

Key terms for Genetic Epidemiology

Words, words, words! I'm so sick of words!

Is that all you blighters can do?

-- Eliza Doolittle, My Fair Lady

Introduction

- Genetic epidemiology is a hybrid discipline, combining the analytical tools of epidemiology & genetics to identify genes underlying complex diseases.
- Complex diseases result from both genes & environmental factors that control risk.
 - It is not “Nature vs Nurture” but “Nature & Nurture”
- Genetic terms have very specific meanings but are often poorly understood & mis-used.
- Words do matter.

Summarizing History of Genetics

Century	19 th	20 th				21 st	
Landmarks	Mendel	Mendel Rediscovered		DNA Structure	Human Genome Project		
Disease		Rare Mendelian		Multifactorial or Complex		All?	?
Mapping				Linkage (2 point → genome wide)		Genome wide association	?
Models		4 Simple models (AD, AR, XR & XD)		One gene, one protein	Models for complex traits	Multiple genes & interaction	?
Gene Expression				Gene action through mutants		Gene Regulation	?
Testing				Neonatal Screening	Diagnostic Testing	Screening for Susceptibility	?

Terms to watch for:

1. Allele, genotype, haplotype, diplotype
2. Familial correlation, heritability, variance components models
3. Linkage, recombination, allele sharing, identical-by-descent (IBD), Quantitative Trait Locus (QTL)
4. Association, allelic effects, linkage disequilibrium (LD), Type I error, confounding/population stratification
5. Case-control, case-only, & case-family designs; family studies
6. Odds ratios (OR), logistic regression, linear models

Remember definitions:

Haplotypes are combinations of alleles at different markers

- 2 markers A & B, with 2 alleles each

- 3 genotypes each

- AA, Aa & aa
 - BB, Bb & bb

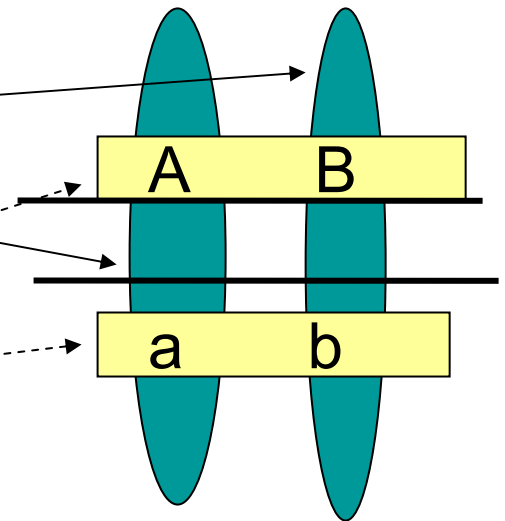
- For n alleles, $\text{genotypes} = \frac{n(n+1)}{2}$

- 4 haplotypes

- AB, Ab, aB & ab

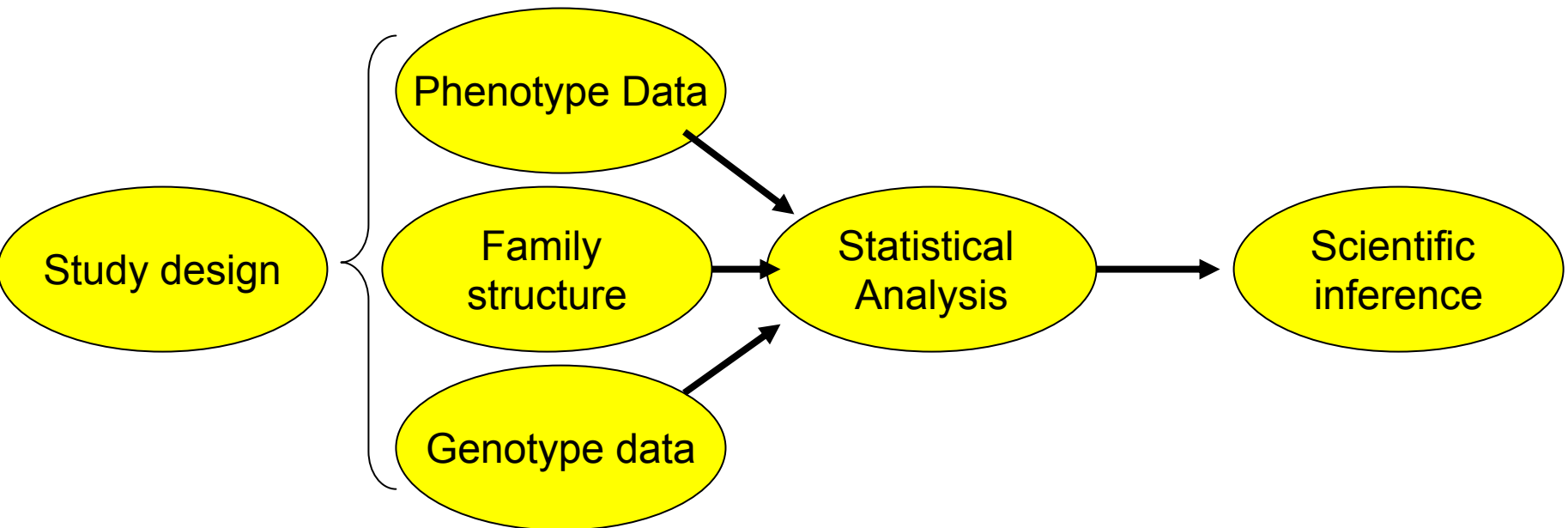
- Haplotypes \rightarrow alleles

- Diploypes \rightarrow genotypes



Do not confuse

- **Study design:** case-control (all unrelated), case-family designs, family studies
 - Population based ascertainment vs. selective sampling
- **Genotype data:** SNP, STR markers, sequencing
 - candidate gene vs. genome wide
- **Analytical model:** Linkage, Association, Modeling inheritance of phenotype
- **Hypotheses being tested:** Inferences are build upon these

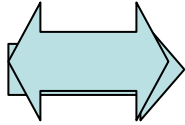


Linkage vs. Association

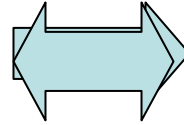
Genetics

Epidemiology

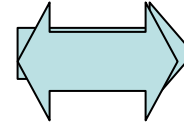
Parametric linkage LODs



Non-Parametric Linkage using Allele sharing



Family based Association tests



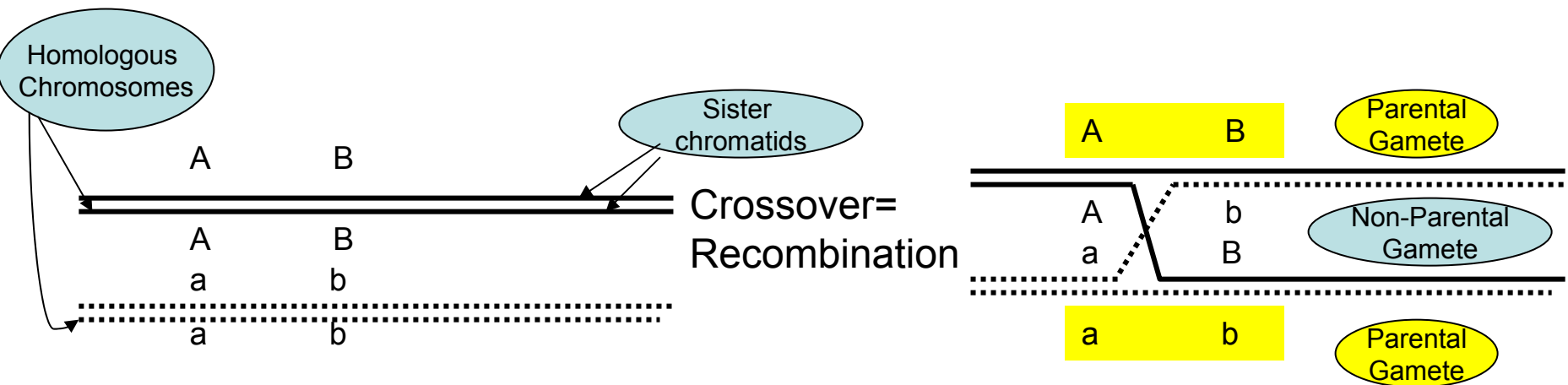
Case-Control Designs

1. Requires multiplex families
 - Bigger is better
2. Guaranteed to work for Mendelian diseases
 - Genome wide studies are feasible
3. Still useful for complex diseases
 - Locus heterogeneity (linked & unlinked families) is a problem
4. Meta-analysis is possible

1. Unrelated cases & controls can be used
2. Can incorporate tests for G, E, GxE, GxG, etc.
3. Meta-analysis can measure consistency across studies
 - Or lack thereof
4. Allelic heterogeneity is a problem.
5. Genome wide studies are now feasible (but expensive)
 - Interpreting will be a challenge

Linkage is 'co-segregation'

- Evidence of **recombination fraction** $\theta < 0.5$ implies co-segregation of 2 genes within a family
 - H_0 : no linkage \rightarrow independent assortment of 2 genes (e.g. marker & gene controlling trait)
 - Recombination fraction (θ) = 0.5
 - This is Mendel's 2nd law
 - H_A : linkage $\rightarrow \theta < 0.5$ reduced recombination
 - 2 genes are syntenic (on the same chromosome) & linked
 - Due to their close physical proximity, 2 genes recombine less often



Linkage can involve

- 2 genetic markers
 - Markers can include genetic phenotype (ABO blood type), anonymous DNA variant (STR, SNP, etc.)
- Genetic marker & causal gene for Mendelian disease
 - This maps a causal gene to a known region of the genome
- Genetic marker & putative causal gene underlying complex phenotype
 - Here evidence for linkage provides both evidence that a causal gene does exist & its location in the genome

Two broad types of linkage analysis

- *Model based or parametric* linkage analysis
 - Specify a model & try to estimate key parameters
 - Maximum likelihood methods are used
 - Test hypothesis & estimate θ in one analysis
- *Model free or non-parametric* linkage analysis
 - Test for consequences of linkage: excess sharing of marker alleles among relatives with same phenotype (e.g. affected pairs of relatives)
 - No particular model of inheritance is specified
 - No attempt to estimate underlying recombination

Linkage is based on meiosis

- Parametric linkage relies on reconstructing meiotic events in an informative parent
 - Only double heterozygotes are informative
 - Allele frequency determines information content
 - Count *recombinant* (non-parental phase) & *non-recombinant* (parental phase) offspring
- Non-parametric linkage tests H_0 : normal IBD sharing for affected relatives. Excess IBD sharing \rightarrow linkage between marker & unobserved causal locus
- Lots of terminology, often carelessly used

Association between Disease & Marker

Unrelated People

	Case	Control
Genotype +	A	B
Genotype -	C	D

H_0 : Allele/genotype frequencies are equal in cases & controls

Test statistics:

- Chi-square tests for independence
- Odds ratio of being a case given genotype $OR=AD/CB$
 - OR is a measure of association, $H_0: OR=1$
- Logistic regression to add covariate effects

$$\ln[P(\text{case})/\{1-P(\text{case})\}] = \beta_0 + \beta_1 X$$

$$\rightarrow OR(\text{case}|X=0)=\exp(\beta_0) \quad \& \quad OR(\text{case}|X=1)=\exp(\beta_0+\beta_1)$$

Allelic or Genotypic Comparison

Marker alleles

	1	2	...	K
Case				
Control				

2xK table
Total=2N

Marker Genotypes

	11	12	13	KK
Case					
Control					

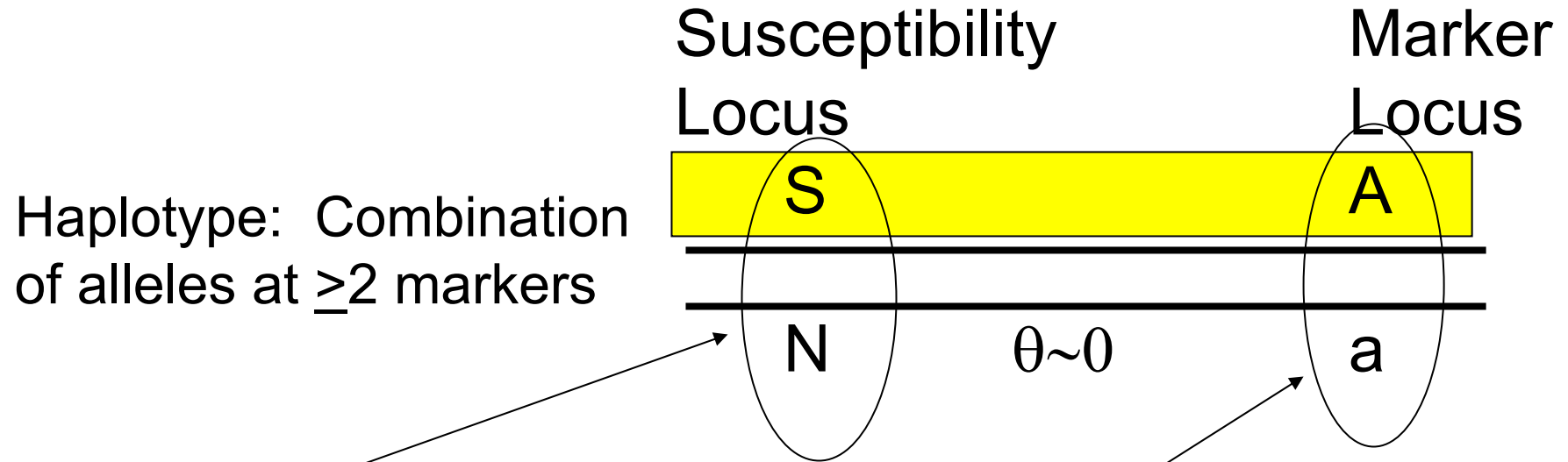
$2 \times \{K(K+1)\} / 2$
table
Total=N

Sparse data will eventually become a problem

Possible explanation for statistically significant disease-marker association

1. Type I error
2. Marker allele is in the causal pathway (“direct effect”)
 - PiZ homozygotes & emphysema (RR=20)
3. Linkage disequilibrium between observed marker & “high risk” allele at *unobserved* susceptibility locus
 - Must consider population history; disequilibrium could be due to genetic drift, selection or admixture
4. Confounding between unrecognized sub-groups (“Simpson’s paradox” or population stratification)
 - Both marker allele frequency & disease prevalence vary across strata of population

3. Indirect effect: Linkage disequilibrium



- Unobserved susceptibility locus with alleles S & N
 - frequency $P(S)=p$ & $q=1-p=P(N)$
- Observed marker locus with alleles A & a
 - frequency $P(A)=r$ & $s=1-r=P(a)$

3. Linkage disequilibrium (cont'd)

- Disequilibrium occurs when alleles at different loci (haplotypes across ≥ 2 markers or between a susceptibility locus & a marker) occur in frequencies other than expected under Hardy Weinberg equilibrium (HWE)
- Gametic equilibrium (HWE)
 - $P(SA) = P(S)P(A) = pr$
- Gametic disequilibrium
 - $P(SA) = P(S)P(A) + D = pr + D$

Deviation from HWE



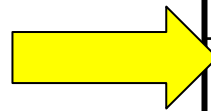
Two locus gametic equilibrium vs. disequilibrium at population level

	A	a	Total
S	pr	ps	p
N	qr	qs	q
Total	r	s	1.0

Disease locus: Allele 'S' with freq. p;
 'N' with freq q;
 Marker locus: Allele 'A' with freq. r;
 'a' with freq. s

← HW equilibrium

HW disequilibrium

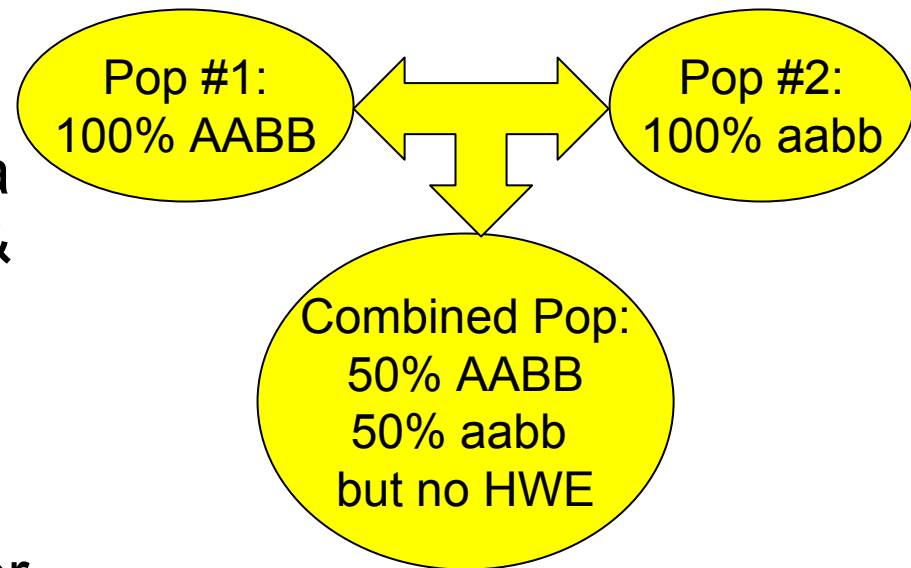


	A	a	Total
S	pr+D	ps-D	p
N	qr-D	qs+D	q
Total	r	s	1.0

Because p remains constant
 Increase of one haplotype implies
 decrease of another

Gametic Disequilibrium in a population could be due to:

1. Selection for/against a haplotype
2. Genetic Drift
3. Migration between populations & admixture
 - Will approach HWE within a few generations (<10) if A & B are unlinked ($\theta=0.5$)
4. Tight linkage between 2 loci
 - Will stay in disequilibrium for many generations if θ is small



Several measures of LD

- $D = P(SA) - P(S)P(A) = P(SA) - pr$
 $= P(SA)P(Na) - P(Sa)P(NA) =$
 $= ad - bc$

(range depends on allele frequency)

- **Standardized $D' = D / \max D$**

*(range -1 to 1, where $\max D = \min(ps, qr)$ if $D > 0$,
or $\max D = \min(pr, qs)$ if $D < 0$)*

- $OR = ad/bc$

- $P_{\text{excess}} = (ad - bc) / rd$

(like attributable risk)

- **$r^2 = [(ad - bc) / (pqrs)]^2$**

- $d = a/r - b/s = (ad - bc) / rs$

- $Q = OR - 1 / OR + 1 = (ad - bc) / (ad + bc)$

	A	a	
S	P(SA) =a	P(Sa) =b	P(S) =p
N	P(NA) =c	P(Na) =d	P(a) =q
	P(A) =r	P(a) =s	1.0

D' vs. r^2 as measures of LD

- D'
 - Easy to understand (standardized D)
 - Not highly dependent on allele frequency
 - Very dependent on sample size
 - Decay rate is slower with more noise
 - D'=1 implies complete LD
 - One haplotype (of 4 possible) is not present
 - Possible haplotypes = 2^n for n SNP (biallelic)
- r^2
 - Correlation between alleles at different loci
 - $r^2 = 1$ implies knowing 1 allele predicts the allele at another locus
 - Two haplotypes (of 4 possible) are not present
 - One marker becomes redundant
 - Not so dependent on sample size
 - More dependent on allele frequency
 - Decay rate is faster with less noise

Linkage Disequilibrium (LD)

- LD is deviation from independence ($D \neq 0$) of alleles at different loci
- LD can be measured as correlation between alleles at different loci
- $D' \neq 0$ implies some haplotypes are more common than expected in the population
- LD is population specific, with substantial variation across populations

Most common hap in EA

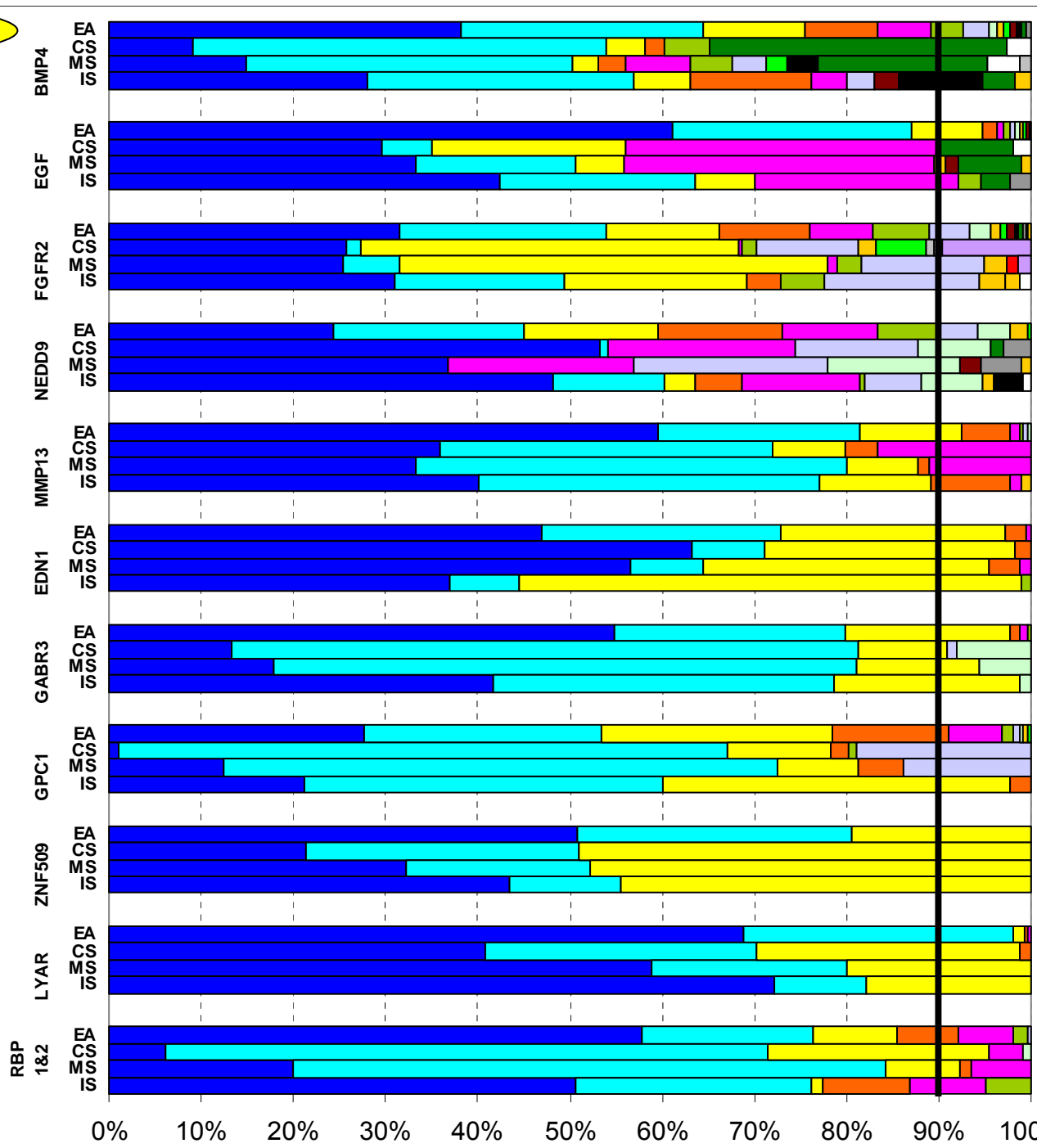
Haplotype frequencies vary across populations:

8 genes in 4 populations

Beaty et al

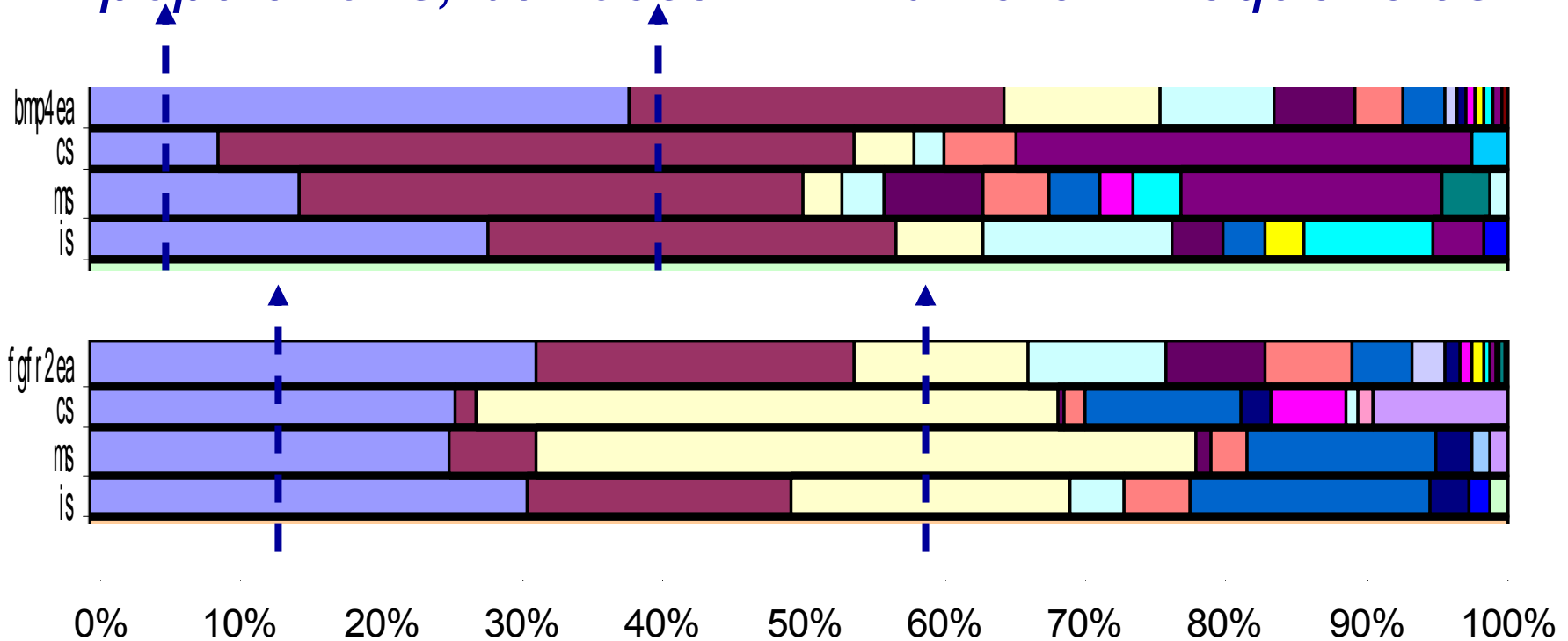
(2005) GENETICS

171:259-267



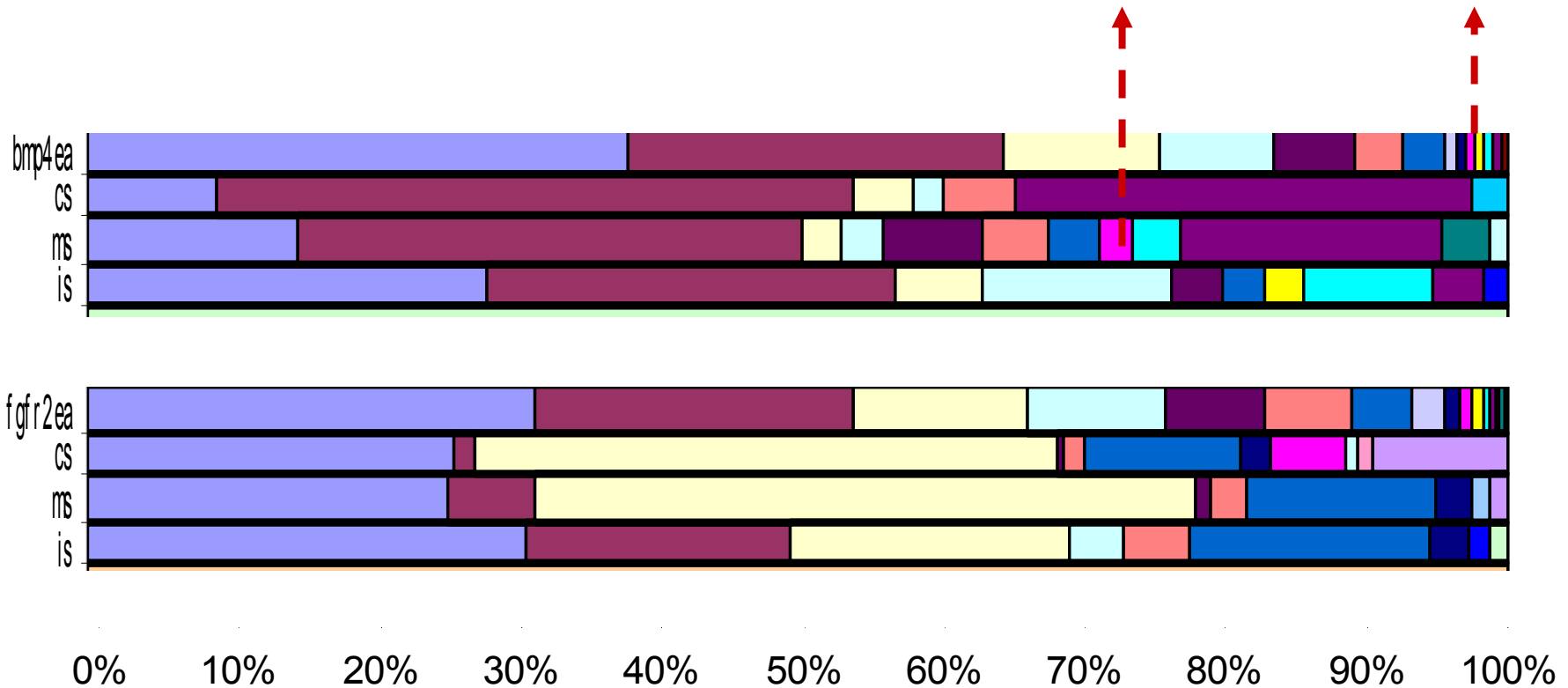
Haplotype frequencies/diversity differ by population

Some haplotypes are common to several populations, but occur with different frequencies



Haplotype frequencies/diversity differ by population

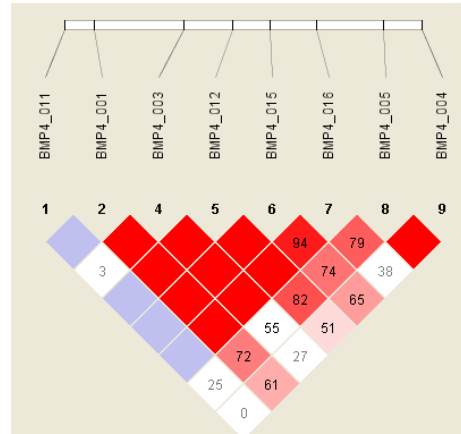
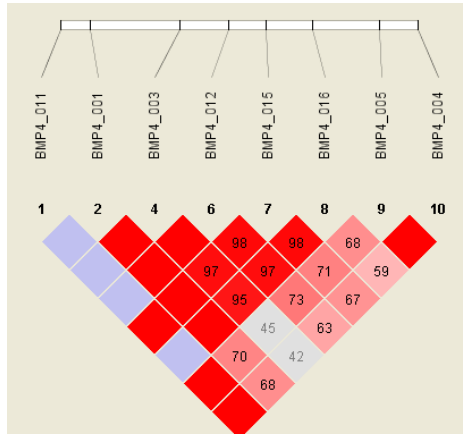
Some haplotypes are unique to particular populations



SNPs within gene show strong LD

BMP4

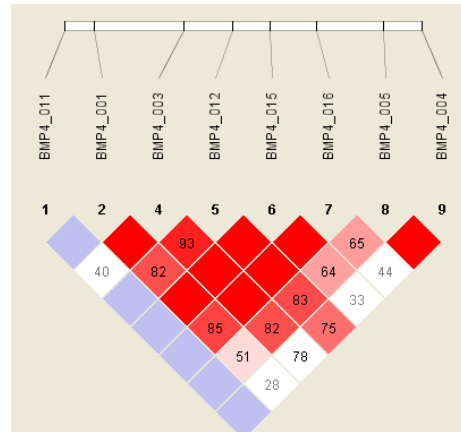
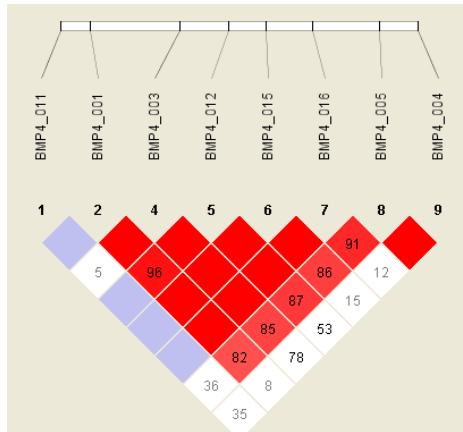
European American
(n=135)



Red: Significant LD
Dots: Almost complete LD
Blue: Not significant LD
White: Markers are independent

Chinese Singaporean
(n=57)

Indian Singaporean
(n=46)



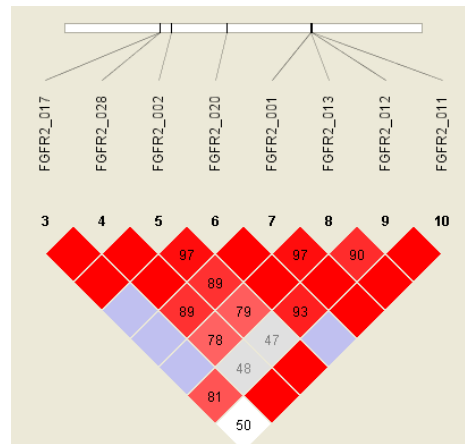
Malay Singaporean
(n=45)

Variation in haplotype frequencies: 6.6% among populations, 93.4% within population

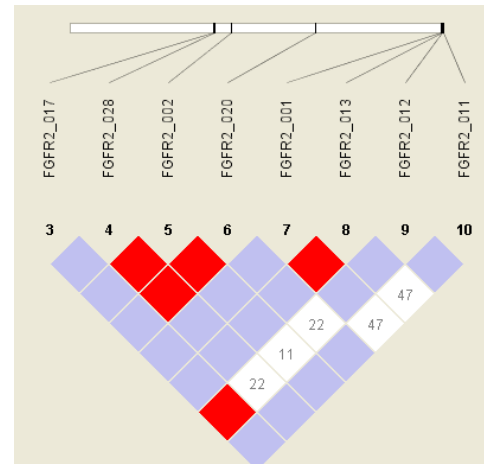
Another example of LD within gene

FGFR2

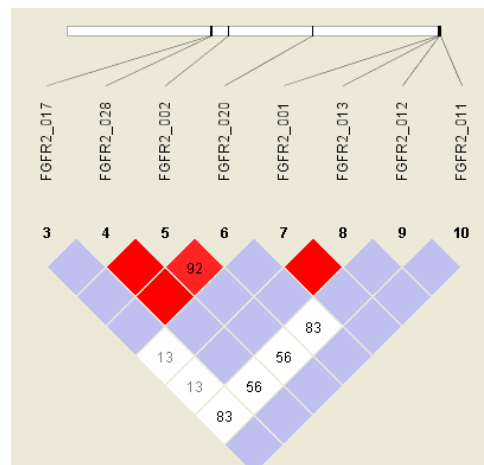
European
American



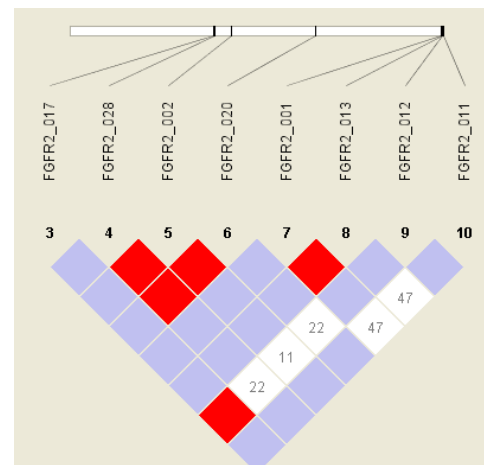
Chinese
Singaporean



Indian
Singaporean

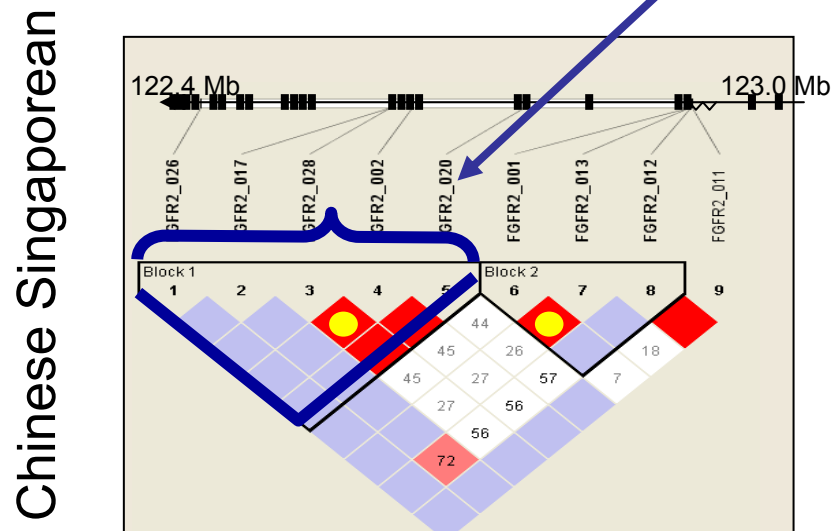
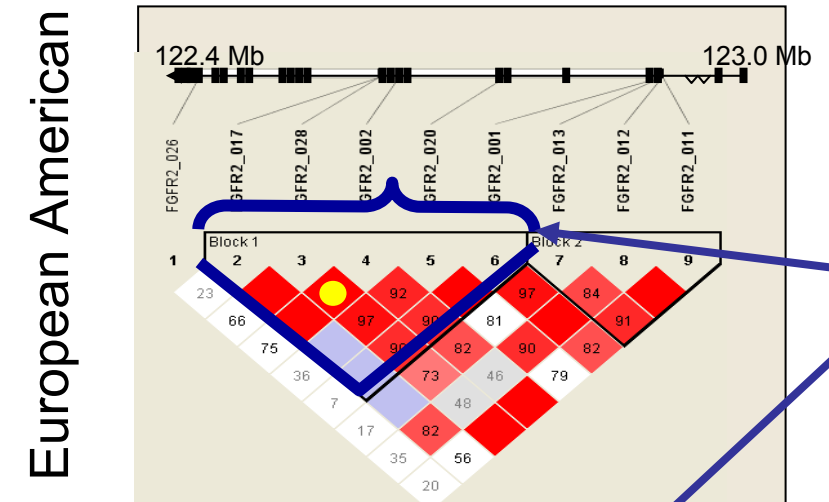


Malay
Singaporean



Variance in haplotype frequencies: 6% among populations, 94% within populations

Block structure & SNPs to genotype depend on population



How can population genetics help in genetic epidemiology?

- Case-control designs & their variations test genes as risk factors
- If a genetic marker is a risk factor, what does this mean?
 - Is it causal or is it in LD with a causal gene?
- How does evidence of LD depend on population?
- How do we prove causality in genetic epidemiology?