

PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations

Peristera Paschou¹, Elad Ziv², Esteban G. Burchard³, Shweta Choudhry⁴, William Rodriguez-Cintron⁵, Michael W. Mahoney⁶, Petros Drineas⁷

1) Dept. of Molecular Biology and Genetics, Democritus University of Thrace, Greece, 2) Division of General Internal Medicine, Institute for Human Genetics, University of California San Francisco, USA, 3) Depts. of Biopharmaceutical Sciences and Medicine, University of California San Francisco, USA, 4) Lung Biology Center, Dept. of Medicine, University of California San Francisco, USA, 5) Pulmonary/CCM Veterans Caribbean Healthcare System, University of Puerto Rico School of Medicine, USA, 6) Yahoo Research, USA, 7) Dept. of Computer Science, Rensselaer Polytechnic Institute, USA

Existing methods to ascertain small sets of markers for the identification of human population structure require prior knowledge of individual ancestry. Based on Principal Components Analysis (PCA), and recent results in theoretical computer science, we present a novel algorithm that, applied on genomewide data, selects small subsets of SNPs (PCA-correlated SNPs) to reproduce the structure found by PCA on the complete dataset, without use of ancestry information. Evaluating our method on a previously described dataset (10,805 SNPs, 11 populations), we demonstrate that a very small set of PCA-correlated SNPs can be effectively employed to assign individuals to particular continents or populations, using a simple clustering algorithm. We validate our methods on the HapMap populations and achieve perfect intercontinental differentiation with 14 PCA-correlated SNPs. The Chinese and Japanese populations can be easily differentiated using less than 100 PCA-correlated SNPs ascertained after evaluating 1.7 million SNPs from HapMap. We show that, in general, structure informative SNPs are not portable across geographic regions. However, we manage to identify a general set of 50 PCA-correlated SNPs that effectively assigns individuals to one of nine different populations. Compared to analysis with the measure of informativeness for assignment, our methods, although unsupervised, achieved similar results. We proceed to demonstrate that our algorithm can be effectively used for the analysis of admixed populations without having to trace the origin of individuals. Analyzing a Puerto Rican dataset (192 individuals, 7,257 SNPs), we show that PCA-correlated SNPs can be used to successfully predict structure and ancestry proportions. We subsequently validate these SNPs for structure identification in an independent Puerto Rican dataset. The algorithm that we introduce runs in seconds and can be easily applied on large genome-wide datasets, facilitating the identification of population substructure, stratification assessment in multi-stage whole-genome association studies, and the study of demographic history in human populations.