

# SCIENTIFIC REPORTS



OPEN

## The Connection of the Genetic, Cultural and Geographic Landscapes of Transoxiana

Maxat Zhabagin<sup>1,2</sup>, Elena Balanovska<sup>3</sup>, Zhaxylyk Sabitov<sup>4</sup>, Marina Kuznetsova<sup>3</sup>, Anastasiya Agdzhoyan<sup>2,3</sup>, Olga Balaganskaya<sup>2</sup>, Marina Chukhryaeva<sup>2,3</sup>, Nadezhda Markina<sup>2</sup>, Alexey Romanov<sup>3</sup>, Roza Skhalyakho<sup>2,3</sup>, Valery Zaporozhchenko<sup>2,3</sup>, Liudmila Saroyants<sup>5</sup>, Dilbar Dalimova<sup>6</sup>, Damir Davletchurin<sup>6</sup>, Shahlo Turdikulova<sup>7</sup>, Yuldash Yusupov<sup>8</sup>, Inkar Tazhigulova<sup>9</sup>, Ainur Akilzhanova<sup>1</sup>, Chris Tyler-Smith<sup>10</sup> & Oleg Balanovsky<sup>3</sup>

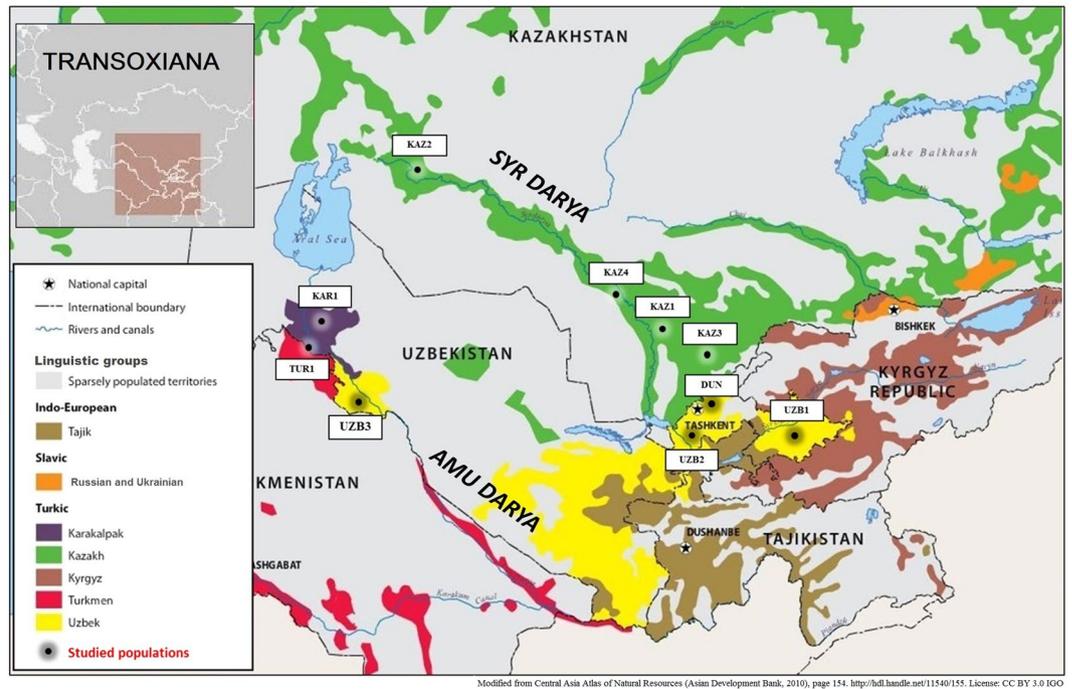
We have analyzed Y-chromosomal variation in populations from Transoxiana, a historical region covering the southwestern part of Central Asia. We studied 780 samples from 10 regional populations of Kazakhs, Uzbeks, Turkmens, Dungans, and Karakalpaks using 35 SNP and 17 STR markers. Analysis of haplogroup frequencies using multidimensional scaling and principal component plots, supported by an analysis of molecular variance, showed that the geographic landscape of Transoxiana, despite its distinctiveness and diversity (deserts, fertile river basins, foothills and plains) had no strong influence on the genetic landscape. The main factor structuring the gene pool was the mode of subsistence: settled agriculture or nomadic pastoralism. Investigation of STR-based clusters of haplotypes and their ages revealed that cultural and demic expansions of Transoxiana were not closely connected with each other. The Arab cultural expansion introduced Islam to the region but did not leave a significant mark on the pool of paternal lineages. The Mongol expansion, in contrast, had enormous demic success, but did not impact cultural elements like language and religion. The genealogy of Muslim missionaries within the settled agricultural communities of Transoxiana was based on spiritual succession passed from teacher to disciple. However, among Transoxianan nomads, spiritual and biological succession became merged.

Transoxiana is a historical region of Central Asia (Fig. 1). It covers the territories of five modern countries: Uzbekistan, western Tajikistan, western Kyrgyzstan, northwestern Turkmenistan and southern Kazakhstan. The peculiar features of its geographical landscape and abrupt shifts of cultural landscapes in the course of the history of the region allow us to use it as a model for investigating the connection between genetic, cultural and geographic landscapes.

*Geographic landscape of Transoxiana.* This includes a desert located between the Amu Darya and Syr Darya river basins. The riversides are densely populated by Kazakhs, Uzbeks, Karakalpaks, Kyrgyz and Turkmens (Fig. 1). The Tian Shan mountains are located in the southeast of Transoxiana, while desert plains expand to the northwest and are bordered by the Aral Sea. As geographical landscapes and barriers tend to shape the gene pool<sup>1-3</sup>, it is important to assess their role in the geographically heterogeneous Transoxiana.

*Cultural landscape of Transoxiana.* Two distinct modes of subsistence with contrasting traditional cultures - settled agriculture and nomadic pastoralism - have been practiced in the region for thousands of years. Located on the Silk Road, Transoxiana's history has been influenced by both the West and the East. The first known impact came from Western Asia. During the rule of the Achaemenid Empire (6th century BC) the region was at the

<sup>1</sup>National Laboratory Astana, Nazarbayev University, Astana, Republic of Kazakhstan. <sup>2</sup>Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow, Russia. <sup>3</sup>Research Centre for Medical Genetics, Moscow, Russia. <sup>4</sup>L.N.Gumilyov Eurasian National University, Astana, Republic of Kazakhstan. <sup>5</sup>Leprosy Research Institute, Astrakhan, Russia. <sup>6</sup>Institute of Bioorganic Chemistry, Tashkent, Uzbekistan. <sup>7</sup>Center of High Technologies, Tashkent, Uzbekistan. <sup>8</sup>Institute of Strategic Research of the Republic of Bashkortostan, Ufa, Russia. <sup>9</sup>Forensic science centre of the Ministry of Justice of the Republic of Kazakhstan, Astana, Republic of Kazakhstan. <sup>10</sup>The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom. Maxat Zhabagin and Elena Balanovska contributed equally to this work. Correspondence and requests for materials should be addressed to M.Z. (email: [mzhabagin@gmail.com](mailto:mzhabagin@gmail.com)) or O.B. (email: [balanovsky@inbox.ru](mailto:balanovsky@inbox.ru))



**Figure 1.** Map of Transoxiana and the populations studied. This figure is a derivative of Central Asia atlas of natural resources<sup>76</sup> (<http://hdl.handle.net/11540/155>) by Asian Development Bank, used under CC BY 3.0 IGO. Areas with low population density (deserts and high mountains) are shown in grey.

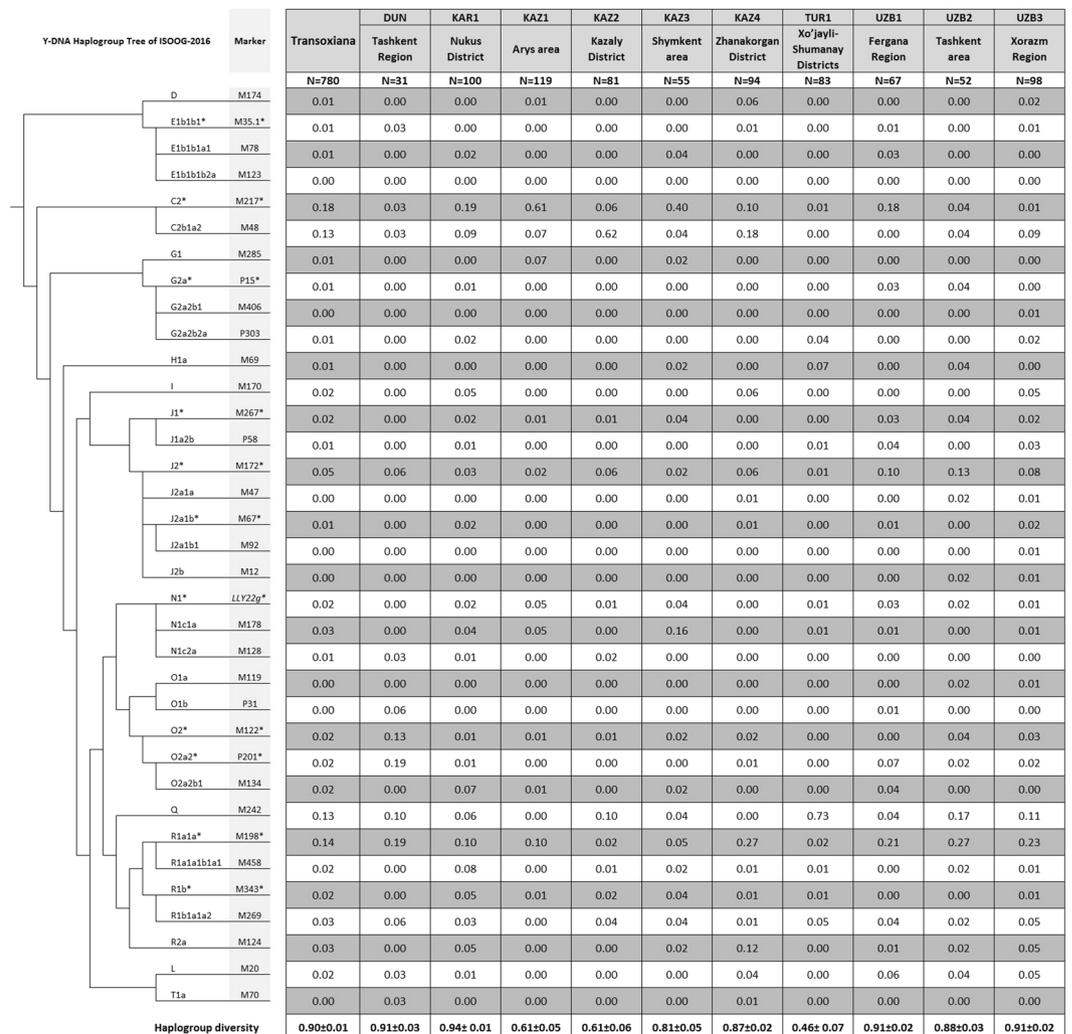
center of the Sogdian culture. In the 4th century BC Alexander the Great turned it into a Hellenistic province, naming it Transoxiana (“area beyond the Ox river”; Ox is the ancient name for Amu Darya). Later the region was part of the Seleucid Empire (4th century BC), the Greco-Bactrian Kingdom (3rd - 2nd centuries BC), the Kushan Empire (1st century BC – 5th century AD), the Hephthalite Empire (5th century AD), and the Sasanian Empire (6th century AD). In the 5th century AD the direction of migration changed to Central Asian sources: with the expansion of Turkic tribes, the region becomes part of the Turkic Khaganate (7th century AD) and succeeding realms. The source of migration returned to Western Asia in the 8th century AD: due to the expansion of Islam, the region fell under the influence of the Arab culture and became known as Mawarannahr. In the 13th century AD the migration vector switched again to a Central Asian source: most of the region was under Mongol cultural influences, becoming the Chagatai Khanate - part of the Mongol Empire<sup>4-6</sup>. Besides these major cultural influences, minor ones such as a small settlement of Han Chinese mentioned by the Buddhist monk Xuanzang in 630 BC took place<sup>4</sup>.

Thus the major landmarks in Transoxiana’s history were: the development of settled agriculture and nomadic pastoralism, the influence of Western Asian Empires, Turkic nomads, Arabs and Mongols. The expansion of Islam by Arabs and preservation of both types of economy (settled agriculture and nomadic pastoralism) had the biggest impact on the cultural landscape. Though cultural expansion itself is of course nothing more than changes of cultural elements (language, religion, technology), it is often associated with demic expansion (human migration) that has direct influence on the gene pool. A number of studies have exemplified influences of religious<sup>7,8</sup> or linguistic shifts<sup>9-11</sup> on genetic structure.

*Genetic landscape of Transoxiana.* Previous genetic studies in Central Asia have used autosomal<sup>12,13</sup>, mtDNA<sup>14-16</sup> or Y-chromosomal markers<sup>17-24</sup>. All these studies stressed the pronounced heterogeneity of the genetic landscape in Central Asia. Most recent studies were dedicated to Y-chromosomal variation, in particular regions or clans within Central Asia<sup>25-33</sup>.

In line with these studies, we also concentrate here on a single region within Central Asia. The Y-chromosomal variation in this Transoxiana region has been insufficiently studied so far. 133 samples from Kyrgyz and Kazakh populations were studied using 8 Y-STRs<sup>17</sup>. 150 samples from Kazakh, Kyrgyz, Uzbeks, Turkmen, and Tajiks were studied using 16 STRs and 16 SNPs<sup>20</sup>. Subsequent studies<sup>22-24</sup> were updated recently<sup>30</sup>, presenting variation of 8 Y-STRs and 31 SNPs in 461 samples. Here, we double this number by genotyping 17-STRs and 35 SNPs in 780 samples, including four regional Kazakh populations along the Syrdarya river, three regional Uzbek populations, Karakalpaks, Turkmen, and Dungans.

The key element of Transoxiana populations that influences both genetic and cultural landscapes is the tribal-clan structure prevalent among nomadic populations. Many modern Kazakh and Turkmen individuals identify themselves not only with an ethnic group, but with a clan as well. Genealogical lineages are combined into clans, and clans are combined into tribes. The tribal-clan structure is based on the traditional belief in the unified bloodline of an ethnic group and has long regulated the ethnosocial order of nomadic populations. This tradition has endured through all the cultural expansions and has incorporated the new elements into the



**Figure 2.** Frequencies of Y-chromosomal haplogroups in Transoxiana populations.

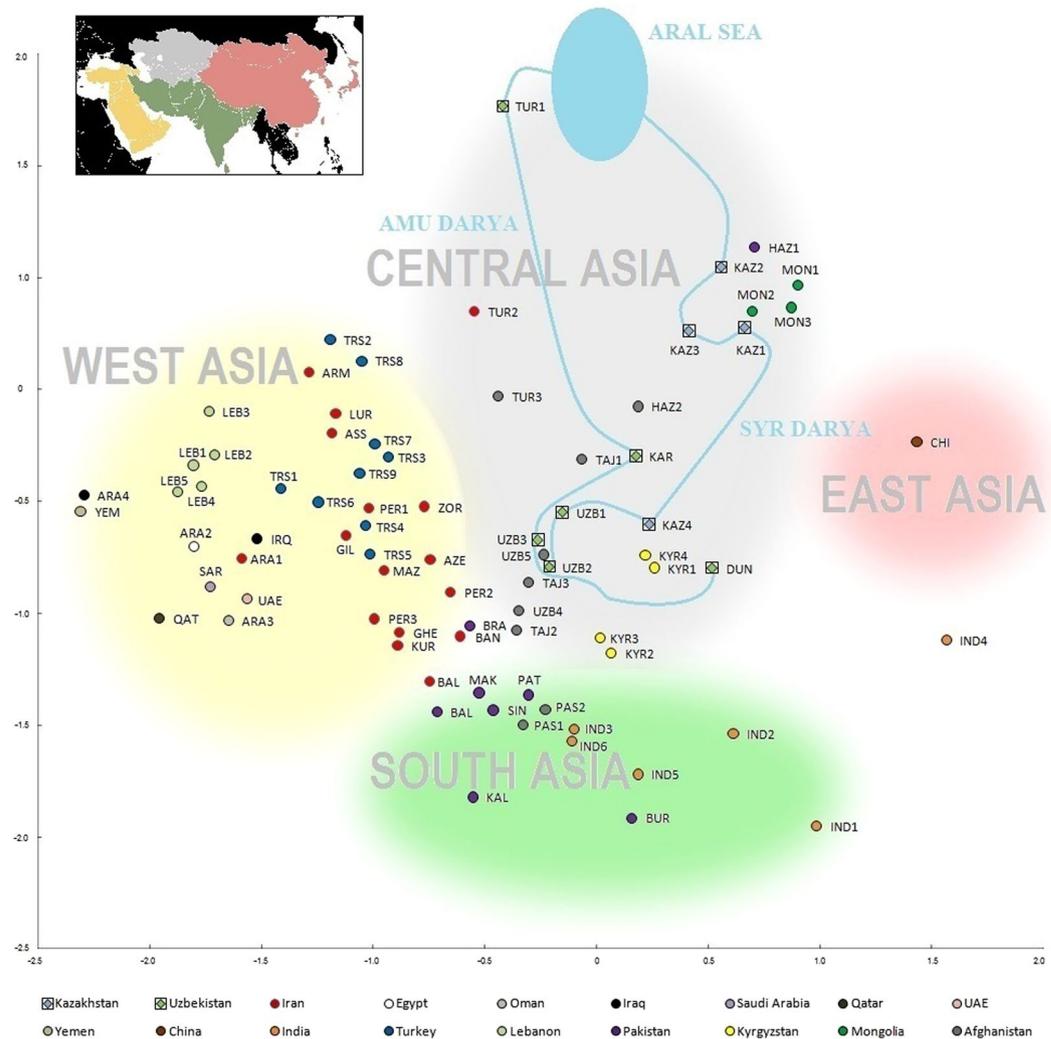
tribal-clan structure. Often the name of the lineage, clan and tribe are inherited through the male line just as the Y-chromosome is. Therefore, it is important to study them together<sup>22, 33</sup>. The areas studied here are predominantly populated by the Kazakh tribe Konyrat, the Kazakh clans Alimuly, Kozha (Khoja) and Sunak, as well as the Turkmen tribe Yomut. A brief summary of the genealogy and history of these clans is given in the Supplementary Text.

This study aims to examine the genetic landscape of Transoxiana and explore its connection to geographical and cultural landscapes. To achieve this aim, we examined Y-chromosomal variation in Kazakhs, Uzbeks, Karakalpaks, Turkmens and Dungans with reference to their cultural and geographical landscapes. The results were used to determine whether or not the two last major expansions (Arab and Mongol) were demic and influenced the Y-chromosomal variation in Transoxiana.

## Results and Discussion

**Genetic portraits of populations and tribal/clan groups of Transoxiana.** 780 samples from 10 populations - Kazakhs (4 regional populations, coded from KAZ1 to KAZ4, see Fig. 2 and Supplementary Table 2), Uzbeks (3 regional populations, UZB1 to UZB3), Karakalpaks (KAR1), Turkmens (TUR1) and Dungans (DUN) - were studied using 35 binary and 17 STR Y-chromosomal markers (Figs 1 and 2, Supplementary Table 1). Data on 5,218 samples from 69 populations, representing 16 Asian countries ranging from Turkey to China (Supplementary Table 2), were used to provide a genetic context. More than half of the Transoxianan Y-chromosomes (58%) falls into three haplogroups (Fig. 2): C2-M217 (31%), R1a1a-M198 (16%), and Q-M242 (13%), although the distribution of these haplogroups in Transoxiana is uneven. C2-M217 comprises almost two-thirds of the south Kazakh gene pool (61% C2\*-M217(xM48) among KAZ1, 62% C2b1a2-M48 among KAZ2). R1a1a-M198 is common among Uzbek (UZB1, UZB2, UZB3, 21-29%), in one Kazakh population (KAZ4, 28%) and among Dungans (DUN, 19%). Q-M242 is found in more than two-thirds of the Turkmen sample (TUR1, 73%).

The prevalence of specific haplogroups is even more pronounced for tribal-clan groups than for geographic populations (Supplementary Fig. 1): C2\*-M217(xM48) comprises 88% of the Y-chromosomes of the Konyrat



**Figure 3.** Genetic relationships of Asian (including Transoxiana) populations using 30 Y-SNPs. Multidimensional scaling plot; stress = 0.17. Populations from 18 countries are marked by colors. The ten populations from this study are shown as rhombuses within squares, while populations from the literature are indicated by circles. Blue lines link populations located along the Amu Darya and Syr Darya rivers. Population codes are explained more fully in Supplementary Table 2. Colored cloud areas represent geographic clusters, with colors on the main plot following colors on the inset (Asian regions according to UN classification).

tribe, C2b1a2-M48 reaches 75% in the Kazakh clan Alimuly, and Q-M242 accounts for 71% in the Turkmen tribe Yomut. Based on haplogroup frequency, the Konyrat tribe is the most homogenous ( $HD = 0.23$ ), while the Kozha-Sunak clan group is the most heterogeneous ( $HD = 0.94$ ). The specificities of the clan pools of paternal lineages are the reason for the specificities of the geographic populations: the clan Alimuly prevails in the KAZ2 population (79% samples are from this clan), the tribe Konyrat predominates in the KAZ1 population (62%), and the tribe Yomut predominates in the TUR1 population (88%).

**The Transoxianan paternal heritage in the Asian context.** The 10 populations from Transoxiana were analyzed along with 69 other Asian populations typed by the same panel of 30 Y-chromosomal SNPs (Supplementary Tables 2 and 3). Clusters corresponding to geographic parts of Asia were revealed in the multidimensional scaling plot (Fig. 3). The Western Asian cluster was represented by Arab, Turkish and Iranian populations. Populations of India, Pakistan and Afghanistan made up Southern Asian cluster. Chinese form the Eastern Asia cluster. All Transoxianan populations lie in the Central Asian cluster.

Analysis on a narrower geographic scale (Transoxiana and the neighboring regions) is available in Supplementary Fig. 3 (Supplementary Table 4). This PCA plot is based on a smaller number of haplogroups, but includes more Central Asian populations. Both an MDS plot of 30 haplogroups and a PC plot of 19 haplogroups (Fig. 3, Supplementary Fig. 2, Supplementary Fig. 3) demonstrate the four following patterns.

First, Uzbek and Tajik populations practicing settled agriculture, as well as Kyrgyz, are genetically distant from most nomadic populations (Mongol, Kazakh, Hazaras). Second, despite originating from three countries (Uzbekistan, Iran, Afghanistan), Turkmen populations form their own firmly separated cluster. The reason lies

Basis of Division	Groups	Percentage of Variation		
		Within populations	Among populations within groups	Among groups
Geography 1	Amu-Darya Syr-Darya	84.18*	16.04*	-0.22 (p-value = 0.39)
Geography 2	Foothills Plain	82.43*	16.52*	1.05 (p-value = 0.27)
Mode of Subsistence	Settled agriculture Nomadic pastoralism	82.82*	14.33*	2.85 (p-value = 0.096)

**Table 1.** Variation in Y-Chromosomal haplogroup frequencies between groups. \*p-value < 0.01.

in high frequency of haplogroup Q-M242 (Supplementary Table 2) in most Turkmen populations, which in particular forms the third PC (Supplementary Fig. 2), though this haplogroup is absent from the fourth Turkmen population<sup>30</sup> despite their sample coming from the same region of Uzbekistan as our sample. Third, Dungan populations of Uzbekistan (DUN) are genetically closer to the populations of China (genetic distance  $d = 0.178$ ) and northeastern India ( $d = 0.152$ ) than to their neighboring Uzbek populations ( $d = 0.228$ ;  $d = 0.410$ ;  $d = 0.425$ ). This is explained by the historically recent migration of Dungsans from China and the maintenance of their Sino-Tibetan language, prevalent in China and northeastern India. Fourth, most of the Kazakh populations studied cluster with Mongols, Pakistani Hazaras (HAZ1) and Afghan Hazaras (HAZ2) due to their high frequency of haplogroup C2-M217. This correlates with the historically well-known Mongol origin of the Hazaras<sup>34,35</sup>. In addition, Fig. 3 shows the populations located along a stretch of the Amu Darya and Syr Darya rivers linked by blue lines symbolizing the rivers. However, the positions of these genetic “rivers” only loosely correlate with their geographical prototypes.

**The lack of relationship between genetics and geography.** To determine the driving forces that shaped the Y-chromosomal variation in Transoxiana, we examined patterns of genetic variation by AMOVA (Supplementary Table 5). Populations were arranged into groups in three ways: (a) Geography – river basins: populations from the Amu Darya or Syr Darya basins; (b) Geography - altitude: plain or foothill populations (400 meters altitude was used as a threshold); (c) Mode of subsistence: settled agriculturalists or nomadic pastoralists.

Both ways of geographic grouping had little to no influence on the genetic structure (Table 1). A Mantel test ( $Yr = -0.006$ ,  $p = 0.44$ ) further supports the idea that, unlike most other regions, in Transoxiana genetic distances between populations do not correlate with the geographic distances. But the mode of subsistence had a significant impact on explaining the genetic structure: the differences between settled and nomadic populations accounts for 2.85% of the total genetic variation, which is almost three times larger than the differences between the geographic groups of populations (Table 1).

From the 5<sup>th</sup> to 2<sup>nd</sup> millennia BC a complete transition to a cattle-raising and agricultural tribal existence occurred in Transoxiana populations<sup>36</sup>. Since that time, the mode of subsistence - settled agriculture or nomadic pastoralism - was the main cultural distinction within Central Asia. This lets us conclude that the influence of geography on the genetic structure was mediated by a combination of subsistence and traditional culture. One may suppose that such relationships of cultural and geographical factors have persisted for thousands of years. It underlines the important role which technical innovations and culture often play in shaping the genetic landscape<sup>37</sup>.

**Arab and Mongol expansions: migration of cultures or populations?** In order to search for signs of male demic expansions, we identified four modal STR-haplotypes of Transoxiana (those present in more than 10 samples in our dataset, Table 2). For each modal haplotype we then identified related haplotypes. We considered haplotypes which were fewer than 5 mutational steps from the modal haplotype and belong to the same haplogroup. Five mutations - considering 15 Y-STRs and mutation rate 0.0021 per locus per generation - might occur within roughly two thousand years, which covers the time interval important for our analysis. The search for related haplotypes was performed in a database of 4495 Y-STR Asian haplotypes using the Haplomatch software<sup>38</sup>. This methodology is similar to that applied by Balaesque and colleagues<sup>30</sup> in their search for Asian primary descent clusters.

The modal haplotype 1 (Table 2) and 257 related haplotypes belonging to haplogroup C2b1a2-M48 were used to construct a phylogenetic network (Supplementary Fig. 4A). Two clusters can be distinguished: cluster  $\alpha$  (which includes the modal haplotype) and cluster  $\beta$ . Cluster  $\alpha$  is  $600 \pm 200$  years old and its modal haplotype is most prominent among the Kazakh clan Alimuly (33%). Cluster  $\beta$  is mostly present among Mongols and Mongolian-speaking Kalmyks. Cluster  $\beta$  is older ( $800 \pm 200$  years using the rho estimate and 660 years using ASD, Table 2), suggesting the gene flow took place from Mongolia to Transoxiana rather than in the reverse direction. The age of the cluster overlaps with the formation of the Mongol Empire (13th century AD) making this suggestion plausible.

The modal haplotype 2 (Table 2) and 138 related haplotypes belonging to the C2\*-M217(xM48) haplogroup were arranged into a second phylogenetic network (Supplementary Fig. 4B). Here as well, two clusters can be distinguished:  $\gamma$  and  $\sigma$ . Cluster  $\gamma$  is prevalent among Mongolian-speaking Kalmyks and in Mongolia itself. The age of this cluster ( $600 \pm 200$  years) overlaps with the time of migration of the Kalmyk ancestors (Oyrats) from Mongolia and the following back migration of some Kalmyk groups. Cluster  $\sigma$  is specific to the Kazakh tribe

Modal haplotype	N*	N**	Haplotype cluster	N***	SNP-marker	DYS	DYS	DYS	DYS	DYS	DYS19-2****	DYS	DYS	DYS	DYS	DYS	DYS	GATA	DYS	DYS	DYS	DYS	DYS	TMRCAs (years) of cluster	
						3891	389b	390	456	19	385a	385b	458	437	438	448	H4	391	392	393	439	635	Rho	ASD	
Modal haplotype1	26	257	α	67	M48	14	17	25	15	16	—	12	12	17	14	10	20	10	10	11	13	11	23	600 ± 200	580
			β	188	M48	14	17	24	15	16	17	12	12	18	14	10	20	10	9	11	13	11	23	800 ± 200	659
Modal haplotype2	15	138	σ	76	M217(xM48)	14	15	23	15	15	—	11	20	18	14	10	21	11	10	11	14	12	21	1100 ± 400	1161
			γ	53	M217(xM48)	13	15	23	15	15	—	11	18	18	14	10	21	11	10	11	14	11	22	600 ± 200	704
Modal haplotype3	12	189	μ	185	M217(xM48)	13	16	25	15	16	—	12	13	18	14	10	22	11	10	11	13	10	21	1100 ± 300	1298
			λ	19	M217(xM48)	13	16	25	15	16	—	12	13	17	14	10	22	12	10	11	13	10	21	400 ± 100	407
Modal haplotype4	11	98	δ	97	M242	13	15	23	17	13	—	13	16	19	14	11	22	11	10	16	13	13	23	1400 ± 500	1360

**Table 2.** Features of the primary descent clusters. Notes: \*Number of samples carrying the modal haplotype; \*\*number of samples carrying related haplotypes (fewer than 5 mutational steps from the modal haplotype); \*\*\*number of samples within the given cluster. \*\*\*\*Duplication of the *DYS19* locus was observed only in some of M48 haplotypes. Each ASD estimate falls within confidence interval of the corresponding rho estimate, so we mention mainly rho estimates in the text.

Konyrat and modal haplotype accounts for 17% of the tribal paternal pool. The age of this cluster (1100 ± 400 years old) suggests a fairly early migration from Mongolia followed by an expansion within the single tribe.

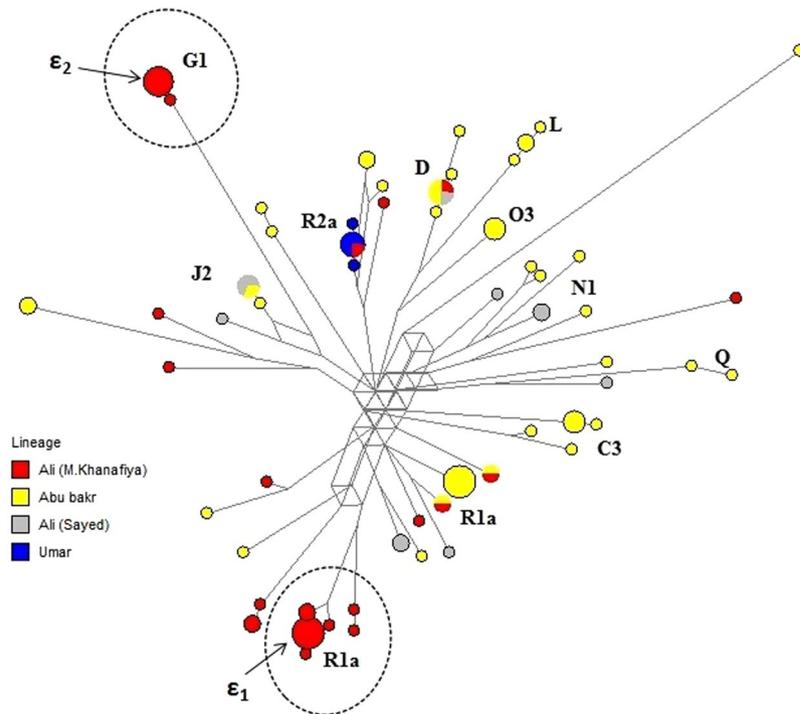
The modal haplotype 3 (Table 2) and 189 related haplotypes belonging to the C2\*-M217(xM48) haplogroup were plotted similarly (Supplementary Fig. 4C). This haplotype coincides with a previously-described haplotype, putatively connected to Genghis Khan's relatives, collectively forming the "C3\* star-cluster" (μ)<sup>21</sup>. From Abilev *et al.*<sup>26</sup> it is known that 76.5% of the Kazakh tribe Kerey belong to the star-cluster, including the 16% that fall within the third modal haplotype in our classification. Within Transoxiana, this founder haplotype is most common among the Kazakh clan Tore (11%), tribe Uysun (6%) and Karakalpaks (5%). The estimated age of the μ cluster (1100 ± 300 years) aligns with previous estimations of ~1000 years<sup>21, 30, 39</sup>. It may be assumed that modal haplotype 3 was the "proto-Mongolian haplotype", inherited, among others, by Genghis Khan, his descendants and patrilineal relatives. It is important to mention that Temujin (Genghis Khan) belonged to the Kiyat clan, which in turn is a branch of the Borjigin tribe, part of the Nirun Mongols. Subcluster λ, aged 400 ± 100 years old, is specific for Hazara from various countries and can be distinguished within the cluster.

The modal haplotype 4 (Table 2) and 97 related haplotypes belonging to the Q-M242 haplogroup were again plotted on a network (Supplementary Fig. 4D). The overwhelming majority of these haplotypes came from Turkmen populations from several countries. The cluster δ is 1400 ± 500 years old, making it older than the Mongol expansion. Despite a small part of the confidence interval overlapping with the period of Arab expansion, haplogroup Q-M242 accounts for just 1.5% of the population of the Arabian Peninsula, which means that expansion of this cluster in Turkmen populations is more likely caused by a local founder effect predating both Arab and Mongol influences.

Thus, three out of four signals of expansion in Transoxiana are connected to Mongol populations and likely reflect the migration to Transoxiana from Mongolia or neighboring regions which was followed by rapid growth of the migrants' descendants (Supplementary Text). Notably, such successful demic expansion was not accompanied by cultural expansion (language change) - most populations of present-day Transoxiana speak not Mongolian, but Turkic languages. The factor that unifies not just most, but all, populations of Transoxiana is Islam. However, our analysis has not revealed any signs of significant demic expansions linked to the Arabs. In a more direct attempt to uncover signs of such expansions, we have analyzed the Y-chromosomes of nomadic Islamic clergy.

**In search of Arab ancestry in Transoxiana.** The spread of Islam was one of the most powerful cultural expansions of West, South and Central Asia. There are ethnographic and genealogical findings identifying the demic traces of Arabs in Transoxiana. In particular, the clans Kozha and Sunak are traditionally considered as descendants of the Prophet Muhammad's close paternal-line relatives (Supplementary Text). These clans have maintained a privileged position among the nomadic populations of Transoxiana with a status similar to Genghis Khan's relatives, who were treated as nobility. While our previous study of the Kozha clan<sup>39</sup> was based on haplogroup frequencies, in this study we have analyzed STR haplotypes of the Kozha and Sunak clans.

Y-chromosomal haplotypes of the Kozha-Sunak tribe are shown in Fig. 4. Many separate individual haplotypes, or sometimes mini-clusters, can be observed. Therefore, unlike most Transoxianan clans, the Kozha and Sunak clans do not have a predominant paternal common ancestor. This is confirmed by the high haplogroup variation (HD = 0.86) among the Kozha-Sunak, which is 2–4 times higher than all other Transoxianan lineages studied (Supplementary Fig. 2). Due to the fact that subclans can have different origins, we divided the Kozha-Sunak into four groups based on traditional genealogy, and likewise divided all other Transoxianan clans. In the course of this analysis (Supplementary Table 6) we have discovered a pattern: the Kozha-Sunak lineages are highly heterogeneous based on the mean number of pairwise differences between haplotypes (PD between 8 and 9), while subclans of other Transoxianan clans are relatively homogeneous (PD between 1 and 7). The absence of one principal male root in the Kozha-Sunak tribe is indicative of its origin not being one Arab ancestor. Furthermore, haplogroup J1-M267, considered a marker of Arab expansion<sup>40</sup>, was not found among the Kozha-Sunak.



**Figure 4.** Haplotypic diversity of genealogical lineages within the Kozha-Sunak tribal-clan group, represented on a median-joining network.

In order to trace the origin of the haplotype mini-clusters identified (Fig. 4), we searched for related haplotypes (fewer than 5 mutational steps) in other Asian populations. Mini-cluster R1a ( $\epsilon_1$ ) had no related haplotypes, but mini-cluster G1 ( $\epsilon_2$ ) has related haplotypes among the Kazakh tribe Argyn<sup>33</sup>. This suggests that the origin of this mini-cluster is local and has not migrated with the Arab expansion.

Sayeds (lineage within the Kozha and Sunak clans) are known as descendants of the Prophet Muhammad on the paternal side, and reside beyond Transoxiana as well. Sayeds of Pakistan, for example, were analyzed by Belle *et al.*<sup>41</sup>, who reported that they are genetically closer to Arabs than to the surrounding populations of Pakistan and India, but did not find a founder effect. Furthermore, the Pakistani Sayeds are quite different genetically from the Transoxianan Sayeds. Thus, despite traditionally attributing their paternal ancestry to one common root among Arab missionaries, the pronounced Y-chromosomal variation suggests that Transoxiana's lineages have descended from several unrelated local ancestors.

A possible explanation is that nomadic clergy genealogy was based on *silsila*, a spiritual legacy passed from teacher to disciple, rather than a biological relationship. In Central Asia, Islam was spread by the Sufi orders Yasawiyya, Naqshbandi and Bektashi. In these orders, the leadership was based on *silsila*, a sequence of teachers who taught the succeeding leader of the Sufi order. However, even in agriculturalists spiritual succession was sometimes passed from father to son<sup>42</sup> and when occurring within a nomadic patronymic tradition could become a patrilineal biological legacy. The age of mini-cluster  $\epsilon_1$  ( $600 \pm 200$  years old) corresponds to the time when the Golden Horde adopted Islam as the official religion, and the rise of the Kozha-Sunak tribal-clan group in social status. This may have facilitated the transition from spiritual *silsila* to biological genealogy in order to maintain the privileged social status within the tribal-clan group. This conclusion coincides with the supposition of Heyer *et al.*<sup>43</sup> that cultural transmission of reproductive success could play an important role in shaping genetic diversity in Central Asia.

## Conclusions

We have analyzed human Y-chromosomal variation in ten populations from Transoxiana, a historical region covering Uzbekistan, western Tajikistan, western Kyrgyzstan, northwestern Turkmenistan and southern Kazakhstan. Considering the peculiar features of the geographical landscape of the region, abrupt shifts of cultural landscapes in the course of its history, and presence of patrilineal tribal-clan groups, we jointly analyzed the patrilineal genetic variation, patrilineal genealogies and historical data. We identified three features of the genetic landscape of Transoxiana and its connection to geographical and cultural landscapes.

First, cultural and demic expansions of Transoxiana were not closely connected with each other. Arab cultural expansion introduced Islam to the region but did not leave a significant mark on the Y-chromosomal pool. The Mongol expansion, in contrast, had enormous demic success, but did not impact on cultural elements like language and religion.

Second, the geographic landscape of Transoxiana, despite its peculiarity and diversity (deserts, fertile river basins, foothills and plains) had no strong influence on the genetic landscape. The main factor structuring the Y-chromosomal variation was the mode of subsistence: settled agriculture or nomadic pastoralism.

Third, the genealogy of Muslim missionaries within the settled agricultural community was based on spiritual succession passed from teacher to disciple, rather than on biological relationship. However, among nomads, spiritual and biological succession merged, leading to the formation of haplotype mini-clusters among nomadic clergy.

## Methods

**Samples.** Blood samples were collected in 2009–2012 by the expeditions of the Laboratory of Human Population Genetics of the Research Centre for Medical Genetics, Genome Geography Laboratory of the Vavilov Institute of General Genetics, The National Laboratory Astana of Nazarbayev University, Forensic science centre of the Ministry of Justice of the Republic of Kazakhstan (Astana), Center of High Technologies and Institute of Bioorganic Chemistry (Tashkent), partially under the auspices of the Genographic Project. All expeditions were supervised by Elena Balanovska and followed the same sampling rules<sup>44</sup>.

10 populations of 5 ethnic groups were studied (Supplementary Table 1). Only individuals who had all their ancestors for at least three generations descending from the specific population, and not related to each other, were selected. All 780 subjects provided their written informed consent in a form approved by the Ethics Committee of the Research Centre for Medical Genetics (Moscow, Russia). When performing genotyping and data analyses, we followed lab protocols approved by the same Ethics Committee.

The MDS analysis dataset consisted of 5,998 samples from 79 populations, including 780 samples reported here for the first time (Supplementary Table 2)<sup>7, 29, 45–52</sup>. PCA included data on 1,944 samples from 33 Central Asian populations (Supplementary Table 4)<sup>29, 30, 50</sup> and this study.

The phylogenetic analysis dataset included 15-STR profiles of 4,495 samples from Central Asia and neighboring regions<sup>2, 25–29, 33, 45, 48, 53–55</sup> and this study (Supplementary Fig. 4A–D). Both the MDS and the phylogenetic dataset were extracted from the in-house *Y-base* database, compiling published data on Y-chromosomal variation in human populations.

**Genotyping.** DNA was extracted from white blood cells of peripheral blood using standard methods<sup>56</sup>. The effective DNA concentration was determined by real-time PCR using the Quantifiler Human DNA Kit (Applied Biosystems), followed by normalization of DNA to a concentration of 2 ng/μl.

To determine the Y-chromosome haplogroup, 35 SNP markers were genotyped (M217, M48, M174, M35.1, M78, M123, M285, P15, P303, M406, M69, M170, M267, P58, M172, M47, M67, M12, M92, M20, LLY22g, M128, M178, M119, P31, M122, P201, M134, M242, M198, M458, M343, M269, M124, M70) using the TaqMan probes on the 7900HT instrument (Applied Biosystems) according to the manufacturer's protocol. Haplogroups were classified according to ISOGG-2016 (Version: 11.208; Date: 30 July 2016)<sup>57</sup> (Fig. 2, Supplementary Fig. 1, Supplementary Table 1).

17 Y-STR loci were genotyped in all samples (DYS389I, DYS389II, DYS390, DYS456, DYS19, DYS385a, DYS385b, DYS458, DYS437, DYS438, DYS448, GATA\_H4, DYS391, DYS392, DYS393, DYS439, DYS635) using the Yfiler PCR Amplification Kit (Applied Biosystems) on the ABI 3130xl genetic analyzer (Applied Biosystems). The results were analyzed using the GeneMapper Software v. 4.1 (Applied Biosystems).

**Statistical methods.** Genetic distances between 79 populations were calculated using Nei's method and DJ software (Supplementary Table 3)<sup>58</sup>. Multidimensional scaling, cluster analysis (Ward's method), and principal component analysis were conducted using the Statistica v.7.1 software<sup>59</sup>. Genetic differentiation within and among groups of populations (AMOVA) was performed in Arlequin 3.5.1.3 software<sup>60</sup>.

15-STR-haplotypes were analyzed (DYS385a and DYS385b were excluded; for DYS389, DYS389I and DYS389b [=DYS398II-DYS389I] were used). Haplotype variability statistics were calculated using the Arlequin 3.5.1.3 software. The search for related haplotypes was conducted using the Haplomatch software<sup>38</sup>. Phylogenetic analysis was conducted using the Reduced-Median method<sup>61</sup> in Network 5 software<sup>62</sup> (<http://www.fluxus-engineering.com/sharenet.htm>) with reduction limit = 1. The resulting phylogenetic networks were then edited using Network Publisher<sup>63</sup> (<http://www.fluxus-engineering.com/nwpub.htm>).

In distinguishing the clusters on the networks, we followed the procedure described earlier<sup>11, 64</sup>. Briefly, we first looked for a zone in the network carrying mostly haplotypes from a single population; then the most recent common ancestral haplotype for all haplotypes in the zone was identified in the network; finally, all haplotypes downstream to this ancestral haplotype were attributed to the cluster, though some very distant haplotypes were ignored. This procedure transforms the initial arbitrary "zone" to a monophyletic – to the best of the network's performance – clade. In previous studies<sup>64</sup> we selected the ancestral haplotype as cluster's founder, but<sup>65</sup> found that using the modal haplotype works much better, so here we selected modal haplotypes as founders. Cluster age was determined using the rho-statistic<sup>66, 67</sup> and, because rho was shown to introduce a systematic bias<sup>68</sup>, we also used the average squared distance (ASD) estimator<sup>69, 70</sup>. Rho was calculated using Network 5, and ASD was calculated by Y TMRCA Calculator (<http://ehelix.pythonanywhere.com/init/default/about>), which is a derivative of Matlab/Octave program Ytime<sup>71</sup>. To convert the number of mutations into number of generations, the "genealogical" mutation rate of  $2.1 \times 10^{-3}$  mutations per STR per generation was used<sup>72, 73</sup> as the analysis in Karmin<sup>74</sup> indicated that for clusters younger than 30,000 years, this rate is consistent with full Y-chromosomal sequence data. When converting the number of generations into an age in years, the male generation time was set to 30 years<sup>75</sup>. To determine whether or not most haplotypes in the same genealogical lineage originated from a common ancestor, we used phylogenetic networks, haplotype variability statistics and the mean number of pairwise differences within each lineage (calculated using Arlequin 3.5.1.3 software).

## References

- Rosser, Z. H. *et al.* Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics* **67**, 1526–1543, doi:10.1086/316890 (2000).
- Haber, M. *et al.* Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. *Eur J Hum Genet* **19**, 334–340, doi:10.1038/ejhg.2010.177 (2011).
- Balanovsky, O. *et al.* Genetic Differentiation between Upland and Lowland Populations Shapes the Y-Chromosomal Landscape of West Asia. *Human Genetics* **136**, 437–450, doi:10.1007/s00439-017-1770-2 (2017).
- Rtveladze, E. V. *Civilizations, states, and cultures of Central Asia*. (University of World Economy and Diplomacy, 2008).
- Baumer, C. *The history of Central Asia: The Age of the Steppe Warriors*. Vol. 1 (I.B.Tauris & Co Ltd, 2012).
- Baumer, C. *The history of Central Asia: The Age of the Silk Roads*. Vol. 2 (I.B.Tauris & Co Ltd, 2014).
- Zalloua, P. A. *et al.* Y-chromosomal diversity in Lebanon is structured by recent historical events. *American Journal of Human Genetics* **82**, 873–882, doi:10.1016/j.ajhg.2008.01.020 (2008).
- Agrawal, S., Khan, F., Pandey, A., Tripathi, M. & Herrera, R. J. YAP, signature of an African-Middle Eastern migration into northern India. *Current Science* **88**, 1977–1980 (2005).
- Beleza, S., Gusmão, L., Amorim, A., Carracedo, A. & Salas, A. The genetic legacy of western Bantu migrations. *Hum Genet* **117**, 366–375, doi:10.1007/s00439-005-1290-3 (2005).
- de Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc Biol Sci* **279**, 3256–3263, doi:10.1098/rspb.2012.0318 (2012).
- Balanovsky, O. *Genepool of Europe*. (In Russian, KMK Scientific Press Ltd., 2015).
- Martinez-Cruz, B. *et al.* In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *European Journal of Human Genetics* **19**, 216–223, doi:10.1038/ejhg.2010.153 (2011).
- Yunusbayev, B. *et al.* The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet* **11**, e1005068, doi:10.1371/journal.pgen.1005068 (2015).
- Comas, D. *et al.* Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet* **12**, 495–504, doi:10.1038/sj.ejhg.5201160 (2004).
- Quintana-Murci, L. *et al.* Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am J Hum Genet* **74**, 827–845, doi:10.1086/383236 (2004).
- Irwin, J. A. *et al.* The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. *Int J Legal Med* **124**, 195–204, doi:10.1007/s00414-009-0406-z (2010).
- Perez-Lezaun, A. *et al.* Sex-specific migration patterns in central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *American Journal of Human Genetics* **65**, 208–219, doi:10.1086/302451 (1999).
- Hammer, M. F. *et al.* Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* **18**, 1189–1203 (2001).
- Wells, R. S. *et al.* The Eurasian Heartland: A continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10244–10249, doi:10.1073/pnas.171305098 (2001).
- Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *American Journal of Human Genetics* **71**, 466–482, doi:10.1086/342096 (2002).
- Zerjal, T. *et al.* The genetic legacy of the mongols. *American Journal of Human Genetics* **72**, 717–721, doi:10.1086/367774 (2003).
- Chaix, R. *et al.* The genetic or mythical ancestry of descent groups: Lessons from the Y chromosome. *American Journal of Human Genetics* **75**, 1113–1116, doi:10.1086/425938 (2004).
- Chaix, R. *et al.* From social to genetic structures in central Asia. *Curr Biol* **17**, 43–48, doi:10.1016/j.cub.2006.10.058 (2007).
- Heyer, E. *et al.* Genetic diversity and the emergence of ethnic groups in Central Asia. *Bmc Genetics* **10**, doi:10.1186/1471-2156-10-49 (2009).
- Dulik, M. C., Osipova, L. P. & Schurr, T. G. Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS One* **6**, e17548, doi:10.1371/journal.pone.0017548 (2011).
- Abilev, S. *et al.* The Y-Chromosome C3\*Star-Cluster Attributed to Genghis Khan's Descendants Is Present at High Frequency in the Kerey Clan from Kazakhstan. *Human Biology* **84**, 79–89 (2012).
- Tarlykov, P. V. *et al.* Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. *Croat Med J* **54**, 17–24 (2013).
- Haber, M. *et al.* Afghanistan's Ethnic Groups Share a Y-Chromosomal Heritage Structured by Historical Events. *Plos One* **7**, doi:10.1371/journal.pone.0034288 (2012).
- Di Cristofaro, J. *et al.* Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. *Plos One* **8**, doi:10.1371/journal.pone.0076748 (2013).
- Balaresque, P. *et al.* Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. *Eur J Hum Genet* **23**, 1413–1422, doi:10.1038/ejhg.2014.285 (2015).
- Biró, A. Z., Zalán, A., Völgyi, A. & Pamjav, H. A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). *Am J Phys Anthropol* **139**, 305–310, doi:10.1002/ajpa.20984 (2009).
- Biró, A., Fehér, T., Bárányi, G. & Pamjav, H. Testing Central and Inner Asian admixture among contemporary Hungarians. *Forensic Sci Int Genet* **15**, 121–126, doi:10.1016/j.fsigen.2014.11.007 (2015).
- Balanovsky, O. *et al.* Deep Phylogenetic Analysis of Haplogroup G1 Provides Estimates of SNP and STR Mutation Rates on the Human Y-Chromosome and Reveals Migrations of Iranian Speakers. *Plos One* **10**, doi:10.1371/journal.pone.0122968 (2015).
- Bellew, H. W. *The races of Afghanistan: being a brief account of the principal nations inhabiting that country*. (Thacker, Spink & Co., 1880).
- Qamar, R. *et al.* Y-chromosomal DNA variation in Pakistan. *Am J Hum Genet* **70**, 1107–1124, doi:10.1086/339929 (2002).
- Dani, A. H. & Masson, V. M. *History of civilizations of Central Asia: The Dawn of Civilization: Earliest Times to 700 B.C.*, Vol. 1 (UNESCO, 1992).
- Palstra, F. P., Heyer, E. & Austerlitz, F. Statistical inference on genetic data reveals the complex demographic history of human populations in central Asia. *Mol Biol Evol* **32**, 1411–1424, doi:10.1093/molbev/msv030 (2015).
- Chukhryaeva, M. I. *et al.* The haplomatch program for comparing Y-chromosome STR-haplotypes and its application to the analysis of the origin of Don Cossacks. *Russian Journal of Genetics* **52**, 521–529, doi:10.1134/S1022795416050045 (2016).
- Zhabagin, M. K. *et al.* The relation between the Y-chromosomal variation and the clan structure: the gene pool of the steppe aristocracy and the steppe clergy of the Kazakhs. *Moscow University Anthropology Bulletin* **1**, 96–101 (2014).
- Tofanelli, S. *et al.* J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur J Hum Genet* **17**, 1520–1524, doi:10.1038/ejhg.2009.58 (2009).
- Belle, E. M. S., Shah, S., Parfitt, T. & Thomas, M. G. Y chromosomes of self-identified Syeds from the Indian subcontinent show evidence of elevated Arab ancestry but not of a recent common patrilineal origin. *Archaeological and Anthropological Sciences* **2**, 217–224, doi:10.1007/s12520-010-0040-1 (2010).
- Prozоров, S. M. *Islam in the territories of the Former Russian Empire: Encyclopaedic Lexicon*. Vol. 1 (In Russian, Oriental Literature, 2006).
- Heyer, E. *et al.* Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *Am J Phys Anthropol* **157**, 537–543, doi:10.1002/ajpa.22739 (2015).

44. Balanovska, E. V. *et al.* Population Biobanks: Organizational Models and Prospects of Application in Gene Geography and Personalized Medicine. *Russian Journal of Genetics* **52**, 1227–1243, doi:10.1134/S1022795416120024 (2016).
45. Abu-Amero, K. K. *et al.* Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions. *Bmc Genetics* **10**, 10.1186/1471-2156-10-59 (2009).
46. Al-Zahery, N. *et al.* Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Molecular Phylogenetics and Evolution* **28**, 458–472, doi:10.1016/s1055-7903(03)00039-3 (2003).
47. Al-Zahery, N. *et al.* In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. *BMC Evol Biol* **11**, 288, doi:10.1186/1471-2148-11-288 (2011).
48. Cadenas, A. M., Zhivotovsky, L. A., Cavalli-Sforza, L. L., Underhill, P. A. & Herrera, R. J. Y-chromosome diversity characterizes the Gulf of Oman. *European Journal of Human Genetics* **16**, 374–386, doi:10.1038/sj.ejhg.5201934 (2008).
49. Cinnioglu, C. *et al.* Excavating Y-chromosome haplotype strata in Anatolia. *Human Genetics* **114**, 127–148, doi:10.1007/s00439-003-1031-4 (2004).
50. Grugni, V. *et al.* Ancient Migratory Events in the Middle East: New Clues from the Y-Chromosome Variation of Modern Iranians. *Plos One* **7**, doi:10.1371/journal.pone.0041252 (2012).
51. Luis, J. R. *et al.* The Levant versus the Horn of Africa: Evidence for bidirectional corridors of human migrations. *American Journal of Human Genetics* **74**, 532–544, doi:10.1086/382286 (2004).
52. Sengupta, S. *et al.* Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet* **78**, 202–221, doi:10.1086/499411 (2006).
53. Dulik, M. C. *et al.* Mitochondrial DNA and Y Chromosome Variation Provides Evidence for a Recent Common Ancestry between Native Americans and Indigenous Altaians. *American Journal of Human Genetics* **90**, 229–246, doi:10.1016/j.ajhg.2011.12.014 (2012).
54. El-Sibai, M. *et al.* Geographical structure of the Y-chromosomal genetic landscape of the Levant: a coastal-inland contrast. *Ann Hum Genet* **73**, 568–581, doi:10.1111/j.1469-1809.2009.00538.x (2009).
55. Malyarchuk, B. *et al.* Y-chromosome diversity in the Kalmyks at the ethnical and tribal levels. *Journal of Human Genetics* **58**, 804–811, doi:10.1038/jhg.2013.108 (2013).
56. Green, M. R. & Sambrook, J. *Molecular cloning: a laboratory manual*. 4th edn, (Cold Spring Harbor Laboratory Press, 2012).
57. *International Society of Genetic Genealogy*. <http://isogg.org/tree/index.html>. Accessed 1 June 2016.
58. Balanovsky, O. *et al.* Two sources of the Russian patrilineal heritage in their Eurasian context. *Am J Hum Genet* **82**, 236–250, doi:10.1016/j.ajhg.2007.09.019 (2008).
59. Statistica Software, version 7.1 (StatSoft Inc., Tulsa, OK, USA, 2005).
60. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**, 564–567, doi:10.1111/j.1755-0998.2010.02847.x (2010).
61. Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human-populations using median Networks. *Genetics* **141**, 743–753 (1995).
62. Network 5 <http://www.fluxus-engineering.com/sharenet.htm> (Fluxus Technology Ltd., 2016).
63. Network Publisher <http://www.fluxus-engineering.com/nwpub.htm> (Fluxus Technology Ltd., 2016).
64. Balanovsky, O. *et al.* Parallel Evolution of Genes and Languages in the Caucasus Region. *Molecular Biology and Evolution* **28**, 2905–2920, doi:10.1093/molbev/msr126 (2011).
65. Hallast, P. *et al.* The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol Biol Evol* **32**, 661–673, doi:10.1093/molbev/msu327 (2015).
66. Forster, P., Harding, R., Torroni, A. & Bandelt, H. J. Origin and evolution of native American mtDNA variation: A reappraisal. *American Journal of Human Genetics* **59**, 935–945 (1996).
67. Saillard, J., Forster, P., Lynnnerup, N., Bandelt, H. J. & Norby, S. mtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. *American Journal of Human Genetics* **67**, 718–726, doi:10.1086/303038 (2000).
68. Cox, M. P. Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models. 2008. *Hum Biol* **81**, 911–933, doi:10.3378/027.081.0631 (2009).
69. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463–471 (1995).
70. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA* **92**, 6723–6727 (1995).
71. Behar, D. M. *et al.* Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. *Am J Hum Genet* **73**, 768–779, doi:10.1086/378506 (2003).
72. Gusmao, L. *et al.* Mutation rates at Y chromosome specific microsatellites. *Human Mutation* **26**, 520–528, doi:10.1002/humu.20254 (2005).
73. Ge, J. *et al.* Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Science International-Genetics* **3**, 179–184, doi:10.1016/j.fsigen.2009.01.007 (2009).
74. Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res* **25**, 459–466, doi:10.1101/gr.186684.114 (2015).
75. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* **128**, 415–423, doi:10.1002/ajpa.20188 (2005).
76. *Central Asia atlas of natural resources*. (Asian Development Bank, 2010).

## Acknowledgements

We gratefully acknowledge all sample donors who participated in this study. We thank Aslan Abdilla for helpful discussions of the Kozha-Sunak origins. MZ has been funded by funding of Ministry of Education and Science of Republic Kazakhstan (№0114PK00492, №0115PK01931). AAg, RS, and OlegB were supported by the Russian Scientific Foundation grant 17-14-01345 (analysis of demic expansions from Central Asia/FarEast). CTS was supported by The Wellcome Trust (098051).

## Author Contributions

Conceived and designed the experiments: M.Z., Oleg B., E.B.; Performed the experiments: M.K., A.Ag, Olga B., M.C., N.M., A.R., R.S., I.T., V.Z.; Analyzed the data: M.Z., Oleg B.; Contributed reagents/materials/analysis tools: A. Ak, I.T., Z.S., Y.Y., Di D., Da D., S.T., L.S., E.B.; Wrote the paper: M.Z., E.B., Oleg B., C.T.S.; Study initiation: M.Z.; Read and approved the final version of the paper: all coauthors.

## Additional Information

**Supplementary information** accompanies this paper at doi:[10.1038/s41598-017-03176-z](https://doi.org/10.1038/s41598-017-03176-z)

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017