Visualizing the Geography of Genetic Variants

Joseph H. Marcus[*,1] and John Novembre[†,1]

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

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 one 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 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.

[*]josephhmarcus@gmail.com
[†]jnovembre@uchicago.edu

Genetics researchers often face the following problem: the researcher has identified one or more genetic variants of interest (e.g. from a genome-wide association, eQTL, or pharmacogenomic study) and would like to know the geographic distribution of the variant. For example, the researcher may hope to address: 1) implications for genomic medicine (e.g. is a risk allele geographically localized to a certain patient population?); 2) design of follow-up studies (e.g. what population should be studied to observe variant carriers?); or 3) the biology of the variant in question (e.g. does the variant correlate with a known environmental factor in a manner suggestive of some geographically localized selection pressure?) (Novembre and Di Rienzo 2009).

A simple geographic map of the distribution of a genetic variant can be incredibly insightful for these questions, yet generating such maps is time-consuming for the average researcher as it requires a combination of data wrangling and use of geographic plotting software that is unfamiliar to most. Our aim here is to produce a tailored system for rapidly constructing informative geographic maps of allele frequency variation.

Acknowledgements

References

Novembre, John, and Anna Di Rienzo. 2009. "Spatial Patterns of Variation Due to Natural Selection in Humans." *Nature Reviews Genetics* 10 (11). Nature Publishing Group: 745–55.