OXFORD

Visualizing the Geography of Genetic Variants

Joseph H. Marcus [1] and John Novembre [1, 2,*]

[1] Department of Human Genetics, University of Chicago, Chicago, 60637, US and
[2] Department of Ecology and Evolution, University of Chicago, Chicago, 60637, US.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Abstract

One of the key characteristics of any genetic variant is its geographic distribution. The geographic distribution can shed light on where an allele first arose, what populations it has spread to, and in turn on how migration, genetic drift, and natural selection have acted. The geographic distribution of a genetic variant can also be of great utility for medical/clinical geneticists and collectively many genetic variants can reveal population structure. Here we develop an interactive visualization tool for rapidly displaying the geographic distribution of genetic variants. Through a REST API and dynamic front-end the *Geography of Genetic Variants (GGV)* browser provides maps of allele frequencies in populations distributed across the globe.

Contact: jnovembre@uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genetics researchers often face the problem that they have identified one or many genetic variants of interest using an approach such as a genome-wide association study and then 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? What population should be studied to observe variant carriers? Rosenberg *et al.*, 2010); or 2) the evolutionary history 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; Coop *et al.*, 2010). A simple geographic map of the distribution of a genetic variant can be incredibly insightful for these questions.

Contemporary population genetics researchers are also faced with the challenge of large, high-dimensional datasets. For example, it is not uncommon for a researcher in human genetics to have a dataset comprised of thousands of individuals measured at hundreds of thousands or even millions of single nucleotide variants (SNVs). One common approach to visualizing such high-dimensional data is to compress the SNV dimensions down to a small number of latent factors, using a method such as principal components analysis (Price *et al.*, 2006; Patterson *et al.*, 2006), or a model-based clustering method such as STRUCTURE (Pritchard *et al.*, 2000) or ADMIXTURE (Alexander *et al.*, 2009). While these methods are

extremely valuable, researchers can use them too often without inspecting the underlying variant data in more detail. A natural approach to gaining more insight to the overall structure of a population genetic dataset is to visually inspect what geographic patterns arise in allele frequency maps.

Unfortunately, generating geographic allele frequency maps is time-consuming for the average researcher as it requires a combination of data-wrangling methods (Kandel *et al.*, 2011) and map-making techniques that are 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 past tools such as the ALFRED database (Rajeevan *et al.*, 2012) and the maps available on the HGDP Selection browser (Pickrell *et al.*, 2009) whose allele frequency output and plots have been used in research articles (e.g. Pickrell *et al.*, 2009; Coop *et al.*, 2009), books (e.g. Dudley and Karczewski, 2013), and have been made available on the UCSC Genome Browser (available under the HGDP Allele Freq track of the browser Kent *et al.*, 2002).

Taking advantage of recent advances in web-based visualization tools (Bostock *et al.*, 2011), we aim to address the significant visualization challenges that are inherit in the production of geographic allele frequency maps for large population genomic datasets (The 1000 Genomes Project Consortium, 2015), (Lazaridis *et al.*, 2014), including dynamic interaction, display of rare genetic variation, and representation of uncertainty in estimated allele frequencies due to variable sample sizes.
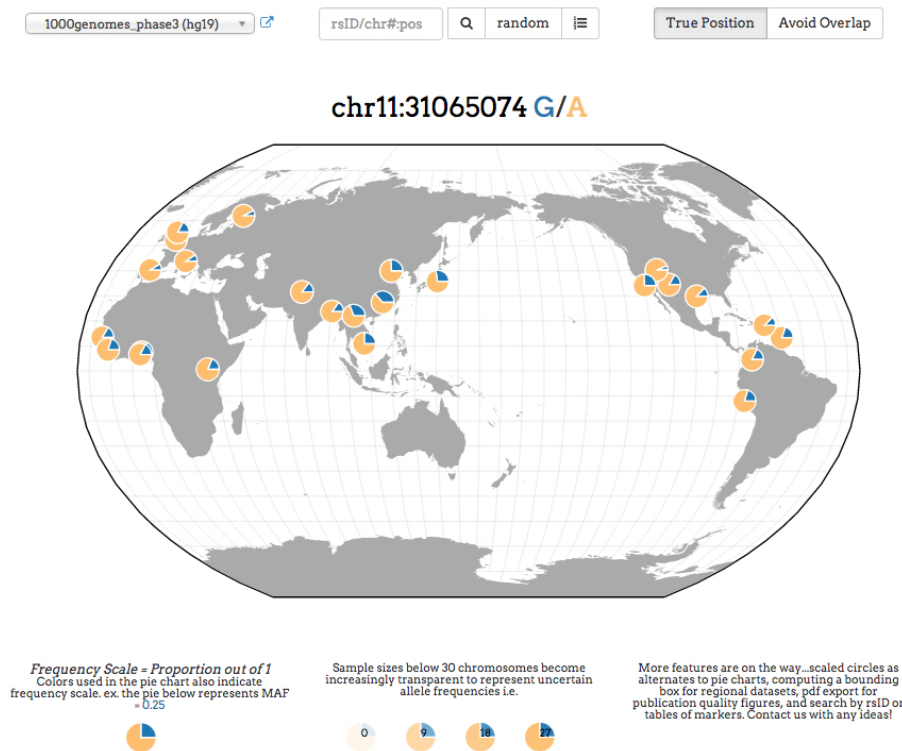
1

Fig. 1. Example screenshot from the Geography of Genetic Variants browser using The 1000 Genomes Project Consortium (2015) data. Each pie chart represents a population with the blue slice of the pie displaying the frequency of the global minor allele and the yellow slice of the pie displaying the frequency of the global major allele in each population.

2 Approach

The Geography of Genetic Variants browser (GGV) uses the scalable vector graphics and mapping utilities of D3.js (Bostock *et al.*, 2011) to generate interactive frequency maps, allowing for quick and dynamic displays of the geographic distribution of a genetic variant. The front-end provides legends for the map and various configuration boxes to allow users to query different datasets or choose visualization options.

In order to allow for easy access to commonly used public genomic datasets, such as the 1000 Genomes project (The 1000 Genomes Project Consortium, 2015) or Human Genome Diversity project (Li *et al.*, 2008), we have developed a REST API (Grinberg, 2014) for accessing data. The API allows querying of allele frequencies by chromosome and position, by reference SNP identifier (Sherry *et al.*, 2001), or randomly sampled SNPs. While many applications require inspection of the distribution of a specific variant, from our experience, it can be very helpful to view the geographic distribution of several randomly chosen variants to quickly gain a sense of structure in a dataset. We find this to be especially useful in teaching contexts, as it provides a highly visual way for learners to understand human genetic variation.

After a query, the GGV displays the allele frequencies for a set of populations as a collection of pie charts where each represents the minor and major allele frequency in a single population. Pie charts are displayed as points at a latitude and longitude associated with a population and the map boundaries are chosen based off of the geographic configuration of populations in a given dataset [Figure 1].

3 Features

Rare variants: Representing the frequency of rare variants across populations can be difficult as allele frequencies between variants often differ greatly, sometimes by orders of magnitude in a single dataset (e.g. The 1000 Genomes Project Consortium, 2015; Nelson *et al.*, 2012; Tennessen *et al.*, 2012). To address this challenge we re-scale frequencies for rare variants, so that small frequencies become visible. Specifically, we use a frequency scale that is indicated in a legend below the map and represented by varying color in the pie charts [Figure 1, Supplementary Figure 1, Supplementary Table 1].

Uncertainty in frequency data: One under-appreciated problem with allele frequency maps is that not all data points have equal levels of certainty. To address this issue, we use varying transparency in a population's pie chart: estimated frequencies with higher levels of sampling error (e.g. those from samples with $n < 30$) are made more transparent, and hence less visible, on the map [Figure 1, Supplementary Figure 2].

Overlapping populations: 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 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 [Figure 1, Supplementary Figure 3]. Also, by hovering the mouse cursor over any population, a user can see the population labels and precise frequency information.

REST API: To provide an interface to the population minor allele frequency data, we use a REST API implemented in the python library Flask-RESTful (Grinberg, 2014). The front-end D3.js visualization uses the API to obtain the data, though users can also interface with it directly. Genetic variants can be queried by chromosome position, rsid, or randomly (see supplementary data for examples).

4 Conclusion

By allowing rapid generation of allele frequency maps, we hope to facilitate the interpretation of variant function and history by practicing geneticists. We emphasize that the overwhelming majority the human genome does not differ across populations (see a classic treatment in Lewontin, 1972). We thus hope that the ability to query random variants from major human population genetic samples will allow students to appreciate the structure of human genetic diversity in an approachable and intuitive form.

Acknowledgements

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(9), 1655–1664.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2301–2309.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., and Pritchard, J. K. (2009). The role of geography in human adaptation. *PLoS Genet*, **5**(6), e1000500.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**(4), 1411–1423.

Dudley, J. T. and Karczewski, K. J. (2013). *Exploring personal genomics*. Oxford University Press.

Grinberg, M. (2014). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, **10**(4), 271–288.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, **12**(6), 996–1006.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., *et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513**(7518), 409–413.

Lewontin, R. C. (1972). The apportionment of human diversity. In *Evolutionary biology*, pages 381–398. Springer.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**(5866), 1100–1104.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., *et al.* (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**(6090), 100–104.

Novembre, J. and Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics*, **10**(11), 745–755.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, **2**(12), e190.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., *et al.* (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, **19**(5), 826–837.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8), 904–909.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959.

Rajeevan, H., Soundararajan, U., Kidd, J. R., Pakstis, A. J., and Kidd, K. K. (2012). ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Research*, **40**(D1), D1010–D1015.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, **11**(5), 356–366.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**(1), 308–311.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**(6090), 64–69.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.