Topological data analysis of SNP array data exposes the genetic differentiation between Iberians and Canary Islanders

Jose M. Lorenzo-Salazar[1], Ana Díaz-de Usera[1], **Adrián Muñoz-Barrera[1]**, Luis A. Rubio-Rodríguez[1], Beatriz Guillen-Guio[2], Almudena Corrales[2,3], Itahisa Marcelino-Rodríguez[2], David Comas[4], Rafaela González-Montelongo[2], Santos Alonso[5], Carlos Flores[1,2,3]

[1]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [2]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [3]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [4]Department of Experimental and Health Sciences, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain; [5]Department of Genetics, Phsyical Anthropology and Animal Physiology, University of the Basque Country UPV/EHU, Leioa, Bizkaia, Spain.

Introduction

Unraveling global patterns of human genetic variation is of main interest for the scientific community. The 1000 Genomes Project (1KGP) highlighted that a typical genome differs roughly at 4.1-5.0 million positions from the reference human genome[1]. The population from the Canary Islands (CAN), a Spanish archipelago situated in the Atlantic Ocean 100 Km off the NW African coast, has a unique genetic pool due to isolation, local adaptation and recent admixture of Europeans (EUR), North-Africans (NAF) and sub-Saharan Africans (SSA)[2].

Assessing the genetic structure of populations often requires a multidimensionality reduction approach, typically assessed by Principal Component (PC) Analysis (PCA)[3]. However, such procedure most commonly focuses on few main dimensions limiting the possibilities to excavate fine-grained strata. Here we used Topological Data Analysis (TDA)[4] to explore the genetic dissimilarity of Iberians and Canary Islanders by embedding high-dimensionality of SNP array and whole-genome sequencing (WGS) data to explore the genetic differentiation between populations into a low-dimensional space. New WGS data from NAF were also included for comparative purposes.

Materials and Methods

Sample data: WGS data from 46 Canary Islanders (CAN) and 23 North Africans (NAF) obtained with a HiSeq 4000 (Illumina) to an average of 30x, together with 740 subjects genotyped for the Spain Biobank Array (SBA, Thermo-Fisher Scientific). Additionally, data from 478 individuals from 1KGP were included as reference of EUR and SSA populations. All individuals were unrelated.

Quality control: See the procedures shown on the **Workflow** diagram below.

Statistical analyses: PCA and TDA were assessed on WGS and SBA data using PLINK[5] v1.9 and umap[4] v0.2.0 library for R.

Workflow

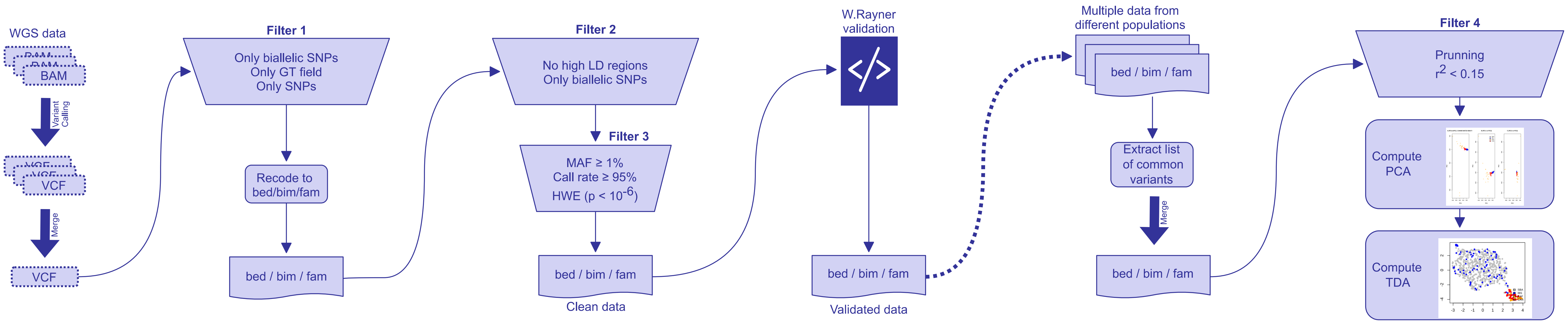

Figure 1. Workflow stages for data preparation: Filter and recode data to bed/bim/fam format, validate it using W.Rayner script and merge data from multiple population to compute Principal Components and Topological Data Analysis.

Results

Comparing CAN, NAF and 1KGP populations

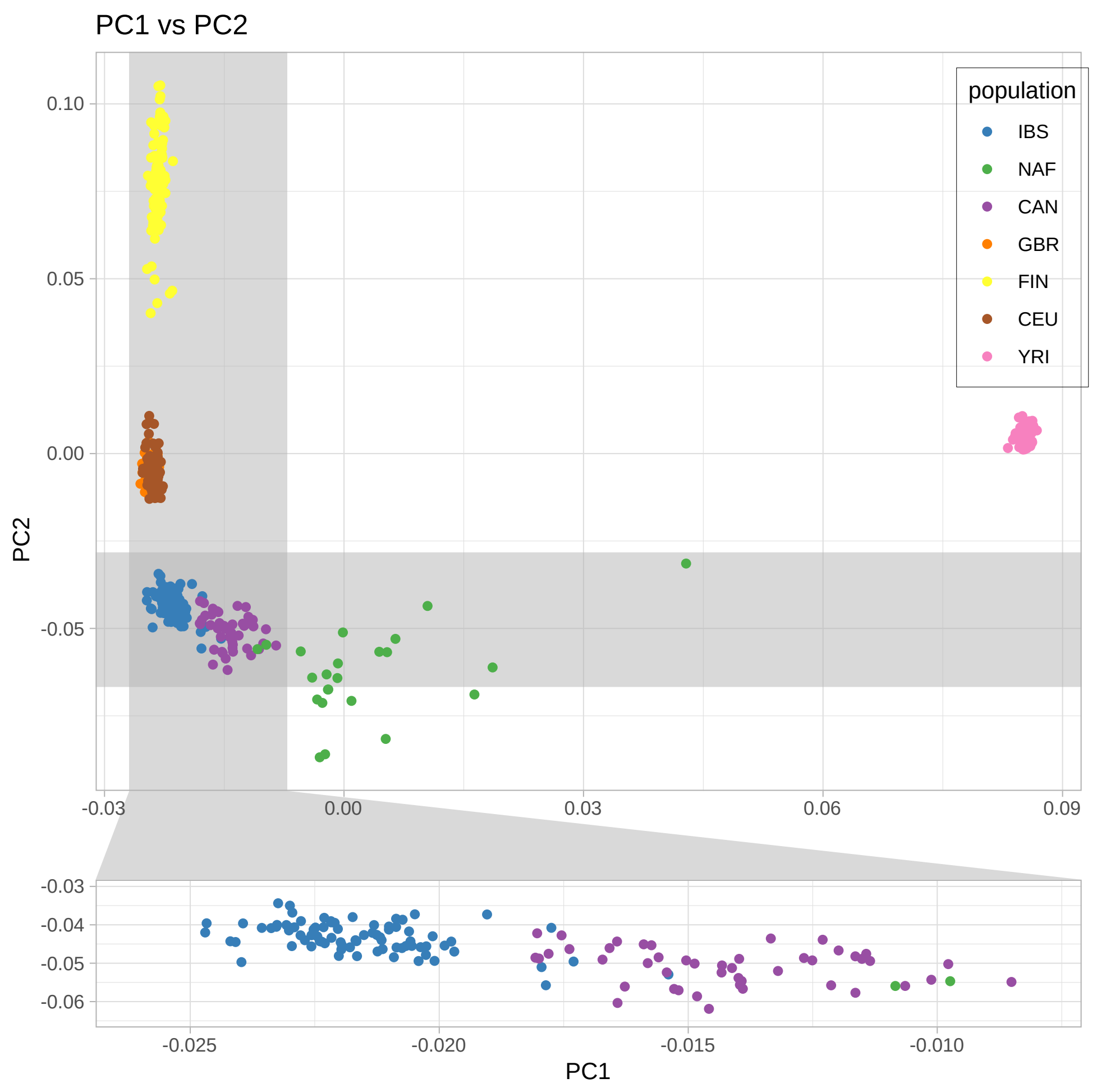

Fig. 2. PC1 vs PC2 representation.

When 1KGP population were included, the first two PCs revealed 3 clusters: a NAF-CAN cluster, a EUR cluster and a SSA cluster (**Fig. 2**).

TDA based on 6 PCs allowed to reveal further fine grained substructure and well-separated clusters, although CAN and NAF still clustered together (**Fig. 3**).

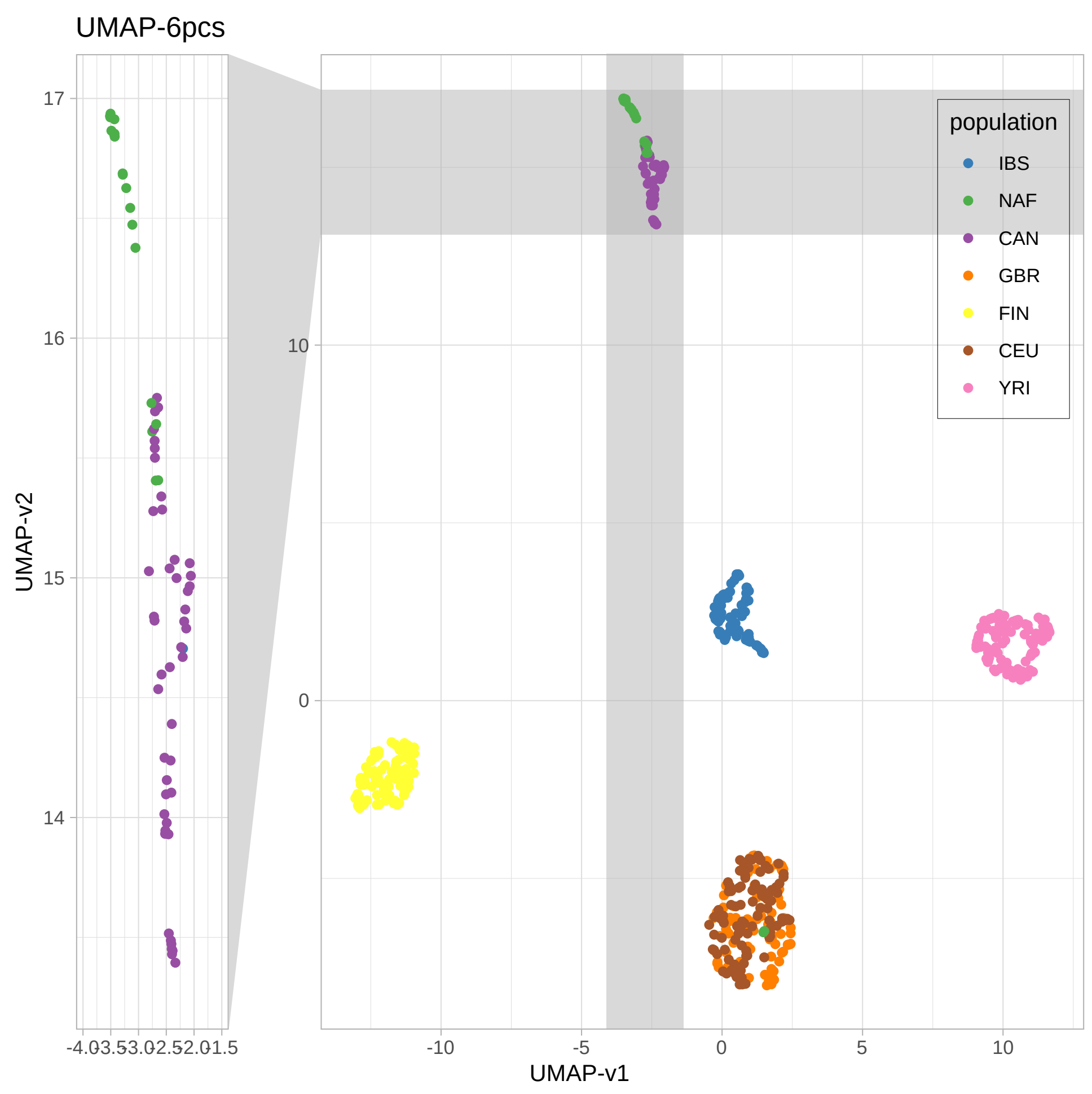

Fig. 3. TDA representation based on 6 PCs.

Comparing CAN, NAF and SBA/IBS populations

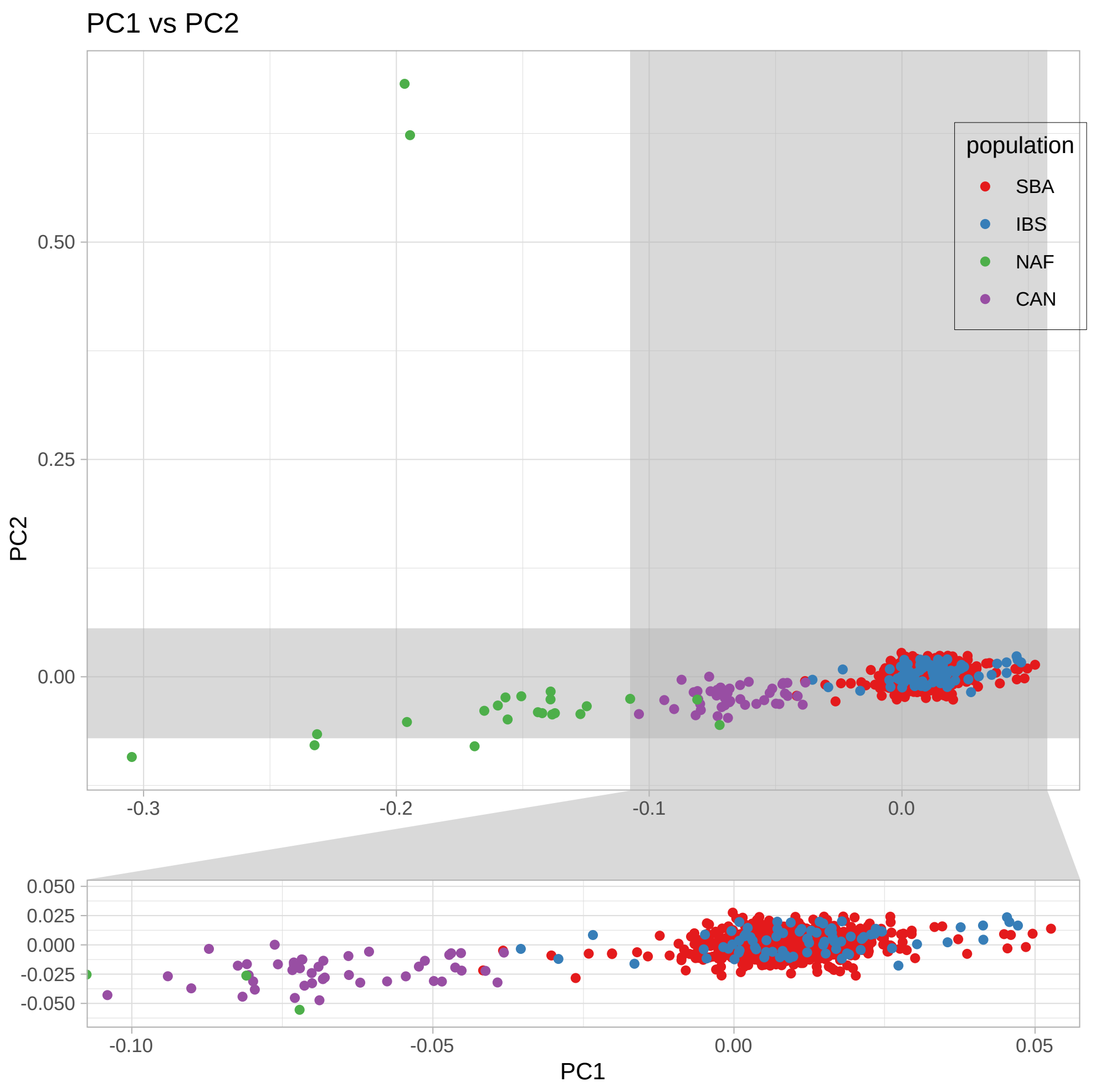

Fig. 4. PC1 vs PC2 representation.

Considering first two PCs in PCA, the vicinity of SBA/IBS and CAN were noticeable with an evident separation from NAF (**Fig. 4**).

This behavior was more remarkable in TDA based on 6 PCs (**Fig. 5**). However, the CAN population clustered closest to NAF than to SBA/IBS.

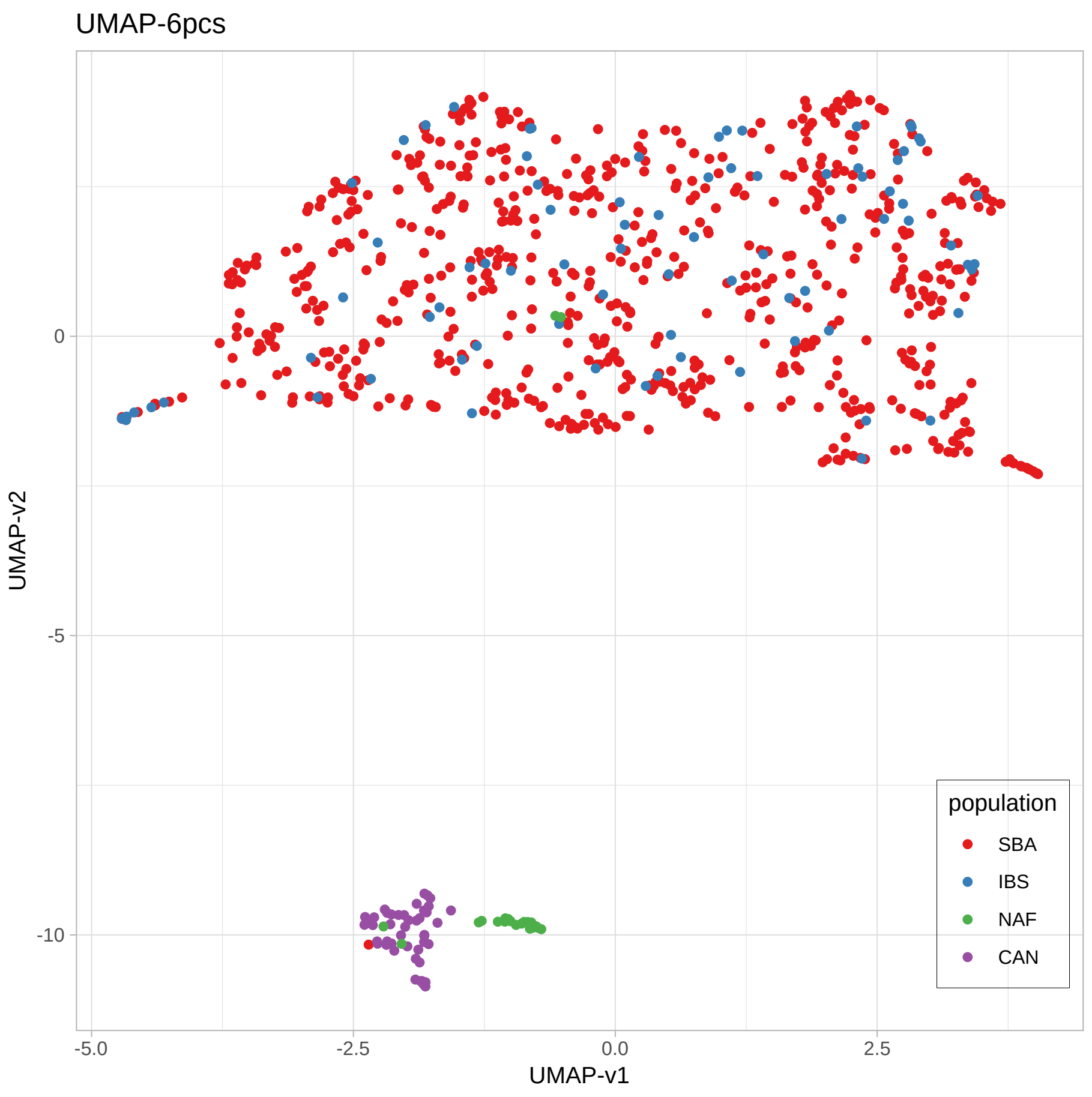

Fig. 5. TDA representation based on 6 PCs.

Conclusions

TDA provides an optimal alternative to reveal previously unrecognized fine structure separating IBS individuals from CAN, a result compatible with genetic drift and African admixture in the latter. Co-clustering of CAN both with NAF and IBS supports wide interindividual variation in ancestries.

In addition, the observed structure of present CAN-IBS and the genetic distance with the rest of EUR populations highlights the unique genetic features of current Canary Islanders.

Contact

Funding

Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE), agreement OA17/008 with ITER. Fellowship by Spanish Ministry of Education, Culture, and Sports to ADU (FPU16/01435) and ACIISI co-funded by European Social Fund to B.G.G (TESIS2015010057).

The authors declare no conflict of interest.

References

1. The 1000 Genomes Project Consortium. *Nature* 2015; 526: 68-74.
2. Guillen-Guio et al. Mol Biol Evol 2018; 35: 3010-3026.
3. Novembre et al. Nature 2008; 456: 98-101.
4. McInnes et al. *arXiv 2018*; 1802.03426v2.
5. Chang et al. *GigaScience 2015*; 4: 7.

Still interested in more details? More information at posters: **P18.68D**, **P18.79C** and **P18.80D**