Visualizing the geography of genetic variants

Joseph H. Marcus[1*], John Novembre[1*]
1 Department of Human Genetics, University of Chicago, Chicago, IL, USA
∗ E-mail: jhmarcus@uchicago.edu jnovembre@uchicago.edu

Abstract

One of the key characteristics 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 spread within, and in turn can help one learn about historical patterns of migration, genetic drift, 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 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 allele frequencies in populations distributed across the globe.

Introduction

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?) [1]. 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, 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 vairants (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 [2], [3], a model-based clustering method such as STRUCTURE [4] or ADMIXTURE [5]. While these approaches are extremely valuable, researchers can use them to 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 [6] and geographic plotting 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 [7] and the maps available on the HGDP Selection browser [8]. One of us (J.N.) developed the scripts for the HGDP Selection Browser maps using The Generic Mapping Tools (GMT) [9], a powerful system of geographic plotting scripts with a sbustantial learning curve. The plots from the HGDP Selection Browser have proved useful, and have appeared in research articles e.g **???**, books e.g. [10], and have been ported and made avialable on the

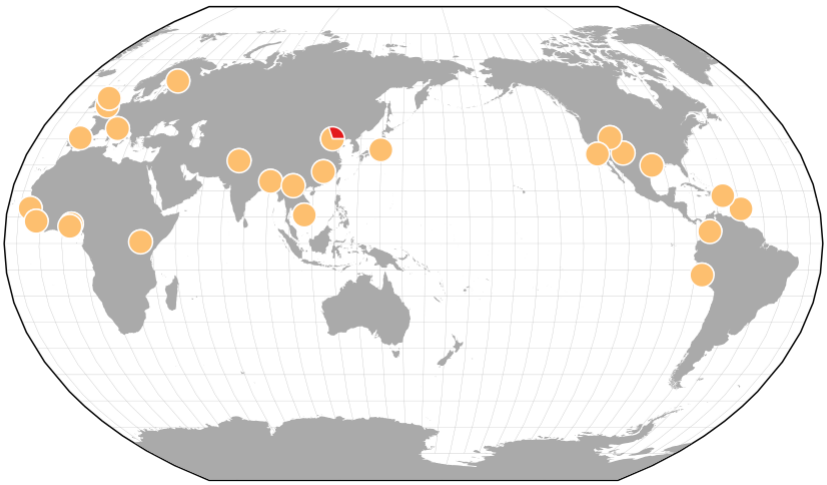

Figure 1. Figure caption

UCSC Genome Browser (available under the XX track of the browser).

Reference datasets for population genetic variation has greatly expanded since the release of the HGDP Illumina 650Y dataset [11] that formed the basis of the HGDP Selection Browser maps. The most notable advance is the publication of the 1000 Genomes phase 3 data [12] though additional datasets are coming online [13], [14].

In addition, novel approaches for data visualization have become more widely available. In particular, 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 [15].

Taking advantage of these recent advances, we aim to address significant visualization challenges that are inherit in the production of geographic allele frequency maps, including dynamic interaction, display of rare genetic variation, and representation of uncertainty in estimated allele frequencies due to variable sample sizes.

Fundamental Approach

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 1000 Genomes project, Human Genome Diversity project and POPRES project, we have developed an SQL database and RESTful API for querying allele frequencies by chromosome and position or reference SNP identifier [Figure 1].

Users can query single variants by chromosome and position identifiers, by rsids [16], or users can simply choose a random variant from within a dataset. While many applications require inspection of the distribution of a specific variant, 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 structure in a dataset. For instance, in data with deep population subdivision, it is obvious in the consistent patterns of differentiation observed across markers. We also expect this will be 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 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].

Representing variable certainty in frequency data

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].

Representing rare variants in frequency data

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, it is common for datasets to contain rare variants [12], [17], [18].

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].

Interface features

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. Lines anchoring the points visibly to their original sampling locations will be used to make sure true sampling locations are indicated [Figure 5].

Underlying Frequency API

[Joe - worth saying more about the API structure and giving example calls and structure of JSON?]

Caveats

A major challenge of using a geographic representation of genetic variation in humans is that the samples must be associated with a geographic location. While doing so is generally immensely helpful, it has inherent complexity. For example, pracitioners must make choices regarding representing where an individual was sampled for the study (e.g. the city of a major research center) or choosing a location that is more representative of an individual's ancestral origins (e.g. as determined in practice by the birthplaces

of recent ancestors, such as grandparents). We do not proscribe a general solution to this problem and instead use a set of locations for each dataset that aligns with those used by the initial analysts of the data. A future feature will be to allow alternative location schema to be used for the populations in a dataset.

Conclusion

By allowing rapid generation of allele frequency maps, we hope to facilitate the interpretation of variant function and history by practicing geneticists. We also hope the ability to query random variants from major human population genetic samples will allow students to appreciate the structure of human genetic diversity in a more approachable and intuitive form than alternative visualizations.

Acknowledgements

Support for this work was provided by the NIH-BD2K initiative (1U01 CA198933-0). The authors would also like to thank XX for supportive conversations.

Acknowledgements

References

1. Novembre J, Di Rienzo A. Spatial patterns of variation due to natural selection in humans. Nature Reviews Genetics. Nature Publishing Group; 2009;10: 745–755.

2. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. Nature Publishing Group; 2006;38: 904–909.

3. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genet. Public Library of Science; 2006;2: e190.

4. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. Genetics Soc America; 2000;155: 945–959.

5. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. Cold Spring Harbor Lab; 2009;19: 1655–1664.

6. Furche T, Gottlob G, Libkin L, Orsi G, Paton NW. Data wrangling for big data: Challenges and opportunities.

7. Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK. ALFRED: An allele frequency resource for research and teaching. Nucleic acids research. Oxford Univ Press; 2011; gkr924.

8. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome research. Cold Spring Harbor Lab; 2009;19: 826–837.

9. Wessel P, Smith WH, Scharroo R, Luis J, Wobbe F. Generic mapping tools: Improved version released. Eos, Transactions American Geophysical Union. Wiley Online Library; 2013;94: 409–410.

10. Dudley JT, Karczewski KJ. Exploring personal genomics. Oxford University Press; 2013.

11. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. science. American Association for the Advancement of Science; 2008;319: 1100–1104.

12. Consortium 1GP, others. A global reference for human genetic variation. Nature. Nature Publishing Group; 2015;526: 68–74.

13. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome

sequence from an archaic denisovan individual. Science. American Association for the Advancement of Science; 2012;338: 222–226.

14. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day europeans. Nature. Nature Publishing Group; 2014;513: 409–413.

15. Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents. Visualization and Computer Graphics, IEEE Transactions on. IEEE; 2011;17: 2301–2309.

16. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. DbSNP: The nCBI database of genetic variation. Nucleic acids research. Oxford Univ Press; 2001;29: 308–311.

17. Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. American Association for the Advancement of Science; 2012;337: 100–104.

18. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. American Association for the Advancement of Science; 2012;337: 64–69.