



Statistical Tools for Disease Gene Mapping and Association Studies

Marcella Devoto

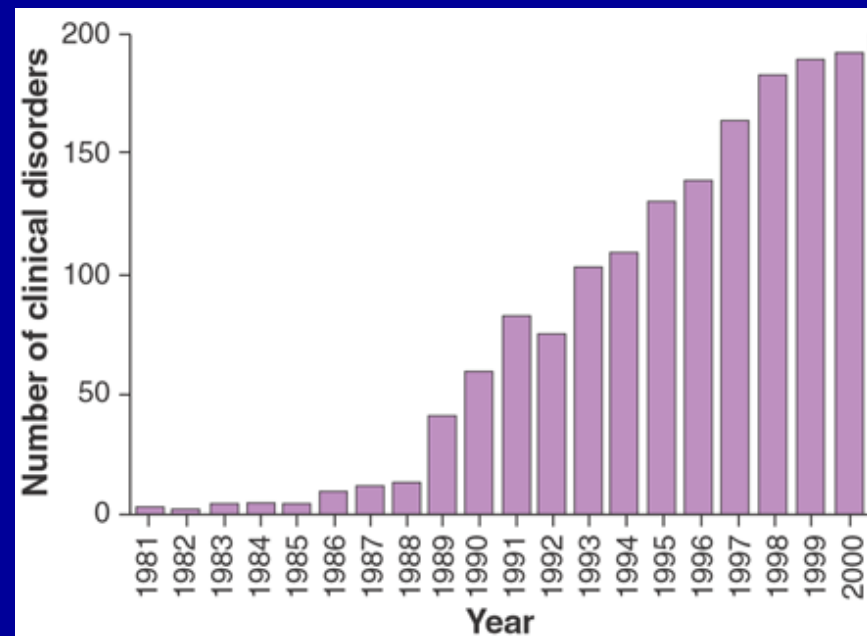
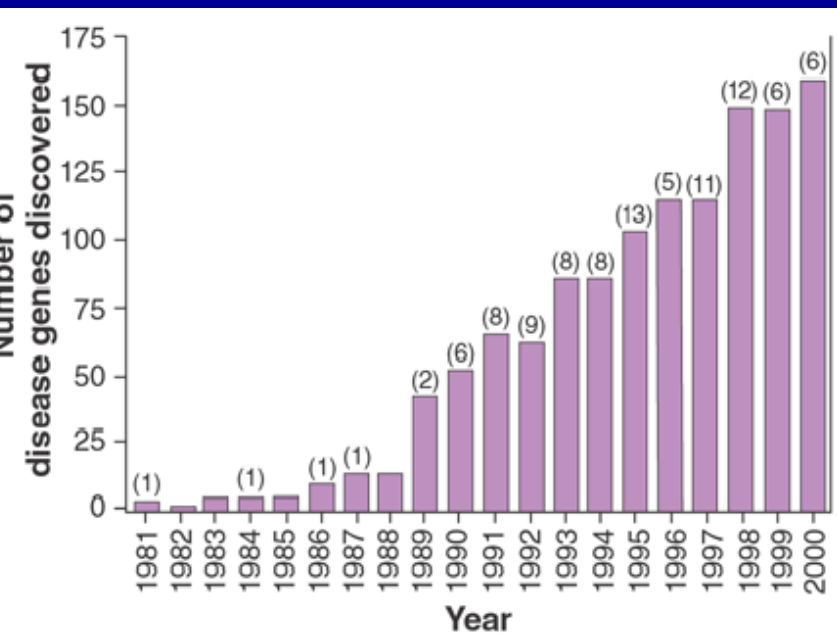
Department of Biomedical Research

Nemours Children's Clinic

Wilmington, DE



Pace of disease gene discovery (1981 to 2000)



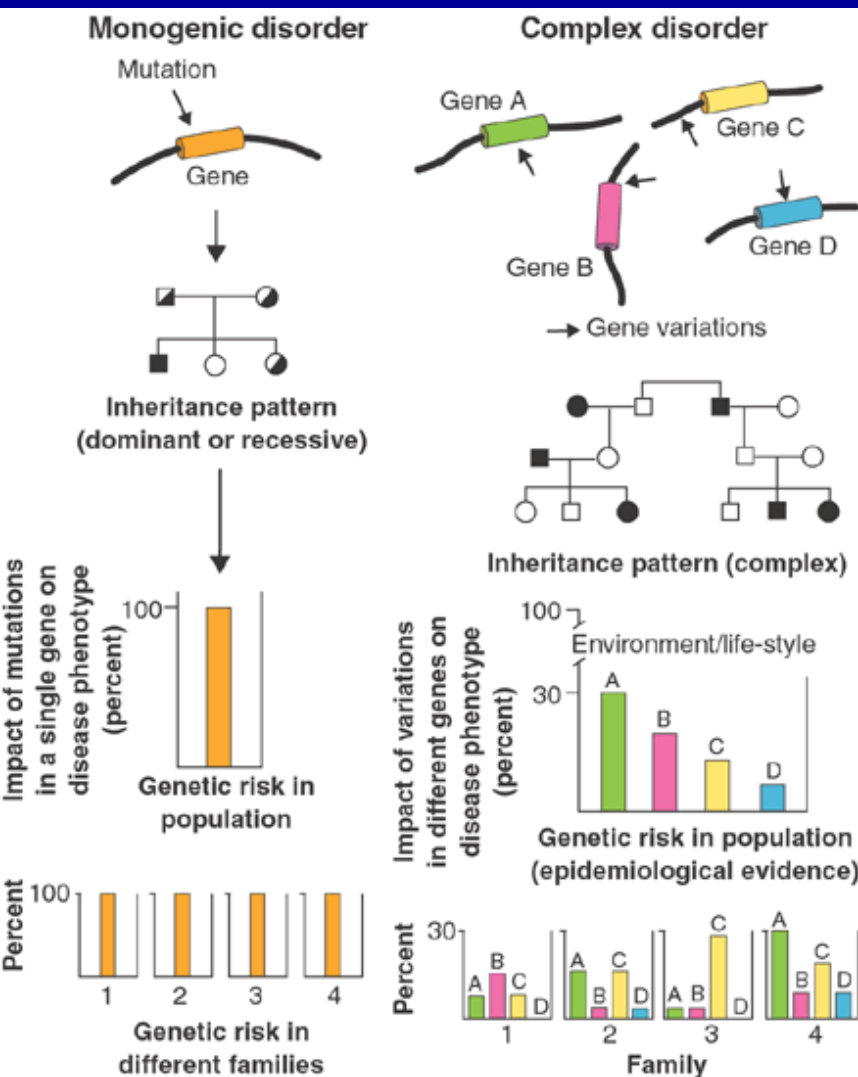
The number of disease genes discovered (and the number of disease entities associated to a molecular defect) has increased dramatically between 1981 and 2000.

Numbers in parentheses indicate disease-related genes that are polymorphisms ("susceptibility genes").

from L. Peltonen and V.A. McKusick: Dissecting Human Disease in the Postgenomic Era. Science, 2001.



Monogenic versus complex disease



Inheritance of monogenic and complex (multifactorial) disorders. In monogenic diseases, mutations in a single gene are both necessary and sufficient to produce the clinical phenotype and to cause the disease. The impact of the gene on genetic risk for the disease is the same in all families. In complex disorders with multiple causes, variations in a number of genes encoding different proteins result in a genetic predisposition to a clinical phenotype. Pedigrees reveal no Mendelian inheritance pattern, and gene mutations are often neither sufficient nor necessary to explain the disease phenotype. Environment and life-style are major contributors to the pathogenesis of complex diseases. In a given population, epidemiological studies expose the relative impact of individual genes on the disease phenotype. However, between families the impact of these same genes might be totally different. (...)

from L. Peltonen and V.A. McKusick: *Dissecting Human Disease in the Postgenomic Era*. Science, 2001.



Polymorphic susceptibility loci in common diseases

Common disorder	Lifetime prevalence	Susceptibility locus	Common risk allele	Population frequency
ischaemic heart disease (>40 y)	1 in 2 males	<i>DCP1 (ACE)</i>	D	0.30–0.70
	1 in 3 females	<i>APOE</i>	e4	0.06–0.37
essential hypertension	1 in 4–10	<i>AGT</i>	M235T T174M	0.34–0.84 0.11
neural tube defect	1–2 in 1,000	<i>MTHFR</i>	677C-T	0.32–0.38
Alzheimer disease (>65 y)	1 in 10–20	<i>APOE</i>	e4	0.06–0.37
insulin-dependent diabetes mellitus (>20 y)	1 in 300	<i>INS</i>	5' VNTR class I	0.76
		<i>HLA-DR3/DR4</i>	DR3/DR4	0.35
ankylosing spondylitis	1 in 200 males >20 y	<i>HLA-B</i>	B27	0.08
venous thrombosis (APC resistance)	1 in 1,000	<i>FV</i>	R506Q	0.02–0.08

Linkage studies of complex traits have been disappointing

- ◆ The positional approach has not been a successful strategy in the identification of genes responsible for complex traits
- ◆ Few loci responsible for complex traits have been consistently mapped; fewer genes have been identified
- ◆ Strategies to improve chances of success have focused on selection of families/patients, markers, and alternative statistical methods

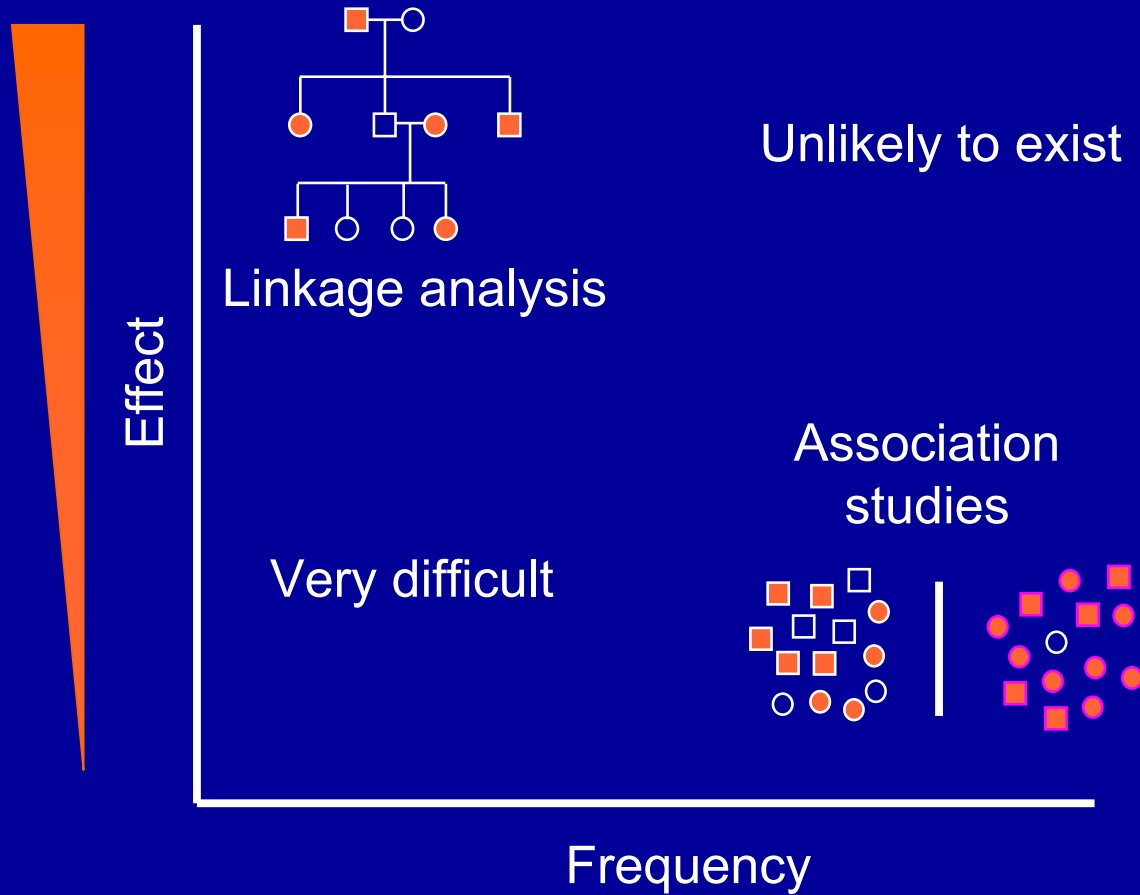


Association studies are a powerful tool for detecting complex trait susceptibility loci

- ◆ In a 1996 Science paper, Risch and Merikangas showed that TDT studies have more power than ASP linkage studies to detect genes for complex traits
- ◆ All TDT computations were based on the optimal assumption that the analyzed allele was the disease allele itself or in complete disequilibrium with it
- ◆ A more common situation is the analysis of polymorphisms with a low prior probability of being the disease allele even if they are within the actual disease gene
- ◆ If linkage disequilibrium is less than maximal, and if allele frequencies greatly differ between disease and marker loci, then power can be low



Association studies are a powerful tool for detecting complex trait susceptibility loci



modified from D. Altschuler



Methods for genetic analysis of complex traits:

Association studies

- ◆ **Association between risk factor and disease:**
risk factor is significantly more frequent among affected than among unaffected individuals
- ◆ Genetic association exists when certain **alleles/genotypes/haplotypes** are found at increased frequency in patients compared to healthy controls



Case-control studies: χ^2 test

2x3 contingency table

	Genotypes			
	AA	Aa	aa	
Cases	n_{AA}	n_{Aa}	n_{aa}	N
Controls	m_{AA}	m_{Aa}	m_{aa}	M
	t_{AA}	t_{Aa}	t_{aa}	N+M

Test of independence:

$$\chi^2 = \sum (O-E)^2 / E \text{ with 2 df}$$



Case-control studies: χ^2 test

2x2 contingency table

	Alleles		
	A	a	
Cases	n_A	n_a	2N
Controls	m_A	m_a	2M
	t_A	t_a	2(N+M)

Test of independence:

$$\chi^2 = \sum (O-E)^2 / E \text{ with 1 df}$$



Hardy-Weinberg Equilibrium

Biallelic locus: A, a \rightarrow genotypes AA, Aa, aa

Allele frequencies: A $P(A) = p$

a $P(a) = q$

Genotype frequencies are in HWE if:

AA $P(AA) = p^2$

Aa $P(Aa) = 2pq$

aa $P(aa) = q^2$

- ◆ If marker genotypes are not in HWE, allelic test is not a valid test of association
- ◆ Deviation from HWE is often indication of problems (genotyping error, population stratification, etc.)



Haplotypes

GENOTYPES

Locus 1	2	13
Locus 2	1	6
	9	15
	4	17
	1	9
	2	6
	9	17
	2	12
	7	12
	6	14
	1	7
	18	18
	1	4
Locus N	10	10

Identification of
phase →

HAPLOTYPES

2	●	●	13
6	●	●	1
9	●	●	15
17	●	●	4
1	●	●	9
6	●	●	2
9	●	●	17
2	●	●	12
12	●	●	7
14	●	●	6
7	●	●	1
18	●	●	18
1	●	●	4
10	●	●	10



Motivation for haplotype-based analysis

- ◆ Increased ability to identify regions that are shared identical by descent among affected individuals, and therefore more informative
- ◆ Haplotype may be the causative “composite allele” rather than a particular nucleotide at a particular SNP
- ◆ Haplotype analysis is meaningful only if SNPs are in themselves in LD – requires preliminary marker LD analysis



Haplotype determination options

- ◆ Collect and genotype family members
- ◆ Laboratory-based techniques
- ◆ Statistical estimation in unphased individuals
 - Likelihood-based E-M and related algorithms
 - Software: Phase, Haplotyper, EH



Case-control studies: χ^2 test

	Haplotypes					2xr contingency table
	1	2	3	...	r	
Cases	n_1	n_2	n_3	...	n_r	2N
Controls	m_1	m_2	m_3	...	m_r	2M
	t_1	t_2	t_3	...	t_r	2(N+M)

Test of independence:

$$\chi^2 = \sum (O-E)^2 / E \text{ with } r-1 \text{ df}$$



Measures of association

Alleles: A, a genotypes AA, Aa, a,a

Genotype relative risk:

$$GRR_{AA} = \frac{\text{risk for AA}}{\text{risk for aa}}$$

$$GRR_{Aa} = \frac{\text{risk for Aa}}{\text{risk for aa}}$$

$$GRR_{a/a} = 1$$

Allele relative risk:

$$\Phi_A = \frac{\text{risk for A}}{\text{risk for a}}$$

$$\Phi_a = 1$$



Measures of association

genotype	disease	
	+	-
A/A	n_{11}	n_{12}
A/a	n_{21}	n_{22}
a/a	n_{31}	n_{32}

$$GRR_{A/A} \sim OR_{A/A} = n_{11}n_{32}/n_{12}n_{31}$$

$$GRR_{A/a} \sim OR_{A/a} = n_{21}n_{32}/n_{22}n_{31}$$

allele	disease	
	+	-
A	n_{11}	n_{12}
a	n_{21}	n_{22}

$$\Phi_A \sim OR_A = n_{11}n_{22}/n_{12}n_{21}$$



Measures of association

- ◆ Genotypes

- Dominant/recessive/codominant

- » e.g.: $GRR_{AA} \sim GRR_{Aa} \rightarrow A$ dominant

- ◆ Alleles/haplotypes

- Multiplicative or additive model

- » $GRR_{ij} = \Phi_i \Phi_j$, or $GRR_{ij} = \Phi_i + \Phi_j$



Tests and measures of association

◆ Software

- Any standard statistical package!
- HAPLOVIEW (Barrett et al, 2004)
 - » www.broad.mit.edu/mpg/haploview/index.php
- Finetti (T. Wienker and T. Strom)
 - » ihg.gsf.de/cgi-bin/hw/hwa1.pl



Association studies in complex traits

- ◆ Association between trait and allele may occur if:
 - allele is responsible for the disease (increases risk of disease)
 - allele is in linkage disequilibrium with disease risk allele (i.e., it occurs at increased frequency on chromosomes carrying a disease risk allele)
 - sample is made of individuals coming from populations with different allele frequencies



Causes of genetic association

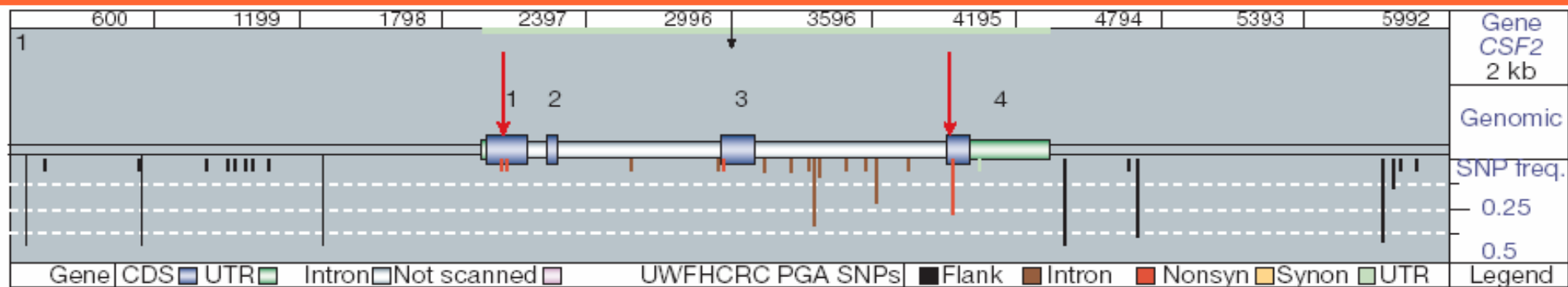
◆ Direct association



Association studies by direct analysis

◆ Direct analysis:

Association studies between disease and functional polymorphism (causative of disease) of candidate gene



Causes of genetic association

◆ **Direct association**

◆ **Indirect association due to LD**



Linkage disequilibrium

- ◆ Non random association between alleles at different loci
- ◆ Loci are in LD if alleles are present on haplotypes in different proportions than expected based on allele frequencies



Linkage disequilibrium

Locus 1: alleles A and a; frequencies : p_A and p_a

Locus 2: alleles B and b; frequencies : p_B and p_b

Possible haplotypes	A \perp B \perp	A \perp b \perp	a \perp B \perp	a \perp b \perp
Expected frequencies	$p_A p_B$	$p_A p_b$	$p_a p_B$	$p_a p_b$
Observed frequencies	p_{AB}	p_{Ab}	p_{aB}	p_{ab}

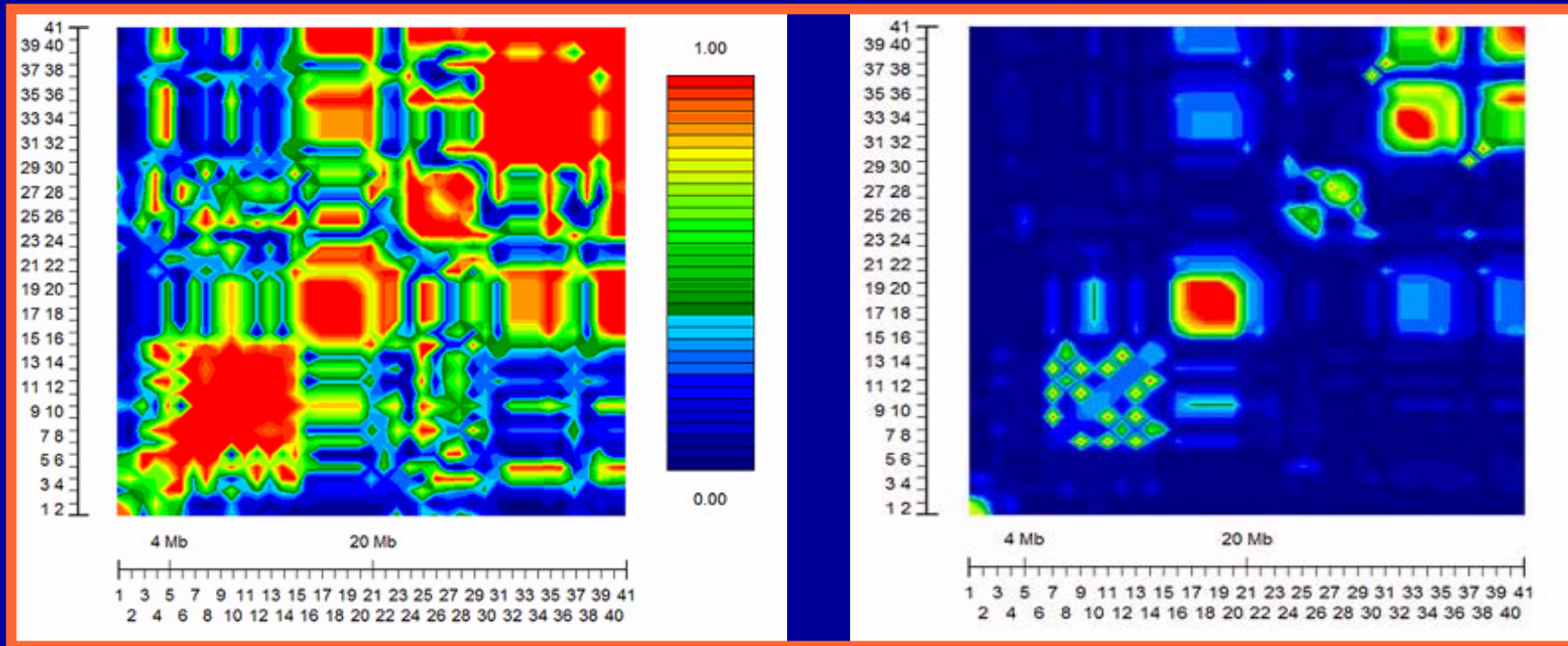
$$D = p_{AB} - p_A p_B \neq 0$$



Graphic representation of LD

D'

r^2



GOLD (Abecasis and Cookson, 2000)

<http://www.sph.umich.edu/csg/abecasis/GOLD/index.html>



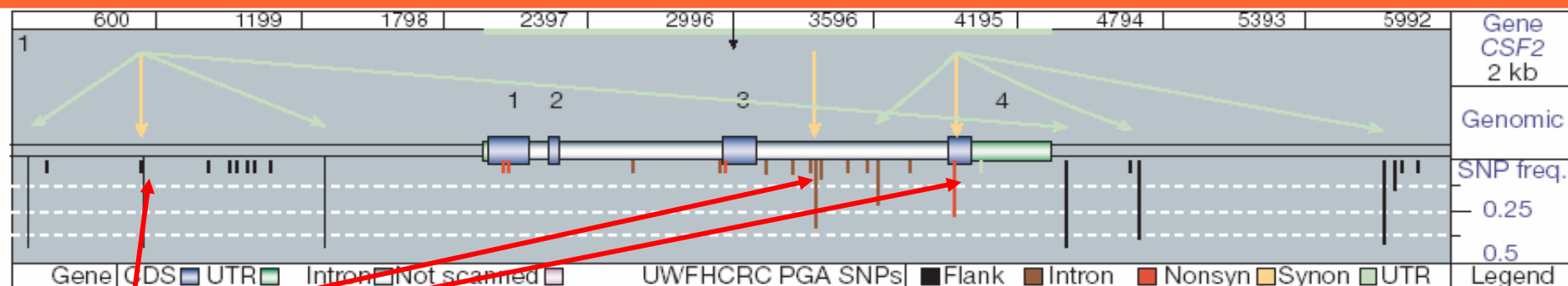


Association studies by indirect analysis

◆ Indirect analysis:

Association studies between disease and “random” SNPs (within or near candidate gene; in linked region; or across the genome)

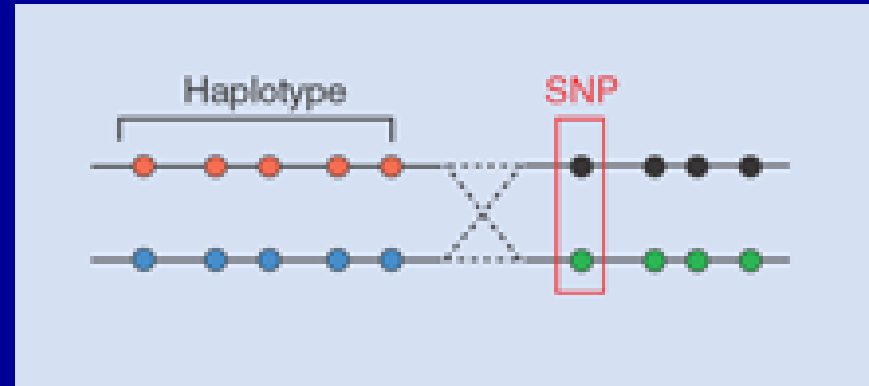
➔ Linkage Disequilibrium mapping



TagSNPs

Haplotype map: Map of the Human Genome 3.0

▶ Alleles of SNPs in LD with each other tend to be transmitted together as a haplotype

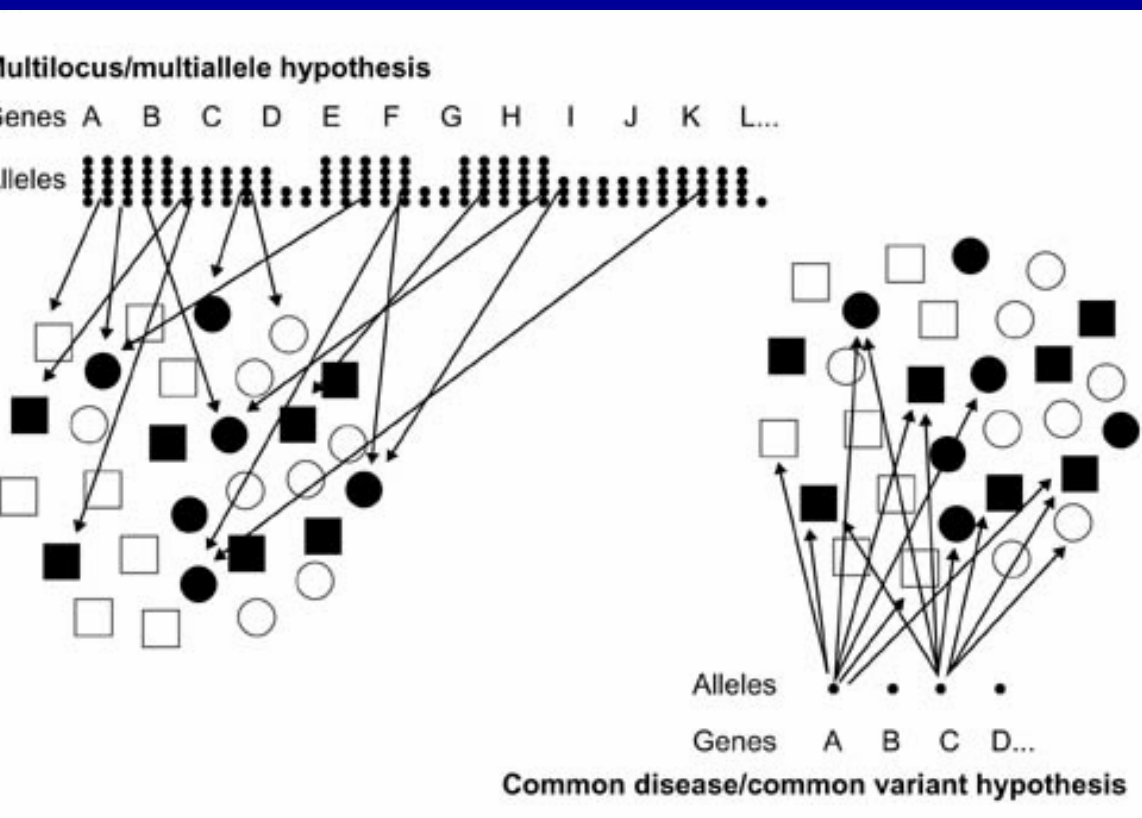


▶ In some regions of the human genome as long as 50Kb, 4 or 5 haplotypes account for up to 80 or 90% of the total variation

▶ This means that fewer SNPs will be needed to scan the corresponding regions



Few or many disease alleles?



from AF Wright and ND Hastie,
<http://genomebiology.com/2001/2/8/comment/2007/?mail=0000104>

Supporters of the SNP association studies argue that common disease susceptibility alleles must be relatively neutral to selective pressure and subject to founder effect, and thus frequent (common disease/common variant hypothesis).

In contrast, other argue that there is an inverse relationship between frequency of an allele and the magnitude of its genetic effect. Thus few variants of clinical consequence will be common, and excessive locus and allelic heterogeneity will pose an obstacle to SNP association studies.

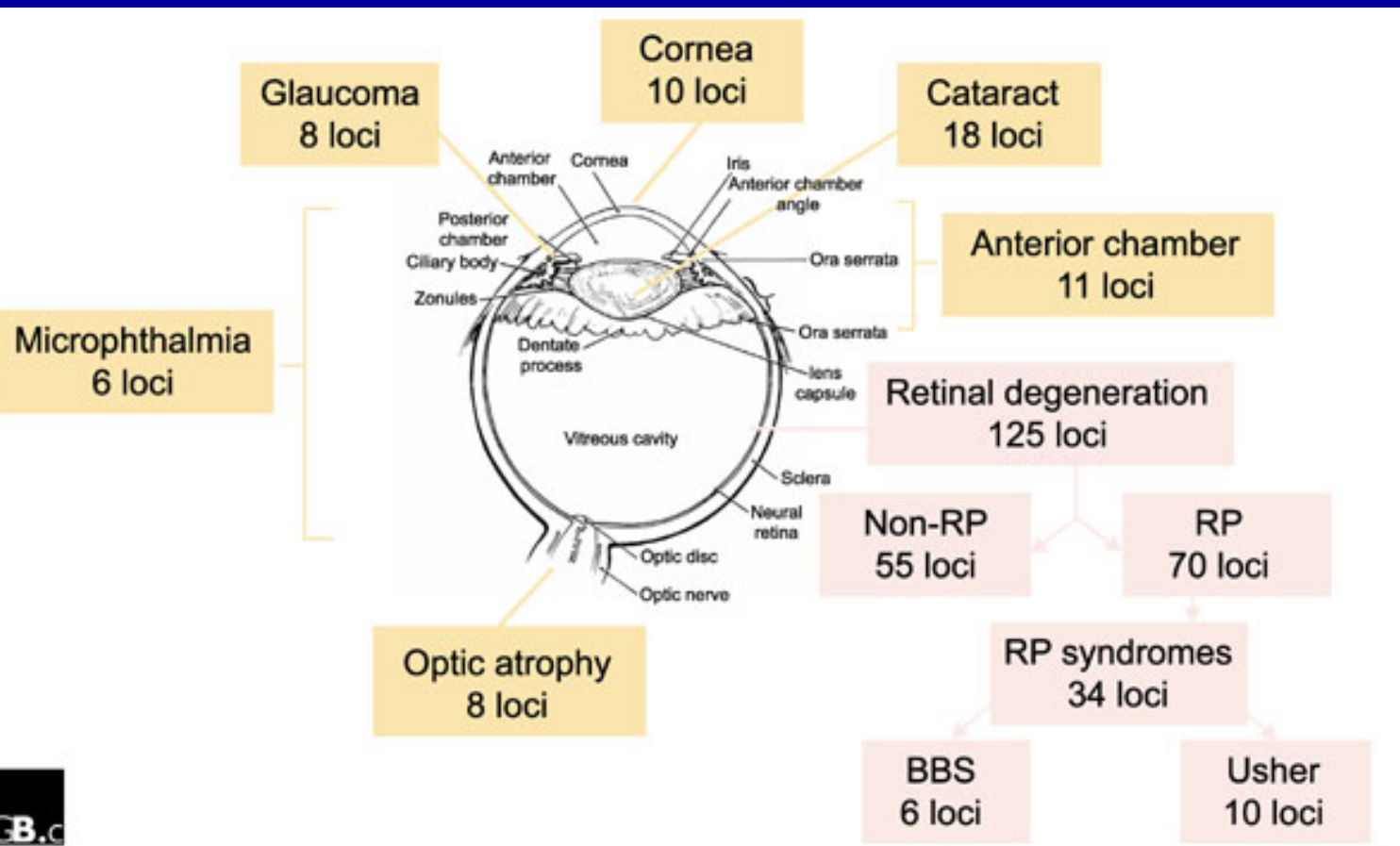


Polymorphic susceptibility loci in common diseases

Common disorder	Lifetime prevalence	Susceptibility locus	Common risk allele	Population frequency
ischaemic heart disease (>40 y)	1 in 2 males	<i>DCP1 (ACE)</i>	D	0.30–0.70
	1 in 3 females	<i>APOE</i>	e4	0.06–0.37
essential hypertension	1 in 4–10	<i>AGT</i>	M235T T174M	0.34–0.84 0.11
neural tube defect	1–2 in 1,000	<i>MTHFR</i>	677C-T	0.32–0.38
Alzheimer disease (>65 y)	1 in 10–20	<i>APOE</i>	e4	0.06–0.37
insulin-dependent diabetes mellitus (>20 y)	1 in 300	<i>INS</i>	5' VNTR class I	0.76
		<i>HLA-DR3/DR4</i>	DR3/DR4	0.35
ankylosing spondylitis	1 in 200 males >20 y	<i>HLA-B</i>	B27	0.08
venous thrombosis (APC resistance)	1 in 1,000	<i>FV</i>	R506Q	0.02–0.08



Locus heterogeneity in Mendelian disorders

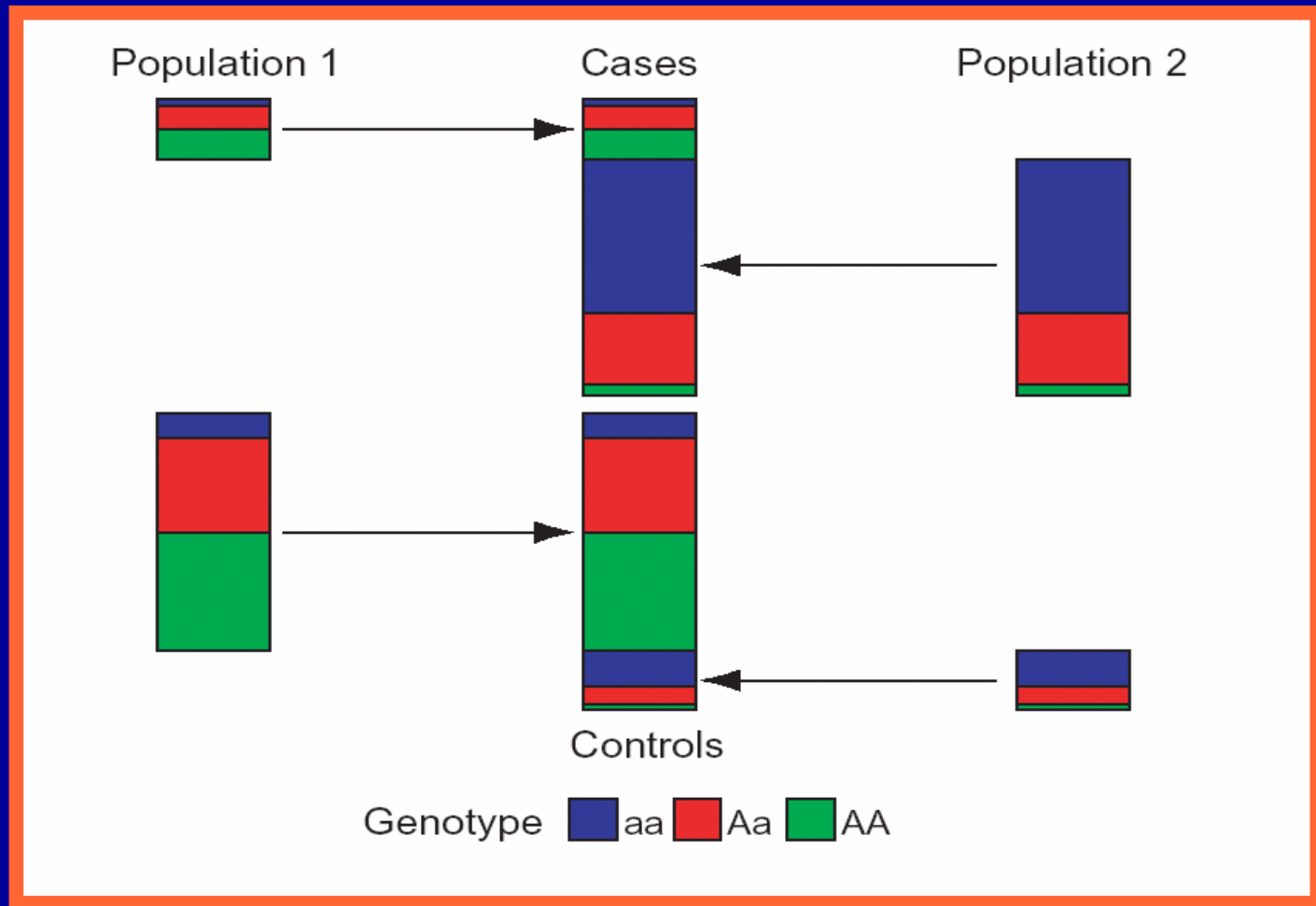


Causes of genetic association

- ◆ Direct association
- ◆ Indirect association due to LD
- ◆ Spurious association due to confounding factors (e.g., population stratification)



Association due to population stratification



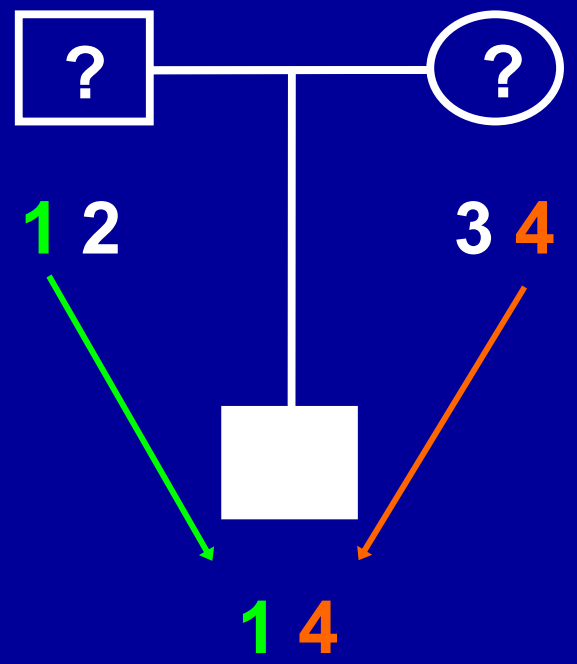
Population stratification

Possible solutions

- ◆ Stratify sample based on confounding variable
- ◆ Apply test correction (Genomic Control)
- ◆ Use related controls in family-based association studies



Family-based association studies



1 4 transmitted

2 3 non transmitted



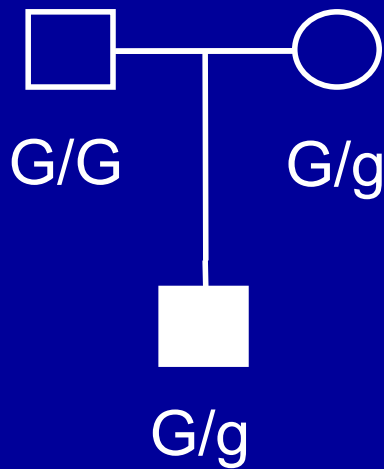
“control”



2 3



TDT: Transmission Disequilibrium Test



		non transmitted	
		G	g
transmitted	G	a	b
	g	c	d

$$\begin{aligned}
 \text{TDT}_G &= (T_G - NT_G)^2 / (T_G + NT_G) \\
 &= (b - c)^2 / (b + c) \sim \chi^2_1
 \end{aligned}$$



TDT: Transmission Disequilibrium Test

- Multiallelic markers
 - ETDT (Sham & Curtis, 1995)
- Missing parent genotypes
 - TRANSMIT (Cayton, 1999)
- Haplotypes
 - TDTHAP (Clayton & Jones, 1999)
- Sibs
 - TDT/STDT (Spielman & Ewens, 1998)
- Pedigrees
 - PBAT (Martin et al, 2000)



References for genetic association studies

- ◆ For all type of statistical genetics software:
<http://linkage.rockefeller.edu/soft/>
- ◆ Hirschhorn and Daly: Nature Review Genetics 6: 95-108
- ◆ Wang, Barratt, Clayton, and Todd: Nature Review Genetics 6: 109-118

