A Genetic Focus on the Peopling History of East Asia: Critical Views

Alicia Sanchez-Mazas · Da Di · María Eugenia Riccio

Abstract Population genetic studies may provide crucial information for our knowledge on human peopling history and have been extensively applied to reconstruct East Asian prehistory in the last 10 years. However, different genetic investigations are not always consistent with each other and some results are conflicting or misinterpreted. This represents a main obstacle for scholars of other disciplines like archaeologists and linguists who try to relate the genetic information on past human migrations to their own results on the spread of domesticated crops or animals or on the dispersal of the main language families. In this paper, we review the current genetic evidence related to the peopling history of East Asia with a critical view on some interpretations. In this way, we hope to provide a useful reference for further interdisciplinary studies on our past.

Abbreviations

HLA	Human leucocyte antigen
RH	Rhesus
GM	IgG immunoglobulins' genetic marker
SNPs	Single-nucleotide polymorphisms
NRY	Non-recombining portion of the Y chromosome
mtDNA	Mitochondrial DNA
STR	Short tandem repeat
PCA	Principal component analysis
MDS	Multidimensional scaling analysis
AMOVA	Analysis of molecular variance
MRCA	Most recent common ancestor
NEAS	Northeast Asian populations
SEAS	Southeast Asian populations
CAS	Central Asians
WAS	West Asians

A. Sanchez-Mazas (✉) · D. Di · M. E. Riccio
Laboratory of Anthropology, Genetics and Peopling history,
Anthropology Unit, Department of Genetics and Evolution,
University of Geneva,
Geneva, Switzerland
e-mail: alicia.sanchez-mazas@unige.ch

Introduction

East Asia is a very large geographic area of the world currently inhabited by more than 1.5 billion people, which represents about 22% of the world population. According to the current fossil record and as supported by genetic evidence, it is likely that modern humans did not originate in East Asia (Jin and Su 2000). However, this vast continental region was probably settled very early in the Paleolithic when the first *Homo sapiens* spread throughout the world, presumably from East Africa; although southern Africa has recently been proposed as another possible original homeland (Henn et al. 2011) and very early well-dated skeletons are also known in the Levant (Stringer et al. 1989). Old human fossil remains are yet very scarce in East Asia (with a very uncertain date of 68,000 years for the most ancient fossil known to date, *Liujiang*, that was found by chance (Shen et al. 2002)) and do not allow on their own to reconstruct East Asian prehistory. During the last decades, different disciplines, i.e., archaeology, linguistics and population genetics, have been involved together in the reconstruction of human peopling history, bringing some indisputable results but also revealing the extraordinary complexity of East Asian past human settlement history (Sagart et al. 2005a; Sanchez-Mazas et al. 2008). As a matter of fact, distinct scenarios of human migrations in

East Asia are still disputed today, like the first arrival of *H. sapiens* from West Asia, either through a single southern route along the sea coast or through two independent routes via the southern and northern edges of the Himalayas. Also, human population migrations in East Asia have been investigated in relation to the spread of agriculture and the main linguistic families, leading to quite different views on the subject.

In this paper, our objective is to make clear the current genetic evidence related to East Asian peopling history. To that aim, we have described the available information—including our own conclusions based on human leucocyte antigen (HLA) genetic studies—within two distinct sections to distinguish the raw genetic results from their interpretation, the latter being more subjective and open to criticism. We also underline the major limitations of such genetic studies, as a complement to what was previously explained by Blench et al. (2008). Finally, we propose some perspectives in this area for future genetic studies. We hope that this clarification will be useful to scholars of other disciplines who are often confused by the very specialized and often contradictory information provided by population genetics studies.

Raw results from population genetic studies

North–south genetic differentiations in East Asia

One of the most robust results found in genetic studies focusing on East Asian populations is the marked genetic differentiation between Northeast Asian and Southeast Asian populations (NEAS and SEAS), respectively (Cavalli-Sforza et al. 1994; Chu et al. 1998; Du et al. 1997; Xue et al. 2005). This general structure is observed for all genetic markers studied so far and is characterized either by sharp differences of gene frequencies or by the occurrence of distinct genetic lineages. For classical markers, the best specific examples are provided by the distribution of Rhesus (RH) and IgG immunoglobulins' genetic marker (GM) haplotypes showing very high frequencies of RH*R2, GM*1,17;21 and GM*1,2,17;21 in Northeast Asia, and RH*R1 and GM*1,3;5* in Southeast Asia (Poloni et al. 2005; Sanchez-Mazas 2008). Overall analyses on a set of classical markers have also shown such differentiations (Cavalli-Sforza et al. 1994). At the HLA loci, distinct alleles and/or allelic lineages display contrasted frequencies between NEAS and SEAS, leading to the definition of "group 1" and "group 2" alleles (Di and Sanchez-Mazas 2011). Through principal coordinate analyses, a general differentiation between northern and southern populations is also observed at most HLA loci (HLA-A, -B, -C, -DPB1, -DRB1) (Di and Sanchez-Mazas (2011) *The north–south differentiation of East Asian populations*; in preparation). Gender-specific polymorphisms show a contrast

between northeastern and southeastern populations as well, e.g., for mitochondrial DNA (mtDNA), A, C, D and G haplogroups are more frequent in the north, while B and F are more frequent in the south (Kivisild et al. 2002; Stoneking and Delfin 2010; Yao et al. 2002). For the Y chromosome, northeastern and southeastern populations are clearly discriminated both in a principal component analysis (PCA) using 19 non-recombining Y (NRY) markers (Su et al. 1999) and in a multidimensional scaling analysis (MDS) using 52 markers (Karafet et al. 2001). More recently, PCA based on genome-wide analyses of tens of thousands of single-nucleotide polymorphisms (SNPs) confirm that genetic differentiation vary principally with latitude in East Asia (Abdulla et al. 2009; Suo et al. 2011).

### Genetic boundaries versus continuous patterns

Whereas NEAS and SEAS clearly exhibit contrasted genetic profiles, this result alone does not indicate how genetic diversity is structured among the two areas. The existence of a clear-cut genetic boundary between northern and southern populations has been debated: based on classical markers, a boundary located at the vicinity of the Yangtze River has been suggested (Du et al. 1997; Xiao et al. 2000; Xue et al. 2005). Several significant boundaries are found among northern and southern Han populations according to mtDNA haplogroups, the sharpest genetic contrast appearing along Huai River and Qin Mountain that are north to Yangtze River, and two others south to Yangtze River and north to Yellow River, respectively (Xue et al. 2008). On the other hand, no significant genetic barrier is found when both Han and non-Han populations are included. An automatic search for a genetic frontier has also been performed for HLA, with a similar result: a significant boundary emerges for HLA-A, -B, and -DRB1 only when Han populations alone are considered; in this case, the boundary appears near the Yangtze River (Di and Sanchez-Mazas 2011). Boundaries are also detected when using Y chromosome data but in a much more fragmented way with no significant uninterrupted barrier between north and south (Xue et al. 2008). Note that the significant level of genetic variation ($\Phi ct=0.16$) observed by Karafet et al. (2001) between SEAS and NEAS for the Y chromosome cannot be interpreted as a genetic boundary as it is the result of an analysis of molecular variance (AMOVA) with an a priori choice of the groups compared. Through genome-wide association studies, population substructure of Han populations into north, central and south subgroups is observed but with very small levels of genetic differentiation (Xu et al. 2009), thus resembling much more a continuous pattern. Such continuity has clearly been put forward by Chen et al. (2009) using over 350,000 genome-wide autosomal SNPs in over 6,000 Han Chinese samples from ten provinces of China.

Thus, except in some cases where only Han populations are considered, the genetic pattern of East Asian populations is definitely not a sharp bipartite subdivision. On the contrary, many studies indicate the existence of genetic clines along the latitude. In addition to multidimensional scaling and/or spatial autocorrelation analyses, continuous patterns of gene frequencies have been evidenced by their correlation with latitude for several genetic markers, e.g., HLA (Di D, and Sanchez-Mazas A. 2011 In prep. The north-south differentiation of East Asian populations) and autosomal SNPs (Abdulla et al. 2009; Suo et al. 2011). Frequency clines were also observed for classical markers, e.g., RH and GM haplotypes (Poloni et al. 2005; Sanchez-Mazas 2008) and mtDNA haplogroups like F1, B, and D4 (Yao et al. 2002).

North and south substructures

This continuous pattern of genetic differentiations in East Asia is also characterized by changes in the levels of genetic diversity and substructure between NEAS and SEAS, although with a hard dispute between different authors. On the basis of 19 NRY markers, Su et al. (1999) claim that SEAS are more diversified than NEAS, while Karafet et al. (2001) sustain the opposite view on the basis of 52 NRY markers. Actually, Su et al. (1999) used more than two times as many SEAS (20) than NEAS (9); moreover, sample sizes are very small (less than 30 individuals) except in two cases ($N=82$ for one northern Han and $N=280$ for one southern Han population), which is clearly a source of bias: in this study, the number of haplotypes detected is highly significantly correlated to the number of individuals tested (Fig. 1).

Shi et al. (2005) reanalyzed Y chromosome markers and sustained Su et al.'s (1999) conclusions. However, because, as they say, *~80% of the Chinese ethnic populations live in southern regions with inhabitation histories longer than 3,000 years*, (Wang 1994 cited by Shi et al. 2005), populations from southern regions of China were overrepresented in their study: on the contrary, Hui, Uygur, and Mongolian, which represented northern populations, were removed from the analyses because they were considered as recently established (<1,000 years ago) with extensive admixture with European and Central Asian populations (Wang 1994 cited by Shi et al. 2005). Likewise, the short tandem repeat (STR) network shown in this work was built after excluding the Tibeto-Burman, Altaic, Hmong-Mien, and southern Han populations *to remove the influence of relatively recent population admixture*; the network then showed that the major STR haplotypes occurred in southern populations (Daic and Austroasiatic), leading them to support a southern origin of the O3-M122 lineage in East Asia, with an age of about 25,000 to 30,000 years (see
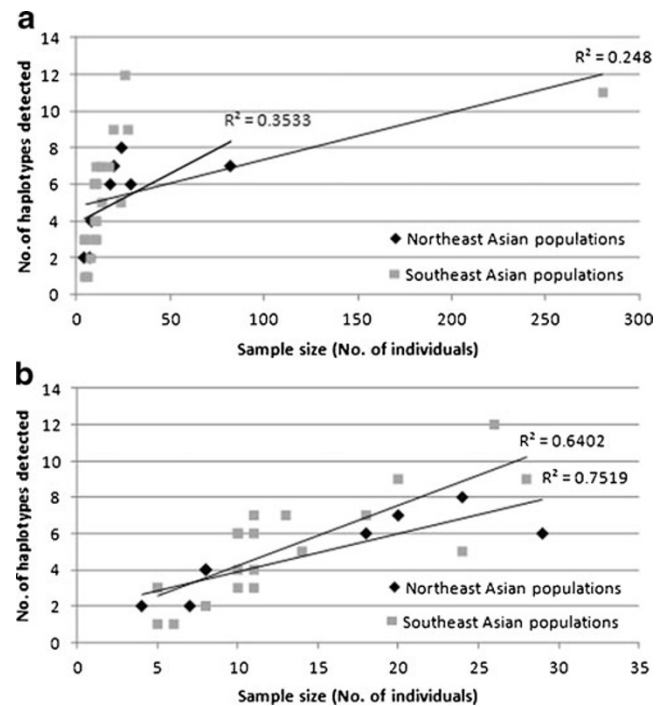
Fig. 1 Graphs showing a high and significant correlation (indicated by the coefficient of determination R^2) between the number of haplotypes detected and the number of individuals tested for the set of population samples analyzed by Su et al. (1999) for Y chromosome markers. *Top*: all population samples; *bottom*: after removing the two samples with highest sample sizes (82 and 280 individuals, respectively).

below). Similar methodological problems related to sampling occurred in the study of Abdulla et al. (2009) based on autosomal SNPs; the authors found that haplotype diversity was strongly correlated with latitude ($R^2=0.91$, $P<0.0001$), with genetic diversity decreasing from south to north. However, this result was obtained through an analysis of 10 "combined" populations (1, Indonesian; 2, Malay; 3, Philippine; 4, Thai; 5, Southern Chinese minorities; 6, Southern Han Chinese; 7, Japanese and Korean; 8, Northern Han Chinese; 9, Northern Chinese minorities; and 10, Yakut), with Yakut as the only population representing Altaic in the north, and several mixed population samples representing southern populations, the genetic diversity of which being then probably inflated. By using a better population sampling, Xue et al. (2006) found a higher STR diversity in the north than in the south, a finding that is not easily reconciled with a largely or exclusively southern origin for the northern populations. More recently, Zhong et al. (2011) also suggested that some Y chromosome haplogroups were introduced in East Asia through postglacial colonization (around 18,000 years ago) from West Asia via a northern route. Our recent analysis of the HLA polymorphism in about 127,000 individuals of 84 populations also indicate that when NEAS are accurately represented with no

preliminary exclusion of particular population samples, NEAS exhibit a higher level of internal diversity than SEAS, in agreement with Karafet et al. (2001) and Xue et al. (2006) for the Y chromosome and with our own observations for classical markers (not shown). In agreement with Zhong et al. (2011), this greater diversity is due in part to alleles and/or lineages that are also observed in Central Asians (CAS), West Asian, and Europeans, whereas SEAS exhibit many lineages that are more specifically represented in East Asia. Then, when only these "East Asian specific" markers are considered, southern populations are more diversified than northern populations. We conclude that the apparent discrepancies between the studies described above are due to differences in either the sets of populations represented or the markers or lineages considered for the analyses. Overall, when all populations and all markers are used, genetic diversity tends to be higher in the north. Now, to interpret these results (see next section), we have to bear in mind that a high level of genetic diversity is not synonymous of an old population origin or differentiation but may also result from a greater permeability to gene flow from genetically diverse populations.

This might have been the case for NEAS. Studies performed on different genetic markers are indeed congruent in showing a genetic relationship with a rather continuous pattern between NEAS and CAS, whereas SEAS are more peculiar in relation to other geographic areas. The link between NEAS and CAS is first illustrated by the more widespread distribution of some alleles/lineages in NEAS, as described above. AMOVA analyses are also relevant: for the Y chromosome data analyzed by Karafet et al. (2001), the among-group variance component between CAS and NEAS is not statistically significant (Φct=0.04), whereas the highest value is found between SEAS and CAS (Φct=0.28), followed by the value between SEAS and NEAS (Φct=0.16). Clinal variation was observed at classical markers by Barbujani et al. (1994) among Altaic populations extending over a large area encompassing CAS and NEAS, and by Karafet et al. (2001) in NEAS, while random genetic variation is found among SEAS, even at small geographic distances. When considering pairwise differences among Y chromosome haplogroups, larger values are found within NEAS populations, whereas 85% of SEAS Y chromosomes belong to a few closely related haplogroups (e.g., M175); such a set of highly divergent haplogroups observed in the north may reflect greater contributions from different populations. Altogether, these results indicate that the genetic pool of NEAS is related to that of CAS and exhibits signatures of gene flow from multiple sources, while that of SEAS indicates greater isolation and population subdivision, although with a low level of differentiation among populations. Another crucial result related to these observations is the very high level of genetic diversity

found in Central Asian populations (Comas et al. 1998, 2004; Hammer et al. 2001; Quintana-Murci et al. 2004; Wells et al. 2001; Zerjal et al. 2002), compatible with their connection to NEAS through gene flow, Central Asia being considered either as a source (Wells et al. 2001) or as a receiver (Comas et al. 1998, 2004; Quintana-Murci et al. 2004; Zerjal et al. 2002) of human migrations.

Genetic variation in relation to linguistic diversity

Current methods in population genetics are not very powerful in discriminating between geography and linguistics to explain the observed genetic patterns in East Asia, most of all because the distribution of linguistic families itself is geographically structured (Blench et al. 2008); actually, East Asian populations tend to display genetic similarities according both to their linguistic relatedness and geographic proximity (Poloni et al. 2005; Sanchez-Mazas et al. 2005). However, some peculiar results emerge depending on each linguistic family.

– Altaic: Altaic-proper (Altaic hereafter), Korean and Japanese generally segregate together at one end of multivariate analyses performed on East Asian populations. However, Altaic populations differ significantly from Japanese and Koreans. A remarkable result is the very high level of internal genetic diversity within Altaic populations at HLA loci, while inter-population diversity (F_{ST}) is relatively low (Sanchez-Mazas et al. 2005). According to the predictions that we summarized in a previous paper (Sanchez-Mazas et al. 2005), these features suggest intensive gene flow after differentiation from a highly diversified population.

– Sino-Tibetan: for HLA, both Sinitic (Han) and Tibeto-Burman populations are geographically structured. Han populations are less diversified (lower F_{ST}) than Tibeto-Burman but a significant genetic boundary is found between northern (mostly Mandarin-speakers) and southern (mostly speakers of Southern languages) Han populations (Di and Sanchez-Mazas 2011; Poloni et al. 2005), although Mandarin populations from Southwest China show smaller genetic distances to SEAS than to NEAS. According to Wen et al. (2004a), northern Han and southern Han also differ significantly for their maternal mtDNA lineages (F_{ST}=0.006, $P<10^{-5}$) but not for their paternal Y chromosome lineages (F_{ST}=0.006, $P>0.05$). Northern Tibeto-Burman (Tujia and populations from Tibetan Plateau, i.e., Tibetan, Monba, Luoba, and Lachung) differ from southern Tibeto-Burman (mainly from Yunnan) for HLA. For mtDNA and the Y chromosome, a sex-biased pattern is also observed for Tibeto-Burman (Wen et al. 2004b).

– Tai-Kadai and Hmong-Mien from East and Southeast Asia: inter-population diversity is much higher in both Tai-Kadai and Hmong-Mien than in Han according to the Y chromosome, and in Hmong-Mien for mtDNA (Wen et al. 2005). For HLA, the populations speaking languages of these three linguistic phyla are related genetically to each other and generally exhibit a low level of internal diversity (except, for example, Thai and Kinh) (Di and Sanchez-Mazas 2011). They are also very close to each other according to GM and mtDNA, but more differentiated according to the Y chromosome (Poloni et al. 2005; Sanchez-Mazas 2008; Wen et al. 2004a). A PCA performed at the individual level on the basis of genome-wide autosomal SNPs discriminates relatively well the speakers of these linguistic families (Abdulla et al. 2009).

– Austroasiatic deserves a specific attention as this linguistic family is widely distributed between Northeast India (in addition to a few other Indian regions like Madhya Pradesh) and Southeast Asia. Populations speaking languages of different Austroasiatic branches are well differentiated from each other for mtDNA, with a pronounced differentiation of the Indian Munda which are genetically close to surrounding populations in India (Reddy and Kumar 2008). According to HLA, the Munda exhibit a unique genetic profile with a rather low level of polymorphism (Riccio et al. 2011). However, they share common genetic features with non-Austroasiatic populations in India (at all HLA loci), but also a few characteristics with Austroasiatic populations from Southeast Asia. The analysis of Y chromosome markers indicates a high frequency of haplogroup M95 (O2a) in Austroasiatic populations including the Munda (Chaubey et al. 2011; Kumar et al. 2007; Sengupta et al. 2006; Thangaraj et al. 2003). However, based on different levels of haplotypic

diversity, several independent genetic studies present opposite views regarding the geographic origin of this haplotype, which has been suggested either in India (Basu et al. 2003; Kumar et al. 2007) or in Southeast Asia (Chaubey et al. 2011; Sahoo et al. 2006; Sengupta et al. 2006).

Times estimation

A key issue in the reconstruction of human peopling history is dating the events related to past human migrations. Population genetics is rather limited in this field compared to archaeology and paleoanthropology which can provide direct absolute dates of human settlements, although with large confidence intervals. However, the genetic literature is full of absolute dates for human prehistory. The main reason is that geneticists use the molecular clock theory to infer the time to the most recent common ancestor (TMRCA) of each set of haplotypes clustered together (i.e., each haplogroup) in the molecular phylogenies obtained for uniparentally inherited mtDNA and Y chromosome genetic markers. Depending on differences in frequencies or molecular diversity, a geographic origin and a geographic spread are also often inferred for each lineage (hence the term "phylogeography" for this kind of approach).

Molecular dating may nevertheless result in contrasting estimations, as illustrated by some examples given below (see also Fig. 2):

– According to Yao et al. (2002a), most mtDNA lineages are very ancient in East Asia, with an age greater than 50,000 years, the oldest ones being most frequent in the south (81,000 and 75,000 years for R9 and B, respectively). Thangaraj et al. (2006) also estimate an old age of 46,300 (±10,900) years for the mtDNA

Fig. 2 Some examples of contrasting results for the dating of mtDNA and Y chromosome lineages (see text). EA East Asian, ky kilo-years, BP before present; confidence intervals are indicated within brackets.
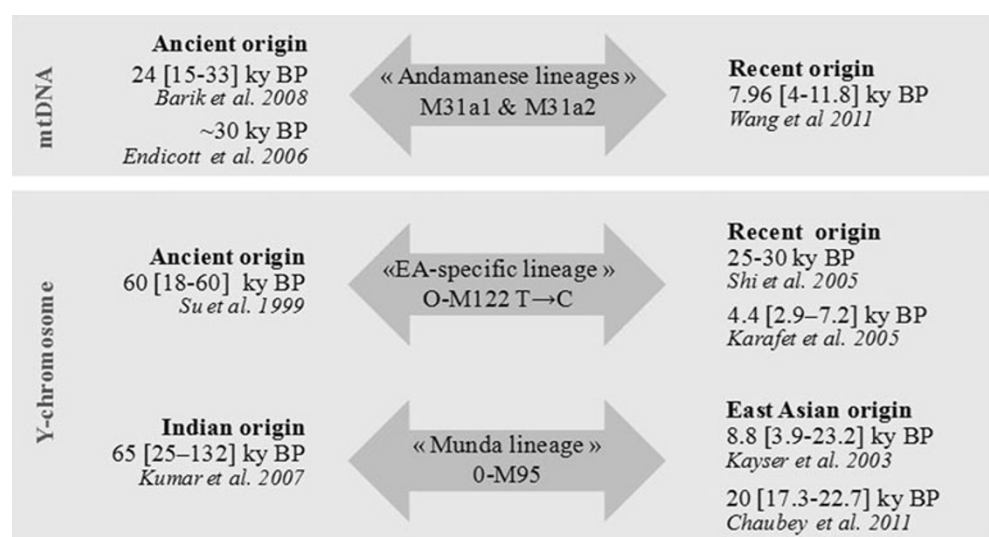
lineage M31 observed in the Andaman Isles (Bay of Bengal). M31 would be derived from lineage M (TMRCA of 63,000 years) predominant in Eurasia, itself derived from haplogroup L3 which is believed to originate in Africa 84,000 years ago. Based on these estimations, a rapid coastal dispersal from ~65,000 years ago around the Indian littoral is suggested (Macaulay et al. 2005). However, Barik et al. (2008) estimate a recent date for Andamanese-specific lineages M31a2 (<12,000 years) and only 24,000 years for lineages shared between Andamanese and Indian populations (M31a). Based on multiplex SNP mtDNA typing, Endicott et al. (2006) also find a coalescent date of about 30,000 years before present for the M31a mtDNA lineage shared by populations of the Andaman islands and the Indian sub-continent. Moreover, by updating the M31 phylogenetic tree, a much younger date was recently estimated by Wang et al. (2011) for M31a1 (−7,960±3,910 years), suggesting that Andamanese arrived from Southeast Asia across a land-bridge around the Last Glacial Maximum (LGM), but that this haplogroup originated in northeast India.

– On the basis of 19 Y chromosome biallelic loci, Su et al. (1999) estimate an age of 18,000 to 60,000 years for the O-M122T→C mutation shared by "Asian-specific" haplotypes H6–H8. However, using both morphological data analyzed by Turner (1993) and archaeological evidence for early settlements in Siberia (Vasil'ev 1993) and New Guinea (Brown et al. 1992; Swisher et al. 1996), they retain the upper boundary of 60,000 years for a bottleneck event leading to the entrance of modern humans into eastern Asia through a southern route. In contrast with this conclusion, Shi et al. (2008, 2005) estimate an older northward expansion of Y chromosome haplogroup D-M174 (60,000 years ago) than the above-mentioned O-M122 haplogroup (25–30,000 years ago), after an origin in southern East Asia. An even younger estimate of 4,400 years before present (BP) was obtained for O-M122 in Balinese populations (Karafet et al. 2005).

– As the Munda exhibit a high frequency and diversity of Y chromosome M95 (O2a) haplotypes (Karafet et al. 2001; Kumar et al. 2007; Reddy and Kumar 2008; Sengupta et al. 2006; Su et al. 2000, 1999), the origin of the Austroasiatic phylum has been claimed to occur in India around 65,000 years BP according to the age estimated for this haplogroup (Kumar et al. 2007), by contrast to the young age of 8,800 years previously given by Kayser et al. (2003). More recently, an age of about 20,000 years has been established for O-M95, resulting in an opposite interpretation: Austroasiatic populations would have a Southeast Asian origin, and those migrating to Northeast India would have exten-

sively admixed with Indian populations (Chaubey et al. 2011). This latter view is in close agreement with our own results based on the HLA polymorphism (Riccio et al. 2011).

Besides phylogeography, another approach based on genetics to date past events is to estimate population expansion times. This may be performed by using either distributions of pairwise nucleotidic differences among DNA sequences ("mismatch" distributions) assuming the infinite site mutation model (e.g., in the case of mtDNA sequences) or other specific estimators (e.g., the variance in repeat length in the case of STR):

– Asian populations show signals of Pleistocene expansions about 70,000 years before present (73,000 with 95% confidence interval of 46,000–87,000 years) according to the mtDNA mismatch distributions analyzed by Excoffier and Schneider (1999), although the heterogeneous composition of the "Asian" sample used in this study may have inflated the estimated date. Different expansion times were obtained by Chaix et al. (2008) depending on the mutation rates used for the analyses: when using pedigree-based mutation rates, the authors find expansion times in East Asia of about 29,000–30,000 years for mtDNA and 14,000–19,000 years for the Y chromosome; when using phylogeny-based mutation rates, they obtain 61,000–63,000 years for mtDNA and 31,000–40,000 years for the Y chromosome. In Central Asia, similar expansion times are found for the Y chromosome (16,000 and 36,000 years, depending on the model), while slightly younger dates are obtained for mtDNA (26,000 and 54,000 years).

– East Asian male demographic history has also been investigated by Xue et al. (2006) through a Bayesian full-likelihood analysis to data from 988 men representing 27 populations from China, Mongolia, Korea, and Japan typed with 45 SNPs and 16 STR markers from the Y chromosome. The authors showed that the northern populations started to expand in number between 34,000 and 22,000 years ago, thus before the LGM, while the southern populations did so between 18,000 and 12,000 years ago, but then grew faster.

Interpretation of genetic results and methodological issues

The genetic results described above have been the subject of multiple interpretations on the peopling history of East Asia. A long-standing debate is that of the first arrival of modern humans in East Asia after their expansion out of Africa, either through a single southern route towards

Southeast Asia with later migrations towards the north (Chu et al. 1998; Shi et al. 2005, 2008; Su et al. 1999), or through two independent routes, a southern and a northern, with later bi-directional migrations and admixture in East Asia (pincer and overlapping models) (Cavalli-Sforza et al. 1994; Di and Sanchez-Mazas 2011; Ding et al. 2000; Karafet et al. 2001; Xiao et al. 2000; Zhong et al. 2011). Two kinds of genetic arguments have been used to sustain the first hypothesis: a very old age for the M lineage in Southeast Asia and its derivatives M31 and M32 in the Andaman islands, and a greater genetic diversity in SEAS, as compared to NEAS.

However, none of these arguments constitutes a definitive proof. Firstly, as described above, very different TMRCAs were obtained for M lineages (Fig. 2). Also, TMRCA estimates often display very large confidence intervals and, unfortunately, non-genetic (e.g., archaeological) data are sometimes used to adopt a final estimation closer to the upper or lower bond of the interval, as exemplified above for the age of the O-M122 T→C allele (Fig. 2). Secondly, we have shown that the heterogeneity of the sample sets used in different studies may explain contradictory results concerning the level of genetic diversity of northern and southern populations, respectively, with crucial consequences in this debate; in the two examples described above, the study design (in this case, the choice of the samples, where many northeast Asian populations were excluded) was built according to non-genetic (e.g., historical) information, thus matching the expected result, i.e., the identification of the most ancient

layer of human migrations in East Asia, which was then taken as the unique migration event (the southern route). Actually, northern populations are found to be genetically more diverse than southern populations, which of course does not mean that the peopling of Northeast Asia was more ancient. Based on different genetic evidence, it merely seems that this diversity reflects a network-like genetic structure of northeast Asian populations in relation to Central Asian populations, while Southeast Asia would have remained more isolated. A pincer or overlapping model (Di and Sanchez-Mazas 2011) suggesting independent migrations along a southern and a northern route, yet at distinct prehistoric periods (i.e., Paleolithic and post-glacial periods, respectively) is more compatible with the observed data.

A remarkable result of most studies cited above is the very old dates inferred from phylogenetic studies for some molecular lineages, although the estimated dates strongly depend upon the method used (Blench et al. 2008). Given the estimated time ranges, the oldest dates inferred from genetic studies are compatible with old settlements of modern humans in East Asia attested by fossil or archaeological remains (Fig. 3) (Chen et al. 1989; Mijares et al. 2010; Shang et al. 2007; Shen et al. 2002; Sun et al. 2000; Vasil'ev 1993; Wu et al. 2006). However, ancient molecular lineages may just represent limited heritages of undefined ancestral populations rather than real indications on the origin and migration history of present populations. This illustrates very well one of the main methodological problems discussed by Blench et al. (2008), the fact that

Fig. 3 Possible route(s) of modern human migrations towards East Asia according to different hypotheses proposed by geneticists ("pincer" or "overlapping" model: both the northern and the southern routes; "southern origin" model: only the southern route), along with representative archaeological sites during the critical period (100,000–20,000 BP), knowing that the shallow parts of the sea (light gray/blue on the map) were postulated as land area with the lower sea level of last ice age (Sun et al. 2000). References for the archaeological sites: Mal'ta: Vasil'ev (1993); Upper Cave: Chen et al. (1989); Tianyuan Cave: Shang et al. (2007); Huanglong Cave: Wu et al. (2006); Liujiang: Shen et al. (2002); Callao Cave: Mijares et al. (2010).
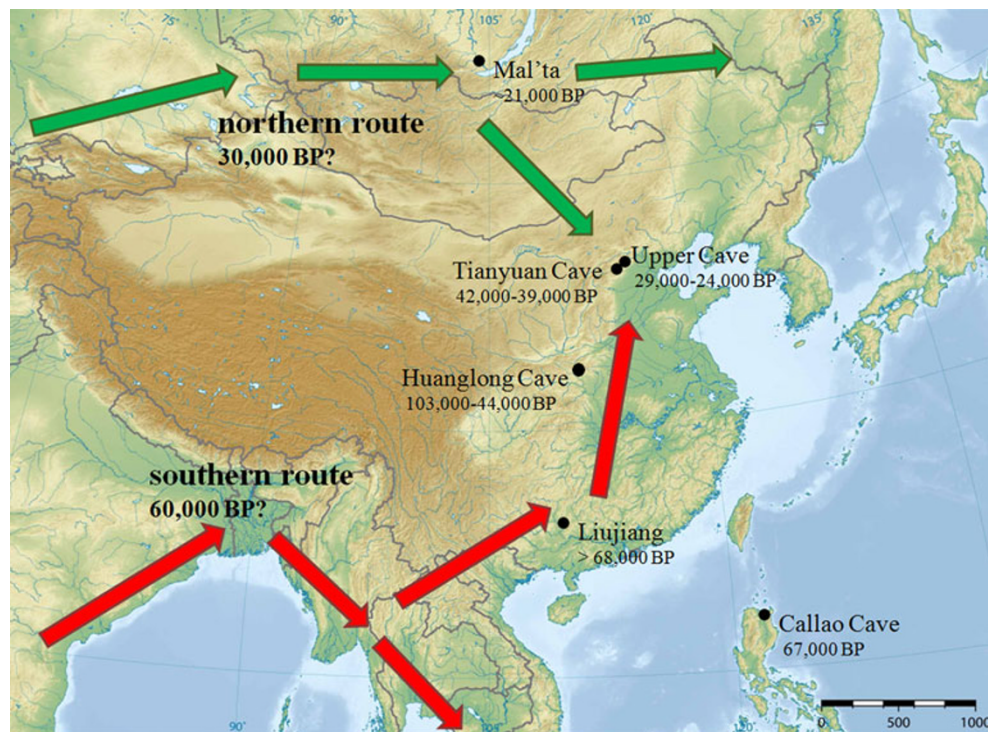
genetic tree nodes do not correspond to identifiable events in population history, and are generally older than population events. Then such dates might not be useful to depict extensive human migrations like those occurring in the Neolithic. This period was probably characterized by wide demographic expansions, long-range migrations and recurrent gene flow between neighboring populations, and other kinds of genetic signals should be explored. This is why specific approaches capable of detecting demographic expansions have been used. Here again, however, very old dates have been inferred, i.e., corresponding to Paleolithic times (~70,000 years) or to different periods predating or closely following the LGM, but, in any case, older than the Neolithic. Such signals may correspond, respectively, to the first expansion of modern humans throughout the world and to postglacial recolonizations, while more recent events would not be easily disentangled by using such approaches. Note, however, that older signals of population expansions are detected for northern Asian populations (Xue et al. 2006), and also for paternally rather than maternally inherited markers (Chaix et al. 2008). These results may be relevant for further inter-disciplinary studies.

Descriptive approaches like PCA, MDS, spatial autocorrelation analyses aiming at detecting genetic clines, specific statistical analyses used to identify genetic boundaries, as well as correlation analyses allowing the comparison of genetic variation with either geographic or linguistic data are still very useful to understand how the current genetic pool of East Asian populations is structured. We have stressed the fact that the identification of genetic boundaries is highly dependent upon the sample set available for the analyses; that is, significant genetic barriers are susceptible to be detected through uneven sampling along genetic clines! However, we may conclude from the different studies cited above that the north to south continuous genetic pattern observed in East Asia crosses a region of sharper variation around the Yangtze or Huai Rivers; actually, as this boundary appears to be significant only when Han populations are considered, it may correspond to a recent (<1,500 years) linguistic subdivision
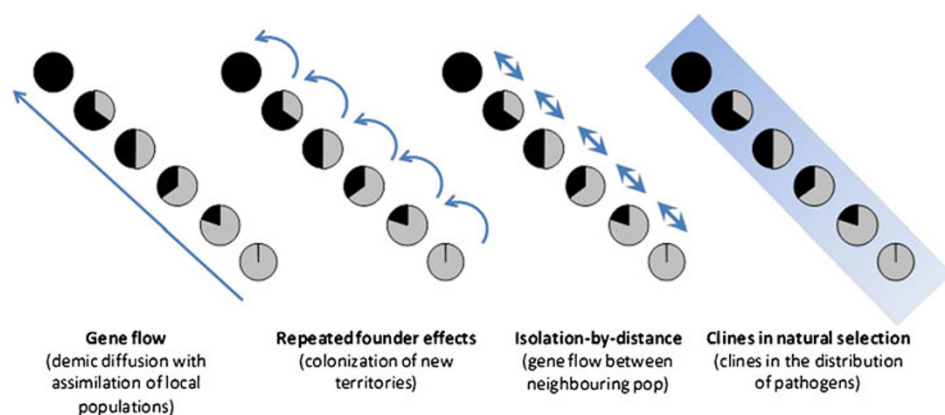
between Mandarin and southern Chinese speakers, as proposed by Sagart et al. (2005b)

The overall pattern of genetic variation in East Asia is yet continuous, characterized by many genetic clines along the latitude, and, to a lesser extent, along the longitude between Central and Northeast Asia. It is tempting, of course, to relate those clines to the expansion of specific linguistic families: e.g., Altaic, to explain the continuous pattern between CAS and NEAS; and Sino-Tibetan, from the Yellow River in the North to southwest and southern regions, corresponding to the expansion of Tibeto-Burman and Sinitic-speakers, respectively. However, to interpret genetic clines is delicate as such clines may be explained by very different mechanisms (Fig. 4): demic diffusion with admixture between genetically distinct populations (Ammerman and Cavalli-Sforza 1984), serial founder effects (Deshpande et al. 2009), isolation-by-distance where gene flow happens between neighboring populations (Novembre and Stephens 2008; Reich et al. 2008), or even differential adaptation to distinct environments, including varying prevalence of infectious diseases (Suo et al. 2011). In the current state of research, no definitive conclusion on the genetic clines observed in East Asia has been reached. This issue is yet crucial to understand the demographic impact of Neolithic migrations like those probably related to the expansion of linguistic families and/or rice and millet domestication in East Asia. Also, specific models have to be considered to investigate the expansion patterns of discontinuously dispersed linguistic families like Austroasiatic. Although several independent genetic studies sustain a Southeast Asian origin of this family, with later migration to India where populations underwent intensive gene flow, the evidence is still weak as no signals of such scenario have been found for mtDNA.

Conclusion and perspectives

We have presented a brief summary of current genetic evidence on the peopling history of East Asia by

Fig. 4 Four different situations generating genetic clines (see text).

Gene flow (demic diffusion with assimilation of local populations)

Repeated founder effects (colonization of new territories)

Isolation-by-distance (gene flow between neighbouring pop)

Clines in natural selection (clines in the distribution of pathogens)

dissociating some raw genetic results from their interpretation, and by pointing out some important methodological problems. A recurrent problem in all kinds of genetic studies is of course insufficiency in the set of population samples analyzed and one should be aware that this may have crucial consequences on the interpretation of the results. Also, misinterpretation may be due to ascertainment bias in the choice of markers. Another critical issue is the interpretation of molecular phylogenies and TMRCAs; to our view, such genetic approaches are useful as long as they ask questions adapted to the data to which they apply, i.e., questions related to the genealogy of *molecules* and not to the history of *populations*.

Although the present paper is not supposed to present an exhaustive review of the literature on the subject, the main results that we have described above indicate a general lack of genetic evidence related to the expansion of the main linguistic families or the diffusion of farming in East Asia during the Neolithic. This is probably because specific hypotheses on these issues are generally not formulated a priori; rather, genetic analyses are performed on sets of available population samples and the results interpreted a posteriori in relation to other disciplines. A main pitfall is that several alternative explanations commonly match the genetic results; then, it is tempting to choose the one that corresponds better to the hypothesis defended a priori by the researcher. A more robust—or, at least, complementary—approach would be first to establish alternative scenarios of peopling history on the basis of different, non-genetic, disciplines, and then to test those scenarios by using genetic approaches.

Computer simulation studies are very appropriate in this respect and represent an interesting perspective, as they may even accommodate models where the genetic loci are submitted to natural selection (like HLA). This can be useful to test scenarios where not only demographic, but also environmental factors are taken into account. This approach has already provided relevant results to understand the peopling history of specific geographic regions (Currat et al. 2010) and is currently being applied to East Asia to test the "southern route" hypothesis versus the "pincer" or "overlapping" models (Di et al. (2011) In prep. *Testing the peopling history of East Asia through computer simulation*). However, it also needs a deep collaboration between scholars of different disciplines to establish the scenarios to be tested and to propose acceptable values for the parameters needed in the simulations, e.g., demographic parameters. This is our future goal to reconstruct the peopling history of East Asia, and we encourage researchers to participate.

References

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, et al. Mapping human genetic diversity in Asia. Science. 2009;326(5959):1541–5.

Ammerman A, Cavalli-Sforza L. The neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press; 1984.

Barbujani G, Pilastro A, De Dominico S, Renfrew C. Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. Paleolithic colonisation. Am J Phys Anthropol 1994;95(2):137-54.

Barik SS, Sahani R, Prasad BV, Endicott P, Metspalu M, Sarkar BN, et al. Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands. Am J Phys Anthropol. 2008;136(1):19–27.

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, et al. Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res. 2003;13(10):2277–90.

Blench R, Ross M, Sanchez-Mazas A. Introduction: methodological issues: linking genetic, linguistic and archaeological evidence. In: Sanchez-Mazas A, Blench R, Ross MD, Peiros I, Lin M, editors. Past human migrations in east asia. London: Routledge; 2008. p. 3–19.

Brown P. Recent human evolution in East Asia and Australia. Philos Trans R Soc. 1992;337:235–42.

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

Chaix R, Austerlitz F, Hegay T, Quintana-Murci L, Heyer E. Genetic traces of east-to-west human expansion waves in Eurasia. Am J Phys Anthropol. 2008;136(3):309–17.

Chaubey G, Metspalu M, Choi Y, Magi R, Romero IG, Soares P, et al. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol. 2011;28(2):1013–24.

Chen T, Hedges REM, Yuan Z. Accelerator radiocarbon dating for the upper cave of Zhoukoudian. Acta Anthropologica Sin. 1989;8:216–21.

Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet. 2009;85(6):775–85.

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA. 1998;95:11763–8.

Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Martinez-Arias R, et al. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. Am J Hum Genet. 1998;63(6):1824–38.

Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, Calafell F, et al. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. Eur J Hum Genet. 2004;12(6):495–504.

Currat M, Poloni ES, Sanchez-Mazas A. Human genetic differentiation across the Strait of Gibraltar. BMC Evol Biol. 2010;10:237.

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL. A serial founder effect model for human settlement out of Africa. Proc Biol Sci. 2009;276(1655):291–300.

Di D, Sanchez-Mazas A. Challenging views on the peopling history of East Asia: the story according to HLA markers. Am J Phys Anthropol. 2011;145:81–96.

Ding YC, Wooding S, Harpending HC, Chi HC, Li HP, Fu YX, et al. Population structure and history in East Asia. Proc Natl Acad Sci USA. 2000;97(25):14003–6.

Du RF, Xiao CJ, Cavalli-Sforza LL. Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. Sci Chi C Life Sci. 1997;40(6):613–21.

Endicott P, Metspalu M, Stringer C, Macaulay V, Cooper A, Sanchez JJ. Multiplexed SNP typing of ancient DNA clarifies the origin of Andaman mtDNA haplogroups amongst South Asian tribal populations. PLoS One. 2006;1:e81.

Excoffier L, Schneider S. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc Natl Acadf Sci USA. 1999;96:10597–602.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, et al. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 2001;18(7):1189–203.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA. 2011;108(13):5154–62.

Jin L, Su B. Natives or immigrants: modern human origin in east Asia. Nat Rev Genet. 2000;1(2):126–33.

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet. 2001;69(3):615–28.

Karafet TM, Lansing JS, Redd AJ, Reznikova S, Watkins JC, Surata SP, et al. Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. Hum Biol. 2005;77(1):93–114.

Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill P, Shen P, et al. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet. 2003;72(2):281–302.

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol. 2002;19(10):1737–51.

Kumar V, Reddy AN, Babu JP, Rao TN, Langstieh BT, Thangaraj K, et al. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol. 2007;7:47.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 2005;308 (5724):1034–6.

Mijares AS, Détroit F, Piper P, Grün R, Bellwood P, Aubert M, et al. New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. J Hum Evol. 2010;59:123–32.

Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40 (5):646–9.

Poloni ES, Sanchez-Mazas A, Jacques G, Sagart L. Comparing linguistic and genetic relationships among east Asian populations: a study of the RH and GM polymorphisms. In: Sagart L, Blench R, Sanchez-Mazas A, editors. The peopling of East Asia: putting together archaeology, linguistics and genetics. London: Routledge Curzon; 2005. p. 252–72.

Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, et al. Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet. 2004;74(5):827–45.

Reddy BM, Kumar V. Origins of the Austro-Asiatic populations, Encyclopedia of Life Sciences. Chichester: Wiley; 2008. p. 1–9.

Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet. 2008;40(5):491–2.

Riccio ME, Nunes JM, Rahal M, Kervaire B, Tiercy JM, Sanchez-Mazas A. The Austroasiatic Munda population from India and Its enigmatic origin: a HLA diversity study. Hum Biol. 2011;83 (3):405–35. doi:410.3378/3027.3083.0306.

Sagart L, Blench R, Sanchez-Mazas A, editors. The peopling of East Asia: putting together archaeology, linguistics and genetics. London: RoutledgeCurzon; 2005a. p. 323.

Sagart L, Sanchez-Mazas A, Poloni ES, Arredi B. The formation of East Asian language families. ESF EUROCORES workshop: languages and genes: recent work and emerging results. Aussois (France), September 22–25. 2005b; p 1–14.

Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, et al. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. Proc Natl Acad Sci USA. 2006;103(4):843–8.

Sanchez-Mazas A. Evidences génétiques du peuplement de l'Asie orientale. Arch Sci. 2008;61:53–68.

Sanchez-Mazas A, Poloni ES, Jacques G, Sagart L. HLA genetic diversity and linguistic variation in East Asia. In: Sagart L, Blench R, Sanchez-Mazas A, editors. The peopling of east asia: putting together archaeology, linguistics and genetics. London: Routledge Curzon; 2005. p. 273–96.

Sanchez-Mazas A, Blench R, Ross M, Peiros I, Lin M, editors. Past Human Migrations in East Asia and taiwan: matching genetics, linguistics and archaeology. London: Routledge; 2008.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, et al. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet. 2006;78(2):202–21.

Shang H, Tong H, Zhang S, Chen F, Trinkaus E. An early modern human from Tianyuan Cave, Zhoukoudian, China. Proc Natl Acad Sci USA. 2007;104:6573–8.

Shen G, Wang W, Wang Q, Zhao J, Collerson K, Zhou C, et al. U-Series dating of Liujiang hominid site in Guangxi, Southern China. J Hum Evol. 2002;43(6):817–29.

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, et al. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet. 2005;77 (3):408–19.

Shi H, Zhong H, Peng Y, Dong YL, Qi XB, Zhang F, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol. 2008;6:45.

Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Curr Biol. 2010;20(4):R188–93.

Stringer CB, Grun R, Schwarcz HP, Goldberg P. ESR dates for the hominid burial site of Es Skhul in Israel. Nature. 1989;338 (6218):756–8.

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, et al. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet. 1999;65(6):1718–24.

Su B, Jin L, Underhill P, Martinson J, Saha N, McGarvey ST, et al. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA. 2000;97(15):8225–8.

Sun X, Li X, Luo Y, Chen X. The vegetation and climate at the last glaciation on the emerged continental shelf of the South China Sea. Palaeogeogr Palaeoclimatol Palaeoecol. 2000;160:301–16.

Suo C, Xu H, Khor C-C, Ong RT, Sim X, Chen J, et al. Natural positive selection and north–south genetic diversity in East Asia. Europ J Hum Genet. 2011;2011:1–9.

Swisher III CC, Rick WJ, Anton SC, Schwarcz HP, Curtis GH, Suprijo A, Widiasmoro. Latest Homo erectus of Java: potential

contemporaneity with Homo sapiens in southeast Asia. Science 1996;274(5294):1870–4

Thangaraj K, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, et al. Genetic affinities of the andaman islanders, a vanishing human population. Curr Biol. 2003;13(2):86–93.

Thangaraj K, Chaubey G, Singh VK, Vanniarajan A, Thanseem I, Reddy AG, et al. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. BMC Genomics. 2006;7:151.

Turner CG. Shifting continuity: modern human origin. In: Brenner S, Hanihara K, editors. THe origin and past of modern humans as viewed from DNA. Singapore: World Scientific; 1993. p. 216-43.

Vasil'ev SA. The upper paleolithic of northern Asia. Curr Anthropol. 1993;34:82–92. 34:82–92.

Wang HW, Mitra B, Chaudhuri TK, Palanichamy MG, Kong QP, Zhang YP. Mitochondrial DNA evidence supports northeast Indian origin of the aboriginal Andamanese in the Late Paleolithic. J Genet Genomics. 2011;38(3):117–22.

Wang ZH. History of nationalities in China. Beijing:China Social Science Press; 1994.

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, et al. The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA. 2001;98(18):10244–9.

Wen B, Li H, Lu D, Song X, Zhang F, He Y, et al. Genetic evidence supports demic diffusion of Han culture. Nature. 2004a;431 (7006):302–5.

Wen B, Xie X, Gao S, Li H, Shi H, Song X, et al. Analyses of genetic structure of tibeto-burman populations reveals sex-biased admixture in southern tibeto-burmans. Am J Hum Genet. 2004b;74 (5):856–65.

Wen B, Li H, Gao S, Mao X, Gao Y, Li F, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol. 2005;22(3):725–34.

Wu X, Liu W, Gao X, Yin G. Huanglong cave, a new late pleistocene hominid site in Hubei Province, China. Chin Sci Bull. 2006;51:2493–9.

Xiao CJ, Cavalli-Sforza LL, Minch E, Du R. Geographic distribution maps of human genes in China. J Genet Genomics (Yi Chuan Xue Bao). 2000;27(1):1–6.

Xu S, Yin X, Li S, Jin W, Lou H, Yang L, et al. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet. 2009;85(6):762–74.

Xue F, Wang J, Hu P, Ma D, Liu J, Li G, et al. Identification of spatial genetic boundaries using a multifractal model in human population genetics. Hum Biol. 2005;77(5):577–617.

Xue Y, Zerjal T, Bao W, Zhu S, Shu Q, Xu J, et al. Male demography in East Asia: a north–south contrast in human population expansion times. Genetics. 2006;172(4):2431–9.

Xue F, Wang Y, Xu S, Zhang F, Wen B, Wu X, et al. A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. Eur J Hum Genet. 2008;16(6):705–17.

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet. 2002;70(3):635–51.

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. Am J Hum Genet. 2002;71(3):466–82.

Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, et al. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol. 2011;28(1):717–27.