A Study of Surnames in China Through Isonymy

Yan Liu,[1] Liujun Chen,[1] Yida Yuan,[2] and Jiawei Chen[1]*

[1]*Department of Systems Science, School of Management, Beijing Normal University, Beijing 100875, People's Republic of China*
[2]*Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, People's Republic of China*

KEY WORDS Chinese; surname distribution; isolation by distance; migration

ABSTRACT The isonymy structure of 1.28 billion people registered in China's National Citizen Identity Information System was studied at the provincial, prefectural, and county administrative division levels. The isonymy was 0.026 for China as a whole. The average value of isonymy was 0.033 for the 30 provinces, 0.035 for the 334 prefectures, and 0.040 for the 2811 counties. The isonymy in China was much higher than in other countries. This finding may be partly explained by the low number of surnames in the Chinese language. Two regional features can be identified from the geographic distributions of isonymy. One feature is that the middle and lower reaches of the Yangtze River had the lowest values of isonymy at both the provincial and county levels. The second feature is that most counties with the highest values of isonymy were distributed in the provinces with high proportions of ethnic minorities. According to the dendrogram of surname distances, several clusters could be identified. Most provinces in a cluster were conterminous with one another. The one exception could be explained by demic migration called "braving the journey to the northeast of China." Isolation by distance could be detected because the correlation coefficients between Nei's distance and the geographic distances at the provincial, prefectural, and county levels were 0.64, 0.43, and 0.37, respectively. Human behaviors in Chinese history that may have caused these results have been discussed, including cultural origin, migration, residential patterns, and ethnic distribution. Am J Phys Anthropol 148:341–350, 2012.

Surnames are inherited through the male line and can be considered alleles of a gene on the Y chromosome (Zei et al., 1983). Therefore, surnames satisfy the expectations of the neutral theory of evolution, which is described by random genetic drift, mutation and migration (Kimura, 1983). Surname distribution among ethnic groups or geographic areas can differ significantly and can provide quantitative information about the genetic structure of different groups. Furthermore, social behaviors, such as living habits and migrations, play important roles in surname distribution. Thus, a study on surname distribution can serve as a bridge for topics related to social behaviors and genetic structure. Such studies have been performed in many countries through isonymy theory, which provides a useful tool for exploring population structure by extracting important features of surname distribution (Branco and Mota-Vieira, 2003, 2005; Colantonio et al., 2003; Bronberg et al., 2009; Rodriguez-Larralde et al., 2011; and references therein). In these works, genetic features, such as the effective allele number and consanguinity are related to the isonymy parameters, which are determined by the evolution of human groups.

The dynamics of population structure are a central consideration. The relative importance of drift and migration—whether drift predominates over migration or migration predominates over drift—can be identified through isonymy analysis. In most European countries, drift has a dominant effect because populations have settled for long enough to permit drift and some local dispersion of surnames, as indicated by the existence of isolation by distance (Rodriguez-Larralde et al., 1998a,b; 2003; Scapoli et al., 2005). In contrast, recent immigration in the 20th century has played a determining role in the population structure of the United States, as indicated by the lack of relevant isolation (Barrai et al., 2001). In other countries, such as Venezuela (Rodriguez-Larralde et al., 2000), Argentina (Dipierri et al., 2005), Yakutia (Tarskaia et al., 2009), Bolivia (Rodriguez-Larralde et al., 2011), and Paraguay (Dipierri et al., 2011), due to significant isolation by distance and some signs of migration, the population structure should be the result of the joint action of drift and migration.

China has a 4,000-year history of recorded surnames, extending as far back as the Xia Dynasty (ca. 21–16 centuries BC), which have undergone a long evolutionary process. Chinese surnames have been well preserved through generations due to the prevalence of Confucian culture, in which people do not change their surnames unless there are special reasons to do so, such as receiving a noble surname from the emperor (Du et al., 1992). Research on the 100 most frequent surnames in China showed that the Zipf plots were exponential and have been maintained since the time of the Song Dynasty (Yuan and Zhang, 2002; Baek et al., 2007). The stability

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Ministry of Education Humanities and Social Sciences; Grant number: 11YJC840006.

*Correspondence to: Jiawei Chen, School of Management, Beijing Normal University, Beijing 100875, People's Republic of China.
E-mail: chenjiawei@bnu.edu.cn

Received 3 November 2011; accepted 20 February 2012

DOI 10.1002/ajpa.22055
Published online 28 March 2012 in Wiley Online Library (wileyonlinelibrary.com).

of surnames indicates that the historical inheritance of Chinese surnames has been continuous, approaching drift-migration equilibrium after thousands of years of surname evolution. Therefore, surnames in China, as a cultural genetic factor, may be a significant and remarkable resource for studying population structure.

China is a multiethnic country, with the Han nationality as the largest ethnic group. Over the evolution of Chinese surnames, the most important factors affecting surname frequencies have been ethnic assimilation and migration (Du and Yuan, 1995). Chinese/Han civilization, which originated in the central plains on the middle and lower reaches of the Yellow River, was diffused and blended through sustained migration. The duration, mass, and geographical scale of such migrations are rare in world history. Migration in Chinese history can be classified into two categories. In the first form, people migrated from outside into the central plains, coming primarily from the north and secondarily from the west. This migration was mostly in the form of military invasions by non-Han ethnic groups. The ethnic minorities that invaded or migrated into central China were assimilated by and integrated into the Han people; thus, most of them adopted the surnames of the Han nationality. Another category involves the migration of Han people outward in all directions from the central plains. The major part of these migrations was realized by movements in search of livelihood, from densely populated areas to sparsely populated areas.

This surname evolution resulted in several specific features of surname structure. The first feature is that there are a smaller number of surnames and a larger number of people sharing the same surname. It has been estimated that there are currently approximately 3000 surnames in use for the Han nationality (Yuan et al., 2000a). It is surprising that the 100 most common surnames account for ~85% of the total population. Furthermore, the three most common surnames have consistently been Wang, Li and Zhang, which cover 21, 17, 18, and 21% of the total population in the Song, Yuan, and Ming Dynasties and the present, respectively (Yuan and Zhang, 2002).

The second feature of Chinese surnames is that there is remarkable regionality in the geographic distribution of surnames. Although the surnames of the Han nationality spread from central China throughout the country during periods of population migration and ethnic assimilation, these migrations were quite uneven. Therefore, some surnames may be regional due to specific migrations. Furthermore, there is a custom of developing concentrated communities with the same surnames, which has resulted in regional features of surnames, especially at smaller scales.

The third feature is that most of the commonly used surnames are polyphyletic. As Crow (1983) noted, for isonymy to be given a simple interpretation of inbreeding coefficient, each name must trace to a single individual. This requirement cannot be satisfied for Chinese surnames with 4,000-year histories. Yuan and Zhang (2002) observed that 97 of the 100 most common surnames originated in the Spring and Autumn Period (722–476 BC) or the Warring States Period (476–221 BC). During these periods, changing one's surname was more common because Confucian ideas did not prevail. Furthermore, ethnic minorities often changed their surnames to surnames of the Han nationality over thousands of years of surname evolution. In general, the more common a surname is the more origins the surname has. Therefore, the isonymy parameters in this paper should not be interpreted as an indicator of inbreeding coefficient. However, there are genetic meanings of the surname distribution in China because these surnames were handed down through generations after they were adopted. The structure of the Chinese population and the regional consanguinity of the population, such as the genetic difference between northern and southern China, were discussed in an isonymy analysis of Chinese Han surnames (Du et al., 1992; Yuan et al., 1999; Yuan et al., 2000a,b).

In this article, we extend these studies to a larger sample of the Chinese population, including ethnic minorities at three administrative division levels, the province, prefecture, and county. The aim of this study was to explore the isonymy structure in China and extract its geographic features to determine the relative importance of drift and migration in determining the evolution of Chinese surnames. This study determines isonymy and presents its corresponding geographic distribution on maps. A matrix of surname distances among provinces and the corresponding dendrogram are constructed, and clusters are identified from the dendrogram. Isolation by distance is discussed based on the correlation of surname distance and geographic distance at all three levels.

MATERIALS AND METHODS

Data and material

The surname dataset in this work was obtained from China's identity information system, which was constructed by the National Citizen Identity Information Center (NCIIC). A large population of 1.28 billion people with 7,327 surnames in 2007 was included in this dataset, which listed the number of individuals for each surname. The original dataset presents 7,329 surnames and excludes Hong Kong, Macao, and Taiwan. This paper also excluded the Tibet Autonomous Region, for reasons that will be explained later. To protect privacy, the NCIIC replaced each surname with a 5-digit number, called a surname ID. It is worth noting that our dataset differed from the study by Du et al. (1992), which used the 1982 Sample Census of 537,429 Han people with 1,054 surnames.

Three levels of administrative divisions were considered in this article. The first level in China is called the provincial level (or province), which consists of provinces, municipalities, autonomous regions, and special administrative regions. Below the provincial level is the prefectural level (or prefecture), which consists of prefectures, prefecture-level cities, autonomous prefectures, and leagues. Below the prefecture level is the county level (or county). Our dataset included 30 provinces that were divided into 334 prefectures and subdivided into 2811 counties. Detailed information for all 30 provinces is provided in Table 1. The administrative map of China used in this article was developed by ArcGIS and the longitudes and latitudes of all prefectures and counties were obtained from Google Maps to calculate geographic distance.

There are 56 ethnic groups in China, of which the largest is the Han nationality. The proportion of ethnic minority populations in each province, as shown in Table 1 was obtained from the website (www.stats.gov.cn) of the

TABLE 1. Isonymy analysis of Chinese surnames

Items	SS (10^6)	S	M (%)	C	I	AI	RI
Country	1,277	7,327	8.4	2811	0.026	0.040	1.54
Beijing[a]	11.89	1,997	4.3	18	0.038	0.040	1.05
Tianjin[a]	9.10	1,657	2.6	17	0.044	0.045	1.02
Hebei	68.73	3,405	4.3	174	0.042	0.044	1.05
Shan1xi	32.99	3,326	0.3	119	0.038	0.044	1.16
Neimenggu[b]	23.52	3,634	20.8	101	0.034	0.034	1.00
Liaoning	41.92	2,698	16	101	0.036	0.038	1.06
Jilin	26.57	2,718	9	60	0.037	0.039	1.05
Heilongjiang	37.45	2,764	5	132	0.037	0.037	1.00
Shanghai[a]	13.68	1,614	0.6	19	0.023	0.024	1.04
Jiangsu	72.66	3,116	0.3	119	0.025	0.028	1.12
Zhejiang	45.89	2,415	0.8	89	0.023	0.033	1.43
Anhui	65.15	4,451	0.6	105	0.027	0.031	1.15
Fujian	33.68	2,162	1.7	73	0.043	0.053	1.23
Jiangxi	43.75	2,607	0.3	99	0.022	0.034	1.55
Shangdong	92.89	3,429	0.7	140	0.039	0.043	1.10
Henan	100.94	4,282	1.2	159	0.037	0.040	1.08
Hubei	59.44	4,058	4.3	102	0.024	0.029	1.21
Hunan	67.61	3,331	10.2	122	0.025	0.041	1.64
Guangdong	79.08	2,991	1.4	129	0.032	0.044	1.38
Guangxi[b]	49.17	2,872	38.3	108	0.031	0.054	1.74
Hainan	8.08	1,798	17.3	22	0.043	0.063	1.47
Chongqing[a]	31.98	2,360	6.4	40	0.025	0.029	1.16
Sichuan	86.60	4,330	5	181	0.025	0.032	1.28
Guizhou	38.49	3,333	37.8	88	0.028	0.044	1.57
Yunnan[b]	42.69	4,350	33.4	129	0.035	0.055	1.57
Shan3xi	37.10	3,329	0.5	107	0.034	0.038	1.12
Gansu	25.96	3,191	8.7	87	0.036	0.058	1.61
Qinghai	5.01	2,905	45.5	43	0.030	0.038	1.27
Ningxia[b]	5.86	2,044	34.5	21	0.046	0.052	1.13
Xinjiang[b]	18.95	3,500	59.4	107	0.024	0.035	1.46

SS, sample size; S, surnames; M, proportion covered by the population of minority nationalities; C, number of counties in a unit; I, isonomy; AI, average isonymy for counties; and RI, ratio of AI to I.
[a] Municipalities.
[b] Autonomous regions.

National Bureau of Statistics of China (NBSC). A population of 108 million ethnic minorities is distributed throughout China, accounting for 8.4% of the total population. There are seven provinces in the west and south of China, including Tibet, Xinjiang, Ningxia, Qinghai, Guangxi, Guizhou, and Yunnan, where ethnic minorities account for more than 30% of the population.

The naming system in Chinese culture, in which the surname is first and the given name follows, differs from Western cultures. Generally, the Han people follow this naming system. However, the naming system of ethnic minorities may differ significantly from the naming system of the Han people (Qian, 1989). There are 34 ethnic minority groups, such as the Zhuang, Hui, and Man ethnic groups, with surnames. Some of these surnames are similar to those of the Han people, whereas others are unique. In 14 ethnic minority groups, such as the Yi, Miao, Tibetan, and Mongolian ethnic groups, most people have no surnames. Nine ethnic minority groups, such as the Uyghur ethnic group, have no surnames. In groups without surnames, children are named according to a patronymic linkage naming system or a matronymic linkage naming system, sometimes even irregularly (such as in Tibetan ethnic groups). In these cases, surnames are represented in our dataset by the first character in the name and have almost no correlation with origin. In the Tibet Autonomous Region, the Tibetan ethnic group comprises 94.1% of the population, and most people have no surname; this is why the Tibet Autonomous Region was excluded from this article.

Isonymy theory

This paper used a methodology similar to that described in Rodriguez-Larralde et al. (2011), involving a series of parameters, including isonymy, Fisher's α, and surname distance. Isonymy is a statistic that can predict the inbreeding frequency in a given region (Crow and Mange, 1965; Crow, 1980). Considering the polyphyletism in Chinese surnames, it is inappropriate to interpret isonymy simply as an indicator of inbreeding frequency. However, these isonymy parameters are helpful for measuring the structure and regional consanguinity of the Chinese population, as shown in previous studies (Du et al., 1992; Yuan et al., 1999; Yuan et al., 2000a,b). Below, we define some statistics derived from the surname distribution.

Isonymy within and between groups. The isonymy within a group is defined as $I = \sum_{k=1}^{S} p_{kj}^2$, where S is the number of surnames and p_{kj} is the relative frequency of surname k in a group j, which is the proportion of the population with surname k to the entire population. Isonymy between two groups is a statistic that assesses similarities in populations at the limit of common origin. This statistic is defined as $I_{ij} = \sum_{k=1}^{S} p_{ki}p_{kj}$, where S is the number of the same surnames in the two groups and p_{ki} and p_{kj} are the relative frequencies of surname k in the groups i and j, respectively.

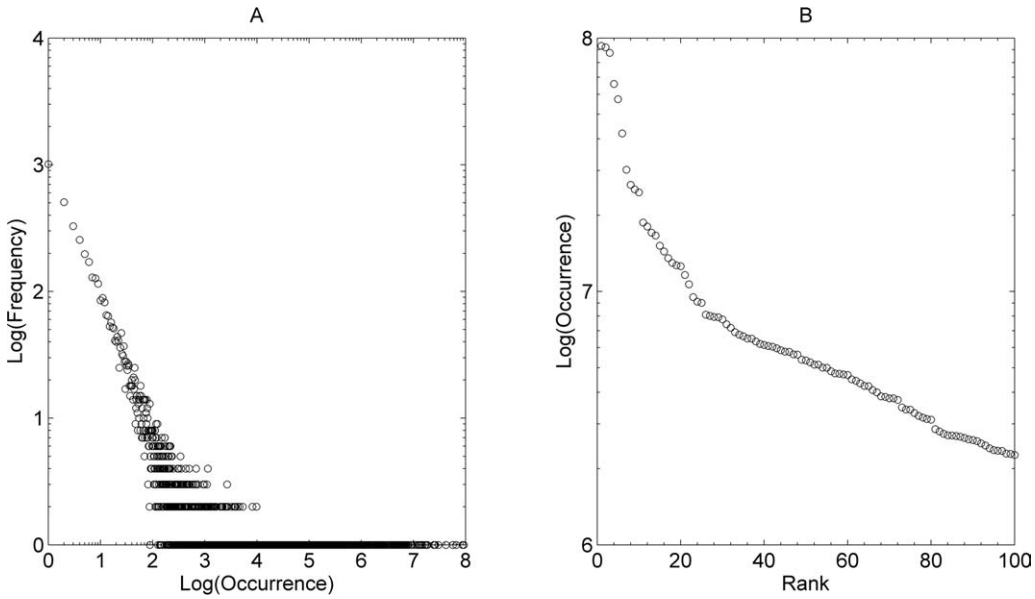

Fig. 1. Surname distribution. **A**: The log–log distribution of the frequency of occurrence of surnames in China. **B**: The occurrence of the 100 most common surnames, arranged by rank (the Zipf plot).

Fisher's alpha. Fisher's α can be directly estimated from $\alpha = \frac{1}{I}$, according to Barrai et al. (1996). The value of α estimates the number of surnames with an equal frequency and is often defined as "the Effective Surname Number". A small α value indicates large genetic drift, whereas a large value indicates migration.

Isolation by distance. There are three kinds of surname distance between group i and j: Lasker's distance, Euclidean distance and Nei's distance. Lasker's distance (Rodriguez-Larralde et al., 1998b) is defined as $L = -\log(I_{ij})$. Euclidean distance (Cavalli-Sforza and Edwards, 1967) is defined as $E = \sqrt{1 - \sum_{k}^{S} \sqrt{p_{ki}\, p_{kj}}}$, where the summation is over all surnames. Nei's distance (Nei, 1973) is defined as $N = -\log\left(I_{ij} / \sqrt{I_i I_j}\right)$.

Isolation by distance can be studied through the linear correlation of surname distances and geographic distances at the provincial, prefectural and county levels. At the provincial level, the distances between the provincial capital cities of the provinces were used as geographic distances.

RESULTS

Frequency distribution and the most frequent surnames

The log–log frequency distribution of the occurrence of surnames (Fox and Lasker, 1983) is shown in Figure 1A. Compared with other countries, there are two striking features in this graph. One feature is that the linear part obtained by truncating the long tail is flattened, which could be attributed to the low number of Chinese surnames, with only 7,327 surnames among 1.28 billion people. Even throughout Chinese history, the total number of Chinese surnames collected in related literature was just over 11,000, which is far smaller than the number of European surnames (Du et al., 1992). This small number of surnames is related to the small number of Chinese characters in the Chinese language; Chinese surnames often consist of a single Chinese character, and only several thousand Chinese characters are commonly used. Chinese words often combine two or more Chinese characters; thus, a large number of words are based on a limited number of characters.

The second feature in the graph is that there is an extremely long tail of approximately four orders of magnitude. The long-tailed distribution implies that the few most common surnames account for a large proportion of the total population. The 100 most common surnames account for 85% of the total population in China. This is a high percentage compared with other countries: the 100 most common surnames account for 8.1% of the population in France (Scapoli et al., 2005), 16% in the United States (Barrai et al., 2001), and 29.5% in Argentina (Dipierri et al., 2005). The phenomenon of a small number of surnames shared by a large number of people may be the result of the evolution of Chinese surnames, as noted by Du et al. (1992), suggesting a strong effect of drift.

The Zipf plot of the 100 most common surnames is shown in Figure 1B and is approximately exponential. This distribution is qualitatively different from other countries and may be explained by the fact that the number of new surnames in China is dependent on time rather than on population size (Baek et al., 2007). The 100 most common surnames in the Song Dynasty (AD 960–1279), the Yuan Dynasty (AD 1271–1368), the Ming Dynasty (AD 1368–1644), and the present (AD 2007) are shown with their corresponding frequencies in Supporting Information Table 1. The most common surname has consistently been Wang, from the Song Dynasty to the present, followed by the next four most common surnames: Li, Zhang, Liu, and Chen.

Isonymy and the prefecture effect

We calculated the isonymy for each of the provinces, prefectures, counties. The values are summarized in Table 1, and the histograms are presented in Figure 2. The isonymy was 0.026 for China as a whole. The

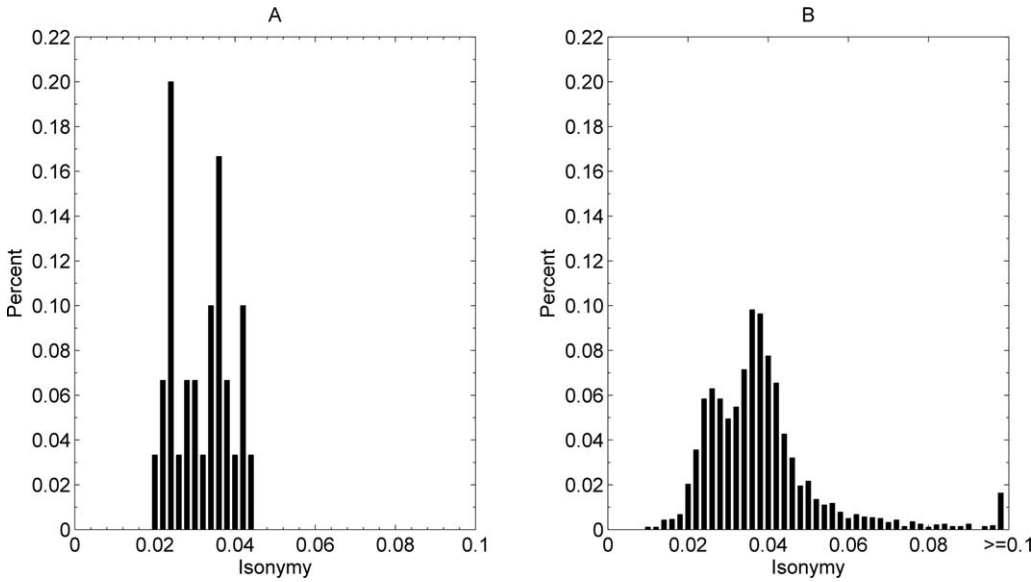

Fig. 2. Histogram of isonymy. **A**: For the 30 provinces. **B**: For the 2811 counties.

isonymy for the 30 provinces ranged from 0.022 to 0.046, with an average value of 0.033. The isonymy for the 334 prefectures ranged from 0.0125 to 0.1943, with an average value of 0.035. The isonymy for the 2,811 counties varied significantly, with an average value of 0.040, ranging from 0.01 to 0.1 in most counties and exceeding 0.1 in 44 counties.

The prefecture effect, named by Scapoli et al. (2007), can be identified qualitatively and quantitatively. A comparison of the isonymic histogram for provinces and counties qualitatively indicates that provincial-level isonymy is generally lower than county-level isonymy because the latter is more flattened and skewed to the right than the former, as shown in Figure 2. Quantitatively, the prefecture effect can be measured as the ratio of AI to I for each province, where AI denotes the average value of a set of county-level isonymies within a province. The results show that the prefecture effect is obvious for provinces that are significantly inhabited by ethnic minority groups. Guangxi, Hunan, and Gansu are the three provinces with the most significant prefecture effect.

Geographic distribution of isonymy

The provincial-level geographic distribution of isonymy is shown in Figure 3. The values of isonymy are classified into four grades on the map. The five provinces with the highest grade of isonymy are Ningxia, Fujian, Hainan, Tianjin, and Hebei, which are distributed throughout the country. The seven provinces with the second-highest grade are situated in the northeastern and northern areas of China. The nine provinces with lower grade of isonymy are situated in the northwestern and southern areas of China. The nine provinces with the lowest grade of isonymy are situated in the middle area of China, on the middle and lower reaches of the Yangtze River.

The county-level geographic distribution of isonymy is shown in Figure 4. The values of isonymy are also divided into four grades on the map. The geographic distribution is qualitatively similar to the distribution at the provincial level, and the regions situated in the middle and lower reaches of the Yangtze River have the lowest grade of isonymy. Other counties with the lowest grade of isonymy are located in several provinces, including Xinjiang Uyghur Autonomous Region, Qinhai, and Neimenggu Autonomous Region. This may be explained by the fact that the ethnic minorities in these regions have no surnames, as described above. Furthermore, some of the counties with the highest grade of isonymy are not located in the provinces with the highest grade of isonymy. In fact, most of these counties are distributed within provinces high proportions of ethnic minorities, such as Guangxi Zhuang Autonomous Region and Yunnan.

One regional feature in the geographic distributions is that the middle and lower reaches of the Yangtze River have the lowest values of isonymy at both the provincial and county levels. This phenomenon may be due to the multiple large-scale migrations in different periods of Chinese history. Yuan et al. (2002) noted that 97 of the 100 most common surnames originated in the Spring and Autumn Period (722–476 BC) and the Warring States Period (476–221 BC), when the territory was limited to the central plains on the middle and lower reaches of the Yellow River. The most frequent surnames in the Yangtze River Basin, such as Chen, Lin, and Huang, originated in the central plains during the Wu Hu Period (AD 304–439) and the Southern and Northern Dynasties (AD 420–589). There was another significant southward migration of the Han ethnic group during the Song Dynasty, which resulted in the development of the Southern Song Regime. The population of the Yangtze River Basin has consisted of local citizens and migrant groups from the central plains in different periods. Consequently, the Yangtze River Basin has the lowest level of isonymy.

Another regional feature is that most of the counties with the highest levels of isonymy are distributed within provinces with a high proportion of ethnic minorities. This phenomenon is related to the characteristics of residence patterns and ethnic distribution in China. There is a custom of developing concentrated communities with

Fig. 3. Provincial-level geographic distribution of isonymy.

the same surnames in China, which may result in a high isonymy value. Ethnic minorities, with their own unique surnames, generally marry within their own ethnic groups and tend to develop residential centers even more than the Han.

Nei's distances and clustering results

Nei's distances between any two provinces were calculated. The dendrogram obtained from the matrix of Nei's distances is shown in Figure 5. Six compact clusters can be identified from the dendrogram, A–F. The map of these clusters is also shown in Figure 6, where it can be seen that all of the clusters (except one) consist of conterminous provinces. Cluster A includes three northeastern provinces (Heilongjiang, Jilin, and Liaoning) and Shandong. These four provinces in cluster A are the most closely connected in the dendrogram, but they are separated geographically. Cluster B (Beijing, Hebei, Tianjin, and Neimenggu Autonomous Region) is situated in the northern area of China. Cluster C (Shanxi, Henan, and Shan3xi) is situated in the central plains on the middle and lower reaches of the Yellow River, which was the cultural birthplace of China. The provinces in clusters A, B, and C were quite closely connected in the dendrogram, whereas the provinces in clusters D, E, and F were not. Cluster D (Shanghai, Jiangsu, Anhui, and Zhejiang) and cluster E (Jiangxi, Hubei, Chongqing, Sichuan, and Hunan) were situated in the lower and middle streams of the Yangtze River, respectively. The three provinces (Fujian, Guangdong, and Hainan) in cluster F, situated in the south area of China, were the most loosely connected.

Our clustering result from the dendrogram shows that the surname structures in provinces with a high proportion of ethnic minorities differ from other provinces. That is, the six provinces with proportions of ethnic minorities that exceed 30%, Xinjiang, Ningxia, Qinghai, Guangxi, Yunnan, and Guizhou, are the outliers of all of these clusters. This result may be explained by the fact that surnames for ethnic minorities are unique and are quite different from those of the Han ethnic group, as mentioned above.

Furthermore, the six compact clusters constitute three major clusters, labeled G, H, and F in the dendrogram. The major cluster G consists of clusters A, B, and C and

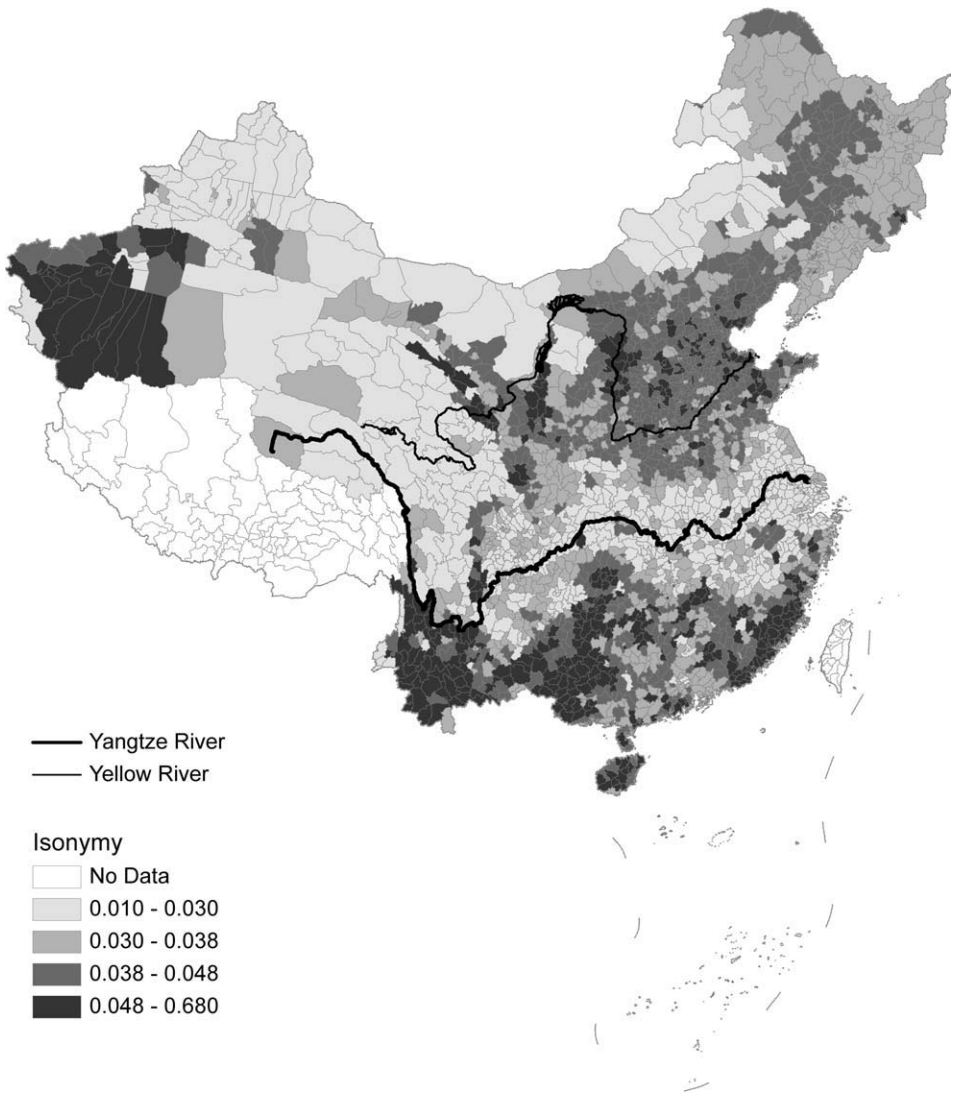

Fig. 4. County-level geographic distribution of isonymy.

the three outliers (Gansu, Yunnan, and Guizhou). Clusters A and B form a whole with a surname structure that is quite similar to that of cluster C. The major cluster H includes clusters D and E. The major clusters G, H, and F roughly correspond to the northern, middle, and southern areas of China, respectively. This clustering result is qualitatively different from the results by Du et al. (1992), which identified two major clusters, northern and southern China. Furthermore, the provinces with a high proportion of ethnic minorities could not be put into any clusters in our study. The reason for this difference may be that surnames used by ethnic minorities were included in our study but excluded in the study by Du et al. (1992).

Our clustering results show that the provinces in the same cluster are almost conterminous, indicating that Nei's distance is correlated with the geographic distance at the provincial level. However, there is an interesting exception in cluster A, the most closely related cluster: the three northeastern provinces and Shandong are not adjacent because the Bohai Sea separates them, but Nei's distances between them are the shortest. This phenomenon is caused by the large and long-lasting migration called "braving the journey to the northeast of China", which occurred from the Qing Dynasty to the period of the Republic of China in the last two centuries. During this migration, more than 20 million people moved from Shandong province to the three northeastern provinces, which had previously been sparsely populated.

Isolation by distance

Isolation by distance was studied through the correlations between surname distance and geographical distances at the provincial, prefectural, and county levels in China. The correlation coefficients of Nei's, Euclidean, and Lasker's distances with geographic distance among the 30 provinces were 0.64, 0.67, and 0.52, respectively (with high significance, $P < 0.0001$). For the 2,811 counties, the correlation coefficients were 0.37, 0.40, and 0.38, slightly smaller than those at the provincial level, which might be due to the presence of more data points at the county level than at the provincial level. Given the high correlation between the three measures of distance ($r_{[\text{Nei-Euclidean}]} = 0.86$, $r_{[\text{Nei-Lasker}]} = 0.85$, and

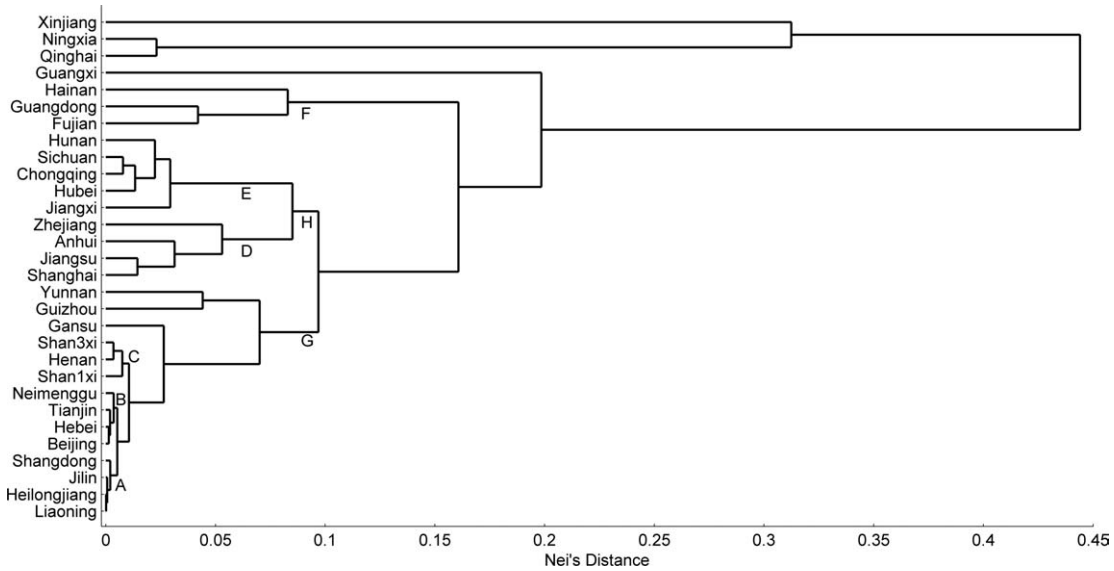

Fig. 5. Dendrogram obtained from the matrix of Nei's distances.

Fig. 6. Map of the main clusters identified by the dendrogram.

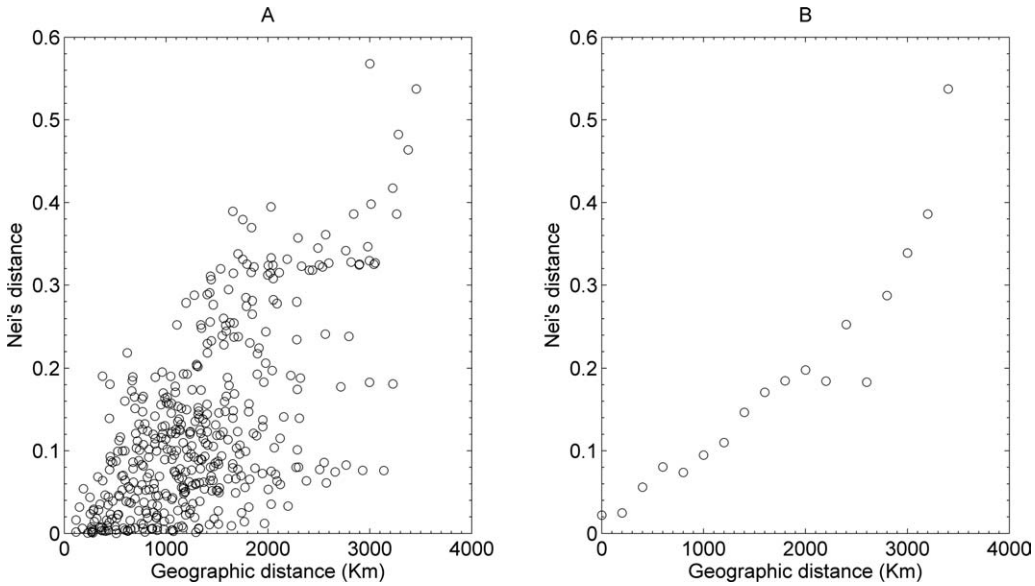

Fig. 7. Variation of Nei's distance between provinces as a function of geographic distance in China. **A**: Scatter diagram including 435 points. **B**. Signal extraction from the scatter diagram, with each point representing the average value of Nei's distance for every 200 km.

TABLE 2. Comparison of isonymic structure in four European countries, Yakutia, China, the United States, and four South American countries

Country	SS (10^6)	S	Level	Divisions	α (average)	Points	r_L	r_E	r_N
Austria	1	140,766	Towns	120	854	7,140	0.565	0.44	0.35
France	6	495,104	Regions	21	4,229	210	0.692	0.546	0.610
			Departments	94	3,546	4,371	0.646	0.502	0.576
			Towns	809	1,615	–	–	–	–
Germany	5.2	462,526	Towns	106	1,596	5,565	0.51	0.48	0.51
Spain P[a]	3.6	94,886	Towns	283	134	39,903	0.128	0.205	0.029
M[a]		110,034			144		0.180	0.263	0.082
Yakutia	0.5	44,625	Districts	35	311	595	0.513	0.629	0.693
			Towns	497	106	–	–	–	–
China	1,276.8	7,327	Provinces	30	32	435	0.52	0.67	0.64
			Prefectures	334	31	55,611	0.40	0.47	0.43
			Counties	2,811	28	3,948,210[b]	0.38	0.40	0.37
United States	18	899,585	States	48	–	1,128	0.24	0.16	0.17
			Towns	247	1,366	30,381	0.21	–	–
Argentina	22.6	414,441	Districts	24	422	276	0.248[c]	0.474[c]	–
Venezuela	3.9	68,665	States	22	122	231	0.35[c]	0.78[c]	–
Bolivia	23.2	174,922	Provinces	112	122	6,216	0.56	0.55	0.5
Paraguay	4.8	39,047	Departments	18	141	153	0.582	0.713	0.597
			Districts	237	108	27,966	0.422	0.320	0.235

SS, sample size; S, surnames; α, average value of Fisher alpha; and r_L, r_E, and r_N, correlation coefficients by Lasker's, Euclidean, and Nei's distances.
[a] P, paternal; M, maternal.
[b] Some points were excluded because the geographic distance between two counties with indistinguishable longitudes and latitudes is zero.
[c] The correlation coefficients were between the logarithmic transformation of geographic and surname distances.

$r_{[\text{Euclidean-Lasker}]} = 0.78$), we used Nei's distance for the subsequent analysis.

The scatter diagram of Nei's distance over kilometers at the provincial level is shown in Figure 7A. The minimum Nei's distance between two provinces was observed between Liaoning and Heilongjiang, with a distance of 0.00019 Nei units and 510 km apart. The maximum Nei's distance was between Guangxi and Xinjiang, with a distance of 0.57 Nei units and 3,000 km apart. An interesting phenomenon was that Nei's distances between two provinces could be very small even though the provinces were located far apart. Therefore, many points were centralized near the horizontal axis, indicating some long distance migrations in Chinese history, such as "braving the journey to the northeast of China", which was from Shandong to the three northeastern provinces.

The signal extracted from the scatter diagram is shown in Figure 7B. It can be observed that the points within 2,000 km follow linearity, whereas the points with distances over 2,000 km deviate from the linearity. This deviation from linearity can also be found at the prefectural and county levels. It may be explained by the local dispersion of surnames and the fact that it was

difficult to migrate over long distances of more than 2,000 km.

CONCLUSIONS

In this study, the population structure in China was investigated through isonymy analysis at the provincial, prefectural, and county levels. A comparison of the isonymic structure between China and other countries is shown in Table 2. There are two significant features in Chinese surname, the scarcity of effective surname number, as indicated by the extremely low value of Fisher's α, and the noticeable isolation by distance, as indicated by the relatively large value of r. The scarcity of effective surname number can be partly attributed to the particular frequency distribution of Chinese surnames, as indicated by that the 100 most common surnames account for 85% of the population. Furthermore, this scarcity is also related to the low number of surnames in the Chinese language, with only 7,327 surnames among 1.28 billion people. The noticeable isolation by distance may be explained by the long history of Chinese surnames, which permit drift and the local dispersion of surnames.

The specific features of the geographic distribution of isonymy may be related to human behaviors in Chinese history, including the origin of Chinese culture, large-scale migrations, residential patterns, and ethnic distribution. Although the Chinese population has had a long time to drift, there are many migrations in Chinese history, and several regional centers of Chinese civilization have developed. Therefore, the population structure in China may be the result of the joint action of drift and migration. Drift has a main effect on surnames in general terms, as indicated by the frequency distribution of Chinese surnames and the detected isolation by distance in China. However, the effects of drift and migration vary by region. Specifically, migration predominated over drift in the middle and lower reaches of the Yangtze River, whereas drift predominated over migration in the Yellow River Basin. Demic migrations, in particular, have played a decisive role in the surname structure of the three northern provinces.

LITERATURE CITED

Baek SK, Kiet H, Kim BJ. 2007. Family name distributions: master equation approach. Phys Rev E 76:046113.

Barrai I, Rodriguez-Larralde A, Mamolini E, Manni F, Scapoli C. 2001. Elements of the surname structure of the USA. Am J Phys Anthropol 114:109–123.

Barrai I, Scapoli C, Beretta M, Nesti C, Mamolini E, Rodriguez-Larralde A. 1996. Isonymy and the genetic structure of Switzerland I. The distributions of surnames. Ann Hum Biol 23:431–455.

Branco CC, Mota-Vieira L. 2003. Population structure of Sao Miguel Island, Azores: a surname study. Hum Biol 75:929–939.

Branco CC, Mota-Vieira L. 2005. Surnames in the Azores: analysis of the isonymy structure. Hum Biol 77:37–44.

Bronberg RA, Dipierri JE, Alfaro EL. 2009. Isonymy structure of Buenos Aires City. Hum Biol 81:447–461.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785–807.

Crow JF. 1980. The estimation of inbreeding from isonym. Hum Biol 52:1–14.

Crow JF. 1983. Discussion. Hum Biol 55:383–397.

Crow JF, Mange AP. 1965. Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugen Q 12:199–203.

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodriguez-Larralde A, Barrai I. 2005. Surnames in Argentina. A population study through isonymy. Am J Phys Anthropol 128:199–209.

Dipierri J, Rodriguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, Salvatorelli G, Caramori G, De Lorenzi S, Sandri M, Carrieri A, Barrai I. 2011. A study of the population of Paraguay through isonymy. Ann Hum Genet 75:678–87.

Du RF, Yuan YD. 1995. The evolution of Chinese surnames and surname frequency in different dialect zones. Soc Sci China 16:171–184.

Du RF, Yuan YD, Juliana H, Joanna M, L.Luca Cavalli-Sforza. 1992. Chinese surnames and the genetic differences between North and South China. J Chin Ling Monogr Ser5.

Fox WR, Lasker GW. 1983. The distribution of surname frequencies. Int Stat Rev 51:81–87.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Nei M. 1973. The theory and estimation of genetic distance. In: Morton NE, editor. Genetic structure of populations. Honolulu: University Hawaii Press. p45–54.

Qian CC. 1989. Study on surnames for ethnic minorities in China. J Cent Univ Nationalities 6:13–16. (in Chinese)

Rodriguez-Larralde A, Barrai I, Nesti C, Mamolini E, Scapoli C. 1998a. Isonymy and isolation by distance in Germany. Hum Biol 70:1041–1056.

Rodriguez-Larralde A, Dipierri J, Gomez EA, Scapoli C, Mamolini E, Salvatorelli G, Lorenzi SD, Carrieri A, Barrai I. 2011. Surnames in Bolivia: a study of the population of Bolivia through isonymy. Am J Phys Anthropol 144:177–184.

Rodriguez-Larralde A, Gonzalez-Martin J, Scapoli C, Barrai I. 2003. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol 121:280–292.

Rodriguez-Larralde A, Morales J, Barrai I. 2000. Surname frequency and the isonymy structure of Venezuela. Am J Hum Biol 12:352–362.

Rodriguez-Larralde A, Scapoli C, Beretta M, Nesti C, Mamolini E, Barrai I. 1998b. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. Ann Hum Biol 25:533–540.

Scapoli C, Goebl H, Sobota S, Mamolini E, Rodriguez-Larralde A, Barrai I. 2005. Surnames and dialects in France: population structure and cultural evolution. J Theor Biol 237:75–86.

Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I. 2007. Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Pop Biol 71:37–48.

Tarskaia L, El'chinova GI, Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I. 2009. Surnames in Siberia: a study of the population of Yakutia through isonymy. Am J Phys Anthropol 138:190–198.

Yuan YD, Jin F, Zhang C. 1999. The study of the distribution of Chinese surnames and the diversity of genetic population structure in the Song Dynasty. Acta Genet Sin 26:187–197.

Yuan YD, Zhang C. 2002. Chinese surnames: community heredity and population distribution. Shanghai: East China Normal University Press (in Chinese).

Yuan YD, Zhang C, Ma Q, Yang H. 2000a. Population genetics of Chinese surnames I. Surname frequency distribution and genetic diversity in Chinese. Acta Genet Sinica 27:471–476 (in Chinese).

Yuan YD, Zhang C, Ma Q, Yang H. 2000b. Population genetics of Chinese surnames II. Inheritance stability of surnames and regional consanguinity of population. Acta Genet Sinica 27:565–572 (in Chinese).

Zei G, Guglielmino Matessi R, Siri E, Moroni A, Cavalli-Sforza L. 1983. Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. Ann Hum Genet 47:329–352.