Principal Component Analysis in Genomic Data

Seunggeun Lee

Department of Biostatistics
University of North Carolina at Chapel Hill

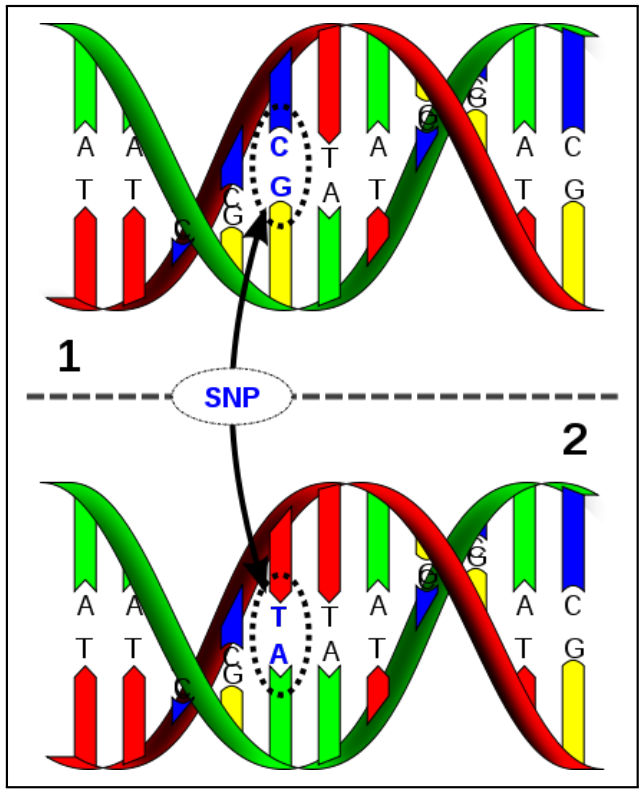
March 4, 2010

Bio

- Korean
- Undergraduate Major : Biology/Statistics
- Worked 3 and 1/2 years as a Software Engineer
- Came to UNC at 2005, admitted to MS and then progressed to PhD
- Dissertation Advisors : Dr. Fei Zou and Dr. Fred A. Wright

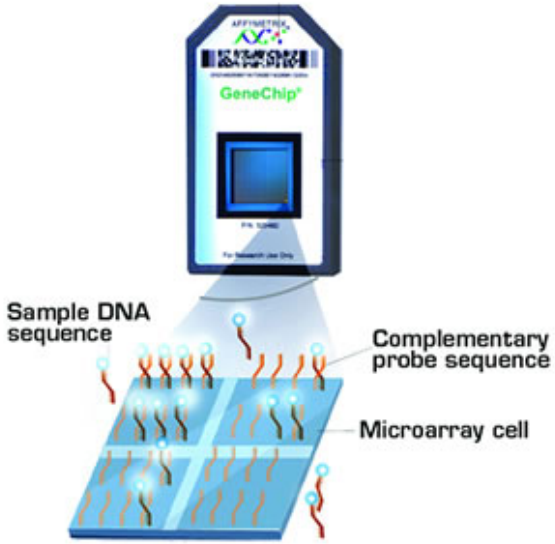
Genomewide Association Studies

- Single Nucleotide Polymorphism (SNP)
  - Single nucleotide variation
  - Occur every 100 to 300 bases
- Genomewide association Study
  - Goal : To find SNPs associated with case-control or quantitative traits.
  - Typically have $> 1,000$ samples
  - Test each SNPs

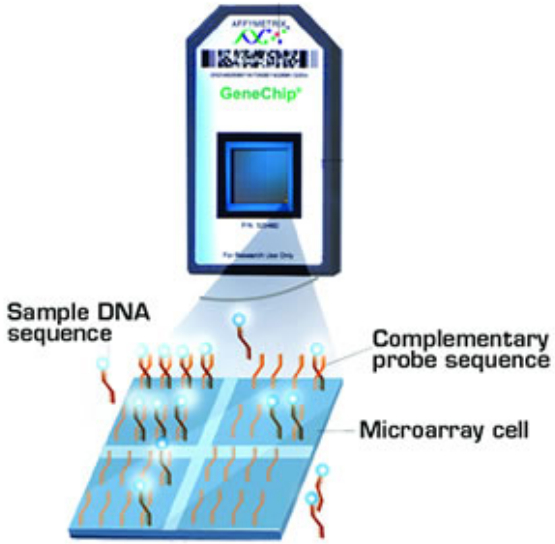


Genomewide Association Studies

- Obtain genotype using SNP microarray.
  - SNP chips have $500k \sim 1$ million SNPs



Genomewide Association Studies



- Obtain genotype using SNP microarray.
  - SNP chips have $500k \sim 1$ million SNPs
- For each SNPs (A vs a)

	AA	Aa	aa
Case	320	160	20
Control	245	210	45

  - P-value $= 2 \times 10^{-6}$
  - Compute p-values of all $5 \times 10^5 \sim 10^6$ SNPs

Genomewide Association Studies

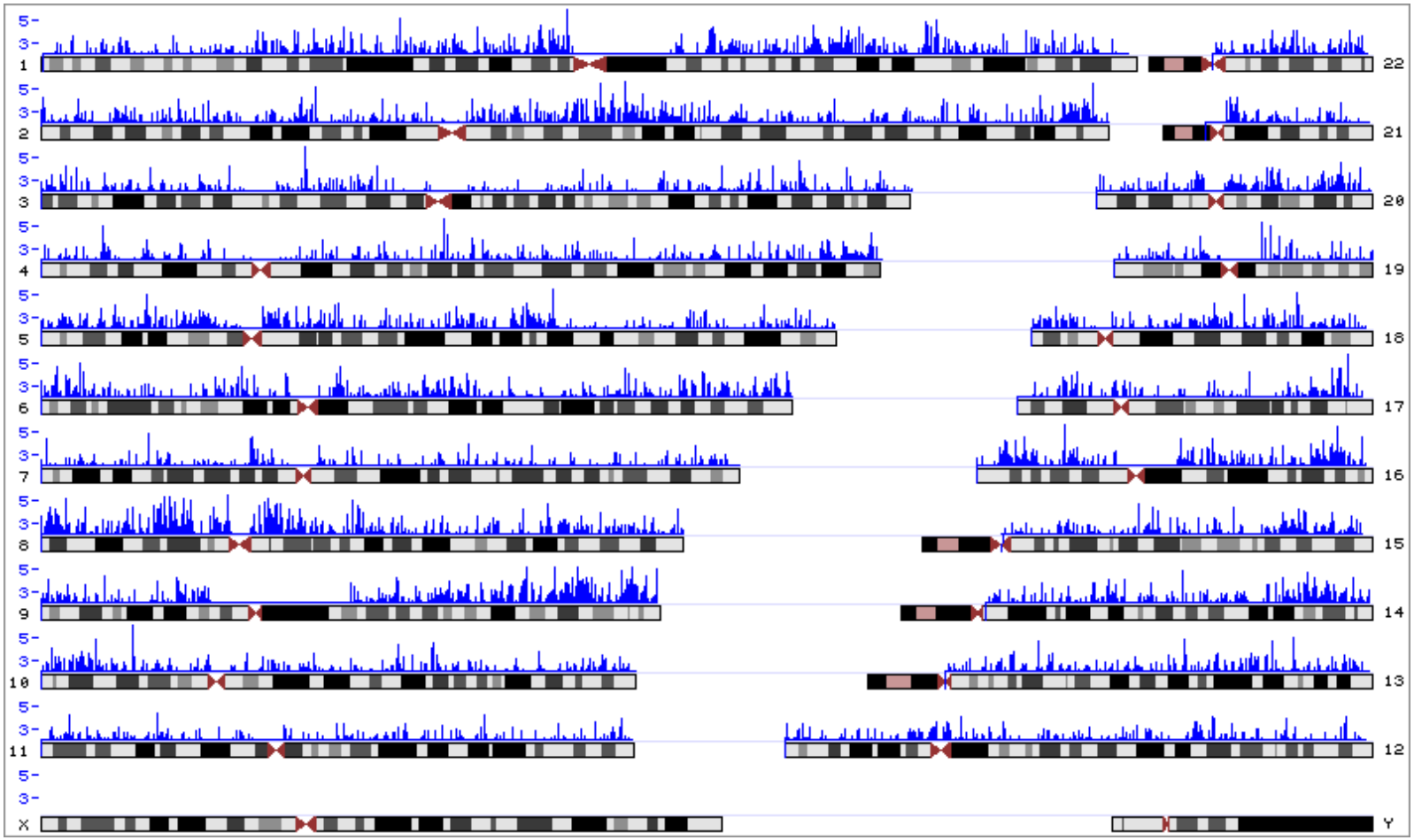
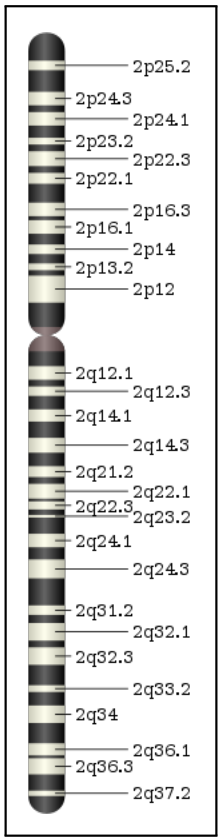


Figure: $-log_{10}$ P-values, from the GAIN Schizophrenia Data

Population Stratification

Population stratification is the presence of a systematic difference in allele frequencies between subpopulations. It is a major confounding factor of GWAS.

- Study : Genetic components of Height
- Samples were collected in Europe
  - Researchers have found that Chr 2q 21 region is associated with height
  - This region encodes the lactase gene (LCT)
- Relationship between Height and Lactase tolerance?



Northern vs. Southern European

Why ?

- Study was conducted in Europe
- Northern European vs. Southern European

	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

- Northern Europeans are taller than Southern Europeans
- Northern Europeans are lactose tolerant, but Southern Europeans are not.

Northern vs. Southern European

Why ?

- Study was conducted in Europe
- Northern European vs. Southern European

	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

- Northern Europeans are taller than Southern Europeans
- Northern Europeans are lactose tolerant, but Southern Europeans are not.

$\Rightarrow$ Population stratification (PS) is a major confounding factor in GWAS

PCA in GWAS data

- To adjust PS, we need to know the accurate ethnicity information. PCA is typically used for this purpose.
- *Price et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*

European

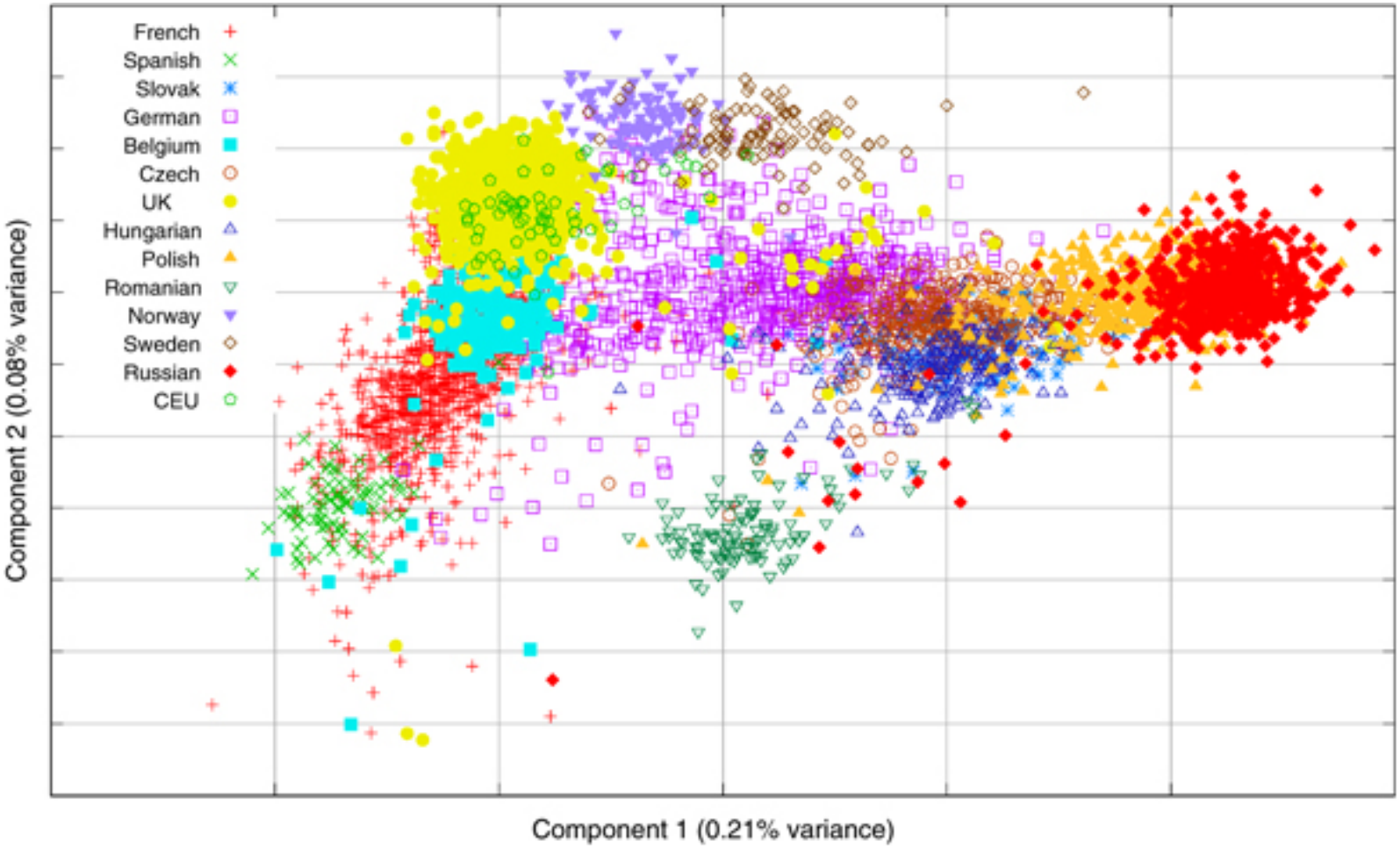


Figure: European Journal of Human Genetics (2008) 16, 14131429

Asian

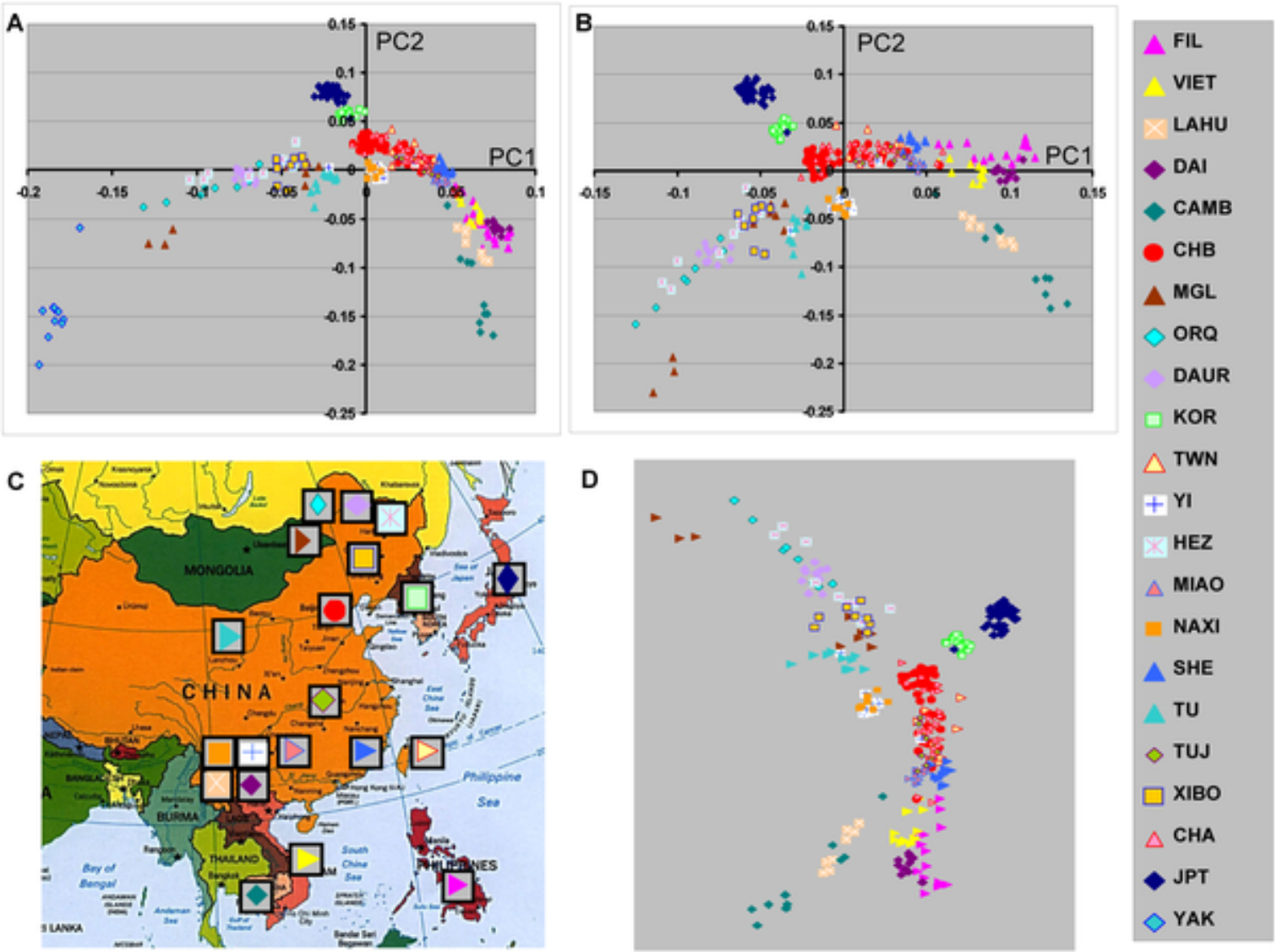


Figure: PLoS ONE. 2008; 3(12): e3862.

What I am doing

My research focuses on

- Linkage Disequilibrium Adjustment
  - Control of population stratification using correlated SNPs by shrinkage principal components. *Human Heredity* (accepted)

What I am doing

My research focuses on

- Linkage Disequilibrium Adjustment
  - Control of population stratification using correlated SNPs by shrinkage principal components. *Human Heredity* (accepted)
- Principal Component Selection
  - Control of population stratification by correlation-selected principal components. *Biometrics* (under revision)

What I am doing

My research focuses on

- Linkage Disequilibrium Adjustment
  - Control of population stratification using correlated SNPs by shrinkage principal components. *Human Heredity* (accepted)
- Principal Component Selection
  - Control of population stratification by correlation-selected principal components. *Biometrics* (under revision)
- Asymptotic behaviors of Principal Component
  - Convergence and prediction of principal component scores in high dimensional settings. *Annals of Statistics* (accepted)

What I am doing

My research focuses on

- Linkage Disequilibrium Adjustment
  - Control of population stratification using correlated SNPs by shrinkage principal components. *Human Heredity* (accepted)
- Principal Component Selection
  - Control of population stratification by correlation-selected principal components. *Biometrics* (under revision)
- Asymptotic behaviors of Principal Component
  - Convergence and prediction of principal component scores in high dimensional settings. *Annals of Statistics* (accepted)
- Other reserch papers: published and submitted in American Journal of Human genetics, Molecular Psychiatry, and Genetics.

Welcome to Chapel Hill!!