Genotype Coding

DNA Sequences at a SNP Position (T is risk allele)

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5
CC	CT	TT	CC	CT

Genotype Coding Models

Coding Model	Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Mean
Additive (0,1,2) Count risk alleles	0	1	2	0	1	0.8
Dominant (0,1,1) 1 if any risk allele	0	1	1	0	1	0.6
Recessive (0,0,1) 1 if both alleles are risk alleles	0	0	1	0	0	0.2

The additive model is particularly valuable as genetic effects often scale with the number of risk alleles.

Standardized Genotype Vector $X = \frac{X_{raw} - mean}{sd}$

Model	Individual 1	Individual 2	Individual 3	Individual 4	Individual 5	Mean	Sd
Additive	-0.96	0.24	1.43	-0.96	0.24	0.8	0.84

Minor Allele Frequency (MAF) Calculation

Step 1: Raw Genotype Data

Individual 1	Individual 2	Individual 3	Individual 4	Individual 5
CC	CT	TT	CC	CT

C (Cytosine) T (Thymine)

Step 2: Count Alleles

All alleles extracted from genotypes from all individuals:

CC CT TT CC CT

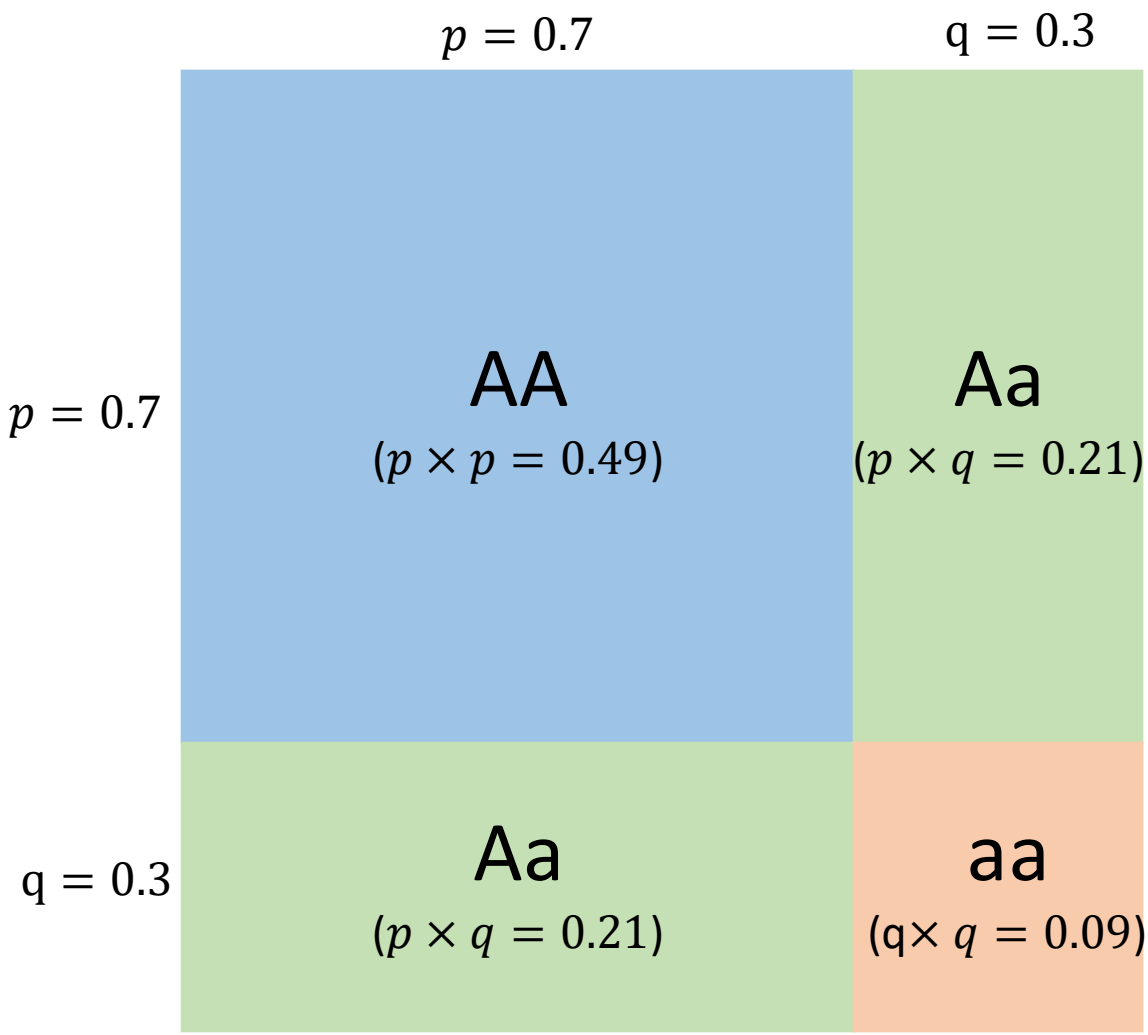
Count: C: 7 (70%) T: 3 (30%)

Step 3: Calculate MAF

Major allele: C (70%)
Minor allele: T (30%)

Minor Allele Frequency (MAF) = 0.3

Hardy-Weinberg Equilibrium



Example:
$p = 0.7$ (A allele)
q = 0.3 (a allele)

Legend:

- AA (Homozygous dominant)
- Aa (Heterozygous)
- AA (Homozygous recessive)

$$\boldsymbol{p^2 + 2pq + q^2 = 1}$$

$$0.49 + 0.42 + 0.09 = 1$$

Linkage Disequilibrium (LD)

Non-random association between alleles at different genetic loci,
where certain combinations occur more or less frequently than expected by chance

Raw Genotype Matrix

	Variant 1	Variant 2	Variant 3
Individual 1	CC	CT	AT
Individual 2	TT	TT	AA
Individual 3	CT	CT	AA
Individual 4	CC	TT	AA
Individual 5	CC	CC	TT

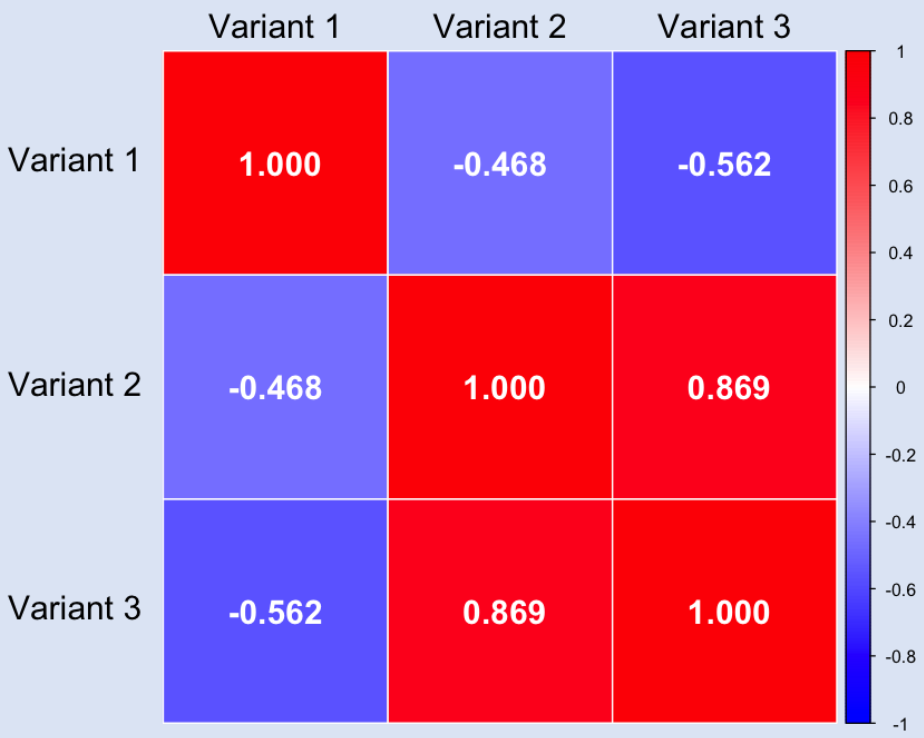
Standardize

Standardized Genotyped Matrix

	Variant 1	Variant 2	Variant 3
Individual 1	-0.671	0.239	0.447
Individual 2	1.565	-0.956	-0.671
Individual 3	0.447	0.239	-0.671
Individual 4	-0.671	-0.956	-0.671
Individual 5	-0.671	1.434	1.565

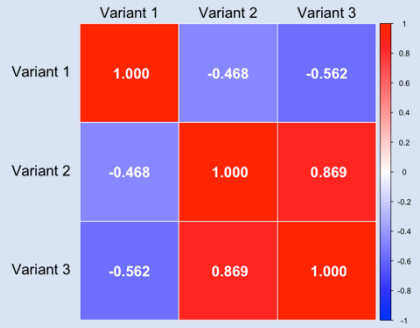
$$LD = R = \frac{X^T X}{N}$$

LD matrix (correlation)

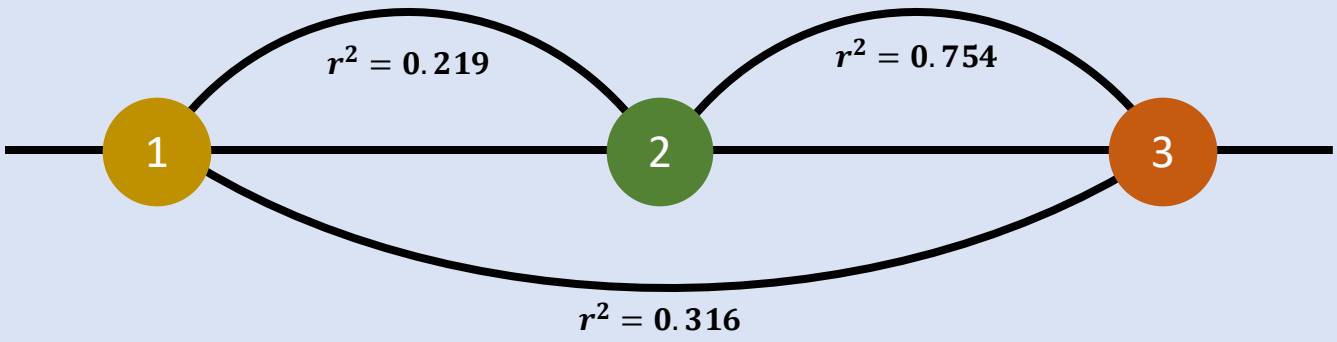


Linkage Disequilibrium (LD) Scores

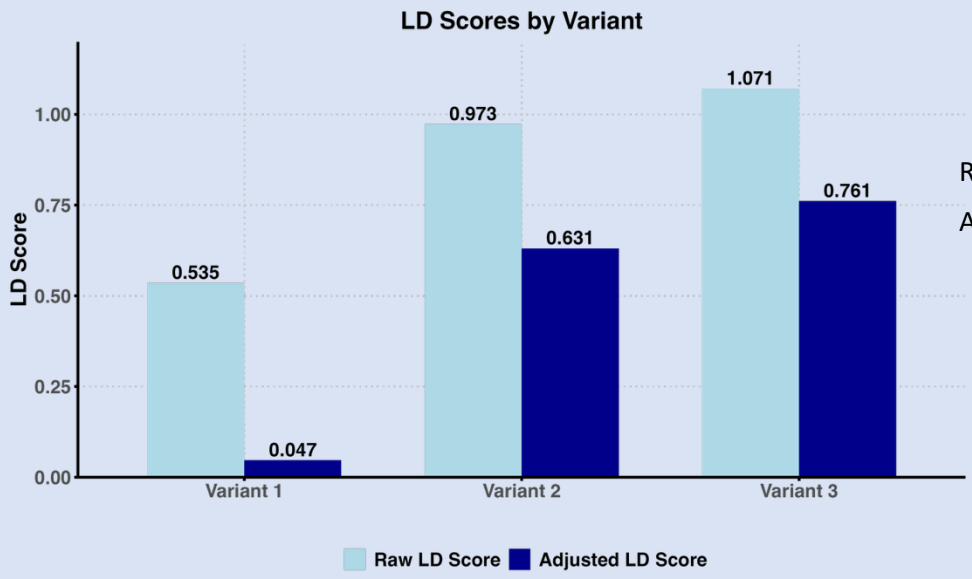
LD matrix (correlation)



Squared Correlations (r^2)



LD Scores



LD Scores Formula

Raw LD score: $l_j = \sum_{k \neq j} r_{kj}^2$

Adjusted LD score: $r_{adj}^2 = r^2 - \frac{1-r^2}{N-2}$

Results

Variant 3 shows the highest connectivity.

Genetic Relationship Matrix (GRM)

Quantifying Genetic Similarity Between Individuals

Population with 5 individuals

Standardized Genotyped Matrix X

	Variant 1	Variant 2	Variant 3
Individual 1	-0.671	0.239	0.447
Individual 2	1.565	-0.956	-0.671
Individual 3	0.447	0.239	-0.671
Individual 4	-0.671	-0.956	-0.671
Individual 5	-0.671	1.434	1.565

Genetic Relationship Matrix (G)

$$G = \frac{XX^T}{M}$$

	Individual 1	Individual 2	Individual 3	Individual 4	Individual 5
Individual 1	0.236	-0.526	-0.181	-0.026	0.498
Individual 2	-0.526	1.271	0.307	0.105	-1.157
Individual 3	-0.181	0.307	0.236	-0.026	-0.336
Individual 4	-0.026	0.105	-0.026	0.605	-0.657
Individual 5	0.498	-1.157	-0.336	-0.657	1.652

Single Marker Linear Regression and OLS

Single Variant Regression

$$Y = X_j \times \beta + \varepsilon$$

Y	X_j	β	ε
$N \times 1$ Trait	$N \times 1$ Genotype	scalar Genetic Effect	$N \times 1$ Error

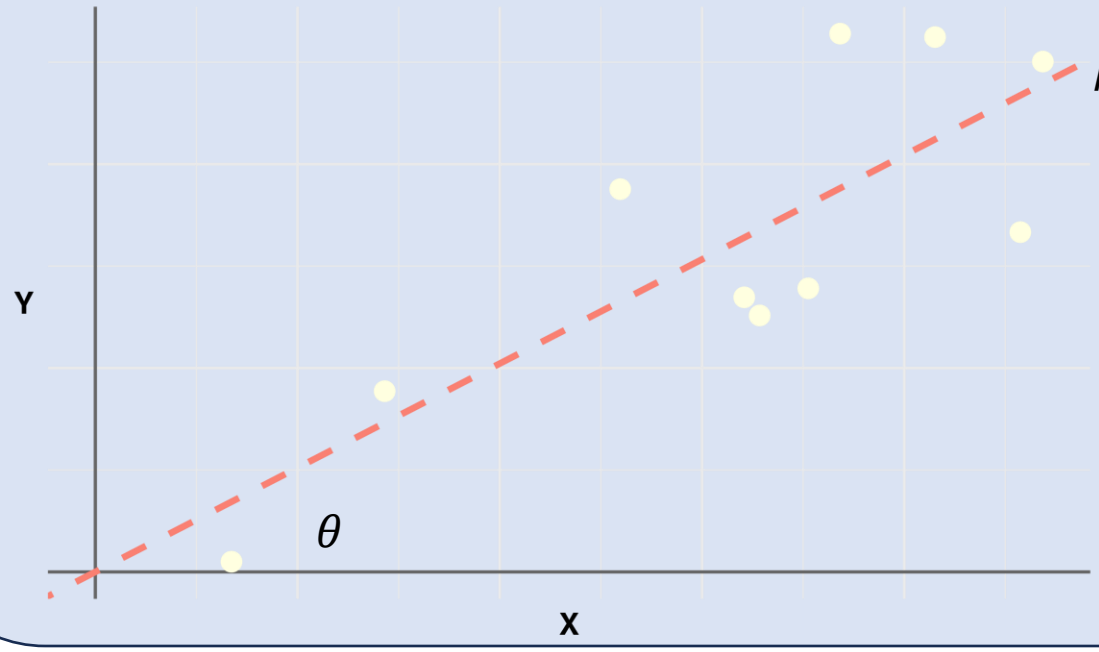
Ordinary Least Squares (OLS)

$$\hat{\beta}_{OLS} = (X_j^T X_j)^{-1} X_j^T Y$$

$$\hat{\beta} = (X_j^T X_j)^{-1} \times X_j^T \times Y$$

$\hat{\beta}$	$(X_j^T X_j)^{-1}$	X_j^T	Y
scalar Genetic Effect	Scalar (scale for X)	$N \times 1$ Transpose of X_j	$N \times 1$ Trait

Ordinary Least Squares



$$\hat{\beta}_{OLS} = \tan\theta$$

OLS finds $\hat{\beta}$ that minimizes the sum of the squared residuals

Odds, Odds Ratio and Logistic Regression

Odds

Definition:

$$Odds = \frac{p}{1-p}$$

Example:

If disease risk = 20%, then

Odds = 0.2/0.8=0.25 (1:4)

Odds Ratio

Definition:

$$OR = \frac{Odds_1}{Odds_2}$$

Interpretation:

- OR = 1: No association
- OR > 1: Increased risk
- OR < 1: Decreased Risk

Logistic Regression

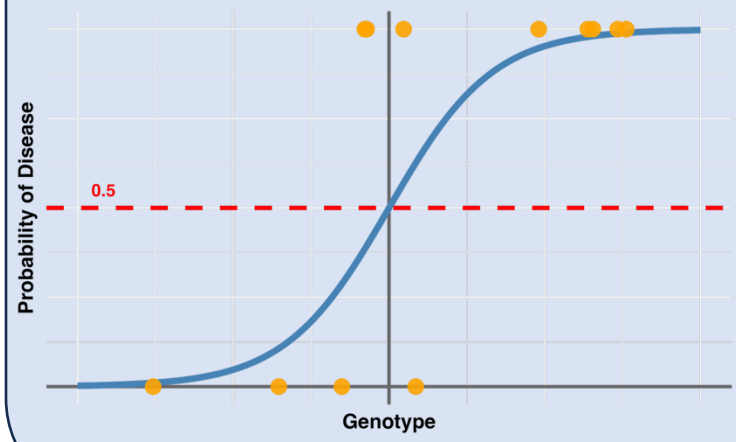
Model:

$$\text{logit}(\text{p}(X)) = \ln\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$$

Relationship to OR:

$$OR = e^{\beta_1}$$

Logistic Function

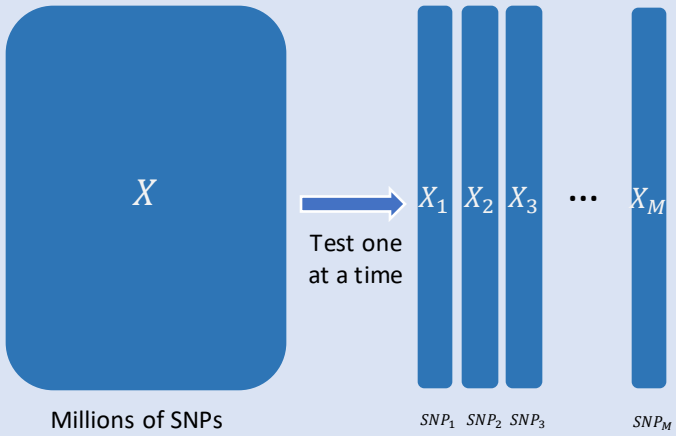


Application in Statistical Genetics

- Case-control studies: control (0) and case (1)
- GWAS: often identify the genetic variants associated with disease (0 or 1) using logistic regression

Summary Statistics

Individual-level Data



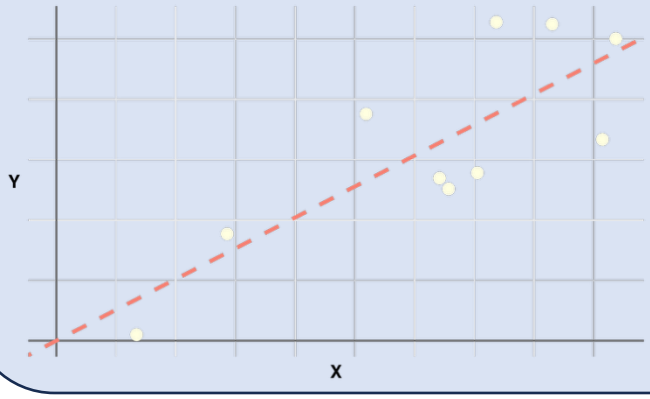
Large file size (GB - TB)

Summary Statistics

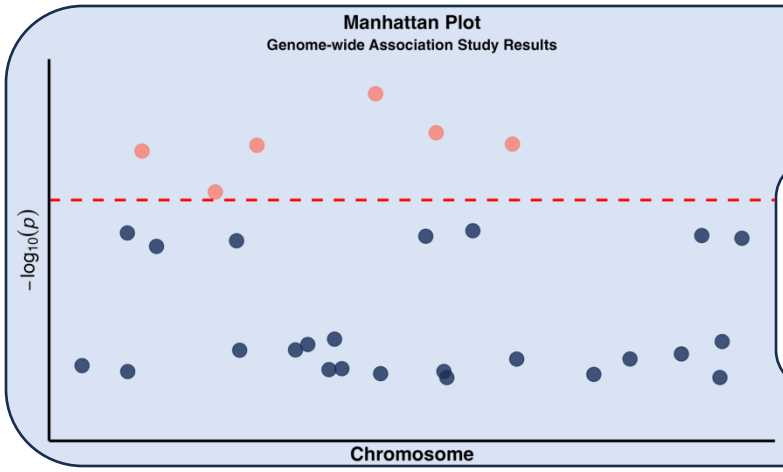
SNP	BETA	SE	Z	P	MAF
rs1	0.85	0.32	2.64	0.02	0.40
rs2	-0.31	0.28	-1.10	0.27	0.30
rs3	0.42	0.30	1.40	0.16	0.25

Small file size (MB)

BETA, Z and P



- **Beta** in the summary statistics can be $\hat{\beta}_{OLS}$ or $\hat{\beta}$ using any other method.
- **BETA>0**: the (risk of) trait increases if one carries the risk allele, and vice versa.
- $Z = \frac{\text{BETA}}{SE}$ and shares the same direction as BETA
- P suggests if the association is significant.

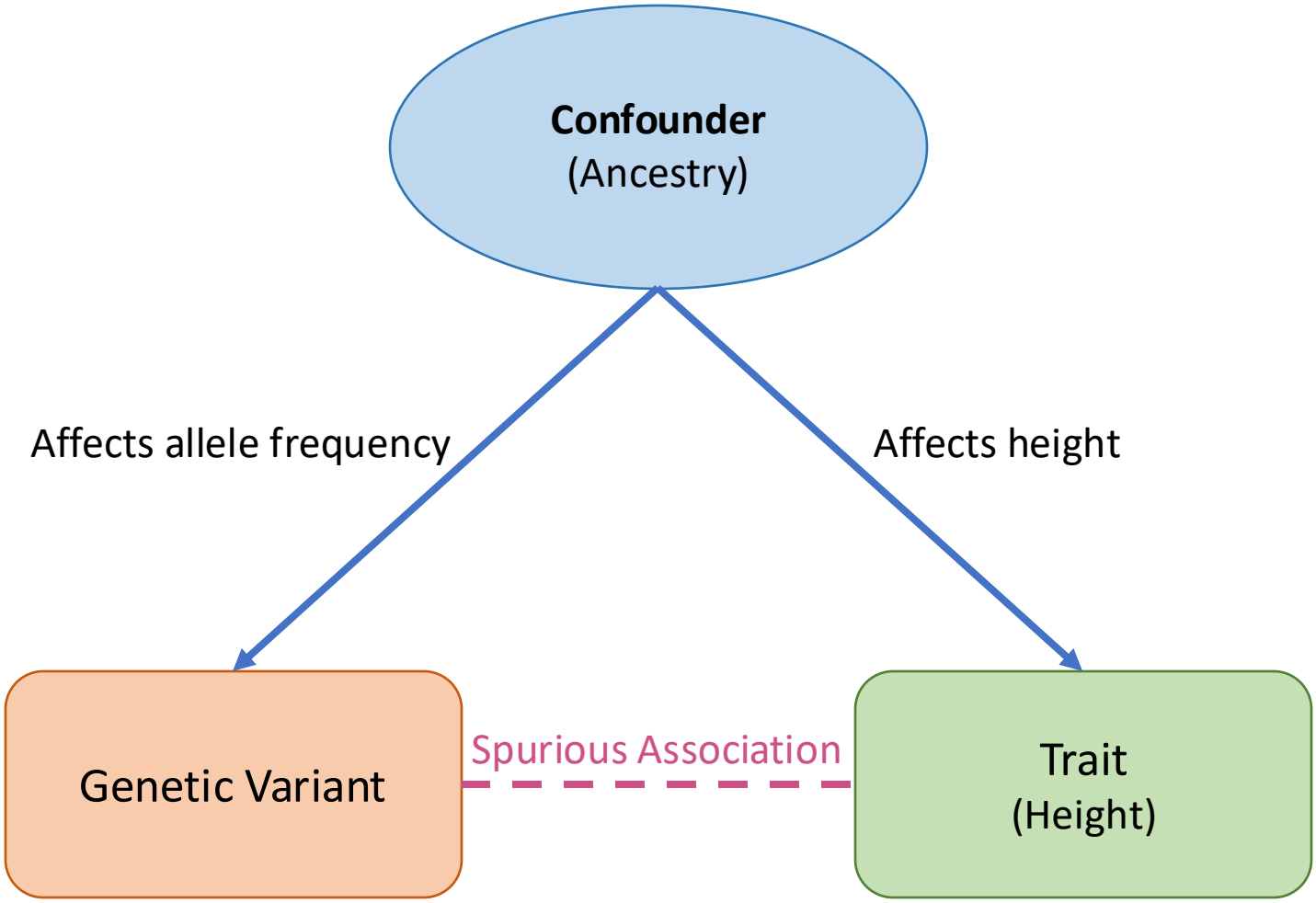


Compact Storage
MB vs. GB-TB

Privacy Protection
Fewer Constraints

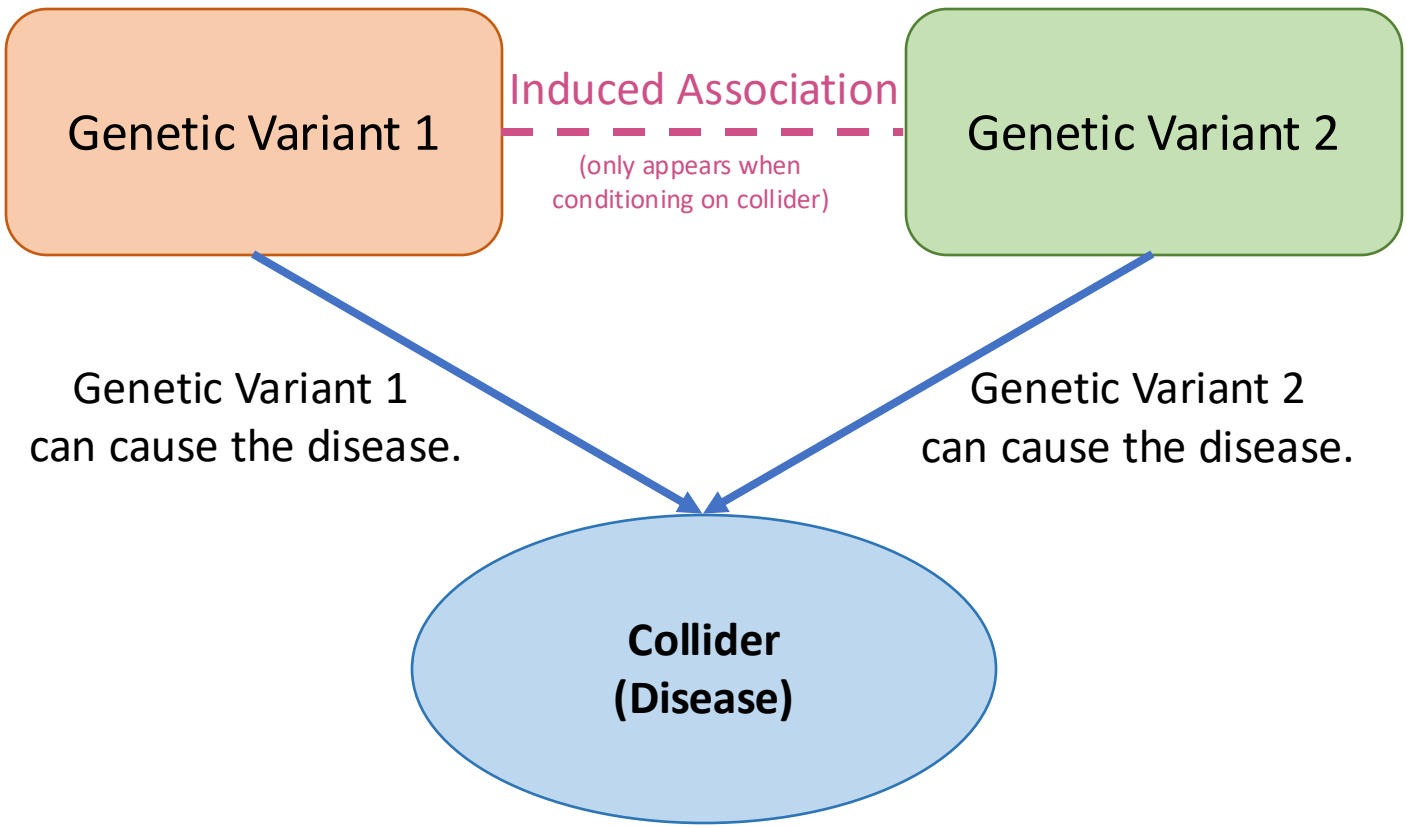
Association Information
P below 5×10^{-8}

Covariates --- Confounder



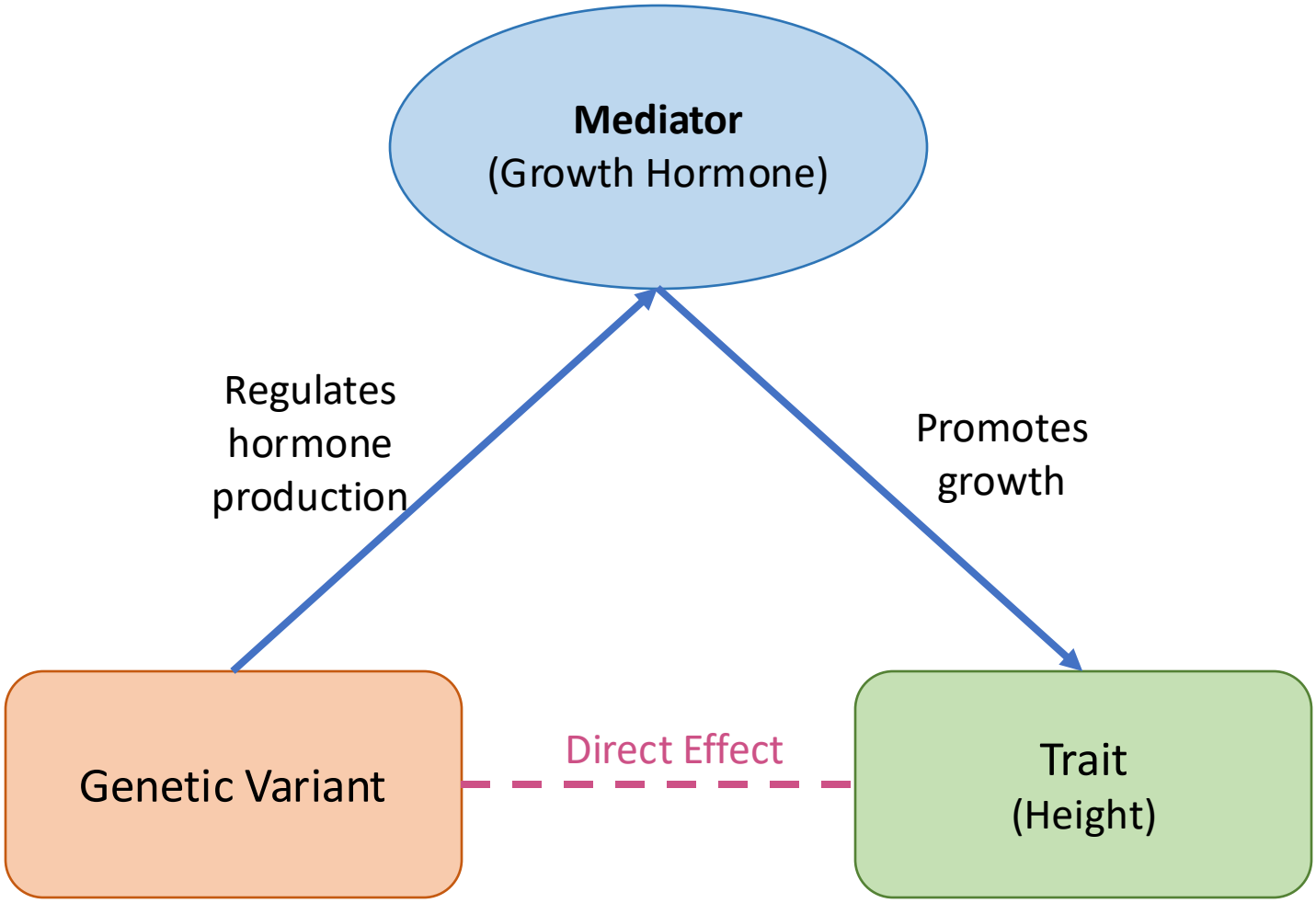
When ancestry is not controlled for in the analysis,
it might create a misleading association between genotype and trait

Covariates --- Collider



When we look at the cases (conditional on collider),
a spurious association may appear between the two genetic variants.

Covariates --- Mediator

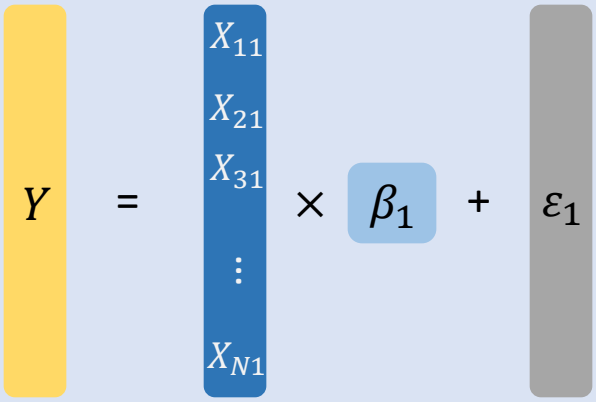


- Growth hormone mediates the relation between genetic variants and height, representing the causal pathway between them.
- When the mediator is controlled, one may fail to detect the association between genetic variant and the trait.

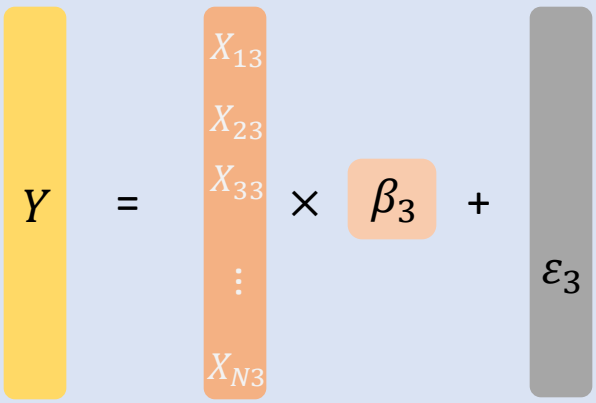
Marginal vs. Joint Effects

Marginal Effects

(Three separate models, $\boldsymbol{Y = X_j \beta_j + \varepsilon_j}$



$N \times 1$ Trait | $N \times 1$ Genotype | scalar Genetic Effect | $N \times 1$ Error

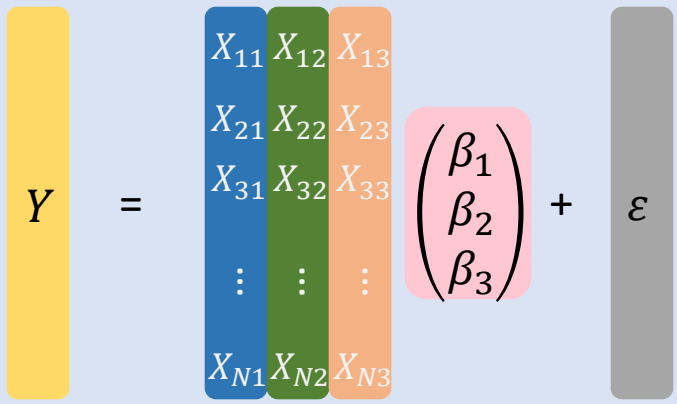


$N \times 1$ Trait | $N \times 1$ Genotype | scalar Genetic Effect | $N \times 1$ Error

Each variant is analyzed independently.

Joint Effects

(One model with all variants, $\boldsymbol{Y = X\beta + \varepsilon}$)



Accounts the correlations between variants (i.e., LD)

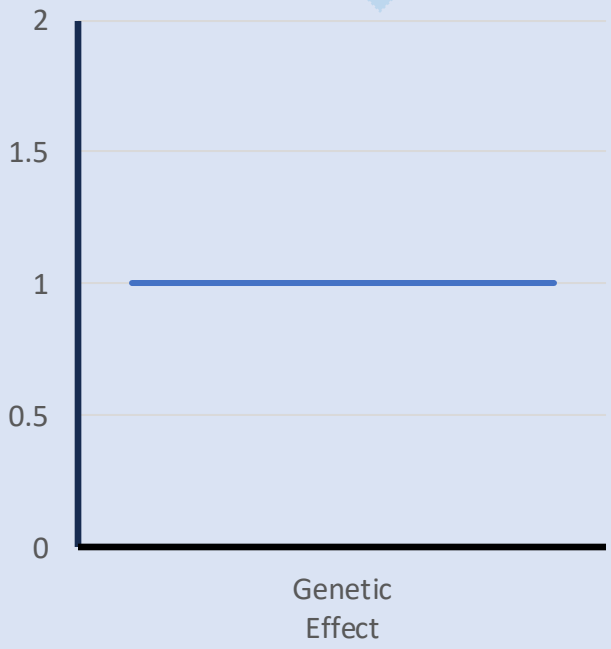
Random Effect

Fixed Effect

Consider only the first variant and the effect is fixed

$$Y = \begin{bmatrix} X_{11} \\ X_{21} \\ X_{31} \\ \vdots \\ X_{N1} \end{bmatrix} \times \beta_1 + \varepsilon_1$$

$N \times 1$ Trait | $N \times 1$ Genotype | $N \times 1$ Error



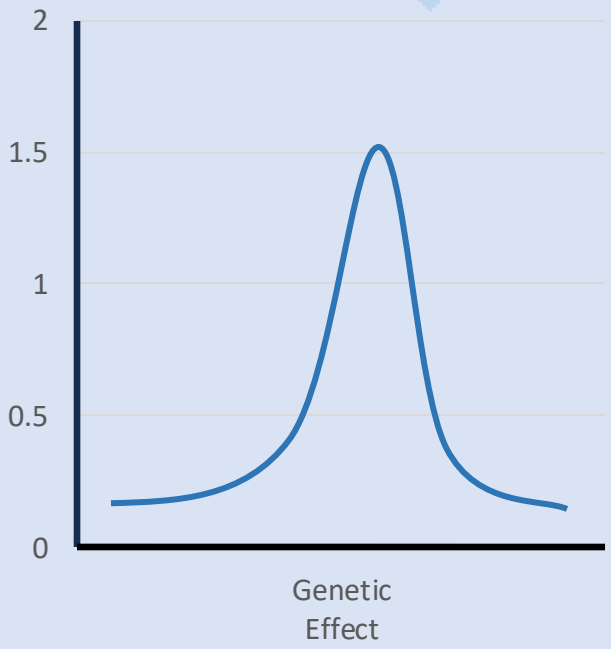
The true genetic effect is a fixed value.

Random Effect

Consider only the first variant and the effect is random

$$Y = \begin{bmatrix} X_{11} \\ X_{21} \\ X_{31} \\ \vdots \\ X_{N1} \end{bmatrix} \times \beta_1 + \varepsilon_1$$

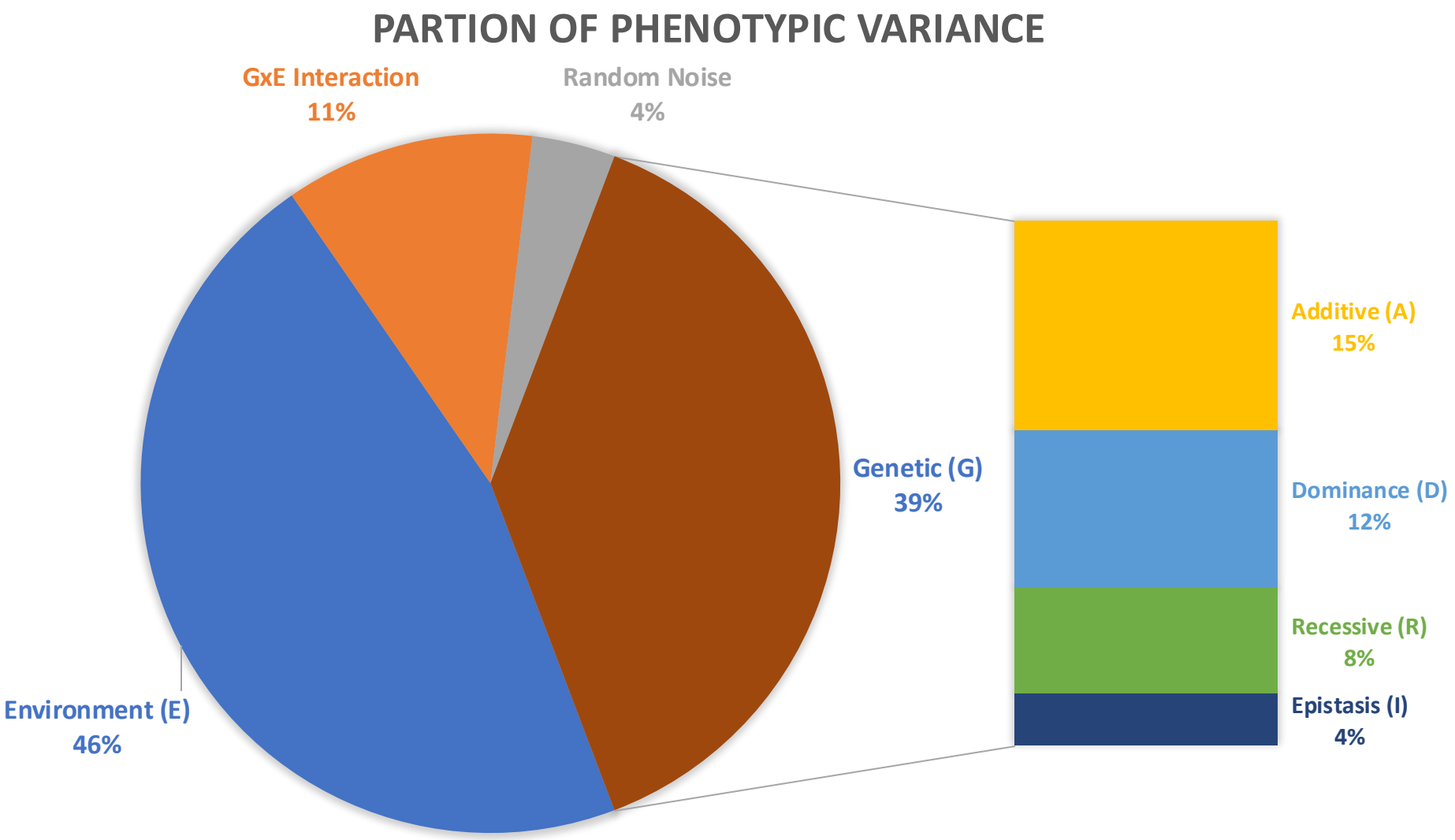
$N \times 1$ Trait | $N \times 1$ Genotype | $N \times 1$ Error



The true genetic effect comes from a distribution

Proportion of Variance Explain (PVE) and Heritability

measures how much of the total variation in a trait (like height or disease risk) can be attributed to specific variables in the statistical model (e.g., genetic variants).



PVE and Heritability

Phenotypic Variance:

$$Var(P) = Var(G)+Var(E)+2Cov(G,E)+Var(\varepsilon)$$

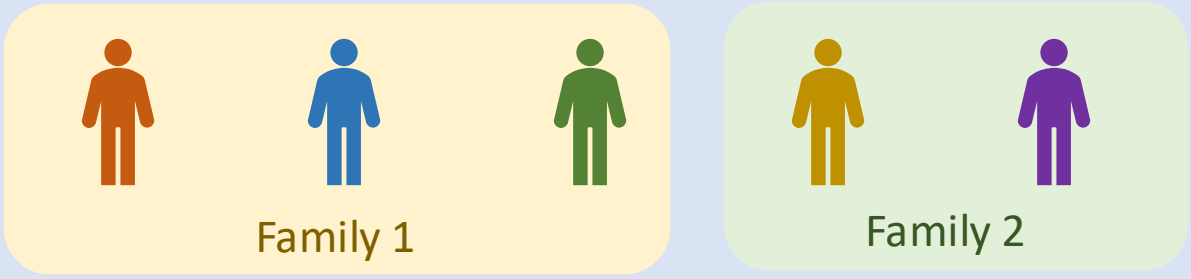
If we assume G and E are independent from each other:

- Genetic Variance (Broad-sense heritability H^2):= Var(A)+Var(D)+Var(R)+Var(I)
- Narrow-sense heritability h^2:= Var(A)

Linear Mixed Model

Accounts for correlation between samples due to shared genetic background

Population with 5 individuals

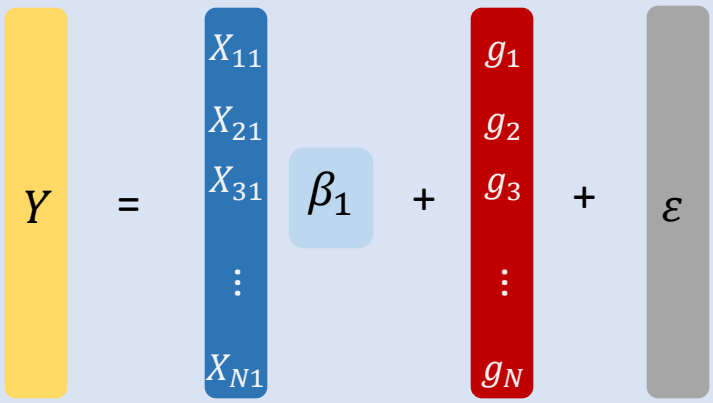


Linear Mixed Model

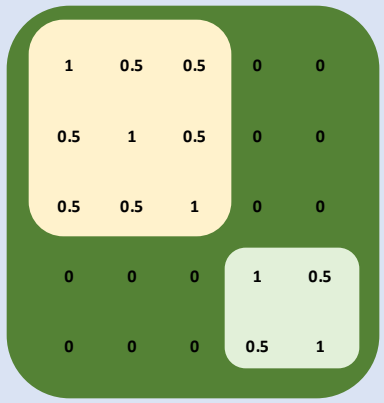
$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{g} + \boldsymbol{\varepsilon}$$

- β: fixed effect (genetic effect)
- $\boldsymbol{g}$: random effect (grouping of samples)
  - $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{u}$ where $\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_u^2 G})$, G is the GRM

Components Visualization



GRM



Captures the relatedness between samples

Fixed Effect Meta-Analysis

Synthesizes statistical evidence across multiple independent studies using weighted averaging techniques

Several Independent Cohorts

Study 1
N1=5000

Sumstats 1

β_1
SE_1
N_1
...

Study 2
N2=8000

Sumstats 2

β_2
SE_2
N_2
...

What we believe about β_1 and β_2 ?

Since this is a fixed effect model, the underlying true β is a fixed value, i.e., β_1 and β_2 should be fixed effect instead of a random effect.

Meta-Analysis

Merged set
N=13000

- $\beta = \frac{\sum_i w_i \beta_i}{\sum_i w_i}$
- w_i is the weight for study I, which can be:
  - Sample Size Weighting
    $$w_i = \frac{N_i}{\sum_i N_i}$$
  - Inverse Variance Weighting
    $$w_i = \frac{1}{SE^2}$$
- Equivalent to merging individuals from two studies together

Random Effect Meta-Analysis

Synthesizes statistical evidence across multiple independent studies assuming the true effect is a random variable

Several Independent Cohorts

Study 1
N1=5000

Sumstats 1

β_1
SE_1
N_1
...

Study 2
N2=8000

Sumstats 2

β_2
SE_2
N_2
...

What we believe about β_1 and β_2 ?

Since this is a random effect model, the underlying true β is NOT a fixed value, and comes from a distribution, i.e., $\beta \sim N(\beta_0, \sigma_0^2)$. And β_1 and β_2 both comes from this distribution.

Meta-Analysis