Visualizing the shared nature of human genetic variation

James Kitchens[1*] and Graham Coop[1]

[1]Department of Evolution & Ecology and Center for Population Biology, University of California - Davis
[*]Corresponding author: jkitchens@ucdavis.edu

May 16, 2023

Abstract

Due to the complexity of the data involved, understanding and visualizing patterns of human genetic variation is often challenging. Many of our tools for data visualization focus on understanding genetic population structure, PCA and STRUCTURE, emphasize the differences among people. Though appropriate for specific applications, these figures are readily misinterpreted by broader audiences. Here, we present a set of Euler diagram based visualizations as a simple tool for demonstrating the shared nature of human genetic variation.

Main

A key insight from human genetics is that, as a species, we are all very genetically similar to one another and share much of our genetic variation. Our genome can be depicted as a string of letters (A, T, G, and C), referring to the four nucleobases found in DNA. Two human genomes picked at random are identical at ∼99.9% of sites (e.g. [7]).[1] In that small fraction that doesn't match (∼1/1000 sites), your chromosome might carry an A while the other person's chromosome carries a T. The majority of sites with variation have no known function; indeed, carrying an A instead of an T may have no discernible effect on your traits. Much of the common genetic variation is shared among human groups [5]. Human geneticists are interested both in understanding which sites in the genome are functional and in unraveling the subtle differences between individuals and groups that highlight our shared history.

[1] This number accounts for only single nucleotide variants and would go down slightly if copy number variants were included.

Due to the complexity of the data involved, understanding and visualizing patterns of human genetic variation is often challenging. One helpful place to start is to visualize the global frequencies of variants at individual sites within the genome to see how variation is shared - see the Geography of Genetic Variants Browser from the Novembre Lab for a nice interactive tool [8]. However, because the human genome contains approximately 3 billion sites, it would take a few lifetimes to walk through the genome in this manner, so researchers often turn to genome-wide summary statistics to capture patterns of genetic variation. Population structure is commonly visualized using approaches like principal component plots which separate individuals along the major axes of genotypic variation. As is their purpose, these plots highlight the differences between individuals and groups, and so it can be easy to forget that their axes explain a relatively small proportion of the genetic variation observed in the subset of base pairs that vary between individuals.

Here, we share some resources for teaching human genetics using data from the 1000 Genomes Project, inspired by [3] and [1]. These visualizations first center on the variation in a set of diverse samples from the Americas (see Figure 2) before expanding to include more globally distributed examples. In a small sample of people, we expect that they vary at only a small fraction of sites in their entire sequenced genomes.[2] Most of this variation is rare, and though these rare variants can be medically salient, they are the properties of specific people and their immediate families, rather than of the larger human groups. To learn about more widely shared variation and following methods similar to those in [1], we defined a variant as “common” in a sample if it was found in more than 5% of people’s chromosomes and then filtered the data based on this criterion.

The small blue circle in the above figure captures just how little variation rises to this frequency in the Americas. As the rest of this manuscript focuses on the sharing of these common variants, it’s important to maintain perspective regarding the scale of these differences relative to the size of the human genome.

There are seven different samples from the Americas in the 1000 Genomes Project dataset (as described in [1]), each sample being made up of 60-105 people, and we counted the number of common variants found in each sample.[3]

The levels of genetic diversity, shown as differences in the number of common variants, vary between samples: African Caribbean in Barbados (ACB) and African Ancestry in Southwest US (ASW) display the highest levels of variation. Similar to Figure 1 from [3], we implement an Euler diagram to visualize the amount of overlap in common genetic variation between samples (Figure 3). This style of visualization is like a Venn diagram, with the added property that the areas and overlaps of the shapes are proportional to the number of common variants in the corresponding samples.

It’s clear from this figure that the majority of common variants are not

[2] If we sequence the entire population of the world, we’d see nearly every site being variable in some one. But these variants would be vanishingly rare in the population, overall.

[3] It would be interesting to explore rarefaction approaches to account for the differences in the sample size [2].

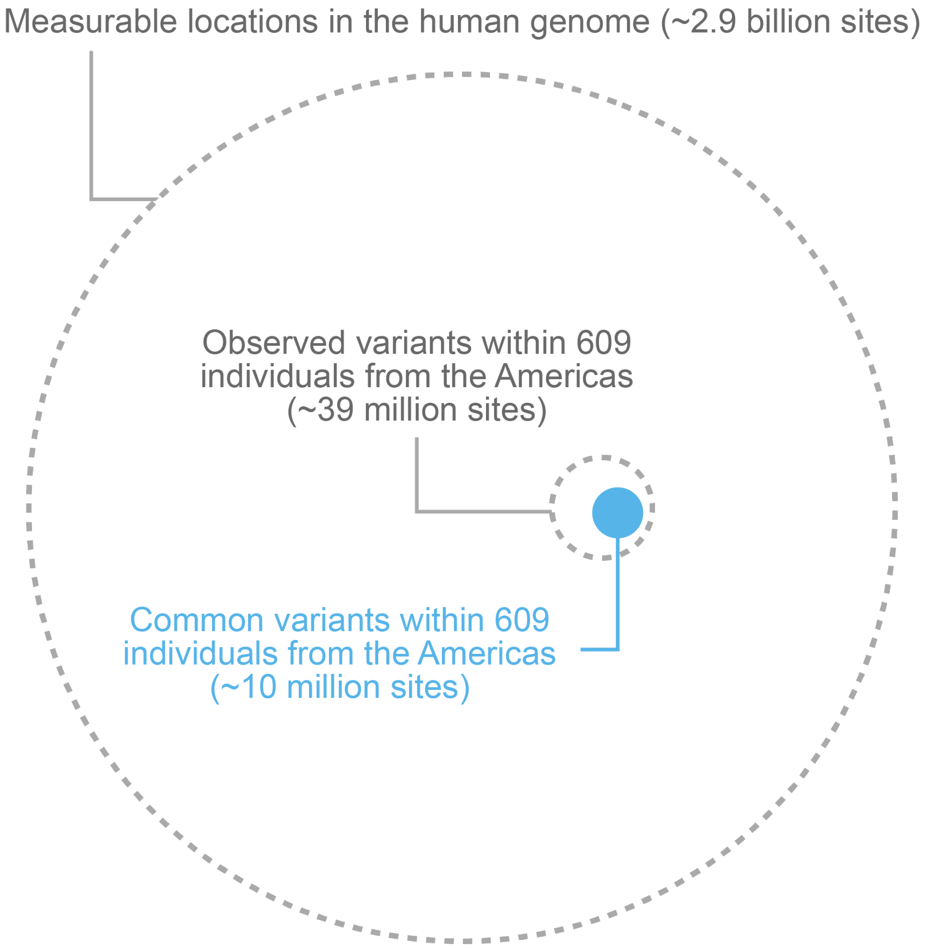
Measurable locations in the human genome (~2.9 billion sites)
Observed variants within 609 individuals from the Americas (~39 million sites)
Common variants within 609 individuals from the Americas (~10 million sites)

Figure 1: The scale of common variants in the Americas compared to the human genome. The area of each circle is scaled proportionally by the number of sites in that category. The small blue circle corresponds with the number of common variants; “common” is defined as having a minor allele frequency of greater than 5% in at least one of the samples.

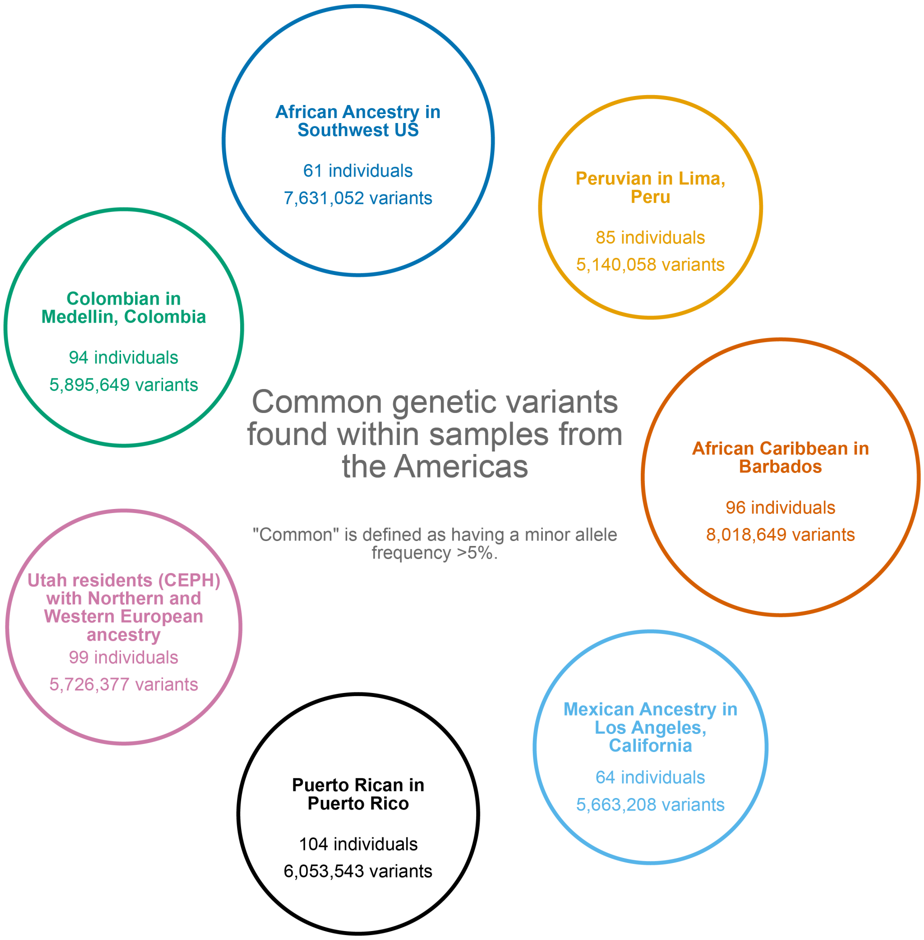

Figure 2: Number of common variants in seven samples from the Americas. The area of each circle is proportional to the number of common variants within that sample from the 1000 Genomes Project. A "common" variant is defined as having a minor allele frequency of greater than 5%, where the minor allele identity is determined by its global allele frequency (its frequency across all samples in the 1000 Genomes Project). The number of individuals within each sample has also been included to ensure that this quantity is relatively consistent between samples.

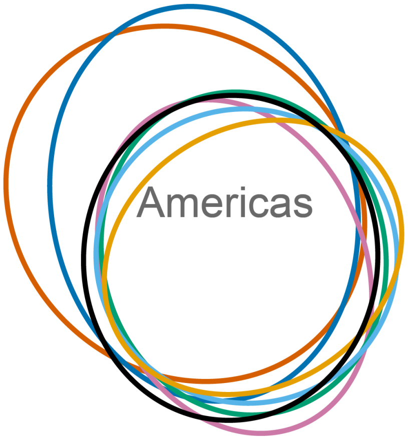

Figure 3: Overlap in common variants between samples from the Americas. See Figure 2 for the color legend. It is not mathematically possible to generate ellipses with a given overlap without distortions to the areas. See the Technical details section (below) for statistics quantifying the slight errors in this and following Euler diagrams.

unique to a single sample. Instead, they are often widely distributed and shared between samples, resulting in a large degree of overlap between ellipses. The African Caribbean and African American (ACB and ASW) samples share nearly all of the common variation found in other samples. However, as noted above, they also have greater amounts of genetic variation compared to that found in the other samples (larger area), and some of that variation is not common in the other samples from the Americas. This does not mean that these variants are completely absent from the other groups, but instead, that these variants are rare or undetected in the other samples included in the figure. For example, maybe 10% of people's chromosomes in the ACB sample carry a T instead of an A at a particular site, but this T is found in only 1% of the CEU sample.

To look at the overlap in a different way, we first considered the variation that is common (>5%) in a given sample and then identified in which other samples the variant is also common.

This method of filtering results in an Euler diagram where the ellipse of the highlighted sample completely encircles the other ellipses. A sample with greater numbers of common variants that are not common in other samples will show a larger disparity in size compared with the other ellipses. As before, these figures illustrate the high degree of sharing of variation among samples in the Americas. The African Caribbean in Barbados (ACB) and African Ancestry in Southwest US (ASW) samples contain the most genetic diversity, with some of this variation being shared only between those two samples. In comparison, there is somewhat less common variation (small diagram size) in the other samples and nearly all of it is shared.

Zooming back out and putting Figure 3 back onto the scale of the whole genome, the Euler diagram shrinks down to match the fraction of common

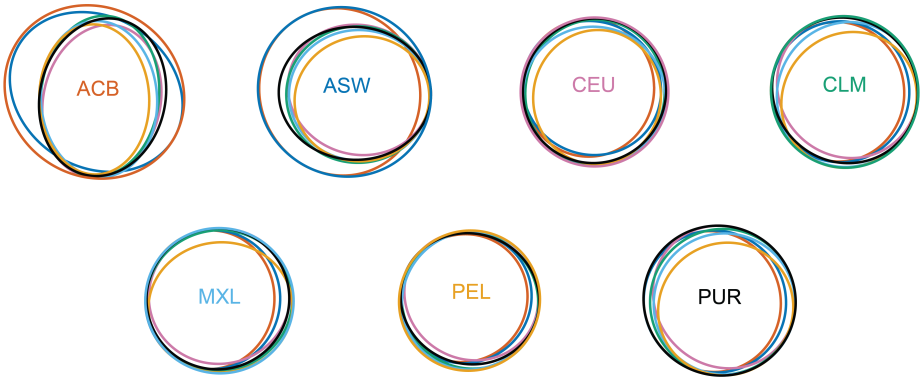

Figure 4: Sharing of common variants found in each sample from the Americas. Each diagram highlights a different sample, identified in the title, that was used to filter the variants down to only those that were common in the sample. The sizes of each plot are proportional to the number of variants included in the analysis (sizes are not proportional to previous figures). See Figure 8 near the bottom of this manuscript for an alternative visualization of this figure.

variants in the genome.

Genetic diversity in the Americas reflects the history of colonialism and the transatlantic slave trade, which has moved people from across the globe into the region over the past few hundred years. Given this, you may wonder whether the high degree of overlap reflects this recent history of the Americas or whether it is representative of sharing that is present in geographically distant samples. To look into this question, we created an Euler diagram with five samples, one from each of the broad geographic groupings used by [1] (Figure 6).

Overall, this diagram has a very similar structure to the diagram created with the samples from the Americas. There is a high degree of overlap between all of the samples, with the higher genetic diversity of the Yoruba in Ibadan, Nigeria sample resulting in a larger ellipse that stretches outside of the cluster of other ellipses. This pattern matches the one of high diversity in the African Caribbean and African American (ACB and ASW) samples from the Americas described above. Even when considering quite geographically distant samples of humans, the dominant pattern is that of shared genetic variation.

Lastly, given this global view, we can zoom in and look at how variation is partitioned at finer geographic scales by using all 26 samples within the 1000 Genomes Project dataset. We see that samples from Africa contain the greatest amount of genetic diversity. Much of that common genetic variation is shared, but each sample contains some variation not found in other samples. There's a slight reduction in the variation present in samples whose recent ancestors lived outside Africa, consistent with the view that humans evolved in Africa, and when humans first migrated out of Africa, they took with them only a subset of the genetic diversity present in Africa.

It's easy for us to fall into the trap of thinking that humans are very genet-

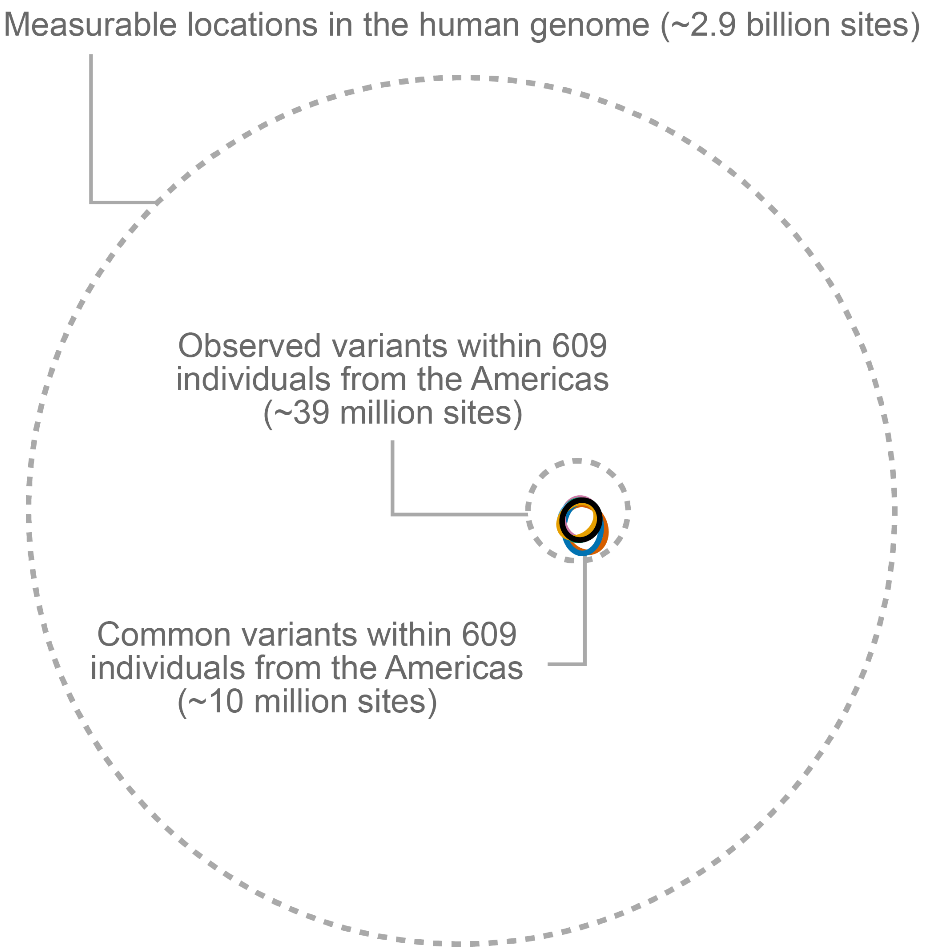
Measurable locations in the human genome (~2.9 billion sites)
Observed variants within 609 individuals from the Americas (~39 million sites)
Common variants within 609 individuals from the Americas (~10 million sites)

Figure 5: Common variants in perspective. An Euler diagram of the common variants in samples located in the Americas relative to the scale of the human genome. As a small note, the positions and orientations of ellipses within the Euler diagram differ slightly from Figure 3. This is because the ‘eulerr‘ package gives varied results with each run due to random starting conditions within the algorithm.

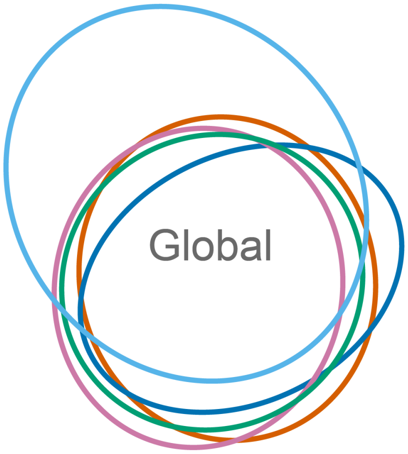

Figure 6: Overlap in common variants between samples from global samples. An Euler diagram of the common variants in five geographically distant samples: Bengali in Bangladesh (BEB), Han Chinese in Beijing, China (CHB), British in England and Scotland (GBR), Mexican Ancestry in Los Angeles, California (MXL), and Yoruba in Ibadan, Nigeria (YRI).

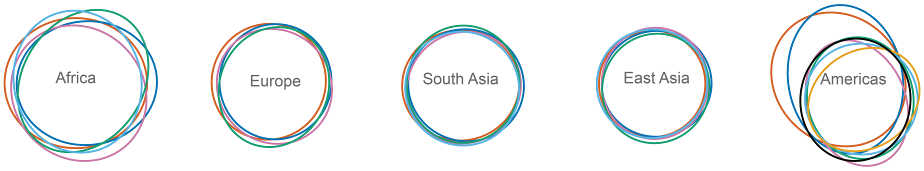

Figure 7: Sharing of common variation within geographic regions. Five Euler diagrams of the 26 global samples using the broad geographic groupings from [1]. See Table 2 in the Technical details section for the color legend for each subfigure.

ically different. Historically, our ideas about the structure of human biological variation have been shaped by a few visible physical traits, notably skin color, that have a geographic pattern. But the genetic variants contributing to skin pigmentation are unrepresentative of the more general patterns of genetic sharing present among groups of people sampled from across the world. The genetic changes involved in skin pigmentation differences can show striking geographic patterns (e.g SLC24A5), but that is because they have been shaped by strong local adaptation to the climatic conditions that people encountered as they moved around the world. These loci are fascinating examples of adaptation but are also the exception in comparison to the high degree of sharing that we see for most of human genetic variation.

Acknowledgements

We thank John Novembre, Arjun Biddanda, Molly Przeworski, Doc Edge, and the Coop Lab for their help during the editing of this manuscript.

Technical details

We used the ‘geovar‘ package in Python to group the $\sim$92 million variants included in the 1000 Genomes Project based on minor allele frequency (MAF) [1]. Variants were separated into five bins based (MAF=0%, 0%<MAF<1%, 1%<MAF<5%, 5%<MAF<10%, and MAF>10%), though two bins would have sufficed for this analysis (MAF<5% and MAF>5%). We used the ‘eulerr‘ package in R to calculate the position and orientation of ellipses in the Euler diagrams [4]. Unfortunately, exactly proportionally scaling the area of every region of this diagram becomes difficult to impossible as you increase the number of sets, or samples. Because of this, we have included two goodness-of-fit measurements provided by the ‘eulerr‘ package and described in further details in the package’s tutorial. For both measurements, values closer to zero have less error. Below is a table with these measurements for all of the Euler diagrams presented in this manuscript:

Table 1: Euler diagram statistics

figure	stress	diagError
3	4.53E-04	0.019573681
4_ACB	1.09E-04	0.008677686
4_ASW	1.52E-04	0.0133913503
4_CEU	1.04E-04	0.016955049
4_CLM	3.08E-04	0.0260526896
4_MXL	2.53E-04	0.0236936357
4_PEL	1.89E-04	0.0181390286
4_PUR	3.58E-04	0.0288044056
5	1.09E-09	0.0001789465
6	2.30E-03	0.0198055721
7_Africa	1.19E-03	0.027619266
7_Europe	6.39E-07	0.0004395497
7_South_Asia	6.41E-04	0.0458260097
7_East_Asia	5.77E-04	0.0410841878
7_Americas	4.53E-04	0.019573681

The package also breaks down error by set overlap to better understand exactly which sections are over-/underrepresented by the visualization, though that is not included here. With all of that being said, these diagrams offer a unique visualization method that can be particularly useful for more qualitative interpretations of the population relationships. We converted the output of ‘eulerr‘ into a JSON format and passed this to JavaScript for plotting using D3.js. Plotting is possible directly from R, but we used D3.js for its customizability and support of interactive figures. Interactive versions of these figures are available at https://james-kitchens.com/blog/visualizing-human-genetic-diversity, and all of the figures (alongside the code we used to generate them) can be found at https://github.com/kitchensjn/visualizing-human-genetic-diversity.

Table 2: Figure 7 Color Legends

Africa	Esan in Nigeria (ESN), Gambian in Western Division, The Gambia - Mandinka (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), Yoruba in Ibadan, Nigeria (YRI)
Europe	Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), Toscani in Italy (TSI)
South Asia	Bengali in Bangladesh (BEB), Gujarati Indians in Houston, TX (GIH), Indian Telugu in the UK (ITU), Punjabi in Lahore, Pakistan (PJL), Sri Lankan Tamil in the UK (STU)
East Asia	Chinese Dai in Xishuangbanna, China (CDX), Han Chinese in Beijing, China (CHB), Han Chinese South (CHS), Japanese in Tokyo, Japan (JPT), Kinh in Ho Chi Minh City, Vietnam (KHV)
Americas	African Caribbean in Barbados (ACB), African Ancestry in Southwest US (ASW), Utah residents (CEPH) with Northern and Western European ancestry (CEU), Colombian in Medellin, Colombia (CLM), Mexican Ancestry in Los Angeles, California (MXL), Peruvian in Lima, Peru (PEL), Puerto Rican in Puerto Rico (PUR)

Additional figures

The following figures offer alternative methods of visualization to those within this manuscript. Details about these figures are provided in the figure captions.

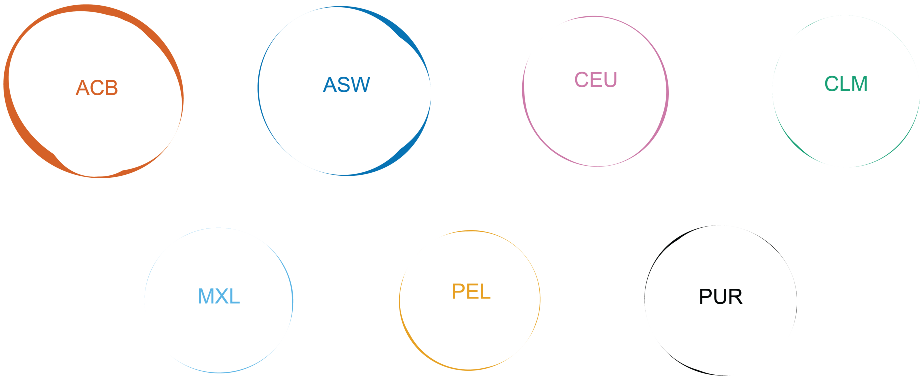

Figure 8: Sharing of common variants found in each sample from the Americas. Seven "coffee stain" diagrams, an alternate visualization of Figure 4. The colored area is proportional in size to the number of common variants within the highlighted sample, identified in the title, that aren't shared with another sample. Within each subfigure, the ellipse on the bottom corresponds with the highlighted sample and is filled in with that sample's respective color. All other ellipses are filled in with white and stacked on top, thus giving the appearance of cutting out the area and leaving only the common variants that aren't shared with another sample.

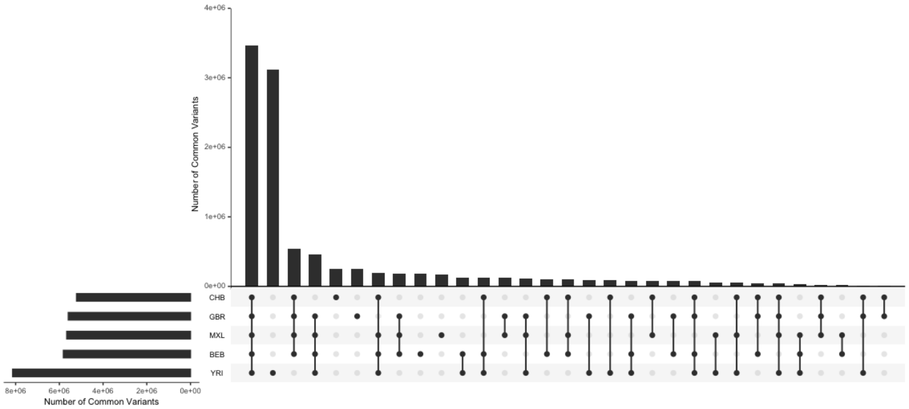

Figure 9: Sharing of common variation within geographic regions. An UpSet plot, an alternative visualization of Figure 6. UpSet plots, created by [6], are useful for handling large numbers of sets. They can communicate the exact overlap between sets, unlike Euler diagrams (as discussed in the Technical details section), but are also a bit more challenging to read as there are multiple subfigures. To draw comparisons with the Euler diagrams, the horizontal bar graph on the bottom left depicts the areas of the ellipses and the vertical bar graph shows the areas of the overlaps between ellipses referenced usings dots in the bottom subfigure.

References

[1] Arjun Biddanda, Daniel P Rice, and John Novembre. A variant-centric perspective on geographic patterns of human allele frequency variation. *eLife*, 9:e60107, December 2020.

[2] Daniel J Cotter, Elyssa F Hofgard, John Novembre, Zachary A Szpiech, and Noah A Rosenberg. A rarefaction approach for measuring population differences in rare and common variation. *GENETICS*, 224(2):iyad070, May 2023.

[3] Brian M. Donovan, Rob Semmens, Phillip Keck, Elizabeth Brimhall, K. C. Busch, Monica Weindling, Alex Duncan, Molly Stuhlsatz, Zoë Buck Bracey, Mark Bloom, Susan Kowalski, and Brae Salazar. Toward a more humane genetics education: Learning about the social and quantitative complexities of human genetic variation research could reduce racial bias in adolescent and adult populations. *Science Education*, 103(3):529–560, May 2019.

[4] Johan Larsson and Peter Gustafsson. A Case Study in Fitting Area-Proportional Euler Diagrams with Ellipses using eulerr. *Proceedings of International Workshop on Set Visualization and Reasoning*, 2116:84–91, 2018.

[5] R. C. Lewontin. The Apportionment of Human Diversity. In Theodosius Dobzhansky, Max K. Hecht, and William C. Steere, editors, *Evolutionary Biology*, pages 381–398. Springer US, New York, NY, 1972.

[6] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, December 2014.

[7] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P. Spence, Yun S. Song, Giovanni Poletti, Francois Balloux, George Van Driem, Peter De Knijff, Irene Gallego Romero, Aashish R. Jha, Doron M. Behar, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L. Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M. Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M. Beall, Anna Di Rienzo, Choongwon Jeong, Elena B. Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M. Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T. S. Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F. Hammer, Toomas Kivisild, William Klitz, Cheryl A. Winkler, Damian Labuda, Michael Bamshad, Lynn B. Jorde, Sarah A. Tishkoff, W. Scott

Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, October 2016.

[8] Joseph H Marcus and John Novembre. Visualizing the geography of genetic variants. *Bioinformatics*, 33(4):594–595, February 2017.