Visualizing the distribution of neanderthal-derived alleles across the globe

Task description

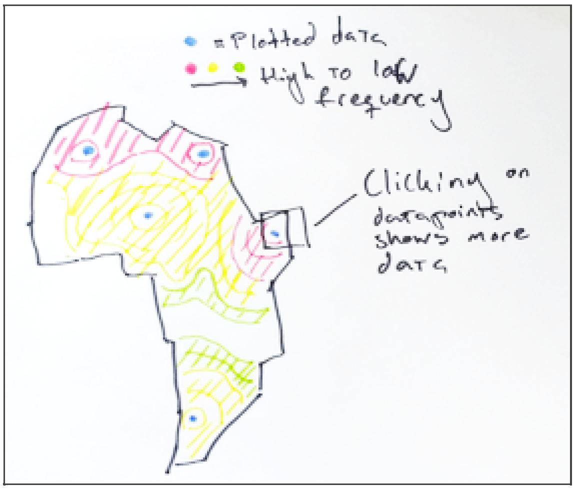
The aim of my data visualization project is to create a geospatial visualization of 42 SNP variant frequencies that were identified in a 2010 publication of a draft neanderthal genome. Technically speaking, the aim of my project is to reduce the human genome diversity project (HGDP) data from over 600,000 SNPs to the 42 SNPs identified as being potentially indicative of neanderthal ancestry. Then I aim to create datasets that describe the frequency of the "Out Of Africa" alleles for those 42 loci that are grouped by the geographic regions represented in the 1,043 samples, and visualize the total combined frequency of all the neanderthal-derived (ND) alleles, or the frequency of the individual alleles in any of the 42 loci. A user should be able to get a good overview of the frequency in different regions across the world and be able to find out more about a specific region.

In short, I aim to visualize how frequencies of a selection of 42 ND alleles vary across the globe — both the combined and individual allele frequencies. The final visualization can be viewed online.

Design

The two first ideas for the project each involved somehow visualizing the distribution of ND alleles over the world, by categorizing the samples from the HGDP dataset by population or geographic location. The third idea revolved around representing allele frequencies without the aid of a map.

Heatmap

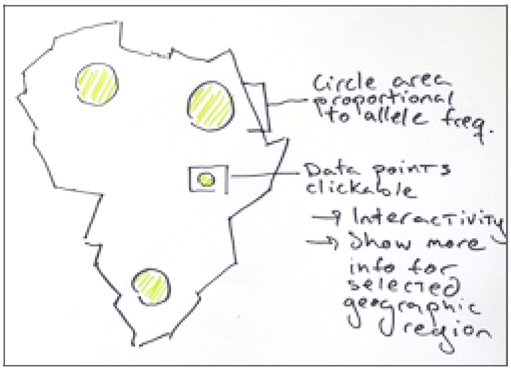

● = Plotted data
● ● ● → High to low frequency
Clicking on datapoints shows more data

The first idea was to represent the allele frequencies in the form of a heatmap. Color gradients overlaid on a map would indicate the frequencies in different locations on the map. In the figure above, this concept can be seen implemented over a sketch of the African continent. The frequencies for one or more alleles are visualized over the map, and a gradient going from red or green depicts high to low frequencies respectively. The interactive component of the map could then be a menu with a slider or some other form of selection that would enable the user to choose which SNP to visualize, or to switch between visualizing individual or combined allele frequencies. The location center of the data (blue points) could be made clickable, which could allow for more interactivity, like displaying more information about the selected region.

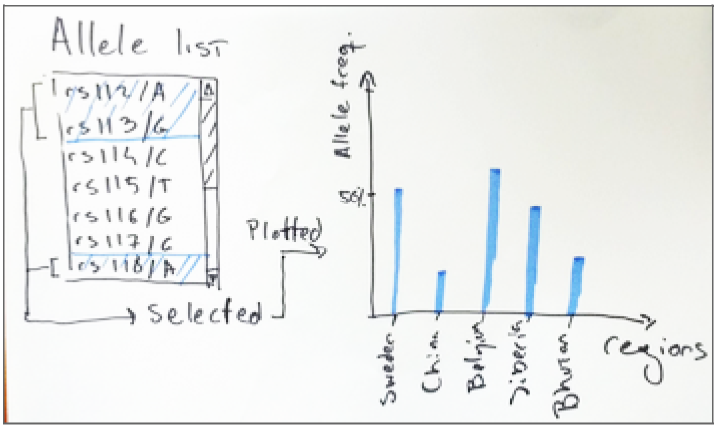
The upside of this approach is that it is a very intuitive visualization, that can be used to identify allelic hotspots on a map, the downside is that in order to be truly useful it probably requires data with high geo-spatial resolution. That is to say, creating this visualization using observations grouped into a few discrete regions would probably be better achieved using a bubble chart.

Bubble map

The second idea is similar to the heatmap idea, but instead uses bubble charts overlayed on the map to visualize allele frequencies. The area of the bubble chart corresponds to the frequency represented by the observations. Bubbles can be made clickable for added interactivity, or the user can hover over them with a cursor for added data display. It is a simpler visualization when compared to the heatmap, but it works better when the observations have lower geo-spatial resolution, as is the case with my data. Interactivity could perhaps work the other way, with the user being able to expand or restrict the dataset in a menu, and have the visualization on the map change accordingly. For example, instead of showing the combined frequency, a user can choose to display the frequency of a subset of alleles.

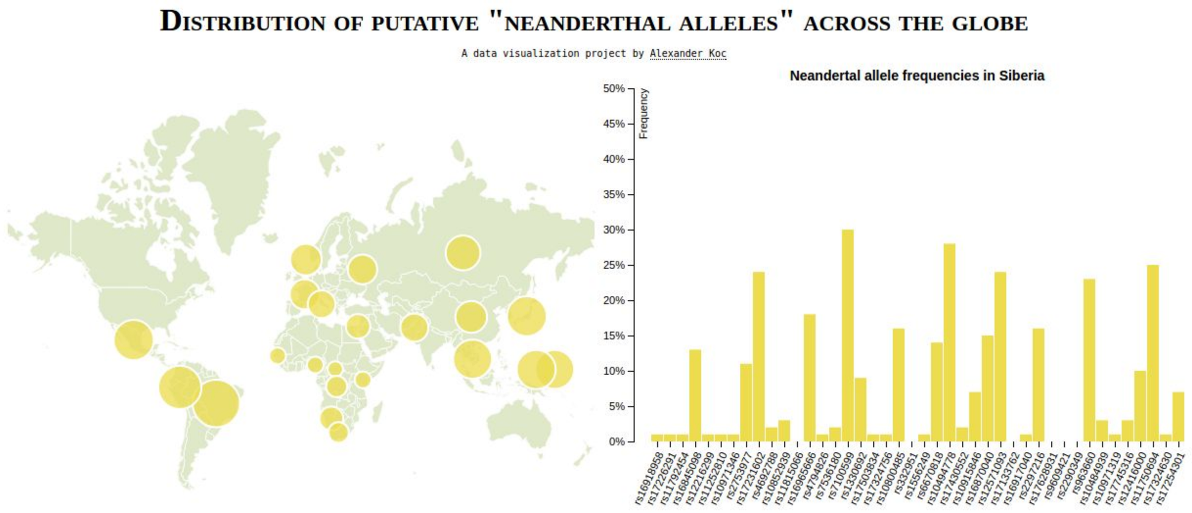

Circle area proportional to allele freq.
Data points clickable
→ Interactivity
→ Show more info for selected geographic region

The good old bar plot



My third idea was to create a simple visualization of allele frequencies per country, which does not rely on a map, but instead features a small number of selection menus, in which the user selects a subset of the data they are interested in, and that subset would be visualized in simple plots, like a bar plot. Users could study and compare the frequencies of a subset of alleles between a subset of countries, and ask questions like: do regions with similar total neanderthal frequencies have a similar distribution.

Implementation



DISTRIBUTION OF PUTATIVE "NEANDERTHAL ALLELES" ACROSS THE GLOBE

A data visualization project by Alexander Koc

In my implementation, which can previewed on a public page hosted on Dropbox, I choose to combine two of the proposed designs above: **the bubble chart map** and **the bar plot**. The rationale behind this design decision is that it allows the user to explore both the combined and individual allele frequencies side by side. The two figures are interactive and contain elements that display more info when hovering over them with the cursor. To implement the designs, I primarily used the javascript library D3.js and the D3-derived javascript library Datamaps.

The visualization takes the shape of a simple webpage enhanced with javascript. A world map is implemented using the DataMaps javascript library. A bubble chart is overlayed on top of the world map, where each bubble is tied to a specific geographic location and reflects the frequency of ND alleles in that location. The radius of each bubble is calculated by treating the ND allele frequency value as a circle area value, from which the circle radius is calculated and scaled up by a constant M: $r = \sqrt{\frac{A}{\pi}} \times M$

The second part of the visualization comes in the form of a bar-chart to the right of the world map. The bar chart is implemented using D3.js and allows the user to study the individual frequencies of each of the 42 alleles for a chosen geographic location. To display the data for a region of interest, the user can click on one of the bubbles on the world map. This calls a javascript function that identifies the geographic location the bubble belongs to, and tells the bar-chart to access the appropriate allele dataset.

Further interactive elements include the SNP ids in the barchart being clickable, redirecting to the appropriate SNPedia page. Hovering over the bubbles or the bars in the bar-chart displays tooltips showing the frequency associated with the datapoint.

Insights

The resulting visual shows a few interesting things about the distribution of the ND alleles across the world. First off, the frequency of all ND alleles in Africa is low compared to the rest of the world. The African frequencies range from 4 to 8% counting all 42 markers.The highest frequencies can be found in South-East Asia (22-23%) and in South and Central America (24-27%), while Europe and the Middle East fall somewhere in between (11-15%). It seems that the frequency of the ND alleles increases approximately along the human migration patterns.

The visualization echoes observations already made elsewhere[1]: Sub-Saharan populations show considerably less Human-Neanderthal admixture than the rest of the world, which makes sense as the ND alleles are all “Out of Africa” variants. Besides, neanderthals — from what we know — were mainly found in western Eurasia.

East-Asian and American populations show a higher neanderthal admixture compared to Europeans: could it be that there was a second admixture event as modern humans migrated to the East? Or Did the frequency of neanderthal-derived alleles decrease in Eurasia due to a second migration of humans from Africa into Eurasia after a first admixture event?

There are some insights or leads for further research that can be found when comparing the frequencies of individual alleles between the regions. In the East Asian populations, four allelles (rs10800485, rs6670818, rs963660, rs11750694) stand out above the rest and seem to increase in frequency when moving east. Result of selection? Genetic drift? In contrast, European and West-Asian regions seem to have a more even distribution of the ND alleles, in their population (around 0-15% per ND allele). The samples found in Central and South America indicate a higher frequency of ND alleles even compared to South-East Asia. The topology of the frequency bar-plot for Mexico seems to resemble Siberia more than the plots in South-East Asia.

In short, this visualization enables a simple visualization of a selection of SNP variant frequencies in a population by geographic region. The geospatial nature of the visualization enables identifying populations with similar frequencies by eye, and perhaps allows for making hypotheses about migratory routes. The current version covers 42 putative neanderthal alleles, but could be adapted to visualize any other subset of SNPs.

The finished assignment can be previewed at:
https://dl.dropboxusercontent.com/u/16264689/neanderthal-map/index.html

The associated screencast can be viewed at:
https://www.youtube.com/watch?v=0_DYL0jI59Y

[1] The Wikipedia article gives a good overview:
https://en.wikipedia.org/wiki/Archaic_human_admixture_with_modern_humans#Neanderthals