Genes Reveal Geographical Relationships Worldwide: a replication of the main results in "Genes Mirror Geography within Europe", using the data from 1000 Genomes Project

Yifei Ning
y3ning@ucsd.edu
UC San Diego
GitHub: https://github.com/Couson/gene-mirror-geo

1. Introduction

1.1 Overview [1] [2]

In the article called Genes Mirror Geography within Europe, interested in revealing how genetic and geographic distances correspond to one another, John et al conduct the study on a sample of 3192 European individuals and collected millions of variable DNA sites in the human genome. They found a close association exists between the geographical map of Europe and a two-dimensional representation of genetic variations in Europeans. Most importantly, their research suggests that by using genotyping technologies, one could discover how population structures are associated with one's genetic information. In this project, I will replicate the results from Genes Mirror Geography within Europe while using the data from "The 1000 Genomes Project". Instead of studying European individuals based on their genetic variation sites, I am interested in how accurately we could associate the sample's origins from "The 1000 Genomes Project" with their genetic information. My project will be stored in a GitHub repository so that any client could run this repository to replicate the main results in "Genes Mirror Geography within Europe" from end-to-end.

1.2 Background Knowledge

Firstly, genotyping refers to the process of determining the genetic variants/ differences which could lead to changes in phenotype. And sequencing refers to the process that constructs the human reference genome. In the 1000 Genomes Project, the sequencing was done based on the sample collected: 2504 individuals' DNA sequences. It is known that the DNA sequence of humans is determined by four types of nucleotides: A, T, C, and G.

Additionally, it's would be beneficial to keep in mind some background knowledge about how sequences of DNA are read, aligned, and stored. When a bio-specimen is collected from an individual, biologists will male millions of copies of DNA from that sample (this process is called Polymerase chain reaction or PCR). Later on, those sequences of DNA will be divided into segments of a suitable size. In this stage, many fragments that overlap each other will be generated. FASTQ files are used to store the raw data—fragmental reads. After that genetics experts will assemble the sequence data from those fragments to determine overlap and filter out those with low quality, making a contiguous sequence with confidence scores measured at each position. At this stage, BAM is used to represent the ("B" stands for binary) aligned sequences. At last, when the results of the BAM files are compared against the reference sequence, the genetic variation calls (either 0 or 1) for an individual will be stored in VCF files.

2. Data

2.1 Data Overview [4]

The data I will use is from "The 1000 Genomes Project" (1000 Genomes). It studies the human genetic variation from a diverse sample of 2504 individuals coming from 26 different populations Africa (AFR),

East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR). Those data are appropriate to study the human genetic variants since the samples collected take into consideration the population's size, geometrical factors, population ancestry. There are mainly two concerns when collecting demographic data. The size of each sample data collected requires roughly a representation of the population size. So, more samples are drawn from regions with higher populations. Also, the sample drawn needs to be representative – needs to cover a wide geometry range.

The major three file types I'll encounter are FASTQ, BAM, and VCF which are all available through the "1000 Genomes" Data Portal. In this project, to establish a contiguous pipeline, I will develop programs to warp existing Genome Analysis toolkits (like GATK and bcftools) and convert FASTQ to BAM and BAM to VCF. Among the three types of files, I will be mainly dealing with VCF of chromosomes 1 to 22 because the question I am interested in is how genetic ancestry is associated with genetic information regardless of gender. Starting from those VCF files, I will do a range of operations including filtrations, combinations, Principal Component Analysis, and Clustering techniques to answer my research question above.

2.2 Data Ingestion

"1000 Genomes" Data are available using the FTP server at http://ftp.1000genomes.ebi.ac.uk. Different formats of files are located differently in the server. FASTQ (raw data) files are stored in the sub-path vol1/ftp/phase3/data/fastq/SAMPLE/sequence_read; BAM (alignment) files are stored in the path: vol1/ftp/phase3/data/SAMPLE/ alignment; VCF (variant call) files are stored at vol1/ftp/release/20130502. Since the files are stored at different locations by their types, the general data ingestion process should reflect the logistics as such:

1. Request single data through three function calls; each function should only get one type of files among FASTQ/ BAM/ VCF since they are at different locations on the ftp server
2. Request a batch of data by parameters or inputs specified by a configuration file.
3. If the data files requested already exist in the local disk, discard the ftp request and use the data file existed. This is to minimize the disk or memory loads.
4. Convert FASTQ to BAM if applicable using BWA
5. Convert BAM to FASTQ if applicable using GATK
6. Parameters that are used to specify the file format(s), sample(s), or chromosome(s) should be stored separately in a JSON file. So, each time when pulling data, the parameter JSON file is the only file that needs to be changed.
7. Data requests are reproducible

2.3 Concerns

There are no potential privacy issues since all the data collected was not associate with the privacy information of any individuals. Therefore, all data currently remains safe in terms of privacy issues. Moreover, all data is accessible for all scientists for academic purposes, so it has been coded and preserves to not be used for any private purpose.

Another concern about pulling data is the storage and pulling speed. Since the data size is too large. Each VCF file on average is in GBs and requires hours to pull all of them. The planned method is to request the data on the FTP servers and pull a section of them. The idea is that the majority of the data won't be stored locally. It will be downloaded per usage. The validity of this data ingestion plan will also be tested through the VCF file of the 22nd chromosome and by a sample FASTQ and BAM file.

3. Methodology

3.1 Filtering VCFs [5] [6]

SNPs in VCF files can be filtered according to a bunch of parameters available in PLINK2 (a genomic analysis toolkit). For example, by allele frequency, using "—maf". Other possible parameters are summarized in this table:

Feature	As summary statistic	As inclusion criteria
Missingness per individual	`--missing`	`--mind` *N*
Missingness per marker	`--missing`	`--geno` *N*
Allele frequency	`--freq`	`--maf` *N*
Hardy-Weinberg equilibrium	`--hardy`	`--hwe` *N*
Mendel error rates	`--mendel`	`--me` *N M*

In this project, the allele frequency threshold is set to be 0.05 which is the most common threshold of the significant level. This means the variants with minor allele frequency below 5% will be filtered out. Also, the missingness per marker is set to be 0.1 which means that only SNPs with a 90% genotyping rate are included and the rest 10% is excluded. Individuals with too much missing genotype data are filtered out too by setting the missingness per individual to be 0.05. So only 95% of individuals are kept after this filtering. After applying those filters on each of the 22 chromosomes, SNPs and individuals with high missingness, or minor frequency rate are dropped. Then the VCF files will be re-coded back to VCF format for merging.

3.2 Merging VCFs

To merge multiple VCF across chromosomes into one single VCF, the first step is to convert each VCF to binary format and then merge those files into a single VCF. This process could be done by using bcftools (a genomic analysis toolkit). The reason to use bcftools is that it is very fast especially dealing with large data sets. After filtering SNPs with low quality, the SNPs that are left could be merged quickly.

3.3 Principal Component Analysis [5]

After merging the VCF of 22 chromosomes, the next step is to get a certain number of SNPs that contain the most information among the "1000 Genomes" samples. Since VCF files are the binary results of aligned DNA sequences compared against the reference genome, the SNPs that have the greatest variance will be the ones that will contain the most information. In another word, the SNPs that vary the most can reveal the information we want to study. To get more informed results from PCA, linkage pruning that removes the dependence among SNPs is required since PCA will assume that the SNPs are independent of each other. Using plink tools, running PCA with a specified number of components to keep will generates two CSV files named eigenvectors and eigenvalues.

Three arguments are required to conduct PCA on the merged VCF. In this project, the first argument of linkage pruning was set to have a window size of 50,000 bases. This means that a sliding window that contains SNPs of 50,000 bases will be used to conduct the pruning. The window step, the second argument, is 10, meaning that when calculating linkage 10 base pairs will be used at each time. The last argument is the threshold of the linkage which is set to be 0.1. So, any variables with more than a 0.1 linkage threshold will be pruned.

3.4 Visualization

After we have the pca.eigenval and pca.eigenvec we will notice that some eigenvalues are greater than the others. This means that the corresponded eigenvectors have a greater total amount of variance. Select appropriate numbers of PC to use. Since we need to plot the data on a 2-D map, where I'll choose 2 PCs that have the largest variance.

3.5 Outliers Detection

Some samples may contribute to an extremely large part of the variance or more precisely, they are outside the plus and minus 2 standard deviation away from the means (outside 95% of the data points). The resultant PCA from step 4 will have 20 eigenvalues, but usually, we would only need 10 of them by default. By calculating those 10 means and standard deviations for each PC, we could filter out the outliers. Another way to discard the outliers is by using more advanced clustering techniques like minimum cluster membership that will automatically keep only the significant SNPs within each cluster. This will be discussed more later.

4. Analysis

4.1 Explorative Data Analysis

To test the validity of the proposed pipeline, the VCF of the 22nd chromosome was used. VCFs are filtered by the threshold specified in section 3. PCA was conducted. Only 2 principal components are kept and plotted as the x and y-axis. Two distinctive clusters on the chart above could be identified using clustering technologies. Meanwhile, there are some outliers outside the two clusters that may need further treatment.

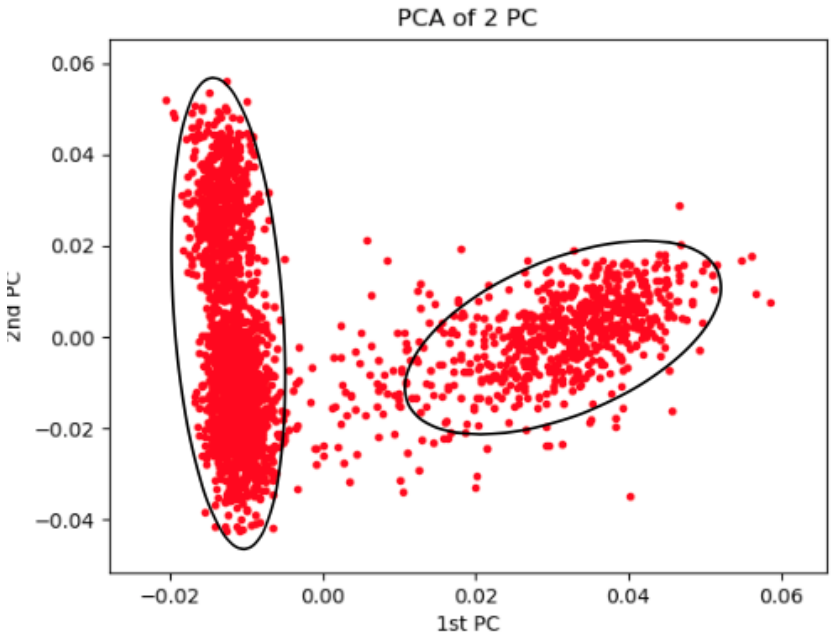

Figure 1 two clusters indicated by plotting 2 PCs

The validity of the pipeline is proven, so the next step is to generalize such results on the whole "1000 Genome" Data. Firstly, SNPs and samples with high missingness or minor frequency rate are dropped using plink tools (see more in Section 3.1). Then, Chromosome 1 to Chromosome 22 will be merged (see more in Section 3.2). The resultant merged VCF will take more than 65GBs disk space. After that, the linkage disequilibrium sites were removed, and the 2 largest principal components are kept and plotted as the x and y-axis. The non-clustered results are visualized below.

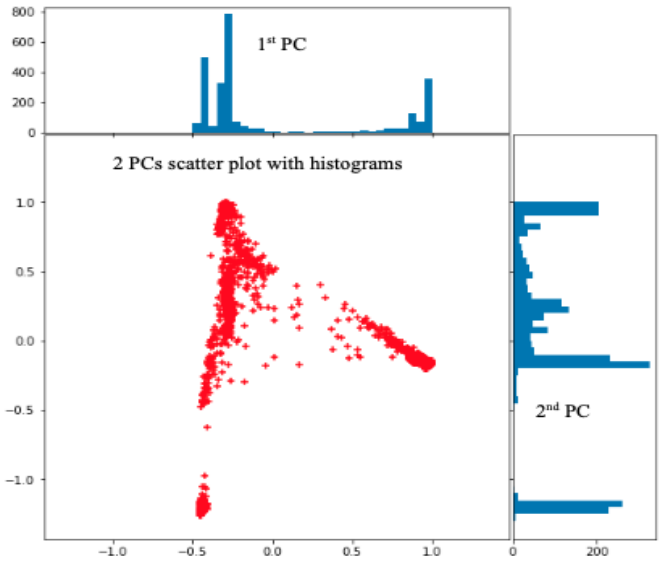

Figure 2 1000 Genome Data histograms and scatter plot

To better visualize how each principal component varies, two histograms are plotted on the x and y-axis to show how many individuals are within each specified bin. Note that for the histogram, each principal component is scaled by its maximum value and the size of the bins is set to be 0.05. Those treatments could ensure the plotted histograms will retain only necessary information. In the histogram on the x-axis, 2 distinctive groups could be identified—they are distinctive from each other as they clustered at each end of the univariate plots. On the other hand, the histogram on the y-axis reveals potentially 3 to 4 distinctive groups. Among those uncertain groups, the lower-left cluster in the scatter plots are discovered to be the most separate as its first and second principal component are the most isolated from the rest. Those plots also reveal that outlier detections might not be appropriate in this case since the distribution for each principal component is not normally distributed. Simply dropping the extreme values will not generate reliable results. The explorative analysis in this section leads me to explore more using clustering tools.

4.2 Clustering

Section 4.1 provides some insights on how to cluster the data points in the scatter plot. For example, the number of clusters I am interested in will set to be 4. This is out of two concerns. Firstly, the histogram on the y-axis reveals potentially 4 distinct clusters. Also, the "1000 Genomes" Data was collected across 5 regions: African, American, East Asian, South Asian, and European. In those regions, it is known that American ancestry is closely related to European ancestry. So, the number of clusters will be 4: African, European, Asian, and others.

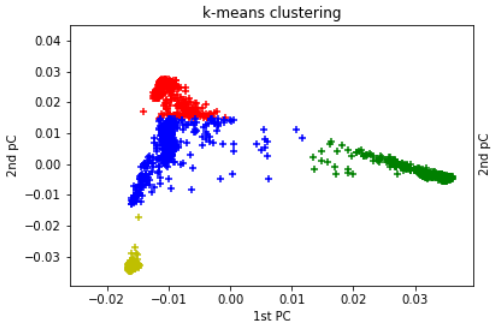

Figure 5 k-mean clustering

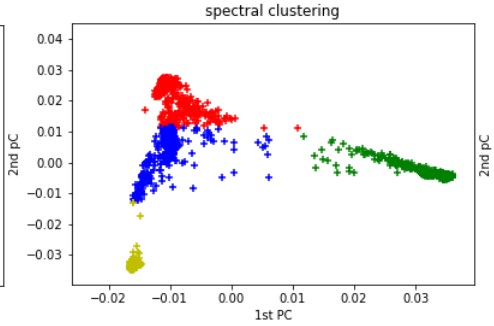

Figure 3 spectral clustering

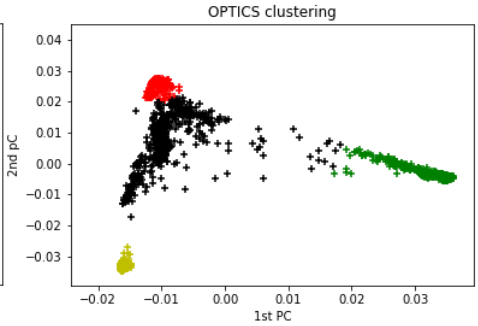

Figure 4 OPTICS clustering

The clustered results are shown above. By setting the number of clusters to be 4, K-means and spectral clustering give similar results. OPTICS clustering shows the best results among the three. It adopts the logic of minimum cluster membership and works well in uneven cluster sizes and density. As shown in Figure 5, the red cluster, green cluster and yellow cluster are distinct from one another. Interestingly, Figure 5 indicates that the yellow cluster seems more isolated from the red and green cluster as fewer dots are between the yellow cluster and either the red or green cluster.

4.3 Future Research

This project can shed light on intriguing questions like identifying an individual's genetic ancestry given his or her gene information. Training on the "1000 Genomes" Data, the clustering model can predict one's genetic ancestry given one's genetic information by testing the closest members of the given DNA data. In a broader sense, using more advanced whole genome sequencing technology, we could even assess if a region of DNA is associated with a certain disease and if the disease is associated with the individual's genetic ancestry and so on.

5. Conclusion

The world map (Figure 6) from the "1000 Genomes" Data Portal visualizes how samples from "1000 Genomes" are associated with their geometrical locations. Since the samples themselves are not related with any demographic information, we can only take some educational guesses on matching the map with the clustered classes. The red triangle on the world map indicates a similar pattern as the clustered principal components plot, each angle represents a major genetic ancestry source from Asia, Europe, or Africa. Based on the results from OPTICS clustering, I would consider the yellow cluster, the group of people with Asian ancestries; the red group, people with European ancestries; and the green group, people with African ancestries. The black cluster might be populations of mixed ancestry. Firstly, the Asian ancestry remains more isolated, so the yellow cluster might represent samples of Asian ancestry. Since most of the dots are clustered within or around the red group and most of the data are collected from

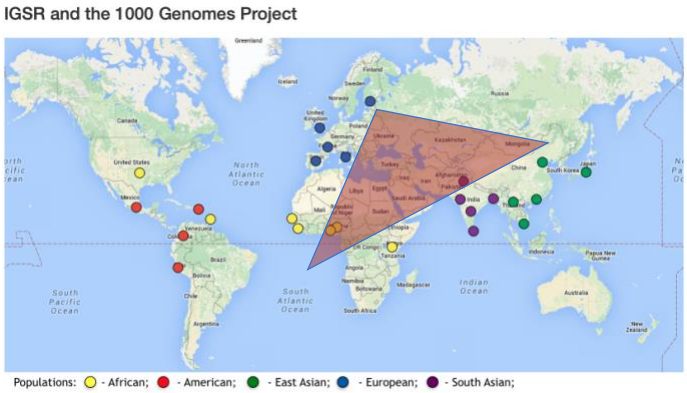

Figure 6 world map with major genetic ancestry sites [4]

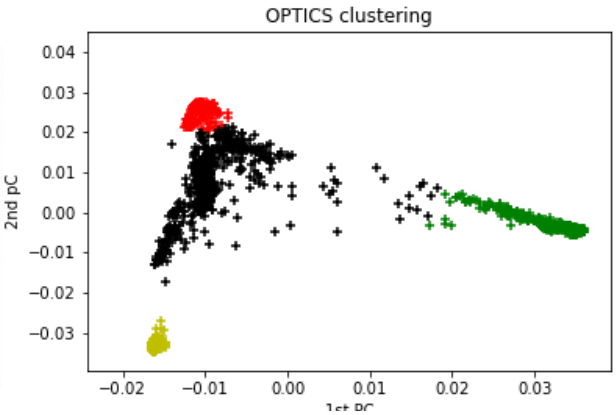

Figure 7 final output with three major genetic ancestry sites

Europe, it is fair to extrapolate that the red cluster might represent samples of European ancestry. The green cluster might represent samples of African ancestry since the samples collect are relatively small and they are not that isolated.

Cluster	# samples in the cluster
red	537
green	657
yellow	515
black	839

6. Evaluation

There are some potential shortcomings of the data from “1000 Genomes”. For capturing more diverse types of variants, the sample size needs to be larger and more representative, considering the scope and size of human population distributions. Samples draw only represent countable regions while ignoring the others.

Second, each VCF file stored on the FTP server is too large to be pulled. If the developing scripts could be stored online like on DSMLP, the data could then be pulled much quicker. During the later stage of developing this project, I have moved my project onto DSMLP where all the data available in a parent folder. This saves me a lot of time than downloading the data of GBs.

6. Reference

[1] A global reference for human genetic variation https://www.nature.com/articles/nature15393
[2] Genes mirror geography within Europe https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/
[3] DNA Sequencing Technologies Key to the Human Genome Project
https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/
[4] Figure 6: 1000 Genomes https://www.internationalgenome.org
[5] Population structure: PCA, https://speciationgenomics.github.io/pca/
[6] Whole-genome association analysis toolset http://zzz.bwh.harvard.edu/plink/dataman.shtml#mergelist