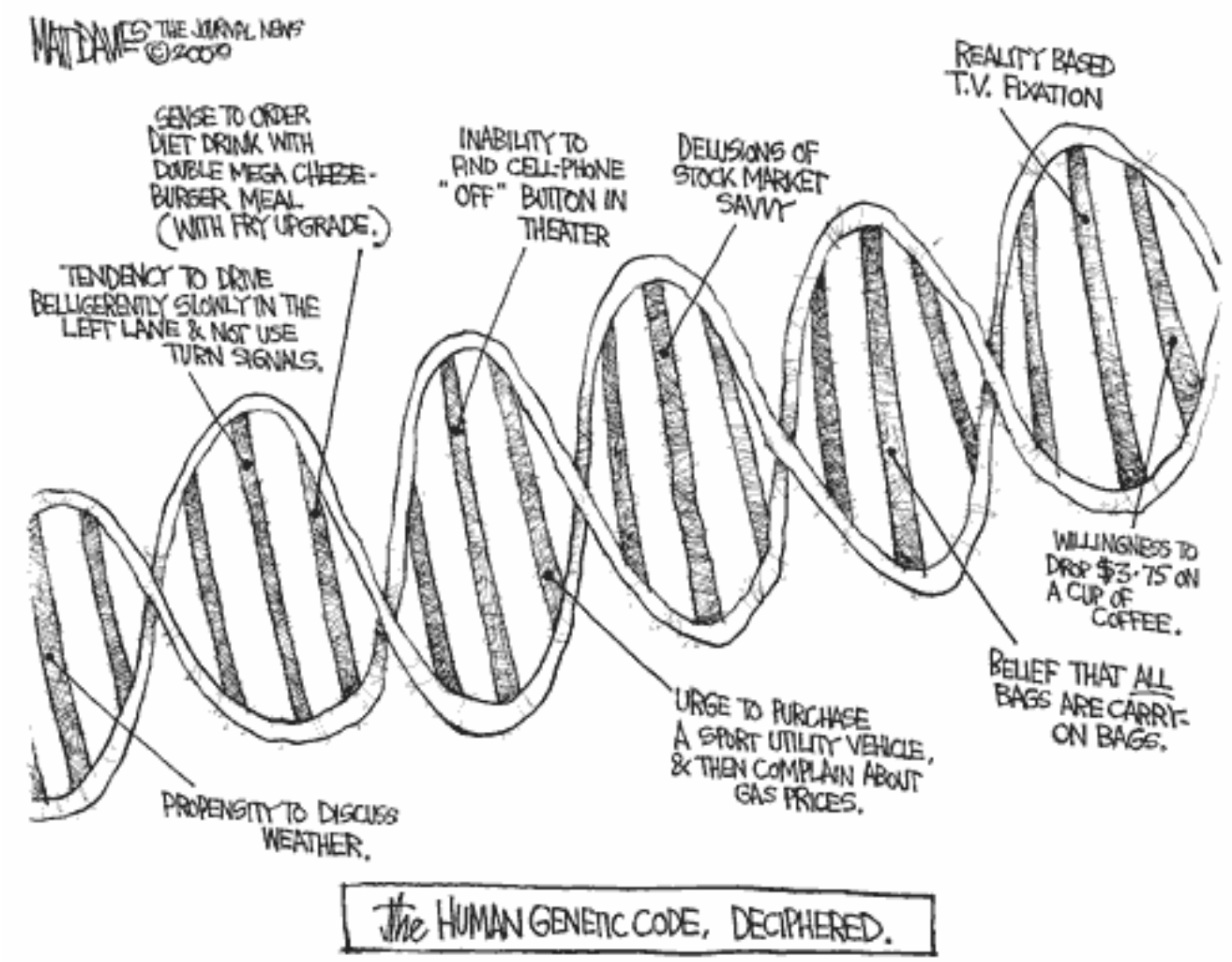
Epidemiology 217
Lecture #2: Population Genetics



Outline

Several methods to answer the question: Is the trait genetic? [Austin Ch. 2]

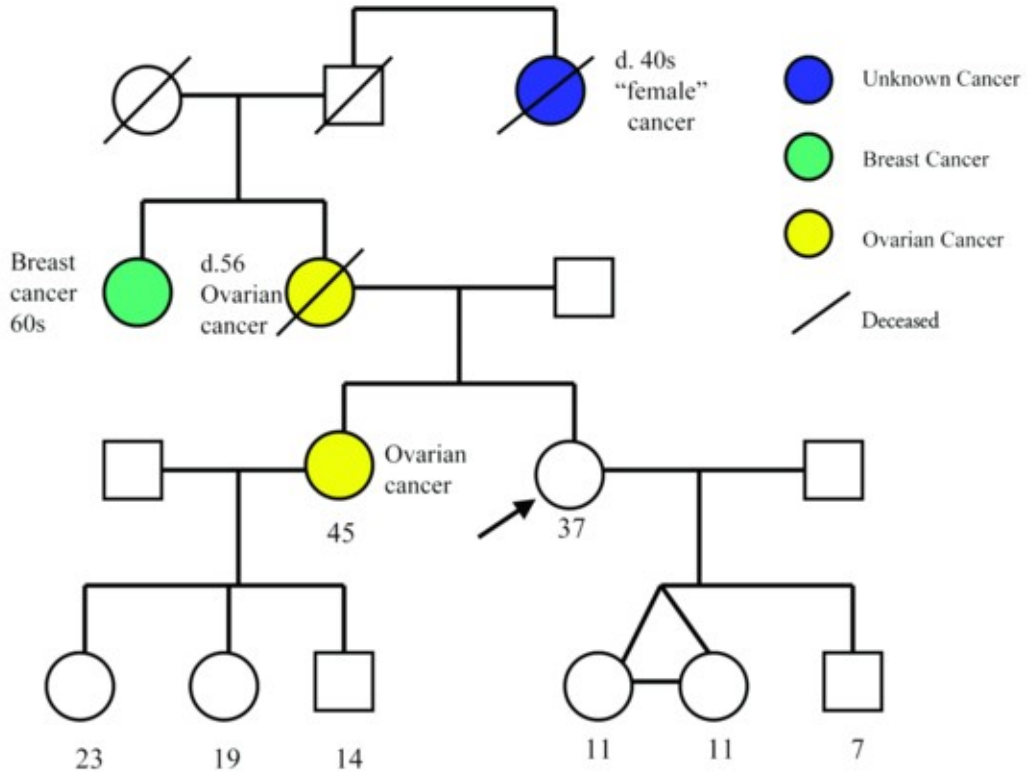
1. Familial Aggregation and recurrence risks
2. Heritability

Genetic concepts [Austin Ch. 3]

3. Allele Frequency Estimation
4. Hardy-Weinberg equilibrium (HWE)
5. Population Substructure/Stratification

1. Familial Aggregation

- Does the phenotype tend to run in families?
- Ascertain family members by “probands”, i.e., affected individuals



Familial aggregation

"Family case-control" approach

Family history of disease

	Yes	No
# Cases	a	b
# Controls	c	d

Odds Ratio = OR
= (a/b)/(c/d) = ad/bc

"Family cohort" approach

Relative has disease?

	Yes	No
# Relatives of Cases	a	b
# Relatives of Controls	c	d

Relative Risk = RR
= (a/(a+b))/(c/(c+d))

- Be careful with "family history"
  - E.g., breast cancer, only females
  - E.g., late onset diseases, relative may not generally have disease
- Generally done by interview (cost reasons), though medical records would be more accurate

Khoury et al., 1993, and in Austin

Familial aggregation

	No of Cases (%)	No of Controls (%)	Adjusted OR (95% CI)
Family history of pancreatic cancer in 1st-degree relative			
No	1,107	1,152	1.0 (ref)
Yes	76	43	1.76 (1.19, 2.61)
Number of affected 1st-degree Relatives with pancreatic cancer			
One	70	42	1.70 (1.14, 2.53)
Two or more	6	1	4.26 (0.48, 37.7)

Jacobs et al., 2010; Austin

Recurrence ('Familial') Risk Ratios

- Compares the probability a subject is affected given they have an affected family member to the population risk:

  $\lambda_R = K_R/K$,

  where K_R is the risk to relatives of type R

  K is the population risk

λ_S = recurrence risk to siblings of probands versus the general population risk.

(Recurrence Risks are not covered in Austin)

Estimating Recurrence Risk Ratios (RRR)

- With case-control data, calculate RRR as:

  Proportion of affected relatives of the cases (observed) /

  Proportion of affected relatives of controls (expected) (assumed to estimate K, the population prevalence of disease)

$$\text{Using } \lambda_R = P(Y_2 = 1 \mid Y_1 = 1) / K$$

- The higher the value of λ_R the stronger the genetic effect

Examples of λ_s

• Alzheimer Disease	3-4
• Rheumatoid Arthritis	12
• Schizophrenia	13
• Type I Diabetes	15
• Multiple Sclerosis	20-30
• Neural Tube Defects	25-50
• Autism	75-150

Limitations of Recurrence Risks

- Depend on mode of inheritance and disease frequency.
- Single gene diseases have high recurrence risks.
- More common complex diseases have lower values (e.g., CHD).
- Hard to distinguish genetic versus environmental effects.

2. Heritability Analysis

- Evaluates the genetic contribution to a trait in terms of variance explained.

- Phenotype = Mean + Genetics + Environment

$$P = \mu + G + E$$

- Overall variation in Phenotype P

$$Var(P) = \sigma^2_P = \sigma^2_G + \sigma^2_E$$

Broad Sense Heritability

- Proportion of the overall phenotypic variance attributable to genetic influences

$$H^2=\sigma^2_G/\sigma^2_P$$

Narrow Sense Heritability

- Proportion of variance explained only by additive genetic effects.
- Most commonly calculated estimate since additive explains most variation.

  (Recall: $\sigma^2_P = \sigma^2_G + \sigma^2_E$)

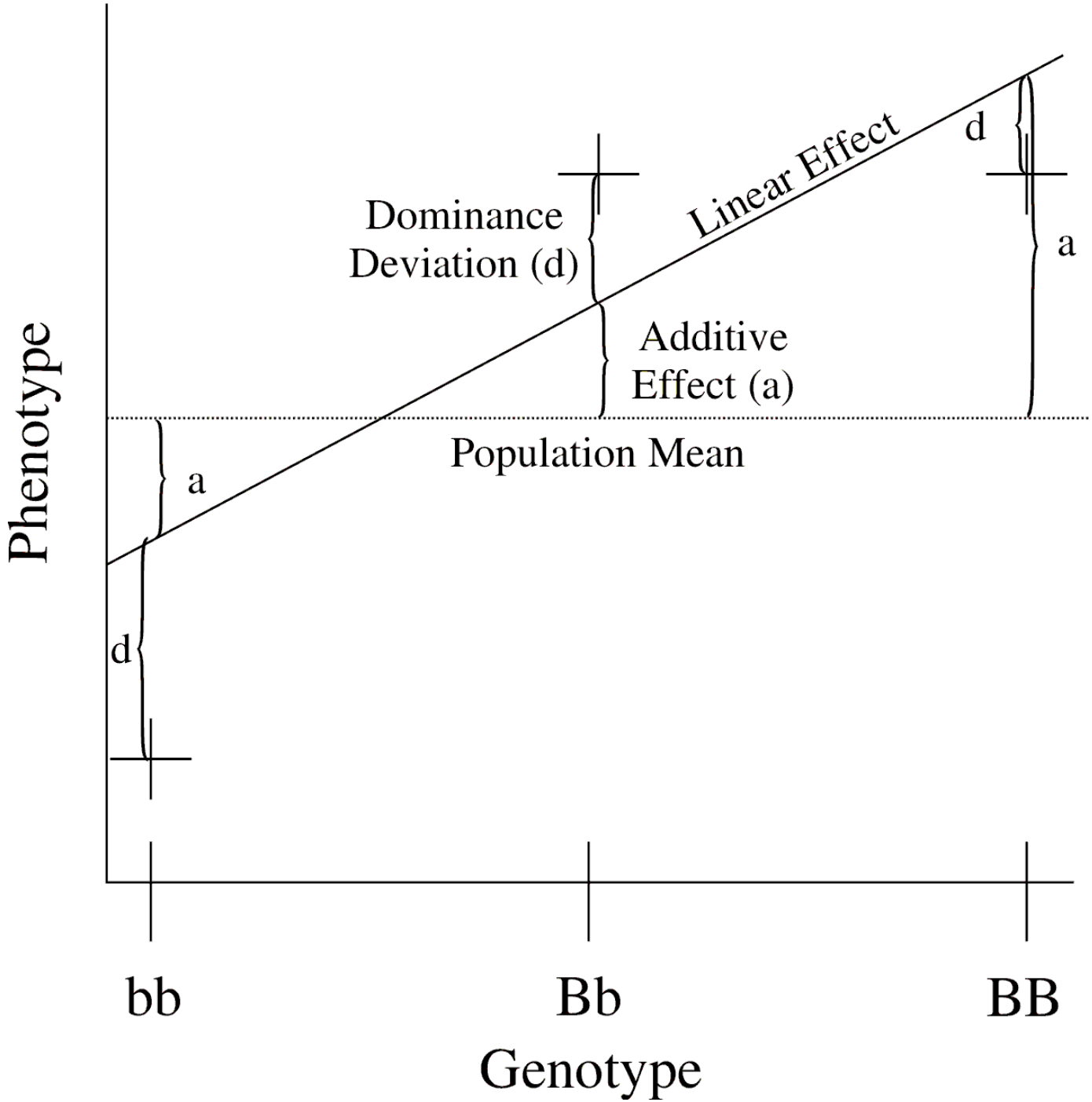
  $\sigma^2_G = \sigma^2_A + \sigma^2_D$

  A: Additive effect

  D: Dominance effect

"Broad sense": $H^2=\sigma^2_G/\sigma^2_P=(\sigma^2_A + \sigma^2_D)/\sigma^2_P$

"Narrow sense": $h^2 = \sigma^2_A/\sigma^2_P$


Phenotype
Linear Effect
d
Dominance
Deviation (d)
a
Additive
Effect (a)
Population Mean
a
d
bb
Bb
BB
Genotype

Narrow Sense Heritability

- Can estimate from regression:
  - Additive: 0, 1, 2 for alleles
  - Dominance: 0, 1, 0 for departure
- *Or estimate from twin studies...*

Liability model

- Two unlinked loci, with allele a and risk allele A; and allele b and risk allele B.
- Increase the risk by 1 unit on liability scale.

Two copies of capital letter (risk allele)

Table 2.3. Genotypes, liability scores, and genotype frequencies under a polygenic model with two loci.

Genotype(s)	Liability score	Frequency
aabb	0	1
Aabb, aAbb, aaBb, aabB	+1	4
AAbb, AaBb, AabB, aABb, aAbB, aaBB	+2	6
AABb, AAbB, AaBB, aABB	+3	4
AABB	+4	1

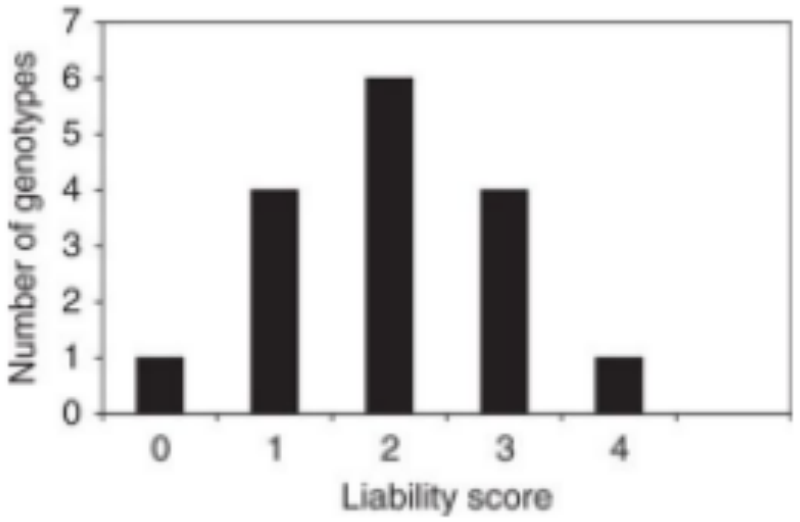


Fig. 2.1. Frequency distribution of genotypes by liability score under a polygenic model with two loci.

Austin 2013

- *Three* unlinked loci Aa, Bb, Cc (capital is risk allele)
- Increase the risk by 1 unit on liability scale
- If dichotomous, threshold determines disease status

Table 2.4. Genotypes, liability scores, and genotype frequencies under a polygenic model with three loci.

Genotype(s)	Liability score	Frequency
aabbcc	0	1
Aabbcc, aAbbcc, aaBbcc, aabBcc, aabbCc, aabbcC	+1	6
AAbbcc, AaBbcc, AabBcc, ...	+2	15
AABbcc, AAbBcc, AaBBcc, ...	+3	20
AABBcc, AABbCc, AABbcC, ...	+4	15
AABBCc, AABBcC, AABbCC, ...	+5	6
AABBCC	+6	1

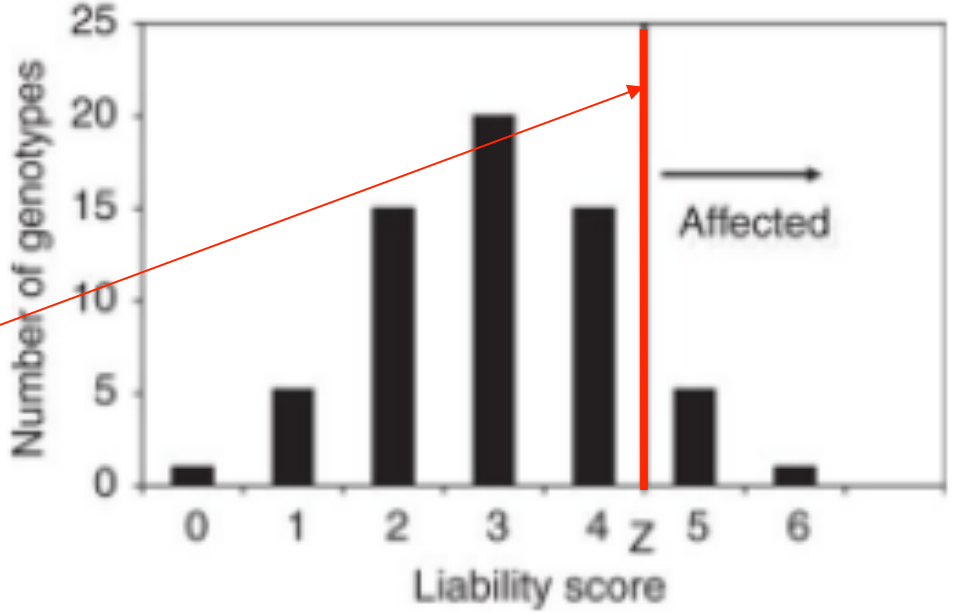
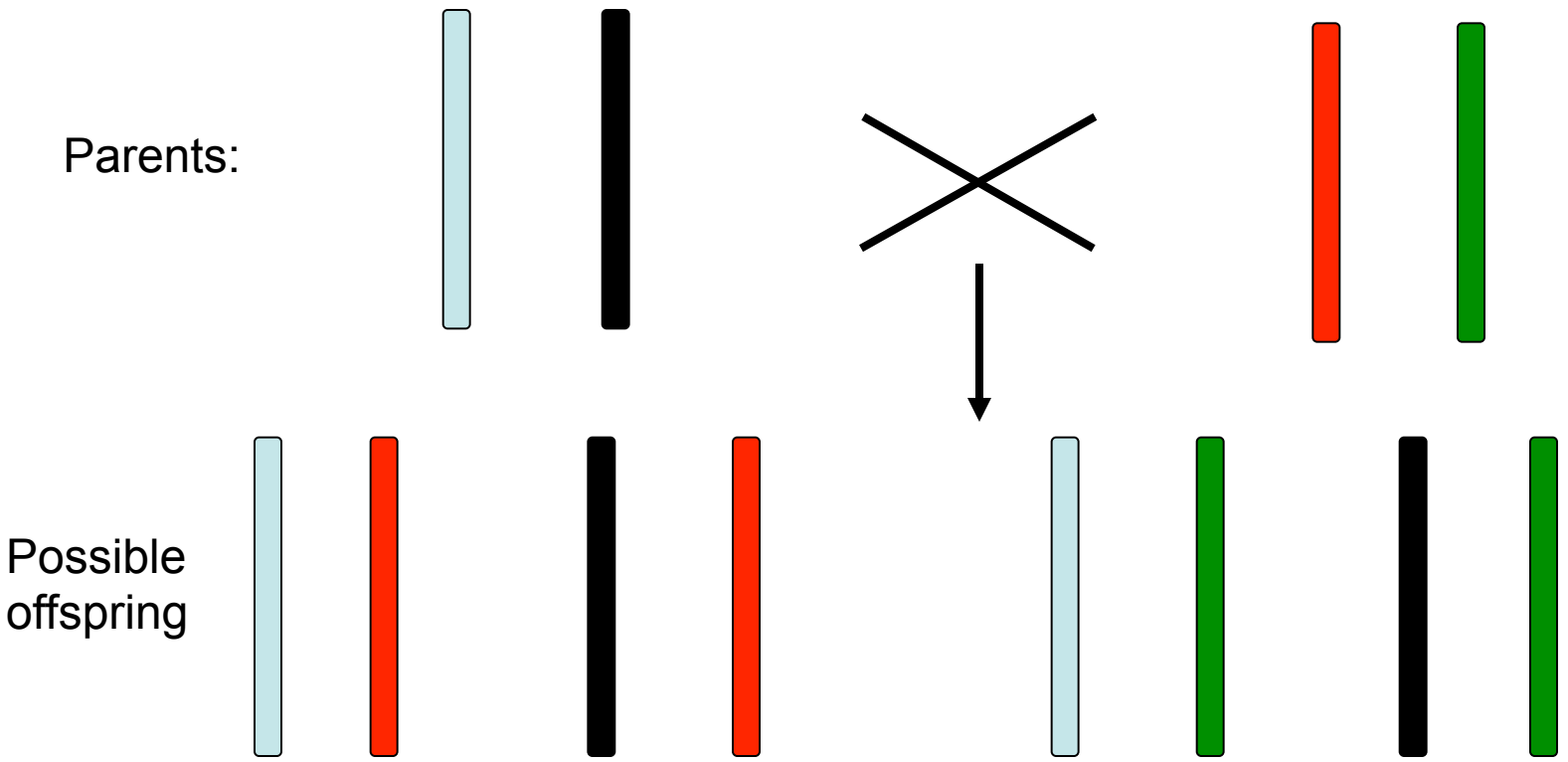


Fig. 2.2. Frequency distribution of genotypes by liability score under a polygenic model with three loci and threshold value Z for dichotomous traits.

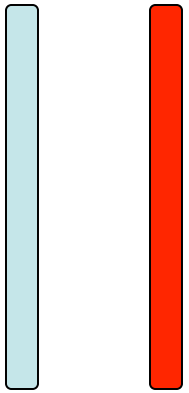
Twin Studies

- Compare the phenotype correlation or disease concordance rates of MZ (identical) and DZ (fraternal) twins.



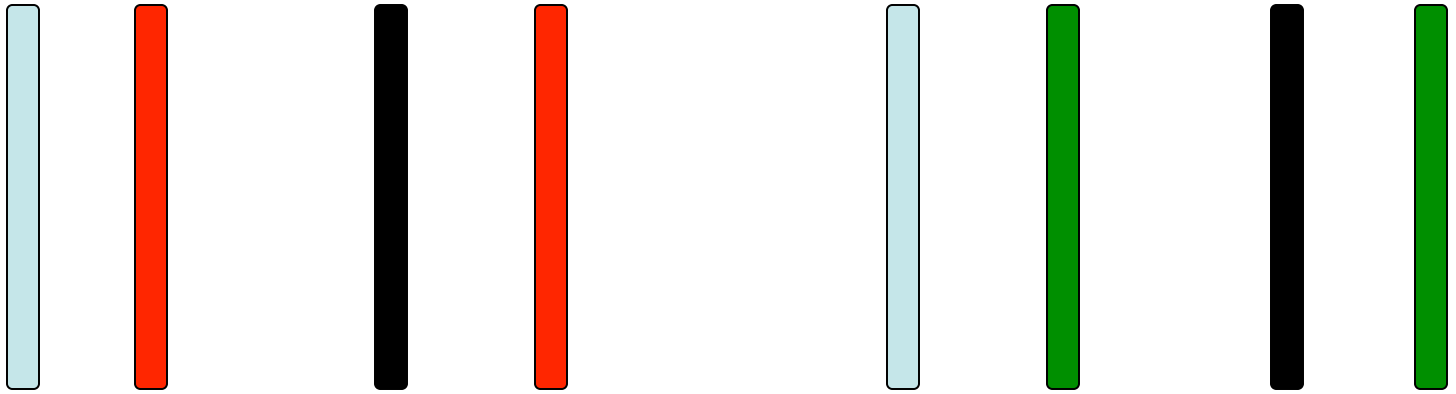
MZ Twins (Identical)

If twin 1 is



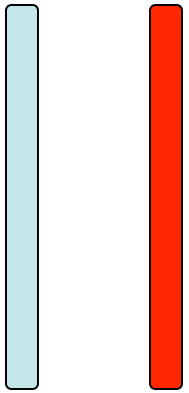
Both alleles are shared identical by descent (IBD)

Then twin 2 must be:



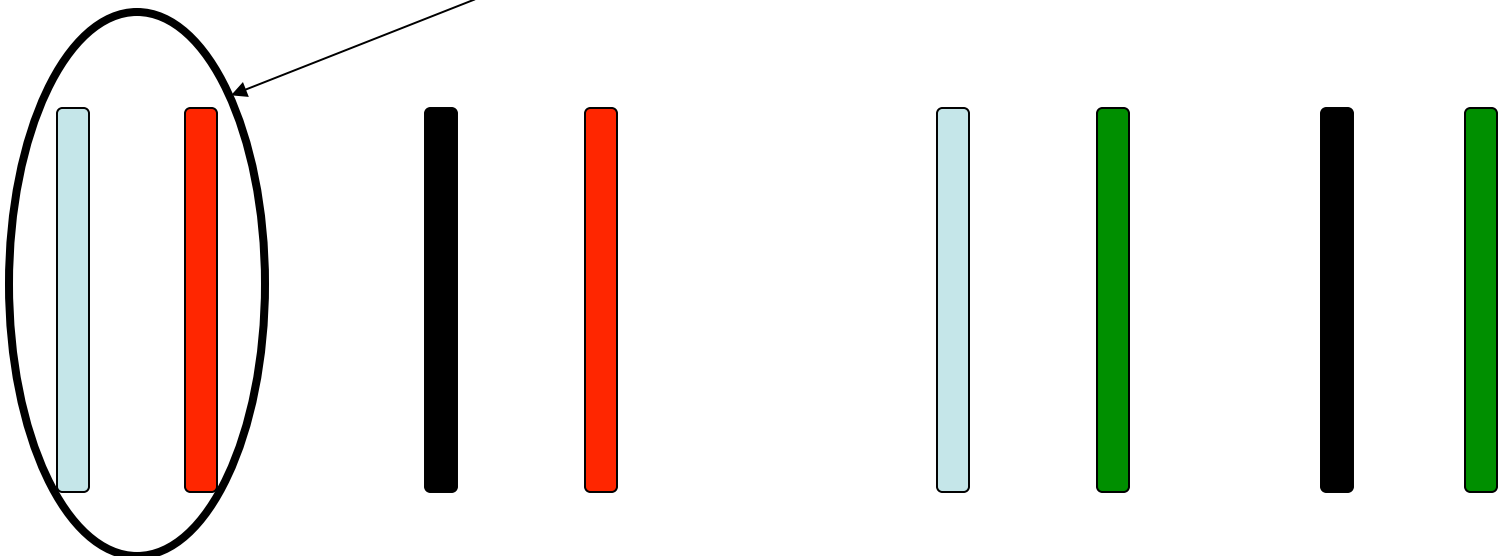
MZ Twins (Identical)

If twin 1 is

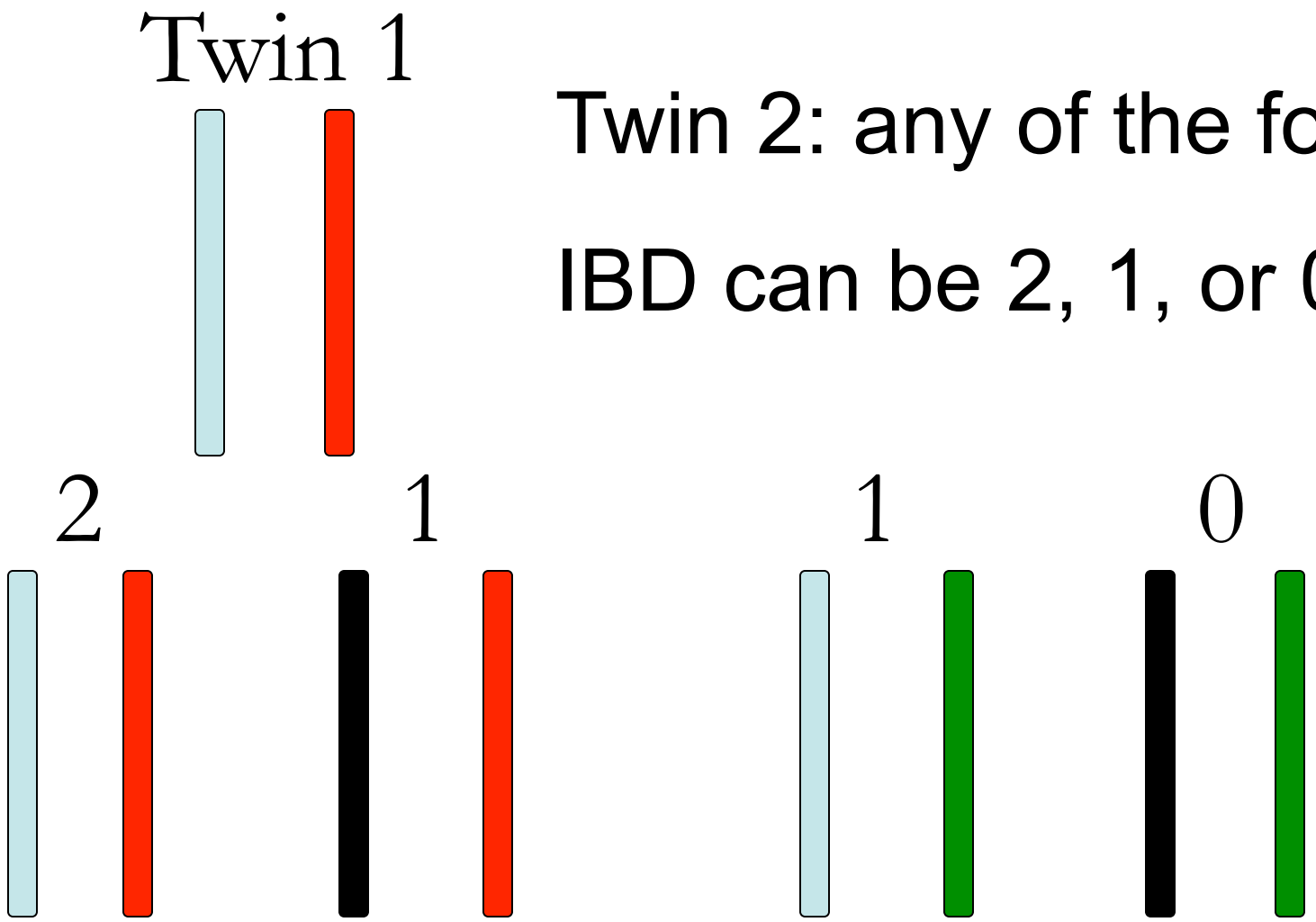


Both alleles are shared identical by descent (IBD)

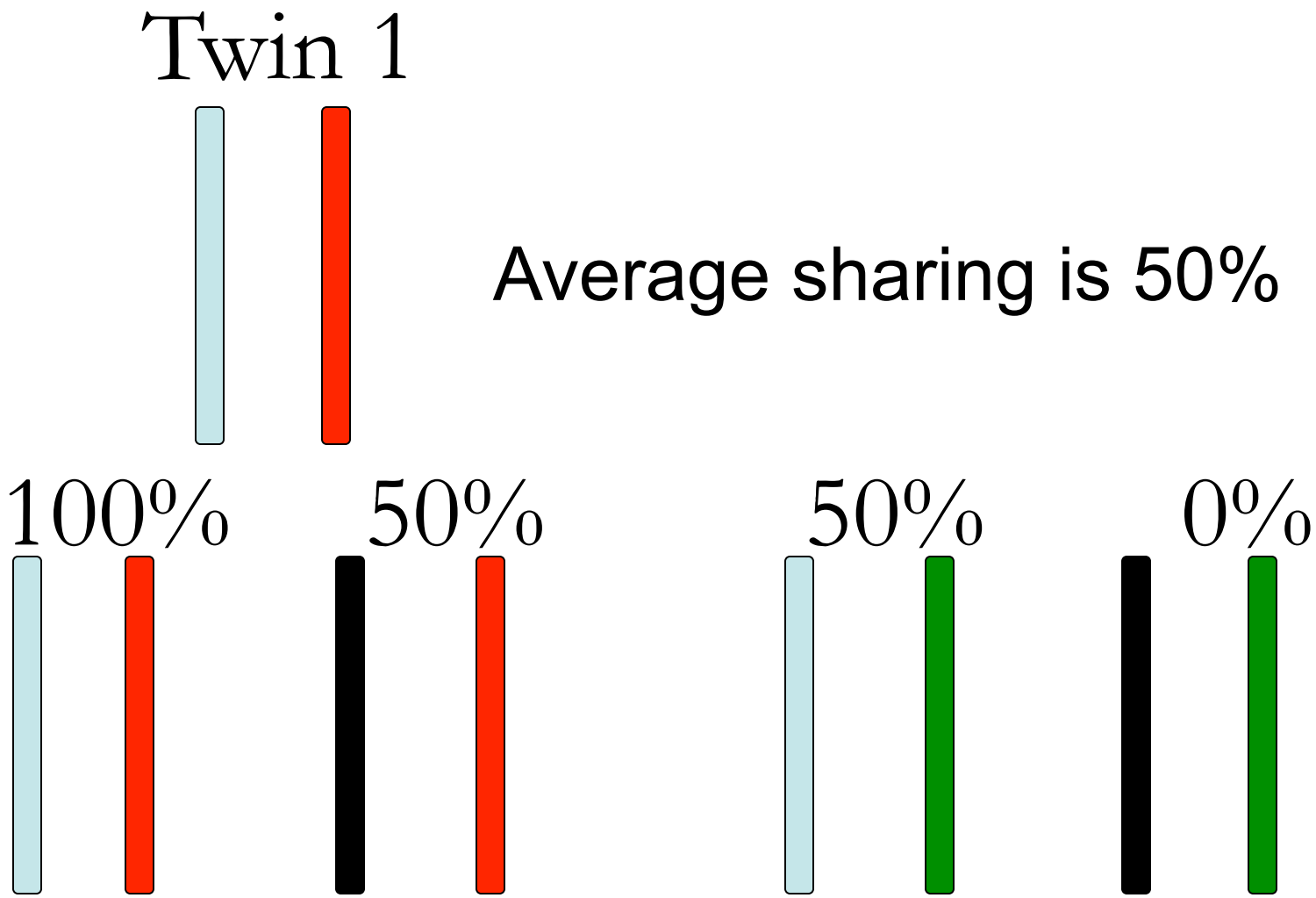
Then twin 2 must be:



DZ Twins (Fraternal)



DZ Twins (Fraternal)



IBD Sharing

	# of alleles shared IBD			
	2	1	0	
	Pr(2)	Pr(1)	Pr(0)	Prop IBD
Relationship				
Self, MZ twins	1	0	0	1
Parent, Offspring	0	1	0	1/2
Full siblings	1/4	1/2	1/4	1/2
Gr-child, Gr-prt	0	1/4	3/4	1/4
First cousins	0	1/4	3/4	1/8

Proportion of alleles shared IBD

= # alleles x Pr(# alleles) / 2

Twin Studies

ACE Model: $\sigma^2_P = \sigma^2_A + \sigma^2_C + \sigma^2_E$

A: Additive genetics

C: Common Environment

E: Unique Environment

Correlation in phenotype (P_1, P_2) among twins:

- $Corr_{mz}(P_1, P_2) = r_{mz} = A + C$ [100% genes + Env]
- $Corr_{dz}(P_1, P_2) = r_{dz} = ½A + C$ [50% genes + Env]

Heritability: $h^2 = 2(r_{mz} - r_{dz})$

Example of Twin Study: PCa

- Twin registry (Sweden, Denmark, and Finland)

Twin	Concordant pairs (cp)	Discordant pairs (dp)	Concordance 2cp / (2cp+dp)
MZ	197	807	0.33 (r_{mz})
DZ	148	1719	0.15 (r_{dz})

heritability = $2(r_{mz} - r_{dz}) = 2(0.33-0.15)$
$= 0.36$

Mucci et al. JAMA 2016;315:68
(actual paper estimate uses Structural Equation Modeling, so differs)

Heritability Extras

- Can also model interactions (see in text)
  - ACE model will overestimate the heritability in the presence of interactions
- Main point is that the ACE model is based on some assumptions that might not be true, and bias our heritability estimates.

Heritability from GWAS

- Extensive recent work using linear mixed models to estimate heritability from GWAS array data.
- Called ‘chip heritability.’

Outline

Several methods to answer the question: Is the trait genetic? [Austin Ch. 2]

1. Familial Aggregation and recurrence risks
2. Heritability

Genetic concepts [Austin Ch. 3]

3. Allele Frequency Estimation
4. Hardy-Weinberg equilibrium (HWE)
5. Population Substructure/Stratification

3. Allele Frequency Estimation

- Diploid, autosomal locus with 2 alleles: **A** and **a**
- Allele frequency is the fraction:

$$\frac{\text{No. of particular allele}}{\text{No. of all alleles in population}}$$

Allele (Gamete) Frequency

- Let p = Freq(A) frequency of the dominant allele
- Let q = Freq(a) frequency of the recessive allele

Then, $p + q = 1$

Genotype Frequency

- p^2 = frequency of homozygous dominant genotype
- q^2 = frequency of homozygous recessive genotype
- 2pq = frequency of heterozygous genotype

Then, $p^2 + 2pq + q^2 = 1$

Estimating Allele Frequencies from Genotype Frequencies

Genotypes:	AA	Aa	aa
Frequency:	p^2	$2pq$	q^2

Frequency of **A** allele = $p^2 + ½ (2pq)$

Frequency of **a** allele = $q^2 + ½ (2pq)$

Ex. Calculation: Allele Frequencies

Assume N=200 in each of two populations

- Pop 1: 90 AA 40 Aa 70 aa (N=200)
- Pop 2: 45 AA 130Aa 25 aa (N=200)

In Pop 1:

- p = 90/200 + ½ (40/200) = 0.45 + 0.10 = 0.55
- q = 70/200 + ½ (40/200) = 0.35 + 0.10 = 0.45

In Pop 2:

- p = 45/200 + ½ (130/200) = 0.225 + 0.325 = 0.55
- q = 25/200 + ½ (130/200) = 0.125 + 0.325 = 0.45

Take home points

- $p + q = 1$ (sum of the allele frequencies = 1)
- $p^2 + 2pq + q^2 = 1$ (sum of the genotype frequencies = 1)
- Two populations with markedly different genotype frequencies can have the same allele frequencies

Outline

Several methods to answer the question: Is the trait genetic? [Austin Ch. 2]

1. Familial Aggregation and recurrence risks
2. Heritability

Genetic concepts [Austin Ch. 3]

3. Allele Frequency Estimation
4. Hardy-Weinberg equilibrium (HWE)
5. Population Substructure/Stratification

4. Hardy-Weinberg Equilibrium

Hardy–Weinberg principle:
Both allele and genotype frequencies in a population remain constant (i.e., in equilibrium) from generation to generation unless specific disturbing influences are introduced.

$$p^2 + 2pq + q^2 = 1$$

Hardy-Weinberg Assumptions

Allele frequencies do not vary IF:

- Large population
- Random mating
- No in or out migration
- No isolated groups within the population
- No mutation
- No selection (no allele is advantageous)

Example: Test of Hardy-Weinberg Equilibrium (HWE)

Assume we have the following observed genotype frequencies: 100 GG, 30 AG, 20AA

Frequency of the G allele:

$$p = 100/150 + 0.5(30/150) = 0.767$$

Frequency of the A allele

$$q = 20/150 + 0.5(30/150) = 0.233 \quad = 1-p$$

Example

Calculate expected genotype frequencies based on HW: $p^2 + 2pq + q^2 = 1$

GG

p^2	= 0.767*0.767	= 0.588
$n(p^2)$	= 150(0.588)	= 88.2

AG

$2pq$	= 2*0.767*0.233	= 0.357
$n(2pq)$	= 150(0.357)	= 53.6

AA

q^2	= 0.233*0.233	= 0.054
$n(q^2)$	= 150(0.054)	= 8.1

Example

Compare expected genotype counts (E) to observed genotype counts (O)

	E	O	(O-E)^2/E
GG	88.2	100	1.58
AG	53.6	30	10.39
AA	8.1	20	17.48

			29.45

Chi-square test

$= \Sigma_i (O_i - E_i)^2/E_i = 29.4$ (chi square distribution with 1 degree of freedom)

$p = 6.6 \times 10^{-8}$ > Out of H-W

HWE can be easily expanded to account for any number of alleles at a locus

- 3 allele case (p_1, p_2, p_3)

  Allele frequencies: $p_1 + p_2 + p_3 = 1$

  Genotype frequencies:

  $p_1^2 + p_2^2 + p_3^2 + 2p_1p_2 + 2p_1p_3 + 2p_2p_3 = 1$

- 4 allele case (p_1, p_2, p_3, p_4)

  Allele frequencies: $p_1 + p_2 + p_3 + p_4 = 1$

  Genotype frequencies:

  $p_1^2 + p_2^2 + p_3^2 + p_4^2 + 2p_1p_2 + 2p_1p_3 + 2p_2p_3 + 2p_3p_4 = 1$

Application of HWE

For genetic association studies:

- Used as QC measure to assess the accuracy of the genotyping method
- Expect SNPs to be in HWE among control populations (ethnic-specific)
- Violations of HWE could indicate genotyping errors or bias in data

HWE Game

1. Everyone receives ~5 pairs of cards
2. Two allele model: Red (R allele) & Black (B allele)
3. Random Mating: Exchange one card from each pair with another person (keep cards face down)
4. Determine genotype frequency: RR, RB, BB
5. Determine allele frequency: R, B

Outline

Several methods to answer the question: Is the trait genetic? [Austin Ch. 2]

1. Familial Aggregation and recurrence risks
2. Heritability

Genetic concepts [Austin Ch. 3]

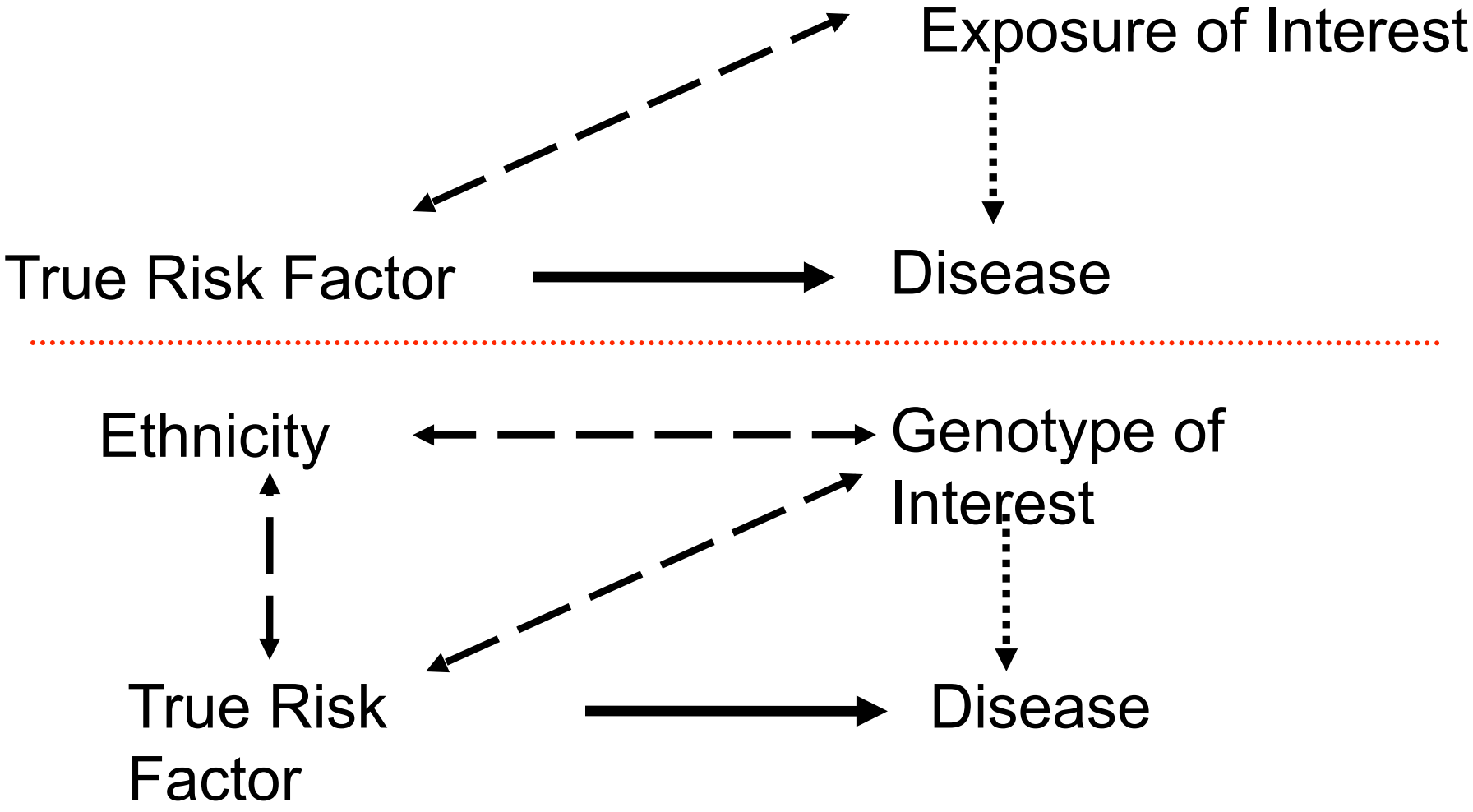
3. Allele Frequency Estimation
4. Hardy-Weinberg equilibrium (HWE)
5. Population Substructure/Stratification

5. Population Substructure/ Stratification

Population stratification:

Confounding due to gene having marked variation in allele frequency across subgroups of a population and these subgroups differ in their baseline risk of disease.

Population Stratification: Confounding



Wacholder, *JNCI*, 2000

Example

Study Population: 4,290 Pima and Papago Native Americans

Genetic Variant: Gm 3;5,13, 15 haplotype (Gm system of human immunoglobulin G)

Outcome: Type 2 diabetes

Question: Is the Gm 3; 5,13, 15 haplotype associated with Type 2 diabetes?

Knowler, *AJHG*, 1998

Population Stratification: Gm3;5,13,14 in admixed sample of Native Americans of the Pima and Papago tribes

Full heritage American Indian population

	+	-
Gm3;5,13,14	~1%	~99%

NIDDM prevalence ~40%

Caucasian population

	+	-
Gm3;5,13,14	~66%	~34%

NIDDM prevalence ~15%

Gm3,5,13,14 haplotype	Cases	Controls
+	7.80%	29.00%
-	92.20%	71.00%

Unadjusted for ethnic background

OR = 0.27 (95% 0.18-0.40)

Different genotype frequency, different phenotype frequency

Population Stratification: Gm3;5,13,14 in admixed sample of Native Americans of the Pima and Papago tribes

Index of Indian heritage	Gm3;5,13,14 haplotype	% Diabetes
0	65.8%	18.5%
4	42.1%	28.5%
8	1.6%	39.2%

Gm3,5,13,14 haplotype	Cases	Controls
+	7.80%	29.00%
-	92.20%	71.00%

Adjusted for ethnic background

OR = 0.83 (95% 0.58-1.18)

Previous result just picked out race/ ethnicity!

Ancestry Informative Markers

- Polymorphisms with known allele frequency differences across ancestral groups
- Useful in estimating ancestry in admixed individuals
- Example: Duffy locus (codes for blood group)

  - 100% sub-Saharan Africans vs. other groups

  - Protects against *P. vivax* (malaria)

Example AIM: Duffy locus (rs2814778)

	Sample Ascertainment				Genotype Detail					Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G	G/G	HWP	A	G
ss105436776	P1		204	GF	0.706	0.118		0.176	0.001	0.765	0.235
	CAUC1		62	GF	1.000					1.000	
	AFR1		48	GF		0.292		0.708	0.752	0.146	0.854
	HISP1		46	GF	0.783	0.174		0.043	0.273	0.870	0.130
	PAC1		48	GF	0.958	0.042			1.000	0.979	0.021
	HAPMAP-ASW		98	IG	0.041	0.327		0.633	1.000	0.204	0.796
	HAPMAP-LWK		180	IG		0.011		0.989	1.000	0.006	0.994
	HAPMAP-MEX		100	IG	0.920	0.080			1.000	0.960	0.040
	HAPMAP-MKK		286	IG	0.007	0.126		0.867	0.752	0.070	0.930
	HAPMAP-TSI		176	IG	0.977	0.023			1.000	0.989	0.011
	P3		532	GF	0.737	0.019		0.244	0.001	0.746	0.254
	CAUC3		122	GF	0.984	0.016			1.000	0.992	0.008
	AFR3		148	GF	0.095	0.027		0.878	0.001	0.108	0.892
	HISP3		92	GF	0.957	0.043			1.000	0.978	0.022
	PAC3		170	GF	1.000					1.000	
	ENSEMBL_Watson		2	IG	1.000					1.000	
	ENSEMBL_Venter		2	IG	1.000					1.000	
ss119047871	YRI		2	IG				1.000			1.000
ss164197181	YRI	Sub-Saharan African	2	IG				1.000			1.000

http://www.ncbi.nlm.nih.gov/projects/SNP