Mitochondrial Phylogenomics of Modern and Ancient Equids

Julia T. Vilstrup¹, Andaine Seguin-Orlando¹, Mathias Stiller², Aurelien Ginolhac¹, Maanasa Raghavan¹, Sandra C. A. Nielsen¹, Jacobo Weinstock³, Duane Froese⁴, Sergei K. Vasiliev⁵, Nikolai D. Ovodov⁶, Joel Clary⁷, Kristofer M. Helgen⁸, Robert C. Fleischer⁹, Alan Cooper¹⁰, Beth Shapiro², Ludovic Orlando¹*¹

¹Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. ²Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, California, United States of America. ³Faculty of Humanities, University of Southampton, Southampton, United Kingdom. ⁴Department of Earth and Atmospheric Sciences, University of Alberta, Alberta, Canada. ⁵Institute of Archaeology and Ethnography, Russian Academy of Sciences, Novosibirsk, Russia. ⁶Laboratory of Archaeology and Paleogeography of Central Siberia, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Novosibirsk, Russia. ⁷Centre de Conservation et d’Étude des Collections, Musée des Confluences, Lyon, France. ⁸Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington D.C., United States of America. ⁹Center for Conservation and Evolutionary Genetics, Smithsonian National Zoological Park, Smithsonian Institution, Washington D.C., United States of America. ¹⁰Australian Centre for Ancient DNA, School of Earth and Environmental Sciences, The University of Adelaide, South Australia, Australia

Abstract

The genus Equus is richly represented in the fossil record, yet our understanding of taxonomic relationships within this genus remains limited. To estimate the phylogenetic relationships among modern horses, zebras, asses and donkeys, we generated the first data set including complete mitochondrial sequences from all seven extant lineages within the genus Equus. Bayesian and Maximum Likelihood phylogenetic inference confirms that zebras are monophyletic within the genus, and the Plains and Grevy’s zebras form a well-supported monophyletic group. Using ancient DNA techniques, we further characterize the complete mitochondrial genomes of three extinct equid lineages (the New World stilt-legged horses, NWSLH; the subgenus Sussemionus; and the Quagga, Equus quagga quagga). Comparisons with extant taxa confirm the NWSLH as being part of the caballines, and the Quagga and Plains zebras as being conspecific. However, the evolutionary relationships among the non-caballine lineages, including the now-extinct subgenus Sussemionus, remain unresolved, most likely due to extremely rapid radiation within this group. The closest living outgroups (rhinos and tapirs) were found to be too phylogenetically distant to calibrate reliable molecular clocks. Additional mitochondrial genome sequence data, including radiocarbon dated ancient equids, will be required before revisiting the exact timing of the lineage radiation leading up to modern equids, which for now were found to have possibly shared a common ancestor as far as up to 4 Million years ago (Mya).

Introduction

The family Equidae, along with Rhinocerotidae and Tapiridae, is one of the three extant families of odd-toed ungulates (Perissodactyla, Mammalia). The Equidae are richly represented in the fossil record throughout the past 55 Million years (My), starting with dog-sized taxa from the North American Eocene [1]. Equid lineages then spread globally and experienced successive episodes of radiations and extinctions during the early Miocene, the late Miocene, and at the end of the Pleistocene [2], becoming adapted to a variety of environments with remarkable variations in body size [1].

While several dozen extinct equid genera have been described [2], all extant equid species are classified in the same genus, Equus. It includes the African wild ass, Equus africanus, which is the progenitor of the Domestic donkey, E. asinus [3], and three living species of zebra, all endemic to Africa. The zebras include the Plains zebra, also called Burchell’s zebra, Equus quagga (previously E. burchellii); Grevy’s zebra, E. grevyi; and the Mountain zebra, E. zebra (with two subspecies, E. z. zebra of South Africa, and E. z. hartmannae of Namibia and Angola) [4]. The genus also includes the Asian wild asses (subgenus Hemionus), E. hemionus and E. kiang, with various recognized subspecies of E. hemionus (e.g. E. h. kilan and E. h. onager). Together, the zebras, donkeys, and Asian asses make up the non-caballine lineages. The remaining extant lineages, the domesticated horses (Equus caballus) and wild populations sometimes referred to as E. ferus (Tarpan and/or the Przewalski horse E. przewalskii), are classified as caballine horses.

While the phylogenetic relationships among modern equids have received considerable attention in the past few years [3–14],
no phylogenetic study has been undertaken using complete mitochondrial sequences of all seven extant species of *Equus*. Complete mitochondrial genome (mitogenome) data is currently available for only three species: domestic (142 individuals) and Przewalski (seven individuals) horses, [5–7,13–17], one Tibetan Kiang [18], and one domestic donkey [19]. A nearly complete mitochondrial sequence has also been made available recently for one wild donkey individual [6]. Using second-generation sequencing technologies we generate complete mitogenome sequences from 2–4 individuals of all other living equid species. This results in a comprehensive dataset of 14 novel and complete mitogenomes comprising all of the taxonomic diversity of extant equids. Additionally we sequence complete mitogenomes from three extinct equid lineages: the New World Stilt-Legged horses (NWSLH), the Sussemionus (*E. ovoidus*), and the Quagga.

NWSLH refer to a group of equids that were endemic to North America. Despite having gracile limbs similar to the Asiatic wild asses, partial mitochondrial DNA sequences revealed no particular genetic affinity with hemiones, instead the NWSLH were nested within caballine horses [20]. Sussemiones (subgenus *Sussemionus*) consist of a lineage of Eurasian equids that survived until the Late Pleistocene, at least up to 46 thousand years ago (Kya) [21,22]. No particular genetic relationship with any of the living equid lineages was supported by partial mitochondrial sequences [22], and their evolutionary origin is still debated [21,22]. The Quagga, *E. quagga quagga*, a morphological variant of the Plains zebra found in South Africa, became extinct in the wild in the late 1800s [23], and is now commonly accepted as a southern variant of the Plains zebra [4,24].

Inferring the evolutionary relationships among extinct and living equids is an active area of research. So far, analyses have used mainly short fragments of mitochondrial DNA, with highly controversial results suggesting that previously defined lineages should be collapsed into single taxa [20,22,25–27]. In this study we aim to (1) resolve the phylogenetic relationships among all of the extant as well as three extinct lineages of equids, and (2) to investigate the evolutionary timing of lineage radiation within the genus *Equus*.

**Materials and Methods**

**Overall Experimental Design**

We isolated one partial and 17 complete mitogenomes representing all extant species within the genus *Equus* (14 modern samples), including three extinct lineages (four ancient samples) (Table 1). For five out of the 14 modern samples, mitogenomes were generated by PCR amplification followed by a combination of Sanger and GS FLX sequencing. The DNA extracts from the remaining nine modern samples and three of the ancient samples were converted into sequencing libraries and shotgun-sequenced using the Illumina HiSeq2000 platform. The mitogenome of the remaining ancient sample (NWSLH sample MS272) was recovered from a genomic library using the MySelect in solution target enrichment kit (MYcroarray, USA) and sequenced on the Illumina HiSeq2000 platform.

**Ethics Statement**

Tissue samples were collected by Hans Siegismund [28], or provided by the Smithsonian Institution (Washington D.C.); the Musée des Confluences (Lyon), the Russian Academy of Sciences, the Government of Yukon, or zoological gardens following official agreements with the Natural History Museum of Denmark in the framework of the COSE certificate DK003 of the host institution, which covers CITES-protected species. Permission was obtained from all museums and institutions to access the collections and all samples were on loan for scientific purposes.

**Modern Samples**

For PCR amplified mitogenomes, we collected 30–50 mg of hair or dried alcohol-preserved tissue and extracted DNA using the DNeasy Blood and Tissue kit (Qiagen, USA) following the manufacturer’s instructions except that 360 μl of Buffer ATL and 40 μl Proteinase K were used, and samples were left to lyse overnight at 55°C. We performed PCR amplification using three to six overlapping primer sets (Table S5) [16,17] as in Vilstrup et al. [29] (Table S4). We pooled purified PCR products from each sample, except for a 1.5 kb fragment, in equimolar amounts and nebulized for 1 min at 21 bar to get approximately 600 bp fragments. We then built the fragmented DNA samples into tagged dA-tailed libraries using NEBNext Quick DNA Library Prep Master Mix set for 454 (New England BioLabs, ref: E6090) using a final adaptor concentration of 9 nM. We then pooled the tagged libraries in equimolar amounts and distributed them on one eighth of a plate on the GS FLX. We followed a similar procedure for PCR amplicons from primer sets Pr1 and Pr2, which span the hypervariable region. All tagged reads were sorted, trimmed and mapped to both a donkey and a horse reference genome (Accession Nt, NC001788 and NC001640, respectively) using GS Reference Mapper with default parameters (454 Life Sciences). We allowed for a maximum of one mismatch during index sorting and a minimum size of 80 bp was required during sequence assembly. Consensus sequences were aligned to a reference using Sequencer v4.8 (Genes Code Corporation, Ann Arbor, MI, USA), and the resulting mitogenomes aligned by eye using SE-AL v2.0a11 Carbon (A. Rambaut, Univ. of Oxford). Sanger sequencing of the short 1.5 kb fragment was performed at the Macrogen facility (Seoul, South Korea) with several primers (Table S4).

For shotgun-sequenced mitogenomes, we collected 2 ml blood samples and extracted DNA using the QIAamp DNA Blood Midi kit (Qiagen, USA) following the manufacturer’s protocol except that blood digestion was performed with Proteinase K for 30 min at 70°C, and the final product was eluted into 150 μl EB buffer. We then sonicated the DNA extracts for seven cycles of 15 secs/ 90 secs (ON/OFF cycles) on mode High (H) using a Bioruptor XL (Diagenode, Belgium). We selected fragments of 200–300 bp using the EZNA Gel Extraction Kit (Omega Bio-Tek, ref: D2500) and converted each sample into an Illumina library using the NEBNext Quick DNA library prep Master mix set for 454 (New England BioLabs, ref: E6090). We amplified the libraries in a 50 μl volume reaction, with two parallel PCR set-ups per library using half (16 μl) of each DNA library. The final PCR reaction consisted of 5U Taq Gold (Invitrogen, Life Technologies), 4 mM MgCl2, 1 mg/ml BSA, 1 mM dNTPs, 1 μM of Primer inPE1.0, 20 nM of Primer inPE2.0, and 1 μM of an Multiplexing Index Primer containing a unique 6 nucleotide index tag (Illumina Inc.). PCR cycling conditions consisted of initial denaturation for 10 min at 92°C, followed by 10–15 cycles of 30 secs denaturation at 94°C, 30 secs annealing at 60°C, and 40 secs elongation at 72°C, and a final elongation step at 72°C for 7 min. We QIAquick (Qiagen, USA) purified each amplified library and then quantified them on an Agilent 2100 Bioanalyser. The libraries were then pooled in equimolar ratios, and sequenced in paired-end mode (2x100 bp) on the Illumina HiSeq2000 platform at the Danish National High-throughput Sequencing Centre.

We trimmed the Illumina reads using default settings in AdapterRemoval [30], except using a minimal read length of 25 bp, and aligned reads against both published mitogenomes and
those generated here via PCR amplification using BWA [31],
aborting the seed and relaxing the edit distance (option –n 0.05) as
suggested by Schubert et al. [32]. We removed reads that mapped
to multiple positions and with mapping quality scores <25 using
SAMtools [31]. We removed sequence duplicates using both read
start and read end coordinates for collapsed paired-end reads, and
we used only the read start coordinates for filtering potential
sequence duplicates for uncollapsed read pairs as implemented in
MarkDuplicates from the Picard package (http://picard.
sourceforge.net). A final alignment was generated for each
mitogenome and visually corrected for potential local misalign-
ments. Finally, we called the sequence of each mitogenome using
a consensus approach requiring a minimum base coverage of 2
and over 50% of sequence match among reads.

Ancient Samples and Museum Specimens
We extracted three ancient samples (JW328, ACAD2304, QH1)
and two museum-preserved dried tissue specimens (CGG10086
and CGG10096) at the Centre for GeoGenetics in laboratory
capabilities dedicated to the analysis of fossil material and geo-
graphically separated from laboratories with amplified PCR
products. We used three different extraction protocols depending
on the tissue type, and all extractions were accompanied by
appropriate controls.

We extracted the two bone samples JW328 (20 mg) and sample
ACAD2304 (464 mg) following the silica-based protocol described
in Orlando et al. [22]. We extracted the two tissue specimens
(CGG10086 and CGG10096) described in Gilbert et al. [33],
adding 2 mg/mL Protease K (New England Biolabs) to the
extraction buffer, which we digested overnight at 37°C, followed
by QIAquick purification and elution into 50 μL. We extracted
~50 mg of the hair sample (QH1) as in Gilbert et al. [33] except
digestion occurred at room temperature (RT) over three
days, after which undigested hair was removed and incubated at RT
for an additional three days with more digestion buffer. We
concentrated both digests to ~50 μL using 30 KDa centricons and
and purified the extracts using QIAquick columns, and eluted the
concentrated DNA into 50 μL after a five-minute incubation in
elution buffer at RT.

We built indexed Illumina libraries for samples JW328,
CGG10086, CGG10096, and QH1 using the NEBNext Quick
DNA Library Prep Master Mix set for 454 (New England BioLabs,
and used the NEBNext Library Prep Master Mix set for 454 (New
England BioLabs, ref: E6090) following the manufacturer’s instruc-
tions, except that
SPRI bead purification was replaced by Qiagen MinElute DNA
purifications. The libraries were amplified in two steps, as
described in Kampmann et al. [34]. We purified the resulting
libraries through QIAquick columns, quantified them using an
Agilent 2100 Bioanalyzer, and pooled them in equimolar ratios for
Illumina sequencing performed in single-read mode (75 bp) on an
Illumina HiSeq2000 platform at the Danish National High-
throughput DNA Sequencing Centre.

We built two independent DNA libraries using 21 μL each of
DNA extract for sample ACAD2304 using the NEBNext DNA
library prep Master mix set for 454 (New England BioLabs, ref:
E6070) and following Meyer and Kircher [35], except working
with 25 μL reaction volumes and using a final concentration of
0.5 μM Illumina multiplex blunted-end adaptors. At the end of
the final fill-in reaction, Bst Polymerase was inactivated following
20 min incubation at 80°C. We then amplified the two libraries
for 12 cycles using PCR conditions as described above for the
seven modern samples. We purified the PCR products through
QIAquick columns and re-amplified them in four parallel
reactions (5 μL each) for ten additional cycles. The four amplifi-
cation products generated per library were QIAquick purified,
quantified using an Agilent 2100 Bioanalyzer and pooled in
equimolar amounts before sequencing in single-read mode
(100 bp) on a HiSeq2000 Illumina platform at the Danish
National High-throughput DNA Sequencing Centre.

A second Quagga library was also enriched for mitochondrial
fragments following the procedure described by Maricic et al. [36]
and using 1.5 μg of library as input. Mitochondrial baits consisted
of a mix of fragmented, amplified mitogenomes of a Grevy’s zebra,
a Somali wild ass, and a Kiang (SF, LF and Primer1 primer sets;
Table S4). The captured library was purified through MinElute
column and 16 μL was amplified for 12 cycles in a 50 μL volume
under the same PCR conditions as the seven modern samples
using 0.2 mM postCap primer inPE1.0 and 0.2 mM post index
primer. The captured library was sequenced in single-read mode
(75 bp) on an Illumina HiSeq2000 platform at the Danish
National High-throughput DNA Sequencing Centre.

We assembled complete mitogenomes following the procedure
described for assembling Illumina reads from modern equid
specimens, except that for sample JW328 we also attempted to
map reads against previously reported partial mitochondrial
sequences of hippidiforms (accession numbers: GQ324601,
GQ324596 and AX152859) and the complete nucleotide sequence
of another NWSLH generated in this study following library
target-enrichment (see below). The authenticity of our ancient
DNA data was then assessed using the mapDamage package [37]
(Figure S3).

The NWSLH sample MS272 was processed separately in a
dedicated ancient DNA laboratory at the Pennsylvania State
University, USA. We powdered ~500 mg of bone using a
Mikrodismembrator (Braun) and extracted DNA using the
column based silica extraction protocol as described in Rohland
et al. [38]. DNA was eluted into 60 μL of 1×TET (1×TE
including 0.05% Tween20). We prepared an indexed Illumina
library using 15 μL of the ancient DNA extract according to the
protocol described in Meyer and Kircher [35] with 40 μL reaction
volumes. We amplified the library for 25 cycles using a unique
indexing primer, and subsequently purified the library using
the Agencourt AMPure XP PCR purification kit. We eluted the
library in 30 μL of 1×TET as described in Meyer and Kircher
[35]. We then enriched the indexed library for the mitogenome
using the MYSelect (now called MYbaits) target enrichment kit
from MYcroarray (USA) following the manufacturer’s instruc-
tions. Custom bait molecules were designed based on the modern
Equus caballus mitogenome sequence (horse genome version
equCab2.0 http://genome.ucsc.edu/cgi-bin/
hgGateway?db=equCab2 provided by The Broad Institute,
Cambridge, USA) at 2X tiling (where each base is covered by
two unique bait molecules). To provide enough input DNA for the
capture reaction, the library was amplified for another 30 cycles
using Phusion Hot Start polymerase (New England BioLabs)
and primers IS3_reamp.P5 and IS6_reamp.P7 [35]. The target
capture itself was performed to the manufacturer’s specifications.
We then amplified the enriched libraries once more using the same
Phusion Hot Start polymerase setup including primers IS5_-
reamp.P5 and IS6_reamp.P7, Following a purification using the
Agencourt AMPure XP PCR purification kit, we eluted the final
library in 30 μL of 1×TET and visualized the results on a 2% agarose
gel.

The post-capture library enriched for the NWSLH mitochon-
drial DNA was then sequenced as part of a larger pool containing
47 additional libraries at equimolar ratios on one lane of an
Illumina HiSeq2000 platform. Raw reads were used for read-
merging using Kircher’s MergeReadsFastQ_cc.py script [39],
resulting in 8.2 millions of reads that were mapped to the horse
<table>
<thead>
<tr>
<th>Sample</th>
<th>Origin</th>
<th>Sample type</th>
<th>Name fr. origin</th>
<th>Species</th>
<th>Seq. approach</th>
<th>Amplicons/reads used</th>
<th>Cov. (Cov$)</th>
<th>Length (bp)</th>
<th>Genbank acc. no</th>
</tr>
</thead>
<tbody>
<tr>
<td>JW328</td>
<td>Mineral Hill Cave, NV, USA</td>
<td>Bone</td>
<td>JW328</td>
<td>E. sp. NWSLH</td>
<td>Illumina shotgun</td>
<td>Sh./411</td>
<td>1.8 (1.8)</td>
<td>7,108</td>
<td>JX312726</td>
</tr>
<tr>
<td>ACAD 2304</td>
<td>Proskuriakov Cave, Russia</td>
<td>Bone</td>
<td>ACAD2304</td>
<td>E. ovodovi</td>
<td>Illumina shotgun</td>
<td>Sh.+C/18,679</td>
<td>78.8 (79.9)</td>
<td>16,195</td>
<td>JX312734</td>
</tr>
<tr>
<td>MS272</td>
<td>Upper Quartz Creek, YT, Canada</td>
<td>Bone</td>
<td>YG 401268</td>
<td>E. sp. NWSLH</td>
<td>Illumina+capture</td>
<td>C./5,740</td>
<td>35.6 (36.3)</td>
<td>16,074</td>
<td>JX312727</td>
</tr>
<tr>
<td>JW328</td>
<td>Mineral Hill Cave, NV, USA</td>
<td>Bone</td>
<td>JW328</td>
<td>E. sp. NWSLH</td>
<td>Illumina shotgun</td>
<td>Sh./411</td>
<td>1.8 (1.8)</td>
<td>7,108</td>
<td>JX312726</td>
</tr>
<tr>
<td>ACAD 2304</td>
<td>Proskuriakov Cave, Russia</td>
<td>Bone</td>
<td>ACAD2304</td>
<td>E. ovodovi</td>
<td>Illumina shotgun</td>
<td>Sh.+C/18,679</td>
<td>78.8 (79.9)</td>
<td>16,195</td>
<td>JX312734</td>
</tr>
<tr>
<td>MS272</td>
<td>Upper Quartz Creek, YT, Canada</td>
<td>Bone</td>
<td>YG 401268</td>
<td>E. sp. NWSLH</td>
<td>Illumina+capture</td>
<td>C./5,740</td>
<td>35.6 (36.3)</td>
<td>16,074</td>
<td>JX312727</td>
</tr>
</tbody>
</table>

Number of amplicons refers to the number of amplicons sequenced on the GS FLX, and does not include amplicons sequenced by Sanger sequencing. For total number of amplicons see Table S5. Sh. stands for shotgun sequencing and refers to cases where no amplicons have been generated. Sh.+C refers to cases where shotgun and target capture data have been merged. C. refers to in solution MYSelect capture. The number of reads used is the number of reads that mapped successfully to the reference used. Coverage is provided as the total number of bases sequenced and aligned against the horse reference mitogenome divided by the length of the horse reference mitogenome (Cov.) or the horse reference mitogenome after excluding a region of tandem repeats (Cov$). Column mitogenome length provides the total sequence length of the mitogenome that is covered by a minimal read depth of 2 and excluding tandem repeats (see Methods). doi:10.1371/journal.pone.0055950.t001
Figure 1. Phylogenetic RAxML trees (GTR+G+I) with 500 bootstraps and MrBayes (GTR+G+I) 50M generations on the full data set. Posterior probabilities are given in proportions and bootstrap support as a percentage on each branch of interest. * Branch is supported by maximum posterior probability and bootstrap (1/100). A: Including outgroups and based on 5 partitions. B: Excluding outgroups and based on 4 partitions.
doi:10.1371/journal.pone.0055950.g001

Data Partitioning and Model Selection
We included 13 previously published mtogenomes in our analyses (Table S1), and used the online Sequence Manipulation Suite [41] to partition protein-coding genes into first, second, and third codon positions. Two additional partitions were generated: one comprising the two rRNA genes (16S and 12S) and another limited to the control region. A highly variable and repetitive section of the control region consisting of tandem repeats was disregarded (from position 16,125–16,368 on horse ref genome JN390390). Phylogenetic analyses were run independently for each gene as well as on the five partitions data set (15,262 bp), the dis- cussing each gene unpartitioned, and support nodes were estimated using both 500 bootstrap pseudo-replicates and an approximate Likelihood Ratio Test (SH-like) [47]. Topological tests (Approximate Unbiased, AU; and Kishino Hasegawa, KH) were performed using CONSEL v0.3 [40] and site likelihood estimates recovered from PhyML with the – print_tree option. A total of 12 topologies were tested and sorted according to the p-values recovered from AU tests.

Phylogenetic Analyses
Phylogenetic analyses were run using PhyML online [43,46]. The topology and branch lengths were optimized, and support nodes were estimated using both 500 bootstrap pseudo-replicates and an approximate Likelihood Ratio Test (SH-like) [47]. Topological tests (Approximate Unbiased, AU; and Kishino Hasegawa, KH) were performed using CONSEL v0.3 [40] and site likelihood estimates recovered from PhyML with the – print_tree option. A total of 12 topologies were tested and sorted according to the p-values recovered from AU tests.

RaxML GUI 1.0 [49] was used to compute maximum likelihood analyses on the four and five partitioned data sets using 500 bootstraps. In cases where the best model chosen by jModelTest was not available in RaxML or PhyML, the next best model implemented was chosen instead (Table S3a–b). Bayesian phylogenetic analyses were performed using MrBayes v3.1.2 [50]. Two runs of 50 million generations were used for each dataset, and the first 25% of samples from each run was excluded as burn-in.

Table 2. Average node ages from BEAST.

<table>
<thead>
<tr>
<th>Node</th>
<th>With outgroup</th>
<th>Without outgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root</td>
<td>5.43E+07</td>
<td>N/a</td>
</tr>
<tr>
<td>Plains zebra</td>
<td>1.06E+07</td>
<td>6.88E+05</td>
</tr>
<tr>
<td>Grevy’s+Plains (B)</td>
<td>2.80E+08</td>
<td>1.46E+06</td>
</tr>
<tr>
<td>Grevy’s zebra</td>
<td>7.41E+05</td>
<td>3.93E+05</td>
</tr>
<tr>
<td>Horses</td>
<td>6.37E+05</td>
<td>3.64E+05</td>
</tr>
<tr>
<td>Horses+NW/HL</td>
<td>5.91E+06</td>
<td>2.59E+06</td>
</tr>
<tr>
<td>Mountain zebra</td>
<td>8.86E+05</td>
<td>4.84E+05</td>
</tr>
<tr>
<td>Rhino+Tapir</td>
<td>4.40E+07</td>
<td>N/a</td>
</tr>
<tr>
<td>Rhinos</td>
<td>2.08E+07</td>
<td>N/a</td>
</tr>
<tr>
<td>Sum.+Wool.</td>
<td>1.01E+07</td>
<td>N/a</td>
</tr>
<tr>
<td>Donkey (H)</td>
<td>1.24E+06</td>
<td>7.13E+05</td>
</tr>
<tr>
<td>Equids</td>
<td>9.38E+06</td>
<td>4.27E+06</td>
</tr>
<tr>
<td>Non-caballines</td>
<td>5.91E+06</td>
<td>2.92E+06</td>
</tr>
<tr>
<td>Kiang (E)</td>
<td>6.91E+05</td>
<td>3.69E+05</td>
</tr>
<tr>
<td>Kulan/onager (G)</td>
<td>1.11E+06</td>
<td>5.94E+05</td>
</tr>
<tr>
<td>Asses (F)</td>
<td>1.24E+06</td>
<td>6.72E+05</td>
</tr>
<tr>
<td>Sussemone</td>
<td>6.39E+06</td>
<td>2.92E+06</td>
</tr>
<tr>
<td>Zebras (A)</td>
<td>4.74E+06</td>
<td>1.75E+06</td>
</tr>
<tr>
<td>Zeb+Don+Ass (C)</td>
<td>5.92E+06</td>
<td>2.87E+06</td>
</tr>
<tr>
<td>Donkey+Ass (D)</td>
<td>5.42E+06</td>
<td>2.62E+06</td>
</tr>
</tbody>
</table>

Analyses run with and without outgroups (see Table S9a–b). All dates are in years with 95% confidence interval given in parentheses. N/a = not applicable.

Mitochondrial Phylogenomics of Equids

genome version equCab2.0 http://genome.ucsc.edu/cgi-bin/hgGateway?db=equCab2 using BWA [31] and SAMtools [40]. The final consensus sequence was called using SAMtools [40] requiring a minimum coverage of 2 x.
Molecular Dating

To place the estimated phylogeny on a calendar time-scale, we ran several additional phylogenetic analyses using the software package BEAST v.1.5.4 [51]. For each analysis, we assumed the relaxed uncorrelated lognormal molecular clock, first with Birth and Deaths, and then with Yule speciation. We ran separate analyses for the four- and five-partition data sets both with and without a non-equid outgroup (see above). MCMC chains were run for 250–500 million generations with 25% burn-in, and with convergence to stationarity determined by visual inspection using Tracer v1.5 [52].

In the analyses including the non-equid outgroup, the molecular clock was calibrated using a normal prior of 55.6 ± 5 Mya [12,20,53] on the root of the Perissodactyla tree. In addition, we ran separate analyses for the four- and five-partition data sets both with and without a non-equid outgroup (see above). MCMC chains were run for 250–500 million generations with 25% burn-in, and with convergence to stationarity determined by visual inspection using Tracer v1.5 [52].

In the analyses including the non-equid outgroup, the molecular clock was calibrated using a normal prior of 55 ± 5 Mya [12,20,53] on the root of the Perissodactyl tree. In addition, we ran separate analyses using a second normal prior calibration of 0.7 ± 0.1 Mya for the emergence of the Plains zebra [22] and based on the fossil record of E. mauritanicus [54,55]. In analyses excluding outgroups, only the Plains zebra calibration was used. In the absence of precise carbon dating information, ages for both the NWSLH and the Sussemione were sampled from a uniform prior spanning 12 to 100 Kya.

Monophyly was constrained for species represented in the analysis by more than one individual. The subspecies E. h. kulan and E. h. onager were not constrained to a single clade. However, following the results of the ML and Bayesian analyses above, we constrained monophyly on the clade comprising all the zebras, the clade comprising all horses (E. caballus and E. przewalskii), the clade

---

**Table 3. Mean substitution rate averaged across the whole tree for each analysis run in BEAST with 95% HPD and 25% burn-in.**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Mean rate</th>
<th>Lower and upper 95% HPD of mean rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4pBD1d</td>
<td>1.96E-08</td>
<td>1.56E-08–2.38E-08</td>
</tr>
<tr>
<td>4pBD2d</td>
<td>2.04E-08</td>
<td>1.68E-08–2.40E-08</td>
</tr>
<tr>
<td>4pBDn1d</td>
<td>3.55E-08</td>
<td>1.76E-08–5.71E-08</td>
</tr>
<tr>
<td>4pY1d</td>
<td>1.17E-08</td>
<td>6.46E-09–1.69E-08</td>
</tr>
<tr>
<td>4pY2d</td>
<td>1.80E-08</td>
<td>1.43E-08–2.18E-08</td>
</tr>
<tr>
<td>4pYn1d</td>
<td>3.96E-08</td>
<td>1.94E-08–6.31E-08</td>
</tr>
<tr>
<td>5pBD1d</td>
<td>1.82E-08</td>
<td>1.48E-08–2.19E-08</td>
</tr>
<tr>
<td>5pBD2d</td>
<td>1.95E-08</td>
<td>1.63E-08–2.27E-08</td>
</tr>
<tr>
<td>5pY1d</td>
<td>1.52E-08</td>
<td>1.17E-08–1.87E-08</td>
</tr>
<tr>
<td>5pY2d</td>
<td>1.77E-08</td>
<td>1.45E-08–2.10E-08</td>
</tr>
<tr>
<td>5pYn1d</td>
<td>3.57E-08</td>
<td>1.92E-08–5.47E-08</td>
</tr>
</tbody>
</table>

Given in number of substitutions per site per million years 4p = 4 partitions; 5p = 5 partitions; BD = Births and deaths model; Y = Yule model; 2d = 2 calibration dates (Perissodactyla and Plains zebra); 1d = 1 calibration date (Perissodactyla); n = no outgroup (1d = Plains zebra).

---

**Figure 2. Node ages as estimated from BEAST analyses given in years and with 95% HPD.** Shown is the averaged results from the three analyses excluding outgroups (Table 2).

doi:10.1371/journal.pone.0055950.g002
comprising the African asses (E. africanus somalicus and E. asinus), and the clade containing all equids.

**Branch Tests for Natural Selection**

To identify regions of the mitogenome that may be under selection in any given lineage, we performed branch tests for each individual protein-coding gene as implemented in PAML v.4.5 [56]. Based on the branch test results, branch-site tests were also run for the NADH dehydrogenase 4 protein gene (ND4) to test the caballine (E. caballus and NWSLH) branch for sites under positive selection.

**Results and Discussion**

We successfully recovered the complete mitogenome of three extinct equid taxa (NWSLH, Sussemione, and Quagga). We also characterized partial mitochondrial contigs (11–283 bp in length, for a total of 7,108 bp) for a second NWSLH [JW325]. Of the modern equids we sequenced 14 full mitogenomes (Table 1). The complete mitogenomes of the NWSLH, Sussemione, E. hemionus kulan, and the three species of zebra plus the Quagga had not been sequenced prior to this study. The average depth-of-coverage is 387.3 for modern samples and 57.2 for ancient samples (Table 1).

Minimal pairwise distances were observed between conspecific individuals, suggesting that the sequences reported here are in agreement with the known sequence diversity among equids (Table S8). We evaluated nucleotide misincorporation patterns through comparison of Illumina reads with the horse reference mitogenome following the procedure described in Ginolhac et al. [37]. In contrast to modern samples, typical accumulation of C to T and G to A mismatches were observed at sequence ends in the ancient samples, suggesting that the latter were affected by significant levels of post-mortem DNA damage, in particular cytosine deamination (Figure S3). This, together with the base composition of our sequences (Table S6), higher substitution rates at third codon positions over first and second codon positions respectively (data not shown), and the blank experimental controls used during the whole laboratory procedure highlights the quality our sequence data.

**Topological Relationships**

Mitogenome alignments were used to perform phylogenetic analyses under a Bayesian and a Maximum Likelihood framework. All methods and sequence partitions considered supported the same topology (Figure 1a–b; Table S2; Figure S2), where all nodes except C, D and G received maximal support (Figure 1a–b). Interestingly, the mitochondrial topology appeared very similar to a study based on 22 partial mitochondrial and nuclear genes [57] and one reconstructed from ca. 55,000 nuclear SNPs [8]. This appears in strong contrast to other recent cases reported among mammals where the phylogenetic signal retrieved from nuclear and mitochondrial genes were in conflict (e.g. for brown and polar bears, [50]; for Denisovans and Neanderthals, [59]).

Rooted phylogenetic analyses supported a major division between caballine horses and non-caballine equids, where horses and NWSLH appeared as sister species within caballine horses, in agreement with previous results based on partial mitochondrial sequences [20,22]. This suggests that Asiatic wild asses and NWSLH did not originate from a single ancestral population with gracile limbs, but instead that similarities in their post-cranial skeleton [25,60] derive from convergent evolution. However, another possibility is that gracile limbs is the ancestral state which has been preserved in the Asian asses, while it has been lost in all other lineages, resulting in limb states derived from a more gracile one.

Zebras were found to be monophyletic (node A, Figure 1a–b), with the Mountain zebra (E. zebra) diverging before the Plains (E. quagga) and Grevy’s (E. grevyi) zebras, which are sister species. This topology is supported by other studies using a variety of genetic markers [8,12,24,57,60,61]. The now extinct Quagga nests within the Plains zebra clade, confirming their conspecific relationship [24]. Asiatic wild asses (node F, Figure 1a–b) can also be confirmed as monophyletic. The subspecies E. h. onager and E. h. kulan clustered together (albeit with relatively low bootstrap support) in a group divergent from E. kiang.

Sussemiones (as represented by E. vossii) were confirmed as a distinct branch within non-caballine equids (Figure 1a–b; Figure S2), as suggested by previous analyses based on partial mitochondrial sequence information [22]. The Sussemione lineage showed greater genetic distance to caballine horses (Mean = 0.0554±0.0036) than to non-caballine equids (Mean = 0.0427±0.0026) (Table S2; Table S8). Approximately Unbiased and Shimodaira-Hasegawa topological tests showed significantly more support for a topology nesting Sussemiones within non-caballine equids (Figure S2; item 3 in Table S2). However, even though the full length of the mitogenome was characterized, the exact placement of Sussemiones within non-caballines could not be resolved with high confidence (Figure 1a–b, Figure S2; and topological tests presented in Table S2), suggesting that additional phylogenetic information from nuclear genes and/or the identification of e.g. rare genomic rearrangements or insertion sites of transposable elements will be required for unraveling the phylogenetic origin of this enigmatic equid lineage [21,62].

In Bayesian analyses, African wild asses and domestic donkeys were supported as a sister lineage to Asiatic wild asses, a pattern compatible with various past suggestions that the subgenus Asinus should incorporate both African and Asiatic asses; however this relationship was not mirrored in ML analyses (node D, Figure 1a–b). Therefore, the mitochondrial information should be regarded as inconclusive with regards to the relationships between the main non-caballine groups (Sussemiones, Asiatic wild asses, African wild asses and donkeys, and zebras). Interestingly, SNP chip data comparing >40K autosomal markers [8] and a study combining nuclear and mitochondrial genes [57] could not resolve these relationships with high support either. This suggests that non-caballine equids have likely experienced an extremely rapid radiation. Alternatively, incomplete lineage sorting within the non-caballine ancestral population might explain the persistence of donkey-like mitochondrial sequences in the Asiatic wild ass lineage, resulting in spurious phylogenetic reconstructions.

The emergence of new approaches using NGS, including whole exome capture, or even complete genome characterization, will soon be available to determine whether the non-caballine phylogeny corresponds to a hard or soft polytomy [63,64], or incomplete lineage sorting [65]. For now, we note that many other studies have also found a sister relationship between African wild asses and Asiatic wild asses [8,57,66–69] (but see [12]).

**Branch Tests for Natural Selection**

We used the PAML package [36] to estimate non-synonymous to synonymous rate ratios (dN/dS) along the main branches of the topology (Figure S1) and compared this model to a null model that assumes one conserved ratio across the whole phylogeny (Table 2). For all genes except COX2, ND2, ND4 and ND5, the null model was better supported and dN/dS ratios were found to be significantly lower than 1 (Table S7), in agreement with the
presence of purifying selection acting on mitochondrial genes. For ND4 (NADH dehydrogenase 4), the dN/dS ratio of the ancestral branch leading to caballine horses (horses and NWSLH) was found to be significantly superior to 1 (Table S7), suggesting positive selection for non-synonymous variation during the early evolutionary history of that lineage. In order to identify which amino acid(s) could have been functionally advantageous, we performed branch-site tests in the branch ancestral to caballine horses compared to the rest of the topology for gene ND4. However, this analysis did not identify any amino acids that were supported as having evolved under positive selection (p-value: >0.25).

Divergence Times

We used relaxed molecular clock (log-uncorrelated) methods as implemented in BEAST [51] to date recent lineage divergence events along the evolutionary history of equids. A total of eleven analyses were run including different data set partitions (4-way or 5-way partitioned), different calibration points (the origin of Perissodactyla at 53 Mya, and/or a recent emergence of the Plains zebra lineage at 0.7 Mya), and different models of speciation (Yule versus Birth/Death). The resulting average estimated ages are displayed in Figure 2, Table 2 and Table S9a-b. The estimated mitochondrial TMRCA for the Equidae varied considerably between our different analyses. When the two outgroup taxa are excluded from the analysis and only the more recent calibration point is used, we estimate the mitochondrial TMRCA of all equids around 4.3 Mya (95% CI: 4.0–4.7 Mya). When the outgroup taxa are included, the average mitochondrial TMRCA increases to 8.6 Mya (95% CI: 6.7–11.9 Mya) (see Table S9a-b). The younger TMRCA is similar to previously reported divergence estimates for the Equidae (3.8 Mya: [12]; 3.9 Mya: [60]; 4.0 Mya: [22]), however ages bracketing that range have also been proposed (2.3 Mya: [10]; 5.8 Mya: [20]). We believe the older age estimates likely result from the contrast between the deep phylogenetic distance separating equids and other living perissodactyls, and the relatively young origin of equids as a whole. Evolutionary rate estimates have been shown to vary in a time-dependent manner [70,71], with deep calibrations leading to slower evolutionary rate estimates. The posterior distribution of mutation rates recovered from our analyses show this trend (Table 3). This may be due to a variety of effects, including substitution saturation, purifying selection, and model misspecification, especially with regard to how rate heterogeneity across sites is compensated [72,73]. The older calibration may therefore be less appropriate for estimating recent divergence times, such as those we aim to address here. Assuming only the more recent calibration, we find that caballine and non-caballine equids most likely diverged ~4 Mya. This estimate is older than the age of the oldest fossil unambiguously identified as Equus in the paleontological record, which has been dated to 2.1 Mya [74]. According to our estimates, the mitochondrial TMRCA of the lineage leading to domestic horses (E. caballus) is ~364 Kya (95% CI: 336–389 Kya). This is older than that reported in Steiner and Ryder [12] (250 Kya), Achilli et al. [5] (150 Kya), and Lippoldt et al. [7] (100 Kya), all of whom assumed that the genus Equus emerged ~2 Mya. We should caution that our estimates assume 700 Ky as a calibration date for the TMRCA of Plains zebras. In the current dataset, the latter is estimated based on only three lineages, which likely leads to an under-estimation of the true phylogenetic distance among Plains zebras, and consequently of the mutation rate. Therefore, our molecular clock analyses are expected to provide upper bound estimates for divergence times and TMRCA dates. Additional complete mitochondrial genome data from modern Plain zebras coupled with tip calibration based on a series of radiocarbon dated ancient equid samples [75] will be needed to revisit the exact timing of lineage radiation in equids. For now, we note that most of the extant equal lineages radiate at the transition from the Tertiary to the Quaternary at 2.6 Mya, with a TMRCA for caballine horses estimated at 2.6 Mya (95% CI: 2.4–2.9 Mya) (Table 2; Table S9b). The mitochondrial TMRCA for zebras dates back to 1.75 Mya (95% CI: 0.7–2.5 Mya) with the Grevy’s and Plains zebras diverging 1.5 Mya; this is similar to estimates by Steiner and Ryder [12] of 1.2 Mya and George and Ryder [60] of 1.6 Mya. We estimate that African asses shared a TMRCA 713 Kya (95% CI: 706–723 Kya), in agreement with the date Kruger et al. [3] estimated using nuclear data (0.5–1.5 Mya). Finally, within Asian wild asses, we estimate that the mitochondrial divergence between the Kiang and Onager/Kulan lineages occurred 672 Kya (95% CI: 649–688 Kya). This time depth supports a separate species status for the Kiang, previously proposed on the basis of morphological [76] and chromosomal distinctions (2n = 51–52 in E. kiang versus 2n = 54–56 in E. hemionus) [77].

Concluding Remarks

Here, we provide new data to assess the timing of the appearance of extant and several Late Pleistocene equid lineages, and to investigate their evolutionary relationships to one another based on complete mitogenome sequences. Our results are relevant to taxonomic boundaries: (1) While we confirm the monophyly of zebras, we find a surprisingly deep evolutionary divergence between the Mountain zebra and Plains zebra; (2) our results support recognition of the Kiang as an evolutionarily distinct species, rather than part of a single radiation that includes the Onager and Kulan; (3) we confirm that New World stilts-legged horses are closely related to caballine horses; (4) we show that the enigmatic Late Pleistocene Sussemione is a non-caballine equid that is only distantly related to extant equid lineages; and (5) we confirm that the recently extinct Quagga of South Africa was a subspecies of the widespread Plains zebra. Our study has also revealed a rapid radiation within non-caballine equids, within which details regarding order and pattern of diversification could not be resolved from complete mitogenomes. The massive throughput of current NGS platforms now enables the characterization of the complete nuclear genome of non-model organisms [78]. Such genome-wide approaches will likely provide enough information to resolve the nodes that are presently still problematic.

Supporting Information

Figure S1 Constrained topology used for positive selection tests in PAML. Branch numbers are shown and are equivalent to branch numbers used in Table S7. (TIFF)

Figure S2 The 12 topologies tested in the topological test. Topology number 1 to 12 is equivalent to item number in Table S2. Suss = Sussemione (E. soudoric); Burch = Plains zebra (E. quagga); Grev = Grevy’s zebra (E. grevyi); Moun = Mountain zebra (E. zebra); Ass = African wild ass and domestic donkey (E. africanus and E. asinus); Ona/Kul = E. hemionus (Onager and Kulan). (PDF)

Figure S3 DNA fragmentation and Nucleotide misincorporation patterns. Panel A: Ancient sample ACAD2304 E. soudoric. Panel B: Ancient sample QH1 E. quagga quagga (only
shotgun reads are considered in order to avoid the base composition bias resulting from the target-enrichment approach.

Panel C: **MS272 E. sp. NWSLH** Panel D: Modern sample H111 *E. zebra hartmannae*. The base composition of the reads is reported for the first 10 nucleotides sequenced (left: 1–10) as well as for the 10 nucleotides located upstream of the genomic region aligned to the reads (left: −1 to −10). In addition, the base composition of the last 10 nucleotides sequenced (right: −10 to −1) and of the 10 nucleotides located downstream from the reads (right: 1–10) is also provided, all in relation to the mitochondrial horse reference genome. Nucleotide positions located within reads are reported with a gray frame. Each dot reports the average base composition per position. The figure shows two parallel base compositions; these correspond to the base composition of the two mitochondrial DNA strands, one being relatively enriched in purines (H, heavy), compared to the other one (L, light). The frequencies of all possible mismatches and indels observed between the horse genome and the reads are reported in gray as a function of distance from 5′- to 3′-ends (first 25 nucleotides sequenced) and 3′- to 5′-last 25 nucleotides), except for C>T and G>A that are reported in red and blue, respectively. These frequencies are calculated by dividing the total number of occurrences of the modified base at a given position in a read by the total number of the unmodified base at the same position in the horse genome.

(PDF)

Table S1 **Published mitogenomes from Genbank used in analyses.** Includes both equids and non-equid outgroups.

(PDF)

Table S2 **Topological test results.** Table displaying results of the 12 different topologies tested. The topologies are visually illustrated in Figure S2. Item = Topology number (Figure S2).

(PDF)

Table S3 **Selected models from jModelTest.** The best model selected by jModelTest per gene or codon is displayed according to analysis. A: On full dataset. B: On dataset excluding non-equid outgroup.

(PDF)

Table S4 **Primer pairs used to amplify some of the modern mitogenomes.** Ta = annealing temperature, Ext. = extension time, Size = size of the amplicon generated by the primer pair, Location = a rough estimate of the part of the mitogenome the primers amplify based on the horse reference mitogenome [JN908398]. * SF and LF are universal mammalian mitogenome amplifying primers (designed in-house by Sandra Abel Nielsen), Pr1 and Pr2 are from study Xu et al. [17], and Pr3 was designed in-house by Ludovic Orlando.

(PDF)

Table S5 **Primer sets used to amplify 5 modern mitogenomes for FLX sequencing.** Primer set names as from Table S4, primer sets from Pr3 to 14.3_Pr3 make up shorter regions of the LF fragment. Pr1 and Pr2 are shorter primers to cover the hypervariable region of the control region, while the last four are to fill in gaps throughout the mitogenome.

(PDF)

Table S6 **Nucleotide frequency composition of all samples.** Nucleotide frequencies in percent of total number of nucleotides (excluding tandem repeats).

(PDF)

Table S7 **Results from branch selection tests.** dN/dS for each branch given per gene, calculated from PAML under Model 2 (with Model 0 values in the second to last bottom row). Branch numbers are as in Figure S1. Values superior to 1 are highlighted in bold. P-values from LRT are given in the bottom row, with p-values ≤0.02 highlighted in bold.

(PDF)

Table S8 **Pair-wise Distance results.** Uncorrected pairwise distances observed between Sussemione (*E. ovodovi*) and other equid species.

(PDF)

Table S9 **Mean node ages estimated from BEAST.** Total number of states is measured in millions of states [M]. Node letter in parentheses corresponds to Figure 1. 4p = 4 partitions; 5p = 5 partitions; BD = Births and Deaths; Y = Yule model; Sum. = Sumatran rhino; Woolly = woolly rhino; NWSLH = New World stilt-legged horse; Asses = *E. hemionus* and *E. kiang*; Zeb = zebras; Donkey = *E. africanus* and *E. asinus*, Don = Donkey. A: Results are with a 25% burn-in for eight different datasets including outgroup. 1d = 1 date (Perissodactyla); 2d = 2 dates (Perissodactyla and Plains zebra). B: Results are with a 25% burn-in for 3 different datasets without outgroup. 1d = 1 date (Plains zebra); n = no outgroup.

(PDF)

Acknowledgments

We thank Eline Lorenzen and Hans Siegismund for providing access to tissue samples from African zebras biopsied in the wild. We would also like to thank Paul Clausen and Jens Sigaard from Aalborg Zoo, Julia Stagegaard from Ree Park, Torsten Møller fra Kolmårdens Zoo, Günter Straus from Tierpark Berlin, Mads Frost Bertelsen from Copenhagen Zoo, and the National Museum of Natural History (Smithsonian Institution) for providing modern equid samples. We gratefully acknowledge the Institute of Archaeology and Ethnography and the Russian Academy of Sciences for the Sussemione sample; Grant Zazula from the Government of Yukon and Patrick J. Wrixon from the University of Arizona for ancient samples. We thank the Danish National High-Throughput DNA Sequencing Centre, and especially Lillian Petersen and Kim Magnusen, who provided technical support related to GS FLX and Illumina sequencing. We thank Tina Brand, Pr. Eske Willerslev, Pr. M. Tom P. Gilbert, and Gianella Garcia-Hughes for technical assistance and and Pr. Eske Willerslev, Dr. Ruth Fernandez-Garcia and Mikkèl Schubert for fruitful discussion.

Author Contributions

Provided taxonomic and nomenclatural input: KMH. Conceived and designed the experiments: JTV LO MS BS. Performed the experiments: JTV ASO MS MR LO. Analyzed the data: JTV AG LO. Contributed reagents/materials/analysis tools: AC BS DF KMH JC JW RCF SCAN SKV NDO LO. Wrote the paper: JTV LO. 4. Groves C, Grubb P (2011) Ungulate taxonomy: Johns Hopkins University Press.

References

17. Xu XF, Arnason U (1994) The Complete Mitochondrial-DNA Sequence of the