PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2007) **362**, 987–995
doi:10.1098/rstb.2007.2028
Published online 22 February 2007

Genetic studies of human diversity in East Asia

Feng Zhang[1], Bing Su[2], Ya-ping Zhang[2,4] and Li Jin[1,3,*]

[1]*Institute of Genetics, School of Life Sciences, Fudan University, Shanghai 200433, People's Republic of China*
[2]*Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, People's Republic of China*
[3]*CAS-MPG Partner Institute of Computational Biology, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China*
[4]*Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, People's Republic of China*

East Asia is one of the most important regions for studying evolution and genetic diversity of human populations. Recognizing the relevance of characterizing the genetic diversity and structure of East Asian populations for understanding their genetic history and designing and interpreting genetic studies of human diseases, in recent years researchers in China have made substantial efforts to collect samples and generate data especially for markers on Y chromosomes and mtDNA. The hallmark of these efforts is the discovery and confirmation of consistent distinction between northern and southern East Asian populations at genetic markers across the genome. With the confirmation of an African origin for East Asian populations and the observation of a dominating impact of the gene flow entering East Asia from the south in early human settlement, interpretation of the north–south division in this context poses the challenge to the field. Other areas of interest that have been studied include the gene flow between East Asia and its neighbouring regions (i.e. Central Asia, the Sub-continent, America and the Pacific Islands), the origin of Sino-Tibetan populations and expansion of the Chinese.

Keywords: East Asian populations; genetic structure; origin of modern humans; migrations; gene flow; admixture

1. INTRODUCTION

East Asia, which is located at the east end of the continent of Eurasia, is one of the most important regions for studying evolution and genetic diversity of human populations (Cavalli-Sforza 1998). Its importance is associated with the extensive presence of humans and their claimed ancestors over the last 2 Myr, and with being the crossroads connecting America and the Pacific Islands.

China, one of the centres of human civilization, comprises most of the geographical span, ethnic groups and languages of East Asia, as well as one-fifth of the human species. There are 56 officially identified ethnic groups in China, with Han being the largest in the world. Over 200 languages belonging to seven linguistic families (Altaic, Austroasiatic, Austronesian, Daic, Hmong-Mien, Sino-Tibetan and Indo-European) are spoken in China.

In the past two decades, much effort has been made by researchers in China and their international collaborators to characterize the structure of genetic diversity of human populations in China. The most significant progress of such studies started with the observation of genetic distinction between the southern and the northern East Asian populations (Zhao *et al.* 1987). This finding was confirmed by subsequent studies and the understanding of its genesis has become the jewel of population studies of East Asians.

Recognizing the importance of studies of genetic diversity in human populations, Chinese researchers were among the first few communities in the world who embraced the idea of the Human Genome Diversity Initiative (Cavalli-Sforza *et al.* 1991; Cavalli-Sforza 2005). The Chinese Human Genome Diversity Project (CHGDP) became the core of the Chinese Human Genome Project in 1993 and it aimed at understanding human genetic diversity in the Chinese and its applications in medical research. Such a nationwide corporation led to an impressive collection of samples encompassing all ethnic groups and contributed a large body of data of genetic variations on autosomes, the Y chromosome and mtDNA to the literature. This effort won the international recognition that it deserves (Cavalli-Sforza 1998).

2. THE NORTH–SOUTH DIVISION

Genetic markers are the tools in studying genetic variations. The most important genetic markers in human genetic diversity research (Du 2004) are: (i) blood groups that can be detected in red blood cells, including ABO, Rh and MNSs, (ii) human lymphocyte antigens and immunoglobulins, including Gm, Km and Am, (iii) isozyme markers, (iv) classic DNA polymorphisms using restriction fragment length polymorphism (RFLP), and (v) contemporary DNA markers, including short tandem repeat (STR or microsatellite) and single

* Author for correspondence (ljin007@gmail.com).

One contribution of 14 to a Theme Issue 'Biological science in China'.

nucleotide polymorphisms (SNPs). However, it is the introduction of mtDNA and Y chromosome markers that has made a profound impact on our understanding of the genetic diversity of human populations (Wallace *et al.* 1999; Jobling & Tyler-Smith 2003; Pakendorf & Stoneking in press).

Zhao *et al.* (1987) and Zhao & Lee (1989) studied the Gm and Km alleles (or allotypes) in 74 Chinese populations and found that there is an obvious genetic distinction between the southern and northern Chinese. By analysing a comprehensive dataset comprising 38 classical markers, Du & Xiao (Du *et al.* 1997) validated the genetic differentiation of southern and northern Chinese and showed that they are separated approximately by the Yangtze River. Chu *et al.* (1998) showed that such a north–south division can also be observed in Chinese populations using DNA markers (i.e. microsatellites). This genetic division is also consistent with multidisciplinary evidence in archaeology, physical anthropology, linguistics and surname distribution (Du *et al.* 1991, 1992; Jin & Su 2000). Furthermore, the results from Chu *et al.* (1998) demonstrated that the north–south division is not limited to Chinese populations and is in fact a reflection of a north–south division of East Asian populations. In the last few years, genetic data on mtDNA and Y chromosomes have been accumulated at an unprecedented pace for East Asian populations. Again, a north–south division of East Asians was observed not only with Y chromosome markers (Su *et al.* 1999; Shi *et al.* 2005), but also with mtDNA data (Kivisild *et al.* 2002; Yao *et al.* 2002). These observations provided convincing evidence of a north–south division in East Asian populations.

However, this well-established fact was not accepted without being challenged. Karafet *et al.* (2001) did not observe the north–south division in East Asians using a set of Y chromosome markers that are less polymorphic in East Asian populations and that over-represent the lineages brought in by recent admixture. In a different study, Ding *et al.* (2000) examined mtDNA, Y chromosome and autosomal variations and failed to observe a major north–south division. The southern populations in their study (Ding *et al.* 2000) are primarily the Tibeto-Burman (TB) populations, which have a recent northern origin, and therefore would blur the north–south distinction (Shi *et al.* 2005). A more extensive study of mtDNA lineages provided a much higher resolution and consequently a strong north–south division emerged (Yao *et al.* 2002).

The north–south division raises the question of whether the southern and northern East Asians (NEAS) are descendants of the same ancestral population in East Asia or originated from different populations that arrived in East Asia via different routes. To date, three main hypotheses have been brought forward on the entry of modern humans into East Asia: (i) entry from Southeast Asia followed by northward migrations (Turner 1987; Ballinger *et al.* 1992; Chu *et al.* 1998; Su *et al.* 1999; Yao *et al.* 2002; Shi *et al.* 2005), (ii) entry from northern Asia followed by southward migrations (Nei & Roychoudhury 1993), and (iii) southern and NEAS are derived from different ancestral populations, i.e. southern populations from Southeast Asia and northern populations from Central Asia (Cavalli-Sforza *et al.* 1994; Xiao *et al.* 2000; Karafet *et al.* 2001). Therefore, to understand the mechanism of genesis and maintenance of the north–south division, much needs to be learnt about the origin and migration of the East Asians.

3. Y CHROMOSOME AND mtDNA: POWERFUL TOOLS FOR STUDYING HUMAN EVOLUTION

Autosomes are inherited from both the parents and recombination makes it hard to trace a particular autosomal segment in human evolution. In the past decade, the utilities of mitochondrial DNA (mtDNA) and the non-recombining region of the Y chromosome (NRY) in studying population origin and divergence have been well recognized by both geneticists and anthropologists. Their advantages, including small effective population size and population-specific haplotype distribution, are powerful tools in tracing the demographic events of human evolution.

(a) *Non-recombining region of the Y chromosome and paternal transmission*

NRYs are paternally inherited without recombination; therefore, they are transmitted completely to the next generation. In addition, the effective population size (*Ne*) of NRYs is smaller than that of autosomes. Therefore, genetic drift would result in more dramatic shifts of the frequencies of Y chromosome markers than autosomal markers during human migration and, in turn, lead to more prominent specific regional distributions of NRY variations. These characters of the NRY make it a useful tool in tracing the origin and migration of human populations.

(b) *Y-single nucleotide polymorphisms and Y-short tandem repeats*

Underhill *et al.* (1996, 1997) have made significant contributions to identifying informative polymorphisms on NRY using denaturing high-performance liquid chromatography. Over 160 bi-allelic polymorphisms were detected in their studies. With the aid of these Y-SNPs, a genealogy of 116 haplotypes, which delineated the contemporary genetic structure of modern humans, was generated in a worldwide representative group of 1062 males (Underhill *et al.* 2000). These aforementioned markers and phylogeny of the NRY made it possible to trace the paternal lineages of the human beings.

Differing from SNPs, STRs are highly mutable and have multiple alleles. They provide more information and reveal more recent events than SNPs in population history. By analysing five STR loci on NRY in 15 diverse human populations, Deka *et al.* (1996) were among the first in evaluating the usefulness of Y-STRs in human evolution studies. In a more systematic study, Kayser *et al.* (1997) evaluated 13 Y-STRs (DYS19, DYS288, DYS385, DYS388, DYS389I/II, DYS390, DYS391, DYS392, DYS393, YCAI, YCAII, YCAIII and DXYS156Y) in 3825 males from 48 populations, and their results suggested that Y-STR loci are useful markers to trace paternal lineages. Furthermore, the higher mutation rate of Y-STRs (approx. 2.80×10^{-3} on average; Kayser *et al.* 2000) makes it possible to

distinguish the closely related Y chromosomes and, more importantly, to estimate the age of the Y-chromosomal lineages. Therefore, by combining the data of Y-SNPs and Y-STRs, NRYs can delineate the population history of modern humans.

(c) *mtDNA and maternal transmission*

MtDNA is a circular genome located in the mitochondria, which is transmitted to the next generation through oogenesis. At present, it has been widely accepted that mtDNA is maternally inherited (Pakendorf & Stoneking in press).

Similar to NRYs, mtDNA also lacks recombination and is more likely to generate population-specific mtDNA polymorphisms, which allow the maternal lineages in human evolution to be traced (Pakendorf & Stoneking in press). Furthermore, the high mutation rate of mtDNA makes it more informative than other genetic markers such as those of autosomes and Y chromosomes. However, the high mutation rate of mtDNA so frequently leads to recurrent mutations in high variation sequence (HVS) that it can be misleading (Yao *et al.* 2000).

The main part of human mtDNA is the coding region, which codes for 13 polypeptides essential in the process of oxidative phosphorylation, 2 ribosomal RNAs (12s and 16s rRNAs) and 22 transfer RNAs (tRNAs). Replication of mtDNA is regulated by a non-coding region, which is also called the control region or HVS (including HVSs I and II). Owing to low selection pressure, the control region is at a higher mutation rate and can give more information on the evolutionary history than the coding region. However, the high presence of recurrent mutations in the control region makes HVS data complex and hard to interpret without the assistance of variations in the coding region.

Kivisild *et al.* (2002) determined the mtDNA phylogeny of the East Asians by analysing both the coding and control region polymorphisms in 69 Han individuals from southern China. By analysing the HVS and coding variations of mtDNA in 263 Han individuals from six provinces in China, Yao *et al.* (2002) obtained a similar mtDNA genealogy. These studies, which generated a basal mtDNA phylogeny and a database of mtDNA variants in coding region and HVSs, can be regarded as the foundation of the population studies on maternal lineage of the East Asians.

According to the Cambridge reference sequence (Andrews *et al.* 1999), the human mtDNA is only approximately 16 569 bp in length. Therefore, sequencing whole mitochondrial genomes on a large scale is possible for the study of the genetic structure of various populations (Kong *et al.* 2003; Tanaka *et al.* 2004). The high-resolution mtDNA phylogeny generated can serve as a basis for genetic studies on human diversity in East Asia.

4. PROGRESS OF THE GENETIC DIVERSITY STUDY OF EAST ASIANS

(a) *Evidence supporting the 'out of Africa' hypothesis*

Regarding the origin of modern humans (*Homo sapiens sapiens*), there are two main hypotheses. One is the multiregional hypothesis, which insists independent evolutions from *Homo erectus* to *Homo sapiens* and to *H. s. sapiens* in different regions of the world. The 'out of Africa' hypothesis asserts that modern humans originate from an ancestral population in Africa, which expanded and spread out of Africa and completely replaced archaic human populations (*H. sapiens* or *H. erectus*) outside Africa (Cann *et al.* 1987; Wilson & Cann 1992).

Some anthropologists have claimed that anthropological evidence in China supported a regional evolution from *H. erectus* to modern human (Wu 1988). However, the phylogenetic tree of human mtDNA variations suggested that the ancestor of modern humans came out of Africa (Cann *et al.* 1987). From then on, more and more genetic data have been accumulated, most of which supported the out of Africa hypothesis (e.g. Bowcock *et al.* 1994; Hammer 1995; Tishkoff *et al.* 1996; Chu *et al.* 1998; Quintana-Murci *et al.* 1999; Su *et al.* 1999; Ingman *et al.* 2000; Ke *et al.* 2001).

As the mtDNA data supporting the out of Africa hypothesis were formerly restricted to the control region and RFLP in the coding region, Ingman *et al.* (2000) utilized the complete mtDNA sequences of 53 humans of diverse origins in tracing human evolution. As before, the newly generated mtDNA phylogeny indicated an African origin of modern humans (Ingman *et al.* 2000).

By analysing the markers on NRYs in over 1000 male individuals, Underhill *et al.* (2000) constructed a parsimonious genealogy to trace the evolution of modern humans. According to the phylogenetic tree, all the tested men from outside Africa share the same mutation (M168), which arose in Africa between 35 000 and 89 000 years ago. There are three parallel mutations (M89, M130 and Y Alu polymorphism, YAP) downstream of M168. Ke *et al.* (2001) examined these markers in 12 127 men from 163 populations in Asia and found that every sample had inherited one of these markers. This finding indicates that all of them are descendants of a relatively recent common ancestor in Africa, which supports a complete replacement of local archaic populations by modern humans from Africa from the perspective of paternal lineages.

(b) *Migration routes into East Asia*

While more and more genetic evidence supports the out of Africa hypothesis, the migration route which leads to the peopling of Eurasia remains controversial. As mentioned previously, there are two possible migration routes (from Central Asia and Southeast Asia) that were involved in three hypothetical models of the entry of modern humans into East Asia. By the phylogenetic analysis of 30 autosomal microsatellite loci in 28 populations, Chu *et al.* (1998) showed that northern and southern Chinese belong to distinct clusters and indicated that the colonization of East Asia might be mainly attributed to the northward migration of the settlers from Southeast Asia. Yao *et al.* (2002) analysed the mtDNA gene pool of Han Chinese and observed a south-to-north cline based on the haplogroup frequency. In addition, the haplogroups in southern East Asians (SEAS) were found to be

more ancient than those in the northern population (Yao *et al.* 2002). This evidence suggested that the southern route might play an important part in the peopling of East Asia.

On the paternal side, Su *et al.* (1999) examined 19 Y-SNPs and 3 Y-STRs in 925 males from a wide area of Asia and observed that SEAS are much more polymorphic than NEAS, which suggested the southern origin of northern populations and the migratory route from south to north after the initial Palaeolithic peopling of East Asia. Recently, in a systematic sampling and genetic screening of one East Asian-specific Y chromosome haplogroup (O3-M122) in 2332 males from diverse East Asian populations, Shi *et al.* (2005) showed that this haplotype was more polymorphic in SEAS than in NEAS, which suggested a southern origin of the O3-M122 mutation. According to the Y-STR data, it was estimated that the early northward migration of the O3-M122 lineages in East Asia occurred *ca* 25 000–30 000 years ago (Shi *et al.* 2005).

(c) *Admixture of northern East Asian and West Eurasian*

Though the evidence of various genetic markers suggests that the southern route makes the main contribution to the gene pool of East Asians, the effect of genetic admixture in Central Asia (including the northwest part of China) cannot be neglected. Different proportions of lineages in the admixture from Central Asia appears to be a reason for the division between NEAS and SEAS.

Similarity of some NEAS to Central Asians indicates that the genetic admixture associated with trade along the Silk Road might have played an important role in the diversity in East Asia. Zhao & Lee (1989) studied the Gm allotypes and found some Caucasian-related haplotypes in populations of Northwest China. Similarly, Yao *et al.* (2000) studied the HVS-I region of mtDNA and melanocortin 1 receptor-gene polymorphisms in the ethnic populations of Northwest China (including Uighur, Kazak and Tu), and extensive gene admixture of northern East Asian and West Eurasian along the Silk Road was indicated. In a consequent study, Yao *et al.* (2004) analysed 252 mtDNAs of five ethnic groups (Uygur, Uzbek, Kazak, Mongolian and Hui) from Xinjiang, China and divided them into the eastern and western Eurasian pools according to previous studies. Their results suggested that Central Asia is the main location of genetic admixture of the east and west. In addition, the frequency of the western Eurasian-specific haplogroups ranges from 42.6 to 6.7% across populations, indicating the different contributions of the west and east gene pools in the admixture process (Yao *et al.* 2004).

Recently, by examining mtDNA and physical characters of 134 human remains excavated from nine sites (dating from 2500 BC to AD 200) in Northwest China, Jin and colleagues found that both the genetic and physical characters of the East and West Eurasians could be observed in some individuals of these Bronze Age populations, indicating a genetic interaction of the East and West Eurasians before the rise of the Silk Road (Li Jin 2005, unpublished data).

(d) *Studies on ancient DNA*

Ancient human DNA can improve our knowledge of human evolution. The desiccation of DNA in ancient specimens makes it challenging to study. However, analysis of mtDNA sequence became the main tool adopted in studies on ancient DNA due to its relatively higher copy number (Hofreiter *et al.* 2001).

Wang *et al.* (2000) studied the genetic structure of human populations in Linzi (Shandong, China) during the past 2500 years by comparing the mtDNA sequences of three populations in different periods (2500 years ago, 2000 years ago and the present day), and strikingly showed that the 2500-year-old Linzi populations had greater genetic similarity to present-day European populations than to present-day East Asian populations. Later, Yao *et al.* (2003) reanalysed the data of Wang *et al.* (2000) and showed that the ancient populations in Linzi are similar to the modern Chinese populations rather than to the Europeans. The false interpretation of the mtDNA sequences of the ancient Linzi populations might be caused by insufficient data (Wang *et al.* 2000; Yao *et al.* 2003). For this reason, it is advisable to combine the sequencing data of the fast-evolving HVS and the RFLP data of the coding region on mtDNA in the study of mtDNA variations (Yao *et al.* 2003).

(e) *Genetic affinity in Sino-Tibetan populations*

Sino-Tibetan languages include Chinese and TB, and the linguistic connection between these two subfamilies is well established (Martisoff 1991). Based on the archaeological findings, the ancestors who spoke Proto-Sino-Tibetan were estimated to live over 6000 years ago (Matisoff 1991; Cavalli-Sforza *et al.* 1994). In the genetic studies using classical autosomal markers (Du *et al.* 1997) and microsatellite markers (Chu *et al.* 1998), it has been confirmed that Tibetans diverged from the NEAS.

By analysing 19 Y-SNPs and 3 Y-STR markers in 607 individuals from 31 Sino-Tibetan-speaking populations residing in East, Southeast and South Asia, Su and other colleagues (Qian *et al.* 2000; Su *et al.* 2000*a*,*b*) showed that a T to C mutation at the M122 locus is highly prevalent in almost all of the Sino-Tibetan populations, which implied a strong genetic affinity among populations in the same language family. In addition, Su *et al.* (2000*a*,*b*) also suggested that the ancient people living in the upper–middle Yellow River basin *ca* 10 000 years ago were the ancestors of modern Sino-Tibetan populations.

(f) *Population study of Tibeto-Burman speakers*

TB is one of the two subfamilies of the Sino-Tibetan language family. There are 351 living languages in this subfamily, primarily distributing in East, South and Southeast Asia. In China, TB-speaking populations mainly reside in Qinghai, Tibet, Sichuan, Yunnan and Hunan. According to historical records, the TB populations were derived from the ancient *Di-Qiang* tribes in Northwest China. About 2600 BP, the TB populations embarked on a large-scale southward migration by the Tibetan–Burman Corridor (Wang 1994). This is consistent with genetic evidence based on Y chromosome markers that almost all the TB

populations share a high frequency of M122-C and M134-deletion (Su *et al.* 2000*a*,*b*).

Wen *et al.* (2004*a*,*b*) analysed 10 Y-SNPs in 965 individuals from 23 TB populations, and HVS-1 sequence and a few coding region variants of mtDNA in 756 individuals from 21 TB populations. Principal components analysis showed that the northern TB populations and the southern native groups played a significant role in shaping the gene pool of the southern TB populations with an unequal contribution of male and female lineages from the parental populations.

(g) *Han culture and its expansion*

The spread of culture in human populations can be explained by two alternative models. The demic diffusion model involves mass movement of people, while the cultural diffusion model refers to cultural impact between populations (Cavalli-Sforza *et al.* 1994). Historical records show that the Hans originated from the ancient *Huaxia* tribes in northern China and experienced a continuous expansion into southern China over the past two millennia (Ge *et al.* 1997).

To test this hypothesis of demic diffusion, Wen *et al.* (2004*a*,*b*) examined genetic variations on both NRY and mtDNA in 28 Han populations in China. According to the NRY data, northern (NH) and southern Hans (SH) share similar haplogroup frequencies. The M122-C mutation is prevalent in almost all the Han populations studied (53.8% in NHs and 54.2% in SHs), while M119-C and M95-T, prevalent in southern natives (SNs), are more frequent in SHs (19%) than in NHs (5%). Some haplogroups prevalent in SNs, such as O1b-M110, O2a1-M88 and O3d-M7, are only observed in some SHs. According to the mtDNA lineages, NHs and SHs are significantly different in their mtDNA lineages. The frequency of haplogroups dominant in the NEAS (A, C, D, G, M8a, Y and Z) is 55% in NHs, which is much higher than that in SHs (36%). In contrast, the frequency of the haplogroups dominant in SNs (B, F, R9a, R9b and N9a) is much higher in SHs (55%) than in NHs (33%).

These observations of Wen *et al.* (2004*a*,*b*) are consistent with historical records, in which the continuous southward migration of the Hans caused by warfare and famine is mentioned. Taking this genetic and historic evidence into account, it can be concluded that the migration into South China is one of the main causes of the expansion of Han culture.

(h) *Summary of progress in genetic diversity studies in East Asians*

In the past decade, the NRY and mtDNA markers have been used to analyse the genetic structure of almost all the 56 officially identified ethnic groups and other unsorted populations. As well as the major members of the CHGDP, more and more Chinese research groups have been joining this promising field of scientific research.

In Korea and Japan, human genetic diversity projects have also been launched and some interesting findings have been published. Hammer & Horai (1995) found that the insertion allele of YAP (also called DYS287) is prevalent (approx. 42%) in Japanese populations. Differing from the E-YAP+ haplogroup in Africans, West Asians and Europeans, the Y chromosomes of YAP+ belong to the D haplogroup (defined by M174), which is at a high frequency in Japanese and Tibetans but is rare in many other East Asian populations, such as the Han Chinese (Jobling & Tyler-Smith 2003). It is suggested that YAP+ chromosomes might have migrated to Japan with the Jomon people over 10 000 years ago (Hammer & Horai 1995). Tajima *et al.* (2002) used seven biallelic Y-SNPs (DYS257108, DYS287, SRY4064, SRY10831, RPS4Y711, M9 and M15) to analyse 610 males from 14 global populations; their results suggested that three major groups with different paternal ancestries separately migrated to prehistoric East and Southeast Asia. Jin *et al.* (2003) examined eight Y-SNP markers (YAP, RPS4Y711, M9, M175, LINE1, SRY+465, 47z and M95) and three Y-STR markers (DYS390, DYS391 and DYS393) in 738 males (including 160 Koreans and 108 Japanese) in East Asia to study the paternal lineage history of Korea. The distribution pattern of Y-chromosomal haplogroups suggested a dual origin for Koreans (a northern Asian settlement and expansion from southern into northern China).

By sequencing the complete mitochondrial genomes of 672 Japanese individuals, Tanaka *et al.* (2004) constructed an mtDNA phylogeny with high resolution and found some new haplogroups. This phylogeny will be very helpful for analysing the mtDNA diversity and tracing the migration of the maternal lineage of East Asians. In addition, Tanaka *et al.* (2004) combined their data with mtDNA sequences from other populations of Asia and revealed that present-day Japanese have the closest genetic affinity to the northern Asian populations.

5. RELATED TOPICS WITH POPULATION STUDIES IN EAST ASIANS

(a) *Ancestral Asian lineages in India*

South Asia, including India, is an important corridor for modern human dispersal out of Africa to East Asia and Oceania. In this area, there are many diverse populations with different morphological, cultural and linguistic characteristics. Using mtDNA data mainly from HVS and RFLPs of the coding region, the genetic structure of South Asians has been partially delineated (Passarino *et al.* 1996; Kivisild *et al.* 1999, 2003; Bamshad *et al.* 2001; Quintana-Murci *et al.* 2004).

To obtain a phylogeny of mtDNA with higher resolution and to study the relationship between the Indian and the western Eurasian more precisely, Palanichamy *et al.* (2004) sampled 75 mtDNA lineages in haplogroup N lineages from over 800 samples (including Reddy, Thogataveera, Brahmin, Rajbhansi and the Khasi population) across India, to sequence the complete mtDNA genome. In that study, five new autochthonous haplogroups (R7, R8, R30, R31 and N5) were identified and some previously described autochthonous haplogroups (R5, R6, N1d, U2a, U2b and U2c) were further characterized with the complete sequence data. By carefully constructing the phylogeny of macro-haplogroup N, Palanichamy *et al.* (2004) showed that the Indian mtDNA pool harbours at least as many deepest-branching lineages as the western

Eurasian mtDNA pool. Furthermore, the evidence of the indigenous haplogroup R lineages in India suggested a common initial spread of the root haplotypes of M, N and R along the southern route, along the Asian coastline, some 60–70 kyr ago, which will be meaningful for the colonization of Southeast Asia, East Asia and Oceania. Recently, Sun *et al.* (2006) selected 56 mtDNAs from over 1200 samples across India for complete sequencing, with the intention of covering all Indian autochthonous M lineages. As a result, the phylogenetic status of previously identified haplogroups based on control-region and/or partial coding-region information, such as M2–M6, M30 and M33, was solidified or redefined. Moreover, seven novel basal M haplogroups (M34–M40) were identified and yet another five singular branches of the M phylogeny were discovered. The comparison of matrilineal components from India, East Asia, Southeast Asia and Oceania at the deepest level yielded a star-like and non-overlapping pattern, reflecting a rapid dispersal of modern humans along the Asian coast after the initial 'out of Africa' event.

(b) *Genetic legacy of Mongols*

Zerjal *et al.* (2003) identified a Y-chromosomal haplogroup C* (×C3c) with high frequency (approx. 8%) in a large region of Asia, which constitutes approximately 0.5% of the worldwide populations.

With the aid of Y-STRs, the age of the most recent common ancestor of this haplogroup was estimated to be only *ca* 1000 years. How can this lineage expand at such a high rate? Taking the historical records into account, Zerjal *et al.* (2003) suggested that the expansion of this C* haplogroup across East Eurasia is linked to the establishment of the Mongol empire by Genghis Khan (1162–1227). Genghis Khan and his male relatives are expected to bear the Y chromosomes of C*. Considering their high social status, this Y chromosome lineage was probably enlarged by the reproduction of numerous offspring. In the course of expeditions, this special lineage spread, partially replaced the local paternal gene pool and developed in the subsequent rulers. Interestingly, Zerjal *et al.* (2003) have found that the boundaries of the Mongol empire match the distribution of the C* lineage well. It is a good example of how social factors, as well as biological selection effects, can play an important role in human evolution.

(c) *Origin of Americans*

America is the last continent settled by modern humans. There are three linguistically identified groups of population: Amerind, Eskimo-Aleut and Na-Dene. mtDNA haplogroups of Native America include four Asian haplogroups (A, B, C and D) and one European haplogroup (X; Mulligan *et al.* 2004). Wallace *et al.* (1985) studied Amerind populations and showed that the sequence diversity of haplogroup B is much lower than those of haplogroups A, C and D. Furthermore, haplogroup B is absent in Siberia, while A, C and D are prevalent. These two observations imply that the Amerind linguistic group might have been derived from two migrations.

Lell *et al.* (2002) analysed 12 Y-SNPs in 549 individuals from Siberia and the Americas. Three major Y lineages of Native American populations have been found: M3 (66%), M45 (25%) and M130 (5%). M3, also known as DYS119 (Underhill *et al.* 1996), was confined to the Chukoka peninsula in Siberia. M45 was divided into two subgroups; one subgroup (M45a) is found throughout the Americas, and another (M45b) is prevalent in North and Central America. These two sub-haplogroups have different distribution patterns in Siberia (M45a in middle Siberia and M45b in eastern Siberia). The C-M130 haplogroup has a similar distribution to that of M45b in Siberia and in North America. They hypothesized that there were two independent migrations into America from Siberia, which is consistent with the mtDNA evidence (Wallace *et al.* 1985). M242 is a polymorphism, which was introduced after M74 (arising in Asia) but before M3 (arising in America) in the phylogeny of the human Y chromosome (Underhill *et al.* 1996, 2000), and can be used to date the entry into the Americas. Based on the diversity of 15 Y-STRs in 69 Eurasian M242-T samples, the time of first entry into the Americas was estimated to be close to 15 000–18 000 years BP (Seielstad *et al.* 2003).

(d) *Origin of Polynesians*

There has been controversy regarding the origin of Polynesian populations, which have been classified as a part of the Austronesian linguistic family. The express train hypothesis, a well-accepted theory on the origin of Austronesian (Diamond 1988), postulates that Proto-Austronesian originated in Taiwan and began to expand southward *ca* 5000–6000 years ago, by way of the Philippines and eastern Indonesia, and eventually navigated eastward to Micronesia and Polynesia. The 'express train' refers to the swift migration in the last leg of this journey starting from eastern Indonesia. Pertaining to East Asian diversity studies, the hypothesis of Taiwanese origin (referred to as the Taiwan homeland hypothesis) requires careful examination.

To test the Taiwan homeland hypothesis, Su *et al.* (2000*a*,*b*) examined 19 Y-SNPs in 551 males from 36 populations living in Southeast Asia, Taiwan, Micronesia, Melanesia and Polynesia. Surprisingly, there is a virtual absence of the Formosan haplotypes in Micronesia and Polynesia. However, the presence of all the Polynesian, Micronesian and Formosan haplotypes in Southeast Asians suggested that Southeast Asians might be the ancestral population for Formosan and Polynesian (Su *et al.* 2000*a*,*b*). Recently, Jin and colleagues examined 20 Y-SNPs and 7 Y-STRs in 1325 males from 29 Daic, 23 Polynesian and 11 Formosan populations, and showed that Taiwan is unlikely to be the homeland of Austronesian; and that Austronesian is not a genetically monophyletic group. Furthermore, the NRY evidence supported the idea that Polynesian and Formosan derived from Daic separately (Li Jin 2005, unpublished data).

By assessing mtDNA variations in 640 individuals from nine tribes from Taiwan, Trejaut *et al.* (2005) showed the prevalence of several haplogroups (B4, B5a, F1a, F3b, E and M7) in the Formosan

populations, which indicated that Taiwan was the common origin of the Austronesian populations. In addition, a new sub-haplogroup (B4a1a) was defined according to the sequence data, which supported the origin of Polynesian migration as being from Taiwan (Trejaut *et al.* 2005).

One explanation for the inconsistent results, mainly between the NRY evidence and the mtDNA data, is that the migration pattern of the Proto-Austronesian populations may be different for the paternal and maternal lineages.

6. PERSPECTIVES

While the origin of East Asians is still being debated, it bears very little relevance to the genetic diversity of populations in East Asia today, so a firm step has been taken to move beyond this controversy. An understanding of the mechanisms of the genesis and maintenance of the north–south division constitutes the frontiers of science. As relevant as they can be, the origin of genetic structure, languages and ethnic groups will come under the spotlight. Much effort has been invested in minority ethnic groups, but not in the bulk of major populations, including the Han Chinese. Systematic characterization of the genetic structure of East Asian populations is therefore called for, considering that the renewed goals of genetic diversity studies in East Asia will include the application of our knowledge of genetic structure to medical research.

REFERENCES

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. 1999 Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147. (doi:10.1038/13779)

Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., Chen, K. H. & Wallace, D. C. 1992 Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics* **130**, 139–152.

Bamshad, M. *et al.* 2001 Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11**, 994–1004. (doi:10.1101/gr.GR-1733RR)

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. 1994 High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457. (doi:10.1038/368455a0)

Cann, R. L., Stoneking, M. & Wilson, A. C. 1987 Mitochondrial DNA and human evolution. *Nature* **325**, 31–36. (doi:10.1038/325031a0)

Cavalli-Sforza, L. L. 1998 The Chinese human genome diversity project. *Proc. Natl Acad. Sci. USA* **95**, 11 501–11 503. (doi:10.1073/pnas.95.20.11501)

Cavalli-Sforza, L. L. 2005 The human genome diversity project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340. (doi:10.1038/nrg1596)

Cavalli-Sforza, L. L., Wilson, A. C., Cantor, C. R., Cook-Deegan, R. M. & King, M. C. 1991 Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. *Genomics* **11**, 490–491.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994 *The history and geography of human genes*. Princeton, NJ: Princeton University Press.

Chu, J. Y. *et al.* 1998 Genetic relationship of populations in China. *Proc. Natl Acad. Sci. USA* **95**, 11 763–11 768. (doi:10.1073/pnas.95.20.11763)

Deka, R., Jin, L., Shriver, M. D., Yu, L. M., Saha, N., Barrantes, R., Chakraborty, R. & Ferrell, R. E. 1996 Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. *Genome Res.* **6**, 1177–1784.

Diamond, J. M. 1988 The express train to Polynesia. *Nature* **336**, 307–308. (doi:10.1038/336307a0)

Ding, Y. C. *et al.* 2000 Population structure and history in East Asia. *Proc. Natl Acad. Sci. USA* **97**, 14 003–14 006. (doi:10.1073/pnas.240441297)

Du, R. F. 2004 *Population genetics in Chinese*. Beijing, China: Science Press.

Du, R., Yuan, Y., Hwang, J., Mountain, J. & Cavalli-Sforza, L. L. 1991 *Chinese surnames and the genetic differences between north and south China*. Stanford, CA: Morrison Institute for Population Resource Studies.

Du, R., Yuan, Y., Hwang, J., Mountain, J. & Cavalli-Sforza, L. L. 1992 Chinese surnames and the genetic differences between north and south China. *J. Chinese Linguist* Monograph Series No. 5.

Du, R. F., Xiao, C. J. & Cavalli-Sforza, L. L. 1997 Genetic distance between Chinese populations calculated on gene frequencies of 38 loci. *Sci. China (Series C)* **40**, 613–621.

Ge, J. X., Wu, S. D. & Chao, S. J. 1997 *The migration history of China*. Fuzhou, China: Fujian People's Publishing House. [In Chinese.]

Hammer, M. F. 1995 A recent common ancestry for human Y chromosomes. *Nature* **378**, 376–378. (doi:10.1038/378376a0)

Hammer, M. F. & Horai, S. 1995 Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet.* **56**, 951–962.

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. 2001 Ancient DNA. *Nat. Rev. Genet.* **2**, 353–359. (doi:10.1038/35072071)

Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. 2000 Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713. (doi:10.1038/35047064)

Jin, L. & Su, B. 2000 Natives or immigrants: modern human origin in East Asia. *Nat. Rev. Genet.* **1**, 126–133. (doi:10.1038/35038565)

Jin, H. J. *et al.* 2003 Y-chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. *Hum. Genet.* **114**, 27–35. (doi:10.1007/s00439-003-1019-0)

Jobling, M. A. & Tyler-Smith, C. 2003 The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4**, 598–612. (doi:10.1038/nrg1124)

Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S., Redd, A. J., Zegura, S. L. & Hammer, M. F. 2001 Paternal population history of east Asia: sources, patterns, and microevolutionary process. *Am. J. Hum. Genet.* **69**, 615–628. (doi:10.1086/323299)

Kayser, M. *et al.* 1997 Evaluation of Y-chromosomal STRs: a multicenter study. *Int. J. Legal Med.* **110**, 125–133. (doi:10.1007/s004140050051)

Kayser, M. *et al.* 2000 Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* **66**, 1580–1588. (doi:10.1086/302905)

Ke, Y. *et al.* 2001 African origin of modern humans in east Asia: a tale of 12,000 Y chromosomes. *Science* **292**, 1151–1153. (doi:10.1126/science.1060011)

Kivisild, T. *et al.* 1999 Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.* **9**, 1331–1334. (doi:10.1016/S0960-9822(00)80057-3)

Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H. J. & Villems, R. 2002 The emerging limbs and twigs of the East Asian mtDNA tree. *Mol. Biol. Evol.* **19**, 1737–1751.

Kivisild, T. *et al.* 2003 The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* **72**, 313–332. (doi:10.1086/346068)

Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L. & Zhang, Y. P. 2003 Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am. J. Hum. Genet.* **73**, 671–676. (doi:10.1086/377718)

Lell, J. T., Sukernik, R. I., Starikovskaya, Y. B., Su, B., Jin, L., Schurr, T. G., Underhill, P. A. & Wallace, D. C. 2002 The dual origin and Siberian affinities of Native American Y chromosomes. *Am. J. Hum. Genet.* **70**, 192–206. (doi:10.1086/338457)

Martisoff, J. A. 1991 Sino-Tibetan linguistics: present state and future prospects. *Annu. Rev. Anthropol.* **20**, 469–504. (doi:10.1146/annurev.an.20.100191.002345)

Mulligan, C. J., Hunley, K., Cole, S. & Long, J. C. 2004 Population genetics, history, and health patterns in native Americans. *Annu. Rev. Genomics Hum. Genet.* **5**, 295–315. (doi:10.1146/annurev.genom.5.061903.175920)

Nei, M. & Roychoudhury, A. K. 1993 Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* **10**, 927–943.

Pakendorf, B. & Stoneking, M. 2005 Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* **6**, 165–183.

Palanichamy, M. G. *et al.* 2004 Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am. J. Hum. Genet.* **75**, 966–978. (doi:10.1086/425871)

Passarino, G., Semino, O., Bernini, L. F. & Santachiara-Benerecetti, A. S. 1996 Pre-Caucasoid and Caucasoid genetic features of Indian population revealed by mtDNA polymorphisms. *Am. J. Hum. Genet.* **59**, 927–934.

Qian, Y. P. *et al.* 2000 Multiple origins of Tibetan Y chromosomes. *Hum. Genet.* **106**, 453–454. (doi:10.1007/s004390000259)

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. 1999 Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat. Genet.* **23**, 437–441. (doi:10.1038/70550)

Quintana-Murci, L. *et al.* 2004 Where west meets east: the complex mtDNA landscape of the southwest and central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845. (doi:10.1086/383236)

Seielstad, M., Yuldasheva, N., Singh, N., Underhill, P., Oefner, P., Shen, P. & Wells, R. S. 2003 A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas. *Am. J. Hum. Genet.* **73**, 700–705. (doi:10.1086/377589)

Shi, H., Dong, Y., Wen, B., Xiao, C. J., Underhill, P. A., Shen, P., Chakraborty, R., Jin, L. & Su, B. 2005 Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am. J. Hum. Genet.* **77**, 408–419. (doi:10.1086/444436)

Su, B. *et al.* 1999 Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. *Am. J. Hum. Genet.* **65**, 1718–1724. (doi:10.1086/302680)

Su, B. *et al.* 2000*a* Polynesian origins: new insights from the Y-chromosome. *Proc. Natl Acad. Sci. USA* **97**, 8225–8228. (doi:10.1073/pnas.97.15.8225)

Su, B. *et al.* 2000*b* Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* **107**, 582–590. (doi:10.1007/s004390000406)

Sun, C. *et al.* 2006 The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol. Biol. Evol.* **23**, 683–690. (doi:10.1093/molbev/msj078)

Tajima, A. *et al.* 2002 Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. *Hum. Genet.* **110**, 80–88. (doi:10.1007/s00439-001-0651-9)

Tanaka, M. *et al.* 2004 Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14**, 1832–1850. (doi:10.1101/gr.2286304)

Tishkoff, S. A. *et al.* 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387. (doi:10.1126/science.271.5254.1380)

Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., Li, Z. Y. & Lin, M. 2005 Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol.* **3**, e247. (doi:10.1371/journal.pbio.0030247)

Turner, C. G. 1987 Late Pleistocene and Holocene population history of East Asia based on dental variations. *Am. J. Phys. Anthropol.* **73**, 305–321. (doi:10.1002/ajpa.1330730304)

Underhill, P. A., Jin, L., Zemans, R., Oefner, P. A. & Cavalli-Sforza, L. L. 1996 A pre-Columbian human Y chromosome-specific transition and its implications for human evolution. *Proc. Natl Acad. Sci. USA* **93**, 196–200. (doi:10.1073/pnas.93.1.196)

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L. & Oefner, P. J. 1997 Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7, 996–1005.

Underhill, P. A. *et al.* 2000 Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**, 358–361. (doi:10.1038/81685)

Wallace, D. C., Garrison, K. & Knowler, W. C. 1985 Dramatic founder effects in Amerindian mitochondrial DNAs. *Am. J. Phys. Anthropol.* **68**, 149–155. (doi:10.1002/ajpa.1330680202)

Wallace, D. C., Brown, M. D. & Lott, M. T. 1999 DNA mitochondrial variation in human evolution and disease. *Gene* **238**, 211–230. (doi:10.1016/S0378-1119(99)00295-4)

Wang, Z. H. 1994 *History of nationalities in China*. Beijing, China: China Social Science Press. [In Chinese.]

Wang, L., Oota, H., Saitou, N., Jin, F., Matsushita, T. & Ueda, S. 2000 Genetic structure of a 2500-year-old human population in China and its spatiotemporal changes. *Mol. Biol. Evol.* **17**, 1396–1400.

Wen, B. *et al.* 2004*a* Genetic evidence supports demic diffusion of Han culture. *Nature* **431**, 302–305. (doi:10.1038/nature02878)

Wen, B. *et al.* 2004*b* Analyses of genetic structure of Tibeto-Burman populations reveals sexbiased admixture in southern Tibeto-Burmans. *Am. J. Hum. Genet.* **74**, 856–865. (doi:10.1086/386292)

Wilson, A. C. & Cann, R. L. 1992 The recent African genesis of humans. *Sci. Am.* **266**, 68–73.

Wu, X. 1988 Comparative study of early *Homo sapiens* from China and Europe. *Acta Anthropol. Sin.* **7**, 287–293. [In Chinese.]

Xiao, C. J., Du, R. F., Cavalli-Sforza, L. L. & Minch, E. 2000 Principal component analysis of gene frequencies of Chinese populations. *Sci. China (Ser. C)* **43**, 472–481.

Yao, Y. G., Lu, X. M., Luo, H. R., Li, W. H. & Zhang, Y. P. 2000 Gene admixture in the silk road of China: evidence from mtDNA and melanocortin 1 receptor polymorphism. *Genes Genet. Syst.* **75**, 173–178. (doi:10.1266/ggs.75.173)

Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T. & Zhang, Y. P. 2002 Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* **70**, 635–651. (doi:10.1086/338999)

Yao, Y. G., Kong, Q. P., Man, X. Y., Bandelt, H. J. & Zhang, Y. P. 2003 Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. *Mol. Biol. Evol.* **20**, 214–219. (doi:10.1093/molbev/msg026)

Yao, Y. G., Kong, Q. P., Wang, C. Y., Zhu, C. L. & Zhang, Y. P. 2004 Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China. *Mol. Biol. Evol.* **21**, 2265–2280. (doi:10.1093/molbev/msh238)

Zerjal, T. *et al.* 2003 The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72**, 717–721. (doi:10.1086/367774)

Zhao, T. M. & Lee, T. D. 1989 Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. *Hum. Genet.* **83**, 101–110. (doi:10.1007/BF00286699)

Zhao, T. M., Zhang, G. L., Zhu, Y. M., Zheng, S. Q., Lui, D. Y., Chen, Q. & Zhang, X. 1987 The distribution of immunoglobulin Gm allotypes in forty Chinese populations. *Acta Anthropol. Sin.* **6**, 1–9. [In Chinese.]