Supplementary Information

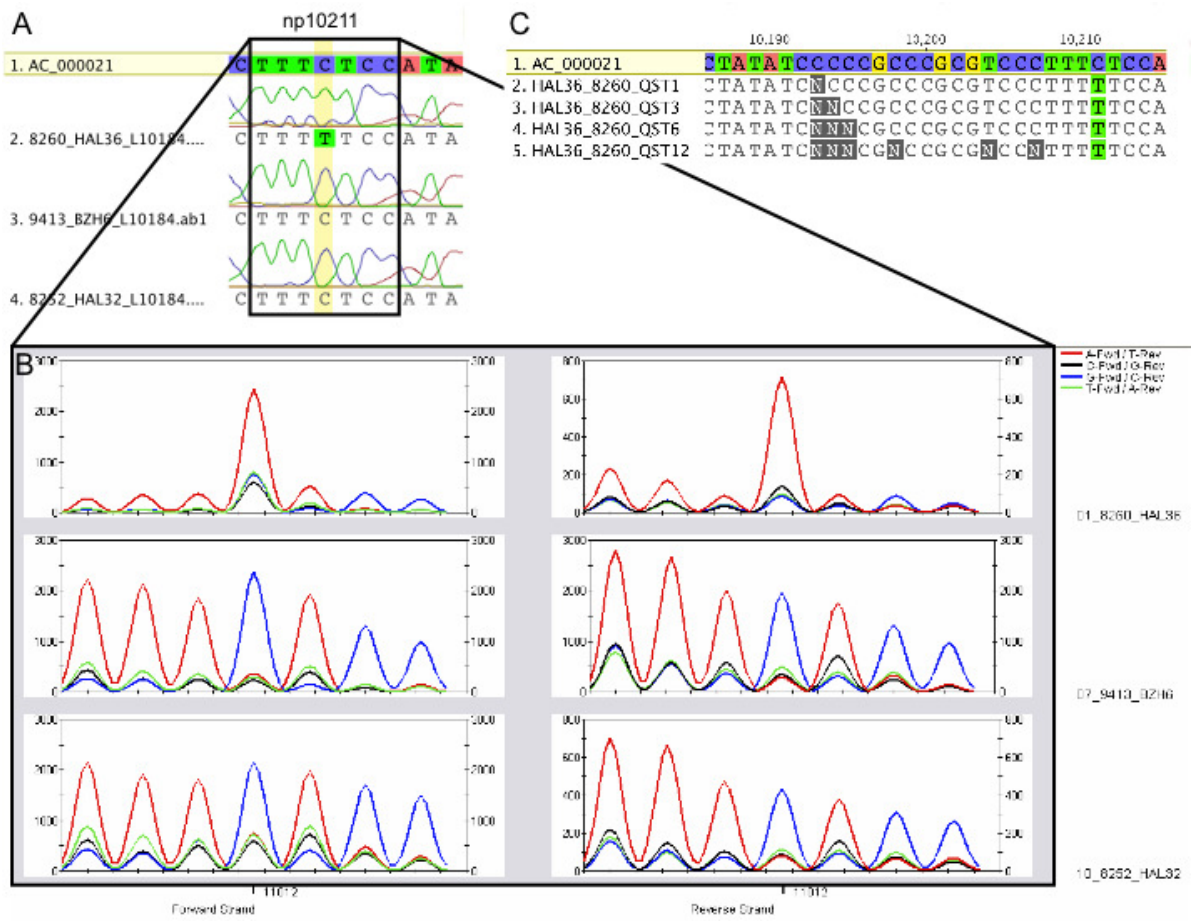
Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans

Paul Brotherton, Wolfgang Haak, Jennifer Templeton, Guido Brandt, Julien Soubrier, Christina J. Adler, Stephen M. Richards, Clio Der Sarkissian, Robert Ganslmeier, Susanne Friederich, Veit Dresely, Mannis van Oven, Rosalie Kenyon, Mark B. Van der Hoek, Jonas Korlach, Khai Luong, Simon Y. W. Ho, Lluis Quintana-Murci, Doron M. Behar, Harald Meller, Kurt W. Alt, Alan Cooper, & The Genographic Consortium

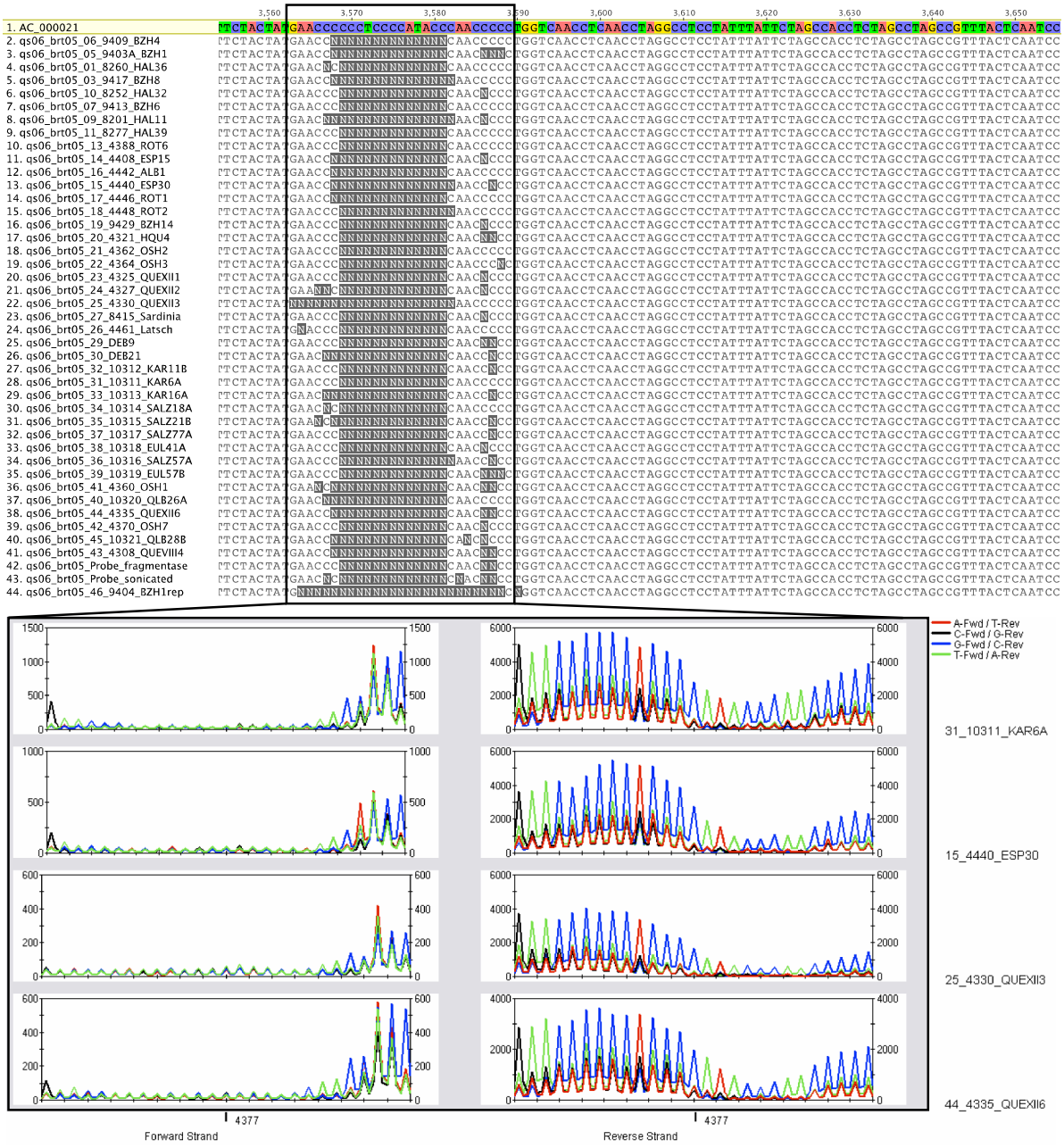
Members of The Genographic Consortium

Syama Adhikarla, ArunKumar GaneshPrasad, Ramasamy Pitchappan & Arun Varatharajan Santhakumari, Madurai Kamaraj University, Madurai, Tamil Nadu, India; Elena Balanovska & Oleg Balanovsky, Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia; Jaume Bertranpetit, David Comas, Begoña Martínez-Cruz & Marta Melé, Universitat Pompeu Fabra, Barcelona, Spain; Andrew C. Clarke & Elizabeth A. Matisoo-Smith, University of Otago, Dunedin, New Zealand; Matthew C. Dulik, Jill B. Gaieski, Amanda C. Owings, Theodore G. Schurr & Miguel G. Vilar, University of Pennsylvania, Philadelphia, Pennsylvania, United States; Angela Hobbs & Himla Soodyall, National Health Laboratory Service, Johannesburg, South Africa; Asif Javed, Laxmi Parida, Daniel E. Platt & Ajay K. Royyuru, IBM, Yorktown Heights, New York, United States; Li Jin & Shilin Li, Fudan University, Shanghai, China; Matthew E. Kaplan & Nirav C. Merchant, University of Arizona, Tucson, Arizona, United States; R. John Mitchell, La Trobe University, Melbourne, Victoria, Australia; Lluis Quintana-Murci, Institut Pasteur, Paris, France; Colin Renfrew, University of Cambridge, Cambridge, United Kingdom; Daniela R. Lacerda & Fabrício R. Santos, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; David F. Soria Hernanz & R. Spencer Wells, National Geographic Society, Washington, District of Columbia, United States; Pandikumar Swamikrishnan, IBM, Somers, New York, United States; Chris Tyler-Smith, The Wellcome Trust Sanger Institute, Hinxton, United Kingdom; Pedro Paulo Vieira, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil; Janet S. Ziegle, Applied Biosystems, Foster City, California, United States.

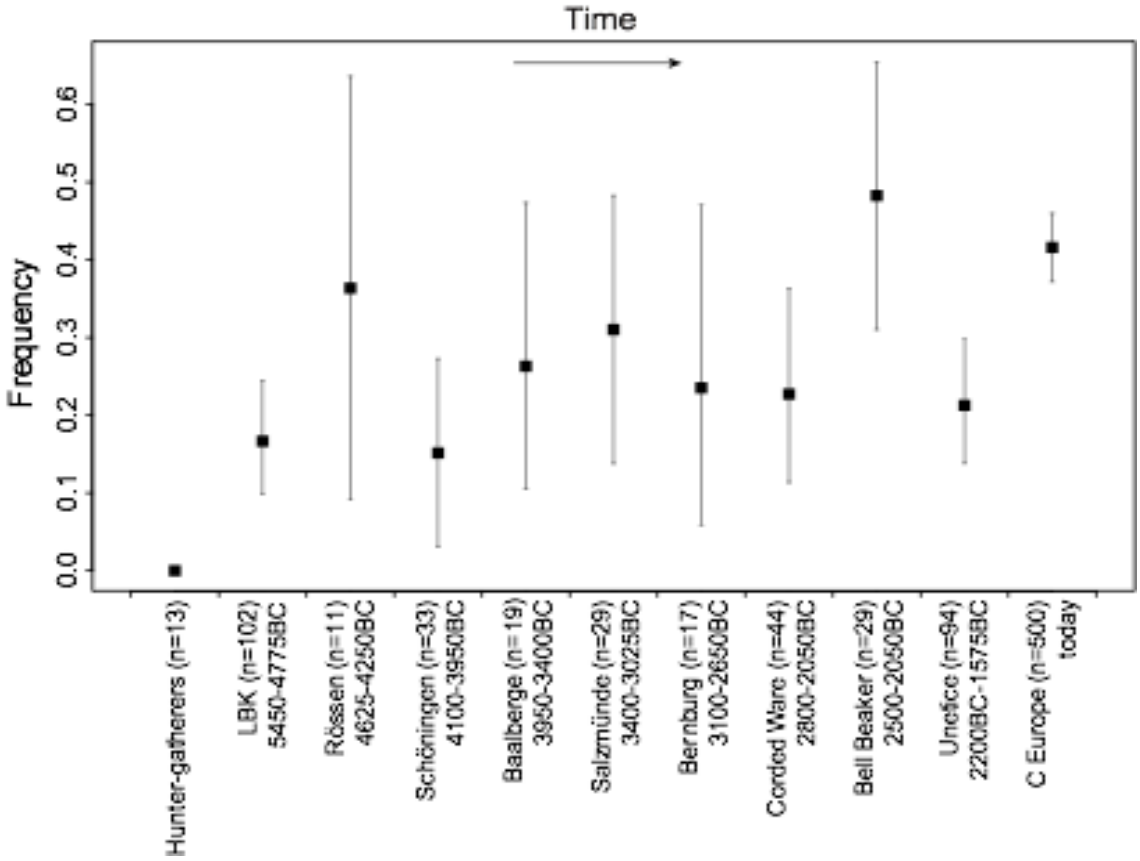
Supplementary Figures



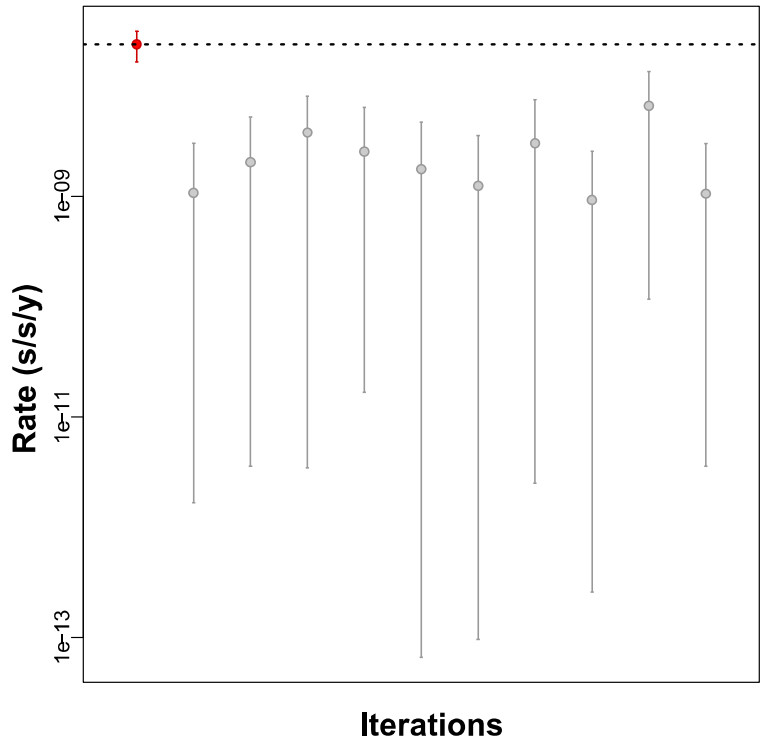
Supplementary Figure S1. Examples of data from different sequencing platforms. Zoom in on a diagnostic SNP at np 10211 in individual HAL36 (hg H23). Classical 'Sanger' or dye-terminator sequencing (a), trace view of probe intensity as consecutive bases from the Affymetrix v2.0. CEL files (b) and alignment of four different QSTs (1, 3, 6 and 12) against the rCRS (c).



Supplementary Figure S2. Reclaiming cytosine-stretches for final mt genome sequence calls. Alignment of QST 6 output for all individuals illustrates N calls for C-stretches, whereas in all cases the tiling array for the reverse strand reveals readable data (zoom in for four selected samples).



Supplementary Figure S3. Mitochondrial haplogroup H frequencies. Line bars indicate 95% confidence intervals based on 10,000 bootstrapped replicates. Samples from Central European archaeological cultures and 500 randomly drawn, present-day individuals are presented on the x-axis in chronological order from left to right (LBK=Linear Pottery Culture, C Europe=Present-day Central Europe). Numbers in brackets indicate the total number of successfully typed individuals per culture[61].



Supplementary Figure S4. Date randomisation test for the dated phylogeny. Estimated rates were replicated ten times, each with randomised dates (in grey). These are significantly different from the rate estimated from the main analysis (in red), confirming the presence of sufficient signal from the ancient tip dates to calibrate the tree.

Supplementary Table S1. Detailed archaeological information for each individual.

Individual	Site	Culture	Radiocarbon dates, uncalibrated	Collection details				
				Museum no.	feature	find	grave	area
HAL36	Halberstadt-Sonntagsfeld	Linear Pottery culture LBK (5450-4775 BC)		2000:4338a	1114		40	05
HAL11	Halberstadt-Sonntagsfeld			2000:3988a	340		9	02
HAL32	Halberstadt-Sonntagsfeld			2000:4307a	1059		36	04
HAL39	Halberstadt-Sonntagsfeld		6145 ± 30 BP (KIA40343)	2000:4014a	413.1		11	01
DEB9	Derenburg-Meerenstieg II			1998:1175a	420		9	5896
DEB21	Derenburg-Meerenstieg II		6,151+/-27 BP (KIA30404)	1998:1288a	600		32	5897
KAR6a	Karsdorf			2006:14423a	170			2006
KAR11b	Karsdorf			2004:26267a	430			2004
KAR16a	Karsdorf			2004:26374a	611			2004
OSH2	Oberwiederstedt-Schrammhöhe	Rössen (4625-4475/4250 BC)			195			
OSH3	Oberwiederstedt-Schrammhöhe				206			
OSH1	Oberwiederstedt-Schrammhöhe				225			
OSH7	Oberwiederstedt-Schrammhöhe		5669 ± 54 (Erl-8395)		95			
SALZ18a	Salzmünde	Schöningen (4100-3950 BC)		2007:7446	3932	2		
SALZ21b	Salzmünde			2006:4805	4090	1		
ESP30	Esperstedt	Baalberge (3950-3400 BC)	5061 ± 62 (Er17784)	2004:22538	6220			A38
HQU4	Halle-Queis			2002:2328a	957			
SALZ57a	Salzmünde	Salzmünde (3400-3100/3025 BC)	4498 ± 27 (KIA-31459)	2006:5445	3833	1		
SALZ77a	Salzmünde			2006:6405	5533	2		
ESP15	Esperstedt	Corded Ware (2800-2200/2050 BC)	3904 ± 47 (Er17257)	2004:21022	6			A38
BZH6	Benzingerode-Heimburg			2003:2314	1287	1036	2	
BZH4	Benzingerode-Heimburg	Bell Beaker (2500-2200/2050 BC)		2003:2589	4607	2267	7	
ROT6	Rothenschirmbach		3953 ± 47 (Erl 8710)	2005:1685	10044			A38
ALB1	Alberstedt		3858 ± 57 (Er18537)	2005:3369	7136			A38
ROT1	Rothenschirmbach		3818 ± 48 (Erl 8715)	2005:4167	10294			A38
ROT2	Rothenschirmbach			2005:4168	10293			A38
QUEXII1	Quedlinburg XII		3792 ± 50 (Erl-7042)	2004:11864a	1266.1			12
QUEXII2	Quedlinburg XII		3655 ± 48 (Erl-7041)	2004:11857	1265			12
QLB26a	Quedlinburg		3839 ± 55 (Erl-8558)	2004:40935	19614	6a		7.1

Individual	Site	Culture	Radiocarbon dates, uncalibrated	Collection details				
				Museum no.	feature	find	grave	area
QUEXII3	Quedlinburg XII		3820 ± 42 (Erl-7038)	2004:9470a	6256			12
QLB28b	Quedlinburg			2004:40957	19617	9a		7.1
BZH1	Benzingerode-Heimburg	Unetice (2200-1575 BC)		2003:2641	6035	3479	23	
BZH8	Benzingerode-Heimburg		3626 ± 27 (KIA27958)	2003:2650	5236.1	2380	12	
BZH14	Benzingerode-Heimburg		3511 ± 29 (KIA27953)	2003:2662	3109.2	3102b	24	
EUL41a	Eulau				882	13a		4
EUL57b	Eulau				1911.5	312, XA		5
QUEVIII4	Quedlinburg VIII		3587 ± 55 (Erl-7045)	2004:10747a	3646	5a		8
Sardinia	Su Grutta ‘e is Bittuleris, Sardinia, Italy	Nuragic Bronze Age (1624 BC)	3398±26 (OxA-22193), associated remains	732				
Latsch	Latsch, South Tyrol, Italy	Iron Age (~500 BC)	2567± 45 (LTL2778A), associated remains 2412± 45 (LTL2781A), associated remains	305, 103 III Tag				

Supplementary Table S2. Mitochip v2.0 call rates (in percentages) and absolute numbers of bases not called per ancient mt genome under different quality threshold settings.

ACAD	Individual	Callrate QST1	Ns QST1	Callrate QST3	Ns QST3	Callrate QST6	Ns QST6	Callrate QST12	Ns QST12	Ns after editing QST6	Final Call rate
8260	HAL36	99.5	90	98.7	220	97.9	351	96.7	548	172	99.0
9417	BZH8	99.3	118	98.4	259	97.6	396	96.5	578	180	98.9
9403	*BZH1*	98.7	*215*	97.9	*355*	96.8	*525*	95.3	*780*	318	98.1
9409	BZH4	99.3	121	98.5	247	97.6	404	96.0	658	214	98.7
9413	BZH6	99.2	131	98.3	278	97.6	403	96.3	611	201	98.8
8201	HAL11	99.3	114	98.3	277	97.4	425	96.2	632	205	98.8
8252	HAL32	99.3	117	98.6	227	97.8	362	96.6	563	181	98.9
8277	HAL39	99.4	100	98.7	222	97.9	350	96.7	549	174	98.9
4388	ROT6	99.3	110	98.6	232	97.8	371	96.5	571	179	98.9
4408	ESP15	99.2	125	98.4	270	97.5	408	96.4	601	224	98.6
4440	ESP30	99.2	135	98.4	272	97.4	426	96.3	618	251	98.5
4442	ALB1	99.3	121	98.5	240	97.8	365	96.7	552	199	98.8
4446	ROT1	99.2	132	98.4	263	97.6	389	96.6	558	221	98.7
4448	ROT2	99.5	88	98.7	218	97.8	366	96.7	550	186	98.9
9429	BZH14	99.0	168	98.2	305	97.2	465	95.7	718	262	98.4
4321	HQU4	99.0	165	98.1	312	97.2	462	96.0	661	290	98.2
4362B	OSH2	99.0	159	98.4	267	97.6	402	96.5	575	207	98.7
4364A	OSH3	99.0	162	98.1	307	97.3	444	96.2	631	250	98.5
4325A	QUEXII1	98.9	176	98.2	302	97.4	437	96.2	632	243	98.5
4327B	QUEXII2	98.9	174	97.9	341	96.9	509	95.5	750	274	98.3
4330C	QUEXII3	98.8	204	97.8	368	96.9	518	95.2	797	273	98.3
4461A	Iron Age	99.2	133	98.4	268	97.5	420	96.2	624	195	98.8
8415A	Sardinia	99.3	113	98.5	240	97.8	369	96.7	545	213	98.7
Mainz	DEB9	99.2	125	98.4	271	97.5	412	96.4	598	240	98.5
Mainz	DEB21	99.1	141	98.4	258	97.6	404	96.4	594	228	98.6
10311	KAR 6A	99.5	86	98.7	209	97.9	340	96.8	523	177	98.9
10312	KAR 11B	99.3	113	98.7	215	98.0	338	96.8	524	172	99.0
10313	KAR 16A	98.8	203	97.8	357	97.0	500	95.7	717	303	98.2
10314	SALZ 18A	99.4	102	98.6	235	97.8	359	96.8	528	183	98.9
10315	SALZ 21B	98.7	210	97.9	352	97.0	494	95.7	707	308	98.1

ACAD	Individual	Callrate QST1	Ns QST1	Callrate QST3	Ns QST3	Callrate QST6	Ns QST6	Callrate QST12	Ns QST12	Ns after editing QST6	Final Call rate
10316	SALZ 57A	98.3	286	97.1	475	96.2	635	94.4	923	449	97.3
10317	SALZ 77A	99.0	165	98.2	297	97.4	425	96.3	616	246	98.5
10318	EUL 41A	99.1	146	98.3	273	97.6	397	96.4	593	228	98.6
10319	EUL 57B	99.2	140	98.4	269	97.5	413	96.3	618	216	98.7
10320	QLB 26A	99.3	115	98.6	228	97.8	356	96.7	541	191	98.8
4360B	OSH1	98.9	174	98.0	324	97.2	465	95.9	674	276	98.3
4370B	OSH7	99.4	98	98.7	221	98.0	338	96.8	536	176	98.9
4308B	QUEVIII4	98.9	176	98.2	299	97.4	436	96.1	639	270	98.4
4335B	QUEXII6	99.3	112	98.6	225	97.9	353	96.8	532	190	98.9
10321	QLB 28b	98.9	185	98.1	315	97.2	466	95.9	674	289	98.3
9404A	BZH1	96.6	565	94.7	883	92.9	1171	90.4	1590	799	95.2
PROBE	fragmentase	99.3	115	98.4	265	97.5	415	96.3	607	201	98.8
PROBE	sonicated	98.8	195	97.8	371	96.7	549	94.9	848	251	98.5
Average		**99.1**	**154**	**98.2**	**294**	**97.4**	**441**	**96.1**	**650**	**248**	**98.5**

Supplementary Table S3. List of problematic regions of the human mt genome as defined by N calls from an alignment of all archaeological mt genome sequences at QST 6.

Problematic regions by np	location	description
208	D-loop	AT-rich
301-312	D-loop	C-stretch
435-436	D-loop	C-stretch
495-501	D-loop	C-stretch
803-805	RNR1	C-stretch
956-965	RNR1	C-stretch
1683-1684	RNR2	C-stretch
2490	RNR2	C-stretch
3169-3171	RNR2	C-stretch
3211	RNR2	C-stretch
3485-3486	ND1	C-stretch
3527-3529	ND1	C-stretch
3568-3586	ND1	C-stretch
3894-3896	ND1	C-stretch
3960-3962	ND1	C-stretch
4138-4140	ND1	C-stretch
4251-4252	ND1	C-stretch
4761-4775	ND2	AT-rich
5084-5094	ND2	AT-rich
5233-5235	ND2	C-stretch
5304-5305	ND2	C-stretch
5439	ND2	C-stretch
5450	ND2	C-stretch
5495-5517	ND2/TRNW	AT-rich
5897	non-coding	C-stretch
6314	COX1	C-stretch
6568	COX1	C-stretch
6847	COX1	C-stretch
7297-7312	COX1	AT-rich
7332-7334	COX1	unclear
7399-7406	COX1	C-stretch
7468-7469	TRNS	C-stretch
7492-7493	TRNS	C-stretch
7517-7535	non-coding/TRND	AT-rich
7540-7569	TRND	AT-rich
7817	COX2	C-stretch
7819	COX2	C-stretch
8030	COX2	C-stretch
8496-8505	ATP8	AT-rich
8560-8561	ATP8/ATP6	C-stretch
9527-9528	COX3	C-stretch
9556-9558	COX3	C-stretch
9573	COX3	C-stretch
9909	COX3	unclear
9998-10005	COX3	AT-rich
10112-10113	COX3	AT-rich
10193-10195	ND3	C-stretch

Problematic regions by np	location	description
10280	ND3	C-stretch
10450-10459	TRNR	AT-rich
10487-10495	ND4L	AT-rich
10675-10676	ND4L	unclear
10939	ND4	C-stretch
10949-10950	ND4	C-stretch
11142	ND4	C-stretch
11235-11236	ND4	C-stretch
11377-11380	ND4	AT-rich, palindrome
11428-11249	ND4	C-stretch
11675	ND4	C-stretch
11869	ND4	C-stretch
12086-12087	ND4	C-stretch
12110	ND4	C-stretch
12387	ND5	C-stretch
12970	ND5	C-stretch
13027-13028	ND5	C-stretch
13054-13060	ND5	C-stretch
13648-13650	ND5	C-stretch
13678-13679	ND5	C-stretch
13755-13765	ND5	C-stretch
14063	ND5	CA-rich
14156-14158	ND5	C-stretch
14245-14247	ND5	C-stretch
14342	ND6	C-stretch
14419-14427	ND6	C-stretch
14491-14515	ND6	C-stretch/AT-rich
14532	ND6	C-stretch
14777-14781	CYTB	AT-rich
14809-14816	CYTB	C-stretch
15265-15268	CYTB	C-stretch
15527-15528	CYTB	C-stretch
15538-15544	CYTB	C-stretch
16186-16191	D-loop	C-stretch
16260-16262	D-loop	C-stretch
16377	D-loop	C-stretch

Supplementary Table S4. Summary of SMRT sequencing data and read assembly.

	BLASR assembly							Geneious Pro re-assembly (20bp cut-off)			
Sample	**BLASR no. reads mapped**	**BLASR mean read length**	**BLASR 95 % read length**	**Unused no. of reads**	**Unused mean read length**	**Unused 95% read length**	**BLASR Alignment Accuracy (%)**	**Reassembly (20bp cut-off) no. unused reads**	**Reassembly no. reads**	**Coverage (%)**	**Mean redundancy (min/max)**
H1_BZH1	51,945	39	68	16,410	43	66	95.9	17,543	34,402	98.5	110 (0/724)
H2_BZH4	44,040	41	66	6,551	41	63	96.9	10,072	33,968	99.1	99 (0/449)
H3_BZH6	62,646	43	69	8,672	41	62	96.8	12,839	49,807	99.8	154 (0/632)
H4_OSH2	46,802	41	67	4,791	38	63	96.2	7,883	38,919	98.7	110 (0/548)
H5_QUEXII3	45,093	43	75	9,039	45	73	95.5	12,930	32,163	99	107 (0/1500)
H6_DEB9	51,808	42	68	6,324	37	63	96.7	8,720	43,088	99.5	125 (0/632)

Supplementary Table S5. Primer sequences for additional coding region SNP confirmation via direct PCR and Sanger sequencing.

SNP	Name	Sequence 5' to 3'	PCR target size (bp)
3010 (H1)	L02988	CAACAATAGGGTTTACGACCTC	71
	H03017	AACGAACCTTTAATAGCGGCTG	
10211 (H23)	L10184	TTACGAGTGCGGCTTCGAC	80
	H10213	AGAAGGTAATAGCTACTAAGAAGAATTTTATGG	
11152 (H26)	L11148	AACCACACTTATCCCCACCTT	78
	H11187	AAGTATGTGCCTGCGTTCA	
14060 (discordant)	L14031	CCTGACTAGAAAAGCTATTACCTAAAACA	80
	H14062	GCCTTTTTGGGTTGAGGTGAT	
14063 (discordant)	L14062	ACCAAATCTCCACCTCCATCA	79
	H14098	AGTGGGAAGAAGAAAGAGAGGAA	
2772 (H46)	L02771	AATGCAAACAGTACCTAACAAACC	72
	H02797	CGCCCCAACCGAAATTTTTAATG	
11893 (private)	L11867	TCGCTAACCTCGCCTTAC	66
	H11895	ACGTGGTTACTAGCACAGAGA	
10675 (discordant)	L10671	TTGCCATACTAGTCTTTGCCG	76
	H10703	CCATATGTGTTGGAGATTGAGACT	
10521 (discordant)	L10520	CTAGCATTTACCATCTCACTTCTA	67
	H10542	ATAGTAGGGAGGATATGAGGTG	

Supplementary Table S6. Results for HVS I, HVS II and selected coding region SNPs from direct PCR followed by Sanger sequencing.

Individual	hg mt genome	HVS I Haplotype	np	HVS II Haplotype	np	Coding SNPs
HAL36	H23	rCRS	15997-16409			10211
HAL11	H	16093C, 16129A	15997-16409			
HAL32	H26	rCRS	15997-16409			11152
HAL39	H1e	rCRS	15997-16409			
DEB9	H88	rCRS	15997-16409			
DEB21	H1j	rCRS	15997-16409			
KAR 6A	H1bz	rCRS	16046-16402	263G, (315.1C)	34-397	
KAR 11B	H	rCRS	16046-16402	152C, 263G, (315.1C)	34-397	
KAR 16A	H46b	rCRS	15997-16409	263G, (315.1C)	34-397	2772, 11893
OSH2	H89	rCRS	15997-16409			
OSH3	H1	rCRS	16017-16409			
OSH1	H16	rCRS	16056-16409			
OSH7	H5b	16304C	15997-16409			
SALZ 18A	H10i	16093C	16046-16402			
SALZ 21B	H1e7	rCRS	15997-16409	263G, (315.1C)	34-397	
ESP30	H1e1a5	rCRS	15997-16401			
HQU4	H7d5	rCRS?	16288-16409			
SALZ 57A	H3	rCRS	15997-16409	152C, 263G, (315.1C)	34-397	
SALZ 77A	H3	rCRS (ACAD) 16150T (Mainz)	15997-16409			
ESP15	H6a1a	16362C	15997-16401			
BZH6	H1ca1	16189C	16056-16193, 16210-16409			3010
BZH4	H1e8	16293G	16056-16409			3010
ROT6	H5a3	16304C	15997-16409			
ALB1	H3b	rCRS	15997-16409			
ROT1	H3ao2	16256T	15997-16409			
ROT2	H5a3	16304C	15997-16409			
QUEXII1	H4a1	rCRS	15997-16409			
QUEXII2	H4a1	rCRS	15997-16409			
QLB 26A	H1	rCRS	16046-16402	263G (309.1C, 309.2C, 315.1C)	34-397	
QUEXII3	H13a1a2c	rCRS	16056-16409			
QLB 28b	H1	rCRS	16046-16402	263G, (309.1C, 309.2C, 315.1C)	34-397	
BZH1	H11a	16293G, 16311C	16056-16409			
BZH8	H2a1a3	16240t, 16354T	16056-16409			
BZH14	H82a	16220G	16056-16409			
EUL 41A	H4a1a1a2	rCRS	15997-16409	73G, 263G, (309.1C, 315.1C)	34-397	
EUL 57B	H3	rCRS	15997-16409	152C, 263G, (315.1C)	34-397	
QUEVIII4	H7h	16213A	15997-16409			
Sardinia	H1aw1	rCRS	15997-16409			
Iron Age	H90a	rCRS	15997-16409			

Supplementary Table S7. Overview of data groupings for summary statistics and NP-MANOVA in Table 2.

Culture	Four time periods	Three time periods (PCA)	Two time periods	Ind.	Sub-hg	H	H 1	H 1a	H 1b	H 2a	H2 a1	H 3	H 4	H 5	H 5a	H 6	H 6a	H 8	H 7	H 11
LBK n=9	Early Neolithic n=13	LBK n=9	Early Neolithic (ENE) n=13	HAL36	H23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				HAL11	H17'27	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				HAL32	H26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				HAL39	H1e	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				DEB9	H88	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				DEB21	H1j	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				KAR6A	H1bz	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				KAR11B	H9'H69	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				KAR16A	H46b	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rössen n=4				OSH2	H89	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				OSH3	H1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				OSH1	H16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				OSH7	H5b	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Schöningen n=2	Middle Neolithic n=6	Middle Neolithic n=10	Not included	SALZ18 A	H10i	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				SALZ21B	H1e7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Baalberge n=2				ESP30	H1e1a5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				HQU4	H7d5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Salzmünde n=2				SALZ57 A	H3a'j	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
				SALZ77 A	H3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Corded Ware n=2	Late Neolithic n=9	Not included	Late Neolithic/ Bronze Age (LNBA) n=16	ESP15	H6a1a	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
				BZH6	H1_TB D	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Bell beaker n=7		Bell beaker n=7		BZH4	H1e7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				ROT6	H5a3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

				ALB1	H3b	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
				ROT1	H3ao2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
				QUEXII1 +2	H4a1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
				QLB 26A	H1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
				QUEXII3	H13a1a 2c	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Unetice n=6	Early Bronze Age n=6	Not included		BZH1	H11a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
				BZH8	H2a1a3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
				BZH14	H82a	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				EUL41A	H4a1a1 a5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
				EUL57B	H3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
				QUEVIII 4	H7h	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Supplementary Table S8. Details of comparative Hg H sub-clade frequency data and geographic coordinates used for principal component analyses.

Population	Abbreviation	n	lat	long	H*[1]	H1	H1a	H1b	H2a	H2a1	H3	H4	H5	H5a	H6	H6a	H8	H7	H11
Galicia[62]	GAL	124	42.50	-8.10	15.3[2]	35.5	2.4	0.0	4.0	2.4	20.2	7.3	0.0	6.5	0.0	4.8	0.8	0.8	0.0
Cantabria[62]	CNT	53	43.33	-4.00	11.3	37.7	7.5	0.0	5.7	0.0	17.0	5.7	3.8	5.7	0.0	5.7	0.0	0.0	0.0
Catalonia[62]	CAT	40	41.82	1.47	22.5	32.5	2.5	2.5	0.0	2.5	12.5	5.0	5.0	10.0	0.0	5.0	0.0	0.0	0.0
Galicia/Asturia[63]	GAS	65	43.33	-6.00	23.1	41.5	0.0	1.5	9.2	0.0	10.8	1.5	1.5	6.2	1.5	3.1	0.0	0.0	0.0
Cantabria2[63]	CAN	31	43.33	-4.00	16.1	41.9	0.0	3.2	0.0	0.0	16.1	0.0	6.5	3.2	0.0	12.9	0.0	0.0	0.0
Cantabria3 (Potes)[63]	POT	38	43.15	-4.62	21.1	57.9	0.0	2.6	0.0	0.0	18.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cantabria4 (Pasiegos)[63]	PAS	22	43.23	-3.81	9.1	63.6	9.1	0.0	0.0	0.0	0.0	18.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Basques (Vizcaya)[63]	VIZ	37	43.25	-2.98	8.1	59.5	0.0	5.4	2.7	0.0	10.8	8.1	0.0	2.7	2.7	0.0	0.0	0.0	0.0
Basques (Guipuzcoa)[63]	GUI	63	43.17	2.17	7.9	42.9	0.0	3.2	20.6	1.6	19	0.0	0.0	4.8	0.0	0.0	0.0	0.0	0.0
Basques[63]	BMI	85	42.88	-1.92	14.1	43.5	0.0	3.5	9.4	1.2	17.6	2.4	1.2	7.1	0.0	0.0	0.0	0.0	0.0
Iberian Peninsula NE[63]	IPNE	52	39.47	-0.38	26.9	28.8	0.0	7.7	9.6	1.9	7.7	1.9	0.0	3.8	0.0	3.8	0.0	5.8	1.9
Turkey[64]	TUR	90	39.92	32.83	58.9	10.0	2.2	1.1	2.2	1.1	0.0	5.6	7.8	2.2	1.1	0.0	1.1	5.6	1.1
Armenia[64]	ARM	54	40.18	44.52	57.4	7.4	0.0	0.0	1.9	7.4	0.0	9.3	3.7	1.9	3.7	0.0	1.9	5.6	0.0
Georgia[64]	GEO	30	41.72	44.78	60.0	6.7	0.0	0.0	0.0	3.3	0.0	3.3	23.3	0.0	0.0	0.0	0.0	3.3	0.0
Northwest Caucasus[64]	NWC	69	44.00	40.00	59.4	13.0	0.0	1.4	1.4	4.3	1.4	0.0	8.7	0.0	1.4	4.3	0.0	2.9	1.4
Dagestan[64]	DAG	69	43.10	46.88	39.1	10.1	0.0	0.0	29.0	5.8	1.4	2.9	0.0	1.4	2.9	4.3	1.4	1.4	0.0
Ossetia[64]	OSS	45	42.23	43.97	62.2	11.1	0.0	2.2	0.0	2.2	0.0	6.7	8.9	0.0	0.0	4.4	0.0	0.0	2.2
Syria[64]	SYR	28	33.50	36.30	57.1	0.0	0.0	0.0	10.7	0.0	0.0	3.6	7.1	0.0	7.1	3.6	7.1	3.6	0.0
Lebanon[64]	LBN	34	33.90	35.53	50.0	20.6	0.0	0.0	2.9	2.9	0.0	2.9	11.8	5.9	0.0	0.0	0.0	0.0	2.9
Jordan[64]	JOR	33	31.95	35.93	75.8	9.1	0.0	0.0	3.0	3.0	0.0	3.0	6.1	0.0	0.0	0.0	0.0	0.0	0.0
Arabian Peninsula[64]	ARB	50	24.65	46.77	48.0	4.0	0.0	0.0	2.0	14.0	0.0	10.0	4.0	0.0	12.0	4.0	0.0	2.0	0.0
Arabian Peninsula2[65]	ARE	24	25.25	55.30	75.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0	4.2	4.2	0.0	0.0	8.3
Karachay-Balkaria[64]	KBK	50	43.58	43.40	38.0	0.0	8.0	10.0	4.0	8.0	4.0	0.0	24.0	0.0	0.0	0.0	0.0	2.0	2.0
Macedonia[65]	MKD	82	42.00	21.43	50.0	12.2	0.0	3.7	0.0	2.4	1.2	6.1	12.2	2.4	0.0	3.7	0.0	4.9	1.2
Volga-Ural region[66]	VUR	50	57.28	52.75	42.0	28.0	6.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	4.0	0.0	0.0	8.0	8.0

Population	Abbreviation	n	lat	long	H*[1]	H1	H1a	H1b	H2a	H2a1	H3	H4	H5	H5a	H6	H6a	H8	H7	H11
Finland[66]	FIN	31	62.75	22.50	48.4	6.5	12.9	0.0	6.5	6.5	6.5	0.0	3.2	9.7	0.0	0.0	0.0	0.0	0.0
Estonia[66]	EST	50	59.42	24.75	38.0	20.0	4.0	12.0	2.0	6.0	6.0	0.0	2.0	2.0	4.0	0.0	0.0	2.0	2.0
Eastern Slavs[66]	ESV	165	55.75	37.62	31.5	21.8	4.2	5.5	3.0	8.5	4.2	1.8	1.8	4.2	6.1	0.0	0.0	1.2	6.1
Slovakia[66]	SVK	50	48.15	17.12	40.0	6.0	8.0	4.0	0.0	2.0	4.0	4.0	8.0	2.0	6.0	0.0	2.0	2.0	12.0
France[66]	FRA	50	48.86	2.35	34.0	22.0	4.0	0.0	4.0	0.0	12.0	0.0	8.0	2.0	6.0	0.0	0.0	8.0	0.0
Balkans[66]	BLK	50	41.33	19.82	46.0	8.0	0.0	4.0	0.0	0.0	8.0	0.0	6.0	10.0	8.0	0.0	0.0	6.0	4.0
Germany[65]	DEU	26	48.40	9.98	42.3	11.5	3.8	7.7	0.0	0.0	7.7	0.0	7.7	7.7	0.0	7.7	3.8	0.0	0.0
Austria[65]	AUT	959	47.27	11.38	41.0	20.6	3.3	3.8	4.0	2.4	7.2	2.5	4.3	4.8	0.3	2.9	0.1	2.4	0.4
Romania[65]	ROU	102	46.55	24.56	21.6	26.5	6.9	4.9	0.0	0.0	0.0	1.0	12.7	4.9	0.0	2.9	0.0	16.7	2.0
France Normandy[63]	FRM	37	49.44	1.10	37.8	29.7	5.4	2.7	0.0	0.0	10.8	0.0	2.7	0.0	2.7	5.4	2.7	0.0	0.0
Western Isles[63]	WIS	39	53.33	-8.00	20.5	35.9	0.0	5.1	0.0	0.0	12.8	5.1	2.6	5.1	0.0	7.7	0.0	2.6	2.6
Czech Republic[63]	CZE	31	50.08	14.47	22.6	32.3	3.2	3.2	0.0	0.0	6.5	0.0	6.5	9.7	0.0	9.7	0.0	0.0	6.5
Linear Pottery culture	LBK	9	51.48	11.97	66.7	33.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bell Beaker	BBC	7	51.48	11.97	14.3	28.6	0.0	0.0	0.0	0.0	28.6	14.3	0.0	14.3	0.0	0.0	0.0	0.0	0.0
Middle Neolithic	MNE	10	51.48	11.97	30.0	30.0	0.0	0.0	0.0	0.0	20.0	0.0	10.0	0.0	0.0	0.0	0.0	10.0	0.0

[1]binning category summarizing all remaining H sub-hg
[2]values given as percentage %

Supplementary Methods

Archaeological background

The Neolithic in Central Europe was an epoch of cultural and social complexity and dynamism, which is reflected in a succession of archaeologically distinct cultural layers[67]. Many of these are temporally interwoven and the assemblages of artefacts of certain cultures can resemble features of preceding ones. At other times, new cultures appear/emerge without a close resemblance to preceding cultures and carry elements in style and/or artefacts which appear 'exotic' (i.e. have their origins outside Central Europe). This succession of distinct cultural layers provides the basis for our main underlying question: whether particular cultural changes were triggered or accompanied by genetic/population changes.

Our area of interest, the Mittelelbe-Saale (MES) region of Saxony-Anhalt, Germany, has long been an attractive settlement area, with a tight network of waterways and close to long-established trade routes. Climatically and geologically, the MES region benefits from its location in the rain-shadow of the nearby Harz Mountains, its deposits of ores and salts, and its fertile black and para-brown soils. This explains why this settlement cluster was highly sought after by many successive groups. Archaeological research postulates recurrent streams of colonisation and a range of demographic models both for the MES region and Europe in general[67-69].

The Neolithic in Central Europe lasted for roughly 3500 years and is commonly divided into an early, intermediate and later phase (the Early Neolithic, the Middle Neolithic and the Late Neolithic). Our rationale was to type samples from all three phases in a 'transect through time'. Archaeological details for each of the individuals and their cultures are given in Supplementary Table S1.

Early Neolithic:

The Early Neolithic (5450–3950 BC) in MES is represented by four archaeological cultures: the *Linienbandkeramik* (linear pottery culture or LBK), the *Stichbandkeramik* (stroke ornamented culture), and the following cultures *Rössen* and *Gatersleben*, named after eponymous sites in MES. A detailed archaeological record traces the spread of agriculture via LBK farmers back to the Carpathian Basin, the Balkans, through Greece and back into Anatolia and the Near East where farming originated ~12,000 years ago[70,71]. For the three subsequent Early Neolithic cultures with a smaller geographical distribution, a cultural origin outside the MES region (but nevertheless in Central Europe) is assumed, which explains the temporal overlap of the beginning and ending of subsequent cultures:
Linienbandkeramik (LBK): 5450–4775 BC,
Stichbandkeramik (SBK): 4925–4625 BC,
Rössen (RSK): 4625–4475/4250 BC and
Gatersleben (GLK): 4475–4250 BC.

The LBK is the oldest Neolithic culture in Central Germany and represents the first people in the region to exploit agriculture and animal husbandry. The LBK was identified earliest in Western Hungary (Transdanubia), where it incorporated novel technologies and ideas from Anatolia, the Levant and the Near East during the Neolithic transition. From there, the LBK expanded relatively rapidly along the major waterways and fertile Loess towards Central Europe, reaching as far as the Paris Basin in the West and Ukraine in the East[72]. The terminal Early Neolithic is characterised by *Epi-Rössen* and *Epi-Lengyel*, two horizons (Late-*Lengyel*-horizon; 4250-3950 BC) that are defined by distinct ceramics found at the sites *Gröna* and *Schiepzig*, respectively.

Middle Neolithic:

The Middle Neolithic (4250–2725 BC) in Saxony-Anhalt encompasses eight cultures: *Michelsberg, Baalberge*, *Salzmünde*, *Tiefstichkeramik*, *Walternienburg*, *Bernburg*, *Elb-Havel* and *Kugelamphorenkultur* (Globular Amphorae). Most of these cultures are only present in distinct regions of Saxony-Anhalt and emerged locally, whereas others originated elsewhere. This explains the large overlap of temporal successions and partial contemporaneity/co-existence in neighbouring regions:
Michelsberg (MBK): 4250–3500 BC
Jordansmühle (JMK): 4100–3650 BC
Baalberge (BAC): 3950–3400 BC
Salzmünde (SMC): 3400–3100/3025 BC
Tiefstichkeramik (TSC): 3650–3325 BC
Walternienburg (WBC): 3325–3100 BC
Bernburg (BEC): 3100–2650 BC
Elb-Havel (EHC): 3100–2725 BC

By and large, the Mid-Neolithic in Saxony-Anhalt is defined by two phases of successive cultures unified by the overarching theme of the Funnel beaker tradition, an older phase (3950–3100 BC) and younger one (3100–2650 BC). Some of these cultures are interlocked with previous ones, whereas others seem to replace existing ones, arguing for differential processes over time and space. Towards the end of the Mid-Neolithic, a seemingly foreign culture (*Fischbeck* group) whose origin is not yet resolved can be observed in eastern parts of the Mittelelbe-Saale area. It resembles the late Globular Amphorae culture and subsequently gives rise to the *Schönfeld* culture.

Late Neolithic:

The Late Neolithic in MES is represented by five cultures, including the
Schnurkeramik or *Corded Ware* culture (CWC): 2800–2200/2050 BC,
Einzelgrab or *Single-grave* culture (SGC): 2800–2200/2050 BC,
Schönfeld culture in the Northeast (SFC): 2725–2050 BC,
Ammensleben: 2650–2200/2050 BC, and
Glockenbecher or *Bell Beaker* culture/phenomenon (BBC): 2500–2200/2050 BC.

The Late Neolithic horizon is defined as 2725–2200 BC, even though earliest signs of the Corded Ware culture can be found around 2800 BC, whereas remains of beaker cultures can last as long as 2050 BC. The Bell Beaker culture is evident in the archaeological record from 2500 BC onwards in the South of MES, when it starts to move into settlement areas previously occupied by Corded Ware people, the latter of which have affinities to archaeological groups further east[73]. The settlement density in MES increases during its later phase (2300-2050 BC), when the Corded Ware is superseded by Bell Beaker elements, often at the same sites. This Late Neolithic Bell Beaker phenomenon is of particular interest, since archaeological evidence suggests it originated in the Tagus region of Western Iberia around 2800-2700 BC before spreading to become one of the first major pan-European cultures[74]. It has been traced archaeologically over large parts of Western Europe (including enclaves in North Africa) as far as Hungary, Ireland, and southern Scandinavia. Earlier Neolithic cultures were overlain/infiltrated by discernable Bell Beaker elites with a cultural package endowed with rich grave goods (including the eponymous bell-shaped ceramic beakers).

During the transition to the Bronze Age, early Bronze Age cultural elements of the Unetice culture appear contemporaneously to late elements of the Bell Beaker culture, again sometimes also at the same site. The site of Eulau, famously known for its oldest nuclear family graves, represents a good example of cultural dynamics during the Late Neolithic, as it shows the presence of both Corded Ware and Bell Beaker cultures, and later of the early Bronze Age Unetice culture.

Ancient DNA work

All ancient hg H individuals in this study were selected from a large pool (currently 378 Neolithic samples) from the Mittelelbe-Saale region in Saxony-Anhalt, Germany (Brandt et al., in preparation). This sample collection forms the core of an interdisciplinary, multi-centre project lead by the State Office for Heritage Management and Archaeology Saxony-Anhalt / State Museum for Prehistory Halle and the Bioarchaeometry group of the Johannes Gutenberg University of Mainz, Germany, and includes the Australian Centre for Ancient DNA (ACAD). All samples reported in this study are from recent excavations (2000 and younger), except for samples from the site of Derenburg, which were excavated from 1996-1999). A minimum of two samples per individual were collected under 'DNA-free' conditions and/or largely in situ following established protocols in collaboration with staff from the State Office in Halle[75]. Samples were not washed or treated after excavation and were kept refrigerated and/or in cooled conditions.

The majority of the ancient DNA (aDNA) work for this study was carried out at the specialized facilities at ACAD following appropriate criteria to prevent/minimise contamination with present-day DNA. DNA extractions, and sequencing of the mitochondrial control region HVS I for samples DEB9 and DEB21, were carried out at the aDNA facilities of the Johannes Gutenberg University of Mainz, Germany. In addition, sample preparation, HVS-I sequencing and coding region SNP-typing of samples KAR6, KAR11, KAR 16, SALZ 18, SALZ 21, SALZ 57, SALZ77, EUL 41 and EUL 57 were also carried out in Mainz.

Frequency estimates from our entire Neolithic time transect dataset (Supplementary Fig. 3) indicate an arrival of hg H with early LBK farmers[61,76,77], relatively stable levels of hg H in subsequent cultures, followed by a remarkable increase of hg H frequencies in individuals belonging to the Bell Beaker culture which expanded out of Iberia in the Late Neolithic. Based on these observations (and due the lack of further resolution of mtDNA hg H via existing approaches), we decided to sequence whole mt genomes from a subset of individuals already typed as hg H. Our aim was to test whether this hg H frequency increase over time (in particular in Bell Beaker individuals) also carried a phylogeographic signal in the form of a discernible distribution of H sub-hg. We selected hg H individuals randomly with the aim of gaining a balanced representation of individuals assigned to Early, Mid- and Late Neolithic and Early Bronze Age cultures from our transect (Supplementary Tables S1).

DNA extractions

A silica suspension was prepared by adding 6g of silicon dioxide (Sigma-Aldrich) to 50mL of DNA-free distilled water. The suspension was left for one hour to pellet larger grain sizes, before 40mL of the supernatant containing the finest particles was transferred to a new tube and kept overnight for further settling. Finally, a working silica suspension was created by discarding ~30mL of supernatant, retaining 10mL of the medium sized silica particles.

Preparation of tooth and bone samples for DNA extraction was carried out as previously described[78]. In previous studies we routinely used a phenol-chloroform based DNA extraction protocol that involved washing and concentration steps on Amicon filter units with a molecular weight cut-off of 30 and/or 50kDa, which results in a gradual loss of double-stranded DNA smaller than ~125bp (Amicon Ultra 4, User guide 2011). To recover DNA fragments of all sizes, and especially from shorter fragments <100 bp, we designed a customised DNA extraction protocol based on a standard silica-based extraction. On average 0.2g of tooth/bone powder were incubated overnight under constant rotation at 37°C in 4.44mL of lysis buffer consisting of 0.5M EDTA, pH 8.0; 0.5% N-lauroylsarcosine; and 0.25mg/μL proteinase K. After lysis, samples were centrifuged at 4,600 rpm for 1min and the supernatant transferred to a new 50mL tube. 125uL of medium-sized silica suspension (see

above) and 16mL of in-house binding buffer (13.5mL QG buffer (Qiagen), 2.86mL of 1X Triton, 20mM NaCl, 0.2M acetic acid (all Sigma-Aldrich)) were added and DNA was left to bind to silica overnight at room temperature under slow and constant rotation. The pH indicator included in the QG buffer ensured we were maintaining the optimal pH conditions necessary for the binding of DNA to silica. On the third day the sample was centrifuged for 1min to pellet the silica particles and the supernatant was poured off. The pellet was transferred to a 1.5mL tube and washed three times by resuspension in 1mL 80% ethanol, centrifuged for 1min at 13,000rpm and the supernatant removed. The pellet was left to dry for 30min and subsequently resuspended in 200µL of pre-warmed (to 50°C) TE buffer (10mM Tris, 1mM EDTA) and incubated for 10min. After pelleting for 1min at 13,000rpm the supernatant was collected, aliquoted and stored at -18°C until further use.

PCR amplifications, HVS I sequencing and coding region SNP typing

All ancient hg H individuals in this study were selected from a large pool of Neolithic samples (n=378) that were genotyped by direct sequencing of the mitochondrial hypervariable segment I (HVS I) and minisequencing of 22 coding region SNPs using a multiplex approach.

DNA was extracted from two independent samples for each individual. HVS I was amplified using a minimum of four short overlapping primer pairs, following established protocols as described previously[77,78]. Multiplex SNP typing of 22 haplogroup informative SNPs (GenoCoRe22) was carried out using a SNaPshot based protocol as described previously[77]. The GenoCoRe22 multiplex typing approach provided an ideal monitoring system for contamination of the ancient samples and non-template controls by present-day DNA, as the PCR multiplex directly targets SNPs of all the major Eurasian haplogroups likely to constitute potentially contaminating lineages. Mitochondrial results were considered genuine and authentic when all sequences from replicated overlapping PCRs (a minimum of 6-8 fragments) produced unambiguous results in accordance with the GenoCoRe22 multiplex typing results from two independent extractions. One extract from each successfully typed individual was subsequently used for DNA library preparation. We also designed primer pairs targeting selected single nucleotide variants (discordant calls) from multiple quality score threshold settings, as well as known characteristic sub-haplogroup SNPs (e.g. H1, H23, etc.) in order to confirm or exclude these via direct sequencing from the original and/or independent extract (Supplementary Table S5). In addition, the complete mt genomes of all staff at ACAD involved directly in the handling of the samples (P.B., W.H., C.J.A., and J.T.) and downstream steps of this study were sequenced to monitor potential contamination. DNA was extracted from swab samples and directly sequenced using standard protocols routinely used at the University of Arizona Genetics Core (UAGC[79]). We found no overlap of characteristic SNPs between our Neolithic hg H samples and staff members. Haplotypes and polymorphic sites in mt genomes from staff members were as follows (aligned against rCRS)[80]:

ADL#3 (hg H1a4) 73, 263, (309.1C, 315.1C), 750, 1438, 3010, 4769, 8860, 9341t, 15326, 16162, 16519; ADL#4 (hg H1be) 263, (309.2C, 315.1C), 750, 1438, 3010, 4769, 8860, 10750, 13035, 15326, 16192, 16519; ADL5#5 (hg H56c) 263, (309.2C, 315.1C), 750, 1438, 4769, 8860, 11788, 14129, 15326, 16519; and ADL#6 (hg H1z1) 263, (309.2C, 315.1C), 327, 750, 1438, 3010, 4769, 8860, 10632, 11428, 15326, 16189, 16311.

Ancient DNA Library preparation

DNA polishing/phosphorylation reactions were performed at 100µl final volume with 5 to 25µl of aDNA extract added to reactions comprising 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 1mM ATP, 10mM Dithiothreitol, 250µg/ml rabbit serum albumin (RSA; Sigma), 400µM of each dNTP (Invitrogen), 50U T4 Polynucleotide Kinase (New England Biolabs, NEB), 10U

DNA Polymerase I, Large (Klenow) Fragment (NEB), and 15U T4 DNA Polymerase (NEB). Thermocycling profiles consisted of 25°C for 15 min, 37°C for 15min, and 12°C for 15min. At 12°C, 10µl of 0.5M EDTA pH 8.0 (Sigma) was added, followed by 550µl Qiagen Buffer PB1. DNA was purified using MinElute spin columns (Qiagen) as per the manufacturer's instructions (Figure 3).

Adaptor ligation reactions were performed at 60µl final volume with reactions comprising 62.8mM Tris-HCl pH 7.6, 10mM $MgCl_2$, 1mM ATP, 2.8mM Dithiothreitol, 6% Polyethylene Glycol (PEG 6000), 2µM Adaptor UniHyb-A, 2µM Adaptor UniHyb-B, and 4,000U T4 DNA Ligase (NEB). The thermocycling profiles consisted of 20 cycles of 24°C for 1min, 16°C for 30sec, and 8°C for 30sec. Then 300µl Qiagen Buffer PB1 was added to each reaction. DNA was isolated from the rest of the reaction components using MinElute spin columns (Qiagen) as per the manufacturer's instructions. The partially double stranded adaptor UniHyb-A comprised the oligonucleotides UniHyb-Af GGTGTTGTTAGGAATGCGAGA and UniHyb-Ar TCTCGCATTCCTAA. The partially double stranded adaptor UniHyb-B comprised the oligonucleotides UniHyb-Bf AGGATAGGTCGTTGCTGTGTA and UniHyb-Br TACACAGCAACGA. UniHyb-A and UniHyb-B were formed by hybridisation with a thermocycling profile of 95°C for 2min, then 75°C for 20sec, followed by a ramp from 75°C to 10°C at 2°C/min increments.

Polymerase 'fill-in' reactions, to remove nicks and to create fully double-stranded adaptor-tagged aDNA, were performed at 30µl final volume with reactions comprising 20mM Tris-HCl pH 8.8, 10mM $(NH_4)_2SO_4$, 10mM KCl, 2mM $MgSO_4$, 0.1% Triton X-100, 250µM of each dNTP, and 16U Bst DNA Polymerase, Large Fragment (NEB). The thermocycling profile was 37°C for 30min. Then 150µl Qiagen Buffer PB1 was added to each reaction. DNA was purified using MinElute spin columns (Qiagen) and eluted into 21µl as per the manufacturer's instructions.

PCR amplification reactions, to create 'master' DNA 'fill-in' reaction libraries, were performed in 3 x 44µl volumes per original sample/extract, with 7µl of eluted DNA added per tube to reactions comprising 1x AmpliTaq Gold buffer II, 2.5mM $MgCl_2$, 2.5U AmpliTaq Gold (Applied Biosystems), 250µM of each dNTP (Invitrogen), and 0.5µM of PCR primers UniHyb-PCR-A (GGTGTTGTTAGGAATGCGAGA) and UniHyb-PCR-B (AGGATAGGTCGTTGCTGTGTA), see also Figure 3. The thermocycling profile consisted of 94 °C for 11 min, followed by 12 cycles of 30sec at 95°C, 30sec at 60°C and 1min (+2 sec/cycle) at 72°C, followed by a final 10min at 72°C. The 3 x 44µl volume reactions were pooled. From each original sample/extract 'master' DNA library pool, 2.5µl was then added to 4 x 35µl reactions comprising: 1X AmpliTaq Gold buffer II, 2.5mM $MgCl_2$, 250µM of each dNTP, 1.0U AmpliTaq Gold (Applied Biosystems); and 0.5µM of PCR primers UniHyb-PCR-A and UniHyb-PCR-B. The remainders of the 'master' DNA libraries were archived at -80°C. Once reamplified using the above thermal profile and pooled, these 4 x 35µl amplification reactions were purified using MinElute spin columns (Qiagen) and eluted into 15µl as per the manufacturer's instructions. These comprised the 'primary' DNA libraries and amplification products were sized and quantified via gel electrophoresis against quantified size markers (HyperLadder™ V, Bioline) and a Nanodrop 2000 (Thermo Scientific).

Hybridisation-based enrichment of human mtDNA target sequences within libraries

The basic design for the hybridisation of tracer (in this case library) DNA and biotinylated driver (in this case human mitochondrial probe) DNA sequences was previously described by Patel and Sive (2001)[81], see also Figure 3. The main differences were as follows. First, in addition to the described HEPES/NaCl-based hybridisation conditions, SSC-based hybridisation conditions were also used with Na^+ concentrations between 660-990mM. Second, hybridisation reactions were carried out in a total volume of 30µl. Third, 300-400ng of

MinElute-concentrated 'primary' library DNA and 85-120ng of biotinylated human mitochondrial probe DNA were used. Fourth, the thermocycling profiles used for hybridisation were 95°C for 5min, followed by 14-18h at 50 or 55°C. Fifth, and finally, the two library primers UniHyb-PCR-A and UniHyb-PCR-B were included as part of the hybridisation mix at 0.67-1.0µM.

There were two reasons behind the inclusion of the library amplification PCR primers in the hybridisation mix. First, by annealing to their complementary sequences in the adaptors flanking the ancient DNA insert sequences in the PCR amplified libraries, UniHyb-PCR-A and UniHyb-PCR-B could act as 'blocking oligos'[82] to minimise unwanted hybridisation between the adaptor-tagged flanking regions of otherwise unrelated single-stranded library DNA molecules. Second, following the binding of biotinylated probe molecules to magnetic Streptavidin beads and stringency washes to remove non- or weakly-hybridised single-stranded library DNA molecules, incubation with dNTPs and a DNA polymerase with strand-displacing activity would allow primer extension to neatly disrupt the double-stranded region of stable hybridisation between human mitochondrial probe DNA sequences and single-stranded library DNA molecules that had inserts with complementary sequences. The net effect of this polymerase 'filling-in' reaction would therefore be to free the originally single-stranded library DNA molecules captured by the probe molecules by hybridisation: the biotinylated probe molecules would remain bound to the magnetic Streptavidin beads and the discrete now-'filled-in' double-stranded library DNA molecules could be removed cleanly in the free supernatant using a magnetic rack (Invitrogen) which would keep the bead-bound probe molecules pelleted adjacent to the magnet.

Following the 14 to 18h overnight hybridisation step, the biotinylated probe molecules in the 30µl hybridisation reactions were bound and immobilised to magnetic Streptavidin beads as previously described[83]. Successive stringency washes, to remove non- or weakly-hybridised single-stranded library DNA molecules, used decreasing salt and increasing temperature and followed the following profile: 2X SSC/0.1% SDS at 37°C for 1min; 2X SSC/0.1% SDS at 42°C for 10min; 1X SSC/0.1% SDS at 43°C for 10min; 0.5X SSC/0.1% SDS at 44°C for 10min; 0.5X SSC/0.1% SDS at 45°C for 10min. At this point, the beads (with attached biotinylated probe molecules and stably hybridised single-stranded library DNA molecules with complementary insert sequences) were attached to the tube side via the magnet and the supernatant was removed via pipetting and discarded. The beads were then immediately resuspended in buffer conditions to suit a strand-displacing polymerase activity.

Two strand-displacing polymerases, the Klenow Fragment (3´→5´ exo⁻) and the *Bst* DNA Polymerase, Large Fragment (NEB) were used. With the Klenow Fragment (3´→5´ exo⁻), the reactions were performed at 60µl final volume comprising 15mM Tris-HCl pH 8.0, 50mM KCl, 5.5mM $MgCl_2$, 250µM of each dNTP, and 5U Klenow Fragment (3´→5´ exo⁻). The tubes were incubated at 37°C for 15-30min with regular agitation to keep the beads in suspension. The reaction tube was then applied to the magnetic rack and the 60µl supernatant transferred to a fresh 200µl tube. This tube was immediately incubated at 80°C for 20min to inactivate the enzyme. With the *Bst* DNA Polymerase, Large Fragment, the reactions were performed at 35µl final volume comprising 15mM Tris-HCl pH 8.0, 50mM KCl, 10.0mM $MgCl_2$, 200µM of each dNTP, and 100µg/ml of BSA. The beads were resuspended in the buffer – minus the *Bst* DNA Polymerase, Large Fragment enzyme – and pre-heated to 60°C. The enzyme (16U) was added and the tubes were incubated at 60°C for 5min with regular agitation to keep the beads in suspension. The reaction tube was then applied to the magnetic rack at 60°C and the 35µl supernatant transferred to a fresh 200µl tube. This tube was immediately incubated at 80°C for 20min to inactivate the enzyme.

Using either approach, the heat-inactivated supernatant was split between several (generally 4-8) PCR reamplification reactions (total combined volume 140µl), designed so that

upon the addition of the sub-portion of Klenow or *Bst* buffer, the final composition of the reactions would be 1X AmpliTaq Gold buffer II, 2.5mM $MgCl_2$, 250µM of each dNTP, 1.0U AmpliTaq Gold (Applied Biosystems), and 0.5µM of PCR primers UniHyb-PCR-A and UniHyb-PCR-B using the above thermal profile used to create 'primary' DNA libraries. Amplification reactions were pooled and library amplicons purified using MinElute spin columns (Qiagen) and eluted into 15µl as per the manufacturer's instructions. These comprised the 'first enrichment' DNA libraries and amplification products were sized and quantified via gel electrophoresis against size markers (HyperLadder™ V, Bioline) and a Nanodrop 2000 (Thermo Scientific). In general, the whole hybridisation / enrichment / reamplification cycle was repeated three times, to produce 'third enrichment' DNA libraries highly enriched for short fragments of endogenous mtDNA sequence within the ancient DNA libraries.

The above system underwent development and optimisation over the period of this research. A number of parameters were varied in order to optimise the enrichment and coverage of human mtDNA sequences within libraries. Overall, a number of observations can be made about the above hybridisation / enrichment / reamplification steps. Most importantly, of all the parameters varied, the only ones that appear to have made a substantial difference to the experimental outcome were the stringency wash steps. High temperatures and/or low salt concentrations generally led to poor subsequent library reamplification (presumably as many of the short human mtDNA inserts within libraries – generally 20-70 bases – did not remain bound to probe DNA efficiently beyond a certain point). Lower temperatures / higher salt concentration led to poorer coverage of the mt genome following enrichment, presumably due to the retention of significantly higher proportions of non-mtDNA sequences during lower stringency washes (data not shown). Empirically, the stringency wash conditions described above seemed to work well with DNA extracted and made into libraries from generally well-preserved archaeological remains. Switching between HEPES/NaCl-based versus SSC-based hybridisation conditions, using between 300-400ng of MinElute-concentrated 'primary' library DNA, using between 85-120ng of biotinylated probe, overnight hybridisation for 14 to 18h at either 50 or 55°C, or using between 0.67-1.0µM UniHyb-PCR-A and UniHyb-PCR-B primers all appeared to make little or no discernable difference.

Preparation of biotinylated human mitochondrial DNA probes

Total genomic DNA from a present-day individual belonging to mitochondrial hg J1c8 (with SNP differences from the revised Cambridge Reference Sequence (rCRS)[80] at nucleotide positions: 73G, 185A, 228A, 295, 462, 489, 2706, 3010, 4216, 7028, 10084, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15452A, 16069, 16126, 16261, 16265 and 16319) was extracted from a buccal swab using the DNeasy Blood & Tissue Extraction kit (Qiagen). This hg J1c8 mitochondrial genome was amplified in two overlapping fragments by long-range PCR (LR-PCR) using the PCR primer pair combinations L06363/H14799 (8476 bp) and L14759/H06378 (8340 bp). Primer details are: L06363 (ACCATCTTCTCCTTACACCTAGCAG), H14799 (GGTGGGGAGGTCGATGA), L14759 (AGAACACCAATGACCCCAATAC) and H06378 (GATGAAATTGATGGCCCCTAA). Probe was generated and used in one of two ways: (a) via LR-PCR with Biotin-dNTPs included in the dNTP mix, followed by DNA fragmentation; (b) via LR-PCR without Biotin-dNTPs, followed by DNA fragmentation then 3'-biotinylation with Biotin-ddUTP using the enzyme Terminal Transferase (TdT), see also Figure 3.

(a) Biotinylated probe DNA was generated by LR-PCR (Expand LR dNTPack, Roche) according to the manufacturer's recommendations. Biotin-dNTPs (NEB Biotinylated dNTP Mixture) were incorporated into the PCR reaction at a ratio of 1:10 with regular dNTPs. The whole mitochondrial genome was amplified in eight separate 25µl LR-PCR reactions using the two primer sets L06363/H14799 (8476 bp) and L14759/H06378 (8340

bp). In total 42 cycles of PCR were carried out using the Bio-Rad Tetrad 2 Thermal Cycler. Cycling conditions included: initial denaturation at 92°C for 2mins; 10 cycles of 92°C for 10sec, 60°C for 15sec, 68°C for 8min 40sec; followed by 32 cycles of 92°C for 10sec, 60°C for 15sec, 68°C for 8 min 40sec – increasing by 15 sec/cycle; final extension at 68° for 10mins; followed by a hold at 4°C. PCR reactions were pooled for each product and loaded onto 1% Agarose gels. Bands were visualized by ethidium bromide staining prior to excision of the correct-sized band under UV light with a clean, sharp scalpel blade. Purification of bands was carried out using the QIAquick Gel Extraction Kit (Qiagen) and eluted in 30µl as per manufacturer's instructions. Quantification was performed using gel electrophoresis against quantified size markers (HyperLadder™ V, Bioline) and a Nanodrop 2000 (Thermo Scientific). DNA was diluted to 120µl total volume with distilled water as a requirement for the sonicator prior to shearing. Fragmentation of DNA to the desired size range was achieved by discontinuous sonication (Microtip sonicator, Thomas Optical & Co. Pty. Ltd) at high speed, amplitude 6, for 4 x 1min intervals, no ice. Sonication was performed in time intervals to avoid successive heat build-up that can damage the DNA and the sample was subjected to centrifugation during each time interval, using a bench top centrifuge. Post sonication, DNA was purified and concentrated using the QIAquick PCR Purification Kit (Qiagen) and eluted in 30µl as per the manufacturer's instructions. The resulting biotinylated probe DNA was analysed on a 1% Agarose gel, with the resulting smear in the size range 200-600bp. DNA concentration was estimated using the Nanodrop. DNA from both overlapping halves of the mitochondrial genome was pooled to produce an equimolar concentration of biotinylated probe DNA across the genome, ready for hybridisation.

(b) Biotinylated probe DNA was generated by LR-PCR (Expand LR dNTPack, Roche) according to the manufacturer's recommended conditions. The whole mitochondrial genome was amplified in eight separate 25µl LR-PCR reactions using the two primer sets L06363/H14799 (8476 bp) and L14759/H06378 (8340 bp). In total 42 cycles of PCR were carried out using the Bio-Rad Tetrad 2 Thermal Cycler. Cycling conditions included: initial denaturation at 92° for 2 mins; 10 cycles of 92°C for 10sec, 60°C for 15sec, 68°C for 8min 40sec; followed by 32 cycles of 92°C for 10sec, 60°C for 15sec, 68°C for 8min 40sec – increasing by 15sec/cycle; final extension at 68° for 10mins; followed by a hold at 4°C. PCR reactions were pooled for each product and loaded onto 1% Agarose gels. Bands were visualized by ethidium bromide staining prior to excision of the correct-sized band under UV light with a clean, sharp scalpel blade. Purification of bands was carried out using the QIAquick Gel Extraction Kit (Qiagen) and eluted in 30 µl as per the manufacturer's instructions. Quantification was performed using a Nanodrop 2000 (Thermo Scientific). DNA was diluted to 120ul total volume with distilled water as a requirement for the sonicator prior to shearing. Fragmentation of DNA to the desired size range was achieved by sonication (Microtip sonicator, Thomas Optical & Co. Pty. Ltd) at high speed, amplitude 6, for 4min, no ice. Post sonication, DNA was purified and concentrated using the QIAquick PCR Purification Kit (Qiagen) and eluted in 30µl as per the manufacturer's instructions. The resulting DNA was analysed on a 1% Agarose gel, with the resulting smear in the size range 200-600bp. DNA concentration was estimated using gel electrophoresis against quantified size markers (HyperLadder™ V, Bioline) and a Nanodrop 2000 (Thermo Scientific). DNA from both overlapping halves of the mitochondrial genome was pooled to produce an equimolar concentration of biotinylated probe DNA across the genome, ready for biotinylation via Biotin-16-ddUTP and the enzyme Terminal Transferase (TdT). Prior to 3' end-labelling, sonicated probe DNA was made single stranded by heating at 95°C for 5min, then immediately placed on wet ice for 5min. Reactions for the 3' end-labelling of probe DNA were performed at 50µl final volumes comprising: 50mM Potassium Acetate; 20mM Tris-Acetate pH 7.9; 10mM Magnesium Acetate; 0.25mM Cobalt Chloride; sonicated

mitochondrial probe DNA at 10pmol of 3' ssDNA ends; 0.1mM Biotin-16-ddUTP (Enzo); 40U Terminal Transferase enzyme (NEB). The thermocycling profiles were: 37°C for 60min; then 70°C for 10min. DNA was isolated from the rest of the reaction components using Qiagen Nucleotide Removal Kit spin columns and eluted into 30ul as per the manufacturer's instructions.

Considerations of probe biotin labelling and library elution via strand-displacing polymerases are as follows. As described above, there were two versions of probe synthesis and preparation: (a) via LR-PCR with Biotin-dNTPs included in the dNTP mix, followed by DNA fragmentation; and (b) via LR-PCR without Biotin-dNTPs, followed by DNA fragmentation then 3'-biotinylation with Biotin-16-ddUTP using the enzyme Terminal Transferase (TdT). In addition, two strand-displacing polymerases were used to disrupt the double-stranded region of stable hybridisation between human mitochondrial probe DNA sequences and single-stranded library DNA molecules that had inserts with complementary sequences, and thereby elute library DNA molecules: (a) the Klenow Fragment (3´→5´ exo⁻); and (b) the *Bst* DNA Polymerase, Large Fragment. Initially, (a) and (a) were used together. Although this approach worked extremely well, on one occasion (probe batch October 2010) probe sequences could be identified in enriched library DNA sequences (almost exclusively in the overlapping region between the probe amplicons). Presumably, this occurred due to interactions between library molecules with mtDNA inserts, library oligos and free 3'-ends of probe molecules. Fortunately, we were able to eliminate these occasional hg J1c8 intrusions via direct PCR on independent ancient DNA extracts from the same archaeological individuals. However, this encouraged the development of the combined (b) and (b) approach. The rationale was that 3'-Biotin-16-ddUTP labelling the probe would prevent probe molecules from extending under any circumstances. In addition, using the *Bst* DNA Polymerase, Large Fragment at 60 °C should also drastically reduce any potential spurious, non-specific, hybridisation of library primers to probe (compared to those that might occur at 37°C). Once this approach had been adopted, we didn't see a single further instance of any identifiable probe SNP in any enriched libraries, for either human or non-human samples (data not shown).

Preparation of enriched libraries as Mitochip probes

Following the final (third) round of hybridisation / enrichment / reamplification, ~1/10 of the enriched library was reamplified with the dU-containing primers: U-UniHyb-Af (GGTGTTGUTAGGAAUGCGAGA) and U-UniHyb-Bf (AGGATAGGUCGTTGCUGTGTA). Following reamplification, the dU bases could subsequently be digested with a combination of Uracil DNA glycosylase (UDG) and the DNA glycosylase-lyase Endonuclease VIII. (NEB's USER Enzyme is a pre-mixed combination of these two enzymes.) UDG catalyses the excision of uracil bases, forming an abasic (apyrimidinic) site while leaving the phosphodiester backbone intact, whereas the lyase activity of Endonuclease VIII breaks the phosphodiester backbone at the 3´ and 5´ sides of the abasic site so that base-free deoxyribose is released. Effectively then, treatment of the U-UniHyb-Af / U-UniHyb-Bf amplified libraries with the USER Enzyme trims off all but 6 bases (A) or 5 bases (B) of the UniHyb-PCR-A and UniHyb-PCR-B library adaptors. This minimises any non-mtDNA hybridisation between unrelated single-stranded library DNA molecules via their adaptor sequences, maximizing the opportunities for authentic enriched human mtDNA sequences to hybridise to the Mitochip. The specific PCR amplification conditions used were 1X AmpliTaq Gold buffer II, 2.5mM $MgCl_2$, 250µM of each dNTP, 1.0U AmpliTaq Gold (Applied Biosystems), and 0.5µM of PCR primers U-UniHyb-Af and U-UniHyb-Bf. The thermocycling profile was 94°C for 10min; followed by 25-30 cycles of 30sec at 95°C, 30sec at 60°C and 30sec (+1 sec/cycle) at 72°C; followed by a final 20min at 72°C. DNA was purified using MinElute spin columns (Qiagen) as per the

manufacturer's instructions. USER Enzyme digestion took place for 2-3 hours at 37°C in 60µl final volume reactions comprising 15mM Tris-HCl pH 8.0, 50mM KCl, 1.5mM $MgCl_2$, 3U USER Enzyme. The entire reaction was then cleaned up by passing through a BioRad P-30 column as per the manufacturer's instructions and purified DNA sized and quantified as described above. The isolated DNA was then biotin-labelled as described below.

Affymetrix Mitochip v2.0 background and sample preparation

The Affymetrix Mitochip v2.0 is an oligonucleotide-tiling array for high-throughput resequencing analysis of the human mitochondrial (mt) genome[84]. For each nucleotide position (np) of the mt genome interrogated, the array possesses eight 25-mer probes - four each to match the heavy and the light strands, respectively, of the rCRS mt genome[80]. Each of these strand-specific 25-mers varies only at the central position. For each np, this allows all four possible alleles (A, T, C or G) to be interrogated on both strands within a tight local sequence context – via the hybridisation of fluorescently labelled fragments of sample DNA. The Mitochip v2.0 also carries additional local context probes to interrogate mtDNA sequence variants for 500 of the most common haplotypes (e.g. particular insertions, deletions, and adjacent or closely spaced SNPs) based on the FBI database[84]. There are a number of reasons why resequencing via oligonucleotide tiling arrays might be an attractive genotyping strategy for researchers dealing with samples that yield highly damaged and degraded DNA.

With some forensic, environmental, or formalin-fixed paraffin-embedded (FFPE) archived samples and medical biopsies, DNA degradation and physical fragmentation can be considerable and irreversible. Forensic, environmental, and in particular ancient/archaeological samples (including archived museum specimens covering unique and irreplaceable now-lost genetic diversity from both extant and extinct species), can contain minute amounts of highly damaged and degraded endogenous DNA. Worse still, up to 99% or more of DNA templates extracted from samples like these can be non-endogenous post-mortem and soil bacterial DNA sequences. DNA extracts with a high background complexity like these sometimes contain few or no endogenous DNA templates >100bp in length (as determined via PCR amplification with a primer pair targeting a 100bp region). However, a key finding in recent years has been that as one interrogates shorter and shorter DNA templates in ancient DNA extracts, the effective copy numbers of DNA targets in that size range increases exponentially. Evidence from a number of independent studies has conclusively demonstrated this negative exponential relationship between fragment size and template copy number[83,85-88].

This relationship between aDNA fragment size and copy number provides the rationale behind the use of the Mitochip v2.0. Provided short fragments of endogenous human mtDNA sequences of ~20-25 bases and above could be significantly enriched (over and above fragments of non-mt human DNA and the high levels of non-endogenous bacterial DNA sequences typically intrinsic to archaeological samples), the Mitochip v2.0 could interrogate every nucleotide position (np) of the mt genome via its overlapping 25 base windows. The chip's ability to SNP-type fragments of aDNA of 20-25 bases and above would maximise the chances of genotyping mt genomes from even highly marginal archaeological samples thousands of years old. Prior to the use of the Mitochip v2.0 as an aDNA genotyping tool, therefore, we had to develop and optimise new methodologies designed: (a) to efficiently extract short fragments of damaged and degraded DNA from archaeological remains; (b) to immortalise this extracted DNA as amplifiable libraries; (c) to enrich for human mtDNA target sequences within these libraries via hybridisation-based DNA-capture protocols; and (d) to label these mt-enriched DNA libraries for use as probes in Mitochip-based resequencing analyses of the human mt genome[84].

In everyday use, with DNA extracts from modern human samples, the 16569 bp mtDNA genome is typically amplified via two or three overlapping long-range (LR-)PCR targets[84,89-91]. Following LR-PCR, amplification products undergo fragmentation, labelling, and finally hybridisation, as described in the Affymetrix GeneChip CustomSeq Resequencing Array Protocol (http://www.affymetrix.com/support/mas/index.affx#1_2). Clearly however, LR-PCR with target lengths over 5kb is impossible for highly damaged and degraded endogenous DNA extracted from marginal forensic or archaeological samples. For example, in the range 126-179 bp used to investigate the control region in short overlapping fragments [78], one of our ancient DNA extracts (from individual HQU4) failed to consistently yield amplicons, but could nevertheless be placed into hg H7d5 using the whole mtDNA Mitochip v2.0 approach described herein.

Both theoretical considerations and empirical evidence indicate two opposing influences on library insert sizes during our mtDNA enrichment process. Hybridisation during the targeted DNA-capture step tends towards more efficient enrichment of longer mtDNA insert sequences, due to the increased stability of longer annealed hybrids between library and probe sequences at any given hybridisation, stringency wash temperatures and salt concentrations. In contrast, post-hybridisation library PCR reamplification preferentially amplifies smaller amplicons (i.e. library constructs with shorter aDNA sequence fragments inserted between library adaptors)[92]. As shown previously, damaged and degraded ancient DNA fragments are (on average) shorter than non-ancient molecules, meaning adaptor-tagged ancient mtDNA fragments will generally be strongly preferentially amplified over any present-day / 'modern' adaptor-tagged contaminant mtDNA present[82,86].

Between these 'push-and-pull' influences on library insert size, our hybridisation-based DNA-capture and library (adaptor) PCR reamplification protocols allowed us to enrich human mtDNA target sequences largely in the 20-70 bp range. Over 95% were within this size range based on a total of 232,347 reads from Pacific Biosystems data from six libraries enriched for human mtDNA sequences. Here, only 10666 (4.6%) were longer than 70 bp with a maximum read length reaching 158 bp.

The vast swathe of endogenous ancient human mtDNA fragments between ~20-to-50 bases targeted in this study would have been too short to be amplified directly by traditional PCR-based approaches (assuming typical forward and reverse PCR primers of ~25 bases). In contrast to traditional PCR-based approaches, therefore, hybridisation-based DNA-capture vastly expands both the potential aDNA template copy numbers available (via targeting <50 base aDNA molecules) and the mitochondrial genome coverage (via targeted enrichment using a whole human mitochondrial genome probe).

Endogenous ancient mtDNA molecules largely within the 20-70 bp range have no need for a fragmentation step prior to labelling as Mitochip v2.0 probe, as they are already ideally sized for a chip-based resequencing/genotyping approach based on hybridisation to the strand-specific 25-mers on the array. Assuming the exponential increase in copy number with decreasing target size applies, our approach therefore maximises the chances of assembling mt genome sequences from the most difficult, poorly preserved, samples[83,87]. However, DNA molecules fragmented to <20 bases are effectively beyond recovery and analysis, as fragments smaller than this cannot be captured efficiently by hybridisation or meaningfully tackled bioinformatically[93].

Following DNA extraction, library immortalization, DNA-capture enrichment for human mtDNA target sequences within these libraries, enriched library reamplification, and pre-labelling preparation, we therefore bypassed the fragmentation step in the Affymetrix CustomSeq Resequencing Array Protocol. Instead we directly labelled prepared libraries using the following protocol:

1. 4-6µg of enriched library DNA underwent biotin labelling using Terminal Deoxynucleotidyl Transferase (TDT) as per the Affymetrix GeneChip Whole Transcript Sense Target Labelling Assay Manual (P/N 701880, rev. 4).
2. Labelled samples were hybridised to Affymetrix GeneChip Human Mitochondrial Resequencing 2.0 Arrays for 17 hours at 49°C.
3. Arrays were washed, stained and scanned as per the GeneChip CustomSeq Resequencing Array Protocol (P/N 701231, rev. 5).
4. Affymetrix GeneChip Command Console software (v3.2) was used to generate CEL files, which were then analysed using GeneChip Sequence Analysis Software (GSEQ v4.1, Affymetrix).

An intrinsic aspect of resequencing arrays is an inevitable trade-off between call rate and accuracy of analyses[84,91,94]. Previous studies using both the Mitochip v2.0 and other resequencing microarrays have empirically optimised parameters. For example, after using quality score threshold settings (QST) of 2, 3, 6, 9, 12, and 30, Hartmann et al. (2008) settled on the haploid model with a QST of 3 and no reliability rule filter as the favoured parameters for their analysis of 93 worldwide (present-day) mitochondrial genomes[91]. However, although previous studies that used the Mitochip v2.0 and other resequencing microarrays are instructive and informative, there are at least two key aspects of our study that do not reflect any previous analyses by resequencing microarrays and which meant that we needed to empirically establish the most favourable parameters for ourselves.

First, in previous Mitochip studies 100% of the biotin-labelled input DNA corresponded to human mtDNA sequences from the sample-of-interest However, in our case, even with successive rounds of targeted DNA-capture enrichment of human mtDNA sequences, followed by library re-amplification, the final enrichment libraries used to generate probe will inevitably be composed of <100% human mtDNA sequences. In our case, cloning and next-generation library sequencing of enriched DNA libraries suggested that ~10-30% of input mt-enriched library probe onto chips is likely to correspond to unique/non-redundant ancient human mtDNA sequences. Second, as has been shown in many studies, the amplification of ancient DNA generates sequence changes due to miscoding lesion DNA damage[83,95-97]. This background of damage derived DNA sequence variation within the labelled Mitochip probe therefore added an additional layer of complexity.

Affymetrix Mitochip v2.0 validation for ancient DNA

The performance of the Mitochip v2.0 on ancient DNA libraries enriched for human mtDNA sequences was assessed by several criteria.

First, for all 39 archaeological individuals, the original DNA extracts from which the initial master and primary libraries were prepared, as well as duplicate extracts from separate bone/tooth samples from the same archaeological individuals (for replication), were tested independently over 413bp of HVSI via short overlapping PCR amplicons. In one case, an ancient DNA extract failed to yield every HVSI PCR amplicon (HQU4). In total, duplicate direct PCR and Sanger sequencing analyses of all 39 hg H individuals identified 16 SNP differences from the rCRS in HVSI (Supplementary Table S6). From the Mitochip runs using our mtDNA-enriched libraries as probes, 15 of these 16 SNPs were also identified for all levels, QST 1, 3, 6 and 12. One mutation (16213A in individual QUEVIII4) was called correctly for QST 1, but called as 'N' by QST 3, 6 and 12.

Second, 363 bp of HVSII were also amplified by PCR in a number of short overlapping fragments from 9 of the ancient DNA extracts and typed by Sanger sequencing as described previously[98]. Direct PCR and Sanger sequencing analyses of data from these individuals identified a further 13 SNP differences from the rCRS (ignoring unstable C-

stretch insertions; Supplementary Table S6). From the Mitochip v2.0 runs using our mtDNA-enriched libraries as probes, 12 of these SNPs were also identified for all levels, QST 1, 3, 6 and 12. One mutation (00152C from SALZ 57A) was called correctly for QST 1, but called as 'N' by QST 3, 6 and 12.

Third, an additional six (previously described) subhaplogroup-defining SNPs identified via the Mitochip v2.0 were also independently confirmed (Supplementary Table S6). All six SNPs were independently tested using newly designed primer pairs (Supplementary Table S5) via direct PCR from the extract used to make the original library (and/or the duplicate extract) followed by Sanger sequencing. All six confirmed the SNPs identified via the Mitochip v2.0 runs. This independent confirmation was particularly important for haplogroups defined by a single SNP and phylogenetic branching point, such as hg H23 (defined by 10211T, Supplementary Fig. S1) and hg H26 (defined by 11152C) for the Early Neolithic LBK individuals HAL36 and HAL32, respectively. In total, 35 out of 159 SNPs (31%) were directly sequenced.

Fourth, the 25-mer based Mitochip v2.0 resequencing array is based on the rCRS [80]. Due to historical developments, current haplogroup nomenclature places the rCRS in hg H2a2a[80,99,100]. This means that any human mt genome sequence outside hg H2 must necessarily have a minimum of 6 SNP variants compared to the rCRS: at nps 263, 8860, 15326 ('out of hg H2a2a'); np 750 ('out of hg H2a2'); np 4769 ('out of hg H2a'); and np 1438 ('out of hg H2'). Of the 39 hg H archaeological individuals investigated, 38 lie outside hg H2 (Table 1): with the remaining individual (BZH8) belonging to hg H2a1a3. With the two duplicate archaeological samples included (i.e. 41 Mitochip v2.0 runs), this means in total there should be (40 x 6) + (1 x 4) = 244 SNP variants to test the performance and accuracy of the Mitochip v2.0 when used with probes made from ancient DNA libraries enriched for human mtDNA sequences. Of these expected variants, 242/244 gave the correct expected calls for all levels, QST 1, 3, 6 and 12. Two samples, BZH1a and SALZ 57A, gave N at SNP 4769 for QST 1, 3, 6 and 12, due to missing data (i.e. the region including the SNP could not be read). However, the replicate sample of BZH1a (BZH1b) gave the correct 4769G at all levels, QST 1, 3, 6 and 12, showing that separate samples from the same archaeological individual can produce different quality outcomes, presumably due to both differential preservation of DNA and stochastic variation.

Fifth, in addition to these generic 'out of H2a2a' SNPs common across our sample set, there were also those pivotal diagnostic SNPs and hierarchies of SNPs that place archaeological individuals into particular established subgroups of hg H according to the most recent release of PhyloTree, Build 14 (Table 1)[100]. In total there are 159 SNP variants in this category. Of these 159 SNP variants, 96 correspond to previously characterised subhaplogroup-defining SNPs. Many of these Mitochip-identified SNPs belong to root-to-tip hierarchies of phylogenetic branching points. The observation of continuous root-to-tip chains, with no missing SNPs, is another key test of the accuracy and reliability of our strategy to type prehistoric mt genomes via the Mitochip v2.0.

For example, for archaeological individual ESP15, beginning with the rCRS (hg H2a2a) basis of the Mitochip v2.0, the hierarchical flow is thus:

- nps 263, 8860, 15326 (out of hg H2a2a);
- np 750 (out of hg H2a2)
- np 4769 (out of hg H2a)
- np 1438 (out of hg H2) – arriving at the root of hg H
- np 16362 (into pre-hg H6)
- nps 239, 16482 (into hg H6)
- np 3915 (into hg H6a)
- nps 4727, 9380 (into hg H6a1)

- np 11253 (into hg H6a1a)

All these key root-to-tip SNPs were successfully identified via our Mitochip v2.0 run. Moreover, SNPs defining derived twigs on the H6a1a branch at nps 5460 (hg H6a1a1) and 7325, 9362, 11611 and 16311 (hg H6a1a1a) were not detected. This means we can unambiguously place archaeological individual ESP15 within hg H6a1a at a basal position. It is important to bear in mind that there were nine phylogenetic branching points involved in placing individual ESP15 within hg H6a1a, involving 13 positive SNPs and dozens of exclusionary ones.

The remaining 63 (out of 159) SNPs are private, of which 25 are located in the coding region and 38 in the D-loop (Table 1). Lastly, 29 of the latter correspond to 16519C, which is a mutational 'hotspot' in evolutionary terms[101,102], although it appears to be stable and repeatable in individual terms, based on duplicated samples and sequence data derived independently from the Mitochip v2.0.

Overall then, the reliability of the Mitochip v2.0 data is confirmed by the fact that there are no missing links in the hierarchical phylogenetic flow and that there is only one root-to-tip route for every ancient individual. Over the whole sample set (plus duplicates) there were no absences of predicted SNP variants or a 'missing links' at previously characterised phylogenetic branching points. Every archaeological individual occupies a single position on the phylogenetic tree. Along with the independent confirmation of SNPs described above (and below), these data provide strong evidence that using ancient DNA libraries enriched for human mtDNA sequences as probes, the Mitochip v2.0 can generate robust and reliable mt genome sequences and can play a very useful role in the detailed classification of mt haplogroups from prehistoric human populations.

Sixth, six selected enriched libraries were run individually on the Pacific Biosciences platform as described below and SMRT reads were compared to results generated via the Mitochip v2.0. We deliberately chose samples that either had a high number of private SNPs, such as OSH2, BZH4, and BZH6 (four, four and six private SNPs, respectively) or from samples where we initially had observed a background level of SNPs that correspond to the probe (e.g. BZH1, DEB9 from the Oct 2010 batch). All private SNPs observed consistently in the Mitochip v2.0 data (at all QST levels) were confirmed by deep sequencing using the SRMT platform. The mean percentage of majority allele calls for private SNPs (86.2%) is not significantly different (i.e. lower) from the expected (i.e. previously characterised) SNPs (89.9%) when compared for these three samples ($p=0.109$; Nonparametric Wilcoxon Signed Ranks Test), indicating that the allele calls of private SNPs are as robust and reliable as ones that reflect examples previously characterised in the established nomenclature.

Affymetrix Mitochip v2.0 data analysis

We visually inspected all 16544 nucleotide positions of the GSEQ outputs at QSTs 1, 3, 6 and 12 as an alignment against the rCRS and in parallel examined trace views of probe intensities in the respective trace CEL files. Redundant probes tiled to cover most common control region variants were not analysed.

We identified and characterised regions, which were not, or only poorly, covered (Supplementary Table S3). These can largely be described in two categories, AT-rich regions and poly-cytosine stretches (C-stretches). It is well known that the hybridisation of AT-rich regions of DNA is less efficient due to the lower melting temperature and potential secondary DNA structure of AT-rich regions[103]. Over three consecutive rounds of hybridisation and enrichment, as well as during the final hybridisation to the Mitochip v2.0, certain restricted AT-rich regions were clearly selected against. In addition, even without prior enrichment/hybridisation rounds Hartmann et al.[91] observed that the vast majority of non-called nucleotides (N calls) occurred in regions with a run of ≥4Cs, within a 25-base window

of a mutation, and within a 25-base window of ≥15 As and Ts. By visual inspection of all samples using the probe intensity view, we could observe that for all C-stretch regions at least one direction of the tiling arrays (forward or reverse) was readable. We therefore used QST 6 outputs to manually reclaim C-stretch regions, and could decrease the number of N calls to an average of 248, resulting in a final call rate of 98.5% across all samples. This corresponds to a mean coverage of 98.5% of the mt genome considering the 12 bases at the beginning and end of the mt genome that are not interrogated by the Mitochip v2.0 (Supplementary Table S2). Judging from the probe intensity view in GSEQ, long AT-rich regions are typically poorly covered in either direction, indicating that these regions are successively lost/decreased during hybridisation rounds (including to the Mitochip), whereas C-stretches are well covered, but can generate readability difficulties on one strand within the 25 base window that is interrogated by the Mitochip v2.0 tiling array (Supplementary Fig. S2).

We realised that by individual visual inspection of probe intensities and by applying signal intensity thresholds, the number of remaining N or discordant calls could be further reduced, as had been described recently[104]. In our case, however, under the assumption that the final round enriched library does not only contain 100% human mtDNA sequences (and damage derived base changes in a sub-fraction of these), but also a proportion of potentially similar microbial analogous sequences, we chose a conservative approach and kept QST 6 outputs, that were manually corrected for C-stretch region only, as final calls.

The final designation of mitochondrial genome sequences into sub-haplogroups of hg H – via identified SNP differences to the rCRS – was therefore the result of several lines of evidence: the Mitochip v2.0 runs themselves (with a comparison between the QST outputs at 1, 3, 6 and 12); independent direct PCR and Sanger sequencing of previously characterised and novel 'private' SNPs from DNA extracts and via SMRT sequencing; and a final phylogenetic validation using the – at present – largest available dataset for hg H mt genomes (provided by DMB and LQ). The rationale of the PhyloTree database is to list subgroup-defining SNPs if they have been observed at least twice from two independent, unrelated sequences/individuals. Most of the end-tip sequences on the mt phylogeny contain a number of additional private SNPs that are not listed in the tree, but are potentially informative for a higher resolution pending future confirmation as additional sequences become available. In the light of the ever-growing nature of this database, the status of these private SNPs will remain provisional for now. However, the fact that all 16 private SNPs that were observed on the Mitochip v2.0 could unambiguously be confirmed by next generation sequencing, allows us to conclude that the detection of private SNPs can be relied upon when sufficiently high QSTs are applied to Mitochip-derived data. CEL files are available upon request and will be made available at http://www.adelaide.edu.au/acad/publications/data.

Pacific Biosciences SMRT® sequencing and data analyses

Six of the mt-enriched DNA libraries were converted to SMRTbell template libraries for sequencing on a Pacific Biosciences *RS*. First, the samples were blunted by incubating 1µg of enriched amplicons with 1X Template Prep Buffer (Pacific Biosciences, Menlo Park, CA), 1mM ATP, 0.4mM dNTPs, 0.2mg/mL rabbit serum albumin, 20U T4 Polynucleotide Kinase (NEB, Ipswich, MA), 4.5U T4 DNA Polymerase (NEB) and water to 40µL for 15min at 25°C. The blunted samples were then purified with GenCatch spin columns (Epoch Life Science) and A-tailed. To A-tail, samples were incubated in solutions containing the blunted amplicons, 1x Template Prep Buffer, 0.4mM dATP, 8U Klenow Fragment exo– (NEB) and water to 40µL for 60min at 37°C. After tailing, the samples were GenCatch purified and circularised by ligation to hairpin adaptors. Ligations were carried out in reactions containing the A-tailed amplicons, 1X Fermentas T4 Ligase Buffer, 60U T4 Ligase (Fermentas), 0.5µM T-tailed hairpin adaptor (Pacific Biosciences), and water to 60µL. All ligations were

incubated at 25°C for 16h and then heated at 65°C for 10min to inactivate the ligase. Each sample was then treated with 20U of Exonuclease I (NEB) and 10U Exonuclease III (NEB) for 60min at 37°C to digest linear, non-circularised, DNA molecules. Finally, the samples were GenCatch purified and an aliquot of each was assayed on a Bioanalyzer using a High Sensitivity DNA chip.

Single-Molecule, Real-Time (SMRT) sequencing was carried out on the PacBio *RS* at the Pacific Biosciences lab (Menlo Park, USA). Reads were processed and mapped to the respective reference sequences for each plasmid using the Basic Local Alignment with Successive Refinement (BLASR) mapper (http://www.pacbiodevnet.com/smrtanalysis/software/blasr) and the Pacific Biosciences SMRT Analysis pipeline using the standard mapping protocol (http://www.pacbiodevnet.com/smrtanalysis/software/smrtpipe). Circular consensus reads[105] were trimmed to remove the library PCR adapter sequences before mapping to the rCRS (NC_012920) using a maxScore = -50 filter and exported as SAM.files (Supplementary Table S4). At ACAD, SAM.files were imported into Geneious Pro 5.5.6[106] and re-assembled to the rCRS discarding reads <20 bp in length and using a conservative mapping method. Custom sensitivity settings were: 10% maximum gaps per read and a maximum gap size of 3, 20 bases minimum overlap with a word length of 14 (ignoring repeats >5 times) and index word length of 12, 10% maximum mismatches per read and a maximum ambiguity of 4 bases (Supplementary Table S4).

Mitochondrial DNA contamination estimates

To estimate the ratio of endogenous human mtDNA, we calculated the average rate of redundancy at expected SNP sites, i.e. we counted the number of reads that correspond to the allele (or the SNP variant) present in the Mitochip data from the same individual (Supplementary Data). Average results across all six samples indicate that the majority of the reads (83.8%) were human mtDNA sequences consistently derived from a single endogenous source. This number increases when two samples containing traces of probe DNA sequence are removed (87.4%). The second largest fraction of the reads (11%, and 10.6% for four samples) corresponds to any other states (transitions, transversions and indels) at these particular nps, indicating varying levels of: a) potential background contamination of exogenous human mtDNA; b) post-mortem miscoding lesion DNA damage; and/or c) read inaccuracies. Finally, we also calculated the proportion of potential contamination with the probe by checking the 29 SNPs that distinguish our probe (haplotype J1c8) from the rCRS. Of these, six SNPs leading out of the H2 branch are shared with all archaeological individuals (except individual BZH8), and SNP 3010A is shared mutation between subhaplogroups J1 and H1 and are therefore not informative, leaving 22 nps that could be checked for the presence of reads that share an identical allele with the probe. The average proportion of reads that resemble the probe is 1.9% across all four samples, for which we did not have an indication from the Mitochip v2.0 data, and higher (5.1%) when BZH1 and DEB9 were included. This proportion can be considered a maximum, since the coverage is not evenly spread across the mt genome and appears random, and secondly, DNA damage, esp. C>T and G>A transitions could mimic read inaccuracies. In addition, the potential contribution of the probe is significantly lower than the average background ($p=0.028$, Nonparametric Wilcoxon Signed Ranks test), therefore not affecting the overall allele call and the final haplogroup designation. With updated probe preparation and mt-enriched library recovery protocols (see above), probe background was reduced to undetectable levels.

SAM and Assembly files are available upon request and will be made available at http://www.adelaide.edu.au/acad/publications/data.

Haplogroup designation according to the "Copernican" reassessment of the human mitochondrial tree

Access to complete mt genome data from present-day samples from two recent studies[107,108] has provided the basis for the phylogenetic network and demographic reconstructions of mtDNA hg H for this study. Several thousand novel mt genomes have extended the human mtDNA phylogeny and have deepened the resolution of hg H to 87 sub-hg. At the same time, during the final stages of this study, a reassessment of the mtDNA tree from its root was published, that recommended the use of the Reconstructed Sapiens Reference Sequence (RSRS)[107] over the traditional use of the revised Cambridge Reference Sequence (rCRS)[80]. We therefore report our ancient mt genome data using the new RSRS, anticipating a rapid reception of the new reference and reporting formats (Table 1). However, since our study was conducted and validated against the rCRS, we describe the Mitochip validation in the traditional reporting format.

When compared to the phylogenetic root (RSRS) all our samples display the following sequence variants leading into branch hg H (in numerical order): G73A (back mutation A73G in EUL41A), C146T, C152T (back mutation T152C! in KAR11B, SALZ57A, EUL57B, OSH1), C182T/T182C!, C195T, (back mutation T195C! in BZH14), A247G, 522.1AC, A769G, A825t, A1018G, G2706A, A2758G, C2885T, T3594C, G4104A, T4312C, T7028C, G7146A, T7256C, A7521G (not covered due to AT-rich region), T8468C, T8655C, G8701A, (back mutation A8701G! in Sardinia), C9540T, G10398A, T10664C, A10688G, C10810T, C10873T, C10915T, A11719G, A11914G, T12705C, G13105A, G13276A, T13506C, T13650C (T13650N in BZH1, HAL32, HAL39, ALB1, HQU4, OSH1, QUEXII6, QUEVIII4 due to C-stretch), T14766C, G15301A/A15301G!, A16129G, (back mutation G16129A! in HAL11), T16187C, (T16187N in HAL11, ROT1, OSH3, DEB9, KAR6A, QUEXII6 due to C-stretch), C16189T (back mutation T16189C! in BZH6), T16223C, G16230A, T16278C and C16311T.

Supplementary References

61. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science (New York, N.Y.)* **326**, 137-40 (2009).
62. Alvarez-Iglesias, V. *et al.* New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS ONE* **4**, e5112 (2009).
63. Garcia, O. *et al.* Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge. *Heredity* **106**, 37-45 (2011).
64. Roostalu, U. *et al.* Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: The near eastern and Caucasian perspective. *Molecular Biology and Evolution* **24**, 436-448 (2007).
65. Brandstätter, A. *et al.* Dissection of mitochondrial superhaplogroup H using coding region SNPs. *Electrophoresis* **27**, 2541-50 (2006).
66. Loogväli, E.L. *et al.* Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Molecular Biology and Evolution* **21**, 2012-2021 (2004).
67. Schwarz, R. Chronologie der neolithischen Kulturen in Mitteldeutschland (Abfolge und Verbreitung). in *Kataloge zur Dauerausstellung im Landesmuseum für Vorgeschichte Halle, 2. Neolithikum* (ed. Meller, H.) (Halle/Saale, in the press).
68. Bocquet-Appel, J.P., Naji, S., Vander Linden, M. & Kozlowski, J.K. Detection of diffusion and contact zones of early farming in Europe from the space-time distribution of 14C dates. *Journal of Archaeological Science* **36**, 807-820 (2009).
69. Shennan, S. & Edinborough, K. Prehistoric population history: from the late glacial to the late neolithic in central and northern Europe. *Journal of Archaeological Science* **34**, 1339-1345 (2007).
70. Price, T.D. *Europe's first farmers*, 395 (Cambridge University Press, Cambridge, 2000).
71. Whittle, A.W.R. & Cummings, V. *Going over: the mesolithic-neolithic transition in North-West Europe*, 632 (Oxford University Press, Oxford, 2007).
72. Gronenborn, D. A variation on a basic theme: the transition to farming in southern central Europe. *J.World Prehistory* **13**, 123-210 (1999).
73. Bogucki, P.I. & Crabtree, P.J. *Ancient Europe 8000 B.C.--A.D. 1000 : Encyclopedia of the Barbarian World*, 1221p. (Charles Scribner's Sons, 2004).
74. Case, H. Beakers and Beaker Culture. in *Beyond Stonehenge. Essays in honour of Colin Burgess* (eds. Burgess, C., Topping, P. & Leach, F.) (Oxford, 2007).
75. Brandt, G., Knipper, C., Roth, C., Siebert, A. & Alt, K.W. Beprobungsstrategien für aDNA und Isotopenanalysen an historischem und prähistorischem Skelettmaterial. in *Anthropologie, Isotopie und DNA, 2. Mitteldeutscher Archäologentag* 3 edn (eds Meller, H. & Alt, K.W.) 17-32 (Landesmuseum for Vorgeschichte Halle (Saale), Halle/Saale, 2010).
76. Fu, Q., Rudan, P., Pääbo, S. & Krause, J. Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe. *PLoS ONE* **7**, e32473. doi:10.1371/journal.pone.0032473 (2012).
77. Haak, W. *et al.* Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. *PLoS Biology* **8**, e1000536 (2010).
78. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science (New York, N.Y.)* **310**, 1016-8 (2005).
79. Taylor, R.W., Taylor, G.A., Durham, S.E. & Turnbull, D.M. The determination of complete human mitochondrial DNA sequences in single cells: implications for the

study of somatic mitochondrial DNA point mutations. *Nucleic Acids Research* **29**, E74-4 (2001).
80. Andrews, R.M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat.Genet.* **23**, 147 (1999).
81. Patel, M. & Sive, H. PCR-based subtractive cDNA cloning. *Current protocols in molecular biology* **Chapter 25**, Unit 25B 2 (2001).
82. Tao, S.C., Gao, H.F., Cao, F., Ma, X.M. & Cheng, J. Blocking oligo--a novel approach for improving chip-based DNA hybridization efficiency. *Molecular and cellular probes* **17**, 197-202 (2003).
83. Brotherton, P. *et al.* Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res* **35**, 5717-28 (2007).
84. Zhou, S. *et al.* An oligonucleotide microarray for high-throughput sequencing of the mitochondrial genome. *The Journal of molecular diagnostics : JMD* **8**, 476-82 (2006).
85. Paabo, S. *et al.* Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645-79 (2004).
86. Brotherton, P., Sanchez, J.J., Cooper, A. & Endicott, P. Preferential access to genetic information from endogenous hominin ancient DNA and accurate quantitative SNP-typing via SPEX. *Nucleic Acids Research* **38**, e7 (2010).
87. Adler, C.J., Haak, W., Donlon, D. & Cooper, a. Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science* **38**, 956-964 (2011).
88. Malmström, H. *et al.* More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human DNA. *Mol Biol Evol* **24**, 998-1004 (2007).
89. Maitra, A. *et al.* The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Research* **14**, 812-9 (2004).
90. Mithani, S.K. *et al.* Mitochondrial resequencing arrays detect tumor-specific mutations in salivary rinses of patients with head and neck cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 7335-40 (2007).
91. Hartmann, A. *et al.* Validation of Microarray-Based Resequencing of 93 Worldwide Mitochondrial Genomes. *Human Mutation* **30**, 115-122 (2009).
92. Walsh, D.J. *et al.* Isolation of deoxyribonucleic acid (DNA) from saliva and forensic science samples containing saliva. *Journal of forensic sciences* **37**, 387-95 (1992).
93. Green, R.E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330-6 (2006).
94. Bruce, C.K. *et al.* Design and validation of a metabolic disorder resequencing microarray (BRUM1). *Human Mutation* **31**, 858-65 (2010).
95. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Paabo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* **29**, 4793-9 (2001).
96. Hansen, A.J., Willerslev, E., Wiuf, C., Mourier, T. & Arctander, P. Statistical Evidence for Miscoding Lesions in Ancient DNA Templates. *Mol.Biol.Evol.* **18**, 262-265 (2001).
97. Briggs, A.W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A* **104**, 14616-21 (2007).
98. Haak, W. *et al.* Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. *Proc Natl Acad Sci U S A* **105**, 18226-31 (2008).

99. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465 (1981).
100. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* **30**, E386-94 (2009).
101. Bandelt, H.J., Kong, Q.P., Richards, M. & Macaulay, V. Estimation of mutation rates and coalescence times: some caveats. in *Human Mitochondrial DNA and the Evolution of Homo Sapiens* (eds. Bandelt, H.J., Macaulay, V. & Richards, M.) 47-90 (Springer, Berlin, 2006).
102. Behar, D.M. *et al.* The Genographic Project public participation mitochondrial DNA database. *PLoS Genet* **3**, e104 (2007).
103. Strachan, T. & Read, A.P. *Human Molecular Genetics*, (Wiley-Liss, New York, 1999).
104. Xie, H.M. *et al.* Mitochondrial genome sequence analysis: A custom bioinformatics pipeline substantially improves Affymetrix MitoChip v2.0 call rate and accuracy. *BMC bioinformatics* **12**, 402 (2011).
105. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* **38**, e159 (2010).
106. Drummond, A.J. *et al.* Geneious v5.4, Available from http://www.geneious.com/. (2011).
107. Behar, D.M. *et al.* A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. *American Journal of Human Genetics* **90**, 675-684 (2012).
108. Behar, D.M. *et al.* The Basque Paradigm: Genetic Evidence of a Maternal Continuity in the Franco-Cantabrian Region since Pre-Neolithic Times. *American Journal of Human Genetics* **90**, 486-493 (2012).