Introduction to Genetic Epidemiology

Erwin Schurr
McGill International TB Centre
McGill University

Methods of investigation in humans

Phenotype	Rare (very severe forms)	Common (infection/affection status)
Sample	Small	Large
Causality	monogenic	complex
Main tools	Mendelian Genetics	Genetic Epidemiology

Rare mutation
Strong effect

Common polymorphism
Modest effect

Complex phenotypes

- In contrast to monogenic disease
- Complex trait :
  - Environmental factors
  - Genetic factors
    - major gene
    - other genes

↕ gene*environment interactions

↕ gene*gene interactions

- Examples:
  - Cancers
  - Cardiovascular diseases
  - Neurological diseases
  - Infectious diseases …

Genetic epidemiology: objectives / tools

Do genetic factors play a role ?

$\Rightarrow$ Epidemiological observations / Experimental model

What is their nature ?

$\Rightarrow$ Segregation analysis

What is their chromosomal location ?

$\Rightarrow$ Linkage analysis

Which allelic variant is implicated ?

$\Rightarrow$ Association studies

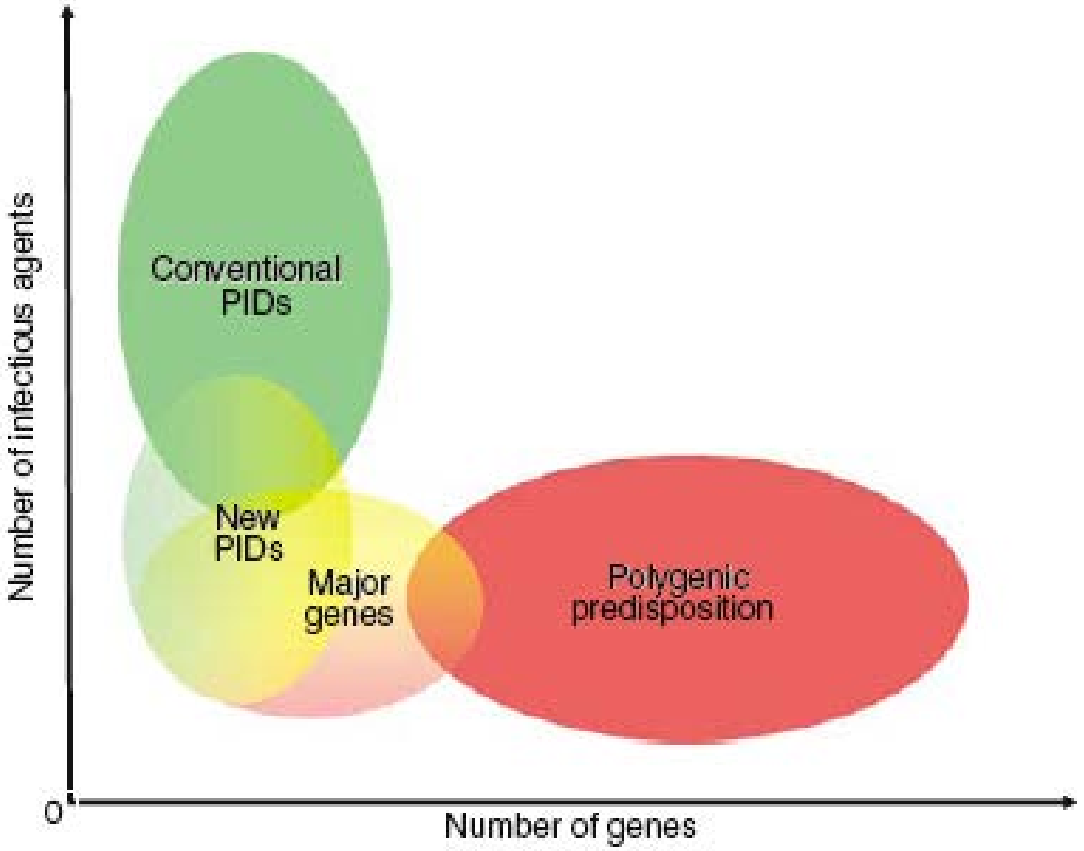
What is its function ?

$\Rightarrow$ Functional studies

Genetic epidemiology: overview

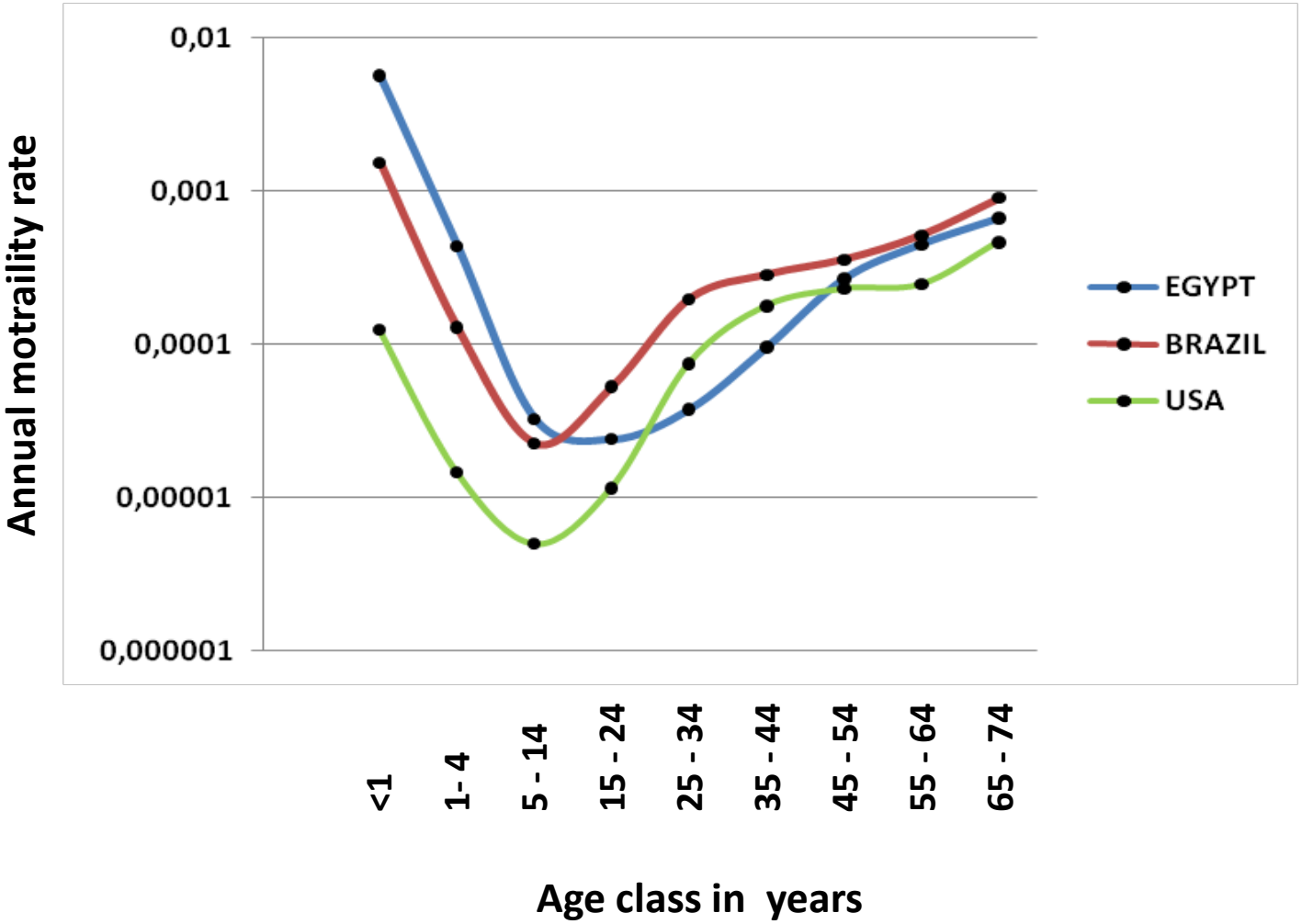
	Sample	# affected sibs	DNA	Markers	Main goal
Segregation analysis	**Families**	**0 → n**	**No**	**-**	**genetic model**
Linkage analysis	**Families**	**2 → n**	**Yes**	**Microsat/ SNPs**	**candidate regions**
Association studies	**Families**	**1 → n**	**Yes**	**SNPs**	**candidate alleles**
	Cases/controls	**-**	**Yes**	**SNPs**	**candidate alleles**

Spectrum of genetic predisposition



Casanova, Abel EMBO J 2007

Interplay of age and genetics



Do genetic factors play a role?

→ Epidemiological observations

What is their nature?

→ Segregation analysis

What is their chromosomal location?

→ Linkage analysis

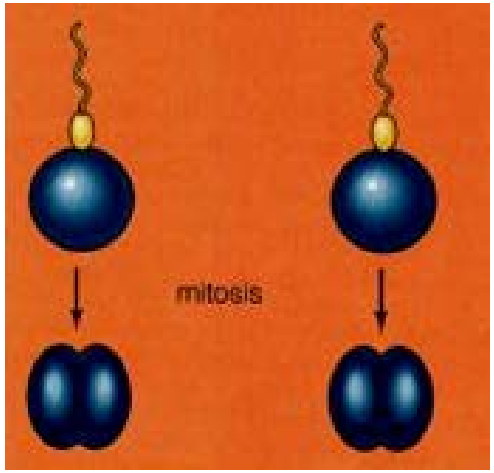
What is the causal variant?

→ Association studies

What is the function?

Family level – Twin studies

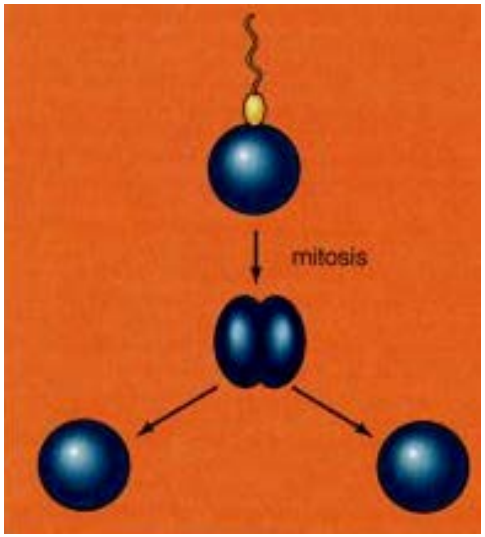
DZ TWINS



2 fertilizations

Share **50%** of genetic background

MZ TWINS



1 fertilization

Share **100%** of genetic background

MZ Twins

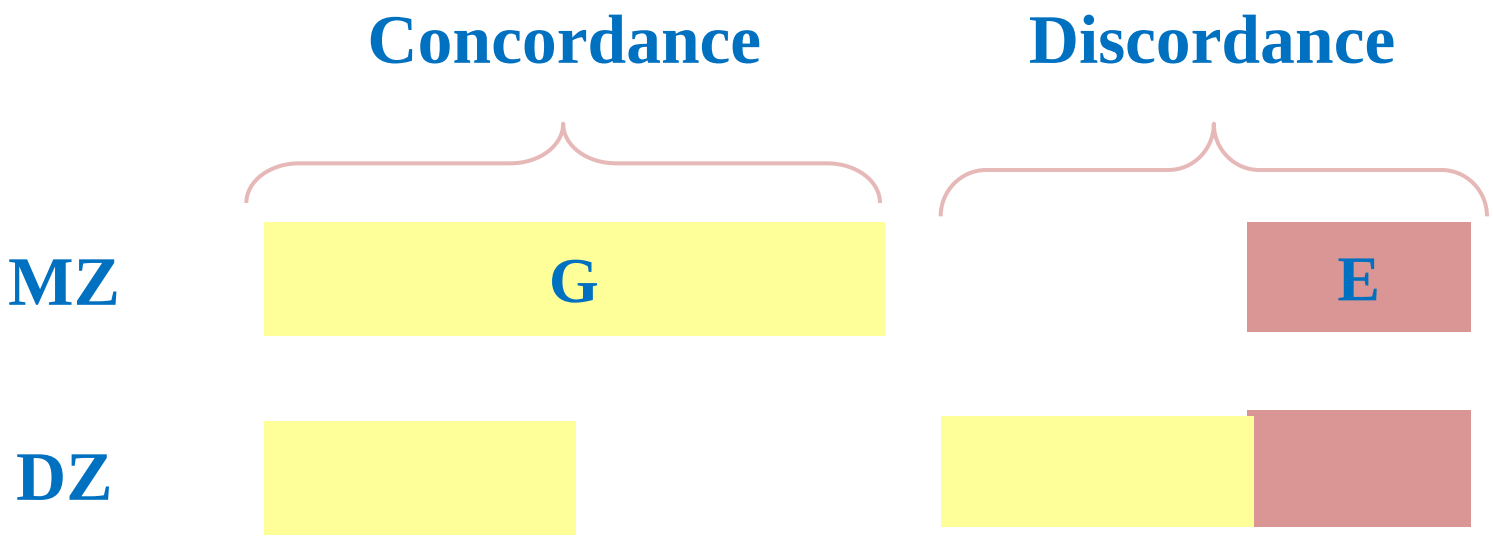
Twin 2 \ Twin 1	+	-
+	A	B
-	C	D

DZ Twins

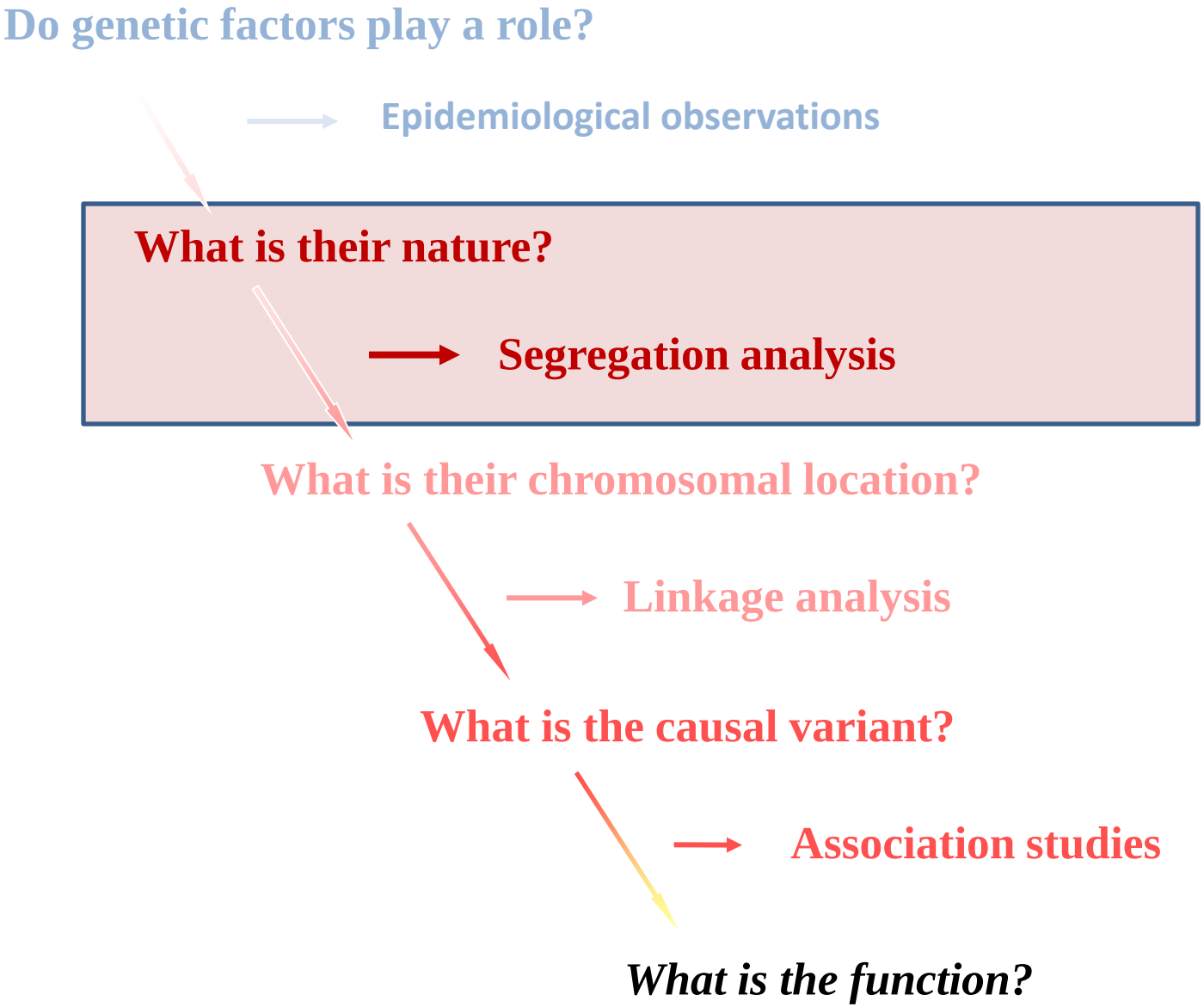
Twin 2 \ Twin 1	+	-
+	A	B
-	C	D

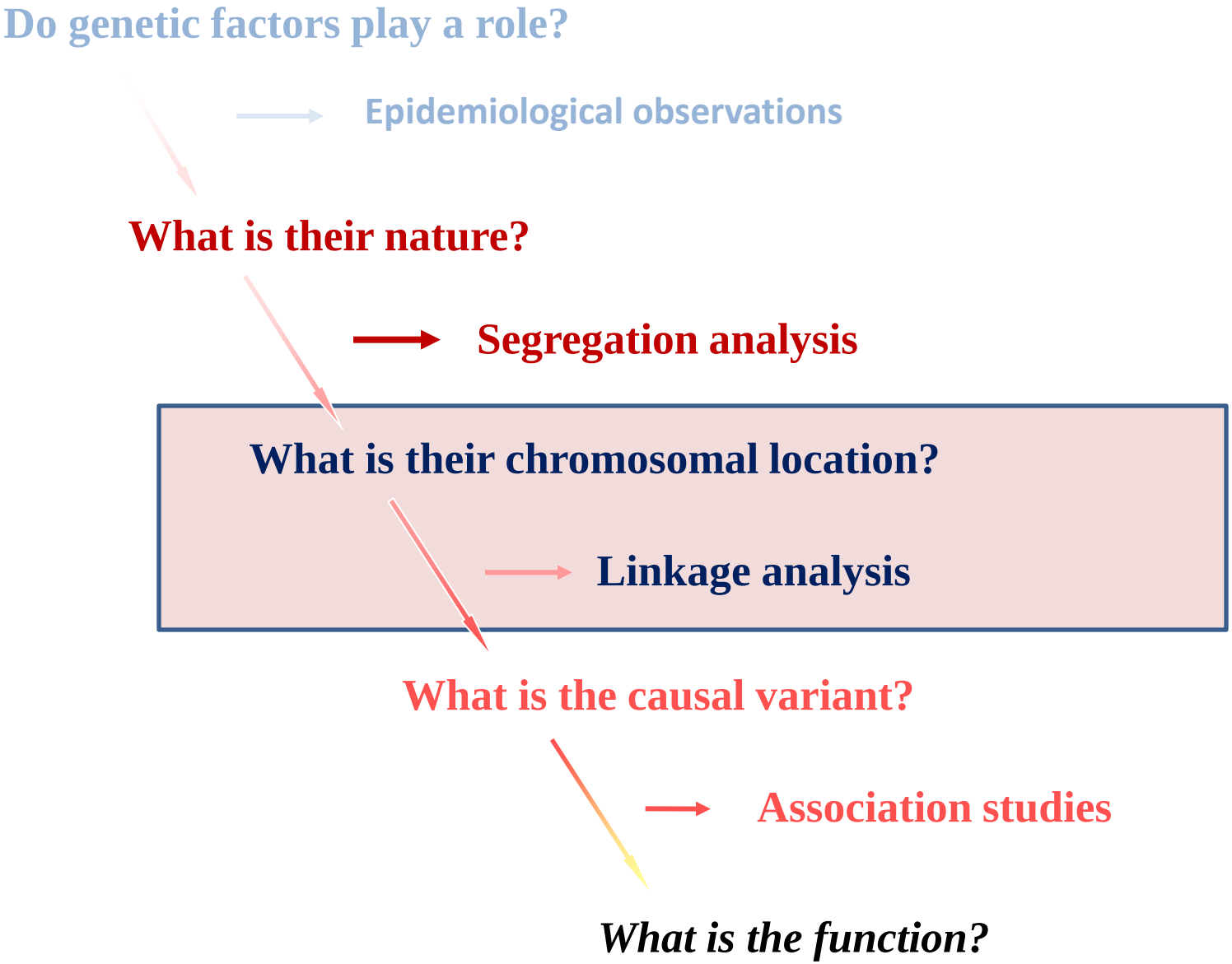
Concordance Rate = 2A / (2A + B + C)

Genetic contribution: C_{MZ} vs. C_{DZ}



Genetic contribution: $C_{MZ} > C_{DZ}$

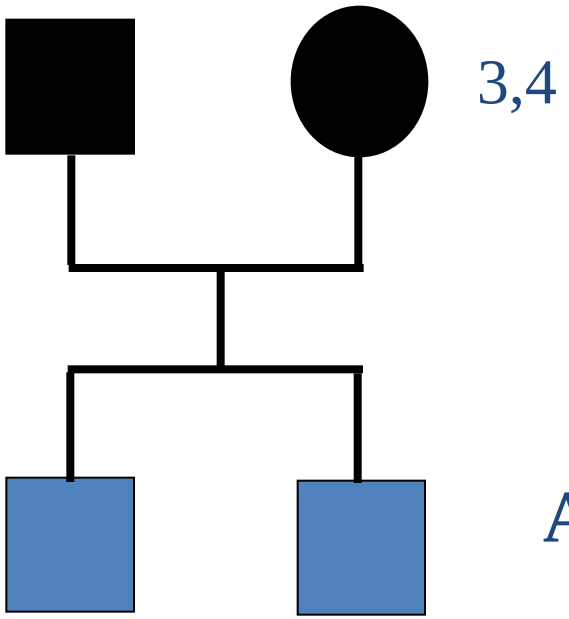

Do genetic factors play a role?
Epidemiological observations
What is their nature?
Segregation analysis
What is their chromosomal location?
Linkage analysis
What is the causal variant?
Association studies
What is the function?


Do genetic factors play a role?
Epidemiological observations
What is their nature?
Segregation analysis
What is their chromosomal location?
Linkage analysis
What is the causal variant?
Association studies
What is the function?

Model-based linkage analysis

- Need to specify the relation between the phenotype and the genotype
  - frequency of the disease allele
  - probability to be affected given genotype and risk factors
- Most powerful method *IF* the genetic model is correct
- Estimation of the *recombination fraction* θ between the 'phenotype' locus (to locate) and the marker locus (known location)
- Linkage test = $\theta < 0.5$?
- Example: schistosomiasis (infection intensities, severe hepatic fibrosis)

Model-free linkage analysis



RESOLUTION FOR GENE LOCALIZATION:
~ 10 million base pairs

Alleles shared	Probability
0	¼
1	¼
1	¼
2	¼

Linkage only look at few meiosis

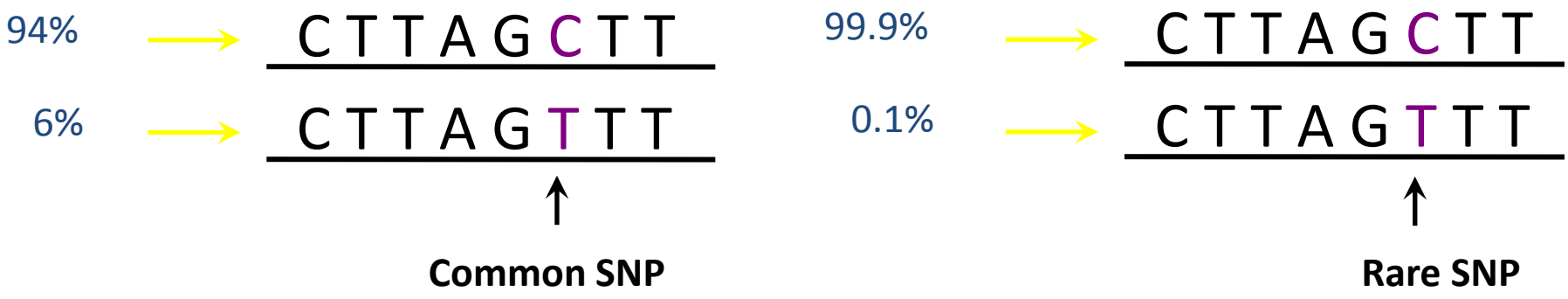
Can we look at more … can we see dead people ?

Yes … by studying Linkage Disequilibrium


SNP
SATURDAY NIGHT
PROJECT

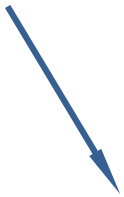
Single Nucleotide Polymorphism

In a population the same building block (=nucleotide) of DNA can occur in two alternative forms – i.e. at a given DNA position two different nucleotides (=alleles) can occur. If in a population the less frequent allele occurs with >2% we call it common variation.



Univariate analysis

One binary phenotype



One candidate SNP in a candidate gene

One association study

Genotypic analysis

	cases	controls
AA	c_0	t_0
AB	c_1	t_1
BB	c_2	t_2

Goodness-of-fit test = Chi-square 2 df

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Hypothesis testing – general strategy

1. **Formulate null (H_0) and alternative (H_1) hypothesis**
2. **Build a test statistic according to the data to come**
3. **Identify distribution of the test statistic under H_0**
4. **Define a decision rule (i.e. type I error)**
5. **Make the experiment and compute the test statistic**
6. **Conclude, i.e. reject or not H_0 and precise p-value**
7. **Interpret the conclusion**

Hypothesis testing – genetic association

1. H_0 : cases = controls H_1: cases ≠ controls

2. $$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3. Under H_0, χ^2 is distributed as a chi-square with 2 df

4. Type I error 5% ⇔ reject H_0 if $\chi^2 > 5.99$

5. $\chi^2 = 348$

6. $\chi^2 > 5.99$ therefore we reject H_0 (p-value<0.001)

7. The genotypic distribution is significantly different in cases and in controls

Genotypic analysis

	cases	controls	
AA	200	500	*700*
AB	200	300	*500*
BB	600	200	*200*
	1,000	*1,000*	*2,000*

Expected AA cases = 1,000*700 / 2,000 =350

Expected AB cases = 1,000*500 / 2,000 =250

Etc …

Goodness-of-fit test = Chi-square 2 df

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

X^2=[(200-350)²/350] + [(200-250)²/250] + [(600-400)²/400] + [(500-350)²/350] + [(300-250)²/250] + [(200-400)²/400]

X^2=348.5 with 2 df

Genotypic analysis

	cases	controls
AA	c_0	t_0
AB	c_1	t_1
BB	c_2	t_2

Odds ratio AB vs. AA = $c_1 * t_0 / c_0 * t_1$

Genotypic analysis

	cases	controls
AA	200	500
AB	200	300
BB	600	200

Odds ratio AB vs. AA = 200*500 / 200*300=1.66

Odds ratio BB vs. AA = 600*500 / 200*200=7.5

Genotypic analysis – general strategy

	cases	controls
AA	c_0	t_0
AB	c_1	t_1
BB	c_2	t_2

Estimate OR

Estimate OR

Optimize coding scheme

Genotypic analysis – dominance effect

B dominant

	cases	controls
AA	c_0	t_0
AB or BB	$c_1 + c_2$	$t_1 + t_2$

B recessive

	cases	controls
AA+AB	$c_0 + c_1$	$t_0 + t_1$
BB	c_2	t_2

Goodness-of-fit test = chi-square 1 df

Example

	cases	controls	OR	P-value
AA	**100**	**200**	**1.00**	
AB+BB	**200**	**100**	**4.00**	**<0.001**

Interpretation ?

Type I error

Allele B $\Rightarrow$ phenotype = B is the causal allele

Allele B is in *linkage disequilibrium* with the causal allele

Genetic linkage between disease locus and marker locus
PLUS
Allele B is preferentially associated with the causal allele

Linkage is a relation between loci

Linkage disequilibrium is a relation between alleles

Descriptors of Linkage Disequilibrium

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

D’

	Linkage equilibrium (expected for distant loci)	Linkage disequilibrium (expected for nearby loci)
P_{AB}	$= P_A P_B$	$\neq P_A P_B$
P_{Ab}	$= P_A P_b$	$\neq P_A P_b$
P_{aB}	$= P_a P_B$	$\neq P_a P_B$
P_{ab}	$= P_a P_b$	$\neq P_a P_b$

$$D_{AB} = P_{AB} - P_A P_B$$

$$D_{AB} = P_{AB} - P_A P_B$$

sign is arbitrary

range ∝ allele frequencies

Hardly allows comparisons

Scaled version

$$D'_{AB} = D_{AB} / Dmax$$

$Dmax = \min (P_A P_b; P_a P_B)$

$$r^2_{AB} = D^2_{AB} / (P_A P_B P_a P_b)$$

Inflated Type I Error: Stratification

Replicate the results in an independent sample/population

Incorporate genomic information in the analysis

genome records demography history

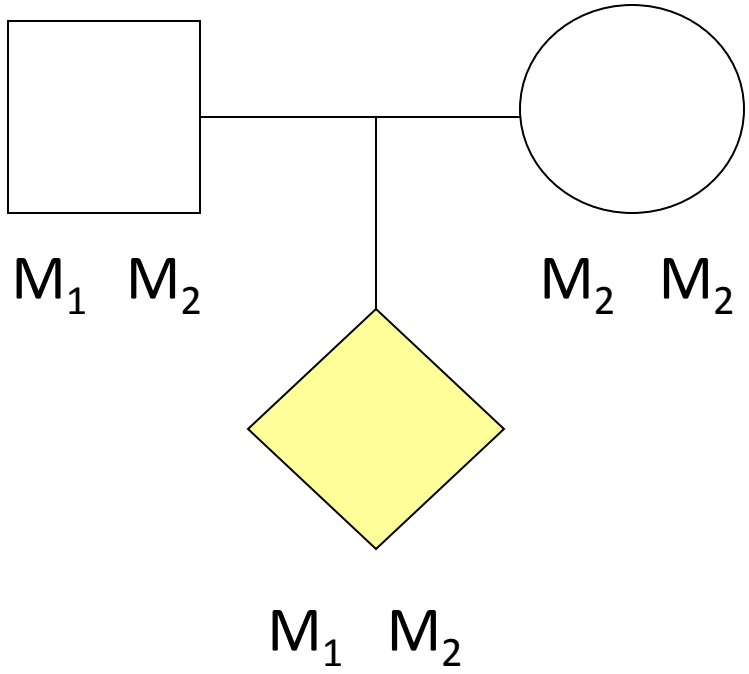
use genome to discover hidden structure

by looking at ‘null’ markers

Use familial controls = family-based association studies

Allelic controls – TDT

Transmitted alleles vs. non-transmitted alleles

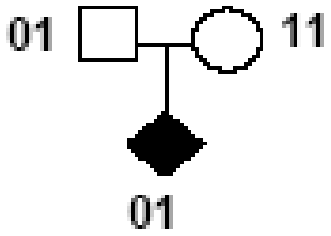


Transmitted alleles vs. non-transmitted alleles

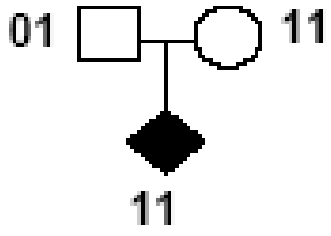
	Non-Transmitted Allele		
Transmitted		M_1	M_2
	M_1	n_{11}	n_{12}
	M_2	n_{21}	n_{22}

$$\mathbf{TDT} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \sim \chi^2 \text{ (1 df)}$$

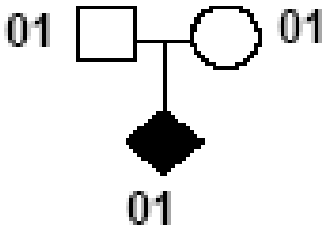
- 10 families



15 families



5 families



		Non-transmitted alleles	
		1	0
Transmitted allele	1	*10+15*	*15+5*
	0	*10+5*	*0*

- TDT statistic = $X^2 = \frac{(15-20)^2}{15+20} = 0.71 < 3.84$, so do not reject H_0
- ***Comment:*** n_{11} and n_{00} do not contribute

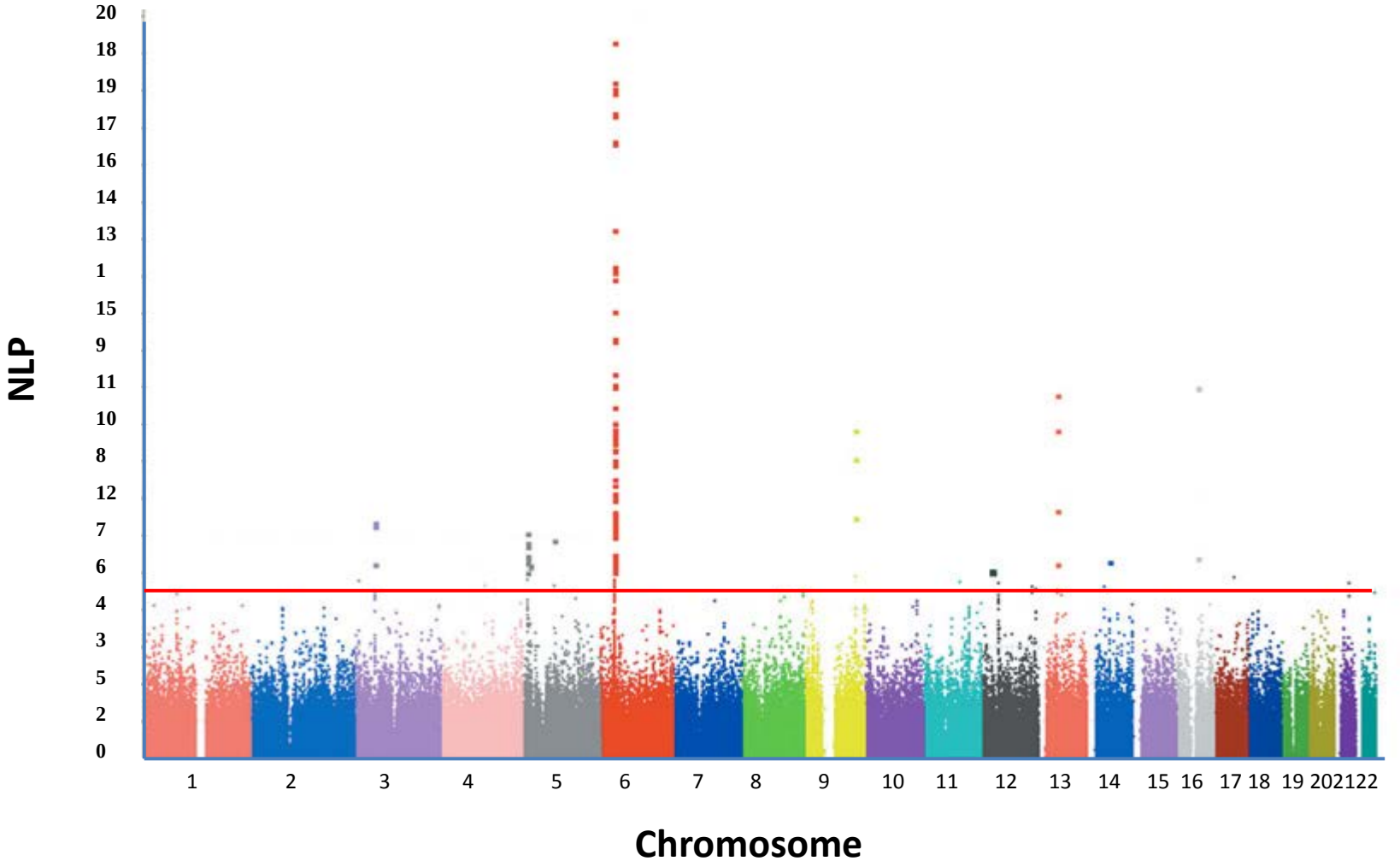
  →so homozygous parents do not provide information

Genome-wide association studies (GWAS)

"Study of the common genetic variation across the entire human genome designed to identify genetic association with observable traits"

NIH 2006

GWAS in leprosy



Zhang et al N Engl J Med 2009;361:2609-18.

GWAS in Infectious Diseases

Table 1 | **Markers significantly associated with infectious disease phenotypes in genome-wide studies**

Disease	Phenotype	Population	Sample size*	Most significant marker or markers	SNP location	*P* value*	Odds ratio	Refs
HIV-1 and AIDS	Viral load at set point‡	European	2,554	rs9264942	*HLA-C*	5.9×10^{-32}	NA	33,34
				rs2395029	*HLA-B, HCP5*	4.5×10^{-35}	NA	33,34
	Viral load at set point§	African American	515	rs2523608	*HLA-B*	5.6×10^{-10}	NA	38
	HIV-1 control‖	European	1,712	rs9264942	*HLA-C*	2.8×10^{-35}	2.9	35
				rs4418214	*MICA*	1.4×10^{-34}	4.4	
				rs2395029	*HLA-B, HCP5*	9.7×10^{-26}	5.3	
				rs3131018	*PSORS1C3*	4.2×10^{-16}	2.1	
		African American	1,233	rs2523608	*HLA-B*	8.9×10^{-20}	2.6	
				rs2255221	Intergenic	3.5×10^{-14}	2.7	
				rs2523590	*HLA-B*	1.7×10^{-13}	2.4	
				rs9262632	Intergenic	1.0×10^{-8}	3.1	
	Disease progression¶	European	1,071	rs9261174	*ZNRD1, RNF39*	1.8×10^{-6}	NA	33,34
	Progression to AIDS 1987#	European American	755	rs11884476	*PARD3B*	3.4×10^{-9}	NA	41
	Long-term nonprogression‡	European	1,627	rs2395029	*HLA-B, HCP5*	6.8×10^{-10}	3.47	42
	Long-term nonprogression§	European	1,911	rs2234358	*CXCR6*	9.7×10^{-10}	1.85	43
Hepatitis C	Spontaneous clearance	European	1,362	rs8099917	*IL28B*	6.1×10^{-9}	2.31	53
Hepatitis B	Chronic infection	Japanese, Taiwanese	6,387	rs3077	*HLA-DPA1*	2.3×10^{-16}	0.56	60
				rs9277535	*HLA-DPB1*	6.3×10^{-19}	0.57	
Dengue	Dengue shock syndrome	Vietnamese	8,697	rs3132468	*MICB*	4.4×10^{-11}	1.34	65
				rs3765524	*PLCE1*	3.1×10^{-10}	0.80	
Severe malaria	Susceptibility	African (Gambian)	5,900	rs11036238	*HBB*	3.7×10^{-11}	0.63	70
Tuberculosis	Susceptibility	African (Ghana, The Gambia, Malawi)	11,425	rs4334126	18q11.2 (*GATA6, CTAGE1, RBBP8, CABLES1*)	6.8×10^{-9}	1.19	72
Leprosy	Susceptibility	Chinese	11,140	rs3764147	*LACC1*	3.7×10^{-54}	1.68	76
				rs9302752	*NOD2*	3.8×10^{-40}	1.59	
				rs3088362	*CCDC122*	1.4×10^{-31}	1.52	
				rs602875	*HLA-DR-DQ*	5.4×10^{-27}	0.67	
				rs6478108	*TNFSF15*	3.4×10^{-21}	1.37	
				rs42490	*RIPK2*	1.4×10^{-16}	0.76	
Meningococcal disease	Protection	European	7,522	rs1065489	*CFH*	2.2×10^{-11}	0.64	85
				rs426736	*CFHR3*	4.6×10^{-13}	0.63	
Variant Creutzfeldt–Jakob disease	Susceptibility	European, Papua New Guinea	5,183	rs1799990	*PRNP*	2.0×10^{-27}	NA	91