

ANNOTATION

research work for the degree of Doctor of Philosophy (PhD) in the specialty
"6D060700-Biology"

Ashirbekov Eldar Erlanovich

on the topic: «Ethnogenetic features of Y-chromosome of Kazakhs»

General characteristics of the work

The research work is devoted to the analysis of Y-chromosome variability of modern Kazakhs to obtain molecular genetic data that will be applicable for describing the ethnogenesis of Kazakhs. The main method is comparing Y-chromosome polymorphism data with information about the clan-tribal structure of Kazakhs.

Relevance of the research topic

One of the main tasks facing the state is the upbringing of new generation with a developed national consciousness, a sense of patriotism and social memory. The urgency of this task has only increased over time.

Many pages of the history of Kazakhstan remain insufficiently developed, in particular, the question of the formation of the Kazakh nationality. The main reason for the limited understanding of the ethnogenesis of Kazakhs is the scarcity of preserved written sources.

The historian-orientalist N. A. Aristov believed that the main obstacle to determining the origins of ethnic groups is the limited knowledge of their clan-tribal composition. Since Kazakhs still traditionally trace their ancestry, their clan-tribal structure is accessible for study. Researchers of Kazakh ethnogenesis have done some work in this direction, but many questions are awaiting an answer, and hypotheses remain unconfirmed.

Thus, due to the limited historical written sources, there is a need for new approaches to reconstructing the ethnogenesis of Kazakhs, including drawing from non-historical disciplines. One such approach is interdisciplinary research, using the capabilities of molecular genetics on the one hand and genealogical and historical data on the other.

The advancement of molecular biology has opened up unprecedented opportunities for studying the population-genetic structure of living organisms. Analyzing the genetic diversity of the Y-chromosome of modern human populations is a powerful tool for describing the genetic characteristics of different ethnic groups and reconstructing their historical relationships. Because the tradition of tracing paternal genealogy among Kazakhs has persisted to this day, the Y-chromosome serves as a convenient object for research, enabling the comparison of information about Kazakh genealogies with data on genetic polymorphism.

The study of the gene pool of Kazakhs is of great interest to the global scientific community because Kazakhstan's territory is situated at the very center of the continent, between Europe and Asia, where the paths of the significant ancient migrations took place, and was a process of mixing the Caucasoid and

Mongoloid races. In addition, many Asian ethnic groups, including Kazakhs, remain relatively understudied from a molecular genetic perspective.

The purpose of the research is to characterize the population genetic structure of Kazakhs based on the analysis of Y-chromosome variability, in order to obtain molecular genetic data for describing the ethnogenesis of Kazakhs.

Research objectives:

1. Form a collection of DNA samples from male Kazakhs across various regions of Kazakhstan, representing the clan-tribal diversity of Kazakhs;
2. Determine microsatellite haplotypes;
3. Identify belonging to Y-chromosomal haplogroups;
4. Describe the population genetic structure of Kazakhs based on the analysis of Y-chromosomal haplogroup distribution, assess the contribution of ancient populations to the gene pool of Kazakh;
5. Describe the population genetic structure of Kazakhs based on the analysis of microsatellite variability of Y-chromosome, estimate the age of specific lineages;
6. Compare the clan-tribal and genetic structure of Kazakhs. Assess the reliability of information from Kazakh shezhire;
7. Determine the contribution of the genetic component to the zhuz structure.

The object of research: Y-chromosome of Kazakhs.

Subject of research: population genetic structure of Kazakhs based on the analysis of Y-chromosome lineage diversity.

Research methods: DNA isolation, polymerase chain reaction, gel electrophoresis, capillary electrophoresis (fragment analysis), analysis of restriction fragment length polymorphism, analysis of molecular variation, construction of median haplotype networks, estimation of time to most recent common ancestor (TMRCA), principal component analysis, multidimensional scaling.

Scientific novelty of research

Researches of genetic structure of the Kazakh Y-chromosome have been conducted earlier by both domestic and foreign researchers. Within the framework of large regional studies of Asia, as a rule, a limited number of samples from any one region of Kazakhstan were typically included. Given the territorial clan-tribal differentiation and inter-clan genetic differentiation among Kazakhs, the data obtained from these studies cannot be extrapolated to the entire ethnic group. Furthermore, the small set of studied markers and the incompatibility of microsatellite panels do not allow for comparisons and generalizations based on their results.

In the majority of studies conducted by domestic researchers, either a small number of samples were examined, samples were collected from specific regions without considering clan-tribal differences, or a particular clan was studied. An exception are the works of M.K. Zhabagin et al. and E.M. Khusainova et al. The generalized work was M.K. Zhabagin's dissertation, where on a sample of more than 2000 samples the analysis of the relationship of polymorphism of the Y-

chromosome and the clan-tribal structure in the Kazakh population was conducted. The primary focus was on the analysis of haplogroup diversity, with less emphasis on microsatellite polymorphism analysis, possibly due to the limited resolving power of the 17-locus haplotype used.

In our work to study the population-genetic structure of the Kazakhs, we used for the first time the long 27-locus STR-haplotype, which allows for a detailed investigation of clan-tribal relationships. We tested 2,171 DNA samples of Kazakhs from 11 regions of Kazakhstan, representing 21 tribes from three zhuz, as well as 5 distinct clans. Of these, 1,793 samples were examined using the long 27-locus haplotype. Based on microsatellite variability, we calculated the time to the most recent common ancestor (TMRCA) of haplotype clusters, allowing us to estimate the age of individual lineages.

Scientific and practical significance

The research has yielded new data that will be valuable for addressing the issues related to the ethnic history of Kazakhs and other populations that have historical ties with Kazakhs. These findings contribute to the development of the scientific field of ethnogenomics and related areas of science, including ethnography and history, in Kazakhstan. The results of this work will be applicable to specialists in related fields of science such as historians, anthropologists, ethnographers, archaeologists, linguists, and others.

The main provisions submitted for protection:

1. 29 Y-chromosomal haplogroups have been identified. The majority (approximately 65%) of Kazakhs belong to eastern haplogroups, with the two most high-frequency ones being C2-M48 and C2-F4002. The remaining haplogroups, considered western, southwestern, northern, and southern, as well as those with disputed origins, are present in the sample at moderate to low frequencies (the vast majority with frequencies less than 1-2%).

2. Based on 27-locus haplotype, median networks were constructed, allowing for the identification of genetic lineages and a detailed examination of inter- and intra-clan genetic relationships; based on the analysis of microsatellite variability, lineage ages were determined. The ages of predominant lineages within tribes range from approximately 350 to 550 years; the ages of lineages uniting 2-4 related tribes fall within the range of approximately 560 to 720 years; the age of the lineage within C2-M48, which unites the major tribes Alim and Baiuly (known as the Alshyn clan), is approximately 810 years; the age of the lineage within C2-F4002, uniting the majority of tribes in the Great zhuz (known as the Uisyn clan) and the Kerei tribe, is approximately 875 years. The age of ancient lineages within C2-F1756, C2-M48, and R1a-M198, encompassing multiple tribes, are estimated to be around 1200, 1400, and 2000 years, respectively.

3. The obtained data on the population genetic structure of Kazakhs overall align with the concepts of the clan-tribal structure. In most tribes, a specific lineage predominates, indicating a founder effect. Additionally, there are two to three lineages with intermediate frequencies and numerous lineages with low frequencies. As expected, at the clan level, the pattern is less diverse. However, singular lineages are typically not observed; instead, one predominant lineage and

one to three secondary lineages are usually identified. Thus, in the vast majority of cases, the shezhire data about the origin of a tribal group from a common ancestor are not corroborated by genetic data. For an individual, shezhire can reliably describe kinship up to a certain level (when considered from bottom to top): individual, clan, tribal, and supra-tribal).

4. The zhuz structure is only partially based on genetic relatedness among tribes within the zhuzes. In most tribes of the Great zhuz, a lineage within C2-F4002 predominates, indicating kinship ties among these tribes; notably, the tribes of Sirgeli, Ysty, and Kanly stand out. In the Junior zhuz, the tribes Alim and Baiuly are united by a common lineage within C2-M48, with the tribe Zhetiru standing out separately. The Middle zhuz proves to be the most diverse of the three, where each studied tribe has its distinctive set of prevalent lineages, allowing them to be considered as separate groups; the Kerei tribe is distantly related to the tribes of the Great zhuz.

Main research findings and conclusions

1. A collection of DNA samples from Kazakhs has been assembled. During expeditions to the North Kazakhstan, Kostanay, Akmola, Pavlodar, East Kazakhstan, Zhetysu, Almaty, Zhambyl, Turkestan, Ulytau, and Atyrau regions, biomaterial was collected from 2171 Kazakhs representing 21 tribes and 5 separate clans. DNA has been extracted from the biomaterial, and a collection of DNA samples has been formed.

2. Y-chromosomal haplogroups have been identified. In the result of determining 25 single nucleotide polymorphisms using the PCR-RFLP method, as well as utilizing predictor programs and comparing data with previous studies, a total of 29 Y-chromosomal haplogroups have been identified.

3. Microsatellite haplotypes have been determined. Using the AmpFLSTR Yfiler PCR Amplification Kit (Applied Biosystems, USA) microsatellite haplotypes for 17 loci have been determined for 378 samples from the collection; using the Yfiler Plus PCR Amplification Kit (Applied Biosystems) microsatellite haplotypes for 27 loci have been determined for 1793 samples from the collection.

4. Median networks of haplotypes have been constructed, and an estimation of the age of lineages has been provided. Using the Network program within each lineage median networks of microsatellite haplotypes have been constructed, which made it possible to identify lineages. Using the Batwing program (based on the coalescent theory and Bayesian algorithm) and the Network program (utilizing the rho statistic function), an estimate of the age of the most recent common ancestor (TMRCA) of the resulting clusters has been provided. The ages of predominant lineages within tribes range from approximately 350 to 550 years. The ages of lineages uniting 2-4 related tribes fall within the range of approximately 560 to 720 years. The age of the lineage within C2-M48, which unites the major tribes Alim and Baiuly (known as the Alshyn clan), is approximately 810 years. The age of the lineage within C2-F4002, uniting the majority of tribes in the Great zhuz (known as the Uisyn clan) and the Kerei tribe, is approximately 875 years. The ages of ancient lineages within C2-F1756, C2-

M48, and R1a-M198, present in numerous tribes, are estimated to be around 1200, 1400, and 2000 years, respectively.

5. An analysis of inter- and intra-clan genetic relationships has been conducted, an assessment of the reliability of the information from shezhire has been provided. Based on the analysis of haplogroup distribution and microsatellite variability in the studied clans and tribes, utilizing principal component analysis (PCA), multidimensional scaling, and analysis of molecular variance (AMOVA), taking into account the obtained median networks, an analysis of inter- and intra-clan genetic relationships has been conducted. The obtained data on the population genetic structure of Kazakhs overall align with the concepts of the clan-tribal structure. In most tribes, a specific lineage predominates, indicating a founder effect. Additionally, there are two to three lineages with intermediate frequencies and numerous lineages with low frequencies. As expected, at the clan level, the pattern is less diverse. However, singular lineages are typically not observed; instead, one predominant lineage and one to three secondary lineages are usually identified. Thus, in the vast majority of cases, the shezhire data about the origin of a tribal group from a common ancestor are not corroborated by genetic data. For an individual, shezhire can reliably describe kinship up to a certain level (when considered from bottom to top): individual, familial, clan, tribal, and supra-tribal).

6. The contribution of the genetic component to the zhuz structure has been assessed. By visualizing the genetic differentiation among tribes using principal component analysis (PCA) and multidimensional scaling methods, as well as employing analysis of molecular variance (AMOVA), it has been demonstrated that the zhuz structure is only partially based on genetic relatedness among tribes within the zhuzes. In most tribes of the Great zhuz, a lineage within C2-F4002 predominates, indicating kinship ties among these tribes; notably, the tribes of Sirgeli, Ysty, and Kanly stand out. In the Junior zhuz, the tribes Alim and Baruly are united by a common lineage within C2-M48, with the tribe Zhetiru standing out separately. The Middle zhuz proves to be the most diverse of the three, where each studied tribe has its distinctive set of prevalent lineages, allowing them to be considered as separate groups; the Kerei tribe is distantly related to the tribes of the Great zhuz.

7. An assessment of the contribution of ancient lineages to the gene pool of Kazakhs has been provided. The comparison of haplogroup spectra among different tribes of Kazakhs with global populations indicates that the most high-frequency haplogroups likely mark the contribution of ancient populations from the east to the Kazakh gene pool. Other lineages, considered western, southwestern, northern, and southern, as well as those with disputed origins, were found in the sample at moderate to low frequencies, the majority being less than 1-2%. The extensive diversity of identified haplogroups likely reflects the formation of the Kazakh ethnic group in the very center of Eurasia, where various ancient populations met and intermingled for a long time.

The study demonstrates that the chosen subject and research methodology are effective in investigating the ethnogenesis of Kazakhs. Considering the significant potential of such research, there is a need to continue these studies,

particularly in populations from unexplored regions of Kazakhstan, to elucidate a comprehensive picture of the genetic diversity of the Y-chromosome among Kazakhs.

The obtained data will prove valuable for historians, ethnographers, population geneticists, and other specialists addressing the issues related to the ethnogenesis of Kazakhs.

Author's personal contribution

The author made a direct contribution at all stages of the research, including formulating the goals and objectives, project development, collecting biomaterial, isolating DNA, determining single nucleotide polymorphisms, identifying microsatellite haplotypes, primary data processing, database creation, statistical data analysis, interpreting the obtained results, and formulating conclusions.

Connection with the plan of the main scientific works

The research was carried out within the framework of projects GF0226 “Molecular-genetic research of ethnogenetic polymorphism of Y-chromosome of modern Kazakhs and ancient people of the Desht-i-Kipchak region”, 4050/GF4 “Molecular-genetic description of Kazakhs from South-Kazakhstan, Zhambyl and Almaty regions based on single-nucleotide and microsatellite variability of Y chromosome”, AP09259527 “Historical, genealogical and genetic research of the descendants of Genghis Khan”, AP09058538 “Historical and genetic study of Kazakh shezhire”, AP09259560 “Genetic study of connections between Kazakh clans, Kyrgyz clans, Altaians, Khakassians and Shors through phylogenetic analysis of Y-chromosome from Altai to the Pamirs”, and AP08855823 «The study of genetic connections between clans of Kazakhs, Mongols and Kalmyks through phylogenetic analysis of Y-chromosome from Mongolia to the Urals”.

Approbation of the work

The research results were presented at the following scientific conferences:

1. 40th FEBS Congress “The Biochemical Basis of Life” (Berlin, Germany, July 4–9, 2015).
2. International scientific-practical conference "Youth, science and Innovation," devoted 25-anniversary of Independence of the Republic of Kazakhstan and the 50th anniversary of Aktobe regional state University after the name of K. Zhubanov (Aktobe, April 8, 2016).
3. VII International school of young scientists on molecular genetics "Genomics and biology of living systems" (Zvenigorod, Russia, November 14–18, 2016).
4. The international scientific conference of young scientists “Fundamental research and innovations in molecular biology, biotechnology, biochemistry” dedicated to the 80th anniversary of academician Murat Aitkhozhin (Almaty, November 28–29, 2019).
5. International scientific conference of young scientists “Basic and applied research in molecular biology, biochemistry, biotechnology” dedicated to the 40th anniversary of the founding of the M. A. Aitkhozhin Institute of Molecular Biology and Biochemistry (Almaty, November 17, 2023).

Publications

Regarding the research topic, 13 scientific works have been published, including: 4 articles in scientific journals ranked in the 1-2 quartiles by impact factor in the Web of Science database; 1 article in a scientific journal ranked in the 4th quartile by impact factor in the Web of Science database; 3 articles in domestic journals, including 1 article in a publication recommended by the Committee for Quality Assurance in the Field of Science and Higher Education; 5 international conference abstracts, including 1 abstract in the proceedings of a foreign conference.