Uncovering Ancient DNA in Petroleum: Insights into Earth's Hidden Life Events and Geological History

Wan-Qian Zhao[1,3*], Zhan-Yong Guo[2], Yu-Qi Guo[3*], Mei-Jun Li[4*], Gang-Qiang Cao[1], Zeng-Yuan Tian[1], Ran Chai[5], Li-You Qiu[6*], Jin-Hua Zeng[7], Xin-Ge Zhang[8], Tian-Cang Qin[9], Jin-Yu Yang[10], Ming-Jie Chen[11], Mei-Rong Song[12], Fei Liang[13], Jun-Hui Geng[1], Chun-Yan Zhou[14,15], Shu-Jie Zhang[3], Li-Juan Zhao[3]

Affiliations:
[01]School of Agricultural Sciences, Zhengzhou University, Zhengzhou, China
[02]College of Agronomy, Henan Agricultural University, Zhengzhou, China
[03]School of Life Sciences, Zhengzhou University, Zhengzhou, China
[04]National Key Laboratory of Petroleum Resources and Engineering, College of Geosciences, China University of Petroleum (Beijing), Beijing, China
[05]Yellow River Conservancy Technical Institute, Zhengzhou, China
[06]College of Life Sciences, Henan Agricultural University, Zhengzhou, China
[07]School of Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, China
[08]College of Tobacco, Henan Agricultural University, Zhengzhou, China
[09]Zhengzhou ToYou Bioengineering Co., Ltd, Zhengzhou, China
[10]Haining BoShang Biotechnology Co., Ltd, Haining, China
[11]NewCore Biotechnology Co., Ltd, Shanghai, China
[12]College of Sciences, Henan Agricultural University, Zhengzhou, China
[13]College of Paleontology, Shenyang Normal University, Shenyang, China
[14]Jiesai Biotechnology Co., Ltd, Haining, China
[15]Jiangsu Jiayu Biomedical Technology Co., Ltd, Jiangyin, China

*Corresponding authors, Emails:
wqzhao@zzu.edu.cn, guoyuqi@zzu.edu.cn, meijunli@cup.edu.cn, qliyou@henau.edu.cn

Abstract

This study employed nanoparticle affinity bead technology to isolate DNA from petroleum. The research identified 3,159,020 DNA sequences from petroleum extracts (pDNA), primarily from environmental DNA (eDNA). The study demonstrates how oil captures a vast array of eDNA through molecular exchanges with its surroundings, including the fragments from ancient periods (paeDNA) and more recent times (preDNA). An academically rigorous “mega screening method” was established to identify these fragments, revealing significant loss of the most original *in situ* DNA (oriDNA). Interestingly, the persistence of paeDNA provides valuable ecological and evolutionary insights that surpass traditional fossils. Petroleum, regarded as newly discovered fossils, reveals Earth's hidden history by uncovering the ancient existence of marine species, ancestral birds, and unclassified species of ancient hominids. Additionally, the study sheds light on the tracks of locally extinct animals, including yaks, turkeys, and kiwi birds. Notably, these ancient DNA (aDNA) fragments do not exhibit signs of “deamination”, contradicting Pääbo's standards for ancient, *in situ* DNA authenticity. The study also highlights significant geological events, such as marine invasions during the Paleocene epoch. This work has the potential to revolutionize both petroleum geology and paleontology.

Keywords: nanoparticle affinity bead, petroleum, DNA, mega screening method, marine, hominid, bird

Introduction

Petroleum is a fossil fuel derived from fossilized organic material, such as plankton and bacteria. The remains of these organisms were deposited with sediments to oceans or lakes and were converted to petroleum under thermal stress[1,2]. GMs, also known as molecular fossils or biomarkers, are organic molecules found in geological samples that provide information about the biological and environmental conditions of the past. They are typically preserved in sediments, rocks, crude oils, and coals. They show little or no change in structure from their parent organic molecular in living organisms. Therefore, they can provide reliable geological interpretations to help solve exploration, development, production, environmental, and archeological or paleontological problems[3]. Previous studies show that the source rocks of the Biyang Sag oil field in the Nanxiang Basin are mainly located in the third member (Eh3) of the Paleogene Hetaoyuan Formation[4,5,6,7,8]. Research suggests that the organic matter within the Eh3 shales contains bio-precursors mainly from phytoplankton, bacteria, and benthic algae, with relatively small contributions from animal and higher plant materials[6]. Furthermore, the presence of marine organisms 24-*n*-propylcholestanes indicates the occurrence of marine incursions in the Nanxiang Basin[6,7,8].

To date, paleontological research can successfully obtain *in situ* aDNA, the oriDNA, from certain fossils dating back to 1 million years ago (Ma)[9]. However, fossil distributions are relatively scarce, geographically dispersed, and temporally discontinuous. Obtaining more sources of aDNA could lead to a great breakthrough in the discipline. Many experiments have shown that vegetable oil refined at 400°C contains amplifiable DNA fragments[10,11,12], suggesting that a certain amount of DNA can be stored and protected in the deoxygenated environment of oil-phase liquids. Organic matter begins to mature and hydrocarbons form when the hydrocarbon source rock reaches a certain temperature (threshold temperature) in the formation conditions of petroleum. The main stages of oil production generally begin at temperatures no lower than 60°C and end at temperatures no higher than 180°C[1]. This raises the question of whether amplifiable DNA can be extracted from petroleum. If the reply is yes, it would be a major advance in the petroleum and paleontological belt, especially in revealing the genetic information stored within petroleum during its formation, migration, accumulation, and preservation. To solve this problem, we successfully extracted DNA from crude oil samples from the Nanyang Oilfield, using a new technology called "nanoparticle affinity bead DNA extraction technology"[13], constructed DNA libraries, and sequenced them. Because GMs and DNA have different physicochemical properties, the two types of molecules are accessed, degraded, lost, and stored in reservoirs in completely different ways.

An amount of 3,159,020 DNA fragments were extracted from oil samples, categorized, and analyzed using the "mega screening method". The results showed that most of the preDNA, mainly bacterial fragments, and species involved in human activities. A small amount of aDNA, including paeDNA, has been discovered and is used to validate important local historical events, such as the local marine invasion, activities of unclassified *Homo species*, and the presence of ancient local bird species. Surprisingly, the oriDNA is nearly absent, making it challenging to determine the species origin of the petroleum, based on the pDNA data. However, this discovery opens up a new resource

for collecting aDNA fragments from non-remains that work as fossils. It provides an exciting opportunity to trace the evolutionary path of ancient local life in the distant past within the context of local paleontology and geological evolution. Notably, most of the DNA fragments showed no sign of “deamination”, which negates the Pääbo team's use of “deamination” as the primary criterion for aDNA authenticity. This study emphasizes the need to create a global pDNA database, serving as a new platform for research in interdisciplinary fields, such as geology, industrial genomics, and paleontology.

Results and Discussion

Petroleum Geology of the Nanyang area

In the early-middle Triassic, the area where the Nanyang field is located was a trough in the Qinling region of southern China. It formed into a complete land mass after the collision of the Sino-Indian plate in the middle-late Triassic[14]. These oil wells are situated in the Biyang Depression between the North China Plate and the Yangtze Plate. Analysis of the Eh3 shales through geochemical studies has revealed an average organic matter abundance of 2.96 wt%, indicating low to mature stages of maturity, with prevalent type I to type II kerogen dominance. The Eh_3 stratum in the Biyang Sag has been proven to be the most favorable hydrocarbon source rocks for the discovered oil in the Nanyang Oilfield[15]. The reservoirs are buried at depths of 90 to 1000 m, with oil layer thicknesses ranging from 1 to 4 m and assemblage thicknesses ranging from 2 to 15 m. The petroleum reservoirs in the Nanyang Oilfield experienced a complex filling history, including an early oil filling, significant uplift, a second phase of filling, and late-stage biodegradation. Therefore, the oils with a burial depth of less than 1100 m are characterized by higher density, higher viscosity, and depletion of light hydrocarbon fraction[16,17,18]. The significant uplift of stratum and erosion made the reservoired oil susceptible to be incorporative to the earth's surface organisms. Fossil evidence suggests that the region was inhabited by *Homo erectus* and a variety of mammals during the Middle Pleistocene about 0.5 Mya ago[19]. It is generally believed that the skeletal morphology of *Homo erectus* is characteristic of both modern humans and apes. By now, the region has a humid and semi-humid continental climate, characterized by abundant rainfall, dense forests, and numerous rivers. It has been densely populated since the arrival of modern agricultural civilization.

The oil samples underwent geochemical analysis to assess their maturity, secondary alteration, sedimentary environment, and organic matter input of their relative source rocks. The results show that these oils and related source rocks have a low thermal maturity. The organic matter source of source rocks for the crude oil is dominant algae and the contributions of bacteria and terrigenous higher plants are minor. The source rocks of these studied crude oil samples were deposited under a high-salinity and strong-reduced environment with a stratified water column.

The GMs work as indicators of life

In this study, the following chemical molecules, including *n*-alkanes, acyclic isoprenoids (pristane and phytane), β-carotene, tricyclic terpanes, hopanes, gammaceranes, and steranes have been detected as the petroleum biomarkers (Fig S1). Generally, these GMs are difficult to degrade in the environment and are soluble in oil but insoluble in water. It can be assumed that from the source rocks in the Paleogene ages to the subsequent migration, charging, and accumulation of oil reservoirs, it is difficult for the GMs to return to the environmental water. This is because the adsorption capacity of the solid phase of the soil and rock is very limited to these molecules, so most of them are still preserved in the oil. Moreover, when crude oil encounters organic residues in the environment, the GMs can be absorbed into the reservoir.

GMs originating from different species show an overlap distribution pattern; however, GMs

originating from specific species indicate specific depositional environments. For example, β-carotene indicates cyanobacteria or algae origin, and 3β-methylhopanes represent the biological source of methanotrophic bacteria. In comparison, gammacerane in oil and sedimentary organic matter was derived from tetrahymanol in ciliates feeding on bacteria. The lower pristane/phytane ratios and higher relative abundance of gammacerane are generally related to anoxic, saline, and stratified water bodies[3]. The lower pristane-to-phytane ratio, higher cammacerane index, and the distribution of other GMs for the oils from the Nanyang Oilfield in this study indicate dominant input of aquatic organisms with less contribution of higher plants (Fig. S1) in the saline lacustrine basin. Our results are in general agreement with the findings of previous authors[6,7].

The distribution of tricyclic terpanes (TT) with the high frequency of the C_{23}TT indicates a marine or saline lacustrine depositional environment[20]. The distribution of other GMs confirmed that the oils are mainly derived from lower aquatic organisms with less contribution from higher plants (Fig. S1). Our results are in general agreement with the findings of previous authors[6,7]. To further illustrate these findings, relevant academic papers and paleogeographic maps were referenced in Fig. S2. It is crucial to understand that most GMs in oils suggest their domain or Kingdom but may not differentiate their specific species. For instance, the hopanes indicate the source of prokaryotes, while steranes are the origin of eukaryotes. Hence, it is important to gather solid evidence, such as original (*in situ*) DNA from marine organisms in the source rocks, while ruling out the possibility of paeDNA and preDNA contamination. This will confirm whether local marine intrusions were involved in source rock formation.

The prevailing belief in the academic community is that most petroleum molecules originate from lower plants and plankton, a belief consistent with our findings[1,2,3,21]. Since the early Paleogene before the formation of Hetaoyuan Eh3 shales, angiosperms have been the dominant plant group. Approximately three thousand years ago, the surface of the oil field was transformed into farmland. Today, the once expansive local lake area has nearly vanished, and the prevailing vegetation primarily consists of terrestrial higher plants. Our results suggest that most of the GMs detected belong to the original (*in situ*) GMs of the source rock (oriGMs). Given the lower contribution of higher plants over time, a question arises: Are these higher plant GMs from the source rocks or later-age species, particularly recent/present environmental ones? We suggest that there are three potential sources for these GMs: original (*in situ*) organisms in the source rocks, ancient/post-depositional environmental species (paeGMs), and recent/present environmental organisms (preGMs), such as terrestrial higher plants in the current environment, all of which could potentially enter the oil reservoir. However, the GMs derived from higher plants do not dominate the total GMs, so they enter the reservoir very slowly. Consequently, the GMs exhibit a general trend of "more original molecules and fewer later ones". As a result, we have developed hypothetical patterns representing changes in petroleum genetic markers (Fig. 1)

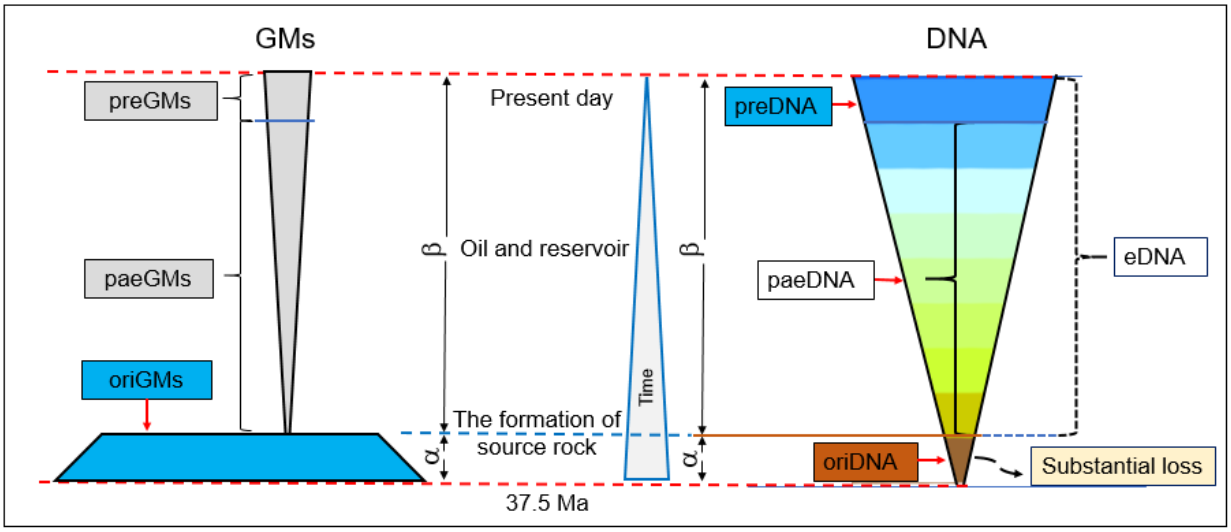

Fig 1. The preservation patterns of two types of biological indicators

Extraction of DNA information by the "mega screening method"

The 3,159,020 pDNA sequences were categorized and further analyzed using the "mega screening method" to extract biological information (see Materials and Methods). They originated from different species and sources and can be broadly divided into oriDNA, paeDNA, and preDNA.

(1) Establishing lineage subsets of pDNA sequences

Non-discriminatory alignment without predefined target genomes: Using the "minimum E-value model," over 3.15 million sequences (TS) were aligned with the NCBI database (see Materials and Methods). They were then divided into subsets by lineage, with 51.29% (1,620,178 fragments) assigned to the bacterial genome and 33.16% (1,047,532 fragments) to the human genome. 3.85% (121,735 fragments) were assigned to the fungal genome and 0.29% (9,252 fragments) to the avian genome. Non-primate mammals and non-human primates accounted for 0.15% (4785 entries) and 1.11% (35108 entries), respectively. Algae entries accounted for just 5, with a negligible proportion of the total, and no red or brown algae were found (Fig. 2). These results are not entirely consistent with the expected findings of previous authors on this point[6,7]. The GM markers in Fig. S1 suggest that Nanyang crude oil is primarily formed by the decomposition of lower plants and plankton, aligning with the prevailing view of the current academic community[1,2,3,21]. If we consider the DNA of lower plants and plankton as the main source of petroleum oriDNA and the DNA of other species as the main source of eDNA, then the pDNA data above does not highlight this view.

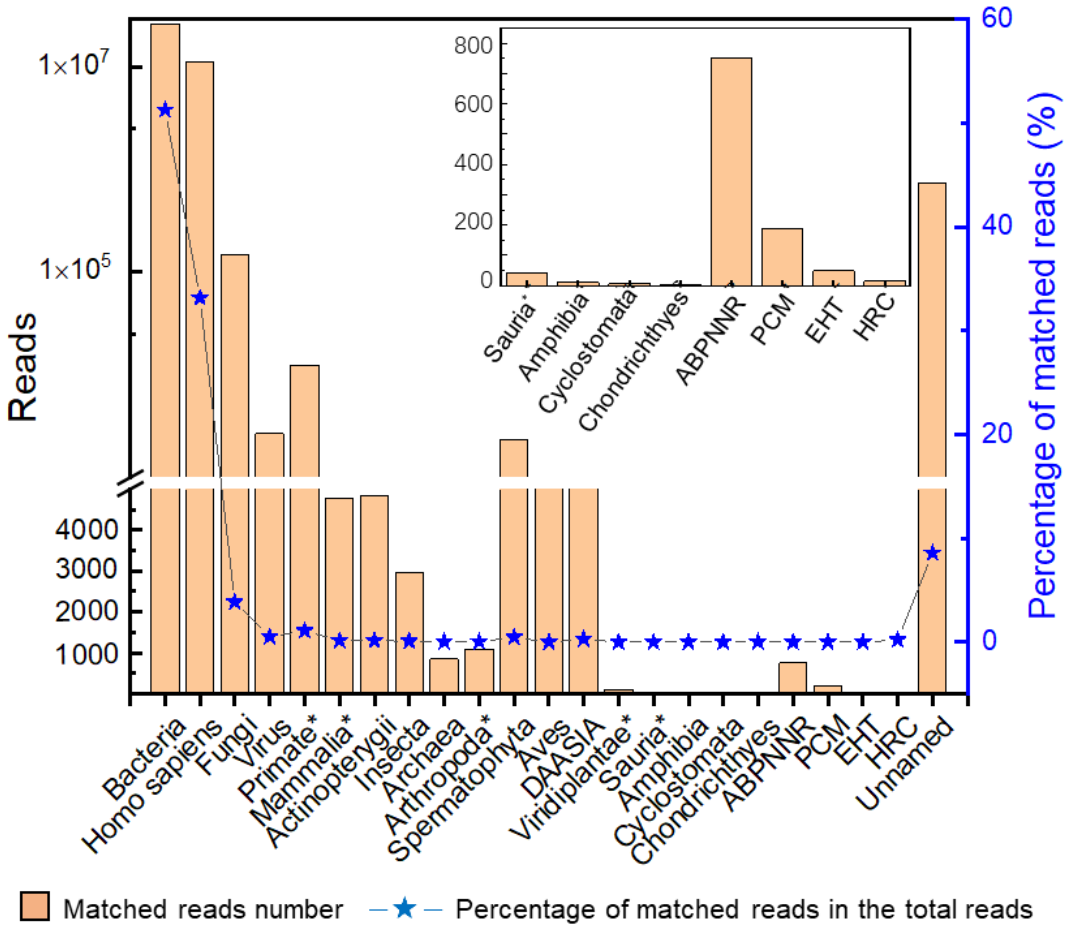

Fig 2. An analysis of pDNA sequences corresponding to different lineages

The bars represent the number of pDNA sequences that match a lineage, while the lines represent the percentage of pDNA fragments in that lineage compared to the total number of pDNA fragments. The abbreviations used are as follows: ABPNNR (Rotifera, Nemertea, Nematoda, Platyhelminthes, Annelida, and Bryozoa), PCM (Porifera, Cnidaria, and Mollusca), EHT (Hemichordata and Tunicata), HRC (Haptophyta, Rhodophyta, Cryptophyceae), and DAASIA (Discoba, Apusozoa, Amoebozoa, Sar, Ichthyosporea, and Apicomplexa). It's important to note that Viridiplantae* excludes Spermatophyta, Arthropoda* excludes Insecta, Sauria* excludes Aves, Primates* excludes *Homo*, and Mammalia* excludes Primates*, as detailed in Table S1.

(2) Identify the unique lineage source of each sequence

We employed the "MS mode" (see Materials and Methods) to determine the specific genome(s) that each pDNA sequence corresponded to and identify its distinct lineage origin. Due to the lack of genome data for certain lower species, some pDNA sequences could not be genealogically determined. Through this approach, 402 pDNA sequences have been successfully identified and categorized into nine groups based on their morphological characteristics and habitat (Table 1). These sequences originated from lower plants (Group G), marine organisms (Group M), primates (Groups C, P, and T), and birds (Group L). Although these sequences cannot trace major oil-forming species, they

can provide valuable data reflecting local geology, history, and evolution. They represent an unexplored reservoir for investigating local geological and species evolution.

Table 1. Summary of the pDNA sequences matched genomes

Group	Target	ID	Lineages	pDNA Count	Affinity	
					Average	SD
G	Plastid + Nuclear DNA	1G-8G	*Chlorophyta, Streptophyta*	8	90.23%	16.05%
H	Primate mtDNA	1H-40H	*Home sapiens*	40	96.12%	15.44%
N	Non-primate mtDNA	1N	*Gallus gallus, Galliformes*	1	99.32%	-
		2N	*Sus scrofa, Artiodactyla*	1	99.48%	-
		3N-5N	*Mylopharyngodon piceus, Cypriniformes*	3	100.00%	-
T	Primate nuclear gene	1T-275T	*Pan troglodytes*	275	96.91%	4.47%
P		21P-28P	*Pan paniscus*	8	98.23%	2.15%
		29P-36P	*Gorilla gorilla*	8	98.42%	2.03%
		6P-20P	*Cercopithecidae* (Old World monkeys)	15	94.31%	2.42%
		2P-5P	*Platyrrhini* (New World monkeys)	4	94.31%	5.80%
		1P	*Carlito syrichta* (*Tarsiiformes*)	1	70.67%	-
C		1C-9C	Primate protein-coding sequences	9	92.11%	4.04%
A	Nuclear gene of non-indigenous organism	1A-7A	*Bos mutus, Artiodactyla*	7	93.67%	5.51%
		8A-15A	*Meleagris gallopavo, Galliformes*	8	92.06%	8.02%
M	Nuclear gene of marine organisms	1M	*Scyliorhinus canicula, Chondrichthyes*	1	50%	-
		2M-5M	*Mercenaria mercenaria, Mollusca*	4	70.89%	17.23%
		6M-10M	Marine fishes, *Actinopterygii*	5	57.79%	18.61%
L	Others	1L-4L	*Matches with very low similarity*	4	< 50%	-

Note: The sequences between 7C and 28P, and between 8C and 4P, are identical.

(3) Explore local geological changes and species evolution

Genomic Analysis of Lower organism DNA: The analysis revealed three pDNA matches to Chlorophyta, and five pDNA matches to the Streptophyta genome, with two matching the Klebsormidium genome and three matching the Bryophytes genome from the pDNA data (Table S1). Notably, sequences ID_1G, ID_2G, and ID_3G appear distinct from known genomes and may belong to aDNA. With numerous unsequenced species of algae, it is currently not feasible to determine their specific species affiliation. Furthermore, the rapid evolution of genomes in lower organisms makes it difficult to determine if these sequences originated from ancient times. As a result, it remains uncertain whether they belong to oriDNA, paeDNA, or preDNA derived from post-source rock formation. Nevertheless, these findings strongly suggest that DNA from the original lower species that formed the hydrocarbon source rock has been largely lost. Scholars found morphological evidence of marine invasion in the core samples from the Hetaoyuan Formation (37.5 Ma) in the Biyang Depression, which

contain remains of red algae (Rhodophyta) and brown algae (Ochrophyta)[6,7]. Despite this, our data did not yield any DNA from these algae. One possible explanation is that the organisms are too small, which makes them susceptible to rapid dissolution and rupture upon entering the reservoir, releasing DNA. In addition, it is known that pDNA in the reservoir is always lost over time.

Detection of marine organism DNA signatures: The analysis identified DNA sequences containing genetic signatures of marine organisms, including Chondrichthyes, Actinopterygii, Tunicata, and Mollusca, as shown in Table S2, indicating the occurrence of local marine transgressions. Four sequences correspond to ray-finned fish species, with two sequences attributed to Scombriformes and two to Eupercaria. These fish species emerged during the mid-to-late Cretaceous period[22], suggesting that at least one marine invasion occurred after this era. While the pDNA evidence cannot determine the timing of these transgressions, such as whether they occurred in the Paleocene (approximately 37.5 Ma) or at some other time, it also cannot ascertain the location and stratigraphy of the hosts of the DNA fragments or whether they originated from other sources. Nonetheless, the current evidence serves as circumstantial support for Xia's view[6].

DNA evidence of ancient species migration and continental drift: Sequences associated with non-native species, such as yaks in the northern Himalayan foothills, turkeys in the Americas, and kiwi birds in the South Pacific Islands, were discovered (Table S3). It is hypothesized that these species are descendants of ancient Pangaea ancestors who migrated and reproduced in their local habitats due to continental drift. Their eventual disappearance left molecular evidence underground, offering new insights into the theory of continental drift.

Unveiling species transitional forms in avian ancestry: Generally, aDNA with lower-affinity values (see Materials and Methods) can serve as DNA fossils, providing valuable information about species evolution. For example, the ID_1L sequence shows some degree of similarity to several bird genome sequences, with an affinity value of 61.91 to the mRNA transcript sequence of *Calidris pugnax*. This suggests that the sequence may have originated from its ancient avian ancestor, indicating its ancient identity (Table S4). It comprises the protein-coding sequence of exons 3 and 4, along with a portion of exon 5, without any intronic insertions between the exons, hinting that it may function as a pseudogene. Moreover, the sequence can be divided into two distinct parts: the initial 25 base pairs upstream closely resemble the intergenic sequence of the non-native reptile *Podarcis lilfordi* and the remaining part bears similarity to various bird species (Fig. 3). The "composite" nature of this sequence suggests that the avian and reptilian genes originate from the LCA genome. Our analysis indicates that this DNA fragment does not align with any known bird DNA sequence, suggesting that the host organism may have become extinct or that its descendants have undergone significant genomic divergence due to long-term evolution.

The Late Mesozoic strata beneath the Hetaoyuan Formation are believed to have existed during the period of dinosaur breeding and the emergence of primitive birds, likely during the Cretaceous period or later. Although DNA fossils can reveal the time of origin of DNA sequences, they cannot provide evidence of the historical existence of the host organism. Therefore, it is challenging

to determine whether this sequence belongs to the oriDNA or eDNA from a certain period. To address this question, additional information is needed, such as the host's life cycle, the location of the fossilized body, and the specific stratum in which it was found. Importantly, the pDNA sequence exhibits both dinosaurian and avian characteristics, which could provide the first molecular evidence of transitional species within Darwin's proposed evolutionary spectrum. This finding also supports the theory that the Nanyang Basin served as an ancestral habitat for birds.

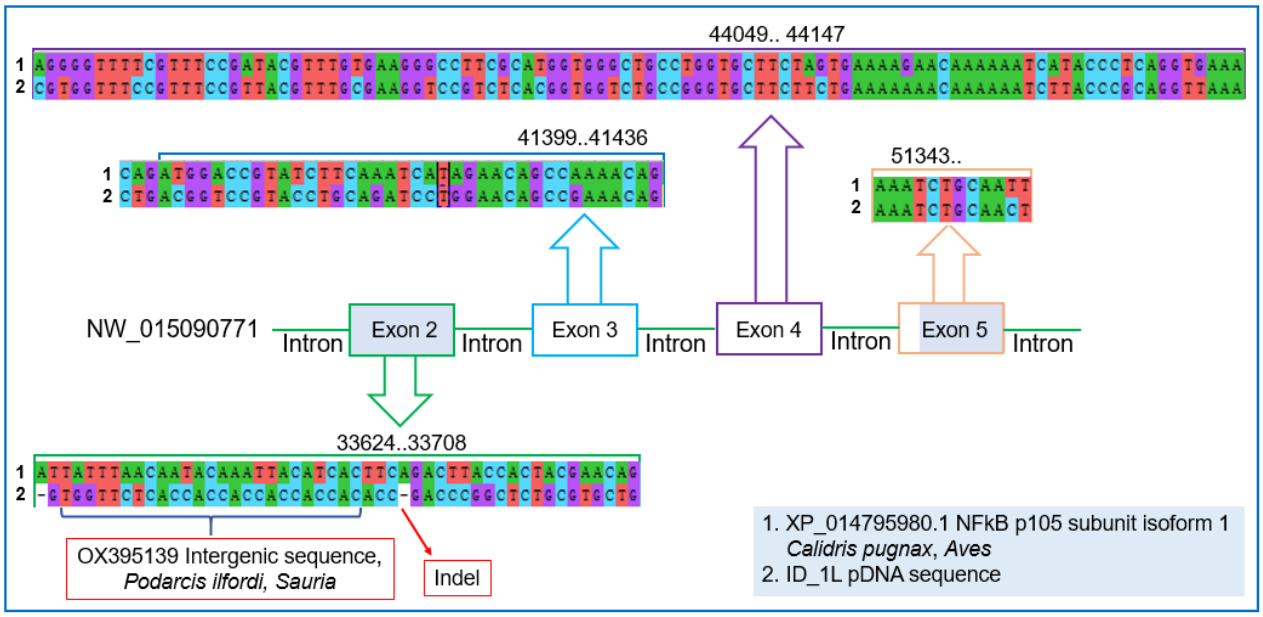

Fig 3. The Schematic diagram of *NFκB* p105 subunit-like DNA sequence

We used MS mode to analyze 1,047,532 sequences in a subset of humans and identify sequences that differed from the human mitogenome and nuclear genome (Fig. 2). Furthermore, we discovered five non-primate mitogenomic sequences in the dataset (Table S5), with the absence of non-human primate mitogenomic sequences. Among these, 846 sequences (affinity value: 100%) were identified to correspond to the human mitogenome and were not subject to further discussion. Additionally, there were 41 sequences displaying an affinity range of 91.48% to 99.5% with the human mitogenome. Given that NCBI houses comprehensive data on the entire human mitogenome, it can be inferred that these sequences are likely of ancient human origin. Nevertheless, potential individual misclassifications stemming from the approximate existence of a few unknown loci differences in modern humans cannot be ruled out (Table S6). Upon scrutiny of two specific sequences, ID_17H and ID_20H, it was discerned that they exhibited 100% coverage of each other, with distinct identities of 96.98% and 99.50%, respectively, implicating that they originate from disparate maternal lineages. The pDNA fragment ID_18H displayed an affinity value of 96.80%, and phylogenetic analyses indicated its positioning on a node with *Homo sapiens*, hinting at the possibility that its host may pertain to a previously unidentified species within the genus *Homo* (Fig. S3, above). Fragment ID_19H, with an affinity value of 83.85%, aligned with human ND4. Phylogenetic analysis suggested that ID_19H lies between the human and chimpanzee genera, potentially representing the earliest locally occurring species within the genus *Homo*; RelTime-ML analysis indicated a divergence time of 0.49 Ma from

human (Fig. S3, middle & bottom). Notably, the above pDNA sequences do not appear in the modern human genome, indicating that the ancient humans carrying these fragments are likely extinct.

Explore the existence of ancient hominids: Several pDNA sequences, exhibiting low affinities, were identified to align with nuclear genome sequences of non-human primates at varying frequencies: 1 in the Tarsiidae family, 4 in Platyrrhini (New World monkeys), 15 in Macaca (Old World monkeys), 8 in Gorillas, 8 in pygmy chimpanzees (Table S7), and 271 in chimpanzees (Table S8). Among the 271 sequences for chimpanzees, the second hit species were modern humans in 241 (88.93%), other primates in 29 (10.70%), and non-primates in the remainder. Given the presence of local fossil records of *Homo erectus* and the absence of that for non-human primates, we speculate that these sequences may be associated with ancient hominids, particularly considering the genomic connections between modern humans and primates.

Human burial practices effectively prevented DNA decomposition, leading to its accumulation in the environment at burial sites. Although other hominids also engaged in burial activities for their companions, the frequency, duration, and depth of these burials were not as significant as those observed in the species of *Homo*. Researchers have discovered *Homo erectus* fossils dating back to 0.5 Ma in Nanzhao County near the oil field[19]. Consequently, it is inferred that the ID_19H pDNA fragment discussed in this paper, along with most of the 271 pDNA fragments mentioned earlier, likely originated from unsequenced *Homo erectus*. Since these DNA fragments are significantly "younger" than the age of the source rock (37.5 Ma), it can be concluded that these sequences do not belong to oriDNA but to paeDNA.

The study suggests that the proposed *Homo erectus* shared some nuclear genes with apes. The difference between the mitogenomic sequences of the *Homo erectus* and modern humans is much smaller than that between the *Homo erectus* and chimpanzees. This indicates a lack of close mitogenomic connection between the *Homo erectus* and chimpanzees, suggesting that during hominin evolution, the mitogenome led the transition from apes to humans and the nuclear genome underwent a slower evolutionary process. Based on extensive research, it is widely accepted that mitochondrial genomic sequences exhibit a higher level of conservation compared to cell nuclear sequences. Previous studies revealed that ancient humans experienced severe population bottlenecks and limited mitochondrial haplogroup diversity[23,24,25]. Thus, we proposed that early species of the genus *Home* may originate from a small group of chimpanzees, possibly with a small number of females. They constituted small clan tribes with a female-centered structure and kinship ties. Later, reproductive isolation may have occurred due to karyotypic changes, preventing descendants from acquiring mitochondrial genomes from apes. Genetic isolation had likely happened before the time of the *Homo erectus*. Determining the exact time when this reproductive isolation occurred, while humans and apes separated, is a primary scientific question that always calls for further studies. The logical steps involved in utilizing the "mega screening method" to search for ancient *Homo* DNA are illustrated in Fig. S4.

DNA Migration and Preservation in Oil Reservoirs: Because DNA is primarily soluble in water and only slightly lipid-soluble, the lipid envelope surrounding the cells and organelles of living organisms aids in their entry into the oil. When the lipid envelope ruptures, DNA mixes with alkane molecules. After encountering water, the DNA leaves the oil phase because of water extraction on the water-oil interface. It is hypothesized that tissues, cells, or casein fragments of remains can enter the oil reservoir. In larger remains (e.g., animal and plant), the body's multiple layers of tissue can be encapsulated by the oil, resulting in a phenomenon, "Oil immersion preservation". This process allows the gradual release of DNA from tissue adjacent to the oil phase. When one layer of tissue is degraded, cells in other layers continue to release DNA until the organism is completely dissolved. Moreover, the fat tissue of the organism contains DNA, and when the fat liquefies, the DNA can flow into the oil reservoir, leading to a long-term, slow, and continuous release of DNA from the organism. Additionally, small organisms such as algae and plankton have cells that can rupture quickly, releasing their DNA into the reservoir. As a result, pDNA is continuously lost over time (Fig. 1). Here arises the interesting question: Are the eDNA (paeDNA and preDNA) hosts also involved in oil formation? While their contribution may seem small, the answer is positive.

Some scholars have observed a decrease in oriDNA and an increase in eDNA, mainly from microorganisms such as bacteria, in bone fossils[26, 27]. It is challenging to determine the exact source of paeDNA in the distant period before and after the formation of hydrocarbon source rocks due to the lack of knowledge about the paleoenvironment at ancient times. During the reservoir formation process, the oriDNA molecules undergo continuous degradation and loss, resulting in only a small amount of their remaining, while eDNA gradually infiltrates the oil. Human urban habitation, burial practices, and widespread cultivation of angiosperm crops have contributed to the ongoing introduction of DNA from human-associated species into the environment, serving as the primary source of preDNA. Utilizing the data in Fig. 2 and adding in our assumptions, the pattern of pDNA changes in oil was depicted (Fig. 1).

Below are a few instances of the "More new molecules and fewer old ones" phenomenon. First, the DNA from lower-life species that contributed to the formation of crude oil has nearly disappeared, as the DNA representing modern species has steadily increased (Fig. 2). Second, out of 886 human mitogenome sequences, 845 show a 100% affinity value with the human genome (preDNA, modern humans), while 41 sequences, possibly ancient human mitogenome sequences (paeDNA), exhibit variant divergences from the human mitogenome (Table S6). The sequence ID_19H displays the most significant divergence from the human mitogenome (Fig. S3) and is likely derived from an unclassified *Homo species*.

The challenge for Pääbo's aDNA Authenticity Criteria: Notably, the above pDNA sequences do not show the "deamination" often documented in aDNA studies by Pääbo's team. We hypothesize that two factors may contribute to this phenomenon. Firstly, the entry of eDNA into the reservoir necessitates specific conditions. One key condition is encapsulation in lipids, which fosters an anaerobic and anhydrous environment within the oil-forming materials and the reservoir. These unique environmental conditions can preserve DNA over time. Secondly, DNA sequences undergoing the

“deamination” may have been rejected by the “mega screening method”. Thus, Pääbo's primary criterion for authenticating aDNA is not applicable when conducting aDNA research with oil samples.

Methods

The geochemical survey

The oils were deasphalted with petroleum ether and then fractionated by column chromatography using a 2:1 (v/v) mixture of silica gel and alumina. The fractions were separated into saturated, aromatic, and resin components by sequential elution with 60 mL of petroleum ether, 40 mL of a dichloromethane/petroleum ether mixture (2:1, v/v), and 30 mL of a dichloromethane/methanol solution (93:7, v/v), respectively.

Gas chromatography (GC) analysis was conducted on the saturated fractions using an Agilent Model 6890 gas chromatograph, with a silica column (HP-1, 30 mm × 0.25 mm i.d.) installed. The oven temperature was programmed to increase from 100°C to 300°C for 10 minutes at a rate of 4°C/min. Helium was employed as the carrier gas.

Gas chromatography-mass spectrometry (GC-MS) analysis of the saturated and aromatic fractions was performed using an Agilent Model 6890 gas chromatograph coupled to a DB-5MS capillary column. The temperature program for the saturated fraction was as follows: the initial temperature was set at 100°C for 1 minute, followed by a ramp of 4°C per minute to 220°C, and then increased to 300°C at a rate of 3°C per minute and held for 15 minutes. The mass spectrometer was operated in electron impact mode, with an ionization energy of 70 eV and a scan range of 50-600 Da. The biomarker parameters of the saturated and aromatic hydrocarbons were determined by integrating the peak areas on the mass chromatograms.

DNA extraction, DNA library construction, sequencing, and NCBI nucleotide BLAST

The procedure was conducted in a BSL-2 industrial laboratory (Biosafety Level 2), following rigorous ancient DNA extraction protocols[29]. The process involved UV sterilization of all items used and the utilization of Geobio® DNA Remover (Jiaxing Jiesai Biotechnology Co., Ltd., Jiaxing, China) on clean benches and solid equipment. It is important to note that this laboratory has never been exposed to fish, non-human primates, or plant samples before this work.

To extract DNA from petroleum samples, we followed the recommended protocol of the Geobio® DPPS-PMNP DNA extraction kit developed by Jiaxing Jiesai Biotechnology Co[30]. Due to various long-term factors, some pDNA molecules are cross-linked with other macromolecules, forming complexes. To estimate the extract concentration, we measured the optical density (OD) at the UV 260 nm peak of the DNAmix, containing DNA complexes and free DNA molecules. We performed the extraction experiments on the samples five times, and the results were consistent (Table S9), providing sufficient DNA (Table S10). We used NEB PreCR®Mix for DNA repair.

The NGS DNA library construction followed the SSDLSP standard procedure (Sangon Biotech, CORP). Sequencing was conducted on Illumina NovaSeq 6000 (S-4 Reagent Kit) using the “Dual index sequencing” standard procedure, and the reads met the quality control standards with satisfactory quality distribution. Sometimes, it's necessary to manually remove extra adapter sequences at both

ends of the read to obtain accurate information. After merging the forward and reverse reads, we performed local NCBI Nucleotide BLAST, Size 11, using the database (Version 5, Nucleotide sequence Database), identifying the amount of 3,159,020 valid sequences. The sequences were classified according to their biological characteristics, and their lengths were measured (Table S11). It is important to note that the processes of secondary oil collection may inadvertently lead to the loss of many aDNA fragments.

The "mega screening method": In the first step, we employed a "minimum E-value mode" to group sequences by aligning fossil DNA with all known sequences from the NCBI database (Version 5, Nucleotide Sequence Database). We applied an E-value cutoff below 1E-07 to filter for qualified sequences (QS) and identify the best match at a specific taxonomic level. In the second step, we used the "MS mode" by selecting "exclude" option to conduct a search that omits the top result. If the E-value difference between two search results at the same taxonomic level (species, genus, family, order, and class) exceeds 1E-02, it suggests that the test sequence (query) belongs to the species identified in the first search result, suggesting a unique origin. Conversely, if the difference does not surpass the established threshold (1E-02), the sequence is shared between two taxa, rather than originating from a single one. In this scenario, we cannot determine the true source of the sequence, making it impossible to confirm whether the fragment belongs to the target species or irrelevant environmental species. Thus, we had to forgo any further analysis of the sequence.

Additionally, the "mega screening method" can be employed to reveal relationships among similar sequences discovered across different genomic organisms, providing valuable insights into the patterns of genome evolution.

The whole process involves the following steps:

1) Nucleotide Blast is employed to align sequences to the entire NCBI database without any limitation.

2) Creation of subsets for sorting the sequences: Using the "minimum E-value mode," results are matched and sequences are grouped into the appropriate subsets, determining the total number of sequences (TS). Subsequently, an E-value threshold of < 1E-07 is established to filter for qualified sequences (QS).

3) The subset primarily containing ancient DNA (aDNA) is selected.

4) Each sequence within this subset is screened individually using the "MS mode" to identify those originating from a single lineage, categorizing the “unique” sequences.

5) Considering various factors, including local climate shifting, geological changes, and evolutionary principles of the host species, we can ascertain if the sequence is oriDNA.

Sequence affinity (Affinity): The metrics we set up to assess the similarity relationship between subject sequences and hit genomes were obtained by multiplying the two values of Identity and Cover obtained from NCBI BLAST and then converted to percentages (Identity × Cover × 100%). Mutations

in the genomic and mitogenomic sequences occur continuously during the evolutionary process, leading to significant sequence differences between ancient and modern species. However, this variation is smaller in conserved sequences than in non-conserved sequences. A high-affinity value indicates that the sequence is highly coherent and related to the genome of the modern species, which is likely to be either a conserved sequence or a modern eDNA. A low-affinity value indicates a low similarity, and the sequence may be from either a distantly related species or other modern species that have not yet been sequenced. It is necessary to consider various factors, including the sequence composition and conservation, host species information, host genome sequencing data, etc., rather than relying on the Affinity value alone, when determining whether a sequence is an oriDNA.

Data and software availability

The sequencing data have been deposited in the NCBI BioProject (PRJNA1091869). The pDNA sequences utilized in this manuscript are listed in the Appendix (Table S12). In this study, MEGA11 was used to construct the RelTime-ML local timetree[31].

References

1. Tissot, B. P. & Welte, D. H. Petroleum Formation and Occurrence. (Springer-Verlag Berlin Heidelberg GmbH, New York. 1984).
2. Hunt, J. M. Petroleum Geochemistry and Geology, 2nd Edition. (W.H. Freeman and Company. 1996).
3. Peter, K. E., Walters, C. C. & Molodwan, J. M. The Biomarker Guide, 2nd Edition. (Cambridge University Press. 2005).
4. Philp, R. P., et al. A geochemical investigation of crude oils and source rocks from Biyang Basin, China, Organic Geochemistry. 18, 933-945 (1992).
5. Xia, L. et al. Organic geochemistry, petrology, and conventional and unconventional hydrocarbon resource potential of Paleogene saline source rocks in eastern China: The Biyang Sag of the Nanxiang Basin, Marine and Petroleum Geology. 101, 343-354 (2019).
6. Xia. L. et al. How marine incursion influence the quality of lacustrine source rocks: The Paleogene Nanxiang Basin, eastern China. AAPG Bulletin. 103, 1071-1096 (2019).
7. Song Y.et al. Salinized lacustrine organic-rich shale influenced by marine incursions: Algal-microbial community, paleoenvironment and shale oil potential in the Paleogene Biyang Depression, East China, Palaeogeography, Palaeoclimatology, Palaeoecology. 580, 110621 (2021).
8. Su, A. et al. Integrated fluid inclusion analysis and petrography constraints on the petroleum system evolution of the central and southern Biyang Sag, Nanxiang Basin, Eastern China, Marine and Petroleum Geology. 118, 104437 (2020).
9. Orlando, L., Gilbert, M. T. & Willerslev, E. Reconstructing ancient genomes and epigenomes. Nat. Rev. Genet. 16, 395-408 (2015).
10. Vahdani, M., Sahari, M. A. & Tanavar, M. Quantitative and qualitative analysis of three DNA extraction methods from soybean, maize, and canola oils and investigation of the presence of genetically modified organisms (GMOs). Food Chem (Oxf). 8, 100201 (2024).
11. Gu, M. et al. Study on detection of soybean components in edible oil with ladder-shape melting temperature isothermal amplification (LMTIA) assay. Anal Methods 15, 581-586 (2023).
12. Duan, Y. et al. An optimized procedure for detection of genetically modified DNA in refined vegetable oils. Food Sci Biotechnol. 30, 129-135 (2021).
13. Su, T. et al. Development of nucleic acid isolation by non-silica-based nanoparticles and real-time PCR kit for edible vegetable oil traceability. Food Chem. 300, 125205 (2019).
14. Ren, J. The continental tectonics of China. J. Southeast Asian Earth Sci. 13,197-204 (1996).
15. Luo, S., Men, F. & Chen, X. The quality of parent rocks of lower Eh3 formation in source Area of Biyang Depression. J. Oil Gas Technol. 31, 62-6471 (2009).
16. Dong, Y. et al. Seismic geomorphology study of the Paleogene Hetaoyuan Formation, central-south Biyang Sag, Nanxiang Basin, China. Marine and Petroleum Geology, 64, 104-124 (2015).
17. Li, J.et al. Effects of paleoenvironment on continental shale oil enrichment and producibility in the Biyang depression. AAPG Bulletin. 106, 2043–2071 (2022).
18. Wang, H. et al. Henan oilfield shallow thin-layer heavy oil thermal chemical vapor huffing technology. Pet. Geol. Eng. 28, 132-134 (2014).

19. Qiu, Z. et al. A human fossil tooth and fossil mammals from NanZhao, Henan. Acta Anthropol. Sinica 1, 109-117 (1982).
20. Xiao, H., Li, M. J. & Nettersheim, B. J. Short chain tricyclic terpanes as organic proxies for paleo-depositional conditions, Chemical Geology, 652, 122023 (2024).
21. Schobert H. Chemistry of fossil fuels and biofuels. (Cambridge University Press, Cambridge, pp. 103-130. 2013).
22. Hughes, L. C. et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. PNAS 115, 6249-6254 (2018).
23. Hu, W. et al. Genomic inference of a severe human bottleneck during the Early to Middle Pleistocene transition. Science 381, 979-984 (2023).
24. Chan, E.K.F., et al. Human origins in a southern African palaeo-wetland and first migrations. Nature. 575, 185-189 (2019).
25. Pontzer, H. Overview of Hominin Evolution. Nature Education Knowledge 3, 8 (2012).
26. Liang, R. et al. Genome-centric resolution of novel microbial lineages in an excavated Centrosaurus dinosaur fossil bone from the Late Cretaceous of North America. Environmental Microbiome 15, 8 (2020).
27. Miao B, et al. Assessment of contaminants associated with gold-standard ancient DNA protocols. Science Bulletin 68, 5-9 (2023).
28. Briggs, A. W. et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38(6), e87 (2010).
29. Gilbert, M. T. P., et al, Assessing ancient DNA studies. Trends in Ecology and Evolution. 20(10):541-544 (2005)
30. Su, T. et al. Development of nucleic acid isolation by non-silica-based nanoparticles and real-time PCR kit for edible vegetable oil traceability. Food Chem. 300, 125205 (2019).
31. Mello, B. Estimating TimeTrees with MEGA and the TimeTree Resource. Mol. Biol. Evol. 35, 2334–2342 (2018).

Acknowledgments

We extend our gratitude to the "Chao-Xiang Talent Program" foundation of Haining City, the support that covers the costs of petroleum DNA extraction and DNA sequencing from 2018 to 2022. The collection of petroleum samples from 2018 to 2021 received support from the Key Research & Development Special Program of Henan Province (231111111200), with the indispensable help of numerous workers and administrative staff from Nanyang oilfield (Sinopec Corp.), including Ms. Yu Liu, Mr. Pin-Lei Cai, and Mr. Peng Wei. Special thanks to Haining BoShang Biotechnology Corporation and Sangon Biotech (Shanghai) Corporation for their collaborative support in the success of this project, especially in the construction of the NGS DNA library using the SSDLSP standard procedure and sequencing. We also appreciate the contributions of Mr. HAN-Yu Zhu in creating a paleogeographic map and the assistance of Ms. Yun-Li Wang and Ms. Yue Zhang in processing the data. We extend our heartfelt appreciation to the following individuals for their invaluable contributions to this project: Dr. Hua-Jun Zheng from the Shanghai Institute for Biomedical and Pharmaceutical Technologies, Dr. Jun-Ye Ma from the Nanjing Institute of Geology and Paleontology, at the Chinese Academy of Sciences, Dr. Alida M. Bailleul from the Institute of Vertebrate Paleontology and Paleoanthropology, at the Chinese Academy of Sciences, Dr Shun-Ping He from the Institute of Hydrobiology, also at Chinese Academy of Sciences. Their stimulating discussions regarding the manuscript have greatly enriched our work.

Author contribution

The study was conceptualized and designed by W.Q.Z. and Y.Q.G., Y.Q.G., T.C.Q., and W.Q.Z. supervised the transportation of oil samples. M.J.L. performed the geochemical marker analysis. Geographical and paleogeological data analysis, as well as map creation, were conducted by F.L. and H.Y.Z. Petroleum DNA extraction from the samples was carried out by W.Q.Z., C.Y.Z., and J.Y.Y. Alignment between the sequences and present genomes was completed by Z.Y.G. and M.J.C. DNA laboratory analysis, interpretations, taxonomic profiling, and annotation were conducted by Z.Y.G., Z.Y.T., S.J.Z, L.J.Z, and W.Q.Z. The statistical analyses were performed and finished by G.Q.C., X.G.Z., R.C., and J.H.Z. Phylogenetic analyses of mitogenomic DNA sequences were performed by J.H.G., under the supervision of G.Q.C. and W.Q.Z. Figures were designed and finished by R.C., J.H.Z., L.Y.Q., and M.R.S. The manuscript was written by W.Q.Z. and corrected by Z.Y.T., M.J.L., and L.Y.Q. The Project Coordinator was T.C.Q.

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

All data are available in the main text or the supplementary materials.

Supplementary Materials

Figs. S1 to S4; Tables S1 to S12