Elżbieta Karpińska (index no: 384283)

Project MCSB

Abstract

The main goal of this project was to find correspondence between genetic and geographic distances and to better understand population structure in the world. To do so, genetic variation of 1092 genetic samples from IGSR[1] (The International Genome Sample Resource) from all around the world was characterized. IGSR is one of the largest publicly available dataset of human variation and genotype data made possible by collecting whole-genome sequencing from healthy individuals from 26 populations. In this study, the chromosome 21 has been chosen for analysis due to the size of the data. This project was inspired by paper[2] analysing genetic variation within Europe corresponding to Population Reference Sample (POPRES) project[3]. In our case, PCA and sPCA gave similar results and UMAP and t-SNE also gave similar results. The achieved results show some interesting correlations between populations and superpopulations.

Introduction

In this project dimensionality reduction techniques for Data Visualization such as: PCA, sPCA (sparse PCA), UMAP (Uniform Manifold Approximation and Projection) and t-SNE (T-distributed stochastic neighbour embedding) were compared. The main goal was to check how these low-dimentional graphs preserved the high-dimentional clusters and relationships between them. Genetic structure information is used in fields such as medicine or anthropology. It helps to better understand human population structure and fluctuations between the continents. It may answer some critical questions such as how genetic clusters reflect the population distribution in the world and how distinct the populations are.

Data characterization

Data has been download from: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/?fbclid=IwAR2knQcec8Z0YSukuoshOJSGrR9erxugcj3v6pMvPPtd3gwJz85ID3zIQrM

Table 1. List of the used representants of each population and superpopulation from IGSR from chromosome 21 (data comes from attached file igsr_populations.tsv)

Superpopulation name	**Superpopulation code**	**Population code**	**Population name**	**Number of samples**
European Ancestry	EUR	GBR	British in England and Scotland	88
European Ancestry	EUR	FIN	Finnish in Finland	93
East Asian Ancestry	EAS	CHS	Han Chinese South	100

American Ancestry	AMR	PUR	Puerto Rican in Puerto Rico	56
American Ancestry	AMR	CLM	Colombian in Medellin, Colombia	62
European Ancestry	EUR	IBS	Iberian populations in Spain	14
European Ancestry	EUR	CEU	Utah residents (CEPH) with Northern and Western European ancestry	85
African Ancestry	AFR	YRI	Yoruba in Ibadan, Nigeria	40
East Asian Ancestry	EAS	CHB	Han Chinese in Beijing, China	97
East Asian Ancestry	EAS	JPT	Japanese in Tokyo, Japan	137
African Ancestry	AFR	LWK	Luhya in Webuye, Kenya	97
African Ancestry	AFR	ASW	African Ancestry in Southwest US	61
American Ancestry	AMR	MXL	Mexican Ancestry in Los Angeles, California	65
European Ancestry	EUR	TSI	Toscani in Italy	97

Methodology

PCA[4] is an unsupervised method widely used for dimensionality reduction. It factorizes the matrix characterizing the data. Rows of the matrix represent the data points and columns the features of the data. PCA finds the most important direction in the data and keeps the direction of the biggest spread of the data to get the desired number of dimensions of the space. The PC1 (Principal Component) axis explains the maximum amount of variance in the

original dataset, and PC2 (Second Principal Component) axis is orthogonal to PC1 axis. The axes are hierarchical, so PCn is showing the bigger variance in the data than PC(n-1). PCs are a linear combination of variables.

The sparse PCA[5] method is a modification of classical PCA that retains consistency even if the number of variables is much larger than the number of samples. The method finds linear combinations of small amounts of input variables by computing the least Squares Estimates of sparse PCs, which reduces the data complexity. The R package maximizes the variance of the explained data, which is reflected in the scree plot.

Contrarily to PCA, axes in the low-dimensional t-SNE[6] graph are not hierarchical. The t-SNE cost function wouldn't change if rotating the points. One should look at the the distances between the points, as neighbouring points from the original high-dimensional space will be neighbours in the low-dimensional space. One should bear in mind though that larger distances in the low-dimensional space might be irrelevant to the ones in high-dimensional space (unlike in PCA). One should pay attention to the perplexity as especially low values can cause t-SNE breaking data into pieces and arbitrarily separating them. The rule is that low values of perplexity is supposed to preserve local structure better, and high values of perplexity are about to keep global structure. Furthermore, when one changes the perplexity settings, the results change completely, unlike in UMAP.

UMAP[7] (Uniform Manifold Approximation and Projection for Dimension Reduction) and t-SNE build the neighbour graph in the original space of the data and try to find the similar graph in lower dimensions.

UMAP and t-SNE essentially work the same as most of the differences are very minor. I would highlight the main differences though. The t-SNE always starts with a random initialization of the low-dimensional graph, which means that one starts every time with a different low-dimensional graph of the same dataset. On the other hand, UMAP uses “Spectral Embedding” to initialize the low-dimensional graph, so one always starts with the same low-dimensional graph, using the same dataset. T-SNE moves each point a little bit each iteration, whereas UMAP moves just one point (or small subset of points) each time. The advantage of UMAP in this case when one works with a big dataset, is better scalability.

The result of the UMAP method depends on the value of *k* (number of neighbours). The choice of the value *k* determines the length of the rays around each data point, which affects if a point is assign to a given cluster. If the rays are too big, then a very small number of big clusters, composed of large amounts of data will be created, so one cannot see the details. On the other hand, a low number of neighbours results in small, independent clusters, so one can see the details but not the bigger scale. So, the number of neighbours (parameter k) determines if the structure of the data is local or global.

The other parameter is minimum distance, which specifies how tightly the algorithm will map points into the target low-dimensional space. Simply, the higher the distance, the more spread the points are.

The change of both parameters will change the results (graphs) gradually. As in t-SNE, the axes can be rotated as the results are rotationally symmetric, so projections can swap the sides of the clusters.

The entire data analysis was conducted using R studio and the results are presented below. All the plots were made using *ggplot* package. Data were pre-processed using the code

preprocessing_data.R. The number of dimensions of the matrix has been reduced (left samples and snips but no number of chromosome) by counting the number of the non-zero alleles (number of alt alleles per person).

Results

Fig 1. PCA plot of populations

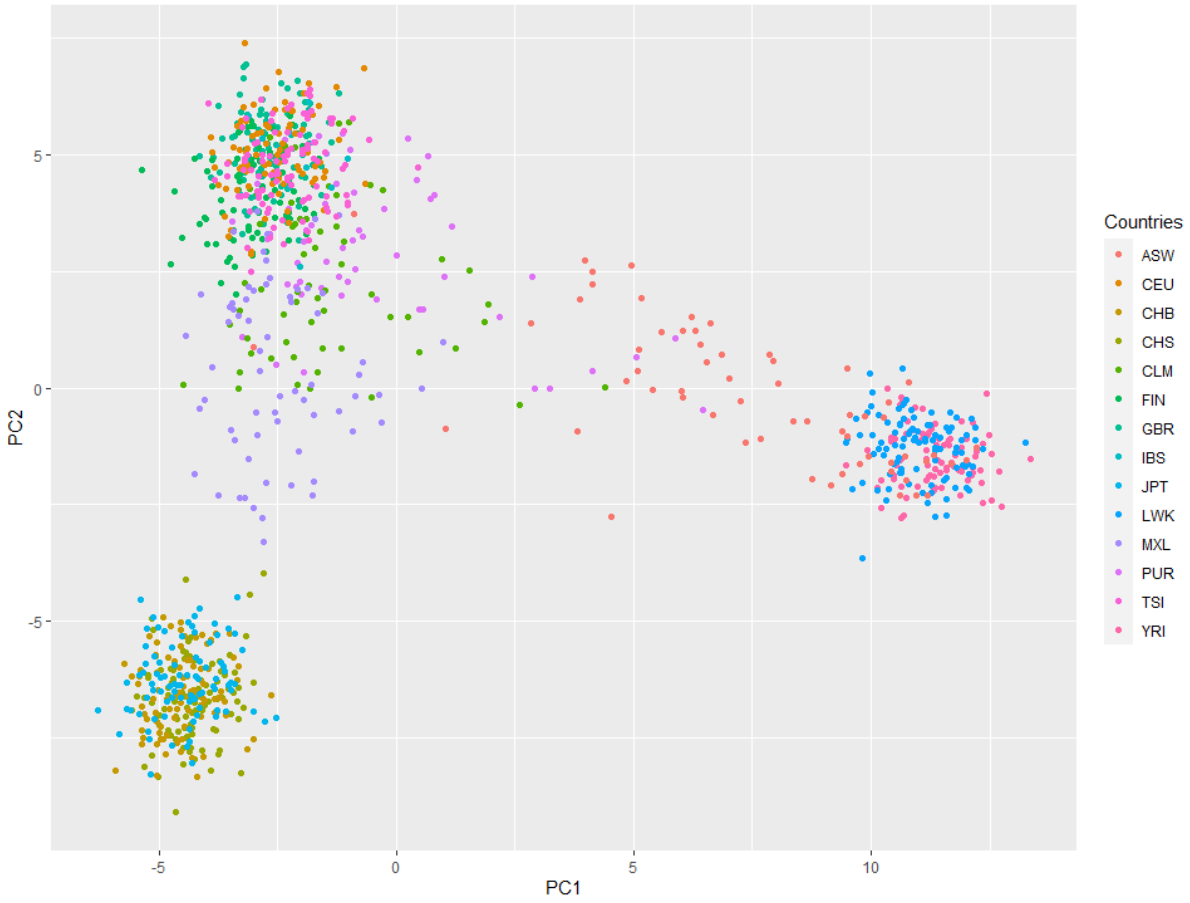


Fig 2. PCA plot of superpopulations

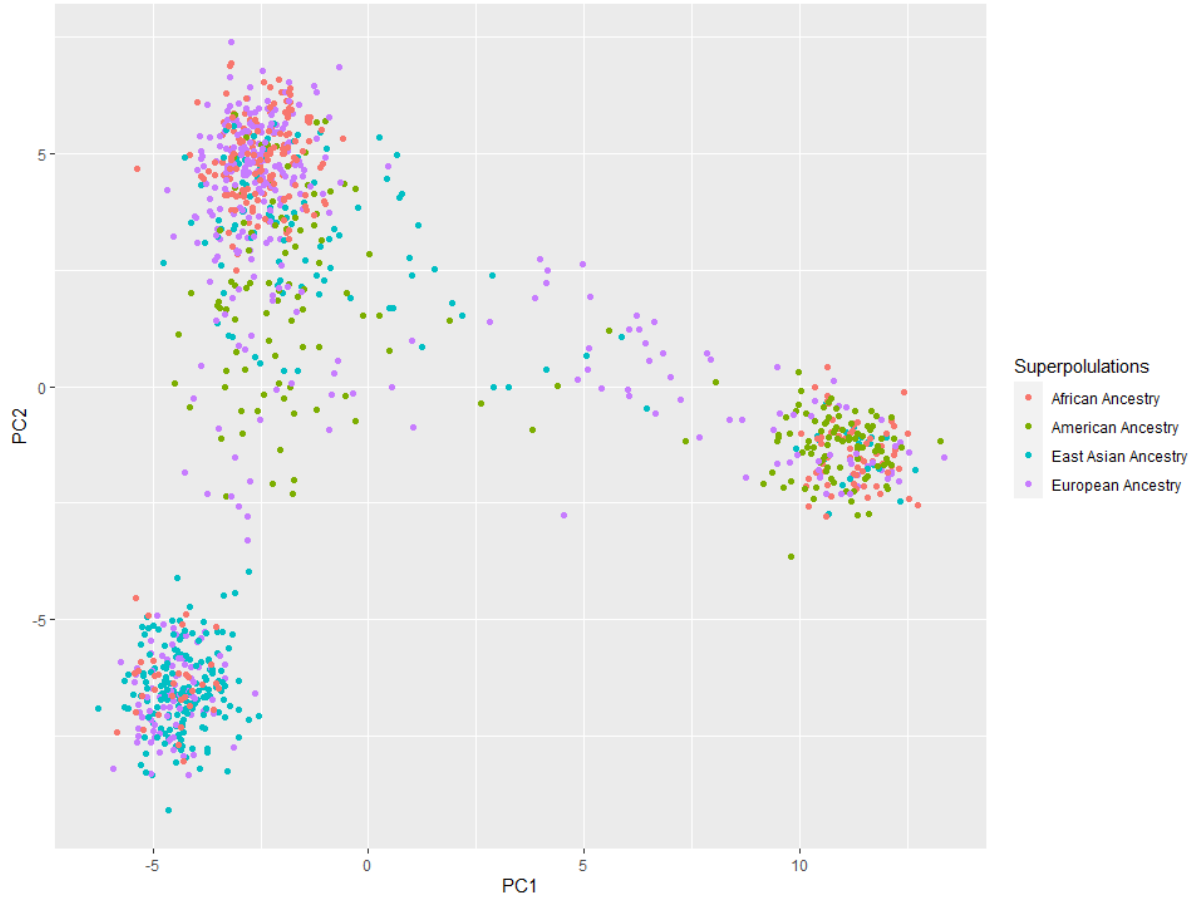


Fig 3. Scree plot of PCA

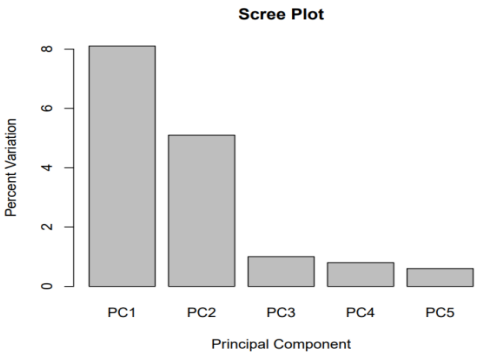


Fig 4. K-means plot of PCA with k=3

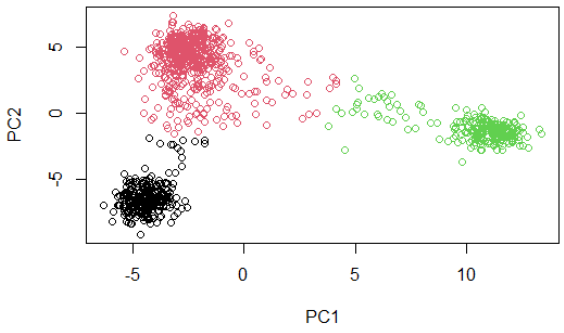


Fig 5. sPCA plot of populations

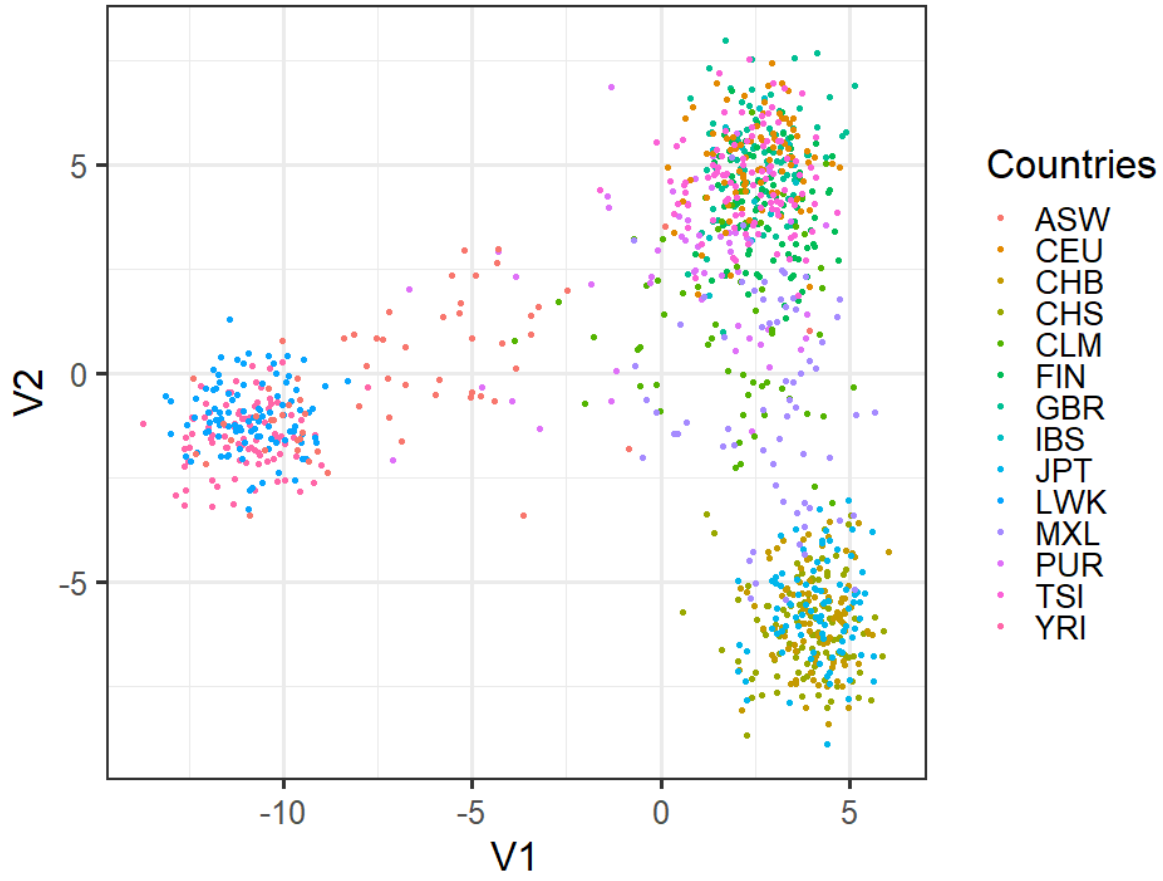


Fig 6. sPCA plot of superpopulations

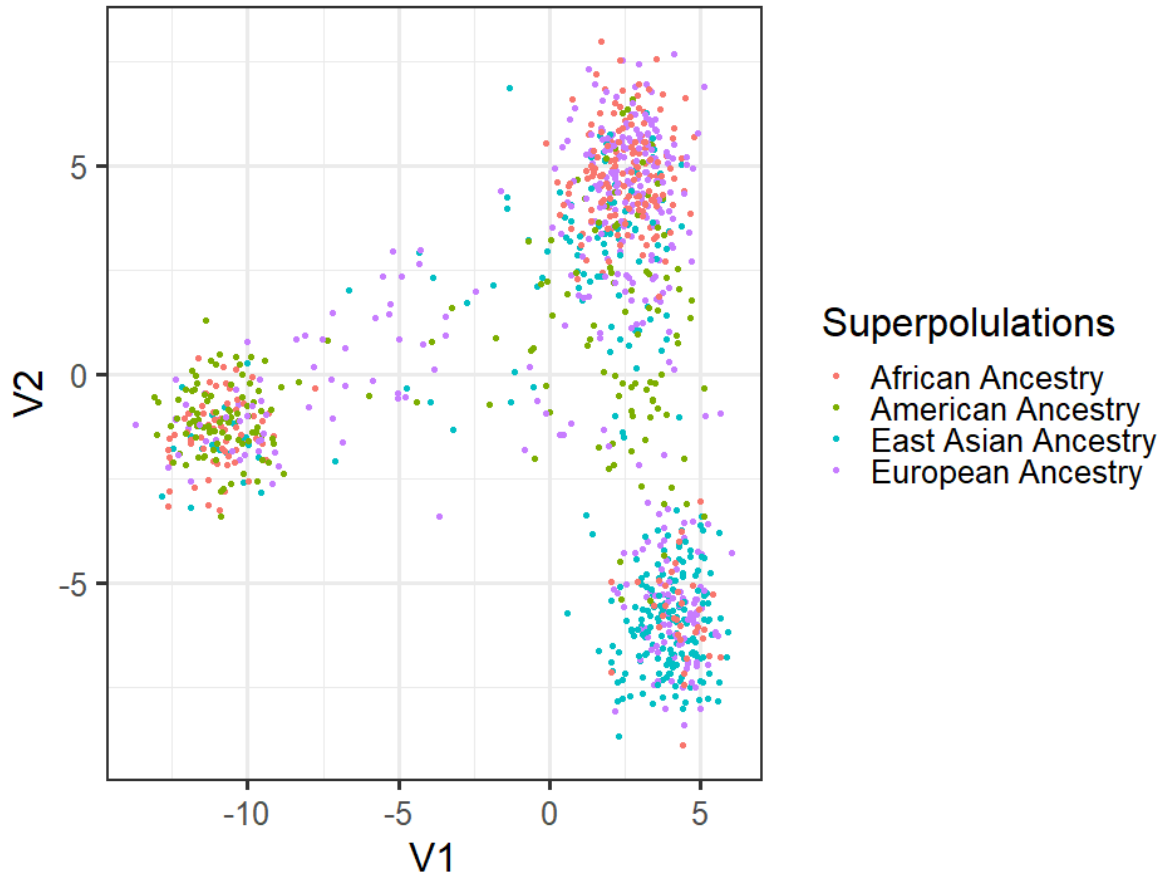


Fig 7. Scree plot of sPCA

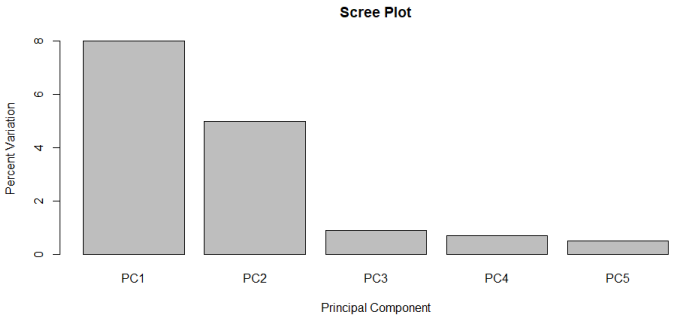


Fig 8. t-SNE plot of populations (default parameters)

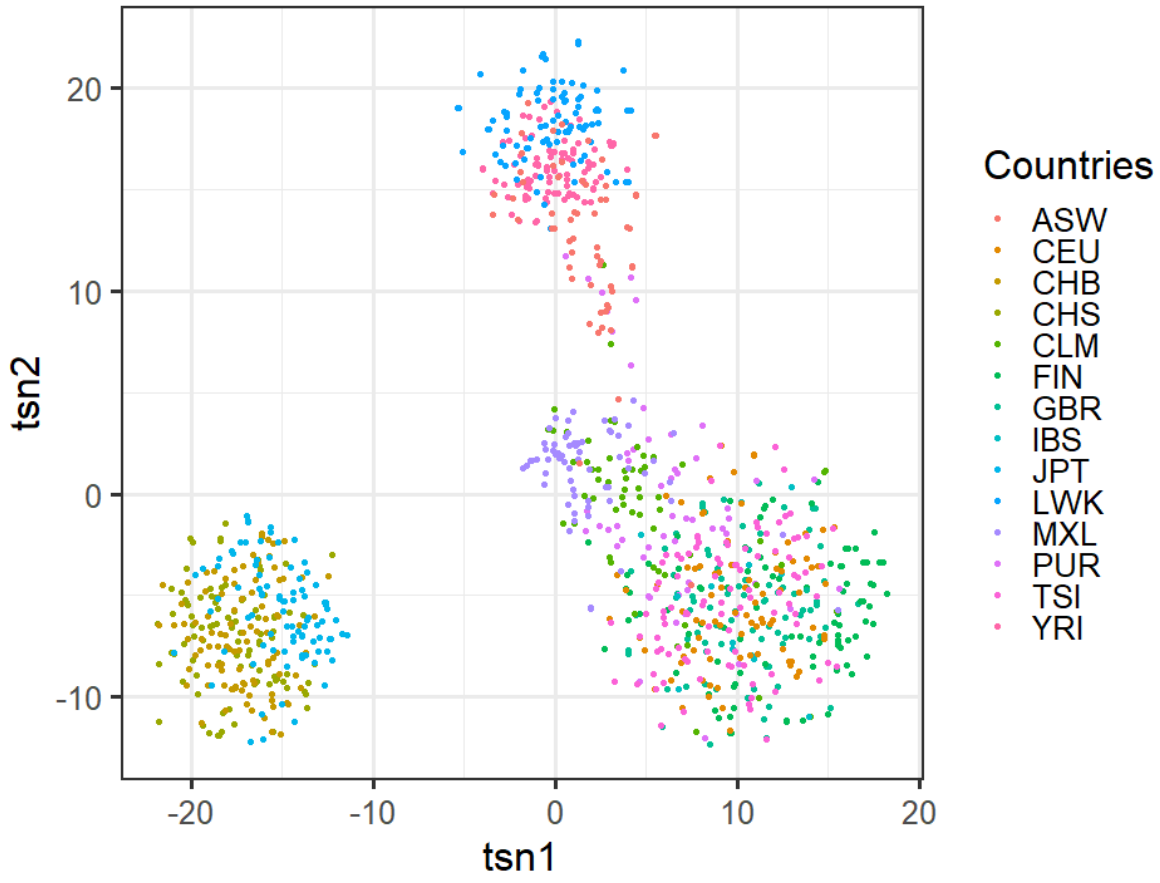


Fig 9. t-SNE plot of superpopulations

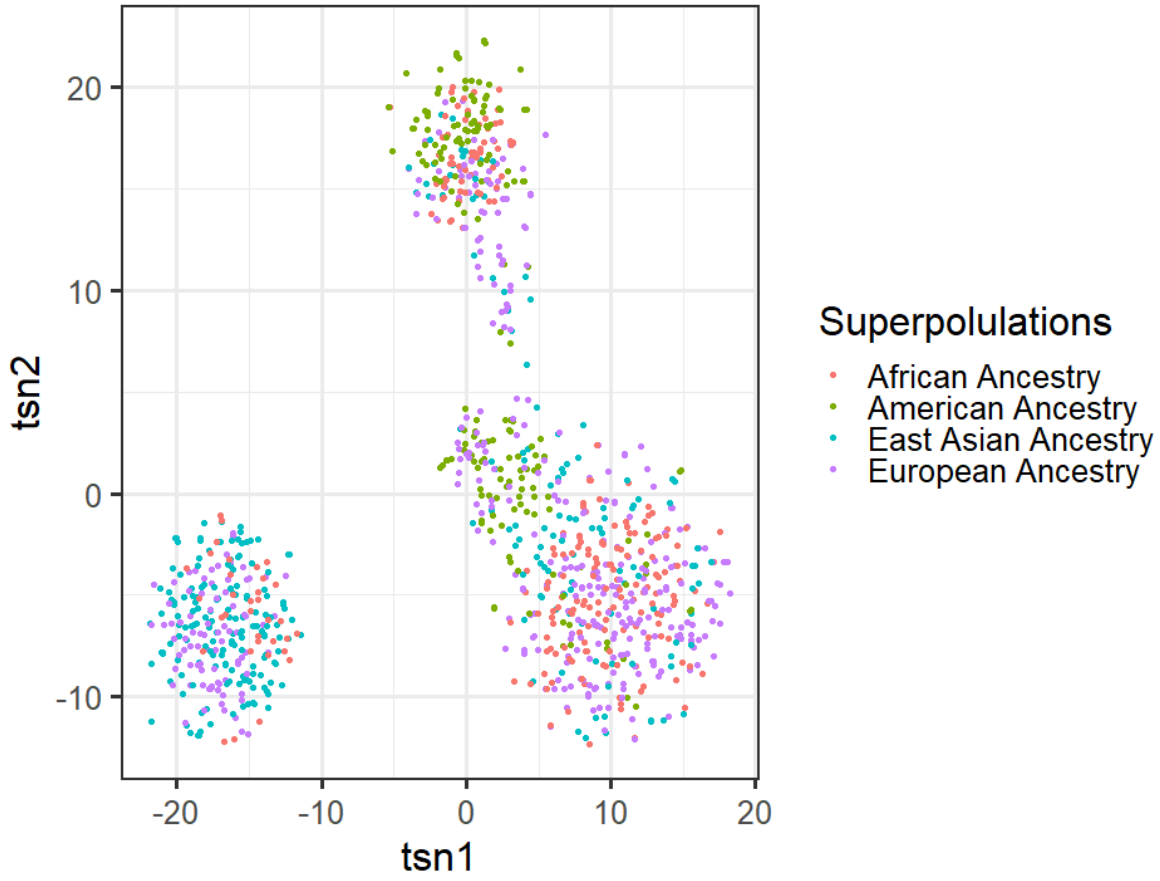


Fig 10. K-means plot of t-SNE for k=3

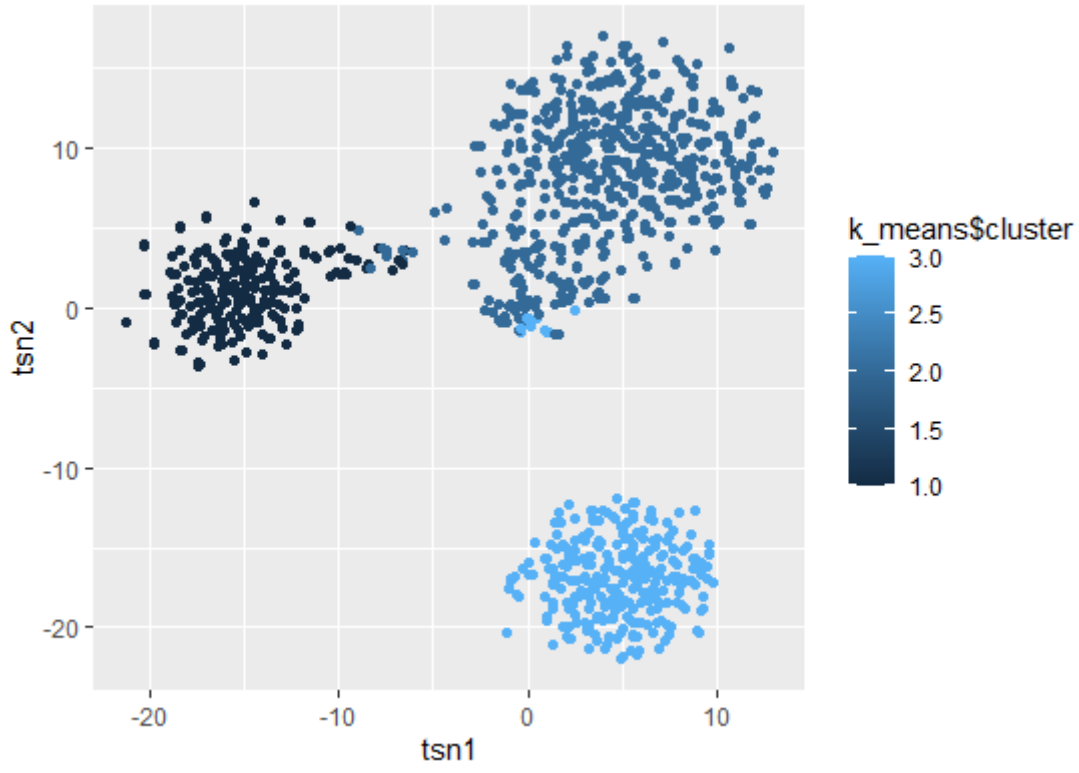


Fig 11. UMAP plot of populations (used parameters: k=15 min_dist=0.5)

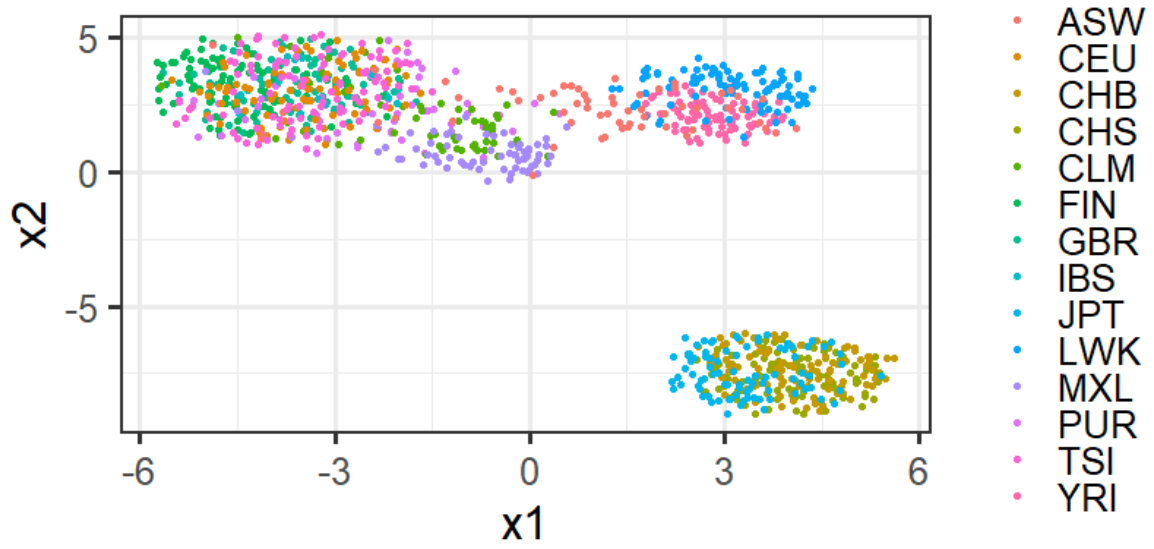


Fig 12. UMAP plot of populations (used parameters: k=15 min_dist=0.1)

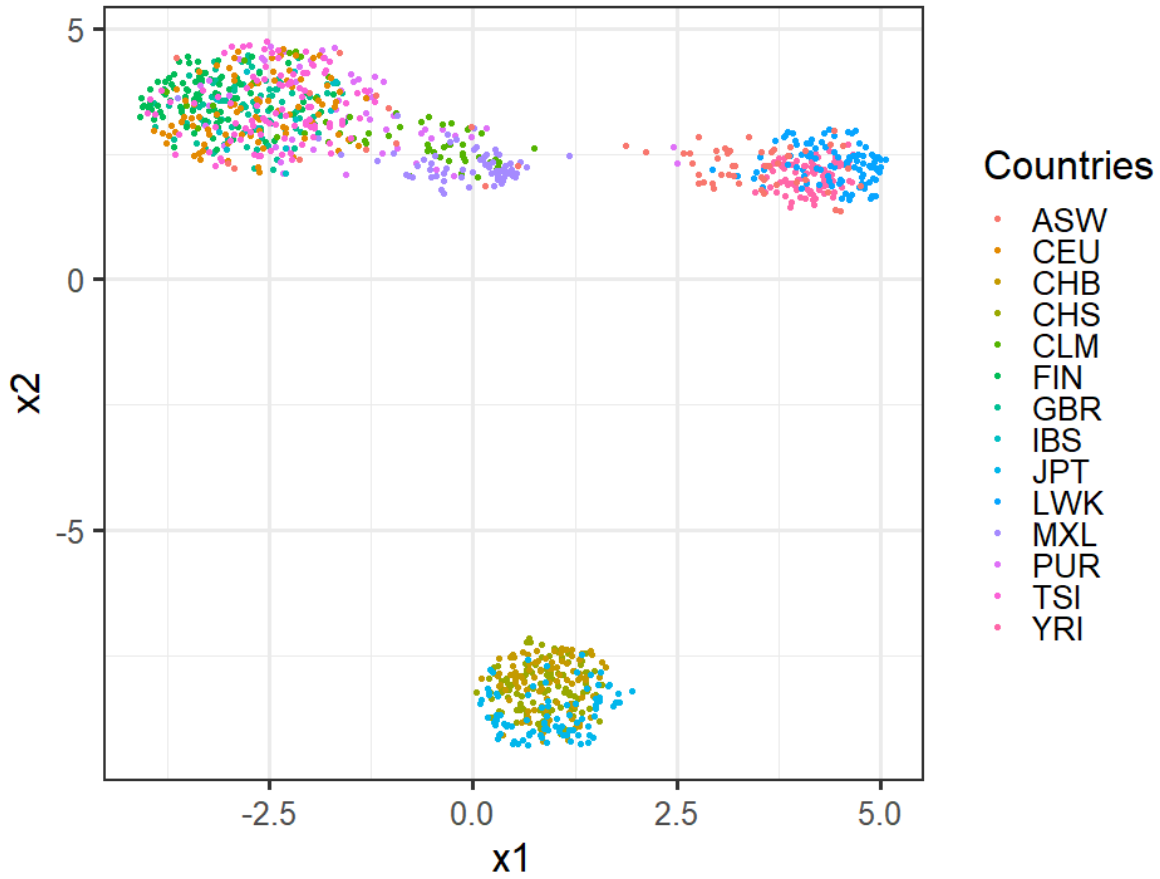


Fig 13. UMAP plot of superpopulations

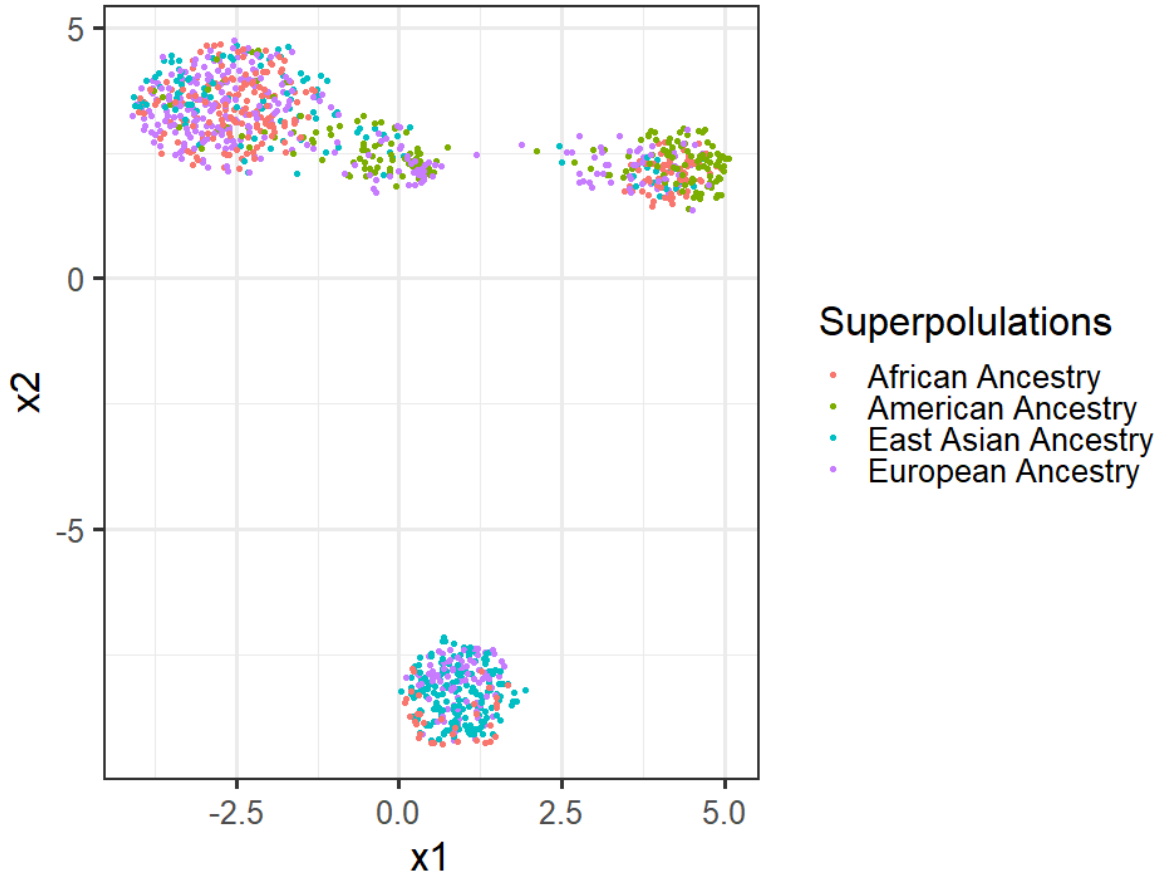
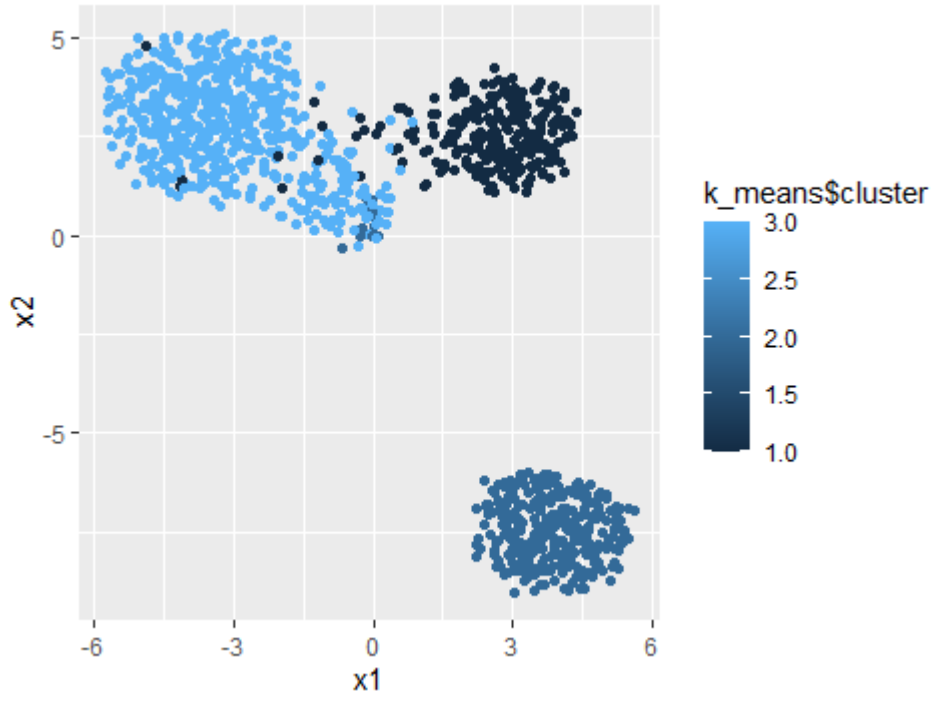


Fig 14. K-means plot of UMAP for k=3



The number of PCs is either the number of variables or number of samples, depending which number is smaller. In our case the amount of samples is higher than the number of variables, as PCA and sPCA graphs are the same. And the amount of snip is higher than the amount of variables, which is reflected in PCA graphs.

In the PCA (Fig 1.) and sPCA (Fig 5.) graph one can notice that Chinese (CHS and CHB) and Japanese (JPT) cluster together. Chinese are connected by Mexicans, mixing with Colombian, then Finish, British, IBS CLM and then CEU, ASW, TSI, PUR. The last cluster of Africans (LWK, YRI, ASW) is connected to them by mostly ASW, and less Colombian and Puerto Rican. The differences between PCA and sPCA are minor, it is the spread of CHS, CLM and MXL around the Japanese and Chinese cluster. The distances between clusters are kept unchanged. The scree plots of PCA and sPCA are very similar, in both cases PC1 and PC2 explain most variability in the original data .

T-SNE (Fig 8.) and UMAP (Fig 12.) graphs are very similar, so I would analyse them together. One can see that Chinese (CHS and CHB) and Japanese (JPT) cluster together, the same as in case of PCA graph, but the main difference here is the distance and not many groups in between the biggest cluster made up from Americans and Europeans. Africans cluster together (LWK, YRI, ASW). The last, biggest cluster is connected with African one by PUR and ASW, later we can notice LWK, CLM as well as ASW, PUR and MXL, then rest all mixed up are CEU,TSI, YRI,CLM IBS, TSI, GBR, FIN.

Conducting UMAP default value of neighbours (parameter k=15) was maintained as changing the value wasn't improving the clarity of data clusters.

When it comes to graphs showing the relationships between ancestries (superpopulations) UMAP (Fig 13.) and t-SNE (Fig 9.), and also PCA (Fig 2.) and sPCA (Fig 6.) are showing very similar results. All the graphs make the same clusters but on PCA methods one can see more fluctuations between the clusters made of European, East Asian and American Ancestries. In case of t-SNE and UMAP clusters are clearly separated with barely any fluctuations in between them. One cluster is made of East Asian, European and African Ancestries. The other cluster shows that American and African Ancestry is close together, then come the Europeans with some East Asian Ancestry fluctuating in between. The last, biggest cluster shows European mixed with Americans, then East Asian are close to African and European, and a few American in between.

Plots of k-means cover the data distribution into clusters. In UMAP plot (Fig 14.) one can notice some fluctuations in top clusters.

Conclusions

To explain why the data obtained with t-SNE is more separated from each other between clusters than in PCA method, we have to remind ourselves how t-SNE works. The t-SNE method is based on calculating the distance between data points, the distance between two points belonging to two different clusters varies depending on which cluster the measurement begins in, so t-SNE calculates the average of the two distances. If a point would be placed in a cluster depends on matching the measured distances to the t distribution.

When it comes to PCA and sPCA, just PC1 and PC2 were used in data analysis as they cover the most percent variation as shown in scree plots (Fig 3. and Fig 7.).

Personally, I find all the results interesting and reflecting prior assumptions based on literature knowledge[8,9]. The UMAP method is pleasant to work with as it is easy to achieve balance between the local and global structure of the data and the graph is changing gradually as one changes the parameters. UMAP and t-SNE gave similar results but UMAP is faster due to higher processing speed and keeps better balance between locality and globality in clustering. sPCA shows very similar results as classical PCA which might be a result of the fact that the amount of variables is much lower than the number of samples.

Extending this project with new methods, pre-processing or analysis tools might reveal some new correlations between populations and superpopulations, so more in-depth analysis is needed as in this study just four Data Visualisation techniques has been used. Another approach would be similar to one used by authors of paper[2], so taking more samples which are closer distributed geographically (more dense) might improve results. Also collecting information about previous generations, so correlating genetic and geographic origins would be valuable. One has to remember that the average level of differentiation across Europe is smaller than around the world so the analysis would be more challenging.

References:

1. Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, Paul Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D941–D947.
2. *Novembre, J., Johnson, T., Bryc, K. et al. Genes mirror geography within Europe. Nature 456, 98–101 (2008)*.
3. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. 2008 Sep;83(3):347-58.
4. Jolliffe Ian T. and Cadima Jorge 2016Principal component analysis: a review and recent developmentsPhil. Trans. R. Soc. A.3742015020220150202
5. Iain M Johnstone; Arthur Yu Lu (2009). "On Consistency and Sparsity for Principal Components Analysis in High Dimensions". Journal of the American Statistical Association. 104 (486): 682–693. doi:10.1198/jasa.2009.0121. PMC 2898454. PMID 20617121.
6. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9(Nov):2579-2605, 2008.
7. McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861.
8. Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, Paul Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D941–D947.
9. Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, Paul Flicek, The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000

Genomes Project data, *Nucleic Acids Research*, Volume 45, Issue D1, January 2017, Pages D854–D859.