National Cancer Institute

Population Substructure and Control Selection in Genome-wide Association Studies

Kai Yu, Ph.D.

Division of Cancer Epidemiology and Genetics, NCI

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Acknowledgements

CGEMS & DCEG

Gilles Thomas
Zhaoming Wang
Stephen Chanock
Sholom Wacholder
Qizhai Li
Robert Hoover
Kevin Jacobs
Meredith Yeager
Joseph Fraumeni
Daniela Gerhard
Xiang Deng
Nick Orr
Robert Welch
Nilanjan Chatterjee
Richard Hayes
Margaret Tucker
Marianne Rivera-Silva

HSPH

David Hunter
Peter Kraft
David Cox
Sue Hankinson

ACS

Michael Thun
Heather Feigelson
Eugenia Calle

CeRePP, France

Olivier Cussenot
Geraldine Cancel-Tassin
Antoine Valeri

NPHI, Finland

Jarmo Virtamo

Wash. U., St Louis

Gerald Andriole

Background

- Genome wide association studies (GWAS) based on case-control design
  - Compare genotype frequency at each genetic markers (SNP)
- Population stratification (PS)
  - Genotype frequency differences at a given SNP between cases and controls due to ancestry differences (confounding by ethnicity).

PS example: *LCT* and height (Campbell et al., 2005)

		Matching on four grandparents Ancestry		
	All	Four US-born	Southeastern Euro	Northwestern Euro
Tall	161:474:489	66:265:314	54:55:18	41:154:157
Short	231:444:380	76:278:282	128:86:13	27:79:86
P-value	3.6 x 10-7	0.098	0.0016	0.71

Note: after adjustment for the three classes, the P-value is 0.0074

More on PS

- PS can occur in a case-control study conducted in a non-homogeneous population
  - Due to disease risk heterogeneity across (hidden) subpopulations
  - Due to sampling bias that results into ancestry background difference between cases and controls

Motivation

- Longstanding debate on the impact of PS on well-designed genetic studies
- The temptation to use a external controls to save costs (using controls from another study, using shared controls)

Focus of This Talk

Using empirical data from CGMES

- Evaluate the impact of PS in GWAS conducted in ***European Americans*** with different sample selection strategies
  - Nested case-control design
  - The use of external controls

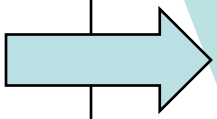
- How to effectively correct for PS

Identifying Genetic Markers for Prostate & Breast Cancer

CGEMS
Cancer Genetic Markers of Susceptibility

Genome-Wide Analysis
Public Health Problem
- **Prostate (1 in 8 Men)**
- **Breast (1 in 9 Women)**

Analyze Long-Term Studies
- **NCI PLCO Study**
- **Nurses' Health Study (NHS)**

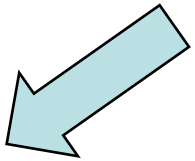


Initial Study

Follow-up #1

Follow-up #2

Establish Loci



Fine Mapping
Functional Studies
Validate Plausible Variants
Possible Clinical Testing

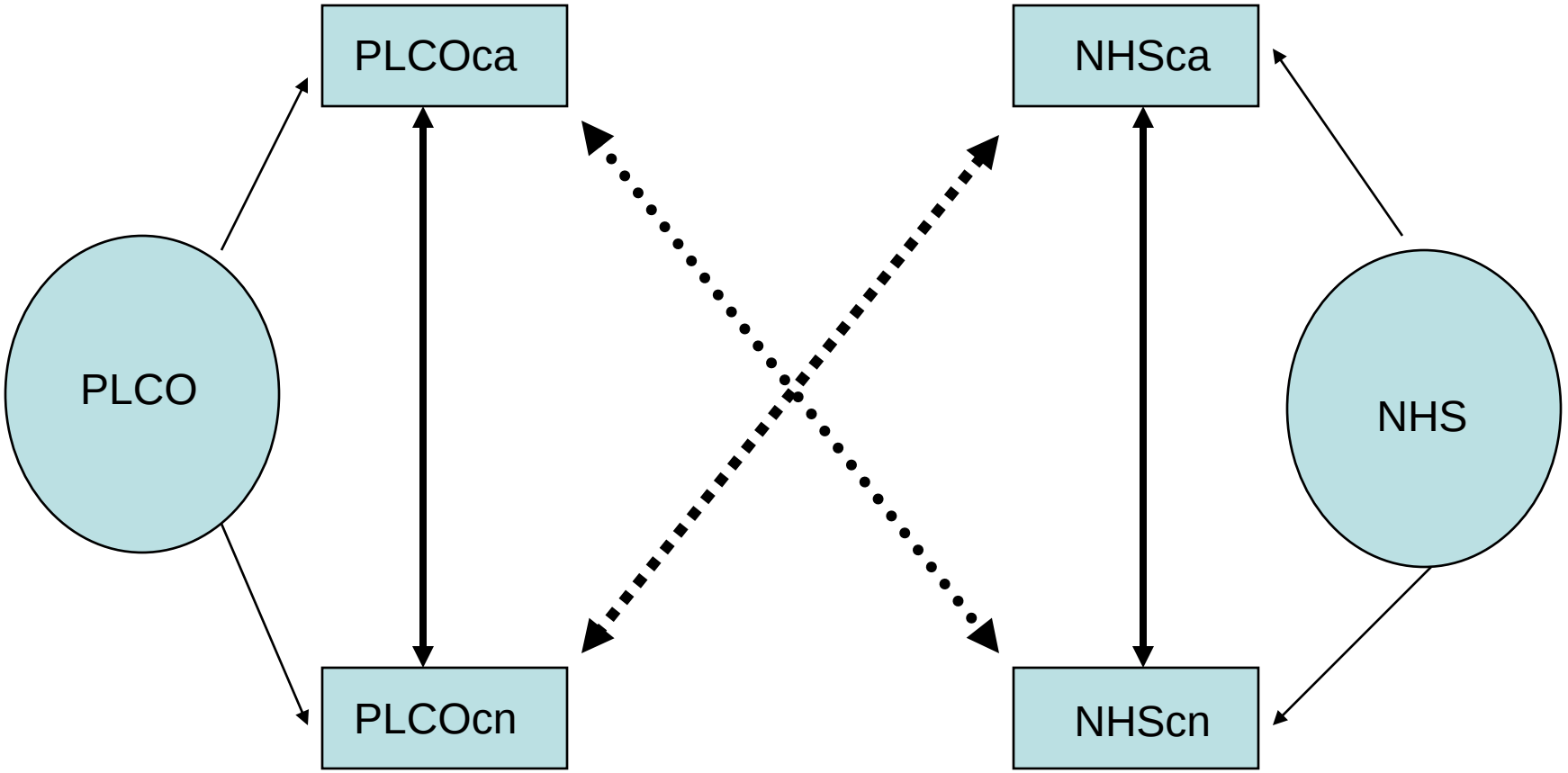
http://cgems.cancer.gov

Material for Analysis

- PLCO (Prostate, Lung, Colorectal and Ovarian cancer screening trial)
  - Men from a randomized trial for cancer prevention
  - Removing subjects with European admixture coefficient <90%
  - 1,171 prostate cancer cases
  - 1,094 controls

- NHS (Nurses' Health Study)
  - Women from a prospective cohort study on nurses
  - Removing subjects with European admixture coefficient <90%
  - 1,140 breast cancer cases
  - 1,138 controls

- # testing autosomal SNP: 450K
  - >5% minor allele frequency in PLCO and in NHS
  - <5% missing rate in PLCO and in NHS

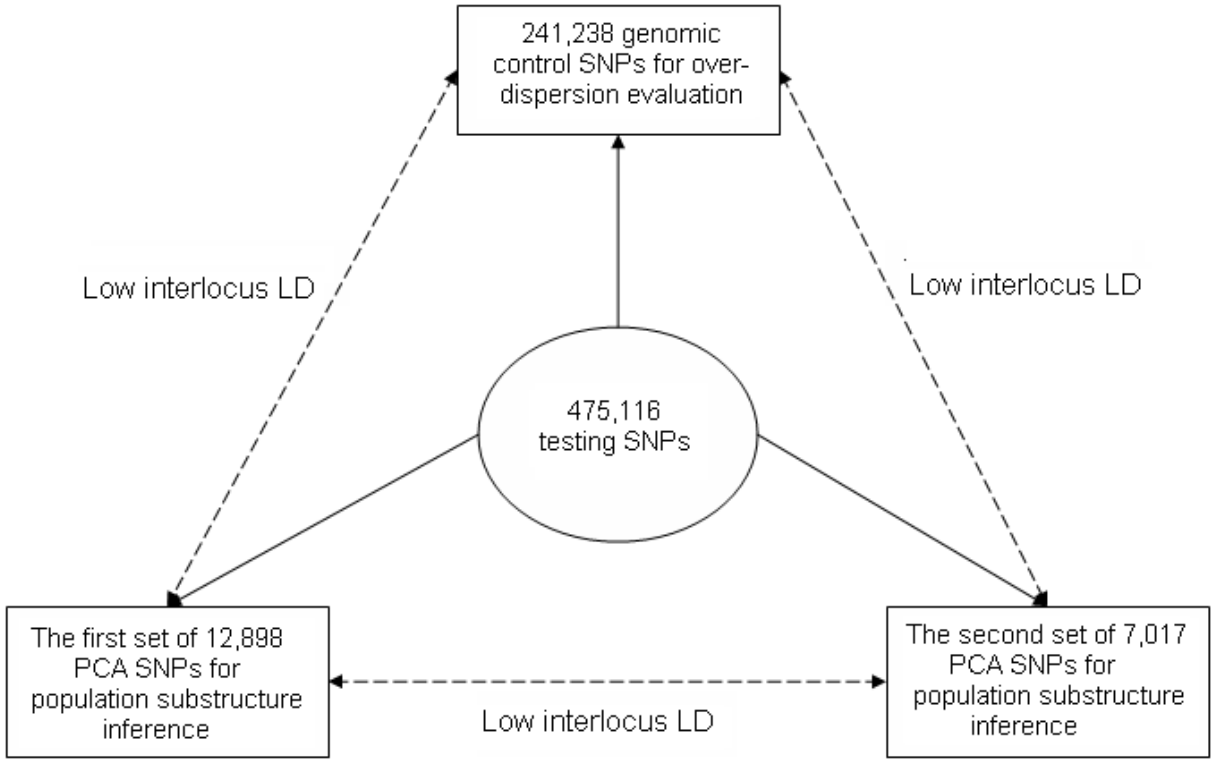

PLCOca
NHSca
PLCO
NHS
PLCOcn
NHScn

Null markers are useful

Because of the availability of many ***null***

SNPs in GWAS

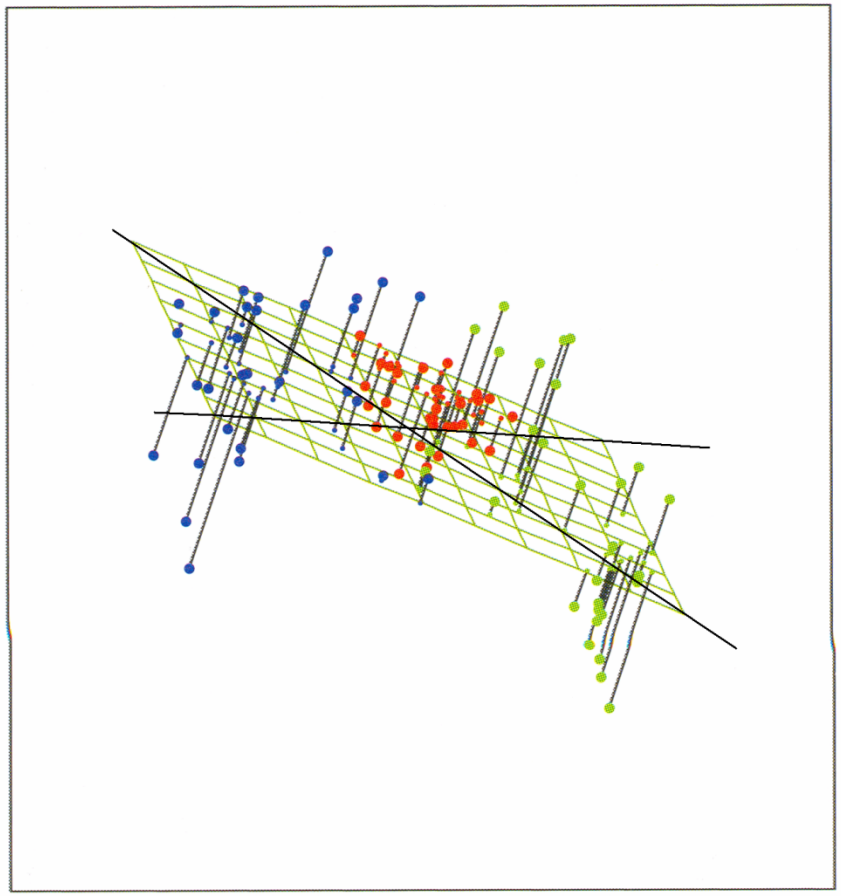
- Monitor extent of PS
  - Q-Q plot, inflation factor
- Estimate the population ancestry and correct for PS (at the cost of power)
  - PCA: capture correlation between genotypes to identify axes with large genetic variation
  - STRUCTURE: Attempts to interpret the correlation between genotypes in terms of admixture among a defined number of ancestral populations

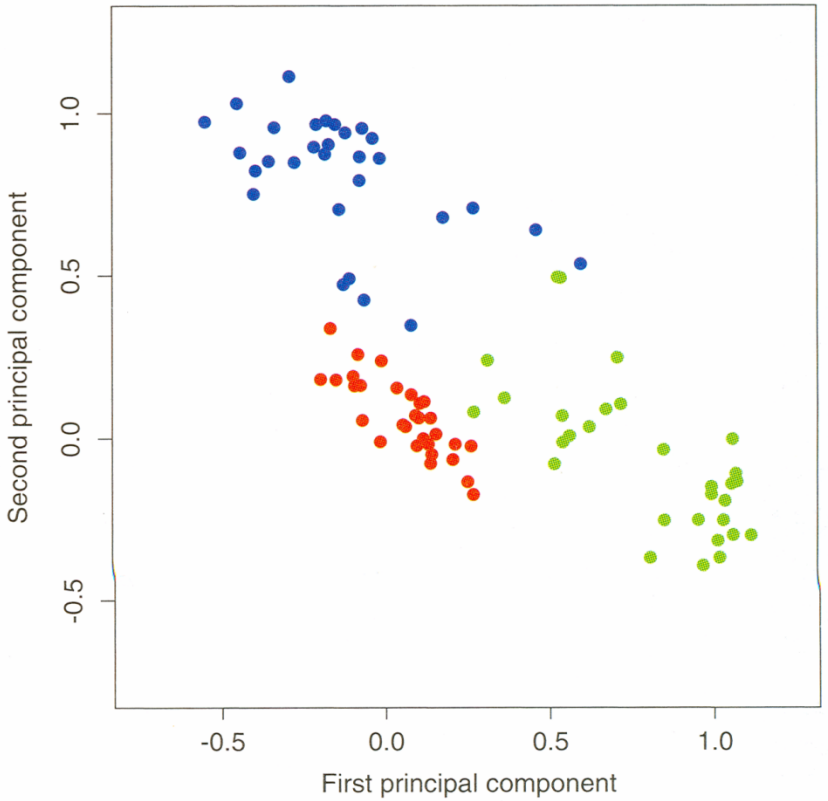

241,238 genomic control SNPs for over-dispersion evaluation
Low interlocus LD
Low interlocus LD
475,116 testing SNPs
The first set of 12,898 PCA SNPs for population substructure inference
Low interlocus LD
The second set of 7,017 PCA SNPs for population substructure inference

Using PCA to study population substructure

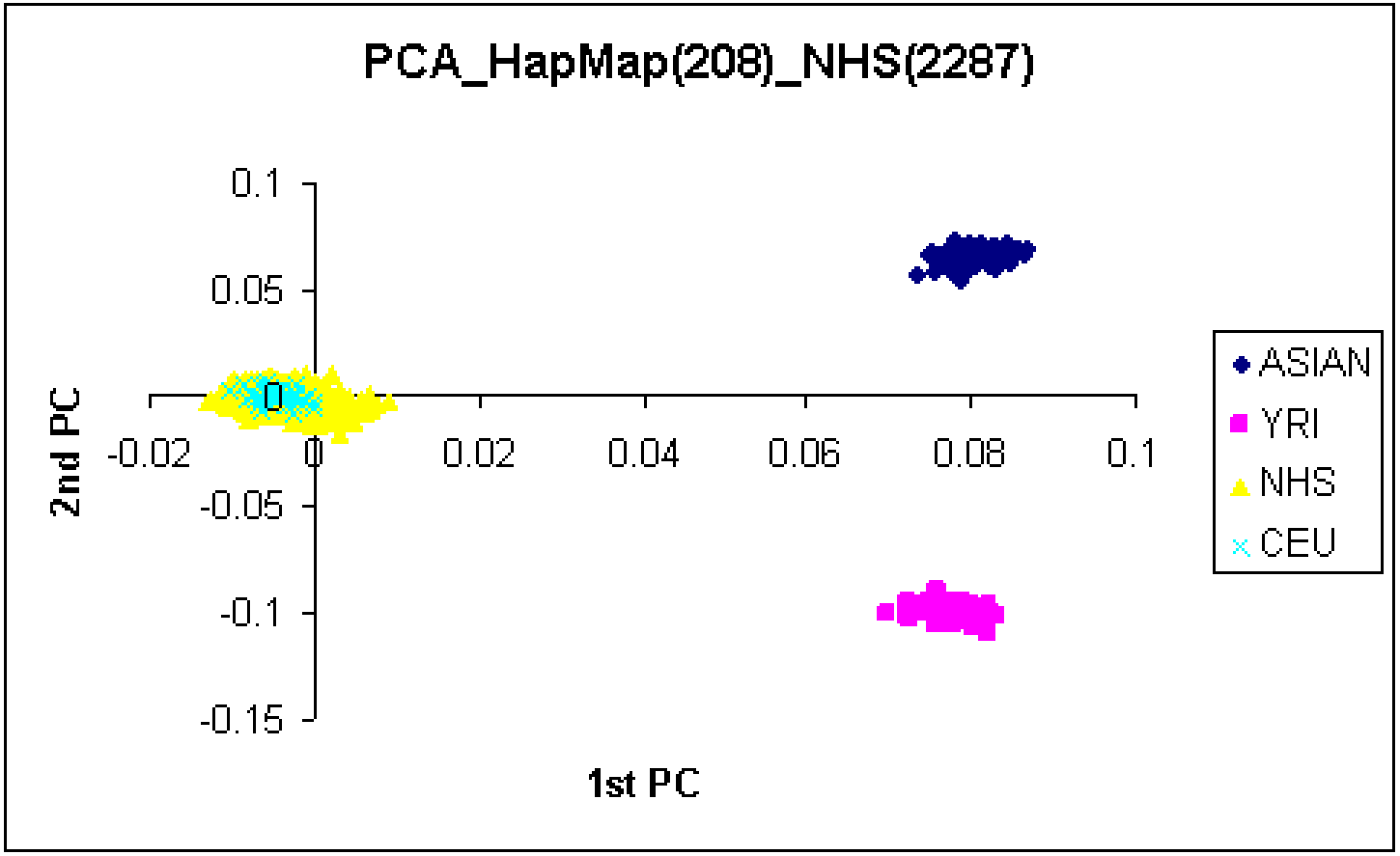
Summarize the information measured on *N* structure inference SNPs and represents study participants in a lower dimensional space so that the Euclidean distance between two subjects represents their genetic difference.

An Illustration for PCA

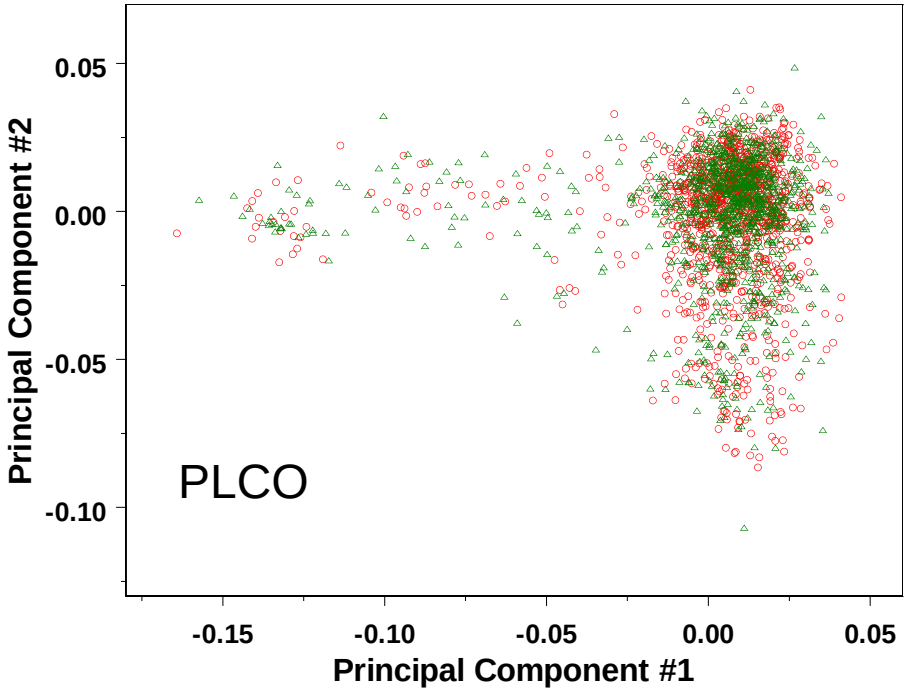


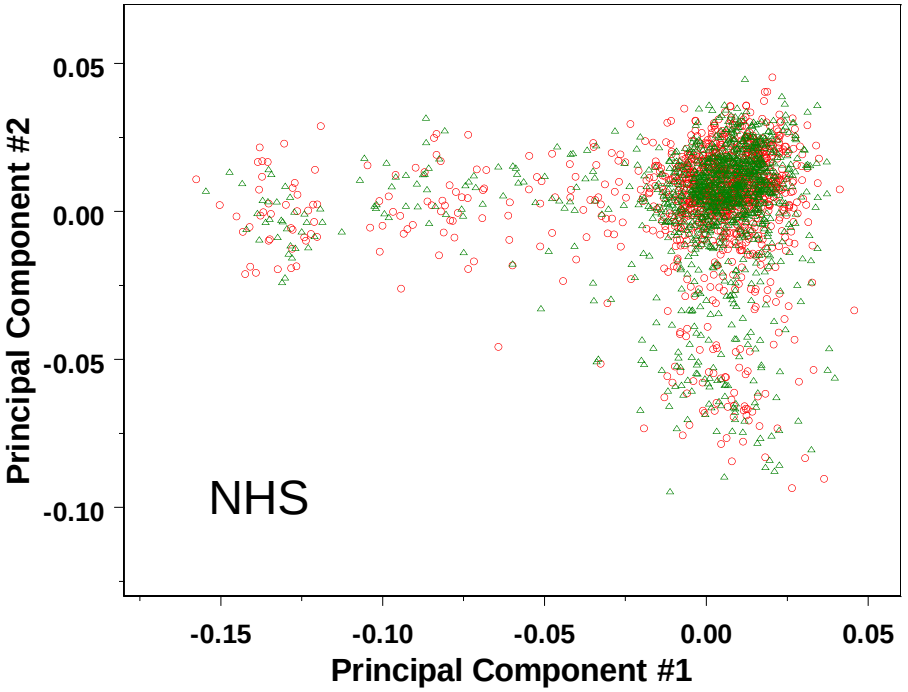


PCA of joint sample of HapMap and NHS



PCA in CGEMS PLCO and NHS GWAS





Principal component comparisons (P-values) between cases and controls based on the Wilcoxon rank-sum test

	PLCOca-PLCOco	PLCOca-NHSco	NHSca-NHSco	NHSca-PLCOco
PC #1	0.294	4.5×10^{-8}	0.664	4.3×10^{-6}
PC #2	0.871	2.2×10^{-7}	0.289	6.9×10^{-12}
PC #3	0.340	0.282	0.036	4.0×10^{-3}
PC #4	0.588	1.2×10^{-4}	0.015	0.191
PC #5	0.490	0.385	0.943	0.157

Observations I

- Similar population sub-structure patterns in GWAS conducted in PLCO and NHS
  - The exchange of controls may be feasible
- Demonstrable genetic background difference between the two GWAS, partially due to
  - Difference in geographic locations of the two source populations

Inflation factor (IF)

Let $T_i,\ i=1,\ \ldots,\ M,$ be the association test statistics for all testing markers, the inflaction factor λ can be defined as

$$\lambda = \frac{\text{median}\{T_i,\ i=1,\ \ldots,\ M\}}{\text{median of } T \text{ under the null}}.$$

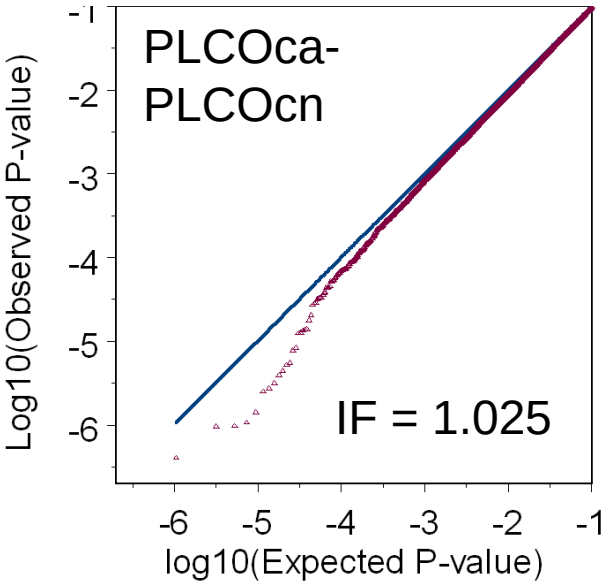
For example, if using 1 df Chi-squared test,

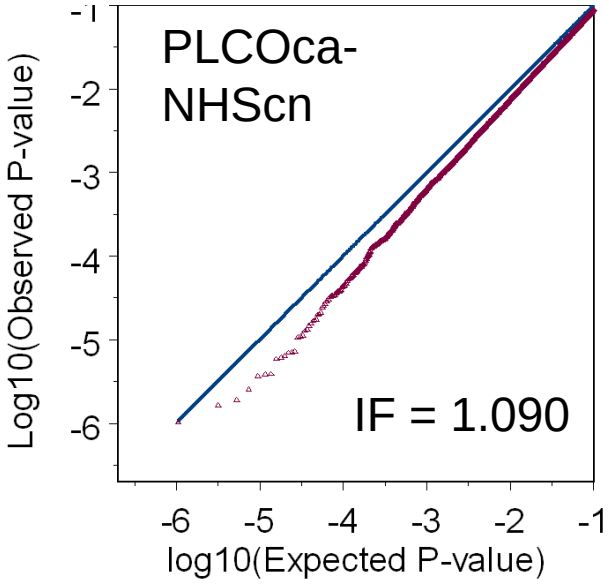
$$\lambda = \frac{\text{median}\{T_i,\ i=1,\ \ldots,\ M\}}{0.455}.$$

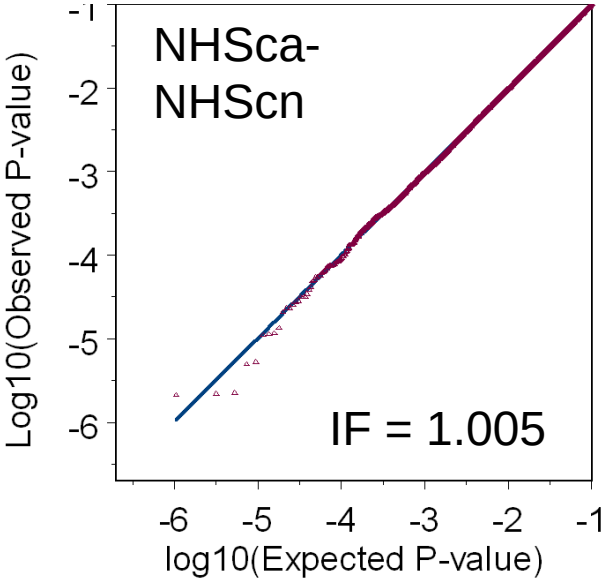
If using 2 df Chi-squared test,

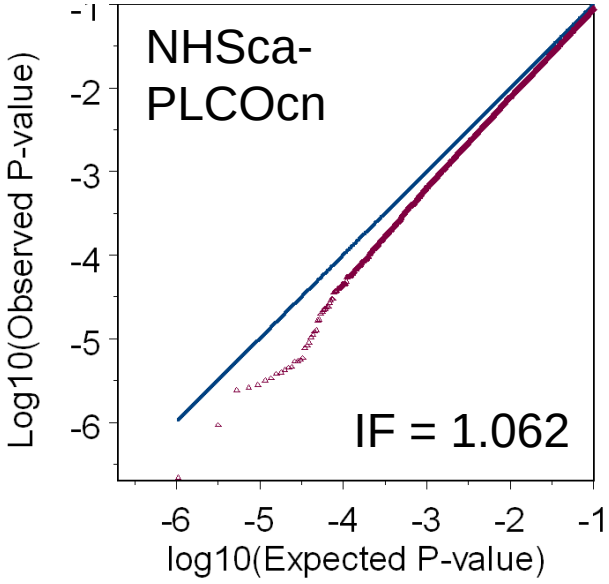
$$\lambda = \frac{\text{median}\{T_i,\ i=1,\ \ldots,\ M\}}{1.386}.$$

Q-Q Plot for the test without PC adjustment









PC selection strategies for the correction of PS

$$\log \frac{p}{1-p} = \alpha + \mathbf{u}_1\beta_1 + \mathbf{u}_2\beta_2 + \mathbf{g}\gamma$$

- Select a fixed number of PCs (e.g., top 10 PCs)
- Select PCs with “large” genetic variations (e.g., PCs with Tracy-Widom test P-value < 0.05)
- Select PCs correlated with the outcome

A Algorithm to Select PCs for PS correction

Order the top L PCs according to their Wilcoxon rank-sum statistics, define them as $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_L$. First, to evaluate whether to include $\mathbf{u}_1$ for the PS adjustment.

1. Obtain the inflation factor λ_1 based on

$$\log \frac{p}{1-p} = \alpha + \mathbf{u}_1 \beta_1 + \mathbf{g}\gamma .$$

2. Random permute $\mathbf{u}_1$ to $\mathbf{u}_1^{(b)}$, $b = 1, \ldots, B$, Based on $\mathbf{u}_1^{(b)}$, obtain $\lambda_1^{(b)}$ from

$$\log \frac{p}{1-p} = \alpha + \mathbf{u}_1^{(b)} \beta_1 + \mathbf{g}\gamma . \ \lambda_1^{(b)}, \ b = 1, \ \ldots, \ B .$$

3. Estimate the empirical P-value as $p = \#\left\{\lambda_1 > \lambda_1^{(b)}, \ b = 1, \ \ldots, \ B\right\} / B$.

4. Include $\mathbf{u}_1$ if p is small (say <0.05), not include $\mathbf{u}_1$ otherwise.

Algorithm (cont.)

Suppose we have chosen $\mathbf{u}_2$ and $\mathbf{u}_3$, and try to decide whether to include $\mathbf{u}_5$, do

1. Obtain the inflation factor λ_5 based on

$$\log \frac{p}{1-p} = \alpha + \mathbf{u}_2\beta_2 + \mathbf{u}_3\beta_3 + \mathbf{u}_5\beta_5 + \mathbf{g}\gamma .$$

2. Random permute $\mathbf{u}_5$ B times to have $\mathbf{u}_5^{(b)}$, $b = 1, ..., B$. Based on $\mathbf{u}_5^{(b)}$, obtain $\lambda_5^{(b)}$ from

$$\log \frac{p}{1-p} = \alpha + \mathbf{u}_2\beta_2 + \mathbf{u}_3\beta_3 + \mathbf{u}_5^{(b)}\beta_5 + \mathbf{g}\gamma . \; b = 1, \; ..., \; B.$$

3. Estimate the empirical P-value as $p = \#\left\{\lambda_5 > \lambda_5^{(b)}, \; b = 1, \; ..., \; B\right\} / B$

4. Include $\mathbf{u}_5$ if p is small (say <0.05), not include $\mathbf{u}_5$ otherwise.

PCs selected

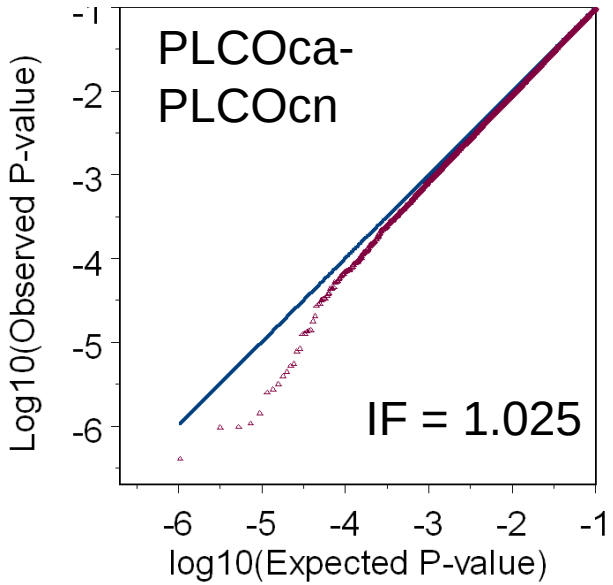
	PLCOca-PLCOco	*PLCOca-NHSco*	*NHSca-NHSco*	*NHSca-PLCOco*
PC #1	Yes	Yes	No	Yes
PC #2	No	Yes	Yes	Yes
PC #3	No	No	No	Yes
PC #4	No	Yes	No	No
PC #5	No	No	No	No

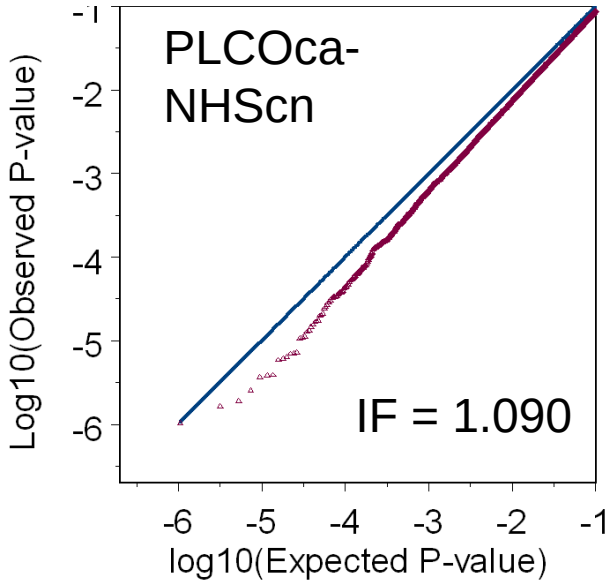
Over-dispersion factor for association tests with adjustment for various numbers of PCs

PCs chosen for the adjustment	PLOCca-PLCOco	PLCOca-NHSco	NHSca-NHSco	NHSca-PLCOco
0 PC	1.025	1.090	1.005	1.062
1st PC	1.020	1.055	1.006	1.040
1-2 PCs	1.022	1.040	1.004	1.013
1-3 PCs	1.021	1.040	1.005	1.006
1-4 PCs	1.021	1.032	1.005	1.007
1-5 PCs	1.023	1.032	1.006	1.008
1-10 PCs	1.025	1.036	1.008	1.010
Selected PCs	1.020	1.032	1.003	1.006

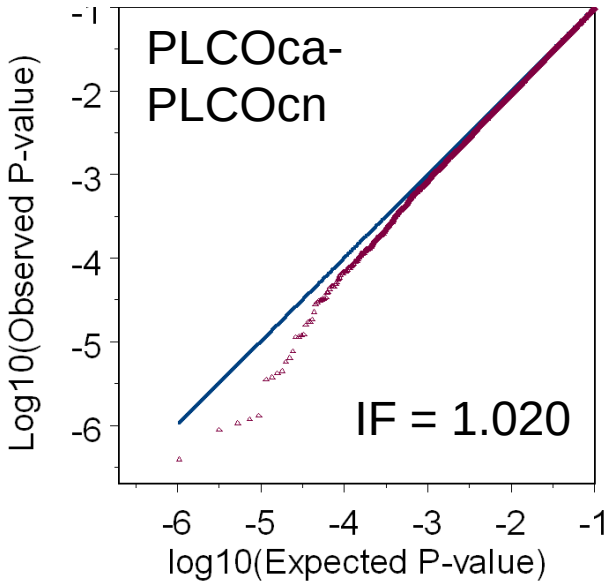
Q-Q Plot for the test with and without PC adjustment

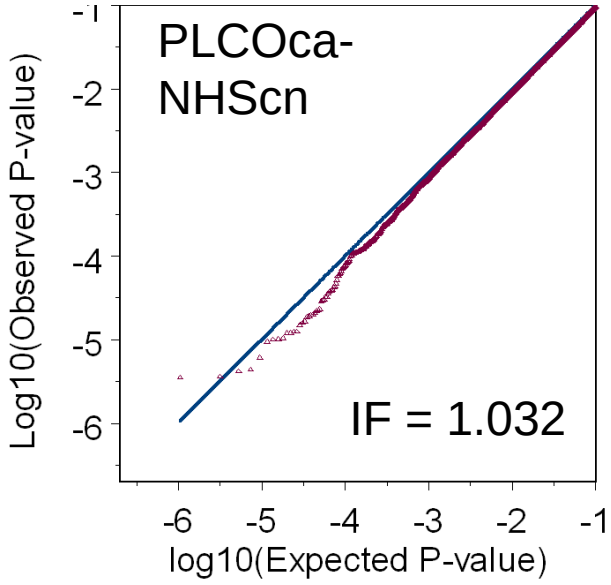
unadjusted





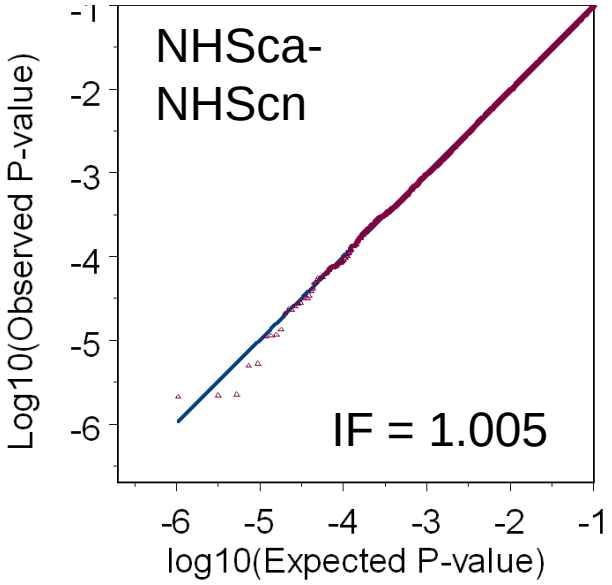
adjusted

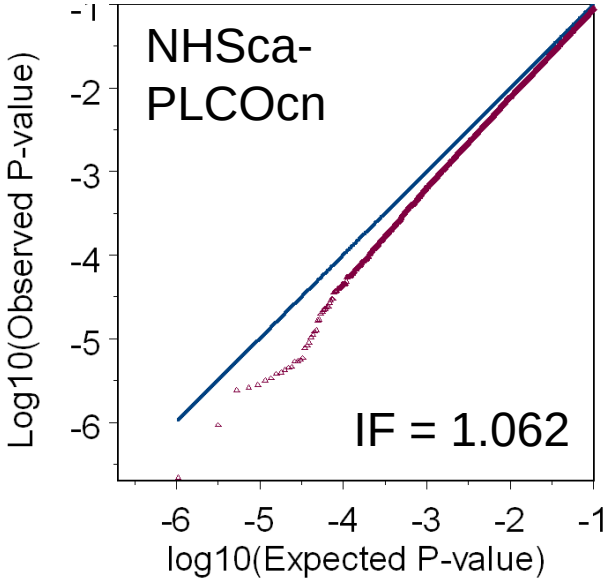




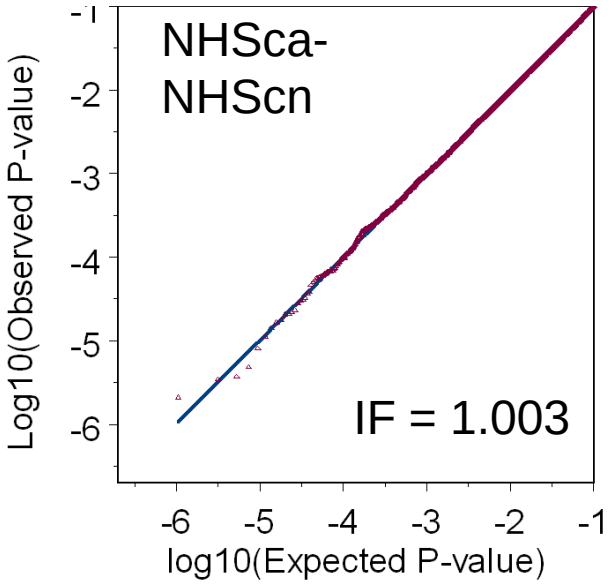
Q-Q Plot for the test with and without PC adjustment

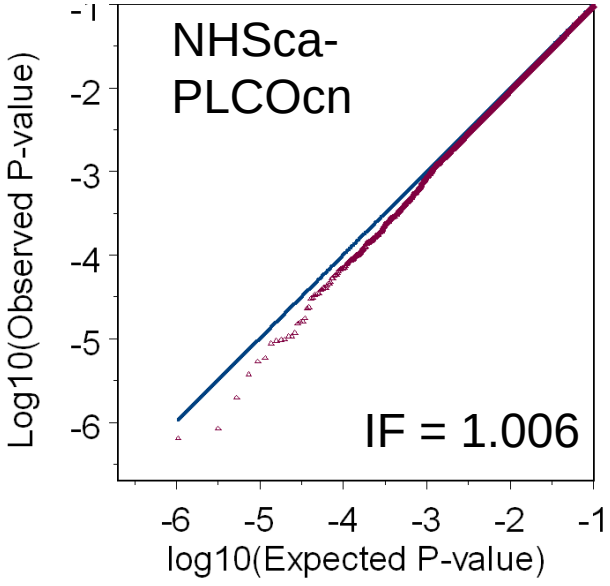
unadjusted





adjusted



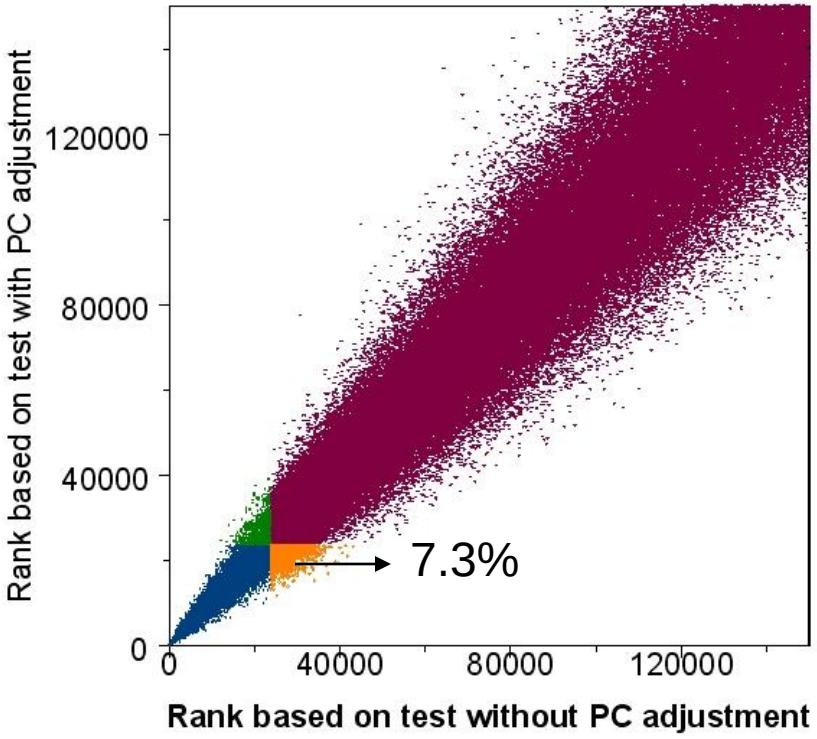


Discussions

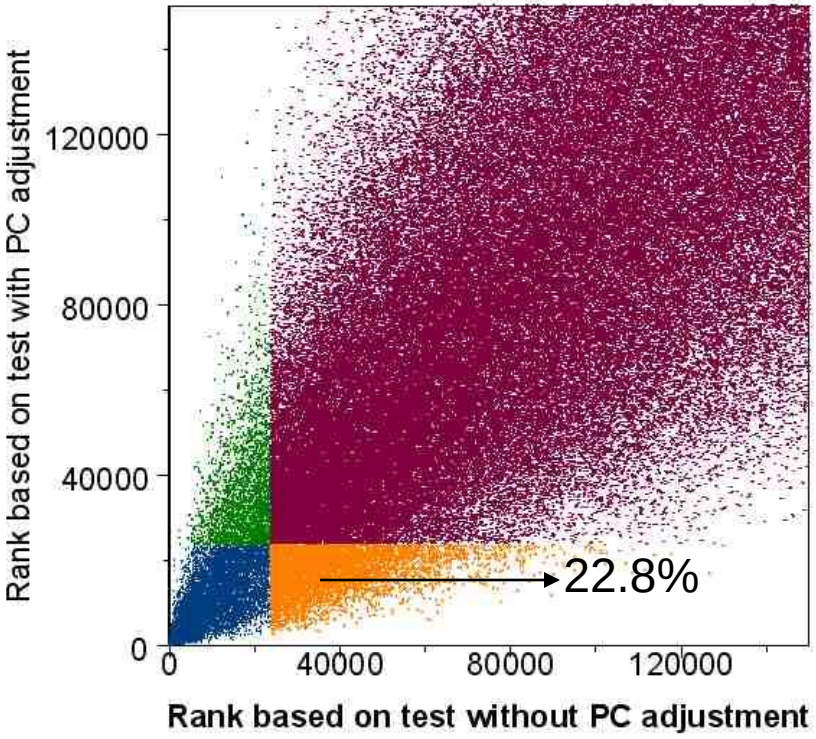
- We observed population heterogeneity exists within the European American population
- PS does not appear to be a problem in well-design studies
- Problem of PS is more extensive when external controls are used, but the adjustment of PCs can help
- We used empirical data for European Americans, what about other populations, such as African Americans?
- More issues to be considered when using “external controls”, such as,
  - Power issue
  - Covariate measurement harmonization

Discrepancy in SNP selection before and after PC adjustment

(selecting top 5% ranked SNPs)

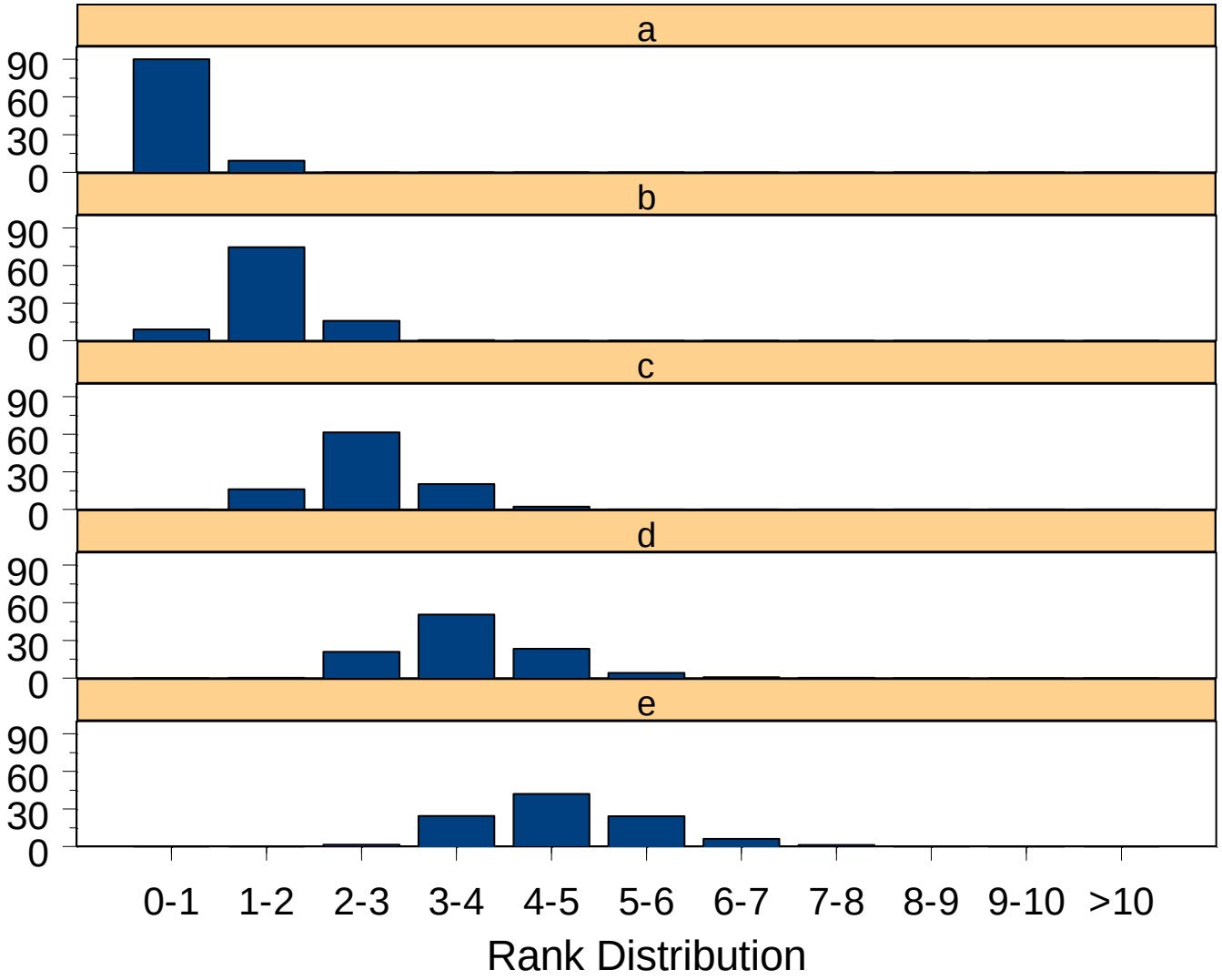


PLCO cases vs. PLCO controls

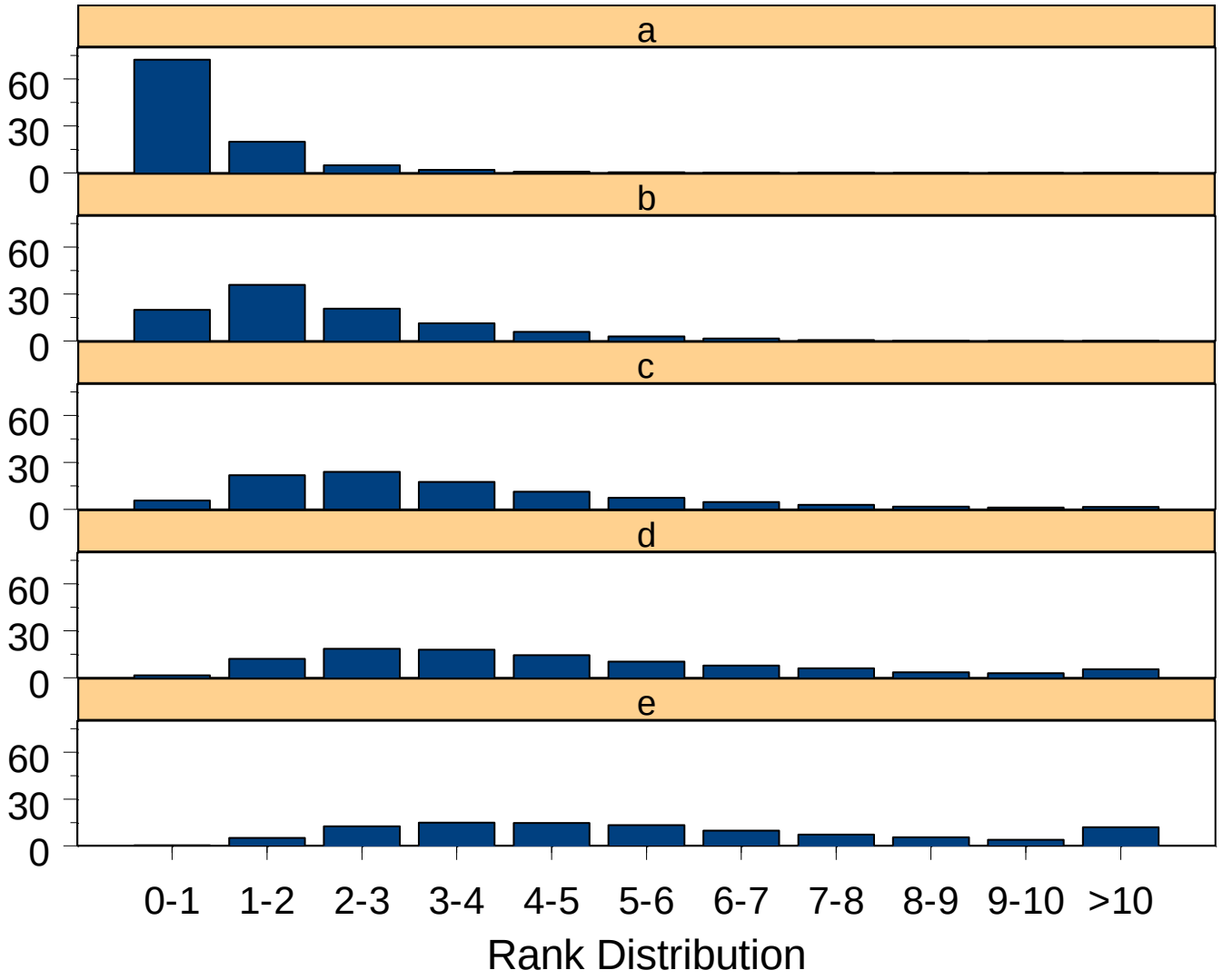


PLCO cases vs. NHS control

Rank shuffling in PLCOca-PLCOcn



Rank shuffling in PLCOca-NHScn

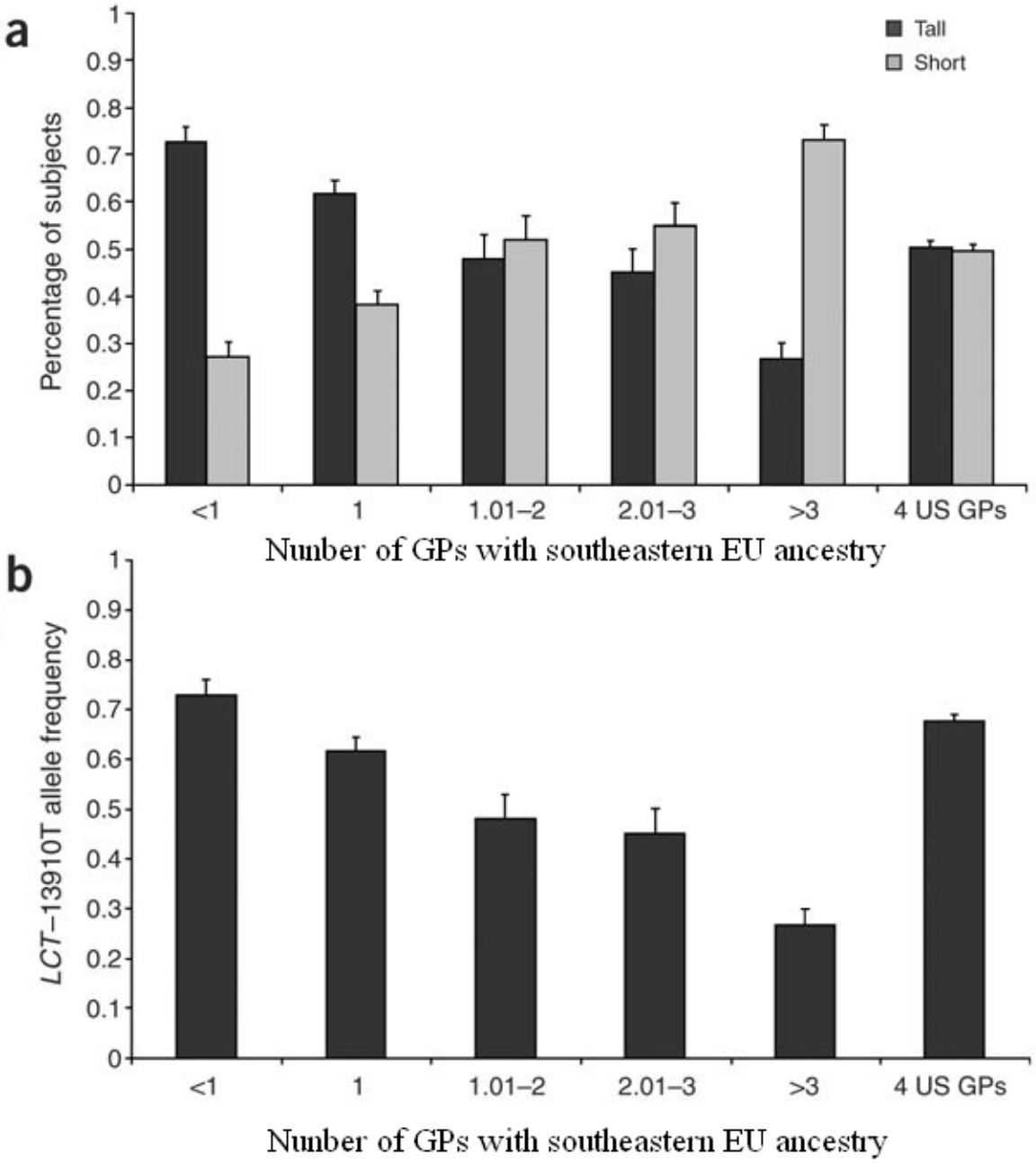


PS-I example: *LCT* and height

		Matching on four grandparents Ancestry		
	All	Four US-born	Southeastern Euro	Northwestern Euro
Tall	161:474:489	66:265:314	54:55:18	41:154:157
Short	231:444:380	76:278:282	128:86:13	27:79:86
P-value	3.6 x 10-7	0.098	0.0016	0.71

Note: after adjustment for the three classes, the P-value is 0.0074

Campbell et al. (NG, 2005)

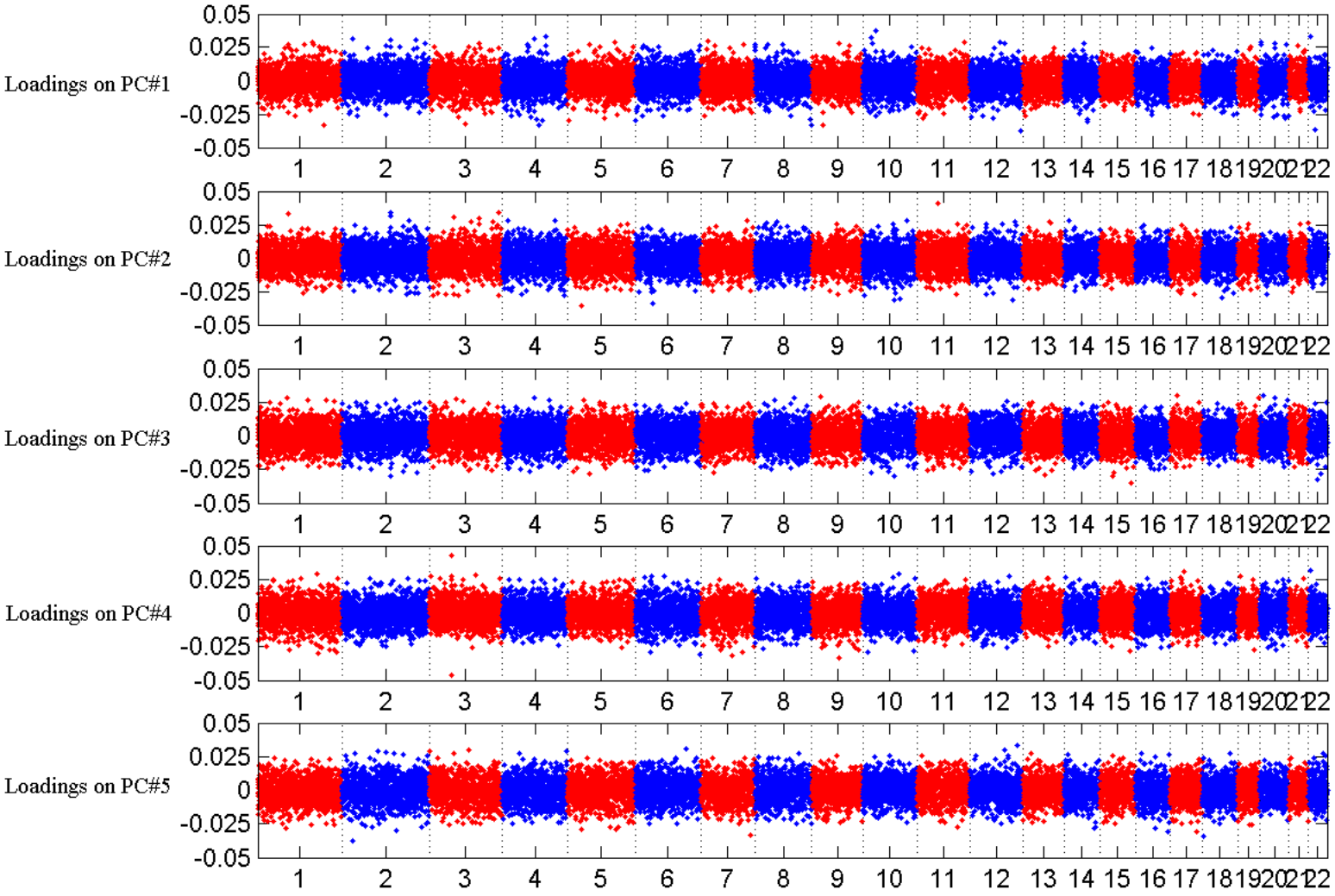


Sample selection and PS-II

Assuming common disease risk, any sampling bias that leads to ancestral difference will cause PS-II.

- Nested case-control design
  - the source population (cohort) is well defined
  - Minimal systematic bias in case control collection
- Standard case-control design
  - source population is not well defined
  - Control participation rate difference across subpopulations
- External controls (shared controls, freezer controls)
  - Cases and controls could represent different populations

Check of loadings (r2<0.004)



Check of loadings (r2<0.01)