Current Biology, Volume *22*

Supplemental Information

Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers

Federico Sánchez-Quinto, Hannes Schroeder, Oscar Ramirez, María C. Ávila-Arcos, Marc Pybus, Iñigo Olalde, Amhed M.V. Velazquez, María Encina Prada Marcos, Julio Manuel Vidal Encinas, Jaume Bertranpetit, Ludovic Orlando, M. Thomas P. Gilbert, and Carles Lalueza-Fox

Supplemental Inventory

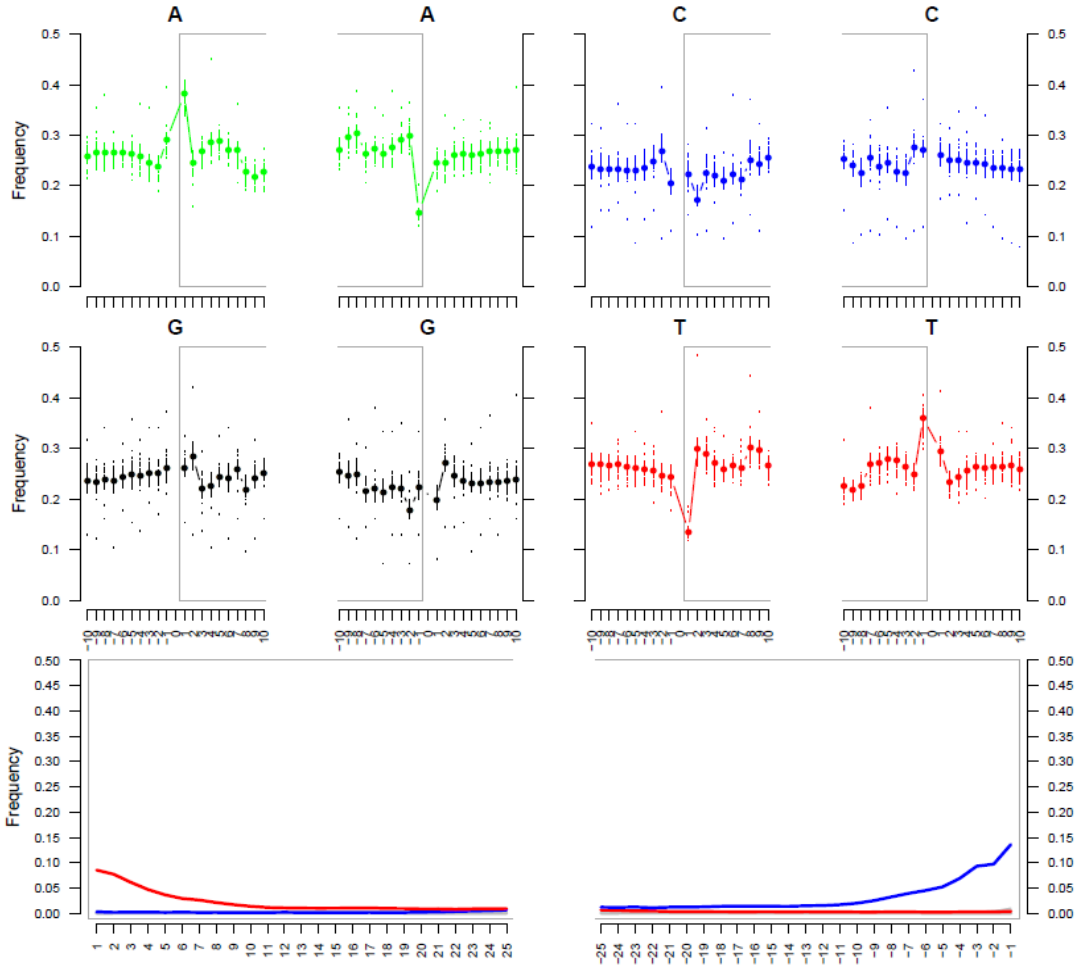

Figure S1. Sequence Features of La Braña 1 mtDNA Reads, Related to Figure 2

Nucleotide base frequencies (above) and misincorporation patterns (below) of La Braña 1 mtDNA reads. The base composition is plotted as a function of distance from the 5'-ends (right) and 3'-ends (left). Previously described increase of C→T's at the 5' breaking point (lower right) as well as increase of G→A's at the 3' breaking point (upper left) can be observed.

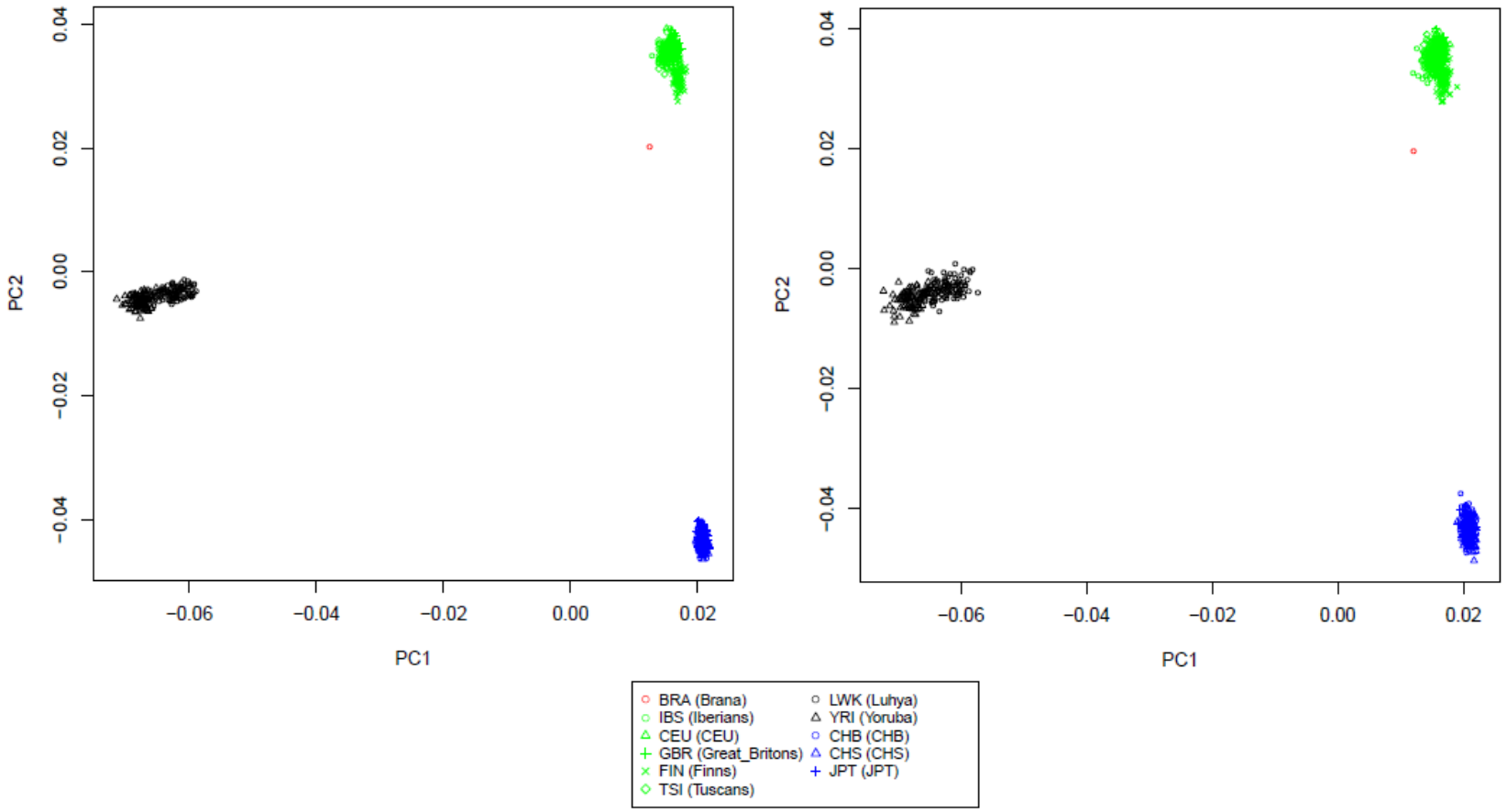

Figure S2. PCA with the Shotgun Data from La Braña 1 (Left) and La Braña 2 (Right) and the Worldwide Data Set from 1,000 Genomes Project, Related to Figure 3

The Mesolithic individual is placed near but outside the current European gene pool.

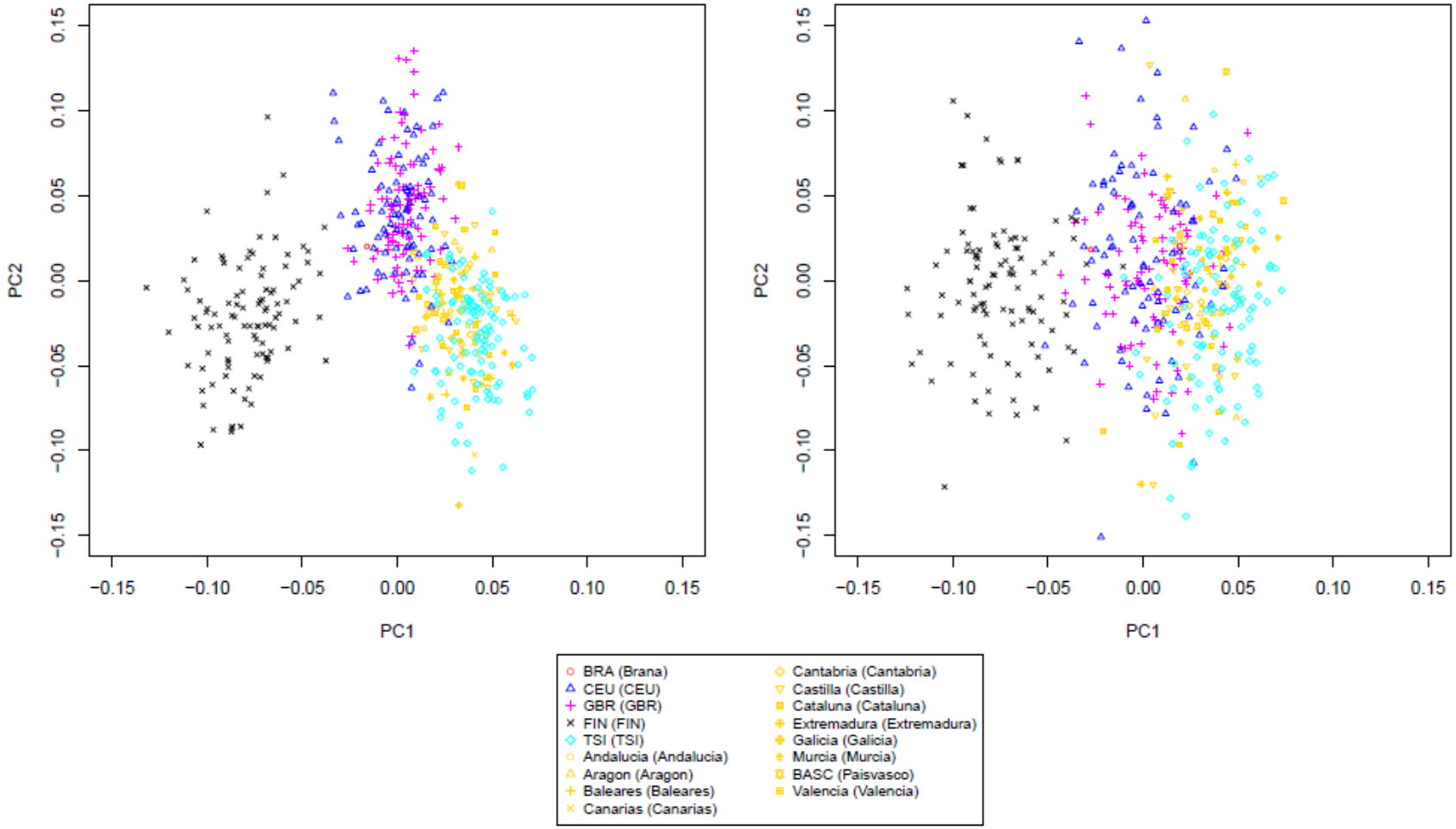

Figure S3. PCA with 1KGPomni Chip Data (European Populations) and La Braña 1 (left) and La Braña 2 (right), Related to Figure 3

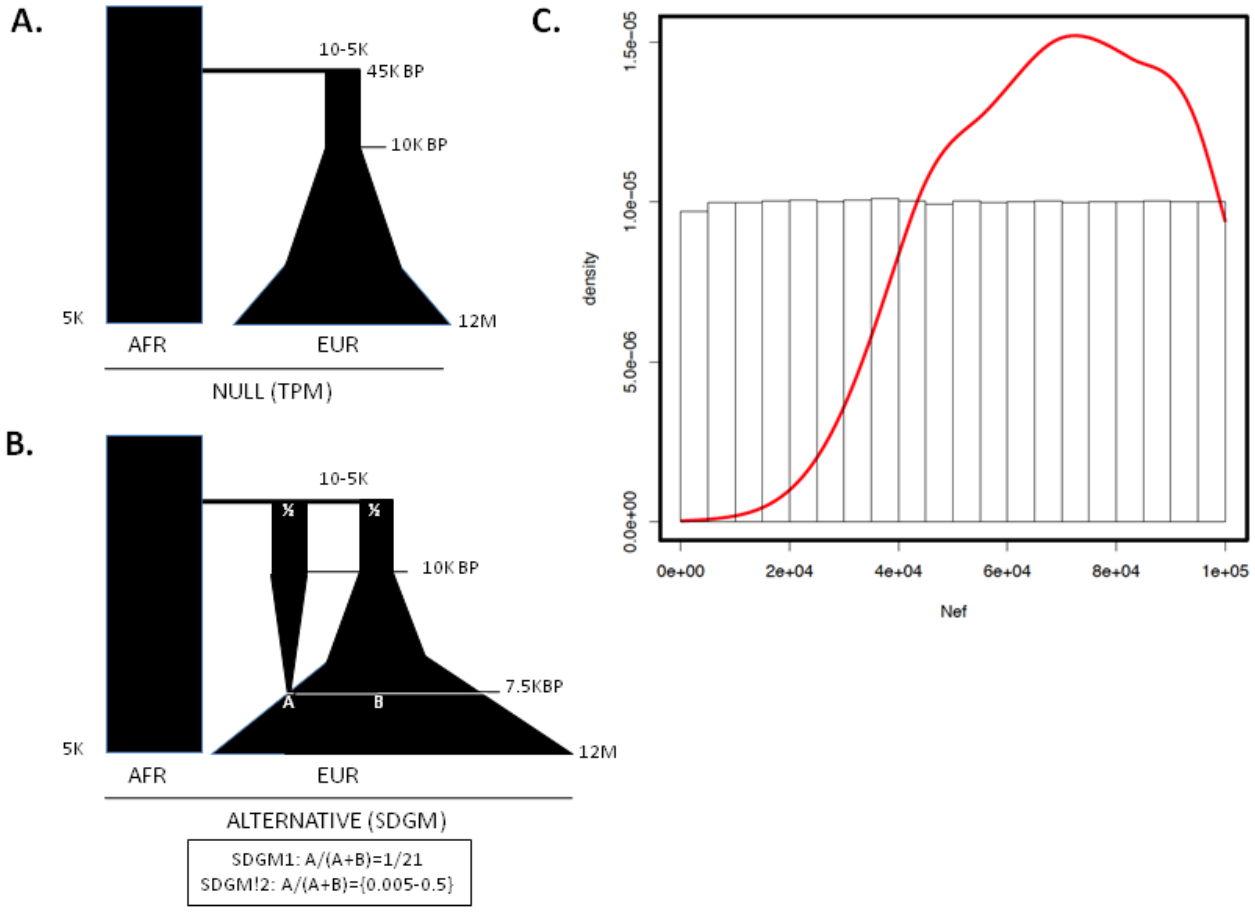

Figure S4. Serial Coalescent Simulations and Approximate Bayesian Computation, Related to Figure 4

(A**)** Total Panmixia Model, TPM. The origin of the European population was sampled at random from a population evolving under a stable demography with 5,000 individuals. The total number of founders was modeled using a prior flat distribution spanning 10 to 5,000 individuals.

(B**)** Split with Differential Growth Models, SDGM. Panel C**.** Distributions of the population size at 7,500 BP. The sum of the population sizes of both demes (A+B) is reported. Posterior distribution = red curve. Prior flat distribution = histograms.

Table S1. Sequence Clones Generated by PCR from La Braña 1 and 2 Skeletons (B: Barcelona, C: Copenhagen) , Related to Figure 2

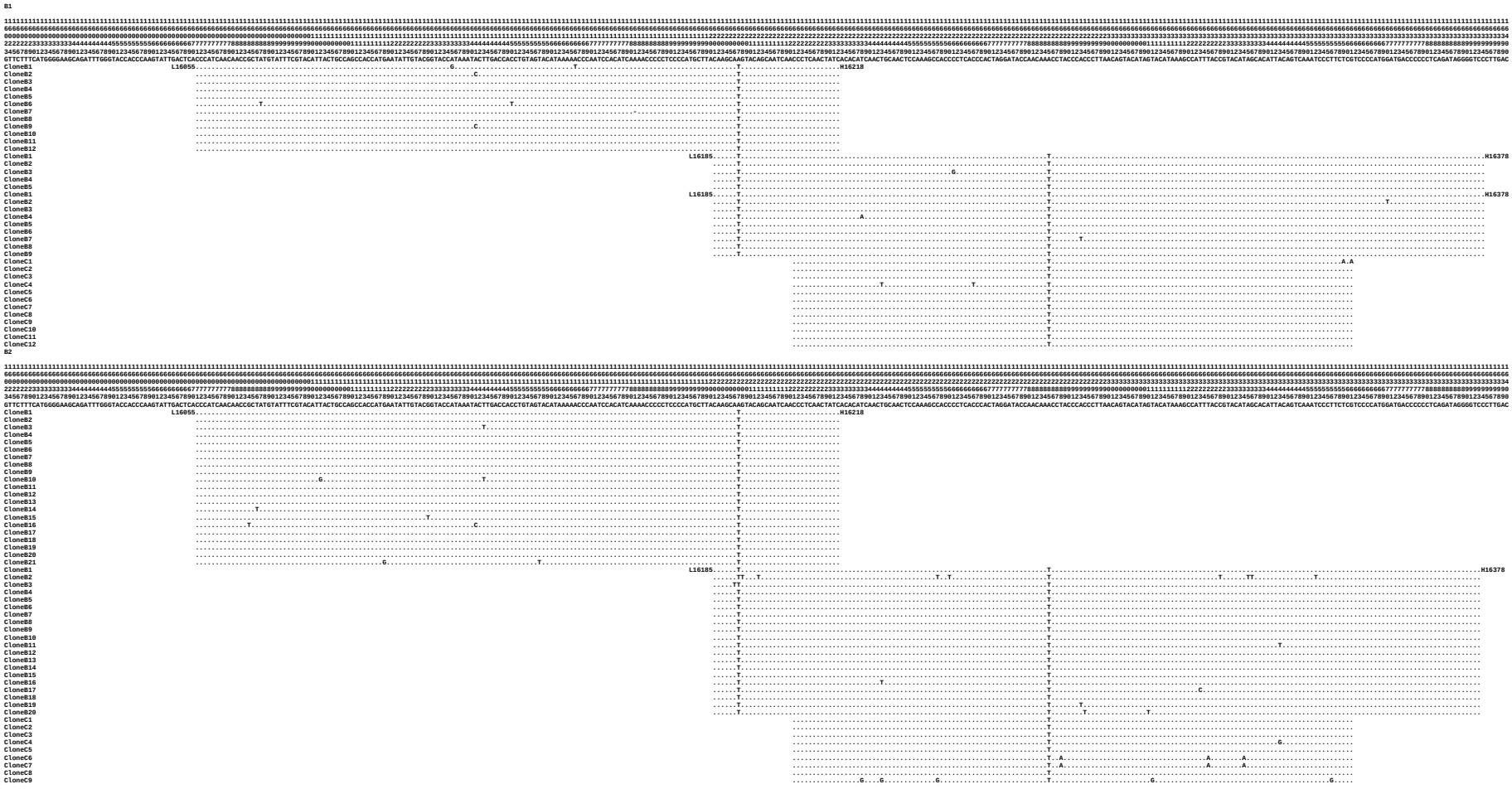

Table S2. Coding Region Mutations and Control Region Mutations Defining U5b2c1 Haplotype in PhyloTree, All of Them Present in the La Braña 1 Consensus, Related to Figure 2

Nucleotide position	rCRS	La Braña consensus	number of reads	Haplogroup	Consensus/rCRS	Contamination estimate
12372	G	A	18	U	13/18	0.28
3197	T	C	12	U5	12/12	0
9477	G	A	34	U5	30/34	0.12
13617	T	C	14	U5	14/14	0
16192	C	T	14	U5	14/14	0
16270	C	T	12	U5	8/12	0.34
150	C	T	14	U5b	14/14	0
7768	A	G	30	U5b	30/30	0
14182	T	C	24	U5b	24/24	0
1721	C	T	35	U5b2	30/35	0.15
13637	A	G	14	U5b2	14/14	0
723	A	G	25	U5b2c	25/25	0
960.1C	-	C	15	U5b2c	15/15	0
13017	A	G	24	U5b2c	24/24	0
6920	C	A	21	U5b2c1	17/21	0.19
13434	A	G	28	U5b2c1	26/28	0.07

Heterogeneities in these positions have been used to obtain a contamination estimate in the La Braña 1 captured mtDNA genome.

Table S3. Targeted Resequencing Results for the mtDNA from the Mesolithic Specimen of La Braña 1, Related to Figure 2

raw reads[1]	filtered reads[2]	mapped reads[3]	mt (%)[4]	collapsed reads[5]	read length[6]	bases >1x[7]	coverage[8]	haplogroup[9]
44,581,347	37,486,041	19,993,417	53%	5,488	85.7	16,450	28	U5b2c1

[1] number of raw sequences generated after capture
[2] number of filtered sequences
[3] number of uniquely mapped mtDNA reads
[4] % of mtDNA reads
[5] number of unique mtDNA reads after collapsing
[6] average read length
[7] number of bases with more than 1x coverage
[8] average coverage of mtDNA genome
[9] haplogroup assignment based on PhyloTree

Table S4. Shotgun Sequencing Results for the Mesolithic Specimen of La Braña 1 and La Braña 2, Related to Figure 3

Sample	raw reads[1]	filtered reads[2]	mapped reads[3]	nu (%)[4]	read length[5]	overlap with 1k SNP data[6]	filtered SNPs[7]
La Braña 1	42,396,337	6,113,535	728,880	1.7	74.7	116,220	47,742
La Braña 2	15,670,532	1,036,759	364,578	2.3	59.5	48,513	20,750

[1] number of raw sequences
[2] number of filtered sequences
[3] number of reads uniquely mapping to hg18
[4] % of nuclear reads
[5] average read length
[6] number of sites overlapping with 1k SNP data
[7] number of SNPs after filtering (Q 40, MAF < 0.5%, damage and triallelic sites)

Supplemental Experimental Procedures

La Braña Site

The main current entrance of the site is located at 1,489 meters of altitude, in the Cantabrian mountain Range; the cave is close to a natural crossing (pass of Vegarada, with an altitude of 1,567 meters) that connects the Cantabrian coast with the river Duero basin. Due to the present difficulties in reaching the remains, it is clear that there was originally another entrance, nowadays obstructed. The area has a strong continental climate, with mild summers and very cold winters (average minimum temperature during winter months is -4.3ºC and average temperature during all year is 8.1ºC). Besides, the remains were placed at about 30 meters from the entrance, thus ensuring a very stable environmental temperature all along the year. All these factors may help explain the notable DNA preservation of La Braña skeletons.

Mesolithic Samples

In 2006, two complete human skeletons were accidentally found in a karst system at La Braña-Arintero (León, Spain). The two individuals (La Braña 1 and 2), were found a few meters apart, in a crouched position. With the exception of 24 perforated atrophic red deer canines in La Braña 2 used as personal ornaments embroidered on a cloth, no other artifacts were discovered on or around the bodies. Both individuals were sexed as males based on the examination of their pelvic bones [1]. Both skeletons were dated; La Braña 1 yielded a C14 date of 6,980±50 years BP (Beta 226472) and La Braña 2 a C14 date of 7,030±50 years BP (Beta 226473). Considering the radiocarbon error margins, it is plausible that both individuals were contemporaneous.

DNA Extraction and HVR1 Sequencing

The La Braña samples (two tooth root tips) were extracted in a dedicated ancient DNA laboratory at the Institute of Evolutionary Biology in Barcelona, following established protocols to control for contamination. The extraction protocol, published elsewhere [2], is based on a standard proteinase K digestion, followed by a phenol-chlorophorm extraction and microcolumn concentration. Following extraction, we amplified and sequenced the complete HVR1 (hypervariable region 1) of the mtDNA from the two La Braña specimens using a two-step PCR protocol. The PCR products were cloned with a TOPO-TA cloning kit (Invitrogen) and sequenced in a Applied BioSystems 3100 DNA sequencer, following a methodology previously described [2]. L16055-H16218 and L16185-H16378 primer couples were used. A fragment amplified with the L16205-H16347 primers was independently replicated in both samples at the Center for Geogenetics in Copenhagen.

Library Preparation

The library was prepared by using a blunt-end library preparation kit (E6070) from NEB, skipping the fragmentation step. End-repair and adapter ligation was performed as per manufacturer's instructions using 30 µL of DNA extract and an adapter mix described in [3]. Subsequently, an adapter fill-in reaction was performed without prior size selection using 30 µL of adapter-ligated DNA, 13 µL of sterile H_2O, 5 µL of adapter fill-in reaction buffer, and 2 µl *Bst* DNA Polymerase. The mix was incubated for 30 mins at 37°C followed by 20 mins at 80°C to inactivate the enzyme. The entire library was then amplified for 15 cycles using the HiFi enzyme (0.1 U/µl) (Lifetech), 1X PCR Buffer, 2 mM $MgSO_4$, 0.8 µM BSA, 0.2 µM dNTP, and 0.2 µM indexing primers in a 100 µl PCR reaction and the following cycling conditions: 5 mins at 94°C, followed by 10-20 cycles of 30 sec at 94°C, 30 sec at 60°C and 30 sec at 68°C, and a final extension of 7 min at 72°C and 10.0°C for ever. The amplified library was then purified using a Qiagen MinElute spin column and eluted in 30 µl EB.

MtDNA Capture and Sequencing

To enrich for target mtDNA, we used long-range PCR products as bait for molecular capture through hybridization [4]. This hybridization capture technique has several advantages over traditional PCR-based approaches. Firstly, it enables the retrieval of short fragments of aDNA that are below the minimal amplification length of PCR. Arguably, this also enables us to retrieve a larger amount of authentic aDNA sequences as opposed to the larger contaminating molecules. Secondly, in combination with high-throughput sequencing this new enrichment technique enables us to sequence large amounts of target molecules.

Two 9kb fragments covering the entire human mitochondrial genome were produced using the methods described in ref. [5]. The PCRs were performed using 1X PCR Buffer, 1.4 mM $MgSO_4$, 0.2 μM dNTP, and 0.2 μM long-range PCR primers and 0.1 U/μL HiFi polymerase. The thermal profile consisted of an initial denaturation at 94 °C for 2 min followed by 25 cycles consisting of a denaturation for 20 s at 94 °C, a 20 s annealing at 60 °C, and a 10 min elongation at 68 °C and a final extension at 72 °C. PCR products were purified over Qiagen Qiaquick spin columns and quantified by Qubit. The two mitochondrial PCR products were then pooled in equimolar amounts to a total amount of 3 μg and sonicated using the Bioruptor (Diagenode), producing fragments of 150–700 bp as observed on a 2% agarose gel. The sonicated products were then biotinylated by ligation to a biotin-carrying adapter and immobilized on streptavidin-coated magnetic beads. Subsequently, the amplified libraries were single-stranded by incubating them at 95°C for 3 min and added to the streptavidin beads coated with the fragmented long-range PCR products. The mixture was then incubated under rotation at 65 °C in a hybridization oven (Model 777; SciGene). After 48 h, the beads were washed, and library molecules were eluted following incubation at 95°C for 3 min.

The captured target library was then amplified for 15 cycles using the same PCR conditions as during the initial library amplification. The mtDNA library was then pooled with other capture products and sequenced together in one fourth of an Illumina HiSeq2000 lane (single read). The result is a well covered (> 20x) mtDNA genome, which allowed to screen for damage patterns characteristic of aDNA sequences. This, in particular, is one of the key improvements of the newly developed technology as it enables us identify ancient damaged sequences, thus reducing the risk of mistaking modern contaminants for authentic ancient DNA.

Sequence Analysis and SNP Filtering

Raw reads were first filtered with an in-house script to remove low quality and adapter sequences as well as low quality stretches at the 3’ ends. Filtered reads were then mapped to the human reference genome (hg18) using BWA version 0.5.9 [6] requiring a mapping quality of >25. Clonal reads were removed using the rmdup program of the Samtools (version 0.1.18 [7]) suite. Ambiguously mapped reads were also filtered out using Samtools and controlling for XT and XA tags.

Sites were overlapped with the low coverage SNPs Phase1 dataset from the 1,000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release) [8].

To exclude potentially damaged sites from the analysis, the first and last five nucleotides in each read were “masked” with the lowest quality and only bases with base quality of >40 were consider for downstream analysis. Due to limited coverage, most positions were covered with a single read; in positions covered by >1 reads a random read was sampled. To compensate for the fact that no heterozygous/homozygous state could be ascertained in the ancient data due to the low coverage, one allele was randomly sampled from all SNPs in the modern reference individuals. Subsequently, the information in both ancient and modern datasets was duplicated to generate homozygote individuals. All triallelic sites (N=35 for La Braña 1 and N=11 for La

Braña 2), likely to be sequencing or alignment errors, were excluded from the SNP dataset. All SNPs with C/T and G/A changes in La Braña 1 sequences were also excluded, to avoid possible artifacts derived from the typical ancient DNA base modifications. Once all SNP positions were identified, sites were pruned using PLINK software, with the LD-based SNP pruning option with default parameters (pairwise genotypic $r^2 > 0.5$ within sliding windows of 50 Kb) in order to only retain informative positions and avoid the possible linkage biases.

Several reference sets were considered for comparison with the La Braña 1 and 2 specimens; the first included only European populations while the second included African, Asian and European populations from 1,000 Genomes Project. Positions in these sets were filtered using plink (http://pngu.mgh.harvard.edu/purcell/plink/) [9] and only SNPs with a MAF >0.5% considered. For La Braña 1, 67,401 sites overlapped with SNPs from European individuals and 106,120 SNPs with the populations worldwide. After quality filters were applied, these figures were reduced to 47,742 and 73,347, respectively. For La Braña 2, the overlapped SNPs were 109,960 and 48,513, respectively. After filtering, the number of SNPs was reduced to 32,339 and 20,750, respectively.

Principal Component Analysis

We performed a principal component analysis (PCA) on the SNP data from the La Braña 1 and 2 specimens, using data from the 1,000 Genomes Project for comparison. This dataset includes 379 European individuals and 850 individuals worldwide from contemporary populations. In order to control for possible bias in the subset of SNPs mapped from the sequence data, a Site Frequency Spectrum plot for both datasets (whole and thinned) was generated. No major differences were found at any population in the thinned dataset pointing that randomly distributed SNPs were selected from the Mesolithic individuals.

From the 1KGPomni chip data, 476 Europeans were included for La Braña 1 analysis. 16,918 SNPs were found to overlap with 1KGPomni European populations, and after filtering, only 7,305 SNPs remained. For La Braña 2, 7,279 SNPs overlapped with European populations, of which 3,475 were kept for the PCA analysis

Approximate Bayesian Computation

Mitochondrial HVR-1 sequence datasets were simulated under three different population models (Figure S4) using a serial coalescent framework as implemented in the software BayeSSC available for download at http://www.stanford.edu/group/hadlylab/ssc/index.html [10]. In the first model, we assumed a genetic continuity in European population from the Mesolithic to the Neolithic culture; this model was referred as a Total Panmixia Model, TPM, in a previous study [11]. The two other population models were referred as Split with Differential Growth Models (SDGM) in [11]. These models (SDGM1 and SDGM2) were designed to account for a structure in Paleolithic and Mesolithic European populations. In both models one deme will mainly contribute to the Neolithic population and it is assumed it will grow exponentially at 10 ky, while the other deme will decline.

For the Total Panmixia Model, TPM [9], we sampled 10 to 5,000 individuals as the source of the European Upper Paleolithic population in Europe from a hypothetical African population of a constant size of 5,000 individuals. The European population was assumed to grow exponentially from 10,000 BP onwards and to reach a current effective size of 12 millions of individuals. These demographic assumptions are in agreement with the simulations presented in [12]. The mutation rate was assumed at $7.5\ 10^{-6}$ mutations per site per generation, the transition bias was of 0.9841 and the generation time was 25 years. We also used a uniform gamma distribution of mutations with a shape parameter of 0.205. Our sampling scheme was designed in order to match the sequence dataset presented in [13] which represents the most

complete HVR-1 sequence dataset reported so far, except that 26 sequences from Les Treilles were considered (instead of 29) in order to remove related individuals [14]. Also, one individual from Donkalnis and one individual from Dudka were removed due to uncertainty in age estimates [12]. In addition, the sequence dataset reported in [11] was added. In total, the sampling scheme consisted of 2 individuals from the Upper Paleolithic, 43 from the Mesolithic (including the two La Braña specimens) and 121 from the Neolithic.

The Differential Growth Models (SDGM) models [11] account for a structure in Paleolithic and Mesolithic European populations. In both models, one deme (B) will mainly contribute to the Neolithic population. The latter deme (B) was assumed to start growing exponentially at 10,000 BP while the other (deme A) was declining. Hence, the demographic contribution of deme A will be lower than for deme B at the time the two demes merged around 7,500 BP and gave rise to the Neolithic population. Two versions of SDGM were ran: one (SDGM1) assuming that the demographic contribution of deme B was 20X larger than the contribution of deme A, following the model that received the highest support in [11]; in the second SDGM (SDGM2), different possibilities were explored using a uniform prior ranging 0.005 to 0.5 for the ratio A/(A+B).

For each model, one million of simulations were performed. We then performed Approximate Bayesian Computation (ABC) analyses using nucleotide diversity, haplotype diversity, tajimas D and Fst as a vector of summary statistics using the R makepd4() function and a tolerance region of 0.001 of all simulations (1,000 simulations). This function performs ABC following the rejection, local linear regression and smooth weighting procedure described in [15]. Posterior distributions for deme sizes as recovered from the ABC procedure were found different than prior uninformative distributions (data not shown). The posterior probability of the three models tested was estimated using categorical regression and the R calmod function following [16]. This procedure takes advantage of the weighted regression framework and treats a model indicator as a categorical variable that can take values of 1 (null model, TPM), 2 (SDGM1) and 3 (SDGM2). The densities of the categorical variable, as estimated from the simulations captured within the tolerance region, provide a direct estimate of the posterior probability of the model. R functions are available online at http://www.rubic.rdg.ac.uk/~mab/stuff/.

In agreement with previous findings, we found that model SDGM1 was best supported with a posterior probability of 0.619 (see below). Interestingly, the total panmixia model received an extremely low support, showing a posterior probability 1,655-2,691-fold lower than SDGM models. This strongly argues against genetic continuity between Mesolithic and Neolithic cultures in Europe.

Model	**Post. Probability**
TPM	0.00023
SDGM1	0.61905
SDGM2	0.38072

Supplemental References

1. Prada Marcos, M.E. (2010). Estudio antropológico de los hombres mesolíticos de La Braña-Arintero. In Los hombres mesolíticos de la cueva de La Braña-Arintero (Valdelugueros, León) J.M. Vidal Encinas and M.E. Prada Marcos, Eds. (León: Junta de Castilla y León), pp. 92-118.

2. Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C., Catalano, G., Hughes, D., Rohland, N., Pilli, E., Longo, L., Condemi, S., et al. (2007). A melanocortin 1 receptor suggests variying pigmentation among Neanderthals. Science 318, 1453-1455.

3. Meyer, M., and Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. doi:10.1101/pdb.prot5448.

4. Maricic, T., Whitten, M., and Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One. 5, e14004.

5. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., et al. (2011). An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94-98.

6. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

7. Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760.

8. Altshuler, D., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De la Vega, F.M., Donnelly, P., Egholm, M., et al. (The 1,000 Genomes Project Consortium). (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

9. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575.

10. Anderson, C.N., Ramakrishnan, U., Chan, Y.L., and Hadly, E.A. (2005). Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics, 21, 1733-1734.

11. Gamba, C., Fernández, E., Tirado, M., Deguilloux, M.F., Pemonge, M.H., Utrilla, P., Edo, M., Molist, M., Rasteiro, R., Chikhi, L., et al. (2012). Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. Mol. Ecol. 21, 45-56.

12. Bramanti, B., Thomas, M.G., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M.N., Jankauskas, R., Kind, C.J., et al. (2009). Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326, 137–140.

13. Deguilloux, M.F., Leahy, R., Pemonge, M.H., and Rottier, S. (2012) European neolithization and ancient DNA: an assessment. Evol. Anthropol. 21, 24-37.

14. Lacan, M., Keyser, C., Ricaut, F.X., Brucato, N., Duranthon, F., Guilaine, J., Crubézy, E., and Ludes, B. (2011). Ancient DNA revealed male diffusion through the Neolithic Mediterranean route. Proc. Natl. Acad. Sci. USA 108, 9788-9791.

15. Beaumont, M.A., Zhang, W., and Balding, D.J. (2002). Approximate Bayesian Computation in Population Genetics. Genetics 162, 2025-2035.

16. Beaumont, M. (2008). Joint determination of topology, divergence time and immigration in population trees. In Simulations, Genetics and Human Prehistory, S. Matsumura, P. Forster and C. Renfrew, eds. (Cambridge: McDonald Institute for Archaeological Research), pp. 134-154.