Present Y chromosomes refute the Roma/Gypsy origin of the Xuejiawan people in Northwest China

Shaoqing Wen, Dan Xu, Hongbing Yao, Hui Li

Shaoqing WEN, Dan XU, Hongbing YAO, Hui LI
Present Y chromosomes refute the Roma/Gypsy origin of the Xuejiawan people in Northwest China
In Dan Xu and Hui Li (eds.) 2017. *Languages and Genes in Northwestern China and Adjacent Regions*. pp107-120. Singapore: Springer Nature

Present Y chromosomes refute the Roma/Gypsy origin of the Xuejiawan people in Northwest China

Shaoqing WEN, Dan XU, Hongbing YAO, Hui LI

1. Introduction

Northwestern China is rich in human genetic and linguistic resources. Notably, it is here that the ancient Silk Road snaked from its eastern terminus in Xi'an across the mountains and deserts to the West, leaving behind a number of ethnic groups with different religious faiths, cultures and customs due to East-West intercommunications. Early in 1947, the Chinese journalist Zhu Tian proposed that the Xuejiawan people in Yongdeng County, once an important town on the ancient Silk Road, were eastern Gypsies (Zhu, 1947). Thereafter, compared to the neighboring Han farmers, this special population who has made their living mainly by fortune-telling has generated enormous publicity. So far, dozens of articles have been published discussing the ethnic origins of the Xuejiawan people from various perspectives (Tuo, 2006; Yang, 1991; Wu *et al.*, 2001; Wu, 1990; Wu, 1991; Guan and Zhang, 2012). Based on folkloric and historic comparisons, some scholars hold the ancient Gypsy origin hypothesis, i.e. that the Xuejiawan people were the descendents of gypsies who appeared in China in the Yuan Dynasty (Yang, 1991) or Qing Dynasty (Zhu, 1947). Others support the native Chinese origin hypothesis, which is further divided into two alternative hypotheses, the southern Hmong-Mien and Han Chinese origin hypotheses. The former was propounded on the basis of incomplete lexical comparison (Wu *et al.*, 2001), while the latter was presented based on local historical anthropologists' long-term fieldwork (Wu, 1990; Wu, 1991; Guan and Zhang, 2012). Among these hypotheses, the Gypsy origin hypothesis is the most famous as the exotic image it creates caters to popular taste. A best-selling novel, *Belly Drum*, was written to vividly reflect the imaginative history and real experiences of these so-called Eastern Gypsies. Like DNA, some cultural items such as language are normally transmitted from parents to offspring (Pagel, 2009). Therefore, when lacking written records, current genetic and linguistic data can be used to trace the ancestry information of a subject population (Hunley *et al.*, 2008; Balanovsky *et al.*, 2011; Karafet *et al.*, 2016). Here, to address the questions of the ethnic origin and population history of the Xuejiawan people, we, a joint team formed by Chinese geneticists and French linguists, performed an interdisciplinary investigation in Xuejiawan village, focusing on a wide-ranging comparison of paternal lineages and word lists.

2. Materials and Methods

-Subjects.

According to the local chronicles, there are four representative clans of Xuejiawan people, Liú (刘), Liǔ (柳), Gao (高) and Hé (何), accounting for the bulk of the population. The rest, such as Hao (郝) and Guo (郭), are minor clans. Accordingly, we collected oral samples of 118 healthy male individuals from six clans, including 31 individuals with the surname Liú (刘)，39 with Liǔ (柳), 30 with Gao (高), 15 with He (何) , 2 with Hao (郝)，and 1 with Guo (郭), in Xuejiawan village, Yongdeng County, Gansu Province

(Figure 1). This study was approved by the Ethics Committee for Biological Research at Fudan University, and all the samples were collected with informed consent.

-Genetic Data.

Genomic DNA was extracted using DP-318 Kit (Tiangen Biotechnology, Beijing), and Y chromosomes were characterized using two marker systems of different mutability. Initially, we amplified 17 Y-STRs (DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA H4, DYS385a/b) using Y-Filer kit (Life Technologies, CA, USA). Moreover, using a hierarchical genotyping strategy, we first genotyped phylogenetically relevant Y-chromosomal SNPs as listed in the latest Y-chromosomal tree described in previous studies (Wang *et al.*, 2014). The O3a1c*-002611 and O3*-M134+,M117- individuals were then genotyped respectively by two panels successfully used in our previous studies (Wang *et al.*, 2013a; Ning *et al.*, 2016). According to the inferred SNP haplogroups of each sample via our private Y-chromosome database, the C3*-M217 individuals were then subjected to further typing of six biallelic markers, F1067, F2613, F1396, F6733/FGC16362, F3918, F1756, using SNaPshot multiplex kit (ABI, Carlsbad, California, US).

-Linguistic Data.

We carried out an investigation of the Xuejiawan language, covering varied aspects of phonology and morpho-syntax. Because the Xuejiawan language faces extinction, we tried to retrieve as much linguistic information as possible. In particular, about 400 words were collected.

3. Statistical Analysis

Networks of Y chromosomal STR data were constructed by the reduced median-joining method using NETWORK v. 5.0.0.0 (Fluxus-engineering.com). Notably, except for two lineages (O3a1c1-F11 and O3a1c2-F238) which were constructed based on ten Y-STR loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439), the networks for the remaining five lineages employed 15 STRs data (excluding DYS385a and DYS385b). Time estimations for each Y-chromosomal lineage in the Xuejiawan people were made by ASD and BATWING methods as we did before (Wang *et al.*, 2014; Wang and Li, 2015), assuming a generation time of 25 years. Reference population data on the Y chromosomes belonging to the detected haplogroups found in Xuejiawan people were retrieved from the published reports (Kim *et al.*, 2011; Haber *et al.*, 2012; Roewer *et al.*, 2007; Dulik *et al.*, 2011; Abilev *et al.*, 2012; Zhaxylyk *et al.*, 2012; Cai *et al.*, 2011; Wang *et al.*, 2014; Yan *et al.*, 2011; Trejaut *et al.*, 2014; Ning *et al.*, 2016; Wang *et al.*, 2013a), and some unpublished data including Han Chinese, Hui, Yugur, Dongxiang, Baonan, Kazakh, Mongolian and Tibetan were drawn from our lab's database (Wang *et al.*, 2015).

4. Genetic structure of the Xuejiawan people

According to the nomenclature of the Y Chromosome Consortium (Yan *et al.*, 2011; Karafet *et al.*, 2008; Yan *et al.*, 2014), seven paternal lineages were determined from the 118 individual Xuejiawan samples (Figure 2 and Table S1). Haplogroup C3-M217 gave rise to two downstream haplogroups, a southern clade (C3-S) defined by F1067, especially frequent in Han Chinese, and a northern clade (C3-N) defined by F1396, prevalent in Altaic-speaking populations (Yan *et al.*, 2014). C3-S-F2613+,M407-, derived from the southern clade, was quite common in northern Han Chinese, 14.29% and 9.87% in Northeastern and Northern Han, respectively (Zhong *et al.*, 2010; our unpublished data), and was only found with moderate

frequency in the Liú (刘) clan (12.9%). C3-N-F1756 derived from the northern clade, whose STR haplotypes display the null allele at DYS448 (Park *et al.*, 2012), was a brother haplogroup of C3* star-cluster, a famous paternal lineage associated with Genghis Khan (Zerjal *et al.*, 2003). C3-N-F1756 was commonly seen in Kazakhstan (Kazakhs, 11.11%), Hailar in Inner Mongolia (Mongolian, 9.26%) and the Altai Republic in Russia (Altain, 12.50%; Teleuts, 9.09%), but was absent or found with very low frequency in Han Chinese (Zhong *et al.*, 2010; our unpublished data). This subclade was exclusively detected with high frequency (94.87%) in the Liŭ (柳) clan. The basal O2*-M268+,PK4- containing Emperor Cao Cao's paternal lineage (Wang *et al.*, 2012; Wang *et al.*, 2013b), which was abundant in Northern and Eastern parts of China, accounts for about 5% of the Han Chinese (Yan *et al.*, 2011). In this study, it was detected once in the Liŭ (柳) clan, represents about half of the Y chromosomes observed in He (何) clan, and reaches the highest frequency in the Gao (高) clan (100%), showing a common descent from a single paternal ancestor. O3a1c-002611, O3a2c1-M134 and O3a2c1a-M117, three main subclades of O3-M122 were seen as the three super-grandfathers of about 40% of modern Chinese (Yan *et al.*, 2014). According to the east-to-west pattern of phylogeographic distribution, O3a2c1*-M134 and O3a2c1a-M117 exhibit high frequencies in northern Han Chinese and Tibeto-Burman populations, whereas O3a1c-002611 is more frequent in the eastern Han Chinese (Wang *et al.*, 2013a; Wang *et al.*, 2014). In the Xuejiawan people, O3a2c1a-M117 and O3a2c1*-M134 were exclusively detected in the Liú (刘) clan at high (70.97%) and moderate levels (16.13%), respectively. O3a1c1-F11, one subclade of O3a1c-002611, was observed at considerable frequencies in the He (何) clan (47%), and was separately observed once in Liŭ (柳) clan and Guo (郭) clan. O3a1c2-F238, a brother subclade of O3a1c1-F11, was only found in Hao (郝). In summary, the dominant paternal lineages in various clans have a definite East Asian origin, for instance C3-S-F2613+,M407-, O3a2c1*-F2887 and O3a2c1a-M117 in the Liú (刘) clan, C3-N-F1756 in the Liŭ (柳) clan, O2*-M268+,PK4-,M176- in the Gao (高) clan, O2*-M268+,PK4-,M176- and O3a1c1-F11 in the He (何) clan, O3a1c2-F238 in the Hao (郝) clan and O3a1c1-F11 in the Guo (郭) clan.

5. Network analysis and time estimation.

To discern the detailed relationships between seven East Asian Asian lineages in the Xuejiawan people and other related populations, median-joining networks constructed based on Y-STR haplotypes of those haplogroups were shown in Figure 3 and Doc S1. In the network C3-N-F1756, 24 Liŭ (柳) clan individuals rooted in the central zone were shared by a northern Han individual, and were closely related with the remaining Liŭ (柳) clan individuals and the individuals from northwestern populations, especially Kazakhs, Gansu Hui and Shaanxi Han, implying a native northwestern Chinese origin of Liŭ (柳) clan. A clear Xuejiawan-specific cluster comprising one Liŭ (柳) clan, eight He (何) clan and 30 Gao (高) clan individuals can be identified from the O2*-M268+,PK4-,M176- network, which was located on the branch mainly containing Pinghua populations in Southwestern China, suggesting a common descent from a single ancestor and a tie between the O2*-M268+,PK4-,M176- individuals in Xuejiawan people and the Southwestern Chinese populations. In the upper part of the O3a1c1-F11 network, a cluster contained one Liŭ (柳) clan, seven He (何) clan individuals and some Northwestern and Northern Han individuals, and demonstrated a star-like pattern with a central founder haplotype and a few subfounders. Notably, the Guo (郭) clan individuals belonged to a different cluster in the lower part of the network, revealing a different origin. The Liú (刘) clan individuals were found in a small cluster in the upper part of the O3a1c1-F11 network, which mainly comprises Pinghua populations in Southwestern China. Reduced median networks for the remaining lineages, O3a2c1a-M117, C3-S-F2613+,M407-, O3a1c2-F238 and O3a2c1*-F2887 (Doc S1), displayed similar patterns of branching. Accordingly, the 22 Liú (刘) clan individuals and northwestern

Han individuals formed a unique cluster, the four Liú (刘) clan individuals were much closer to northern Han individuals, the two Hao (郝) clan individuals were clustered with northwestern Han and northern Han individuals, and the five Liú (刘) clan individuals showed a close genetic relationship with eastern Han individuals.

Having inferred the putative origins of the clan-based Y chromosomal lineages in the Xuejiawan people at individual level, we then estimated the coalescence time for each paternal lineage using both the ASD and BATWING methods (Table 1 and Table S2). In our previous case studies evaluating Y-STR dating in deep-rooted pedigrees, we found that the Y-chromosomal genealogical mutation rates (OMRB and lmMR) in the BATWING method can give the best-fit estimation for historical lineage dating (Wang and Li, 2015). Hence, in the Xuejiawan people, four paternal lineages, O3a2c1a-M117, C3a1-F2613+,M407-, C3-N-F1756, and O3a1c1-F11, can trace their common ancestor to the Tang and Song dynasties. The coalescence of subhaplogroup O2*-M268+,PK4-,M176- was more likely in the late Yuan and early Ming Dynasties. Because they share a Y chromosomal haplotype, the lineages O3a2c1*-F2887 and O3a1c2-F238 cannot be meaningfully used for Y-STR dating, suggesting that their coalescence times were quite recent.

6. Linguistic affinity of the Xuejiawan language

The Xuejiawan language is phonetically similar to Yongdeng dialect, which belongs to the Jincheng group of Lanyin Mandarin. Additionally, compared to Northwestern Chinese dialects, the Xuejiawan language has some specific phonetic features. Firstly, in addition to its front-back nasal merger like the surrounding Chinese dialects, the Xuejiawan language sometimes leaves out the nasal element, such as pronounting today (今闲) as $tɕi^{11}ɕiɛ^{13}$ (allophones of $tɕĩ^{11}ɕiɛ̃^{13}$) and woman (娘娘) as $ȵia\ ȵia$ (instead of $ȵiã\ ȵiã$). Secondly, plosives and fricatives/affricates, as well as fricatives and affricates, can sometimes be pronounced interchangeably without misunderstanding; for instance, both $tĩ^{11}\ kuɛ^{33}$ and $ɕĩ^{11}\ kuɛ^{13}$ mean 'thing' (东西), $liu^{11}\ thĩ^{31}$ and $liu^{11}\ tɕhĩ^{31}$ mean 'kerchief' (头巾), $phiɛ^{13}\ suə^{31}$ and $phiɛ^{13}\ tsuə^{31}$ mean 'cry/blow' (哭,吹) etc. Syntactically, the Xuejiawan language is grouped with Sinitic languages rather than with minority languages. Some nearby Chinese dialects in Gansu province, such as Hezhou, Tangwang and Xining, have changed their word order from SVO (subject-verb-object) to SOV (subject-object-verb), being affected by surrounding Altaic languages. However, the Xuejiawan language has kept the word order SVO, which is common in all Chinese dialects. Interestingly, the Xuejiawan language has a peculiar pronoun system (e.g. $xə^{11}tɕhĩ^{33}$ (贺秦) means 'I/me', $tɕhĩ^{11}\ tɕia^{33}$ (秦家) designates 'you' and 'he/she'), the numeral system (e.g. $lio^{33}\ tʂə^{11}\ tsɿ$ 'one', $mi^{11}\ tʂə^{33}\ tsɿ$ 'two', etc.) and interrogative system (e.g. WH question words are expressed with the same form $nuo^{11}tu^{33}$ (挪都) or $nə^{11}tuə^{13}$ (呢多)). In a word, despite having its own characteristics, the Xuejiawan language unequivocally belongs to the Sinitic languages from phonetic and syntactic perspectives.

Then we must ask the question, what linguistic elements make the Xuejiawan language unintelligible for the peripheral Han farmers? In other words, how is the Xuejiawan language so different from the neighboring Chinese dialects? To address this issue, a word list containing 406 words from the Xuejiawan language has been dissected and compared with other potentially associated languages, including Altaic, Hmong and Chinese languages. Actually, the Xuejiawan people employed Chinese means of word formation. Among the 406 collected words, 318 words have been derived from gang languages of various historic periods, covering 13 words from the Song and Yuan dynasties, 10 words from the Ming and Qing Dynasties, 48 words from the end of the Qing Dynasty, and 247 words from the contemporary era; 55 words come from Northwestern Chinese dialect, 17 words from Swadesh's 200-word list have been lost by the Xuejiawan people, 6 words come from Altaic languages and the remaining 10 words could not be identified

(Figure 4).

On the one hand, 78.32% commonly used words in the Xuejiawan language are derived from various argots and some argot words can be dated back to the Song dynasty, making it hard for the neighboring Han farmers to understand. On the other hand, the Xuejiawan people have created many new words. This word-building rule is to create new words based on the limited existing vocabulary. For example, sɿ^{33} ma^{11} (司马) means 'vehicle' (车) in the Xuejiawan language; subsequently, they created the new words xũ^{33} tsɿ^{31}sɿ^{33} ma^{11} (fire + vehicle, 红子司马), tɕiɛ̃^{13}sɿ^{33} ma^{31} (little + vehicle, 尖司马) and xe^{33} sɿ^{33} ma^{11} (big + vehicle 海司马), instead of saying 'train' (火车), 'bicycle' (自行车) and 'car' (大汽车), where the first syllables mean fire (火), small (小) and large (大), respectively. In addition, the Xuejiawan people sometimes assembled the existing words into a phrase in order to represent a new thing. For example, they put three words together, jũ^{33} tʂɑ^{31} ('water' 云掌), li (postposition like other Chinese dialects 'in' 里) and fa^{11} la^{33} ('to play' 耍拉), to express the meaning 'swimming'. Therefore, the unique argot words, as well as the created new words based on the extant vocabulary, have caused the public and some scholars to misunderstand the Xuejiawan language as a foreign language or minority language.

7. Discussion

The Roma (Gypsies), which represent a population of 10-15 million living throughout Europe and West Asia, have no nation-state, speak different languages, belong to many religions and comprise a mosaic of socially and culturally divergent endogamous groups (Kalaydjieva *et al.*, 2005). Linguistically, the Gypsies speak more than 60 dialects called *Romani*, which are most closely related to the Northwestern Indian languages like Punjabi or Kashmiri or Central Indian languages like Hindi, suggesting an Indian origin of the Gypsies (Turner, 1984). Previous genetic studies also supported the scenario that the Gypsies originated after their exodus from India about 1,000-1,500 years ago (Moorjani *et al.*, 2013; Mendizabal *et al.*, 2012; Gomez-Carballa *et al.*, 2013). Y-chromosome haplogroup H1a1a-M82 (Kalaydjieva *et al.*, 2001, Pamjav *et al.*, 2011, Rai *et al.*, 2012), mtDNA haplogroup M5a1, M18 and M35b (Mendizabal *et al.*, 2011), and several disease-causing mutations (for instance, the congenital myasthenia 1267delG mutation) (Morar *et al.*, 2004), found on the same ancestral chromosomal background in Gypsy, Indian and Pakistani subjects, has demonstrated a strong genetic link between the proto-Gypsies and Northwestern Indian. Furthermore, it is clear that, in the Romani paternal gene pool, Y-chromosomal lineages were from two different putative origins, ancestral Indian (H1a-M82) and later mixed present-day Eurasian (J2a2-M67, J2*-M172, E1b1b1a-M78, I1-M253, I2a-P37.2, R1a1-M198 and R1b1-P25) during their migration route (Zalan *et al.*, 2011).

In this study, seven East Asian lineages assigned to two major clades C and O were found in the Xuejiawan people, definitively disproving the Gypsy origin hypothesis. Furthermore, due to the evidences from network analysis and lexical comparison, we found that the formation of the Xuejiawan people was a dynamic process and was mainly divided into three stages, albeit with some uncertainty: the mixing of paternal lineages O3a1c1-F11, C3-N-F1756, C3-S-F2613+,M407- and O3a2c1a-M117 in the Song dynasty when they might have primarily spoken Northern varieties; the participation of lineage O2*-M268+,PK4-,M176- in the late Yuan and early Ming Dynasties when some Southern vocabulary might have been introduced, and the immigration of lineages O3a2c1*-F2887 and O3a1c2-F238 in recent times as the Northwestern Chinese dialects have gradually expanded their influence on the Xuejiawan people.

With their nomadic lifestyle and endogamous social practices, the Xuejiawan people have been socially marginalized and historically persecuted in the past, resembling the same historical experiences of the Gypsies in the West. Coincidentally, the Xuejiawan people have chosen a means of living by fortune-

telling, instead of farming. Therefore, the Xuejiawan people could be easily misinterpreted as Eastern Gypsies. In addition, they speak a language derived from the gang languages of various historic periods, emphasizing their mystique and thus meeting the demands of their vocation. There are two similar cases from elsewhere in the world, in which Uisai speakers in Papua New Guinea and Quechua speakers in Peru have changed the linguistic elements to make their language differ from the surrounding languages, due to various social and cultural causes (Thomason, 2001; Thomason, 2003).

8. Supporting Information

Table S1. Y-chromosome SNP and STR data for the Xuejiawan people.

Table S2. The detailed results of coalescence time estimations for each paternal lineage in the Xuejiawan population using both BATWING and ASD methods (time in years).

Doc S1. Median-joining networks of Y-STR haplotypes for the remaining three paternal lineages (C3-S-F2613+,M407-, O3a1c2-F238 and O3a2c1*-F2887).

Acknowledgments

We thank all the volunteers and the local guide Xiaosheng Li for sample collection. This work was partly supported by NSFC for Excellent Young Scholars (31222030), MOE Scientific Research Project (113022A), the Shanghai Shuguang Project (14SG05), the French National Research Agency (No. ANR-12-BSH2-0004-01), the Natural Science Foundation of Gansu province (1308RJZA190), and the Scientific Research Project for Colleges of Gansu Province (2014A-085).

References

Abilev, S., Malyarchuk, B., Derenko, M., Wozniak, M., Grzybowski, T. and Zakharov, I., 2012. The Y-chromosome C3* star-cluster attributed to Genghis Khan's descendants is present at high frequency in the Kerey clan from Kazakhstan. *Hum Biol* 84 (1), 79-89.

Balanovsky, O., Dibirova, K., Dybo, A., Mudrak, O., Frolova, S., Pocheshkhova, E., Haber, M., Platt, D., Schurr, T., Haak, W., Kuznetsova, M., Radzhabov, M., Balaganskaya, O., Romanov, A., Zakharova, T., Soria, H.D., Zalloua, P., Koshel, S., Ruhlen, M., Renfrew, C., Wells, R.S., Tyler-Smith, C. and Balanovska, E., 2011. Parallel evolution of genes and languages in the Caucasus region. *Mol Biol Evol* 28 (10), 2905-2920.

Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., Wei, L., Wang, C., Li, S., Huang, X., Jin, L. and Li, H., 2011. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. *PLoS One* 6 (8), e24282.

Deng, Q. Y., Wang, C. C., Wang, X. Q., Wang, L. X., Wang, Z. Y.,and Wu, W.J., and Li, H., 2013. Genetic affinity between the kam-sui speaking chadong and mulam people. *Journal of Systematics & Evolution* 51(3), 263–270.

Dulik, M.C., Osipova, L.P. and Schurr, T.G., 2011. Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS One* 6 (3), e17548.

Gomez-Carballa, A., Pardo-Seco, J., Fachal, L., Vega, A., Cebey, M., Martinon-Torres, N., Martinon-Torres, F. and Salas, A., 2013. Indian signatures in the westernmost edge of the European Romani diaspora: new insight from mitogenomes *PLoS One* 8 (10), e75397.

Guan, S.X., and Zhang, H., 2012. The research and analysis into the nationality - belonging of "Xuejiawan Fortune-teller". *Journal of Gansu Lianhe University: social Sciences* 28(1), 81-85.

Haber, M., Platt, D.E., Ashrafian, B.M., Youhanna, S.C., Soria-Hernanz, D.F., Martinez-Cruz, B., Douaihy, B., Ghassibe-Sabbagh, M., Rafatpanah, H., Ghanbari, M., Whale, J., Balanovsky, O., Wells, R.S., Comas, D., Tyler-Smith, C. and Zalloua, P.A., 2012. Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. *PLoS One* 7 (3), e34288.

Hunley, K., Dunn, M., Lindstrom, E., Reesink, G., Terrill, A., Healy, M.E., Koki, G., Friedlaender, F.R. and Friedlaender, J.S., 2008. Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genet* 4 (10), e1000239.

Kalaydjieva, L., Calafell, F., Jobling, M.A., Angelicheva, D., de Knijff, P., Rosser, Z.H., Hurles, M.E., Underhill, P., Tournev, I., Marushiakova, E. and Popov, V., 2001. Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. *Eur J Hum Genet* 9 (2), 97-104.

Kalaydjieva, L., Morar, B., Chaix, R. and Tang, H., 2005. A newly discovered founder population: the Roma/Gypsies. *Bioessays* 27 (10), 1084-1094.

Karafet, T.M., Bulayeva, K.B., Nichols, J., Bulayev, O.A., Gurgenova, F., Omarova, J., Yepiskoposyan, L., Savina, O.V., Rodrigue, B.H. and Hammer, M.F., 2016. Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan. *J Hum Genet* 61 (3), 181-191.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L. and Hammer, M.F., 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 18 (5), 830-838.

Kim, S.H., Kim, K.C., Shin, D.J., Jin, H.J., Kwak, K.D., Han, M.S., Song, J.M., Kim, W. and Kim, W., 2011. High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. *Investig Genet* 2 (1), 10.

Lu, Y., Pan, S. L., Qin, S. M., Qin, Z. D., Wang, C. C., Gan, R. J., and Li, H., 2013. Genetic evidence for the multiple origins of pinghua chinese. *Journal of Systematics & Evolution*, 51(3), 271-279.

Mendizabal, I., Lao, O., Marigorta, U.M., Wollstein, A., Gusmao, L., Ferak, V., Ioana, M., Jordanova, A., Kaneva, R., Kouvatsi, A., Kucinskas, V., Makukh, H., Metspalu, A., Netea, M.G., de Pablo, R., Pamjav, H., Radojkovic, D., Rolleston, S.J., Sertic, J., Macek, M.J., Comas, D. and Kayser, M., 2012. Reconstructing the population history of European Romani from genome-wide data. *Curr Biol* 22 (24), 2342-2349.

Mendizabal, I., Valente, C., Gusmao, A., Alves, C., Gomes, V., Goios, A., Parson, W., Calafell, F., Alvarez, L., Amorim, A., Gusmao, L., Comas, D. and Prata, M.J., 2011. Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. PLoS One 6 (1), e15988.

Moorjani, P., Patterson, N., Loh, P.R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kadasi, L., Riess, O., Berger, B., Reich, D. and Melegh, B., 2013. Reconstructing Roma history from genome-wide data. *PLoS One* 8 (3), e58633.

Morar, B., Gresham, D., Angelicheva, D., Tournev, I., Gooding, R., Guergueltcheva, V., Schmidt, C., Abicht, A., Lochmuller, H., Tordai, A., Kalmar, L., Nagy, M., Karcagi, V., Jeanpierre, M., Herczegfalvi, A., Beeson, D., Venkataraman, V., Warwick, C.K., Reeve, J., de Pablo, R., Kucinskas, V. and Kalaydjieva, L., 2004. Mutation history of the roma/gypsies. *Am J Hum Genet* 75 (4), 596-609.

Ning, C., Yan, S., Hu, K., Cui, Y.Q. and Jin, L., 2016. Refined phylogenetic structure of an abundant East Asian Y-chromosomal haplogroup O*-M134. *Eur J Hum Genet* 24 (2), 307-309.

Pagel, M., 2009. Human language as a culturally transmitted replicator. Nat Rev Genet 10 (6), 405-415.

Pamjav, H., Zalan, A., Beres, J., Nagy, M. and Chang, Y.M., 2011. Genetic structure of the paternal lineage

of the Roma people. *Am J Phys Anthropol* 145 (1), 21-29.

Park, M.J., Lee, H.Y., Yang, W.I. and Shin, K.J., 2012. Understanding the Y chromosome variation in Korea-relevance of combined haplogroup and haplotype analyses. *Int J Legal Med* 126 (4), 589-599.

Rai, N., Chaubey, G., Tamang, R., Pathak, A.K., Singh, V.K., Karmin, M., Singh, M., Rani, D.S., Anugula, S., Yadav, B.K., Singh, A., Srinivasagan, R., Yadav, A., Kashyap, M., Narvariya, S., Reddy, A.G., van Driem, G., Underhill, P.A., Villems, R., Kivisild, T., Singh, L. and Thangaraj, K., 2012. The phylogeography of Y-chromosome haplogroup h1a1a-m82 reveals the likely Indian origin of the European Romani populations. *PLoS One* 7 (11), e48477.

Roewer, L., Kruger, C., Willuweit, S., Nagy, M., Rodig, H., Kokshunova, L., Rothamel, T., Kravchenko, S., Jobling, M.A., Stoneking, M. and Nasidze, I., 2007. Y-chromosomal STR haplotypes in Kalmyk population samples. *Forensic Sci Int* 173 (2-3), 204-209.

Thomason, S., 2001. Language contact: An introduction. Edinburgh: Edinburgh University Press.

Thomason, S., 2003. Contact as a source of language change. In Joseph, Brian D.& Richard D. Janda, (eds.), *The handbook of Historical Linguistics*. Blackwell Publishing.

Trejaut, J.A., Poloni, E.S., Yen, J.C., Lai, Y.H., Loo, J.H., Lee, C.L., He, C.L. and Lin, M., 2014. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. *BMC Genet* 15, 77.

Tuo, A., 2006. The vestige and cultural phenomenon of minority nationality languages in Yongdeng Dialect. *Journal of Gansu Lianhe University: social Sciences* 22(6), 92-95.

Turner, R.L., 1984. The position of Romani in Indo-Aryan. *Journal of the Gypsy Lore Society* 3: 145–94.

Wang, C., Yan, S., Hou, Z., Fu, W., Xiong, M., Han, S., Jin, L. and Li, H., 2012. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. *J Hum Genet* 57 (3), 216-218.

Wang, C.C. and Li, H., 2015. Evaluating the Y chromosomal STR dating in deep-rooting pedigrees. *Investig Genet* 6, 8.

Wang, C.C., Wang, L.X., Shrestha, R., Wen, S., Zhang, M., Tong, X., Jin, L. and Li, H., 2015. Convergence of Y Chromosome STR Haplotypes from Different SNP Haplogroups Compromises Accuracy of Haplogroup Prediction. *J Genet Genomics* 42 (7), 403-407.

Wang, C.C., Wang, L.X., Shrestha, R., Zhang, M., Huang, X.Y., Hu, K., Jin, L. and Li, H., 2014. Genetic structure of Qiangic populations residing in the western Sichuan corridor. *PLoS One* 9 (8), e103772.

Wang, C.C., Yan, S., Qin, Z.D., Lu, Y., Ding, Q.L., Wei, L.H., Li, S.L., Yang, Y.J., Jin, L., Li, H., the Genographic Consortium, 2013a. Late Neolithic expansion of ancient Chinese revealed by Y chromosome haplogroup O3a1c-002611. *Journal of Systematics and Evolution* 51(3), 280–286.

Wang, C.C., Yan, S., Yao, C., Huang, X.Y., Ao, X., Wang, Z., Han, S., Jin, L. and Li, H., 2013b. Ancient DNA of Emperor CAO Cao's granduncle matches those of his present descendants: a commentary on present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. *J Hum Genet* 58 (4), 238-239.

Wu, J.S., 1990. Analyzing the origin of Xuejiawan people in Yongdeng County, Gansu Province. *Journal of the Central University for Nationalities* (1), 48-53.

Wu, J.S., 1991. A review of the linguistic affiliation of the Xuejiawan language. *Journal of Northwest University for Nationalities: social Sciences* (2), 111-116.

Wu, R.Z., Wu, S.G., Zhao, Y.Y., 2001. Preliminary Research into the Ethnic Source of "Xuejiawan Residents" in Yongdeng County, Gansu Province. *Journal of South-Central University for Nationalities: Humanities and Social Science* 21(3), 39-43.

Yan, S., Wang, C.C., Li, H., Li, S.L. and Jin, L., 2011. An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *Eur J Hum Genet* 19 (9), 1013-1015.

Yan, S., Wang, C.C., Zheng, H.X., Wang, W., Qin, Z.D., Wei, L.H., Wang, Y., Pan, X.D., Fu, W.Q., He, Y.G., Xiong, L.J., Jin, W.F., Li, S.L., An, Y., Li, H. and Jin, L., 2014. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. *PLoS One* 9 (8), e105691.

Yang, Z.J., 1991. Luri Huihui - the Gypsies people in Yuan Dynasty. Historical Research (3):40-47.

Zalan, A., Beres, J. and Pamjav, H., 2011. Paternal genetic history of the Vlax Roma. *Forensic Sci Int Genet* 5 (2), 109-113.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du R, Yang, H., Hurles, M.E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S.Q. and Tyler-Smith, C., 2003. The genetic legacy of the Mongols. *Am J Hum Genet* 72 (3), 717-721.

Zhaxylyk, S., Turuspekov, Y., Daulet, B., Sadykov, M., and Khalidullin, O., 2012. The kazakhstan DNA project hits first hundred y-profiles for ethnic kazakhs. *Russian Journal of Genetic Genealogy*, 2(1), 1920-2989.

Zhong, H., Shi, H., Qi, X.B., Xiao, C.J., Jin, L., Ma, R.Z. and Su, B., 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *J Hum Genet* 55 (7), 428-435.

Zhu, T., 1947. The ‘eastern Gypsies’ - a record of the old barbarian women in Yongdeng County, Gansu Province. Communication on Borderland (4)

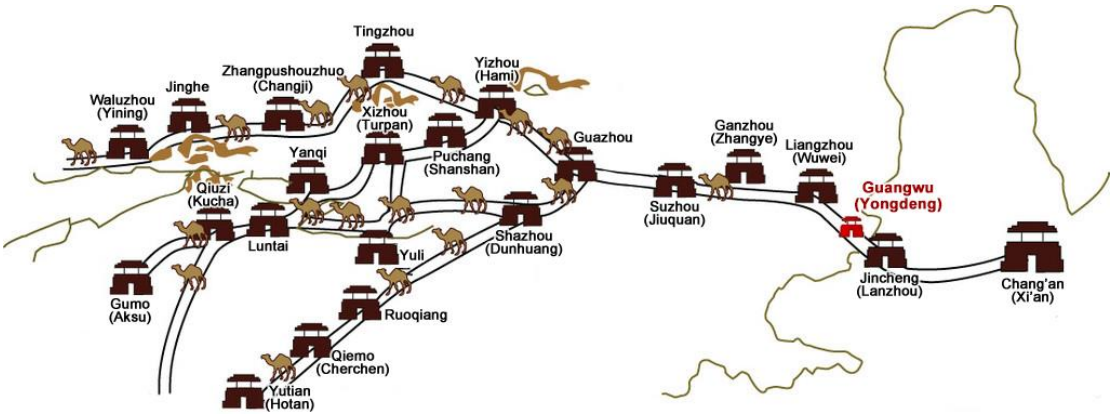
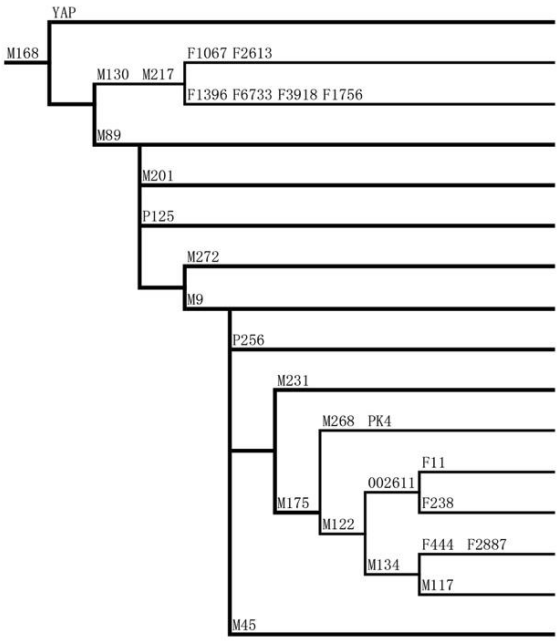


Figure 1. The Xuejiawan village is located in Yongdeng County, which was once called Guangwu County in Tang Dynasty, an important town on the ancient Silk Road.



Clan	Liu 刘	Liu 柳	Gao 高	He 何	Hao 郝	Guo 郭
Sample size	31	39	30	15	2	1
(DE)	–	–	–	–	–	–
C3-S-F2613+, M407-	12.90	–	–	–	–	–
C3-N-F1756	–	94.87	–	–	–	–
(F)	–	–	–	–	–	–
(G)	–	–	–	–	–	–
(IJ)	–	–	–	–	–	–
(T)	–	–	–	–	–	–
(K)	–	–	–	–	–	–
(M)	–	–	–	–	–	–
(N)	–	–	–	–	–	–
02*-M268+, PK4-, M176-	–	2.56	100.00	53.00	–	–
03a1c1-F11	–	2.56	–	47.00	–	100.00
03a1c2-F238	–	–	–	–	100.00	–
03a2c1*-F2887	16.13	–	–	–	–	–
03a2c1a-M117	70.97	–	–	–	–	–
(P)	–	–	–	–	–	–

Figure 2. The phylogenetic relationship of Y-chromosome haplogroups surveyed in this study and their clan-based frequencies among the Xuejiawan people. The marker names are shown along the branches, and haplogroup names are shown on the right side. Potentially paraphyletic undefined subgroups are distinguished from recognized haplogroups by the asterisk symbol. Haplogroups tested for but not seen in this study are enclosed in parentheses.

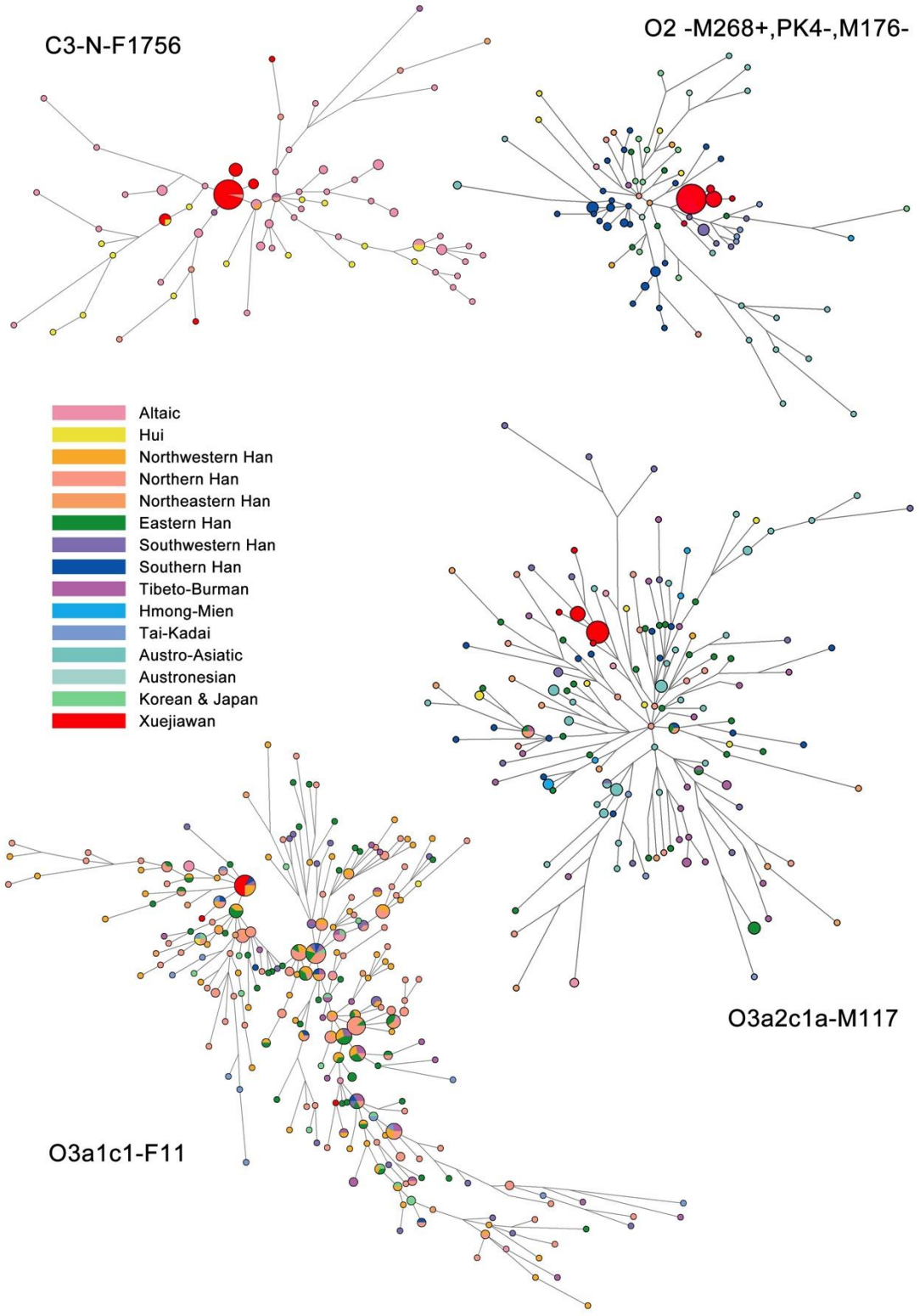


Figure 3. Median-joining networks of Y-STR haplotypes for the four dominant haplogroup lineages (C3-N-F1756, O2*-M268+,PK4-,M176-, O3a1c1-F11 and O3a2c1a-M117) in the Xuejiawan people.

Haplotypes are represented by circles with area proportional to the number of individuals. Colors indicate geographic origin. Notably, northwestern Han refers to the ethnic Han individuals originating from the provinces of Shaanxi, Gansu and Xinjiang; northern Han refers to Hebei, Henan, Shandong, Shanxi, Tianjin and Beijing; northeastern Han refers to Jilin, Liaoning, Heilongjiang and Inner Mongolia; eastern Han refers to Jiangsu, Jiangxi, Zhejiang, Anhui and Shanghai; southern Han refers to Guangdong, Hainan, Hunan, Hubei and Fujian; southwestern Han refers to Yunnan, Chongqing, Guangxi, Guizhou and Sichuan.

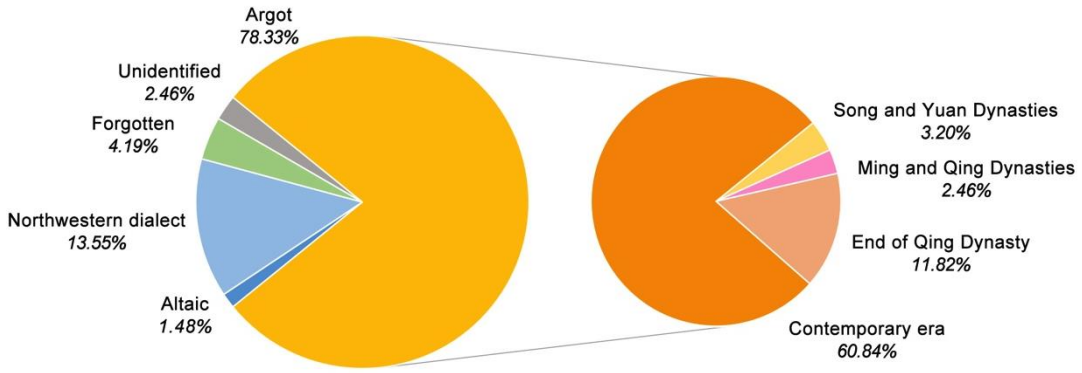


Figure 4. The lexical composition of the Xuejiawan's 406 word-list.

Table 1. The partial results of coalescence time estimations for each paternal lineage in the Xuejiawan people using BATWING (time in years)

Haplogroup	Clan	Putative origins	lmMR		OMRB		Period
			TMRCA	95% CI	TMRCA	95% CI	
O2*-M268+,PK4-,M176-	Gao	southwest	504.4	59.4-3455.9	474.4	55.7-3265.4	Ming Dynasty
	He	southwest	559	33.5-6286.8	534.7	32.1-5998.1	Ming Dynasty
O3a2c1a-M117	Liú	northwest	977.6	157.5-5764.4	945.2	151.5-5641.1	Song Dynasty
C3-S-F2613+,M407-	Liú	north	1093.4	48.7-17359.2	1066	46.3-17172.1	Song Dynasty
O3a2c1*-F2887	Liú	east	-	-	-	-	Recent times
C3-N-F1756	Liǔ	northwest	1181	208.3-6520	1140.9	199.9-6356.7	Five Dynasties
O3a1c1-F11	He	Northwest/North	1332.9	133.2-11387.7	1283.8	127.9-10987.8	Tang Dynasty
	Guo	Northwest	-	-	-	-	Recent times
O3a1c2-F238	Hao	Northwest/North	-	-	-	-	Recent times