Title

Mapping Populations with Genomes from the 1,000 Genomes Project

Context

The 1,000 Genomes Project set out to understand genetic variation throughout the world by comparing the DNA sequences of people in 26 populations across 5 continental regions including the nations: China, Italy, Japan, Kenya, Nigeria, Peru, the United Kingdom, and the United States. In doing this, they generated over four terabytes of DNA sequence data from 2,504 genomes.

This project's goal is to use some of that data to map a correlation between DNA and geographic location. This is possible because DNA within populations are more similar to each other than to DNA outside their populations. So DNA can be grouped by similarity, and these groups have a geographic basis.

There have been similar projects in the past: The Human Genome Project and the HapMap Project. The Human Genome Project was the first effort to sequence an entire genome. The HapMap project is similar to the 1,000 Genomes Project in that its effort was with sequencing multiple genomes, although not the entire genome. Some of the HapMap data was used for the 1,000 Genomes Project research.

One limitation to the data is that the researchers used 'surrogate' populations, for example using Indian Telugu in the U.K. to represent Indians, or those of African descent in the USA. A sampling of these people may not be representative of the population. Another limitation is the sheer amount of data that the project has generated. Having that much data may be worse than having less.

Data

The data originates from volunteer donors, as well as from the Human Genome Project and the HapMap project. There were originally supposed to be 1,000 individual genomes sequenced, but with the advancement of sequencing technology and methodology, over 2,000 individual sequences were used over the three-year project.

The sequence data has been anonymized, however that is like saying a person's fingerprint is anonymized. It can still be traced back to the person.

A major theme of the project is global collaboration. The data was generated from globally diverse individuals, and was contributed to and researched by globally diverse people. The purpose of the project is to better understand the variation of the human genome, and so it is fitting that the project is global. A requirement for the project researchers was to make their data available within 72 hours of its generation so that it is shared. With the intended availability of the data being free and open, there are likely no issues with accessing the data.

We may be using up to four terabytes of data, and we can store this either on a local hard disk or on an online storage cloud.

Available data includes: use raw, aligned, and variant. These come in file paths: fastq, cram, and vcf, respectively. There are software written to read sequence data that are available online for free.

To acquire this data we can use lists of ftp file paths that have been supplied by the database website which lead straight to file downloads. The pipeline follows a number of those paths, depending on how it is configured. Once the data is stored, the data is fed into a conversion process to bring the data into vcf format. Finally, with the vcfs, the data can be filtered and have pca applied to reduce it to two dimensions for final analysis.

Data Cleaning & Explanation

Plink is a genome analysis toolset. It will be used with VCF genome data. It will filter the data's SNPs by a MAF of .05, a maximum per-SNP missing of .1, and a maximum per-person missing of .1, output the files: bed, bim, fam, ped, and map, and compute the top two principal components.

Minor allele frequency (MAF) is frequency at which the second most common allele occurs in a population. A filter value of .05 is the value used by the HapMap project, but I went with .10. This ended

Maximum per-SNP missing will filter the SNPs that are missing no more than a percentage of the time, in our case, 10%, which is the default value for Plink.

Maximum per-person missing will filter out the genomes of people who are missing too much data, in our case, any more than 10%, which is the default value for Plink.

Files from all of the chromosomes will be used.

UCSD's DSMLP server will be used. Plink comes installed, and the necessary VCF files are graciously available offline.

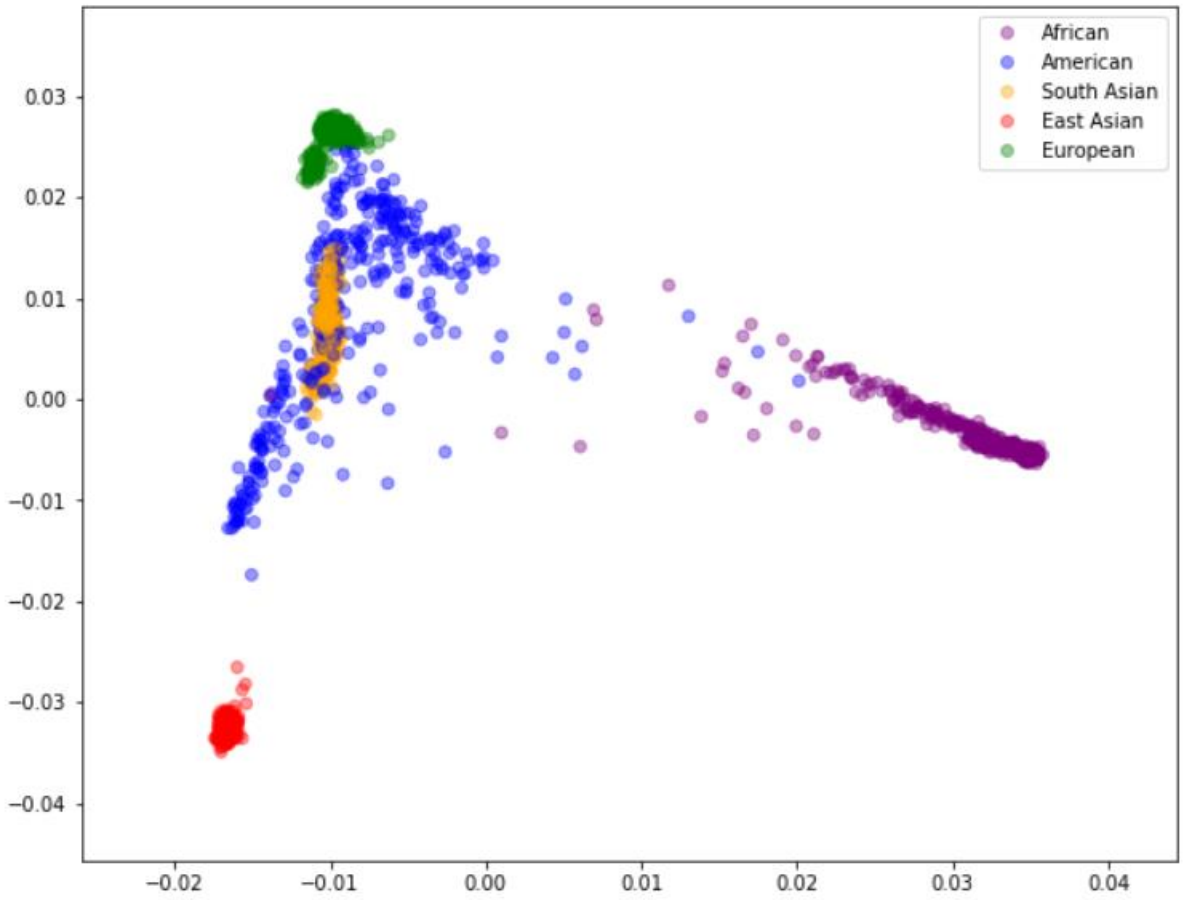
I concatenate the files using the bcftools concat command, then filter and convert the files to binary format using plink2, and finally compute the eigenvectors with plink2.

I started with 73,159,510 variants from 2,548 people. Nothing was removed due to the –mind filter, and Nothing was removed due to the –geno filter, but 67961911 variants were filtered due to the minor allele frequency filter. So after filtering, I was left with 5197599 variants and 2548 people.

However, there were 44 individuals who did not have their populations labeled, so I removed those from my plot. So here is my PCA plot of the 2504 remaining people.

Outlier Discussion / PCA Visualization

Each circle is a person, and they are colored according to the super-population they came from. For example, the red circles from East Asia could be from different parts of China, or from Japan. Or, the blue circles could be from Puerto Rico, Peru, or Columbia.



Analysis

We see that East Asians, Europeans, and Africans are the most distinct populations, being furthest apart. We find lots of overlap between Americans and South Asians. If we add a third dimension, the overlap between Americans and South Asians may disappear, but this plot can show that the difference between the three distinct groups is greater than any other differences between populations, which is why the PC axes plotted the way that they did.

Limitations of approach

In addition to the data limitations discussed in the context section at the beginning, there are some limitations of my approach overall.

Firstly, I am not completely familiar with the genome analysis tools, and so for the sake of completing the project, I went with the safest commands and filters. But with more expertise, I could better fine-tune or experiment with filters or other commands and may find other results.

Second, similar to the previous point, I am not used to handling this much data, and so I went with the offline data files that were graciously available. But ideally, I should bring all of the data from their raw sequence states in FASTQ format all the way to VCF, and then maybe repeat the same steps I took.

Conclusion

I think that, overall, the project was a success. I was able to plot the distinctions between populations based on their genomes, which was the goal of the project.

In reflection, I was surprised that the groups plotted as distinctly as they did. Conceptually, it makes sense that different geographic origin would account for the most genetic diversity, and I've seen the paper discussing the same results I found, but to actually see the process of the distinctions coming from the sequence data was cool. Instead of taking the scientist's word that this is what they did and this is what they found, I was able to generally replicate what they did and can see what they found (of course, I'm using the same data that they released, but if I were to totally replicate the project, I would need volunteer donors and sequencing machines).

It's amazing that we can know the genetic sum of a person just from little samples of them. I hope that we will use this new technology with wisdom and care – that we respect people's genomes as we would respect their identity, and never use their data without their consent.