Journal, YYYY, 0–0
doi: 10.1093/*Journal*/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Analyzing Population Migration using the Geographic Population Structure Algorithm

Corresponding Shijie Niu[1*]

[1] Department of Biology, Box 118, 221 00, Lund University, Sweden.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The geographic population structure (GPS) is a good method to reflect the historical gene flow and the potential environmental change. It also could be used to identification of genetic ancestry and the recent genetic admixture is crucial for admixture mapping. Several mass population migrations have occurred throughout history, this research aims to investigate the relationship between migration (use the migration between China, Japan and Vietnam as an example) and the migration distance by GPS algorithm.

In this research, I divide the data into four groups by movement distance: the close group (less than 2000 km) median group (2000 km ~ 5000 km), the far group (5000 km ~ 10000 km) and the very far group (more than 10000 km).

Whether could use this method to trace the migration of the population in Asia and the proportion of different distances is an interesting question.

Based on the cost of movement and cultural similarity, the hypothesis of this research is more migration distance will mean less proportion of migration

Results: The population moving from China to Japan and Vietnam is significantly large, which is suitable to the historical records, and the proportion of each distance group is significantly different, and the trend of it also fits the hypothesis, except for the very far group, the reason of it need more research.

Availability: Analyzing Population Migration using the Geographic Population Structure Algorithm

Contact: sh2887ni-s@student.lu.se

Supplementary information: https://github.com/Sh2887ni-s/BINP29_RESIT.git

1 Introduction

The geographic population structure (GPS) is a good method to reflect the historical gene flow and the potential environmental change (J. C. Avise et.al. 1987). It also could be used to identification of genetic ancestry and the recent genetic admixture is crucial for admixture mapping (Oscar Lao, 2006). So that, it could show the historical migration of the human population in the world.

As we know, there are several human-being migrations in the world or say cultural influences. I will choose one of them that the Chinese culture influences East Asia and south-east Asia as an example. The migration is hypothesized to lie at the very origin of the Japanese population, and exchanges between Japan and China go back at least to the Han Dynasty (R. Achenbath, 2016), and at a similar historical period, there is migration from China to Southeast Asia because the part land of the part of north Vietnam belongs to China and Chinese culture still influence to Southeast Asia. Whether could use this method to trace the migration of the population in Asia and the proportion of different distances is an interesting question. Based on this background knowledge, I will verify whether the GPS could be used to trace the migration of the populations. And I also will show the proportion of different distances in these migrations in my data.

2 Material and methods

The data of this project are offered by Prof. Eran Elhaik, the data are saved in three files named 'gen.csv', 'geo.csv' and data.csv'. And the method of the project could be divided into two parts, one is the GPS algorithm, and another one is the select the record and calculating the distance.

2.1 GPS algorithm

The GPS algorithm is encoded by Python from the reference script which is written by R.

2.2 The migration direction and the distance

The migration and the distance calculation parts are written in another script to filter the record of which population and the predicted place are different, and then, based on the longitudes and latitudes of the location of the population and predict the location, calculating the distance of those locations

3 Results

3.1 The accuracy of the GPS algorithm

To test the accuracy of the script which I wrote by GPS algorithm, I do this test and the result is in Fig. 1

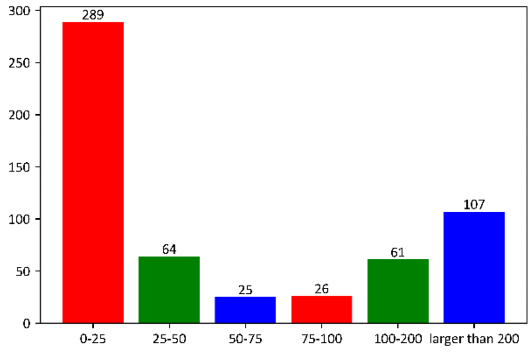

Fig.1 The distance of prediction latitude and longitude and the actual location with each population. The X-axis shows the distance range (km) and the Y-axis shows the count of the population in this range.

In this graph, The most distances between the prediction location and the actual location of the populations are within 25 km. there are 289 populations in this range which occupies around 50% of all of the population, and around 71% of the population is located within 100 km.

3.2 The model that is created by the distance matrix in the GPS algorithm

There is a linear model created between the two distance matrixes in the GPS algorithm. The model is: Y=67.79X+1.08, the X is from the gene information distance matrix, and Y is from the latitude and longitude distance matrix. The figure is plotted in Fig. 2

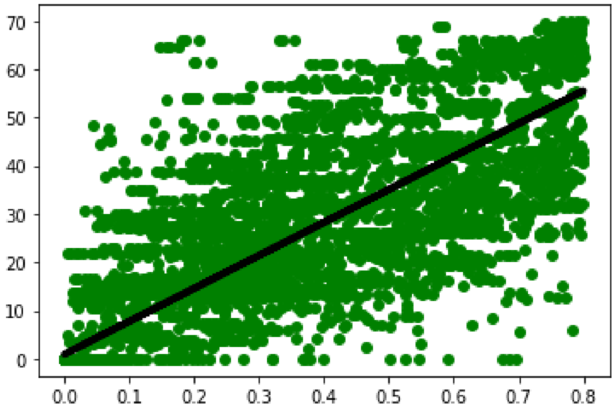

Fig.2 The plot of the model prediction, the green point is the real point in the two matrixes, and the black line is the predicted linear model.

3.3 The migration between China and Japan and Vietnam

Based on the data, I found there are 24 samples come from Japanese and Vietnamese in total, and 22 of these samples are close to the Chinese so the proportion of migration is around 91.97% (fig.3). This proportion is significantly higher than the samples which closer than themselves and the other of the world.

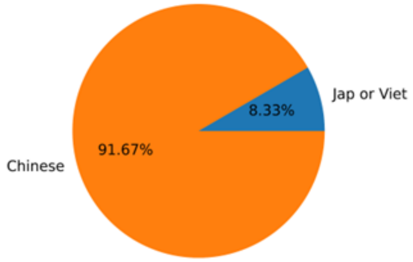

Fig3. The proportion of Vietnamese and Japanese samples is close to themselves and Chinese.

3.4 The different proportions in different migration distance group

Secondly, I divide the distance into four groups: the close group which is closer than 2000 kilometres (km) median group which is between 2000 km and 5000 km, the far group which is between 5000 km and 10000 km and the very far group which is longer than 10000 km.

In my research, there are 297 samples in the close group (67.5%), 91 samples in the middle group (20.68%), 16 samples in the far group (3.64%) and 35samples in the very far group (8.18%) (fig.4).

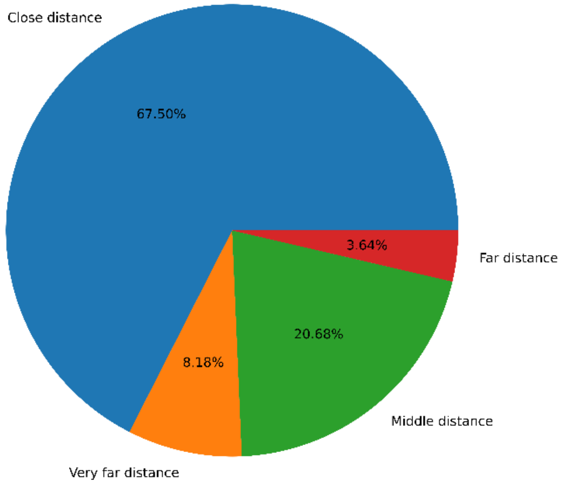

Fig.4 The proportion between each group.

Discussion

In the first part, we could find that the samples in Japanese and Vietnamese are significantly closer to the Chinese than themselves. That could indicate how deep the Chinese influence they are, and also could show the history of influence, because a higher percentage usually means more people migrating in history, especially since there is not a large migration happening in recent centuries. Because if it happens in recent several centuries, the gene cannot spread to so many people, and the gene will be closer to the normal gene in Japanese and Vietnamese. However, the data are not only collected in Asia, so that, more accurate migration information should also be researched after getting more data in the future

The second part could show the most distance the migration usually happens within 2000 km, the reason for that would be the cultural similarity and the migration cost. The migration cost is easy to understand, longer travel usually means more expenses. The culture similarity is considered because a similar culture usually could help them fit into the local society and easily learn the local language, for example, the Irish usually prefer to move to Sweden instead of China, even though the cost is the same. This research lacks data on culture and costs at different distances, so that, these parts need more research.

However, the very far group is almost triple that of the far group. That would be because of the colonial movement and refugee migration, but it still needs to research.

Acknowledgement

In this project, my teacher Eran Elhaik and teaching assistant Sara Behnamian gave me very useful help and suggestions.

And I also pretty much appreciate the code sharer about the GPS algorithm in R language: https://github.com/homologus/GPS.git, it is pretty useful and easy to understand for me.

References

Achenbach, Ruth(2016), Chinese Migration to Japan. Return Migration Decisions. Return Migration Decisions[M]. Springer 99-115

Avise, J. C. et.al. (1987) Geographic population structure and species differences in mitochondrial and DNA of mouthbrooding marine catfishes (Aiidae) and demersal spawning toadfishes (Batrachoididae). Evolution, 41(5), 991-1002.

Lao, O et. al(2006), Proportioning Whole-Genome Single-Nucleotide–Polymorphism Diversity for the Identification of Geographic Population Structure and Genetic Ancestry. The American Journal of Human Genetics, 78, 680-690.