Lecture 1: Association Tests and Whole-Genome Regression

Instructors: Andrey Ziyatdinov and Timothy Thornton

The Unidad de Medicina Experimental, Faculty of Medicine,
National Autonomous University of Mexico

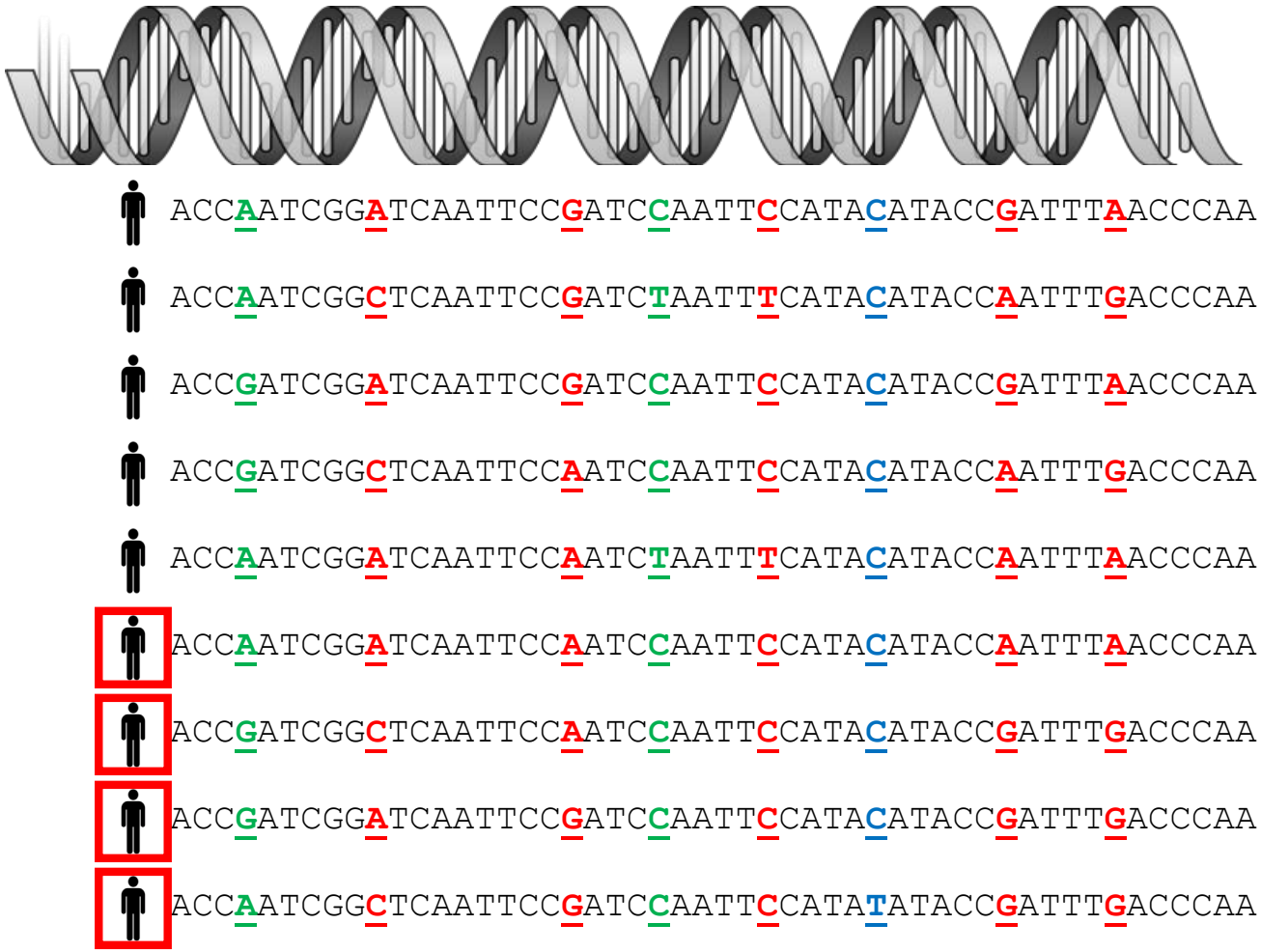
Lecture Overview

1. Quantitative Genetic Model
2. Association Tests for Quantitative Traits
3. Association Tests for Binary Traits
4. Whole Genome Regression (Regenie)
5. Overview of Plink

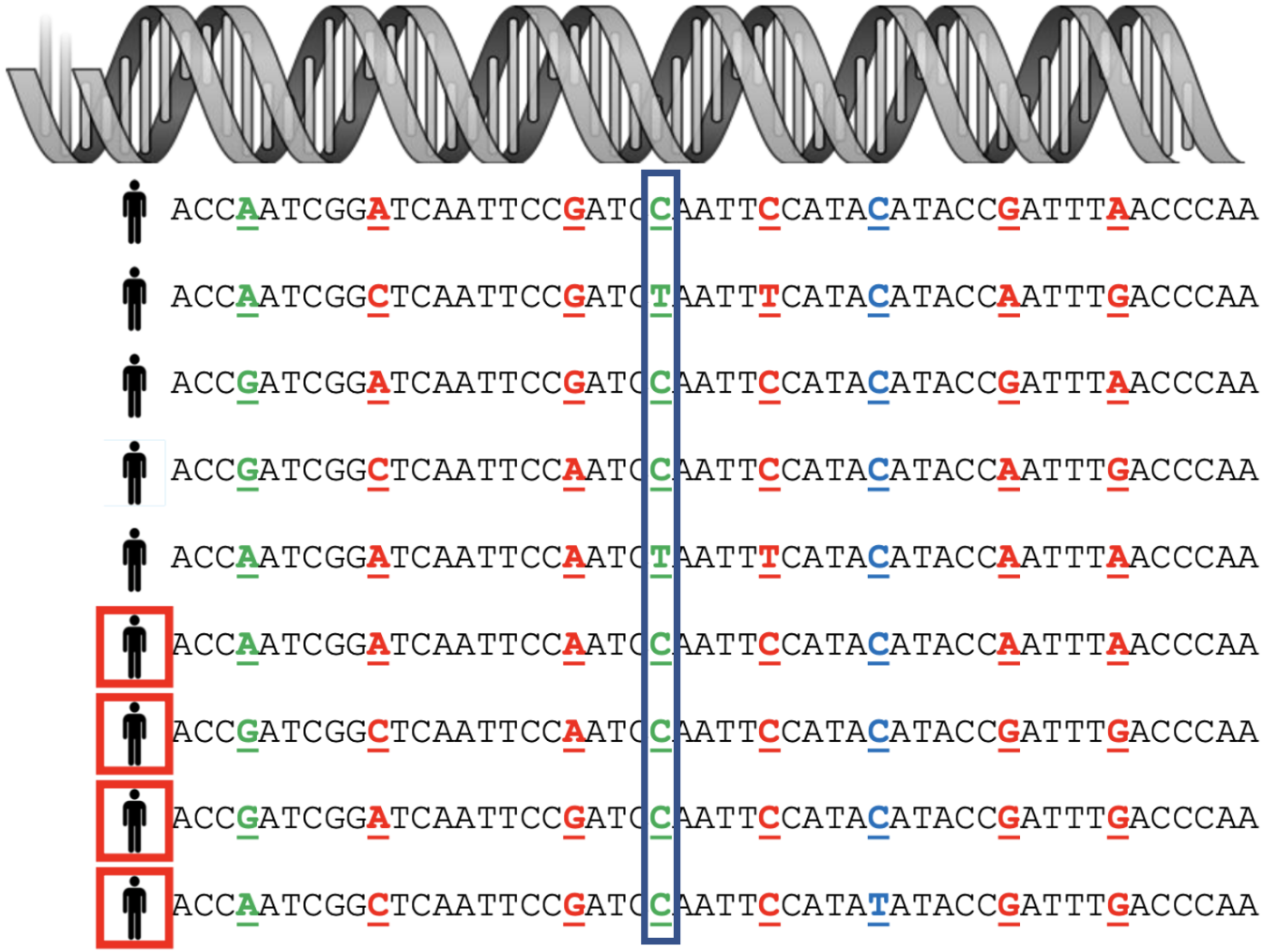
Introduction

- Genome-wide association studies (GWASs) aim to identify loci involved with complex traits.
- **Genotypes:** Technological advances have made it feasible to perform association studies on a genome-wide basis with hundreds of thousands of markers in a single study.
- **Phenotypes:** We consider testing a genetic marker for association with a disease (e.g. affected/unaffected) or a quantitative trait (e.g. height) in a sample of *unrelated subjects*.

Phenotypic vs. genotypic variation



GWAS: test one marker at a time



Quantitative Genetic Model

- The classical quantitative genetics model introduced by Ronald Fisher (1918) is

$$Y = G + E$$

  where Y is the phenotypic value, G is the genetic value, and E is the environmental deviation.
- G is the combination of all genetic loci that influence the phenotypic value and E consists of all non-genetic factors that influence the phenotype (mean set to 0)

Components of Genetic Variance

- Consider a single locus. Fisher modeled the genotypic value G with a linear regression model (least squares) where the genotypic value can be partitioned into an additive component (A) and deviations from additivity as a result of dominance (D), where

$$G = A + D,$$

$$\underbrace{Var(G)}_{\sigma^2_G} = \underbrace{Var(A)}_{\sigma^2_A} + \underbrace{Var(D)}_{\sigma^2_D}$$

- σ^2_A is the **additive genetic variance**. It is the genetic variance associated with the average additive effects of alleles
- σ^2_D is the **dominance genetic variance**. It is the genetic variance associated with the dominance effects.

Heritability

- Remember

$$\begin{aligned} Y &= G + E \\ &= A + D + E, \\ \underbrace{Var(Y)}_{\sigma^2_Y} &= \underbrace{Var(A)}_{\sigma^2_A} + \underbrace{Var(D)}_{\sigma^2_D} + \underbrace{Var(E)}_{\sigma^2_E} \end{aligned}$$

- **Narrow-sense heritability** (or simply heritability) is

$$h^2 = \frac{\sigma^2_A}{\sigma^2_Y}$$

- h^2 is the proportion of the total phenotypic variance due to additive effects.
- It can also be viewed as the extent to which phenotype is determined by the alleles transmitted from the parents.

Heritability

- The **broad-sense heritability** is defined to be

$$H^2 = \frac{\sigma^2_G}{\sigma^2_Y} = \frac{\sigma^2_A + \sigma^2_D}{\sigma^2_Y}$$

- H^2 is the proportion of the total phenotypic variance that is due to all genetic effects (additive and dominance)
- Heritability can vary over time and with the study population as it depends also on environmental effects

QTL Mapping

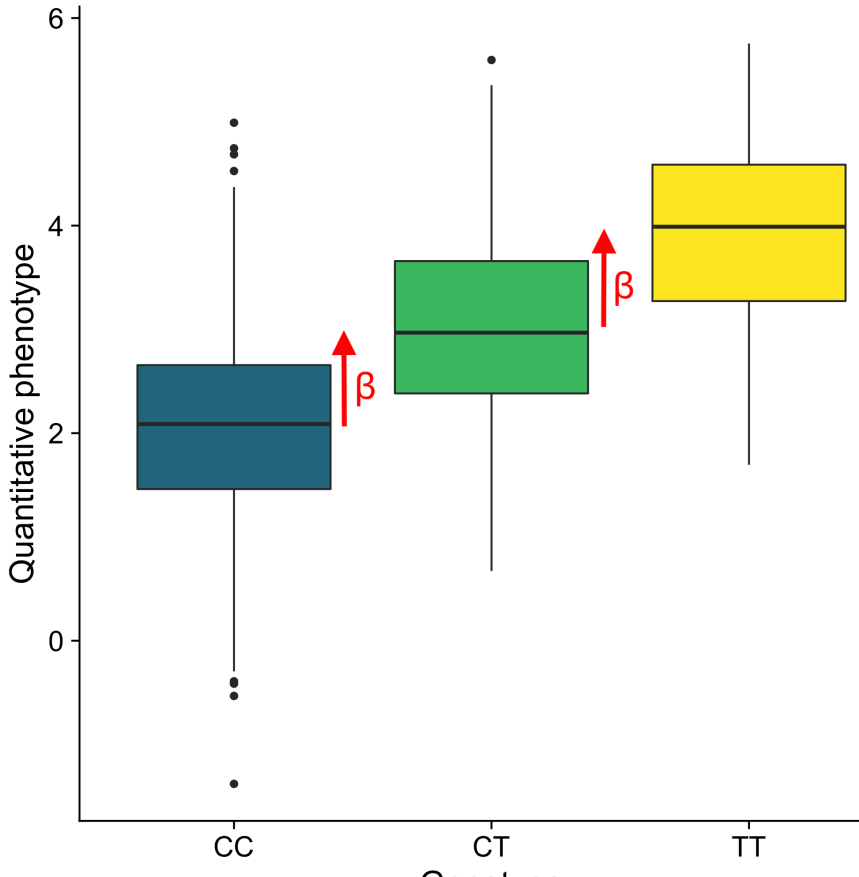
- For traits that are heritable, i.e., traits with a non-negligible genetic component that contributes to phenotypic variability, identifying (or mapping) QTLs that influence the trait is often of interest.
- Linear regression models are commonly used for QTL mapping
  - They will often include a single genetic marker (e.g., a SNP) as predictor in the model, in addition to other relevant covariates (e.g. age, sex), with the quantitative phenotype as the response

Linear regression with SNPs

Many analyses fit the 'additive model'

Let a SNP have C (reference) and T (alternate) alleles

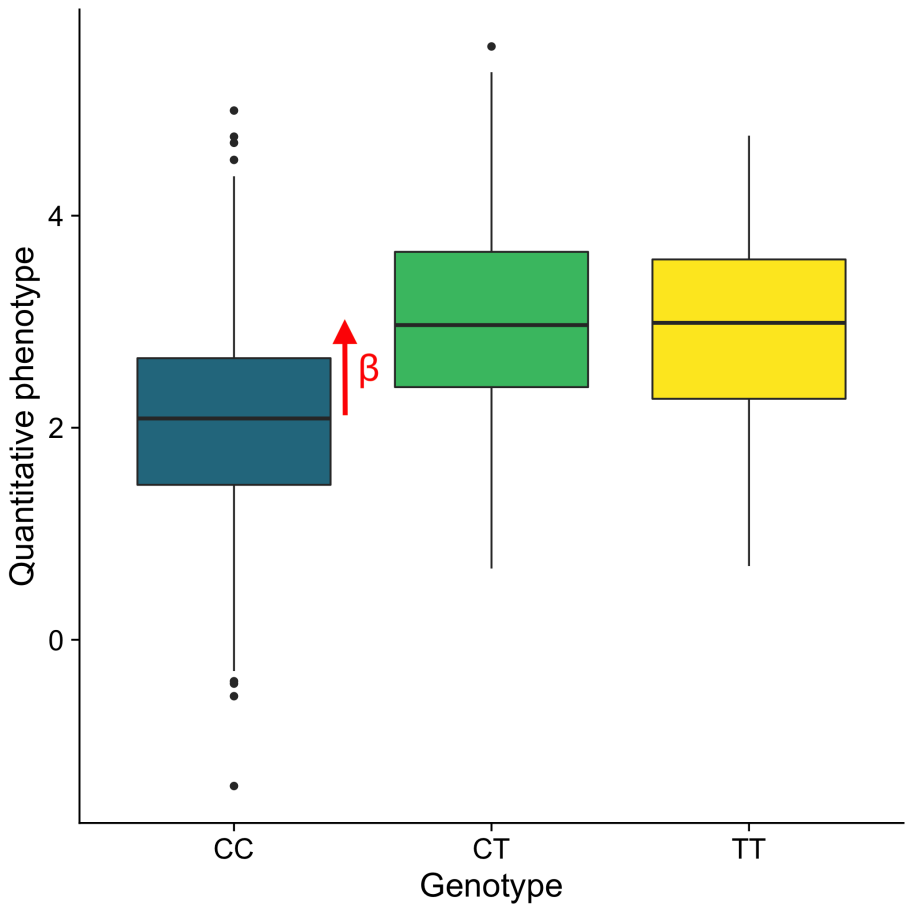
$$y = \beta_0 + \beta \times \#\text{T alleles}$$



Linear regression, with SNPs

An alternative is the 'dominant model';

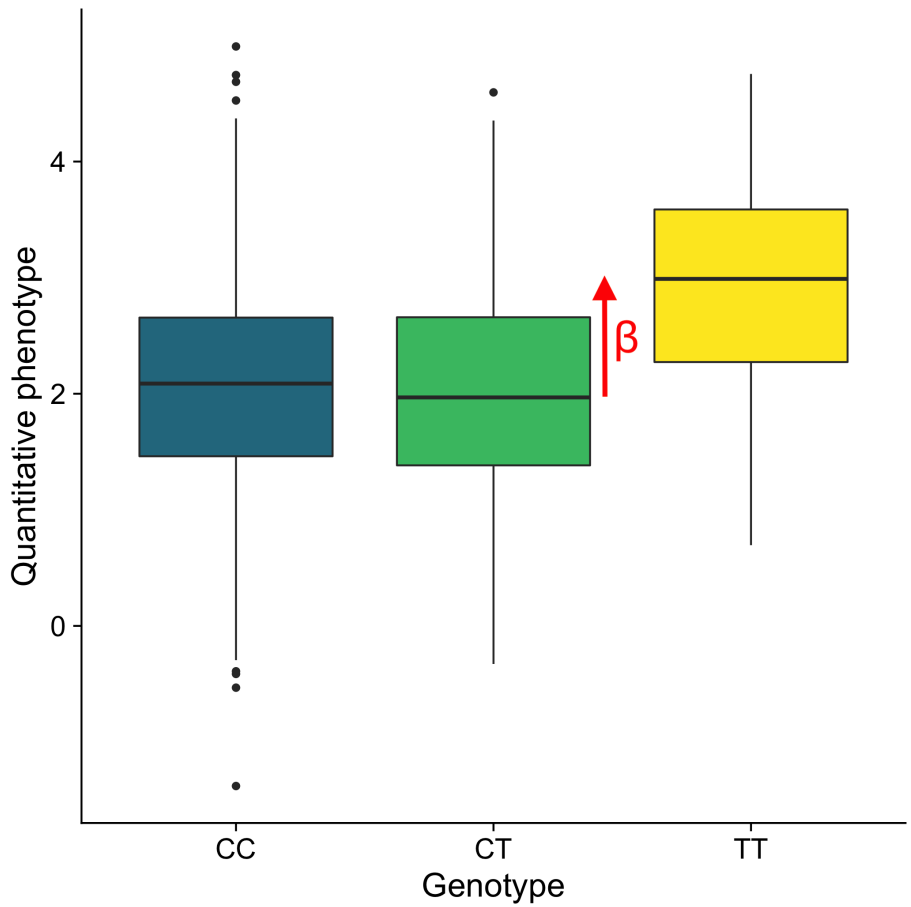
$$y = \beta_0 + \beta \times I\{G \neq CC\}$$



Linear regression, with SNPs

or the 'recessive model';

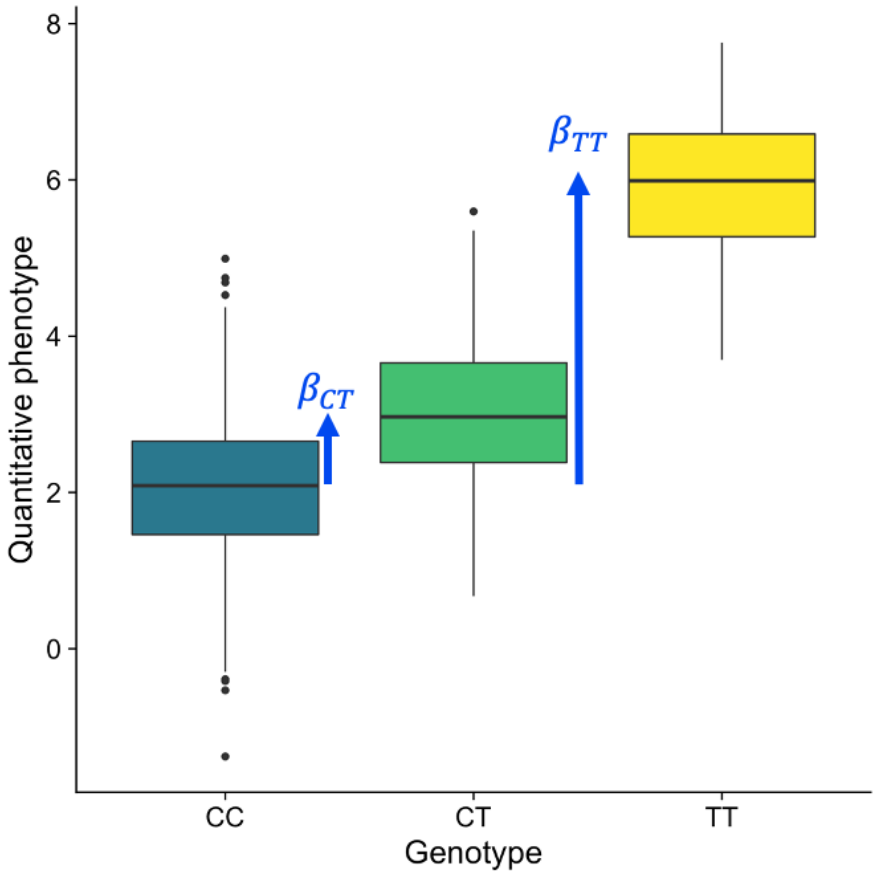
$$y = \beta_0 + \beta \times I\{G == TT\}$$



Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{CT} \times I\{G == CT\} + \beta_{TT} \times I\{G == TT\}$$



Additive Genetic Model

- **Most GWAS perform single SNP association testing with linear regression assuming an additive model.**
- The coefficient of determination (r^2) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the "SNP heritability"

Additive Genetic Model

- Consider the following additive model for association testing with a quantitative trait and a SNP with alleles C and T:

$$Y = \beta_0 + \beta_1 G + \epsilon$$

  where G is the number of copies of the allele T.
- How would you interpret ϵ in this model?

Association Testing with Additive Model

$$Y = \beta_0 + \beta_1 G + \epsilon$$

- Two test statistics for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim \mathbf{t}_{N-2} \approx N(0,1) \text{ for large } N$$

$$T^2 = \frac{\hat{\beta}_1^2}{var(\hat{\beta}_1)} \sim \mathbf{F}_{1,N-2} \approx \chi_1^2 \text{ for large } N$$

where

$$var(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{GG}}$$

and S_{GG} is the corrected sum of squares for the G_i's

Logistic regression for a Binary Trait

- Logistic regression is generally used to get odds ratios and confidence intervals for genotypes.
  - Allows to include other relevant covariates (e.g., age, sex)
- Let π_i be the probability that individual i is affected with the disease and let G_i be the genotype for individual i at the SNP:

$$\log(\text{odds of disease for individual } i | G_i)$$

$$= log\left(\frac{\pi_i}{1-\pi_i}\middle| G_i\right)$$

$$= \beta_0 + \beta_{CT} I\{G_i = CT\} + \beta_{TT} I\{G_i = TT\}$$

where $I\{G_i = CT\}$ is 1 if $G_i = CT$ and 0 otherwise, and similarly for $I\{G_i = TT\}$.

Logistic Regression

- The coefficient estimates for $\hat{\beta}_{CT}$ and $\hat{\beta}_{TT}$ can be used to calculate odds ratios:

$$OR_{CT} = exp(\hat{\beta}_{CT})$$

$$OR_{TT} = exp(\hat{\beta}_{TT})$$

- 95% CI for OR_{CT} is

$$exp(\hat{\beta}_{CT} \pm 1.96 \times s.e.(\hat{\beta}_{CT}))$$

Let's pause ... and discuss the two models

Let $M \sim$ 500,000 SNPs across the genome and $\mathbf{Y}$ is a quantitative trait.

GWAS Model

$$\mathbf{Y} = G_l \alpha_l + \boldsymbol{\epsilon}_l$$

for $l = 1, 2, ..., M$

Whole Genome Regression Model

$$\mathbf{Y} = \sum_{l=1}^{M} G_l \theta_l + \boldsymbol{\epsilon}$$

- Which model is simpler?
- How to fit each model?
- What is interpretation of the model parameters?

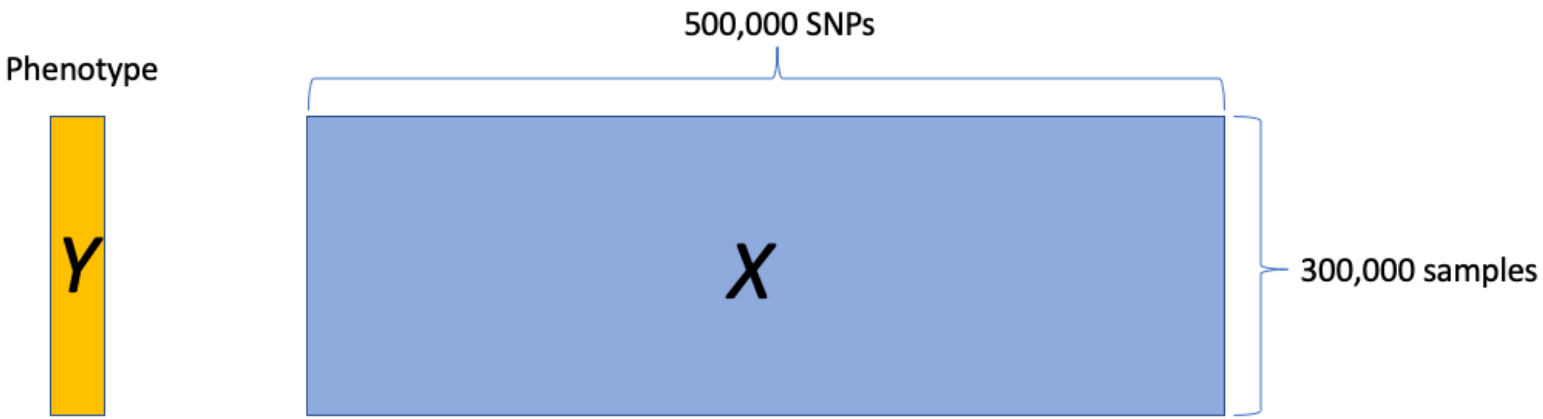
Regenie: Whole Genome Regression

- Step 1: computationally efficient whole genome regression

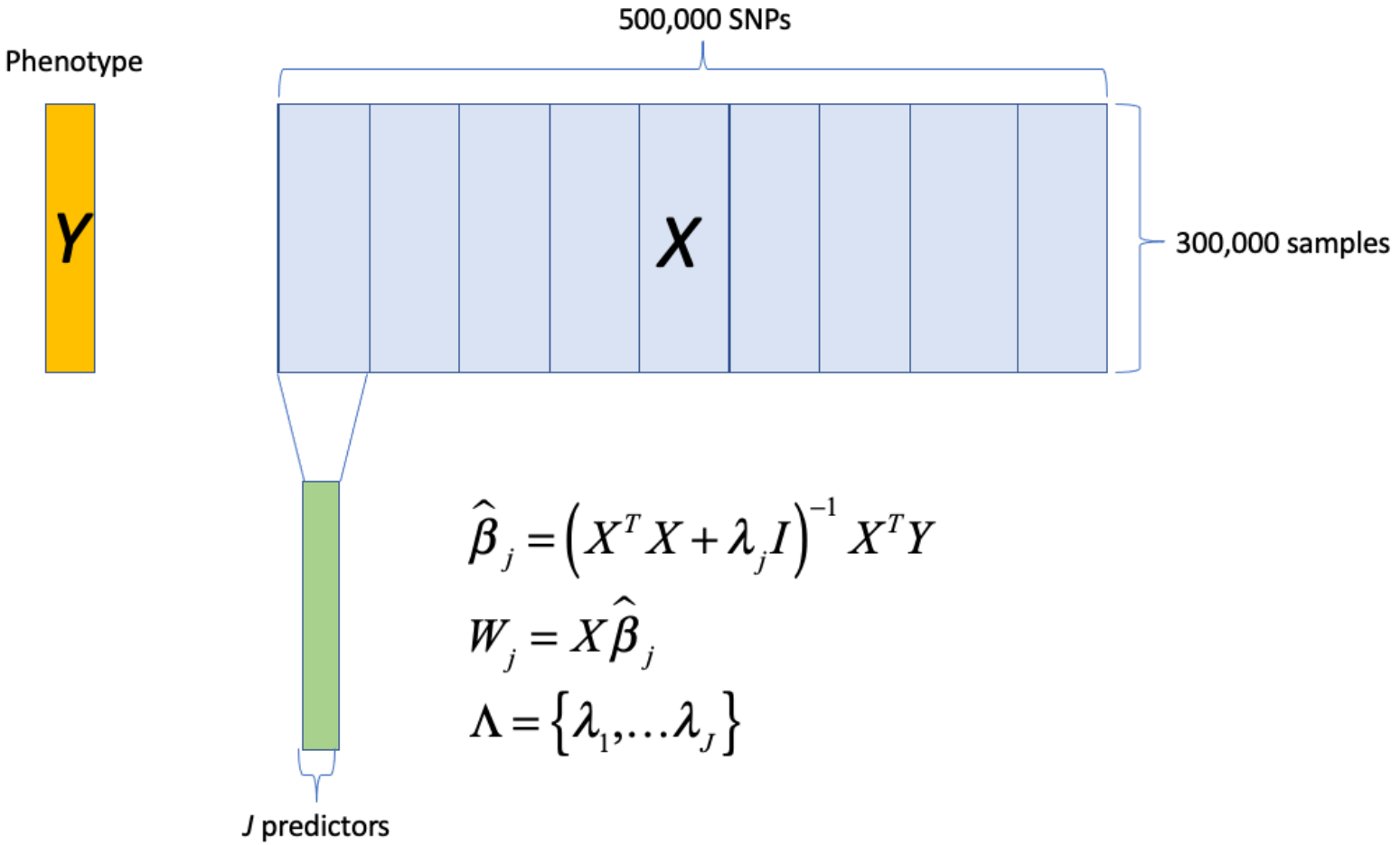
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{M} G_l \theta_l + \boldsymbol{\epsilon}$$

- M is usually $\sim$ 500,000 SNPs across the genome
- Regenie splits genetic data into blocks and runs local regressions in each block to obtain local genetic scores

Regenie: Whole Genome Regression



Regenie: Whole Genome Regression

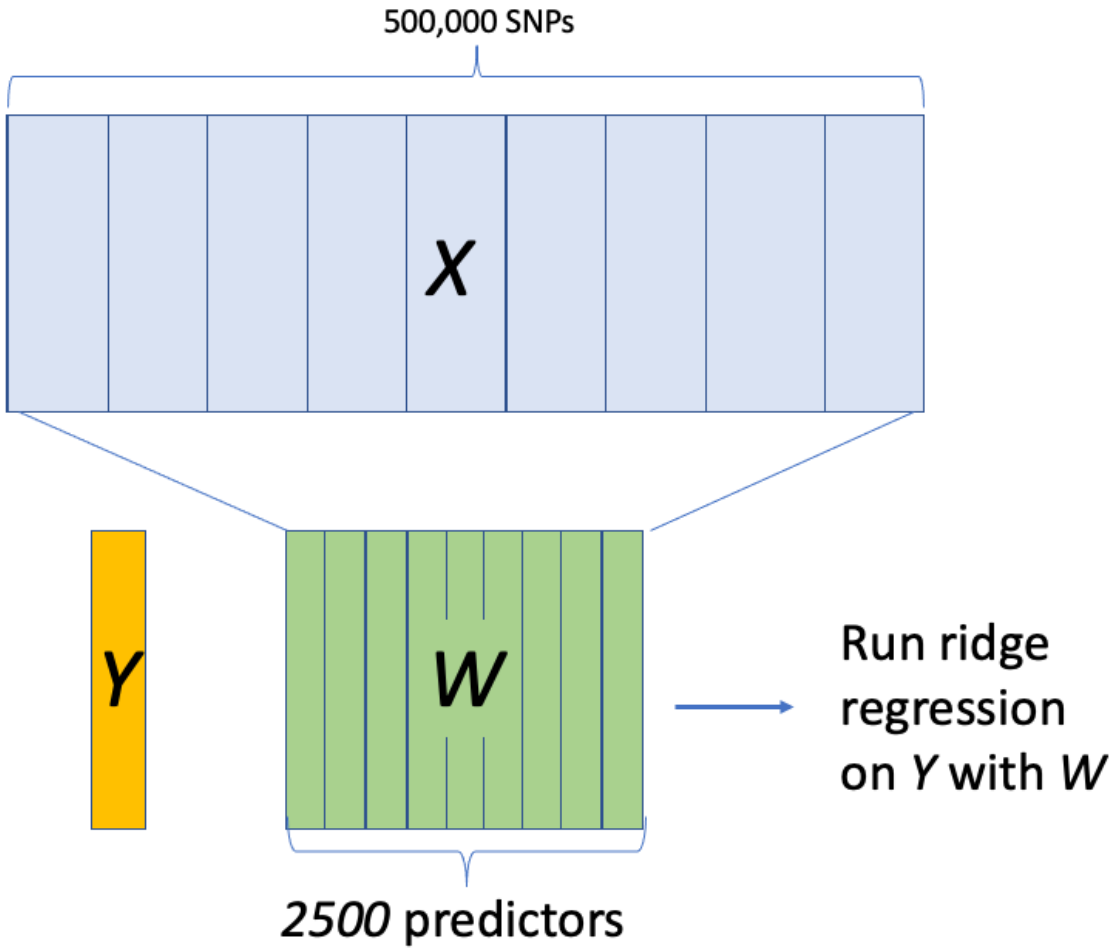


$$\hat{\beta}_j = \left(X^T X + \lambda_j I\right)^{-1} X^T Y$$

$$W_j = X\hat{\beta}_j$$

$$\Lambda = \{\lambda_1, \ldots \lambda_J\}$$

Regenie: Whole Genome Regression



Regenie: Whole Genome Regression

- Step 1: computationally efficient whole genome regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{M} G_l\theta_l + \boldsymbol{\epsilon}$$

- Divide into two levels of regressions
  - Reads genetic data in blocks and within each block fits ridge regression (penalized linear regression)
  - Fit another round of ridge regression on all the block predictors
- Polygenic predictions ($\sum_{l=1}^{M} G_l\hat{\theta}_l$) capture population structure, relatedness as well as polygenicity

Regenie: Whole Genome Regression

- Step 2: test the association parameter γ under the null hypothesis of $H_0 : \gamma = 0$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + G_s\gamma + \sum_{l=1}^{M} G_l\hat{\theta}_l + \boldsymbol{\epsilon}$$

- Test on millions of genetic variants (array/imputed/exome)
- Also works on binary traits where logistic regression is used instead of linear regression

`https://rgcgithub.github.io/regenie/`

PLINK Overview

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner:

  https://www.cog-genomics.org/plink/1.9/
  https://www.cog-genomics.org/plink/2.0/

- PLINK has numerous useful features for genetic data analysis
  - data management: data I/O, support for multiple formats
  - quality control and statistic report
    - allele frequencies, missing genotype rates, HWE test, etc
  - basic association tests (for samples of unrelated subjects)

Input Files

PLINK BED

BIM

Chr ID CM Pos. Ref Alt

```
1 1:12030946:T:C 0 12030946 T C
1 1:12032428:A:C 0 12032428 A C
1 1:12057950:C:T 0 12057950 C T
1 1:12095233:A:C 0 12095233 A C
1 1:12100532:T:C 0 12100532 T C
```

Variant info

FAM

FID. IID Fa Mo Sex Y

```
1432 HGDP00702 0 0 2 -9
1433 HGDP00703 0 0 1 -9
1434 HGDP00704 0 0 2 -9
1436 HGDP00706 0 0 2 -9
1438 HGDP00708 0 0 2 -9
```

Sample info

BED

Compressed binary file (bytes) storing 0/1/2/NA

```
00000000: 01101100 00011011 00000001 11111111 11111110 11111111  l.....
00000006: 11111110 11111111 11111111 11111111 11111111 11111111  ......
0000000c: 11111111 11111111 11111111 11111111 11111110 11111111  ......
00000012: 11111111 11111111 11111111 11111111 11111111 11111111  ......
```

Genotype data

Summary

- Most GWASs perform association tests using linear model + additive coding for genotypes
  - Logistic regression for binary traits
  - Linear regression for quantitative traits
- The whole genome regression approach in Regenie
  - Combines local **prediction** models within blocks of SNPs
  - Fit the final model on all the block predictors
  - $\rightarrow$ polygenic predictions that capture population structure, relatedness and polygenicity

References

- Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097-1103 (2021).