Genome-wide association mapping

Brian Kissmer

USU Department of Biology

Nov. 14th, 2024

What is genome-wide association mapping?

Genome-wide association (GWA) mapping is a set of methods used to identify genetic variants associated with variation in particular traits or disease susceptibility.

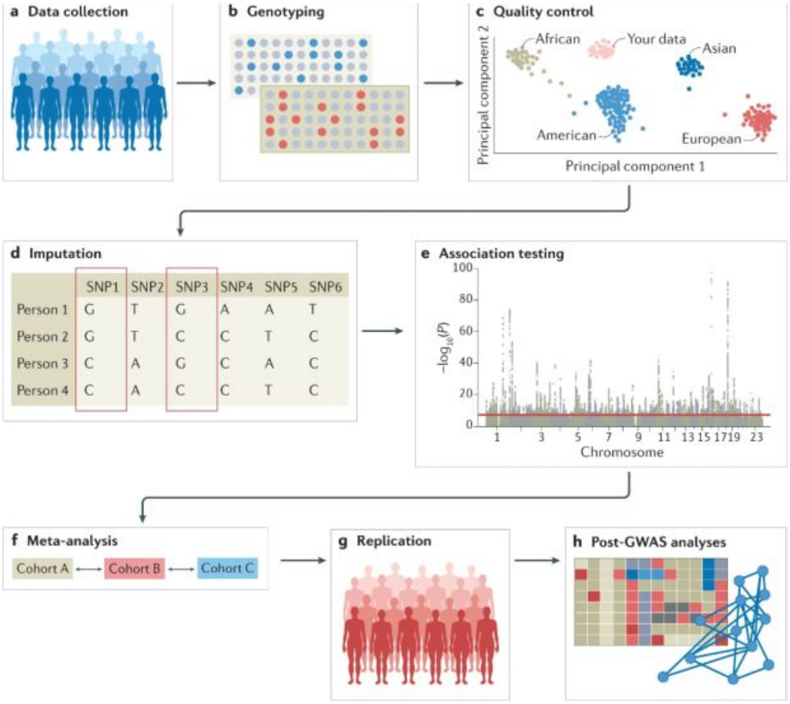
Why is genome-wide association mapping useful?

1. Provides insights into the genetic basis of complex traits and diseases
2. Medical applications: helps identify potential therapeutic targets.
3. Evolutionary biology: aids in understanding the genetic architecture of traits in diverse populations and species.

Basic principles of genome-wide association mapping

- ➢ Observational study of a genome-wide set of genetic variants in different individuals.
- ➢ Tests for statistical association between genetic variants and traits.
- ➢ Typically use single-nucleotide polymorphisms (SNPs) genetic variants.
- ➢ Statistical association is NOT equivalent to causal effect.

Overview of steps for conducting GWA mapping



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

[Uffelmann et al., 2021]

Linear regression models for GWA mapping

Standard linear model for phenotype (y_i) as a function of genotype (g_i) for individual i:

$$y_i = \beta_0 + \beta_{SNP}\, g_i + \epsilon_i$$

Linear regression models for GWA mapping

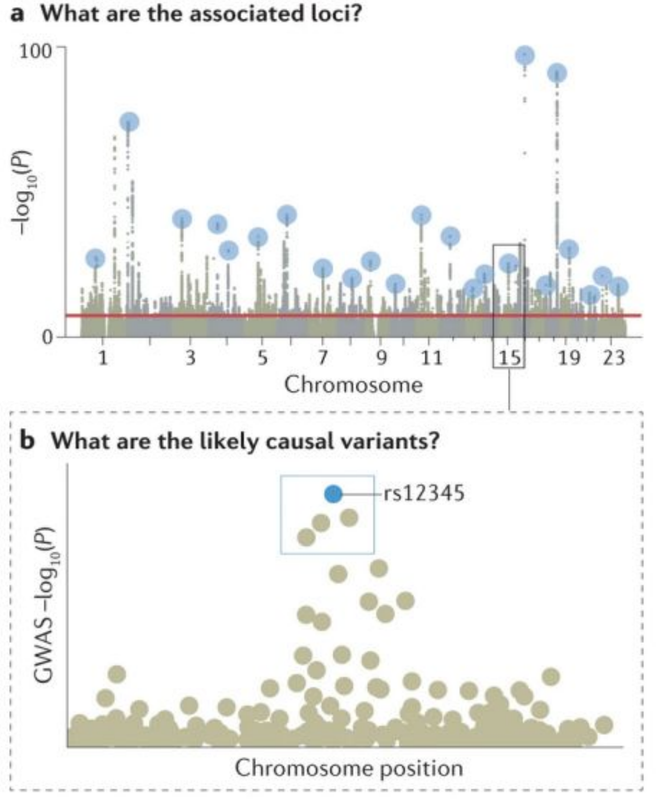
Standard linear model for phenotype (y_i) as a function of genotype (g_i) for individual i:

$$y_i = \beta_0 + \beta_{SNP}\, g_i + \epsilon_i$$

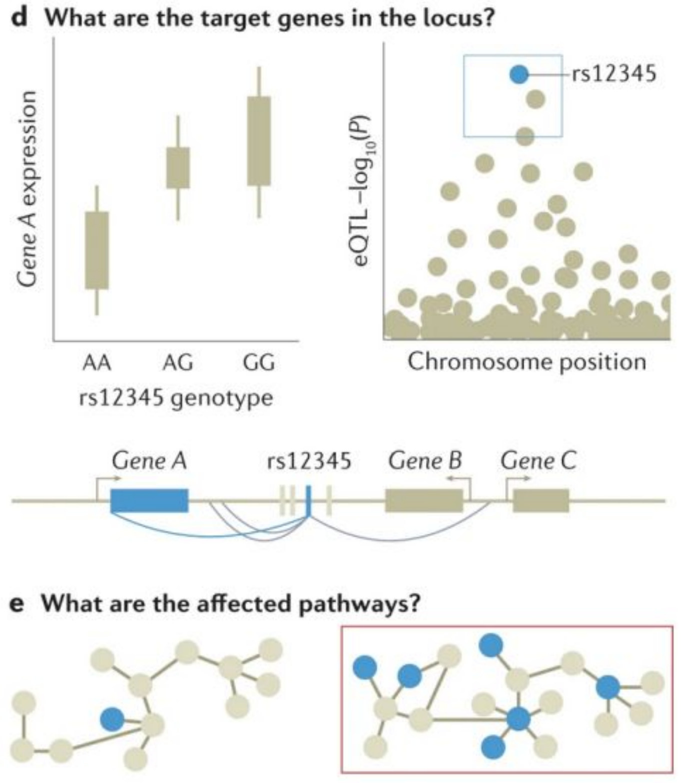
Models can include additional covariates (x), such as environmental effects, organism attributes (sex, age, etc.), or genetic background:

$$yi = \beta_0 + \beta_{SNP}\, g_i + \alpha_1\, x_{1i} + \ldots + \alpha_k\, x_{ki} + \epsilon_i$$

Putative causal variants prioritized based on patterns of association



GWA signals can be associated with changes in gene expression and molecular pathways



Challenges for genome-wide association mapping

- Many traits are polygenic, i.e., influenced by many genes often with small and contingent effects.
- Population stratification and environmental influences can lead to false associations.
- Very large sample sizes are often needed to increase the power to detect “true” (useful) associations.

GWA mapping methods to increase power

Common methods test one genetic variant (SNP) at a time, this leads to low power

$$y_i = \beta_0 + \beta_{SNP}\, g_i$$

Multi-locus models that test many genetic variants simultaneously can increase power and better account for redundant associations:

$$yi = \beta_0 + \beta_1\, g_{1i} + . . . + \beta_k\, g_{ki}$$

How to fit a multilocus model

- Traditional regression methods do not work when the number of parameters exceeds the number of observations, which is often the case for multi-locus GWA mapping analyses.
- Two possible solutions:
  - Use penalized regression, e.g., LASSO (last week)
  - Use machine learning, e.g., Random Forest (more on this later)

GWA in practice

See programming project 5