Analysis of Paternal Lineages in the Chinese Populations of Macau and Shanghai

José Guilherme Moreira Alexandre
Mestrado em Genética Forense
Departamento de Biologia
2013

Orientador
Professora Doutora Maria João Prata Martins Ribeiro, Professora Associada c/ Agregação na Faculdade de Ciências da Universidade do Porto e Investigadora no Instituto de Patologia e Imunologia da Universidade do Porto

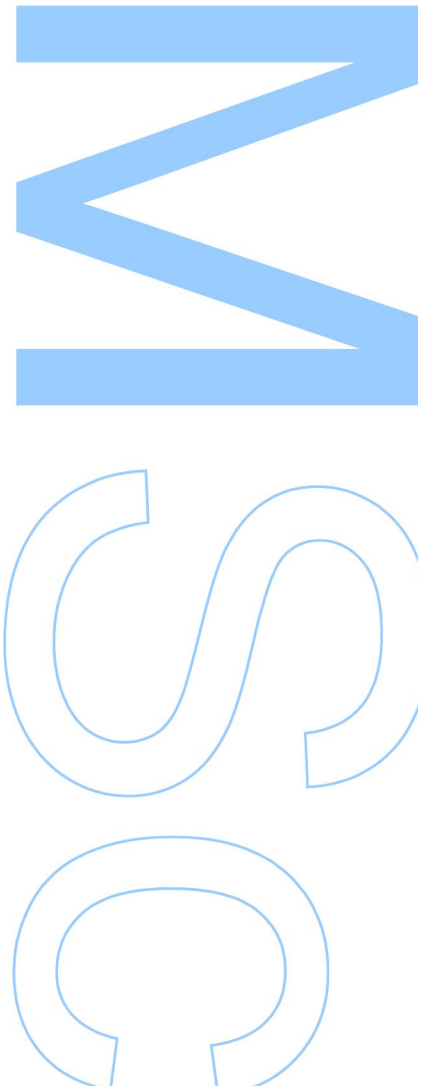
Coorientador
Doutor Luis Alvarez Fernandez, PhD, Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



José Guilherme Moreira Alexandre

Analysis of Paternal Lineages in the Chinese Populations of Macau and Shanghai

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Instituto de Patologia e Imunologia Molecular da Universidade do Porto

Faculdade de Ciências da Universidade do Porto

Mestrado em Genética Forense

2013

“If I have seen further, it is by standing on the shoulders of Giants”

Isaac Newton

Dissertação de candidatura ao grau de Mestre em Genética Forense submetida à Faculdade de Ciências da Universidade do Porto.

O presente trabalho foi desenvolvido no Instituto de Patologia e Imunologia da Universidade do Porto e sob orientação da Professora Doutora Maria João Prata Martins Ribeiro.

Dissertation for applying to a Master's Degree in Forensic Genetics submitted to the Faculty of Sciences of the University of Porto.

The present work was developed at the Institute of Molecular Pathology and Immunology of the University of Porto and under the scientific supervision of Professor Maria João Prata Martins Ribeiro.

AGRADECIMENTOS

À Professora Maria João Prata, minha orientadora, gostaria de agradecer em primeiro lugar, por toda a ajuda e incentivos que me deu ao longo do trabalho, com preciosas sugestões, e pela oportunidade que me deu de participar neste projecto.

Ao Luís, meu co-orientador, pela disponibilidade e boa disposição com que sempre me ajudou quando necessário, e pela contribuição que deu para a realização da tese.

Ao Professor António Amorim, pela oportunidade que me concedeu de frequenter este Mestrado e por tudo o que me ensinou nestes últimos dois anos.

Ao grupo de Genética Populacional da Universidade Complutense de Madrid, pela forma como me receberam e por todo o apoio que me prestaram, nomeadamente a Ana Lopez, o Eduardo Arroyo e o Qi Huang.

A todo o grupo de Genética Populacional, pela forma como me recebeu e por estar sempre disponível quando era necessário. À Cíntia, à Sofia Marques e à Sofia Quental.

A todos os meus colegas de Genética Forense, grupo com uma grande disponibilidade e entre ajuda, com quem foi um prazer partilhar estes dois últimos anos.

ABSTRACT

East Asia, the geographical region bordered by the Plateau in the southeast, by the Bering strait in the northeast, and extending into island southeast Asia is still a scarcely studied, with uncertainty surrounding the when and how the first settlements of humans reached the are, and especially the region now known as China.

Currently, 56 ethnic groups are recognized in China, though the concept of ethnicity is not necessarily correlated with genetic ancestry, since this often failed to recognize pre-existing anthropological and demographic realities.

In our work, we aimed to explore the genetic background of the Shanghai and Macau populations, in the broader context of population diversity in China, as well as to enrich the knowledge of the region, combining genetics with ethnic patterns, linguistic affiliation and historical records. Furthermore, we compared the results obtained in Shanghai and Macau, in order to evaluate whether signs were retained of the different geographical, historical and cultural factors that might have shaped the Y chromosome diversity in both populations, as well as China in general.

In order to do so, we have sampled genetic material from 135 unrelated males, 85 from Shanghai and 50 from Macau. Samples were tested for 25 Y-SNPs using two multiplex systems – Multiplex O and Multiplex 1 and in addition for the Yap element. They were also genotyped for 17 different Y-STR markers using the AmpF*ℓ*STR® Yfiler® PCR Amplification Kit (Applied Biosystems).

As expected, the large majority of the Y lineages in this study belonged to the O haplogroup (M175): 89.5% in Macau and 81.5% in Shanghai. Furthermore, high haplogroup diversity was shown to be high for the studied populations. The levels of haplotype diversity were similarly very high in Macau and Shanghai, a feature which was also shared by other Chinese populations used to perform comparisons. Among those, the Muslim and Tibetan populations presented, however, the lowest levels of diversity, probably due to the peculiarities of their history and religious beliefs that led to a certain isolation of these populations.

Although populations from China were shown to be clearly substructured, no correlation was found between geographical distance was and degree of genetic relatedness between populations.

Concerning the population from Macau, despite the long shared history with Portugal, no evidence was found testifying a male-mediated genetic influence of the Portuguese.

As for the Shanghai, its population revealed to be highly diverse, which may be a consequence of the metropolitan characteristics of Shanghai to where people since long were attracted to migrate.

Keywords: China, Ethnicity, Y-SNP, Y-STR

RESUMO

O Este Asiático, a região geográfica demarcada pelo planalto a sudoeste, pelo estreito de Bering a noroeste, é ainda actualmente uma zona escassamente estudada, persistindo grande incerteza em torno das migrações de humanos que chegaram à área, especialmente à região actualmente designada da China.

Correntemente, 56 grupos étnicos são reconhecidos na China, embora o conceito de etnicidade apenas esteja vagamente correlacionado com a ascendência genética, uma vez que esta designação falha em reconhecer realidades demográficas e antropológicas preexistentes.

Este trabalho tivemos como objectivo principal explorar o *background* genético das populações de Xangai e Macau, no contexto geral da diversidade populacional Chinesa. Tentamos também comparar os resultados obtidos de Xangai e Macau com outras populações Chinesas para tentar compreender as interacções entre geografia, história e cultura que podem ter moldado a diversidade do cromossoma Y em ambas as populações estudadas, assim como na China em geral.

Nesse sentido, 135 homens não aparentados foram estudados, 85 de Xangai e 50 de Macau. As amostras foram testadas para 25 SNPs do cromossoma Y usando dois sistemas de multiplex, Multiplex O e Multiplex 1, assim como para o elemento YAP. Foram também caracterizadas quanto a 17 Y-STRs diferentes, usando o Kit de Amplificação de PCR AmpFℓSTR® Yfiler® (Applied Biosystems).

Como esperado, a larga maioria das linhagens do cromossoma Y detectadas neste estudo pertenciam ao haplogrupo O (M175): 89.5% em Macau e 81.5% em Xangai. Ademais, elevados níveis de diversidade haplotípica foram encontrados em ambas as populações, o que também se verificou noutras populações chinesas usadas nos estudos comparativos. Entre estas, as populações Muçulmanas e Tibetanas apresentavam, no etanto, níveis de diversidade comparativamente mais baixos, o que deve ter explicação na história peculiar e crenças religiosas, que levaram a um certo nível de isolamento das mesmas.

Apesar de se ter detectado clara subestruturação entre as populações chinesas, não se encontrou relação significativa entre distância geográfica e grau de relacionamento genético entre populações.

Quanto à população de Macau, apesar do passado recente partilhado com Portugal; não se registaram sinais de influências genéticas mediadas por homens portugueses.

Sobre Xangai, de referir a elevada diversidade genética da amostra estudada, que pode ser consequente das características metropolitanas de Xangai, que desde há muito atraiu a ida de pessoas para a região.

Palavras-Chave: Etnia, China, Y-SNP, Y-STR

TABLE OF CONTENTS

TABLES

FIGURES

ABREVIATIONS

AMOVA	Analysis of Molecular VAriance
bp	base pair
BP	Before Present
ddNTP	dideoxyNucleotide-TriPhosphate
DNA	DeoxyriboNucleic Acid
Mb	Megabases
MDS	MultiDimensional Scaling
MNPD	Mean Number of Pairwise Differences
MSY	Male Specific region of the Y chromosome
NRY	NonRecombining region of the Y chromosome
PAGE	PolyAcrylamide Gel Electrophoresis
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
SBE	Single Base Extension
SMM	Stepwise Mutation Model
SNP	Single Nucleotide Polymorphism
SRY	Sex determining gene
STR	Short Tandem Repeat
YAP	Y *Alu* Polymorphic element
YBP	Years Before Present
YCC	Y Chromosome Consortium
YHRD	Y chromosome Haplotype Reference Database
Yp	Y chromosome short arm
Yq	Y chromosome long arm

INTRODUCTION

Y-Chromosome

"The properties of the Y chromosome read like a list of violations of the rulebook of human genetics. It is not essential for the life of an individual, males have it but females do well without it. One half consists of tandemly repeated satellite DNA and the rest carries few genes, and most of it does not recombine. However, it is because of this disregard for the rules that the Y chromosome is such a superb tool for investigating recent human evolution from a male perspective and has specialized, but important, roles in medical and forensic genetics." [1]

Origin/Structure

Nowadays, the two human sex chromosomes, X and Y, look very different from each other. Their origin, however, is believed to have been the same, evolving from a pair of homologous autosomes that followed two different evolutionary routes, around 240-320 million years before present. [1-3]

Evidence indicates that the X-Y differentiation occurred only after X-Y recombination ceased, as a consequence of successive chromosomal inversions. This was likely a stepwise process that started with the acquisition of the sex-determining gene (SRY) by one of the autosome pair. [1, 4]. This successive repression of recombination between the two chromosomes led to an expansion of the nonrecombining region (NRY), which now represents 95% of the chromosome.

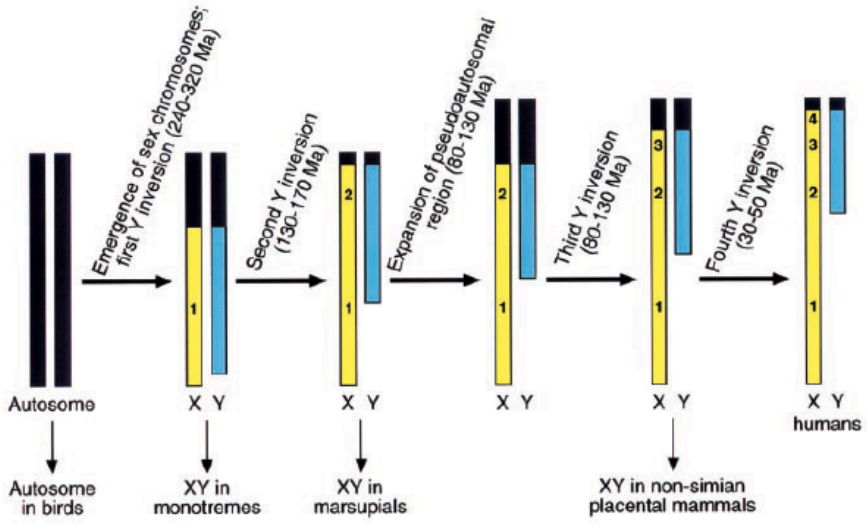


Figure 1 – Progressive X-Y differentiation, leading to loss in recombination from stratum 1 to strata 2, 3 and 4, four inversions are postulated. Each inversion reduced the size of the pseudoautosomal (X-Y recombining) region. [Adapted from Bruce T. Lahn and David C. Page, 1999]

The lack of recombination along with the action of other evolutionary forces led the Y chromosome to become one of the smallest chromosomes in size (60 million base pairs) and also in gene content, since the absence of recombination incapacitates the Y chromosome to eliminate deleterious mutations, a characteristic known as Muller's Ratchet. [2]

Although, as mentioned before, 95% of the chromosome does not recombine during male meiosis, X and Y chromosomes still undergo recombination in two specific regions called pseudoautosomal regions, regions that still behave as do autosomes. These two regions, located in each tip of the Y chromosome, are named PAR 1 - located on the short arm spanning approximately 2.6 Mb - and PAR 2 - located on the long arm and with approximately 0.32 Mb - and are fundamental in male meiosis, since they have a role in chromosome alignment (Figure 2).

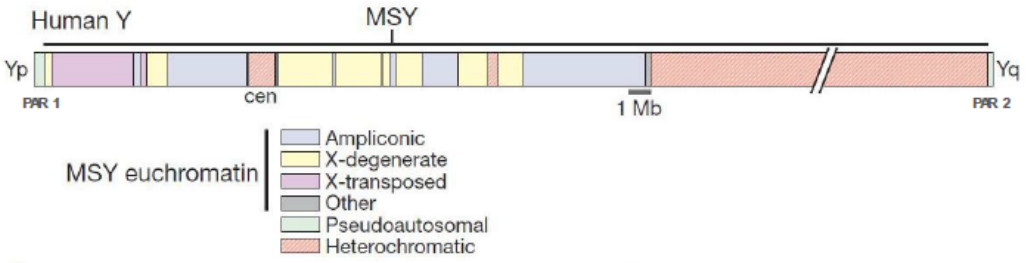


Figure 2 – Y chromosome structure. The Y chromosome is divided into a short (Yp) and a long arm (Yq), divided by a centromeric region (cen). Inside the MSY region can be distinguished the heterochromatic region and the eurochromatic region, further divided in X-degenerated, X-transposed and ampliconic regions. (Adapted from Hughes *et al.*, 2010)

Recently, the NRY has been renamed to male specific region (MSY). This happened because, even though X-Y recombination doesn't occur, abundant intrachromosomal recombination has been reported. The MSY contains two different regions, the heterochromatic and the euchromatic region. [5]

The heterochromatic region is composed mainly by long repetitive sequences that can extend for around 40 Mb. The euchromatic region includes all the genes identified in the Y-chromosome and ranges for about 23Mb, divided in both arms of the chromosome. This region includes the X-transposed sequence, originated from an X to Y transposition and therefore possesses great homology with the X chromosome, the X-degenerated region, a remnant of the ancient autosomes from

which the X and Y chromosomes evolved, and ampliconic sequences that allow intrachromosomal recombination due to their great similarity to other sequences in the MSY. [5, 6]

Uniparental Markers

The vast majority of the human genome is inherited from both parents and undergoes recombination. It suffers, therefore, a complex reshuffling process of the genetic information in each generation. There are, however, two exceptions, the Male Specific region of the Y-chromosome (MSY) and the mtDNA. These types of genetic systems, unlike the autosomes, are transmitted from generation to generation without variation, unless when mutation occurs (Figure 3). [7]

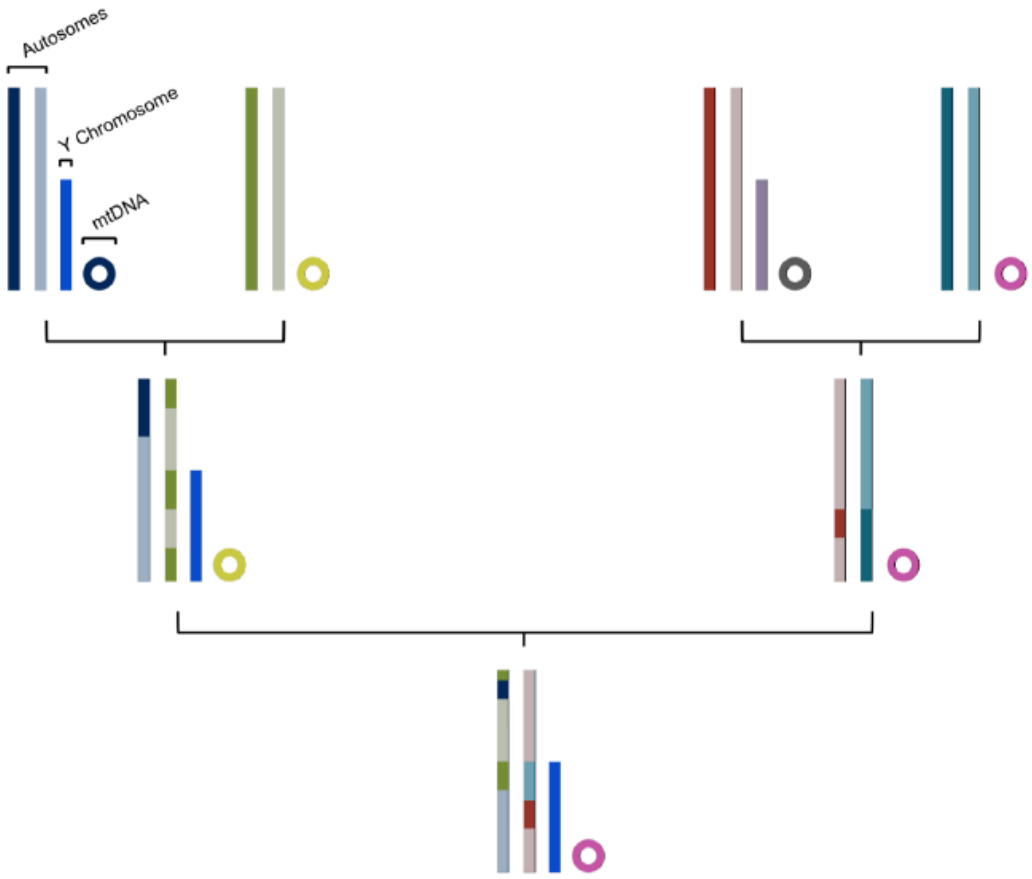


Figure 3 – Inheritance patterns of autosomes, Y chromosome and mtDNA. The mtDNA and MSY pass through generations escaping meiotic recombination, being transmitted as a haploid single block. Autosomes pass through generations suffering recombination.

Thus, since the Y chromosome is paternally inherited and the mtDNA is maternally inherited, uniparental markers are not appropriated for individual discrimination. They are however capable of defining male and female lineages, a characteristic that represents a noticeable advantage to population ancestry studies.

Y Chromosome: Genetic Markers

Different types of polymorphisms can be found in the Y chromosome, each one with particular characteristics. The two main classes are the bi-allelic markers, which define haplogroups, and the multi-allelic markers, that define haplotypes. One of their main differences is the mutation rate, which results in distinct characteristics that permits a wide-range of different analysis.

Bi-allelic markers (Y-SNPs)

Bi-allelic markers on the Y chromosome (Y-SNPs), include Single Nucleotide Polymorphisms (Y-SNPs) and an *Alu* element insertion (Y *Alu* Polymorphism – YAP). The *Alu* element was the first Y bi-allelic marker to be discovered, and since then a large amount of new polymorphisms has been described, given that SNPs are thereafter the most frequent class of polymorphisms in the human genome. [8] SNPs present a low frequency of mutation, around 10^{-8} per generation, and therefore most usually arise by unique events: when two alleles are present, it is accepted that the derived allele arose, as a unique event, from an ancestral form. [9]

In the forensic field SNPs have a number of characteristics that make them very appropriate for forensic studies, as the very low mutation rate and together with the short of DNA segments necessary to amply for their detection afford them desirable characteristic not only in the analysis of degraded samples, but also in general in the field of population genetics, due to their power to differentiate major human populations. Furthermore, Y chromosome haplogroups present a very well studied geographic distribution, revealing high population specificity, allowing to infer about the demographic events and evolutionary history of a population. [6, 10]

On the other hand, SNPs have some limitations. In order for their power of discrimination meet that of the nowadays-used STRs in the new multiplexes, 60 SNPs would be necessary. [6, 11, 12]

More than 600 bi allelic Y-chromosome markers have now been characterized and integrated in the phylogenetic tree of Y-haplogroups [13], defining 311 haplogroups. Nonetheless, new mutations are constantly being discovered and

await to be included in the YCC tree, which is being regularly updated. [14]

Multiallelic Polymorphisms (Y-STRs)

The human genome is overflowing with repetitive sequences, scattered all over the autosomes and sexual chromosomes. [9] Among these sequences, the Short Tandem Repeats (STRs) or Microsatellites are the most commonly used in Population and Forensic Genetics. STRs consist of 1 to 6 bp repetitions units, whose number varies between individuals. Repetitions from different STRs can be dealt together, since they define haplotypes, increasing this way the informative power.

These markers present higher mutation rates when compared to the SNPs, which underlies their common multiallelism. This characteristic gives the STRs a higher discrimination power between individuals, and therefore, can be used to differentiate Y-chromosome haplotypes with fairly high resolution due to their higher mutation rates. Furthermore, STRs typing usually requires simpler techniques than SNPs and can be simultaneous tested PCR using multiplex systems (Figure 4). Although the power of discrimination of each Y-SNP marker is lower than that of Y-STR markers, a broad combination of both types of markers contributes highly to increase the discrimination capacity. Y-SNP markers provide, however, an additional advantage over Y-STRs because they have more capacity for distinguishing human population groups and individuals within specific populations. [9, 15]

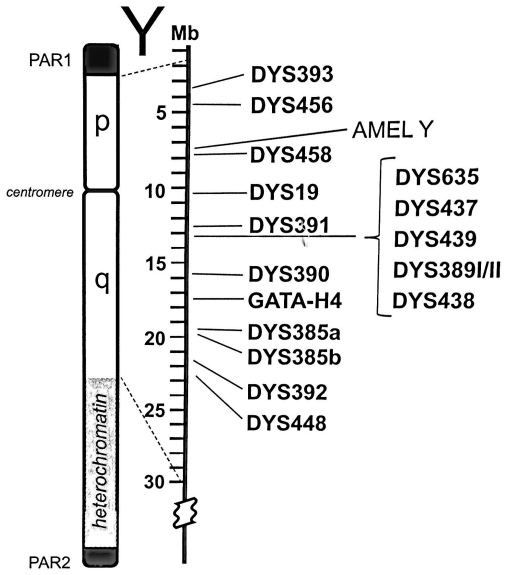


Figure 4 – Relative position of the 17 Y-STR loci commonly used in Y Chromosome testing, such as the commercial AmpF*ℓ*STR® Yfiler® PCR Amplification Kit (Applied Biosystems).

The stepwise mutation model (SMM) [16] is the generally accepted model to explain the formation of new STR alleles. This mutation model is a consequence of DNA replication slippage, where mutation events can add or remove one repeat unit (one step) to the length of the polymorphism. DNA stretch, STR mutation, however, depends on other factors, since rates generally get higher as the allele size increases and with higher number of uninterrupted repeats. [17, 18]

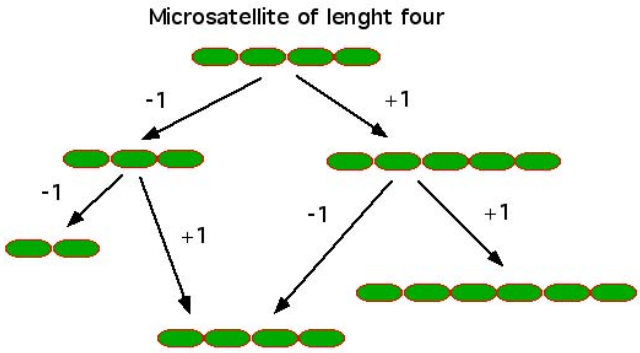


Figure 5 – Stepwise Mutation Model (SMM). A microsatellite, by the means of one step mutations, changes the number of repetitions, either by +1 or -1, depending on the pathway.

One of the advantages of STRs, as mentioned before, is the possibility to combine information from different markers to increase their discrimination power. Nevertheless, both in forensic and in evolutionary genetics it is also very important to be able to compare results from different studies and from different populations. Therefore, databases with information from Y-STRs haplotypes from different populations – such as the Y chromosome Haplotype Reference Database (YHRD) [19] - are of the most importance. YHRD is a reference anonymous database, developed for the forensic community, which contains information ranging from 7 to 23 STRs, the recommended minimum to the maximum number of Y-STRs currently found in commercial kits.

Y-Phylogeography

Phylogeography is a relatively new discipline that deals with the spatial arrangements of genetic lineages, especially within and among closely related species. [20] The discipline uses genetic data to clarify the history of population in combination with different fields of study that include extremely broad and diverse areas, such as History and Linguistics. The bridging between different approaches relies on the assumption that the present geographical distribution is based, not solely on the current biological and ecological factors, but also on past evolutionary and historical processes. [21]

In the most recent years, Phylogeography has taken important steps, mainly due to the recent advances in the study of genetic information and in improvement of analytical tools.. With the progresses in the study of Y-SNPs, patterns of similar sequences with a common ancestor – haplogroups – have been found, which present a remarkable specific geographic distribution (Figure 6).

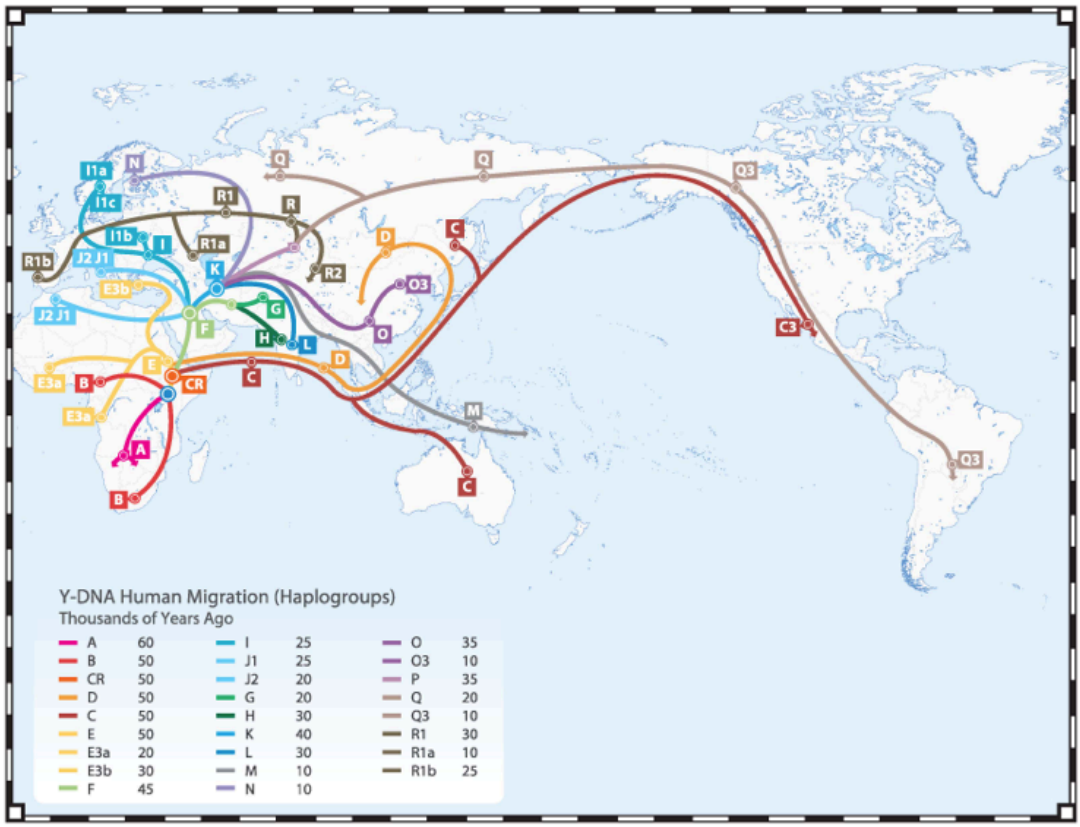


Figure 6 – Y chromosome haplogroups spatial distributions and phylogeographic relations. (Adapted from http://www.familytreedna.com)

Genetic drift, as mentioned before, is one of the fundamental mechanisms behind the populations' differentiation process. It involves random changes in the frequency of alleles/haplotypes owing to random sampling from one generation to the next. Since there is one Y chromosome to every four autosomes, the effective population size of the Y chromosome represents one quarter of that of any autosome. Assuming that the same mutational processes act more or less similarly on all chromosomes, it is therefore expected lower diversity on the Y chromosome than elsewhere in the nuclear genome, as indeed is observed. Y chromosome is, therefore, expected to be more susceptible to genetic drift, which accelerates the differentiation between groups of Y-chromosomes in different populations, but, for the same reason, the frequencies of haplotypes can change rapidly through times. [22]

These geographic patterns are also heavily influenced by social factors, as stated before. Approximately 70% of modern societies practice pratilocality, this is, if a man and a woman get together but they don't belong to the same place, it is the woman who tends to move, rather than the man. As a consequence, most men live closer to their birthplaces than do woman, and local differentiation of Y-chromosome is enhanced. By contrast, mtDNA, which is transmitted only by woman, is expected to show more reduced geographical clustering. [22]

The Y-chromosome became, therefore, a popular tool for tracing historical human migration patterns through male lineages. Anthropological, historical, and genealogical questions can be answered through Y-chromosome results (Figure 7).

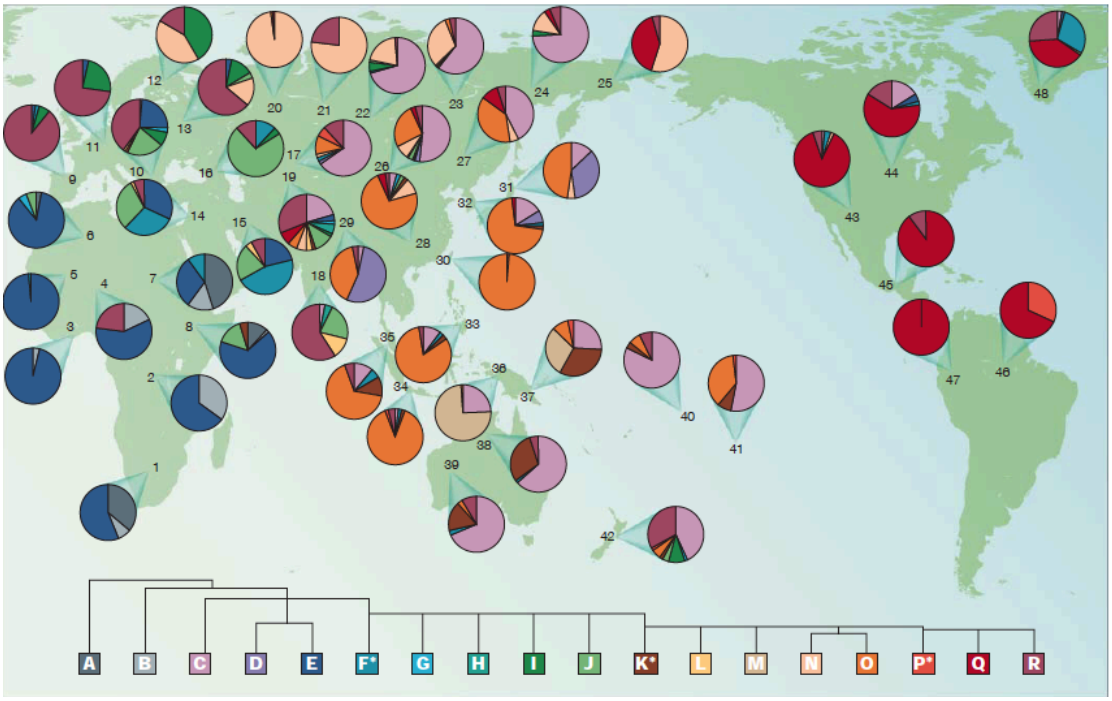


Figure 7 – Global distribution of Y haplogroups. Each circle represents a population sample with the frequency of the 18 main Y haplogroups identified by the Y Chromosome Consortium (YCC), indicated by the color sectors. Comparing the distribution of the different Y-chromosome haplogroups in different populations, it is noticeable a remarkable degree of population specificity. (Jobling *et al.*, 2003)

Applications

The unique characteristics of the Y-chromosome led it to become an important tool in many different fields, such as forensic evidence examination, paternity testing, medical and historical investigations, besides those more commonly addressed in the field of population genetics. [23]

With respect to the medicine field, the Y chromosome study has been used in order to understand genetic diseases, such as the male infertility, which might be caused by intra chromosomal recombination that leads to deletions in the MSY. [5]

In the scope of forensic studies, the Y chromosome can be very informative for some specific cases. The primary value of the Y chromosome in forensic DNA is that it is found only in males. The analysis of the Y chromosome is, therefore, becoming regular in cases such as sexual assault, where usually the aggressor is a man and the victim is a woman, having the sample contributions from both individuals. If more than one man perpetuates the sexual assault the number of contributors to the sample can also be inferred, only failing if a kinship between the

aggressors exists. [9, 24] Hence males commit the vast majority of violent crimes, a sample left in a crime scene by an inductee, and typed with Y-Chromosome specific markers, will likely be informative. [24] Y-STR analysis allows a more accurate and easier detection of the number of males in mixtures; multiple alleles at single copy locus give a good indication of the number of male contributors. The Y chromosome can also be used in investigations where the autosomal markers alone fail. [25, 26]

Equally, the Y-chromosome testing can have a decisive role in incomplete paternity testing, such as the cases when the alleged father is not available to be tested. When that is the case, his relatives belonging to the paternal lineage can be tested instead, since they will share the same profile, besides if mutation occurs. [9]

Population Genetics

The main goal of the Population Genetics is to study the differences between populations, at both intra and inter-population levels, seeking an understanding of the fundamental mechanisms behind their differentiation process.

These mechanisms include mutations, the primary source of diversity, and more or less complex evolutionary processes, like natural selection, genetic drift or gene flow (migration), all of which might be highly influenced by historical and demographic circumstances. Population genetics tries, therefore, to comprehend the driving forces behind the variation, and also to interpret how the accumulation of genetic diversity happened over time. [27, 28]

In the last few decades, Y chromosome analysis has become a crucial tool to investigate a number of questions involving the human origin and the history of human populations.

Origin of Modern Humans

Molecular anthropology is a fairly new field of science, which incorporates and combines data from broad areas, such as palaeontology, archaeology, history and geography, with genetic data and DNA analysis. [29]

During the last decades much interest has been devoted to the origin of the modern humans. From the combined analysis of genetic data and other fields of research, two major theories emerged as an attempt to explain the origin of *Homo sapiens*: the Out of Africa (OOA) and the Multiregional Evolution Hypothesis (MEH).

The OOA theory states that the *Homo sapiens* evolved in Africa, and later, around 50 to 200 thousand years before present (YBP), migrated from the African continent, replacing earlier hominid species such as the *Homo erectus*.

The other proposed theory, the MEH theory, reasons that the *Homo sapiens* evolved directly from archaic hominid species in several widely dispersed geographical locations. The nowadays homogeneity of the *Homo sapiens* species

would have arise from gene flow among the archaic populations, and from natural selection for favourable genes and characteristics. [29, 30]

Nowadays, however, the “Out of Africa” theory is the most consensually accepted, when it comes to explain the origin of the modern human being. According to this model the transition from *Homo sapiens* occurred less than 200,000 years BP, in Africa, and only after this transition the dispersal towards the rest of the world occurred, replacing the archaic human forms present in those areas. [31]

East Asia Framework

Once it became generally accepted that modern humans evolved recently in Africa, the times and routes of migration to East Asia – the geographical region bordered by the Himalayan Plateau in the southeast, by the Bering strait in the northeast, and extending into island southeast Asia - remained controversial. Different researchers insisted upon three main scenarios. One postulated that northern populations of East Asia migrated to the south, mixing then with the Australian ancestors who had settled in Southeast Asia. Contrarily, another model suggested that the northern populations of East Asia evolved from the southern settlers. However, an explanation also exists holding that northern and southern East Asia populations evolved independently since the late Pleistocene, more than 10,000 years ago. [31-33]

To obtain insights into this question, Y chromosome diversity has been widely explored. There are four dominant Y chromosome haplogroups in East Asia – O-M175, C-M130, D-M174 and N-M231 – accounting for about 93% of the East Asia Y chromosomes. Other haplogroups, such as E-SRY4064, G-M201, H-M69, I-M170, J-P209, L-M20, Q-M242, R-M207 and T-M70, account for around 7% of the males in East Asia. [32]

Haplogroup O-M175 is the largest haplogroup in East Asia, comprehending around 75% of the Chinese population. O-M175 gave rise to three downstream haplogroups, O1a-M119, O2-M268 and O3-M122, totaling around 60% of the males in East Asia. [33, 34]

While the current Y chromosome diversity suggests multiple early migration

of modern humans from Africa via Southeast Asia to East Asia, and that after the initial settlements, northward migrations during the Paleolithic Age shaped the genetic structure in East Asia, many uncertainties still persist on the population history of the region.

China

First Settlements

Uncertainty also surrounds the when and how the first settlements of humans reached the geographical region now known as China.

The Y chromosome capability to infer the age of genealogies has been a support for the OOA hypothesis, as before mentioned. In combination with archaeological evidence, it is believed that the first modern man arrived in China, via Southeast Asia, around 60 to 45 thousand YBP. Evidence also indicates that very likely a second wave of migration from the north occurred 10 to 20 thousand years later. [35, 36] A major north-south genetic differentiation in the Chinese populations is being revealed, though the northern populations retain signs of the genetic origins in the south of the continent.

However, a few recent studies pointed out that East Asian populations' origin might be more complex than assumed in the previous works. According to those studies, the East Asian migrations can be explained by a multidirectional pattern, with both south and north migrations and with a great migration from central Asia, mainly into northern China, therefore suggesting a frame with multiple ancient migrations, into and within China. [35, 37-39]

Important information about earlier migrations is being obtained from the analysis of Y chromosome haplogroups. As mentioned before. In East Asia the dominant haplogroup is the O, defined by the O-M175 mutation. It emerged from an earlier NO clade, in which the haplogroups O and N shared a common mutation. It is believed that soon after the arrival in East Asia, a further mutation occurred in carriers of the NO lineages, separating them into N and O (~30.000 YBP). Later on, carriers of the N haplogroup individuals would mainly travel northwards, into Siberia,

Russia and Eastern Europe (~12.000 YBP), while individuals with the O chromosomes would establish mainly in East Asia.

Ethnicity concept in China – *Minzu*

Definitions of ethnicity and ethnic identification in China have a long and complex history.

The People's Republic of China was established in 1949, after the victory of the Chinese Communist Party in the post World War II civil war. One of the early steps undertaken by the new government was to recognize the existence of non-Han ethnic minorities. Then in PRC, "ethnic group" was connoted with the term minzu, an ambiguous word, whose meaning changed over time accompanying the symbiotic development of the concepts of "nation" and "ethnicity" in China.

Table 1 – Total population of each Ethnic group considered in this study (2010 Census results). The Han are the most numbered ethnic group in China, as well in the world. The Hui, Manchu, Tibetan and Salar follow them, in terms of absolute population number.

***Minzu* Name**	**Population (2010)**	**National share (%)**
Han	1.220.844.520	91.65
Hui	10.586.087	0.79
Manchu	10.387.958	0.78
Tibetan	6.282.187	0.47
Salar	130.607	0.0098

Currently, there are 56 officially recognized ethnic groups, or *minzu*, within the China population of ~1.300 million, with the majority Han population forming approximately 91% of the total (Table 1). The other 55 minority populations mostly live in the peripheral and boundary regions of the country and they range widely in size from some thousand to millions. [40]

The concept of ethnicity in China is not necessarily correlated with genetic ancestry, and it is probably more appropriate to consider *minzu* as a political construct, rather than a real ethnological identity.

This happens because after the formalization of the *minzu* concept, many Han communities, who lived in close association with ethnic minorities, voluntarily

adopted minority ethnic status, in order to benefit from the political and personal advantages that the ethnic minorities have. Furthermore, most of the 55 *minzu* minorities have not been subject to the One Child Certificate programme, and as a result, their number have increased disproportionately since the inception of this programme in 1979. [41, 42]

It is tempting, when conducting population genetic studies within China, to rely on the official *minzu* categories to define each population. This system, however, often failed to recognize pre-existing anthropological and demographic realities, which in turn meant that many *minzu* exhibit significant internal genetic heterogeneity. This blend of political, demographic and historical aspects implies that the concept of ethnicity in China is a complex and intricate subject, being necessary to account for all the factors that might underlie population structure, including history, when performing and interpreting the meaning of genetic analyses.

Han People

The term Han is used collectively to define the majority of the Chinese population. It was brought to common usage after the fall of the Eastern Han Dynasty, often referred to as the Chinese history golden age (202 BC – AD 220).

According to the 2000 Census, the Han numbered approximately 1.200 million people, and, although resident throughout most of the Chinese territory, they are more numerous in the more densely populated east region of the country.

Language wise, most Han people speak the Mandarin Chinese, the official language of PRC, whose written form is uniform throughout all the population. However, the spoken Mandarin language differs from province to province, with eight recognized major dialects.

As stated before, the ethnic concept in China is fluid and complex. As so, there is historical evidence of ethnic self-identification between Han population, mainly due to changing political fortunes. During the middle periods of the Qing dynasty (1700-1800 AD), some Han migrants changed their family names to Man surnames, and thus changed their perceived ethnic status. Mutually, the opposite

also occurred during the final stages of the Qing dynasty, with the intent of taking profit from economic and political advantages. [42, 43]

Given the large number of individuals and degree of geographical dispersal, it the effect of historical events on Han genetic diversity should be considered at least on a province-by-province basis.

Chinese Muslims

The arrival of Islam into China had a lasting effect. During the Tang Dynasty (618-907 AD), people of Islamic faith, mostly men, are believed to have entered China as soldiers, merchant and political emissaries from Central Asia, Arabia and Persia. Some of these men eventually settled in China and intermarried with local Han women, founding this way the present-day Chinese Muslim communities. [44]

There are ten officially recognized Muslim minorities in the Peoples Republic of China – the Bonan, Dongxiang, Hui, Kazakh, Kirghiz, Salar, Tatar, Tajik, Uygur and Uzbek – with a combined population of 91 million people.

Salar People

The Salar people are one of the official Muslim *minzu,* accounting for around 130.000 individuals, according to the last Census. The population mainly inhabit in the Autonomous County of Jisishan, in the Gansu province.

Their language is of Turkic origin, similar to Uzbek and Turkman. It is believed that the Salar originated from an Oghuz tribe, which travelled east from Samarkand (currently Uzbekistan), in the 14th or 15th century. Salar people ancestors are thought to have merged with Tibetans, Han people and Mongolians to form the present-day community. [45]

Hui People

The Hui people are other of the official Muslim *minzu*, present in China. Their total population, as for 2010, was of more than 10 million individuals. Despite practicing the Islam, the Hui share many ethnic affinities with the Han people. The

majority of the Hui people is also Mandarin speaker, although retaining some of the Persian words. They are mainly concentrated in the northwestern provinces and in the Central Plain, the area on the lower reaches of the Yellow River, in an Autonomous Region called Ningxia. [44, 45]

The Hui Chinese have diverse origins, and many of them are direct descendants of Silk Road travellers. The Silk Road was a series of trading routes, from Xian through Northwest China and Central Asia, to Constantinople (Istanbul), and as far as Italy. Hence, it is not surprising the ethnic diversity present in the regions of the old Silk Road, and the cultural influence that this trade route brought to China, as a result of the on going migrations that occurred. [44]

Tibetan People

The Tibetans account for more than 6 million people in the present Chinese population. Their ethnic group that is native of Tibet, but significant minorities also live in India, Nepal and Bhutan. Tibetans speak tibetic languages, many mutually unintelligible, which belong to the Tibeto-Burman languages family. Most Tibetans practice the Tibetan Buddhism, another religion present in the Chinese territory. [44]

The traditional, or mythological, explanation of the Tibetan peoples' origin is that they are descendants of the monkey Pha Trelgen Changchup Sempa.

The real history, however, of the human occupation of the Tibetan plateau remains largely unknown. Archaeological evidence points toward to around 23-50 000 years ago, as the period when the peopling of the Tibetan plateau first occurred. [46]

Recently, strong genetic evidence appeared of two major migrations of modern humans into the plateau, with the permanent occupation likely occurring in the early Upper Paleolithic before the Last glacial maximum, and a recent migration and population expansion beginning in the early Neolithic, coincident with the emergence of farming and yak pastoralism of the plateau. [44, 47, 48]

Manchu People

Manchus are the third largest ethnic minority group in China, with over 10 million people. They are distributed throughout China, for over 31 Chinese provincial regions, making the Manchu people the largest minority group in China without an autonomous region. At least half of the population, however, lives in the Liaoning province. Most Manchu today speak Standard Chinese, while the Manchu language is only spoken by elderly people in remote northeastern China.[44]

Liaoning province, the main Manchu territory, was for many centuries the border between the Chinese Empire and the 'barbarian' hordes. The tribes in this area absorbed, that way, a lot from the Chinese culture, through contact with the Chinese Empire, resulting in the formation of the Man (Manchu) nationality. In the time that the Chinese definitely established in Liaoning, mass migration and settlement of Han communities was recorded. Within a generation, Han immigrants outnumbered the local Manchu inhabitants. Over the times, many Manchu people changed their names to Han surnames, in order to obtain more economic and political advantages. For this reason, historically the Manchu and Han ethnicities are considered to have had a very fluid and strong connection, over the time. [43, 44]

Shanghai and Macau in the context of China

In the present, China, officially known as People's Republic of China, is a sovereign state located in East Asia and also the most populated country in the world, with a population of 1.35 billion people. China is a Socialist Republic, ruled by the Chinese Communist Party, and has jurisdiction over twenty two provinces, five autonomous regions – Xinjiang, Interior Mongolia, Tibet, Ningxia and Guangxi – four municipalities – Peking, Tianjin, Shanghai and Chongqing – and two Special Administrative Regions, with a high degree of autonomy – Macau and Hong Kong.

Shanghai

Shanghai city is located in the Yangtze River Delta, in the Chinese east coast, being bordered by the provinces of Jiangsu and Zhejiang to the north, south and west, and bounded to the east by the East China Sea.

During the Qing Dynasty (17th century), Shanghai became one of the most important seaports in the Yangtze Delta region. International attention to Shanghai, however, just grew importance in the 19th century, due to European recognition of its economic and trade potential at the Yangtze River. During the First Opium War (1839-1842), British forces occupied the city. The war ended with the 1842 Treaty of Nanjing, which allowed the British to dictate opening the treaty ports, Shanghai included, for international trade. [49]

Between 1860 and 1862, rebels attacked Shanghai twice and destroyed the city's eastern and southern suburbs, but failed to take the city. In 1862, the British settlement, to the south, and the American settlement, to the north, joined in order to form the Shanghai International Settlement. The French opted out of the Shanghai Municipal Council and maintained it's own concession, to the southwest of Shanghai. [50]

Citizens from many countries and from all continents migrated to Shanghai to live and work during the following decades, calling themselves *Shanghailanders*. In the 1920s and 1930s, almost 20 thousand "White Russians" and Russian Jews fled to Shanghai from the newly established Soviet Union and took shelter in Shanghai. In the 1930s, around 30 thousand Jewish refugees from Europe also arrived in the city. By 1932, Shanghai had become the world's fifth largest city and home to 70 thousand foreigners. Shanghai is, thereupon, the most 'occidentalized' city in China, having a long and lasting history of different cultures and people, from all over the world. [50]

Nowadays, Shanghai is, according to the last census, the house for more than 23 million people, earning this way the title for biggest Chinese city. Also according to census data, 98.8% of Shanghai's residents are of the Han Chinese ethnicity, while 1.2% belongs to various minority groups.

Macau

Macau is one of the two special administrative regions of the PRC. Geographically, Macau lies on the western side of the Pearl River Delta, across Hong Kong to the east, bordered by Guangdong province to the north and facing the South China Sea to the east and south.

Macau was a former Portuguese colony, having been under the administration of Portugal from the mid-16th century until 1999. Portuguese traders first settled in Macau in the 1550s, and in 1557 Macau was rented to Portugal, by the Chinese empire, as a trading port. Portugal administered the city under Chinese authority until 1887, when Macau officially became a colony of the Portuguese empire. During the 17th century, around 5 thousand slaves lived in Macau, in addition to 2 thousand Portuguese and 20 thousand Chinese. [51]

Nowadays, live in Macau over 509 thousand Chinese people, among whom the Portuguese account for only around 1% of the population.

Aims

This study involved the characterization of the Y chromosome genetic pool in samples of two Chinese populations, from Shanghai and Macau, aiming at achieving the following main goals:

- Compare the results obtained from Shanghai and Macau in order to evaluate whether signs were retained of the different factors (geographical, historical and/or cultural) that have shaped Y chromosome diversity in both populations.

- Explore the genetic background of the studied populations in the broader context of population diversity in China.

- Enrich the historical knowledge of the region, combining genetics with ethnic patterns, linguistic affiliation and historical records.

- Assess the global pattern of population genetic substructure in China, and investigate if it had any correlation with:
    1. Ethnic and linguistic affiliations
    2. Geography
    3. Pre-historical, historical and other recent migrations.

MATERIAL AND METHODS

Populations Samples and DNA Extraction

In this study, genetic material of 135 unrelated males, 85 from Shanghai and 50 from Macau, were sampled. All donors participated in the study according to the principles outlined in the “Helsinki Declaration”.

In the sample from Macau total DNA was extracted using the commercial Generation® Capture Card kit (Gentra Systems Inc, Minneapolis, USA), following the original manufacturer’s instructions with the adaptations: the incubation time, was extended to an overnight period in the case of DNA Purification Solution and to one hour with the DNA Elution Solution.

The Shanghai samples used in this work are a subset of those previously studied by Huang *et al.* (2013), who performed total DNA extraction using the standard phenol-chloroform methodology with proteinase K [52] as referred in material and methods section [Huang 2013].

Y-STR Genotyping

AmpFℓSTR® Yfiler® PCR Amplification Kit (Applied Biosystems) was used in order to analyze 17 STR markers - DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS393, Y GATA H4, DYS437, DYS438, DYS448 - located in the male-specific region of the Y Chromosome (MSY). This kit allowed the simultaneous analysis of the 17 Y-STR loci in a single reaction.

PCR amplifications were performed following the manufacturer's specifications, adjusted to a final volume of 5 µl, instead of 10 µl, with 1.84 µl of Yfiler Kit PCR Reaction Mix, 1 µl of Y filer Kit Primer Set and 0.16 µl of AmpliTaq Gold® DNA Polymerase. The volume of added DNA of each sample was 0.75 µl and 1.25 µl of water was added to the mix, in order to obtain the 5 µl final volume.

The PCR cycling conditions samples were one cycle at 95ºC for 11 minutes, followed by 30 cycles of 94ºC for 1 minute, 61ºC for 1 minute and 72ºC for 1 minute. The final extension step was performed at 60ºC for 80 minutes.

PCR products were run on the ABI 3130 Genetic Analyser and the results analysis was performed using GeneMapper® ID Software v4.0. Genotyping of samples were conducted by comparisons with sequenced ladders included in the kit. The *locus* nomenclature and allele designation were performed according to the ISFG guidelines [53, 54].

Y-SNP Genotyping

Based on the knowledge that the majority of East and Southeast Asia Y chromosomes belong to haplogroup O, which is characterized by a 5-bp deletion known as M175, a multiplex system was implemented, hereinafter referred to as multiplex O, originally designed by *van Oven, et al.,* 2012, who have taken advantage of the recent advances and discoveries in Y-SNP typing to improve substantially the phylogenetic resolution of haplogroup O sublineages. The multiplex contains the following 16 different SNPs: M175, M134, M119, P203, M110, M268, M95, M88, M176, M122, M324, KL1, 002611, P201, M7, PS23. They allow the discrimination of 16 distinct O sublineages (Figure 8).

Another multiplex was used, referred to as Multiplex 1, which accounted for 9 SNPs: P25, 92R7, SRY1532, M70, M173, Tat, M213, M9, M13.

The YAP system was also analysed, although isolate in a single PCR reaction. Yap plus Multiplex 1 turned able to discriminate 10 different haplogroups (Figure 8).

In total the combination of the results obtained from all bi-allelic markers allowed us to define 28 different haplogroups (Figure 8), named according to the nomenclature proposed by Karafet *et al.* (2008).

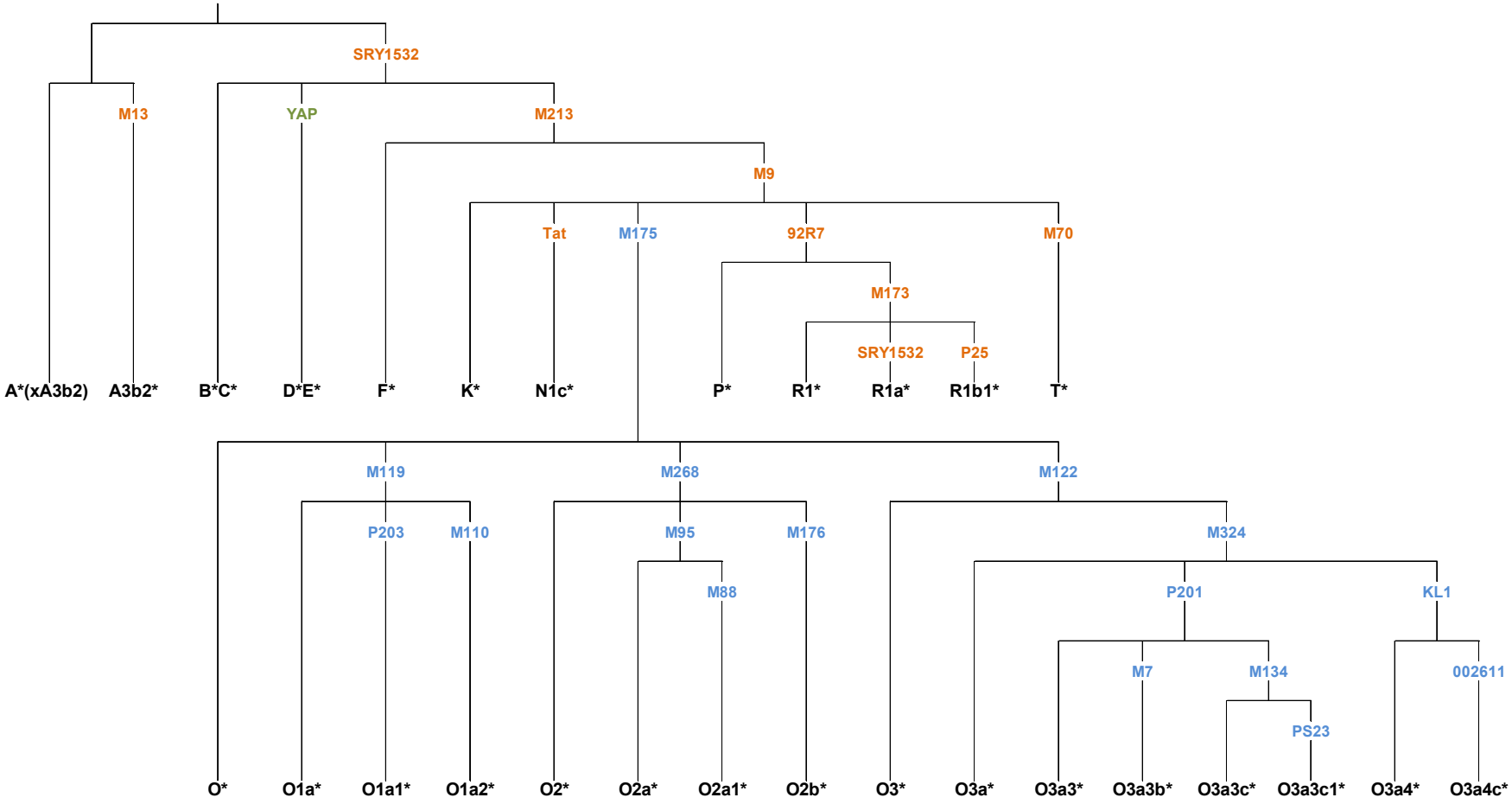


Figure 8 – Phylogenetic tree of the Y chromosome haplogroups studied. Biallelic markers are displayed in each branch, where the colors correspond to the different multiplexes used (blue – multiplex O and orange – multiplex 1), the YAP polymorphism, genotyped in a singleplex reaction is shown in green. Haplogroups are named according to Karafet et *al.* (2008)

Genotyping Strategy

The typing strategy involved three sequential steps:

1. The multiplex O was tested in all samples, given the previous knowledge that O lineages present very high incidence in East Asia. Since besides M175, the multiples also includes other 15 SNPs, it allowed to identify sublineages to which belonged O individuals.

2. Samples excluded to be O were all analysed with multiplex 1, which includes, 9 different previously mentioned SNPs.

3. Samples that did not fell into any haplogroup able to be detected with both multiplexes were lastly tested for the presence of the YAP element.

Figure 9 – Schematic representation of the three step strategy followed in this study. In red are the step relying in multiplex PCR and SBE techniques while in orange that based in single PCR and PAGE.

The typing strategy was determined by the known structure of the Y chromosome phylogenetic tree (Figure 8). After tested with multiplexes O and 1, a likely possibility for samples showing a derived status for SRY1532 polymorphism and ancestral status for the M213 polymorphism, was to belong to the haplogroups B*C* or DE*. This was deduced because the absence of the mutation defining the F haplogroup – M213 polymorphism– included in multiplex 1, indicated that those samples could belong to an older and more basal haplogroup in the tree, since in addition they didn't accumulate any of the more recent mutations integrated in the YCC tree (excluding the possibility of a back mutation event).

Therefore, we test the YAP polymorphism, given that it could differentiate between two likely options, haplogroup B*C* and haplogroup DE*. Since haplogroup B is found mainly in sub-Saharan Africa populations and haplogroup E is also only well represented across the African continent their presence in East Asia seemed unlikely.

Genotyping Methods

SNaPshot

Concerning the referred two multiplexes, a mini-sequencing reaction was performed, using the SNaPshot™ kit (Applied Biosystems). This technique involves an amplification of the surrounding region of the polymorphic site, primer annealing, adjacent to the SNP position, and primer extension of a single ddNTP (fluorescent dye-labelled dideoxynucleotide triphosphate). Since the ddNTPs are labelled with specific fluorescent dyes, it was possible to combine different polymorphisms in the same reaction.

The PCR for the O haplogroup multiplex and for the multiplex 1 plus M13 were performed in a final volume of 10 µl, with 5 µl of 2x QIAGEN® Multiplex PCR Kit and 1 µl of primer mix at 2.0 µM. The volume of added DNA for each sample was of 0.5 µl, and 3.5 µl of water was added in order to complete the 5 µl volume.

As for the PCR conditions, samples were submitted to a denaturation step of 15 minutes with 95ºC, followed by 30 cycles of 94ºC for 30 seconds, 60ºC for 1 minute and 30 seconds and 72º for 1 minute. Each one of the different temperatures in the cycles phase corresponds to the denaturation, annealing and elongation step, respectively. A final extension step was performed at a temperature of 72ºC for 15 minutes.

Purification of the PCR products was done in order to degrade the remaining primers and nucleotides of the PCR reaction, was done using ExoSAP-IT® (USB Corp.), following the manufacturer's specification.

The SNaPshot reaction was performed in a final volume of 5 µl, with 1 µl of SNaPshot™ Multiplex Mix (Applied Biosystems), 1.5 µl of SBE (Single Base Extension) Mix and 1.5 µl of PCR product, previously purified. Finally, 1 µl of water was added to the mix to obtain the 5 µl final volumeThe samples went through 25 cycles of 96ºC for 10 seconds, 50ºC for 5 seconds and 60ºC for 30 seconds.

To remove unincorporated ddNTPs from the SNaPshot reaction, a product purification was performed using SAP™ (USB Corporation, Clevelend, OH, USA), according to the manufacturer's specifications.

The separation of the mini-sequencing reaction products was done by capillary electrophoresis using the ABI 3130 Genetic Analyser (Applied Biosystems).

The data obtained from the capillary electrophoresis was analysed using the GeneMapper® Software v4.0 (Applied Biosystems). Results were visualized in a single eletropherogram, corresponding each peak to a different polymorphism, with the color defining the status (Figure 10).

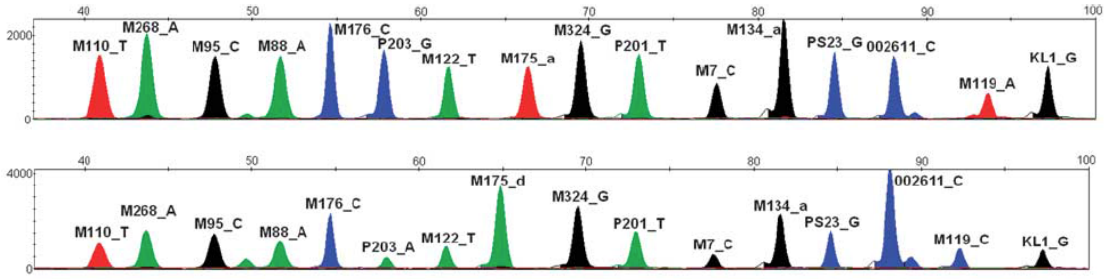


Figure 10 – Example of an electropherogram obtained from the O specific multiplex. Each nucleotide is associated with a particular dye color, A as green, C as yellow but displayed as black, G as blue and T as red. **On the top** is represented the profile of an individual negative for all the polymorphisms tested with this multiplex. **On the bottom** is represented an individual with the O-P203 mutation. [55]

Direct Visualization

For the detection of the YAP element, a simple PCR amplification was performed and then the amplified product was submitted to the Polyacrylamide Gel Electrophoresis (PAGE).

The PCR for YAP was performed in a final volume of 5 µl, with 2.5 µl of 2x QIAGEN® Multiplex PCR Kit and 0.4 µl of each forward and reverse primers at 2.5 µM. The volume of added DNA for each sample was of 1.2 µl, and 0,5 µl of water was added in order to complete the 5 µl volume.

Samples were submitted to an initial denaturation at the temperature of 95ºC for 15 minutes, followed by 35 cycles at 94ºC for 30 seconds, 52ºC for 90 seconds and 72ºC for 60 seconds. The final extension step was performed at a temperature of 72ºC for 10 minutes.

The difference in the size of the two amplified allele products (YAP+ with 455bp and YAP- with 150bp) was clearly visualized after applying to the polyacrylamide gel with the silver staining method (Figure 11). [56] The silver staining is a sensitive method of nucleic acids detection, which permits the visualization of the electrophoresis results without any special equipment. In silver staining, polyacrylamide gels are impregnated with the soluble silver ion (Ag^{+}) and developed by treatment with a reductant. Macromolecules in the gel promote the reduction of silver ion to metallic silver (Ag^{0}), which is insoluble and visible, allowing bands containing nucleic acids to be seen.

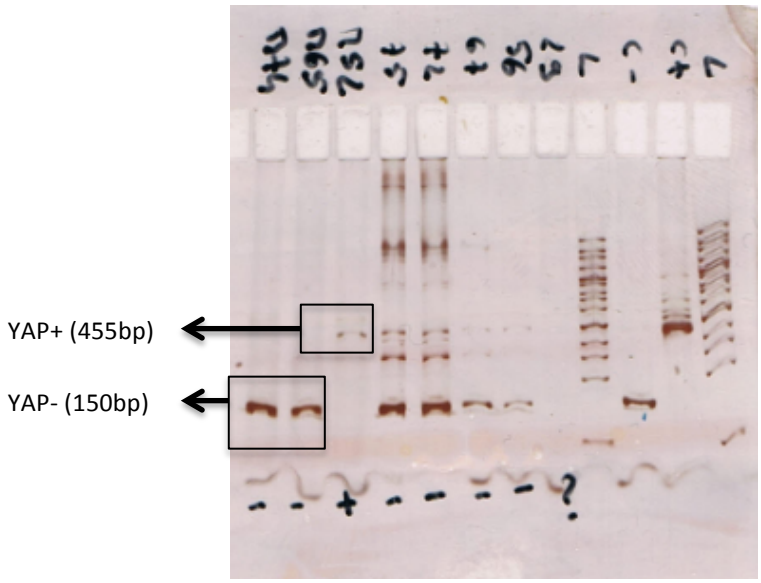


Figure 11 – Discrimination of the two YAP alleles in a polyacrylamide gel after application of the silver staining method. YAP positive (YAP+), Y chromosome with an inserted copy of the Alu element is represented in the image by the 455bp band and the 150bp band represents the negative YAP (YAP-).

Populations Data

In order to conduct comparative analyses, previously published population data were recruited from the literature for this study. Concerning STRs only, populations from China were used (Table 2 and Figure 12).

Table 2 – Populations used for comparative analysis with STRs and SNPs.

Fig. Code	Population	N	Ethnia	Language	Reference
1	Shanghai	85	Han	Shanghainese	This work
2	Macau	50	Han	Cantonese	This work
3	Henan	276	Han	Henan dialect (Dialect of Mandarin)	Wu et al. (2011)
4	Shandong	131	Han	Jiaoliao and Zhongyuan (Dialects of Mandarin)	Yan et al. (2007)
5	Xinbin	231	Manchu	Jiaoliao Mandarin and Manchu	Juan He (2013)
6	Ningxia Hui	141	Hui	Zhongyuan and Lanyn (Dialects of Mandarin)	Hua Guo et al. (2008)
7	Shanxi	222	Han	Jin and Zhongyuan (Dialects of Mandarin)	Mei-Sen et al. (2011)
8	Qinghai Tibetan	167	Tibetan	Khams language dialect	Bofeng Zhu et al. (2008)
9	Taiwan	200	Han	Mandarin	Tsun-Ying et al. (2008)
10	Lhasa Tibetan	351	Tibetan	Khams language Lhasa dialect	Qingxia Zhang et al. (2006)
11	Qinghai Salar	133	Salar	Turkic branch of the Altaic language family	Bofeng Zhu et al. (2007)
12	Luzhou, Sichuan	424	Han	Sichuanese	Long Bing et al. (2013)
a	Korean	506	Korean	Korean	Kim et al. (2010)
b	Khalkh	45	Mongol	Khalkha Dialect	Jin et al. (2003)
c	Buryat	36	Mongol	Buryat Dialect	Kim et al. (2000)
d	Beijing	51	Han	Mandarin	Kim et al. (2000)
e	Xian	34	Han	Jin and Zhongyuan (Dialects of Mandarin)	Kim et al. (2000)
f	Yunnan	60	Han	Linguistic diversity	Jin et al. (2003)

Figure 12 – Map of China showing the location of Shanghai, Macau and other sample populations used in the comparative analysis with STR data. Correspondence between each number and population is described in Table 2.

For SNPs, a sample from Korea and two from Mongolia – Buryat and Khalkh – were used besides three samples from different Chinese populations – Han from Beijing, Han from Yunnan and Xian. (Table 2 and Figure 12)

Four samples from Shanghai and two from Macau did not perform well in the Y-STR typing and were not used in this analysis, since it was not possible to obtain a complete Y-STR profile. Four other different samples from Shanghai and two from Macau were not possible to type, haplogroup wise, and were not used in the Y-SNP tests.

Statistical Analysis

The haplogroup and haplotype frequencies, as well as the number of shared and unique haplotypes of the studied populations were calculated using Arlequin v3.5.1.3 software (Excoffier and Lischer, 2010). The Mean Number of Pairwise differences (MNPD) for the STR data and gene diversity values (Nei, 1987) for bi-allelic and STR *loci* data were estimated, as well as the matrixes of pairwise genetic distances and the corresponding p-values, by Arlequin v3.5.1.3 software. Pairwise F_{ST} distances results were computed using the Slatkin's linearization (Slatkin, 1995) for STR data (R_{ST}) and Reynolds linearization (Reynolds *et al.* 1983) for bi-allelic data (F_{ST}). A statistical significance of P<0..05 was also initially considered, although a Bonferroni correction was also applied – P<0.00415 for STR data and 0.00625 for bi allelic data.

Arlequin software was also used to perform analysis of molecular variance (AMOVA), considering both types of information, haplotypes and haplogroups. Such analysis allowed the estimation of the percentage of genetic variation that was observed among groups, among populations within groups and within populations. The groups for this test were chosen following geographic and ethnic criteria. For these tests SAMOVA v1.0 (Dupanloup, I., Schneider, S., Excoffier, L. 2002) was also used. SAMOVA is software which, based on a number of previously chosen groups, will then divide the studied populations following the mentioned pattern. This separation is based on three parameters of population differentiation measurement – FCT, FST and FSC.

Finally, pairwise genetic distances were analysed through MultiDimensional Scale (MDS), using the SPSS v21.0 (SPSS inc.) program and the PROXCAL algorithm, where the fitness of the results was measured by the s-stress value. The pairwise genetic distances were also measured using the Barrier v2.2 (Manni *et al*. 2004), a software that combines genetic and geographic distances. This software computes geographic barriers by taking into account the correlation between the two types of distances, by using the Monmonier algorithm (Monmonier, 1973), assigning barriers when a high genetic differentiation between neighbouring populations exists.

RESULTS
AND DISCUSSION

Haplogroup Frequencies, Diversity Values and Phylogeography

Out of the samples tested for Y-SNPs, in four from Shanghai and one from Macau no successful amplifications were obtained.

In the remaining samples, 15 different haplogroups were identified, based on the analysis of the panel of 25 Y-SNPs.

The samples were at first tested for the M175 and other mutations and other mutation within haplogroup O with a specifically designed multiplex, which permitted to classify the O-positive chromosomes into 11 different O subtypes.

Further testing, using initially a broader multiplex and next, when though needed, YAP allowed the classification of the samples previously classified as negative for the M175 mutation in four new different haplogroups.

In nine samples from Shanghai and one from Macau it was only possible to exclude that they fell into any of the haplogroups identifiable with the 25 Y-SNPs examined. Due to time limitations, no additional characterization was performed.

In Table 3 are presented the distributions of Y-haplogroups detected in Shanghai and Macau.

Table 3 – Haplogroup frequencies in Shanghai and Macau. N represents the number of individuals and the frequency is also represented. Total N is 81 and 49 for Shanghai and Macau, respectively. Last SNP genotyped in the branches is showed. In bold the main clades are represented.

Haplogroup	N_S	Frequency Shanghai	N_M	Frequency Macau
O1a* (M119)	1	0.01235	1	0.02083
O1a1* (P203)	15	0.18519	9	0.18750
O2* (M268)	5	0.06173	1	0.02083
O2a* (M95)	1	0.01235	4	0.08333
O2a1* (M88)	3	0.03704	4	0.08333
O3* (M122)	1	0.01235	1	0.02083

Table 3 (Continued) – Haplogroup frequencies in Shanghai and Macau. N represents the number of individuals and the frequency is also represented. Total N is 81 and 49 for Shanghai and Macau, respectively. Last SNP genotyped in the branches is showed. In bold the main clades are represented.

O3a3* (P201)	10	0.12345	6	0.12500
O3a3b* (M7)	1	0.01235	2	0.04167
O3a3c* (M134)	8	0.09875	5	0.10417
O3a3c1* (PS23)	11	0.13580	8	0.16667
O3a4c* (002611)	10	0.12345	2	0.04167
BC*	4	0.04938	2	0.04167
DE* (YAP)	0	0.00000	1	0.02083
N1c* (Tat)	1	0.01235	1	0.02083
R1* (M173)	1	0.01235	0	0.00000
Y*	9	0.11111	1	0.02084

The large majority of the Y lineages in this study belonged, as expected, to the O haplogroup (M175): 89.58% in Macau and 81.5% in Shanghai.

As a matter of fact, this is the most widespread haplogroup not only in China but in East Asia as a whole, where it has been reported with frequencies as high as 75% [33], 76.5% [32], or even to 80.9% [55], and is therefore associated with the major Neolithic migrants. Still, the frequencies of O chromosomes here found in Macau and Shanghai are higher than up to now described in Chinese populations.

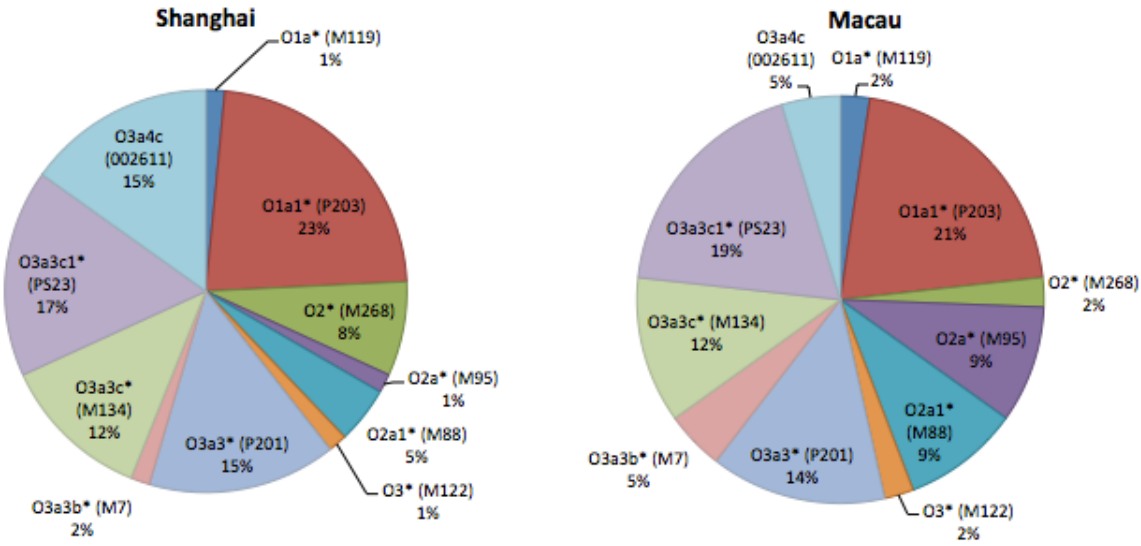


Figure 13 – Frequency of the O-M175 lineages in Shanghai and Macau.

O-M175 contains three major downstream haplogroups – O1a-M119, O2-M268 and O3-M122. The O3*-M122 clade was the most common, being carried by 50.61% and 50.02% of the individuals for Shanghai and Macau, respectively. This values are also in accordance with previously studies showing that 50 to 60% of males from Han Chinese populations can belong to the O3*-M122 clade [33, 34] Haplogroup O3 is prevalent throughout East and Southeast Asia, and also highly prevalent in almost all the populations of the Sino-Tibetan linguistic family. Branches defined by 002611, M134 and PS23 (M117), the three main subclades of O3, have been reported to account for 16.9%, 11.4% and 16.3%, respectively, of the Han Chinese populations. [33] The relative proportions of these three subclades are rather dissimilar in our samples. Shanghai presented 9.88%, 13.6% and 12%, respectively, for each previously type. In Macau the pattern was not far from that: O-M134 and O-PS23 occurred at frequency of, respectively, 10.42% and 16.67%; only for O-002611 a considerable decreased frequency was observed since it accounted for only 4.17%. O3a2b-M7 was another of the detected O lineages at frequency of 1.2% in Shanghai and 4.0% in Macau. Both values are in the range typically found in Han Chinese, among whom O3a2b-M7 reaches less than 5% of the Han Chinese. [33, 34, 57, 58] The dissection of O3 Y-lineages is proving crucial genetic evidence to reconstruct a picture of the dispersal and expansion patterns of Han Chinese in the late Neolithic age. Carriers of haplogroup O3a1c-002611 and O3a1c1-F11 started their northward migration about 12 KYA, from Southeast Asia, along with carriers of other O3-M122 lineages, reaching the upper and middle Yellow river

basin. About 6 KYA Han Chinese split and started their migration to the east and south. [59]

The high frequency of O3 lineages present in our samples is in accordance with the pattern commonly found in Han Chinese spread throughout over China.

Two other O clades were here detected: O1a*-M119, represented at 19.75% in Shanghai and 20.8% in Macau, being so the second most common clade in our samples; and O2*-M268 present at 11.11% and 18.75%, in Shanghai and Macau, respectively. From these two clades, O1a-M119 has been reported to be more prevalent clade along the southeast coast of China. Previous data from Han Chinese from all over China revealed that of O1a* (M119) accounts for 3.2%-24% of males, whereas O2*-M268 accounts for 7.2%-13.2% Y chromosomes. [57] So, whilst the results here obtained for O1a* are in the range up to now found in Han Chinese, but for O2* in Macau the frequency is considerably higher than usual.

Samples that were not assigned to haplogroup O after analysis with Multiplex O were further tested for additional SNPs, which lead to identify four other haplogroups. In Shanghai the non-O chromosomes were C, N and R1, which together accounted for 7.4%; In Macau C, D and N lineages were found at a combined frequency of 10.21%. In both samples, the most common non-O haplogroup was C with 4.9% and 4.0% frequency, respectively in Shanghai and Macau.

Apart from O, the next three dominant haplogroups in East Asia are C, D and N, whose presence is common across populations from the region. [32]

In Table 3 the samples are labelled as BC* and DE*, were described in that manner because the most ancestral mutation they harboured was common to the C and B haplogroups or to the D and E haplogroups, respectively. The known patterns of worldwide geographic distribution of these four haplogroups indicate that only C and D are haplogroups usually found in East Asia, although at values varying widely from population to population. [32], In fact, along with N, C and D represent the most common non-O lineages in the region.

On the contrary, the macrohaplogroup B is only typically found in Africa. So then, the six lineages here classified as BC* are very likely C chromosomes, although the presumption still needs to be demonstrated.

Haplogroup C, defined by M130, is an ancient lineage, widely dispersed in Asia and Oceania. Its distribution pattern suggest that C-M130 might have arisen somewhere in mainland Asia before modern humans arrived in Southeast Asia. It is believed that populations with haplogroup C settled in East Asia around 10 KYA earlier than the individuals with the haplogroup O. Nowadays, C-M130 accounts for around 10% of Y chromosomes in Eat Asian populations [13, 58, 60], a value which is higher than here found in Shanghai and Macau, assuming that the BC* types are indeed B.

As for the unique sample classified as DE*, given that haplogroup E has only been residually detected in East Asia, most probably it belongs to haplogroup D.

The dispersion history of the D haplogroup, defined by M174, still remains highly mysterious. This haplogroup is derived from the African haplogroup DE-M1 (YAP insertion), being thus brother of haplogroup E. It is believed that, while haplogroup E spread westwards to Africa, haplogroup D was carried eastwards to East Asia. Shi *et al.* (2008) proposed that the northward expansion of D-M174 to western China might predate the migrations of other major East Asian lineages, where D-M174 in East Asia was probably displaced from Eastern China by the later northward migration of haplogroup O, and the Neolithic expansion of Han Chinese. [32, 58, 61, 62]

Haplogroup D-M174 has the highest frequencies in the Andaman Negritos, the northern Tibeto-Burman populations and the Ainu of Japan. In some populations it can reach values up to 60% as it occurs in Tibetan populations. However in China its common frequency varies between 0 and 5%. [58]

So, in Macau the frequency of D at 2% (if confirmed), is a usual value for a Chinese population. No doubt that is necessary to confirm the absence of the B and E haplogroups, especially in Macau, where a lasting history of slave trade during the Portuguese rule of the territory might have facilitated the introduction of male lineages unusually found in East Asia.

Admitting that B and E will indeed be ruled out, the combined frequency of C, D plus N accounts in Shanghai and Macau for values somehow lower than usual in Eurasian populations. But that is compensated by the high frequency O lineages in the two samples, resulting in a combined frequency of C, D, N and O chromosomes reaching the common range in populations from Eurasia.

N1c-Tat was also detected in one male from Shanghai and another from Macau. The haplogroup arose probably in China around 14 thousand years ago, experiencing afterwards a series of founder effects or strong bottlenecks in Siberia, accompanied by a secondary expansion in East Europe. Studies trace the origin of haplogroup N to Southwestern China or Southeast Asia, and the dispersion of haplogroup N has been argued as evidence for the origin of the East Asian populations. The N haplogroup emerged from an earlier NO clade, in which the haplogroups N and O shared a common mutation. It is believed that soon after the arrival in East Asia, a further mutation occurred in the NO lineages, separating them into N and O (~30.000 YBP). Later on, individuals with N chromosomes would mainly travel northwards, into Siberia, Russia and Eastern Europe (~12.000 YBP), while carriers of O haplogroup would establish in East Asia. Haplogroups O and N are, therefore, brother haplogroups.

The scarce representativeness of N1c both in Macau and Shanghai fits well the low frequency of the haplogroup found across East Asia, including China where the highest frequencies were reported to be 5-7%. [62-64]

One individual from Shanghai was found to belong to the R1 haplogroup, which was identified by the presence of the derived allele at M173. Two main clades within R1-M173, defined by M420 and M343-R1b, are widely spread throughout Eurasia, with the exception of the Southeast and East Asia. R1-M173 contains, however, other minor clades, still remaining incipiently characterized by known SNPs and being collectively referred to as R-M173*, which have shown to be present in the Americas, Asia and Oceania. In this study, because just a single chromosome felt in R1-M173 and given time limitation, no further SNP characterization of that lineage was performed, which would be necessary to clarify whether it could signal a possible European influence in the Han population from Shanghai. Anyway, since no other males lineages of likely European origin were detected in Shanghai, our data indicate that it was virtually null the genetic impact of the presence of European

people in what was known since the early nineteen century as “The Paris of the East”.

The absence of male European lineages in Macau is still more remarkable given that for more than four centuries Macau represented the centre of Portuguese trade and culture on the South China coast.

Haplogroup diversities here estimated for Shanghai and Macau are presented in Figure 14, where are also shown previously reported values for other East Asian populations.

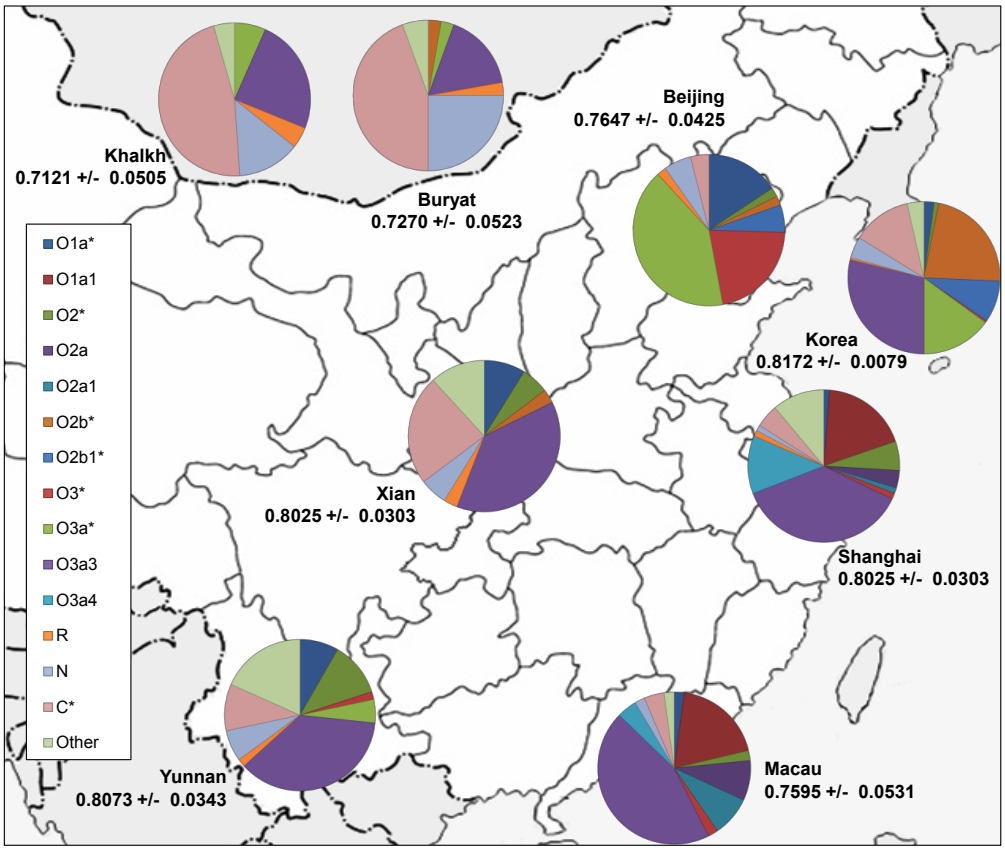


Figure 14 – Y-SNP gene diversity values. In the figure below are represented the values of Y-SNP diversity, obtained considering a set of 13 common haplogroups.

Levels of haplogroup diversity are high in all populations shown in Figure 14, with the Koreans presenting the most elevated value of Y haplogroup diversity, followed by the populations from Yunnan and Shanghai that show quite similar high values 0.8073 +/- 0.0343 and 0.8025 +/- 0.0303, respectively, when compared to the other populations considered in the comparison (Figure 14). In the opposite lowest

range are the two populations from Mongolia, the Buryat and the Khalkh, presenting diversities of 0.7270 +/- 0.0523 and 0.7121 +/- 0.0505, respectively.

High diversity of male lineages in the Koreans probably relies in their intricate and complex history. Although Koreans are in overall integrated in the genetic context of other Northeast Asian groups, it has been proposed a dual origin for the peopling of the Korean peninsula, with contributions from both southern and northern parts of East Asia, since the population from Korea contains lineages typical of both Southeast and Northeast Asian populations. [65]

Within China, the more diverse populations are those from Shanghai, Yunnan and Xian, located in Central or Southern China. Comparatively, the population from Beijing, in the Northeast region, reveals a considerable reduced diversity.

As mentioned before, Underhill *et al.* (2001) sustained a southeastern origin for the populations of East Asia. After the arrival of modern man in China via Southeast Asia, they began migrating toward north having made a substantial contribution to extant Northeastern Asia populations, which still retain signs not only of the southern origin but also of bottlenecks and genetic drift during the northwards migration, with consequent lose of diversity. Even though other migrations in Postglacial and Neolithic times also importantly contributed to current East Asian populations, compelling evidence suggest that the southern route shaped a south-to-north clinal structure that is being demonstrated in East Asia (Zhong et *al.* 2011). Recently, a trend to diversity decrease from South to North East Asia has been clearly shown by genome-wide SNP data (Abdulla 2009), which seemingly is also supported by our Y-SNP based results notwithstanding the drawback of the small number of populations here considered.

Other factors might have also contributed for the high diversities in Shanghai and Yunnan. For a long time ago, both cities are known to be important trade centres of goods. Shanghai was a well-established strategic point for trade, mainly because of the seaport. Yunnan, on the other hand, constituted the cornerstone of the Silk Road Route. The long-lasting history of migratory movements, into and out of both cities, as well as the constant renewed settlement of the cities, likely accounted for their broader spectrum of Y chromosomes.

Concerning diversity in Macau, it represents indeed the lowest value compared to the Chinese populations. For that it must likely has accounted the recent, while centenary, history of human movements into this small regions that still constitutes a special administrative region in China.

Haplotype Frequencies and Diversity Values

In this study, a total of 127 different haplotypes (79 in Shanghai and 48 in Macau) defined by 17 Y-STR markers were identified. Only one shared haplotype was observed between two individuals from Shanghai samples. Accordingly, haplotype diversity inferred from the entire set of markers was 0.9997 +/- 0.0020 in Shanghai, while in Macau reached the maximum values 1.0000 +/- 0.0043. Both values are in range of those commonly reported for Chinese populations tested with the same set of Y-STR *loci.* [66-69]

Table 4 – Y-STR haplotype diversity data of the 12 populations used in this study. Mean Number of Pairwsise Differences (MNPD) value on the right.

Population	N	STR haplotype diversity	MNPD
Shanghai	81	0.9991 +/- 0.0021	9.537346 +/- 4.419859
Macau	48	1.0000 +/- 0.0043	9.515071 +/- 4.442024
Henan	276	0.9998 +/- 0.0003	10.268590 +/- 4.701417
Shandong	131	0.9998 +/- 0.0010	9.158191 +/- 4.239352
Xinbin Manchu	231	0.9996 +/- 0.0005	9.689253 +/- 4.455636
Ningxia Hui	141	0.9933 +/- 0.0010	10.057244 +/- 4.624071
Shanxi	222	0.9999 +/- 0.0004	10.183115 +/- 4.668170
Qinghai Tibetan	167	0.9998 +/- 0.0007	9.115576 +/- 4.215253
Taiwan	200	0.9996 +/- 0.0006	9.360804 +/- 4.317102
Lhasa Tibetan	351	0.9999 +/- 0.0002	9.074562 +/- 4.186760
Qinghai Salar	133	0.9983 +/- 0.0013	10.096605 +/- 4.642639
Luzhou (Sichuan)	424	1.0000 +/- 0.0002	9.663522 +/- 4.437385

The levels of haplotype diversity were similar and very high in all Chinese populations (Table 4) ranging from 0.9933 +/- 0.0010 in the Ningxia Hui to 1.0000 +/- 0.0002 in the Luzhou and Macau. For Y-STR lineages, no signs emerged of diversity structuring according to geography in China.

Interestingly, however, the Hui and the Salar presented the lowest levels of Y-haplotype diversity. Both are Muslim populations, with a history of relative isolation from other populations, mainly due to the religious affiliation. The presence of Islamic groups in China is associated to people of Islamic faith, most of them men, who have

entered China from Central Asia, Arabia and Persia. It is believed that there are Muslim in China since around 1.400 years ago [45] and from then on the Islam expanded gradually across the entire China. Despite the multiple accounts of continuous interaction between Muslims and other people from China, the religious beliefs might have favoured some isolation, and consequently genetic drift, in certain groups, contributing to decrease levels of diversity. Whereas that can be indeed the case of the Hui and the Salar, it is noteworthy the elevated mean number of pairwise differences between haplotypes in both populations, which otherwise may reflect the assimilation of quite differentiated lineages as a consequence of interactions with other Chinese populations.

Also of note that the two lowest values of MNPD are shown by the Tibetan populations from Qinghai and Lhasa, which indicated that despite the high haplotype diversity (a common feature even in other Chinese ethnic minorities [44, 59]), on average haplotypes from Tibetans and molecularly more related between each other than in other populations from China.

One of the largest minority ethnic groups in China, the Manchu, presents a very high level of diversity, result that was expected from the recent extensive admixture between the Manchu with the Northern Asian and Han Chinese populations.

Population Comparisons

In order to investigate the affinities between the samples here studied from Shanghai and Macau and other East Asian populations, genetic distances were calculated, R_{ST} for STR data and F_{ST} for SNP data, using the data available for surrounding populations. However, since East Asia is still a scarcely studied area, data for the comparative analysis was quite limited for the markers tested, especially when it comes to SNP data.

Y-STR

For the comparative analysis using Y-STRs, a set of 15 different *loci* was used, in order to maintain the same level of resolution for all populations.

The matrix of R_{ST} genetic distances obtained through AMOVA is displayed in Table 5.

Table 5 – Matrix of the R_{ST} genetic distances for the 12 populations used in the STR study. All the statistical significant values are highlighted in bold. Those significant (P<0.05) are labelled with an asterisk and those values that remain significant after the Bonferroni correction (P<0.0042) with two asterisks.

	Shanghai	Macau	Shandong	Ningxia Hui	Shanxi	Qinghai Tibetan	Taiwan	Lhasa Tibetan	Q. Salar	Henan	Xinbin Manchu
Shanghai	*	-	-	-	-	-	-	-	-	-	-
Macau	**0.02065***	*	-	-	-	-	-	-	-	-	-
Shandong	**0,02410****	0.00789	*	-	-	-	-	-	-	-	-
Ningxia Hui	**0.06286****	**0.06732****	**0.02120****	*	-	-	-	-	-	-	-
Shanxi	**0.03964****	**0.01895****	0.00000	**0.02226****	*	-	-	-	-	-	-
Qinghai Tibetan	**0.26783****	**0.26442****	**0.21807****	**0.13227****	**0.19414****	*	-	-	-	-	-
Taiwan	0.00331	0.00000	0.00000	**0.07414****	**0.04030****	**0.28784****	*	-	-	-	-
Lhasa Tibetan	**0.15883****	**0.14741****	**0.09957****	**0.06621****	**0.11289****	**0.02337****	**0.17000****	*	-	-	-
Qinghai Salar	**0.12294****	**0.12601****	**0.09101****	**0.02084****	**0.06211****	**0.12658****	**0.14041****	**0.08575****	*	-	-
Henan	**0.01952****	0.00936	0.00000	**0.07359****	**0.03626****	**0.25675****	**0.01496****	**0.15728****	**0.13951****	*	-
Xinbin Manchu	**0.01969****	0.00000	0.00000	**0.02570****	**0.01470****	**0.19410****	**0.01261****	**0.11188****	**0.07211****	**0.02084****	*
Sichuan Luzhou	**0.08211****	**0.00936****	**0.06625****	**0.09094****	**0.06039****	**0.28188****	**0.06956****	**0.18021****	**0.16460****	**0.07477****	**0.05989****

Between the two studied populations in this work, Macau and Shanghai, the R_{ST} value was 0.02065, with no significant statistical differences after the Bonferroni correction, attesting, therefore, for a rather high homogeneity between the two populations in which respect Y-STRs.

Furthermore, among the compared samples of Han populations – which were Macau, Shanghai, Henan, Shandong, Taiwan, Shanxi, Luzhou, it was noticeable the low pairwise values of genetic distances, even though some reached statistical significance, when compared to distances involving non-Han populations, confirming previous evidence that the Han constitute a coherent genetic cluster of populations.

As for the two Tibetan ethnic groups, from Lhasa and Qinghai, not surprisingly they presented very lower level of genetic differentiation, and the same happened with the two Muslim populations, the Hui from Ninxgia and the Salar ethnic populations. Historically, minorities living in China are closed populations, were the people tend to get together with other culturally and ethnic similar people, even some Muslim communities that live in close proximity have little contact or intermarriage [70], probably from the arise of the ethnic nationalism between the different minorities in the country. [44, 71]. Therefore, even though groups with identical ethnical or cultural affiliation might live geographically quite far in different provinces, such as the Tibetans Lhasa and Qinghai or the Muslims Salar and Hui, they retain firm genetic signatures of the shared history and/or ancestry.

The Manchu population, from Xinbin, even though commonly categorized as a different ethnic other than the Han, genetically reveals clear affinities with the assumed Han populations. The values of genetic distances reflect the close relationship between Manchu *minzu* and Han throughout the Chinese history.

Overall, with genetic distances ranging from 0.0000 to 0.28784, with most of them assuming statistical significance, the results provided by Y-STRs data testify that current China encompasses remarkable genetic diversity,

MDS Representation

In order to obtain a visual representation of the genetic affinities between populations, the matrix of pairwise R_{ST} values was submitted to MDS and the obtained plot is presented in the Figure 15, where a color code is used to represent the different ethnic affiliations of the populations.

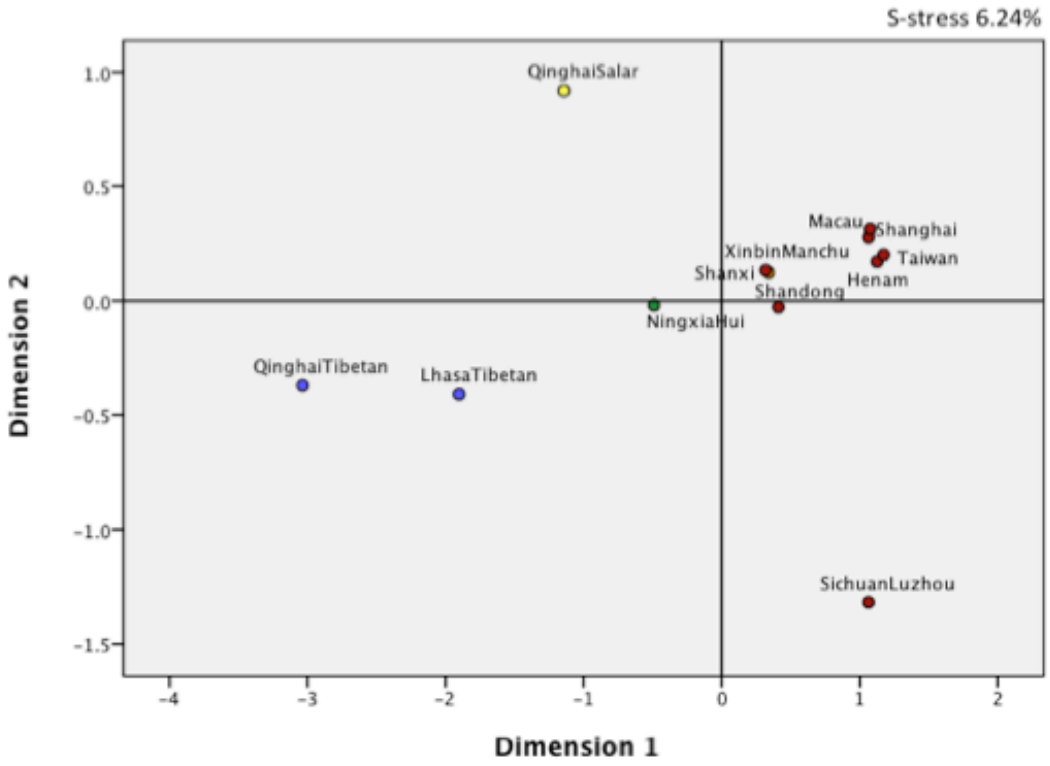


Figure 15 – **R_{st} Genetic Distance MDS representation**. In dark red are represented the Han ethnic populations, blue the Tibetan populations, in yellow and green are the Muslim populations, Salar and Hui, respectively. In black is represented the Manchu ethnic population.

Multidimensional scaling is widely used to assess and to visualize similarities or dissimilarities among a set of objects. The fitness of these results is measured by the value of S-stress, which evaluates the diagram's stress to the dimensional fit. As described by Jorge Rocha (2000) in "A Multidimensional Scaling Stress Evaluation", when working with twelve objects in a two dimensions Multidimensional scale, the value of S-stress should fall under 18.3%, a value high above the S-stress presented by our results, conferring them a high power of discrimination and significance.

A major cluster of populations appears located in the up right side of the plot. This cluster integrates the majority of Han populations plus the Xinbin Manchu. As did indicate the analysis of values of genetic distances, the Xinbin Manchu are

positioned close to the Han populations, which is in accordance with the known historical interactions between Manchu and Han.

The two Tibetan populations form a distinct cluster occupying the centre left side of the plot. Even though being from different regions – Qinghai and Tibet – the two populations evidence clear genetic ties, meaning therefore that geographical distance has not neutralized the influence of a shared culture and history.

The two Muslim populations, Hui and Salar, are also clearly separated from the remaining Chinese, lying in the upper left quadrant of the plot.

As for the Sichuan population, despite being aligned with other Han populations when it comes to the first dimension of the plot, the second dimension sharply discriminates this group, pushing the Sichuan alone from Hans to the lower right quadrant of the MDS.

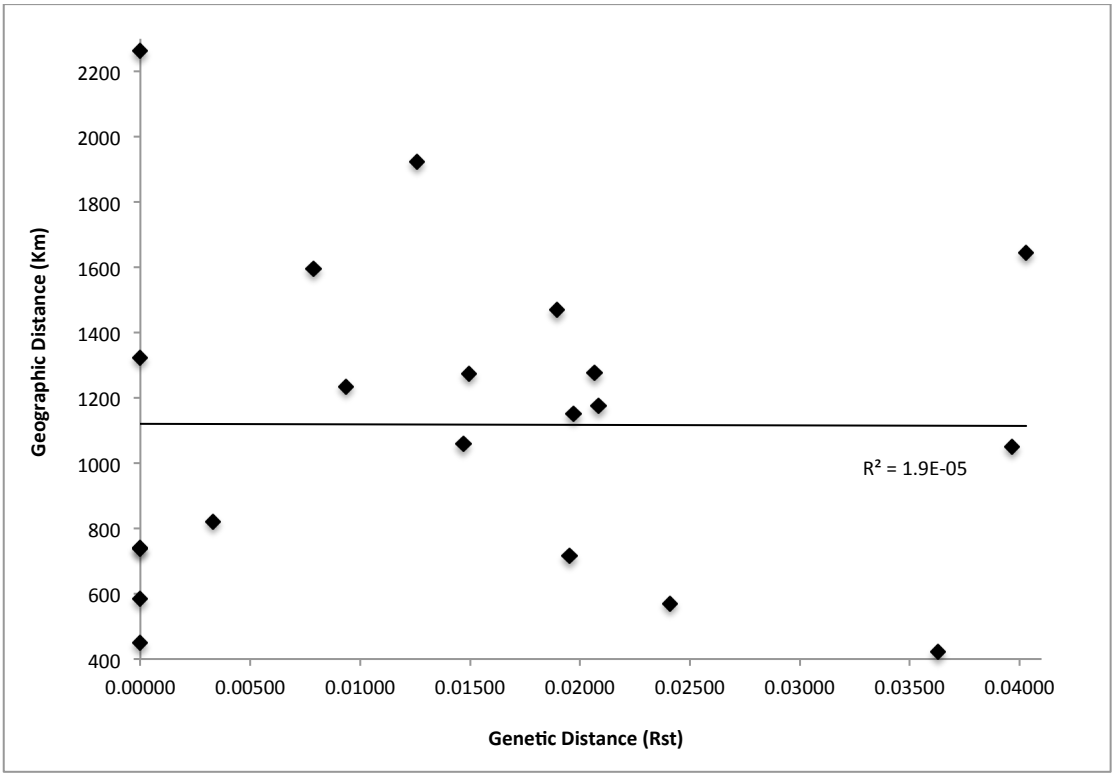


Figure 16 – Correlation map between genetic (R_{st}) distance and geographic distance (Km) among the 12 populations used for this study.

Ding *et al.* proposed, in the year 200, that the differences between North and South populations might be due mainly to geographic distance, than to a reflection of southern origin of the Chinese populations.

In the Figure 16 is graphically presented the correlation between genetic distance (R_{st}) and the geographic distance between the twelve populations used in this study. For this analysis, in order to try to rule out the ethnic factor, only Han populations were used. Also the population from Sichuan, Luzhou, was excluded from this analysis due to their particular history peculiar genetic profile, as was shown in the previously MDS analsysis.

Analysis of Figure 16, indicates that the geographical distance gradient doesn't have an influence in the genetic distances. By other words, there is no correlation between the genetic distances and how far, geographically, populations are from each other. It is, therefore, not clear that the geographic distance might be a factor of differentiation between populations, as previously proposed by Ding *et al.* (2000). This results are, however, not surprising, since China one of the biggest countries in the world, where more than one factor has to be taken into account, from the cultural factors, to the ancient and modern migrations inside the territory.

SAMOVA Analysis

The program implemented in SAMOVA runs in a way dependent on a number of populations groups chosen *a priori*, grouping then populations according to the pre-established pattern. The clustering relies on three parameters of population differentiation measurements – FCT, FST and FSC.

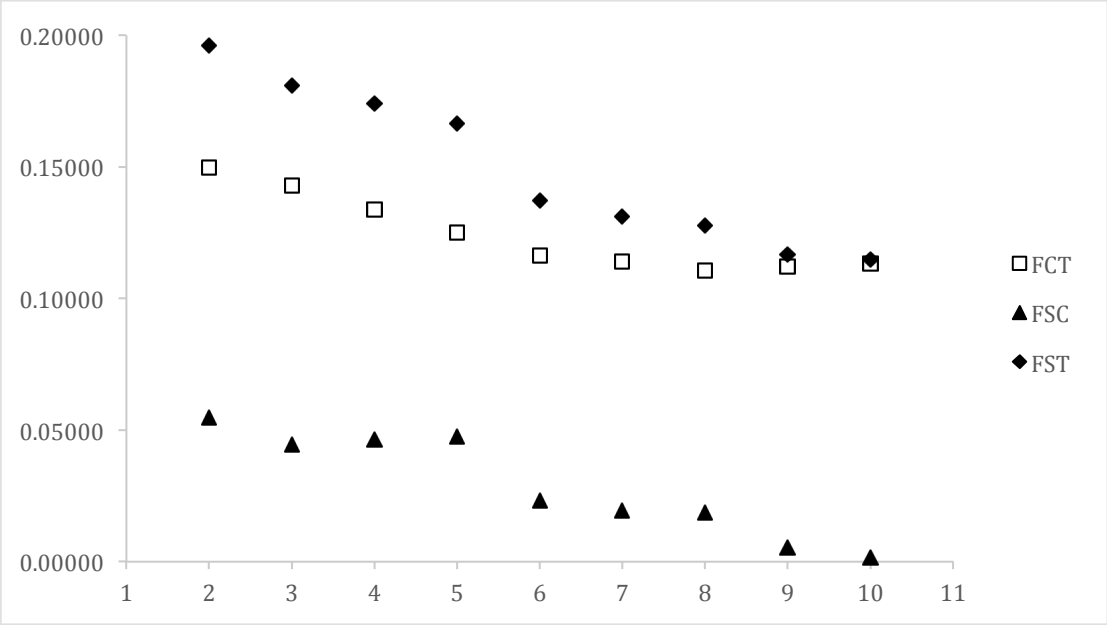


Figure 17 – Graphic representation of the FCT, FSC and FST values for 9 different possible group divisions of the 12 populations used in this study.

The FCT measures the level of differentiation among groups of populations, while the FSC regards the differentiation among populations within groups. It is possible to infer which model of populations clustering fits better the data looking for an optimum proportion between FCT and FSC and also accounting for the number of groups that must not be excessively high. That is achieved when the FCT is high enough, and consequently the differences between groups are elevated, while at the same time the populations inside each groups comparatively show poor differentiation, which is translated into a low FSC value.

We conducted iteratively SAMOVA, with the available Y-STR data for 12 populations, considering 2 to 10 as number of clusters, Comparing the obtained values of FCT, FSC and FST, which are graphically represented in Figure 17, we found that beginning with 2 groups and augmenting the number assumed, at 6 groups a qualitative leap occurred in the relative magnitude of FCT, FST and FSC,

since the FSC value substantially decreased, while the FCT diminished but much less sharply.

Another leap in the relative proportions of FCT, FST and FSC when 9 groups were assumed, this number was represented but in this case 9 groups were observed and this number appeared too high to allocate 12 populations.

So, we selected the 6-group system as the most adequate to capture structure in our data, and the produced distribution of populations *per group* presented a cluster of Han populations – Shanghai, Shandong, Shanxi, Taiwan, Henan, Xinbin Manchu and Macau – in the first group, where only the Hans from Sichuan Luzhou were allocated alone in a different group. These results are fully consistent with the data from the MDS study. The other remaining four groups are constituted by the two Tibetan populations – Qinghai and Lhasa - and the two Muslim populations – Salar and Hui.

Table 6 – SAMOVA results with the three parameters of population differentiation measurements for the 6-group division of the populations.

	% Of Variation	F-statistics	P-Value
Among Groups (FCT)	11.65	0.11646	0.00098
Among Populations Within Groups (FSC)	2.05	0.02321	0.00000
Within Populations (FST)	86.3	0.13696	0.00000

These results indicate that genetic similarities between populations are much influenced by ethnic and cultural affinities. Geography or geographical barriers did not emerged as playing significant role in the structuring pattern of current day Chinese populations (Table 6).

Furthermore, we compared the 6-group system obtained with the SAMOVA software with a forced Arlequin AMOVA, where the ethnicity of the populations was the main criteria to divide them. In that manner, the populations were grouped into three different clusters, the first composed by the assumed Han populations, and the two other composed by the Tibetan and Muslim populations.

Table 7 – Arlequin AMOVA results with the three parameters of population differentiation measurements for the proposed 3-group division of the populations, based solely on ethnicity.

	% Of Variation	F-statistics	P-Value
Among Groups (FCT)	12.50	0.12502	0.00098
Among Populations Within Groups (FSC)	2.64	0.03016	0.00000
Within Populations (FST)	84.86	0.15141	0.00000

The results obtained, present in Table 7, did not reveal major alterations, when compared to those obtained with the SAMOVA analysis (Table 6), given that the apportioning of variation and F-statistic indices were similar to the ones observed after SAMOVA.

Once again, also AMOVA indicate that population differentiation in China relies more in ethnic/cultural factors, than in geographic distances.

Barrier Analysis

If the multidimensional scaling (MDS) approach has provided a panorama on the proximity or dissimilarity between the studied populations, to obtain further insights on the factors that might had influenced that genetic landscape, we have performed an analysis of barriers, based also on R_{ST}.

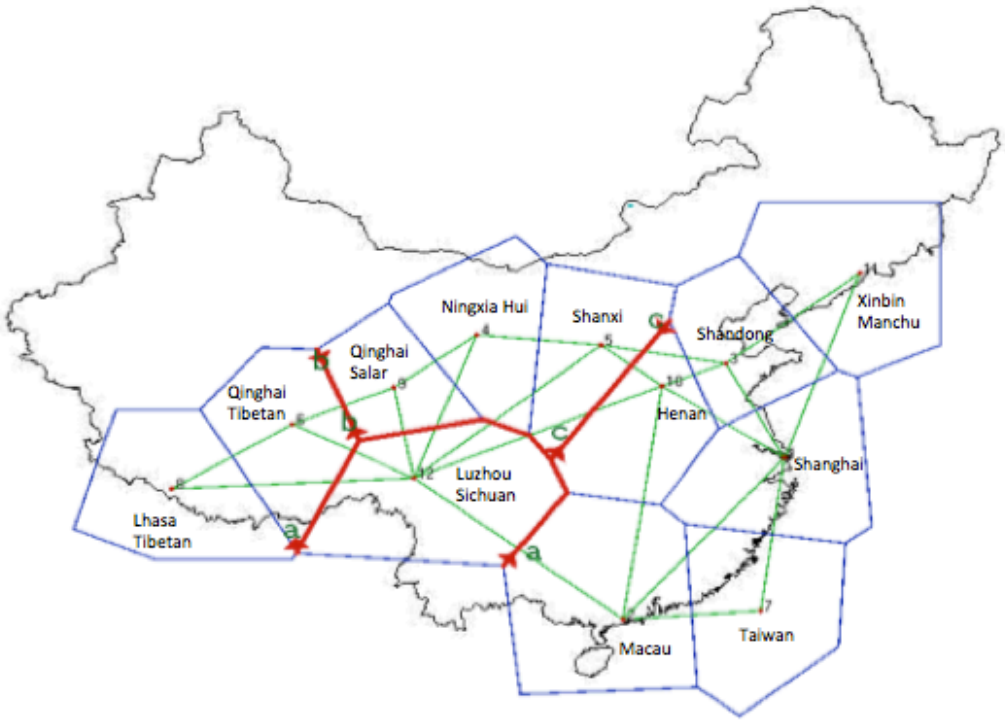


Figure 18 – Barrier representation of the R_{ST} genetic distances. The red arrows – a, b, c - represent the three calculated barriers.

The first barrier isolates the population of Luzhou, Sichuan (the a. red arrow on the Barrier analysis), which is consistent with the previously performed analyses that also pinpointed the differentiation of the Luzhou population, comparatively to other Han populations, despite the commonly assumed identical ethnic affiliation.

As well not surprisingly, the second assigned barrier totally isolated the Tibetan populations, a finding not only in full agreement with previous performed analyses but also with what is acknowledged to be the origin and the cultural background of this ethnic group.

The third barrier isolated a cluster that contained the two Muslim and the Shanxi populations, although all isolated from each other. Since these three

populations are neighbouring populations, and also since they share a dialect of Mandarin (Zhongyuan), probably gene flow between the Shanxi and the Muslim populations has contributed to erase differences between them.

Final Considerations and Conclusion

The major dispute about the peopling of East Asia among researchers is rooted in the different interpretations of the early migration history, which are based on the observed genetic divergence between northern (NEAS) and southern (SEAS) East Asian populations. [72, 73]

Su *et al*. (2001), examining Y chromosome SNPs from Chinese populations found out that the southern populations were more diversified than the northern populations. They concluded, therefore, that the northern populations derived from the southern populations after the initial Neolithic peopling of East Asia, an interpretation that was also supported in other studied. [74, 75]

However, Karafet *et al*. (2001) based on the lack of enough genetic divergence between SEAS and NEAS, did not sustain the hypothesis of early northward migration in East Asia.

Ding *et al.* (2000), claiming that the genetic divergence between SEAS and NEAS could be due only to isolation by distance, maintained a northern origin as a possible explanation. The authors also argued that the southern areas are heavily populated, whereas the northern areas are parsley populated, and consequently, between regions migration accompanied by high rates of genetic drift and lineage loss in northern groups could had accounted for an asymmetry in lineage composition, implying that a northern origin could not be ruled out.

More recent studies, such as that of Abdulla *et al.* (2009), that was based on a large scale survey of autosomal single nucleotide polymorphisms, showed that East Asian populations have a clinal structure, with haplotype diversity decreasing from south to north. The south/north cline was also demonstrated by previous mitochondrial DNA and Y chromosome studies by Kivisild *et al*. (2002), Shi *et al*. (2005) and Shi *et al*. (2008). The distribution of O haplogroup and subsequent clades has been interpreted as a mere reflection of the initial founder effect of the early northward effect, followed by the consequent geographic isolation. [65]

Diversity in East Asia was also strongly modelled by the demographic history of the past 5000 years, during which the major population migration in the region

occurred from north to south, following the expansion of the Han culture, although, as pointed out by Hong Shi *et al.* (2005), more recent population movements and admixture have to be taken into account, since they might have erased signatures of the early population movements [34]. For example, Karafet *et al*. (2001), provided evidence that the Hui and Uygurs are recently established NEAS with various degrees of admixture among them.are recently established NEAS with various degrees of admixture among them.

Viewing to extend the knowledge of Y chromosome diversity in China, in this work we have studied two Chinese populations, one from Shanghai and another from Macau. In both populations the vast majority of the Y chromosomes belonged to the O haplogroup - 81.5% in Shanghai and 89.58% in Macau, which was in fact expected since the O haplogroup is very ancient and overwhelmingly represented in East Asia.

Our Y-SNP results, seemingly supported the model initially proposed by Underhill *et al.* (2001) of a southeastern origin for the East Asian populations, with a afterward north migration, in the sense the southern populations presented, in some way, a pattern of higher haplogroup diversity.

Other factors, nevertheless, have also a key influence in the patterns of male diversity in China. Shanghai and Yunnan, two populations characterized by high levels of diversity, share the condition of having been, in the course of history, important trade points, with intense contact with other cultures and populations, which likely has played a major influence in their Y chromosome spectrum.

The haplotype diversity levels were, on the other hand, very high and similar across all Chinese populations. The Muslim populations presented the lower diversities, probably due to their religious beliefs that might have instigated certain isolation around these populations. They presented, however, high MNPD, which is probably a reflection of the differentiated lineages that entered China with the Muslims, who, after settling in the territory, assimilated lineages from neighbour populations. Contrarily, Tibetan populations were characterized by low MNPD values, meaning high molecular relatedness between haplotypes, which can be a sign of a rather homogeneous set of lineages that gave rise to the nowadays Tibetan populations.

The Manchu population, showed clear genetic affinities with the other Han populations, probably signalling the recent history of admixture between Manchu and neighbour Han people.

Also of note, the population from Luzhou, which despite considered to be Han people, consistently appeared quite differentiated in the cluster of Han populations.

The Luzhou populations, and especially the Sichuan province, have a very peculiar modern history, which might in part explain these results. In the middle of the 17th century the Sichuan province was conquered by rebels from the North – the province had already gone throw several years of war after a Manchu invasion – whom in response to the local resistance, massacred a great deal of the native population. [76] In the 18th century an earthquake resulted in the death of 100.000 people, and in the 20th century, between 1959 and 1961, the Great Chinese Famine also deeply affected this province, resulting in the death of over 9 million people (13% of the population at that time). [77, 78] Our data might be, accordingly, the result of the modern history of the region, which was shaped by warfare, famine and natural disasters.

Chinese populations were found to be highly substructured, however, the role of geographical distance was shown to be not significantly correlated with the degree of genetic relatedness between populations. Distant populations did not tended to be genetically more differentiate than neighbouring populations. The results revealed, however, that level of differentiation between populations was fairly related with ethnic-cultural aspects. Populations from the same ethnic group, such as Han, Tibetans or Muslim populations, tended to be more genetically related even when living quite far from each other, than were with the neighbouring populations of different ethnic affiliation.

The Macau population, with a long shared history with the Portuguese, who ruled the region for over 400 years, roused special interest. Contrarily to the reasonable expectation, no evidence was found testifying the genetic influence of the Portuguese settlers in Macau or the African slaves that were brought by them, according to the historical documentation. [51] It is remarkable that, despite being so long under the Portuguese rule, the population from Macau had not assimilated any European male lineages. This can likely be explained by the fact that the effective number of Portuguese in the territory was never much expressive.

As for the Shanghai population, the “Paris of the Orient” presented a higher Y-SNP gene diversity value, when compared to the Macau population, result that may be a consequence of the metropolitan characteristics of Shanghai, to where people were attracted to migrate, favouring population admixture and consequently increment in genetic diversity.

In conclusion, China harbours remarkable genetic diversity, but also anchors equally remarkable cultural, ethnic and linguistic diversity. China is the house for over 1 billion people, with different demography, culture and history.

Consequently, a very cautious approach needs to be taken when investigating the patterns of genetic diversity of Chinese populations. The gap between the official *minzu* categories and genetic structure is still understudied, which might introduce strong bias in the interpretation of overall patterns of genetic diversity. So, a much more fine genetic coverage of populations from China, not selected only by the *minzu* categorization, is needed to turn possible to infer a better picture of the global history of the region. Furthermore, without taking into account cultural and historical ties it will be difficult to understand the net of genetic relationships between Chinese populations.

1. Ohno, S., *Sex Chromosomes and Sex-linked genes*. Annals of Internal Medicine, 1968. **68**: p. 1375.
2. Jobling, M.A. and P. Gill, *Encoded evidence: DNA in forensic analysis.* Nat Rev Genet, 2004. **5**(10): p. 739-51.
3. Lahn, B.T. and D.C. Page, *Four evolutionary strata on the human X chromosome.* Science, 1999. **286**(5441): p. 964-7.
4. Charlesworth, B., *Evolution. When to be diploid.* Nature, 1991. **351**(6324): p. 273-4.
5. Skaletsky, H., et al., *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.* Nature, 2003. **423**(6942): p. 825-37.
6. Hughes, J.F., et al., *Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.* Nature, 2010. **463**(7280): p. 536-9.
7. Cavalli-Sforza, L.L., *The DNA revolution in population genetics.* Trends Genet, 1998. **14**(2): p. 60-5.
8. Sobrino, B., M. Brion, and A. Carracedo, *SNPs in forensic genetics: a review on SNP typing methodologies.* Forensic Sci Int, 2005. **154**(2-3): p. 181-94.
9. Goodwin, W., Linacre, A. & Hadi, S., *An Introduction to Forensic Genetics*, 2007, John Wiley & Sons Ltd.: England.
10. Brion, M., et al., *Hierarchical analysis of 30 Y-chromosome SNPs in European populations.* Int J Legal Med, 2005. **119**(1): p. 10-5.
11. Sobrino, B. and A. Carracedo, *SNP typing in forensic genetics: a review.* Methods Mol Biol, 2005. **297**: p. 107-26.
12. Amorim, A., Pereira, L., *Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs*. Forensic Science International, 2005. **150**(1): p. 17-21.
13. Karafet, T.M., et al., *New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree.* Genome Res, 2008. **18**(5): p. 830-8.
14. Cruciani, F., et al., *About the X-to-Y gene conversion rate.* Am J Hum Genet, 2010. **86**(3): p. 495-7; author reply 497-8.
15. Kayser, M., et al., *A comprehensive survey of human Y-chromosomal microsatellites.* Am J Hum Genet, 2004. **74**(6): p. 1183-97.
16. Kimura, M.a.O., T., *Stepwise mutation model and distribution of allelic frequencies in a finite population*. Proc Natl Acad Sci U S A, 1978. **75**(6): p. 2868–2872.
17. Ellegren, H., *Microsatellites: Simples sequences with complex evolution.* Nature Review Genetics, 2004. **5**: p. 435-445.
18. SCHLÖTTERER, C., *Evolutionary dynamics of microsattelite DNA*. Chromosoma, 2000. **109**(365-371).
19. Willuweit, S., L. Roewer, and Y.C.U.G. International Forensic, *Y chromosome haplotype reference database (YHRD): update.* Forensic Sci Int Genet, 2007. **1**(2): p. 83-7.
20. Avise, J.C., *Phylogeography: retrospect and prospect.* Journal of Biogeography, 2009. **36**: p. 3-15.
21. Avise, J.C., Arnold, J., Ball, R. M., Berminghan, E., Lamb, T., et al., *Intraspecific Phylogeography - the Mitochondrial-DNA Bridge between Population-Genetics and Systematics*. Annual Review of Ecology and Systematics, 1987. **18**: p. 489-522.

22. Jobling, M.A. and C. Tyler-Smith, *The human Y chromosome: an evolutionary marker comes of age.* Nat Rev Genet, 2003. **4**(8): p. 598-612.
23. Butler, J.M., *Forensic DNA typing: Methodology*. Academic Press, ed. Elsevier2012, San Diego, California.
24. Jobling, M.A., A. Pandya, and C. Tyler-Smith, *The Y chromosome in forensic analysis and paternity testing.* Int J Legal Med, 1997. **110**(3): p. 118-24.
25. Daniels, D.L., A.M. Hall, and J. Ballantyne, *SWGDAM developmental validation of a 19-locus Y-STR system for forensic casework.* J Forensic Sci, 2004. **49**(4): p. 668-83.
26. Prinz, M. and M. Sansone, *Y chromosome-specific short tandem repeats in forensic casework.* Croat Med J, 2001. **42**(3): p. 288-91.
27. Relethford, J.H., *Ancient DNA and the origin of modern humans.* Proc Natl Acad Sci U S A, 2001. **98**(2): p. 390-1.
28. Cavalli-Sforza, L.L. and M.W. Feldman, *The application of molecular genetic approaches to the study of human evolution.* Nat Genet, 2003. **33 Suppl**: p. 266-75.
29. Cavalli-Sforza, L.L.a.B., W.F., *The Genetics of Human Populations*. Vol. XVI. 1971, San Francisco: W.H. Freeman. 965.
30. L. Luca Cavalli-Sforza, P.M., & Alberto Piazza, *The History and Geography of Human Genes*. Vol. XI. 1994, Princeton, N. J.: Princeton University PRess. 518.
31. Jobling, M.A., Hurles, M. E. and C. Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples & Disease*2004: Garland Publishing.
32. Zhong, H., 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): p. 717-27.
33. Wang, C.-C.a.L., Hui, *Inferring human history in East Asia from Y chromosomes*. Investigative Genetics, 2013. **4**(11): p. 1-10.
34. Shi, H., et al., *Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122.* Am J Hum Genet, 2005. **77**(3): p. 408-19.
35. Underhill, P.A., et al., *The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations.* Ann Hum Genet, 2001. **65**(Pt 1): p. 43-62.
36. Underhill, P., *A Synopsis of extant Y chromosome diversity in East Asia and Oceania, in the Peopling of East Asia, Putting Together Archaeology, Linguistics and Genetics*2004, Routledge Curzon, London: L. Sagart, R. Blench and A. Sanchez-Mazas.
37. Su, B., et al., *Y chromosome haplotypes reveal prehistorical migrations to the Himalayas.* Hum Genet, 2000. **107**(6): p. 582-90.
38. Wen, B., et al., *Genetic evidence supports demic diffusion of Han culture.* Nature, 2004. **431**(7006): p. 302-5.
39. Wen, B., et al., *Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans.* Am J Hum Genet, 2004. **74**(5): p. 856-65.
40. PRB, *World Population Data Sheet*. Population Reference Bureau, 2007.
41. Peters, H.A., Journal of East Asian Archaeology, 2001. **3**: p. 75-103.

42. Bittles, A.a.C., Y.-Y., *Eugenics and population policies, in Human Biology and Social Inequality,* ed. C.U. Press. Vol. Cambridge University Press. 1998, Cambridge: Strickland and P.S. Shetty (Eds.). 272-287.
43. Fairbank, J.K.a.R., E.O., *China, Tradition & Transformation,* ed. A. Unwin1989, Sydney: Allen & Unwin.
44. Du, R.a.Y., V.F., *Ethnic Groups In China,* ed. S. Press1993, Beijing.
45. Gladney, L., *Besides the Hui people, nine other officially recognized ethnic groups of PRC are considered predominantly Muslim. .* 1997. **XXIII**: p. 18-20.
46. Gayden, T., et al., *The Himalayas: barrier and conduit for gene flow.* Am J Phys Anthropol, 2013. **151**(2): p. 169-82.
47. Beall, C.M., et al., *Higher offspring survival among Tibetan women with high oxygen saturation genotypes residing at 4,000 m.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 14300-4.
48. Qi, X., et al., *Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau.* Mol Biol Evol, 2013. **30**(8): p. 1761-78.
49. Danielson, E.N., *Shanghai and the Yangtze Delta.* 2004: p. 10-11.
50. W., W., City Profile Shanghai, 1999. **16**(3): p. 207-216.
51. Souza, G.B., *The Survival of the Empire: Portuguese trade and society in China*. Cambridge University Press. Vol. 32. 2004.
52. Sambrook, E.F.F., Maniatis, T., *Molecular Cloning: A Laboratory Manual,* in *Cold Spring Harbor Laboratory Press*1989: New York.
53. Gusmao, L., et al., *DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis.* Int J Legal Med, 2006. **120**(4): p. 191-200.
54. Mulero, J.J., et al., *Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system.* J Forensic Sci, 2006. **51**(1): p. 64-75.
55. van Oven, M., N. van den Tempel, and M. Kayser, *A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia.* J Hum Genet, 2012. **57**(1): p. 65-9.
56. Budowle, B., et al., *Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE.* Am J Hum Genet, 1991. **48**(1): p. 137-44.
57. Yan, S., et al., *An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4.* Eur J Hum Genet, 2011. **19**(9): p. 1013-5.
58. Hammer, M.F., et al., *Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes.* J Hum Genet, 2006. **51**(1): p. 47-58.
59. Chuan-Chao Wang, S.Y., Zhen-Dong Qin, Yan Lu, Qi-Liang Ding, Lan-Hei Weir, Shin-Lin Li, Ya-Jun Yang, Li Jin, Hui Li, *Late Neolithic Expansion of Ancient Chinese revealed by Y chromosome haplogroup O3a1c-002611.* Journal of Systematics an Evolution, 2013. **51**(3): p. 280-286.
60. Zhong, H., et al., *Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia.* J Hum Genet, 2010. **55**(7): p. 428-35.

61. Shi, H., 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**: p. 45.
62. Karafet, T., et al., *Paternal population history of East Asia: sources, patterns, and microevolutionary processes.* Am J Hum Genet, 2001. **69**(3): p. 615-28.
63. Rootsi, S., et al., *A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe.* Eur J Hum Genet, 2007. **15**(2): p. 204-11.
64. Derenko, M., et al., *Y-chromosome haplogroup N dispersals from south Siberia to Europe.* J Hum Genet, 2007. **52**(9): p. 763-70.
65. Kim, S.H., et al., *High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea.* Investig Genet, 2011. **2**(1): p. 10.
66. Zheng, L., et al., *Y chromosomal STR polymorphism in northern Chinese populations.* Biol Res, 2009. **42**(4): p. 497-504.
67. Bai, R., et al., *Haplotype diversity of 17 Y-STR loci in a Chinese Han population sample from Shanxi Province, Northern China.* Forensic Sci Int Genet, 2013. **7**(1): p. 214-6.
68. Yan, J., et al., *Genetic polymorphisms of 17 Y-STRs haplotypes in Chinese Han population residing in Shandong province of China.* Leg Med (Tokyo), 2007. **9**(4): p. 196-202.
69. Wu, W., et al., *Population genetics of 17 Y-STR loci in a large Chinese Han population from Zhejiang Province, Eastern China.* Forensic Sci Int Genet, 2011. **5**(1): p. e11-3.
70. Black, M.L., Wang, W., Bittles, A. H., *Unity and diversity: Genetic studies on the population of China*. Recent Advances in Molecular Biology and Evolution: Applications to Biological Anthropology2007. 347-371.
71. Tapp, N., *Asian Ethnicity*. Vol. 3. 2002. 63-84.
72. Su, B., 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): p. 1718-24.
73. Jin, L. and B. Su, *Natives or immigrants: modern human origin in east Asia.* Nat Rev Genet, 2000. **1**(2): p. 126-33.
74. Underhill, P.A., et al., *Y chromosome sequence variation and the history of human populations.* Nat Genet, 2000. **26**(3): p. 358-61.
75. Ke, Y., et al., *African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes.* Science, 2001. **292**(5519): p. 1151-3.
76. Dai, Y., *The Sichuan Frontier and Tibet: Imperial Strategy in the Early Qing*, ed. U.o.W. Press2009.
77. Shuji, C., *大饑荒：1959－1961 年的中国人口*, 2005: Hong Kong.
78. Schuster, R.L.a.G.F.W., *Landslide triggers and types*, in *Landslides: Proceedings of the First European Conference on Landslides*2002, A.A. Balkema Publishers.

SUPPLEMENTARY DATA

Table S1 – Multiplex O PCR and SBE primers.

Mutations	Primers	SBE (5'-3')	Reference
M175	F - CCCAAATCAACTAACTCCAG R - TTCTACTGATACCTTTGTTTCTGTTCA	F - CACATGCCTTCTCACTTCTC	(van Oven et al. 2012)
M119	F - CAAACCGCAGTGCTATGTGT R - TGGGTTATTCCAATTCAGCA	R - CCAATTCAGCATACAGGC	(van Oven et al. 2012)
P203	F - GGCTACATGGAAATGGTTGG R - TTCTCACTTAGCACATATACAAAAGGT	F - GGCTATTGAGTTAGCATAATCA	(van Oven et al. 2012)
M110	F - CGAGAACGTTCCTGTCACAA R - AAAATCCAACGACAAATGTGC	F - GGATGCCGGTACAATGTATT	(van Oven et al. 2012)
M268	F - CATGCCTAGCCTCATTCCTC R - AACCCTGATCATTCCCCTTC	F - CCTAGCCTCATTCCTCTAAAAT	(van Oven et al. 2012)
M95	F - CCTTCTTGGGATCAAATGGA R - GCCTACAGGTTGGAAAGGCTA	F - GATAAGGAAAGACTACCATATTAGTG	(van Oven et al. 2012)
M88	F - GCTATGGCCTAGGTGCTTTTC R - CACAGGCCTTAGAGAGGTAGTCA	F - CTTATTCCTGCTTCTTCTGC	(van Oven et al. 2012)
M176	F - ATCCCGCTTCGGTACTCTG R - TCTTGAGTGTGTGGCTTTCG	R - TGTTGTCCAGTTGCACTTC	(van Oven et al. 2012)
M122	F - TTAGTTGCCTTTTGGAAATGAA R - TCAGATTTTCCCCTGAGAGC	R - AGATTTTCCCCTGAGAGC	(van Oven et al. 2012)
M324	F - GATTTGATCTACCTGCCCTTTC R - ATACATGGGCTGCAACAAGA	R - ATGGGCTGCAACAAGA	(van Oven et al. 2012)
KL1	F - TAGATGGTTGAGACACATCTTCA R - ACTCCAGGATGTTTGGGAAC	R - GGTTGCTAGAATTGCACAAT	(van Oven et al. 2012)
002611	F - GCCCTGCTAGTAGGCACCA R - AATCCAATGACCCTTTGCAG	R - AAGTGCAGCAGTGGCC	(van Oven et al. 2012)
P201	F - GTGCTGTGCAAGTTGTGTGA R - TGGGTGCAGTTAAGCAATGA	R - GAGAGCCAGTTAAAGCCC	(van Oven et al. 2012)
M7	F - CATCACCAAAGGGCATGTAAT R - TTGTCCCTGCAGCCTTGT	F - CCTCTCTGTATGTCAGATCTAACAA	(van Oven et al. 2012)
M134	F - AAGAAAAGGCCCAGGAAAGT	R - CTTTTGATCCCCACCAAT	(van Oven et al. 2012)

Table S2 – Multiplex 1 PCR and SBE primers.

Mutations	Primers	SBE (5'-3')	Reference
M13	F - TCCTAACCTGGTGGTCTTTCA R - GCCATGATTTTATCCAACCAC	F - AAGGGCAAGACGGTTA	(Brion et al., 2005)
SRY1532	F - TCCTTAGCAACCATTAATCTGG R - AAATAGCAAAAACTGACACAAGGC	F - TTGTATCTGACTTTTTCACACAGT	(Brion et al., 2005)
M213	F - GGCCATATAAAAACGCAGCA R - TGAATGGCAAATTGATTCCA	R - TCAGAACTTAAAACATCTCGTTAC	(Brion et al., 2005)
M9	F - GCAGCATATAAAACTTTCAGG R - AAAACCTAACTTTGCTCAAGC	F - GGCCTAAGATGGTTGAAT	(Brion et al., 2005)
Tat	F - GACTCTGAGTGTAGACTTGTGA R - GAAGGTGCCGTAAAAGTGTGAA	R - CTCTGAAATATTAAATTAAAACAAC	(Gomes et al., 2010)
92R7	F - TGCATGAACACAAAAGACGTA R - GCATTGTTAAATATGACCAGC	R - GCATGAACACAAAAGACGTAGAAG	(Brion et al., 2005)
M173	F - GCACAGTACTCACTTTAGGTTTGC R - GCAGTTTTCCCAGATCCTGA	F - CTTACAATTCAAGGGCATTTAGAAC	(Brion et al., 2005)
M70	F - TCATAGCCCACTATACTTTGGAC R - CTGAGGGCTGGACTATAGGG	R - TAGGGATTCTGTTGTGGTAGTCTTAG	(Brion et al., 2005)
P25	F - GGACCATCACCTGGGTAAAGT R - AGTGCTTGTCCAAGGCAGTA	F - TCTGCCTGAAACCTGCCTG	(Brion et al., 2005)

Table S3 – YAP PCR primers.

Mutations	Primers	Reference
YAP	F – ACTGCTAAAAGGGGATGGAT R - CAGGGGAAGATAAAGAAATA	(Hammer and Horai, 1995)

Table S4 – Macau haplogroup data. The haplotypes are presented with the following Y-STR order: DYS456, DYS389 I, DYS390, DYS389 II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS392, YGATA H4, DYS437, DYS438 and DYS448.

Haplotype Code	Haplotype Frequency	Y-STR Haplotypes																	
M1	1	16	13	24	29	16	17	15	12	18	14	11	12	22	13	11	15	10	20
M2	1	15	12	19	28	16	17	15	13	13	13	11	12	20	13	12	14	10	18
M3	1	15	12	24	28	16	18	17	12	19	12	10	11	23	13	12	16	10	20
M4	1	15	12	24	28	16	18	14	13	18	12	10	13	20	14	11	15	11	20
M5	1	15	14	25	30	16	20	14	13	18	12	9	11	21	13	11	15	10	22
M6	1	15	13	25	29	16	17	16	14	18	15	10	11	22	13	10	14	10	18
M7	1	14	12	25	28	16	17	16	14	19	12	10	12	23	13	11	15	11	20
M8	1	16	12	23	28	16	16	15	13	13	13	11	11	19	14	12	15	10	18
M9	1	16	12	23	27	15	20	15	12	12	12	10	13	20	12	11	15	10	19
M10	1	15	12	24	28	16	19	13	13	17	12	10	13	20	14	12	15	11	20
M11	1	16	12	23	28	16	16	15	13	13	13	11	11	19	14	12	14	10	18
M12	1	15	12	25	29	17	16	14	13	19	12	11	12	20	14	12	15	11	20
M13	1	15	12	26	28	16	18	16	13	21	12	10	14	24	13	11	14	11	18
M14	1	14	12	23	29	17	17	15	12	17	12	10	13	19	12	12	15	10	19
M15	1	15	13	23	29	16	15	14	13	13	13	11	12	21	14	12	15	10	18
M16	1	15	12	24	28	16	18	15	14	19	12	10	12	20	14	12	15	11	20
M17	1	15	12	24	29	17	16	17	17	17	12	10	11	21	13	11	14	10	20
M18	1	14	13	23	32	19	16	15	12	19	12	10	10	21	13	11	13	10	19
M19	1	16	12	23	29	17	15	15	13	13	13	10	11	21	14	12	14	10	18
M20	1	15	12	24	28	16	21	15	15	16	12	10	11	20	12	11	14	11	19
M21	1	15	12	26	27	15	17	14	13	19	12	11	12	20	14	12	15	11	20
M22	1	15	13	25	29	16	18	15	13	20	12	9	12	21	13	11	15	10	19
M23	1	15	13	23	30	17	16	15	11	13	13	11	11	21	15	13	14	10	18
M24	1	17	12	23	29	17	15	15	13	13	13	10	11	19	14	11	14	10	18
M25	1	15	12	23	27	15	19	14	15	20	12	10	12	21	14	12	15	11	20
M26	1	16	13	23	29	16	21	15	11	18	13	10	11	21	13	11	15	10	18
M27	1	15	13	24	29	16	17	15	14	18	14	10	11	24	13	10	14	10	18
M28	1	15	14	22	30	16	15	14	11	11	14	12	11	22	14	11	14	10	19
M29	1	17	12	23	28	16	16	16	13	13	13	10	11	19	14	12	14	10	?
M30	1	15	13	24	28	15	18	16	13	19	14	11	13	22	13	11	14	10	17
M31	1	15	15	25	32	17	16	15	13	14	13	11	12	23	11	12	14	9	21

Table S4 (Continued) – Macau haplogroup data. The haplotypes are presented with the following Y-STR order: DYS456, DYS389 I, DYS390, DYS389 II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS392, YGATA H4, DYS437, DYS438 and DYS448.

M32	1	15	12	25	27	15	17	15	14	19	13	10	14	20	14	12	16	11	20
M33	1	15	14	24	30	16	18	16	13	18	12	10	12	21	14	12	15	10	20
M34	1	15	14	24	31	17	19	15	13	19	14	11	12	24	14	11	14	10	18
M35	1	15	12	25	28	16	21	14	13	17	12	10	13	20	14	12	15	11	20
M36	1	17	12	23	29	17	17	15	12	19	12	10	12	21	12	12	15	10	19
M37	1	15	13	22	28	15	19	15	12	16	12	10	12	21	13	11	15	10	19
M38	1	15	13	25	30	17	16	15	13	19	14	11	12	23	13	11	14	10	18
M39	1	17	12	23	30	18	14	15	13	14	14	9	11	21	14	12	14	10	17
M40	1	15	13	24	30	17	15	15	11	19	16	10	11	21	11	11	14	10	21
M41	1	15	12	24	27	15	21	14	13	21	12	10	11	20	14	12	15	11	19
M42	1	15	12	24	28	16	18	15	15	20	14	10	11	21	13	10	14	10	18
M43	1	16	12	23	28	16	19	14	11	16	12	10	13	20	12	13	16	10	19
M44	1	15	13	23	29	16	15	14	13	13	13	10	13	20	14	12	15	10	18
M45	1	15	12	23	28	16	18	14	11	19	12	10	12	20	15	13	15	11	20
M46	1	15	13	24	28	15	18	16	13	13	14	10	13	21	14	11	16	10	?
M47	1	15	12	23	28	16	15	15	12	13	13	10	13	20	11	12	14	10	18
M48	1	14	12	24	28	16	16	15	12	19	12	11	13	20	13	12	14	10	19

Table S5 – Shanghai haplogroup data. The haplotypes are presented with the following Y-STR order: DYS456, DYS389 I, DYS390, DYS389 II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS392, YGATA H4, DYS437, DYS438 and DYS448.

Haplotype Code	Haplotype Frequency	Y-STR Haplotypes																
S1	1	17	14	22	17	15	14	11	12	13	10	11	22	12	12	14	10	19
S2	2	17	12	24	17	16	16	12	20	12	10	11	21	13	11	15	10	19
S3	1	15	11	25	17	18	16	13	19	14	11	11	21	13	11	15	10	18
S4	1	16	12	23	16	18	15	12	17	12	10	12	19	12	12	14	10	19
S5	1	15	12	25	15	18	16	12	20	12	10	12	22	15	12	13	10	19
S6	1	15	12	23	18	19	16	12	17	12	10	12	19	12	12	16	10	19
S7	1	16	12	23	16	20	14	13	17	12	10	12	21	14	12	15	11	20
S8	1	17	12	23	17	15	16	13	13	13	11	11	19	14	12	14	10	18
S9	1	17	12	23	17	15	16	13	13	13	11	11	19	14	12	14	10	18
S10	1	15	13	22	17	20	15	11	11	12	10	12	21	13	12	15	10	19
S11	1	15	12	24	15	19	14	13	21	12	10	12	20	14	12	15	10	20
S12	1	15	12	24	17	19	15	12	20	12	10	11	21	13	11	15	10	20
S13	1	14	13	24	16	17	14	13	13	12	10	12	20	14	12	15	11	21
S14	1	17	13	23	17	15	15	13	13	16	11	12	19	14	13	14	10	20
S15	1	17	12	23	15	15	16	12	13	13	11	12	21	14	12	14	10	19
S16	1	16	12	23	16	16	14	13	19	12	10	12	20	14	13	15	11	20
S17	1	18	12	23	15	15	15	12	13	13	11	13	20	14	12	14	10	18
S18	1	15	12	24	16	16	14	13	20	12	10	13	21	14	11	15	11	20
S19	1	15	12	25	16	20	17	13	19	12	9	12	21	13	12	14	10	19
S20	1	16	12	24	17	16	16	12	20	12	10	11	21	13	11	15	10	19
S21	1	16	12	23	17	15	15	13	14	13	11	11	19	14	12	14	10	18
S22	1	15	12	25	15	19	15	12	21	12	10	12	23	15	12	13	10	19
S23	1	17	13	22	16	16	15	12	17	14	10	12	21	13	12	14	11	18
S24	1	16	12	24	17	15	16	13	13	12	11	11	23	12	12	15	10	19
S25	1	15	14	24	16	15	15	12	17	15	10	12	21	13	12	14	10	18
S26	1	15	12	25	16	19	17	15	21	12	9	12	21	13	12	14	10	19
S27	1	15	14	23	16	15	16	12	18	15	9	11	21	11	11	15	10	21
S28	1	17	12	23	15	15	17	12	13	13	11	12	20	14	12	14	10	18
S29	1	15	14	24	17	14	15	12	17	14	10	11	21	13	12	14	10	18
S30	1	16	12	22	18	15	15	13	14	13	12	13	21	14	12	14	10	18

Table S5 (Continued) – Shanghai haplogroup data. The haplotypes are presented with the following Y-STR order: DYS456, DYS389 I, DYS390, DYS389 II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS392, YGATA H4, DYS437, DYS438 and DYS448.

S31	1	17	12	23	15	15	17	12	13	13	11	12	20	14	12	14	10	18
S32	1	17	12	23	17	15	16	13	14	13	11	11	19	14	12	14	10	18
S33	1	14	13	24	16	17	14	12	12	12	10	12	20	14	12	15	11	21
S34	1	15	13	24	16	19	14	14	18	12	10	12	20	14	12	15	11	20
S35	1	18	12	23	15	15	16	12	13	13	11	12	20	14	12	14	10	18
S36	1	16	12	24	16	20	15	16	17	14	10	11	22	13	10	13	10	18
S37	1	17	12	23	16	15	16	13	13	13	12	11	19	13	12	14	10	18
S38	1	16	12	23	16	19	15	12	17	12	10	11	20	12	12	14	10	19
S39	1	16	12	24	18	16	16	12	18	12	10	10	21	13	11	15	9	19
S40	1	15	13	23	16	18	15	14	20	15	10	11	22	11	10	14	10	21
S41	1	15	14	24	16	15	15	12	17	14	10	11	21	13	12	14	10	18
S42	1	16	12	23	16	17	16	13	14	12	10	11	20	13	13	14	10	18
S43	1	14	12	25	16	20	17	12	18	12	10	11	22	13	11	14	10	18
S44	1	18	12	23	16	15	15	13	13	13	10	11	19	14	12	14	10	18
S45	1	14	12	24	18	16	15	13	21	12	10	12	21	13	13	15	10	20
S46	1	14	12	25	17	16	15	13	18	12	11	11	21	13	13	15	10	20
S47	1	15	13	23	17	16	15	12	17	14	11	12	22	13	12	14	10	18
S48	1	15	14	24	15	16	15	12	17	14	10	11	21	13	12	14	10	18
S49	1	15	12	25	16	17	15	11	16	12	10	11	19	12	12	15	10	19
S50	1	17	12	24	16	17	14	14	18	12	10	12	20	14	12	16	11	20
S51	1	15	12	23	17	17	14	13	19	12	10	11	20	14	11	14	11	20
S52	1	15	13	23	18	16	15	10	19	13	10	11	21	11	11	14	10	21
S53	1	15	12	23	17	15	15	13	13	13	10	11	19	14	13	14	10	18
S54	1	15	14	23	16	16	15	11	12	13	10	11	24	14	12	14	11	19
S55	1	15	12	24	15	16	15	11	17	12	10	11	21	14	11	14	11	20
S56	1	15	13	25	18	17	15	11	19	13	10	11	21	13	11	14	10	19
S57	1	15	13	23	17	13	15	12	18	12	10	10	21	13	13	15	9	19
S58	1	15	14	22	17	18	15	11	14	13	10	11	21	14	12	14	10	18
S59	1	14	13	25	15	19	17	13	19	12	10	12	24	13	13	15	10	19
S60	1	15	13	25	16	17	17	13	19	14	11	12	22	13	11	14	10	18
S61	1	15	13	23	16	15	15	11	12	13	11	11	23	14	12	14	12	19
S62	1	16	14	23	17	17	15	11	16	12	11	12	19	13	12	15	10	19
S63	1	15	14	24	16	16	15	11	17	15	10	12	21	11	12	14	10	22

Table S5 (Continued) – Shanghai haplogroup data. The haplotypes are presented with the following Y-STR order: DYS456, DYS389 I, DYS390, DYS389 II, DYS458, DYS19, DYS385a, DYS385b, DYS393, DYS391, DYS439, DYS635, DYS392, YGATA H4, DYS437, DYS438 and DYS448.

S64	1	15	14	23	20	14	14	11	12	13	11	12	22	14	12	14	11	20
S65	1	15	14	24	16	16	16	12	13	13	10	11	23	13	12	14	11	19
S66	1	15	12	23	15	19	15	12	16	13	10	12	21	12	11	15	10	20
S67	1	15	13	23	15	19	15	11	18	13	10	12	22	11	11	14	10	21
S68	1	14	12	25	16	17	15	12	20	12	9	11	22	13	12	14	10	19
S69	1	15	15	23	17	17	16	11	16	15	10	11	21	11	12	14	10	21
S70	1	17	13	22	17	14	15	11	12	14	10	10	22	14	11	14	10	19
S71	1	15	12	21	16	18	15	11	18	12	10	13	24	11	11	14	10	20
S72	1	15	12	24	15	17	14	13	15	12	10	12	20	14	11	16	11	20
S73	1	14	12	25	19	16	15	13	21	12	10	11	21	13	13	15	10	20
S74	1	15	13	24	16	18	15	13	18	12	10	12	20	14	11	15	11	20
S75	1	16	12	23	17	16	15	12	14	13	10	13	22	14	12	14	10	18
S76	1	14	14	23	16	17	14	11	12	13	10	12	20	14	10	14	10	18
S77	1	16	12	23	17	15	15	13	14	13	10	11	20	15	12	14	10	18
S78	1	16	14	22	17	17	15	11	11	12	10	12	21	13	11	15	10	19
S79	1	17	12	23	15	15	16	12	13	13	11	12	20	14	13	14	10	18
S80	1	15	12	23	16	18	15	12	16	12	11	11	19	12	12	15	10	19

Table 6 – AMOVA results. All the statistical significant values are highlighted in bold (P<0.05).

	Y-STR (%)	Significance Test (P-value)
Geographic Distinction: North/South		
Among Groups	0.00	0.65103 +- 0.00000
Among Populations Within Groups	**10.64**	**0.00000 +- 0.00000**
Within Populations	**90.31**	**0.00000 +- 0.00000**
Geographic Distinction: Cost/Interior		
Among Groups	2.37	0.13196 +- 0.01129
Among Populations Within Groups	**8.86**	**0.00000 +- 0.00000**
Within Populations	**88.78**	**0.00000 +- 0.00000**
Ethnic Distinction: Muslims/Tibetans/Han		
Among Groups	**12.50**	**0.00000 +- 0.00000**
Among Populations Within Groups	**2.64**	**0.00000 +- 0.00000**
Within Populations	**84.86**	**0.00000 +- 0.00000**
Ethnic Distinction: Muslims/Tibetans/Han/Manchu		
Among Groups	**9.63**	**0.00782 +- 0.00280**
Among Populations Within Groups	**3.28**	**0.00000 +- 0.00000**
Within Populations	**87.09**	**0.00000 +- 0.00000**