GWAS.

9/11

* About human genome → 3 billion base pairs

* De oxyribonucleic Acid

* Single Nucleotide Polymorphisms (SNPs)

→ A DNA variation when a single A, T, C, G differ between species or paired chromosomes

* Basic Definitions

I Locus → A specific location in genome, Could be a single DNA base to gene to bend

II Variant → A genetic difference relative to human reference genome.

III Polymorphism → A variant common in population (1% Allel, not individual-specific

IV Allele → A genetic variant on a specific locus

V Genotype → The combination of allele in a individu

A SNP with two allele, A, a.

P = Percentage of A

q = Percentage of a

$p + q = 1$. → Variant

↓

$p^2 + 2pq + q^2 = 1$. → Genotype

↓

Could used to estimate the genotype divived disease incident rate

* Hardy-Weinberg Equilibrium

An ideal population, the observed genotype frequency is identical to expected frequency.

No mutation, no assortative mating, no migration no natrual selection

→ Most population stay in HWE, HWE ~ χ^2

	AA	Aa	aa	Total
observation	N_{AA}	N_{Aa}	N_{aa}	N
Expected	Np^2	$N \times pq$	Nq^2	N

$$\sum_{g=AA,Aa,aa} \left(\frac{(O_g - E_g)^2}{E_g} \right) \sim \chi^2_1$$

Linkage Disequilibrium

Loci are associated with each other in Non-random level.

Ex

Haplotype

$A_1 B_1$	x_{11}
$A_1 B_2$	x_{12}
$A_2 B_1$	x_{21}
$A_2 B_2$	x_{22}

P_1 = Prop of A_1

P_2 A_2

q_1 B_1

q_2 B_2

$\downarrow$ $P_1 + P_2 = q_1 + q_2 = 1$

$x_{11} + x_{12} = P_1$

$x_{11} + x_{21} = q_1$

$\vdots$

Define $D = x_{11} - P_1 \cdot q_1 = P(AB) - P(A) \cdot P(B)$

$\downarrow$

the deviation of expected and observed frequency

$\downarrow$

Serve as Covariance

→ Next

The formula of correlation between two variables

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} \rightarrow D$$

→ In Binary data

Variance is $P \cdot (1-P)$

↓

$$r = \frac{D}{\sqrt{P_1 \cdot P_2 \cdot q_1 \cdot q_2}}$$

The correlation between Pairs of Loci

To test LD → by Chi-square

Halotype	Observed	Expected
$A_1 B_1$	N_{11}	$N \cdot P_1 \cdot q_1$
$A_1 B_2$	N_{12}	$N \cdot P_1 \cdot q_2$
$A_2 B_1$	N_{21}	$N \cdot P_2 \cdot q_1$
$A_2 B_2$	N_{22}	$N \cdot P_2 \cdot q_2$

$$\chi^2 = \sum \frac{(O-E)^2}{e}$$

The meaning of the "D"

① $D = P(AB) - P(A) \cdot P(B) = 0$

$P(AB) = P(A) \cdot P(B)$ → The loci do not affect each other

② $D = P(AB) - P(A) \cdot P(B) > 0$

$P(AB) > P(A) \cdot P(B)$ → The loci influence each other in positive association

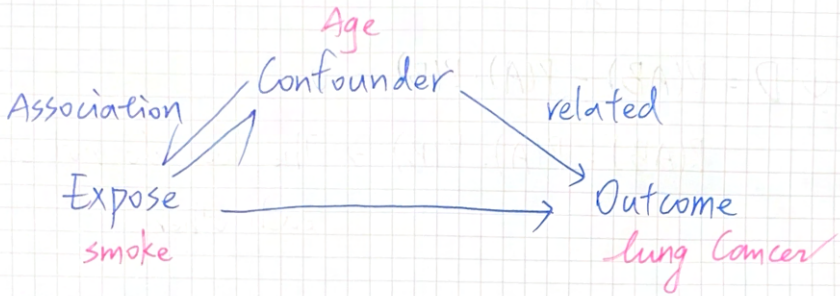
Tend to be together

③ $D = P(AB) - P(A) \cdot P(B) < 0$

$P(AB) < P(A) \cdot P(B)$ → The loci influence each other in negative association

Tend to be Separative

Confounder



GWAS ?

A Study associated with genotype and phenotype (trait)

Type I Error → False positive

① Avoid Post hoc subgroup analysis → multiple test

② Statistical adjustment

I Bonferroni adjustment

II Perturbation test

Type I Error $= 1-(1-\alpha)^M = \alpha M$

Why M is huge, more type I Error

Type II Error False negative findings

Genotype and Environment

		Case	Control
G+	E+	a	b
G+	E-	c	d
G-	E+	e	f
G-	E-	g	h

$$OR_{ge} = \frac{a \cdot h}{b \cdot g}$$

$$OR_g = \frac{c \cdot h}{d \cdot g}$$

$$OR_e = \frac{e \cdot h}{f \cdot g}$$

Addictive interaction

$$OR_{ge} = OR_g + OR_e - 1$$

↳ minus baseline

Multiplication interaction

$$OR_{ge} = OR_g \cdot OR_e$$

Ex: Hyposize Case is a Genotype and environment interaction disease

For Case

	G+	G-
E+	a	b
E-	c	d

$$OR = \frac{a \cdot g}{b \cdot c} \neq 1$$

For Control

	G+	G-
E+	b	f
E-	g	h

$$OR = \frac{b \cdot h}{g \cdot f} = 1$$