Visualizing the Geography of Genetic Variants

Joseph Marcus*[1] and John Novembre*[1]

[1]Department of Human Genetics, University of Chicago, Chicago, IL,USA

Abstract

One of the core features of any genetic variant, beyond its potential phenotypic effects or its frequency, is its geographic distribution. The geographic distribution of a genetic variant can shed light on where the variant first arose, in what populations it survived and spread within, and in turn help us learn about historical patterns of migration and natural selection. Collectively the geographic distribution of genetic variants can help to explain how populations have been related through time (e.g. levels of gene flow and divergence). For variants with large effects, it can also help us understand the geographic distribution of spatially-varying phenotypes. For these reasons, visual inspection of geographic maps for genetic variants is common practice in genetic studies. Here we develop a series of reusable interactive visualizations for illuminating the geographic distribution of genetic variants. We specifically address several non-trivial challenges of this type of visualization; in particular, how to represent non-uniform levels of uncertainty in allele frequencies due to variable sample sizes; how to represent results from data with over 10,000 individuals in which allele frequencies can vary over 4 orders of magnitude; how to display data for regions of the globe with dense sampling of populations; and how to quickly access frequency data from large samples. To meet these challenges, we implement a flexible REST API for allowing for easy access to allele frequency and sample size data from large scale public genomic datasets. Built upon this API we develop a web-based browser, entitled the Geography of Genetic Variants (GGV) browser for visualizing the geographic distribution of genetic variants. The GGV browser rapidly provides maps of derived allele frequencies in populations distributed across the globe. The GGV browser builds upon past tools such as the HGDP Selection browser by allowing for more interactive features, new representations of rare variation, as well as incorporating uncertainty in allele frequency estimation. As ancillaries, we also develop a research visualization toolkit that includes a method for displaying high Fst outlier SNPs from the joint site frequency spectrum and an interactive version of commonly used PCA figures. We hope the GGV browser will be a valuable research and education tool for exploring population genetics data.

*Address correspondence to JM (josephhmarcus@gmail.com) or JN (jnovembre@uchicago.edu).

Introduction