Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 474–481
GENOMICS
www.elsevier.com/locate/ygeno

mtDNA evidence: Genetic background associated with related populations at high risk for esophageal cancer between Chaoshan and Taihang Mountain areas in China

Xiao-Yun Li [a,1], Min Su [a,*,1], Hai-Hua Huang [a], Hui Li [b], Dong-Ping Tian [a], Yu-Xia Gao [a]

[a] *Department of Pathology, Key Immunopathology Laboratory of Guangdong Province, Shantou University Medical College, Shantou, Guangdong 515031, China*
[b] *State Key Laboratory of Genetic Engineering and Center for Anthropological Studies, School of Life Sciences, Fudan University, Shanghai, Jiangsu 200433, China*

Received 4 April 2007; accepted 20 June 2007
Available online 9 August 2007

Abstract

There are three major geographic regions in China known for their high incidences of esophageal cancer (EC): the Taihang Mountain range of north-central China, the Minnan area of Fujian province, and the Chaoshan plain of Guangdong province. Historically, waves of great population migrations from north-central China through coastal Fujian to the Chaoshan plain were recorded. To study the genetic relationship among the related EC high-risk populations, we analyzed mitochondrial DNA (mtDNA) haplogroups based on 30 EC patients from Chaoshan and used control samples from the high-risk populations, including 48, 73, and 89 subjects from the Taihang, Fujian, and Chaoshan areas, respectively. The principal component of all haplogroups, correlation analysis of haplogroup frequency distributions between populations, and haplogroup D network analysis showed that compared with other Chinese populations, populations in the three studied areas are genetically related. The highest haplogroup frequency shared by all studied populations was haplogroup D, with much higher frequency in the Chaoshan area EC patients. The majority of haplogroup D individuals among the Chaoshan area EC patients belonged to subhaplogroups D4a and D5a, with the total frequency of these two haplogroups significantly higher than that in the high-risk population in the same area ($\chi^2 = 9.017$, $p < 0.01$). In conclusion, EC high-risk populations in these three areas share a similar matrilineal genetic background, and D4a and D5a might be candidate genetic markers for screening populations susceptible to EC in the Chaoshan area. Ours is the first report to show the association between mtDNA haplogroups (D4a and D5a) and esophageal cancer.

Keywords: mtDNA haplogroup; Esophageal cancer high-risk population; Genetic background; Candidate genetic markers

Esophageal cancer (EC) is one of the most common fatal cancers worldwide. There are particular geographical hot spots with high EC incidences. Most EC patients live in the so-called "esophageal cancer belt," which stretches from central China westward through Central Asia to northern Iran [1,2]. In north-central China, the Taihang Mountain area between Henan, Hebei, and Shanxi provinces is a well-known high-risk region for EC. In contrast, EC high-risk areas located in the south-eastern littoral of China, including the Chaoshan plain in Guangdong province and the southern part of Fujian province (Minnan area), are less known to the world. The Chaoshan area, next to the Minnan area, is surrounded by the Lianhua Mountains and South China Sea and is thus relatively geographically isolated from the inner part of China. Our previous study [3] reported an annual average crude incidence rate for EC to be 71.07/100,000 from 1995 to 2003 on Nanao Island, which represents a relatively isolated district within the Chaoshan area.

We and others have reported familial aggregation of EC patients and increased EC risk in family members in north-central China and the Chaoshan area [4–7]. These findings support the potentially important role of genetic susceptibility in EC etiology in high-risk populations. Ascertaining genetic

Abbreviations: EC, esophageal cancer; mtDNA, mitochondrial DNA; HVS-I, hypervariable segment I; PC, principal component.

* Corresponding author. Fax: +86 754 8900429.
E-mail address: minsu@stu.edu.cn (M. Su).

[1] These authors are co-first authors of this article.

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.06.006

background of EC patients will facilitate the clarification of molecular genetic mechanisms in esophageal carcinogenesis, the risk evaluation of individuals from high-risk populations, and the establishment of effective screening to identify EC-susceptible individuals and preventive measures.

According to historical records [8], Han inhabitants of north-central China (Henan and Shanxi Hans) continuously migrated into the Chaoshan area via Fujian due to warfare and famines and gradually became the predominant inhabitants of the Chaoshan area. Although the Taihang Mountain, Fujian, and Chaoshan EC high-risk areas are geographically distant, we hypothesize that these EC high-risk populations could share common genetic traits. mtDNA is characterized by a strictly maternal mode of inheritance, absence of recombination, rapid rate of mutation, and high level of population-specific polymorphisms. These make mtDNA widely applicable for studying evolutionary relationships among human ethnic groups [9–11]. Because mtDNA is strictly maternally inherited, the mtDNA sequence has evolved by the sequential accumulation of base substitutions along radiating maternal lineages [12]. Therefore, the characteristics of mtDNA enable researchers to trace related lineages back through time, highlighting the maternal ancestry of a population, without the confounding effects of biparental inheritance and recombination inherent in nuclear DNA [13]. The common methods used in these studies are restriction enzyme analysis of mtDNA coding regions in conjunction with sequence analysis of the mtDNA D-loop region, which has apparently evolved several times faster than other regions [9–11,14–16]. The combined information from coding region and D-loop region contributes to the phylogenetic analysis of mtDNA.

In this study, we investigated the mtDNA polymorphism distribution among populations from the Taihang Mountain, Fujian, and Chaoshan EC high-risk areas. Our data suggested a close genetic relationship between the Chaoshan and the Taihang Mountain EC high-risk populations. Compared with the other mtDNA haplogroups, the D4a and D5a haplogroups were found to be closely associated with EC in the Chaoshan area.

Table 1
mtDNA haplogroup distribution among all populations in three studied areas

Haplogroup	Taihang Mountain EC high-risk population (%) (n=48)	Fujian EC high-risk population (%) (n=73)	Chaoshan EC high-risk population (%) (n=89)	Chaoshan area EC patients (%) (n=30)
A	6.25	8.22	7.87	0.00
B4	0.00	2.74	0.00	0.00
B4a	4.17	5.48	3.37	0.00
B4b1	2.08	2.74	3.37	0.00
B4c1	0.00	1.37	5.62	6.67
B5[a]	0.00	0.00	2.25	0.00
B5a	2.08	0.00	2.25	0.00
B5b	10.42	1.37	1.12	6.67
C	6.25	5.48	3.37	3.33
D[a]	16.67	10.96	5.62	3.33
D4a	0.00	4.11	6.74	20.00
D4b	0.00	1.37	1.12	0.00
D5	2.08	2.74	2.25	0.00
D5a	6.25	4.11	3.37	13.33
F[a]	2.08	0.00	3.37	0.00
F1a	6.25	10.96	3.37	3.33
F1b	2.08	1.37	2.25	0.00
F1	0.00	1.37	0.00	0.00
F1c	2.08	0.00	0.00	0.00
F2a	0.00	1.37	1.12	0.00
F2[a]	0.00	0.00	0.00	0.00
G[a]	10.42	1.37	5.62	6.67
M[a]	0.00	1.37	0.00	0.00
M7[a]	0.00	0.00	1.12	0.00
M7b	0.00	4.11	5.62	6.67
M7b1	0.00	6.85	2.25	0.00
M7b2	2.08	0.00	1.12	3.33
M7c	0.00	1.37	0.00	6.67
M8a	6.25	4.11	4.49	10.00
M9a	2.08	2.74	3.37	0.00
N[a]	0.00	1.37	1.12	0.00
N9a	4.17	4.11	3.37	0.00
M10	2.08	0.00	0.00	0.00
R9a	0.00	2.74	4.49	0.00
R9b	0.00	1.37	1.12	0.00
R11	2.08	1.37	1.12	0.00
Y	2.08	0.00	3.37	0.00
Z	0.00	1.37	3.37	10.00

[a] Haplogroup status cannot be further specified into subhaplogroups.

Results

mtDNA haplogroup frequencies

Our primary goal was to investigate the genetic affinity between the EC high-risk populations and the potential association between mtDNA haplogroup variations and EC occurrence. mtDNA haplogroups identified in 30 unrelated EC patients from the Chaoshan area and 210 unrelated EC high-risk individuals from the Taihang Mountain, Fujian, and Chaoshan areas are presented in Table 1. A total of 38 haplogroups were studied. We estimated haplogroup diversity of each population (Table 2). Compared with the EC high-risk populations from the Taihang Mountain, Fujian, and Chaoshan areas, the mtDNA haplogroups in EC patients from the Chaoshan area was the least diverse. According to previous reports [17–20], A, C, D, G, M8a, Y, and Z variations accounted for much higher overall frequencies in northern Hans than in their southern counterparts and thus were northern Han dominant haplogroups, whereas B, F, R9a, R9b, N9a, M7b, and M7b1 constituted the southern native dominant haplogroups and showed much higher overall frequencies in southern Hans than in northern Hans. Thus, we further summarized the haplogroup distribution in the Chaoshan area EC patients and EC high-risk populations from the Taihang Mountain, Fujian, and Chaoshan areas according to southern and northern haplogroup distributions among the general Chinese population (Table 3).

Table 3 shows, in the Taihang Mountain area EC high-risk populations, that the overall frequency of the northern Han dominant haplogroups (A, C, D, G, M8a, Y, and Z) was 56.25%. In striking similarity to this, the northern Han dominant haplogroups constituted 66.67% of the Chaoshan area EC

Table 2
Haplogroup diversity in all populations in three studied areas

	Haplogroup diversity
Taihang Mountain EC high-risk population	0.94236
Fujian EC high-risk population	0.95699
Chaoshan EC high-risk population	0.96859
Chaoshan area EC patients	0.92643

patient haplogroups. However, in the Fujian and Chaoshan EC high-risk populations, the overall frequencies of northern Han dominant haplogroups were 43.84 and 47.19%, respectively, while those of southern native dominant haplogroups were 47.95 and 44.95%, respectively. All of the seven southern native dominant haplogroups, B, F, R9a, R9b, N9a, M7b, and M7b1, were present in the Fujian and Chaoshan area EC high-risk populations, whereas only three of them were found in the Taihang Mountain EC high-risk population and Chaoshan area EC patients. Of particular note, the highest frequency of haplogroups (19.10–36.67%) shared by populations of all three studied areas was haplogroup D (one of the northern Han dominant haplogroups). Haplogroup D accounted for 36.67% of the total haplogroups among the Chaoshan area EC patients, followed by haplogroups Z, M8a, M7c, M7b, G, B5b, B4c1, M7b2, F1a, and C. Haplogroup D in Chaoshan area EC patients comprised mainly the subhaplogroups D4a and D5a (20 and 13.33%, respectively) (Table 1), and their total percentage was significantly higher than that of the high-risk population in the same area ($\chi^2=9.017$, $p<0.01$).

PC analysis and correlation analysis of mtDNA data

All of the mtDNA haplogroup profiles identified in the populations of all three studied areas (Table 1) plus 23 additional Chinese Han population data items were treated as input vectors for PC analysis. The PC map (Fig. 1) was constructed using the first three principal components, which together accounted for 68% of the total variations. The PC map displays the genetic divergence of populations of all three studied areas from other Chinese populations. Those from the three studied areas clustered together in the uppermost part of the PC map and were separated from other populations by the second principal component (PC2). Northern Han Chinese, labeled using white squares in the PC map, clustered together, and southern Han Chinese, labeled using black squares, formed two clusters with one including Han Chinese from Zhejiang province, Hunan province, and Zhanjiang of Guangdong province and another comprising the rest of the southern Han Chinese.

We conducted a correlation analysis of haplogroup frequency distributions between populations (Table 4). The results showed that haplogroup frequency distribution in the Chaoshan EC high-risk population was significantly positively correlated to those in the Taihang Mountain and Fujian high-risk populations and Chaoshan area EC patients, and correlation coefficients were 0.446 ($p<0.01$), 0.610 ($p<0.01$), and 0.466 ($p<0.01$), respectively, higher than those between the Chaoshan high-risk population and the 23 additional Chinese populations mentioned above. Another significantly positive correlation with $p<0.01$ was found between the Chaoshan EC high-risk population and the Xi'an population (correlation coefficient 0.404). The Xi'an population belongs to the ancient northern Hans in China.

Network analyses for haplogroup D

Fig. 2 displays the haplogroup D network. According to the phylogenetic tree of East Asian mtDNAs [17], the "OUT" was defined by site 16362 of the hypervariable segment I (HVS-I) and haplogroup D was defined under the OUT. The ancestral node leading to the OUT was represented by two Taihang Mountain EC high-risk individuals, three Chaoshan high-risk individuals, two Liaoning Han individuals, and one Gansu Han individual. All of the other haplogroup D individuals came from this ancestral node. The ancestral node was connected to two one-step neighbors, with one neighbor representing one Taihang Mountain high-risk individual above it and another representing one Taihang Mountain and two Fujian high-risk individuals below. The remaining nodes representing Chaoshan, Fujian, and Taihang Mountain high-risk individuals and Chaoshan area EC patients were generated largely by these two one-step neighbors and thus were clustered mainly in two areas above or below the ancestral node (these two areas are circled in Fig. 2). The ancestral node-derived haplogroup D individuals from the northern Han Chinese (from Liaoning, Inner Mongolia, and Gansu provinces) and southern natives (including Hmong,

Table 3
The distribution of northern Han-dominating haplogroups and southern native-dominating haplogroups among all populations in three studied areas

Haplogroup	Taihang Mountain EC high-risk population (%) ($n=48$)	Fujian EC high-risk population (%) ($n=73$)	Chaoshan area EC high-risk population (%) ($n=89$)	Chaoshan area EC patients (%) ($n=30$)
A	6.25	8.22	7.87	0.00
C	6.25	5.48	3.37	3.33
Total haplogroup D	25.00	23.29	19.10	36.67
G*	10.42	1.37	5.62	6.67
M8a	6.25	4.11	4.49	10.00
Y	2.08	0.00	3.37	0.00
Z	0.00	1.37	3.37	10.00
Total northern Han dominant haplogroups	56.25	43.84	47.19	66.67
Total haplogroup B	18.75	13.70	17.98	13.33
Total haplogroup F	12.50	15.07	10.11	3.33
N9a	4.17	4.11	3.37	0.00
R9a	0.00	2.74	4.49	0.00
R9b	0.00	1.37	1.12	0.00
M7b	0.00	4.11	5.62	6.67
M7b1	0.00	6.85	2.25	0.00
Total southern native dominant haplogroups	35.42	47.95	44.95	23.34

Northern Han dominating haplogroups include haplogroups A, C, D, G*, M8a, Y, and Z; southern native dominating haplogroups include haplogroups B, F, N9a,R9a, R9b, M7b, and M7b1.

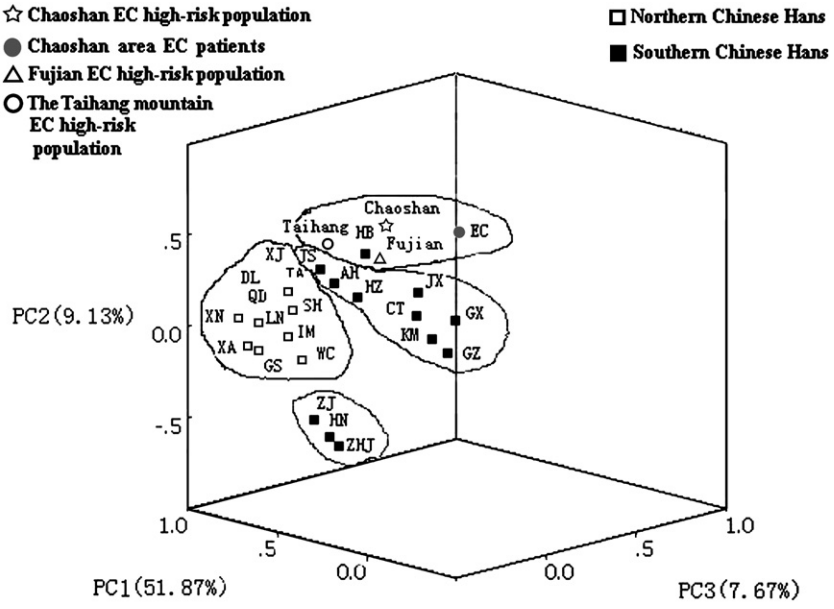

Fig. 1. The principal component analysis of mtDNA haplogroups among all populations in three studied areas and the Han Chinese populations derived from the major regions in China. The Chaoshan area EC patients and three EC high-risk populations clustered together; the northern Han Chinese populations form another cluster. However, the southern Han Chinese populations are divided into two clusters. Abbreviation of samples: XJ, Xinjiang; DL, Dlian; TA, Taian; QD, Qindao; SH, Shanghai; XN, Xining; LN, Liaoning; IM, Inner Mongolia; XA, Xi'an; GS, Gansu; WC, Weicheng; ZJ, Zhejiang; HN, Hunan; ZHJ, Zhanjiang; JS, Jiangsu; AH, Anhui; HZ, Huize; JX, Jiangxi; CT, Changting; KM, Kunming; GX, Guangxi; GZ, Guangzhou; HB, Hubei.

Dai, and Mien ethnic groups) were located outside these two areas.

Discussion

Daic ethnic groups formed the earliest settlement of the Guangdong district of southern China in modern history. According to historical records, before 216 BC the main natives living in the Chaoshan littoral were the Minyue population, one branch of the Daic ethnic groups [8]. The north-to-south strategic expansion started by Emperor Qin Shi Huang initiated a precedent for massive southward migrations of central China Hans beginning in 216 BC. Gradually they became the major population in the Chaoshan area, called Helao or Fulao, as they came mostly from Henan and Shanxi via Fujian, with well-maintained language and customs from north-central China during the past 2000 years [8]. Geographic isolation plus difficulty in traveling in the past made the Helao or Fulao become a relatively closed population. Therefore, we put forward a hypothesis that the Chaoshan littoral is an EC high-risk area due to the common genetic background shared between the Chaoshan population and those in north-central China. This study provides genetic evidence supporting this hypothesis.

The genetic affinity among the Taihang Mountain, Fujian, and Chaoshan EC high-risk populations and Chaoshan area EC patients

Chinese Han populations are divided into two genetically differentiated groups, northern Han and southern Han, by the Yangtze River [21–24]. Historically the Hans originated from the ancient Huaxia tribes of northern China. A study by Wen et al. [20] demonstrated that Han expansion toward the south was primarily because of massive southward movements of northern Hans, not simply because of Han cultural southward diffusion and assimilation. During the expansion process, gene flow between Hans and southern natives occurred to a certain extent.

In the present study, we analyzed the distribution of northern Han dominant haplogroups and southern native dominant haplogroups among all populations of the three studied areas (Table 3). Similar to the Taihang Mountain EC high-risk population, the Chaoshan area EC patients have northern Han dominant haplogroups as their major haplogroups. Compared with the Fujian and Chaoshan high-risk populations, the Chaoshan area EC patients had fewer southern native dominant haplogroups. Thus, the Chaoshan area EC patients appeared to have the mtDNA haplogroup distribution of northern Hans. In contrast, the Chaoshan high-risk population possessed maternal genetic structures characterized by similar proportions of northern and southern haplogroups, indicating exposure to more maternal influence from southern natives compared with Chaoshan area EC patients. The highest frequency of haplogroups shared by all populations of the three studied areas was haplogroup D. The frequencies of haplogroup D tended to decrease from north to south, suggesting that the high frequency of haplogroup D might be characteristic of northern Hans [18].

We further used PC analysis, correlation analysis, and network analysis to compare the genetic relationship among all populations of the three studied areas. PC analysis is a common statistical method to study the resemblance of haplogroup distribution among different populations. The closer the positions of two populations on the PC map, the more intimate the genetic relationship between the two. Compared with other

Table 4
Correlation analysis on haplogroup frequency distributions between the Chaoshan EC high-risk population and other populations (numbers represent correlation coefficients)

	Chaoshan EC high-risk population
Taihang Mountain EC high-risk population	0.446**
Fujian EC high-risk population	0.610**
Chaoshan area EC patients	0.466**
Gansu	0.237
Liaoning	0.263
Dalian	0.269
Inner Mongolia	0.226
Xining	0.282
Qindao	0.338*
Taian	0.278
Xi'an	0.404**
Xinjiang	0.386*
Anhui	0.349*
Changting	0.265
Guangxi	0.287
Hunan	0.241
Jiangsu	0.261
Jiangxi	0.107
Shanghai	0.207
Weicheng	0.264
Huize	0.330*
Zhejiang	0.322*
Guangzhou	0.121
Zhanjiang	0.130
Hubei	0.390*
Kunming	0.112

* Correlation is significant at the 0.05 level.
** Correlation is significant at the 0.01 level

Chinese Han populations, all populations of the three studied areas clustered more closely in the PC map (Fig. 1), and correlation analysis of haplogroup frequency distribution documented that the matrilineal genetic structure of the Chaoshan high-risk population was more similar to those in the other two high-risk populations and EC patients. Network analysis for haplogroup D (Fig. 2) suggested that the matrilineal ancestral lineage of haplogroup D individuals existed in the Taihang Mountain and Chaoshan high-risk individuals who constituted the ancestral node, and the rest of the haplogroup D individuals from the three studied areas were generated largely by two one-step neighbors containing the Taihang Mountain high-risk individuals. Therefore, we think that there is an obvious matrilineal genetic affinity between the geographically separated EC high-risk populations in China. Network analysis documented that haplogroup D individuals of the Fujian and Chaoshan high-risk populations and Chaoshan area EC patients originated mainly from the Taihang Mountain EC high-risk area, which is consistent with the migration history of the Chaoshan population [8]. Part of the Taihang Mountain people migrated to the Chaoshan area following a stay in Fujian. As was recorded in pedigrees and ancient inscriptions, since the Northern Song Dynasty, large numbers of southern Fujian people, especially from Quanzhou and Putian, made settlements toward Chaoshan in batches and soon spread all over the Chaoshan area [25].

The main genetic background resulting in high incidence of EC shared by the Taihang Mountain and Chaoshan populations

Recently, the role of mtDNA variation in the occurrence of tumors has received increasing attention, and people have detected mtDNA somatic mutations in many cancers, e.g., EC [26–30], gastric cancer [27], and breast cancer [31]. However, most studies focused on mtDNA somatic mutation; only a few have investigated the maternal genetic background in cancer formation. To explore whether EC shows strong mtDNA haplogroup bias, our samples in this study contain 30 Chaoshan area EC patients. Because the maternal genetic information in each individual represents that of all family members with maternal consanguinity, information provided by 30 patients was sufficient for genetic background analysis.

We found that the haplogroup types identified in Chaoshan area EC patients (13 haplogroups, see Table 1) were evidently fewer than those in the three EC high-risk populations (21 in the Taihang Mountain area, 29 in Fujian, and 31 in Chaoshan, see Table 1), and haplogroup diversity of EC patients was also the lowest (Table 2), indicating that the occurrence of esophageal cancer is closely associated with certain mtDNA haplogroups, rather than a random event.

The evidence provided in the present study does not support the obvious association of southern native dominant haplogroups with the high prevalence of EC in Chaoshan. Compared with the other three populations, the southern native dominant haplogroups identified in Chaoshan area EC patients showed the fewest types and the lowest overall frequency (23.34%), which is by far lower than the overall frequency (66.67%) of northern Han dominant haplogroups (Table 3), and none of the southern native dominant haplogroups possessed evidently higher frequency in Chaoshan area EC patients. In contrast, the northern Han dominant haplogroups, although fewer in type, showed much higher overall frequency in Chaoshan area EC patients compared with the three high-risk populations. This was due mainly to the obviously higher haplogroup D frequency (36.67%) in EC patients. This suggested that haplogroup D might be one of the main candidate genetic factors associated with the high incidence of EC shared by the Taihang Mountain and Chaoshan EC high-risk areas.

We thus put forward a "bottleneck effect" hypothesis, namely that esophageal cancers do not occur randomly in the Chaoshan EC high-risk population, but are closely associated with certain types of northern Han dominant haplogroups, especially haplogroup D, which makes the haplogroup types of EC patients less diverse than those of EC high-risk population. The majority of haplogroup D individuals in Chaoshan area EC patients belonged to the D4a and D5a subhaplogroups, with the total frequency of these two haplogroups significantly higher than that in high-risk population in the same area. Therefore, haplogroups D, especially D4a and D5a, are associated with the high risk of EC in Chaoshan and may represent candidate genetic background markers for screening individuals susceptible to EC. We find only scant literature on maternal genetic background and cancer formation. Haplogroup D was documented as likely to play a genetic role in predisposing to endometrial cancer in southwest China [32], haplogroup N is considered a risk factor

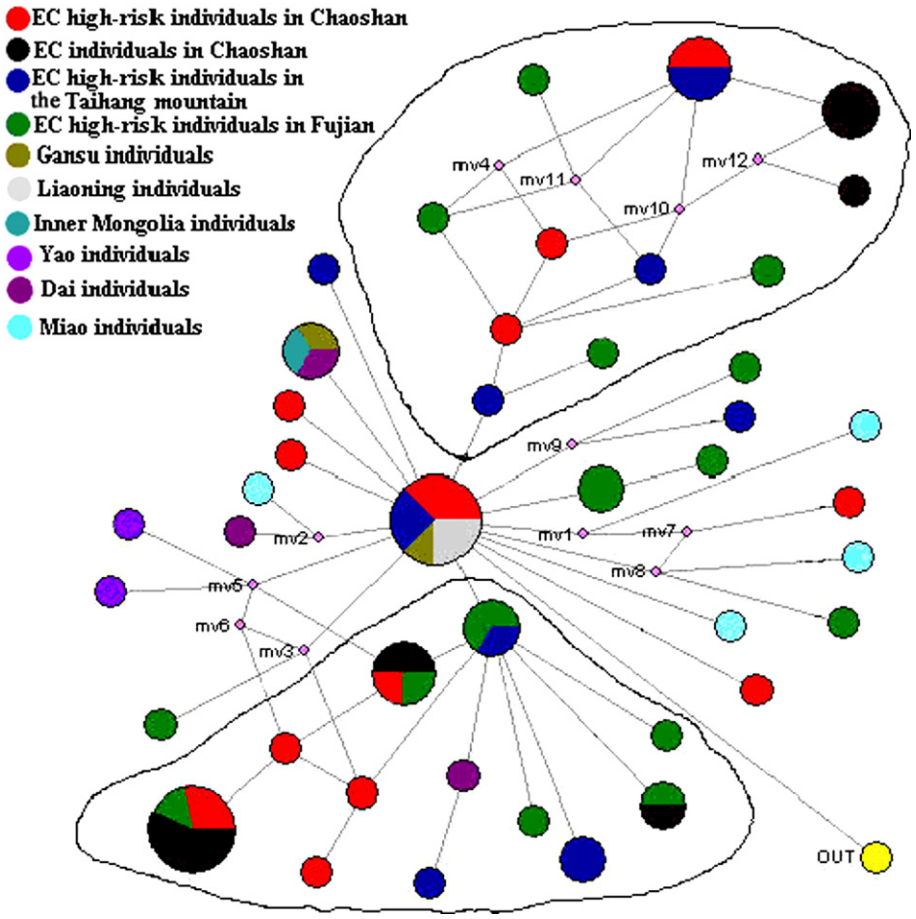

Fig. 2. Network analysis of haplogroup D. The ancestral node linked to "OUT" represents the ancestral origin of the haplogroup D population. The other nodes are derived from the ancestral node. Individuals with the same HVS-I mutations are clustered into the same node and the node area is proportional to the number of individuals in such node. Among the haplogroup D individuals of all populations in three studied areas, two Taihang Mountain EC high-risk individuals and three Chaoshan EC high-risk individuals are clustered into the ancestral node, and the other haplogroup D individuals are largely clustered into the regions above or below the ancestral node marked with a circle.

for breast cancer and EC [33], and haplogroup U is associated with increased risk of prostate cancer and renal cancer in white North American individuals [34]. Therefore, this is the first report of maternal genetic background documenting the involvement of haplogroups D4a and D5a in carcinogenesis of esophageal cancer. To explore this aspect further, a large-scale study on EC patients and high-risk populations is warranted.

In summary, the matrilineal genetic structure study of the Taihang Mountain, Fujian, and Chaoshan EC high-risk populations and Chaoshan area EC patients showed that our hypothesis on the genetic background of the Chaoshan EC high-risk population is relatively creditable. Chaoshan and the Taihang Mountain EC high-risk populations might share similar matrilineal genetic backgrounds, and D4a and D5a might be candidate genetic background markers for screening populations susceptible to EC in the Chaoshan area.

Materials and methods

Samples

Blood samples of 240 unrelated adult male Han individuals from three EC high-risk areas in China were collected from 2002 to 2004. Informed consent was obtained from each individual. These included: (1) Chaoshan populations—89 individuals without EC and 30 patients with pathologically confirmed EC from the Chaoshan EC high-risk area; (2) a Fujian EC high-risk population—73 individuals without EC from the Fujian EC high-risk area (Minnan area); and (3) a Taihang Mountain EC high-risk population—48 individuals without EC from the Taihang Mountain EC high-risk area. Maternal genetic background information from an individual represents that of all family members with maternal consanguinity, and thus the sample size required in this kind of population genetics study is far smaller than in other types of studies. A sample size of more than 24 individuals can provide complete information, and a sample size of more than 30 individuals qualifies as a big sample [35]. Because most mtDNA mutations reflect mainly evolutionary information of human ethnic groups from long ago, and were selectively neutral or near neutral [12,36], they would not likely be influenced by age. Thus, we have taken no account of the age of the subjects.

Typing of mtDNA polymorphisms

Genomic DNA was extracted from whole blood by standard phenol/chloroform methods. The primer pair L15974 (TCCACCATTAGCACCCAAAG) and H16488 (AGGAACCAGATGTCGGATACAG) was designed to amplify HVS-I from the D-loop region of mtDNA in a 15-μl final volume. PCR was performed using GeneAmp PCR System 2700 (Applied Biosystems, Foster City, CA, USA) under the following conditions: an initial step of 94°C for 3 min followed by 30 cycles of 94°C for 30 s, 62°C for 40 s, and 72°C for 50 s and a final elongation step of 72°C for 5 min. After purification and treatment with shrimp alkaline

Table 5
PCR primers, restriction enzymes, and the pattern of polymorphism for the six mtDNA coding region variations

Locus	PCR primer pair	Restriction endonuclease	Wild type	Mutation type
9-bp deletion	L8215: 5′-ACAGTTTCATGCCCATCGTC3′- H8297: 5′-ATGCTAAGTTAGCTTTACAG-3′	/	No deletion	Deletion
9824	L9766: 5′-CATTTCCGACGGCATCTAC-3′ H10165:5′-GGTGGATTTTTCTATGTAGCC-3′	HinfI	Cut (−)	Cut (+)
4831	L4576:5′-AACATGCTAGCTTTTATTCCAG-3′ H4841: 5′-AGAAGAAGCAGTCCGGATGT-3′	HhaI	Cut (−)	Cut (+)
5176	L5126:5′-AACTTAAACTCCAGCACCAC-3′ H5464:5′-TAGGTAGGAGTAGCGTGGTA-3′	AluI	Cut (+)	Cut (−)
10310	L10310:5′-TGAGCCCTACAAACAACTAACAT-3′ H10310:5′-ATACCAATTCGGTTCAGTCTAATC-3′	NlaIII	Cut (−)	Cut (+)
10397	L10296: 5′-AAACAACTAACCTGCCACTA-3′ H10498: 5′-GAAGTGAGATGGTAAATGCT-3′	AluI	Cut (−)	Cut (+)

phosphatase (SAP) and exonuclease I, the PCR products were subjected to direct DNA sequencing using the Big-Dye Terminator v3.1 cycle sequencing kit and ABI Prism 3100 genetic analyzer (Applied Biosystems).

In addition, based on the phylogenetic tree of East Asian mtDNAs [17], six coding region variations, including the COII-tRNA^{Lys} 9-bp deletion, 9824HinfI, 4831HhaI, 5176AluI, 10310NlaIII, and 10397AluI, were genotyped using the PCR-based restriction fragment length polymorphism method. PCR primers, restriction enzymes, and the pattern of polymorphism for these six variations are listed in Table 5. All primers were synthesized by Sangon Co. Ltd., Shanghai, China. Restriction endonuclease, SAP, and restriction exonuclease were purchased from New England Biolabs (Beverly, MA, USA).

Data analysis

The sequence of HVS-I was aligned and analyzed with DNA STAR and BioEdit software and compared with the revised Cambridge Reference Sequence [37]. Based on the phylogeny of East Asian mtDNAs [17], both the HVS-I motif and the coding region variations were used to infer mtDNA haplogroups. The mtDNA haplogroup can be understood as a monophyletic clade in the rooted mtDNA phylogenetic tree, i.e., a group of haplotypes that comprises all descendants of their most recent common ancestor, as inferred from the shared mutations [38]. For example, East Asian mtDNAs belong to two superhaplogroups, M and N; M is subdivided into five major subhaplogroups D, G, M8, M7, and M9, and D further encompasses D4 and D5 [17].

The genetic relationship among all populations in the three studied areas was investigated by PC analysis, which was conducted using SPSS11.5 software (SPSS, Inc.) based on the frequencies of all mtDNA haplogroups identified in this study (see Table 1). In addition, previously published mtDNA haplogroup data from 23 Chinese Han populations [18] were also included in the PC analysis for comparison. These Han populations are representative of all major regions in China. PC analysis is the classical technique to reduce the dimensionality of the data set by transforming a new set of variables to summarize the features of the data.

The highest haplogroup frequency shared by all populations in the three studied areas was haplogroup D. The network for haplogroup D was further constructed using Network 4.1.0.8 software (www.fluxus-engineering.com) based on all of haplogroup D (including D*, D4a, D4b, D5, D5a) individual HVS-I sequences for analyzing the origin of D individuals among the EC high-risk areas. In addition, haplogroup D data from Han Chinese populations in the northern part of China (Liaoning, Inner Mongolia, and Gansu provinces) and the southern China non-Han Chinese populations, including Hmong nationality, Dai nationality, and Mien nationality (provided by the State Key Laboratory of Genetic Engineering and Center for Anthropological Studies, School of Life Sciences, Fudan University), were included for comparison. In the network map, individuals with the same mutations of HVS-I were present in the same node and one node can generate other nodes below due to gradual mtDNA mutations.

Moreover, correlation analysis and χ^2 test were performed using SPSS11.5 software (SPSS, Inc.). Haplogroup diversity was estimated based on the formula

$$h = (n/n-1)\left(1-\sum_i p_i^2\right),$$

where p_i is the sample frequency of the ith haplogroup and n is the number of individuals in the sample [39].

Acknowledgments

This study was sponsored by the China Natural Science Fund (Certificate 30210103904) and the Grand Natural Science Fund of Guangdong Province in China (Certificate A1080203).

References

[1] D.M. Parkin, E. Laara, C.S. Muir, Estimates of the worldwide frequency of sixteen major cancers in 1980, Int. J. Cancer 41 (1988) 184–197.
[2] D. Schottenfeld, Epidemiology of cancer of the esophagus, Semin. Oncol. 11 (1984) 92–100.
[3] M. Su, M. Liu, D.P. Tian, X.Y. Li, H.L. Yang, H.H. Huang, H.F. Yan, C.Q. Zou, Epidemiological investigations of the morbidity rate of malignant tumors and their diet habitats among residents of Nanao island in South China Sea, J. Environ. Occup. Med. (China) 22 (2005) 312–316.
[4] N. Hu, S.M. Dawsey, M. Wu, G.E. Bonney, L.J. He, X.Y. Han, M. Fu, P.R. Taylor, Familial aggregation of oesophageal cancer in Yangcheng County, Shanxi Province, China, Int. J. Epidemiol. 21 (1992) 877–882.
[5] J. Chang-Claude, H. Becher, M. Blettner, S. Qiu, G. Yang, J. Wahrendorf, Familial aggregation of oesophageal cancer in a high incidence area in China, Int. J. Epidemiol. 26 (1997) 1159–1165.
[6] N. Hu, S.D. Dawsey, M. Wu, P.R. Taylor, Family history of esophageal cancer in Shanxi Province, China, Eur. J. Cancer 27 (1991) 1336.
[7] Y.P. Wang, X.Y. Han, W. Su, Y.L. Wang, Y.W. Zhu, T. Sasaba, K. Nakachi, Y. Hoshiyama, Y. Tagashira, Esophageal cancer in Shanxi Province, People's Republic of China: a case-control study in high and moderate risk areas, Cancer Causes Control 3 (1993) 107–113.
[8] M. Su, S.M. Lu, D.P. Tian, H. Zhao, X.Y. Li, D.R. Li, Z.C. Zheng, Relationship between ABO blood groups and carcinoma of esophagus and cardia in Chaoshan inhabitants of China, World J. Gastroenterol. 7 (2001) 619–657.
[9] S. Horai, R. Kondo, Y. Nakagawa-Hattori, S. Hayashi, S. Sonoda, K. Tajima, Peopling of the Americas, founded by four major lineages of D.NA. mitochondrial, Mol. Biol. Evol. 10 (1993) 23–47.
[10] S. Horai, K. Murayama, K. Hayasaka, S. Matsubayashi, Y. Hattori, G. Fucharoen, S. Harihara, K.S. Park, K. Omoto, I.H. Pan, mtDNA polymorphism in East Asian population, with special reference to the peopling of Japan, Am. J. Hum. Genet. 59 (1996) 579–590.
[11] D.A. Merriwether, A.G. Clark, S.W. Ballinger, T.G. Schurr, H. Soodyall, T. Jenkins, S.T. Sherry, D.C. Wallace, The structure of human mitochondrial DNA variation, J. Mol. Evol. 33 (1991) 543–555.
[12] D.C. Wallace, M.D. Brown, M.T. Lott, Mitochondrial DNA variation in human evolution and disease, Gene 238 (1999) 211–230.
[13] B. Pakendorf, M. Stoneking, Mitochondrial DNA and human evolution, Annu. Rev. Genomics Hum. Genet. 6 (2005) 165–183.

[14] L. Vigilant, R. Pennington, H. Harpending, T.D. Kocher, A.C. Wilson, Mitochondrial DNA sequences in single hairs from a southern African population, Proc. Natl. Acad. Sci. USA 86 (1989) 9350–9354.
[15] A. Torroni, Y.S. Chen, O. Semino, A.S. Santachiara-Beneceretti, C.R. Scott, M.T. Lott, M. Winter, D.C. Wallace, mtDNA and Y chromosome polymorphisms in four Native American populations from Southern Mexico, Am. J. Hum. Genet. 54 (1994) 303–318.
[16] O. Rickards, C. Martínez-Labarga, J.K. Lum, G.F. De Stefano, R.L. Cann, mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations, Am. J. Hum. Genet. 65 (1999) 519–530.
[17] T. Kivisild, H.V. Tolk, J.R. Parik, Y.M. Wang, S.S. Papiha, H.J. Bandel, V. Richard, The emerging limbs and twigs of the East Asian mtDNA tree, Mol. Biol. Evol. 19 (2002) 1737–1751.
[18] Y.G. Yao, Q.P. Kong, H.J. Bandelt, T. Kivisild, Y.P. Zhang, Phylogeographic differentiation of mitochondrial DNA in Han Chinese, Am. J. Hum. Genet. 70 (2002) 635–651.
[19] D.C. Wallace, M.D. Brown, M.T. Lott, Mitochondrial DNA variation in human evolution and disease, Gene 238 (1999) 211–230.
[20] B. Wen, H. Li, D. Lu, X.F. Song, F. Zhang, Y.G. He, Genetic evidence supports demic diffusion of Han culture, Nature 431 (2004) 302–305.
[21] T.M. Zhao, T.D. Lee, Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation, Hum. Genet. 83 (1989) 101–110.
[22] R.F. Du, C.J. Xiao, L.L. Cavalli-Sforza, Genetic distances calculated on gene frequencies of 38 loci, Sci. China Ser. C 40 (1997) 613.
[23] J.Y. Chu, W. Huang, S.Q. Kuang, J.M. Wang, J.J. Xu, Z.T. Chu, Genetic relationship of populations in China, Proc. Natl. Acad. Sci. USA 95 (1998) 11763–11768.
[24] C.J. Xiao, L.L. Cavalli-Sforza, E. Minch, R.F. Du, Principal component analysis of gene frequencies of Chinese populations, Sci. China Ser. C 43 (2000) 472–481.
[25] C.G. Xie, The transplantation and spreading of Min-nan Culture of Song dynasty in Chao-shan areas, J. Hanshan Teachers Coll. (China) 24 (2003) 1–9.
[26] D.J. Tan, J. Chang, L.L. Liu, R.K. Bai, Y.F. Wang, K.T. Yeh, L.J. Wong, Significance of somatic mutations and content alteration of mitochondrial DNA in esophageal cancer, BMC Cancer 8 (2006) 93.
[27] K. Kose, T. Hiyama, S. Tanaka, M. Yoshihara, W. Yasui, K. Chayama, Somatic mutations of mitochondrial DNA in digestive tract cancers, J. Gastroenterol. Hepatol. 20 (2005) 1679–1684.
[28] H. Kumimoto, Y. Yamane, Y. Nishimoto, H. Fukami, M. Shinoda, S. Hatooka, K. Ishizaki, Frequent somatic mutations of mitochondrial DNA in esophageal squamous cell carcinoma, Int. J. Cancer 108 (2004) 228–231.
[29] K. Hibi, H. Nakayama, T. Yamazaki, T. Takase, M. Taguchi, Y. Kasai, K. Ito, S. Akiyama, A. Nakao, Mitochondrial DNA alteration in esophageal cancer, Int. J. Cancer 92 (2001) 319–321.
[30] C.C. Abnet, K. Huppi, A. Carrera, D. Armistead, K. McKenney, N. Hu, Z.Z. Tang, P.R. Taylor, S.M. Dawsey, Control region mutations and the 'common deletion' are frequent in the mitochondrial DNA of patients with esophageal squamous cell carcinoma, BMC Cancer 4 (2004) 30.
[31] Paola Parrella, Yan Xiao, Makiko Fliss, et al., Detection of mitochondrial DNA mutations in primary breast cancer and fine-needle aspirates, 1, Cancer Res. 61 (2001) 7623–7626.
[32] L. Xu, Y. Hu, B. Chen, W. Tang, X. Han, H. Yu, C. Xiao, Mitochondrial polymorphisms as risk factors for endometrial cancer in southwest China, Int. J. Gynecol. Cancer 16 (2006) 1661–1667.
[33] K. Darvishi, S. Sharma, A.K. Bhat, E. Rai, R.N. Bamezai, Mitochondrial DNA G10398A polymorphism imparts maternal haplogroup N a risk for breast and esophageal cancer, Cancer Lett. 249 (2007) 249–255.
[34] L.M. Booker, G.M. Habermacher, B.C. Jessie, Q.C. Sun, A.K. Baumann, M. Amin, S.D. Lim, C. Fernandez-Golarz, R.H. Lyles, M.D. Brown, F.F. Marshall, J.A. Petros, North American white mitochondrial haplogroups in prostate and renal cancer, J. Urol. 175 (2006) 468–472.
[35] L.L. Cavalli-Sforza, P. Menozzi, A. Piazza, The History and Geography of Human Genes, Princeton Univ. Press, Princeton, NJ, 1994.
[36] D.C. Wallace, Mitochondrial DNA sequence variation in human evolution and disease, Proc. Natl. Acad. Sci. USA 99 (1994) 8739–8746.
[37] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human DNA mitochondrial, Nat. Genet. 23 (1999) 147.
[38] A. Torroni, M. Riciiards, V. Macaulay, P. Forster, R. Villems, S. Norby, M.L. Savontaus, K. Huoponen, R. Scozzari, H.J. Bandelt, mtDNA haplogroups and frequency patterns in Europe, Am. J. Hum. Genet. 66 (2000) 1173–1177.
[39] M. Nei, Molecular Evolutionary Genetics, Columbia Univ. Press, New York, 1987.