Main causes of population structure

1. **Lack of random mating (geographic or cultural isolation). Allele frequencies are not homogenous among demographic subgroups.**
2. **Recent population migrations**
3. **Presence of cryptic familial relationships between individuals**

Most Genome wide-association studies rely on the assumption that cases and controls are selected from the same homogeneous population.

An ascertainment bias in the collection of cases and controls from a structured population may lead to inadequate matching of the two groups, thus resulting in decreased statistical power for the detection of true associations

Two Approaches to Characterizing Population Structure

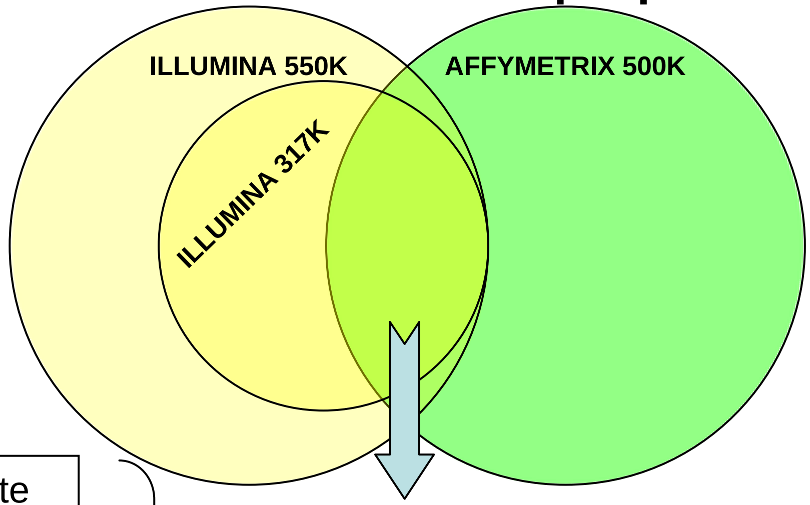
1. Principal Component Analysis (Price et al. Nat Gen 2006)
   - Captures correlation between genotypes.
   - Ranks the detected correlations.

2. STRUCTURE (Pritchard et al. Genetics 2000)
   - Attempts to interpret the correlation between genotypes in terms of admixture between a defined number of ancestral populations

- Both approaches rely on the use of a set of SNPs that do not demonstrate background LD.

- Interest in identifying a group of SNPs that may be used to compare Genome Wide Association Studies.

Selection of a set of SNPs for population stratification



50 374 SNP

Remove SNPs with call rate < 90% in either Illumina or Affymetrix

Remove untyped or monomorphic SNPs in YRI or (JPT+CHB)

Remove SNPs with P-values for HW proportion < 0.01

40 829 SNP

Select a set of SNPs with parwise $r^2 < 10^{-3}$

10 095 SNPs

These SNPs are expected to provide reliable genotypes and will be included in the SNP set typed in most GWAS.

When ancestral populations known optimization possible :
Pfaff et al. Genetic Epi 26:305-315(2004)

A model of a structured population

Population studied :

-Europe : CEPH founders	=>	60	individuals	HapMap
-African : YRI founders	=>	59	individuals	HapMap
-Asian : CHB	=>	44	individuals	HapMap
-Asian : JPT	=>	45	individuals	HapMap
-Native American : Mexican	=>	30	individuals	Penn State U.*
-Native American : Mayan	=>	25	individuals	Penn State U.*
-African Americans	=>	15	individuals	Penn State U.*
- "Latino"	=>	7	individuals	SNP500

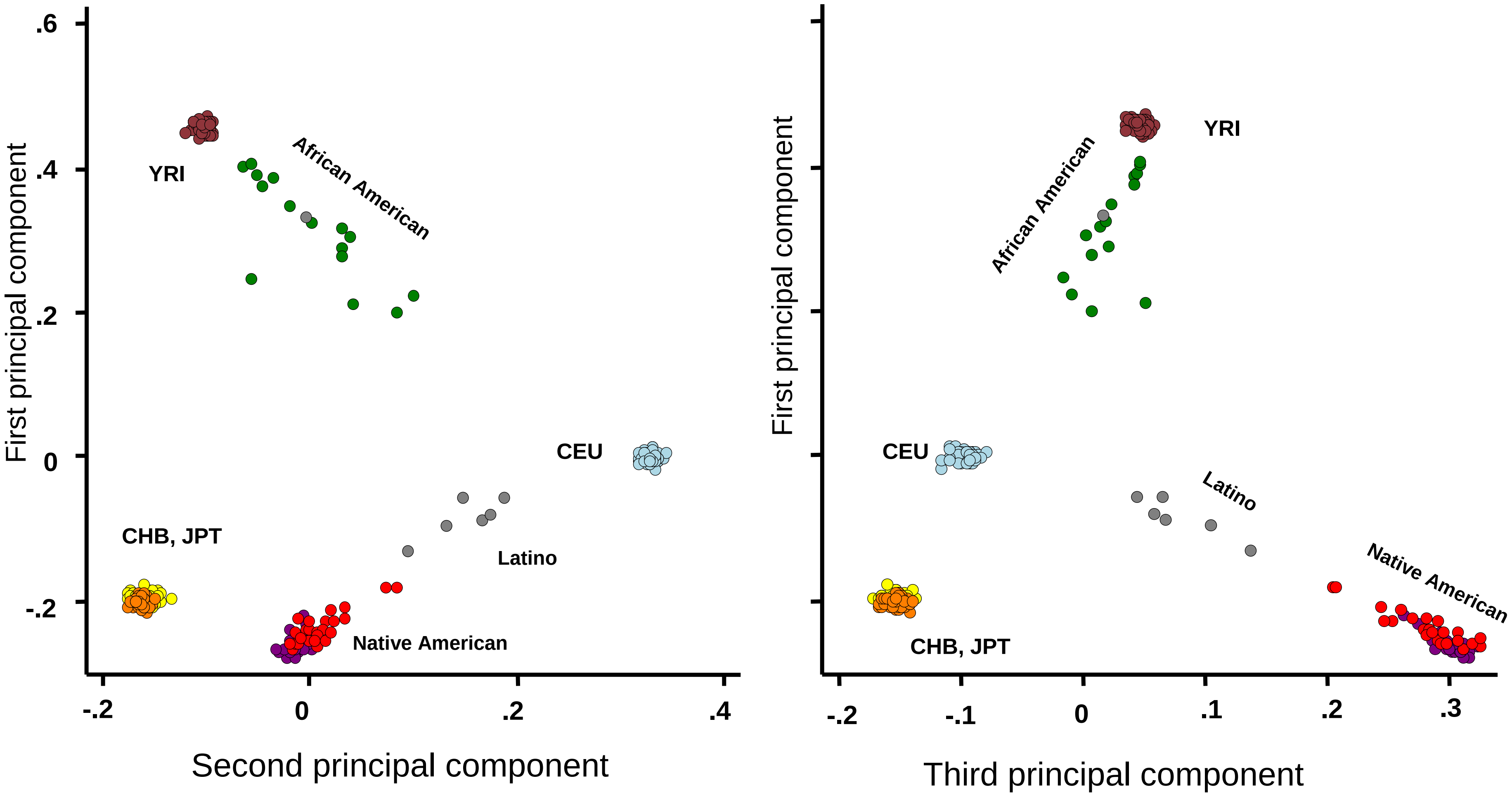
Total of 285 individuals

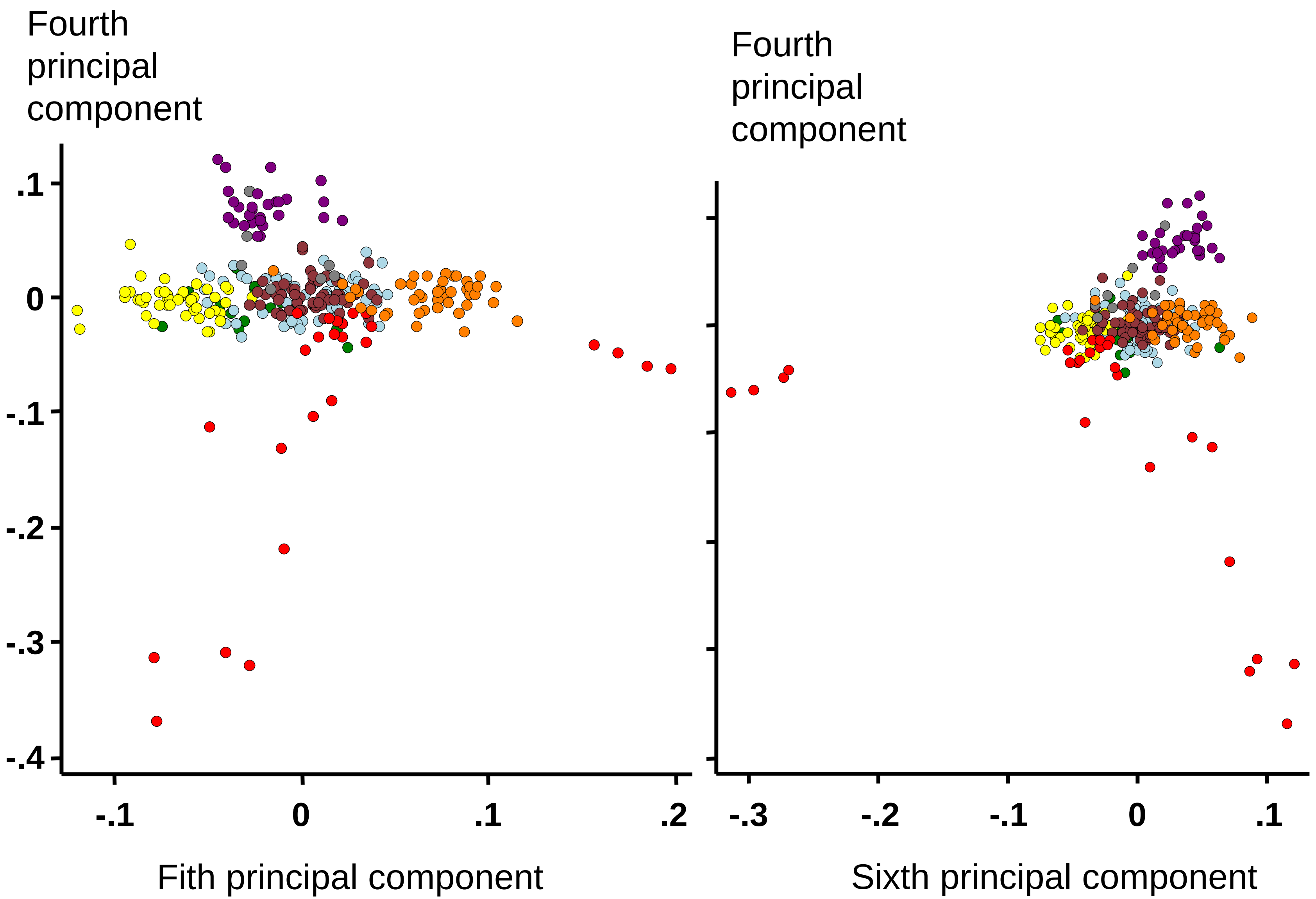
*Courtesy of X. Mao , E. Parra and M. Shriver

PCA analysis
Significance of observed principal components

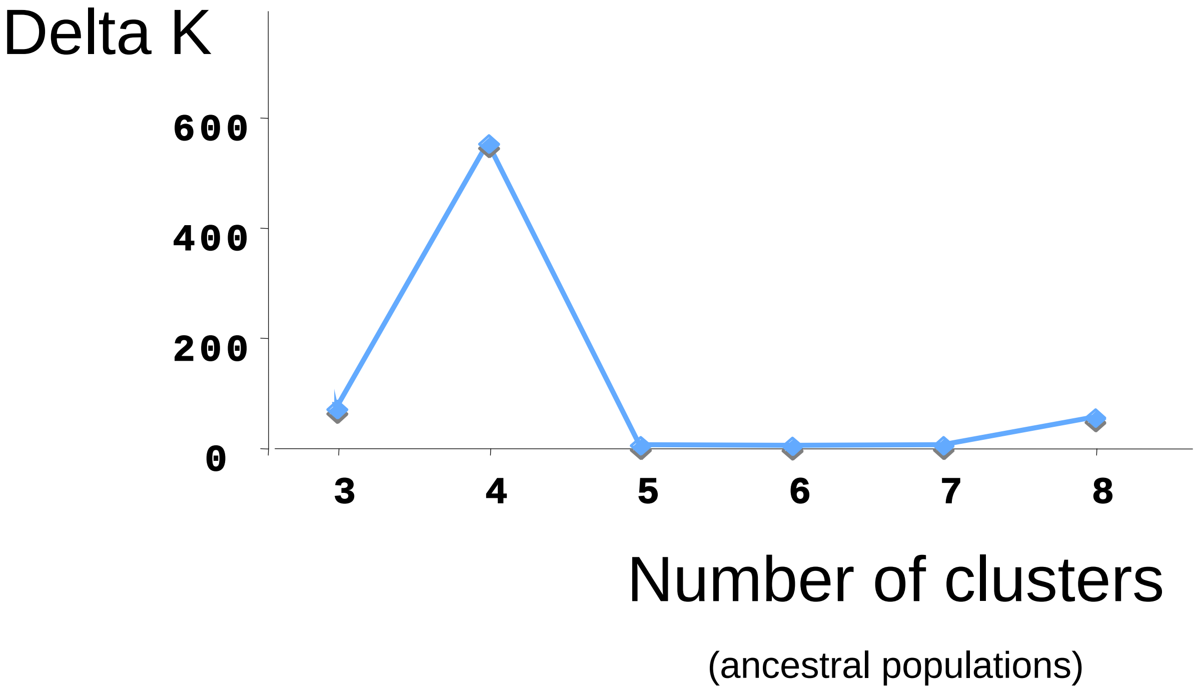
rank	eigen val	p-value
1	21.03	$< 10^{-20}$
2	9.66	$< 10^{-20}$
3	7.57	$< 10^{-20}$
4	0.80	$< 10^{-20}$
5	0.71	$< 10^{-20}$
6	0.70	$< 10^{-20}$
7	0.64	$5.1\ 10^{-9}$
8	0.57	0.5
9	0.57	0.9

First to third components



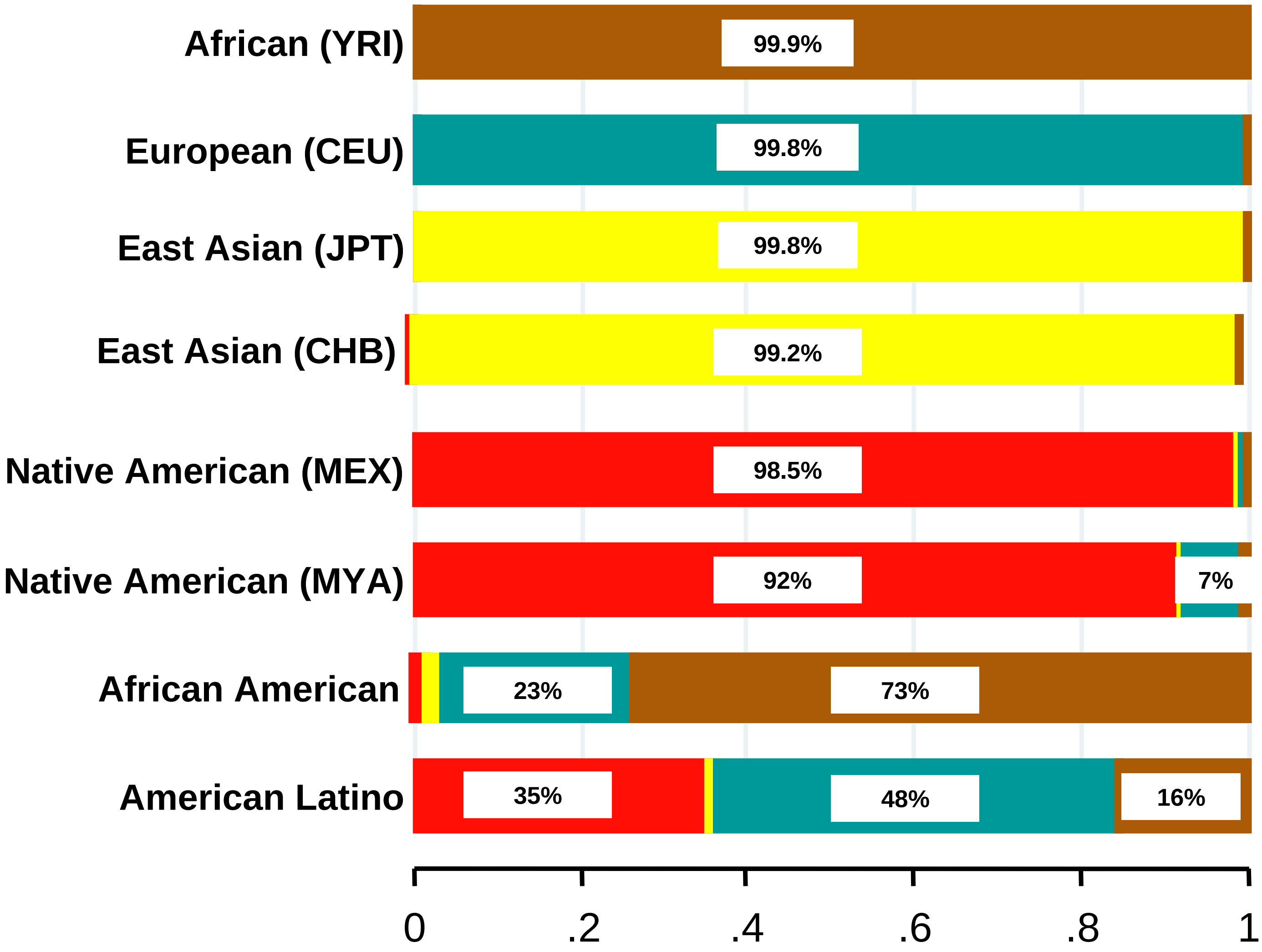

Fourth
principal
component
.1
0
-.1
-.2
-.3
-.4
-.1
0
.1
.2
Fith principal component
Fourth
principal
component
-.3
-.2
-.1
0
.1
Sixth principal component

Evaluation of the optimal number of clusters to be used by STRUCTURE for the group of 285 individuals

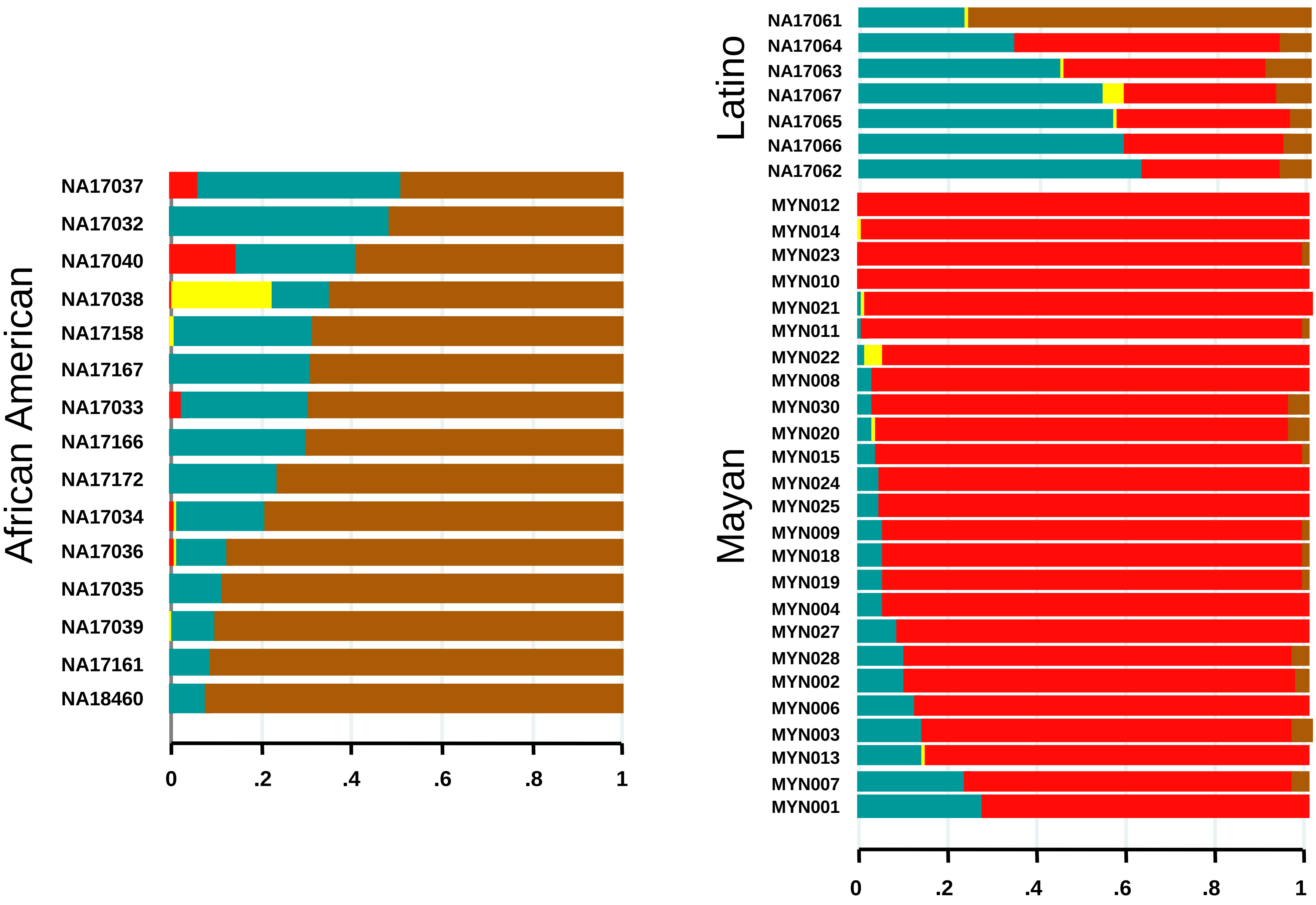


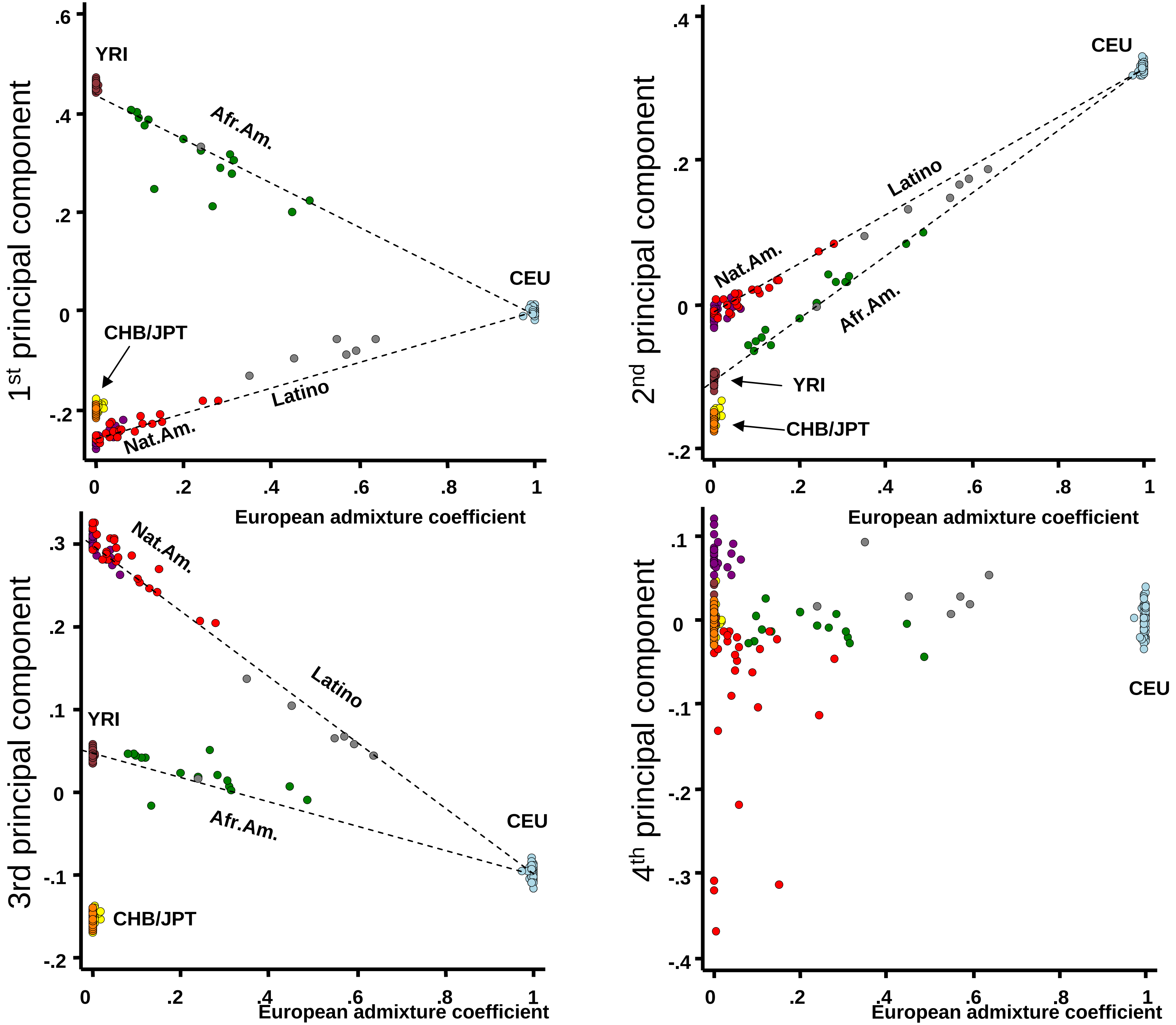
Evanno et al. Molecular Ecology 14:2611-2620, 2005

Proportion of continental origin in population samples

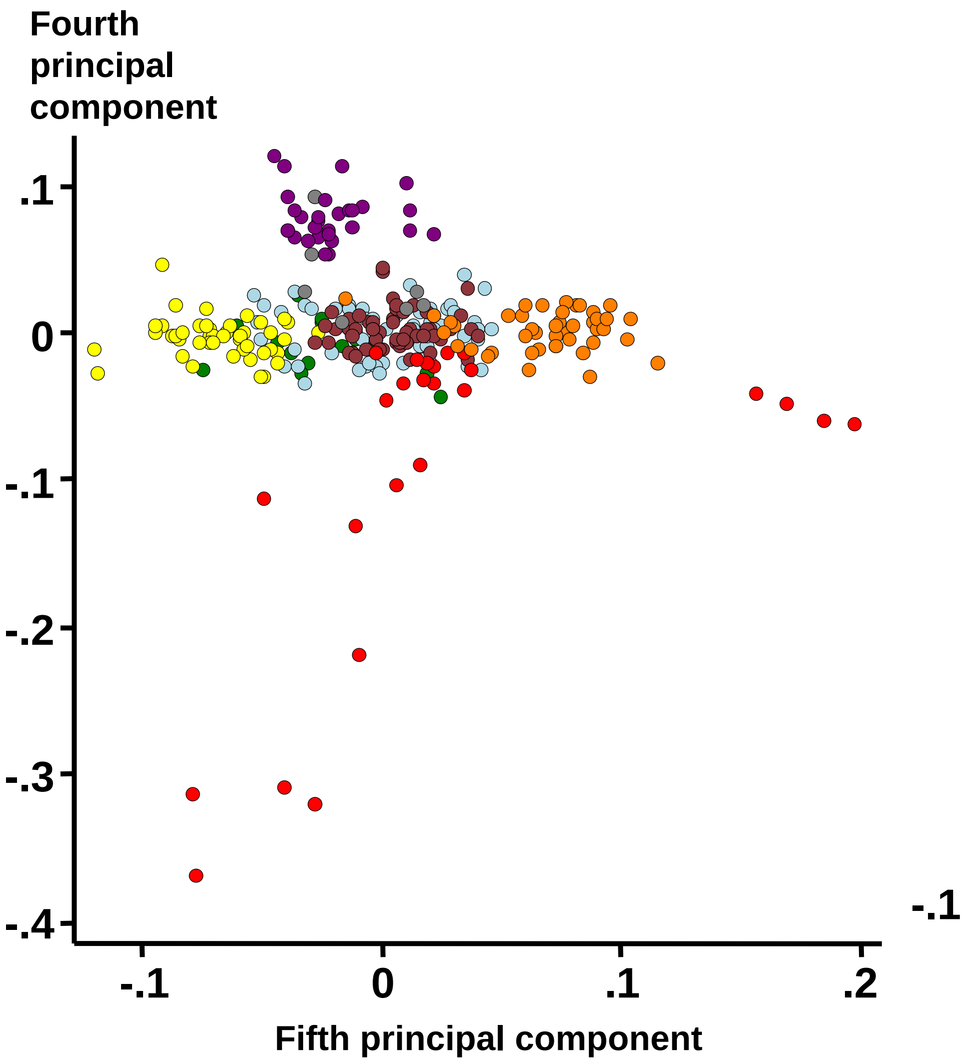


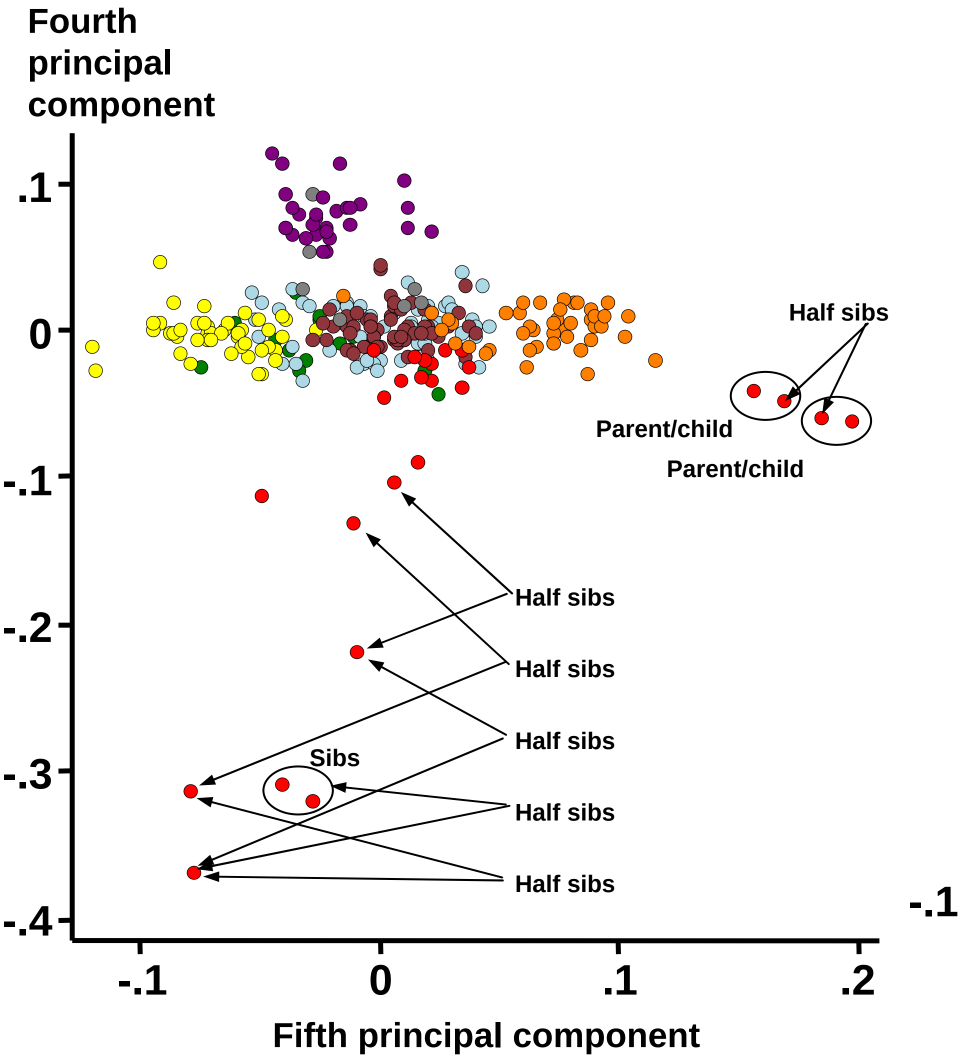
Proportion of continental origin in self identified African-American



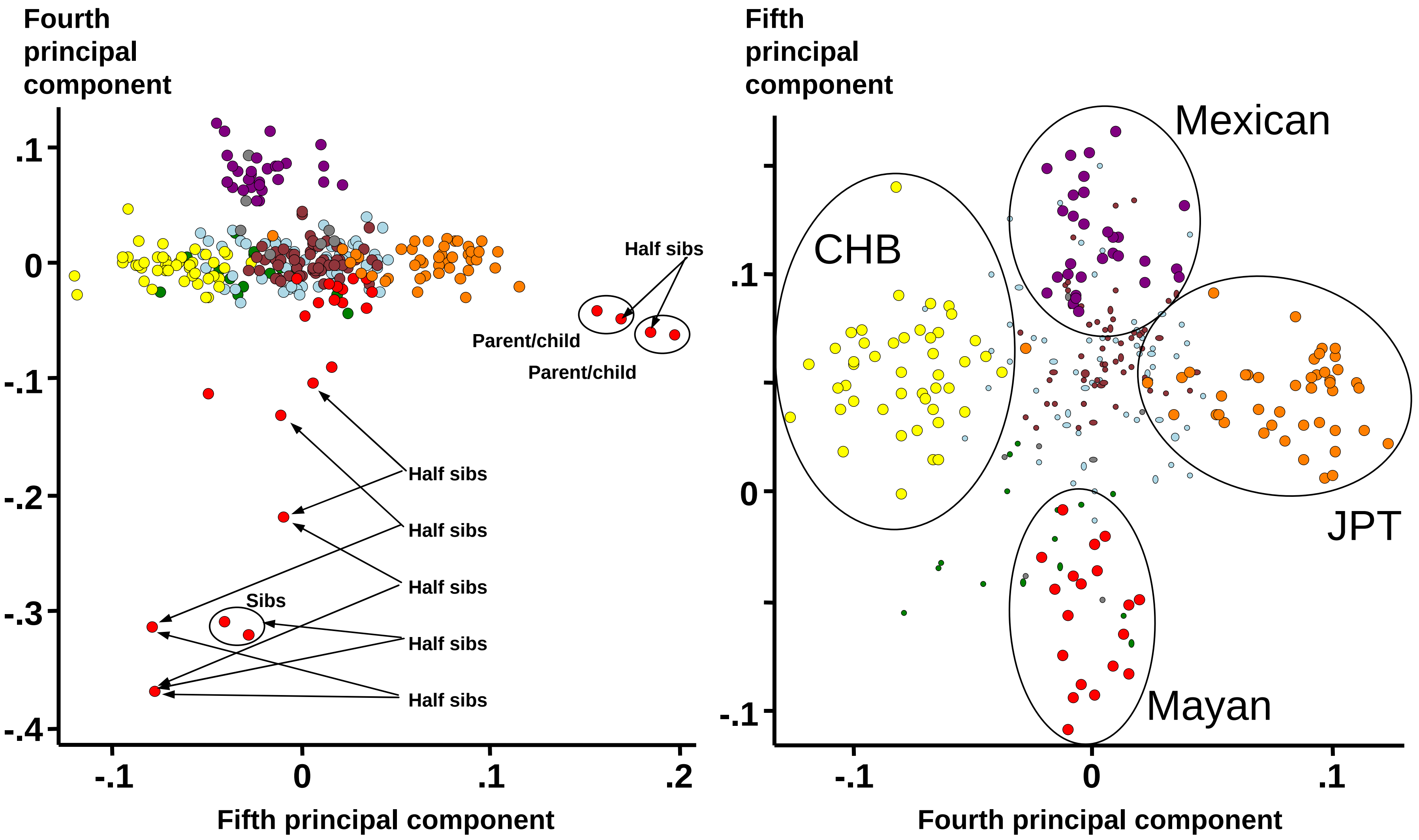


1st principal component
.6
.4
.2
0
-.2
YRI
Afr.Am.
CEU
CHB/JPT
Latino
Nat.Am.
0
.2
.4
.6
.8
1
European admixture coefficient
2nd principal component
.4
.2
0
-.2
CEU
Latino
Nat.Am.
Afr.Am.
YRI
CHB/JPT
0
.2
.4
.6
.8
1
European admixture coefficient
3rd principal component
.3
.2
.1
0
-.1
-.2
Nat.Am.
Latino
YRI
Afr.Am.
CEU
CHB/JPT
0
.2
.4
.6
.8
1
European admixture coefficient
4th principal component
.1
0
-.1
-.2
-.3
-.4
CEU
0
.2
.4
.6
.8
1
European admixture coefficient


Fourth principal component
.1
0
-.1
-.2
-.3
-.4
-.1
0
.1
.2
Fifth principal component
-.1


Fourth principal component
.1
0
-.1
-.2
-.3
-.4
-.1
0
.1
.2
Fifth principal component
-.1
Half sibs
Parent/child
Parent/child
Half sibs
Half sibs
Half sibs
Sibs
Half sibs
Half sibs

Influence of relatedness on principal component analysis



Conclusions on model population

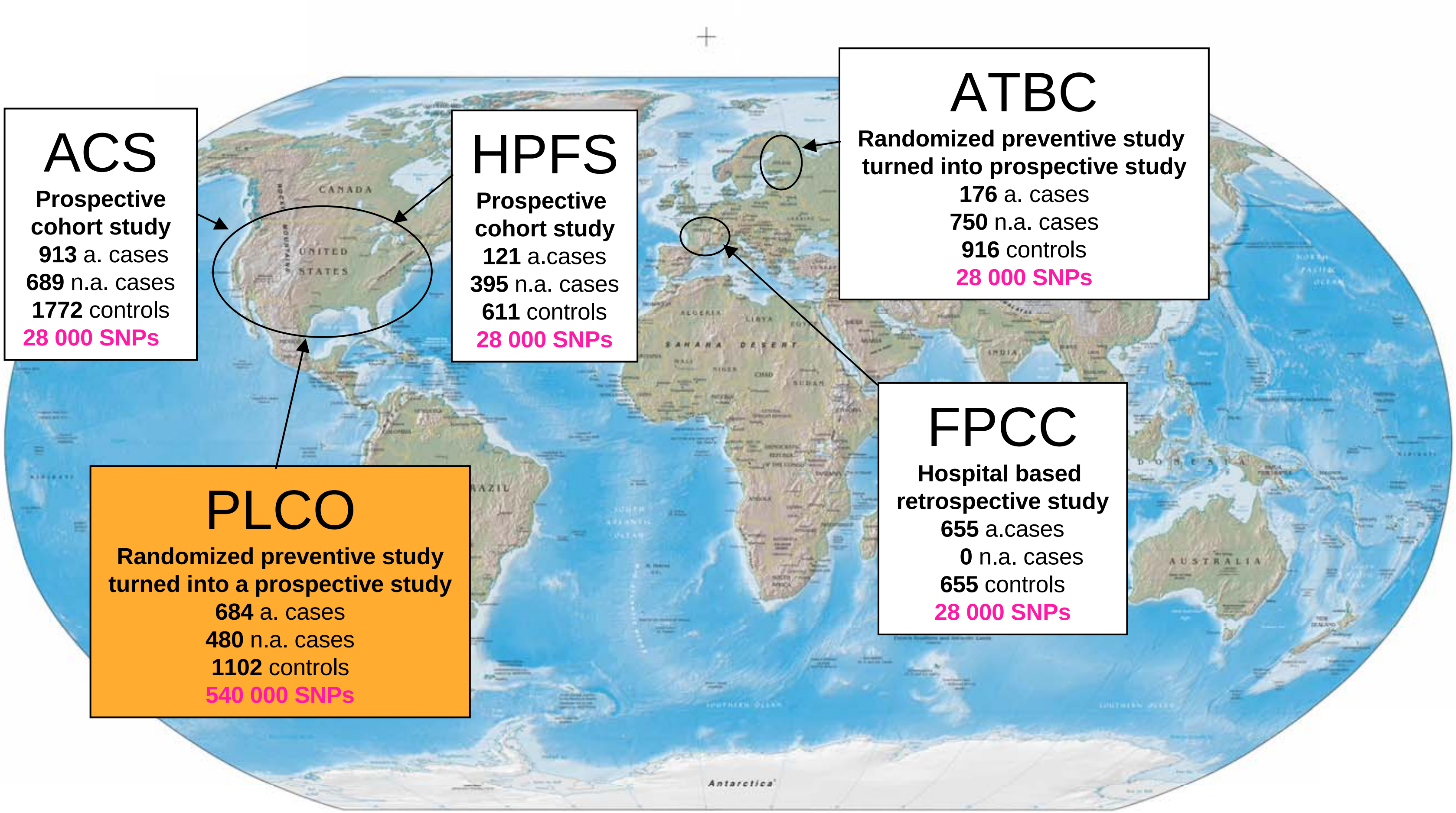
- PC analysis with 10 000 uncorrelated SNPs reliably identifies continental subpopulations

- Cryptic relationships may significantly interfere with PC analysis.

- Good correlation between the admixture coefficient evaluated by the program Structure and the components along the major principal directions

CGEMS
Prostate cancer scan

Total :agressive cases 2549, non-aggressive cases 2314, controls 5056.



ACS

Prospective cohort study

913 a. cases
689 n.a. cases
1772 controls
28 000 SNPs

HPFS

Prospective cohort study

121 a.cases
395 n.a. cases
611 controls
28 000 SNPs

ATBC

Randomized preventive study turned into prospective study

176 a. cases
750 n.a. cases
916 controls
28 000 SNPs

FPCC

Hospital based retrospective study

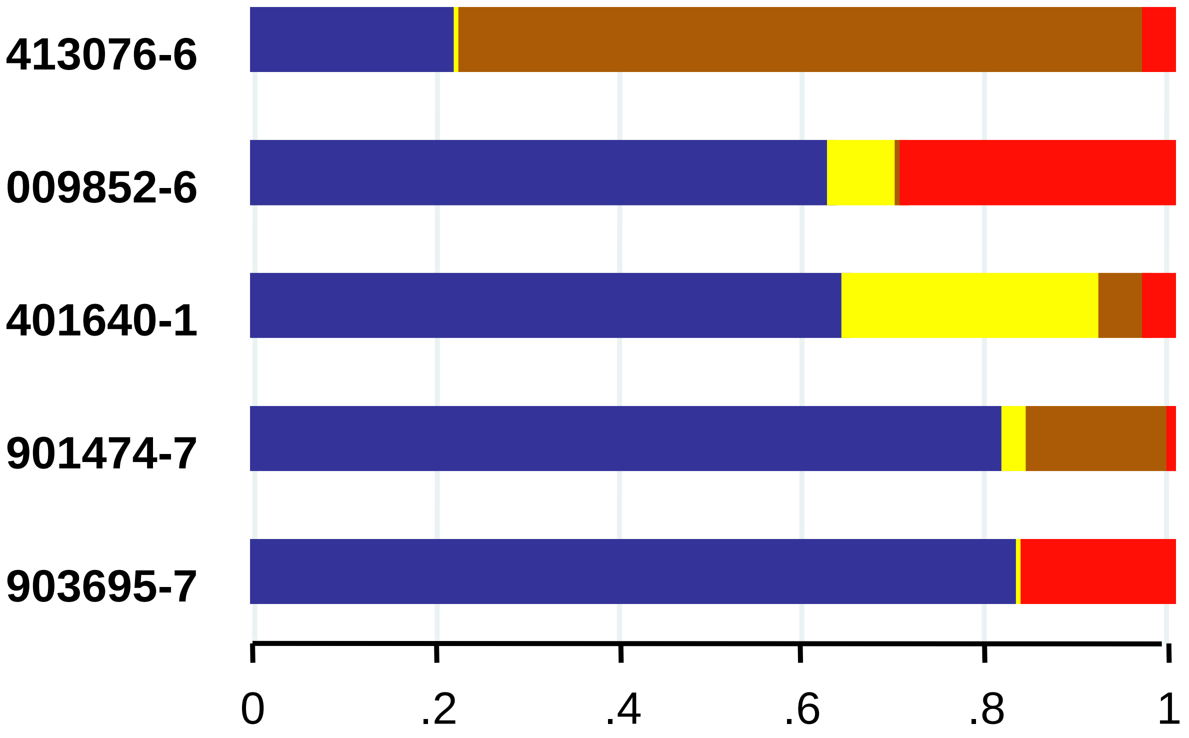
655 a.cases
0 n.a. cases
655 controls
28 000 SNPs

PLCO

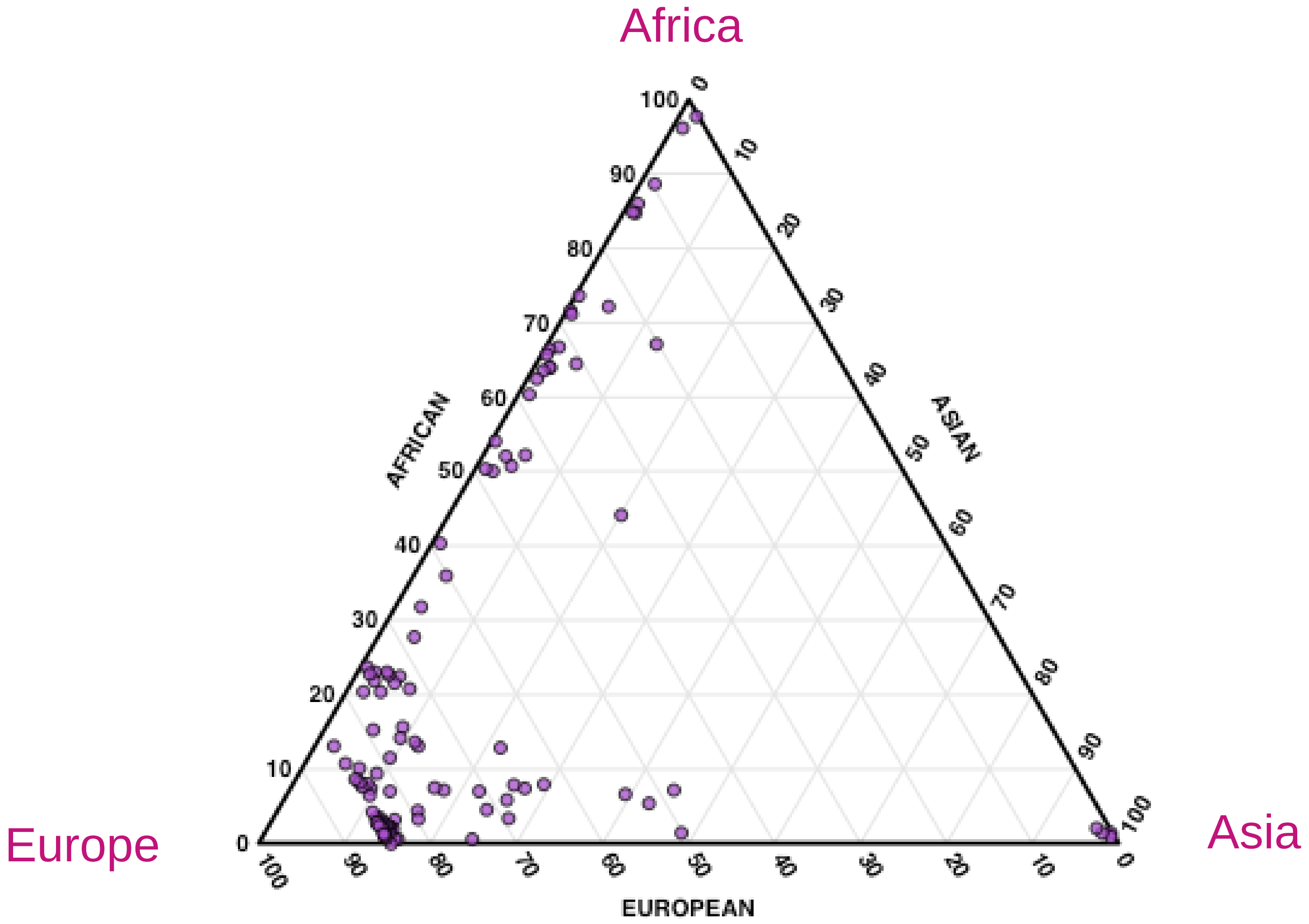
Randomized preventive study turned into a prospective study

684 a. cases
480 n.a. cases
1102 controls
540 000 SNPs

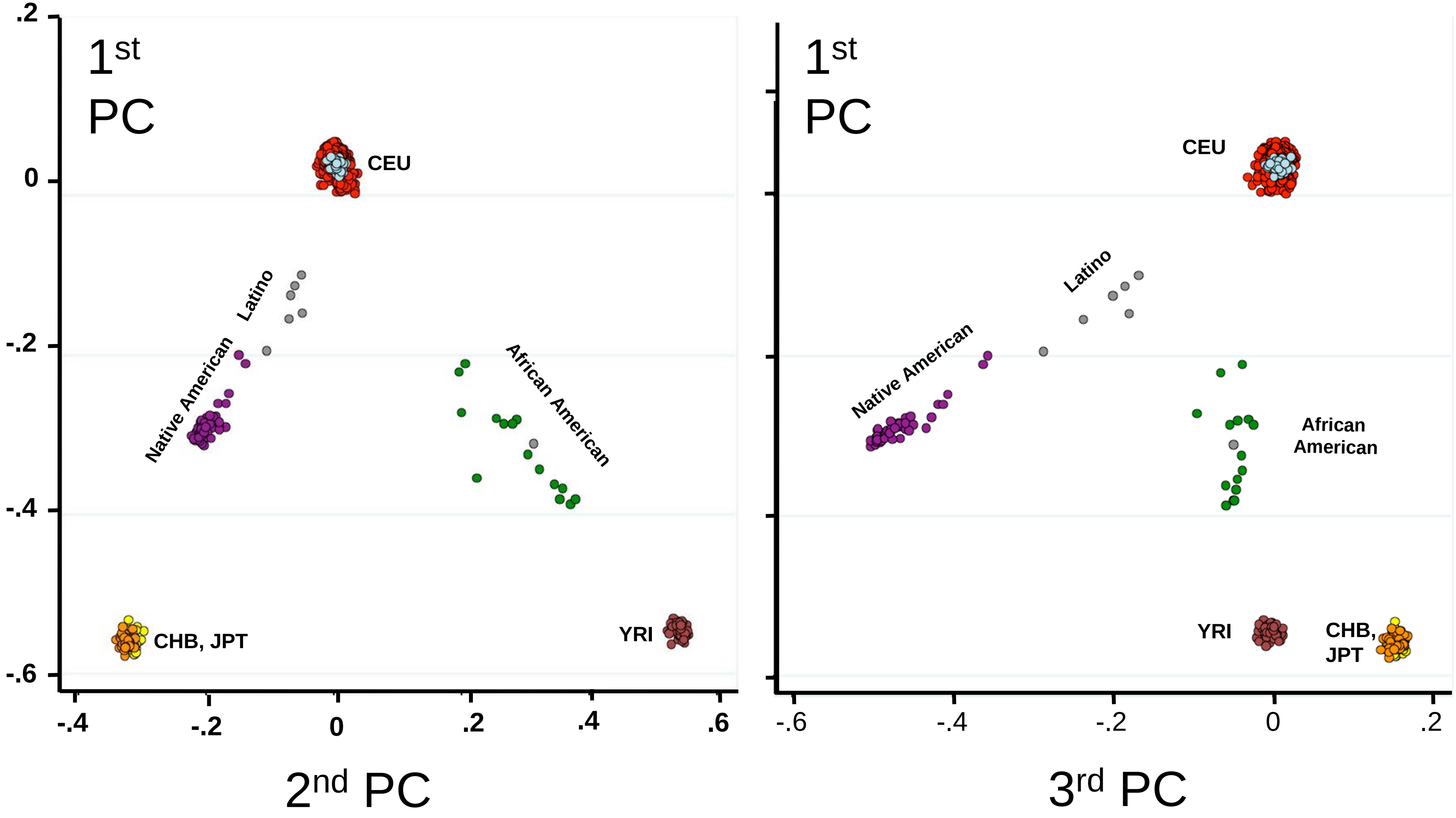
DNAs with large admixture coefficient in the PLCO study (all controls)



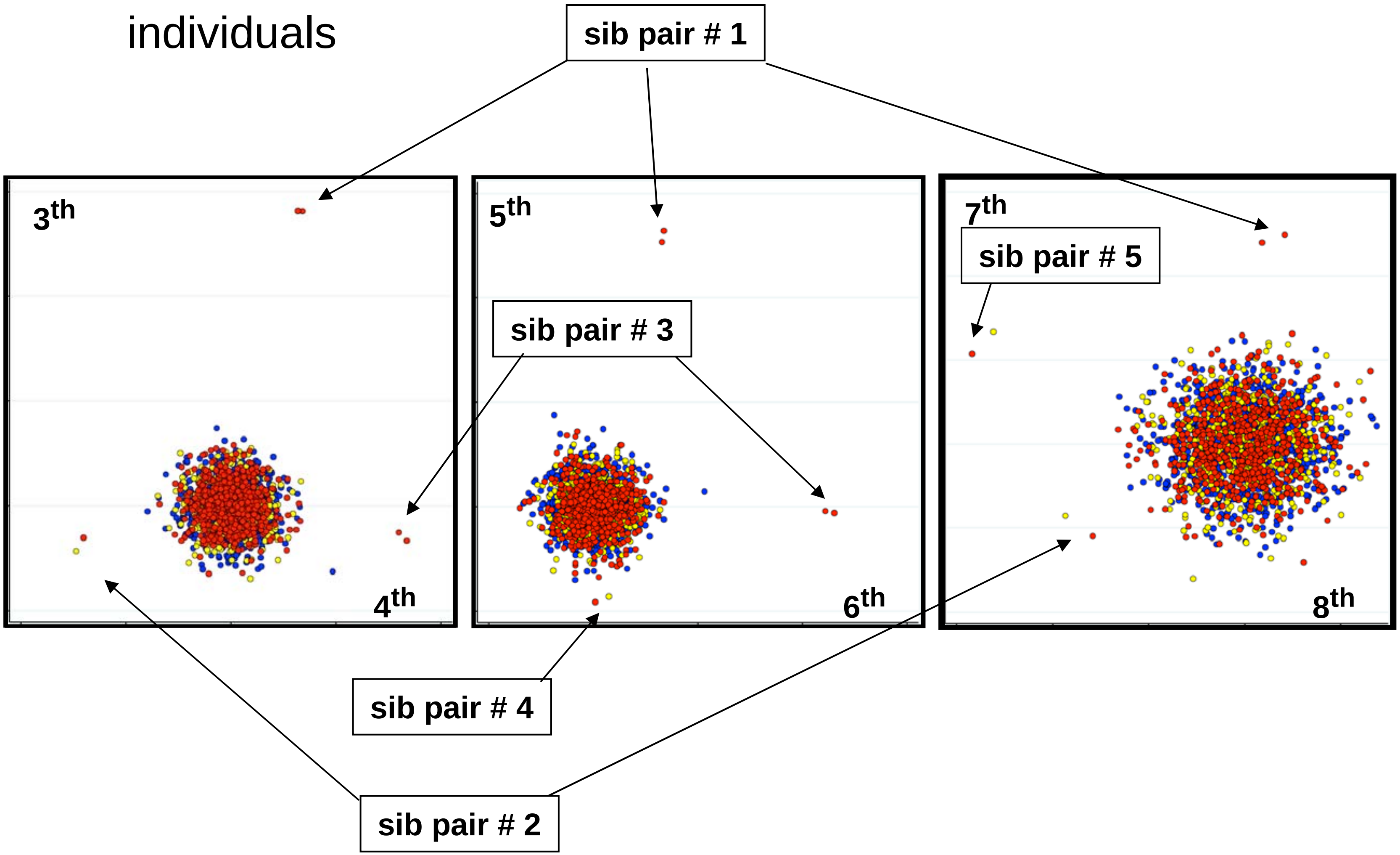
Admixture coefficients of 109 DNAs with less than 85% European origin found in CGEMS prostate cancer follow-up studies



PCA analysis of the cases and controls of the CGEMS-PLCO study



PCA analysis of PLCO cases and controls after removal of admixed individuals



sib pair # 1

sib pair # 5

sib pair # 3

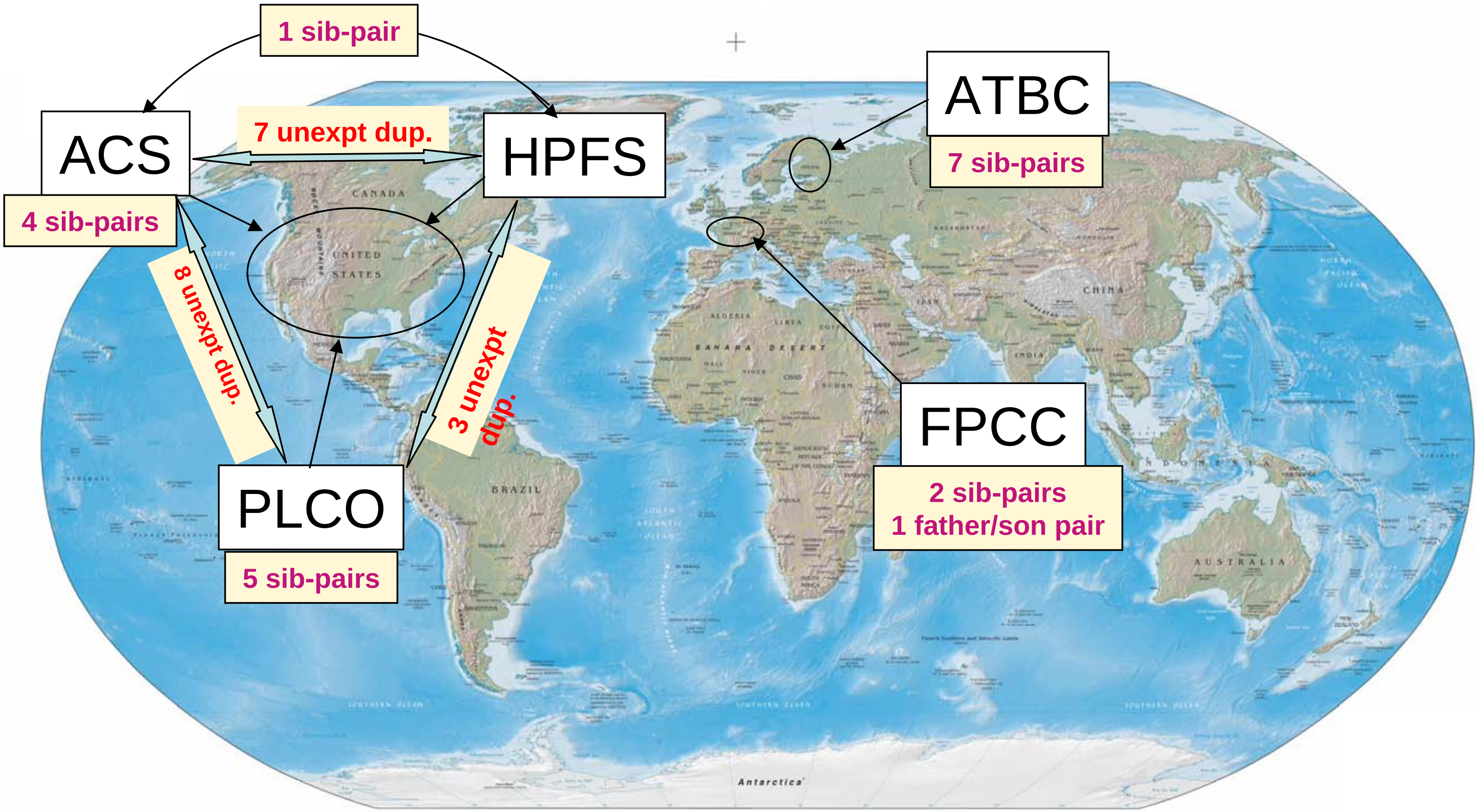
sib pair # 4

sib pair # 2

Unexpected relatedness

18 unexpected duplicates and

20 pairs of 1st degree relatives



1 sib-pair

ACS

7 unexpt dup.

HPFS

ATBC

7 sib-pairs

4 sib-pairs

8 unexpt dup.

3 unexpt dup.

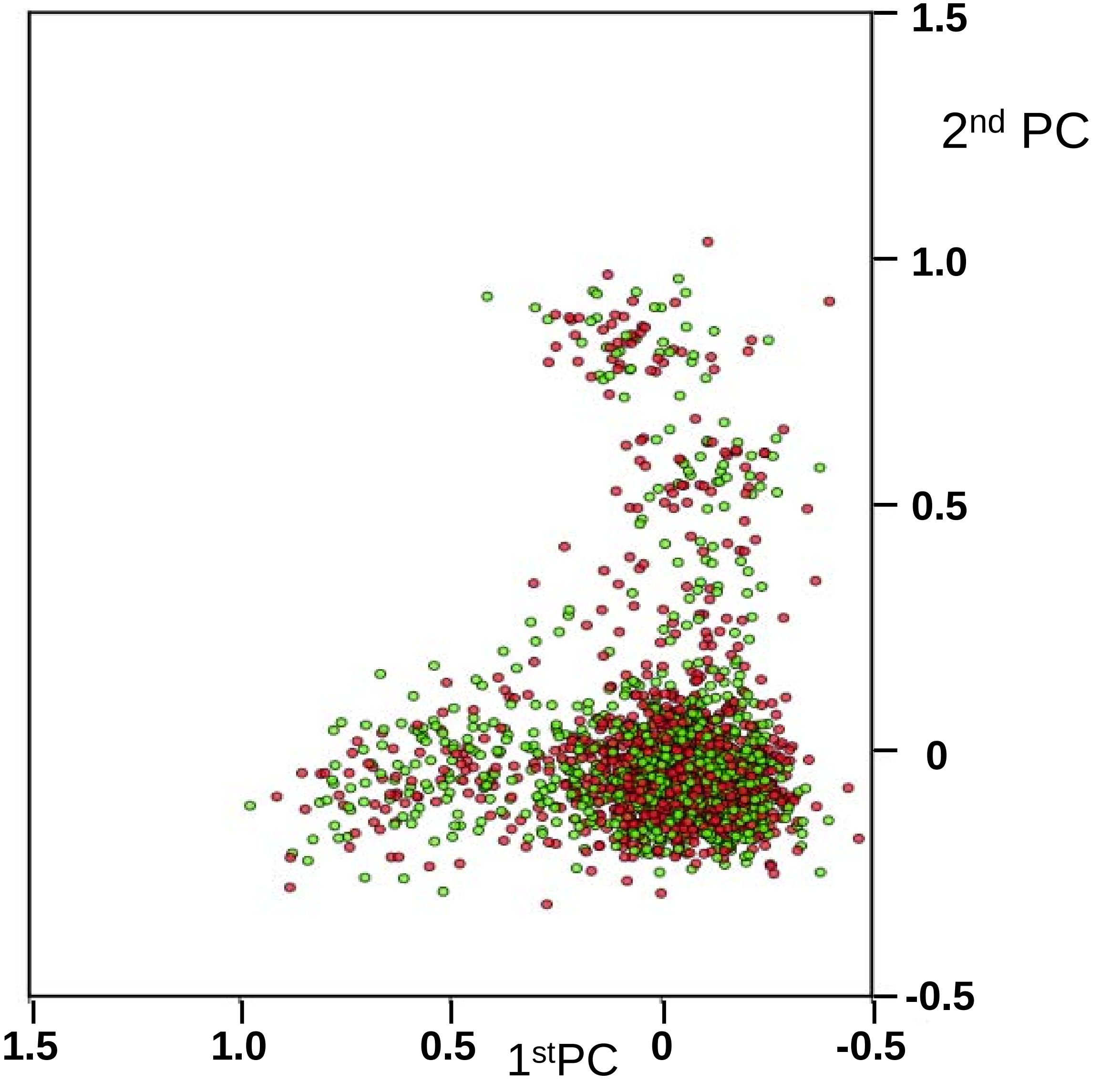
FPCC

2 sib-pairs
1 father/son pair

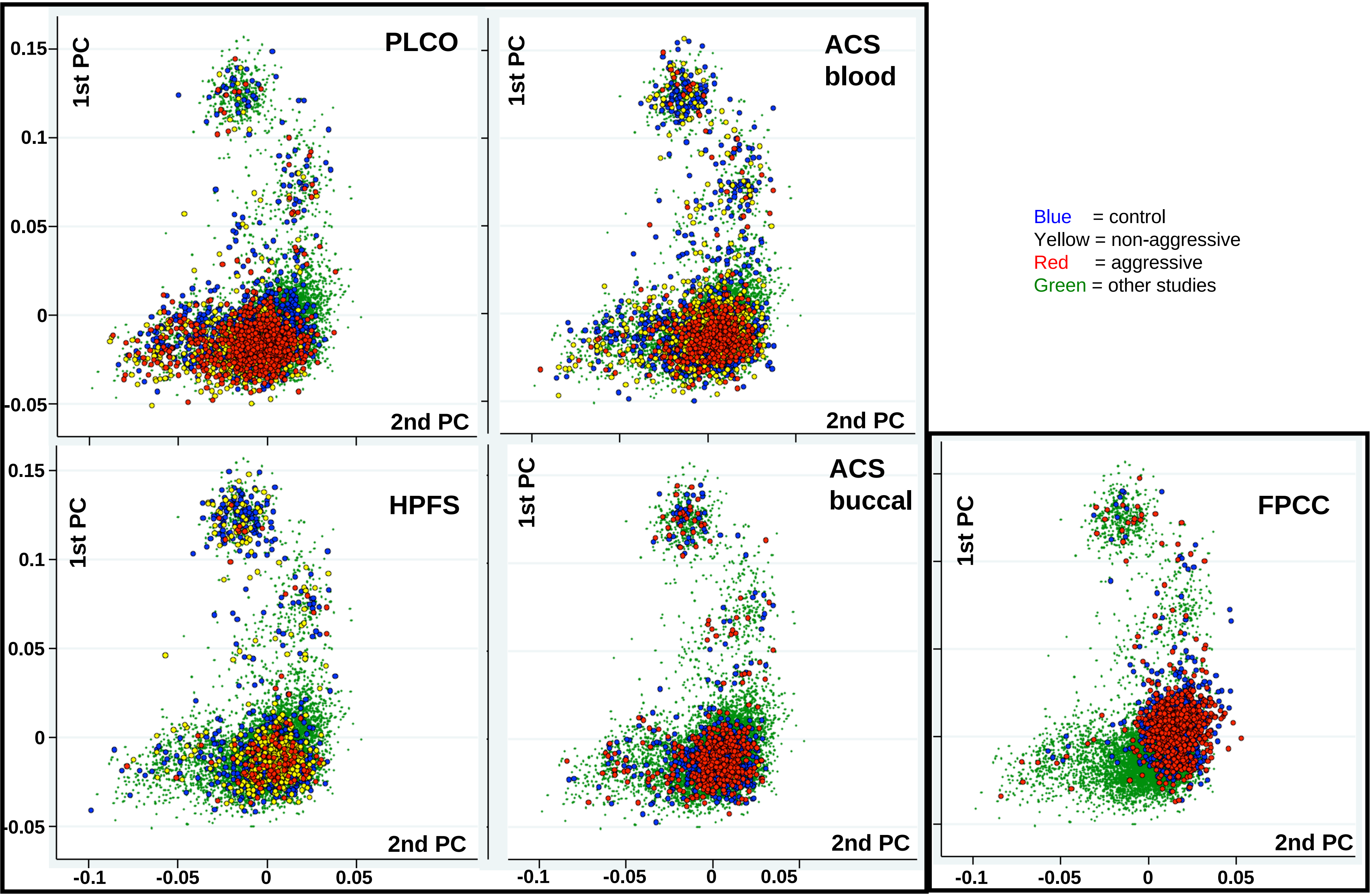
PLCO

5 sib-pairs

PCA analysis of PLCO cases and controls after removal of admixed individuals and one member of each first degree relative pairs



PCA on pooled US-based and French Studies.



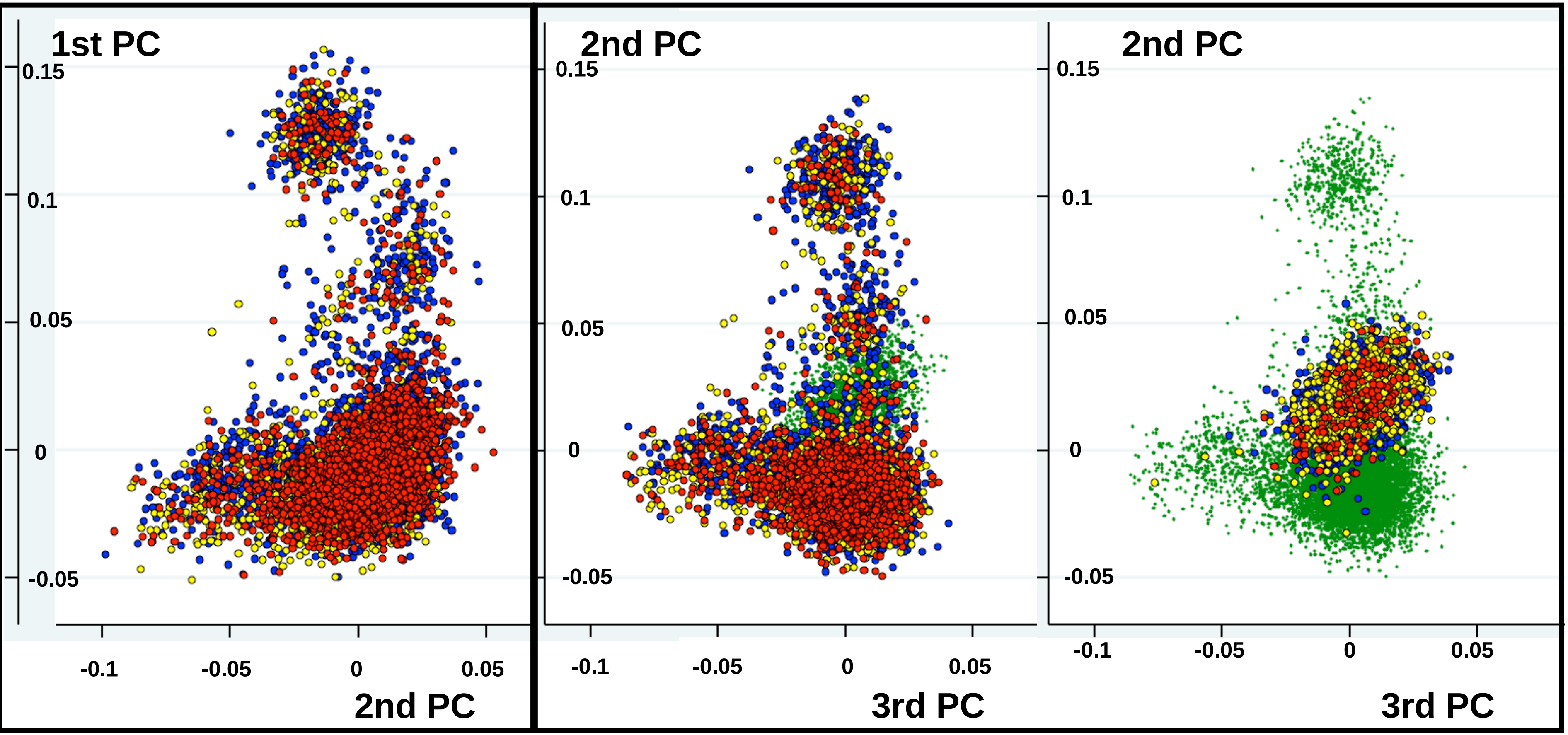
Blue = control
Yellow = non-aggressive
Red = aggressive
Green = other studies

PCA on pooled
US and French studies only .
- 1st and 2nd PC -

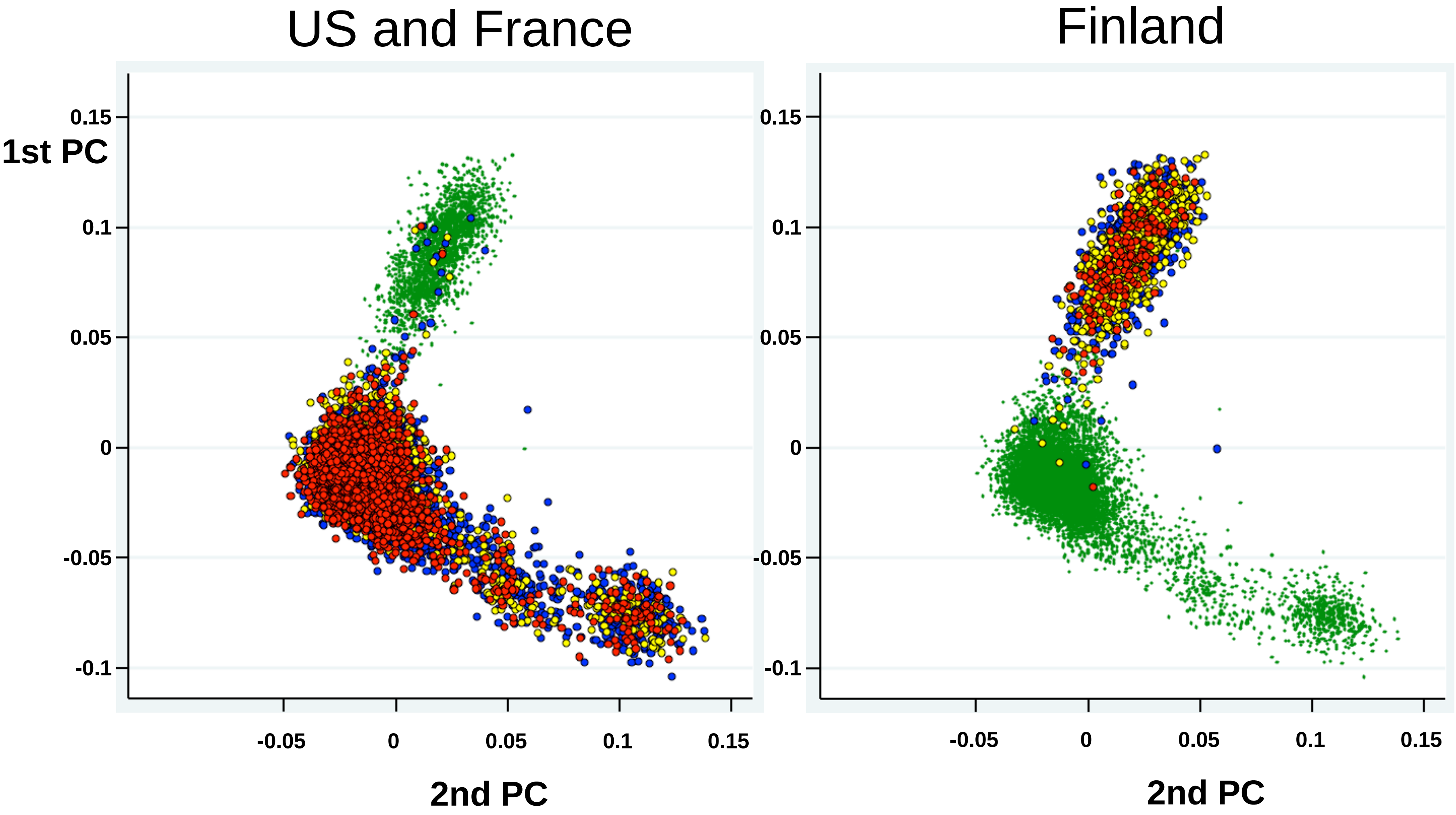
PCA on all Studies.
2nd and 3rd PC

US and France

Finland

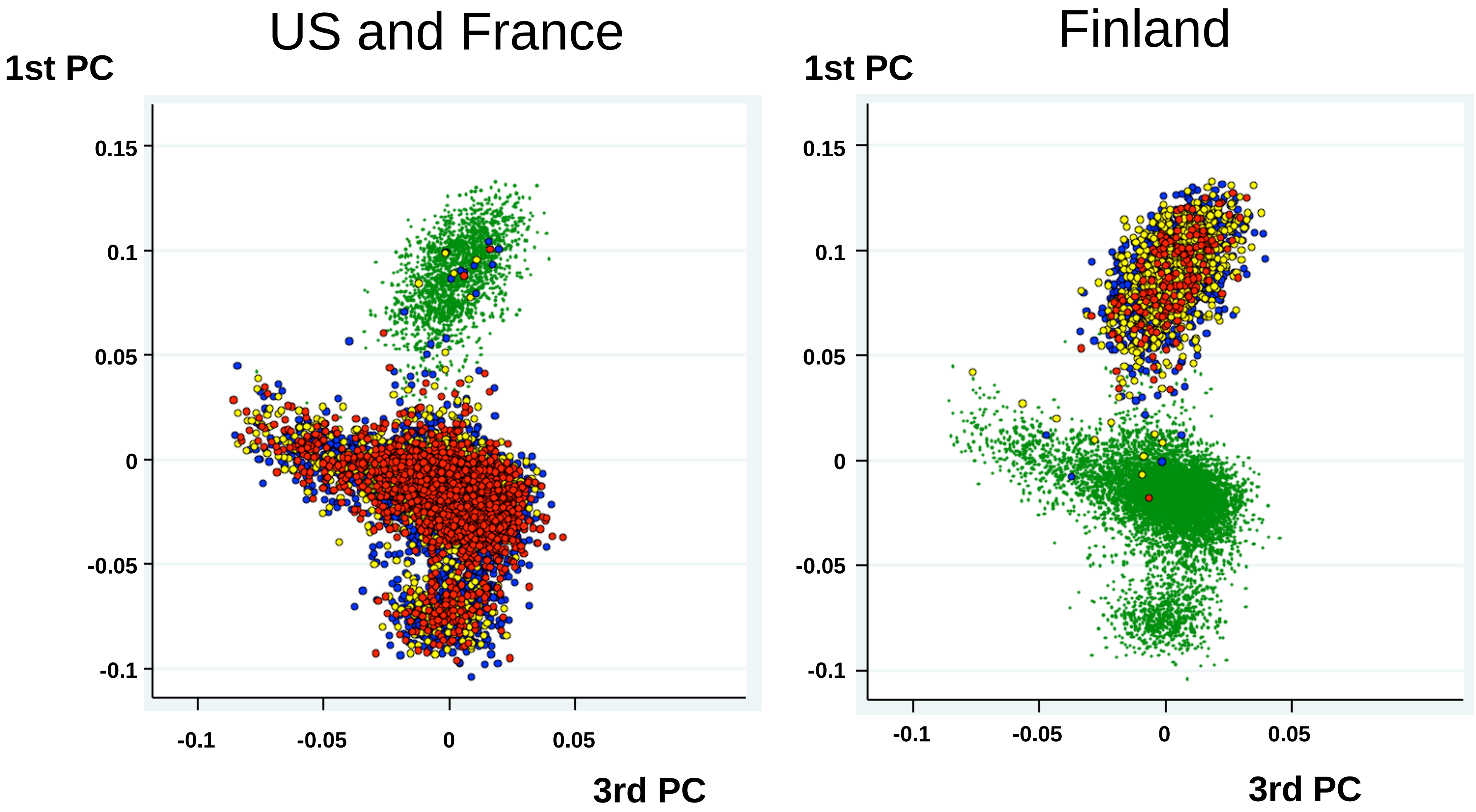


PCA on all Studies.
1st and 2nd PC



Blue = control
Yellow = non-aggressive
Red = aggressive
Green = other studies

PCA performed on all Studies.
1st and 3rd PC

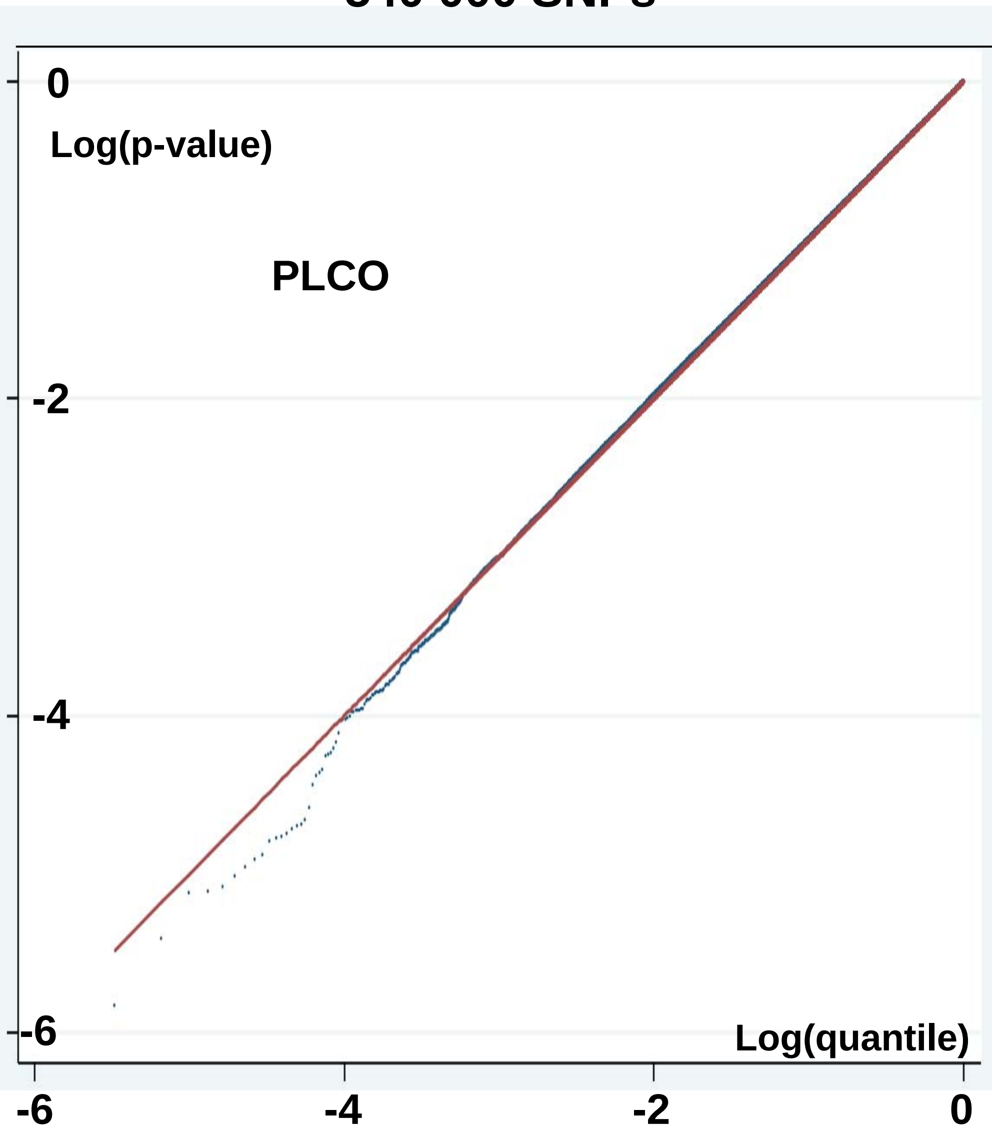

US and France

Finland

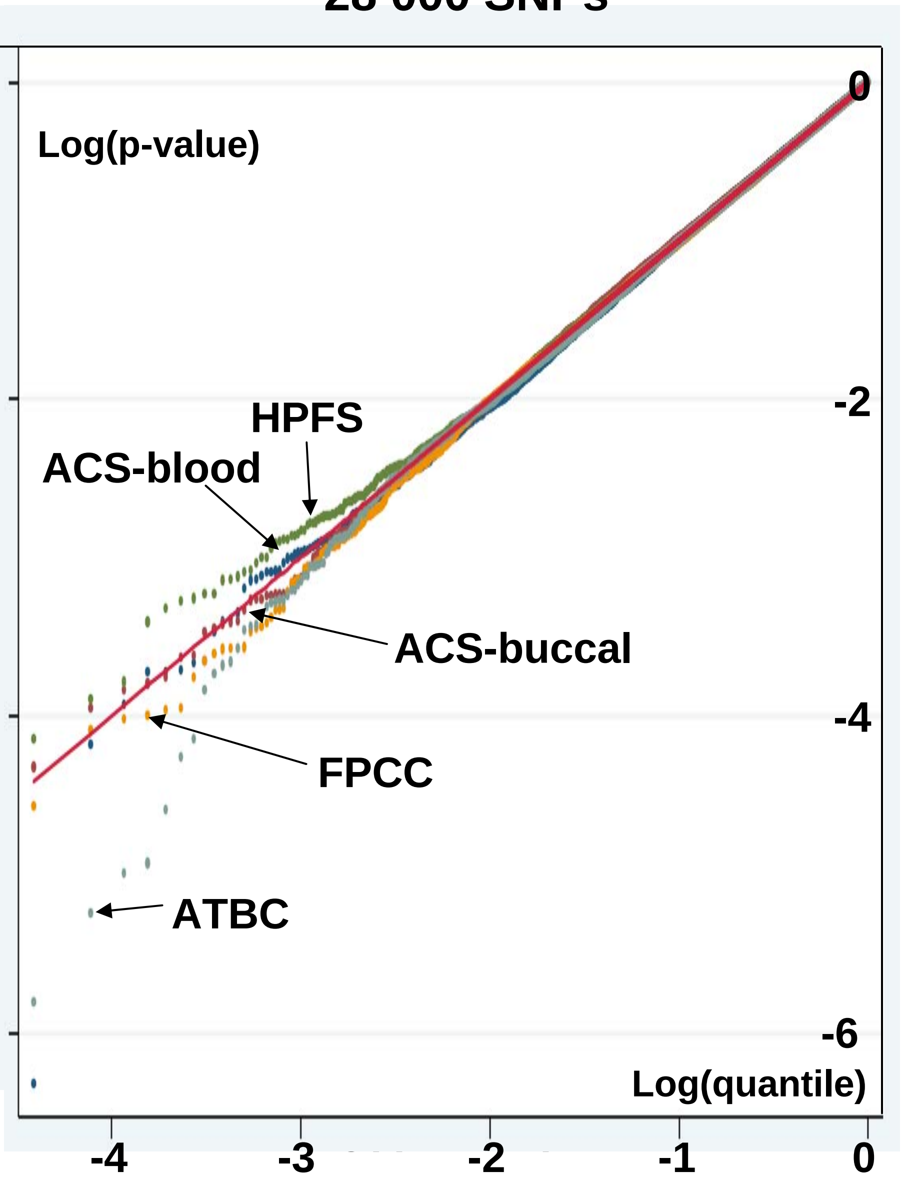
Blue = control
Yellow = non-aggressive
Red = aggressive
Green = other studies

Individual QQ plots for each study

Initial Genome Wide Screen
540 000 SNPs

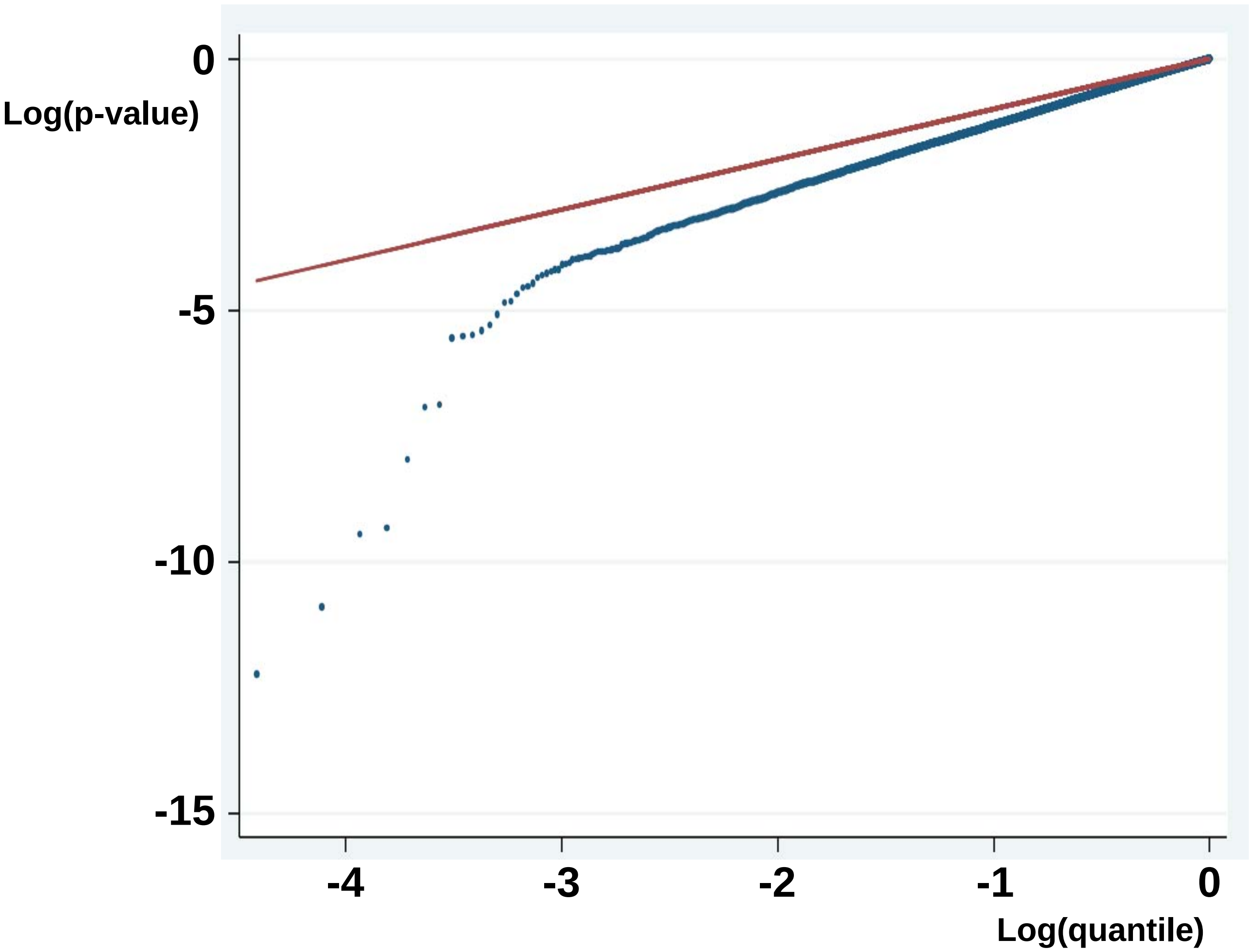


Follow-up studies
28 000 SNPs

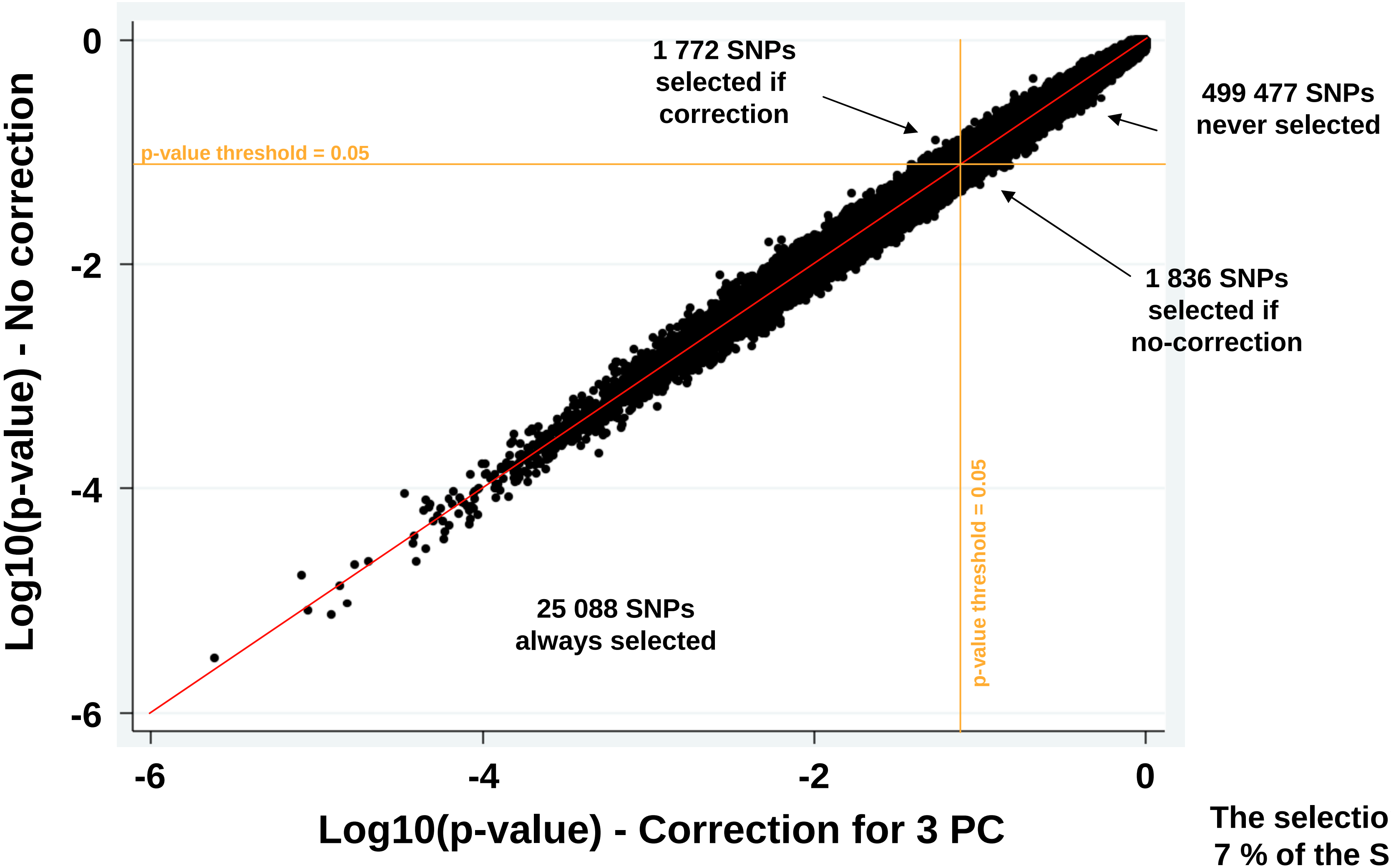


QQ plot on the combined analysis of all studies

(SNPs from the 8q24 region have been removed)



Change in the selection outcome when 3 PCs are taken into account for population stratification in the PLCO study



1 772 SNPs selected if correction

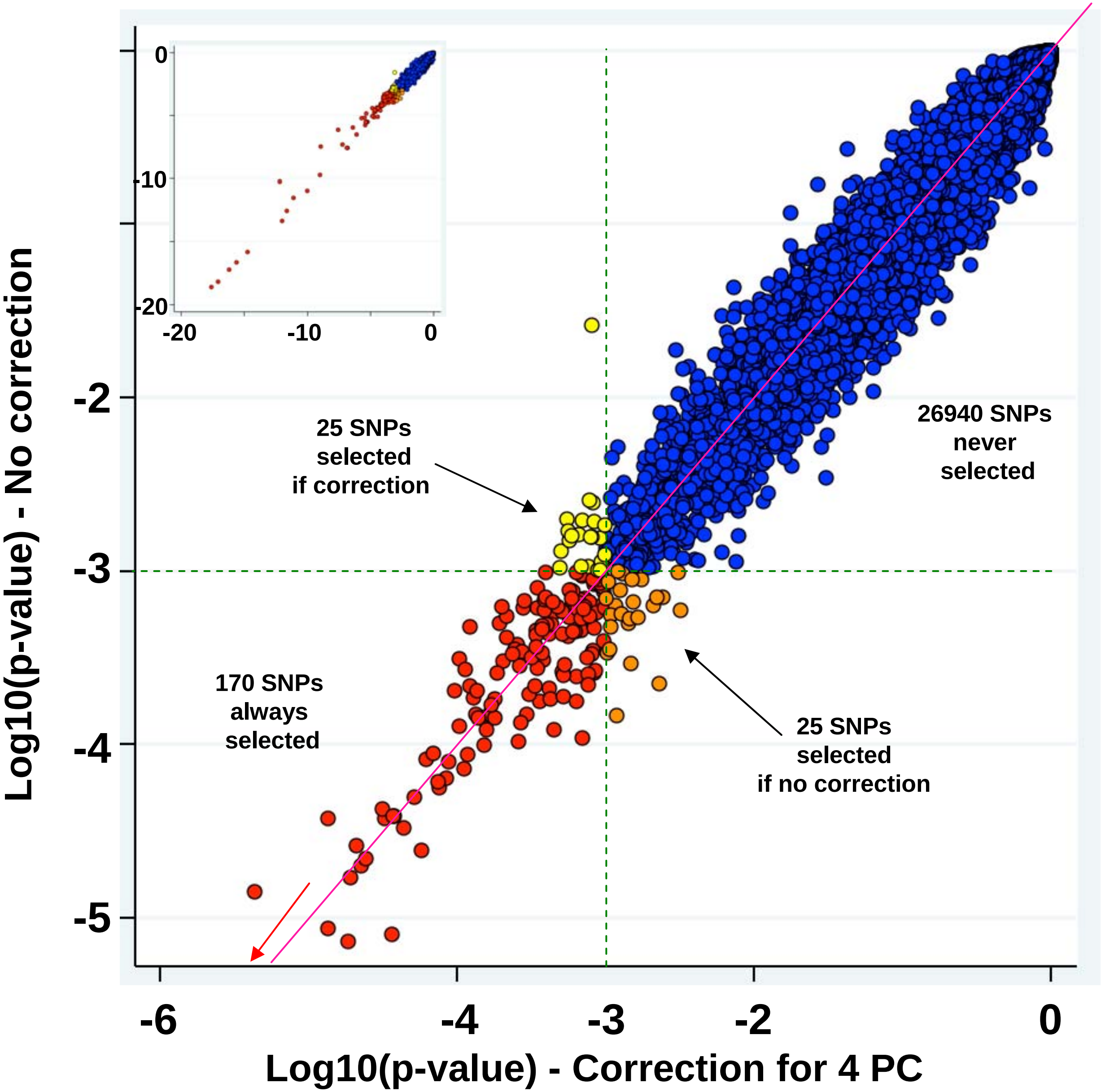
499 477 SNPs never selected

1 836 SNPs selected if no-correction

25 088 SNPs always selected

The selection of 7 % of the SNPs depends on the PC correction

Change in the selection outcome when 4 PCs are taken into account in the joint analysis of all studies



25 SNPs selected if correction

26940 SNPs never selected

170 SNPs always selected

25 SNPs selected if no correction

The selection of 13 % of the SNPs depends on the PC correction

Conclusion

- Search for population structure in the CGEMS prostate cancer study revealed :
  - Individuals that did not meet the inclusion criteria :
    - 1.1 % individuals with less than 85% European origin.
    - 18 individulas that participated in two independent studies.
    - 20 pairs of first degree relatives.
  - A very significant population structure in the combined cases and controls groups.

- Accounting for population structure changed the status of :
  - 7% of the □□□SNP to be taken from the intital genome wide scan to the first follow-up study.
  - 13% of the SNPs to be taken from the first follow-up study to the second follow-up study.

Acknowledgements

- **NCI**

Stephen Chanock
Zhaoming Wang
Kai Yu
Meredith Yeager
Kevin Jacobs
Robert Welch
Robert Hoover
Joseph Fraumeni
Sholom Wacholder
Nilanjan Chatterjee
Daniela Gerhard
Richard Hayes
Margaret Tucker
Marianne Rivera-Silva

HSPH, Boston

David Hunter
Peter Kraft

ACS, Atlanta

Heather Feigelson
Carmen Rodriguez
Eugene Calle
Michael Thun

Wash. U., St Louis

Gerald Andriole

CeRePP, France

Olivier Cussenot
Geraldine Cancel-Tassin
Antoine Valeri

NPHI, Finland

Jarmo Virtamo

Penn State U.

Xianyun Mao
Esteban Parra
Mark Shriver