%
8	35.9%	39.3%

Reviewed in Cardon & Palmer, Lancet 2003

'Control' Samples in Human Genetics ≤ 2000

- Because of fear of stratification, complex trait genetics turned away from case/control studies
  - *fear may be unfounded*
- Moved toward family-based controls (flavour is TDT: transmission/disequilibrium test)
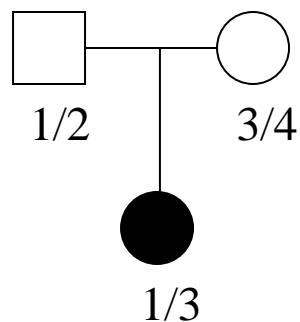
"Case"   = transmitted alleles
          = 1 and 3

"Control" = untransmitted alleles
          = 2 and 4

TDT Advantages/Disadvantages

Advantages

Robust to stratification
Genotyping error detectable via Mendelian inconsistencies
Estimates of haplotypes possible

Disadvantages

Detection/elimination of genotyping errors causes bias (Gordon et al., 2001)
Uses only heterozygous parents
Inefficient for genotyping

> 3 individuals yield 2 founders: 1/3 information not used

Can be difficult/impossible to collect

> Late-onset disorders, psychiatric conditions, pharmacogenetic applications

Association studies < 2000: TDT

- TDT virtually ubiquitous over past decade

  Grant, manuscript referees & editors mandated design

- View of case/control association studies greatly
  diminished due to perceived role of stratification

Association Studies 2000+ :
Return to population

- Case/controls, using extra genotyping
  - +families, when available

Detecting and Controlling for
Population Stratification with Genetic Markers

Idea

• Take advantage of availability of large N genetic markers

• Use case/control design

• Genotype genetic markers across genome
   (Number depends on different factors)

• Look if any evidence for background population substructure exists and account for it

• Shaun Purcell to describe in 'Genomic Control' lecture

Outline

Current Association Study Challenges

1) Genome-wide screen or candidate gene

Genome-wide screen

- Hypothesis-free
- High-cost: large genotyping requirements
- Multiple-testing issues
  - Possible many false positives, fewer misses

Candidate gene

- Hypothesis-driven
- Low-cost: small genotyping requirements
- Multiple-testing less important
  - Possible many misses, fewer false positives

Current Association Study Challenges

2) What constitutes a replication?

GOLD Standard for association studies

Replicating association results in different laboratories is often seen as most compelling piece of evidence for 'true' finding

But…. in any sample, we measure
> Multiple traits
> Multiple genes
> Multiple markers in genes

and we analyse all this using multiple statistical tests

<p style="text-align:center">What is a true replication?</p>

What is a true replication?

Replication Outcome

- Association to same trait, but different gene
- Association to same trait, same gene, different SNPs (or haplotypes)
- Association to same trait, same gene, same SNP – but in opposite direction (protective ⬅➡ disease)
- Association to different, but correlated phenotype(s)
- No association at all

Explanation

- Genetic heterogeneity

- Allelic heterogeneity

- Allelic heterogeneity/popln differences

- Phenotypic heterogeneity

- Sample size too small

Measuring Success by Replication

- Define objective criteria for what is/is not a replication *in advance*

- Design initial and replication study to have enough power
  - 'Lumper': use most samples to obtain robust results in first place
    - Great initial detection, may be weak in replication

  - 'Splitter': Take otherwise large sample, split into initial and replication groups
    - One good study → two bad studies.
    - Poor initial detection, poor replication

Current Association Study Challenges

3) Do we have the best set of genetic markers

There exist 6 million putative SNPs in the public domain. Are they the right markers?

Allele frequency distribution is biased toward common alleles

Expected frequency in population

Frequency of public markers
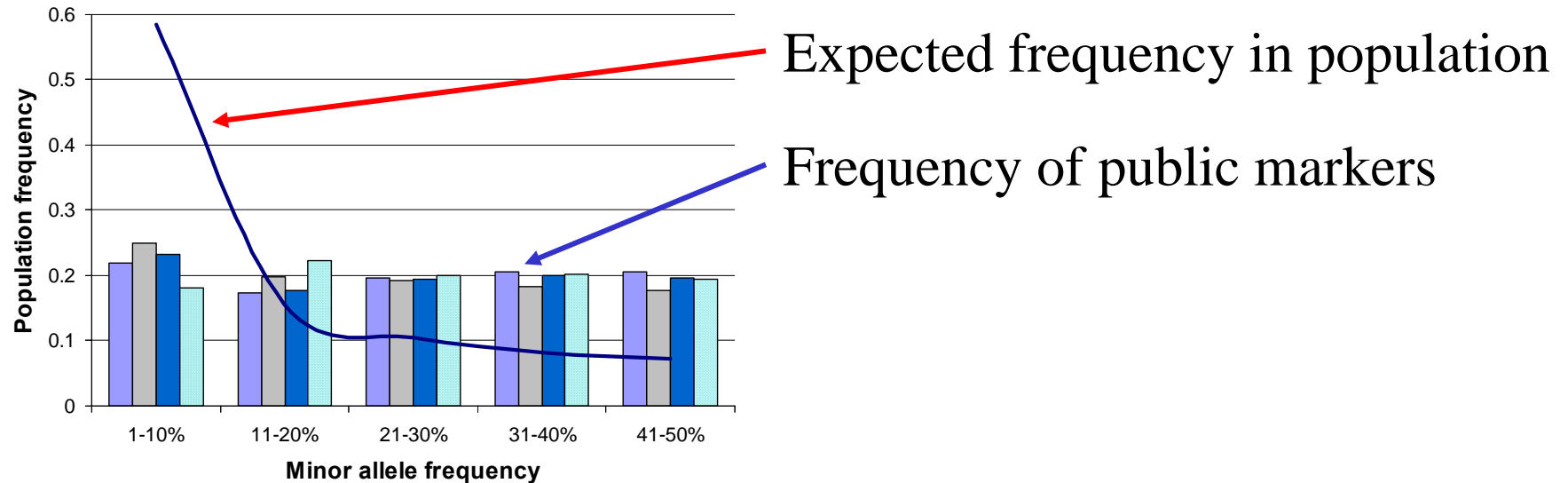
Current Association Study Challenges

3) Do we have the best set of genetic markers

Table 1 | **Priorities for single-nucleotide-polymorphism selection**

Type of variant	Location	Functional effect	Frequency in genome
Nonsense	Coding sequence	Premature termination of amino-acid sequence	Very low
Missense/ non-synonymous (non-conservative)	Coding sequence	Changes an amino acid in protein to one with different properties	Low
Missense/ non-synonymous (conservative)	Coding sequence	Changes an amino acid in protein to one with similar properties	Low
Insertions/deletions (frameshift)	Coding sequence	Changes the frame of the protein-coding region, usually with very negative consequences for the protein	Low
Insertions/deletions (in frame)	Coding or non-coding	Changes amino-acid sequence	Low
Sense/synonymous	Coding sequence	Does not change the amino acid in the protein — but can alter splicing	Medium
Promoter/regulatory region	Promoter, 5' UTR, 3' UTR	Does not change the amino acid, but can affect the level, location or timing of gene expression	Low to medium
Splice site/intron–exon boundary	Within 10 bp of the exon	Might change the splicing pattern or efficiency of introns	Low
Intronic	Deep within introns	No known function, but might affect expression or mRNA stability	Medium
Intergenic	Non-coding regions between genes	No known function, but might affect expression through enhancer or other mechanisms	High

Tabor et al, Nat Rev Genet 2003

Greatest power comes from markers that match allele freq with trait loci

Disease Allele Frequency	Marker Allele Frequency				
	0.1	0.3	0.5	0.7	0.9
0.1	**248**	626	1306	2893	10830
0.3	1018	**238**	466	996	3651
0.5	2874	702	**267**	556	2002
0.7	9169	2299	925	**337**	1187
0.9	73783	18908	7933	3229	**616**

$\lambda s = 1.5$, $\alpha = 5 \times 10^{-8}$, Spielman TDT
(Müller-Myhsok and Abel, 1997)

Current Association Study Challenges

4) Integrating the sampling, LD and genetic effects

Questions that don't stand alone:

How much LD is needed to detect complex disease genes?

What effect size is big enough to be detected?

How common (rare) must a disease variant(s) be to be identifiable?

What marker allele frequency threshold should be used to find complex disease genes?

Complexity of System

•In any indirect association study, we measure marker alleles that are *correlated* with trait variants…

  We <u>do not</u> measure the trait variants themselves

•But, for study design and power, we concern ourselves with frequencies and effect sizes *at the trait locus*….

  This can only lead to underpowered studies and inflated expectations

•We <u>should</u> concern ourselves with the apparent effect size at the marker, which results from

  1) difference in frequency of marker and trait alleles
  2) LD between the marker and trait loci
  3) effect size of trait allele

Decay in power to detect effect (α=0.001) by MAF and LD in 1000 cases + 1000 controls
- Crohn's & NOD2 (DAF = 0.06) -
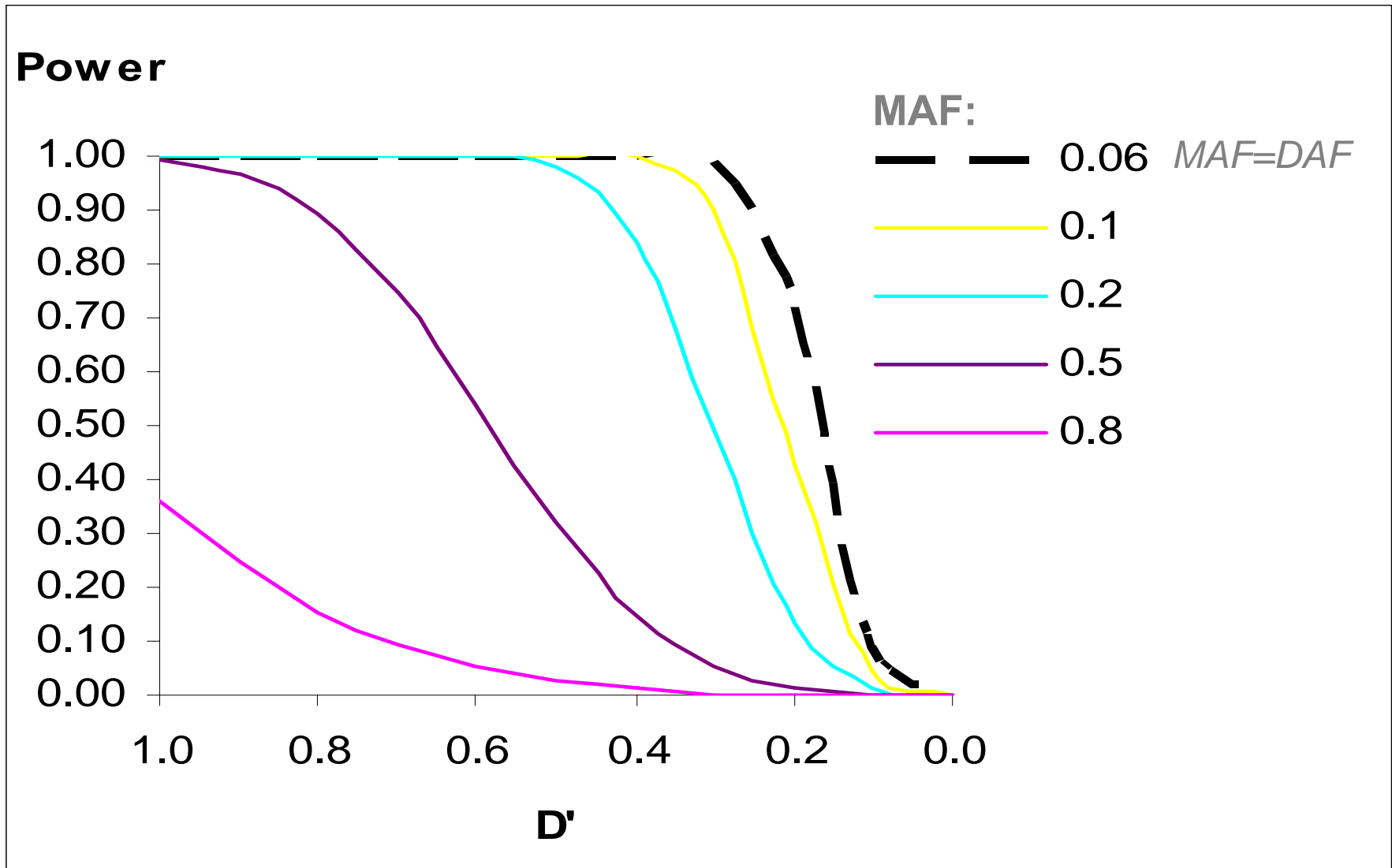
Decay in power to detect effect (α=0.001) by MAF and LD in 5000 cases + 5000 controls
- Type II Diabetes & PPARG (DAF = 0.85) -
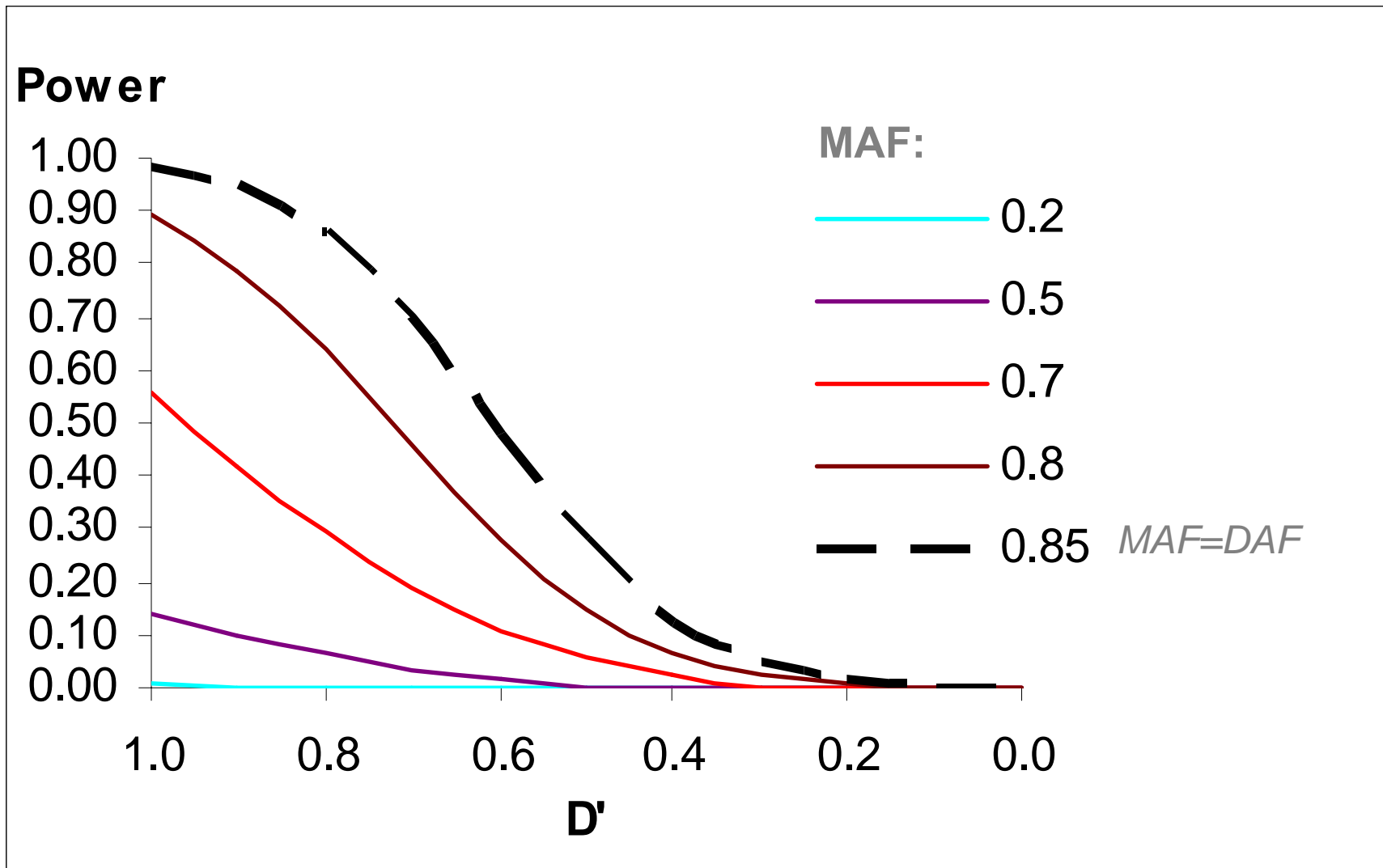
Practical Implications of Allele Frequencies

- 'Strongest argument for using common markers is not CD-CV. It is practical:

  For small effects, common markers are the only ones for which sufficient sample sizes can be collected

$\Rightarrow$ There are situations where indirect association analysis will not work
  - Discrepant marker/disease freqs, low LD, heterogeneity, …
  - Linkage approach may be only genetics approach in these cases

At present, no way to know when association will/will not work
  - Balance with linkage

Current Association Study Challenges

5) How to analyse the data

- **Allele based test?**
  - 2 alleles → 1 df
    - $E(Y) = a + bX$ $X = 0/1$ for presence/absence

- **Genotype-based test?**
  - 3 genotypes → 2 df
    - $E(Y) = a + b_1A + b_2D$ $A = 0/1$ additive (hom); $W = 0/1$ dom (het)

- **Haplotype-based test?**
  - For M markers, 2^M possible haplotypes → $2^M - 1$ df
    - $E(Y) = a + \Sigma bH$ H coded for haplotype effects

- **Multilocus test?**
  - Epistasis, G x E interactions, many possibilities

Current Association Study Challenges

6) Multiple Testing

- Candidate genes: a few tests (probably correlated)

- Linkage regions: 100's – 1000's tests (some correlated)

- Whole genome association: 100,000s – 1,000,000s tests (many correlated)

- What to do?
  – Bonferroni (conservative)
  – False discovery rate?
  – Permutations?
  ….Area of active research

Despite challenges: upcoming association studies hold some promise

- Large, epidemiological-sized samples emerging
  - ISIS, Biobank UK, Million Women's Study, …

- Availability of millions of genetic markers
  - Genotyping costs decreasing rapidly
    - Cost per SNP: 2001 ($0.25) → 2003 ($0.10) → 2004 ($0.01)

- Background LD patterns being characterized
  - International HapMap and other projects

Realistic expectations and better design should yield success

Genetic analysis of genome-wide variation in human gene expression

Michael Morley[1,3]*, Cliona M. Molony[2]*, Teresa M. Weber[1,3], James L. Devlin[2], Kathryn G. Ewens[2], Richard S. Spielman[2] & Vivian G. Cheung[1,2,3]

[1]Department of Pediatrics and [2]Department of Genetics, University of Pennsylvania,
[3]The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA

*These authors contributed equally to this work

- Examined expression levels of ~8000 genes on CEPH families
  - Used expression levels as 'phenotypes'
  - Linked expression phenotypes with CEPH microsatellites
- Found evidence for linkage for many phenotypes
- Follow-up SNP genotyping also showed some association

- Found many cis- linkages (linkage region overlaps location of gene whose expression is phenotype), but also many trans

Genome-wide Association

- Most of the CEPH families phenotyped by Cheung are also being genotyped by HapMap

- Can integrate all genotypes for the 1 million current HapMap SNPs with Cheung expression phenotypes
  - Estimate heritabilities, examine 100 most heritable expression traits
  - Genome-wide linkage analysis (+4500 STRs)
  - Genome-wide association analysis (1 million SNPs)
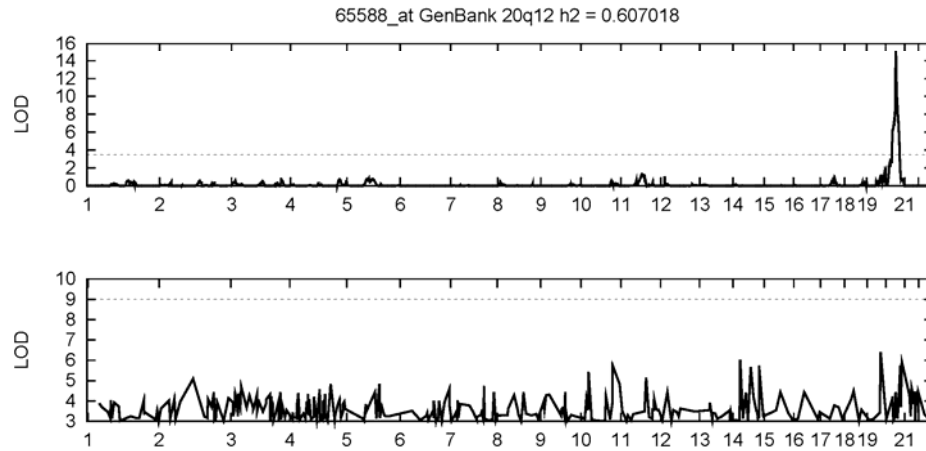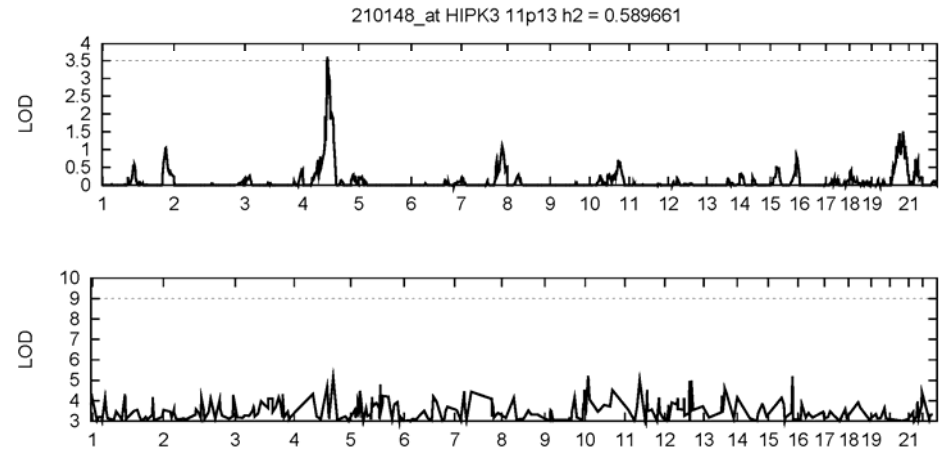
No Linkage
No Association

201938_at CDK2AP1 12q24.31 h2 = 0.582474

33646_g_at GM2A 5q31.3-q33.1 h2 = 0.631579
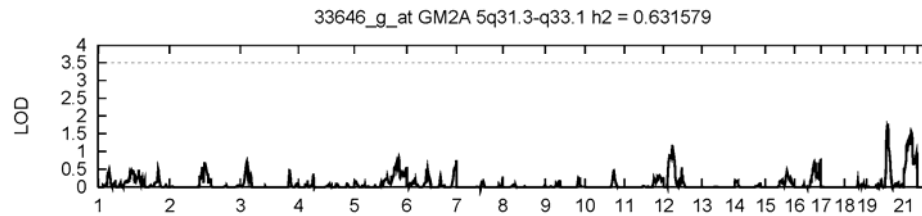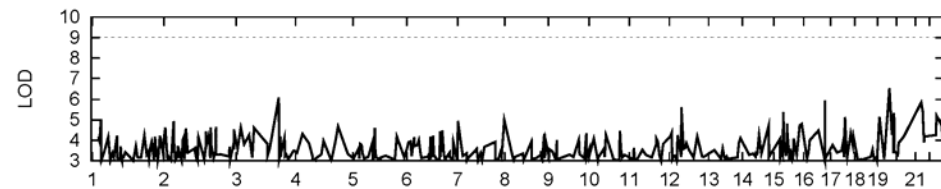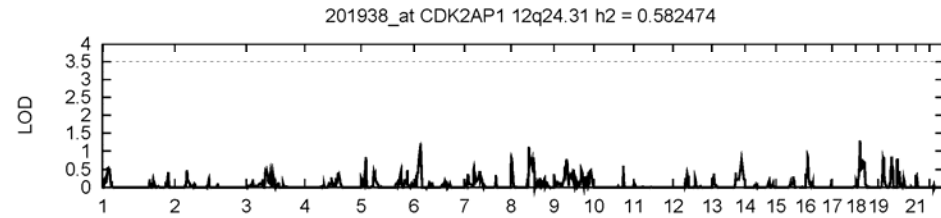
Linkage genome scan
4,000 highly polymorphic markers
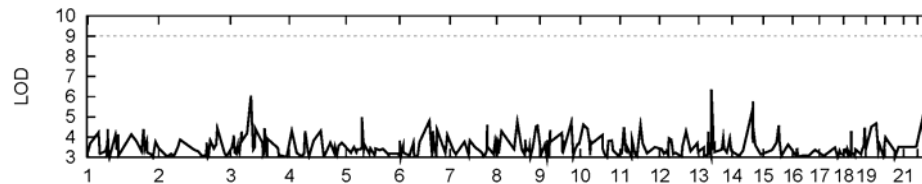
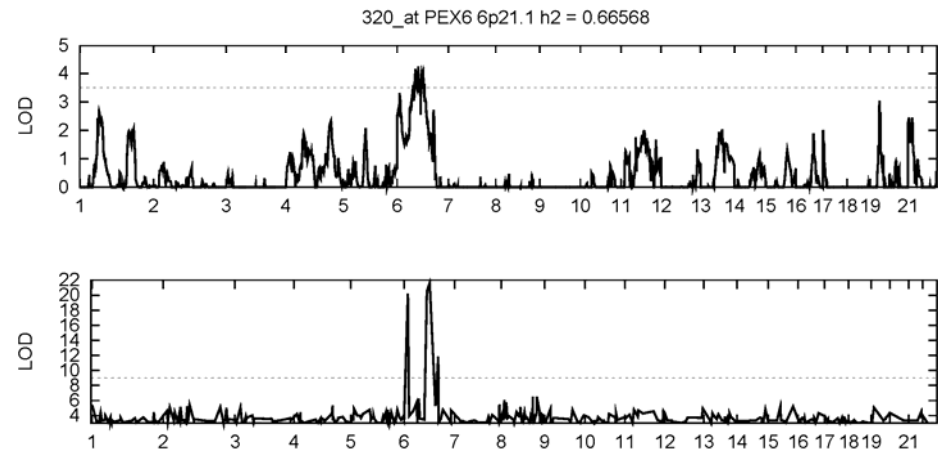Association genome scan
1,000,000 diallelic markers

+ Linkage
No Association

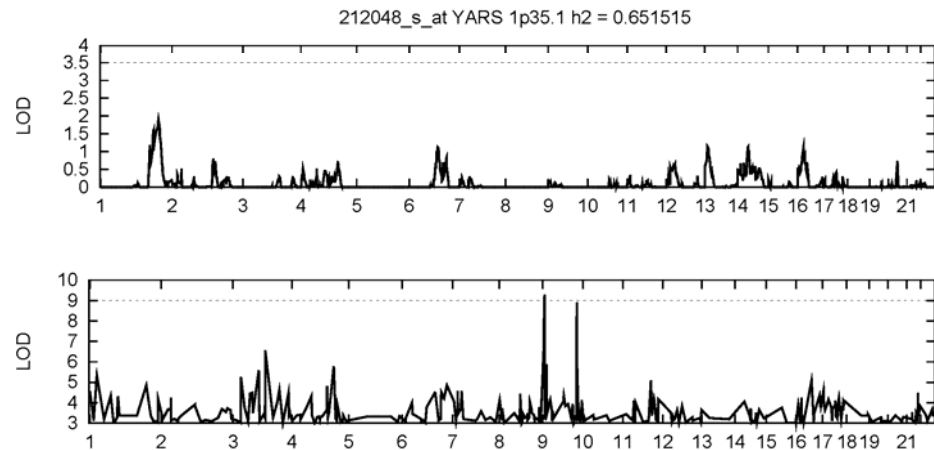210148_at HIPK3 11p13 h2 = 0.589661

65588_at GenBank 20q12 h2 = 0.607018

+ Linkage
+ Association

320_at PEX6 6p21.1 h2 = 0.66568

217869_at HSD17B12 11p11.2 h2 = 0.583333

No Linkage + Association

212048_s_at YARS 1p35.1 h2 = 0.651515
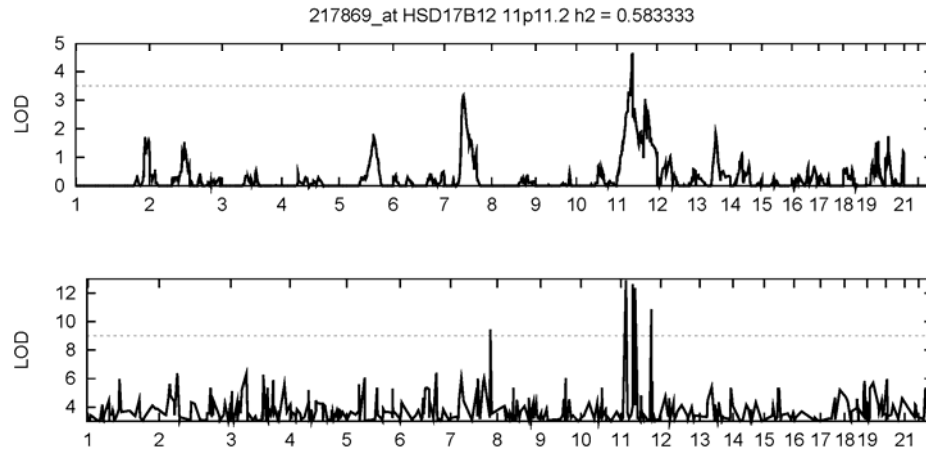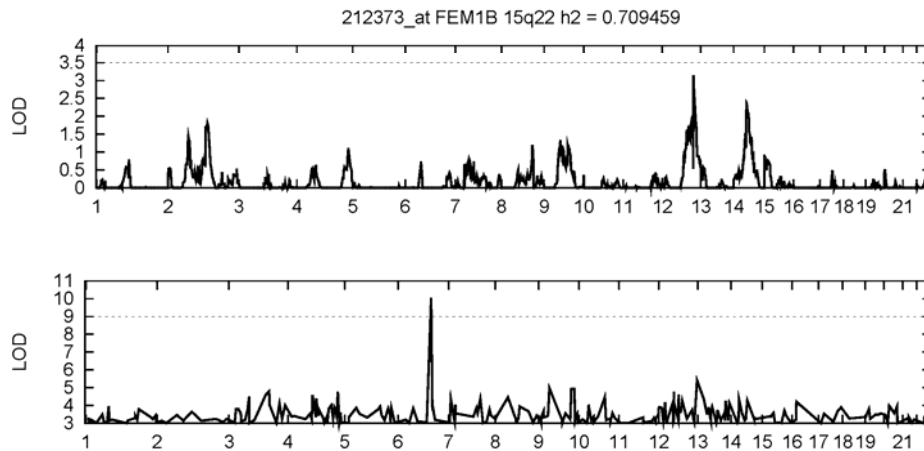
Yes, genome-wide association will work

(…sometimes…)

212373_at FEM1B 15q22 h2 = 0.709459

Challenges to come?

205048_s_at PSPH 7p15.2-p15.1 h2 = 0.589911

208151_x_at DDX17 22q13.1 h2 =
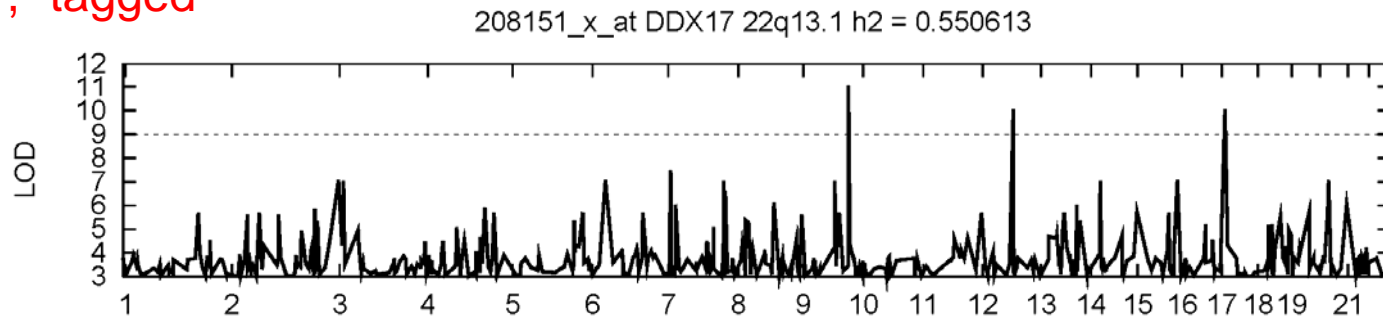
Caution with Tagging

208151_x_at DDX17 22q13.1 h2 = 0.550613
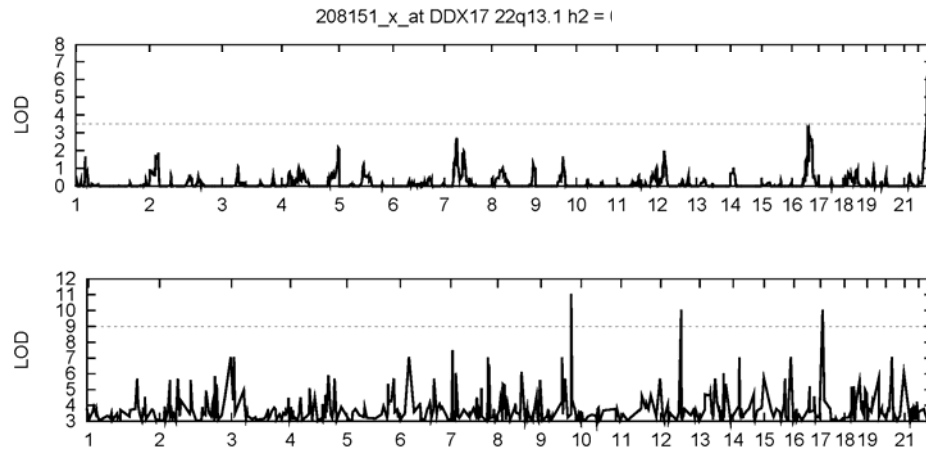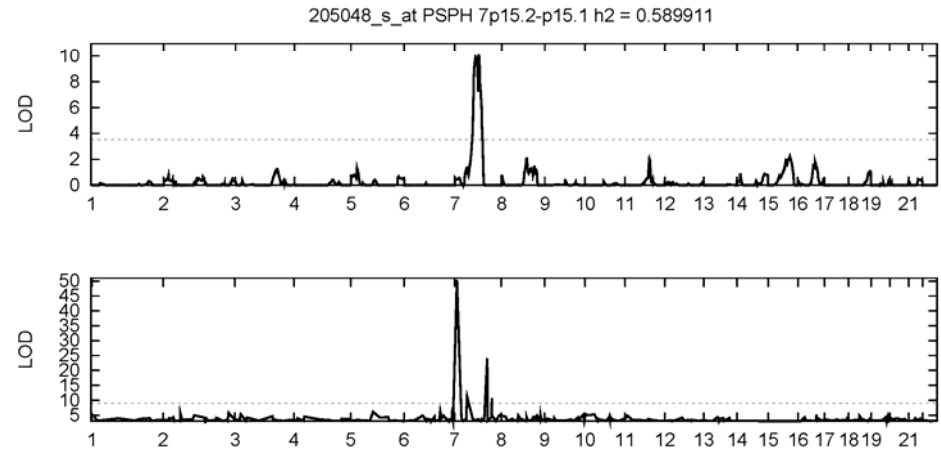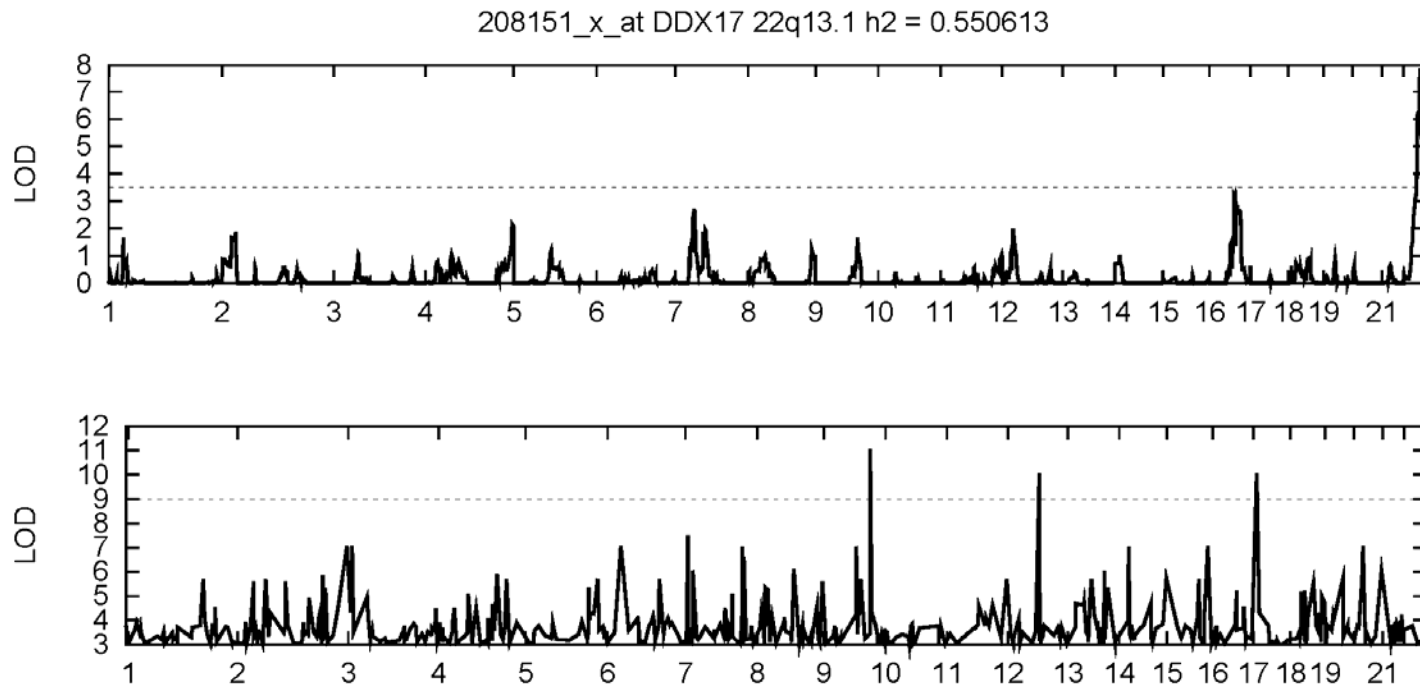
Here: excluded all SNPs with $r^2 = 1$
What effect does this exclusion have?

Caution with Inferences Based on Tagging
- localization-

No r^2 = 1, "tagged"

208151_x_at DDX17 22q13.1 h2 = 0.550613

All markers, "untagged"

Which ones are 'true' loci?