Gene Mirror Geography within Europe

Darshan Barhate

Key Questions

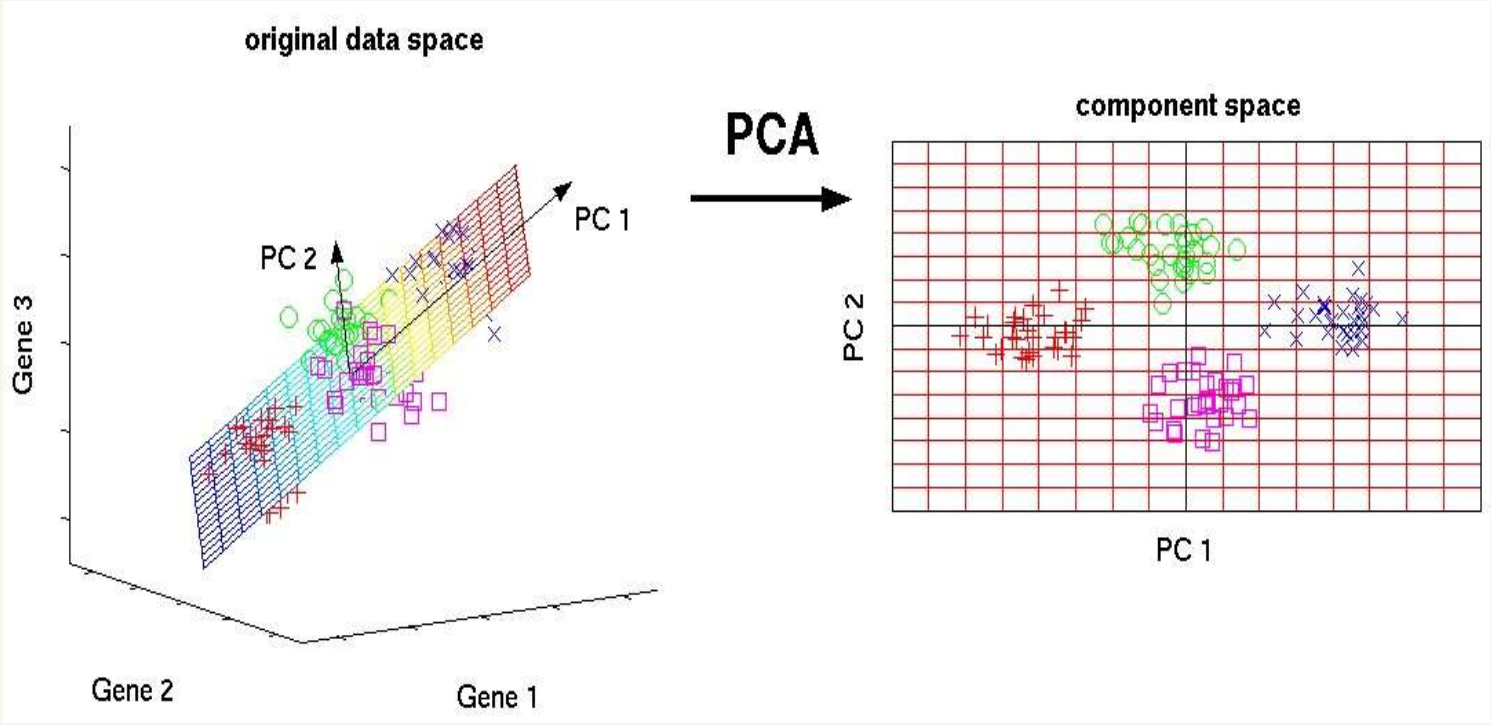
- Do populations within continental regions exist as discrete genetic clusters or continuum?
- Can one assign an individual to a geographic location based on genetic information alone ?

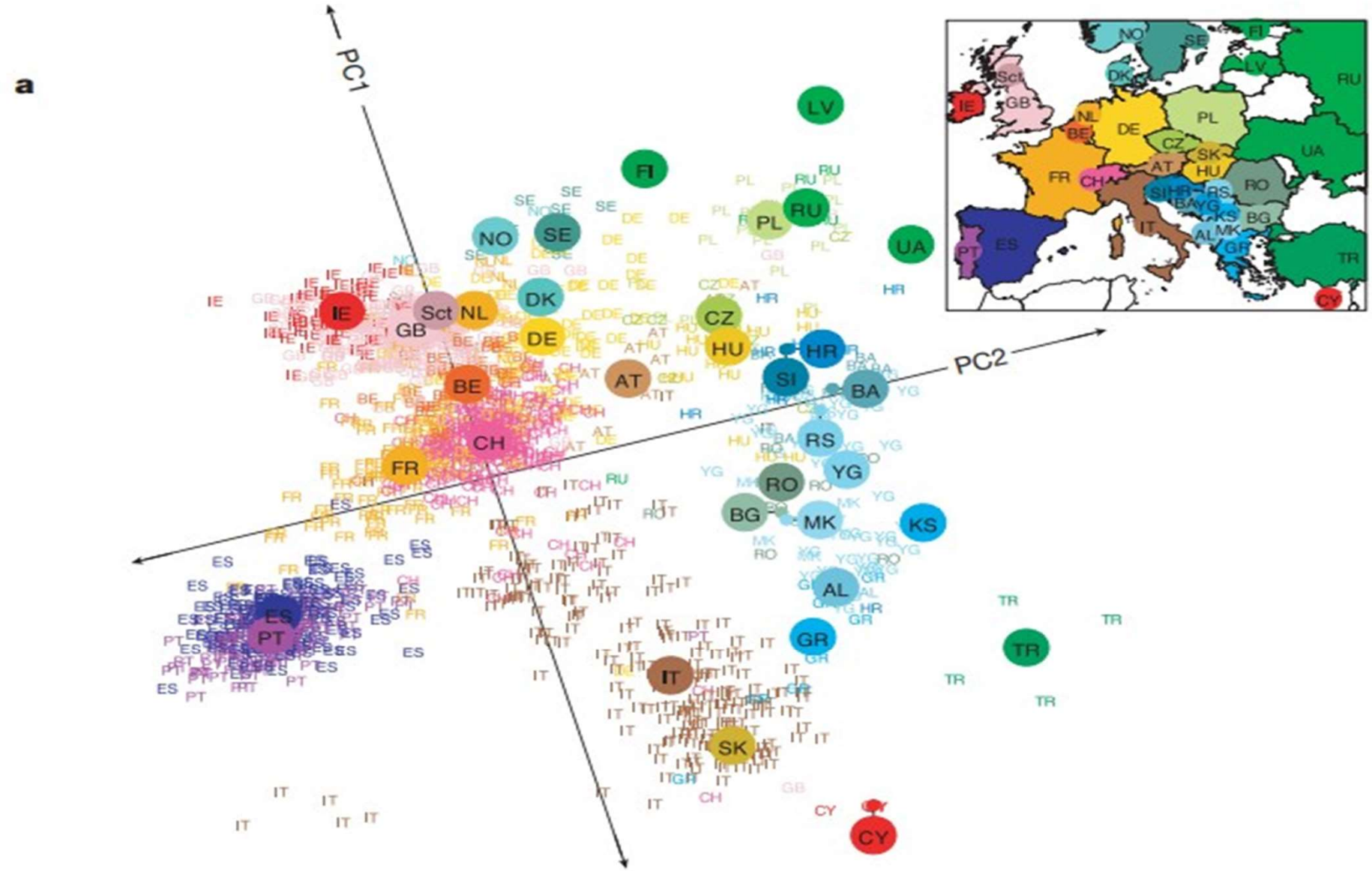
Genotyping Individuals

Stringent criteria is applied to the individuals to avoid potential complications of SNPs in HIGH LINKAGE DISEQUILIBRIUM.

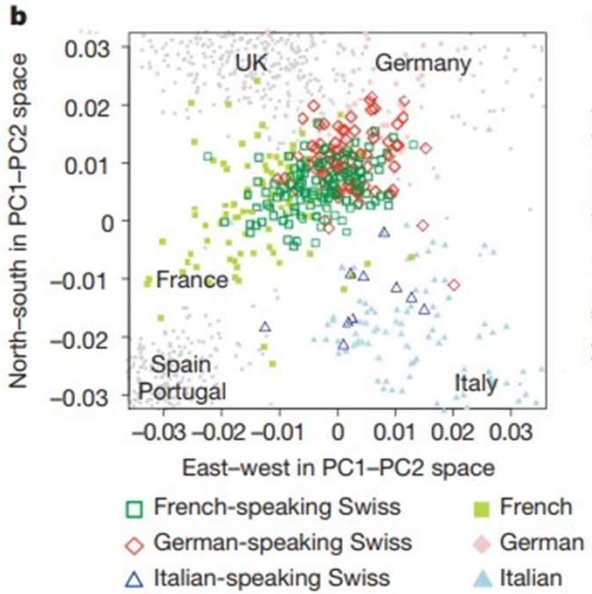
Genotype data from 197,146 loci in 1,387 individuals was obtained, for whom we have high confidence of individual origins.

PRINCIPLE COMPONENT ANALYSIS




a
PC1
PC2
LV
FI
RU
PL
NO
SE
UA
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
AT
SI
BA
BE
CH
RS
FR
YG
RO
BG
MK
KS
AL
ES
PT
GR
TR
IT
SK
CY

RESULTS AFTER DIMENSIONALITY REDUCTION



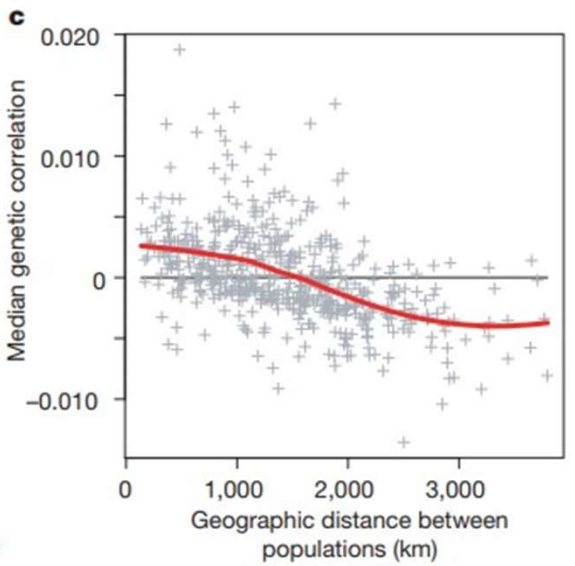


Figure 1 | Population structure within Europe. **a**, A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia. **b**, A magnification of the area around Switzerland from **a** showing differentiation within Switzerland by language. **c**, Genetic similarity versus geographic distance. Median genetic correlation between pairs of individuals as a function of geographic distance between their respective populations.

RESULTS AFTER DIMENSIONALITY REDUCTION

- Using a multiple linear regression-based assignment approach, one can place 50% of individuals within 310 km of their reported origin and 90% within 700 km of their origin . Across all populations, 50% of individuals are placed within 540 km of their reported origin, and 90% of individuals within 840 km.

LIMITATIONS

Only the LARGE NUMBER OF LOCI and DENSE GEOGRAPHIC SAMPLING of individuals used here reveal the clear map-like structure to European genetic variation.

POSSIBILITY OF IMPROVEMENTS

- Current SNP genotyping platforms under-represent variation at low-frequency alleles. Low-frequency alleles tend to be the result of a recent mutation and are expected to geographically cluster around the location at which the mutation first arose.

- Soon-to-be available whole-genome re-sequencing will give us access to informative low-frequency alleles, and further statistical method development will allow us to leverage patterns of haplotype variation