PCAviz: A Principal Component Analysis Visualization Package for R

Richard Williams IV[1], Peter Carbonetto[1], Hossein Pourreza[1], Yuexi Wang[1], and John Novembre[2]

[1]Research Computing Center, [2]Human Genetics, University of Chicago, Chicago, IL, USA

THE UNIVERSITY OF CHICAGO | Research Computing Center

INTRODUCTION

Professor John Novembre and his research group perform studies of human genomic variation using a variety of methods. A key method they employ is principal component analysis (PCA). To efficiently use PCA methods to examine large datasets and visualize the results of these analyses, RCC developed an R statistical programming language package, PCAviz, for quickly creating evocative, interactive visualizations from PCA and accompanying data. PCAviz is specifically designed for visualizing three types of information jointly: (1) PCA results; (2) continuous covariates such as geographical co-ordinates; and (3) categorical data such as group labels.

RCC Support

RCC computational scientists developed a package that rewrites and expands upon the utility of a collection of functions previously written by John Novembre for PCA plotting and visualizations (Novembre et al. 2008). Additional plotting functions for PCA were incorporated in the software package (violin plots, scree plots, PC loading plots). Multiple features and utilities were also added to the original functions (mirroring of PCA plots, axis rotation for PCA plots, hover-over descriptors and IDs for individual points, map inset paired to PCA data). The images presented here are examples of plots produced with the PCAviz R package.

Summary

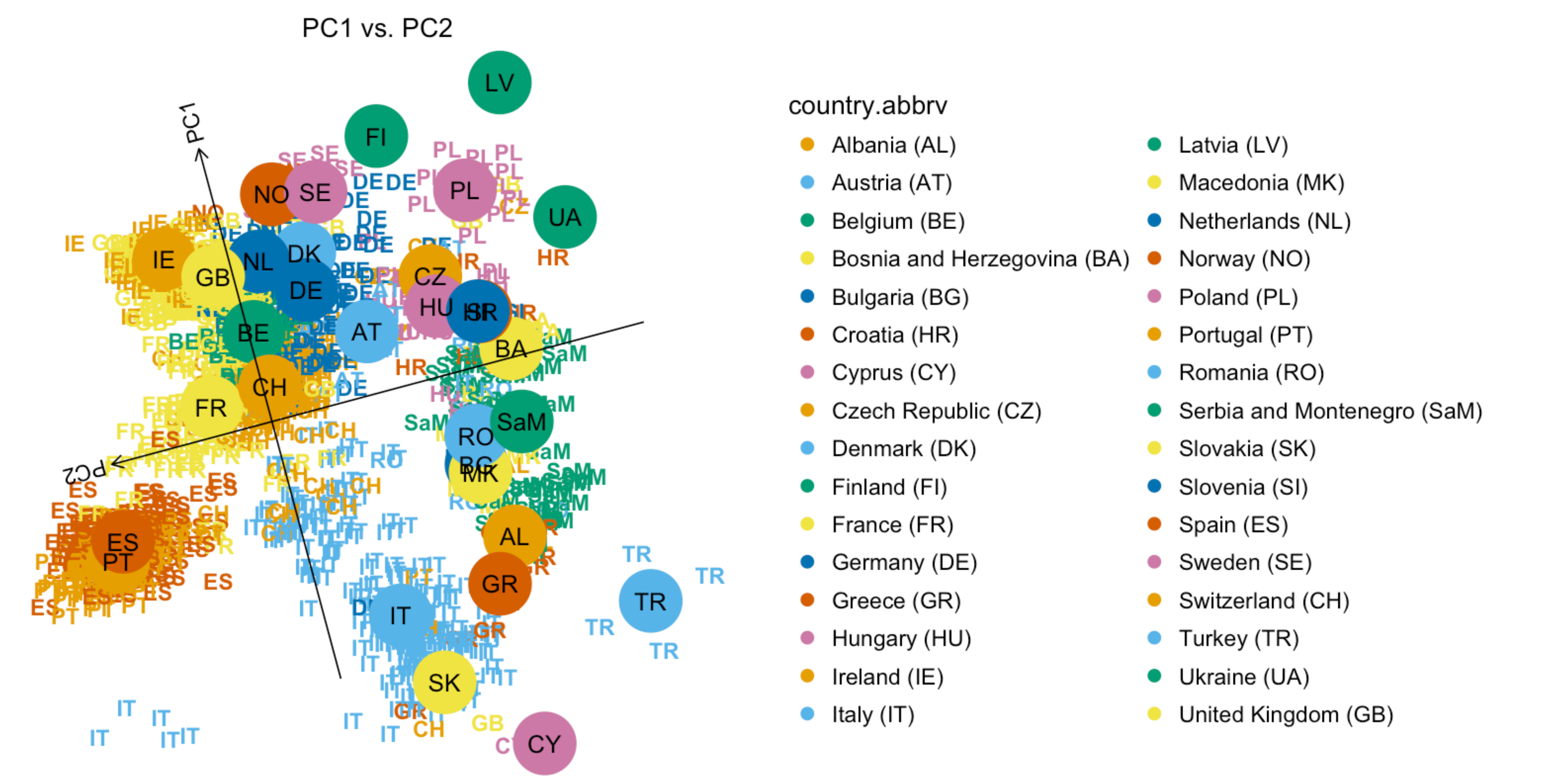
In this project analytical and software development expertise from RCC produced a software package that enables researchers to create publication quality PCA visualizations as well as interactive PCA visualizations. The tools provided in this package are broadly applicable to visualizing PCA of datasets regardless of discipline and will contribute to scientific community's ability explore data and share research findings.

References and Acknowledgements

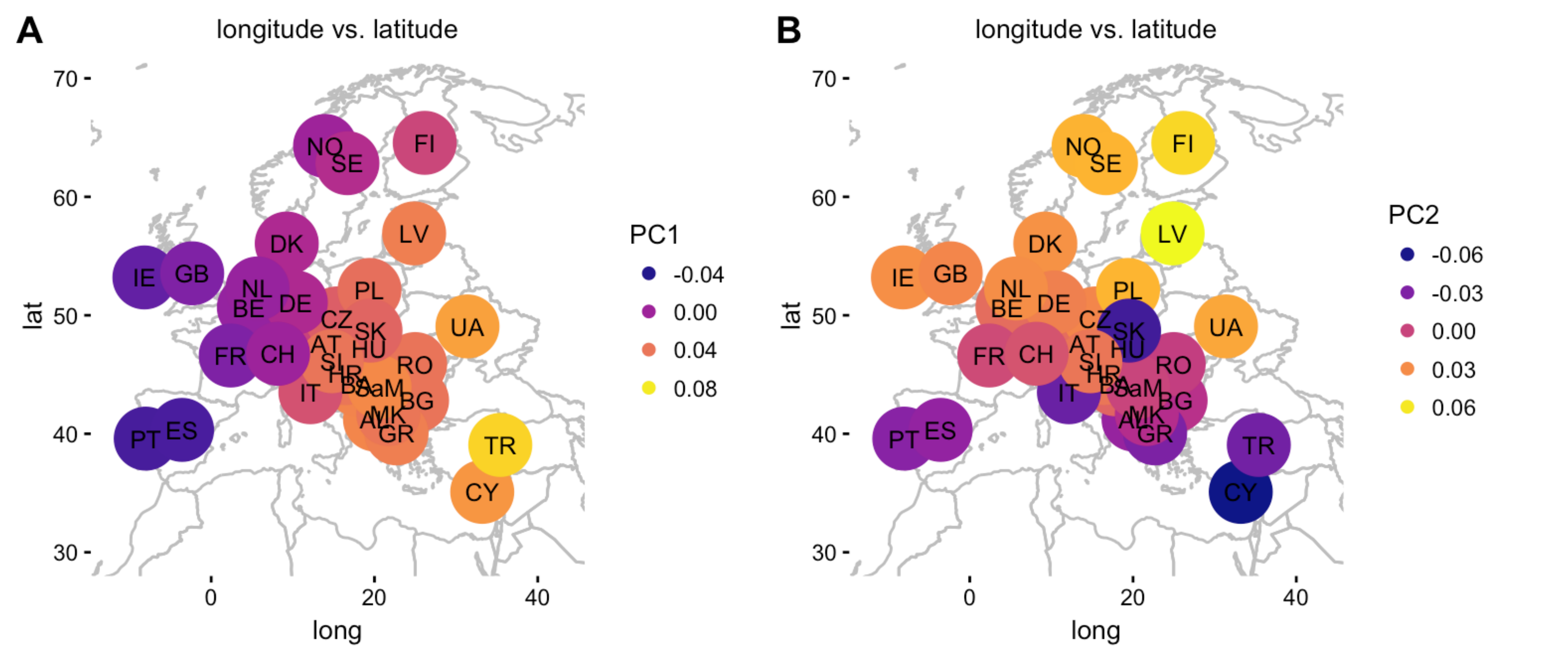
Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... & Stephens, M. (2008). Genes mirror geography within Europe. Nature, 456(7218), 98-101.

Funding for this project was provided by John Novembre. We thank the Novembre research group for valuable feedback and discussions of this work.

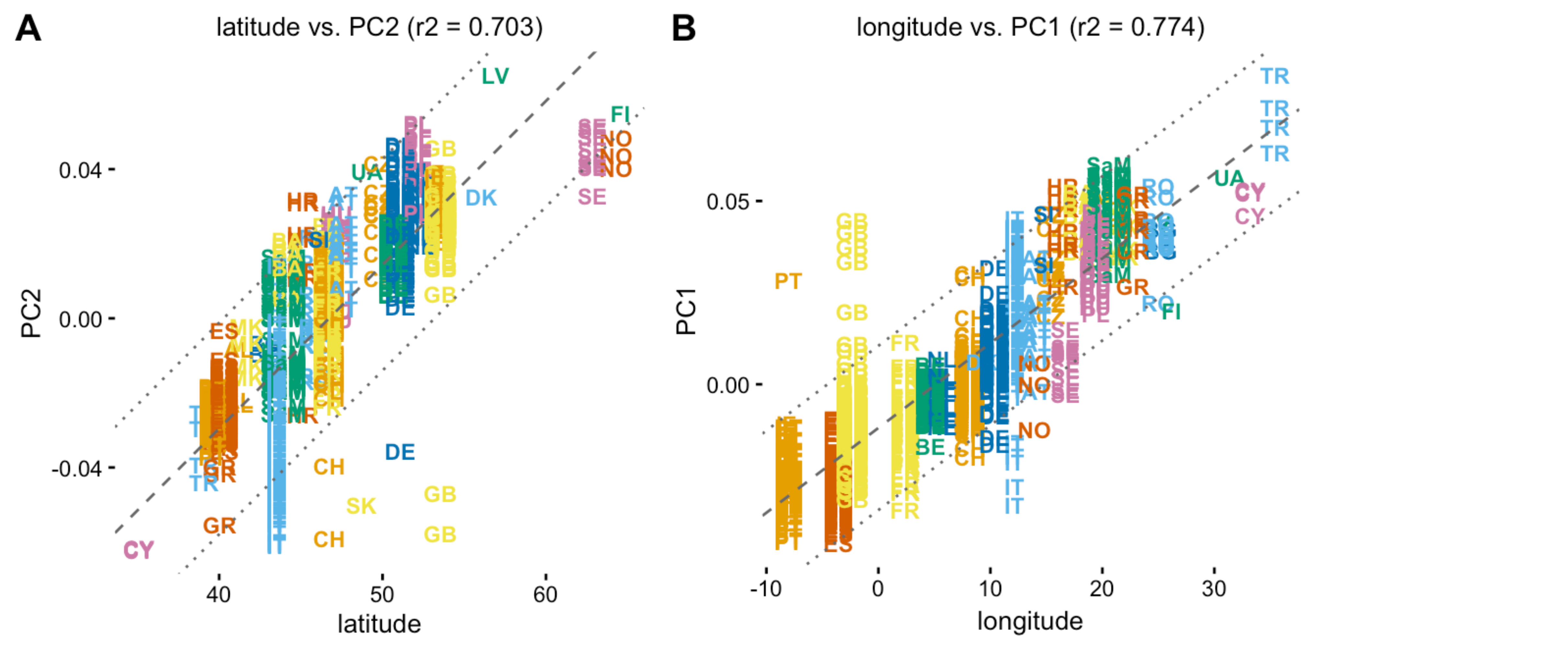
1. Plot of SNP samples plotted in PC space (axes rotated). Individual samples are represented by country abbreviation, group means are represented by circles.



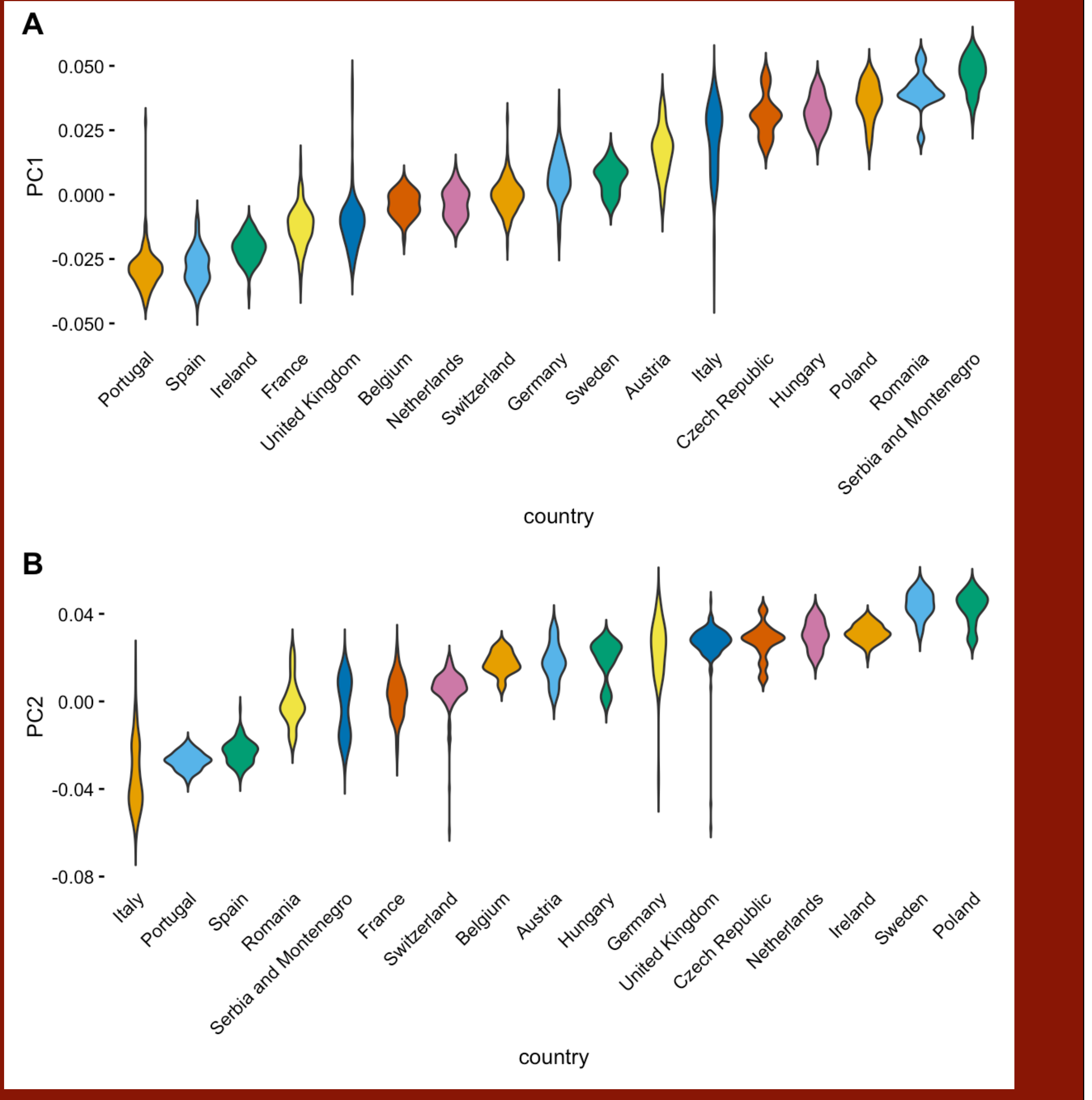
2. Visualization of east-west and north-south trends for PC1 (figure A.) and PC2 figure (B.).



3. Plots showing the relationship between PC2 and latitude (A) and PC1 and .



4.Violin plot of a subset of countries showing the distribution and probability density of the data. The distribution of PC1 values for samples from each country are shown in Figure A, the distribution of PC2 values for samples from each country are shown in Figure B.


A
PC1
country
Portugal
Spain
Ireland
France
United Kingdom
Belgium
Netherlands
Switzerland
Germany
Sweden
Austria
Italy
Czech Republic
Hungary
Poland
Romania
Serbia and Montenegro
B
PC2
country
Italy
Portugal
Spain
Romania
Serbia and Montenegro
France
Switzerland
Belgium
Austria
Hungary
Germany
United Kingdom
Czech Republic
Netherlands
Ireland
Sweden
Poland