**Mitochondrial genomes reveal the extinct *Hippidion* as an outgroup to all living equids.**

Clio Der Sarkissian1, Julia T. Vilstrup1, Mikkel Schubert1, Andaine Seguin-Orlando1,2, David Eme3, Jacobo Weinstock4, Maria Teresa Alberdi5, Fabiana Martin6, Patricio M. Lopez7, Jose L. Prado8, Alfredo Prieto7, Christophe J. Douady3, Tom W. Stafford1, Eske Willerslev1, Ludovic Orlando1\*.

1 Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark.

2 DanishNational High-throughput DNA Sequencing Centre, Øster Farimagsgade 2D, 1353 Copenhagen K, Denmark.

3 Laboratoire d’Écologie des Hydrosystèmes Naturels et Anthropisés, Université Claude Bernard Lyon 1, UMR CNRS 5023, 6 rue R. Dubois, 69622 Villeurbanne, France.

4 Faculty of Humanities, University of Southampton, Avenue Campus, Highfield, Southampton, SO17 1BF, United Kingdom.

5 Departamento de Paleobiología, Museo Nacional de Ciencias Naturales, CSIC, José Gutiérrez Abascal 2, 28006-Madrid, Spain.

6 Universidad de Magallanes, Instituto de la Patagonia, Centro de Estudios del Hombre Austral, Av. Bulnes 01890, Punta Arenas, CP 6200000, Chile.

7 Universidad Católica del Norte, Instituto de Investigaciones Arqueológicas y Museo Gustavo Le Paige, Calle Gustavo Le Paige Nº 380, San Pedro de Atacama, Chile.

8 INCUAPA, CONICET-UNICEN, Del Valle 5737, Olavarría, B7400JWI, Argentina.

\* Author for correspondence: Ludovic Orlando, Email: Lorlando@snm.ku.dk.

**Abstract**

Hippidions were equids with very distinctive anatomical features that lived in South America 2.5 million years ago (mya) until their extinction ~10 thousand years ago. The evolutionary origin of the three *Hippidion* morphospecies is still disputed. Based on paleontological data, *Hippidion* could have diverged from the lineage leading to modern equids prior to 10 mya. In contrast, a much later divergence date, with *Hippidion* nesting within modern equids, was indicated from partial ancient mitochondrial DNA sequences. Here, we characterised eight *Hippidion* complete mitochondrial genomes at 3.4-386.3-fold coverage by using target-enrichment capture and next-generation sequencing. Our dataset reveals that the two morphospecies sequenced (*H. saldiasi* and *H. principale*) formed a monophyletic clade, basal to extant and extinct *Equus* lineages. This contrasts with previous genetic analyses and supports *Hippidion* as a distinct genus, in agreement with paleontological models. We date the *Hippidion* split from *Equus* at 4.6-7.6 mya, suggesting an early divergence in North America prior to the colonization of South America after the formation of the Panamanian Isthmus 3.5 mya and the Great American Biotic Interchange.

**Keywords**

Ancient DNA, *Hippidion*, mitochondrial genomes, Great American Biotic Interchange.

**Background**

All contemporary equids belong to a single genus, *Equus* [1]. Their Most Recent Common Ancestor (MRCA) lived in North America ~4.0-4.5 million years ago (mya) [2], before spreading into other continents, and radiating into a wide range of extant and now-extinct forms. The ancestors of South American equids are believed to have left North America through inter-continental migrations after the formation of the Panamanian Isthmus approximately 3.0-3.7 mya [3, 4], which promoted the Great American Biotic Interchange. No American equid survived the Megafaunal Extinction ~10 thousand years ago (kya) [2].

Two main equine lineages are found in the South American fossil record from 2.5 mya [5]. The first defines a subgenus of its own, *Equus (Amerhippus)* (but see [6]), and shows horse-like anatomical features and genetic affinities [6]. Hippidiforms represent the second lineage, which includes three morphospecies: two genetically-similar generalists (*Hippidion principale* and *H. saldiasi*) [7] and one genetically-distinct high-altitude specialist (*H. devillei)* [7-9].

 Like *Amerhippus*, hippidiforms were stocky, but displayed very long nasal bones and extremely deep nasoincisival notches [5]. Paleontological data support affinities with the Northern American *Pliohippus* [3, 8, 10], which lived 6-14.5 mya. This considerably predates the time when hippidiforms first entered South America [3] implying an early divergence from the *Equus* lineage. However, recent molecular studies based on short mitochondrial sequences from 20 fossil specimens have suggested clustering of hippidiforms within *Equus* [6-7, 11-12], or possible affinities with *Dinohippus* [12], a North American lineage that lived 5-8 mya [13]. Therefore, the evolutionary origin of *Hippidion* remains unclear.

 Mitochondrial genomes (mitogenomes) have recently resolved the evolutionary tree underlying the main radiation of equids [1], which was later confirmed by whole-exomes [14]. This established mitogenomes as reliable phylogenetic markers for equids. We therefore used target-enrichment capture and next-generation sequencing to characterize mitogenomes of *Hippidion* and resolve its evolutionary origin.

**Methods**

*Molecular analyses*

All experimental procedures, including drilling (94-485 mg of sample powder), DNA library building and PCR amplification set up, were performed in ancient DNA facilities following published methods described in supplementary material S1 and table S1. As shotgun sequencing on Illumina HiSeq 2000 platforms revealed a limited number of mitochondrial hits, we performed target enrichment capture for equine mitogenomes, following [15] and building probes from mitochondrial PCR amplicons from living equids (supplementary material S1). Enriched libraries were amplified for an additional 12 cycles post-capture before being quantified on an Agilent 2100 BioAnalyzer, pooled and sequenced on an Illumina HiSeq2000 platform. DNA contamination was monitored by using mock controls during DNA extraction, library construction, PCR amplification and target enrichment. Contamination levels were also estimated from sequence reads using the Maximum Likelihood (ML) method from [16] and a database of 245 equine mitogenomes (tables S2).

*Sequence analyses*

DNA read processing (adapter trimming), mapping (read alignment, PCR duplicate removal, indel realignment) and damage analyses were performed using the PALEOMIX pipeline [17]. Seeding was disabled for mapping, a relaxed edit distance (-n 0.03) was used and reads showing mapping qualities <25 were disregarded. To ensure that all equine-like reads could be identified in the absence of a mitochondrial reference for *Hippidion*, we first recovered, for each sample, high-quality hits mapping uniquely against at least one equine mitogenome and/or partial sequences available for *Hippidion* (table S3). We then used a majority rule and a minimum depth-of-coverage of 2 to call a preliminary consensus sequence for each sample, which was used for read mapping with default parameters. A new and final consensus was then called using a majority rule, a minimum depth-of-coverage of 3 and filtering for base scores inferior to 30. Alignment statistics and DNA damage parameters were calculated based on a final mapping against each individual mitochondrial sequence.

*Phylogenetic analyses*

We constructed a 40-taxon mitochondrial dataset by aligning *Hippidion* mitochondrial sequences against a set of equine orthologous sequences, adding the white and black rhinoceroses as outgroups (table S4). The dataset was divided into 6 partitions (1st, 2nd, 3rd codon positions, rRNA, tRNA, control region CR), which were also concatenated into a merged dataset. modelgenerator was used to determine the best substitution model for each dataset/partition. We ran ML analyses on the merged dataset (with and without outgroups) using PhyML and treating the alignment as a single partition, as well as Approximate Unbiased (AU) and Shimodeira-Hasegawa (SH) topological tests against five trees differing in their *Hippidion* placement in CONSEL. We also performed Bayesian phylogenetic analyses assuming a Birth-Death serially-sampled tree model and treating partitioned datasets (with and without outgroups) as unlinked in BEAST and for each, fitting the best substitution model according to their Bayesian Information Criterion. Divergence dates were estimated in BEAST using a log-uncorrelated molecular clock model together with tip-sampling and assuming a time for the MRCA at 4.0-4.5 mya for *Equus* [2]*.* ML analyses of the CR under a GTR+I+8 model were run to compare our *Hippidion* sequences to those characterized previously (table S5). See supplementary material S1 and tables S6-8 for parameters and references to the programs used in the sequence analyses.

**Results and discussion**

We characterized eight *Hippidion* mitogenomes at 3.4-386.3-fold coverage (tables 1 and S9). Compared to shotgun sequencing, target-enriched libraries showed a 4.6-144.7-fold increase (average 41.7-fold, table S9) in the number of high-quality hits mapping uniquely against the *Hippidion* mitogenome. An additional sample (EQB) showed poor DNA preservation and resulted in limited mitogenome coverage (0.6-fold). Several lines of evidence support our data as authentic: i) presence of typical nucleotide mis-incorporation and ancient DNA fragmentation patterns [18] (figures S1-S9), ii) 6.9-32.5-fold higher cytosine deamination rates at overhangs than in double stranded DNA (table S10), iii) contamination levels ≤4.5% (per library) and 1.4% (per sample; table S11), iv) ML phylogenetic clustering with all *H. saldiasi* and *H. principale* CR sequences previously characterized (figure 1A). In addition, albeit only partially characterized due to limited coverage, the CR sequence obtained for sample JW251 clustered with the sequence previously published for this specimen [12].

ML analyses (figure 1B) and Bayesian phylogenetic inference rooted on rhinoceros outgroups revealed *Hippidion* as a monophyletic assemblage basal to all *Equus*. Similar phylogenetic placement was found using BEAST and excluding rhinoceros outgroups (data not shown). We furthermore used AU and SH topological tests to show that this topology provided a significantly better alternative than four other topologies where *Hippidion* was forced to nest within different *Equus* clades (*p-values* ≤ 0.001 and ≤ 0.044; figure S10), following all possible alternative topologies highlighted in [7]. Our analyses therefore contrast with previous results based on partial mitochondrial sequences [6-7, 11-12] and support paleontological models, which place *Hippidion* outside the range of variation of all extinct and extant *Equus* species. Additionally, we found that *H. devillei* clusters outside a paraphyletic assemblage consisting of *H. principale* and *H. saldiasi*. Whether the latter should still be described as two distinct morphospecies remains to be determined with nuclear data.

We estimated the divergence between *Hippidion* and *Equus* at ~4.6-7.6 mya (table 2), before *Hippidion* entered South America following the formation of the Panamanian Isthmus 3.0-3.7 mya [[5](#_ENREF_5)]. This is supported by fossils found in southern North America, which exhibit a more primitive nasal notch than South American hippidions, and could represent earlier hippidiforms [19].

**Conclusions and Perspectives**

In this study, we sequenced for the first time complete mitogenomes of *Hippidion saldiasi* and *H. principale*. We reject previous claims that *Hippidion* is nested within *Equus*, and estimate an early split between *Hippidion* and *Equus* 4.6-7.6 mya. Although the Equidae family flourished during the Miocene, all present-day equids originate from a MRCA that lived 4.0-4.5 mya [2]. With closest living relatives diverging 55 mya, the early evolutionary stages of the genus *Equus* are difficult to reconstruct based on present-day genetic information alone. The survival of *Hippidion* until 10 kya and the possibility to sequence whole genomes from extinct species [20] offer a unique opportunity to access a closer phylogenetic outgroup for equids, and thereby reveal the foundation for the *Equus* genetic makeup.

**Data accessibility**

*Hippidion* mitogenome sequences can be accessed on Genbank (accession numbers KM881671-KM881679).

**Authors’ contributions**

LO conceived and designed the study. CDS, JV, ASO and LO performed ancient DNA extraction, DNA library construction and amplification. CDS performed target-enrichment capture experiments. MS and LO performed sequence analyses. DE, JW, MTA, FM, PML, JLP, AP, CJD, TWS, EW and LO contributed samples, reagents and methods. All authors participated to the interpretation of the data. CDS, JV and LO wrote the article.

**Competing interests**

We have no competing interests.

**Acknowledgements**

We thank the Danish National High-throughput DNA Sequencing Centre; Tina Brand and Pernille Selmer Olsen for laboratory assistance; Philip Johnson for sharing R-scripts for estimating contamination levels.

**Funding statement**

This research was supported by the Marie Curie Career-Integration Grant (293845), the Danish Council for Independent Research, Natural Sciences, the Danish National Research Foundation (DNRF94), and the Lundbeck Foundation (R52-A5062). FM was supported by the FONDECYT 110822 and CD MAG0901 grants.

**References**

1. Vilstrup JT, *et al.* 2013. Mitochondrial phylogenomics of modern and ancient equids. *PLoS One* **8**, e55950. (doi:10.1371/journal.pone.0055950)
2. Orlando L, *et al.* 2013. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78. (doi:10.1038/nature12323)
3. MacFadden BJ. 1997. Pleistocene horses from Tarija, Bolivia, and validity of the genus †*Onohippidium* (Mammalia: Equidae). *J. Vert. Paleontol.* **17**, 199-218. (doi:10.1080/02724634.1997.10010964)
4. Ibaraki M. 1997. Closing of the Central American Seaway and Neogene coastal upwelling along the Pacific coast of South America. *Tectonophysics* **281**, 99-104. (doi:10.1016/S0040-1951(97)00161-3)
5. Alberdi MT, *et al.* 2005. Considerations on the Paper “Morphological Convergence in *Hippidion* and *Equus (Amerhippus)* South American Equids Elucidated by ancient DNA analysis". *J. Mol. Evol.* **61**, 145-147. (doi:10.1007/s00239-004-0216-3)
6. Orlando L, *et al.* 2008. Ancient DNA clarifies the evolutionary history of American late Pleistocene equids. *J. Mol. Evol.* **66**, 533-538. (doi:10.1007/s00239-008-9100-x)
7. Orlando L, *et al.* 2009. Revising the recent evolutionary history of equids using ancient DNA. *Proc. Natl. Acad. Sci. USA* **106**, 21754-21759. (doi:10.1073/pnas.0903672106)
8. Prado JL, Alberdi MT. 1996. A cladistic analysis of the horses of the tribe Equini. *Palaeontology* **39**, 663-680.
9. Borrero LA. 2009. The elusive evidence: the archeological record of the South American extinct megafauna. American megafaunal extinctions at the end of the Pleistocene: Springer. pp. 145-168.
10. Alberdi MT, *et al.* 1998. Late Pleistocene *Hippidion saldiasi* Roth, 1899 (Mammalia, Perissodactyla) from Calama, northern Chile. *Rev. Chli. Hist. Nat.* **80**, 157-171. (doi:10.4067/S0716-078X2007000200003)
11. Orlando L, *et al.* 2003. Morphological convergence in *Hippidion* and *Equus (Amerhippus)* South American equids elucidated by ancient DNA analysis. *J. Mol. Evol.* **57**, S29-S40. (doi:10.1007/s00239-003-0005-4)
12. Weinstock J, *et al.* 2005. Evolution, systematics, and phylogeography of Pleistocene horses in the New World: A molecular perspective. *Plos Biol.* **3**, 1373-1379. (doi:10.1371/journal.pbio.0030241)
13. Groves C, Ryder OA. 2000. Systematics and phylogeny of the horse. In: Bowling AV, Ruvinsky A (eds.) The Genetics of the Horse. CABI Publishing, UK and USA pp. 1-24
14. Jónsson H, *et al.* 2014 Speciation with gene flow in equids despite extensive chromosomal plasticity. *Published online.* (doi:10.1073/pnas.1412627111)
15. Maricic T, et al. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* **5**, e14004 (doi: 10.1371/journal.pone.0014004)
16. Fu Q, et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553-559 (doi:10.1016/j.cub.2013.02.044)
17. Schubert M, *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056-1082 (doi:10.1038/nprot.2014.063)
18. Dabney J, *et al.* 2013. Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* **5**, pii:a012567 (doi: 10.1101/cshperspect.a012567)
19. MacFadden BJ, Skinner MF. 1979. Diversification and Biogeography of the One-Toed Horses “*Onohippidium*” and “*Hippidion*”. Peabody Museum of Natural History, Yale University.
20. Miller W, *et al.* 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**, 387-390. (doi:10.1038/nature07446)

**Figure 1. Maximum likelihood phylogenetic reconstructions.**

A. Control region. B. Mitochondrial genome (rooted using rhinoceroses as outgroups, not shown). The *Hippidion devillei* clade is represented in brown. The other *Hippidion* clade including *H. saldiasi, H. principale* and *H. sp.*, is shown in red. In bold, *Hippidion* sequences generated in this study. See table 2 for node supports and Bayesian divergence estimates. New World Stilt Legged horse= extinct Pleistocene North American equid [12].

**Table 1. Sample and sequencing information.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Species** | **Location** | **Age (cal. BP)** | **#Reads** | **#Hits** | **Cov. (X)** |
| CDM4 | *H. saldiasi* | Cueva del Milodón, Magallanes, Chile | 13,575-13,810 | 18.6M | 37,151 | 186.6 |
| CE99208-31 | *H. saldiasi* | Cueva Escondida, Magallanes, Chile | 16,870-17,020 | 26.5M | 11,718 | 51.5 |
| AE57-17 | *H. saldiasi* | Cueva Nordenskjöld, Magallanes, Chile | 16,809-17,400 | 34.6M | 57,640 | 293.1 |
| CM66 | *H. saldiasi* | Cueva del Medio, Magallanes, Chile | LP | 36.1M | 75,624 | 386.3 |
| EQF | *H. sp.* | Cueva Baño Nuevo-1, Chile | 15,602-16,724 | 7.7M | 18,522 | 78.7 |
| EQB | *H. sp.* | Quebrada Opache, Calama, Chile | LP | 5.5M | 138 | 0.6 |
| CM-A37 | *H. saldiasi* | Cueva del Medio, Magallanes, Chile | 12,570-12,640 | 37.4M | 16,441 | 69.8 |
| JW251 | *H. principale* | Provincia de Buenos Aires, Argentina | LP | 47.1M | 1,028 | 3.4 |
| OH | *H. saldiasi* | Cueva del Milodón, Magallanes, Chile | 13,197-13,468 | 29.7M | 2,286 | 7.1 |

cal. BP = calibrated years Before Present. #Reads= Total number of reads; M= million; #Hits= Number of high-quality hits; Cov= Average depth-of-coverage, both when mapping against the sample’s mitogenome. LP = Late Pleistocene.

**Table 2. Clade phylogenetic support and divergence dates.**

|  |  |  |
| --- | --- | --- |
|  | **ML inference** | **Bayesian inference** |
|  | **Support** | **Support** | **Node Age (mya)** |
| **Node** | **CR** | **MG +O** | **MG -O** | **MG +O** | **MG -O** | **MG +O** | **MG -O** |
| A | NA | 1 | NA | 1 | 1 | 6.502[5.551-7.608] | 5.596[4.599-6.629] |
| B | 0.891 | 1 | 0.915 | 1 | 1 | 0.377[0.276-0.494] | 0.322[0.247-0.425] |
| C | 0.994 | 0.860 | 1 | 1 | 1 | (4.0-4.5)\* | (4.0-4.5)\* |
| D | 0.888 | 0.999 | 0.996 | 1 | 1 | 3.126[2.648-3.603] | 2.833[2.314-3.331] |
| E | - | 1 | 1 | 1 | 1 | 2.965[2.600-3.340] | 2.650[2.290-3.017] |
| F | 0.895 | NA | NA | NA | NA | NA | NA |

See figure 1 for a definition of nodes A-F. CR= Control Region. MG= Mitochondrial Genome. O= Outgroups. Node supports inferior to 0.80 are masked. NA = Not Applicable. Node ages correspond to the median of Bayesian posterior distributions (5% and 95% quantiles in brackets). \*Calibration point.