Lecture 3

GWAS Explained: Theoretical Underpinnings and Analysis Strategies

by Dr. Mustafa İsmail Özkaraca

Contents

1. Association Analysis by Linear Regression

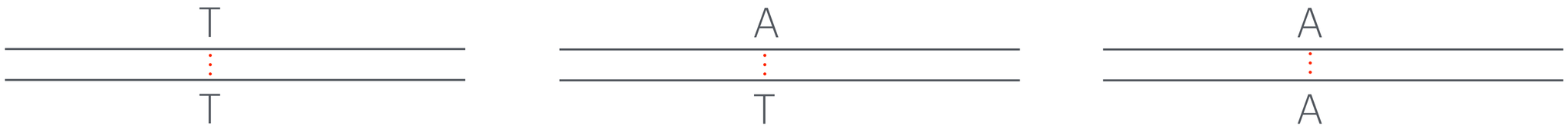
2. Why Linear Regression does not work?

3. Association Analysis by Linear Mixed Models

4. Advantages and (further) Challenges

Association Analysis by Linear Regression

Three cases:



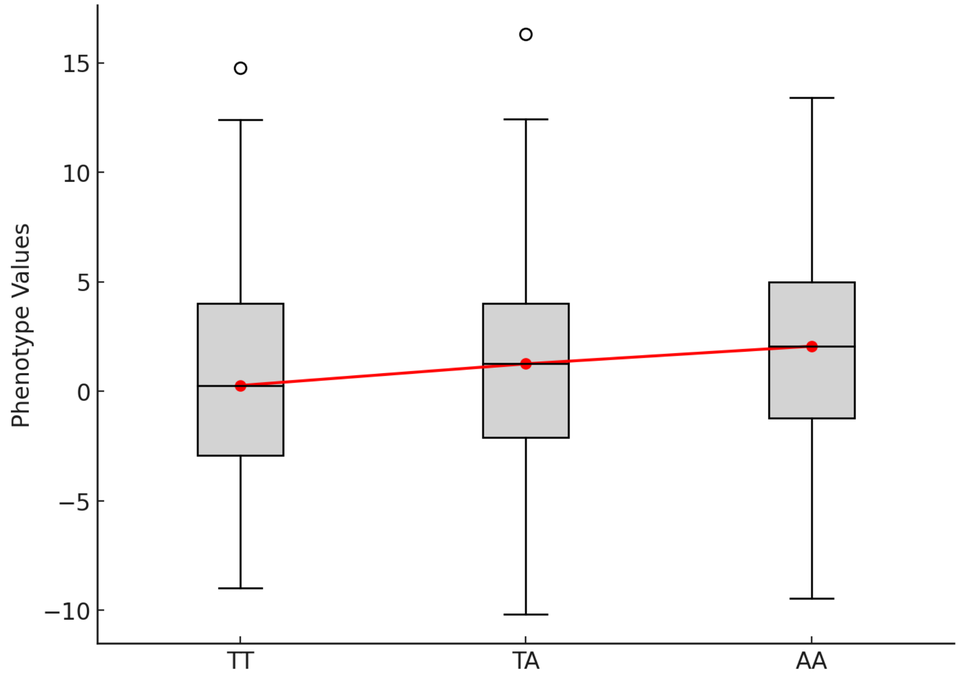
$$y = g\beta + \epsilon$$

y Trait/Phenotype value (vector)

g Number of As (vector)

β Effect size of g (number)

ϵ Random noise (vector) $\sim N(0,\sigma^2)$



Why Linear Regression does not work?

Confounders:

1. Population Structure

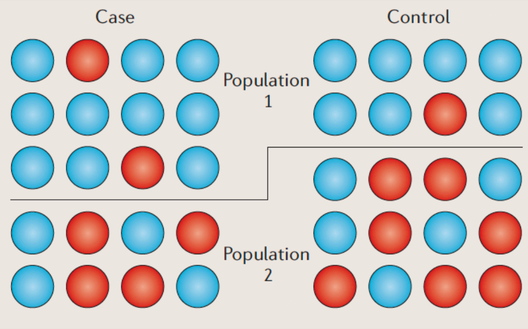


Figure from: Balding, D. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781–791 (2006).

2. Related Individuals

John Krasinski (191cm) and his brothers (203 cm, 205 cm)

Info from: https://healthyceleb.com/john-krasinski

3. Linkage Disequilibrium (LD)

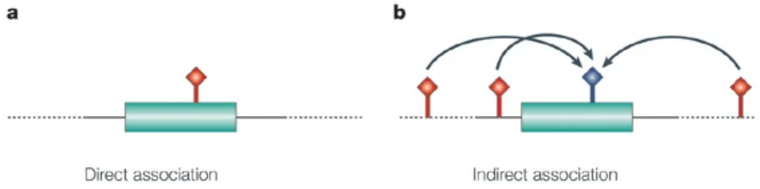


Figure from: Hirschhorn, J., Daly, M. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95–108 (2005).

Why Linear Regression does not work?

Multivariate Linear Regression **is not suitable** because

1. Multicollinearity: High correlations among predictor variables

2. Low Statistical Power: High degree of freedom due to large number of predictors

Association Analysis by Linear Mixed Models (LMM)

$$y = g\beta + X_c\alpha_c + G + E \quad [1]$$

$$GLS(\beta) = \hat{\beta} = \frac{g^t V^{-1} y}{g^t V^{-1} g}$$

$$Var(\hat{\beta}) = \frac{1}{g^t V^{-1} g} \text{ with } V = \pi\sigma_G^2 + I\sigma_E^2$$

Hypothesis Testing: $\frac{\hat{\beta}^2}{Var(\hat{\beta})} \sim \chi_1^2(0)$ (Null)

g :Variant to be tested,

β :Effect size of g,

X_c :Matrix of covariates,

α_c :Effects of covariates,

G :Total genetic effects, $G \sim N(0, \pi\sigma_G^2)$, where π is the GRM,

E :Residuals, $E \sim N(0, I\sigma_E^2)$.

[1] Jiang, L., Zheng, Z., Qi, T. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* **51**, 1749–1755 (2019).

Advantages and (further) Challenges

Linear Mixed Models (LMMs) control for population structure.

$$y = g\beta + X_c\alpha_c + G + E$$

X_c typically contains top PCA components of genetic values

Other common components of X_c:

Age, Sex, Batch Centre

Figure 1: Population structure within Europe.

From: Genes mirror geography within Europe

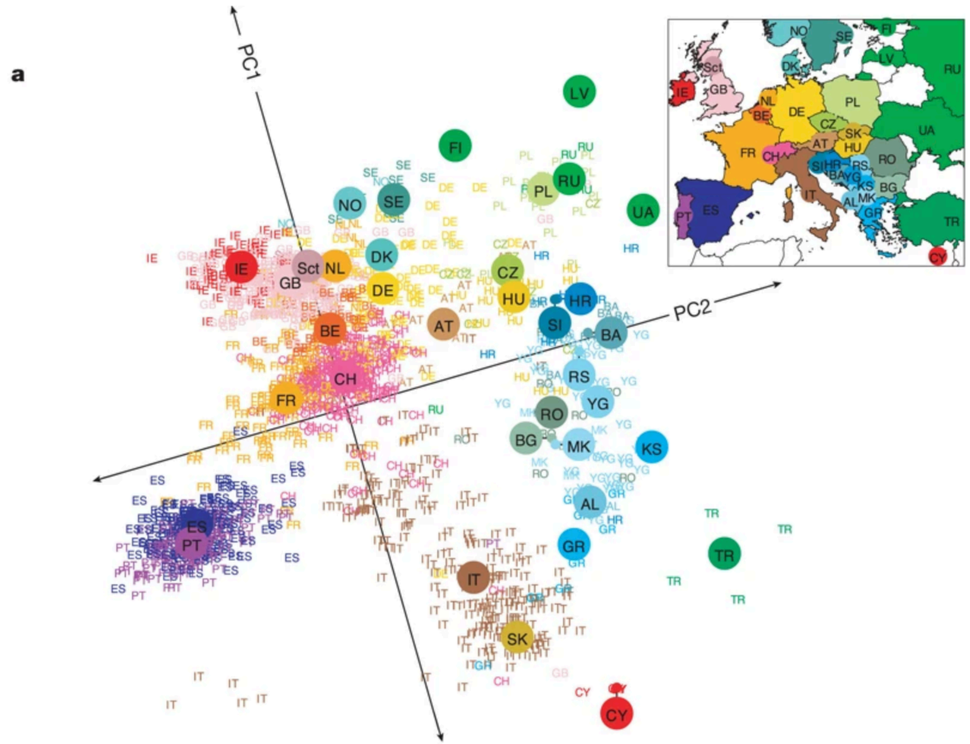


Figure from: Novembre, J., Johnson, T., Bryc, K. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

Advantages and (further) Challenges

Linear Mixed Models (LMMs) control for family relatedness.

$$y = g\beta + X_c\alpha_c + G + E$$

G :Total genetic effects, $G \sim N(0, \pi\sigma_G^2)$, where π is the GRM,

E :Residuals, $E \sim N(0, I\sigma_E^2)$.

A. Reduces to Linear Regression whenever $\pi = I$ because $G + E = \tilde{E}$ with $\tilde{E} \sim N(0, I\sigma_{\tilde{E}}^2)$

B. Genetically similar individuals have similar environmental/residual variance contributions

Advantages and (further) Challenges

Linear Mixed Models (LMMs) control for LD.

$$y = g\beta + X_c\alpha_c + G + E$$

G :Total genetic effects, $G \sim N(0, \pi\sigma_G^2)$, where π is the GRM

Leave-One Chromosome Out (LOCO)

If variant to be tested is in Chromosome 1, then π is the GRM generated by variants from Chromosome 2-22.
If variant to be tested is in Chromosome i, then π is the GRM generated by variants from Chromosome 1-22 except Chromosome i.

That is, 22 many π matrices are computed/generated (Computationally Tractable).

Advantages and (further) Challenges

Linear Mixed Models (LMMs) control for LD.

Why LOCO?

1. LOCO removes LD effects of markers (confounders) from the same chromosome.

2. Excluding only high-LD variants linked to the variant being tested is computationally intractable.

3. Using all chromosomes can reduce power due to “proximal contamination” [1].

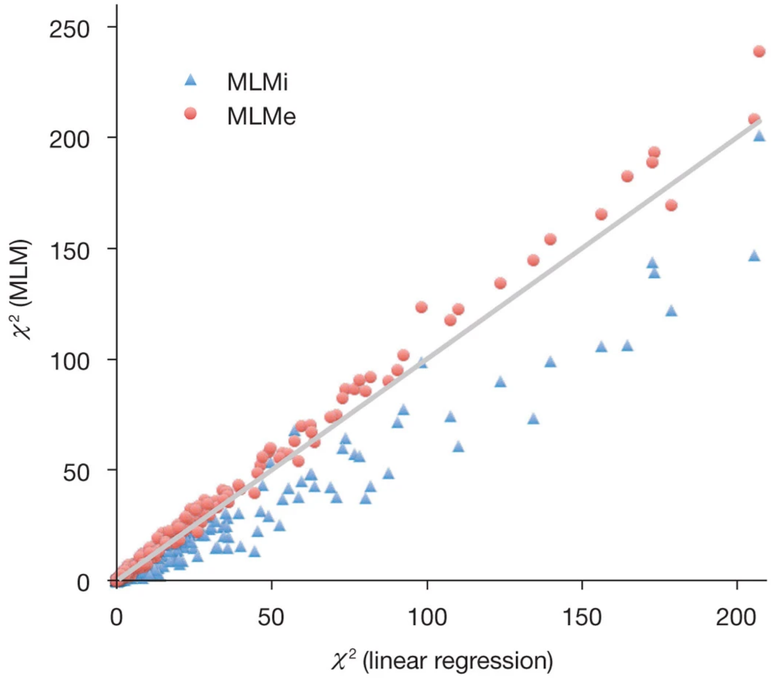


Figure from:
Yang, J., Zaitlen, N., Goddard, M. *et al.*
Advantages and pitfalls in the application of mixed-model association methods.
Nat Genet **46**, 100–106 (2014).

[1] Listgarten, J., Lippert, C., Kadie, C. *et al.* Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**, 525–526 (2012).

Advantages and (further) Challenges

Further Challenges:

Multiple Testing Burden — 5 mistakes on 100 questions → 20,000 mistakes on 1,000,000 questions

GWAS requires large-scale datasets

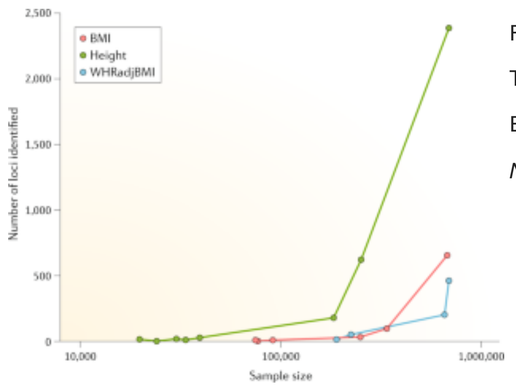


Figure from:

Tam, V., Patel, N., Turcotte, M. *et al.*

Benefits and limitations of genome-wide association studies.

Nat Rev Genet **20**, 467–484 (2019).

GWAS software capable handling large-scale datasets

What’s Next

1. What is Meta-Analysis?

2. Why Meta-Analyse in GWAS?

3. Types of Meta-Analysis in GWAS