Visualizing the Geography of Genetic Variants

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

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

Abstract

One of the key characteristics of any genetic variant, beyond its potential phenotypic effects or its global 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 has spread to, and in turn how migration, genetic drift, and natural selection have acted on the variant. Collectively the geographic distribution of many genetic variants can inform on how populations have been related through time (e.g. levels of gene flow and divergence). The distribution of a genetic variant can also be informative background knowledge for medical / clinical geneticists. As a result, 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 SQL database, REST API and dynamic front-end the *Geography of Genetic Variants (GGV)* browser rapidly provides maps of allele frequencies in populations distributed across the globe.

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

Introduction

Genetics researchers often face the problem that they have identified one or many 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?) (Rosenberg *et al.*, 2010; 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 approaches 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 (Furche *et al.*, 2016) 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). One of us (J.N.) developed the scripts for the HGDP Selection Browser maps using The Generic Mapping Tools (GMT) (Wessel *et al.*, 2013), a powerful system of geographic plotting scripts for making static plots. The plots from the HGDP Selection Browser have proved useful, and have appeared in research articles (Pickrell *et al.*, 2009; Coop *et al.*, 2009, e.g.), books (Dudley and Karczewski, 2013, e.g.), and have been ported and made available on the UCSC Genome Browser (available under the HGDP Allele Freq track of the browser) (Kent *et al.*, 2002).

Reference datasets for population genetic variation have greatly expanded since the release of the HGDP Illumina 650Y dataset (Li *et al.*, 2008) that formed the basis of the HGDP Selection Browser maps. The most notable advance is the publication of the 1000 Genomes Phase 3 data (The 1000 Genomes Project Consortium, 2015) though additional datasets are continually coming online (Meyer *et al.*, 2012; Lazaridis *et al.*, 2014). 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 useful methods for interactivity, the advantages of software development in modern web-browsers, a large open-source development community, and ease of sharing (Bostock *et al.*, 2011).

Taking advantage of these recent advances, we aim to address the 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 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.

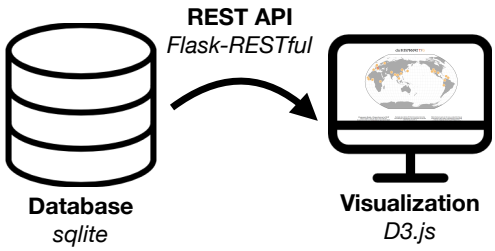

Figure 1: Schematic diagram of the Geography of Genetic Variants (GGV) browser. We compute minor allele frequencies from publicly available population genetic datasets from the 1000 Genomes Project, Human Genome Diversity Project (HGDP) and Population Reference Sample (POPRES). These allele frequencies are stored in a SQL database along side with geographic coordinates associated with each population. The front-end visualization was created using D3.js and accesses allele frequency and population meta-data from a REST API implemented using the python library flask.

In order to allow for easy access to large commonly used public genomic datasets, such as the 1000 Genomes project (The 1000 Genomes Project Consortium, 2015), Human Genome Diversity project (Li *et al.*, 2008), we have developed an SQL database and REST API (Grinberg, 2014) for querying allele frequencies by chromosome and position, by reference SNP identifier (Sherry *et al.*, 2001), or randomly sampled SNPs [Figure 1]. 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 and the geographic range displayed is chosen based off of the geographic configuration of populations in a given dataset [Figure 2].

Representing uncertainty 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: sampled populations with higher levels of sampling error (i.e. $n < 30$) are 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 a pervasive problem until recently, as most population genetic samples were genotyped on SNP arrays, which have been biased towards variants that are common in human populations (5-50 % in minor allele frequency). With the combination of next generation sequencing technologies, new array designs focusing on rarer variants, and large samples of thousands of individuals, it is now routine for datasets to contain rare variants (The 1000 Genomes Project Consortium, 2015; Nelson *et al.*, 2012; Tennessen *et al.*, 2012). A new challenge is that in visualization schemes using proportional area to represent frequency (such as standard pie charts), rare variants would be represented as narrow slivers, nearly invisible to the naked eye.

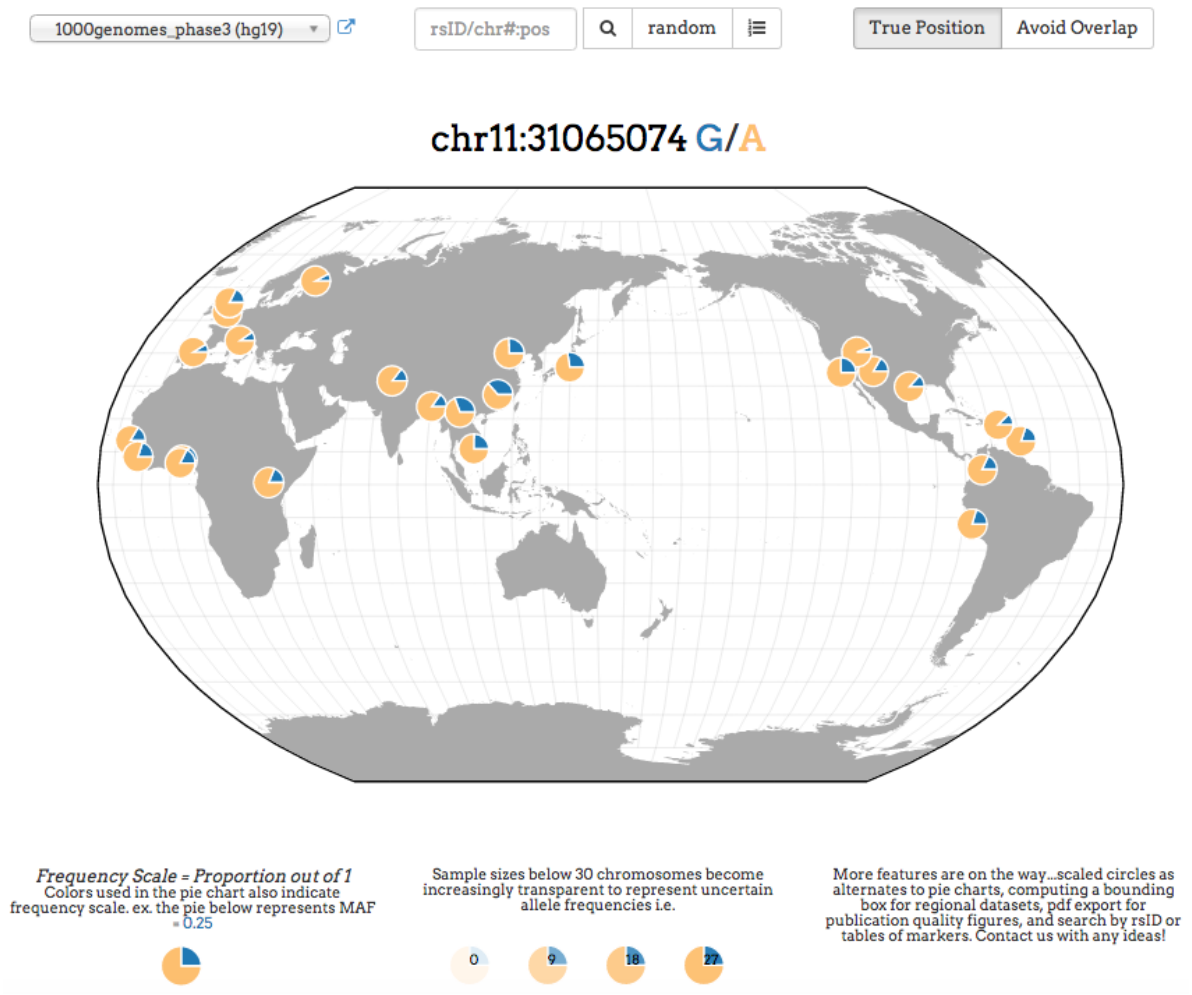

Figure 2: Example screenshot from the Geography of Genetic Variants browser using the (The 1000 Genomes Project Consortium, 2015) data. Each pie chart represents a population with the blue slice of the pie displaying the frequency of the global minor allele and the yellow slice of the pie displaying the frequency of the global major allele in each population.

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 redundantly by using color, that indicates the scale for the frequencies displayed [Figure 4]. Much like scientific notation, this allows a wide range of frequencies to be displayed (Table 1).

Additional features of the interface

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

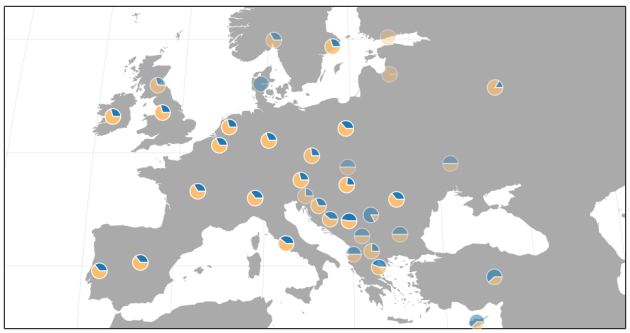

Figure 3: Example map from the Geography of Genetic Variants browser displaying the use of varying transparency of population pie charts to represent uncertainty in allele frequencies. The transparency is scaled in proportion to the number of observed chromosomes in each population for a particular variant. The frequency data and population identifiers are from (Novembre *et al.*, 2008).

Sample Frequency	Display Frequency Scale	Displayed Image
0.25	1	
0.025	0.1	
0.0025	0.01	
0.00025	0.001	

Table 1: Rare variants present a challenge for display. To address this challenge, the GGV browser changes the displayed image and the frequency scale of the map depending on the input sample frequency. As an example, a variant with a frequency 0.0025 is shown as a pie-chart that is 25% full and a frequency scale of 0.01 is marked in the legend of the map.

will be pulled towards their true origins [Figure 5]. Also, by hovering the mouse cursor over any population, a user can see the population labels and precise frequency information.

Access to the underlying frequency data

To provide an interface to the population minor allele frequency data stored in the SQL database 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. For the front-end, GET requests are submitted to the database via HTTP, returning json formatted allele frequency data and the meta-data associated with each population and genetic variant e.g. latitude, longitude, population label, sample size, and frequency scale. Genetic variants can be queried by chromosome position, rsid, or randomly. Example HTTP requests and json response can be seen in Examples 1,2,3 (below).

Discussion

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.

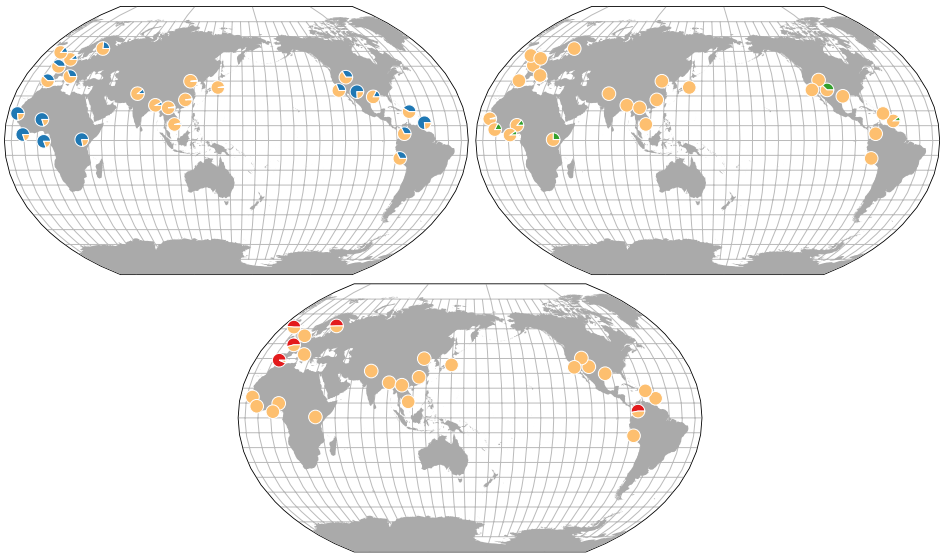

Figure 4: Example maps from the Geography of Genetic Variants browser displaying the use of frequency scales for more expressive representations of rare variation on geographic maps. The blue pie charts convey a give minor allele frequency out of 100 percent, the green out 1 percent, and the red out of .01 percent. The data are from (The 1000 Genomes Project Consortium, 2015).

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 and limitations. For example, practitioners 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 for the current defaults we use locations based on the approach taken in the source publications. A future feature will allow alternative location schema to be used for the populations in a dataset.

We also envision a variety of future extensions to the GGV that would allow for further dissection of geographical structure in large-scale population genomic datasets. Providing an interactive means of browsing neighboring variant sites near a SNP of interest would offer a unique view into patterns of linkage disequilibrium around that focal SNP. This feature would be relevant to both medical geneticists conducting genome-wide association studies with in inteterests in fine mapping as well as population geneticists interested in scanning the genome to detect signatures of positive selection. We imagine that incorporating a chromosomal browser such as jbrowse or a visualization inspired by the UCSC Genome Browser within the GGV would be greatly utilized by researchers and educators alike (Kent *et al.*, 2002).

Acknowledgements

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

Example 1: Query by rsid

```
  http://popgen.uchicago.edu/ggv_api/freq_table?data="1000genomes_phase3_table"&rsID=rs1834640
[
  {
```

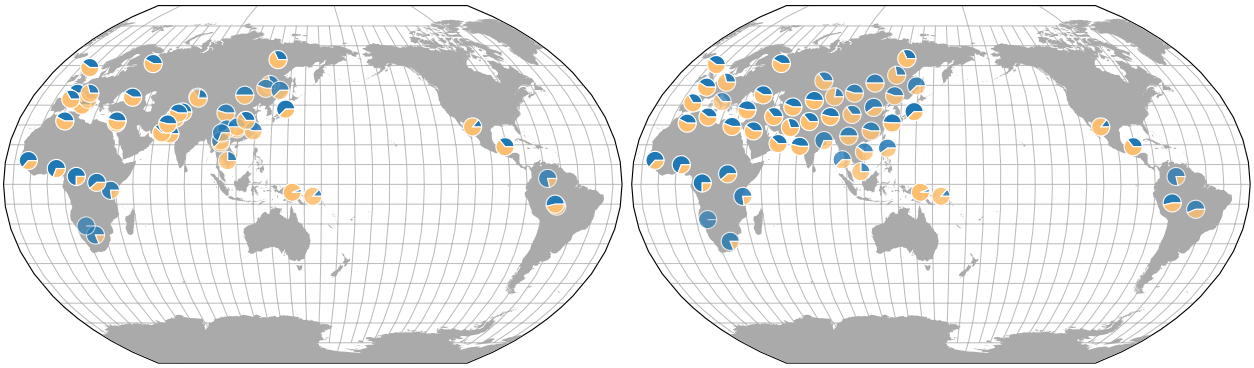

Figure 5: Example maps from the Geography of Genetic Variants browser displaying the use of a force directed layout to limit visual clutter when many populations overlap in geographic position. The left map shows the original population locations while the right shows the application of the force directed layout.


```
    "alleles": ["A", "G"],
    "pos": ["-15.310139", "13.443182"],
    "pop": "GWD",
    "nobs": "226",
    "xobs": "17",
    "freqscale": 1,
    "freq": [0.0752212389381, 0.9247787610619],
    "chrom_pos": "15:48392165",
    "rawfreq": 0.0752212389381
  },
  ...
]
```


Example 2: Query by chromosome position

http://popgen.uchicago.edu/ggv_api/freq_table?data="1000genomes_phase3_table"&chr=14&pos=37690093


```
[
  {
    "alleles": ["G", "A"],
    "pos": ["-15.310139", "13.443182"],
    "pop": "GWD",
    "nobs": "226",
    "xobs": "0",
    "freqscale": 0.01,
    "freq": [0.0, 1.0],
    "chrom_pos": "14:37690093",
    "rawfreq": 0.0
  }
  ...
]
```


Example 3: Random query

http://popgen.uchicago.edu/ggv_api/freq_table?data="1000genomes_phase3_table"&random_snp=True

```
[
  {
    "alleles": ["T", "C"],
    "pos": ["-15.310139", "13.443182"],
    "pop": "GWD",
    "nobs": "226",
    "xobs": "0",
    "freqscale": 0.01,
    "freq": [0.0, 1.0],
    "chrom_pos": "5:42452893",
    "rawfreq": 0.0
  },
  ...
]
```

References

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

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

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, *et al.*, 2009 The role of geography in human adaptation. PLoS Genet **5**: e1000500.

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

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

Furche, T., G. Gottlob, L. Libkin, G. Orsi, and N. W. Paton, 2016 Data wrangling for big data: Challenges and opportunities. In *Proceedings of the 19th International Conference on Extending Database Technology*. 473–478.

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

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, *et al.*, 2002 The human genome browser at UCSC. Genome Research **12**: 996–1006.

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

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319**: 1100–1104.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, *et al.*, 2012 A high-coverage genome sequence from an archaic denisovan individual. Science **338**: 222–226.

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science **337**: 100–104.

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

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, *et al.*, 2008 Genes mirror geography within europe. Nature **456**: 98–101.

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

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Research **19**: 826–837.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**: 904–909.

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

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

Rosenberg, N. A., L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, *et al.*, 2010 Genome-wide association studies in diverse populations. Nature Reviews Genetics **11**: 356–366.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, *et al.*, 2001 dbsnp: the ncbi database of genetic variation. Nucleic Acids Research **29**: 308–311.

Tennessen, J. A., A. W. Bigham, T. D. OConnor, W. Fu, E. E. Kenny, *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science **337**: 64–69.

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

Wessel, P., W. H. Smith, R. Scharroo, J. Luis, and F. Wobbe, 2013 Generic mapping tools: Improved version released. Eos, Transactions American Geophysical Union **94**: 409–410.