Current Biology, Volume *23*

Supplemental Information

A Revised Timescale for Human Evolution

Based on Ancient Mitochondrial Genomes

Qiaomei Fu, Alissa Mittnik, Philip L.F. Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, Liane Giemsch, Ralf Schmitz, Joachim Burger, Anna Maria Ronchitelli, Fabio Martini, Renata G. Cremonesi, Jiří Svoboda, Peter Bauer, David Caramelli, Sergi Castellano, David Reich, Svante Pääbo, and Johannes Krause

Supplemental Experimental Procedures

SI 1. Skeletal Samples. These included three individuals of the triple burial from the Czech site of Dolni Vestonice where directly associated charcoal was radiocarbon dated to 26,640 ± 110 BP (31,500 calBP[1]), a bone sample from the original Cro-Magnon 1 type specimen[2], 43 samples from Upper Paleolithic sites in Italy (Grotta Paglicci, Grotta Continenza, Grotta delle Veneri, Grotta Romanelli, Grotta del Romito, Grotta di San Teodoro), two individuals of the Oberkassel double burial near Bonn, Germany, radiocarbon dated to 11,570 ± 100 and 12,180 ± 100 BP (14,020 and 13,430 calBP respectively), three Upper Paleolithic remains from Sandalja Cave in Croatia dated to 12,320 ± 100 BP (14,500 calBP), and one Mesolithic individual from the Loschbour rock shelter in Luxembourg radiocarbon dated to 7,205 ± 50 BP (8,000 calBP[3]). From Asia we obtained a bone sample from the Boshan 11 fossil from Boshan, Shandong, China, radiocarbon dated to 7,368 ± 34 BP (8,180 calBP). Our analysis also included the published complete mtDNAs of Upper Paleolithic humans from Tianyuan cave from China radiocarbon dated to 34,430 ± 510 BP (39,475 calBP[4]) and Kostenki 14 from Russia radiocarbon dated to 33,250 ± 500 BP (37,985 calBP[5]), the Tyrolean iceman complete mtDNA from Italy dated to 4,550 calBP[6] and the mtDNA sequence from the Saqqaq individual from Greenland dated to 4,044 ± 31 BP (3,600-4,170 calBP[7]) (Table 1 and Table S6).

SI 2. DNA extraction and enrichment. All extraction and library preparation steps before amplification were performed in clean-room facilities at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig, Germany. Using a dental drill, 40–160 mg of bone powder was collected from each sample from which DNA was then extracted following an established protocol using a guanidinium-silica based method[8]. A 20 µl aliquot of each extract was used to produce indexed libraries according to a modified Illumina multiplex protocol[9]. The libraries were enriched for human mtDNA in a bead-capture method using long-range PCR products as bait for hybridization as described previously[10]. One negative control each was carried along for every step of DNA extraction and library preparation.

SI 3. Illumina sequencing and analysis. High-throughput DNA sequencing for the enriched library pools was carried out on the Illumina Genome Analyzer IIx platform using 2 × 76 + 7 cycles according to the manufacturer’s instructions for multiplex sequencing (FC-104-400x v4 sequencing chemistry and PE-203-4001 v4 cluster generation kit). The manufacturer’s protocol was followed with the exception that the raw reads were aligned to the PhiX 174 reference sequence to obtain a training data set for the base caller Ibis[11]. Raw reads called by Ibis 1.1.1 were filtered according to the individual indices. Adapter and index sequences were removed and paired-end reads overlapping for at least 11 nucleotides were collapsed to one fragment where the base with the higher quality score was called in the overlapping sequence. The sequences were mapped to the revised Cambridge Reference Sequence (rCRS, NC_012920) using a custom iterative mapping assembler[12, 13].

Authenticity of the sequences was assessed by an analysis of DNA damage patterns expected for ancient DNA as well as by identifying diagnostic positions that differ from a set of 311 modern human mtDNAs[12]. The scarcity of diagnostic positions led us to develop a more powerful contamination estimator that leverages information from sites that vary within the 311 modern human mtDNAs as well as fixed diagnostic positions (see SI text 5 for details).

SI 4. Likelihood ratio test

The molecular clock test was performed in MEGA5 by comparing the ML value for the given topology with and without the molecular clock constraints under General Time Reversible model (+G+I) using a dataset of 54 complete worldwide mtDNAs as well as including 7 Neandertals, 1 Denisova and 2 chimpanzee mtDNAs. Differences in evolutionary rates among sites were modeled using a discrete Gamma (G) distribution (shape parameter shown) and allowed for invariant (*I*) sites to exist (estimate of percent invariant sites shown). The null hypothesis of equal evolutionary rate throughout the tree was not rejected at a 5% significance level for both the test on just the 54 modern humans ($P < 0.517$) and including Neandertals, Denisova and chimpanzee ($P < 0.181$). The analyses involved 54 and 64 nucleotide sequences with a total of 16548 and 15618 positions, respectively. All positions containing gaps and missing data were eliminated.

SI 5. Likelihood-based method for contamination estimation

We want to estimate contamination in mitochondrial ancient DNA extracted from the bones of early modern humans. We assume that contamination is less than 50% as determined by another means (e.g. determined from the amount of aDNA damage). Given this assumption, the primarily challenges are that we do not know:

1. the number of distinct contaminating individuals
2. the frequency of present-day human mtDNA haplotypes in the contaminator population

We solve this problem by deriving a probabilistic model and using Markov chain Monte Carlo (MCMC) to estimate the proportion authentic.

Data

Our input data consist of mtDNA reads from the ancient sample and a set of current human full-length mitochondrial genomes that encompass all plausible contaminating sequences. The ancient reads are combined to form a consensus sequence. Here we assume that this consensus represents the true aDNA sequence (i.e., contamination + error is < 50% and the depth of coverage is sufficiently high such that the majority base is correct).

Notation and parameters

We have n reads from our aDNA sample, the consensus ancient mitochondrial genome, and m present-day human mitochondrial genomes ($m = 311$ for our data). Essentially, we consider the mtDNA reads to be drawn from a mixture of these $m+1$ genomes. For each aDNA read $i \in \{1,\ldots,n\}$ and each mitochondrial genome $j \in \{0,\ldots,m\}$ where $j=0$ represents the consensus aDNA genome, we summarize the data with three numbers:

- $M_{i,j} \rightarrow$ the number of bases in read i for which the read matches genome j (i.e. evidence that this read comes from genome j).
- $N_{i,j} \rightarrow$ the number of bases in read i for which the read does not match genome j (i.e. error or evidence that this read does not come from j). This does not include bases involved in insertions or deletions.
- $I_{i,j} \rightarrow$ whether (1) or not (0) read i contains an insertion or deletion relative to genome j (i.e. error or strong evidence this read does not come from j).

For notational convenience, define $D_{i,j} = \{M_{i,j}, N_{i,j}, I_{i,j,}\}$ and $D_{i,\cdot} = \{D_{i,j} : j \in \{0,\ldots,m\}\}$. We discard reads that contain insertions or deletions not found in any of the $m+1$ potential source genomes, on the assumption that these changes arose through a difficult-to-model error process. We assume single base sequencing errors arise with probability ε per base and that reads are drawn from a multinomial (mixture) distribution with unknown proportions $p = \{p_0,\ldots,p_m\}$ where $\sum_j p_j = 1$.

Likelihood

We wish to calculate the likelihood of the data (D) given the parameters (ε, p). We will assume sequencing error operates at the level of individual bases within a read with an error equally likely to occur at any base. In contrast, contamination operates at the level of the read, with the entire read being from a single source genome (whether authentic or contaminant). We begin by using the knowledge that reads are independent:

$$\Pr(D \mid \varepsilon, p) = \prod_i \Pr(D_{i,\cdot} \mid \varepsilon, p)$$

Then we condition on the source genome of the read, j, and calculate the probability of the observed data given this source ($D_{i,j}$):

$$\begin{aligned}\Pr(D_{i,\cdot} \mid \varepsilon, p) &= \sum_j p_j \Pr(D_{i,j} \mid \varepsilon) \\ &= \sum_j p_j \binom{M_{i,j} + N_{i,j}}{M_{i,j}} (1-\varepsilon)^{M_{i,j}} \varepsilon^{N_{i,j}} (1 - I_{i,j})\end{aligned}$$

Parameter estimation

We estimate the error rate ε before estimating the full mixture proportions p. Most of the information about the error rate comes from regions of the mtDNA where all $m+1$ genomes are identical --- the rate at which reads differ from this fixed sequence is the error rate. I estimate this rate by dividing the number of bases that differ in these regions by the total number of bases in these regions. For any realistic amount of data, this estimate has negligible remaining uncertainty.

Given the error rate, we use a Markov chain Monte Carlo algorithm to estimate the proportion authentic under the more general model in which contamination can arise from an arbitrary number of individuals ($m > 1$). Note that this is feasible in part because the individual proportions of contaminants are nuisance parameters – we care only about the total contamination proportion (or, equivalently, the proportion authentic: $\sum_{j=1}^{m} p_j = 1 - p_0$). We apply a two-stage Gibbs sampler in which we augment the model with latent variables Z_i assigning each read i to one of the $m+1$ genomes. Since we know little about the potential contaminants, we chose an uninformative prior using a uniform marginal distribution on the proportion authentic (p_0) and a symmetric Dirichlet joint distribution over all possible contaminant proportions $(p_1,\ldots,p_m)$:

$$\Pr(p \mid \alpha) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{m} p_j^{\alpha_j - 1} / (1 - p_0)$$

where the vector α is the concentration hyperparameter of the Dirichlet distribution. We set $\alpha_{1 \le j \le m} = 1$.

Now we need to draw samples from a Markov chain with stationary distribution corresponding to our desired posterior distribution:

$$\Pr(p \mid D, \varepsilon, \alpha) \propto \Pr(D \mid p, \varepsilon) \Pr(p \mid \alpha)$$

We draw samples $p^{(1)},\ldots,p^{(t)},\ldots$ from the posterior by iterating the following two steps:

1. Draw an assignment of reads (i) to genomes (j), conditional on the current contamination proportions ($p^{(t)}$):

$$Z_i^{(t+1)} \sim \Pr(j \mid p^{(t)}, \varepsilon, D_{i,\cdot}) \propto \Pr(D_{i,j} \mid \varepsilon) p_j^{(t)}$$

2. Draw a new sample from the posterior, conditional on these assignments:

$$\begin{aligned} p^{(t+1)} &\sim \Pr(p^{(t+1)} \mid Z^{(t+1)}, \alpha) = \Pr(p_{1,\ldots,m}^{(t+1)} \mid p_0^{(t+1)}, Z^{(t+1)}, \alpha) \Pr(p_0^{(t+1)} \mid Z^{(t+1)}, \alpha) \\ &= f_D(p_{1,\ldots,m}^{(t+1)} / (1 - p_0^{(t+1)}); \alpha_{1,\ldots,m} + \eta_{1,\ldots,m}) \cdot f_B(p_0^{(t+1)}; 1 + \eta_0; 1 + n - \eta_0) \end{aligned}$$

$$\eta_j = \sum_i 1_{Z_i = j}$$

where f_D is the density of a Dirichlet distribution (conjugate to Dirichlet prior), f_B is the density of a Beta distribution (conjugate to uniform prior), and η summarizes the Z variables by counting the number of assignments to each genome j.

We start the chain with mixture proportions drawn at random from the prior distribution. After running the chain sufficiently long to reach stationarity (i.e., after ``burn-in"), the 2.5% and 97.5% quantiles of subsequent $p_0^{(t)}$ samples correspond to our 95% credibility interval for the proportion authentic, p_0.

SI 6. Calculation of mtDNA substitution rates using radiocarbon dated ancient modern humans and Neanderals The four Neandertal individuals Vindija33.16[14], Sidron1253[15], Feldofer1[16] and Feldofer2[16] used are radiocarbon dated to 38,310 ± 2,130 BP, 38,790 BP, 39,900 ± 620 BP and 39,240 ± 670 BP (44,003 , 43,129, 43,926, 43,507 calBP) respectively. The 10 ancient human mtDNAs are identical to the ones used in the Bayesian analysis in the main text (Table S1). To calculate substitution rates with inclusion of the Neandertals, we used the contemporary dataset comprising 54 instead of the 311 globally distributed mtDNAs for the coding region and whole mtDNA only (Table S5). The calculated substitution rates for the mtDNA coding region and whole mtDNA largely overlaps with our previous estimates using the radiocarbon dated AMHs only and the dataset of 311mtDNAs for relaxed and fixed clock as well as the partitioned datasets. In theory, mitochondrial substitution rates could have changed between Neandertals and modern humans, we therefore concentrate on the susbtitution rates determined with the AMH only in the main text of the manuscript.

Supplemental Data

Table S1 (related to Table 2). Substitution rates using fossil mtDNAs separately for coding region.

Coding region	Substitution rate	HPD 95%	
Sample	**mean**	**lower**	**upper**
Tianyuan	1.55E-08	6.59E-09	2.36E-08
Kostenki	1.14E-08	4.42E-09	1.85E-08
DolniVestonice13	1.26E-08	4.22E-09	2.04E-08
DolniVestonice14	2.19E-08	1.45E-08	2.85E-08
Oberkassel998	1.89E-08	6.12E-09	3.04E-08
BS11	4.50E-08	9.55E-11	7.67E-08
Loschbour	3.62E-08	5.70E-09	6.60E-08
Iceman	1.89E-08	6.12E-09	3.04E-08
Eskimo	1.45E-08	8.73E-11	3.73E-08
CroMagnon	3.78E-08	3.87E-11	1.10E-07

Table S2 (related to Figure 1). Haplogroup assignment.

Sample	Haplogroup	Additional substitutions
Dolni Vestonice 13	U8	151,3480,7031,10398,16189
Dolni Vestonice 14	U	16192, 16270
Dolni Vestonice 15	U	16192, 16270
Continenza 7	U5b2b1	
Paglicci Accesso sala 2 Rim P	U2'3'4'7'8'9	146, 150, 152, 5999, 6152, 6498C→A, 8860, 10020, 14152, 15326, 15466, 16274, 16297
Paglicci Str. 4b	H1	1346, 4084, 6044, 9110
Oberkassel 998	U5b1	16192!
Oberkassel 999	U5b1	16189, 16166
Tianyuan	B	5836,5348,11257,16293
Boshan11	B4c1a	14133!, 16311!, 16519
Kostenki	U2	13269,15262,542,711
Iceman	K1	16519

Saqqaq	D2a1	5178!,16092 14226 11234
Cro Magnon	T2b	15148,16519,6620,1871,15884,4491,235
Loschbour	U5b1a	16189!, 6701

Mutations toward a base identical by state to the rCRS are indicated with an exclamation mark (!)

Table S3 (related to Evolutionary analysis in main text). MtDNA diversity in present-day humans, ancient Europeans and Neandertals.

Sample	Length[1]	N[2]	MPWD[3]	σ[4]	Θ_{π} (%)[5]
Modern humans	16,537	54	60.4	26.1	0.365
African	16,550	21	76.7	24.3	0.463
Non-African	16,548	33	38.2	9.4	0.231
European	16,563	9	27.2	8	0.164
Ancient Europeans	15,897	8	12.5	5.6	0.079
Pre-LGM	15,902	4	9.6	1.7	0.060
Post-LGM	16,455	4	13.7	8.0	0.083
Neandertals	16,565	7	18.6	17.2	0.112
Upper Paleolithic Neandertals	16,565	6	8.1	3.5	0.049

[1] Number of aligned positions excluding alignment gaps
[2] Number of sequences
[3] Mean number of pairwise differences
[4] Standard deviation of MPWD
[5] Average percentage of pairwise difference per site

Table S4 (related to SI 4. Likelihood ratio test). Results from a test of molecular clocks using the Maximum Likelihood method.

		ln*L*	**Parameters**	**(+*G*))**	**(+*I*)**
54 modern humans	**with clock**	-29193.503	62	0.050	0.00
	without clock	-29168.551	113	0.05	0.00
54 modern humans, Neandertals, Denisova, Chimpanzee	**with clock**	-39373.743	73	0.804	0.65
	without clock	-39337.748	135	0.76	0.64

Table S5 (related to Table 2). Estimated substitution rate using 10 ancient modern humans (AMH) only and dataset including 10AMH+4 radiocarbon dated Neandertals and 54 present-day global mtDNAs assuming a constant population size and a relaxed molecular clock.

Region	Ancient samples	mean	95% HPD lower	95% HPD upper
Coding region	10 AMH	9.98E-09	5.41E-09	1.47E-08
Coding region	10 AMH+4Nea	1.12E-08	6.65E-09	1.57E-08
Whole mtDNA	10 AMH	1.76E-08	1.23E-08	2.29E-08
Whole mtDNA	10 AMH+4Nea	1.86E-08	1.35E-08	2.37E-08

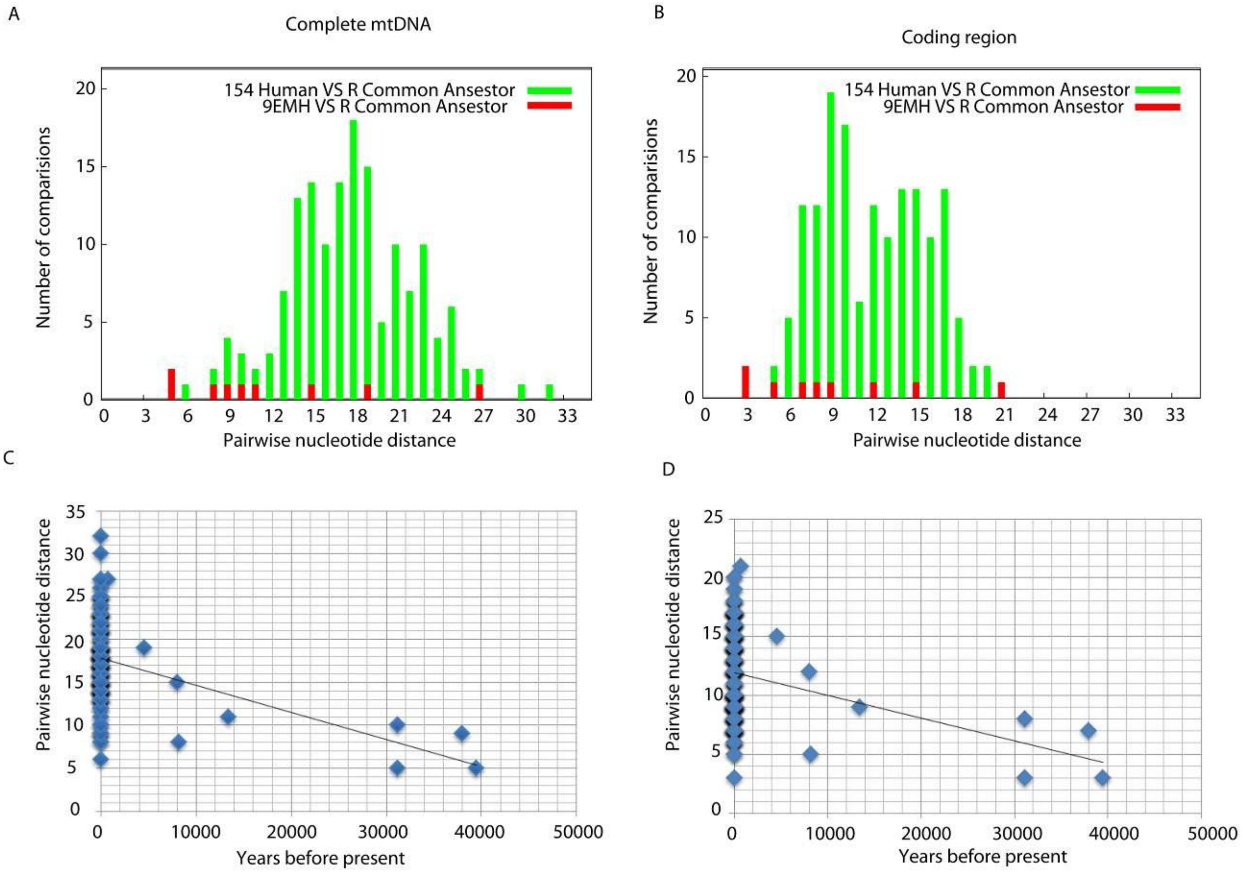

Figure S1 (related to Substitution rate estimates in main text).Nucleotide distance to the root of hg R for 154 modern humans and 9 ancient modern humans that fall into hg R, for the complete mtDNA (A) and coding region (B). Plot of nucleotide distance against the age of the sequence gives the substitution rate as slope of the linear regression: 1.92±0.76 x10-8 substitutions bp-1 yr-1 for the complete genome (C), 1.25±0.68 10-8 substitutions bp-1 yr-1 for the coding region (D).

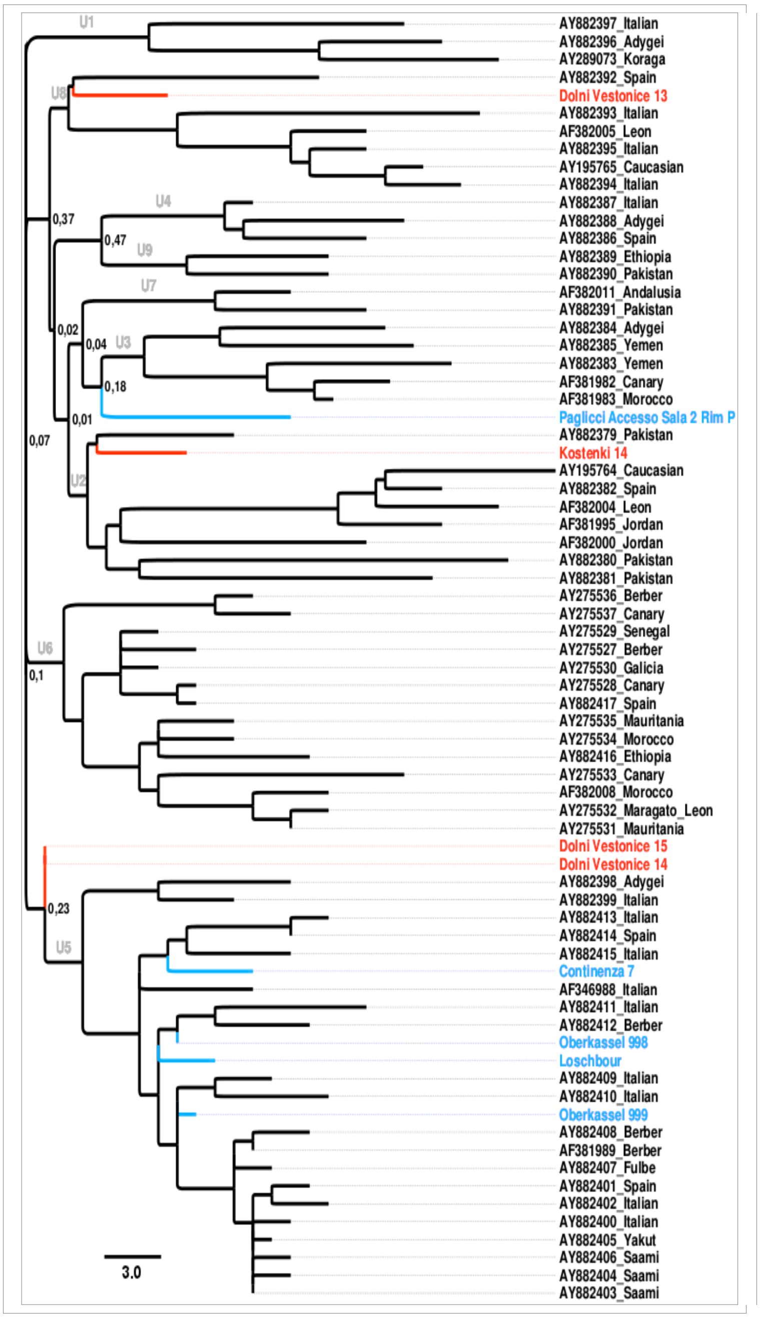

Figure S2 (related to Evolutionary analysis in main text). Relationship of early modern European mtDNAs to 63 modern mtDNAs of haplogroup U from a worldwide population. Maximum Parsimony tree, pre-LGM branches are indicated in red, post-LGM in blue. Scale shows the branch length representing 3 substitutions.

Supplemental References

1. Svoboda, J., van der Plicht, J., and Kuz Elka, V. (2002). Upper Palaeolithic and Mesolithic human fossils from Moravia and Bohemia (Czech Republic): some new 14C dates. Antiquity *76*, 957–962.
2. Henry-Gambier, D. (2002). Les fossiles de Cro-Magnon (Les Eyzies-de-Tayac, Dordogne): Nouvelles donnees sur leur Position chronologique et leur attribution culturelle. Bull. et Mém. de la Société d'Anthropologie de Paris *14,* 89-112.
3. Delsate, D., Guinet, J.M., and Saverwyns, S. (2009). De l'ocre sur le crâne mésolithique (haplogroupe U5a) de Reuland-Loschbour (Grand-Duché de Luxembourg)? Bull. Soc.

Préhist. Luxembourgeoise *31*, 7-30.

4. Shang, H., Tong, H., Zhang, S., Chen, F., and Trinkaus, E. (2007). An early modern human from Tianyuan Cave, Zhoukoudian, China. Proceedings of the National Academy of Sciences of the United States of America *104*, 6573-6578.
5. Marom, A., McCullagh, J.S., Higham, T.F., Sinitsyn, A.A., and Hedges, R.E. (2012). Single amino acid radiocarbon dating of Upper Paleolithic modern humans. Proceedings of the National Academy of Sciences of the United States of America *109*, 6878-6881.
6. Hedges, R.E.M., Housley, R.A., Bronk, C.R., and van Klinken, G.J. (1992). Radiocarbon dates from the Oxford AMS system: Archaeometry datelist 15. Archaeometry *34/2*, 337-357.
7. Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature *463*, 757-762.
8. Rohland, N., and Hofreiter, M. (2007). Ancient DNA extraction from bones and teeth. Nat Protoc *2*, 1756-1762.
9. Meyer, M., and Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc *in press*.
10. Maricic, T., Whitten, M., and Paabo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PloS one *5*, e14004.
11. Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome biology *10*, R83.
12. Green, R.E., Malaspinas, A.S., Krause, J., Briggs, A.W., Johnson, P.L., Uhler, C., Meyer, M., Good, J.M., Maricic, T., Stenzel, U., et al. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell *134*, 416-426.
13. Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science *325*, 318-321.
14. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710-722.
15. Lalueza-Fox, C., Sampietro, M.L., Caramelli, D., Puder, Y., Lari, M., Calafell, F., Martinez-Maza, C., Bastir, M., Fortea, J., de la Rasilla, M., et al. (2005). Neandertal evolutionary genetics: mitochondrial DNA data from the iberian peninsula. Molecular biology and evolution *22*, 1077-1081.
16. Schmitz, R.W., Serre, D., Bonani, G., Feine, S., Hillgruber, F., Krainitzki, H., Paabo, S., and Smith, F.H. (2002). The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. Proceedings of the National Academy of Sciences of the United States of America *99*, 13342-13347.