Inference after human genetic clustering

Javier González-Delgado*†[1] and Simon Gravel*‡[2]

[1]Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz] – Groupe des Écoles Nationales d'Économie et Statistique – France
[2]Department of Human Genetics, McGill University – Canada

Abstract

Human genetic variation is both highly complex and deeply hierarchical, making its characterization challenging. Clustering techniques play a central role in how we interpret genetic variation and reconstruct population history, being widely used in genome-wide association studies. However, genetic research has shown that humans cannot be divided into biologically discrete subgroups: most genetic variation is distributed along continuous gradients, and clear boundaries between populations cannot be objectively defined. Consequently, the use of clusters risks reinforcing typological thinking, that is, the mistaken assumption that humans fall into a limited set of distinct types. It may also artificially amplify within-group homogeneity and between-group differences.

In this context, equipping clustering methods (which are inherently exploratory) with inferential tools may help strengthen the reliability of genetic data analyses. Such tools would make it possible to assess whether individuals assigned to different clusters truly differ genetically or, conversely, whether two clusters returned by an algorithm should be regarded as representing the same group. In this work, we investigate the suitability of post-clustering inference in the context of human genetic clustering. To this end, we examine how well the data conform to admissible models and assess the robustness of the methods to deviations from those models. We illustrate the gap between defining a statistical tool, often under simplified or idealized assumptions, and applying it with guarantees to real data. Our results demonstrate the utility of post-clustering inference when genotypes are classified using certain families of algorithms, while also identifying the intrinsic limitations of existing theory when applied to more complex clustering methods that are better suited to genetic datasets.

Keywords: Population genetics, clustering, post, clustering inference, selective inference

*Speaker
†Corresponding author: javier.gonzalez-delgado@ensai.fr
‡Corresponding author: simon.gravel@mcgill.ca