

Introduction to GWAS and Polygenic Risk Score

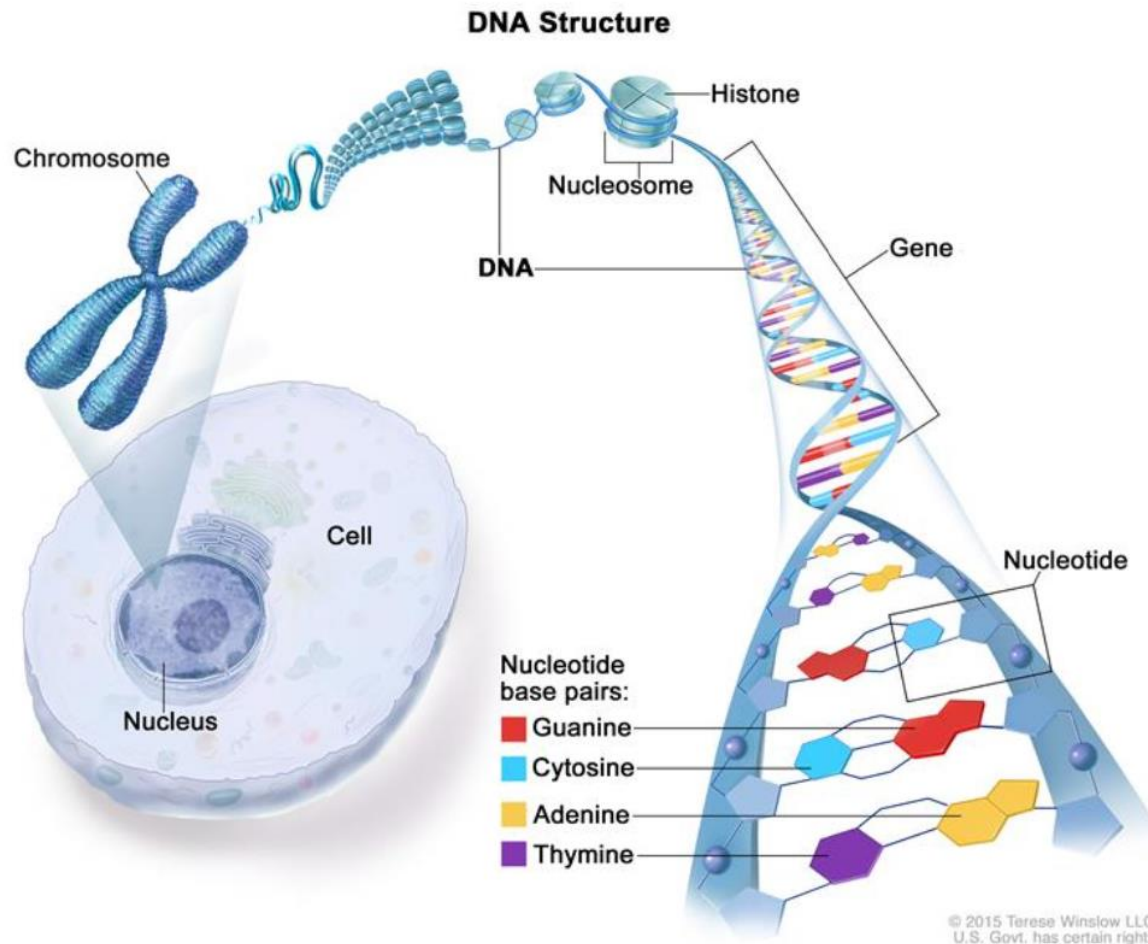
Kipoong Kim

Department of Statistics, Changwon National University

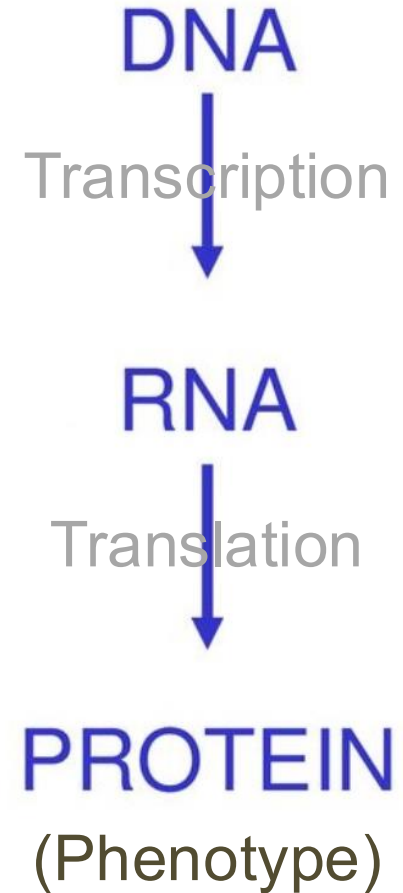
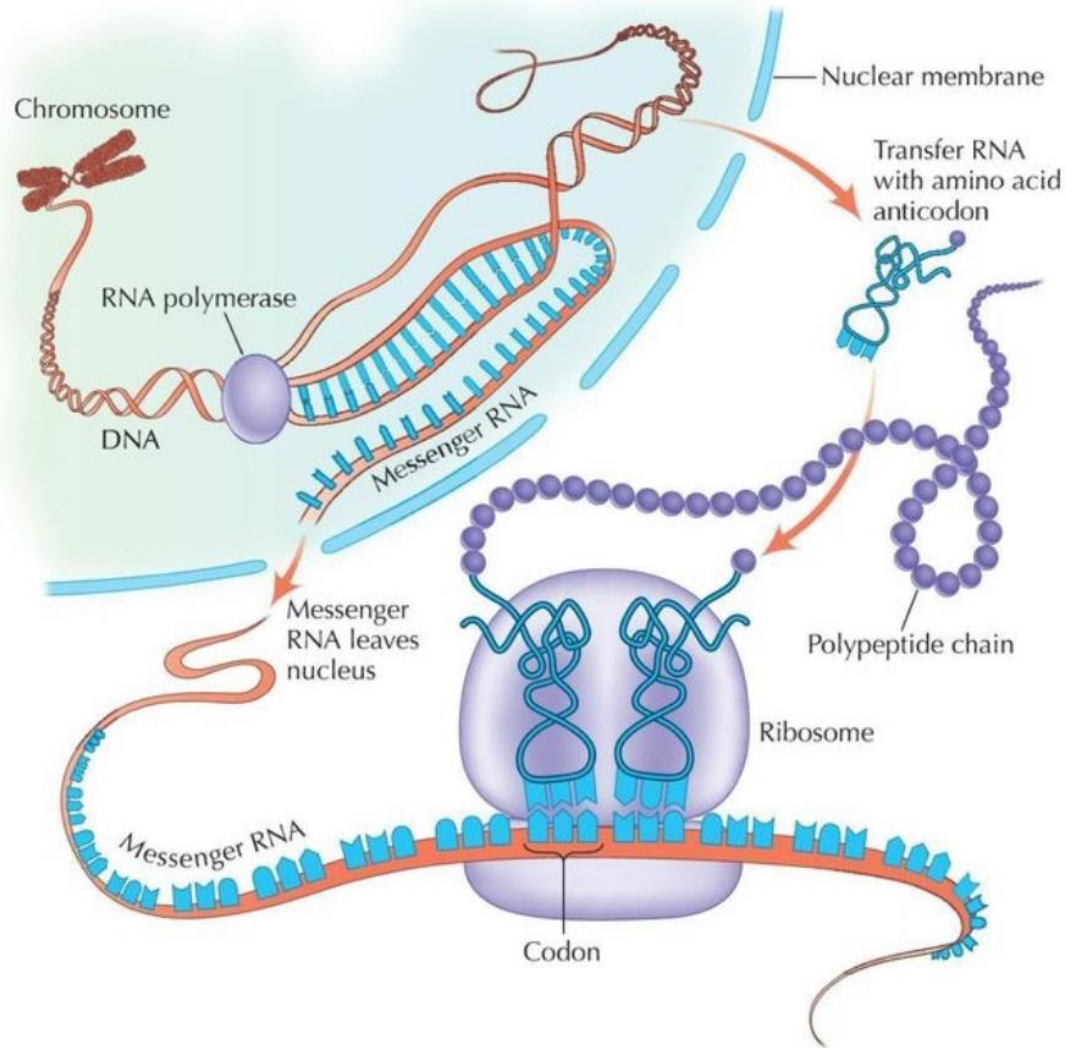
January 2026

DNA...?

- Cell → Nucleus → Chromosome → DNA



The central dogma of molecular biology



Human Genome Overview

- Total Genome Length \approx 3 billion base pairs
- Inter-individual Genomic Similarity \approx 99.9%
Genomic Differences \approx 0.1% (3 million base pairs)
- These differences are called “Single Nucleotide Polymorphisms (SNPs)”



Single Nucleotide Polymorphism (SNP)

- Human Genome Project

- Collected allele data across nearly the entire human genome.

Reference Genome

5' -	A	G	C	T	G	A	T	A	G	C	T	C	T	G	A	C	G	A	G	C	C	C	G	A	T	C	-3'
MOM	A	G	C	T	G	A	T	A	G	C	T	C	T	G	A	C	G	A	G	C	C	C	G	A	T	C	
DAD	A	G	C	T	G	A	T	A	G	C	T	A	T	G	A	C	G	A	G	C	C	C	G	A	T	C	

A diploid genome

(Homozygote Reference) CC

(Homozygote Alternate) AA

(Heterozygote) AC

} Genotype

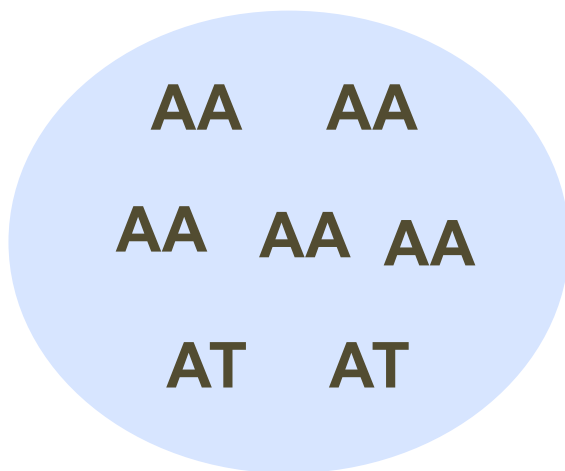
- SNP genotyping

- At each SNP location, we observe genotype information that reflects the combination of alleles inherited from both parents.

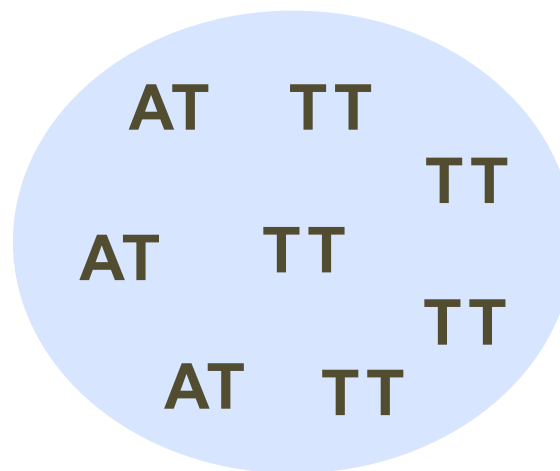
Genome-Wide Association Study (GWAS)

- Identify genetic variants (e.g., **SNPs**) associated with specific **traits or diseases** of interest.
 - e.g. cholesterol levels or well-being outcomes
- e.g., at a certain SNP, we observed the following **genotypes**

High-cholesterol group



Low-cholesterol group



Statistical Testing in SNP Analysis

- Testing methods
 - Continuous traits:
 - Two-sample comparison: t-test
 - Multiple group comparison: ANOVA
 - Simple linear regression
 - Categorical traits:
 - Chi-squared test, Fisher's exact test
 - Logistic / Multinomial regression
 - etc. (more details on this later.)
- We can prioritize SNPs through statistical significance based on their **p-values**.

SNP data structure

Sample ID	SNPs					
	1	2	3	4	5	...
1	AA	GC	CC	TT	GC	
2	AG	GC	CC	TT	GC	
3	GG	CC	TT	AT	CC	
4	AG	CC	TC	TT	CC	
5	AG	CC	TT	AT	CC	...
6	GG	GC	TC	AT	CC	
7	GG	CC	TC	TT	CC	
8	AG	CC	CC	TT	CC	
9	GG	CC	CC	TT	CC	
10	GG	GG	CC	AT	CC	
...						

GG=0 CC=0 ... CC=0
 AG=1 GC=1 ... GC=1
 AA=2 GG=2 ... GG=2

Dependent variable

- Disease
- Phenotypes
- Psychological outcomes
- etc.

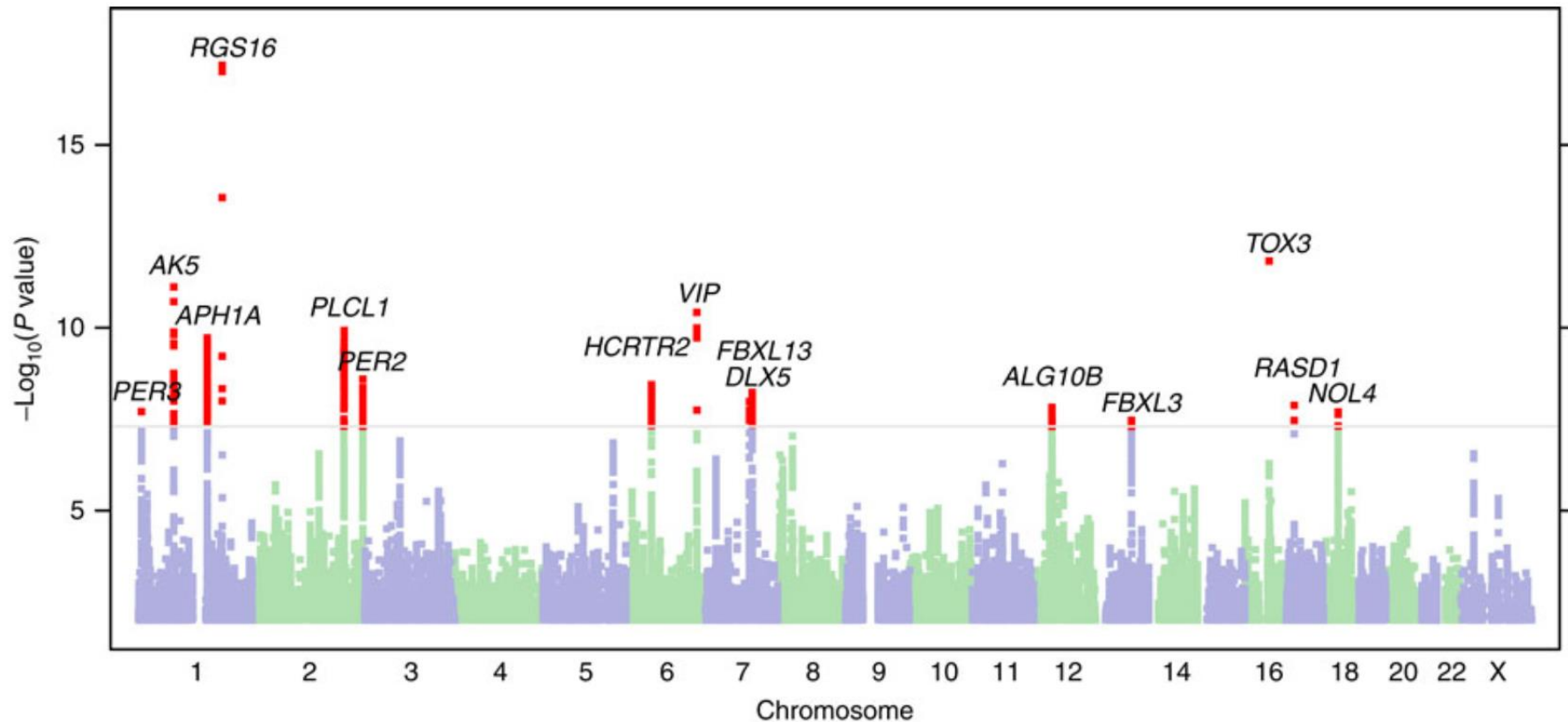
Covariates

- Age
- Gender
- Genomic PCA
- etc.

PC adjustment for GWAS

- What is PCA?
 - A simple math tool that finds the biggest patterns in genetic data.
- Goal:
 - Adjust for population stratification and confounding intrinsic to genomic data.
- Concept:
 - Extract principal components (PCs) from genomic data representing major genetic variation.
 - Each PC summarizes an individual's genetic background.
- Why It's Needed:
 - Minimizes false positives due to population structure.

Manhattan plot



- In GWAS, it is important to detect truly causal SNPs correctly from limited sample sizes.

Large sample theory in association test

- Total Bias in GWAS association test

$$\text{Total Bias} = \text{Systematic Bias} + \text{Estimation Bias}$$

- Systematic bias

- Arises from the selection of testing methods and study design.
- Choosing appropriate testing methods → systematic bias ↓

- Estimation bias

- Occurs when estimating the true parameter from a limited sample size.
- The sample size ↑ → consistency of estimators → estimation bias ↓

- Total Bias ↓ ≡ Statistical Power ↑ & False Discovery ↓

- Detects subtle effects of individual SNPs that contribute small increments to phenotypic variance.

Introduction to Polygenic Risk Scores (PRS)

- Research Question:

- Quantify complex traits (e.g., [happiness](#)) for each individual using SNP data.

- Regression model:

$$\text{Happiness} = \beta_1 \times \text{SNP}_1 + \cdots + \beta_p \times \text{SNP}_p + \text{error}$$

- Each β_j represents the effect of SNP_j on the trait.

- Polygenic Risk Score (PRS):

- Quantifies the cumulative effect of many genetic variants (usually SNPs) on an individual's predisposition to a particular trait or disease.

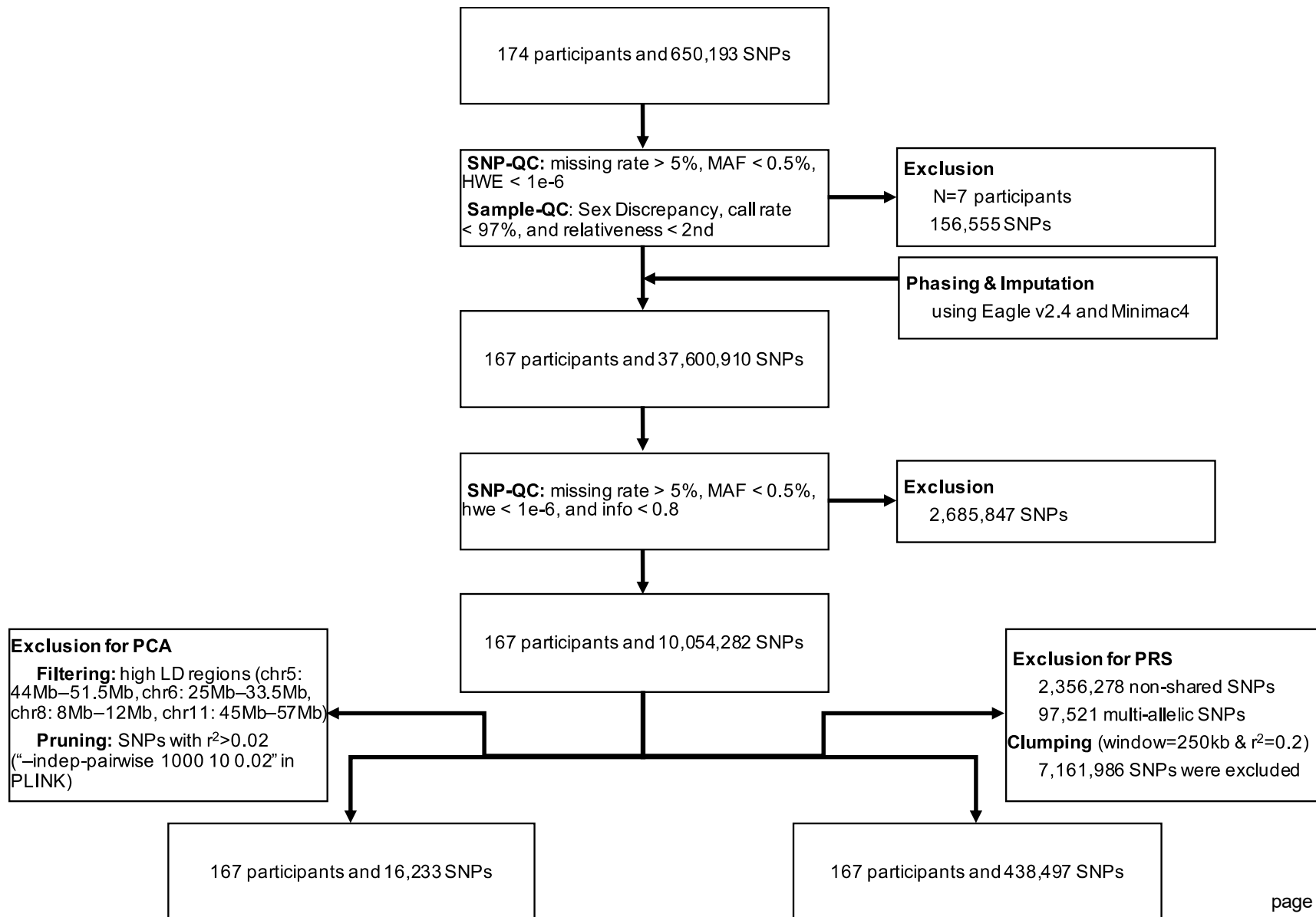
$$\text{PRS} = \sum_{j=1}^p \hat{\beta}_j \times \text{SNP}_j \quad \text{for } j \in \{j: \text{p-value}_j < \alpha\}$$

※ α is a significance level.

PRS calculation using large-scale dataset

- Typically,
 - Each SNP's **effect size** $\hat{\beta}_j$ is directly taken from the estimates derived from **large-scale GWAS results (summary statistics)**.
- Why **large-scale**?
 - Higher Statistical Power: Large sample sizes enable more precise estimation of SNP effect sizes.
- + LD Clumping:
 - Prune SNPs in high **Linkage Disequilibrium (LD)** to avoid redundancy.

GWAS workflow



GWAS workflow

1. Data preparation

- Collect and integrate genomic and clinical data.

2. Sample QC / SNP QC

- High missingness, gender disparity, outliers
- High missingness, low MAF (rare variants), Hardy-Weinberg Disequilibrium

3. Phasing & Imputation → SNP QC

- Estimate haplotype structures by leveraging SNP correlations.
- Predict missing genotypes using reference panels.

4. GWAS & LD Clumping

- Estimate regression coefficients for SNPs.
- Select representative SNPs, reducing redundancy.

5. PRS calculation

- Compute PRS by weighting SNP effects (e.g., from large-scale GWAS results).

Thank you for your attention ! 