Principal Component Analyses in Genetic Studies

GENOME 560, Spring 2026

Saori Sakaue (sakaue@uw.edu)

Major Topics for Today

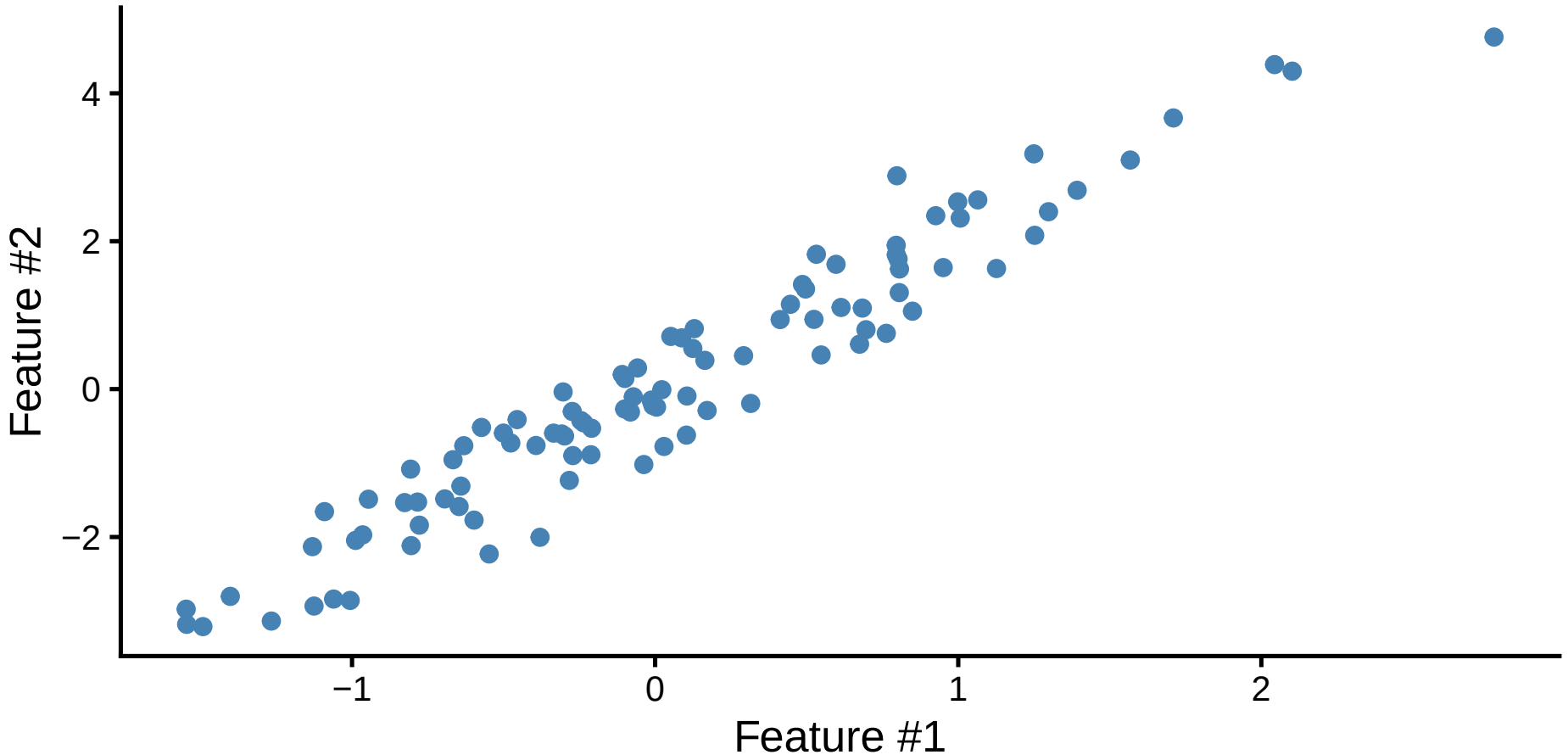
- Recap of last week – on PCA concept and calculation
- Why we need PCA in genetic studies?
- Computational challenges in large-scale genetic data
- Practical examples and tutorial to calculate and use PCA in GWAS

Section 1

Recap: PCA

We Do PCA For

Understanding the data structure and deriving biological (or any) insights.



The Curse of Dimensionality

In high dimensions, all points become approximately equidistant. Distance-based methods lose discriminating power.

"You really shouldn't try to use your 20K gene expression data to compare and characterize your cells!" — without dimensionality reduction

{https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02/_kNN.html}

What Do You Want for Reduced Representation?

Definition of PCA:

Low-dimensional representation of a dataset as a sequence of linear combinations of the variables that have **maximal variance**, and are **mutually uncorrelated** and **ordered in variance**.

Steps to Perform PCA

Given a sample (N) $\times$ features (M) matrix $\mathbf{X}$:

1. **Center** (and optionally normalize) the matrix $\mathbf{X}$

$$\mathbf{X}_c = \mathbf{X} - \mu$$

Steps to Perform PCA (continued)

2. Compute covariance matrix and perform eigen-decomposition:

$$\mathbf{C} = \frac{1}{N-1}\mathbf{X}_c^T\mathbf{X}_c$$

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^T$$

where Λ = diagonal matrix of eigenvalues, $\mathbf{V}$ = eigenvectors.

Steps to Perform PCA (continued)

3. Sort eigenvectors by decreasing eigenvalue: $\lambda_1 \geq \lambda_2 \geq \cdots$
4. Select top S eigenvectors
5. **Project:** $\mathbf{X}_{\text{projected}} = \mathbf{X}_c \cdot \mathbf{V}_S$

Does the First Eigenvector Maximize Variance?

Yes! Maximize $\mathbf{v}^T \mathbf{C} \mathbf{v}$ subject to $\|\mathbf{v}\| = 1$.

By Lagrange multipliers: $\mathbf{C}\mathbf{v} = \lambda \mathbf{v}$

The maximum of $\mathbf{v}^T \mathbf{C} \mathbf{v} = \lambda$ is achieved at the largest eigenvalue.

Section 2

PCA in Genetic Studies

PCA Can Inform Us of Genetic Population Structure

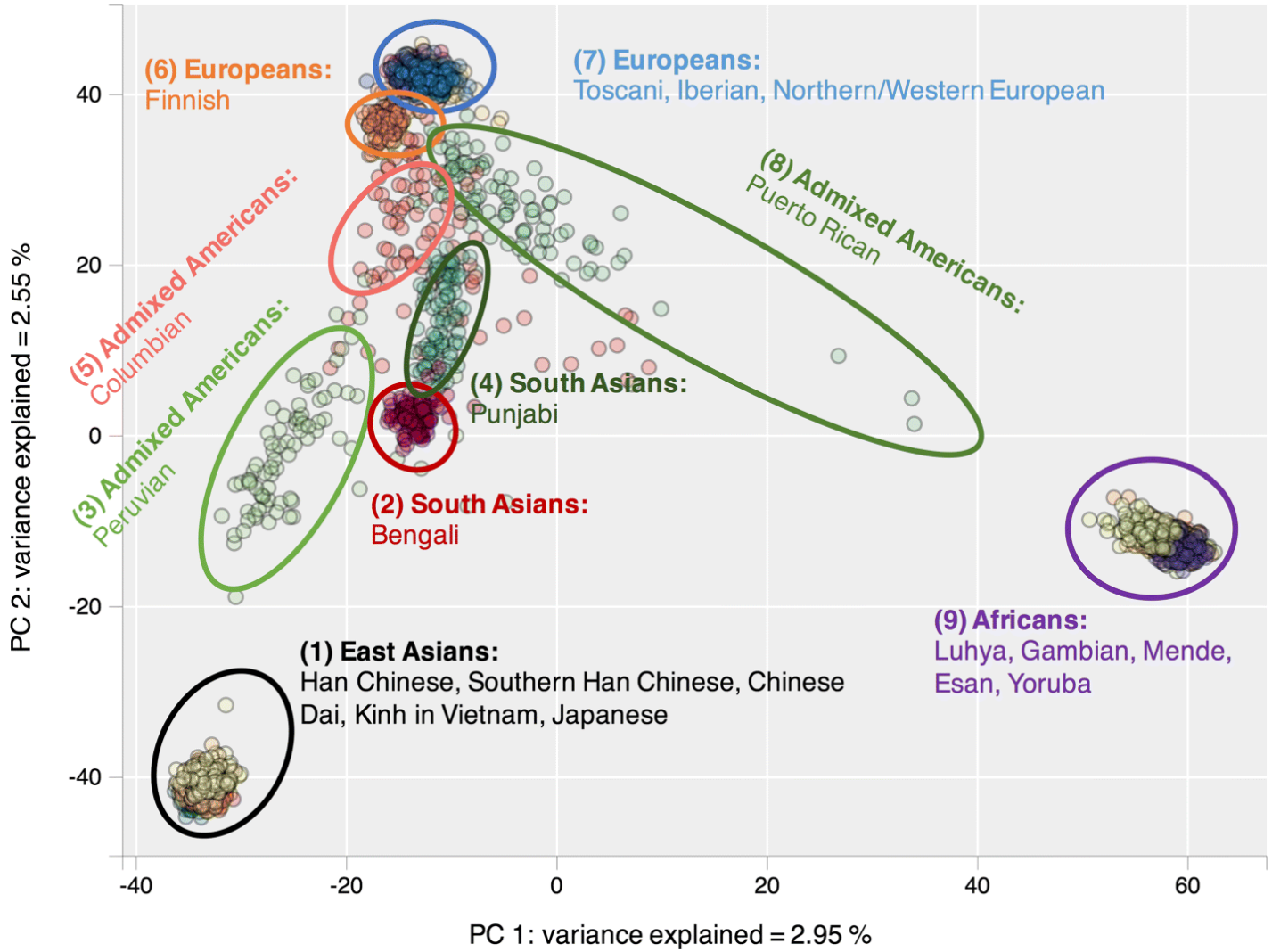


Figure 1: Gasper et al. BMC Bioinformatics 2019

Population Stratifications Could Bias Genetic Associations

Some phenotypes vary with population structure. Representative examples include...

- RA risk, MS risk, height vary along PC1 in European cohorts
- Allele frequencies (e.g., LCT) vary with ancestry

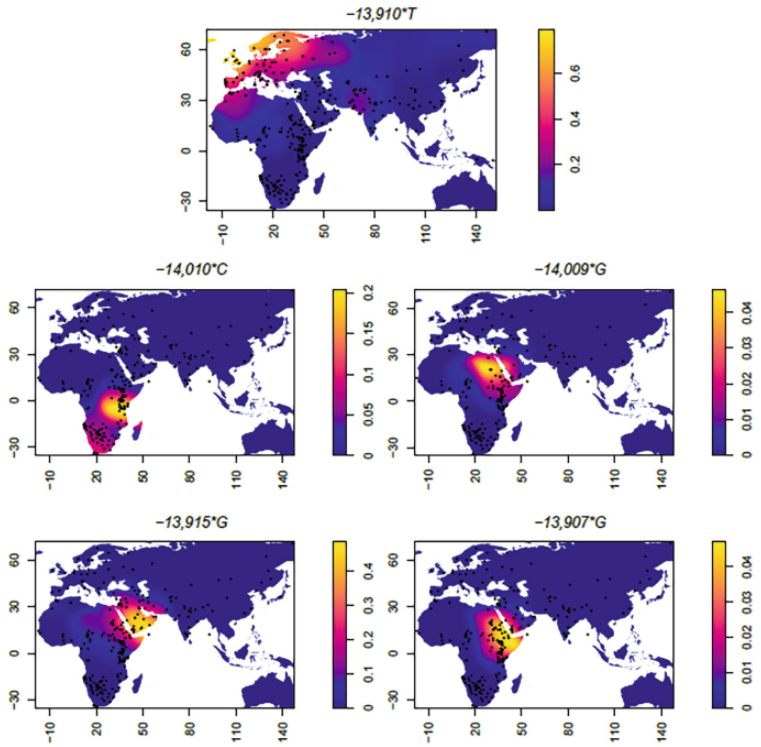


Figure 2: LCT allele frequency distribution across Europe

Historical Ways to Adjust for Systematic Inflation

Genomic control (GC) method:

$$\lambda_{\text{GC}} = \frac{\text{median}(\chi^2_{\text{observed}})}{0.455}$$

where $0.455 = \chi^2_{0.5,1}$ (median of χ^2_1 distribution)

- $\lambda \approx 1$: well-calibrated
- $\lambda \gg 1$: systematic inflation from confounding

Historical Ways to Adjust for Systematic Inflation

Genomic control corrects by dividing all χ^2 statistics by λ_{GC}.

Limitation: assumes uniform inflation across the genome, which may not hold.

PCA to Account for Population Stratification

Price et al. Nature Genetics 2006

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

Figure 3: Price et al. 2006

PCA to Account for Population Stratification

Price et al. Nature Genetics 2006

- Apply PCA to genotype data to infer continuous axes of genetic variation
- For each inferred axis k, use the embeddings as covariates in the association model:

$$\text{pheno} \sim \text{geno} + \text{sex} + \cdots + \text{PC}_1 + \text{PC}_2 + \cdots + \text{PC}_k$$

PCA to Account for Population Stratification

$$Y_i = \beta_0 + \beta_{\text{SNP}} g_i + \beta_{\text{sex}} \text{sex}_i + \sum_{k=1}^{K} \gamma_k \text{PC}_{ik} + \epsilon_i$$

The genotype effect β_{SNP} is now estimated **after controlling for ancestry**.

Genotype Data

Typical whole-genome genotype data in biobanks:

- $M =$ tens of millions of variants (many in LD $\rightarrow$ \$ \$50K independent)
- $N =$ hundreds of thousands or millions of individuals
- Coded as: $a/a = 0$, $A/a = 1$, $A/A = 2$

Section 3

Computational Challenges

Computational Complexity of Eigen-decomposition

- Computing the covariance matrix $\mathbf{C}$: $O(NM \times \min(N, M))$
- Eigen-decomposition of $\mathbf{C}$: $O(M^3)$ in the worst case

For genotype data: $M \gg N$ typically $\rightarrow$ compute $N \times N$ GRM instead of $M \times M$.

Power Iteration to Rapidly Estimate the First Few PCs

Instead of full eigen-decomposition ($O(M^3)$), use **power iteration**:

1. Start with random vector $\mathbf{v}_0$
2. $\mathbf{v}_{t+1} = \mathbf{C}\mathbf{v}_t$
3. Normalize: $\mathbf{v}_{t+1} = \mathbf{v}_{t+1}/\|\mathbf{v}_{t+1}\|$
4. Repeat until convergence

Power Iteration — Why It Works

By repeated matrix multiplication by $\mathbf{C}$:

$$\mathbf{C}^t \mathbf{v}_0 = \sum_i c_i \lambda_i^t \mathbf{u}_i$$

Because $\lambda_1 > \lambda_2 > \cdots$, the dominant eigenvector $\mathbf{u}_1$ dominates as $t \to \infty$.

Convergence rate: $\sim (\lambda_2/\lambda_1)^t$

Power of Power Iteration

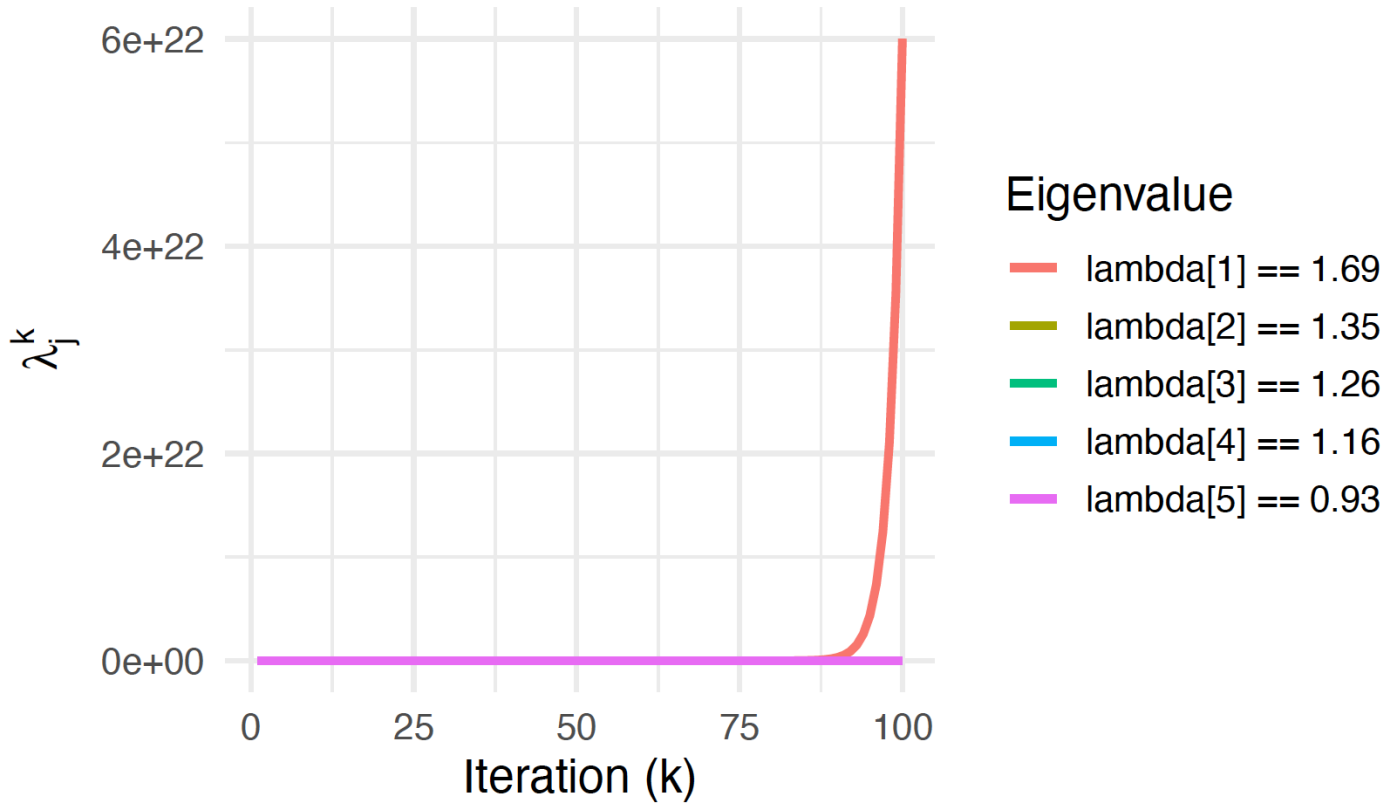


Figure 4: Convergence of power iteration

Power Iteration in Practice

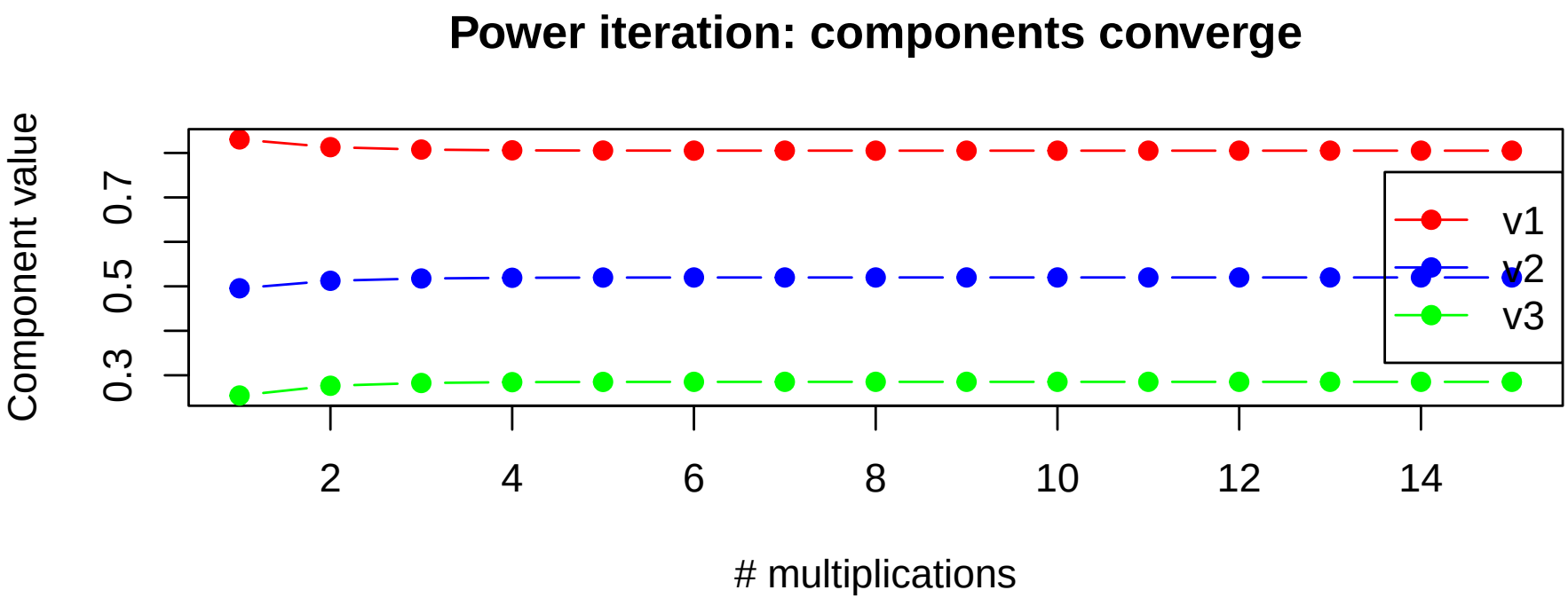
Start with any random vector $\mathbf{v}$ of appropriate dimension.

$$\mathbf{v}_0 = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{(any random initialization)}$$

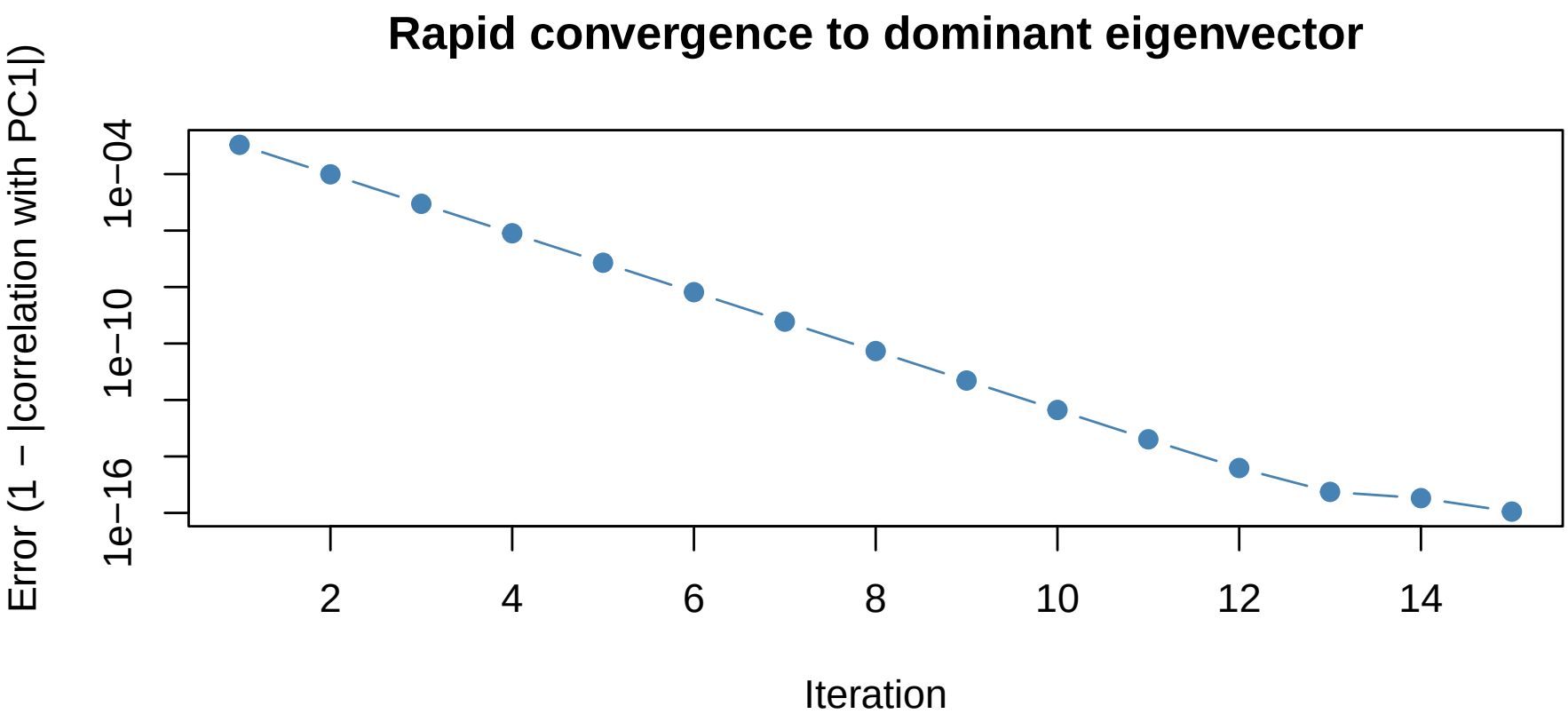
Power Iteration in Practice (continued)

Repeatedly multiply by $\mathbf{C}$ and normalize:

$$\mathbf{v}_{t+1} = \frac{\mathbf{C}\mathbf{v}_t}{\|\mathbf{C}\mathbf{v}_t\|}$$



Power Iteration in Practice (convergence)



Section 4

When PCA May Not Work

When PCA May Not Work

Non-orthogonal structure. The constraint that PCs are orthogonal may make higher PCs difficult to interpret — the 5th component is FORCED to be orthogonal to the top 4, which is rather artificial.

Non-linear structure. PCA finds linear structure. If data has non-linear low-dimensional structure (e.g., $y \sim x^2$), PCA will not find this.

Other Methods for Dimensionality Reduction

Other linear / matrix factorization methods:

- Non-negative matrix factorization (NMF), ICA, CCA, ...

Non-linear dimensionality reduction:

- Isomap, t-SNE, UMAP, ...
- Autoencoder, Variational autoencoder (VAE)

Section 5

Practical PCA with Genotype Data

Performing PCA with Genotype Input

Typical genotype data in the format of plink *.{bim, fam, bed} files:

- **bim:** SNP information (chromosome, position, alleles)
- **fam:** sample information (family ID, individual ID, sex, phenotype)
- **bed:** binary genotype matrix

Performing PCA with Genotype Input — QC

QC your genotype data before doing anything!!

1. Remove low quality samples
2. Remove low quality genotypes
3. Select only common alleles (e.g., MAF $> 5\%$)
4. Exclude HLA region and other long-range LD regions

Long-range LD Regions Can Confound PCA

Especially in admixed populations:

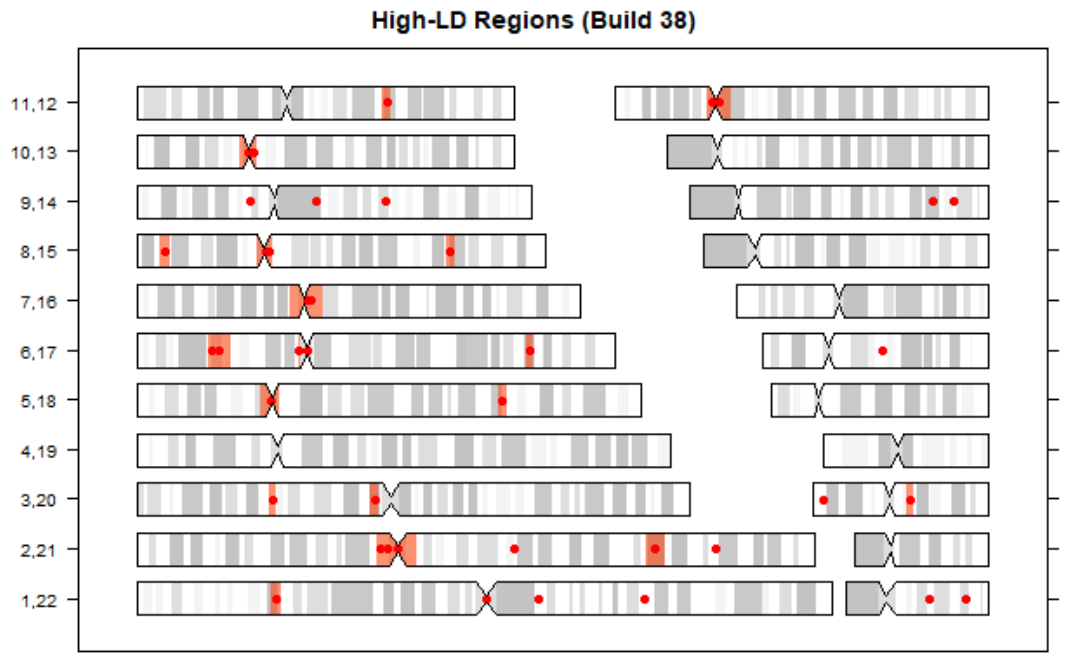
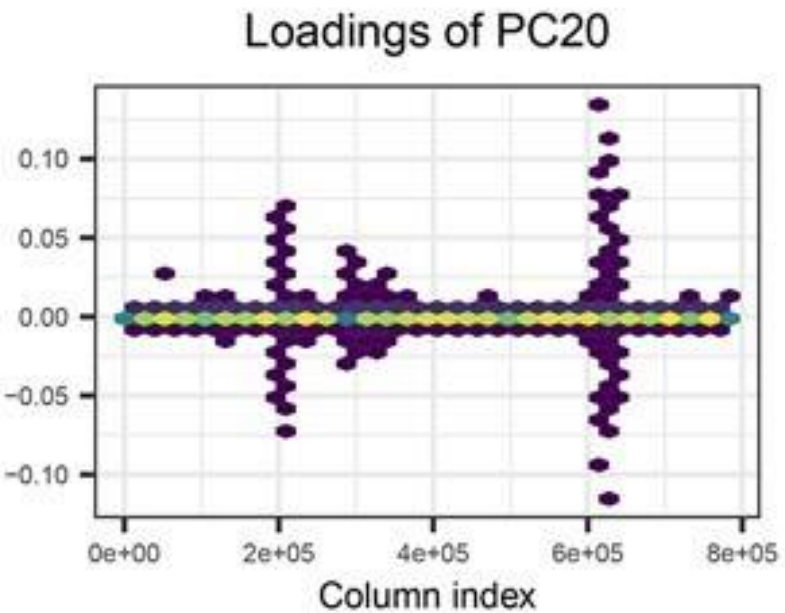
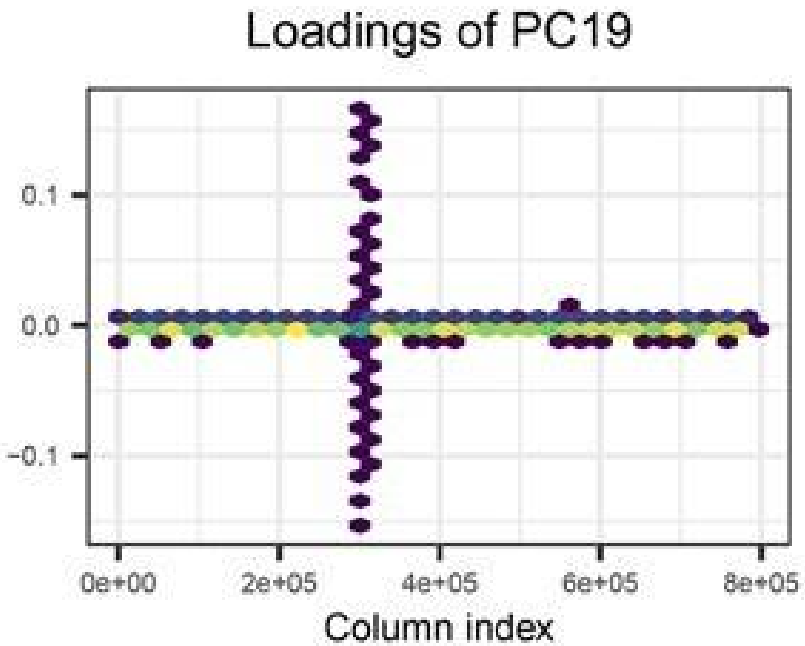


Figure 5: Galinsky et al. 2016 Am J Hum Genet

...Which Can Be Reflected on PCs





Performing PCA with Genotype Input — LD Pruning

- Perform **LD pruning** to obtain LD-independent SNPs (to select independent markers)
- Perform PCA on LD-pruned matrix
  - Genotype will be **centered and normalized**
  - You can specify how many PCs you'd like to obtain (faster for power iteration)

PCA Output

Output: samples $\times$ PCs matrix

samples	PC1	PC2	PC3	PC4
sample1	0.023	-0.041	0.005	0.012
sample2	0.019	-0.038	-0.002	0.008
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Top PCs Capture Global Genetic Ancestries

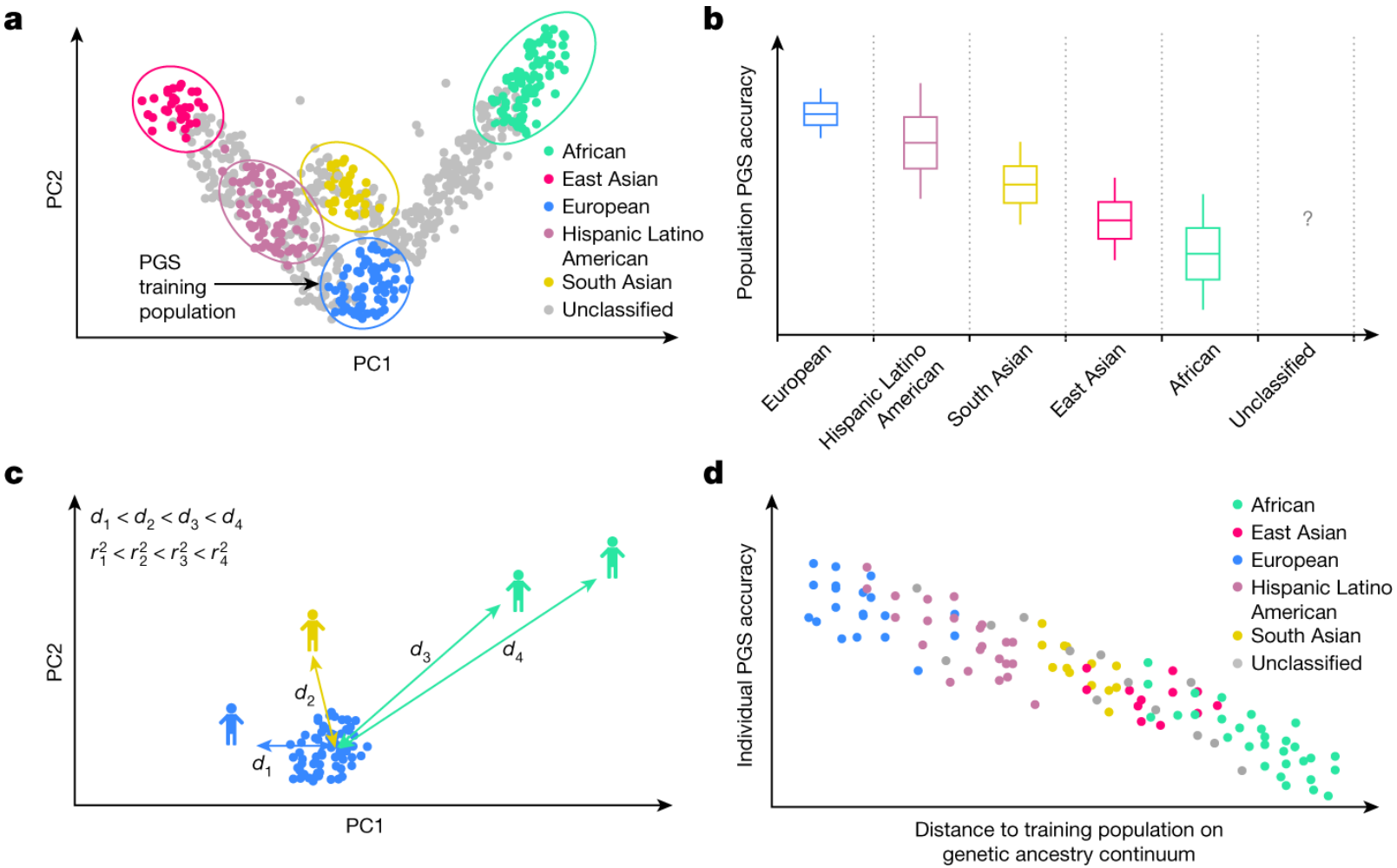


Figure 6: Ding et al. Nature 2023

Genetic Risk Prediction Accuracy Varies Across Ancestry

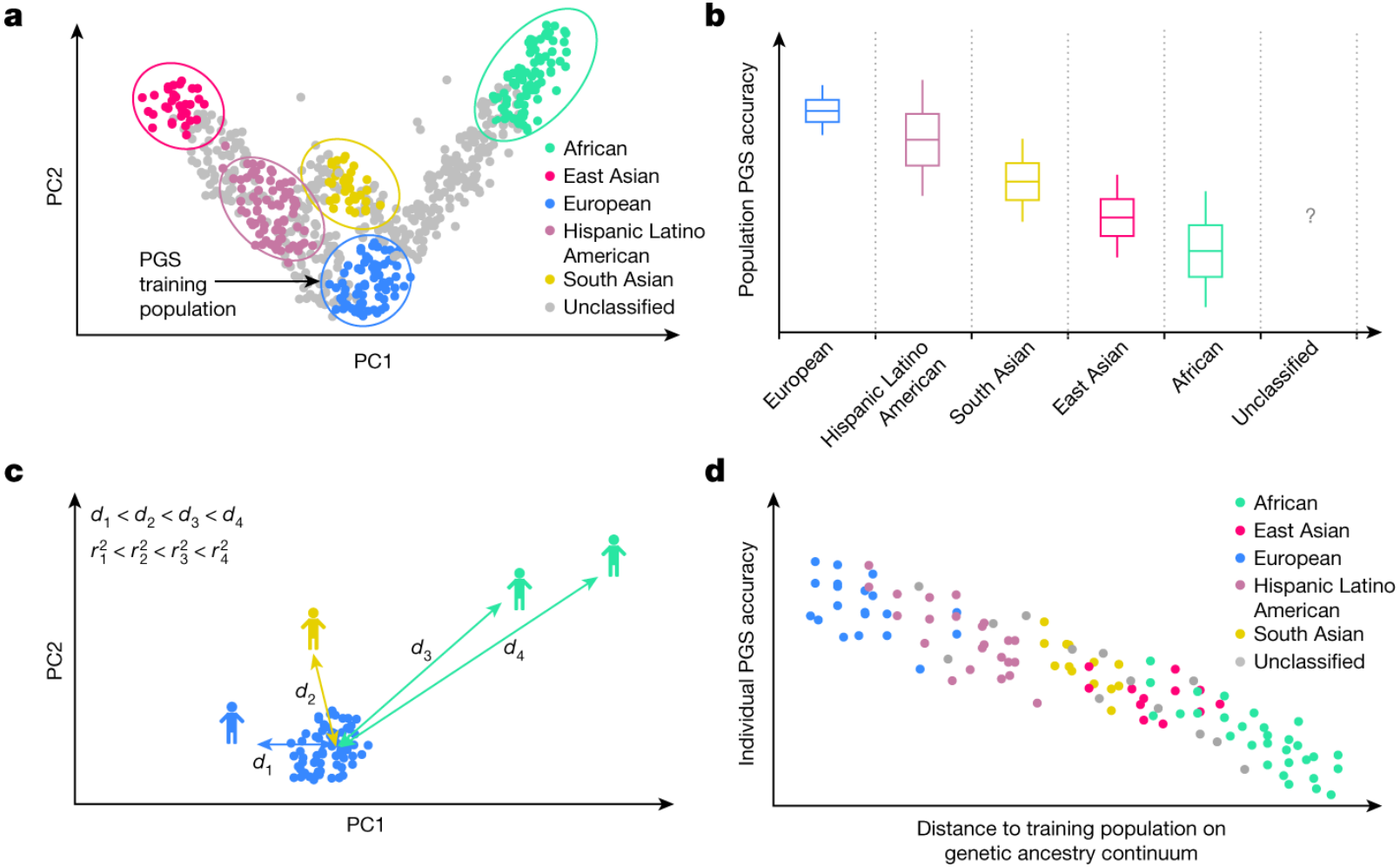
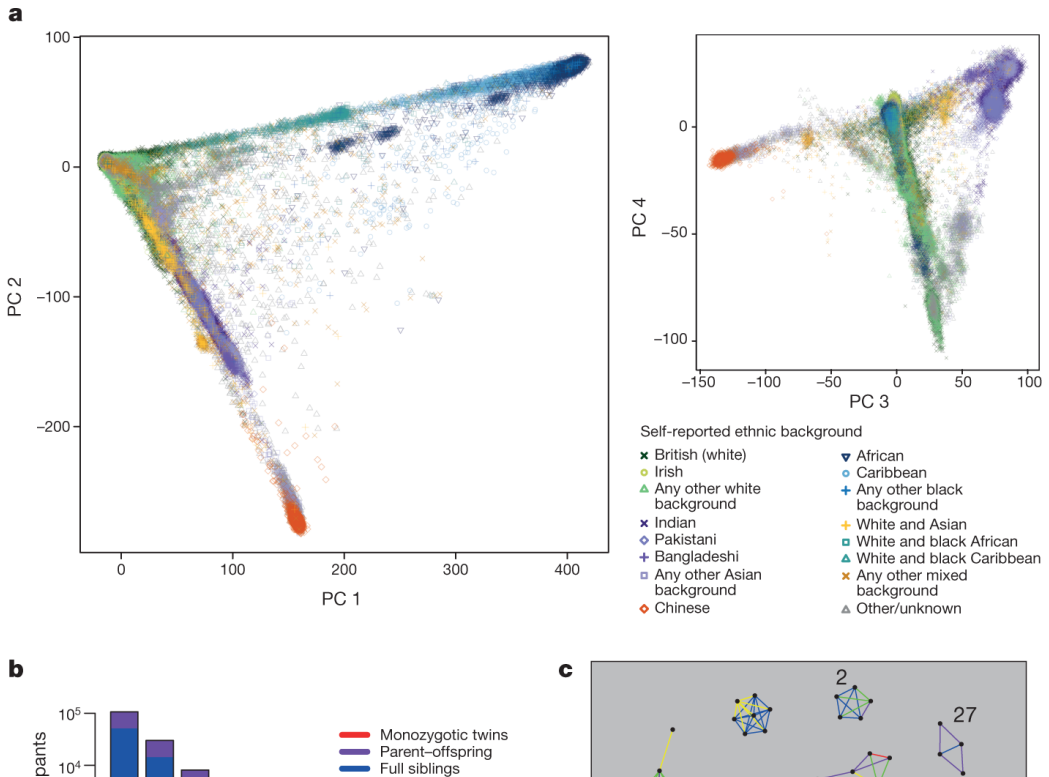


Figure 7: Ding et al. Nature 2023

PCA May Not Fully Account for Fine Population Structure

Extremely large datasets such as UK Biobank ($n \approx 500K$) have subtle population structure and cryptic relatedness:

"A total of 147,731 UK Biobank participants (30.3%) are inferred to be related (third degree or closer) to at least one other person in the cohort"
— Bycroft et al. Nature 2018



Handling Fine Structure and Relatedness

- **Straightforward:** Remove all potentially related individuals $\rightarrow$ decrease in sample size
- **New approach:** Mixed model association — accounts for both population stratification and cryptic relatedness

Section 6

Mixed Model Association

Linear Model Association (Fixed Effects)

$$\mathbf{Y} = \mathbf{X}\beta + g_j\beta_j + \epsilon$$

$$\epsilon \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

where $\mathbf{X}$ = covariates (age, sex, PCs), g_j = SNP genotype.

Mixed Model Association

$$\mathbf{Y} = \mathbf{X}\beta + g_j\beta_j + \mathbf{u} + \epsilon$$

- **Fixed effects:** $\mathbf{X}\beta + g_j\beta_j$
- **Random effects:** $\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{K})$

where $\mathbf{K}$ = genetic relationship matrix (GRM), capturing genome-wide relatedness.

Mixed Model Association (continued)

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{K})$$

The GRM $\mathbf{K}$ encodes:

- **Fine population structure** (continuous ancestry gradients)
- **Cryptic relatedness** (close relatives, e.g., sib-pairs)
- Both are modeled simultaneously

Mixed Model Accounts for Fine Structure and Relatedness

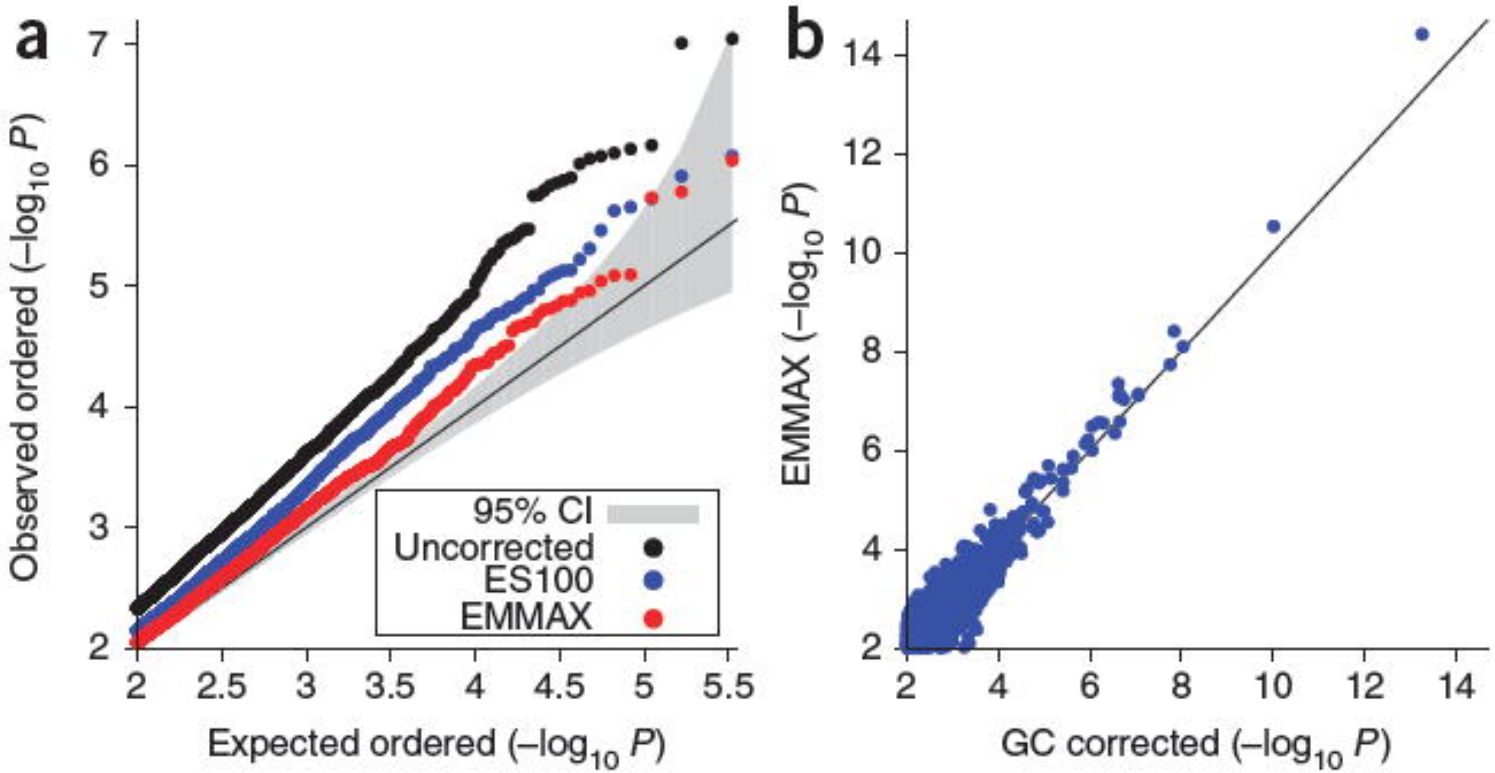


Figure 9: Kang et al. Nature Genetics 2011

Left: QQ plot showing uncorrected (black), PCA-corrected (blue), and mixed model (red). Right: Mixed model vs. GC-corrected p-values.