Whole Mitochondrial Genome Analysis of the Daur Ethnic Minority from Hulunbuir of the Inner Mongolia Autonomous Region, China

Chi-Zao Wang

Department of Radiology, the First Affiliated Hospital of Shantou University Medical College, Shantou 515041

Xue-Er Yu

Fudan University

Mei-Sen Shi

Criminal Justice College of China University of Political Science and Law

Hui Li

Fudan University

Shu-Hua Ma (✉ shuhuama6699@163.com)

Department of Radiology, the First Affiliated Hospital of Shantou University Medical College, Shantou 515041

Research Article

Keywords: Daur, Mongolic-speaking groups, whole mtDNA sequence, Heilongjiang River basin, northern East Asia

Posted Date: September 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-809269/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Mitochondrial DNA (mtDNA) variations are often associated with bioenergetics, disease, and speciation and can be used to track the history of women. Although advances in massively parallel sequencing (MPS) technology have greatly promoted our understanding of the population's history (especially genome-wide data and whole Y chromosome sequencing), the whole mtDNA sequence of many important groups has not been fully studied. In this study, we employed whole mtDNA genomes of 209 healthy and unrelated individuals from the Daur group, a Mongolic-speaking representative population of the indigenous groups in the Heilongjiang River basin (also known as the Amur River basin).

Results: The dataset presented 127 distinct mtDNA haplotypes, resulting in a haplotype diversity of 0.9933. The majority of haplotypes were assigned to eastern Eurasian-specific lineages, such as D4 (19.62%), B4 (9.09%), D5 (7.66%) and M7 (4.78%). We collected whole mitochondrial genomes from the 1000 Genomes Project, the Human Genome Diversity Project and published papers for population comparisons and phylogenetic analysis, and the results showed that the Daurians do have certain connections with the ancient populations in the Heilongjiang River basin but the matrilineal genetic composition of the Daur group was also greatly influenced by other non-Mongolic groups from neighboring areas.

Conclusions: Collectively, the whole mtDNA data generated in the present study will augment the existing mtDNA database and deepen our understanding of the history of the Daurians as well as other populations from northern East Asia.

Background

The Daur minority is one of the important members of the Mongolic-speaking population, recognized as one of the 56 official ethnic groups in China. They originally lived on the north beach of the Heilongjiang River (Amur River)[1]. After the 17th century, they gradually moved to Hulunbuir, Qiqihar and other settlements on the south beach of the Heilongjiang River[2]. A small number of them even migrated to Tacheng Prefecture in Xinjiang Province, as Qing government troops accompanied by their families[3]. The Daur ethnic group has long intermingled with the Ewenk and Oroqen ethnic groups, two other officially recognized ethnic groups in China, who are members of the Tungusic-speaking ethnic populations. They were known as the Suolun in the Qing Dynasty and are now called the Three Minorities in Inner Mongolia[3].

The Daur ethnic group has appeared in many important genetic studies as one of the representatives of Mongolic-speaking populations and indigenous groups from the Heilongjiang River basin[4-6]. Our previous studies elaborated on the paternal phylogenetic relationship between the Daur group, other Mongolic-speaking populations and Tungusic-speaking populations (including the Aisin Gioro family) by analyzing their Y-STR genetic polymorphisms and whole Y-chromosome sequences[7-11]. Recent

genome-wide studies have further revealed the high level of genetic continuity of indigenous populations from the Heilongjiang River basin (including the Daur ethnic group) over at least the last 14,000 years and their distinct phylogenetic position in the genetic structure of human populations in East Asia[6,12,13]. However, previous studies on the Daur group from the perspective of maternal inheritance were relatively limited in terms of sample size and merely based on partial sequence polymorphisms, such as hypervariable segments I and II (HVS-I and HVS-II, respectively) and the control region (CR)[14-16]. Two early genetic studies established a certain genetic relationship between the modern Daur group and the ancient Khitan, which is one of the most significant findings of ethnic studies in China[15,16]. Therefore, expanding the sample size and introducing whole mitochondrial genome analyses will undoubtedly contribute to a more comprehensive understanding of the maternal genetic background of the Daurians.

In this study, the whole mitochondrial genomes of 209 healthy and unrelated Daur individuals from Northeast China were sequenced by massive parallel sequencing (MPS) on the HiSeq X Ten system (Illumina, San Diego, CA, USA). Based on the sequencing data, we analyzed the haplogroup distribution and genetic diversity of the maternal genetic structure of the Daur group. To shed more light on the genetic relationship of the Daur group with worldwide populations, especially other neighboring/linguistically close populations and some related ancient groups, we conducted comprehensive population genetic analyses via Principal Component Analysis (PCA)[17], multidimensional scaling (MDS) analysis, the construction of a neighbor-joining phylogenetic tree and network analysis.

Methods

Samples, DNA extraction and quantification

A cohort of 209 maternally unrelated Daur individuals (84 females and 125 males) was collected after receiving informed consent. The individuals were considered autochthonous if their ancestors had lived in Hulunbuir, Inner Mongolia Autonomous Region of China for at least three generations. Written informed consent was obtained from the all participants, and the ethics committee of School of Life Sciences, Fudan University, Shanghai, People's Republic of China approved this study.

Genomic DNA was extracted using a DP-318 Kit (Tiangen Biotechnology, Beijing, China) according to the manufacturer's protocol. The quantity of gDNA was measured with a NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE, USA) according to the manufacturer's protocol. In consideration of the requirements of downstream processing, the gDNA was normalized to 0.1 ng/ μ L and stored at -20 °C until amplification.

Library construction and workflows for next-generation sequencing

DNA libraries were constructed using an MtDNA Library Preparation Kit 2.0 (Enlighten Biotech, Shanghai, China) and a WhoChrMT kit (Enlighten Biotech, Shanghai, China). PCR amplification was performed in a final volume of 30 μ L containing 10 ng of template DNA, 5 μ L RealCapChrMT Mix and 10 μ L

3×EnzymeHF. Total reaction volumes were adjusted with nuclease-free water. The PCR was performed under the following conditions: enzyme activation for 3 min at 98 °C, 13 cycles of 20 s at 98 °C and 4 min at 58 °C, 7 cycles of 20 s at 98 °C and 1 min at 72 °C, 2 min at 72 °C followed by a 10 °C hold. The PCR products were purified with Agencourt AMPure XP beads (Beckman Coulter). Then, a second round of PCR amplification was carried out to introduce adapters and barcodes. The reaction volume (30 µL) was comprised of 10 µL 3×EnzymeHF, 18 µL nuclease-free water, 1 µL primer mix and 1 µL barcode mix. The PCR was performed under the following conditions: enzyme activation for 2 min at 98 °C, 7 cycles of 15 s at 98 °C, 15 s at 58 °C and 30 s min at 72 °C, extension for 2 min at 72 °C followed by a 10 °C hold. After purification, the libraries were pooled to a final concentration of 20 pM. Sequencing was performed on the Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA) with the corresponding Reagent Kit (PE150).

Sequencing data analysis

The sequence data obtained from the Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA) were automatically analyzed by base recognition and converted into the original sequences in FASTQ format. First, redundant primers and indexes in the initial offline data were removed by cutadapt software[18]. Second, low-quality reads were filtered by Trimmomatic software[19]. To ensure the successful alignment of the loop amplification captured sequence, the final cleaned files were mapped to the revised Cambridge Reference Sequence[20] plus 64 bp (rCRS + 64 bp) using the Burrows-Wheeler Aligner[21] to generate the binary alignment/map (BAM) file. The sequences were also compared with the human reference genome hg19 to filter nuclear copies of mtDNA (NUMTs)[22]. We used Bedtools[23] to extract all reads that were successfully mapped to the HG19 reference genome from the BAM files in the previous step and then realigned them to rCRS + 64 bp to generate new BAM files using Bowtie2 software[24]. Then, SAMtools[25] and VarScan[26] were used to identify the mutation sites and output variants in VCF format files. Finally, BCFTools [25] was used to generate the consensus sequence (FASTA).

Haplogroup assignment and genetic diversity analysis

Sequencing performance was evaluated by read depth. The mtDNA haplogroups were determined using HaploGrep 2[27] based on PhyloTree build 17[28] and reconfirmed using the updated query engine (SAM2) built into EMPOP[29]. With reference to PhyloTree build 17, we constructed a simplified phylogenetic tree that showed the distribution of the coarse haplogroups. Haplogroup frequencies were estimated by direct gene counting. Haplotype diversities were calculated according to Nei's formula[30]. The discrimination capacity (DC) was also calculated as an important diversity parameter[31]. To show the differences in the genetic diversity of the different mitochondrial regions, haplogroup-based analyses were repeated for the control region (CR, 16024 to 576) and hypervariable segment I (HVS1, 16024 to 16488).

Population comparisons

To investigate the genetic relationship between the Daur group and other populations around the world, 49 worldwide populations were collected from the 1000 Genomes Project[32] and the Human Genome Diversity Project[33] (Table S1). Subsequently, the genetic background of the Daur was analyzed by typical Principal Component Analysis (PCA) with the R statistical package (<https://www.r-project.org/>) based on haplogroup frequencies (Table S2). In particular, we required the group size to be greater than 20 in the PCA to avoid artificially low genetic diversity.

When we focused on the matrilineal genetic relationships in and around East Asia, we further sorted the reference groups by language classification. For the purpose of enlarging the reference dataset as much as possible, partial sequence (16024-16383) analysis was carried out first, followed by analysis involving only the whole mtDNA sequence dataset. We also used the HVS1 sequence from the whole mtDNA sequence dataset for a comparative analysis. Detailed information and cited references of the populations are listed in Table S3. In this part of the analysis, we also required the group size to be greater than 20. For groups with better data sources, we did not select samples from genome projects. For example, Qing-Peng Kong et al. provided 21,668 Han samples from virtually all provinces in China (the average for each province was over 600)[34]. It is usually difficult to collect representative datasets of such large sample sizes for current genome projects. Pairwise population comparisons and AMOVA analyses were executed using Arlequin[35], and the *Fst* matrix was imported into the R statistical package for multidimensional scaling (MDS) analysis and plotting heatmaps. Based on the pairwise *Fst* values, a phylogenetic tree was built with neighbor-joining (N-J) methods by MEGA-X[36].

For some MtDNA haplogroups of particular interest, we collected relevant full sequence data from the 1000 Genomes Project[32], the Human Genome Diversity Project[33] and published papers (including recently published ancient genomes, Table S4) for network analysis. The network was constructed using the median-joining method in the Popart software[37,38]. In particular, since network itself is not suitable for showing the connections among a large number of samples, the samples we collect here do not include Han samples from Qing-Peng Kong[34].

Results And Discussion

Sequence performance

The average mapped reads were 139,681 per sample, and the overall mean read depth was $1,260X \pm 422X$ (mean $\pm$ SD) per individual. The variants recommended by EMPOP as well as the haplogroup information and the mean sequencing depth of 209 Daur individuals are presented in Table S5.

Haplogroup distribution

Figure 1 presents a simplified phylogenetic tree that shows the distribution of the coarse haplogroups, and the detailed typing results are shown in Table S5. In general, the matrilineal component of the Daur group was predominantly comprised of the eastern Eurasian-specific component (89.21%), represented by haplogroups D (28.24%), G (10.54%), B (10%), C (8.62%), R9 (7.65%), N9 (6.92%), Z (6.23%), A (4.79%),

M7 (4.78%) and M9 (1.44%) [39,40]. The remaining samples consisted of haplogroups U (1.44%), T (1.92%) and H (1.44%), which are generally confined to the European region[41,42], and a few root types (R* and M*). Among these haplogroups, C and D have distinct Asian characteristics, and more than half of the northern Asian pool of human mtDNA is fragmented into their subclades[39,43]. In the Daur population we studied, haplogroup C consisted of four sister subclades, C1 (0.48%), C4 (2.39%), C5 (3.83%) and C7 (1.92%), while haplogroup D consisted of three sister subclades, D2 (0.96%), D4 (19.62%) and D6 (7.66%). Notably, haplogroup D4 not only has a high frequency but also contains a total of 28 abundant downstream clades (Table S5). Some subbranches of haplogroup D4 have very distinctive geographical distributions and are of great significance for the study of the demographic history of Asia[34,43]. For example, haplogroup D4j (2.87% in this study) demonstrated a more southern geographic distribution, and haplogroup D4e4a (0.48% in this study) was mostly found in the Subarctic and Arctic regions[44]. According to previous studies, haplogroups B (10% in this study) and G (10.54% in this study) are also frequent in Mongolic-speaking groups[39,45].

On the whole, the Daur population in this study embodies distinct regional and ethnic characteristics. Compared with earlier studies on Daur mitochondria [14-16], our research showed some changes in some haplogroup frequency distributions and detected some types that were not previously found in Daur mitochondria (U, F, H, etc.), which could be attributed to the larger sample size and more advanced full mtDNA sequence methods used in this study.

Genetic diversity analysis

Based on whole mtDNA sequence data, a total of 127 different haplogroups were identified from the 209 unrelated Daur samples, of which 81 (63.78%) were unique. Although close matrilineal relatives (first to three degrees) were excluded, 61.24% of the total samples still shared haplogroups with others. It is worth noting that the haplogroups belonging to M7b1a1+(16192), G2a1 and Z3d were shared by 6 individuals. Moreover, one haplogroup was shared between five individuals, seven were shared between four individuals, seven were shared between three individuals and twenty-eight were shared between two individuals. The overall haplogroup diversity was calculated as 0.9933 with a discrimination capacity of 60.77%. Table S6 summarizes the above results. Repeated analysis based on CR and HVS1 showed that whole mtDNA sequence data decreased the number of shared haplogroups and increased the number of unique haplogroups. This is reflected in the discriminatory capacity increasing from 53.11% with the HVS1 haplogroups and 54.55% with the CR haplogroups to 60.77% with the whole mtDNA sequence for the Daur samples (Table S6). These results indicate that the whole mtDNA sequence data offer a high power of discrimination and can be useful for genetic investigation and maternal lineage research in the Daur minority.

Of course, the genetic diversity of maternal genetic markers was slightly lower than that of paternal genetic markers, which is more due to the limitations of mitochondrial genetic markers themselves. In our previous study of genetic polymorphisms of 27 Yfiler® Plus loci in the Daur group, a total of 196 different haplotypes were observed in the sample of 203 Daur individuals, and the overall haplotype diversity was

calculated as 0.9997 with a discrimination capacity of 0.9655[7]. Our other two studies based on Y-STR/Y-SNP and Y-chromosome sequencing provided rich details on the paternal genetic diversity of the Daur group[8-10].

Population comparisons and phylogenetic analysis

We first performed a series of genetic relationship and structure analyses among 51 populations based on haplogroup frequencies (Table S2). In our PCA results, 59.2% of the genetic variations were extracted by the first three components (Figure 2). The African ancestry (AFR) and American ancestry (AMR) populations can be separated clearly by PC1 and PC2, while the four large groups from Eurasia, East Asian ancestry (EAS), European ancestry (EUR), South Asian ancestry (SAS) and Middle East (Middle_Est), are closely related and even overlap. When using PC2 and PC3 as references, the PCA showed a genetic affinity cline, an east-west cline, which consisted of EAS, SAS, Middle_Est and EUR. Our Daur population was located on one side of EAS between Yakut and CHB (Beijing Han) individuals. Whole mtDNA sequence analysis based on worldwide populations further illustrates the correlation between the maternal genetic background and geographic factors, and the position of Daur in the PCA plot was generally consistent with its geographic origin.

To clarify the genetic relationship of the Daur group with East Asian populations, partial sequences (16024-16383) of all 55 populations (Table S3) were selected for further genetic analysis. Pairwise *Fst* values (Table S7) were calculated based on partial sequence variation results that are displayed as a heatmap in Figure S1. The results ranged from 0.00148 (for the LK group, Lowland Kyrgyz from Artux, Xinjiang, China) to 0.09088 (for the Balochi group, from Pakistan). The Daur group showed higher similarity levels with LK, Yakut, JPT, LNH, UzbT, MHN, LU, Tib_LB, Gelao, SouthKorea, Turk, Hazara and Burusho ($Fst < 0.01$, $P > 0.0009$, after Bonferroni's correction), which may indicate a smaller genetic difference.

MDS plots based on pairwise *Fst* value data were drawn for the obtained data, as shown in Figure 3. To a certain extent, the genetic relationship patterns reconstructed here also correspond to their geographical origin or linguistic affinities. In terms of rough categories, populations from the Indo-European languages (Iranian, Indo-Aryan and Slavic) are gathered on the left side, people from the Sino-Tibetan languages family (Chinese and Tibeto-Burman) are clustered mainly on the right side, while Mongolic, Tungusic and Turkic (used to be known collectively as the “Altaic language”) speaking groups are mainly located at the bottom of the plot. The Daur group (Mongolic-speaking) is in marginal position of the “Altaic language” group and is closest to the LNH (Han in Liaoning Province, also located in Northeast China). Notably, the Daur group is also close to Gelao (Gelao in Guizhou Province) and MHN (Miao in Hunan Province), two groups from South China. We have not come up with a good explanation for this so far, but in our previous studies on the Y chromosome of the Daur ethnic group, we found that the proportion of four haplogroups mainly distributed in Southern China and Southeast Asia (O1b1a1-M95, C2a1b-F845, O2a2a1a2-M7 and O1a-M119) was also not low (14.49%)[11]. Further research is needed on the

phenomenon that certain southern characteristic elements appear in both the Daur paternal and maternal lineages.

A cladogram was also drawn applying N-J methods, as presented in Figure 4. There were four main branches and the relatively independent LK population in the resulting phylogenetic tree, in which the first branch populations consisted of populations speaking the Indo-European languages (Iranian, Indo-Aryan and Slavic) and Turkic languages, the second and third branches and the relatively independent LK population mainly came from Mongolic-, Tungusic- and Turkic-speaking groups, and the bottom branch was mainly comprised of people from the Sino-Tibetan language family (Chinese and Tibeto-Burman) and low latitude regions. In the bottom branch, the Daur group was first clustered with LNH and JPT (Japanese in Tokyo, Japan) and then with Tibeto-Burman speaking populations and low latitude populations. Although the Daur group is representative of Mongolic-speaking populations, it is not genetically close to the others, as shown in Figure 4. This indicates that the maternal genetic composition of the Daur group is greatly influenced by other groups, especially a genetic admixture from northern East Asia.

The heat maps, MDS plots and cladograms involving only the whole mtDNA sequence and the HVS1 sequence taken from the whole mtDNA sequence (Table S8-S9) are shown in Figure S2-S4. Despite including fewer groups, the patterns of genetic relationships reconstructed here are also generally similar to the results of the partial sequence dataset, and populations with linguistic or regional associations clustered more closely in the MDS plot of the whole mtDNA sequence. In other words, this also reflects that the whole mitochondrial sequence data increase the resolution and offer a higher power of discrimination than previous maternal typing systems.

As mentioned above, haplogroup D4 not only has a high frequency (19.62%) but also contains abundant downstream clades in the Daur samples. According to previous studies based on partial sequences, D4 is also the high-frequency type of several ancient ethnic groups in Northeast China[46-48]. In the latest genome-wide study of northern East Asia, D4 also accounted for the majority of the detected samples in the ancient Heilongjiang River basin(66.67%, 16/24)[13]. We collected relevant available full sequence data (Table S4) and constructed networks(Figure 5 and Figure S5). In Figure 5A, the Daur samples came from scattered sources, showing connections with multiple regions of Asia. When we focused on the genetic connection between the Daur samples and ancient samples, we found that most samples from the ancient Heilongjiang River basin had close connections with samples of Daur (Figure 5A and 5B), and concentrated in haplogroups D4m, D4o, D4g and D4c. Haplogroup D4h, another high-frequency type in ancient Heilongjiang River basin populations, has not been detected in the modern Daur group which also makes sense that D4h is a distinctive native American type that may not have been involved in the late demographic history of northern East Asia[49]. In other words, the network analysis shows that the Daurians do have certain connections with the ancient populations in the Heilong River basin, but in the development process of the Daurians, they also absorbed a large number of female population from other sources. As to whether the modern Daur group has the closest matrilineal genetic connection with

the ancient Heilongjiang population, we will collect more complete mitochondrial sequence data and carry out it in detail in follow-up studies.

Conclusion

The present study provided the first set of whole mitochondrial genome data of 209 Daur individuals residing in Northeast China. The investigation of the Daur maternal lineages revealed that the vast majority of haplogroups belong to the eastern Eurasian-specific component. Population analyses showed that the Daurians do have certain connections with the ancient populations in the Heilongjiang River basin but the matrilineal genetic composition of the Daur group was also greatly influenced by other non-Mongolic groups from neighboring areas. This study also shows that whole mitochondrial sequence data can improve the resolution and offers a high power of discrimination in maternal studies by comparison of whole and partial sequence data in genetic diversity and population comparative analyses. Overall, the mitogenomes generated in the present study will augment the existing Daur mtDNA database, which provides a deeper understanding of the genetic composition of the Daur group and could potentially be useful for regional-specific and prerequisite references for forensic, genealogical, and evolutionary purposes.

Abbreviations

MtDNA: Mitochondrial DNA, MPS: Massively parallel sequencing, HVS-I and HVS-II: Hypervariable segments I and II, CR: Control region, PCA: Principal component analysis, MDS: multidimensional scaling, DC: Discrimination capacity, N-J: Neighbor-joining.

Declarations

Ethics and Consent to Participate

Written informed consent was obtained from the all participants, and the ethics committee of School of Life Sciences, Fudan University, Shanghai, People's Republic of China approved this study. All methods were performed in accordance with the Declaration of Helsinki.

Consent for publication

Not Applicable.

Availability of data and materials

The 209 novel Daur complete mtDNA sequences are being uploaded to the Genome Sequence Archive (GSA) in the BIG Data Center (Members BIGDC 2017), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences(<http://bigd.big.ac.cn/gsa-human>).

Author Contributions

C.Z.,W. conceived the study and wrote the original draft. X.E.,Y. performed calculations, interpreted results and contributed to writing the manuscript. C.Z.,W. and X.E.,Y. were major contributors in writing the manuscript. M.S.,S. and S.H.,M. helped in conceiving the study and contributed to the manuscript. M.S.,S. and H.L. provided the funding support. All authors read and approved the final manuscript.

Acknowledgments

We thank all sample donors for their contributions to this work and all those who helped with sample collection. We are particularly grateful to Mr. A-Li Aola for his support of this study.

Competing Interest

The authors declare no conflict of interest.

Funding

This study was funded by the National Natural Science Foundation of China (91731303).

References

1. Narangoa L, Cribb R. Historical Atlas of Northeast Asia,1590–2010: Korea, Manchuria, Mongolia, Eastern Siberia. New York: Columbia University Press,2014.
2. Dmytryshyn, B, Crownhart-Vaughan, EAP, Vaughan, T. Russia's conquest of Siberia, 1558–1700: a documentary record, vol. 1. Portland: The Press of the Oregon Historical Society,1985.
3. Aola A-Li. The western expedition and defense of Solon. Beijing,China: Minzu University of China Press,2017. In Chinese
4. 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(3):717-21.
5. Xue Y, Zerjal T, Bao W, Zhu S, Shu Q, Xu J, et al. Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics.* 2006,172(4):2431-9.
6. Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, et al. Genomic insights into the formation of human populations in East Asia. *Nature.* 2021,591(7850):413-9.
7. Wang CZ, Su MJ, Li Y, Chen L, Jin X, Wen SQ, et al. Genetic polymorphisms of 27 Yfiler® Plus loci in the Daur and Mongolian ethnic minorities from Hulunbuir of Inner Mongolia Autonomous Region, China. *Forensic Sci Int Genet.* 2019,40:e252-e5.
8. Wei LH, Yan S, Yu G, Huang YZ, Yao DL, Li SL, et al. Genetic trail for the early migrations of Aisin Gioro, the imperial house of the Qing dynasty. *J Hum Genet.* 2017,62(3):407-11.
9. Wang CZ, Wei LH, Wang LX, Wen SQ, Yu XE, Shi MS, et al. Relating Clans Ao and Aisin Gioro from northeast China by whole Y-chromosome sequencing. *J Hum Genet.* 2019,64(8):775-80.
10. Liu BL, Ma PC, Wang CZ, Yan S, Yao HB, Li YL, et al. Paternal origin of Tungusic-speaking populations: Insights from the updated phylogenetic tree of Y-chromosome haplogroup C2a-M86.

Am J Hum Biol. 2021,33(2):e23462.

11. Wang CZ, Shi MS, Li H. The Origin of Daur from the Perspective of Molecular Anthropology. Journal of North Minzu University.2018,5:110-117. [In Chinese]
12. Siska V, Jones ER, Jeon S, Bhak Y, Kim HM, Cho YS, et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. Sci Adv. 2017,3(2):e1601877.
13. Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, et al. The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell. 2021,184(12):3256-66 e13.
14. 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(5):391-405. [In Chinese]
15. Wu DY, Ma SS, Liu CY, Yang HM, Liu FZ, Chen ZC, et al. Study on the molecular archaeology of Khitan ancient cadavers. Journal of Yunnan University (Natural Science Edition). 1999(S3):300. (In Chinese)
16. Xu Yue, Zhang XL, Zhang QC, Cui YQ, Zhou H, Zhu H. Genetic Relationship between Ancient Khitan and Modern Daur. Journal of Jilin University (Science Edition). 2006(06):997-1000.(In Chinese)
17. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006,2(12):e190.
18. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet Journal. 2011,17(1).
19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014,30(15):2114-20.
20. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999,23(2):147.
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010,26(5):589-95.
22. Just RS, Irwin JA, Parson W. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int Genet. 2015,18:131-9.
23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010,26(6):841-2.
24. Langmead B, Salzberg SL. Langmead B, Salzberg SL.. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. Nature Methods. 2012,9(4):357-9.
25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009,25(16):2078-9.
26. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009,25(17):2283-5.
27. Weissensteiner H, Pacher D, Kloss-Brandstatter 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(W1):W58-63.

28. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* 2009,30(2):E386-94.
29. Huber N, Parson W, Dur A. Next generation database search algorithm for forensic mitogenome analyses. *Forensic Sci Int Genet.* 2018,37:204-14.
30. Clegg MT. Molecular evolution: molecular evolutionary genetics. *Science.* 1987,235(4788):599.
31. Ip SCY, Lin SW, Lam TT. Haplotype data of 27 Y-STR loci in Hong Kong Chinese. *Forensic Sci Int Genet.* 2019,38:e14-e5.
32. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 2020,48(D1):D941-D7.
33. Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet.* 2006,70(Pt 6):841-7.
34. Li YC, Ye WJ, Jiang CG, Zeng Z, Tian JY, Yang LQ, et al. River Valleys Shaped the Maternal Genetic Landscape of Han Chinese. *Mol Biol Evol.* 2019,36(8):1643-52.
35. Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, et al. The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol.* 2002,19(10):1737-51.
36. Sudhir K, Glen S, Li M, Christina K, Koichiro T. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology & Evolution.* 2018(6):6.
37. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 1999,16(1):37-48.
38. Leigh JW, Bryant D. PopART: Full-Feature Software for Haplotype Network Construction. *Methods in Ecology and Evolution.* 2015,6(9).
39. 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(2):e32179.
40. 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(6):637-47.
41. 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(6):637-47.
42. 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.* 2014,14:217.
43. 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(12):e15214.

44. Ko MS, Chen CY, Fu Q, Delfin F, Ko YC. Early Austronesians: Into and Out Of Taiwan. *American Journal of Human Genetics*. 2014,94(3):426-36.
45. 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(10):e00934.
46. Wang H, Liu W, Yuqin FU, Zhang X, Zhou H, Zhu H. Molecular biological analysis of remains from Jiangjunggou Cemetery in Inner Mongolia. *Progress in Natural Science*. 2006,16(7):727-31.
47. Molecular genetic analysis of remains from Lamadong cemetery, Liaoning, China. Wiley Subscription Services, Inc, A Wiley Company. 2007,134(3):404-11.
48. Yu C, Xie L, Zhang X, Hui Z, Hong Z. Genetic analysis on Tuoba Xianbei remains excavated from Qilang Mountain Cemetery in Qahar Right Wing Middle Banner of Inner Mongolia. *FEBS Letters*. 2006,580(26):6242-6.
49. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, et al. Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr Biol*. 2009,19(1):1-8.

Figures

Figure 1

The phylogenetic relationship of coarse mtDNA haplogroups in this study and their clan-based frequencies among the Daur group.

Figure 2

Principal component (PC) analyses between the Daur group and other worldwide reference populations included in the 1000 Genomes Project and HGDP.

Figure 4

. Phylogenetic tree built with neighbor-joining methods based on pairwise F_{st} values for mtDNA partial sequence (16024-16383) variations of 55 populations.

Figure 5

Median-joining haplogroup D4 networks. (A) Ancient and modern samples. (B) The modern Daur and ancient samples. According to the naming method of the original author, samples from the ancient Heilongjiang River basin have been specially labeled [13]. The larger the number, the older the age.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables20210812.xlsx](#)
- [SupplementaryFigures20210808.docx](#)