nature

https://doi.org/10.1038/s41586-019-1714-1

Supplementary information

Human origins in a southern African palaeo-wetland and first migrations

In the format provided by the authors and unedited

Eva K. F. Chan, Axel Timmermann, Benedetta F. Baldi, Andy E. Moore, Ruth J. Lyons, Sun-Seon Lee, Anton M. F. Kalsbeek, Desiree C. Petersen, Hannes Rautenbach, Hagen E. A. Förtsch, M. S. Riana Bornman & Vanessa M. Hayes

Human origins in a southern African palaeo-wetland and first migrations

Eva K.F. Chan[1,2], Axel Timmermann[3,4], Benedetta F. Baldi[1], Andy E. Moore[5], Ruth J. Lyons[1], Sun-Seon Lee[3,4], Anton M.F. Kalsbeek[1], Desiree C. Petersen[1], Hannes Rautenbach[6,7], Hagen E.A. Förtsch[8], M.S. Riana Bornman[7], & Vanessa M. Hayes[1,2,7,9,10]

[1]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia; [2]St Vincent's Clinical School, University of New South Wales, Randwick, NSW, Australia; [3]Center for Climate Physics, Institute for Basic Science, Busan, South Korea; [4]Pusan National University, Busan, South Korea; [5]Department of Geology, Rhodes University, Grahamstown, South Africa; [6]Climate Change and Variability, South African Weather Service, South Africa; [7]School of Health Systems and Public Health, University of Pretoria, South Africa; [8]Windhoek Central Hospital, Windhoek Khomas, Namibia; [9]Faculty of Health Sciences, University of Limpopo, Turfloop Campus, South Africa; [10]Central Clinical School, University of Sydney, Camperdown, NSW, Australia.

Supplementary Methods

Haplogroup Pre-screening

Haplgroup pre-screening was performed using a 2,673 bp region of the mitogenome, amplified with primers 3322-F (5'CTCCTACTCCTCATTGTACC-3') and 5995-R (5'-GTGCCTAGGACTCCAGCTCA-3'). The products were initially screen one of the L0 defining-variants, T5442C, using primer 4838-F (5'-TCTGACATCCGGCCTGCTTC-3'). L0 samples were subsequently sequenced with 3322-F, 4040-F (5'-CAACATATGACGCACTCTCC-3') and/or 5995-R to further delineate into L0d (T4232C), L0d1 (G3438A), L0d1b (T3618C), L0d1c (C4197T), L0d1'2 (A3756G), L0d2 (A3981G, C205T, A4044G), L0d2a (A5153G), L0d2d (G5147A, G5231A), L0d2C (A4038g, T4937C) and L0d3 (G5460A, G5773A).

Whole Mitogenome Sequencing

A 7.2 kb (12,250 – 3,005) and a 9.7 kb (2,583 – 12,337) fragment of the circular mitogenome were isolated and amplified using primer pairs GGCTTTCTCAACTTTTAAAGGATA(F)/TGTCCTGATCCAACATCGAG(R) and CCGTGCAAAGGTAGCATAATC(F)/TTACTTTTATTTGGAGTTGCACCA(R), respectively.

Public Data

Three main search routes were used to identify all publicly available L0 mitogenomes, between the year 2015 and 2017:

NCBI Nucleotide Database was interrogated with the following search terms, filtering on sequence lengths between 16,000 and 17,000 for "complete mitochondrion sequences":

- "L0k" identified 78 sequences, of which 75 belongs to the PubMed 23332919 and 3 to PubMed 17194802.
- "L0g" identified a single mitogenome: MD7, which we had previously reported.
- "L0f" identified 11 belonging to PubMed records 15890885 (n=1), 17194802 (n=6), 24901532 (n=3), and 26567083 (n=1).
- "L0a" identified 28, belonging to PubMed IDs 17194802 (n=8), 19083815 (n=8), 22130972 (n=10), 28177087 (n=1), and 28669760 (n=1).
- "L0d" identified 7 sequences, all from PubMed ID 17194802.
- "L0d1" identified 293 sequences, all from PubMed ID 23332919.
- "L0d2" identified 114 sequences, from PubMed 17194802 (n=3) and 23332919 (n=111).
- "L0d3" identified 5 sequences, all from PubMed 23332919.

NCBI PubMed database was interrogated with the search term "human L0 mitochondrial genome", which returned three records:

1. PubMed ID 24236171 did not have an associated PopSet, but has 42 associated nucleotide sequences, all 16000-17000 long.
2. PubMed ID 23428049 did not have an associated PopSet or any GenBank Nucleotide Accession. There was also no supplemental data accompanying the publication.
3. PubMed ID 17194802 corresponded to PopSet 133854527 containing 62 complete mitogenomes.

From the **L0 phylogenetic tree of PhyloTree** (Build 17) <http://www.phylotree.org/tree/L0.htm>, all 204 accessions numbers indicated as representative mitogenomes to the corresponding haplogroups were obtained.

- Eight of these were based on Coriell cell lines and were excluded for the purpose of this study.
- The remaining 196 mitogenomes corresponded to 15 different PubMed studies.

There is overlap in studies and mitogenomes across the three search routes. In all, 26 studies were identified containing a total of 6,334 mitogenomes. From this, 1,019 mitogenomes were confirmed, using HaploGrep2[44] with PhyloTree v17[12], as belonging to the L0 haplogroup (*Supplemental Table 2*).

Phylogenetic Inference

Due to computational burden, BEAST was performed on several different subsets of mitogenomes, all selected to include (i) the novel 198 mitogenomes from this study, (ii) only complete mitogenomes (excluding 27 mitogenomes containing only sequences of the coding region[48,49]), (iii) rare L0, particularly L0g, L0f, L0d2d, L0d1d, and L0k, and (iv) the 121 L0 mitogenomes from our previous studies: 77 from Chan *et al*. 2015[6], the 2,330 year old StHe skeleton from Morris et al. 2014[17] that defined the new haplogroup L0d2c1c, 37 of 87 mitogenomes from McCrow et al. 2016[41], and six mitogenomes from Schuster et al. 2010[50].

The specifics of the phylogenetic model set up in BEAUTi v2.4.2 were:

- Alignment were not partitioned by site.
- Gamma Site Model was used with 6 gamma categories (no significant difference was noted with fewer (4) gamma categories) and zero proportion invariant sites (due to high correlation with rate variation among sites).
- The GTR (generalized time reversible) substitution model was assumed, estimating all rates, except the rate of CT, and estimating all frequencies. The mean substitution rate was not fixed as calibration data was used (below).
- A strict constant clock model was used with a normal prior of with $\mu = 1.665 \times 10^{-8}$ and $\sigma = 1.479 \times 10^{-9}$, based on previously estimates for whole mitogenome[51].
- For the purpose of estimating timelines, a Coalescent Constant Population model was assumed.
- Times were calibrated on the Neanderthal genomes. Specifically, the data was divided into two taxon sets:
    - Set "Neanderthals" included all seven *Homo neanderthalensis* mitogenomes, with tip dates set to their archeological dating (in years before present): Feldhofer1=40000, Vindija=38000, Croatia=38310, ElSidron=39000, Feldhofer2=40000, Altai=50000, Mezmaiskaya=65000 (See *Supplementary Table 10*). No prior was set on the most recent common ancestor of this taxon set, and calibration was applied to the leaves instead of the most recent common ancestor.
    - Set "L0" included all contemporary mitogenomes, with a normal prior set on the coalescent time $N(\mu = 200000, \sigma = 50000)$. A tip date for the StHe mitogenome was set to 2,330 ybp as previously estimated from carbon dating[17]. This taxon was assumed to be monophyletic.

Timeline estimates were performed on several different data subsets, differing predominantly in the number and individual samples included for the "common" haplogroups. No dramatic difference in tree topology or coalescent times was noted between different subsets, except for the splitting of L0d'k, L0d1a'b'c'd, L0d2a'b'd. In the paper, the final results were derived from a subset of 461 mitogenomes (all 198 samples from this study (*Supplementary Table 1*) and 263 samples from *Supplementary Table 2*). As only two L0b mitogenomes have been reported, we also tested the exclusion of L0b in our analysis and confirmed it did not significantly impact the overall tree topology or estimated timelines.

Bayesian Skyline Plot Analysis

For haplogroups with fewer than 100 samples, all mitogenomes were included in the analyses. For haplogroups with more than 100 mitogenomes, five subsets of 100 randomly selected samples were examined. BSP analyses were performed using BEAST2 with model setup via BEAUTi 2 as before but with the following key differences:

- Gamma shape of the Gamma Site Model was estimated with an exponential prior with mean = 1.0 and offset = 0.0.
- The molecular clock was fixed (not estimated) at 1.665×10^{-8} (Soares *et al*. 2009).
- The phylogenetic tree prior was set to Coalescent Bayesian Skyline. Different number of intervals between the root of the tree and the present time was used for different haplogroups, depending on the number of mitogenomes included.
- Each BSP BEAST2 run was performed with 50 million MCMC chains, sampling every 5,000.

BEAST runs with the best overall effective sample sizes are reported in the paper, as well as in *Supplementary Table 12*.