Principal component analysis of gene frequencies of Chinese populations

XIAO Chunjie (肖春杰)[1], L. L. Cavalli-Sforza[2], E. Minch[2] & DU Ruofu (杜若甫)[1]

1. Institute of Genetics, Chinese Academy of Sciences, Beijing 100101, China;
2. Department of Genetics, Stanford University, CA 94305, USA
Correspondence should be addressed to Xiao Chunjie: Department of Biology, Yunnan University, Kunming 65091, China (email: cjxiao@public.km.yn.cn)

Received January 13, 2000

Abstract Principal components (PCs) were calculated based on gene frequencies of 130 alleles at 38 loci in Chinese populations, and geographic PC maps were constructed. The first PC map of the Han shows the genetic difference between Southern and Northern Mongoloids, while the second PC indicates the gene flow between Caucasoid and Mongoloids. The first PC map of the Chinese ethnic minorities is similar to that of the second PC map of the Han, while their second PC map is similar to the first PC map of the Han. When calculating PC with the gene frequency data from both the Han and ethnic minorities, the first and second PC maps most resemble those of the ethnic minorities alone. The third and fourth PC maps of Chinese populations may reflect historical events that allowed the expansion of the populations in the highly civilized regions. A clear-cut boundary between Southern and Northern Mongoloids in the synthetic map of the Chinese populations was observed in the zone of the Yangtze River. We suggest that the ancestors of Southern and Northern Mongoloids had already separated before reaching Asia. The ancestors of the Southern Mongoloids may result from the initial expansion from Africa or the Middle East, via the south coast of Asia, toward Southeast Asia, and ultimately South China. Upon reaching the Yangtze River, they might even have crossed the river to occupy the nearby regions for a period of time. The ancestors of the Northern Mongoloids probably expanded from Africa via the Northern Pamirs, first went eastward, then towards the south to reach the Yangtze River. The expansion of the Northern Mongoloids toward the south of the Yangtze River happened only in the last 2 or 3 thousand years.

Keywords: principal components, synthetic map, Chinese populations.

Genetic differences among populations generally consist only of variations of gene frequencies. The geographic distribution of gene frequencies among populations is especially useful for understanding the effects of migration, admixture, natural selection, genetic drift, geographic and social isolations, and mutation on human microevolution, as well as the origin and development of modern humans.

A great deal of gene frequency data on Chinese populations has been accumulated and can be used to analyze their geographic distributions. Because the distributions of alleles of different loci among populations change greatly, it is difficult to discover the trends and patterns common to many genes that are the outcome of events influencing their geographic distribution.

The principal components (PCs) analysis is of great help in solving the problem mentioned above. It was first developed by Hottelling in 1933, and first applied to anthropometrics by Rao

(1948) and to human gene frequencies by Cavalli-Sforza and Edwards (1964)[1].

The PC and synthetic maps of Chinese populations were constructed with the Genography program in the laboratory of L. L. Cavalli-Sforza at Stanford University, using the gene frequencies of 130 alleles at 38 loci.

1 Materials and methods

1.1 Materials

Thirty-eight loci used in the current study were the same as used for calculation of genetic distances among Chinese populations[2]. They were: ABO, MNSs, P, Rh, Kell, Kidd, Lewis, Lutheran, Diego, Duffy red cell blood group systems; HLA-A, -B, -C, -D; acid phosphatase, adenine deaminase, adenylate kinase, esterase D, glyoxalase, phosphoglucomutase 1, glutamate-pyruvate-transaminase, 6-phosphogluconate dehydrogenase, glucose-6-phosphate dehydrogenase, α-antitrypsin, components of complement C2, C3, C4, C6 and C7, properdin factor B, group-specific component, haptoglobin, immunoglobulin Gm and Km, transferrin, secretion type, PTC taste blindness, and earwax type. Among them, the MNSs and Rh systems were treated as single loci, and their haplotypes as alleles.

Each set (gene frequency data on 1 locus in 1 population are considered 1 set) was tested for Hardy-Weinberg equilibrium, and those with c^2 value showing $P<0.05$ were rejected. Gene frequencies from populations with the same geographic coordinates were averaged (weighting by sample size).

From the Han subpopulations living in 31 provinces, cities or autonomous regions, there are 704 sets and from 55 ethnic groups[3] 1190 sets of data which can be used for analysis. When we put the data of Han subpopulations and ethnic minorities together for analysis, we calculated the weighted averages of gene frequencies from Han and ethnic groups on the same longitude and altitude. Thus, we got 1606 sets of data for analysis.

The gene frequency data were collected from more than 600 papers. Therefore, it is impossible to list them here and they are omitted from the list of literature below.

1.2 The principle and procedure of PC analysis

They have been described previously by Cavalli-Sforza[1].

2 Results

2.1 Han populations

The geographic map of the first PC of the Han (fig. 1(a)), accounting for about one-third of the total variance, has a clear north to south gradient, indicating the genetic differences between Southern and Northern Mongoloids. It is worth noting that only two lines appear north of the Yangtze River, which demonstrates a relatively small genetic difference among Northern Han populations. Two reasons can easily explain this result. One of them is that the most Northern parts of China are plains, and allow less geographic isolation; and people can move from one

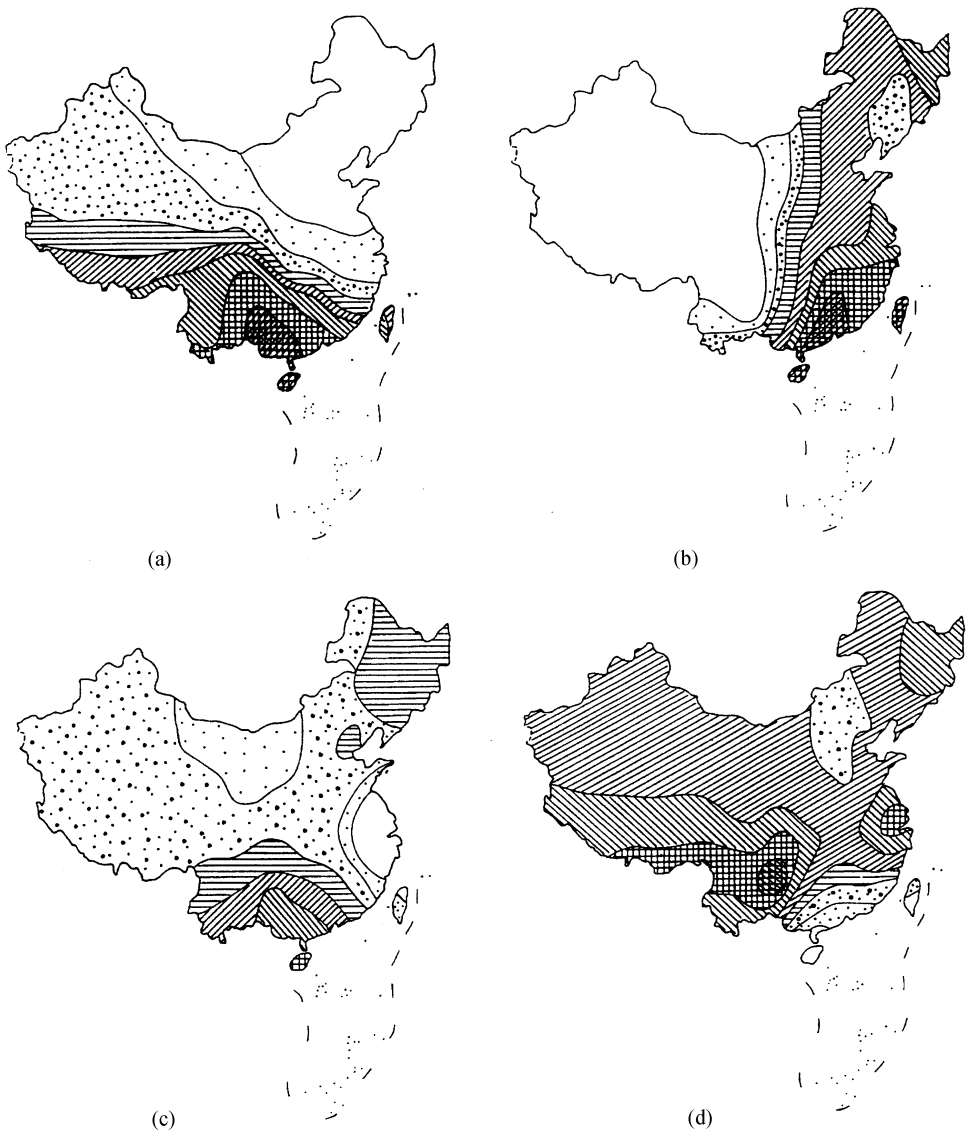

Fig. 1. The sketch maps of PCs of Han. (a) 1st PC; (b) 2nd PC; (c) 3rd PC; (d) 4th PC.

place to another very easily even in winter, because the Yellow River freezes during that time. Another reason is that large-scale demographic migrations often happened there during the last 5 thousand years due to wars and famines. This is why the Han in the northern provinces have a common genetic structure to a large extent. But the situation in the South is quite different from that in North China. Many steep gradients were observed in the first, as well as second, third and fourth, PC maps, indicating more heterogeneity among Southern Han populations living south of the Yangtze River. The more blood component of Southern Mongoloids and the less blood component of Northern Mongoloids were found in Han subpopulations with going further south. The most southern regions, including Guangxi, Hainan and western Guangdong, were the highest.

More severe geographic isolations owning to more mountains and rivers and fewer major migrations in South China must have contributed to this great genetic diversity among the southern Han populations.

The second PC map of the Han conveys 23.9% of the total variance and shows a clear west to east, and southeast gradient. The extreme PC values are in the extreme west and in the extreme south (fig. 1(b)). It may be considered to reflect gene flow between Caucasoid and Mongoloid populations. Among the Han subpopulations, the most important genetic difference is the differentiation between Southern and Northern Mongoloids, with the differentiation between Caucasoids and Mongoloids only listed second. This is because, although the genetic difference between Caucasoids and Mongoloids is much greater than that between Southern and Northern Mongoloids, most blood component of the Han is Mongoloid and only a little of Caucasoid is in the Han. In other words, the gene flows from Caucasoids into the Han are few, which has only a little effect on the whole genetic structure of the Han populations in different provinces or regions of China.

The third and fourth PC maps convey 8.6% and 5.5% of the total variance, respectively (fig.1(c) and (d)). There are extremely dark or light areas on these maps in Guangdong, Guangxi and Hainan; Yangtze delta; middle and Northern Innermongolia, Shanxi and Ningxia; southern Sichuan and so on. We think that these regions with advanced techniques in making stoneware, bronze and porcelain, or with the developed agriculture and animal husbandry, might have experienced the substantial population growth, which led to their expansion into neighboring regions at some ancient time.

The location of these regions might be not very exact because the distribution of sampling places for analyzing gene frequency was not dense enough. Also, the date of these events could be reached on by using combined research in archaeology, anthropology, linguistics, history and others.

2.2 Ethnic minorities

The PC geographic maps of Chinese ethnic minorities are very similar to those of the Han. This may demonstrate that substantial gene flow has occurred from ethnic minorities into the neighboring Hans, and at the same time, some gene from the Han may have mixed into the local ethnic minorities[4]. But some significant differences can still be observed in these two sets of map. The first PC map of ethnic minorities reflects the gene flow between Caucasoids and Mongoloids, while the second PC map shows the gene flow between Southern and Northern Mongoloids. This is just the reverse of the Hans.

The first PC map of ethnic minorities has a clear northwest to east, then to south gradient (fig. 2(a)), which is definitely caused by gene flow between Caucasoids and Mongoloids. The studies on gene frequency of many loci and anthropological characters have showed clearly that among the Chinese ethnic minorities, those living in northwest region have the highest fraction of Caucasoid genes. The same result is also shown in the first PC map of ethnic minorities. One pole is in

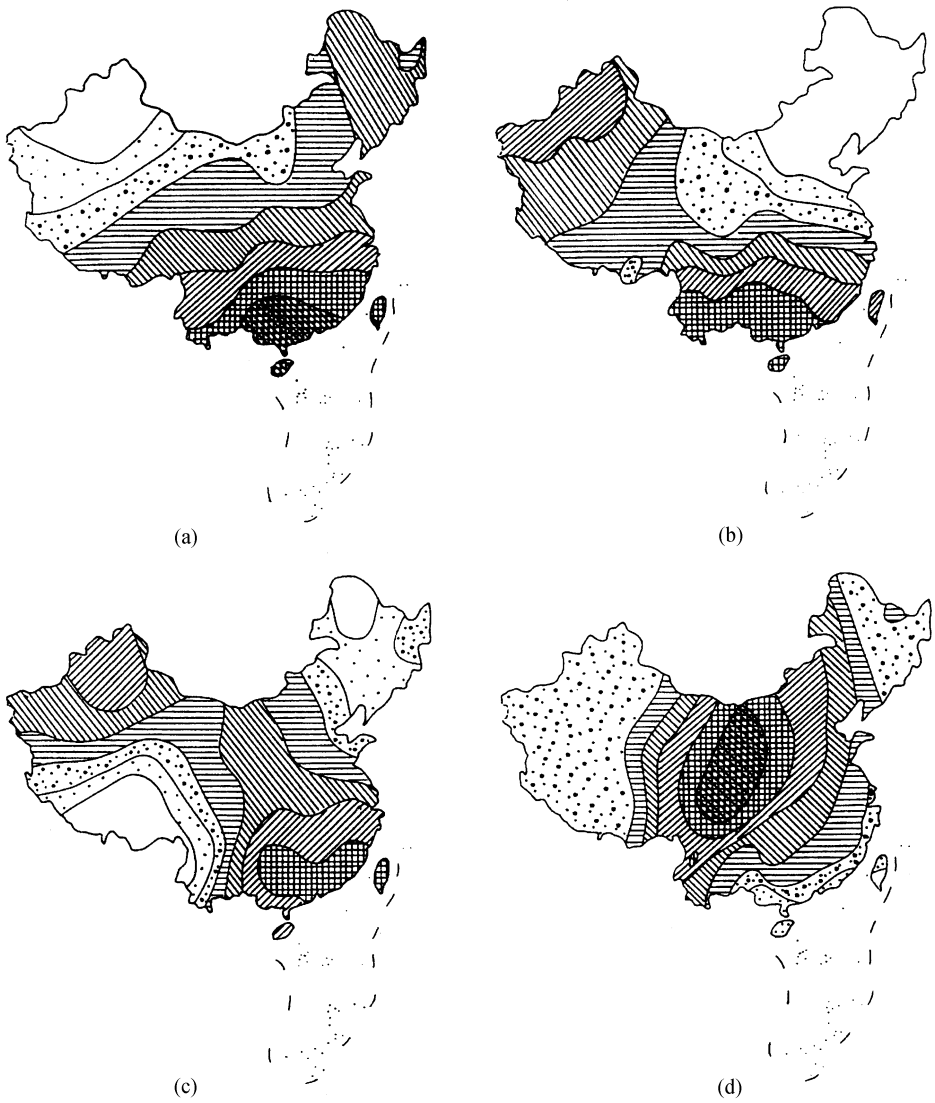

Fig. 2. The sketch maps of PCs of Chinese ethnic minorities. (a) 1st PC; (b) 2nd PC; (c) 3rd PC; (d) 4th PC.

the northwest, another in the farthest south corner including Guangdong, Guangxi and Hainan, where the least evidence of Caucasoid admixture is found. The first PC map of ethnic minorities conveys 30.1% of the total variance.

The second PC map of ethnic minorities accounts for 18.6% of the total variance. The gradient in this map starts from the northeast, first descending towards the west, then bending towards the south. The extreme light color is in northeast, while the darkest in Guangdong, Guangxi, Yunnan and Southern Fujian. This result, like the first PC map of the Han, reflects gene flow between Southern and Northern Mongoloids (fig. 2(b)).

The third PC map of ethnic minorities, accounting for 13.6% of the total variance, shows that

Northeast and Southwest have the extreme light color, while Southeast and Northwest the darkest (fig. 2(c)). These poles, perhaps, are the major origin places of modern Chinese ethnic minorities. At some time in history, the populations in these regions grew drastically and then expanded toward the central area of China. This is only primary conjecture which needs more data from gene frequencies and other information toward doing further analysis.

The fourth PC map conveys 7.5% of the original variation. It significantly shows a peak with concentric gradients toward the middle portion of the Yellow River, which corresponds approximately to area of early development of the Han composed of Yan and Huang tribes to form Hua-Xia group some 5 000 years ago (fig. 2(d)). Thus, the fourth PC map may reflect the result of the radiation of the Han to their neighboring ethnic groups.

But the distribution maps of 1st, 2nd, 3rd, and 4th principle components of Han ethnic group do not show any evident influence of Hua-Xia group on the gene frequencies of Han subpopulations in different places of China. The most possible explanation of this result may be that the gene frequencies of Han subpopulations were effected greatly by the local neighboring ethnic groups other than Han, but not by the Hua-Xia group. It has been demonstrated also by the genetic distances between Han subpopulations and ethnic minorities in China as the average genetic distance between Han subpopulations and their neighboring ethnic minorities is much smaller than that between Han subpopulations.

2.3 Chinese populations including Han and ethnic minorities

When the gene frequency data from both Han and ethnic minorities were analyzed together, the first, second, third and fourth PC maps of the Chinese populations convey 30.4%, 17.2%, 12.2%, and 6.0% of the total variance, respectively. These maps (fig. 3) are very similar to those of the Chinese ethnic minorities (fig. 2), but two major differences can still be observed. One is that the difference between northwest and northeast in the second map of Chinese populations was narrowed. This is not surprising, since the northwestern Han mostly came from the North China. When the data from the Han were added to the analysis, the difference became smaller, compared with that in the second PC map of ethnic minorities. Another is that the two-third PC maps look very similar, except that the dark region in one map was replaced by the light region in the other (fig. 2(c) and fig. 3(c)). The choice of dark or light is totally arbitrary, and it could be reversed if desired, without any loss of information. They both indicate the presence of a population with extreme values in the area being analyzed.

2.4 Synthetic maps of Asia

The combination of the first most important PCs in a color map is called a synthetic map. The synthetic map of the first three PCs of Asia was constructed by analyzing all Asian data collected both in China and in the laboratory of Cavalli-Sforza.

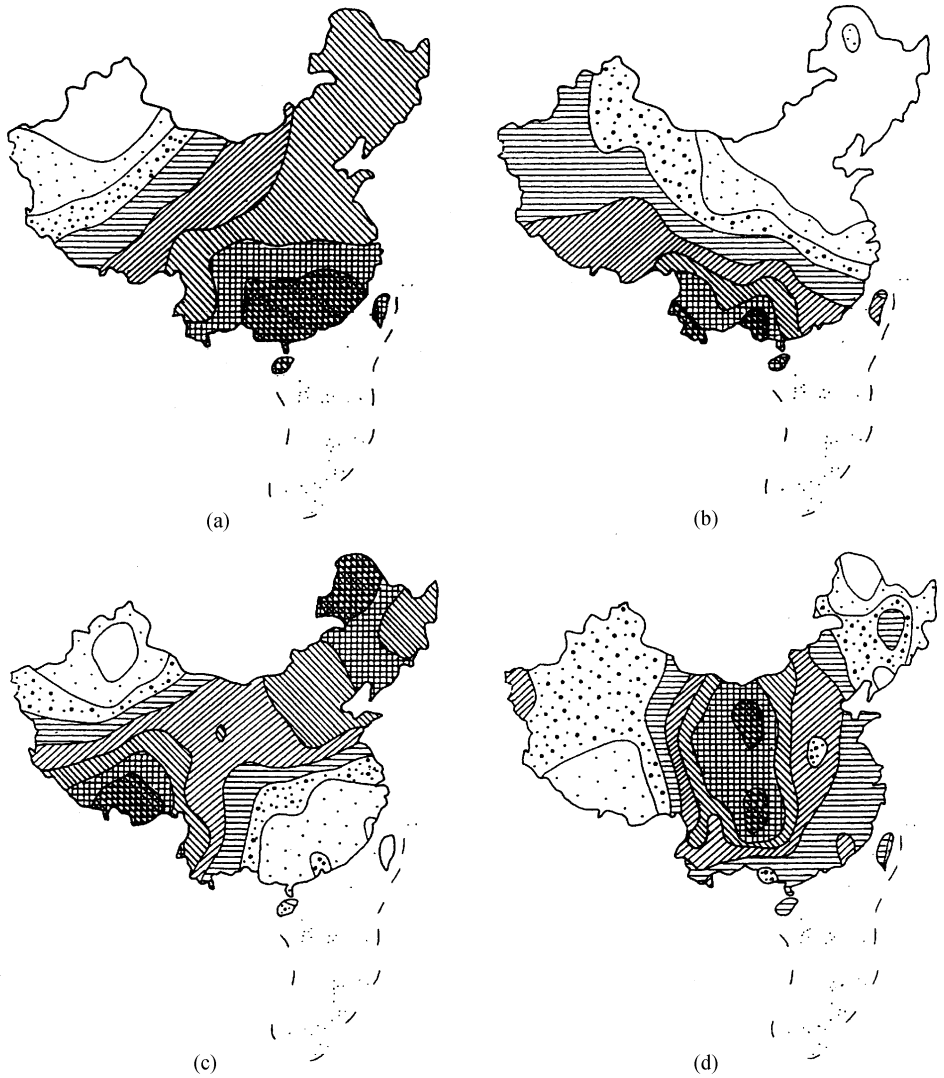

Fig. 3. The sketch maps of PCs of both Han and ethnic minorities. (a) 1st PC; (b) 2nd PC; (c) 3rd PC; (d) 4th PC.

The map (plate Ⅰ) obviously shows that there is a green zone in South China, inhabited by the Southern Mongoloids, which is connected with Southeast Asia; and a purple zone in North China, occupied by the northern Mongoloids, merges with all of Siberia, and turns to have a light blue or greenish purple tinge in Middle Asia which is in conjunction with Europe[1] occupied by Caucasoid. On the synthetic map of the world drawn by Cavalli-Sforza the purple region of East Siberia is in conjunction with America, and the yellow region including Indonesia, the Philippines and New Guinea, is in conjunction with Australia[1]. It is easy to conclude that the Yangtze River as the demarcation line between Southern and Northern Mongoloids in China is so conspicuous.

3 Discussion

3.1 Gene flow between Southern and Northern Mongoloids

It is surprising to find that a very clear boundary between Southern and Northern Mongoloids in the synthetic map of Asia is still seen in the zone of the Yangtze River, although substantial population migrations caused by war or famine happened in the last 2 000 or 3 000 years. The results are the same as revealed by mt-DNA, microsatellites,Y-chromosome polymorphism and dermatoglyphic study[4–7].

In the last 4 000 or 5 000 years, the major direction of population migration in China was from the middle or lower reaches of the Yellow River to south, after crossing the Yangtze River, first toward further south and then west. But in prehistoric times, Southern Mongoloids perhaps crossed the Yangtze River and occupied some places in the north part of the river for sometime, and then were driven back to the south by Northern Mongoloids. According to early historical records, the earliest ancestor (the Yellow Emperor) of the Han once fought with Chi You who was the leader of the southern Miao people, in the zone of Zhoulu of today' s Hebei Province. Chi You was defeated and retreated with his troops to cross the Yangtze River and inhabited the area of today' s the Poyang Lake and the Dongting Lake[8,9].

Some archaeological evidence and the results of cluster analysis based on anthropometric data from paleoanthropological study strongly show the existence of a difference between Southern and Northern Mongoloids in the Neolithic age [10]. Then a question arises: how far can this difference be traced back? Or in other words, when did the southern and northern types of Mongoloids begin to exist separately? This question definitely concerns the origins of modern human, and must be answered by both theories of African and Polycentric origin of modern humans.

The synthetic map of Asia shows that South China, inhabited by Southern Mongoloids, is connected with Southeast Asia, and the north part of the Yangtze River inhabited by Northern Mongoloids merges with Siberia and Northeast Asia. This indicates that the Southern and Northern Mongoloids separated at the early time after they were formed as Mongoloids. If the Northern Mongoloids in China came from South China, or in contrast, the Southern Mongoloid came from the north, it would be impossible to form the clear-cut visible boundary between them near the Yangtze River, which was also reached by a variety of other approaches, archeological, craniometric and dental. Therefore, from the point view of African origin, the probability is that the ancestor of Mongoloids came into China from Africa or the Middle East by two routes and then met in the zone of the Yangtze River. One route might be from Africa via the Southern coast of Asia, toward South Asia and Southeast Asia, then turning to the north to reach the Yangtze River. As mentioned above, the early settlers might have crossed the River and occupied some eastern region of North China in prehistoric times before being forced back to the south side of the Yangtze River. The other route to get into China was through the north part of the Pamirs, toward the east and north to reach Northeast Asia. Later they went toward the south to reach the Yangtze

River, and arrived in Japan via the Korean Peninsula and America via the Bering Strait.

We do not think that the hypotheses of entrance of Mongoloid from North or South into China are definitely wrong, but only propose a third possibility of the entrance of Mongoloid from both south and north into China. Our proposal is on the basis of evident and sharp difference in genetic construction between Southern and Northern Mongoloids with Yangtze River as a boundary that is clearly shown on the synthetic map of principle components of Asia. The conclusive clarification of the origin and migration of Mongoloids needs more data from human population genetics studies, as well as more archaeological findings, especially about the period from 20 to 100 thousand years ago, and also results from linguistic, historical, ethnological and paleogeographic studies.

3.2 Gene flow between Caucasoids and Mongoloids

There are two points in this paper to be sure. First, the genetic difference caused by gene flow between Caucasoids and Mongoloids is only of secondary importance among Han subpopulations, while the most important is the difference between Southern and Northern Mongoloids. But among Chinese ethnic minorities or Han plus ethnic minorities, the difference caused by gene flow between Caucasoids and Mongoloids is in the first position, which is in contrast with that of the Han only.

Second, among Chinese populations, the northwestern populations have the highest component of Caucasoid blood, while those in the most south of China, including Guangdong, Guangxi and Hainan provinces, the least. The genetic gradient caused by gene flow between Caucasoids and Mongoloids tends to show an expansion from Northwest China, first towards the east and then south, and ultimately, most southern China. This indicates that a very small component of Caucasoid genes were brought by the Persian merchants in Guangzhou and Quanzhou into Chinese populations after the Tang and Song Dynasties, while nearly no Caucasoid blood from India mixed into Chinese populations. The flow of Caucasoid genes into China happened nearly all in north China, from west to east, then towards the south, and mainly in the prehistoric age, not after the Silk Road had been opened.

Acknowledgements We thank Prof. M. W. Feldman, Department of Biological Science, Stanford University, for his financial support and Dr. Mark T. Seiestad , Program for Population Genetics, Harvard School of Public Health, for his help.

References

1. Cavalli-Sforza, L. L., Menozzi, P., Piazza, A., The History and Geography of Human Genes, Princeton: Princeton Univ. Press, 1994, 541, 518.
2. Du, R. F., Xiao, C. J., Cavalli-Sforza, L. L., Genetic distances calculated on gene frequencies of 38 loci, Science in China, Ser. C, 1997, 40(6): 613.
3. Du, R. F., Yip, V. F. -S., Ethnic Groups in China, Beijing: Science Press, 1993, 318.
4. Chu, J. Y., Huang, W., Kuang, J. M. et al., Genetic relationship of populations in China, Proc. Natl. Acad. Sci. USA, 1998, 95:11763.
5. Xiao, J. H., Hu, F., Xu, H. Y. et al., Provincial distribution of three HIV-1 resistant polymorphisms (CCR5-Δ32, CCR2-64I, and SDF1-3′ A) in China, Science in China, Ser. C, 2000, 43(1): 16.

6. Xu, L. P., Xu, J. J., Zhu, S. L. et al., Distribution of YAP in 10 populations of China, Chinese Science Bulletin, 1998, 43(12): 1023.
7. Zhang Hai-guo, Ding Ming, Jiao Yunping et al., A dermatoglyphic study of the Chinese population III—Dermatoglyphics cluster of fifty-two nationalities in China, Acta Genetica Sinica, 1998, 25(5): 381.
8. Wu, X., Long, B., History of Miao Ethnic Group (in Chinese), Chengdu: Sichuan National Press, 1992, 2–38.
9. Fan, W., Comprehensive History of China (in Chinese) (Revised Edition), vol.1, Beijing: People's Press, 1965, 88–94.
10. Zhang, Z., Human remains of the Neolithic Age in China, in The Ancient Human Beings in China (ed. Wu Rukang) (in Chinese), Beijing: Science Press, 1989, 62–80.

The synthetic sketch map of the first three PCs of China.