Introduction to GWAS and Polygenic Risk Score

Kipoong Kim

Department of Statistics, Changwon National University

February 19, 2025

Kipoong Kim

- History
  - PhD in Statistics, Pusan National Univ.
  - PostDoc. in Statistics, Seoul National Univ.
  - Assistant Prof. in Statistics, Changwon National Univ.

- Research Area
  - High-dimensional data analysis
  - Multi-source data integration
  - Non-Euclidean data analysis
    - e.g. compositional, spherical

Previous research topics

- Statistical variable selection methods for (epi-)genomic data
  - Gene expression data, SNP data, DNA methylation data
  - Pleiotropy

- Multi-omics data integration with multi-response outcomes
  - $Y_1, \ldots, Y_q \sim [X_1, \ldots, X_{p_1}] + [X_1, \ldots, X_{p_2}] + \cdots + [X_1, \ldots, X_{p_d}]$

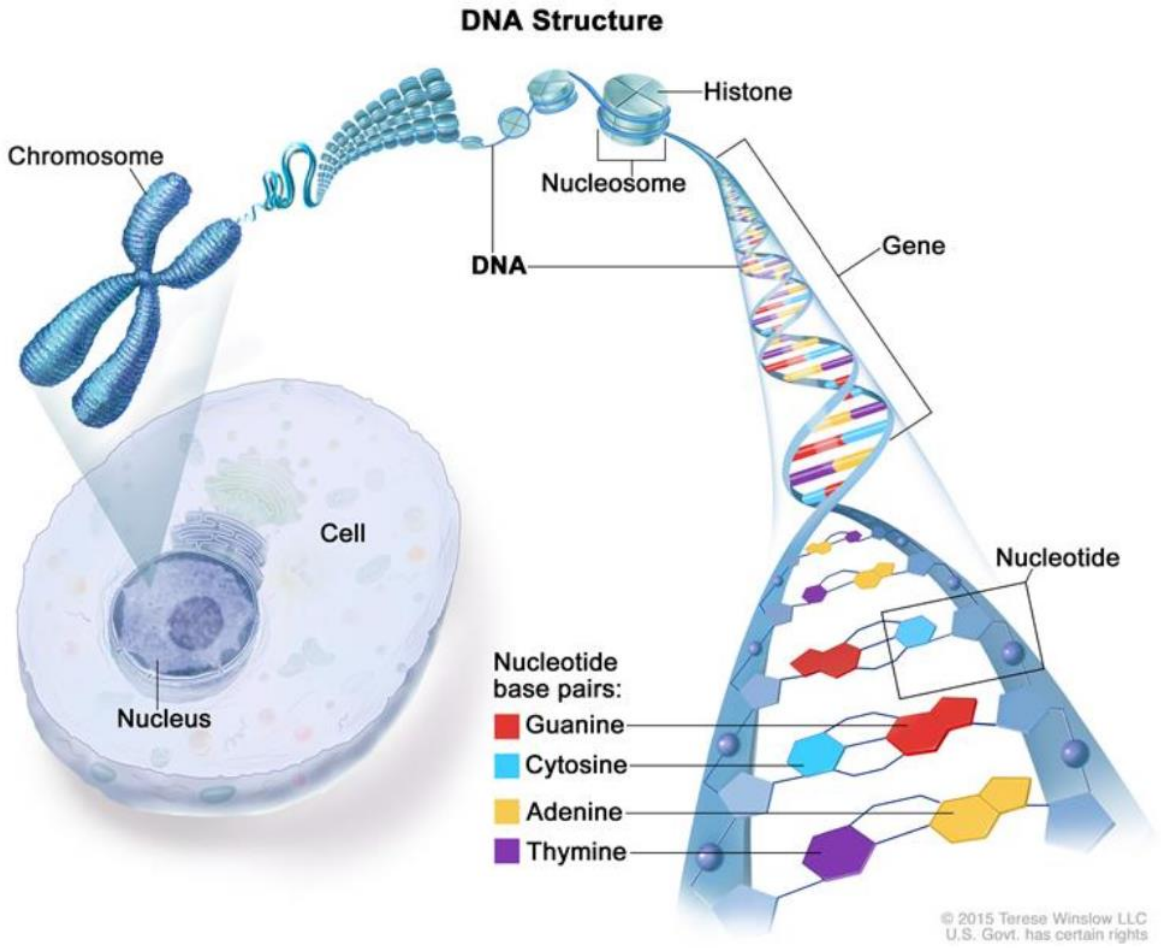
- Principal component analysis for zero-inflated microbiome data

Ongoing projects

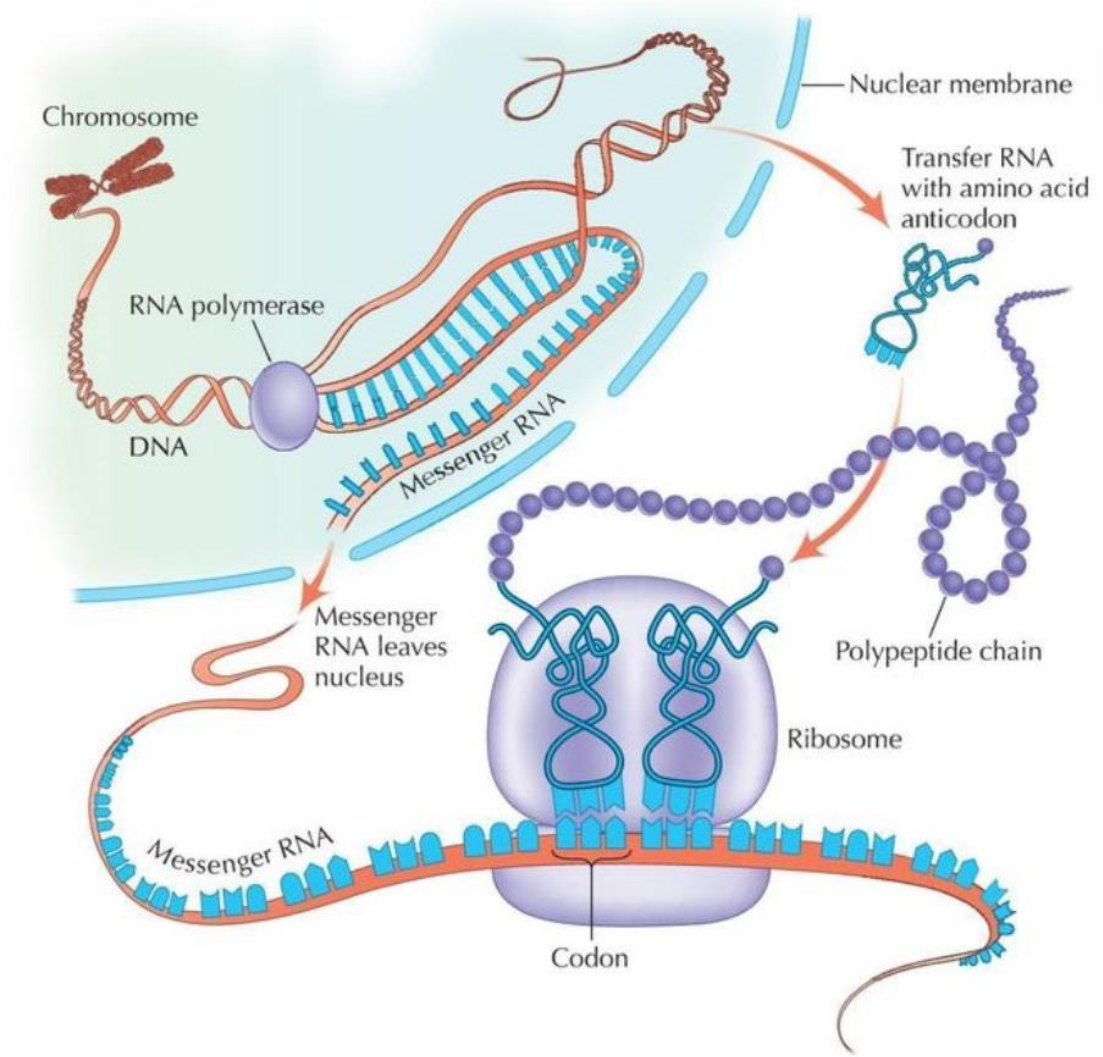
- Microbiome data integration
  - Integrative analysis of multiple sources: [Urine, Serum, Stool]
- Regression for Diffusion Tensor Imaging (DTI) data
  - Direction of water molecule ~ Covariates (age, gender, …)
- Transfer learning with UK Biobank data
  - Source data: 500K UK Biobank
  - Target data: A dataset with limited samples in the specific domain
- Latent profile analysis based on five factor model
- Genome-wide association studies (GWAS) for well-being outcome

DNA…?

- Cell → Nucleus → Chromosome → DNA



The central dogma of molecular biology



DNA

Transcription

RNA

Translation

PROTEIN

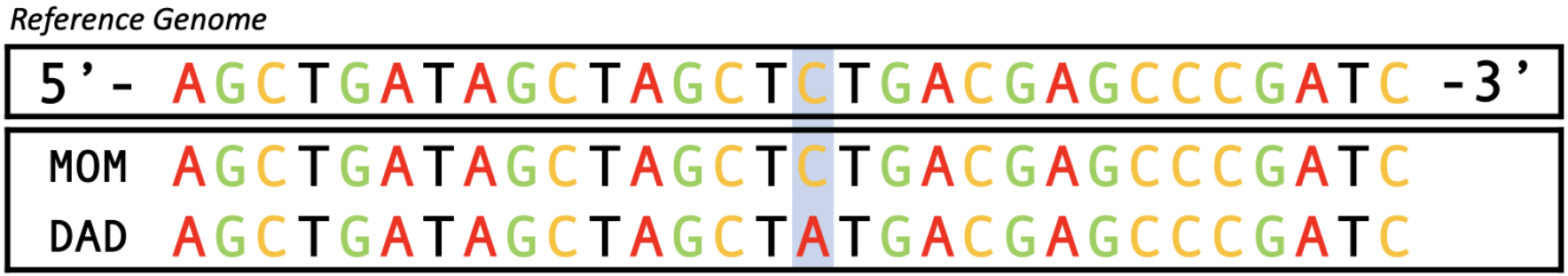
(Phenotype)

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)”

Sigle Nucleotide Polymorphism (SNP)

- Human Genome Project
  - Collected allele data across nearly the entire human genome.



(Homozygote Reference) CC
(Homozygote Alternate) AA
(Heterozygote) AC

} Genotype

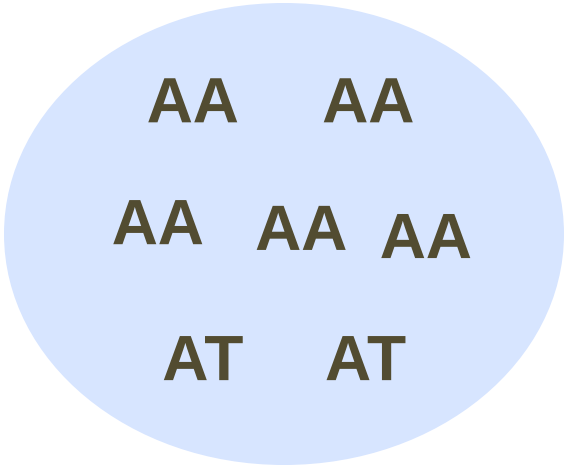
- SNP genotyping
  - At each SNP location, we observe genotype information that reflects the combination of alleles inherited from both parents.

Genome-Wide Association Study (GWAS)

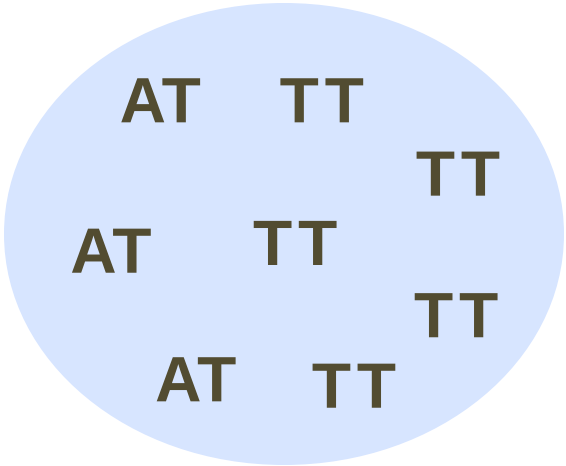
- Identify genetic variants (e.g., SNPs) associated with specific traits or diseases of interest.
  - e.g. cholesterol levels or well-being outcomes

- 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

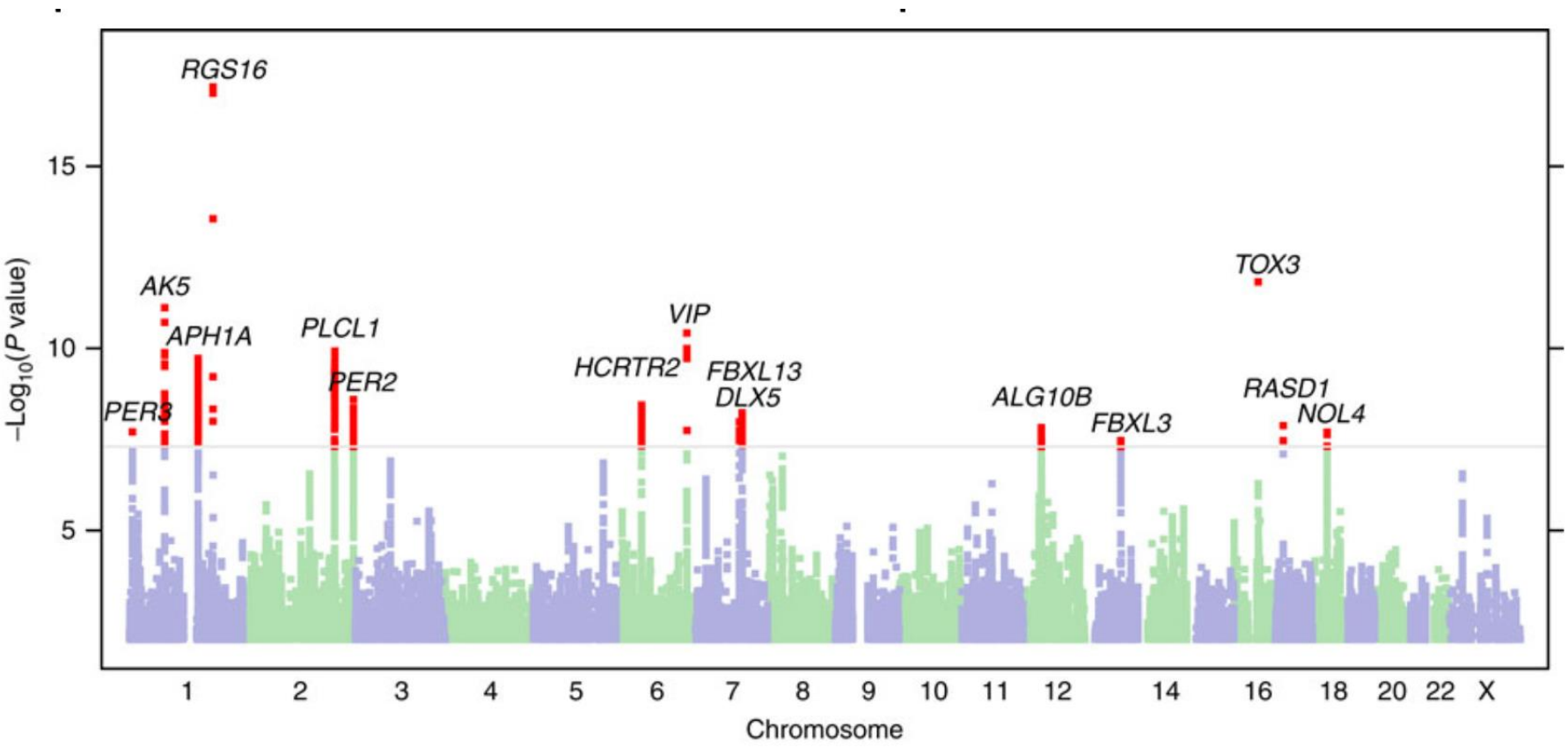
Dependent variable
- Disease
- Phenotypes
- Psychological outcomes
- etc.

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

  Total Bias = Systematic Bias + 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.

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

※ α is a significance level.

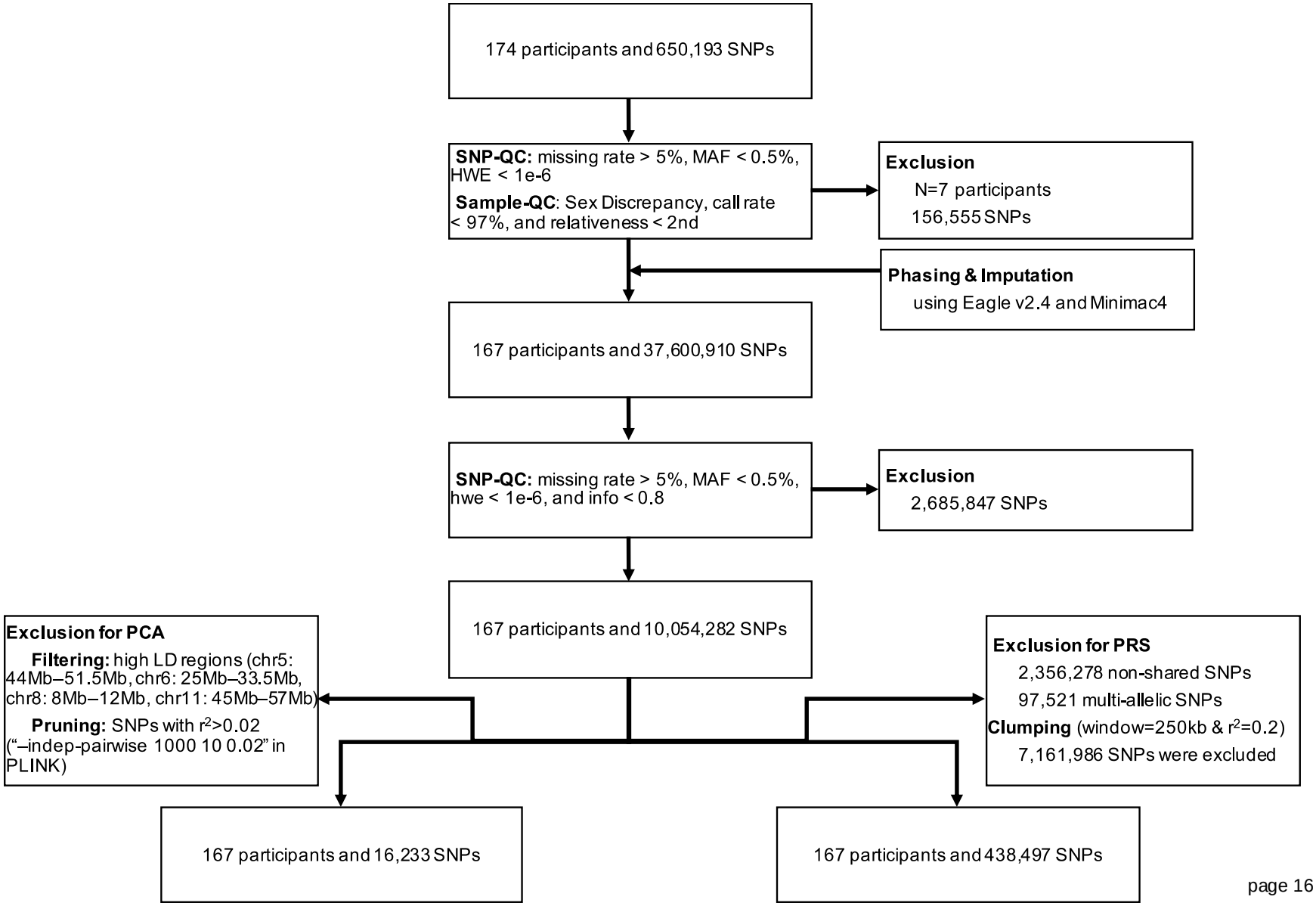
PRS calculation using large-scale dataset

- 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).

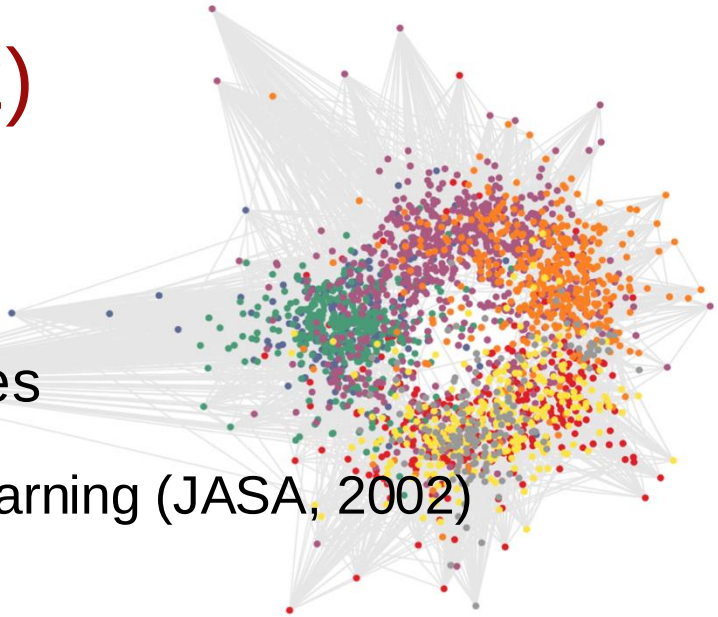
Connection to Transfer Learning

- What is Transfer Learning?
  - A machine learning strategy that utilizes models pre-trained on large datasets to improve performance on a new, related task with limited data.
- Connection to PRS Calculation:
  - Pre-trained GWAS Models:
    - Effect sizes ($\hat{\beta}_j$) are estimated from large-scale GWAS
  - Knowledge Transfer:
    - Transfer learning allows us to adapt these pre-trained effect sizes to target dataset with relevant phenotypes.
- Benefits:
  - Efficient Use of Data
  - Improved Predictive Accuracy:

Future Research Example (1)

- Cross-Population PRS
  - Source dataset: UK Biobank data with 500K individuals
  - Target dataset: KSAH data
- AJHG(2022), PLoS Genet.(2023), Nat.Rev.Genet.(2023), Brief. Bioinfo.(2025)

Future Research Example (2)



- Motivating data: Social network
- GWAS for network-structured responses
  - (1) Statistical network representation learning (JASA, 2002)
  - (2) Variational Graph Auto-Encoders

Latent Space Approaches to Social Network Analysis

Peter D. HOFF, Adrian E. RAFTERY, and Mark S. HANDCOCK

Network models are widely used to represe
emphasis has been placed on random graph m
presence of a specified relation between actors
on the positions of individuals in an unobserv
Bayesian frameworks, and propose Markov c
observed covariates. We present analyses of t
alternative stochastic blockmodeling approach
and interpretable model-based spatial represen
uncertainty in the social space to be quantifie

KEY WORDS: Conditional independence m

Variational Graph Auto-Encoders

Thomas N. Kipf
University of Amsterdam
T.N.Kipf@uva.nl

Max Welling
University of Amsterdam
Canadian Institute for Advanced Research (CIFAR)
M.Welling@uva.nl

Bioinformatics, 38(14), 2022, 3621–3628
https://doi.org/10.1093/bioinformatics/btac369
Advance Access Publication Date: 31 May 2022
Original Paper

OXFORD

Genetics and population analysis

transferGWAS: GWAS of images using deep transfer learning

Matthias Kirchler [1,2,*], Stefan Konigorski [1,3], Matthias Norden[4,5], Christian Meltendorf[6], Marius Kloft[2], Claudia Schurmann[3,4] and Christoph Lippert[1,3,*]

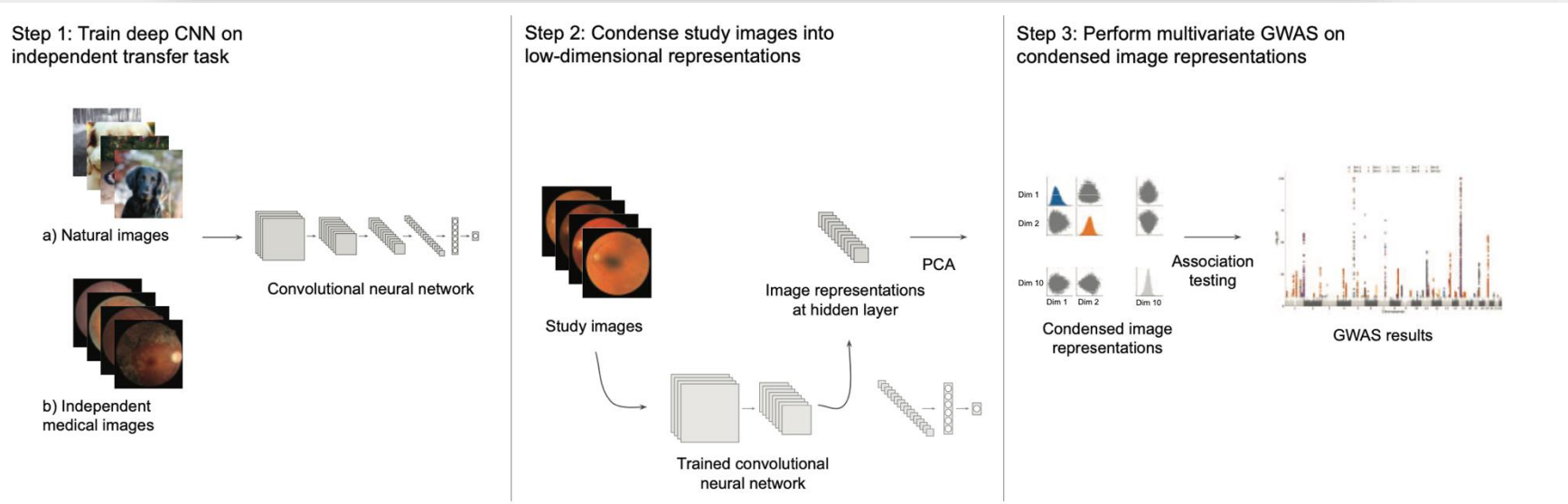


Fig. 1. Overview of the *transferGWAS* approach which consists of three steps. First, a convolutional neural network is trained on an independent transfer task to prime the network. Second, *transferGWAS* uses the trained network and a principal component analysis to condense study images into low-dimensional embeddings. Last, a linear mixed model association analysis is performed on the image representations

Thank you for your attention !