Proceedings of the Royal Society B

Supplementary Data

Reconciling evidence from ancient and contemporary genomes: A major source for the European Neolithic within Mediterranean Europe

Joana B. Pereira, Marta D. Costa, Daniel Vieira, Maria Pala, Lisa Bamford, Nourdin Harich, Lotfi Cherni, Farida Alshamali, Jiři Hatina, Sergey Rychkov, Gheorghe Stefanescu, Turi King, Antonio Torroni, Pedro Soares, Luísa Pereira, Martin B. Richards

Methods

We performed the complete sequencing of 203 mtDNAs previously classified into haplogroup JT by the characterisation of the HVS-I region, as described elsewhere [1]. The population surveyed includes samples from Europe, especially from the Mediterranean Basin and Iberia, the Caucasus, the Near East and North Africa (Supplementary Data 1). We sequenced the samples using Sanger sequencing and scored whole-DNA sequencing results against the rCRS [2, 3]. We also included 1746 published JT sequences (883 in haplogroup J and 863 in haplogroup T). We inferred a preliminary phylogeny with Network 4.6 software using reduced-median network analysis [4] together with PhyloTree [5] which we resolved into a putative most parsimonious tree using estimates of site-specific rate variation [6]. We used mtDNA-GeneSyn to convert files [7]. The phylogenetic trees for haplogroup J and T mitogenomes are shown in Supplementary Data 1.

For the time estimates we used two different methods: the rho (ρ) statistic [8] and maximum likelihood (ML) [9]. For both methods, we used a mutation rate estimate for the whole-mtDNA (mitogenome) sequence of one substitution every 3,624 years, corrected for purifying selection [6]. We also used ρ with a synonymous mutation rate of one substitution every 7,884 years [6]. For the ML age estimates, we used the PAML software, assuming the HKY85 model [10] with gamma-distributed rate variation across sites (with 0.1 as the initial value of the shape parameter estimated for a gamma distribution with 32 categories) and an enforced molecular clock. Due to limitations of the PAML software, we only used 1700 sequences of our tree in order to obtain the age estimates. For the calculation of the ML time estimates, we considered two partitions: HVS-I (positions 16051-16400) plus HVS-II (positions 68-263) and the remainder, due to the high mutation rate of the hypervariable regions when compared with the remaining molecule. Age estimates for the main nodes in the trees are shown in Table S1 (in Supplementary Data 2).

Founder analysis [11-13] is an approach to estimating migration times from genetic data. It partitions the data into putative source and sink regions, and subtracts the diversity within the sink

dataset that arose in the source region. The founder clusters diversifying from each founder lineage identified in this way are then dated by means of the ρ statistic and the mtDNA molecular clock. By this means we effectively reset the molecular clock to zero for each founder lineage. Several procedures are used to reduce the confounding impact of back-migration and back-mutation. The procedure is conservative in that it provides only minimum estimates for the arrival time of each founder lineage, since the arrival necessarily predates the origin and expansion of the corresponding founder cluster.

We performed founder analysis using in-house Founder Analysis software that applies an algorithm for choosing the best tree. We calculated the age estimate for the migration of each founder using the ρ statistic and obtained the effective number of samples associated with each founder as before [13-15] by multiplying the number of samples in each founder cluster by a ratio of the variance assuming a star-like network and the variance calculated with the method of Saillard et al. (Supplementary Data 2) [16]. We performed the founder analysis of mitogenome data in three ways: (1) from the Near East (including Anatolia) to Mediterranean Europe (including Iberia); (2) from the Near East and eastern/central Mediterranean Europe to Iberia; and (3) from the Near East and Mediterranean Europe to central/northern Europe (excluding the British Isles and Volga Tatars from Russia).

Additionally, we performed the founder analysis considering Mediterranean Europe and the Near East as independent sources to Iberia as well as to central and northern Europe. We also performed a reciprocal founder analysis, reversing the migration direction of all the models described to confirm that all sources and sink regions were correctly assigned.

In order to avoid being misled by false founders due to back-migration, we specified that the founder clade must include at least one specific basal branch in the source. We estimated the age of each founder using the ρ statistic with an independent mutation rate for each migration event, considering the timeframe of the event to allow for the effects of purifying selection, as before [13].

We scanned the distribution of the founders at equally spaced 200-year intervals from 0–70 ka [13-15], and we also calculated the proportion of the founder lineages in a two-migration model that approximates to potential Late Glacial and Neolithic migration events from the Near East to the Mediterranean Europe (defined at 13.0 and 9.0 ka). We further assumed a third dispersal at 0.5 ka to allow for any recent/historical gene flow.

We obtained Bayesian skyline plots (BSPs) [17] using the BEAST software and assuming the HKY nucleotide substitution model with gamma-distributed rates and an uncorrelated lognormal relaxed molecular clock. The mutation rate used was the one obtained previously for the U6 haplogroup, calculated using BEAST and taking into account particular demographic events internal to the phylogeny [18]. The MCMC chain length was 100,000,000 and parameters were drawn every 10,000 interactions with a 10% burn-in discarded. We used a generation time of 25 years to rescale the vertical axis of the BSP into years [19].

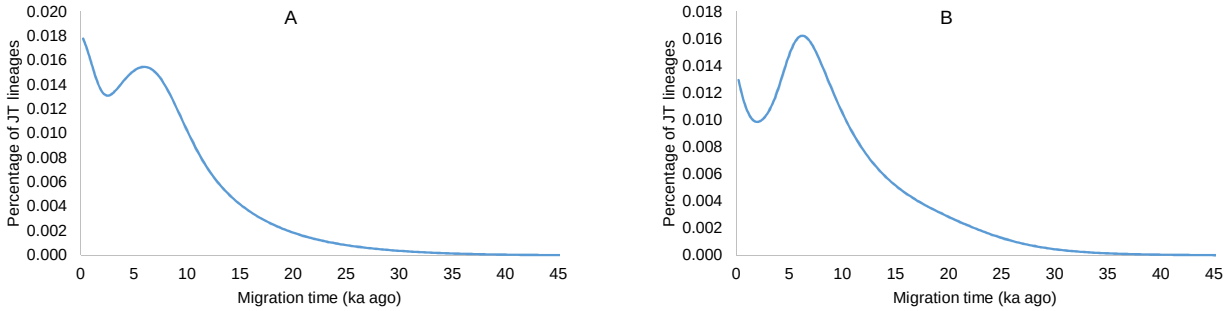

Figure S1. Founder analysis results for whole mtDNA genomes. Probabilistic distribution across migration times scanned at 200-year intervals from 0-70 ka. (A) From Mediterranean Europe to Iberia. (B) From the Near East to Iberia.

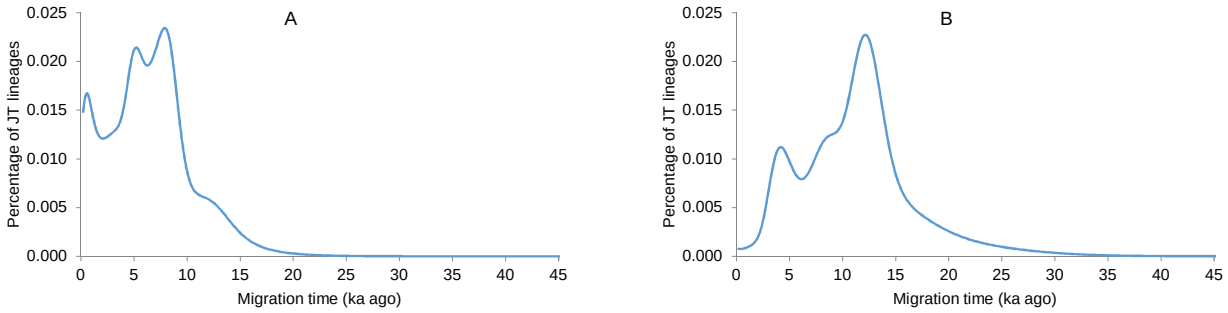

Figure S2. Founder analysis results for whole mtDNA genomes. Probabilistic distribution across migration times scanned at 200-year intervals from 0-70 ka. (A) From Mediterranean Europe to northern/central Europe. (B) From the Near East to northern/central Europe.

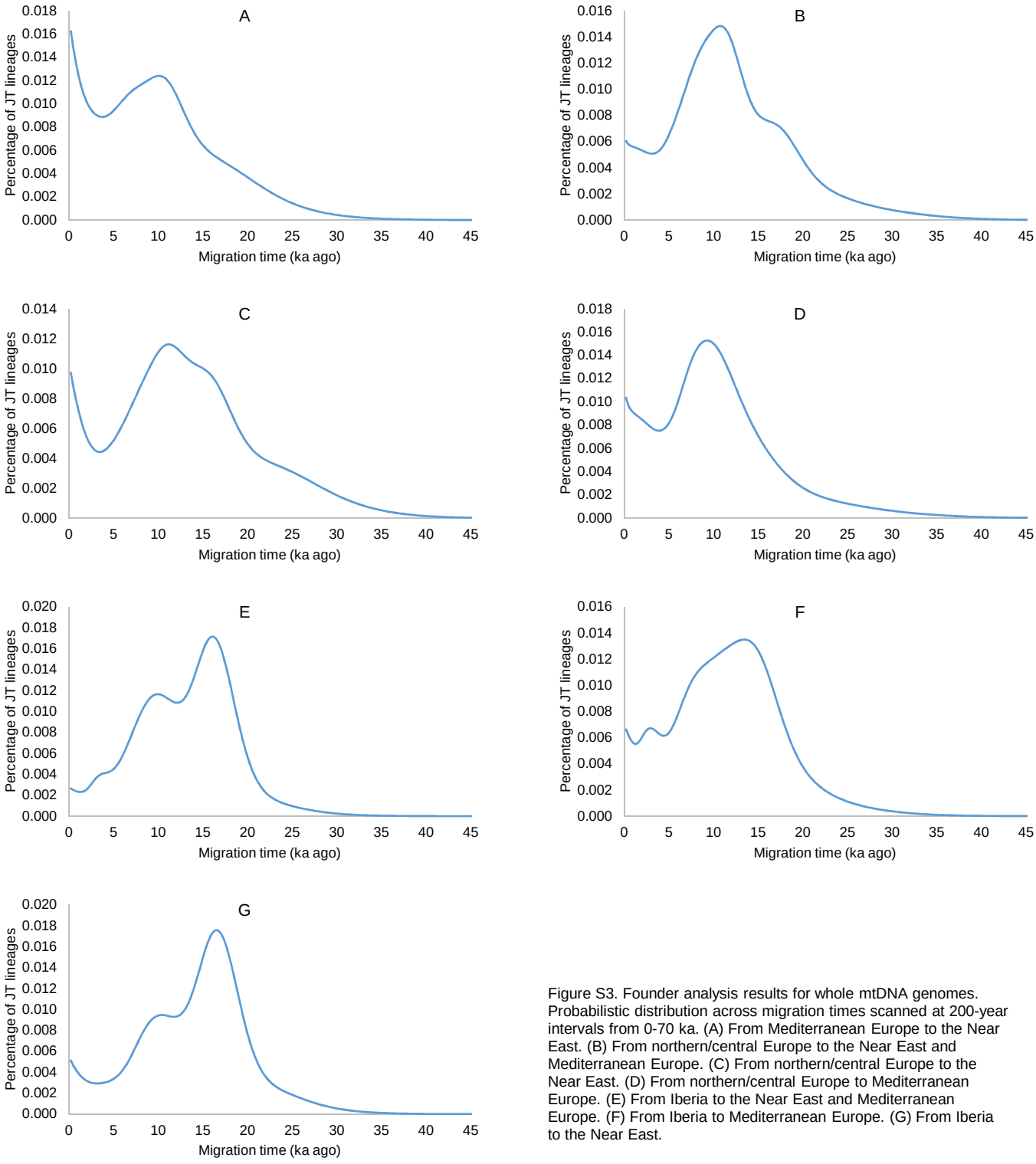

Figure S3. Founder analysis results for whole mtDNA genomes. Probabilistic distribution across migration times scanned at 200-year intervals from 0-70 ka. (A) From Mediterranean Europe to the Near East. (B) From northern/central Europe to the Near East and Mediterranean Europe. (C) From northern/central Europe to the Near East. (D) From northern/central Europe to Mediterranean Europe. (E) From Iberia to the Near East and Mediterranean Europe. (F) From Iberia to Mediterranean Europe. (G) From Iberia to the Near East.

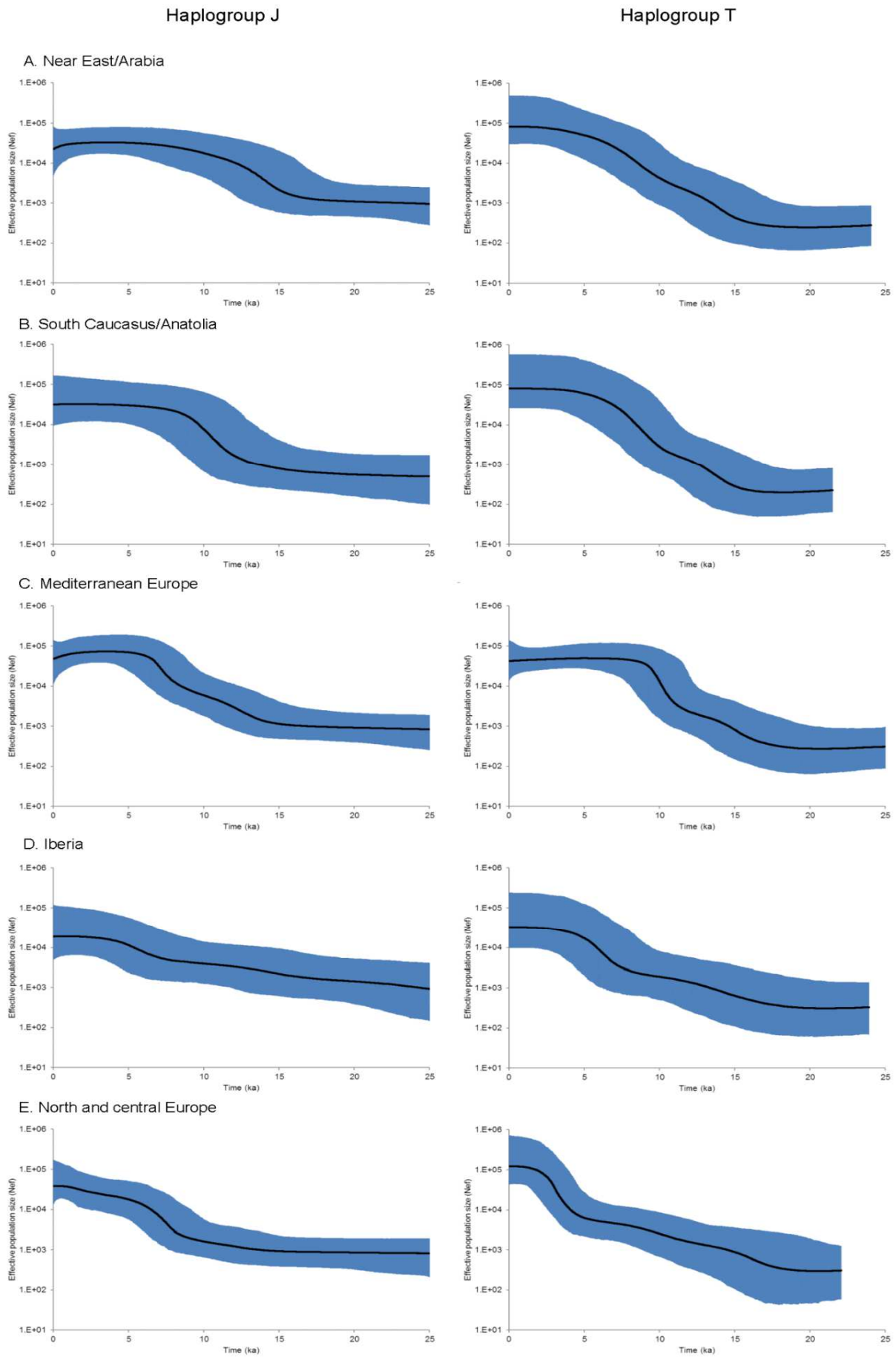

Figure S4. BSPs indicating the median of the effective population size trough time based on haplogroup J (left panel) and T (right panel), assuming a generation time of 25 years. The blue area represents the 95% confidence intervals. (A) Near East/Arabia; (B) South Caucasus/Anatolia; (C) Mediterranean Europe; (D) Iberia; (E) North and central Europe.

References

1. Costa M.D., Pereira J.B., Pala M., Fernandes V., Olivieri A., Achilli A., Perego U.A., Rychkov S., Naumova O., Hatina J., et al. 2013 A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. *Nature communications* **4**, 2543. (doi:10.1038/ncomms3543).
2. Andrews R.M., Kubacka I., Chinnery P.F., Lightowlers R.N., Turnbull D.M., Howell N. 1999 Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics* **23**(2), 147. (doi:10.1038/13779).
3. Pala M., Olivieri A., Achilli A., Accetturo M., Metspalu E., Reidla M., Tamm E., Karmin M., Reisberg T., Hooshiar Kashani B., et al. 2012 Mitochondrial DNA signals of Late Glacial recolonization of Europe from Near Eastern refugia. *Am J Hum Genet* **90**(5), 915-924. (doi:10.1016/j.ajhg.2012.04.003).
4. Bandelt H.J., Forster P., Sykes B.C., Richards M.B. 1995 Mitochondrial portraits of human populations using median networks. *Genetics* **141**(2), 743-753.
5. van Oven M., Kayser M. 2009 Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30**(2), E386-394. (doi:10.1002/humu.20921).
6. Soares P., Ermini L., Thomson N., Mormina M., Rito T., Rohl A., Salas A., Oppenheimer S., Macaulay V., Richards M.B. 2009 Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* **84**(6), 740-759. (doi:10.1016/j.ajhg.2009.05.001).
7. Pereira L., Freitas F., Fernandes V., Pereira J.B., Costa M.D., Costa S., Maximo V., Macaulay V., Rocha R., Samuels D.C. 2009 The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet* **84**(5), 628-640. (doi:10.1016/j.ajhg.2009.04.013).
8. Forster P., Harding R., Torroni A., Bandelt H.J. 1996 Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* **59**(4), 935-945.
9. Yang Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**(5), 555-556.
10. Hasegawa M., Kishino H., Yano T. 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**(2), 160-174.
11. Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V., Rengo C., Sellitto D., Cruciani F., Kivisild T., et al. 2000 Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**(5), 1251-1276. (doi:dx.doi.org/10.1016/S0002-9297(07)62954-1).
12. Macaulay V., Richards M.B. 2013 Mitochondrial genome sequences and their phylogeographic interpretation. In *Encyclopedia of Life Sciences (ELS)* (Chichester, John Wiley & Sons, Ltd.
13. Soares P.A., Trejaut J.A., Rito T., Cavadas B., Hill C., Eng K.K., Mormina M., Brandão A., Fraser R.M., Wang T.Y., et al. 2016 Resolving the ancestry of Austronesian-speaking populations. *Hum Genet* **135**(3), 309-326. (doi:10.1007/s00439-015-1620-z).
14. Soares P., Alshamali F., Pereira J.B., Fernandes V., Silva N.M., Afonso C., Costa M.D., Musilova E., Macaulay V., Richards M.B., et al. 2012 The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol Biol Evol* **29**(3), 915-927. (doi:10.1093/molbev/msr245).
15. Rito T., Richards M.B., Fernandes V., Alshamali F., Cerny V., Pereira L., Soares P. 2013 The first modern human dispersals across Africa. *PLoS One* **8**(11), e80031. (doi:10.1371/journal.pone.0080031).
16. Saillard J., Forster P., Lynnerup N., Bandelt H.-J., Nørby S.S. 2000 mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**(3), 718-726. (doi:10.1086/303038).
17. Drummond A.J., Rambaut A., Shapiro B., Pybus O.G. 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**(5), 1185-1192. (doi:10.1093/molbev/msi103).
18. Pereira L., Silva N.M., Franco-Duarte R., Fernandes V., Pereira J.B., Costa M.D., Martins H., Soares P., Behar D.M., Richards M.B., et al. 2010 Population expansion in the North African Late

Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evol Biol* **10**(1), 390. (doi:10.1186/1471-2148-10-390).
19. Atkinson Q.D., Gray R.D., Drummond A.J. 2008 mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol* **25**(2), 468-474. (doi:10.1093/molbev/msm277).