PLOS ONE

RESEARCH ARTICLE

Rectifying long-standing misconceptions about the ρ statistic for molecular dating

Vincent Macaulay[1], **Pedro Soares**[2,3,4]*, **Martin B. Richards**[5]

1 School of Mathematics and Statistics, University of Glasgow, Glasgow, United Kingdom, **2** CBMA (Centre of Molecular and Environmental Biology), Department of Biology, University of Minho, Campus de Gualtar, Braga, Portugal, **3** Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Campus de Gualtar, Braga, Portugal, **4** Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal, **5** Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield, United Kingdom

* pedrosoares@bio.uminho.pt

Citation: Macaulay V, Soares P, Richards MB (2019) Rectifying long-standing misconceptions about the ρ statistic for molecular dating. PLoS ONE 14(2): e0212311. https://doi.org/10.1371/journal.pone.0212311

Editor: Jun Gojobori, SOKENDAI (The Graduate University for Advanced Studies), JAPAN

Received: March 30, 2018

Accepted: January 19, 2019

Published: February 19, 2019

Data Availability Statement: All relevant data are within the paper.

Funding: P.S. and M.B.R. acknowledge FCT support through project PTDC/EPH-ARQ/4164/2014 partially funded by FEDER funds (COMPETE 2020 project 016899). P.S. is supported by FCT, European Social Fund, Programa Operacional Potencial Humano and the FCT Investigator Programme (IF/01641/2013) and acknowledges FCT/MEC for support to CBMA through Portuguese funds (PIDDAC)—PEst-OE/BIA/UI4050/2014. M.B.R. received support from a Leverhulme Doctoral

Abstract

When divided by a given mutation rate, the ρ (rho) statistic provides a simple estimator of the age of a clade within a phylogenetic tree by averaging the number of mutations from each sample in the clade to its root. However, a long-standing critique of the use of ρ in genetic dating has been quite often cited. Here we show that the critique is unfounded. We demonstrate by a formal mathematical argument and illustrate with a simulation study that ρ estimates are unbiased and also that ρ and maximum likelihood estimates do not differ in any systematic fashion. We also demonstrate that the claim that the associated confidence intervals commonly estimate the uncertainty inappropriately is flawed since it relies on a means of calculating standard errors that is not used by any other researchers, whereas an established expression for the standard error is largely unproblematic. We conclude that ρ dating, alongside approaches such as maximum likelihood (ML) and Bayesian inference, remains a useful tool for genetic dating.

Introduction

Archaeogenetics has been described as "the study of the human past using the techniques of molecular genetics" [1]. Mitochondrial DNA (mtDNA), particularly when analysed phylogeographically, was pivotal to the pioneer phase of archaeogenetics [2] and continues to play an important role, even as ancient DNA and genome-wide studies, which now allow for direct checking of age estimates and dispersal models, have become central [3,4]. The value of mtDNA is due to its high mutation rate, allowing the accumulation of diversity within the time frame of recent human evolution, and lack of recombination, allowing the reconstruction of extremely well-resolved phylogenetic trees. More particularly, alongside the paternally inherited Y-chromosome variation, the maternally inherited mtDNA is invaluable for assessing sex-specific dispersal patterns, which are now understood to have had a major impact in recent prehistory [5].

Scholarship programme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Of course, like the male-specific part of the Y chromosome (MSY), the mtDNA is inherited as single locus, subject to the vagaries of one realization of genetic drift. They capture merely some shadow of human evolution. For example, neither preserves signals of admixture with archaic species of human, which required autosomal data to detect it [6]. In the future, haplotype blocks within the autosomes should provide a goldmine of phylogeographic insight. Nevertheless, due to the high density of non-recombining markers they carry, the phylogeography of these two genetic systems are uniquely fine-grained, with genealogical trees that reflect processes in human history in important ways. For example, the discovery, based on mtDNA variation, of an origin for modern humans in Africa by around 200 ka [7,8] followed by an out-of-Africa migration between 60 and 70 ka, has been refined and substantiated by other lines of evidence over the last three decades [9–11].

This molecular dating, whether based on the mtDNA, the MSY or haplotype blocks elsewhere in the genome, is based on a calibrated molecular clock. Properties of the mtDNA clock have been painstakingly worked out over many years [12], including the evaluation of and correction for the effects of purifying selection on the time estimates [13–16]. Some consensus seems to have been achieved, as current molecular clock estimates calibrated using very different approaches, using paleontological data [13,14], archaeological data [4] and radiometrically dated ancient DNA sequences [1,3,8,17–19] now provide largely compatible results. Moreover, the most widely used whole-mtDNA clock at present [14] has provided reliable, independently verifiable age estimates for the settlement of the Pacific [20], the Bantu expansion into southern Africa [7,21] and the colonization of the American continent [22], all of which are rather well dated radiometrically.

A separate issue to the clock calibration, however, is the method used to estimate the age of clades in the tree. Alongside maximum-likelihood and Bayesian approaches, one very simple method that remains widely used is based on the ρ (rho) statistic [23], which measures the average number of mutations from the supposed root of the clade (the ancestral sequence) to each of the sampled sequences in the clade. When divided by the mutation rate for the whole sequence per unit time, it provides an estimate of the age of a given clade in those time units. The mathematics of the approach, and the means by which confidence intervals can be estimated, were explained by Saillard et al. [24] and subsequently implemented in the Network software package (http://www.fluxus-engineering.com/sharenet.htm). However, the ρ approach has received more than its fair share of criticism over the years.

Some of the perceived weaknesses of ρ can, in fact, be seen as strengths. The fact that ρ is a statistic that does not use the tree topology in its calculation, let alone any explicit demographic model, has sometimes been argued to be a drawback, but this feature rather provides a robust and, as we shall see, unbiased estimate against which more assumption-heavy estimates can be compared. In general, with few recent exceptions, whole-mtDNA genome articles published these days rarely use ρ alone but use it alongside maximum likelihood and/or Bayesian inference estimates [21], where it consistently gives similar results. Note though that, as we will discuss below, the topology of the tree has important consequences for the uncertainty of the estimates of age derived from ρ or, for that matter, any other estimator.

However, nearly a decade ago, Cox [25] claimed that the ρ statistic (even when scaled by an appropriate estimate of the mutation rate) produces biased estimates of the time to the most recent common ancestor (TMRCA) of a set of sequences, and that associated confidence intervals do not estimate the uncertainty appropriately. A simple mathematical proof that ρ is unbiased in this context–as, for example, presented by Saillard et al. [24] and Thomson et al. [26], and repeated below–deals with the first claim: any simulation that displays a contradictory result must therefore be flawed. The second claim rests on an expression for the estimated standard error of ρ (equation (3) in [23]) that has not appeared elsewhere.

The ρ statistic has continued to be usefully employed by numerous diverse researchers working across the areas of mtDNA phylogeography [27–30] and disease studies [31], and the use of Y-chromosome [32–37] and X-chromosome [38] variation to study human evolution. Despite this, Cox's arguments continue to be cited, even in widely-used textbooks covering human evolutionary genetics [39,40], as well as research papers [41] and high-profile reviews [42].

Our present purpose is therefore to: (a) prove again that ρ is unbiased; (b) to try to reproduce some of the simulations performed by Cox [25]; and (c) to illustrate the value of ρ by drawing age estimates from the literature where both ρ and other methods were employed for the same clades and the same datasets. We will (d) also explore the question raised regarding the coverage of confidence intervals derived from ρ.

Methods

We carried out a simulation aiming for the same conditions as simulations described before that yielded an apparent bias for ρ [23], up to an ambiguity about population and sample sizes being for haploids or diploids (10,000 realizations of a coalescent process, with haploid population size $N = 1000$, haploid sample size $n = 100$, mutation rate across the whole sequence $\mu = 0.00234$ per generation [coming from Cox's assumption of a transition rate of 1.8×10^{-7} per base-pair per year, a generation time of 26 years and a sequence length of 500], giving $\vartheta = 2N\mu = 4.68$), coded in R [43] (scripts available at http://www.stats.gla.ac.uk/~vincent/rho). Trees were generated using the constant-size coalescent process and mutations assigned to its edges under the infinite sites model.

For comparison of real data, we collected data from several published manuscripts that displayed a comparison between ρ and maximum likelihood [9,10,11,44–47] using the same mtDNA dataset. In order for all the ages to be comparable we used age estimates based only on the time-dependent mtDNA molecular clock we developed [14]. This earlier work also addressed many of the problems also raised by Cox concerning the accuracy and precision of the mtDNA mutation rate, including the issue of selection, and we have also discussed previously how uncertainty in the estimates of the mutation rate can affect the outcomes [10], but the comparison is fully independent of the molecular clock used, as the objective is to compare the methods of estimating branch lengths and not the actual age estimate. ML age estimates in the different studies were performed with PAML [48] using the HKY85 model with gamma-distributed rates. We note that other criticisms of ρ made by Cox, such as mtDNA mutation rate heterogeneity, are addressed by the comparisons with ML, which explicitly takes the tree structure into account; and the issue of mtDNA homoplasy has been addressed in the last decade by focusing on highly informative molecular sequences, such as whole mtDNA genomes, which provide better resolved molecular phylogenies.

Results and discussion

Mathematical demonstration

Suppose that the genealogy of a sample of n sequences consists of a tree of m links, the jth of which has length T_j generations and bears R_j mutations. The sets Ψ_i ($i = 1,\ldots,n$) have as members the indices of the links between the MRCA and the ith leaf. Let the total mutation rate of the gene segment be μ. Then $R_j|T_j \sim \mathrm{Po}(\mu T_j)$, independently. The statistic ρ is the mean number of mutations across all the paths from the MRCA to the leaves: $\rho = \frac{1}{n}\sum_{i=1}^{n} L_i$, where $L_i = \sum_{j\in\Psi_i} R_j$. Hence, $L_i|\{T_j\} \sim \mathrm{Po}(\mu T)$, where T is the time to the MRCA (TMRCA), by the reproductive property of the Poisson distribution, and the fact that each path is of length T. Note, however, that in general the L_is are not independent (unless the Ψ_is are disjoint). So

$E[\rho] = \frac{1}{n}\sum_{i=1}^{n} E[L_i] = \mu T$. Hence ρ/μ is unbiased for T. This argument goes through whether the T_js are considered random (*e.g.*, the result of a coalescent process) or fixed parameters (as in the classical phylogenetic setting).

Comparing age estimates in simulations and real data

In light of the mathematical demonstration, the bias reported by Cox [25] (their Fig 2) is puzzling and we are unable to reproduce it. A simulation aiming for the same conditions as that figure yielded Fig 1, where the TMRCA estimated using ρ/μ is plotted against the true TMCRA of the simulated trees. The line of equality (red solid line) and the least-squares regression line through the origin (black dashed line) are virtually indistinguishable, and the slope of the regression line is not significantly different from 1. The observed bias in this finite sample of 10,000 runs is just 1.7 generations.

Although there is therefore no evidence that ρ is a biased estimator (and indeed a proof that it is not), we compared age estimates from the literature. We used ages that were estimated with both ρ and ML using the same dataset (Fig 2). It is clear that the mtDNA coalescence age estimates are very similar between the two estimators. The same observation has been made for Y-chromosome variation by Batini et al. [37], when comparing ρ and Bayesian inference. There is no observed trend where the age estimates based on ρ are systematically higher or lower (the latter as suggested by Cox [25]) than ML estimates. A correlation between age estimates using both methods displayed a relationship of nearly 1 (0.9925, $R^2 = 0.9951$). Haplogroup L3 is the only one whose age estimates were substantially different between ρ and ML. As previously discussed [8,11], this is due to the high frequency of L3e, associated with the Bantu expansion. However, this in itself shows the random nature of the difference and not a directional bias: if a branch with higher average length like L3a or L3h were the most frequent it could easily lead to an over-estimate in relation to ML.

Coverage of confidence intervals

The coverage of confidence intervals derived from ρ is a pertinent issue, but Cox's discussion of this is compromised at the outset by his expression (3) for the estimated squared standard

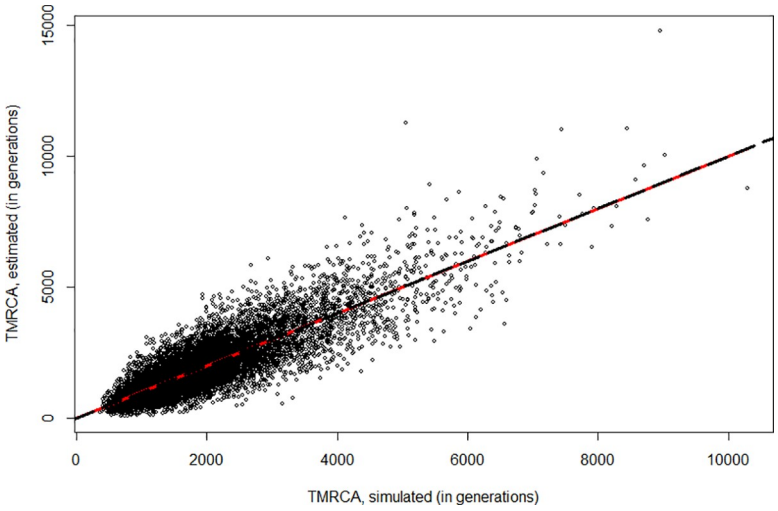

Fig 1. Scatterplot comparing the estimated time to the most recent common ancestor (TMRCA) using ρ and the true simulated time, across 10,000 simulations using a constant-size coalescent process. The line of equality (red) and the least-squares regression (black dashes) are superimposed, meaning that estimated TMRCA with ρ shows no bias.

https://doi.org/10.1371/journal.pone.0212311.g001

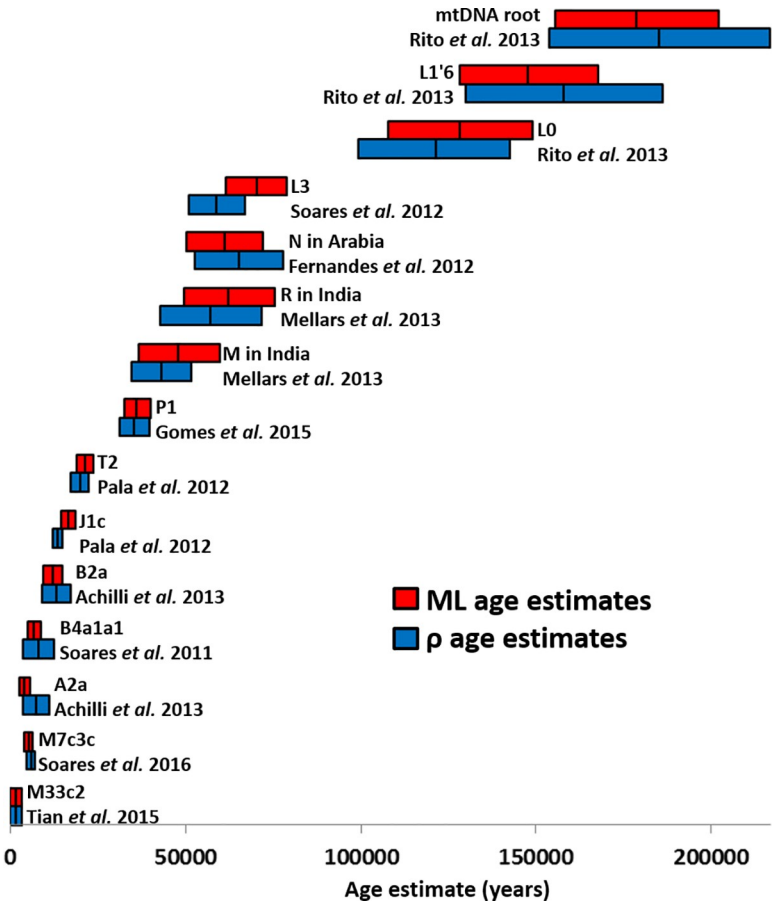

Fig 2. Comparison of age estimates from the literature based on ρ and maximum likelihood (ML) using the molecular clock developed by Soares et al. [14]. ML age estimated were generated using the HKY85 evolution model and gamma distributed-rates using the PAML software [48]. Size of the bars correspond to the 95% confidence interval of each age estimated based in the standard error as obtained by PAML in ML [48] or using the Saillard et al. calculation in ρ estimates [24].

https://doi.org/10.1371/journal.pone.0212311.g002

error of ρ [25], which, to our knowledge, has never before been used to assess the error in ρ. It corresponds neither to the expression proposed by Saillard et al. [24], and used extensively thereafter and herein, which incorporates the dependence of the L_is, nor to the lower bound given by ρ/n, which assumes a perfectly star-like genealogy for the sequences and as a result can seriously under-estimate the error. Cox's expression is a halfway house between these two: it corresponds to a tree which is star-like in the distinct *haplotypes*. Most crucially, it does not describe the increased uncertainty arising from the mutations on internal edges of the tree. Or, to put it another way, it assumes the number of mutations from the root to each distinct haplotype are independent random variables, which since the haplotypes are related by a tree, they in general are not.

We have explored the coverage of the commonly used Wald-style confidence intervals provided by the end points $(\rho \pm 1.96\ \text{e.s.e.}[\rho])/\mu$, where the estimated standard error (e.s.e.) is that given by Saillard et al. [24]. Given that this expression reflects the shape of the underlying genealogy, which is itself influenced by the demography, we should expect different coverage properties under different demographic scenarios. Fig 3 illustrates the estimated coverage as a function of ϑ in the simplest, constant-size, model described above (with 10,000 realizations at

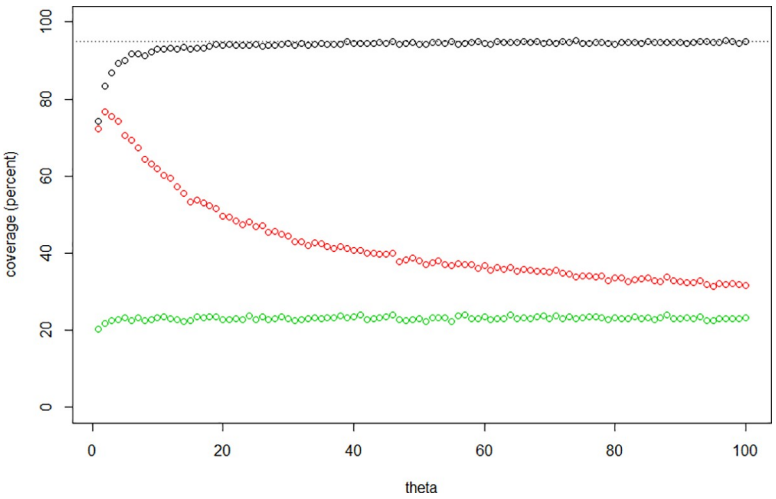

Fig 3. Coverage of Wald confidence intervals for the estimated time to the most recent common ancestor (using ρ) as a function of ϑ, based on Saillard et al. [24] (black circles), Cox [25] (red circles) and lower-bound (green circles) estimates of standard error, obtained from 10,000 simulations using a constant-size coalescent process. The nominal coverage of 95% is indicated with a dotted line.

https://doi.org/10.1371/journal.pone.0212311.g003

each ϑ value). Coverage is indeed anti-conservative for small values of ϑ (when the sampling distribution of ρ is very skewed), but is acceptable if $\vartheta \gtrsim 10$. Note, however, the opposite behaviour of the coverage of the Wald confidence intervals derived from Cox's expression for estimated standard error: it decreases with increasing ϑ, and is always smaller than the Saillard coverage. Not surprisingly, the coverage provided by the lower bound on the standard error, mentioned above, is everywhere extremely poor.

There is no denying that the precision with which ρ/μ estimates the TMRCA depends on the (unknown) demography, since the demography influences the correlation of the L_is (via the tree) and hence the standard error of ρ. A demography that leads to long internal edges in the tree (e.g. constant population size) will lead to much more correlation between certain L_is and hence to data sets that have much less information about the TMRCA, whereas star-like trees (e.g., coming from population expansion) lead to much less correlation and to more informative data sets. No method of estimation could (or should!) get around that. Hence it is doubly important to have a method of estimating the standard error of ρ that accounts well for the correlation of the L_is, as the Saillard standard error estimator does, and Cox's does not (since they are assumed either to be perfectly correlated if they lead to the same haplotype or independent if they lead to distinct haplotypes, and nothing in between).

In this paper we have, like Cox, explored the variability in ρ. The transformation of ρ into an estimate of the age of a node in the phylogeny requires division by a well-calibrated mutation rate. Obtaining this is not a trivial task, and much effort has been invested in it over the years. Any calibration should, at the very least, supply some quantification of error in the estimated mutation rate. Given that, say in the form of an estimated standard error, the delta method [49] provides a quick route to a crude estimate of the standard error of the age of the node of interest [10].

Conclusions

In summary, we have shown that ρ is an unbiased estimator through a mathematical proof; but we additionally supported this conclusion by means of simulations and, empirically, by comparing age estimates for clades simultaneously obtained through ρ and ML. We have also

shown that the coverage of the confidence intervals is only problematic for lower values of ϑ, contrary to previous suggestions. Overall, this shows that ρ should not be dismissed, as suggested; it can play an important role in genetic dating. This is a crucial first step in many lines of research based on phylogenetic analysis, but it is only the first step–discussion of how the estimated dates of nodes in a tree can be interpreted, for example in drawing conclusions about gene flow and population history, is a much larger topic [19].

Author Contributions

Conceptualization: Vincent Macaulay, Pedro Soares, Martin B. Richards.

Data curation: Vincent Macaulay.

Formal analysis: Vincent Macaulay, Pedro Soares.

Funding acquisition: Pedro Soares, Martin B. Richards.

Investigation: Vincent Macaulay, Pedro Soares, Martin B. Richards.

Methodology: Vincent Macaulay, Pedro Soares.

Project administration: Vincent Macaulay, Pedro Soares, Martin B. Richards.

Software: Vincent Macaulay, Pedro Soares.

Supervision: Vincent Macaulay.

Validation: Vincent Macaulay, Pedro Soares, Martin B. Richards.

Visualization: Vincent Macaulay, Pedro Soares.

Writing – original draft: Vincent Macaulay, Pedro Soares, Martin B. Richards.

Writing – review & editing: Vincent Macaulay, Pedro Soares, Martin B. Richards.

References

1. Pala M, Chaubey G, Soares P, Richards MB (2014) The archaeogenetics of European ancestry. Encyclopedia of Life Sciences (ELS). Chichester: John Wiley and Sons.
2. Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31–36. https://doi.org/10.1038/325031a0 PMID: 3025745
3. Posth C, Renaud G, Mittnik A, Drucker Dorothée G, Rougier H, Cupillard C, et al. (2016) Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. Curr Biol 26: 827–833. https://doi.org/10.1016/j.cub.2016.01.037 PMID: 26853362
4. Pereira JB, Costa MD, Vieira D, Pala M, Bamford L, Harich N, et al. (2017) Reconciling evidence from ancient and contemporary genomes: a major source for the European Neolithic within Mediterranean Europe. Proc Biol Sci 284:20161976 https://doi.org/10.1098/rspb.2016.1976 PMID: 28330913
5. Silva M, Oliveira M, Vieira D, Brandão A, Rito T, Pereira JB, et al. (2017) A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. BMC Evol Biol 17: 88. https://doi.org/10.1186/s12862-017-0936-9 PMID: 28335724
6. Sankararaman S, Mallick S, Patterson N, Reich D (2016) The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Curr Biol 26: 1241–1247. https://doi.org/10.1016/j.cub.2016.03.037 PMID: 27032491
7. Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, et al. (2013) The first modern human dispersals across Africa. PLoS One 8: e80031. https://doi.org/10.1371/journal.pone.0080031 PMID: 24236171
8. Soares P, Rito T, Pereira L, Richards MB (2016) A genetic perspective on African prehistory. In: Jones SC, Stewart BA, editors. Africa from MIS 6–2: Population Dynamics and Paleoenvironments. Dordrecht: Springer Netherlands. pp. 383–405.

9. Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, et al. (2012) The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. Am J Hum Genet 90:347–55. https://doi.org/10.1016/j.ajhg.2011.12.010 PMID: 22284828
10. Mellars P, Gori KC, Carr M, Soares PA, Richards MB (2013) Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc Natl Acad Sci U S A 110: 10699–10704. https://doi.org/10.1073/pnas.1306043110 PMID: 23754394
11. Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, et al. (2012) The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol 29: 915–927. https://doi.org/10.1093/molbev/msr245 PMID: 22096215
12. Macaulay VA, Richards MB, Forster P, Bendall KE, Watson E, Sykes B, et al. (1997) mtDNA mutation rates—no need to panic. Am J Hum Genet 61: 983–990. https://doi.org/10.1016/S0002-9297(07)64211-6 PMID: 9382113
13. Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373–387. https://doi.org/10.1534/genetics.105.043901 PMID: 16172508
14. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. (2009) Correcting for purifying selection: An improved human mitochondrial molecular clock. Am J Hum Genet 84: 740–759. https://doi.org/10.1016/j.ajhg.2009.05.001 PMID: 19500773
15. Cavadas B, Soares P, Camacho R, Brandão A, Costa MD, Fernandes V, et al. (2015) Fine time scaling of purifying selection on human nonsynonymous mtDNA mutations based on the worldwide population tree and mother–child pairs. Hum Mutat 36: 1100–1111. https://doi.org/10.1002/humu.22849 PMID: 26252938
16. Soares P, Abrantes D, Rito T, Thomson N, Radivojac P, Li B, et al. (2013) Evaluating purifying selection in the mitochondrial DNA of various mammalian species. PLoS One 8:e58993 https://doi.org/10.1371/journal.pone.0058993 PMID: 23533597
17. Endicott P, Ho SYW, Stringer C (2010) Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. J Hum Evol 59: 87–95. https://doi.org/10.1016/j.jhevol.2010.04.005 PMID: 20510437
18. Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, et al. (2013) A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 23: 553–559. https://doi.org/10.1016/j.cub.2013.02.044 PMID: 23523248
19. Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, et al. (2016) Resolving the ancestry of Austronesian-speaking populations. Hum Genet 135: 309–326. https://doi.org/10.1007/s00439-015-1620-z PMID: 26781090
20. Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, et al. (2011) Ancient voyaging and Polynesian origins. Am J Hum Genet 88: 239–247. https://doi.org/10.1016/j.ajhg.2011.01.009 PMID: 21295281
21. Silva M, Alshamali F, Silva P, Carrilho C, Mandlate F, Trovoada MJ, et al. (2015) 60,000 years of interactions between central and eastern Africa documented by major African mitochondrial haplogroup L2. Sci Rep 5: 12526. https://doi.org/10.1038/srep12526 PMID: 26211407
22. Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Kashani BH, et al. (2013) Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc Natl Acad Sci U S A 110: 14308–14313. https://doi.org/10.1073/pnas.1306290110 PMID: 23940335
23. Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of native American mtDNA variation: A reappraisal. Am J of Hum Genet 59: 935–945.
24. Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S (2000) mtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. Am J of Hum Genet 67: 718–726.
25. Cox MP (2008) Accuracy of molecular dating with the ρ statistic: Deviations from coalescent expectations under a range of demographic models. Hum Biol 80: 335–357. https://doi.org/10.3378/1534-6617-80.4.335 PMID: 19317593
26. Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc Natl Acad Sci U S A 97: 7360–7365. PMID: 10861004
27. Kulichova I, Fernandes V, Deme A, Novackova J, Stenzl V, Novelletto A, et al. (2017) Internal diversification of non-Sub-Saharan haplogroups in Sahelian populations and the spread of pastoralism beyond the Sahara. Am J Phys Anthropol 164: 424–434. https://doi.org/10.1002/ajpa.23285 PMID: 28736914
28. Li YC, Wang HW, Tian JY, Li RL, Rahman ZU, Kong QP (2018) Cultural diffusion of Indo-Aryan languages into Bangladesh: A perspective from mitochondrial DNA. Mitochondrion 38: 23–30. https://doi.org/10.1016/j.mito.2017.07.010 PMID: 28764911

29. de Saint Pierre M (2017) Antiquity of mtDNA lineage D1g from the southern cone of South America supports pre-Clovis migration. Quat Int 444: 19e25.
30. Nagle N, van Oven M, Wilcox S, van Holst Pellekaan S, Tyler-Smith C, Xue Y, et al. (2017) Aboriginal Australian mitochondrial genome variation—an increased understanding of population antiquity and diversity. Sci Rep 7: 43041. https://doi.org/10.1038/srep43041 PMID: 28287095
31. Wei W, Gomez-Duran A, Hudson G, Chinnery PF (2017) Background sequence characteristics influence the occurrence and severity of disease-causing mtDNA mutations. PLoS Genet 13: e1007126. https://doi.org/10.1371/journal.pgen.1007126 PMID: 29253894
32. Zhabagin M, Balanovska E, Sabitov Z, Kuznetsova M, Agdzhoyan A, Balaganskaya O, et al. (2017) The connection of the genetic, cultural and geographic landscapes of Transoxiana. Sci Rep 7: 3085. https://doi.org/10.1038/s41598-017-03176-z PMID: 28596519
33. D'Atanasio E, Trombetta B, Bonito M, Finocchio A, Di Vito G, Seghizzi M, et al. (2018) The peopling of the last Green Sahara revealed by high-coverage resequencing of trans-Saharan patrilineages. Genome Biol 19: 20. https://doi.org/10.1186/s13059-018-1393-5 PMID: 29433568
34. Sole-Morata N, Villaescusa P, Garcia-Fernandez C, Font-Porterias N, Illescas MJ, Valverde L, et al. (2017) Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ. Sci Rep 7: 7341. https://doi.org/10.1038/s41598-017-07710-x PMID: 28779148
35. Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, et al. (2015) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol 32: 661–673. https://doi.org/10.1093/molbev/msu327 PMID: 25468874
36. Scozzari R, Massaia A, Trombetta B, Bellusci G, Myres NM, Novelletto A, et al. (2014) An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res 24: 535–544. https://doi.org/10.1101/gr.160788.113 PMID: 24395829
37. Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, et al. (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun 6:7152. https://doi.org/10.1038/ncomms8152 PMID: 25988751
38. Maisano Delser P, Neumann R, Ballereau S, Hallast P, Batini C, Zadik D, et al. (2017) Signatures of human European Palaeolithic expansion shown by resequencing of non-recombining X-chromosome segments. Eur J Hum Genet 25: 485–492. https://doi.org/10.1038/ejhg.2016.207 PMID: 28120839
39. Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C (2013) Human Evolutionary Genetics. Oxford: Garland Scince.
40. Stoneking M (2016) An Introduction to Molecular Anthropology. Oxford: Wiley Blackwell.
41. Balanovsky O (2017) Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. Hum Genet 136: 575–590. https://doi.org/10.1007/s00439-017-1805-8 PMID: 28455625
42. Stoneking M, Delfin F The human genetic history of East Asia: Weaving a complex tapestry. Curr Biol 20: R188–R193. https://doi.org/10.1016/j.cub.2009.11.052 PMID: 20178766
43. Team RDC (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: the R Foundation for Statistical Computing.
44. Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, Reidla M, et al. (2012) Mitochondrial DNA signals of late glacial recolonization of Europe from Near Eastern refugia. Am J Hum Genet 90: 915–924. https://doi.org/10.1016/j.ajhg.2012.04.003 PMID: 22560092
45. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, et al. (2005) Saami and Berbers—An unexpected mitochondrial DNA link. Am J Hum Genet 76: 883–886. https://doi.org/10.1086/430073 PMID: 15791543
46. Tian JY, Wang HW, Li YC, Zhang W, Yao YG, van Straten J, et al. (2015) A genetic contribution from the Far East into Ashkenazi Jews via the ancient Silk Road. Sci Rep 5: 8377. https://doi.org/10.1038/srep08377 PMID: 25669617
47. Gomes SM, Bodner M, Souto L, Zimmermann B, Huber G, Strobl C, et al. (2015) Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocenic mtDNA diversity. BMC Genomics 16: 1–20. https://doi.org/10.1186/1471-2164-16-1
48. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13: 555–556.
49. Dorfman R (1938) A note on the delta-method for finding variance formulae. The Biometric Bulletin 1: 129–137