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 an interactive web-based visualization tool for illuminating the geographic distribution of genetic variants. Through an efficient RESTful API and dynamic front-end the Geography of Genetic Variants (GGV) browser rapidly provides maps of minor allele frequencies in populations distributed across the globe.

[*] 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 cleaning methods 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. Our work is inspired by the past tools such as the HGDP Selection browser but aims to address significant visualization challenges that are inherit in the production of such "frequency maps", including dynamic interaction, display of rare genetic variation, and representation of uncertainty in estimated allele frequencies due to variable sample sizes. Web based visualization tools, such as Data Driven Documents (d3.js), offer powerful approaches to realize these aims through useful methods for interactivity, advantages of software development in modern web-browsers, a large open-source development community, and ease of share-ability (Bostock, Ogievetsky, and Heer 2011).

The Geography of Genetic Variants browser (GGV) uses the SVG and mapping utilities of d3.js to generate interactive "frequency maps", allowing for quick and dynamic displays of the geographic distribution of a genetic variant. In order to allow for easy access to large commonly used public genomic datasets, such as the 1000genomes project, Human Genome Diversity project and POPRES project, we develop a multi-billion row SQL database and RESTful API for querying allele frequencies by chromosome and position or reference SNP identifier [Figure 1]. The GGV displays the frequency of a variant in a given population as a pie chart where each slice represents minor and major allele frequencies. Pie charts are displayed as points positioned at the population's region of origin where the projection/map-view is chosen based off of the geographic proximity of populations present in a given dataset [Figure 2].

One under-appreciated problem with allele frequency maps is that not all data points have equal levels of certainty. For some locations, sample sizes are small, and the reported allele frequency may be quite far from the true population frequency due to sampling error. To address this issue, we use varying transparency in a population's pie chart: sample allele frequencies with higher levels of sampling error will be made more transparent, and hence less visible on the map [Figure 3].

An additional challenge is that allele frequencies between variants often differ greatly, sometimes by orders of magnitude in a single dataset. This has not been an pervasive problem until recently, as most population genetic samples were based on genotype arrays biased towards variants that are common in human populations (5-50% in minor allele frequency). With the advent of next generation sequencing technologies and large samples of thousands of individuals the research community discovered that the vast majority of genetic variation is rare ((@1000genomes2012integrated, Nelson et al. 2012, Tennessen et al. (2012)). In current visualization schemes, such as the HGDP Selection Browser, rare variants would be represented as narrow slivers in a pie chart, nearly invisible to the naked eye. To address this challenge we re-scale frequencies for rare variants, so that small frequencies become visible. Specifically, we will use a "frequency scale" that is indicated below the map, and redundantly using color, that will indicate a constant scale for the frequencies displayed [Figure 4]. Much like scientific notation, this allows a wide range of frequencies to be displayed].

In many datasets where populations are sampled densely in geographic space, one problem is that allele frequency plots begin to overlap each other and obscure information. To address this issue, we will use force-directed layouts of the populations such that no two points are overlapping each other, and yet the points will be pulled towards their true origins. Lines anchoring the points visibly to their original sampling locations will be used to make sure true sampling locations are indicated [Figure 5].

Finally, from our experience, it can be very helpful to quickly view the geographic distribution of several randomly chosen variants to quickly gain a sense of a dataset . For instance, in data with deep population subdivision, it is obvious in the consistent patterns of differentiation observed across markers. To help facilitate this we will include a button to query for a random variant. We also expect this will be useful in teaching contexts – as it provides a highly visual way for learners to understand human genetic variation.

To Add

- Provide information about links to ggv, freq_api, and github for contributions
- Fill up citation and fix citation formatting .bst?
- Acknowledgments
- Fill in figures

Acknowledgements

References

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. "D^3 Data-Driven Documents." *Visualization and Computer Graphics, IEEE Transactions on* 17 (12). IEEE: 2301–9.

Nelson, Matthew R, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, et al. 2012. "An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People." *Science* 337 (6090). American Association for the Advancement of Science: 100–104.

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.

Tennessen, Jacob A, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, et al. 2012. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes." *Science* 337 (6090). American Association for the Advancement of Science: 64–69.