Analysis and Application of European Genetic Substructure Using 300K SNP Information

Mikk Eelmets

Journal Club

19.02.2008

Introduction

- Differences in population genetic structure between cases and controls can lead to false positive association tests
- Population substructure can be explored using principal component anaysis or non-hierarchical cluster analysis

Results

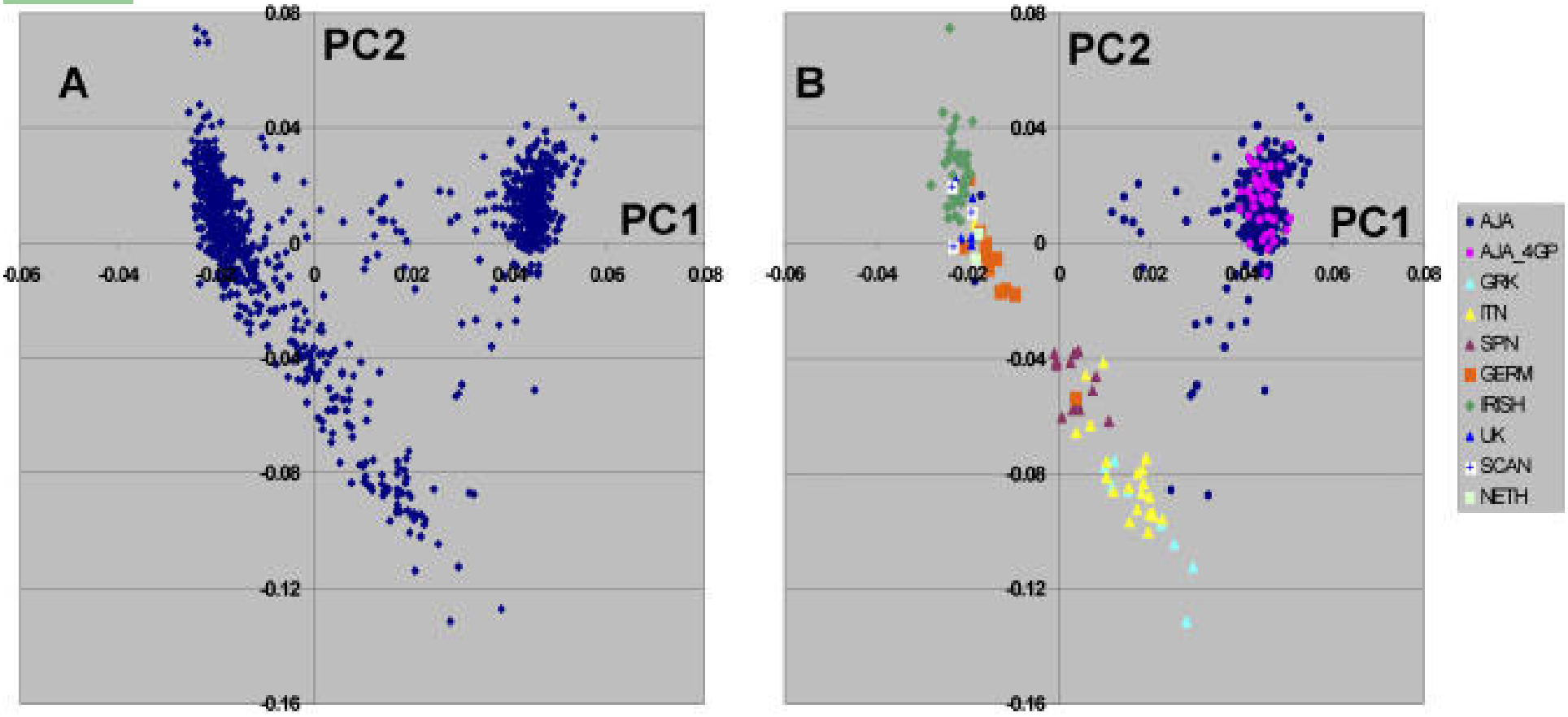
- 952 self-identified (European decent)
- >300K SNPs
- Genotypes were examined using PCA algorithm implemented in the EIGENSTAT program

Results

Principal Component	Percent Eigenvalue[a] (Top 10)	SHT[b] r^2	ANOVA[c] r^2	ANOVA *p* Value	NL DIST[d] *p* Value
PC1	42.42%	0.991+/−0.001	0.983	2.95E−121	1.14E−11
PC2	8.32%	0.559+/−0.044	0.936	4.66E−80	2.00E−12
PC3	6.66%	0.009+/−0.011	0.068	4.96E−01	1.45E−01
PC4	6.36%	0.024+/−0.023	0.766	5.50E−40	2.50E−06
PC5	6.13%	0.034+/−0.041	0.253	9.97E−06	1.10E−02
PC6	6.06%	0.015+/−0.009	0.143	1.84E−02	2.96E−01
PC7	6.03%	0.004+/−0.003	0.045	8.14E−01	1.10E−01
PC3 (no Inv)[e]	6.70%	0.047+/−0.017	0.773	7.06E−41	1.88E−06
PC4 (no Inv) [e]	6.36%	0.067+/−0.107	0.256	7.50E−06	9.63E−03

[a]The % Eigenvalue is the percentage of the total variance in the first ten PCs.
[b]The Spearman-Brown split half reliability test (SHT) [41] r^2 is the mean +/− SD from the adjusted correlations between: (1) every other chromosomes; (2) half chromosomes (first half each chromosome and second half each chromosome); and (3) first half genome and second half genome (see Methods). These correlations, ANOVA, and test for normality of distribution were determined after PCA of each individual set.
[c]ANOVA results are based on prior self-identified ethnic group assignments.
[d]The *p* values determined using Shapiro and Wilk's W test [42] indicate whether the probability that the null hypothesis, normal distribution is consistent with the observed data.
[e]Results for PC3 and PC4 changed after removal of SNPs within Chromosome 8 inversion (see text).
doi:10.1371/journal.pgen.0040004.t001

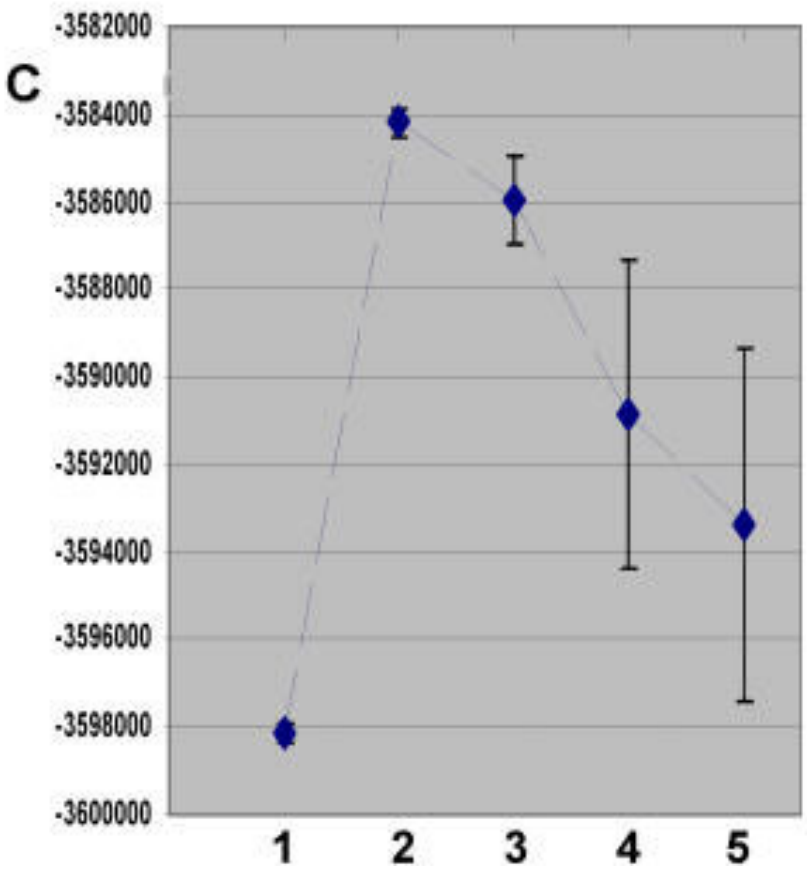
Results

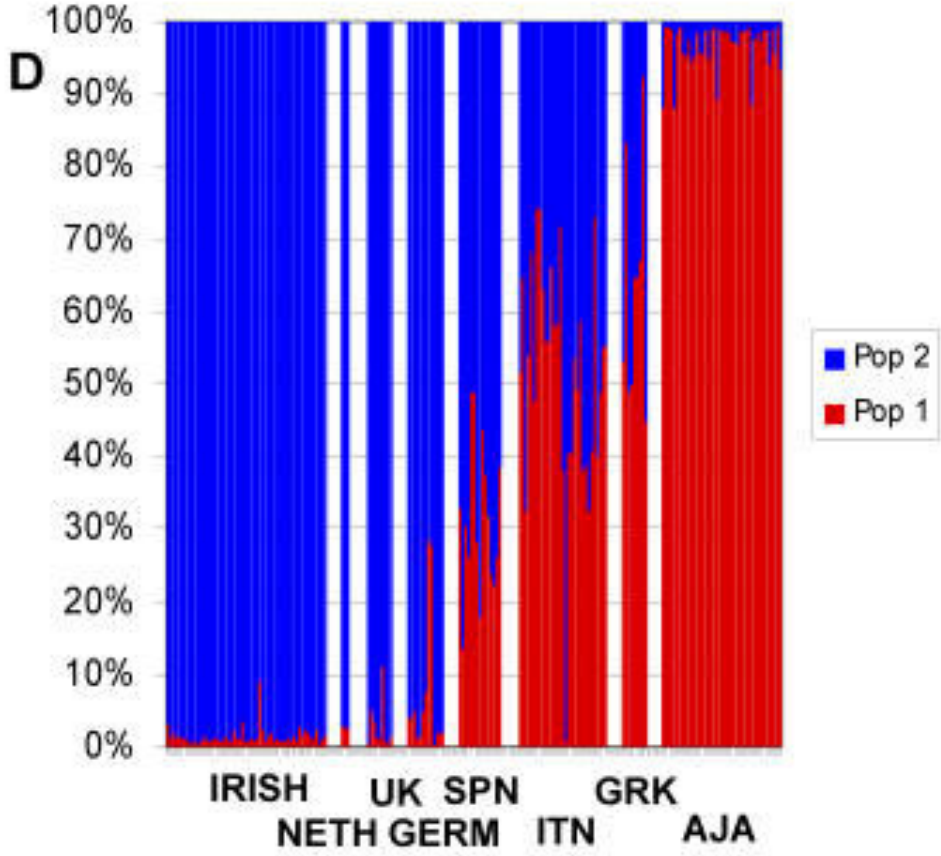

A
PC2
PC1
0.08
0.04
0
-0.04
-0.08
-0.12
-0.16
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
B
PC2
PC1
AJA
AJA_4GP
GRK
ITN
SPN
GERM
IRISH
UK
SCAN
NETH

Results

- 952 self-identified (European decent)
- Three sets >3500 SNPs
- Genotypes were examined using Bayesian clustering algorithm in the STRUCTURE program

Results


C
-3582000
-3584000
-3586000
-3588000
-3590000
-3592000
-3594000
-3596000
-3598000
-3600000
1
2
3
4
5


D
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
IRISH
NETH
UK
GERM
SPN
ITN
GRK
AJA
Pop 2
Pop 1

Results

- WGA study for RA in European Americans
- IRF4 would appear as strong candidate
- After using algorithm based on PCs, EIGENSTRAT, IRF4 no longer associate with RA

Reference

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK, Seldin MF.

Analysis and Application of European Genetic Substructure Using 300K SNP Information

PLoS Genet. 2008 Jan;4(1):e4.

THANK YOU