Journal club: Correcting methods of Stratification of population in genome-wide association studies

Safia Safa-tahar-henni

Review

February 2, 2018

Overview

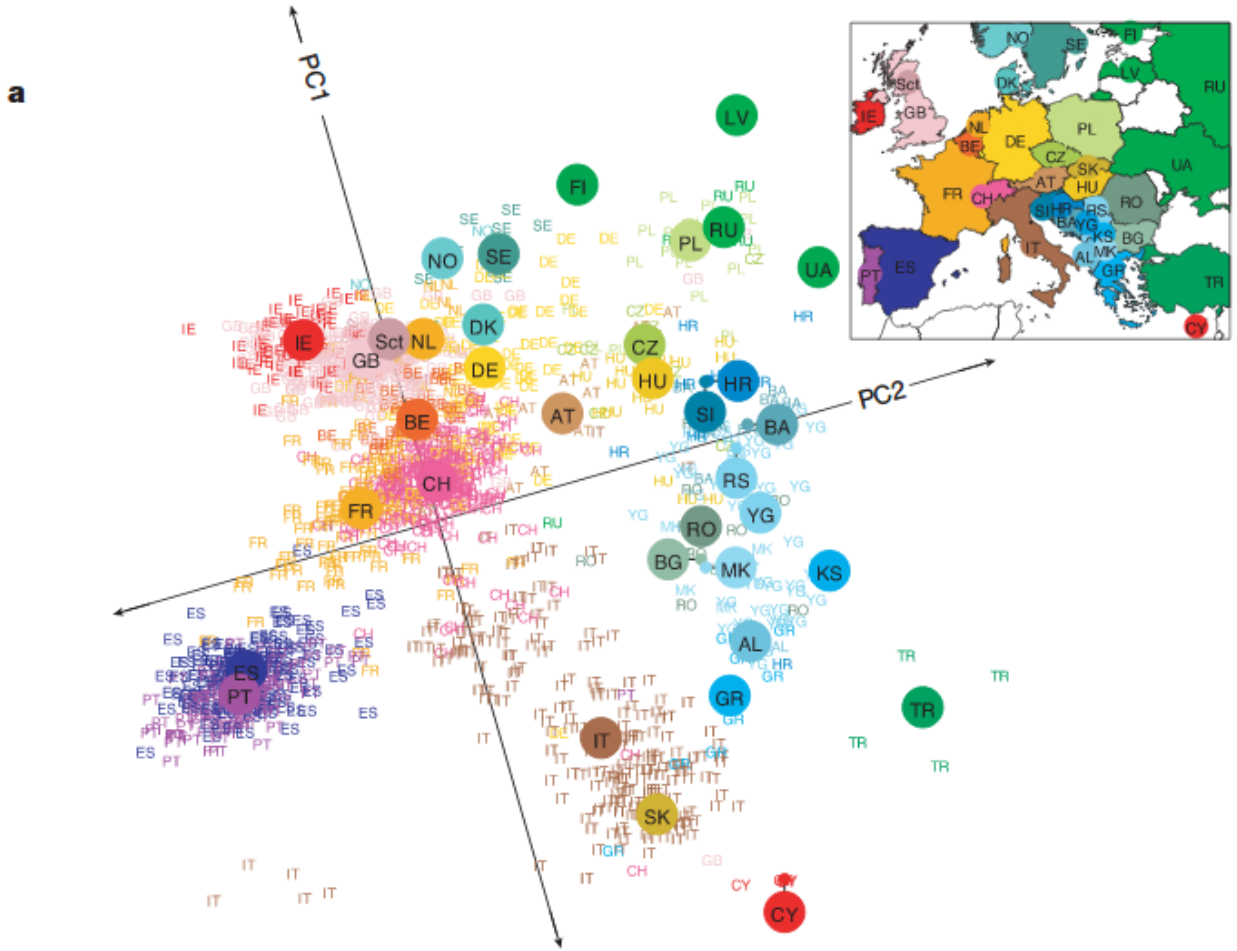
1. Introduction
   - Context
   - Corrects population stratification in GWAS
   - Genomic Control
   - Structured association
   - Principal components analysis
   - Linear mixed model

Context

Genome-wide association studies (GWAS) have become routine for unraveling the genetic variants underlying complex phenotypes in humans and many other species.

Population stratification(PS):
Allele frequency differences between cases and controls due to systematic ancestry differences.

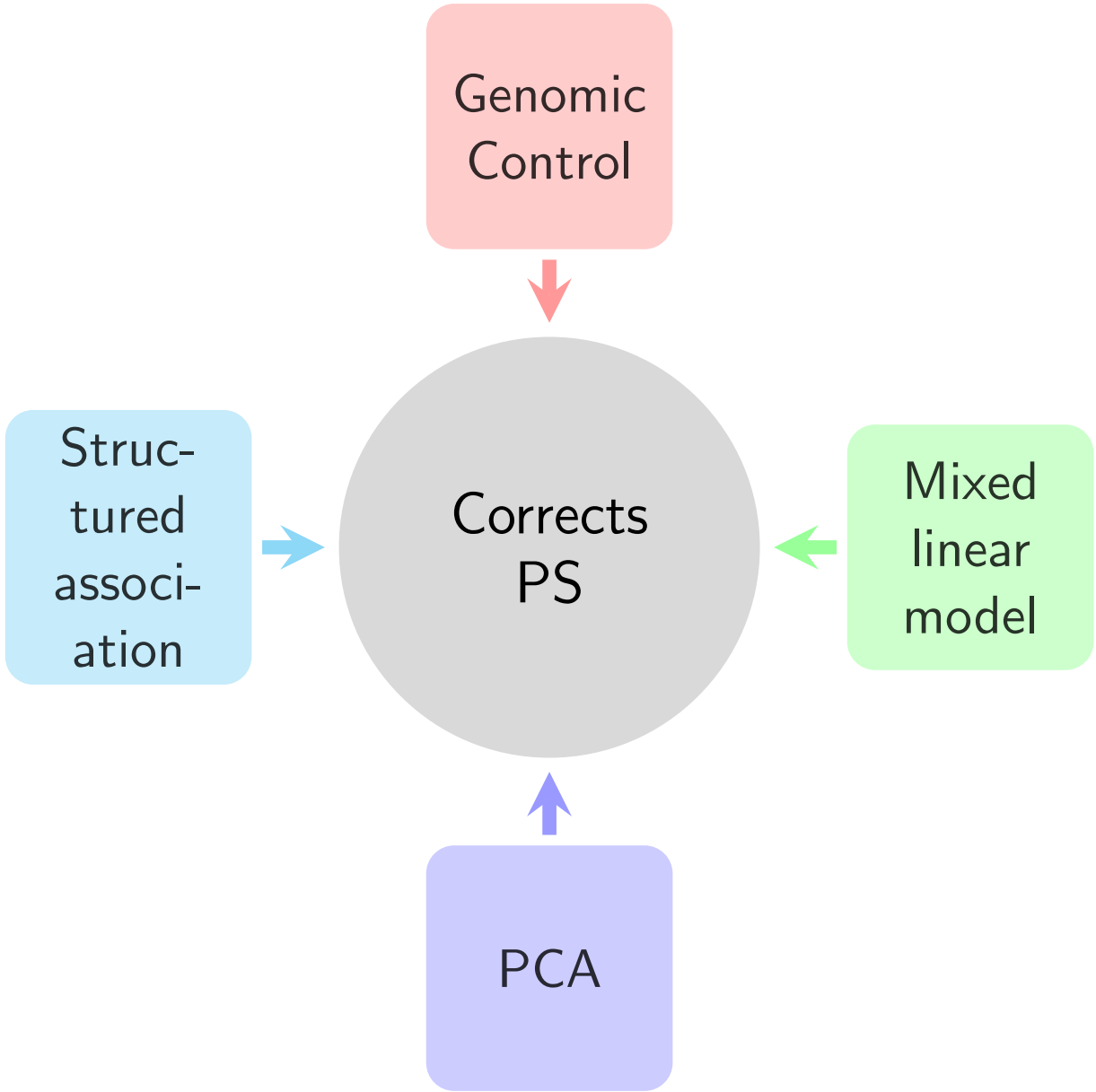
Context (2)



The effects of stratification vary in proportion to the number of samples.

Corrects population stratification in GWAS

Problem: Population stratification often produces spurious genetic associations.



Genomic Control

- Theory:
  Adjusting association statistics at each marker by a uniform overall inflation factor (λ).

  Used to test the association with a test of χ^2_1

  Test statistic $= \chi^2$ Pearson test or Cochran-Armitage trend test (Y^2).

Genomic Control (2)

- Theory:
  Without stratification:
  - the test statistic follows a distribution of χ^2_1

  With stratification:
  - the test statistics : $Y^2 \sim \lambda\chi^2_1$.

  In GWAS, $\lambda =$
  - the median of all the χ^2_{obs} statistics $\div 0,4549$ (the median of the χ^2_1 distribution),
  - or average value of all the χ^2_{obs} statistics .

Genomic Control (3)

- Disadvantage:

  Some markers differ in their allele frequencies across ancestral populations more than others.

  Uniform adjustment may be:
  - insufficient at markers having unusually strong differentiation across ancestral populations
  - superfluous at markers devoid of such differentiation,

  $\Rightarrow$ Loss in power.

Structured association

- Theory:

  Assign the samples to discrete subpopulation clusters (STRUCTURE program) and then aggregates evidence of association within each cluster.

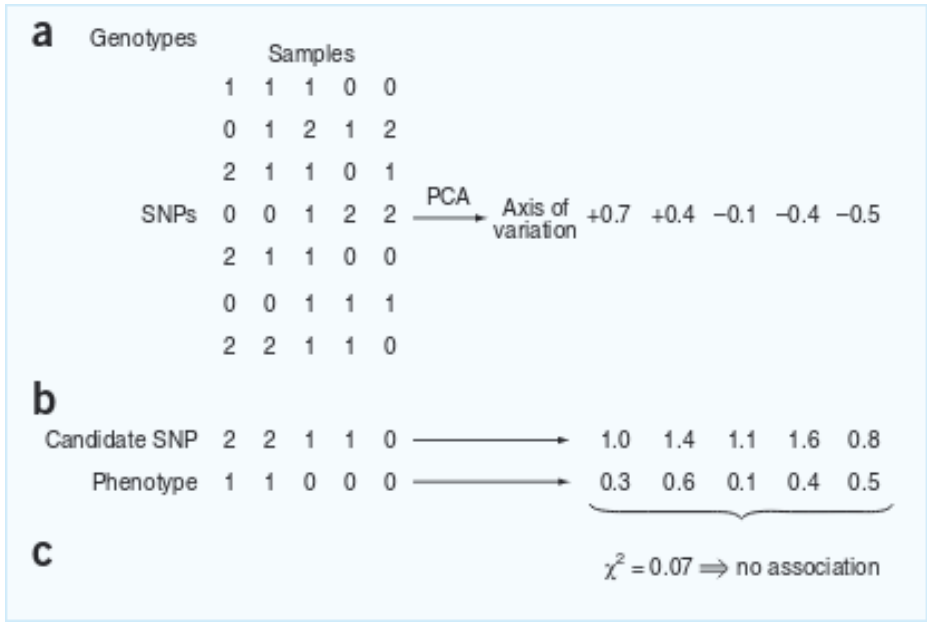
- Disadvantage:

  - intensive computational cost on large data sets.
  - assignments of individuals to clusters are highly sensitive to the number of clusters (not well defined)

Principal components analysis

- Principal components analysis (PCA)
  $\rightarrow$ EIGENSTRAT (Price & al.)

(a) Principal components analysis
(b) Ancestry adjustment
(c) Association statistic



Principal components analysis (2)

Correcting for stratification using continuous axes of variation has several advantages:

- Provide the most useful description of within-continent genetic variation
- Continuous axes orthogonal $\rightarrow$ results are insensitive to the number of axes inferred
- Computationally tractable on a genome-wide scale.

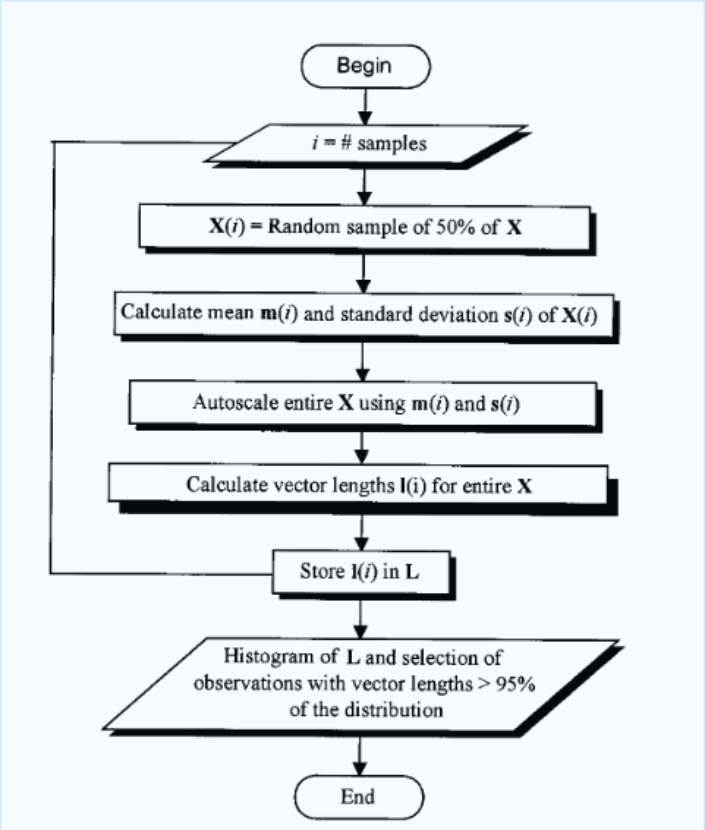
Principal components analysis (3)

- Multidimensional scaling (MDS):
  Li Q & al extension of EIGENSTRAT [Price et al.]:
  - multidimensional scaling (MDS)
  - clustering analysis

  Advantage:
  - requires a smaller number of markers
  - gives a more appropriate correction for the stratification of the population.

Principal components analysis (4)

- Robust PCA based on resampling by half means (RPCA-RHM): Detect outliers by studying the distribution of observation vector lengths obtained by sampling without replacement from the original data set.



Principal components analysis (5)

- Robust PCA based on the projection pursuit (RPCA-PP)

Projecting multivariate data on a lower-dimensional subspace.

C. Croux & al demonstrate that the currently available algorithm performs poor in the presence of many variables.

Linear Mixed Model (LMM)

Use pairwise genetic relationships among individuals with abundant genotype data.

A standard MLM for GWAS :

$$Y = Wv + X\beta + Zu + e$$

Hyunju Ryoo & al study suggests an underestimated heritability in GWAS upon using the mixed model methodology with an excessively larger number of variants versus causal variants.

Thank you for your attention