Genes Mirror Geography Within Europe

CSC 334

Hee Jin

Contents

Introduction

- High-throughput genotyping technologies + dense geographic samples
- Determine global patterns of variation among closely spaced populations
- Genetic ancestry -> infer geographic origin with high accuracy

Sample collection and genotyping

- Data
  - Strict Consensus Approach
    - Country of Origin of grandparents
    - If from two different countries -> eliminated
    - No info -> individual's country of birth
- Geographic locations
  - Geographic central point
  - Exceptions : capital

Principal Components analysis

- Smartpca software
- Conservative approach for filtering
  - avoid artefacts due to linkage disequilibrium
  - genome-wide patterns achieved

Spatial Assignment

- Independent linear model

$$\mathbf{x} = \beta_{x1}\mathbf{u}_1 + \beta_{x2}\mathbf{u}_2 + \beta_{x11}\mathbf{u}^2{}_1 + \beta_{x22}\mathbf{u}^2{}_2 + \beta_{x12}\mathbf{u}_1\mathbf{u}_2 + \boldsymbol{\varepsilon}$$

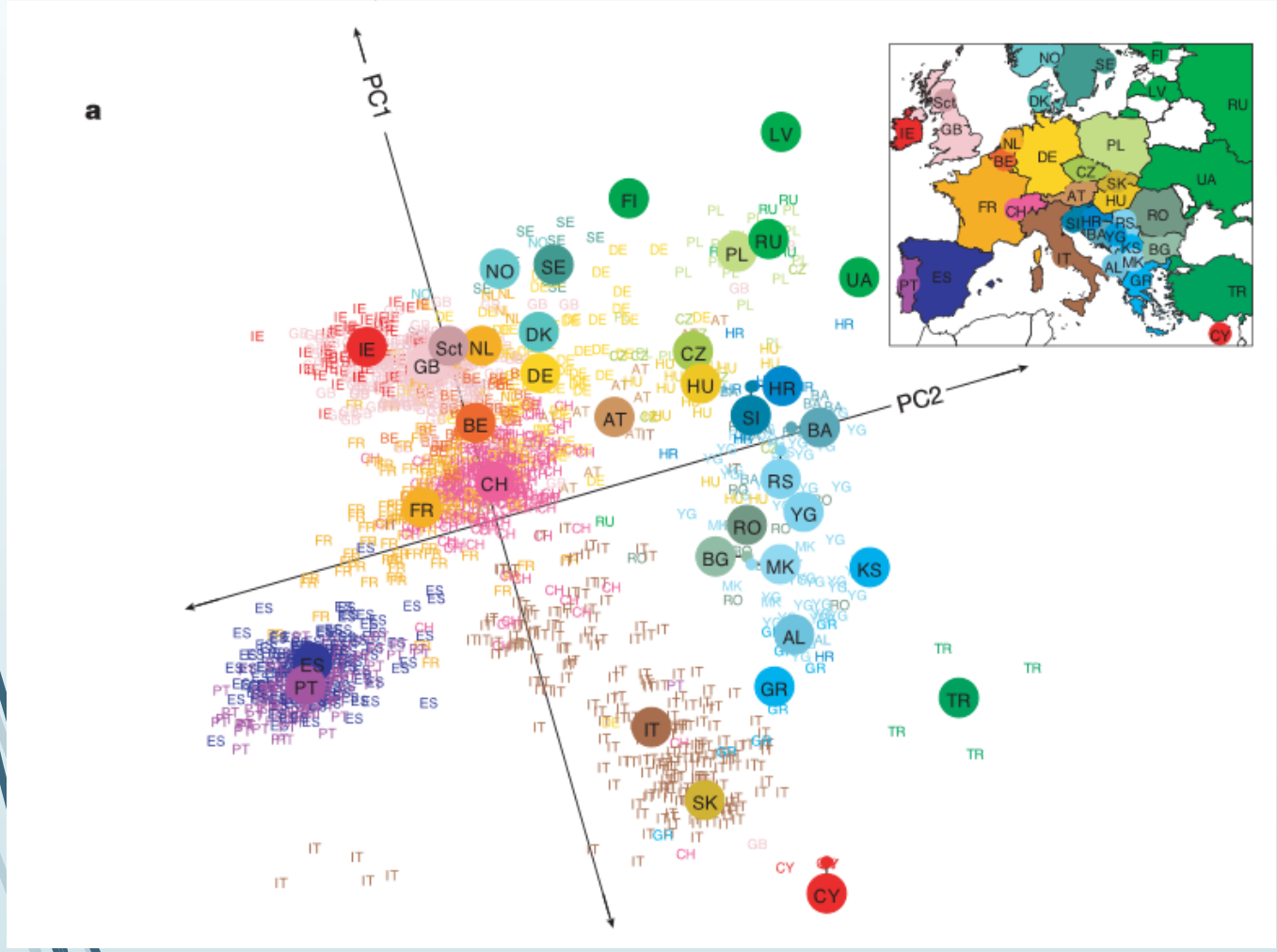
$$\mathbf{y} = \beta_{y1}\mathbf{u}_1 + \beta_{y2}\mathbf{u}_2 + \beta_{y11}\mathbf{u}^2{}_1 + \beta_{y22}\mathbf{u}^2{}_2 + \beta_{y12}\mathbf{u}_1\mathbf{u}_2 + \boldsymbol{\varepsilon}$$

- Continuous assignment
  - Predict location by PC1 and PC2
- Discrete assignment
  - Assign individuals to country

Genome-wide association simulations

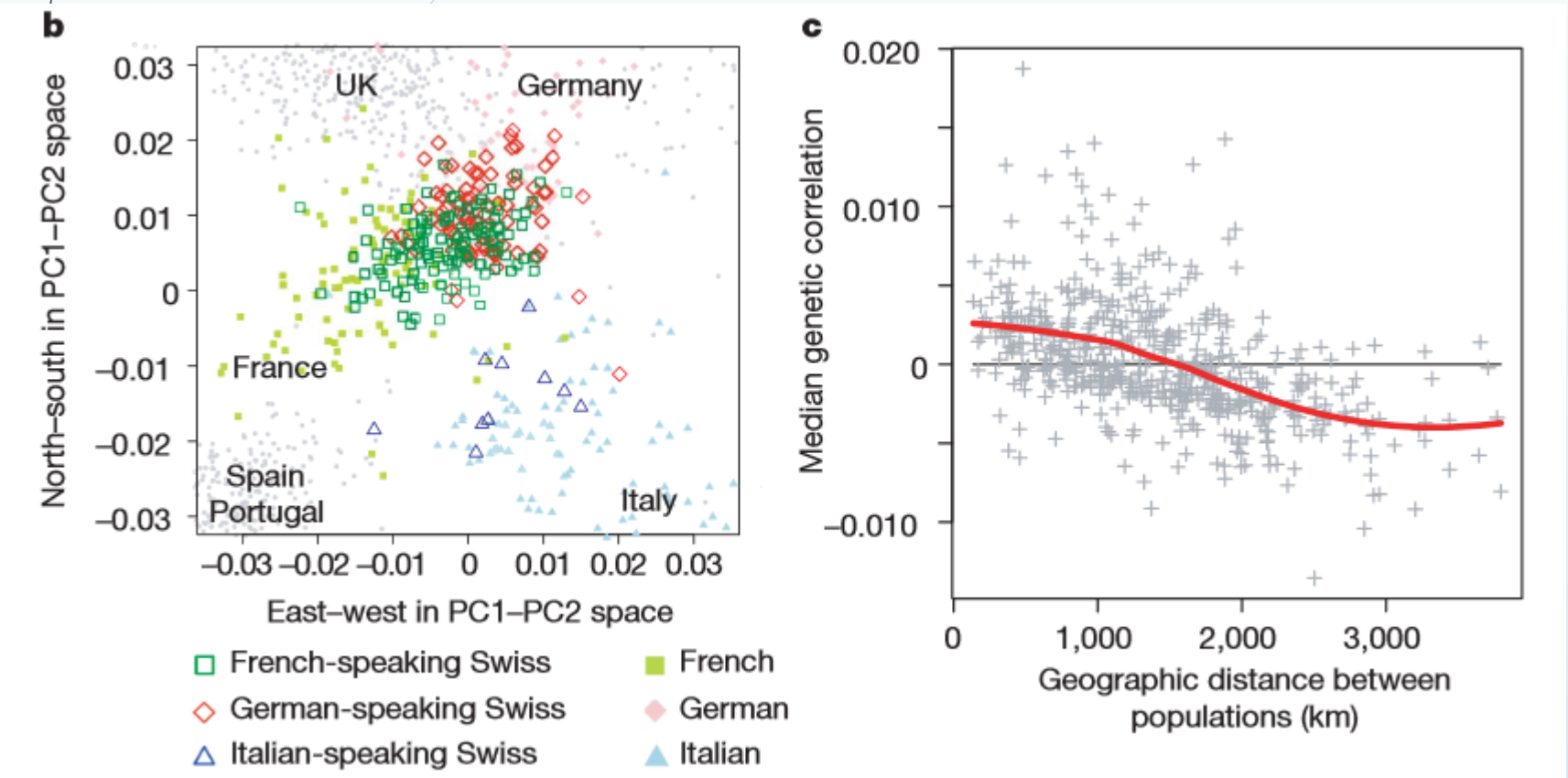
- For disease-association studies
- Examine Spatial variation in phenotype
- Multiple linear regression with PC1 & PC2 as covariates
- Quantitative traits Phenotype simulated
- PC-based correction for p-value inflation

Results



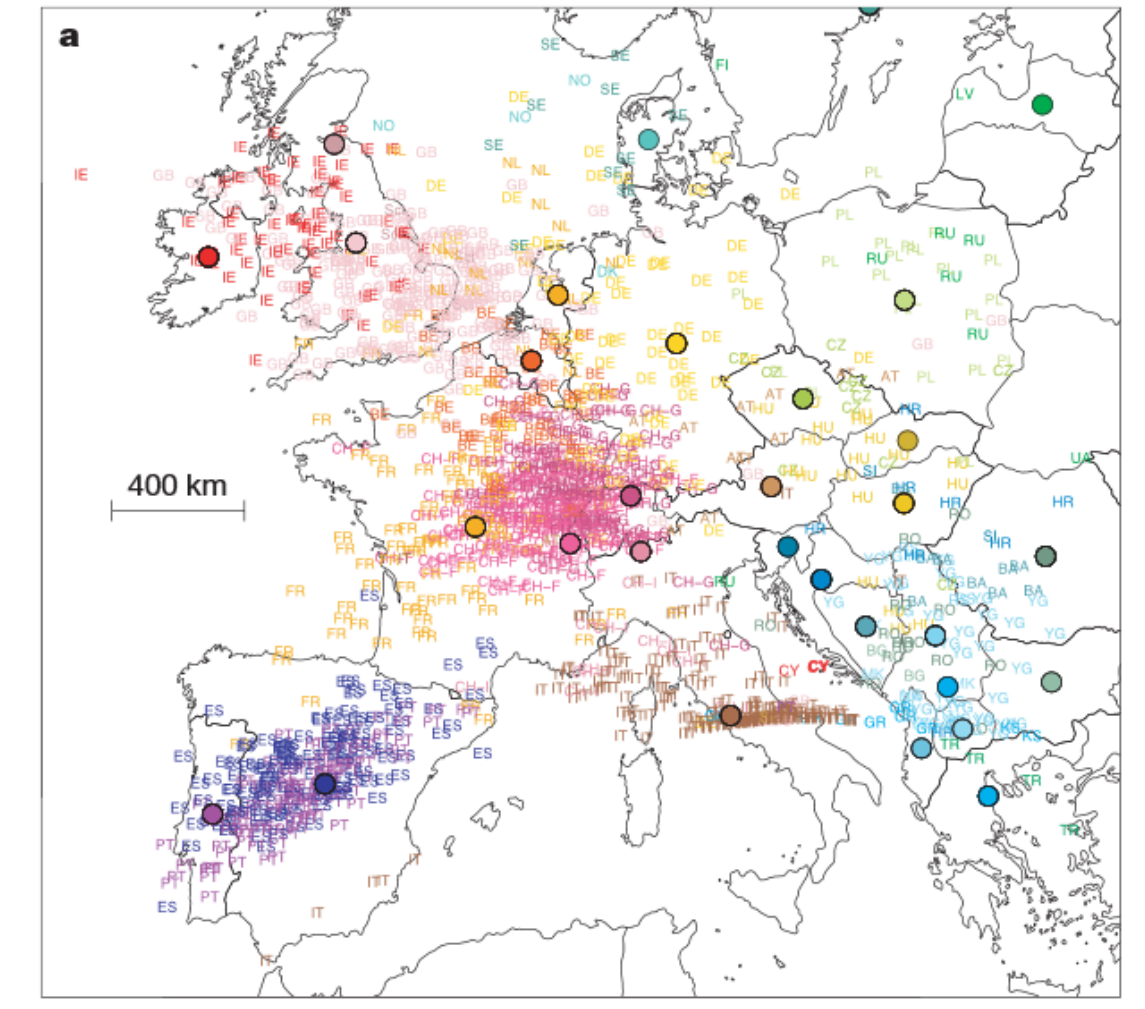
- PCA
- Similar to geographic map of Europe
- Theoretical expectation
  - Genetic similarity ↓ as distance ↑
  - PC1 & PC2 correlate with perpendicular geographic axes
  - Separation of discrete clusters along PCs -> X

Results



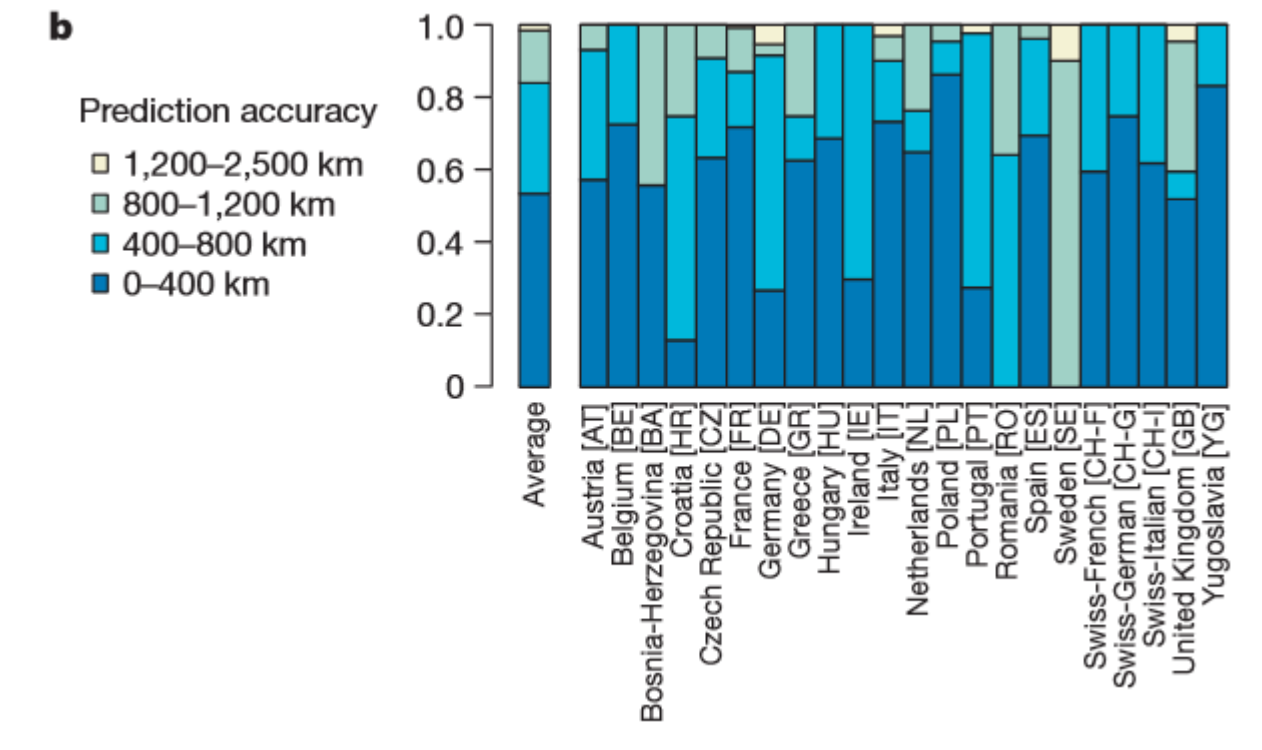
- Differentiation in population by language
- Correlation ↓ as distance ↑

Results



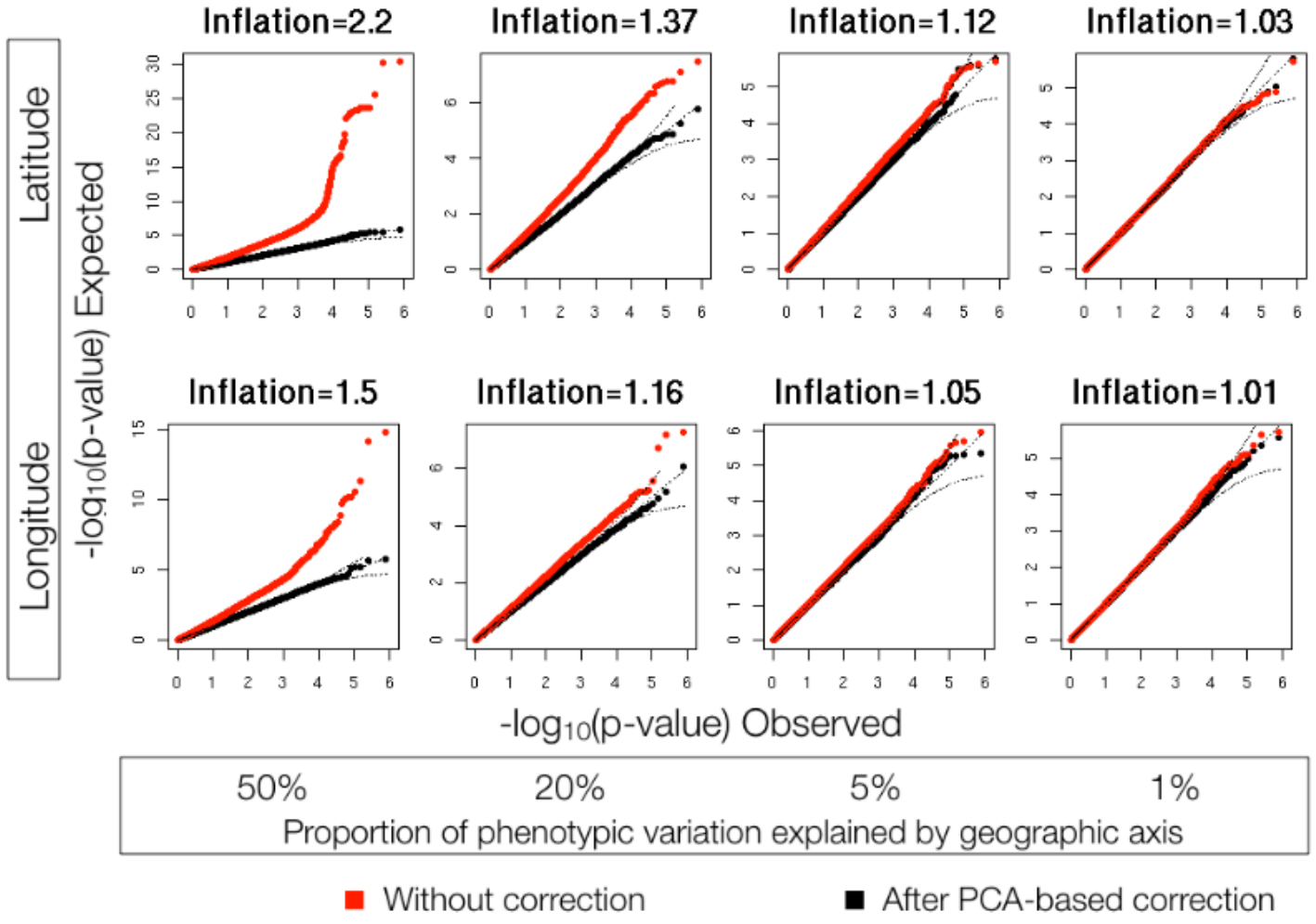
- Spatial assignment
- Most of the individuals placed near reported origin

Result



- Distribution of prediction accuracy
- Performance ↓ as population sample size ↓

Results



- PCA based correction controls p-value inflation
- PCA could also be used to explain spatial variation in phenotype
- But caution
  - Which phenotypes
  - Sample size
  - Distribution of sampling locations

Discussion

- Under-represent variation at low-frequency alleles
- PCA based on genotypic patterns of variation
- Do not take advantage of patterns of haplotype variation
- Future research
  - Access to informative low-frequency alleles