1) What data was used for this analysis.

- The data comes from a survey of 3,192 European individual whose data was collected and genotyped as part of the Population Reference Sample (POPRES) project[i]. Each individual's data was then genotyped at a locus of 500,568 using a single nucleotide polymorphism (SNP) chip on the Affymetrix microarray 500K by Applied Biosystems.

2) What was the purpose/goal of the analysis

- **Understand Differentiation Status**: Many countries data had a wide range of data values just within their own borders showing transient populations of a nationality will stay within the country. This indicates that many people of French, German and Italian have highly correlated social structures to remain within their own border's vs emigrate elsewhere.
- **Segmentation of Nationality**: Can analysis based on phenotype DNA provide location-based insights to the dispersion of genes in Continental Europe and the UK?
- **Increased Origin Density**: Is there a correlation of geography to phenotypes with localized density? Are the origin densities correlated directly to local longitude or latitude? Does this also highlight a large distribution of the populated surveyed provided by larger or smaller values increasing the size of the standard distribution values. Does this suggest a tightly knitted population that is related and again stays close vs emigrating to other countries?

3) How was the data structured/framed?

- **Data Framing:** The data is framed around the country of reported origin of everyone's grandparents. Then they seek to provide quite clear axial direction of significant strength, which could suggest a data driven causal role of the geographic axis in the role of European history. The next logical question is how can the axis of local geography impact dispersion or differentiation of population across and entire country let alone an entire continent as large and as diverse as UK and Europe?

4) What was the intuition behind the method of analysis?

- **Method Intuition:** PCA was used because there was an alignment between the correlation of locality between the DNA haplotypes and their desperation within continental Europe. This means that people from Europe tend to stay in Europe vs moving elsewhere, this is a well-known fact that people call "Old Family". For example, my ancestors on my mother's side have been in America for over 380 years, my father's family was in the same region of Germany as far back as 1100 AD. The paper found direct evidence that "PC1 aligns north-northwest/south-southeast, and accounts for approximately twice the amount of variation of PC2" PCA produced a solid two-dimensional plot of the data that not only generated insights to the population DNA but also creates opportunity for further analysis, so the intuition was highly accurate to the outcome found. Multivariate regression gave us the insight that 90% of the population was found within 700 km of their origin. Because they used PCA along a 2-dimensional axis that matches the geographic layout of the country they found very clear alignment between DNA and its source consistently across many countries.

5) In what other use cases or applications can this be helpful.

- **Use Cases:** Elon Musk has talked about how the Mechanical Engineering world uses PCA for power distribution, improving engine efficiency. I have personally used a similar form in data visualization to find insights where data is incredibly dense in Gene expression clustering and in Elisa kit assay chemical dispersion accuracies. PCA is also used in optimal starting, optimal stopping as well as in microchips miniaturization but need to maintain power to weight ratio of chip vs silicon wafer size.

iNelson, M. R. et al. The Population Reference Sample (POPRES): a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet.. (in the press).