PLOS ONE

Kazak mitochondrial genomes provide insights into the human population history of Central Eurasia

Ayken Askapuli[1,2,3,4]*, Miguel Vilar[5,6], Humberto Garcia-Ortiz[7], Maxat Zhabagin[1,2,3], Zhaxylyk Sabitov[8], Ainur Akilzhanova[3], Erlan Ramanculov[1,2], Uli Schamiloglu[1], Angelica Martinez-Hernandez[7], Cecilia Contreras-Cubas[7], Francisco Barajas-Olmos[7], Theodore G. Schurr[9], Zhaxybay Zhumadilov[3,10], Marlen Flores-Huacuja[7], Lorena Orozco[7], John Hawks[4,11], Naruya Saitou[12,13,14]*

1 School of Sciences and Humanities, Nazarbayev University, Astana, Kazakhstan, 2 National Center for Biotechnology, Astana, Kazakhstan, 3 National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan, 4 Department of Integrative Biology, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, 5 The Genographic Project, National Geographic Society, Washington, DC, United States of America, 6 Department of Anthropology, University of Maryland, College Park, Maryland, United States of America, 7 Immunogenomics and Metabolic Diseases Laboratory, National Institute of Genomic Medicine, Mexico City, Mexico, 8 L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, 9 Department of Anthropology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 10 School of Medicine, Nazarbayev University, Astana, Kazakhstan, 11 Department of Anthropology, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, 12 Population Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka, Japan, 13 Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, Japan, 14 Advanced Medical Research Center, Faculty of Medicine, University of the Ryukyus, Okinawa Ken, Japan

* ayken.askapuli@wisc.edu (AA); saitounr@nig.ac.jp (NS)

Abstract

As a historical nomadic group in Central Asia, Kazaks have mainly inhabited the steppe zone from the Altay Mountains in the East to the Caspian Sea in the West. Fine scale characterization of the genetic profile and population structure of Kazaks would be invaluable for understanding their population history and modeling prehistoric human expansions across the Eurasian steppes. With this mind, we characterized the maternal lineages of 200 Kazaks from Jetisuu at mitochondrial genome level. Our results reveal that Jetisuu Kazaks have unique mtDNA haplotypes including those belonging to the basal branches of both West Eurasian (R0, H, HV) and East Eurasian (A, B, C, D) lineages. The great diversity observed in their maternal lineages may reflect pivotal geographic location of Kazaks in Eurasia and implies a complex history for this population. Comparative analyses of mitochondrial genomes of human populations in Central Eurasia reveal a common maternal genetic ancestry for Turko-Mongolian speakers and their expansion being responsible for the presence of East Eurasian maternal lineages in Central Eurasia. Our analyses further indicate maternal genetic affinity between the Sherpas from the Tibetan Plateau with the Turko-Mongolian speakers.

Introduction

Central Asia is a region that has witnessed important events in modern human population history, including the westward migration of nomadic pastoralists [1, 2], the eastward expansion of Indo-European speakers [3–5], and admixture between anatomically modern humans and archaic humans [6–9]. Notwithstanding the historical significance of the region, in-depth, comprehensive genetic studies of human populations in Central Asia are generally sparse. While the distribution of paternal lineages in Central Asia and Siberia is associated with language and ethnic identity [10, 11], maternal lineages do not exhibit any associations with language and ethnic group, except for an increase of West Eurasian mtDNA haplogroups in an east-to-west direction [12].

Previous studies have showed that human populations in Central Asia, including Kazaks, harbor elevated frequencies of both East and West Eurasian maternal lineages [12–14]. To date, there are several studies of maternal lineages in Kazak populations from the Altay Republic (Russia) [12, 15], from Xinjiang (China) [16, 17], and from Kazakhstan [13, 14, 18, 19]. While most of these studies sequenced the mitochondrial DNA (mtDNA) control region and genotyped coding region diagnostic SNPs, few studies analyzed the mitochondrial genomes (mtGenomes) of Kazaks (N = 19) [15, 20–22]. These studies again revealed the presence of both East and West Eurasian lineages in Kazak populations. Ingman and Gyllensten (2007) sequenced mitochondrial genomes (haplogroups A, D, and T) from six Kazaks. Derenko et al. (2012, 2014) presented whole mitochondrial genome sequences for twelve Kazaks, which belong to haplogroups M, H, HV, and U. Besides, Sahakyan and colleagues published a single Kazak mtGenome of haplogroup U in 2017 [21].

Historically speaking, there are three major divisions of Kazak populations. They include the Great Jüz, Middle Jüz, and Junior Jüz. The three Jüz occupy different, albeit overlapping, regions of Kazakhstan and bordering countries. While the Junior Jüz resides in the west, the Middle Jüz inhabits central, north, and east Kazakhstan, and the Great Jüz lives in southern regions of the country.

To evaluate the maternal lineages of Kazaks at a finer scale, and to better understand the origins and dispersal of maternal lineages in Central Eurasia, we analyzed the mitochondrial genome data of 200 Kazak individuals, including full mtGenomes from 120 individuals and microarray data based pseudo-mtGenomes for 80 individuals, all living in the Jetisuu region of southern Kazakhstan (Fig 1).

These individuals belong to four clans of the Great Jüz, including Jalayir, Alban, Shapir-ashti, and Suan. Our results show that while a diverse array of haplogroups of Eastern and Western Eurasian origins comprise Kazak maternal lineages, Eastern Eurasian haplogroups, including A, B, C, D, F and G, account for about 50% of maternal genetic diversity of Jetisuu Kazaks. These findings indicate that mtDNA can help delineate historical processes of population formation in communities that are patrilineal in their social organization. We also performed comparative analyses of mtDNA variation in Jetisuu Kazaks and previously published mtGenomes (N = 2074) from 21 ethnic groups from across Central Eurasia. Our analyses indicate that mtDNA haplogroups are associated with language background at the language family level, while the presence of East Eurasian lineages in Central Eurasia is associated with the expansion of Turko-Mongolian speakers.

Results

mtDNA haplogroups in Jetisuu Kazaks

As shown in Table 1, Jetisuu Kazaks harbor at least 56 different haplogroups. Seven of them (C4, D4, F1, G2, H2, T2, and U5) occur at an average frequency of 3.5% or higher and
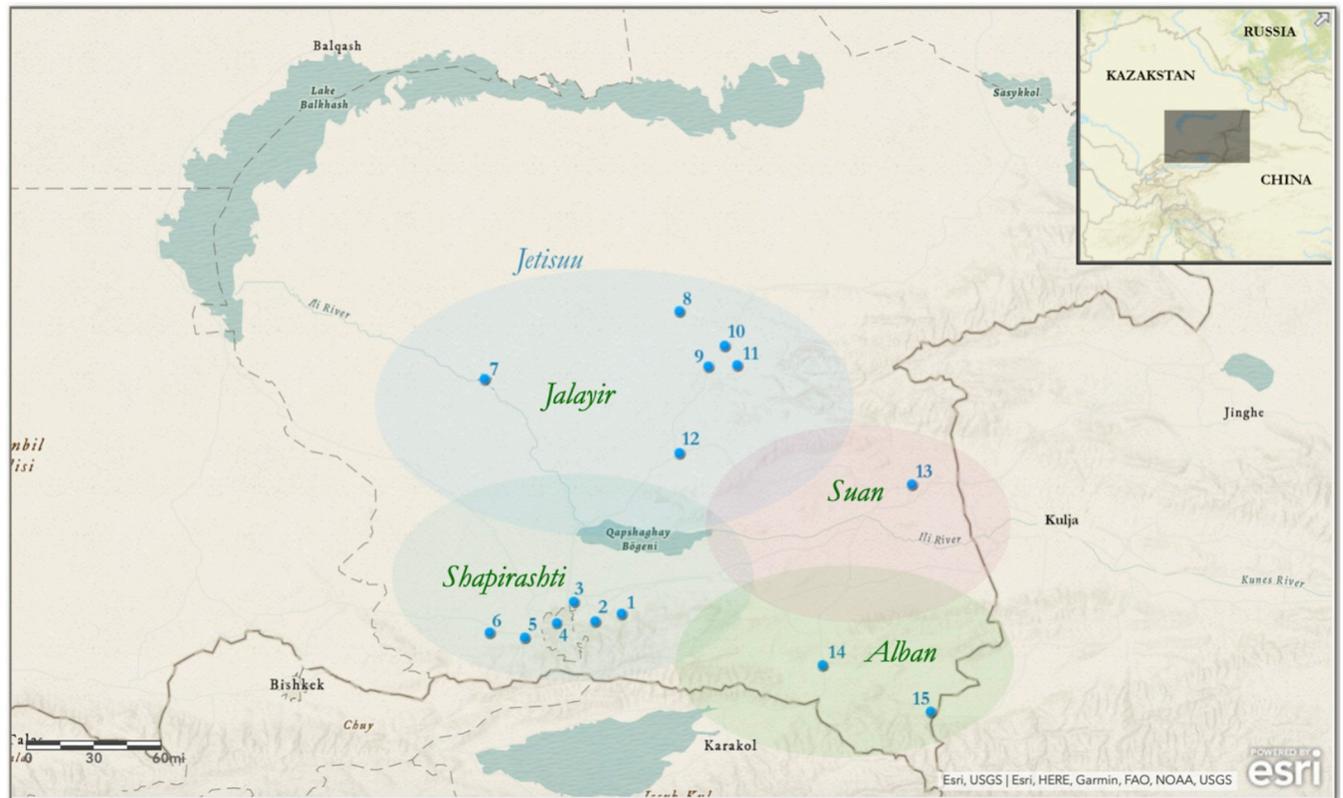
Fig 1. Map of Jetisuu and the sampling sites. 1. Engbekshi Kazak; 2. Talghar; 3. Ile; 4. Almati; 5. Karasay; 6. Jambil; 7. Balkash; 8. Karatal; 9. Köksuu; 10. Taldikorghan; 11. Eskeldi; 12. Kerbulak; 13. Panfilov; 14. Kegen; 15. Rayimbek.

https://doi.org/10.1371/journal.pone.0277771.g001

appear in at least three of the clans surveyed in this study. Another six haplogroups (A5, B4, C5, D5, N9, and U2) appeared at a frequency of 2.0–3.0% and occurred in at least two of the clans. The remaining 43 haplogroups were represented by just one to three individuals (0.5–1.5%). Only B4, D4, G2, H2, and U5 mtDNAs were found in all four Jetisuu Kazak populations. In addition, haplogroup and genetic diversity was associated with the sample size in the Jetisuu Kazak populations (Tables 1 and 2), with Jalayirs (N = 76) having the highest number of haplogroups and Suans (N = 26) having the lowest haplogroup diversity (Tables 1 and 2).

mtDNA control region based PCA analysis

We compared mtDNA profiles of the Jetisuu Kazaks with geographically adjacent populations based on haplogroup frequencies determined from the mtDNA control region sequences, since only control region data were available for most of the populations analyzed. As shown in the resulting PCA (Principal Component Analysis) plot, three groupings of populations appear in this analysis: (1) Turko-Mongolians clustered around indigenous Altaians; (2) Sino-Tibetan speaking groups; and (3) Uyghurs-Uzbeks (Fig 2). The Turko-Mongolian cluster was comprised of Turkic speaking Altaians at the center, surrounded by Turkic speakers (Kazak populations from Jetisuu (Kazakhstan), from Xinjiang (China), and from the Altay Republic (Russia), and Kirghiz from Kyrgyzstan) and Mongolic speakers (Mongols from Inner Mongolia and Xinjiang (China), and Barghuts from Inner Mongolia). Notwithstanding the vast geographic distance between them, Mongols and Barghuts from

Table 1. Mitochondrial DNA haplogroup frequencies in Kazak populations from Jetisuu, Kazakhstan.

Hg	Alban		Jalayir		Shapirashti		Suan		JetiKaz	
A	1	2.86	2	2.63	1	1.69			4	2
A12			1	1.32	1	1.69			2	1
A14					1	1.69			1	0.5
A5	1	2.86	1	1.32					2	1
B4	1	2.86	1	1.32	1	1.69	1	3.85	4	2
B5					2	3.39	1	3.85	3	1.5
C4	1	2.86	6	7.89	4	6.78			11	5.5
C5					3	5.08	1	3.85	5	2.5
C7					1	1.69			1	0.5
D1			1	1.32					1	0.5
D4	5	14.29	10	13.16	11	18.64	5	19.23	32	16
D5			3	3.95	3	5.08			6	3
F1			6	7.89	2	3.39	3	11.54	11	5.5
F2							1	3.85	1	0.5
G2	2	5.71	7	9.21	5	8.47	1	3.85	15	7.5
G3			1	1.32					1	0.5
H	4	11.43			1	1.69	1	3.85	6	3
H1	1	2.86			2	3.39			3	1.5
H10			1	1.32					1	0.5
H13			2	2.63	1	1.69			3	1.5
H2	2	5.71	8	10.53	3	5.08			13	6.5
H3	1	2.86							1	0.5
H44							1	3.85	1	0.5
H5							1	3.85	1	0.5
H6			1	1.32					1	0.5
H8			1	1.32	1	1.69			2	1
H9	1	2.86							1	0.5
HV					1	1.69			1	0.5
HV1					1	1.69			1	0.5
HV14	1	2.86							1	0.5
HV2			2	2.63					3	1.5

Hg	Alban		Jalayir		Shapirashti		Suan		JetiKaz	
I1					1	1.69			1	0.5
I2					1	1.69			1	0.5
I4			1	1.32					1	0.5
J1	1	2.86			1	1.69	1	3.85	3	1.5
K1			1	1.32					1	0.5
M10	1	2.86							1	0.5
M5			1	1.32					1	0.5
M65					1	1.69			1	0.5
M7					1	1.69	1	3.85	2	1
N1	1	2.86	1	1.32					2	1
N9	2	5.71	1	1.32			1	3.85	5	2.5
R0	1	2.86							1	0.5
R1			1	1.32					1	0.5
R2	1	2.86							1	0.5
R9	1	2.86	1	1.32					2	1
T1	1	2.86	1	1.32					2	1
T2			4	5.26	5	8.47	2	7.69	11	5.5
U2	1	2.86	3	3.95					4	2
U4	1	2.86	1	1.32			1	3.85	3	1.5
U5	2	5.71	2	2.63	2	3.39	1	3.85	7	3.5
U7	1	2.86			1	1.69			2	1
V14							1	3.85	1	0.5
W3			1	1.32					1	0.5
W6			1	1.32					1	0.5
X2i					1	1.69			1	0.5
Z1							2	7.69	2	1
Z3			1	1.32					1	0.5
Z4			1	1.32					1	0.5
Hg	24		33		26		18		59	
N	35		76		59		26		200*	

Notes: Haplogroups in this table are based on the haplogroup calls obtained via Haplogrep 2.0 [23] and mtPhyl 5.003 [24];

*The total number of mtDNA analyzed is *N* = 200, including four samples with ambiguous clan identity, but are from the four clans (their haplogroups are D4, HV2, C5, and N9, respectively) and 196 samples with unambiguous clan identities; Frequencies are shown in italics (%); **Hg:** Haplogroups; **N:** Number of samples; **JetiKaz**: all 200 Kazak samples from Jetisuu; Total number of haplogroups: 59.

Inner Mongolia clustered close to Mongols, Kazaks, and Kirghiz in Central Asia. By contrast, Uyghurs and Uzbeks are genetically distant from their geographic, religious, and linguistic kin in Central Asia.

With regard to Kazak populations, they did not form a unique (single) cluster, but instead were scattered around the PCA plot (Fig 2). This distribution reveals an elevated maternal genetic diversity in Kazaks, as noted previously [12]. Mitochondrial haplogroup frequencies in Kazaks also varied depending on sampling sites and sample sizes. The most extreme disparities were observed between Kazak populations from Xinjiang (XJ_Kaz_a and XJ_Kaz_b), while an analogous situation was also observed in Altaian Kazaks (Alt_Kaz_a and Alt_Kaz_b), indicating the pronounced effect of sampling (size, coverage, location etc.).

Table 2. Results from Tajima's Neutrality Test [33].

Populations	Seq. Type	N	S	Ps	Θ	π	D
Alban	Pseudo	15	160	0.0096566	0.0029698	0.0016957	-1.8932561
	Genuine	20	192	0.0115879	0.0032663	0.0018621	-1.7843685
	Combined	**35**	**308**	**0.0185889**	**0.0045138**	**0.0019007**	**-2.2005190**
Jalayir	Pseudo	25	208	0.0125536	0.0033246	0.0019530	-1.6451853
	Genuine	51	443	0.0267367	0.0059425	0.0021242	-2.3341813
	Combined	**76**	**526**	**0.0317460**	**0.0064770**	**0.0021572**	**-2.3233549**
Shapirashti	Pseudo	27	221	0.0133382	0.0034605	0.0019516	-1.7178648
	Genuine	32	276	0.0166576	0.0041362	0.0020340	-1.9549262
	Combined	**59**	**383**	**0.0231155**	**0.0049751**	**0.0021047**	**-2.0590438**
Suan	Pseudo	13	142	0.0085702	0.0027617	0.0018354	-1.5346953
	Genuine	13	153	0.0092341	0.0029757	0.0020451	-1.4321028
	Combined	**26**	**244**	**0.0147263**	**0.0038591**	**0.0020544**	**-1.8558244**
Jetisuu Kaz	Pseudo	80	412	0.0248657	0.0050204	0.0019058	-2.1460627
	Genuine	120	689	0.0415837	0.0077574	0.0020337	-2.4686793
	Combined	**200**	**854**	**0.0515420**	**0.0087761**	**0.0020823**	**-2.4548640**

Notes: This analysis involved 200 mitochondrial sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were 16569 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [34].

Abbreviations: N = number of sequences, **S** = Number of segregating sites, Ps = S/n, **n** = total number of sites, $Θ$ = Ps/a, $π$ = nucleotide diversity, and D is the Tajima test statistic [35].

Although living in proximity to each other, the four Kazak populations from Jetisuu exhibited considerable diversity in mtDNA lineages (Figs 2 and 3 and Table 1). While Albans had the highest number of West Eurasian lineages, Suans had the fewest. Jalayir and Shapirashti were positioned between Albans and Suans in the PCA plot. As indicated before, mtDNA haplogroup frequency distribution in Jetisuu Kazaks might have been influenced by sample size. Overall, no pronounced differences were observed in the phylogenetic affiliation of the four populations (SF-1).

As shown in Fig 3, East Eurasian haplogroups A, B, C, D, F, and G occurred in almost all populations in Central Asia, with D being the most frequent one. West Eurasian haplogroups (H, HV, V, J, T, U, W, and X) including several subbranches of R were also found in Turko-Mongolian speakers, albeit at relatively low frequencies. Haplogroups H and U were the most common West Eurasian haplogroup in Central Asia, and also occurred in the Jetisuu Kazaks at appreciable frequencies. In contrast, the West Eurasian haplogroups were almost absent in Sino-Tibetan speaking groups, but they harbored high frequencies of East Eurasian haplogroups M7 and M8.

In the PCA plot, Han Chinese and Dungan (Hui) from Xinjiang clustered together. Yet, the frequencies of the maternal lineages in Dungans considerably differed from those in Han Chinese. This result suggested that the maternal gene pool of Dungans did not exclusively derive from Han Chinese, even though Dungans supposedly shared a common maternal ancestry with them and speak Chinese. In Central Asia, intermarriages between Dungans and Turkic speakers are not uncommon because of shared religious beliefs. As an example, Dungans harbored West Eurasian maternal lineages such as U2, X2, T1, and HV [29], which are virtually absent in Han Chinese. The Han Chinese from Xinjiang also exhibited different maternal lineages compared to other Turko-Mongolian groups in Central Asia, indicating their recent arrival in Xinjiang. Despite the overall similarity of mtDNA diversity in Han Chinese and Tibetans, the latter clustered closer to Mongols and Kazaks.
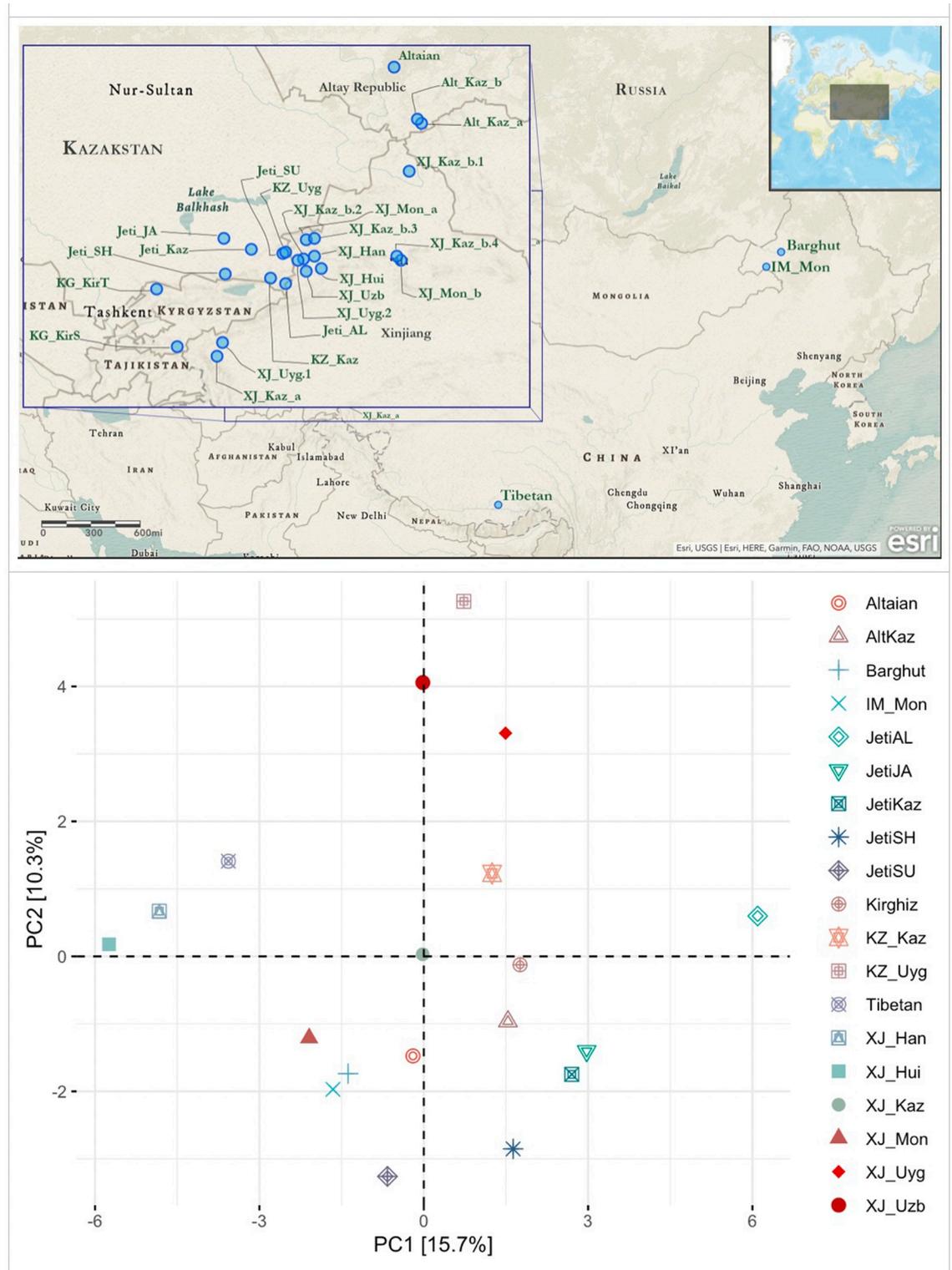
Fig 2. Geographic distribution of the human populations and the mtDNA control region haplogroup frequency based PCA.
Notes on Fig 2: **Jeti_AL** [Alban (N = 35)], **Jeti_JA** [Jalayir (N = 76)], **Jeti_SH** [Shapirashti (N = 59)], and **Jeti_SU** [Suan (N = 26] are
the four Kazak populations from Jetisuu, Kazakhstan; **Jeti_KZ** [Four populations from Jetisuu combined, plus 4 individuals with
ambiguous clan identity (N = 200)]; **Altaian** [Altaians (N = 490) from the Altay Republic, Russia] [25]; **AltKaz_a** [Kazaks (N = 237)
from the Altay Republic, Russia] [12]; **AltKaz_b** [Kazaks (N = 98) from the Altay Republic, Russia] [15]; **Barghut** [Barghuts (N = 149)

from Hulunbuir, Inner Mongolia, China][15]; **IM_Mon** [Mongols (N = 48) from Inner Mongolia, China] [26]; **KZ_Uyg** [Uyghurs (N = 55) from Penjim, Panfilov, Almati, Kazakhstan]**, KG_KirT** [Kirghiz (N = 48) from Talas, Kirghizstan]**, KG_KirS** [Kirghiz (N = 47) from Sari-Tash, Kirghizstan]**, KZ_Kaz** [Kazaks (N = 55) from Kegen, Almati, Kazakhstan] [13]; **Tibetan** [Tibetans (N = 6109) from across Tibet, China] [27]; **XJ_Hui** [Dungans (Hui) (N = 45)], **XJ_Mon_a** [Mongols (N = 49)], **XJ_Kaz_a** [Kazaks (N = 53)], **XJ_Uzb** [Uzbeks (N = 58)], **XJ_Uyg** [Uyghurs (N = 47)], and **XJ_Han** [Han Chinese (N = 47)] [All six populations from Xinjinag, China] [17]; **XJ_Kaz_b** [Kazaks (N = 151) from several locations in Xinjiang, China, including b.1: Altay, b.2: Ile (Kulja), b.3: Buratala, and b.4: Urumqi, Sanji, and Kumul] [16]; **XJ_Mon_b** [Mongols (N = 106) from Xinjiang, China, but exact sampling site is unknown] [28].

Mongols from Xinjiang (XJ_Mon_a and XJ_Mon_b) and Inner Mongolia (IM_Mon) cluster together with Barghuts who speak a Mongolian language. Although characterized predominantly by East Eurasian maternal lineages, Mongolian speakers also harbored certain West Eurasian haplotypes (e.g., U, H, and HV, as shown in Fig 3) [15] at minimal frequencies, most likely due to their links with Turkic speakers in Central Asia. Xinjiang Mongols further shared multiple maternal lineages with Kazaks from Xinjiang and Kazakhstan. This finding might be explained by their common Altaic ancestry, geographic proximity, similarity in subsistence patterns, and intermarriage during the periods of Jungar, Kalmak, and Kazak Khanates.

By contrast, Mongols from Inner Mongolia shared a certain number of maternal haplotypes with Kazaks. This finding may imply their descent from a common ancestral group at the time of Genghis Khan or even before. Among Turkic speakers from Central Asia, Uyghurs and Uzbeks exhibited elevated frequencies of West Eurasian lineages, indicating that a substantial number of maternal lineages may have been contributed by Indo-European speaking peoples to the gene pools of Uyghurs and Uzbeks, even though they are Turkic speakers.

Analysis of mtDNA variation

We subjected the mtDNA sequences to statistical analysis to evaluate patterns of diversity in Kazaks. Values of Tajima's D are significantly deviated from neutrality, implying that population expansion and/or natural selection might have played a role in shaping genetic diversity in Kazaks (Table 2). The disparity between the number of segregating sites and the average number of nucleotide differences was more pronounced when the number of sequences was higher. However, it seems that negative Tajima's D characterizes mitochondrial sequences (both control region and whole genome sequences) from many human populations [12, 30–32].

Comparative phylogenetic analysis of mtGenomes from Central Eurasia

Here we discuss the genetic features of maternal lineages present in Kazaks from Jetisuu in the context of central Eurasia, including Central Asia, Siberia, and Eastern Europe. We retrieved over 2000 full mtGenome sequences from GenBank [36], and used them to conduct a comparative analysis of the maternal genetic diversity in the four language families in Eurasia, namely, Altaic, Uralic, Sino-Tibetan, and Indo-European (Table 3).

Altaic speakers were the main focus of the analyses in the current study, since we aimed to understand their origins from the maternal perspective and evaluate the maternal gene pool of Kazaks in relation to those of other Altaic speakers. The mtDNA sequences of Slavs [37–39] and Tajiks [32] were included in the analysis in an effort to trace the origins of the West Eurasian maternal lineages (e.g., H, V, and U) found in Kazaks. We further included Hungarian sequences [40] to investigate whether Hungarians had any maternal genetic affinity with Kazaks and other Altaic speakers, since Hungarians have historical connections with Central Asia. Lastly, we analyzed mtGenomes of Tibetans [41] and Sherpas [42], because Tibetan
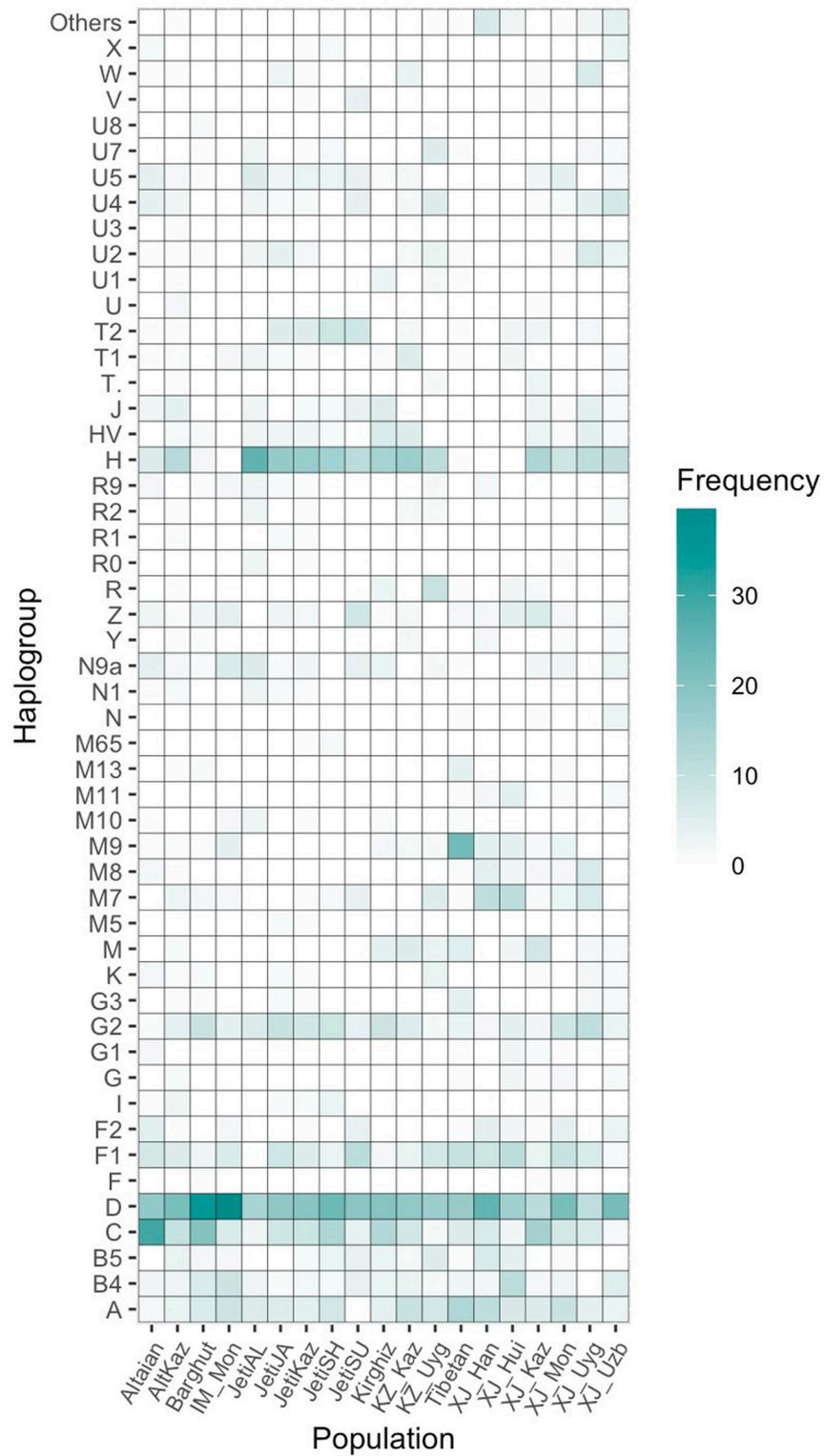
Fig 3. A heatmap for the mtDNA haplogroups and their relative frequencies. Notes on Fig 3: the heatmap was generated for the haplogroup frequency data that was used for PCA analysis in Fig 2.

https://doi.org/10.1371/journal.pone.0277771.g003

speakers have been neighbors of Altaic speakers for millennia and have interacted with them during various time periods of history.

In the current study, we were able to build high resolution of phylogenetic trees from the mtGenome sequences of the Jetisuu Kazak populations and evaluate their maternal genetic affinities with populations from Central Asia and other regions. Comparative analyses were carried out using 2074 full mtGenome sequences from 21 ethnic groups (Tables 3 and S4). The analyses were performed in three rounds: (1) all sequences (N = 2274), including those from Jetisuu Kazaks and other Central Eurasian populations, were analyzed together to identify their affinities and haplogroup affiliations; (2) Central Eurasian and Jetisuu Kazak sequences that belonged to the haplogroups A, B, C, D, F, G, H, T, and U, were selected and aligned together with sequences from the Phylotree Build 17.0 [46] (N = 1155; Tables 4 and S5 and S6). Subsequent analyses were separately performed for each of the haplogroups. Below, we present the results of the final round of phylogenetic analyses, focusing on each of the major haplogroups which characterize maternal genetic composition of the Jetisuu Kazaks.

In order to evaluate genetic distance among populations, we estimated pairwise F_{ST} values and performed PCA analysis with them (Fig 4). In the PCA plot shown in Fig 4, Kirghiz, Hungarians, and Tibetans were positioned at the tips of a triangular pattern, indicating they were genetically quite distant from each other, while other populations were scattered largely around the edges of the imaginary triangle. Altaic speakers clustered together in the top left quarter of the PCA plot (Fig 4), with Tubas and Tatars being exceptions. While Tubas were positioned close to Sherpas, Tatars separated from the other Turkic speaking populations and clustered together with Tajiks. An unexpected observation was the separation of Sherpas from

Table 3. Information on whole mitochondrial genome sequences which used for comparison.

Nr.	Language		Ethnicity	N	Reference
1	Altaic	Turkic	Altaian	35	Derenko et al. 2012; Derenko et al. 2010; Derenko et al. 2007;
2			AltKaz	11	Derenko et al. 2014;
			Kazak	7	Sahakyan et al. 2017; Ingman and Gyllensten 2007;
3			Kirghiz	125	Peng et al. 2018
4			Tatar	73	Malyarchuk et al. 2010
5			Tuba	33	Sukernik et al. 2012; Zhao et al. 2009; Volodko et al. 2008;
6			Uyghur	734	Peng et al. 2018; Zheng et al. 2017 [43]; Zhao et al. 2009
7			Uzbek	5	Sahakyan et al. 2017; Zhao et al. 2009; Ingman et al. 2000;
8		Mongolic	Barghut	14	Derenko et al. 2014; Derenko et al. 2012;
9			Buryat	231	Derenko et al. 2007; Derenko et al. 2010; Derenko et al. 2012; Derenko et al. 2014; Derenko et al. 2017/2018;
10			Kalmak	3	Palanichamy et al. 2015 [44]; Derenko et al. 2014; Derenko et al. 2010;
11			Mongol	7	Ingman & Gyllensten 2007; Sahakyan et al. 2017; Kong et al. 2003; Protasova et al. 2016 [45];
12	Uralic	Ugric	Hungarian	95	Malyarchuk et al. 2018
13	Sino-Tibetan	Tibetic	Tibetan	231	Zhao et al. 2009
14			Sherpa	76	Kang et al. 2013
15	Indo-European	Persian	Tajik	223	Peng et al. 2018
16		Slavic	Belrussian	7	Mielnik-Sikorska et al. 2013; Malyarchuk et al. 2008;
17			Czech	13	Mielnik-Sikorska et al. 2013; Malyarchuk et al. 2008;
18			Polish	30	Mielnik-Sikorska et al. 2013; Malyarchuk et al. 2008;
19			Slovak	22	Mielnik-Sikorska et al. 2013; Malyarchuk et al. 2008;
20			Russian	98	Mielnik-Sikorska et al. 2013; Malyarchuk et al. 2008; FTDNA;
21			Ukranian	1	Malyarchuk et al. 2008
Total number of mtG sequences (N):				2074	

https://doi.org/10.1371/journal.pone.0277771.t003

Table 4. Central Eurasian mtGenomes analyzed together with phylotree mtGenomes.

Count	Haplogroup	Jetisuu	Central Eurasia	Phylotree	Total N
1	A	9	53	26	**88**
2	B	4	13	7	**24**
3	C	17	68	22	**107**
4	D	37	112	57	**206**
5	F	12	42	22	**76**
6	G	17	54	28	**99**
7	H	28	126	116	**270**
8	T	13	60	38	**111**
9	U	15	123	36	**174**

Tibetans, with the former being closer to Turkic speakers, especially Tubas. Interestingly, Hungarians were clustered together with Russians in the bottom right quarter of the PCA plot.

Major East Eurasian Lineages

Haplogroup A. Haplogroup A is one of the founding haplogroups for Asia and the Americas. Haplogroup A was first identified by Schurr et al. [47] and further described by Torroni et al. [48] as one of the major maternal lineages of Native Americans. In our sample set from Jetisuu, we observed three major subbranches of A, namely A5, A12, and A14. The diverse branching pattern of haplogroup A suggested that it was one of the founding maternal lineages of Kazaks.

Haplogroup A5 contained mtDNA sequences from Buryats, Kazaks, Uyghurs, and Japanese (Fig 5), with the Japanese sequences AP010701 and AP008874 being identical to those in Uyghurs and Buryats. This result suggested that Japanese have some maternal genetic affinities with Turkic speakers (Kazaks and Uyghurs) and Mongolic speakers (Buryats), as suggested by some models of Altaic language family evolution, e.g., Trans-Eurasian language model [49].

Fig 4. Geographic distribution of the human populations included in the analysis of mitochondrial genomes from Central Eurasia and the mtGenome pairwise F$_{ST}$ value based PCA. Notes on Fig 4: Populations having more than 30 sequences were included in the PCA analysis, and thus the dataset contained 1954 sequences from previous studies (Table 3) and 120 Kazak mtGenome sequences from Jetisuu. Sample IDs of the sequences were given in S7 Table. The pseudo-mtGenomes (N = 80) were excluded from the PCA analysis. **Kazak** [Kazaks from Jetisuu, Kazakhstan, i.e., **Jeti_Kaz on the map**]; **Alt_Kaz** [Kazaks from the Altay Republic, Russia]; **Buryat** [Buryats from the Buryat Republic, Russia]; **Hungarian** [Hungarians from Hungary]; **Kirghiz** [Kirghiz from Xinjiang, China (**TJ_Kir**) and Kirghiz from Tajikistan (**TJ_Kir**)]; **Russian** [Russians from Russia]; **Sherpa** [Sherpas from Zhangmu, Tibet, China]; **Tajik** [Tajiks from Xinjiang, China (**XJ_Taj**) and Tajiks from Tajikistan (**TJ_Taj**)]; **Tatar** [Tatars from the Tatar Republic, Russia]; **Tibetan** [Ethnic Tibetans and other ethnic groups from Tibet, China (**Tib_1–5**)]; **Tuba** [Tubas from the Tuba/Tuva Republic, Russia]; **Uyghur** [Uyghurs from Xinjiang, China (Uyg_**1–2**)].
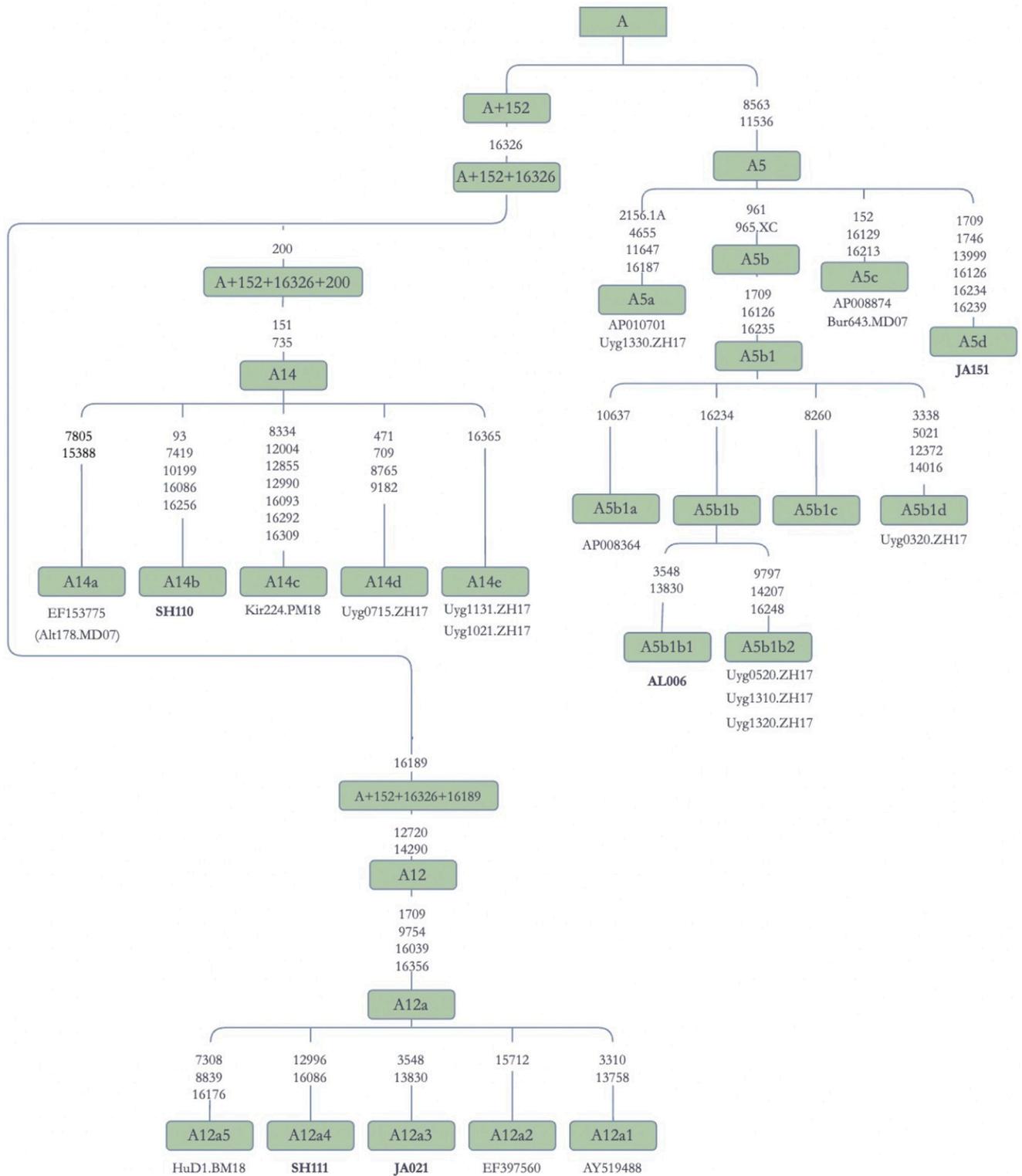
Fig 5. Schematic phylogenetic tree for the haplogroup A mitochondrial genomes. Note: Alt: Altaian; Uyg: Uyghur; Kir: Kirghiz; Hu: Hungarian.

https://doi.org/10.1371/journal.pone.0277771.g005

Kazaks, Uyghurs, and Japanese had additional mtDNA lineages representing multiple branches of A5, implying that the possibility of recent admixture was highly unlikely.

In the Phylotree Build 17.0 [46], A14 does not have any sub-branches. It contains only a single Altaian sequence (EF153775), which is included in our dataset (Alt178.MD07). Our phylogenetic analysis identified new subbranches of A14, which were formed by sequences from Turkic speaking Altaians (A14a), Kazaks (A14b), Kirghiz (A14c), and Uyghurs (A14d and A14e). This configuration suggested that these sequences might derive from a common Turkic maternal ancestor.

Two sequences from Jetisuu, along with a sister branch from Hungary, formed a distinct clade (A12a). The two Kazak sequences (JA021 and SH111) were named A12a3 and A12a4, respectively, whereas the Hungarian sequence (HuD1.BM18) was named A12a5. One of the two sequences representing A12a in the Phylotree Build 17.0 [46] came from an Evenk (EF397560), and the second from a Mansi (AY519488).

Haplogroup B. Haplogroup B is ubiquitous in Asia, Oceania, and Americas [12, 15, 25, 50–53]. It may have originated in Southeast Asia [50, 51], and is estimated to be 50 kyr old [54]. In support of this view, an mtDNA with the basal sequence motive of haplogroup B was found in a 40-kyr-old ancient human remain excavated from Tianyuan Cave near Beijing, China [55].

Two of its sub-branches, B4 and B5, were found in Jetisuu Kazak populations. In Fig 6, B4j contains three sequences, one each from a Kazak (B4j1), a Uyghur (B4j3), and a Buryat (B4j2), indicating that the haplogroup is shared among Turko-Mongolian speakers. With these data, we confirm the motif of B4j [*3548 4080 5300 6122 11893 11941 13911 14248 15172 16223 16362*] that was given a preliminary status in the Phylotree Build 17.0 [46].

A subclade of B4, B4b1a3, which was first described and named by Derenko and colleagues [15], was found in Altaians, Buryats, and Uyghurs (S4 Table). Found in Turko-Mongolian speakers, this haplotype is phylogenetically closest to the Native American B2 branch. B4c1b2b, a rare branch of B4 previously found in Chuvash people [15], has its sister branch (B4c1b2a) in Uyghurs (S4 Table), indicating the possible existence of the lineage in the Proto-Turkic stock population.

In our sample set, B5b was represented predominantly by sequences from Turko-Mongolian speakers (N = 13), including Turkic speaking Altaians, Kazaks, Kirghiz, and Uyghurs, and Mongolic speaking Buryats, except for a single sequence in a Tibetan. Two Kazak and one Uyghur sequences also did not belong to any previously known branches in the clade of B5b. Therefore, the Kazak sequences, SH017 and SH020, were named B5b7a and B5b7b, respectively, and the Uyghur sequence was named B5b6.

In the Phylotree Build 17.0 [46], we noted Japanese sequences which represented several sub-branches of B5b. For instance, Jetisuu Kazak sequence SU112 showed similarities with Japanese sequence AP008546 as shown in Fig 6. Although it is evident that Japanese mtDNAs show similarities to those in Turko-Mongolian speakers from Central Asia and Siberia, B5b seems to have a wide geographic distribution. B5b5 [KF540694] comes from Hakka, a Sinitic speaking native group from Taiwan [56], while B5b4 [KF849981] is (mostly likely) from Han Chinese [57].

Haplogroup C. Haplogroup C mtDNAs are found in human populations throughout Asia and the Americas [12, 15, 25, 51, 58]. In Jetisuu Kazak populations, seventeen individuals harbored haplogroup C mtDNAs, which represented 8.5% of the total sample size (Table 1). Except for C1, which is an American branch [59], all major sub-branches of haplogroup C appeared in the Jetisuu Kazak populations, including C4, C5, and C7. Most of the Jetisuu Kazak C4 sequences belonged to C4a, with one C4b and one C4d mtDNA also being present.
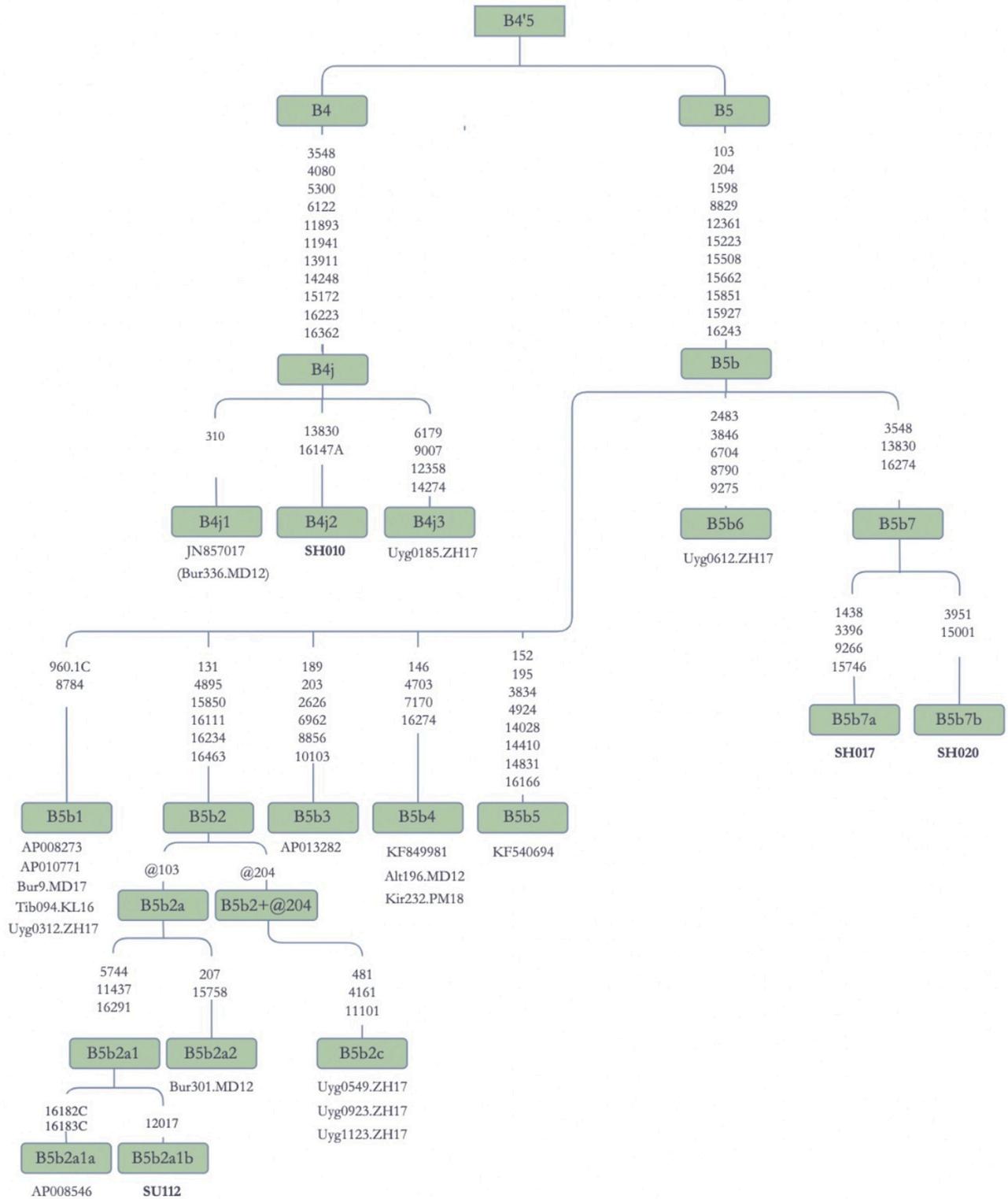
Fig 6. Schematic phylogenetic tree for the haplogroup B mitochondrial genomes. Note: Alt: Altaian; Bur: Buryat; Uyg: Uyghur; Kir: Kirghiz; Tib: Tibetan.

https://doi.org/10.1371/journal.pone.0277771.g006

A refined nomenclature was used for the subbranches of C4 after constructing a phylogenetic network for the mtGenome sequences.

C4a1a encompassed sequences from mainly Turko-Mongolian speakers, but also contained sequences from Indo-European speakers, such as Russians (Rus32), Poles (Pol422), and Tajiks (Taj126 and others), Uralic speaking Hungarians (HuS22), and Sino-Tibetan speakers (Tib034), indicating much deeper temporal origin and wide dispersal of the clade (S4 Table and Fig 7). Similarly, C4a1a4 contained four Jetisuu Kazak sequences in addition to multiple sequences from various ethnic groups, including Altaians, Buryats, Kirghiz, Tibetans, and Uyghurs. The five sub-branches of C4a1a4 were named as C4a1a4a through C4a1a4e, of which C4a1a4e was represented by a Jetisuu Kazak sequence (Fig 7). The tree topology suggested that C4a1a4 could also be one of the founding maternal lineages for Turko-Mongolians. In the Phylotree Build 17.0 [46], C4a1a4 is represented by just two sequences from Siberia, one from the Shors and another from the Buryats, the latter being included in our dataset.

Haplogroup C5 has four sub-branches, including C5a, C5b, C5c, and C5d, each of which also having their own subclades. There are both C5a and C5b sequences in our Jetisuu Kazak populations (Fig 8). C5a1 is represented by a single branch in the Phylotree Build 17.0 [46]. Two of our Kazak sequences, SU013 and KZ103, clustered within this clade, and were named C5a1c1 and C5a1c2, respectively. Altaian, Buryat, and Uyghur sequences also grouped within this cluster, with the Altaian sequence (Alt121) being defined as C5a1b in the Phylotree Build 17.0 [46].

C5b1 was represented by sequences from Kazaks, Uyghurs, Buryats, and Yakuts (Fig 8). C5b1a included one Kazak sequence (SH026), and several other sequences from Altaians, Buryats, and Kirghiz, while Kazak sequence (SH026) belonged to C5b1a3. Two Jetisuu Kazak sequences, SH108 and SH011, belonged to subclade C5b1b1, which has a single branch represented by two Yakut sequences (FJ951438 and KF148480) in the Phylotree Build 17.0 [46]. Its sister branch C5b1b2 appeared in Uyghurs. The Kazak and Uyghur sequences were more derived compared to the ones found in Yakuts and Buryats, as shown in Fig 8. The diagnostic mutation for C5b, an insertion at np 1719 (1719.1G), should be reconsidered, since all 15 mtGenomes in our data set possessed a back mutation at this position (@1719.1G).

In light of this evidence, haplogroup C5 can be considered as one of the founder lineages of Turko-Mongolian speakers, or Altaic speakers more generally. All but one of the sequences in C5b belong to Turko-Mongolian speakers, including Altaians, Buryats, Kazaks, Kirghiz, Uyghurs, and Yakuts. The remaining sequence (AY615359) appeared in Nganasans, who speak a Samoyedic language of the Uralic language family but have mixed with Tungusic speaking and other groups in the past [60].

Haplogroup D. Haplogroup D, first discovered in Native Americans [47, 48, 61], is a major East Eurasian lineage, one that occurs in Northeast and Southeast Asia, and Siberia but also the Americas [15, 26, 27, 31, 32, 41, 48, 50–52, 58]. In Jetisuu Kazaks, we found three branches of haplogroup D, namely D1, D4, and D5. These lineages account for 19.5% of the maternal genetic diversity in Jetisuu Kazaks, making it the most common haplogroup in this population.

Within D4m (Fig 9), we observed one Jetisuu Kazak sequence. Together with close relatives in Uyghurs and Kirghiz, the Kazak sequence formed a new subbranch (D4m3). As shown in Fig 9, D4m was represented by sequences from Turkic (Kazak, Kirghiz, and Uyghur), Mongolian (Buryat) and Tungusic (Even, KF148265) speakers, along with Japanese (AP008432), implying a common Altaic origin for them.

Most of the D5 sequences analyzed in this study belonged to D5a2a1 (Fig 10). The branching patterns of the D5 clades were complex, and the branches are formed by the sequences from Turko-Mongolian speakers, Tibetans, and Japanese.

Fig 7. Schematic phylogenetic tree for the haplogroup C4a1 mitochondrial genomes. Note: Alt: Altaian; Bur: Buryat; Uyg: Uyghur; Kir: Kirghiz; Tib: Tibetan.

https://doi.org/10.1371/journal.pone.0277771.g007
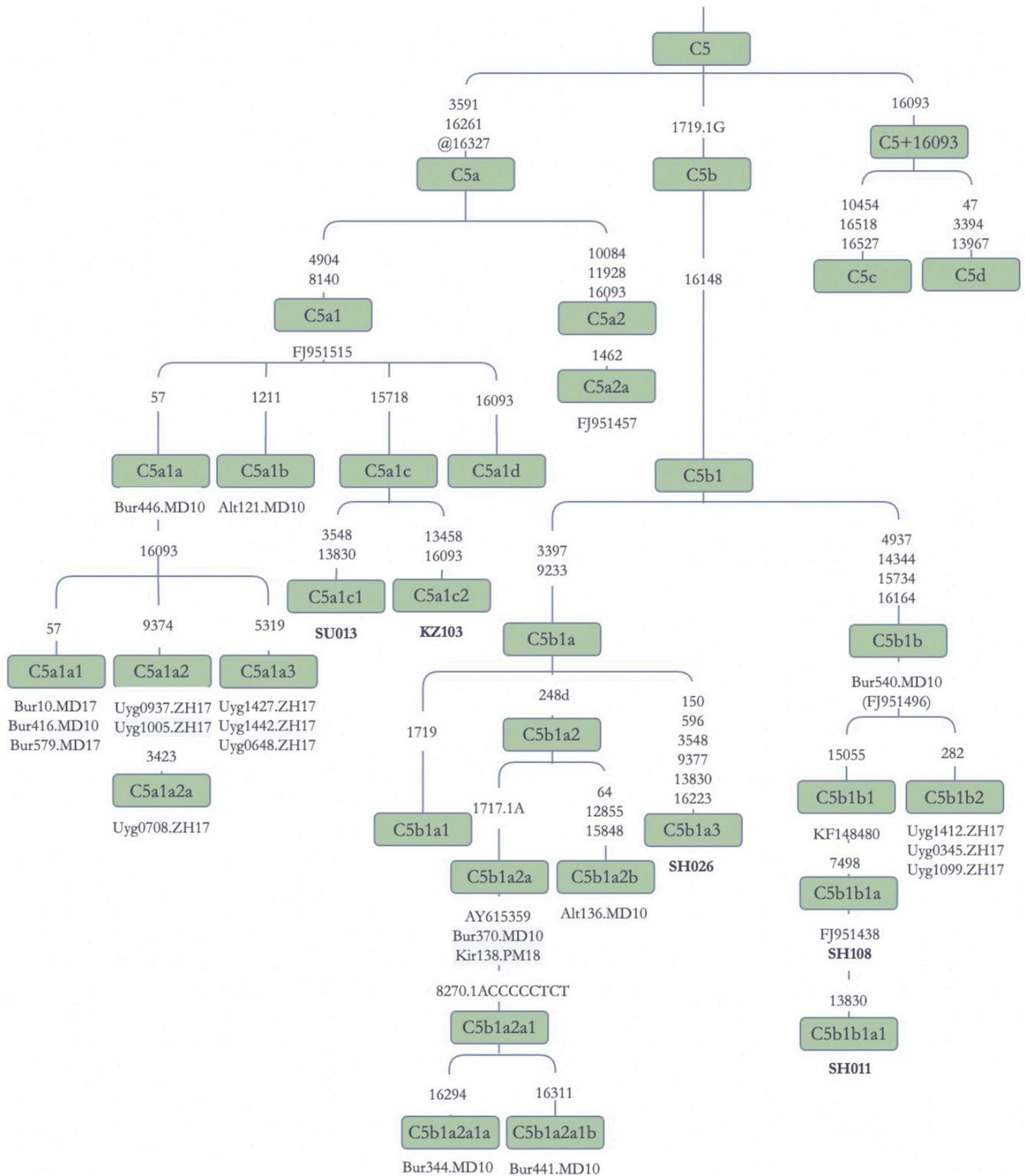
Fig 8. Schematic phylogenetic tree for the haplogroup C5 mitochondrial genomes. Note: Alt: Altaian; Bur: Buryat; Uyg: Uyghur; Kir: Kirghiz.

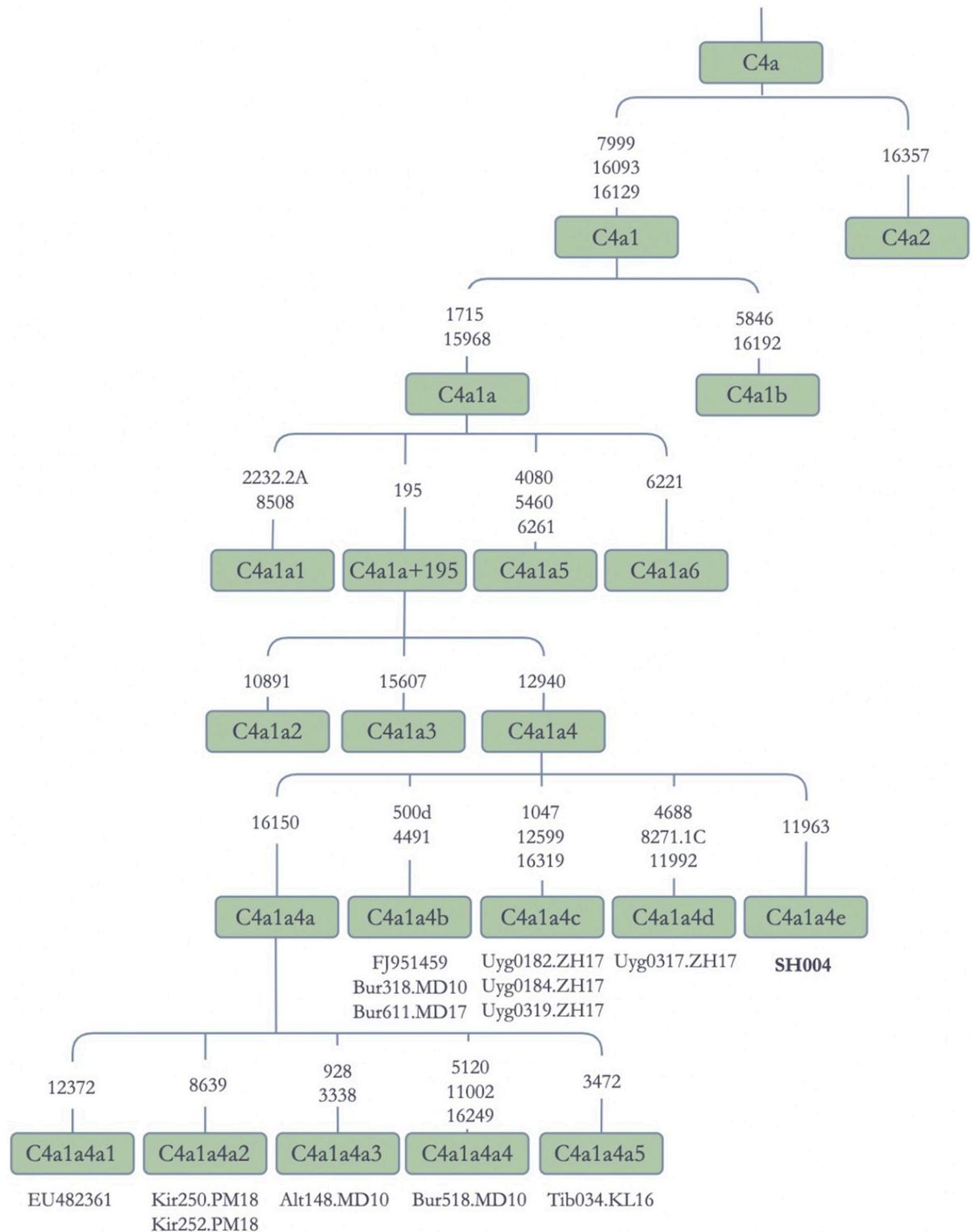https://doi.org/10.1371/journal.pone.0277771.g008
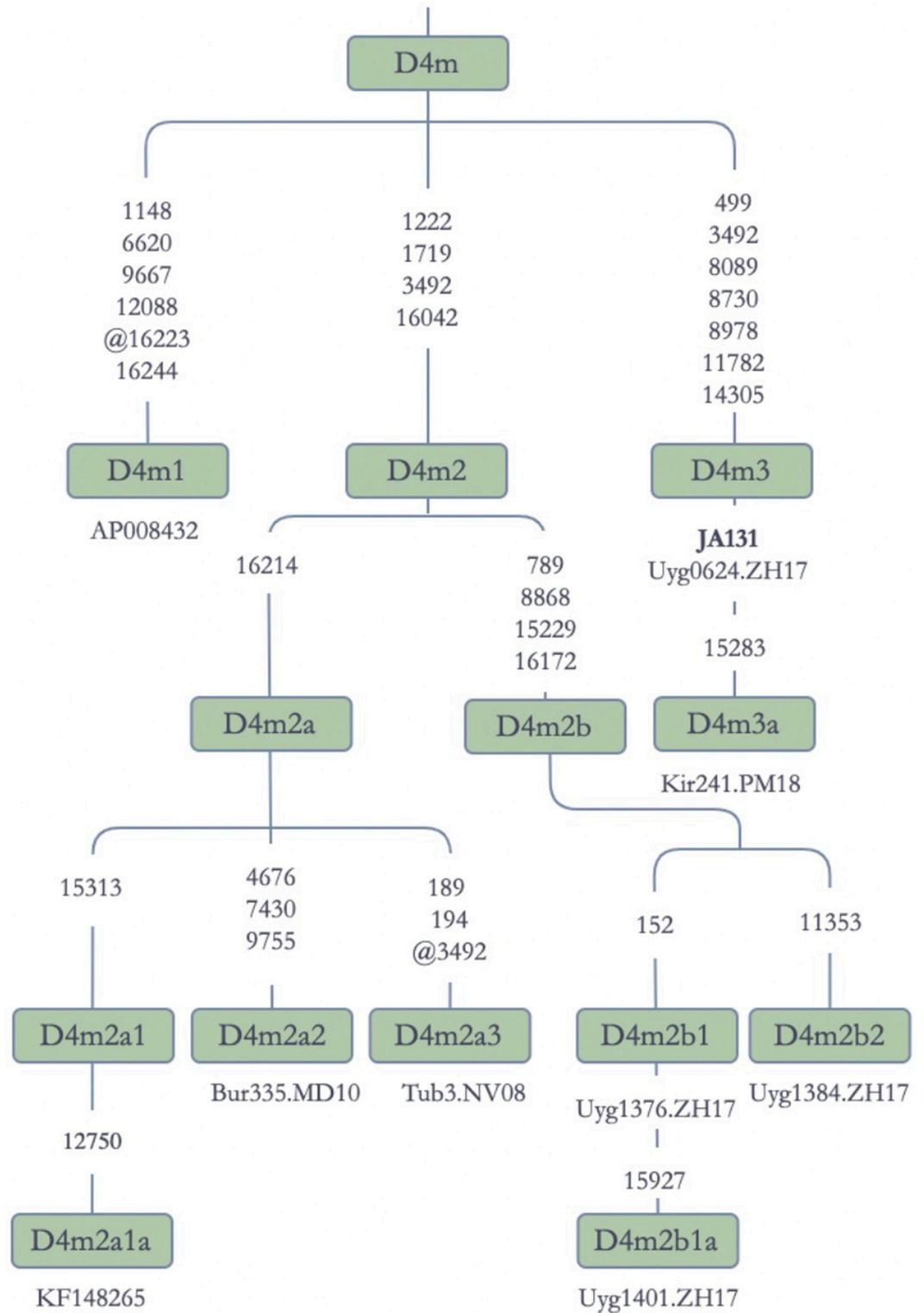
Fig 9. Schematic phylogenetic tree for the haplogroup D4m mitochondrial genomes. Note: Alt: Altaian; Bur: Buryat; Uyg: Uyghur; Kir: Kirghiz; Tib: Tibetan; Tub: Tuba/Tuva;. KF148265 (Even) and AP008432 (Japanese) were from the Phylotree Build 17.0 [46].

https://doi.org/10.1371/journal.pone.0277771.g009
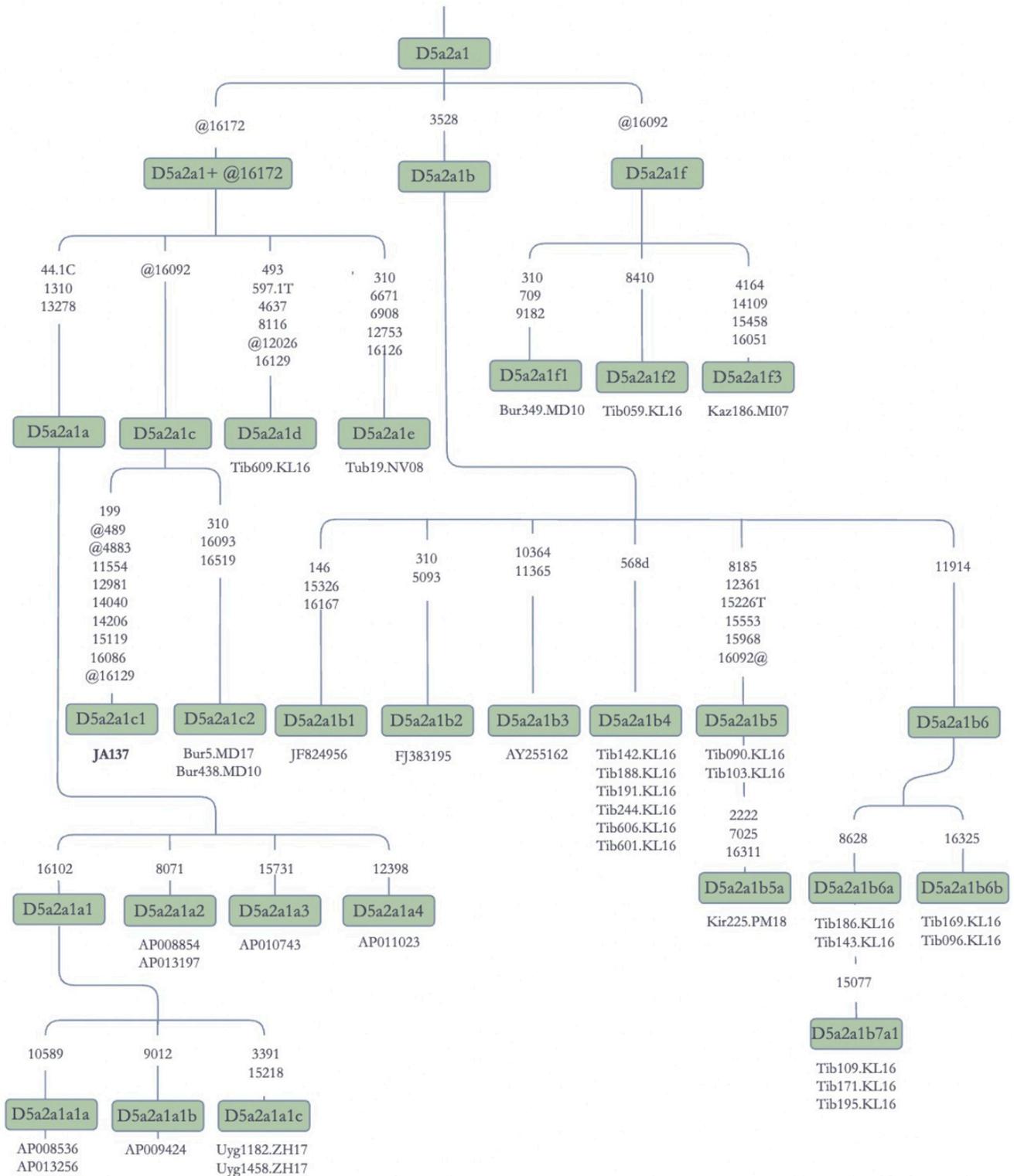
Fig 10. Schematic phylogenetic tree for the haplotype D5a2a1. Note: Bur: Buryat; Kaz: Kazak; Kir: Kirghiz; Tib: Tibetan; Uyg: Uyghur; The following sequences were obtained from the Phylotree Build 17.0 [46]: AP013256, AP008854, AP013197, AP010743, AP011023, AP008536, AP013256, and AP009424 are Japanese sequences; JF824956 is from China, but unknown ethnic origin; FJ383195 is from India; AY255162 is a Han Chinese sequence.

https://doi.org/10.1371/journal.pone.0277771.g010

The D5a2a1+@16172 lineage contained sequences from Buryats, Kazaks, Tibetans, Tuba, and Uyghurs. This clade had four subbranches, which included three new ones (D5a2a1c, D5a2a1d, and D5a2a1e), in addition to one (D5a2a1a) already recorded in the Phylotree Build 17.0 [46]. We also propose the new branch of D5a2a1 defined by a back mutation at 16092, but not the back mutation at 16172, which occurs in D5a2a1f (Fig 10). The newly named clade has three subbranches formed by Buryat, Kazak, and Tibetan sequences, respectively, and implying that it has a wide geographic area of distribution.

Other haplogroups of significance

Here, we discuss the mitochondrial haplogroups F, G, H, T, and U, which are found relatively high frequencies in Jetisuu Kazaks. Haplogroups F and G are common in populations from Siberia and Central Asia [12, 15], but they occur at considerably higher frequencies in East Asia [51]. While the West Eurasian haplogroups H and T are also found in Central Asia and Siberia, their frequencies decrease in the eastward direction.

Haplogroup F is distributed in central, eastern, southern regions of Asia [12, 15, 25, 50, 51, 62, 63]. F1b1f occurs in populations that belong to Turkic (Kazak and Uyghur), Mongolic (Buryat), and Tungusic (Evenks) branches of the Altaic language family. While most subbranches of F1b1 are found in Turko-Mongolian speakers, including Buryats, Kazaks, Kirghiz, and Uyghurs, some sublineages also occur in Japanese. Similarly, F1a1 mtDNAs are found in Turko-Mongolian speakers, Japanese, and Tibetans (S4 Table), but also occur in populations in southeast Asia (i.e. Cambodia and the Orchid Island of Taiwan) [46]. Haplogroup G, which was defined by Schurr et al. [64], also reflects maternal genetic connection between Turko-Mongolian speakers and Japanese. For instance, in the clade of G2a5, Kazak and Buryat sequences are derived from a Japanese sequence (AP008897). In the same way, multiple Turko-Mongolian sequences cluster together with Japanese sequences in the clades of G2b2 and G3a2 (S1 Fig).

In our dataset, haplogroup H lineages comprised about 17% of the maternal linages in Jetisuu Kazaks, with H2 being the most common haplogroup (6.5%, Table 1). While Turkic speaking Tatars shared multiple West Eurasian haplotypes with Russians and other Slavic populations, most of the West Eurasian haplotypes found in Kazaks and other Turko-Mongolian speakers were distantly related to those occurring in the Slavic groups. In addition, some West Eurasian haplotypes exclusively occur in Turko-Mongolian speakers (e.g., H8b1). These findings strongly support the notion that the existence of West Eurasian haplotypes in modern Turko-Mongolian speakers are not simply due to Russian expansions in the recent centuries. For instance, in S4 Table, the subclade of H8b1 is comprised of Buryat, Kazak, Kirghiz, and Tuba sequences. In addition, H8b1 also includes a sequence (KF148188) from a Tungusic speaking Evenk.

It is interesting to note that, in the phylogeny of haplogroup T, Tatar sequences seemed to be ancestral to subbranches appearing in Central Asia. As shown in Fig 11, Tat42.MB10 is ancestral to T2b lineages found in Buryats, Kazaks, Kirghiz, Russians, and Hungarians. Another Tatar sequence (Tat7.BM10) is ancestral to T1a1 subbranches (S1 Text). This finding may reflect the geographic origin of haplogroup T in Central Asia and Siberia. In support of this view, T1a and T2b mtDNAs were discovered in archaeological samples of Yamna Culture sites [65] in Volga region, Russia, whence the Tatars (Tat7 and Tat42) were sampled [66].

Excluding U6 and U9, all subclades of U (U1-5 and U7-8) occurred in Turko-Mongolian speakers in Central Asia (S4 Table). While U7a was found mostly in Turko-Mongolian speakers and in the Iranian speaking Tajiks, U7b, previously believed to be absent in Central Asia [21], was discovered only in Uyghurs (S4 Table). While U4a seemed to characterize Slavic
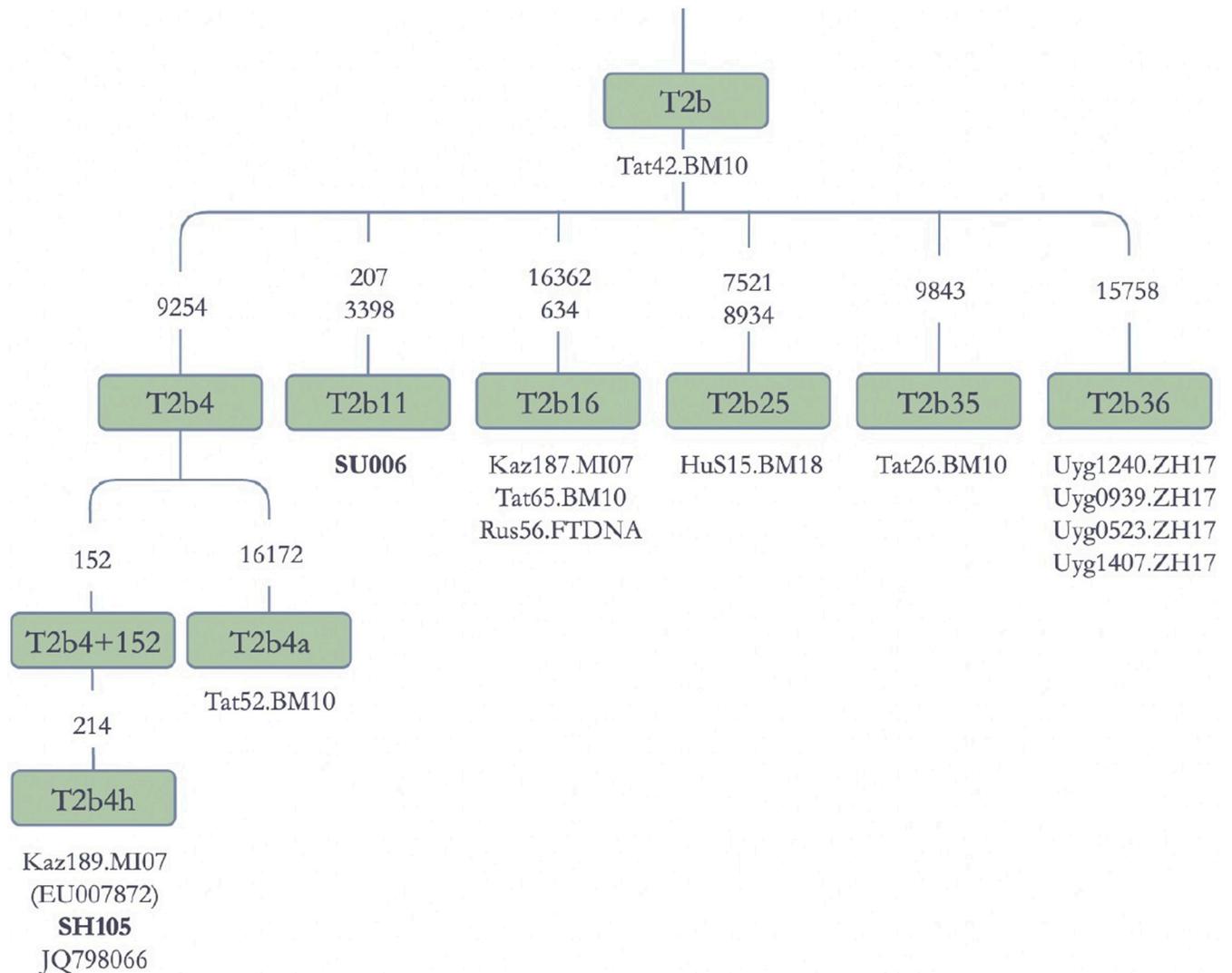
Fig 11. Schematic phylogenetic tree for the haplotype T2b. Note: Kaz: Kazak; Hu: Hungarian; Rus: Russian; Tat: Tatar; Uyg: Uyghur; **SU006** and **SH105** are Jetisuu sequences; EU007872 and JQ798066 are sequences from the Phylotree Build 17.0 [46].

speakers, all U8 sequences in our dataset were from Turko-Mongolians, with the exception of one Hungarian sequence (S4 Table).

Discussion

We characterized maternal lineages of Kazak populations in Jetisuu at the mtGenome level. To our knowledge, this is the first mtGenome study of human populations in Kazakhstan. Notwithstanding the small size and uniparental inheritance of mtDNA, it has provided important insights into the modern human evolution [67, 68]. The mtGenome data produced in this study will therefore be invaluable for reconstructing human population history and can serve as a reference for identification of pathogenic mutations in patients with mitochondrial disorders in Central Asians.

In previously published studies, the mitochondrial genetic profiles of Kazaks and other ethnic groups in Central Asia and Siberia were characterized through the analyses of mtDNA

control region sequences and coding region SNPs [12, 14, 19, 25, 47, 61, 69, 70] and complete mtDNA sequencing [15, 20, 63, 71]. A mtDNA study of Kazaks from the Altay-Sayan region [12] revealed that they exhibit highly diverse maternal genetic profiles, harboring both West and East Eurasian maternal lineages. Our results confirmed the same pattern of maternal genetic variations in Jetisuu Kazaks, with almost all 200 individuals surveyed having distinct haplotypes (S2 and S3 Tables).

West Eurasian lineages could have become part of the maternal gene pool of Kazaks during Turko-Mongolian expansions (i.e., Turkic Khanates and Golden Horde) and/or Slavic expansions during Russian Empire and the Soviet Union, or perhaps even deeper in time, due prehistoric expansion of human populations. The comparative analyses of mtGenomes from multiple ethnic groups in Central Eurasia indicate that West Eurasian haplogroups (e.g. H, T, and U) have diverse subbranches in Kazaks and other ethnic groups in Central Asia and Siberia, with most of the subbranches being distinct from those haplotypes found in Russians and other Slavic populations. This observation indicates that Slavic expansions may not fully account for the occurrence of West Eurasian maternal lineages in Central Asia. For instance, multiple subbranches of U, including U2, U4, U5, and U7, are found in Jetisuu Kazaks, and the haplogroup U has the largest TMRCA (~ 40,000 years) among the haplogroups found in Kazaks (Table 1 and Table 2 in S1 Text). The mtDNA lineage of the 24-thousand-year-old Malta individual also belongs to basal U [72], suggesting a deep origin for this maternal lineage in Central Eurasia. Recent ancient DNA studies also found West Eurasian lineages in Iron Age populations or earlier inhabitants of the Altay Mountains [5]. Interestingly, Jetisuu Kazaks harbor the same set of West Eurasian lineages (HV, N1, J, T, U, K, W, I, and X) found in the Iron Age populations of Central Asia [5].

It can be deduced from the frequencies of these ancestral lineages in various ethnic groups in Central Asia that, the arrival of East Eurasian lineages (e.g., Haplogroups A, B, C, and D) in southern Central Asia, i.e. the Pamir Plateau, could possibly be attributed to Turko-Mongolian expansions. Examples of these expansions could include early movements of Turks and Mongols, then southward movement of Mughals, and later the expansion of the Jungarian Khanate [1]. The subbranches of haplogroups A, B, C, and D found in Uyghurs from Xinjiang and Tajiks from the Pamir are closely related to those of Altaians, Buryats, Kirghiz, and Kazaks (S4 Table). However, such a hypothesis may require further lines of evidence, given the fact that haplogroup C was found among Tarim Basin mummies [73, 74] and haplogroup D was found in an Iron Age population in Xinjiang [75]. These ancient populations predate the Turko-Mongolian expansions. In addition, haplogroup C7 occurs in northeast India [76], most likely due to prehistoric migrations.

East Eurasian haplogroups found in eastern Europe are associated with those found in Turko-Mongolians as well. A previous study tried to demonstrate paternal genetic affinities between Hungarian and Kazak populations [77], but the coalescence time of the Y-chromosome haplogroups G1 and G2 predated the historical events linking Hungarians to Central Asians. In addition, the arrival of haplogroup G1 in the Kazak steppes may date to recent historical time periods, as the age of this lineage in Kazaks was estimated to be ~ 600 years old [78]. On the other hand, it appears that maternal lineages A12a, C4a1a, and U8b1b may link Hungarians with Kazaks and other Turko-Mongolian speakers (Fig 5 and S4 Table).

In the F_{ST}-based PCA analysis, Sherpa population unexpectedly exhibited a closer affinity with Turko-Mongolians than with Tibetans (Fig 4). Our analysis indicates that certain mtDNA lineages (D4j13 and D4j1a1) and some other haplotypes clearly connect Sherpas in Tibet (and Nepal) and Turko-Mongolians in Central Asia and Siberia. D4j13 occurs in Sherpas and Turkic speaking Kirghiz, Tatars, and Uyghurs, while D4j1a1 appears in Sherpas and Tibetans, as well as Mongolic speaking Buryats and Turkic speaking Kirghiz and Uyghurs (S4 Table). The similar

genetic affinities can be observed between Sherpas and Turko-Mongolian speakers through haplotypes D5a2a1, C4a2, G3a, M5, M9a1a1c1, M11, M13, and the West Eurasian lineages H10, U2, and W. For instance, a Sherpa individual (Shr035.KL13) shares the basal W with a Uyghur (S4 Table). The existence of Y-chromosome haplogroups Q and R in Sherpas [79] may further suggest their connections with Turkic peoples through paternal lineages as well. However, the exact nature of the genetic affinity needs to be further evaluated by Y-chromosome and whole genome studies.

It has been noted by previous studies that Tibetans show genetic affinities with Japanese through their paternal [80] and maternal [51] lineages. In the current study, we identified several mitochondrial lineages that connect Kazaks and other ethnic groups from Central Asia with both Tibetans and Japanese. Haplogroup M10 occurs in Tibetans, but it is also found in Turkic speaking Altaians, Kazaks (e.g., AL104), and Uyghurs. Their shared maternal genetic ancestry may explain most of the maternal genetic connections between populations from the Tibetan Plateau and Turko-Mongolian groups.

This pattern of genetic diversity may have been partly shaped by recent genetic admixture. Within haplogroup A15, Uyghur sequences clustered together with those from Tibetans and Sherpas, revealing gene flow between Tibetans and Uyghurs despite their dissimilarities in language and religion. This genetic affinity could be attributed to geographic proximity, Tibetan Empire, or once flourished Buddhism in Xinjiang. An analogous genetic link was observed between Kirghiz and Tibetans in the clade of A17 (S4 Table). The directionality of gene flow, as inferred from the branching pattern, was from Tibetans into Kirghiz.

Some clades of mtDNA (D4a3b, and D4b1, D4m, and M7) present in our dataset link Japanese with Turko-Mongolian speakers. It is known that Ryukyuans from Japan harbor a high frequency of M7 mtDNAs [51]. All three subclades of M7, namely M7a, M7b, and M7c, are found in Turko-Mongolian speakers (S4 Table), including Kazaks from Jetisuu (S2 and S3 Tables). In addition, N9a mtDNAs are present in Turko-Mongolian speakers (e.g., N9a1, N9a3, N9a8, and N9a9) and occur at a considerably high frequency (2.5%) in Jetisuu Kazaks, while its sister branch N9b occurs at elevated frequencies in Ainu and Ryukyuans from Japan [51]. N9a mtDNAs are mainly found in continental Asia, including Japan [15, 31, 51, 81]. Haplogroup Y1, a subclade of N9, is also present in Buryats and Uyghurs (S4 Table), and occurs at a high frequency in Ainu people from Japan [51]. It was hypothesized that N9a might have reached East Asia through northern route after the dispersal of macrohaplogroup N from Africa [51]. Kazaks and other ethnic groups harbor several basal N and R sequences (S2–S4 Tables), and this may reflect that the basal branches of macrohaplogroup N have existed in Central Eurasia for a very long time. The discovery of a basal R mtDNA from a 45-kyr-old individual in Siberia (Ust-Ishim) [82] and a basal N mtDNA from a 34-kyr-old individual from Mongolia (Salkhit) [83] further supports the notion that some subbranches of N expanded towards East Asia through a northern route.

We found a limited number of haplogroups that exclusively characterize Turkic speakers in our dataset (e.g., A14, B4c1b2, and C4d). Interestingly, haplogroups HV1a1a, K1a17, and M13a1b1 are only found in Mongolic speaking Buryats and Barghuts. Many haplogroups, such as B4j, B4b1a3, C4a1a4, C4a2, C4b6, C5b, D4j5b, and D4j8 to name a few, are shared among Turkic and Mongolic speakers, i.e., Turko-Mongolians. Furthermore, a few other haplogroups, such as A5, B5b, D4b2b, D5a2a1, F1a1, F1b1, G2a5, G2b2, and G3a2, are found in human populations from a wide geographic area spanning from the Japanese archipelago in the East and the Kazak steppes in the West.

Some SNPs identified in certain pseudo-mtGenome sequences of Jetisuu Kazaks may potentially lead to identification of novel mtDNA lineages in haplogroup D. These may include coding region SNPs at nucleotide positions (nt) 3548 and 13830, and a control region

back mutation at nt 16223 in Jetisuu Kazak mtDNA (e.g., JA005, SU009, SH012, and SH007), although we could not rule out the possibility of incorrect SNP calls on the GenoChip microarray. In addition, SNPs at nt 2396 and nt 16519 in the pseudo-mtGenomes seem to be ambiguous, as well. Therefore, the authenticity of these SNPs needs to be evaluated and confirmed by future DNA analysis. In addition to these variants, Haplogrep reported multiple back mutations for some mtGenome sequences (AL108, AL109, JA116, and SH119 etc.), suggesting that such mtGenome sequences may require additional evaluation.

Mass migration, demographic fluctuations, and genetic drift can hinder the ability to accurately pinpointing the geographic origins and TMRCA of mitochondrial lineages. Thus, to confirm our current findings regarding the TMRCA and demographic history of Jetisuu Kazak maternal lineages (S1 Text), additional analyses of the West and East Eurasian lineages found in contemporary Kazaks and other ethnic groups in Central Asia, along with an expanded analysis of ancient DNA samples, should be undertaken.

Materials and methods

Fieldwork and DNA preparation

Blood samples were collected from ethnic Kazak individuals during several fieldwork sessions conducted in Almati Oblast, Kazakhstan, in summer 2014 (S1 Table). All blood samples were drawn via vacutainer (BD) by medical nurses. Written informed consent was obtained from the research participants. Prior to taking bloods, all of the research participants signed an informed consent form and filled out a genealogical questionnaire. The informed consent form and questionnaire were reviewed and approved by the Ethics Committee at the Center for Life Sciences (CLS), National Laboratory Astana (NLA), Nazarbayev University (NU). After fieldwork, blood samples were stored at -86˚C, and DNAs were extracted with a salt extraction method [84] with modification, or a commercial DNA extraction kit (e.g., Qiagen) at CLS, NLA, NU.

Microarray genotyping

The genomic DNAs of 80 individuals were genotyped via the GenoChip 2.0 microarray (Illumina iSelect HD custom genotyping bead array) [85] under the framework of the of Geno-graphic Project. The GenoChip 2.0 surveys about 750,000 SNPs from autosomes, sex chromosomes, and the mtDNA. Accordingly, in the current study, we analyzed about 3153 mitochondrial SNPs located across mitochondrial genome, i.e., from both control and coding regions of the mtDNA. Among the SNPs genotyped, about 400 of them were different from the revised Cambridge Reference Sequence (rCRS) [86], as shown in Table 2.

Pseudo-mitochondrial genome reconstruction

Pseudo-mitochondrial genomes (pseudo-mtGenomes) were reconstructed from the SNP genotype data for 80 Kazak individuals from Jetisuu via a Python script modified from GitHub [87] (S2 Table). For the reconstruction of pseudo-mtGenomes, rCRS [86] was used as a template, and nucleotides at certain positions of rCRS were replaced with SNP calls from the GenoChip.

Mitochondrial genome sequencing

In addition to the 80 genotyped samples, we sequenced the whole mitochondrial genomes of 120 Kazak individuals from Jetisuu. Whole mtGenome sequencing was performed on the NGS platform of Iron Torrent at the National Institute of Genomic Medicine, Mexico City, Mexico.

Sequence coverages of the newly generated mtGenomes were evaluated by QUALIMAP v2.2.1 after converting the raw sequence files into BAM files [88]. The average sequence coverage of the mtGenomes was 67.5x.

Sequence alignment

mtGenomes from Jetisuu. In the initial analysis, the pseudo-mtGenomes (N = 80) were aligned together with the genuine mitochondrial genomes (N = 120) using the online version of MAFFT v7.0 [89, 90]. rCRS [86] and the Reconstructed Sapiens Reference Sequence (RSRS) [91] were also included in the multiple sequence alignment (MSA). The aligned mtGenomes were further processed by the program of SNP-Sites [92] to generate an MSA VCF file. In the final stage of data analysis, BWA [93] was used for the MSA of both pseudo and genuine mtGenome sequences from Jetisuu (N = 200), for the sake of accuracy and consistency with the subsequent analyses.

mtGenomes from Central Eurasia. The Jetisuu Kazak mtGenome sequences (N = 200) were aligned together with 2074 mtGenomes from multiple ethnic groups across Central Eurasia (Tables 3 and S4) via BWA [93]. This step produced a SAM format alignment of 2274 mtGenome sequences. Subsequent file format conversions were performed through SAMtools [94], BAMtools [95], and BCFtools [96, 97].

Selection of mtGenomes for phylogenetic analysis. From the first round MSA of mtGenomes from Central Eurasia (N = 2274), subsets were selected based on haplogroup identities and their affinity with Jetisuu Kazak sequences. Thereafter, an MSA alignment was performed for each of the macro-haplogroups (A, B, C, D, F, G, H, T, and U) (Table 4), with sequences from the Phylotree Build 17.0 [46] representing major subbranches of the haplogroups also being added (Tables 4 and S5 and S6). All alignments were carried out via BWA [93], and file format conversions were performed through SAMtools [94], BAMtools [95], and BCFtools [96, 97].

Extended Bayesian skyline plot (EBSP)

Jetisuu mtDNA sequences (N = 200) were first partitioned into HVS-I, HVS-II, and coding regions to estimate the demographic history of Jetisuu Kazaks using the EBSP method (S1 Text). The coding region was divided again based on codon start positions. Different mutation rates were used to estimate the coalescence times based on the coding and control regions of mtDNA. Two mutation rates, 31.43×10^{-8} μ/site/year for the control region [98] and 1.71×10^{-8} μ/site/year for the coding region [54] were used (Table 1 in S1 Text). EBSP was performed on BEAST 2.0 [99] with the Markov Chain Monte Carlo (MCMC) chain length of 30,000,000 and sampling every 3000 steps. Initial 3,000,000 iterations were considered pre-burn-in and excluded. EBSP results were reviewed with Tracer v1.7 [100] and plotted via R [101].

Identification of mitochondrial haplogroups

Mitochondrial haplogroups were identified by analyzing either VCF or FASTA file via the online version of Haplogrep 2.0 [23]. In parallel, mitochondrial haplogroups were also obtained by analyzing FASTA format sequence files on mtPhyl 5.003 [24], which does not require MSA. No discrepancy was observed between Geno 2.0 SNP based haplogroup calls and the pseudo-mtGenome based haplogroup calls.

PCA using mtDNA control region haplogroup frequencies

PCA analysis (Fig 2) was performed using mtDNA haplogroup frequencies for several populations in Central Asia, together with the four Kazak populations from Jetisuu. Other

populations included Altaians from the Altay Republic, Russia [25]; Kazaks and Uyghurs from Kazakhstan, Kirghiz from Talas and Sari-Tash, Kyrgyzstan [13]; Mongols from Inner Mongolia, China [26]; Dungans (Hui), Mongols, Kazaks, Uyghurs, and Han Chinese from Xinjiang, China [16, 17]; Kazaks from the Altay Republic, Russia [12, 15]; Barghuts from Inner Mongolia, China [15]. Most of the comparative studies identified mtDNA haplogroups based on control region sequences, i.e., HVS-I and HVS-II, and coding region SNPs. Therefore, mtDNA haplogroup frequencies of Jetisuu Kazaks used in this PCA plot are mostly from the frequencies of basal branches of the mitochondrial phylogeny rather than terminal haplotypes. The prcomp function of base R [101] was used for the PCA analysis, and the ggplot2 [102] and factoextra packages of R [101] were used to plot PCA results.

A heatmap was created using the R packages reshape [103] and ggplot2 [102] for the haplogroup frequency data used for the PCA.

mtGenome pairwise F_{ST} value based PCA

Pairwise F_{ST} values were calculated with pairwise_Gst_Nei function of the mmod package of R [104]. PCA was performed using the prcomp function of base R [101]. PCA plot was made by the R packages factoextra and ggplot2 [102].

Phylogenetic analysis

Phylogenetic networks were constructed from mtGenome sequences with Network 10.1.0. [105]. Phylogenetic trees were also obtained through Haplogrep 2.0 [23] and mtPhyl 5.003 [24]. New branches and twigs, which are not available in the current version of Phylotree (v17.0) [46] and cannot be identified by Haplogrep 2.0, were reconstructed and named based on the comparative analyses of the phylogenetic trees from Haplogrep 2.0 and mtPhyl 5.003, as well as the phylogenetic networks constructed by Network v10.1.0. Schematic phylogenetic trees were drawn manually using the draw.io program [106]. mtGenome and pseudo-mtGenome sequences with ambiguous SNPs were excluded from the schematic phylogenetic trees. Jetisuu Kazak sequences were highlighted in bold. Ethnicity was used as prefix and the initials of the first author followed by the year of publication were used as suffix for the sample ID of the sequences used for comparative analysis. The mtDNA sequences from Phylotree [46] were named after their GenBank accession number, as given in the Phylotree Build 17.0 [46]. If a Phylotree sequence was also included in our initial dataset, it is shown in parenthesis in the schematic phylogenetic trees.

Coalescence time estimation

Coalescence times (TMRCA) were estimated via the method provided by Soares and colleagues [54] with a whole mtGenome mutation rate of one mutation per 3624 years (S1 Text). Rho and sigma values were obtained on Network v10.1.0 [105].

Supporting information

S1 Text. Additional analysis of mitochondrial sequences.
(PDF)

S1 Fig. A phylogenetic tree of Jetisuu mitochondrial DNA (N = 200). The phylogenetic tree of the Jetisuu Kazak mtDNA was constructed via Haplogrep 2.0 [23].
(TIFF)

S1 Table. Number of individuals surveyed from each of the four clans and the sampling sites.
(XLSX)

S2 Table. Information on the Jetisuu samples genotyped via the GenoChip microarray (N = 80).
(XLSX)

S3 Table. Information on the Jetisuu samples subjected to whole mitochondrial genome sequencing (N = 120).
(XLSX)

S4 Table. Mitochondrial genomes (N = 2074) from the Central Eurasia for comparative analyses.
(XLSX)

S5 Table. Mitochondrial genomes from phylotree (N = 369) for comparative analyses.
(XLSX)

S6 Table. Sample ID List of the mitochondrial genomes (N = 1155) used for the second round mtGenome analysis. Ethnicity and GenBank ID information of the samples can be found in S2–S4 Tables.
(XLSX)

S7 Table. Sample ID list of mitochondrial genomes from the Central Eurasia (N = 1954) and Jetisuu (N = 120, excluding pseudo-mtGenomes) that were used for the mtGenome F_{ST} value based PCA analysis. Ethnicity and GenBank ID information of the samples can be found in S3 and S4 Tables.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Ayken Askapuli, Lorena Orozco.

Data curation: Miguel Vilar, Humberto Garcia-Ortiz, Cecilia Contreras-Cubas, Francisco Barajas-Olmos, Marlen Flores-Huacuja, Lorena Orozco.

Formal analysis: Ayken Askapuli, Humberto Garcia-Ortiz, Francisco Barajas-Olmos, Marlen Flores-Huacuja.

Funding acquisition: Maxat Zhabagin, Ainur Akilzhanova, Erlan Ramanculov, Uli Schamiloglu, Zhaxybay Zhumadilov.

Investigation: Ayken Askapuli, Maxat Zhabagin, Zhaxylyk Sabitov, Angelica Martinez-Hernandez, Lorena Orozco.

Methodology: Ayken Askapuli, Humberto Garcia-Ortiz, Cecilia Contreras-Cubas.

Project administration: Ayken Askapuli, Miguel Vilar, Maxat Zhabagin, Angelica Martinez-Hernandez, Lorena Orozco.

Resources: Ainur Akilzhanova, Erlan Ramanculov, Lorena Orozco.

Software: Lorena Orozco.

Supervision: John Hawks.

Writing – original draft: Ayken Askapuli.

Writing – review & editing: Ayken Askapuli, Miguel Vilar, Uli Schamiloglu, Theodore G. Schurr, John Hawks, Naruya Saitou.

References

1. Grousset R, Walford N. The Empire of the Steppes: A history of Central Asia. 1st Ed. New Brunswick, New Jersey, and London: Rutgers University Press; 1970.

2. Golden P. An introduction to the history of the Turkic Peoples. Wiesbaden, Germany: Harrassowitz Publishing House; 1992.

3. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015; 522: 207–211. https://doi.org/10.1038/nature14317 PMID: 25731166

4. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015; 528: 499–503. https://doi.org/10.1038/nature16152 PMID: 26595274

5. Unterländer M. Ancestry and demography and descendants of Iron Age nomads of the Eurasian steppe. Nat Commun. 2017; 8:14615. https://doi.org/10.1038/ncomms14615 PMID: 28256537

6. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505: 43–9. https://doi.org/10.1038/nature12886 PMID: 24352235

7. Lowery RK, Uribe G, Jimenez EB, Weiss M a, Herrera KJ, Regueiro M, et al. Neanderthal and Denisova genetic affinities with contemporary humans: introgression versus common ancestral polymorphisms. Gene. 2013; 530: 83–94. https://doi.org/10.1016/j.gene.2013.06.005 PMID: 23872234

8. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell. 2018; 173: 53–61. https://doi.org/10.1016/j.cell.2018.02.031 PMID: 29551270

9. Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature. 2018; 561: 113–116. https://doi.org/10.1038/s41586-018-0455-x PMID: 30135579

10. Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, et al. The Eurasian Heartland: A continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA. 2001; 98: 10244–10249. https://doi.org/10.1073/pnas.171305098 PMID: 11526236

11. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The genetic legacy of the Mongols. Am J Hum Genet. 2003; 72: 717–721. https://doi.org/10.1086/367774 PMID: 12592608

12. Gokcumen O, Dulik MC, Pai AA, Zhadanov SI, Rubinstein S, Osipova LP, et al. Genetic variation in the enigmatic Altaian Kazakhs of south-central Russia: Insights into Turkic population history. Am J Phys Anthropol. 2008; 136: 278–293. https://doi.org/10.1002/ajpa.20802 PMID: 18322915

13. Comas D, Calafell F, Mateu E, Pe A, Bosch E, Martí R, et al. Trading Genes along the Silk Road: mtDNA Sequences and the Origin of Central Asian Populations. Am J Hum Genet. 1998; 63: 1824–1838. https://doi.org/10.1086/302133 PMID: 9837835

14. Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, Calafell F, et al. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. European Journal of Human Genetics. 2004; 12: 495–504. https://doi.org/10.1038/sj.ejhg.5201160 PMID: 14872198

15. Derenko M, Malyarchuk B, Denisova G, Perkova M, Rogalla U, Grzybowski T, et al. Complete mitochondrial DNA analysis of eastern Eurasian haplogroups rarely found in populations of northern Asia and eastern Europe. PLoS One. 2012; 7: 1–12. https://doi.org/10.1371/journal.pone.0032179 PMID: 22363811

16. Shan W, Ren Zh, Wu W, Hao H, Abulimiti a., Chen K, et al. Maternal and paternal diversity in Xinjiang Kazakh population from China. Russ J Genet. 2014; 50: 1218–1229. https://doi.org/10.7868/s0016675814110149 PMID: 25739291

17. Yao YG, Kong QP, Wang CY, Zhu CL, Zhang YP. Different matrilineal contributions to genetic structure of ethnic groups in the Silk Road region in China. Mol Biol Evol. 2004; 21: 2265–2280. https://doi.org/10.1093/molbev/msh238 PMID: 15317881

18. Aytkhozhina N, Dzisuk N, Ludvikova E. Polymorphism of the noncoding region of the mitochondrial genome in three Kazakh populations inhabiting different areas of Kazakhstan and in DNA samples from ancient populations of the Kazakhstani Altai. Mol Biol. 2004; 38: 494–502. https://doi.org/10.1023/B:MBIL.0000037000.59356.20

19. Tarlykov P, Zholdybayeva E, Akilzhanova A, Nurkina Z, Sabitov Z, Rakhypbekov T, et al. Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. Croat Med J. 2013; 54: 17–24. https://doi.org/10.3325/cmj.2013.54.17 PMID: 23444242

20. Derenko M, Malyarchuk B, Denisova G, Perkova M, Litvinov A, Grzybowski T, et al. Western Eurasian ancestry in modern Siberians based on mitogenomic data. BMC Evol Biol. 2014; 14: 1–11. https://doi.org/10.1186/s12862-014-0217-9 PMID: 25301575

21. Sahakyan H, Kashani BH, Tamang R, Kushniarevich A, Francis A, Costa MD, et al. Origin and spread of human mitochondrial DNA haplogroup U7. Sci Rep. 2017; 7: 1–9. https://doi.org/10.1038/srep46044 PMID: 28387361

22. Ingman M, Gyllensten U. Rate variation between mitochondrial domains and adaptive evolution in humans. Hum Mol Genet. 2007; 16: 2281–2287. https://doi.org/10.1093/hmg/ddm180 PMID: 17617636

23. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44: W58–W63. https://doi.org/10.1093/nar/gkw233 PMID: 27084951

24. Elstov. mtPhyl: Software tool for human mtDNA analysis and phylogeny reconstruction. 2016. Available: https://sites.google.com/site/mtphyl/home

25. Dulik MC, Zhadanov SI, Osipova LP, Askapuli A, Gau L, Gokcumen O, et al. Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and indigenous Altaians. The American Journal of Human Genetics. 2012; 90: 229–246. https://doi.org/10.1016/j.ajhg.2011.12.014 PMID: 22281367

26. Kong QP, Yao YG, Liu M, Shen SP, Chen C, Zhu CL, et al. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Hum Genet. 2003; 113: 391–405. https://doi.org/10.1007/s00439-003-1004-7 PMID: 12938036

27. Qi X, Cui C, Peng Y, Zhang X, Yang Z, Zhong H, et al. Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the Tibetan plateau. Mol Biol Evol. 2013; 30: 1761–1778. https://doi.org/10.1093/molbev/mst093 PMID: 23682168

28. Lan Q, Xie T, Jin X, Fang Y, Mei S, Yang G, et al. mtDNA polymorphism analyses in the Chinese Mongolian group: Efficiency evaluation and further matrilineal genetic structure exploration. Mol Genet Genomic Med. 2019; 7: 1–14. https://doi.org/10.1002/mgg3.934 PMID: 31478599

29. Chen C, Li Y, Tao R, Jin X, Guo Y, Cui W, et al. The genetic structure of Chinese Hui ethnic group revealed by complete mitochondrial genome analyses using massively parallel sequencing. Genes (Basel). 2020; 11: 1–17. https://doi.org/10.3390/genes11111352 PMID: 33202591

30. Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Martínez-Arias R, et al. Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. Am J Hum Genet. 1998; 63: 1824–1838. https://doi.org/10.1086/302133 PMID: 9837835

31. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, et al. Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet. 2007; 81: 1025–1041. https://doi.org/10.1086/522933 PMID: 17924343

32. Peng MS, Xu W, Song JJ, Chen X, Sulaiman X, Cai L, et al. Mitochondrial genomes uncover the maternal history of the Pamir populations. European Journal of Human Genetics. 2018; 26: 124–136. https://doi.org/10.1038/s41431-017-0028-8 PMID: 29187735

33. Tajima F. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics. 1989; 123: 585–595.

34. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018; 35: 1547–1549. https://doi.org/10.1093/molbev/msy096 PMID: 29722887

35. Nei M, Kumar S. Molecular Evolution and Phylogenetics. 1st ed. Oxford University Press; 2000.

36. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2020; 48: D84–D86. https://doi.org/10.1093/nar/gkz956 PMID: 31665464

37. Mielnik-Sikorska M, Daca P, Malyarchuk B, Derenko M, Skonieczna K, Perkova M, et al. The History of Slavs Inferred from Complete Mitochondrial Genome Sequences. PLoS One. 2013; 8: 1–11. https://doi.org/10.1371/journal.pone.0054360 PMID: 23342138

38. Malyarchuk B, Grzybowski T, Derenko M, Perkova M, Vanecek T, Lazur J, et al. Mitochondrial DNA phylogeny in eastern and western Slavs. Mol Biol Evol. 2008; 25: 1651–1658. https://doi.org/10.1093/molbev/msn114 PMID: 18477584

39. FTDNA. Gene by Gene mtDNA Sequences. In: GenBank [Internet]. Available: https://www.ncbi.nlm.nih.gov/nuccore

40. Malyarchuk B, Derenko M, Denisova G, Litvinov A, Rogalla U, Skonieczna K, et al. Whole mitochondrial genome diversity in two Hungarian populations. Molecular Genetics and Genomics. 2018; 293: 1255–1263. https://doi.org/10.1007/s00438-018-1458-x PMID: 29948329

41. Zhao M, Kong Q-P, Wang H-W, Peng M-S, Xie X-D, Wang W-Z, et al. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. Proc Natl Acad Sci U S A. 2009; 106: 21230–21235. https://doi.org/10.1073/pnas.0907844106 PMID: 19955425

42. Kang L, Zheng HX, Chen F, Yan S, Liu K, Qin Z, et al. MtDNA lineage expansions in sherpa population suggest adaptive evolution in Tibetan Highlands. Mol Biol Evol. 2013; 30: 2579–2587. https://doi.org/10.1093/molbev/mst147 PMID: 24002810

43. Zheng HX, Li L, Jiang XY, Yan S, Qin Z, Wang X, et al. MtDNA genomes reveal a relaxation of selective constraints in low-BMI individuals in a Uyghur population. Hum Genet. 2017; 136: 1353–1362. https://doi.org/10.1007/s00439-017-1829-0 PMID: 28799012

44. Palanichamy MG, Mitra B, Zhang CL, Debnath M, Li GM, Wang HW, et al. West Eurasian mtDNA lineages in India: an insight into the spread of the Dravidian language and the origins of the caste system. Hum Genet. 2015; 134: 637–647. https://doi.org/10.1007/s00439-015-1547-4 PMID: 25832481

45. Protasova MS, Grigorenko AP, Tyazhelova T v., Andreeva T v., Reshetov DA, Gusev FE, et al. Whole-genome sequencing identifies a novel ABCB7 gene mutation for X-linked congenital cerebellar ataxia in a large family of Mongolian ancestry. European Journal of Human Genetics. 2016; 24: 550–555. https://doi.org/10.1038/ejhg.2015.139 PMID: 26242992

46. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009; 30: E386–94. https://doi.org/10.1002/humu.20921 PMID: 18853457

47. Schurr TG, Ballinger SW, Gan YY, Hodge J a, Merriwether D a, Lawrence DN, et al. Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am J Hum Genet. 1990; 46: 613–23. PMID: 1968708

48. Torroni A, Schurr G, Yang C, Williams RC, Schanfield MS, Troup GA, et al. Native American Mitochondrial DNA Analysis Indicates That the Amerind and the Nadene Populations Were Founded by. 1992.

49. Robbeets M, Bouckaert R, Conte M, Savelyev A, Li T, An DI, et al. Triangulation supports agricultural spread of the Transeurasian languages. Nature. 2021; 599: 616–621. https://doi.org/10.1038/s41586-021-04108-8 PMID: 34759322

50. Jinam TA, Hong L-C, Phipps ME, Stoneking M, Ameen M, Edo J, et al. Evolutionary History of Continental Southeast Asians: "Early Train" Hypothesis Based on Genetic Analysis of Mitochondrial and Autosomal DNA Data. Mol Biol Evol. 2012; 29: 3513–3527. https://doi.org/10.1093/molbev/mss169 PMID: 22729749

51. Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, et al. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res. 2004; 1832–1850. https://doi.org/10.1101/gr.2286304 PMID: 15466285

52. Yao Y-G, Kong Q-P, Bandelt H-J, Kivisild T, Zhang Y-P. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet. 2002; 70: 635–651. https://doi.org/10.1086/338999 PMID: 11836649

53. Volodko N v, Starikovskaya EB, Mazunin IO, Eltsov NP, Naidenko P v, Wallace DC, et al. Mitochondrial genome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. J Hum Genet. 2008; 1084–1100. https://doi.org/10.1016/j.ajhg.2008.03.019 PMID: 18452887

54. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009; 84: 740–59. https://doi.org/10.1016/j.ajhg.2009.05.001 PMID: 19500773

55. Fu Q, Meyer M, Gao X, Stenzel U, Burbano H a, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A. 2013; 110: 2223–7. https://doi.org/10. 1073/pnas.1221359110 PMID: 23341637

56. Ko AMS, Chen CY, Fu Q, Delfin F, Li M, Chiu HL, et al. Early Austronesians: Into and out of Taiwan. Am J Hum Genet. 2014; 94: 426–436. https://doi.org/10.1016/j.ajhg.2014.02.003 PMID: 24607387

57. Jiang C, Cui J, Liu F, Gao L, Luo Y, Li P, et al. Mitochondrial DNA 10609T promotes hypoxia-induced increase of intracellular ROS and is a risk factor of high altitude polycythemia. PLoS One. 2014;9. https://doi.org/10.1371/journal.pone.0087775 PMID: 24498190

58. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Rogalla U, Perkova M, et al. Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in northern Asia. PLoS One. 2010; 5: 1–9. https://doi.org/10.1371/journal.pone.0015214 PMID: 21203537

59. Schurr TG. Molecular genetic diversity in siberians and native Americans suggests an early colonization of the new world. Entering America: Northeast Asia and Beringia Before the Last Glacial Maximum. 2004.

60. Chard CS. The Nganasan: Wild Reindeer Hunters of the Taimyr Peninsula. Arctic Anthropol. 1963; 1: 105–121.

61. Torroni A, Sukernik RI, Schurr TG, Starikorskaya YB, Cabell MF, Crawford MH, et al. mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet. 1993; 53: 591–608. Available: http://www.ncbi.nlm.nih.gov/pubmed/7688933%0Ahttp://www.pub medcentral.nih.gov/articlerender.fcgi?artid=PMC1682407 PMID: 7688933

62. Derenko M v, Grzybowski T, Malyarchuk BA, Dambueva IK, Denisova GA, Czarny J. Diversity of Mito-chondrial DNA Lineages in South Siberia. Ann Hum Genet. 2003; 391–411. https://doi.org/10.1046/j. 1469-1809.2003.00035.x PMID: 12940914

63. Derenko M, Denisova G, Malyarchuk B, Dambueva I, Bazarov B. Mitogenomic diversity and differenti-ation of the Buryats. J Hum Genet. 2018; 63: 71–81. https://doi.org/10.1038/s10038-017-0370-2 PMID: 29215085

64. Schurr TG, Sukernik REMI, Starikovskaya YB, Wallace DC. Mitochondrial DNA Variation in Koryaks and Itel ' men: Population Replacement in the Okhotsk Sea–Bering Sea Region During the Neolithic. Am J Phys Anthropol. 1999; 39: 1–39.

65. Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A. 2014; 111: 4832–4837. https://doi.org/10.1073/pnas.1316513111 PMID: 24616518

66. Malyarchuk B, Derenko M, Denisova G, Kravtsova O. Mitogenomic diversity in Tatars from the Volga-Ural region of Russia. Mol Biol Evol. 2010; 27: 2220–2226. https://doi.org/10.1093/molbev/msq065 PMID: 20457583

67. Ingman M, Kaessmann H, Pääbo S, Gyllensten U, A Z. Mitochondrial genome variation and the origin of modern humans. Nature. 2000; 408: 708–13. https://doi.org/10.1038/35047064 PMID: 11130070

68. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987; 325: 31–36. https://doi.org/10.1038/325031a0 PMID: 3025745

69. Pakendorf B, Wiebe V, Tarskaia LA, Spitsyn VA, Soodyall H, Rodewald A, et al. Mitochondrial DNA evidence for admixed origins of central Siberian populations. Am J Phys Anthropol. 2003; 224: 211–224. https://doi.org/10.1002/ajpa.10145 PMID: 12567375

70. Pakendorf B, Novgorodov IN, Osakovskij VL, Stoneking M. Mating Patterns Amongst Siberian Rein-deer Herders: Inferences From mtDNA and Y-Chromosomal Analyses. Am J Phys Anthropol. 2007; 1027: 1013–1027. https://doi.org/10.1002/ajpa.20590 PMID: 17492671

71. Sukernik RI, Volodko N v., Mazunin IO, Eltsov NP, Dryomov S v., Starikovskaya EB. Mitochondrial genome diversity in the Tubalar, Even, and Ulchi: Contribution to prehistory of Native Siberians and their affinities to Native Americans. Am J Phys Anthropol. 2012; 148: 123–138. https://doi.org/10. 1002/ajpa.22050 PMID: 22487888

72. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014; 505: 87–91. https://doi. org/10.1038/nature12736 PMID: 24256729

73. Li C, Ning C, Hagelberg E, Li H, Zhao Y, Li W, et al. Analysis of ancient human mitochondrial DNA from the Xiaohe cemetery: Insights into prehistoric population movements in the Tarim Basin, China. BMC Genet. 2015; 16: 1–11. https://doi.org/10.1186/S12863-015-0237-5/FIGURES/4

74. Li C, Li H, Cui Y, Xie C, Cai D, Li W, et al. Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. BMC Biol. 2010; 8: 1–12. https://doi.org/10.1186/1741-7007-8-15/COMMENTS

75. Gao S-Z, Zhang Y, Wei D, Li H-J, Zhao Y-B, Cui Y-Q, et al. Ancient DNA Reveals a Migration of the Ancient Di-Qiang Populations Into Xinjiang as Early as the Early Bronze Age. https://doi.org/10.1002/ajpa.22690 PMID: 25546319

76. Chandrasekar A, Kumar S, Sreenath J, Sarkar BN. Updating Phylogeny of Mitochondrial DNA Macro-haplogroup M in India: Dispersal of Modern Human in South Asian Corridor. October. 2009;4. https://doi.org/10.1371/journal.pone.0007447 PMID: 19823670

77. Bíró a. Z, Zalán a., Völgyi a., Pamjav H. A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). Am J Phys Anthropol. 2009; 139: 305–310. https://doi.org/10.1002/ajpa.20984 PMID: 19170200

78. Balanovsky O, Zhabagin M, Agdzhoyan A, Chukhryaeva M, Zaporozhchenko V, Utevska 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 Iranic speakers. PLoS One. 2015; 10: e0122968: 1–20. https://doi.org/10.1371/journal.pone.0122968 PMID: 25849548

79. Bhandari S, Zhang X, Cui C, Bianba Liao S, Peng Y, et al. Genetic evidence of a recent Tibetan ancestry to Sherpas in the Himalayan region. Sci Rep. 2015;5. https://doi.org/10.1038/srep16249 PMID: 26538459

80. Hammer MF, Horai S. Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet. 1995; 56: 951–962. PMID: 7717406

81. Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, et al. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet. 2004; 5: 26. https://doi.org/10.1186/1471-2156-5-26 PMID: 15339343

82. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014; 514: 445–449. https://doi.org/10.1038/nature13810 PMID: 25341783

83. Massilani D, Skov L, Hajdinjak M, Gunchinsuren B, Tseveendorj D, Yi S, et al. Denisovan ancestry and population history of early East Asians. Science (1979). 2020; 370: 579–583. https://doi.org/10.1126/science.abc1166 PMID: 33122380

84. Lahiri DK, Nurnberger JI. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. Nucleic Acids Res. 1991; 19: 5444. https://doi.org/10.1093/nar/19.19.5444 PMID: 1681511

85. Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, et al. The GenoChip: A New Tool for Genetic Anthropology. Genome Biol Evol. 2013; 5: 1021–1031. https://doi.org/10.1093/gbe/evt066 PMID: 23666864

86. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics 1999 23:2. 1999; 23: 147–147. https://doi.org/10.1038/13779 PMID: 10508508

87. Lipworth S. BaseChanger. 2017. Available: https://github.com/samlipworth/base_changer

88. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics. 2012; 28: 2678–2679. https://doi.org/10.1093/bioinformatics/bts503 PMID: 22914218

89. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059–3066. https://doi.org/10.1093/nar/gkf436 PMID: 12136088

90. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30: 772. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

91. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am J Hum Genet. 2012; 90: 675. https://doi.org/10.1016/j.ajhg.2012.03.002 PMID: 22482806

92. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016; 2: e000056. https://doi.org/10.1099/mgen.0.000056 PMID: 28348851

93. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

95. Barnett DW, Garrison EK, Quinlan AR, Stïmberg MP, Marth GT. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011; 27: 1691–1692. https://doi.org/10.1093/bioinformatics/btr174 PMID: 21493652

96. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27: 2987–2993. https://doi.org/10.1093/bioinformatics/btr509 PMID: 21903627

97. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10. https://doi.org/10.1093/gigascience/giab008 PMID: 33590861

98. Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, et al. Improved calibration of the human mitochondrial clock using ancient genomes. Mol Biol Evol. 2014; 31: 2780. https://doi.org/10.1093/molbev/msu222 PMID: 25100861

99. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014; 10: e1003537: 1–6. https://doi.org/10.1371/journal.pcbi.1003537 PMID: 24722319

100. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018; 67: 901–904. https://doi.org/10.1093/sysbio/syy032 PMID: 29718447

101. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available: https://www.R-project.org/

102. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. Houston, TX, USA: Springer; 2016.

103. Wickham H. Reshaping Data with the reshape Package. J Stat Softw. 2007; 21: 1–20. https://doi.org/10.18637/JSS.V021.I12

104. Winter DJ. MMOD: an R library for the calculation of population differentiation statistics. Mol Ecol Resour. 2012; 12: 1158–1160. https://doi.org/10.1111/j.1755-0998.2012.03174.x PMID: 22883857

105. Fluxus Engineering. Network: Phylogenetic Network Software. 2020. Available: https://fluxus-engineering.com/sharenet.htm

106. JGraph Ltd. diagrams.net. Northampton, UK; Available: https://www.diagrams.net/about