

# A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia

Johannes Krause,<sup>1,\*</sup> Adrian W. Briggs,<sup>1</sup> Martin Kircher,<sup>1</sup> Tomislav Maricic,<sup>1</sup> Nicolas Zwyns,<sup>1</sup> Anatoli Derevianko,<sup>2</sup> and Svante Pääbo<sup>1</sup>

<sup>1</sup>Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany

<sup>2</sup>Paleolithic Department, Institute of Archaeology & Ethnography, Russian Academy of Sciences, Siberian Branch, Lavrentieva Prospekt, 17 Novosibirsk, 630090 Russia

## Summary

The recovery of DNA sequences from early modern humans (EMHs) could shed light on their interactions with archaic groups such as Neandertals and their relationships to current human populations. However, such experiments are highly problematic because present-day human DNA frequently contaminates bones [1, 2]. For example, in a recent study of mitochondrial (mt) DNA from Neolithic European skeletons, sequence variants were only taken as authentic if they were absent or rare in the present population, whereas others had to be discounted as possible contamination [3, 4]. This limits analysis to EMH individuals carrying rare sequences and thus yields a biased view of the ancient gene pool. Other approaches of identifying contaminating DNA, such as genotyping all individuals who have come into contact with a sample, restrict analyses to specimens where this is possible [5, 6] and do not exclude all possible sources of contamination. By studying mtDNA in Neandertal remains, where contamination and endogenous DNA can be distinguished by sequence, we show that fragmentation patterns and nucleotide misincorporations can be used to gauge authenticity of ancient DNA sequences. We use these features to determine a complete mtDNA sequence from a ~30,000-year-old EMH from the Kostenki 14 site in Russia.

## Results and Discussion

DNA extracted from fossil remains is invariably degraded to a low average fragment size [1, 7–10]. Until recently, the only way to gauge the length of DNA molecules from the organism of interest has been to compare the success rates of polymerase chain reactions (PCRs) performed with primer pairs that target templates of different lengths [11]. However, because PCR requires two primers of length ~20 nt and usually targets a segment of > 10 nt between the primers, the abundance of template fragments below a length of ~50 bp remain unknown. This situation has changed with the application of single primer extension-based approaches [12] and high-throughput DNA sequencing to ancient DNA, e.g., the Illumina Genome Analyzer II (GA2) [13] and Roche 454 platforms [9]. In particular, the adaptor-ligation and direct-sequencing approach of the two latter technologies allow DNA fragments of short lengths to be completely sequenced. This has shown that more than half of DNA fragments

extracted from, e.g., Neandertal remains, are of a size less than 50 nt [7] and therefore not retrievable by PCR. By contrast, modern DNA that contaminates a specimen may often be of larger average size [7, 14–16].

In order to test whether the extent of contamination differs systematically between the longer molecules analyzed by PCR and the bulk of the molecules analyzed by direct high-throughput sequencing, which also include shorter molecules, we analyzed 21 Neandertal DNA extracts by PCR-based contamination assays [7, 17] with up to three primer pairs used to amplify regions of the mtDNA that differ between Neandertals and present-day humans. Between 30 and 2,118 amplified fragments per extract were sequenced on the 454 platform, and the ratios of Neandertal-like and modern-human-like mtDNA sequences were determined (Figure 1A). Four extracts yielded less than 5% contaminating modern human sequences whereas the remaining 17 extracts contained between 9% and 100% contaminating modern human sequences. We then analyzed the same 21 Neandertal extracts with a recently published method for ancient DNA sequence capture (PEC) [18]. This method uses 5'-biotinylated oligonucleotide primers that hybridize to sequences of interest. The primers are then extended by a DNA polymerase to capture ancient target sequences, allowing fragments of size ~20 nt and above to be retrieved. We used up to six PEC primers designed to anneal to conserved mtDNA regions adjacent to diagnostic differences between human and Neandertal mtDNA, followed by direct sequencing on the 454 platform. This generated between 30 and 465 independent informative fragments per extract (Figure 1B). In 14 of the 21 extracts, exclusively Neandertal and no contaminating molecules were detected, and the remaining seven extracts had contamination levels lower than or equal to 1%. A major reason for this discrepancy in the apparent extent of contamination is evident from the size distributions of ancient DNA fragments retrieved by PEC (see Figure S2, available online), which shows that only 27% of the fragments are larger than the smallest size amplified by PCR (67 nt). Thus, the direct sequencing taps into a majority of short molecules that cannot be analyzed by PCR. In this population, endogenous Neandertal molecules dominate quantitatively over contaminating molecules to such an extent that in many cases contaminating molecules are not even seen among the molecules sequenced.

In addition to a reduction in size, cytosine deamination is the most common modification readily detectable in ancient DNA [19]. With sequencing on high-throughput platforms, it has been shown to occur particularly toward the ends of the molecules and to result in C-to-T substitutions close to 5'-ends and in the complementary G-to-A substitutions at 3'-ends as a result of end polishing before sequencing adaptor ligation [12, 20]. In order to directly study this feature as well as fragment size in large numbers of not only endogenous but also contaminating modern human DNA, we captured informative mtDNA molecules [7, 14] by PEC from libraries prepared from Neandertal individuals that had earlier been found to be highly contaminated and had therefore been excluded from further study. They come from individual 2 [21] from Feldhofer, Germany; individual 2 [14] from Mezmaiskaya, Russia; and

\*Correspondence: krause@eva.mpg.de

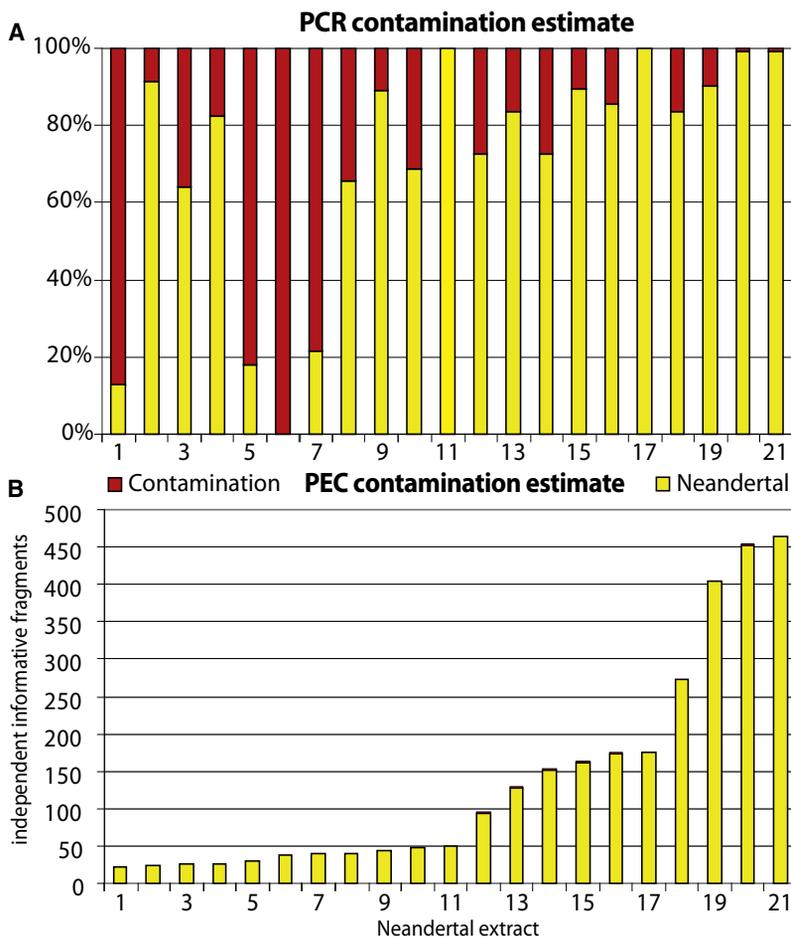


Figure 1. PCR and PEC Contamination Estimate in Neandertal Extracts

Estimates of the level of modern human DNA contamination (red) compared to endogenous Neandertal DNA (yellow) in 21 Neandertal extracts (see also Table S1). (A) For each extract the proportion of human to Neandertal clones for the PCR contamination assay is shown. (B) For each extract, the numbers of human and Neandertal DNA fragments retrieved with the PEC are shown (see also Figure S2).

It has been shown that fragmentation of ancient DNA occurs predominantly at purine bases [20]. In order to study whether this feature differs between endogenous and contaminating DNA, we analyzed the base composition close to 5'- and 3'-ends of mtDNA sequences from the Feldhofer, Mezmaiskaya, and Teshik Tash libraries by aligning each fragment to a reference mtDNA and gauging the base composition at the first 10 bp outside the terminal base of the fragment. As previously described, for the endogenous Neandertal fragments an elevation of G and A in frequency was found at the 5'-end, whereas C and T were found elevated at the 3'-end (Figure S3), consistent with a preferential fragmentation of ancient DNA at purine bases [20]. In contrast, the human contaminants in the Feldhofer and Mezmaiskaya libraries as well as the fragments in the highly contaminated Teshik Tash library show no increase in frequency of purines or pyrimidines on either side of the fragment (Figure S3). In conclusion, fragment length, deamination-induced

sequence errors at ends of molecules, and purine-associated fragmentation represent features by which endogenous and contaminating populations of DNA molecules can be distinguished in at least some late Pleistocene specimens.

from Teshik Tash, Uzbekistan [22]. For the two former libraries, 1,492 and 534 contaminating molecules were found among 1,865 and 1,063 informative molecules sequenced, respectively, and for the latter library 220 informative molecules were retrieved of which only five were of Neandertal origin (Table S2).

In the two former cases, the mode of the size distributions of contaminating molecules seems to be higher than 81 nt, which is the maximum length of molecules sequenced. In contrast, the mode of the endogenous DNA fragments is ~60 and ~45 nt, respectively (Figure 2). In Teshik Tash, where practically all mtDNA molecules captured represent contamination, no direct comparison to the endogenous DNA can be made, but it should be noted that the mode of the size distribution of the contaminating molecules is similar to the endogenous DNA found in the Feldhofer specimen (~60 bp, Figure 2), suggesting that fragment size alone cannot be used to gauge the authenticity of ancient DNA [23]. In the endogenous DNA molecules from the Feldhofer and Mezmaiskaya specimens, cytosine residues at the 5'-ends of DNA fragments appear as thymine residues in ~28% and ~35% of cases, respectively, and similar levels of G-to-A substitutions are seen at the 3'-ends. By contrast, in contaminating DNA molecules cytosine residues at the 5'-ends appear as thymine residues in ~4% and ~2% of cases, respectively (Figure 2). In the Teshik Tash specimen, where contamination greatly dominates, C-to-T misincorporations are seen at less than 3% of 5'-ends (Figure 2).

Given that sequencing of shorter molecules appears to reduce the problem of contamination and that these features may help distinguish endogenous from contaminating DNA, we attempted to sequence the mitochondrial genome of an EMH. To do this, we extracted DNA from the Markina Gora skeleton found in the Kostenki 14 site in Russia, which was excavated in 1954 [24]. Based on stratigraphic evidence, it is older than 30,000 years but younger than 33,000 years (Supplemental Information). All attempts to produce direct radiocarbon dates failed or gave unreliable results, probably because of previous treatment of the bones with organic compounds. We used a total of 574 PEC primers in four multiplex mixes to retrieve the complete mtDNA from a sequencing library produced from the DNA extract. The primers were chosen to match Neandertal mtDNA sequences so that the capture would be equally biased for all modern human mtDNA, given that Neandertal mtDNAs fall outside the variation of modern human mtDNAs [7]. From the retrieved molecules a total of 4,066,269 sequences were produced on the Illumina GA2 platform and processed with a custom mapping assembler. Approximately 30% of the raw sequences represented mtDNA fragments. Sequences with the same start and end coordinates were collapsed because they may stem from the same ancient molecules, yielding 10,664 distinct mtDNA

sequence errors at ends of molecules, and purine-associated fragmentation represent features by which endogenous and contaminating populations of DNA molecules can be distinguished in at least some late Pleistocene specimens.

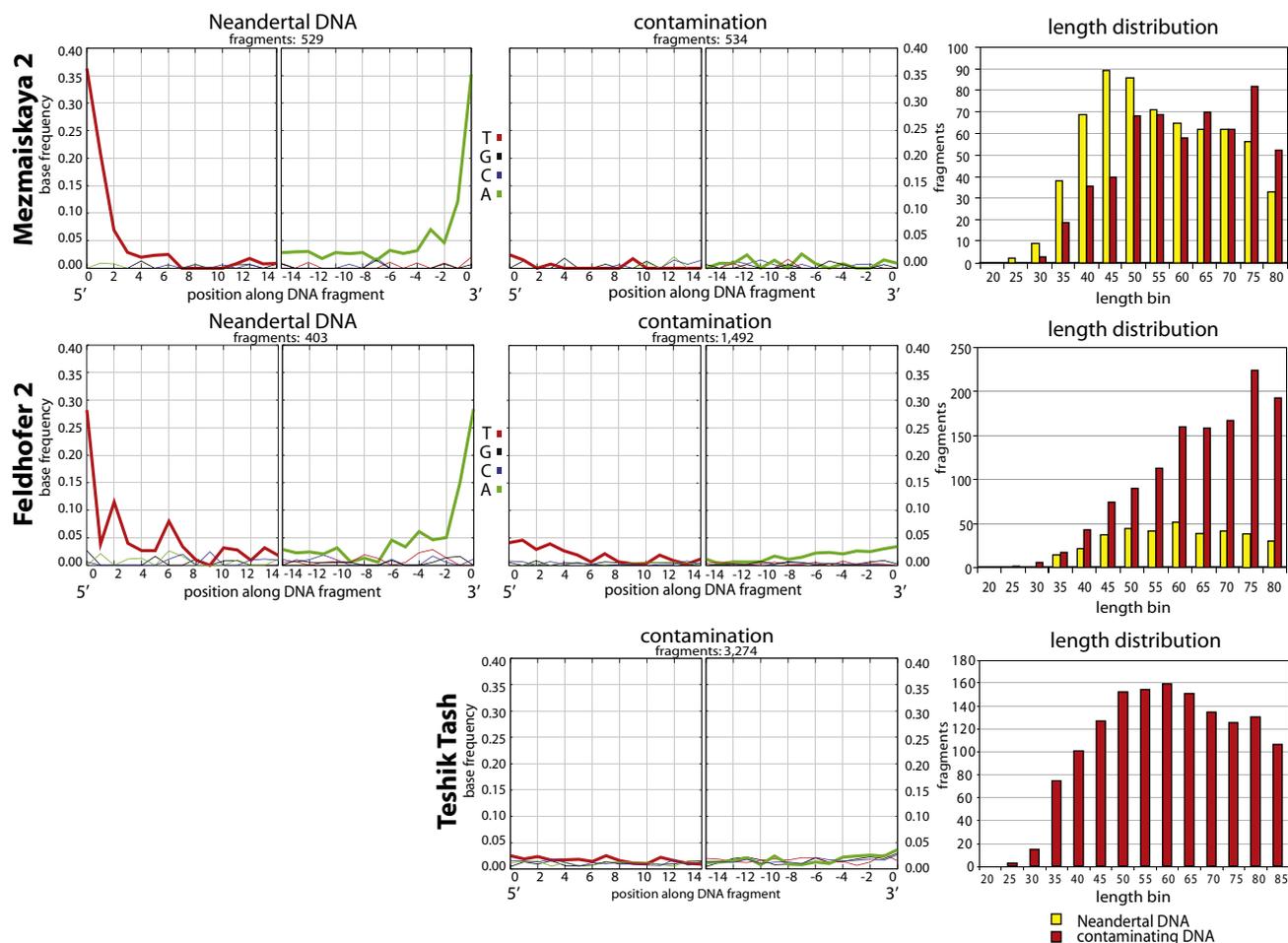


Figure 2. Misincorporation Patterns and Length Distribution of Neandertal DNA and Contaminating DNA

Nucleotide misincorporation patterns and length distributions for Neandertal and contaminating mtDNA from Feldhofer 2, Mezmaiskaya 2, and Teshik Tash specimens. The maximum sequence lengths determined were 83 bp for Mezmaiskaya and Feldhofer and 87 bp for Teshik Tash (see also Figure S3).

sequences. The average length of these sequences is 54 bp (Figure 3A). They were used to construct a consensus mtDNA sequence where the average coverage is 31.2 and over 99% of positions have 4-fold and more coverage (Figure 3E). Cytosine residues in the consensus sequence appear as thymine residues at 5'-ends of sequences, and guanidine residues appear as adenine residues at 3'-ends in ~38% of cases (Figure 3B). Furthermore, purines are substantially elevated at the base 5' to the mtDNA sequences and pyrimidines are elevated at the base 3' to sequences (Figure 3C), indicating preferential breakage at purines [20]. Thus, the mtDNA sequences determined from the Kostenki 14 individual are similar to Neandertal molecules of a similar or slightly older age with respect to length, nucleotide misincorporations, and fragmentation patterns.

Next, we investigated whether mtDNA sequences from one or more individuals are present in the Kostenki 14 DNA library. In order to maximize the power to detect the presence of multiple mtDNA sequences, we identified three positions where the Kostenki 14 consensus sequence differs from a world-wide sample of 311 complete mtDNAs [7] (542C > T, 711T > C, and 15,262T > C) as well as one position (13,269A > G) that differs from more than 99% of these mtDNAs. At these positions, mtDNA from another individual is particularly

likely to differ from the consensus sequence [25]. Among the 77 distinct DNA fragments that overlap these positions, only one shows a base different from the consensus sequence (Figure 4), indicating that the vast majority 98.7% (95% C.I. 93.1%–99.9%) of DNA comes from one single individual.

Hence, if the consensus sequence would derive from a contaminating DNA source, the Kostenki 14 specimen would be contaminated by DNA from a single individual that carries an mtDNA sequence that is unusual in the current human gene pool. Furthermore, the DNA from that individual would be degraded and chemically modified to an extent and in a pattern typical of endogenous mtDNA sequences in Neandertals rather than contaminating DNA in such remains. Although this is possible, we suggest that the more plausible alternative is that the consensus sequence is endogenous to the Kostenki specimen.

We next analyzed how the Kostenki 14 mtDNA sequence relates to Neandertal and current human mtDNAs. It is clearly distinct from Neandertal mtDNAs and carries five diagnostic substitutions that define mtDNA haplogroup U2 (Figure S4) (rCRS positions 11,467, 12,308, 12,372, 1,811, and 16,051, based on Phylotree.org-mtDNA, build 5). In a phylogenetic tree relating all 63 mtDNAs from the 311 present-day mtDNA dataset in haplogroup U as well as the Kostenki mtDNA, the

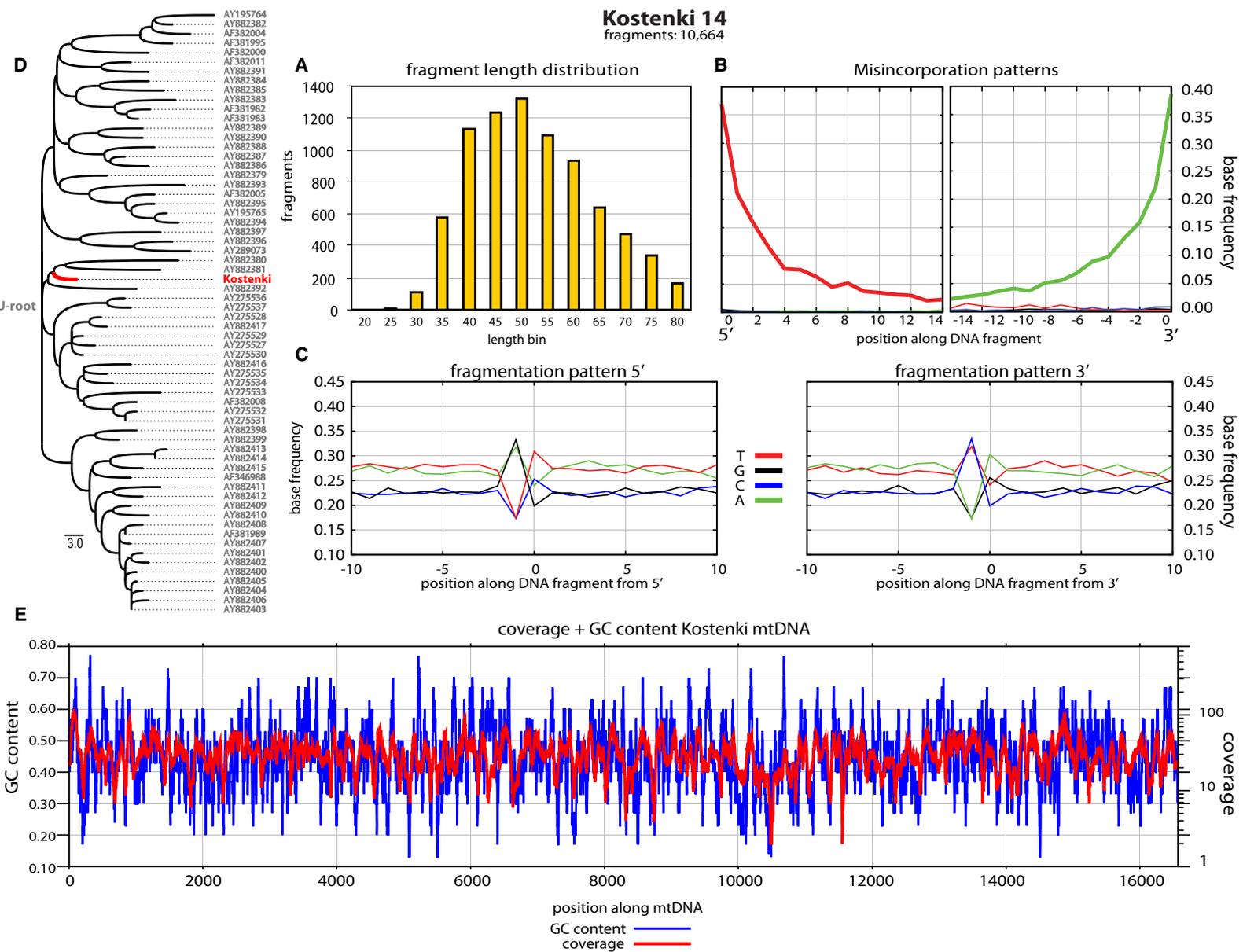


Figure 3. Sequence Features of Kostenki 14 mtDNA  
 (A) Length distribution. (B) nucleotide misincorporation patterns, and (C) nucleotide base frequency of mtDNA fragments captured by PEC from the Kostenki 14 library. (D) Maximum parsimony tree of the Kostenki 14 consensus sequence and 63 mtDNAs that fall into haplogroup U1s shown (see also Figures S4 and S5). (E) Redundant coverage of Kostenki 14 (red, after clustering unique molecules) and GC content (blue) along the complete mtDNA are shown. Coverage is strongly correlated with GC content as was shown before [7, 14] (see also Figure S6).



2. Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G., and Pääbo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* 2, E57.
3. Ermini, L., Olivieri, C., Rizzi, E., Corti, G., Bonnal, R., Soares, P., Luciani, S., Marota, I., De Bellis, G., Richards, M.B., and Rollo, F. (2008). Complete mitochondrial genome sequence of the Tyrolean Iceman. *Curr. Biol.* 18, 1687–1693.
4. Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tänzer, M., Villems, R., Renfrew, C., Gronenborn, D., Alt, K.W., and Burger, J. (2005). Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* 310, 1016–1018.
5. Caramelli, D., Milani, L., Vai, S., Modi, A., Pecchioli, E., Girardi, M., Pilli, E., Lari, M., Lippi, B., Ronchitelli, A., et al. (2008). A 28,000 years old Cro-Magnon mtDNA sequence differs from all potentially contaminating modern sequences. *PLoS ONE* 3, e2700.
6. Sampietro, M.L., Gilbert, M.T., Lao, O., Caramelli, D., Lari, M., Bertranpetit, J., and Lalueza-Fox, C. (2006). Tracking down human contamination in ancient human teeth. *Mol. Biol. Evol.* 23, 1801–1807.
7. Green, R.E., Malaspina, A.S., Krause, J., Briggs, A.W., Johnson, P.L., Uhler, C., Meyer, M., Good, J.M., Maricic, T., Stenzel, U., et al. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134, 416–426.
8. Pääbo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. USA* 86, 1939–1943.
9. Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., et al. (2006). Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* 311, 392–394.
10. Richards, M.B., Macaulay, V.A., Bandelt, H.J., and Sykes, B.C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.* 62, 241–260.
11. Hagelberg, E., Bell, L.S., Allen, T., Boyde, A., Jones, S.J., and Clegg, J.B. (1991). Analysis of ancient bone DNA: Techniques and applications. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 333, 399–407.
12. Brotherton, P., Endicott, P., Sanchez, J.J., Beaumont, M., Barnett, R., Austin, J., and Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* 35, 5717–5728.
13. Blow, M.J., Zhang, T., Woyke, T., Speller, C.F., Krivoschapkin, A., Yang, D.Y., Derevianko, A., and Rubin, E.M. (2008). Identification of ancient remains through genomic sequencing. *Genome Res.* 18, 1347–1353.
14. Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325, 318–321.
15. Noonan, J.P., Coop, G., Kudravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J.K., and Rubin, E.M. (2006). Sequencing and analysis of Neandertal genomic DNA. *Science* 314, 1113–1118.
16. Malmström, H., Svensson, E.M., Gilbert, M.T., Willerslev, E., Götherström, A., and Holmlund, G. (2007). More on contamination: The use of asymmetric molecular behavior to identify authentic ancient human DNA. *Mol. Biol. Evol.* 24, 998–1004.
17. Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., and Pääbo, S. (2006). Analysis of one million base pairs of Neandertal DNA. *Nature* 444, 330–336.
18. Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., and Pääbo, S. (2009). Primer extension capture: Targeted sequence retrieval from heavily degraded DNA sources. *J. Vis. Exp.* 3, 1573.
19. Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A.V., and Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* 29, 4793–4799.
20. Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. USA* 104, 14616–14621.
21. Schmitz, R.W., Serre, D., Bonani, G., Feine, S., Hillgruber, F., Krainitzki, H., Pääbo, S., and Smith, F.H. (2002). The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl. Acad. Sci. USA* 99, 13342–13347.
22. Krause, J., Orlando, L., Serre, D., Viola, B., Prüfer, K., Richards, M.P., Hublin, J.J., Hänni, C., Derevianko, A.P., and Pääbo, S. (2007). Neandertals in central Asia and Siberia. *Nature* 449, 902–904.
23. Green, R.E., Briggs, A.W., Krause, J., Prüfer, K., Burbano, H.A., Siebauer, M., Lachmann, M., and Pääbo, S. (2009). The Neandertal genome and ancient DNA authenticity. *EMBO J.* 28, 2494–2502.
24. Rogachev, A.N. (1955). Pogrebenie drevnekamennogo veka na stoyanke kostenki XIV (markina gora). *Sovetskaya Etnografiya* 1955, 29–38.
25. Gilbert, M.T., Kivisild, T., Grønnow, B., Andersen, P.K., Metspalu, E., Reidla, M., Tamm, E., Axelsson, E., Götherström, A., Campos, P.F., et al. (2008). Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Science* 320, 1787–1789.
26. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* 100, 171–176.
27. Bramanti, B., Thomas, M.G., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M.N., Jankauskas, R., Kind, C.J., et al. (2009). Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* 326, 137–140.
28. Malmström, H., Gilbert, M.T., Thomas, M.G., Brandström, M., Storå, J., Molnar, P., Andersen, P.K., Bendixen, C., Holmlund, G., Götherström, A., and Willerslev, E. (2009). Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr. Biol.* 19, 1758–1762.